Statistique descriptive

Introduction :

La statistique est une méthode scientifique qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à analyser, à commenter et à critiquer ces données. Il ne faut pas confondre *la* statistique qui est la science qui vient d'être définie et *une* statistique qui est un ensemble de données chiffrées sur un sujet précis.

Les premières statistiques correctement élaborées ont été celles des recensements démographiques. Ainsi le vocabulaire statistique est essentiellement celui de la démographie.

Les ensembles étudiés sont appelés population. Les éléments de la population sont appelés individus ou unités statistiques. La population est étudiée selon un ou plusieurs caractères.

Les statistiques descriptives peuvent se résumer par le schéma suivant :

Echantillonnage statistique :

Pour recueillir des informations sur une population statistique, l'on dispose de deux méthodes :

- La méthode exhaustive ou recensement où chaque individu de la population est étudié selon le ou les caractères étudiés.
- La méthode des sondages ou échantillonnage qui conduit à n'examiner qu'une fraction de la population, un échantillon.

Définition :

L'échantillonnage représente l'ensemble des opérations qui ont pour objet de prélever un certain nombre d'individus dans une population donnée.

Pour que les résultats observés lors d'une étude soient généralisables à la population statistique, l'échantillon doit être représentatif de cette dernière, c'est à dire qu'il doit refléter fidèlement sa composition et sa complexité. Seul l'échantillonnage aléatoire assure la représentativité de l'échantillon.

Un échantillon est qualifié d'aléatoire lorsque chaque individu de la population a une probabilité connue et non nulle d'appartenir à l'échantillon.

Le cas particulier le plus connu est celui qui affecte à chaque individu la même probabilité d'appartenir à l'échantillon.

Echantillonnage aléatoire simple :

L'échantillonnage aléatoire simple est une méthode qui consiste à prélever au hasard et de façon indépendante, n individus ou unités d'échantillonnage d'une population à N individus.

Chaque individu possède ainsi la même probabilité de faire partie d'un échantillon de n individus et chacun des échantillons possibles de taille n possède la même probabilité d'être constitué.

L'échantillonnage aléatoire simple assure l'indépendance des erreurs, c'est-à-dire l'absence d'autocorrélations parmi les données relatives à un même caractère. Cette indépendance est indispensable à la validité de plusieurs tests statistiques.

Exemple :

Les données météorologiques ne sont pas indépendantes puisque les informations recueillies sont d'autant plus identiques qu'elles sont rapprochées dans le temps et dans l'espace. *

Il existe d'autres techniques d'échantillonnage que nous ne développerons pas dans un premier temps dans ce cours comme l'échantillonnage systéma-

tique ou l'échantillonnage stratifié qui répondent à des problématiques biologiques spécifiques.

Les caractères statistiques

Définition :

On appelle caractère statistique simple toute application : $\boldsymbol{X}: \boldsymbol{P} \longrightarrow \mathbb{R}$ avec ***P*** un ensemble fini appelé **population** ; tout élément $\boldsymbol{\omega}$ de ***P*** s'appelle un **individu**

Définition :

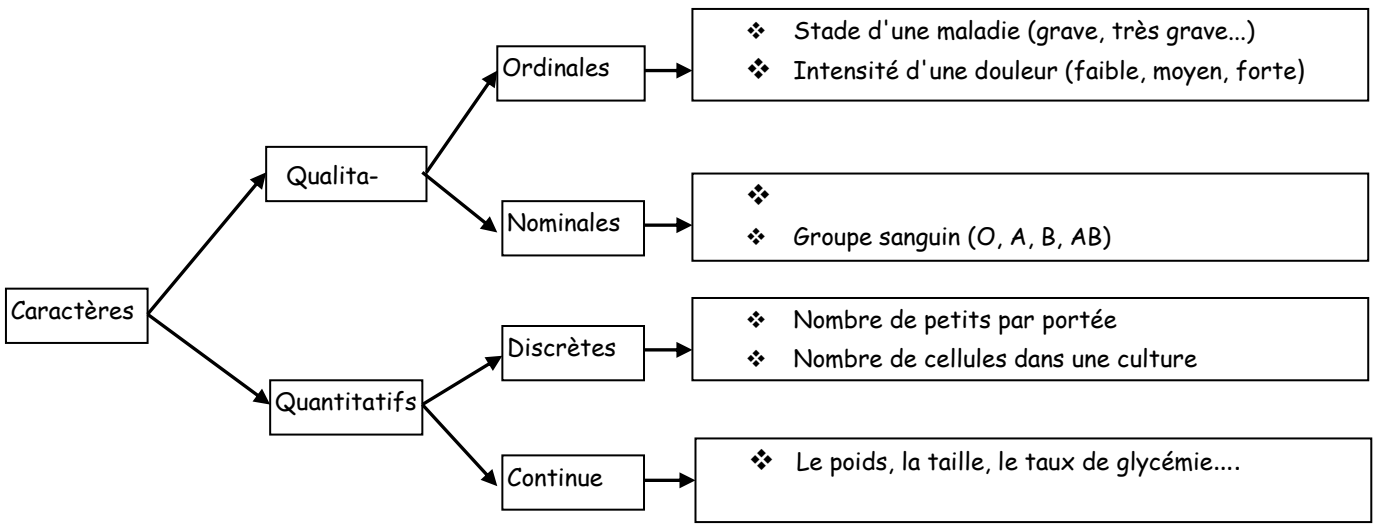
On appelle **modalité** (نمط) toute valeur : $x_i \in X(P)$ telle que $\boldsymbol{X(P)} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3} \ldots, \boldsymbol{x_i}, \ldots \boldsymbol{x_k}\}$ avec k nombre de modalités différentes de *X*.

Remarque : seuls les caractères quantitatifs ont valeurs dans $\mathbb{R}$, les caractères qualitatifs s'y ramenant par un codage.

Exemple :

Lors des recensements, les caractères étudiés sont l'âge, le sexe, la qualification professionnelle, etc. Le caractère « sexe » présente deux modalités alors que pour la qualification professionnelle, le nombre de modalités va dépendre de la précision recherchée.

On peut schématiser les caractères par :



- Qualitatif, c'est-à-dire non numérique
- Quantitatif discret (nombre fini de valeurs)
  continu (infinité de valeurs)

Liens avec les concepts probabilistes :

Les concepts qui viennent d'être présentés sont les homologues de concepts du calcul des probabilités et il est possible de disposer en regard les concepts homologues (voir table ci-dessous).

probabilités	statistique
Espace fondamental (univers)	Population
Épreuve	Tirage (d'un individu), expérimentation
Évènement élémentaire	Individu, observation
Variable aléatoire	Caractère (variable statistique)
Épreuves répétées	Échantillonnage
Nombre de répétition d'une épreuve	Taille de l'échantillon, effectif total
probabilité	Fréquence observé
Loi de probabilité	Distribution observée ou loi empirique
Espérance mathématique	Moyenne observée
Variance	Variance observée

Représentation des données

Il existe plusieurs niveaux de description statistique : la représentation brute des données, des présentations par tableaux numériques, des représentations graphiques et des résumés numériques fournis par un petit nombre de paramètres caractéristiques.

Séries statistiques

Une série statistique correspond aux différentes modalités d'un caractère sur un échantillon d'individus appartenant à une population donnée. Le nombre d'individus qui constituent l'échantillon étudié s'appelle la taille de l'échantillon.

Exemple :

Afin d'étudier la structure de la population de **gélinottes huppées** (Bonasa *umbellus*) abattues par les chasseurs canadiens, une étude du **dimorphisme sexuel** (الفروق بين الذكر والأنثى من نفس النوع) de cette espèce a été entreprise. Parmi les caractères mesurés figure **la longueur de la rectrice centrale** (plume de la queue). Les résultats observés exprimés en millimètres sur un échantillon de **50** mâles juvéniles (الذكور الأحداث) sont notés dans la série ci-dessus :

La gélinotte huppée

153	165	160	150	159	151	163
160	158	149	154	153	163	140
158	150	158	155	163	159	157
162	160	152	164	158	153	162
166	162	165	157	174	158	171
162	155	156	159	162	152	158
164	164	162	158	156	171	164
158						

Tableaux statistiques :

Le **tableau de distribution de fréquences** est un mode synthétique (طريقة مزج) de présentation des données. Sa constitution est immédiate dans le cas d'un caractère discret mais nécessite en revanche une transformation des données dans le cas d'un caractère continu.

Fréquences absolues, relatives et cumulées :

A chaque **modalité** x_i du caractère X, peut correspondre un ou plusieurs individus dans l'échantillon de taille n.

- On appelle **effectif (fréquence absolue)** de la modalité x_i, le nombre n_i où n_i est le nombre d'individu ω tel que $X(\omega) = x_i$
- On appelle **fréquence (fréquence relative)** de la modalité x_i, le nombre f_i tel que $f_i = \frac{n_i}{n}$
- Le **pourcentage** est une fréquence exprimée en pour cent. Il est égal à **100** f_i.
- On appelle **effectif cumulé croissante** de la modalité x_i, la somme des effectifs partiels des valeur

$$x_1, x_2, \dots, x_i \quad n_i^{\nearrow} = n_1 + n_2 + \cdots + n_i = \sum_{k=1}^{i} n_k$$

- On appelle **effectif cumulé décroissante** de la modalité x_i le nombre

$$n_i^{\searrow} = n - (n_1 + n_2 + \cdots n_{i-1}) = \sum_{p=1}^{p=i-1} n_p$$

$$n_i^{\searrow} = n_1 + n_2 + \cdots n_{i-1} + n_i + \cdots + n_k - (n_1 + n_2 + \cdots n_{i-1}) = n_i + \cdots + n_k$$

- On appelle **fréquences cumulées ou fréquences relatives cumulées** en x_i, le nombre $f_i cum$ tel que

$$f_i cum = \sum_{p=1}^{p=i} f_p$$

Remarque :

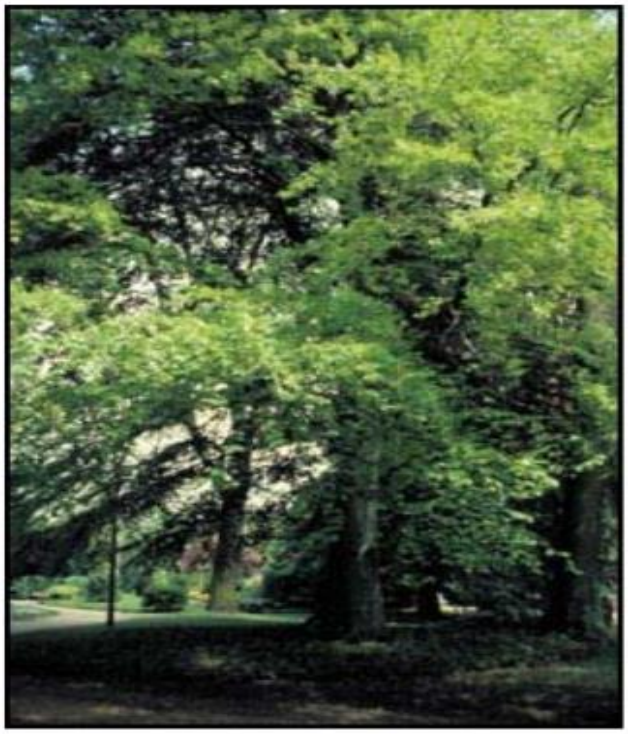
$$\sum_{i=1}^{i=k} n_i = n \qquad \sum_{i=1}^{k} f_i = 1$$

Caractères quantitatifs discrets

Dans le cas d'un **caractère quantitatif discret**, l'établissement de la distribution des données observées associées avec leurs fréquences est immédiate.

Exemple : La **cécidomyie** du hêtre (ذبابة شجرة الزان) provoque sur les feuilles de cet arbre des galles dont **la distribution de fréquences observées** est la suivante :

Caractère X x_i: *nombre de galles par feuille*	0	1	2	3	4	5	6	7	8	9	10	∑
n_i: *nombre de feuille portant* x_i *galles*	182	98	46	28	12	5	2	1	0	1	0	375
f_i: *fréquence relative*	0,485	0,261	0,123	0,075	0,032	0,013	0,005	0,003	0	0,003	0	1
f_i *cum: fréquence relative cumulée*	0,485	0,746	0,869	0,944	0,976	0,0989	0,994	0,997	0.997	1	1	



Caractères quantitatifs continus

Dans le cas d'un caractère quantitatif continu, l'établissement du tableau de fréquences implique d'effectuer au préalable **une répartition en classes** des

données. Cela nécessite de définir le nombre de classes attendu et donc l'amplitude associée à chaque classe ou **intervalle de classe**.

En règle générale, on choisit des classes de même **amplitude**. Pour que la distribution en fréquence est un sens, il faut que chaque classe comprenne un nombre suffisant de valeurs (n_i).

Diverses formules empiriques permettent d'établir le **nombre de classes** pour un échantillon de taille *n*.

- La règle de **STURGE** : Nombre de classes = **1+ 3,3 *log n***
- La règle de **YULE :** Nombre de classes = $2,5\sqrt[4]{n} = 2,5\sqrt{\sqrt{n}}$
- L'intervalle entre chaque classe est obtenu ensuite de la manière suivante :

Intervalle de classe = (X_{max} - X_{min}) / *Nombre de classes*

Avec X_{max} et X_{min}, respectivement la plus grande et la plus petite valeur de *X* dans la série statistique.

A partir de X_{min} on obtient les limites de classes ou **bornes de classes** par addition successive de l'intervalle de classe. En règle générale, on tente de faire coïncider **l'indice de classe** ou valeur centrale de la classe avec un nombre entier ou ayant peu de décimales.

Exemple :

Dans le cadre de l'étude de la population de **gélinottes_huppées** (Bonasa *umbellus*), les valeurs de la longueur de la rectrice principale peuvent être réparties de la façon suivante :

- **Définition du nombre de classes :**

Règle de Sturge : 1 + 3,3 log 50 = 6,60

Règle de Yule : $2,5\sqrt[4]{50} = 6,64$ les deux valeurs sont très peu différentes

- **Définition de l'intervalle de classe :** $IC = \frac{174-140}{6.6} = 5,15\ mm$ que l'on arrondit à 5 mm par commodité (facilité)

Tableau de distribution des fréquences

Caractère X : longueur	[140, 145[	[145, 150[	[150, 155[	[155, 160[	[160, 165[	[165, 170[	[170, 175[

de la rectrice x_i *: bornes des classes*							
c_i *:Valeurs médianes des classes (centres des classes)*	142,5	147,5	152,5	157,5	162,5	167,5	172,5
n_i *: nombre d'individu par classe*	1	1	9	17	16	3	3
N_i *: effectif cumulée*	1	2	11	28	44	47	50
f_i *: fréquence relative*	0,02	0,02	0,18	0,34	0,32	0,06	0,06
f_i **cum.** *: fréquence relative cumulée*	0,02	0,04	0,22	0,56	0,88	0,94	1

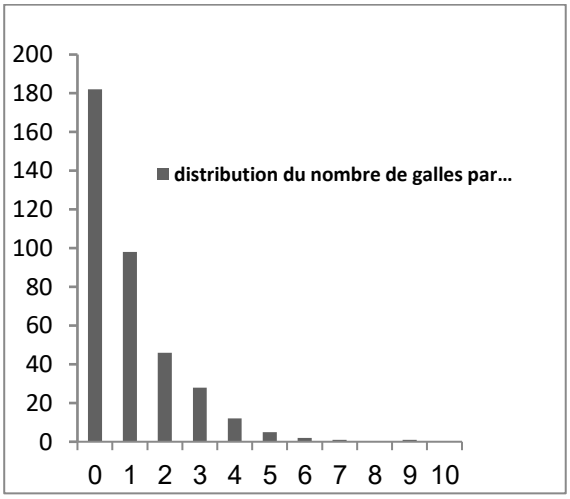
Représentations graphiques

Les représentations graphiques ont l'avantage de renseigner immédiatement sur l'allure générale de la distribution. Elles facilitent l'interprétation des données recueillies.

Caractères quantitatifs discrets

Pour les caractères quantitatifs discrets, la représentation graphique est le **diagramme en bâtons** où la hauteur des bâtons correspond à l'effectif n_i associé à chaque modalité du caractère x_i.

Exemple

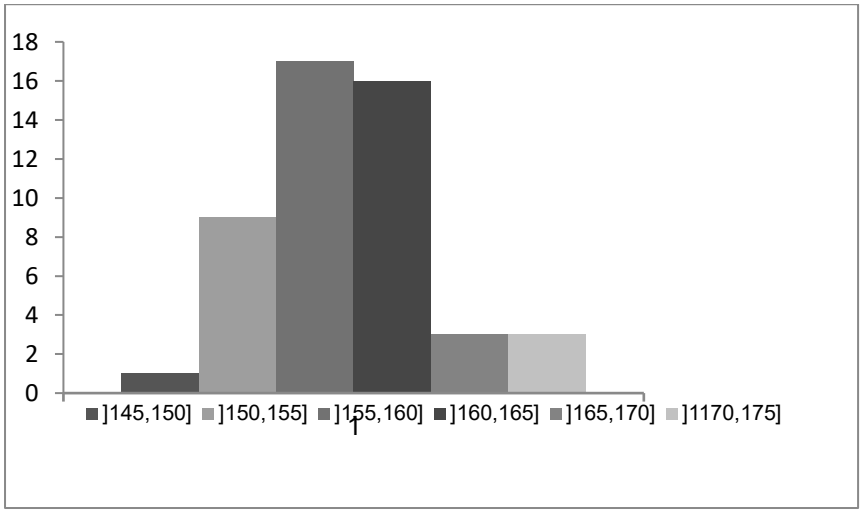


Caractères quantitatifs continus

Pour les caractères quantitatifs continus, la représentation graphique est **l'histogramme** où la hauteur du rectangle est proportionnelle à l'effectif n_i.

Ceci n'est vrai que si l'intervalle de classe est constant. Dans ce cas l'aire comprise sous l'histogramme s'avère proportionnelle à l'effectif total. En revanche lorsque les intervalles de classe sont inégaux, des modifications s'imposent pour conserver cette proportionnalité. Dans ce cas, en ordonnée, au lieu de porter l'effectif, on indique le rapport de la fréquence sur l'intervalle de classe. Ainsi la superficie de chaque rectangle représente alors l'effectif associé à chaque classe.

Exemple



Indicateurs numériques

Le dernier niveau de description statistique est le résumé numérique d'une distribution statistique par des **indicateurs numériques** ou **paramètres caractéristiques**.

Indicateurs de position

Ces paramètres ont pour objectif dans le cas d'un caractère quantitatif de caractériser **l'ordre de grandeur** des observations.

La moyenne arithmétique

Soit un échantillon de n valeurs observées $x_1, x_2, \ldots, x_i, \ldots, x_n$ d'un caractère quantitatif X, on définit sa moyenne observée comme la moyenne arithmétique des n valeurs :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Remarque :

$$\sum_{}^{n}(x_i - \bar{x}) = 0$$

Si les données observées x_i sont regroupées en k classes d'effectif n_i (caractère continu regroupé en classe ou caractère discret), il faut les pondérer par les effectifs correspondants :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i \quad avec \ n = \sum_{i=1}^{k} n_i$$

Exemple :

Dans le cas de l'étude du dimorphisme sexuel de la **gélinotte huppée**, la longueur moyenne de la rectrice principale du mâle juvénile est :

- Dans le cas des **données non groupées** :

$$\bar{x} = \frac{153 + 165 + 1609 \ldots + 158}{50} = 158{,}9\ mm$$

- Dans le cas des **données groupées** où les valeurs x_i correspondent aux valeurs médianes des classes,

$$\sum_{i=1}^{n} n_i x_i = 7960 \ d'ou \ \bar{x} = \frac{7960}{50}$$

La médiane :

La médiane, M_e, est la valeur du caractère pour laquelle la fréquence cumulée est égale à 0,5 ou 50%. Elle correspond donc au centre de la série statistique classée par ordre croissant, ou à la valeur pour laquelle 50% des valeurs observées sont supérieures et 50% sont inférieures.

- Cas discret
  - $Si\ n = 2m + 1 (impair) \qquad M_e = x_{m+1}$
  - $Si\ n = 2m \quad (pair) \qquad M_e = \frac{x_m + x_{m+1}}{2}$
- Cas continue

$$M_e = x_m + (x_{m+1} - x_m)\left[\frac{\frac{n}{2} - N_i}{n_i}\right]$$

Avec

x_m : limite inférieure de la classe dans laquelle se trouve le $n^e/2$ individu (classe médiane).
x_{m+1} : limite supérieure de la classe dans laquelle se trouve le $n^e/2$ individu (classe médiane).
n_i : effectif de la classe médiane
N_{i-} : effectif cumulé inférieur à x_m
n : taille de l'échantillon

Exemple :

Dans le cas de la distribution de la longueur de la rectrice centrale de la **gélinotte huppée**, la valeur de la médiane est :

- Cas des **données non groupées** :

$$n = 50 = 2 \times 25 \; alors, M_e = \frac{x_{25} + x_{26}}{2} = \frac{158 + 159}{2} = 158{,}5 \; mm$$

- Cas des **données groupées** :

$$\frac{n}{2} = 25 \; donc \; M_e \in [155{,}160[\,, \quad M_e = 155 + \frac{5}{17}\left(\frac{50}{2} - 11\right) = 159{,}11 \; mm$$

Remarque

Si la distribution des valeurs est **symétrique**, la valeur de la médiane est proche de la valeur de la moyenne arithmétique. $M_e \approx \bar{x}$

Le mode

Le mode, M_o d'une série statistique est la valeur du caractère la plus fréquente ou dominante dans l'échantillon. Le mode correspond à la classe de fréquence maximale dans la distribution des fréquences.

On peut identifier le mode comme la valeur médiane de la classe de fréquence maximale ou bien effectuer une interpolation linaire pour obtenir la valeur exacte du mode comme suit :

$$M_o = x_m + \frac{i\Delta i}{\Delta s + \Delta i}$$

avec

x_m: limite inférieure de la classe d'effectif maximal (Classe modale)
i: Intervalle de classe ($x_{m+1} - x_m$)

Δi: Écart d'effectif entre la classe modale et la classe inférieure la plus proche.
Δs: Écart d'effectif entre la classe modale et la classe supérieure la plus proche

Exemple :

Dans le cas de la distribution de la longueur de la rectrice centrale de la géli-notte huppée, la valeur du mode est :

- **Valeur approchée :** La classe de fréquence maximale est [155,160[avec n_i = **17** d'où M_o = **157,5 mm**
- **Valeur exacte :**

$$M_o = 155 + \frac{5 \times 8}{(1+8)} = 159.44\, mm$$

Avec

$x_m = 155mm,\ \Delta_i = 17 - 9 = 8, \quad \Delta_s = 17 - 16 = 1 et\ i = 5mm$

Remarque :

- Une distribution de fréquences peut présenter un seul mode (distribution unimodale) ou plusieurs modes (distribution bi ou trimodale).
- Si la distribution des valeurs est symétrique, la valeur du mode est proche de la valeur de la moyenne arithmétique. $M_o = \overline{x}$

Comparaison des indicateurs de position

	Avantages	Inconvénients
Moyenne arithmétique	♦ Facile à calculer, ♦ Répond au principe des moindres carrés.	♦ Fortement influencée par les valeurs extrêmes de la v.a., ♦ Représente mal une population hétérogène (polymodale).
Médiane	♦ Pas influencée par les valeurs extrêmes de la v.a., ♦ Peu sensible aux variations d'amplitude des classes, ♦ Calculable sur des caractères	♦ Se prête mal aux calculs statistiques, ♦ Suppose l'équi-répartition des données ♦ Ne représente que la valeur qui sépare l'échantillon en 2 parties égales.

	cycliques (saison, etc.) où la moyenne a peu de significa-tion.	
Mode	♦ Pas influencée par les valeurs extrêmes de la v.a., ♦ Calculable sur des caractères cycliques (saison, etc.) où la moyenne a peu de significa-tion, ♦ Bon indicateur de population hétérogène.	♦ Se prête mal aux calculs statistiques, ♦ Très sensible aux variations d'amplitude des classes, ♦ Son calcul ne tient compte que des indi-vidus dont les valeurs se rapprochent de la classe modale.

Indicateurs de dispersion

Ces paramètres ont pour objectif dans le cas d'un caractère quantitatif de caractériser **la variabilité des données** dans l'échantillon. Les indicateurs de dispersion fondamentaux sont **la variance observée** et **l'écart-type observé**.

La variance observée

Soit un échantillon de n valeurs observées $x_1, x_2, \ldots, x_i, \ldots, x_n$ d'un caractère quantitatif **X** et soit $\bar{x}$ sa moyenne observée. On définit la variance observée notée s^2 comme la moyenne arithmétique des carrés des écarts à la moyenne.

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Pour des commodités de calcul, on se sert du **théorème de Kœnig** que nous démontrons dans un cas particulier.

Voici pourquoi :

Soit

$$A = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$= \sum_{i=1}^{n} \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)$$

$$= \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n} \bar{x}^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 \; or \; \sum_{i=1}^{n} x_i = n\bar{x}$$

$$A = \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

La formule de la variance qui résulte du théorème de **Kœnig** est donc :

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

Dans le cas de **données regroupées en *k* classes d'effectif n_i** (variable continue regroupée en classes ou variable discrète), la formule de la variance est la suivante :

$$S^2 = \frac{1}{n}\sum_{i=1}^{i=k} n_i\left(x_i^2 - \bar{x}^2\right)$$

Pour des commodités de calcul, on utilisera la formule développée suivante

$$S^2 = \frac{1}{n}\sum_{i=1}^{i=k} n_i x_i^2 - \bar{x}^2 \; avec \; n = \sum_{i=1}^{i=k} n_i$$

L'**écart-type observé** correspond à la racine carrée de la variance observée :

$$S = \sqrt{S^2}$$

Exemple :

(1) Dans le cas de l'étude du dimorphisme sexuel de la gélinotte huppée, la variance observée de la longueur de la rectrice centrale du mâle juvénile est :

❖ **cas des données <u>non groupées :</u>**

$$\sum_{i=1}^{i=n} x_i^2 = 126347 \quad \text{et } \bar{x} = 158.86 \text{ mm} \quad ; \quad S^2 = \frac{1}{50}(1263647) - (158.86)^2$$

$$= 36{,}44 \quad \text{d'où } S^2 = 36{,}44 \text{ et S } = 6{,}04 \text{ mm}$$

❖ **cas des données <u>groupées</u>**

$$\sum_{i=1}^{i=n} n_i x_i^2 = 1269012.5 \quad \text{et } \bar{x} = 159.20 \text{ mm} \quad ; \quad S^2 = \frac{1}{50}(1269012.5) - (159.20)^2$$

$$= 35{,}61 \quad \text{d'où } S^2 = 35{,}61 \text{ et S } = 5.97 \text{ mm}$$

Exercice: Quelle est la variance et l'écart-type observée du nombre moyen de galles par feuille pour la <u>cécidomyie</u> du hêtre

Remarque

De part sa définition, l**a variance est toujours un nombre positif**. Sa dimension est le carré de celle de la variable. Il est toutefois difficile d'utiliser la variance comme mesure de dispersion car le recours au carré conduit à un changement d'unités.

Elle n'a donc pas de sens biologique direct contrairement à l'écart-type qui s'exprime dans les mêmes unités que la moyenne.

Le coefficient de variation

La **variance** et l'**écart-type observée** sont des paramètres de **dispersion absolue** qui mesurent la variation absolue des données indépendamment de l'ordre de grandeur des données.

Le **coefficient de variation** noté *C.V.* est un indice de **dispersion relatif** prenant en compte ce biais et est égal à :

$$CV = \frac{S}{\bar{x}} 100$$

Exprimé en pour cent, il est indépendant du choix des unités de mesure permettant la comparaison des distributions de fréquence d'unité différente.

Exemple :

Le coefficient de variation des longueurs de la rectrice centrale des gélinottes huppées mâles juvéniles est égal à :

$$CV = \frac{6.09}{158.86} 100 = 3.83\%$$

Quel est le coefficient de variation du nombre de galles par feuille pour la cécidomyie du hêtre ?

Paramètres de forme

Nous définissons les paramètres de forme pour une variable statistique quantitative, discrète ou
continue, à valeurs réelles.

Coefficient d'asymétrie.

a) Définition.

Il existe plusieurs coefficients d'asymétrie. Les principaux sont les suivants.
Le coefficient d'asymétrie de **Pearson** fait intervenir le mode Mo
: quand il existe, il est définie par

$$P = \frac{\overline{X} - M_o}{s(X)}.$$

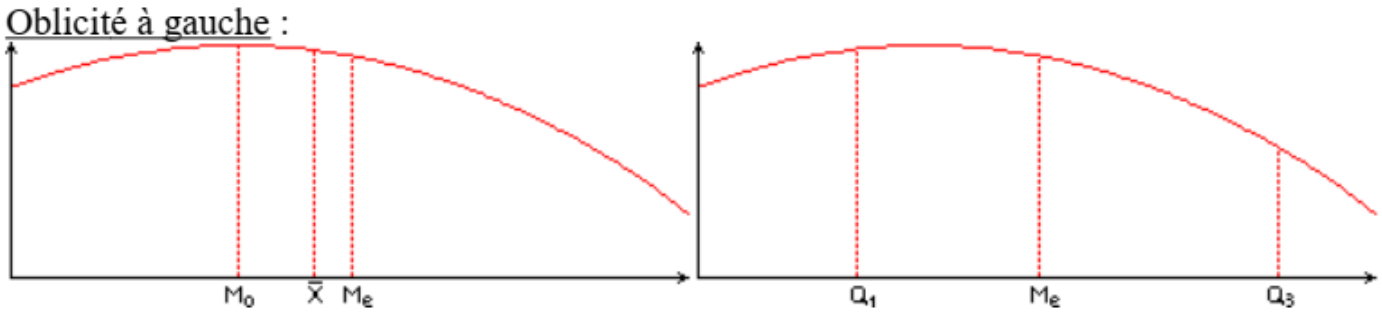
Le coefficient d'asymétrie de Yule fait intervenir la médiane et les quartiles, il est défini par

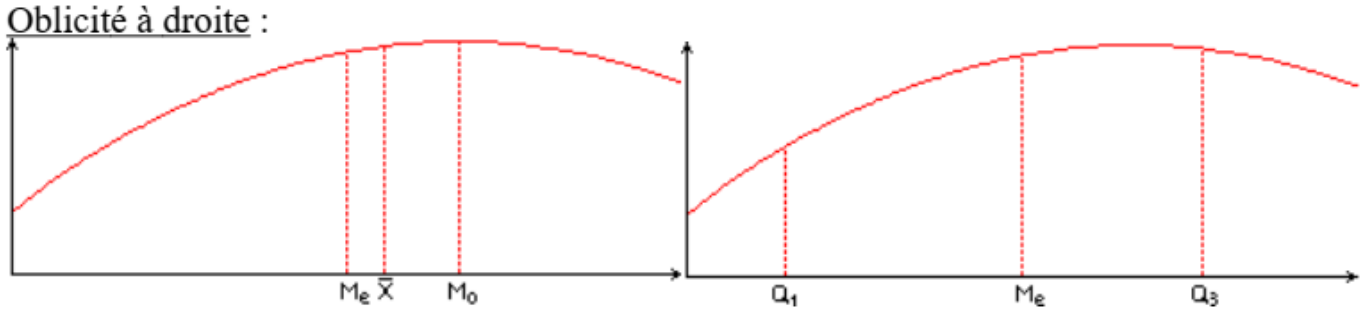
$$Y = \frac{Q_1 + Q_3 - 2M_e}{2(Q_3 - Q_1)}.$$

Le coefficient d'asymétrie de **Fisher** fait intervenir les moments centrés, il est défini par

$$F = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{\mu_3}{s^3(X)}.$$

Lorsque le coefficient d'asymétrie est positif, la distribution est plus étalée à droite : on dit qu'il y a **oblicité à gauche**.
Lorsque le coefficient d'asymétrie est négatif, la distribution est plus étalée à gauche : on dit qu'il y a **oblicité à droite**.





On utilise souvent un coefficient d'asymétrie de **Pearson basé sur les moments centers**

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}.$$

Ce coefficient d'asymétrie est toujours positif. Il est nul pour une distribution à densité de fréquence symétrique, telle la loi de Gauss.
b) Exemples.
1°/ Considérons la variable statistique *X* de distribution:

x_i	-1	4
n_i	4	1

$$M_o = -1 \; ; \mu_3 = \frac{1}{5}(4 \times (-1)^3 + 1 \times 4^3) = 12 \; ; \mu_2 = \frac{1}{5}(4 \times (-1)^2 + 1 \times 4^2) = 4.$$

$$P = \frac{\overline{X} - M_o}{s(X)} = \frac{1}{2} > 0 : \text{oblicité à gauche.}$$

$$F = \frac{\mu_3}{s^3(X)} = \frac{3}{2} > 0 : \text{oblicité à gauche.}$$

$$\beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3} = \frac{9}{4}.$$

2°/ Considérons la variable statistique *X* de distribution:

x_i	-4	1
n_i	1	4

$$M_o = 1 \; ; \mu_3 = \frac{1}{5}(1 \times (-4)^3 + 4 \times 1^3) = -12 \; ; \mu_2 = \frac{1}{5}(1 \times (-4)^2 + 4 \times 1^2) = 4.$$

$$P = \frac{\overline{X} - M_o}{s(X)} = -\frac{1}{2} < 0 : \text{oblicité à droite.}$$

$$F = \frac{\mu_3}{s^3(X)} = -\frac{3}{2} < 0 : \text{oblicité à droite.}$$

$$\beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3} = \frac{9}{4}.$$

Coefficient d'aplatissement.

Là encore plusieurs définitions sont possible

Le coefficient d'aplatissement de Pearson est

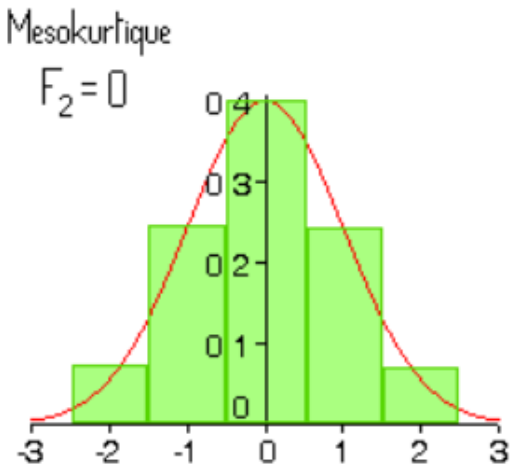
$$\beta_2 = \frac{\mu_4}{{\mu_2}^2}.$$

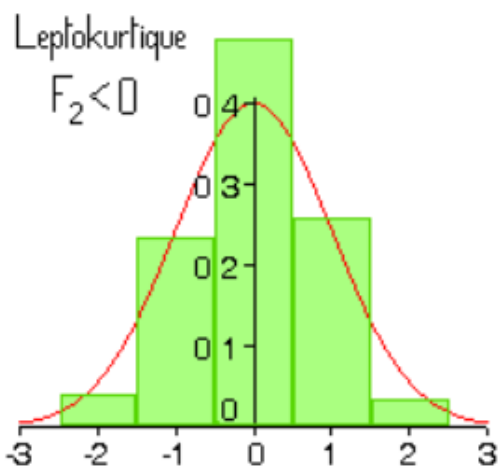
Le coefficient d'aplatissement de Yule est

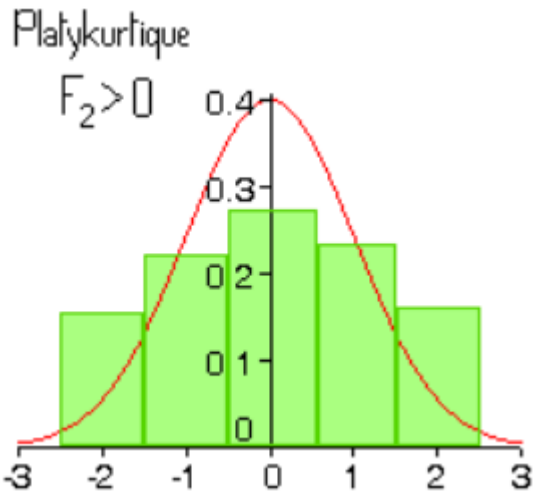
$$F_2 = \frac{\mu_4}{{\mu_2}^2} - 3.$$

On peut se demander pourquoi – 3?
C'est parce que, en Probabilités, on peut démontrer que le coefficient d'aplatissement de Pearson pour une variable aléatoire réelle qui suit une loi de Gauss, est égal à 3.

Il est alors naturel, pour comparer l'applatissement d'une distribution statistique à l'aplatissement d'une variable de Gauss, d'introduire le coefficient $F_2 = \beta_2 - 3$.
Si F_2 est égal à 0, le polygone statistique de la variable réduite a le même aplatissement qu'une courbe en cloche, on dit que la variable est **mésokurtique**.
Si F_2 est > 0, le polygone statistique de la variable réduite est moins aplati qu'une courbe en cloche, on dit que la variable est **leptokurtique**.
Si F_2 est < 0, le polygone statistique de la variable réduite est plus aplati qu'une courbe en cloche, on dit que la variable est **platykurtique**.






Platykurtique
$F_2 > 0$
0.4
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3

Lois continues

I. Loi uniforme :

1) Définition :

La variable X est de loi uniforme sur le segment [a, b] si et seulement si

$$X \rightsquigarrow \mathfrak{U}_{[a,b]} \Leftrightarrow f(x) = \begin{cases} \dfrac{1}{b-a} \ si\ a \leq x \leq b \\ 0\ sinon \end{cases}$$

2) Fonction de répartition :

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0\ si\ x < a \\ \displaystyle\int_a^x \frac{1}{b-a}dt = \frac{x-a}{b-a}\ si\ a \leq x < b \\ \displaystyle\int_a^b \frac{1}{b-a}dt = 1\ \ si\ x \geq b \end{cases}$$

$$F(x) = \begin{cases} 0\ si\ x > a \\ \dfrac{x-a}{b-a}\ si\ a \leq x \leq b \end{cases}$$

3) Espérance et variance :

$$X \rightsquigarrow \mathfrak{U}_{[a,b]} \Rightarrow \begin{cases} E(x) = \dfrac{a+b}{2} \\ V(x) = \dfrac{(b-a)^2}{12} \end{cases}$$

$$E(x) = \int_a^b xf(x)dx = \int_a^b x \times \frac{1}{b-a}dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{a+b}{2}$$

$$V(x) = E(X^2) - E^2(X)\ tel\ que\ E(X^2) = \int_a^b x^2 \times \frac{1}{b-a}dx$$

II. Loi normale ou de Laplace-Gauss ou de Gauss :

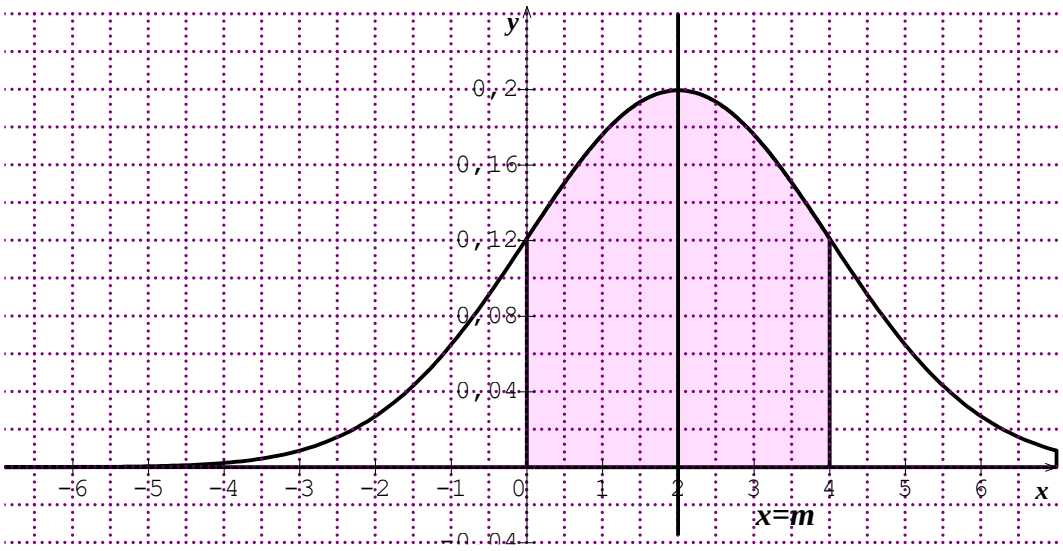
Loi normale générale $\mathcal{N}(m, \sigma^2)$

Définition :

$$X \rightsquigarrow \mathcal{N}(m, \sigma^2)\ \ ou\ m \in \mathbb{R}\ et\ \sigma > 0\ ssi\ f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(x-m)^2}\ \forall x \in \mathbb{R}$$

Remarque : $m = E(X)$ $\quad$ $\sigma^2 = V(X)$

Graphe de f

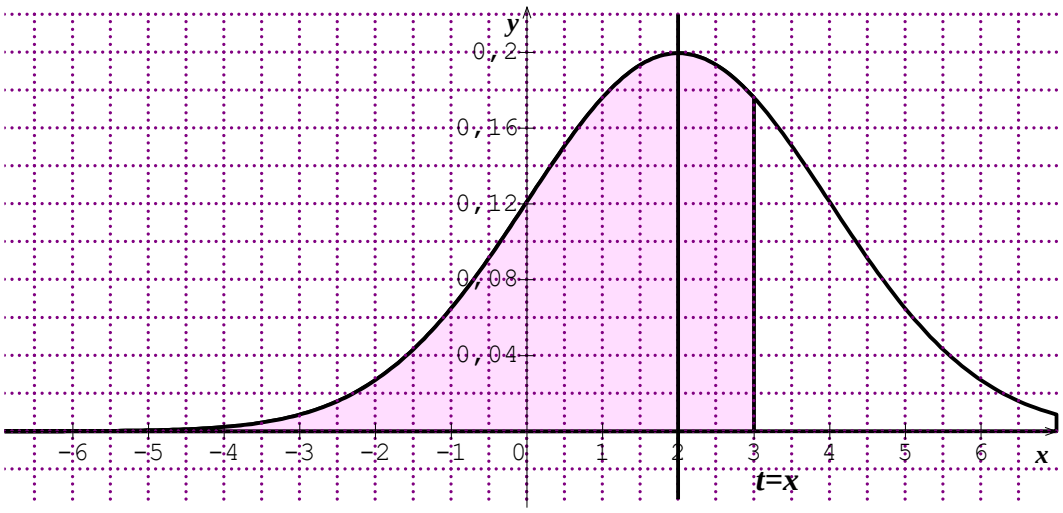


- *Courbe en cloche*
- *x = m axe de symétrie*
- *2 points d'inflexion (نقطتا انعطاف)* $x = \mp\sigma$

Fonction de répartition :

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(x-m)^2}$$

Graphiquement :



Propriétés : règle des 3σ

$$P(m - \sigma < X < m + \sigma) = 0.6827 \cong 68\%$$

$$P(m - 2\sigma < X < m + 2\sigma) = 0.9545 \cong 95\%$$

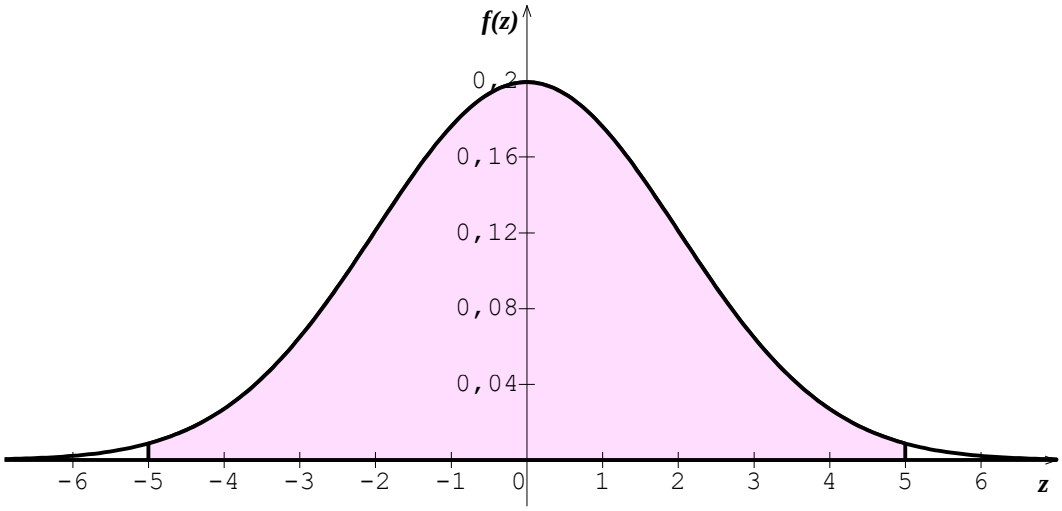
$$P(m - 3\sigma < X < m + 3\sigma) = 0.9973 \cong 99.7\%$$

Loi normale centré réduite $\mathcal{N}(0,1)$

Définition

$$X \rightsquigarrow \mathcal{N}(0,1) \quad ssi\ f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}z^2}\ \forall z \in \mathbb{R}$$

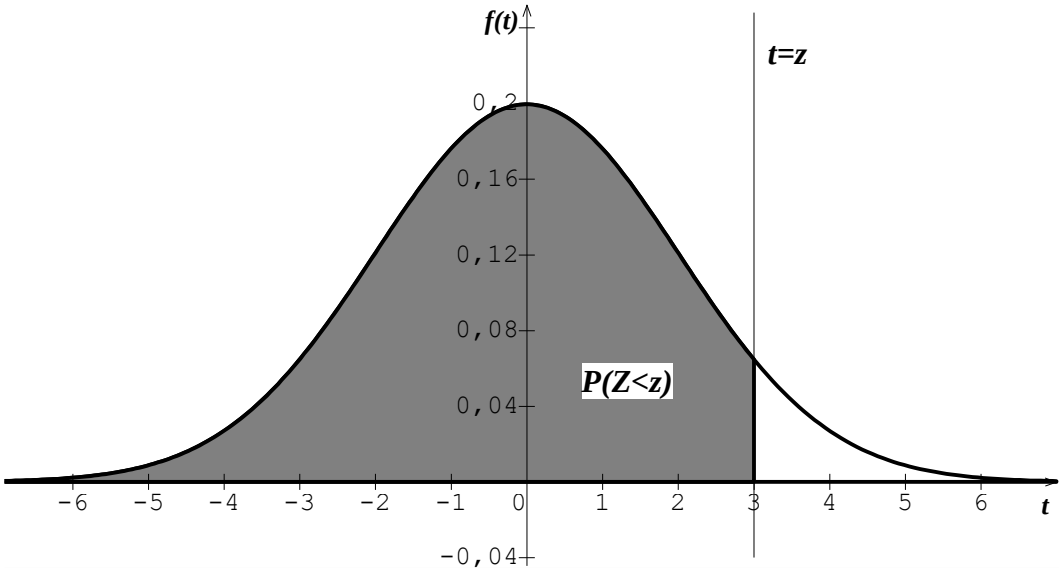
Graphe :



Fonction de répartition :

$$\boldsymbol{\varphi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}t^2} dt}$$

Graphiquement :



Analytiquement :

L'intégrale précédente a été calculée par interpolation et donnée sous forme de table statistique.

Propriétés de $\varphi(z)$

i. $\varphi(0) = P(Z \leq 0) = \frac{1}{2} = P(Z \geq 0)$

ii. $\varphi(-z) = 1 - \varphi(z)$

$$\varphi(Z \leq -z) = P(Z \geq z) = 1 - P(Z \leq z)$$

iii. $Si\ z > 0 \Rightarrow \varphi(z) > \frac{1}{2}$

$Si\ z < 0 \Rightarrow \varphi(z) < \frac{1}{2}$

Lecture de la table :

- *Lecture directe z donné* $\rightarrow lire\ \varphi(z) = P(Z \leq z)$
- $si\ 0 \leq z < 4 \rightarrow lire\ \varphi(z) sur\ la\ table$
- $si\ z \geq 4 \rightarrow \varphi(z) = 4$
- $si\ z < 0 \rightarrow \varphi(z) = 1 - \varphi(-z)$
- *Lecture inverse* $\varphi(z) = \alpha\ connu \rightarrow lire\ z$

  $\varphi(z) > \alpha > \frac{1}{2} \rightarrow lire\ z$

Exemple :

- $\varphi(z) = \alpha = 0.6700 \rightarrow z = 0.44$
- $\varphi(z) = \alpha = 0.8930 \rightarrow n'est\ pas\ dans\ la\ table$

$0.8925 < 0.8930 < 0.8944 \rightarrow \varphi(1.24) < \varphi(z) < \varphi(1.25)\ ou\ \varphi(z) = \varphi(1.24) \equiv lavaleur\ la\ plus\ proche \rightarrow z = 1.24$

- $\varphi(z) = \alpha < \frac{1}{2} \rightarrow z!$

Exemple : $\varphi(z) = 0.25 < \frac{1}{2} \rightarrow z < 0. On\ calculera\ \varphi(-z) = 1 - \varphi(z)$

$$= 1 - 0.25 = 0.75$$

$0.7486 < 0.75 < 0.7517 \rightarrow \varphi(0.67) < \varphi(-z) < \varphi(0.68)$

$\varphi(-z) = \varphi(0.67) \rightarrow -z = 0.67 \rightarrow z = -0.67$

Théorème :

$$soit \begin{cases} X \rightsquigarrow \mathcal{N}(m, \sigma^2) \\ Z = \dfrac{X - m}{\sigma} \end{cases} \Rightarrow Z \rightsquigarrow \mathcal{N}(0,1)$$

$X \rightsquigarrow \mathcal{N}(m, \sigma^2) \rightarrow P(X \leq x) = ?$

$(X \leq x) \Leftrightarrow (X - m \leq x - m) \Leftrightarrow (\frac{X - m}{\sigma} \leq \frac{x - m}{\sigma})$

$P(X \leq x) = P(Z \leq z) = \varphi(z)\ \ avec\ z = \frac{x - m}{\sigma}$

$P(X \geq x) = P(Z \geq z) = 1 - P(Z \leq z) = 1 - \varphi(z) = \varphi(-z)$

Exemple :

$X \rightsquigarrow \mathcal{N}(m, \sigma^2)\ m = 2, \qquad \sigma^2 = 9 \rightarrow$

- $P(X \leq 5) = ?$
- $x_0 = ?\ si\ P(x \geq x_0) = 0.05$

$$P(X \leq 5) = P\left(\frac{X-m}{\sigma} \leq \frac{5-m}{\sigma}\right) = P\left(Z \leq \frac{5-2}{3}\right) = P(Z \leq 1) = 0.8413 \quad avec\ Z \rightsquigarrow \mathcal{N}(0,1)$$

$$P(X \geq x_0) = 0.05 \Leftrightarrow P\left(Z \geq \frac{x_0-2}{3}\right) = 0.05 \Leftrightarrow 1 - P\left(Z \leq \frac{x_0-2}{3}\right) = 0.05$$
$$\Leftrightarrow P\left(Z \leq \frac{x_0-2}{3}\right) = 0.95 \Leftrightarrow \frac{x_0-2}{3} = 1.65 \Leftrightarrow x_0 = 6.95$$

Espérance et variance :

$$si\ Z \rightsquigarrow \mathcal{N}(0,1) \quad \Longrightarrow \begin{cases} E(Z) = 0 \\ V(Z) = 0 \end{cases}$$

$$si\ X \rightsquigarrow \mathcal{N}(m, \sigma^2) \Longrightarrow \begin{cases} E(X) = m \\ V(X) = \sigma^2 \end{cases}$$

Approximation de la loi binomiale par la loi normale :

*Soit **X** une loi binomiale de paramètre n, p* $X \rightsquigarrow \mathcal{B}(n,p)$

Dans la pratique si $n > 30\ np \geq 5\ et\ nq > 5 \Longrightarrow \mathcal{B}(n,p) \cong \mathcal{N}(m, \sigma^2)\ ou\ m = np\ et\ \sigma^2 = npq$

Exercice :

*Selon la loi de l'hérédité Mendélienne, la population théorique de sourds-muets de naissance est **p = 0.25** lorsque les parents sont des consanguins porteurs d'un certain gène récessif.* عندما يكون الوالدان قريبان يحملان جينًا متنحيًا *On considère une population de nouveau-nés issus de tels parents.*

1. *Soit **X** le nombre d'enfants sourds-muets sur un échantillon de 10 enfants, calculer* $P(X > 1)\ ; P(X \geq 1)$.
2. *Soit Y le nombre d'enfants sourds-muets sur un échantillon de 300 enfants.*
   - *Déterminer la probabilité d'avoir au moins 60 sourds-muets parmi les 300 enfants*
   - *Déterminer le nombre y_0 d'enfants tel que* $P(Y > y_0) = \frac{1}{2}$.

Solution :

1. $X \rightsquigarrow \mathcal{B}(n,p)\ avec\ n = 10\ et\ p = 0.25$

$$P(X > 1) = 1 - P(X \leq 1) = 1 - [P(X = 0) + P(X = 1)] = 1 - 0.056 - 0.187 = 0.757$$

$P(X=0)=C_{10}^{0}(0.25)^{0}(0.75)^{10}=0.75^{10}=0.056$

$P(X=1)=C_{10}^{1}(0.25)^{1}(0.75)^{9}=10\times0.25\times0.075=0.187$

2. ***Y :*** *nombre de sourds-muets sur un échantillon de 300.*
   $Y\rightsquigarrow\mathcal{B}(300,0.25)\cong Y\rightsquigarrow\mathcal{N}(m,\sigma^2)\ avec\ m=np=300\times0.25=75>5\ et\ \sigma^2=npq=300\times0.25\times0.75=56.25$

- $P(Y\geq60)=P\left(Z\geq\frac{60-m}{\sigma}\right)=P\left(Z\geq\frac{60-75}{7.5}\right)=P(Z\geq-2)=P(Z\leq2)=\varphi(2)=0.9772$
- $P(Y>y_0)=\frac{1}{2}\Leftrightarrow P\left(Z>\frac{y_0-75}{7.5}\right)=\frac{1}{2}\Leftrightarrow1-P\left(Z\leq\frac{y_0-75}{7.5}\right)=\frac{1}{2}\Leftrightarrow P\left(Z\leq\frac{y_0-75}{7.5}\right)=0.5\Leftrightarrow\frac{y_0-75}{7.5}=0\Leftrightarrow y_0=75$

<u>Loi de la somme :</u>

Soit $(X_i)_{i=1,\dots,n}$ *une suite de variables aléatoires indépendantes telles que :*$X_i\rightsquigarrow\mathcal{N}\left(m_i,\sigma_i^2\right)$ *alors :*

$$\begin{cases} X=\sum_{i=1}^{n}X_i\rightsquigarrow\mathcal{N}\left(\sum_{i=1}^{n}m_i,\sum_{i=1}^{n}\sigma_i^2\right)\\ Y=aX+b\rightsquigarrow\mathcal{N}(am+b,a^2\sigma^2)\end{cases}$$

<u>Échantillonnage :</u>

Soit $(X_i)_{i=1,\dots,n}$ *une suite de variables aléatoires indépendantes telles que :*$X_i\rightsquigarrow\mathcal{N}(m,\sigma^2)$ *alors :*

- $X=\sum_{i=1}^{n}X_i\rightsquigarrow\mathcal{N}(nm,n\sigma^2)$
- $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\rightsquigarrow\mathcal{N}\left(m,\frac{\sigma^2}{n}\right)$ $\quad\bar{X}\ moyenne\ échantillonnage$

<u>Définition :</u>

La suite $(X_1,X_2,\dots,X_n)$ *constitue un échantillon aléatoire d'une population de la loi normale.*

<u>Théorème central limite :</u>

Soit $(X_i)_{i=1,\dots,n}$ *une suite de variables aléatoires indépendantes telles que :*

$$E(X_i)=m\ et\ V(X_i)=\sigma^2\ \ \forall i\Rightarrow_{n>30}\sum_{i=1}^{n}X_i\rightsquigarrow\mathcal{N}(nm,n\sigma^2)\ et$$

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\rightsquigarrow\mathcal{N}\left(m,\frac{\sigma^2}{n}\right)\qquad\bar{X}$$

<u>Remarque :</u>

$B(n,p)\approx P(np)$

$B(n,p) \approx \mathcal{N}(np, npq)$

Lois relatives à la loi normale

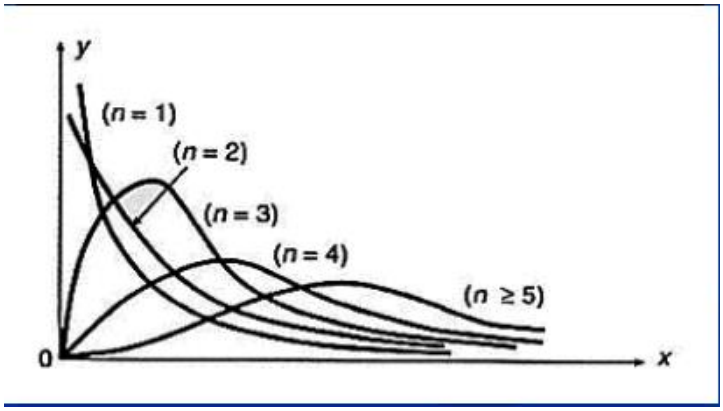
Loi de Khi-deux χ^2

La loi de Pearson ou loi de χ^2 (Khi deux) trouve de nombreuses applications dans le cadre de la comparaison de proportions, des tests de conformité d'une distribution observée à une distribution théorique et le test d'indépendance de deux caractères qualitatifs. Ce sont les test du khi-deux.

Définition : *On appelle χ^2 à* ***n*** *degrés de liberté la variable aléatoire définie par :*

$$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2 \quad avec\ X_i \rightsquigarrow \mathcal{N}(0,1)$$

Graphe :



Espérance et variance :

$E(X) = n\ et\ V(X) = 2n$

Lecture de la table :

La loi du khi-deux n'est pas symétrique

Exemple : $X \rightsquigarrow \chi^2_{10}$

$a = ?\quad P(X \geq a) = 0.95$

$b = ?\quad P(X \leq b) = 0.95$

$P(X \geq b) = 0.05$

$Si\ n > 30 \Rightarrow \chi_n^2 \approx \mathcal{N}(m, \sigma^2) avec\ m = n\ et\ \sigma^2 = 2n$

Théorème :

$$\left.\begin{array}{l} Si \quad X_1 \rightsquigarrow \chi^2_{n_1} \\ \quad\ X_2 \rightsquigarrow \chi^2_{n_2} \\ X_1 indép\ X_2 \end{array}\right\} \Rightarrow X_1 + X_2 \rightsquigarrow \chi^2_{n_1+n_2}$$

Théorème :

$Soit\ Z \rightsquigarrow \mathcal{N}(0,1)\ et\ soit\ X = Z^2 \Rightarrow X \rightsquigarrow \chi_1^2$

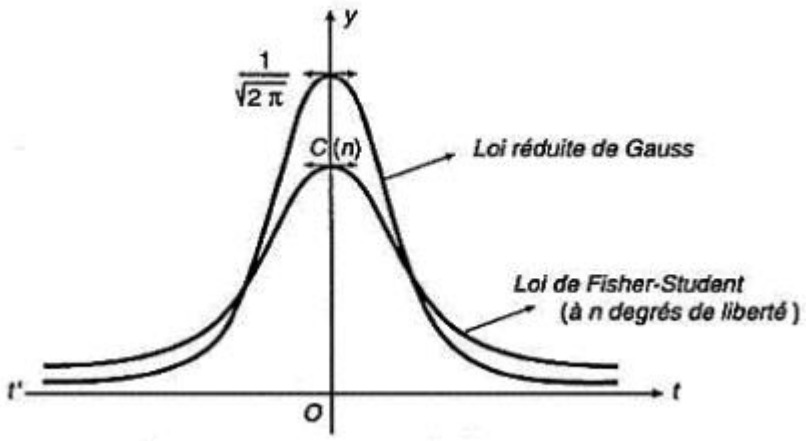
Loi de Student :

La loi de ***Student*** *(ou loi de* ***Student-Fisher****) est utilisée lors des tests de comparaison de paramètres comme la moyenne et dans l'estimation de paramètres de la population à partir de données sur un échantillon (****Test de Student****). Student est le pseudonyme du statisticien anglais Gosset qui travaillait comme conseiller à la brasserie Guinness et qui publia en 1908 sous ce nom, une étude portant sur cette variable aléatoire.*

Définition :

Soit ***X*** *une variable aléatoire suivant une loi normale réduite* $\mathcal{N}(0,1)$ *et* ***Y*** *une variable aléatoire suivant une loi de khi-deux (Pearson) à* ***n*** *degrés de liberté* χ_1^2 *,* ***X*** *et* ***Y*** *étant indépendantes, on dit alors que* $T_n = \frac{X}{\sqrt{\frac{Y}{n}}}$ *suit une loi de Student à* ***n*** *degrés de liberté.*

Graphe :



Espérance et variance :

$E(T) = 0\ si\ n > 1 \quad V(T) = \frac{n}{n-2} \quad si\ n > 2$

Loi de Fisher-Snedecor

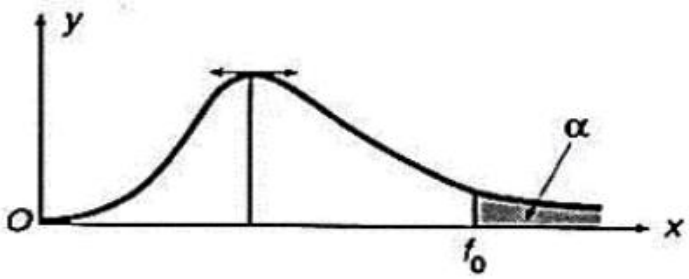
La ***loi de Fisher-Snedecor*** *est utilisée pour comparer deux variances observées et sert surtout dans les très nombreux tests d'analyse de variance et de covariance.*

Définition :

Soit X et Y deux variables aléatoires indépendantes suivant une ***loi de Pearson (khi-deux)*** *respectivement à* ***n*** *et* ***m*** *degrés de liberté.*

On dit que $F = \frac{X/n}{Y/m}$ *F suit une loi de* ***Fisher-Snedecor*** *à n, m degrés de liberté.*

Graphe :



$P(F > f_0) = \alpha$

Espérance et variance :

$E(F) = \frac{m}{m-2}$ *si* $m > 2$ $\qquad V(F) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$ *si* $m > 4$

Exercice :

Dans une certaine pathologie, la durée d'hospitalisation est distribuée selon un χ^2_{18}

QCM

1. *L'écart-type de la durée d'hospitalisation est 6 jours*
2. *50% des sujets ont une durée d'hospitalisation supérieure à 18 jours*
3. *Plus de 5% des patients restent hospitalisés au moins 4 semaines*
4. *Plus de 10% des patients restent hospitalisés au moins 4 semaines*
5. *Moins de 25% des patients restent hospitalisés entre 3 et 4 semaines*

Répence :

1. $V(X) = 2n = 2 \times 18 = 36 \rightarrow \sigma(X) = \sqrt{36} = 6$

2. $P(X > 18) = 0.5 \qquad faux$

$P(X > x_0) = 0.5 \longrightarrow x_0 = m$

$X \rightsquigarrow \chi^2_{18} \longrightarrow E(X) = 18\, jours$

$P(X > x_0) = P(X < x_0) = 0.5 \longrightarrow x_0 = médiane$

$la\ loi\ \chi^2 n'estpas\ symétrique\ alors\ Me \neq m$

Loi symétrique Me= m=Mo

3. $P(X > a) = 0.05$

Séries statistiques doubles (à deux variables)

Définition : c'est l'étude simultanée de 2 variables quantitatives X et Y d'une même population.

A chaque individu i ($1 \leq i \leq n$) correspond un couple (x_i , y_i) où x_i est la modalité du caractère X et y_i la modalité du caractère Y. L'ensemble des couples (x_i ; y_i) définit une série statistique double

Exemple :

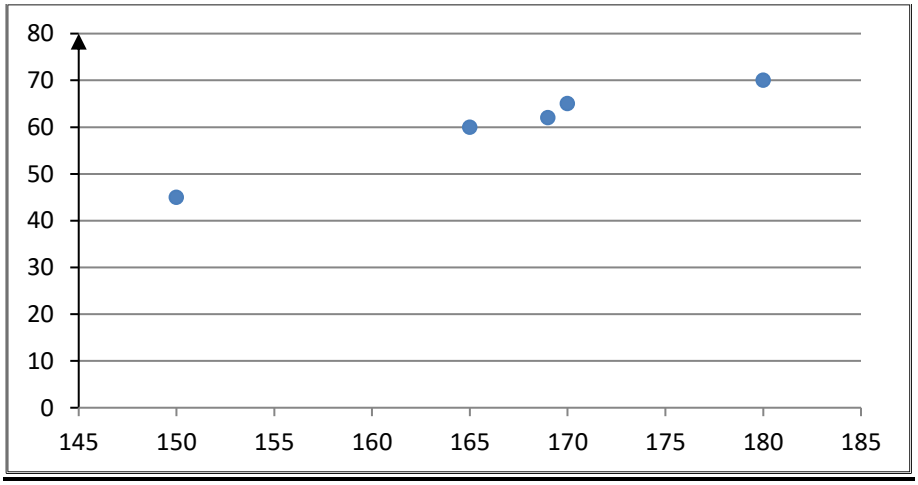
Le poids X et la taille Y d'une personne

Représentation graphique :

La série statistique double est représentée par un nuage de points dans un repère du plan constitué des points M_i (x_i , y_i).

Exemple :

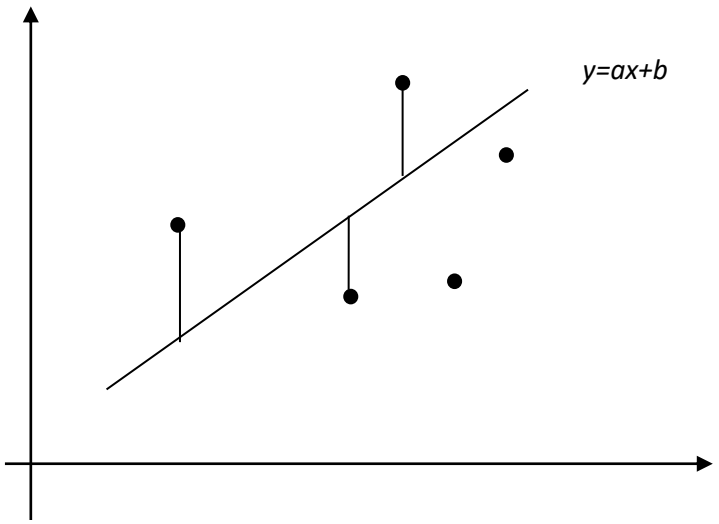
X : taille en cm	170	150	165	169	180
Y : poids en kg	65	45	60	62	70

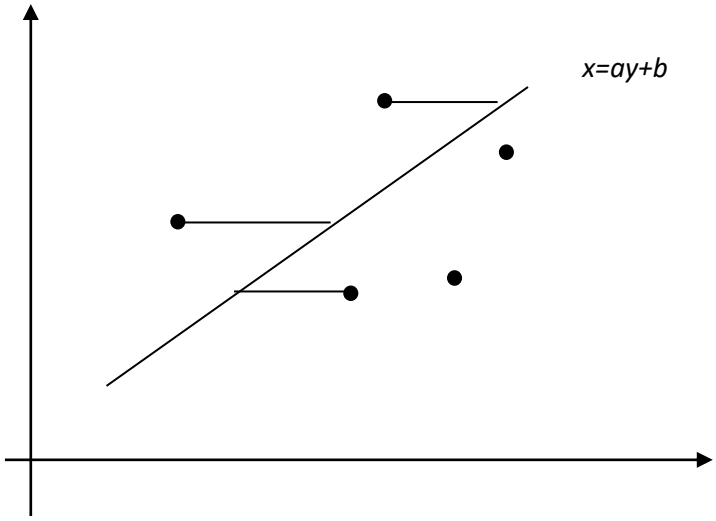


Ajustement linéaire (droite de régression)

On cherche une droite $y = ax + b$ qui rapproche le mieux les points du nuage.

Droite de régression de y en x $D_y(x)$





Supposons les couples (x_i , y_i) de variables (X , Y) approximativement alignés

La méthode des moindres carrés permet de déterminer les coefficients $\hat{a}$ et $\hat{b}$ de la droite $\hat{y} = \hat{a}x + \hat{b}$ avec

$$\hat{a} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \ , \hat{b} = \bar{y} - \hat{a}\bar{x} \ , \qquad \bar{x} = \frac{\sum x_i}{n} \ , \bar{y} = \frac{\sum y_i}{n}$$

<u>Droite de régression de x en y D_x(y</u>

Elle est obtenue en minimisant la somme des carrés des écarts des points à la droite par rapport à l'axe (OX) on obtient une droite d'équation $\hat{x} = \widehat{a'}y + \widehat{b'}$ avec

$$\widehat{a'} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (y_i - \bar{y})^2}, \qquad \widehat{b'} = \bar{x} - \hat{a}\bar{y}$$

Covariance

Définition :

On appelle covariance de (X,Y) le nombre noté cov(X,Y) tel que :

$$cov(x,y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}$$

Remarque :

1) En posant $X = Y$ on retrouve l'expression de la variance
2) $\hat{a} = \frac{cov(X,Y)}{{\sigma_X}^2}$ $\quad \widehat{a'} = \frac{cov(X,Y)}{{\sigma_y}^2}$

Le coefficient de corrélation :

Il Est calculé à partir de l'expression

$$r_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Proposition :

1) Si $r_{XY} = 0$ alors les caractères associent aux variables (X, Y) sont indépendant
2) Si $0 < r_{XY} < 1$ alors la corrélation est positive (x et y varient au même sens)
3) Si $-1 < r_{XY} < 0$ la corrélation est négative (x et y varient en sens contraire)

Remarque :

- $\hat{a}\widehat{a'} = r_{XY}^2$
- Si $|r_{XY}| \cong 1 \Rightarrow$
  la corrélation est maximale , X et Y sont fortement corrélés (dépendants)

On considérera $|r_{XY}| \cong 1$ si $r_{XY} \geq 0.81$

- Si $r_{XY} = 0$, la corrélation est nulle (X et Y sont indépendants)
- On a $\hat{b} = \bar{y} - \hat{a}\bar{x}$ et $\widehat{b'} = \bar{x} - \widehat{a'}\bar{y}$ implique que $D_x(Y)$ et $D_Y(X)$ se correspondent en un point $G(\bar{x},\bar{y})$ appelé centre de gravité en formant un angle θ

Exemple :

Supposons un échantillon aléatoire de 4 firmes pharmaceutiques présentant les dépenses de recherche X et les profits Y suivants (en milliers de dollars) :

X	Y
40	50
40	60
30	40
50	50

Trouvez la droite de régression et le coefficient de corrélation.

Calculons tout d'abord $\overline{X}$ et $\overline{Y}$:

$$\overline{x} = \frac{1}{n}\sum x_i = \frac{1}{4}(40 + 40 + 30 + 50) = \frac{160}{4} = 40$$

$$\overline{y} = \frac{1}{n}\sum y_i = \frac{1}{4}(50 + 60 + 40 + 50) = \frac{200}{4} = 50$$

Complétons le tableau suivant :

x_i	y_i	$x - \overline{x}$	$y - \overline{y}$	$(x - \overline{x})^2$	$(y - \overline{y})^2$	$(x - \bar{x}).(y - \overline{y})$
40	50	0	0	0	0	0
40	60	0	+10	0	+100	0
30	40	−10	−10	+100	+100	+100
50	50	+10	0	+100	0	0
$\sum$				200	200	100

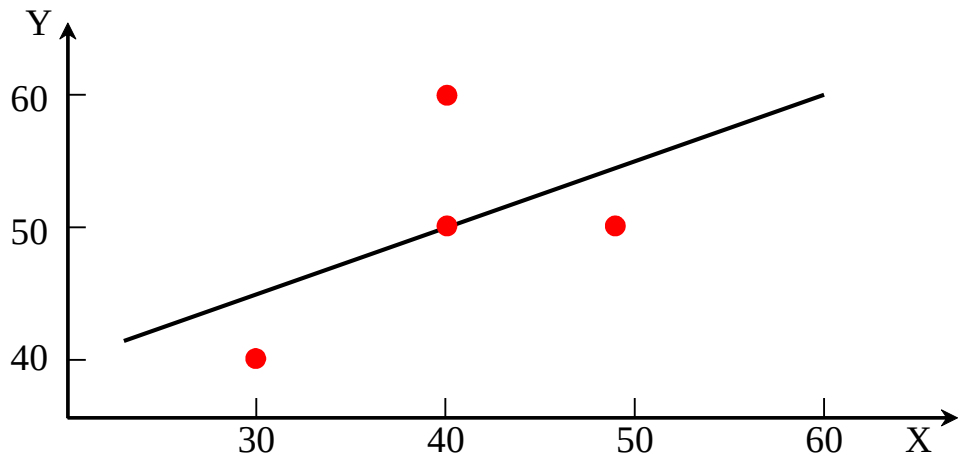
On a donc :

$$\sum (x_i - \overline{x})^2 = 200$$
$$\sum (y_i - \overline{y})^2 = 200$$
$$\sum (x_i - \overline{x}).(y_i - \overline{y}) = 100$$

Les coefficients de la droite de régression sont :

$$\hat{a} = \frac{\sum (x_i - \overline{x}).(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{100}{200} = 0{,}5$$
$$\hat{b} = \overline{y} - \hat{a}.\overline{x} = 50 - 0{,}5 \times 40 = 50 - 20 = 30$$

Et le coefficient de corrélation :

$$r_{XY} = \frac{\sum (x_i - \overline{x}).(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2} \times \sqrt{\sum (y_i - \overline{y})^2}} = \frac{100}{\sqrt{200} \times \sqrt{200}} = \frac{100}{200} = 0{,}5$$

La corrélation est positive et de qualité moyenne