SÉRIES STATISTIQUES

Une population est un ensemble d'**individus** sur lesquels on étudie un **caractère** ou une **variable**, qui prend différentes valeurs ou modalités.
Nous nous intéresserons uniquement aux **variables quantitatives**.
Les modalités sont mesurables et prennent des valeurs numériques.

Une variable quantitative peut être :

- **Discrète**, quand elle prend des **valeurs entières.**
- **Continue** quand elle prend n'importe quelle **valeur sur un intervalle donné**.
- **Effectif total** noté **N**, est le nombre d'individus qui composent la population.
- **Effectif d'une valeur** noté $\mathbf{n_i}$ d'une valeur $\mathbf{x_i}$ est le nombre d'individus associé à cette valeur.
- **Fréquence** notée $\mathbf{f_i}$ est le rapport entre l'effectif de cette valeur et l'effectif total : $f_i = \frac{n_i}{N}$
- **Effectif (fréquence) cumulé** croissant d'une valeur $\mathbf{x_i}$ est égal à la somme des effectifs (ou fréquences) des valeurs inférieures ou égales à $\mathbf{x_i}$.

Paramètres de position :

- Mode est la (ou les valeurs) de la variable ayant le plus grand effectif.
- Médiane est la valeur qui partage la population en deux sous ensembles de même effectif. Elle correspond à la fréquence cumulée croissante de 50%.
- Moyenne : $\overline{X} = \frac{\sum_{i=1}^{n} n_i \ x_i}{N} = \sum_{i=1}^{n} f_i \ x_i$

Paramètres de dispersions :

- **Étendue** est la différence entre la plus grande valeur et la plus petite valeur de la variable.
- **Variance** : $v(x) = \frac{\sum_{i=1}^{n} n_i \ (x_i - \overline{X})^2}{N}$
- **Écart type** : $\sigma(x) = \sqrt{v(x)}$

Série Statistiques à deux variables :

Quand il semble exister dans certains cas, un lien entre deux caractères **x** et **y** d'une même population, par exemple entre le poids et la taille d'un individu.
On les étudie simultanément en vue de faire des prévisions.
À chaque individu **i** correspond alors le couple ($\mathbf{x_i}$; $\mathbf{y_i}$) dans lequel $\mathbf{x_i}$ est une donnée de la variable **x** et $\mathbf{y_i}$ est une donnée de la variable **y**.

- L'ensemble des n couples (x_i ; y_i) s'appelle une série statistique à deux variables d'effectif total **n**.
- Cette série statistique à deux variables peut être présentée sous forme de tableau, ou représentée graphiquement dans le plan muni d'un repère par le nuage des points M_i de coordonnées (x_i ; y_i)
- Dans le plan muni d'un repère, l'ensemble des points M_i de coordonnées (x_i ; y_i) est appelé nuages de points de la série statistiques des (x_i ; y_i).
- On appelle point moyen d'un nuage de n points M_i (x_i ; y_i) le point **G** de coordonnées :

$$G(\ x\ ;\ y\ )\ \text{ avec : }\ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Droite de régression linéaire :

La droite de régression linéaire (où droite des moindres carrés) est la droite qui passe à travers le nuage le pus près possible de chaque point.
Pour cela on utilise la méthode des moindres carrés qui consiste à trouver la droite telle que la somme des carrés de la distance entre cette droite et chacun des points soit minimale.

La droite de régression D de y en x a pour équation y = a x + b avec :

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

Cette droite permet d'extrapoler la valeur que devrait avoir la variable y en fonction de la variable x.

On peut aussi, à l'inverse étudier la droite de régression D' de x en y qui a l'inverse permettra d'extrapoler la valeur que devrait avoir la variable x en fonction de la variable y.
Elle aura pour équation x = a' y + b' avec :

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

En théorie on peut toujours calculer une droite de régression linéaire quel que soit le nuage de points, mais si les points sont trop éloignés de la droite ou ne sont pas groupés autour, cette droite n'a aucun sens.
On appelle coefficient de corrélation affine des variables x et y d'une série statistiques à deux variables le nombre noté r tel que :

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\ \sqrt{\sum (y_i - \bar{y})^2}} = \frac{\sigma_{x\,y}}{\sigma_x\ \sigma_y}$$

Si : $\mathbf{0{,}75 \le r \le 1}$ on convient de dire qu'il y a une bonne corrélation, donc la droite est bien représentative de la série statistique.
Si : $\mathbf{0 \le r < 0{,}75}$ on convient de dire qu'il y a une faible corrélation, donc la droite n'est pas représentative de la série statistique.
r = 1 $\Rightarrow$ totale dépendance linéaire entre les 2 variables.
r = 0 $\Rightarrow$ aucune dépendance linéaire entre les 2 variables.