En tant que science, les statistiques ne se limitent pas à une description empirique des propriétés numériques d'un objet. Elles jouent un rôle bien plus important en proposant, à partir des données, des méthodologies et des outils qui permettent d'établir des hypothèses et d'en mesurer la fiabilité.

Les statistiques sont une science non deterministe par nature, qui donne une interprétation probabiliste de l'ensemble des solutions. Les mêmes expériences pouvant aboutir à des résultats différents, l'analyse des données peut donner lieu à des controverses. L'objectif de cet enseignement est de vous familiariser avec les statistiques afin que vous soyez capable d'apporter des réponses scientifiques à ce type de controverses.

Statistiques descriptives

Au sein de ce chapitre, nous allons présenter différents outils suceptibles de faciliter l'analyse et la description des données recueillies lors d'échantillonages ou d'expérimentations. L'objectif ne sera pas de tirer des conclusions sur les différentes populations statistiques (modèles et prédictions), il s'agira de présenter les données accumulées sous forme synthétique.

1 Présentation des données

Lors de la collecte des données, les valeurs observées se trouvent évidemment sans ordre. Pour un nombre important d'échantillons, il s'agira d'organiser les données :

tableaux de distributions de fréquence
représentations graphiques

1.1 Série statistique simple

Une série statistique simple est un ensemble de données relative à une variable mesurée sur un échantillon d'éléments.

1.1.1 Tableaux de distribution de fréquence

Pour des variables qualitatives, il s'agira de présenter chaque catégorie ou de regrouper ces catégories en classes.

Pour une population de n variables quantitatives, il s'agira de regrouper les échantillons dans des classes. Pour cela on détermine

la valeur maximum v_{max}
la valeur minimum v_{min}
le nombre de classe n_c
l'intervalle de classe Δ_c
les indices de classe v_i

v_{max} et v_{min} sont souvent les maxima de la série. Le nombre de classe n_c peut être déterminer en fonction du nombre d'échantillons

$$\begin{array}{ll} n_c = 1 + 3.3\log(n) & \text{Règle de Sturge} \\ n_c = 2.5\sqrt[4]{n} & \text{Règle de Yule} \end{array}$$

L'intervalle de classe est alors

$$\Delta_c = \frac{v_{max} - v_{min}}{n_c}.$$

Ces valeurs permettent de déterminer les indices de classes, le plus souvent au centre de la classe.

Définissons quelques variables :

La fréquence absolue f_i est le nombre d'élément appartenant à une classe.

La fréquence relative d'une classe est le rapport de son effectif à l'effectif total (f_i/n).

Le pourcentage est la fréquence relative exprimée en % $(100 f_i/n)$.

La fréquence cumulée correspond à l'effectif total des valeurs plus petites que la borne supérieure de la classe considérée.

La fréquence relative cumulée est le rapport de la fréquence cumulée à l'effectif total.

Le pourcentage cumulé est la fréquence relative cumulée exprimée en pourcentage.

1.1.2 Représentation graphique

le diagramme en bâton
polygone de fréquence

histogramme
courbe de fréquence
les diagrammes circulaires

1.2 Séries statistiques à deux variables

Chaque variable pouvant être qualitative ou quantitative il existe trois combinaisons (qualitative-qualitative, qualitative-quantitative, quantitative-quantitative).

1.2.1 Tableaux de distribution de fréquence

Si x_i et y_j sont les i^eme et j^{eme} classes des variables x et y, on note f_{ij} la fréquence absolue des éléments appartenants à ces deux classes.

	x_1	x_2	$\cdots$	x_l
y_1	f_{11}	f_{12}	$\cdots$	f_{1l}
y_2	f_{21}	f_{22}	$\cdots$	f_{2l}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_k	f_{k1}	f_{k2}	$\cdots$	f_{kl}

Tableau de corrélation

Si les deux variables sont qualitatives, il s'agit d'un tableau de contingences.

1.2.2 Représentation graphique

Si les deux variables sont quantitatives on peut tracer un
→ *stéréogramme* c'est à dire histogramme à trois dimensions.
→ *diagramme de dispersion* c'est à dire un nuage de point.
Si les deux variables sont qualitatives, la représentation graphique est souvent superflue. S'il y a une variable quantitative et une variable qualitative, on peut tracer une série de courbes de fréquence ou d'histogramme pour chaque catégorie de la variable qualitative.

1.3 Séries statistiques multiples

On dresse un tableau en associant à chaque élément une ligne et à chaque variable une colonne. Les représentations graphiques à deux ou trois dimensions sont possibles mais ne contiennent pas toute l'information.

2 Description des séries statistiques

Définissons quelques variables d'une distribution:
Les paramètres de position renseignent sur l'ordre de grandeur et sur les possibles valeurs centrales.
Les paramètres de dispersion renseignent sur l'étalement de la distribution et du degré de dispersion autour d'une valeur centrale.
Les paramètres de totalisation indiquent, pour une population finie, la valeur cumulée jusqu'au i^{eme} élément.
Les paramètres d'ajustement renseignent sur le degré de conformité d'une série à un modèle donné.
Les paramètres de covariation iindiquent le degré de corrélation ou d'interrelation existant entre deux ou plusieurs variables.
Les paramètres de similitude renseignent sur le degré ressemblance de divers variables.

2.1 Les paramètres de position

2.1.1 La moyenne arithmétique

$\bar{y}$ est la moyenne arithmétique de y_i:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Si les données sont groupées en k classe la formule s'écrit

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{k} f_i y_i,$$

avec f_i la féquence de la classe i, et y_i la valeur centrale de la classe i. En définissant

$$p_i = \frac{f_i}{n}$$

la probabilité d'appartenir à une classe, on obtient

$$\bar{y} = \sum_{i=1}^{k} p_i y_i$$

. Cette formule peut être appliqué à toute loi de probabilité d'une variable aléatoire.

La moyenne arithmétique est le paramètre de position qui répond au principe des moindres carrés:

$$\frac{d\sum_{i=1}^{n}(p-x_i)^2}{dp}=0$$

2.1.2 La médiane

La médiane sépare deux sous-populations d'égale importance. Il est facile de retrouver cette médiane graphiquement ou en interpollant à partir de la fonction cumulée. La médiane s'écrit

$$M_e = L_m + \frac{i}{f_m}\left(\frac{n}{2} - f_{cum}\right),$$

avec L_m la limite inférieure de la classe à laquelle appartient le $i^{eme}/2$ élément, f_m la fréquence de la classe médiane, i l'intervalle de la classe, n l'effectif de l'échantillon, f_{cum} la fréquence cumulée jusqu'à la limite inférieure de la classe médiane.

2.1.3 Le mode ou la valeur dominante

Le mode est la valeur de la variable qui a la plus forte fréquence. C'est le sommet de la courbe. Il s'agit alors de calculer la valeur modale

$$M_o = L + i\left(\frac{\Delta i}{\Delta i + \Delta s}\right),$$

avec L la limite inférieure de la classe modale (classe ayant la fréquence la plus élevée), i l'intervalle de la classe, Δi l'excédent de fréquence entre la classe modale et la classe inférieure et Δs l'excédent de fréquence entre la classe modale et la classe supérieure.
Pour les variables qualitatives le mode est le seul paramètre de position.

2.1.4 La moyenne géométrique

La moyenne géométrique s'écrit

$$M_g = \sqrt[n]{\prod_{i=1}^{n} x_i},$$

et

$$\log(M_g) = \frac{1}{n}(\log(x_1) + \log(x_2) + \ldots + \log(x_n)$$

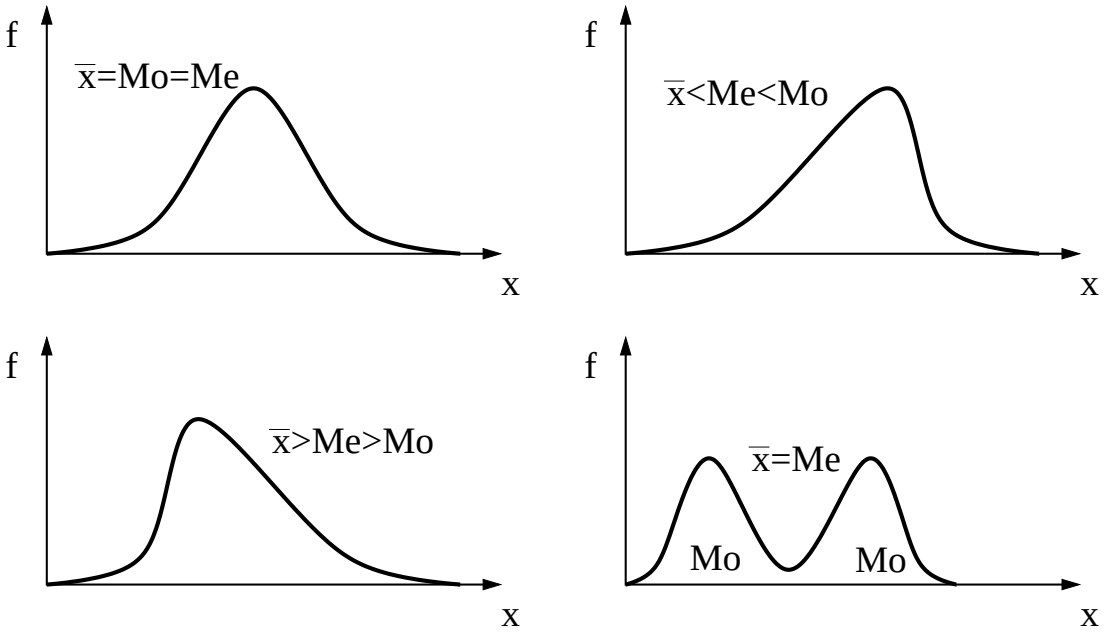

Figure 1: Comparaisons entre différents types de distributions: symétrique, asymétrique à gauche et à droite, bimodale et symétrique

Très utile si on a une progression géométrique ou pour les valeurs nécessitant une transformation logarithmique pour suivre une distribution approximativement normale.

2.1.5 La moyenne quadratique

La moyenne quadratique s'écrit

$$M_q = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}.$$

Elle sert essentiellement à définir des surfaces moyennes.

2.1.6 La moyenne des rapports

Une série de variables r_i résulte du rapport entre deux variables x_i et y_i. La moyenne des rapports s'écrit

$$\bar{r} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{r_i}},$$

ou

$$\bar{r} = \frac{\bar{x}}{\bar{y}}$$

2.1.7 Les paramètres de position de séries statistiques multiples

On calcule le centre de gravité

$$[\bar{x_j}] = \frac{\sum_{i=1}^{n} x_{ij}}{n}$$

et les différents modes

2.2 Les paramètres de dispersion

2.2.1 Variables qualitatives

$var(x)$ ou s_x^2 est **la variance** de x_i:

$$var(x) = s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right).$$

Si x est une variable discrète pouvant prendre les valeurs $x_1,\ x_2,\ \ldots,\ x_n$ avec les probabilités $p_1,\ p_2,\ \ldots,\ p_n$,

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

Pour la loi de probabilité de variable aléatoire x, nous avons
La variance corespond au carré de **l'écart type** s_x.

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

Le coefficient de variation est le rapport entre l'écart type et la moyenne exprimé en pourcentage:

$$C_v = \frac{100 s_x}{\bar{x}}$$

Le moment du troisième ordre s'écrit

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Le coefficient d'asymétrie α_3,

$$\alpha_3 = \frac{m_3}{s_x^3},$$

permet de mesurer la symétrie d'une distribution.
$\alpha_3 \to 0$ la courbe est symétrique par rapport à $\bar{x}$.
$\alpha_3 \to \infty$ la courbe est allongée vers la droite par rapport à $\bar{x}$.
$\alpha_3 \to -\infty$ la courbe est allongée vers la gauche par rapport à $\bar{x}$.
Le moment de quatrième ordre s'écrit

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

Le coefficient d'aplatissement α_4,

$$\alpha_4 = \frac{m_4}{s_x^4},$$

permet de mesurer l'aplatissement relatif de différentes distributions.

2.2.2 Variables quantitatives

La richesse est le nombre de catégories d'une population.
La richesse estimée est l'espérence mathématique du nombre de catégories représentées au sein d'une sous-population

$$E(rich_{n'}) = \sum_{k=1}^{rich}\left(1 - \frac{C_{n-f_k}^{n'}}{C_n^{n'}}\right)$$

La diversité H est un indicateur de la répartition des éléments dans les différentes catégories représentées. Pour cela on peut utiliser l'indice de Shannon

$$D_{shannon} = -\sum_{k=1}^{rich} p_k \ln(p_k)$$

La régularité permet de mesurer l'étalement d'une distribution par rapport à la richesse maximum de cette distribution:

$$R = -\frac{\sum_{k=1}^{rich} p_k \ln(p_k)}{\ln(rich)}$$

L'indice de diversité de Shannnon est biaisé par la richesse de l'échantillon. Ce n'est pas le cas de **la diversité de Simpson** qui calcule la probabilité que deux échantillons successif n'appartiennent pas à la même catégorie:

$$D_{simpson} = 1 - \sum_{k=1}^{rich} \frac{f_k(f_k - 1)}{n(n-1)}$$

On défini la concentration de Simpson comme le complément de la diversité

$$C_{simpson} = \sum_{k=1}^{rich} \frac{f_k(f_k - 1)}{n(n-1)}$$

2.2.3 Les séries statistique multiples

s_{xy} est la covariance des deux variables x et y:

$$\begin{aligned} s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{t=1}^{n} y_i}{n}\right). \end{aligned}$$

La matrice de covariance s'écrit

$$S = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{pmatrix}$$

avec

$$s_x^2 = s_{xx}$$

2.3 Les paramètres de totalisation

L'estimateur de la quantité totale s'écrit

$$\hat{x} = n\bar{x}$$

L'estimateur du nombre d'élements au sein d'une population est

$$\hat{A} = np$$