STATISTIQUE DESCRIPTIVE

Ce cours développe essentiellement des notions qui se retrouveront dans le cours de probabilités, avec un vocabulaire proche de celui des probabilités (le vocabulaire statistique varie d'un ouvrage à l'autre). Une partie de ce programme est traité au lycée (2^{de} et $1^{ère}$).

1. les étapes d'une étude statistique - vocabulaire

LA statistique est une méthode scientifique permettant d'analyser des données préalablement assemblées et portant sur de grands ensembles. L'ensemble (fini) étudié est appelé **Population [statistique]** (Ω) et ses éléments **individus** (ω) ou **"unité statistique"** ; ils doivent être rigoureusement définis.
UNE [étude] statistique est l'étude d'un ensemble de données sur un sujet précis concernant la population considérées et comportant généralement les étapes suivantes :

- partie descriptive :
  - collecte des données,
  - classement des données,
  - synthèse des résultats,
- partie interprétative.

2de

1.1. collecte des données (relevé statistique)

Les données à collecter sont des renseignements sur les individus de la population et correspondent à un ou plusieurs **caractères** pouvant prendre chacun différentes valeurs appelées **modalités**. L'ensemble des modalités d'un caractère est appelé **espace des résultats** ou encore **espace des modalités**.
Si l'espace des modalités est une partie de $\mathbb{R}$ (ou de $\mathbb{R}^n$), le caractère est dit **quantitatif** de type **discret** (nombre fini ou dénombrable de valeurs) ou bien de type **continu**, sinon il est dit **qualitatif**.

La collecte peut être exhaustive par **recensement** de la population (réalisation d'une **enquête** statistique). Quand la population est trop nombreuse, l'étude peut, au contraire, être restreinte à un sous-ensemble appelé **"échantillon"** sensé être **représentatif** de la population (la répartition du caractère doit être la même dans l'échantillon et dans la population) ; la collecte est alors réalisée par **sondage**.

Les relevés statistiques peuvent être systématiques (*exp : état civil*), périodiques (*exp : recensement, inventaires,...*) ou occasionnels (*exp : dégâts d'une catastrophe*).

On appelle **variable statistique** (=statistique) l'application X qui à chaque individu ω de la population fait correspondre la valeur observée (la modalité) $X(\omega)$ du caractère étudié.

1.2. classement des données ("dépouillement")

Les données doivent être organisées et présentées sous une forme synthétique, claire et exploitable : un **tableau statistique** (appelé par abus de langage "des statistiques") indique, pour chaque modalité effectivement prise par le caractère étudié, l'**effectif** (nombre d'individus ayant cette valeur). Si le caractère est continu (ou à valeurs ordonnables), les valeurs sont regroupées en **intervalles** d'amplitudes identiques ou différentes ; des regroupements en **classes** peuvent aussi être faits sur des caractères discrets ou qualitatifs à valeurs trop nombreuses (*par exemple plus de 20*). Le tableau construit (à une ligne et k colonnes) correspond à une application de l'ensemble des classes vers un ensemble d'effectifs, appelée **série statistique [groupée** (en continue) / **discrète** (par modalité)**] des effectifs**. Ce dépouillement des données qui passe de **la** variable statistique à **une** série statistique s'accompagne d'une perte d'information "anonymisant" plus ou moins fortement les individus (Population $\longrightarrow$Espace des modalités $\longrightarrow$Espace des effectifs).

On peut construire une **série statistique des fréquences** en remplaçant, pour chaque classe du tableau, l'effectif par la **fréquence** (effectif de la classe / effectif total de la population).

Si le caractère est ordonné, on peut construire une série statistique des effectifs (ou des fréquences) **cumulés**. 1ère

La **répartition des effectifs (ou des fréquences)** d'une variable statistique ordonnée X est la fonction F qui à chaque modalité m_i associe le nombre d'individus (ou la fréquence correspondante) dont la modalité ne dépasse pas m_i; dans le cas discret, F est une fonction en escalier croissante et continue à droite, dans le cas continu, F est une fonction affine par morceaux croissante et continue

Si les classes n'ont pas la même amplitude, on remplace les effectifs ou les fréquences par les **densités d'effectifs** ou **de fréquences** (fréquence divisée par l'amplitude de la classe).

1.3. synthèse des résultats

La variable statistique pouvant prendre un grand nombre de valeurs, il est nécessaire de résumer la série statistique par un petit nombre de paramètres caractéristiques, numériques ou graphiques.

1.4. interprétation des résultats

Si la population est abordée par des techniques d'échantillonnages, on utilise des méthodes scientifiques issues de la théorie des probabilités (on observe une "stabilité statistique" des expériences aléatoires répétitives...).

2. Etude des séries statistiques simples (à valeurs réelles)

Population Ω, $|\Omega|=N$; nombre fini k de classes c_i , intervalles $[a_i,b_i]$ de **centre** $x_i=(a_i+b_i)/2$ ou de valeurs x_i, d'effectifs n_i et de fréquences $f_i=n_i/N$ avec $\sum_{i=1}^{k} f_i = 1$. Utilisation de l'informatique (tableurs-grapheurs).

2.1. paramètres caractéristiques numériques

2.1.1. paramètres de position

Mode : pour une variable statistique discrète, toute modalité ayant un effectif (donc une fréquence) maximale ("l'abscisse d'un bâton le plus long du diagramme). **Classe modale** : pour une variable statistique groupée, toute classe ayant une densité d'effectif (donc de fréquence) maximale ("l'abscisse du rectangle le plus haut de l'histogramme", pas forcément la classe d'effectif max).

Médiane : valeur M (calculée au besoin par interpolation linéaire) de X qui fait franchir le seuil de 50% de fréquence cumulée à F, partageant la population en deux parties de même taille.

Quartiles : valeurs Q1, Q2=M et Q3 de X qui font respectivement franchir les seuils 25%, 50% et 75%. On peut définir des déciles, centiles... et autres quantiles.

Moyenne : $\bar{x} = \frac{\sum_{i=1}^{k} n_i\, x_i}{N} = \sum_{i=1}^{k} f_i\, x_i$. Avec une série groupée en k intervalles $[a_{i-1},a_i]$, on a $x_i = \frac{a_{i-1}+a_i}{2}$.

Propriété de conservation de la moyenne : pour α et β "bien choisis", et en posant $Y=\frac{x_i-\beta}{\alpha}$, on a : $\bar{y} = \frac{\bar{x}-\beta}{\alpha}$, et donc $\bar{y}=0$ quand $\bar{x}=\beta$. On a donc un changement de variable et d'échelle dans un calcul barycentrique. Ceci permettait de calculer plus facilement la moyenne "à la main" avant l'arrivée des calculatrices. On a aussi : $\sigma_Y = \frac{\sigma_X}{|\alpha|}$.

Mode, moyenne et médiane forment les **caractéristiques de tendance centrale**.

2.1.2. paramètres de dispersion

Etendue E d'une série ordonnée (après éventuelle élimination de valeurs extrêmes) : $E=\max\{x_i\}-\min\{x_i\}$.

Variance : $V = \mathrm{Var}(X) = \frac{\sum n_i \left(x_i - \bar{x}\right)^2}{N}$.

Formule de Koenig : $V = \left(\frac{\sum n_i \, x_i^{\,2}}{N} \right) - \left(\bar{x} \right)^2$.

Ecart-type : $\sigma = \sqrt{V}$, moyenne quadratique des écarts à la moyenne.

Propriété : $0 \leq \sigma \leq E$.

"Ecart-type empirique" : $S = \frac{\sum n_i \left(x_i - \bar{x} \right)}{N - 1}$. Cette valeur est en général calculée comme approximation de l'écart-type par les calculatrices avec fonctions financières (en 1990). On montre que l'écart-type empirique d'un échantillon donne une meilleure estimation de l'écart-type de la population que ne le ferait le calcul de l'écart-type de l'échantillon.

Inégalité de Bienaymé-Tchebychev : les valeurs xi s'écartent de la moyenne $\bar{x}$ d'une valeur supérieure ou égale à un réel $e \in R_+^*$ donné avec une fréquence inférieure ou égale à $(\sigma/e)^2$.
Conséquence : plus 75% des valeurs x_i appartiennent à $]\, \bar{x} - 2\sigma, \bar{x} + 2\sigma [$.

Ecart-moyen : $\frac{\sum n_i \left| x_i - \bar{x} \right|}{N}$, moyenne des écarts à la moyenne.

Ecarts médians : $\max\{x_i\} - M$ et $M - \min\{x_i\}$. **Ecarts interquantiles** sur le même modèle ($Q_j - Q_i$ avec $j > i$ quelconques).

2.2. paramètres caractéristiques graphiques

Tracés des **fonctions de répartition** des effectifs, des fréquences.

"Camemberts"

2.2.1. cas "discret"

Diagramme en baton.

2.2.2. cas "continu"

Histogramme (fonction en escalier) : pour chaque rectangle, la surface est proportionnelle à l'effectif et à la fréquence de la classe et la hauteur est proportionnelle à la densité d'effectif ou de fréquence.

3. Etude des séries statistiques doubles

L'étude conjointe de plusieurs caractères sur une même population peut poser des problèmes : triangles rectangles $3^2+4^2=5^2$ et $5^2+12^2=13^2$, les valeurs moyennes ne forment pas un triangle rectangle : $4^2+8^2 \neq 9^2$. En cas de caractères multiples, il faut donc étudier le degré et la nature de leur interdépendance.

Deux variables statistiques simples X et Y sur une même population Ω forment une **variable statistique double** (X,Y) à laquelle on associe une **série statistique double** $(x_i,y_i) \longrightarrow n_{ij}$ qui correspond à un tableau à double entrée (x_i et y_i). **Distributions marginales** : *on peut rajouter au tableau une ligne et une colonne donnant respectivement les séries statistique de X et de Y*.

La série statistique double peut être représentée graphiquement par un **nuage pondéré de points** : X en abscisse et Y en ordonnée, chaque case du tableau correspond à un point de coordonnées (x_i,y_j) pondéré par n_{ij}. L'allure de ce nuage conduit à choisir, par **ajustement**, une relation mathématique Y=f(X) (une courbe continue : droite, puissance, logarithme, exponentielle,...) qui matérialise le plus exactement possible l'interdépendance des deux caractères.

Ajustement linéaire d'une série statistique double

L'ajustement se fait dans ce cas sur une fonction affine Y=f(X)=aX+b qui correspond à une droite, d'équation Y=aX+b, dite **droite de régression.** On calcule a et b de manière à minimiser la distance pondéré de chaque point du nuage à cette droite (**méthode des moindres carrés**).

Covariance de X et Y : $\mathrm{cov}(X,Y)=\sum\left(n_{ij}\frac{x_j y_i}{n}\right)-\bar{x}\,\bar{y}$.

N.B. : cov(X,X)=Var(X) ; cov(X,Y)=cov(Y,X) ; on peut avoir cov(X,Y)<0.

La droite d'ajustement <u>de Y en X</u> est la droite d'équation : $y-\bar{y}=\left(\frac{\mathrm{cov}(X,Y)}{\mathrm{Var}(X)}\right)(x-\bar{x})$

c'est-à-dire la droite d'équation $y=a_1x+b_1$ avec $a_1=\left(\frac{\mathrm{cov}(X,Y)}{\mathrm{Var}(X)}\right)$ et $b_1=(\bar{y}-a_1\bar{x})$.

Cette droite passe par le point de coordonnées $(\bar{x},\bar{y})$.

La droite d'ajustement <u>de X en Y</u> est la droite d'équation : $x-\bar{x}=\left(\frac{\mathrm{cov}(X,Y)}{\mathrm{Var}(Y)}\right)(y-\bar{y})$

c'est-à-dire la droite d'équation $x=a_2y+b_2$ avec $a_2=\left(\frac{\mathrm{cov}(X,Y)}{\mathrm{Var}(Y)}\right)$ et $b_2=(\bar{y}-a_2\bar{x})$

ou encore d'équation : $y-\bar{y}=\left(\frac{\mathrm{Var}(Y)}{\mathrm{cov}(X,Y)}\right)(x-\bar{x})$.

Coefficient de corrélation du couple (X,Y) : $r=\frac{\mathrm{cov}(X,Y)}{\sigma(X)\,\sigma(Y)}$.

<u>Propriétés</u> : $|r|\leq 1$.
Quand |r|=1, les deux droites d'ajustement sont confondues, les points du nuage de (X,Y) sont alignés.
L'ajustement est bon quand |r| est proche de 1 (supérieur à 0,7).
Quand r=0, les deux droites d'ajustement sont perpendiculaires, X et y ne sont pas corrélés.
L'ajustement est mauvais quand |r]<0,7, il faut rechercher alors une corrélation non linéaire.