STATISTIQUE DESCRIPTIVE

1. MÉTHODE STATISTIQUE

1.1. HISTORIQUE ET DÉFINITION

Aussi loin que l'on remonte dans le temps et dans l'espace (en Chine et en Égypte, par exemple), les États ont toujours senti le besoin de disposer d'informations sur leurs sujets ou sur les biens qu'ils possèdent et produisent. Mais les recensements de population et de ressources, les statistiques (du latin *status* : état) sont restées purement descriptives jusqu'au $17^{ème}$ siècle.

Puis s'est développé le calcul des probabilités et des méthodes statistiques sont apparues en Allemagne, en Angleterre et en France. Beaucoup de scientifiques de tous ordre ont apporté leur contribution au développement de cette science : PASCAL, HUYGENS, BERNOULLI, MOIVRE, LAPLACE, GAUSS, MENDEL, PEARSON, FISCHER etc....

Actuellement, beaucoup de domaines utilisent les méthodes statistiques (médecine, agronomie, sociologie, industrie etc....).

Définition : La Statistique, c'est l'étude des variations observables. C'est une méthode qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à les analyser et à les interpréter.

1.2. MÉTHODES STATISTIQUES

- $1^{ère}$ étape :On collecte des données :
  - ◊ soit de manière exhaustive
  - ◊ soit par sondage
- $2^{ème}$ étape : On trie les données que l'on organise en tableaux, diagrammes, etc...
- $3^{ème}$ étape : On interprète les résultats : on les compare avec ceux déduits de la théorie des probabilités.

On pourra donc :

⇒ évaluer une grandeur statistique comme la moyenne ou la variance (estimateurs, intervalles de confiance).

⇒ savoir si deux populations sont comparables (tests d'hypothèses).

⇒ déterminer si deux grandeurs sont liées et de quelle façon (corrélation, ajustement analytique).

Les conclusions, toujours entachées d'un certain pourcentage d'incertitude, nous permettent alors de prendre une décision.

2. SÉRIES STATISTIQUES A UNE VARIABLE

2.1. TERMINOLOGIE

POPULATION : Ensemble que l'on observe et qui sera soumis à une analyse statistique. Chaque élément de cet ensemble est un **individu** ou **unité statistique**.

ÉCHANTILLON : C'est un sous ensemble de la population considérée. Le nombre d'individus dans l'échantillon est la **taille** de l'échantillon.

CARACTÈRE : C'est la propriété ou l'aspect singulier que l'on se propose d'observer dans la population ou l'échantillon. Un caractère qui fait le sujet d'une étude porte aussi le nom de **variable statistique**.

Différents types de variables statistiques :

- Lorsque la variable ne se prête pas à des valeurs numériques, elle est dite **qualitative** (exemple : opinions politiques, couleurs des yeux...) .Elle peut être ordonnée ou non, dichotomique ou non.
- Lorsque la variable peut être exprimée numériquement, elle est dite **quantitative** (ou mesurable). Dans ce cas, elle peut être discontinue ou continue.
  - ♦ Elle est **discontinue** si elle ne prend que des valeurs isolées les unes des autres. Une variable discontinue qui ne prend que des valeurs entières est dite discrète (exemple : nombre d'enfants d'une famille).
  - ♦ Elle est dite **continue** lorsqu'elle peut prendre toutes les valeurs d'un intervalle fini ou infini (exemple : diamètre de pièces, salaires...).

2.2. COMMENT ORGANISER LES DONNÉES

On regroupe toutes les données de la série statistique dans un tableau indiquant la répartition des individus selon le caractère étudié. Le regroupement s'effectue par **classes** :

- Si le caractère est qualitatif ou discontinu, une classe contient tous les individus ayant la même modalité ou la même valeur du caractère.
- Si le caractère est continu, une classe est un intervalle.
  - ◊ Pour construire ces intervalles, on respecte les règles suivantes :
    1. Le nombre de classes est compris entre 5 et 20 (de préférence entre 6 et 12)
    2. Chaque fois que cela est possible, les amplitudes des classes sont égales.
    3. Chaque classe (sauf la dernière) contient sa borne inférieure mais pas sa borne supérieure.
  - ◊ Dans les calculs, une classe sera représentée par son centre, qui est le milieu de l'intervalle.

◊ Une fois la classe constituée, on considère les individus répartis uniformément entre les deux bornes (ce qui entraîne une perte d'informations par rapport aux données brutes).

◊ Que faut-il indiquer pour chaque classe ?

1. L'**effectif** : nombre d'individus de la classe : on le note n_i (i est l'indice de la classe).

2. La **fréquence** : proportion d'individus de la population ou de l'échantillon appartenant à la classe : on la note f_i.

f_i et n_i sont liés par : $\boxed{f_i = \frac{n_i}{N}}$ où N est le nombre total d'individus dans la population.

Remarque : On peut remplacer f_i par $f_i \times 100$ qui représente alors un pourcentage.

On a toujours : $\boxed{\sum_{i=1}^{k} n_i = N}$ $\boxed{0 \le f_i \le 1}$ $\boxed{\sum_{i=1}^{k} f_i = 1}$

où k représente le nombre de classes

3. L'**effectif** (ou la fréquence) **cumulé** (e) : effectif (ou fréquence) de la classe augmenté (e) de ceux (ou celles) des classes précédentes(lorsque la variable statistique est quantitative). La fréquence cumulée est une fonction F de la borne supérieure de la classe (dans le cas d'une variable statistique continue).

2.3. DIAGRAMMES

Ils servent à visualiser la répartition des individus.

- Pour une variable statistique qualitative :

  On utilise des **diagrammes à secteurs circulaires**, des **diagrammes en tuyaux d'orgue**, des **diagrammes en bandes**. Le principe est de représenter des aires proportionnelles aux fréquences de la variable statistique.

- Pour une variable statistique discrète :

  On utilise un **diagramme différentiel en bâtons**, complété du diagramme des fréquences cumulées appelé **diagramme cumulatif.** Le diagramme cumulatif est la représentation graphique d'une fonction F, appelée fonction de répartition de la variable statistique.

Exemple : nombre d'erreurs d'assemblage sur un ensemble d'appareils

nombre d'erreurs	nombre d'appareils	fréquences cumulées
0	101	0.26
1	140	0.61
2	92	0.84
3	42	0.94
4	18	0.99
5	3	1

Diagramme cumulatif

nombre d'erreurs d'assemblage

- Pour une variable statistique continue :

  1. Le diagramme représentant la série est un **histogramme** : ce sont des rectangles juxtaposés dont chacune des bases est égale à l'intervalle de chaque classe et dont la hauteur est telle que l'aire de chaque rectangle soit proportionnelle aux effectifs(histogramme des effectifs) ou aux fréquences de la classe correspondante (histogramme des fréquences).

  2. On obtient le **polygone des effectifs** (ou des fréquences) en reliant les milieux des bases supérieures des rectangles.

  3. La **courbe cumulative** (ou **polygone des fréquences cumulées**) est obtenue en portant les points dont les abscisses représentent la borne supérieure de chaque classe et les ordonnées les fréquences cumulées correspondantes, puis en reliant ces

points par des segments de droite. Son équivalent dans la théorie probabiliste est la **fonction de répartition.**

Exemple : nombre de ventes effectuées en un mois par 50 employés d'une compagnie

Dans cet exemple la variable statistique(le nombre de ventes), quoique discrète, doit être traitée comme une variable continue car elle prend un grand nombre de valeurs.

HISTOGRAMME

nombre de ventes : x	nombre d'employés	fréquences cumulées
$80 \le x < 90$	2	0.04
$90 \le x < 100$	6	0.16
$100 \le x < 110$	10	0.36
$110 \le x < 120$	14	0.64
$120 \le x < 130$	9	0.82
$130 \le x < 140$	7	0.96
$140 \le x < 150$	2	1

médiane

On remarque que : → F est une fonction croissante.
→ On a toujours : $\boxed{0 \le F(x) \le 1}$.

3. CARACTÉRISTIQUES NUMÉRIQUES D'UNE SÉRIE QUANTITATIVE

3.1. CARACTÉRISTIQUES DE POSITION

3.1.1. Le mode

Le **mode**, désigné par Mo est la valeur de la variable statistique la plus fréquente.

Dans le cas d'une variable statistique continue, on parle plutôt de **classe modale**.

NB : Le mode ou la classe modale n'est pas obligatoirement unique.

3.1.2. La médiane

La **médiane**, désignée par Me, est la valeur de la variable telle qu'il y ait autant d'observations, en dessous d'elle qu'au dessus ou, ce qui revient au même, la valeur correspondant à 50% des observations.

Comment la déterminer?

- Si la variable est discrète :

  On désigne par n le nombre d'observations .

  ⇒ Si n est impair : Me est la $(\frac{n+1}{2})$ ème observation.

  ⇒ Si n est pair : n = 2k. Me est la moyenne arithmétique des deux observations centrales.

$$Me = \frac{k^{ème}\,observation + (k+1)^{ème}\,observation}{2}$$

- Si la variable est continue, Me vérifie F(Me) = 0.5 ,où F est la fonction de répartition de la variable. On détermine alors un **intervalle médian**(intervalle contenant la médiane), puis on procède à l'intérieur de cette classe à une interpolation linéaire.

Généralisation : notion de **quantiles**

Quantile d'ordre 1/4 : C'est la valeur Q_1 tel que $F(Q_1) = 0.25$.
Quantile d'ordre 3/4 : C'est la valeur Q_3 tel que $F(Q_3) = 0.75$ (on a $Me = Q_2$).
Déciles d'ordre 1/10, 2/10.... : $F(D_1)=0.1$, $F(D_2)=0.2$...

Remarque : Ces éléments se déterminent facilement à partir des courbes cumulatives, en cherchant les abscisses des points d'ordonnées $\frac{n}{2}$ pour Me, $\frac{n}{4}$ pour Q_1...

3.1.3. La moyenne

Lorsque x désigne la variable statistique, la valeur moyenne, ou **moyenne** de la série se note m ou $\overline{x}$. Elle est l'analogue d'un centre de gravité.

1er cas : si les observations ne sont pas groupées (la série est dite **non classée**)

$$\boxed{\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j}$$ n = effectif total $x_j = j^{ème}$ valeur de la variable

2ème cas : si les observations sont groupées (la série est dite **classée**)

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

x_i = centre de la classe i

n_i = effectif de la classe i

n= effectif total

f_i = fréquence de la classe i

On effectue en fait ici une moyenne arithmétique pondérée.

NB : Dans le cas d'une variable continue, cette moyenne pondérée n'est qu'une valeur approchée de la vraie valeur moyenne de la série car on remplace chaque x_j par le centre de la classe à laquelle il appartient.

Pourquoi utiliser la moyenne arithmétique?

Elle a été choisie parmi d'autres types de moyenne (géométrique, harmonique...) car elle possède une propriété extrêmement intéressante:

Lorsqu'on se livre à des observations scientifiques, les mesures ne sont pas toujours exactement identiques d'une fois sur l'autre, même lorsque les conditions semblent être similaires. Il se produit ce que l'on appelle une erreur d'observation . On a la relation suivante :

valeur observée = valeur exacte + erreur d'observation

avec: x_i = valeur observée x_e = valeur exacte

$x_i - x_e$ = erreur d'observation

On décide alors de prendre pour x_e la valeur qui minimise les erreurs d'observation , en fait la moyenne des carrés de ces erreurs (critère des moindres carrés) . Le calcul prouve que la meilleure valeur estimant x_e suivant ce critère est $\bar{x}$.

Propriété : La moyenne $\bar{x}$ des valeurs observées d'une grandeur x correspond à la meilleure estimation de x_e .

Cela ne signifie pas que $\bar{x}$ soit la valeur exacte x_e de la grandeur observée mais que c'est la meilleure évaluation possible que l'on puisse en faire selon le critère des moindres carrés.

3.2. CARACTÉRISTIQUES DE DISPERSION

3.2.1. L'étendue

L'étendue, notée e, représente la différence entre les valeurs extrêmes de la distribution : $e = x_n - x_1$.

3.2.2. L'intervalle interquartile

L'intervalle interquartile, noté I, est la différence entre les deux quartiles Q_3 et Q_1 :

$$I = Q_3 - Q_1$$

Cet intervalle contient 50% de la population en en éliminant 25% à chaque extrémité. Cette caractéristique est nettement meilleure que l'étendue.

3.2.3. La variance

C'est la caractéristique de dispersion la plus utilisée avec l'écart quadratique moyen.

1er cas : série non classée

$$V_x = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2$$

2ème cas : série classée

$$V_x = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2 = \sum_{i=1}^{k} f_i(x_i - \overline{x})^2$$

Dans le cas d'une variable statistique continue, x_i représente le centre de la $i^{ème}$ classe.

La **variance** est donc toujours positive ou nulle. Les formules ci-dessus imposent de calculer les différences $(x_i-\overline{x})^2$ ce qui est assez fastidieux. On peut éviter cet inconvénient en utilisant le théorème de Koenig.

Autre expression de la variance **: Théorème de KOENIG**

1^{er} cas: série non classée

$$V_x = (\frac{1}{n}\sum_{j=1}^{n} x_j{}^2) - \overline{x}^2$$

$2^{ème}$ cas: série classée

$$V_x = (\frac{1}{n}\sum_{i=1}^{k} n_i x_i{}^2) - \overline{x}^2 = (\sum_{i=1}^{k} f_i x_i{}^2) - \overline{x}^2$$

Démonstration:

$$V_x = \sum_{i=1}^{k} f_i(x_i - \overline{x})^2 = \sum_{i=1}^{k} f_i(x_i{}^2 - 2x_i\overline{x} + \overline{x}^2) = \sum_{i=1}^{k} f_i x_i{}^2 - 2\overline{x}(\sum_{i=1}^{k} f_i x_i) + \overline{x}^2\sum_{i=1}^{k} f_i = \sum_{i=1}^{k} f_i x_i{}^2 - \overline{x}^2$$

car : $\sum_{i=1}^{k} f_i x_i = \bar{x}$ et $\sum_{i=1}^{k} f_i = 1$

3.2.4. Écart quadratique moyen

Par définition, l'**écart quadratique moyen** d'une série statistique est la racine carrée de la variance. On le note s_x

A la différence de la variance qui correspond à un carré, l'écart quadratique moyen est homogène à la variable statistique et s'exprime dans les mêmes unités. Il permet de mesurer la dispersion de la distribution statistique autour de sa valeur moyenne.

3.3. DÉTERMINATION GRAPHIQUE DE LA MOYENNE ET DE L'ÉCART QUADRATIQUE MOYEN D'UNE DISTRIBUTION GAUSSIENNE A L'AIDE DE LA DROITE DE HENRY

On connaît plusieurs distributions statistiques particulières donnant la fréquence théorique d'apparition d'une valeur x en fonction de x (on reviendra en détail sur ces notions dans les chapitres suivants) . L'une des plus importantes est la distribution gaussienne ou distribution normale.

La fréquence théorique d'apparition d'une valeur x_i est donnée par :

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)^2}$$

où m est la moyenne théorique et σ l'écart-type théorique de la distribution gaussienne.

La représentation graphique de cette fonction f' est la fameuse courbe "en cloche".

On associe à la variable X , la variable $T = \frac{X - m}{\sigma}$,appelée variable gaussienne centrée réduite (sa moyenne est nulle et son écart-type égal à 1).

Méthode de la DROITE DE HENRY :

Elle permet à la fois :

1) De tester si la distribution donnée est gaussienne.

2) Si elle l'est, de déterminer m et σ.

Il faut pour cela utiliser du **papier gausso-arithmétique** contenant :

1) en abscisses : les valeurs x, prises par la variable X.

2) en ordonnées :

- à droite : les valeurs t de la variable gaussienne centrée réduite T.
- à gauche : les valeurs de la fonction de répartition de la loi T.

Que faire?

1. On calcule pour chaque valeur de x_{is}, borne supérieure d'une classe, la fréquence cumulée correspondante $F(x_{is})$.

2. On porte les points de coordonnées(x_{is} , $F(x_{is})$ sur le papier gausso-arithmétique (en utilisant l'échelle de gauche des ordonnées).

Si les points sont alignés :

- La droite obtenue est la **droite de Henry** de la distribution.

- On en déduit que la variable statistique a une distribution gaussienne. Comme les fonctions de répartition des variables T et X, notées respectivement F et Π se correspondent, on a : $F(x_{is}) = \Pi(t_{is}) = \Pi(\frac{x_{is} - m}{\sigma})$ et la relation affine entre T et X est de la forme : $T = \frac{X - m}{\sigma}$

- On peut alors déterminer graphiquement m et σ :

L'abscisse du point d'intersection de la droite de Henry avec la droite d'équation t = 0 est m.

L'abscisse du point d'intersection de la droite de Henry avec la droite d'équation t = 1 est m+σ.