

1 Concepts fondamentaux de la statistique

Mots clés

Population, individus, échantillon, observations, caractères, données brutes, statistique descriptive, statistique inférentielle.

1. POPULATION – INDIVIDUS – ÉCHANTILLON

a) Population

L'ensemble sur lequel porte l'activité statistique s'appelle la **population**. Elle est généralement notée Ω pour rappeler la notation des probabilités mais par exemple dans la théorie des sondages elle est notée U , U comme Univers.

Remarques

1. Le terme de « population » est employé aussi bien lorsqu'il s'agit d'un ensemble d'êtres humains que d'un ensemble d'objets inanimés.
2. Il est recommandé de toujours bien définir la population étudiée.
3. Le nombre d'éléments contenus dans Ω est généralement noté N , ce qui s'écrit également $\text{Card}(\Omega) = N$ ou encore $|\Omega| = N$.

Exemple

Dans les [Fiches 7, 8, 9](#) et [10](#), le temps de travail moyen a été recensé sur l'ensemble des pays européens. Il faudra donc bien faire attention : dans cette application, tout ce qui est calculé l'est sur la **population** et non sur un **échantillon**.

b) Individus ou unités statistiques

Les éléments qui constituent la population sont appelés les **individus** ou encore les **unités statistiques**. Un individu est noté ω lorsque la population est notée Ω ou u lorsque la population est notée U .

Remarque

Ces « individus » peuvent être de natures très diverses.

Exemples

Ensemble de personnes, mois d'une année, pièces produites par une usine, résultats d'expériences répétées un certain nombre de fois...

c) Échantillon

Un **échantillon**, noté généralement S , S comme « *sample* », qui signifie échantillon en anglais, est une partie de la population prélevée soit de façon aléatoire, soit de façon déterministe. Les trois notions de population, d'individus et d'échantillon sont représentées sur la Figure 1.1.

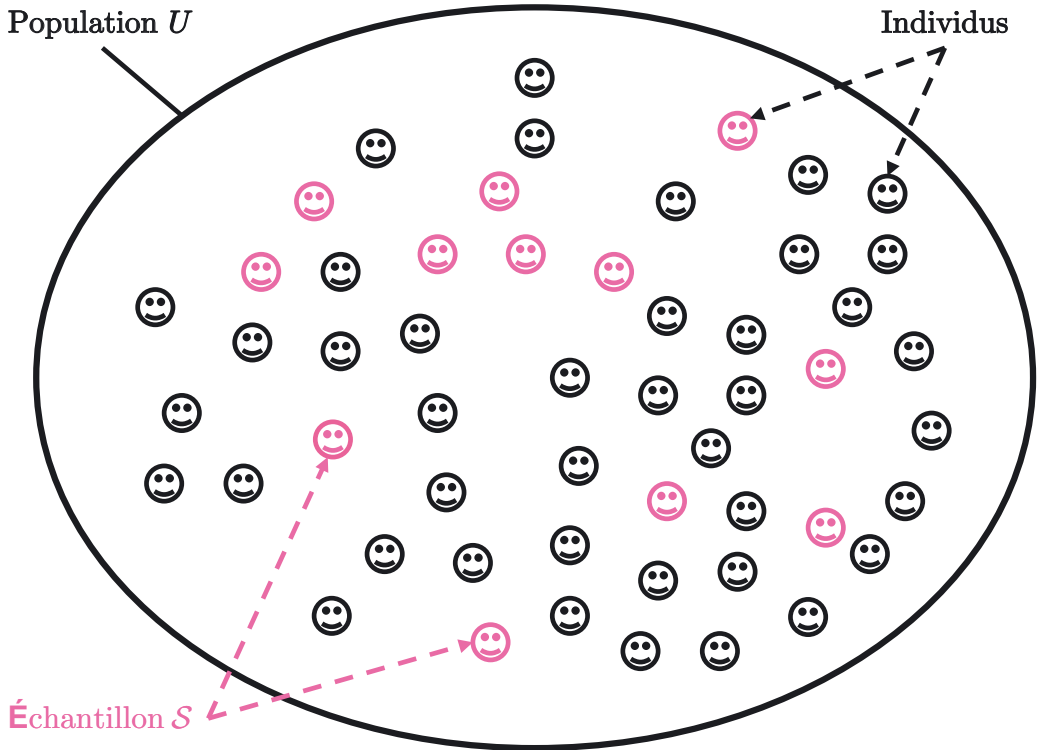


Figure 1.1 – Population, individus et échantillon.

2. OBSERVATIONS – CARACTÈRES – DONNÉES BRUTES

a) Observations

Le statisticien fait des relevés sur les unités statistiques : ce sont les **observations**.

b) Caractères

Les caractéristiques étudiées sur les individus d'une population sont appelées les caractères. Un **caractère** est une application χ d'un ensemble Ω fini de cardinal N (la population) dans un ensemble C (l'ensemble des valeurs possibles du caractère), qui associe à chaque individu ω de Ω la valeur $\chi(\omega)$ que prend ce caractère sur l'individu ω . Nous considérons deux types de caractères :

1. Les **caractères quantitatifs** : leur détermination produit un nombre ou une suite de nombres. Nous distinguons :

- les **caractères simples** ou **univariés** : leur mesure sur un individu produit un seul nombre. Un caractère quantitatif simple est dit **continu** si l'ensemble C des valeurs possibles du caractère est un intervalle. Il est dit **discret** si l'ensemble C des valeurs possibles du caractère est discret.

Exemples

Taille, poids, salaire, température, ...

- les **caractères multiples** : leur mesure sur un individu produit une suite finie de nombres. L'ensemble de leurs valeurs est donc \mathbb{R}^n ou une partie de \mathbb{R}^n .

Exemples

Relevé de notes d'un(e) étudiant(e), fiche de salaire, ...

2. Les **caractères qualitatifs** simples ou multiples.

Exemples

Profession, adresse, sexe, numéro de téléphone, ...

Remarques

1. La distinction entre les deux types est très importante. En effet, les méthodes d'analyse d'une population diffèrent suivant la nature du caractère étudié.

Exemple

Les représentations graphiques ne sont pas les mêmes.

2. Les caractères qualitatifs peuvent toujours être transformés en quantitatifs par codage. C'est ce qui se fait le plus généralement. Mais un tel codage est purement conventionnel et n'a pas vraiment un sens quantitatif.

Exemple

Nous ne pouvons pas calculer le sexe moyen.

3. Certains caractères qualitatifs s'expriment à l'aide de nombres (par exemple, un numéro de téléphone), mais ils n'ont pas non plus de sens quantitatif.

Exemple

Calculer un numéro de téléphone moyen n'est pas pertinent.

c) Données brutes

La suite des valeurs $\chi(\omega)$ prises par χ lorsque ω décrit toute la population Ω s'appelle les **données brutes**. C'est une suite finie (X_1, \dots, X_N) d'éléments de l'ensemble C des valeurs possibles du caractère. Ces valeurs X_i ne sont pas nécessairement distinctes : un caractère peut prendre la même valeur sur deux individus différents.

3. STATISTIQUE DESCRIPTIVE – STATISTIQUE INFÉRENTIELLE

a) Statistique ou statistiques ?

- **La statistique** désigne à la fois un ensemble de données et l'ensemble des activités consistant à collecter ces données, à les traiter et à les interpréter.
- **Les statistiques**, l'ensemble des données numériques, interviennent pratiquement dans tous les domaines d'activité : gestion financière (états, banques, assurances, entreprises, ...), démographie, contrôles de qualité, études de marché, sciences expérimentales (biologie, psychologie, etc.)...

b) Statistique descriptive

Le traitement des données, pour en dégager un certain nombre de renseignements qualitatifs ou quantitatifs généralement à des fins de comparaison, s'appelle la **statistique descriptive**. Elle ne s'applique que si les données ont été collectées sur toute la population.

c) Statistique inférentielle

Un autre but de la statistique consiste à extrapoler à partir d'un échantillon de la population à étudier, le comportement de la population dans son ensemble. C'est la **statistique inférentielle** également appelée **statistique inductive**.

Exemples

Sondages, contrôle de qualité comportant un test destructif et plus généralement l'analyse des résultats de toute expérience pour laquelle il n'a pas été possible d'étudier la population en entier.

2 Modalités, classes et tableaux statistiques

Mots clés

Modalités, classes, tableaux statistiques, forme empilée d'un tableau statistique.

1. MODALITÉS ET CLASSES

Le choix d'un caractère détermine le critère qui sert à comparer les individus entre eux.

a) Modalités

Dans le cas d'un caractère qualitatif, cela revient à classer les individus en plusieurs sous-groupes. Ils correspondent aux différentes valeurs que peut prendre le caractère qualitatif étudié. Ces dernières sont appelées **modalités**.

Remarques

1. Le nombre de modalités peut être fixé plus ou moins conventionnellement suivant la nature du caractère qualitatif étudié.

Exemples

Pour le caractère sexe, il y a deux modalités. Pour la profession, la répartition peut se faire suivant plusieurs critères. En effet, elle peut, par exemple, se faire soit avec le nombre d'années d'étude après le baccalauréat, soit suivant les classes socio-professionnelles.

2. Les modalités du caractère étudié doivent être simultanément exhaustives (chaque individu appartient à une modalité car tous les cas ont été prévus) et incompatibles (un individu ne peut appartenir à deux modalités ou plus).

Exemple

Un individu est soit marié, soit il ne l'est pas.

b) Classes

Dans le cas d'un caractère quantitatif simple, discret ou continu, les individus sont souvent regroupés dans des **classes** c'est-à-dire au sein d'intervalles de valeurs que peut prendre le caractère. L'utilisation de classes est plus fréquente dans le cas d'un caractère quantitatif simple continu.

Exemple

Lors d'une étude du caractère quantitatif simple continu « taille » sur la population française, il est possible de regrouper les observations en classes de 10 cm de largeur de 0 à 230 cm. Ce découpage aboutirait aux classes suivantes :]0 cm ; 10 cm],]10 cm ; 20 cm], ...,]220 cm ; 230 cm].

2. TABLEAUX STATISTIQUES

En fonction du contexte, il peut exister plusieurs possibilités pour représenter les données statistiques d'un échantillon à l'aide de tableaux. La solution qui peut être employée dans tous les cas,

et qui l'est le plus souvent par les logiciels statistiques, est de présenter les tableaux de données sous la **forme empilée**. Dans cette forme un **tableau statistique** est constitué :

- d'une ligne par individu,
- d'une colonne par caractère.

Concrètement, dans le cas de N observations d'un seul caractère nous pouvons avoir le **tableau de données brutes** suivant :

	X
Individu 1	X_1
Individu 2	X_2
...	...
Individu N	X_N

Dans le cas de N observations de deux caractères nous pouvons avoir le **tableau de données brutes** suivant :

	X	Y
Individu 1	X_1	Y_1
Individu 2	X_2	Y_2
...
Individu N	X_N	Y_N

Pour diminuer la place occupée par les tableaux, certaines personnes inversent malheureusement le rôle traditionnel des lignes et des colonnes dans les tableaux empilés. Si nous utilisons un logiciel pour réaliser nos analyses statistiques, il faudra penser à rétablir la convention précédente.

	Individu 1	Individu 2	...	Individu N
X	X_1	X_2	...	X_N
Y	Y_1	Y_2	...	Y_N

3 Statistique descriptive univariée

Mots clés

Séries statistiques simples quantitative et qualitative, distributions statistiques simples groupée ou non groupée, effectif, effectif cumulé, fréquence, fréquence cumulée.

1. SÉRIE STATISTIQUE

Les observations d'un ou plusieurs caractères récoltées sur toute la population forment une **série statistique**.

Les séries statistiques les plus élémentaires sont les séries à un seul caractère (**série statistique simple**) univarié (**série statistique simple univariée**) quantitatif ou qualitatif. Lorsque le caractère de la série statistique simple univariée est quantitatif discret (respectivement continu), la série est dite **série statistique simple discrète** (respectivement **continue**).

En général, ces séries sont représentées sous la forme d'un tableau statistique à une seule entrée puisqu'il n'y a qu'un seul caractère simple étudié. Parfois, la lecture de ces tableaux se révèle difficile. C'est pourquoi les représentations graphiques, qui sont présentées dans les **Fiches 4, 5, 6 et 9**, sont utilisées très fréquemment pour faciliter leur analyse.

2. DESCRIPTION D'UNE SÉRIE STATISTIQUE QUANTITATIVE

Soit X un caractère quantitatif simple discret ou continu. L'ensemble des valeurs atteintes par le caractère $X(\Omega) = \{X(\omega), \omega \in \Omega\} = \{X_1, \dots, X_N\}$ est un ensemble fini de valeurs distinctes $\{x_1, \dots, x_p\}$. Le fait que telle valeur soit relative à tel individu est un renseignement qui n'intéresse pas le statisticien. Seuls l'ensemble des valeurs atteintes et le nombre de fois que chacune d'elle est atteinte lui sont utiles.

Nous supposons que les valeurs x_1, \dots, x_p sont ordonnées dans l'ordre croissant : $x_1 < \dots < x_p$.

a) Distribution statistique non groupée

• Définitions

Soit une série statistique simple univariée, discrète ou continue, (X_1, \dots, X_N) qui prend pour valeurs x_1, \dots, x_p avec $x_1 < \dots < x_p$. Nous appelons :

1. **effectif de la valeur x_i** : le nombre n_i de fois que la valeur x_i est prise, c'est-à-dire le cardinal de l'ensemble $X^{-1}(x_i)$;
2. **effectif cumulé en x_i** : la somme $\sum_{j=1}^i n_j$;
3. **fréquence de la valeur x_i** : le rapport $f_i = \frac{n_i}{N}$ de l'effectif de x_i à l'effectif total N de la population, c'est-à-dire le cardinal de Ω ou encore la somme des n_i ;
4. **fréquence cumulée en x_i** : la somme $\sum_{j=1}^i f_j$.
5. La suite de couples $((x_i, n_i))_{i=1, \dots, p}$ ou $((x_i, f_i))_{i=1, \dots, p}$ est appelée **distribution statistique** (simple, discrète ou continue) **non groupée**.

• Remarques

1. Notons que, par définition, la somme des effectifs n_i est égale à N .
2. Notons que, par définition, la somme des fréquences est égale à un.

Il est bien sûr évident que tous ces indicateurs peuvent figurer dans un même tableau statistique. Leur lecture est alors plus rapide car les informations sont concentrées dans un seul élément statistique.

b) Distribution statistique groupée

- **Classes**

Lorsque le caractère X quantitatif discret ou continu comprend un grand nombre de valeurs, il est préférable de regrouper ces valeurs en intervalles appelés **classes** pour rendre la statistique plus lisible. L'ensemble C des valeurs du caractère est alors partagé en classes $]a_i; a_{i+1}]$ avec $a_i < a_{i+1}$.

- **Définitions**

Soit une série statistique $(]a_i; a_{i+1}], n_i)_{i=1, \dots, p}$. Nous appelons :

1. **effectif de la classe $]a_i; a_{i+1}]$** : le nombre n_i de valeurs prises par le caractère dans l'intervalle $]a_i; a_{i+1}]$, c'est-à-dire $X^{-1}(]a_i; a_{i+1}])$;
2. **effectif cumulé en a_i** : le nombre de valeurs prises dans l'intervalle $] - \infty; a_i]$;
3. **fréquence de la classe $]a_i; a_{i+1}]$** : le rapport $f_i = \frac{n_i}{N}$;
4. **fréquence cumulée en a_i** : la somme $\sum_{j=1}^i f_j$.
5. La série statistique $(]a_i; a_{i+1}], n_i)_{i=1, \dots, p}$ ou $(]a_i; a_{i+1}], f_i)_{i=1, \dots, p}$ est appelée **distribution statistique** (simple, discrète ou continue) **groupée**.

3. DESCRIPTION D'UNE SÉRIE STATISTIQUE QUALITATIVE

Pour un caractère qualitatif simple, les notions d'effectif et de fréquence s'appliquent également en considérant les différentes modalités que peut prendre le caractère qualitatif. Si les modalités du caractère qualitatif peuvent être ordonnées, il est également naturel de se servir des effectifs cumulés et des fréquences cumulées. Il est aussi vivement recommandé de résumer toutes les informations concernant le caractère sous forme d'un tableau statistique.