

PRÉSENTATION DES SÉRIES STATISTIQUES

→ *Séries statistiques*
 → *Regroupement en classes*
 → *Effectifs, fréquences simples*
 → *Fréquences corrigées*

→ *Fréquences cumulées*
 → *Diagrammes de fréquences*
 → *Histogrammes*

REPÈRES

1. Définitions

• Une *population* est l'ensemble de référence (ou ensemble fondamental) qui constitue l'enjeu de l'investigation statistique (que l'on cherche donc à connaître). Elle est composée de membres ou d'éléments. Pour des raisons techniques ou financières, on doit fréquemment se contenter d'étudier seulement une partie de celle-ci. Ce sous-ensemble s'appelle un *échantillon*.

• Le *caractère* (ou *variable statistique*) est le fait que l'on désire analyser. Cette notion peut être quantitative (taille, âge, production d'un atelier...) ou qualitative (sexe, profession, etc.).

On appelle *valeur du caractère* le nombre (ou l'index) qui permet de mesurer ou de repérer le caractère relatif à chaque membre de l'ensemble étudié (population ou, plus fréquemment, échantillon de population).

• Une *série statistique* est la suite des valeurs du (ou des) caractère(s) observé(s) sur chaque membre de l'ensemble étudié (population ou échantillon).

2. Séries statistiques

a) Types de données

Une série statistique est la suite des observations d'un (voire plusieurs) caractère, ou variable, relevées sur les individus d'une population. Exemples : la couleur des yeux des étudiants inscrits au cours de statistique, le PIB des différents pays de l'OCDE en 1988 mesuré en dollars, la production annuelle de vin d'appellation Saint-Émilion mesurée en milliers d'hectolitres de 1960 à 1987, etc.

Il est souhaitable que la suite des observations soit accompagnée des indications utiles pour une bonne compréhension (méthode, date, unité de mesure, etc.).

L'objectif est de mettre en évidence et d'étudier la distribution du caractère observé sur la population.

On distingue les variables qualitatives (par exemple la couleur des yeux, la catégorie socioprofessionnelle ou CSP, la tendance politique) des variables quantitatives ou numériques (par exemple le produit intérieur brut ou PIB, le poids, la note de statistique). Par ailleurs, certaines variables ne peuvent prendre qu'un petit nombre de valeurs isolées (la CSP, le nombre d'enfants...) et d'autres une infinité, ou du moins un très grand nombre (un poids, une durée, le revenu imposable...). Il est alors commode de regrouper les observations.

b) Regroupement en classes

Une série statistique numérique pouvant prendre un grand nombre de valeurs différentes est regroupée par intervalles de valeurs. En ne retenant ainsi d'une observation que l'intervalle auquel elle appartient, on perd une certaine partie de l'information initiale, mais on rend plus aisés sa manipulation et son examen.

Le découpage doit être défini sans ambiguïté, en particulier quant aux bornes des intervalles, afin de pouvoir affecter une observation à un intervalle unique.

On définit la *largeur* (ou l'*amplitude*) et le *centre* de chaque intervalle, qui peut alors être représenté par cette valeur centrale.

On peut regrouper une série par intervalles d'amplitudes égales ; on préfère en général découper plus finement les zones où se concentrent les valeurs observées et regrouper les données à l'intérieur d'intervalles d'amplitudes inégales.

3. Effectifs, fréquences

a) Effectifs, fréquences simples

Soit une série regroupée par classes de valeurs : x_1, x_2, \dots, x_k (éventuellement par intervalles, en retenant les valeurs centrales).

L'effectif n_1 de la première classe est le nombre d'observations valant x_1 , l'effectif n_2 est le nombre d'observations égales à x_2, \dots , l'effectif n_k le nombre d'observations égales à x_k . L'effectif total est :

$$N = n_1 + n_2 + \dots + n_k$$

c'est-à-dire le nombre d'individus de la population observée.

Les effectifs n_i étant liés à la taille de la population, on mesure mieux l'importance des différentes classes par les *fréquences*.

La fréquence f_i de la classe numéro i est le rapport :

$$f_i = \frac{n_i}{N}$$

c'est-à-dire la part de la classe numéro i dans l'ensemble de la population.

La somme des fréquences est égale à 1. Comme les fréquences sont des nombres inférieurs à 1, elles sont souvent données en pourcentages, c'est-à-dire sous la forme de fractions de dénominateur cent.

b) Fréquences corrigées

Lorsqu'il s'agit d'une série numérique regroupée en classes d'amplitudes inégales, les fréquences ne permettent pas d'apprécier la distribution du caractère (ainsi la fréquence d'un intervalle « étroit » ne peut être directement comparée à celle d'un intervalle dix fois plus large).

On ramène toutes les classes à une largeur standard, en calculant par proportionnalité les *fréquences corrigées* correspondantes : soit a l'amplitude standard (choisie librement), si la classe numéro i a pour fréquence f_i et pour amplitude a_i , sa fréquence corrigée est :

$$f'_i = f_i \cdot \frac{a}{a_i}$$

c) Fréquences cumulées

Lorsque les classes sont ordonnées, on définit les *fréquences cumulées croissantes* ; la fréquence cumulée de la classe numéro i est le rapport :

$$f_{C_i} = \frac{n_1 + n_2 + \dots + n_i}{N}$$

qui mesure la part cumulée des i premières classes dans l'ensemble de la population.

Lorsque l'on détermine les fréquences cumulées croissantes, la dernière calculée est égale à 1.

4. Représentations graphiques

Une grande variété de procédés sont employés pour visualiser la distribution d'une série statistique : indiquons les plus répandus.

a) Diagrammes de fréquences

Pour figurer la distribution en fréquences d'un caractère qualitatif ou d'un caractère quantitatif discontinu, on peut construire un *diagramme* en bâtonnets (fig. 1) ou en rectangles (fig. 2), dont les hauteurs traduisent les fréquences des différentes valeurs de la variable.

Figure 1

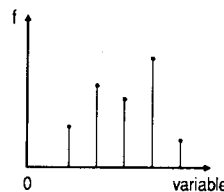
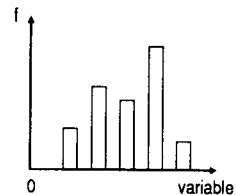


Figure 2



b) Histogramme

Pour une variable quantitative regroupée en classes, on construit les rectangles en prenant pour base les intervalles sur l'axe horizontal (fig. 3). Dans le cas d'amplitudes inégales, ce sont les fréquences corrigées qu'il faut figurer (fig. 4) ; ainsi, dans tous les cas, l'allure de l'*histogramme* traduit bien la distribution de la variable ; les rectangles les plus élevés indiquent les régions de forte densité des valeurs observées.

Précisément, la propriété fondamentale des histogrammes qu'il importe de retenir est que l'*aire* de chaque rectangle est proportionnelle à l'effectif (ou à la fréquence) de la classe correspondante.

Figure 3

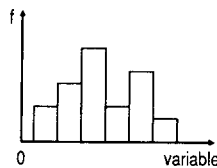
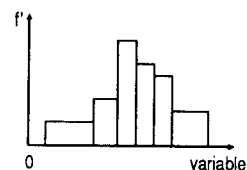
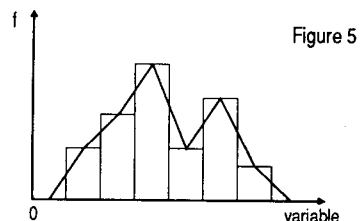


Figure 4

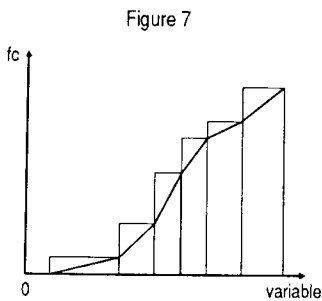
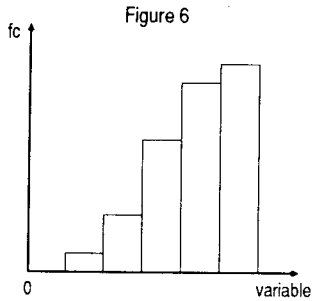


Dans le cas de classes de même amplitude, on peut construire le *polygone des fréquences*, en joignant les points situés au milieu de chaque intervalle (fig. 5). De manière à tracer le polygone des fréquences à partir de l'axe horizontal, on crée deux classes vides fictives, ayant l'amplitude commune, aux deux extrémités de l'histogramme et l'on fait passer la « courbe » par le milieu de ces deux classes.



c) Représentation des fréquences cumulées

Pour une variable quantitative regroupée en classes, on construit de même l'histogramme des fréquences cumulées en prenant pour base les intervalles. On obtient alors un « escalier », et le polygone des fréquences cumulées joint les points situés aux bornes de chaque intervalle (fig. 6 et 7).



d) Usages

La détermination des fréquences, puis leur représentation graphique, permettent d'observer la répartition d'un caractère dans une population et de comparer les formes de plusieurs distributions.

On peut mentionner quelques formes types : distribution *unimodale* (fig. 8), distribution *bimodale* (fig. 9), distribution *monotone descendante* (fig. 10), distribution *symétrique* (fig. 11), distribution *asymétrique* (fig. 12), etc.

Figure 8

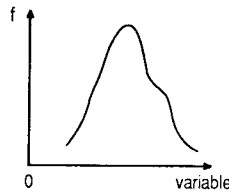


Figure 9

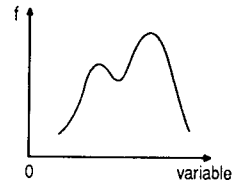


Figure 10

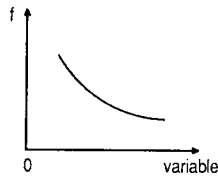


Figure 11

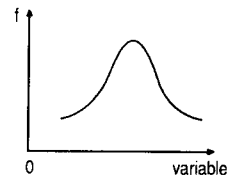


Figure 12

