Statistiques descriptives

Chapitre 1 - Unités statistiques, caractères et rappels mathématiques importants

Ce cours vous est proposé par Olivier Baron, Maître de conférences, Université de Bordeaux et par AUNEGe, l'Université Numérique en Économie Gestion.

Table des matières

Préambule ***2***

Introduction et vocabulaire ***2***

Les différents types de caractères et de variables statistiques ***4***

Les caractères qualitatifs **4**

Les caractères quantitatifs **5**

Les variables statistiques discrètes 5

Les variables statistiques continues 5

Rappel de notions mathématiques simples utiles en statistique descriptive ***7***

Propriétés des logarithmes **7**

Les opérations sur les exposants **8**

Les pourcentages **8**

Exemples 9

Les opérateurs Σ et Π **10**

Propriétés remarquables **12**

Interpolation linéaire **12**

Binôme de Newton **13**

Références ***15***

Préambule

Objectifs

Dans ce chapitre premier nous allons présenter les notions essentielles permettant d'aborder sereinement l'étude et l'analyse des distributions statistiques à un caractère. La présentation du vocabulaire dédié, celle des différents types de variables et de caractères ainsi que quelques rappels mathématiques importants feront l'objet des sections à suivre.

Introduction et vocabulaire

La statistique descriptive vise à étudier et analyser de manière quantitative des ensembles nombreux et variés. On appellera **population statistique** l'ensemble sur lequel porte l'analyse statistique qui convient d'être menée. Cet ensemble peut prendre différentes formes : êtres humains, ménages, entreprises, secteurs d'activité, pays, stock ou flux de biens de consommation ou de production, biens matériels ou immatériels, ensembles concrets ou non concrets observés à un moment donné ou sur une période de temps fixée.

Cette population est constituée d'éléments dénombrables, ou pas, appelés **individus statistiques**. Ces individus peuvent bien entendu être des êtres humains mais aussi tout autre éléments constitutifs de la population étudiée.

Pour décrire la population étudiée on observe un ou plusieurs **caractères**[1] des individus la constituant. Ces caractères sont bien entendu en lien avec la population d'individus concernés par l'étude statistique engagée : âge, genre, catégorie socio-professionnelle pour des individus au sens strict, chiffre d'affaire ou taille pour des entreprises, diamètre pour des pièces de fonderie d'une certaine qualité, mode de transport pour l'étude des mobilités urbaine, montant, durée et taux d'intérêt pour l'étude des prêts immobiliers dans une certaine région…

Ces caractères prennent généralement plusieurs valeurs, plusieurs situations différentes sur l'ensemble des individus de la population étudiée. Ces valeurs, numériques ou non, seront nommées modalités. Le nombre de modalités d'un caractère peut être fini (2 modalités

[1] On parlera aussi de variables, terme générique pour nommer le ou les caractères étudiés.

pour le genre : masculin / féminin ; plusieurs modalités pour le mode transport en milieu urbain : voiture / tram / Bus / vélo / marche à pieds / …) ou infini (revenu mensuel d'un ménage ; chiffre d'affaires d'une entreprise ; …). Lorsque le nombre de modalités d'un caractère est fini, celles-ci doivent être exhaustives (toutes les situations possibles doivent avoir été prévues) et chaque individu doit présenter une et une seule modalité de ce caractère (on dit que les modalités du caractère sont incompatibles).

Le nombre d'individus de la population présentant la même modalité d'un caractère donné est appelé l'**effectif** de cette modalité. Généralement on représente l'ensemble des informations associées à un caractère sous la forme d'un tableau. Ce tableau est dit à une dimension puisqu'il ne concerne qu'un seul caractère et, une fois remplit, décrira la **distribution statistique** du caractère au sein de la population étudiée.

On peut synthétiser l'ensemble de la terminologie statistique qui vient d'être présentée à l'aide de l'exemple suivant :

Population :	Ensemble des entreprises françaises	E
Individu :	Une entreprise	
Caractère :	Secteur d'activité	X
Modalités :	Secteurs primaire, secondaire, tertiaire	x_1 , x_2 , x_3
Effectifs :	Nombre d'entreprises du secteur primaire	n_1
	Nombre d'entreprises du secteur secondaire	n_2
	Nombre d'entreprises du secteur tertiaire	n_3

Distribution statistique de X dans E

Modalités x_i	Effectifs n_i
x_1	n_1
x_2	n_2
x_3	n_3
	n

Les différents types de caractères et de variables statistiques[2]

Un caractère peut être soit qualitatif, soit quantitatif. Dans ce premier cas on peut associer au caractère une variable nominale ou ordinale et dans le second, une variable statistique discrète ou continue. Cette distinction est importante car les méthodes d'analyse statistique diffèrent suivant la nature du caractère étudié. Les modalités de classement et de représentation graphique ne sont pas les mêmes et seuls les caractères quantitatifs se prêtent au calcul de caractéristiques de tendance centrale, de dispersion, etc…

Les caractères qualitatifs

Un caractère est dit qualitatif lorsque ses modalités ne peuvent pas être traduites par une mesure, mais peuvent seulement être constatées. Le sexe, la nationalité, la profession, la situation matrimoniale, la zone de résidence, … sont des caractères qualitatifs[3]. Comme cela a été précisé dans la section précédente les modalités d'un tel caractère doivent être exhaustives et mutuellement incompatibles de telle sorte qu'un individu quelconque puisse être classé dans une et une seule de celles-ci. Avant de procéder au classement des unités statistiques suivant un caractère qualitatif, il sera nécessaire d'établir ou d'adopter une nomenclature. Pour établir celle-ci, il faut d'abord dresser la liste de tous les cas possibles. Pour certains caractères, le sexe, la situation matrimoniale par exemple, l'opération est simple : on peut être un homme ou une femme ; célibataire, marié, veuf ou divorcé. Pour d'autres, l'activité économique, la profession…, il y a des milliers de cas possibles. Il alors nécessaire de procéder à des regroupements des cas élémentaires pour constituer les rubriques de la nomenclature[4].

---

[2] En statistiques, les termes "caractère statistique" et "variable statistique" sont souvent utilisés pour décrire des concepts différents mais interdépendants. Lorsqu'on parle d'un caractère statistique, il s'agit d'une description qualitative ou quantitative d'un phénomène. Une variable statistique représente les différentes valeurs que peut prendre un caractère statistique.

[3] On pourra bien entendu coder les modalités de tels caractères par des chiffres ou des nombres. Par contre, aucune opération algébrique n'est possible sur ces nombres.

[4] Pour un exemple de nomenclature, consulter le site Insee.fr. On y trouve, en particulier, la NAF (Nomenclature des Activités Françaises). Nous avons ici un exemple de nomenclature, c'est à dire de tableau qualitatif à une dimension, dont les rubriques s'emboitent en divers niveaux : 21 lignes pour le niveau 1 et jusqu'à 732 lignes pour le dernier niveau proposé.

On distinguera au sein des caractères qualitatifs, ceux décrits par une variable nominale de ceux décrits par une variable ordinale. Dans le premier cas, aucune relation d'ordre ne peut être établie sur l'ensemble des modalités (le sexe, la couleur des yeux, …), alors que dans le second cas une telle relation d'ordre existe (niveau de satisfaction – mauvais, bon, très bon – niveau d'éducation, …).

Les caractères quantitatifs

On dit qu'un caractère est quantitatif lorsqu'il est mesurable. A chaque unité statistique correspond alors un nombre qui est la mesure du caractère. A ce nombre, on donne le nom de variable statistique et les modalités du caractère sont les valeurs possibles de la variable statistique correspondante. Ces variables statistiques peuvent être de deux types différents : discrètes ou continues.

Les variables statistiques discrètes

Une variable statistique est discrète[5] lorsqu'elle ne peut prendre que certaines valeurs isolées dans son intervalle de variation. Il s'agit, en général, de valeurs entières. Le nombre de personnes d'un ménage, le nombre d'accidents du travail survenus dans un établissement, le nombre de pièces d'un appartement, le nombre de salariés d'une entreprise sont des variables statistiques discrètes. Les modalités du caractère associé seront soit les valeurs possibles de la variable lorsque celles-ci sont peu nombreuses (1, 2, 3, … pièces dans un appartement), soit des regroupements, des classes, de valeurs possibles lorsque celles-ci sont très nombreuses. Par exemple, dans le cas du nombre de salariés d'une entreprise, les classes, qui constituent les modalités du caractère pourront être : moins de 5 salariés, de 5 à 9 salariés, de 10 à 19 salariés, de 20 à 49 salariés, de 50 à 499 salariés, 500 salariés et plus.

Les variables statistiques continues

Une variable statistique est continue lorsqu'elle peut prendre toutes les valeurs à l'intérieur de son intervalle de variation. Ce nombre de valeurs possibles est alors infini puisqu'il dépend de la précision de la mesure effectuée. De façon générale les grandeurs liées à l'espace (longueur, distance, surface), au temps (âge, vitesse, durée), à la masse (poids, teneur) ou à la valeur monétaire (revenus, chiffre d'affaires) sont des variables continues.

[5] Du latin discretus, qui signifie « séparé » ; dans un ensemble discret, on peut séparer les éléments.

Les variables statistiques continues pouvant prendre un si grand nombre de valeurs différentes, leur présentation nécessite de regrouper ces valeurs en classes[6].

La notion de classe implique la connaissance d'un certain vocabulaire. L'**amplitude** d'une classe est la « longueur » de l'intervalle choisi. Elle peut être constante ou variable. On la note a_i et, pour une classe regroupant les valeurs de la variable allant de x_i à x_j, on aura :

$$a_i = x_j - x_i$$

Les **extrémités** (ou **bornes**) de classe sont les valeurs qui encadrent les classes. Il convient de les définir de façon rigoureuse de manière à pouvoir classer sans ambiguïté chaque individu de la population étudiée. Par convention, on présente les intervalles « borne comprise à gauche – borne non comprise à droite », soit de la façon suivante :

$$[x_i\ ;\ x_j[$$

Le **centre**[7] de classe, noté c_i, est la moyenne arithmétique des bornes de la classe. Ainsi, pour la classe précédente $[x_i\ ;\ x_j[$, on aura :

$$c_i = \frac{x_i + x_j}{2}$$

Le choix du nombre de classes et de leur amplitude se fait en fonction de l'effectif de la population, de façon à ce que le nombre d'unités statistiques dans chaque classe soit suffisant pour éliminer les variations accidentelles qui se produisent lorsqu'on considère de trop faibles effectifs. En règle générale, on choisit les amplitudes de classes de façon à ce que les effectifs de chaque classe soient environ du même ordre de grandeur. Ce souci conduit, pour les distributions les plus fréquentes (distributions de revenus, de chiffre d'affaires, …), à adopter des classes de faible amplitude au centre de la distribution, là où les observations sont les plus nombreuses, et de plus grande amplitude aux extrémités. Par ailleurs, le nombre de classes doit être suffisant et les amplitudes assez faibles pour ne pas masquer certaines particularités de la distribution. Toute diminution du nombre de classes et toute augmentation de l'amplitude de celles-ci conduit, en effet, à une perte d'information. Un compromis doit donc être trouvé entre les risques d'irrégularités

[6] Attention, si les variables continues sont nécessairement regroupées en classes, l'existence de classes n'implique pas que l'on ait nécessairement à faire à une variable continue. Certaines variables discrètes présentant un grand nombre de modalités possibles, comme le nombre de salariés dans une entreprise par exemple, peuvent, elles aussi, être présentées en classes.

[7] Le centre de classe est appelé à jouer un grand rôle dans les calculs, car le regroupement en classes constitue une perte d'information importante.

accidentelles, l'ampleur des calculs résultant de la prise en compte d'un trop grand nombre de classes et la perte d'information provenant de regroupements excessifs. En général, on évitera de constituer plus d'une dizaine de classes.

En conclusion de ces deux premières sections, toute étude de statistique descriptive devra être précédée d'une identification claire de la population, du caractère étudié et de sa nature, à savoir qualitatif ou quantitatif et, dans le cas quantitatif, discret ou continu.

Rappel de notions mathématiques simples utiles en statistique descriptive

Propriétés des logarithmes

Définition

Il existe une unique fonction dérivable sur $\mathbb{R}_+^*$ telle que $f'(x) = \frac{1}{x}$ et $f(1) = 0$. Cette fonction est le **logarithme népérien** et est notée **ln**.

Propriétés :

$\ln 1 = 0$ et $\ln e = 1$ où e est un nombre irrationnel, appelé « base » des logarithmes naturels.

$\ln ab = \ln a + \ln b$ (avec $a > 0$ et $b > 0$)

$\ln \frac{a}{b} = \ln a - \ln b$ (avec $a > 0$ et $b > 0$)

$\ln a^n = n \ln a$ (avec $a > 0$ et $n \in \mathbb{Z}$)

Fonction réciproque : La fonction g définie sur $\mathbb{R}$ par $g(x) = e^x$ (fonction exponentielle) est la fonction réciproque de la fonction ln. On aura : $y = e^x \Leftrightarrow x = \ln y$.

Les opérations sur les exposants

Définition

Le produit d'un nombre a, m fois par lui-même s'écrit a^m.

$$a^m = \underbrace{a \times a \times \ldots \times a}_{m}$$

L'exposant m peut être un nombre entier positif ou négatif, ou un nombre fractionnaire.

Par exemple : $a^{1/2} = \sqrt{a}$; $a^{-m} = \frac{1}{a^m}$; $a^{1/m} = \sqrt[m]{a}$; $a^{-1/2} = \frac{1}{\sqrt{a}}$.

Règles de calcul :

$a^m \times a^n = a^{m+n}$

$\frac{a^m}{a^n} = a^{m-n}$

$a^0 = 1$

$(a^m)^n = a^{mn}$

. $(a \times b)^m = a^m \times b^m$

$\left(\frac{a}{b}\right)^m = \frac{a^m}{b^m}$

Les pourcentages

Définitions :

Le signe % (pourcentage) revient à diviser par 100 : $7\% = \frac{7}{100} = 0{,}07$

Si y_1 et y_2 sont les valeurs prises par une variable y respectivement aux instants 1 et 2 : $(y_2 - y_1)$ est la variation absolue de y entre 1 et 2, $\frac{(y_2 - y_1)}{y_1}$ est la variation relative (ou taux de variation) de y entre 1 et 2, $(\frac{y_2}{y_1} * 100)$ est l'indice base 100 de y entre 1 et 2 et $\frac{y_2}{y_1}$ est le coefficient multiplicateur.

Remarques :

Si l'on note t le taux de variation de y entre 1 et 2 et CM le coefficient multiplicateur alors $CM = t + 1$.

Pour augmenter de *t* % on multiplie par $\left(1+\frac{t}{100}\right)$ et pour diminuer de t % on multiplie par $\left(1-\frac{t}{100}\right)$.

Exemples

Exemple 1

Années	2018	2020	2022
Prix du baril de pétrole en $	50,38	40,34	77,94

Calculer la variation absolue et la variation relative du prix du baril de pétrole entre 2018 et 2020 puis entre 2020 et 2022.

Entre 2018 et 2020 :

Variation absolue : $40{,}34-50{,}38=-10{,}04$ donc le prix du baril a diminué de 10,04 $. Variation relative : $\frac{40{,}34-50{,}38}{50{,}38}=-0{,}1993$ donc le prix du baril a diminué de 19,93 %.

Entre 2020 et 2022 :

Variation absolue : $77{,}94-40{,}34=37{,}6$ donc le prix du baril a augmenté de 37,6 $. Variation relative : $\frac{77{,}94-40{,}34}{40{,}34}=0{,}9321$ donc le prix du baril a augmenté de 93,21 %.

Exemple 2

Un article de 145 € a baissé de 4,5%, son nouveau prix est de :$145\times\underbrace{\left(1-\frac{4{,}5}{100}\right)}_{CM}=138{,}48€$

Un article de 67 € a augmenté de 12,3%, son nouveau prix est de : $67\times\underbrace{\left(1+\frac{12{,}3}{100}\right)}_{CM}=75{,}24€$.

Évolutions successives – taux de variation global

Il ne faut jamais additionner les taux de variation successifs pour obtenir le taux de variation global. La méthode à suivre est la suivante :

- Transformer les taux de variation successifs en coefficients multiplicateurs,
- Multiplier entre eux les coefficients multiplicateurs,

- Transformer le résultat obtenu en taux de variation pour obtenir le taux de variation global.

En reprenant les données du tableau précédent sur l'évolution du prix du baril de pétrole on obtient : Notons CM_1 le coefficient multiplicateur de 2018 à 2020 et CM_2 le coefficient multiplicateur de 2020 à 2022. Nous avons $CM_1 = 1 - 0{,}1993 = 0{,}8007$ et $CM_2 = 1 + 0{,}9321 = 1{,}9321$. Le coefficient multiplicateur de 2018 à 2022 vaut $CM = CM_1 \times CM_2 = 0{,}8007 \times 1{,}9321 = 1{,}5470$. Pour obtenir le taux de variation global, on enlève 1 au résultat obtenu et on trouve 0,547. Le prix du baril a donc augmenté de 54,7% entre 2018 et 2022.

Remarque

Des hausses et des diminutions identiques de taux de variation ne sont pas symétriques. En effet, si le prix d'un produit augmente de 15% puis diminue de 15%, le taux de variation global du prix de ce produit sera de :$[(1{,}15) \times (0{,}85)] - 1 = -0{,}0225$ soit une baisse de 2,25%.

Taux réciproque : La question posée est la suivante : Comment revenir à la situation initiale après une évolution de taux *t* ? La solution de ce problème est simple : Il suffit que le coefficient multiplicateur global des deux évolutions successives soit égal à 1. En notant t' le taux réciproque au taux de variation t, on aura : $(1 + t) \times (1 + t') = 1$ soit :

$$(1 + t') = \frac{1}{(1 + t)}$$

Ainsi, pour compenser une hausse de 15% au cours d'une période, il faut appliquer une baisse de 13,04% à la période suivante $\left(\frac{1}{1{,}15} - 1 = -0{,}1304\right)$.

Les opérateurs Σ et Π

Il s'agit de symboles de notation souvent utilisés en statistique descriptive.

Définition 1

Quand une variable statistique prend les valeurs $x_1, x_2, \ldots, x_n$, on symbolise leur somme par :

$$x_1 + x_2 + \cdots + x_n = \sum_{i=1}^{n} x_i$$

Propriétés du symbole Σ :

Si chaque terme est multiplié par une constante a alors :

$$\sum_{i=1}^{n} a.x_i = a.\sum_{i=1}^{n} x_i$$

Décomposition des sommes :

$$\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

Si $\exists k \in \mathbb{N}$ tel que $1 < k < n$ alors :

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{k} x_i + \sum_{i=k+1}^{n} x_i$$

Sommation d'une constante :

$$\sum_{i=1}^{n} a = \underbrace{a + a + \cdots + a}_{n} = n.a$$

Définition 2

Quand une variable statistique prend les valeurs $x_1, x_2, \ldots, x_n$, on symbolise leur produit par :

$$x_1 \times x_2 \times \ldots \times x_n = \prod_{i=1}^{n} x_i$$

Propriétés du symbole Π :

Si chaque terme est multiplié par une constante a alors :

$$\prod_{i=1}^{n} a.x_i = a^n . \prod_{i=1}^{n} x_i$$

Décomposition des produits :

$$\prod_{i=1}^{n} x_i . y_i = \prod_{i=1}^{n} x_i . \prod_{i=1}^{n} y_i$$

Si $\exists k \in \mathbb{N}$ tel que $1 < k < n$ alors :

$$\prod_{i=1}^{n} x_i = \prod_{i=1}^{k} x_i \cdot \prod_{i=k+1}^{n} x_i$$

Multiplication d'une constante :

$$\prod_{i=1}^{n} a = a^n$$

Propriétés remarquables

La somme des *n* premiers entiers est égale à :

$$\sum_{i=1}^{n} i = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$

La somme des n premiers carrés est égale à :

$$\sum_{i=1}^{n} i^2 = 1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

La somme des n premiers cubes est égale à :

$$\sum_{i=1}^{n} i^3 = 1^3 + 2^3 + \cdots + n^3 = \frac{n^2(n+1)^2}{4} = \left[\frac{n(n+1)}{2}\right]^2$$

Si q est un nombre réel positif différent de 1, alors :

$$\sum_{i=0}^{n} q^i = q^0 + q^1 + q^2 + \cdots + q^n = \frac{1 - q^{n+1}}{1 - q}$$

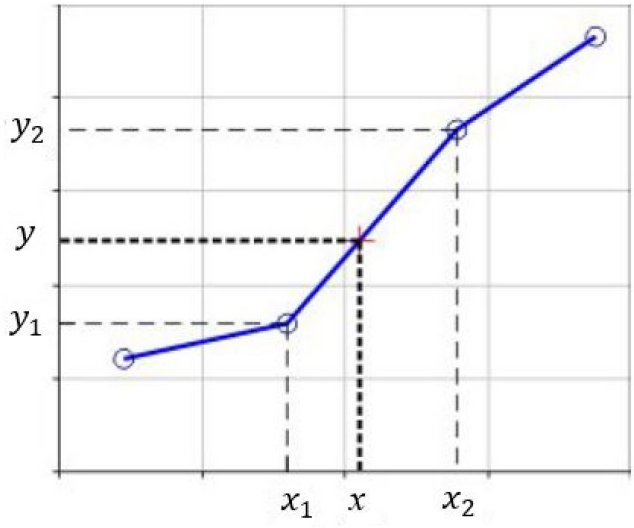
Double somme :

$$\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} = a_{11} + a_{12} + \cdots + a_{1m} + a_{21} + \cdots + a_{2m} + \cdots + a_{n1} + \cdots + a_{nm}$$

Interpolation linéaire

L'interpolation linéaire est la méthode la plus simple pour estimer la valeur prise par une fonction continue entre deux points déterminés.

Figure 1 : Interpolation linéaire



Si les valeurs connues sont (x_1, y_1) et (x_2, y_2) alors la valeur de *y* en un certain point x compris entre x_1 et x_2 est donnée par :

$$y = y_1 + (y_2 - y_1).\frac{(x - x_1)}{(x_2 - x_1)}$$

Binôme de Newton

Les identités suivantes sont souvent utilisées :

$$(a+b)^0 = 1$$

$$(a+b)^1 = a + b$$

$$(a+b)^2 = a^2 + 2ab + b^2$$

$$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$$

Ces formules sont des cas particuliers d'une formule plus générale, appelée formule du binôme de Newton, qui donne le développement de $(a+b)^n$, où *n* est un entier naturel quelconque :

$$(a+b)^n = \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{n-1}ab^{n-1} + \binom{n}{n}b^n$$

Dans cette expression, les termes $\binom{n}{k}$ sont appelés coefficients binomiaux, et sont définis pour $n = 1, 2, \ldots$ et $k = 1, \ldots, n$ par :

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

où $n! = 1 \times 2 \times 3 \times \ldots \times (n-1) \times n$ et avec la convention $0! = 1$.

Références

Comment citer ce cours ?

Statistiques descriptives, Olivier Baron, AUNEGe (http://aunege.fr), CC – BY NC ND (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Table des illustrations :

Figure 1 : Interpolation linéaire 13