L1 MASS 2013-2014

Statistique descriptive
Notes de cours

Hélène Boistard
Université Toulouse 1 - Capitole
www.boistard.fr

Table des matières

Chapitre 1

Les données statistiques

1.1 Les variables statistiques - éléments de vocabulaire

On observe un **échantillon** composé de n **individus** appartenant à une même **population** de taille N. Chaque individu de l'échantillon est observé à travers des caractéristiques, caractères ou indicateurs appelés **variables**. Une **série statistique** $\{x_1, x_2, \ldots, x_n\}$ est la suite des valeurs prises par une ou plusieurs variables pour chacun des individus de l'échantillon.
Exemple : un questionnaire est distribué à 20 personnes. Il comporte diverses questions. La population = l'échantillon = les étudiants ayant répondu au questionnaire. Les individus sont les personnes interrogées. Les variables correspondent aux questions posées : l'âge, la taille, la couleur des yeux, etc.
Schéma :

1.2 Les types de variables

1.2.1 Variables qualitatives

Une variable est appelée **qualitative** lorsque les réponses possibles à la question posée, ou les valeurs prises par la variable, ne correspondent pas à une quantité mesurable par un nombre mais appartiennent à un groupe de **catégories**. On les appelle **modalités** de la variable.
Exemple : le sexe, la couleur des yeux, la mention au baccalauréat, la fréquence d'une activité (jamais, rarement, parfois, souvent, très souvent).
On distingue :

- les variables **qualitatives nominales** : il n'y a pas de hiérarchie entre les différentes modalités ; exemple : sexe, couleur des yeux.
- les variables **qualitatives ordinales** : les différentes modalités peuvent être ordonnées de manière naturelle ; exemple : la mention au baccalauréat, la fréquence d'une activité.

Remarque : certaines variables nominales peuvent être désignées par un code numérique, qui n'a pas de valeur de quantité. Exemple : le code postal, le sexe (1=garçon, 2=fille).

1.2.2 Variables quantitatives

Les réponses correspondent à des quantités mesurables et sont données sous forme de nombre.
On distingue :

- les variables quantitatives discrètes : elles prennent leurs valeurs dans un ensemble discret, le plus souvent fini ; exemple : le nombre d'enfants, la pointure du pied.
- les variables quantitatives continues : elles peuvent prendre toutes les valeurs d'un intervalle réel ; exemple : la taille des individus, une note à un examen.

Remarque : l'âge peut être vu et traité comme une variable quantitative discrète ou continue suivant la précision que l'on choisit et le nombre de valeurs qu'il prend au sein de la population. Il peut également exister des variables basées sur l'âge qui sont qualitatives. Si dans un sondage on pose la question "quelle est votre tranche d'âge parmi les possibilités suivantes : - de 25 ans, entre 25 et 40, entre 40 et 60 et + de 60 ans", on peut voir la variable "tranche d'âge" comme une variable qualitative ordinale.

1.3 Les variables qualitatives : tableaux de fréquence et représentation graphique

Exemple : On s'intéresse à la variable "couleur des yeux" sur un groupe de 20 personnes. On code chaque modalité de la manière suivante : M=marron, V=vert, N=noir, B=bleu. On obtient la série statistique suivante :
M, V, M, M, M, N, M, B, M, B.

1.3.1 Tableaux de distribution de fréquences absolues, relatives et cumulées

Exemple : Pour l'exemple précédent, on remplit le tableau suivant :

Couleur des yeux	M	V	N	B	Total
Effectif					
Proportion					

Tableau-type : On choisit une notation pour la variable, par exemple : X. n désigne le nombre d'individus dans l'échantillon. On note C_1, ..., C_k les k modalités de la variable. Pour $1 \leq j \leq k$, on note

- n_j l'effectif associé à la modalité C_j (le nombre d'individus pour lesquels la valeur prise par la variable est C_j),
- $f_j = n_j/n$ la fréquence relative ou proportion associée à cette modalité,

– et si la variable est qualitative **ordinale** : $\Phi_j = f_1 + f_2 + \cdots + f_j$ la fréquence relative cumulée pour cette modalité (avec la convention : $\Phi_0 = 0$). Elle n'a de sens que si la variable est qualitative ordinale et si les modalités C_1, ..., C_k sont ordonnées suivant l'ordre croissant naturel (ou hiérarchique ascendant) qui règne parmi ces modalités.

Le tableau suivant est un tableau-type qui permet de résumer les données.

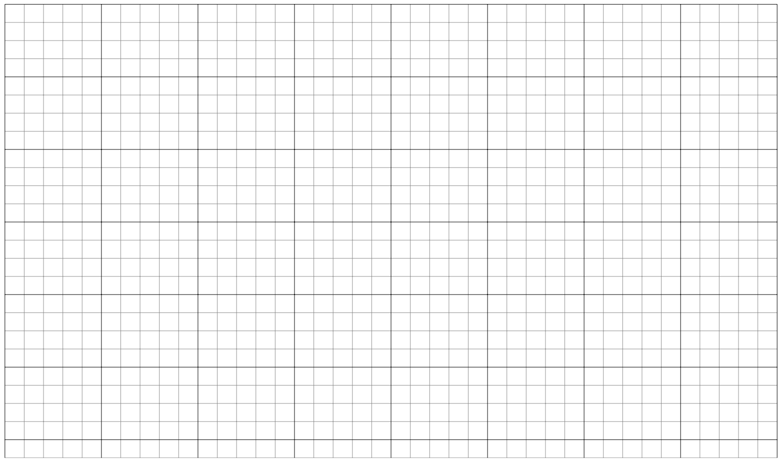
Variable X	C_1	C_2	...	C_k	Total
Fréquence absolue ou effectif	n_1	n_2	...	n_k	n
Fréquence relative ou proportion	$f_1 = n_1/n$	$f_2 = n_2/n$	...	$f_k = n_k/n$	1
Fréquence relative cumulée*	$\Phi_1 = f_1$	$\Phi_2 = f_1 + f_2$	...	$\Phi_k = f_1 + f_2 + \cdots + f_k = 1$	pas de sens

$*$ Attention : uniquement dans le cas de variables qualitatives ordinales.

1.3.2 Représentation graphique : diagrammes en secteurs et diagrammes en tuyaux d'orgue

1. **Diagramme en secteurs** : chaque modalité est représentée par un secteur d'un disque dont l'angle est proportionnel à la fréquence de la modalité (ou au pourcentage), l'angle 360 degrés équivalant à la fréquence relative 1 (ou au pourcentage 100%).
   Exemple :

2. **Diagramme en tuyaux d'orgue** : en abscisse sont disposées les différentes modalités auxquelles on associe des rectangles espacés entre eux, de largeur constante, dont les hauteurs (en ordonnée) sont proportionnelles à l'effectif ou à la fréquence relative de chaque modalité. Préciser le nom des axes, le nom du graphique et la source des informations. Dans le cas d'une variable qualitative ordinale, on peut également construire le diagramme en tuyaux d'orgue des effectifs ou des proportions cumulés.
   Exemple :



Remarque : cette représentation graphique est plus adaptée dans le cas d'une variable qualitative ordinale car elle rend compte de la structure d'ordre entre les modalités, disposées de gauche à droite par ordre croissant. C'est impossible de suggérer une structure d'ordre dans un diagramme en secteurs.

1.4 Les variables quantitatives discrètes

Exemple : pour 20 individus, on a relevé le nombre de fois où chacun a assisté à une séance de cinéma durant le mois d'août 2010. Pour simplifier, on nomme « ciné »la variable « nombre de séances de cinéma pendant le mois d'août ». La variable « ciné »sera notée C. La série statistique est résumée sous la forme du tableau suivant :

C	0	1	2	3	4
Effectif	4	6	7	2	1

1.4.1 Tableaux de distribution de fréquences

Exemple : pour la variable C, on remplit le tableau suivant :

C	0	1	2	3	4
Effectif					
Proportion ou fréquence relative					
Proportion cumulée ou fréquence relative cumulée					

On note $v_1, \ldots, v_k$ les k valeurs différentes que peut prendre la variable (remarque : on n'en rencontrera pas d'exemple dans ce cours, mais une variable discrète peut prendre une infinité de valeurs). Pour $1 \leq j \leq n$, on note n_j l'effectif des individus pour lesquels la variable prend la valeur v_j. On note f_j la fréquence relative ou proportion pour la valeur v_j et $\Phi_j = f_1 + \cdots + f_j$ la j-ème fréquence relative cumulée (avec la convention : $\Phi_0 = 0$). On résume habituellement les données comme dans le tableau-type suivant :

Valeurs prises par la variable	v_1	v_2	$\ldots$	v_k	Total
Fréquence absolue	n_1	n_2	$\ldots$	n_k	n
Fréquence relative	$f_1 = n_1/n$	$f_2 = n_2/n$	$\ldots$	$f_k = n_k/n$	1
Fréquence relative cumulée	$\Phi_1 = f_1$	$\Phi_2 = f_1 + f_2$	$\ldots$	$\Phi_k = f_1 + f_2 + \cdots + f_k = 1$	pas de sens

1.4.2 Représentation graphique : diagramme en bâtons

On trace un graphique avec

- sur l'axe des abscisses les différentes valeurs prises par la variable, placées **en respectant une échelle**,
- en ordonnée les fréquences relatives ou les fréquences absolues.
- Pour chaque valeur v_j on construit un bâton vertical à l'abscisse v_j, de hauteur proportionnelle à la fréquence de la valeur v_j.

Exemple : ciné.

1.4.3 Autre représentation graphique : fonction de répartition empirique

La fonction de répartition empirique permet de décrire la série statistique de manière complète. Elle est définie sur $\mathbb{R}$ et prend ses valeurs dans $[0, 1]$. Pour x dans $\mathbb{R}$, elle est définie par :

$$F(x) = \begin{cases} 0 & \text{si } x < v_1 \\ \Phi_j & \text{si } v_j \leq x < v_{j+1} \\ 1 & \text{si } v_k \leq x. \end{cases}$$

Exemple : ciné.

1.5 Les variables quantitatives continues

Exemple : on s'intéresse à la taille, notée T et exprimée en mètres, de 20 individus. On a obtenu la série statistique suivante :
1,72 ; 1,87 ; 1,66 ; 1,73 ; 1,64 ; 1,77 ; 1,80 ; 1,81 ; 1,60 ; 1,78 ; 1,83 ; 1,75 ; 1,70 ; 1,58 ; 1, 68 ; 1,66 ; 1,93 ; 1,75 ; 1,80 ; 1,85.

1.5.1 Tableaux de distribution de fréquences - fréquences cumulées

Les données brutes de la variable pour chaque individu sont notées x_1, ..., x_n. Elles peuvent prendre n'importe quelle valeur dans un intervalle de $\mathbb{R}$ et il est très rare d'avoir deux fois la même valeur pour deux individus différents. Il serait donc inutile de tracer un diagramme en bâtons comme dans le cas d'une variable discrète : il consisterait en un amoncellement illisible de bâtons de hauteur $1/n$. On choisit donc de faire un **regroupement en classes**.

Regroupement en classes :

- L'intervalle où la variable prend ses valeurs est divisé en k classes : $[b_0, b_1[$, $[b_1, b_2[$, ..., $[b_{k-1}, b_k[$ (il est possible d'avoir des bornes infinies).
- Pour $1 \leq j \leq k$, on note n_j l'effectif associé à la classe $[b_{j-1}, b_j[$, $f_j = n_j/n$ la fréquence relative associée à cette classe et $\Phi_j = f_1 + \cdots + f_j$ la j-ième fréquence cumulée (avec la convention : $\Phi_0 = 0$).
- On note $a_j = b_j - b_{j-1}$ l'amplitude de la classe $[b_{j-1}, b_j[$.
- On note $d_j = f_j/a_j$ la densité de proportion pour la classe $[b_{j-1}, b_j[$.

Exemple de la taille :

T	[1,50 ; 1,65[	[1,65 ; 1,75[	[1,75 ; 1,85[	[1,85 ; 2,00[
Effectif	3	6	8	3
Proportion				
Proportion cumulée				
Amplitude				
Densité de proportion				

Remarques :

- la densité de proportion permet de comparer les effectifs dans chaque classe en tenant compte de la taille de ces classes (cf. la notion de densité de population en géographie).
- Dans le cas de classes qui ont toutes la même longueur, il n'est pas nécessaire de calculer la densité de proportion, il est suffisant d'étudier les fréquences relatives ou absolues (qui sont directement proportionnelles à la densité de proportion).

Tableau-type :

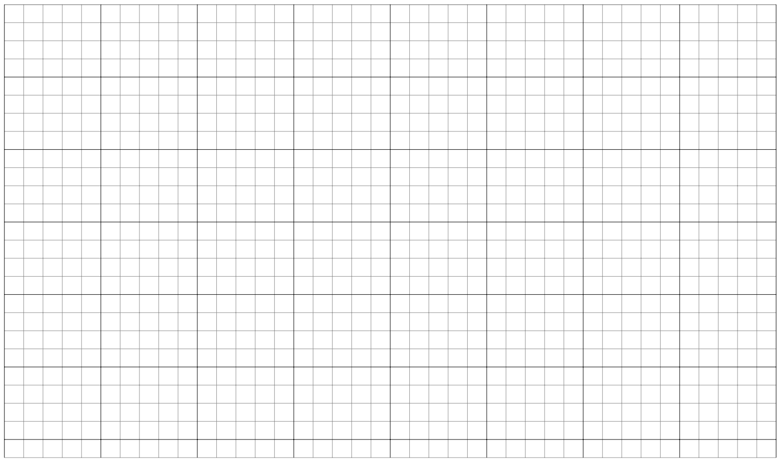
Variable X	$[b_0, b_1[$	$[b_1, b_2[$	...	$[b_{k-1}, b_k[$	Total
Fréq. absolue	n_1	n_2	...	n_k	n
Fréq. relative	$f_1 = n_1/n$	$f_2 = n_2/n$	...	$f_k = n_k/n$	1
Fréq. relative cumulée	$\Phi_1 = f_1$	$\Phi_2 = f_1 + f_2$	...	$\Phi_k = 1$	
Amplitude	$a_1 = b_1 - b_0$	$a_2 = b_2 - b_1$	...	$a_k = b_k - b_{k-1}$	
Densité de proportion	$d_1 = f_1/a_1$	$d_2 = f_2/a_2$	...	$d_k = f_k/a_k$	

Remarque : Ce tableau contient-il toute l'information apportée par les données brutes ou bien représente-t-il une perte d'information ? Quel est l'intérêt d'un tel tableau ?

1.5.2 Représentation graphique : histogramme et fonction de répartition empirique

Sur l'axe des abscisses sont placées les bornes des classes en respectant une échelle. Pour chaque classe, on élève un rectangle de hauteur proportionnelle à la densité de proportion.

Exemple de la taille T :



Remarque : on représente la **densité de proportion** et non pas les fréquences relatives ou absolues.
Conséquence : l'aire d'un rectangle est proportionnelle à la fréquence (relative ou absolue) de la classe correspondante. En effet, pour le rectangle correspondant à la classe $[b_{j-1}, b_j[$, l'aire est

$$(b_j - b_{j-1}) \times d_j = f_j.$$

Approximation de proportions : pour x une valeur dans l'intervalle $[b_{j-1}, b_j[$, on approche la proportion d'individus pour lesquels la variable est inférieure ou égale à x par l'aire de l'histogramme entre les abscisses b_0 et x, notée $F(x)$:

$$F(x) = f_1 + f_2 + \cdots + f_{j-1} + (x - b_{j-1}) \times d_j = \Phi_{j-1} + (x - b_{j-1}) \times d_j.$$

On a ainsi défini une fonction Φ qui vaut 0 sur $]-\infty, b_0[$, et 1 sur $[b_k, +\infty[$. Elle vaut Φ_j en b_j. Sur $[b_{j-1}, b_j[$, c'est une fonction affine de pente d_j. Cette fonction, affine par morceaux, est appelée **fonction de répartition empirique**.

Fonction de répartition empirique de la variable T :