Statistiques

I. Présentation

Statistiques mathématiques

La partie authentiquement mathématique des statistiques consiste pour l'essentiel en des probabilités. On essaye parfois de deviner une loi cachée ou des corrélations. On parle alors de **statistiques inférentielles**. Plus généralement, la partie probabiliste des statistiques, s'appelle aujourd'hui **statistiques mathématiques**.

Statistiques descriptives

Mais les statistiques, à l'origine, ne sont pas une discipline mathématique. Aujourd'hui, la partie des statistiques qui n'est pas proprement mathématique (puisqu'elle ne contient ni théorèmes ni démonstrations, mais seulement des formules, des définitions et de zolis dessins), s'appelle les **statistiques descriptives**.

II. Vocabulaire

Population

Une **population** statistique est un *ensemble* dont les éléments sont appelés **individus**. (Des hommes, ou des huîtres, ou des chaises…)
Une **classe** est un regroupement d'*individus* (c'est un sous-ensemble de la *population*). Le nombre d'individus dans une classe s'appelle l'**effectif** de cette *classe*.

Fréquence

La **fréquence** d'une classe est le rapport de l'effectif de cette classe sur l'effectif total. (Un *rapport* est un *quotient* de deux grandeurs de même nature.) C'est un nombre compris entre 0 et 1 qu'on peut donner sous forme de pourcentage.

Caractère

Pour chaque individu, on considère un **caractère**, qui est une propriété particulière. La fonction qui, à chaque individu, associe son caractère s'appelle une **variable statistique**. Le caractère peut être **qualitatif** (la couleur des yeux) ou **quantitatif** (l'âge). S'il est quantitatif, il peut être **discret** ou **continu**. Discret : il ne peut prendre que certaines valeurs bien séparées les unes des autres. Continu : il peut prendre toutes les valeurs réelles sur un certain intervalle. Exemple de caractère *discret* : le nombre d'enfants. Exemple de caractère *continu* : la taille. Un caractère continu peut être regroupé par intervalles. Par exemple, pour relever le salaire mensuel, on pourra grouper par tranches de 100 euros. On regroupera tous les individus dont le salaire mensuel est compris entre 1500 euros et 1600 euros.

L'**amplitude** d'une classe ou d'un intervalle, c'est l'écart entre la plus petite valeur et la plus grande. Ici, les intervalles ont des amplitudes de 100 euros. L'amplitude de la population totale, s'appelle son **étendue**.

Série statistique

Une **série statistique** est la liste des valeurs associées aux individus.

Par exemple, dans une classe de 25 élèves, on s'intéresse aux notes obtenues par les élèves à un contrôle donné. Ces notes sont :

16 ; 9 ; 9 ; 16 ; 10 ; 12 ; 18 ; 9 ; 15 ; 15 ; 13 ; 9 ; 6 ; 12 ; 15 ; 6 ; 4 ; 13 ; 9 ; 8 9 ; 9 ;10 ; 8 ; 14

On ne perdrait pas d'information si l'on donnait la série sous la forme suivante, appelée **tableau d'effectifs** :

Valeur	Effectif
4	1
6	2
8	2
9	7
10	2
12	2
13	2
14	1
15	3
16	2
18	1
Total :	25

À chaque valeur (ou à chaque intervalle), on peut ainsi associer une classe, formée de l'ensemble des individus dont le caractère considéré prend cette valeur. La *fréquence* d'une valeur donnée est la fréquence de la classe associée à cette valeur. Ainsi, en reprenant le tableau qui précède, on peut constater que la note 9 est assez fréquente. Sa fréquence est de $\frac{7}{25} = 0{,}28 = \underline{28\%}$.

III. STATISTIQUES DÉSCRIPTIVES

Dans ce paragraphe, nous prendrons encore comme exemple des notes obtenues par des élèves à un contrôle. Mais, pour simplifier les expressions numériques, on imaginera un groupe de seulement treize élèves. Les notes sont les suivantes :

7 ; **11** ; **14** ; **11** ; **12** ; **15** ; **4** ; **14** ; **11** ; **11** ; **15** ; **14** ; **14**.

MOYENNE

On obtient la moyenne d'une série statistique en additionnant toutes les valeurs et en divisant par le nombre de valeurs.

Ici :

$$\frac{7+11+14+11+12+15+4+14+11+11+15+14+14}{10}$$

Si une valeur apparaît plusieurs fois, il faut la compter autant de fois qu'elle apparaît. Ainsi, si la série statistique est présentée en *tableau d'effectifs* :

Valeur	Effectif
4	1
7	1
11	4
12	1
14	4
15	2

Alors, il faut tenir compte des effectifs pour exprimer la moyenne :

$$\frac{1\times 4+1\times 7+4\times 11+1\times 12+4\times 14+2\times 15}{1+1+4+1+4+2}$$

Les effectifs sont, au numérateur, des *coefficients*.

MÉDIANE

La médiane est « la » valeur qui sépare la série en deux classes de même effectif : il doit y avoir autant d'individus au dessus qu'en dessous.

Soit n l'effectif total d'une série statistique.
Si n est pair, la médiane est la valeur de rang $\frac{1+n}{2}$.
Si n est impair, nous prenons pour médiane la moyenne des deux valeurs centrales, c'est-à-dire de celles dont les rangs encadrent le nombre (pas entier) $\frac{1+n}{2}$.

QUARTILES

Ce sont les trois valeurs qui partagent la série en quatre classes de même effectif. Notons-les, par ordre croissant, Q_1, Q_2 et Q_3.
Le quartile du milieu, Q_2 est la médiane. Les deux autres ont des définitions qui ressemblent à celle de la médiane, mais simplifiées :

Le quartile Q_1 est la plus petite valeur telle que $\frac{1}{4}$ des données lui soient inférieures ou égales. Le quartile Q_3 est la plus petite valeur telle que $\frac{3}{4}$ des données lui soient inférieures ou égales.

Par exemple, en reprenant notre groupe de 13 élèves :

rang	1	2	3	4	5	6	7	8	9	10	11	12	13
valeur	**4**	**7**	**11**	**11**	**11**	**11**	**12**	**14**	**14**	**14**	**14**	**15**	**15**

$\frac{13}{4} = 3{,}25$. Donc Q_1 est la valeur du terme de rang 4. Q_1 = **11**.

Q_2 est la médiane. Q_2 = **12**.

$\frac{3\times13}{4} = 9{,}75$. Donc Q_3 est la valeur du terme de rang 10. Q_3 = **14**.

EFFECTIFS CUMULÉS, FRÉQUENCES CUMULÉES

L'effectif cumulé d'une valeur donnée, c'est le nombre d'individus dont la valeur est inférieure ou égale à cette valeur donnée. On calcul cela très facilement à l'aide des effectifs. Voici un tableau fictif donnant le nombre d'enfants par femme d'un groupe (non représentatif) de 205 mères :

Nombre d'enfants	Effectif	Effectif cumulé
1	25	25
2	75	100
3	54	**154**
4	26	180
5	18	198
6	4	202
7	2	204
8	1	205

← **154** = 25+75+54

De la même façon, les **fréquences cumulées** s'obtiennent en additionnant les fréquences correspondant aux valeurs inférieures ou égales à la valeur considérée :

Nombre d'enfants	Effectif	Effectif cumulé	Fréquences	Fréquences cumulées
1	25	25	**0,12**	0,12
2	75	100	**0,37**	0,49
3	54	154	**0,26**	**0,75**
4	26	180	0,13	0,88
5	18	198	0,09	0,97
6	4	202	0,02	0,99
7	2	204	0,01	1
8	1	205	0	1

0,75 = **0,12**+**0,37**+**0,26**

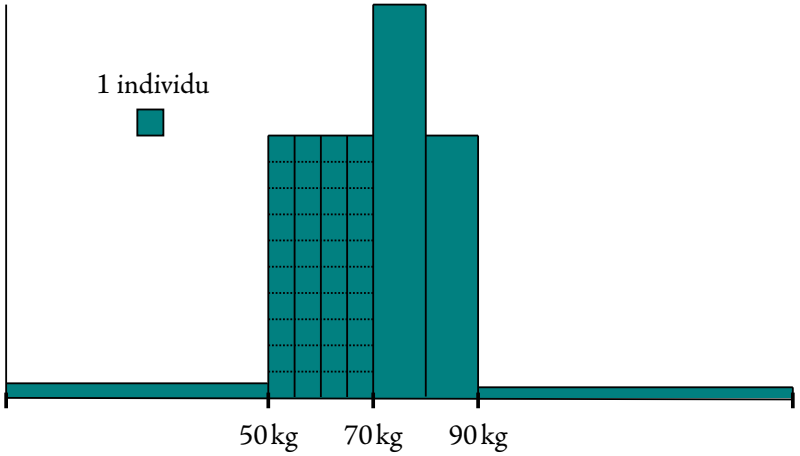
Représentations graphiques

Nous n'allons pas passer en revue toutes les formes de représentations graphiques. Contentons-nous de parler des *histogrammes*.

L'**histogramme** est particulièrement utile lorsqu'on représente un *caractère continu*, regroupé par classes d'*amplitudes* inégales. Supposons par exemple qu'on cherche à représenter graphiquement la façon dont se répartissent les poids d'un groupe de 100 personnes et que les données soient regroupées ainsi par classes d'amplitudes inégales :

Poids en Kg	Effectif
[0 ; 50[	5
[50 ; 55[	10
[55 ; 60[	10
[60 ; 65[	10
[65 ; 70[	10
[70 ; 80[	30
[80 ; 90[	20
[90 ; 150]	5

Le principe de l'**histogramme** est de représenter l'effectif non pas par la hauteur des rectangles (en ordonnée), mais par leur aire. La largeur des rectangles étant, elle, proportionnelle à l'amplitude de l'intervalle.



III. Statistiques inférentielles

Intervalle de fluctuation

Définition

On répète n fois une même expérience aléatoire à deux issues A et B, dans laquelle la probabilité d'obtenir l'issue A est toujours la même. Notons p cette probabilité. On note alors f la fréquence de l'issue A dans les n tirages. On l'appelle la **fréquence observée**. On appelle **intervalle de fluctuation au seuil de** 95% tout intervalle (inclus dans [0 ;1]) dans lequel f a au moins 95% de chances de se trouver.

Formule

Lorsque $0{,}2 \leq p \leq 0{,}8$ et que $n \geq 25$, on obtient (approximativement) un intervalle de fluctuation au seuil de 95% par la formule : $\left[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}} \right]$.

Résumé

Lorsqu'on répète suffisamment une même expérience aléatoire à deux issues A et B, la fréquence f réellement observée de l'événement A se rapproche de sa probabilité p. Les formules sur les intervalles de fluctuation nous renseignent sur la rapidité de ce rapprochement.