Tables des matières

Chapitre 1

Introduction et Vocabulaire.

Commençons par établir un peu de vocabulaire.

Définitions 1.0.1.

1. Une population statistique est l'ensemble sur lequel on effectue des observations.
2. Les individus sont les éléments de la population statistique étudiée.
3. Echantillon: Une partie de la population.
4. Caractères (variables) statistique: C'est ce qui est observé ou mesuré sur les individus d'une population statistique.
5. Modalité: valeur prise par le caractère pour un individu de la population.
6. Un caractère statistique est quantitatif si ces valeurs sont des nombres exprimant une quantité, sur lesquels les opérations aritghmetiques (sommes,etc...) ont un sens. Parmi les caractères quantitatifs ont discernent
   - les variables disctrètes: prenant des valeurs isolées (entières par exemple)
   - les variables continues: peut prendre toutes les valeurs d'un intervalle.
7. Un caractère est qualitatif si ces modalités sont seulement repérables (couleur, forme,marque,...)

Exemples 1.0.2.

1. Taille des élèves d'une classe.
2. Couleur des voitures d'une marque donnée.
3. Nombre d'habitants de chaque département français.
4.

1. Effectifs Nombre d'individus prenant une valeur d'un caractère.
2. classe
3. effectif total (ou taille): C'est le nombre d'individus de la population.
4. fréquences (d'une valeur ou d'une classe): C'est le quotient de l'effectif (d'une valeur ou d'une classe) par l'effectif total.

5. fréquences cumulées.
6. Différentes représentations d'une série statistique à une variable.

Chapitre 2

Les paramètres de position.

BUT : Synthétiser l'information contenue dans un tableau par un graphique est la première étape réalisée en statistique. Par la suite, on cherche à synthétiser encore plus l'information en la réduisant à une seule valeur numérique. Les caractéristiques de tendance centrale essayent de donner la valeur la plus représentative d'un ensemble de valeurs numériques.

Remarque: A l'exception du mode, les paramètres définis par la suite n'ont de sens que pour les caractères quantitatifs.

2.1 Mode

C'est la valeur observée d'effectif maximum.

Variable discrète : Il est fortement conseillé d'utiliser le diagramme en bâtons pour déterminer le mode. En effet, deux valeurs consécutives x_i, x_{i+1} peuvent avoir le même effectif maximum; on parlera d'intervalle modal $[x_i, x_{i+1}]$. Il peut aussi y avoir un mélange de deux populations qui conduit à un diagramme en bâtons où apparaissent deux bosses; on considérera deux modes. Il est déconseillé, sauf raison explicite, d'envisager plus de deux modes.

Exemple 2.1.1. Considérons les notes obtenues par 30 étudiants lors d'un examen.

Notes	2	3	4	5	6	7	8	9	10
Effectifs	1	0	2	3	8	9	3	2	2

Le mode est 7.

Variable classée : la classe modale correspond à la classe ayant l'effectif maximum par unité d'amplitude (il faut donc considéré les effectifs corrigés ou les fréquences corrigées dans le cas de classes d'amplitudes inégales). Il est fortement conseillé d'utiliser l'histogramme pour déterminer le mode. Comme pour le cas discret, on peut avoir deux classes modales. Toutes les valeurs de la classe pouvant à priori se réaliser, on ne se contentera pas

de déterminer la classe modale. Une des valeurs de cette classe sera le mode. On préconise parfois, simplicité de prendre le centre de la classe modale. Il est cependant préférable de tenir compte des classes adjacentes de la manière suivante: Si la classe modale est déterminée par l'intervalle $[x_i, x_{i+1}]$ h_i étant la différence des effectifs entre la classe modale et la classe qui la précède et h_{i+1} étant la différence d'effectif entre la classe modale et la classe suivante on a

$$M_0 = \frac{x_i h_{i+1} + x_{i+1} h_i}{h_i + h_{i+1}}.$$

Exemple 2.1.2. Soit le caractère X de distribution statistique

Classes	$[0, 10[$	$[10, 20[$	$[20, 25[$	$[25, 30[$	$[30, 40[]$
Effectifs	15	30	25	15	15
fréquences	$0,15$	$0,30$	$0,25$	$0,15$	$0,15$
Amplitudes	2	2	1	1	2
Fréquences corrigées	$0,075$	$0,15$	$0,25$	$0,15$	$0,075$

La classe $[10, 20[$ a une fréquence de 30 % alors que la classe $[20, 25[$ a une fréquence de 25%. Pourtant c'est cette dernière qui est la classe modale. On prendra pour mode la valeur $M_0 = \frac{20(25-15)+25(25-15)}{10+10} = \frac{450}{20} = 22,5$. On constate que, dans ce cas le mode est exactement la valeur centrale de la classe modale...Il n'en est pas toujours ainsi !

Remarque 2.1.3. La valeur obtenue reste une approximation. Cette valeur est intiment liée au choix des limites de classe. Deux répartitions en classes différentes forniront en général des estimations différentes.

2.2 Médiane

Les valeurs étant rangées par ordre croissant, c'est la valeur de la variable qui sépare les observations en deux groupes d'effectifs égaux.

Variable discrète: la détermination peut s'obtenir à partir du tableau statistique en recherchant la valeur de la variable correspondant à un effectif cumulé égale à $n/2$ (n étant l'effectif global) ou à une fréquence cumulée égale à $1/2$ (fréquence cumulée). Il est encore plus facile de lire sur les graphiques cumulatifs les abscisses des points d'ordonnée n/2 (effectif cumulé) ou $1/2$ (fréquence cumulée). Si tout un intervalle a pour image n/2 ($1/2$ pour la fréquence), on parlera d'intervalle médian (on peut prendre le milieu de l'intervalle comme médiane). On remarque que si le nombre d'observations

n est impair alors la médiane est la valeur de la série *ordonnée* située à la position $(n+1)/2$. Si, par contre, n est impair la médiane sera le centre de l'intervalle médian lui-même déterminé par les deux valeurs centrales situées aux positions $n/2$ et $(n+1)/2$.

Exemple 2.2.1. Considérons le nombre de chatons par portée :

n° de la chatte	1	2	3	4	5	6	7	8	9	10	11	12
Nombre de chatons	5	3	2	3	6	3	5	4	7	2	1	4

La médiane se situe entre la sixième et la septième valeur du tableau ordonné. La médiane est donc la moyenne de ces deux valeurs $M_e = (3+4)/2 = 3,5$. Cette valeur n'est pas (bien sur !) une valeur observée.

Variable classée: l'abscisse du point d'ordonnée n/2 (1/2 pour la fréquence)se situe en général à l'intérieur d'une classe. Pour obtenir une valeur plus précise de la médiane, on procède à une interpolation linéaire. La valeur de la médiane peut être lue sur le graphique ou calculée analytiquement.

De manière générale, si a et b sont les bornes de la classe contenant la médiane, $F(a)$ et $F(b)$ les valeurs de la fréquence cumulée croissante en a et b, alors

$$M_e = a + (b-a)\frac{0,5 - F(a)}{F(b) - F(a)}.$$

Exemple 2.2.2. Une entreprise réalise une enquête sur le montant de 125 factures impayées.

Classes(en euros)	$[0,100[$	$[100,200[$	$[200,300[$	$[300,500[$	$[500,1000[$
Effectifs	8	27	36	35	19
Fréquences	$0,064$	$0,216$	$0,288$	$0,28$	$0,152$
Fréqu. cumu. croissantes	$0,064$	$0,280$	$0,568$	$0,848$	1

La classe médiane est $[200,300[$. On a $M_e = 200 + \frac{(0,5-0,280)}{(0,568-0,280)}(300-200) = 276,39$

2.3 Moyennes

a)**Moyenne arithmétique**

Si x_i sont les observations d'un caractère discret ou les centres de classe d'une variable classée, la moyenne arithmétique $\overline{x}$ est égale à

$$\sum_{i=1}^{k} \frac{n_i x_i}{n} = \sum_{i=1}^{k} f_i x_i$$

où n_i et f_i désigne respectivement les effectifs et les fréquences de la modalité x_i (ou le centre de la classe dans le cas d'une variable classée)

La moyenne arithmétique est un paramètre de tendance centrale plus utilisé que les autres de par ses propriétés algébriques:

Proposition 2.3.1.

1. *Soient plusieurs populations d'effectifs $n_1, n_2, \ldots, n_k$, de moyennes respectives $\overline{x_1}, \ldots, \overline{x_k}$. La moyenne globale est la moyenne des moyennes i.e.*

$$\overline{\overline{x}} = \frac{\sum_{i=1}^{k} n_i \overline{x_i}}{n}.$$

2. *La moyenne arithmétique conserve les changements d'échelle et d'origine i.e. pour $a, b \in \mathbb{R}$, $\overline{ax+b} = a\overline{x} + b$.*
3. *La somme algébrique des écarts de tous les termes de la série à la moyenne est nulle: i.e. $\sum_{i=1}^{k} n_i(x_i - \overline{x}) = 0$.*
4. *La somme des carrés des différences de tous les termes d'une série statistique à un nombre quelconque est minimal lorsque ce nombre es la moyenne arithmétique de cette série: i.e. $\sum_{i=1}^{k} n_i(x_i - x)^2$ est minimale si et seulement si $x = \overline{x}$.*

b)**Moyenne géométrique**

Si x_i sont les observations d'une variable quantitative, la moyenne géométrique $G(x)$ est égale à $(x_1^{n_1} \ldots x_k^{n_k})^{1/n}$ où $n = \sum_{i=1}^{k} n_i$.

Ce type de moyenne est surtout utilisé pour calculer des pourcentages moyens. r étant un taux d'accroissement, $1 + r$ est appelé coefficient multiplicateur; et le coefficient multiplicateur moyen est alors égal à la moyenne géométrique des coefficients multiplicateurs.

Exemple 2.3.2. Dans un certain pays l'inflation mensuel des six premier mois de l'année est donnée par les valeurs 1% en janvier, $1,8\%$ en février, $1,7\%$ en mars, $0,2\%$ en avril, $0,4\%$ en mai et $0,9\%$ en juin. Si un article vaut 1000 euros en début janvier il vaudra donc $1000 \times 1,01 = 1010$ euros en fin janvier... En fin juin il vaudra donc $1000 \times 1,01 \times 1,018 \times 1,017 \times 1,002 \times 1,004 \times 1,009 = 1.061,41$ euros. Le taux moyen d'inflation durant ce premier semestre est le nombre t tel que $1000 \times (1+t)^6 = 1.061,41$. On a donc $1 + t = (\frac{1061,41}{1000})^{1/6} = 1,00998$.

le taux d'inflation moyen est donc la moyenne géométrique des taux d'inflation.

c)**Moyenne harmonique**

Si x_i sont les observations d'une variable quantitative, la moyenne harmonique est égale à

$$H(x) = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} n_i/x_i}$$

On peut donc énoncer : l'inverse de la moyenne harmonique est égal à la moyenne arithmétique des inverses.

La moyenne harmonique est utilisée lorsqu'on demande une moyenne de valeurs se présentant sous forme de quotient de deux variables x/y (km/h, km/litre,...). Attention, il faut cependant bien étudier le problème car il peut aussi s'agir d'une moyenne arithmétique.

d)**Moyenne quadratique**

Si x_i sont les observations d'une variable quantitative, la moyenne quadratique est égale à

$$Q(x) = (n_1 x_1^2 + n_2 x_2^2 + \cdots + n_k x_k^2)^{1/2}$$

Remarque 2.3.3. Pour une série statistique x, on a $H(x) \leq G(x) \leq \overline{x} \leq Q(x)$.

2.4 Quantiles

Ce sont des caractéristiques de position.

Il y a 1a médiane M_e qui sépare les observations en 2 groupes d'effectifs égaux 3 quartiles $Q1, Q2, Q3$ qui séparent les observations en 4 groupes d'effectifs égaux 9 déciles $D1, D2, ..., D9$ qui séparent les observations en 10 groupes d'effectifs égaux 99 centiles $C1, C2, ..., C99$ qui séparent les observations en 100 groupes d'effectifs égaux

La détermination de ces caractéristiques est identique à celle de la médiane.

Les quartiles sont obtenus lorsqu'on a cumulé $25, 50, 75$ pourcent de la population. Les déciles sont obtenus lorsqu'on a cumulé $10, 20, ..., 90$ pourcent de la population Les centiles sont obtenus lorsqu'on a cumulé $1, 2, ..., 99$ pourcent de la population Remarque: la notion de déciles et de centiles n'a de sens que s'il y a beaucoup d'observations et donc essentiellement pour une variable classée.

Exemple 2.4.1. On reprend l'exemple,2.2.2 : le tableau statistique montre que le premier quartile set compris entre 100 et 200 car on a F(100)= 0,064 <0,25<0,280=F(200) et donc en effectuant une interpolation linéaire on a :

$$Q_1 = 100 + \frac{0,25 - 0,064}{0,280 - 0,064}(200 - 100) = 186,11.$$

Le tableau montre que le troisième quartile est compris entre 300 et 500 et une interpolation linéaire donne $Q_3 300 + \frac{0,75-0,568}{0,848-0,568}(500 - 300) = 430$.

Chapitre 3

Paramètres de dispersion

Comme leur nom l'indique, ces caractéristiques essayent de synthétiser par une seule valeur numérique la dispersion de toutes les valeurs observées.

On considère un caractère statistique X dont les k observations distinctes (ou les centres de classe dans le cas d'une variable classée) notées $x_1, \ldots, x_k$ ont des effectifs respectifs $n_1, \ldots, n_k$ on note $n := \sum_{i=1}^{k} n_i$.

3.1 Étendue, écarts interquartiles, boîtes à moustaches

L'étendue est la différence entre la plus grande et la plus petite observation.

l'écart interquartile est la différence entre le troisième et le premier quartile.

La bôîte à moustache : pour une série statistique donnée $(x_1, x_2, \ldots, x_n)$ un rectangle (=une boîte) sur base du premier et du troisième quartile, coupée en deux parties (généralement de longueur inégales) par la médiane. Cette boîte est ensuite prolongée à sa gauche et à sa droite par deux moustaches jusqu'aux valeurs minimale et maxiamle.

Remarque : On calcule parfois des valeurs pivots : $a_1 := x_{1/4} - 1,5(x_{3/4} - x_{1/4})$ et $a_2 := x_{3/4} + 1,5(x_{3/4} - x_{1/4})$. Elle sont situées de part et d'autre de la boîte et en sont distantes d'une fois et demie sa longueur. Leur raison d'être résulte d'une constatation : la plupart des séries qui ne contiennent pas de valeurs aberrantes, se situent dans l'intervalle $[a_1, a_2]$.

3.2 L'écart moyen absolu et l'écart médian absolu

L'écart moyen absolu, noté e_m, est la moyenne des valeurs absolues des différences entre les observatiuons et la moyenne $\overline{x}$:

$$e_m := \frac{\sum_{i=1}^{k} n_i|x_i - \overline{x}|}{n}.$$

L'écart médian absolu, noté e_M est la moyenne des valeurs absolues des dfférences entre les observations et la médiane M_e:

$$e_M := \frac{\sum_{i=1}^{k} n_i|x_i - M_e|}{n}.$$

3.3 Variance et écart type

Si x_i sont les observations d'une variable discrète X ou les centres de classe d'une variable classée, la variance $V(X)$ est égale à

$$V(X) := \sum_{i=1}^{k} \frac{n_i(x_i - \overline{X})^2}{n} = \sum_{i=1}^{k} f_i(x_i - \overline{X})^2.$$

On utilise plus couramment l'écart-type, noté σ, qui est la racine carrée de la variance :

$$\sigma(X) := V(X)^{1/2}.$$

L'écart type a l'avantage d'être un nombre de même dimension que les données (contrairement à la variance qui en est le carré).

La variance est un paramètre de dispersion plus utilisé que les autres de par ses propriétés algébriques:

Proposition 3.3.1. *Soit X un caractère statistique et $a, b \in \mathbb{R}$. On pose $Y = aX + b$. (i.e. les modalités de Y sont donnés par $ax_1 + b, \ldots, ax_k + b$ avec des effectifs correspondants $n_1, \ldots, n_k$.*

1. *$V(X) = \sum_{i=1}^{k} \frac{n_i x_i^2}{n} - \overline{X}^2$, (Formule de Koenig).*
2. *$V(Y) = a^2 V(X)$.*
3. *$\sigma(Y) = |a|\sigma(X)$.*

3.4 Coefficient de variation

Le coefficient de variation C_v d'un caractère statistique X est le nombre :

$$C_v(X) = \frac{\sigma(X)}{\overline{X}}.$$

C'est un coefficient qui permet de relativiser l'écart-type en fonction de la taille des valeurs. Il permet ainsi de comparer la dispersion de séries de mesures exprimées dans des unités différentes.