Chapitre 2

Résumés numériques d'une variable quantitative

Dans ce chapitre, X désigne une variable quantitative.

2.1 Paramètres de position

2.1.1 Le mode

Le mode rend compte de l'endroit où les données sont le plus concentrées.

Pour une variable **discrète**, le mode est la ou les valeurs de la variable qui correspond(ent) à l'**effectif maximal** (ou à la fréquence relative maximale).

Pour une variable **continue** regroupée en classes, le mode est la ou les classe(s) de **densité de proportion maximale**.

Exemples : ciné, taille.

2.1.2 La moyenne

On note $\{x_1 \ldots, x_n\}$ la série statistique. La moyenne est définie par :

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

Exemple : ciné, taille.

Cas d'une variable discrète : si v_1, ..., v_k sont les k valeurs prises par la variable X, n_j l'effectif et f_j la fréquence relative correspondant à la valeur v_j, on peut réécrire :

$$\bar{x} = \frac{n_1 v_1 + n_2 v_2 + \cdots + n_k v_k}{n} = \frac{1}{n} \sum_{i=1}^{n} n_j v_j = \sum_{i=1}^{n} f_j v_j.$$

Exemple : ciné.

Cas d'une variable continue regroupée en classes : la variable X est regroupée dans les classes $[b_{j-1}, b_j[$ $(1 \leq j \leq n)$, les fréquences relatives associées à ces classes sont notées f_j, $1 \leq j \leq n$. Lorsque les données brutes ne sont plus accessibles et qu'on ne dispose que des données regroupées en classes, on calcule une **moyenne approchée** grâce à des représentants des classes (leurs centres) : $c_j = (b_{j-1} + b_j)/2$, par la formule :

$$\bar{x}_{app} = f_1 c_1 + f_2 c_2 + \cdots + f_k c_k = \sum_{i=1}^{n} f_j c_j.$$

Exemple : calcul d'une moyenne approchée de la variable « taille » à partir du regroupement en classes.

Propriétés de la moyenne : si on fait le changement de variable $Y = aX + b$ (traduction sur les séries statistiques : $y_i = ax_i + b$, $1 \leq i \leq n$), alors

$$\bar{y} = a\bar{x} + b.$$

Exemple : calcul de la taille moyenne en centimètres.

2.1.3 La médiane

"En gros", le calcul de la médiane revient à ranger les observations par ordre croissant et trouver un point au-dessous duquel se situent 50 % des observations et au-dessus duquel se situent 50 % des observations.

a) Cas d'une variable discrète.

- Si n est **impair**, la médiane est la $\frac{n+1}{2}$-ième observation.
- Si n est **pair**, il y a plusieurs façons convenables de définir la médiane. Nous choisirons la suivante : la médiane est la plus petite valeur observée v_j telle que l'effectif cumulé en v_j dépasse $n/2$ (dépasse au sens large : est supérieure ou égale). Autrement dit, c'est la plus petite valeur v_j pour laquelle la proportion cumulée dépasse $1/2$. **Remarque** : cette définition est encore vraie pour n **impair**.

La détermination de la médiane se fait donc à l'aide des effectifs cumulés, des proportions cumulées ou de la fonction de répartition empirique (graphiquement).

Exemple : ciné.

b) Cas d'une variable continue. La médiane est définie comme la solution Q_2 de l'équation :

$$F(Q_2) = 0.5,$$

où F est la fonction de répartition empirique de la variable. On sait que cette solution existe parce que F est continue, et $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to+\infty} F(x) = 1$. Si de plus F est strictement croissante, la solution Q_2 est unique. La méthode pratique est la suivante :

1. S'il existe une borne de classe b_j telle que la proportion cumulée sur la classe $[b_{j-1}, b_j[$ est exactement 0.5, autrement dit : $F(b_j) = 0.5$, alors **la médiane est ce** b_j.
2. Sinon, alors il existe une classe $[b_{j-1}, b_j[$ telle que

   $$F(b_{j-1}) < 0.5 < F(b_j).$$

   Cette classe est la première sur laquelle la fréquence cumulée dépasse 0.5. Pour $x \in [b_{j-1}, b_j[$, $F(x) = \Phi_{j-1} + (x - b_{j-1}) \times d_j$. Mais en particulier :

   $$F(Q_2) = \Phi_{j-1} + (Q_2 - b_{j-1}) \times d_j = 0.5$$

   D'où

   $$Q_2 = \frac{0.5 - \Phi_{j-1}}{d_j} + b_{j-1}.$$

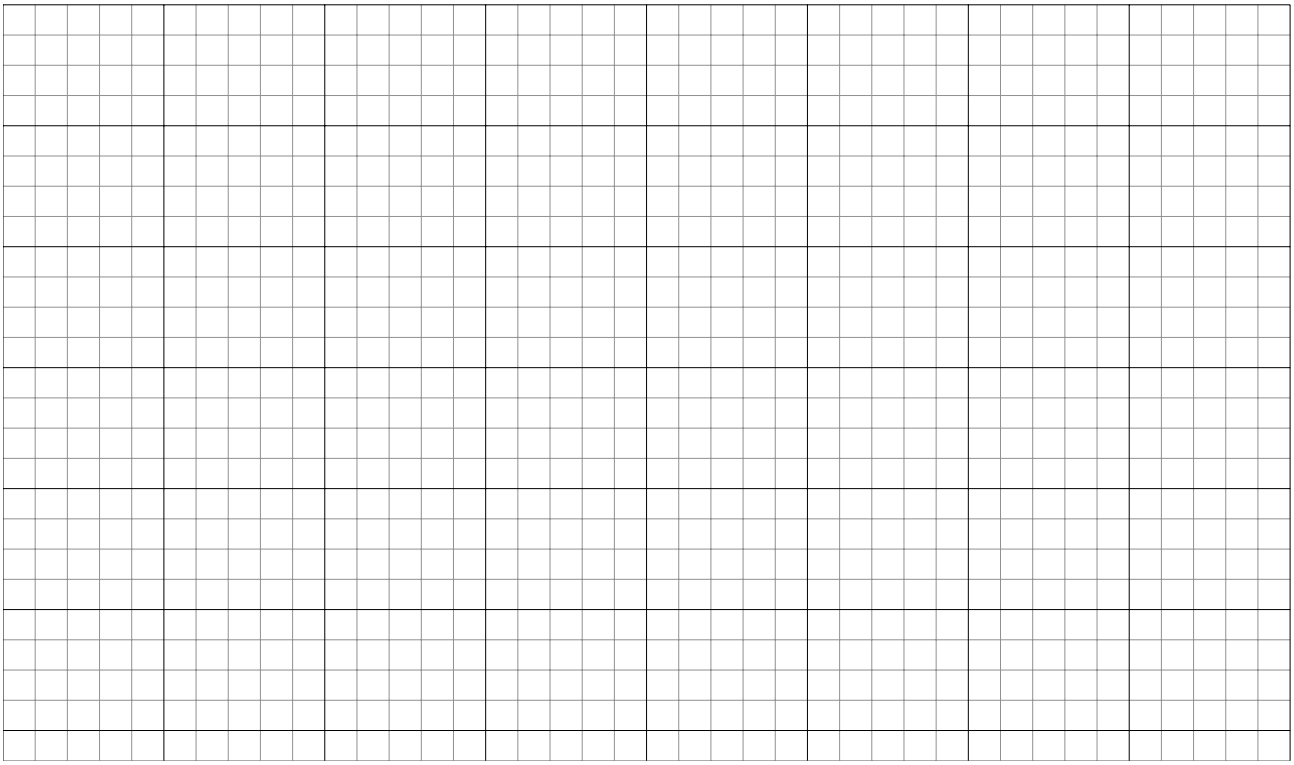
   Ou encore, en termes des b_j et de F :

   $$Q_2 = \frac{0.5 - F(b_{j-1})}{F(b_j) - F(b_{j-1})} \times (b_j - b_{j-1}) + b_{j-1}.$$

Cette méthode peut se traduire graphiquement en utilisant le graphe de la fonction de répartition empirique et le théorème de Thalès.

Exemple : médiane de la variable « taille », regroupée en classes.

Méthode graphique avec la fonction de répartition empirique :



2.1.4 Quantiles

a) Cas d'une variable continue

Soit X une variable quantitative continue, de fonction de répartition empirique F. On suppose qu'on dispose de la répartition en classes des observations.

Le **quantile d'ordre** p de X est la solution notée q_p de :

$$F(q_p) = p.$$

Cela signifie qu'une proportion d'environ p des observations est inférieure à q_p et qu'une proportion d'environ $1-p$ des données est supérieure à q_p.

Quantiles particuliers

- Quartiles : quantiles correspondant aux proportions multiples de 0.25 (un quart). On note Q_1 le premier quartile, qui correspond à $q_{0.25}$, Q_3 le troisième quartile, qui correspond à $q_{0.75}$. La médiane est le deuxième quartile $Q_2 = q_{0.5}$.
- Déciles : quantiles correspondant aux proportions multiples de 0.1 : $q_{0.1}$ (premier décile), $q_{0.2}$ (deuxième décile), etc.
- Percentiles ou centiles : quantiles correspondant aux proportions multiples de 0.01. Par exemple, le 65ème percentile est le quantile $q_{0.65}$.

Calcul du quantile q_p : même méthode que pour le calcul de la médiane.

1. S'il existe une borne de classe b_j telle que la proportion cumulée sur la classe $[b_{j-1}, b_j[$ est exactement p, autrement dit : $F(b_j) = p$, alors $q_p = b_j$.
2. Sinon, alors il existe une classe $[b_{j-1}, b_j[$ telle que

$$F(b_{j-1}) < p < F(b_j).$$

Cette classe est la première sur laquelle la fréquence cumulée dépasse p. Pour $x \in [b_{j-1}, b_j[$, $F(x) = \Phi_{j-1} + (x - b_{j-1}) \times d_j$. Mais en particulier :

$$F(q_p) = \Phi_{j-1} + (q_p - b_{j-1}) \times d_j = p$$

D'où

$$q_p = \frac{p - \Phi_{j-1}}{d_j} + b_{j-1}.$$

Ou encore, en termes des b_j et de F :

$$q_p = \frac{p - F(b_{j-1})}{F(b_j) - F(b_{j-1})} \times (b_j - b_{j-1}) + b_{j-1}.$$

Exemple : troisième quartile de la variable « taille ».

b) Cas d'une variable discrète

Comme pour la médiane, il existe diverses manières de définir les quantiles d'une loi discrète : comme la fonction de répartition empirique n'est pas continue mais a des paliers, elle ne prend pas toutes les valeurs entre 0 et 1. Pour une proportion p fixée, on cherche donc une valeur x telle que $F(x)$ s'approche, en un certain sens, de p. Nous choisissons la définition suivante :

$$q_p = \begin{cases} v_1 & \text{lorsque } 0 < p \leq \Phi_1 = f_1, \\ v_2 & \text{lorsque } \Phi_1 < p \leq \Phi_2, \\ \dots, & \\ v_j & \text{lorsque } \Phi_{j-1} < p \leq \Phi_j, \\ \dots, & \\ v_k & \text{lorsque } p = \Phi_k (= 1). \end{cases}$$

Exemple : troisième quartile de la variable « ciné ».

2.1.5 Utilisation des paramètres de tendance centrale

Robustesse
La médiane est plus **robuste** que la moyenne : une ou plusieurs données erronnées ne font pratiquement, voire pas du tout, changer la médiane, alors qu'elles peuvent affecter considérablement la moyenne.

Assymétrie
La comparaison de la médiane et de la moyenne permet de détecter des assymétries dans les données :

2.2 Paramètres de dispersion

2.2.1 L'étendue

Soit x_{min} la plus petite observation et x_{max} la plus grande. On définit **l'étendue** $e = x_{max} - x_{min}$. Elle a la même unité que l'unité de la variable. Elle n'est pas très informative car elle ne tient pas du tout compte de la répartition des données à l'intérieur de l'intervalle $[x_{min}, x_{max}]$.

Exemple : étendue de la variable « taille ».

2.2.2 L'intervalle inter-quartile

On appelle **intervalle interquartile** l'intervalle $[Q_1, Q_3]$, qui contient environ 50% des observations. La **distance interquartile** $Q_3 - Q_1$ est une mesure de dispersion.

Exemple : intervalle inter-quartile de la variable « taille ».

2.2.3 La variance et l'écart-type

La **variance** est définie par :

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

L'expression suivante est plus pratique pour le calcul de la variance :

$$Var(X) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - (\bar{x})^2.$$

Preuve : en développant le carré dans la définition de la variance.

Pour une **variable quantitative discrète** prenant la valeur v_j un nombre n_j de fois (ou avec la fréquence f_j), pour $1 \leq j \leq k$:

$$\begin{aligned} Var(X) &= \frac{1}{n}\sum_{j=1}^{k} n_j (v_j - \bar{x})^2 = \sum_{j=1}^{k} f_j (v_j - \bar{x})^2 \\ &= \left(\frac{1}{n}\sum_{j=1}^{k} n_j v_j^2\right) - (\bar{x})^2 = \left(\sum_{j=1}^{k} f_j v_j^2\right) - (\bar{x})^2. \end{aligned}$$

Dans le cas d'une variable continue pour laquelle on dispose seulement des **données regroupées en classes**, on peut faire un calcul approché similaire à celui de la moyenne approchée $\bar{x}_{app}$. On calcule une valeur approchée de la variance, notée $Var_{app}(X)$. Toutes les expressions qui suivent sont équivalentes.

$$\begin{aligned} Var_{app}(X) &= \frac{1}{n}\sum_{j=1}^{k} n_j (c_j - \bar{x}_{app})^2 = \sum_{j=1}^{k} f_j (c_j - \bar{x}_{app})^2 \\ &= \left(\frac{1}{n}\sum_{j=1}^{k} n_j c_j^2\right) - (\bar{x}_{app})^2 = \left(\sum_{j=1}^{k} f_j c_j^2\right) - (\bar{x}_{app})^2, \end{aligned}$$

où c_j est le centre de la j-ème classe, dotée de l'effectif n_j (ou de la fréquence relative f_j).

Propriétés de la variance

- La variance est toujours positive ou nulle. Elle est nulle si et seulement si toutes les observations sont identiques :

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \Leftrightarrow \forall i,\ x_i - \bar{x} = 0.$$

- L'unité de la variance est l'unité de X au carré.

L'**écart-type** σ_X est défini par :

$$\sigma_X = \sqrt{Var(X)}.$$

Propriété : l'unité de σ_X est l'unité de X.

Exemple : variance et écart-type de la variable « ciné », de la variable « taille ».

2.3 Changement de variable linéaire ou affine - Variable centrée réduite

2.3.1 Changement de variable linéaire ou affine

On considère une variable quantitative X et on lui fait subir une application affine qui la transforme en une variable Y. a et b sont des constantes réelles.

Nouvelle variable Y	Observations y_i	Moyenne de Y	Variance de Y	Ecart-type de Y
$Y = aX$	$y_i = ax_i$	$\bar{y} = a\bar{x}$	$Var(Y) = a^2 Var(X)$	$\sigma_Y = \lvert a \rvert \sigma_X$
$Y = X + b$	$y_i = x_i + b$	$\bar{y} = \bar{x} + b$	$Var(Y) = Var(X)$	$\sigma_Y = \sigma_X$
$Y = aX + b$	$y_i = ax_i + b$	$\bar{y} = a\bar{x} + b$	$Var(Y) = a^2 Var(X)$	$\sigma_Y = \lvert a \rvert \sigma_X$

Exemple :

2.3.2 Variable centrée réduite

On considère une variable X de moyenne $\bar{x}$ et de variance $Var(X)$, d'écart-type $\sigma_X = \sqrt{Var(X)}$. On définit une nouvelle variable

$$Y = \frac{X - \bar{x}}{\sigma_X}.$$

Elle est **sans unité**. Cette variable est appelée variable **centrée réduite associée à X**. En effet, elle est :

- **centrée** : $\bar{y} = \frac{\bar{x} - \bar{x}}{\sigma_X} = 0$.
- **réduite** $Var(Y) = \frac{Var(Y)}{Var(Y)} = 1$.

Quand on transforme une variable en la variable centrée réduite associée, on retire à cette variable toute l'information concernant son échelle ou unité, et sa *localisation*. Il ne reste plus que des informations sur la **forme** de la distribution. Cette transformation permet de comparer plusieurs variables sur le plan de la forme, même si ce sont des variables exprimées dans des échelles différentes ou qui ont des moyennes complètement différentes.

Exemple : Variable centrée réduite associée à la variable « ciné », à la variable « taille ».

Autre utilisation : Etant donné un individu i pour lequel la variable prend la valeur x_i, on peut situer cet individu dans l'ensemble des observations en calculant son écart à la moyenne réduit :

$$\frac{x_i - \bar{x}}{\sigma_X}.$$

Exemple : quel est l'écart à la moyenne, mesuré en écart-types, d'un individu mesurant 177 cm ?

2.4 Boîtes à moustaches

La boîte à moustaches est une représentation graphique qui permet de visualiser les quartiles ainsi que la dispersion des données et de repérer les données extrêmes ou *outliers*. Elle se fait couramment pour les variables quantitatives continues ou pour les variables quantitatives discrètes prenant un grand nombre de valeurs différentes. En revanche, elle n'a pas beaucoup d'intérêt pour une variable discrète prenant peu de valeurs différentes.

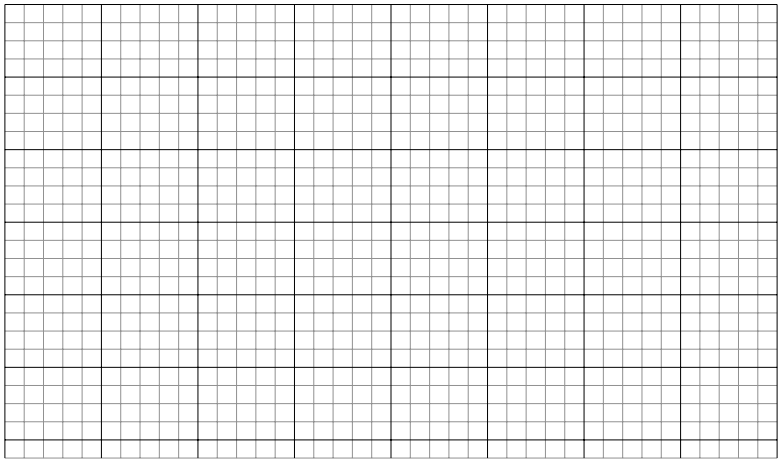
Elle est constituée :

- d'une **boîte** dont les bornes sont les premier et troisième quartile Q_1 et Q_3. A l'intérieur de la boîte figure la médiane Q_2.
- de **moustaches**. On définit tout d'abord deux bornes : $b_- = Q_1 - 1.5(Q_3 - Q_1)$ et $b_+ = Q_3 + 1.5(Q_3 - Q_1)$. On note m_{inf} la plus petite observation supérieure à m_-, et m_{sup} la plus grande observation inférieure à m_+. Soit :

$$\begin{aligned} m_{inf} &= \min\{x_i : x_i \geq m_-\}, \\ m_{sup} &= \max\{x_i : x_i \leq m_+\}, \end{aligned}$$

  La moustache inférieure est le segment $[m_{inf}; Q_1]$. La moustache supérieure, de la même manière, est le segment $[Q_3; m_{sup}]$.
- des **données extrêmes** éventuelles : les observations qui sont en dehors de la boîte et des moustaches, c'est à dire : supérieures à m_+ ou inférieures à m_-. On place ces données une à une quand on en dispose.

Exemple : Boîte à moustaches de la variable « taille »à partir de la série statistique de 20 observations.



Dans le cas où on ne dispose pas des données brutes mais seulement des données regroupées en classes, on utilise les extrémités b_0 et b_k de la première et de la k-ème classe.

- La limite inférieure m_{inf} de la moustache inférieure est $\max\{m_-, b_0\}$ et la limite supérieure m_{sup} de la moustache supérieure est $\min\{m_+, b_k\}$.
- On ne peut pas placer les données extrêmes, sauf si elles sont fournies en plus.

Exemple : Boîte à moustaches de la variable « taille »à partir des données regroupées.