Paramètres de position et caractéristiques de dispersion

Annexe de la la faculté des sciences médicales
Université d'Oum El Bouaghi

Octobre 2023

Introduction

L'étude descriptive d'une variable statistique quantitative (aussi bien discrète que continue) se fait en trois étapes :

- Description préliminaire.

Introduction

L'étude descriptive d'une variable statistique quantitative (aussi bien discrète que continue) se fait en trois étapes :

- Description préliminaire.
- Caractéristiques de position centrale.

Introduction

L'étude descriptive d'une variable statistique quantitative (aussi bien discrète que continue) se fait en trois étapes :

- Description préliminaire.
- Caractéristiques de position centrale.
- Caractéristiques de dispersion.

Paramètres de position centrale

- Les paramètres de position centrale sont un ensemble de valeurs caractéristiques qui permettent une représentation de l'information contenue dans la série statistique.
- Certaines de ces caractéristiques nous renseigneront sur la position (mode, médiane, Quantile).
- D'autres, en plus de cette information, serviront à résumer les données en notre possession (moyenne).

Le mode M_0

Dans le cas d'une variable discrète, le mode est la valeur la plus fréquente de la variable statistique, c'est à dire celle qui correspond au plus grand effectif.

Le mode M_0

Dans le cas d'une variable discrète, le mode est la valeur la plus fréquente de la variable statistique, c'est à dire celle qui correspond au plus grand effectif.

Remarque : Le mode peut ne pas être unique.

Le mode M_0 dans le cas continu

Nous définirons pour le cas d'une variable continue "la classe modale" comme suit : C'est la classe qui correspond au plus grand rapport $\frac{n_i}{a_i}$, a_i étant l'amplitude de la classe $[e_i, e_{i+1}[$.

La moyenne arithmétique

- Quand la série statistique est discrète, de taille N, on appelle moyenne de X le nombre :

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{i=k} n_i x_i = \sum_{i=1}^{i=k} f_i x_i.$$

k étant le nombre de modalités x_i de la variable X.

La moyenne arithmétique

- Quand la série statistique est discrète, de taille N, on appelle moyenne de X le nombre :

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{i=k} n_i x_i = \sum_{i=1}^{i=k} f_i x_i.$$

  k étant le nombre de modalités x_i de la variable X.
- Quand la série statistique est continue, de taille N, pour calculer la moyenne, on utilise la formule précédente en remplaçant x_i par le centre c_i de l'intervalle $[e_i, e_{i+1}[$.

Un tableau statistique

X_i	n_i	f_i	n_iX_i	f_iX_i	$n_iX_i^2$	$f_iX_i^2$
8	12	0,080	96	0,64	768	5,12
10	23	0,153	230	1,53	2300	15,3
16	41	0,273	656	4,368	10496	69,88
20	24	0,160	480	3,2	9600	64
24	22	0,147	528	3,528	12672	84,67
32	16	0,107	512	3,424	1638	109,56
42	12	0,080	504	3,36	21168	141,12
Total	150	1	3006	**20,05**	73388	**489,66**
Total/N			**20,04**		**489,253**	

Figure – Tableau récapitulatif de l'exemple 1.

Les quantiles

Definition

On appelle quantiles les valeurs du caractère qui définissent les bornes d'une partition en classes d'effectifs égaux.

La médiane

Definition

La médiane est la valeur du caractère qui sépare la série ordonnée en valeurs croissantes en deux groupes de même effectif.

La médiane

Definition

La médiane est la valeur du caractère qui sépare la série ordonnée en valeurs croissantes en deux groupes de même effectif.

Pour la trouver dans le cas discret on distingue deux cas :

La médiane

Definition

La médiane est la valeur du caractère qui sépare la série ordonnée en valeurs croissantes en deux groupes de même effectif.

Pour la trouver dans le cas discret on distingue deux cas :

- si l'effectif total N est un nombre impair, la médiane est le terme de rang $\frac{N+1}{2}$.

La médiane

Definition

La médiane est la valeur du caractère qui sépare la série ordonnée en valeurs croissantes en deux groupes de même effectif.

Pour la trouver dans le cas discret on distingue deux cas :

- si l'effectif total N est un nombre impair, la médiane est le terme de rang $\frac{N+1}{2}$.
- si l'effectif total N est un nombre pair, la médiane est la moyenne des termes de rang $\frac{N}{2}$ et $\frac{N+1}{2}$.

La médiane dans le cas continu

Quand les données sont regroupés en classes (cas continu), on peut déterminer la médiane par "interpolation linéaire".

Le premier et le troisième quartile

- Le premier quartile Q_1 d'une série est la valeur du caractère qui correspond à la fréquence cumulée de 25%.

Le premier et le troisième quartile

- Le premier quartile Q_1 d'une série est la valeur du caractère qui correspond à la fréquence cumulée de 25%.
- Le troisième quartile Q_3 d'une série est la valeur du caractère qui correspond à la fréquence cumulée de 75%.

Le premier et le troisième quartile

- Le premier quartile Q_1 d'une série est la valeur du caractère qui correspond à la fréquence cumulée de 25%.
- Le troisième quartile Q_3 d'une série est la valeur du caractère qui correspond à la fréquence cumulée de 75%.

Comment calculer les quartiles ?

- Le deuxième quartile Q_2 coïncide avec la médiane.

Comment calculer les quartiles ?

- Le deuxième quartile Q_2 coïncide avec la médiane.
- Si $N/4$ est un entier, Q_1 est le terme de rang $N/4$ et Q_3 est le terme de rang $3N/4$.
- Si n/4 n'est pas un entier, Q1 et Q3 sont respectivement les termes de rang immédiatement supérieur à $N/4$ et $3N/4$.

Comment calculer les quartiles ?

- Le deuxième quartile Q_2 coïncide avec la médiane.
- Si $N/4$ est un entier, Q_1 est le terme de rang $N/4$ et Q_3 est le terme de rang $3N/4$.
- Si n/4 n'est pas un entier, Q1 et Q3 sont respectivement les termes de rang immédiatement supérieur à $N/4$ et $3N/4$.
- Exemple : La série statistique suivante :
  9, 10, 10, 12, 13, 15, 16, 17, 17, 8, 13, 14.

Les déciles et les centiles

- On définit aussi les déciles (la série est partagée en 10 sous ensembles de même effectif) On obtient : $D_1, D_2, , D_9$ (même technique).

Les déciles et les centiles

- On définit aussi les déciles (la série est partagée en 10 sous ensembles de même effectif) On obtient : $D_1, D_2, , D_9$ (même technique).
- Les centiles partagent la série en 100 parties de même effectif. Les centiles sont les 99 valeurs de X qui permettent de découper la distribution en 100 classes d'effectifs égaux. On les note $C_1, C_2, , C_{99}$.

Moyenne et médiane avantages et inconvénients

- Dans les petits groupes, la médiane est en général plus représentative que la moyenne.

Moyenne et médiane avantages et inconvénients

- Dans les petits groupes, la médiane est en général plus représentative que la moyenne.
  Exemple 1 : 2,2,3, 4, 19.
- Dans les groupes modérés (plusieurs dizaines ou centaines de données) la médiane et la moyenne se confondent.

Moyenne et médiane avantages et inconvénients

- Dans les petits groupes, la médiane est en général plus représentative que la moyenne.
  Exemple 1 : 2,2,3, 4, 19.
- Dans les groupes modérés (plusieurs dizaines ou centaines de données) la médiane et la moyenne se confondent.
- Dans les grands groupes, c'est toutefois la médiane qui peut donner une image peu représentative de la réalité.

Les paramètres de dispersion

Des paramètres qui servent à tirer des conclusion concernant la dispersion des valeurs de la variable autour du centre, ces indices sont :

- L'étendue.

Les paramètres de dispersion

Des paramètres qui servent à tirer des conclusion concernant la dispersion des valeurs de la variable autour du centre, ces indices sont :

- L'étendue.
- L'écart interquartile.

Les paramètres de dispersion

Des paramètres qui servent à tirer des conclusion concernant la dispersion des valeurs de la variable autour du centre, ces indices sont :

- L'étendue.
- L'écart interquartile.
- La variance.

Les paramètres de dispersion

Des paramètres qui servent à tirer des conclusion concernant la dispersion des valeurs de la variable autour du centre, ces indices sont :

- L'étendue.
- L'écart interquartile.
- La variance.
- L'écart-type.

L'étendue

- Quand la série statistique est discrète, l'étendue est la différence entre la plus grande valeur et la plus petite valeur de la série.

$$E = x_{max} - x_{min}$$

L'étendue

- Quand la série statistique est discrète, l'étendue est la différence entre la plus grande valeur et la plus petite valeur de la série.

$$E = x_{max} - x_{min}$$

- Quand la série statistique est continue, l'étendue est la longueur de l'intervalle sur lequel se disperse la variable.

L'écart interquartile

C'est la différence entre le troisième quartile et le premier quartile.
$I_Q = Q_3 - Q_1$.

L'écart interquartile

C'est la différence entre le troisième quartile et le premier quartile.
$I_Q = Q_3 - Q_1$.
L'intervalle interquartile mesure la dispersion de la série, Il correspond à 50% de la population .

La Variance

LA variance d'une série est la moyenne des carrés des écarts de chaque valeur à la moyenne $\bar{X}$. C'est un nombre positif. On la note $Var(X)$. La variance est donnée par :

$$Var(X) = \frac{1}{N}\sum_{i=1}^{i=k} n_i(x_i - \bar{X})^2 = \sum_{i=1}^{i=k} f_i(x_i - \bar{X})^2.$$

La Variance

LA variance d'une série est la moyenne des carrés des écarts de chaque valeur à la moyenne $\bar{X}$. C'est un nombre positif. On la note $Var(X)$. La variance est donnée par :

$$Var(X) = \frac{1}{N}\sum_{i=1}^{i=k} n_i(x_i - \bar{X})^2 = \sum_{i=1}^{i=k} f_i(x_i - \bar{X})^2.$$

Pour simplifier les calculs on préfère la formule :

$$Var(X) = \frac{1}{N}\sum_{i=1}^{i=k} n_i x_i^2 - \bar{X}^2 = \sum_{i=1}^{i=k} f_i x_i^2 - \bar{X}^2.$$

La Variance

LA variance d'une série est la moyenne des carrés des écarts de chaque valeur à la moyenne $\bar{X}$. C'est un nombre positif. On la note $Var(X)$. La variance est donnée par :

$$Var(X) = \frac{1}{N}\sum_{i=1}^{i=k} n_i(x_i - \bar{X})^2 = \sum_{i=1}^{i=k} f_i(x_i - \bar{X})^2.$$

Pour simplifier les calculs on préfère la formule :

$$Var(X) = \frac{1}{N}\sum_{i=1}^{i=k} n_i x_i^2 - \bar{X}^2 = \sum_{i=1}^{i=k} f_i x_i^2 - \bar{X}^2.$$

Si la série est regroupée en classes, on remplace les x_i par les centres des classes.

L'écart-type

L'écart-type d'une série est égal à la racine carrée de la variance et est notée $\sigma(X) = \sqrt{Var(X)}$.

Paramètres de dispersion relative

La comparaison des paramètres de dispersion absolue de deux caractères n'a de sens que si les deux caractères sont de même nature et de même ordre de grandeur.

Paramètres de dispersion relative

La comparaison des paramètres de dispersion absolue de deux caractères n'a de sens que si les deux caractères sont de même nature et de même ordre de grandeur.
Adoption des paramètres qui mettent en considération l'ordre de grandeur et la nature des données.

Paramètres de dispersion relative

La comparaison des paramètres de dispersion absolue de deux caractères n'a de sens que si les deux caractères sont de même nature et de même ordre de grandeur.
Adoption des paramètres qui mettent en considération l'ordre de grandeur et la nature des données.
Dispersion relative = Paramètre de dispersion absolue/Valeur centrale.
Les plus courants sont :

- Le coefficient de variation C.V. = écart-type/moyenne.
- Le coefficient interquartile relatif C.I.R. = (Q3-Q1)/Q2.

Paramètres de dispersion relative

La comparaison des paramètres de dispersion absolue de deux caractères n'a de sens que si les deux caractères sont de même nature et de même ordre de grandeur.
Adoption des paramètres qui mettent en considération l'ordre de grandeur et la nature des données.
Dispersion relative = Paramètre de dispersion absolue/Valeur centrale.
Les plus courants sont :

- Le coefficient de variation C.V. = écart-type/moyenne.
- Le coefficient interquartile relatif C.I.R. = (Q3-Q1)/Q2.

Interprétation du C.V.

Le coefficient de variation est un indicateur de l'homogénéité de la population.

- Un coefficient de variation inférieur à 15% indique que la population est homogène.

Interprétation du C.V.

Le coefficient de variation est un indicateur de l'homogénéité de la population.

- Un coefficient de variation inférieur à 15% indique que la population est homogène.
- Une valeur du C.V. supérieure à 15% indique que les valeurs sont relativement dispersées.

Diagramme en boite

- Le diagramme en boite, la boite à moustaches ou boîte de Tukey (Box-Plot, en anglais), est une invention de TUKEY (1977) pour représenter schématiquement une distribution.
- Cette représentation graphique peut être un moyen pour approcher les concepts abstraits de la statistique.
- Le diagramme en boite utilise 5 valeurs qui résument des données : le minimum, les 3 quartiles Q1, Q2 (médiane), Q3, et le maximum.

Exemple de Box-Plot

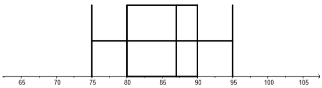

Figure – Exemple 1

- La ligne au milieu de la boîte représente la médiane, qui est la valeur centrale des données triées.

Exemple de Box-Plot

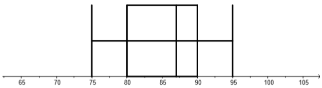

Figure – Exemple 1

- La ligne au milieu de la boîte représente la médiane, qui est la valeur centrale des données triées.
- La boîte elle-même représente Q_1 et Q_3.

Exemple de Box-Plot

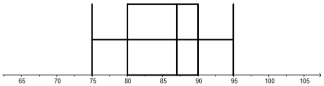

Figure – Exemple 1

- La ligne au milieu de la boîte représente la médiane, qui est la valeur centrale des données triées.
- La boîte elle-même représente Q_1 et Q_3.
- La longueur de la boîte est donc la distance interquartile I_Q, qui mesure la dispersion des données au sein du 50% central de l'échantillon.

Exemple de Box-Plot

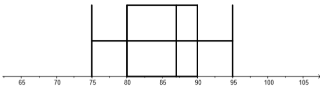

Figure – Exemple 1

- La ligne au milieu de la boîte représente la médiane, qui est la valeur centrale des données triées.
- La boîte elle-même représente Q_1 et Q_3.
- La longueur de la boîte est donc la distance interquartile I_Q, qui mesure la dispersion des données au sein du 50% central de l'échantillon.
- Les segments (moustaches) qui s'étendent à partir de la boîte représentent l'étendue des données à l'extérieur des quartiles.

Un autre exemple de Box-Plot

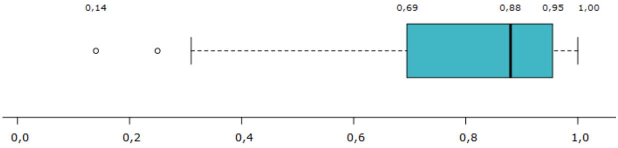

Figure – Exemple 2

- La distribution est asymétrique, car la portion gauche de la boîte et la moustache gauche sont plus longues que du côté droit.

Un autre exemple de Box-Plot

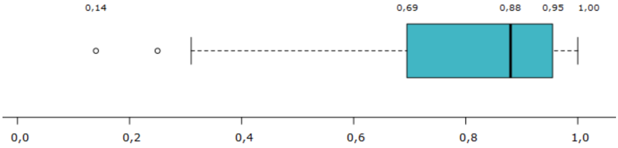

Figure – Exemple 2

- La distribution est asymétrique, car la portion gauche de la boîte et la moustache gauche sont plus longues que du côté droit.
- La distribution en question inclut des valeurs potentiellement extrêmes.

Un autre exemple de Box-Plot

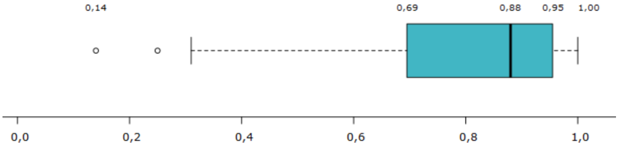

Figure – Exemple 2

- La distribution est asymétrique, car la portion gauche de la boîte et la moustache gauche sont plus longues que du côté droit.
- La distribution en question inclut des valeurs potentiellement extrêmes.
- Toute valeur qui dépasse $Q_3 + 1.5(Q_3 - Q_1)$ (resp. inférieure à $Q_1 - 1.5(Q_3 - Q_1)$) se trouve à l'extérieur de la moustache et est indiquée par un cercle.

Merci pour votre attention !