LES STATISTIQUES DESCRIPTIVES

La description statistique consiste en un résumé numérique d'une distribution statistique par des indicateurs numériques ou paramètres caractéristiques. Elle représente une transition entre les statistiques purement descriptives et les statistiques inductives.

Définition

Les statistiques descriptives sont l'ensemble des méthodes et des techniques permettant de présenter, de décrire et de résumer, des données nombreuses et variées ; elles nous permettent de connaître les caractéristiques de notre échantillon.
L'échantillon tente de nous donner des informations sur la population, tels que la moyenne et la variance.

I. Les indicateurs numériques

A. Les indicateurs de position

Un paramètre statistique est dit de position s'il s'agit d'un nombre clé permettant de préciser où se répartit une certaine fraction des observations ; il a pour objectif dans le cas d'un caractère quantitatif de caractériser l'ordre de grandeur des observations.
Nous traitons le cas des paramètres statistiques de tendance centrale ; un paramètre statistique est dit de tendance centrale s'il s'agit d'un nombre clé autour duquel les observations sont réparties comme le mode, la médiane et la moyenne.

1. La moyenne

Une moyenne est une valeur caractéristique ou représentative d'un ensemble de données. Dans la mesure où cette valeur caractéristique a tendance à se situer au milieu d'un ensemble de données rangées par ordre croissant, on dit aussi que l'on a une mesure de tendance centrale.
On peut définir plusieurs types de moyennes, les plus courantes étant la moyenne arithmétique ou plus simplement la moyenne, la moyenne géométrique et la moyenne harmonique. Chacune d'elles présente des avantages et des inconvénients, qui dépendent des données et de ce que l'on veut en faire.

Rq : la somme algébrique des écarts d'un ensemble de nombres à leur moyenne arithmétique est nulle.

- **Moyenne arithmétique :**

Soit un échantillon de n valeurs observées x_1, x_2, ….,x_i,….,x_n d'un caractère quantitatif X,

on définit sa moyenne observée $\overline{X}$ comme **la moyenne arithmétique** des n valeurs. C'est le quotient $\overline{X}$ de la somme d'une série d'observations par leur nombre.

- **Pour une série brute :**

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- **Pour une série groupée :** de k classes (x_i, n_i), où x_i la valeur centrale de la classe et n_i l'effectif de chaque classe :

Si les données observées x_i sont regroupées en k classes d'effectif ni (caractère continu regroupé en classe ou caractère discret), il faut les pondérer par les effectifs correspondants:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{K} n_i x_i$$

avec n (effectif total), ni (effectif de chaque classe) et k (nbre des classes)

Rq : Une des propriétés de la moyenne arithmétique est que la somme des écarts à la moyenne est nulle.

Rq : La moyenne obtenue après regroupement des données en classe diffère légèrement en raison d'une perte d'information. Si l'échantillonnage n'est pas de type aléatoire simple, les deux moyennes peuvent être très différentes.

- **moyenne arithmétique pondérée**

Souvent on associe aux nombres X_1, X_2 ..., X_k des facteurs d'importance dépendant de la signification ou de l'importance que l'on donne aux nombres.

La moyenne des nombres x_1 , x_2 , ... , x_n , pondérée par les poids p_1 , p_2 , ... , p_n est égale à :

$$\overline{X} = \sum_{i=1}^{n} p_i x_i$$

2. La médiane

La médiane d'un ensemble de nombres rangés par ordre de grandeur croissante est la valeur du milieu ou la moyenne arithmétique des valeurs centrales. Géométriquement la médiane est la valeur de X (l'abscisse X) correspondant à la verticale qui divise un histogramme en deux parties d'aires égales.

La médiane M_e est aussi la valeur du caractère pour laquelle la fréquence cumulée est égale à 50% de l'ensemble des effectifs ; elle correspond donc au centre de la série statistique classée par ordre croissant, ou à la valeur pour laquelle 50% des valeurs observées sont supérieures et 50% sont inférieures.

- Dans le cas où les valeurs prises par le caractère étudié (variable) ne sont pas regroupées en classe :

- si n est **impair**, avec m=(n-1)/2 ; la médiane serait la valeur du milieu Me = xm+1.
- si n est **pair**, avec m=n/2 ; la médiane serait une valeur quelconque entre xm et xm+1, dans ce cas il peut être commode de prendre le milieu

- Dans le cas où les valeurs prises par le caractère étudié sont groupées en classe, on cherche la classe correspondant ne/2 de l'échantillon où (ne) est l'effectif cumulé.

Exemple 1: L'ensemble des nombres 3, 4, 4, 5, 6, 8, 8, 8, 10 a pour médiane 6.
Exemple 2: L'ensemble des nombres 5, 5, 7, 9, 11, 12, 15, 18 a pour médiane (9+11)/2=10.

Rq : Si la distribution des valeurs est symétrique, la valeur de la médiane est proche de la valeur de la moyenne arithmétique.

3. Le mode

Le mode d'un ensemble de nombres est le nombre que l'on rencontre le plus fréquemment, c'est-à-dire celui qui a la plus grande fréquence. Le mode peut ne pas exister, et s'il existe, il peut ne pas être unique.
Le mode M_o d'une série statistique est la valeur du caractère la plus fréquente ou dominante dans l'échantillon. Le mode correspond à la classe de fréquence maximale dans la distribution des fréquences.
Rq: Une distribution de fréquences peut présenter un seul mode (distribution unimodale) ou plusieurs modes (distribution bi ou trimodale).
Rq: Si la distribution des valeurs est symétrique, la valeur du mode est proche de la valeur de la moyenne arithmétique. $M_o \approx x$

❖ Avantages et inconvénients des différents indicateurs de position

• Moyenne arithmétique

- Facile à calculer,
- Fortement influencée par les valeurs extrêmes de la distribution des variables
- Représente mal une population hétérogène (polymodale).

• Médiane

- Pas influencée par les valeurs extrêmes de la distribution des variables
- Peu sensible aux variations d'amplitude des classes,
- Calculable sur des caractères cycliques où la moyenne a peu de signification.
- Se prête mal aux calculs statistiques,
- Suppose l'équi-répartition des données
- Ne représente que la valeur qui sépare l'échantillon en 2 parties égales.

• Mode

- Pas influencée par les valeurs extrêmes de la distribution des variables
- Calculable sur des caractères cycliques où la moyenne a peu de signification,

- Bon indicateur d'une population hétérogène.
- Se prête mal aux calculs statistiques,
- Très sensible aux variations d'amplitude des classes,

RQ : Dans le cas où le caractère étudié se distribue selon une loi normale Laplace-Gauss, la moyenne x, la médiane Me et le mode M_o prennent la même valeur.

B. Les indicateurs de dispersion

Un paramètre statistique est dit de dispersion s'il s'agit d'un nombre clé résumant la plus ou moins grande disparité des observations et leur variabilité de part et d'autre de la tendance centrale.
Un ensemble de données numériques tend généralement à s'étaler autour d'une valeur centrale, que l'on appelle dispersion ou variabilité des données. Il existe plusieurs mesures de la dispersion ; les plus courantes sont l'étendue, l'écart moyen, l'écart interquartile, la distance entre le 1er et le $9^{ème}$ décile, la distance entre le 10` et le 90` centile, l'écart-type, la variance et le coefficient de variation.

1. L'étendue

L'étendue d'un ensemble de nombres est la différence entre le plus grand et le plus petit de ces nombres.
Exemple : L'étendue de l'ensemble 2, 3, 3, 5, 5, 5, 8, 10, 12, est 12 - 2 = 10.
On note souvent l'étendue en indiquant simplement le plus petit et le plus grand des nombres considérés. Ainsi dans l'exemple précédent l'étendue peut être notée : 2 - 12 ou de 2 à 12

2. l'écart moyen ou l'écart à la moyenne

On définit aussi l'écart moyen en fonction des écarts absolus à la médiane ou à tout autre indicateur de tendance centrale. C'est l'ensemble des écarts à la moyenne ou à la médiane sur l'effectif.
Rq : L'écart moyen par rapport à la médiane est inférieur à l'écart moyen par rapport à tout autre indicateur.

3. L'écart-type

On définit l'écart-type s d'un ensemble de n nombres comme la racine carrée de la somme des carrés des écarts à la moyenne, ou encore la racine carrée de la variance.

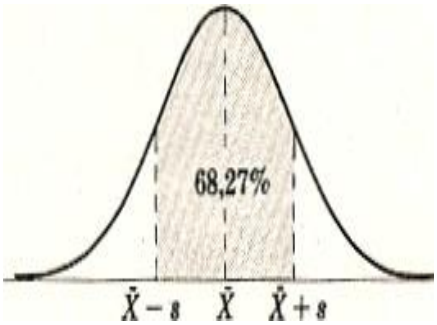
$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{ou} \quad \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}$$

Rq : Une distribution aura un écart-type d'autant plus faible (proche de 0) qu'elle sera ramassée autour de la moyenne, avec des valeurs très peu différentes les unes des autres.

❖ propriétés de l'écart-type

- 68,27 % des cas sont compris entre X - s et X + s (un écart-type de part et d'autre de la moyenne)
- 95,45 % des cas sont compris entre X - 2s et X+ 2s (deux fois l'écart-type de part et d'autre de la moyenne)
- 99,73 % des cas sont compris entre X- 3s et X+ 3s (trois fois l'écart-type de part et d'autre de la moyenne)

Rq : Ces pourcentages sont approximativement valables pour des distributions légèrement dissymétriques.



4. L'intervalle semi-interquartile ou écart des quartiles

Si un ensemble de nombres est rangé par ordre de grandeur croissante, le nombre du milieu (ou la moyenne des nombres centraux) divisant l'ensemble en deux parties égales est la médiane. Par extension, on peut penser aux valeurs qui divisent l'ensemble en quatre parties égales. On note ces valeurs **Q_1, Q_2 et Q_3** et on les appelle respectivement le premier, le deuxième et les troisièmes quartiles, Q_2 étant en fait la médiane.
De même, on appelle déciles les valeurs qui divisent les données en 10 parties égales et on les note $D_1, D_2, \ldots, D_9$, tandis que les valeurs divisant les données en 100 parties égales sont appelées quantiles d'ordre 100 centiles ou percentiles**.** Les 25^e et 75^e centiles correspondent respectivement au 1^{er} et aux 3èmes quartiles.
L'intervalle semi-interquartile ou écart des quartiles d'un ensemble de données est défini par Intervalle semi-interquartile $Q = {}^{Q3-Q1}$ où Q_1 et Q_3 sont les premier et troisième quartiles des données.
L'intervalle entre les 10' et 90e centiles d'un ensemble de données est défini par l'intervalle entre le 10e et 90e centile = $P_{90} - P_{10}$ où $_{P10,}$ et P_{90} sont les 10e et 90e centiles des données. Le demi-intervalle entre le 10e et 90e centile, $(P_{90} - P10)/2$ peut être pris en considération mais demeure d'un emploi rare.

5. La variance

La variance et l'écart-type observée sont des paramètres de dispersion absolue qui mesurent la variation absolue des données indépendamment de l'ordre de grandeur des données. Ces

paramètres ont pour objectif dans le cas d'un caractère quantitatif de caractériser la variabilité des données dans l'échantillon.

On définit la variance d'un ensemble de données par le carré de l'écart-type. On utilise les symboles a pour distinguer l'écart-type d'une population et le symbole s pour distinguer l'écart-type d'un échantillon, de sorte que a^2 et s^2 sont respectivement la variance de la population et la variance de l'échantillon.

Soit un échantillon de n valeurs observées x1, x2, ….,xi,….,xn d'un caractère quantitatif X et soit x sa moyenne observée. On définit la variance observée notée $\mathbf{s^2}$ comme la moyenne arithmétique des carrés des écarts à la moyenne. Pour des commodités de calcul, on se sert du théorème de Koenig :

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \overline{x}^2$$

Rq : De part sa définition, la variance est toujours un nombre positif. Sa dimension est le carré de celle de la variable. Il est toutefois difficile d'utiliser la variance comme mesure de dispersion car le recours au carré conduit à un changement d'unités. Elle n'a donc pas de sens direct, contrairement à l'écart-type qui s'exprime dans les mêmes unités que la moyenne.

6. La covariance

On appelle covariance de deux variables statistiques X et Y sur un nombre d'individus d'effectif (n) la "Moyenne des produits moins le produit des moyennes"

$$\mathrm{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \left(\frac{1}{n}\sum_{i=1}^{n}x_i y_i\right) - \overline{x}\,\overline{y}$$

Rq : Ce nombre est positif si X et Y ont tendance à varier dans le même sens, et négatif si elles ont tendance à varier en sens contraire.

7. Le coefficient de variation

Le coefficient de variation noté C.V. est un indice de dispersion relatif, il est exprimé en pourcentage, il est indépendant du choix des unités de mesure. Il est le rapport de l'écart-type à la moyenne, il est exprimé en pourcentage, il permet la comparaison de distributions de valeurs dont les échelles de mesure ne sont pas comparables.

$$C.V. = 100*s/x$$

Rq : Plus la valeur du coefficient de variation est élevée, plus la dispersion autour de la moyenne est grande.

Rq : Lorsque l'on dispose de valeurs estimées, le CV rapporte l'écart-type de l'estimation à la valeur de cette estimation. Plus la valeur du coefficient de variation est faible, plus l'estimation est précise.

8. La variable centrée réduite

La variable Z= (X-Xi)/s qui mesure l'écart de la moyenne en unités d'écart-type est appelée variable centrée réduite. C'est une variable sans dimension, elle est indépendante du choix des unités.

Quand les écarts à la moyenne sont donnés en unités d'écart-type, on dit qu'ils sont exprimés en unités centrées réduites. Ces unités sont d'un grand intérêt pour comparer des distributions.

❖ Relations empiriques entre les différentes mesures de la dispersion

Pour des distributions légèrement dissymétriques, on a des relations empiriques comme pour la distribution normale ; nous trouvons que l'écart-moyen et l'intervalle semi-interquartile sont respectivement 0,7979 et 0,6745 fois l'écart-type.
La variation effective ou dispersion qu'on détermine à partir de l'écart-type ou de toute autre mesure de la dispersion est appelée dispersion absolue.
La dispersion relative est définit par :
Dispersion relative = dispersion absolue/moyenne
La dispersion relative est appelée coefficient de variation ou coefficient de dispersion qui est commode pour comparer des distributions où les unités sont différentes.