EPITECH
PROBABILITES ET STATISTIQUES
Cours206
Statistique descriptive

Dominique Neveu

Année 2009-2010

Table des matières

Résumé du cours

La partie des statistiques qui permet la description et l'analyse d'un groupe donné, sans vouloir en tirer de conclusions à propos d'un groupe plus important s'appelle la statistique descriptive.

On définit d'abord la population, composée d'individus, qui forme le champ d'analyse d'une étude particulière. Les individus sont étudiés suivant un ou plusieurs caractères que l'on observe. On dispose alors de données. L'objet de ce chapitre est de montrer comment étudier ces données, pour obtenir des caractéristiques de leur répartition.

Pour caractériser une répartition de données, on utilise une mesure de leur tendance centrale. Plusieurs mesures de ce type sont couramment utilisées : différentes formules de moyennes, la médiane et le mode.

Enfin, en complément de la mesure de tendance centrale, l'allure de la répartition des données est caractérisée par les mesures de dispersion.

Chapitre 1

Statistique descriptive

1.1 Définitions

1.1.1 Population

La statistique est une étude des données numériques recueillies sur des groupements d'êtres, de choses ou même de faits. Elle s'est occupée en premier lieu de questions démographiques et les premiers groupements étudiés ont été des populations. Nous désignons sous le nom de **population** tout ensemble soumis à une étude statistique. Par exemple, dans une étude sur la production d'une usine, la population est l'ensemble des pièces produites sur une certaine période.

Les unités composant une population sont soit des êtres humains, soit des êtres vivants, soit des objets inanimés. La population est donc constituée d'un ensemble d'éléments que l'on appelle **individus** ou **unités statistiques**.

1.1.2 Echantillon

Parfois, on est amené à étudier un petit groupe de membres d'une population plus large. Ce groupe est alors appelé un **échantillon**. Par exemple, cent pièces prélevées parmi l'ensemble des pièces fabriquées en usine constituent un échantillon de la production d'une machine.

1.1.3 Caractère

Dans l'étude d'une population, l'attention se porte en général sur un trait déterminé commun à tous ses membres et appelé **caractère**. Les traits étudiés sont de nature variée : renseignements démographiques (naissances, décès, mariages...) ; données fiscales (impôts, salaires, revenus...) ; mesures physiques (tailles, âges, poids...).

Le caractère que l'on étudie ne se présente pas de la même façon chez tous les membres de la population. Parfois c'est son intensité qui varie, parfois c'est sa nature. Ainsi, on est amené à considérer deux sortes de caractères :
1° les caractères **quantitatifs** comme l'âge d'une personne, le poids d'un objet, la production d'une usine...
2° les caractères **qualitatifs** tels que le sexe, la profession, la couleur des yeux d'une personne.

Caractères quantitatifs

Les caractères quantitatifs se subdivisent eux-mêmes en deux espèces : Les caractères discrets et les caractères continus. Les caractères discrets sont ceux qui ne prennent que des valeurs isolées souvent entières. Par exemple, le nombre des enfants d'une famille, le nombre de pièces d'un logement... Les caractères continus sont ceux qui peuvent prendre n'importe quelle valeur dans un intervalle déterminé ou infini. Par exemple, la taille d'une personne, la vitesse du vent... Lorsque les valeurs que peut prendre un caractère quantitatif sont trop abondantes, il est commode de regrouper ces valeurs en classes. Par exemple, la taille d'un enfant sera défini suivant les classes 0-30 cm, 30-60 cm, 60-90 cm...

Caractères qualitatifs

Lorsque le caractère est qualitatif, il est l'objet d'une énumération. L'énumération sera de préférence courte (dizaine de valeurs).

1.2 Mesures de tendance centrale

Une distribution de valeurs peut souvent être caractérisée par la mesure de l'emplacement du centre et une mesure de la dispersion autour de ce centre. Dans ce paragraphe, nous examinerons la première de ces caractéristiques.

Une moyenne est une valeur caractéristique d'un jeu de données. Puisque de telles valeurs ont tendance à se trouver dans la zone centrale d'un jeu de données rangées par ordre croissant, les moyennes sont aussi appelées **mesures de tendance centrale**.

On peut définir plusieurs types de moyennes, les plus courants étant :

- la moyenne arithmétique
- la moyenne quadratique
- la moyenne géométrique
- la moyenne harmonique

- la médiane
- le mode

Chacun présente des avantages et des inconvénients, selon les données disponibles et l'objectif attendu.

1.2.1 Moyenne arithmétique

La moyenne arithmétique de plusieurs données est le quotient de leur somme par le nombre de ces données. Soient n données notées x_1, x_2, ...,x_n, leur moyenne arithmétique notée $\overline{x}$ est définie par l'égalité :

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

Cette moyenne est la plus utilisée. Elle est fondamentale dans la théorie et très commode dans la pratique.

Exemple 1 *La moyenne des cinq nombres :*

$$8, 10, 9, 12, 13$$

s'obtient en faisant leur somme, 52, et en divisant par 5, ce qui donne 10,4.

Lorsque chaque donnée se présente plusieurs fois dans la série, c'est à dire que chaque donnée x_i admet un effectif e_i, l'expression de la moyenne arithmétique devient :

$$\overline{x} = \frac{e_1.x_1 + e_2.x_2 + ... + e_n.x_n}{e_1 + e_2 + ... + e_n} = \frac{1}{N}\sum_{i=1}^{n} e_i.x_i$$

où N est l'effectif total de la série égal à :

$$N = e_1 + e_2 + ... + e_n$$

Moyenne arithmétique pondérée

On associe parfois aux nombres x_1, x_2, ...,x_n, des facteurs de pondération (ou poids) notés w_1, w_2, ...,w_n, en fonction de l'importance accordée aux nombres. Dans ce cas, on calcule une moyenne dite pondérée donnée par la formule :

$$\overline{x} = \frac{w_1.x_1 + w_2.x_2 + ... + w_n.x_n}{w_1 + w_2 + ... + w_n} = \frac{\sum_{i=1}^{n} w_i.x_i}{\sum_{i=1}^{n} w_i}$$

Exemple 2 *Si un examen final a un coefficient trois fois plus fort que les deux examens partiels et que l'étudiant a obtenu les notes de 18, 9 et 15, la note moyenne de cet étudiant est :*

$$\overline{x} = \frac{1.18 + 1.9 + 3.15}{1 + 1 + 3} = \frac{1.18 + 1.9 + 3.15}{1 + 1 + 3} = \frac{72}{5} = 14,4$$

1.2.2 Moyenne quadratique

La **moyenne quadratique** d'un jeu de données x_1, x_2, ...,x_n, notée Q, est définie par :

$$Q = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + ... + x_n^2)}$$

Ce type de moyenne est souvent utilisé dans les applications de physique.

Exemple 3 *La moyenne quadratique des cinq nombres :*

$$8, 10, 9, 12, 13$$

est égale à :

$$Q = \sqrt{\frac{1}{5}(8^2 + 10^2 + 9^2 + 12^2 + 13^2)} = \sqrt{\frac{558}{5}} = \sqrt{111,6} \simeq 10,564$$

1.2.3 Moyenne géométrique

On définit la **moyenne géométrique** d'un jeu de données x_1, x_2, ...,x_n, positives notée G de la façon suivante :

$$G = \sqrt[n]{x_1.x_2...x_n} = (x_1.x_2...x_n)^{\frac{1}{n}}$$

Exemple 4 *La moyenne géométrique des cinq nombres :*

$$8, 10, 9, 12, 13$$

est égale à :

$$G = \sqrt[5]{8.10.9.12.13} = \sqrt[5]{112320} \simeq 10,235$$

La moyenne géométrique est couramment utilisée en économie. La moyenne géométrique admet la propriété que son logarithme s'exprime simplement en fonction des logarithmes de chaque donnée. On a la formule suivante :

$$log(G) = \frac{1}{n}.(log(x_1) + log(x_2) + ... + log(x_n)) = \frac{1}{n}.\sum_{i=1}^{n} log(x_i)$$

1.2.4 Moyenne harmonique

On définit la **moyenne harmonique** notée H d'un jeu de données x_1, x_2, ...,x_n, ne prenant pas la valeur nulle de la façon suivante :

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}}$$

Exemple 5 *La moyenne harmonique des cinq nombres :*

$$8, 10, 9, 12, 13$$

est égale à :

$$H = \frac{5}{\frac{1}{8} + \frac{1}{10} + \frac{1}{9} + \frac{1}{12} + \frac{1}{13}} \simeq \frac{5}{0,496} \simeq 10,073$$

Comme la moyenne géométrique, la moyenne harmonique est largement utilisée en économie.

1.2.5 Comparaison entre moyennes

Les moyennes arithmétique et quadratique attribuent plus d'influence aux valeurs élevées des jeux de données (la moyenne quadratique plus que la moyenne arithmétique). Les moyennes géométrique et harmonique réduisent l'influence des valeurs les plus grandes et augmentent celle des plus petites (la moyenne harmonique plus que la moyenne géométrique).

On peut classer les moyennes arithmétique $\overline{x}$, quadratique Q, géométrique G et harmonique H de la manière suivante :

$$H \leq G \leq \overline{x} \leq Q$$

Exemple 6 *On considère le jeu de données suivant :*

$$8, 10, 9, 12, 13$$

Les différentes moyennes sont égales à :

$$\overline{x} = 10,4$$

$$Q \simeq 10,564$$

$$G \simeq 10,235$$

$$H \simeq 10,073$$

On retrouve bien l'ordre croissant des moyennes :

$$10,073 \leq 10,235 \leq 10,4 \leq 10,564$$

1.2.6 Médiane

La médiane, symbolisée par *med*, est le point qui partage la distribution d'une série d'observations en deux parties égales. Considérons un jeu de n données x_1, x_2, ...,x_n, rangées par ordre croissant :

$$x_1 \leq x_2 \leq ... \leq x_n$$

Pour définir la médiane de ce jeu de données, deux cas sont à envisager suivant la parité de n. Si n est impair, n=2p+1 la médiane est la valeur située au milieu de la série, soit :

$$med = x_{p+1}$$

Si n est pair, n=2p, la médiane est la demi somme des deux valeurs situées au milieu de la série, soit :

$$med = \frac{x_p + x_{p+1}}{2}$$

Exemple 7 *On considère le jeu de cinq données suivant :*

$$8, 9, 10, 12, 13$$

La médiane est la valeur 10. On considère un autre jeu de six données suivant :

$$2, 5, 8, 10, 14, 15$$

La médiane est la demi somme des valeurs 8 et 10, soit la valeur 9.

D'après sa définition, la médiane est toujours définie, mais elle n'a d'utilité que s'il y a réellement autant de valeurs en dessous d'elle qu'au dessus d'elle. Lorsque plusieurs valeurs coïncident avec la médiane, sa valeur significative devient douteuse.

La médiane est un paramètre très utile dans bon nombre d'analyses statistiques notamment dans la théorie des erreurs de mesure.

1.2.7 Mode

Le mode, symbolisé par *mod*, est la valeur qui possède l'effectif le plus élevé. Le mode peut ne pas exister, et s'il existe, il peut ne pas être unique. Le mode n'est valable, pour être un bon indicateur du centre du jeu de données, que lorsqu'un seul effectif domine.

Lorsque l'on considère un caractère quantitatif, ni les moyennes ni la médiane ne s'appliquent. On utilise alors le mode comme mesure de tendance centrale. Dans le cas d'un caractère quantitatif discret, le mode peut être trouvé immédiatement au vu des effectifs de chaque valeur. Si la variable est continue, et si les données sont regroupées en classes, on parle plutôt de classe modale : la classe ayant l'effectif le plus élevé.

1.2.8 Conclusion

La moyenne est le paramètre le plus utilisé. Il prend en compte chaque valeur du jeu de données. Par contre, ce paramètre a l'inconvénient d'être sensible aux valeurs aberrantes.

La médiane a pour avantage d'être peu sensible aux valeurs extrêmes qui peuvent ne pas être fiables. Dans certains cas, elle peut être beaucoup plus pertinente qu'une moyenne. Cependant, elle a le très gros inconvénient de mal se prêter aux calculs.

Le mode indique une seule valeur de la distribution, celle qui a l'effectif le plus élevé.

1.3 Mesures de dispersion

Les paramètres de tendance centrale ne suffisent pas à caractériser une série statistique. Dans certains cas, les données sont resserrées autour de leur mesure de tendance centrale, dans d'autres cas, elles sont dispersées sur une large étendue. On parle alors de faible dispersion ou de forte dispersion. Dans ce paragraphe, nous décrivons comment caractériser la dispersion à l'aide d'un nombre que l'on appelle mesure de dispersion. Il en existe plusieurs :

- la variance
- l'écart-type
- l'écart moyen
- l'écart médian

1.3.1 Variance

La **variance** d'un jeu de données x_1, x_2, ...,x_n, notée Var(x) est la somme des carrés des écarts des valeurs à leur moyenne arithmétique $\overline{x}$. La variance est définie par l'égalité :

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

La formule peut se simplifier en la formule équivalente suivante :

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

Exemple 8 *La moyenne arithmétique des cinq nombres :*

$$8, 10, 9, 12, 13$$

est égale à 10,4. La variance est donnée par :

$$Var(x) = \frac{8^2 + 10^2 + 9^2 + 12^2 + 13^2}{5} - 10,4^2 = \frac{558}{5} - 108,16 = 3,44$$

La variance a toujours une valeur positive ou nulle. La variance est nulle si toutes les valeurs sont identiques et donc égales à leur moyenne arithmétique. La variance est d'autant plus élevée que les valeurs sont fortement dispersées autour de leur moyenne arithmétique. Au contraire, une variance faible signifie une faible dispersion des valeurs.

1.3.2 Ecart-type

L'écart-type d'un jeu de données x_1, x_2, ...,x_n, noté σ est égal à la racine carrée de sa variance :

$$\sigma = \sqrt{Var(x)}$$

Exemple 9 *Reprenons le jeu de cinq valeurs de l'exemple 8 :*

$$8, 10, 9, 12, 13$$

dont on avait obtenu une variance égale à 3,44. L'écart-type est alors donné par :

$$\sigma = \sqrt{3,44} \simeq 1,855$$

L'écart-type exprime la même caractéristique que la variance mais est donné dans la même unité de mesure que les valeurs du jeu de données.

L'écart-type est l'indice de dispersion le plus utilisé. Comme pour la variance, une faible valeur de σ indique une accumulation forte des valeurs autour de la moyenne, une valeur grande, un étalement important. L'écart-type est un élément utile dans la définition d'autres constantes statistiques et dans la comparaison de celles-ci entre elles.

Pour les distributions normales, c'est à dire réparties suivant une loi de Gauss, on peut montrer que :

* 68,27% des valeurs sont comprises entre $\overline{x}$ - σ et $\overline{x}$ + σ,
* 95,45% des valeurs sont comprises entre $\overline{x}$ - 2σ et $\overline{x}$ + 2σ,
* 99,73% des valeurs sont comprises entre $\overline{x}$ - 3σ et $\overline{x}$ + 3σ.

1.3.3 Ecart moyen

L'**écart moyen** d'un jeu de données x_1, x_2, ...,x_n, noté $\overline{E}$ est égal à la moyenne arithmétique des écarts de ces valeurs par rapport à leur moyenne arithmétique $\overline{x}$. L'**écart moyen** est donné par la formule :

$$\overline{E} = \frac{1}{n}.\sum_{i=1}^{n} |x_i - \overline{x}|$$

Exemple 10 *Reprenons le jeu de cinq valeurs de l'exemple 8 :*

$$8, 10, 9, 12, 13$$

dont on avait obtenu une moyenne égale à 10,4. L'écart moyen est alors donné par :

$$\overline{E} = \frac{1}{5}.(|8-10,4|+|10-10,4|+|9-10,4|+|12-10,4|+|13-10,4|) = \frac{8,4}{5} = 1,68$$

L'écart moyen exprime l'ordre de grandeur des déviations autour de la moyenne arithmétique. L'écart moyen n'est pas une mesure de dispersion très importante, quoiqu'il soit souvent employé dans les séries de mesures de laboratoire.

1.3.4 Ecart médian

L'écart médian noté E_m se calcule comme l'écart moyen, mais à partir de la médiane *med* :

$$E_m = \frac{1}{n}.\sum_{i=1}^{n} |x_i - med|$$

Exemple 11 *Reprenons le jeu de cinq valeurs de l'exemple 8 :*

$$8, 10, 9, 12, 13$$

On obtient la médiane en choisissant la valeur milieu de la série de valeurs ordonnée :

$$8 \leq 9 \leq 10 \leq 12 \leq 13$$

La médiane est donc égale à 10 et l'écart médian est donné par :

$$E_m = \frac{1}{5}.(|8-10| + |10-10| + |9-10| + |12-10| + |13-10|) = \frac{8}{5} = 1,6$$

L'écart médian exprime l'ordre de grandeur des déviations autour de la médiane.