Partie 3
Les statistiques

Chapitre 3-1
Statistiques

1. Objectif du chapitre

Les statistiques regroupent un ensemble de méthodes dédiées à l'échantillonnage de données ainsi qu'à leur analyse afin de tirer des conclusions et de comprendre les phénomènes sous-jacents à ces données. Ces méthodes statistiques font partie intégrante de la Data Science.

Il est quasiment impossible d'aborder l'ensemble des méthodes statistiques en un seul ouvrage vu leur diversité. Il existe plusieurs livres qui traitent des statistiques. L'objectif de ce chapitre est double : le premier est la présentation des outils statistiques élémentaires que tout Data Scientist devrait connaître, et le deuxième objectif est d'attirer l'attention du lecteur sur l'intérêt des statistiques et leur relation avec la Data Science. Ainsi, nous allons porter une attention particulière à la partie inférentielle des statistiques.

À la fin de ce chapitre, le lecteur aura abordé :

– les statistiques descriptives,
– les lois de probabilité,
– la loi normale et la loi normale centrée réduite,
– le principe de l'échantillonnage,
– le théorème central limite,
– l'estimation ponctuelle,

– l'estimation par intervalle de confiance,
– les tests d'hypothèses,
– le paradoxe de Simpson,
– les séries temporelles.

2. Les statistiques descriptives

Les statistiques descriptives permettent de résumer un ensemble de données de manière concise. Avec les statistiques descriptives, et pour un échantillon de valeurs $E = \{x_1, x_2, \dots x_n\}$, nous pouvons calculer certains paramètres afin de cerner la nature de la distribution associée aux valeurs x_i.

Ainsi, nous distinguons deux types de paramètres que nous pouvons calculer sur une série statistique de type quantitative : les paramètres de position et les paramètres de dispersion présentés dans les sous-sections suivantes.

2.1 Paramètres de position

Les paramètres de position permettent d'avoir une idée précise sur la nature du domaine de définition d'un ensemble de données. Ces paramètres de position, également appelés indicateurs de position, sont des nombres réels utilisés comme référence pour un ensemble $E = \{x_1, x_2, \dots x_n\}$.

2.1.1 La moyenne

La moyenne $\overline{X}$ associée à une série de valeurs $S = (x_1, x_2, \dots, x_n)$ se calcule comme suit $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

La moyenne ainsi calculée correspond à la moyenne arithmétique. Il existe d'autres types de moyennes telles que la moyenne harmonique, la moyenne quadratique ou encore la moyenne géométrique. Généralement, en statistiques, la moyenne utilisée est la moyenne arithmétique.

Remarque

Si la série S est un échantillon issu d'une population plus grande, alors il ne faut pas confondre la moyenne $\overline{X}$ calculée sur cet échantillon avec la moyenne de la population ! Dans la suite de ce chapitre, nous allons revenir sur la relation entre la moyenne d'un échantillon et la moyenne de la population.

2.1.2 Le mode

Le mode d'une série de valeurs est tout simplement la valeur qui apparaît le plus fréquemment.

Par exemple, dans la série de valeurs $S=(1, 2, 5, 2, 5, 5, 6, 8, 5, 9, 5)$, on dira que le mode est la valeur 5, car c'est bien cette valeur qui apparaît avec le plus d'occurrences. La valeur 5 apparaît cinq fois, la valeur 2 deux fois, puis les autres valeurs apparaissent une fois chacune.

La série S de notre exemple est dite unimodale, car elle dispose d'un seul mode. La série $S^1=(1, 2, 5, 2, 5, 5, 2, 2, 5, 2, 5)$ est dite bimodale, car elle dispose de deux modes, à savoir le mode 5 et le mode 2.

2.1.3 La médiane

La médiane associée à une série de valeurs rangée dans l'ordre croissant $S = (x_1, x_2, \ldots, x_n)$ avec $x_i < x_{i+1}\ \forall\, i$ est une valeur qui partage toutes les valeurs de S en deux groupes de valeurs de taille égale. La valeur de la médiane peut ou pas faire partie de S.

Si les valeurs x_i de S sont des valeurs discrètes, alors la médiane se calcule comme suit :

– Si la taille n de la série S est impaire, alors nous pouvons écrire $n=2p+1$. Comme les valeurs de S sont ordonnées, la médiane est la $(p+1)^{\text{ième}}$ valeur. Dans ce cas, la médiane fait partie de la série S.

  Par exemple, pour la série $S=(1, 2, 5, 8, 15, 55, 68, 82, 125, 199, 500)$, la médiane est la valeur 55, alors que la moyenne est égale à 96,36.

– Si la taille n de la série S est paire, alors nous pouvons écrire $n=2p$. Comme les valeurs de S sont ordonnées, la médiane est la moyenne entre la $(p+1)^{\text{ième}}$ valeur et la $p^{\text{ième}}$ valeur. Dans ce cas, la médiane ne fait pas partie de la série S.

  Par exemple, pour la série $S=(1,2,5,8,15,55,68,82,125,199)$, la médiane est la valeur $\frac{15+55}{2} = 35$ et la moyenne est égale à 56.

Remarquez que dans des deux derniers exemples, les différences entre les moyennes et les médianes étaient importantes !

Si les valeurs x_i de S sont des valeurs continues, alors la médiane correspond à la valeur centrale que nous pouvons calculer par interpolation linéaire du centre des effectifs cumulés sur les x_i.

Par exemple, soit le tableau récapitulatif des notes obtenues par un groupe de 100 étudiants :

Notes	Effectifs
[0;5[	15
[5;7[	30
[7;11[	20
[11;16[	25
[16;20[	10
Total	100

À partir de ce tableau, nous allons procéder au calcul des effectifs cumulés comme suit :

Notes	Effectifs	Effectifs cumulés
[0;5[	15	15
[5;7[	30	45
[7;11[	20	65

Notes	Effectifs	Effectifs cumulés
[11;16[	25	90
[16;20[	10	100

D'après la colonne des effectifs cumulés, nous avons :

– 15 étudiants qui ont une note en dessous de 5.

– 45 étudiants qui ont une note en dessous de 7.

– 65 étudiants qui ont une note en dessous de 11.

– 90 étudiants qui ont une note en dessous de 16.

– 100 étudiants qui ont une note en dessous de 20.

Le nombre total des étudiants est égal à 100, donc la moitié est égale à 50 étudiants. L'intervalle sur l'axe des effectifs cumulés où se situe la médiane est l'intervalle [45;65], puisque 50 appartient à cet intervalle.

Nous pouvons calculer la valeur de la médiane par interpolation linéaire comme sur la figure 6-1 suivante :

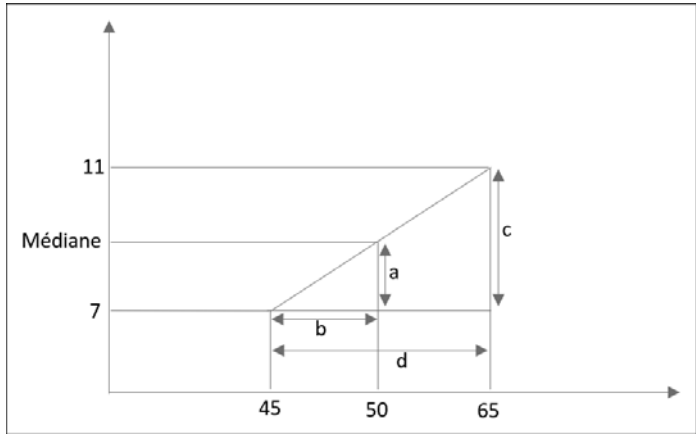

Figure 6-1 : Calcul de la médiane par interpolation linéaire du centre des effectifs cumulés

Grâce au théorème de Thalès, nous savons que $\frac{a}{b} = \frac{c}{d}$ (voir la figure précédente).

a = médiane - 7.

b = 50 - 45 = 5.

c = 11 - 7 = 4.

d = 65 - 45=20.

En procédant au remplacement de a, b, c et d par leurs valeurs respectives, on obtient $\frac{médiane - 7}{5} = \frac{4}{20}$.

Donc la médiane est égale à 8.

2.1.4 Les quartiles

Pour une série de valeurs rangées dans l'ordre croissant $S = (x_1, x_2, \ldots, x_n)$ avec $x_i < x_{i+1} \ \forall \ i$, les quartiles sont définis par trois valeurs qui partagent les valeurs de la série S en quatre groupes de valeurs de même taille. Ces trois valeurs sont définies comme suit :

– Le deuxième quartile correspond à la valeur de la médiane définie ci-dessus.

– Le premier quartile partage les valeurs de S situées entre la première valeur x_1 et la médiane en deux groupes de valeurs de même taille. En d'autres termes, le premier quartile est la médiane de la série S_1 constituée des valeurs entre x_1 et la médiane.

– De même que pour le premier quartile, le troisième quartile partage les valeurs de S situées entre la médiane et la dernière valeur xn en deux groupes de valeurs de même taille. En d'autres termes, le troisième quartile est la médiane de la série S_2 constituée des valeurs entre la médiane et x_n.

2.2 Paramètres de dispersion

Les paramètres de dispersion permettent de comprendre la variabilité associée à une série de données quantitatives.

2.2.1 La variance

La variance associée à un échantillon de données est une mesure qui nous renseigne sur la moyenne des carrés des écarts à la moyenne. On peut également dire que la variance est une valeur qui nous donne une idée sur la dispersion d'un ensemble de valeurs autour de leur moyenne.

Pour une série statistique $X = (x_1, x_2, \dots, x_n)$, la variance *Var(X)* se calcule comme suit : $Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2$.

Avec $\overline{X}$ la moyenne de la série X et qui, pour rappel, se calcule comme suit $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$

À partir de la formule de *Var(X)*, on voit bien que la variance est une distance moyenne au carré qui sépare chaque élément x_i de sa moyenne $\overline{X}$.

2.2.2 Calcul de la variance avec la formule de Koenig

Le calcul de la variance *Var(X)* avec la formule vue précédemment peut être compliqué à réaliser sans l'aide d'un logiciel. La formule de Koenig est une simplification de la formule précédente et qui permet de calculer la variance comme suit : $Var(X) = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \overline{X}^2$.

Avec cette formule, la variance est la différence entre la moyenne des carrés des x_i et le carré de la moyenne x_i.

Avec la formule de Koenig, le calcul de la variance devient plus facile et une simple calculatrice serait largement suffisante.

2.2.3 L'écart-type

Nous avons vu que la variance est la distance moyenne au carré d'un ensemble de valeurs par rapport à leur moyenne. L'écart-type σ est tout simplement la distance moyenne, ou la distance type, qui sépare les valeurs de leur moyenne et il se calcule à partir de la variance comme suit : $\sigma = \sqrt{Var(X)}$.

2.2.4 L'écart interquartile

L'écart interquartile correspond à la distance entre le troisième et le premier quartile.

Par exemple, pour la série $S=(1, 2, 5, 8, 15, 55, 68, 82,125, 199, 500)$, les quartiles sont les suivants :

– Le premier quartile est égal à 5.

– Le deuxième quartile est égal à 55, ce qui est également la médiane.

– Le troisième quartile est égal à 125.

Pour cette série S, l'espace interquartile est donc égal à $125\text{-}5 = 120$.

La valeur de l'écart interquartile, aussi appelé l'espace interquartile ou bien l'étendue interquartile, nous renseigne sur l'ampleur de la dispersion des valeurs d'une série.

En combinant les mesures de la moyenne, du mode, de la médiane, de la variance, de l'écart-type et de l'écart interquartile, nous obtenons un résumé de l'ensemble des valeurs étudiées. Toutes ces mesures peuvent être calculées sur un ensemble de valeurs associées à une seule variable quantitative. Dans le chapitre Analyse en composantes principales, nous aborderons l'analyse en composantes principales, qui est une méthode permettant de résumer un ensemble de données définies avec plusieurs variables.

3. Les lois de probabilité

Une loi de probabilité permet de cerner le comportement d'une variable aléatoire. Dans le domaine des probabilités, une variable aléatoire dépend du hasard. Justement, c'est le comportement de ce hasard que l'on tente de décrire avec une loi de probabilité. Avec une loi de probabilité, nous pouvons calculer la probabilité qu'une variable aléatoire soit fixée à une valeur donnée.

Par exemple, si nous considérons une variable X associée au résultat obtenu après le lancer d'un dé à six chiffres, alors cette variable X sera appelée une variable aléatoire, puisque la survenue de l'un des six chiffres est un événement aléatoire.

Si nous supposons que notre dé est parfait, c'est-à-dire que chacun des six chiffres est équiprobable avec une probabilité de $\frac{1}{6}$, alors la loi de la variable X est tout simplement $F(X) = \frac{1}{6}$.

Le choix d'une loi de probabilité est en fonction de la nature de la variable aléatoire étudiée et en fonction du phénomène associé à cette variable aléatoire. En effet, une variable aléatoire X peut être discrète ou continue et elle peut être définie dans un intervalle fini, semi-fini ou infini. Le phénomène associé à une variable aléatoire peut concerner des durées, des quantités comme le poids, des longueurs comme la taille ou toute autre mesure qui peut être observée et relative à un événement quelconque.

En résumé, une loi de probabilité décrit la manière dont sont distribuées les valeurs possibles d'une variable aléatoire X.

Il existe un nombre important de lois de probabilité. Loin d'être exhaustifs, parmi les plus usuelles nous pouvons citer :

- Loi de Bernoulli : décrit la distribution d'une variable aléatoire associée à un événement à deux issues possibles.
- Loi binomiale : décrit la distribution d'une variable aléatoire associée à un événement à deux issues possibles répété plusieurs fois avec remise, c'est-à-dire que le même événement peut survenir plusieurs fois.
- Loi de Poisson : décrit la distribution d'une variable aléatoire associée au nombre d'événements se produisant dans un intervalle de temps donné. Par exemple, le nombre de personnes supplémentaires dans une file d'attente au bout d'un temps donné.
- Loi uniforme : décrit la distribution d'une variable aléatoire associée à un ensemble d'événements équiprobables.
- Loi exponentielle : décrit la distribution d'une variable aléatoire associée au délai de la survenue d'un événement. Par exemple, le délai de l'arrivée d'une personne supplémentaire dans une file d'attente.