Éléments de compréhension des statistiques

Jeffery P.

Doctorant au Laboratoire des Sciences du Numérique de Nantes (LS2N)

2019

Le quantilage

Le quantilage est le découpage de la population en proportions égales. Il s'effectue nécessairement pour des **variables quantitatives**

Par exemple un effectif de 12 observations peut être divisé :

- En deux groupes de 6 modalités (quantiles d'ordre 6)
- En trois groupes de 4 modalités (quantiles d'ordre 4 nommés **quartiles**)
- ...

$\rightarrow$ Les quantiles ne font pas nécessairement référence à des groupes entiers, c'est même rarement le cas. On peut vouloir découper un effectif de 12 en 10 groupes (quantiles d'ordre 10 nommés **déciles**)

Le quantilage

Les **quantiles** sont des modalités (possiblement non observées) qui correspondent à des séparation de l'ensemble des modalités observées.

Un quantile est toujours précisé avec deux indices :

- p : le nombre total de groupes
- r : l'index de la séparation avec $1 \leq r \leq p-1$

Si on note Q_p^r le quantile r d'ordre p pour une variable, alors :

Il y a une proportion au moins égale à $\frac{r}{p}$ d'observation $\leq Q_p^r$ Et une proportion au moins égale à $1 - \frac{r}{p}$ d'observation $\geq Q_p^r$

Le quantilage

Remarques

- L'ordre p d'un quantile correspond au nombre de groupes que l'on fait
- Il y a toujours $p-1$ quantiles d'ordre p
- Les quantiles sont ordonnés $Q_p^1 < Q_p^2 < \cdots < Q_p^{p-1}$
- Bien retenir que :
- les quantiles d'ordre 4 sont appelés **quartiles** (Q_4)
- les quantiles d'ordre 10 sont nommée **déciles** (Q_{10})
- les quantiles d'ordre 100 sont nommés **centiles** (Q_{100})
- Le deuxième quartile correspond à la médiane ($m_e = Q_4^2$)

Le quantilage en pratique

En pratique, on procède un peu comme pour la médiane. Il y a deux étapes :

1. On commence par déterminer le rang du quantile recherché avec la formule suivante :

$$rang(Q_p^r) = r \times \frac{\mathrm{N} + 1}{p}$$

2. On recherche la modalité associé au rang (c'est là que ça peut se corser !)

La recherche d'une modalité associé à un rang

Cette recherche peut arriver dans deux grands cas souvent fréquents :

- On dispose des données brutes mais le rang n'est pas entier
- Les données ont été regroupées en classe

On utilise alors la méthode d'**interpolation linéaire**

L'interpolation linéaire sur données brutes

Facile ! Imaginons qu'on obtienne un rang égal à $2,7$

Supposons d'autre part que l'on dispose des 3 observations ordonnées suivantes : 4, 10, 18 . . .

- Il ne serait pas logique de privilégier la modalité 10 plutôt que 18 (et inversement) $\rightarrow$ on imagine bien que la modalité associée au rang 2.7 se situerait entre 10 et 18 mais plus prêt de 18 quand même. . .

On utilise alors le calcul d'interpolation suivant :

$$10 + (2,7 - 2) \times (18 - 10) = 15,6$$

L'interpolation linéaire sur données brutes

Le calcul énoncé :

$$10 + (2,7 - 2) \times (18 - 10) = 15,6$$

Correspond à la formule générale :

$$V_i + (\tilde{i} - i) \times (V_{i+1} - V_i)$$

où :

- V_i et V_{i+1} : sont les valeurs des rangs entiers i et $i + 1$
- $\tilde{i}$: est le rang du quantile recherché tel que $i < \tilde{i} < i + 1$

Remarque :

Bien que le résultat obtenu soit plus précis avec l'interpolation sur données brutes, on ne fait pas toujours ce calcul. . .souvent, on se contente de faire la moyenne entre V_i et V_{i+1}

L'interpolation linéaire sur données regroupées en classe

Supposons maintenant que l'on recherche le quantile associé au rang 8 et que l'on ait les observations suivantes :

Poids	[68 ; 72[	[72 ; 76[	[76 ; 80[
n_i	5	4	7
n_{ic}	5	9	16

- On sait d'après la ligne des effectifs cumulés (n_{ic}) que l'individu 8 se situe dans la classe [72; 76[mais où exactement ?

L'interpolation linéaire sur données regroupées en classe

Poids	[68 ; 72[	[72 ; 76[	[76 ; 80[
n_i	5	4	7
n_{ic}	5	9	12

$\rightarrow$ L'idée est de se dire que cette classe contient 4 observations et que ces dernières sont réparties *uniformément* de manière ordonée entre 72kg et 76kg. Cela signifie qu'ils ne sont pas tous proches de 72kg, ni tous proches de 76kg mais qu'au contraire, il y en a proches de 72kg d'autres proches de 76kg et d'autres plus proches de 74kg (centre de la classe).

En suivant cette idée l'individu 8 serait alors l'avant dernière valeur de cette classe, donc relativement proche de 76kg. . .

L'interpolation linéaire sur données regroupées en classe

Poids	[68 ; 72[	[72 ; 76[	[76 ; 80[
n_i	5	4	7
n_{ic}	5	9	12
n'_{ic}	1, 2, 3, 4, 5	6, 7, 8, 9	...

On utilise alors le calcul suivant :

$$72 + (8 - 5) \times \frac{(76 - 72)}{4} = 75kg$$

- On part de la borne inférieure (B^i_{inf}) de la classe, ici égale à 72
- On cherche le 8ème individu en sachant qu'il y a 5 observations $\leq B^i_{inf}$
- L'amplitude (a_i) de la classe est initialement divisée en 4 (n_i) proportions

L'interpolation linéaire sur données regroupées en classe

Avec les notations, le calcul :

$$72 + (8 - 5) \times \frac{(76 - 72)}{4}$$

Est équivalent à :

$$B^i_{inf} + \left(\tilde{i} - n_{(i-1)c}\right) \times \frac{(B^i_{sup} - B^i_{inf})}{n_i} = B^i_{inf} + \left(\tilde{i} - n_{(i-1)c}\right) \times \frac{a_i}{n_i}$$

Où :

- $\tilde{i}$ est le rang du quantile recherché tel que $n_{(i-1)c} < \tilde{i} < n_{ic}$
- n_i est l'effectif de la classe i
- B^i_{inf} et B^i_{sup} sont les bornes de la classe i
- a_i est l'amplitude de la classe i
- $n_{(i-1)c}$ est l'effectif cumulé de la classe inférieure

L'interpolation linéaire sur données regroupées en classe : remarques

- Le calcul précédent est intuitif (non ?)
- Le calcul reste valable peu importe la nature du rang recherché, qu'il soit entier ou décimal
- L'interpolation linéaire nécessite les deux types de valeur d'effectif (non cumulé et cumulé)

Digression

- Le calcul de la médiane est une interpolation linéaire