L'ANALYSE DE DONNÉES PONDÉRÉES

. QU'EST-CE QU'UN POIDS ?

. L'USAGE DE PONDÉRATION DANS L'ANALYSE

. ESTIMER LA MOYENNE . ESTIMER LA VARIANCE DE LA POPULATION

. CRÉER DES DISTRIBUTIONS DE FRÉQUENCES

. PROPORTIONS DE LA POPULATION . ERREURS TYPES

À LA FIN DE CE CHAPITRE VOUS DEVEZ ÊTRE CAPABLE DE :

- COMPRENDRE CE QUE SONT LES POIDS DANS UNE ENQUÊTE
- COMPRENDRE POURQUOI ON LES UTILISE
- APPLIQUER LES POIDS POUR ESTIMER LES PLANS DE POPULATION, LES VARIANCES ET LES FRÉQUENCES

CHAPITRE 7

L'ANALYSE DE DONNÉES PONDÉRÉES

Introduction

une analyse différente pour les données de l'échantillon

J'usqu'à présent, toutes les techniques d'analyse ont traité de la production de moyennes et de variances à partir des données brutes. Mais que fait-on quand on a des données pondérées ? Dans ce chapitre on verra d'abord ce qu'est un poids, et ensuite comment on peut calculer des statistiques à partir de données pondérées. Enfin, on étudiera l'erreur-type et les différences entre l'erreur-type et l'écart-type.

Qu'est-ce qu'un poids ?

l'échantillon représente la population

On a vu au chapitre 1 qu'il vaut souvent mieux prendre un échantillon plutôt qu'étudier la population entière. À la fin de l'enquête on doit souvent faire des estimations sur la population en utilisant l'échantillon.

Par exemple, on vient de terminer une enquête par échantillonage qui demande combien de shells de kava ont été bus la semaine dernière. On peut facilement appliquer les techniques des six premiers chapitres pour obtenir la moyenne, la médianne et la variance **de cet échantillon**. Mais comment peut-on estimer le montant total d'argent dépensé pour le kava à Vanuatu, ou le nombre total de personnes qui ont bu du kava ?

utiliser les poids pour les estimations de la population

Pour produire des estimations sur la population totale, on alloue un poids à chaque personne de l'échantillon.

Un poids indique combien de personnes dans la population représente une personne de l'échantillon.

Par exemple, dans l'enquête sur le kava, si on interroge 1.500 personnes sur un total de 150.000, alors chaque personne de l'enquête recevra une pondération de $^{150,000}/_{1,500}$ =100. Chaque personne de l'échantillon représentera ainsi 100 personnes de la population totale.

les poids dépendent de la sélection de l'échantillon

Dans l'enquête sur le kava, chaque personne a reçu la même pondération, mais ça n'est pas toujours le cas. Les poids de chaque unité statistique dépendront de la manière dont les unités auront été sélectionnées, et du nombre d'unités qui n'auront pas répondu. Cependant, ce qui suit devrait toujours se répéter :

3 'règles' pour les poids

1. Si on aditionne tous les poids le total est égal au nombre d'unités dans la population;
2. le poids d'une unité indique le nombre d'unités qu'elle représente dans une population; et
3. les poids sont toujours supérieurs à 1. Si vous êtes choisi dans une enquête, alors vous devez au moins représenter une personne.....vous-même !

les poids ne sont pas nécessairement des chiffres entiers

Bien que ça puisse paraître étrange, les poids ne sont pas nécessairement des chiffres entiers. C'est-à-dire qu'on peut avoir une pondération de 4,53 ou de 88,763. Bien que ça paraisse absurde qu'un individu représente 88,763 personnes, c'est simplement notre "meilleure estimation" de combien de personnes auront des caractéristiques similaires dans la population.

L'usage de la pondération dans l'analyse

les valeurs de l'échantillon donnent une estimation de la population

La grande différence quand on analyse des données pondérées, c'est qu'on ne connait pas la valeur exacte de la population. Si on extrait la moyenne pour un recensement, on connait chaque valeur de la population, alors le résultat sera exact. Mais si on prend un échantillon on ne fait pas des recherches sur toute la population, juste sur quelques personnes ou quelques unités dont on pense qu'elles représentent la population.

Parfois l'échantillon sélectionné fournira une bonne représentation de la population, ou une représentation précise. D'autres fois il se peut que la représentation ne soit pas si bonne. Il est important de sélectionner l'échantillon de manière à représenter au mieux la population.

ASTUCE

QUAND ON CALCULE LA MOYENNE, LA MÉDIANE OU LA VARIANCE EN UTILISANT DES DONNÉES PONDÉRÉES, ON DIT QU'ON FAIT UNE ESTIMATION DE LA VALeUR DE LA POPULATION. C'EST-À-DIRE QU'ON FAIT UNE ESTIMaTION DE LA MOYENNE, DE LA MÉDIANE OU DE LA VARIANCE DE LA POPULATION.

au mieux une estimation

Si on travaille sur des données de recensement, alors on peut dire qu'on calcule la moyenne, mais si on utilise un échantillon pour calculer la moyenne d'une population, alors on fait **une estimation de la moyenne**.

l'erreur-type est la mesure de la représentation de l'échantillon

Ces estimations des valeurs de la population sont l'objet d'**erreurs-types**. Ce sont des erreurs, parce qu'on a un échantillon, et pas la population entière. Si on a choisi correctement l'échantillon, les

estimations seront très proches, ou mêmes identiques aux valeurs réelles de la population. Dans ce cas l'erreur d'échantillonage est très faible. Rappelez-vous que les estimations de population sont aussi affectées par des erreurs non dues à l'échantillonage, qui sont très difficiles à calculer.

on suppose que les poids ont été définis pour vous

Lorsqu'on travaille sur des données pondérées, les poids sont fournis avec les données. Les poids sont déterminés par un certain nombre de facteurs, comme la manière dont on a choisi l'échantillon, les taux de réponses, et d'autres critères spécifiés en général par un statisticien mathématique. Ne vous tracassez pas à calculer les poids, c'est souvent complexe et hors du domaine de ce cours.

Estimer la moyenne

formule

Pour estimer la moyenne à partir de données pondérées, on applique la formule :

Estimer la moyenne de la population à partir de l'échantillon =

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

où w_i est le poids de chaque unité x_I est la valeur de la variable

exemple

Par exemple, si on obtient les résultats suivants d'une enquête :

Tableau 7.1 Estimer la moyenne d'une population à partir des données de l'échantillon

Valeur (x_i)	Poids (w_i)	Poids × Valeur ($w_i x_i$)
5	10	50
3	20	60
6	15	90
7	10	70
9	5	45
Total	**60**	**315**

Source : données fictives

estimation de la moyenne

Alors, l'estimation de la moyenne de la population est :

$$= \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad = \frac{315}{60} \qquad = 5.25$$

Estimer la variance de la population

estimer la variance

Pour des données pondérées, estimer la variance de la population paraît une tâche compliquée, car les poids sont intégrés dans les calculs :

$$\textbf{Estimation de la variance} = \frac{\sum_{i=1}^{n} w_i x_i^2 - \frac{(\sum_{i=1}^{n} w_i x_i)^2}{(\sum_{i=1}^{n} w_i)}}{(\sum_{i=1}^{n} w_i) - 1}$$

exemple

On peut utiliser les mêmes données d'enquête que celles avec lesquelles on a estimé la variance de la population. Bien sûr, il est plus facile de produire des estimations en utilisant un ordinateur que de faire les calculs manuellement en utilisant des formules.

Tableau 7.2 Estimer la variance d'une population avec des données pondérées

Valeur (x_i)	Poids (w_i)	Poids × Valeur ($w_i x_i$)	Poids × Valeur au carré ($w_i x_i^2$)
5	10	50	250
3	20	60	180
6	15	90	540
7	10	70	490
9	5	45	405
Total	60	315	1,865

Source : données fictives

estimation de la variance

Alors, l'estimation de la variance de la population sera :

$$\textbf{Estimation de la variance} = \frac{\sum_{i=1}^{n} w_i x_i^2 - \frac{(\sum_{i=1}^{n} w_i x_i)^2}{(\sum_{i=1}^{n} w_i)}}{(\sum_{i=1}^{n} w_i) - 1} = \frac{1.865 - \frac{(315)^2}{60}}{60 - 1} = \frac{211.25}{59} = 3{,}58$$

Créer des distributions de fréquences

additionner les poids

On peut aussi créer des distributions de fréquences de la population en utilisant des poids. Le procédé ressemble beaucoup à la création des distributions de fréquences du chapitre 3. La différence majeure est que la fréquence est maintenant la somme des poids de l'intervalle de classe.

exemple

Par exemple, on a les données suivantes :

Tableau 7.3 Créer des distributions de fréquences à partir de données d'un échantillon

Variable (x_i)	Poids (w_i)
2	5
3	6
3	6
4	5
4	5
4	5
5	4
5	7
7	8
8	5
9	4

Source : données fictives

intervalles de classes

On utilisera les classes : 2–3, 4–5, 6–7, 8–9

fréquence

Les fréquences de notre population seront :

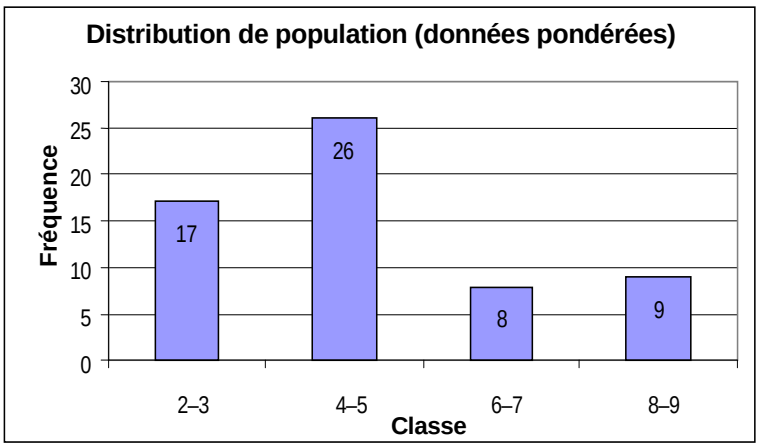
Tableau 7.4 Créer des distributions de fréquences à partir de données d'un échantillon

Classe	Poids dans les classes (w_i)	Fréquence (Σw_i)
2–3	5 + 6 + 6	17
4–5	5 + 5 + 5 + 4 + 7	26
6–7	8	8
8–9	5 + 4	9

Source : données fictives

diagrammes

Une fois qu'on a la distribution de fréquences, on met les données en diagramme de la même manière qu'on met en diagramme des données non pondérées, en utilisant les mêmes repères pour les données quantitatives et les données qualitatives.



Proportions de la population

estimer les proportions de la population

On veut souvent connaître le nombre de personnes ayant telle ou telle caractéristique. Comme pour de nombreuses mesures montrées dans ce chapitre, on utilise la somme des poids plutôt que la fréquence de la valeur ou la variable.

Par exemple, on peut vouloir estimer le nombre de personnes qui boivent du kava dans la population. On estime cette proportion de la population à partir de l'échantillon en appliquant la formule :

formule

Estimation d'une proportion de population (p) =

$$\frac{\sum w_{\text{i (avec caractéristique)}}}{\sum w_{i\text{ (toutes les unités)}}}$$

On prend la somme des poids des variables qui ont cette caractéristique, et on la divise par la somme de tous les poids dans l'échantillon.

exemple

On a l'échantillon suivant de personnes, et on leur demande s'ils ont bu du kava la veille :

Tableau 7.5 Échantillon de personnes - ont-ils bu du kava la veille ?

Personne	Poids	A bu du kava
Jean	10	Oui
Simon	9	Oui
Howard	7	Non
Jacques	8	Oui
Theto	10	Non
Jean-Marc	9	Non
Meryline	4	Oui

Source : données fictives

travail

L'estimation de la proportion de population qui a bu du kava la veille sera :

$$\text{Estimation du nombre de buveurs de kava } (p) = \frac{\sum w_{\text{i (avec caractéristique)}}}{\sum w_{i\text{ (toutes les unités)}}}$$

$$= \frac{(10+9+8+4)}{(10+9+7+8+10+9+4)} \qquad = \frac{31}{57} \quad = 0.5438$$

résultat

On peut alors dire qu'on estime à 54,38% la proportion de population qui a bu du kava la nuit dernière.

Erreurs-types

les données de l'échantillon estiment la population

Quand on mène une enquête par échantillonage on introduit une autre erreur, du fait d'échantilloner uniquement un sous-ensemble de population. La théorie relative à ce type d'erreur est bien développée, et la capacité à calculer l'erreur introduite par cette méthode de collecte est une des caractéristiques intéressantes des enquêtes par échantillonage.

La quantité calculée en général pour mesurer l'exactitude de l'estimation d'un échantillon est l'**erreur-type** de l'estimation. L'erreur-type d'une estimation nous permet d'établir certaines conjectures sur l'estimation, si certaines conditions sont réunies (ces conditions ne sont pas restrictives tant que la taille de l'échantillon n'est pas trop petite). On peut dire qu'on est certain à 95% que la valeur réelle de ce qu'on est en train d'essayer de mesurer se trouvera entre deux erreurs-types de notre estimation d'échantillon.

minimiser les erreurs non dues à l'échantillonage

Rappelez-vous que l'erreur due à l'échantillonage représente seulement une des composantes de l'erreur totale de l'enquête. On peut introduire des erreurs de multiples façons dans des enquêtes, autres que l'utilisation de méthodes d'échantillonage. L'erreur de dénombrement sur le terrain, l'erreur du répondant, les défauts de conception du questionnaire et les erreurs de dépouillement sont autant de sources d'erreurs pour l'enquête toute entière (notez que ces erreurs sont indépendantes du processus de l'échantillonage, et se produiraient même dans le cadre d'un recensement).

ASTUCE

Le plus important C'est DE minimiser les erreurs non dues à l'échantillonage, et DE s'assurer que les erreurs qui introduisent une distortion systématique DES résultats de l'enquête sont évitées.

quelle est l'exactitude des données de l'échantillon ?

L'erreur-type ne doit pas être confondue avec l'écart-type.

- ☆ **Erreur-type :** mesure de l'exactitude d'une estimation.
- ☆ **Écart-type :** mesure de l'étendue des valeurs dans une population.

Plus grande est l'erreur-type, **moins bonne** est l'estimation.

formule

Pour calculer l'erreur-type de l'estimation de la moyenne d'une population à partir d'un échantillon simple pris au hasard, on applique la formule :

$$SE \text{ (estimation de la moyenne de la population)} = \sqrt{\{(1-(\frac{n}{N}))*\frac{s^2}{n}\}}$$

$$SE \text{ (estimation de la proportion de la population)} = \sqrt{\{(1-(\frac{n}{N}))*\frac{(p(1-p))}{n}\}}$$

Où s^2 est l'estimation de la variance de la population

n est la taille de l'échantillon

N est le nombre d'unités dans la population

p est l'estimation de la proportion de la population

exemple du kava

Dans l'exemple du kava, le proportion est 0,5438, la taille d'échantillon 7, et la population 57. L'erreur-type est alors :

***SE* (estimation de la proportion de la population)** =

$$= \sqrt{\{(1-(\frac{7}{57}))*\frac{(0.5438(1-0.5438))}{7}\}} \qquad = 0.1763$$

On dit que l'erreur-type de la proportion de l'échantillon est 0,1763.

zéro erreurs dans un recensement

On note aussi que si on mène un recensement, la taille de l'échantillon égale la taille de la population (n = N). L'estimation de la valeur de la population est juste parce qu'on a inclus tous les individus de la population dans l'échantillon.

On peut aussi voir dans la formule que quand n = N, alors :

$$(\frac{1-n}{N}) \quad = \quad (\frac{1-N}{N}) \quad = \quad 1-1 \quad = \quad 0$$

Il n'y a aucune erreur dans l'estimation.