Variables

. CONCEPTS ET DÉFINITIONS . TYPES DE DONNÉES

. TYPES DE VARIABLES QUANTITATIVES ET QUALITATIVES

. LES ÉCHELLES DE MESURE DES DONNÉES

. DONNÉES UNIDIMENSIONNELLES, BIDIMENSIONNELLES OU MULTIDIMENSIONNELLES

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

- DÉFINIR DIFFÉRENTS TYPES DE DONNÉES STATISTIQUES
- EXPLIQUER LA DIFFÉRENCE ENTRE VARIABLE CONTINUE ET VARIABLE DISCRÈTE
- COMPRENDRE LES DIFFÉRENTES ÉCHELLES DE MESURE

CHAPITRE 2

LES VARIABLES

Introduction

concepts et définitions

Dans ce chapitre on définira quelques concepts et définitions statistiques, et les grands types de données statistiques, dans le but de savoir de quelle manière les concepts et définitions utilisés pour collecter des données auront un impact sur l'analyse, et de se familiariser avec les différents types de données et les termes et concepts utilisés pour définir les données.

Concepts et définitions

première chose à prendre en compte

Avant d'entreprendre toute analyse statistique, on doit examiner les concepts et définitions utilisés pour collecter les données. On se pose un certain nombre de questions :

- ? Quelle est la **population cible** ?
- ? Quelle est la **population de l'enquête** ?
- ? Quelle est **l'unité statistique** ?
- ? Comment sont pratiquées les **observations** ?
- ? Quels sont les **standards** et **classifications** utilisés, s'il y en a ? Quelles sont les décisions de codage ?
- ? Quels sont les **corrections**, ou les contrôles de logique appliqués, s'il y en a ?

nombre d'observations

On peut aussi estimer si le nombre d'observations est suffisant pour l'analyse (par exemple, dans certains sondages certaines analyses statistiques détaillées de variables peuvent ne pas être possibles). Regardez aussi le taux de non-réponse – chapitre 3.

définitions

La **population de l'enquête**, ou le **champ d'observation** des données, peut avoir un impact non négligeable sur l'analyse que l'on peut faire, ou pas. Comme population d'enquête on peut avoir des personnes, des chefs de famille, des visiteurs, des personnes âgées de plus de 15 ans, des personnes qui travaillent ou des personnes sans emploi. On doit ensuite étudier comment et pourquoi la population de l'enquête a été définie. Par exemple :

- ? **Qu'est-ce qu'un 'visiteur'** – combien de temps une personne est-elle en 'visite' avant d'être considérée comme une personne 'résidant' dans une famille ?
- ? **Que veut dire 'employé'** – est-ce que ce terme inclue à la fois le travail rémunéré et le travail non rémunéré (par exemple l'économie de subsistence) ? Combien d'heures doit-on consacrer au travail avant qu'il soit considéré comme un 'emploi' ?

? **Que veut dire 'sans emploi'** – dans certaines collectes, pour être sans emploi une personne doit être à la recherche d'un emploi et disponible pour travailler, dans d'autres, les personnes disent simplement qu'elles sont sans emploi.

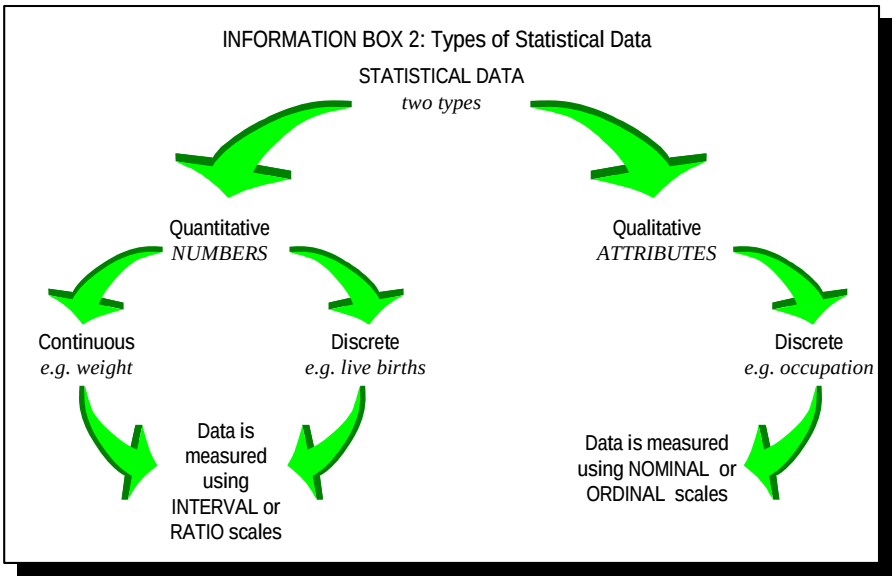
ATTENTION

BIEN SOUVENT LES DÉFINITIONS EMPLOYÉES POUR LES POPULATIONS D'ENQUÊTES SONT PLUS COMPLIQUÉES QUE VOUS NE VOUS Y ATTENDEZ. LA PLUPART DU TEMPS, LA BONNE MANIÈRE DE COMMENCER L'ÉTUDE D'UNE POPULATION D'ENQUÊTE EST D'INTERROGER LES SERVICES DE LA STATISTIQUE, OU LES PERSONNES QUI ONT COLLECTÉ LES DONNÉES.

Types de données

différents types de données

On va déterminer aussi bien les concepts et définitions que le type de donnée à collecter pour chaque variable, étant donné qu'il a également une incidence sur l'analyse. Le diagramme suivant résume la manière de catégoriser différents types de données :



définitions

Le mot 'statistique' a trois grandes définitions : le sujet, les données, et les figures résumées dérivées des données collectées. Pour éviter toute confusion entre ces trois définitions, on utilisera le terme 'donnée statistique' pour les valeurs observées lors de la collecte de l'information.

deux types de données statistiques

La donnée statistique peut être soit **quantitative**, soit **qualitative**, c'est-à-dire que ça peut être soit un chiffre, soit un attribut quelconque d'une unité particulière. Par exemple, si on doit récolter des informations sur les participants de ce cours sur leur âge et leur sexe, le jeu de données statistiques ayant trait à l'âge serait quantitatif. Il consisterait en une série de chiffres représentant les âges de différents individus répartis en années et en mois. Le jeu de données statistiques sur le sexe serait

qualitatif. Il enregistrerait ‘homme’ ou ‘femme’ pour chaque personne.

les collectes utilisent les deux types de données

On peut observer les valeurs de nombreuses variables dans une unité statistique. Certaines de ces variables peuvent être quantitatives et d’autres qualitatives. Ainsi, dans une enquête sur les employés d’une industrie, les observations pour chaque employé peuvent comporter l’âge, le sexe, la profession, le niveau d’études, les rentes ou le salaire, et le temps passé à occuper leur poste. De ces six caractéristiques, les première, cinquième et sixième sont quantitatives, alors que les deuxième, troisième et quatrième sont des variables qualitatives.

Types de variables quantitatives et qualitatives

variables quantitatives continues

Pour des raisons pratiques, on peut faire la distinction entre deux types de variables quantitatives, et on doit s’assurer que l’on sait bien quel type de variable on rencontre dans une situation donnée. Le premier type est la variable qui peut prendre n’importe quelle valeur dans une fourchette donnée. On utilise le terme **continue** pour décrire ce type de variable. Si on mesure la superficie des terres d’une ferme, la superficie sera une variable continue. On mesure la superficie en hectares (ha) et, en théorie, on pourrait avoir n’importe quelle valeur comme résultat dans la fourchette de disons 0 à 5000 ha. Les résultats 40,63 ha, 7,405 ha et 53 ha seraient tous valides. Si on plaçait toutes les superficies possibles sur un axe, le résultat serait une ligne continue.

variables quantitatives discontinues

Le deuxième type de variable quantitative est celui qui prend les valeurs telles que le nombre de travailleurs dans un établissement industriel. On dit que ce type de variable quantitative est **discontinu**. Si on mène une enquête sur les ménages et qu’on récolte des informations sur le nombre de personnes qui vivent habituellement dans le ménage, alors on sait qu’on va seulement obtenir les réponses 1, 2,3,4,5, etc. Pour cette variable, les valeurs telles que 4,2 , 1,3 ou 3,6 ne sont pas valides. Ça n’a aucun sens de parler de 4,62 personnes. Si on met toutes les valeurs possibles d’une variable discontinue sur un axe, on obtient une série de points. Dans la plupart des exemples pratiques, les variables quantitatives discontinues prennent en général les valeurs des nombres entiers.

variables qualitatives discontinues

Toutes les variables qualitatives sont discontinues. C’est parce que les variables qualitatives décrivent des attributs qui, par leur nature, sont discontinus. Dans l’exemple ci-dessus ça n’avait aucun sens de dire que 4,62 personnes vivent habituellement dans un ménage. De même, la couleur des yeux d’une personne, sa profession ou sont adresse ne peuvent qu’être discontinus.

quelques exemples de variables continues

la taille d’une personne
son poids
la température journalière maximum
l’âge d’une personne
la longueur d’une route
la pluviométrie journalière

Combien de pluie ?

quelques exemples de variables discontinues

nombre d'enfants nés par femme
nombre de chambres dans une maison
nombre de cochons possédés par un ménage
nombre d'employés dans une société
nombre d'élèves dans une école
nombre de fermes dans plusieurs unités de superficie
nombre de taxis dans une ville

Combien de chambres ?

revenus ?

Les revenus sont une variable intéressante à observer. La plupart des manuels disent que les revenus sont une variable continue. Étudiez cette variable pendant quelques instants et discutez-en lors du cours.

Les revenus sont une variable discontinue en termes de la somme que vous touchez réellement – vous pouvez compter la somme et vous n'êtes pas payé en ¼ ou ½ franc – si vous touchez un salaire convertissez-le en salaire horaire – maintenant vous aurez des ¼, des ½ ou d'autres fractions de francs.

nombre d'employés ?

Le nombre d'employés est une autre variable intéressante – une fois de plus la manière dont il est calculé détermine s'il s'agit d'une variable continue ou discontinue. Comment est-ce possible ?

Manifestement le compte des employés donne un nombre discontinu – une société pourrait avoir 15 employés à plein temps, 5 employés à temps partiel, et 2 employés occasionnels. Cela donnerait un total de 22 employés, mais que se passe-t-il si on s'intéresse uniquement aux employés à plein temps ? On pourrait former des équivalents plein temps – il y aurait maintenant 15 + 5/2 = 17,5 employés équivalents plein temps. Mais comment recenser les employés occasionnels ? Nous devons en apprendre plus sur leurs horaires de manière à pouvoir en faire des équivalents temps complet. Cela pourrait nous donner un chiffre comme 18,25764 employés équivalents plein temps – une variable continue.

quelquefois dur à distinguer

Lorsqu'on observe les variables continues et discontinues dans la vie réelle, le problème est un peu plus compliqué. Bien qu'en théorie la superficie d'une ferme puisse avoir n'importe quelle valeur, en pratique il se peut que nous soyons seulement capables de mesurer la surface la plus proche d'un hectare. Alors, nous pourrions avoir une série d'observations comme 0,6 hectare, 7,4 hectares, 3,4 hectares, 18,9 hectares, 11,4 hectares, 73,6 hectares, 80 hectares. Au premier coup d'oeil ça peut ressembler à une variable discontinue, puisque ces valeurs ont été enregistrées au dixième d'hectare près (par exemple, 7,38 hectare sera enregistré sous 7,4). On sait que toutes ces valeurs peuvent exister, mais nous réalisons que le manque d'accuité de notre système de mesure fait que les valeurs sont arrondies aussitôt qu'elles sont enregistrées.

ASTUCE

essayez de 'compter' la variable pour voir si elle est continue ou discontinue — on peut compter chaque enfant qui nait (discontinue) mais COMMENT peut-on "COMPTER' LA TAILLE ? EN GÉNÉRAL Les variables continues sont mesurées'.

en général les variables continues sont approximatives

Les variables continues sont en général approximatives et le degré d'approximation dépend largement de ce qui est observé, la manière dont c'est observé, et la précision requise. Par exemple, quand une hauteur est mesurée, c'est en général au centimètre près ou au pouce près, la mesure n'est pas très précise.

mais les données discontinues peuvent tout autant prêter à confusion

Une erreur différente peut se produire quand on observe des variables discontinues. Si on mène une enquête sur les entreprises et qu'on enregistre le nombre d'employés dans différentes sociétés on peut classer les résultats de l'enquête en groupes ou en unités (par exemple, on a 26 entreprises dans le groupe de 50 à 99 employés). Ceci semble impliquer une fourchette de valeurs possibles, continues pour les entreprises, mais on doit savoir qu'elle aura les valeurs 50, 51, 53 et jusqu'à 99. On ne pourra pas avoir une entreprise de 76,38 employés. Même si on cite une fourchette de valeurs, la variable est quand même discontinue. On doit faire attention, dans ce cas, à ne pas regarder simplement les valeurs des données enregistrées, mais aussi à examiner la variable pour déterminer si elle est continue ou discontinue.

Les échelles de mesure des données

les échelles sont utilisées pour mesurer les données

Les variables continues et discontinues sont mesurées de différentes façons. Quand on fait des observations sur les variables d'unités statistiques, les données observées peuvent être mesurées grâce à un grand nombre d'échelles. Ces échelles se classent en 'nominales' ou 'ordinales' pour les données qualitatives, et 'intervalles' ou 'échelles d'évaluation' pour les données quantitatives.

Données Qualitatives

échelle nominale

L'échelle nominale est la forme la plus primaire de mesure de données. C'est le résultat d'observations sur des variables classifiées dans une catégorie particulière. Par exemple, le sexe, la situation de famille, la profession, la branche d'activité, la religion, le pays de naissance ou la couleur des cheveux. Il n'y a pas d'ordre particulier dans cette catégorie de données nominales, donc aucune catégorie n'est considérée comme plus petite ou plus grande (par exemple il n'y a pas d'ordre d'importance pour les femmes ou pour les hommes). Quand les variables sont mesurées avec une échelle nominale, il est important que les catégories de la variables soient **mutuellement exclusives** (c'est-à-dire que toutes les valeurs de la variable sont classifiées dans une catégories seulement) et exhaustives (c'est-à-dire que toutes les valeurs possibles de la variable peuvent être classifiées dans chacune des catégories).

*On utilise une **échelle nominale** quand les observations sont mutuellement exclusives et exhaustives et quand il n'existe aucun ordre dans les observations.*

échelle ordinale

L'échelle ordinale est la mesure qui permet aux données d'être classées (c'est-à-dire que différentes valeurs peuvent être identifiées comme étant plus grandes ou plus petites). Néanmoins, la différence entre les valeurs ne peut pas être mesurée de manière significative. On peut par exemple évaluer la

performance d'un travailleur comme étant 'bonne', 'moyenne' ou 'pauvre'; on peut classifier des personnes en 'grands', 'de taille moyenne' ou 'petits'; on peut évaluer votre santé comme étant 'au-dessus de la moyenne', ou 'en-dessous de la moyenne'. À partir de ces exemples, il est clair que les différentes observations de la variable peuvent être classées (par exemple 'bon' est meilleur que 'moyen' et 'moyen' est meilleur que 'pauvre'), mais les différences entre ces valeurs ne peuvent pas être mesurées de manière significative. Les échelles ordinales sont les plus courantes dans les enquêtes d'attitudes — comme le questionnaire d'évaluation de ce cours.

*On utilise une **échelle ordinale** quand les valeurs des observations peuvent être mises en ordre ou classées, mais quand on ne peut pas mesurer la différence entre les observations.*

Données Quantitatives

échelle d'intervalle

L'**échelle d'intervalle** intervient quand la différence entre les valeurs peut être mesurée, mais qu'il n'y a pas de relativité entre les valeurs (c'est-à-dire qu'on ne peut pas dire qu'une des valeurs est n fois plus grande ou plus petite qu'une autre valeur). Par exemple l'échelle d'intervalle qui sert à mesurer le temps ou la température.

température

Si les températures maximum et minimum d'aujourd'hui sont 15°C et 30°C, alors on peut dire que la température maximum est 15°C de plus que la température minimum. Néanmoins, on ne peut pas dire que la temperature maximum est deux fois la température minimum. La caractéristique principale d'une échelle d'intervalle est que le point zéro est simplement un point parmi d'autres sur l'échelle. Zéro degré Celsius ne veut pas dire absence de chaleur. Pour s'en apercevoir on peut convertir les températures à des degrés Fahrenheit (c'est-à-dire 59°F et 86°F). On peut maintenant se rendre compte que ça n'a aucun sens de dire que la température maximum est égale à deux fois la température minimum.

le temps

Un groupe de personnes a commencé une tâche à 13 heures qui leur a pris 20 mn. Un autre groupe a commencé à 14h10 qui leur a pris 10 mn. On peut dire que le premier groupe a été 10 mn plus lent, ou deux fois plus lent que le deuxième groupe. Mais on ne peut pas dire que 13h20 est deux fois plus lent que 14h20. C'est la différence en minutes entre les deux temps qui peut être mesurée, pas le temps.

*On utilise une **échelle d'intervalle** quand on peut mesurer la différence entre les observations mais qu'il n'y a pas de relativité entre les observations et qu'il n'y a pas de point fixe zéro.*

échelle de rapport

L'**échelle de rapport** intervient lorsque les données sont mesurées pour que la relativité entre les valeurs puisse être établie de manière significative. Comme dans l'échelle d'intervalle, les observations sont classées, et la différence entre les observations est significative. En plus, l'échelle de rapport utilise le chiffre zéro pour indiquer l'absence de caractéristiques à mesurer — 0 dollar veut dire que vous n'avez pas d'argent. C'est la caractéristique distinctive des données de rapport, le point fixe zéro.

Ceci permet des comparaisons relatives significatives entre les observations. Par exemple, si vous avez 5 dollars et que votre ami en a 10, votre ami a deux fois plus d'argent que vous. Alors, le rapport entre les observations prend un sens. On trouve beaucoup d'exemples de données de rapport : la taille, l'âge, la distance, les revenus, parmi tant d'autres. Non seulement on peut dire qu'une personne de 40 ans est de 30 ans plus agée que quelqu'un de 10 ans, mais on peut également dire qu'elle est 4 fois plus agée. Une hauteur zéro, un âge zéro, un poids, une distance zéro ou des revenus zéro ne peuvent être interprétés que d'une seule manière.

*On utilise une **échelle de rapport** quand on peut mesurer la différence entre les observations et qu'il existe une relativité entre les observations.*

Données unidimensionnelles, bidimensionnelles ou multidimensionnelles

termes les plus couramment utilisés

Le groupe de termes utilisés pour décrire les variables se rapporte à la manière dont les données sont présentées ou analysées. Les termes 'unidimensionnel', 'bidimensionnel' et 'multidimensionnel' sont fréquemment utilisés en statistique. Il est alors essentiel de savoir ce que ces termes veulent dire, bien que pour ce cours on se tiendra seulement à étudier les données unidimensionnelles et bidimensionnelles.

une variable

'Unidimensionnel' veut simplement dire que l'on n'a affaire qu'à une variable. Par exemple, si on a des données statistiques d'une école sur les tailles des élèves, les données sont unidimensionnelles.

deux variables

'Bidimensionnel' veut dire à deux variables. Si on prend à nouveau l'exemple d'une école, on peut regarder le rapport entre les tailles et les poids des élèves, ou le rapport entre les poids et les âges.

Dans la vie de tous les jours, les statisticiens sont en général impliqués dans l'analyse de données bidimensionnelles. Par exemple, on peut répartir des tailles et des poids sur un diagramme pour établir la relation entre la taille et le poids. On utilise souvent un diagramme en nuage de points pour établir la relation entre deux diagrammes — on étudiera les diagrammes en nuages de points au chapitre 4.

plus de deux variables

'Multidimensionnel' signifie que l'on a affaire à plus de deux variables. Si on veut connaître par exemple la relation entre la taille, le poids et l'âge, ça implique une analyse plus compliquée, multidimensionnelle des données. Les analyses multidimensionnelles recourent à des techniques statistiques complexes et nous n'avons pas l'intention de nous y pencher dans ce cours. Les personnes qui voudront étudier les données multidimensionnelles devront consulter des textes et documents appropriés sur les méthodes statistiques.