Cours L3 MIAGE

Introduction à la statistique descriptive.

Télécharger les données Classe.data.txt depuis Celene.
Charger ces données sous R.

Supposons que l'on vous demande de dégager les caractéristiques principales de ces données. Il s'agit d'utiliser des outils de statistique descriptive afin de les décrire de manière synthétique (nous verrons comment dans les séances qui viennent).

A votre avis, quelle est la première chose à faire? Quelle caractéristique doit on regarder en premier?

Avant toute étude statistique il est important d'obtenir plus de précision sur la nature des données et la manière dont elles ont été recueillies. Cette étape est très importante si l'on veut pouvoir choisir les méthodes adaptés et pouvoir tirer des conclusions pertinentes de nos analyses.

On cherche notamment à préciser:

- population (souvent notée Ω): ensemble concerné par une étude statistique. Ce terme de population peut faire références à des ensembles de toute nature: étudiants d'une académie, production d'une usine, entreprises d'un secteur donné, …
- individu w $\in\Omega$: un élément de la population
- échantillon: sous-ensemble de la population sur lequel sont effectivement réalisées les observations
- enquête statistique: opération qui consiste à observer l'ensemble des individus d'un échantillon
  1. Recensement: l'échantillon est toute la population => Description exhaustive de la population.
  2. Sondage: l'échantillon est seulement une partie de la population => Cherche à mieux connaître des caractéristiques de la population à partir de l 'échantillon.

On appelle variable aléatoire X (statistique) une fonction de Ω dans V qui correspond à une caractéristique définie sur la population.
Lors d'une enquête statistique, on observe les valeurs x_1,...,x_n prises par cette variable sur les n individus de l'échantillon. On appelle n la taille de l 'échantillon. Ces observations sont couramment appelées « données » ou statistiques.

On distingue différents types de variables aléatoires (et donc de données):

- On dit qu'une variable est quantitative si elle prend des valeurs numériques sur lesquelles cela a un sens de faire des opérations arithmétiques (addition, soustraction, moyenne)
  Parmi ces variables certaines on différencie les variables:
  discrètes parce qu'elles ne prennent qu'un nombre fini ou au plus dénombrable de valeurs (généralement valeurs entières, décimales)
  continues qui peuvent prendre n'importe quelle valeur sur un intervalle de R ou de R^p (taille exacte, distance, …)
  Remarque: En statistique on parle de généralement données discrètes lorsque la variable ne prend qu'un petit nombre de valeur (en général moins de 20) et de données continues dans le cas contraire.

- On dit sinon qu'une variable est qualitative. Elle prend des valeurs qui en général ne sont pas numériques. Il peut arriver que des variables qualitatives soient codées avec des valeurs numériques mais pour lesquelles cela n'a pas de sens de faire des opérations arithmétiques
  On distingue parmi ces variables les variables
  nominales dont les valeurs représentent différents « états » qui ne peuvent pas être ordonnés.
  ordinales dont les valeurs représentent différents « états » qui peuvent être ordonnés.
  Attention, pour prendre en compte cet ordre on encode souvent les données qualitatives nominales de manière numérique mais cela n'a aucun sens de faire des calculs dessus (somme, soustraction, moyenne, variance, …)

Les données apparaissent souvent sous forme de série brute x_1, …,x_n qui correspond aux observations réalisées sur les n individus de l'échantillon. Cependant lorsque la taille n de l'échantillon est grande cette série est illisible. On va proposer des méthodes (statistique descriptive) pour en dégager les caractéristiques principales.

Quelques notations:
modalités: ensemble des valeurs que peut prendre la variable étudiée (on se limite en général aux valeurs prises sur l'échantillon sauf si on a plus d'information). On les note m_i .
effectif associé à la modalité m_i: nombre de fois qu'apparaît la modalité m_i dans les données. On le note e_i.
fréquence associée à la modalité m_i: proportion d'apparition de la modalité m_i dans les données. On les note généralement f_i=e_i/n.
mode: modalité qui apparaît avec la plus forte fréquence dans les données.

EXERCICE 1: préciser la population, les individus, la taille de l'échantillon et le type des variables des données contenues dans le fichier Classe.data.txt .

EXERCICE 2: indiquer la nature des caractères suivant et la population concernée. Age des employés d'une entreprise X - place sur laquelle a lieu la compensation banquaire d'un chèque - solvabilité d'un client - nombre de part d'imposition d'un ménage - espèce animale - chiffre d'affaire des entreprises Européenne - catégorie socio-professionnelle des habitants d'Orléans - nombre de pièces défectueuses dans un lot de résistances produit par l'entreprise Y - durée de vie des lave-vaisselle modèle x de la marque m - longueur de la queue à quatre quarante cinq au guichet du bureau de poste d'Olivet .

Il arrive que les données n'apparaissent pas sous la forme de la série brute mais sous la forme d'un tableau donnant l'effectif correspondant à chaque modalité (ou classe s'il s'agit de variable continue).
Pour pouvoir traiter ce genre de données avec R il faut générer un vecteur ou chaque modalité (ou centre de classe s'il s'agit de variable continue) apparaît le bon nombre de fois.

A. Analyse de Données qualitatives

i) nominales

On obtient les différentes modalités de la variable ainsi que les effectifs correspondants avec la commande **table()** (la première ligne correspond aux modalités, la deuxième aux effectifs).
On représente généralement les fréquences (resp. les effectifs) sous forme de diagramme en bâton pour lequel chaque barre a pour longueur la fréquence, (resp. l'effectif). Lorsque l'on vous demande de tracer la distribution des données, on vous demande de tracer les fréquences. Il s'agit d'une version empirique de la distribution de la variable aléatoire X. Sous certaines conditions, lorsque n

est grand cette distribution empirique et la vraie distribution de X sont proches.
Pour obtenir les fréquences, il suffit de diviser les effectifs par le nombre de données.

EXERCICE:

1. Charger les données contenues dans le fichier Classe.data dans R avec la commande read.table dans une table nommée data.
2. Definir les vecteurs sexe, departement, et serie.
3. Représenter les fréquences des variables sexe, departement, et serie en utilisant respectivement un digramme en bâtons, en digramme en barres, un diagramme par secteurs.
4. Déterminer les modes de ces séries.

ii) variables ordinales

Lorsque les variables sont ordinales, cela a un sens d'encoder les données de manière numérique de manière à prendre en compte l'ordre qu'il existe entre les différentes modalités. Toutefois, il faut bien comprendre que cela n'a pas de sens de chercher à faire des calculs sur ces données. On peut toutefois aller un peu plus loin que ce que nous avons appris à faire pour les données qualitatives nominales au pargraphe précédent. En dehors des représentations et caractéristiques évoquées dans la paragraphe précédent, cela a un sens de considérer:

- la fréquence cumulée associée à la modalité m_i: proportion de données qui sont inférieure ou égales à m_i.
- les valeurs extrêmes: modalité maximale et modalité minimale.
- la médiane: valeur centrale des données telle qu'il y ait approximativement 50% des données qui lui soient inférieures et 50% qui lui soient supérieures. Notons x_(1),...,x_(n) les données rangées par ordre croissant. Par convention, lorsque la taille n de l'échantillon est impaire (n=2s+1) la médiane est la valeur x_(s+1) et lorsque la taille est paire (n=2s) la médiane est (x_(s)+x_(s+1))/2.
- On peut défnir de manière semblable les quartiles qui correspondent à d'autres pourcentages:
  premier quartile: approximativement 25% lui sont inférieures et 75% lui sont supérieures.
  deuxième quartile: la médine
  troisème quartile: approximativement 75% lui sont inférieures et 25% lui sont supérieures.
  On utilise les mêmes onventions que pour la médiane.
- Les déciles correspondent à des pas de 10%

On représente souvent les fréquences cumulées fc qui correpondent à une courbe constante par morceaux qui représente les proportions de données inférieures ou égales à x lorsque x varie.
On peut calculer les fréquences cumulées (resp. les effectifs cumulés) à partir des fréquences (resp. des effectifs).

On présente souvent la boîte à moustache associée au données qui donne un résumé des autres caractéristiques. La base correspond à la valeur minimale tandis que le sommet correspod à la valeur maximale, Le bas de la boîte représente le premier quartile et le haut le troisème. Le trait central représente la médiane. Dans certains cas les valeurs extrêmes (situées à plus de 3/2 de la longueur de la boîte par rapport à la médiane) sont représentées par des points et ne sont pas prises en compte pour tracer le graphique.

EXERCICE:
1, Définir le vecteur difficulte
2, Calculer les fréquences cumulées associées
3. Tracer la boîte à moustache associée.
4. Comparer les difficultés des étudiants provenant des filières “S” et “ES” au travers de leurs boîtes à moustache.