TP 3

Statistiques descriptives univariées

Compétences attendues.

- ✓ Regrouper une série statistique par modalités ou par classes.
- ✓ Connaître les indicateurs de position (moyenne, médiane, quartiles) et les commandes associées.
- ✓ Connaître les indicateurs de dispersion (écart-type, étendue, distance inter-quartile) et les commandes associées.
- ✓ Représenter graphiquement une série statistique.

Objectifs. L'objet des statistiques descriptives univariées (ou unidimensionnelles) est de fournir des résumés synthétiques, graphiques et numériques, de séries de valeurs observées sur une population ou un échantillon. On présente ici les indicateurs les plus couramment employés pour décrire une série statistique.

Mathieu Mansuy - Professeur de Mathématiques en ECS2 au Lycée Louis Pergaud (Besançon)
mansuy.mathieu@hotmail.fr

1 Principales notions en statistiques descriptives

1.1 Présentation des données

Soit $\Omega = \{\omega_1, \dots, \omega_n\}$ un ensemble fini. Un tel ensemble sera appelé *population* en statistique descriptive. On appellera ses éléments ω des *individus*, et son cardinal n l'*effectif de la population*.

Exemple. Ω = l'ensemble de la population française, Ω = l'ensemble des voitures immatriculées en France.

On étudie un *caractère* de cette population.

Définition.

Un *caractère* (ou *variable*) sur la population Ω est une application $X : \Omega \to E$, où E désigne un ensemble quelconque.

Si E est un ensemble de nombres, on dit que X est un caractère *quantitatif*. Dans le cas contraire, on parle de caractère qualitatif.

Exemple. Un caractère possible sur la population française est la taille (caractère quantitatif) ou encore la couleur des yeux (caractère qualitatif).

Nous ne traiterons que du cas des caractères quantitatifs.

Définition.

- On appelle *série statistique* de la population Ω pour le caractère X la donnée de la liste $X(\Omega) = (X(\omega_1), \dots, X(\omega_N))$ des valeurs prises par X sur Ω.
- Les valeurs prises par X sont appelées *modalités*.

  L'*effectif d'une modalité* est le nombre de fois où cette modalité apparait dans la série statistique.

  La *fréquence d'une modalité* est son effectif divisé par l'effectif total.

Le saviez vous ?

Les statistiques sont nées en Angleterre, au début du $XVII^e$ siècle pour décompter les décès lors d'une épidémie de peste. Ce n'était à l'époque que des données numériques, sans outil théorique pour les analyser. Il faut attendre le XIX^e siècle pour voir l'apparition de méthodes mathématiques pour l'étude de telles données. Ce n'est qu'à la fin du XIX^e siècle que la statistique devient une discipline à part entière des mathématiques sous l'impulsion des savants anglais Karl Pearson et Udny Yule.

Exemple. On considère la série statistique suivante :

$$2, 11, 7, 2, 15, 4, 5, 5, 5, 13, 5, 15, 7, 7, 8, 10, 10, 10, 11, 13, 7, 2, 15, 15$$

L'ensemble des modalités est $\{2, 4, 5, 7, 8, 10, 11, 13, 15\}$. L'effectif de la modalité 2 est $n_2 = 3$ et sa fréquence est $\dfrac{n_2}{n} = \dfrac{3}{24}$.

Représentation informatique. Sous `Scilab`, nous représenterons les séries statistiques par des vecteurs (lignes ou colonnes). L'effectif de la série est obtenue à l'aide de la commande `length`.

Regroupement par modalités

Pour présenter les données, on regroupe la série statistique par *modalités - effectifs*, c'est-à-dire qu'on donne :

- la liste (x_i) des modalités du caractère X,
- les effectifs (n_i) correspondants : pour tout i, n_i est le nombre d'individus ω de l'échantillon tels que $X(\omega) = x_i$.

On peut aussi choisir de présenter cette série regroupée par *modalité - fréquence*, en donnant les modalités (x_i) et les fréquences des modalités (f_i) correspondantes.

Pour passer d'une série statistique *brute* à une série statistique groupée par modalités ordonnées, on peut utiliser l'instruction `tabul`.

Définition.

Si `x` est un vecteur, `y=tabul(x,"i")` est une matrice à deux colonnes, la première contenant les valeurs prises par les composantes de `x` rangées dans l'ordre croissant (décroissant sans `"i"`) et la seconde contenant le nombre d'occurrences de chaque valeur.

Exercice 1

En utilisant `Scilab`, grouper la série statistique de l'exemple précédent par modalités - effectifs, puis par modalités - fréquences.

Définition.

On appelle *fréquence cumulée d'une modalité* la somme de toutes les fréquences des modalités qui lui sont inférieures.

Exercice 2

En utilisant la commande `cumsum`, déterminer le vecteur des fréquences cumulées de la série statistique précédente.

Regroupement par classes

Lorsque $x = (x_1, \dots, x_n)$ est une série statistique dont le nombre de modalités est très grand, plutôt que de conserver toutes les valeurs, il est plus intéressant de les regrouper par classes :

- on considère une suite de réels $c = (c_0 < \cdots < c_k)$ définissant les *classes* $I_1 = [c_0, c_1]$, $I_2 =]c_1, c_2]$, $\dots$, $I_k =]c_{k-1}, c_k]$, l'*amplitude* de la classe I_i étant $c_i - c_{i-1}$;
- On note n_i le nombre d'éléments de x appartenant à l'intervalle I_i pour $1 \leq i \leq k$.

On se ramène ainsi à une série statistique de taille k, dont les modalités sont les milieux $y_i = \dfrac{c_{i-1} + c_i}{2}$ des classes et d'effectifs correspondants les n_i.

Pour grouper une série par classes, on peut utiliser l'instruction `dsearch` (*dichotomic search*).

Définition.

Soit `x` un vecteur, `c` un vecteur constitué d'une suite de réels strictement croissante.
L'instruction `[a,b]=dsearch(x,c)` crée deux vecteurs :

- `a` qui est de même taille que `x`, et qui à la position x_i indique le numéro j de l'intervalle I_j auquel il appartient ;
- `b` qui est de taille k (où $k+1$ est la taille de `c`) et qui contient en j-ème position le nombre d'éléments de `x` dans l'intervalle I_j.

Exercice 3

La commande `grand(q,r,'unf',a,b)` permet de simuler une matrice de taille $q \times r$ dont les entrées suivent la loi uniforme continue $\mathcal{U}([a,b])$.

À l'aide de la commande `grand`, simuler 10000 nombres suivant la loi uniforme sur $[0,10]$. Regrouper cette série statistique par modalités, puis par classes (choisir 5 classes de même amplitude).
Quelle présentation de la série statistique préférez vous ? Que peut on dire de l'effectif de chaque classe ? Était ce prévisible ?

1.2 Indicateurs de position

Définition.

On appelle *moyenne* de la série statistique $x = (x_1, \dots, x_n)$ le réel :

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Remarque. Si la série statistique est groupée par modalités - effectifs (modalités $(m_1, \dots, m_p)$ d'effectifs $(n_1, \dots, n_p)$) alors on a :

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{p} n_i \cdot m_i.$$

Définition.

La *médiane* d'une série statistique ordonnée un réel m partageant la série en deux séries d'effectifs égaux. Si $(x_1 \leq x_2 \leq \cdots \leq x_n)$ est la série statistique ordonnée, m est défini par :

- si $n = 2p-1$ est impaire, $m = x_p$ (la valeur du milieu) ;
- si $n = 2p$ est paire, $m = \dfrac{x_p + x_{p+1}}{2}$.

Définition.

Soit `x` un vecteur.

- `mean(x)` donne la moyenne du vecteur `x`.
- `median(x)` donne une médiane du vecteur `x` (non nécessairement ordonné).

Exercice 4

Déterminer la moyenne et la médiane de la série statistique.

Définition.

Le *premier quartile* q_1 d'une série statistique x est la plus petite modalité de x ayant une fréquence cumulée supérieure ou égale à $\frac{1}{4}$. Autrement dit, c'est la plus petite valeur de la série telle que 25 % des valeurs lui soient inférieures ou égales.

Le *troisième quartile* q_3 est la plus petite modalité de la série telle que 75 % des valeurs de la série lui soient inférieures ou égales.

De même, on définit les *déciles* et les *centiles* d'une série statistique.

Exercice 5

À l'aide du vecteur des fréquences cumulées de la série statistique, déterminer le premier quartile, le troisième quartile et le huitième décile.

Définition.

On appelle *mode* d'une série statistique toute modalité pour laquelle l'effectif est maximal (il peut y en avoir plusieurs).

Exercice 6

En utilisant l'instruction `find` (dont le fonctionnement a été expliqué lors du TP1), déterminer le(s) mode(s) de la série statistique.

1.3 Indicateurs de dispersion

Définition.

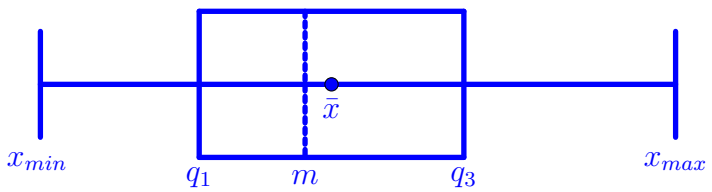
On appelle *écart-type* d'une série statistique $x = (x_1, \dots, x_n)$ le réel :

$$\sigma_{n-1}(x) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Définition.

- On appelle *étendue* d'une série statistique la différence entre la plus grande et la plus petite modalité.
- On appelle *distance inter-quartile* le réel $q_3 - q_1$.

Remarque. La distance inter-quartile est un indicateur de dispersion : c'est la longueur de l'*intervalle inter-quartile* $[q_1, q_3]$, lequel contient la moitié des valeurs de la série, réparties autour de la médiane. On représente parfois la boîte à moustache de la série statistique :



Définition.

Soit `x` un vecteur.

- `stdev(x)` (pour standard deviation) donne l'écart-type du vecteur `x`.
- `max(x)-min(x)` donne l'étendue du vecteur `x`.

Exercice 7

Déterminer l'écart-type de la série statistique, et représenter son diagramme à moustache.

2 Représentation graphique

Aux notions de statistiques descriptives introduites dans la section précédente, on associe des représentations graphiques de trois types :

- *Diagramme en bâtons / diagramme cumulatif en bâton.*

  On représente la série statistique **groupée par modalités** en plaçant sur l'axe horizontal les modalités et en dressant à la verticale de chacune un bâton de hauteur égale à son effectif ou sa fréquence (resp. sont effectif cumulé ou sa fréquence cumulée).

- *Histogramme.*

  On représente la série statistique **groupée par classes** en plaçant les c_i sur un axe horizontal et en traçant à la verticale un rectangle de base $[c_i, c_{i+1}]$ d'aire égale à la fréquence de la classe correspondante.

- *Diagramme circulaire.*

  Chaque modalité (ou classe) est représentée par un secteur circulaire dont l'angle au centre est égal à la fréquence de la modalité (ou de la classe) multipliée par 360°.

Pour toutes ces représentations graphiques, on donne l'instruction `Scilab` à utiliser.

Définition.

- *Diagramme en bâtons.* On utilise l'instruction `bar`.

  Si `a` et `b` sont des vecteurs, `bar(a,b)` trace un diagramme en bâtons d'abcisse `a` et d'ordonnée `b`.

  On suppose que le vecteur `x` contient une série statistique brute. Pour obtenir un diagramme en bâtons, on peut alors utiliser la syntaxe :

  ```
  y=tabul(x,"i")
  ```

  `bar(y(:,1),y(:,2))` (effectifs) / `bar(y(:,1),y(:,2)/length(x))` (fréquences)

- *Histogramme.* On utilise l'instruction `histplot`.

  `histplot(c,x)` trace l'histogramme des données `x` où `c` est un vecteur aux composantes strictement croissantes définissant les classes.

  `histplot(n,x)` trace l'histogramme des données `x` où `n` désigne le nombre de classes (qui sont alors équiréparties entre la plus petite et la plus grande valeur de `x`).

- *Diagramme circulaire.* On utilise l'instruction `pie`.

  On suppose que le vecteur `x` contient une série statistique brute. Pour obtenir un diagramme circulaire, on peut alors utiliser la syntaxe :

  ```
  y=tabul(x,"i")
  ```

  `pie(y(:,2))` ou `pie(y(:,2),['x1','x2',...,'xp'])` si on veut rajouter des légendes.

Exercice 8
Représenter le diagramme en bâtons des effectifs de la série statistique, ainsi que le diagramme circulaire par modalité.

Exercice 9
La commande `grand(q,r,'nor',μ,σ)` permet de simuler une matrice de taille $q \times r$ dont les entrées suivent la loi normale $\mathcal{N}(\mu, \sigma)$.

À l'aide de la commande `grand`, simuler 10000 nombres suivant la loi normale $\mathcal{N}(3, 4)$. Regrouper cette série par classes (à vous de choisir des classes appropriées), puis la représenter à l'aide d'un diagramme circulaire et d'un histogramme.

3 Étude de la population mondiale

À partir de mon site `mathieu-mansuy.fr/ecs2`, télécharger et exécuter le fichier `population_mondiale.sce`, qui définit plusieurs matrices lignes :

- `pays` contient les noms des pays ;
- `superficie` contient la surface terrestre en milliers de km^2 de chaque pays ;
- `population` contient le nombre d'habitants en millions de chaque pays ;
- `naissance` contient le nombre de naissances sur 1000 habitants ;
- `mort` contient le nombre de décès sur 1000 habitants ;
- `homme` contient l'espérance de vie des hommes ;
- `femme` contient l'espérance de vie des femmes ;

à partir des données contenues dans l'étude 2017 de l'Institut National d'Études Démographiques (disponible également sur mon site). Pour les exercices qui suivent, vous trouverez en annexe de ce TP la liste des pays et de leurs index.

Exercice 10 (Interrogation de la base de données)
Écrire une fonction `donnees(n)` qui affiche le nom, la superficie, le nombre d'habitants et la densité de population du pays d'index n.

Exercice 11

1. Calculer la surface terrestre mondiale, le nombre d'habitants mondial et la densité moyenne d'habitants au km^2.
2.
3. Calculer la surface terrestre, le nombre d'habitants et la densité moyenne d'habitants au km^2 pour chaque continent.
4. Pour chacune des données étudiées en (a), représenter la répartition de la surface terrestre et du nombre d'habitants par continents sous la forme de diagrammes en camembert à l'aide de l'instruction `pie`.

Exercice 12

On considère l'espérance de vie des hommes (ou des femmes) par pays.

1. Calculer la moyenne sur l'ensemble des pays.
2. Calculer l'écart-type et la médiane.
3. Calculer les espérances de vie minimale et maximale en précisant les pays correspondant à ces valeurs extrémales (utiliser l'instruction `find`).
4. Représenter l'histogramme de l'espérance de vie des hommes sur l'intervalle $[0, 100]$ avec 20 classes. Quelle est la classe modale de l'espérance de vie des hommes ?
5. Trier le tableau `homme` par ordre croissant et en déduire :
   (a) les valeurs du premier et du troisième quartile ainsi que l'écart inter-quartile ;
   (b) les valeurs du premier et du neuvième décile ainsi que la liste des pays dont l'espérance de vie est inférieure au premier décile ou supérieure au neuvième décile.

Exercice 13

On rappelle que le taux d'accroissement naturel est la différence entre la natalité et la mortalité.

1. Quels sont les accroissements minimaux et maximaux ? Préciser les pays.
2. Faire afficher la liste des pays pour lesquels l'accroissement est négatif.
3. Déterminer l'accroissement mondial moyen.
4. Dans ses projections, l'INED prévoit une population mondiale de 9731 millions d'habitants en 2050. Cela est-il conforme à l'hypothèse d'un taux d'accroissement constant ?

Annexe. liste des pays et de leurs index

Afrique

Afrique septentrionale

1. Algérie
2. Égypte
3. Libye
4. Maroc
5. Sahara occidental
6. Soudan
7. Tunisie

Afrique occidentale

8. Bénin
9. Burkina Faso
10. Cap-Vert
11. Côte d'Ivoire
12. Gambie
13. Ghana
14. Guinée
15. Guinée-Bissau
16. Liberia
17. Mali
18. Mauritanie
19. Niger
20. Nigeria
21. Sénégal
22. Sierra Leone
23. Togo

Afrique orientale

24. Burundi
25. Comores
26. Djibouti
27. Érythrée
28. Éthiopie
29. Kenya
30. Madagascar
31. Malawi
32. Maurice
33. Mayotte
34. Mozambique
35. Ouganda
36. Réunion
37. Rwanda
38. Seychelles
39. Somalie
40. Sud-Soudan
41. Tanzanie
42. Zambie
43. Zimbabwe

Afrique centrale

44. Angola
45. Cameroun
46. Centrafricaine (Rép.)
47. Congo
48. Congo(Rép. dém.)
49. Gabon
50. Guinée équatoriale
51. Sao Tomé-et-Principe
52. Tchad

Afrique australe

53. Afrique du Sud
54. Botswana
55. Lesotho
56. Namibie
57. Swaziland

Amérique

Amérique septentrionale

58. Canada
59. États Unis

Amérique centrale

60. Belize
61. Costa Rica
62. Guatemala
63. Honduras
64. Mexique
65. Nicaragua
66. Panama
67. Salvador

Caraïbes

68. Antigua-et-Barbuda
69. Aruba
70. Bahamas
71. Barbade
72. Cuba
73. Curaçao
74. Dominicaine (Rép.)
75. Dominique
76. Grenade
77. Guadeloupe
78. Haïti
79. Jamaïque
80. Martinique
81. Porto Rico
82. Sainte Lucie
83. St Vincent Grenadines
84. St.Kitts-et-Nevis
85. Trinité-et-Tobago
86. Vierges (Îles)

Amérique du sud

87. Argentine
88. Bolivie
89. Brésil
90. Chili
91. Colombie
92. Équateur
93. Guyana
94. Guyane (française)
95. Paraguay
96. Pérou
97. Surinam
98. Uruguay
99. Venezuela

Asie

Asie occidentale

100. Arabie saoudite
101. Arménie
102. Azerbaïdjan
103. Bahreïn
104. Chypre
105. Émirats arabes unis
106. Géorgie
107. Irak
108. Israël
109. Jordanie
110. Koweït
111. Liban
112. Oman
113. Palestine (Territoires)
114. Qatar
115. Syrie
116. Turquie
117. Yémen

Asie centrale

118. Kazakhstan
119. Kirghizistan
120. Tadjikistan
121. Turkménistan
122. Ouzbékistan

Asie du sud

123. Afghanistan
124. Bangladesh
125. Bhoutan
126. Pakistan
127. Inde
128. Iran
129. Maldives
130. Népal
131. Sri Lanka

Asie du sud-ouest

132. Brunei
133. Cambodge
134. Indonésie
135. Laos
136. Malaisie
137. Myanmar (Birmanie)
138. Philippines
139. Singapour
140. Thaïlande
141. Timor-Est
142. Viêt Nam

Asie orientale

143. Chine
144. Chine-Hong Kong
145. Chine-Macao
146. Corée du Nord
147. Corée du Sud
148. Japon
149. Mongolie
150. Taïwan

Europe

Europe septentrionale

151. Danemark
152. Estonie
153. Finlande
154. Irlande
155. Islande
156. Lettonie
157. Lituanie
158. Norvège
159. Royaume-Uni
160. Suède

Europe occidentale

161. Allemagne
162. Autriche
163. Belgique
164. France (métropolitaine)
165. Liechtenstein
166. Luxembourg
167. Monaco
168. Pays-Bas
169. Suisse

Europe orientale

170. Biélorussie
171. Bulgarie
172. Hongrie
173. Moldavie
174. Pologne
175. Roumanie
176. Russie
177. Slovaquie
178. Tchèque (République)
179. Ukraine

Europe méridionale

180. Albanie
181. Andorre
182. Bosnie-Herzégovine
183. Croatie
184. Espagne
185. Grèce
186. Italie
187. Kosovo
188. Macédoine
189. Malte
190. Monténégro
191. Portugal
192. Saint-Marin
193. Serbie
194. Slovénie

Océanie

195. Australie
196. Fidji
197. Guam
198. Kiribati
199. Marshall (Îles)
200. Micronésie (États fédérés de)
201. Nouvelle-Calédonie
202. Nouvelle-Zélande
203. Papouasie-Nouvelle Guinée
204. Polynésie française
205. Salomon (Îles)
206. . Samoa occidentales
207. Tonga
208. Vanuatu