

Informatique

TP 7

Statistiques descriptives

L'objet de la statistique descriptive est la description de données à travers leur *présentation* (la plus comode et la plus synthétique possible), leur *représentation graphique* et le calcul de *résumés numériques*. On parle également de *statistique exploratoire* et d'*analyse des données*.

PARTIE 1 : STATISTIQUES UNIVARIÉES

On souhaite étudier, chez les *individus* ω d'une *population*¹ Ω , l'expression d'une *variable*² ou d'un *caractère* X prenant différents états dans un ensemble E . On supposera que $E \subset \mathbb{R}$, c'est-à-dire que le caractère étudié est numérique³.

On peut distinguer deux types de caractères quantitatifs selon les valeurs qu'ils prennent :

- La variable X est dite *discrète* lorsqu'elle prend ses valeurs dans un ensemble E fini ou dénombrable (par exemple, des valeurs entières). Pour présenter les données, on considère une énumération $(x_i)_{1 \leq i \leq r}$ de $X(\Omega)$ et l'on définit, pour tout $i \in \llbracket 1, r \rrbracket$, l'*effectif* n_i associé à la valeur x_i , i.e. le nombre d'individus $\omega \in \Omega$ pour lesquels le caractère $X(\omega)$ prend la valeur x_i . On obtient alors une *série statistique* $((x_i, n_i))_{1 \leq i \leq r}$. La donnée des effectifs n_i équivaut à celle des *fréquences* $f_i = \frac{n_i}{n}$, où $n = \text{Card } \Omega = \sum_{i=1}^r n_i$ est l'effectif total. On a bien sûr $\sum_{i=1}^r f_i = 1$.

On définit ensuite les *fréquences cumulées* $F_i = \sum_{j \leq i} f_j$.

- La variable X est dite *continue* lorsqu'elle prend ses valeurs dans un ensemble E non dénombrable, par exemple un intervalle de \mathbb{R} . On lui associe une série statistique discrète en choisissant une famille de classes $C_1 = [a_0, a_1]$, $C_2 =]a_1, a_2]$, ..., $C_r =]a_{r-1}, a_r]$ recouvrant $X(\Omega)$, où $(a_i)_{0 \leq i \leq r}$ est une suite strictement croissante. On identifie alors chaque classe C_i à son milieu $x_i = \frac{a_{i-1} + a_i}{2}$ et l'on définit l'effectif associé n_i d'individus ω pour lesquels le caractère $X(\omega)$ prend une valeur de C_i .

Comme on l'a déjà vu lors des TP de simulation, une variable statistique discrète X peut être représentée par un diagramme en bâtons (instruction `bar` en Scilab) et une variable continue par un histogramme (instruction `histplot` en Scilab).

PARTIE 2 : INDICATEURS DE POSITION ET DE DISPERSION

On considère une série statistique $X = ((x_i, n_i))_{1 \leq i \leq r}$.

- On définit la *moyenne* de la série :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^r n_i x_i = \sum_{i=1}^r f_i x_i,$$

sa *variance empirique* :

$$\sigma^2(X) = \frac{1}{n} \sum_{i=1}^r n_i (x_i - \bar{X})^2 = \sum_{i=1}^r f_i (x_i - \bar{X})^2$$

et son écart-type empirique $\sigma(X) = \sqrt{\sigma^2(X)}$.

On établit sans difficulté la formule de Koëning-Huygens :

$$\sigma^2(X) = \frac{1}{n} \sum_{i=1}^r n_i x_i^2 - \left(\frac{1}{n} \sum_{i=1}^r n_i x_i \right)^2 = \overline{X^2} - \bar{X}^2.$$

- On définit, pour tout $x \in \mathbb{R}$, la *fréquence cumulée* jusqu'à x :

$$F(x) = \sum_{\substack{1 \leq i \leq r : \\ x_i \leq x}} f_i,$$

1. La terminologie est à considérer dans un sens très large : Ω est un ensemble fini quelconque.

2. Une variable est donc une application $X : \Omega \rightarrow E$.

3. Ces caractères sont dits *quantitatifs*, par opposition aux caractères *qualitatifs* que l'on n'étudiera pas dans ce TP.

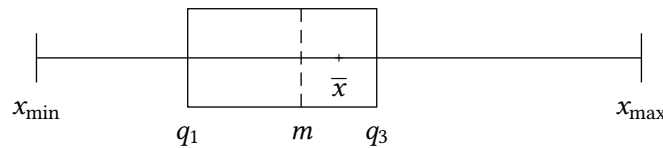
égale à la proportion d'individus ω pour lesquels $X(\omega) \leq x$. On définit alors le *quantile* d'ordre $\alpha \in [0, 1]$ par $t_\alpha = \min\{x \in X(\Omega) : F(x) \geq \alpha\}$: c'est la plus petite valeur de X supérieure ou égale à une proportion α d'éléments de la série.

En particulier, la *médiane* m est le quantile d'ordre $\frac{1}{2}$: elle partage la population de deux sous-populations d'effectifs égaux ou presque. En supposant la série ordonnée $x_1 \leq \dots \leq x_n$:

- > si $n = 2p - 1$ est impair, la médiane est la valeur centrale x_p de la série ;
- > si $n = 2p$ est pair, on prend $m = \frac{x_p + x_{p+1}}{2}$ comme médiane de la série plutôt que x_p (qui serait donnée par une application stricte de la définition).

On définit également le *premier quartile* $q_1 = t_{1/4}$, le *troisième quartile* $q_3 = t_{3/4}$ ainsi que les déciles, centiles, etc.

L'*écart inter-quartile* $q_3 - q_1$ 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 :



- Le (ou les) *mode(s)* (aussi appelé(s) *valeur(s) modale(s)*) est (sont) la (les) valeur(s) de X pour laquelle (lesquelles) l'effectif est maximal.

L'*étendue* est la différence entre la plus grande et la plus petite valeur de la série.

Moyenne, médiane, mode, quantiles sont des indicateurs de position de la série. Étendue, variance et écart-type empiriques, écart inter-quartile sont des indicateurs de dispersion.

En Scilab, l'instruction `mean(X)` permet de calculer la moyenne, `median(X)` la médiane, `quart(X)` les quartiles, `min(X)` et `max(X)` les valeurs minimale et maximale d'une série dont les valeurs sont contenues dans le vecteur X (avec répétition). L'instruction `stdev(X)` renvoie l'écart-type normalisé par $n - 1$:

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

Une option permet d'obtenir l'écart-type $\sigma(X)$ (on renvoie à l'aide).

Si f désigne le vecteur des fréquences associées aux valeurs de la série contenues dans le vecteur X (sans répétition), alors les instructions `meanf(X, f)` et `stdevf(X, f)` renvoient la moyenne et l'écart-type de la série.

PARTIE 3 : ÉTUDE DE LA POPULATION MONDIALE

Télécharger l'archive `tp7.zip` sur le site `www.rblld.fr` (rubrique 2015-2016/informatique) 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 ;
- > `deces` 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 contenus dans l'étude 2013 de l'INED (en annexe, ainsi qu'une liste des pays et de leurs index).

Exercice 1 : 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 2

- Calculer la surface terrestre mondiale, le nombre d'habitants mondial et la densité moyenne d'habitants au km^2 .
- Calculer la surface terrestre, le nombre d'habitants et la densité moyenne d'habitants au km^2 pour chaque continent.
 - Pour chacune des données étudiées en **a.**, représenter la répartition par continent sous la forme d'un diagramme en camembert à l'aide de l'instruction `pie`.

Exercice 3

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

- Calculer la moyenne sur l'ensemble des pays.
- Calculer l'écart-type.
- Calculer la médiane.
- Calculer les espérances de vie minimale et maximale en précisant les pays correspondant à ces valeurs extrêmes (utiliser l'instruction `find`).
- 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 ?
- Trier le tableau homme par ordre croissant et en déduire :
 - les valeurs du premier et du troisième quartile ainsi que l'écart inter-quartile ;
 - 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 4 : accroissement naturel et prévision

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

- Quels sont les accroissements minimal et maximal ? Préciser les pays.
- Faire afficher les listes des pays pour lesquels l'accroissement est négatif.
- Déterminer l'accroissement mondial moyen.
- Dans ses projections, l'INED prévoit une population mondiale de 9 731 millions d'habitants en 2050. Cela est-il conforme à l'hypothèse d'un taux d'accroissement constant ?

PARTIE 4 : STATISTIQUES BIVARIÉES

Dans la suite du TP, on s'intéresse à l'étude simultanée de deux caractères numériques X et Y se rapportant à une même population. On considère donc une série statistique bivariée, i.e. une application $\Omega \rightarrow \mathbb{R}^2$, que l'on représentera sous la forme d'une liste de couples (x_i, y_i) , $1 \leq i \leq n$, correspondant aux valeurs des caractères étudiés pour les différents individus.

La représentation des points de coordonnées (x_i, y_i) , $1 \leq i \leq n$, dans le plan constitue le *nuage* de points associé à la série statistique double étudiée. Le *point moyen* du nuage est le point de coordonnées (\bar{X}, \bar{Y}) . Scilab propose l'instruction `plot2d(X, Y, -1)` pour tracer le nuage de points de la série double (X, Y) , où l'option `-1` a pour effet de ne pas relier les points.

Exercice 5

On considère de nouveau les données étudiées dans la troisième partie. Représenter le nuage de points associé à la série double formée par l'espérance de vie des hommes d'une part et celles des femmes d'autres part.

Lorsque les caractères sont discrets, si l'on note $x_i, 1 \leq i \leq r$, les valeurs de X et $y_j, 1 \leq j \leq s$, celles de Y, alors on peut caractériser la série en donnant la fréquence d'apparition $f_{i,j}$ de chaque couple (x_i, y_j) . La matrice $(f_{i,j})_{i,j} \in \mathbf{M}_{r,s}(\mathbb{R})$ est appelée *matrice de contingence* (parfois présentée sous forme de tableau). Pour tout $i \in \llbracket 1, r \rrbracket$, la *fréquence (marginale)* de x_i est donnée par $f_{i,\cdot} = \sum_{j=1}^s f_{i,j}$ et l'on a $\sum_i f_{i,\cdot} = 1$. De même pour les fréquences marginales $f_{\cdot,j}$ des $y_j, 1 \leq j \leq s$.

La *covariance* de la série double est par définition le réel

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X} \bar{Y}$$

c'est-à-dire, avec les fréquences :

$$\text{cov}(X, Y) = \sum_{\substack{1 \leq i \leq r \\ 1 \leq j \leq s}} f_{i,j} (x_i - \bar{X})(y_j - \bar{Y}) = \sum_{\substack{1 \leq i \leq r \\ 1 \leq j \leq s}} f_{i,j} x_i y_j - \bar{X} \bar{Y}$$

Scilab propose l'instruction `covar(X, Y, f)` pour calculer la covariance du couple (X, Y) , où X, Y contiennent les valeurs prises par chacune des séries et f est la matrice de contingence (qui peut être remplacée par la matrice des effectifs).

Si la série statistique bivariée est donnée sous la forme d'une liste de couples $(x_i, y_i), 1 \leq i \leq n$, on utilisera donc l'instruction `covar(X, Y, eye(n, n))`.

Exercice 6

Calculer la covariance du couple étudié dans l'exercice précédent.

PARTIE 5 : CORRÉLATION LINÉAIRE

Soit (X, Y) une série statistique bivariée donnée sous forme de liste $(x_i, y_i), 1 \leq i \leq n$. Dans ce paragraphe, on essaie d'expliquer la variable Y (dite variable *expliquée*) à partir de la variable X (dite *explicative*) en décidant si les données sont compatibles avec une relation du type $Y = aX + b$.

Si la variable X prend au moins deux valeurs, on a vu en TD sur les espaces euclidiens qu'il existe un unique couple (\hat{a}, \hat{b}) minimisant la fonctionnelle

$$(a, b) \in \mathbb{R}^2 \mapsto \delta^2(a, b) = \frac{1}{n} \sum_{i=1}^n (y_i - ax_i - b)^2$$

La droite d'équation $y = \hat{a}x + \hat{b}$ est appelée *droite de régression* (ou *droite des moindres carrés*) de Y en X. Les formules obtenues pour \hat{a} et \hat{b} en TD montrent que son équation peut encore s'écrire sous la forme

$$y - \bar{Y} = \frac{\text{cov}(X, Y)}{\sigma^2(X)} (x - \bar{X}),$$

ce qui montre que cette droite passe par le point moyen de la série. On peut également montrer⁴ que

$$\sigma^2(Y) = \hat{a}^2 \sigma^2(X) + \delta^2(a, b)$$

4. En effet, en munissant \mathbb{R}^n de sa structure euclidienne canonique,

$$n\delta^2(\hat{a}, \hat{b}) = \min_{(a,b) \in \mathbb{R}^2} \sum_{i=1}^n (y_i - ax_i - b)^2 = \min_{(a,b) \in \mathbb{R}^2} \|y - ax - b\mathbf{1}\|^2$$

représente le carré de la distance $d(y, F)$ du vecteur $y = (y_1, \dots, y_n)$ au sous-espace vectoriel F engendré par les vecteurs $x = (x_1, \dots, x_n)$ et $\mathbf{1} = (1, \dots, 1)$. Par théorème, cette distance est atteinte au projeté orthogonal $p_F(y) = \hat{a}x + \hat{b}\mathbf{1}$ de y sur F.

Dans la formule ci-dessus, le terme $\widehat{a}^2\sigma^2(X)$ est la composante de la variance de Y expliquée par la régression. Le terme $\delta^2(\widehat{a}, \widehat{b})$, quant à lui, est la variance résiduelle : il provient du fait que $Y = \widehat{a}X + \widehat{b} + \varepsilon$, où ε est une fonction inconnue.

Pour se satisfaire de la formule $Y = \widehat{a}X + \widehat{b}$, il faut que ε ait une distribution « sans structure particulière » autour de 0, et que la part de la variance de Y expliquée par la régression soit la plus forte possible. Cette part est donnée par le carré du *coefficient de corrélation linéaire*

$$\rho_{Y/X} = \frac{\widehat{a}\sigma(X)}{\sigma(Y)} = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)}.$$

Ce coefficient de corrélation linéaire appartient à $[-1, 1]$. Il est égal à ± 1 si, et seulement si, les points de la série sont alignés.

L'instruction Scilab `correl(X, Y, f)` renvoie le coefficient de corrélation de la série statistique (X, Y) de matrice de contingence *f*.

L'instruction `[a, b]=reglin(X, Y)` calcule les coefficients *a* et *b* de la droite de régression linéaire de Y en X.

Exercice 7 : un « bon » coefficient de corrélation n'assure pas une relation affine

On considère des séries X et Y définies par les instructions

```
X=1:20;
```

```
Y=X.^2+rand(1,20);
```

1. Représenter le nuage de points associés.
2.
 - a. Calculer le coefficient de corrélation linéaire. Semble-t-il bon ?
 - b. Calculer les coefficients *a* et *b* de l'équation de la droite de régression de Y par rapport à X.
 - c. Superposer la droite de régression linéaire au nuage de points.
 - d. Sur une autre figure, représenter les écarts résiduels $y_i - ax_i - b$, $1 \leq i \leq 20$.
3.
 - a. Vérifier que le nuage de points se superpose bien avec la parabole d'équation $y = x^2$.
 - b. Étudier la corrélation linéaire de Y par rapport à X^2 .

Exercice 8 : morts sur la route

Exécuter le script `morts.sce`.

On s'intéresse aux facteurs influençant la mortalité sur la route aux États-Unis. Pour ce faire, on a consigné dans plusieurs vecteurs les données suivantes pour chacun des 26 états :

- > le vecteur `morts` contient le nombre de morts ;
- > le vecteur `conducteurs` contient le nombre de conducteurs ;
- > le vecteur `densite` contient la densité de population ;
- > le vecteur `routes` contient la longueur des routes ;
- > le vecteur `temp` contient la température annuelle moyenne ;
- > le vecteur `fuel` contient la consommation de fuel.

1. Déterminer la corrélation empirique entre le nombre de morts et la consommation de fuel. Quelle relation de causalité existe entre ces deux quantités ?
2.
 - a. Déterminer, parmi toutes les variables explicatives, la variable X_1 pour laquelle la régression linéaire est la plus pertinente.

Le théorème de Pythagore assure alors que :

$$\|y - \bar{Y}\mathbf{1}\|^2 = \|\mathcal{P}_F(y - \bar{Y}\mathbf{1})\|^2 + d(y - \bar{Y}, F)^2 = \|\widehat{a}x + (\widehat{b} - \bar{Y})\mathbf{1}\|^2 + d(y, F)^2 = \|\widehat{a}(x - \bar{X}\mathbf{1})\|^2 + d(y, F)^2$$

d'où le résultat en divisant par *n*.

- b.** Déterminer alors la droite de régression correspondante $M = a_1X_1 + b_1$, où M désigne la variable donnant le nombre de morts.
- 3.** On souhaite étendre le modèle linéaire à un modèle à deux variables $M = a_1X_1 + a_2X_2 + b_2$. Déterminer, parmi toutes les variables explicatives restantes, la variable X_2 pour laquelle la régression $M - a_1X_1 = a_2X_2 + b_2$ est la plus pertinente.
- 4.** Recommencer ainsi de suite jusqu'à épuiser toutes les variables explicatives.
- 5.** Critiquer le modèle final.

Annexe : liste des pays et de leurs index

AFRIQUE

Afrique septentrionale

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

Afrique occidentale

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

Afrique orientale

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

Afrique centrale

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

Afrique australe

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

AMERIQUE

Amérique septentrionale

- | | | |
|------------|----------------|--|
| 58. Canada | 59. États-Unis | |
|------------|----------------|--|

Amérique centrale

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

Caraïbes

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

Amérique du sud

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

ASIE**Asie occidentale**

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

Asie centrale

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

Asie du sud

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

Asie du sud-ouest

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

Asie orientale

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

EUROPE**Europe septentrionale**

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

Europe occidentale

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

Europe orientale

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

Europe méridionale

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

OCEANIE

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

Numéro 503
Septembre 2013



Population & Sociétés

English
Version

Tous les pays du monde (2013)

Gilles Pison*

Tous les deux ans, *Population & Sociétés* publie un numéro intitulé *Tous les pays du monde* présentant un tableau de la population mondiale**. Celle-ci compte un peu plus de 7 milliards d'habitants en 2013. Elle a été multipliée par sept au cours des deux derniers siècles, et devrait continuer à croître jusqu'à atteindre 10 à 11 milliards à la fin du XXI^e siècle.

Les données du grand tableau central concernent toutes les entités géopolitiques dont la population atteint ou dépasse 150 000 habitants, et quelques autres. Les États souverains y voisinent ainsi avec d'autres territoires, dont les départements français d'outre-mer. Pays et entités sont classés géographiquement, selon la pratique des annuaires des Nations unies, par région et continent.

Les indicateurs démographiques sont les mêmes que dans les éditions précédentes : superficie, population estimée à la mi-2013, taux de natalité et de mortalité, taux de mortalité infantile, indice synthétique de fécondité, pourcentage des moins de 15 ans et des 65 ans ou plus dans la population totale, espérance de vie masculine et féminine à la naissance, revenu national brut 2012 par habitant en parité du pouvoir d'achat, et proportion en 2011 de personnes infectées par le virus du sida (VIH) parmi celles ayant entre 15 et 49 ans (sont inclus à la fois les malades du sida et les personnes infectées par le virus mais n'ayant pas développé la maladie). À noter que les indicateurs sont des projections, mais sur quelques années seulement, les statistiques d'état civil ou d'enquête n'étant pas encore disponibles pour l'année 2013 elle-même.

Dans dix-huit petits tableaux, les pays ou entités sont classés selon quelques indicateurs par ordre décroissant. Dans les huit premiers tableaux, un total mondial est indiqué et une ligne sépare les pays dont le cumul dépasse la moitié du total mondial. Par exemple, les six pays les plus peuplés (Chine, Inde, États-Unis, Indonésie, Brésil, Pakistan) totalisent 3,59 milliards d'habitants, plus de la moitié du total mondial estimé à 7,14 milliards. Dans le

dix-septième et avant-dernier petit tableau, les pays sont classés selon la proportion des 15-64 ans dans la population totale. Elle donne une idée de l'importance de la population d'âge actif. Elle est particulièrement élevée dans les petits États du Golfe persique, qui accueillent une importante population de travailleurs immigrés venus sans leur famille, et dans les pays où la fécondité a fortement baissé pour atteindre des niveaux très bas (Russie, Pologne). Les personnes d'âge actif sont également proportionnellement très nombreuses dans les pays du Sud qui ont aussi connu récemment une baisse très rapide de leur fécondité (Chine, Iran, Viêt Nam) : leur pyramide des âges s'est rétrécie à la base, alors que leur sommet est encore très étroit. Cette situation ne devrait pas durer et la proportion de 15-64 ans devrait revenir à des niveaux plus faibles au fur et à mesure du vieillissement de la population.

* Institut national d'études démographiques.

** Les données proviennent essentiellement de la World Population Data Sheet publiée par le Population Reference Bureau (PRB) [1]. Cet organisme indépendant synthétise chaque année les chiffres émanant de plusieurs sources, essentiellement la Division de la Population des Nations unies [2], le Bureau of the Census des États-Unis, le Conseil de l'Europe et... l'Ined, qui s'efforcent de rassembler l'ensemble des données démographiques publiées par les offices nationaux de statistique et les organisations internationales.

Références

[1] Carl Haub et Toshiko Kaneda – 2013, World Population Data Sheet, Population Reference Bureau, 1875 Connecticut Ave. NW Suite 520, Washington DC 20009-5728, États-Unis (www.prb.org).

[2] Nations unies, Division de la population – 2013, *World Population Prospects : The 2012 Revision*, New York (<http://esa.un.org/unpd/wpp/>)