Informatique
TP 7

Statistiques descriptives

L'objet de la statistique descriptive est la description de données à travers leur *présentation* (la plus commode 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 numérique[3].

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 $(x_i)_{1\leqslant i\leqslant r}$ de $X(\Omega)$ et l'on définit, pour tout $i \in [\![1, r]\!]$, 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* $\big((x_i, n_i)\big)_{1\leqslant i\leqslant r}$.
  Cette présentation de la série est parfois qualifiée de *dépouillée*, par opposition à la présentation *brute*, qui consisterait à faire la liste des valeurs $X(\omega)$ lorsque ω parcourt Ω.
  La donnée des effectifs n_i équivaut à celle des *fréquences* $f_i = \frac{n_i}{n}$, où $n = \operatorname{Card}\Omega = \sum_{i=1}^{r} n_i$ est l'effectif total. On a bien sûr $\sum_{i=1}^{r} f_i = 1$.
  On définit ensuite les *fréquences cumulées* $F_i = \sum_{j\leqslant i} f_j$.
- 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 $(a_i)_{0\leqslant i\leqslant r}$ 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

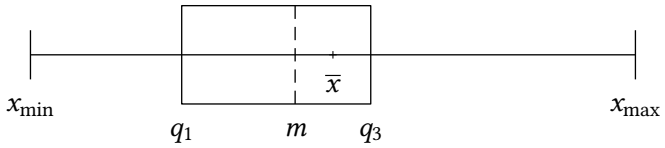
On considère une série statistique $X = \big((x_i, n_i)\big)_{1\leqslant i\leqslant r}$.

- On définit la *moyenne* de la série :
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{r} n_i x_i = \sum_{i=1}^{r} f_i x_i,$$
sa *variance empirique* :
$$\sigma^2(X) = \frac{1}{n}\sum_{i=1}^{r} n_i (x_i - \overline{X})^2 = \sum_{i=1}^{r} f_i (x_i - \overline{X})^2$$
et son écart-type empirique $\sigma(X) = \sqrt{\sigma^2(X)}$.
On établit sans difficulté la formule de Koënig-Huygens :
$$\sigma^2(X) = \frac{1}{n}\sum_{i=1}^{r} n_i x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{r} n_i x_i\right)^2 = \overline{X^2} - \overline{X}^2.$$

1. La terminologie est à considérer dans un sens très large : Ω est un ensemble fini quelconque.
2. Une variable est donc une application $X : \Omega \longrightarrow E$.
3. Ces caractères sont dits *quantitatifs*, par opposition aux caractères *qualitatifs* que l'on n'étudiera pas dans ce TP.

- On définit, pour tout $x \in \mathbb{R}$, la *fréquence cumulée* jusqu'à x :
$$\mathrm{F}(x) = \sum_{\substack{1 \leqslant i \leqslant r : \\ x_i \leqslant x}} f_i,$$
égale à la proportion d'individus ω pour lesquels $\mathrm{X}(\omega) \leqslant x$. On définit alors le *quantile* d'ordre $\alpha \in [0,1]$ par $t_\alpha = \min\{x \in \mathrm{X}(\Omega) : \mathrm{F}(x) \geqslant \alpha\}$: 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 $x_1 \leqslant \cdots \leqslant x_n$:
  - ➢ 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 brute dont les valeurs sont contenues dans le vecteur `X` (avec répétition). L'instruction `stdev(X)` renvoie l'écart-type normalisé par $n-1$:
$$\sigma_{n-1}(\mathrm{X}) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{\mathrm{X}})^2} = \sqrt{\frac{n}{n-1}}\sigma(\mathrm{X}).$$
Une option permet d'obtenir l'écart-type $\sigma(\mathrm{X})$ (on renvoie à l'aide).
Pour une série dépouillée, 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 2017-2018/informatique) et éxécuter le fichier `population_mondiale.sce`, qui définit plusieurs vecteurs :

- ➢ `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 2017 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 `function donnees(n)` qui affiche le nom, la superficie, le nombre d'habitants et la densité de population du pays d'index n.

Exercice 2

1. Calculer la surface terrestre mondiale, le nombre d'habitants mondial et la densité moyenne d'habitants au km^2.
2. **a.** Calculer la surface terrestre, le nombre d'habitants et la densité moyenne d'habitants au km^2 pour chaque continent.
   b. 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.

1. Calculer la moyenne sur l'ensemble des pays.
2. Calculer l'écart-type.
3. Calculer la médiane.
4. Calculer les espérances de vie minimale et maximale en précisant les pays correspondant à ces valeurs extrémanes (utiliser l'instruction `find`).
5. 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 ?
6. 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 4 : acroissement naturel et prévision

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

1. Quels sont les accroissements minimal et maximal ? Préciser les pays.
2. Faire afficher les listes 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 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 $\Omega \longrightarrow \mathbb{R}^2$, qui admet une présentation *brute* sous la forme d'une liste de couples (x_i, y_i), $1 \leqslant i \leqslant n$, correspondant aux valeurs des caractères étudiés pour les différents individus.
La représentation des points de coordonnées (x_i, y_i), $1 \leqslant i \leqslant n$, 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 $(\overline{X}, \overline{Y})$.
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, $1 \leqslant i \leqslant r$, les valeurs de X et y_j, $1 \leqslant j \leqslant s$, celles de Y, alors on peut présenter la série sous forme *dépouillée* en donnant la fréquence $f_{i,j}$ d'apparition 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 $i \in [\![1, r]\!]$, 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, $1 \leqslant j \leqslant s$.
On définit alors la *covariance* de la série double par la formule

$$\operatorname{cov}(\mathrm{X}, \mathrm{Y}) = \sum_{\substack{1 \leqslant i \leqslant r \\ 1 \leqslant j \leqslant s}} f_{i,j}(x_i - \overline{\mathrm{X}})(y_j - \overline{\mathrm{Y}}) = \sum_{\substack{1 \leqslant i \leqslant r \\ 1 \leqslant j \leqslant s}} f_{i,j} x_i y_j - \overline{\mathrm{X}}\,\overline{\mathrm{Y}}$$

c'est-à-dire, dans le cas d'une présentation brute,

$$\operatorname{cov}(\mathrm{X}, \mathrm{Y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{\mathrm{X}})(y_i - \overline{\mathrm{Y}}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{\mathrm{X}}\,\overline{\mathrm{Y}}.$$

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 présentée sous forme brute (x_i, y_i), $1 \leqslant i \leqslant n$, on pourra utiliser l'instruction `covar(X,Y,eye(n,n))`.

Exercice 6

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

Partie 5 : régression linéaire

Soit (X, Y) une série statistique bivariée donnée sous forme brute, i.e. d'une de liste (x_i, y_i), $1 \leqslant i \leqslant n$.
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 $\mathrm{Y} = a\mathrm{X} + b$.
Si la variable X prend au moins deux valeurs, on a vu en TD sur les espaces euclidiens qu'il existe un unique couple $(\widehat{a}, \widehat{b})$ minimisant la fonctionnelle

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

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

$$y - \overline{\mathrm{Y}} = \frac{\operatorname{cov}(\mathrm{X}, \mathrm{Y})}{\sigma^2(\mathrm{X})}(x - \overline{\mathrm{X}}), \tag{1}$$

ce qui montre que cette droite passe par le point moyen de la série. On peut également montrer [4] que

$$\sigma^2(\mathrm{Y}) = \widehat{a}^2\sigma^2(\mathrm{X}) + \delta^2(\widehat{a}, \widehat{b}). \tag{2}$$

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

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

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

Dans la formule ci-dessus, le terme $\widehat{a}^2\sigma^2(\mathrm{X})$ 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 $\mathrm{Y} = \widehat{a}\mathrm{X} + \widehat{b} + \varepsilon$, où ε est une fonction inconnue.

Pour décider d'accepter ou non la formule $\mathrm{Y} = \widehat{a}\mathrm{X} + \widehat{b}$, on examine les deux points suivants :

- Il est important que la proportion de la variance de Y expliquée par la régression soit la plus forte possible. D'après (2), cette proportion est donnée par

$$\mathrm{R}^2 = 1 - \frac{\delta^2(\widehat{a},\widehat{b})}{\sigma^2(\mathrm{Y})} = 1 - \frac{\sum_{i=1}^n (y_i - \widehat{a}x_i - \widehat{b})^2}{\sum_{i=1}^n (y_i - \overline{\mathrm{Y}})^2}.$$

  D'après (1), il s'agit aussi du carré du *coefficient de corrélation linéaire*

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

  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.
- On s'assure également d'avoir capturé au mieux la dépendance de Y en X en vérifiant que le nuage de points de la série double (X,ε) ne présente pas de « structure particulière ».

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 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 $y_i - ax_i - b$, $1 \leqslant i \leqslant 20$. Que met-on ainsi en évidence ?

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.
c. Effectuer la régression linéaire de Y par rapport à X^2. Vérifier la qualité de l'approximation graphiquement et étudier les écarts résiduels. Conclusion ?

Exercice 8 : une régression linéaire multiple pour illustrer l'absence de causalité

Une entreprise vend un produit sur 200 marchés différents. Elle a mis en place une campagne publicitaire sur trois supports : télévision, radio et journaux dont elle souhaite étudier l'impact sur les ventes.

En notant Y le nombre de ventes (en milliers d'unités) et X_1, X_2 et X_3 les budgets publicitaires (en milliers d'euros) dépensés respectivement sur les trois supports précédents, on s'interroge sur la pertinence d'un modèle linéaire de la forme

$$\mathrm{Y} = a_1\mathrm{X}_1 + a_2\mathrm{X}_2 + a_3\mathrm{X}_3 + b.$$

Le théorème de Pythagore assure alors que :

$$\left\|y - \overline{\mathrm{Y}}\mathbb{1}\right\|^2 = \left\|p_{\mathrm{F}}(y - \overline{\mathrm{Y}}\mathbb{1})\right\|^2 + d(y - \overline{\mathrm{Y}}\mathbb{1},\mathrm{F})^2 = \left\|\widehat{a}x + (\widehat{b} - \overline{\mathrm{Y}})\mathbb{1}\right\|^2 + d(y,\mathrm{F})^2 = \left\|\widehat{a}(x - \overline{\mathrm{X}}\mathbb{1})\right\|^2 + d(y,\mathrm{F})^2$$

d'où le résultat en divisant par n.

L'intérêt d'un tel modèle, s'il est acceptable, est d'expliquer la contribution sur les ventes des campagnes publicitaires sur chacun des trois supports, mais aussi de prédire quelles seront les ventes lors d'une nouvelle campagne publicitaire avec une répartition différente du budget ou sur de nouveaux marchés.

Le fichier `campagne_publicitaire.sce` contenu dans l'archive téléchargée en ligne définit les vecteurs `ventes`, `TV`, `radio` et `journaux` contenant respectivement les valeurs des variables Y, X_1, X_2 et X_3 sur les 200 marchés.

1. Effectuer une régression linéaire simple de Y par rapport à X_3. Que dire de l'impact de la campagne publicitaire dans les journaux sur les ventes ?

2. Sans chercher à donner de formules explicites, justifier l'existence de coefficients $\widehat{a}_1$, $\widehat{a}_2$, $\widehat{a}_3$ et $\widehat{b}$ optimaux en un sens que l'on précisera.

Le coefficient R^2 permettant de mesurer la pertinence du modèle vaut ici

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{a}_1 x_{1,i} - \widehat{a}_2 x_{2,i} - \widehat{a}_3 x_{3,i} - \widehat{b})^2}{\sum_{i=1}^{n}(y_i - \overline{Y})^2}.$$

L'instruction Scilab `[a,b]=reglin([X1;X2;X3],Y)` fournit les valeurs de $\widehat{a}_1$, $\widehat{a}_2$, $\widehat{a}_3$ et $\widehat{b}$.

3. **a.** Effectuer la régression linéaire de Y par rapport à X_1, X_2 et X_3.

b. Que dire à présent de l'impact de la campagne publicitaire dans les journaux sur les ventes ? Comment expliquer la différence avec la conclusion de la question **1.** ?

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

AMERIQUE

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. Kitts-et-Nevis
84. St Vincent Grenadines
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. Georgie
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. Ouzbékistan
121. Tadjikistan
122. Turkménistan

Asie du sud

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

OCEANIE

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

Numéro 547
Septembre 2017

Population & Sociétés

Tous les pays du monde (2017)

English Version

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 7,5 milliards d'habitants en 2017. Elle a été multipliée par sept au cours des deux derniers siècles, et devrait continuer à croître jusqu'à atteindre peut-être 11 milliards à la fin du XXIe 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, territoires, régions et collectivités 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-2017, 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 2016 par habitant en parité du pouvoir d'achat. Il faut noter que plusieurs indicateurs sont des projections, car les statistiques d'état civil ou d'enquête ne sont pas encore disponibles pour l'année 2017 elle-même.

Dans dix-huit petits tableaux, les pays ou entités sont classés selon quelques indicateurs par ordre décroissant. Dans les sept 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 sept pays les plus peuplés (Chine, Inde, États-Unis, Indonésie, Brésil, Pakistan, Nigeria) totalisent 3,93 milliards d'habitants, plus de la moitié du total mondial estimé à 7,54 milliards. Dans le tableau 10, les pays sont classés selon le taux de mortalité. Il peut paraître étonnant qu'avec 9 décès pour 1 000 habitants en 2017, le Burkina Faso soit mieux classé que le Japon qui en compte 10. Le relativement faible nombre de décès au Burkina Faso vient du fait que sa population est jeune et que la proportion de personnes âgées est très faible, alors qu'à l'inverse elle est élevée au Japon. Le calcul de l'espérance de vie, qui tient compte de la répartition par âge de la population, donne une idée plus juste des contrastes de mortalité. Le Japon se retrouve alors en tête du classement avec l'espérance de vie la plus élevée du monde (84 ans), alors que le Burkina Faso se situe presque en fin du classement (60 ans). 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 (Moldavie, Russie). Les personnes d'âge actif sont proportionnellement plus nombreuses dans les pays du Sud qui ont aussi connu récemment une baisse rapide de leur fécondité (Chine, Iran, Vietnam) : 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 revenir à des niveaux plus faibles au fur et à mesure du vieillissement de la population.

* Muséum national d'histoire naturelle et 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] Toshiko KANEDA et Geneviève DUPUIS, 2017, *World Population Data Sheet*, Population Reference Bureau, Washington DC, États-Unis (www.prb.org).

[2] Nations unies, Division de la population, 2017, World Population Prospects : The 2017 Revision, New York (http://esa.un.org/unpd/wpp/)