Université de Liège

Étude statistique des taux de natalité et mortalité dans différents pays du monde

Élément de statistiques

Groupe 19

Bastien Hoffmann (20161283)
Maxime Meurisse (20161278)

3e année de Bachelier ingénieur civil
Année académique 2018-2019

1 Analyse descriptive

1.a Histogrammes des taux de natalité et de mortalité

On travaille sur une base de données de 100 pays contenant leur nombre de naissances et de décès (par 1000 habitants) en 2013. Les histogrammes, dont les taux en abscisse sont en ‰[1], ont été générés par le script `Q1a`[2] grâce à la fonction `histogram`.

Tout d'abord, l'histogramme du taux de natalité met en évidence, à la figure 1, que la répartition des naissances est éparse, malgré une plus grande tendance vers les taux faibles (environ 10‰).

À l'inverse, celui du taux de mortalité montre une concentration centrée autour d'une valeur d'environ 8‰.

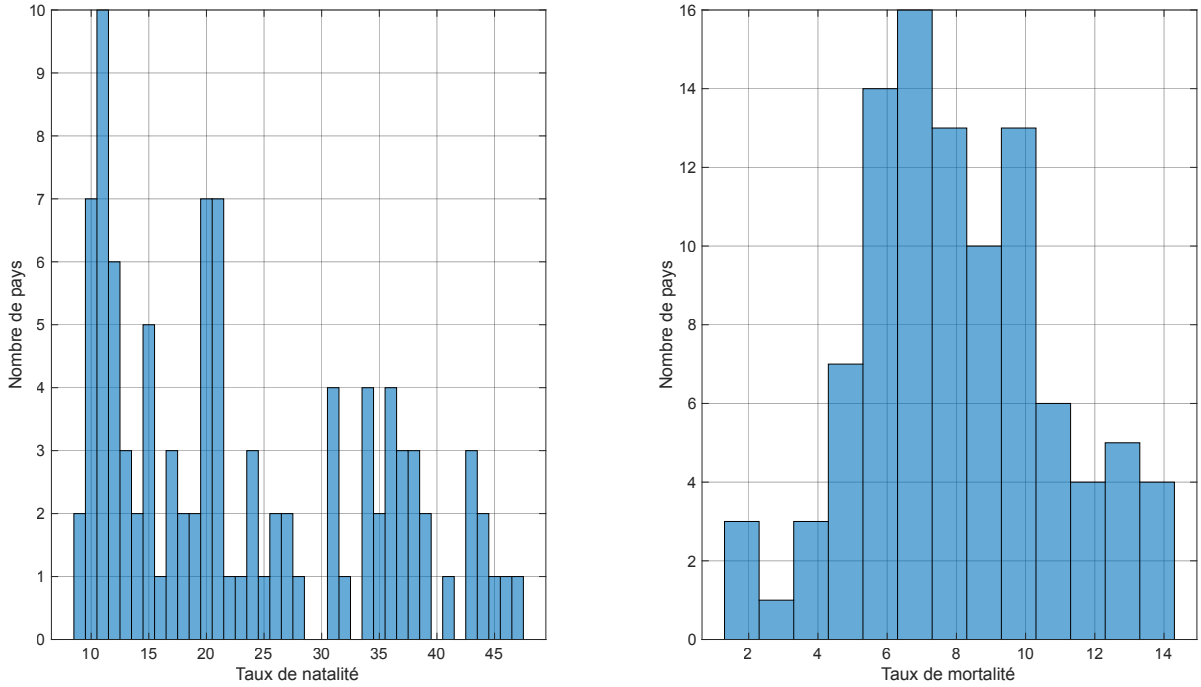

Figure 1 – Histogrammes des taux de natalité et de mortalité dans le monde.

Il est intéressant de mettre ces deux figures en parallèle afin de constater que le nombre maximum de naissances est largement (environ 4 fois) supérieur au nombre maximum de décès, traduisant, d'un point de vue démographique, une augmentation massive de la population mondiale. Une cause possible serait l'évolution technologique et qualitative des soins de santé, permettant à l'Homme de vivre bien plus longtemps qu'auparavant.

[1] Tous les taux de ce rapport, et notamment ceux des figures, même si cela n'est pas partout précisé et sauf indication contraire, sont en ‰.

[2] Tous les scripts et fonctions mentionnés dans ce rapport se trouvent en annexe et ont été exécutés via le logiciel Matlab.

1.b Statistiques descriptives des taux mondiaux

Les statistiques descriptives des taux mondiaux ont été calculées via le script `Q1b` avec les fonctions `mean`, `median`, `mode` et `std`[3] et sont présentées dans la table 1.

On remarque que pour le taux de natalité, la moyenne et la médiane sont un peu distantes l'une de l'autre et toutes deux distante du mode, confirmant cette répartition éparse observée à la section 1.a.

Ces 3 valeurs, pour le taux de mortalité, sont au contraire assez proches l'une de l'autre. L'écart-type du taux de mortalité est relativement faible, contrairement à celui du taux de natalité qui assez élevé, traduisant une fois de plus l'étalement des données autour de la moyenne.

Statistique descriptive	**Taux de natalité**	**Taux de mortalité**
Moyenne [‰]	23,1510	8,0270
Médiane [‰]	20,5000	7,8000
Mode [‰]	11,3000	9,6000
Écart-type [‰]	11,1497	2,8567
Belgique	11,7000	9,9000

Table 1 – Statistiques descriptives des taux mondiaux.

La Belgique présente un taux de natalité largement inférieur à la moyenne mondiale et un taux de mortalité, quant à lui, légèrement supérieur. Le taux de natalité faible s'explique en partie par le fait que la Belgique soit un pays développé, présentant souvent un nombre de naissance compensant à peine le nombre de décès. On peut également remarquer qu'elle est proche du mode du tableau pour les deux taux, signifiant que ce type de proportion est courante dans le monde.

1.c Caractéristiques d'un taux normal

Un taux normal, au sens de la loi normale, est un taux qui est compris dans l'intervalle

$$[\mu - \sigma; \mu + \sigma]$$

avec

- μ, la moyenne de la population;
- σ, l'écart-type de la population.

Ces deux valeurs ayant été calculées à la section 1.b, le script `Q1c` détermine l'intervalle pour chaque taux. Ceux-ci sont présentés à la table 2.

[3]Sauf indication contraire, tous les écarts-types calculés sont les écarts-types classiques, issus des variances non corrigées.

Taux	**Intervalle** [‰]
Taux de natalité	[12,0013; 34,3007]
Taux de mortalité	[5,1703; 10,8837]

Table 2 – Intervalles des taux normaux (au sens de la loi normale).

Le script `Q1c` calcule également que 54% des pays ont un taux de natalité normal et 68% un taux de mortalité normal.

Le taux de natalité prenant des valeurs très étendues avec une répartition éparse comme visible sur l'histogramme, il est donc logique que la proportion associée soit importante mais pas très élevée. À l'opposé, les valeurs du taux de mortalité étant déjà approximativement concentrées autour de leur moyenne, son histogramme se rapprochait déjà de la loi normale et la proportion de pays appartenant à l'intervalle est directement plus grande.

Concernant la Belgique, son taux de natalité se situe à gauche de l'intervalle des taux normaux et son taux de mortalité se situe dans l'intervalle. Son taux de natalité n'est donc pas normal au sens de la loi normale, mais bien son taux de mortalité.

1.d Boîtes à moustaches

Les boîtes à moustaches des taux de natalité et de mortalité, générées par le script `Q1d` avec la fonction `boxplot`, sont présentées à la figure 2.

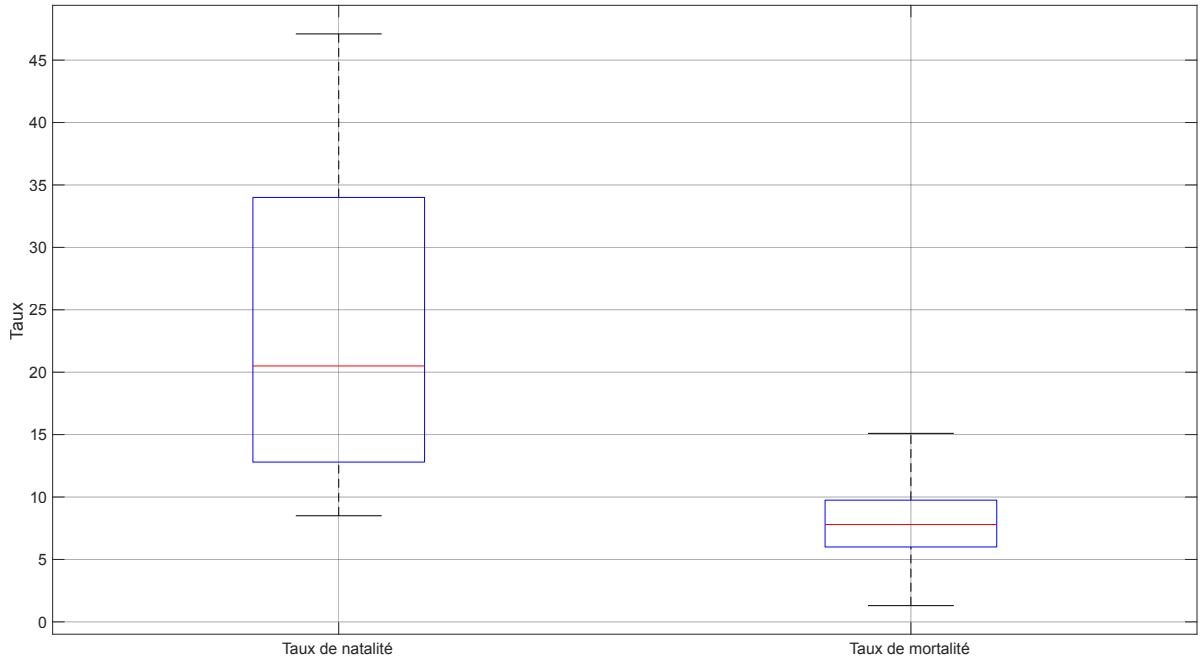

Figure 2 – Boîtes à moustaches relatives aux taux de natalité et de mortalité.

On constate immédiatement l'absence de données aberrantes pour les deux taux (aucun symbole "+" sur les figures, confirmé également par le script). Les quartiles, calculés avec la fonction `prctile`, sont explicités à la table 3. À noter que les deuxièmes quartiles, en rouge sur les boîtes à moustaches, sont les médianes, déjà calculées à la section 1.b.

Taux	**Premier quartile** [‰]	**Troisième quartile** [‰]
Taux de natalité	12,8000	34,0000
Taux de mortalité	6,0000	9,7500

Table 3 – Quartiles du taux de natalité et de mortalité.

On constate que les deux quartiles concernant le taux de natalité sont fort écartés, reprenant donc un grand nombre de pays dans la boîte, tandis que celle du taux de mortalité voit ses quartiles être très resserrés autour de la valeur médiane. Concernant les moustaches, appelées également bornes aberrantes, celles de la natalité sont asymétriques et la valeur maximale est très élevée, alors que celles du taux de mortalité sont symétriques par rapport à la médiane, traduisant une nouvelle fois l'allure de gaussienne de son histogramme.

1.e Polygone des fréquences cumulées du taux de natalité

Le polygone des fréquences cumulées du taux de natalité, généré par le script `Q1e` via la fonction `cdfplot`, est présenté à la figure 3.

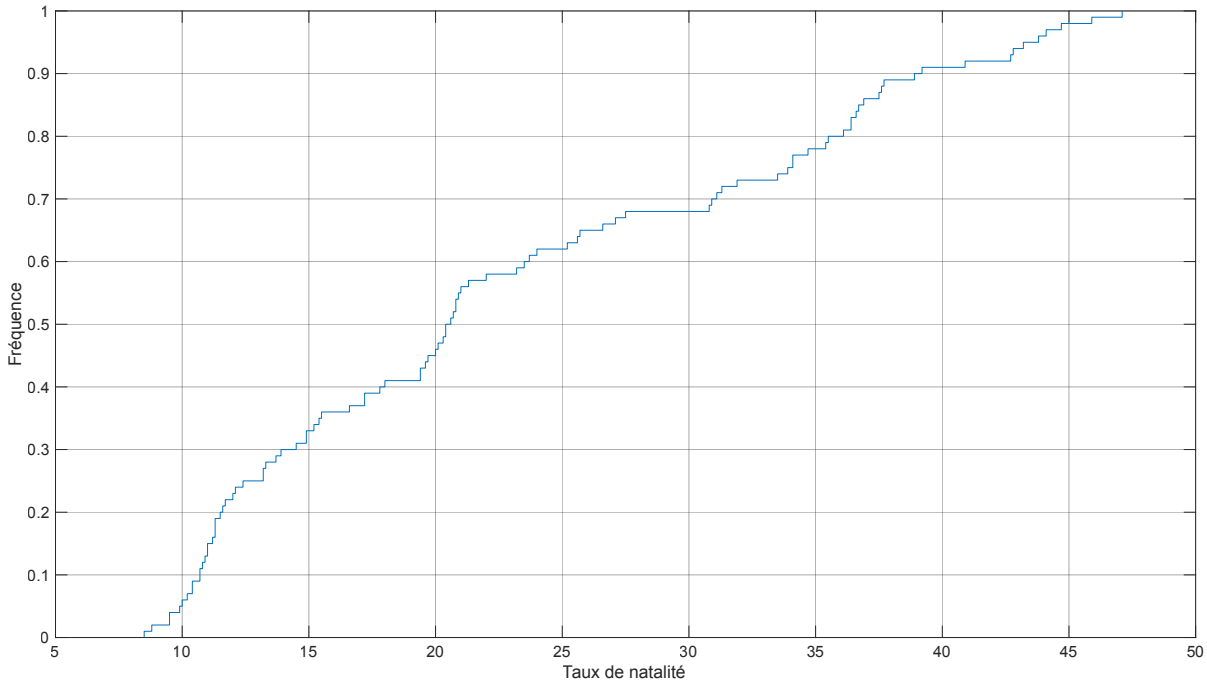

Figure 3 – Polygone des fréquences cumulées du taux de natalité.

L'estimation de la proportion de pays ayant un taux de natalité inférieur ou égal à 20 pour 1000 habitants et supérieur à celui de la Belgique est donnée par

$$\begin{aligned} F(20) - F(11{,}7) &= 0{,}4600 - 0{,}2200 \\ &= 0{,}2400 \end{aligned}$$

En effet, la valeur du polygone au point $x = 20$ ($F(20)$, calculée avec la fonction `ecdf`) donne la proportion de pays ayant un taux de natalité inférieur ou égal à 20. En retirant à cette proportion celle des pays ayant un taux de natalité inférieur ou égal à celui de la Belgique ($F(11{,}7)$), on obtient bien la proportion de pays appartenant à l'intervalle demandé.

1.f Nuage de points

Le nuage de points (*scatterplot*) comparant les deux taux, généré par la fonction `scatter` du script `Q1f`, est présenté à la figure 4.

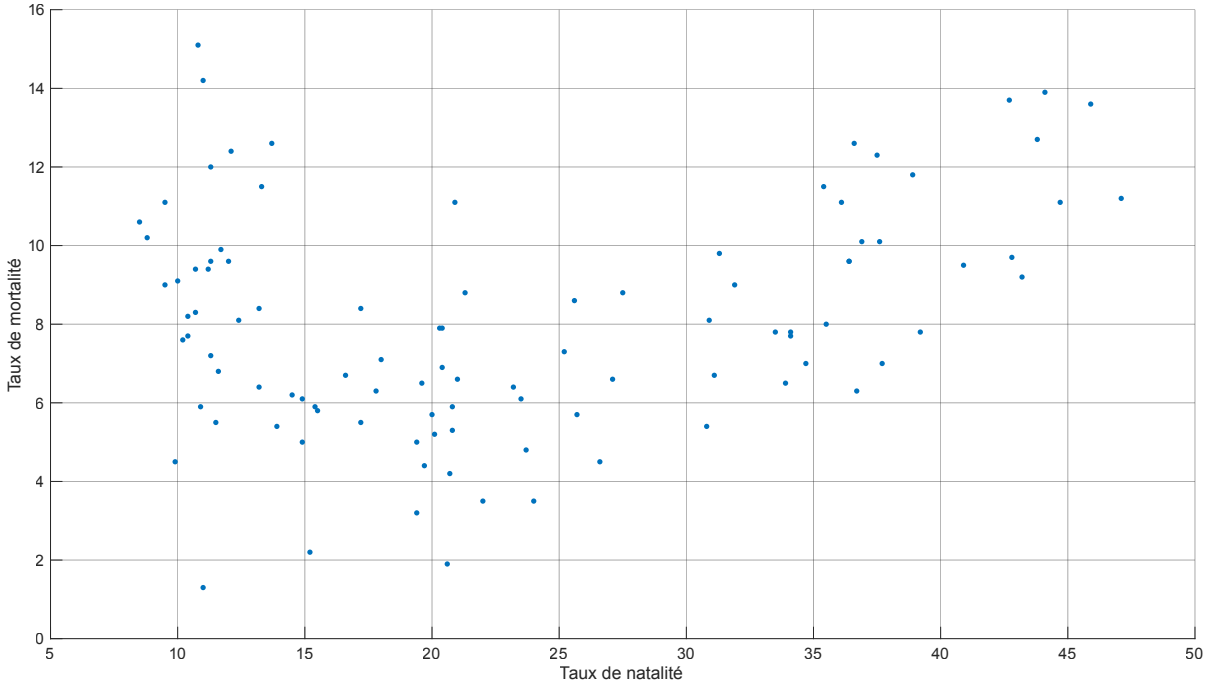

Figure 4 – Nuage de points corrélant le taux de natalité et le taux de mortalité.

Il est difficile de tirer une relation entre les deux taux sur base de cette figure. On constate tout de même que les pays ayant un taux de natalité élevé ont également tendance à avoir un taux de mortalité élevé lui aussi.

Le coefficient de corrélation entre le taux de natalité et le taux de mortalité, calculé avec `corrcoef`, vaut 0,2803. Ces deux taux sont donc faiblement corrélés.

2 Génération d'échantillons i.i.d.

Remarque préliminaire Dans toutes les sections et sous-sections suivantes concernant des échantillons i.i.d., les résultats présentés sont issus d'un tirage particulier. Ces valeurs peuvent varier en fonction du tirage effectué. Il est à noter également que, pour une cohérence dans la comparaison des résultats, les mêmes échantillons ont été utilisés pour toutes les sections travaillant sur un même tirage.

2.a Échantillon i.i.d. de 20 pays

On tire, grâce à la fonction auxiliaire `getsample`, un échantillon i.i.d. de 20 pays.

2.a.i Statistiques descriptives

Les résultats, obtenus grâce au script `Q2a`, sont explicités à la table 4.

Statistique descriptive	**Taux de natalité**	**Taux de mortalité**
Moyenne [‰]	21,0500	8,1450
Médiane [‰]	19,7500	8,1500
Écart-type [‰]	11,6061	2,6541

Table 4 – Statistiques descriptives d'un échantillon i.i.d. de 20 pays.

En comparant ces valeurs avec celles de la population (table 1), on constate qu'elles sont plutôt proches l'une de l'autre. Les variations par rapport aux résultats de la population sont encore trop élevées que pour considérer ces valeurs comme précises. Cependant, dans un cadre général, celles-ci permettent de dégager une tendance et sont donc tout à fait acceptables.

2.a.ii Boîtes à moustaches

De manière générale, les boîtes obtenues (présentées à la figure 5), via la fonction `boxplot` du script `Q2a`, s'apparentent au boîtes à moustaches de la population (figure 2).

La taille de l'échantillon étant assez petite par rapport à la population, il est normal que les valeurs des quartiles varient; surtout celles concernant le taux de natalité dont, de base, les valeurs de la population sont assez éparses. Les moustaches sont, elles aussi, plus restreintes (et seront forcément toujours plus petites ou égales à celle de la population).

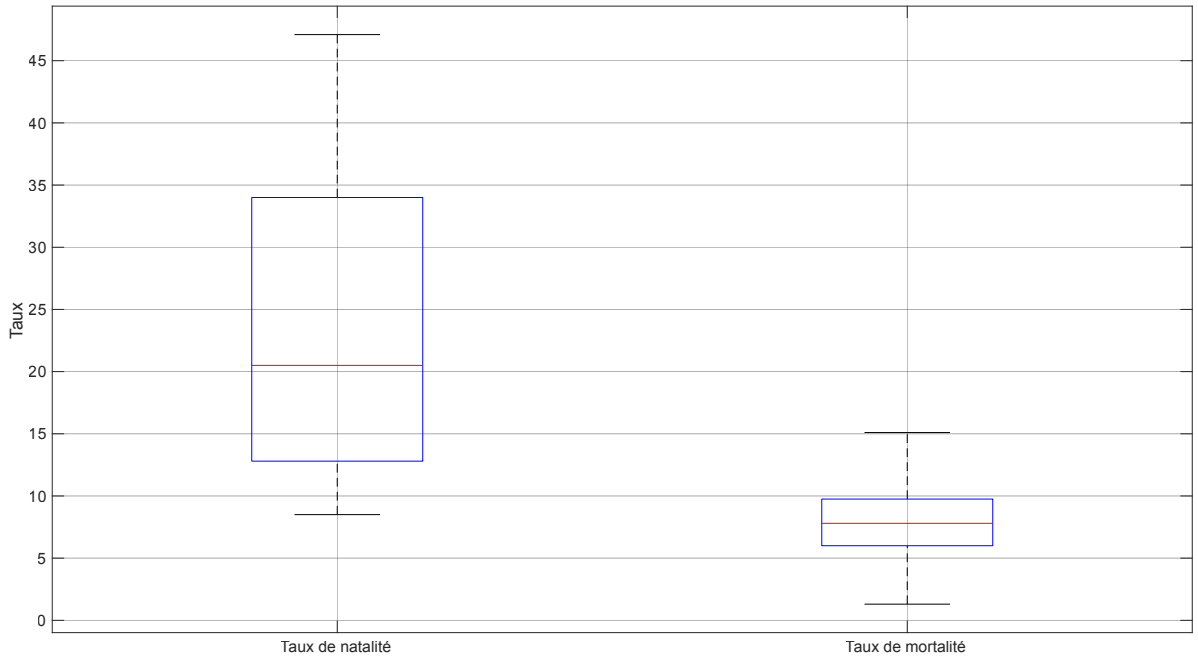

Figure 5 – Boîtes à moustaches d'un échantillon i.i.d. de 20 pays.

2.a.iii Polygones des fréquences cumulées

Les deux graphes des fréquences cumulées, obtenus via le script `Q2a`, sont présentés à la figure 6.

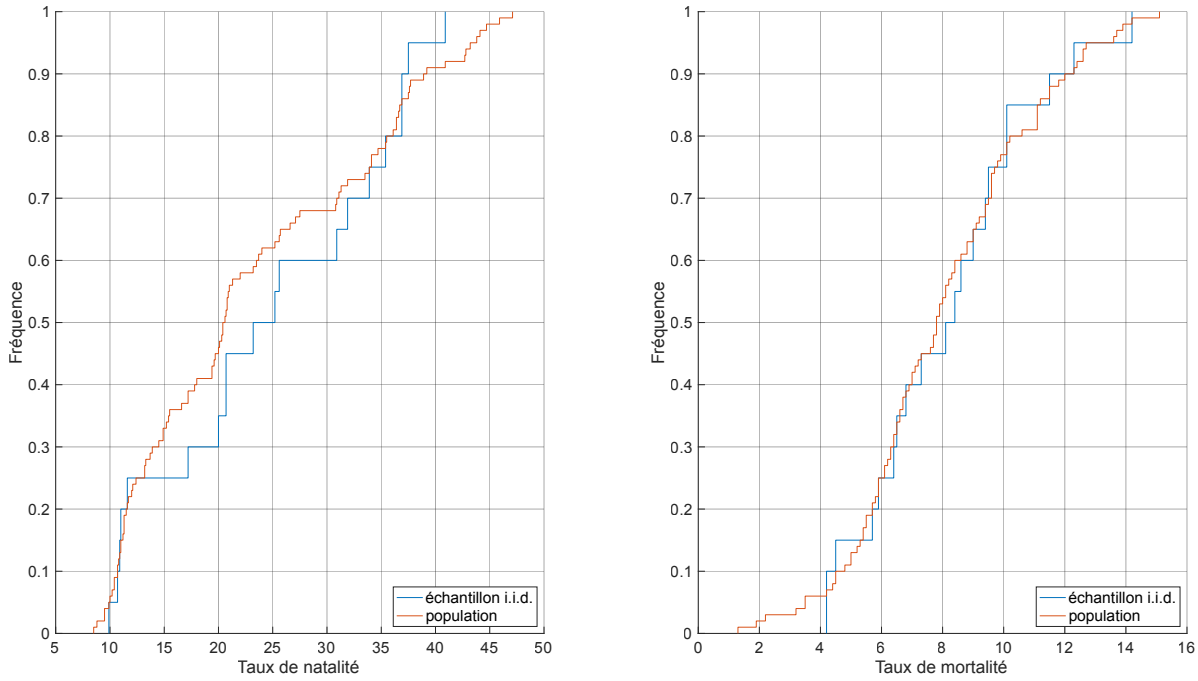

Figure 6 – Polygones des fréquences cumulées du taux de natalité d'un échantillon i.i.d. de 20 pays.

Concernant le polygone relatif au taux de natalité, on remarque que l'allure de l'échantillon est relativement semblable à celle de la population. Cependant, le peu de valeur utilisées conduit à un graphe beaucoup moins lisse avec notamment des paliers par endroit. Cette allure dépendra de l'échantillon tiré, notamment pour la présence de paliers, mais conservera la même allure générale.

Le polygone du taux de mortalité de l'échantillon est également proche de celui de la population, avec des écarts parfois important, mais qui sont dûs, eux aussi, au nombre de valeurs utilisées.

Les deux graphes de l'échantillon se rapprochent donc de ceux de la population mais ne permettent pas de s'en affranchir et de se baser uniquement sur ceux-ci.

Les distances de Kolmogorov-Smirnov sont explicitées à la table 5.

Taux	**Distance de Kolmogorov-Smirnov**
Taux de natalité	0,2400
Taux de mortalité	0,1100

Table 5 – Distances de Kolmogorov-Smirnov de l'échantillon i.i.d.

Ces valeurs varient entre 0,1500 et 0,2500 en fonction du tirage, confirmant ainsi les écarts observables sur la figure 6. On peut noter que ces valeurs ne seront jamais nulles puisqu'un échantillon de 20 pays ne coïncidera jamais exactement avec la population de 100 pays. Ces distances sont relativement faibles, confirmant ainsi le fait que les résultats de l'échantillons sont acceptables.

2.b 500 échantillons i.i.d. de 20 pays

On tire, grâce à la fonction auxiliaire `getsample`, 500 échantillons i.i.d. de 20 pays. Les figures et résultats explicités dans les sous-sections suivantes ont été obtenus par le script `Q2b`.

2.b.i Étude du taux moyen de natalité et de mortalité

L'allure des deux histogrammes se rapproche d'une loi normale comme on peut le voir à la figure 7.

Les moyennes des nouvelles variables sont données à la table 6.

En comparaison avec les taux moyens de la population, ces deux moyennes de taux moyens des échantillons sont toutes deux extrêmement proches des valeurs de la population. Alors que les taux pour un échantillon étaient acceptables par rapport à ceux de la population, ceux calculés sur 500 échantillons sont très précis. Ce résultat est logique puisque l'on travaille sur un nombre de données beaucoup plus grand, donc se rapprochant beaucoup plus de la population.

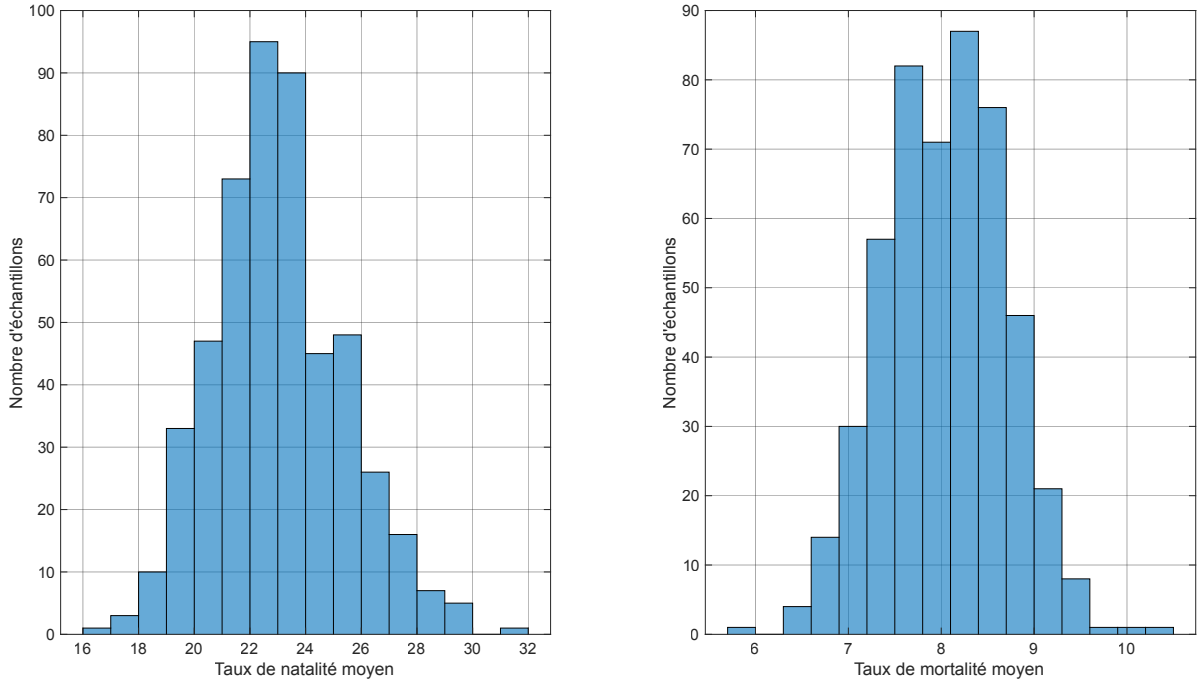

Figure 7 – Histogrammes des taux moyens de 500 échantillons i.i.d.

Taux	**Valeur moyenne des échantillons** [‰]
Taux de natalité	23,1286
Taux de mortalité	8,0051

Table 6 – Valeurs moyennes des taux moyens de 500 échantillons i.i.d.

2.b.ii Étude de la médiane du taux de natalité et de mortalité

Les deux histogrammes des médianes sont visibles à la figure 8. De nouveau, on distingue clairement des allures de loi normales. À noter que celle du taux de natalité est fort piquée.

Les moyennes des nouvelles variables sont données à la table 7.

Taux	**Valeur moyenne des échantillons** [‰]
Taux de natalité	20,8635
Taux de mortalité	7,7869

Table 7 – Valeurs moyennes des médianes de 500 échantillons i.i.d.

On constate que ces valeurs sont moins proches des moyennes de la population que les valeurs calculées à la section 2.b.i. La médiane est donc, dans ce cas, un moins bon estimateur du taux moyen de la population que la moyenne.

On note que cela n'est pas toujours le cas, les données aberrantes pouvant influencer les

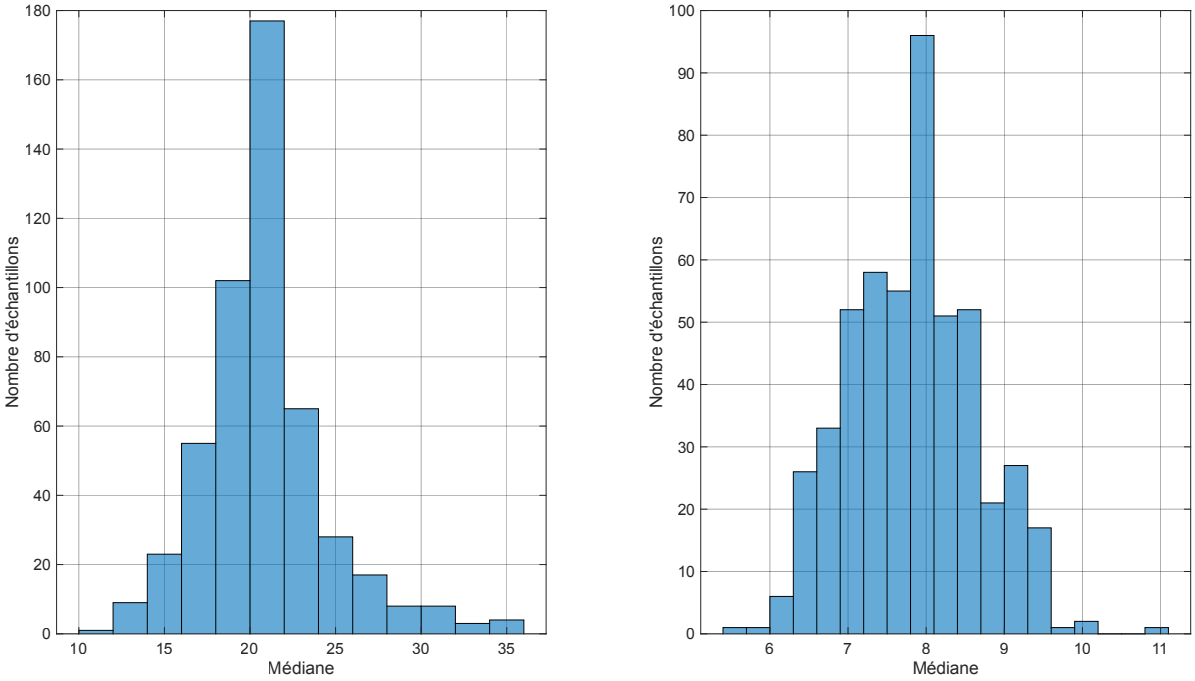

Figure 8 – Histogrammes des médianes de 500 échantillons i.i.d.

résultats. En effet, la médiane étant peu touchée par les données aberrantes, elle est souvent meilleur estimateur que la moyenne. Cependant, dans la base de données étudiée ici, il n'y a pas de données aberrantes.

2.b.iii Étude de l'écart-type du taux de natalité et mortalité

Les deux graphes sont présentés à la figure 9.

Bien que leurs allures fassent penser à une loi normales, ce n'est pas le cas ici. En effet, il a été vu que le carré de s_X suit une loi *Chi-carré* à $n-1$ degrés de liberté. Il n'est donc pas possible que s_X suive une loi normale.

Les moyennes des nouvelles variables sont données à la table 8.

Taux	**Valeur moyenne des échantillons** [‰]
Taux de natalité	10,8078
Taux de mortalité	2,7802

Table 8 – Valeurs moyennes des écarts-types de 500 échantillons i.i.d.

Les valeurs moyennes des écarts-types des taux de natalité et de mortalité de 500 échantillons se rapprochent des écarts-types de la population. On remarque cependant que ces estimations sous-estiment les valeurs de la population. Pour amoindrir cela, on aurait pu utiliser les variances corrigées s^2_{n-1} des échantillons, qui sont des estimateurs non biaisés

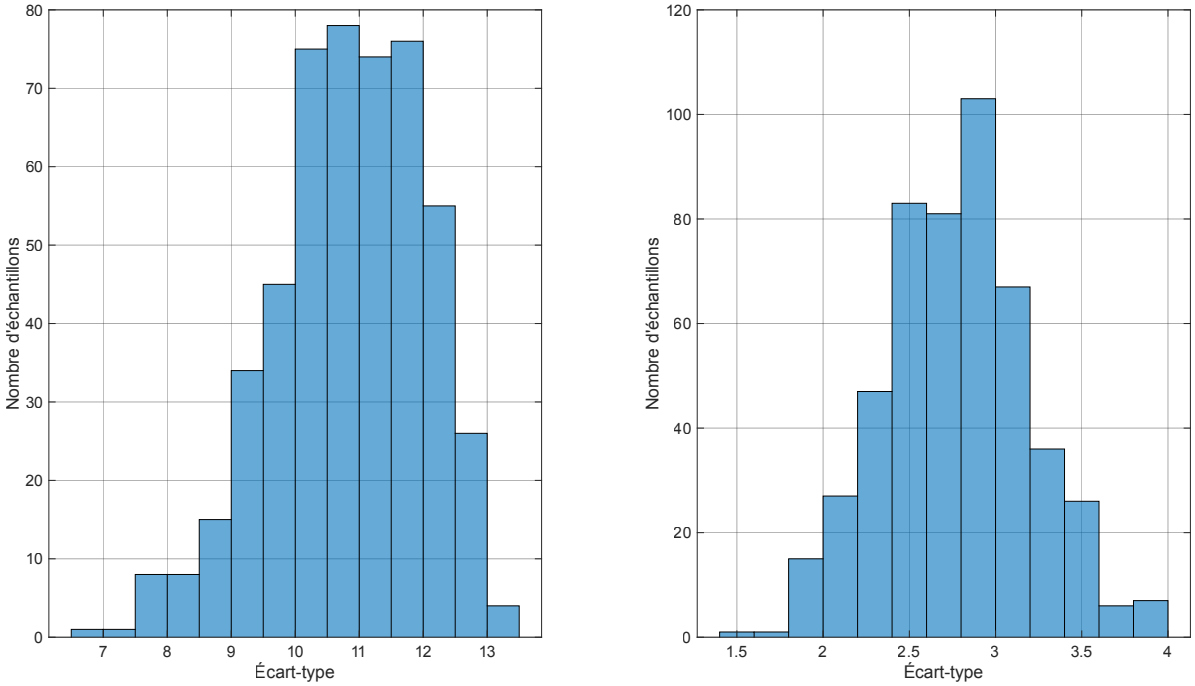

Figure 9 – Histogrammes des écarts-types de 500 échantillons i.i.d.

de la variance de la population. Cependant, vu l'inégalité de Jensen, l'écart-type s_{n-1} sous estime également l'écart-type de la population.

2.b.iv Étude de la distance de Kolmogorv-Smirnov

Les graphes présentés à la figure 10 explicitent les distances de Kolmogorov-Smirnov entre les polygones des fréquences cumulées de la population et des échantillons.

L'allure de ces graphes s'apparente à une loi normale, mais légèrement asymétrique par rapport aux valeurs les plus représentées. Concernant le taux de naissance, la distance la plus courante semble valoir environ 0,15, signifiant, qu'en moyenne, il y a 0,15 d'écart entre la population et un échantillon aléatoire i.i.d. quelconque de 20 pays. Pour le taux de mortalité, cet écart moyen semble être d'environ 0,8.

3 Estimation

On tire, grâce à la fonction auxiliaire `getsample`, 100 échantillons i.i.d. de 20 pays. On ne considère, pour les sous-sections suivantes, que le taux de natalité.

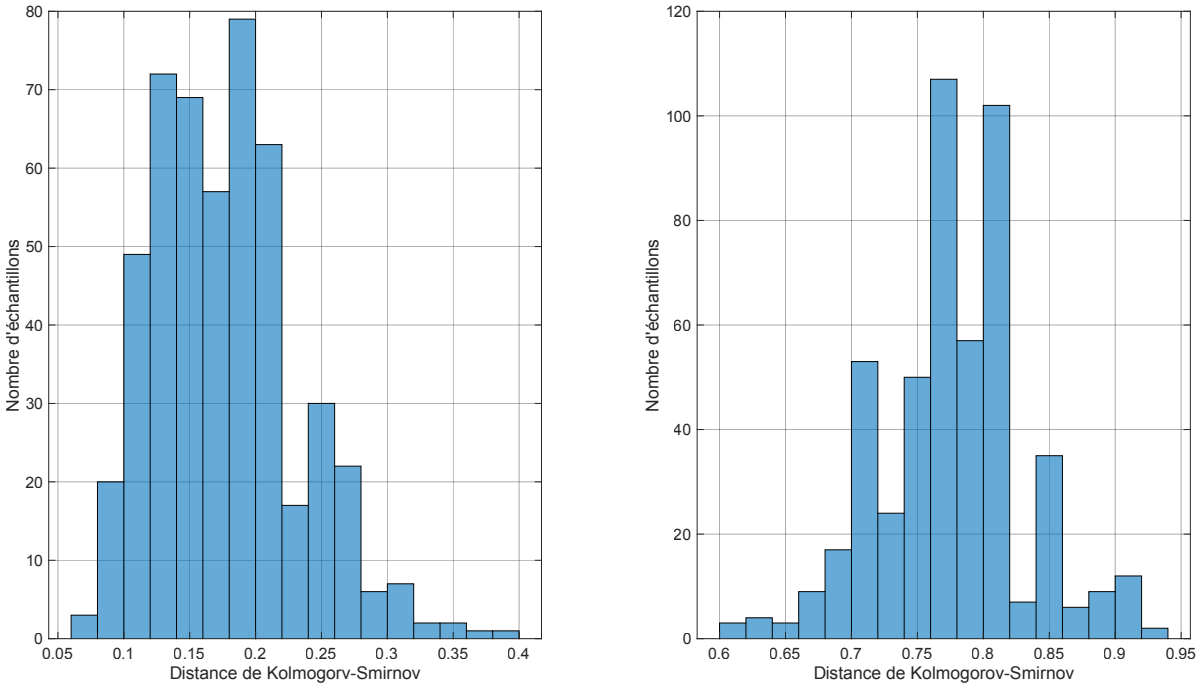

Figure 10 – Histogrammes des distances de Kolmogorov-Smirnov de 500 échantillons i.i.d.

3.a Utilisation de la moyenne comme estimateur

On choisit la moyenne m_X des échantillons comme estimateur. Le script `Q3abd` calcule ces moyennes et les enregistre dans une nouvelle variable. Le biais et la variance de l'estimateur sont calculés par le biais de cette nouvelle variable.

L'erreur quadratique d'une estimation est donnée par

$$\begin{aligned} E\left\{(\mathcal{T}_n - \theta^*)^2\right\} &= V\left\{\mathcal{T}_n\right\} + \left(E\left\{\mathcal{T}_n - \theta^*\right\}\right)^2 \\ &= \text{Variance} + \text{Biais}^2 \end{aligned} \tag{1}$$

avec

- $\mathcal{T}_n$, l'estimateur;
- θ^*, la vraie valeur à estimer.

Dans ce cas-ci, l'estimateur est la moyenne m_X et la vraie valeur à estimer est la moyenne μ de la population.

On peut alors, en reprenant les expression de l'équation (1), calculer le biais et la variance de l'estimateur m_X:

- biais: $\left|E\left\{m_X - \mu\right\}\right|$;
- variance: $V\left\{m_X\right\} = E\left\{\left(m_x - E\left\{m_x\right\}\right)^2\right\}$.

La moyenne μ de la population étant connue (calculée à la section 1.b), on obtient les résultats présentés à la table 9.

Estimateur	**Biais** [‰]	**Variance** [‰^{2}]
Moyenne m_X	0,1254	6,4093

Table 9 – Biais et variance de l'estimateur m_X du taux de natalité moyen de la population.

Comme attendu au vu des résultats de la section 2.b.i, on constate que le biais de l'estimateur moyenne est très faible par rapport à la moyenne de la population. La variance, quant à elle, semble tourner autour de la valeur 6,5. L'écart-type, obtenu en prenant la racine carrée de cette valeur, est plutôt petit traduisant un étalement assez faible des valeurs de m_X.

L'erreur quadratique peut également être calculée et vaut 7,8014.

3.b Utilisation de la médiane comme estimateur

On choisit la médiane $median_X$ des échantillons comme estimateur. Le script `Q3abd` calcule ces médianes et les enregistre dans une nouvelle variable. Le biais et la variance de l'estimateur sont calculés par le biais de cette nouvelle variable.

En procédant de la même manière qu'à la section 3.a, on trouve:

- biais: $|E\{median_X - \mu\}|$;
- variance: $V\{median_X\} = E\left\{(median_x - E\{median_x\})^2\right\}$.

On obtient alors les résultats présentés à la table 10.

Estimateur	**Biais** [‰]	**Variance** [‰^{2}]
Médiane $median_X$	0,7430	15,6342

Table 10 – Biais et variance de l'estimateur $median_X$ du taux de natalité moyen de la population.

On constate que le biais et la variance de $median_X$ sont plus élevés que ceux de m_X. L'étalement des valeurs de $median_X$ est plus important.

L'erreur quadratique vaut, quant à elle, 15,1499. À première vue, on observe que $median_X$ est un moins bon estimateur que m_X.

3.c Estimateurs sur des échantillons i.i.d. de taille 50

En procédant de la même manière qu'aux sections 3.a et 3.b avec des échantillons i.i.d. de 50 pays, on obtient, via le script `Q3c`, les résultats présentés à la table 11.

On constate tout d'abord que le biais et la variance de l'estimateur m_X sont plus petits que ceux de $median_X$, confirmant l'observation de la section 3.b comme quoi m_X est un meilleur estimateur que $median_X$.

Estimateur	**Biais** [‰]	**Variance** $\left[‰^2\right]$
Moyenne m_X	0,0796	2,2881
Médiane $median_X$	0,1740	3,0515

Table 11 – Biais et variance des estimateurs m_X et $median_X$ du taux de natalité moyen de la population pour des échantillons i.i.d. de 50 pays.

On constate également que les 4 valeurs obtenues sont plus petites que celles calculées aux sections 3.a et 3.b, traduisant des résultats d'une précision plus importante. Cette diminution des valeurs est la conséquence logique de l'augmentation de la taille des échantillons.

Plus la taille de l'échantillon augmente, plus le biais de l'estimateur m_X diminue et tend en fait vers 0. La moyenne de cet estimateur tend donc vers la moyenne de la population.

Enfin, on remarque que, pour chacun des estimateurs, la variance diminue lorsque la taille de l'échantillon augmente. En effet, il parait logique que les valeurs des moyennes et médianes soient moins dispersées lorsque l'on considère de plus grands échantillons.

3.d Construction d'intervalles de confiance

Dans cette partie, on travaille de nouveau avec 100 échantillons i.i.d. de 20 pays. Les données des sous-sections suivantes ont été obtenues grâce au script `Q3abd`.

Les intervalles de confiance, centré sur m_X, étant à 95% du taux de natalité de la population, on peut d'ors et déjà définir que le paramètre α vaut $1 - 0{,}95 = 0{,}05$.

3.d.i Loi de Student

L'intervalle de confiance construit avec la loi de Student est donné par

$$\left[m_X - t_{1-\frac{\alpha}{2}} \frac{s_{n-1}}{\sqrt{n}}; m_X + t_{1-\frac{\alpha}{2}} \frac{s_{n-1}}{\sqrt{n}}\right]$$

En prenant un degré de liberté égale à la taille des échantillons - 1, à savoir 19, dans le cas demandé, on a

- $n = 20$, la taille des échantillons;
- m_X, la moyenne des échantillons;
- $t_{1-\frac{\alpha}{2}} = t_{0{,}975} = 2{,}093$, calculé avec la fonction `tinv`. Cette valeur pourrait également être lue dans une table de données;
- s_{n-1}, l'écart-type corrigé des échantillons, calculée avec la fonction `std`.

On peut dès lors construire un intervalle de confiance pour chaque échantillon. On constate, en effectuant plusieurs tests, que le nombre d'intervalles de confiance contenant la valeur de la population varie entre 88 et 98.

3.d.ii Loi de Gauss

L'intervalle de confiance construit avec la loi de Gauss est donné par

$$\left[m_X - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; m_X + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

Dans le cas demandé, on a

- $n = 20$, la taille des échantillons;
- m_X, la moyenne des échantillons;
- $u_{1-\frac{\alpha}{2}} = u_{0,975} = 1{,}96$, calculé avec la fonction `normiv`. Cette valeur pourrait également être lue dans une table de données;
- σ, l'écart-type de la population, calculée à la section 1.b.

On peut dès lors construire un intervalle de confiance pour chaque échantillon. On constate, dans ce cas, que le nombre d'intervalles de confiance contenant la valeur de la population varie entre 93 et 99.

Bien que le nombre d'intervalles contenant la valeur de la population varie dans les deux cas, il semblerait que les intervalles construits avec la loi de Gauss sont, de manière générale, plus proche des 95% attendus.

Une loi de Student utilise moins d'informations qu'une loi de Gauss et est donc par conséquent plus générale et moins précise. Au vu des résultats des intervalles construits avec la loi de Gauss, il semblerait qu'il était raisonnable de supposer que la variable parente était Gaussienne.

4 Tests d'hypothèse - proportion

On tire, grâce à la fonction auxiliaire `getsample`, 100 fois 5 échantillons i.i.d. de 40 pays. Le premier échantillon de chaque tirage est celui de la Belgique; les autres sont ceux des instituts. Tous les résultats des sous-sections suivantes ont été obtenus par le biais du script `Q4ab`.

Au vu des hypothèses à tester, il s'agit de réaliser un test d'hypothèses unilatéral à droite sur une proportion. L'hypothèse H_0 est la suivante: "la proportion de pays ayant un taux de natalité plus faible que la Belgique est de $x\%$".

Premièrement, on calcule cette valeur x, à savoir la vraie proportion de pays, dans la population, ayant un taux de natalité plus faible que la Belgique. On trouve $x = 21\%$.

Deuxièmement, on calcule la borne supérieure de l'intervalle favorable à H_0, avec un seuil de signification de $\alpha = 5\%$ et une taille d'échantillon $n = 40$:

$$\begin{aligned} x + u_{0,975}\sqrt{\frac{x\,(1-x)}{n}} &= 0{,}21 + 1{,}96\sqrt{\frac{0{,}21\,(1-0{,}21)}{40}} \\ &= 0{,}3362 \end{aligned}$$

Troisièmement, pour chaque échantillon de chaque tirage, on calcule la proportion p de pays ayant un taux de natalité inférieur à celui de la Belgique. L'hypothèse H_0 sera rejetée si la proportion p est supérieure à la valeur de la borne supérieure, c'est-à-dire si $p > 0.3362$.

Si on définit une variable aléatoire $\mathcal{X}$ qui vaut 1 si un pays tiré au hasard possède un taux de natalité inférieur à celui de la Belgique et qui vaut 0 dans les autres cas, cette variable $\mathcal{X}$ sera une variable binomiale. De ce fait, si l'hypothèse H_0 est vérifiée, la moyenne et l'écart-type de $\mathcal{X}$ seront donnés par:

- moyenne: x;
- écart-type: $\sqrt{x\,(1-x)}$.

Lors de la construction de la borne supérieur de l'intervalle, on a approximé une loi binomiale par une loi normale. Cela n'est valable que si $\min\left(nx; n\,(1-x)\right) \geq 5$, ce qui est bien le cas.

4.a Rejet de l'hypothèse par l'État belge

De multiples tests montrent que l'État belge rejette l'hypothèse entre 0 et 7 fois sur 100. En moyenne, le pourcentage de rejet est proche du seuil α fixé mais lui reste néanmoins pratiquement toujours inférieur. Cette différence par rapport à α peut être due à l'approximation de la loi binomiale par une loi normale.

4.b Rejet de l'hypothèse par l'O.M.S.

Afin de connaître le nombre de rejet de l'hypothèse H_0 par l'O.M.S., on comptabilise le nombre de tirages où au moins un des quatres instituts a rejeté l'hypothèse H_0.

De multiples tests montrent que le nombre de fois où l'O.M.S. a considéré que les belges n'ont pas un faible taux de natalité varie entre 5 et 17 fois sur 100.

On constate premièrement que cette proportion est, en moyenne, supérieure à α. Ce résultat parait logique puisque les 4 instituts possèdent chacun un échantillon différent et qu'il suffit qu'un seul des 4 rejette l'hypothèse. La probabilité que cela arrive est supérieure à la probabilité que l'État belge rejette l'hypothèse.

En effet, la probabilité qu'un échantillon rejette l'hypothèse étant de $\alpha = 0{,}05$, on calcule

$$1 - (1 - 0.05)^4 = 0{,}1855 > \alpha \tag{2}$$

On constate néanmoins que la probabilité (2) calculée ne correspond pas tout à fait au nombre de fois où l'O.M.S. a considéré que l'État belge avait un faible taux de naissance. Cela reflette, une fois de plus, l'imprécision des résultats due à l'approximation de la loi binomiale par une loi normale.

4.c Méthodes alternatives

Pour éviter que les instituts de statistique indépendants soient avantagés par rapport à l'État belge, plusieurs méthodes auraient pu être utilisées:

- on aurait pu imposer à chaque institut de travailler sur le même échantillon;
- on aurait pu, tout en gardant le seuil de signification α pour l'État belge, diminuer le seuil de signification des autres instituts, réduisant ainsi la taille de leur intervalle favorable à H_0. On pourrait, par exemple, choisir un seuil tel que la probabilité (2) soit égale à la probabilité que l'État belge rejette l'hypothèse H_0;
- à l'inverse, on aurait pu augmenter le seuil de signification de l'État belge tout en gardant le seuil α pour les autres instituts;
- on aurait pu également augmenter le nombre de données sur lesquelles a travaillé l'État belge, c'est-à-dire soit augmenter la taille de l'échantillon, soit travailler sur plusieurs échantillons. En effet, plus elle aura accès à un grand nombre de données, plus les résultats trouvés seront proches de la réalité et donc fiables, réduisant ainsi les inégalités avec les autres instituts.

À noter que les solutions augmentant un seuil de signification réduisent la fiabilité des résultats.

Code Matlab

Les codes Matlab se trouvant dans des dossiers différents, il est impératif de lancer le script `startup`[4] en premier.

Scripts

Question 1

```
%% Loading data
loaddata;

%% Histograms
figure;

% Birth rate
subplot(1, 2, 1)
histogram(birth, min(birth):1:(max(birth) + 1))

% Death rate
subplot(1, 2, 2)
histogram(death, min(death):1:(max(death) + 1))

%% Deleting unnecessary variables
clearvars
```

Listing 1 – Script `Q1a`.

```
%% Loading data
loaddata;

%% Descriptive statistics
% Mean
birth_mean = mean(birth);
death_mean = mean(death);

% Median
birth_median = median(birth);
death_median = median(death);

% Mode
birth_mode = mode(birth);
death_mode = mode(death);

% Standard deviation
birth_std = std(birth, 1);
death_std = std(death, 1);

%% Belgium
birth_be = birth(9, 1);
```

[4]Théoriquement, Matlab exécute ce script automatiquement à l'ouverture du projet, mais cela ne pourrait pas être le cas.

```
23 death_be = death(9, 1);
24
25 %% Deleting unnecessary variables
26 clearvars -except data birth death birth_mean death_mean
       birth_median...
27     death_median birth_mode death_mode birth_std death_std birth_be...
28     death_be
```

Listing 2 – Script `Q1b`.

```
1  %% Loading previous data
2  Q1b;
3
4  %% Normal rates and country proportions
5  countries_number = size(data, 1);
6
7  % Birth
8  birth_normal = [birth_mean - birth_std, birth_mean + birth_std];
9
10 birth_number = sum(birth >= birth_normal(1) & birth <=
       birth_normal(2), 1);
11 birth_proportion = birth_number / countries_number;
12
13 % Death
14 death_normal = [death_mean - death_std, death_mean + death_std];
15
16 death_number = sum(death >= death_normal(1) & death <=
       death_normal(2), 1);
17 death_proportion = death_number / countries_number;
18
19 %% Belgium
20 birth_normal_be = birth_be >= birth_normal(1) & birth_be <=
       birth_normal(2);
21 death_normal_be = death_be >= death_normal(1) & death_be <=
       death_normal(2);
22
23 %% Deleting unnecessary variables
24 clearvars -except birth_normal birth_proportion death_normal...
25     death_proportion birth_normal_be death_normal_be
```

Listing 3 – Script `Q1c`.

```
1  %% Loading data
2  loaddata;
3
4  %% Boxplots
5  figure;
6  boxplot([birth, death])
7
8  %% Quartiles
9  birth_25 = prctile(birth, 25);
10 birth_75 = prctile(birth, 75);
11
12 death_25 = prctile(death, 25);
13 death_75 = prctile(death, 75);
14
```

```
15 %% Outliers
16 birth_outliers = sum(isoutlier(birth), 1);
17 death_outliers = sum(isoutlier(death), 1);
18
19 %% Deleting unnecessary variables
20 clearvars -except birth_25 birth_75 death_25 death_75 birth_outliers...
21     death_outliers
```

Listing 4 – Script `Q1d`.

```
1 %% Loading data
2 loaddata;
3
4 %% Polygon of cumulative frequencies (birth rate)
5 figure;
6 cdfplot(birth)
7
8 %% Proportion of countries with a birth rate in [be_rate, 20]
9 be_rate = birth(9, 1);
10
11 [f, x] = ecdf(birth);
12 birth_proportion = f(x == 20) - f(x == be_rate);
13
14 %% Deleting unnecessary variables
15 clearvars -except birth_proportion
```

Listing 5 – Script `Q1e`.

```
1 %% Loading data
2 loaddata;
3
4 %% Scatter plot
5 figure;
6 scatter(birth, death)
7
8 %% Correlation coefficient
9 corr_coef_m = corrcoef(birth, death);
10 corr_coef = corr_coef_m(1, 2);
11
12 %% Deleting unnecessary variables
13 clearvars -except corr_coef
```

Listing 6 – Script `Q1f`.

Question 2

```
1 %% Loading data
2 loaddata;
3
4 %% Sample of 20 countries
5 sample_set = getsample(1, 20, data);
6
7 sample_birth = sample_set{1, 1}(:, 1);
8 sample_death = sample_set{1, 1}(:, 2);
```

```
 9
10 %% Descriptive statistics
11 % Mean
12 birth_mean = mean(sample_birth);
13 death_mean = mean(sample_death);
14
15 % Median
16 birth_median = median(sample_birth);
17 death_median = median(sample_death);
18
19 % Standard deviation
20 birth_std = std(sample_birth, 1);
21 death_std = std(sample_death, 1);
22
23 %% Boxplots
24 figure;
25 boxplot([sample_birth, sample_death])
26
27 %% Polygon of cumulative frequencies
28 % Birth rate
29 figure;
30
31 subplot(1, 2, 1)
32
33 hold on
34
35 cdfplot(sample_birth) % sample
36 cdfplot(birth) % population
37
38 hold off
39
40 % Death rate
41 subplot(1, 2, 2)
42
43 hold on
44
45 cdfplot(sample_death) % sample
46 cdfplot(death) % population
47
48 hold off
49
50 %% Kolmogorov-Smirnov distances
51 [~, ~, birth_ks] = kstest2(sample_birth, birth);
52 [~, ~, death_ks] = kstest2(sample_death, death);
53
54 %% Deleting unnecessary variables
55 clearvars -except birth_mean death_mean birth_median death_median...
56     birth_std death_std birth_ks death_ks
```

Listing 7 – Script `Q2a`.

```
1 %% Loading data
2 loaddata;
3
4 %% 500 samples of 20 countries
5 sample_number = 500;
```

```
sample_size = 20;

sample_set = getsample(sample_number, sample_size, data);

%% Means of samples
birth_sample_mean = zeros(sample_number, 1);
death_sample_mean = zeros(sample_number, 1);

for i = 1:sample_number
    birth_sample_mean(i, 1) = mean(sample_set{i}(:, 1));
    death_sample_mean(i, 1) = mean(sample_set{i}(:, 2));
end

%% Histograms of means of samples
figure;

% Birth rate
subplot(1, 2, 1)
histogram(birth_sample_mean)

% Death rate
subplot(1, 2, 2)
histogram(death_sample_mean)

%% Mean of the means of the samples
birth_mean = mean(birth_sample_mean);
death_mean = mean(death_sample_mean);

%% Median of samples
birth_sample_median = zeros(sample_number, 1);
death_sample_median = zeros(sample_number, 1);

for i = 1:sample_number
    birth_sample_median(i, 1) = median(sample_set{i}(:, 1));
    death_sample_median(i, 1) = median(sample_set{i}(:, 2));
end

%% Histograms of medians of samples
figure;

% Birth rate
subplot(1, 2, 1)
histogram(birth_sample_median)

% Death rate
subplot(1, 2, 2)
histogram(death_sample_median)

%% Mean of the medians of the samples
birth_median = mean(birth_sample_median);
death_median = mean(death_sample_median);

%% Standard deviation of samples
birth_sample_std = zeros(sample_number, 1);
death_sample_std = zeros(sample_number, 1);

```

```
for i = 1:sample_number
    birth_sample_std(i, 1) = std(sample_set{i}(:, 1), 1);
    death_sample_std(i, 1) = std(sample_set{i}(:, 2), 1);
end

%% Histograms of standard deviations of samples
figure;

% Birth rate
subplot(1, 2, 1)
histogram(birth_sample_std)

% Death rate
subplot(1, 2, 2)
histogram(death_sample_std)

%% Mean of the standard deviations of the samples
birth_std = mean(birth_sample_std);
death_std = mean(death_sample_std);

%% Kolmogorov-Smirnov distances
birth_ks = zeros(sample_number, 1);
death_ks = zeros(sample_number, 1);

for i = 1:sample_number
    [~, ~, birth_ks(i, 1)] = kstest2(sample_set{i}(:, 1), birth);
    [~, ~, death_ks(i, 1)] = kstest2(sample_set{i}(:, 1), death);
end

% Histograms
figure;

% Birth rate
subplot(1, 2, 1)
histogram(birth_ks)

% Death rate
subplot(1, 2, 2)
histogram(death_ks)

%% Deleting unnecessary variables
clearvars -except birth_sample_mean death_sample_mean birth_mean...
    death_mean birth_sample_median death_sample_median birth_median...
    death_median birth_sample_std death_sample_std birth_std
        death_std...
    birth_ks death_ks
```

Listing 8 – Script `Q2b`.

Question 3

```
%% Loading data
loaddata;

```

```
4  %% 100 samples of 20 countries
5  sample_number = 100;
6  sample_size = 20;
7
8  sample_set = getsample(sample_number, sample_size, data);
9
10 %% Mean (Q3a) and median (Q3b)
11 birth_mean = zeros(sample_number, 1);
12 birth_median = zeros(sample_number, 1);
13
14 for i = 1:sample_number
15     birth_mean(i, 1) = mean(sample_set{i, 1}(:, 1));
16     birth_median(i, 1) = median(sample_set{i, 1}(:, 1));
17 end
18
19 %% Bias and variance
20 % Question 3 - a)
21 mean_bias = mean(birth_mean - mean(birth));
22 mean_var = var(birth_mean, 1);
23
24 % Question 3 - b)
25 median_bias = mean(birth_median - median(birth));
26 median_var = var(birth_median, 1);
27
28 %% Confidence intervals (Q3d)
29 % General data
30 pop_mean = mean(data(:, 1));
31 p = 0.95;
32 alpha = 1 - p;
33
34 % Student law
35 birth_student_count = 0;
36
37 t = tinv(1 - (alpha / 2), sample_size - 1);
38
39 birth_student_ci = zeros(sample_number, 2);
40
41 for i = 1:sample_number
42     part(1) = mean(sample_set{i, 1}(:, 1));
43     part(2) = t * (std(sample_set{i, 1}(:, 1), 0) / sqrt(sample_size));
44
45     birth_student_ci(i, 1) = part(1) - part(2);
46     birth_student_ci(i, 2) = part(1) + part(2);
47
48     if (pop_mean >= birth_student_ci(i, 1)) && (pop_mean <=
       birth_student_ci(i, 2))
49         birth_student_count = birth_student_count + 1;
50     end
51 end
52
53 % Gaussian law
54 birth_gaussian_count = 0;
55
56 u = norminv(1 - (alpha / 2));
57
58 birth_gaussian_ci = zeros(sample_number, 2);
```

```
59
60 for i = 1:sample_number
61     part(1) = mean(sample_set{i, 1}(:, 1));
62     part(2) = u * (std(birth, 1) / sqrt(sample_size));
63
64     birth_gaussian_ci(i, 1) = part(1) - part(2);
65     birth_gaussian_ci(i, 2) = part(1) + part(2);
66
67     if (pop_mean >= birth_gaussian_ci(i, 1)) && (pop_mean <=
           birth_gaussian_ci(i, 2))
68         birth_gaussian_count = birth_gaussian_count + 1;
69     end
70 end
71
72 %% Deleting unnecessary variables
73 clearvars -except mean_bias mean_var median_bias median_var...
74     birth_student_ci birth_student_count birth_gaussian_ci...
75     birth_gaussian_count
```

Listing 9 – Script `Q3abd`.

```
1  %% Loading data
2  loaddata;
3
4  %% 100 samples of 50 countries
5  sample_number = 100;
6  sample_size = 50;
7
8  sample_set = getsample(sample_number, sample_size, data);
9
10 %% Mean and median
11 birth_mean = zeros(sample_number, 1);
12 birth_median = zeros(sample_number, 1);
13
14 for i = 1:sample_number
15     birth_mean(i, 1) = mean(sample_set{i, 1}(:, 1));
16     birth_median(i, 1) = median(sample_set{i, 1}(:, 1));
17 end
18
19 %% Bias and variance
20 mean_bias = mean(birth_mean - mean(birth));
21 mean_var = var(birth_mean, 1);
22
23 median_bias = mean(birth_median - median(birth));
24 median_var = var(birth_median, 1);
25
26 %% Deleting unnecessary variables
27 clearvars -except mean_bias mean_var median_bias median_var
```

Listing 10 – Script `Q3c`.

Question 4

```
1 %% Loading data
```

```
loaddata;

%% 100 times 5 samples of 40 countries
sample_time = 100;
sample_number = 5;
sample_size = 40;

sample_set = cell(sample_time, 1);

for i = 1:sample_time
    sample_set{i, 1} = getsample(sample_number, sample_size, data);
end

%% Parameters
alpha = 0.05;
u = norminv(1 - (alpha / 2));

%% True proportion of countries with a birth rate lower than Belgium
birth_rate_belgium = birth(9, 1);
x = sum(birth < birth_rate_belgium) / size(birth, 1);

%% Maximum bound of the confidence interval
CI_max = x + (u * sqrt((x * (1 - x)) / sample_size));

%% Proportion of countries with a lower birth rate than Belgium for
   each sample
prop = cell(sample_time, 1);

reject_belgium = 0;
reject_OMS = 0;

for i = 1:sample_time
    for j = 1:sample_number
        prop{i, 1}(j, 1) = 0;

        for k = 1:sample_size
            if sample_set{i, 1}{j, 1}(k, 1) < birth_rate_belgium
                prop{i, 1}(j, 1) = prop{i, 1}(j, 1) + 1;
            end
        end

        prop{i, 1}(j, 1) = prop{i, 1}(j, 1) / sample_size;

        % Number of rejections
        if prop{i, 1}(j, 1) > CI_max
            if j == 1
                reject_belgium = reject_belgium + 1; % Belgium
            else
                reject_OMS = reject_OMS + 1; % OMS
                continue
            end
        end
    end
end

%% Deleting unnecessary variables
```

```
57 clearvars -except x CI_max reject_belgium reject_OMS
```

Listing 11 – Script `Q4ab`.

Scripts et fonctions auxiliaires

```
1  %% Additional function
2  % Function that pulls a number of samples i.i.d of a fixed size from
3  % the data.
4
5  % Arguments:
6  %   - number: the number of samples to pull
7  %   - sample_size: the size of samples to pull
8  %   - data: the data from which the samples are pullen
9
10 % Returned data:
11 %   - sample_set: a cell containing the samples
12
13 function sample_set = getsample(number, sample_size, data)
14
15 sample_set = cell(number, 1);
16
17 for i = 1:number
18     draw = randsample(size(data, 1), sample_size, true);
19     sample_set{i, 1} = data(draw, :);
20 end
21
22 end
```

Listing 12 – Fonction `getsample`.

```
1  %% Additional script
2  % Reads the data provided in the CSV file.
3
4  filename = 'db_stat19.csv';
5
6  data = csvread(filename, 1, 1, [1, 1, 100, 2]);
7
8  birth = data(:, 1);
9  death = data(:, 2);
10
11 clearvars filename
```

Listing 13 – Script `loaddata`.

```
1  %% Additional script
2  % Script that add to path the folders of the project.
3
4  addpath(genpath('additional'));
5  addpath('Q1');
6  addpath('Q2');
7  addpath('Q3');
8  addpath('Q4');
9  addpath(genpath('resources'));
```

Listing 14 – Script `startup`.