MATH0487 : Rapport du devoir

Romain LAMBERMONT (s190931) Arthur LOUIS (s191230)

18 décembre 2021

1 Analyse descriptive

(a) Ce tableau permet de représenter aisément les données de notre population. Les trois variables représentées ci-dessous sont :

- Top10 : Proportion du revenu national détenu par les 10% les plus riches.
- CO2 / habitant
- PIB / habitant

Pays	Top 10%	CO2 / habitant	PIB / habitant
USA	0.4546	17.061747	47757.5109
Belgique	0.3289	15.282899	39506.0410
Chine	0.4166	6.535790	15417.9174
Togo	0.4798	0.998398	1234.2999

Table 1 – Données extraites de **data.csv** pour les USA, la Belgique, la Chine et le Togo

En analysant les données, on remarque que disparités entre pauvres et riches sont moins marquées en Belgique quae dans les autres pays, et également que la Belgique et les USA sont les pays les plus polluants et riches par rapport à la taille de leur population. En effet, même si la Chine peut paraître moins polluante et moins riche, elle compte largement plus d'habitants que les deux pays précédents. Pour ce qui est du Togo, qu'on peut comparer avec Belgique (population semblable), qu'ils sont largement moins polluants et riches.

(b) i. Dans ce tableau, on retrouve l'écart-type et la moyenne des variables explicitées précédemment. Pour ce qui est du "Top10", on remarque que les disparités sont généralement élevées et les valeurs restent proche de 0.45. Par contre pour ce qui est du CO2 et PIB par habitant, les valeurs sont beaucoup moins concentrées au vu de l'écart type extrêmement élevé (même plus grand que la moyenne) qui montre une répartition disparate des donnée.

	Moyenne	Écart-Type
Top10	0.450072	0.089464
CO2 / habitant	5.241130	5.632340
PIB / habitant	19057.331583	27206.714860

Table 2 – Moyenne et écart-type des variables de **data.csv**

ii. Dans ce tableau, on retrouve la médiane et les quartiles des variables.

	Médiane	1er Quartile	3ème Quartile
Top10	0.4547	0.3792	0.49475
CO2 / habitant	3.00205112	0.893557638	7.99829507
PIB / habitant	11053.4877	3984.48210	25457.3043

Table 3 – Médiane et quartiles des données de **data.csv**

On réunit ensuite ces données dans des graphiques appelés "boites à moustache" qui permettent de se répresenter facilement la médiane et les quartiles.

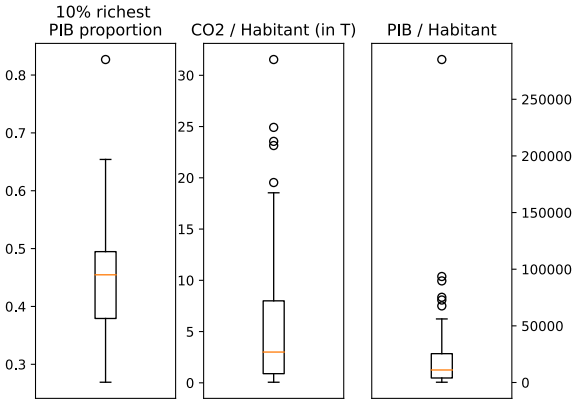

FIGURE 1 – Boite à moustache des données de **data.csv**

En analysant cette boîte à moustache on remarque bien des données aberrantes qui très éloignées des quartiles. On peut pour ça calculer l'intervalle de validité des données (les minima = 0 sont choisis car des résultats négatifs, logiquement impossibles ont été calculés) :

- Top10 $\in [0.2058749; 0.668075]$
- CO2 / habitant $\in [0; 18.6554]$
- PIB / habitant $\in [0; 57666.537475]$

En dehors de ces intervalles on peut conclure que les En comparant les boîtes à moustache des trois variables, on se rend compte qu'ils sont assez semblables (en prenant en compte l'échelle, au niveau de la répartition autour de la médiane et également au niveau des données aberantes qui sortent de notre intervalle de validité.

iii. On retrouve dans les six graphiques ci-dessous, on retrouve les histogrammes et les fonctions de répartition de chaque variable :

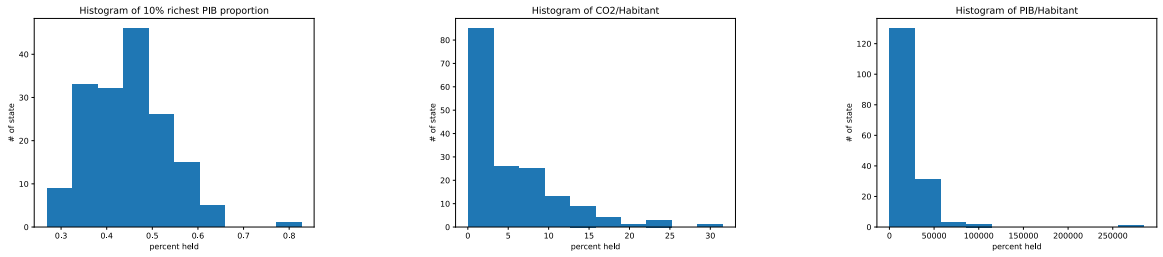

FIGURE 2 – default

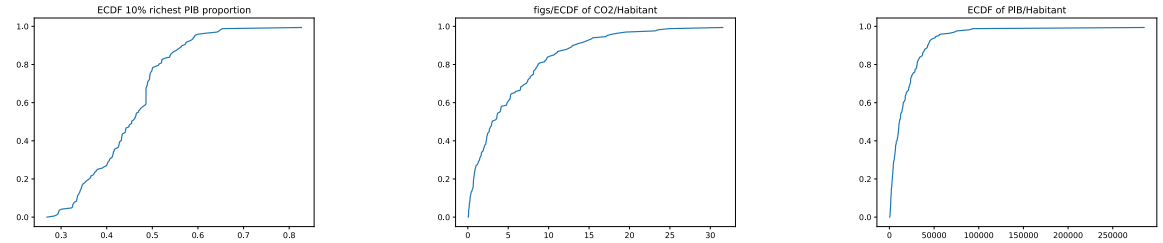

FIGURE 3 – default

En comparant les histogrammes des trois variables on se rend compte que la variable Top10 est beaucoup plus distribuée que les deux autres et cela se voit clairement sur les distributions

des variables. En effet, on voit bien que la courbe de répartition de la variable Top10 grimpe plus doucement que les deux autres variables

(c) Pour analyser les relations entre les variables, on décide de mettre en place un graphique de type "matrice" comme ci-dessous :

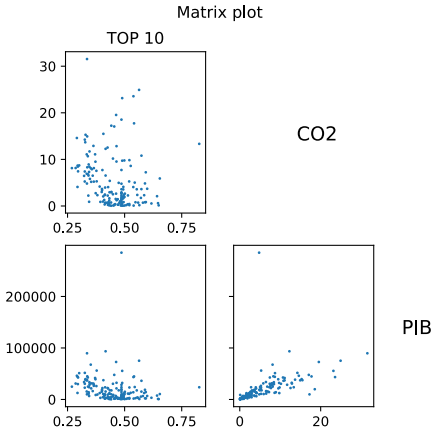

FIGURE 4 – Graphique "matrice" des variables de **data.csv**

En observant ce graphique "matrice", on remarque que il existe une relation linéaire entre le PIB/habitant et le CO2/habitant dans un pays. En effet, on remarque une droite dans le graphique en bas à droite. Pour ce qui est des autres relations, on ne peut rien distinguer de remarquable.

2 Estimation ponctuelle

(a) $\begin{cases} E = \frac{\hat{a}}{\hat{a}+\hat{b}} & (1) \\ V = \frac{\hat{a}\hat{b}}{(\hat{a}+\hat{b})^2(\hat{a}+\hat{b}+1)} & (2) \end{cases}$

En utilisant (1) :

$$\begin{aligned} E(\hat{a}+\hat{b}) &= \hat{a} \\ E\hat{a} + E\hat{b} &= \hat{a} \\ E\hat{b} &= \hat{a} - E\hat{a} \\ \hat{b} &= \hat{a}(\frac{1-E}{E}) \qquad (3) \end{aligned}$$

On injecte (3) dans (2) :

$$\begin{aligned} V &= \frac{\hat{a}\hat{a}(\frac{1-E}{E})}{(\hat{a}+\hat{a}(\frac{1-E}{E}))^2(\hat{a}+\hat{a}\frac{1-E}{E}+1)} \\ V &= \frac{\hat{a}^2(1-E)}{(\hat{a}(1+\frac{1-E}{E}))^2(\hat{a}(1+\frac{1-E}{E})+1)E} \\ 1+\frac{1-E}{E} &= \frac{1}{E} \\ V &= \frac{\hat{a}^2(1-E)}{\hat{a}^2\frac{1}{E^2}(\frac{\hat{a}}{E}+1)E} \end{aligned}$$

$$V = \frac{(1-E)E}{\frac{\hat{a}}{E}+1} = \frac{(1-E)E^2}{\hat{a}+E}$$
$$\hat{a}+E = \frac{(1-E)E^2}{V}$$
$$\hat{a} = \frac{(1-E)E^2}{V} - E$$
$$\hat{a} = E(\frac{(1-E)E}{V} - 1) \quad (4)$$

On injecte (4) dans (3) :

$$\hat{b} = \hat{a}\frac{1-E}{E}$$
$$\hat{b} = E(\frac{(1-E)E}{V} - 1)(\frac{1-E}{E})$$
$$\hat{b} = (\frac{(1-E)E}{V} - 1)(1-E)$$

(b) Ces résultats sont calculés par la fonction **Q2** de main **main.py** et apparaissent dans le terminal.

(c)

$$\begin{aligned}
\log L(a,b;\mathbf{x}) &= \sum_{i=1}^{n} \log(f_{x_i}(x_i,a,b)) \\
&= \sum_{i=1}^{n} \log(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x_i^{a-1}(1-x_i)^{b-1}) \\
&= \sum_{i=1}^{n} \log(x_i^{a-1}) + \sum_{i=1}^{n} \log((1-x_i)^{b-1}) + \sum_{i=1}^{n} \log(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}) \\
&= (a-1)\sum_{i=1}^{n} \log(x_i) + (b-1)\sum_{i=1}^{n} \log(1-x_i) - n\log(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}) \\
&= (a-1)\sum_{i=1}^{n} \log(x_i) + (b-1)\sum_{i=1}^{n} \log(1-x_i) - n\log(\beta(a,b))
\end{aligned}$$

(d) Ces résultats sont calculés par la fonction **Q2** de main **main.py** et apparaissent dans le terminal.

(e) En superposant les données de notre population et la distributions Beta(a,b), on obtient ce graphique : On remarque alors facilement que les données "collent" bien à la distribution Beta(a,b)

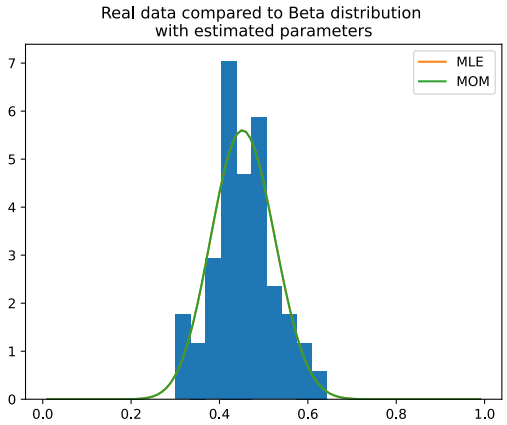

FIGURE 5 – Superposition des données de notre population et de la distribution Beta(a,b)

(f) Ces résultats sont calculés par la fonction **Q2** de main **main.py** et apparaissent dans le terminal.

(g) Ces résultats sont calculés par la fonction **Q2** de main **main.py** et apparaissent dans le terminal.

(h) En comparant les résultats qui apparaissent sur le terminal, on conclut que les deux méthodes ont une différence négligeable et qu'elles fournissent toutes deux des bons approximateurs.

Bonus

(i) Les résultats obtenus pour les autres tailles d'échantillons sont également disponibles dans le terminal. En les observant, on remarque que plus grande est la taille de l'échantillon, meilleurs sont les résultats (erreur, variance et biais diminuent).

3 Estimation par intervalle

(a) On pose le fait que la variable **PIB / habitant** suit une distribution exponentielle de paramètre λ inconnu. Nous allons créer des intervalles de confiance 95% par la méthode du pivot et du bootstrap. Par la méthode du pivot l'intervalle de confiance vaut : $100\%(1-\alpha) = 95\%$:

$$P(Q(Y,\lambda) \in \mathcal{A}) \geq 1-\alpha$$

Dans notre cas $Q(Y,$
$lambda)$ la quantité pivot et notre intervalle peut est décrit comme ceci :

$$\{\lambda : Q(Y,\lambda) \in \mathcal{A}\}$$

On décrit maintenant notre quantité pivot grâce au fait que notre variable suit une distribution exponentielle :

$$Y_1, ..., Y_n \sim \text{Expo}(\lambda), \text{avec } Y_i\lambda \sim \text{Expo}(1)$$

$$2T\lambda \sim \Gamma(n, \frac{1}{2}) = \chi^2_{2n} \text{avec } T = \sum_{i=1}^{n} Y_i$$

On obtient donc l'intervalle de confiance : $[\frac{\chi^2_{2n,\frac{\alpha}{2}}}{2T}; \frac{\chi^2_{2n,1-\frac{\alpha}{2}}}{2T}]$ Et dans le cas précis de $\alpha = 0.05$: $[\frac{\chi^2_{2n,0.025}}{2T}; \frac{\chi^2_{2n,1-0.975}}{2T}]$

(b) Ces résultats sont calculés par la fonction **Q3** de main **main.py** et apparaissent dans le terminal.

(c) La méthode du bootstrap approxime la distribution en trois étapes :

- Tirer un échantillon de bootstrap de $Y_1, ..., Y_n$, $\hat{F}$
- Calculer $\hat{\lambda} = T(Y_1, ..., Y_n)$ la réalisation de l'estimateur
- Répéter les deux premières étapes m fois pour obtenir : $\hat{\lambda}_1, ..., \hat{\lambda_m}$.

On construit ensuite l'intervalle de confiance grâce aux quantiles de distribution de l'estimateur obtenus précédemment : $[\mathcal{Q}_n(\frac{\alpha}{2}); \mathcal{Q}_n(1-\frac{\alpha}{2})]$

(d) En repartant des définitions faites au point précédent, on choisit les valeurs suivantes :
$\begin{cases} \alpha = 0.05 \text{Pour que l'intervalle de confiance soit à } 95\% \\ m = 100 \text{Pour réaliser nos 100 échantillons} \end{cases}$

(e) On marque l'évolution de la taille de l'intervalle pour les deux méthodes en fonction de la taille de l'échantillon dans le graphique ci-dessous : En comparant les deux méthodes, on remarque que la taille de l'intervalle diminue avec la méthode du pivot alors que dans le cas du bootstrap, celle-ci est faible et reste constante avec l'agrandissement de la taille de l'échantillon.

(f) On marque l'évolution de la proportion d'intervalles contenant la vraie valeur de lambda en fonction de la taille de l'échantillon dans le graphique ci-dessous : En comparant les deux méthodes, on remarque encore une fois que pour le pivot, cela reste constant alors que pour le bootstrap, cette proportion augmente.

(g) Oui car on rencontre plusieurs fois la vraie valeur de lambda avec nos méthodes du pivot et du bootstrap dans nos intervalles de confiance, il est donc raisonable de supposer que notre variable suivait une distribution exponentielle.

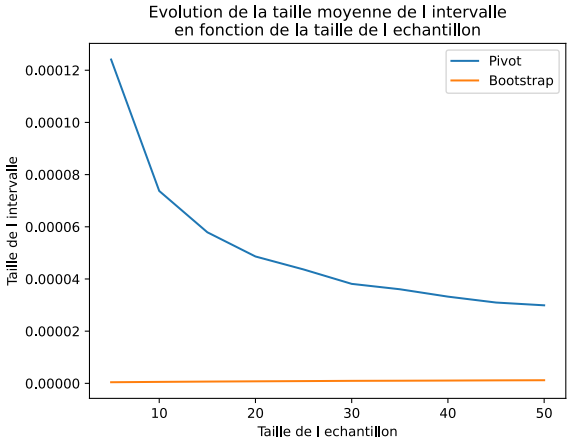

FIGURE 6 – Évolution de la taille de l'intervalle en fonction de la taille de l'échantillon

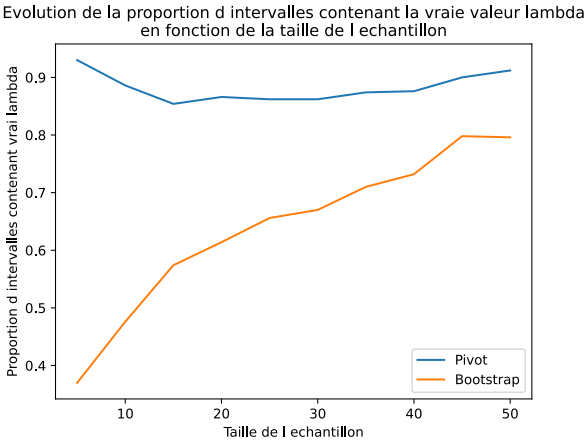

FIGURE 7 – Évolution de la proportion d'intervalles contenant la vraie valeur de lambda en fonction de la taille de l'échantillon

4 Test d'hypothèse

(a) Nous formulons les hypothèses suivantes pour vérifier nos intuitions concernant le rapport possible entre les variables **PIB / habitant** et **CO2 / habitant**. Nous créons donc deux hypothèses, une nulle et une alternative. Pour cela on définit $\Delta_{rel} = \text{emissions}_riche - \text{emissions}_pauvre$

- Hypothèse nulle : H_0 : Delta est supérieur ou égal à la différence entre la moyenne des émissions des pays riches et des pays pauvres : $Delta \geq \Delta_{reel}$
- Hypothèse alternative : H_1 : Delta est inférieur ou égal à la différence entre la moyenne des émissions des pays riches et des pays pauvres : $Delta \geq \Delta_{reel}$

Dans notre population, on ne peut réfuter l'hypothèse nulle car l'erreur sur Delta est quasi-nulle. Il faut donc réaliser d'autre tests pour valider ou réfuter des hypothèses.

(b) Nous commençons par calculer les variances des échantillons des populations riches et pauvres. Grâce à celles-ci, on réalise une interpolation afin d'estimer σ^2. Pour conclure, on utilise la distribution student-t afin de déterminer l'intervalle de confiance par la formule qui suit :

$$-t < \frac{(\bar{X} - \bar{Y}) - \Delta_{reel}}{S_{pooled}\sqrt{\frac{1}{n} + \frac{1}{m}}} < t$$

Ce qui nous permet d'obtenir les bornes de l'intervalle sur Δ_{reel} qui permet de valider ou refuter une hypothèse :

$$[-t \times S_{pooled}\sqrt{\frac{1}{n} + \frac{1}{m}}(\bar{X} - \bar{Y}); t \times S_{pooled}\sqrt{\frac{1}{n} + \frac{1}{m}}(\bar{X} - \bar{Y})]$$

(c) Pour vérifier nos hypothèses, nous procédons au test d'hypothèse sur 100 échantillons. À chaque itération, on vérifie si Δ_{reel} appartient à l'intervalle décrit plus haut et dans notre cas, on rejete 39% des essais.

(d) Dans ce deuxième, cas on ré applique notre fonction **test_hypothesis** de **main.py** et nous devons donc rejeter 25% des essais. Pour conclure, on peut dire que notre α n'était pas assez représentatif.