MATH0487 : Rapport du devoir

Romain LAMBERMONT (s190931) Arthur LOUIS (s191230)

16 août 2022

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.

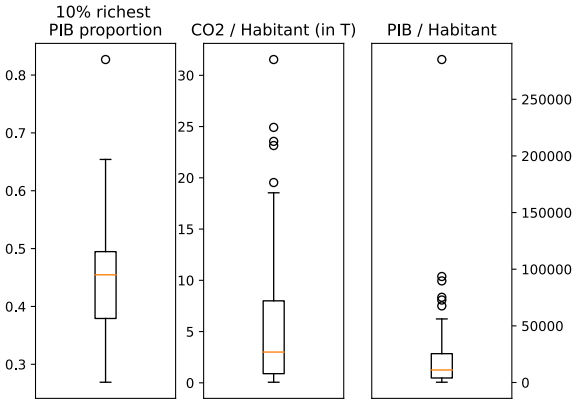


FIGURE 1 – Boite à moustache des données de **data.csv**

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 avec les formules suivantes $Q_n(0.5), Q_n(0.25), Q_n(0.75)$ représentant respectivement la médiane, le premier et troisième quartile :

$$\text{borne}_{\min} = Q_n(0.25) - 1.5 * [Q_n(0.75) - Q_n(0.25)]$$
$$\text{borne}_{\max} = Q_n(0.75) + 1.5 * [Q_n(0.75) - Q_n(0.25)]$$

On regroupe les bornes dans les intervalles de confiance suivants avec les données négatives égalées à 0, des données ne faisant aucun sens dans nos données. :

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

En comparant les 3 boîtes à moustache, on se rend compte que celles du CO_2 et du PIB par habitant sont assez semblable au niveau des données aberrantes localisées assez proche de la borne maximale de l'intervalle de confiance. Dans les 3 graphiques, on remarque une donnée largement abberante par graphique. Pour la proportion des 10% les plus riche, c'est l'Oman, pour le CO_2 le Luxembourg et pour le PIB le Venezuela.

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

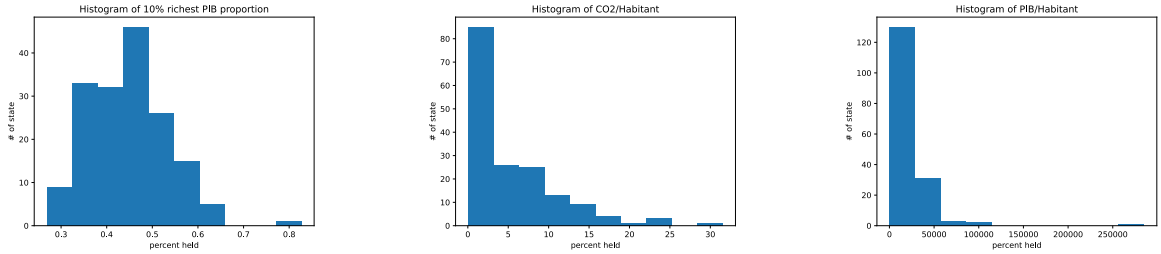


FIGURE 2 – Histogrammes des différentes variables

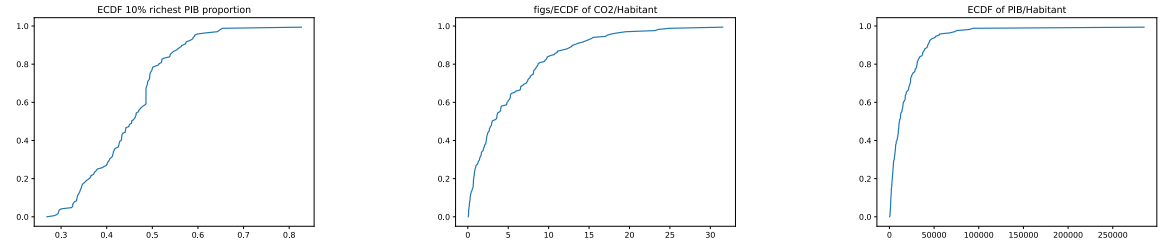


FIGURE 3 – ECDF des différentes variables

En comparant les 3 groupes de graphiques, on remarque aisément que les histogrammes et CDF relatifs aux variables CO_2 et PIB par habitant se ressemblent énormément. On en déduit donc que ces variables sont certainement distribuées de la même manière, ici une exponentielle. Pour ce qui est de la variable Top10, on remarque que la distribution est totalement différente et nous fait penser à une loi normale.

(c) Pour analyser les relations entre les variables numériquement, on décide d'utiliser les coefficients de corrélation. En effet, grâce à la fonction **corr** de Pandas, on calcule les différents coefficients de corrélation regroupés dans ce tableau :

	Top10	CO2	PIB
Top10	1	-0.226302	-0.216884
CO2	-0.226302	1	0.536511
PIB	-0.216884	0.536511	1

TABLE 4 – Coefficients de corrélation

On remarque donc un coefficient de corrélation positif assez élevé (corrélation forte) entre les données de CO_2 et PIB par habitant ce qui semble assez logique car plus un pays s'enrichit, plus il a tendance à produire du CO_2. Pour les deux autres corrélations, on remarque un coefficient de corrélation négatif assez faible. On peut donc remarquer que moins les richesses sont distribuées entre tous les habitants (Top10 augmente), moins le pays est riche et produit du CO_2.

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

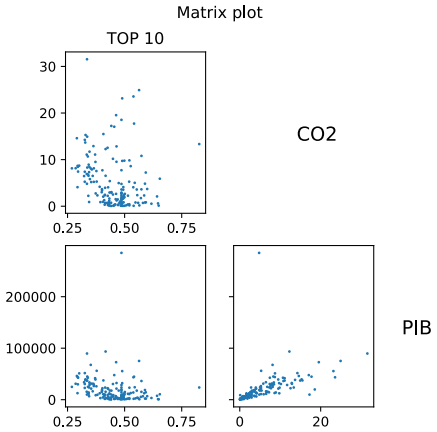


FIGURE 4 – Graphique "matrice" des variables de **data.csv**

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.

En comparant les résultats obtenus entre la méthode numérique et graphique, on remarque bien la corrélation forte entre le CO_2 et le PIB par habitant prévue par le coefficient de corrélation grâce à la droite présente dans le graphique correspondant de la matrice. En ce qui concerne les deux autres relations, on remarque bien la corrélation faible prévue par les coefficients dans les graphiques respectifs de la matrice.

2 Estimation ponctuelle

(a) Nous allons supposer que notre variable **Top10** suit une distribution Beta(a,b) de paramètres a et b inconnus. Nous allons déterminer ceux-ci par la méthode des moments en supposant l'espérance et la variance comme connues et nous allons partir de leurs expressions pour isoler nos paramètres et calculer leurs valeurs :

$$\text{Top10} \equiv X \sim \text{Beta(a,b)} \qquad \text{E(X)} = \frac{a}{a+b} \qquad \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

Nous pouvons également noter :

$$\text{E}(X) = \bar{x} \equiv \mu \qquad \text{Var}(X) = \text{E}(X^2) - \text{E}(X)^2 = \bar{x^2} - \bar{x}^2 \equiv v$$

Résolvons le système :

$$\begin{aligned}
&\begin{cases} \mu(\hat{a}+\hat{b}) = \hat{a} \\ v(\hat{a}+\hat{b})^2(\hat{a}+\hat{b}+1) = \hat{a}\hat{b} \end{cases} \\
\Leftrightarrow &\begin{cases} \hat{a} = \frac{\mu}{1-\mu}\hat{b} \\ v\left(\frac{\mu}{1-\mu}\hat{b}+\hat{b}\right)^2\left(\frac{\mu}{1-\mu}\hat{b}+\hat{b}+1\right) = \frac{\mu}{1-\mu}\hat{b}^2 \end{cases} \\
\Leftrightarrow &\begin{cases} \hat{a} = \frac{\mu}{1-\mu}\hat{b} \\ \hat{b} = \left[\frac{\mu}{v(1-\mu)}\frac{1}{\left(\frac{\mu}{1-\mu}+1\right)^2} - 1\right]\frac{1}{\frac{\mu}{1-\mu}+1} \end{cases} \\
\Leftrightarrow &\begin{cases} \hat{a} = \left[\frac{\mu(1-\mu)}{v} - 1\right]\mu \\ \hat{b} = \left[\frac{\mu(1-\mu)}{v} - 1\right](1-\mu) \end{cases}
\end{aligned}$$

On obtient donc ces deux estimateurs pour la méthode des moments :

$$\hat{a}_{\text{MOM}} = \left[\frac{\mu(1-\mu)}{v} - 1\right]\mu \qquad \hat{b}_{\text{MOM}} = \left[\frac{\mu(1-\mu)}{v} - 1\right](1-\mu)$$

(b) Pour notre échantillon de 50 pays, ces paramètres prennent les valeurs suivantes :

$$\hat{a}_{\text{MOM}} = 15.80411 \qquad \hat{b}_{\text{MOM}} = 18.96703$$

(c) On sait que la fonction $f_{x_i}(x_i, a, b)$ suit une distribution $\beta(a, b)$ ce qui nous donne :

$$f_{x_i} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$$

On peut alors calculer que :

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

On obtient donc bien la formule du log de la vraisemblance de la distribution étudiée grâce aux propriétés des logarithmes.

(d) On se sert de la fonction **scipy.stats.optimize** de Python et de la fonction **beta_log_likelihood** fournie dans le projet pour calculer les deux paramètres a et b en partant du point (1,1) pour converger vers les bons résultats :

$$\hat{a}_{MLE} = 16.19073 \qquad \hat{b}_{MLE} = 19.41809$$

(e) En superposant les données de notre population et la distributions Beta(a, b), on obtient ce graphique avec les méthodes MOM et MLE :

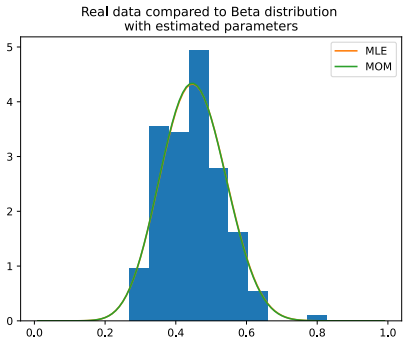


FIGURE 5 – Superposition des données de notre population et de la distribution Beta(a, b)

Les deux méthodes produisent un résultat assez semblable, les deux méthodes semblent donc être appropriées pour approximer correctement les paramètres de la distribution Beta.

(f) Nous allons tirer 500 échantillons i.i.d de 50 pays de la population et ensuite approximer les paramètres $\hat{a}$ et $\hat{b}$ de la distribution selon la méthode des moments. Nous allons également calculer le biais, la variance et l'erreur quadratique moyenne par rapport à la valeur réelle de la variable. On définit le biais et l'erreur quadratique moyenne comme suit :

$$\text{Biais}(\hat{\theta}) \equiv E(\hat{\theta}) - \theta \qquad \text{MSE}(\hat{\theta}) \equiv E((\hat{\theta} - \theta)^2)$$

On obtient donc les résultats suivants :

$$\hat{a}_{MOM} = 14.61235 \qquad \hat{b}_{MOM} = 17.84364$$

	$\hat{a}$	$\hat{b}$
Biais	1.26235	1.53364
Var	11.34227	17.80979
MSE	12.93583	20.16187

TABLE 5 – Comparaison des estimateurs à ceux de la méthode des moments

(g) De la même manière on obtient les résultats suivants grâce à la méthode du maximum de vraisemblance :

$$\hat{a}_{MLE} = 14.58236 \qquad \hat{b}_{MLE} = 17.79154$$

	$\hat{a}$	$\hat{b}$
Biais	1.23236	1.48154
Var	11.91723	19.08437
MSE	13.43596	21.27933

TABLE 6 – Comparaison des estimateurs à ceux de la méthode du maximum de vraisemblance

(h) En comparant les résultats obtenus par les deux méthodes, on remarque que pour une taille d'échantillon de 50 pays, elles produisent des résultats similaires. Les valeurs de biais, variance et MSE sont du même ordre de grandeur. On peut donc conclure que pour cette taille d'échantillon les deux méthodes se valent.

Bonus

(i) Nous allons maintenant effectuer la même expérience sur différentes tailles d'échantillons pour les deux méthodes et regarder l'évolution des différentes valeurs (biais, variance, MSE) par rapport aux tailles des échantillons.

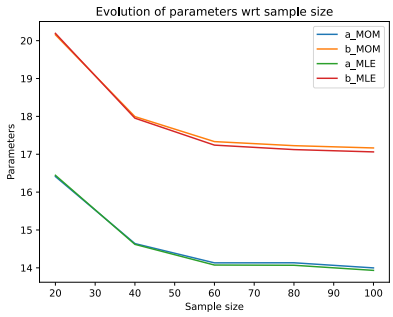
Taille échantillon	$\hat{a}$	$\hat{b}$	Biais $\hat{a}$	Biais $\hat{b}$	Var $\hat{a}$	Var $\hat{b}$	MSE $\hat{a}$	MSE $\hat{b}$
20	16.41	20.15	3.06	3.84	39.39	61.40	48.76	76.18
40	14.64	17.99	1.29	1.68	15.03	24.51	16.69	27.35
60	14.13	17.33	0.78	1.02	9.59	15.72	10.20	16.77
80	14.13	17.22	0.78	0.91	7.16	11.60	7.77	12.44
100	13.99	17.16	0.64	0.85	5.18	8.44	5.60	9.17

TABLE 7 – Méthode des moments appliquée à différentes tailles d'échantillons

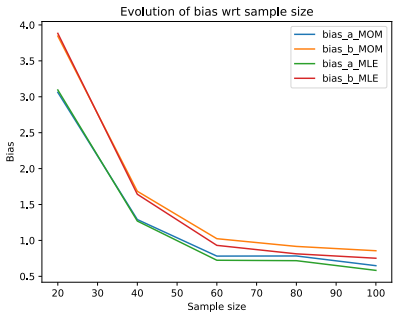
Taille échantillon	$\hat{a}$	$\hat{b}$	Biais $\hat{a}$	Biais $\hat{b}$	Var $\hat{a}$	Var $\hat{b}$	MSE $\hat{a}$	MSE $\hat{b}$
20	16.44	20.19	3.09	3.88	39.12	61.70	48.70	76.77
40	14.61	17.95	1.26	1.64	15.73	26.09	17.34	28.79
60	14.07	17.24	0.72	0.93	10.22	17.09	10.75	17.96
80	14.06	17.12	0.71	0.81	7.77	12.82	8.29	13.48
100	13.93	17.06	0.58	0.75	5.59	9.30	5.93	9.86

TABLE 8 – Méthode du maximum de vraisemblance appliquée à différentes tailles d'échantillons

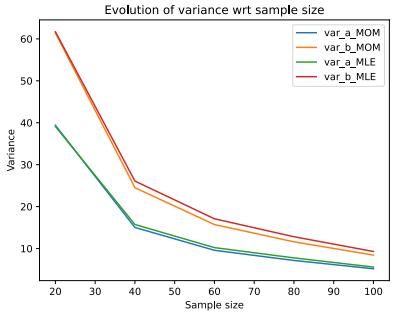
Plus la taille de l'échantillon est grande, plus les résultats sont précis et convergent vers les vraies valeurs de $\hat{a}$ et $\hat{b}$. Les différents graphiques suivants illustrant l'évolution de toutes les valeurs (biais, variance, MSE) en fonction de la taille de l'échantillon nous permettent de montrer que les deux méthodes et évoluent de la même façon. Cela nous conforte dans l'idée exprimée plus tôt que les deux méthodes sont équivalentes.



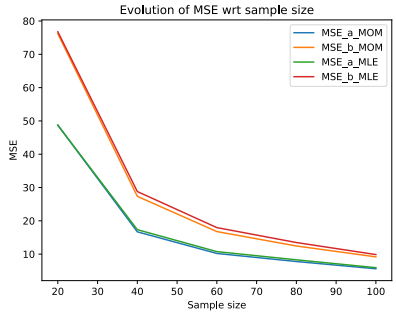
(a)



(b)



(c)



(d)

FIGURE 6 – Évolution des différentes valeurs en fonction de la taille de l'échantillon

3 Estimation par intervalle

(a) Nous allons faire l'hypothèse que la variable **PIB/habitant** suit une distribution exponentielle de paramètre λ inconnu. Nous allons, pour vérifier notre hypothèse, construire des intervalles de confiance à 95% à l'aide de la méthode du pivot et du bootstrap. Commençons par le pivot. Si on désire un intervalle de confiance à 95% $= (1-\alpha)100\%$, on définit $\theta = \frac{1}{\lambda}$ et $Q(Y;\theta)$ la quantité pivot tel que :

$$P(Q(Y;\theta) \in A) \geq 1-\alpha$$

Notre intervalle de confiance devient alors :

$$\{\theta : Q(Y;\theta) \in A\}$$

Calculons maintenant $Q(Y;\theta)$ grâce à notre hypothèse de distribution exponentielle il vient pour n échantillons :

$$\begin{aligned}
\textbf{PIB/habitant} \equiv Y_i &\overset{\text{i.i.d}}{\sim} \text{Expo}(\lambda) \quad i = 1, ..., n \\
&\Leftrightarrow \frac{Y_i}{\theta} \overset{\text{i.i.d}}{\sim} \text{Expo}(1) \\
&\Leftrightarrow \frac{2\sum_{i=1}^{n} Y_i}{\theta} \sim \Gamma\left(n, \frac{1}{2}\right) \\
&\Leftrightarrow \frac{2n\mu}{\theta} \sim \chi^2_{2n}
\end{aligned}$$

On obtient donc finalement notre intervalle de confiance :

$$\begin{aligned}
1-\alpha &= P(\chi^2_{2n,1-\alpha/2}) \\
&\Leftrightarrow \theta \in \left[\frac{2n\mu}{\chi^2_{2n,\alpha/2}}, \frac{2n\mu}{\chi^2_{2n,1-\alpha/2}}\right] \\
&\Leftrightarrow \lambda \in \left[\frac{\chi^2_{2n,1-\alpha/2}}{2n\mu}, \frac{\chi^2_{2n,\alpha/2}}{2n\mu}\right]
\end{aligned}$$

(b) Pour notre population de 50 pays, notre intervalle de confiance est donc :

$$\lambda \in \left[4.88037 \times 10^{-5}; 8.51914 \times 10^{-5}\right]$$

(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) Pour notre population de 50 pays, notre intervalle de confiance est donc :

$$\lambda \in \left[3.98483 \times 10^{-5}; 6.73827 \times 10^{-5}\right]$$

(e) On représente l'évolution de la taille de l'intervalle en fonction de la taille de l'échantillon dans la figure 7. On y remarque aisément que plus la taille de l'échantillon augmente, plus la taille de l'intervalle diminue. La décroissance est d'autant plus importante en ce qui concerne la méthode du Bootstrap. Ces résultats sont attendus. En effet, plus un échantillon est grand moins il est impacté par les données aberrantes car selon le théorème central limite les valeurs se regroupent autour de l'espérance. Pour un petit échantillon, on privilégiera la méthode du pivot. Une fois que la taille de l'échantillon dépasse 10, les deux méthodes se valent et plus aucune amélioration significative ne se fait remarquer quand on dépasse un échantillon de 25 pays.

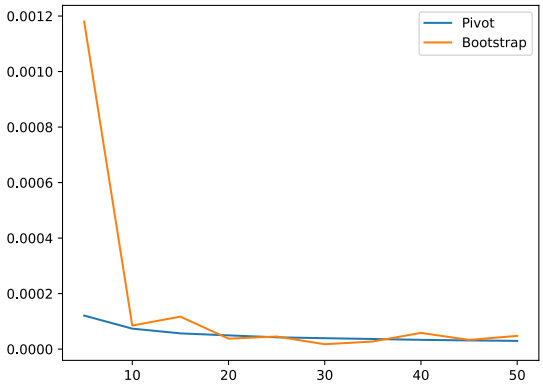


FIGURE 7 – Taille de l'intervalle en fonction de la taille de l'échantillon

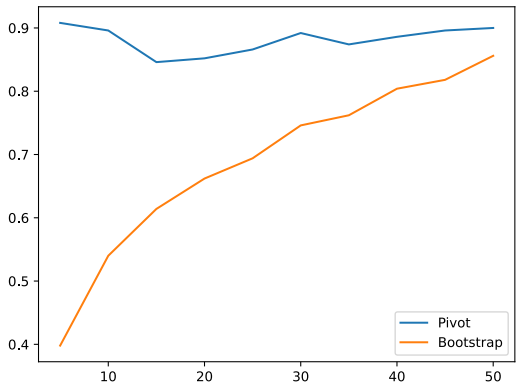


FIGURE 8 – Proportion d'intervalles contenant le vrai lambda en fonction de la taille de l'échantillon

(f) On représente l'évolution de la proportion d'intervalles contenant la vraie valeur de λ dans la figure 8. On y remarque aisément que la méthode du pivot nous fournit peu importe la taille de l'échantillon, une proportion supérieure à celle du Bootstrap qui commence par être assez mauvaise mais qui s'améliore grandement avec l'augmentation de la taille de l'échantillon. Cependant, peu importe la taille de l'échantillon, on privilégiera la méthode du pivot.

(g) On peut conclure que notre supposition initiale était correcte. En effet, la méthode du pivot se basant sur la distributio proposée nous rend de meilleurs résultats que la méthode du Bootstrap qui se base sur les données brutes avec un échantillon plus petit alors que cette dernière rejoint la méthode du pivot plus l'échantillon grandit, ce qui nous confirme encore dans notre hypothèse.

4 Test d'hypothèse

(a) Nous allons tenter de vérifier l'affirmation des scientifiques en réalisant deux hypothèses. Soit μ_r et μ_p les moyennes d'émissions de respectivement les pays riches et pauvres et Δ calculé par la fonction **scientific_delta** :

$$H_0 = \mu_r - \mu_p = \Delta$$
$$H_1 = \mu_r - \mu_p > \Delta$$

Nous allons maintenant vérifier laquelle des 2 hypothèses est vraie pour notre population. Pour ce faire nous isolons les pays riches et pauvres en comparant le PIB par habitant de chaque pays à la médiane. Nous calculons ensuite le $\Delta_{\text{réel}}$ comme la différence des moyennes d'émission de CO_2 des pays riches et pauvres. On obtient les valeurs suivantes :

$$\Delta_{\text{réel}} = \mu_r - \mu_p = 7.474667084167256$$
$$\Delta_{\text{scientifique}} = 7.474667084167256$$
$$\Leftrightarrow \Delta_{\text{réel}} = \Delta_{\text{scientifique}}$$

Cette dernière relation est vérifiée sur Python qui nous retourne bien une différence égale à 0.0 pour notre population. L'hypothèse correcte pour notre population est donc bien H_0 proposée par les scientifiques.

(b) Nous allons tester l'hypothèse nulle H_0 qui dit $\Delta_{\text{réel}} = \Delta_{\text{scientifique}}$ en considérant l'hypothèse H_1 en supposant une distribution $\mathcal{N}(\mu, \sigma^2)$ avec les variances supposées égales pour les moyennes d'émission de CO_2. Nous allons également utiliser les moyennes d'émission observées $\bar{\mu_r}$ et $\bar{\mu_p}$ avec les propriétés de la distribution normale :

$$\bar{\mu_r} - \bar{\mu_p} \sim \mathcal{N}\left(\mu_r - \mu_p, \sigma^2\left(\frac{1}{n_r} + \frac{1}{n_p}\right)\right)$$

On peut ensuite normaliser comme suit sachant $\Delta = \mu_r - \mu_p$:

$$\frac{\bar{\mu_r} - \bar{\mu_p} - \Delta}{\sigma\sqrt{\frac{1}{n_r} + \frac{1}{n_p}}} \sim \mathcal{N}(0, 1)$$

Nous allons donc estimer la variance σ sous l'hypothèse que les variances des pays riches et pauvres sont égales :

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

On sait également que :

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}$$

Et :

$$\sum_i \chi^2 = \chi^2_{\sum_i}$$

Ce qui nous donne :

$$\frac{(n_r-1)S_r^2}{\sigma^2} + \frac{(n_p-1)S_p^2}{\sigma^2} \sim \chi^2_{n_r+n_p-2}$$

En effectuant un pooling :

$$S_{\text{pooled}^2} = \frac{(n_r-1)S_r^2 + (n_p-1)S_p^2}{n_r+n_p-2}$$

On relie ainsi les observations et les moyennes exactes grâce à une distribution student-t dans notre variable de test T sous H_0 :

$$T = \frac{\bar{\mu_r} - \bar{\mu_p} - \Delta}{S^2_{\text{pooled}}\sqrt{\frac{1}{n_r} + \frac{1}{n_p}}} \sim t_{n_r+n_p-2}$$

Notre test d'hypothèse permet donc de rejeter H_0 pour H_1 au seuil $\alpha = 0.05$ si :

$$T > t_{n_r+n_p-2,1-\alpha}$$
$$\Leftrightarrow \Delta_{\text{réel}} = \bar{\mu_r} - \bar{\mu_p} > t_{n_r+n_p-2,1-\alpha} S_{\text{pooled}}\sqrt{\frac{1}{n_r} + \frac{1}{n_p}}$$

(c) Nous avons commencé par tirer les 100 échantillons demandés et nous obtenions des résultats assez volatils au vu du peu de nombre de test. Nous avons donc décidé de faire 10000 tests avec un échantillon de 75 pays de notre population. Nous avons obtenu un résultat oscillant entre 0.4% et 0.7%. Ce nombre représentant la proportion d'hypothèses nulles rejetées par le test, on en conclut que l'hypothèse nulle ne peut pas être rejetée dans notre population.

(d) Cette fois-ci, nous avons effectué le test en prenant des échantillons de 25 pays de notre population et nous obtenons un résultat oscillant entre 2.3% et 2.6%. Ceux-ci sont encore une fois inférieurs à 5%, ce qui montre que H_0 passe les tests et que l'hypothèse nulle ne peut être réfutée. Néanmoins, ces résultats sont moins bons, en effet cela est logique au vu de la taille réduite de l'échantillon. Un échantillon de 25 pays implique deux choses : le test est moins précis car affecté plus lourdement par les données aberrantes et notre proposition de distribution normale des émissions est discutable au vu du théorème centrale limite.