Statistiques Appliquées
Correction du Contrôle Continu N°2

TD N°3-4-5

Gwenn PARENT

Questions de cours : (3 points)

1. **Rappelez les propriétés des estimateurs MCO ?**
   - Les estimateurs des MCO des coefficients du modèle linéaire sont sans biais.
   - Ces estimateurs sont linéaires par rapport aux variables dépendantes (généralement notées y_i).
   - Ils sont efficaces, c'est à dire de variance minimale, dans l'ensemble des estimateurs linéaires sans biais.

2. **Donnez leur formule algébrique pour un ajustement simple.**
   Modèle linéaire simple avec constante : $y_t = ax_t + b + \epsilon_t$; $t = 1, 2, ..., T$
   L'estimateur des MCO du paramètre a est :

$$\widehat{a} = \frac{\sum_{t=1}^{T} x_t \; y_t - T \; \overline{x} \; \overline{y}}{\sum_{t=1}^{T} x_t^2 - T \; \overline{x}^2}$$

   où $\overline{x}$ (resp. $\overline{y}$) est la moyenne empirique des x_t (resp. y_t).
   Posons $S_{xy} = \sum_{t=1}^{T} x_t \; y_t - T \; \overline{x} \; \overline{y}$ et $S_{xx} = \sum_{t=1}^{T} x_t^2 - T \; \overline{x}^2$. On a alors d'après ce qui précède : $\widehat{a} = \frac{S_{xy}}{S_{xx}}$
   Notons $cov(x, y)$ la covariance empirique entre $(x_1, x_2, ..., x_T)$ et $(y_1, y_2, ..., y_T)$ et $V(x)$ la variance empirique de $(x_1, x_2, ..., x_T)$. On a :

$$cov(x, y) = \frac{1}{T} \sum_{t=1}^{T} x_t y_t - \overline{x}\overline{y} = \frac{1}{T} S_{xy}$$

$$V(x) = \frac{1}{T} \sum_{i=1}^{T} x_t^2 - \overline{x}^2 = \frac{1}{T} S_{xx}$$

   ce qui nous fournit une seconde expression pour $\widehat{a}$:

$$\widehat{a} = \frac{cov(x, y)}{V(x)}$$

   L'estimateur des MCO de b est : $\widehat{b} = \overline{y} - \widehat{a} \; \overline{x}$

3. **Pourquoi ces estimateurs sont-ils des variables aléatoires ?**
   Dans le modèle linéaire, les variables dépendantes y_t sont des variables aléatoires (elles sont généralement obtenues par échantillonage aléatoire dans une population). Les estimateurs MCO sont des fonctions de ces variables, et par conséquent, sont également des variables aléatoires.
   Pour plus de détails, regardez le corrigé du TD3 sur le site du cours :
   (http ://www.pse.ens.fr/junior/parent/tdstat.htm).

Problème : Evaluation du prix moyen des cigarettes, du pain et du riz sur un échantillon de 26 pays. (17 points)

"Tabac et pauvreté forment à eux deux un cercle vicieux. Dans la plupart des pays, le tabagisme est généralement plus répandu parmi les plus pauvres. C'est pourquoi les dépenses de tabac représentent une part importante du revenu des familles défavorisées. Or, l'argent qui passe dans le tabac ne peut être dépensé pour des besoins essentiels comme l'alimentation, le logement, l'éducation et les soins de santé. Le tabac peut en outre aggraver la pauvreté des fumeurs et de leurs familles du fait que ces derniers sont beaucoup plus susceptibles de tomber malade et de mourir prématurément d'un cancer, d'une crise cardiaque, d'une maladie respiratoire ou d'autres maladies liées au tabagisme, privant leurs familles d'un revenu très précieux et leur imposant des dépenses supplémentaires pour les soins de santé. Par ailleurs, même si l'industrie du tabac emploie des milliers de personnes, la grande majorité d'entre elles gagnent très peu tandis que les grandes compagnies de tabac engrangent d'énormes bénéfices." Organisation Mondiale de la Santé (2000).

Toutes les questions du problème sont indépendantes, vous pouvez les traiter séparemment (sauf les questions 6 et 11 qui sont des applications numériques des questions précédentes).

Nous souhaitons étudier le prix moyen des cigarettes, du pain et du riz dans le monde. Nous disposons pour cela des prix (calculés en temps de travail pour un ouvrier) d'un paquet d'une marque présente dans tous les pays sélectionnés, d'un paquet d'une marque locale, ainsi que d'un kg de pain et d'un kg de riz pour un échantillon de 26 pays (13 pays développés et 13 pays en développement). (cf. le tableau ci-joint).
Nous faisons l'hypothèse que les temps moyens qu'un ouvrier de chaque pays doit travailler pour se payer ces différents biens suivent une loi normale d'espérance m et de variance σ^2. Nous adopterons par la suite les notations suivantes :

$$X_1, X_2, \dots X_N \approx \mathcal{N}\left(m_1, \sigma_1^2\right) \textit{ pour les } Marlboro$$

$$Y_1, Y_2, \dots Y_N \approx \mathcal{N}\left(m_2, \sigma_2^2\right) \textit{ pour les marques locales}$$

$$Z_1, Z_2, \dots Z_N \approx \mathcal{N}\left(m_3, \sigma_3^2\right) \textit{ pour le pain}$$

$$T_1, T_2, \dots T_N \approx \mathcal{N}\left(m_4, \sigma_4^2\right) \textit{ pour le riz}$$

1. **En reprenant les notations définies ci-dessus, indiquez à quoi correspondent les chiffres $25,77$ puis $15,67$ du tableau 1. Expliquez quelle est la différence entre un paramètre, et son estimateur.** ***(1,5 points)***

   ATTENTION ! ! !
   La question précisait bien en reprenant les notations précédemment définies !

   $25,77 = \frac{\sum_{i=1}^{26} Y_i}{26}$: moyenne empirique sur l'échantillon des 26 pays du prix des cigarettes de marque locale.

   $15,67 = \sqrt{\frac{\sum_{i=1}^{26} (T_i - \overline{T})^2}{25}}$: écart-type empirique sur l'échantillon des 26 pays du prix du kilogramme de riz.

   Le paramètre m_1 est l'espérance de la loi suivie par le prix des paquets de Marlboro, c'est une constante :

   $$X_1, X_2, \dots X_N \approx \mathcal{N}\left(m_1, \sigma_1^2\right)$$

   L'estimateur $\widehat{m_1}$ de ce paramètre est l'estimation qui peut être faite de ce paramètre sur un échantillon particulier. Cet estimateur est donc aléatoire car il dépend de l'échantillon sélectionné.

2. **Qu'observez-vous à première vue en comparant le temps de travail moyen nécessaire à un ouvrier pour l'achat d'un paquet de Marlboro et d'un kg de riz ? Quelle analyse succinte pouvez-vous faire du tableau 2 ?** ***(1 point)***

   - Il semble qu'en moyenne le temps de travail nécessaire à un ouvrier pour acquérir un paquet de Marlboro soit 1,5 fois plus important quecelui nécessaire pour acquérir un kg de riz.
   - Il existe des distinctions importantes dans ces prix entre pays développés et pays en développement : le prix du paquet de Marlboro est deux fois supérieur dans les pays en développement par rapport aux pays développés, le prix du riz, lui, est 2,5 fois supérieur.
   - Il semble donc qu'il y ait en fait deux modèles à prendre en compte :

$$X_{1devps}, X_{2devps}, \ldots X_{Ndevps} \approx \mathcal{N}\left(m_{1devps}, \sigma^2_{1devps}\right)$$

$$X_{1dvt}, X_{2dvt}, \ldots X_{Ndvt} \approx \mathcal{N}\left(m_{1dvt}, \sigma^2_{1dvt}\right)$$

   avec $m_{1devps} = \frac{1}{2} m_{1dvt}$.

3. **Calculez l'estimateur du maximum de vraisemblance du paramètre m_1. Comment s'appelle l'estimateur que vous trouvez ?** *(3 points)*
   Indication : Fonction de densité de la loi normale $\mathcal{N}\left(m, \sigma^2\right)$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

   Nous avons donc un N-échantillon d'une loi de Normale : X_1, X_2, ..., $X_N \approx$ *i.i.d.* $N(m_1, \sigma_1^2)$.
   Les deux premiers moments de la loi Normale sont :
   $E\left(X_i\right) = m_1 \qquad V\left(X_i\right) = \sigma_1^2, \qquad$ donc le paramètre est (m_1, σ_1).

   La vraisemblance du paramètre est $l(m_1, \sigma_1, x_1, x_2, ..., x_N)$
   $l\left(m_1, \sigma_1, x_1, x_2, ..., x_N\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2\sigma_1^2}\left(x_i - m_1\right)^2\right] = \left(2\pi\sigma_1^2\right)^{-\frac{N}{2}} \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{N}\left(x_i - m_1\right)^2\right]$

   La log-vraisemblance du paramètre est $L(m_1, \sigma_1, x_1, x_2, ..., x_N) = ln\ l(m_1, \sigma_1, x_1, x_2, ..., x_N)$
   $lnl\left(m_1, \sigma_1, x_1, x_2, ..., x_N\right) = L\left(m_1, \sigma_1, x_1, x_2, ..., x_N\right) = -\frac{N}{2}\ln\left(2\pi\right) - \frac{N}{2}\ln\left(\sigma_1^2\right) - \frac{1}{2\sigma_1^2}\sum_{i=1}^{N}\left(x_i - m_1\right)^2$

   L'estimateur du maximum de vraisemblance $\widehat{m_1}$ réalise le maximum en m_1 de $L\left(m_1, \sigma_1, x_1, x_2, ..., x_N\right)$.
   La fonction objectif est deux fois continument dérivable : nous pouvons appliquer les résultats classiques d'optimisation.
   - La condition nécessaire d'ordre 1 (CN1) est $\frac{\partial}{\partial m_1} L\left(m_1, \sigma_1, x_1, x_2, ..., x_N\right) = 0$ en $m_1 = \widehat{m_1}$
   - Une condition suffisante d'ordre 2 (CS2) est $\frac{\partial^2}{\partial m_1^2} L\left(m_1, \sigma_1, x_1, x_2, ..., x_N\right) < 0$ en $m_1 = \widehat{m_1}$

> **ATTENTION !!! Dérivez par rapport à m_1 (le paramètre dont vous cherchez un estimateur) !!!**
> Je ne veux pas voir de dérivée par rapport à σ_1, et encore moins par rapport à θ, qui n'apparait nulle part dans cette interro !

   CN1 :

$$\frac{\partial L}{\partial m_1} = -\frac{1}{2\sigma_1^2}\sum_{i=1}^{N}(-2)\left(x_i - m_1\right) = 0$$

$$\frac{1}{\sigma_1^2}\sum_{i=1}^{N}(x_i - m_1) = 0$$

$$\sum_{i=1}^{N} x_i \; - \; N.m_1 = 0$$

$$solution \quad \hat{m_1} = \frac{\sum_{i=1}^{N} x_i}{N}$$

et n'oubliez pas la CS2 :

$$\frac{\partial^2 L}{\partial m_1^2} = -\frac{\sum_{i=1}^{N} x_i}{\sigma_1^2} < 0$$

La solution trouvée correspond bien à un maximum.

L'estimateur du maximum de vraisemblance est donc $\hat{m_1}$, moyenne arithmétique des X_i.

4. **Vous disposez dans le tableau 1 de deux indicateurs statistiques, rappellez comment ces indicateurs ont été trouvés (donnez les formules de calcul) ? Quel est le lien entre écart-type et variance ? Quel est l'intérêt de l'écart-type par rapport à la variance ? Quels autres outils statistiques connaissez-vous, qui permettent d'apporter des indications supplémentaires sur la véritable moyenne m_4 du temps nécessaire à l'achat d'un kg de riz dans le monde ?** ***(2 points)***

   - Les 2 indicateurs statistiques sont la moyenne empirique $m = \frac{\sum_{i=1}^{N} x_i}{N}$, et l'écart-type empirique $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}} = \sqrt{variance} = \sqrt{\sigma^2}$
   - L'intérêt de l'écart-type par rapport à la variance est qu'il est exprimé dans la même unité de mesure que la moyenne (ici des minutes de temps de travail, et non des $minutes^2$ comme pour la variance).
   - Autre outil statistique pour avoir des indications sur la moyenne m_4 : calculer un intervalle de confiance à $\alpha\%$ de confiance pour m_4. En renouvellant l'échantillonage un grand nombre de fois, dans $\alpha\%$ des cas l'estimation trouvée pour m_4 sera située dans cet intervalle.

5. **Vous vous interrogez sur la significativité de la différence entre le prix moyen d'un kg de riz, et celui d'un paquet de Marlboro. Vous souhaitez donc connaître (avec une confiance de $\alpha\%$) une borne supérieure de la moyenne du temps nécessaire à l'achat d'un kg de riz. Comment procédez-vous ?** ***(1,5 points)***

   Soit $\widehat{m_4}$ le meilleur estimateur du paramètre m_4.
   La méthode du maximum de vraisemblance a conduit précédemment à l'estimateur suivant : $\widehat{m_4} = \overline{T} = \frac{\sum_{i=1}^{N} T_i}{N}$ pour l'espérance du prix d'un kilogramme de riz.
   $E(\overline{T}) = E(\frac{\sum_{i=1}^{N} T_i}{N}) = \frac{1}{N}\sum_{i=1}^{N} E(T_i) = \frac{1}{N} N m_4 = m_4$
   $V(\overline{T}) = V(\frac{\sum_{i=1}^{N} T_i}{N}) = \left(\frac{1}{N}\right)^2 \sum_{i=1}^{N} V(T_i) = \left(\frac{1}{N}\right)^2 N\sigma_4^2 = \frac{\sigma_4^2}{N}$ car les T_i sont indépendants.

donc $\overline{T} \approx \mathcal{N}\left(m_4, \frac{\sigma_4^2}{N}\right)$.
Soit U_4 la fonction pivotale associée à $\overline{T}$:

$$U_4 = \frac{\overline{T} - m_4}{\sqrt{\frac{\sigma_4^2}{N}}} \approx \mathcal{N}(0,1)$$

Pour trouver une borne supérieure pour m_4, nous cherchons une borne inférieure pour U_4 :

Lecture dans la loi $N[0;1]$: $P\{u_\alpha \leq U_4\} = \alpha\%$ (Remarque, ici u_α sera négatif)

$$P\left\{u_\alpha \leq \frac{\overline{T} - m_4}{\sqrt{\frac{\sigma_4^2}{N}}}\right\} = \alpha\%$$

$$P\left\{u_\alpha \sqrt{\frac{\sigma_4^2}{N}} \leq \overline{T} - m_4\right\} = \alpha\%$$

$$P\left\{u_\alpha \sqrt{\frac{\sigma_4^2}{N}} - \overline{T} \leq -m_4\right\} = \alpha\%$$

$$P\left\{m_4 \leq \overline{T} - u_\alpha \sqrt{\frac{\sigma_4^2}{N}}\right\} = \alpha\% \quad avec\ u_\alpha \leq 0$$

La borne supérieure pour l'espérance du prix d'un kg de riz est donc : $\overline{T} - u_\alpha \sqrt{\frac{\sigma_4^2}{N}}$ avec $u_\alpha \leq 0$.

6. **Application numérique avec $\alpha = 80\%$. *(0,5 point)***

Application numérique avec $\alpha = 80\%$: $u_\alpha = -0,84179$.
(u_α obtenu par approximation linéaire : $\frac{0.8023-0.7995}{0.8023-0.80} = \frac{0.85-0.84}{0.85-u_\alpha}$)
Nous avons donc : $P\left\{m_4 \leq \overline{T} - (-0,8417)\sqrt{\frac{\sigma_4^2}{N}}\right\} = 80\%$.
Soit : $P\{m_4 \leq 23,4369\} = 80\%$.

7. **Afin de pouvoir comparer cette borne supérieure de confiance pour m_4, avec le prix d'un paquet de Marlboro, calculez une <u>borne inférieure de confiance</u> (toujours à 80% de confiance) pour m_1. Que concluez-vous sur la différence de prix moyens entre ces deux biens ? *(1,5 points)***

Même procédé qu'à la question 5 :

le meilleur estimateur du paramètre m_1 est $\widehat{m_1} = \overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$.

$E(\overline{X}) = E(\frac{\sum_{i=1}^{N} X_i}{N}) = \frac{1}{N}\sum_{i=1}^{N} E(X_i) = \frac{1}{N} N m_1 = m_1$

$V(\overline{X}) = V(\frac{\sum_{i=1}^{N} X_i}{N}) = \left(\frac{1}{N}\right)^2 \sum_{i=1}^{N} V(X_i) = \left(\frac{1}{N}\right)^2 N\sigma_1^2 = \frac{\sigma_1^2}{N}$ car les X_i sont indépendants.

donc $\overline{X} \approx \mathcal{N}\left(m_1, \frac{\sigma_1^2}{N}\right)$.
Soit U_1 la fonction pivotale associée à $\overline{X}$:

$$U_1 = \frac{\overline{X} - m_1}{\sqrt{\frac{\sigma_1^2}{N}}} \approx \mathcal{N}(0,1)$$

Pour trouver une borne inférieure pour m_1, nous cherchons une borne supérieure pour U_1 :

Lecture dans la loi $N[0;1]$: $P\{U_1 \leq u_\alpha\} = \alpha\%$ (Remarque, ici u_α sera positif)

$$P\left\{\frac{\overline{X} - m_1}{\sqrt{\frac{\sigma_1^2}{N}}} \leq u_\alpha\right\} = \alpha\%$$

$$P\left\{\overline{X} - m_1 \leq u_\alpha \sqrt{\frac{\sigma_1^2}{N}}\right\} = \alpha\%$$

$$P\left\{\overline{X} - u_\alpha \sqrt{\frac{\sigma_1^2}{N}} \leq m_1\right\} = \alpha\% \quad avec \;\; u_\alpha > 0$$

La borne inférieure pour l'espérance du prix d'un paquet de Marlboro est donc : $\overline{X} - u_\alpha \sqrt{\frac{\sigma_1^2}{N}}$ avec $u_\alpha > 0$.
Application numérique : $u_\alpha = 0,84179$, donc la borne inférieure pour l'espérance du prix d'un paquet de Marlboro à 80% est : 27,4183.

$$P\{27,4183 \leq m_1\} = 80\%$$

8. **La variance du prix moyen d'un kg de pain est inconnue, donnez un encadrement à 90% de confiance de la moyenne m_3.** Indication : variance empirique corrigée : $S^2 = 634$ *(2 points)*

Si l'écart-type n'est pas connu, on ne peut utiliser la variable normale $\frac{\overline{Z} - m_3}{\sigma_3/\sqrt{N}}$ comme fonction pivotale, car elle contient l'inconnue σ_3.

Propriété utilisée : soient $Z_1, ..., Z_n \approx i.i.d.N\left(m_3, \sigma_3^2\right)$: nous notons $\overline{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$ et $S^2 = \frac{\sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)^2}{N-1}$.
Alors : $U = \frac{\overline{Z} - m_3}{S/\sqrt{N}} \approx T_{(N-1)}$, loi indépendante de m_3 et de σ_3^2.
C'est cette variable U que nous devons utiliser pour construire l'intervalle de confiance pour m_3 (la variable S^2 est l'estimation sans biais de σ_3^2).
Lecture dans la table de la loi $T_{(26-1)}$: $P\{|Z| \leq 1,708\} = 0,90$
On en déduit : $P\left\{-1,708 \leq \frac{\overline{Z} - m_3}{S/\sqrt{N}} \leq 1,708\right\} = 0,90$
et donc : $P\left\{\overline{Z} - 1,708\frac{S}{\sqrt{N}} \leq m_3 \leq \overline{Z} + 1,708\frac{S}{\sqrt{N}}\right\} = 0,90$
Pour $N = 26$ et $S^2 = 634$, on obtient $19,915 \leq m_3 \leq 39,784$ Euros.
Rappel : ne pas écrire la probabilité que m_3 soit compris entre 19,915 et 39,784, car il n'y a plus rien d'aléatoire dans l'inégalité une fois l'observation faite : l'événement correspondant est soit réalisé, soit non-réalisé.

Nous souhaitons désormais PREVOIR le prix d'un kg de riz au Kenya.

9. **Quelles hypothèses doit-on poser pour étudier cette question à partir des élèments dont nous disposons ? Ces hypothèses vous paraissent-elles justifiées, expliquez.** ***(1 point)***

   – Hypothèse que le prix du riz du Kenya suit la même loi que le prix du riz dans les autres pays :

   $$T_{Kenya} \approx \mathcal{N}\left(m_4, \sigma_4^2\right)$$

   – Hypothèse que le prix du riz au Kenya est indépendant du prix du riz dans les autres pays :

   $$T_{Kenya} \;\; independant\ de\ T_1, T_2, \ldots T_N$$

   – La seconde parait justifiée si le Kenya est un pays fermé et que le prix mondial du riz n'a pas d'influence sur son prix national, ce qui est évidemment faux.
   – la première semble mise en défaut d'après le tableau 2 : En effet, les prix dans les pays en développés et les pays en développement ne semblent pas suivre la même loi.

10. **Nous considérons que les hypothèses nécessaires pour l'étude de cette question sont vérifiées malgré tout, donnez un encadrement à 90% de confiance du prix PREVU d'un kg de riz au Kenya :** T_k**.** ***(2 points)***

    Le meilleur estimateur de T_{Kenya} est $\overline{T}$ la moyenne empirique du prix du riz sur les autres pays (résultat obtenu en minimisant l'erreur quadratique moyenne MSE). **Cf Correction TD2 exo 9 sur le site du cours.**
    On s'intéresse à l'erreur de prévision : $B = T_{Kenya} - \overline{T}$. Si les hypothèses précédentes sont vérifiées alors :

    $$B \approx \mathcal{N}\left(0, \sigma_4^2 + \frac{\sigma_4^2}{N}\right)$$

    En effet :
    $E(B) = E(T_{Kenya} - \overline{T}) = E(T_{Kenya}) - E(\overline{T}) = m_4 - m_4 = 0$
    $V(B) = V(T_{Kenya} - \overline{T}) = V(T_{Kenya}) + V(\overline{T}) - 2cov(T_{Kenya}, \overline{T}) = \sigma_4^2 + \frac{\sigma_4^2}{N}$ car T_{Kenya} et $\overline{T}$ sont indépendantes.

    La fonction pivotale associée à B est : $U_3 = \frac{B - E(B)}{\sqrt{V(B)}} = \frac{T_{Kenya} - \overline{T}}{\sqrt{\sigma_4^2 + \frac{\sigma_4^2}{N}}} \approx \mathcal{N}(0,1)$
    On cherche donc l'intervalle $[-u_N, u_N]$ qui vérifie :

    $$p\Big(-u_N < \frac{T_{Kenya} - \overline{T}}{\sqrt{\sigma_4^2 \frac{N+1}{N}}} < u_N\Big) = 0,90$$

    D'après les tables statistiques, on a :

    $$p\Big(-1,645 < \frac{T_{Kenya} - \overline{T}}{\sqrt{\sigma_4^2 \frac{N+1}{N}}} < 1,645\Big) = 0,90$$

    On résoud les inégalités et on obtient :

    $$p\left(\overline{T} - 1,645\sqrt{\sigma_4^2 \frac{N+1}{N}} < T_{Kenya} < \overline{T} + 1,645\sqrt{\sigma_4^2 \frac{N+1}{N}}\right) = 0,90$$

D'où l'**intervalle de PREVISION** suivant :

$$p\Big(-5,418 < T_{Kenya} < 47,118 \Big) = 0,90$$

Avec $\overline{T} = 20,85$, $\sigma_4^2 = (15,67)^2$ et $N = 26$.

11. **Procédez également à l'application numérique à partir des informations disponibles dans le tableau 2 (nous considérerons ici que le Kenya est un pays en développement) ? Expliquez d'où provient la différence avec la prévision de la question précédente.** ***(1 point)***

Si l'on considère le tableau 2, et le fait que le Kenya est un pays en développement, l'application numérique devient pour $\overline{T} = 29,23$, $\sigma_4^2 = (18,34)^2$ et $N = 13$:

$$p\Big(-2,078 < T_{Kenya} < 60,538 \Big) = 0,90$$

La différence vient de :
- une moyenne plus élevée pour les pays en développement : $m_{4tot} < m_{4dvt}$
- un écart-type plus élevé pour les pays en développement : $\sigma_{4tot} < \sigma_{4dvt}$
- un échantillon plus petit si l'on prend en compte uniquement les pays en développement
  La borne inférieure négative vient du faible nombre d'observations et de la variance élevée qui est associée à cet échantillon.