Corrigé du partiel 2006/2007

Partie I : Question de cours (1.5 point)

A. On considère le modèle linéaire $y = X\beta + u$. L'estimateur des MCO du paramètre β est une variable aléatoire car son expression $\hat{\beta} = (X'X)^{-1} X'y$ fait intervenir les variables dépendantes (ou expliquées) y_i qui sont elles-mêmes des variables aléatoires. On peut justifier le fait que les y_i soient des variables aléatoires de deux façons:
- Soit on raisonne à partir des hypothèses du modèle (sans les justifier). Dans ce cas, il suffit de dire que les variables y_i dépendent des résidus u_i qui sont, par hypothèse, des variables aléatoires.
- Soit on justifie la modélisation des y_i comme variables aléatoires en disant que les y_i sont généralement obtenues par échantillonnage aléatoire dans une population donnée. Ceci revient à justifier la modélisation aléatoire des résidus si on considère, comme c'est fait dans ce cours, que les X_i ne sont pas des variables aléatoires.

B. Enoncé du théorème central limite :
Si (X_i) est une suite de variables aléatoires indépendantes, de même loi et admettant des moments d'ordre 1 et d'ordre 2 (ou de façon équivalente admettant une espérance m et une variance σ^2 finies) alors, en notant $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, on a :

$$\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[loi]{} N(0,1)$$

Application : deux façons de répondre :
- Soit on donne une ou plusieurs applications du théorème en statistique : intervalles de confiance (lorsque les observations sont suffisamment nombreuses et non normales), tests (détermination de la constante correspondant au risque de première espèce choisi en utilisant la loi asymptotique de la statistique)...
- Soit on applique le théorème à une loi donnée, le cas le plus classique étant celui de la loi de Bernouili qui permet de justifier l'approximation normale de la loi binômiale.

Partie II : Question de cours (2.5 points)

A. 1. La méthode des MCO appliquée au modèle linéaire $y = X\beta + u$ consiste à minimiser la somme des carrés des résidus $\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - X_i\beta)^2$ par rapport au paramètre β (qui peut éventuellement être multidimensionnel) .

2. La méthode du maximum de vraisemblance consiste à maximiser, par rapport aux paramètres du modèle, la vraisemblance de l'échantillon observé, c'est–à-dire la probabilité d'observer cet échantillon (l'échantillon étant supposé i.i.d).

B. 1. Deux façons de justifier le recours à la distance D_n :
- La première consiste à dire que si on remplace les fréquences marginales empiriques par les fréquences marginales théoriques dans l'expression de D_n alors sous l'hypothèse d'indépendance $H_0 : p_{ij} = p_i p_j$, on obtient $D_n = 0$ alors que sous l'hypothèse alternative, on a $D_n \neq 0$. L'intuition derrière ce test est que sous l'hypothèse H_0, il est peu probable que la distance D_n soit "très grande".
- La seconde (plus rigoureuse) consiste à rappeler qu'on connaît la loi asymptotique de la statistique D_n sous l'hypothèse $H_0 : p_{ij} = p_i p_j$, à savoir une loi du χ^2 à $(r-1)(s-1)$ degrés de liberté. On peut alors déterminer, pour un seuil de test donné, la constante C apparaissant dans la région critique de ce test qui est de la forme $D_n > C$.

2. Exemples économiques où ce test peut-être utilisé :
- Dépendance entre l'inflation anticipée (définie qualitativement, c'est-à-dire pour des items de réponse du type : l'inflation va augmenter, diminuer...) et l'inflation antérieure observée (également définie qualitativement).
- Dépendance entre l'épargne (regroupée en tranches) et l'âge (également en tranches).

Partie III : Exercice (5 points)

1. Défintions :
- Une suite de variable aléatoires réelles (X_n) converge en probabilité vers la variable aléatoire X si pour tout $\varepsilon > 0$, on a :

$$P\left(|X_n - X| > \varepsilon\right) \underset{n\to+\infty}{\longrightarrow} 0$$

ou de façon équivalente :

$$P\left(|X_n - X| \leq \varepsilon\right) \underset{n\to+\infty}{\longrightarrow} 1$$

Remarque : les inégalités peuvent être strictes ou larges. En effet, les énoncés avec égalité stricte ou large sont équivalents car la condition doit être vérifiée *pour tout* $\varepsilon > 0$.
- Une suite de variables aléatoires réelles (X_n) converge en moyenne quadratique vers vers la variable aléatoire X si

$$E\left(|X_n - X|^2\right) \underset{n\to+\infty}{\longrightarrow} 0$$

Remarque : la valeur absolue dans cette définition n'est pas nécessaire puisque $|X_n - X|^2 = (X_n - X)^2$ mais elle permet de faire le lien avec le cas général d'une suite de variables aléatoires (X_n) définie sur un espace métrique quelconque. Dans ce cas, il faut remplacer la valeur absolue par la norme $\| \, . \, \|$ définie sur cet espace.

2. La convergence en moyenne quadratique entraîne la convergence en probabilité. L'implication inverse est fausse.

3. Si on note E_{C^P/Y^P} la vraie valeur de l'élasticité de la consommation permanente C^P par rapport au revenu permanent Y^P alors l'hypothèse d'absence de biais de T_n s'écrit

$$E(T_n) = E_{C^P/Y^P}$$

Cette propriété peut s'interpréter de la façon suivante : si on tire de façon aléatoire un très grand nombre d'échantillons puis qu'on fait la moyenne des valeurs prises par l'estimateur T_n sur ces échantillons alors il est très probable que cette moyenne soit très proche de la vraie valeur du paramètre qu'on cherche à estimer, à savoir E_{C^P/Y^P}. A la limite, si on calcule la moyenne de T_n sur tous les échantillons possibles alors on trouve E_{C^P/Y^P}.

4. Vu qu'il n'a pas été stipulé explicitement dans la question qu'on devait utiliser l'inégalité de Tchebychev, on peut répondre à cette question de trois façons différentes:
- Première méthode : on applique l'inégalité de Tchebychev à T_n, ce qui nous donne: pour tout $\varepsilon > 0$,

$$P\left(|T_n - E(T_n)| \geq \varepsilon\right) \leq \frac{V(T_n)}{\varepsilon^2}$$

or $E(T_n) = E_{C^P/Y^P} = 1$ donc

$$P\left(|T_n - 1| \geq \varepsilon\right) \leq \frac{V(T_n)}{\varepsilon^2}$$

mais par hypothèse $V(T_n) \underset{n\to+\infty}{\longrightarrow} 0$ donc, sachant que $P\left(|T_n - 1| \geq \varepsilon\right) \geq 0$, on obtient, en passant à la limite :

$$P\left(|T_n - 1| \geq \varepsilon\right) \underset{n\to+\infty}{\longrightarrow} 0$$

Ceci étant vrai pour tout $\varepsilon > 0$, il s'ensuit (par définition) que la suite (T_n) converge en probabilité vers 1 qui est la vraie valeur du paramètre qu'on cherche à estimer. Ceci veut justement dire que T_n est un estimateur convergent de ce paramètre.
- Deuxième méthode : puisque $E(T_n) = 1$ alors

$$E\left(|T_n - 1|^2\right) = E((T_n - E(T_n))^2) \overset{par\ définition}{=} V(T_n)$$

or par hypothèse $V(T_n) \underset{n\to+\infty}{\longrightarrow} 0$ donc

$$E\left(|T_n - 1|^2\right) \underset{n\to+\infty}{\longrightarrow} 0$$

ce qui veut dire que la suite (T_n) converge en moyenne quadratique vers 1, donc d'après la question 2, elle converge en probabilité vers 1, ce qui permet de conclure : T_n est bien un estimateur convergent de l'élasticité E_{C^P/Y^P} dont la vraie valeur est 1.

-Troisième méthode : on utilise un résultat de cours (Cf formulaire) qui stipule que tout estimateur sans biais et de variance tendant vers 0 quand $n \longrightarrow +\infty$ est convergent. C'est le cas de T_n.

5. Notons N le nombre d'observations dont nous disposons. Soit X le vecteur à N lignes et trois colonnes dont la $i^{ème}$ ligne est donnée par le vecteur $X_i = (\ln Y_i^P, D_i, 1)$, u le vecteur colonne dont la $i^{ème}$ composante, $i = 1, ..., N$, est u_i, et notons y le vecteur colonne dont la $i^{ème}$ composante, $i = 1, ..., N$, est $\ln C_i^P$. En utilisant les variables ainsi définies (ce qui revient à compiler les équations correspondant aux observations), on obtient donc le modèle :

$$y = X\beta + u$$

L'estimateur des MCO de β a pour expression:

$$\hat{\beta} = (X'X)^{-1}X'y$$

6. Sous les hypothèses habituelles du modèle linéaire, cet estimateur est :
- sans biais
- linéaire par rapport à y
- de variance minimale dans la classe des estimateurs linéaires sans biais.

Remarque : si on suppose que les résidus u_i suivent une loi normale, alors $\hat{\beta}$ suit également une loi normale.

Partie IV : Question de cours (2 points)

1. Le risque de premier espèce est le risque (la probabilité) de rejeter *à tort* l'hypothèse nulle H_0. Le risque de seconde espèce est le risque (la probabilité) d'accepter *à tort* l'hypothèse nulle H_0. On privilégie l'hypothèse H_0 dans le sens où on veut que le risque de la rejeter alors qu'elle est vraie soit faible (par exemple 5%). Ce risque d'erreur est justement ce qu'on appelle le risque de première espèce, d'où l'intérêt particulier qu'on lui accorde par rapport au risque de seconde espèce (qui est par ailleurs plus difficile à quantifier en général).

2. Le modèle économétrique de la question 5 de la partie précédente est dérivé du modèle théorique

$$\ln C^P = \beta_1 \ln Y^P + \beta_2 D + \beta_3$$

Dans le cadre de ce modèle, on a $\beta_1 = \frac{\partial \ln C^P}{\partial \ln Y^P}$ or par définition $\frac{\partial \ln C^P}{\partial \ln Y^P} = \frac{\partial C^P}{\partial Y^P}\frac{Y^P}{C^P}$ n'est autre que l'élasticité de C^P par rapport à Y^P , donc β_1 est l'élasticité de C^P par rapport à Y^P . Par conséquent l'hypothèse nulle est :

$$H_0 : \beta_1 = 1$$

et l'hypothèse alternative est :

$$H_1 : \beta_1 \neq 1$$

L'erreur de première espèce consiste à rejeter l'hypothèse H_0 alors qu'elle est vraie, c'est-à-dire à considérer que l'élasticité β_1 est différente de 1 alors qu'en fait cette élasticité est égale à 1.

Partie V : Exercice (6 points)

Le modèle statistique considéré est le suivant : on note X_i, $i = 1, ..., 25$, les dettes des 25 étudiants de l'échantillon (tiré aléatoirement) et on fait l'hypothèse $X_i \overset{i.i.d.}{\rightsquigarrow} N(m, \sigma^2)$. On fait un tirage sans remise mais la taille de la population étudiante ayant souscrit un emprunt est très grande par rapport à la taille de l'échantillon; on peut donc supposer que les variables X_i sont indépendantes (comme si le tirage avait été fait avec remise).

1. Notons $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ la moyenne empirique de la dette sur un échantillon de taille n. Puisque les X_i sont indépendantes et suivent une loi normale, alors $\bar{X}_n$ suit également une loi normale. Plus précisément, il est aisé de montrer que $\bar{X}_n \rightsquigarrow N\left(m, \frac{\sigma^2}{n}\right)$. Lorsqu'on centre et réduit cette variable, on obtient :

$$\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N(0,1)$$

Nous pouvons maintenant déterminer un intervalle de confiance pour le paramètre m lorsque l'écart-type σ est connu. Commençons par poser $U_n = \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}$. Nous cherchons un intervalle bilatéral de confiance au niveau 90%; on doit donc déterminer le nombre β tel que $P(|U_n| \leq \beta) = 0.90$. En notant F la fonction de répartition de la loi $N(0,1)$, il est aisé de montrer, en utilisant la relation $P(|U_n| \leq \beta) = 2F(\beta) - 1$ ou en exploitant graphiquement le caractère symétrique de la densité de $N(0,1)$, que $F(\beta) = 0.95$. Autrement dit, β est égal au fractile d'ordre 0.95 de la loi $N(0,1)$, qu'on note $u_{0.95}$. Ainsi, on a

$$P\left(\left|\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}\right| \leq u_{0.95}\right) = 0.90$$

qu'on peut réécrire sous la forme :

$$P\left(-u_{0.95}\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n - m \leq u_{0.95}\frac{\sigma}{\sqrt{n}}\right) = 0.90$$

ce qui donne

$$P\left(\bar{X}_n - u_{0.95}\frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X}_n + u_{0.95}\frac{\sigma}{\sqrt{n}}\right) = 0.90$$

Le fractile $u_{0.95}$ n'est pas directement donné par la table statistique de $N(0,1)$. Par contre, la table donne les valeurs $F(1.64) = 0.9495$ et $F(1.65) = 0.9505$, or $0.95 =$

$\frac{1}{2}(0.9495 + 0.9505)$ donc par interpolation linéaire, on obtient : $u_{0.95} \simeq \frac{1}{2}(1.64 + 1.65) = 1.645$. L'intervalle $\left[\bar{X}_n - 1.645\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.645\frac{\sigma}{\sqrt{n}}\right]$ est donc un intervalle bilatéral de confiance à 90% pour le paramètre m.
A.N : On a $\sigma = 2500$ euros, $n = 25$ et la valeur observée de $\bar{X}_{25}$ est 10290 euros. On obtient l'intervalle de confiance (en euros) suivant :

$$[9467.5, 11112.5]$$

2. Un raisonnement analogue à celui de la question précédente permet d'aboutir à :

$$P\left(\bar{X}_n - u_{0.995}\frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X}_n + u_{0.995}\frac{\sigma}{\sqrt{n}}\right) = 0.99$$

où $u_{0.995}$ désigne le fractile d'ordre 0.995 de la loi $N(0,1)$. La seule chose qui change par rapport à la question précédente est qu'on remplace le fractile $u_{0.95}$ qui correspond à un intervalle bilatéral de confiance à 90% par le fractile $u_{0.995}$ qui correspond à un intervalle bilatéral de confiance à 99%. Le fractile $u_{0.995}$ n'est pas directement donné par la table statistique de $N(0,1)$. Par contre, nous savons que $F(2.57) = 0.9949$ et $F(2.58) = 0.9951$, or $0.995 = \frac{1}{2}(0.9949 + 0.9951)$ donc par interpolation linéaire, on obtient : $u_{0.995} \simeq \frac{1}{2}(2.57 + 2.58) = 2.575$. L'intervalle $\left[\bar{X}_n - 2.575\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 2.575\frac{\sigma}{\sqrt{n}}\right]$ est donc un intervalle bilatéral de confiance à 99% pour le paramètre m.
A.N : On obtient l'intervalle (en euros) suivant :

$$[9002.5, 11577.5]$$

3. L'augmentation du niveau de confiance entraîne l'augmentation de la largeur de l'intervalle de confiance. Ceci peut être expliqué de deux façons:
- Première façon : on peut utiliser l'expression formelle d'un intervalle de confiance bilatéral au niveau de confiance $1 - \alpha$:

$$\left[\bar{X}_n - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

où $u_{1-\frac{\alpha}{2}}$ désigne le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi $N(0,1)$. La largeur de l'intervalle est donc donnée par $2u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$. Une augmentation du niveau de confiance $1-\alpha$, c'est-à-dire une diminution de α, entraîne une augmentation de $1 - \frac{\alpha}{2}$ ce qui entraîne une valeur plus élevée de $u_{1-\frac{\alpha}{2}}$ (puisque u_p est une fonction croissante de p du fait du caractère croissant de la fonction de répartition F). Il en découle alors une largeur plus élevée de l'intervalle de confiance.
- Seconde façon : on fait appel à notre intuition qui suggère que, de manière générale en statistique, l'obtention d'une assurance (ou confiance) plus grande se fait au détriment de la précision, d'où l'arbitrage classique entre niveau de confiance et précision. En effet,

une assurance plus élevée s'obtient généralement par l'inclusion de plus d'évènements, ce qui nuit à la précision du résultat. La précision dans le cas de l'estimation par intervalle de confiance étant inversement liée à la largeur de l'intervalle de confiance, on s'attend donc naturellement à ce qu'un niveau de confiance plus élevé donne lieu à un intervalle de confiance plus large.

4. Nous disposons de la variance empirique $s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$ et nous avons besoin de la variance empirique modifiée (ou corrigée) $s'^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$ qui est un estimateur sans biais de σ^2 (contrairement à s^2). Il est clair qu'on peut obtenir s' à partir de s en utilisant la relation :

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

qu'on peut réécrire sous la forme :

$$s' = s\sqrt{\frac{n}{n-1}}$$

Dans le cas qui nous intéresse, on a $n = 25$ et la valeur observée de s est 2000 euros, donc la valeur observée de s' est $2041,2$ euros.
Par ailleurs, nous savons que :

$$Z_n := \frac{\bar{X}_n - m}{\frac{s'}{\sqrt{n}}} \rightsquigarrow T_{n-1}$$

où T_{n-1} désigne la loi de Student à $n-1$ degrés de liberté (24 pour la taille de l'échantillon dont on dispose). On cherche un intervalle de confiance à 90% pour m. On doit donc commencer par déterminer le nombre γ tel que $P(|Z_n| \leq \gamma) = 0.90$. En notant G la fonction de répartition de la loi T_{n-1}, il est aisé de montrer, en utilisant la relation $P(|Z_n| \leq \gamma) = 2G(\beta) - 1$ ou en exploitant graphiquement le caractère symétrique de la densité de T_{n-1}, que $G(\beta) = 0.95$. Autrement dit, γ est le fractile d'ordre 0.95 de la loi T_{n-1}, qu'on note $t_{0.95}$. Ainsi, on a

$$P\left(\left|\frac{\bar{X}_n - m}{\frac{s'}{\sqrt{n}}}\right| \leq t_{0.95}\right) = 0.90$$

ce qui donne :

$$P\left(\bar{X}_n - t_{0.95}\frac{s'}{\sqrt{n}} \leq m \leq \bar{X}_n + t_{0.95}\frac{s'}{\sqrt{n}}\right) = 0.90$$

La table de la loi de Student fournit le fractile d'ordre 0.95 de la loi de Student à 24 degrés de libertés T_{24}, à savoir $t_{0.95} = 1.711$. Ainsi, en utilisant le fait que $n = 25$, on peut affirmer que l'intervalle $\left[\bar{X}_{25} - \frac{1.711}{5}s', \bar{X}_{25} + \frac{1.711}{5}s'\right]$ est un intervalle bilatéral de confiance à 90% pour le paramètre m. Pour les valeurs observées de $\bar{X}_{25}$ et de s' (respectivement 10290 euros et 2041.2 euros), on trouve l'intervalle de confiance (en euros) suivant :

$$[9591.5, 10988.5]$$

Déterminons maintenant un intervalle bilatéral de confiance à 95% pour m. Un raisonnement analogue à celui que nous venons de faire permet d'aboutir à:

$$P\left(\bar{X}_n - t_{0.975}\frac{s'}{\sqrt{n}} \leq m \leq \bar{X}_n + t_{0.975}\frac{s'}{\sqrt{n}}\right) = 0.95$$

où $t_{0.975}$ désigne le fractile d'ordre 0.975 de la loi T_{n-1}. La table de Student fournit le fractile d'ordre 0.975 de la loi de Student T_{24}, à savoir $t_{0.975} = 2.064$. Ainsi, en utilisant le fait que $n = 25$, on peut affirmer que l'intervalle $\left[\bar{X}_{25} - \frac{2.064}{5}s', \bar{X}_{25} + \frac{2.064}{5}s'\right]$ est un intervalle bilatéral de confiance à 95% pour le paramètre m. Pour le valeurs observées de $\bar{X}_{25}$ et de s' (respectivement 10290 euros et 2041.2 euros), on trouve l'intervalle de confiance (en euros) suivant :

$$[9447.4, 11132.6]$$

5. L'hypothèse nulle est $H_0 : m \leq 9300$ et l'hypothèse alternative est $H_1 : m > 9300$. Il s'agit d'un test unilatéral contrairement aux intervalles de confiance précédemment calculés qui, eux, étaient bilatéraux. Dans le cas d'un écart-type théorique σ connu, la règle de décision suivante fournit une procédure de test de H_0 contre H_1 au seuil α : on rejette l'hypothèse H_0 si et seulement si:

$$\frac{\bar{X}_n - 9300}{\frac{\sigma}{\sqrt{n}}} > u_{1-\alpha} \tag{1}$$

où $u_{1-\alpha}$ désigne le fractile d'ordre $1-\alpha$ de $N(0,1)$.
- Le seuil $\alpha = 5\%$ correspond au fractile $u_{0.95}$ qui a été calculé précédemment : $u_{0.95} \simeq 1.645$
- Le seuil $\alpha = 1\%$ correspond au fractile $u_{0.99}$ qui n'est pas fourni directement par la table statistique de $N(0,1)$. On peut l'obtenir par interpolation linéaire en utilisant les deux valeurs $F(2.32) = 0.9898$ et $F(2.33) = 0.9901$. En effet, si on cherche le nombre x tel que $0.99 = 0.9898 + x(0.9901 - 0.9898)$, on trouve $x = \frac{2}{3}$, ce qui permet de faire l'approximation linéaire suivante : $u_{0.95} \simeq 2.32 + \frac{2}{3}(2.33 - 2.32) \simeq 2.327$ (si on arrondit supérieurement et 2.326 si on arrondit inférieurement).
- Pour $n = 25$, $\sigma = 2500$ et la valeur observée 10290 de la moyenne empirique, la statistique $\frac{\bar{X}_n - 9300}{\frac{\sigma}{\sqrt{n}}}$ prend la valeur 1.98.

Conclusion :

- Au seuil $\alpha = 5\%$, l'inégalité (1) est vérifiée ($1.98 > 1.645$). On rejette donc l'hypothèse $H_0 : m \leq 9300$ au seuil $\alpha = 5\%$.
- Au seuil $\alpha = 1\%$, l'inégalité (1) n'est pas vérifiée ($1.98 < 2.327$). On accepte donc l'hypothèse $H_0 : m \leq 9300$ au seuil $\alpha = 1\%$.

6. D'après ce qui précède, pour obtenir une réponse négative au test de $H_0 : m \leq 9300$ contre $H_1 : m > 9300$, il faut (et il suffit) que l'inégalité (1) soit vérifiée. Pour $\alpha = 1\%$, cette inégalité peut se réécrire sous la forme :

$$\sqrt{n}\frac{\bar{X}_n - 9300}{\sigma} > u_{0.99}$$

ce qui donne :

$$\sqrt{n} > \frac{\sigma.u_{0.99}}{\bar{X}_n - 9300}$$

c'est-à-dire :

$$n > \left(\frac{\sigma.u_{0.99}}{\bar{X}_n - 9300}\right)^2$$

On sait que $u_{0.99} \simeq 2.327$, $\sigma = 2500$, donc pour une valeur observée de la moyenne empirique $\bar{X}_n$ égale à 10290, on rejette l'hypothèse $H_0 : m \leq 9300$ si et seulement si :

$$n > \left(\frac{2500 \times 2.327}{10290 - 9300}\right)^2 = 34.53$$

Conclusion : Si l'écart-type théorique $\sigma = 2500$ est connu alors pour pouvoir rejeter au seuil $\alpha = 1\%$ l'hypothèse d'une dette moyenne inférieure ou égale à 9300 euros à partir d'un échantillon sur lequel la moyenne empirique est de 10290 euros, il faut que cet échantillon compte au moins 35 individus.

Partie VI : Exercice (3 points)

1. La vraisemblance d'un échantillon $(x_1, x_2, ..., x_n)$ tiré dans la loi normale $N(m, \sigma^2)$ est :

$$\begin{aligned} L(m, \sigma^2; x_1, x_2, ..., x_n) &= \prod_{i=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x_i - m)^2\right)\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2\right) \end{aligned}$$

On en déduit la log-vraisemblance de l'échantillon:

$$\ln L(m, \sigma^2; x_1, x_2, ..., x_n) = -\frac{n}{2}\ln 2\pi - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2$$

donc

$$\frac{\partial \ln L}{\partial m}(m, \sigma^2; x_1, x_2, ..., x_n) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - m) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - nm\right)$$

d'où

$$\frac{\partial \ln L}{\partial m}(m, \sigma^2; x_1, x_2, ..., x_n) = 0 \Longleftrightarrow \sum_{i=1}^{n} x_i - nm = 0 \Longleftrightarrow m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ceci assure qu'en $\hat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i$ la fonction $m \longrightarrow \ln L(m, \sigma^2; x_1, x_2, ..., x_n)$ atteint un extrémum (un maximum ou un minimum) à $(\sigma^2; x_1, x_2, ..., x_n)$ donnés. Pour s'assurer qu'il s'agit bien d'un maximum, il suffit de vérifier que la condition du second ordre $\frac{\partial^2 \ln L}{\partial m^2}(\hat{m}, \sigma^2; x_1, x_2, ..., x_n) < 0$ est vérifiée. Or il est aisé de voir que $\frac{\partial^2 \ln L}{\partial m^2}(\hat{m}, \sigma^2; x_1, x_2, ..., x_n) = -\frac{n}{\sigma^2} < 0$ donc $\hat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i$ est bien un maximum de la fonction $m \longrightarrow \ln L(m, \sigma^2; x_1, x_2, ..., x_n)$. Ainsi, l'estimateur du maximum de vraisemblance du paramètre m n'est autre que la moyenne empirique:

$$\hat{m} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Remarque (et compléments): Le fait de parler d'estimateur du maximum de vraisemblance de m présuppose que m est LE paramètre du modèle statistique considéré, ce qui est le cas par exemple si on considère que la variance σ^2 est connue. Dans ce cas, le paramètre du modèle étant unidimensionnel, la condition du second ordre prend la forme habituelle. Par contre, si la variance σ^2 est inconnue et qu'on cherche à l'estimer également alors LE paramètre du modèle statistique est (m, σ^2) qui est bi-dimensionnel. Dans ce cas, il n'est pas rigoureux mathématiquement de procéder comme on a fait même si on cherche uniquement à estimer m : il faut absolument maximiser la log-vraisemblance par rapport au paramètre du modèle qui est (m, σ^2). La maximisation de la log-vraisemblance par rapport à (m, σ^2) donne lieu à une condition du premier ordre formée des deux équations $\frac{\partial \ln L}{\partial m} = 0$ et $\frac{\partial \ln L}{\partial \sigma^2} = 0$. La résolution de ce système fournit un couple $\left(\hat{m}, \hat{\sigma}^2\right)$. Pour vérifier que la condition du second ordre est satisfaite dans ce cadre bidimensionnel, il faut alors s'assurer que la matrice $\begin{pmatrix} \frac{\partial^2 \ln L}{(\partial m)^2} & \frac{\partial^2 \ln L}{\partial \sigma^2 \partial m} \\ \frac{\partial^2 \ln L}{\partial m \partial \sigma^2} & \frac{\partial^2 \ln L}{(\partial \sigma^2)^2} \end{pmatrix}$ est bien définie négative au point $\left(\hat{m}, \hat{\sigma}^2; x_1, x_2, ..., x_n\right)$.

2. Soit Y une variable aléatoire tirée dans la loi de densité $f(y; a) = ay^{a-1}$ si $0 < y < 1$ et 0 sinon. Pour pouvoir trouver un estimateur de a par la méthode des moments, nous devons obtenir une expression de a en fonction de moments du type $E(g(Y))$, où g est une fonction. Pour cela, essayons d'abord d'exprimer $E(Y)$ en fonction de a. Par définition de l'espérance d'une variable aléatoire continue sur R, on a :

$$E(Y) = \int_{-\infty}^{+\infty} yf(y; a)dy = \int_0^1 yay^{a-1}dy = a\int_0^1 y^a dy = a\left[\frac{y^{a+1}}{a+1}\right]_0^1 = \frac{a}{a+1}$$

A partir de cette relation, on peut exprimer a en fonction de $E(Y)$. En effet, on a :

$$(a+1)\,E(Y) = a$$

d'où

$$a(1 - E(Y)) = E(Y)$$

et donc :

$$a = \frac{E(Y)}{1 - E(Y)}$$

La méthode des moments consiste à estimer les moments théoriques par les moments empiriques. L'estimateur par la méthode des moments de a est donc donné par :

$$\hat{a} = \frac{\bar{Y}}{1 - \bar{Y}}$$

où $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ est la moyenne empirique des observations Y_i.