LFSAB 1105
Probabilité et statistiques
APP 2013-2014

Colognesi Victor (82151100)
Goreux Nicolas (90361100)
Jacques François (82221000)
Libert Alexis (32081100)
Van Verdeghem Joachim (45591100)

20 décembre 2013

Chapitre 1

Question 1

1.1 Méthode des moments

a) Pour cette partie, nous avons choisi un $c = 0.5$ et un $k = 2$. Nous avons utilisé la fonction `histfit` de MATLAB pour tracer l'histogramme de l'échantillon (voir figure 1.1). Le code MATLAB est disponible à l'annexe A

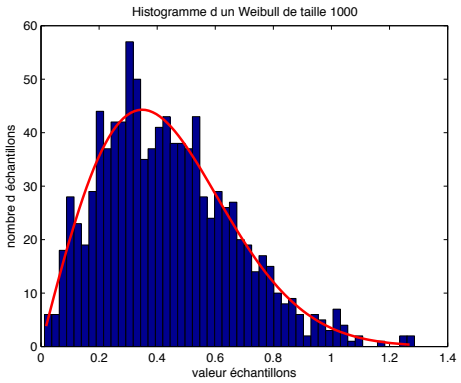

FIGURE 1.1 – Histogramme de 1000 échantillons avec $c = 0.5$ et $k = 2$

Nous trouvons une moyenne, la variance et le coefficient de variation :

$$\text{Moyenne} = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = 0.4431 \tag{1.1}$$

$$\text{Variance} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = 0.0537 \tag{1.2}$$

$$\text{Coefficient de variation} = cv = \frac{S}{\overline{X}} = 0.5227 \tag{1.3}$$

b) Pour trouver l'estimateur $\hat{\theta}_{MM}$, on utilise la méthode des moments. On cherche d'abord le moment d'ordre 1 qui est égal à l'espérance d'ordre 1 et ensuite le moment d'ordre 2 qui est égal à l'espérance

d'ordre 2

$$\mu_1 = \mathbb{E}(X) = \int_0^\infty x\frac{k}{c}\left(\frac{x}{c}\right)^{k-1}\exp\left(\left(-\frac{x}{c}\right)^k\right)dx \tag{1.4}$$

$$= c\Gamma\left(1+\frac{1}{k}\right) \tag{1.5}$$

$$\mu_2 = \mathbb{E}(X^2) = \int_0^\infty x^2\frac{k}{c}\left(\frac{x}{c}\right)^{k-1}\exp\left(\left(-\frac{x}{c}\right)^k\right)dx \tag{1.6}$$

$$= c^2\Gamma\left(1+\frac{2}{k}\right) \tag{1.7}$$

Connaissant la moyenne trouvée précédemment et la moyenne quadratique qui est :

$$\text{Moyenne quadratique} = \frac{\sum_{i=1}^n X_i^2}{n} = 0,2551 \tag{1.8}$$

Nous égalisons la variance d'ordre 1 à la moyenne et l'espérance d'ordre 2 à la moyenne quadratique. On isole c :

$$c = \frac{\overline{X}}{\Gamma\left(1+\frac{1}{k}\right)} \tag{1.9}$$

Nous devons dès lors résoudre l'équation suivante :

$$\frac{\sum_{i=1}^n X_i^2}{n} = \left(\frac{\overline{X}}{\Gamma\left(1+\frac{1}{k}\right)}\right)^2\Gamma\left(1+\frac{2}{k}\right) \tag{1.10}$$

c) Nous avons utilisé la méthode de la sécante pour trouver la valeur de k. On remplace cette valeur dans l'équation avec c (voir équation 1.9) On trouve alors un $\hat{k}_{MM}$ et un $\hat{c}_{MM}$
On peut alors calculer l'erreur quadratique :

$$\text{ERT} = (\hat{k}_{MM}-k)^2 + (\hat{c}_{MM}-c)^2 = 0.0403 \tag{1.11}$$

d) Nous faisons maintenant 500 réplications pour trouver 500 $\hat{k}_{MM}$, $\hat{c}_{MM}$ et ERT. Nous mettons tous ces résultats dans 3 vecteurs. On calcule ensuite la moyenne et la variances de chaques séries.

	Moyenne	Variance
$\hat{k}_{MM}$	2.0223	0.0291
$\hat{c}_{MM}$	0.4996	6.9557e-07
ERT	0.0295	0.002

e) On trace les box-plot et l'histogramme (voir figure 2)

f) Pour les simulations *Matlab* que nous avons effectué, nous avons pris 1000 échantillons (n). On trouve pour la méthode des moments un $Bias(\hat{\theta})$ de (0.0223,0.004) ce qui est proche de (0,0). On peut encore réduire celui-ci en augmentant le nombre d'échantillons. Enfin, on peut extrapoler qu'en tendant n à l'infini, on sera unbiased ce qui fait de notre estimateur un estimateur asymptotically unbiasad.
Pour la variance, elle est relativement faible, on sait donc que nos points seront pour la plupart concentrés.
Il est également consistant car l'erreur (ERT) tend vers 0 quand n tend vers l'infini.

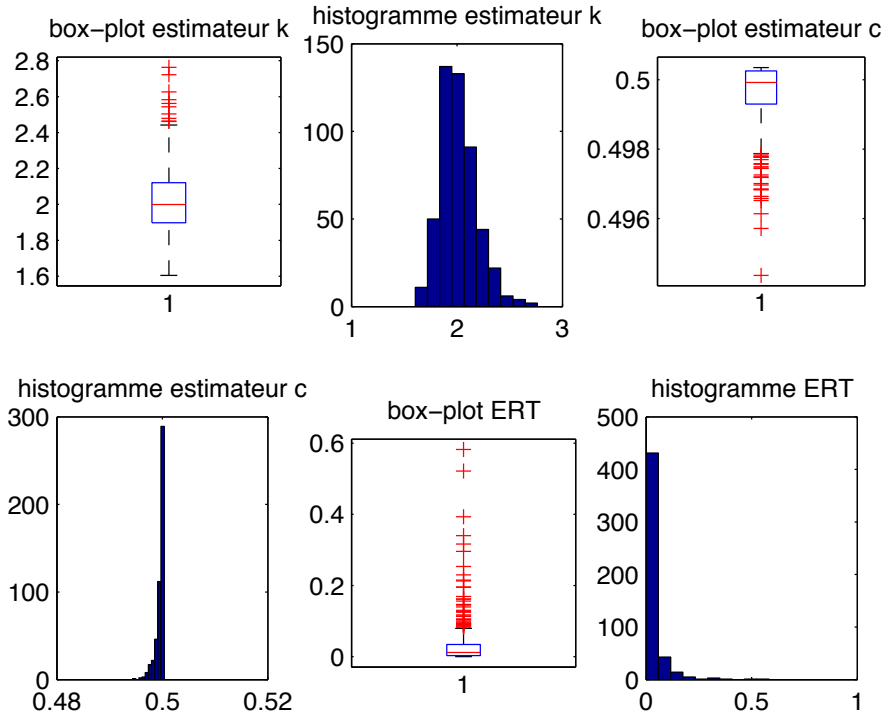

FIGURE 1.2 – Histogramme et box-plot des $\hat{k}_{MM}, \hat{c}_{MM}$ et ERT par la méthode des moments

1.2 Méthode graphique

Ici, nous voulons trouver les coefficients $\hat{k}$ et $\hat{c}$ de manière à approcher au mieux k et c. Nous savons que k et c vérifient :

$$\ln\left(-\ln\left(1-F(x)\right)\right) = k\cdot\ln\left(x\right) + k\cdot\ln\left(c\right)$$

On voudra donc que les estimations soient telles que l'erreur soit la plus petite au sens des moindres carrés. Autrement dit, nous voudrions que l'expression suivante soit minimale :

$$\sum_{i=1}^{n}\left[\hat{k}\cdot\ln\left(x_i\right) + \hat{k}\cdot\ln\left(\hat{c}\right) - \ln\left(-\ln\left(1-\hat{F}(x_i)\right)\right)\right]^2$$

où $\hat{F}(x_i)$ est l'équivalent empirique de la fonction de répartition :

$$\hat{F}(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n}$$

En faisant le changement de variable $y = \ln(x)$ et en posant $-\ln\left(-\ln\left(1-\hat{F}(x_i)\right)\right) = U_i$, l'expression à minimiser devient :

$$\sum_{i=1}^{n}\left[\hat{k}\cdot y_i + \hat{k}\cdot\ln(\hat{c}) - U_i\right]^2$$

On voit bien qu'il s'agit d'une approximation linéaire par les moindres carrés (les coefficients de la droite sont respectivement $a = \hat{k}$ et $b = \hat{k}\cdot\ln\hat{c}$) et sans entrer dans de longs et fastidieux développements,

nous pouvons utiliser les acquis du cours de méthodes numériques et énoncer que les coefficients seront donnés par le système suivant :

$$\mathbf{A}^t \cdot \mathbf{A} \cdot \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{A}^t \cdot \mathbf{U}$$

dans lequel $\mathbf{A}$ est la matrice de *Vandermonde* pour une approximation linéaire :

$$\mathbf{A} = [\mathbf{Y}\ \mathbf{1}]$$

et $\mathbf{Y}$ et $\mathbf{U}$ sont simplement les vecteurs colonnes contenant les valeurs des y_i et des U_i et $\mathbf{1}$ est un vecteur colonne composé de n entrées égales à 1.

Le code MATLAB résolvant ce système linéaire est disponible à l'annexe B. Les résultats sont représentés sous forme de boxplot et d'histogramme pour 500 échantillons de 1000 essais chacun à la figure 1.3.

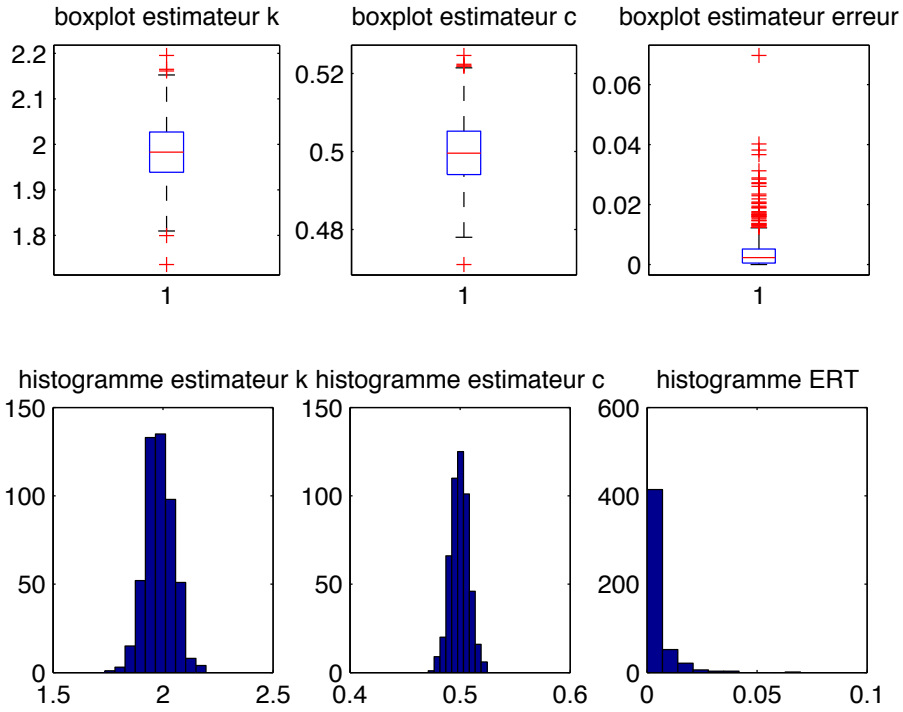

FIGURE 1.3 – Boxplots et histogrammes de (de gauche à droite) $\hat{k}$, $\hat{c}$ et l'erreur quadratique totale obtenus avec une méthode graphique

Pour ces trois séries, nous avons repris la moyenne et la variance dans la table suivante :

	Moyenne	Variance
$\hat{k}_{MLE}$	1.9833	.004001
$\hat{c}_{MLE}$	.4997	0.000067
ERT	.00434	.0000428

f) A nouveau, on prend un 1000 échantillons (n). On trouve pour la méthodes graphiques un $Bias(\hat{\theta})$ de (-0.0167,0.003) ce qui est proche de (0,0). On peut encore réduire celui-ci en augmentant le nombre

d'échantillons. Enfin, on peut extrapoler qu'en tendant n à l'infini, on sera unbiased ce qui fait de notre estimateur un estimateur asymptotically unbiasad.
Pour la variance, elle est relativement faible, on sait donc que nos points seront pour la plupart concentrés.
Il est également consistant car l'erreur (ERT) tend vers 0 quand n tend vers l'infini.

1.3 Méthodes de maximum de vraisemblance

a) La fonction de vraisemblance pour une loi de Weibull de paramètres (k, c) pour n essais est donnée par :

$$L(k,c) = \prod_{i=1}^{n} \frac{k}{c}\left(\frac{x_i}{c}\right)^{k-1} \exp\left[-\left(\frac{x_i}{c}\right)^{k}\right]$$

Au vu de l'expression, il est plus aisé d'utiliser la logarithme népérien de cette fonction :

$$\begin{aligned} LL(k,c) &= \sum_{i=1}^{n} \ln\left(\frac{k}{c}\left(\frac{x_i}{c}\right)^{k-1} \exp\left[-\left(\frac{x_i}{c}\right)^{k}\right]\right) \\ &= \sum_{i=1}^{n} \left[\ln\left(\frac{k}{c}\right) + (k-1)\ln\left(\frac{x_i}{c}\right) - \left(\frac{x_i}{c}\right)^{k}\right] \end{aligned}$$

Afin trouver le maximum de cette fonction de vraisemblance, nous dérivons d'abord cette fonction par rapport à c et nous égalons le résultat à zéro :

$$\begin{aligned} \frac{\partial LL}{\partial c} &= \sum_{i=1}^{n} \left[-\frac{1}{\hat{c}} - \frac{\hat{k}-1}{\hat{c}} - (-\hat{k})\frac{x_i^{\hat{k}}}{\hat{c}^{\hat{k}+1}}\right] \\ &= -\frac{\hat{k}}{\hat{c}} \cdot \sum_{i=1}^{n} \left[1 - \left(\frac{x_i}{\hat{c}}\right)^{\hat{k}}\right] \\ &= 0 \\ &\Updownarrow \\ 0 &= \sum_{i=1}^{n} \left[1 - \left(\frac{x_i}{\hat{c}}\right)^{\hat{k}}\right] \\ &\Updownarrow \\ n &= \tfrac{1}{\hat{c}^{\hat{k}}} \cdot \sum_{i=1}^{n} x_i^{\hat{k}} \end{aligned}$$

De cette relation, nous trouvons :

$$\hat{c} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\hat{k}}\right)^{1/\hat{k}}$$

Nous dérivons à présent par rapport l'expression LL par rapport à k :

$$
\begin{aligned}
\frac{\partial LL}{\partial k} &= \sum_{i=1}^{n}\left[\frac{1}{\hat{k}}+\ln\left(\frac{x_i}{\hat{c}}\right)-\ln\left(\frac{x_i}{\hat{c}}\right)\cdot\left(\frac{x_i}{\hat{c}}\right)^{\hat{k}}\right] \\
&= \sum_{i=1}^{n}\left[\frac{1}{\hat{k}}+\ln\left(\frac{x_i}{\hat{c}}\right)\cdot\left(1-\left(\frac{x_i}{\hat{c}}\right)^{\hat{k}}\right)\right] \\
&= \frac{n}{\hat{k}}+\sum_{i=1}^{n}\ln(x_i)\cdot\left(1-\left(\frac{x_i}{\hat{c}}\right)^{\hat{k}}\right)-\ln(\hat{c})\cdot\underbrace{\sum_{i=1}^{n}\left(1-\left(\frac{x_i}{\hat{c}}\right)^{\hat{k}}\right)}_{=0 \text{ par la relation sur } \hat{c}} \\
&= \frac{n}{\hat{k}}+\sum_{i=1}^{n}\ln(x_i)\cdot\left(1-\left(\frac{x_i}{\hat{c}}\right)^{\hat{k}}\right) \\
&= \frac{1}{\hat{k}}+\frac{\sum_{i=1}^{n}\ln(x_i)}{n}-\sum_{i=1}^{n}\ln(x_i)\frac{x_i^{\hat{k}}}{\sum_{j=1}^{n}x_j^{\hat{k}}} \\
&= \frac{1}{\hat{k}}+\frac{\sum_{i=1}^{n}\ln(x_i)}{n}-\frac{\sum_{i=1}^{n}\ln(x_i)\,x_i^{\hat{k}}}{\sum_{i=1}^{n}x_i^{\hat{k}}} \\
&= 0
\end{aligned}
$$

De cette relation, nous trouvons aisément :

$$
\hat{k}=\left[\frac{\sum_{i=1}^{n}x_i^{\hat{k}}\ln(x_i)}{\sum_{i=1}^{n}x_i^{\hat{k}}}-\frac{\sum_{i=1}^{n}\ln(x_i)}{n}\right]^{-1}
$$

b) Pour calculer l'estimateur du maximum de vraisemblance, il nous suffit donc de résoudre les deux équations dérivées au point précédent. Pour ce faire, nous avons utilisé la méthode de la sécante. Nous n'avons pas utilisé la marche à suivre proposée dans l'énoncé parce qu'il nous semblait trop arbitraire de choisir une grille sur aucun critère et aussi moins précis que l'usage des équations vérifiées par l'estimateur du maximum de vraisemblance.

Le programme qui nous a permis de calculer les estimateurs est disponible à l'annexe C. Nous y avons directement réalisé les sous-questions de c) à e). Les boxplots et diagrammes obtenus à la sous question e) sont représentés à la figure 1.4.

La moyenne et la variance des trois séries sont disponible au tableau suivant :

	Moyenne	Variance
$\hat{k}$	2.0027	.002759
$\hat{c}$	.4994	.00006533
ERT	.002826	.00001392

f) A nouveau, on prend un 1000 échantillons (n). On trouve pour la méthodes graphiques un $Bias(\hat{\theta})$ de (0.0027,0.006) ce qui est proche de (0,0). On peut encore réduire celui-ci en augmentant le nombre d'échantillons. Enfin, on peut extrapoler qu'en tendant n à l'infini, on sera unbiased ce qui fait de notre estimateur un estimateur asymptotically unbiasad.
Pour la variance, elle est relativement faible, on sait donc que nos points seront pour la plupart concentrés.
Il est également consistant car l'erreur (ERT) tend vers 0 quand n tend vers l'infini.

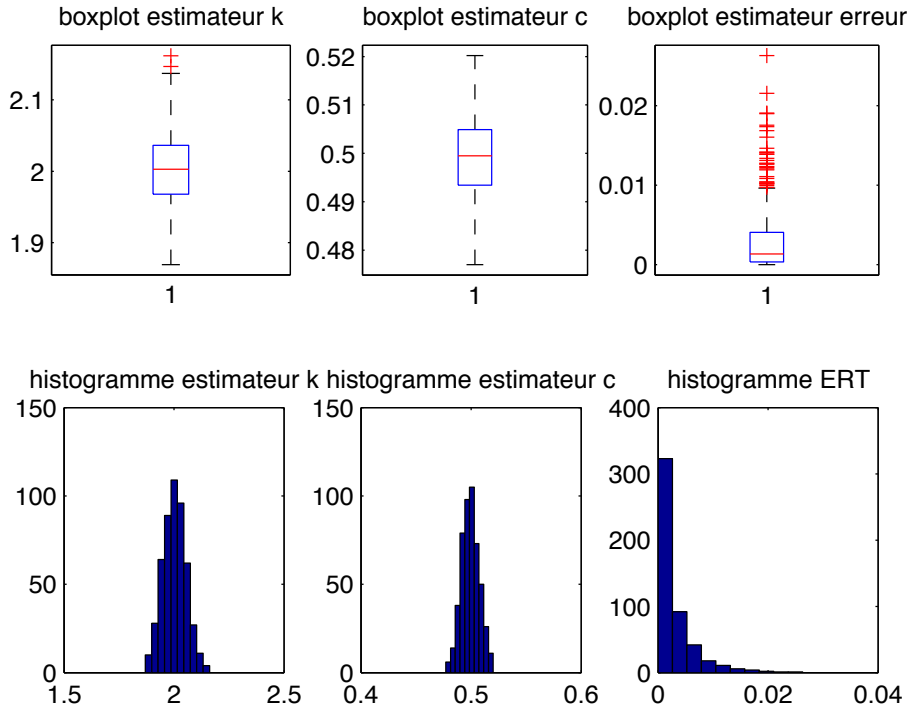

FIGURE 1.4 – Boxplots et histogrammes de (de gauche à droite) $\hat{k}$, $\hat{c}$ et l'erreur quadratique totale obtenus avec la méthode du maximum de vraisemblance

1.4 Comparaison

Le tableau suivant reprend les valeurs de la moyenne et de la variance de $\hat{k}$, $\hat{c}$ et ERT pour les trois méthodes utilisées :

	Moyenne	Variance
$\hat{k}_{MM}$	2.0223	0.0291
$\hat{c}_{MM}$	0.4996	6.9557e-07
ERT_{MM}	0.0295	0.002
$\hat{k}_{MG}$	1.9833	.004001
$\hat{c}_{MG}$	.4997	0.000067
ERT_{MG}	.00434	.0000428
$\hat{k}_{MLE}$	2.0027	.002759
$\hat{c}_{MLE}$	.4994	.00006533
ERT_{MLE}	.002826	.00001392

On y voit directement que la méthode du maximum de vraisemblance est la meilleure au niveau de l'erreur quadratique totale. C'est en effet pour cette méthode que cette erreur sera la plus petite. D'ailleurs, les variances des différentes séries sont également plus petites (sauf pour le $\hat{c}$ de la méthode des moments), ce qui veut dire que les valeurs sont plus concentrées autour de la moyenne. Et cette dernière est également plus proche de la bonne valeur (rappelons nous, nous avons tout fait avec $k = 2$ et $c = .5$) avec la méthode du maximum de vraisemblance.

Chapitre 2

Question 2

1) La moyenne, la médiane, l'écart-type, l'étendue et le coefficient de variation sont donnés par :

$\overline{X}$	10.5952
Médiane	10.5
S	2.6427
Etendue	12.5
cv	0.2494

2) En appliquant la méthode des moments pour une loi Normale, nous trouvons des estimateurs pour la moyenne et la variance :

$$\hat{\mu} = \overline{X} = 10.5952$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} = 6.9839$$

Tout d'abord, observons que le paramètre m correspond au k du premier exercice tandis que $\alpha = c^k$. Nous décidons d'utiliser la méthodes des moments : nous réutilisons donc les mêmes équations que pour la question 1.1.b. Grâce au code MATLAB disponible en annexe D, nous trouvons des estimateurs pour les paramètres α et m de la Weibull :

$$\hat{m} = 4.5637$$

$$\hat{c} = 11.6008$$

$$\hat{\alpha} = 7.2114 \cdot 10^4$$

Il nous est maintenant possible de représenter les fonctions de probabilités des lois normale et Weibull avec les paramètres calculés :

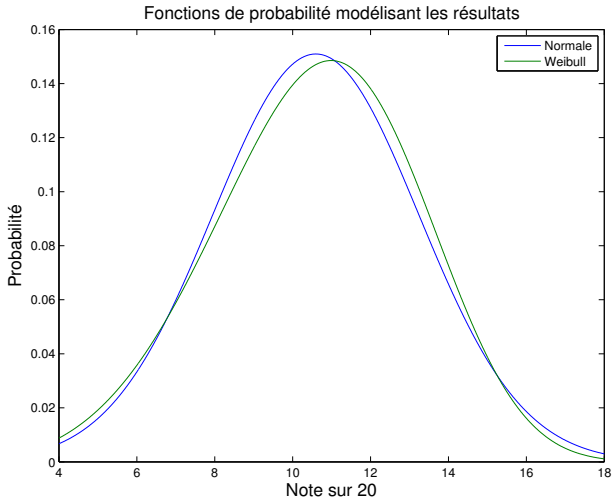

FIGURE 2.1 –

3) Nous observons que les deux fonctions sont extrêmement proches l'une de l'autre : nous utiliserons donc la loi normale dans nos calculs, possédant des tables pour celle-ci.

4)

5) L'intervalle de confiance pour un paramètre θ avec θ_L et θ_U la limite de confiance inférieur et supérieur doit être tel que :

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha \tag{2.1}$$

Nous utilisons intervalle de confiance pour des grand échantillons ($n > 30$). Nous pouvons donc utiliser le théorème de la *Central Limit* :

$$\frac{\hat{\theta} - \theta}{\sigma/\sqrt{n}} = Z \sim N(0,1) \tag{2.2}$$

En injectant cela dans l'équation 2.1, on a :

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha \tag{2.3}$$

Et on arrive à :

$$P\left(\hat{\theta} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{2.4}$$

Nous avons donc finalement avec $\alpha = 0.05$, $\hat{\theta} = \overline{y} = 12$ et $\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} = 2.6427$:

$$z_{\alpha/2} = 1.96\frac{2.6427}{\sqrt{1000}} \tag{2.5}$$

L'intervalle de confiance est :

$$[11.8362; 12.16] \tag{2.6}$$

6) $H_0 : p = 0.8$ vs l'alternative $H_1 : p < 0.8$

Le statistique test, lequel est basé sur $\hat{p} = Y/n$ est donné par :

$$Z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \tag{2.7}$$

On utilise $\sqrt{p_0(1-p_0)/n}$ car on utilise la distribution sous H_0, c'est plus approprié. Nous avons donc trouvé que $\hat{p} = 87/252 = 0.3452$ qui est le nombre de personne avec une note au dessus de 12. $p_0 = 0.8$ et $n = 252$. On trouve que $Z = -18.049$. On prend par défaut un $\alpha = 0.05$. Pour RH_0 il faut que $Z < -z_\alpha$ $-18.049 < -2.57$. Nous refusons dès lors H_0