TP Principes et Méthodes Statistiques

Gabriel Sarrazin, Nejmeddine Douma, Simon Rabourg

Avril 2015

1 Analyse des défauts de cuves

2 Vérifications expérimentales à base de simulations

1. Il est possible de simuler n échantillons de la loi Pa(a,b) car nous connaissons sa fonction de répartiton.
   $P_a(a,b)$ est une fonction continue, elle peut donc s'apparenter à une loi uniforme. Dans un premier temps, simuler n échantillons de cette loi va donc consister à tirer, au hasard, n valeurs aléatoires sur l'intervalle [0,1]. Connaissant la fonction de répartition de la loi $P_a(a,b)$,nous allons ensuite calculer l'image inverse $F^{-1}(Ui)$ pour obtenir un échantillon de loi $P_a(a,b)$ et nous ferons cela pour les n valeurs obtenues sur [0,1].

$$\begin{aligned} U &= 1 - b^a/F^{-1}(U)^a \\ \Longrightarrow -F^{-1}(U)^a &= -b^a/U - 1 \\ \Longrightarrow F^{-1}(U)^a &= b^a/1 - U \\ \Longrightarrow -F^{-1}(U) &= b/(1-U)^{(1/a)} \end{aligned}$$

   Nous pouvons représenter cette méthode sous forme d'un graphique : en mettant en ordonnée les n valeurs de la loi Ui et en abscisse la projection pour chacune de ses valeurs de son image inverse ($F^{-1}(Ui)$).

2. En suivant la méthode décrite précedemment, nous avons simuler m échantillon de taille n avec différentes valeurs pour m,n et a.
   Nous avons ensuite, pour chaque échantillon de taille n, calculer l'intervalle de confiance bilatéral. Pour cela nous avons utiliser L'intervalle de confiance trouvé en première partie qui s'utilise avec $Y = \ln \dfrac{X}{2}$. A chaque fois que a est bien contenu dans cet intervalle, nous incémentons une variable compteur. La proportion Pr d'IC contenant a est donc $Pr = Compteur/m$.

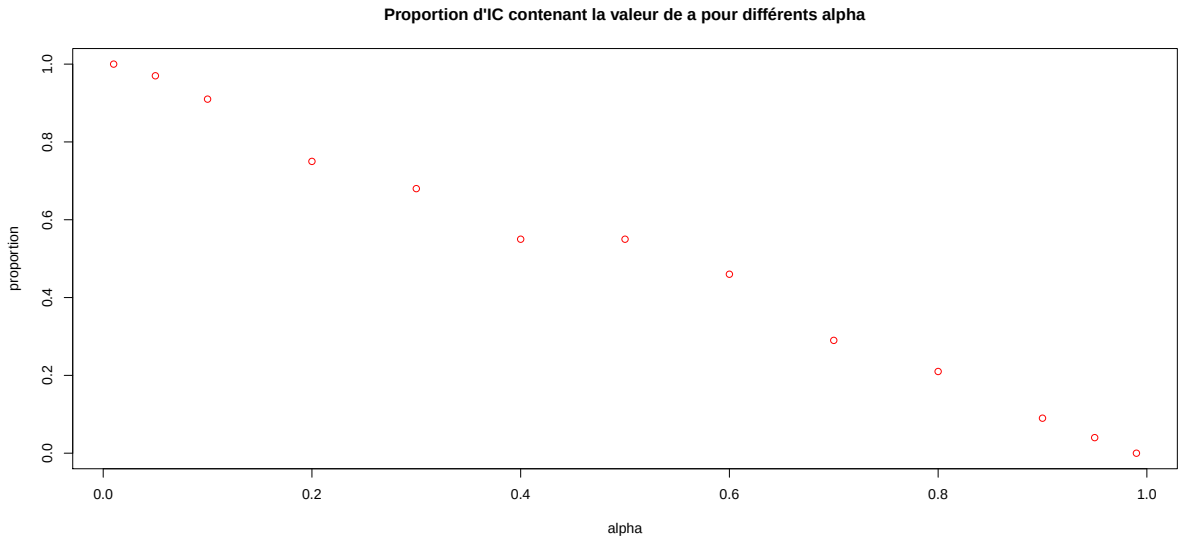

FIGURE 1 – Graphe des proportions d'IC(α) contenant la valeur exacte de a pour différentes valeurs de α

Afin que cela soit plus représentatif,nous avons tracé un graphique avec les différentes proportions obtenues en fonction des α. (Voir figure page 2)

Nous pouvons voir qu'il existe une linéarité entre la proportion et les α. Plus on augmente le nombre d'échantillon m et leur taille n et plus l'approximation est exacte. Ces points ont été obtenus pour les valeurs de $m = 100$, $n = 10, 30, 50, 100, 200, 500$, $a = 3, 5, 10, 20, 50$ et $\alpha = 0.01, 0.05, 0.10, 0.20..., 0.80, 0.90, 0.95, 0.99$.
Quand on simule un grand nombre m déchantillons de taille n de la loi $P_a(a, 2)$ alors la proportion d'IC(α) contenant a est aproximativement égale à $1 - \alpha$.

3. Afin d'estimer le paramètre a, nous disposons de trois méthodes : la méthode des moments, la méthode du maximum de vraisemblance qui nous mènent au même résultat pour l'éstimation du paramètre a et nous pouvons aussi déterminer l'estimateur sans bais de variance minimale si celui-ci existe.

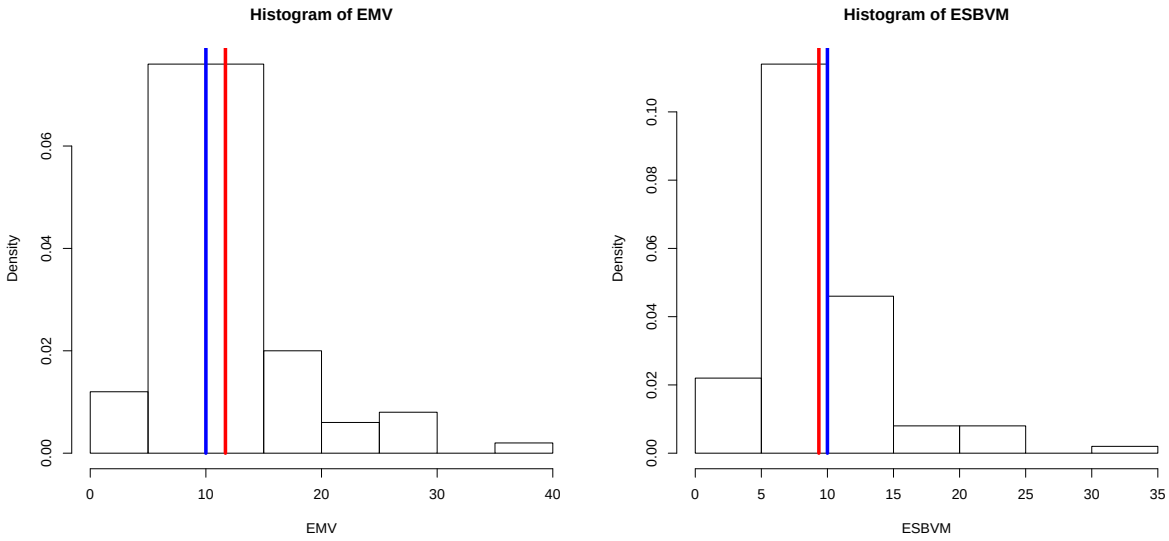

FIGURE 2 – Histogramme des EMV et ESBVM pour $m = 100, n = 5, a = 10$

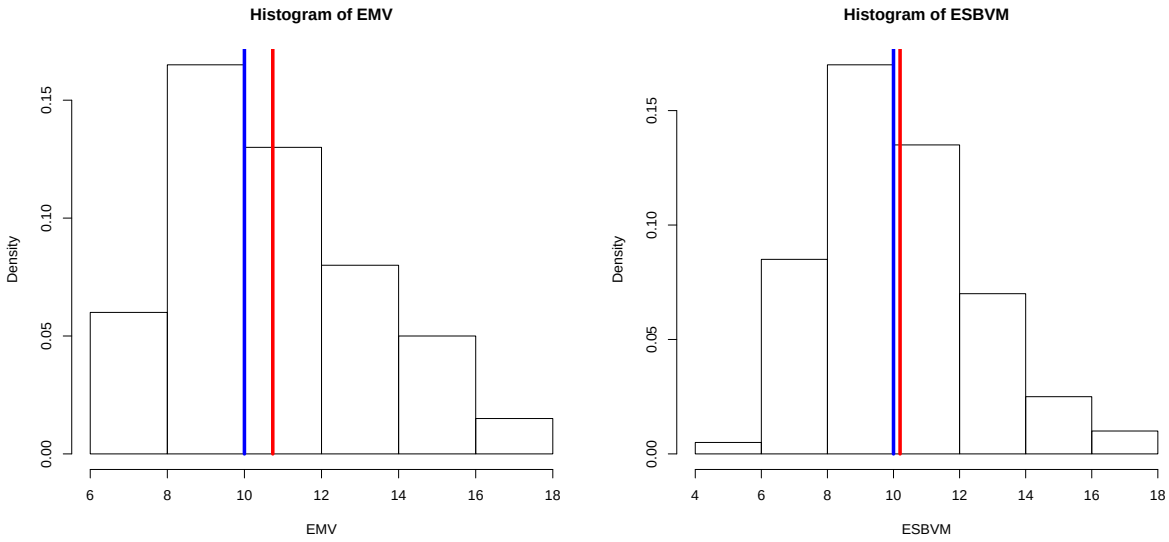

FIGURE 3 – Histogramme des EMV et ESBVM pour $m = 100, n = 20, a = 10$

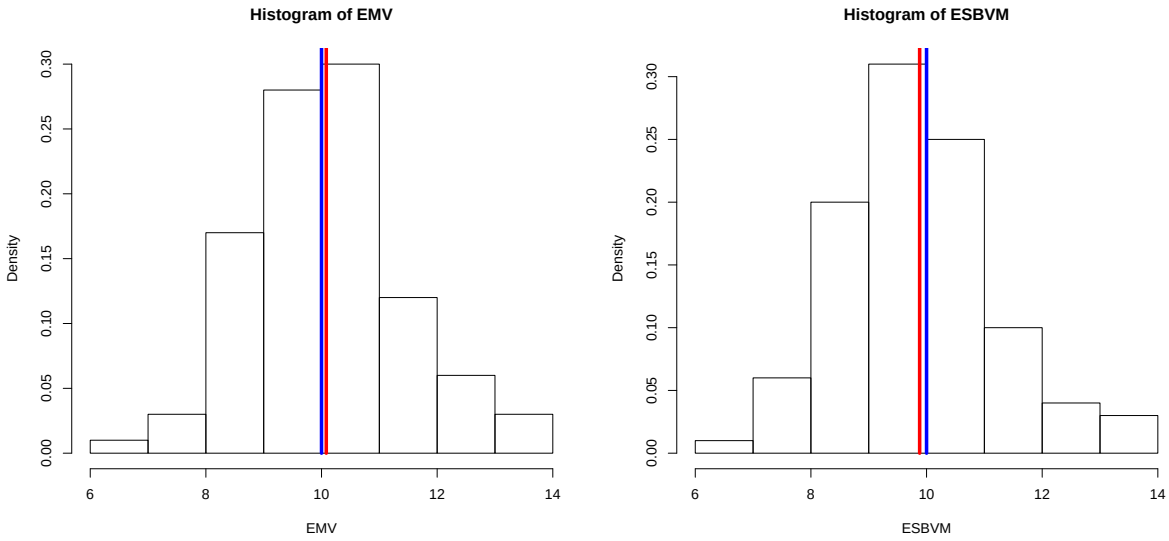

FIGURE 4 – Histogramme des EMV et ESBVM pour $m = 100, n = 50, a = 10$

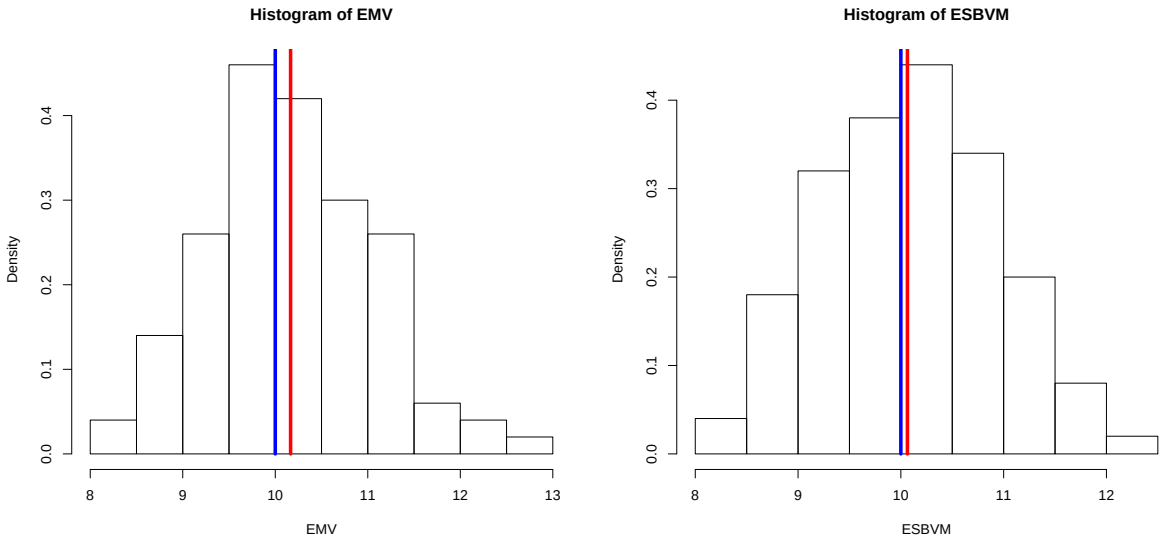

FIGURE 5 – Histogramme des EMV et ESBVM pour $m = 100, n = 100, a = 10$

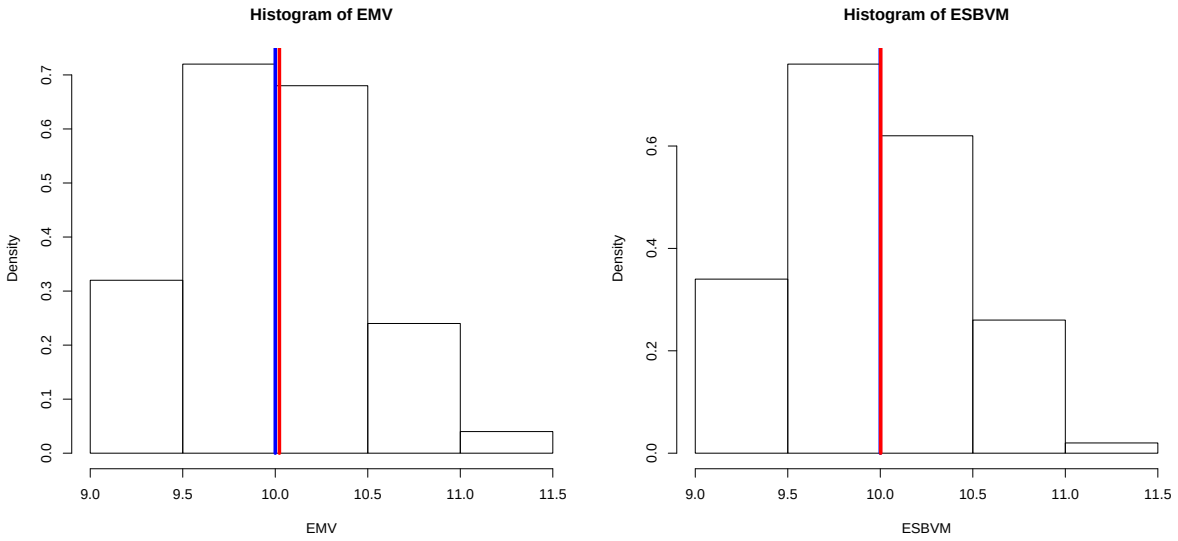

FIGURE 6 – Histogramme des EMV et ESBVM pour $m = 500, n = 5, a = 10$

Pour chaque échantillons (m en tout) nous avons mis au point sur R une fonction qui calcule les estimateurs proposés , qui estime le biais et l'erreur quadratique de chaque estimateur, et qui fait une moyenne pour chacun de ces résultats. Par ces résultats, nous pouvons tracer un histogramme des estimateurs obtenus avec en rouge : la moyenne de ces estimateurs et en bleu : la valeur exacte de a pour laquelle il faut être la plus proche (voir figures pages 4 , 5) .

Pour chaque graphique voici les moyennes des biais et des EQM obtenus pour chaque estimateur :

n	$\bar{biais}EMV$	$\bar{biais}ESBVM$	$\bar{EQM}EMV$	$\bar{EQM}ESBVM$
5	1.691318	-0.6469459	38.3953	23.16078
20	0.7355605	0.1987825	6.691049	5.589889
50	0.08228447	-0.1193612	1.867297	1.801097
100	0.1657237	0.06406643	0.8215394	0.7823774
500	0.02238977	0.002344989	0.207069	0.2057478

Après analyse des résultats et des graphiques, nous pouvons en déduire que le meilleur estimateur est l'$ESBVM$ car il possède le biais et l'erreur quadratique moyen le plus faible sur l'ensemble des échantillons. Il faut noter que plus la taille des échantillons est élevée, plus les estimateurs sont précis. Nous pouvons tout de même constater que quelque soit le taille de l'échantillon, l'$ESBVM$ est le meilleur estimateur.

4. Dans cette partie nous simulons à nouveau m échantillons de taille n suivant la loi $P_a(a, 2)$. Nous calculons la moyenne empirique à l'aide de la fonction *mean* disponible sur R et nous calculons son Esperance. Pour chaque valeur n allant de 5 à 500 nous calculons la difference absolue de ces deux paramètres. Nous avons ensuite enregistrer le nombre de fois où la valeur dépasse une valeur $\epsilon(Ici, \epsilon = 1)$. Sur la figure qui suit, nous pouvons remarquer que plus la taille de l'échantillon

grandit, moins la différence entre la moyenne et l'esperance de cet échantillon est grande.

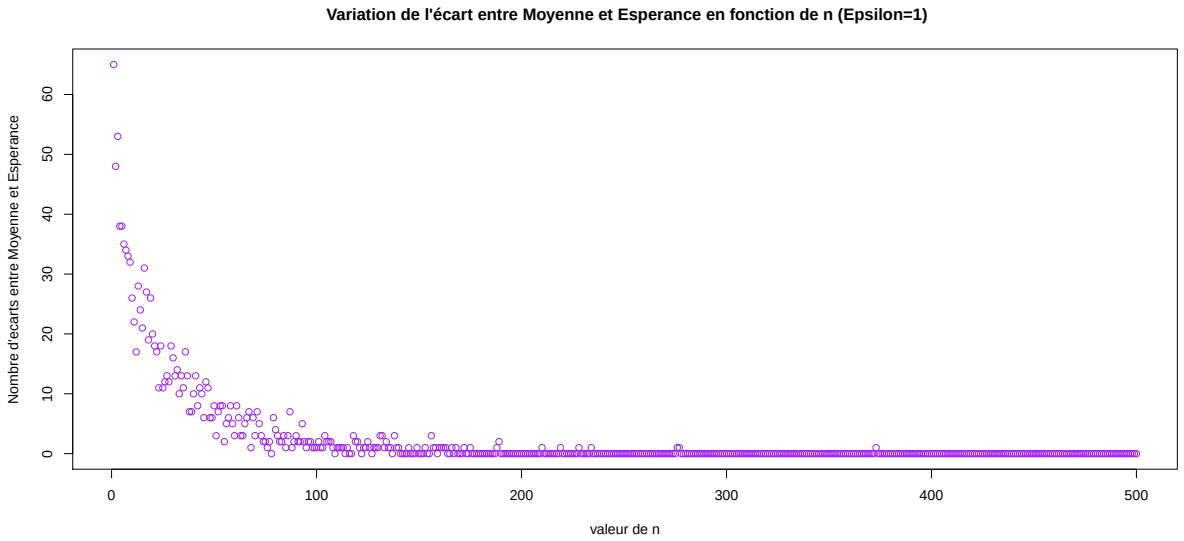

FIGURE 7 – Variation de la différence absolue entre Moyenne est Esperance en fonction de n

Par conséquent, plus n est grand, moins la moyenne empirique $\bar{X}_n$ s'éloigne de l'Esperance $E(X)$ d'au moins ϵ. On a donc

$$\forall \varepsilon > 0, \quad \lim_{n \to +\infty} \mathbb{P}\left(\left| \frac{X_1 + X_2 + \cdots + X_n}{n} - E(X) \right| \geqslant \varepsilon \right) = 0$$

Autrement dit, (X_n) converge en probabilité vers $E(X)$. La moyenne empirique est bien un estimateur convergent de l'ésperance.

5. Après avoir simuler m échantillons de taille n suivant la loi $P_a(a, 2)$, nous calculons leur moyenne. Nous obtenons donc un échantillon de m moyennes empiriques. Pour différentes valeurs de n (Voir figures), nous avons tracé un histogramme et un graphe de probabilités pour la loi normale à l'aide de la function R *qqnorm* qui permet de comparer graphiquement la distribution de l'échantillon des m moyennes empiriques avec une distribution normale.

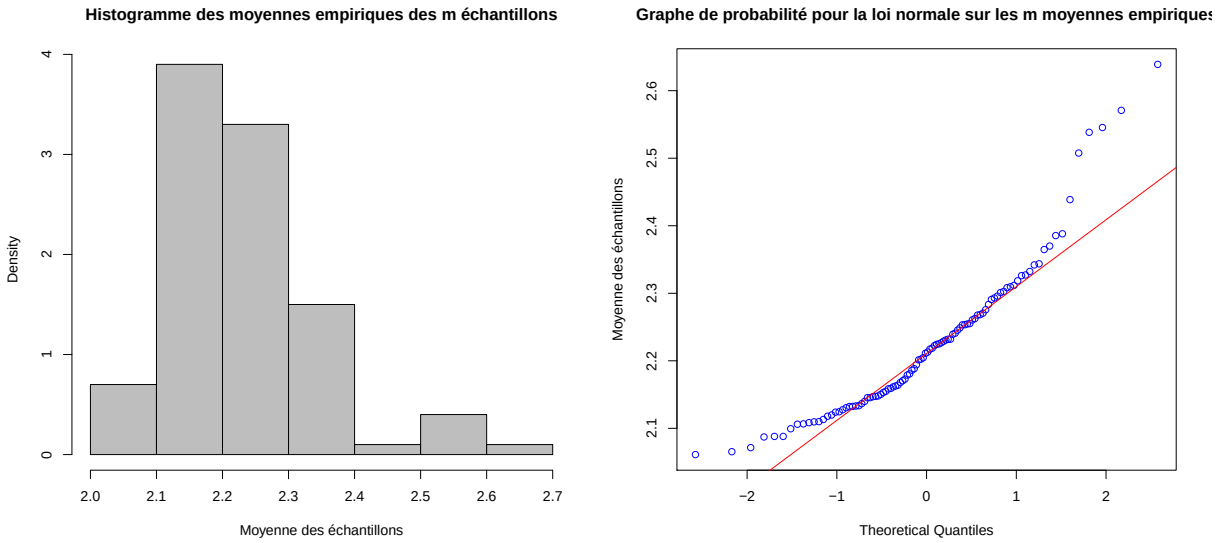

FIGURE 8 – Etude de la distribution $\bar{X}_n$ par une distribution normale pour n=5

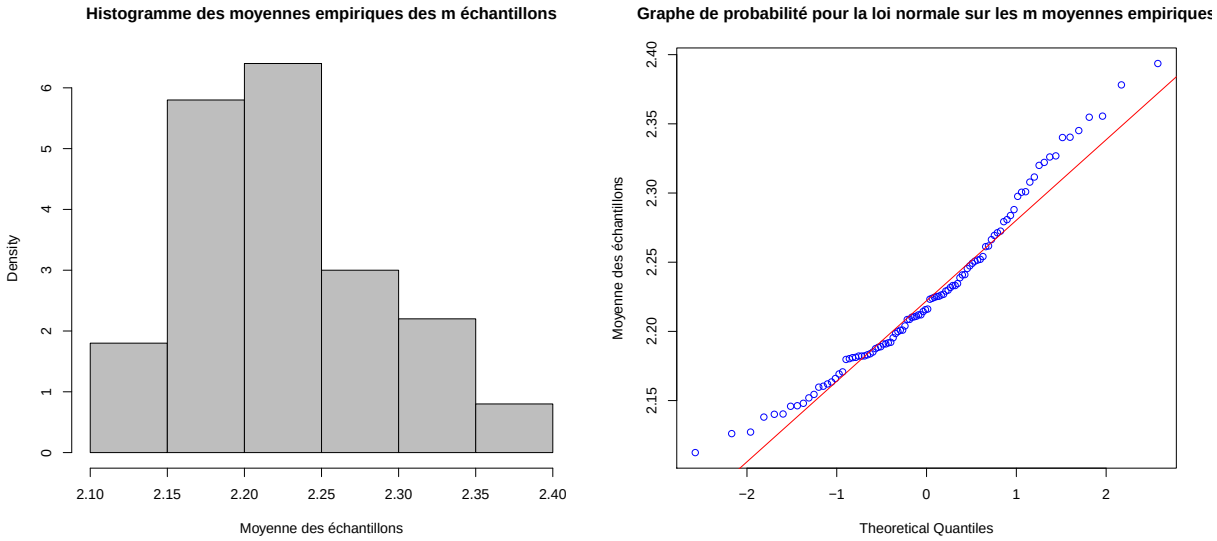

FIGURE 9 – Etude de la distribution $\bar{X}_n$ par une distribution normale pour n=20

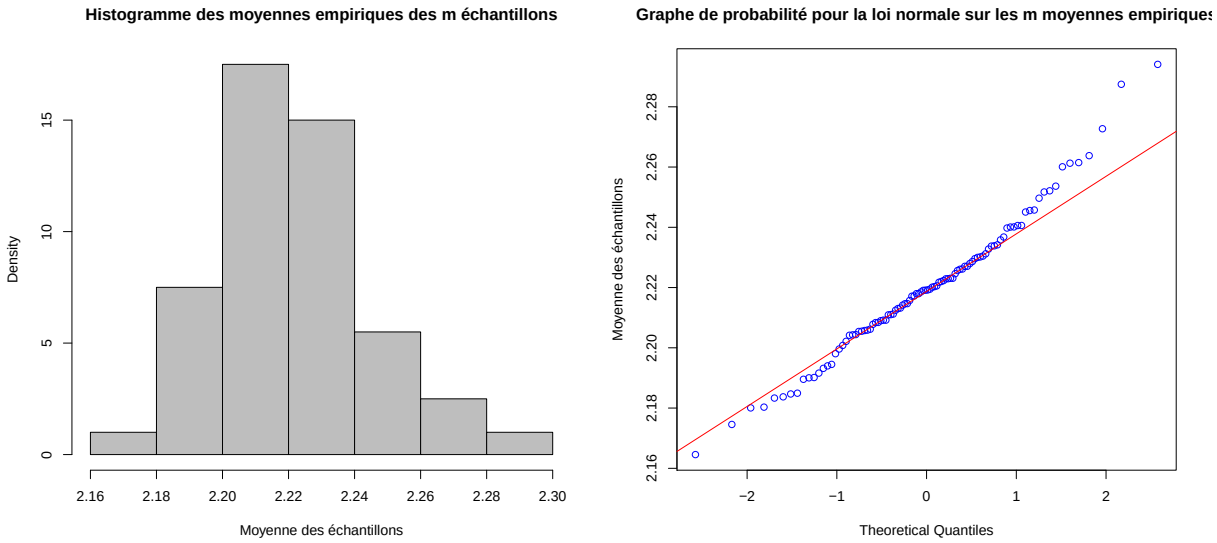

FIGURE 10 – Etude de la distribution $barX_n$ par une distribution normale pour n=100

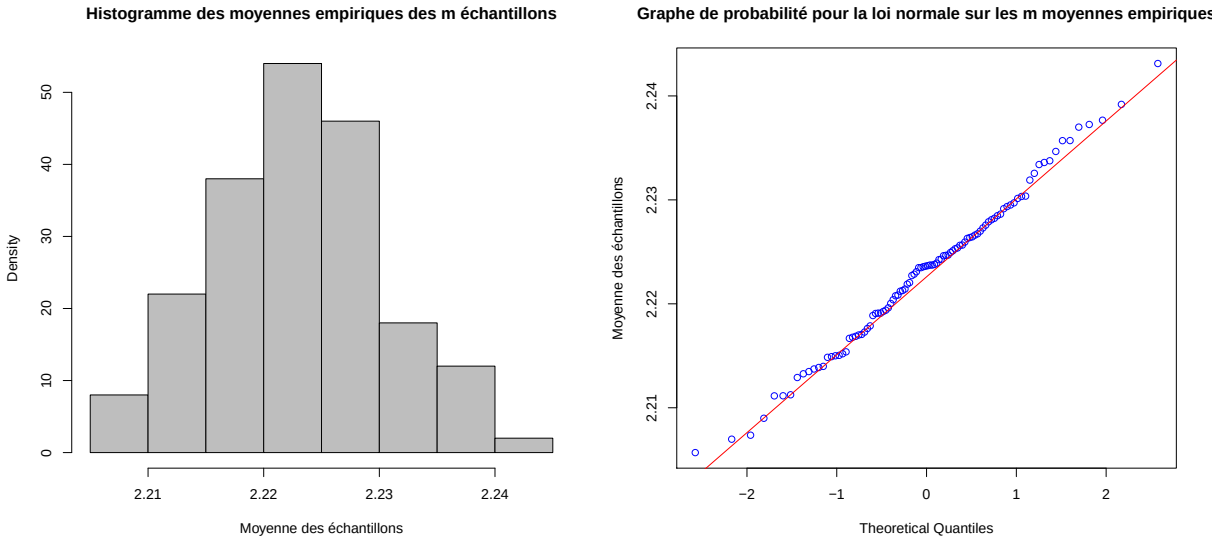

FIGURE 11 – Etude de la distribution $\bar{X}_n$ par une distribution normale pour n=1000

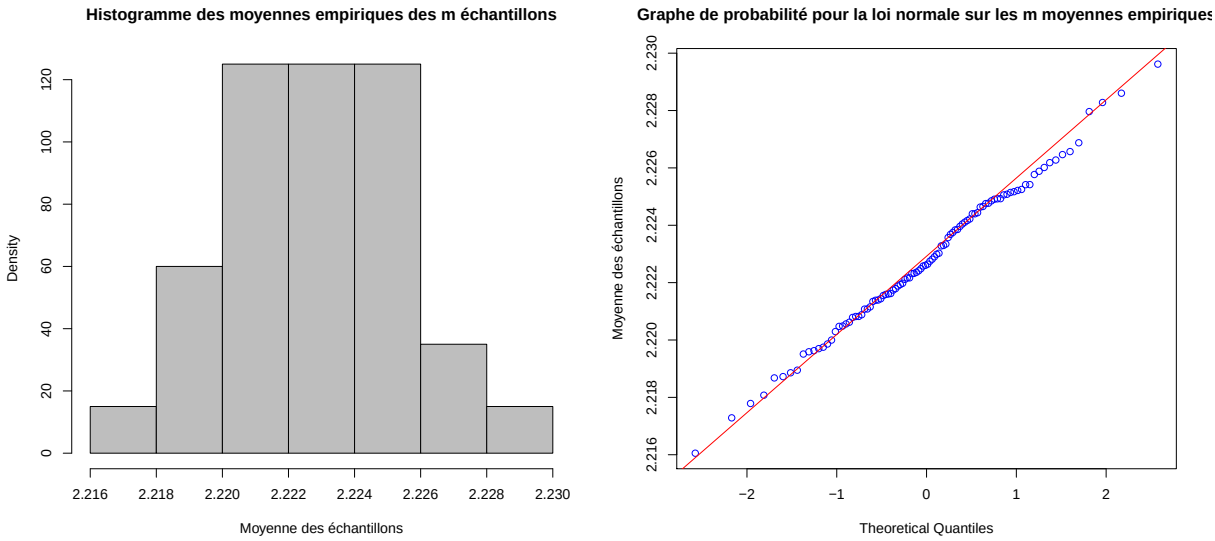

FIGURE 12 – Etude de la distribution $\bar{X}_n$ par une distribution normale pour n=10000

Nous pouvons constater que pour $n = 5$ il est difficile de dire que la loi normale est un modèle approrié car la courbe a une allure logarithmique. Cependant, plus n augmente et plus les point sont alignés. De plus nous pouvons remarquer plus n augmente et plus l'histogramme tend à avoir une allure de la densité $N(0, 1)$
Nous pouvons donc en déduire que

$$\forall n \geq 1, \sqrt{n}\frac{\overline{X_n} - E[X]}{\sigma[X]} \xrightarrow[L]{} N(0, 1)$$

est vérifié expérimentalement.

6. Fixons pour cette partie a=0.5.
Pour étudier la convergence en loi des estimateurs, nous étudions la fonctions de répartition empirique. Nous vérifions expérimentalement que plus n est grand, plus la fonction de répartition emprique obtenue a une allure de la fonction de répartition $1 - (2/t)^a$ qui suit la loi $P(a, 2)$

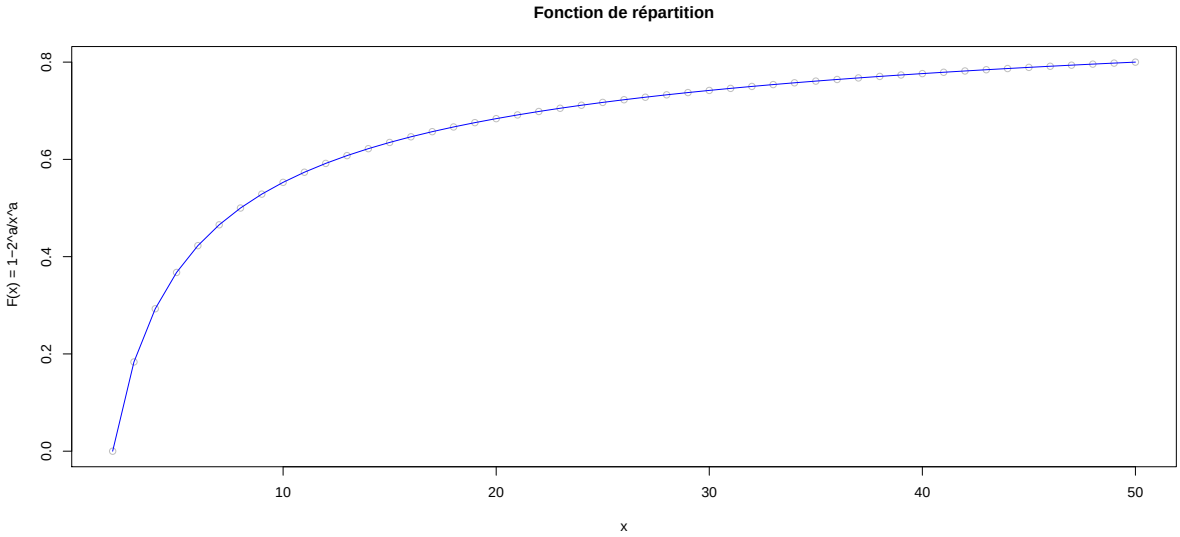

FIGURE 13 – Fonction de répartition $F(X) = 1 - (2/t)^a$

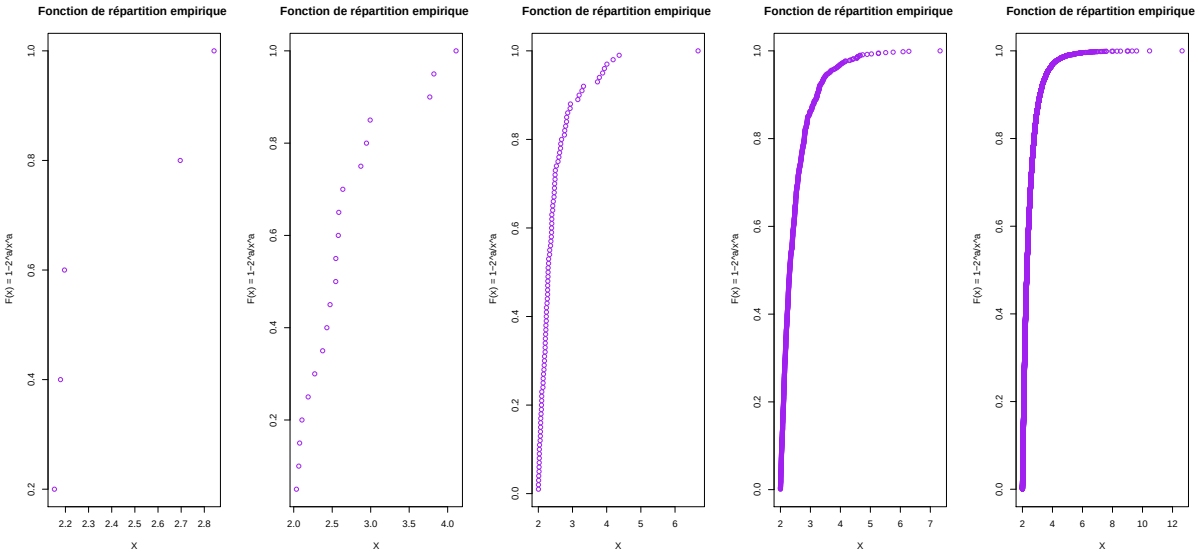

FIGURE 14 – Fonctions de répartition avec $n = 5, 20, 100, 1000, 10000$

Ci dessus vous pouvez voir en premier la fonction de répartition puis les fonctions de répartitions empiriques pour différentes valeurs de n. Nous avons donc bien

$$\forall x \lim_{n \to +\infty} F_n(x) = F(x)$$