TP Principes et Méthodes Statistiques

Gabriel Sarrazin, Nejmeddine Douma, Simon Rabourg

Avril 2015

1 Analyse des défauts de cuves

1. Les mesures des trois cuves présentent des valeurs minimums assez proches les unes des autres: 2.007, 2.006 et 2.059 ce qui s'explique par la précision de l'appareil A qui ne détecte que les défauts de taille supèrieure à 2 mm. La cuve 1 s'empare de la valeur maximale 6.416, la variance la plus grande 1.046943 , le maximum des écat-types 1.023202 et du maximum du coefficient de variation empirique 0.3563262. Tandis que la cuve 2 possède le minimum de la valeur médiane 2.362 et de la valeur moyenne 2.592. Les mesures de la cuve 3 présentent le plus de régularité avec le minimum de variation 0.15907528, le minimum d'écart-type 0.4163554 et de coefficient de variation empirique 0.1475989 .

   D'après les allures des histogrammes des mesures de la cuve 1 (figures 1 et 2) et celles des mesures de la cuve 2 (figures 3 et 4), ces deux échantillions sont vraisemblablement de loi exponentielle. Les figures 5 et 6 montrent que les mesures de la cuve 3 sont vraisemblalement de loi normale.

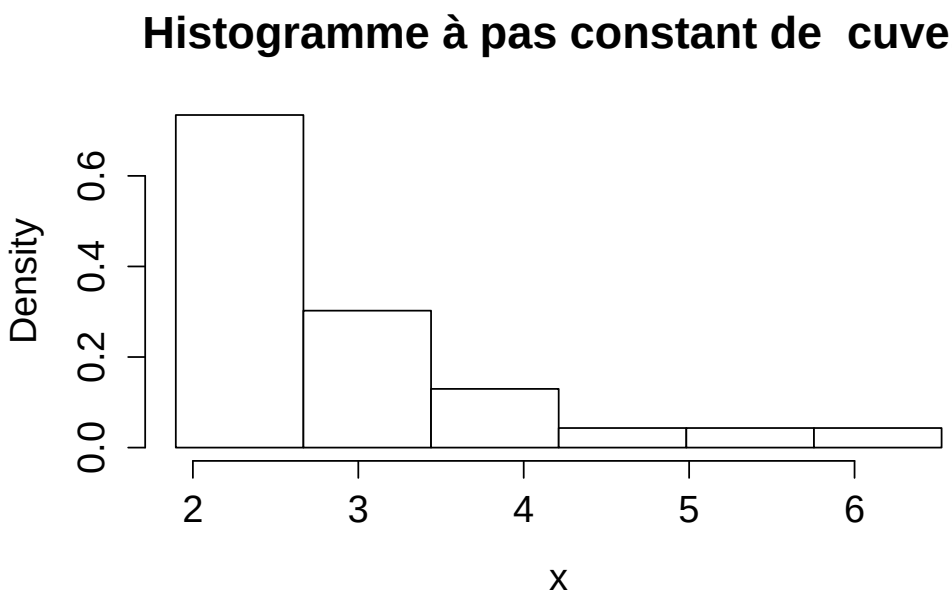

Figure 1: Histogramme à pas constant de cuve1 obtenu dans R

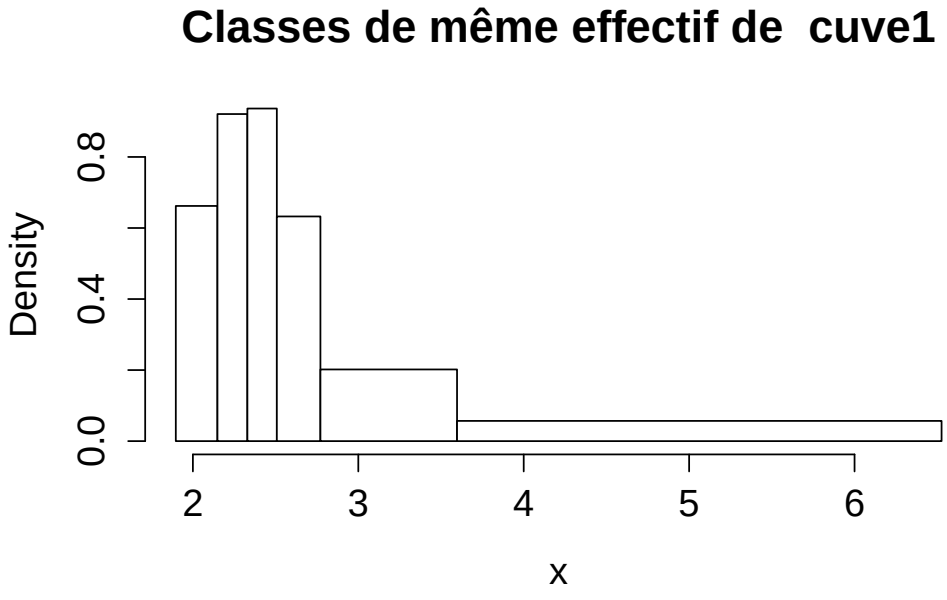

Figure 2: Histogramme à classe de même effectif de cuve1 obtenu dans R

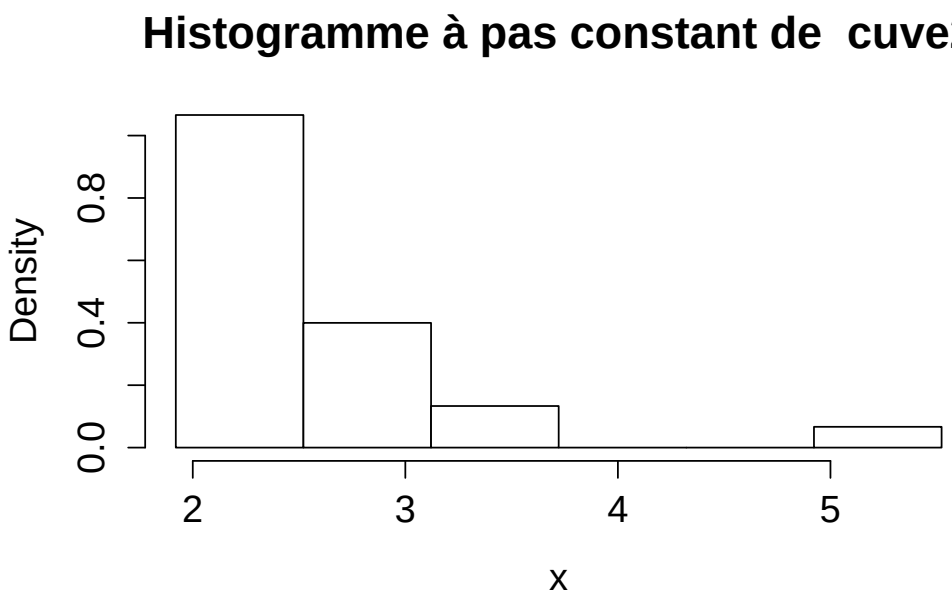

Figure 3: Histogramme à pas constant de cuve2 obtenu dans R

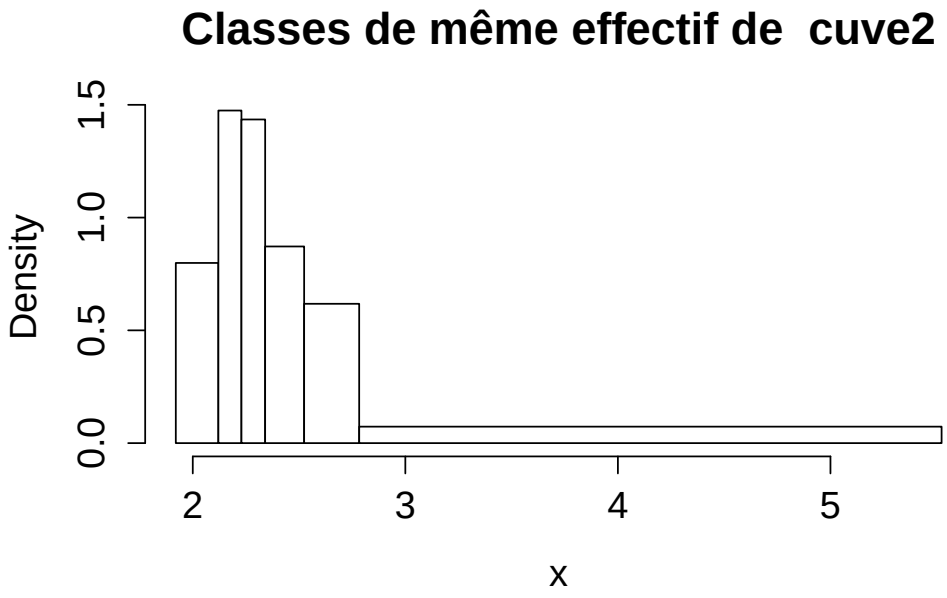

Figure 4: Histogramme à classe de même effectif de cuve2 obtenu dans R

2. Caculons F_x, la fonction de répartition de X.

   X est une variable aléatoire de loi $\mathcal{P}a(a,2)$, sa densité est :

$$f(x) = \frac{a\,2^a}{x^{1+a}}\,\mathbb{1}_{[2,+\infty[}(x)$$

   Donc,

$$\begin{aligned} F_x(x) &= \int_{-\infty}^{x} \frac{a\,2^a}{t^{1+a}}\,\mathbb{1}_{[2,+\infty[}(t)\,\mathrm{d}t \\ &= \begin{cases} \int_2^x a 2^a\, t^{-(1+a)}\,\mathrm{d}t & si\, x > 2 \\ 0 & sinon. \end{cases} \\ F_x(x) &= \begin{cases} 1 - 2^a\, x^{-a} & six > 2 \\ 0 & sinon. \end{cases} \end{aligned}$$

   Le théorème de transfert donne:

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{+\infty} t\,\frac{a\,2^a}{t^{1+a}}\,\mathbb{1}_{[2,+\infty[}(t)\,\mathrm{d}t \\ &= a\,2^a \int_2^{+\infty} \frac{1}{t^a}\,\mathrm{d}t \end{aligned}$$

   Donc,

$$\mathcal{P}a(a,2)\,\text{admet une espérance finie} \Leftrightarrow a > 1$$

$$\begin{aligned} Var(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= a\,2^a \int_2^{+\infty} \frac{1}{t^{a-1}}\,\mathrm{d}t - (a\,2^a \int_2^{+\infty} \frac{1}{t^a}\,\mathrm{d}t)^2 \end{aligned}$$

   Donc,

$$\mathcal{P}a(a,2)\,\text{admet une variance finie} \Leftrightarrow a > 2$$

3. Caculons F_Y, la fonction de répartition de Y=ln $\frac{X}{2}$.

$$\begin{aligned}
F_Y(x) &= \mathbb{P}(Y < x) \\
&= \mathbb{P}(\ln \frac{X}{2} < x) \\
&= \mathbb{P}(\ln X - \ln 2 < x) \\
&= \mathbb{P}(X < \exp(x + \ln 2)) \\
&= \mathbb{P}(X < 2\exp(x)) \\
&= F_X(2\exp(x)) \\
&= \begin{cases} 1 - 2^a(2\exp(x))^{-a} & si\, x > 2 \\ 0 & sinon. \end{cases} \\
F_Y(x) &= \begin{cases} 1 - \exp(-ax) & si\, x > 2 \\ 0 & sinon. \end{cases}
\end{aligned}$$

Donc Y suit la loi $\mathcal{E}$(a).

4. Trouvons une fonction pivotale pour déterminer l'expression d'un intervalle de confiance de seuil α pour a. On a besoin des trois lemmes suivant:

Lemme 1: Si λ>0, μ >0 et X une variable aléatoire réelle de loi $\mathcal{E}(\lambda)$ alors $\mu X \sim \mathcal{E}(\frac{\lambda}{\mu})$

Lemme 2: Si $X_1, X_2, ..., X_n$ une famille de variables aléatoires réelles indépendantes et identiquement distriuées de loi $\mathcal{E}(\lambda)$ alors $\sum_{i=0}^{n} X_i \sim$ Γ(n,λ).

Lemme 3: Γ(n, $\frac{1}{2}$) et χ^2_{2n} décrivent la même loi de probabilité.

Soit $Y_1, Y_2, ..., Y_n$ une famille de variables aléatoires réelles indépendantes et identiquement distriuées de loi $\mathcal{E}$(a). $2aY_1, 2aY_2, ..., 2aY_n$ sont donc de loi $\mathcal{E}(\frac{1}{2})$ (Lemme 1). D'où, 2a$\sum_{i=0}^{n} Y_i \sim \Gamma$(n,$\frac{1}{2}$) (Lemme 2). Donc, 2a$\sum_{i=0}^{n} Y_i \sim \chi^2_{2n}$. Or χ^2_{2n} ne dépend pas du paramètre a, donc 2a$\sum_{i=0}^{n} Y_i$ est une fonction pivotale.

En notant $z_{n,\alpha}$ le (1-α)-quantile de la loi χ^2_{2n} on a:

$$\mathbb{P}(z_{2n,1-\frac{\alpha}{2}} \leq 2a \sum_{i=0}^{n} Y_i \leq z_{2n,\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$$

D'où:

$$\mathbb{P}(a \in [\frac{z_{2n,1-\frac{\alpha}{2}}}{2a\sum_{i=0}^{n} Y_i}, \frac{z_{2n,\frac{\alpha}{2}}}{2a\sum_{i=0}^{n} Y_i}]) = 1-\alpha$$

Donc:
$[\frac{z_{2n,1-\frac{\alpha}{2}}}{2a\sum_{i=0}^{n} Y_i}, \frac{z_{2n,\frac{\alpha}{2}}}{2a\sum_{i=0}^{n} Y_i}]$ est un intervalle de confiance de seuil α pour a.

5. On a $F_Y(x) = 1-\exp(-ax)$, d'où $\ln(1-F_Y(x)) = -ax$. Par conséquent, le graphe de probabilités est le nuage de points $(Y_i, \ln(1-\frac{i}{n})), i \in 1, ..., n-1$.

   Le tracé des graphes de probabilités sous `R` (figure 7) des échantillions de la cuve 1 et la cuve 2 permettent de considérer que les points sont approximativement allignés sur une droite de pente négative et passants par l'origine, donc on peut considérer qu'il est vraisemblable que les mesures des faissures des cuves 1 et 2 suivent une loi exponentielle. Le graphe de probabilité des échantillions de la cuve 3 semble plus proche d'un logarithme que d'une droite. On en conclue que la loi exponentielle n'est pas un modèle approprié pour ces données.

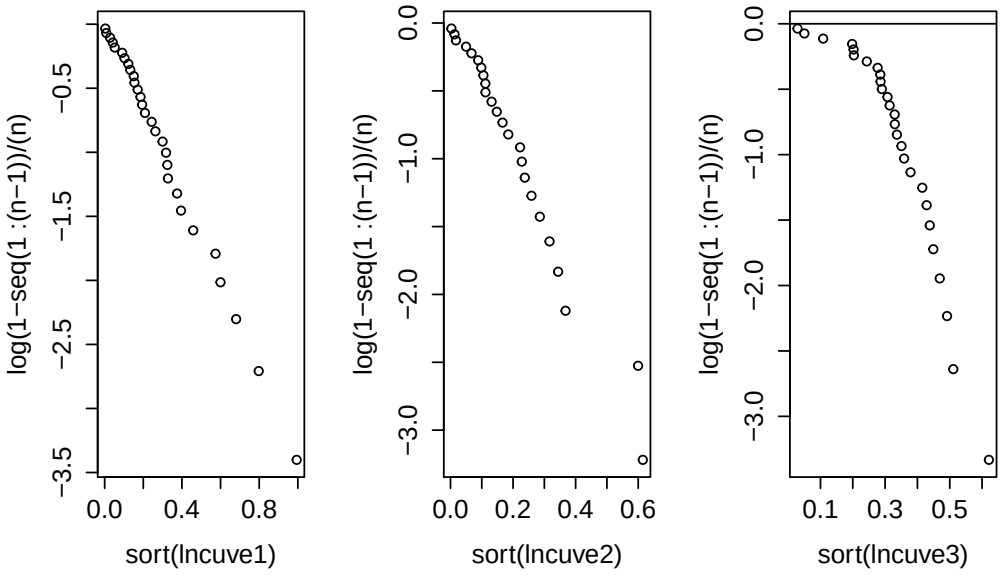

Figure 7: Graphes de probabilités

On sait que l'estimateur de maximum de vraisemblance et l'estimateur des moments de λ pour un échantillion $Y_1, Y_2, ..., Y_n$ qui suit une loi exponentielle de paramètre λ sont égaux à $\hat{\lambda_n} = \frac{1}{\overline{Y_n}}$. Ce qui donne les valeurs suivantes de $\hat{a_n}$ pour chacune des trois cuves:

$\hat{a_{cuve1}} = 3.170032$

$\hat{a_{cuve2}} = 4.331652$

$\hat{a_{cuve3}} = 2.998926$

6. D'après la question 1, les données des cuves 1 et 2 sont vraisemblablement de loi exponentielle et celles de la cuve 3 est vraisemblablement de loi normale. Pour vérifier ces résultats on trace les graphes de probabilités dans `R` et on obtient les figures 8 et 9, qui sont en adéquation avec les résultat de la première question.

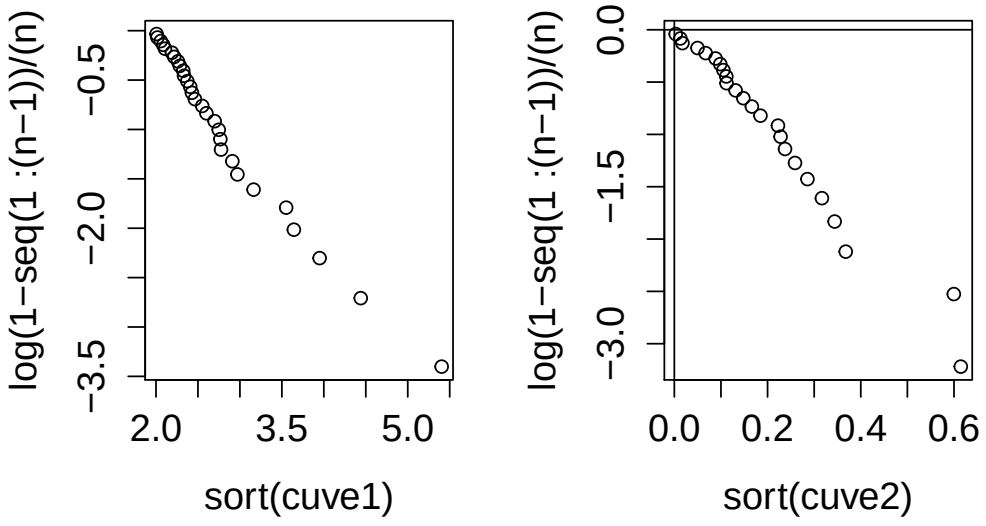

Figure 8: Graphes de probabilités des données de cuve 1 et cuve 2

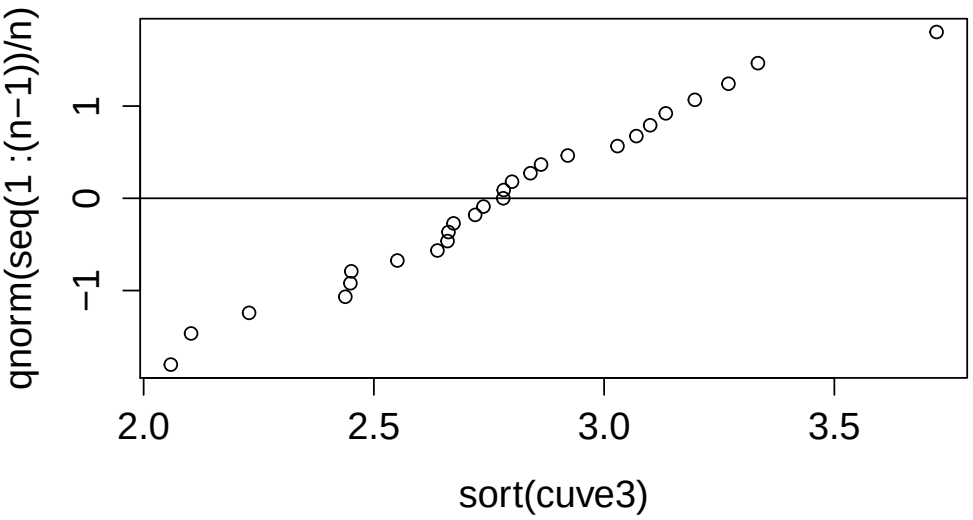

Figure 9: Graphes de probabilités des données de cuve 3

7. (a) Les défauts sont classés dangereux lorsque leur taille est supérieure à 5 mm. Le constructeur assure que ses cuves après 5 années d'utilisation ne présenteront pas une proportion de défauts dangereux supérieure à 5%. Ce qui équivaut à dire: $\mathbb{P}(X > 5) < 0.05$
Or, d'après la question 2,

$$\mathbb{P}(X \leq 5) = 1 - 2^a\, 5^{-a}$$

D'où,

$$\mathbb{P}(X > 5) = 1 - \mathbb{P}(X \leq 5) = 2^a\, 5^{-a}$$

Donc,

$$\mathbb{P}(X > 5) < 0.05 \Leftrightarrow 2^a\, 5^{-a} < 0.05 \Leftrightarrow a > \ln(0.05)/\ln(\frac{2}{5}) = 3.269412$$

D'après les estimations de a trouvées à la question 5, on peut dire que les affirmations du constructeur ne sont valides que pour les données de la cuve 2.

(b) Sur les 30 mesures faites pour la cuve 1, 2 seulement présentent des mesures supèrieures à 5 mm et seront signaler par l'appareil

B. Ce qui fait 6.7% $>$ 5% et contredit les affirmations du constructeur. Cette imprécision dans les estimations du constructeur peut être lié à plusieurs facteurs tels que: un mauvais choix de la loi de probabilité, condition extèrieures différentes par rapport aux expèriences du constructeur (température, pression, contenus des cuves, ...)

Avec l'appareil B on obtient des valeurs qui suit une loi binomiale. C'est une loi discrète et on perd donc de l'information sur l'état des défauts. Par exemple savoir les défaults proche de 5 mm peut donner à l'entreprise une idée sur la durée de vie restante de ces cuves et donc savoir qu'il faut penser à les changer bientôt ce qui garantit des meilleures prévisions budgétaires.

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 $Y = \ln \frac{X}{2}$. 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 $Pr = Compteur/m$.

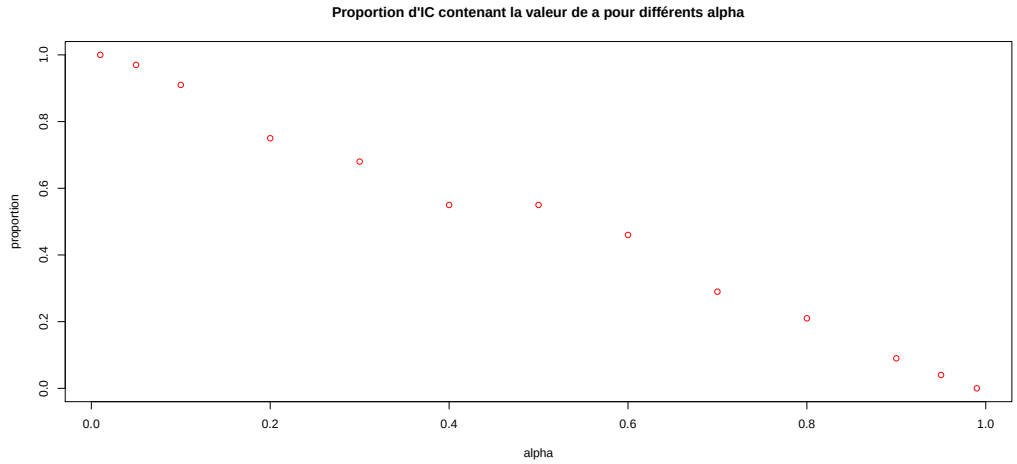

Figure 10: Graphe des proportions d'IC(α) contenant la valeur exacte de a pour différentes valeurs de α

Afin que cela soit plus représentatif,nous avons tracé un graphique avec les différentes proportions obtenues en fonction des α. (Voir figure 10)

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.

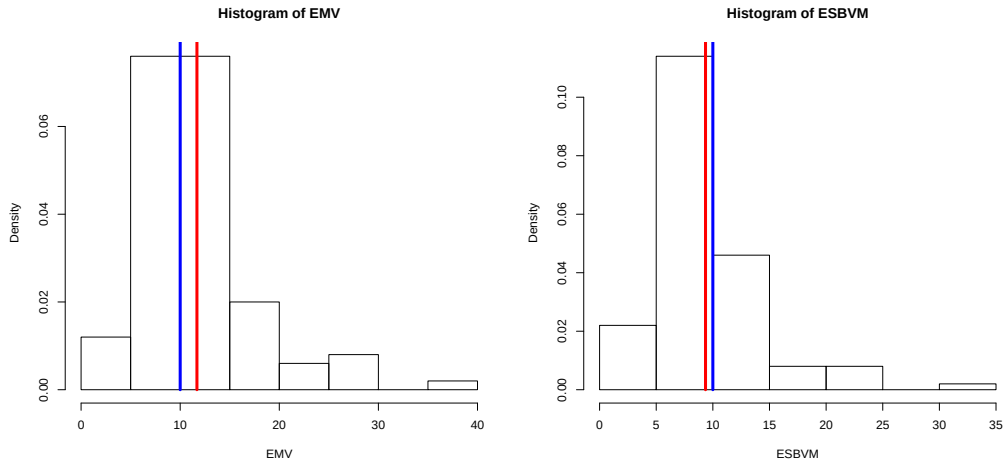

Figure 11: Histogramme des EMV et ESBVM pour $m = 100, n = 5, a = 10$

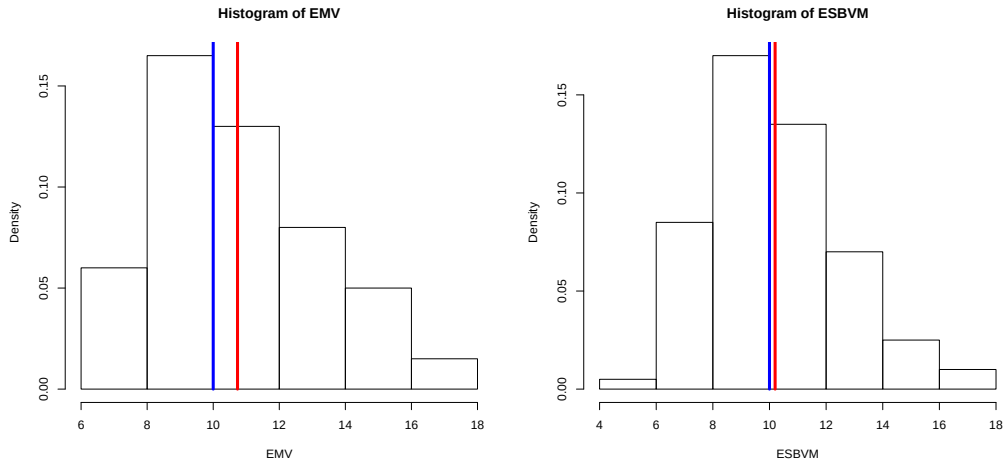

Figure 12: Histogramme des EMV et ESBVM pour $m = 100, n = 20, a = 10$

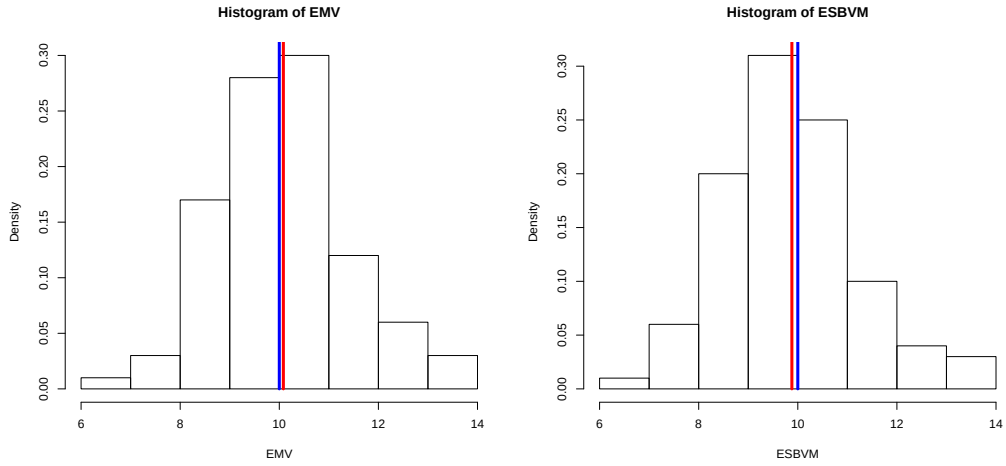

Figure 13: Histogramme des EMV et ESBVM pour $m = 100, n = 50, a = 10$

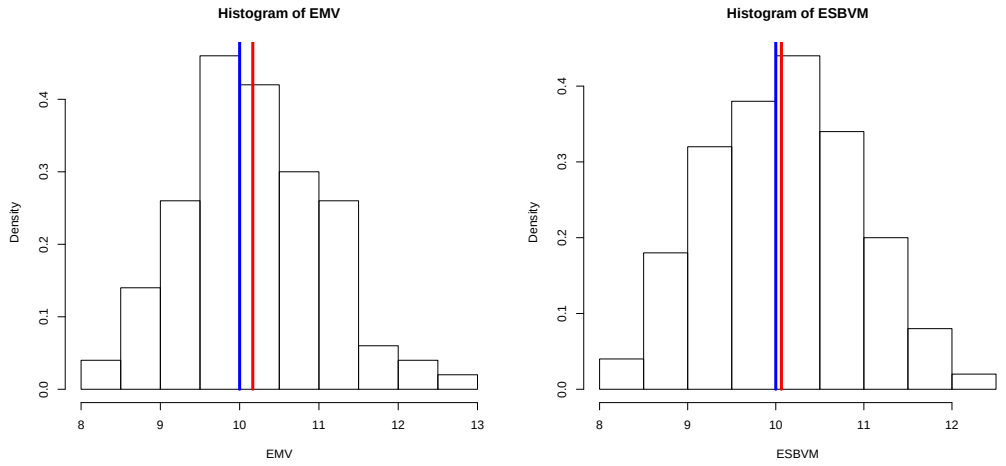

Figure 14: Histogramme des EMV et ESBVM pour $m = 100, n = 100, a = 10$

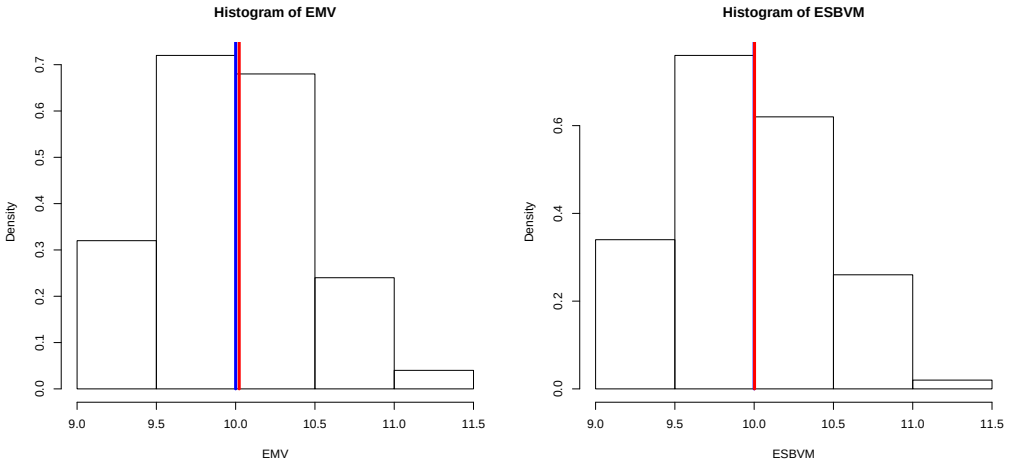

Figure 15: Histogramme des EMV et ESBVM pour $m = 500, n = 5, a = 10$

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 11 à 15) .

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.

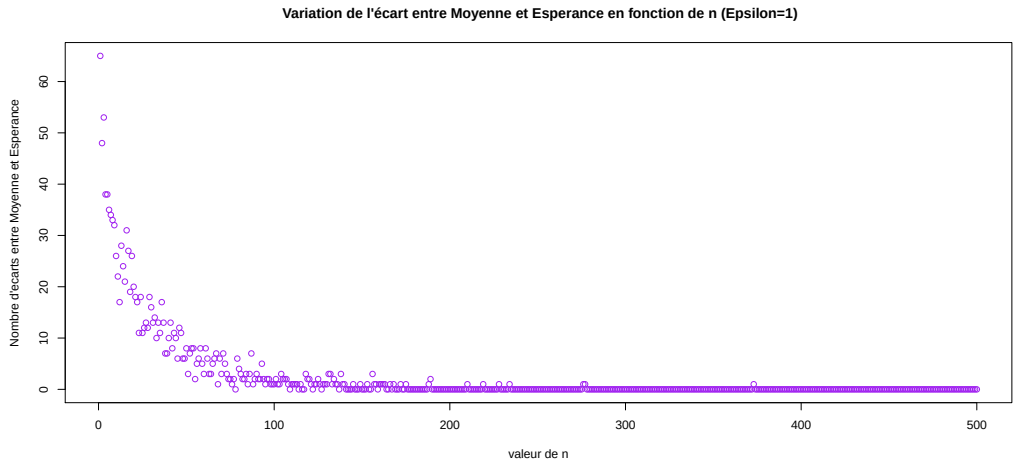

Figure 16: Variation de la différence absolue entre Moyenne est Esperance en fonction de n

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.

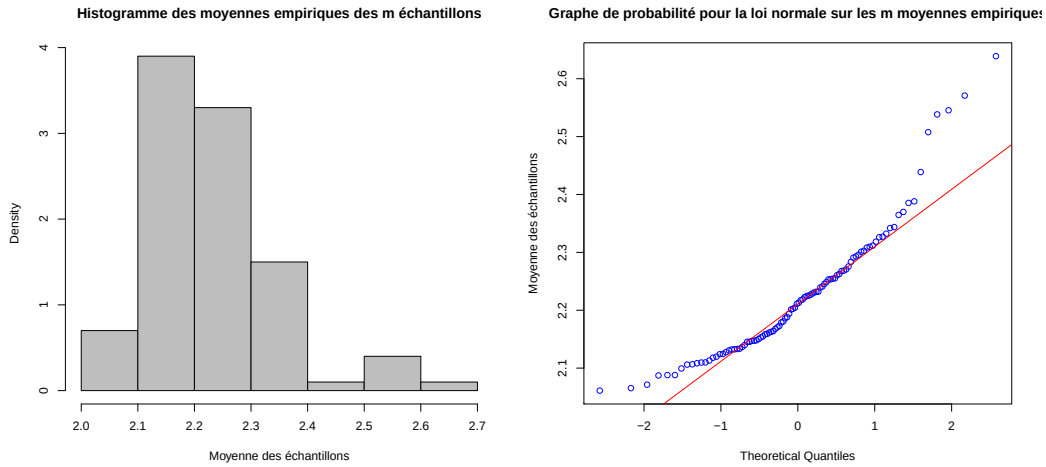

Figure 17: Etude de la distribution $\bar{X}_n$ par une distribution normale pour n=5

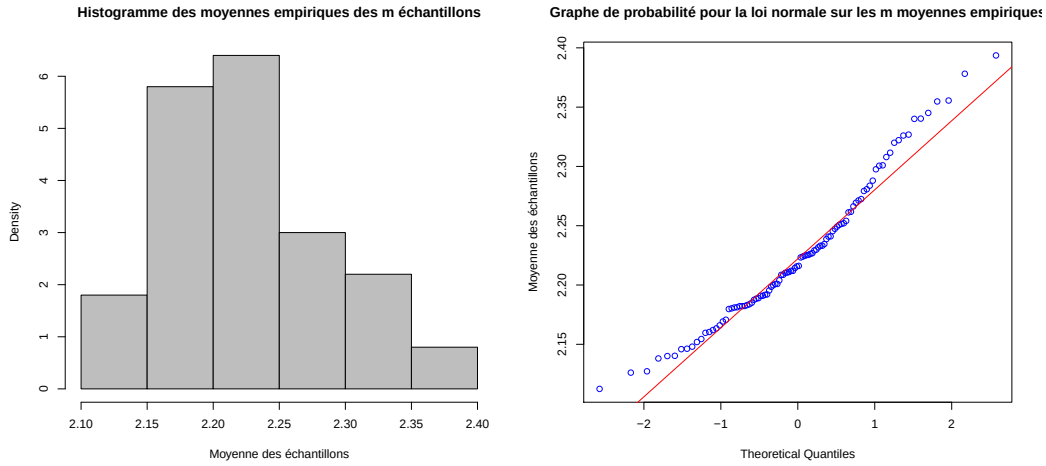

Figure 18: Etude de la distribution $\bar{X}_n$ par une distribution normale pour n=20

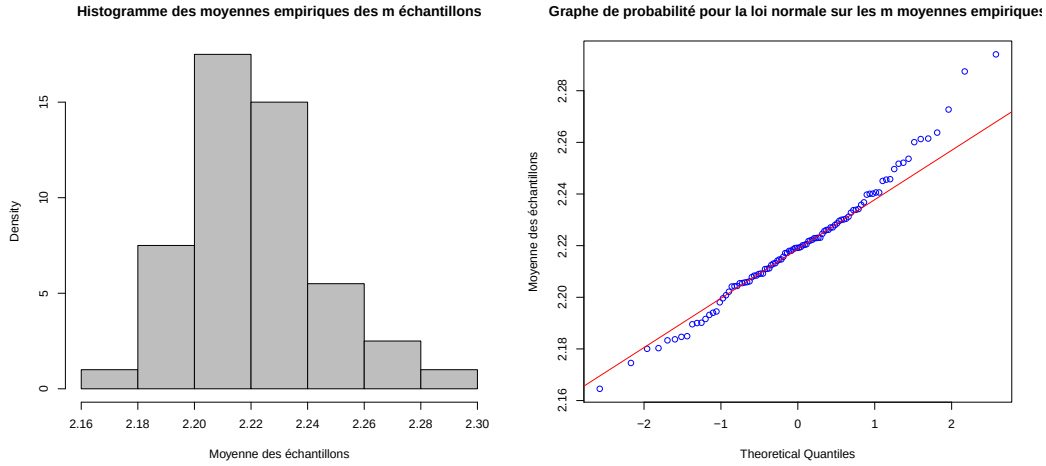

Figure 19: Etude de la distribution $barX_n$ par une distribution normale pour n=100

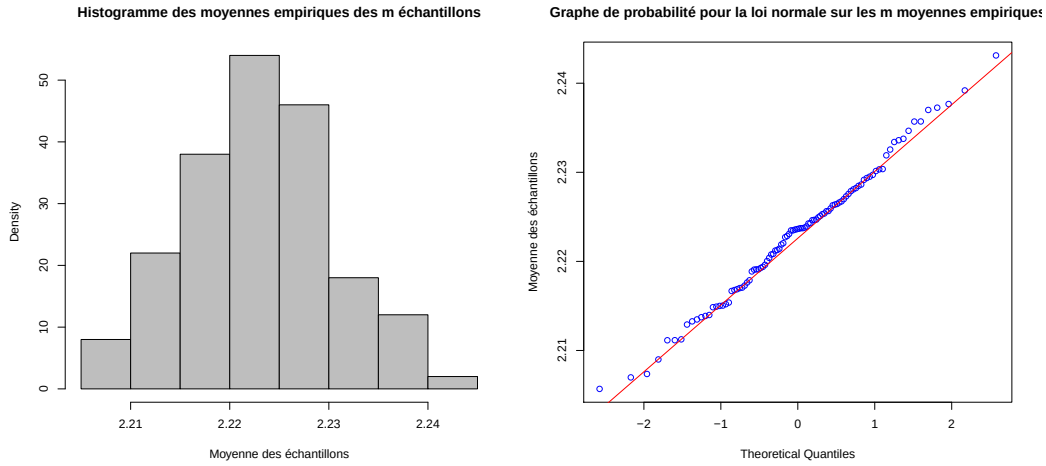

Figure 20: Etude de la distribution $\bar{X}_n$ par une distribution normale pour n=1000

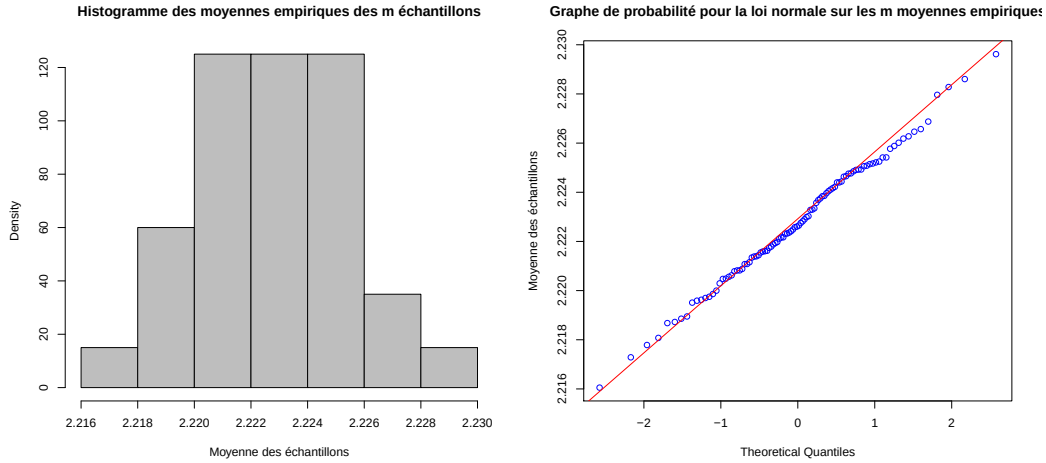

Figure 21: Etude de la distribution $\bar{X}_n$ par une distribution normale pour n=10000

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 $P(a,2)$

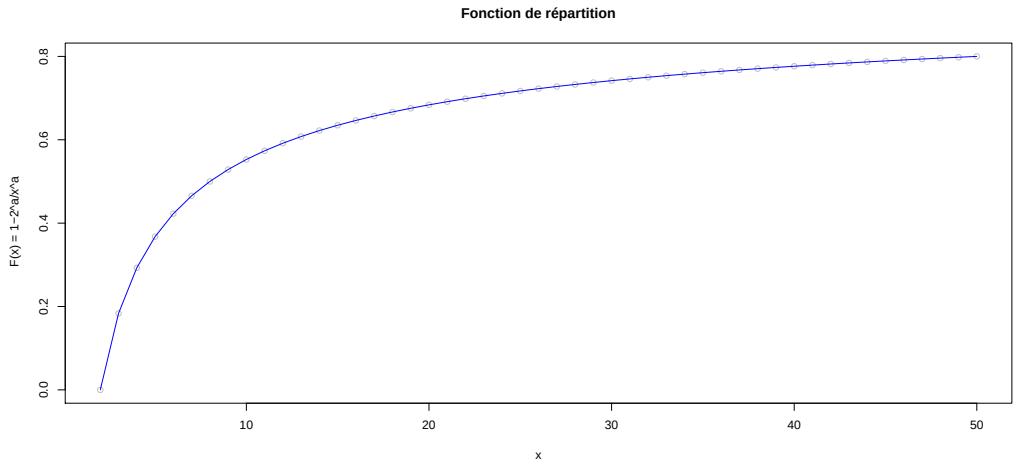

Figure 22: Fonction de répartition $F(X) = 1 - (2/t)^a$

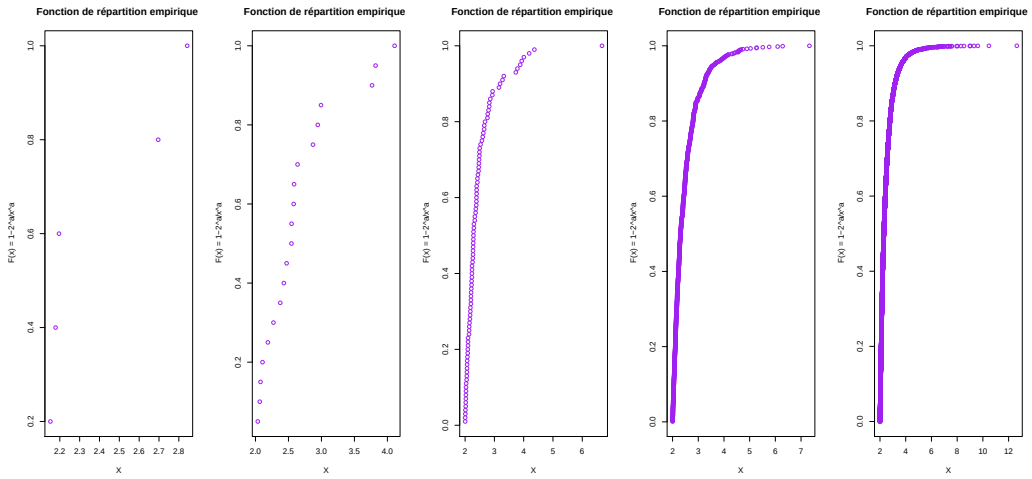

Figure 23: Fonctions de répartition avec $n = 5, 20, 100, 1000, 10000$

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

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