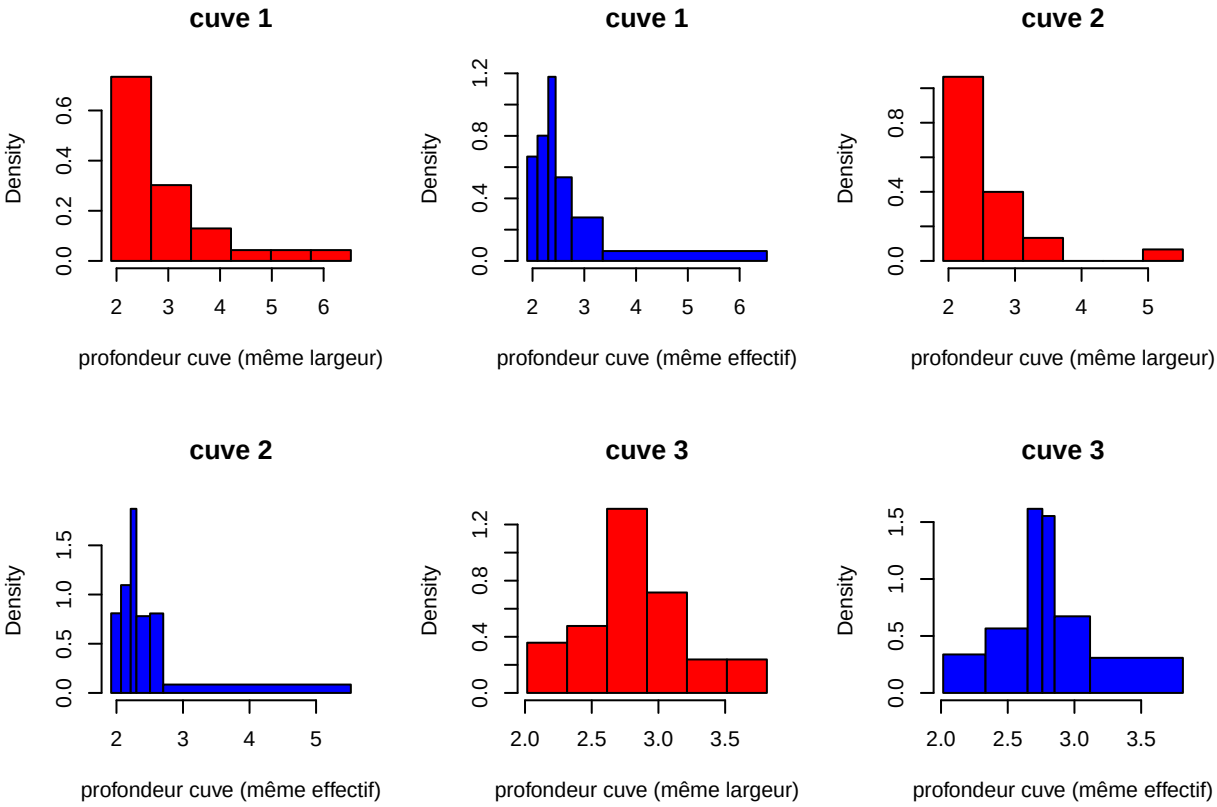
TP Principes et Méthodes Statistiques

Omar BENCHEKROUN
Mohamed SGHIOUAR IDRISSI
Andrew MCDONALD

1 - Analyse des défauts de cuves

Question 1



Les indicateurs de la cuve1:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.007   2.267   2.507   2.872   2.955   6.416
```

```
## variance de la cuve 1 est:  1.046943
```

```
## l'écart-type de la cuve 1 est:  1.023202
```

```
## le coefficient de variation empirique de la cuve1 est 0.3503371
```

Les indicateurs de la cuve2:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.006   2.207   2.362   2.592   2.661   5.437
```

```
## variance de la cuve 2 est:  0.5412769
```

```
## l'écart-type de la cuve 2 est:  0.7357152
```

```
## le coefficient de variation empirique de la cuve2 est 0.2781317
```

Les indicateurs de la cuve3:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.059   2.616   2.781   2.821   3.078   3.770
```

```
## variance de la cuve 3 est:  0.1733518
```

```
## l'écart-type de la cuve 3 est:  0.4163554
```

```
## le coefficient de variation empirique de la cuve3 est 0.1449392
```

On remarque d'après les histogrammes que les données des cuves 1 et 2 semblent suivre la même loi, cependant les données de la cuve 3 elles, suivent une loi différente des deux premières cuves.

Question 2

La fonction de répartition :

$$F_X(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} \int_2^x \frac{a2^a}{t^{1+a}}dt & \text{si } x > 2 \\ 0 & \text{sinon} \end{cases}$$

et :

$$\begin{aligned} \int_2^x \frac{a2^a}{t^{1+a}}dt &= a2^a \int_2^x t^{-1-a}dt \\ &= a2^a \left[\frac{t^{-a}}{-a}\right]_2^x \\ &= 1 - \left(\frac{2}{x}\right)^a \end{aligned}$$

d'où :

$$F_X(x) = \begin{cases} 1 - \left(\frac{2}{x}\right)^a & \text{si } x > 2 \\ 0 & \text{sinon} \end{cases}$$

L'espérance de X : Si $a > 1$ alors X admet une espérance finie et :

$$E[X] = \int_{\mathbb{R}} xf(x)dx \quad = \int_2^{+\infty} x\frac{a2^a}{x^{1+a}}dx = \frac{2a}{a-1}$$

La variance de X : Si $a > 2$ alors X admet une variance finie

$$\begin{aligned} V(X) &= E[X^2] - E[X]^2 \\ &= \int_{\mathbb{R}} x^2 f(x)dx - E[X]^2 \\ &= \frac{4a}{a-2} - \left(\frac{2a}{a-1}\right)^2 \end{aligned}$$

Question 3

Pour a et b dans $\bar{\mathbb{R}}$:

$$\begin{aligned} F_Y(x) &= P(ln(\frac{X}{2}) < x) \\ &= P(X < 2e^x) \\ &= F_X(2e^x) \\ &= \begin{cases} 1 - e^{-ax} & \text{si } x > 0 \\ 0 & \text{sinon} \end{cases} \end{aligned}$$

d'où Y est de densité :

$$f_Y(x) = F'_Y(x) = ae^{-ax}\mathbb{1}_{[0,+\infty[}$$

Question 4

On a $Y \sim \exp(a)$. Pour

$$Y_i,\ i \in [1, n]$$

On a

$$\sum_{i=1}^{n} Y_i \sim \Gamma(n, a)$$

Et donc $2a\sum_{i=1}^{n} Y_i \sim \Gamma(n, \frac{1}{2})$. Notre fonction pivotale est donc $2a\sum_{i=1}^{n} Y_i$. On pose

$$\begin{cases} P(2a\sum_{i=1}^{n} Y_i < a_1) = \frac{\alpha}{2} \\ P(2a\sum_{i=1}^{n} Y_i < a_2) = 1 - \frac{\alpha}{2} \end{cases}$$

D'où d'abord:

$$\begin{cases} a_1 = F^{-1}_{\chi^2_{2n}}(\frac{\alpha}{2}) \\ a_2 = F^{-1}_{\chi^2_{2n}}(1 - \frac{\alpha}{2}) \end{cases}$$

Et d'autre part:

$$\begin{aligned} & P(a_1 < 2a\sum_{i=1}^{n} Y_i < a_2) = 1 - \alpha \\ \Rightarrow\ & P(\frac{a_1}{2\sum_{i=1}^{n} Y_i} < a < \frac{a_2}{2\sum_{i=1}^{n} Y_i}) = 1 - \alpha \Rightarrow \quad P(\frac{F^{-1}_{\chi^2_{2n}}(\frac{\alpha}{2})}{2\sum_{i=1}^{n} Y_i} < a < \frac{F^{-1}_{\chi^2_{2n}}(1 - \frac{\alpha}{2})}{2\sum_{i=1}^{n} Y_i}) = 1 - \alpha \end{aligned}$$

On déduit l'intervalle de confiance suivant :

$$IC(\alpha) = \left[\frac{F^{-1}_{\chi^2_{2n}}(\frac{\alpha}{2})}{2n\bar{Y_n}}, \frac{F^{-1}_{\chi^2_{2n}}(1 - \frac{\alpha}{2})}{2n\bar{Y_n}}\right]$$

Question 5

On utilisera les trois méthodes vues en cours :

Graphe de probabilités :

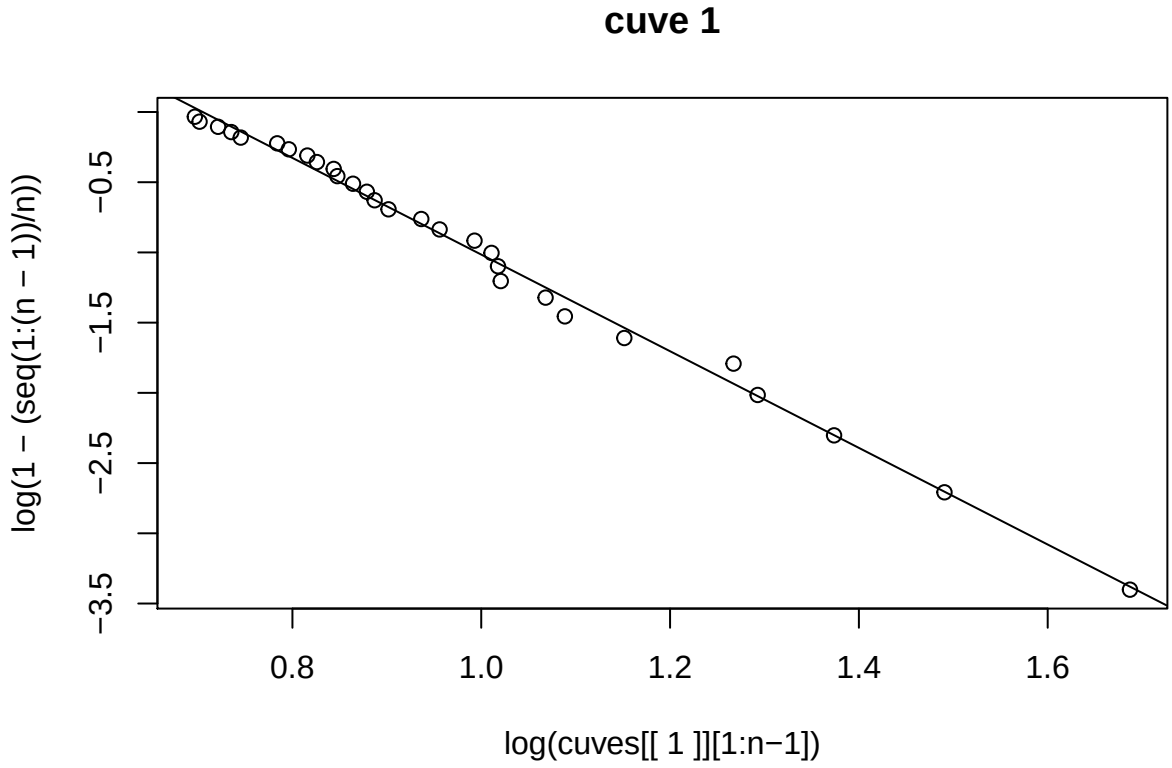
On a :

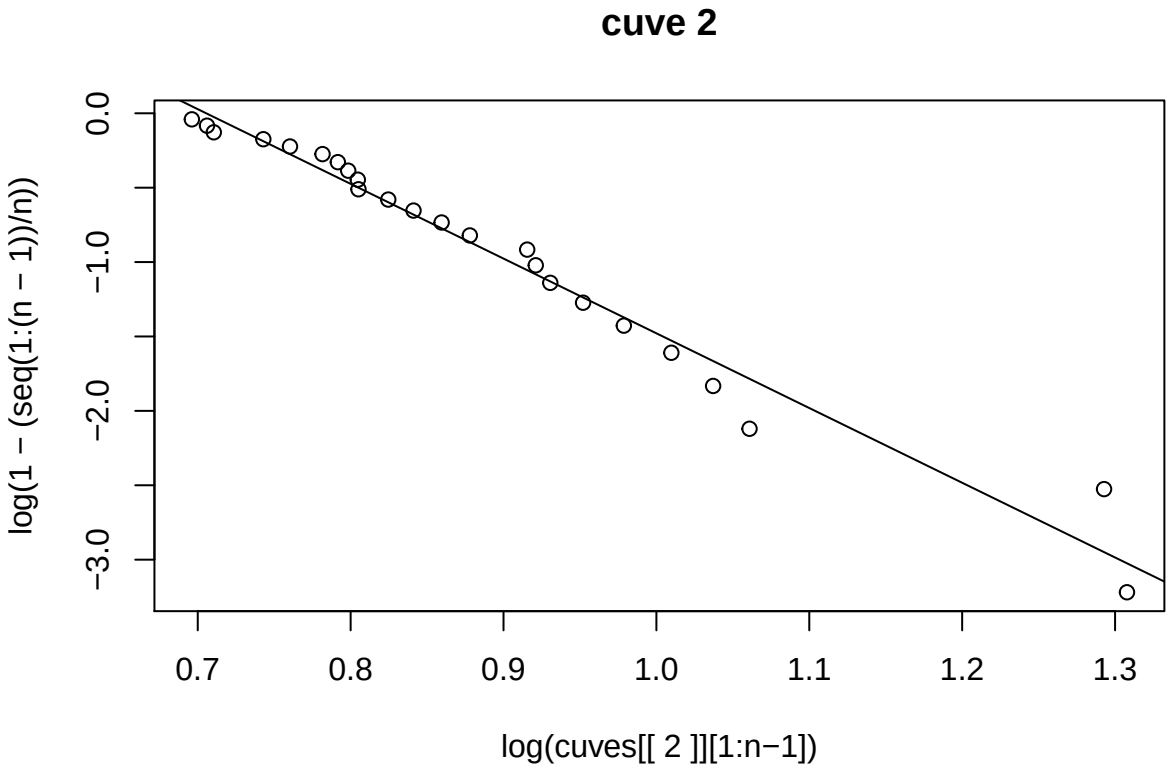
$$\log(1 - F_X(x)) = -a\log(x) + a\log(2)$$

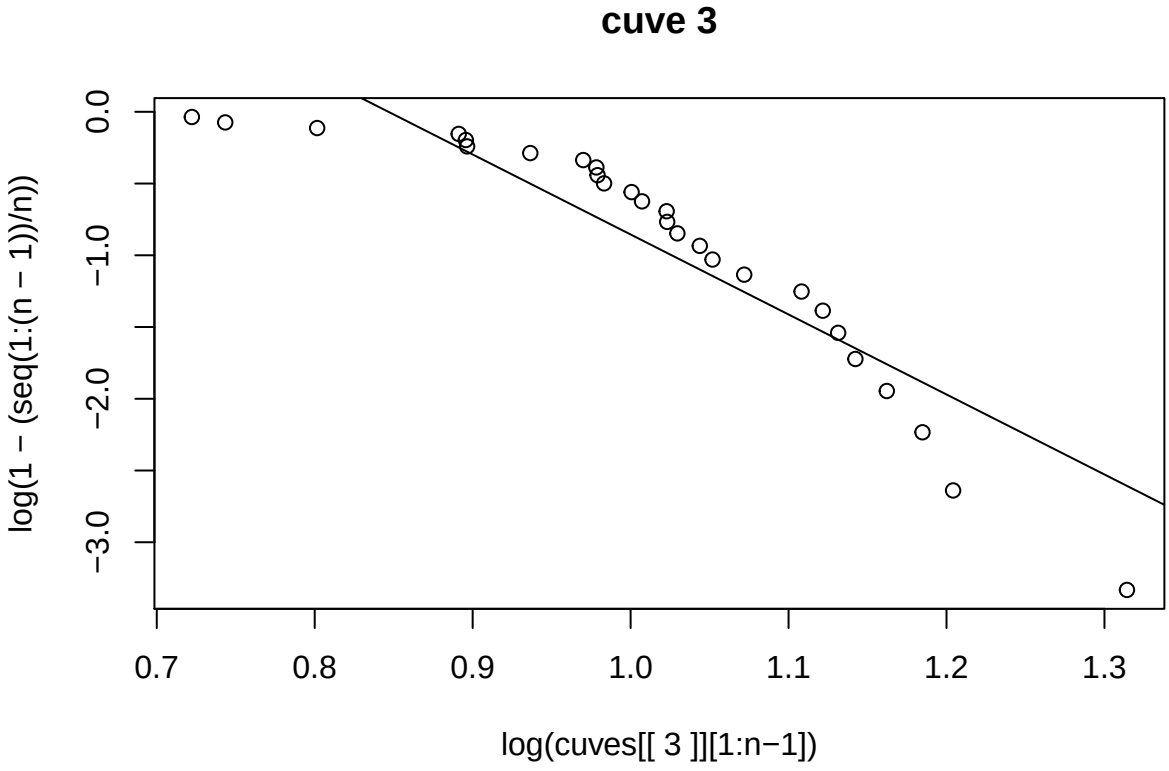
Donc notre graphe de probabilité est le nuage de points

$$(log(x_i^*), \log(1 - \frac{i}{n})), i \in \{1, \ldots, n\}$$

et dont la pente correspond à $-a$.







On remarque ici que le graphe de probabilités nous donne des points alignés pour les cuves 1 et 2, mais pas pour la cuve 3 ainsi on peut en déduire que la loi Pa(a,2) est pertinente pour les données des deux premières cuves mais pas de la troisième. Cela rejoint donc notre hypothèses sur les histogrammes des trois cuves. On ne calcule donc la valeur de a que pour les cuves 1 et 2.

```
## 3.441092
```

```
## 5.021564
```

Estimateur de moments EMM :

La méthode consiste à approcher l'espérance (théorique) de la loi $Pa(a, 2)$ par la moyenne empirique de l'échantillon:

$$E[X] = \frac{2a}{a-1} = \bar{X_n}$$

Ce qui nous donne notre estimateur EMM:

$$\hat{a}_n = \frac{\bar{X_n}}{\bar{X_n} - 2}$$

En calculant les estimations pour chacune des 3 cuves :

```
## Estimateurs des moments
```

```
## cuve 1  :  3.294806
## cuve 2  :  4.379749
```

Estimateur de maximum de vraisemblance EMV :

Pour une observation $(x_1, x_2, \ldots, x_n)$ quelconque, on a :

$$\begin{aligned}
\log \mathcal{L}(x_1, x_2, \ldots, x_n, a) &= \sum_{i=1}^{n} \log f(x_i) \\
&= \sum_{i=1}^{n} \log \frac{a 2^a}{x_i^{1+a}} \\
&= n \log(a) + na \log(2) - (1+a) \sum_{i=1}^{n} \log(x_i)
\end{aligned}$$

dont la dérivée s'annule quand :

$$a = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} \log(\frac{x_i}{2})}$$

Ce qui nous donne notre estimateur EMV:

$$\tilde{a}_n = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} \log(\frac{X_i}{2})} = \frac{1}{\bar{Y_n}}$$

```
## Estimateurs de maximum de vraisemblance
```

```
## cuve 1  :  3.170032
## cuve 2  :  4.331652
```

2 - Vérifications expérimentales à base de simulations

Question 1

Pour simuler un échantillon de taille n de la loi $Pa(a, b)$, on peut simuler d'abord un échantillon de loi $exp(a)$, et utiliser le fait que si $X \sim Pa(a, b)$ et si on pose :

$$Y_b = \ln(\frac{X}{b})$$

Alors par des calculs similaires à ce qu'on a fait précédemment, on a $Y \sim exp(a)$.
Synthèse : On simule un premier échantillon y suivant la loi $exp(a)$, et on calcule $x = 2\exp(y)$ qui est un échantillon qui suit la loi $Pa(a, b)$.

Question 2

Avec $\alpha = 0.05$, on effectue les simulations suivantes :

```
## Pour m= 100 n= 1000 a= 1  Proportion d'intervalles contenant le paramètre 0.98
## Pour m= 100 n= 1000 a= 5  Proportion d'intervalles contenant le paramètre 0.97
## Pour m= 100 n= 5000 a= 1  Proportion d'intervalles contenant le paramètre 0.97
## Pour m= 100 n= 5000 a= 5  Proportion d'intervalles contenant le paramètre 0.92
## Pour m= 100 n= 10000 a= 1  Proportion d'intervalles contenant le paramètre 0.98
## Pour m= 100 n= 10000 a= 5  Proportion d'intervalles contenant le paramètre 0.95
## Pour m= 500 n= 1000 a= 1  Proportion d'intervalles contenant le paramètre 0.938
## Pour m= 500 n= 1000 a= 5  Proportion d'intervalles contenant le paramètre 0.956
## Pour m= 500 n= 5000 a= 1  Proportion d'intervalles contenant le paramètre 0.948
## Pour m= 500 n= 5000 a= 5  Proportion d'intervalles contenant le paramètre 0.95
## Pour m= 500 n= 10000 a= 1  Proportion d'intervalles contenant le paramètre 0.958
## Pour m= 500 n= 10000 a= 5  Proportion d'intervalles contenant le paramètre 0.958
## Pour m= 1000 n= 1000 a= 1  Proportion d'intervalles contenant le paramètre 0.951
## Pour m= 1000 n= 1000 a= 5  Proportion d'intervalles contenant le paramètre 0.938
## Pour m= 1000 n= 5000 a= 1  Proportion d'intervalles contenant le paramètre 0.962
## Pour m= 1000 n= 5000 a= 5  Proportion d'intervalles contenant le paramètre 0.944
## Pour m= 1000 n= 10000 a= 1  Proportion d'intervalles contenant le paramètre 0.957
## Pour m= 1000 n= 10000 a= 5  Proportion d'intervalles contenant le paramètre 0.95
```

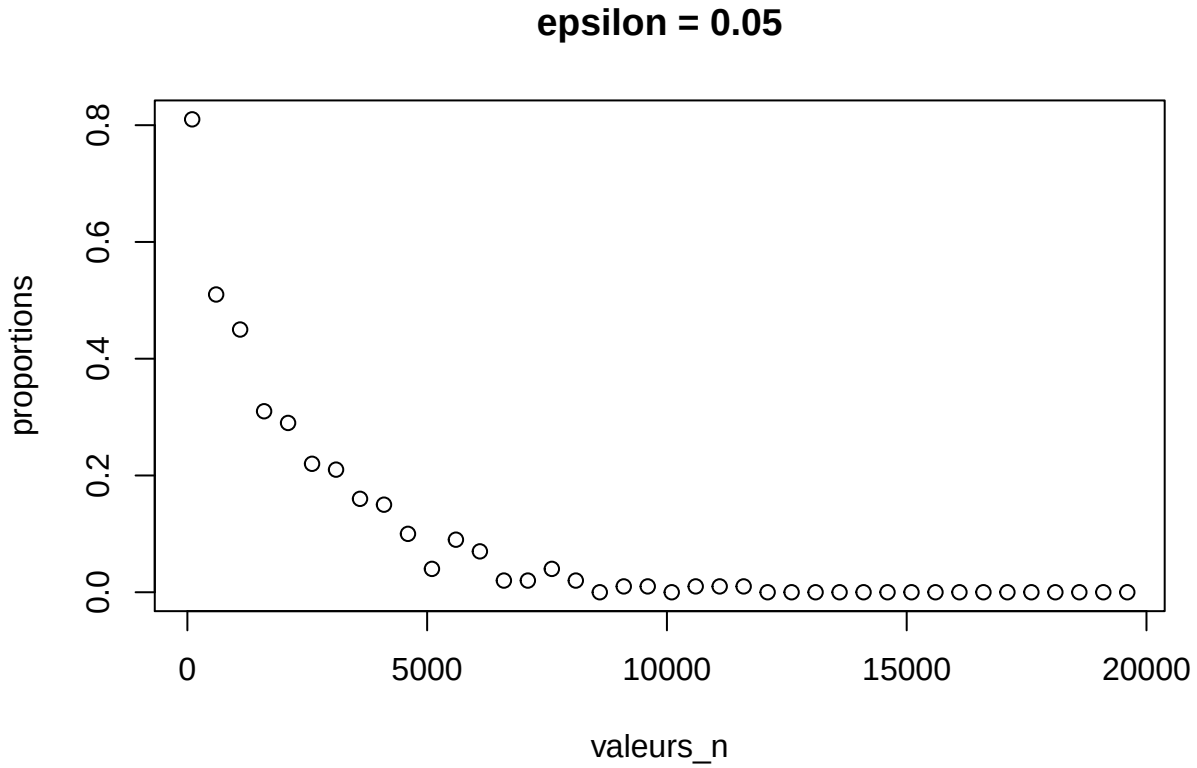
Question 3

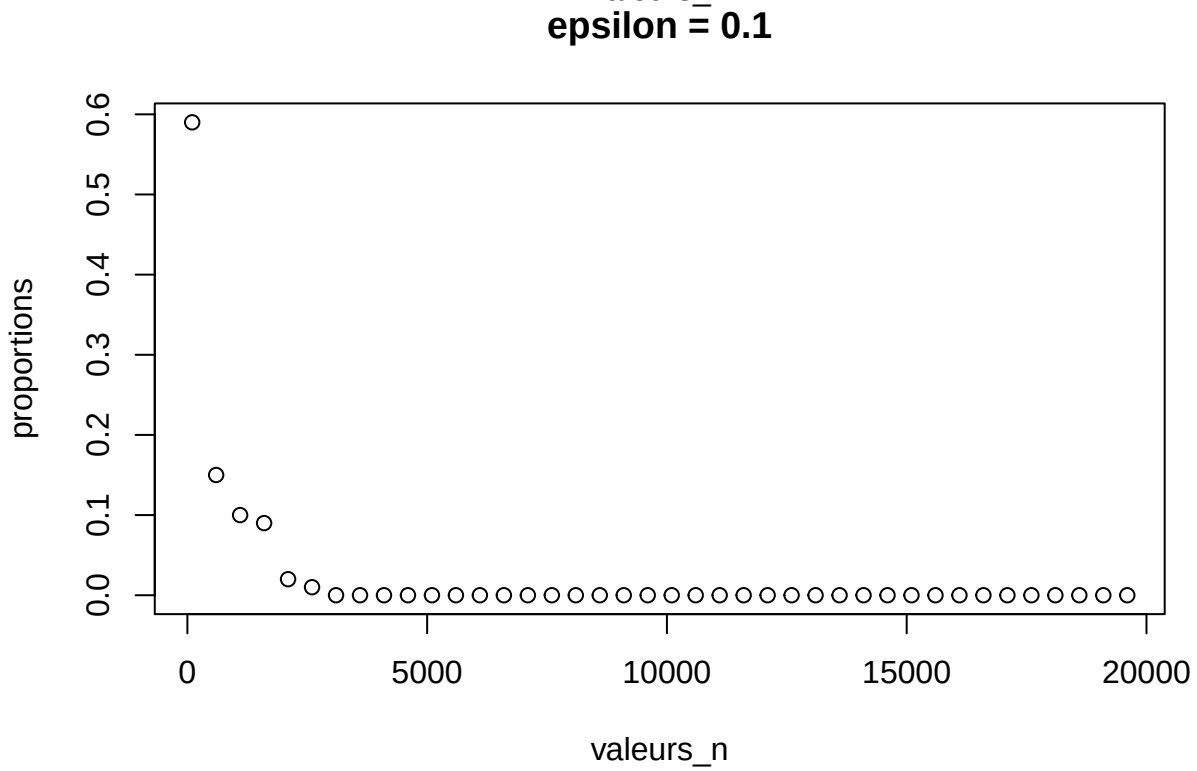
```
## biais emm 0.001345966
## biais emv -0.01716327
## erreur quadratique emm 4.51494e-05
## erreur quadratique emv 0.0006139183
```

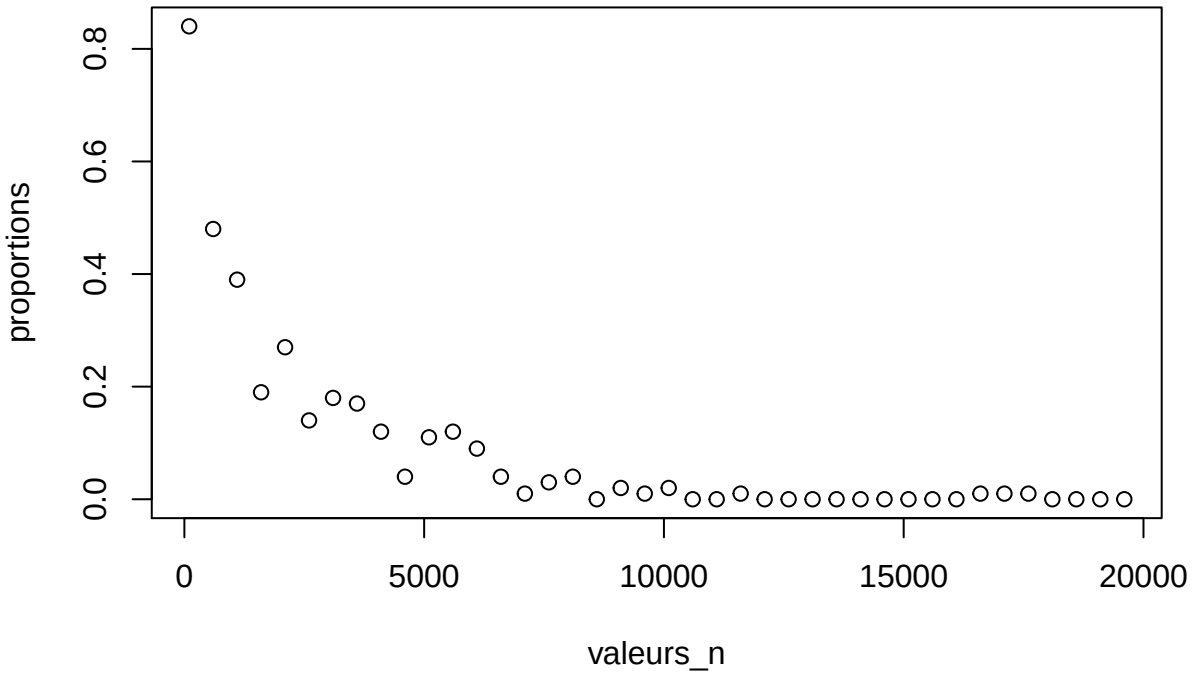
Conclusion : On remarque ici que le meilleur estimateur est l'estimateur de maximum de vraisemblance, car il a le biais et l'erreur quadratique les moins élevés.

Question 4

On choisit la méthode de l'estimateur de maximum de vraisemblance :

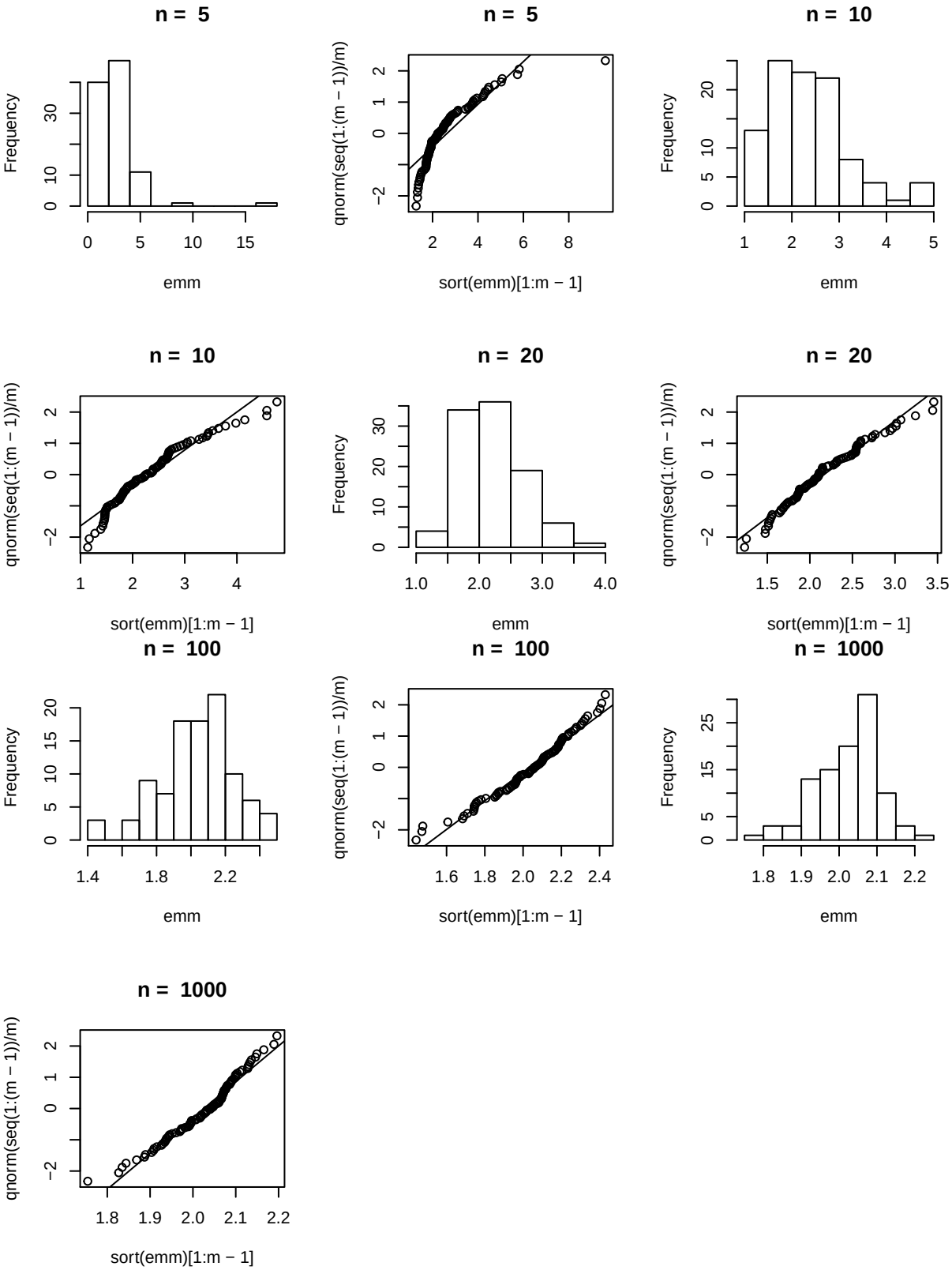

epsilon = 0.05
proportions
0.0
0.2
0.4
0.6
0.8
0
5000
10000
15000
20000
valeurs_n


epsilon = 0.1
proportions
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
5000
10000
15000
20000
valeurs_n



$\Rightarrow$ On conclut la convergence faible de l'estimateur de maximum de vraisemblance.

Question 5



⇒ Sur les différents graphes de probabilités,on remarque que plus n est grand, plus les points sont alignés.

Aussi les histogrammes se rapprochent de la cloche d'une loi normale centrée en 2, qui est la valeur de a choisie. On conclut que l'estimateur converge asymptotiquement vers la loi normale centrée en a lorsque n tend vers l'infini.

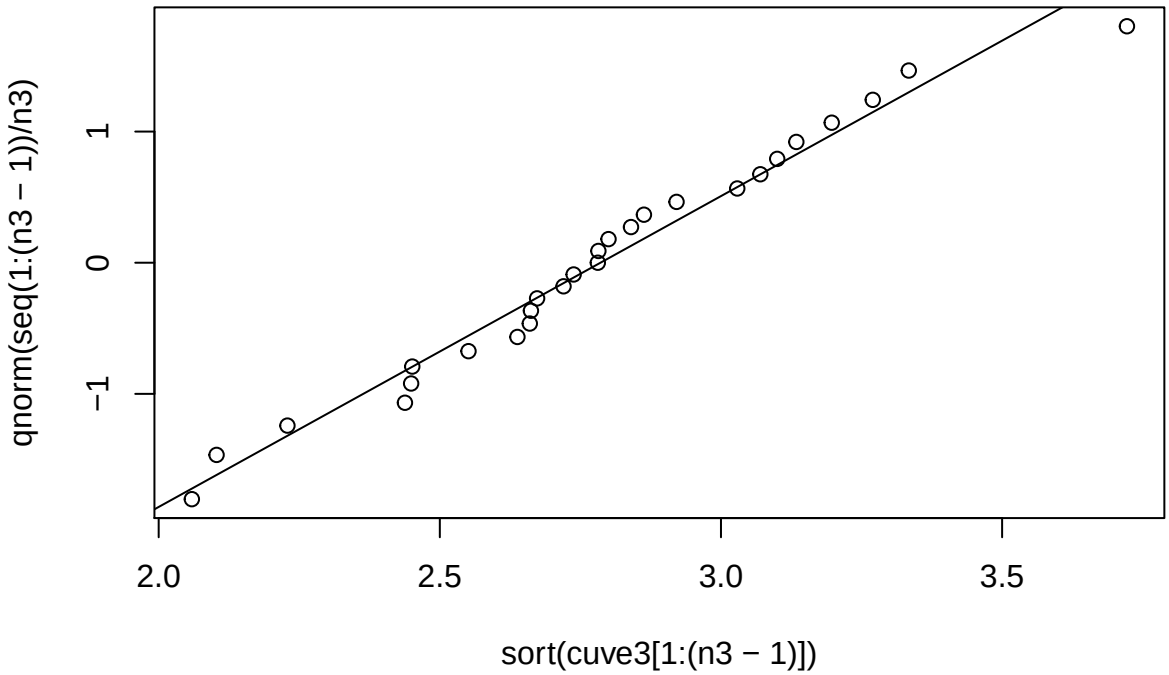
3 - Comparaison de modèles et certification des cuves

Question 1

Pour la cuve 1 : Grace au résultats obtenues dans la partie précédente, il semble que la loi $\mathcal{P}a(a, b)$ soit une bonne approximation. Pour les paramètres de $\mathcal{P}a(a, 2)$, on va prendre comme dans la partie précédente, b=2 .Quant au paramètre a, on peut utiliser l'estimateur de maximum de vraisemblance On modélisera donc les défauts de la cuve 1 par la loi $\mathcal{P}a(3.17, 2)$.

Pour la cuve 2 : Idem pour la cuve 2, $\mathcal{P}a(a, 2)$ semble pertinent. On estime à l'aide de l'estimateur du maximumu de vraisemblance le paramètre a. On modélisera donc les défauts de la cuve 2 par la loi $\mathcal{P}a(4.33, 2)$.

Pour la cuve 3 : La loi $\mathcal{P}a(a, b)$ n'est en revanche pas applicable pour modéliser les défauts de la cuve 3 (exercice 1, question 5). Au vu de l'allure de l'histogramme des données de la cuve 3, on peut penser à modéliser ces données par une loi normale. On vérifie la pertinence de ce choix à l'aide d'un graphe de probabiltés pour la loi normale. Le nuage de points est donc : $(x_i*,\ \phi^{-1}(i/n))$.



On voit que l'on obtient une droite. La loi normale semble donc adaptée pour modéliser les défauts des données de cuve3. Utilisons l'estimateur des moments pour déterminer les paramètres de cette loi normale.

```
## [1] 2.820857
```

```
## [1] 0.1733518
```

Par conséquent, on modélisera les défauts de la cuve 3 par une loi normale de paramètres : $\mathcal{N}(2.82,\ 0.17)$

Question 2 - a

Il s'agit d'un test d'hypothèse. Si on note $X = X_i$ la variable aléatoire associé à la taille du défaut dans la cuve. On prend $H_0 : E[X] \geq 5$ et $H_1 : E[X] < 5$, car l'erreur de première espèce est de décider que les défauts ne sont pas dangereux alors qu'elles le sont, ce qui est plus grave que de décider que les défauts sont dangereux alors qu'ils ne sont pas.

On se rapporte encore une fois à la variable aléatoire $Y = ln(\frac{X}{2})$ avec $Y \sim exp(a)$. Puisque $E[X] = \frac{2a}{a-1}$, les hypothèses deviennent : $H_0 : a \leq \frac{5}{3}$ et $H_1 : a > \frac{5}{3}$. On détermine donc la région critique d'une manière presque identique à celle de l'exo 3 Fiche 5, en utilisant la fonction pivotale $2a\sum_{i=1}^{n} Y_i \sim \chi_{2n}$ On pose

$$
\begin{aligned}
\alpha &= sup_{a\leq\frac{5}{3}} P(\hat{a}_n > l_\alpha, a) \\
&= sup_{a\leq\frac{5}{3}} P(\frac{n}{\sum_{i=1}^{n} Y_i} > l_\alpha, a) \\
W = \{\hat{a}_n > l_\alpha\} \quad &= sup_{a\leq\frac{5}{3}} P(2a\sum_{i=1}^{n} Y_i < \frac{2an}{l_\alpha}) \\
&= sup_{a\leq\frac{5}{3}} F_{\chi_{2n}}(\frac{2an}{l_\alpha}) \\
&= F_{\chi_{2n}}(\frac{10}{3}.\frac{n}{l_\alpha})
\end{aligned}
$$

Ce qui donne $l_\alpha = \frac{10}{3}\frac{n}{F^{-1}_{\chi 2n}(\alpha)}$, et enfin $W = \{\hat{a}_n > \frac{10}{3}\frac{n}{F^{-1}_{\chi 2n}(\alpha)}\}$.

Vérifions l'affirmation du constructeur pour la cuve 1 et 2 qui suivent la loi Pa(a,2):

```
## la borne de la région critique pour la cuve 1 est : 2.31546
## la borne de la région critique pour la cuve 2 est : 2.397098
```

Et on rappelle les estimations (méthode EMV) pour les cuve 1 et 2:

```
## cuve 1  :  3.170032
## cuve 2  :  4.331652
```

On déduit que l'affirmation du constructeur est vraie.

Question 2 - b

On étudie ici un test d'hypothèse sur une proportion.

On a $p_0 = 0.05$, et on prend $H_0 : p \geq p_0$ et $H_1 : p < p_0$ pour les mêmes raisons de la question 2-a. On a donc la région critique :

$$W = \left\{ \frac{t-np_0}{\sqrt{np_0(1-p_0)}} < -u_{2\alpha} \right\}$$

On essaie de voir si les observations sont dans la région critique pour un $\alpha = 5\%$

```
## x1 :  -0.2294157 x2 :  -1.644854
```

La réponse est bien évidemment non.

Calculons désormais la p-valeur:

```
##
##  Exact binomial test
##
## data:  sum(cuve_b) and length(cuve_b)
## number of successes = 2, number of trials = 30, p-value = 0.8122
```

```
## alternative hypothesis: true probability of success is less than 0.05
## 95 percent confidence interval:
##  0.000000 0.195326
## sample estimates:
## probability of success
##             0.06666667
```

```
##
##  Exact binomial test
##
## data:  sum(cuve_b) and length(cuve_b)
## number of successes = 1, number of trials = 25, p-value = 0.6424
## alternative hypothesis: true probability of success is less than 0.05
## 95 percent confidence interval:
##  0.0000000 0.1761207
## sample estimates:
## probability of success
##                   0.04
```

Ainsi les p-valeurs des cuves 1 et 2 sont respectivement 0.81 et 0.64, ce qui est trop élevé.

On ne peut pas conclure si l'affirmation du constructeur est vraie en utilisant seulement l'appareil B.

4- Conclusion

Finalement ce TP apporte une vision complémentaire à celles vues en cours et en TD pour ce qui est des méthodes statistiques de base. En ne concentrant l'étude que sur un problème en particulier il nous est possible de déployer un éventail fourni de méthodes vues ou non dans le cours ou en TD. Par exemple la simulation d'échantillons comme elle est présentée dans la partie 2 (questions 2 et 3) nous a intéressé. Aussi nous avons apporté une grande importance au côté concret des statistiques c'est pourquoi nous avons le plus possible mis en avant nos analyses graphiques et nos graphes en tous genre.

Pour ce qui est de l'organisation dans le groupe une bonne harmonie s'est dégagé de notre travail. Nous avons traité les parties théoriques chacun de notre côté et nous nous sommes réunis pour les parties en R afin d'enrichir au maximum nos compréhensions respectives du problème.

Pour finir l'analyse et le travail qui a été effectué est évidemment non-exhaustif. Nous pensons que pour réellement obtenir plus d'information il nous faudrait plus de données pour éventuellement ajuster plus précisément le modèle. On pourrait aussi penser à traiter ce problème comme un problème de classification et entraîner un réseau de neurones sur les données.