

TP Statistiques : Compte-rendu

Michel Yoeung, Charles-Frédéric Amaudruz, Alexandre Berrada (ENSIMAG - 1A - Groupe 2)

Avril 2018

1 Première stratégie.

Question 1

$$\forall i = 1, \dots, n, X_i = \begin{cases} 1 & \text{si le } i^{\text{eme}} \text{ poisson est bagué} \\ 0 & \text{sinon} \end{cases}$$

De plus, les résultats des pêches successives sont indépendants et la probabilité que le i^{eme} poisson soit bagué est $p = \frac{n_0}{\theta}$.

Donc les X_i sont indépendants et identiques (iid) car suivent tous une loi de Bernoulli de paramètre de succès $p = \frac{n_0}{\theta}$.

Soit $\theta = 1000$ et $n_0 = 50$. On choisit $n = 60$ pour notre simulation sur R.

On obtient l'échantillon de données suivant :

```
[1] "échantillon :"  
[1] 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
[60] 0
```

On calcule ensuite sur R les moyennes et les variances (empiriques et théoriques) :

```
> source('~/.ensimag1/Statistiques/projet/premiere_strategie.R')  
[1] "moyenne empirique :"  
[1] 0.06666667  
[1] "moyenne théorique :"  
[1] 0.05  
[1] "variance empirique :"  
[1] 0.06222222  
[1] "variance empirique :"  
[1] 0.0475
```

On constate que pour ce jeu de données la moyenne empirique est assez proche de la moyenne théorique et de même pour les variances.

Question 2

T est une variable aléatoire qui compte le nombre de poissons bagués parmi les n poissons pêchés donc $T = \sum_{i=1}^n X_i$ avec X_i qui suit une loi de Bernoulli.

Donc T suit une loi binomiale de paramètres n et $p = \frac{n_0}{\theta}$.

On donne ainsi avec R t le nombre de poissons pêchés sur notre échantillon de données :

```
[1] "nombre de poissons bagués parmi les n poissons pêchés :"
```

```
[1] 4
```

Question 3

Estimateur des moments :

Soit \bar{X}_n la moyenne empirique,

$$\begin{aligned} E[X] &\simeq \bar{X}_n \Rightarrow p \simeq \frac{1}{n} \sum_{i=1}^n x_i \\ \Rightarrow \frac{n_0}{\theta} &\simeq \frac{1}{n} \sum_{i=1}^n x_i \\ \Rightarrow \theta &\simeq \frac{n_0 * n}{\sum_{i=1}^n x_i} \end{aligned} \quad (1)$$

Donc l'estimateur des moments (d'ordre 1) vaut $\tilde{\theta}_n = \frac{n_0 * n}{\sum_{i=1}^n x_i}$.

Estimateur de maximum de vraisemblance :

$$\begin{aligned} \mathcal{L}(X_1 = x_1, \dots, X_n = x_n, \theta) &= \prod_{i=1}^n P(X_i = x_i, \theta) \\ &= p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i} \\ &= \left(\frac{n_0}{\theta}\right)^{\sum_{i=1}^n x_i} \left(1 - \frac{n_0}{\theta}\right)^{n-\sum_{i=1}^n x_i} \end{aligned} \quad (2)$$

$$\ln(\mathcal{L}(X_1 = x_1, \dots, X_n = x_n, \theta)) = \left(\sum_{i=1}^n x_i\right)(\ln(n_0) - \ln(\theta)) + \left(n - \sum_{i=1}^n x_i\right) \ln\left(1 - \frac{n_0}{\theta}\right)$$

En dérivant par rapport à θ :

$$\begin{aligned} \frac{\partial \ln(\mathcal{L}(X_1 = x_1, \dots, X_n = x_n, \theta))}{\partial \theta} &= -\left(\sum_{i=1}^n x_i\right) \frac{1}{\theta} + \left(n - \sum_{i=1}^n x_i\right) \frac{\frac{n_0}{\theta^2}}{1 - \frac{n_0}{\theta}} \\ &= -\left(\sum_{i=1}^n x_i\right) \frac{1}{\theta} + \left(n - \sum_{i=1}^n x_i\right) \frac{n_0}{\theta^2 - \theta n_0} \\ &= 0 \Rightarrow \theta = \frac{n_0 * n}{\sum_{i=1}^n x_i} \end{aligned} \quad (3)$$

Donc l'estimateur de maximum de vraisemblance vaut $\hat{\theta}_n = \frac{n_0 * n}{\sum_{i=1}^n x_i}$.

Donc ces deux estimateurs sont confondus.

On calcule ainsi sur R la valeur de ces estimateurs sur notre échantillon de données :

```
##
[1] "estimateur des moments :"
```

```
[1] 750
```

```
[1] "estimateur de maximum de vraisemblance :"
```

```
[1] 750
```

Sur l'échantillon simulée, on observe un estimateur de θ qui vaut 750 soit 25% de moins que la

vraie valeur de θ pour la simulation (1000).
Cet estimateur n'est pas très précis.

Question 4

intervalle de confiance exact de seuil α pour θ :

On a $p = \frac{n_0}{\theta} \Rightarrow \theta = \frac{n_0}{p}$ donc en estimant p (avec la formule du cours), on obtient directement une estimation de θ .

$$\left[n_0 \left(1 + \frac{n-T}{T+1} f_{2(n-T), 2(T+1), 1-\frac{\alpha}{2}} \right), n_0 \left(1 + \frac{n-T+1}{T} f_{2(n-T+1), 2T, \frac{\alpha}{2}} \right) \right] \quad (4)$$

avec $T = n\bar{X}_n = \sum_{i=1}^n x_i$ et $f_{\nu_1, \nu_2, \alpha} = F_{\mathcal{F}}^{-1}(1 - \alpha, \nu_1, \nu_2)$, $F_{\mathcal{F}}^{-1}$ étant ici la fonction quantile de la loi de Fisher-Snedecor.

intervalle de confiance asymptotique de seuil α pour θ :

On obtient l'intervalle associée avec la même démarche que précédemment.

$$\left[\frac{n_0}{\bar{X}_n + u_\alpha \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}}, \frac{n_0}{\bar{X}_n - u_\alpha \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}} \right] \quad (5)$$

avec $u_\alpha = F_{\mathcal{N}}^{-1}(1 - \frac{\alpha}{2})$, $F_{\mathcal{N}}^{-1}$ étant ici la fonction quantile de la loi normale.

On calcule ainsi ces intervalles de confiance (exacts puis asymptotiques) associés aux valeurs de nos paramètres fixés sur R pour les différentes valeurs de α spécifiées dans l'énoncé :

```
[1] "intervalle de confiance exact pour theta de seuil"
[1] 0.01
[1] 255.9594
[1] 4375.487
[1] "intervalle de confiance exact pour theta de seuil"
[1] 0.05
[1] 308.6672
[1] 2708.298
[1] "intervalle de confiance exact pour theta de seuil"
[1] 0.1
[1] 342.2379
[1] 2165.389
[1] "intervalle de confiance exact pour theta de seuil"
[1] 0.2
[1] 388.1135
[1] 1701.15
```

On remarque que les intervalles exacts fournissent un encadrement plus précis de la vraie valeur de θ que les intervalles asymptotiques ce qui semble logique car les intervalles asymptotiques sont plus efficaces pour un n très grand, or ici on a fixé $n = 60$ seulement.

Question 5

$$P(\hat{\theta}_n = +\infty) = P\left(\frac{n_0 * n}{\sum_{i=1}^n x_i} = +\infty\right) = P\left(\sum_{i=1}^n x_i = 0\right) = P(T = 0) = (1 - p)^n = \left(1 - \frac{n_0}{\theta}\right)^n$$

Cet estimateur n'est pas convergent.

$$\text{Biais}(\hat{\theta}_n) = E[\hat{\theta}_n] - \theta$$

```

[1] "intervalle de confiance asymptotique pour theta de seuil"
[1] 0.01
[1] 334.1883
[1] -3070.703
[1] "intervalle de confiance asymptotique pour theta de seuil"
[1] 0.05
[1] 385.257
[1] 14085.18
[1] "intervalle de confiance asymptotique pour theta de seuil"
[1] 0.1
[1] 417.9345
[1] 3650.342
[1] "intervalle de confiance asymptotique pour theta de seuil"
[1] 0.2
[1] 463.2351
[1] 1968.753

```

Comme $\hat{\theta}_n$ a une probabilité non nulle de valoir $+\infty$ alors $E[\hat{\theta}_n] = +\infty \Rightarrow E[\hat{\theta}_n] - \theta = +\infty$ (car θ est une constante) $\Rightarrow \text{Biais}(\hat{\theta}_n) = +\infty$.

Donc on peut en déduire que le biais de cet estimateur vaut $+\infty$. Sur notre échantillon de données, on peut calculer la probabilité $P(\hat{\theta}_n = +\infty)$:

```

[1] "probabilité que l'estimateur vale +infini :"
[1] 0.0460698

```

Question 6

$$\begin{aligned}
 P(\hat{\theta}_n = +\infty) &> \frac{1}{2} \Rightarrow \left(1 - \frac{n_0}{\theta}\right)^n > \frac{1}{2} \\
 &\Rightarrow n \ln\left(1 - \frac{n_0}{\theta}\right) > \ln\left(\frac{1}{2}\right) \\
 &\Rightarrow n < -\frac{\ln(2)}{\ln\left(1 - \frac{n_0}{\theta}\right)} \\
 &\Rightarrow n \leq \left\lfloor -\frac{\ln(2)}{\ln\left(1 - \frac{n_0}{\theta}\right)} \right\rfloor
 \end{aligned} \tag{6}$$

Toujours avec notre échantillon de données, on calcule sur R cette valeur de n pour laquelle la probabilité $P(\hat{\theta}_n = +\infty)$:

```

-----
[1] "pour que la probabilité que l'estimateur vale +infini soit strictement supérieure à 1/2, n doit être inférieur à :"
[1] 13
[1] "probabilité que l'estimateur vale +infini avec cette valeur de n :"
[1] 0.5133421

```

2 Deuxième stratégie.

Question 1

$\forall j = 1, \dots, m, Y_j$ représente le nombre de poissons pêchés entre le $j^{\text{ème}}$ poisson bagué et le $(j+1)^{\text{ème}}$. On répète successivement l'expérience "on pêche un poisson" jusqu'à ce qu'il soit bagué, de probabilité $\frac{n_0}{\theta}$. Les Y_i suivent donc tous une loi géométrique de paramètre $\frac{n_0}{\theta}$.

D'autre part, ils concernent des pêches différentes, donc ils sont de plus indépendants.

```
> ech
[1] 22 42 5 46 1 11 13 7 12 1 4 37 70 32 4 1 2 17 90 43 50 11 14 30 2 130
[27] 13 100 6 4 31 6 3 35 16 22 8 13 16 24 5 29 13 3 69 77 13 1 5 56 34 18
[53] 4 11 49 18 27 29 6 11 9 32 36 69 32 5 47 5 14 57 31 20 40 10 75 17 14 28
[79] 22 15
```

En théorie, la moyenne et la variance d'une loi géométrique $G(p)$ sont données par $\frac{1}{p}$ et $\frac{1-p}{p^2}$, soit pour $n_0 = 50$ et $\theta = 1000$, $\mathbb{E}(Y) = 20$ et $\text{Var}(Y) = 380$. Pour notre échantillon, nous calculons :

```
> mean(ech)
[1] 25.625

> var(ech)
[1] 638.9462
```

Les résultats que nous obtenons sont proches des résultats théoriques.

Question 2

En notant N le nombre de poissons pêchés, on a $N = \sum_{j=1}^m Y_j$

$\Omega(N) = [m, +\infty[$

Soit $n \geq m$, et $x = (x_1, \dots, x_n) \in (N = n)$.

On sait que $x_n = 1$ (le dernier poisson pêché est bagué par définition de Y_n), $m-1$ des $n-1$ autres pêches sont des poissons bagués. Ceci justifie que $\#(N = n) = \binom{n-1}{m-1}$.

Chacun des éléments de $(N = n)$ étant de probabilité $(\frac{n_0}{\theta})^m (1 - \frac{n_0}{\theta})^{n-m}$, on obtient finalement :

$$\forall n \in \mathbb{N}, \mathbb{P}(N = m + n) = \binom{m+n-1}{m-1} \left(\frac{n_0}{\theta}\right)^m \left(1 - \frac{n_0}{\theta}\right)^n$$

Ici,

```
> n
[1] 2050
```

Question 3

L'estimateur par la méthode des moments est $\bar{\theta}'_m = n_0 \bar{Y}$

On cherche à annuler la dérivée de la fonction de vraisemblance \mathcal{L} :

$$\begin{aligned} \mathcal{L}(\theta; y_1, \dots, y_n) &= \prod_{j=1}^m P(Y_j = y_j; \theta) \\ &= \prod_{j=1}^m \left(1 - \frac{n_0}{\theta}\right)^{y_j-1} \frac{n_0}{\theta} \\ &= \left(\frac{n_0}{\theta}\right)^m \prod_{j=1}^m \left(1 - \frac{n_0}{\theta}\right)^{y_j-1} \end{aligned} \tag{7}$$

On calcule donc $\frac{\partial \ln \mathcal{L}}{\partial \theta}$ et on cherche ensuite à l'annuler pour maximiser la log-vraisemblance :

$$\begin{aligned}\ln \mathcal{L}(\theta; y_1, \dots, y_n) &= m \ln \frac{n_0}{\theta} + \sum_{j=1}^m (y_j - 1) \ln 1 - \frac{n_0}{\theta} \\ \frac{\ln \partial \mathcal{L}}{\partial \theta} &= \frac{-m}{\theta} + \sum_{j=1}^m (y_j - 1) \frac{-\frac{n_0}{\theta^2}}{1 - \frac{n_0}{\theta}} \\ &= \frac{-m}{\theta} - \sum_{j=1}^m (y_j - 1) \frac{n_0}{\theta^2 - \theta n_0}\end{aligned}\quad (8)$$

Avant simplification, on trouve $\hat{\theta}'_m = n_0 \frac{\sum_{j=1}^m (y_j - 1) + m}{m}$, puis le $+m$ au numérateur compense les -1 , et on retrouve $\hat{\theta}'_m = n_0 \bar{Y}$

Ainsi, pour notre échantillon, nous calculons :

```
> est
[1] 1281.25
```

Question 4

Les Y_i sont indépendantes et suivent la même loi, on a donc :

$$\begin{aligned}\mathcal{I}_m &= m\mathcal{I}_1 \\ &= m\text{Var}\left(\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta, Y_1)\right) \\ &= m\text{Var}\left(\frac{\partial}{\partial \theta} \left[\ln \frac{n_0}{\theta}\right] + (Y_1 - 1) \ln\left(1 - \frac{n_0}{\theta}\right)\right) \\ &= m\text{Var}\left(\frac{-1}{\theta} + (Y_1 - 1) \frac{n_0}{\theta^2 - \theta n_0}\right) \\ &= m\left(\frac{n_0}{\theta^2 - \theta n_0}\right)^2 \left(\left(\frac{\theta}{n_0}\right)^2 - \frac{\theta}{n_0}\right)\end{aligned}\quad (9)$$

Finalement, $\mathcal{I}_m = \frac{m}{\theta^2 - \theta n_0}$

Pour notre échantillon, on obtient

```
> I_m(teta)
[1] 8.421053e-05
```

$\mathbb{E}(\hat{\theta}'_m) = n_0 \mathbb{E}(\bar{Y}) = \theta$ donc $\hat{\theta}'_m$ est sans biais. À présent, on déduit de l'inégalité FDCR que cet estimateur est de variance minimale si et seulement si sa variance est égale à $\frac{1}{\mathcal{I}_m(\theta)}$, donc égale à $\frac{\theta^2}{m} - \frac{\theta n_0}{m}$.

$$\begin{aligned}\text{Var}(\hat{\theta}'_m) &= \text{Var}\left(\frac{n_0}{m} \sum_{j=1}^m Y_j\right) \text{ (où les } Y_i \text{ sont indépendantes)} \\ &= \frac{n_0^2}{m} \text{Var}(Y_1) \\ &= \frac{n_0^2}{m} \left[\left(\frac{\theta}{n_0}\right)^2 - \frac{\theta}{n_0}\right] \\ &= \frac{\theta^2}{m} - \frac{\theta n_0}{m}\end{aligned}\quad (10)$$

$\hat{\theta}'_m$ est donc effectivement de variance minimale.

Question 5 *Le calcul des intervalles de confiance pour notre échantillon donne les résultats suivants :*

```
> IC(seuils)
[1] 0.01
[1] 919.5386 1642.9614
[1] 0.05
[1] 1006.022 1556.478
[1] 0.1
[1] 1050.271 1512.229
[1] 0.2
[1] 1101.288 1461.212
```

On obtient le tableau suivant :

α	<i>Intervalle de confiance</i>
1%	[919.5, 1643.0]
5%	[1006.0, 1556.5]
10%	[1050.3, 1512.2]
20%	[1101.3, 1461.2]

3 Application et comparaison des stratégies.

Question 1

La première stratégie consiste à évaluer le ratio $\frac{\text{NombreDePoissonsBagues}}{\text{NombreTotalDePoissons}}$, puis, connaissant le nombre exact de poissons bagués, en déduire une estimation du nombre total de poissons. Sur 1000 poissons pêchés, 35 sont bagués. Il y a donc une proportion $p = \frac{35}{1000} = 0.035$ de poissons bagués parmi les poissons pêchés. On peut donc estimer $\theta = 0.035$.

On calcule les intervalles de confiances :

1. Intervalle de confiance exact : $[0.02449753, 0.02533891]$
2. Intervalle de confiance asymptotique de seuil à 5% : $[0.02544073, 0.04455927]$

Question 2

La deuxième stratégie repose sur la supposition suivante : le nombre de poissons pêchés avant de tomber sur un poisson bagué suit une loi géométrique de paramètre λ à déterminer. Ce paramètre est le ratio précédent, donc :

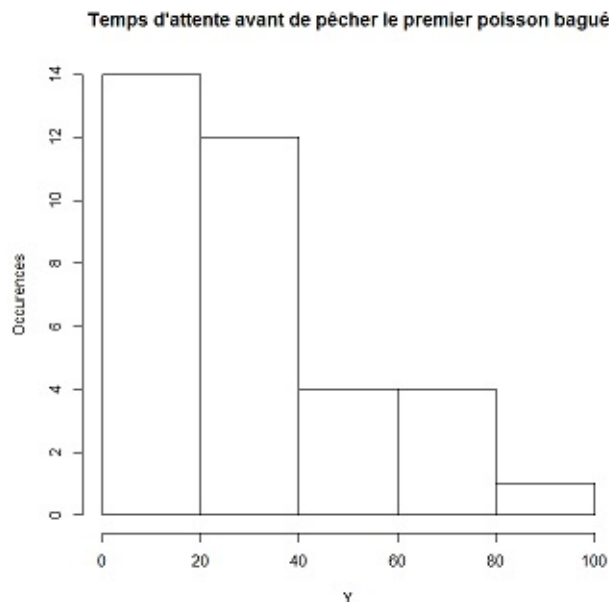
$$\lambda = \frac{\text{NombreDePoissonsBagues}}{\text{NombreTotalDePoissons}}.$$

Donc une estimation de λ conduira de façon inévitable à une estimation du ratio. La deuxième stratégie consiste à évaluer λ .

On trouve $\lambda = 0.03521127$. Calculons l'intervalle de confiance asymptotique de seuil à 5% : $[0.00000000, 0.08645611]$

Question 3

R nous permet de tracer cet histogramme :



Cet histogramme ressemble fort à la représentation d'une variable aléatoire qui suit une loi géométrique. On peut donc considérer que l'hypothèse d'une loi géométrique est vérifiée. En tout cas, elle n'est pas absurde.

Question 4

Pour décider de la meilleure stratégie, comparons les intervalles de confiance asymptotiques. Intéressons nous plus particulièrement aux longueurs de ces intervalles :

- 1. Stratégie 1 : 0.01911855*
- 2. Stratégie 2 : 0.08645611*

Il apparaît que la longueur pour la deuxième stratégie est presque 5 fois plus importante que celui de la première stratégie.

Il est donc préférable de choisir la première stratégie.

4 Vérifications expérimentales à base de simulations.

Question 1

On calcule en R les différentes proportions d'appartenance de θ pour les intervalles de confiance établis en variant les paramètres.

```
[1] "proportions (en pourcentage) en fonction de l'augmentation de theta (intervalles de confiance exacts) :"  
[1] 98 95 96 94 98 96  
[1] "proportions (en pourcentage) en fonction de l'augmentation de theta (intervalles de confiance asymptotiques) :"  
[1] 93 89 91 89 80 74
```

Lorsqu'on augmente θ , on remarque que la proportion d'appartenance aux intervalles de θ diminue lorsqu'on choisit l'intervalle de confiance asymptotique.

En effet, lorsqu'on augmente θ , la probabilité p diminue donc l'intervalle asymptotique a une plus grande probabilité de se "tromper" car le nombre n d'essais n'est pas grand.

```
[1] "proportions (en pourcentage) en fonction de l'augmentation de n0 (intervalles de confiance exacts) :"  
[1] 95 96 99 97 99 97  
[1] "proportions (en pourcentage) en fonction de l'augmentation de n0 (intervalles de confiance asymptotiques) :"  
[1] 87 94 88 97 95 89
```

Lorsqu'on augmente n_0 , on remarque que la proportion d'appartenance augmente très légèrement.

En effet, comme n_0 est proportionnel à p et pour la même raison que précédemment, cela paraît cohérent.

```
[1] "proportions (en pourcentage) en fonction de l'augmentation de n (intervalles de confiance exacts) :"  
[1] 94 98 97 91 97 94  
[1] "proportions (en pourcentage) en fonction de l'augmentation de n (intervalles de confiance asymptotiques) :"  
[1] 95 96 97 96 96 93
```

Lorsqu'on augmente n , on ne constate pas de changement particulier, ce qui peut paraître incohérent car concernant l'intervalle asymptotique, ce dernier devrait être plus précis lorsque n est grand.

```
[1] "proportions (en pourcentage) en fonction de l'augmentation de m (intervalles de confiance exacts) :"
```

[1]	95.00	94.00	95.20	94.60	95.02	95.62
-----	-------	-------	-------	-------	-------	-------

```
[1] "proportions (en pourcentage) en fonction de l'augmentation de m (intervalles de confiance asymptotiques) :"
```

[1]	95.00	93.50	93.80	93.60	92.74	93.31
-----	-------	-------	-------	-------	-------	-------

Lorsqu'on augmente m , on constate une stabilisation au niveau des proportions donc de la précision des intervalles.

En effet, augmenter m permet juste de rendre la simulation plus précise puisqu'on se base sur un plus grand nombre d'essais pour établir les proportions.

```
[1] "proportions (en pourcentage) en fonction de l'augmentation de alpha (intervalles de confiance exacts) :"
```

[1]	98	96	90	85	48	36
-----	----	----	----	----	----	----

```
[1] "proportions (en pourcentage) en fonction de l'augmentation de alpha (intervalles de confiance asymptotiques) :"
```

[1]	85	93	88	79	50	8
-----	----	----	----	----	----	---

Lorsqu'on augmente α , quel que soit le type d'intervalle (exact ou asymptotique), il est cohérent que la proportion diminue car on baisse le niveau de confiance donc la précision de ces intervalles.

Question 2

On simule $m = 100$ échantillons de taille $n = 5, 10, 100, 1000, 10000, 100000$ de loi de Bernoulli, puis on compare les m moyennes empiriques avec les espérances (avec une erreur de $\epsilon = 0.01$) pour illustrer la loi faible des grands nombres :

```
[1] "valeur de n :"
```

[1]	5e+00	1e+01	1e+02	1e+03	1e+04	1e+05
-----	-------	-------	-------	-------	-------	-------

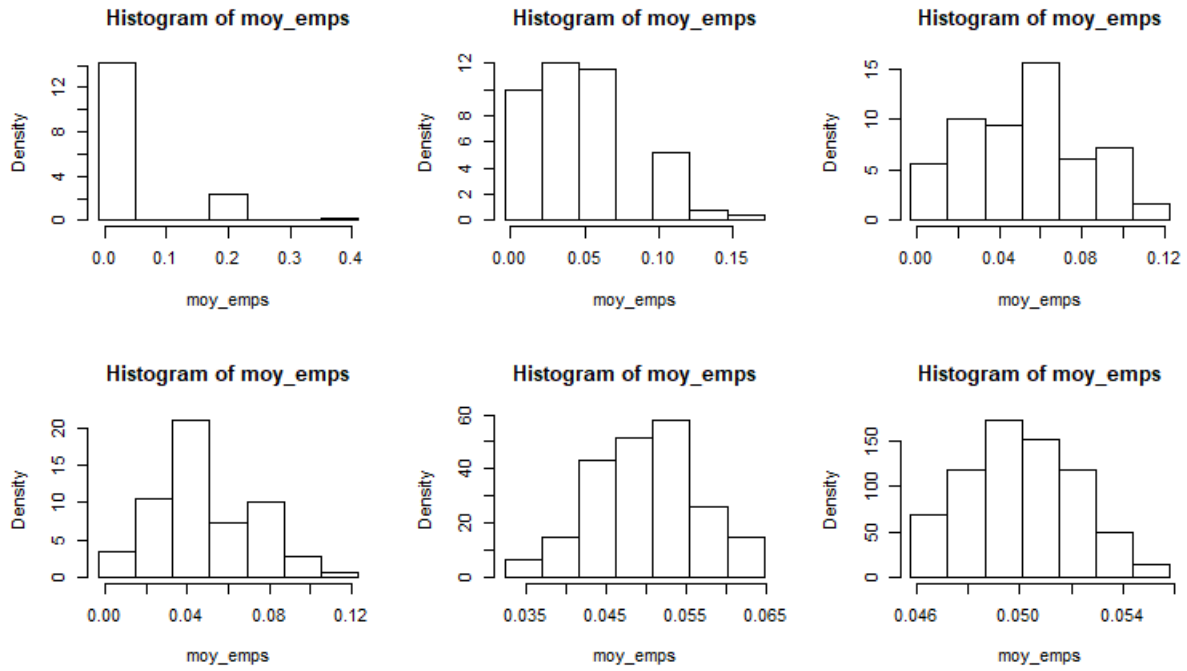
```
[1] "proportions (en pourcentage) en fonction de n :"
```

[1]	0	0	24	85	100	100
-----	---	---	----	----	-----	-----

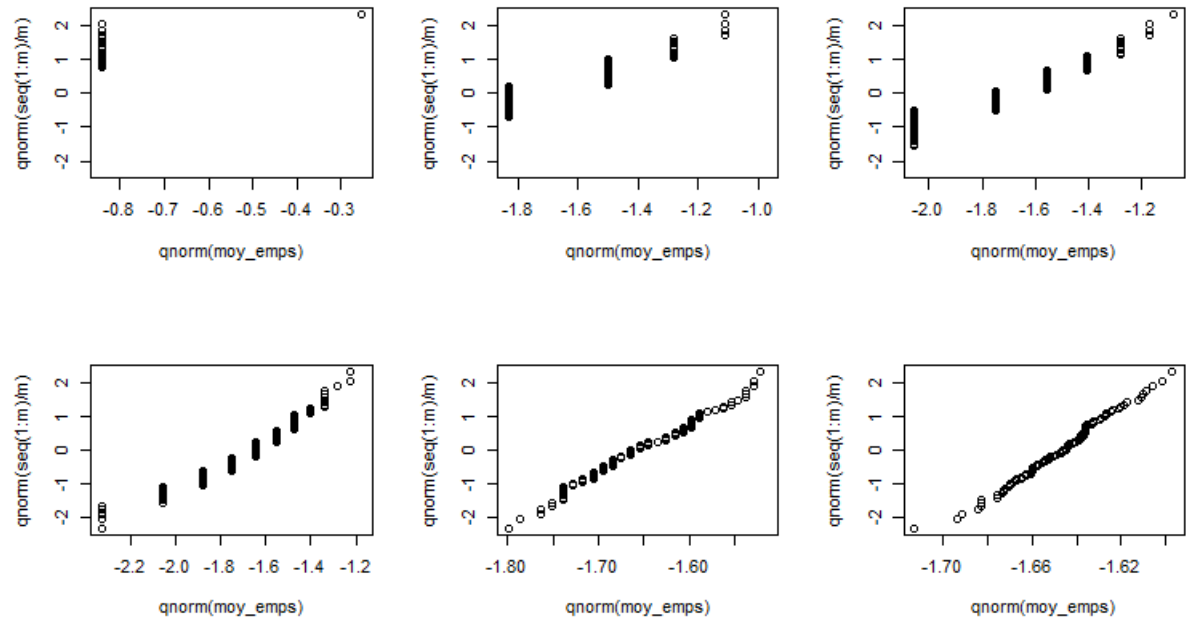
Lorsque n augmente, on remarque que l'écart entre la moyenne empirique et l'espérance diminue. A partir de $n = 1000$ environ, on peut dire que la moyenne empirique peut être approximée par l'espérance. La loi faible des grands nombres est ainsi illustrée.

Question 3

De la même manière que précédemment, on va faire la même simulation avec $m = 100$ et en faisant varier $n = 5, 30, 50, 100, 1000, 10000$, sauf qu'on va tracer les histogrammes (de même largeur en l'occurrence) des moyennes empiriques pour chaque valeur de n ainsi que les graphes de probabilités associés pour la loi normale :



On remarque que la courbe associée à l'histogramme s'apparente à celle modélisant la fonction de densité de la loi normale à partir de $n = 50$ environ.



De même, en traçant les graphes de probabilité $(F_{\mathcal{N}}^{-1}(\bar{x}_i), F_{\mathcal{N}}^{-1}(\frac{i}{m}))$ où $\forall i = 1, \dots, n$ \bar{x}_i est une des m moyennes empiriques de chaque groupe de n x_i (avec les x_i qui sont iid et suivent chacune une loi de Bernoulli) de l'échantillon de données. On remarque de même l'apparence d'une droite à partir de $n = 50$ ce qui confirme le fait qu'à partir d'une certaine valeur, la moyenne empirique des x_i suit une loi normale. Le théorème central limite est ainsi illustré.