Principes et Méthodes Statistiques
Chars d'assaut allemands et iPhones 3G

Roos Bastien, Bouton Paul, Bayard Guillaume

3 mai 2016

1 Tirage avec remise

1)

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=1}^{\theta} k.\mathbb{P}(X=k) \\
&= \sum_{k=1}^{\theta} k.\frac{1}{\theta} \\
&= \frac{1}{\theta}.\frac{\theta(\theta+1)}{2} \\
&= \frac{\theta+1}{2} \\
\mathbb{E}[X^2] &= \sum_{k=1}^{\theta} k^2.\mathbb{P}(X=k) \\
&= \sum_{k=1}^{\theta} k^2.\frac{1}{\theta} \\
&= \frac{1}{\theta}\sum_{k=1}^{\theta} k^2. \\
&= \frac{1}{\theta}\frac{\theta(\theta+1)(2\theta+1)}{6} \\
&= \frac{(\theta+1)(2\theta+1)}{6} \\
Var(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
&= \frac{(\theta+1)(2\theta+1)}{6} - (\frac{\theta+1}{2})^2 \\
&= \frac{2(\theta+1)(2\theta+1) - 3(\theta+1)^2}{12} \\
&= \frac{\theta^2-1}{12}
\end{aligned}$$

Conclusion

$$\begin{aligned}
\mathbb{E}[X] &= \frac{\theta+1}{2} \\
Var(X) &= \frac{\theta^2-1}{12}
\end{aligned}$$

2. $\mathbb{E}[X] = \varphi(\theta)$ ie $\varphi : x \mapsto 2x-1 \implies \tilde{\theta}_n = \varphi^{-1}(\overline{X}_n) = 2\overline{X}_n - 1$

$$\begin{aligned}
\mathbb{E}[\tilde{\theta}_n] &= 2\mathbb{E}[\overline{X}_n] - 1 \\
&= 2\frac{\theta+1}{2} - 1 \\
&= \theta + 1 - 1 \\
&= \theta \\
Var(\tilde{\theta}_n) &= Var(2\overline{X_n} - 1) \\
&= 4Var(\frac{1}{n}\sum_{i=1}^{n} X_i) \\
&= \frac{4}{n^2}\sum_{i=1}^{n} Var(X_i) \\
&= \frac{4}{n^2}\sum_{i=1}^{n} \frac{\theta^2-1}{12} \\
&= \frac{\theta^2-1}{3n}
\end{aligned}$$

D'où $\tilde{\theta}_n$ est sans biais.

Conclusion
L'estimateur $\tilde{\theta}_n$ est sans biais et a pour variance

$$Var(\tilde{\theta}_n) = \frac{\theta^2-1}{3n}$$

3)
Calculons la médiane M de la loi de X.
1er cas : θ est pair.
Alors $F_X\lfloor M \rfloor = \mathbb{P}(X \leq M) = \frac{1}{2} = \frac{1}{\theta}\lfloor M \rfloor \implies \lfloor M \rfloor = \frac{\theta}{2}$
2ème cas : θ est impair

$$\begin{aligned}
& F_X(M-1) < \frac{1}{2} \text{ et } F_X(M) > \frac{1}{2} \\
\implies & \frac{1}{\theta}\lfloor M-1 \rfloor < \frac{1}{2} \text{ et } \frac{1}{\theta}\lfloor M \rfloor > \frac{1}{2} \\
\implies & \frac{1}{2} + \frac{1}{\theta} > \frac{1}{\theta}\lfloor M \rfloor > \frac{1}{2} \\
\implies & \frac{\theta}{2} + 1 > \lfloor M \rfloor > \frac{\theta}{2} \\
\implies & \lfloor M \rfloor = \frac{\theta+1}{2} \\
\Leftrightarrow & \theta = 2\lfloor M \rfloor - 1
\end{aligned}$$

$$
\begin{aligned}
F_X(t) &= \mathbb{P}(X \leq t) \\
&= \mathbb{P}\left(X = \bigcup_{k=1}^{\lfloor t \rfloor} k\right) \\
&= \sum_{k=1}^{\lfloor t \rfloor} \mathbb{P}(X = k) \\
&= \sum_{k=1}^{\lfloor t \rfloor} \frac{1}{\theta} \mathbf{1}_{1..\theta} \\
&= \begin{cases} 0 \text{ si t} \leq 0 \\ \frac{1}{\theta}\lfloor t \rfloor \text{ si } t \in]0;1[\\ 1 \text{ si t} \geq 0 \end{cases}
\end{aligned}
$$

Conclusion

$$
F_X(t) = \begin{cases} 0 \text{ si t} \leq 0 \\ \frac{\lfloor t \rfloor}{\theta} \text{ si } t \in]0;1[\\ 1 \text{ si t} \geq 0 \end{cases}
$$

La médiane M est donnée par

$$
\begin{cases} M = \frac{\theta}{2} \text{ si } \theta \text{ est pair} \\ \lfloor M \rfloor = \frac{\theta+1}{2} \text{ si } t \in]0;1[\end{cases}
$$

4)

$$
\begin{aligned}
(X_n^* = k) &= \bigcup_{j=1}^{n} ((j)X = k \text{ et } (n-j)X < k) \\
&= \bigcup_{j=1}^{n} \binom{n}{j} (X = k)^j (X < k)^{n-j}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{P}(X_n^* = k) &= \sum_{j=1}^{n} \binom{n}{j} (\mathbb{P}X = k)^j (\mathbb{P}(X < k))^{n-j} \\
&= \sum_{j=1}^{n} \binom{n}{j} \frac{1}{\theta^j} \frac{n-j}{\theta^{n-j}} \\
&= \frac{1}{\theta^n} \sum_{j=1}^{n} \binom{n}{j} (k-1)^{n-j} \\
&= \frac{1}{\theta^n} \underbrace{\sum_{j=1}^{n} \binom{n}{j} 1^j (k-1)^{n-j}}_{(1+k-1)^n - \frac{(k-1)^n}{\theta^n}} \\
&= \frac{k^n - (k-1)^n}{\theta^n}
\end{aligned}
$$

$$
\begin{aligned}
F_{X_n^*}(t) &= \mathbb{P}(X_n^* \leq t) \\
&= \mathbb{P}\left(\bigcap_{i=1} nX_i \leq t \right) \\
&= \prod_{i=1}^{n} \mathbb{P}(X_i \leq t)) \\
&= \prod_{i=1}^{n} F_{X_i}(t) \\
&= \begin{cases} 0 \text{ si t} \leq 0 \\ \left(\frac{\lfloor t \rfloor}{\theta} \right)^n \text{ si } t \in]0;1[\\ 1 \text{ si t} \geq 0 \end{cases}
\end{aligned}
$$

Conclusion

$$
\begin{aligned}
F_{X_n^*}(t) &= \begin{cases} 0 \text{ si t} \leq 0 \\ \left(\frac{\lfloor t \rfloor}{\theta} \right)^n \text{ si } t \in]0;1[\\ 1 \text{ si t} \geq 0 \end{cases} \\
\mathbb{P}(X_n^* = k) &= \frac{k^n - (k-1)^n}{\theta^n}
\end{aligned}
$$

5)

$$\begin{aligned} & \mathcal{L}_{x_1,\dots,x_n}(\theta) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) = \frac{1}{\theta^n} \\ \Longrightarrow \quad & \ln\left(\mathcal{L}_{x_1,\dots,x_n}(\theta)\right) = -n.ln(\theta) \\ \Longrightarrow \quad & \frac{d}{d\theta}\left(\ln\left(\mathcal{L}_{x_1,\dots,x_n}(\theta)\right)\right) = -\frac{1}{\theta} \end{aligned}$$

On a θ_{min} pour $\max\mathcal{L}_{x_1,\dots,x_n}(\theta)$ mais $\theta \geq \max(x_i)$
$\Longrightarrow \theta \geq X_n^*$
d'où $\hat{\theta}_n = \min\theta = X_n^*$

$$\begin{aligned} \mathbb{E}[X_n^*] &= \sum_{k=1}^{\theta} k\mathbb{P}(X_n^* = k) \\ &= \frac{1}{\theta^n}\left(\sum_{k=1}^{\theta} k^{n+1} - \sum_{k=1}^{\theta} k(k-1)^n\right) \\ &= \frac{1}{\theta^n}\left(\theta^{n+1} + \sum_{k=1}^{\theta-1} k^n(k-k-1)\right) \\ &= \theta - \frac{1}{\theta^n}\sum_{k=1}^{\theta-1} k^n \\ &\neq \theta \end{aligned}$$

D'où l'estimateur de vraissemblance $\hat{\theta}_n$ est biaisé.

Conclusion

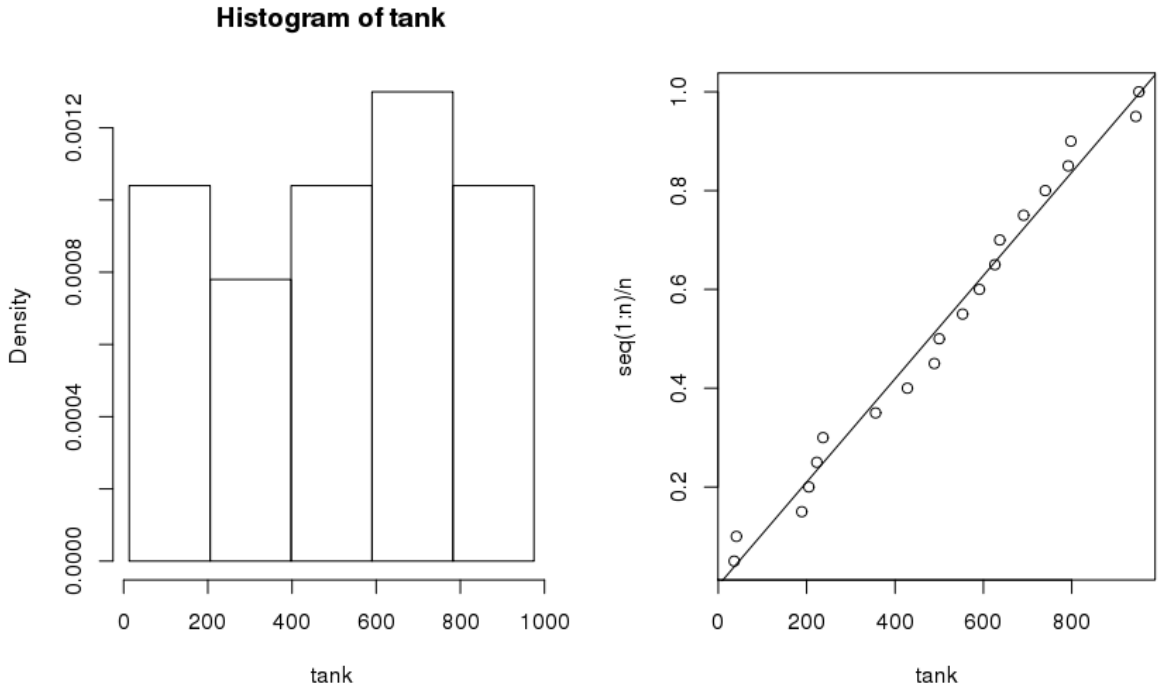
L'estimateur de maximum de vraisemblance de θ est $\hat{\theta}_n = X_n^*$ et est est biaisé.

6) On considère un tirage avec remise dont les résultats (x_i) sont ordonnées par ordre croissant (x_i^*) avec $i \leq \theta$.
On cherche donc h tel que $h(F(x)) = \alpha(\theta).g(x) + \beta(\theta)$.
Or $F(x_i^*) = \frac{\lfloor (x_i^*) \rfloor}{\theta}$ => $\begin{cases} \alpha(\theta) = \frac{1}{\theta} \\ \beta(\theta) = 0 \end{cases}$
Or on a également $F(x_i^*) = \frac{1}{n}$
=> $\frac{i}{n} = \frac{\lfloor x_i^* \rfloor}{\theta}$
On trace la droite $(\lfloor (x_i^* \rfloor; \frac{i}{n})$, on obtient une droite de coefficient $\frac{1}{\theta_g}$.
=> $\theta_g = \frac{1}{\text{coeff directeur}} = \frac{x_n^* - x_1^*}{\frac{n}{n} - \frac{1}{n}}$

Conclusion

$$\theta_g = \frac{n(x_n^* - x_1^*)}{n-1}$$

7)



Concernant l'histogramme, on s'aperçoit qu'il y a peu d'écarts entre les effectifs par classe. De plus, notre graphe de probabilités semble présenter des points alignés. On a donc une adéquation entre la théorie et l'expérimentation (les X_i suivent bien une loi uniforme sur $1...\theta$). Cependant il est important de noter que les effectifs des classes de l'histogramme diffèrent beaucoup lors de certaines simulations. Cela pourrait s'expliquer par un échantillon trop petit.

8)

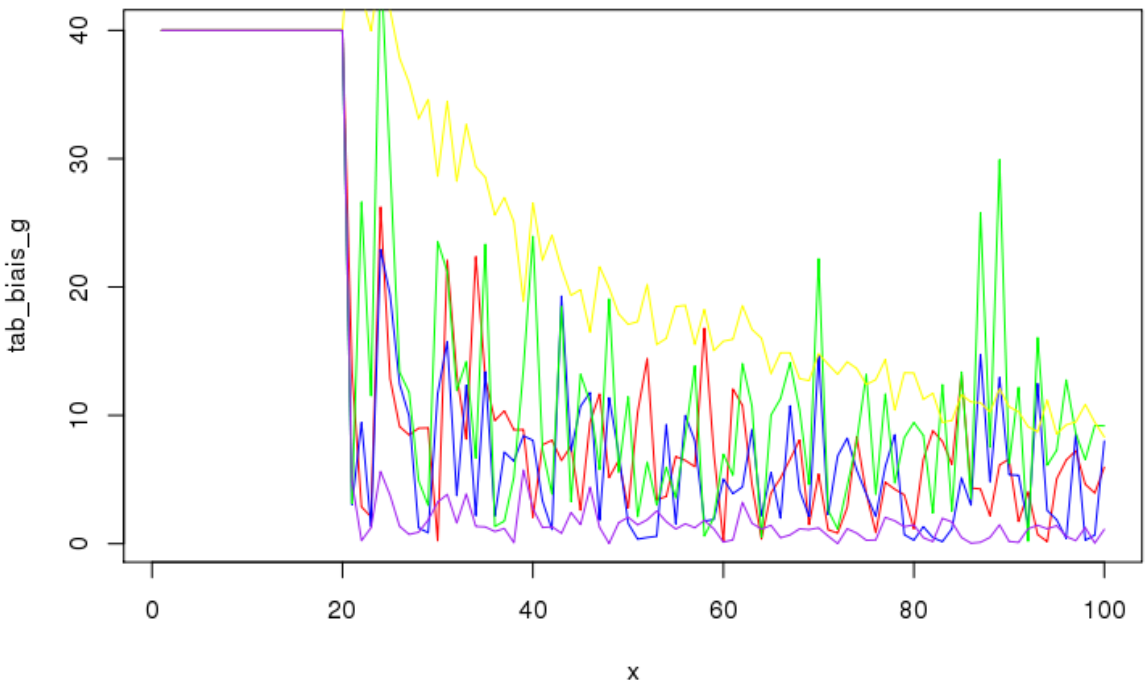


FIGURE 1 – Biais moyen en fonction de la taille de l'échantillon

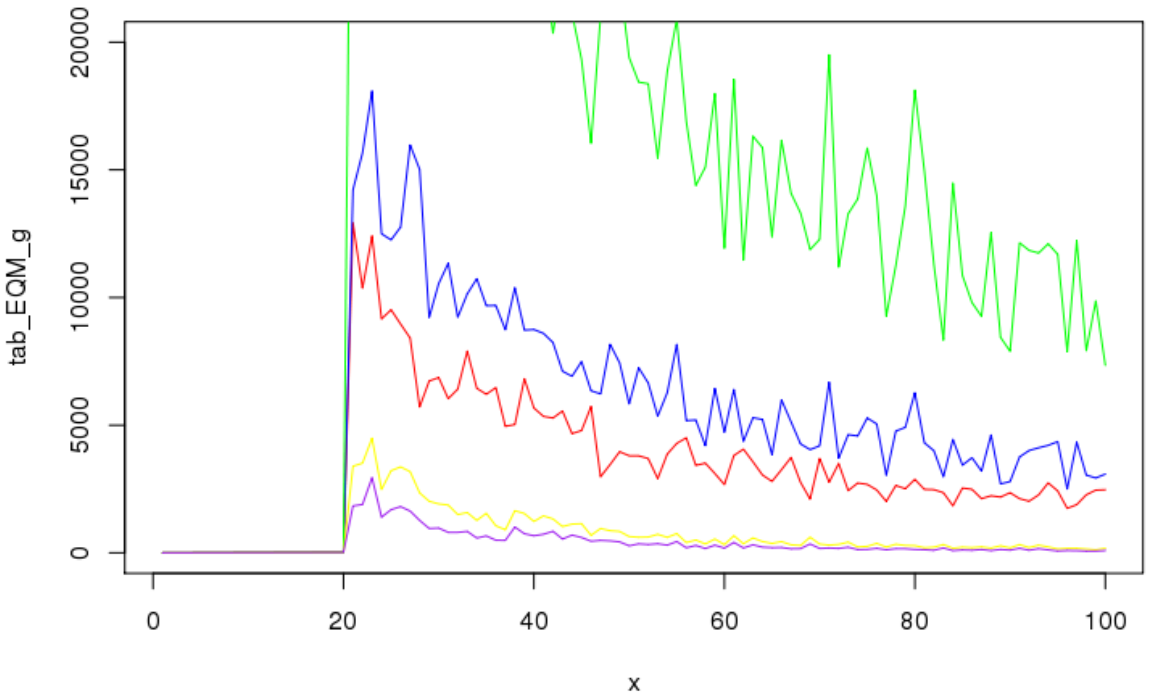


FIGURE 2 – Ecart quadratique moyen en fonction de la taille de l'échantillon

En ce qui concerne le biais, l'estimateur $\tilde{\theta}'_n$ présente de grandes variations de son biais tout en restant fortement biaisé.
En ce qui concerne $\hat{\theta}_n$, le biais diminue avec la taille de l'échantillon mais reste quand encore important pour un échantillon assez grand (100 individus).
Pour l'estimateur graphique et $\tilde{\theta}_n$ le biais diminue de nouveau avec la taille de l'échantillon mais reste cependant très instable.
C'est logiquement l'ESBVM qui présente un biais minimal (quasi-nul pour un échantillon de taille 100). De plus, en s'intéressant à l'écart quadratique moyen, il est clair que $\hat{\theta}_n$ et l'ESBVM sont les deux meilleurs estimateurs parmi les 5.
Ainsi il ressort de nos expérimentations et du croisement de nos deux graphiques que l'ESBVM fourni est le meilleur que l'on puisse utiliser pour notre étude.

9)
$$(\mathbb{P}\left|\sqrt(n)\frac{\overline{X}_n - \mathbb{E}[X]}{\sigma(X)}\right| \leq U_\alpha) = 1 - \alpha$$

$$
\begin{aligned}
& \frac{n(\overline{X}_n - \mathbb{E}[X])^2}{\sigma(X)^2} \leq U_\alpha^2 \\
\Leftrightarrow \quad & \frac{n(\overline{X}_n - 2\mathbb{E}[X]\overline{X}_n + \mathbb{E}[X]^2}{\sigma(X)^2} \leq U_\alpha^2 \\
\Leftrightarrow \quad & 12n(\overline{X}_n^2 - (\theta+1)\overline{X}_n + \frac{(\theta+1)^2}{4}) \leq U_\alpha^2(\theta^2-1) \\
\Leftrightarrow \quad & 12n(\overline{X}_n^2 - \theta\overline{X}_n - \overline{X}_n + \frac{\theta^2+2\theta+1}{4}) \leq U_\alpha^2(\theta^2-1) \\
\Leftrightarrow \quad & 0 \leq (U_\alpha^2 - 3n)\theta^2 + (12n\overline{X}_n - 6n)\theta + (12n\overline{X}_n - 12n\overline{X}_n^2 - 3n - U_\alpha^2) \\
\Leftrightarrow \quad & 0 \leq (U_\alpha^2 - 3n)\theta^2 + 6n\tilde{\theta}_n\theta - (3n\tilde{\theta}_n^2 + U_\alpha^2)
\end{aligned}
$$

On calcule le discriminant Δ(on néglige les termes en U_α) :

$$
\begin{aligned}
\Delta &= 36n^2\tilde{\theta}_n^2 + 4(3n\tilde{\theta}_n^2 + U_\alpha^2)(U_\alpha^2 - 3n) \\
&= 12nU_\alpha^2(\tilde{\theta}_n^2 - 1)
\end{aligned}
$$

$$
\implies R_{12} = \frac{-6n\tilde{\theta}_n \pm \sqrt{12nU_\alpha^2(\tilde{\theta}_n^2-1)}}{2(U_\alpha^2 - 3n)} = \tilde{\theta}_n \pm \sqrt{\frac{U_\alpha^2}{3n}(\tilde{\theta}_n^2 - 1)}
$$

$$
\implies IC = \left[\tilde{\theta}_n - \sqrt{\frac{U_\alpha^2}{3n}(\tilde{\theta}_n^2-1)}; \tilde{\theta}_n + \sqrt{\frac{U_\alpha^2}{3n}(\tilde{\theta}_n^2-1)}\right]
$$

Pour des valeurs de n, m et θ raisonnables ($\theta = 1000, n = 100, m = 100$), on observe dans 94 à 96% des cas que θ appartient à l'intervalle de confiance. Par conséquent, pour de telles valeurs, on a une confiance d'environ 95% dans le fait que θ soit dans cet intervalle. De plus, même pour des valeurs plus extrêmes (n et m petits), ce pourcentage descend rarement en dessous de 70%. Ainsi, notre intervalle reste cependant pertinent malgré les simplifications des termes en U_α lors du calcul du discriminant.

Conclusion
On obtient ainsi un intervalle de confiance IC :

$$
IC = \left[\tilde{\theta}_n - \sqrt{\frac{U_\alpha^2}{3n}(\tilde{\theta}_n^2-1)}; \tilde{\theta}_n + \sqrt{\frac{U_\alpha^2}{3n}(\tilde{\theta}_n^2-1)}\right]
$$

2 Tirage sans remise

1)

$$
\mathbb{P}(X_1 = k) = \frac{1}{\theta} \text{ si k}\in \{1,..,\theta\} \implies X_1 \sim \mathcal{U}_{\{1,...,\theta\}}
$$

$$
\mathbb{P}(X_2 = k | X_1 = x_1) = \frac{1}{\theta - 1} \text{ si } k \neq x_1 \implies (X_2 \text{ sachant } [X_1 = x_1]) \sim \mathcal{U}_{\{1,...,\theta\}\setminus x_1}
$$

$$
\mathbb{P}(X_3 = k | X_1 = x_1, X_2 = x_2) = \frac{1}{\theta - 2} \text{ si } k \neq x_1 \text{ et } k \neq x_2
$$

$$
\implies (X_2 \text{ sachant } [X_1 = x_1]) \sim \mathcal{U}_{\{1,...,\theta\}\setminus x_1}
$$

donc (X_i sachant $[X_1 = x_1, ..., X_n = x_n]$) $\sim \mathcal{U}_{\{1,...,\theta\}\setminus x_1,x_2,..,x_{i-1}}$
$\mathcal{L}(\theta; x_1, ..., x_n) = \frac{1}{\theta} \prod_{i=2}^{n-1} \frac{1}{\theta - (i+1)}$
$\Longrightarrow$ $\mathcal{L}(\theta; x_1, ..., x_n)_max$ est atteint lorsque θ est minimum

mais on a toujours $\theta \geq \max_{i=1..n} x_i = X_n^*$

d'où

$$\hat{\theta}_n = X_n^*$$

Conclusion
L'estimateur de maximum de vraisemblance $\hat{\theta}_n$ de θ est toujours

$$\hat{\theta}_n = X_n^*$$

2) $\mathbb{P}(X_n^*) = n.\frac{1}{\theta}.\prod_{j=1}^{n-1} \frac{k-j}{\theta - j}$
avec
n : position du max
θ : valeur du max
$\frac{k-j}{\theta-j}$:valeur du j-ème

$$\begin{aligned}
\Longrightarrow \mathbb{P}(X_n^* = k) &= n.\frac{1}{\theta}.\frac{(k-1)!}{(k-n)!}.\frac{(\theta-n)!}{(\theta-1)!} \\
&= \frac{n!(k-1)!(\theta-n)!}{(n-1)!\theta!(k-n)!} \\
&= \frac{(k-1)!}{(n-1)!(k-n)!}.\frac{n!(\theta-n)!}{\theta!} \\
&= \frac{\binom{k-1}{n-1}}{\binom{\theta}{n}}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[X_n^*] &= \sum_{k=1}^{\theta} k.\mathbb{P}(X_n^* = k) \\
&= \sum_{k=n}^{\theta} k\binom{k-1}{n-1}\frac{1}{\binom{\theta}{n}} \\
&= \frac{1}{\binom{\theta}{n}}\sum_{k=n}^{\theta} k\frac{(k-1)!}{(n-1)!(k-n)!} \\
&= \frac{n}{\binom{\theta}{n}}\sum_{k=n}^{\theta}\binom{k}{n} \\
&= \frac{n}{\binom{\theta}{n}}\binom{\theta+1}{n+1} \\
&= n.\frac{n!(\theta-n)!(\theta+1)!}{\theta!(\theta-n)!(n+1)!} \\
&= \frac{n}{n+1}(\theta+1)
\end{aligned}$$

donc $\hat{\theta}^{(1)} = \frac{n+1}{n}X_n^* - 1$ est sans biais :

$$\begin{aligned}
\mathbb{E}[\hat{\theta}_n^{(1)}] &= \mathbb{E}[\frac{n+1}{n}X_n^* - 1] \\
&= \frac{n+1}{n}\mathbb{E}[X_n^*] - 1 \\
&= \theta + 1 - 1 \\
&= \theta
\end{aligned}$$

$E[X_n^*]$ et en déduire que $\hat{\theta}_n^{(1)} = \frac{n+1}{n}X_n^* - 1$ est estimateur sans biais de θ

Conclusion
$E[X_n^*] = \frac{n}{n+1}(\theta+1)$ et $\hat{\theta}_n^{(1)} = \frac{n+1}{n}X_n^* - 1$ est estimateur sans biais de θ

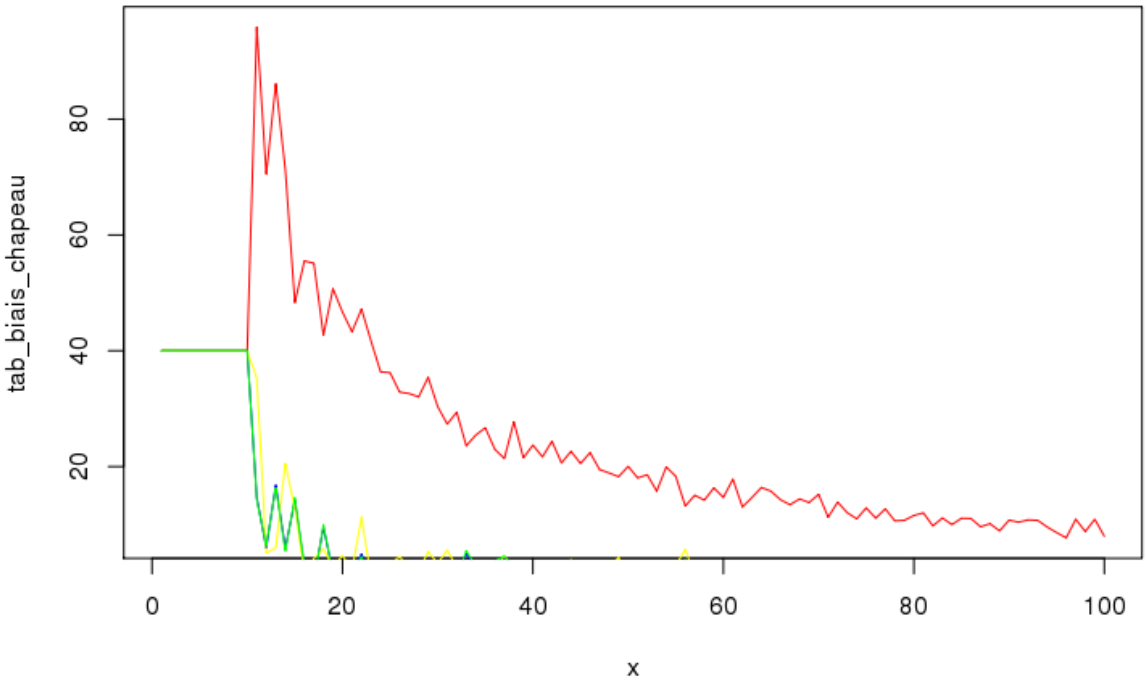


FIGURE 3 – Biais moyen en fonction de la taille de l'échantillon

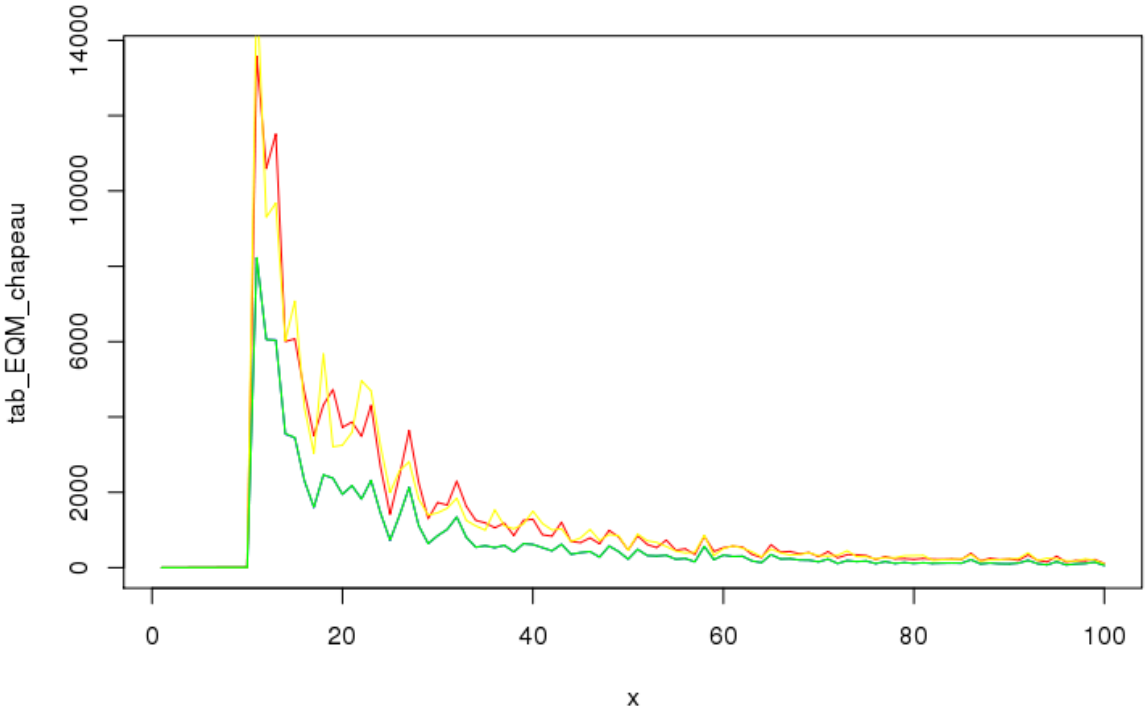


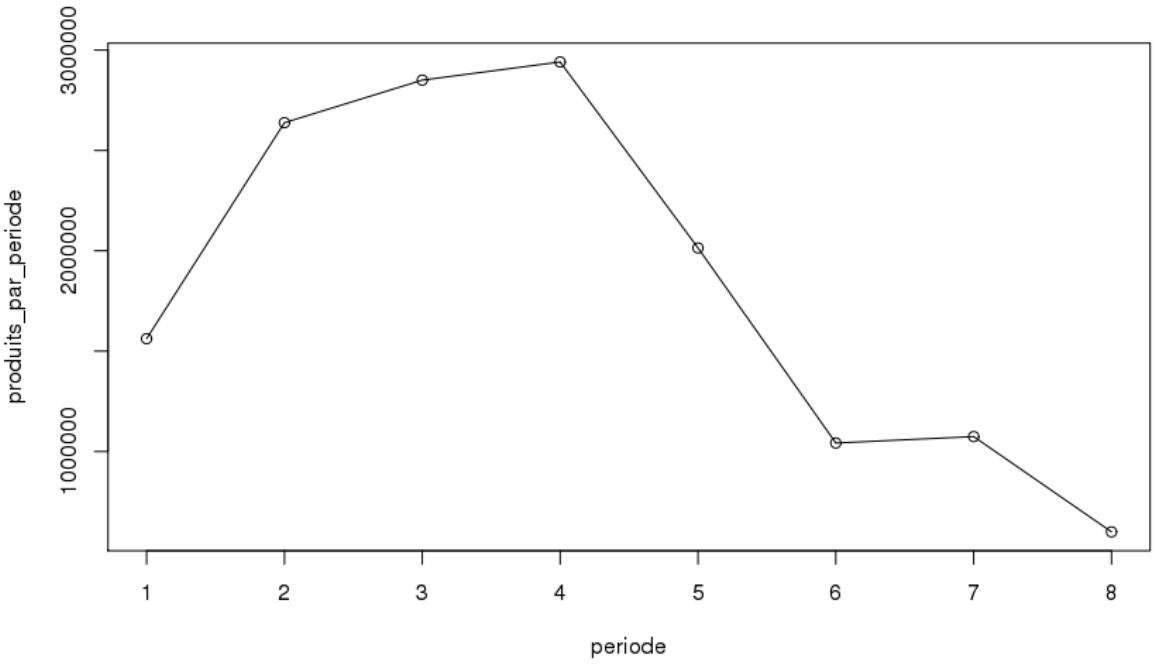
FIGURE 4 – Ecart quadratique moyen en fonction de la taille de l'échantillon

On remarque que $\hat{\theta}_n^1$ et $\hat{\theta}_n^2$ et l'ESBVM sont de bien meilleurs estimateurs dans le cas d'une simulation sans remise (biais et écart quadratique moyens beaucoup plus faibles même pour des échantillons plus petits). Ainsi il vaut mieux considérer un tirage avec remise afin de modéliser notre problème.

3 Estimation du nombre d'iPhones 3G produits

2)
On estime le nombre d'Iphone produits sur la période considérée à 14 887 670 unités (en utilisant $\tilde{\theta_n^{(1)}}$

3)



4)
L'hypothèse n'est valide que si l'on considère une période de production s'étalant de la période 2 à la période 4. Cependant, sur la totalité de la période considérée, cette hypothèse n'est pas vérifiée.

5)
Les résultats obtenus reposent sur l'hypothèse d'une uniformité des numéros de série. En effet notre estimateur est égal à $\frac{n+1}{n}X_n^* - 1$. le terme en facteur du X_n^* permet de surestimer légèrement et linéairement le nombre d'iphones produits par rapport au maximum de l'échantillon. Cependant, s'il y a des pics de production lors de certaines périodes, notre estimateur ne pourra pas en tenir compte. Effectivement, si sur la dernière période de mesure, la production augmente drastiquement et que l'on ne dispose pas des derniers numéros de série, l'estimateur ne surestimera pas assez le nombre d'unités produites En définitive, sur la période considérée, les résultats de notre étude ne peuvent pas être considérés comme totalement fiables.