Rendu du TP de Principes et Méthodes Statistiques.

Leïla Kany, Ayoub Nasr & David Colton Sowers

7 Apr 2018

1 Première Stratégie.

1.1

Quand on pêche un poisson pendant la première stratégie, on le rejette directement à l'eau. On revient a l'état initial. Donc les variables aléatoires X_1, X_2, ..., X_n sont indépendantes.
Soit donc $1 \leq i \leq n$:

$$P(X_i = 1) = \frac{n_0}{\theta}$$
$$P(X_i = 0) = 1 - \frac{n_0}{\theta}$$

X_i suit une loi de benoulli de paramètre $\frac{n_0}{\theta}$

$$\forall i, 1 \leq i \leq n, \boxed{X_i \sim \mathrm{B}\left(\frac{n_0}{\theta}\right)}$$

1.2

On a : $T = \sum_1^n X_i$ C'est une loi Binomiale de paramètres n et $\frac{n_0}{\theta}$.

$$\boxed{T \sim \mathrm{B}\left(n, \frac{n_0}{\theta}\right)}$$

Simulation

La simulation renvoie :

```
>[1] "la moyenne empirique est: 0.051000"
>[1] "la vraie moyenne est: 0.050000"
>[1] "la variance empirique est: 0.048399"
>[1] "la vraie variance est: 0.047500"
>[1] "la realisation de T est t = 51"
```

1.3

On calcule l'estimateur des moments $\tilde{\theta}_n$:
soit $s = \sum_1^n x_i$

$$\frac{n_0}{\tilde{\theta}_n} = \frac{s}{n}$$
$$\tilde{\theta}_n = \frac{n.n_0}{s}$$

On calcule maintenant l'estimateur du maximum de vraisemblance $\hat{\theta}_n$: On a :

$$\begin{aligned}
\mathrm{L}\,(x_1, x_2, \ldots, x_n; \theta) &= \prod_1^n P(X_i = x_i; \theta) \\
&= \prod_1^n (x_i * \frac{n_0}{\theta} + (1 - x_i).(1 - \frac{n_0}{\theta})) \\
&= (\frac{n_0}{\theta})^s.(1 - \frac{n_0}{\theta})^{n-s} ln(\mathrm{L}\,(x_1, x_2, \ldots, x_n; \theta)) \\
&= -n.ln(\theta) + s.ln(n_0) + (n - s).ln(\theta - n0) \\
&= h(\theta) \\
\frac{d}{d\theta} h(\theta) &= -\frac{n}{\theta} + \frac{n - s}{\theta - n_0}
\end{aligned}$$

et puis :

$$\begin{aligned}\frac{d}{d\theta}h(\hat{\theta}_n) = 0 \Rightarrow \frac{n}{\hat{\theta}_n} &= \frac{n-s}{\hat{\theta}_n - n_0} \\ \Rightarrow \frac{n}{\hat{\theta}_n} &= \frac{s}{n_0} \\ \Rightarrow \hat{\theta}_n &= \frac{n.n_0}{s}\end{aligned}$$

Les deux estimateurs sont donc bien confondus.

Pour l'exemple précédent, on ajoute :

```
theta_n <- n0/m
sprintf("l'estimateur de theta pour cette simulation
est: %f", theta_n)
```

ce qui nous rend :

```
>[1] "l'estimateur de theta pour cette simulation
est: 980,392157"
```

1.4

D'après le cours, un intervalle de confiance exact pour la proportion p de seuil α est :

$$\left[\frac{1}{1+\frac{n-s+1}{s}.f_{2(n-s+1),2s,\alpha/2}}, \frac{1}{1+\frac{n-s}{s+1}.f_{2(n-s),2(s+1),1-\alpha/2}}\right]$$

On peut en deduire un intervalle de confiance exact pour $\theta = \frac{n_0}{p}$:

$$\left[n_0.(1+\frac{n-s}{s+1}.f_{2(n-s),2(s+1),1-\alpha/2}), n_0.(1+\frac{n-s+1}{s}.f_{2(n-s+1),2s,\alpha/2})\right]$$

L'intervalle suivant est un intervalle asymptotique de seuil α pour $p = \frac{n_0}{\theta}$:

$$\left[\widehat{p_n} - u_\alpha \sqrt{\frac{\widehat{p_n}(1-\widehat{p_n})}{n}}; \widehat{p_n} + u_\alpha \sqrt{\frac{\widehat{p_n}(1-\widehat{p_n})}{n}}\right]$$

où $\hat{p}_n = \frac{s}{n}$.
On en déduit un intervalle asymptotique de seuil α our θ :

$$\left[\frac{n_0}{\widehat{p_n}+u_\alpha \sqrt{\frac{\widehat{p_n}(1-\widehat{p_n})}{n}}}; \frac{n_0}{\widehat{p_n}-u_\alpha \sqrt{\frac{\widehat{p_n}(1-\widehat{p_n})}{n}}}\right]$$

Simulation

Pour $n_0 = 50$, $\theta = 1000$, et $n = 1000$, la simulation renvoie les résultats suivants :

```
Intervalles exacts:
[1] Seuil 0.010000 : [908.440119; 1141.691996]
[2] Seuil 0.050000 : [932.448446; 1110.057727]
[3] Seuil 0.100000 : [945.070982; 1094.331630]
[4] Seuil 0.200000 : [959.917317; 1076.574387]

Intervalles asymptotiques:
[1] Seuil 0.010000 : [912.889494; 1146.030452]
[2] Seuil 0.050000 : [935.644232; 1112.077761]
[3] Seuil 0.100000 : [947.731154; 1095.472096]
[4] Seuil 0.200000 : [962.060037; 1076.931890]
```

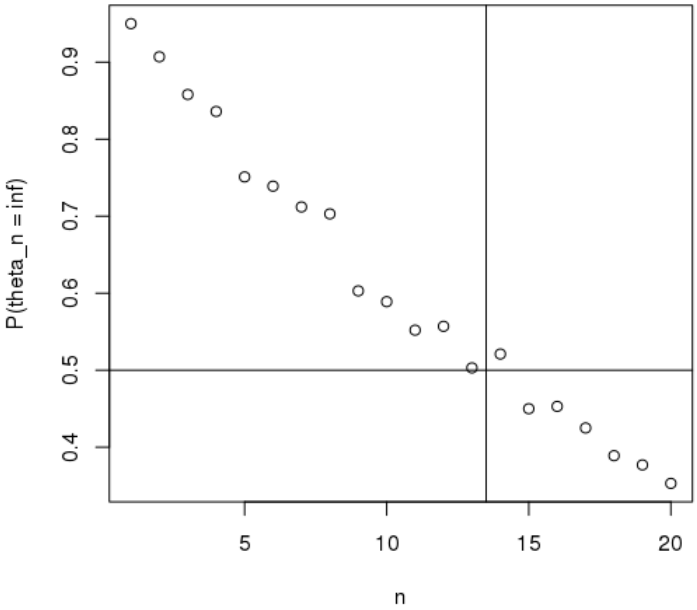
1.5

$$\begin{aligned} P(\hat{\theta} = +\infty) &= P(\bar{X}_n = 0) \\ &= P(\sum_{i=1}^{n} X_i = 0) \\ &= \prod_{i=1}^{n} P(X_i = 0) \\ &= \left(1 - \frac{n_0}{\theta}\right)^n \\ &= (0.95)^n \end{aligned}$$

1.6

$$\begin{aligned} P(\hat{\theta}_n > \frac{1}{2}) &\iff (0.95)^n > \frac{1}{2} \\ &\iff n \ln(0.95) > \ln(0.5) \\ &\iff n < \frac{ln(0.5)}{ln(0.95)} \simeq 13.5 \end{aligned}$$

On a donc $P(\hat{\theta}_n = +\infty) > \frac{1}{2} \iff n \in [\![1; 13]\!]$.


P(theta_n = inf)
0.4
0.5
0.6
0.7
0.8
0.9
5
10
15
20
n

2 Deuxième stratégie

2.1

A chaque fois que l'on pêche un poisson, on le remet à l'eau donc on revient à l'état initila : les Y_i sont donc des variables indépendates de même loi.
Soient $i, k \in \mathbb{N}$

$$\begin{aligned} P(Y_i = k) &= P((X_i = 0) \cap (X_{i+1} = 0) \cap \ldots \cap P(X_{i+1} = 0) \cap P(X_i = 1)) \\ &= \frac{n_0}{\theta}\left(1 - \frac{n_0}{\theta}\right)^{k-1} \end{aligned}$$

$$\forall i \in [\![1; m]\!], \boxed{Y_i \sim \mathcal{G}\left(\frac{n_0}{\theta}\right)}.$$

Simulation

La simulation renvoie :

```
"la moyenne empirique est: 19.380000"
"la vraie moyenne est: 20.000000"
"la variance empirique est: 378.965253"
"la vraie variance est: 380.000000"
```

2.2

Soit N la variable aléatoire du nombre de poissons pêchés.
$N = \sum_{i=1}^{m} Y_i$
N suit une loi de Pascal $\mathcal{P}(m, \frac{n_0}{\theta})$:
$P(N = n) = \binom{n-1}{m-1}(\frac{n_0}{\theta})^m(1 - \frac{n_0}{\theta})^{n-m}$
La simulation donne $n = 1942$.

2.3

$$E[N] = \frac{m\theta}{n_0}$$
$$\theta = \frac{n_0 E[N]}{m}$$

L'estimateur des moments de θ est donc

$$\theta'_m = \frac{n_0}{m} \sum_{i=1}^{m} Y_i.$$
$$= \frac{n_0}{m} N$$

2.4

Estimateur des moments :
$E[N] = \frac{m\theta}{n_0}$
$\theta = \frac{n_0 E[N]}{m}$
L'estimateur des moments de θ est donc :

$$\tilde{\theta}'_m = \frac{n_0}{m} \sum_{i=1}^{m} Y_i.$$
$$= \frac{n_0}{m} N$$

Estimateur du maximum de vraisemblance :

$$\begin{aligned}
\mathcal{L}(\theta, y_1, \ldots, y_m) &= \prod_{i=1}^{m} P(Y_i = y_i) \\
&= \prod_{i=1}^{m} \left(\frac{n_0}{\theta}\right) \left(1 - \frac{n_0}{\theta}\right)^{y_i - 1} \\
&= \left(\frac{n_0}{\theta}\right)^{m} \left(1 - \frac{n_0}{\theta}\right)^{\sum_{i=1}^{m} y_i - 1} \\
&= \left(\frac{n_0}{\theta}\right)^{m} \left(1 - \frac{n_0}{\theta}\right)^{N} \\
\ln(\mathcal{L}(\theta, y_1, \ldots, y_m)) &= m \ln n_0 - m \ln \theta + N \ln(\theta - n_0) - N \ln \theta \\
\frac{\partial}{\partial \theta} \ln(\mathcal{L}(\theta, y_1, \ldots, y_m)) &= \frac{N - m}{\theta - n_0} - \frac{N}{\theta} \\
\frac{\partial}{\partial \theta} \ln(\mathcal{L}(\theta, y_1, \ldots, y_m)) &= 0 \iff \theta = \frac{n_0}{m} N
\end{aligned}$$

D'où : $\hat{\theta}'_m = \frac{n_0}{m} N$.

$$VarN = \left(\frac{\theta}{n_0}\right)^2 m\left(1-\frac{n_0}{\theta}\right)$$

$$\begin{aligned}
\mathcal{I}_n(\theta) &= Var\left[\frac{\partial}{\partial\theta}\ln(\mathcal{L}(\theta, Y_1, \dots Y_n))\right] \\
&= Var\left[\frac{N-m}{\theta-n_0} - \frac{m}{\theta-n_0}\right] \\
&= Var\left[N\left(\frac{1}{\theta-n_0} - \frac{1}{\theta}\right) - \frac{m}{\theta-n_0}\right] \\
&= \left(\frac{1}{\theta-n_0} - \frac{1}{\theta}\right)^2 VarN \\
&= \left(\frac{1}{(\theta-n_0)}\right)^2 m\left(1-\frac{n_0}{\theta}\right) \\
&= \boxed{\frac{m}{\theta(\theta-n_0)}}
\end{aligned}$$

$$\begin{aligned}
E[\hat{\theta}'_m - \theta] &= \frac{n_0}{m}E[N] - \theta \\
&= \frac{n_0}{m}\frac{m\theta}{n_0} - \theta \\
&= 0
\end{aligned}$$

L'estimateur est bien sans biais.

$$\begin{aligned}
Var(\hat{\theta}'_n) &= \left(\frac{n_0}{m}\right)^2 m\frac{1-\frac{n_0}{\theta}}{\frac{n_0^2}{\theta^2}} \\
&= \frac{\theta(\theta-n_0)}{m} \\
&= \frac{1}{\mathcal{I}_m(\theta)}
\end{aligned}$$

L'estimateur est bien de variance minimale.

Simulation

On obtient les résultats suivants :

```
"l'intervalle de confiance de seuil 0.010000
    est [882.990837, 1265.009163]."
"l'intervalle de confiance de seuil 0.050000
    est [928.659981, 1219.340019]."
"l'intervalle de confiance de seuil 0.100000
    est [952.026811, 1195.973189]."
"l'intervalle de confiance de seuil 0.200000
    est [978.967270, 1169.032730]."
```

3 Application et comparaison des stratégies

En applicant la première méthode, la simulation renvoie les résultats suivants :

```
"l'estimation de theta par la premiere strategie
est 2857.142857"
"l'intervalle de confiance exact par la premiere strategie
est [2068.578267, 4082.044541]"
"l'intervalle de confiance asymptotique
par la deuxieme strategie est[2155.610124, 4235.597163]"
```

En applicant la deuxièm stratégie, on obtient les résultats suivants :

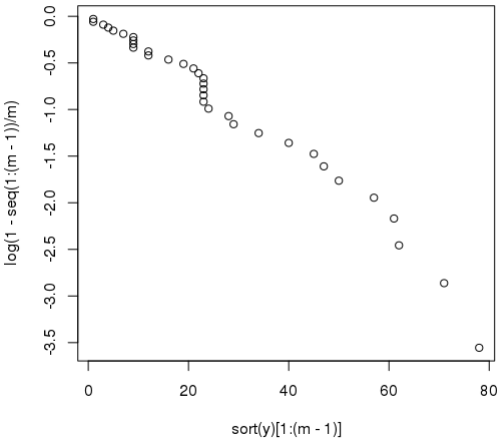
```
"l'estimateur de la deuxieme strategie donne: 2840"
"l'intervalle de confiance asymptotique
par la deuxieme strategie est [2155.610124, 4235.597163]"
```

3.1

La fonction de répartition de la loi géométrique vérifie :

$$F(x) = 1 - (1-p)^{\lfloor x \rfloor}$$
$$\lfloor x \rfloor \ln(1-p) = \ln(1 - F(x))$$

On trace donc le graphe des $\ln(1 - \frac{i}{n})$ en fonction de y_i^*.
On obtient le graphe suivant :



Les points sont à peu près alignés : l'hypothèse de loi géométrique semble donc pertinente.

La largeur de l'intervalle asymptotique vaut 2080 pour la première stratégie et 1849 pour la deuxième. On en déduit que

la deuxième stratégie est meilleure pour estimer θ.

4 Vérifications expérimentales à base de simulations

4.1

```
n <- 1000
theta <- 2000
n0 <- 200
m <- 40
alpha <- .2
```

On obtient les résultats suivants :

```
1 - alpha =  0.8
La proportion des intervalles exacts contenant theta est : 0.825
La proportion des intervalles asymptotiques contenant theta
est : 0.8
```

La valeur de θ translate les intervalles. Plus n est grand, plus l'estimation de θ obtenue est précise. Plus n est grand, plus la proportion d'intervalles contenant la vraie valeur de θ est proche de $1 - \alpha$. Enfin, plus α est petit, plus l'intervalle est large.

4.2

La simulation renvoie les résultats suivants :

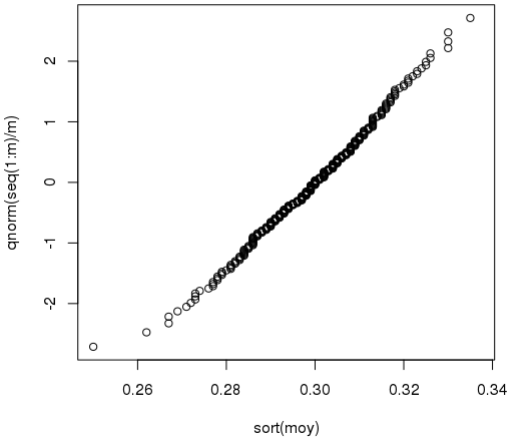
```
"pour epsilon = 0.010000 la proportion de fois ou la moyenne
empirique s'eloigne de l'esperance est : 0.530000 (n = 1000)"
"pour epsilon = 0.010000 la proportion de fois ou la moyenne
empirique s'eloigne de l'esperance est : 0.026667 (n = 10000)"
"pour epsilon = 0.010000 la proportion de fois ou la moyenne
empirique s'eloigne de l'esperance est : 0.000000 (n = 100000)
"
```

```
"pour epsilon = 0.100000 la proportion de fois ou la moyenne
empirique s'eloigne de l'esperance est : 0.000000 (n = 1000)"
"pour epsilon = 0.050000 la proportion de fois ou la moyenne
empirique s'eloigne de l'esperance est : 0.000000 (n = 1000)"
"pour epsilon = 0.020000 la proportion de fois ou la moyenne
empirique s'eloigne de l'esperance est : 0.130000 (n = 1000)"
"pour epsilon = 0.005000 la proportion de fois ou la moyenne
empirique s'eloigne de l'esperance est : 0.790000 (n = 1000)"
```

On observe que quand n augmente, la proportion de fois où la moyenne empirique s'éloigne de l'espérance diminue.

4.3

Plus n est grand, plus l'allure de l'histogramme s'approche de celle d'une loi normale et plus le graphe de probabilité semble linéaire.



histogramme de la simulation