R un outil de simulation

J.Lejeune

Université de Caen, Département de mathématiques et mécanique
lejeune@math.unicaen.fr

21 novembre 2007

PROBLEME BAC 2006

Stand de tir.
Probabilité de crever le ballon avec un tir = p. Ici $p = 0.2$.
Les tirs successifs sont supposés indépendants.

PROBLEME BAC 2006

Stand de tir.
Probabilité de crever le ballon avec un tir = p. Ici $p = 0.2$.
Les tirs successifs sont supposés indépendants.

- Lancement d'un dé tétraédrique régulier numéroté de 1 à 4.
  Résultat = k

PROBLEME BAC 2006

Stand de tir.
Probabilité de crever le ballon avec un tir = p. Ici $p = 0.2$.
Les tirs successifs sont supposés indépendants.

- Lancement d'un dé tétraédrique régulier numéroté de 1 à 4. Résultat = k
- Le tireur a droit au plus à k tirs pour crever le ballon.

PROBLEME BAC 2006

Stand de tir.
Probabilité de crever le ballon avec un tir = p. Ici $p = 0.2$.
Les tirs successifs sont supposés indépendants.

- Lancement d'un dé tétraédrique régulier numéroté de 1 à 4. Résultat = k
- Le tireur a droit au plus à k tirs pour crever le ballon.
- Probabilité de crever le ballon ?

Solution

$$\text{Probabilité} = 1 - \frac{(1-p)(1-(1-p)^4)}{4p}$$

Pour $p = 0.20$, probabilité de crever le ballon $= 0.4096$

Solution

$$\text{Probabilité} = 1 - \frac{(1-p)(1-(1-p)^4)}{4p}$$

Pour $p = 0.20$, probabilité de crever le ballon $= 0.4096$

On va simuler n fois cette expérience.
Succès = on crève le ballon.
La fréquence observée de succès devrait être proche de la probabilité.

Utilisation d'une fonction tirage1 (fichier "tirage.txt")

Utilisation d'une fonction tirage1 (fichier "tirage.txt")

3 arguments pour tirage1 :

- p = probabilité de crever le ballon en 1 coup
- l = nombre de résultats affichés sur la fenêtre de commande de R.
  On affiche les l premières expériences ($(l \leq n)$ avec pour chaque expérience :
  k = tirage du dé tétraédrique régulier puis une suite de 0 et/ou 1 de longueur $\leq k$.
- n = nombre d'expériences à réaliser

Résultat :

- les l premières expériences
- prop = fréquence de succès pour n expériences et proba= probabilité de crever le ballon en 1 coup

Utilisation d'une fonction tirage1 (fichier "tirage.txt")

3 arguments pour tirage1 :

- p = probabilité de crever le ballon en 1 coup
- l = nombre de résultats affichés sur la fenêtre de commande de R.
  On affiche les l premières expériences ($(l \leq n)$ avec pour chaque expérience :
  k = tirage du dé tétraédrique régulier puis une suite de 0 et/ou 1 de longueur $\leq k$.
- n = nombre d'expériences à réaliser

Résultat :

- les l premières expériences
- prop = fréquence de succès pour n expériences et proba= probabilité de crever le ballon en 1 coup

Pour $n = 10000, l = 10$ et $p = 0.2$ faire tirage1(0.2,10,100000)

Exemple d'une sortie de logiciel :

```
>tirage1(0.2,10,100000)
 4        0  1
 2        0  1
 4        0  1
 3        0  0  1
 2        0  0
 4        1
 4        0  1
 3        1
 3        0  1
 3        0  0  0
$ prop
[1] 0.4091
$ proba
[1] 0.4096
```

Une autre solution

On remarque que si X est une variable aléatoire qui indique le nombre de fois qu'il a fallu tirer dans le ballon avant qu'il ne crève alors X suit une loi géométrique modifiée c.a.d.
$P(X = k) = (1 - p)^{k-1}p$ pour $k = 1, 2, ...$
Cette loi est programmée dans R.

Une autre solution

On remarque que si X est une variable aléatoire qui indique le nombre de fois qu'il a fallu tirer dans le ballon avant qu'il ne crève alors X suit une loi géométrique modifiée c.a.d.
$P(X = k) = (1 - p)^{k-1}p$ pour $k = 1, 2, ...$
Cette loi est programmée dans R.

Si Y est la variable qui, à un lancer de dé associe k, $k = 1, 2, 3, 4$, alors :
Succès si $X \leq Y$

D'où une fonction plus simple tirage2 (dans fichier "tirage.txt")

Suite du problème

On lance 200 fois le dé

Face k	1	2	3	4
Nombre de sorties de face k	58	49	52	41

Suite du problème

On lance 200 fois le dé

Face k	1	2	3	4
Nombre de sorties de face k	58	49	52	41

f_k fréquence de sortie de k

On demandait de calculer : $d^2 = \sum_{k=1}^{4} \left(f_k - \frac{1}{4} \right)^2$.

Si le dé est pipé, $f_k \neq \frac{1}{4}$ et d^2 sera "grand".

Avec les résultats du tableau, $d^2 =$

Suite du problème

On lance 200 fois le dé

Face k	1	2	3	4
Nombre de sorties de face k	58	49	52	41

f_k fréquence de sortie de k

On demandait de calculer : $d^2 = \sum_{k=1}^{4} \left(f_k - \frac{1}{4} \right)^2$.

Si le dé est pipé, $f_k \neq \frac{1}{4}$ et d^2 sera "grand".
Avec les résultats du tableau, d^2 =0.25375
Peut-on considérer que la valeur observée soit "grande" ?

Attention ! Si le dé n'est pas pipé, on peut aussi avoir d^2 "grand" par le seul fait du hasard.
On pose $d_{0.90}$ le nombre tel que :
$P(d^2 \geq d_{0.90}) = 0.10$. $d_{0.90}$ est le neuvième décile de la série des résultats quand le dé n'est pas pipé.

Attention ! Si le dé n'est pas pipé, on peut aussi avoir d^2 "grand" par le seul fait du hasard.
On pose $d_{0.90}$ le nombre tel que :
$P(d^2 \geq d_{0.90}) = 0.10$. $d_{0.90}$ est le neuvième décile de la série des résultats quand le dé n'est pas pipé.
On décidera (à tort !) que le dé est pipé si $d^2 \geq d_{0.90}$.

Avec cette régle de décision, la probabilité de se tromper en décidant que le dé est pipé est donc de 10%.

Recherche expérimentale de $d_{0.90}$

Le rédacteur du problème a effectué 1000 simulations des 200 lancers d'un dé tétraédrique bien équilibré et on calcule pour chaque simulation le nombre d^2. L'énoncé du sujet donnait le minimum, le premier décile, le premier quartile, la médiane, le troisième quartile, le neuvième décile et le maximum :

Min	D_1	Q_1	Médiane	Q_3	D_9	Max
0.00124	0.00192	0.00235	0.00281	0.00345	0.00452	0.01015

Recherche expérimentale de $d_{0.90}$

Le rédacteur du problème a effectué 1000 simulations des 200 lancers d'un dé tétraédrique bien équilibré et on calcule pour chaque simulation le nombre d^2. L'énoncé du sujet donnait le minimum, le premier décile, le premier quartile, la médiane, le troisième quartile, le neuvième décile et le maximum :

Min	D_1	Q_1	Médiane	Q_3	D_9	Max
0.00124	0.00192	0.00235	0.00281	0.00345	0.00452	0.01015

Construction d'une fonction tirage3 (dans fichier "tirage.txt") qui effectue la même opération et qui calcule les mêmes paramètres (sauf min et max).

Exemples de 3 sorties de logiciel :

```
> tirage3(1000, c(0.1,0.25,0.50,0.75,0.9))
     10%         25%         50%         75%         90%
 0.0007500   0.0013875   0.0030000   0.0051500   0.0078500
> tirage3(1000, c(0.1,0.25,0.50,0.75,0.9))
   10%       25%       50%       75%       90%
 0.00080   0.00155   0.00295   0.00515   0.00786
> tirage3(1000, c(0.1,0.25,0.50,0.75,0.9))
    10%         25%         50%         75%         90%
 0.000750   0.001550   0.002875   0.005000   0.007750
```

Chaque simulation donne un résultat différent. Normal !

Mais les résultats sont éloignés de ceux du tableau. Est-ce seulement dû au hasard ?

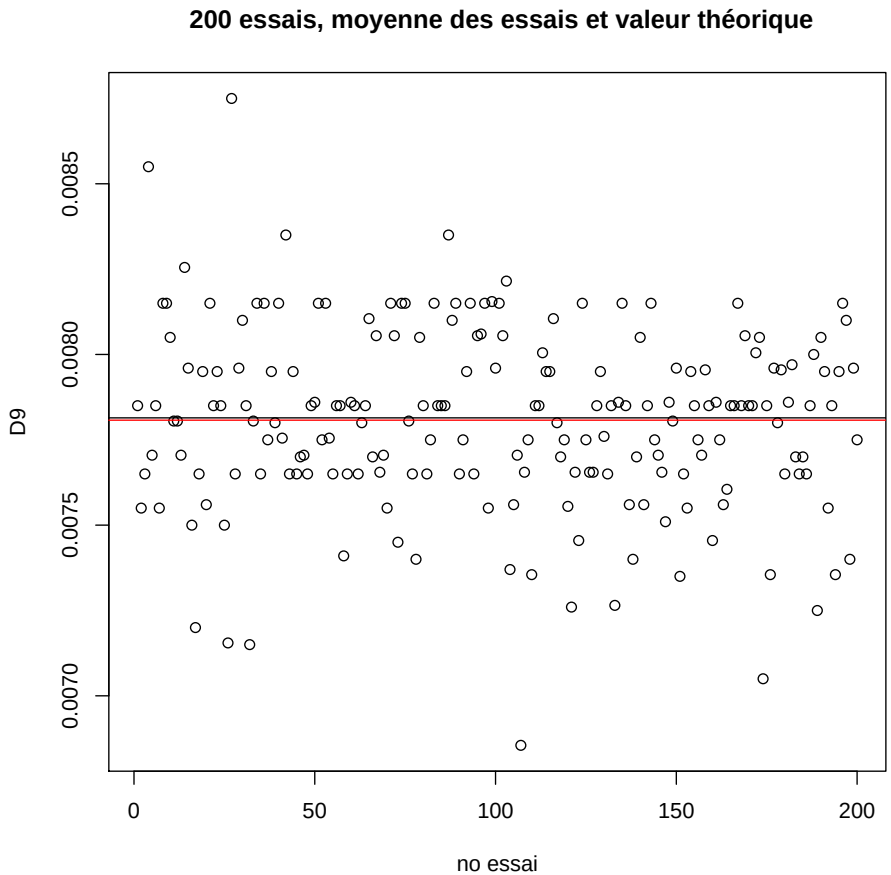
En "moyennant " les résultats issus de "beaucoup" de simulations, obtient-on les résultats du tableau ?
On s'intéresse plus spécialement au neuvième décile $D_9 = d_{0.90}$.

En "moyennant " les résultats issus de "beaucoup" de simulations, obtient-on les résultats du tableau ?
On s'intéresse plus spécialement au neuvième décile $D_9 = d_{0.90}$.
Construction d'une fonction tirage4 (dans fichier "tirage.txt"). On effectue 200 fois les 1000 simulations et à chaque fois on calcule D_9 (en ordonnée sur le graphique).

On calcule la moyenne des D_9.
Droite en rouge horizontale d'ordonnée : moyenne des 200 résultats aléatoires D_9.

La théorie permet de calculer la valeur de $d_{0.90} = 0.007814236$.
Droite en noir horizontale d'ordonnée : 0.007814236.

Un exemple de graphique possible :



La moyenne des D_9 est "proche" de la valeur théorique, et très éloignée de 0.00452. Il doit y avoir un bug dans l'énoncé ! !

Intervalles de confiance d'une proportion

Population de N personnes.

K votent pour X
La proportion réelle de ceux qui votent X est donc $p = \frac{K}{N}$

Intervalles de confiance d'une proportion

Population de N personnes.

K votent pour X
La proportion réelle de ceux qui votent X est donc $p = \frac{K}{N}$ On tire un échantillon de taille n
Parmi ces n personnes tirées au sort, k votent pour X donc une proportion observée de $\widehat{p} = \frac{k}{n}$

Intervalles de confiance d'une proportion

Population de N personnes.

K votent pour X
La proportion réelle de ceux qui votent X est donc $p = \frac{K}{N}$ On tire un échantillon de taille n
Parmi ces n personnes tirées au sort, k votent pour X donc une proportion observée de $\widehat{p} = \frac{k}{n}$

A partir de cette proportion observée, calcul d'un intervalle de confiance :

$[\widehat{p}^-, \widehat{p}^+]$, de niveau $1 - \alpha$, c.a.d.
$P(p \in [\widehat{p}^-, \widehat{p}^+]) = 1 - \alpha$

Ici $\widehat{p}^- = \widehat{p} - e$ et $\widehat{p}^+ = \widehat{p} + e$ avec : $e = z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{N-n}{N-1}} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$
où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale réduite centrée.

Expérimentation avec $\alpha = 0.05$:

Un institut de sondage effectue le tirage de 1000 personnes dans une population de 10000 personnes dont 4000 votent X c.a.d. $p = 0.40$. Il donne une "fourchette" de confiance de 95% pour p.

On simule le résultat de 200 instituts de sondages avec la fonction sondage (dans le fichier tirage.txt").

Expérimentation avec $\alpha = 0.05$:

Un institut de sondage effectue le tirage de 1000 personnes dans une population de 10000 personnes dont 4000 votent X c.a.d. $p = 0.40$. Il donne une "fourchette" de confiance de 95% pour p.

On simule le résultat de 200 instituts de sondages avec la fonction sondage (dans le fichier tirage.txt").
On doit avoir "approximativement" 10 instituts qui se trompent sur les 200 c.a.d. tels que la fourchette de chevauche pas la "vraie valeur" $p = 0.40$. On affiche les 5 premieres "fourchettes".

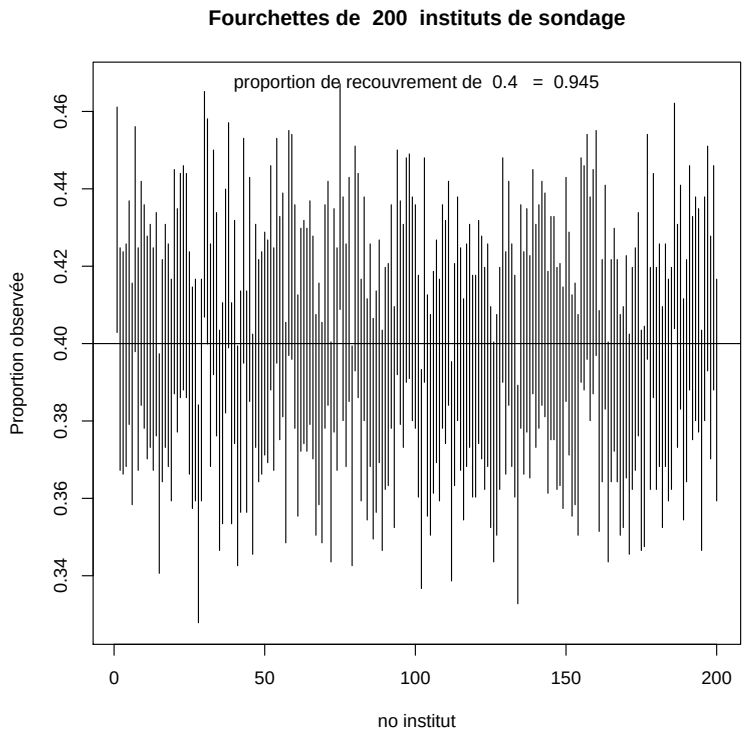
Expérimentation avec $\alpha = 0.05$:

Un institut de sondage effectue le tirage de 1000 personnes dans une population de 10000 personnes dont 4000 votent X c.a.d. $p = 0.40$. Il donne une "fourchette" de confiance de 95% pour p.

On simule le résultat de 200 instituts de sondages avec la fonction sondage (dans le fichier tirage.txt").
On doit avoir "approximativement" 10 instituts qui se trompent sur les 200 c.a.d. tels que la fourchette de chevauche pas la "vraie valeur" $p = 0.40$. On affiche les 5 premieres "fourchettes".
Exemple de sortie de logiciel :

$>$ sondage(10000,4000,1000,200,5,0.05)
0.4029 0.4611
0.3672 0.4248
0.3663 0.4237
0.3682 0.4258
0.3791 0.4369

Un exemple de graphique possible :



Commentaire : le premier institut se trompe mais la proportion de ceux qui ne se trompent pas $(94,5\%)$ est proche de celle attendue par la théorie (95%). L'expérimentation confirme la théorie.

De l'utilité des graphiques

Soit un jeu de données.

$x1$	$x2$	$x3$	$x4$	$y1$	$y2$	$y3$	$y4$
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.10	5.39	12.50
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89

On va calculer la droite de régression estimée de y1 sur x1, puis de y2 sur x2,etc

On va calculer la droite de régression estimée de y1 sur x1, puis de y2 sur x2,etc

Pour les 4 jeux, même résultat :

$y = 3 + 0.5\ x$

On va calculer la droite de régression estimée de y1 sur x1, puis de y2 sur x2,etc

Pour les 4 jeux, même résultat :

$y = 3 + 0.5\ x$

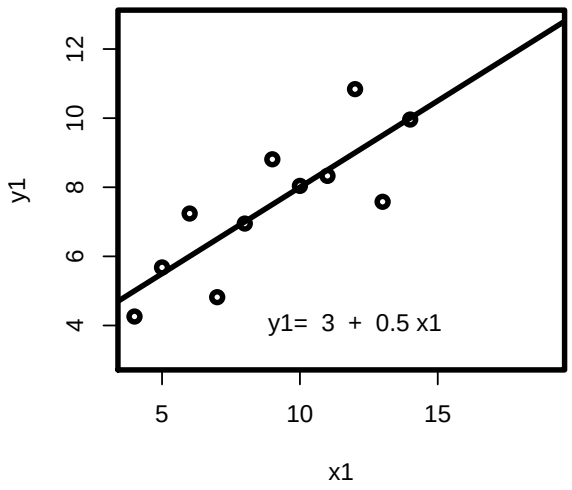
Coefficients de corrélation r très voisins :
0.816, , 0.816, , 0.816, , 0.817
Ces droites ajustent-elles correctement le nuage de points ?

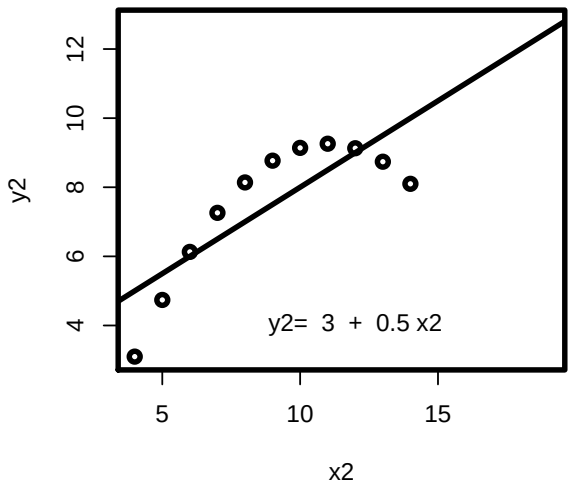
On va calculer la droite de régression estimée de y1 sur x1, puis de y2 sur x2,etc

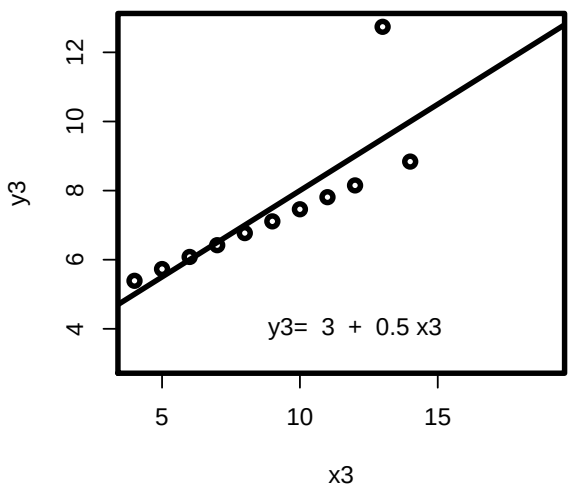
Pour les 4 jeux, même résultat :

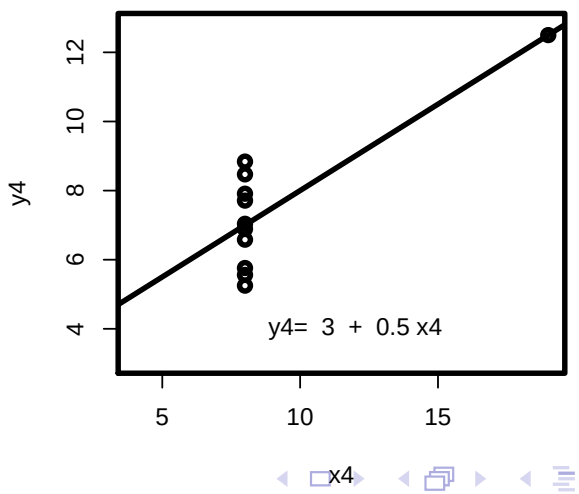
$y = 3 + 0.5\ x$

Coefficients de corrélation r très voisins :
0.816, , 0.816, , 0.816, , 0.817
Ces droites ajustent-elles correctement le nuage de points ?
Graphiques :


y1
x1
y1= 3 + 0.5 x1
4
6
8
10
12
5
10
15


y2
x2
y2= 3 + 0.5 x2
4
6
8
10
12
5
10
15


y3
x3
y3= 3 + 0.5 x3
4
6
8
10
12
5
10
15


y4
x4
y4= 3 + 0.5 x4
4
6
8
10
12
5
10
15