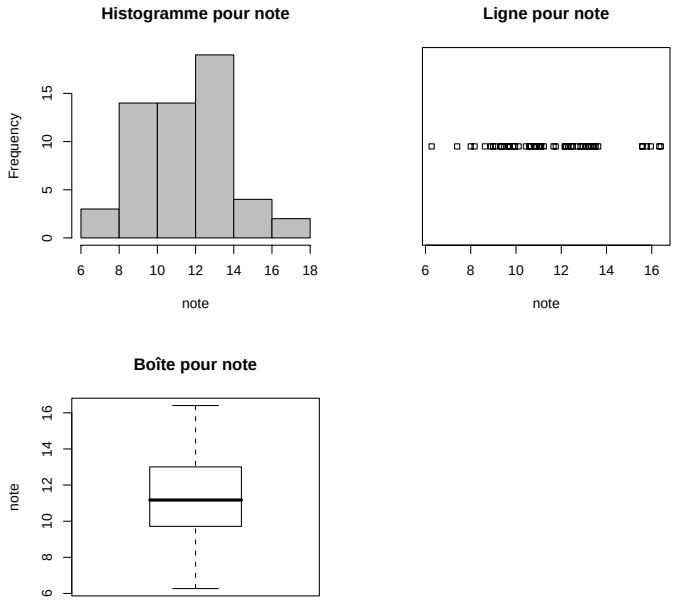
Université Claude Bernard Lyon 1

M1 PPMR (Semestre 1) statistique
Session 1 8 Décembre 2008

Corrigé de l'examen CCF2 de statistique

Correction de l'exercice 1.

(1) On étudie la variable quantitative (ou numérique) 'note'. On procèdera donc comme dans le chapitre 3 du document de cours.



Voir les trois graphiques ci-dessus pour la variable 'note'. Le nombre de données étant important (ici , la ligne de point n'est pas très pertinente.

(2) Les différents résultats déterminés par R sont donnés dans le tableau suivant

noms	valeurs
moyenne	11.508065
sd	2.294819
Q_1 (quartile à 25 %)	9.71661
médiane	11.1719
Q_3 (quartile à 75 %)	12.978262
minimum	6.270278
maximum	16.402698
nombre	56

(3) La taille d'effet relative par rappor à la norme $norm = 10$ est définie dans la section 4.2 page 17 du chapitre 4 par (4.2) :

$$d = \frac{M - norme}{SD},$$

soit ici

$$d = \frac{11.508065 - 10}{2.294819} = 0.657161.$$

D'après l'équation (4.3) page 17 de ce même chapitre, la taille d'effet est moyen. De plus, d est positif. Ainsi, le groupe est supérieur à la norme, de façon moyenne.

Correction de l'exercice 2.

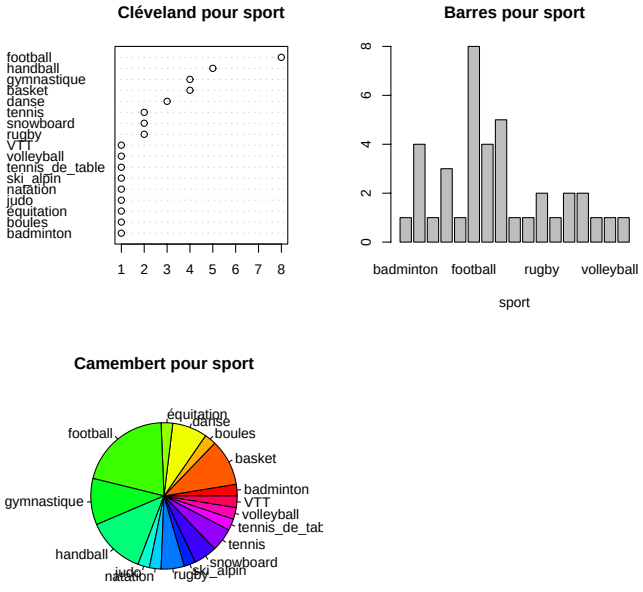
- On étudie la variable qualitative (ou catégorielle) 'sport'. On procèdera donc comme dans le chapitre 5 du document de cours.
- Les effectifs et les pourcentages déterminés par R sont donnés dans le tableau suivant

	effectifs	pourcentages
badminton	1	2.564
basket	4	10.256
boules	1	2.564
danse	3	7.692
équitation	1	2.564
football	8	20.513
gymnastique	4	10.256
handball	5	12.821
judo	1	2.564
natation	1	2.564
rugby	2	5.128
ski_alpin	1	2.564
snowboard	2	5.128
tennis	2	5.128
tennis_de_table	1	2.564
volleyball	1	2.564
VTT	1	2.564

- On peut aussi afficher les effectifs et les pourcentages ordonnés : Les effectifs et les pourcentages déterminés par R sont donnés dans le tableau suivant

	effectifs	pourcentages
badminton	1	2.564
boules	1	2.564
équitation	1	2.564
judo	1	2.564
natation	1	2.564
ski_alpin	1	2.564
tennis_de_table	1	2.564
volleyball	1	2.564
VTT	1	2.564
rugby	2	5.128
snowboard	2	5.128
tennis	2	5.128
danse	3	7.692
basket	4	10.256
gymnastique	4	10.256
handball	5	12.821
football	8	20.513

-



Voir les trois graphiques (ordonnés) ci-dessus pour la variable 'sport'. Ici, le nombre de modalités étant élevés (17), le diagramme en barre et le camembert ne sont pas très lisibles.

Correction de l'exercice 3.

(1)

(a) La loi gouvernant la variable 'nombre' est la loi binomiale, puisqu'on extrait un certain nombre d'électeurs dans une population de "grand" effectif ; chacun des électeur a une probabilité p_0 de succès de voter pour N.S.. Les deux paramètres de la loi binomiale sont donc $\pi = p_0$, inconnu et n, le nombre total d'électeurs ayant répondu pour chaque bureau de vote, ici appelée 'total'.

(b) On utilise la définition 5.50 page 43 du chapitre 5. Pour les bureaux de vote numéro 6, 10 et 14, avec le niveau de confiance NC=0.95, on obtient donc successivement :

$$[0.4524544, 0.6455848] , \tag{1a}$$

$$[0.5118488, 0.6996897] , \tag{1b}$$

$$[0.389586, 0.5895806] . \tag{1c}$$

On constate que le premier intervalle de confiance est le plus grand, le deuxième le plus petit. Seul le deuxième de contient pas le nombre 50 %, c'est donc le plus intéressant. En effet, avec lui, au niveau de confiance choisi, on est sûr que N.S. gagnera, contrairement aux deux autres !

Remarque 1. La formule de la définition 5.50 page 43 donnait comme intervalle de confiance $p \pm z\sqrt{\frac{p(1-p)}{n}}$ où

$$P(-z \leq X \leq z) = NC$$

Pour $NC = 0.95$, on obtenait $z = 1.959964$. Ceux qui ont choisi la valeur approché $z = 2$ devraient avoir trouvé comme intervalle de confiance :

$$[0.4504819, 0.6475574] , \tag{2a}$$

$$[0.5099302, 0.7016082] , \tag{2b}$$

$$[0.3875434, 0.5916233] . \tag{2c}$$

Ces résultats seront comptés justes !

(c) Cette question est très proche de l'exercice 5.59 page 46 du document de cours (chapitre 5) auquel on renvoie. Voir aussi sa correction page 57.

On créant avec Rcmdr les variables
– de nom `SEP` et égale à `sqrt((pr*(1-pr))/total)`
– de nom `prmin` et égale à `pr-1.959964*SEP`
– de nom `prmax` et égale à `pr+1.959964*SEP`
vous devriez obtenir le tableau 1 page 12.

D'après la formule de la définition 5.50 page 43 du document de cours, l'antépenultièle colonne contient la SEP et les deux dernières colonnes créées contiennent la borne inférieure et supérieure des intervalles de confiance au seuil $NC = 0.95$. On retrouve par exemple les intervalles de confiance des bureaux de vote numéros 6, 10 et 14 donnés par (3). On constate grâce à la colonne SEP que l'intervalle de confiance correspondant au bureau de vote numéro 10 (resp. 14) est le plus petit (resp. grand) puisqu'il correspond à la SEP la plus petite (resp. grande).

(d) Ces renseignements ne sont pas intéressants en tant que tels pour l'institut de sondage qui les a collectés. C'est mis ensemble qu'ils apporteront tout leur intérêt. Voir question 2a.

Ces sondages, à la sortie des urnes, sont fondamentaux pour pouvoir annoncer à[1] 20H00 une estimation du nombre de voix.

(2)

(a) Si on enregiste les données avec le nom 'Dataset', les commandes suivantes (dans Rgui ou dans la fenêtre de script de Rcmdr) :

```
sum(Dataset$total)
sum(Dataset$nombre)
sum(Dataset$nombre)/sum(Dataset$total)
```

calculent le nombre total d'élécteurs (sur l'ensemble des bureaux), le nombre total d'électeurs pour N.S. et la proportion d'électeurs pour N.S. (sur l'ensemble des bureaux).

(b) Cela fournit comme résultat (pour la dernière)

$$pr = 0.541308. \tag{3}$$

qui correspond donc à la proportion d'électeurs pour N.S. sur l'ensemble des bureaux de votes.

(c) Grâce à la définition 5.50 page 43 du document de cours, on peut déterminer les intervalles de confiance aux niveaux 0.95, 0.99 et 0.999999 de la proportion d'élécteurs pour N.S. sur l'ensemble des bureaux de vote. La taille de l'échantillon est ici donnée par la première commande de la question 2a, soit

$$n = 2706. \tag{4}$$

On a donc successivement pour les niveaux de confiance 0.95, 0.99 et 0.999999

$$[0.5274952, 0.5551213]\,, \tag{5a}$$

$$[0.5231548, 0.5594617]\,, \tag{5b}$$

$$[0.506834, 0.5757826]\,. \tag{5c}$$

Naturellement, le premier est plus fin que ceux déterminés dans la question 1, puisque la taille de l'échantillon est plus grande.

(d) On donne finalement les résultats finaux officiels :

	nombre
M. Nicolas SARKOZY	18983138
Mme Ségolène ROYAL	16790440

On peut donc connaître *a posteriori* la véritable proportion d'électeurs pour N.S. en tapant dans R

```
18983138/(18983138+16790440)
```

[1] Bien qu'en fait, pour cette élection, certains connaissaient déjà à 13h le résultat final ...

ce qui donne

$$p_0 = 0.530647. \tag{6}$$

On se rend compte *a posteriori* que les trois intervalles de confiance (5) contiennent bien p_0.

(3) *Question facultative*

(a) Si on enregiste les données avec le nom 'Dataset', les commandes suivantes :

```
p0<-18983138/(18983138+16790440)
(Dataset$pmin<=p0)&(Dataset$pmax>=p0)
100*sum((Dataset$pmin<=p0)&(Dataset$pmax>=p0))/50
```

déterminent (comme dans l'exercice 5.60 page 46 du document de cours) le pourcentage de cas où l'intervalle de confiance contient réellement le paramètre p_0 inféré pour la question 1.

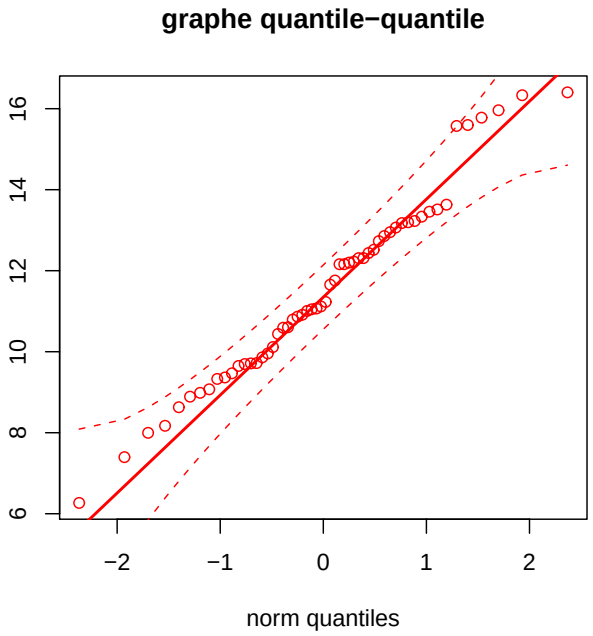
(b) On trouve un pourcentage égal à 0 %, qui est bien proche de niveau de confiance 95 %.

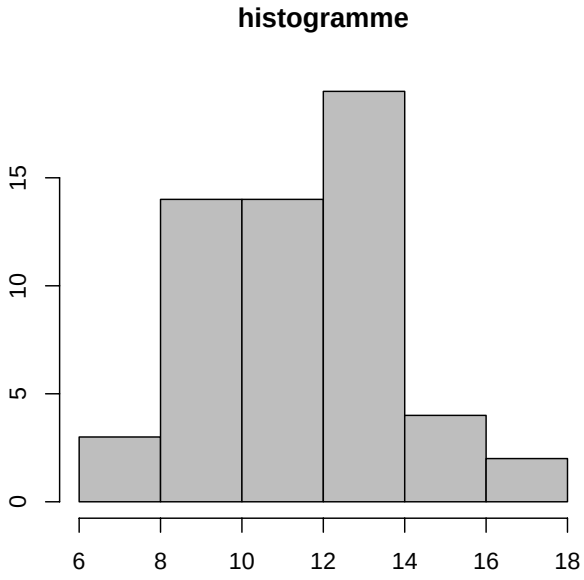
Correction de l'exercice 4.

(1) Comme dans le chapitre 3 du document de cours, on peut tracer un graphe quantile quantile des données du fichier `noteMT40A04.xls`. On pouvait aussi tracer un histogramme.

(2)

On obtient les graphiques de la figure ci-dessus :





Les points s'alignent autour de la droite médiane : on peut donc considérer que ces notes proviennent d'une loi normale.

(3) (a) On renvoie à la section 6.5 page 75 du chapitre 6 page 61. On souhaite tester la moyenne par rapport à la norme $\mu_0 = 10$. On veut montrer que le groupe est "sous" la moyenne pour le "motiver". On teste donc l'hypothèse nulle $H_0 : \mu = \mu_0$ contre $H_1 : \mu < \mu_0$. On choisit le seuil conventionnel $\alpha = 0.05$.

(b) On procède au *test de Student pour un échantillon* (à la moyenne).

On fait l'hypothèse nulle H_0 : $\mu = \mu_0$. avec μ_0 =10. On cherche à montrer que la moyenne de la loi normale, dont proviendraient les données de l'échantillon étudié, est plus petite que μ_0. 0n fait donc l'hypothèse alternative suivante : H_1 : $\mu < \mu_0$.

Grâce à R, on trouve la valeur suivante de la statistique

$$t = \frac{m - \mu_0}{sd/\sqrt{n}} = 4.917739$$

La probabilité critique $P(T \leq t)$ (pour la loi de Student à $ddl = 55$ degrés de libertés) est égale à

$$p_c = 0.999996$$

Puisque p_c est strictement supérieure au niveau de signification 0.05, on accepte l'hypothèse nulle H_0. Ainsi, H_0 est vraie et donc *la moyenne est égale à* $\mu_0 = 10$, au risque 0.05.

Finalement, le groupe a bien la moyenne et il n'est pas nécessaire de le motiver. Cela confirme le résultat de la question 3 de l'exercice 1.

Remarque 2. Attention, on pouvait voir apparaître le message suivant :

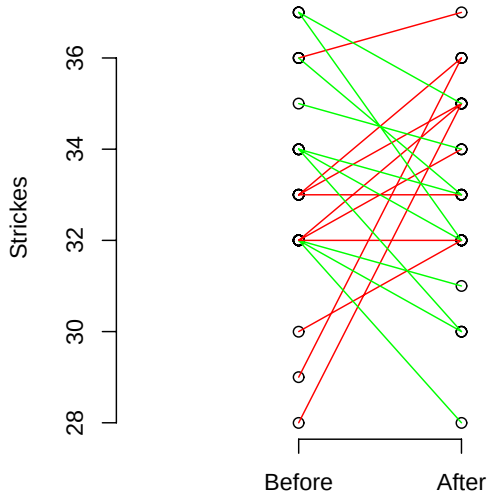
```
        One Sample t-test

data:  data.frame[, 1]
t = 4.9177, df = 55, p-value = 1
alternative hypothesis: true mean is less than 10
95 percent confidence interval:
     -Inf 12.02111
sample estimates:
mean of x
 11.50806
```

et croire que la probabilité critique valait 1 au lieu de 0.999996 ; il s'agit seulement d'un arrondi à l'affichage.

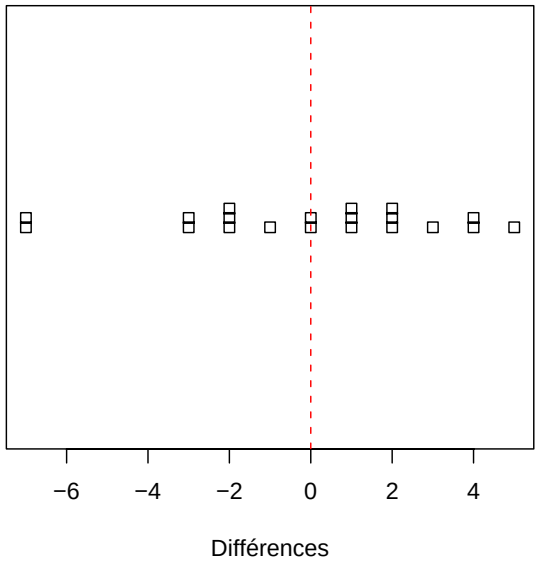
Correction de l'exercice 5.

(1) On renvoie au chapitre 7 page 89 du polycopié de cours.

– Grâce à la fonction `parallelplot.R`, on peut tracer le graphe des différences ci-dessus.



On pouvait aussi afficher une ligne de point



Sur ces graphes, on voit a peu près autant de scores qui ont augmenté que de scores qui ont diminué.

– Confirmons cela par un test d'hypothèse. On veut montrer que le nombre de strickes a augmenté significativement. On choisit le seuil conventionnel $\alpha = 0.05$.
On procède au *test de Student apparié* (à la différence des moyennes).

On fait l'hypothèse nulle H_0 : $\mu_x - \mu_y = 0$. On cherche à montrer que la moyenne de la loi normale, dont proviendraient les différences entre le premier et le second échantillon (qui sont appariés) est strictement négative. 0n fait donc l'hypothèse alternative suivante : $H_{a(2)}$: $\mu_x - \mu_y < 0$
Grâce à R, on trouve la valeur suivante de la statistique

$$t = \frac{m_d}{sd_d/\sqrt{n}} = -0.134583$$

La probabilité critique $P(T \leq t)$ (pour la loi de Student à $ddl = 19$ degrés de libertés) est égale à

$$p_c = 0.447179$$

Puisque p_c est strictement supérieure au niveau de signification 0.05, on accepte l'hypothèse nulle H_0. Ainsi, H_0 est vraie et donc *la moyenne μ_x du premier échantillon est égale à celle du second échantillon μ_y*, au risque 0.05. Ainsi, la moyenne n'a pas significativement bougé et sur ce point, au risque $\alpha = 0.05$, la publicité est mensongère !

(2) Sur l'ensemble des 20 individus, il y a eu après la vision de la vidéo 663 strikes réalisés sur $n = 20 \times 50 = 1000$ essais. La proportion observée de succès est donc $\pi = 0.663$. On cherche à montrer que cette proportion est supérieure à $\pi_0 = 0.6$.

Cela n'était pas traité explicitement dans le cours, mais en appliquant les idées des chapitres 5 et 6 du document de cours avec un peu d'astuce, on pouvait s'en sortir ! On renvoie aussi aux pages 203 et 304 (définition 12.5) de l'ouvrage de S. Chamepely [Cha04].

On choisit un risque $\alpha = 0.05$. Nous allons adapter la définition 6.28 page 78 du document de cours à une variable aléatoire qui n'est pas normale mais qui suit une loi binomiale de paramètres $\pi = 0.663$ et $n = 1000$. Comme dans la section 5.7 page 40, on approchera la loi suivie par la proportion par une loi normale (puisque n est "grand").

On fait donc l'hypthoses H_0 que la proportion de succés (observée) π suit une loi normale de moyenne $\mu = \pi_0$ et d'écart-type $\sqrt{\pi_0(1-\pi_0)/n}$. Ainsi la statistique

$$z = \frac{\pi - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

suit une loi normale centrée réduite. Puisque l'on veut montrer que $\pi > \pi_0 = 0.6$, l'hypothèse alternative choisie est $H_1 : \pi > \pi_0$. Comme dans la définition 6.28 page 78 du document de cours (ou définition 12.5 page 204 de [Cha04]), on calcule la probabilité que $P(Z > z)$, où Z suit une loi normale centrée réduite.

Sous R, on obtient successivement

$$z = \frac{\pi - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.663 - 0.6}{\sqrt{\frac{0.6\times(1-0.6)}{1000}}} = 4.066633,$$

puis

$$p_c = 2.3849e - 05.$$

Enfin, comme dans la définition 6.28 page 78 du document de cours, on constate que $p_c \leq \alpha = 0.05$; ansi, on rejettera H_0 donc on acceptera l'hypothèse alternative, soit $\pi > \pi_0 = 0.6$. Donc, sur ce point la publicité a raison : les joueurs ont dépassé de façon significative le seuil de 60% de strikes.

Cependant, d'après la conclusion de la question 1, le progrès n'est pas significatif !

Remarque 3. Si on calcule la proportion de succés avant la vidéo, on observe

```
sum(STRICKES$Before)
```

soit une proportion observée de

$$\pi = \frac{661}{n} = 0.661.$$

Comme précédemment, on calcule successivement

$$z = \frac{\pi - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.661 - 0.6}{\sqrt{\frac{0.6\times(1-0.6)}{1000}}} = 3.937533,$$

puis

$$p_c = 4.1162e - 05.$$

Enfin, comme dans la définition 6.28 page 78 du document de cours, on constate que $p_c \leq \alpha = 0.05$; ansi, on rejettera H_0 donc on acceptera l'hypothèse alternative, soit $\pi > \pi_0 = 0.6$. Donc, avant vision de la vidéo, les joueurs ont aussi dépassé de façon significative le seuil de 60% de strikes, ce qui corrobore les résultats des deux questions précédente !
Cette pub est donc bien mensongère !

Références

[Cha04] Stéphane Champely. *Statistique vraiment appliquée au sport*. de Boeck, 2004. disponible à la BU de Lyon I sous la cote 519.5 CHA.

	total	nombre	pr	SEP	prmin	prmax
1	99	58	0.5858586	0.0495055	0.4888297	0.6828875
2	101	51	0.5049505	0.0497494	0.4074434	0.6024576
3	98	56	0.5714286	0.0499896	0.4734508	0.6694064
4	100	50	0.5000000	0.0500000	0.4020018	0.5979982
5	99	53	0.5353535	0.0501261	0.4371082	0.6335989
6	102	56	0.5490196	0.0492689	0.4524544	0.6455848
7	98	53	0.5408163	0.0503391	0.4421536	0.6394791
8	99	55	0.5555556	0.0499407	0.4576735	0.6534376
9	101	55	0.5445545	0.0495539	0.4474305	0.6416784
10	104	63	0.6057692	0.0479195	0.5118488	0.6996897
11	100	52	0.5200000	0.0499600	0.4220802	0.6179198
12	98	57	0.5816327	0.0498299	0.4839678	0.6792975
13	101	57	0.5643564	0.0493380	0.4676557	0.6610572
14	96	47	0.4895833	0.0510200	0.3895860	0.5895806
15	96	44	0.4583333	0.0508535	0.3586622	0.5580044
16	101	51	0.5049505	0.0497494	0.4074434	0.6024576
17	104	48	0.4615385	0.0488838	0.3657280	0.5573489
18	99	63	0.6363636	0.0483469	0.5416054	0.7311218
19	99	52	0.5252525	0.0501878	0.4268863	0.6236187
20	99	55	0.5555556	0.0499407	0.4576735	0.6534376
21	99	60	0.6060606	0.0491083	0.5098101	0.7023112
22	98	53	0.5408163	0.0503391	0.4421536	0.6394791
23	100	53	0.5300000	0.0499099	0.4321784	0.6278216
24	101	62	0.6138614	0.0484447	0.5189116	0.7088112
25	100	49	0.4900000	0.0499900	0.3920214	0.5879786
26	102	53	0.5196078	0.0494693	0.4226498	0.6165659
27	99	53	0.5353535	0.0501261	0.4371082	0.6335989
28	100	52	0.5200000	0.0499600	0.4220802	0.6179198
29	100	55	0.5500000	0.0497494	0.4524930	0.6475070
30	102	56	0.5490196	0.0492689	0.4524544	0.6455848
31	97	61	0.6288660	0.0490522	0.5327254	0.7250065
32	99	58	0.5858586	0.0495055	0.4888297	0.6828875
33	103	60	0.5825243	0.0485908	0.4872881	0.6777605
34	99	47	0.4747475	0.0501878	0.3763813	0.5731137
35	100	52	0.5200000	0.0499600	0.4220802	0.6179198
36	101	55	0.5445545	0.0495539	0.4474305	0.6416784
37	103	53	0.5145631	0.0492456	0.4180436	0.6110826
38	101	45	0.4455446	0.0494559	0.3486127	0.5424764
39	99	49	0.4949495	0.0502493	0.3964626	0.5934364
40	99	55	0.5555556	0.0499407	0.4576735	0.6534376
41	99	47	0.4747475	0.0501878	0.3763813	0.5731137
42	101	55	0.5445545	0.0495539	0.4474305	0.6416784
43	100	52	0.5200000	0.0499600	0.4220802	0.6179198
44	100	61	0.6100000	0.0487750	0.5144028	0.7055972
45	101	58	0.5742574	0.0492001	0.4778270	0.6706879
46	100	62	0.6200000	0.0485386	0.5248660	0.7151340
47	102	55	0.5392157	0.0493549	0.4424819	0.6359495
48	100	54	0.5400000	0.0498397	0.4423159	0.6376841
49	100	52	0.5200000	0.0499600	0.4220802	0.6179198
50	100	53	0.5300000	0.0499099	0.4321784	0.6278216

TAB. 1. Intervalles de confiance [pmin,pmax] pour les données sondages.sortie.urnes