UNIVERSITÉ DE NANTES

Université de Nantes – UFR des Sciences et Techniques
Département de Mathématiques

Master 1 : statistique **Année 2006-2007**

TP 1: Statistique descriptive

A. Philippe

Le logiciel utilisé pendant les séances de TP est le logiciel libre `R`. Une présentation de `R` est disponible sur le web à l'adresse suivante

`http://www.math.sciences.univ-nantes.fr/~philippe/index.php?frame=R`

Ex 1. *Statistique descriptive*

Les données `sleep` donnent l'augmentation ou la diminution du temps de sommeil (variable `extra`) chez deux groupes de patients traités par deux médicaments (la variable `group` prend les valeurs 1 ou 2). On dispose de 20 données.

```
 data(sleep)
 sleep
   extra group
1   0.7     1
2  -1.6     1
3  -0.2     1
4  -1.2     1

....

15 -0.1     2
16  4.4     2
17  5.5     2
18  1.6     2
19  4.6     2
20  3.4     2
```

1) Créer un vecteur `g1` contenant la variable `extra` du groupe 1

```
 g1= sleep$extra[1:10]
#plusieurs syntaxes
#ou sleep$extra[sleep$group==1]
#ou sleep$extra[which(sleep$group==1)]
```

puis, un vecteur `g2` contenant la variable `extra` du groupe 2

2) Représenter l'histogramme (en utilisant la fonction `hist` avec l'option `proba=TRUE`, voir l'aide) des données `g1` puis ajouter sur ce graphique des droites verticales de couleurs différentes correspondant

- à la moyenne (fonction `mean`)
- à la médiane (fonction `median`)
- aux quartiles Q_1 et Q_3 (fonction `quantile`)

Ajouter une légende.

```
 hist(g1)
 abline(v=mean(g1))
 -----etc---------
legend(-----à-complèter-----)
```

3) Refaire le même graphique avec les données `g2`

4) Tracer sur un même graphique les densités pour les deux groupes

```
 plot(density(g1))
 lines(density(g2),col=2)
```

Ajuster les axes en utilisant les options `xlim=c(a,b)` et `ylim` de la fonction `plot`

5) Tracer les boxplots des deux séries sur un même graphique

```
boxplot(data.frame(g1,g2))
```

6) Commenter les résultats obtenus.

Ex 2. Étude de la série `nottem` : relevé de températures à Nottingham pendant 20 ans (une donnée par mois)

1) Tracer la série de données

```
plot(nottem)
```

et commenter l'allure de la courbe.

2) On organise les données sous la forme d'une matrice (12 colonnes, 20 lignes) telle que la première colonne contient les données du mois de janvier.

```
DT = matrix(nottem, ncol=12 , byrow =T)
```

3) Représenter les différents mois sur un même graphique. Commenter

```
 matplot(DT, type="b", main="les différents mois")
```

4) Représenter les différentes années sur un même graphique. Commenter

5) Expliquer et commenter le graphique suivant

```
boxplot(data.frame(DT))
```

6) Construire la courbe des températures moyennes à Nottingham

7) Ajouter sur le graphique la courbe des minima et des maxima.

UNIVERSITÉ DE NANTES

Université de Nantes – UFR des Sciences et Techniques
Département de Mathématiques

Master 1 : statistique **Année 2006-2007**

TP 2: Statistique descriptive (suite)

A. Philippe

Ex 3. *Nombre de classes dans un histogramme*

L'option `nclass` (ou `breaks`) de la fonction `hist` permet d'ajuster le nombre de classes de l'histogramme. Par défaut, il est fixé par la formule de Sturges

$$n_c = [1 + Log_2(n)]$$

où [ ] désigne la partie entière.

1) Simuler $n = 100$ variables aléatoires iid suivant la loi gaussienne standard.

2) Sur une même page (`par(mfrow=c(3,3))`) :
tracer l'histogramme de l'échantillon simulé en faisant varier le nombre de classes. Prendre par exemple `nclass` $\in \{3, 5, 8, 10, 15, 20, 25, 30, 50\}$

3) Refaire la question précédente avec un échantillon simulé suivant la loi exponentielle de paramètre 1, puis avec un échantillon simulé suivant la loi de Cauchy

4) Commenter les résultats.

Ex 4. *Boxplot : pourquoi 1.5 ?*

Dans un boxplot, les valeurs aberrantes sont définies comme les observations en dehors de l'intervalle

$$I = [Q_1 - \mathbf{1.5}(Q_3 - Q_1), Q_3 + \mathbf{1.5}(Q_3 - Q_1).$$

D'où vient le **1.5**?

Réponse de l'inventeur du boxplot (Turkey):

> Because 1 is too small and 2 is too large.

On suppose que les observations sont iid suivant une loi normale standard $\mathcal{N}(0, 1)$.

1) Calculer les quartiles théoriques Q_1 et Q_3 d'une loi gaussienne standard (en utilisant la fonction `qnorm`).

2) En déduire l'écart inter quartiles théorique de la loi gaussienne standard et l'intervalle I.

3) Calculer la probabilité de l'évènement $[X \in I]$ quand $X \in \mathcal{N}(0, 1)$

4) Refaire ces calculs en remplaçant 1.5 par 1 puis par 2.

5) Commenter les résultats.

Ex 5. *Robustesse de la médiane*

Pour évaluer la robustesse, on étudie les variations de la médiane et de la moyenne lorsque l'on ajoute des observations aberrantes.

Prenons un $n-$échantillon (n=100) simulé suivant la loi gaussienne standard. Pour évaluer la robustesse, on peut par exemple ajouter m fois l'observation $z = 5$ et représenter l'évolution de la médiane et de la moyenne en fonction de m.

1) Commenter et exécuter le code suivant

```
x=rnorm(500) ;
z=5 ;
moyenne=NULL ; mediane=NULL ;
m.max=25
for( m in 1:m.max)
      { xx=c(x,rep(z,m))
       moyenne=c(moyenne,mean(xx) )
       mediane=c(mediane,median(xx))
      }

 matplot(cbind(moyenne,mediane), type="l",ylim=c(-0.3,.5))
 abline(h=c(mean(x),median(x)) ,col=1:2,lty=1:2)
```

2) Commenter les résultats obtenus.

UNIVERSITÉ DE NANTES

Université de Nantes – UFR des Sciences et Techniques
Département de Mathématiques

Master 1 : statistique **Année 2006-2007**

TP 3: Méthode de Monte Carlo

A. Philippe

Sous linux : On lance le logiciel `R` dans un terminal avec la commande `R`.
Pour sauvegarder les graphiques, utiliser la commande
`dev.print(png, file="nom-du-fichier.png", width=480, height=480)`
Le format `png` est reconnu par Word et OpenOffice.

Les méthodes de Monte Carlo sont des méthodes numériques qui permettent par exemple d'approcher la valeur d'une intégrale, de maximiser une fonction ...

La situation la plus simple est la suivante : on veut estimer une intégrale qui s'écrit sous la forme

$$I = \int h(x)f(x)\ \mathrm{d}x$$

où f est une densité et h une fonction mesurable telle que $\int |h|(x)f(x)\ \mathrm{d}x < \infty$.

L'estimateur de Monte Carlo est construit de la façon suivante.

- On simule $x_1, \ldots, x_n$ des nombres aléatoires suivant la loi de densité f.
- On approche l'intégrale I par $\delta_n(I) = \dfrac{1}{n}\sum_{i=1}^{n} h(x_i)$

La loi forte des grands nombres assure la convergence presque sure de l'estimateur de Monte Carlo vers l'intégrale I. De plus, sous l'hypothèse $\mathrm{Var}_f(h(X)) < \infty$, on montre que

- l'estimateur converge au sens de la convergence L^2
- $\sqrt{n}(\delta_n(I) - I)$ converge en loi vers une variable gaussienne centrée et de variance $\mathrm{Var}_f(h(X))$.

Ex 6. On souhaite estimer l'intégrale[1]

$$\int_0^1 h(x)\ \mathrm{d}x \qquad \text{avec} \qquad h(x) = (\cos(50x) + \sin(20x))^2$$

1) Tracer la fonction h sur $[0, 1]$

[1] la valeur exacte de l'intégrale est 0.965

2) Simuler un échantillon de taille $n = 10000$ suivant la loi uniforme sur $[0, 1]$

3) Créer un vecteur contenant les valeurs de $\delta_j(I)$ pour $j = 1, \ldots, n$.

4) Représenter la courbe reliant les points $\{(j, \delta_j(I)), \quad j = 1, \ldots, n\}$.

5) En déduire une estimation de l'intégrale I

6) Recommencer le calcul des $\delta_j(I)$ pour $j = 1, \ldots, n$ sur 100 échantillons indépendants. Stocker les résultats dans une matrice. (par exemple la k ème colonne contient les valeurs $\{\delta_j(I),\ j = 1, \ldots, n\}$ calculées sur le k ème échantillon simulé. On a donc une matrice 100 colonnes et 10000 lignes)

7) Superposer les 100 courbes reliant les points $\{(j, \delta_j(I)), = j = 1, ..., 10000\}$

8) Commenter. Quelle est la qualité de l'estimation calculée à la question 5)

On a $N = 100$ estimations indépendantes de l'intégrale I construites à partir d'échantillons de taille $j \in \{1, ..., 10000\}$.

On note ces valeurs $\delta_j^1, \ldots, \delta_j^{100}$ (c'est la j ème ligne de la matrice construite à la question 6)

9) Représenter l'histogramme de l'échantillon $\delta_j^1, \ldots, \delta_j^{100}$ pour $j = 5000$ puis pour $j = 10000$

10) En utilisant la fonction `density` donner une estimation de la densité de l'échantillon $\delta_j^1, \ldots, \delta_j^{100}$ pour $j = 5000$, puis pour $j = 10000$. Superposer les deux estimations sur un même graphique.

11) Commenter les résultats obtenus aux questions 9) et 10). Quelles propriétés peut on illustrer à partir de ces graphiques.

12) Calculer la variance (notée Var_j) de échantillon $\delta_j^1, \ldots, \delta_j^{100}$ pour tout $j = 1, \ldots, n$.

13) Représenter Var_j en fonction de la taille de l'échantillon j.

14) Quelle quantité peut on estimer à partir de cette courbe ?

15) Représenter la courbe reliant les points $\{(\log(j), \log(\mathrm{Var}_j)),\ j = 1, \ldots, n\}$ (représentation log-log de la variance estimée).

16) Commenter l'allure de la courbe.

17) Représenter la courbe reliant les points $\left\{ \left(j, \dfrac{\log(\mathrm{Var}_j) - \log(\mathrm{Var}_1)}{\log(j)} \right),\ j = 2, \ldots, n \right\}$.

18) Quel est l'intérêt de ce graphique ?

19) En déduire une estimation de la vitesse de convergence de l'estimateur $\delta_n(I)$ vers l'intégrale I

UNIVERSITÉ DE NANTES

Université de Nantes – UFR des Sciences et Techniques
Département de Mathématiques

Master 1 : statistique **Année 2006-2007**

TP 4: Le modèle uniforme

A. Philippe

Ex 7. On considère un n-échantillon $(X_1, \cdots, X_n)$ d'une loi uniforme sur $[0, \theta]$.
On pose

$$\hat{\theta}_n = \max\{X_1, \cdots, X_n\}.$$

On a prouvé les résultats suivants en TD :

- $\hat{\theta}_n$ est asymptotiquement sans biais.
- $\hat{\theta}_n$ converge presque sûrement et dans L^2 vers θ quand $n \to \infty$.

Comment vérifier ces propriétés par la simulation ?

1) Simuler $N = 500$ échantillons de taille $n = 1000$ de loi uniforme sur $[0, \theta_0]$ par exemple $\theta_0 = 2$. Stocker ces échantillons dans une matrice.

2) Pour chacun des échantillons simulés, calculer les valeurs de $\hat{\theta}_i$ pour $i = 1, \ldots, n$. Stocker les résultats dans une autre matrice.

3) Représenter sur un même graphique, les N courbes reliant les points $\left\{(i, \hat{\theta}_i),\ i = 1, \ldots, n\right\}$

4) Interpréter ce graphique.

5) Utiliser une méthode de Monte Carlo pour estimer le biais de l'estimateur lorsque la taille de l'échantillon varie de 1 à n.

6) Représenter le biais estimé en fonction de la taille de l'échantillon. Commenter.

7) Utiliser une méthode de Monte Carlo pour estimer la variance de l'estimateur lorsque la taille de l'échantillon varie de 1 à n.

8) Représenter la variance estimée (notée $\widehat{\mathrm{Var}}(j)$) en fonction de la taille de l'échantillon $j = 1, \ldots, n$.

9) Représenter le nuage de points

$$\left\{(\log(j), \log(\widehat{\mathrm{Var}}(j))),\ j = 1, \ldots, n\right\}$$

[*représentation log-log*]

10) Commenter les résultats et faire le lien avec la vitesse de convergence de l'estimateur.

11) Reprendre les questions précédentes avec $\theta_0 = 10$. Superposer sur un même graphique, la variance estimée pour les deux valeurs de θ_0 (utiliser la représentation log-log). Comparer les vitesses de convergence.

On veut maintenant comparer deux estimateurs sans biais de θ (vu en TD)

1. $\hat{\theta}'_n = \dfrac{n+1}{n} \max\{X_1, \cdots, X_n\}$
2. $\hat{\theta}''_n = \dfrac{2}{n} \displaystyle\sum_{j=1}^{n} X_j$

12) En utilisant les questions précédentes, calculer et représenter une estimation de la variance de $\hat{\theta}'_j$ en fonction de $j = 1, \ldots, n$.

13) Utiliser une méthode de Monte Carlo pour estimer la variance de l'estimateur $\hat{\theta}''_j$ lorsque la taille de l'échantillon varie de 1 à n.

14) Ajouter sur le graphe précédent, la variance estimée de $\hat{\theta}''_j$ en fonction de $j = 1, \ldots, n$.

15) Commenter les résultats obtenus.

16) Comparer les vitesses de convergence à l'aide de la représentation log-log des deux variances.

17) Conclure.

On souhaite maintenant illustrer la convergence en loi.

18) Mettre en évidence que $\sqrt{n}(\hat{\theta}_n - \theta)$ ne converge pas en loi vers une variable gaussienne $\mathcal{N}(0, \sigma^2)$ $(\sigma > 0)$

19) Illustrer la convergence en loi de $n(\hat{\theta}_n - \theta)$ vers une variable non gaussienne.

20) Donner une estimation de la densité de la loi limite de $n(\hat{\theta}_n - \theta)$.

21) Comparer la loi limite de $n(\hat{\theta}_n - \theta)$ avec celle de $\sqrt{n}(\hat{\theta}''_n - \theta)$.

UNIVERSITÉ DE NANTES

Université de Nantes – UFR des Sciences et Techniques
Département de Mathématiques

Master 1 : statistique **Année 2006-2007**

TP 5 : La vraisemblance

A. Philippe

Ex 8. 1) Simuler un échantillon de taille 100 suivant la loi gaussienne de moyenne zéro et de variance 4.

2) Créer une fonction pour calculer la vraisemblance de l'échantillon simulé.

3) Représenter la vraisemblance

Indications : utiliser la fonction `outer` et les fonctions graphiques `persp` ou `image`

4) Donner une valeur approchée de l'estimateur du MV.

5) Reprendre les questions précédentes en augmentant la taille de l'échantillon (1000, puis 10 000).

6) Commenter les résultats obtenus.

Ex 9.

1) Charger la librairie `MASS` et les données `quine`.

indication : utiliser les fonctions `library,data`

On cherche à modéliser le nombre de jours d'absence. Les observations sont disponibles dans la variable `Days`.

2) Donner une estimation de la moyenne, puis de la variance de cette variable.

3) Représenter l'histogramme des données

4) Déduire des résultats précédents une modélisation possible.

5) Pour le modèle choisi, tracer la fonction de vraisemblance.

6) En déduire une valeur approchée de l'estimateur du MV

7) Superposer l'histogramme et la loi obtenue pour modéliser ces données.

8) Commenter

Ex 10.

1) Simuler un échantillon iid suivant la loi normale centrée, réduite.

2) Comparer les performances de l'estimateur à noyau, lorsque l'on change le noyau K et la fenêtre h_n.

3) Reprendre avec un échantillon iid suivant la loi uniforme sur $[0, 1]$

4) Reprendre avec un échantillon iid suivant la loi qui admet pour densité

$$f(x) = 4x\,\mathbb{I}_{[0,1/2]} + 4(1-x)\,\mathbb{I}_{]1/2,1]}$$

UNIVERSITÉ DE NANTES

Université de Nantes – UFR des Sciences et Techniques
Département de Mathématiques

Master 1 : statistique **Année 2006-2007**

TP 6 : Tests et régions de confiance

A. Philippe

Ex 11.

On jette une pièce 5 fois et on note N le nombre de fois où la pièce tombe sur PILE. Soit p la probabilité que la pièce tombe sur PILE. On veut tester $p = 1/2$ contre $p > 12$. Pour obtenir un test de niveau $\alpha = 5\%$, il suffit de rejeter H_0 avec une probabilité 1 lorsque $N = 5$ et avec une probabilité γ lorsque $N = 4$, ce qui revient à prendre

$$\Phi(N) = \mathbb{I}_{\{5\}}(N) + \gamma \mathbb{I}_{\{4\}}(N).$$

On ajuste la valeur de γ en écrivant que le niveau est 5%, soit

$$\frac{1}{32} + \frac{5\gamma}{32} = \frac{5}{100},$$

soit $\gamma = 3/25$.

1) Écrire une fonction qui retourne la décision du test au niveau 5%.

2) Simuler $N_1, \ldots, N_n$ un n-échantillon distribué suivant la loi binomiale de paramètres $(5, \frac{1}{2})$ (prendre n=1000)

3) Calculer la décision du test pour chacune des observations N_k, $k = 1, ..., n$.

4) Calculer la fréquence de l'évènement *on rejette l'hypothèse nulle* et commenter le résultat obtenu.

5) Reprendre les questions 3) et 4) lorsque la loi de l'échantillon $N_1, \ldots, N_n$ est la loi binomiale de paramètres $(5, p)$ avec $p = 0.6;\ 0.7;\ 0.8;\ 0.9$

6) Commenter les résultats obtenus

7) Écrire une fonction qui calcule γ et c en fonction K le nombre de lancers

8) Reprendre les questions pécédentes pour $K = 10; 50; 100$

9) Commenter les résultats

Ex 12. Soit $X_1, \ldots, X_n$ un n-échantillon distribué suivant la loi normale de paramètres (μ, σ^2).

Situation 1

On suppose que σ^2 est connu et on estime μ

Soit Q_α le quartile supérieur d'ordre α de la loi normale standard
c'est à dire si $Y \sim \mathcal{N}(0,1)$ alors $P(Y > Q_\alpha) = \alpha$

1) Montrer que l'intervalle

$$I_n = [\bar{X}_n - \frac{Q_\alpha \sigma}{\sqrt{n}}; \bar{X}_n + \frac{Q_\alpha \sigma}{\sqrt{n}}]$$

est un intervalle de confiance de niveau $1 - 2\alpha$ c'est à dire

$$P_\mu(\mu \in I_n) = 1 - 2\alpha$$

2) Simuler N échantillons de taille $n = 25$ iid suivant la loi normale de paramètres $(1, 1)$

3) Calculer pour chaque échantillon la région de confiance I_n au niveau 95%

4) Représenter les régions de confiance

`Indications`

```
> lu  # bornes supérieures
> li  # bornes inférieures
> matplot(cbind(li,lu), pch=1,col=1)
> for(i in 1:length(lu)) lines(c(i,i),c(li[i],lu[i]))
```

5) Utiliser des couleurs pour distinguer les intervalles qui contiennent $\mu = 1$

6) Evaluer la fréquence de l'évènement μ *appartient à l'intervalle de confiance* Commenter le résultat.

Situation 2

On suppose que σ^2 est inconnu.

Soit $T_{n,\alpha}$ le quartile supérieur d'ordre α de la loi de Student à n degrés de liberté. On note $\hat{\sigma}_n^2$ l'estimateur du MV de σ^2.

7) Montrer que l'intervalle

$$I'_n = [\bar{X}_n - \frac{T_{n-1,\alpha}\hat{\sigma}_n}{\sqrt{n}}; \bar{X}_n + \frac{T_{n-1,\alpha}\hat{\sigma}_n}{\sqrt{n}}]$$

est un intervalle de confiance de niveau $1 - 2\alpha$ pour le paramètre μ.

8) Reprendre les questions 3)→6) pour les intervalles I'_n..

UNIVERSITÉ DE NANTES

Université de Nantes – UFR des Sciences et Techniques
Département de Mathématiques

Master 1 : statistique **Année 2006-2007**

TP 7 : Tests sur la moyenne

A. Philippe

Ex 13. Un outil graphique pour tester la normalité : `qqnorm` cette fonction represente les quantiles théoriques contre les quantiles empiriques. Si l'échantillon est bien gaussien alors la courbe est une droite.

```
> qqnorm(x) ; qqline(x)
```

1) Simuler des échantillons gaussiens et tracer la courbe `QQ plot`

2) Simuler des échantillons non gaussiens

- Student à 1; 3 ; 5 ; 10 ; 15 ; 50 ; 100 degrés de liberté
- $X + U_a$ où X suit une loi gaussienne standard et U_a une loi uniforme sur $[0, a]$ avec (a= 0.1 ; 0.5 ; 1 ; 2)

et tracer la courbe `QQ plot`

3) Commenter les résultats

Ex 14. Test de Kolmogorov-Smirnov :

$\mathcal{H}_0$: la loi P a pour fonction de répartition F_0 ,

On mesure l'adéquation de la fonction de répartition empirique à la fonction F_0 par la distance de Kolmogorov-Smirnov,

$$D_{KS}(F_0, \widehat{F}) = \max_{i=1,\dots,n} \left\{ \left| F_0(X_{(i)}) - \frac{i}{n} \right|, \left| F_0(X_{(i)}) - \frac{i-1}{n} \right| \right\} .$$

Sous l'hypothèse $\mathcal{H}_0$, la loi de la statistique $D_{KS}(F_0, \widehat{F})$ ne dépend pas de F_0. Mais la fonction de répartition de $D_{KS}(F_0, \widehat{F})$ n'a pas d'expression explicite simple et doit être calculée numériquement.

Si l'hypothèse $\mathcal{H}_0$ est fausse, $D_n = \sqrt{n} D_{KS}(F_0, \widehat{F})$ tend vers $+\infty$ avec n. La région de rejet R_0 du test est donc $\{D_n > C(\alpha)\}$ telle que

$$P_{H_0}(D_n > C(\alpha)) = \alpha$$

1) À l'aide de la fonction `ks.test`, tester la normailté des échatillons simulés dans l'exercice précédent

```
ks.test(X,'pnorm',0,1)  # pour tester si la loi est gaussienne standard
```

Ex 15. 1) Écrire des fonctions permettant de tester l'égalité des moyennes de deux échantillons gaussiens.

Ex 16. On désire tester si un médicament a une influence sur le comportement psychomoteur. On choisit au hasard 20 sujets qu'on répartit au hasard en deux groupes: le groupe témoin et le groupe expérimental. On leur fait subir la même expérience psychomotrice. On a administré auparavant le médicament aux sujets du groupe expérimental et un placebo au groupe témoin. Les résultats sont les suivants:

Groupe témoin	166	167	169	170	174	173	172	170	166	173
Groupe expérimental	167	162	165	168	162	160	164	158	165	169

On suppose que dans chaque groupe les résultats sont distribués selon une loi gaussienne, que la variance est la même pour les deux groupes et que les performances des sujets sont indépendantes.

1) Tester la normailté des échantillons

2) Comparer les variances des deux échantillons

3) Tester au niveau 0.05 l'hypothèse selon laquelle le médicament n'a aucun effet sur le comportement psychomoteurn, en utilisant votre fonction test (voir exercice précédent), puis à l'aide de la fonction `t.test`)

Ex 17. On a fait une numération globulaire à un groupe de 20 personnes à deux périodes différentes de l'année. Pour chaque sujet on note les résultats des deux numérations (à multiplier par 10^5):

Sujet	01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20
Janvier-février	46 38 42 47 48 40 40 43 42 49 45 51 47 52 50 48 47 47 47 45
Sept.-oct.	48 47 44 45 51 44 47 48 47 57 49 55 48 48 46 48 54 54 44 48

On suppose que les sujets sont mutuellement indépendants et suivent une loi gaussienne.

1) Tester la normailté des échantillons

2) Tester au niveau 0.05 l'hypothèse selon laquelle les résultats de la numération sont les mêmes aux deux périodes.