CPP - la prépa des INP (2 ème année). Bordeaux, 16/04/2014.

Devoir de statistiques: CORRIGE
durée 2h

Données: On rappelle que si Z suit une loi $\mathcal{N}(0,1)$, on a

$$\mathbb{P}(Z \leq 1.96) \simeq 0,975 \text{ et } \mathbb{P}(Z \leq 1.65) \simeq 0,95.$$

Exercice 1. On considère une variable aléatoire X de densité f_θ avec $-\frac{1}{2} \leq \theta \leq \frac{1}{2}$ et

$$f_\theta(x) = \begin{cases} \frac{1}{2} - \theta & \text{si } x \in [-1, 0[, \\ \frac{1}{2} + \theta & \text{si } x \in [0, 1], \\ 0 & \text{sinon} \end{cases}$$

1. Représenter f_θ et justifier que f_θ est bien une densité de probabilité pour $-\frac{1}{2} \leq \theta \leq \frac{1}{2}$. f_θ est positive et d'intégrale 1 (et continue par morceaux donc bien intégrable).

2. Calculer l'espérance et la variance de X. (Les résultats sont $\mathbb{E}[X] = \theta$ et $\text{Var}(X) = \frac{1}{3} - \theta^2$). $\mathbb{E}[X] = \int_{\mathbb{R}} x f_\theta(x) dx = \theta$, $\mathbb{E}[X^2] = \int_{\mathbb{R}} x^2 f_\theta(x) dx = 1/3$ et $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{3} - \theta^2$.

On considère maintenant $X_1, \dots X_n$ des variables aléatoires indépendantes et de même loi. On suppose que la loi commune est de densité f_θ avec $-\frac{1}{2} \leq \theta \leq \frac{1}{2}$ inconnu. On va chercher à estimer θ.

3. Proposer un esitmateur $\hat{\theta}$ de θ basé sur la méthode des moments. On prend $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

4. Calculer son biais et son risque quadratique. Il est sans biais ($\mathbb{E}_\theta[\hat{\theta}_n] = \theta$) et de risque quadratique:

$$\mathbb{E}_\theta[(\hat{\theta}_n - \theta)^2] = \text{Var}_\theta(\hat{\theta}_n) = \frac{\text{Var}(X_1)}{n} = \frac{\frac{1}{3} - \theta^2}{n}.$$

On va maintenant s'intéresser à l'estimateur du maximum de vraisemblance de θ.

5. On note $N_1 = \sum_{i=1}^n \mathbf{1}_{\{X_i \geq 0\}}$ et $N_2 = \sum_{i=1}^n \mathbf{1}_{\{X_i < 0\}}$. Soient $x_1, \dots, x_n \in [-1, 1]$ et $-\frac{1}{2} \leq \theta \leq \frac{1}{2}$, écrire la vraisemblance du modèle au point $(x_1, \dots, x_n; \theta)$ en fonction de N_1, N_2 et θ.

En un tel point la vraisemblance s'écrit:

$$L(x_1,\dots,x_n,\theta)=\prod_{i=1}^n f_\theta(x_i)=\prod_{i=1}^n\left(\frac{1}{2}-\theta\right)^{\mathbf{1}_{\{X_i<0\}}}\left(\frac{1}{2}+\theta\right)^{\mathbf{1}_{\{X_i\geq 0\}}}$$
$$=\left(\frac{1}{2}-\theta\right)^{\sum_{i=1}^n \mathbf{1}_{\{X_i<0\}}}\left(\frac{1}{2}+\theta\right)^{\sum_{i=1}^n \mathbf{1}_{\{X_i\geq 0\}}}=\left(\frac{1}{2}-\theta\right)^{N_2}\left(\frac{1}{2}+\theta\right)^{N_1}.$$

6. Montrer que l'estimateur du maximum devraisemblance existe et vaut:

$$\hat{\theta}_{MV}=\frac{N_1-N_2}{2(N_1+N_2)}=\frac{N_1}{n}-\frac{1}{2}.$$

On calcule $\ln L(x_1,\dots,x_n,\theta)=N_2\ln\left(\frac{1}{2}-\theta\right)+N_1\ln\left(\frac{1}{2}+\theta\right)$. Pour chercher, le maximum en θ (s'il existe), on calcule la dérivée en θ. On a:

$$\partial_\theta lnL=-\frac{N_2}{\frac{1}{2}-\theta}+\frac{N_1}{\frac{1}{2}+\theta}$$

Cette dérivée s'annule si et seulement si $\theta=\frac{N_1-N_2}{2(N_1+N_2)}$. On vérifie que l'on a bien un maximum local en ce point. C'est le seul extremum local, donc il s'agit bien d'un maximum global.

7. On pose $Y_i=1_{\{X_i\geq 0\}}$, pour $1\leq i\leq n$. Calculer l'espérance et la variance des variables aléatoires Y_i. En déduire l'espérance et le risque quadratique de $\hat{\theta}_{MV}$.

Y_i ne prend que 2 valeurs 0 et 1. Y_i suit donc une loi de Bernoulli de paramètre:

$$p=\mathbb{P}(Y_i=1)=\mathbb{P}(X_i\geq 0)=\int_0^{+\infty} f_\theta(x)dx=1/2+\theta.$$

Son espérance vaut $p=1/2+\theta$ et sa variance $p(1-p)=1/4-\theta^2$.

8. Quel estimateur vaut-il mieux utiliser entre $\hat{\theta}$ et $\hat{\theta}_{MV}$? Justifier. On a $\hat{\theta}_{MV}=\frac{N_1}{n}-\frac{1}{2}$ avec $N_1=\sum_{i=1}^n Y_i$. Les Y_i étant indépendants, on calcule facilement que $\mathbb{E}_\theta[\hat{\theta}_{MV}]=\theta$ et

$$\mathbb{E}_\theta[(\hat{\theta}_{MV}-\theta)^2]=\frac{1}{n^2}\sum \mathrm{Var}(Y_i)=\frac{1/4-\theta^2}{n}.$$

Les 2 estimateurs sont sans biais mais le risque quadratique de $\hat{\theta}_{MV}$ est plus petit que celui de $\hat{\theta}$. Il vaut donc mieux utiliser $\hat{\theta}_{MV}$.

9. 0n veut maintenant tester : l'hypothèse nulle: $H_0=\{\theta=0\}$ contre l'hypothèse alternative: $H_1=\{\theta>0\}$.

On suppose ici que la taille de l'échantillon est suffisament grande pour pourvoir utiliser le théorème central limite. Proposer un test de niveau

de confiance asymptotique 0,95 basé sur l'estimateur $\hat{\theta}$ ou $\hat{\theta}_{MV}$. On justifiera notamment la forme de la région d'acceptation ou de rejet du test.

On construit ici le test avec $\hat{\theta}_{MV}$. Sous H_0, par la loi forte des grands nombres, $\hat{\theta}_{MV} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ converge presque sûrement vers $\mathbb{E}_{\theta=0}[Y_1] = \frac{1}{2}$ lorsque $n \to +\infty$. De plus, vu que $\mathrm{Var}_{\theta=0}(Y_1) = 1/4$, le théorème central limite donne la convergence en loi lorsque $n \to +\infty$:

$$A_n := 2\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{2}\right) \longrightarrow^{\mathcal{L}} Z \text{ avec } Z \sim \mathcal{N}(0,1).$$

Sous H_1, $\hat{\theta}_{MV} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ converge presque sûrement vers $\mathbb{E}_{\theta}[Y_1] = \frac{1}{2} + \theta > \frac{1}{2}$. Donc A_n tend presque sûrement vers $+\infty$.

On choisit donc la règle de décision suivante : on accepte (ou plutôt on ne rejette pas) H_0 si $A_n \leq k$ et on rejette H_0 si $A_n > k$ avec k un réel à déterminer. La valeur de k est choisie de telle sorte que $\mathbb{P}(\text{rejeter } H_0 | H_0 \text{ est vraie}) \leq 0,05$. On a $\mathbb{P}(\text{rejeter } H_0 | H_0 \text{ est vraie}) = \mathbb{P}_0(A_n > k) \simeq \mathbb{P}(Z > k)$ avec Z de loi $\mathcal{N}(0,1)$. Au vu des données, on choisit donc: $k = 1,65$.

Exercice 2. On considère $(X_1, \ldots, X_n)$ des variables aléatoires indépendantes de loi géométrique de paramètre p inconnu ($0 \leq p \leq 1$). On pose $q = 1 - p$. On rappelle X est à valeurs dans $\mathbb{N}^*$ et que pour $k \geq 1$, $\mathbb{P}(X = k) = q^{k-1}p$.

1. Soit X une variable aléatoire de loi géométrique de paramètre p. Montrer que $\mathbb{E}[X] = \frac{1}{p}$. *Indication:* On rappelle que pour $|x| < 1$, on a

$$\sum_{k\geq 1} k x^{k-1} = \left(\sum_{k\geq 0} x^k\right)' = \left(\frac{1}{1-x}\right)' = \frac{1}{(1-x)^2}$$

Pour la suite, on admettra que $\mathrm{Var}(X) = \frac{q}{p^2}$.

$\mathbb{E}[X] = \sum_{k\geq 1} k\mathbb{P}(X = k) = \sum_{k\geq 1} k q^{k-1} p = \frac{p}{(1-q)^2} = \frac{1}{p}$.

2. Proposer un estimateur $\hat{\theta}_n$ de $\theta = \frac{1}{p}$ à l'aide de la méthode des moments: $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

3. Proposer un intervalle de confiance asymptotique de niveau 0,95 à l'aide du théorème central limite.

Au vu de la question suivante, on va le choisir centré autour de θ_n. Le théorème central limite nous dit que:

$$B_n := \frac{\sqrt{n}}{\sqrt{q/p^2}}\left(\hat{\theta}_n - \frac{1}{p}\right) \longrightarrow^{\mathcal{L}} Z \text{ avec } Z \sim \mathcal{N}(0,1).$$

Soit maintenant $l > 0$,

$$\begin{aligned}\mathbb{P}(\theta \in [\hat{\theta}_n - l, \hat{\theta}_n - l]) &= \mathbb{P}(-l \leq \hat{\theta}_n - \theta \leq l) \\ &= \mathbb{P}\left(-l\frac{p\sqrt{n}}{\sqrt{q}} \leq B_n \leq l\frac{p\sqrt{n}}{\sqrt{q}}\right) \\ &\simeq \mathbb{P}\left(-l\frac{p\sqrt{n}}{\sqrt{q}} \leq Z \leq l\frac{p\sqrt{n}}{\sqrt{q}}\right) \text{ avec } Z \sim \mathcal{N}(0,1).\end{aligned}$$

On veut que la probabilité ci-dessus soit (supérieure ou) égale à 0,95. On choisit donc l de telle sorte que:

$$l\frac{p\sqrt{n}}{\sqrt{q}} = 1,96 \text{ soit } l = \frac{1,96\sqrt{q}}{p\sqrt{n}}.$$

4. 0n veut maintenant tester l'hypothèse nulle: $H_0 = \{\theta = 2\}$ contre l'hypothèse alternative: $H_1 = \{\theta \neq 2\}$. A l'aide de la variable aléatoire $Y_n = \sqrt{\frac{n}{2}}\left(\frac{1}{n}\sum_{i=1}^n X_i - 2\right)$, construire un test de niveau asymptotique 0,95 pour θ.

   On remarque que sous H_0, $p = 1/2$ et $Y_n = B_n$ converge en loi vers une loi $\mathcal{N}(0,1)$ (d'après le TCL), tandis que d'après la LFGN, sous H_1, on a soit $Y_n \to +\infty$, soit $Y_n \to -\infty$ presque sûrement.

   On choisit donc la règle de décision suivante: on accepte (ou plutôt on ne rejette pas) H_0 si $|Y_n| \leq k$ et on rejette H_0 si $|Y_n| > k$ avec k un réel à déterminer. La valeur de k est choisie de telle sorte que $\mathbb{P}(\text{rejeter } H_0 | H_0 \text{ est vraie}) \leq 0,05$. On a $\mathbb{P}(\text{rejeter } H_0 | H_0 \text{ est vraie}) = \mathbb{P}_{\theta=2}(|Y_n| > k) \simeq \mathbb{P}(|Z| > k)$ avec Z de loi $\mathcal{N}(0,1)$. Au vu des données, on choisit donc: $k = 1,96$.

5. Sur un échantillon de taille n, on trouve que $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i = 2,3$. Que peut-on conclure si la taille de l'échantillon est de 70? de 200? Pour $n = 70$, on trouve $Y_n = \sqrt{35} * 0,3 = 1,77$. On ne rejette pas H_0. Pour $n = 200$, on trouve $Y_n = \sqrt{100} * 0,3 = 2,3$. On rejette H_0.

6. Proposer un script Matlab permettant de vérifier que l'erreur de première espèce est bien de 0,05 et calculant la puissance du test lorsque $(\theta \neq 2)$ lorsque la taille de l'échantillon n est de 70. On considerera que la commande *geometrique(N,p)* a déjà été programmée et qu'elle renvoie N variables aléatoires indépendantes de loi géométrique de paramètre p.

   Programme pour tester l'erreur de première espèce:
   On cherche à estimer $\mathbb{P}(\text{rejeter } H_0 | H_0 \text{ est vraie})$. A chacun des N

passages dans la boucle for, on simule un échantillon de taille 70 de loi géométrique de paramètre 1/2. On compte le nombre de fois ou le test est rejeté (alors qu'ici H_0 est vraie :$p = 1/2$). On lancera alors le programme avec un N suffisament grand ($N = 1000$ par exemple.)

```
function E=erreur1(N)
c=0
for i=1:N
A=geometrique (70,1/2)
S=sum(A)
Y= sqrt( 70/2)* (S/70 -1/2)
if abs(Y) > 1,96,
c=c+1
end
end
E= c/N
end
```

Programme pour tester la puissance du test:
On fixe $p \neq 1/2$. A chacun des N passages dans la boucle for, on simule un échantillon de taille 70 de loi géométrique de paramètre p. On compte le nombre de fois ou le test est H_1 est accepté. Ici si on choisit $p \neq 1/2$, c'est H_1 qui est vraie).

```
function P=puissance(N,p)
c=0
for i=1:N
A=geometrique (70,p)
S=sum(A)
Y= sqrt( 70/2)* (S/70-1/2)
if abs(Y) > 1,96,
c=c+1
end
end
P= c/N
end
```