Statistique exploratoire: Examen du 28 août 2012

Préliminaires

Cet examen est à réaliser sur papier, avec l'aide éventuelle d'un ordinateur. Les codes R devront être stockés dans le fichier "examen R" présent sur le bureau du compte temporaire. Ne pas oublier de reporter votre identifiant sur votre copie. Seuls les polys et les livres de R le sont. Les problèmes sont indépendants, peuvent être traités dans n'importe quel ordre et le barème est donné à titre indicatif. **Vous ne devez traiter que cinq problèmes au maximum.**

1 Estimation de constante [4 points]

On souhaite estimer la constante

$$e = \sum_{n=0}^{\infty} \frac{1}{n!}\,.$$

On considère une suite infinie de variables aléatoires $U_1, U_2, \ldots, U_i \ldots$ de loi $\mathcal{U}_{[0,1]}$ et la variable aléatoire K qui indique la première réalisation qui dépasse la réalisation qui la précède, c'est à dire

$$K \;=\; \min_{\{k\geq 2\}} \{k \;:\; U_k \geq U_{k-1}\}\,.$$

On observe que $P(K=k) \;=\; \frac{k-1}{k!}$ car $P(K>k) \;=\; P(U_1 \geq U_2 \geq U_3 \geq \cdots \geq U_k) = \frac{1}{k!}$ et donc

$$\mathbb{E}[K] \;=\; \sum_{k=2}^{\infty} k\,\frac{k-1}{k!} \;=\; \sum_{n=0}^{\infty} \frac{1}{n!} \;=\; e\,.$$

1. [**1.5**] Ecrire une fonction qui simule n réalisations de la variable K.
2. [**.5**] Pour $n = 10^3$ donner une estimation par Monte-Carlo de e et la comparer à sa valeur théorique.
3. [**1**] Donner une estimation de l'intervalle de confiance de niveau $1-\alpha$ avec $\alpha = 0.05$.
4. [**1**] Vérifier graphiquement la convergence de l'estimateur vers sa valeur théorique. Tracer sur le même graphique l'évolution de l'intervalle de confiance.

2 Test d'adéquation [4 points]

On considère le jeu de données `fdeaths` (déjà résident sous R). et on cherche à tester l'adéquation de ces données à une loi exponentielle, $\mathcal{E}(\lambda)$, de densité $\lambda \exp(-\lambda x)$ sur $\mathbb{R}_+$.

1. [**1**] Le résultat de l'appel à la procédure `ks.test` pour tester l'adéquation de ces données à une loi exponentielle de paramètre $\lambda = 1/540$ est

```
> ks.test(fdeaths,"pexp",1/540)

        One-sample Kolmogorov-Smirnov test

data:  fdeaths
D = 0.4573, p-value = 2.864e-14
alternative hypothesis: two-sided
```

En déduisez-vous que les données sont *(encadrer la bonne réponse)*

1. compatibles avec une loi $\mathcal{E}(1/540)$;
2. incompatibles avec une loi $\mathcal{E}(1/540)$.

On définit l'échantillon `drift` par `drift=fdeaths-min(fdeaths)`.

2. **[0.5]** Interpréter le résultat du test de Kolmogorov-Smirnov sur l'adéquation de `drift` à une loi exponentielle de paramètre `lambda=1/mean(drift)` et l'adéquation de `drift` à une loi exponentielle de paramètre `lambda=1/sqrt(var(drift))`

Le fait de choisir λ en fonction des données invalide la p-value fournie ci-dessus par R.

3. **[0.5]** La p-value est-elle *(encadrer la bonne réponse)*

1. la probabilité de dépasser le D observé sous l'hypothèse nulle;
2. la probabilité d'être en dessous du D observé sous l'hypothèse nulle;
3. la probabilité de dépasser le D observé sous l'hypothèse alternative;
4. la probabilité d'être en dessous du D observé sous l'hypothèse alternative.

On construit ci-dessous une `p-value` non-biaisée par bootstrap.

4. **[2]** Si D est la valeur de la statistique retournée par le test de Kolmogorov-Smirnov sous R, par exemple $D = 0.1144$ ci-dessus, détailler la construction d'un programme R évaluant par bootstrap paramétrique la probabilité de dépasser le D observé pour `drift` sous l'hypothèse nulle que les données viennent bien d'une loi exponentielle de paramètre `1/mean(drift))=0.004334999`.

3 Programmation R [4 points]

Ecrire un programme R permettant de résoudre le problème suivant : Soit une loterie comprenant N tickets numérotés de 1 à N. On suppose les N tickets vendus. Les tickets gagnants sont ceux comportant un 1 et un 3 à droite du 1, comme par exemple 123 et 8135. Écrire un programme R permettant de trouver l'unique valeur de $999 < N < 9999$ telle que la proportion de billets gagnants soit exactement 10%.

4 Approximation de π [2 points]

1. Soient (X, Y) deux variables aléatoires indépendantes et identiquement distribuées de loi uniforme sur $[-1, 1]$.
   (a) **[.5]** Montrer que le couple (X, Y) est de loi uniforme sur un domaine $\mathcal{D}$ de $\mathbb{R}^2$ à préciser.
   (b) **[.5]** Donner la probabilité théorique que $P(X^2+Y^2 \leq 1)$ (un dessin et un raisonnement géométrique pourront suffire).
2. **[1]** Proposer une méthode de Monte-Carlo pour calculer la probabilité $P(X^2 + Y^2 \leq 1)$. En déduire une approximation de π avec la précision de cette estimation.

5 Théorème de Wilson-Hilferty [4 points]

Soit une variable aléatoire X suivant une loi $\chi^2(k)$. On souhaite vérifier le théorème limite de Wilson-Hilferty, qui établit que, asymptotiquement en k,

$$\sqrt[3]{\frac{X}{k}} \approx \mathcal{N}\left(1 - \frac{2}{9k}, \frac{2}{9k}\right)$$

1. **[.5]** On rappelle que si $Y_1, \ldots, Y_k$ est un échantillon de loi $\mathcal{N}(0, 1)$ alors la v.a. $X = \sum Y_i^2$ suit une loi $\chi^2(k)$. Écrire une fonction `rchi2(n,k)` permettant de simuler n réalisations indépendantes de la loi $\chi^2(k)$.
2. **[1]** Dans cette question, on choisit $k = 5$ et $n = 1000$. Vérifier le théorème de Wilson-Hilferty, d'abord graphiquement, puis à l'aide de la fonction `ks.test()`.

3. [**1**] Étudier le comportement de la p-value associée au test de normalité de Kolmogorov-Smirnov quand la valeur de k varie entre 1 et 50. On pourra proposer une représentation graphique.

4. [**1.5**] Le théorème central limite établit que, asymptotiquement en k,

$$X \approx \mathcal{N}(k, 2k)$$

Répéter les questions 2 et 3 pour le théorème central limite. Comparer la vitesse de convergence de ces deux théorèmes limites.

6 Extrêmes [4 points]

Soient $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ deux échantillons de lois $\mathcal{N}(0,1)$ et $\mathcal{N}(1,2)$ respectivement.

1. [**1**] Donner la densité de $\max_i X_i$ *(question indépendante de la question suivante)*

2. [**.5**] En déduire que la loi de $\{\max_i X_i - \max_j Y_j\}$ n'est pas disponible sous forme explicite *(question indépendante de la question suivante)*

3. [**1.5**] Indiquer comment simuler la loi de $Z = \{\max_i X_i - \max_j Y_j\}$. Ecrire la fonction R correspondante et fournir un histogramme de cette loi pour $n = 111$ et $m = 77$.

4. [**1**] Tester l'adéquation de cette loi à une loi normale de moyenne nulle et de variance inconnue.

7 Simulation de loi [4 points]

Etant donnée la densité

$$f(x) \propto \exp\{-x^2\sqrt{x}\}[\sin(x)]^2\,,$$

$0 < x < \infty$, correspondant à la variable aléatoire X, on considère les lois de densité

$$g_1(x) = \frac{1}{2}e^{-|x|}, \quad g_2(x) = \frac{1}{2\pi}\frac{1}{1+x^2/4}, \quad g_3(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,.$$

1. [**2**] Pour des échantillons produits suivant chacune de ces trois lois, estimer par une expérience de Monte-Carlo le nombre minimal M de simulations nécessaire pour obtenir une précision de deux décimales sur l'évaluation de $\mathbb{E}_f[X]$. Comparer les valeurs ainsi obtenues en expliquant les possibles divergences entre valeurs numériques.

2. [**1**]En utilisant g_1 comme proposition, générer un échantillon suivant f par acceptation-rejet et en déduire une estimation de la constante de normalisation de f. Comment évaluer la précision de cette estimation ?

3. [**1**] Comparer aux estimations obtenues en utilisant g_2 et g_3.

8 Mélanges de normales [3 points]

Lorsque f correspond à un mélange de lois normales

$$p\,\mathcal{N}(0,1) + (1-p)\mathcal{N}(3,1),$$

on peut simuler un échantillon Z de distribution f de taille `nech` de la manière suivante :

```
p=0.6
nech=1000
Z=replicate(nech, rnorm(1, sample(c(0,3),prob=c(p,1-p)),1))
```

On s'intéresse à la quantité $P = \mathbb{P}(X > 4)$.

a. Déterminer la valeur exacte de P :

1. **[.5]** par une intégration numérique;
2. **[1]** en utilisant le fait que f est un mélange de lois normales.

On estime maintenant la quantité $P = \mathbb{P}(X > 4)$ en utilisant plusieurs méthodes de Monte-Carlo.

b. **[.5]** Simuler $N = 500$ échantillons, chacun de taille `nech`, suivant f. En déduire un intervalle de confiance à 95% sur P en fournissant le code R utilisé.

c. **[1]** A partir d'un unique échantillon de taille `nech` simulé suivant f, et en remarquant que P correspond à l'estimation de la fonction de répartition de f, donner un intervalle de confiance sur P à 95%. Commentez la différence avec l'intervalle précédent.

9 Théorèmes Limites [5 points]

Soit $X_1, \ldots, X_n$ un n-échantillon distribué sous la loi de fonction de répartition F (et de densité correspondante f). Soit $\alpha \in]0,1[$ et $\hat{q}_n$ est l'estimateur empirique du quantile $F^{-1}(\alpha)$.

On voudrait vérifier le théorème limite suivant:

$$\sqrt{n}(\hat{q}_n - F^{-1}(\alpha)) \xrightarrow[\mathcal{D}]{n\to\infty} \mathcal{N}(0, \sigma_q^2) \tag{9.1}$$

où la variance de la loi normale asymptotique est égale à

$$\sigma_q^2 = \frac{\alpha(1-\alpha)}{[f(F^{-1}(\alpha))]^2}.$$

Dans la suite du problème, on supposera que F est la fonction de répartition de $\mathcal{N}(0, 300)$, c'est-à-dire une normale de moyenne 0 et de variance 300, et donc de densité

$$f(x) = \frac{1}{\sqrt{600\pi}} \exp(-\frac{x^2}{600}).$$

On prendra $\alpha = 0.11$.

1. **[1]** Simuler $N = 10^3$ fois des échantillons de taille $n = 10^2$ de la loi $\mathcal{N}(0, 100)$, et calculer $\hat{q}_n$ pour chacun de ces échantillons. On pourra écrire une fonction R prenant comme argument (N, n) et fournissant les $\hat{q}_n$.
2. **[2]** Tracer l'histogramme des $\hat{q}_n$ ainsi obtenus et tester l'adéquation à la loi asymptotique donnée en (9.1) par un test de Kolmogorov-Smirnov.
3. **[2]** Donner un intervalle de confiance à 95% sur l'espérance de $\hat{q}_n$.