ENS Cachan 2016-2017
Probabilités
L3

DM 2

Ce DM comporte des exercices mixant théorie et implémentation. Les parties théoriques seront rendues sous forme manuscrite (ou imprimée). L'implémentation est à rendre en respectant les consignes distribuées avec le fichier squelette. Il est fortement conseillé de tester ses fonctions (en testant différentes entrées, en traçant des graphes, etc), même si le code utilisé pour tester, ainsi que les résultats, ne sont pas à rendre. Le code est à rendre par mail à `axel.davy@ens-cachan.fr` avec comme entête [RENDUL3], et le manuscrit à rendre en cours ou TD, les deux le 21 Avril 2017.

Exercice 1.

1. Soit $Y_i \sim \mathcal{N}(\mu_i, 1), i = 1 \dots n$ des variables aléatoires indépendantes. Montrer que la distribution de $\sum_{i=1}^n Y_i^2$ ne dépend que de $\lambda = \sum \mu_i^2$. On nommera cette loi la loi du chi2 non centrée de degré n et de paramètre λ.
2. Montrer que $Y_1 Y_2$ est de distribution identique à $\frac{(Z_1^2 - Z_2^2)}{2}$ avec Z_1, Z_2 gaussiennes indépendantes de variance 1 et moyennes $\frac{\mu_1+\mu_2}{\sqrt{2}}$ et $\frac{\mu_1-\mu_2}{\sqrt{2}}$.

Exercice 2. Soit Z une variable aléatoire de loi du chi2 non centrée de degré p et paramètre λ, avec λ inconnu. On suppose p connu. On souhaite construire un intervalle de confiance sur λ. Il n'y a pas unicité des choix d'intervalles de confiance, et donc dans les questions suivantes, plusieurs solutions sont possibles (et seront acceptées, sous réserve qu'elles soient raisonnables).

1. Soit $X \sim \mathcal{N}(\mu, 1)$. Donnez un intervalle de confiance $1 - \alpha$ sur μ à partir de réalisations $x_1, \dots, x_N$. On supposera que l'on a accès aux fonction quantiles de la loi normale. Implementez une fonction `gaussian_ci_nu(samples_x, alpha)` qui retourne la réalisation de cet intervalle pour `samples_x` un tableau contenant des réalisations de X.
2. En déduire un intervalle de confiance (contenant 0) sur $|\mu|$. Supposons que l'on ait désormais accès uniquement à des réalisations $y_1, \dots, y_N$ avec $Y = X^2$, proposez à partir de l'intervalle précédent un intervalle de confiance (pas forcément très bon) sur $\lambda = \mu^2$.
Implementez de même `chi2_noncentral_df1_ci_lambda(samples_y, alpha)`.
3. Supposons que l'on ait deux intervalles de confiance $[f_1, f_2]$ et $[g_1, g_2]$ pour des paramètres f et g de degrés de confiance $1 - \alpha_f$ et $1 - \alpha_g$. De quel degré de confiance est $[f_1, f_2] \times [g_1, g_2]$ pour (f, g) si f et g sont totalements indépendants (ainsi que les données utilisées pour les intervalles) ? Si l'on a possible dépendance, quel degré de confiance as-t'on au moins ?

4. Soit A une variable aléatoire de loi du chi2 centrée de degré $p-1$ ($A = \sum_{i=1}^{p-1} B_i^2, B_i \sim \mathcal{N}(0,1)$). On a $Z = A + Y$. Supposons que l'on ait une seule réalisation de Z : z_1. On peut obtenir un intervalle de confiance sur la réalisation y_1 de Y à partir des quantiles de A. En déduire un intervalle de confiance $1-\alpha$ (contenant 0) sur λ. Implementez de même `chi2_noncentral_dfp_ci_lambda(sample_z, p, alpha)`.
5. Les articles [2, 1] proposent différents algorithmes pour obtenir un intervalle de confiance sur λ à partir d'une seule réalisation z_1. Dans l'article [2], Page 9, la Table 1 présente les résultats retournés pour p fixé pour différentes entrées y (où ici $y = \sqrt{z_1}$). Pour $p = 7$ et $y = 1$, une des méthodes retourne un intervalle vide. Expliquez en quoi ce n'est pas une contradiction. Comparer les intervalles obtenus avec votre algorithme avec le tableau, pour quelques valeurs.

Exercice 3. L'algorithme suivant est un algorithme de simulation de variable aléatoire par méthode de rejet. On suppose que l'on sait tirer des variables aléatoires uniformes et des variables aléatoires d'une distribution de densité g sur $\mathbb{R}^d$. On souhaite générer n éléments de distribution de densité f, avec g choisi tel que $\frac{f}{g}$ borné par une constante c (plus c est petit, plus g est bien choisi).

```
rejection_sampling
Entrées: f, g, g_sampler, c, n
Jusqu'à avoir n éléments, faire:
  . tirer un élement x de la distribution définie par g (x <- g_sampler())
  . tirer u de la loi uniforme
  . si u <= f(x)/(cg(x)) alors on garde x
```

1. On cherche à montrer que cet algorithme retourne bien des élements de la distribution de densité f. Montrer les propriétés suivantes :
   — Soit X variable aléatoire de distribution g sur $\mathbb{R}^d$ et $U \sim \mathcal{U}([0,1])$, X et U indépendantes, et $c > 0$, alors $(X, cUg(X))$ est uniforme sur $\{(x, v) \in \mathbb{R}^{d+1}, 0 \leq v \leq cg(x)\}$.
   — Soit X variable aléatoire sur $\mathbb{R}^d$ et V sur $\mathbb{R}$, tels que (X, V) soit uniforme sur $\{(x, v) \in \mathbb{R}^{d+1}, 0 \leq v \leq cg(x)\}$, où $c > 0$ et g densité de variable aléatoire par rapport à μ sur $\mathbb{R}^d$, alors la distribution de X est $g.d\mu$.
   — Soit $X_1, \ldots, X_n$ variables iid sur $\mathbb{R}^d$. Soit A borélien de $\mathbb{R}^d$ tel que $\mathbb{P}(X_1 \in A) = p > 0$. Soit Y :"La première variable X_i qui est dans A", alors $Y = X \mid X \in A$, (ie $\forall B \in \mathcal{B}(\mathbb{R}^d), \mathbb{P}(Y \in B) = P(X \in B | X \in A)$), et si on note $I = \inf\{i, X_i \in A\}$, alors I est une variable aléatoire géométrique de paramètre p.

   En utilisant ces propriétés, justifiez que l'algorithme retourne bien des élements tirés selon f.
2. Soit $a > 0$, on note $\mathcal{N}_{\leq a}(\mu, \sigma^2)$ la loi normale tronquée. La pdf est $f_{a,\mu,\sigma^2}(x) = K g_{\mu,\sigma^2}(x) \mathbb{1}_{\leq a}(x)$, avec g_{μ,σ^2} la densité de la loi $\mathcal{N}(\mu, \sigma^2)$. Déterminez K en fonction de Φ la fonction de répartition de la loi normale standard.
3. Implémenter `rejection_sampling(f, g, g_sampler, c, n)`, et écrire une fonction
   `truncated_gaussian_sampling(mu, sigma, n, a)` qui retourne n élements tirés selon $\mathcal{N}_{\leq a}(\mu, \sigma^2)$, utilisant `rejection_sampling` .

4. Il existe un algorithme de rejet beaucoup plus simple pour simuler la loi normale tronquée. Montrer qu'il est correct et l'implémenter également en le nommant `truncated_gaussian_sampling_simple(mu, sigma, n, a)`.
5. Déterminez l'espérance et la variance. Ecrire une fonction `test_inequalities()` dans laquelle vous testez les inégalités de Hoeffding et Bennett sur la distribution pour différents paramètres. Commentez.

Références

[1] John T Kent. Calculations for the noncentral chi distribution.

[2] John T Kent and Timothy J Hainsworth. Confidence intervals for the noncentral chi-squared distribution. *Journal of statistical planning and inference*, 46(2) :147–159, 1995.