Contrôle des connaissances du cours «Bruits et Signaux» de l'Ecole Doctorale d'Astrophysique.

Corrigé

Le mercredi 19 décembre 2001 de 14h à 16h amphi de l'IAP.

Le contrôle est noté sur 20, le barème est indiqué à la fin de chaque groupe de questions. Le total des points est de 23 points. Les exercices sont indépendants. Le polycopié et les notes de cours sont les seuls documents autorisés.

Du bon usage de l'espérance mathématique.

De nombreux problèmes relevant du traitement des données se trouvent grandement simplifiés si l'on sait bien manipuler l'opérateur espérance mathématique. Le but de cet exercice est de vous aider à faire le point sur son usage. On ne demande pas une longue justification des réponses.

① Soient X et Y deux variables aléatoires, à quelles conditions peut-on écrire que $\mathrm{E}\{X+Y\} = \mathrm{E}\{X\} + \mathrm{E}\{Y\}$?

② De même, à quelles conditions peut-on écrire que $\mathrm{E}\{XY\} = \mathrm{E}\{X\}\,\mathrm{E}\{Y\}$?

③ De même pour $\mathrm{E}\{X^2\} = (\mathrm{E}\{X\})^2$?

④ Donner l'expression de la variance de la variable aléatoire X à l'aide de l'opérateur espérance mathématique.

⑤ Donner l'expression de la covariance des variables aléatoires X et Y à l'aide de l'opérateur espérance mathématique.

⑥ Une variable aléatoire X a pour moyenne μ et pour variance σ^2, que vaut la moyenne de X^2 ? On demande une expression en fonction de μ etσ^2.

($1+1+1+1+1+1 = 6$ points.)

Corrigé.

❶ L'opérateur espérance est linéaire, pour que la propriété soit vraie, il suffit que les espérances existent, c'est-à-dire que $\mathrm{E}\{|X|\} < \infty$ et $\mathrm{E}\{|Y|\} < \infty$ ($\mathrm{E}\{X+Y\} < \infty$ découle de l'inégalité triangulaire). En particulier on ne demande pas que les variables X et Y soient indépendantes.

Par la suite on supposera que toutes les espérances évoquées existent.

❷ Si le variables sont non-corrélées (voir questions suivantes). À plus forte raison si elles sont indépendantes.

❸ Si Var$(X) = 0$, c'est-à-dire si la variable X est presque-partout égale à une constante (voir question suivante).

❹ Var$(X) = \mathrm{E}\{(X - \mathrm{E}\{X\})^2\} = \mathrm{E}\{X^2\} - (\mathrm{E}\{X\})^2$.

❺ Cov$(XY) = \mathrm{E}\{(X - \mathrm{E}\{X\})(Y - \mathrm{E}\{Y\})\} = \mathrm{E}\{XY\} - \mathrm{E}\{X\}\,\mathrm{E}\{Y\}$.

❻ De Var$(X) = \mathrm{E}\{X^2\} - (\mathrm{E}\{X\})^2$ on tire : $\mathrm{E}\{X^2\} = \sigma^2 + \mu^2$.

Loi suivie par l'empan.

Cet exercice aborde un calcul classique semblable à l'un de ceux que vous pourriez être amenés à effectuer un jour ou l'autre.

Soit $(X_1, X_2, \ldots, X_n)$ un échantillon i.i.d (indépendant et identiquement distribué). La fonction de répartition des X_i est $F(x)$. On pose :

$$X_{(1)} = \min_i(X_1, X_2, \ldots, X_n)\,,$$
$$X_{(n)} = \max_i(X_1, X_2, \ldots, X_n)\,.$$

On appelle *empan* de l'échantillon, la quantité : $U = X_{(n)} - X_{(1)}$, c'est-à-dire la plus grande étendue couverte par l'échantillon. Le problème est de calculer la loi de U.

① Comme toujours, lorsque l'on a affaire à un problème de calcul de la loi d'une fonction de deux variables aléatoires, il faut tout d'abord déterminer la fonction de répartition du couple formé par ces variables.
Quelle est la définition de la fonction de répartition du couple $(X_{(1)}, X_{(n)})$? On notera $G(x, y)$ cette fonction.

② Les variables $X_{(1)}$ et $X_{(n)}$ sont-elles indépendantes ?

③ Pouvez-vous justifier l'écriture :

$$G(x, y) = \Pr\{X_{(n)} \le y\} - \Pr\{X_{(1)} > x, X_{(n)} \le y\}\ ?$$

④ Donnez l'expression de $\Pr\{X_{(n)} \le y\}$ en fonction de $F(y)$.

⑤ Donnez l'expression de $\Pr\{X_{(1)} > x, X_{(n)} \le y\}$ dans le cas où $x > y$.

⑥ Donnez l'expression de $\Pr\{X_{(1)} > x, X_{(n)} \le y\}$ dans le cas où $x \le y$.

⑦ Donnez l'expression de la fonction de répartition G du couple $(X_{(1)}, X_{(n)})$.

⑧ Si les X_i admettent une densité de probabilité f, le couple $(X_{(1)}, X_{(n)})$ admet aussi une densité g. Donnez l'expression de cette densité.

⑨ Donnez alors l'expression de la densité de probabilité $h(u)$ de la variable $U = X_{(n)} - X_{(1)}$.

⑩ On suppose maintenant que les X_i suivent la loi uniforme entre 0 et 1. À l'aide de l'expression trouvée ci-dessus, calculez $h(0)$ et $h(1)$.
Finalement, donnez l'expression générale de $h(u)$.

($1 + 1 + 1 + 1 + 1 + 1 + 2 + 1 + 2 + 2 + 2 = 14$ points.)

Corrigé.

❶ C'est une simple question de cours, on a :

$$G(x, y) = \Pr\{X_{(1)} \le x, X_{(n)} \le y\}\,.$$

❷ Il n'y a aucune raison pour qu'elles soient indépendantes. On se contentera de cette réponse intuitive mais, pour qui veut en savoir plus, il suffira de remarquer que : si $x > y$ alors $G(x, y) = \Pr\{X_{(n)} \le y\}$ (pour que le minimum soit inférieur à x alors que $x > y$, il faut et il suffit que le maximum soit inférieur à y). Si les variables étaient indépendantes, on aurait toujours : $G(x, y) = \Pr\{X_{(1)} \le y\} \Pr\{X_{(n)} \le y\}$, ces deux expressions de $G(x, y)$ pour $x > y$ sont en général différentes, les variables$X_{(1)}$ et $X_{(n)}$ ne peuvent pas être indépendantes.

❸ Il suffit pour cela de considérer l'événement : $\{X_{(n)} \leq y\}$, il se scinde en deux événements disjoints $\{X_{(1)} \leq x, X_{(n)} \leq y\}$: et $\{X_{(1)} > x, X_{(n)} \leq y\}$. On rappelle que la notation $\{A, B\}$ veut dire que les événements A et B sont réalisés simultanément. D'où, d'après la règle d'additivité des probabilités :

$$\Pr\{X_{(n)} \leq y\} = \Pr\{X_{(1)} \leq x, X_{(n)} \leq y\} + \Pr\{X_{(1)} > x, X_{(n)} \leq y\}\,,$$

et la formule demandée.

❹ Il s'agit aussi d'une question de cours, la loi de maximum est $\Pr\{X_{(n)} \leq y\} = [F(y)]^n$. Pour que le maximum d'un échantillon soit inférieur ou égal à y il faut et il suffit que toutes les variables de l'échantillon soient inférieures ou égales à y.

❺ Si $x > y$, alors $\Pr\{X_{(1)} > x, X_{(n)} \leq y\} = 0$ car le minimum ne peut pas être supérieur au maximum.

❻ Pour que $\{X_{(1)} > x, X_{(n)} \leq y\}$ alors que $x \leq y$, il faut et il suffit que toutes les variables X_i soient comprises entre x et y. Pour une variable on a : $\Pr\{X_i > x, X_i \leq y\} = F(y) - F(x)$, et pour les n variables indépendantes de l'échantillon il vient : $\Pr\{X_{(1)} > x, X_{(n)} \leq y\} = [F(y) - F(x)]^n$.

❼ En réunissant les résultats précédents, il vient :

$$G(x, y) = \begin{cases} [F(y)]^n & \text{si } x > y\,; \\ [F(y)]^n - [F(y) - F(x)]^n & \text{si } x \leq y\,. \end{cases}$$

❽ Par définition $g(x, y) = \partial^2 G(x, y)/\partial x \partial y$. Il vient :

$$g(x, y) = \begin{cases} 0 & \text{si } x > y\,; \\ n(n-1)[F(y) - F(x)]^{n-2} f(x) f(y) & \text{si } x \leq y\,. \end{cases}$$

❾ Pour obtenir la densité h de la variable $U = X + Y$, on a la formule générale :

$$h(u) = \int_{-\infty}^{+\infty} f(x, x+u)\, dx\,,$$

où f est la densité du couple (X, Y). Dans notre cas, cette formule s'écrit :

$$h(u) = \int_0^{\infty} g(x, x+u)\, dx\,.$$

En remplaçant g par sa valeur, on trouve :

$$h(u) = \begin{cases} 0 & \text{si } u \leq 0\,; \\ n(n-1) \int_0^{\infty} [F(x+u) - F(x)]^{n-2} f(x) f(x+u)\, dx & \text{si } u > 0\,. \end{cases}$$

❿ La fonction de répartition de la loi uniforme entre 0 et 1 est :

$$F(x) = \begin{cases} 0 & \text{si } x < 0\,; \\ x & \text{si } 0 \leq x \leq 1\,; \\ 1 & \text{si } x > 1\,, \end{cases}$$

et sa densité de probabilité est :

$$f(x) = \begin{cases} 0 & \text{si } x \notin [0,1]\,; \\ 1 & \text{si } x \in [0,1]\,. \end{cases}$$

Pour $u = 0$, on a $F(x+u) = F(x)$ et $h(0) = 0$.
Pour $u = 1$, on a $f(x+u) = 0$, sauf en $x = 0$ qui est de mesure nulle, et $h(1) = 0$ également.
En fait, l'intégrale permettant de calculer $h(u)$ se réduit à une intégrale de 0 à $1-u$, en dehors de cet intervalle la fonction à intégrer est nulle parce que $f(x+u) = 0$. D'où : $h(u) = 0$ si $u \notin [0,1]$ et si $n \in [0,1]$, on a : $g(u) = \int_0^{1-u} (x+u-x)^{n-2}\,dx$, d'où

$$h(u) = \begin{cases} 0 & \text{si } x \notin [0,1]\,; \\ n(n-1)u^{n-2}(1-u) & \text{si } u \in [0,1]\,. \end{cases}$$

Cette densité de probabilité est illustrée sur la figure 1 pour $n = 2, \ldots, 30$.

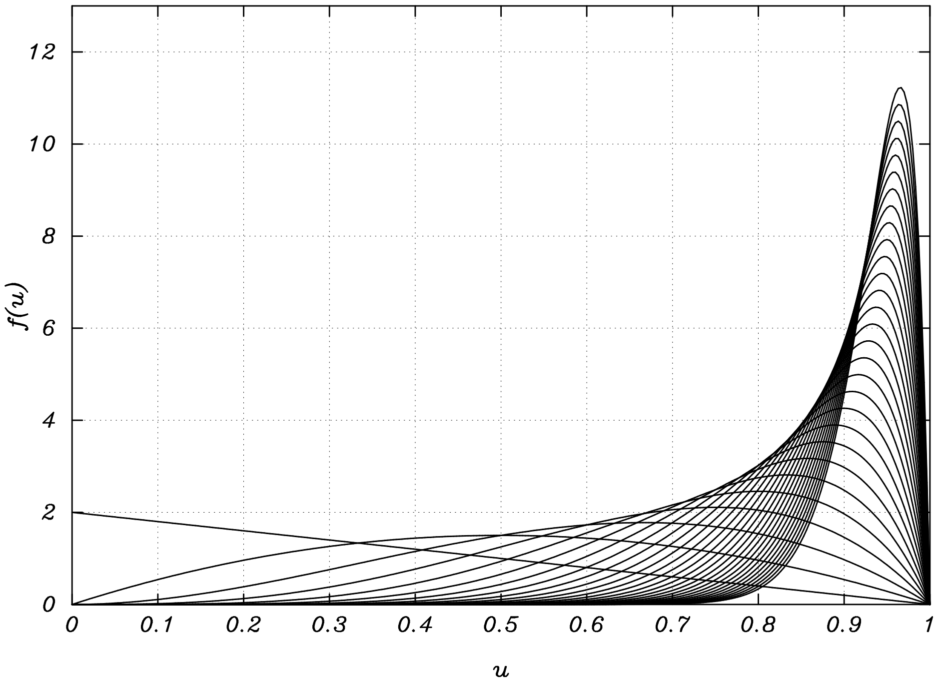

FIG. 1 – Densité de probabilité de l'empan d'un échantillon issu d'une loi uniforme entre 0 et 1, lorsque la taille de l'échantillon passe de $n = 2$ à $n = 30$.

La fonction de répartition de U est donné par :

$$H(u) = \begin{cases} 0 & \text{si } u < 0\,; \\ nu^{n-1} - (n-1)u^n & \text{si } u \in [0,1]\,; \\ 1 & \text{si } u > 1\,. \end{cases}$$

Cette fonction de répartition est illustrée sur la figure 2 pour $n = 2, \ldots, 30$.

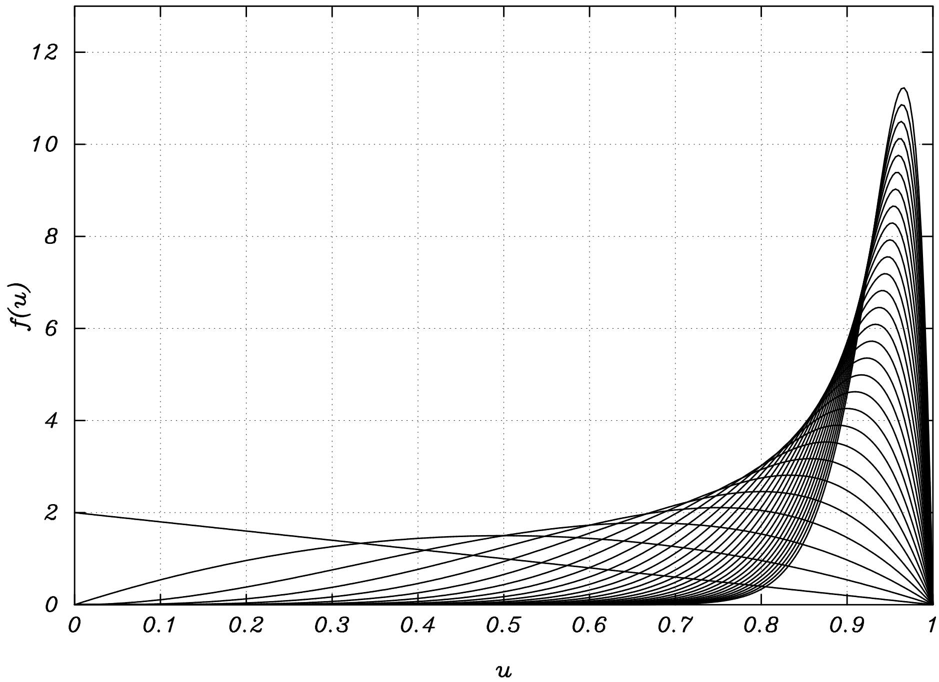

FIG. 2 – Fonction de répartition de l'empan d'un échantillon issu d'une loi uniforme entre 0 et 1, lorsque la taille de l'échantillon passe de $n = 2$ à $n = 30$.

Amélioration de la borne de Rao-Cramér

Soit $(X_1, \dots, X_n)$ un échantillon de taille n et une statistique $T = \phi(X_1, \dots, X_n)$ visant à estimer un paramètre $t = \tau(\theta)$. On suppose que l'estimation de t par T est non-biaisé et que l'on se trouve dans le cas régulier, c'est-à-dire où le domaine d'intégration des espérances mathématiques ne dépend pas de θ.

On sait alors que l'erreur quadratique moyenne d'estimation est égale à la variance de T et que celle-ci est plus grande que la limite de Rao-Cramér. Plus précisément on a :

$$\mathrm{E}\{(T - \tau(\theta))^2\} = \mathrm{Var}(T) \geq \frac{[\tau'(\theta)]^2}{I_n(\theta)} ,$$

où $I_n(\theta)$ est l'information de Fisher, par rapport à l'estimation θ, contenue dans l'échantillon et qui, dans le cas régulier, vaut :

$$I_n(\theta) = \mathrm{E}\{-\frac{\partial^2}{\partial\theta^2} \ln L\} .$$

La fonction L est la fonction de vraisemblance de l'échantillon.

① On remarque que cette borne serait améliorée, c'est-à-dire plus proche de la borne optimale, si on remplaçait I_n par une expression plus petite. Imaginons, par exemple, que l'on ignore toutes les valeurs de l'échantillon et que l'on ne connaisse que la statistique T. Tout se passe alors comme si l'échantillon de taille n était remplacé par un échantillon de taille 1 composé de la seule valeur T. Il doit bien y avoir une information contenue dans cet échantillon vis-à-vis de l'estimation de θ, soit I_T cette information.

D'après-vous, I_T est-elle inférieure, égale ou supérieure à I_n ? Ne donnez qu'un réponse intuitive en la justifiant très brièvement.

② Comment vous y prendriez-vous, en pratique, pour calculer I_T ?

($1 + 2$ points.)

Corrigé.

❶ Ne connaître l'échantillon que par l'intermédiaire d'une fonction de l'échantillon représente sans aucun doute une perte d'information, on doit avoir $I_T(\theta) \leq I_n(\theta)$. Ainsi, si l'on remplace I_n par I_T, on obtiendra une valeur plus réaliste de la variance minimum de T.

❷ Il faut calculer la fonction de vraisemblance de l'échantillon de taille 1 composé de la seule valeur T. Il faut donc calculer la densité de probabilité, disons $g(t)$, de la variable T. Les méthodes classiques de changement de variables aléatoires permettent de faire ce calcul, par exemple en posant :

$$\begin{aligned} T &= \phi(X_1, \ldots, X_n)\,, \\ T_2 &= X_2\,, \\ &\ldots \\ T_n &= X_n\,. \end{aligned}$$

Tout calcul fait on doit obtenir :

$$L(x_1, \ldots, x_n, \theta) = g(t, \theta) h(x_2, \ldots, x_n | t, \theta) \Big| \frac{\partial \phi}{\partial x_1} \Big|\,,$$

où $g(t, \theta)$ est la densité de probabilité de T et h la densité de probabilité de $(x_2, \ldots, x_n)$ connaissant T. On a :

$$I_T = \mathrm{E}\{-\frac{\partial^2}{\partial \theta^2} \ln g(t, \theta)\}\,,$$

d'où l'on tire :

$$I_n(\theta) = I_T(\theta) - \mathrm{E}\{\frac{\partial^2}{\partial \theta^2} \ln h\}\,.$$

Mais h est une densité de probabilité, on donc dans le cas régulier :

$$-\,\mathrm{E}\{\frac{\partial^2}{\partial \theta^2} \ln h\} = \mathrm{E}\{\big(\frac{\partial}{\partial \theta} \ln h\big)^2\} \geq 0\,,$$

qui montre bien que $I_n(\theta) \geq I_T(\theta)$.