Université des Sciences et Technologies de Lille
U.F.R. de Mathématiques Pures et Appliquées

IS Math314 **Année 2010**

Examen, 1er juin 2010, durée 3 heures.

- Ce sujet comporte **4 pages**, dont une table de la loi normale.
- Le barème indiqué est là pour vous aider à gérer votre temps et n'a pas valeur contractuelle.
- Documents autorisés : polycopié du cours IPE, polycopié du cours d'IS, dictionnaire bilingue pour étudiants étrangers.
- Calculatrices autorisées.

Ex 1. *Mise en confiance ? (3 points)*

Dans un centre de tri postal, on voudrait connaître la proportion p de lettres affranchies au tarif « lettre prioritaire 20g » et dont la masse est strictement supérieure à 20g. Dans ce but, on prélève 100 lettres au hasard et on pèse chacune. Les résultats obtenus sont rassemblés dans le tableau suivant.

20,01	19,85	20,04	16,49	19,60	17,80	19,15	17,57	20,67	15,98
16,92	13,53	19,12	20,12	17,74	17,52	17,28	19,11	21,02	17,65
19,15	16,38	18,91	15,90	18,85	21,37	19,73	19,29	16,88	15,69
18,82	16,44	17,11	15,63	23,43	18,29	18,24	18,77	17,21	13,02
18,25	18,45	19,02	18,57	18,89	20,79	15,93	17,02	16,13	14,05
19,69	20,74	16,28	17,53	17,88	18,34	15,81	17,17	20,20	17,20
15,70	15,60	19,41	18,48	13,25	20,02	19,97	18,41	19,60	15,97
19,46	19,95	17,25	15,07	18,45	19,48	16,31	20,97	17,37	18,75
16,30	14,73	14,57	14,26	20,77	20,43	18,78	17,96	16,38	19,29
19,08	20,88	18,23	22,19	15,96	20,21	20,44	18,03	19,21	19,32

Proposez deux intervalles de confiance pour p au niveau 95% en expliquant comment vous les avez obtenus et en indiquant les théorèmes qui les justifient.

Ex 2. *Serait-ce de l'acharnement ? (4 points)*

Soient X_1, X_2, X_3, trois variables aléatoires réelles définies sur le même espace probabilisé, indépendantes et de même loi gaussienne $\mathfrak{N}(0,1)$.

1) Pour chacun des vecteurs aléatoires suivants, indiquez s'il est gaussien en justifiant votre réponse.

$$U = (X_1, X_2, X_3), \quad V = (X_1, X_2, X_2, X_2 + X_3), \quad W = (X_1, X_2, X_3, 7 - X_1).$$

2) Donnez la matrice de covariance de chacun de ces vecteurs, en évitant de détailler tous les calculs, mais en donnant les justifications pertinentes.

Ex 3. *Autour d'un théorème central (5 points)*

Soient $(X_k)_{k\geq 1}$ et $(U_k)_{k\geq 1}$ deux suites de variables aléatoires définies sur le même espace probabilisé. On suppose que ces deux suites sont indépendantes l'une de l'autre et que :

a) les X_k sont indépendantes et de même loi, de carré intégrable et $\mathbf{E}X_1^2 = \tau^2$ avec $\tau > 0$;

b) les U_k sont indépendantes et de même loi, de carré intégrable, d'espérance nulle et de variance σ^2 avec $\sigma > 0$.

On définit pour tout $n \geq 1$ les variables aléatoires :

$$S_n = \sum_{k=1}^{n} U_k X_k, \qquad V_n = \sum_{k=1}^{n} U_k^2.$$

1) Que vaut $\mathbf{E}(U_1 X_1)$? Calculez $\mathrm{Var}(U_1 X_1)$ en fonction de σ et τ.

2) Montrez que $n^{-1/2} S_n$ converge en loi vers une gaussienne dont vous préciserez les paramètres. Étudiez la convergence presque sûre de $(1 + V_n)/n$.

3) Déduisez de ce qui précède que

$$W_n = \sqrt{n} \frac{S_n}{1 + V_n}$$

converge en loi vers une gaussienne dont vous préciserez les paramètres.

Ex 4. *Séparation d'un mélange de lois (8 points)*

On définit sur $\mathbb{R}$ la fonction :

$$F = pG + (1 - p)H,$$

où G et H sont des fonctions de répartition, et p un réel de $]0, 1[$. On suppose que G est la fonction de répartition d'une loi discrète de support l'ensemble *fini* inconnu D (autrement dit, si Y a pour f.d.r. G, $P(Y \in D) = 1$). G est donc une fonction en escaliers avec sauts aux points de D. On suppose aussi que H est continue sur $\mathbb{R}$. Le réel p et les fonctions G, H et F sont inconnus. Le but de cet exercice est de justifier théoriquement une méthode d'estimation [1] de ces inconnues au vu d'un échantillon $(X_1, \ldots, X_n)$ de grande taille de la loi de fonction de répartition F. On rappelle que pour toute fonction de répartition K, on note par $K(x-)$ la limite à gauche de K au point x. On désigne par F_n la *fonction de répartition empirique* construite sur l'échantillon $(X_1, \ldots, X_n)$. La première et la sixième question de cet exercice sont indépendantes de cette méthode d'estimation.

1) Vérifiez que F est bien une fonction de répartition.

1. Il n'est pas nécessaire de connaître le chapitre sur l'estimation pour traiter cet exercice.

2) Expliquez brièvement pourquoi :

$$\forall x \in \mathbb{R}, \quad P(X_1 = x) = F(x) - F(x-) = p(G(x) - G(x-))$$

et

$$p = \sum_{x \in D} (F(x) - F(x-)).$$

3) On définit l'ensemble aléatoire D_n par

$$D_n(\omega) = \left\{ x \in \mathbb{R};\ F_n(\omega, x) - F_n(\omega, x-) > \frac{1}{n} \right\}.$$

Autrement dit, $D_n(\omega)$ est l'ensemble des nombres qui apparaissent plus d'une fois dans la suite de réels $X_1(\omega), \ldots, X_n(\omega)$. Pourquoi ? *Nous admettrons* que D_n est inclus dans D presque-sûrement [2]. On propose alors d'estimer p par la variable aléatoire

$$p_n = \sum_{x \in D_n} \big(F_n(x) - F_n(x-)\big).$$

On esquisse ci-dessous une démonstration de la convergence p.s. de p_n vers p. Explicitez les arguments manquants dans cette preuve et expliquez l'intérêt de sommer sur D au lieu de D_n. Esquisse de preuve :

$$p_n = \sum_{x \in D} \big(F_n(x) - F_n(x-)\big) - R_n \tag{1}$$

$$= \sum_{x \in D} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i = x\}} - R_n. \tag{2}$$

La somme sur D converge presque-sûrement vers p et la v.a. positive R_n se majore facilement par une suite non-aléatoire qui converge vers 0.

4) On pose maintenant pour tout $t \in \mathbb{R}$,

$$p_n G_n(t) = \sum_{\substack{x \in D_n, \\ x \le t}} \big(F_n(x) - F_n(x-)\big),$$

avec la convention $G_n(t) = 0$ lorsque $p_n = 0$. En adaptant la méthode de la question précédente, montrez que pour tout t fixé dans $\mathbb{R}$, $p_n G_n(t)$ converge p.s. vers $pG(t)$ et $F_n(t) - p_n G_n(t)$ converge p.s. vers $(1-p)H(t)$.

5) (Question bonus hors-barème). Montrez que pour la fonction F de cet énoncé (c'est-à-dire n'ayant qu'un ensemble fini D de points de discontinuité), le théorème de Glivenko-Cantelli pour les limites à gauche est vérifié :

$$\sup_{x \in \mathbb{R}} |F_n(x-) - F(x-)| \xrightarrow[n \to +\infty]{\text{p.s.}} 0.$$

En déduire que p.s. les convergences de $p_n G_n$ vers G et de $F_n - p_n G_n$ vers $(1-p)H$ sont *uniformes* sur $\mathbb{R}$.

2. On pourrait le prouver en adaptant la méthode de l'exercice sur les *ex-aequo* d'un échantillon vue en T.D.

6) Un enseignant souhaite illustrer cette méthode d'estimation de p, G et H au cours d'un T.D. sur machine. Pour cela, il lui faut simuler un échantillon de grande taille de la loi de f.d.r. F après avoir choisi p, G, H. Il fournira alors cet échantillon aux étudiants qui eux, ne connaîtront pas ses choix et devront estimer p, G et H. L'enseignant choisit $p = 0,35$, et pour G et H les f.d.r. de lois faciles à simuler, disons la loi binomiale de paramètres 8 et 0,6 pour G et la loi exponentielle de paramètre 2 pour H. Pour simuler une variable aléatoire X de f.d.r. F, il programme alors l'algorithme suivant.

1. On simule une variable aléatoire U de loi uniforme sur $]0,1[$.
2. On teste si $U \leq p$.
3. a) Si oui, on génère X de loi Bin$(8;0{,}6)$;
   b) sinon, on génère X de loi Exp(2).

Justifiez cet algorithme en calculant $P(X \leq t)$ pour t réel quelconque.

Table des valeurs de Φ, f.d.r. de la loi normale standard $\mathfrak{N}(0,1)$ sur l'intervalle $[1,3]$

$$\Phi(x) = P(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(\frac{-t^2}{2}\right) \mathrm{d}t, \quad Z \sim \mathfrak{N}(0,1).$$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.0	0.8414	0.8438	0.8461	0,8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8622
1.1	0.8643	0.8665	0.8687	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0,8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9083	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9193	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9485	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9648	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9895	0.9898	0.9901	0.9903	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9926	0.9928	0.9930	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9944	0.9946	0.9948	0.9949	0,9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9958	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986