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

L3 **Année 2011**

Initiation à la Statistique
D.S. du 25 mars 2011 (durée 2 heures)

- Ce sujet comporte **4 pages**, dont une table partielle de la loi normale.
- Le barème indiqué est là pour vous aider à gérer votre temps et n'a pas valeur contractuelle.
- Documents autorisés : une feuille manuscrite A4 recto verso.
- Calculatrices autorisées.
- La qualité de la rédaction sera un élément important d'appréciation des copies.

Ex 1. *Intervalle de confiance (2 points)*

Le tableau ci-dessous donne un 100-échantillon observé d'une loi de Bernoulli de paramètre inconnu p.

1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	0	0
1	0	0	1	1	0	0	0	0	0	1	0	0	0	1	1	1	0	1	1
0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0

1) Donner un intervalle de confiance au niveau 95 % pour p par la méthode avec *variance majorée*. Quel résultat du cours vous permet de justifier cette méthode ?

Ex 2. *Dé (3 points)*

On lance 3 600 fois un dé équilibré et on s'intéresse au nombre X de « cinq » obtenus. Donnez en justifiant votre réponse, une valeur approchée de $P(578 \leq X \leq 622)$. Indication : commencez par donner la loi *exacte* de X.

Ex 3. *Estimation du paramètre d'une loi géométrique (15 points)*

La loi géométrique de paramètre $\theta \in]0,1[$ est celle du temps d'attente du *premier* succès dans une suite d'épreuves répétées indépendantes avec pour chaque épreuve probabilité de succès θ. Une variable aléatoire X sur l'espace probabilisé $(\Omega, \mathcal{F}, P_\theta)$ suit cette loi si

$$\forall k \in \mathbb{N}^*, \quad P_\theta(X = k) = (1-\theta)^{k-1}\theta. \tag{1}$$

On se propose dans cet exercice d'estimer θ.

En préambule, on rappelle l'énoncé du théorème limite central.

Théorème 1 (théorème limite central)**.** *Soit $(X_k)_{k\geq 1}$ une suite de variables aléatoires définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$, indépendantes, de même loi et de carré intégrable (et non* p.s. *constantes). Notons $\mu := \mathbf{E}X_1$, $\sigma^2 := \operatorname{Var} X_1$ avec $\sigma > 0$ et $S_n = \sum_{k=1}^n X_k$. Alors*

$$S_n^* := \frac{S_n - \mathbf{E}S_n}{\sqrt{\operatorname{Var} S_n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow[n\to+\infty]{\text{loi}} Z, \tag{2}$$

où Z est une variable de loi gaussienne $\mathfrak{N}(0,1)$.

La convergence en loi vers une variable aléatoire Z de fonction de répartition continue (ce qui est le cas de la gaussienne Z ci-dessus), se définit comme la convergence en tout point de la suite des fonctions de répartition. Il résulte donc de ce théorème que pour tous réels $a < b$, $P(a \leq S_n^* \leq b)$ converge quand n tend vers l'infini vers $\Phi(b) - \Phi(a)$ où Φ est la fonction de répartition de Z, calculable numériquement grâce à la table de la loi normale.

1) Calculez $\sum_{k=1}^{+\infty}(1-\theta)^{k-1}\theta$ et en déduire que (1) définit bien une loi de probabilité de variable aléatoire discrète.

2) Vérifiez que

$$\mathbf{E}_\theta X = \frac{1}{\theta}, \tag{3}$$

$$\operatorname{Var}_\theta X = \frac{1-\theta}{\theta^2}. \tag{4}$$

3) On considère désormais le modèle statistique $(\Omega, \mathcal{F}, P_\theta)_{\theta\in]0,1[}$ et un n-échantillon $X_1, \ldots, X_n$ associé. Autrement dit, pour toute valeur de θ, les variables aléatoires X_i sont P_θ-indépendantes et de même loi sous P_θ donnée par (1). On note $S_n = X_1 + \cdots + X_n$ la somme de cet échantillon. Justifiez brièvement les deux convergences suivantes :

$$\forall\theta \in]0,1[, \quad \frac{n}{S_n} \xrightarrow[n\to+\infty]{P_\theta-\text{p.s.}} \theta; \tag{5}$$

$$\forall\theta \in]0,1[, \quad S_n^* := \frac{\theta S_n - n}{\sqrt{n(1-\theta)}} \xrightarrow[n\to+\infty]{P_\theta-\text{loi}} Z, \quad Z \sim \mathfrak{N}(0,1). \tag{6}$$

4) En déduire un intervalle de confiance I au niveau 95% pour θ. Comment peut-on améliorer cet intervalle si on a la certitude que $\theta \geq \frac{3}{4}$?

5) Pour traiter cette question, on admettra la propriété suivante : si $(Z_n)_{n\geq 1}$ et $(Y_n)_{n\geq 1}$ sont deux suites de variables aléatoires définies sur le même espace probabilisé et telles que Z_n converge en loi vers Z et Y_n converge p.s. vers une constante c, alors Y_nZ_n converge en loi vers cZ.

Soit ε_n une suite de réels *strictement positifs* tendant vers 0. Montrez que

$$\forall\theta \in]0,1[, \quad \frac{\sqrt{n(1-\theta)}}{\varepsilon_n + \sqrt{n\left(1 - \frac{n}{S_n}\right)}} \xrightarrow[n\to+\infty]{P_\theta-\text{p.s.}} 1. \tag{7}$$

Déduire alors de (6) la convergence en loi suivante :

$$\forall \theta \in]0,1[, \quad T_n = \frac{\theta S_n - n}{\varepsilon_n + \sqrt{n\left(1 - \frac{n}{S_n}\right)}} \xrightarrow[n\to+\infty]{P_\theta-\text{loi}} Z, \quad Z \sim \mathfrak{N}(0,1). \tag{8}$$

Le seul intérêt de ε_n est d'empêcher que le dénominateur de $T_n(\omega)$ s'annule pour tout ω réalisant l'évènement de probabilité non nulle $\{S_n = n\}$. Pour l'application qui suit, on prendra ε_n suffisamment petit pour être numériquement négligeable, par exemple $\varepsilon_n = 10^{-n}$. Donnez l'intervalle de confiance J pour θ au niveau 95% que l'on peut déduire de (8).

6) Le tableau ci-dessous donne un échantillon observé de taille 200.

1	2	1	1	1	1	1	1	2	1	1	1	1	1	3	2	1	2	1	1
1	1	1	1	1	1	1	1	4	1	1	1	2	1	1	1	1	2	2	1
2	1	1	1	1	1	1	1	2	1	1	1	2	1	1	1	1	1	2	2
2	2	1	1	3	1	4	1	1	1	1	1	1	2	1	2	1	1	1	1
2	1	1	1	1	6	1	1	1	1	2	2	1	1	1	1	1	1	2	1
1	1	1	1	4	1	2	1	1	1	1	1	1	1	1	2	1	2	1	1
1	1	1	3	1	1	1	1	2	1	1	2	2	2	1	1	1	1	2	1
1	1	1	3	2	1	1	1	2	1	3	1	1	1	1	1	1	1	1	2
3	2	1	1	1	2	1	1	2	1	1	1	2	3	2	1	1	1	1	4
1	2	4	2	1	1	1	1	1	1	1	1	1	1	1	1	2	5	1	2

a) On rappelle que la fonction de répartition empirique F_n associée à l'échantillon $X_1, \ldots, X_n$ est définie par

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \le t\}}.$$

Reproduisez et complétez le tableau suivant. Utilisez-le pour dessiner la fonction de répartition empirique associée à cet échantillon. Unité verticale conseillée 10 cm.

Valeur	1	2	3	4	5	6
Fréquence			0,035			

b) Calculez l'estimateur n/S_n de θ sur cet échantillon observé en indiquant comment vous faites pour éviter de rentrer une à une les 200 valeurs dans votre calculatrice.

c) Donnez les intervalles de confiance I et J calculés à partir de cet échantillon observé.

7) Une autre idée pour estimer θ est de se rappeler que c'est le paramètre d'une variable aléatoire de Bernoulli Y_i valant 1 en cas de succès à la i^{e} épreuve et 0 en cas d'échec. En partant de cette idée, pourriez-vous expliquer comment le tableau des $X_k(\omega)$ observés donné à la question 6) permet de reconstituer la suite des résultats $Y_i(\omega)$ des épreuves répétées (on ne demande pas d'écrire les résultats en détail !). Déduisez-en la valeur $\overline{Y}(\omega)$ de la moyenne empirique des $Y_i(\omega)$, puis l'intervalle de confiance à 95% par la méthode avec variance estimée pour θ. Que remarquez vous ?

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

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

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,0	0,841 4	0,843 8	0,846 1	0,848 5	0,850 8	0,853 1	0,855 4	0,857 7	0,859 9	0,862 2
1,1	0,864 3	0,866 5	0,868 7	0,870 8	0,872 9	0,874 9	0,877 0	0,879 0	0,881 0	0,883 0
1,2	0,884 9	0,886 9	0,888 8	0,890 7	0,892 5	0,894 4	0,896 2	0,898 0	0,899 7	0,901 5
1,3	0,903 2	0,904 9	0,906 6	0,908 3	0,909 9	0,911 5	0,913 1	0,914 7	0,916 2	0,917 7
1,4	0,919 3	0,920 7	0,922 2	0,923 6	0,925 1	0,926 5	0,927 9	0,929 2	0,930 6	0,931 9
1,5	0,933 2	0,934 5	0,935 7	0,937 0	0,938 2	0,939 4	0,940 6	0,941 8	0,942 9	0,944 1
1,6	0,945 2	0,946 3	0,947 4	0,948 5	0,949 5	0,950 5	0,951 5	0,952 5	0,953 5	0,954 5
1,7	0,955 4	0,956 4	0,957 3	0,958 2	0,959 1	0,959 9	0,960 8	0,961 6	0,962 5	0,963 3
1,8	0,964 1	0,964 8	0,965 6	0,966 4	0,967 1	0,967 8	0,968 6	0,969 3	0,969 9	0,970 6
1,9	0,971 3	0,971 9	0,972 6	0,973 2	0,973 8	0,974 4	0,975 0	0,975 6	0,976 1	0,976 7
2,0	0,977 2	0,977 8	0,978 3	0,978 8	0,979 3	0,979 8	0,980 3	0,980 8	0,981 2	0,981 7
2,1	0,982 1	0,982 6	0,983 0	0,983 4	0,983 8	0,984 2	0,984 6	0,985 0	0,985 4	0,985 7
2,2	0,986 1	0,986 4	0,986 8	0,987 1	0,987 4	0,987 8	0,988 1	0,988 4	0,988 7	0,989 0
2,3	0,989 3	0,989 5	0,989 8	0,990 1	0,990 3	0,990 6	0,990 9	0,991 1	0,991 3	0,991 6
2,4	0,991 8	0,992 0	0,992 2	0,992 4	0,992 6	0,992 8	0,993 0	0,993 2	0,993 4	0,993 6
2,5	0,993 8	0,994 0	0,994 1	0,994 3	0,994 4	0,994 6	0,994 8	0,994 9	0,995 1	0,995 2
2,6	0,995 3	0,995 5	0,995 6	0,995 7	0,995 8	0,996 0	0,996 1	0,996 2	0,996 3	0,996 4
2,7	0,996 5	0,996 6	0,996 7	0,996 8	0,996 9	0,997 0	0,997 1	0,997 2	0,997 3	0,997 4
2,8	0,997 4	0,997 5	0,997 6	0,997 7	0,997 7	0,997 8	0,997 9	0,997 9	0,998 0	0,998 1
2,9	0,998 1	0,998 2	0,998 2	0,998 3	0,998 4	0,998 4	0,998 5	0,998 5	0,998 6	0,998 6