utbm
université de technologie Belfort-Montbéliard

SQ 20 Final Automne 05

Lundi 16 janvier 2006

Matériel autorisé: une feuille aide-mémoire A4 recto, calculatrice, Tables de statistiques

I. Première partie (7 + 4 + 3 points)

1°) Le temps (en heures par semaine) passé sur Internet par une (grande) population d'étudiants est une variable qu'on supposera normale $\mathcal{N}(M, \sigma^2)$, M et σ² étant des paramètres qu'on cherche à estimer à partir d'un échantillon indépendant $E_n = (X_1, \ldots, X_n)$.

a) Un sondage auprès de 20 étudiants a donné les observations suivantes :

13	11,5	10	7,2	7	8,3	10,1	11,5	7,1	10,3
2,9	4	8,6	4,1	6,8	4,1	5,5	10	4,3	7,7

- Déterminer des estimations ponctuelles de M et de σ².
- En déduire des intervalles de confiance de M et de σ² au niveau 0,95.

b) Quel serait l'intervalle de confiance de M avec la donnée supplémentaire : les 20 étudiants étaient choisis au hasard dans une population de 60 étudiants ?

c) Certains observateurs prétendent que le temps moyen passé sur Internet est de 7 heures, alors que d'autres pensent qu'il est de 9 heures (par semaine).

Effectuer le test H_0 : M = 7 contre H_1 : M = 9. Définir la variable de test, déterminer le domaine d'acceptation, la décision et calculer le risque β.

2°) Soit p la proportion d'étudiants français qui ont un branchement personnel.

a) Une étude sur n = 400 étudiants a montré que 120 d'entre eux avaient leur propre branchement. Déterminer un intervalle de confiance de p au niveau 0,95.

b) Pour quelles valeurs de n aurait-on un intervalle de confiance de longueur inférieure à 0,02 ?

c) Dans un autre pays de l'Union Européenne la même étude a été faite sur 200 étudiants, parmi lesquels 80 étaient connectés chez eux. Est-il raisonnable de penser qu'il y a une différence entre les deux pays ?

3°) On s'intéresse maintenant aux éventuelles différences entre les étudiants et les étudiantes. On sait que les filles sont plus bavardes au téléphone (cf. final de janvier 2003), mais passent-elles moins de temps sur Internet ? On note T_F et T_G les temps passés sur le réseau, les variables étant normales.

Une observation sur 20 étudiantes et 30 étudiants a donné les résultats :

$\overline{x}_F = 6$ et $s_F^2 = 5$ pour les filles, $\overline{x}_G = 8$ et $s_G^2 = 9$ pour les garçons

a) Tester d'abord l'égalité des variances. Calculer une estimation de la variance commune.

b) Peut-on affirmer, au niveau 0,95 que les filles passent moins de temps sur Internet ?

II. Seconde partie (3 + 4 + 2 points)

1°) Un service de vente en ligne met en contact des particuliers : des vendeurs et des acheteurs. Un acheteur passe commande d'un article (livre, disque, DVD, …) et il reçoit une réponse du vendeur après un temps aléatoire T. On supposera que ce temps T en heures est la somme de trois variables exponentielles de même paramètre λ (temps avant consultation du vendeur + temps de réponse + temps avant consultation de l'acheteur).

Remarque : que les trois lois soient exponentielles dans la réalité est possible, mais elles ne doivent certainement pas avoir le même paramètre. Nous considérons que si pour simplifier l'étude.

a) Si X_1 , X_2 … et X_n sont des variables exponentielles $\mathcal{E}(\lambda)$ indépendantes, montrer par récurrence sur n que $S_n = \sum_{k=1}^{n} X_k$ a pour densité $f_n(t) = \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!} 1_{\mathbb{R}_+}(t)$.

b) En déduire la densité $f(t)$ de T.

c) Pour évaluer le paramètre λ, on étudie un échantillon $E_n = \{T_1, T_2, \ldots, T_n\}$ où les T_k suivent la même loi que T. Une observation $e_n(t_1, t_2, \ldots, t_n)$ ayant été obtenue, déterminer un estimateur de maximum de vraisemblance de λ.

d) Déterminer une estimation ponctuelle de λ à partir de 20 valeurs numériques (temps en heures)

7 9,8 14 8 7,8 6,7 25 3,2 10 40 15 20 26 17 14 44 7,4 12,3 8,5 4,3

2°) Le nombre d'acheteurs potentiels qui se rendent sur le site d'un certain vendeur au cours d'une journée est une variable discrète N. On pense que cette variable suit une loi de Poisson $\mathcal{P}(\mu)$, mais il est souhaitable de vérifier cette hypothèse, au risque de 0,05, et de déterminer une estimation du paramètre . Une observation sur un peu plus de trois mois (en réalité 100 jours) a fourni les données :

Nombre de clients	0	1	2	3	4	5	6	≥ 7
Nombre de jours	3	16	21	27	14	7	7	5

a) Déterminer une estimation ponctuelle de μ à partir des données ci-dessus.

b) Vérifier par un test du χ^2 si l'hypothèse sur la variable aléatoire N est acceptable.

3°) Une visite sur le site d'un vendeur se termine par une vente avec une probabilité 0,3. Le site d'un vendeur a été visité 100 fois en décembre 2005. On note V le nombre de ventes de ce mois.

a) Déterminer la loi de V, son espérance et sa variance.

b) Quelle est la probabilité d'avoir entre 25 et 40 ventes en décembre 2005.

Quelques réponses à l'intention de l'infime minorité des étudiants qui n'auraient pas résolu certaines questions : II- 1°) b)- $f(t) = \dfrac{\lambda^3 t^2 e^{-\lambda t}}{2} 1_{R_+}(t)$ 2°) a) estimation de μ : 3

Certaines questions sont un peu plus difficiles, à vous de voir lesquelles, mais le barème scandaleusement généreux en tient compte.

Félicitez-vous d'avoir fait quelque chose d'étrange et d'extravagant qui a brisé la monotonie de votre époque.

Ralph Waldo EMERSON (États-Unis 1803 – 1882)

FINAL SQ20 lundi 16 janvier 2006 - CORRIGÉ -

- PREMIÈRE PARTIE -

1°) a) estimations ponctuelles (calculatrice) : $\bar{x} = 7{,}70$ $s^2 = 2{,}92^2 = 8{,}54$ $n = 20$

b) • Intervalle de confiance de de M : Variable de confiance $Y = \dfrac{\bar{X} - M}{\frac{S}{\sqrt{n}}} \sim T_{n-1} = T_{19}$ (Variance estimée et v.a normale)

table $T_{19} \longrightarrow p(-2{,}093 < Y < 2{,}093) = 0{,}95$

$\Rightarrow \underline{I = \left] \bar{X} - 2{,}093 \frac{S}{\sqrt{20}} , \bar{X} + 2{,}093 \frac{S}{\sqrt{20}} \right[}$ d'observation ($\bar{x} = 7{,}7$, $s^2 = 8{,}54$) $\boxed{I_M =]\,6{,}33 ; 9{,}07\,[}$

• Intervalle de confiance de σ^2 ; variable de confiance $Z = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} = \chi^2_{19}$ (moyenne estimée et v.a. normale)

table $\chi^2_{19} \longrightarrow 0{,}95 = p(8{,}9 < Z < 32{,}8) \Rightarrow$ Intervalle $\underline{I = \left] \frac{19 S^2}{32{,}8} , \frac{19 S^2}{8{,}9} \right[}$ d'observation $\boxed{I_{\sigma^2} =]\,4{,}94 ; 18{,}23\,[}$

b) hypothèse supplémentaire = taille population $N = 60 \rightarrow$ correction d'exhaustivité $1 - \frac{20}{60} = \frac{2}{3}$

variable $Y' = \dfrac{\bar{X} - M}{\frac{S}{\sqrt{n}}\left(1 - \frac{n}{N}\right)} = \dfrac{\bar{X} - M}{\frac{S}{\sqrt{20}} \times \sqrt{\frac{2}{3}}} \sim T_{19} \rightarrow \underline{I'_M = \left] \bar{X} - 2{,}093 \frac{S}{\sqrt{20 \times \frac{3}{2}}} , \bar{X} + 2{,}093 \frac{S}{\sqrt{20 \times \frac{3}{2}}} \right[}$ d'observation $\boxed{I =]\,6{,}58 ; 8{,}82\,[}$

c) Test $H_0 : M = 7$ contre $H_1 : M = 9$

Variable de test $Y_3 = \dfrac{\bar{X} - 7}{\frac{S}{\sqrt{20}}} \sim T_{19}$ table $T_{19} : 0{,}95 \rightarrow 1{,}729 \Rightarrow p(Y_3 < 1{,}729) = 0{,}95$

donc $p\left(\bar{X} < 7 + 1{,}729 \times \sqrt{\frac{8{,}54}{20}}\right) = 1 - \alpha = 0{,}95 \Rightarrow D_0 =]-\infty ; 8{,}13[$ et $7{,}7 \in D_0 \Rightarrow$ $\boxed{\text{on accepte } H_0 : M = 7}$

$\beta = p(\bar{X} < 8{,}13 \mid M = 9) = p\left(\dfrac{\bar{X} - 9}{\frac{S}{\sqrt{20}}} < \dfrac{8{,}13 - 9}{\frac{S}{\sqrt{20}}} \right) = p(T_{19} < -1{,}33) = 1 - p(T_{19} < 1{,}33) = \boxed{0{,}1 = \beta}$

2°) a) 400 étudiants $\begin{cases} p & \text{branchés} \\ 1-p & \cancel{\text{branchés}} \end{cases} \rightarrow$ échantillon $(X_1, X_2, \dots, X_{400})$ $X_k \sim \mathcal{B}(1, p)$

Estimateur de p : $F = \bar{X} = \frac{1}{n} \sum X_k$, estimation $f = \frac{120}{400} = 0{,}3$

Variable de confiance de p : $Y_p = \dfrac{F - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1)$ (th. central limite) avec $p(1-p) \simeq f(1-f) = 0{,}21$

donc $0{,}95 = p(-1{,}96 < Y_p < 1{,}96) = p\left(F - 1{,}96 \sqrt{\frac{p(1-p)}{400}} < p < F + 1{,}96 \sqrt{\frac{p(1-p)}{400}} \right)$

Intervalle de confiance $\underline{I = \left] F - 1{,}96 \sqrt{\frac{p(1-p)}{400}} , F + 1{,}96 \sqrt{\frac{p(1-p)}{400}} \right[}$ d'observation $\boxed{I_p =]\,0{,}255 ; 0{,}345\,[}$

b) Pour n quelconque (assez grand) longueur de l'intervalle $= 2 \times 1{,}96 \sqrt{\frac{0{,}21}{n}} < 0{,}02 \Rightarrow n > 8067{,}4$

<u>taille minimale de l'échantillon $n = 8068$</u>

c) 2 échantillons : $n_1 = 400$ $f_1 = 0{,}3$; $n_2 = 200$ $f_2 = 0{,}4$ $\rightarrow$ estimation de la valeur commune $f = \dfrac{0{,}3 \times 400 + 0{,}4 \times 200}{600} = \dfrac{1}{3}$

test $H_0 : p_1 = p_2 = p$ contre $H_1 : p_1 \neq p_2$ variable de test $Y_{12} = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{3} \times \frac{2}{3}\left(\frac{1}{400} + \frac{1}{200}\right)}} \sim \mathcal{N}(0,1)$

donc $0{,}95 = p\left(-1{,}96 \sqrt{0{,}21\left(\frac{1}{400} + \frac{1}{200}\right)} < \bar{X}_1 - \bar{X}_2 < 1{,}96 \sqrt{\quad}\right)$ donc $D_0 =]-0{,}078 ; +0{,}078[$

Comme $f_1 - f_2 = 0{,}1 \in D_0$ <u>on rejette H_0 : donc il existe une différence entre les 2 pays</u>

3°) G $\bar{x}_G = 8$ $s_G^2 = 9$ $X_G \sim \mathcal{N}(M_G, \sigma_G^2)$

F $\bar{x}_F = 6$ $s_F^2 = 5$ $X_G \sim \mathcal{N}(M_F, \sigma_F^2)$

a) Test sur les variances $H_0 : \sigma_F^2 = \sigma_G^2 = \sigma^2$ contre $H_1 = \sigma_F^2 > \sigma_G^2$ car lois normales et $s_G^2 > s_F^2$

Variable de test $Y_{\sigma^2} = \dfrac{S_G^2}{S_F^2} \sim F(29, 19)$ table $F_{(29,19)} : 0{,}95 \longrightarrow 2{,}075 \Rightarrow D_0 = [1 ; 20{,}75[$

observation $\frac{9}{5} = 1{,}8 \in D_0 \Rightarrow$ <u>Variances égales</u> Valeur commune estimée par S^2 d'observation $s^2 = 7{,}42$

b) Test $H_0 : M_G = M_F$ contre $H_1 : M_G > M_F$ de variable $Y_M = \dfrac{\bar{X}_G - \bar{X}_F}{s^2 \sqrt{\frac{1}{20} + \frac{1}{30}}} \sim T_{48}$

table $T_{48} : 0{,}95 \longrightarrow 1{,}678 \Rightarrow D_0 = \left] -\infty ; 1{,}678 \sqrt{7{,}42\left(\frac{1}{20} + \frac{1}{30}\right)} \right[=]-\infty ; 1{,}296[$ $\bar{x}_G - \bar{x}_G = 2 \notin D_0 \rightarrow$ <u>H_1</u>

- SECONDE PARTIE -

1°) a) $X_k \sim \mathcal{E}(\lambda) \Rightarrow f(x_k) = \lambda e^{-\lambda x_k} \mathbb{1}_{\mathbb{R}_+}(x_k)$

Hypothèse de récurrence H_n : S_n a pour densité $f_n(t) = \dfrac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!} \mathbb{1}_{\mathbb{R}_+}(t)$

- H_1 vérifiée
- Si H_n vérifiée $S_{n+1} = S_n + X_{n+1}$ de densité $f_{n+1}(t) = \displaystyle\int_{-\infty}^{\infty} \underbrace{f_n(x)}_{=0 \text{ si } x<0} \underbrace{f(t-x)}_{=0 \text{ si } x>t} dx = \int_0^t f_n(x) f(t-x) dx$

donc $f_{n+1}(t) = \displaystyle\int_0^t \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} \lambda e^{-\lambda(t-x)} dx$

$= \dfrac{\lambda^{n+1} e^{-\lambda t}}{(n-1)!} \displaystyle\int_0^t x^{n-1} dx = \dfrac{\lambda^{n+1} t^n e^{-\lambda t}}{n!}$ sur $\mathbb{R}_+$ donc $H_n \Rightarrow H_{n+1}$

On a donc H_1 et $H_n \Rightarrow H_{n+1}$ donc $\forall n \in \mathbb{N}^*$ $f_n(t) = \dfrac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!} \mathbb{1}_{\mathbb{R}_+}(t)$

b) T = somme de 3 lois $\mathcal{E}(\lambda)$ indépendantes $\Rightarrow f(t) = f_3(t) = \dfrac{\lambda^3 t^2 e^{-\lambda t}}{2} \mathbb{1}_{\mathbb{R}_+}(t)$ d'après la question a)

c) Estimation de λ : fonction de vraisemblance $L(t_1, t_2, \dots, t_n) = \displaystyle\prod_{k=1}^{n} \frac{\lambda^3 t^2 e^{-\lambda t}}{2} = \left(\frac{1}{2}\right)^n \lambda^{3n} (t_1 t_2 \dots t_n)^2 e^{-\lambda \sum t_k}$

et $\ln L = -n \ln 2 + 3n \ln \lambda + \sum \ln t_k^2 - \lambda \sum t_k$

Equations de vraisemblance $\begin{cases} \dfrac{\partial \ln L}{\partial \lambda} = 0 \\ \dfrac{\partial^2 \ln L}{\partial \lambda^2} < 0 \end{cases} \Rightarrow \begin{cases} \dfrac{3n}{\lambda} - \sum t_k = 0 \\ -\dfrac{3n}{\lambda^2} < 0 \text{ ok!} \end{cases} \Rightarrow \lambda = \dfrac{3n}{\sum t_k} \rightarrow \boxed{\text{estimateur } \Xi = \dfrac{3}{\overline{T}}}$

d) observation $\overline{t} = \frac{1}{20}(7 + 9{,}8 + \dots + 4{,}3) = 15 \Rightarrow$ estimation de λ : $\frac{3}{15} = 0{,}2 = \hat{\lambda}$

2°)

$k =$	n_k	$k n_k$	p_k	table $\mathcal{P}(3)$ np_k	$\dfrac{(n_k - np_k)^2}{np_k}$
0	3	0	0,05	5	0,8
1	16	16	0,15	15	0,067
2	21	42	0,224	22,4	0,0875
3	27	81	0,224	22,4	0,32
4	14	56	0,168	16,8	0,47
5	7	35	0,10	10	0,9
6	7] 12	42	0,05	5] 8,3	1,65
7	5]	35	0,033	3,3]	
	100	307	1		$4{,}3 = d^2$

Estimation du paramètre $\mu = \bar{x} = \frac{1}{100}\left(\sum k n_k\right) \simeq 3$

test H_0 : $X \sim \mathcal{P}$ contre H_1 : $X \nsim \mathcal{P}$

Variable de test :

$D^2 = \sum \dfrac{(N_k - np_k)^2}{np_k} \sim \chi^2_{7-1-1} = \chi^2_5$

→ 1 paramètre estimé

→ nb de classes après regroupement

table χ^2_5 : $0{,}95 \rightarrow 11{,}07 \Rightarrow D_0 = [0; 11{,}07[$

observation $d^2 = 4{,}3 \in D_0 \Rightarrow H_0$

On accepte donc H_0 : X suit une loi de Poisson

3°) a) 1 visite $\begin{cases} \xrightarrow{0,3} \text{vente} \\ \xrightarrow{0,7} \overline{\text{vente}} \end{cases}$ 100 fois indépendantes $\Rightarrow V$ = nb de ventes $\sim \mathcal{B}(n = 100; p = 0{,}3)$

$E(V) = np = 30$ $\text{Var } V = 21$

qu'on approche par une loi normale $\mathcal{N}(30; \sigma^2 = 21)$ avec une correction de continuité.

b) $p(25 \leq V \leq 40) = p(24{,}5 < V < 40{,}5) = p\left(\dfrac{24{,}5 - 30}{\sqrt{21}} < \mathcal{N}(0,1) < \dfrac{40{,}5 - 30}{\sqrt{21}}\right) = p(-1{,}2 < \mathcal{N}(0,1) < 2{,}29)$

donc p(entre 25 et 40 ventes) $= 0{,}874$