Université PANTHÉON - ASSAS (PARIS II) **U.E.F.1**

Droit - Economie - Sciences Sociales **4005**

Guy de la Brosse

Session : Janvier 2018

Année d'étude : Troisième année de Licence économie-gestion mention économie et gestion parcours gestion

Discipline : *Analyse des données appliquée à la gestion*
(Unité d'Enseignements Fondamentaux 1)

Titulaire(s) du cours : M. Dorin MILITARU

Document(s) autorisé(s) : Calculettes

L'épreuve se compose de 4 exercices. Aucun ordre n'est imposé pour les traiter.

Exercice 1

Variables discrètes et variables continues. Vrai ou Faux ?
a. Une variable discrète ne prend que des valeurs positives.
b. Une variable continue est groupée en classes.
c. Le chiffre d'affaires d'une entreprise est une variable discrète.
d. Le nombre de salariés d'une entreprise est une variable discrète.
e. L'âge et la taille sont des variables continues.

Exercice 2

Caractères et modalités. Vrai ou Faux ?
a. Tout caractère peut avoir une infinité de modalités.
b. Un même individu peut appartenir simultanément à deux (ou plus) modalités.
c. La taille est un caractère quantitatif, de même que l'état matrimonial.
d. Un caractère quantitatif est tel que les modalités qui lui sont associées sont mesurables.
e. Le département de naissance ainsi que la nationalité des individus sont des caractères qualitatifs.

Exercice 3

L'étude trimestrielle portant sur les équipements numériques des ménages, menée en partenariat par GfK et Médiamétrie, dont les résultats ont été publiés en novembre 2013, montre que le taux de ménages possédant au moins une tablette tactile (taux d'équipement en tablettes), en France métropolitaine, est de 24,7 % au troisième trimestre (T3) 2013.

On admettra qu'il y a, à cette période, en France métropolitaine, 27 600 000 ménages.

L'échantillon du trimestre T3 est composé de 6000 ménages.

1) Quel est le taux de sondage de l'étude ?

2) En 2012, le taux d'équipement en tablettes des ménages estimé sur une enquête analogue était de 10,9 %.

Quel est le taux d'évolution de l'équipement en tablette des ménages en un an ?

3) Calculer l'écart-type du taux d'équipement fourni par l'échantillon du T3 2013.

En déduire un intervalle de confiance à 95 % p, proportion de ménages possédant au moins une tablette tactile dans l'ensemble de la population au T3 2013.

4) Combien aurait-il fallu interroger de ménages pour avoir la proportion p connue à + ou – 0,5 % près ?

5) Quelle est l'estimation du nombre N(T) de ménages équipés d'au moins une tablette au T3 2013 dans la population totale ?

6) Donner un intervalle de confiance à 95 % pour N(T) dans l'ensemble de la population au T3.

7) On note par MAE les ménages ayant au moins un enfant de moins de 18 ans ; dans la population totale des ménages, les foyers MAE comptent pour 31,4 % (source INSEE).
L'échantillon du T3 2013 a été tiré de façon proportionnelle.
Combien de ménages avec au moins un enfant de moins 18 ans ont été interrogés ?

8) Dans ces ménages MAE, la proportion d'équipes tablette(s) est 40 %.
a) Calculer l'écart-type de ce résultat et donner un intervalle de confiance à 95 % pour p(AE), proportion de ménages avec enfant(s) de moins de 18 ans possédant au moins une tablette.
b) Quel est donc le taux d'équipement en tablettes des ménages sans enfant(s) de moins de 18 ans (personnes seules, couples sans enfant de moins de 18 ans) ?

9) Combien de ménages avec au moins un enfant de moins de 18 ans aurait-il fallu interroger dans l'échantillon pour avoir p(AE) connu à 1 ou -1 % près ?

10) sur la base des résultats du T3 2013, en découpant l'échantillon en deux strates (la première composée des ménages avec moins un enfant de moins de 18 ans et la

deuxième des « autres ménages »), combien faudrait-il interroger de ménages avec au moins un enfant de moins de 18 ans si on applique la méthode de l'échantillon optimal de Nyeman ?

11) Dans son panel de magasins distributeurs, la société GfK observe que, depuis le lancement commercial des tablettes en France, il a eu 10 millions de tablettes vendues au grand public (hors entreprises).
Comment expliquez-vous ce nombre de tablettes vendues et le nombre N(T) de ménages équipés trouvé à la question 5 ?

12) On veut étudier s'il existe un lien entre la présence d'au moins une tablette dans un ménage et le fait que le chef de ménage soit actif ou non.

Sur la base de l'échantillon du T3 2013, le tableau ci-après donne le nombre de possesseurs d'une tablette selon l'activité du chef de ménages :

	Actifs	**Inactifs**	**Total**
Equipés Tablette	886	596	1480
Pas de tablette	2234	2284	4518
Total	3120	2880	6000

Donnée additionnelle :
Le seuil de « petitesse » vaut 3,84

Quelle est votre réponse à la question : y a-t-il un lien entre activité du chef de ménage et équipement du ménage en tablette(s) ?

Exercice 4

Une banque cherche à mieux connaître le comportement des grandes entreprises en matière de placement des excédents temporaires de trésorerie. Pour ce faire, une enquête est effectuée auprès d'un échantillon (que l'on considérera aléatoire) de vingt grandes entreprises clientes de cette banque. Le tableau que vous avez constitué à la demande du directeur commercial de la banque porte sur le chiffre d'affaires mensuel moyen xi et le portefeuille détenu de bons d'épargne à court terme (voir tableau ci-après).

i	$X(i)$	$Y(i)$	$X^2(i)$	$Y^2(i)$	$X(i)Y(i)$
1	25,00	4,50	625,00	20,25	112,50
2	26,30	5,20	691,69	27,04	136,76
3	28,50	9,40	812,25	88,36	267,90
4	28,70	8,40	823,69	70,56	241,08
5	29,20	9,20	852,64	84,64	268,64
6	30,30	9,00	918,09	81,00	272,70
7	33,40	14,50	1115,56	210,25	484,30
8	34,00	15,50	1156,00	240,25	527,00

9	36,20	15,20	1310,44	231,04	550,24
10	37,00	15,50	1369,00	240,25	573,50
11	37,80	18,80	1428,84	353,44	710,64
12	40,10	22,50	1608,01	506,25	902,25
13	41,30	21,70	1705,69	470,89	896,21
14	41,80	21,00	1747,24	441,00	877,80
15	43,50	24,00	1892,25	576,00	1044,00
16	45,50	21,00	2070,25	441,00	955,50
17	46,00	29,00	2116,00	841,00	1334,00
18	46,30	27,00	2143,69	729,00	1250,10
19	46,90	23,00	2199,61	529,00	1078,70
20	50,00	22,00	2500,00	484,00	1100,00
Total	747,80	336,40	29085,94	6665,22	13583,82

1) Calculer le coefficient de corrélation linéaire entre le chiffre d'affaires mensuel moyen et le portefeuille de bons d'épargne.
2) Déterminez la droite de régression linéaire faisant dépendre le portefeuille du chiffre d'affaires : $y = a + bx$
3) Calculer les indicateurs de qualité de la régression : coefficient de détermination, tests de Student.
4) Fournissez un intervalle de confiance à 95% de la pente.
5) Dans l'hypothèse où le chiffre d'affaires d'une entreprise quelconque *i* s'accroîtrait de 15% l'an prochain, quel serait le montant prévisible du portefeuille détenu ? Pour ce faire, on supposera que le comportement des entreprises restera stable et l'environnement économique inchangé.
6) Supposons que l'entreprise ait un chiffre d'affaires mensuel de 35 millions d'euros, donnez un intervalle de confiance à 95% du portefeuille qu'elle détiendra l'an prochain dans les conditions décrites à la question 5.

- La **moyenne** m de X sur P : $m = \sum_i \frac{X(i)}{N} = \bar{X}$
- La **variance** V(X) sur P : $V(X) = \sum \frac{[X(i)-m]^2}{N} = \frac{[\sum X(i)^2]}{N} - m^2$

  Variance = moyenne des carrés – carrés de la moyenne
- La **variance** S^2 (pop°) et S'^2 (échantillon) :

  $S'^2 = \frac{[\sum X(i)^2]}{n} - \left[\sum \frac{X(i)}{n}\right]^2$ et $S^2 = \frac{nS'^2}{n-1}$ (sans biais)
- L'**écart-type** $\sigma(X) = \sqrt{V(X)} \iff V(X) = \sigma^2(X)$
- **Codage** : **distance entre les individus** :

$$d^2(A,B) = \sum_k p^2(k) \sum_m [X_k(m,A) - X_k(m,B)]^2$$

où $X_k (\forall k = 1, \dots, K)$ les K variables observées ;
La variable X_k à M(k) modalités notées $m(\forall m = 1, \dots, M(k))$;

$$\begin{cases} X_k(m,A) = 1 \text{ si l'individu possède la modalité m de la variable } X_k \\ X_k(m,A) = 0 \; sinon \end{cases}$$

- **Moyenne** de l'échantillon : $m^* = \sum_j \frac{X(i(j))}{n}$
- **Taux de sondage** $f = \frac{n}{N}$ où N, effectif P et n, effectif E
- **La stratification d'un échantillon**

  Ecart-type d'un échantillon stratifié = $\sqrt{\frac{[V(Strate\ 1)+V(Strate2]}{4}}$

  Moyenne d'un échantillon stratifié = $\frac{[moy(Strate\ 1)+moy(Strate\ 2)]}{2}$
- **Intervalle de confiance** (pour n grand)

$$Pr[\theta^* - 2\sigma(\theta^*) \leq \theta < \theta^* + 2\sigma(\theta^*)] \approx 0{,}95$$

  - **Précision**

La précision : $\pm 2\sigma(\theta^*)$ dépend de la taille n de E

Soit h la précision attendue, on a :

$$\pm 2\sigma(\theta^*) = h \iff 4\sigma(\theta^*) = h^2$$

La précision d'une moyenne : $\pm \frac{2S}{\sqrt{n}}$

Soit h la précision attendue, on a :

$$\pm \frac{2S}{\sqrt{n}} = h \Leftrightarrow n \approx \frac{4S^2}{h^2}$$

Lois de probabilité

- Lois discrètes : $P(X = x), \forall x \in \mathbb{N}$
- Lois continues :
  Fonction de densité : $f_X(x), \forall x \in \mathbb{R}$
  Fonction de répartition : $F(x) = P(X < x)$

- ***Lois usuelles discrètes***
  - Loi uniforme sur $(x(1), \dots, x(k))$: $P(X = x(j)) = \frac{1}{K}$
  - Loi de poisson : $P(X = x) = \left(\frac{\lambda^x}{x!}\right) e^{-\lambda}$; $V(X) = E(X) = \lambda$
  - Loi binomiale : $P(X = x) = C(n,x)p^x(1-p)^{n-x}$
    $E(X) = np$ et $V(X) = np(1-p)$, $0 < p < 1$ et
    $0 < x < n$

En général : $E(X) = \sum xp$; $E(m) = m$; $V(m) = 0$

$$V(X) = \sigma^2 = \sum (x - E(X))^2 \times p(x)$$
$$V(X) = E(X^2) - E^2(X)$$
$$V(X^2) = E(X^4) - E^2(X^2)$$

- ***Lois usuelles continues***
  - **Loi uniforme** : $f(x) = \frac{1}{b-a}, \forall a < x < b$
  - **Loi normale** $\boldsymbol{X \sim N(m, \sigma)}$:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}, \forall x \in \mathbb{R}$$
$$Z = \frac{X-m}{\sigma} \sim N(0,1)$$
$$Y = X - m \sim N(0,\sigma) \rightarrow V(Y) = \sigma^2$$

Stabilité de la loi Normale par combinaison linéaire

$$X_1 + \cdots + X_n \sim N(nm; \sqrt{n}\sigma)$$

où $m = \sum_{i=1}^n m_i$ et $\sigma = \sum_{i=1}^n \sigma_i$

- **Loi du khi-deux** : $X_i \sim N(0,1) \rightarrow Y = \sum_{i-1}^n X_i^2 \sim \mathcal{X}^2(n)$
  $E(Y) = n$ *et* $V(Y) = 2n$
- **Loi de Student** : $X_i \sim N(0,1) \rightarrow Z \sim \mathcal{X}^2(n) \rightarrow U = \frac{X}{\sqrt{Z/n}} \sim T(n)$

$$E(U) = 0 \; et \; V(U) = \frac{n}{n-2}, \forall n \in \mathbb{N}_+^*$$

Règles de probabilité

$$P(X \geq a) = 1 - P(X \leq a)$$
$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$$
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A\ ET\ B)}{P(B)}$$

- Moyenne : E : $\bar{X} = \sum_i X(i) / n$

Variance moyenne empirique : $\sigma^2(\bar{X}) = \sigma^2/n \approx S^2/n$

- Variance de f (échantillon en proportion : V(f) = f(1 - f)/n

- Intervalle de confiance à 95%

$[\bar{X} - 1{,}96\ \sigma / n^{1/2} ; \bar{X} + 1{,}96\ \sigma / n^{1/2}]$
On peut remplacer 1,96 par 2

- Si σ est inconnu, on l'estime par S (écart-type de l'échantillon):

$0{,}95 = P(\bar{X} - 2\ S / n^{1/2} \leq m < \bar{X} + 2\ S / n^{1/2})$

- **Calcul de la distance (Khi 2)**

d(M, T) = $n \sum_i \sum_j [p(i, j) - p(i, .)p(., j)]^2 / p(i, .)p(., j)$
= « n.(observé - théorique)² / théorique »

- **Échantillon optimal de Neyman de taille n :**

Nombre de femmes : n(F) et salaire moyen = SalM(F)
Nombre d'hommes : n(H) et salaire moyen = SalM(H)

$V(SalM(F)) = \sigma^2(F) / n(F)$
et $V(SalM(H)) = \sigma^2(H) / n(H)$

$V(SalM(H)) = \sigma^2(H) / n(H)$
$= [\sigma^2(F) + \sigma^2(H)] / n = V(SalM(F))$

- **Sondage stratifié :**

Méthode 1 (proportionnalité) :
a=n/N n(h) = a.N(h) n(h) = N(h).n/N
Méthode 2 (Neyman) :
$n. \times \sigma^2(h) / \sum \sigma^2(h)$

- **Modèle linéaire**

Cov (X, Y) = $[\Sigma(X(i) - \bar{X})(Y(i) - \bar{Y})]/n = [\Sigma X(i)Y(i)]/n - \bar{X}\bar{Y}$
OU Cov(X, Y) = « moyenne des produits » - « produit des moyennes »

Puis le coefficient de corrélation linéaire (normé) :
$\rho(X, Y) = Cov(X, Y) / \sigma(X)\sigma(Y)$

Minimiser $Q(a, b) = \sum u(i)^2 = \sum (Y(i) - a - bX(i))^2$

- **Estimation de σ^2 :**

Un estimateur sans biais de σ^2 est donné par :
$\sigma^{2*} = \sum u^{*2}(i) / (n - 2)$
Ou encore : $\sigma^{2*} = n\ S^2(Y)\ (1 - \rho^2)/ (n - 2)$

$V^*(b^*) = \sigma^{2*}(b^*) = \sigma^{2*} / \sum_i (X - \bar{X})^2 = \sigma^{2*} / nS^2(X)$,

$\sigma^{2*}(b^*) = S^2(Y)\ (1 - \rho^2)/ (n - 2)\ S^2(X)$

Équation d'analyse de la variance

$S^2(Y) = S^2(Y^*) + S^2(u^*)$

Variance totale = Variance expliquée par le modèle

+ Variance résiduelle

Qualité de l'ajustement, coefficient de détermination
R^2 : $R^2 = S^2(Y^*) / S^2(Y)$

= Variance Expliquée par le modèle / Variance Totale

$= 1 - S^2(u^*) / S^2(Y)$

Le modèle linéaire multiple
$Y = a + b_1 X_1 + b_2 X_2 + ... + b_K X_K + u$
Ecriture matricielle du modèle (MLM)
$Y = Xb + u$
$E(u) = 0$ $V(u) = \sigma^2 I_n$

Estimateur sans biais de σ^2 est donné par :
$\sigma^{2*} = \sum_i u^{*2}(i) / (n - (K+1))$

$V(b^*) = \sigma^2 (X^t X)^{-1}$

R^2 corrigé:
$R^{2*} = 1 - (n - 1)(1 - R^2) / (n - K)$