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

Droit - Economie - Sciences Sociales **4005**

Guy de la Brosse

Session : Septembre 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 6 exercices. Aucun ordre n'est imposé pour les traiter.

Exercice 1

Mode. Vrai ou Faux ?
a. Le mode correspond à la valeur de la série qui partage la population en deux sous-ensembles d'effectifs égaux.
b. Le mode est la valeur la plus élevée de la série étudiée.
c. Le mode est égal à la somme des observations de la série divisée par le nombre d'observations.
d. Le mode est une caractéristique de dispersion.
e. Le mode est la valeur de la variable qui correspond à l'effectif le plus élevé.

Exercice 2

On considère les notes suivantes (sur 20) obtenues par 7 étudiants à l'examen de microéconomie en première année : 18, 15, 8, 12, 8, 15, 4. Vrai ou Faux ?
a. La moyenne est égale à (18 + 15 + 8 + 12 + 4) / 7, soit 8,14/20.

b. La note médiane est égale à 12/20.
c. La distribution est unimodale, le mode étant égal à 8/20.
d. L'étendue est égale à 12.
e. Le moment simple d'ordre 1 est égal à 11,42.

Exercice 3

Liaison entre deux variables. Vrai ou Faux ?
a. Deux variables indépendantes sont telles que si l'une augmente, l'autre diminue.
b. Deux variables corrélées évoluent dans le même sens.
c. Les moyennes marginales et conditionnelles sont identiques pour chacune des deux variables étudiées si celles-ci sont indépendantes l'une de l'autre.
d. Les moyennes conditionnelles sont égales aux valeurs des variables dans le cas d'une liaison fonctionnelle entre x et y.
e. Lorsque y croît quand x décroît, on dit que les variables sont corrélées négativement.

Exercice 4

Coefficient de corrélation linéaire. Vrai ou Faux ?
a. Un coefficient de corrélation linéaire égal à −1 témoigne d'une absence de corrélation entre les variables étudiées x et y.
b. Un coefficient de corrélation linéaire supérieur à 1 témoigne d'une très forte corrélation entre x et y.
c. Si le coefficient de corrélation linéaire r(x,y) est égal à 1, on peut en déduire que x influence y.
d. Si r(x,y) = 0, les variables x et y n'ont aucun lien entre elles.
e. Plus un nuage de points a une forme allongée, plus la corrélation entre les deux variables est forte.

Exercice 5

On cherche à expliquer la variable Y, probabilité qu'un ménage soit équipé d'au moins une tablette.

En première approche, on reteint les variables ci-après pour tenter d'expliquer Y :

X(1) = nombre de personnes du ménage
X(2) = présence d'enfant(s) de moins de 18 ans
X(3) = activité du « chef de ménage » (actif / inactif)
X(4) = revenu du ménage
X(5) = taux d'épargne du ménage

X(6) = présence d'au moins un smartphone dans le ménage
X(7) = accès à un service de télévision par ADSL
X(8) = abonnement à au moins un titre de presse
X(9) = accès à internet haut débit
X(10) = niveau d'études du « chef de ménage »
X(11) = région d'habitation
X(12) = taille de l'agglomération
X(13) = âge du « chef de ménage »
X(14) = profession du « chef de ménage »
X(15) = assidu du cinéma en salle
X(16) = utilisateur régulier des transports en commun

On postule le modèle linéaire suivant :

$$Y = a + \sum_{k=1}^{16} b(k).X(k) + u$$

u étant la variable d'écart entre la réalité Y et le modèle, supposée suivre une loi normale N(0, 1).

Les données ont été recueillies dans une enquête de 6000 ménages interrogés.

Les résultats de l'ajustement du modèle sur les données sont les suivants :

Variable	**Coefficient**	**Valeur estimée**	**Ecart-type**
Constante	a	1,28	0,43
X(1)	b(1)	3,11	1,11
X(2)	b(2)	2,65	1,09
X(3)	b(3)	0,89	0,24
X(4)	b(4)	3,87	3,18
X(5)	b(5)	- 2,15	1,67
X(6)	b(6)	1,98	1,14
X(7)	b(7)	0,65	0,78
X(8)	b(8)	- 1,14	0,78
X(9)	b(9)	2,08	1,98
X(10)	b(10)	- 1,76	0,99
X(11)	b(11)	2,59	3,15
X(12)	b(12)	- 0,56	0,45
X(13)	b(13)	- 3,78	2,59
X(14)	b(14)	1,49	0,37
X(15)	b(15)	- 0,69	0,78
X(16)	b(16)	2,36	1,88

$R^2 = 0,78$

1) Calculer et justifier l'emploie du coefficient de détermination ajusté R^{2*}. Le résultat vous surprend-il ?

2) Quelles sont les variables à retenir pour expliquer la probabilité qu'un ménage possède au moins une tablette tactile ? Expliquez la méthode et justifiez vos choix.

3) Ecrire l'intervalle de confiance à 95% pour le paramètre b(14).

4) Trois foyers f(1), f(2) et f(3) ont répondu à l'enquête.

On retient les douze variables suivantes : nombre de personnes du foyer (NPF), activité du chef de ménage, présence d'enfants, nombre de postes de télévision, possession d'un smartphone (au moins), possession d'une tablette (au moins), statut vis-à-vis du logement, accès à internet, abonné à un réseau ADSL, abonné à un service de télévision payante, raccordé en haut débit ou à la fibre optique, possesseur d'une SmartTV (téléviseur connecté).

Les modalités des variables sont les suivantes :
Nombre de personnes du foyer (NPF) (3 modalités) :1, 2, 3 ou plus
Activité du chef de ménage (2) : actif, inactif
Présence d'enfants (2) : oui, non
Nombre de postes de télévision (2) : 0, 1, 2, ou plus
Possession d'un smartphone (au moins) (2) : oui, non
Possession d'une tablette (au moins) (2) : oui, non
Statut vis-à-vis du logement (3) : locataire, propriétaire, autre
Accès à internet (2) : oui, non
Abonné à un réseau ADSL (2) : oui, non
Abonné à un service de télévision payante (2) : oui, non
Raccordé en haut débit ou à la fibre optique (2) : oui, non
Possesseur d'une SmartTV (téléviseur connecté) (2) : oui, non

Caractéristiques du foyer f(1) : 3 personnes, chef de ménage actif, 1 enfant, 2 téléviseurs, deux smartphones, une tablette, locataire, a accès à internet, est abonné à l'ADSL, ne reçoit que des chaînes gratuites de télévision, est en haut débit, n'a pas de SamrtTV

Caractéristiques du foyer f(2) : 5 personnes, chef de ménage actif, 3 enfant, 2 téléviseurs, trois smartphones, une tablette, propriétaire, a accès à internet, est abonné à l'ADSL, reçoit des chaînes payantes de télévision, est en haut débit, n'a pas de SamrtTV

Caractéristiques du foyer f(3) : 2 personnes, chef de ménage actif, pas d'enfant, un téléviseurs, deux smartphones, pas de tablette, locataire, a accès à internet, n'est pas abonné à l'ADSL, reçoit des chaînes payantes de télévision, n'a pas accès au haut débit, n'a pas de SamrtTV.

Entre f(2) et f(3), quel est le foyer le plus proche de f(1) ?

Exercice 6

Vous êtes chargé d'études au sein du Département Loisirs d'une société d'études.

Vous avez la responsabilité de réaliser une étude portant sur le comportement des personnes de 15 ans ou plus résidant en France métropolitaine au cours de la période juillet-août 2017.

Cette étude aura lieu par téléphone, la durée moyenne de l'interview sera de 15 mn, auprès d'un échantillon de n personnes.

Le critère de qualité pris pour la taille n de l'échantillon est la précision sur la proportion p de personnes parties en vacances au moins une fois au cours de cette période.

Vous réalisez une première vague d'enquête auprès d'un premier échantillon de taille n(1) = 2000, qui conduit à une estimation f de p égale à 56 %.

1) Donner un intervalle de confiance pour p, au niveau 95 %.

2) Vous souhaitez avoir une précision de +/- 1,5 % pour l'estimation de p.

Combien faudrait-il interroger de personnes n(2) dans une deuxième vague d'enquête pour atteindre cette précision ?

3) on donne les éléments de coût suivants :

- le coût unitaire du recueil des données est de 15 € (ce coût comprend l'accès à la base de sondage, le questionnaire, la réalisation des interviews, les contrôles, et la transmission du fichier)

- le traitement des données (redressement, production des résultats, analyse des données, rédaction du rapport et la note de synthèse) prend 10jours/homme à raison de 700 € la journée.

- la direction financière vous impose une marge de 25 %

- le prix de vente par client est de 5000 €

Combien de clients devez-vous avoir pour cette étude soit rentable ?

- 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** $\sigma(X) = \sqrt{V(X)} \Leftrightarrow V(X) = \sigma^2(X)$
- **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 :

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

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):

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

- **<u>Calcul de la distance (Khi 2)</u>**

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

- **<u>Échantillon optimal de Neyman de taille n :</u>**

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

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

- **<u>Sondage stratifié :</u>**

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

- **<u>Modèle linéaire</u>**

$Cov(X, Y) = [\Sigma(X(i) - \bar{X})(Y(i) - \bar{Y})]/n = [\Sigma X(i)Y(i)]/n - \bar{X}\bar{Y}$
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 $Q(a, b) = \sum u(i)^2 = \sum (Y(i) - a - bX(i))^2$

- **<u>Estimation de σ^2 :</u>**

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 :
$\sigma^{2*} = \sum_i u^{*2}(i) / (n - (K+1))$

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

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