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

Droit - Economie - Sciences Sociales **4005**

Guy de la Brosse

Session : Janvier 2017

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) : Documents papier de cours et de TD, calculettes

L'épreuve se compose d'un exercice et d'un problème. Aucun ordre n'est imposé pour les traiter.

1. Exercice :

Un institut spécialisé en sondages politiques a réalisé en 2004 deux sondages à 15 jours d'intervalle sur les opinions des Français à propos de la ratification du projet de constitution de l'Union Européenne.

Les résultats ont été les suivants :

Premier sondage, réalisé par téléphone les 1 et 2 septembre 2004 auprès d'un échantillon de 885 personnes âgées de plus de 18 ans et inscrites sur les listes électorales.

Oui à la constitution : 69%
Non à la constitution : 31%

Deuxième sondage, réalisé par téléphone les 14 et 15 septembre 2004 auprès d'un échantillon de 834 personnes âgées de plus de 18 ans et inscrites sur les listes électorales.

Oui à la constitution : 67%
Non à la constitution : 33%

Entre ces deux dates, un important homme politique F est intervenu longuement au JT de France 2 pour préconiser un vote négatif. De nombreux médias ont alors conclu que l'intervention de F a joué en la faveur du "non" au référendum.

Qu'en pensez-vous ?
Expliquez et explicitez votre réponse.

2. Problème :

L'indicateur de référence de la mesure d'audience de la radio s'appelle l'audience cumulée : l'audience cumulée A du média radio, pour une journée donnée, est la proportion de la population étudiée ayant écouté la radio, au cours de cette journée, quelle que soit la station écoutée et quelle que soit la durée d'écoute.

La population P objet de l'étude est celle des personnes âgées de 15 à 64 ans, que l'on admettra être égale à 42 millions (source INSEE).
La période étudiée est le trimestre avril-juin 2016, constitué de 13 semaines ; les résultats sont établis sur le « jour moyen de semaine lundi-vendredi », c'est-à-dire la moyenne des 13 x 5 = 65 jours d'observation allant du lundi au vendredi (samedi et dimanche sont donc exclus).

Partie I

1) Dans cette première partie, on interroge 100 personnes différentes chaque jour.
Quelle est la taille n de l'échantillon total E utilisé pour établir des résultats sur le jour moyen lundi-vendredi d'avril-juin ?
Quel est le taux de sondage f ?

2) Sur cette population P des 15 - 64 ans, l'audience cumulée A* de la radio calculée sur l'échantillon E sur le jour moyen L-V d'avril-juin est de 81,7%.

2a - En moyenne quotidienne lundi-vendredi, que vaut NA, nombre de personnes de P écoutant la radio (« auditeurs ») ?

2b - Calculer l'écart-type $\sigma(A^*)$ de l'audience cumulée A*.

2c - Donner alors l'intervalle de confiance à 95% pour l'audience cumulée A sur la population P.

2d - En déduire l'intervalle de confiance à 95% du nombre d'auditeurs NA de la radio sur P.

3) Si on avait voulu que A* soit connu à +-0,5 % près, combien de personnes n_1 aurait-il fallu interroger chaque jour pour avoir le résultat A* = 81,7% avec cette précision souhaitée ?

Partie II

On conserve l'échantillon E d'effectif n.

Concrètement, cet échantillon E est stratifié en croisant les variables activité (actif/inactif) et âge (avec 3 classes d'âge : 15-24 ans, 25-49 ans, 50-64 ans), et il est construit pour être un modèle réduit de cette structure.

La structure (en %) de la population P fournie par l'INSEE est donnée dans le tableau 1 suivant :

	Actifs	*Inactifs*	*Total*
15-24 ans	6 %	12 %	*18 %*
25-49 ans	43 %	8 %	*51 %*
50-64 ans	19 %	12 %	*31 %*
Total	*68 %*	32 %	*100 %*

Tableau 1 : structure de la population des 15 - 64 ans

Pour chaque catégorie (croisant âge et activité) ainsi constituée, les résultats d'audience cumulée A* de la radio sont donnés dans le tableau 2 ci-après :

	Actifs	*Inactifs*	*Total*
15-24 ans	78 %	75,2 %	*76,1 %*
25-49 ans	85,2 %	66,1 %	*82,2 %*
50-64 ans	93,2 %	70 %	*84,2 %*
Total	*86,8 %*	*71,0 %*	*81,7 %*

Tableau 2 : audiences cumulées

4) Donner la répartition de l'échantillon E selon les strates ainsi constituées.

5) En quelques lignes claires et pédagogiques, justifier le recours à cette méthode de modèle réduit.

6) On s'intéresse à l'audience cumulée de la catégorie notée a des actifs, soit A(a).

6a - Combien vaut A*(a) à partir de E ?

6b - Combien y a-t-il, en jour moyen lundi-vendredi, d'auditeurs actifs NA(a) de 15 à 64 ans ?

6c - Donner les intervalles de confiance à 95 % de A(a) et de NA(a).

7) Les actifs sont une des cibles stratégiques pour la radio.

7a - Combien faudrait-il interroger d'actifs pour avoir A*(a) connu avec une précision de +-0,5% près ?

7b - En conservant la logique de modèle réduit, quelle serait alors la taille de l'échantillon total E et sa répartition par catégorie ?

7c - Quels commentaires vous inspirent ces résultats ?

8) En fait, la cible stratégique numéro 1 pour la radio est celle des actifs de 25 à 49 ans, notée (aj).

8a - Que vaut A*(aj) ?

8b - Quel est son écart-type ?

8c - Combien faudrait-il interroger d'actifs de 25 à 49 ans pour avoir A*(aj) connu avec une précision de +-0,5% près ?

8d - En conservant la logique de modèle réduit, quelle serait alors la taille de l'échantillon total E et sa répartition par catégorie ?

8e - Quels commentaires vous inspirent ces résultats ?

Partie III

On abandonne ici la logique de modèle réduit pour passer à celle de l'échantillon optimal de Neyman.

9) De quand date l'approche de Neyman ? En quelques lignes claires et pédagogiques, justifier le recours à cette méthode de Neyman.

10) A partir des audiences cumulées des six strates croisant âge et activité du tableau 2 et de la question 4 de la Partie II, calculer les variances des six audiences cumulées correspondantes.

11) On fixe autoritairement la taille totale de l'échantillon à 25000.

11a - Quelle est la répartition par strate de l'échantillon de Neyman ?

11b - Quels seraient alors l'écart-type de l'audience cumulée A*(aj) des actifs de 25 à 49 ans et l'intervalle de confiance à 95% de A(aj) ?

[on considèrera A*(aj) égale à celle donnée dans le tableau 2]

Partie IV

On sélectionne au hasard uniforme dans E un sous-échantillon E' de 1000 individus respectant les proportions d'actifs et d'inactifs données par l'INSEE (tableau 1). On classe ces 1000 individus de E' en fonction de leur statut actif/inactif et auditeur/non auditeur.

Cela permet de constituer le tableau 3 ci-dessous :

	Actifs	*Inactifs*	*Total*
Auditeurs	600	220	*820*
Non auditeurs	80	100	*180*
Total	*680*	*320*	*1000*

***Tableau 3* : répartition de E' par activité et auditeur/non auditeur**

12) Comment répondre à la question de l'existence (ou non) d'un lien entre l'écoute de la radio et l'activité ? Deux approches sont possibles. En choisir une, l'expliciter et répondre à la question posée.

NB : à toute fin utile, en fonction de la méthode choisie, on donne le seuil de petitesse de la distance existant entre tableau théorique et tableau observé : 3,84 (avec une certitude de 95 %).

Partie V

Dans cette partie, on cherche à expliquer la variable Y, probabilité d'écouter la radio au cours d'une journée. Pour cela, on retient les 9 variables « explicatives » présentées dans le tableau 4 ci-après, et on postule un modèle de la forme :

$$Y = a + \Sigma_k b_k X_k + u,\ k = 1 \text{ à } 9$$

Les données viennent de l'interrogation des 1000 individus de E'.
Les résultats numériques du modèle sont résumés dans le tableau 4.

13) Ecrire l'intervalle de confiance au niveau 95% pour le coefficient de la variable « Age ».

14) Rappeler ce que représente le R^2 d'un modèle.

15) Au vu des résultats présentés, quelles variables explicatives retenez-vous pour le modèle expliquant la probabilité d'écouter la radio au cours d'une journée ?
Vous expliquerez la démarche retenue et la façon de procéder.

Variables	Estimation de a et des b_k	Ecart-type de a et des b_k
Constante a	4,574	1,212
Nombre de personnes du foyer (NPF)	-0,654	0,208
Age	-0,543	0,067
Activité (actif ou inactif)	1,842	0,462
Statut professionnel (CSP+, employé, ouvrier)	1,708	0,662
Nombre de postes de radio au domicile	3,745	0,965
Accès à Internet	-0,561	0,115
Possession d'un smartphone	0,888	0,166
Possession d'une voiture avec autoradio	2,781	0,464
Logement (propriétaire / locataire / autre)	1,456	0,983

R^2 = 69 %

***Tableau 4* : Résultats du modèle linéaire explicatif**