AlterMecatro/Mecatro3 - Analyse Statistique des Données Abena

1 Description des données

L'étude *Abena 2011-2012* a porté sur les adultes de 18 ans ou plus ayant eu recours à l'aide alimentaire au cours de l'hiver 2011-2012, dans six territoires urbains de France métropolitaine : Paris, Marseille, Grand-Dijon, Seine-Saint-Denis, Val-de-Marne et Hauts-de-Seine.
Il s'agit d'une étude transversale, l'inclusion des personnes enquêtées ayant eu lieu de novembre 2011 à mi-avril 2012. Les personnes ont été tirées au sort selon un plan de sondage stratifié à deux degrés (pour la suite du devoir, ne tenez pas compte du plan de sondage, autrement dit, faites comme si le tirage au sort était fait par sondage aléatoire simple dans la population générale des adultes ayant recours à l'aide alimentaire, afin d'utiliser les méthodes classiques vues dans le cours). Les personnes étaient considérées comme éligibles si elles étaient âgées de 18 ans ou plus, capables de comprendre le contenu et les enjeux de l'étude et de répondre au questionnaire, et si aucun autre membre de leur foyer n'avait déjà été interrogé dans le cadre de l'étude.
Les personnes ayant accepté de participer répondaient à un questionnaire permettant de recueillir des données sur les caractéristiques socio-démographiques (âge, sexe, pays de naissance, situation familiale, nombre d'enfants à charge, niveau scolaire), les conditions de logement (type de logement, équipement en moyens de cuisson), la couverture maladie (affiliation à un régime de sécurité sociale et type de couverture maladie complémentaire), l'emploi et les ressources financières du ménage (activité rémunérée, soutien financier) et le recours à l'aide alimentaire.
Sur les bénéficiaires inclus dans l'étude Abena 2011-2012, 668 femmes ont été pesées et mesurées et ne présentent pas de données incomplètes.
Le fichier **Abena.csv** contient des données simulées inspirées de cette étude. Les variables disponibles sont :

Nom variable	Description	Unité / Valeurs
Urbaine	Zone urbaine	Paris / Marseille / Dijon/Hauts-de-Seine/Seine-Saint-Denis/Val-de-Marne
Age	Age	Années
Couple	Situation familiale	Seule / En couple
Enfants	Nombre d'enfants à charge	Nombre
Scolaire	Niveau scolaire	$\geq$ 2e cycle universitaire / $<$ 2e cycle
situation	Situation professionnelle	Active / Chômeuse* / Inactive
Repas	Type d'aide alimentaire	Denrées à emporter / Repas
Duree	Durée du recours à l'aide alimentaire	Années
Assurance	Assurance maladie complémentaire	Mutuelle / Autres**
Imc	Indice de masse corporelle (IMC)	kg/m^2

*Inscrite ou non à Pôle emploi. **Y compris couverture médicale universelle complémentaire (CMUc).

2 Instructions

— Répondre aux questions qui suivent en utilisant Python.
— Il est attendu une justification des outils utilisés (pourquoi tel test plutôt qu'un autre) et une validation de leurs conditions de validité (taille d'échantillon, normalité...)

3 Statistique descriptive

1. Faire une étude exploratoire des mesures quantitatives de la base de données Abena. Pour chacune,
   — détecter les valeurs atypiques éventuelles (Boxplots). Que préconisez-vous ?
   — donner une synthèse graphique (Histogramme)
   — donner des mesures de tendance centrale et de dispersion
   — pour les variables quantitatives **continues**, justifier si elles sont distribuées suivant la loi normale (QQ-plot et test de Shapiro-Wilk)
2. Faire une synthèse des variables qualitatives de la base de données Abena. Pour chacune, donner une table de fréquences et un diagramme à secteurs.

4 Tests à un échantillon

Interroger, au risque de première espèce $\alpha = 5\%$, les affirmations qui suivent. Mettre en forme les résultats issus de Python et donner une représentation des intervalles de confiance.

1. L'âge moyen des femmes qui ont recours à l'aide alimentaire est de 45 ans.
2. La durée moyenne de recours à l'aide alimentaire est inférieure à 2,5 mois.
3. L'IMC moyen des femmes qui ont recours à l'aide alimentaire est supérieur à 25.

5 Tests à deux échantillons

Interroger, au risque de première espèce $\alpha = 5\%$, les affirmations qui suivent. Mettre en forme les résultats issus de Python et donner une représentation des intervalles de confiance.

1. L'âge moyen des femmes diffère selon le type d'aide alimentaire ("denrées à emporter" ou "repas").
2. La durée moyenne de recours à l'aide alimentaire est plus importante chez les femmes ayant un niveau d'études inférieur au 2ème cycle.
3. L'IMC moyen est plus important chez les femmes ayant un niveau d'études inférieur au 2ème cycle.

6 Test d'association entre deux variables catégorielles

Y a-t-il un lien d'association entre les variables catégorielles qui suivent ? Si oui, mesurer l'intensité du lien et indiquer quelles couples de modalités s'attirent/se repoussent.

1. Situation professionnelle et nature de l'aide reçue (repas ou denrées)
2. Situation professionnelle et situation familiale (seule ou en couple)
3. Zone urbaine et niveau d'études

7 Anova

7.1 Durée de recours à l'aide alimentaire et situation professionnelle

La situation professionnelle a-t-elle un effet sur la durée de recours à l'aide alimentaire ? Justifier et commenter.

7.2 IMC et durée de recours à l'aide alimentaire

Problème 1 : On souhaite étudier d'éventuels liens entre l'indice de masse corporelle et la durée du recours à l'aide alimentaire des personnes.
Au préalable, on créera la variable DUREE2 qui vaut 0 si la personne a eu recours à l'aide alimentaire pendant moins de 2 ans et 1 si la personne a eu recours à l'aide alimentaire pendant 2 ans ou plus.
La question que l'on se pose est : l'indice de masse corporelle diffère-t-il selon que les personnes aient eu recours à l'aide alimentaire pendant moins de 2 ans ou plus de 2 ans ? Et plus généralement, l'indice de masse corporelle est-il différent en fonction de la durée du recours à l'aide alimentaire ? Pour cela, nous allons procéder en plusieurs étapes.

1. Quelle(s) analyse(s) pouvez-vous envisager de réaliser pour répondre à la question : « l'indice de masse corporelle diffère-t-il selon que les personnes aient eu recours à l'aide alimentaire pendant moins de 2 ans ou plus de 2 ans» (sous réserve de vérification des hypothèses associées aux analyses) ?
   — Une analyse de la variance
   — Une régression linéaire
   — Un test de comparaison de moyennes ? loi de Student
   — Un test de comparaison de médianes
   — Un test de Mann-Whitney-Wilcoxon
   — Aucune des précédentes
2. Effectuer un test de comparaison de moyennes basé sur la loi normale. Donner la valeur de la statistique de test avec une précision de deux chiffres après la virgule.
3. Donner la valeur de la p-value avec une précision de deux chiffres après la virgule.
4. L'indice de masse corporelle diffère-t-il significativement selon que la durée soit supérieure ou inférieure à 2 ans ?
   — Oui
   — Non
   — Nous ne pouvons pas conclure

Problème 2 : Nous allons chercher à affiner le résultat obtenu. Pour cela, à partir des quartiles de la durée du recours à l'aide alimentaire, créer la variable DUREECL qui prend les valeurs :

$$\begin{array}{ll} 1 & \text{si durée} \leq 1{,}68 \text{ ans} \\ 2 & \text{si } 1{,}68 \text{ ans} < \text{durée} \leq 2{,}05 \text{ ans} \\ 3 & \text{si } 2{,}05 \text{ ans} < \text{durée} \leq 2{,}35 \text{ ans} \\ 4 & \text{si durée} > 2{,}35 \text{ ans.} \end{array}$$

1. Quelle(s) analyse(s) pouvez-vous envisager de réaliser pour répondre à la question : « l'indice de masse corporelle diffère-t-il selon que les personnes aient eu recours à l'aide alimentaire pendant moins de 1,68 ans ou entre

1,68 et 2,05 ou entre 2,05 et 2,35 ou plus de 3,35 ans» (sous réserve de vérification des hypothèses associées aux analyses) ?
   — Une analyse de la variance
   — Un test de comparaison de moyennes ? loi normale
   — Un test de comparaison de moyennes ? loi de Student
   — Un test de comparaison de médianes
   — Un test de Mann-Whitney-Wilcoxon
2. Effectuer un test ANOVA. Quelle est l'hypothèse alternative H_1 du test à réaliser ?
   — Toutes les moyennes de la variable DUREE sont différentes.
   — Au moins une moyenne de la variable DUREE est différente des autres.
   — Toutes les moyennes de la variable IMC sont différentes.
   — Deux moyennes de la variable IMC sont différentes.
   — Au moins une moyenne de la variable IMC est différente des autres.
3. Rappeler l'équation de l'analyse de la variance.
4. Compléter le tableau d'analyse de la variance suivante.

Source de variation	Degrés de liberté (df)	Somme des carrés (Sum Sq)	Carré moyen (Mean Sq)	Statistique F (F value)
DUREECL			450.1	
Résiduelle				
Totale		24131		

5. Donner la valeur de la statistique de test avec une précision de deux chiffres après la virgule.
6. Donner la valeur de la p-value avec une précision de deux chiffres après la virgule.
7. L'indice de masse corporelle diffère-t-il significativement selon que les personnes aient eu recours à l'aide alimentaire pendant moins de 1,68 ans ou entre 1,68 et 2,05 ou entre 2,05 et 2,35 ou plus de 3,35 ans ?
   — Oui
   — Non
   — Nous ne pouvons pas conclure

8 Analyse de corrélation

Problème 1 : Nous souhaitons pousser plus loin les investigations. Pour cela nous allons utiliser les variables DUREE (en tant que variable quantitative) et IMC.

1. Quelle est la valeur du coefficient de corrélation observé entre les variables DUREE et IMC. Donner la valeur de la statistique avec une précision de deux chiffres après la virgule.
2. Quelle est la borne inférieure de l'intervalle de confiance du coefficient de corrélation ? Donner la valeur avec une précision de deux chiffres après la virgule.
3. Quelle est la borne supérieure de l'intervalle de confiance du coefficient de corrélation ? Donner la valeur avec une précision de deux chiffres après la virgule.
4. Que pouvez-vous en déduire ?
   — Il existe un lien significatif entre la durée du recours à l'aide alimentaire et l'indice de masse corporelle.
   — On ne peut pas conclure quant à l'existence d'un lien linéaire significatif entre les deux variables.
   — On ne peut pas conclure quant à l'existence d'un lien significatif entre les deux variables.
5. Qu'est-ce qui différencie un problème de corrélation linéaire d'un problème de régression linéaire ?
   — Il n'y a aucune différence.
   — Le rôle des 2 variables est symétrique dans la corrélation, contrairement à la régression linéaire.
   — Le rôle des 2 variables est symétrique dans la régression linéaire, contrairement à la corrélation.
   — La corrélation porte forcément sur deux variables quantitatives, contrairement à la régression linéaire.
   — La régression linéaire porte forcément sur deux variables quantitatives, contrairement à la corrélation.
   — Les résultats obtenus avec les deux approches peuvent être différents.

Problème 2 : La durée du recours à l'aide alimentaire DUREE et l'âge des bénéficiaires de l'aide alimentaire sont-ils corrélés ? Justifier et commenter.

9 Régression linéaire simple

Problème 1 : Nous souhaitons maintenant réaliser une régression linéaire de l'indice de masse corporelle IMC en fonction de la durée du recours à l'aide alimentaire DUREE.

1. Quelle est la valeur de l'intercepte de la droite de régression ? Donner la valeur avec une précision de deux chiffres après la virgule.
2. Est-elle interprétable ?
   — Non, l'IMC ne peut pas prendre la valeur 0
   — Non, il n'y a pas d'interprétation de ce paramètre de la régression linéaire
   — Non, dans cette étude sur les personnes ayant recours à l'aide alimentaire, la durée du recours à l'aide alimentaire ne peut pas prendre la valeur 0

3. Quelle est la valeur de la pente de la droite de régression ? Donner la valeur avec une précision de deux chiffres après la virgule.
4. Quelle est la borne inférieure de l'intervalle de confiance de la pente de la droite de régression ? Donner la valeur avec une précision de deux chiffres après la virgule.
5. Quelle est la borne supérieure de l'intervalle de confiance de la pente de la droite de régression ? Donner la valeur avec une précision de deux chiffres après la virgule.
6. Peut-on affirmer que la durée du recours à l'aide alimentaire apporte une information significative pour la prédiction de l'indice de masse corporelle ?
   — Oui
   — Non
   — Nous ne pouvons pas conclure
7. Peut-on estimer la valeur de l'IMC pour une personne ayant recours à l'aide alimentaire pendant 2 ans ?
   — Oui
   — Non
   — La régression linéaire ne permet pas cette estimation
8. Si oui, quelle est la valeur prédite par le modèle ? Donner la valeur avec une précision de deux chiffres après la virgule.
9. Quelle est le pourcentage de variation de l'indice de masse corporelle expliqué par la durée du recours à l'aide alimentaire ? Donner la valeur avec une précision de deux chiffres après la virgule.

10 Réduction de dimensionnalité

1. Réalisez une ACP sur les variables quantitatives pour réduire le nombre de dimensions de la base de données.
2. Quelle mesure de distance entre individus utilisez-vous ?
3. Quel critère d'aggrégation des clusters utilisez-vous ?
4. Combien de composantes principales conservez-vous et pourquoi ?
5. Faire une projection des données sur le premier plan principal. Quelle part d'information est perdue avec la distorsion due à la projection ?

11 Partitionnement des données avec K-means

1. Réalisez un partitionnement des données avec k-Means.
2. Une standardisation des données est-elle nécessaire ? Si oui, pourquoi ?
3. Combien de clusters choisissez-vous ? Justifiez.
4. Donnez une représentation des clusters obtenus, dans le premier plan factoriel (une couleur par groupe).
5. Caractérisez les clusters obtenus (statistique descriptive).

12 Classification hiérarchique des données (CAH)

1. Réalisez une classification hiérarchique des données avec CAH.
2. Une standardisation des données est-elle nécessaire ? Si oui, pourquoi ?
3. Combien de clusters choisissez-vous ? Justifiez.
4. Donnez une représentation des clusters obtenus, dans le premier plan factoriel (une couleur par cluster).
5. Caractérisez les clusters obtenus (statistique descriptive).