

Traitement et Analyse de Données Biologiques – L2S4

Examen de la Session 1

Ma 21 février 2017

- Durée : 2 h 00

Paris Sud - Orsay

- Les questions posées sont, pour la plupart, indépendantes
- Tous les documents sont autorisés
- Ordinateurs et tablettes numériques autorisés mais interdit d'envoyer des messages
- Connexion internet autorisée mais interdit d'envoyer des messages
- Usage des téléphones portables interdit (*règlement en vigueur pour tous les examens sur le campus*)
- N'oubliez pas de reporter votre N° d'anonymat sur les intercalaires
- Le texte de cet examen est composé de 6 pages + 2 pages d'annexes.

Le terme QCM est employé pour désigner une question à choix multiple. Il vous faut dans ce cas indiquer, pour chaque proposition, si elle est vraie ou si elle est fausse en reportant le numéro de la question et la réponse choisie sur votre copie (par ex : Q12-a : FAUX; Q12-b : FAUX; Q12-c : VRAI; Q12-d : FAUX; Q12-e : VRAI)

Dans cet examen nous nous intéressons à une étude sur le café produit par des petits producteurs indépendant (concernant, notamment, sa culture, sa consommation et ses qualités gustatives).

La plupart des questions peuvent être traitées indépendamment les unes des autres.

Un ami vous a recommandé « *Chez Ramon* », un café branché du 3e arrondissement où l'on déguste des cafés étiquetés « *petits producteurs écoresponsables* ». Vous vous êtes rendu à ce café et y avez commandé un arabica qui vous a comblé de bonheur.

Ramon, le patron de l'établissement, vous explique qu'il va lui-même chercher son café à L'Union des Communautés Indiennes de la Région Sud du Mexique (UCIRSM) qui regroupe plusieurs villages s'organisant pour vendre un café cultivé organiquement. Ramon vous apprend que les organisations européennes de commerce alternatif achètent directement le café biologique aux paysans pauvres des plantations de la région.

L' UCIRSM a construit sa propre infrastructure de stockage, de transport, de transformation et d'exportation pour le café qu'elle produit. Ceci permet à ses membres de profiter d'une plus-value précédemment usurpée par des intermédiaires. Le profit est réutilisé par les villages pour améliorer la scolarité, le système médical, la distribution des articles de première nécessité, la formation des agriculteurs et bien d'autres choses encore.

Actuellement, 94 communautés sont membres de l'UCIRSM et plus de 6 500 familles sont impliquées dans ses actions. Cet aspect social est respecté par les acheteurs européens pratiquant le commerce alternatif (Fair Trade Organizations) qui valorisent le café biologique à son juste prix en lui accordant une prime à la qualité.

Ramon vous explique, qu'une fois sur place, il fait son choix au feeling. Il préférerait, cependant, disposer d'informations fiables lui permettant de mieux caractériser les produits pour être sûr de sélectionner les meilleurs cafés. Vous lui proposez alors de faire une recherche poussée sur internet pour récolter des données qui vont l'aider à faire son choix. Pour vous exprimer sa reconnaissance, Ramon vous offre un an de consommation de café dans son établissement.

1^{ère} partie : Chez Ramon, à Paris



Figure 1 a/ L'ambiance chez Ramon b/ Chez Ramon le service est simple. Ce qui compte est la qualité des cafés qu'il propose ; ils sont servis dans une tasse en porcelaine. c/ Ramon est intransigent sur la pression délivrée par son percolateur qui peut être réglée de 9 à 30 bars selon le café commandé.

Q1/ Définissez, dans le contexte de cette étude (vous pouvez vous aider des illustrations en figure 1), une variable aléatoire continue qui pourrait suivre une loi normale, une variable aléatoire discrète qui pourrait suivre une loi binomiale ou, au choix, une loi de Poisson ainsi qu'une variable qualitative. Vous indiquerez, pour chacune d'entre elles, si elle est définie dans une population ou un échantillon (population ou échantillon que vous définirez également, en quelques mots).

Q2/ QCM > Indiquez sur votre copie si chacune des propositions suivantes est vraie ou fausse :

- a/ La variété de café consommée par un client est définie par une variable binomiale [Vrai ou Faux ?]
- b/ Le nombre de clients fréquentant chaque jour l'établissement est une variable qualitative [Vrai ou Faux ?]
- c/ La tasse en figure 1-a appartient à la fois à un échantillon et à une population [Vrai ou Faux ?]
- d/ La tasse de café servie à un client est un individu au sens statistique du terme [Vrai ou Faux ?]
- e/ Le nombre de tasses casées en une heure chez Ramon suit une loi de Poisson [Vrai ou Faux ?]
- f/ Le bénéfice réalisé chaque semaine par Ramon est une variable aléatoire continue [Vrai ou Faux ?]

2^{ème} partie : Etude du questionnaire de satisfaction des clients de Ramon

Ramon distribue systématiquement un questionnaire à chacun des clients fréquentant son établissement.

Ce **questionnaire**, outre des questions permettant d'obtenir des informations sur le café consommé (nom, référence) et sur son consommateur, tel son sexe, son âge ou encore son code postal (questions dites « ouvertes »), est composé de **42 questions** proposant chacune un choix codé sur une échelle de Likert à 5 niveaux (variable ordinale allant de 1 à 5). A titre d'illustration, deux de ces questions sont présentées dans le cadre ci-dessous,

....

QRC12 : Concernant l'arôme émanant du café que vous avez consommé vous l'avez trouvé

1 : désagréable ; 2 : il ne m'a pas marqué ; 3 : agréable ; 4 : très agréable ; 5 : exceptionnel

....

QRC29 : Compte-tenu de sa qualité, trouvez-vous que le café que vous avez dégusté est

1 : excessivement cher ; 2 : très cher ; 3 : cher ; 4 : à un prix raisonnable ; 5 : vraiment pas cher

....

L'objet de ce questionnaire est d'obtenir une **note globale du café** consommé.

Pour chaque questionnaire rendu, les valeurs obtenues en réponse aux 42 questions posées sont sommées dans un score global. Ce score est ensuite divisé par le nombre de questions posées dans le questionnaire (42) pour obtenir une note globale évaluant le café dégusté par le client (variable *note globale*). On admettra qu'il s'agit d'une variable continue.

Q3/ Démontrez très simplement que le domaine de définition dans lequel la variable *note globale* du café prend ses valeurs est l'intervalle [1,5]

Le jeune statisticien à qui Ramon a commandé le questionnaire a trouvé un emploi dans un cabinet de Data Science en Californie. Ramon est perdu, il ne sait pas comment exploiter les réponses qu'il a obtenues sans l'aide de ce statisticien. Vous proposez alors à Ramon de l'aider à compiler et à exploiter les données provenant des questionnaires que ses clients ont complétés.

- Q4/** Ramon dispose à ce jour de 439 questionnaires renseignés par ses clients. Cet ensemble constitue-t-il une population ou un échantillon ?
- Q5/** Si l'on veut l'engager dans des statistiques inférentielles, que faut-il vérifier au préalable pour la variable *note globale*? A cette fin, quel test allez-vous faire ?
- Q6/** On recode l'âge des clients en une nouvelle variable *tranche_age* à 4 modalités : « moins de 25 ans », « entre 25 et 40 ans », « entre 40 et 60 ans », « plus de 60 ans ». Quel est le type de la variable *tranche_age* ?
- Q7/** On désire croiser les variables *tranche_age* et *sexe* pour savoir si les hommes et les femmes se répartissent de la même façon dans les différentes tranches d'âge. Quel test allez-vous réaliser ? (N'oubliez pas de poser l'hypothèse nulle). L'analyse de ce test, avec R, a conduit à un non rejet de l'hypothèse nulle. Qu'en concluez-vous ?

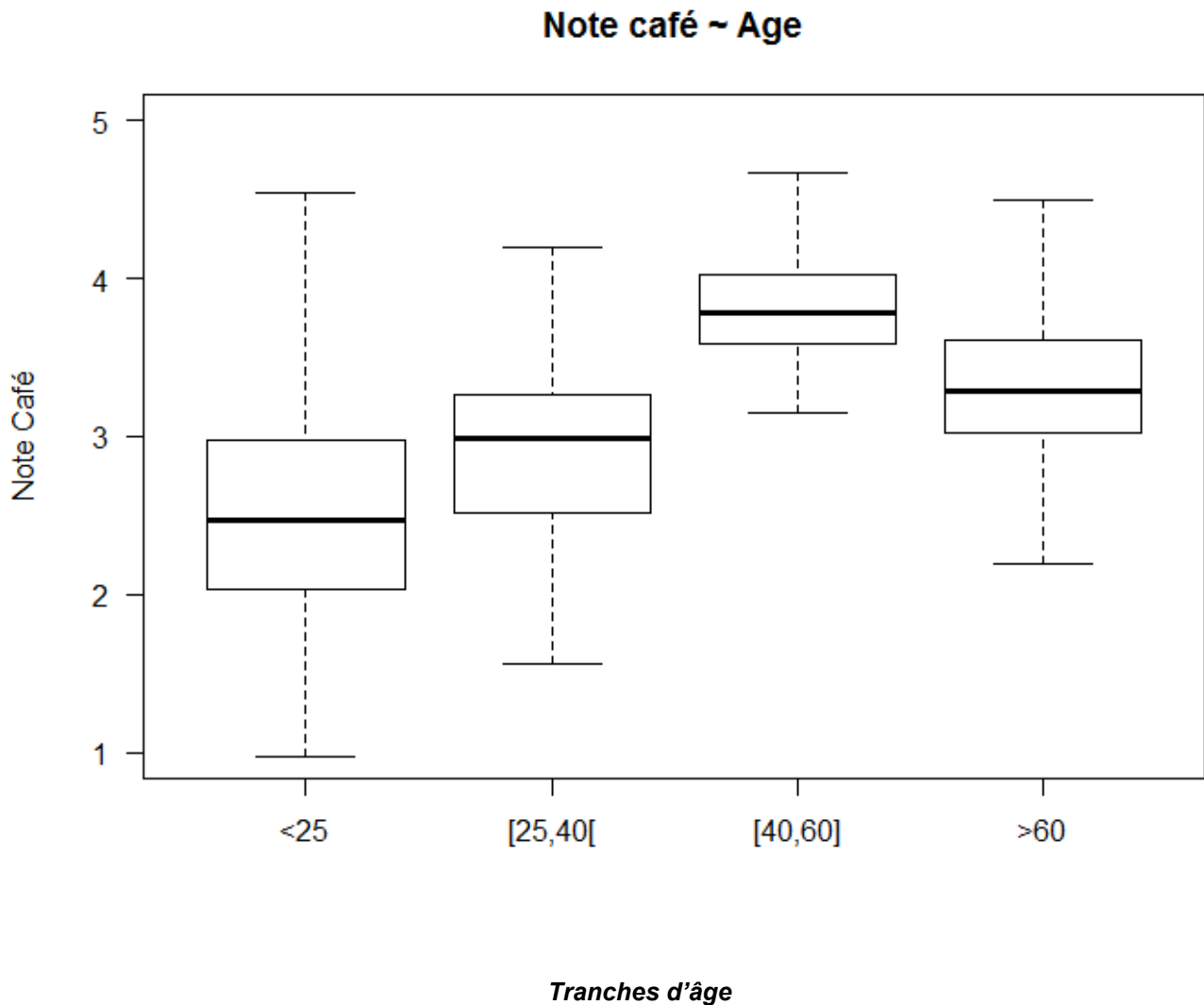


Figure 2: Graphique réalisé avec le logiciel R accompagnant l'étude du lien entre la note globale donnée au café et la tranche d'âge du consommateur

- Q8/** Vous souhaitez maintenant étudier l'éventuel lien entre la note globale donnée au café et la tranche d'âge à laquelle appartiennent les consommateurs.
Commentez brièvement les résultats représentés sur la figure 2.
De quel type de graphique s'agit-il ?
Quel test allez-vous réaliser pour établir le lien entre la note globale donnée au café et la tranche d'âge des consommateurs ? (N'oubliez pas de poser l'hypothèse nulle).
Ce test, effectué avec R, a conduit à une valeur de 0,0000435 pour le degré de signification (p_{value}). Qu'en concluez-vous ?
- Q9/** Vous considérez cette fois l'âge réel du consommateur (réponse ouverte dans le questionnaire), variant dans l'intervalle [0,120], pour tester le lien entre la note globale donnée au café et l'âge réel du consommateur. Quel type d'analyse allez-vous réaliser ?
- Q10/** En supposant que la note globale donnée à un café dégusté chez Ramon par les clientes de plus de 40 ans suit une loi normale de moyenne 3,8 et d'écart-type 0,5 :
- Calculez la probabilité qu'une femme prise au hasard de la population des clientes de plus de 40 ans donne une note globale inférieure à 2 à un café dégusté chez Ramon.
 - Quelle est la probabilité qu'une femme prise au hasard de la population des clientes de plus de 40 ans donne une note globale supérieure à 4,5 à un café dégusté chez Ramon ?

3^{ème} partie : Intégration d'informations sur la culture, la récolte et le stockage du café par l'UCIRSM



Figure 3 Culture, récolte et stockage du café par l'UCIRSM

a/ Plantation des caféiers par des membres de l'UCIRSM b/ Récolte effectuée en suivant un cahier des charges bio c/ Production rassemblée au sein de la coopérative dans des sacs de 60 kg (d'origine brésilienne, ce format de sac est devenu l'unité de mesure internationale dans le commerce du café).

Q11/ La note globale donnée au café par les consommateurs a été recodée en 4 modalités dans une variable appelée *Appréciation_café* :

« pas apprécié » (note globale < 2), « un peu apprécié mais sans plus » ($2 \leq$ note globale < 3) ;
« bien apprécié » ($3 \leq$ note globale < 4,5) ; « beaucoup apprécié » (note \geq 4,5)

On s'est intéressé à l'influence du Volume de la production du café évalué par les clients (mesurée en sacs de 60Kg) dont les 3 modalités sont : « <1000 » ; « entre 1000 et 5000 » ; « >5000 » (en bleu) sur son Appréciation.

L'analyse de l'association entre l'Appréciation du café (quels que soient son origine ou son arôme) et le Volume de la production (en sacs de 60Kg) a été réalisée avec R. Les résultats sont les suivants : $P_{\text{value}}=0,00234$; intensité de l'association : 0,375.

Quelle analyse multivariée a-t-elle ainsi été effectuée sous R ?

Quel(/s) est(/sont) le(/s) type(/s) des variables entrant dans cette analyse ?

Quelle est la variable dépendante ?

Posez l'hypothèse nulle H_0 ainsi que l'hypothèse alternative H_1 (chacune en 1 ligne seulement)

Comment analysez-vous conjointement ces deux résultats ($P_{\text{value}}=0,00234$; intensité de l'association : 0,375) ?

Est-on ici en situation d'observation ou bien en situation d'expérience ?

Selon les éléments dont vous disposez, combien y-a-t-il de degrés de liberté ?

Q12/ Un graphique plus détaillé de l'analyse de cette relation est présenté en figure 4.

Proposez une synthèse **de 7 à 8 lignes, tout au plus**, qui expose les **principaux** résultats que vous pouvez extraire de la figure 4 (n'oubliez pas de commenter les valeurs portées entre parenthèse près des axes). Comme à chaque question à laquelle vous devez répondre, allez à l'essentiel tout en essayant d'exploiter au maximum les informations dont vous disposez.

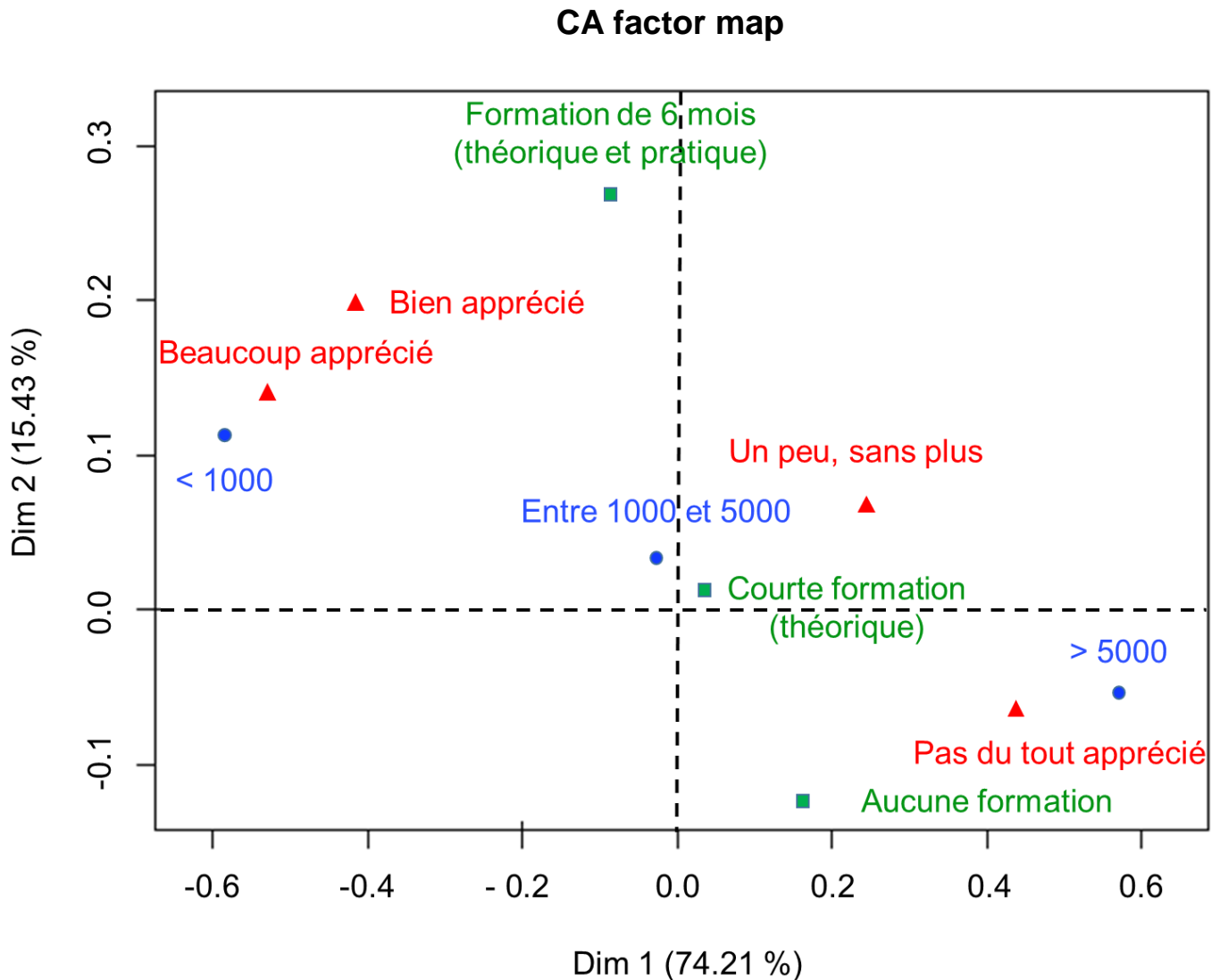


Figure 4 – **Plan factoriel** résultant de l'analyse de 439 questionnaires croisant l'**Appréciation du café** (quels que soient son origine ou son arôme) décrite en 4 modalités : « pas apprécié », « un peu apprécié mais sans plus » ; « bien apprécié » ; « beaucoup apprécié » (en rouge sur la figure) et le **Volume de la production (en sacs de 60Kg)** dont les 3 modalités sont : « <1000 » ; « entre 1000 et 5000 » ; « >5000 » (en bleu).

La **variable illustrative Formation en agronomie des ouvriers agricoles**, déclinée en 3 modalités (« aucune formation reçue », « courte formation théorique », « formation de 6 mois théorique et pratique », en vert sur la figure) a été **ajoutée pour aider à interpréter le plan factoriel** (une variable illustrative aide à l'interprétation des axes mais ne change pas le plan factoriel car elle ne contribue pas aux statistiques calculées lors de l'analyse bivariée réalisée sur les autres variables).

Q13/ La production du café s'effectue rarement dans des conditions idéales.

On s'intéresse ainsi aux répercussions de l'épandage d'un insecticide utilisé dans les plantations de café sur la santé des ouvriers agricoles. Donnez dans ce contexte un exemple d'analyse bivariée que vous pourriez réaliser ; citez la variable dépendante considérée.

(on ne vous demande aucun calcul dans cette question).

4^{ème} partie : Les dessous d'une carte

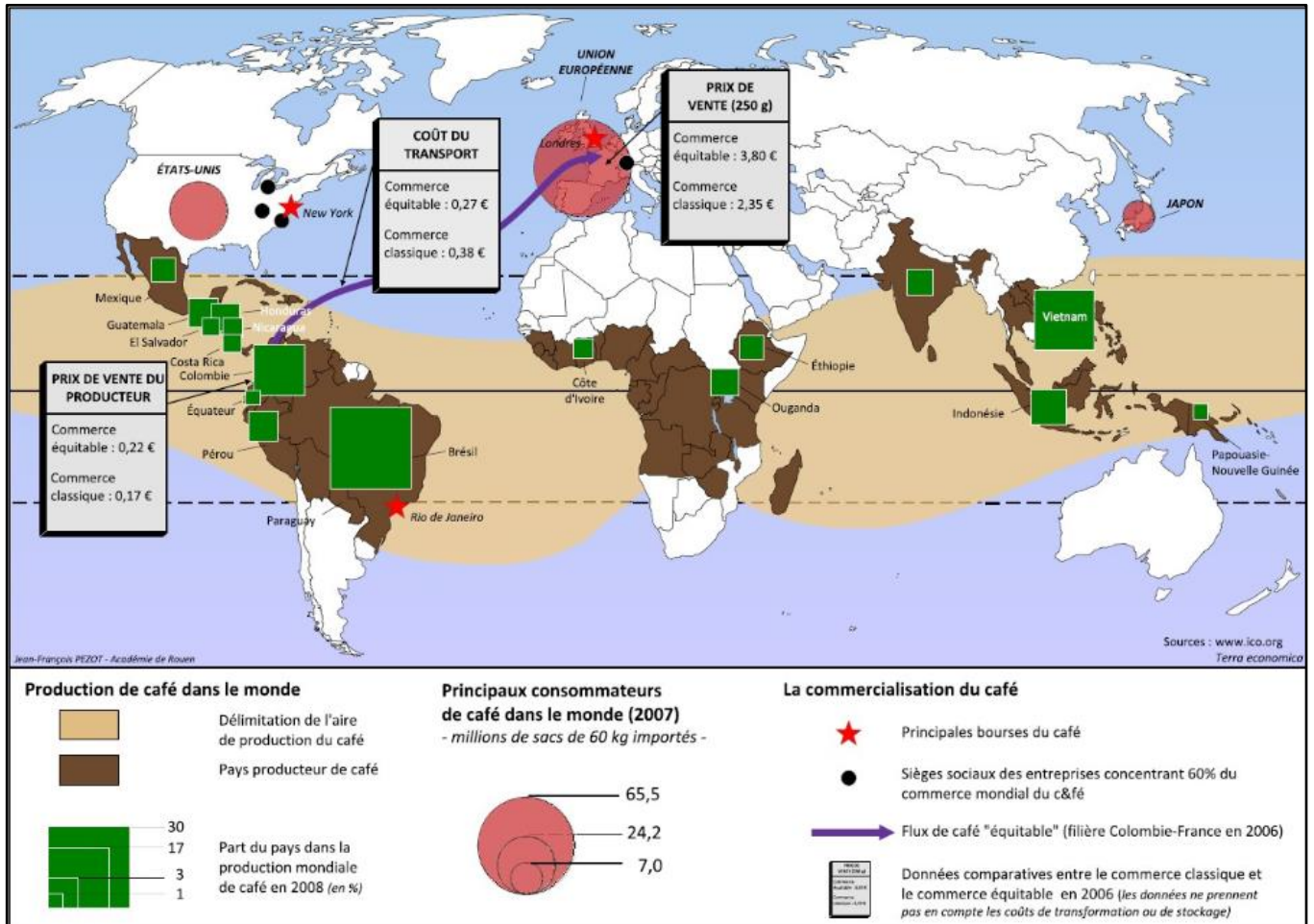


Figure 5 Carte de synthèse sur la production, le transport et la consommation du café dans le monde (en 2007). Le café est la matière première la plus commercialisée dans le monde après le pétrole. Produit exclusivement au sud, sa transformation et sa commercialisation sont très largement dominées par les pays du Nord. L'Amérique latine domine le marché avec 70% de la production mondiale, suivie par l'Asie (20%) et l'Afrique (10%).

Q14/ Analyse cartographique

Proposez une analyse statistique, pertinente et intéressante, se rapportant à la figure 5 (votre réponse ne doit pas dépasser 7 à 8 lignes). Il existe, certes, de nombreuses façons d'exploiter les informations présentées sur la carte de la figure 5 mais on vous demande, ici, de ne proposer qu'une de ces analyses (ne rentrez pas dans les détails, **sachez aller à l'essentiel**).

On ne vous demande aucun calcul en réponse à cette question.

Table de la loi normale centrée réduite

	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

