Université de Perpignan - IUT de Carcassonne
Nathalie VILLA-VIALANEIX
Année 2010/2011
Ce sujet comporte 8 pages

DUT STID, 1ère année & APPC
Statistique descriptive
Devoir 2 du 12 Janvier 2011

Important !

- Les réponses sont à donner directement sur le sujet. N'oubliez pas de noter votre nom.
- **Matériel autorisé** : Crayon, calculatrices (pas d'ordinateur), cerveau (si vous l'avez oublié, inutile de regarder sur le voisin, il n'en a pas non plus). Les téléphones portables même utilisés comme calculatrice, sont formellement interdits sur les tables.
- Sauf indication contraire, les notations utilisées sont celles du cours.
- Il sera tenu compte de la présentation et de la justification des résultats : tout résultat non justifié ne donnera lieu à aucun point.
- Le barème fourni est seulement indicatif : il peut évoluer.

Noms : ..... /40,5

Exercice 1 Bières belges vs bières françaises ..... /12

Cet exercice a été conçu grâce à l'aimable participation de Daniel C. et Étienne A. (STID2 2010/2011) qui ont bien voulu me fournir les données.

Le tableau ci-dessous reproduit **une partie** d'un jeu de données indiquant, pour 529 bières belges et françaises, à la fois, le pays d'origine de la bière ainsi qu'une note concernant son goût (de 0 à 5), apprécié par un jury de goûteurs. L'exercice porte sur **l'intégralité des bières contenues dans le fichier initial** que l'on n'a pas reproduites pour des raisons évidentes.

Numéro	Pays d'origine	Goût
1	France	3,0
2	France	2,5
3	France	1,0
4	France	2,5
5	Belgique	1,5
6	Belgique	2,0
7	Belgique	3,0
8	Belgique	3,5
9	Belgique	2,5
10	Belgique	1,5
⋮	⋮	⋮

Le but de l'exercice est de savoir si il existe une différence notable entre le goût des bières française et le goût des bières belges.[1] Vous pouvez utiliser l'espace supplémentaire page 3 en cas de besoin.

1. Quelle est la population étudiée ? Quelle est sa taille ? Quelles sont les variables étudiées ? Quels sont leurs types ?

1. Les boissons alcoolisées sont à consommer avec modération. L'abus d'alcool est dangereux pour la santé.

Réponse :

La population étudiée est l'ensemble des 529 bières ; sa taille est $N = 529$.
Il y a deux variables étudiées : le pays d'origine de la bière, de type qualitatif nominal, et la note de goût de la bière, de type quantitatif discret.

2. Quels graphiques peut-on utiliser pour représenter la distribution conjointe de ces deux variables ? En citer deux : on demande seulement les noms des graphiques mais pas de les effectuer, ni de les décrire.
Réponse :

On peut utiliser un graphe plan ou des boîtes à moustaches parallèles.

3. On note X la variable "Goût" et on donne les calculs intermédiaires suivants :

$$\sum_{u\in\{\text{Bières françaises}\}} x(u) = 490 \qquad \text{et} \qquad \sum_{u\in\{\text{Bières belges}\}} x(u) = 846,5,$$

ainsi que

$$\sum_{u\in\{\text{Bières françaises}\}} x(u)^2 = 1\ 257,5 \qquad \text{et} \qquad \sum_{u\in\{\text{Bières belges}\}} x(u)^2 = 2\ 689,75.$$

Compléter le tableau des statistiques conditionnelles suivant :

	Bières françaises	Bières belges
N_j	231	298
$\overline{X}_j$	$2,121$	$2,841$
$\sigma_j^2 = \mathrm{Var}_j(X)$	$0,945$	$0,955$

On n'oubliera pas de donner le détail des calculs.
Réponse :

Calcul des moyennes conditionnelles :

$$\overline{X}_1 = \frac{490}{231} \simeq 2,121 \qquad \text{et } \overline{X}_2 = \frac{846,5}{298} \simeq 2,841.$$

Calcul des variances conditionnelles :

$$\sigma_1^2 = \frac{1\ 257,5}{231} - 2,121^2 \simeq 0,945 \qquad \text{et } \sigma_2^2 = \frac{2\ 689,75}{298} - 2,841^2 \simeq 0,955.$$

4. D'après les statistiques conditionnelles précédentes, que peut-on dire des différences de goût entre bières françaises et bières belges ?
Réponse :

Les bières françaises sont en moyenne moins bien notées que les bières belges mais la variabilité dans la notation est légèrement moins forte pour les bières françaises que pour les bières belges.

5. À partir des résultats de la question 3, déterminer la moyenne de la note de goût pour l'ensemble des bières.
Réponse :

$$\overline{X} = \frac{231 \times 2,121 + 298 \times 2,841}{529} \simeq 2,527.$$

6. Calculer la variance inter-groupes, la variance intra-groupes puis la variance totale de la note de goût pour les bières des deux pays.
Réponse :

Variance inter-groupes

$$\mathrm{Var}_{\text{inter}} = \frac{231 \times 2,121^2 + 298 \times 2,841^2}{529} - 2,527^2 \simeq 0,125.$$

Variance intra-groupes

$$\mathrm{Var}_{\text{intra}} = \frac{231 \times 0,945 + 298 \times 0,955}{529} \simeq 0,951.$$

Variance totale

$$\mathrm{Var}(X) = \mathrm{Var}_{\text{inter}} + \mathrm{Var}_{\text{intra}} = 0,125 + 0,951 = 1,076.$$

7. En déduire le rapport de corrélation puis $\eta(X|Y)$ (où Y est le pays d'origine de la bière) et interpréter sa valeur.
Réponse :

$$\eta^2(X|Y) = \frac{\text{Var}_{\text{inter}}}{\text{Var}(X)} = \frac{0,125}{1,076} \simeq 0,116.$$

donc $\eta(X|Y) = \sqrt{0,116} \simeq 0,341$. Ainsi, le rapport de corrélation est plutôt faible : le pays d'origine de la bière (France ou Belgique) a une influence faible sur l'appréciation de son goût.

Exercice 2 Hommes et femmes dans "Les Misérables" ...../12

Le tableau ci-dessous donne pour tous les personnages du roman "Les Misérables" de Victor Hugo, le nombre de fois où chaque personnage rencontre un autre personnage dans le même chapitre, ces co-apparitions étant classées par sexe. [2] Pour des raisons de symétrie, le rôle des personnages (personnage 1 et personnage 2) dans les co-apparitions a été échangé : ainsi, lorsque Cosette est cité' en même temps que Fantine dans un chapitre, la co-apparition apparaît comme (Fantine,Cosette) et comme (Cosette,Fantine). Il y a donc $\frac{110}{2} = 55$ co-apparitions entre deux femmes dans le roman, 520 co-apparitions entre deux hommes et 245 co-apparitions entre un homme et une femme.

Personnage 2 Personnage 1	Homme	Femme	Total
Homme	1 040	245	1 285
Femme	245	110	355
Total	1 285	355	1 640

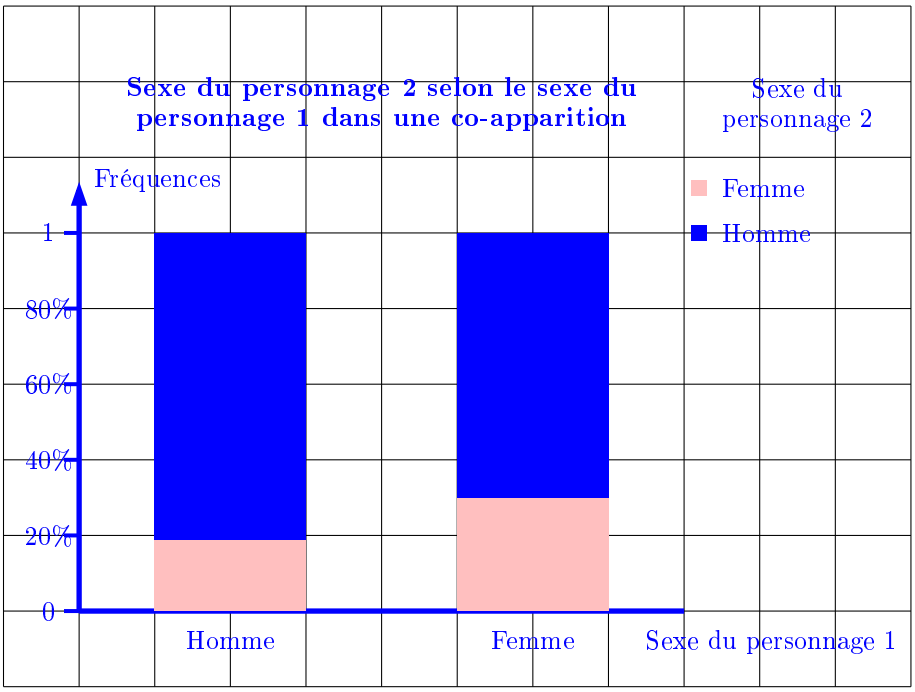
Cet exercice se propose d'étudier la question suivante : "Est-ce que les hommes ont plus tendance se lier avec d'autres hommes que les femmes dans le roman "Les Misérables" de Victor Hugo ?". Vous pouvez utiliser l'espace supplémentaire page 5 en cas de besoin.

1. Compléter le tableau ci-dessus avec **en rouge** la distribution marginale de la variable "Sexe du personnage 1" et **en bleu** la distribution marginale de la variable "Sexe du personnage 2". Précisez en noir l'effectif de la population.
2. Si X est la variable "Sexe du personnage 1" et Y la variable "Sexe du personnage 2", laquelle de ces deux distributions permet de répondre à la question : "Lorsque le personnage 1 est un homme, il a plus tendance à se lier avec d'autres hommes (ou avec des femmes) que lorsqu'il est une femme." (Cocher la bonne réponse)
   - ☒ La distribution de Y conditionnellement à X ;
   - ☐ La distribution de X conditionnellement à Y.

   Remarque : Évidemment, les deux distributions, dans notre cas particulier, sont identiques mais on demande de mettre en valeur celle qui répond précisément à la question posée, y compris si les deux distributions n'avaient pas été identiques.
3. Calculer (dans le tableau ci-dessous) la distribution conditionnelle sélectionnée précédemment et la représenter (sur le quadrillage ci-dessous).

Personnage 2 Personnage 1	Homme	Femme	Total
Homme	$\frac{1\,040}{1\,285} \simeq 80,93\%$	$\frac{245}{1\,285} \simeq 19,07\%$	1
Femme	$\frac{245}{1\,285} \simeq 69,01\%$	$\frac{110}{1\,285} \simeq 30,99\%$	1
Ensemble	$\frac{1\,285}{1\,640} \simeq 78,35\%$	$\frac{355}{1\,640} \simeq 21,65\%$	1

2. Les données relationnelles sont basées sur Knuth D., The Stanford GraphBase : A Platform for Combinatorial Computing, Addison-Wesley, Reading, MA, 1993. Les données sur les sexes des personnages sont extraites de Laurent T., Villa-Vialaneix N., Analysis of the influence of a network on the values of its nodes : the use of spatial indexes. In *MARAMI 2010*, Toulouse, France, October 11-12, 2010.



4. Expliquer pourquoi étudier l'indépendance entre X et Y permet de répondre à la question "Est-ce que les hommes ont plus tendance se lier avec d'autres hommes que les femmes dans le roman "Les Misérables" de Victor Hugo ?".
   Réponse :

X et Y sont indépendantes est équivalent au fait que la distribution de Y conditionnellement à X est identique quelle que soit la modalité de X considérée. Cela revient à dire que les hommes et les femmes ont la même tendance à se lier avec des hommes (ou avec des femmes).

5. Calculer, dans le tableau ci-dessous, les effectifs théoriques d'indépendance des variables X et Y.

Personnage 2 Personnage 1	Homme	Femme	Total
Homme	$\frac{1\ 285 \times 1\ 285}{1\ 640} \simeq 1\ 006,84$	$\frac{1\ 285 \times 355}{1\ 640} \simeq 278,16$	1 285
Femme	$\frac{1\ 285 \times 355}{1\ 640} \simeq 278,16$	$\frac{355 \times 355}{1\ 640} \simeq 76,84$	355
Total	1 285	355	1 640

6. Calculer, dans le tableau ci-dessous, les contributions au χ^2.

Personnage 2 Personnage 1	Homme	Femme
Homme	$\frac{(1\ 006,84 - 1\ 040)^2}{1\ 006,84} \simeq 1,09$	3,95
Femme	3,95	14,31

7. Quelle paire de modalités contribue le plus au χ^2 ? Celle-ci est-elle sur ou sous représentée ? Expliquer ce que cela signifie concrètement.
   Réponse :

La modalité (Femme,Femme) est celle qui contribue le plus au χ^2. Cette modalité est sur-représentée. Cela signifie que les femmes ont plus tendance à se lier à d'autres femmes que ce que l'on pourrait attendre si le sexe n'influençait pas la manière dont hommes et femmes sont liés dans le roman.

8. Calculer le χ^2 puis le C de Cramer. Interpréter cette dernière valeur concrètement.
   Réponse :

$$\chi^2 = 1,09 + 3,95 \times 2 + 14,31 = 22,60$$

et donc

$$C = \sqrt{\frac{22,60}{1\ 640}} \simeq 0,12.$$

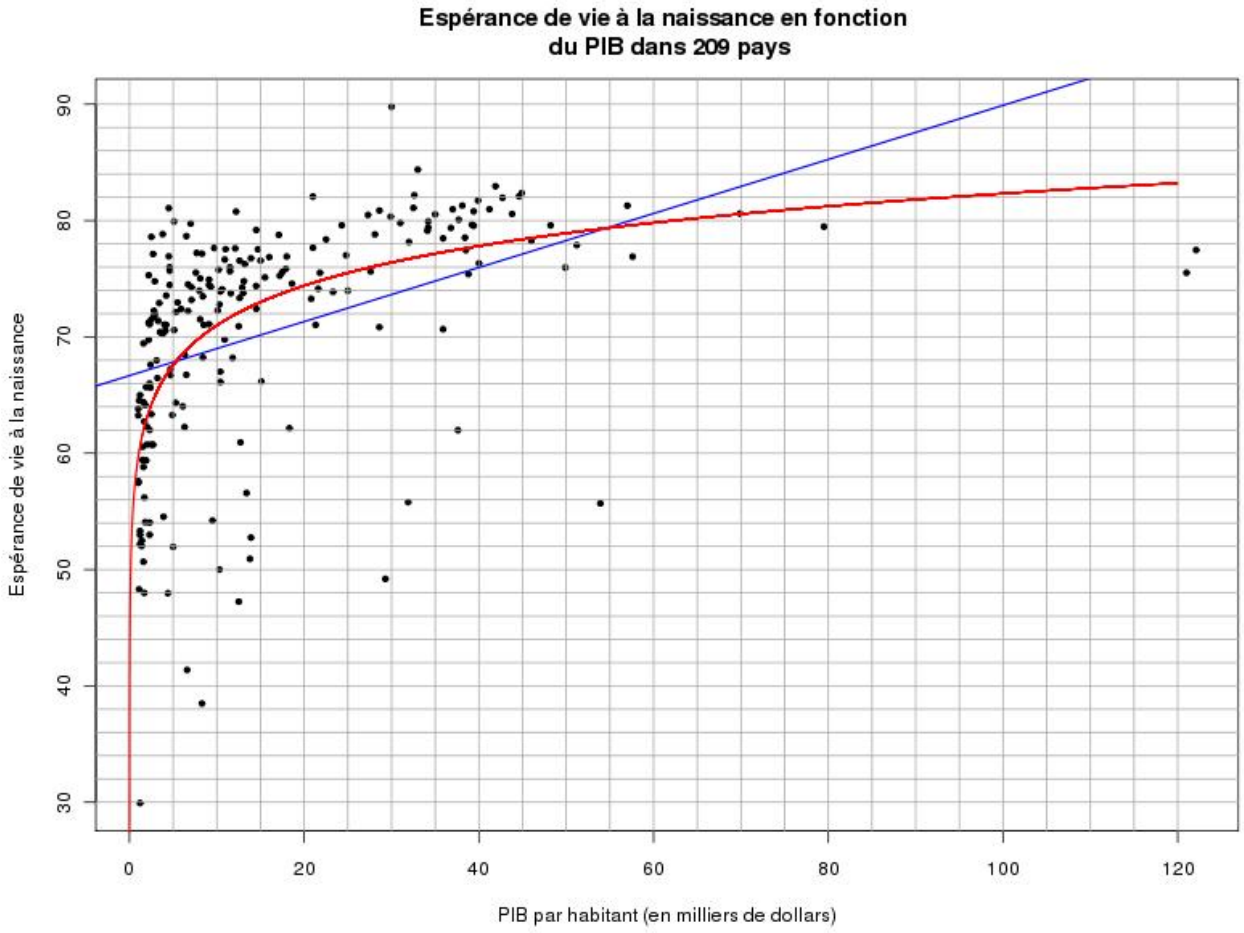
La corrélation entre X et Y est donc faible ce qui est le signe que le sexe du personnage influence peu la manière dont il est lié avec les autres personnages, hommes ou femmes dans le roman.

Exercice 3 Espérance de vie et PIB dans le monde ..... /17,5

Le tableau ci-dessous reproduit **une partie** d'un jeu de données donnant, pour 209 pays, le PIB par habitant (en milliers de dollars, estimation 2009) et l'espérance de vie à la naissance (en années, estimation 2010). La figure qui suit est le nuage de point de l'espérance de vie en fonction du PIB par habitant pour les 209 pays[3]. L'exercice porte sur **l'intégralité des pays contenus dans le fichier initial** que l'on n'a pas reproduits pour des raisons évidentes.

Pays	Espérance de vie	PIB / habitant
Albanie	77,22	7,7
Algérie	74,26	7,1
Samoa américaines	73,97	8,0
Andorre	82,36	44,9
Angola	38,48	8,3
Anguilla	80,77	12,2
Antigua et Barduba	75,26	17,2
Argentine	76,76	13,9
Arménie	72,96	5,5
Aruba	75,51	21,8
⋮	⋮	⋮

3. Les données ont été obtenues sur le site de la CIA (Central Intelligence Agency, USA), "The World Factbook" : https://www.cia.gov/library/publications/the-world-factbook/index.html. Quelques valeurs atypiques ont été supprimées du jeu de données initial.



Le but de l'exercice est de savoir si l'espérance de vie à la naissance d'un pays donné peut être déduite du PIB par habitant de ce pays. Vous pouvez utiliser l'espace supplémentaire page 8 en cas de besoin.

1. Quelle est la population étudiée ? Quelle est sa taille ? Quelles sont les variables étudiées ? Quels sont leurs types ?
   Réponse :
La population étudiée est l'ensemble des 209 pays pour lequel les données sont disponibles ; sa taille est donc $N = 209$. Les variables étudiées sont le PIB par habitant et l'espérance de vie à la naissance qui sont toutes les deux des variables quantitatives continues.

2. D'après la figure ci-dessous et le commentaire qui la suit, quelle est la régression d'intérêt ? (Cocher la bonne réponse)
   - ☒ La régression de l'espérance de vie en le PIB par habitant ;
   - ☐ La régression du PIB par habitant en l'espérance de vie.

3. On note X le "PIB par habitant" et Y l'"Espérance de vie à la naissance" et on donne les calculs intermédiaires suivants :
$$\sum_{i=1}^{N} x_i = 3\ 419,9 \qquad \text{et} \ \sum_{i=1}^{N} y_i = 14\ 728,68,$$
$$\sum_{i=1}^{N} x_i^2 = 126\ 676,8 \qquad \text{et} \ \sum_{i=1}^{N} y_i^2 = 1\ 058\ 461$$
et
$$\sum_{i=1}^{N} x_i y_i = 257\ 436,3.$$

(a) Déterminer les moyennes et les écarts types de X et Y ainsi que la covariance entre X et Y.
Réponse :

Moyennes

$$\overline{X} = \frac{3\ 419,9}{209} \simeq 16,363 \qquad \text{et } \overline{Y} = \frac{14\ 728,68}{209} \simeq 70,472$$

Variances

$$\text{Var}(X) = \frac{126\ 676,8}{209} - 16,363^2 \simeq 338,356 \qquad \text{et Var}(Y) = \frac{1\ 058\ 461}{209} - 70,472^2 \simeq 98,081$$

d'où $\sigma(X) = \sqrt{338,356} \simeq 18,39$ et $\sigma(Y) = \sqrt{98,081} \simeq 9,90$. **Covariance**

$$\text{Cov}(X,Y) = \frac{257\ 436,3}{209} - 16,363 \times 70,472 \simeq 78,605.$$

(b) En déduire le coefficient de corrélation linéaire entre X et Y, $r(X,Y)$. Commenter la valeur de ce coefficient.
Réponse :

$$r(X,Y) = \frac{78,605}{18,39 \times 9,90} \simeq 0,431.$$

La corrélation linéaire entre X et Y a une intensité moyenne : l'espérance de vie à la naissance a une dépendance linéaire moyenne avec le PIB par habitant.

(c) Calculer l'équation de la droite de régression correspondant à la question 2 et la représenter **en bleu** sur la figure du nuage de points. Que pensez-vous de la qualité de ce modèle pour faire des prévisions ?
Réponse :

La droite de régression de Y en X a pour équation $Y = aX + b$ où

$$a = \frac{78,605}{338,356} \simeq 0,232 \qquad \text{et } b = 70,472 - 0,232 \times 16,363 \simeq 66,7.$$

D'après la valeur de $r(X,Y)$ et aussi d'après la figure, le modèle est de faible qualité pour faire des prévisions.

4. On donne les calculs intermédiaires suivants :

$$\sum_{i=1}^{N} \log(x_i) = 459,237 \qquad \text{et } \sum_{i=1}^{N} \log(x_i)^2 = 1\ 290,884,$$

et

$$\sum_{i=1}^{N} \log(x_i) y_i = 33\ 750,48.$$

(a) Déterminer la moyenne et l'écart type de $\log(X)$ ainsi que la covariance entre $\log(X)$ et Y.
Réponse :

Moyenne et variance de $\log(X)$

$$\overline{\log(X)} = \frac{459,237}{209} \simeq 2,197 \qquad \text{et Var}(\log(X)) = \frac{1\ 290,884}{209} - 2,197^2 \simeq 1,348$$

d'où $\sigma(\log(X)) = \sqrt{1,348} \simeq 1,16$. **Covariance**

$$\text{Cov}(\log(X), Y) = \frac{33\ 750,48}{209} - 2,197 \times 70,472 \simeq 6,637.$$

(b) En déduire le coefficient de corrélation linéaire entre $\log(X)$ et Y, $r(\log(X), Y)$. Commenter la valeur de ce coefficient en le comparant à celui trouvé dans la question 3b.
Réponse :

$$r(\log(X), Y) = \frac{6,637}{1,16 \times 9,90} \simeq 0,577.$$

La corrélation linéaire entre $\log(X)$ et Y a une intensité moyenne mais plus forte que la corrélation entre X et Y. L'espérance de vie à la naissance a une dépendance linéaire meilleure avec le logarithme du PIB par habitant qu'avec le PIB par habitant lui-même.

(c) Calculer l'équation de la droite de régression de Y en $\log(X)$ et représenter **en rouge** cette **courbe** de régression sur la figure du nuage de points.
Réponse :

La droite de régression de Y en $\log(X)$ a pour équation $Y = a\log(X) + b$ où

$$a = \frac{6,637}{1,348} \simeq 4,92 \qquad \text{et } b = 70,472 - 4,92 \times 2,197 \simeq 59,66.$$

Pour représenter la courbe de régression sur le nuage de points, on calcule les valeurs prédites pour l'espérance de vie pour un certain nombre de valeurs du PIB par habitant selon la courbe de régression :

PIB/hab	2	5	10	15	20	30	40	60	80	120
Prévision	63,1	67,6	71,0	73,0	74,4	76,4	77,8	79,8	81,2	83,2

(d) Même si la valeur de $r(\log(X), Y)$ n'est pas très élevée, des résultats de statistique inférentielle nous informent de la bonne qualité de la régression linéaire de Y en $\log(X)$. Quelle est l'estimation de l'espérance de vie pour un pays dont le PIB par habitant est égal à \$20 000 ?
Réponse :

L'estimation de l'espérance de vie pour un pays dont le PIB par habitant est

$$\hat{y} = 4,92 \times \log{}^{a}(20) + 59,66 \simeq 74,4 \text{ ans.}$$

a. Il s'agissait du logarithme népérien, noté sur votre calculatrice "ln" ; je n'ai pas compté d'erreur pour la confusion qui aurait pu être détectée toutefois par la valeur trouvée.