Université de Perpignan - IUT de Carcassonne
Durée : 3h00
Ce sujet comporte : **5 page(s)**

Énoncé Statistique descriptive

IUT STID, 1ère année
Devoir 2 : *Correction*

17 janvier 2008

Nom : **Nathalie Villa-Vialaneix**

Attention ! Toutes les questions doivent être effectuées sur la feuille d'énoncé ! **Aucune autre copie ne sera acceptée.**
Par ailleurs, il sera **tenu compte des justifications et de la rédaction** des réponses dans la notation.

Note : Sauf indication contraire, les notations utilisées sont celles introduites en cours.

1 Liaison entre deux variables qualitatives (sur 12 points)

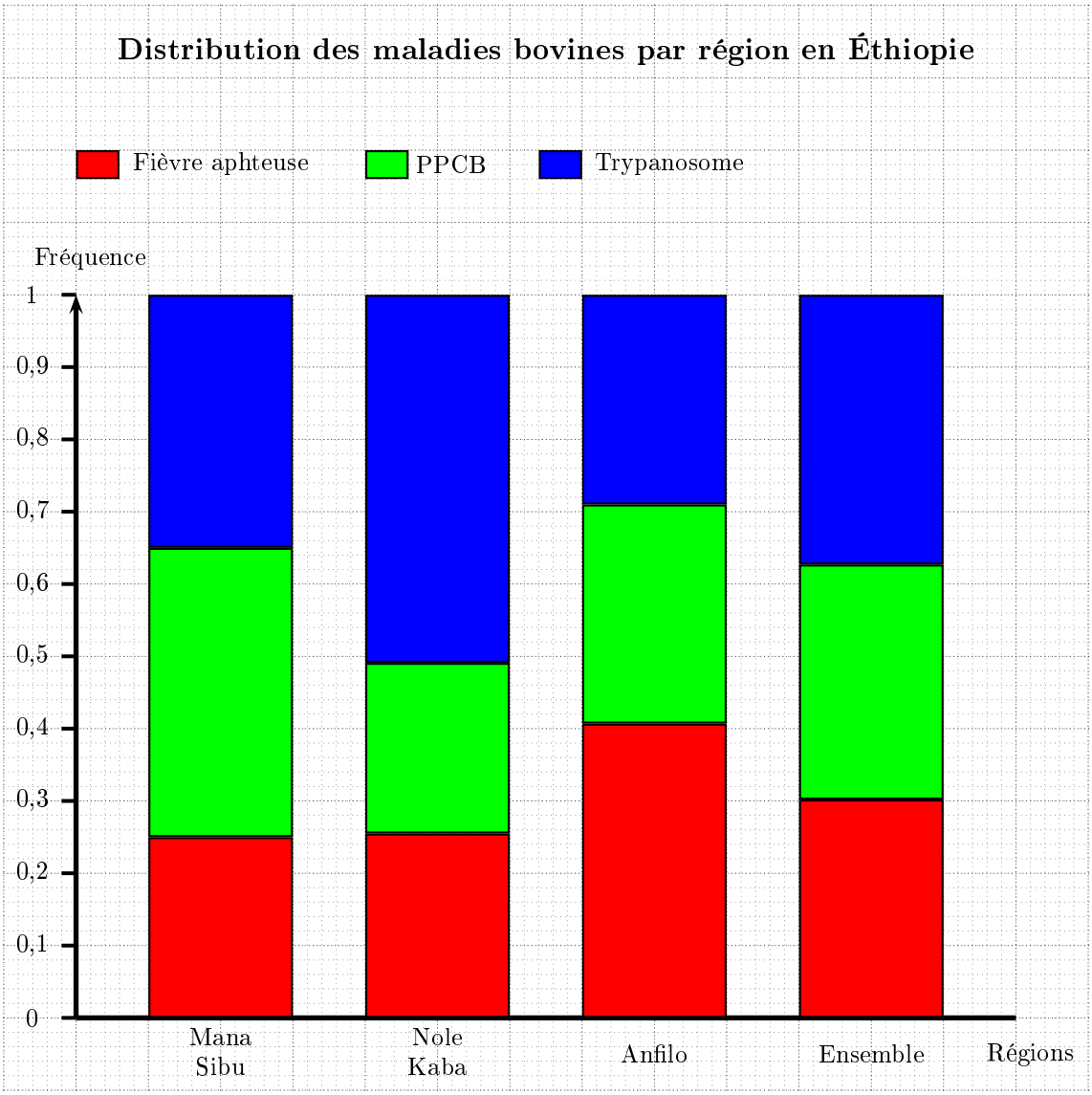
Le Tableau 1 donne le nombre de cas de chacune des trois maladies, fièvre aphteuse, PPCB et trypanosomose, observés sur les bovins de trois régions de l'Éthiopie.

TAB. 1 – Répartition des maladies bovines dans les trois régions éthiopiennes

Maladies Région	Fièvre aphteuse	PPCB	Trypano-somose	Total Total
Mana	150	240	210	600
Sibu	$\frac{150}{600} = 25\%$	$\frac{240}{600} = 40\%$	$\frac{210}{600} = 35\%$	1
Nole	140	130	280	550
Kaba	$\frac{140}{550} \simeq 25,5\%$	$\frac{130}{550} \simeq 23,6\%$	$\frac{280}{550} = 50,9\%$	1
Anfilo	220	180	140	40
	$\frac{220}{540} \simeq 40,7\%$	$\frac{180}{540} \simeq 32,3\%$	$\frac{140}{540} = 25,9\%$	1
Total	510	550	630	1690
Ensemble	$\frac{510}{1690} \simeq 30,2\%$	$\frac{550}{1690} \simeq 32,5\%$	$\frac{630}{1690} = 37,3\%$	1

1. Compléter le Tableau 1 avec, en bleu, les effectifs marginaux de la variable « Région » et en rouge, les effectifs marginaux de la variable « Maladie ». Ajouter également, en noir, la taille de la population.
2. Calculer, dans le Tableau 1, en vert, les distributions de la variable « Maladie » conditionnellement aux modalités de la variable « Région ».
3. Effectuer, sur la Figure 1 ci-dessous, un graphique représentant les distributions de la variable « Maladie » conditionnellement aux modalités de la variable « Région » (plusieurs choix sont possibles).
   Commenter ce graphique : les trois régions éthiopiennes sont-elles affectées de la même manière par ces trois maladies ?
   Les trois régions ne sont pas affectées de la même manière : la fièvre aphteuse est dominante (environ 41%) en Anfilo alors que c'est la trypanosome (environ 51%) qui l'est en Nole Kaba. En Mana Sibu, la répartition des trois maladies est pratiquement équilibrée.
4. Calculer, dans ce Tableau 2, les effectifs théoriques d'indépendance.

FIG. 1 – Graphique des distributions de la variable « Maladie » conditionnellement aux modalités de la variable « Région »



5. Déduire de la question précédence, la valeur du χ^2 et du C de Cramer. Interpréter ce résultat.
Du Tableau 2, on déduit :

$$\begin{aligned}\chi^2 &= \frac{(181,07-150)^2}{181,07}+\frac{(195,27-240)^2}{195,27}+\frac{(223,67-210)^2}{223,67}+\frac{(165,98-140)^2}{165,98}+\frac{(178,99-130)^2}{178,99}\\ &+\frac{(205,03-280)^2}{205,03}+\frac{(162,96-220)^2}{162,96}+\frac{(175,74-180)^2}{175,74}+\frac{(201,3-140)^2}{201,3}\\ &\simeq 100,04\ :\end{aligned}$$

ainsi,

$$C=\sqrt{\frac{100,04}{1690\times 2}}\simeq 0,172.$$

La liaison entre maladie et région est donc faible : les maladies sont sensiblement réparties de la même manière dans les différentes régions.

2 Liaison entre une variable qualitative et une variable quantitative (sur 9 points)

Le Tableau 3 donne le taux de pauvreté par classes d'âge pour les individus appartenant à un ménage

TAB. 2 – Répartition des maladies bovines dans les trois régions éthiopiennes

Maladies Région	Fièvre aphteuse	PPCB	Trypano--somose	*Total*
Mana Sibu	$\frac{510\times600}{1690} \simeq 181,07$	$\frac{550\times600}{1690} \simeq 195,27$	$\frac{630\times600}{1690} \simeq 223,67$	*600*
Nole Kaba	$\frac{510\times550}{1690} \simeq 165,98$	$\frac{550\times550}{1690} \simeq 178,99$	$\frac{630\times550}{1690} \simeq 205,03$	*550*
Anfilo	$\frac{510\times540}{1690} \simeq 162,96$	$\frac{550\times540}{1690} \simeq 175,74$	$\frac{630\times540}{1690} \simeq 201,30$	*540*
Total	*510*	*550*	*630*	*1690*

dont le revenu déclaré est positif ou nul et dont la personne de référence n'est pas étudiante (*Source : INSEE*, `http://www.insee.fr`).

TAB. 3 – Taux de pauvreté par classe d'âges

Classe d'âge	Taux de pauvreté	Nombre d'individus dans la classe d'âge (en milliers)	Variance	$N_j\overline{X}_j$	$N_j\overline{X}_j^2$
Moins de 18 ans	0,1548	13 340,85	*0,1309*	*2065,57*	*319,81*
18 à 24 ans	0,1750	4 809,21	*0,1443*	*841,42*	*147,21*
25 à 34 ans	0,0972	7 844,84	*0,0878*	*762,53*	*74,12*
35 à 44 ans	0,1197	8 592,78	*0,1053*	*1028,16*	*123,02*
45 à 54 ans	0,1105	8 268,12	*0,0983*	*913,77*	*100,99*
55 à 64 ans	0,0972	6 746,58	*0,0878*	*655,87*	*63,76*
65 à 74 ans	0,0695	4 936,18	*0,0647*	*343,03*	*23,84*
75 ans et plus	0,1141	4 611,08	*0,1011*	*526,00*	*60,00*
Total	⋈	59 149,64	⋈	*7 136,34*	*912,76*

1. Quelle est la population ? Sa taille ?
   La population est l'ensemble des individus appartenant à un ménage dont le revenu est positif et dont la personne de référence n'est pas étudiante. Sa taille est $N = 59149,65$ milliers d'individus.
   Dans la suite, on considère les sous-populations $\mathcal{P}_1$, ..., $\mathcal{P}_8$ induites par les 8 classes d'âges (la variable Y, « classe d'âge », est donc traitée comme une variable qualitative) et la variable $X = \mathbb{I}_{\{\text{Être en dessous du seuil de pauvreté}\}}$[1].
2. Expliquer, en justifiant brièvement, quelle est la valeur des moyennes conditionnelles $(\overline{X}_j)_{j=1,\dots,8}$ sur les populations $(\mathcal{P}_j)_j$.
   La moyenne conditionnelle est

$$\begin{aligned}\overline{X}_j &= \frac{1}{N_j}\sum_{i=1}^{N_j} x_{ij} \\ &= \frac{\textit{Nombre de personnes en dessous du seuil de pauvreté dans la classe d'âge } \mathcal{P}_j}{N_j} \\ &= \textit{Taux de pauvreté dans } \mathcal{P}_j\end{aligned}$$

   Ainsi, $\overline{X}_1 = 0,1548$, $\overline{X}_2 = 0,1750$, ...
3. Calculer, dans la quatrième colonne du Tableau 3, la valeur des variances conditionnelles $\mathrm{Var}_j(X)$ sur les populations $\mathcal{P}_j$. On justifiera préalablement, ci-dessous, la méthode de calcul utilisée.

[1] On rappelle que $\mathbb{I}_A = \begin{cases} 1 & \text{si l'individu est dans } A \\ 0 & \text{sinon} \end{cases}$

Pour déterminer la variance dans la population $\mathcal{P}_j$, on effectue le calcul suivant :

$$\begin{aligned} Var_j(X) &= \frac{1}{N_j}\sum_{i=1}^{N_j} x_{ij}^2 - \overline{X}_j^2 \\ &= \frac{1}{N_j}\sum_{i=1}^{N_j} x_{ij} - \overline{X}_j^2 \qquad \textit{car } x_{ij} \textit{ est une indicatrice} \\ &= \overline{X}_j - \overline{X}_j^2 = \overline{X}_j(1-\overline{X}_j) \end{aligned}$$

où $\overline{X}_j$ est donnée par la deuxième colonne du tableau d'après la question précédente.

4. En utilisant, au besoin, les colonnes vides du Tableau 3, déterminer les valeurs de $\text{Var}_{\text{inter}}$ et de $\text{Var}_{\text{intra}}$. En déduire la valeur de $\text{Var}(X)$.
La moyenne de X sur l'ensemble de la population est, d'après le Tableau 3,

$$\overline{X} = \frac{7\ 136,34}{59\ 149,64} \simeq 0,1206\ ;$$

ainsi,

$$Var_{inter} = \frac{912,76}{59\ 149,64} - 0,1206^2 \simeq 0,0009.$$

Par ailleurs, toujours d'après le Tableau 3,

$$\begin{aligned} Var_{intra} &= \frac{1}{59\ 149,64}(13\ 340,65 \times 0,1309 + 4\ 809,21 \times 0,1443 + 7\ 844,84 \times 0,0878 + \\ &\quad 8\ 592,72 \times 0,1053 + 8\ 268,12 \times 0,0983 + 6\ 746,58 \times 0,0878 + 4\ 936,18 \times 0,0647 + \\ &\quad 4\ 611,08 \times 0,1011) \\ &\simeq 0,1052. \end{aligned}$$

Enfin, $Var = Var_{inter} + Var_{intra} =$ $\boxed{0,1061}$.

5. Calculer le rapport de corrélation $\eta(X|Y)$. Interpréter cette valeur de manière concrète.
Le rapport de corrélation est égal à

$$\eta(X|Y) = \sqrt{\frac{0,0009}{0,1061}} \simeq 0,091.$$

Ainsi, la liaison entre taux de pauvreté et classe d'âge est faible : on retrouve sensiblement le même taux de personnes en dessous du seuil de pauvreté dans toutes les classes d'âges.

3 Concentration (sur 11 points)

Le Tableau 4, ci-dessous, donne les valeurs de l'effort en recherche et développement (DIRD) par habitant des 23 régions françaises (Outre-mer incluse), rangées par ordre croissant, ainsi que la population (en fréquence par rapport à la population nationale) de chacune de ces régions.

Source : « La Vie de la Recherche Scientifique », numéro 371, publication de la FSU.

On considère ici la population nationale et la variable X : « part de DIRD de l'individu considéré ». La deuxième colonne du tableau correspond donc à des fréquences, f_i, et la troisième aux modalités de la variable X, x_i.

1. À quoi correspond concrètement la valeur $f_i x_i$? Calculer ces valeurs dans la quatrième colonne du Tableau 4.
$f_i x_i$ est la valeur de la DIRD de la région considérée.

2. Compléter le tableau précédent de manière à obtenir les coordonnées des points de la courbe de Lorenz, (f_i^*, v_i) (vous utiliserez, à votre choix, toutes ou partie des colonnes vides du Tableau 4).
Remarque : On ne demande pas de détailler toutes les opérations effectuées pour toutes les cases du tableau mais simplement de préciser, ci-dessous, les opérations effectuées pour le calcul de [1] et [2] (et des éventuels calculs intermédiaires figurant dans d'autres colonnes du tableau pour cette ligne).
Le tableau a été rempli de la manière suivante :
- *Pour [1], $0,004 + \frac{6,4}{100} = 0,068$;*
- *Pour [2], on calcule d'abord [3] $= 13 + 567 = 580$ puis [2] $= \frac{580}{35\ 179} \simeq 0,0165$.*

TAB. 4 – DIRD et populations régionales

Région	Population (%)	DIRD/hab (en millions € par % pop. nat.)	$f_i x_i$	f_i^*	$(f_i x_i)^*$	v_i
Corse	0,4	32,5	13	*0,004*	*13*	*0,0004*
Nord-Pas-de-Calais	6,4	88,59	567	*0,068*[1]	*580*[3]	*0,0165*[2]
Outre-Mer	2,9	96,55	280	*0,097*	*860*	*0,0244*
Limousin	1,1	107,27	118	*0,108*	*978*	*0,0278*
Champagne-Ardenne	2,1	108,57	228	*0,129*	*1 206*	*0,0343*
Poitou-Charentes	2,7	109,63	296	*0,156*	*1 502*	*0,0427*
Bourgogne	2,6	130	*338*	*0,182*	*1 840*	*0,0523*
Basse-Normandie	2,3	144,35	332	*0,205*	*2 172*	*0,0614*
Lorraine	3,7	146,49	542	*0,242*	*2 714*	*0,0771*
Picardie	3	148	444	*0,272*	*3 158*	*0,0898*
Pays de la Loire	5,4	154,26	833	0,326	*3 991*	0,113
Aquitaine	5	218,4	1092	0,376	*5 083*	0,145
Haute-Normandie	2,9	228,62	663	0,405	*5 746*	0,163
Centre	4,1	230,24	*944*	0,446	*6 690*	0,190
Bretagne	4,8	235,83	*1 132*	0,494	*7 822*	0,222
Alsace	2,9	245,86	*713*	0,523	*8 535*	0,243
France-Comté	1,8	278,89	*502*	0,541	*9 037*	0,257
PACA	7,4	312,16	*2 310*	0,615	*11 347*	0,323
Auvergne	2,1	312,38	*656*	0,636	*12 003*	0,341
Languedoc-Roussillon	3,9	326,92	*1 275*	0,675	*13 278*	0,377
Rhône-Alpes	9,5	438,32	*4 164*	0,770	*17 442*	0,496
Midi-Pyrénées	4,3	646,05	*2 778*	0,813	*20 220*	0,575
Île de France	18,4	812,99	*14 959*	0,997	*35 179*	1,000
Total	100	⋈	*35 179*	⋈	⋈	⋈

3. La Figure 2, ci-dessous, est la courbe de Lorenz de la variable « PIB régional par habitant » pour la population française. Que permet, concrètement, de mesurer cette courbe ? Interpréter la forme de cette figure.
   Cette courbe permet de mesurer la concentration du PIB régional, c'est-à-dire le fait qu'un petit nombre d'individus issus des régions les plus riches, profitent, ou non, de la majorité du PIB régional total. La courbe nous montre que cette concentration est faible, car la courbe est proche de la diagonale $y = x$ (représentée en pointillés) : la répartition du PIB régional est assez égalitaire.
4. Quel pourcentage de la population nationale, située dans les régions les plus riches, concentre 50% du PIB régional ? (On fera apparaître les traits de construction sur le graphique.)
   Environ $1 - 0,58 = 42\%$ de la population nationale des régions les plus riches concentre 50% du PIB régional.
5. Compléter la Figure 2 avec la courbe de Lorenz de la DIRD par habitant. Que peut-on dire en comparant ces deux courbes ?
   La concentration de la DIRD est plus forte que celle du PIB : un faible nombre d'individus des régions les plus investies supportent une part importante de l'effort régional en recherche et développement.
6. Déterminer l'indice de Gini de la DIRD par habitant. Interpréter.
   L'indice de Gini de la DIRD par habitant se calcule à partir des quantités $f_i(v_i + v_{i-1})$; pour la ligne correspondant au Nord-Pas-de-Calais, il s'agit de $0,064 \times (0,0165 + 0,0004) \simeq 0,001$. Ainsi,

$$\begin{aligned} G &= 1 - (0 + 0,001 + 0,001 + 0,001 + 0,001 + 0,002 + 0,003 + 0,003 + 0,005 + 0,011 + 0,013 + 0,009 \\ &\quad + 0,115 + 0,020 + 0,014 + 0,009 + 0,043 + 0,014 + 0,028 + 0,083 + 0,046 + 0,290) \\ &\simeq 0,385 \end{aligned}$$

   La concentration de la DIRD par habitant est moyenne : si l'effort en recherche en développement n'est pas parfaitement également réparti, les différences individuelles restent modérées.

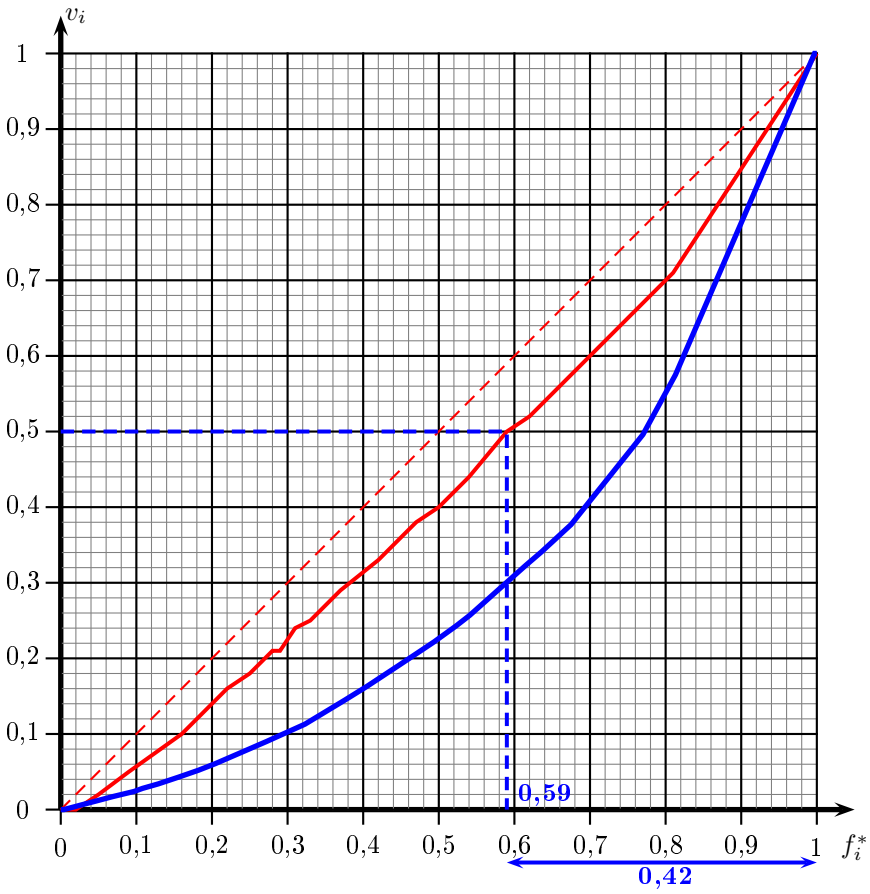


FIG. 2 – Courbe de Lorenz du PIB régional par habitant sur la population française