Nathalie Villa-Vialaneix

Livret d'exercices de

Statistique Descriptive II (M1201)

Année scolaire 2013/2014

Université de Perpignan Via Domitia, IUT
STatistique et Informatique Décisionnelle (STID)

Table des matières

1 Généralités sur la statistique descriptive bivariée

1.1 Distributions jointes et conditionnelles

Ce document est issu des statistiques de l'INSEE et intitulé « La population française selon la nationalité et le lieu de naissance en 1999 » (effectifs en milliers) :

Lieu de naissance Nationalité	France	Étranger
Français de naissance	51 340	1 560
Français par acquisition	800	1 560
Étrangers	510	2 750

1. Quelle est la population ? Sa taille ?
2. Quelles sont les variables étudiées ? Leurs types ?
3. Compléter la table de contingence précédente avec les distributions marginales.
4. Déterminer les distributions conditionnelles des deux variables. Laquelle nous permet de déterminer si le pourcentage de personnes nées en France est plus fort parmi les Français par acquisition ou parmi les étrangers ?

1.2 Distributions jointes et conditionnelles

Ce document est issu des statistiques de l'INSEE et intitulé « Répartition des séjours d'été par mode d'hébergement et par catégorie socio-professionnelle en 2004 ».

Type de vacances CSP du chef de famille	Hôtel	Location	Résidence secondaire	Parents et amis	Camping	Autres
Agriculteurs	17	15	2	27	34	6
Artisans & commerçants	15	20	7	32	19	7
Cadres	14	18	8	43	9	8
Professions intermédiaires	11	18	4	44	16	8
Employés	11	15	3	47	18	6
Ouvriers	7	17	4	45	22	5
Retraités	17	11	14	39	11	8
Autres inactifs	5	10	1	59	20	5

1. Quelle est la population étudiée ?
2. Quelles sont les variables étudiées ? Leurs types ?
3. Quel est le nom de la (des) distribution(s) présentée(s) ici ?
4. Peut-on, à partir de celles-ci, retrouver la table de contingence (des fréquences) associée à ce problème ?
   Quelle information supplémentaire faudrait-il avoir pour y parvenir ?
5. On donne le tableau suivant :

CSP	Fréquence
Agriculteurs	0,020
Artisans et commerçants	0,044
Cadres	0,108
Professions intermédiaires	0,173
Employés	0,217
Ouvriers	0,178
Retraités	0,255
Autres inactifs	0,005

   Retrouver, à partir de ces deux tableaux, la table de contingence des deux variables étudiées. Calculer les distributions marginales.
6. Calculer la distribution de la variable « CSP » conditionnellement à la modalité « Camping ». Au vu de celle-ci et du tableau initial, qui prévilégie le plus le camping comme mode de vacances ? Qui a-t-on le plus de chances de rencontrer lorsque l'on va dans un camping ?

2 Représentations graphiques

2.1 Variables qualitatives

On reprend les données de l'exercice 1.1 page 5 :

Lieu de naissance / Nationalité	France	Étranger	Ensemble
Français de naissance	51 340	1 560	52 900
Français par acquisition	800	1 560	2 360
Étranger	510	2 750	3 260
Total	52 650	5 870	58 520

1. Représenter graphiquement la distribution conjointe des deux variables.
2. Représenter graphiquement les distributions conditionnelles aux modalités de chacune des deux variables. On rappelle :

Distribution de la Nationalité conditionnellement au Lieu de naissance

Lieu de naissance / Nationalité	France	Étranger	Total
Français de naissance	97,51 %	26,58 %	90,40 %
Français par acquisition	1,52 %	26,58 %	4,03 %
Étranger	0,97 %	46,85 %	5,57 %
Total	1	1	1

Distribution du Lieu de naissance conditionnellement à la Nationalité

Lieu de naissance / Nationalité	France	Étranger	Total
Français de naissance	97,05 % %	2,95 %	1
Français par acquisition	33,90 % %	66,10 %	1
Étranger	15,64 %	84,36 %	1 %
Ensemble	89,97 %	10,03 %	1

3. Est-il plus fréquent d'acquérir la nationalité française lorsque l'on est né en France ou bien lorsque l'on est né à l'étranger ? (Quel graphique permet cette conclusion ?)

2.2 Variable qualitative et variable quantitative

On donne, pour les mois allant de novembre 2006 à octobre 2007 (inclus), le cours du supercarburant et celui du blé :

	Super	Blé
Octobre 2007	733,3	853,7
Septembre 2007	729,6	863,0
Août 2007	703,2	691,8
Juillet 2007	751,6	613,3
Juin 2007	763,1	573,5
Mai 2007	773,1	486,0
Avril 2007	722,5	471,2
Mars 2007	620,6	459,5
Février 2007	542,8	464,7
Janvier 2007	492,2	466,1
Décembre 2006	552,1	491,1
Novembre 2006	520,6	487,5

1. Quelle est la population ? Sa taille ? Quelles sont les variables ? Leurs types ?
2. Déterminer les indices base 100 en octobre 2007 des deux séries statistique.
   Remarque : L'indice base 100 en octobre 2007 de la variable X est $\frac{X}{\textit{Valeur de } X \textit{ en octobre 2007}} \times 100$.
3. Déterminer le diagramme chronologique en ligne de ces deux séries statistiques à partir des indices calculés à la question précédente. Commenter.
4. Déterminer le graphique de la distribution conjointe des deux variables. Commenter.

3 Étude de la liaison entre une variable qualitative et une variable quantitative

3.1 Problème

Les données de cet exercice sont issues du site de l'ORAMIP (Observatoire Régional de l'Air en Midi-Pyrénées, http://www.oramip.org).

3.1.1 Influence d'un incinérateur sur le taux de métaux lourds dans l'air

Dans le tableau ci-dessous, sont relevées les valeurs du taux de plomb (en ng/m^3) dans l'air pour trois stations ORAMIP de Toulouse : Eisenhower et Chapitre, qui sont situées à proximité de l'incinérateur de déchets du Mirail, et Berthelot, situé en zone urbaine, qui est caractéristique de l'air respiré par l'ensemble de la population toulousaine.

Station		
Eisenhower	Chapitre	Berthelot
7,6	9,1	7,5
21,4	13,1	9,3
12,4	10,5	10,2
15,8	8,3	20,3
14,8	17,4	10,3
5,9	9,4	16,4
12,5	7,8	13,3
11,3	12,2	14,9

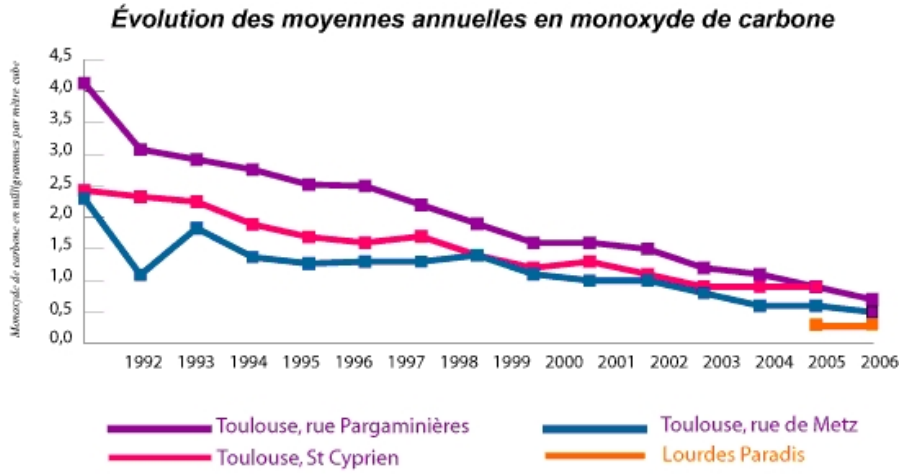
On considère ici la population des relevés de taux de plomb de taille 24 et les variables X, « Taux de plomb dans l'air », et Y « Station ».

1. De quels types sont les variables considérées ?
2. Sur cet échantillon de relevés, peut-on considérer qu'il y a un lien important entre la station de relevé et le taux de plomb dans l'air ?
3. Si on considère maintenant la variable Z, « Proximité de l'incinérateur ? », de type qualitative nominale, peut-on considérer qu'il y a un lien important entre X et Z ?

3.1.2 Influence de la station sur le taux de monoxide de carbone

Sur le graphique ci-dessous, sont données les moyennes annuelles de taux de monoxide de carbone (en mg/m^3) dans l'air pour trois stations toulousaines : Toulouse

Pargaminières (située au centre de Toulouse, rue étroite, immeubles hauts, circulation importante), Toulouse Saint Cyprien (située à proximité du centre ville sur un grand carrefour de circulation) et Toulouse rue de Metz (située au centre de Toulouse, rue peu dégagée, circulation très importante).



On considère ici la population des années (de taille 15) et les variables X, « Moyenne annuelle du taux de monoxide de carbonne », et Y, « Station ».

1. Faire un tableau donnant, pour chacune des trois stations toulousaines, la valeur des relevés effectués.
2. Peut-on considérer qu'il existe un lien important entre le taux de monoxide de carbonne et la localisation géographique du relevé dans Toulouse ?

3.2 Élections

Les données suivantes sont les résultats, obtenus au premier tour de la présidentielle de 2007, par Ségolène Royal (Parti Socialiste) et Jean-Marie Le Pen (Front National) [1] pour divers départements.

Département	Nombre de votants	Taux de votes pour SR	Taux de votes pour JML
Aude	216 882	30,70 %	13,20 %
Corrèze	159 909	29,73 %	7,58 %
Creuse	81 585	29,02 %	8,68%
Hérault	593 411	26,05 %	13,35 %
Gard	416 170	23,31 %	15,44 %
Haute-Vienne	225 765	31,46 %	8,56 %
Lozère	51 925	22,72 %	9,44%
Pyrénées orientales	263 862	24,82 %	14,21 %

1. *Sources* : Site du ministère de l'intérieur http://www.interieur.gouv.fr

On considère, dans cet exemple, la population des votants (suffrages exprimés) des régions Limousin et Languedoc-Roussillon et les variables $X = \mathbb{I}_{\{\text{Avoir voté pour Ségolène Royal}\}}$[2], $X' = \mathbb{I}_{\{\text{Avoir voté pour Jean-Marie Le Pen}\}}$[1], Y, « Département » et Y', « Région ». X et X' sont considérées ici, comme quantitatives (elles « comptent » le nombre de votes pour les deux candidats).

1. Quelle est la taille de la population ?
2. On considère les sous-populations $(\mathcal{P}_i)_{i=1,\dots,8}$ définies par Y. À quoi est égal $\overline{X}_i$ pour $i = 1, \dots, 8$?
3. En déduire $\overline{X}$.
4. Pour une variable valant 0 ou 1 avec une proportion p de 0, la variance est égale à $p(1-p)$. À quoi est égal σ_i^2 pour $i = 1, \dots, 8$?
5. En déduire $\text{Var}_{\text{inter}}$ et $\text{Var}_{\text{intra}}$ puis σ^2.
6. Peut-on dire qu'il existe un lien important entre X et Y ?
7. Mêmes questions entre X' et Y, entre X et Y' et entre X' et Y'. Conclusion ?

2. $\mathbb{I}_A = \begin{cases} 1 & \text{si l'individu appartient à } A \\ 0 & \text{sinon} \end{cases}$

4 Étude de la liaison entre deux variables qualitatives

4.1 χ^2, C de Cramer

On reprend les données de l'exercice 1.1 :

Lieu de naissance / Nationalité	France	Étranger	Total
Français de naissance	51 340	1 560	52 900
Français par acquisition	800	1 560	2 360
Étranger	510	2 750	3 260
Total	52 650	5 870	58 520

Déterminer l'indice du χ^2 et le coefficient C de Cramer. Commenter.

4.2 Un exemple de paradoxe de Simpson

Les données suivantes rapportent les résultats du jugement de 4 764 homicides jugés en Floride entre 1973 et 1979. Ces données ont été publiées dans le New York Times du 11 mars 1979.

		Sentence	
Meurtrier	Victime	Peine de mort	Autre peine
Blanc	Blanc	72	2074
	Noir	0	111
Noir	Blanc	48	239
	Noir	11	2209

1. Quelle est la population ? Sa taille ? Quelles sont les variables ? Leur type ?
2. Effectuer la table de contingence des variables « Couleur du meurtrier » et « Type de peine ». À partir de cette table, déterminer la distribution de la variable « Sentence » conditionnellement à la variable « Couleur du meurtrier ». Calculer le χ^2 puis le C de Cramer. Que pense-t-on pouvoir conclure ?
3. Effectuer la table de contingence des variables « Couleur du meurtrier » et « Couleur de la victime ». Calculer le χ^2 puis le C de Cramer. Conclusion ?

4. Enfin, effectuer, *seulement pour les victimes blanches*, la table de contingence des variables « Couleur du meurtrier » et « Sentence ». En déduire la distribution de la variable « Sentence » conditionnellement à la variable « Couleur du meurtrier ». Comparez à la distribution conditionnelle obtenue dans la Question 2 et, en vous appuyant sur les résultats de la question précédente, expliquez ce phénomène.

5 Régression linéaire

5.1 Série chronologique

Ces données correspondent à l'évolution du nombre d'habitants en France depuis 1990 *(source : INSEE http://www.insee.fr)*.

Année	Population (en milliers)
2007	61542
2006	61167
2005	60825
2004	60462
2003	60067
2002	59660
2001	59249
2000	58850
1999	58497
1998	58299
1997	58116
1996	57936
1995	57753
1994	57565
1993	57369
1992	57111
1991	56841
1990	56577

On considère, dans la suite, la régression de la population (notée Y) par rapport à l'année (notée T).

1. Effectuer le nuage de points des deux variables. Que peut-on en dire ?
2. Déterminer la moyenne et l'écart type de Y.
3. On pose $X = T - 1990$.
   a) Calculer les valeurs de X ainsi que sa moyenne et son écart type.
   b) Trouver une relation entre $\overline{X}$ et $\overline{T}$ ainsi qu'entre σ_X et σ_T.
   c) En déduire $\overline{T}$ et σ_T.

4. Calculer $\text{Cov}(X, Y)$. Après avoir exprimé $\text{Cov}(T, Y)$ en fonction de $\text{Cov}(X, Y)$, en déduire la valeur de $\text{Cov}(T, Y)$.
5. Calculer $r(T, Y)$.
6. Déterminer l'équation de la droite de régression de Y en T.
7. Si l'évolution de la population se poursuit au même rythme, quelle population peut-on prévoir en France en 2010 ? en 2015 ?

5.2 Joueurs de football

Voici la taille et le poids des joueurs de Toulouse (TFC, Ligue 1 de football, en 2007/2008) :

Taille (en m)	Poids (en kg)
1,74	68
1,78	75
1,82	78
1,84	79
1,84	78
1,86	82
1,85	77
1,80	73
1,88	79
1,80	77
1,86	72
1,80	65
1,86	84
1,80	74
1,84	76
1,82	74
1,80	70
1,89	83
1,76	75
1,85	78
1,72	68
1,80	68

1. On note, respectivement, X, la variable « Taille » et Y la variable « Poids ». Construire, sur deux graphiques séparés les nuages de points de Y en fonction de X et de Y en fonction de X^2.
2. Calculez $r(Y, X^2)$. Conclusion ?
3. Déterminer la droite de régression de Y en X^2. On appelle Indice de Masse Corporelle, la quantité

$$IMC = \frac{\text{Poids}}{\text{Taille}^2}.$$

Un individu a une corpulence normale si son IMC est compris entre 18,5 et 25. Les joueurs de football semblent-ils avoir une IMC normale ?

4. Quel poids peut-on prévoir pour un joueur de $1,75$ mètres ?

5.3 Oignons

Les données suivantes, extraites de *Ratkowsky D.A. (1983) Nonlinear regression modeling. Marcel Dekker. New York.*, sont les poids secs (moyens) d'oignons à différentes périodes de leur développement (temps en semaines).

Poids secs	Durée de vie
16,08	1
33,83	2
65,8	3
97,2	4
191,55	5
326,2	6
386,87	7
520,53	8
590,03	9
651,92	10

Dans la suite, on notera Y la variable « Poids sec des oignons » et X la variable « Durée de vie ». Le but est de modéliser le poids sec en fonction de la durée de vie.

1. Effectuer le nuage de points de ces données. Peut-on penser, au vu de celui-ci, qu'une régression linéaire est bien adaptée à la question ?
2. Calculer $r(X, Y)$.
3. Déterminer les valeurs de la variable $Y' = \ln\left(\left(\frac{Y}{700}\right)^{-1,28} - 1\right)$ puis calculer $r(Y', X)$. Conclusion ?
4. Déterminer l'équation de la droite de régression de Y' en X. À quelle courbe de régression cette droite correspond-elle pour les variables Y et X ? Tracer cette courbe sur le nuage de points de la première question.
5. Évaluer le poids d'un oignon sec cueilli à $10,5$ semaines.