Université de Perpignan - IUT de Carcassonne
Nathalie VILLA-VIALANEIX
Année 2009/2010 - Nb de pages : 6

IUT STID, 1ère année et APPC

Statistique descriptive

Devoir du Jeudi 13 janvier 2010

Important : Les réponses sont à donner directement sur le sujet. N'oubliez pas de noter votre nom.
Par ailleurs, les documents sont interdits et les calculatrices autorisées. Sauf indication contraire, les notations utilisées sont celles du cours.

Nom : .....

Exercice 1 Concentration des revenus en France

Remarque : Les réponses doivent être justifiées. Si l'espace fourni pour répondre n'est pas suffisant, vous pouvez utiliser l'espace situé page 2.

Le tableau ci-dessous donne la répartition des revenus salariaux et d'allocation chômage des 9 premiers déciles de la population française (étudiants exclus) en 2007, ainsi que divers calculs permettant la mesure de la concentration des revenus en France.

Source : INSEE (http://www.insee.fr)

Revenus (classe $\mathcal{C}_i$)	c_i	f_i	f_i^*	$f_i c_i$	v_i	$f_i(v_i + v_{i-1})$
[0; 4 786[	*2 393*	11,11%	*11,11%*	*266*	*0,0164*	*0,0018*
[4 786; 9 363[	7 074,5	11,11%	22,22%	786	0,0648	0,0090
[9 363; 12 974[	11 168,5	11,11%	33,33%	1 241	0,1413	0,0229
[12 974; 15 420[	14 197	11,11%	44,44%	1 577	0,2386	0,0422
[15 420; 17 512[	16 466	11,11%	55,56%	1 830	0,3514	0,0655
[17 512; 19 795[	18 653,5	11,11%	66,67%	2 073	0,4791	0,0923
[19 795; 22 585[	*21 190*	11,11%	*77,78%*	*2 354*	*0,6242*	*0,1226*
[22 585; 26 525[	24 555	11,11%	88,89%	2 728	0,7924	0,1574
[26 525; 34 081[	30 303	11,11%	100,00%	3 367	1,0000	0,1992
Total	∅	1	∅	16 222	∅	*0,7129*

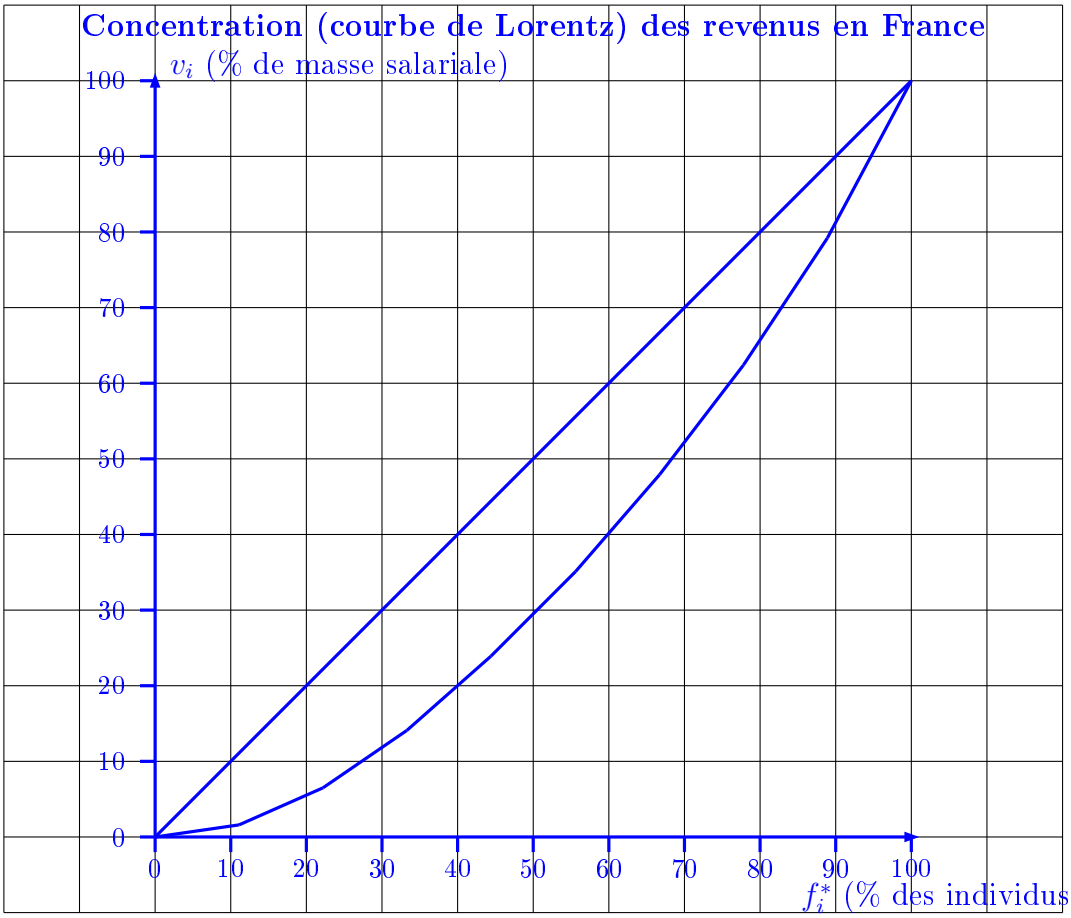
1. Complétez la première et la septième lignes du tableau avec les valeurs de c_i, f_i^*, $f_i c_i$, v_i et $f_i(v_i + v_{i-1})$ manquantes. Vous détaillerez ci-dessous les calculs effectués.
   Réponse :

   Pour la première ligne : $c_1 = \frac{0+4\ 786}{2} = 2\ 393$, $f_1^* = f_1 = 0,1111 = 11,11\%$, $f_1 c_1 = 0,1111 \times 2\ 393 = 266$, $v_1 = \frac{f_1 c_1}{\sum_i f_i c_i} = \frac{266}{16\ 222} \simeq 0,0164$ et $f_1 v_1 = 0,1111 \times 0,0164 \simeq 0,0018$.
   Pour la septième ligne : $c_7 = \frac{19\ 262+22\ 137}{2} = 21\ 190$, $f_7^* = 0,6667 + 0,1111 = 0,7778 = 77,78\%$, $f_7 c_7 = 0,1111 \times 21\ 190 = 2\ 354$, $v_7 = \frac{\sum_{i \leq 7} f_i c_i}{\sum_i f_i c_i} = v_6 + \frac{f_7 c_7}{\sum_i f_i c_i} = 0,4791 + \frac{2\ 354}{16\ 222} \simeq 0,6242$ et $f_7(v_7 + v_6) = 0,1111 \times (0,4791 + 0,6242) \simeq 0,1226$.

2. Comment peut-on interpréter concrètement les nombres $33,33\%$ et $0,1413$ situés dans les colonnes f_i^* et v_i de la troisième ligne du tableau ?
   Réponse :

   Le tiers des personnes ayant les plus faibles revenus totalisent 14,13% du total des revenus français.

3. Tracez, sur le quadrillage ci-dessous, la courbe de Lorentz des revenus en France.



4. Déterminez l'indice de Gini des revenus en France.[1]
   Réponse :
   $G = \sum_i f_i(v_i + v_{i-1}) = 1 - (0,0018 + 0,0090 + \ldots + 0,1992) = 1 - 0,7129 = 0,2871.$

5. Au vu des deux dernières questions, comment peut-on qualifier la concentration des revenus en France ?
   Réponse :
   La courbe de Gini est relativement proche de la diagonale est l'indice de Gini est petit, on peut donc dire que la concentration des revenus en France est relativement faible.

6. L'indice de Gini de la Suède en 2008 est inférieur à 0,25 alors que celui des USA est compris entre 0,45 et 0,50[2]. Que vous suggère, concrètement, la comparaison entre ces deux indices de Gini et celui obtenu à la question 4 ?
   Réponse :
   La concentration des revenus est plus faible en Suède qu'en France et elle est plus faible en France qu'aux USA : cela signifie que les disparités de revenus entre individus sont moins importantes en Suède et plus importantes aux USA, qu'en France.

Exercice 2 Corrélation 1 (dédicacé à Daniel)

Remarque : Les réponses doivent être justifiées. Si l'espace fourni pour répondre n'est pas suffisant, vous pouvez utiliser l'espace situé page 4.

On s'intéresse à la distribution conjointe de deux variables qui correspondent à deux questions posées à un échantillon de la population française lors d'une enquête sur le sentiment d'insécurité :

[1] Notez que, comme ces données ne sont fournies que pour les 9 premiers déciles, l'indice de Gini de la France est légèrement sous-évalué par rapport à la réalité.

[2] Source : Wikipédia, « World Map Gini coefficient » (`http://commons.wikimedia.org/wiki/File:World_Map_Gini_coefficient.svg`)

- Avez-vous parfois peur à votre domicile ? (réponses possibles : oui ou non)
- Avez-vous été
  □ Victime d'une agression ?
  □ Pas victime d'une agression mais témoins de destructions dans votre quartier ?
  □ Pas victime d'une agression ni de destructions mais témoins d'une agression ?
  □ Ni victime d'une agression, ni témoins de destructions ou d'une agression ?

On notera X la variable correspondant aux réponses à la première question et Y la variable correspondant aux réponses à la seconde question. La table de contingence sur l'échantillon de Français interrogés est la suivante[3] :

X \ Y	Victime	Destructions dans quartier	Témoins d'agression	Rien	*Total*
Apeurés	816	1468	163	1 632	*4 079*
Non apeurées	5 083	12 015	1 848	27 727	*46 673*
Total	*5 899*	*13 483*	*2 011*	*29 359*	*50 752*

1. Entre ces deux distributions :
   - Distribution de Y conditionnellement à X
   - Distribution de X conditionnellement à Y

   laquelle permet de déterminer si les personnes apeurées ont plus souvent été victimes d'agressions que les personnes non apeurées ?
   Réponse :
   C'est la distribution de Y conditionnellement à X qui permet de répondre à cette question.

   Déterminer cette distribution dans le tableau ci-dessous :

X \ Y	Victime	Destruction dans quartier	Témoins d'agression	Rien	*Total*
Apeurés	*20,00%*$^{(*)}$	*35,99%*	*4,00%*	*40,01%*	*1*
Non apeurées	*10,89%*	*25,74%*	*3,96%*	*59,41%*	*1*
Ensemble	*11,62%*$^{(**)}$	*26,57%*	*3,96%*	*57,85%*	*1*

   Réponse :
   $^{(*)}$ est calculé par $\frac{816}{4\ 079} \simeq 0,2000$ et $^{(**)}$ est calculé par $\frac{5\ 899}{50\ 752} \simeq 0,1162$.

2. D'après les résultats de la question précédente, les personnes apeurées ont-elles été plus souvent victimes d'agressions que les personnes non apeurées ? (Justifier en reprenant les valeurs adéquates du tableau précédent).
   Réponse :
   20% des personnes apeurées ont été victimes d'une agression contre 10,89% des personnes non apeurées. Les personnes apeurées ont donc été plus souvent victime d'une agression que les personnes non apeurées.

3. Le tableau ci-dessous contient les effectifs théoriques d'indépendance : complétez-le en justifiant vos résultats. Coloriez ensuite en vert les paires de modalités sous-représentées et en rouge les paires de modalités sous-représentées. Que peut-on déduire du coloriage ?

Tableau des effectifs théoriques d'indépendance

X \ Y	Victime	Destruction dans quartier	Témoins d'agression	Rien	Total
Apeurés	474,11	1 083,65	*161,63*$^{(*)}$	*2 359,62*	*4 079*
Non apeurées	5 424,89	12 399,35	*1 849,37*	*26 999,38*	*46 673*
Total	*5 899*	*13 483*	*2 011*	*29 359*	*50 752*

[3] Les données sont reconstituées de la publication de l'INSEE : Thomas Le Jeannic (2006) Insécurité : Perceptions et réalités. Dans *Données sociales, la société française, Édition 2006*, p637-647. Elles concernent la période 2000-2004.

Réponse :

Les effectifs marginaux sont repris du tableau initial. $^{(*)}$ est obtenue par : $\frac{2\ 011 \times 4\ 079}{50\ 752}$.
Le coloriage nous indique que, pour les personnes apeurées, les modalités relatives aux agressions et dégradations, quelles qu'elles soient, sont sur-représentées. À l'inverse, pour les personnes non apeurées, les modalités relatives aux agressions et dégradations sont sous-représentées.

4. Le tableau ci-dessous contient les contributions au χ^2 : complétez-le en justifiant vos résultats. Quelle est la paire de modalités qui contribue le plus au χ^2. Interprétez ce phénomène.

Tableau des contributions au χ^2

Y / X	Victime	Destruction dans quartier	Témoins d'agression	Rien
Apeurés	246,54	136,33	*0,01*$^{(*)}$	*224,37*
Non apeurées	21,55	11,91	0,00	19,61

Réponse :

$^{(*)}$ est calculée par $\frac{(161,63-163)^2}{161,63}$.
La paire de modalités qui contribue le plus au χ^2 est les victimes d'agression qui sont apeurées : cela signifie que celles-ci sont beaucoup plus présentes dans l'échantillon qu'elles ne le devraient si les variables X et Y étaient indépendantes.

5. Calculez le χ^2 puis le C de Cramer. Interprétez cette dernière valeur.
Réponse :

$\chi^2 = 246,54 + 136,33 + 0,01 + 224,37 + 21,55 + 11,91 + 19,61 = 660,32$ et $C = \sqrt{\frac{660,32}{50752 \times (2-1)}} \simeq 0,114$. La corrélation entre la peur et le fait d'avoir été témoins ou victime d'une agression ou de dégradations est faible malgré les différences observées dans les questions précédentes.

Exercice 3 Corrélation 2

Remarque : Les réponses doivent être justifiées. Si l'espace fourni pour répondre n'est pas suffisant, vous pouvez utiliser l'espace situé page 6.

Des expériences ont été effectuées concernant la distance de freinage sur route mouillée : en fonction de la vitesse de la voiture au moment du freinage, on a mesuré la distance nécessaire pour l'arrêt complet du véhicule. Les expériences, répétées 100 fois, ont été reportées dans le tableau ci-dessous. Pour alléger la taille du tableau, seule une partie du tableau et divers calculs associés sont reproduits.[4]

Exp. numéro	Vitesse	Distance	Vitesse^2	Distance^2	Vitesse × Distance	Vitesse^2 × Distance	Vitesse^4
1	102	137	10 404	18 769	13 974	1 425 348	108 243 216
2	42	20	1 764	400	840	35 280	3 111 696
3	73	71	5 329	5 041	5 183	378 359	28 398 241
4	8	3	64	9	24	192	4 096
5	62	53	3 844	2 809	3 286	203 732	14 776 336
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
98	120	185	14 400	34 225	22 200	2 664 000	207 360 000
99	134	227	17 956	51 529	30 418	4 076 012	322 417 936
100	20	9	400	81	180	3 600	160 000
Total	7 759	10 461	801 509	1 850 311	1 185 352	$\simeq 141\ 928 \times 10^3$	$\simeq 10\ 900 \times 10^6$

1. Quelle est la population étudiée ? Quelle est sa taille ?
Réponse :

La population étudiée est l'ensemble des expériences menées. Sa taille est $N = 100$.

2. Quelles sont les variables étudiées ? Quels sont leurs types ?
Réponse :

Les variables étudiées sont la vitesse et la distance. Elles sont toutes les deux quantitatives continues.

[4] Les vitesses sont données en km/h et les distances en mètres.

3. Laquelle de ces deux régressions :

   ☐ régression de la vitesse en la distance de freinage

   ☒ régression de la distance de freinage en la vitesse

   vous paraît être d'intérêt dans ce problème ? (Cochez la case correspondante)

4. Pour la régression choisie à la question précédente, calculez le coefficient de régression linéaire. Commentez sa valeur.
   Réponse :

Statistique pour la variable X_1, la vitesse :

$$\overline{X}_1 = \frac{7\ 759}{100} = 77,59 \text{ km/h} \qquad \text{Var}(X_1) = \frac{801\ 509}{100} - 77,59^2 = 1\ 994,882$$

Statistique pour la variable Y, la distance de freinage :

$$\overline{Y} = \frac{10\ 461}{100} = 104,61 \text{ m} \qquad \text{Var}(Y) = \frac{1\ 850\ 311}{100} - 104,61^2 = 7\ 559,858$$

Covariance

$$\text{Cov}(X_1, Y) = \frac{1\ 185\ 352}{100} - 77,59 \times 104,61 = 3\ 736,83$$

Coefficient de corrélation linéaire

$$r(X_1, Y) = \frac{3\ 736,83}{\sqrt{1\ 994,882}\sqrt{7\ 559,858}} \simeq 0,962.$$

La corrélation linéaire entre la distance de freinage et la vitesse est forte : la distance de freinage peut donc être estimée à partir de la vitesse au moment du freinage.

5. On s'intéresse, à présent, à la régression de la distance de freinage en la vitesse au carré (Vitesse^2). Pour cette régression, calculez le coefficient de régression linéaire. Commentez sa valeur en la comparant à celle trouvée à la question précédente.
   Réponse :

Statistique pour la variable X_2, la vitesse au carré :

$$\overline{X}_2 = \frac{801\ 509}{100} = 8\ 015,09 \qquad \text{Var}(X_2) = \frac{10\ 900 \times 10^6}{100} - 8\ 015,09^2 \simeq 44,758 \times 10^6$$

Covariance

$$\text{Cov}(X_2, Y) = \frac{141\ 928 \times 10^3}{100} - 8\ 015,09 \times 104,61 \times 580,821 \times 10^3$$

Coefficient de corrélation linéaire

$$r(X_2, Y) = \frac{580,821 \times 10^3}{\sqrt{44,758 \times 10^6}\sqrt{7\ 559,858}} \simeq 0,999.$$

La corrélation linéaire entre la distance de freinage et la vitesse au carré est très forte : il est donc préférable d'estimer la distance de freinage à partir de la vitesse au carré.

6. Déterminez l'équation de la droite de régression de la distance de freinage en la vitesse au carré. En déduire la relation obtenue entre la vitesse et la distance de freinage.
   Réponse :

L'équation de la droite de régression est de la forme $y = ax_2 + b$ avec :

$$a = \frac{580,821 \times 10^3}{44,758 \times 10^6} \simeq 0,0130 \qquad \text{et} \qquad b = 104,61 - 0,0130 \times 8\ 015,09 \simeq 0.414.$$

Ainsi,

$$\text{Distance} = 0,0130 \times \text{Vitesse}^2 + 0,414.$$

7. Quelle est la distance de freinage estimée pour une vitesse de 100 km/h ?

Réponse :

La distance de freinage estimée est :

$$\hat{y} = 0,0130 \times 100^2 + 0,414 = 130,414 \text{ m.}$$

8. Sur la figure de la page suivante, représentant le nuage de points des deux variables étudiées, tracez la courbe de la relation trouvée dans la question 6. On fera apparaître quelques points de construction.