TD PS
Statistiques - Etude conjointe de 2 variables statistiques

1. L'enquête effectuée auprès des usagers du Minitel pour suivre les comportements relatifs aux services de Télétel et de l'annuaire électronique comprte près de 70 questions dont de nombreuses à modalités multiples. A partir de ces questions, on a extrait le tableau suivant qui croise les réponses aux questions suivantes :
   - Question 1 : Quelle est la durée moyenne hebdomadaire de consultation du minitel ? Il y a 5 réponses possibles :
     - moins de 5 minutes
     - de 5 à 15 mn
     - de 15 à 30 minutes
     - de 30mn à une heure
     - plus d'une heure.
   - Question 2 : Quel est votre degré de satisfaction du service offert ? Il y a 4 modalités , très satisfait noté Satplus, assez satisfait noté Assat, plutôt pas satisfait noté Passat, Pas satisfait noté Pascont.

	Satplus	Assat	Passat	Pascont	TOTAL
moins 5 mn	110	449	132	13	704
5-15 mn	153	432	37	4	626
15-30 mn	76	146	9	3	234
30-60 mn	42	88	1	0	131
plus 60 mn	31	20	0	0	51
TOTAL	412	1135	179	20	1746

   (a) Les deux tableaux suivants représentent les profils lignes et les profils colonnes. Complétez les valeurs manquantes dans chacun des tableaux.
   Au vu de ces deux tableaux, pensez-vous qu'il y a une relation entre les réponses aux deux questions ? Pourquoi ?

	Satplus	Assat	Passat	Pascont	TOTAL
< 5 mn	15.6	63.8	?	1.8	100.0
5-15 mn	24.4	69.0	5.9	0.6	100.0
15-30 mn	32.5	?	3.8	1.3	100.0
30-60 mn	32.1	67.2	0.8	0.0	100.0
> 60mn	60.8	39.2	0.0	0.0	100.0
TOTAL	23.6	65.0	10.3	1.1	100.0

	Satplus	Assat	Passat	Pascont	TOTAL
< 5 mn	26.7	39.6	?	65.0	40.3
5-15 mn	37.1	38.1	20.7	20.0	35.9
15-30 mn	?	12.9	5.0	15.0	13.4
30-60 mn	10.2	7.8	0.6	0.0	7.5
> 60mn	7.5	1.8	0.0	0.0	2.9
TOTAL	100.0	100.0	100.0	100.0	100.0

(b) Nous avons vu en cours deux indices permettant d'évaluer le degré de liaison entre les modalités de deux variables. Si il n'y avait pas de relation entre les modalités de ces deux questions, c'est-à-dire si les réponses aux questions étaient indépendantes, le tableau original observé devrait être "voisin " du tableau théorique suivant où dans chaque cellule, on a la valeur $\frac{n_{i.}n_{.j}}{n}$ (Les notations sont celles du cours ; n_{ij} sont les valeurs observées dans la cellule (i, j), n=1746 ; les $n_{i.}$ et $n_{.j}$ sont les marginales.) Complétez les valeurs manquantes dans le tableau.

	Satplus	Assat	Passat	Pascont	TOTAL
< 5 mn	166.1	457.6	?	8.1	704.0
5-15 mn	147.7	406.9	64.2	7.2	626.0
15-30 mn	55.2	152.1	24.0	2.7	234.0
30-60 mn	30.9	85.2	13.4	1.5	131.0
> 60mn	12.0	?	5.2	0.6	51.0
TOTAL	412.0	1135.0	179.0	20.0	1746.0

(c) Le tableau suivant contient pour chaque cellule la différence entre la valeur observée n_{ij} et la valeur théorique obtenue sous le modèle d'indépendance $\frac{n_{i.}n_{.j}}{n}$. Complétez les valeurs manquantes dans le tableau.

	Satplus	Assat	Passat	Pascont	TOTAL
< 5 mn	-56.1	-8.6	?	4.9	0.0
5-15 mn	5.3	25.1	-27.2	-3.2	-0.0
15-30 mn	20.8	-6.1	-15.0	0.3	-0.0
30-60 mn	11.1	2.8	-12.4	-1.5	0.0
> 60mn	19.0	?	-5.2	-0.6	0.0
TOTAL	0.0	0.0	0.0	0.0	0.0

(d)

	< 5 mn	5-15 mn	15-30mn	30-60mn	>60mn	TOTAL
Satplus	-0.3	0.0	0.4	0.4	1.6	2.0
Assat	-0.0	0.1	-0.0	0.0	-0.4	-0.4
Passat	0.8	-0.4	-0.6	-0.9	-1.0	-2.1
Pascont	0.6	-0.4	0.1	-1.0	-1.0	-1.7
TOTAL	1.1	-0.8	-0.2	-1.5	-0.8	-2.2

Attention : le tableau ci-dessus est transposé par rapport au précédent, c'est-à-dire que la durée de communication définit les colonnes au lieu des lignes. Il donne les valeurs de

$$\frac{n_{ij}}{\frac{n_{i.}n_{.j}}{n}} - 1$$

Déterminer le tableau des indices d'attraction-répulsion d_{ij} à partir du tableau ci-dessus. Vous mettrez en ligne et en colonne les variables de votre choix.

(e) Quelles sont les cellules (et donc les modalités) responsables du rejet de l'indépendance, c'est-à-dire de la liaison entre les réponses aux deux questions ? Conservez les deux ou 3 valeurs très supérieures à 1 et les quelques valeurs très inférieures à 1.

(f) Si on calcule la seconde statistique,la statistique du χ^2, permettant de tester l'indépendance entre les réponses aux deux questions, on trouve la valeur 161.8. Cette statistique est calculée en faisant une somme de termes sur toutes les cellules du tableau. Dans le tableau suivant figure la contribution de chaque cellule à cette statistique du χ^2. Nous n'avons pas les éléments pour faire le test statistique mais nous pouvons examiner ce tableau de façon descriptive pour déterminer quelles sont les cases qui contribuent le plus au χ^2.

Quelles sont les cellules et donc les modalités des deux variables qui contribuent le plus ? Vérifiez que ces résultats sont comparables à ceux qui ont été obtenus avec les d_{ij} (Ne pas oublier que le tableau précédent est le transposé de celui qui suit).

	Satplus	Assat	Passat	Pascont	TOTAL
moins 5 mn	19.0	0.2	49.6	3.0	71.7
5-15 mn	0.2	1.5	11.5	1.4	14.6
15-30 mn	7.8	0.2	9.4	0.0	17.5
30-60 mn	4.0	0.1	11.5	1.5	17.1
plus 60mn	29.9	5.2	5.2	0.6	40.9
TOTAL	60.8	7.3	87.2	6.5	161.8

2. La moyenne et l'écart-type des notes sur 100 obtenues par des étudiants de deux groupes distincts dans un cours de statistique sont données ci-dessous :

Groupe	**moyenne**	**écart-type**	**taille du groupe**
A	75	10.3	30
B	60	7.2	36

(a) Calculez la moyenne pour les deux groupes réunis.

(b) Calculez l'écart-type pour les deux groupes réunis.

(c) Comparez les coefficients de variation dans chaque groupe.

(d) Calculez le rapport de corrélation . Pouvez-vous facilement conclure à une différence entre les deux groupes ?

(e) Comparez un étudiant ayant obtenu 80 dans le groupe A à un étudiant ayant obtenu 76 dans le groupe B en calculant pour chacun d'entre eux leur note standardisée (centrée réduite) dans leur groupe respectif.

3. Le tableau suivant vous donne les chiffres d'affaires de deux magasins A et B en millions de francs pendant 6 années successives et leurs logarithmes en base 10.

Année	CA de A		CA de B	
1	1121	3.0496	292	2.4654
2	1233	3.0910	321	2.5065
3	1356	3.1323	353	2.5478
4	1492	3.1738	387	2.5877
5	1641	3.2151	428	2.6314
6	1805	3.2565	470	2.6721

(a) Représentez graphiquement les CA en fonction du temps et les log des CA en fonction du temps.

(b) Que concluez-vous au sujet de l'augmentation des CA ? (il peut y avoir des commentaires différents suivant les magasins)

(c) Quel est le magasin qui a le plus fort taux d'accroissement ?

4. On dispose de n observations $(X_i, Y_i), i = 1, n$ sur deux variables statistiques X et Y.

(a) Calculez la moyenne empirique de $X + Y$.

(b) Calculez $cov(X,X)$,$var(X+Y)$,$cov(aX+b,cY+d)$,r_{XX},$r_{aX+b,cY+d}$ où $r_{X\ Y}$ est le coefficient de corrélation entre X et Y et a,b,c,d sont des réels non nuls.

(c) On cherche à déterminer la droite des moindres carrés $y=ax+b$ à partir des n observations (X_i,Y_i) précédentes.Déterminer les coefficients a et b qui minimisent $\sum_{i=1,n}(Y_i-aX_i-b)^2$.Même question pour la droite $x=cy+d$.

(d) On appelle somme des carrés résiduelles SCR la quantité $SCR=\sum_{i=1,n}(Y_i-\hat{Y}_i)^2=\sum_{i=1,n}r_i^2$ où $\hat{Y}$ est la valeur estimée de Y et r_i est le résidu, c'est-à-dire la différence entre la valeur observée et la valeur estimée.
Montrez que la moyenne des r_i est nulle.

(e) On appelle SCT somme des carrés totale la quantité $SCT=\sum_{i=1,n}(Y_i-\bar{Y})^2$.Montrez que $SCR = SCT\times(1-r_{XY}^2)$

(f) On appelle somme des carrés expliquée la quantité $SCE=\sum(\hat{Y}_i-\bar{Y})^2$. Montrez que $SCT = SCE+SCR$. Calculez la variance des r_i.

Remaque : Les résidus normés sont les résidus divisés par leur écart-type. On considère qu'un résidu normé est grand quand en valeur absolue il est supérieur à 2.