FEUILLE DE TRAVAUX PRATIQUES N° 2

séries bi-variées

Enseignants : H. El-Otmany & V. Darrigrand

A.U. : 2014-2015

N.B : Les étudiants ont le choix du système d'exploitation (linux ou Microsoft Windows) et du tableur (Microsoft Excel, OpenOffice Calc ou Gnumeric), dans la mesure des moyens disponibles dans la salle.

Exercice n°1 Le tableau ci-dessous donne la répartition de 2000 individus selon l'âge et le principal sport pratiqué.

age \ sport	équitation	football	golf	natation	tennis
moins de 20 ans	50	140	20	140	150
entre 20 et 30 ans	80	150	50	170	250
entre 30 et 40 ans	80	50	70	100	200
plus de 40 ans	30	20	60	90	100

1. Déterminer les distributions marginales et les différentes distributions conditionnelles. On veillera à ce que les calculs soient effectués le plus rapidement et le plus simplement possible.
2. Calculer le coefficient Φ^2 de Pearson, le coefficient T de Tschuprow et le coefficient C de Cramer.

Exercice n°2 Le tableau ci-dessous donne le produit national brut et la consommation privée pour les années 1960 à 1969 en France (exprimés en francs constants de 1963).

Années	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
PNB (x)	346	365	392	412	439	460	486	508	533	575
Conso. privée (y)	209	222	238	255	269	281	294	309	326	350

1. Préciser la population étudiée et les variables observées sur cette population.
2. Représenter le nuage de points.
3. Déterminer la droite de régression de y sur x (par la méthode des moindres carrés) en utilisant les fonctions `pente` et `ordonnee.origine`. Vérifier les calculs fournis par le tableur. Représenter la droite de régression sur le graphique précédent.
4. Donner le coefficient de corrélation linéaire et le coefficient de détermination à l'aide des fonctions `coefficient.correlation` et `coefficient.determination`. Vérifier les calculs fournis par le tableur.
5. Déterminer la droite de régression de x sur y et la représenter sur le graphique précédent.
6. Déterminer la droite obtenue par la méthode des deux points de Mayer et la représenter graphiquement.
7. Déterminer la droite obtenue par la méthode des moindres distances (régression orthogonale) et la représenter graphiquement.

Exercice n°3 Les deux tableaux ci-dessous indiquent la population respectivement en Afrique et au Canada pour certaines années entre 1950 et 1995 (exprimée en millions d'habitants).

Années	1950	1960	1970	1980	1990	1995
Population en Afrique	222	277	362	470	640	712

Annéées	1950	1960	1970	1980	1990	1995
Population au Canada	13.1	17.7	21	23.8	26.2	27.6

1. Préciser la population étudiée et les variables observées sur cette population.
2. Représenter les deux nuages de points sur deux graphiques séparés.
3. Déterminer la droite de régression pour les deux séries bivariées et leur coefficient de détermination. Représenter les deux droites sur les graphiques précédentes.
4. Proposer une ou plusieurs transformations qui aboutissent à un meilleur ajustement des données (ces transformations ne seront pas forcément les mêmes pour les deux séries bivariées).

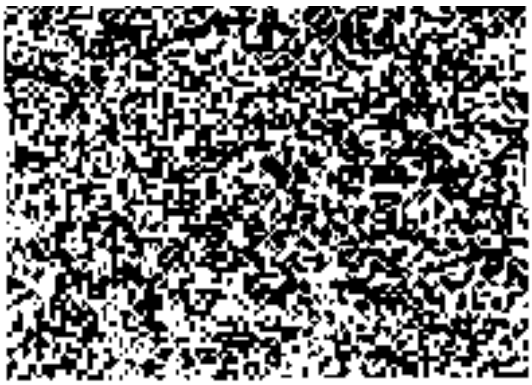
Exercice n°4 Sur un échantillon de douze adolescentes diabétiques, on a relevé, pour chaque adolescente, son score sur une échelle de dépression (CES-D) et son score sur une échelle d'alexithymie [1] (TAS). Les résultats sont les suivants :

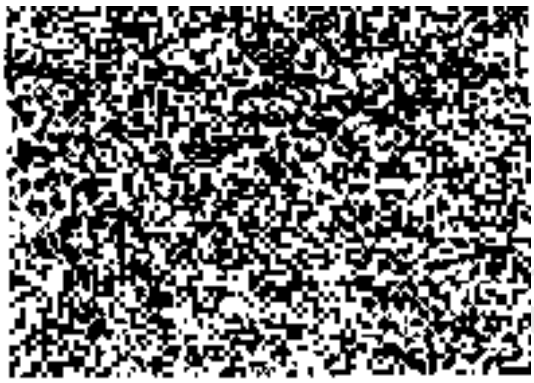
Score CES-D	25	9	21	33	32	21	5	12	19	5	21	13
Score TAS	74	80	72	30	86	72	20	33	82	20	72	55

Un psychologue se pose la question de l'existence d'une éventuelle relation entre le niveau d'alexithymie et le niveau de dépression. A l'aide des différentes notions étudiées dans ce cours (en particulier, à l'aide du coefficient de Bravais-Pearson et/ou du coefficient de Spearman), pouvez-vous fournir une réponse à ce psychologue ?

Exercice n°5 Le fichier `stereo.dat` contient les résultats d'une étude sur la perception visuelle en utilisant des stéréogrammes aléatoires comme les deux images ci-dessous [2]. Ces deux images semblent être composées entièrement de points choisis aléatoirement, mais cependant elles sont construites de sorte qu'une image en trois dimensions (un diamant, en l'occurence) puisse être obtenue à l'aide d'un outil de visualisation stéréo (faisant superposer les deux images). Une autre manière consiste à fixer un point entre les deux images et de défocaliser les yeux. Cette technique demande quelque peu des efforts et l'entrainement [3].

1. Difficulté à verbaliser ses émotions.
2. D'après Cleveland, W. S. (1993). Visualizing Data. Original source : Frisby, J. P. and Clatworthy, J.L., "Learning to see complex random-dot steregrams," Perception, 4, (1975), pp. 173-178
3. Pour plus de détails, on pourra consulter la page sur les stéréogrammes dans wikipédia `fr.wikipedia.org/wiki/Stéréogramme`





L'expérience qui a été menée avait pour but de déterminer si la connaissance de la forme de l'image cachée derrière un stéréogramme influence ou pas le temps nécessaire pour la fusion des images (et donc l'apparition de l'objet). Les 81 individus de l'échantillon ont été divisé en deux sous-échantillons. Le premier groupe (noté NV dans le fichier) n'a reçu aucune information a priori, alors que le second groupe (noté NN dans le fichier) a reçu des informations sur la figure à voir (indications verbales, dessin de l'objet, ...). Pour tous les individus, on a mesuré le temps nécessaire pour la fusion des images et on a noté leur groupe d'appartenance.

1. Afin de comparer les deux sous-échantillons, représenter-les à l'aide de boîtes à moustaches.
2. Effetcuer une transformation logarithmique des observations et représenter la nouvelle version à l'aide de boîtes à moustaches.
3. Commenter.

Exercice n°6 On a vu précédemment comment effectuer un regroupement par modalités lorsqu'on travaille sur une série univariée (cf. exercice 9). Dans cet exercice, on va faire de mettre en considérant deux variables, l'une quantitative et l'autre qualitative. On va produire un tableau croisé.

— Ouvrir le fichier `Enquete.xls` : ce fichier contient les données relevées lors d'une enquête. Les variables relevées sont l'âge de la mère (AGE en années), le poids de la mère (LWT en livres ; 1 livre = 0,483 Kg), si elle a fumé pendant sa grossesse (SMOKE codée 0 si non, 1 si oui), le nombre de ses antécédents de prématurité (PTL), si elle a eu de l'hypertension (HT codée 0 si non, 1 si oui), le nombre de visites à un médecin spécialisé au cours du premier trimestre de sa grossesse (FVT), le poids de son bébé à la naissance (BWT en grammes).

— Sélectionner la plage de données (y compris le titre des colonnes).

— Aller dans le menu `Données`, puis dans le sous-menu `Pilote de données`, et cliquer sur `Démarrer`. Après avoir cliqué sur `Ok` à la première fenêtre, une seconde fenêtre s'ouvre. Dans celle-ci, faire glisser le bouton "AGE" dans la partie `Ligne champs`, le bouton "SMOKE" dans la partie `Champs de colonne` et enfin le bouton "ID" (identifiant) dans la partie `Champs de données`. Si dans cette partie s'affiche `Somme` avant le titre de la colonne, il faut mettre `Nombre` en choisissant dans les options (bouton du haut). Regarder le résultat obtenu en bas du tableau (il est possible de faire afficher le résultat ailleurs en allant dans les options - bouton du bas).

— On peut également effectuer un regroupement en classes pour la variable "AGE" qui peut être traitée comme une variable quantitative continue. Pour cela, il faut cliquer sur une des modalités de cette variable. Ensuite, il faut aller dans le menu `Données`, puis dans le sous-menu `Plan`, et cliquer sur `Grouper`. On indique dans la case `Grouper par` l'amplitude des classes (même amplitude pour toutes les classes), par exemple 10. On peut aussi préciser la borne inférieure de la première classe et/ou la borne supérieure de la dernière classe. Observer le nouveau tableau (à la place de l'ancien).

— Effectuer d'autres tableaux croisés et regroupement en classes à l'aide des autres variables.

Annexe : bugs dans les tableurs ?

Microsoft Excel (et fort probablement d'autres tableurs) contient des bugs qui ont été signalés dans des revues spécialisées. Certains d'entre eux ont été corrigés. Les deux exercices suivants portent sur des erreurs qui ont été signalées. Concernent-elles le tableur que vous utilisez ?

Exercice n°7 Pour différentes valeurs de a, calculer avec votre tableur la quantité suivante $-a^2/2$. Quel ordre de priorité des différents opérateurs impliqués dans cette formule pour votre tableur ? Cela vous semble-t-il correct ? Si non, comment corriger cela ?

Exercice n°8 La fonction `var.p` d'Excel contenait un bug très bête. Cette erreur ayant été repérée, elle a été corrigée dans les versions récentes d'Excel. Votre version d'Excel est-elle assez récente ? Votre tableur a-t-il reproduit cette erreur ?

1. Choisir une cellule et lui affecter la valeur 1. Nommer cette cellule `c`.
2. Dans une plage de trois cellules, saisir les fonctions suivantes : `=c`, `=c+1` et `=c+2`.
3. Que vaut la moyenne et la variance de cette série statistique faite de trois valeurs (sans utiliser le tableur) ?
4. Les résultats ci-dessous ont été obtenus avec une version peu récente d'Excel (depuis le bug a surement été corrigé) :

c	variance
30E6	0.667
40E6	0.889
160E6+1	-3.555
1E10	14563.555
1E10+1	-14563.555

   Commenter ces résultats. Pour chacune des valeurs de c, calculer la variance de cette série statistique, en utilisant la définition puis en utilisant la formule de Koenig-Huyghens.
5. Utiliser la fonction `var.p` pour les mêmes différentes valeurs de c. Commenter et comparer avec les résultats obtenus à la question précédente.