Gilles HUNAULT

2005

Analyses Statistiques et Informatique

(Travaux Dirigés)

Université d'Angers

T.D. 1 : Probabilités élémentaires, comptages

1. Toute union peut être rendue disjointe.
2. Combien y a-t-il de diagonales dans un polygone convexe de n cotés ?
3. Ecriture formelle du calcul de $p(\text{"}Pair\text{"})$ si on lance un dé.
4. Calcul de probabilités par décomposition.
5. La notion d'indépendance est relative à une probabilité donnée.
6. Obtenir un as aux cartes.
7. Chances théoriques de gagner au *Loto*.
8. Comparaison des chances de gain au tiercé et au quarté.

1. Toute union peut être rendue disjointe

Enoncé

Montrer que toute union, disons $U = A \cup B$ peut s'écrire de façon disjointe $U = C \sqcup D$ pour C et D bien choisis.

Solution

Paraphrasant "*les éléments dans A ou B sont soit dans A soit dans B sans être aussi dans A*" nous pouvons écrire $A \cup B = A \sqcup (B \backslash A)$. La décomposition la plus fine et symétrique serait $(A \cap B) \sqcup (A \backslash A \cap B) \sqcup (B \backslash B \cap A)$.

2. Diagonales dans un n-gone régulier convexe

Enoncé

Combien y a-t-il de diagonales dans un polygone convexe régulier de n cotés ? On commencera par dessiner ces polygones pour $n = 1, 2, 3, 4, 5$ et on comptera à la main le nombre de points, de cotés, de diagonales.

Solution

Il ne faut pas se laisser impressionner par l'énoncé ! Un 3-gone convexe régulier ou polygone convexe régulier à 3 cotés n'est jamais qu'un triangle équilatéral. Pour n=4, c'est un carré, pour n=5 un joli pentagone etc.

Nommons a_n le nombre de cotés, b_n le nombre de points, c_n le nombre de droites et d_n le nombre de diagonales.

On peut alors remplir le tableau suivant après avoir fait les dessins à la main en numérotant soigneusement les points, les droites, les cotés et les diagonales :

n	0	1	2	3	4	5
a_n	?	?	2	3	4	5
b_n	0	1	2	3	4	5
c_n	?	0	1	3	6	10
d_n	?	?	0	0	2	5

Il y a bien sur $a_n = n$ cotés, $b_n = n$ points, $c_n = C_n^2 = n(n-1)/2$ droites (une droite correspond à un sous-ensemble à 2 éléments pris dans un ensemble à n éléments). Une diagonale joint deux points non consécutifs c'est donc un droite qui n'est pas un coté. Il y en a $d_n = c_n - a_n = n(n-1)/2 - n$ c'est à dire $n(n-3)/2$.

3. Ecriture formelle de $p(\text{"}Pair\text{"})$ pour un dé

Enoncé

Tout le monde sait qu'il y autant de chances d'avoir un nombre pair que d'avoir un nombre impair avec un dé. Ecrire la démonstration mathématique rigoureuse correspondante. Détailler ensuite ce que peut être un dé "non pipé" à 2, 3, 4,... 7, 12, 32, 365 faces...

Solution

Nommons $I = \{1, 2, 3, 4, 5, 6\}$ l'ensemble des évènements élémentaires ; on identifie le singleton $\{i\}$ à l'évènement "lorsque le dé est lancé, le chiffre inscrit sur la face supérieure est i". Nommons P l'évènement "on obtient un chiffre pair". Alors $P = \{2, 4, 6\}$. Pour appliquer la formule d'additivité, on décompose P en évènements élémentaires, soit :

$$P = \{\ 2\ \} \sqcup \{\ 4\ \} \sqcup \{\ 6\ \}$$

La probabilité de P est alors

$$p(P) = p(\{\ 2\ \}) + p(\{\ 4\ \}) + p(\{\ 6\ \})$$

Avec des hypothéses "raisonnables", "minimalistes" et sans doute rigoureusement fausses, on peut prendre la même probabilité pour chaque face et donc $p(P) = 1/2$. On s'autorisera pour la suite à noter "2" plutot que $\{\ 2\ \}$ et on utilisera les techniques de comptage lorsqu'il y a équiprobabilité pour un dé.

Un dé à 2 faces est une pièce de monnaie ou une feuille de papier (on choisit arbitrairement ce qui est "1" et ce qui est "2"). Un dé à 3 faces n'est certainement pas un objet réel mais intellectuellement, rien n'empêche de l'imaginer. Il n'est pas sûr qu'un dé à 6 faces pour lequel on convient que "4" est comme "1", "5" est comme "2", "6" est comme 3 soit assimiliable à un dé à 3 faces. Pour un dé à 4 faces, on peut penser à un pyramide à base équilatérale.

7 faces, c'est comme 7 jours dans la semaine ; pour 32 faces on peut utiliser un jeu de cartes et enfin pour 365 jours on peut utiliser une année, un calendrier.

Une façon simple de simuler un dé à n faces, c'est de mettre n boules de même couleur numérotées de 1 à n dans un sac. Jeter le dé consiste alors à tirer une boule. A défaut de sac et de boules, on peut prendre des papiers sur lesquels on écrit les nombres de 1 à n.

4. Calcul de probabilité par décomposition

Enoncé

Calculer $p(A \cap \overline{B}) \cup (B \cap \overline{A})$ en fonction de $\alpha = p(A \cap B)$. Discuter sur α si on donne $p(A) = 0,2$ et $p(B) = 0,5$.

Solution

L'expression $p(A \cap \overline{B}) \cup (B \cap \overline{A})$ n'a aucun sens, donc il n'y a aucun calcul à faire ! Plus raisonnablement, essayons de calculer $p(\ (A \cap \overline{B}) \cup (B \cap \overline{A})\ )$.

A partir du calcul classique
$$\begin{aligned} A &= A \cap E \\ &= A \cap (B \sqcup \overline{B}) \\ &= (A \cap B) \sqcup (A \cap \overline{B}) \end{aligned}$$

on déduit par la formule additivité $p(A \cap \overline{B}) = p(A) - p(A \cap B)$. De façon symétrique $p(B \cap \overline{A}) = p(B) - p(B \cap A)$ ce qui permet de trouver, en additionnant : $p(\ (A \cap \overline{B}) \cup (B \cap \overline{A})\ ) = p(A \cap \overline{B}) + p(B \cap \overline{A}) = p(A) + p(B) - 2p(A \cap B)$ car $A \cap \overline{B}$ et $B \cap \overline{A}$ sont disjoints.

Finalement, on peut donc écrire

$$p_\alpha = p(\ (A \cap \overline{B}) \cup (B \cap \overline{A})\ ) = p(A) + p(B) - 2\alpha$$

Toute probabilité est comprise entre 0 et 1, en particulier p_α ce qui impose donc des conditions sur α. Résolvant chacune des inéquations de $0 \leq p_\alpha \leq 1$ on trouve : $-0,15 \leq \alpha \leq 0,35$. Mais α est aussi une probabilité, il faut donc tenir compte de $0 \leq \alpha \leq 1$.

Enfin, puisqu'une probabilité est une fonction croissante, et comme $A \cap B$ est à la fois inclus dans A et dans B, il faut aussi utiliser l'inéquation $\alpha \leq p(A)$.

Toutes inégalités confondues, la réponse est donc $0 \leq \alpha \leq 0.2$

5. Indépendance et probabilité

Enoncé

On lance un dé "normal" qu'on suppose équilibré. On nomme p_1 la probabilité induite par cette hypothèse. On désigne par A le résultat "pair" et par B le résultat $\{5, 6\}$. A et B sont-ils incompatibles ? A et B sont-ils indépendants pour p_1 ?

On utilise désormais un dé non équilibré. Soit p_2 la probabilité associé à ce dé, telle que $p_2(\text{"6"}) = 1/12$, $p_2(\text{"2"}) = p_2(\text{"4"}) = p_2(\text{"5"}) = 1/6$. On reprend les mêmes résultats A et B. A et B sont-ils indépendants pour p_2 ?

Solution

Puisque $A \cap B = \{6\}$, A et B ne sont pas incompatibles.

$p_1(A \cap B) = 1/6$ et comme $p_1(A) = 1/2$, $p_1(B) = 1/3$, on peut vérifier que $p_1(A \cap B) = p_1(A).p_1(B) = 1/3$ donc A et B sont indépendants pour p_1.

Avec le deuxième dé, $p_2(A \cap B) = 1/12$ et $p_2(A) = 5/12$, $p_2(B) = 3/12$, $p_2(A \cap B) \neq p_2(A).p_2(B) = 5/48$ donc A et B ne sont pas indépendants pour p_2.

Conclusion : dire "A et B sont [non]indépendants" n'a pas de sens si on ne précise pas la probabilité sous-jacente.

6. Obtenir un as aux cartes

Enoncé

Quelle est la probabilité d'obtenir un as si on tire un carte dans un jeu ?

Solution

Traitons deux cas usuels avant de voir les formules générales. A la belote, $n = 32$, au bridge, $n = 52$. On pourra remarquer qu'il y a 4 "couleurs" (comme 4 saisons) et "comme par hasard" il y a 52 semaines dans une année (365=7x52+1).

Une structuration possible des cartes consiste à les noter H_i^j avec i une "couleur" de 1 à 4 et j une valeur de 7 à 13 (ou de 1 à 13), 13 étant l'as, 12 le roi etc. On peut aussi noter (i, j) une telle carte. Un as correspond donc aux couples $(i, 13)$ et il y en 4 soit la probabilité $4/n$ dans le cas général (respectivement $1/8 \simeq 0.125$ soit 13 % et $1/13 \simeq 0.0769$ soit 8%).

Si on pose la question "quelle est la probabilité d'obtenir un as si on tire deux cartes", il y a ambiguité sur le type de tirage puisqu'il peut s'agir de tirer deux cartes simultanément (c'est à dire **sans** remise) ou il peut s'agir de tirer une carte puis de la remettre et de retirer une carte qui peut être éventuellement la même (tirage **avec** remise).

Il a aussi d'ambiguité sur le terme "un as". Si on a deux as, est-ce qu'on a un as ? C'est pourquoi on doit employer les termes "un as exactemnt" ou "au moins un as". Nous laissons les 4 calculs (avec ou sans remise, exactement ou au moins) pour n cartes au soin du lecteur.

7. Chances théoriques de gagner au *Loto*

Enoncé

Au jeu de loto, sous controle d'un huissier, 6 numéros parmi 49 numéros de 1 à 49 sont tirés. Un joueur de loto est une personne qui achète un billet sur lequel 6 numéros sont inscrits (on ignore ici la notion de "numéro complémentaire"). On dit qu'on a gagné le gros lot si les numéros du billet acheté correspondent aux numéros tirés.

Quel type de boules peut-on utiliser pour structurer les événements associés aux tirages du loto ? Quelle est la probabilité de gagner le gros lot ? Vous n'oublierez pas de détailler la structuration de votre espace probabilisé.

La société "La Française des Jeux" envisage de passer de 49 numéros à 60 numéros avec un tirage de 8 boules plutot que 6. Aura-t-on plus de chances de gagner le gros lot ? Comment les valeurs 6 et 49 ont-elles été choisies historiquement ?

Solution

La probabilité de gagner au loto est celle de tirer le bon 6-uplet parmi tous les 6-uplets possibles (ce sont des 6-uplets car on tire 6 numéros sans ordre et sans répétition). A défaut d'autre hypothèse et ce qu'assure (sans démonstration) la Française des Jeux, chaque 6-uplet est équiprobable. On vient donc dénombrer des sous-ensembles de 6 éléments dans un ensemble de base à 49 éléments. La probabilité demandée est donc $1/C_{49}^6$ ce qui fait $1/13983816$ soit en en gros $1/1,39.10^7$ donc $0,715.10^-7$.

Jouer avec 8 boules et 60 numéros diminuerait les chances de gagner car les coefficients du binome croissent très rapidement.

On pourra vérifier que le rapport des chances est de 1/183 au loto si on passe de 49 numéros et 6 boules à 60 numéros et 8 boules.

Quant à l'origine des nombres 6 et 49, je n'en ai aucune idée !

8. Chances de gain au tiercé et au quarté

Enoncé

Vaut-il mieux tenter de gagner (dans l'ordre, dans le désordre), au tiercé avec n chevaux ou au quarté avec $n + 1$ chevaux ? On pourra utiliser la valeur numérique $n = 17$.

Solution

Pour le tiercé dans le désordre, on peut utiliser la même structuration mais avec des triplets (3-uplets) où les boules correspondent aux chevaux. Là encore, il faut faire abstraction de la réalité (météo, jockeys, etc.) pour attribuer une équirépartition de chances.

La probabilité de gagner dans le désordre n'est pas $1/C_n^3$ soit 6 / 4080 car parmi les 6 triplets possibles il y a le "bon ordre" mais seulement 5 / 4080 (on ote un seul cas, celui du "bon" ordre).

Numériquement on obtient 1/816 soit 0,001225.

Pour gagner dans l'ordre, il faut trouver le bon cheval arrivé en premier (probabilité $1/n$) puis le second ($1/(n-1)$) et le troisième ($1/(n-2)$). On trouve alors comme probabilité de l'évènement $1/(n.(n-1).(n-2))$.

Numériquement cela donne $1/4080 = 0{,}000245$.

Si on joue au quarté, on obtient comme probabilités $1/57120 = 0.000175$ de gagner dans l'ordre et $1/2380 = 0.0004201$ de gagner dans le désordre.

T.D. 2 : Probabilités discrètes

1. Ces valeurs définissent-elles des probabilités ?
2. Comptage des fichiers en $ghFs$.
3. Formule de Bayes "pour les jeunes".
4. Probabilités conditionnelles avec 2 dés.
5. Suivez le sprite !

1. Ces valeurs définissent-elles des probabilités ?

Enoncé

Soit p_i un ensemble de probabilités discrètes. Quelles conditions doivent vérifier les p_i ?

Je pense après réfléchi à un problème que pour i de 1 à n, les valeurs p_i définies par $p_i = p(E_i) = i/(\ n(n+1)(n+3)\ )$ doivent représenter les probabilités que je cherchais. Est-ce crédible ?

Mon voisin me dit que les n nombres $q_i = C_n^i 0.2^i 0.8^{n-i}$ pour i de 1 à n sont la solution à mon problème. Puis-je lui faire confiance ?

Solution

Les p_i ne définissent pas une probabilité car par exemple pour $n = 2$, $p_1 + p_2 = 3/10 < 1$.

Si on connait la formule du binome $(a+b)^n = \sum_{i=0}^{n} C_n^i 0.2^i 0.8^{n-i}$, on peut se rendre compte que les q_i correspondent aux termes de cette formule pour i de 1 à n seulement avec $a = 0.$ et $b = 0.8$; ce sont donc bien des nombres positifs inférieurs à 1 mais leur somme ne fait pas 1. Donc ces nombres ne définissent pas une probabilité.

2. Comptage des fichiers en $ghFs$

Enoncé

Un enseignant à forte tendance pédagogique veut imposer un nouveau système de fichiers nommé *ghFs*. Dans un tel système, un identificateur de fichier se compose d'un nom et d'une extension reliés par un tiret. Un nom de fichier comporte de 1 à 5 caractères dont le premier est une lettre, les caractères suivants sont soit une lettre soit un chiffre. Une extension comporte de 1 à 3 caractères, le premier est une lettre, les suivants une lettre ou un chiffre. De façon à ne pas avoir des noms trop imprononçables, on n'utilise que 20 lettres (mais on ne dit pas lesquelles, sauf la lettre P).

a) combien de noms de fichiers différents peut-on avoir, sachant qu'on ne distingue pas majuscule et minuscule ?
b) un fichier-programme est un fichier dont l'identificateur a une extension qui commence par la lettre P. Quelle est la proportion de fichiers-programmes dans ce système de fichiers ?

Solution

Si L désigne une lettre et K une lettre ou un chiffre, un nom de fichier est représenté par L, LK, LKK, LKKK ou par LKKKK. L a 20 possibilités et K a 30 possibilités (les 20 de L plus 10 chiffres). De façon plus rigoureuse, pour L on veut un singleton dans un ensemble à 20 éléments et il y a donc $C_{20}^1 = 20$ possibilités.

Chaque L et chaque K sont indépendants, donc on a $20 + 20.30 + 20.30^2 + 20.30^3 + 20.30^4$ noms de fichiers. Numériquement, on trouve donc $20.(30^5 - 1)/(30-1) = 16758620$ noms, soit à peu près 16,75 millions de noms différents.

Le nombre d'identificateurs (ce qui n'est pas demandé) est obtenu en ajoutant L, LK ou LKK à un nom de fichiers. Si N est le nombre de noms, le nombre d'identificateurs est donc $I = N.(20 + 20.30 + 20.30^2) = N.18620$. Numériquement $I = 312045504400$ (en gros 312 milliards d'identificateurs).

Un fichier programme a une extension qui commence par la lettre P. Comme il y a 20 lettres différentes, la proportion de fichiers-programmes est donc 1/20 si on suppose que seuls les fichiers programmes commencent par cette lettre.

Par exemple, et cela ne contredit pas l'énoncé, un fichier de paramètres pourrait aussi commencer par P.

3. Formule de Bayes "pour les jeunes"

Enoncé

Une récente enquête auprès de jeunes "ché(e)brans" montre que les DeuxSontTrois sont un groupe "firimique". L'enquête a porté sur 3 groupes de jeunes, notés A, B et J.

Le nombre de personnes interrogées dans chaque groupe et le nombre de voix pour le groupe est fourni dans le tableau suivant

	Nombre de jeunes	% de "pour" dans le groupe
Groupe A	160	60
Groupe B	240	40
Groupe J	400	48

Justifiez vos réponses aux questions suivantes

a) Quelle est la probabilité qu'un jeune au hasard dans A,B ou J soit pour le groupe ?

b) Sachant qu'un jeune est pour le groupe, quelle est la probabilité qu'il soit dans le groupe A ?

Solution

Soit F l'évènement on est pour le groupe. Comme A, B et J forment un système complet d'évènements

$$F = F \cap (A \sqcup B \sqcup J) = (F \cap A) \sqcup (F \cap B) \sqcup (F \cap J)$$

On en déduit donc que

$$\begin{aligned} p(F) &= p(A)p(F|A) + p(B)p(F|B) + p(J)p(F|J) \\ &= (160/800)0{,}6 + (240/800)0{,}4 + (400/800)0{,}48 \\ &= 0{,}2\text{x}0{,}6 + 0{,}3\text{x}0{,}4 + 0{,}5\text{x}0{,}48 \\ &= 0{,}48 \end{aligned}$$

Comme $p(A|F) = p(A)p(F|A)/(\ p(A)p(F|A) + p(B)p(F|B) + p(J)p(F|J)\ )$, on obtient

$$(160/800)x0,6/((160/800)x0,6 + (240/800)x0,4 + (400/800)x0,48)$$

donc $p(A|F)$ vaut 0,2x0,6/(0,2x0,6+0,3x0,4+0,5x0,48) soit finalement 0,25.

4. Probabilités conditionnelles avec 2 dés.

Enoncé

On dispose de deux dés normaux à 6 faces et d'une pièce de monnaire usuelle. L'un des deux, nommé A comporte 4 faces rouges et 2 blanches ; l'autre, nommé B comporte 2 faces rouges et 4 blanches. On invente la règle suivante : *"on lance la pièce. si on obtient pile, on joue toujours avec le dé nommé* A. *si c'est face, on joue toujours avec le dé nommé* B*".*

a) Quelle est la probabilité d'obtenir rouge en un lancer ?
b) Quelle est la probabilité d'obtenir rouge au troisième lancer du dé alors qu'on a déjà obtenu rouge au premier et au deuxième lancer ?
Notant R_i l'évènement "on a obtenu rouge au i-ème coup", R_1 et R_2 sont-ils indépendants ? idem pour $R_1|A$ et $R_2|A$.
c) Quelle est la probabilité d'avoir utilisé le dé A alors que sur n lancers, on a obtenu rouge n fois rouge ?

Solution

a) Obtenir rouge (noté R) se fait soit avec le dé nommé A soit avec le dé nommé B d'où la décomposition $R = (R \cap A) \sqcup (R \cap B)$. On calcule bien sur $p(R \cap A)$ par $p(R \cap A) = p_{|A}(R).p(A)$ et $p(R \cap B)$ par une formule équivalente. On obtient donc

$$\begin{aligned} p(R) &= p(R|A)p(A) + p(R|B)p(B) \\ &= (4/6)(1/2) + (2/6)(1/2) \\ &= 1/2. \end{aligned}$$

b) Pour obtenir rouge au troisème lancer du dé alors qu'on a déjà obtenu rouge au premier et au deuxième lancer, on nomme R_i l'évènement "on a obtenu rouge au i-ème coup".

La probabilité cherchée est alors $p(R_3|R_1 \cap R_2)^*$, soit tout simplement $p(R_3 \cap R_1 \cap R_2)/p(R_1 \cap R_2)$. Maintenant, comme $p(R_1 \cap R_2) = p(R_1 \cap R_2|A).p(A)+p(R_1 \cap R_2|B).p(B)$, $p(R_1 \cap R_2)$ vaut $(2/3)^2.1/2+(1/3)^2.1/2 = 5/18$.
De même, $p(R_3 \cap R_1 \cap R_2) = (2/3)^3.1/2 + (1/3)^3.1/2 = 1/6$ et donc la probabilité cherchée est $(1/6)/(5/18)$ soit $3/5$.
Puisque $p(R_1) = p(R_2) = p(R_i) = 1/2$ et $p(R_1 \cap R_2) = 5/18$. on en déduit que R_1 et R_2 ne sont pas indépendants. Par contre la question $R_1|A$ *et* $R_2|A$ *sont-ils indépendants ?* n'a aucun sens : $R_1|A$ et $R_2|A$ ne sont pas des évènements.

c) Avec les notations précédentes, on cherche à déterminer $p(A|\bigcap_{i=1}^{n} R_i)$, ce qu'on peut calculer par $p(\bigcap_{i=1}^{n} R_i|A).p(A)/p(\bigcap_{i=1}^{n} R_i)$. En généralisant les formules obtenues en a) et b), on obtient simplement

$$\frac{(2/3)^n.1/2}{(2/3)^n.1/2 + (1/3)^n.1/2}$$

soit encore $2^n/(2^n + 1)$ de limite 1 pour n infini, ce qui est "raisonnable".

5. Suivez le sprite !

Enoncé

Un économiseur d'écran se compose d'une image fixe sur laquelle se déplace un *sprite* qui ressemble à un petit bonhomme. Le bonhomme ne peut aller qu'à trois endroits nommés H (haut), G (gauche) et D (droit). Par exemple on pourra supposer que H, G et D sont les sommets d'un triangle équilatéral centré au milieu de l'écran. A chaque instant, le bonhomme (qui est sur un point) va vers un des deux autres points de façon équitable.

On note α_n la probabilité qu'au bout de n instants le bonhomme soit en H, γ_n la probabilité qu'au bout de n instants le bonhomme soit en G et δ_n la probabilité qu'au bout de n instants le bonhomme soit en D.

* Cette notation n'est pas ambigue et il n'est pas nécessaire d'écrire $p(\ R_3|(R_1 \cap R_2)\ )$ car $p(\ (R_3|R_1)\ \cap R_2\ )$ n'a aucun sens.

On suppose que le bonhomme est en H à l'instant 0, ce que l'on traduit par $\alpha_0 = 1$, $\gamma_0 = 0$ et $\delta_0 = 0$.

Donner les valeurs de α_1, γ_1, δ_1 puis de α_2, γ_2, δ_2.

Calculer α_{n+1} en fonction de γ_n et δ_n puis γ_{n+1} en fonction de α_n et δ_n et enfin δ_{n+1} en fonction de α_n et γ_n.

En déduire que $\alpha_n + \gamma_n + \delta_n$ vaut 1.

Donner α_n, γ_n et δ_n en fonction de n seulement puis leur limite pour n infini. Le résultat obtenu était-il prévisible ?

On détaillera bien l'espace des évènements et la décomposition en évènements qui permet de fournir les formules en α_i, γ_i et δ_i.

Solution

Le bonhomme est à un point donné si à l'instant précédent il est sur l'un des deux autres points. Par exemple, notant H_n l'évènement "le bonhomme est au point H à l'instant n", G_n l'évènement "le bonhomme est au point G à l'instant n", D_n l'évènement "le bonhomme est au point D à l'instant n", l'évènement H_1 ne peut provenir que de l'évènement G_0 ou de l'évènement D_0. On peut donc écrire

$$H_1 = (H_1 \cap G_0) \sqcup (H_1 \cap D_0)$$

d'où, par l'axiome d'additivité disjointe :

$$p(H_1) = p(H_1 \cap G_0) + p(H_1 \cap D_0)$$

et en utilisant les probabilités conditionnelles :

$$p(H_1) = p(H_1|G_0).p(G_0) + p(H_1|D_0).p(D_0)$$

Puisque les points G et D sont équiprobables, $p(H_1|G_0) = p(H_1|D_0) = 1/2$ et compte-tenu des valeurs initiales $p(G_0) = \gamma_0 = 0$, $p(D_0) = \delta_0 = 0$, on déduit que $\alpha_1 = p(H_1) = (1/2).0 + (1/2).0 = 0$.

De même, $\gamma_1 = p(G_1) = p(G_1|H_0).p(H_0) + p(G_1|D_0).p(D_0)$ donc on calcule $\gamma_1 = (1/2).1 + (1/2).0 = 1/2$; par un calcul similaire, $\delta_1 = 1/2$.

Suivant le même principe, $\alpha_2 = p(H_2) = p(H_2|G_0).p(G_0) + p(H_2|D_0).p(D_0)$ vaut donc $(1/2).(1/2) + (1/2).(1/2)$ soit $\alpha_2 = 1/2$, et $\gamma_2 = \delta_2 = 1/4$.

A l'ordre $n+1$, on utilise la même décomposition, soit

$$H_{n+1} = (H_{n+1} \cap G_n) \sqcup (H_{n+1} \cap D_n)$$

donc $\alpha_{n+1} = p(H_{n+1}) = p(H_{n+1}|G_n).p(G_n) + p(H_{n+1}|D_n).p(D_n)$ vaut donc $(1/2).\gamma_n + (1/2).\delta_n$ soit $\alpha_n = (\gamma_n + \delta_n)/2$.

De la même façon, $\gamma_n = (\alpha_n + \delta_n)/2$ et $\delta_n = (\alpha_n + \gamma_n)/2$.

Soit S_n la somme $\alpha_n + \gamma_n + \delta_n$. On démontre par récurrence que $S_n = 1$: à l'ordre 0, on a bien $\alpha_0 + \gamma_0 + \delta_0 = 1 + 0 + 0 = 1$ et si S_{n-1} vaut 1 alors S_n vaut $\alpha_n + \gamma_n + \delta_n = (\gamma_n + \delta_n)/2 + (\alpha_n + \delta_n)/2 + (\alpha_n + \gamma_n)/2$ soit, en regroupant les termes, $2.S_{n-1}/2$ ce qui fait bien 1.

Pour calculer α_n en fonction de n seulement, on remplace $\gamma_n + \delta_n$ par $1 - \alpha_n$ puisque $\alpha_n + \gamma_n + \delta_n = 1$; la suite des α_i est alors définie par $\alpha_0 = 1$ et $\alpha_i = (1/2)(1-\alpha_{i-1})$. C'est donc une suite arithmético-géométrique classique. On lui associe la suite des α_i^* définie par décalage additif constant

$$\alpha_i^* = \alpha_i + h$$

de même raison géométrique que la partie principale des α_i :

$$\alpha_i^* = (-1/2)\alpha_{i-1}^*$$

Cette dernière relation équivaut à

$$\alpha_i + h = (-1/2)(\alpha_{i-1} + h)$$

et compte-tenu de la relation entre α_i et α_{i-1}, h vaut (-1/3).

(α_n^*) est une suite géométrique donc $\alpha_n^* = \alpha_0^*(-1/2)^n$ soit $\alpha_n^* = (2/3)(-1/2)^n$ puisque $\alpha_0^* = \alpha_0 - h = 1$ - $(1/3)$ et donc finalement $\alpha_n = \alpha_n^* - h$ soit

$$\alpha_n = \frac{1}{3} + \frac{2}{3}\left(\frac{-1}{2}\right)^n$$

On trouve de la même façon

$$\gamma_n = \frac{1}{3} - \frac{2}{3}\left(\frac{-1}{2}\right)^n \qquad \text{et} \qquad \delta_n = \frac{1}{3} - \frac{2}{3}\left(\frac{-1}{2}\right)^n$$

Puisque $|(-1/2)| < 1$, les trois suites α_n, γ_n et δ_n tendent toutes trois vers la même valeur $1/3$ ce qui était prévisible puisque les trois points sont équiprobables...

T.D. 3 : Variables aléatoires

1. Un calcul combinatoire
2. Question de vocabulaire...
3. Diverses v.a. (somme, produit...) pour 2 dés à 3 faces
4. Centrage et Réduction
5. Loi de *Bernoulli* généralisée
6. Lois et valeurs comme $V < m$ pour $\mathcal{B}(n, p)$
7. Qu'est-ce qu'une moyenne ?

1. Un calcul combinatoire

Enoncé

Montrer que $C_{2n}^n = \sum_{k=0}^{n} \left(C_\alpha^k\right)^2$ pour α bien choisi dépendant de n. Indication : on pourra utiliser l'identité $(1+x)^{n_1+n_2} = (1+x)^{n_1}.(1+x)^{n_2}$

Solution

Considérons le produit $(1+x)^{n_1}.(1+x)^{n_2}$.

Remplaçons le premier terme par le développement $\sum_{k_1=0}^{n_1} C_n^{k_1} x^{k_1}$ et le deuxième terme par le développement $\sum_{k_2=0}^{n_2} C_n^{k_2} x^{k_2}$.

Pour obtenir x^k dans le produit des deux développements, il faut considérer tous les x^{k_1} et les x^{k_2} avec $k_1 + k_2 = k$.

Puisque le coefficient de x^k dans $(1+x)^{n_1+n_2}$ est $C_{n_1+n_2}^k$ l'identité du départ implique que

$$C_{n_1+n_2}^k = \sum_{k_1+k_2=k} C_{n_1}^{k_1} C_{n_2}^{k_2}$$

Appliquons cette formule à $n_1 = n_2 = n = k$.

On en déduit $C_{2n}^n = \sum_{k_1+k_2=n} C_{n_1}^{k_1} C_{n_2}^{k_2}$.

Remplaçant k_1 par k et donc k_2 par $n-k$, puisque $C_n^{n-k} = C_n^k$, on a finalement $C_{2n}^n = \sum_{k=0}^{n} (C_n^k)^2$: il suffit donc de prendre $\alpha = n$.

2. Question de vocabulaire...

Enoncé

Rappeler comment sont nommés et comment sont définis les indicateurs et phénomènes désignés par les symboles $\mathcal{T}_E$, p, X, p_X, m et σ.

Solution E est l'espace des épreuves et $\mathcal{T}_E$ est l'univers ou espace des évènements. p est une **fonction** de $\mathcal{T}_E$ dans $[0,1]$ souvent définie sur une partition E_i de E. X est une fonction sur $\mathcal{T}_E$ à valeurs dans R. $p_X(A)$ est en fait $p(X^{-1}(A))$.

m est un indicateur numérique global de tendance générale ou centrale, nommé valeur moyenne ou plus simplement moyenne. σ est un indicateur numérique global de dispersion nommé écart-type ou aussi écart-standard.

m se calcule par $\sum\ p_i x_i$ et si $m_c = \sum\ p_i x_i^2$ alors $V = m_c - m^2$ et $\sigma = \sqrt{V}$.

3. Diverses v.a. pour 2 dés à 3 faces

Enoncé

On lance 2 dés à 3 faces. Etudiez les variables S, P, $M = S * P$, D qui correspondent respectivement à la somme, au produit, au produit de S par P et à la valeur absolue de la différence des chiffres inscrits sur le dé.

Solution

Avant se lancer directement dans les calculs (la somme S vaut 2 pour 1+1 seulement, elle vaut 3 pour 1+2 et 2+1, elle vaut 4 pour 1+3,2+2, 3+1...) il faut essayer de voir comment on peut structurer l'ensemble de départ. Avec n faces pour le premier dé et n faces pour le second, il y a n^2 couples de résultats possibles, même si les dés ne sont pas humainement discernables. Ainsi $(1,2)$ et $(2,1)$ sont deux évènements élementaires, alors qu'on ne les identifiera que si par exemple les dés sont de couleurs différentes. Le système complet d'évènements est donc l'ensemble des couples (i,j) pour i et j de 1 à n, chaque couple équiprobable ayant une probabilité de 1/36.

Une autre solution aurait consisté à ne retenir que les couples (x, y) discernables, en les ordonnant par exemple par $x \leqslant y$. Mais alors on aurait moins de couples (combien ?) et ils ne seraient plus équiprobables.

Ainsi pour $n = 3$, au lieu des 9 couples

$$(1,1),\ (1,2),\ (1,3),\ (2,1),\ (2,2),\ (2,3),\ (3,1),\ (3,2),\ (3,3)$$

avec chacun comme probabilité $1/9$, on n'aurait plus que les 6 couples

$$(1,1),\ (1,2),\ (1,3),\ (2,2),\ (2,3),\ (3,3)$$

de probabilité respective

$$1/9,\ 2/9,\ 2/9,\ 1/9,\ 2/9,\ 1/9$$

Le plus simple pour construire tous les cas possibles (et surtout pour éviter d'en oublier) serait d'écrire un programme qui gére tous les calculs. Comme l'écriture d'un tel programme prend beacuoup de temps pours des programmeurs débutants, on peut pour $n = 3$ se content d'écrire les produits cartésiens des couples de valeurs.

```
Etude de 2 dés avec chacun 3 faces
... On passe en revue les 9 cas    possibles.

étude de la variable : S = somme des valeurs

  S  |    1    2    3
-----+--------------------------------
  1  |    2    3    4
  2  |    3    4    5
  3  |    4    5    6

  S  |    2    3    4    5    6
-----+---------------------------
 9p  |    1    2    3    2    1    total : 9
```

```
variable S moyenne    :     4.000  (somme des valeurs :  36)
           variance   :     1.333  (somme des carréss : 156)
           écart-type :     1.155
           cdv        :    28.868 % (coefficient de variation)

étude de la variable : P = produits des valeurs

   P  |    1    2    3
 -----+--------------------------------
   1  |    1    2    3
   2  |    2    4    6
   3  |    3    6    9

   P  |    1    2    3    4    6    9
 -----+--------------------------------
  9p  |    1    2    2    1    2    1    total : 9

variable P moyenne    :     4.000 (somme des valeurs :   36)
           variance   :     5.778 (somme des carrés  :  196)
           écart-type :     2.404
           cdv        :    60.093 % (coefficient de variation)

étude de la variable : M = somme*produit des valeurs

   M  |    1    2    3
 -----+--------------------------------
   1  |    2    6   12
   2  |    6   16   30
   3  |   12   30   54

   M  |    2    6   12   16   30   54
 -----+--------------------------------
  9p  |    1    2    2    1    2    1    total : 9

variable M moyenne    :    18.667 (somme des valeurs :    168)
           variance   :   244.444 (somme des carrés  :   5336)
           écart-type :    15.635
```

```
          cdv        :    83.757 % (coefficient de variation)

étude de la variable : D = valeur absolue de la différence

  D  |    1    2    3
-----+--------------------------------
  1  |    0    1    2
  2  |    1    0    1
  3  |    2    1    0

  D  |    0    1    2
-----+-----------------
 9p  |    3    4    2    total : 9

variable D moyenne    :     0.889 (somme des valeurs :    8)
          variance    :     0.543 (somme des carrés  :   12)
          écart-type  :     0.737
          cdv         :    82.916 % (coefficient de variation)



Matrice de corrélation

       S         P         M         D

 S     1.0000
 P     0.9608    1.0000
 M     0.9355    0.9934    1.0000
 D     0.0000   -0.2509   -0.2828    1.0000

-- fin des calculs
```

Ayant écrit un programme qui effectue ces calculs pour n variable, nous forrnissons les résultats pour $n = 6$:

```
 Etude de 2 dés avec chacun 6 faces
 ... On passe en revue les 36 cas    possibles.

 étude de la variable : S = somme des valeurs

  S  |    1    2    3    4    5    6
-----+--------------------------------
  1  |    2    3    4    5    6    7
  2  |    3    4    5    6    7    8
  3  |    4    5    6    7    8    9
  4  |    5    6    7    8    9   10
  5  |    6    7    8    9   10   11
  6  |    7    8    9   10   11   12

  S  |    2    3    4    5    6    7    8    9   10   11   12
-----+-------------------------------------------------------
36p  |    1    2    3    4    5    6    5    4    3    2    1    total : 36

 variable S moyenne    :     7.000 (somme des valeurs :   252)
            variance   :     5.833 (somme des carrés  :  1974)
            écart-type :     2.415
            cdv        :    34.503 % (coefficient de variation)


 étude de la variable : P = produits des valeurs

    P  |    1    2    3    4    5    6
  -----+--------------------------------
    1  |    1    2    3    4    5    6
    2  |    2    4    6    8   10   12
    3  |    3    6    9   12   15   18
    4  |    4    8   12   16   20   24
    5  |    5   10   15   20   25   30
    6  |    6   12   18   24   30   36
```

```
  P  |    1    2    3    4    5    6    8    9   10  ...
-----+-----------------------------------------------
36p  |    1    2    2    3    2    4    2    1    2  ...
```

```
     (suite)      12   15   16   18   20   24   25   30   36
                 ---------------------------------------------
                    4    2    1    2    2    2    1    2    1    total : 36
```

```
variable P moyenne    :    12.250 (somme des valeurs :    441)
           variance   :    79.965 (somme des carrés  :   8281)
           écart-type :     8.942
           cdv        :    72.999 % (coefficient de variation)
```

étude de la variable : M = somme*produit des valeurs

```
  M  |    1    2    3    4    5    6
-----+--------------------------------
  1  |    2    6   12   20   30   42
  2  |    6   16   30   48   70   96
  3  |   12   30   54   84  120  162
  4  |   20   48   84  128  180  240
  5  |   30   70  120  180  250  330
  6  |   42   96  162  240  330  432
```

```
  M  |    2    6   12   16   20   30   42   48   54   70 ...
-----+-------------------------------------------------------
36p  |    1    2    2    1    2    4    2    2    1    2 ...
```

```
    (suite)   84   96  120  128  162  180  240  250  330  432
             ---------------------------------------------------
               2    2    2    1    2    2    2    1    2    1

total : 36
```

```
variable M moyenne    :   106.167 (somme des valeurs :   3822)
           variance   : 11034.528 (somme des carrés  : 803012)
           écart-type :   105.045
           cdv        :    98.944 % (coefficient de variation)

étude de la variable : D = valeur absolue de la différence

  D  |    1    2    3    4    5    6
-----+--------------------------------
  1  |    0    1    2    3    4    5
  2  |    1    0    1    2    3    4
  3  |    2    1    0    1    2    3
  4  |    3    2    1    0    1    2
  5  |    4    3    2    1    0    1
  6  |    5    4    3    2    1    0

  D  |    0    1    2    3    4    5
-----+--------------------------------
 36p |    6   10    8    6    4    2    total : 36


variable D moyenne    :     1.944 (somme des valeurs :    70)
           variance   :     2.052 (somme des carrés  :   210)
           écart-type :     1.433
           cdv        :    73.679 % (coefficient de variation)



Matrice de corrélation

      S         P         M         D
 S    1.0000
 P    0.9453    1.0000
 M    0.9120    0.9889    1.0000
 D    0.0000   -0.2808   -0.2864    1.0000

-- fin des calculs
```

4. Centrage et Réduction

Enoncé

Soit X une v.a. ; construire une v.a. Y liée linéairement à X dont la moyenne est nulle. On la nomme la variable centrée issue de X. Construire une v.a. Z liée linéairement à X dont l'écart-type vaut 1. On la nomme la variable réduite issue de X. Construire une v.a. T liée linéairement à X telle que $m(T) = 0$ et $\sigma(T) = 1$. On la nomme la variable centrée réduite issue de X. Comparer T avec les variable $X - m(X)/\sigma(X)$ et $X/\sigma(X) - m(X)$.

Solution

Posant $Y = aX + b$, la condition $m(Y) = 0$ impose $a = 1$ et $b = -m_X$. De même pour $Z = cX + d$ où la condition $\sigma(Z) = 1$ impose $d = 0$ et $c = 1/\sigma$. Si la variable $T = eX + f$ vérifie $m(T) = 0$ et $\sigma(T) = 1$, alors $e = 1/\sigma$ et $f = -m$ et T est donc la variable $(X - m)/\sigma$. Par contre $m(X - m(X)/\sigma(X)) = m(X)(1 - 1/\sigma)$ ce qui ne donne rien de remarquable, pas plus que $m(X/\sigma(X) - m(X)) = m(X)(1/\sigma - 1)$.

5. Loi de *Bernoulli* généralisée

Enoncé

Soit U=$b(x, y, p)$ la loi de Bernoulli généralisée, qui prend les valeurs x et y avec les probabilités respectives p et $1 - p$ avec $x \leqslant y$. Calculer directement $m(\mathsf{U})$ et $V(\mathsf{U})$. En remarquant que U=$a.\mathsf{T}$+b où T=$b(p)$, calculer a et b puis retrouver les valeurs de $m(\mathsf{U})$ et $V(\mathsf{U})$.

Solution

Le calcul direct donne : $m(\mathsf{U}) = (1 - p)x + py = x + (y - x)p$; de même $V(\mathsf{U}) = (1 - p)x^2 + py^2 - m(\mathsf{U})^2$ soit $V(\mathsf{U}) = (y - x)^2 p(1 - p)$. On peut aussi écrire $\mathsf{U} = (y - x)\mathsf{T} + x$ d'où, comme $m(aX + b) = a.m(X) + b$ et $V(aX + b) = a^2 V(X)$, $m(\mathsf{U}) = (y - x)m(\mathsf{T}) + x = (y - x)p + x$ et $V(\mathsf{U}) = (y - x)^2 V(\mathsf{T}) = (y - x)^2 p(1 - p)$.

6. Lois et valeurs comme $V < m$ pour $\mathcal{B}(n, p)$

Enoncé

Un élève prétend avoir calculé $m_X = 3.123456$, $V_X = 1.654321$.
Est-ce possible ?

Un autre élève prétend avoir trouvé $m_X = 3.123456$, $m_{X^2} = 1.654321$.
Est-ce possible ?

Un troisième enfin prétend avoir $V_X = 3.123456$ et $m_X = 1.654321$.
Est-ce possible ?

Trouvez une façon simple de démontrer que V_X est toujours positif ou nul.

Solution

Les valeurs $m_X = 3.123456$ et $V_X = 1.654321$, sans aucune autre information, sont possibles. Par contre, $m_X = 3.123456$ et $m_{X^2} = 1.654321$ est impossible car $V(X) = m_{X^2} - m_X^2 \geq 0$ et donc on a toujours $m_{X^2} \geq m_X^2$ ce qui n'est pas le cas ici. Enfin, $V_X = 3.123456$ et $m_X = 1.654321$ sans aucune autre information, est possible. Toutefois, si on sait que X est une loi binomiale, c'est impossible car pour une telle loi $V = np(1-p)$ est forcément inférieur à $m = np$ puisque p (et donc $1-p$) est inférieur à 1.

Pour montrer facilement que V_X est toujours positif ou nul, on utilise non pas la formule $m_{X^2} - m_X^2$ mais la formule $m(\ (X - m_X)^2\ )$ car la moyenne de carrés (donc positifs) est forcément positive...

7. Qu'est-ce qu'une moyenne ?

Enoncé

On appelle moyenne arithmétique de deux valeurs x et y la quantité $(x+y)/2$, moyenne géométrique la quantité $\sqrt{x * y}$, moyenne quadratique $\sqrt{(x^2 + y^2)/2}$, moyenne harmonique $2/((1/x) + (1/y))$.

Généraliser à n valeurs $x_1, x_2 ... x_n$ plutôt que x et y.

Montrer que les différentes fonction-moyennes vérifient les propriétés suivantes :

$$min\{x_i\} \leqslant moy(\ (x_i)\ ) \leqslant max\{x_i\}$$

$$\forall i x_i = c \Rightarrow moy(\ (x_i)\ ) = c$$

$moy(\ (x_i)\ )$ est invariante par permutation des x_i

Comparer m_a, m_g, m_q et m_h pour $x = 2$, $y = 8$. Et dans le cas général ?

Soient α_i des réels. Quelle(s) condition(s) doit-on imposer aux α_i pour que, les m_i désignant des fonctions-moyennes, la combinaison linéaire $\sum \alpha_i m_i$ soit aussi une fonction-moyenne ?

Solution

La moyenne arithmétique de n valeurs x_i pour i de 1 à n est définie par $m_a(\ (x_i)\ ) = \dfrac{1}{n}\sum_{i=1}^{n} x_i$.

Leur moyenne géométrique est $m_g(\ (x_i)\ ) = \sqrt[n]{\prod_{i=1}^{n} x_i}$ pour des x_i positifs.

Leur moyenne quadratique est $m_q(\ (x_i)\ ) = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n} x_i^2}$.

Leur moyenne harmonique est $m_h(\ (x_i)\ ) = \dfrac{n}{\sum_{i=1}^{n} \dfrac{1}{x_i}}$.

Si pour tout i, $x_i = c > 0$, alors $m_a(\ (x_i)\ ) = n.c/n = c$.

De même $m_g(\ (x_i)\ ) = \sqrt[n]{c^n} = c$, $m_q(\ (x_i)\ ) = \sqrt{nc^2/n} = c$, $m_h(\ (x_i)\ ) = n/(n/c) = c$.

Puisque $min\{x_i\} \leq x_i \leqslant max\{x_i\}$, en sommant terme à terme puis en divisant terme à terme par n, on déduit $min\{x_i\} \leqslant m_a(\ (x_i)\ ) \leqslant max\{x_i\}$. Pour les autres moyennes, un calcul analogue (mais plus technique) aboutit aux mêmes conclusions.

L'addition étant commutative, m_a n'est pas sensible à une permutation des x_i. Même remarque pour les autres moyennes.

Pour $x = 2$, $y = 8$, $m_a(x, y) = (8 + 2)/2 = 10/2 = 5$.

$m_g(x, y) = \sqrt{8 + 2} = \sqrt{16} = 4$ $m_h(x, y) = 1/((1/2) + (1/8)) = 8/5 = 1.6$
$m_q(x, y = \sqrt{(2^2 + 8^2)/2} = \sqrt{34} = 5.83$

Donc $m_h < m_g < m_a < m_q$, ce qui est le cas général mais dont la démonstration est plus technique.

Remarque :

La moyenne harmonique est souvent définie par la formule

$$n \times \frac{1}{m_h(\ (x_i)\ )} = \sum_{i=1}^{n} \frac{1}{x_i}$$

ce qui est peut-être plus "parlant".

T.D. 4 : Lois classiques et approximations

1. Conditions sur n sachant $V \geqslant 10$ pour $\mathcal{B}(n, p)$
2. Loi de u erreurs dans un livre de v pages
3. Saturation d'un serveur multiposte
4. Retard annuel d'une montre et hypothése "déraisonnable"
5. Calculs concrets de χ^2 : pièces de fonderie
6. Approximations $\mathcal{B}$ et $\mathcal{P}$: filtrage de substrats

1. Conditions sur n sachant $V \geqslant 10$ pour $\mathcal{B}(n,p)$

Enoncé

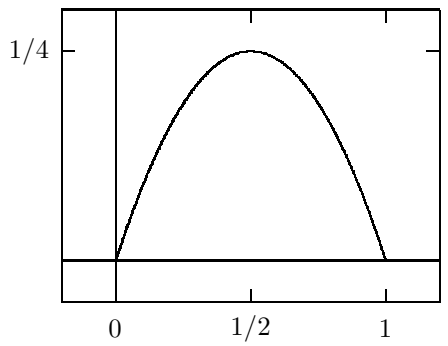
Montrer que $V \geqslant 10 \Rightarrow n \geqslant 40$ pour $\mathcal{B}(n,p)$.

Solution

On sait que la moyenne m de $\mathcal{B}(n,p)$ est $np(1-p)$. L'étude de la fonction $f:\ x \ \rightarrow\ x(1-x)$ sur $[0,1]$ est résumée par le tableau de variation ci-dessous :

x	0		1/2		1
f'(x)		+	0	−	
f(x)	0	↗	1/4	↘	0

Un tracé rapide de f montre que $f(x) \leqslant 1/4$ pour tout x :



D'où : $n < 40$ et $p(1-p) < 1/4 \Rightarrow np(1-p) < 10$. $Cqfd$.

2. Loi de u erreurs d'affichage pour v pages Web

Enoncé

Un serveur Web fournit par programme 1000 pages Web par jour. En phase de tests, on constate globalement 1500 erreurs d'affichage sur ces 1000 pages. On appelle X la v.a. "nombre d'erreurs pour une page donnée". Quelle est la loi de X ? sa moyenne ? son écart-type ?

Solution

En l'absence d'hypothèse, nous déciderons que les erreurs sont idépendantes les unes des autres. Chaque erreur apparait ou n'apparait pas et suit donc une loi binaire (*bernoulli*) de paramètre $p = 1/1000$. X est la loi "compte" du nombre d'erreurs, c'est à dire la somme de $n = 1500$ lois $b(p)$. X est donc la loi binomiale $\mathcal{B}(1500, 0.001)$. Sa moyenne $m = np$ vaut donc 1.5 et sa variance $V = np(1-p)$ vaut 1.4985 d'où un écart-type $\sigma = \sqrt{V}$ de 1.2241.

3. Saturation d'un serveur multiposte

Enoncé

Un serveur/concentrateur dessert 1000 postes via 50 lignes à haut débit. Aux heures de pointe, chaque poste est occupé en moyenne pendant 2,5 secondes par minute. Quelle est la probabilité de saturation du réseau pendant une durée moyenne d'une minute de pointe ?

Solution

L'énoncé ne propose aucune v.a., donc c'est à nous de tout inventer !

Prenons comme évènement élémentaire "un poste est occupé". Cet évènement est binaire : il se produit ou ne se produit pas, en terminologie probabiliste "il est réalisé" ou il ne l'est pas, et on peut donc le modéliser par une loi $b(p)$.

L'énoncé indique que l'on doit prendre $p = 2,5/60$ si on ramène tout en secondes. Maintenant, introduisons la v.a. X définie par X est "le nombre de postes occupés désirant communiquer avec le serveur".

X correspond alors à la somme des $n = 1000$ évènements élémentaires que l'on considérera comme indépendants.

On peut représenter X par $\mathcal{B}(n, p) = \mathcal{B}(1000, 0.0417)$ La saturation correspond à l'évènement $X > 50$. Sa probabilité est

$$p(X \geqslant 51) = \sum_{k=51}^{1000} C_{1000}^{k}\, p^k (1-p)^{1000-k}$$

Elle vaut en gros 0,084.

On peut la calculer directement, sous Maple par exemple avec les expressions :

```
Digits := 15 ;
p := 2.5/60 ; q := 1 - p
v := Sum(binomial(n,k)*p^k*q^(n-k),k=51..n) ;
n := 1000 ;
evalf(v) ;
```

ce qui donne comme résultat numérique .08413259 en quelques secondes.

On verra dans un autre T.D. comment réaliser une approximation de cette probabilité en utilisant la fonction de répartition de la loi normale $\mathcal{N}(0, 1)$.

Remarque :

Volontairement, nous n'indiquons pas ici comment calculer cette valeur avec *Rstat*.

4. Retard annuel d'une montre

Enoncé

Une montre fait une erreur d'au plus 30 secondes par jour (dans un sens ou dans un autre). Quelle est la probabilité que l'erreur soit inférieure à 15 minutes au bout d'un an ?

En quoi cet énoncé est-il "déraisonnable", "irréaliste" ?

Solution

Soit X_i l'erreur en minute au jour i. X_i peut être modélisée par une loi réelle uniformément répartie sur $[-1/2, 1/2]$. Supposant les jours indépendants, l'erreur S au bout d'un an est la somme des X_i pour i de 1 à 365 (on ignore "poliment" les années bisextiles) : $S = \sum_{i=1}^{365} X_i$. Compte-tenu de nos connaissances actuelles, il n'est pas possible de résoudre la question posée, à savoir calculer $p(|S| \leqslant 15)$.

On verra, là encore dans un autre T.D., comment réaliser une approximation de cette probabilité en utilisant la fonction de répartition de la loi normale $\mathcal{N}(0, 1)$.

Les hypothèses "déraisonnables", "irréalistes" sont de supposer "les jours indépendants", ce qui ne veut rien dire. Il serait plus "naturel" de penser que si une montre retarde ou avance, c'est qu'il s'agit d'un problème mécanique et non pas d'une volonté ou d'un caprice.

En conséquence, il y a de fortes chances que ce problème reste constant ou s'aggrave, ce qui implique que le retard (ou l'avance) reste le même ou augmente régulièrement, ce qui signifie donc que les X_i ne sont pas indépendantes.

5. Calculs concrets de χ^2 : pièces de fonderie

Enoncé

Le conditionnement de pièces de fonderie a conduit à ventiler la population totale (180 pièces) en 6 lots contenant respectivement 29, 41, 31, 29, 18, 32 pièces.

Le conditionnement théorique aurait abouti pour la même population aux valeurs 30, 40, 30, 30, 20, 30.

Le calcul du χ^2 est-il possible ? Si oui, combien trouve-t-on ?

Discuter alors la significativité du test sous-jacent.

Solution

Puisque :

- le total des effectifs observés et des effectifs théoriques est le même (il vaut 180 dans les deux cas),
- chaque effectif est positif et supérieur à 5
- le total des effectifs est supérieur à 50,

le calcul du χ^2 est possible.

Le détail du calcul des différences pondérées et de leur somme est le suivant

```
  obs      th   d=obs-th       d*d/th    chideux
29.00   30.00    -1.00      0.0333333    0.03
41.00   40.00     1.00      0.0250000    0.06
31.00   30.00     1.00      0.0333333    0.09
29.00   30.00    -1.00      0.0333333    0.12
18.00   20.00    -2.00      0.2000000    0.33
32.00   30.00     2.00      0.1333333    0.46
```

Le χ^2 est donc 0.46 ; pour 6 classes, le χ^2 théorique avec 6-1=5 degrés de liberté au seuil de 5 % est 11.10 ; au seuil de 5 % on peut donc accepter l'hypothèse que les deux distributions suivent le même modèle.

6. Approximations $\mathcal{B}$ et $\mathcal{P}$

Enoncé

Soit x_i le nombre de fois où on doit filtrer un substrat organique avant d'être sur de sa pureté. Compte-tenu des techniques modernes de filtration, il est très peu probable que ce nombre dépasse 5 fois et on admettra donc que la valeur "5 fois" représente en fait l'évènement "5 fois ou plus". On fournit dans le tableau suivant le nombre n_i de substrats ayant été filtré x_i fois.

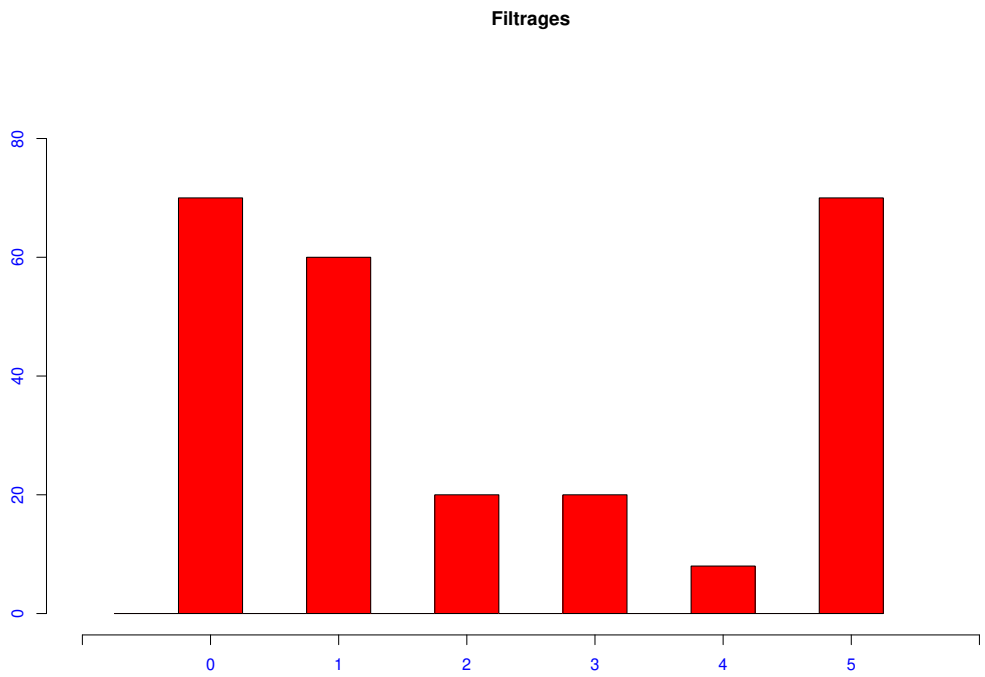
x_i	0	1	2	3	4	5
n_i	70	60	20	20	8	70

– Donner le total, la moyenne et la variance du nombre de filtrages.
– Effectuer une approximation des effectifs n_i par la loi binomiale.
– Effectuer une approximation des effectifs n_i par la loi de Poisson.
– Conclure en comparant ces approximations.

Vous n'oublierez pas de détailler les calculs, de justifier les effectifs théoriques réels et arrondis, de fournir le χ^2 utilisé, le nombre de degrés de liberté *etc.*

Solution

Commençons par remarquer que la distribution des n_i est en "creux" (forme parabolique positive) comme le montre l'histogramme des effectifs



L'approximation par des lois unimodales en "plein" (forme parabolique négative) que sont les lois binomiale et de Poisson sera donc certainement impossible.

La somme des n_i vaut 248.

Leur moyenne pondérée est $m = 2{,}185$ filtrages,

La variance pondérée est 4,103 soit un écart-type de 2,026 filtrages.

L'application du modèle binomial $\mathcal{B}(n, p)$ doit donc de faire avec les valeurs $n = 5$ et $p = m/n = 0{,}437$.

Le détail de la construction des probabilités dans le modèle binomial $\mathcal{B}(5; 0, 437)$ est

xi	c(n,k)	p^k	(1-p)^k	probabilite
0	1	1.000000	0.056516	0.0565
1	5	0.437097	0.100400	0.2194
2	10	0.191054	0.178362	0.3408
3	10	0.083509	0.316860	0.2646
4	5	0.036501	0.562903	0.1027
5	1	0.015955	1.000000	0.0160

ce qui aboutit aux effectifs théoriques suivants

Ni	Xi	Pi	Obs	ThExact	ThArrondi
70	0	0.057	70	14.016	14
60	1	0.219	60	54.417	54
20	2	0.341	20	84.510	85
20	3	0.265	20	65.622	66
8	4	0.103	8	25.478	25
70	5	0.016	70	3.957	4

Il n'est alors pas possible de calculer directement le χ^2 car un effectif est trop faible (inférieur à 5). Il faut donc effectuer des regroupements.

On regroupe les deux dernières classes ensemble et on dispose alors des valeurs observées et théoriques suivantes

obs	70	60	20	20	78
th	14	54	85	66	29

Ces deux séries de valeurs sont "manifestement très différentes" et c'est ce que montre le calcul du χ^2 qui vaut 389,23.

i	obs	th	obs-th	tc	cumul
1	70	14	56	224.00000	224.000
2	60	54	6	0.66667	224.667
3	20	85	-65	49.70588	274.373
4	20	66	-46	32.06061	306.433
5	78	29	49	82.79310	389.226

Or pour 4 degrés de liberté, au seuil de 5 %, le χ^2 théorique maximal autorisé lu dans la table est 9,49. On doit donc refuser l'hypothèse que nos données suivent une loi binomiale.

L'approximation par la loi de Poisson commence par utiliser la valeur 2.185 de la moyenne pondérée de nos données comme paramètre λ de la loi de Poisson. Si on calcule les 6 premières valeurs de cette loi, on obtient les valeurs

xi	ni	pi
0	70	0,11242
1	60	0,24570
2	20	0,26849
3	20	0,19559
4	08	0,10687
5	70	0,04671

La somme de ces probabilités vaut 0,97578.

On doit donc rajouter 1-0,97578=0,02422 à p_5. Après avoir multiplié les probabilités par 248 et après avoir arrondi les valeurs obtenues, on a la mauvaise surprise d'obtenir un total d'effectifs théoriques égal à 250.

xi	ni	pi*	thir	thi*
0	70	0,11242	27,88016	28
1	60	0,2457	60,93360	61
2	20	0,26849	66,58552	67
3	20	0,19559	48,50632	49
4	08	0,10687	26,50376	27
5	70	0,07093	17,59064	18
somme	250	1	248	250

On retranche alors 1 aux effectifs de probabilités les plus proches de $x, 5$ et on peut ensuite calculer le χ^2. Les deux effectifs incriminés sont ici 48,50632 et 26,50376. On leur associe alors respectivement les valeurs 48 et 16.

Finalement, les effectifs à comparer et le calcul du χ^2 est résumé dans le tableau suivant.

xi	obs	thi*	thi	diff	tc	cumul
0	70	28	28	42	63,00000	63,00000000
1	60	61	60	-1	0,01639	63,01639344
2	20	67	67	20	32,97015	95,98654270
3	20	49	48	20	16,33333	112,3198760
4	08	27	26	8	12,46154	124,7814145
5	70	18	18	70	150,22222	275
somme	248	250	248		275	

On obtient 275 ce qui est beaucoup trop élevé pour qu'on puisse accepter l'hypothèse que nos données suivent une loi de Poisson. Il faut donc rédiger la phrase usuelle "*au risque de première espèce* $\alpha = 5\,\%$, *on refuse l'hypothèse que nos données suivent la loi de Poisson* $\mathcal{P}(2.185)$".

T.D. 5 : Programmation probabiliste

1. Convergence de α_n pour le sprite
2. Programmons la loi binomiale
3. Programmons la loi de Poisson
4. Ce fichier induit-il une distance ?

1. Convergence de α_n pour le sprite

Enoncé

Ecrire un algorithme qui suit la syntaxe de GALG pour afficher les valeurs de n, α_n et $|\alpha_n - 1/3|$. On affichera les valeurs pour $|\alpha_n - 1/3| > \varepsilon$ où ε est une précision donnée fixée, par exemple 10^{-5}.

Solution

Tout d'abord, pour ceux et celles qui ne connaissent pas GALG, il faut consulter les pages *Web* qui le présentent, à savoir les pages à partir de l'URL

`http://www.info.univ-angers.fr/pub/gh/Galg/galg.htm`

En deux mots, GALG est à la fois un essai de normalisation de langage algorithmique et un programme de traduction (et d'exécution dans la foulée) des algorithmes en langages de programamtion réels. GALG traduit actuellement les algorithmes en C, C++, Dbase, Java, Pascal, Perl, Rexx et Tcl.

Pour le sprite, on doit programmer l'affichage d'un certain nombre de lignes (non connu à l'avance) en fonction de la distance entre α_n et $1/3$ où α_n se calcule par

$$\alpha_n = \frac{1}{3} + \frac{2}{3}\left(\frac{-1}{2}\right)^n$$

Une boucle `tant que` suffit. On pourrait la réduire à :

```
affecter nbval  <--  0

tant_que abs(alpha(nbval) - 1/3) > epsilonne
     affecter nbval  <--  nbval + 1
fin_tant_que # abs(alpha(nbval) - 1/3) > epsilonne
```

Toutefois l'énoncé précise que l'on veut un bel affichage. Il faut donc utiliser des formats entiers et réels et afficher une en-tête des colonnes,

Ceci nous oblige à stocker le contenu des fonctions et différences dans des variables sous peine de calculer plusieurs fois le même terme. En effet, si nous écrivons

```
tant_que abs(alpha(nbval) - 1/3) > epsilonne
     ecrire nbval, alpha(nbval),  abs(alpha(nbval) - 1/3)
     affecter nbval  <--  nbval + 1
fin_tant_que # abs(alpha(nbval) - 1/3) > epsilonne
```

alors nous faisons calculer 3 fois α_n au programme.

Voici l'algorithme complet

```
# pour le sprite ; auteur (gH) 2001
# remarque : alpha est une fonction externe
#            et epsilonne une constante donnée

écrire " Etude de la convergence pour le sprite "
écrire ""
écrire "  n  alpha(n)   différence avec 1/3 "
écrire ""

affecter nbval  <--  0
affecter valc   <--  alpha(nbval)
affecter dif    <--  abs( valc - 1/3)

tant_que dif > epsilonne

     affecter f_n    <-- formatEntier(nbval,3)
     affecter f_a    <-- formatReel(valc,12,8)
     affecter f_d    <-- formatReel(dif,12,8)

     écrire f_n, f_a, f_d

     affecter nbval  <--  nbval + 1
     affecter valc   <--  alpha(nbval)
     affecter dif    <--  abs( valc - 1/3)

fin_tant_que # dif > epsilonne
```

Pour exécuter en `rexx` l'algorithme traduit par GALG on rajoute ε et la définition de α_n soit :

```
/*  fichier sprite.gex : on a complété la traduction de GALG */

epsilonne = 10**(-5)

/* ##-#   Fichier sprite.rex issu de galg -a sprite.alg -o rexx */
/* ##-#  ================================================== */
/* ##-#      17/03/2001 10:1.54 */

/* # pour le sprite ; auteur (gH) 2001 */
/* # remarque : alpha est une fonction externe */
/* #            et epsilonne une constante donnée */

say " Etude de la convergence pour le sprite "
say ""
say "  n  alpha(n)   différence avec 1/3 "

nbval =  0
valc  =  alpha( nbval )
dif   =  abs(valc  -  1  /  3)

do while dif  >  epsilonne
     f_n =  format(nbval ,  3)
     f_a =  format(valc ,  12 ,  8)
     f_d =  format(dif ,  12 ,  8)
     say f_n f_a f_d
     nbval =  nbval  +  1
     valc  =  alpha( nbval )
     dif   =  abs(valc  -  1  /  3)
end /* # dif > epsilonne */
return

/* ##-# ===========  Fin de traduction par GALG =========== */

alpha: procedure
  parse arg n
  return   (1/3) + (2/3)*(-1/2)**n
/* fin de procédure alpha */
```

2. Programmons la loi binomiale

Enoncé

Programmer "bêtement" l'affichage des valeurs et des probabilités pour la loi Binomiale à l'aide des fonctions $coefbin$ et $puiss$. On utilisera un algorithme qui suit la syntaxe de GALG.

Affiner en trouvant une relation de récurence d'un terme à l'autre.

Compléter enfin l'algorithme par l'affichage du cumul des probabilités, par le calcul d'effectifs entiers pour un effectif total donné.

Solution

L'algorithme "bête et méchant" se trouve dans le cours, soit :

```
# binom1.alg ; auteur (gH)
# remarque les paramètres n et p sont connus
# nommés ici parmn et proba
# version 1 : "bête et méchante"

affecter ump <-- 1 - proba # calcul de 1-p
pour indk de0a parmn
   affecter cnk  <-- coefbin(parmn,indk)
   affecter ppp  <-- puiss(proba,indk )
   affecter qqq  <-- puiss(ump,(parmn-indk))
   affecter valp <-- cnk * ppp * qqq
   écrire indk , valp
fin_pour # indk de1a parmn
```

Si on écrit C_n^k de façon efficace sous la forme

$$C_n^k = \frac{n(n-1)(n-2)...(n-k+1)}{k(k-1)(k-2)...1}$$

au lieu de l'expression $n!/k!(n-k)!$ classique, alors on voit que pour passer de C_n^{k-1} à C_n^k il faut multiplier par $n-k$ et diviser par k.

Bien sûr pour passer de p^{k-1} à p^k il suffit de multiplier par p et pour passer de $q^{n-(k-1)}$ à q^{n-k} il suffit de diviser par q soit l'algorithme :

```
# binom2.alg ; auteur (gH)
# remarque les paramètres n et p sont connus
# nommés ici parmn et proba
# version 2 : on optimise les calculs

affecter ump  <-- 1 - proba # calcul de 1-p
affecter cnk  <-- 1
affecter ppp  <-- 1
affecter qqq  <-- puiss(ump,parmn)
pour indk de0a parmn
   affecter valp <-- cnk * ppp * qqq
   écrire indk , valp
   affecter cnk  <-- cnk * (parmn-indk) / indk
   affecter ppp  <-- ppp * proba
   affecter qqq  <-- qqq / ump
fin_pour # indk de1a parmn
```

Enfin, si l'on veut la fonction de répartition, il suffit de cumuler les valeurs de probabilité. Il n'y a donc qu'une affectation à rajouter dans la boucle `pour`. De même, l'effectif correspondant s'effectue avec une simple multiplication suivie d'un arrondi, soit là encore une simple affectation à rajouter dans la boucle `pour`. Voici finalement le "bel" algorithme :

```
# binom3.alg ; auteur (gH)
# remarque les paramètres n et p sont connus
# nommés ici parmn et proba
# version 3 : on optimise les calculs et on calcule
# la fonction de répartition, ainsi que les effectifs
# associés pour un effectif total nommé efftot (connu)

affecter ump  <-- 1 - proba # calcul de 1-p
affecter cnk  <-- 1
affecter ppp  <-- 1
affecter qqq  <-- puiss(ump,parmn)
```

```
affecter cumu <-- 0

écrire " Loi Binomiale B(" , n , "," , p ,")"
écrire "   effectif total " , efftot
écrire ""
écrire "  k  p_k  cumul effectif"
écrire ""

pour indk de0a parmn

   affecter valp <-- cnk * ppp * qqq
   affecter cumu <-- cumu + valp
   affecter effe <-- arrondi(valp*efftot)

   écrire indk , valp , cumul , effe

   affecter cnk  <-- cnk * (parmn-indk) / indk
   affecter ppp  <-- ppp * proba
   affecter qqq  <-- qqq / ump

fin_pour # indk de1a parmn
```

3. Programmons la loi de Poisson

Enoncé

Adapter l'algorithme précédent à la loi de Poisson.

Si on en fait des programmes et des sous-programmes, vaut-il mieux deux sous-programmes distincts $loiB$ et $loiP$ ou un seul sous-programme $lois$ avec un premier paramètre pour distinguer la loi binomiale qu'on appelerait alors par $loi("B", ...)$ de la loi de Poisson qu'on appelerait alors par $loi("P", ...)$?

Faut-il traduire ces algorithmes en *Rstat* ?

Solution

Il n'y aucune difficulté à programmer la loi de Poisson car c'est le même principe que pour la loi binomiale. D'où l'algorithme

```
# poisson.alg ; auteur (gH)

# on affiche les k premières valeurs de
# la loi de poisson de paramètre lambda
# on optimise les calculs
# et on calcule la fonction de répartition,
# ainsi que les effectifs associés pour un
# effectif total nommé efftot (connu)

# k est noté ici parmk

affecter cst  <-- exp(-parmk)
affecter fac  <-- 1 # pour la factorielle
affecter ppp  <-- 1 # pour la puissance
affecter cumu <-- 0 # pour la fdr
écrire " Loi de Poisson P(" , lamnda ,")"
écrire "   tronquée aux  " , parmk , " premiers termes "
écrire "   effectif total " , efftot
écrire ""
écrire "  i  p_i  cumul effectif"
écrire ""
pour indi de1a parmk
   affecter valp <-- cst * ppp / fac
   affecter cumu <-- cumu + valp
   affecter effe <-- arrondi(valp*efftot)
   écrire (indi-1) , valp , cumul , effe
   affecter ppp  <-- ppp * lambda
   si indi>1
      alors
        affecter fac  <-- fac * (indi-1)
   fin_si # indi>1
fin_pour # indk de1a parmn
```

En ce qui concernce les programmes et sous-programmes, nous pensons que cela n'a aucune importance d'avoir un ou deux sous-programmes. Il est sans doute plus cohérent pour l'utilisateur de disposer de fonctions distinctes puisqu'il y a un nombre de paramètres différents pour ces lois, mais rien n'empêche d'avoir en interne un seul programme de calcul et de fournir deux interfaces d'appel. De même, il est tout à fait possible d'avoir deux sous-programmes de calcul et de fournir une interface d'appel commune.

si on a	*on invente*	*défini par*
loiB(n,p) loiP(m,k)	loip(a,b,c) "	loi("B",n,p) = loiB(n,p) loi("P",m,k) = loiP(m,k)
loi(a,b,c)	loiB(a,b) loiP(a,b)	loiB(n,p) = loi("B",n,p) loiP(m,k) = loi("P",m,k)

Par contre, il ne faut certainement pas réinventer ces fonctions car *Rstat* en dispose déjà, certainement mieux programmées...

4. Ce fichier induit-il une distance ?

Enoncé

On dispose d'un fichier qui contient une matrice triangulaire inférieure, comme par exemple

```
Baboon     0.00000
Gibbon     0.18463 0.00000
Orang      0.19997 0.13232 0.00000
Gorilla    0.18485 0.11614 0.09647 0.00000
PygmyCh    0.17872 0.11901 0.09767 0.04124 0.00000
Chimp      0.18213 0.11368 0.09974 0.04669 0.01703 0.00000
Human      0.17651 0.11478 0.09615 0.04111 0.03226 0.03545 0.00000
```

Ecrire un algorithme qui lit ces valeurs et remplit en conséquence le tableau $tabDist$ où $tabDist[i, j]$ correspond à la ligne i et à la colonne j du fichier (on ignorera poliment ou on stockera ailleurs les identificateurs).

La fonction d induite par $(i, j) \mapsto d(i, j) = tabDist[i, j]$ est-elle une distance ?

Solution

Rappelons qu'une distance est une fonction à valeurs positives ou nulles (ce qui fait un premier test élémentaire) et qu'il y a trois conditions à vérifier pour qu'une fonction soit une distance ce qui fait trois autres tests à effectuer :

- test de dissimilarité, correspondant à la propriété
$$(\forall x, y) \;\; d(x, y) = 0 \Leftrightarrow x = y$$
- test de symétrie, correspondant à la propriété
$$(\forall x, y) \;\; d(x, y) = d(y, x)$$
- test de triangularité, correspondant à la propriété
$$(\forall x, y, z) \;\; d(x, y) \leqslant d(x, z) + d(z, y)$$

Notre algorithme devrait donc se composer de cinq parties, à savoir

- lecture du fichier et remplissage du tableau,
- vérification de la positivité,
- vérification de la dissimilarité,
- vérification de la symétrie,
- vérification de la triangularité.

Toutefois, il n'y a pas à tester la symétrie car on ne lit que la partie inférieure de la matrice.

Une première question se pose qui est de savoir ce qu'il faut faire en cas d'erreur lorsqu'une propriété n'est pas vérifiée. On peut envisager de quitter le programme, sans aller plus loin ou on peut décider de continuer pour montrer toutes les erreurs à corriger. Dans le premier cas il faut utiliser des boucles `tant_que` avec des tests sur les indices de boucle et sur la validité des propriétés alors que dans le deuxième cas il suffit de mettre des boucles `pour` partout.

Nous programmerons une solution mixte, à savoir : le test de positivité est primordial ; s'il échoue, le programme s'arrête. Par contre pour chacun des deux autres tests, on n'affichera que la première erreur rencontrée mais nous enchainerons les tests.

La partie lecture du fichier est sans doute la plus simple à programmer, à condition de disposer de fonctions comme *mots* et *mot* dont la syntaxe est $mots(chaine)$ et $mot(chaine, indice)$ et qui renvoient respectivement le nombre de mots et le i-ème mot d'une chaine.

Après réflexion, nous décidons de tester la positivité au cours de lecture du fichier en conservant les indices de ligne et de colonne éventuels de la dernière valeur négative (si elle existe) et en calculant le nombre de valeurs négatives, ce qui donnera une idée du nombre de termes à corriger.

Pour désigner nos variables il suffit de commenter le fichier des variables fourni par GALG déjà trié par ordre alphabétique.

```
* Liste des variables (par ordre alphabétique)

Variable

 #  ficmatd   variable chaine de caractère qui contient
              le nom du fichier à traiter
 #  icneg     indice colonne du nombre négatif
 #  ilig      indice courant de ligne
 #  ilneg     idice  ligne   du nombre négatif
 #  indm      indice courant de mot
 #  jcol      indice courant de colonne
 #  lig       ligne lue dans le fichier
 #  nblig     nombre courant puis total de lignes
 #  nbml      nombre de mots dans la ligne
 #  nbneg     nombre de valeurs négatives
 #  pbdi      en normal, 1 si problème de diagonale
 #  pbtr      en normal, 1 si problème de triangularité


** Liste des variables de fichier (par ordre alphabétique)

Variable                              Ligne de première     Nombre de
                                         utilisation        références
 #  fpmdd                                    10                 4
```

```
*** Liste des tableaux (par ordre alphabétique)

Tableau                 Dims.        Ligne de première     Nombre de
                                        utilisation        références
 #  tabDist                2                19                 6

**** Liste des modules (par ordre alphabétique)

Module                  Arité        Ligne de première     Nombre de
                                        utilisation        références
 #  fdf                    1                14                 1
 #  mot                    2                19                 2
 #  mots                   1                17                 1
 #  testeTriang            2                80                 1
```

On peut donc écrire notre algorithme en admettant que les fonctions *mot*, *mots*, fdf existent (fdf signifie bien sûr "fin de fichier" soit encore eof en anglais comme dans beaucoup de langages).

Par contre il faut écrire explicitement le sous-programme $testeTriang$, ce que nous ferons après l'algorithme principal que voici :

```
#####################################################################
#                                                                   #
# testedist.alg : teste si le fichier peut être                     #
#                 considéré comme une matrice triangulaire          #
#                 inféreure de distances ; auteur (gH)              #
#                                                                   #
#####################################################################
#                                                                   #
#  attention : on ne teste pas la symétrie car elle                 #
#              est implicite, vu que le fichier contient            #
#              la partie sous-diagonale au sens large de            #
#              la matrice.                                          #
#                                                                   #

écrire " Etude du fichier "  , ficmatd
```

```
   # lecture : on transfère les données dans le tableau tabDist
   #           on teste la positivité au passage

   ouvrir ficmatd en_lecture comme fpmdd
   affecter nblig <-- 0 # nombre de lignes
   affecter nbneg <-- 0 # nombre de valeurs négatives

   tant_que non fdf( fpmdd)
        affecter nblig <-- nblig + 1
        lire lig sur fpmdd
        affecter nbml <-- mots(lig)
        pour indm de1a nbml
             affecter tabDist[ nblig, indm  ] <-- mot(lig,indm)
             # on force la symétrie
             affecter tabDist[ indm,  nblig ] <-- mot(lig,indm)
             si tabDist[ nblig, indm ] < 0
                alors
                  affecter nbneg <-- nbneg + 1
                  affecter ilneg <-- nblig
                  affecter icneg <-- indm
             fin_si # tabDist[ nblig, indm ] < 0
        fin_pour # indm de1a nbml
   fin_tant_que # non fdf( fpmdd)
   fermer fpmdd

   # affichage éventuel en cas de valeurs négatives

   si nbneg > 0
      alors
écrire ""
écrire "L'élément en ligne " , ilneg , " colonne " , icneg , " est négatif"
écrire "car il vaut " , tabDist[ilneg,icneg] , " ; votre fichier " , ficmatd
écrire "ne peut donc PAS être considéré comme le fichier d'une matrice "
écrire "triangulaire inférieure."
écrire ""
écrire " Attention : il y a en tout " , nbneg ," élements négatifs."
écrire ""
        quitter # le programme !
```

```
   sinon
        écrire " votre matrice est positive "
   fin_si # nbneg > 0

   # test de dissimilarité

   affecter ilig  <-- 1
   affecter pbdi  <-- 0
   tant_que ilig <= nblig
        si non (tabDist[i,i]=0)
           alors
écrire "L'élément sur la diagonale en position " , ilig , " est non nul"
écrire "puisqu'il vaut " , tabDist[ilig,ilig] , " ; votre fichier " , ficmatd
écrire "ne peut donc PAS être considéré comme le fichier d'une matrice "
écrire "triangulaire inférieure de distances."
              affecter  pbdi  <-- pbdi + 1
              affecter  ilig  <-- nblig + 1
        fin_si # non (tabDist[i,i]=0)
        affecter  ilig  <-- ilig + 1
   fin_tant_que # ilig <= nblig

   si pbdi=0
      alors
        écrire " votre matrice vérifie le critère de dissimilarité."
   fin_si # pbdi=0

   # test de triangularité

   affecter ilig  <-- 1
   affecter pbtr  <-- 0
   tant_que ilig <= nblig
        affecter  jcol  <-- 1
        tant_que  jcol  <= nblig
           appeler testeTriang(ilig,jcol)
           affecter  jcol  <-- jcol + 1
        fin_tant_que # jcol  <= nblig
        affecter  ilig  <-- ilig + 1
   fin_tant_que # ilig <= nblig
```

```
si pbtr=0
   alors
     écrire " votre matrice vérifie le critère de triangularité."
fin_si # pbtr=0

# -- fin de l'algorithme testedist.alg
```

Passons donc maintenant au sous-programme *testeTriang* qui doit comparer les valeurs `tabDist[i,j]` et `tabDist[i,k] + tabDist[k,j]`. Avec la même démarche que précédemment, nous allons utiliser une boucle `tant_que` utilisant l'indice k (renommé en `indk`). Le passage des paramètres est délicat vu qu'en cas de problème nous voulons intervenir sur les indices de boucle *ilig* et *jcol* : c'est donc un passage de paramètre par adresse qu'il vaut effectuer.

Visiblement, nous aurons besoin d'afficher les valeurs incriminées et donc comme nous n'avons pas passé le nom du tableau en paramètre, il faudra que celui-ci soit global pour notre sous-programme. Ensuite, si on nomme *valx* la valeur `tabDist[i,j]`, *valy* la valeur `tabDist[i,k]` et *valz* la valeur `tabDist[k,j]` il faut déclarer ces variables comme locales.

Enfin, si l'on veut pouvoir utiliser la variable *pbtr* dans l'algorithme principal, il faut aussi supposer qu'elle est globale (une autre solution aurait pu être de faire de *testeTriang* une fonction et d'incrémenter *pbtr* avec la valeur de retour de *testeTriang*).

D'où la déclaration de notre sous-programme

```
Module testeTriang

####  ce sous-programme est appelé par testeDist

 Paramètres

   iilig      # indice de ligne   dans le programme appelant
   jjcol      # indice de colonne dans le programme appelant
```

```
    Globales
      nblig      # nombre de lignes dans tabDist
      pbtr       # indicateur de problème pour triangularité
      tabDist    # tableau-matrice
    Locales
      indk       # indice de boucle
      valx       # correspond à tabDist[iindi,jjcol]
                 # soit encore d(i,j)
      valy       # correspond à tabDist[iindi,indk]
                 # soit encore d(i,k)
      valz       # correspond à tabDist[indk,jjcol]
                 # soit encore d(k,j)
    Début de Module
      affecter indk <-- 1
      tant_que indk <= nblig
            affecter valx <--  tabdist[iicol,jjcol]
            affecter valy <--  tabdist[iicol,indk ]
            affecter valz <--  tabdist[indk, jjcol]
            if (valx>valy+valz) {
écrire "Les éléments x = " x " ligne " iilig " colonne " jjcol
écrire "             y = " y " ligne " iilig " colonne " indk
écrire "             z = " z " ligne " indk  " colonne " jjcol
écrire "ne vérifient pas l'inégalité triangulaire puisque "
écrire "" x " > " y+z " = " y " + " z " soit x > y + z "
écrire "c'est à dire d(i,j) > d(i,k) + d(k,j) "
écrire "pour i = " iilig", j = " jjcol " et k " indk"."
écrire "Votre fichier " ficmatd
écrire "ne peut donc PAS être considéré comme le fichier d'une matrice "
écrire "triangulaire inférieure de distances."
               affecter iilig <-- nblig + 1
               affecter jjcol <-- nblig + 1
               affecter indk  <-- nblig + 1
            } # fin de si sur (x>y+z)
         affecter indk <-- indk + 1
      fin_tant_que # indk <= nblig
    Fin de Module
```

T.D. 6 : Analyse de variables statistiques

1. Types de variables
2. Sont-ce des QT ou des QL ?
3. Calculs concrets de QT : temps de transit
4. Calculs concrets de QL : bande passante
5. Valeurs de a et b pour $|\rho| = 1$
6. Analyse Statistique Générale du dossier VINS
7. Formules de moyennes et de variance pondérées
8. Analyse Statistique Générale du dossier ELF

1. Types de variables

Enoncé

Dans le cours, on utilise principalement les variables QT (quantitatives) et QL (qualitatives). Peut-il y avoir d'autres types de variables ? On pourra par exemple imaginer que les variables correspondent à des questions pour un questionnaire de type enquête, ou que les variables sont des mesures issues de capteurs...

Solution

Il y a bien sûr de nombreux autres types de variable. Citons

les **QM** ou *variables multi-réponses*,
les **QP** ou *variables pourcentages*,
les **QS** ou *variables (anti-)scores*,
les **QF** ou *variables floues*,
les **QH** ou *variables hiérarchiques* (classement),
les **QE** ou *variables d'énonciation.*
les **QX** ou *variables textuelles.*

Pour la plupart de ces variables, les calculs se ramènent à des calculs de comptages comme pour les **QL** ou des sommations pondérées comme pour les **QT**. On laisse au lecteur le soin de vérifier ces dernières affirmations sur des exemples concrets et d'essayer de trouver quel affichage "intelligent" doit alors être utilisé c'est à dire avec quelle présentation on fera au mieux ressortir les variables les plus flagrantes, quel ordre d'affichage sera le mieux à même de mettre en évidence ce qui est "fort" de ce qui est "faible" en termes de résultats.

Rappelons au passage que la plupart des logiciels se contentent d'afficher par ordre de variable, ce qui n'est pas vraiment intelligent, car la plupart des logiciels disposent de fonctions de tri des résultats, encore faut-il penser à savoir les utiliser et à savoir selon quels critères trier.

2. Sont-ce des QT ou des QL ?

Enoncé

Un de nos étudiants doit traiter une variable IMC (indice de masse corporelle) définie par le rapport poids en kg sur taille au carré en m^2.

Est-ce une QT ou une QL ?

Une de nos étudiantes doit traiter une variable $DENSP$ (densité de population) définie par le rapport population en millions d'habitants sur superficie en km^2.

Est-ce une QT ou une QL ?

Enfin, un autre groupe doit traiter un taux d'alphabétisation de pays défini par le rapport nombre d'enfants scolarisés sur nombre d'enfants en tout pour des enfants dont l'age se situe entre 3 et 10 ans.

Est-ce là encore une QT ou une QL ?

Que peut-on en conclure sur l'utilisation de la fonction `mean` sous *Rstat* et la fonction `MOYENNE` sous *Excel* ?

Solution

Aucune de ces variables n'est une QL car leurs valeurs sont des valeurs numériques souvent non entières : ce ne peuvent donc être des codes.

Aucune de ces variables n'est une QT car aucune n'est sommable. Il suffit de prendre un contre-exemple pour s'en rendre compte.

Ainsi pour une population $p_1 = 100$ habitants et une surface $s_1 = 100$ km^2 (quel désert !) la densité d_1 est 1 h/km^2 . Pour un deuxième pays dont la population p_2 est de 100 habitants pour une surface $s_2 = 100$ km^2 on trouve une densité $d_2 = 1$ h/km^2. La densité correspondant à la réunion des pays est $d_3 = (p_1 + p_2)/(s_1 + s_2)$ redonne 1 h/km^2 alors que la somme $d_1 + d_2$ donne 2 h/km^2 : $DENSP$ n'est donc pas sommable.

Il faut donc se méfier des logiciels car faire calculer $d_1 + d_2$ à la machine ou cliquer sur le bouton "Moyenne" ne s'applique pas tout le temps !

3. Calculs concrets de QT : temps de transit

Enoncé

Soit T la variable quantitative "*temps de transit*" exprimée en minutes dont les valeurs sont, dans l'ordre, $[97, 12, 192, 25, 48]$.

Soit maintenant D la variable quantitative "*durée de transport*" exprimée en heures et dont les valeurs sont, dans l'ordre $[16, 2, 32, 4, 8]$.

Effectuez l'analyse séparée puis conjointe des ces deux variables. On présentera les résultats suivant un ordre "intelligent". Calculer aussi le coefficient de corrélation et, si besoin est, les coefficients a et b de la relation linéaire correspondante à savoir T=a.D+b.

Pour ceux et celles qui ont oublié leur calculette, on fournit les résultats numériques suivants :

somme des valeurs de D	62
somme des carrés des valeurs de D	1364
somme des valeurs de T	374
somme des carrés des valeurs de T	49346
somme des produits terme à terme D × T	8204

Solution

On en déduit

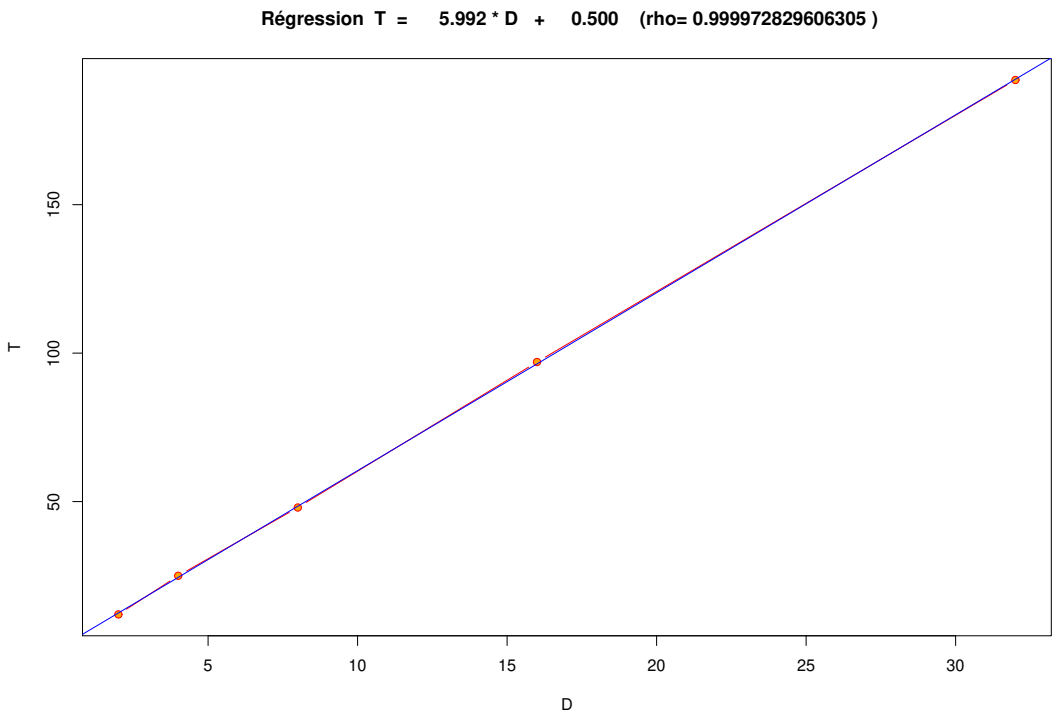
moyenne(D)	=	62/5	=	12.4
moyenne(D^2)	=	1364/5	=	272.8
variance(D)	=	272.8 - 12.4*12.4	=	119.04
écart-type(D)	=	$\sqrt{119.04}$	=	10.9105453575

Tous calculs numériques effectués, on obtient les résultats suivants (une fois D convertie en minutes, soit les valeurs $[960, 120, 1920, 240, 480]$) :

Variable	Moyenne	Ecart-type	σ/m
D	744.0	654.6	87 %
T	74.8	65.4	87 %

D'après la formule du cours la covariance est 8204/5-12.4*74.8 soit 713.28 donc le coefficient de corrélation linéaire ρ vaut 713.28/(10.91*65.4) soit ici 0.99967 c'est à dire pratiquement 1 ce qui indique une liaison linéaire.

Ce résultat est prévisible si l'on trace la courbe des points $(\mathsf{D}_i, \mathsf{T}_i)$:



Utilisant les formules $a = \rho.\sigma_Y/\sigma_X$ et $b = m_y - a.m_X$ pour la liaison linéaire $Y = a.X + b$ avec $Y = \mathsf{D}$ et $X = \mathsf{T}$, on obtient

$$\mathsf{D} = 10,01.\mathsf{T} - 4,75$$

alors que pour $X = \mathsf{D}$ et $X = \mathsf{T}$ on obtient l'équation

$$\mathsf{T} = 0.100 * \mathsf{D} + 0.500$$

ce qui est sans doute plus conforme à la réalité (sous-entendu via la causalité entre durée de transport et temps de transit).

4. Calculs concrets de QL : bande passante

Enoncé

Soient $[10, 12, 17, 12, 10, 17, 10]$ les valeurs (codées) de la variable qualitative B "*Bande Passante*" où 10 correspond à la gamme FM, 12 à la gamme UHF et 17 à la gamme VHF. On considère également la variable R "*Type de Radio*" dont les valeurs codées sont $[1, 1, 1, 2, 2, 2, 2]$. Le code 1 signifie "Radio de qualité moyenne" et le code 2 "Radio de qualité supérieure". Effectuez l'analyse séparée puis conjointe des ces deux variables. On présentera les résultats suivant un ordre "intelligent". Peut-on parler de liaison entre B et R ?

Solution

Le comptage à la main des différents croisements possibles donne le tri croisé suivant (effectifs absolus) :

	FM	UHF	VHF	Total
R q.moyenne	1	1	1	3
R q.superieure	2	1	1	4
Total	3	2	2	7

Puisque 1/7 est à peu-près 14.3 %, on peut donc présenter les variables suivant le classement suivant

Variable	Mode	%	Autres modalites
R	q. moyenne	57	q. superieure (43 %)
B	bande FM	43	UHF (29 %) VHF (29 %)

Compte-tenu des faibles effectifs (7 valeurs en tout), il est difficile de tirer des conclusions. Enfin, la notion de liaison entre QL n'a pas été définie en cours. On ne sait donc pas comment la calculer.

5. Valeurs de a et b pour $|\rho|$ =1

Enoncé

Dans le cours, on affirme que si $|\rho(X,Y)| = 1$ alors X et Y sont liés par la relation linéaire

$$Y = a.X + b$$

avec

$$\begin{aligned} a &= \frac{m(X).m(Y) - m(XY)}{d} \\ &= \rho(X,Y).\sigma(Y)/\sigma(X) \\ b &= \frac{m(X).m(XY) - m(Y)m(X^2)}{d} \\ &= m(Y) - a.m(X) \end{aligned}$$

où $d = m(X)^2 - mt(X,2)$.

Démontrez ces formules.

Solution

Si $Y = a.X + b$ cela signifie que pour tout i, $Y_i = a.X_i + b$. Appliquant l'opérateur moyenne à ces relations, c'est à dire en sommant et en divisant par n, on obtient l'équation

$$\mathsf{E} \ : moy(Y) = a.moy(X) + b$$

De même, si $Y_i = a.X_i + b$ pour tout i, en multipliant par X_i puis en sommant et en divisant par n, on obtient l'équation

$$\mathsf{F} \ : moy(XY) = a.moy(X^2) + b.moy(X)$$

X et Y étant donnés, $\{\mathsf{E},\mathsf{F}\}$ est un système de deux équations linéaires en les deux inconnues a et b qu'on réécrit en

$$\begin{array}{lrcrcl} \text{(E)} : & moy(X).\text{a} & + & \text{b} & = & \text{moy(Y)} \\ \text{(D)} : & moy(X^2).\text{a} & + & \text{moy(X).b} & = & \text{moy(XY)} \end{array}$$

La résolution de ce système donne les premières valeurs indiquées avec

$$d = m(X)^2 - m(X^2)$$

qui se nomme le discriminant du système.

Si l'on écrit explicitement la définition de la covariance et si on écrit les variances comme les carrés des écart-types :

$$\begin{array}{lcl} cov(X,Y) & = & m(XY) - m(X).m(Y) \\ V(Y) & = & \sigma(Y)^2 \\ V(X) & = & \sigma(X)^2 \\ d & = & -V(X) \end{array}$$

Alors, en remplaçant ρ par sa définition à l'aide de la covariance et des écart-types :

$$\rho(X,Y) = cov(X,Y)/\sigma(Y)/\sigma(X)$$

on voit que a se calcule par

$$a = \frac{-cov(X,Y)}{-\sigma(X)^2}$$

soit encore en multipliant au numérateur et au dénominateur par $\sigma(Y)$:

$$a = \rho(X,Y).\sigma(Y)/\sigma(X)$$

et b se déduit alors directement de l'équation E d'où la valeur annoncée

$$b = m(Y) - a.m(X)$$

6. Analyse Statistique Générale du dossier VINS

Enoncé

La Direction Générale des Impots publie régulièrement au Journal Officiel une Statistique Mensuelle des Vins. Le J.O. du 4 novembre 1987 fournit en particulier le tableau de données suivant où sont croisées des catégories de vins avec des pays exportateurs. L'unité commune est l'hectolitre. Les sigles se veulent explicites ; ainsi BOJO signifie Beaujolais, ANJO est mis pour Anjou...

ID	BELG	NEDE	RFA	ITAL	UK	SUIS	USA	CANA
CHMP	7069	3786	12578	8037	13556	9664	10386	206
MOS1	2436	586	2006	30	1217	471	997	51
MOS2	3066	290	10439	1413	7214	112	3788	330
ALSA	2422	1999	17183	57	1127	600	408	241
GIRO	22986	22183	21023	56	30025	6544	13114	3447
BOJO	17465	19840	72977	2364	39919	17327	17487	2346
BORG	3784	2339	4828	98	7885	3191	11791	1188
RHON	7950	10537	7552	24	8172	11691	1369	1798
ANJO	2587	600	2101	0	7582	143	872	131
AOCX	17200	22806	15979	50	20004	1279	4016	944
VDQS	1976	1029	1346	0	2258	212	1017	487
XXXX	38747	19151	191140	7992	101108	1029	26192	38503
PROV	1375	1150	2514	0	284	401	9	236
MUSC	2016	2908	1529	0	12891	18	716	653
RHOF	785	1648	1009	6	775	643	542	35
AOCF	160	246	135	8	1177	26	7	0
XXXF	24	1533	160	0	480	0	0	0
XXFF	2415	74	208	8	1705	12	36	47

Analyser ces variables quantitatives (analyse conjointe et séparée). On présentera les résultats comme convenu pour essayer de comprendre comment le vin français s'exporte aujourd'hui. Pour faciliter les calculs, on fournit diverses sommes et valeurs calculées par ordinateur.

Toutefois, suite à des gesticulations intempestives de *pokemon* (!) certaines valeurs ont disparu et sont remplacées par un ou plusieurs symboles ? et il faut donc retrouver ces valeurs.

Sommes de valeurs

BELG	NEDE	RFA	ITAL	UK	SUIS	USA	CANA
134463	112705	364707	20143	257379	53363	92747	50643

UK*RFA	CANA*RFA	UK*BELG	UK*UK
23597981730	7650378991	5922865383	13719495029

RFA*RFA	CANA*CANA	ITAL*ITAL
43220226841	1506353705	136070627

Résultats partiels

		moyenne	ecart-type	cdv	min	max
2	BELG	7470.167	9993.916	134	24	38747
9	CANA	2813.500	8704.627	???	0	38503
5	ITAL	1119.056	????.???	224	0	8037
3	NEDE	6261.389	8227.603	131	74	22806
4	RFA	20261.50	44616.09	220	135	191140
7	SUIS	2964.611	4882.127	165	0	17327
6	UK	14298.83	23616.47	165	284	101108
8	USA	5152.611	7336.798	142	0	26192

	BELG	NEDE	RFA	ITAL	UK	SUIS	USA
BELG	1.0000						
NEDE	0.8702	1.0000					
RFA	0.8692	0.5818	1.0000				
ITAL	0.5856	0.2895	0.6998	1.0000			
UK	0.9416	0.6997	0.?693	0.6906	1.0000		
SUIS	0.3353	0.5177	0.1984	0.3098	0.2462	1.0000	
USA	0.8699	0.6799	0.8477	0.7172	0.8935	0.4681	1.0000
CANA	0.8143	0.4582	0.9476	0.6585	0.9256	-0.?246	0.7469

Rho max : 0.?693 UK / RFA
Liaison : UK = 0.?13 * RFA + ?903.?87

Solution

Le premier résultat qui manque est le coefficient de variation ou *cdv* du Canada. Pour l'obtenir, il suffit de prendre le pourcentage associé au rapport σ/m. Comme 8704.627/2813.500 vaut 3.09387844 ce *cdv* est donc 309 % si on utilise le même affichage entier que les autres cdv. Il manque ensuite l'écart-type de l'italie. Pour le calculer, on calcule d'abord la variance, soit "*la moyenne des carrés moins le carré de la moyenne*" ce qui s'écrit ici, puisque nous avons les sommes correspondantes pour les 18 catégories de vins :

$$V = \frac{136070627}{18} - \left(\frac{20143}{18}\right)^2$$

On trouve donc 6307193.93 dont la racine nous donne l'écart-type demandé soit 2511.413 pour rester cohérent avec les autres affichages d'écart-type. Il aurait aussi été possible d'utiliser le coefficient de variation pour obtenir 224×1119.056 soit 250668.544 mais c'est un peu moins précis car le *cdv* est affiché avec un arrondi entier.

L'affichage "intelligent" des résultats doit alors reprendre ces valeurs en un premier tri par moyenne décroissante puisque toutes les variables traitent d'un même phénomène (l'importation) décrit dans la même unité (l'hectolitre) de façon à discuter l'effet de taille, soit les valeurs :

Champ	Nom	Moyenne	Ecart-Type	Cdv	Min	Max
4	RFA	20261.50	44616.09	220	135	191140
6	UK	14298.83	23616.47	165	284	101108
2	BELG	7470.167	9993.916	134	24	38747
3	NEDE	6261.389	8227.603	131	74	22806
8	USA	5152.611	7336.798	142	0	26192
7	SUIS	2964.611	4882.127	165	0	17327
9	CANA	2813.500	8704.627	309	0	38503
5	ITAL	1119.056	2511.413	224	0	8037

Un deuxième affichage doit utiliser non pas l'ordre alphabétique des pays comme proposé mais un tri par coefficient de variation décroissant pour montrer à l'intérieur d'un même pays la fluctuation de la quantité de vin importé par catégorie de vin.

D'oùt le deuxième tableau de valeurs :

Champ	Nom	Moyenne	Ecart-Type	Cdv	Min	Max
9	CANA	2813.500	8704.627	309	0	38503
5	ITAL	1119.056	2511.413	224	0	8037
4	RFA	20261.50	44616.09	220	135	191140
6	UK	14298.83	23616.47	165	284	101108
7	SUIS	2964.611	4882.127	165	0	17327
8	USA	5152.611	7336.798	142	0	26192
2	BELG	7470.167	9993.916	134	24	38747
3	NEDE	6261.389	8227.603	131	74	22806

Pour rendre le premier affichage plus lisible, changeons d'unité : 10 000 hectolitres font un million de litre (Ml) donc on peut afficher les moyennes d'importation comme suit :

Pays	Moyenne en Ml
RFA	2.0
UK	1.4
BELG	0.8
NEDE	0.6
USA	0.5
SUIS	0.3
CANA	0.3
ITAL	0.1

On met alors tout de suite en évidence un *très gros importateur* RFA[†], un gros importateur UK et un *très faible importateur* ITALIE.

[†] pour les jeunes qui lisent ce texte, il s'agit d'une partie ce qu'on nomme aujourd'hui l'Allemagne.

Un affichage simplifié du deuxième tableau, à savoir

Pays	Cdv en %
CANA	309
ITAL	224
RFA	220
UK	165
SUIS	165
USA	142
BELG	134
NEDE	131

permet de déduire que CANADA, ITALIE et RFA sont des importateurs très *sélectifs*. Pour savoir quel genre de fortes variations se cache derrière ce résumé, il faut trier chacune des colonnes correspondantes, soit les 3 tableaux indépendants où nous faisons figurer les plus forts pourcentages par rappoprt au total de la colonne

	CANA	%/T
XXXX	38503	76
GIRO	3447	7
BOJO	2346	5
RHON	1798	4
BORG	1188	2
AOCX	944	2
MUSC	653	
VDQS	487	
MOS2	330	
ALSA	241	
PROV	236	
CHMP	206	
ANJO	131	
MOS1	51	
XXFF	47	
RHOF	35	
AOCF	0	
XXXF	0	

	ITAL	%/T
CHMP	8037	40
XXXX	7992	40
BOJO	2364	12
MOS2	1413	7
BORG	98	
ALSA	57	
GIRO	56	
AOCX	50	
MOS1	30	
RHON	24	
AOCF	8	
XXFF	8	
RHOF	6	
ANJO	0	
VDQS	0	
PROV	0	
MUSC	0	
XXXF	0	

	RFA	%/T
XXXX	191140	52
BOJO	72977	20
GIRO	21023	6
ALSA	17183	5
AOCX	15979	4
CHMP	12578	3
MOS2	10439	3
RHON	7552	
BORG	4828	
PROV	2514	
ANJO	2101	
MOS1	2006	
MUSC	1529	
VDQS	1346	
RHOF	1009	
XXFF	208	
XXXF	160	
AOCF	135	

Le Canada est très fortement sélectif puisqu'il importe presque exclusivement des vins de type XXXX. L'Italie est aussi très sélective puisqu'elle importe principalement des vins de type CHMP ou XXXX. Enfin, la RFA est le troisième pays le plus sélectif avec un import massif de vins XXXX puisque la quantité correspond à 18 % du total général importé, tous pays confondus.

Pour apprécier l'influence relative de chaque ligne (vin) et de chaque colonne (pays), une technique possible est de calculer les marges du tableau de données, c'est à dire les pourcentages de chaque ligne, chaque colonne par rapport au total général. On trouve alors :

Vin	%/T
CHMP	6
MOS1	1
MOS2	2
ALSA	2
GIRO	11
BOJO	17
BORG	3
RHON	5
ANJO	1
AOCX	8
VDQS	1
XXXX	39
PROV	1
MUSC	2
RHOF	1
AOCF	0
XXXF	0
XXFF	0

Pays	%/T
BELGIQUE	12
NEDERLAND	10
RFA	34
ITALIE	2
UK	24
SUISSE	5
USA	9
CANADA	5

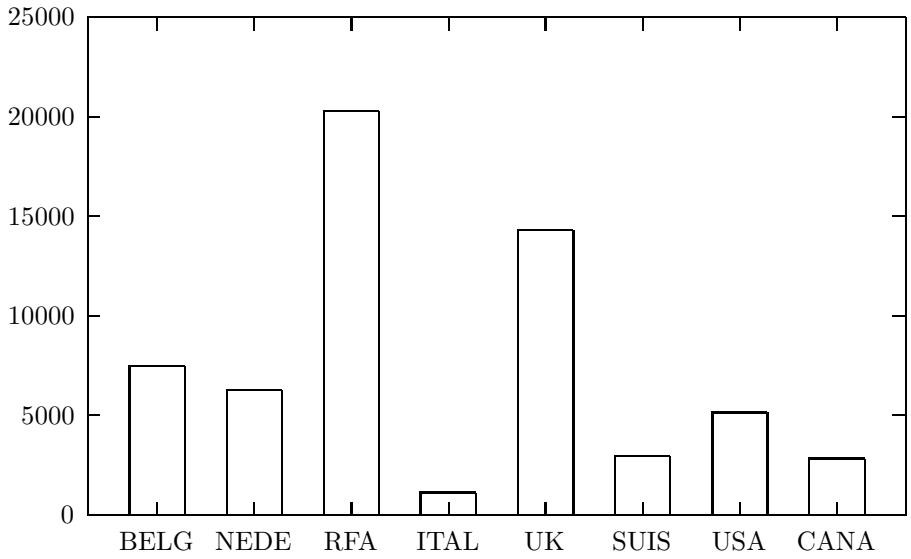
On y voit bien l'importance de RFA et de UK, la domination du vin de type XXXX... car les pourcentages indiqués sont bien sur proportionnels aux moyennes. Ainsi le 34 de la RFA correspond au pourcentage du rapport 364707 (total de la colonne RFA) sur 1086150 (total général) alors que la moyenne de RFA est 364707 / 18 soit 20261,5...

Une autre technique pour dégager à l'oeil nu les "grandes" valeurs est d'arrondir les valeurs. Par exemple voici le tableau des données original, divisé par 1000 et avec un arrondi entier pour lequel la valeur 0 est remplacée par un point :

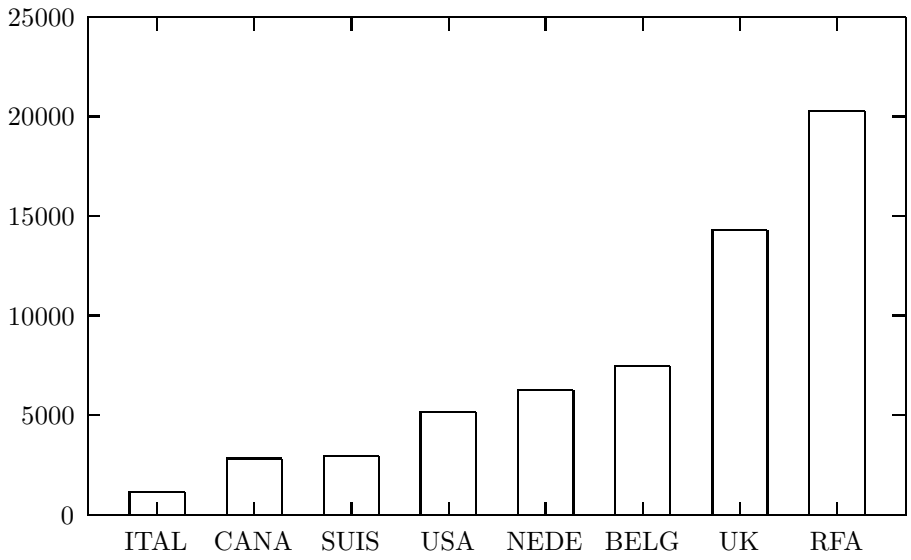
VIN	BELG	NEDE	RFA	ITAL	UK	SUIS	USA	CANA
CHMP	7	4	13	8	14	10	10	.
MOS1	2	1	2	.	1	.	1	.
MOS2	3	.	10	1	7	.	4	.
ALSA	2	2	17	.	1	1	.	.
GIRO	23	22	21	.	30	7	13	3
BOJO	17	20	73	2	40	17	17	2
BORG	4	2	5	.	8	3	12	1
RHON	8	11	8	.	8	12	1	2
ANJO	3	1	2	.	8	.	1	.
AOCX	17	23	16	.	20	1	4	1
VDQS	2	1	1	.	2	.	1	.
XXXX	39	19	191	8	101	1	26	39
PROV	1	1	3	.	.	.	.	.
MUSC	2	3	2	.	13	.	1	1
RHOF	1	2	1	.	1	1	1	.
AOCF	.	.	.	.	1	.	.	.
XXXF	.	2	.	.	.	.	.	.
XXFF	2	.	.	.	2	.	.	.

On y voit bien que seuls deux nombres sont très grands, à savoir 191 et 101 respectivement pour `RFA` et `UK` dans l'importation de vin `XXXX`, on "voit" que la colonne `ITAL` ne contient que des faibles valeurs etc.

Des histogrammes des totaux en lignes, en colonnes seraient certainement tout aussi explicites. Ainsi pour le graphique des moyennes d'importation par pays, on a le choix entre utiliser l'ordre historique des pays



et l'ordre croissant de moyennes



Il reste deux chiffres à trouver pour compléter l'analyse bivariée des pays. Si l'on croit l'indication "Rho max" alors le chiffre manquant est un 9 et le plus fort coefficient de corrélation linéaire est `0.9693` car ce nombre présenté comme `0.?693` doit être supérieur au coefficient `0.9416` pour `UK` et `BELG`.

De façon plus rigoureuse,

$$\rho(UK, RFA) = \frac{cov(UK, RFA)}{\sigma(UK)\ \sigma(RFA)}$$

La covariance étant "*la moyenne du produit moins le produit des moyennes*", $cov(UK, RFA)$ se calcule ici par (23597981730/18)-(364707/18) $\times$ (257379/18) soit 1021283173.42 et ρ vaut donc 1021283173.42 / (44616.09 $\times$ 23616.47) soit 0.969258650549 ce qui est bien la valeur arrondie 0.9693 annoncée.

On pourra vérifier que l'autre coefficient manquant, pour `CANA` et `SUIS` est -0.0246.

Si on s'intéresse maintenant à la plus forte liaison linéaire qui relie `RFA` et `UK`, il faut choisir le sens de l'équation. `RFA` jouant un role particulier puisque c'est le plus fort importateur, nous écrirons donc l'équation sous la forme

$$UK = a.RFA + b$$

On sait que dans le modèle $Y = aX + b$ a vaut $\rho(X, Y)\ \sigma(Y)/\sigma(X)$ et que b vaut $m(Y) - am(X)$ on trouve donc a = 0.9693 * 23616.47 / 44616.09 soit environ `0.513` et b = 14298.83 - 0.513 * 20261.50 soit à peu près `3905`. La relation linéaire est donc $UK = 0.513 \times RFA + 3905$ qu'on interprétera comme "*l'importation de UK se correspond en gros à la moitié de l'importation de RFA augmentée de 4000 hectolitres*".

Il y a d'autres très forts coefficients de corrélation linéaire dont voici la liste

```
0.969 UK    RFA     (rappel)
0.948 CANA  RFA
0.942 UK    BELG
0.926 CANA  UK
```

Les coefficients de corrélation linéaire suivants sont aussi également élevés puisque nous trouvons

```
0.894 USA   UK
0.870 NEDE  BELG
0.870 USA   BELG
0.869 RFA   BELG
0.848 USA   RFA
0.814 CANA  BELG
```

Nous laissons le soin au lecteur de trouver les relations linéaires correspondantes.

Reste à critiquer les données et à remarquer qu'on ne peut rien en conclure quant à la question e*comment le vin français s'exporte aujourd'hui ?*".

7. Formules de moyenne et de variance pondérées

Enoncé

Rappeler les formules de la moyenne et de la variance pour n valeurs x_i où i varie de 1 à n.

Quelle est la meilleure façon d'écrire ces formules si on veut les utiliser sachant que les x_i sont regroupés en r_j fois x_j^* ?

On répartit maintenant les x_i du départ en deux groupes I_1 et I_2. On désigne par n_j pour j de 1 à 2 le nombre d'élement dans I_j. On note de façon "évidente" les sommes des valeurs S_j, les moyennes m_j, les sommes des carrés C_j, les variances V_j et les écart-types σ_j pour les deux groupes.

Quelle sont les les formules qui relient n aux n_i, m aux m_i etc. ?

On fera apparaitre explicitement les rapports $\dfrac{n_i}{n}$.

Généraliser aux cas où I est décomposé en p groupes I_j pour j de 1 à p.

Solution

Si on réécrit les formules du cours en

$$\sum_{i=1}^{n} x_i = n.m \quad \text{et} \quad \sum_{i=1}^{n} x_i^2 = n(V + m^2)$$

alors la généralisation pour les p valeurs distinctes x_j^* s'écrit

$$\sum_{j=1}^{p} r_j x_j^* = n.m \quad \text{et} \quad \sum_{j=1}^{p} r_j (x_j^*)^2 = n(V + m^2)$$

Pour deux groupes I_1 et I_2 la décomposition disjointe $I = I_1 \sqcup I_2$ permet d'écrire

$$S = \sum_{i \in I} x_i = \sum_{i \in I_1 \sqcup I_2} x_i = \sum_{i_1 \in I_1} x_{i_1} + \sum_{i_2 \in I_2\} \sqcup I_2} x_{i_2} = S_1 + S_2$$

On en déduit $n.m = n_1 m_1 + n_2 m_2$ soit encore :

$$m = \frac{n_1}{n}\, m_1 + \frac{n_2}{n}\, m_2$$

De même, la décomposition de $\sum_{i \in I} x_i^2$ permet d'écrire $C = C_1 + C_2$ soit encore

$$n(V + m^2) = n_1(V_1 + m_1^2) + n_2(V_2 + m_2^2)$$

on en déduit

$$V = \frac{n_1}{n}\, V_1 + \frac{n_2}{n}\, V_2 - \left(\frac{n_1}{n}\, m_1^2 + \frac{n_2}{n}\, m_2^2\right) - \frac{1}{n}\, m^2$$

Généralisation :

$$m = \sum_{j=1}^{p} \frac{n_j}{n}\, m_j \quad \text{et} \quad V = \sum_{j=1}^{p} \frac{n_j}{n}\, V_j - \left(\sum_{j=1}^{p} \frac{n_j}{n}\, m_j^2\right) - \frac{1}{n}\, m^2$$

T.D. 7 : ASG de **QT** et de **QL**

1. Analyse Statistique Générale du dossier ELF
2. Analyse Statistique Générale du dossier CHIENS
3. Analyse Statistique Générale du dossier BILAN
4. Tige et feuille à la main
5. Transitivité de la relation Y = aX + b

1. Analyse Statistique Générale du dossier ELF

Enoncé

Dans le cadre d'une enquête linguistique sur la féminisation des noms de métiers le Ministère des Droits de la Femme a établi un questionnaire comprenant un signalétique de 7 variables et 26 questions.

Nous reproduisons ici le codage des 4 variables qualitatives du signalétique et un extrait des données.

Codage

	SEXE (2)	ETUD (5)	REGIONALITE (5)	Us. LANGUE (4)
0	homme	nr	nr	nr
1	femme	primaire	faible	peu fréquent
2		bepc	moyenne	commun
3		bac	forte	très particulier
4		supérieur	très forte	

Extrait des données

Ligne	NUM	SEXE	AGE	PROF	ETUD	REGI	USAG
1	M001	1	62	1	2	2	3
2	M002	0	60	9	3	4	1
3	M003	1	31	9	4	4	1
4	M004	1	27	8	4	1	1
5	M005	0	22	8	4	1	2
...							
96	M096	1	17	12	3	1	0
97	M097	1	39	1	2	1	0
98	M098	0	62	6	3	1	0
99	M100	1	48	9	4	2	0

Les résutats informatiques sont alors

```
TRIS A PLAT

SEXE        0       35 ( 35.35 %)        1       ?? ( ??.?? %)
ETUD        0        3 (  3.03 %)        1        6 (  6.06 %)
            2       30 ( 30.30 %)        3       21 ( 21.21 %)
            4       39 ( 39.39 %)
REGI        0        2 (  2.02 %)        1       35 ( 35.35 %)
            2       14 ( 14.14 %)        3        5 (  5.05 %)
            4       43 ( 43.43 %)
USAG        0       66 ( 66.67 %)        1       18 ( 18.18 %)
            2       13 ( 13.13 %)        3        2 (  2.02 %)

ANALYSE DE L'AGE PAR SEXE
                TOUT        HOM         FEM
  Age   Moy     35.828      36.400      ??.516
        Ect     17.464      ??.650      17.886
        Cdv     48.7        45.7        ??.4
```

Retrouvez les chiffres qui manquent (remplacés par `?`), présentez les tris à plat comme convenu et essayez de commenter tous les résultats.

Solution

La ligne

```
   USAG        0       66 ( 66.67 %)        1       18 ( 18.18 %)
```

doit se comprendre ainsi : "pour la variable `USAG`, on a 66 fois le code 0 soit un pourcentage de 66.67 % et 18 fois le code 1 soit 18.18 %". Il est clair que le premier nombre qu'il manque est celui du nombre de codes 1 pour la variable `SEXE`. Grâce à l'extrait des données et au total des autres modalités, on trouve qu'il y a 99 personnes en tout dans l'enquête. Puisque le codage de la `SEXE` ne met pas en jeu de non-réponses, s'il y a 35 hommes (code 0) il y a donc 99-35 soit 64 femmes (code 0) d'où un pourcentage de 64.655 %.

Un affichage correct n'utilise **aucun code** mais seulement des *labels* et doit être présenté trié : les variables sont présentées par mode décroissant (le mode étant la modalité de plus fort pourcentage) et les modalités à l'intérieur d'une même variable sont ordonnées par pourcentage décroissant. L'affichage correct est donc

```
USAG NR      66    ( 66.67 %)        Peu     18    ( 18.18 %)
     Commun  13    ( 13.13 %)        Très     2    (  2.02 %)
SEXE Femme   64    ( 64.65 %)        Homme   35    ( 35.35 %)
REGI Très    43    ( 43.43 %)        Faible  35    ( 35.35 %)
     Moyenne 14    ( 14.14 %)        Forte    5    (  5.05 %)
     NR       2    (  2.02 %)
ETUD Sup     39    ( 39.39 %)        Bepc    30    ( 30.30 %)
     Bac     21    ( 21.21 %)        Primaire 6    (  6.06 %)
     NR       3    (  3.03 %)
```

Commentaires

La plupart (67 %) des gens dans cette population d'une centaine de personnes n'ont pas répondu à la question sur l'usage de la langue. Il y a eu une majorité de femmes interrogées puisqu'on a en gros une répartition 1/3 d'hommes et 2/3 de femmes. La régionalité se décline nettement (43 %) en forte régionalité et un peu moins nettement en faible régionalité (35 %). On trouve beaucoup (40 %) de gens ayant fait des études supérieures et un nombre relativement élevé de personnes de niveau bepc (30 %) ou bac (21 %).

La moyenne totale est la moyenne pondérée des 35 hommes et des 64 femmes. Si on désigne par m_F la moyenne d'age des femmes, on a donc l'équation

$$99 \times 35.828 = (35 \times 36.4) + (64 \times m_F)$$

On en déduit que l'age des femmes est (99*35.828 - 35*36.4)/ 64 soit 35.5151875 arrondi en `35.516` pour correspondre aux autres affichages. On peut appliquer la même technique pour la variance pondérée, soit

$$99(V + m^2) = 35(V_H + m_H^2) + 64(V_F + m_F^2)$$

où V est la variance globale, m la moyenne globale, V_H est la variance de l'age des hommes, m_H la moyenne de l'age des hommes, V_F est la variance de l'age des femmes et m_F la moyenne de l'age des femmes.

Utilisons les valeurs déja connues :

$$\begin{aligned} V &= \sigma^2 \\ &= 17.464^2 \\ &= 304.9913 \end{aligned}$$

$$\begin{aligned} V_F &= \sigma_F^2 \\ &= 17.886^2 \\ &= 319.9090 \end{aligned}$$

L'équation qui donne V_H est alors

$$99(304.9913 + 35.828^2) = 35(V_H + 36.4^2) + 64(319.9090 + 35.516^2)$$

On trouve donc que V_H vaut 277.1016 ce qui permet d'obtenir pour σ_H la valeur `16,64637` les chiffres manquant sont donc sans doute 1 et 6‡.

Nous ne détaillerons pas le calcul (simple) du *cdv* de l'age des femmes mais nous nous contentons de le donner dans le tableau complet de l'analyse de l'age ci-dessous :

ANALYSE DE L'AGE PAR SEXE

		TOUT	HOM	FEM
Age	Moy	35.828	36.400	35.516
	Ect	17.464	16.650	17.886
	Cdv	48.7	45.7	50.4

Il serait intéressant de se demander si les moyennes et variances pour les hommes et les femmes peuvent être considérées comme égales. Le test proposé en cours pour la comparaison de moyennes s'applique ici et il donne

$$\delta = \frac{|36.4 - 35.516|}{\sqrt{16.65^2/35 + 17.886^2/64}} = \frac{0.884}{3.59433} = 0.25$$

ce qui permet d'affirmer au risque de 5 % que les moyennes sont égales.

‡ la "vraie" valeur de l'écart-type est bien 16.65 ; l'arrondi sur les moyennes et les écart-types ne joue pas trop ici.

2. Analyse Statistique Générale du dossier CHIENS

Enoncé

Un chenil de la région Grand Ouest (pour ne pas dire "Pays de la Loire") nous a fourni des données relatives à des races de chiens. Nous reproduisons ici quelques données, la description des variables et quelques résultats informatiques.

Desciption des Variables

```
RACE Identificateur                 : races de chiens
HMM  Hauteur maximale du male       : entre 00 et 99 cm
HMF  Hauteur maximale de la femelle : entre 00 et 99 cm
PMIN Poids minimum                  : entre 00 et 99 kg
PMAX Poids maximum                  : entre 00 et 99 kg
DVM  Durée de vie moyenne           : entre 00 et 99 ans
```

Extrait des Données

	RACE	HMM	HMF	PMIN	PMAX	MA	LP	TP	OR	DVM	PRIX
1	alaskan malamute	67.0	55.0	30.0	38.0	1	1	1	0	14	0
2	basenji	42.0	41.0	10.0	11.0	1	1	1	0	10	0
3	basset-hound	38.0	38.0	28.0	30.0	0	1	1	2	12	0
...											
49	welsh-terrier	40.0	37.0	8.0	9.5	1	1	0	1	12	1
50	whippet	51.0	47.0	10.0	15.0	1	0	1	1	14	1

Sommes de valeurs

HMM	HMF
2607.7	2318.4

PMIN	PMAX
1022.0	1311.1

DVM	HMM*HMF	PMIN*PMAX	HMM*PMAX	HMM*HMM
642	133940.89	40730.51	81387.77	150249.89

Résutats

```
      champ        m         s        cdv
2     HMM        ??.???   ??.???      ??       20       80       60
3     HMF        46.368   15.881      34       15       75       60
5     PMAX       26.222   19.159      73        2       90       88
4     PMIN       20.440   14.879      73        1       60       59
6     DVM        12.840    1.782      14       10       17        7

            HMM     HMF     PMIN    PMAX    DVM
    HMM     1.0000
    HMF     0.9719  1.0000
    PMIN    0.8195  0.8287  1.0000
    PMAX    0.8045  0.8047  0.9775  1.0000
    DVM    -0.4296 -0.4381 -0.3676 -0.3877  1.0000


Rhos 1 : 0.977 PMAX  PMIN
     2 : 0.972 HMF   HMM
     3 : 0.829 PMIN  HMF

Correlation    0.977  :  PMAX    =    1.?59  *  PMIN    +   0.496
Correlation    0.977  :  PMIN    =    0.?59  *  PMAX    +   0.534
Correlation    0.972  :  HMF     =    0.914  *  HMM     -   1.317
Correlation    0.972  :  HMM     =    1.033  *  HMF     +   4.252
```

Critiquez les affichages, trouvez les résultats manquants (les `?` sont dus ici aux *Digimon* et non aux *Pokemon*) puis commentez les résultats obtenus.

Solution

A partir de la somme de `HMF` et de la moyenne affichée pour `HMF` on déduit qu'il y a $n = 2318.4/46.368 = 50$ lignes de données. La moyenne pour `HMM` est donc 2607.7 / 50 soit 52.154 et la moyenne des carrés 150249.89/ 50 soit 3005. La variance de `HMM` est alors 284.96028 et son écart-type 16.880767 d'où un coefficient de variation d'environ 32 %.

Le coefficient a pour la relation linéaire entre Y = `PMAX` et X = `PMIN` vaut $\rho \times \sigma(\texttt{PMAX})/\sigma(\texttt{PMIN})$ soit 0.977*19.159/14.879 c'est à dire 1.258038 et le

chiffre manquant est sans doute un 2.

On remarquera que la différence entre 1.258 et 1.259 est sans doute imputable aux erreurs d'arrondis.

Pour calculer le deuxième coefficient demandé dans la relation linéaire entre Y = `PMIN` et X = `PMAX` on peut

- soit utiliser le calcul $\rho \times \sigma(\texttt{PMIN})/\sigma(\texttt{PMAX})$ donc 0.977*14.879/19.159 ce qui donne 0.7587443 et le chiffre manquant est sans doute un 7,
- soit inverser la relation linéaire `PMAX = 1.259 * PMIN + 0.496` pour trouver $a = 1/1.259$ ce qui donne ici 0.7942812 et le chiffre manquant est encore un 7 (mais la valeur de a est ici nettement moins bonne).

Les commentaires de premier niveau que l'on peut faire sur les variables sont

- les males sont en moyenne plus grands que les femelles (ce qui doit être caractéristique des quadrupèdes, ou des mammifères...),
- le poids maximal est en moyenne plus grand que le poids minimal (cela parait évident, quoique...),
- le poids minimal et le poids maximal varient fortement sur l'ensemble des chiens considérés (cela revient à dire qu'il varie en fonction de la "race" de chien),
- la durée de vie moyenne de vie varie très peu (ce serait une caractéristique de l'espèce chien ?).

Si on utilise la matrice des corrélations et les formules linéaires pour les meilleures corrélations :

- le poids minimal et le poids maximal et le poids maximal sont liés par une formule linéaire simple énoncée comme *le poids maximal est en gros de 26 % supérieur au poids minimal*,
- la hauteur des males et la hauteur des femelles sont liées par une formule linéaire simple à savoir *les males ont en moyenne 4.3 cm de plus que les femelles.*

Toutefois, on peut s'interroger sur les données, leur stockage, la pertinence des variables.

Ainsi :

- les données auraient gagné à être arrondies ; le poids en kilos "tout ronds" doit être suffisant, de même que la taille en *cm* sans décimale ;
- pourquoi ne pas avoir traité le poids des males et des femelles ? s'il y a lieu de distinguer le poids minimal et le poids maximal, pourquoi ne pas traiter alors le poids minimal des males, le poids minimal des femelles, etc. ?
- la relation linéaire entre le poids minimal et la hauteur des femelles semble indiquer que poids et taille sont liés au sexe de l'animal mais le choix des variables ne permet pas de le révéler ; de plus, si on utise la relation linéaire suivante par ordre de ρ décroissant, le poids minimal est aussi lié à la hauteur des males (c'est une conséquence de la transitivité faible de la liaison linéaire),
- n'y aurait-il pas une confusion entre les chiens du chenil et les caractéristiques de la race des chiens ? chaque ligne de données semble correspondre à une race de chiens et non à un chien particulier du chenil. on aurait été en droit d'attendre un nombre de chiens par race correspondant aux animaux présents dans le chenil (pourquoi n'y aurait-il qu'un seul chien de chaque race dans le chenil ?).
- s'il s'agit de chiens réels, comment peut-on être sûr que l'animal est typique de la race ?
- s'il s'agit d'une valeur moyenne pour la race de chiens considérée, d'où sort cette moyenne ?

3. ASG du dossier BILAN

Enoncé

On trouvera ci-dessous des données d'entreprises extraites d'un magazine mensuel paru en 1996.

Description des variables

```
. PARTIC (PARTICIPATION)
  système obligatoire de répartition
  des profits quand ils atteignent un certain niveau
   0 ---> 0 francs
   1 ---> de 0 à 10 000 francs
   2 ---> plus de 10 000 francs

. FORMAT (DEPENSE DE FORMATION PAR SALARIE)
   0 ---> de 0 à 5000 francs
   1 ---> 5000 à 10 000 francs
   2 ---> 10 000 à 15 000 francs
   3 ---> plus de 15 000 francs

. INTERE (INTERESSEMENT)
  mécanisme de répartition des profits en fonction de critères
  liés aux performances
   0 ---> 0 francs
   1 ---> de 0 à 10 000 francs
   2 ---> plus de 10 000 francs
```

Extrait des Données

Enreg. Nø	NOM	PARTIC	FORMAT	INTERE	EFFECT	TXPREC	TPPART	SLRCAD
1	air li	1	2	0	4263	7.0	7.5	34731
2	alcate	0	3	0	4469	0.8	6.6	30385
3	alumin	0	2	0	3147	7.2	5.9	37659
4	automo	0	1	0	28392	5.1	3.8	24373
5	bertra	1	1	1	4808	15.5	0.7	26982

...

Voici quelques résultats concernant les variables qualitatives décrites. Essayez de compléter, critiquer puis décrire ces résultats.

```
code  PARTIC  FORMAT  INTERE
 0      31      13      41
 1      14      21       7
 2       7      15       7
 3              3      ??
```

Solution

Il doit y avoir en tout 52 lignes de données (52=31+14+7=13+21+15+3). Les valeurs pour `INTERE` sont alors incompréhensibles car 41+7+7 fait 55 donc plus que le total !

Après vérification auprès de l'organisme ayant transmis ces chiffres, il faut lire 41, 7 et 4 pour la colonne `INTERE`.

Le tableau récapitulatif des QL présenté de façon "intelligente" est alors

Variable	*Mode*	*Fréq.*	*Modalité*	*Fréq.*	*Modalité*	*Fréq.*
INTERE	0 kF	79 %	0 à 10 kF	13 %	plus de 10 kF	7 %
PARTIC	0 kF	60 %	0 à 10 kF	27 %	plus de 10 kF	13 %
FORMAT	5 kF à 10 kF	40 %	10 kF à 15 kF	29 %	0 à 5 kF	25 %

Les commentaires "naturels" sont

- une très forte majorité (presque 80 %) d'entreprises ont une faible politique d'intéressement ;
- une grande majorité (60 %) d'entreprises appliquent faiblement la participation ;
- la tendance non majoritaire (40 %) est de dépenser une valeur moyenne pour la formation.

Avec un peu de réflexion, on peut dire que :

- le choix des bornes pour la détermination des modalités n'est pas précisé ; il aurait été meilleur de donner les valeurs quantitatives pour justifier ces bornes ;
- le choix des variables est sans doute criticable : la formation est une volonté de l'entreprise, comme l'intéressement alors que la participation est une obligation légale.

4. Tige et feuille à la main

Enoncé

Donner le diagramme en *tige et feuille* pour les hommes du dossier ELF. On fournit les données :

```
60 22 62 65 78 20 49 28 47 64
26 43 42 16 20 22 52 28 28 52
29 28 30 26 29 32 27 35 33 17
18 25 47 12 62
```

Que peut-on en dire par rapport aux ages des femmes fournis ci-dessous :

```
11 12 13 14 15 15 15 17 17 18
18 19 19 19 19 21 21 22 23 24
24 25 25 25 26 26 27 27 28 28
28 29 30 31 31 31 35 36 37 39
39 40 41 44 44 46 48 48 49 50
50 50 53 59 60 61 61 62 63 70
73 73 73 76
```

On fournit le diagrammes en *tige et feuille* pour les ages des femmes, à savoir

n_i	T_i	F_i
15	1	123455577889999
17	2	11234455566778889
9	3	011156799
8	4	01446889
5	5	00039
5	6	01123
5	7	03336

Solution

Ce serait déjà plus facile si les données étaient triées, soient les valeurs :

```
12 16 17 18 20 20 22 22 25 26
26 27 28 28 28 28 29 29 30 32
33 35 42 43 47 47 49 52 52 60
62 62 64 65 78
```

On obtient alors rapidement :

n_i	T_i	F_i
4	1	2678
14	2	00225667888899
4	3	0235
5	4	23779
2	5	22
5	6	02245
1	7	8

La comparaison est visuelle : il y a plus de femmes que d'hommes partout ; d'autre part, il y a beaucoup de plus de femmes entre 10 et 19 que d'hommes pour les mêmes ages : les classes 1 et 2 pour les femmes sont presque aussi remplies.

Par contre le mode est le même dans les deux cas : entre 20 et 29 ans.

5. Transitivité de la relation Y = aX + b

Enoncé

Soient X et Y deux variables QT de même taille. Montrer que la relation binaire $\mathcal{R}$ définie par

$$X\mathcal{R}Y \Leftrightarrow \exists\ a, b\ ;\ Y = aX + b$$

est une relation d"équivalence.

Pourquoi dit-on que *corrélation n'est pas causalité* ? Quelle conséquence peut avoir la transitivité ?

Solution

Puisque X = 1 *X + 0 la variable X est relation avec elle-même donc $\mathcal{R}$ est réflexive.

L'équation $Y = aX + b$ peut se réécrire $X = (1/b)X - (b/a)$ donc de $X\mathcal{R}Y$ on passe à $Y\mathcal{R}X$: $\mathcal{R}$ est symétrique.

Si $Y = aX + b$ et si $Z = cY + d$ alors $Z = (ac)X + (cb + d)$ donc de $X\mathcal{R}Y$ et $Y\mathcal{R}Z$ on passe à $X\mathcal{R}Z$: $\mathcal{R}$ est transitive.

$\mathcal{R}$ étant réflexive, symétrique et transitive est une relation d'équivalence.

Imaginons que nous analysions la relation entre les variables "C=consommation d'essence" et "D=distance parcourue". S'il y a corrélation linéaire, C et D sont liés et on peut écrire les deux équations $\mathsf{C} = \alpha_1\mathsf{D} + \beta_1$ $\mathsf{D} = \alpha_2\mathsf{C} + \beta_2$.

S'il y a causalité, et c'est certainement le cas, seule une équation est correcte et interprétable physiquement : $\mathsf{C} = \alpha_1\mathsf{D} + \beta_1$.

T.D. 8 : χ^2, rangs et concordance

1. Un calcul progressif
2. Discussion sur m et σ
3. Un χ^2 d'indépendance en usine
4. χ^2 d'indépendance pour une vente de livres
5. Coefficients de corrélation des rangs
6. Coefficient de concordance de Kendall

1. Un calcul progressif

Enoncé

On dispose d'un ensemble de N valeurs $x_i > 0$ pour i de 1 à N. On suppose qu'on a calculé pour $n < N$ la somme s_n des n premières valeurs x_i (i de 1 à n) ainsi que leur moyenne m_n, la somme c_n de leur carré et leur variance v_n.

- exprimer s_{n+1} en fonction de s_n et x_{n+1} ;
- exprimer m_{n+1} en fonction de n, m_n et x_{n+1} ;
- exprimer c_{n+1} en fonction de c_n et x_{n+1} ;
- en déduire l'expression de v_{n+1} en fonction de n, m_n, m_{n+1}, v_n et x_{n+1}.

Application : si n =9, $s_n = 39$, $c_n = 199$ et $x_{n+1} = 10$, donner m_n, v_n puis s_{n+1} , m_{n+1}, c_{n+1} et v_{n+1}.

Un programmeur fait des statistiques une fois par an dans son entreprise pour le bilan annuel. Pour un certain produit, comptabilisé en kE ("kilo-euros") il n'a gardé des années précédentes que le nombre de valeurs `n`, leur moyenne `m` et leur écart-type `s`.

Sachant que la valeur à ajouter cette année est `x`, donner un algorithme accepté par *Galg* qui met à jour `n`, `m` et `s` sans utiliser de tableau.

Solution

Nous donnons sans explication l'enchainement des formules pour s_{n+1} qui ne présente aucune difficulté :

$$\begin{aligned} s_{n+1} &= \sum_{i=1}^{n+1} x_i \\ &= \sum_{i=1}^{n} x_i \ + \ x_{n+1} \\ &= s_n \ + \ x_{n+1} \end{aligned}$$

Puisque la "bonne formule" de la moyenne est $n_i.m_i = s_i$ on en déduit pour $n_i = i$ appliqué à $i = n + 1$ que :

$$m_{n+1} \;=\; \frac{s_{n+1}}{n+1} = \frac{n.m_n + x_{n+1}}{n+1}$$

Le calcul pour c_{n+1} est analogue à celui de s_{n+1} d'où $c_{n+1} = c_n + x_n^2$ et comme la "bonne formule" de la variance est $c_i = n_i(V_i + m_i^2)$ on en déduit pour $n_i = n + 1$ que :

$$c_{n+1} \;=\; (n+1)(V_{n+1} + m_{n+1}^2) = c_n + x_{n+1}^2$$

d'où, en extrayant v_{n+1} et en remplaçant c_n par $n(V_n + m_n^2)$:

$$v_{n+1} \;=\; \frac{n(V_n + m_n^2) + x_{n+1}^2}{n+1} - m_{n+1}^2$$

Pour l'application numérique,

$$\begin{aligned}
m_n &= s_n/n = 39/9 = 13/3 \simeq 4.33, \\
v_n &= c_n/n - m_n^2 = (199/9) - (13/3)^2 = 30/9 \simeq 3.33, \\
s_{n+1} &= 39 + 10 = 49, \\
m_{n+1} &= 49/10 = 4.9, \\
c_{n+1} &= 199 + 10^2 = 299, \\
v_{n+1} &= (299/10)\text{-}(49/10)^2 = 589/10 \simeq 5.89
\end{aligned}$$

Un algorithme de mise à jour avec des identificateurs plus explicites que m, c, n... peut être le suivant :

```
# maj annuelle des valeurs, auteur (gH)


# au début de l'algorithme :
#   moy est l'ancienne valeur, somcar l'ancienne somme des carrés,
#   nbval est l'ancien nombre de valeurs, ectype est l'ancien écart-type
#   et on rajoute la valeur val_x


  AFFECTER   somcar   <--    nbval * ( ectype * ectype ) + moy*moy
  AFFECTER   moy      <--  ( nbval * moy + val_x ) / ( nbval + 1 )
  AFFECTER   nbval    <--    nbval + 1
  AFFECTER   variance <--  ( somcar + val_x * val_x ) / nbval - moy * moy
  AFFECTER   ectype   <--    racine( variance )
```

2. Discussion sur m et σ

Enoncé

Trouver deux nombres x_1 et x_2 dont la moyenne est a et l'écart-type est b ; par exemple $a = 5$ et $b = 1$.

Peut-on trouver deux séries différentes X et Y avec chacune 2 valeurs ordonnées, soit $X = (x_1, x_2)$ et $Y = (y_1, y_2)$ avec $x_i \leqslant x_{i+1}$ et $y_i \leqslant y_{i+1}$ telles que $m(X) = m(Y)$ et $\sigma(X) = \sigma(Y)$?

Si oui, donner un exemple, si non démontrez-le.

Peut-on trouver deux séries différentes X et Y avec chacune 3 valeurs ordonnées, soit $X = (x_1, x_2, x_3)$ et $Y = (y_1, y_2, y_3)$ avec $x_i \leqslant x_{i+1}$ et $y_i \leqslant y_{i+1}$ telles que $m(X) = m(Y)$, $\sigma(X) = \sigma(Y)$?

Reprendre la question avec deux séries de n valeurs. En déduire pourquoi l'écart-type couplé à la moyenne et à la taille est un bon indicateur résumé d'une série de valeurs numériques.

Solution

Clairement, si on se donne x_1 alors x_2 vaut $2m$-x_1 soit encore $2a$ - x_1. Le double de la variance est alors $(x_1 - m)^2 + (2m - x_1 - m)^2$ soit $\sigma = |x_1 - m|$. On peut donc prendre comme solution $x_1 = a - b$ et $x_2 = a + b$. Par exemple pour $a = 5$ et $b = 1$ on trouve $x_1 = 4$ et $x_2 = 6$.

Soient $X = (x_1, x_2)$ et $Y = (y_1, y_2)$ les deux séries de valeurs.

Puisqu'elles ont même moyenne $m = (x_1 + x_2)/2 = (y_1 + y_2)/2$, on peut remplacer x_2 par $2m - x_1$ et y_2 par $2m - x_2$.

Si on impose que les deux séries ont même écart-type σ alors

$$2\ \sigma^2 = (x_1 - m)^2 + (x_2 - m)^2 = (y_1 - m)^2 + (y_2 - m)^2$$

Or $(x_2 - m)^2 = (\ (2m - x_1)\ - m)^2$ soit encore $(x_2 - m)^2 = (x_1 - m)^2$ et donc, si on développe la dernière égalité en remplaçant x_2 et y_2 par les valeurs dépendant de m, on arrive à l'équation :

$$(x_1 - m)^2 = (y_1 - m)^2$$

Si on fixe x_1 cette équation admet deux solutions :

- la première solution de cette équation est $x_1 - m = y_1 - m$ soit $y_1 = x_1$ ce qui n'est pas acceptable car alors les séries X et Y sont égales.
- la deuxième solution mène à $x_1 - m = -(y_1 - m)$ donc $y_1 = 2m - x_1$ soit encore $y_1 = x_2$ et cette solution n'est pas acceptable non plus car alors Y n'est pas ordonnée (c'est X à l'envers puisque $x_1 < x_2$).

Il n'est donc pas possible d'avoir deux séries différentes avec chacune 2 valeurs ordonnées de même moyenne et de même écart-type.

Par contre, et cet exercice le montre bien, la moyenne et l'écart-type sont insensibles à l'ordre des valeurs.

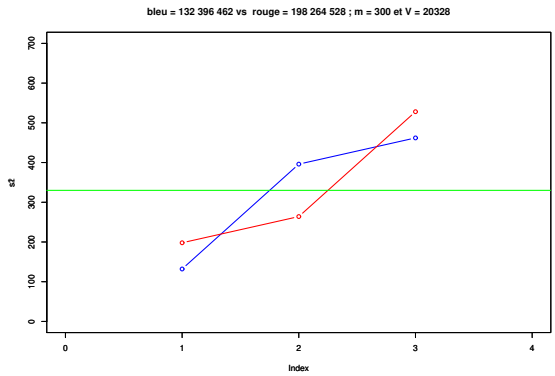
L'écriture des équations précédentes pour deux séries avec trois valeurs mène aux équations

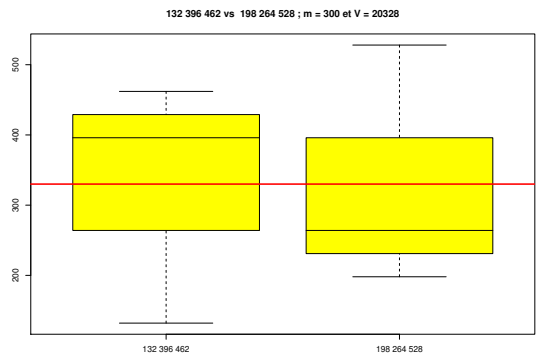
$$\begin{aligned} 3m &= x_1 + x_2 + x_3 = y_1 + y_2 + y_3 \\ 3\sigma^2 &= (x_1 - m)^2 + (x_2 - m)^2 + (x_3 - m)^2 \\ &= (y_1 - m)^2 + (y_2 - m)^2 + (y_3 - m)^2 \end{aligned}$$

Si on remplace x_3 par $3m - x_1 - x_2$ et y_3 par $3m - y_1 - y_2$ alors pour le cas particulier $x_1x_2 = y_1y_2$ le problème est possible si on résoud l'équation

$$(x_1 - y_1)(x_1 + y_1 - 3m) + 2x_1x_2 + (x_2 - y_2)(x_2 + y_2 - 3m) - 2y_1y_2 = 0$$

On arrive facilement à une solution entière comme $X = (132, 396, 462)$ et $Y = (198, 264, 528)$ qui ont toutes deux comme moyenne $m = 330$ et comme variance $\sigma^2 = 20328$. On se rend mieux compte de la différence entre les deux séries de valeurs à l'aide des deux graphiques suivants :

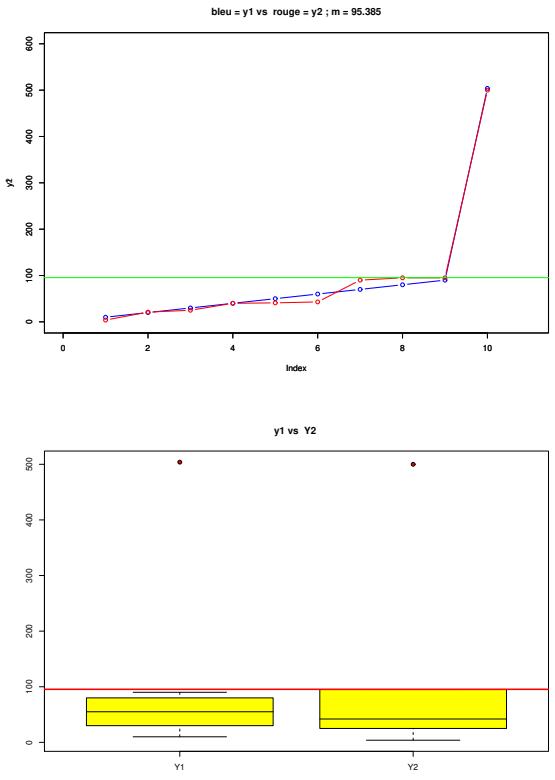




L'exemple précédent montre qu'il est possible de trouver deux séries différentes de 3 valeurs ordonnées de même moyenne et de même écart-type. Si au lieu de 3 valeurs on prend n valeurs on a encore plus de degrés de liberté puisque la relation $\Sigma\ x_i = n.m$ n'impose qu'une seule contrainte (linéaire) sur la dernière valeur. La relation $\Sigma\ (x_i - m)^2 = n.\sigma^2$ impose des conditions (non linéaires) avec suffisament de degrés de liberté pour qu'on puisse, si σ est grand, répartir "comme on veut" les carrés des différences à la moyenne.

Voici par exemple deux séries de $n = 10$ valeurs de même moyenne 95.385 et de même écart-type 138.340 :

Y1	Y2
10	3.846
20	21
30	25
40	40
50	41
60	43
70	90
80	95
90	95
503.846	500



Par contre si σ est faible, la somme de carrés $\Sigma\ (x_i - m)^2$ ne peut pas être décomposée de nombreuses façons et les x_i sont donc contraints à rester faiblement autour de la moyenne m. C'est pourquoi le couple (m, σ) est un bon résumé de l'ensemble des valeurs (tout en restant seulement un résumé).

Voici par exemple deux séries de $n = 10$ valeurs de même moyenne 50.97 et de même écart-type 2.92 ; elles sont presques indiscernables à "l'oeil nul".

```
49.6   49
49.7   49
49.8   49
49.9   50
50.0   50
50.1   50
50.2   50
50.3   50
50.4   54
59.7   58.7
```

3. Un χ^2 d'indépendance en usine

Enoncé

Dans une entreprise de 300 salariés, on trouve 70 % d'hommes et 30 % de femmes ; sachant qu'il y a 20 % de cadres et donc 80 % de non-cadres, donner les effectifs du tri-croisé SEXE/CADRE sous hypothèse d'indépendance.

Sachant maintenant que cette répartition est en fait

	Femmes	*Hommes*	Total
Cadres	10	50	60
Non-cadres	80	160	240
Total	90	210	300

effectuer le calcul du chi-deux d'indépendance. Conclure au seuil $\alpha = 5$ %.

Solution

Le total du nombre d'employés (300) facilite grandement le remplissage du tri-croisé : il y a 90 femmes, 210 hommes, 60 cadres et 240 non-cadres. Le nombre de femmes cadres sous hypothèse d'indépendance est de $90 \times 60 / 300$ soit 18.

Remplir le tableau du tri-croisé sous hypothèse d'indépendance c'est donc compléter le tableau

	Femmes	*Hommes*	Total
Cadres			60
Non-cadres			240
Total	90	210	300

en respectant les proportions 20 % et 80 % dans le sens des lignes, les proportions 30 % et 70 % dans le sens des colonnes.

Tous calculs faits on trouve :

	Femmes	*Hommes*	Total
Cadres	18	42	60
Non-cadres	72	168	240
Total	90	210	300

On en déduit le tableau des contributions signées `TH -OBS\verb`

	Femmes	*Hommes*
Cadres	+ 3.556	- 1.524
Non-cadres	- 0.889	+ 0.381

ce qui fournit comme valeur du chi-deux `6.34921` pour 1 degré de liberté.

Si on ordonne les contributions par ordre décroissant en valeur absolue :

```
+    3.556 ligne 1 colonne 1   Cadres      Femmes
-    1.524 ligne 1 colonne 2   Cadres      Hommes
-    0.889 ligne 2 colonne 1   Non-cadres  Femmes
+    0.381 ligne 2 colonne 2   Non-cadres  Hommes
```

on met en évidence un déséquilibre (pour les effectifs, par pour l'usine !) au niveau des femmes-cadres.

La valeur théorique du χ^2 au risque $\alpha = 5$ % étant de 3.842, on peut en conclure qu'il n'y a pas indépendance entre les modalités des variables `SEXE` et `CADRE` et qu'au contraire les modalités `femme` et `non-cadre` sont liées.

4. χ^2 d'indépendance pour une vente de livres

Enoncé

Voici, extrait d'une enquête de 1998 relative à l'achat de certaines familles de livres pour divers lieux de vente à l'occasion des fêtes de fin d'année, le tri à plat des deux variables `LIEU de vente` et `FAMILLE de livres`

```
Analyse par Lieu

        G     Grand Magasin           370     32.2 %
        H     Librairie générale      220     19.1 %
        I     Féd. Nat. d'Achat       330     28.7 %
        J     Vente par corresp.      230     20.0 %

 Analyse par Famille

        A    Art                       10      0.9 %
        B    Policiers                240     20.9 %
        C    Scientifiques            190     16.5 %
        D    Romans                   250     21.7 %
        E    Bandes dessinées         400     34.8 %
        F    Atlas                     60      5.2 %
```

ainsi que leur tri croisé :

	G	H	I	J
A	10	0	0	00
B	110	40	30	60
C	110	10	70	0
D	60	40	110	40
E	70	120	110	100
F	10	10	10	30

Réorganisez l'affichage des tris à plat puis indiquez s'il y a un lien entre lieu de vente et famille de livre et enfin discutez l'équirépartition de l'achat des livres par lieu et par famille.

Solution

Une "bonne" présentation des tris à plat est celle qui trie par mode et par fréquence, soit :

```
 pour un total de 1150 achats  :

   Famille  BD   35 %  Romans 22 %  Polic. 20 % Scien. 17 %
   Lieu     GM   32 %  Fnac   29 %  Vpc    20 % LibG.  19 %
```

ce qui met en évidence que les bandes dessinées sont les plus achetées et que le lieu le plus fréquent correspond aux grands magasins sans toutefois de modalité majoritaire.

La détection de l'indépendance se fait en calculant les effectifs théoriques sous hypothèse d'indépendance, c'est à dire en ne connaissant que les marges (totaux en ligne et en colonne), soit les valeurs $t_{i,j}$

```
  Effectifs théoriques avec recalcul des marges

     3.2       1.9       2.9       2.0  +      10.0
    77.2      45.9      68.9      48.0  +     240.0
    61.1      36.3      54.5      38.0  +     190.0
    80.4      47.8      71.7      50.0  +     250.0
   128.7      76.5     114.8      80.0  +     400.0
    19.3      11.5      17.2      12.0  +      60.0
++++++++++++++++++++++++++++++++++++++++++++++++++++++
   370.0     220.0     330.0     230.0  +    1150.0
```

Notant $d_{i,j}$ les données originales, le χ^2 d'indépendance se calcule comme la somme des contributions $c_{i,j}$ soit la formule

$$\chi^2 = \sum_{i,j} c_{i,j} \quad \text{où} \quad c_{i,j} = (d_{i,j} - t_{i,j})^2 / t_{i,j})$$

Il est intéressant d'afficher à coté des contributions le signe de la différence $d_{i,j} - t_{i,j}$ afin de savoir dans quel sens se situe le déséquilibre pour les fortes contributions, soient les valeurs :

```
Contributions avec signe de différence

  - 14.298    +  1.913    +  2.870    +  2.000
  - 13.918    +  0.762    + 21.938    -  3.000
  - 39.068    + 19.099    -  4.394    + 38.000
  +  5.192    +  1.281    - 20.406    +  2.000
  + 26.770    - 24.704    +  0.199    -  5.000
  +  4.485    +  0.190    +  3.025    - 27.000
```

Le total des contributions est égal à 281.51.

Le chi-deux maximal théorique pour $(l-1) * (c-1)$=5*3=15 degrés de libertés vaut 25. Notre χ^2 lui est supérieur donc on peut dire qu'il n'y a pas indépendance, ou plutôt qu'il y a liaison entre des modalités.

Pour déterminer cette liaison, il faut regarder les plus fortes contributions. Voici ces plus fortes valeurs, après tri des contributions :

```
  -   39.068 ligne 3 colonne 1  Scientifiques Grand Magasins
  +   38.000 ligne 3 colonne 4  Scientifiques VPC
  -   27.000 ligne 6 colonne 4
  +   26.770 ligne 5 colonne 1
  -   24.704 ligne 5 colonne 2
  +   21.938 ligne 2 colonne 3
  -   20.406 ligne 4 colonne 3
  +   19.099 ligne 3 colonne 2
  -   14.298 ligne 1 colonne 1
  -   13.918 ligne 2 colonne 1
  +    5.192 ligne 4 colonne 1
  -    5.000 ligne 5 colonne 4
  +    4.485 ligne 6 colonne 1
  -    4.394 ligne 3 colonne 3
...
```

Les deux plus fortes de toutes les contributions sont donc à peu près 39 et 38, fournies par la ligne 3, colonne 1 et la ligne 3 colonne 4. Le signe indiqué permet de dire qu'il y a eu plus de ventes de livres scientifiques en Grands magasins qu'on aurait du avoir et qu'il y a eu moins ventes de livres scientifiques en VPC.

Puisque le total général est 1150, s'il y avait équirépartition de la vente des livres par lieu pour chacun des 4 lieux, chaque lieu aurait 287,5 livres.

Nos valeurs, à savoir `370 220 330 230` en sont suffisamment éloignées pour qu'on se doute qu'un χ^2 d'adéquation va montrer qu'il n'y a pas accord avec la loi uniforme. Voici le détail des calculs faits par programme :

```
Valeurs observées :
     370  330  230  220
Valeurs théoriques :
     287.5  287.5  287.5  287.5

Somme des théoriques :  1150.0  = somme des observées.

CHI-DEUX avec  4  valeurs =  57.3043478
         pour  3  ddl (degrés de liberté)

Détail  des calculs :

   Différences     -82.50    -42.50     57.50     67.50
   Cumul            23.67     29.96     41.46     57.30

La valeur théorique du CHI-DEUX est
   7.815  pour 3  ddl au seuil de  5 %
```

Ces résultats confirment qu'au risque de première espèce $\alpha = 5\ \%$ on ne peut pas accepter l'hypothèse que nos données suivent la loi uniforme.

5. Coefficients de corrélation des rangs

Enoncé

Montrer que le coefficient de *Spearman* est en fait le coefficient usuel de corrélation linéaire. Donner les plages de variation de ρ_K et ρ_S.

Calculer ensuite les coefficient de corrélation des rangs de Spearman et de Kendall pour les valeurs A et B correspondants à des rangs de préférence pour 6 types de petits gâteaux pour le gouter

Numero du gateau	Rang A	Rang B
1	2	2
2	5	4
3	6	1
4	3	5
5	1	3
6	4	6

Donner enfin les algorithmes de calcul des deux coefficients supposant n donné, les rangs de A et B étant mis dans des tableaux tels que $A[i]$ et $B[i]$ correspondent à A_i et B_i.

Solution

On rappelle que si ε_i est la différence entre les rangs A_i et B_i, ρ_S vaut $1 - 6.(\Sigma\ \varepsilon_i^2)/n.(n^2-1)$ où n est le nombre d'objets à hiérarchiser.

On rappelle aussi que si r_i désigne le nombre d'inversions ($B_j > B_i$ pour $j > i$) entre A et B après la valeur i lorsque A est trié par ordre croissant, ρ_K, noté aussi τ_K, vaut $4.(\Sigma\ r_i)/n.(n-1) - 1$.

Soient $A = (A_i)$ et $B = (B_i)$ les rangs pour n objets de 1 à n.

Alors $m(A) = m(B) = (n+1)/2$ et $\sigma(A) = \sigma(B) = \sqrt{(n^2-1)/12}$ car A et B correspondent à la variable uniforme discrète $\mathcal{UD}(n)$.

Comme $\Sigma\ (A_i - B_i)^2 = \Sigma\ A_i^2 + \Sigma\ B_i^2 - 2\Sigma\ A_i B_i$, on en déduit que
$\Sigma\ \varepsilon_i^2 = 2n(n+1)(2n+1)/6 - 2\ \Sigma\ A_i B_i$.

Comme $\rho = cov(A,B)/\sigma(A)\sigma(B)$ et $cov(A,B) = (\Sigma\ A_iB_i)/n - m(A)m(B)$ on peut écrire $\Sigma\ A_iB_i = n\ (\ \rho \times \sigma(A)\sigma(B) + m(A)m(B)\ )$. Reportant cette valeur dans l'équation précédente, on obtient

$$\Sigma\ \varepsilon_i^2 = n(n+1)(2n+1)/3 - 2\left(\ \rho n(n^2-1)/12 - n(n+1)^2/4\ \right)$$

Développant ces calculs, avec $n(n+1)/12$ en facteur, on a

$$\Sigma\ \varepsilon_i^2 = n(n+1)/12\left[\ 4(2n+1) - 6(n+1)\right] - n\rho(n^2-1)/6\ \Big]$$

soit finalement $\Sigma\ \varepsilon_i^2 = n(n^2-1)/6\ (\ 1-\rho\ )$. On peut donc écrire le ρ normal comme $1 - 6\Sigma\ \varepsilon_i^2/n(n^2-1)$ ce qui est justement la définition de ρ_S.CQFD.

Si on a les mêmes rangs, alors A_i=B_i donc $\varepsilon_i = 0$ donc $\rho_S = 1$. Ou encore : $A = B$ donc $\rho(A,B) = 1$. Par contre, si les choix sont inverses : $B_i = n+1-A_i$ donc $\rho(A,B) = -1$, la démonstration directe dans ρ_S étant plus longue...

Pour ρ_K, si on a les mêmes rangs, on a $n-1$ inversions puis $n-2$ puis... jusqu'à 1. Alors Σr_i vaut n(n-1)/2 d'où $4\Sigma r_i/n(n-1)$ vaut +2 et donc $\rho_K = +1$.

Pour ρ_K, si les choix sont inverses, il n'y a aucune inversion donc $\rho_K = -1$.

Les ε_i valent ici respectivement 0,1,5,2,2,2 donc la somme de leur carré est 38 pour n=6, d'où ρ_S=1-6.38/6.35 soit -0,086.

L'algorithme du calcul est simplement

```
# Calcul de rhos (corrélation des rangs pour Spearman)

   sdc ← 0
   pour i de1a n
        eps ← A[i]-B[i]
        sdc ← sdc + eps*eps
   finpour
   denom ← n*(n*n-1)

   rhoS  ← 1 - 6*sdc/denom
```

Si on trie A par ordre croissant et B en conséquence, les rangs deviennent

```
Rang A   1  2  3  4  5  6
Rang B   3  2  5  6  4  1
```

et les nombres r_i sont respectivement 3,3,1,0,0,0 de somme 7. ρ_K vaut donc $4.7/6.5-1$ soit -0,067. On remarquera que dans les deux calculs de corrélation des rangs ρ est négatif ce qui semble indiquer plutot un désaccord entre A et B...

Voici un algorithmes possible, avec trois méthodes de calculs différentes, lorsque les valeurs sont triées pour A

```
# Calcul de rhok (corrélation des rangs pour Kendall)

   r1 ← 0 # inversions simples
   r2 ← 0 # comptage algébrique
   r3 ← 0 # dépassements simples
   pour i de 1 a  n-1
      pour j de i+1 a n
         si b[j] < b[i] alors
            r1 ← r1 + 1
            r2 ← r2 - 1
         sinon
            r2 ← r2 + 1
            r3 ← r3 + 1
         finsi
      finpour sur k
   finpour sur j

   rhoK1 ← 1 - (4*r1/(n*(n-1)))
   rhoK2 ← (2*r2/(n*(n-1)))
   rhoK3 ← (4*r3/(n*(n-1))) - 1
```

On remarquera que les algorithmes sont très simples, contrairement aux formules.

6. Coefficient de concordance de Kendall

Enoncé

Les coefficients de corrélation de *Kendall* et de *Pearson* permettent de comparer deux séries de rangs. Si l'on veut comparer m jugements plutôt que deux, on a recours au calcul du coefficient R_K de concordance de *Kendall* qui se calcule comme suit.

Soient n objets à classer par un rang (nombre de 1 à n). Soit $r_{i,j}$ le rang donné à l'objet i par le juge j et s_i la somme des rangs attribués à l'objet i c'est à dire

$$s_i = \sum_{j=1}^{m} r_{i,j}$$

On note S la moyenne des s_i. R est alors calculé par

$$R_K = \frac{12\ T}{m^2(n^3 - n)}$$

où T est la somme des carrés des écarts des s_i à S c'est à dire la quantité

$$T = \sum_{i=1}^{n} (s_i - S)^2$$

Que vaut R_K si tous les juges sont tous d'accord pour mettre le rang i à l'objet i ?

Donner un exemple de jugements avec n=3 objets et m=4 juges tels que R_K vaut 0. On mettra les objets en lignes et les juges en colonnes.

Calculer R_K avec 3 décimales exactes pour le tableau des m=7 juges et n=4 objets suivant

	Juge 1	Juge 2	Juge 3	Juge 4	Juge 5	Juge 6	Juge 7
Objet 1	1	2	1	2	1	2	1
Objet 2	4	3	4	3	2	4	4
Objet 3	3	4	2	4	4	3	3
Objet 4	2	1	3	1	3	1	2

Solution

Si les m juges donnent la note i à l'objet i alors s_i vaut mi. La moyenne S des s_i sur les n objets est alors $m(n+1)/2$. Alors $T = \sum_{i=1}^{n} (s_i - S)^2$ vaut :

$$\begin{aligned} T &= m^2 \left(\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{2} + \frac{n(n+1)^2}{4} \right) \\ &= m^2 \, \frac{n(n+1)}{12} \, (\, 2(2n+1) - 6(n+1) - 3(n+1)) \\ &= m^2 \, \frac{(n^3 - n)}{12} \end{aligned}$$

et R_K vaut donc 1.

Pour avoir R_K =0 il faut que chaque objet ait le meme score que les autres, à savoir 8 en total de ligne. On peut prendre par exemple les jugements

	Juge 1	Juge 2	Juge 3	Juge 4
Objet 1	1	1	3	3
Objet 2	2	2	2	2
Objet 3	3	3	1	1

Pour l'exemple proposé, on trouve numériquement :

$$\begin{aligned} s_1 &= 10, \\ s_2 &= 24, \\ s_3 &= 23, \\ x_4 &= 13, \\ S &= 17.5, \\ T &= 149, \\ R_K &= 1788 \ / \ 2940 \text{ soit } 0,608163265. \end{aligned}$$

T.D. 9 : Comparaisons

1. *Kendall* : inversions après A ou B ?
2. Nombre d'appels sur une "hotline"
3. Comparaison de moyennes : durées de tri
4. Comparaison de pourcentages
5. Comparaison de variances
6. Intervalle de confiance et de variabilité
7. Que font ces programmes ?

1. *Kendall* : inversions après A ou B ?

Enoncé

Un enseignant étourdi écrit parfois que le coefficient de corrélation des rangs de *Kendall* entre A et B se calcule à l'aide du nombre r_i d'inversions calculé comme le nombre de valeurs $B_j > A_i$ pour $j > i$; d'autres fois il écrit que r_i correspond au nombre de valeurs $B_j > B_i$ pour $j > i$. De plus il escamote régulièrement la démonstration de ρ_S = -1 pour le cas $B_i = n + 1 - A_i$.

Comment aider cet enseignant à progresser ?

Solution

Pour savoir si on utilise les $B_j > A_i$ ou les $B_j > B_i$ il suffit d'un peu de bon sens. En effet, l'algorithme fourni au T.D. précédent pour calculer ρ_K sous ses trois formes utilise toujours $B_j > B_i$.

De plus la formule 2 de ce même algorithme qui exprime ρ_K en fonction de la somme algébrique des inversions $B_j > B_i$ à savoir

$$\rho_K = \frac{R_2}{\frac{n(n-1)}{2}}$$

est facile à vérifier et donc à interpréter : on voit que R_2 est divisé par C_n^2.

Le nombre maximum d'inversions positives à partir de B_1 calculé comme $B_j > B_i$ correspond au nombre de couples (i, j) avec $j > i$ lorsque B est en ordre croissant ; il y en a bien sûr C_n^2. On en déduit que R_2 vaut alors 1.

De même le nombre maximum d'inversions négatives à partir de B_1 calculé comme $B_j < B_i$ correspond au nombre de couples (i, j) avec $j > i$ lorsque B est en ordre croissant ; il y en a bien sûr là encore C_n^2. On en déduit que R_2 vaut alors -1.

Quant à la démonstration de ρ_S = -1 pour le cas $B_i = n + 1 - A_i$ ce n'est pas vraiment plus difficile, juste un peu plus calculatoire.

Si $B_i = n + 1 - A_i$ alors $\varepsilon_i = B_i - A_i$ vaut $(2i - 1) - n$. Il n'y a plus qu'à savoir calculer la somme des carrés des n premiers impairs pour conclure.

Désignons par S_n la somme des n premiers nombres entiers, par P_n la somme des n premiers nombres entiers pairs et enfin par I_n la somme des n premiers nombres entiers impairs.

On sait depuis l'école primaire que S_n vaut $n(n+1)/2$. Mettant 2 en facteur dans P_n on a $P_n = 2\ S_n$ donc $P_n = n(n+1)$.

De plus, puisque $S_{2n} = I_n + P_n$ on en déduit que $I_n = n^2$ ce qui était connu des philosophes grecs 3 siècles avant notre ère.

Revenant à notre calcul, désignons par $c_{n,1}$ la somme $\Sigma\ (2i-1)^2$, par $c_{n,2}$ la somme $\Sigma\ 2\ (2i-1)\ n$ et par $c_{n,3}$ la somme $\Sigma\ n^2$. Alors :

$$\Sigma\ \varepsilon_i^2 = c_{n,1} - c_{n,2} + c_{n,3}$$

A l'aide de la somme des n premiers carrés $K_n = \sum_{i=1}^{n} i^2 = n(n+1)(2n+1)/6$ et de la somme des carrés des n premiers nombres pairs $J_n = \Sigma\ (2i)^2 = 4\ K_n$ et puisque $K_{2n} = J_n + c_{n,1}$ on en déduit que $c_{n,1}$ vaut $n(4n^2-1)/3$.

Par contre le calcul de $c_{n,2}$ est simple et vaut $2n\ I_n$ soit $2n^3$ de même que $c_{n,3} = n^3$.

Après réduction au même dénominateur, on trouve que $\Sigma\ \varepsilon_i^2 = n(n^2-1)/3$.

Multipliant par $6/n(n^2-1)$ on obtient la valeur 2 et donc finalement le calcul de ρ_s donne -1 ce qui était le résultat demandé.

2. Nombre d'appels sur une "hotline"

Enoncé

On compte le nombre d'appels obtenus sur une "ligne chaude" pour l'aide en ligne d'un nouveau logiciel de gestion. On obtient les valeurs suivantes

nb jours	2	3	4	5	6	7	8	9
nb appels	03	03	05	38	39	75	26	01

Lors de la dernière mise à jour du logiciel, cette même ligne avait enregistré 115 appels en tout pour une durée moyenne de 6 jours avec un écart-type de 1.2083 jour. Peut-on comparer l'utilisation de la ligne pour les deux séries de valeurs ?

Solution

Bien sûr qu'on peut comparer ! On dispose de toutes les valeurs nécessaires pour comparer les moyennes de durées... σ=1.2083 donne $V = \sigma^2 = 1.46$, et pour cette année, $\Sigma\, n_i = 190$, $m_a = 6.32$, $v_a = 1.587$ donc $\delta = 0.32/\sqrt{0.0211} = 2.20$. Au seuil de 5 %, la différence est donc significative.

3. Comparaison de moyennes entre ordinateurs

Enoncé

On dispose de deux ordinateurs nommés A et B. Des simulations de transferts de fichiers fournissent les valeurs suivantes pour les fichiers transférés

taille (Meg)	ordi. A	ordi. B
3	6 fois	8 fois
4	2	3
8	5	11
11	9	7

Effectuez une comparaison de moyennes pondérées des tailles pour les fichiers temporaires transférés. On donne $\Sigma\, a_i t_i = 165$, $\Sigma\, a_i t_i^2 = 1495$, $\Sigma\, b_i t_i = 201$, $\Sigma\, b_i t_i^2 = 1671$.

Solution

Rappelons les formules de moyennes et de variance pondérées : si les n valeurs x_i pour i de 1 à n sont utilisées directement, alors

$$\sum_{i=1}^{n} x_i = n.m \quad \text{et} \quad \sum_{i=1}^{n} x_i^2 = n(V + m^2)$$

Si on regroupe les x_i en r_j fois x_j^* alors pour les p valeurs distinctes x_j^*

$$\sum_{j=1}^{p} r_j x_i^* = n.m \quad \text{et} \quad \sum_{j=1}^{p} r_j (x_i^*)^2 = n(V + m^2)$$

Tous calculs effectués, on trouve $n_a = \Sigma\ a_i = 22$, $m_a = 165/22 = 7.5$, $v_a = (1495/22)$ - $m_a^2 = 11.7045455$ et $n_b = 29$, $m_b = 6.93103448$, $v_b = 9.5814507$ d'où $|m_a - m_b| = 0.56896552$, $r = \sqrt{0.862419647} = 0.928665519$, soit enfin $\delta = 0.612670018$. Ce nombre est inférieur à 1.96 donc on peut conclure que les deux ordinateurs se comportent de la même façon sur cet exemple.

4. Comparaison de pourcentages

Enoncé

D'après J. Saville (*Les déplacements humains*, Ed. de Monaco, 1962), les recensements de la population anglaise et galloise réunies, tant urbaine que totale furent (en millions d'individus) :

An	Urbain	Total
1851	8,99965972	17,927609
1901	25,04643911	32,527843
1951	35,30197290	43,744700

- Y a-t-il une différence significative au seuil de 5 % pour la proportion urbain/total entre 1851 et 1901 ?
- Y a-t-il une différence significative au seuil de 5 % pour la proportion urbain/total entre 1901 et 1951 ?

Vous indiquerez clairement la valeur de p_a, p_b, celle de r et de ε correspondant au cours avant de conclure.

Solution

Avant d'utiliser les formules du cours, il est bon de se poser la question de la validté des données. Si les valeurs fournies sont exprimées en milions d'individus, la valeur `8,99965972` correspond à

```
8 millions 999 mille 659 individus virgule 72
```

ce qui n'est certainement pas possible ! Voici donc les vraies valeurs :

An	Urbain	Total
1851	8,999659	17,927609
1901	25,046439	32,527843
1951	35,301972	43,744700

Pour la comparaison des pourcentages urbain/total entre 1851 et 1901, avec les notations du cours :

```
ia   8999659  na  17927609  pa  0.501999960
ib  25046439  nb  32527843  pb  0.769999997
i   34046098  n   50455452  p   0.674775404
```

La différence entre p_a et p_b est 0.268000037, r^2 vaut 0.0000000189877 donc r vaut 0.000137796 soit finalement $\varepsilon \simeq 1944.90$; on est très loin de 1.96 donc la différence entre ces pourcentages est significative.

De même pour la comparaison des pourcentages urbain/total entre 1901 et 1951

```
ia  25046439  na  32527843  pa  0.769999997
ib  35304972  nb  43744700  pb  0.807068559
i   60351411  n   76272543  p   0.791259982
```

La différence entre p_a et p_b est 0.037068562, r^2 est 8.85345 10^{-9}, r vaut à peu prés 0.00009409276 soit finalement $\varepsilon = 393.96$; là encore, on est très loin de 1.96 donc la différence entre ces pourcentages est significative.

5. Comparaison de variances

Enoncé

Soient A et B les mesures d'étalonnage de fréquences pour un spectrophotomètre prises respectivement pour 250 $m\mu$ et 260 $m\mu$.

A	120	164	153	148	143	132	155	142	169	144
B	172	210	206	199	192	181	204	190	218	198

Comparer les variances des deux séries.

Solution

Pour nos deux séries de 10 valeurs, on trouve $m_A = 147$, $m_B = 197$, $V_A = 208.67$, $V_B = 188.89$. On ne peut aller plus loin car aucune formule du cours ne traite de la comparaison de deux variances.

Si on admet que l'usage est d'utiliser le du test des rapport des variances $R = V_A/V_B$ où le seuil maximal est donné dans la table de *Snedecor*, on trouve ici $R = 1.10$ mais comme le seuil R_{max} lu dans la table au seuil de α =5 % est 3.18, pour 9 ddl, on considére donc qu'il n'y pas de différence significative entre nos séries de valeurs...

6. Intervalle de confiance et de de variabilité

Enoncé

Soit X une série statistique de moyenne m, de variance V et d'écart-type σ. On appelle intervalle de confiance à α % l'intervalle bilatéral centré défini par $I_m = [m - t\sigma, m + t\sigma]$ où t correspond à $p(\text{"}|\mathcal{N}(0,1)| < \alpha\text{"})$. On appelle intervalle de variabilité l'intervalle centré $I_V = [m - tV, m + tV]$. Enfin, on appelle intervalle de sûreté $I_s = [m - t\sigma/\sqrt{n}, m + t\sigma\sqrt{n}]$ où n désigne le nombre de valeurs.

Soit $X = (12, 15, 17, 50)$, $Y = (25, 31, 35, 101)$ et $Z = (144, 255, 289, 2500)$ trois séries statistiques correspondant à des variables quantitatives, dont les

unités sont respectivement la minute, le kilomètre et la minute-carrée. Donner la matrice des corrrélations de X, Y et Z ainsi que leur intervalle de confiance, leur intervalle de variabilité pour $t = 1.96$ et leur intervalle de sûreté. On fournit, si cela peut aider les sommes suivantes

```
sx      sy      sz
94      192     3158
sx*sx   sy*sy   sz*sz
3158    13012   6404882
sx*sy   sx*sz   sy*sz
6410    135016  273190
```

Solution

On trouve assez facilement $m_X = 23.5\ min$, $m_Y = 48\ km$, $m_Z = 789.5\ min^2$, $\sigma_X = 15.403\ min$, $\sigma_Y = 30.806\ km$ et $\sigma_Z = 988.893\ min^2$ soit des cdv respectifs de 70 %, 64 % et 125 % ; on devrait donc utiliser l'ordre `Z`, `X`, `Y` pour afficher les résultats.

La matrice des corrélations est

```
        X       Y       Z
X       1.0000
Y       1.0000  1.0000
Z       0.9980  0.9980  1.0000
```

S'intéressant à la liaison linéaire entre X et Y (puisque $\rho(X, Y) = 1$), à l'aide des formules du cours on trouve donc que $Y = 2X + 1$. De même, $\rho(X, Z)$ est proche de 1 et on obtient par calcul $Z = 64.071 * X - 716.159$ (alors qu'en fait la "vraie" liaison est $Z = X^2$).

Pour la variable X, $I_m = [m - t\sigma, m + t\sigma] = [23.5 - 1.96 * 15.4, 23.5 + 1.96 * 15.4]$ est donc l'intervalle $[-6.68, 53.68]$. L'intervalle $I_V = [m - t.V, m + t.V]$ est stupide car m et V n'ont pas les mêmes unités.

Pour $I_s = [m - t\sigma/\sqrt{n}, m + t\sigma\sqrt{n}]$, comme n vaut 4, on trouve $[8, 4, 38.6]$

7. Que font ces programmes ?

Enoncé

Voici deux algorithmes et leur traduction respective en perl et en java. Que font ces programmes ? Quelles seraient les sorties si on passe comme paramètres 100 9 1 pour le premier et 100 15 10 pour le second ?

Algorithme 1 dit "simb"

```
affecter iobs <-- 1
tant_que (iobs<=nbobs)
|   affecter som  <-- 0
|   affecter itos <-- 1
|   tant_que (itos<=nbtos)
|     affecter x <-- valeurAleatoire()
|     si (x>0.5)
|        alors affecter y <-- 1
|     sinon
|        affecter y <-- 0
|     fin_si # sur x
|     affecter som <-- som + y
|     affecter itos <-- itos + 1
|   fin_tant_que # sur itos
|   affecter haut <-- href
|   tant_que (som>haut)
|      affecter haut <-- haut + href
|   fin_tant_que # sur som>haut
|   affecter tclass[ haut ] <-- tclass[ haut ] + 1
|   affecter iobs <-- iobs + 1
fin_tant_que # sur iobs

# pour chaque indice v du tableau tclass trié
  écrire_ formatReel(v,5,2)
  écrire  "   " , formatEntier(tclass[v])
# fin pour
```

Algorithme 2 dit "simj"

```
affecter mclass <-- 0
affecter iobs   <-- 1

tant_que (iobs<=nbobs)
|
|    affecter som  <-- 0
|    affecter itos <-- 1
|
|    tant_que (itos<=nbtos)
|      affecter x    <-- valeurLoiNormale()
|      affecter som  <-- som + x *x
|      affecter itos <-- itos + 1
|    fin_tant_que # sur itos
|
|    affecter haut <-- href
|    affecter clas <-- 0
|
|    tant_que (som>haut)
|       affecter haut <-- haut + href
|       affecter clas <-- clas + 1
|    fin_tant_que # sur som>haut
|
|    affecter tclass[ clas ] <-- tclass[ clas ] + 1
|    si clas>mclass
|       alors affecter mclass <-- clas
|    fin_si # clas > mclass
|    affecter iobs <-- iobs + 1
|
fin_tant_que # sur iobs

pour indi de1a mclass
  écrire  format(indi,3) , " : "  , format(tclass[indi],5)
fin_pour # indi
```

programme Perl associé à l'algorithme 1

```
          # @fmt perlp.tex ;
          # @src simb.pl

     0001 die(" il en faut  trois.") unless ($#ARGV>1) ;
                                          # faut ce qu'il faut !
     0002
     0003  ($nbobs,$nbtos,$href) = ($ARGV[0],$ARGV[1],$ARGV[2]) ;
     0004  $iobs = 1 ;
     0005
     0006  while ($iobs<=$nbobs) {
     0007    $som  = 0 ;
     0008    $itos = 1 ;
     0009    while ($itos<=$nbtos) {
     0010      $x    = rand() ;
     0011      if ($x>0.5) {
     0012         $y = 1 ;
     0013      } else {
     0014         $y = 0 ;
     0015      } ; # fin de si
     0016      $som += $y ;
     0017      $itos++ ;
     0018    } ; # fin tant que sur itos
     0019    $haut = $href ;
     0020    while ($som>$haut) { $haut += $href ; } ;
     0021    $tclass{$haut}++;
     0022    $iobs++ ;
     0023  } ; # fin tant que sur iobs

     0024
     0025  foreach $v (sort keys %tclass) {
     0026    print sprintf("%05.2f",$v) ;
     0027    print "   ".sprintf("%4d",$tclass{$v})." \n" ;
     0028  } ; # fin pour chaque clé triée v dans tclass

//#-#   Fin de traduction pour simb.pl via de galg -a simb.alg -o perl
```

programme Java associé à l'algorithme 2

```
0001       class simj {
0004
0005            mclass = 0 ;
0006            iobs = 1 ;

0007            while ((iobs <= nbobs)) {
0008              som = 0 ;
0009              itos = 1 ;
0010              while ((itos <= nbtos)) {
0011                x = random(  ) ;
0012                som = som + x * x ;
0013                itos = itos + 1 ;
0014              } ; // sur itos
0015              haut = href ;
0016              clas = 0 ;
0017              while ((som > haut)) {
0018                 haut = haut + href ;
0019                 clas = clas + 1 ;
0020              } ; // sur som>haut
0021              tclass[clas] = tclass[clas] + 1 ;
0022              if (clas > mclass) {
0023                 mclass = clas ;
0024              } ; // clas > mclass
0025              iobs = iobs + 1 ;
0026            } ; // sur iobs
0027
0028            for (int indi=1; indi<=mclass; indi++) {
0029               System.out.println(format( indi, 3 )
                        +" : "+format( tclass[indi], 5 )) ;
0030
0031            } ; // indi
0032
0033       } // fin de la classe simj

//#-#   Fin de traduction pour simj.java via de galg -a simj.alg -o java
```

Solution

Visiblement ces programmes font des calculs ! Pour le premier programme, l'instruction en ligne 1 vérifie qu'il y a au moins trois arguments (car en *Perl* on commence à compter à partir de 0). Les paramètres sont respectivement le nombre *nbobs* de valeurs (ou "nombre d'observations") , le nombre *nbtos* de "toss" c'est à dire le nombre de répétitions et la hauteur $href$ de référence.

Le programme effectue ensuite une première boucle de simulation (lignes 6 à 23) pour produire *nbobs* résultats. Chaque résultat est construit de la valeur suivante : on effectue *nbtos* tirages aléatoires entre 0 et 1 grâce à la fonction `rand` dans la deuxième boucle (lignes 9 à 18) et pour chaque tirage on compte 0 si le nombre est strictement inférieur à 0.5 et 1 s'il est supérieur ou égal à 0.5 ; le résultat est la somme des *nbtos* valeurs 0 ou 1 nommé *som*.

A chaque valeur *som* on associe une hauteur notée *haut* grâce à la boucle de la ligne 20 ; pour associer une hauteur, on incrémente la hauteur de base par pas de $href$ (qui avait été fourni en paramètre) jusqu'à dépasser la valeur de *som*.

On utilise ensuite en ligne 21 un tableau associatif (ou "hachage") pour comptabiliser combien de valeurs ont la même hauteur.

La dernière boucle du programme (lignes 25 à 28) vient afficher les hauteurs trouvées et leurs effectifs respectifs.

Le tirage aléatoire exécuté par `rand` correspond à la simulation de valeurs uniformément réparties sur $[0, 1]$, nommée `valeurAleatoire` dans l'algorithme. Ces valeurs correspondent à la loi $\mathcal{UC}([0, 1])$ du cours.

Compte-tenu de ce que nous avons appris, il est clair que ce programme simule *nbobs* fois l'addition de *nbtos* tirages d'une loi de bernoulli de paramètre 0.5 ; ce programme fournit donc une simulation de la binomiale $\mathcal{B}(n, p)$ avec comme paramètres $n = nbtos$, $p = 0.5$ sur un effectif total de *nbobs* valeurs pour ensuite afficher les valeurs des effectifs obtenus par classe de largeur $href$.

On notera que si on remplace la valeurs 0.5 de la ligne 11 par une autre valeur p entre 0 et 1, c'est la loi binomiale $\mathcal{B}(n, p)$ qui est alors simulée.

Voici le détail d'une exécution avec les paramètres `5 3 1` ; nous avons ajouté dans le programme des instructions instructions pour afficher les variables dans la boucle la plus interne :

```
 iobs   itos    x          somme

   1      1     0.87602915     1
   1      2     0.53537713     2
   1      3     0.15608289     2
                    somme      2 effectif     1

   2      1     0.57335606     1
   2      2     0.82444119     2
   2      3     0.96009187     3
                    somme      3 effectif     1

   3      1     0.31753912     0
   3      2     0.48113369     0
   3      3     0.52712724     1
                    somme      1 effectif     1

   4      1     0.32291614     0
   4      2     0.46888259     0
   4      3     0.45528799     0
                    somme      1 effectif     2

   5      1     0.56834310     1
   5      2     0.55914746     2
   5      3     0.10163933     2
                    somme      2 effectif     2

  hauteur effectif

   01.00      2
   02.00      2
   03.00      1
```

Voici 4 exemples d'exécution avec les paramètres proposés `100 9 1` pour le programme original :

```
01.00      1          01.00      3
02.00      9          02.00      5
03.00     16          03.00     18
04.00     26          04.00     27
05.00     27          05.00     24
06.00     12          06.00     13
07.00      8          07.00      9
08.00      1          08.00      1

01.00      6
02.00      2          02.00     10
03.00     22          03.00     23
04.00     30          04.00     16
05.00     18          05.00     18
06.00     14          06.00     17
07.00      8          07.00     15
                      08.00      1
```

On remarquera que les valeurs extremes `0` et `9` ne sont jamais atteintes, ce qui est "normal" puisque les valeurs théoriques du modèle sont

```
valeurs        probabilité    cumul   effectif
   0            0.00195       0.00195     0
   1            0.01758       0.01953     2
   2            0.07031       0.08984     7
   3            0.16406       0.25391    16
   4            0.24609       0.50000    25
   5            0.24609       0.74609    25
   6            0.16406       0.91016    16
   7            0.07031       0.98047     7
   8            0.01758       0.99805     2
   9            0.00195       1.00000     0
```

Le deuxième programme n'est pas correct : un programme java ne se réduit pas à une classe sans fonction *main*. En particulier la variable *mclass* n'est pas déclarée, pas plus que les variables *iobs*, *nbobs*... Si ce le deuxième programme doit fonctionner comme le premier, il devrait aussi y avoir une détection des paramètres... Toutefois, en admettant qu'il y ait une déclaration des variables, une gestion des paramètres, il reste des problèmes : la fonction *random*() de la ligne 11 n'existe pas en *Java* : si on veut un tirage aléatoire pour une loi uniforme, il faut écrire `Math.random()`.

Compte-tenu de ce qui a été vu dans le programme précédent, il est clair que celui-ci effectue la simulation de *nbobs* valeurs pour une loi définie comme suit : on fait la somme du carré de *nbtos* valeurs d'une loi uniforme réelle continue sur [0..1]. Rien dans le cours n'indique de quoi il s'agit. Voici cependant un exemple d'éxécution avec les paramètres `5 3 1` (sans l'affichage de fin avec les classes) :

```
01.00      1          01.00      3
02.00      9          02.00      5
03.00     16          03.00     18
04.00     26          04.00     27
05.00     27          05.00     24
06.00     12          06.00     13
07.00      8          07.00      9
08.00      1          08.00      1


01.00      6
02.00      2          02.00     10
03.00     22          03.00     23
04.00     30          04.00     16
05.00     18          05.00     18
06.00     14          06.00     17
07.00      8          07.00     15
                      08.00      1
```

On remarquera que les valeurs extremes `0` et `9` ne sont jamais atteintes.

Pour comprendre pourquoi ces valeurs extremes `0` et `9` ne sont jamais atteintes, il suffit de calculer les valeurs théoriques du modèle soit

valeurs	probabilité	cumul	effectif
0	0.00195	0.00195	0
1	0.01758	0.01953	2
2	0.07031	0.08984	7
3	0.16406	0.25391	16
4	0.24609	0.50000	25
5	0.24609	0.74609	25
6	0.16406	0.91016	16
7	0.07031	0.98047	7
8	0.01758	0.99805	2
9	0.00195	1.00000	0

Le deuxième programme n'est pas correct : un programme java ne se réduit pas à une classe sans fonction *main*. En particulier la variable *mclass* n'est pas déclarée, pas plus que les variables *iobs*, *nbobs*... Si ce le deuxième programme doit fonctionner comme le premier, il devrait aussi y avoir une détection des paramètres... Toutefois, en admettant qu'il y ait une déclaration des variables, une gestion des paramètres, il reste des problèmes : la fonction *random*() de la ligne 11 n'existe pas en *Java* : si on veut un tirage aléatoire pour une loi uniforme, il faut écrire `Math.random()`.

Compte-tenu de ce qui a été vu dans le programme précédent, il est clair que celui-ci effectue la simulation de *nbobs* valeurs pour une loi définie comme suit : on fait la somme du carré de *nbtos* valeurs d'une loi uniforme réelle continue sur [0..1]. Rien dans le cours n'indique de quoi il s'agit. Voici cependant un exemple d'éxécution avec les paramètres `5 3 1` (sans l'affichage de fin avec les classes) :

```
1  1   0.8407723696144608   0.7068981775071155
1  2   0.1370950189583970   0.7256932217303187
1  3   0.7254659310023159   1.2519940387753756
                   somme   1.2519940387753756  classe  1
```

```
2  1   0.8000965749858006   0.6401545293040087
2  2   0.3777321982949924   0.7828361429327761
2  3   0.1058804335728914   0.7940468091463596
                   somme    0.7940468091463596  classe  0

3  1   0.0288798535332235   8.340459401004437E-4
3  2   0.0719946632447126   0.0060172774758200
3  3   0.7631688821801924   0.5884440202039845
                   somme    0.5884440202039845  classe  0

4  1   0.2279294229026123   0.0519518218247179
4  2   0.5971620453488992   0.4085543302299985
4  3   0.8429238758059742   1.1190749906337638
                   somme    1.1190749906337638  classe  1

5  1   0.9507565588424207   0.9039380341818811
5  2   0.5747689722463066   1.2342974056389571
5  3   0.4152608988628521   1.4067390197633412
                   somme    1.4067390197633414  classe  1
```

Dans le cours, la seule loi qui ressemble à ce genre de calculs est la somme des carrés de n lois normales indépendantes, nommé loi du χ^2, ce que semble suggérer la fonction `valeurLoiNormale()` de l'algorithme. Pour que ce programme simule le tirage de valeurs du χ^2 à *nbtos* degrés de liberté pour un effectif total de *nbobs* valeurs il faut remplacer la ligne

```
x =  Math.random() ;
```

par la ligne

```
x =  generateur.nextGaussian();
```

en prenant soin d'importer la classe

```
import java.util.Random;
```

et d'instancier le générateur avec

```
Random generateur = new Random();
```

en début de programme principal.

Notre programme pourra alors ensuite calculer dans quelle classe il faut ranger la somme obtenue à l'aide de la boucle

```
while( (som>haut) ) { ...
```

avant d'afficher les différentes classes et effectifs correspondants. La classe 0 doit alors être incluse dans les affichages pour les valeurs inférieure à $href$.

On trouvera sur la page suivante le programme *Java* complet et correct qui simule la loi du χ^2 avec gestion des paramètres, initialisation des tableaux, déclaration des variables, utilisation de la classe 0...

```
//#-#   Fichier simj.java issu de galg -a simj.alg -o java
//#-#  ================================================
//#-#      11/12/2001 9:22.55

import java.io.* ;
import java.util.Random;

class simjdet {

////////////////////////////////////////////////////////////////////////
////////////////////////////////////////////////////////////////////////

public static String formatEntier(int nombre, int longueur) {

  String chen = Integer.toString(nombre) ;
  while (chen.length() < longueur) { chen = " " + chen ; } ;
  return chen ;

} // fin de format

////////////////////////////////////////////////////////////////////////

static int valeurEntiere(String chen) {

  int valeur_entiere = 100 ;

  try {  valeur_entiere = Integer.valueOf(chen).intValue() ; }
  catch (NumberFormatException erreurConv) {
         System.out.println(" Erreur de conversion en entier ") ;
         System.exit(-1) ;
  } // fin de catch

  return valeur_entiere ;

} // fin de valeurEntiere

////////////////////////////////////////////////////////////////////////
```

```
static double valeurReelle(String chen) {

  double valeur_reelle = -999.99 ;

  try {  valeur_reelle = Double.valueOf(chen).doubleValue() ; }
  catch (NumberFormatException erreurConv) {
         System.out.println(" Erreur de conversion en réel ") ;
         System.exit(-1) ;
  } // fin de catch

  return valeur_reelle ;

} // fin de valeurReellepublic static void main(String args[]) {

////////////////////////////////////////////////////////////////////////
//                                                                    //
//                                                                    //
//            programme principal                                     //
//                                                                    //
//                                                                    //
////////////////////////////////////////////////////////////////////////

public static void main(String args[]) {

int     mclass, iobs, itos, clas ;
double  haut, x,som ;
int[]   tclass ;


int     nbobs = valeurEntiere(args[0]) ;
int     nbtos = valeurEntiere(args[1]) ;
double  href  = valeurReelle(args[2]) ;

tclass = new int[50]  ;

for (int indb=0; indb<40; indb++) {
     tclass[indb] = 0 ;
} ; // indb de1a nbElt-1
```

```
Random generateur = new Random();

    mclass = 0 ;
     iobs = 1 ;
     while ((iobs <= nbobs)) {
       som = 0 ;
       itos = 1 ;
       while ((itos <= nbtos)) {
         x =  generateur.nextGaussian();
         som = som + x * x ;
         itos = itos + 1 ;
       } ; // sur itos
       haut = href ;
       clas = 0 ;
       while ((som > haut)) {
          haut = haut + href ;
          clas = clas + 1 ;
       } ; // sur som>haut
       tclass[clas] = tclass[clas] + 1 ;
       if (clas > mclass) {
          mclass = clas ;
       } ; // clas > mclass
       iobs = iobs + 1 ;
     } ; // sur iobs

     for (int indi=0; indi<=mclass; indi++) {
        System.out.println(formatEntier( indi, 3 )
            +" : "+formatEntier(tclass[indi], 5 )) ;

     } ; // indi


} // fin de la méthode main()
} // fin de la classe simj

//#-#   Fin de traduction pour simj.java via de galg -a simj.alg -o java
```

Voici donc un exemple d'éxécution du programme correct avec les paramètres `5 3 1`

```
1  1    1.579357188390926   2.494369128522093
1  2   -0.627857389977887   2.888574030671938
1  3   -1.136528630051704   4.180271357599141
                   somme   4.180271357599141  classe  4

2  1   -0.156077312518737    0.024360127483071
2  2    3.0067416290908193   9.064855351590785
2  3   -0.9860328474210595  10.037116127784067
                    somme  10.037116127784067 classe 10

3  1    2.248465348970061    5.055596425519061
3  2    1.365134478580089    6.919188570127194
3  3   -1.635628459310937    9.594469027035062
                   somme    9.594469027035062 classe  9

4  1    0.3960156989258405   0.156828433795721
4  2    1.9842817794513368   4.094202614058285
4  3    0.43028727303004666  4.279349751389919
                  somme   4.279349751389919  classe  4
```

et deux exemples de résultat pour les classes d'effectif si on utilise les paramètres `100 15 10`

```
 0 :    15
 1 :    71
 2 :    13
 3 :     1

 0 :    26
 1 :    60
 2 :    12
 3 :     2
```

Il faut se rappeler que le troisième paramètre est la hauteur de classe. Ainsi, il est plus immédiat de comprendre l'affichage pour les paramètres `100 15 1` à savoir

```
 0 :     0
 1 :     0
 2 :     0
 3 :     0
 4 :     2
 5 :     0
 6 :     5
 7 :     4
 8 :     3
 9 :     8
10 :     9
11 :     4
12 :    11
13 :     7
14 :     6
15 :     5
16 :    10
17 :     3
18 :     9
19 :     4
20 :     2
21 :     1
22 :     0
23 :     1
24 :     0
25 :     1
26 :     0
27 :     1
28 :     1
29 :     3
```

T.D. 10 :

1. Table de la loi normale $\mathcal{N}(0,1)$; d'où vient 1.96 ?
2. Approximation de $\mathcal{B}(n,p)$ par $\mathcal{N}(0,1)$
3. Approximation de $\mathcal{P}(\lambda)$ par $\mathcal{N}(0,1)$
4. Approximation de $\mathcal{B}(n,p)$ par $\mathcal{P}(\lambda)$
5. Saturation d'un concentrateur
6. Découpage en classes d'une variable quantitative
7. Algorithme de la loi hypergéométrique
8. Algorithme de m, σ

1. Table de la loi $\mathcal{N}(0,1)$; d'où vient 1.96 ?

Enoncé

Soit F la fonction de répartition de la loi normale unitaire (centrale réduite) c'est à dire la fonction de répartition de la variable aléatoire $U = \mathcal{N}(0,1)$:

$$F(u) = p(\text{"}U < u\text{"}) = p(\text{"}\mathcal{N}(0,1) < u\text{"}) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}.dt$$

Soit a un nombre strictement positif. Exprimer $F(-a)$ en fonction de $F(a)$; en déduire les valeurs de $F(1.23)$ et $F(-1.23)$ à l'aide d'une table de la fonction de répartition de "la" loi normale.

Soit g la fonction définie par $g(u) = p(\text{"}|U| < u\text{"})$. Exprimer $g(u)$ en fonction de $F(u)$.

Résoudre ensuite l'équation $g(u) = 0.95$ à l'aide de la table. Que peut-on en conclure sur la valeur 1.96 ? Comment la trouve-t-on avec *Rstat* et *Excel* ?

Solution

La densité de U est paire donc $F(-a) = p(\text{"}X < -a\text{"}) = p(\text{"}X > a\text{"})$ et en utilisant la condition opposée : $F(-a) = 1 - p(\text{"}X \leqslant a\text{"}) = 1 - p(\text{"}X < a\text{"})$ soit finalement $F(-a) = 1 - F(a)$. C'est pourquoi la table ne donne que les valeurs de $F(u)$ pour u positif.

1.23 se décompose en 1.2 + 0.03 et on lit donc dans la table à l'intersection de la ligne pour 1.2 et de la colonne pour 0.03 la valeur `0.8907` ; on en déduit que $F(1.23) = 0.8907$ et que $F(-1.23)$ vaut 1 - 0.8907 soit 0.1093.

L'intervalle "$x_1 \leqslant X \leqslant x_2$" peut s'écrire comme la différence de l'intervalle "$X \leqslant x_2$" et de "$X \leqslant x_1$" donc $p(\text{"}x_1 \leqslant X \leqslant x_2\text{"}) = F(x_2)$ - $F(x_1)$. Si on prend $x_1 = -a$ et $x_2 = a$ alors $p(\text{"}|X| \leqslant a\text{"}) = p(\text{" } - a \leqslant X \leqslant +a\text{"})$ soit encore $p(\text{"}|X| \leqslant a\text{"}) = F(a)$ - $F(-a) = 2F(a) - 1$ d'où avec les notations proposées $g(u) = 2F(u) - 1$.

$g(u) = 0.95$ équivaut à $F(u) = 1.950/2 = 0.975$; dans la table, c'est la valeur `1.96` qui vérifie $F(u) = 0.975$ d'où son utilisation lors de calculs à 5 %.

Avec *Rstat*, on écrit *qnorm*(0.975) pour avoir la réponse, à savoir `1.959964` ; de même *pnorm*(1.23) renvoie `0.8906514`.

Sous *Excel*, avec = loi.normale.standard.inverse(0,975) on obtient `1,959962787` et =loi.normale.standard(1,23) renvoie `0,890651383`.

2. Approximation de $\mathcal{B}(n,p)$ par $\mathcal{N}(0,1)$

Enoncé

Soit $X = \mathcal{B}(15, 0.3)$. Effectuer un calcul direct de $p(\text{"}X \in [3,6]\text{"})$. Calculer cette même probabilité de façon approchée en utilisant une table de la fonction de répartition de la loi normale puis avec *Rstat*.

Solution

La valeur exacte est $\displaystyle\sum_{k=3}^{6} C_{15}^{k}\ 0.3^{k}\ (1-0.3)^{15-k}$

Cette valeur est assez difficile à calculer avec une petite calculatrice, mais avec une calculatrice sophistiquée, ou avec *Maple* sur ordinateur, on trouve 0.7419.

On a bien sûr $m_X = 15 \times 0.3 = 4.5$, $V_X = m_X \times (1-0.3) = 3.15$ et donc $\sigma_X = \sqrt{V_X} = 1.7748$. Si on utilise la loi normale, on assimile "$X \in [\![\, 3,6\,]\!]$" à "$X \in [2.5, 6.5]$" soit "$U \in [(2.5-m_X)/\sigma_X, (6.5+m_X)/\sigma_X]$" c'est à dire, tous calculs faits, à "$U \in [-1.126, +1.126]$". Utilisant la fonction de répartition F_U de la loi normale $U = \mathcal{N}(0,1)$ puisque

$$p(\text{"}U \in [-a,a]\text{"}) = p(\text{"}|U| < a\text{"}) = 2F_U(a) - 1$$

il suffit de prendre $a = 1.126$ pour avoir $F_U(0.8699]$ donc $p = 0.7398$ soit seulement une erreur de `0,0021` avec une table à 4 décimales.

Pour effectuer ce calcul avec *Rstat*, on écrit

```
k <- 3:6
sum( choose(15,k)*(0.3**k)*((1-0.3)**(15-k)) )
```

et on obtient alors `0.7420297` comme réponse.

3. Approximation de $\mathcal{P}(\lambda)$ par $\mathcal{N}(0,1)$

Enoncé

Soit $X = \mathcal{P}(20)$. Effectuer un calcul direct de $p(\text{"}X \leqslant 10\text{"})$.

Calculer cette même probabilité de façon approchée en utilisant la loi normale et à l'aide de la table. Utiliser enfin *Rstat* pour déterminer sa valeur.

Solution

Le calcul direct donne $\sum_{k=0}^{10} e^{-20}\ 20^k\ /\ k!$ soit 0.01081171882 avec *Maple*.

Le calcul sous *Rstat* s'écrit :

```
k <- 0:10
sum( exp(-20)*(20**k)/factorial(k) )
```

et on obtient `0.01081172` comme résultat.

On a $m_X = 20 = V_X$, $\sigma_X = 4.4721$. Si on assimile "$X \leqslant 10$" à "$X \leqslant 10.5$" alors $p = p(\text{"}(X-m)/\sigma \leqslant (10.5-20)/4.4721\text{"})$ vaut $F_U(-2.1243)$ soit encore $1 - F_U(2.1243) = 1 - 0.09832 = 0.0168$

Cette approximation est mauvaise (`0.017` au lieu de `0.011`) car la condition d'application $\lambda > 30$ n'est pas respectée.

4. Approximation de $\mathcal{B}(n,p)$ par $\mathcal{P}(\lambda)$

Enoncé

Un livre de 1000 pages contient 1500 erreurs. Donner une valeur exacte de la probabilité qu'une page contienne moins de 2 erreurs puis une approximation de cette probabilité en utilisant la loi normale.

Calculer la probabilité de cet événement si on remplace la loi binomiale sous-jacente par une loi de Poisson bien choisie.

Utiliser *Rstat* pour donner les résultats numériques associés.

Solution

Bien sûr, X est la loi $\mathcal{B}(n = 1500, p = 1/1000)$ de moyenne $m = np$ =1.5 et de variance $V = np(1-p)$ =1.4985 soit un écart-type de 1.2241. "$X < 2$" se traduit par l'union disjointe de "$X = 0$" et "$X = 1$" de probabilités respectives $C_n^0\ p^0\ (1-p)^n$ et $C_n^1\ p^1\ (1-p)^{n-1}$ soit 0.22296+0.33478 donc à peu près 0.55774.

L'approximation gaussienne dit que

$$p(\text{"}X < 2\text{"}) = p\left(\text{"}\frac{X-m}{\sigma} \ < \ \frac{2-np}{\sqrt{np(1-p)}}\text{"}\right) \text{ vaut } F_U\left(\frac{1.5-np}{\sqrt{np(1-p)}}\right)$$

soit ici $F_U(0) = 0.5$. L'erreur commise entre le 0.5 d'approximation et le 0.5577 véritable provient des conditions d'applications qui ne sont pas respectées ($n > 30$ est vérifiée mais $np > 5$ ne l'est pas car np vaut 1.5).

Prenant $\lambda = np = 1.5$ et remarquant que les conditions d'application $n > 36$ et $np < 5$ sont respectées, la somme $p(\text{"}X = 0\text{"})+p(\text{"}X = 1\text{"})$ vaut $e^{-\lambda}\lambda^0/0!+ e^{-\lambda}\lambda^1/1!$ soit, puisque $e^{-\lambda} = e^{-1.5} = 0.22313$, la valeur 0.22313+0.33469 ce qui aboutit à la valeur 0.55782 qui elle, est très proche du résultat exact 0.55774.

5. Saturation d'un concentrateur

Enoncé

Un serveur concentrateur dessert 1000 postes via 50 lignes à haut débit. Aux heures de pointe, chaque poste est occupé en moyenne pendant 2.5 secondes. Quelle est la probabilité de saturation de l'ensemble des lignes à un instant donné pendant une minute de pointe ?

Solution

A partir de la variable aléatoire binaire "un poste est occupé", modélisé par $b(p = 2,5/60)$, "le nombre de postes occupés désirant communiquer avec le serveur" correspond alors à la loi Y somme de $n = 1000$ lois $b(p)$ considérées comme indépendantes. C'est donc la loi $\mathcal{B}(n, p) = \mathcal{B}(1000, 0.0417)$.

La saturation correspond à l'évènement "$Y > 50$". Sa probabilité exacte est

$$p(\text{"}Y \geqslant 51\text{"}) = \sum_{k=51}^{1000} C_{1000}^{k}\ p^k\ (1-p)^{1000-k}$$

et elle vaut 0.08413259 (avec *Maple* par exemple). Avec *Rstat* on écrit `sum(dbinom(x=51:1000,size=1000,prob=2.5/60))` et on obtient `0.0841326`.

Si au lieu de la calculer directement on effectue une approximation par la loi normale, puisque $n = 1000 > 36$, $np = 41.66667 > 5$ et $np(1-p) = 39.93056 > 5$, on remplace

$$p(Y > a) = p\left(\text{"}\frac{Y-m}{\sigma} > \frac{a-np}{\sqrt{np(1-p)}}\text{"}\right) \text{ par } \int_a^{+\infty} \frac{1}{\sqrt{2\pi}e^{-t^2/2}}dt$$

soit, pour $a = 50.5$, $p = 1 - F_U(1.397887) = 0.08107356$ ce qui est assez proche du résultat exact.

6. Découpage en classes d'une variable QT

Enoncé

On dispose d'une variable QT comme par exemple le nombre d'hectolitres de champagne importé par pays. Comment découper cette variable en deux classes ? en trois ?

Application numérique :

PAYS	CANADA	NEDERLAND	BELGIQUE	ITALIE	SUISSE	USA	RFA	UK
CHMP	206	3786	7069	8037	9664	10386	12578	13556

On fournit les résultats numériques suivants :

moyenne	8610.3	*hl*
écart-type	4175.2	*hl*
médiane	8850.5	*hl*
33ième centile	7389.4	*hl*
66ième centile	10142.0	*hl*

Solution

Pour découper en deux classes, il suffit de choisir une valeur seuil disons s et de prendre comme éléments de classe 1 les valeurs inférieures à s et comme éléments de classe 2 les valeurs supérieures à s. Si on utilise la médiane pour s alors on a les deux classes qui sont d'effectifs égaux par construction :

```
Classe 1   : CANADA NEDERLAND BELGIQUE ITALIE
Classe 2   : SUISSE USA       RFA      UK
```

Si on utilise la moyenne pour s on trouve ici les mêmes classes car moyenne et médiane sont assez proches.

Si on prend comme critère une forte différence entre deux valeurs consécutives (disons 50 %) alors on trouve 4 classes à savoir :

Liste des 8 valeurs et tableau des différences (seuil 50 %)

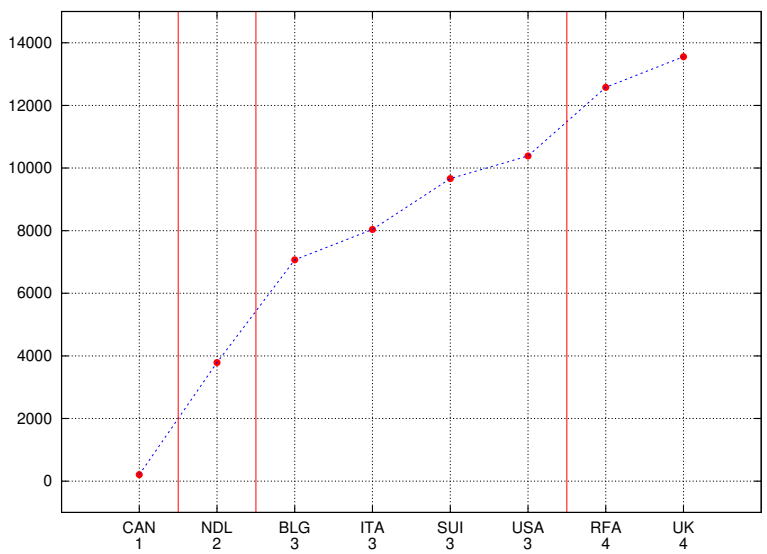
Num	Classe		Nom	Valeur	Nom	Diff.Abs	Diff_%
1	1	*	CAN	206.00	CAN	3580.00	100.0
2	2	*	NDL	3786.00	NDL	3283.00	91.7
3	3	*	BLG	7069.00	BLG	968.00	27.0
4	3		ITA	8037.00	ITA	1627.00	45.4
5	3		SUI	9664.00	SUI	722.00	20.2
6	3		USA	10386.00	USA	2192.00	61.2
7	4	*	RFA	12578.00	RFA	978.00	27.3
8	4		UK	13556.00	UK	0.00	0.0

Description des classes issues de la gestion des différences

Numero	Borne_1	Borne_2	Effectif	Frequ_%	minimum	maximum
1		1996.0	1	12.5	206.00	206.00
2	1996.0	5427.5	1	12.5	3786.00	3786.00
3	5427.5	11482.0	4	50.0	7069.00	10386.00
4	11482.0		2	25.0	12578.00	13556.00

```
Indicateurs numériques de la variable QT

   taille  8  moyenne  8160.3  écart-type  4175.2  cdv =   51  %
   Q1 =  3786.0  ; Q2 = médiane =  8850.5  ; Q3 =  10386.0  ie =  6600.0

Contenu des classes issues de la gestion des différences

 Numero  Eléments

     1        CAN
     2        NDL
     3        BLG   ITA   SUI   USA
     4        RFA   UK
```

Voici ce découpage :



Pour découper en trois classes, il faut définir deux valeurs-seuil disons s_1 et s_2 puis prendre comme éléments de classe 1 les valeurs inférieures à $s1$, comme éléments de classe 2 les valeurs supérieures à $s1$ et inférieures à $s2$, comme éléments de classe 3 les valeurs supérieures à $s2$.

Si on utilise le 33ième centile pour $s1$ et le 66ième centile pour $s2$ alors on a les trois classes qui devraient être d'effectifs égaux par construction (mais ici le faible nombre de valeurs empêche la classe 2 d'avoir le bon effectif) :

```
Classe 1   : CANADA   NEDERLAND  BELGIQUE
Classe 2   : ITALIE   SUISSE
Classe 3   : USA      RFA        UK
```

Si on utilise $m - \sigma$ pour $s1$ soit 4435 hl et $m + \sigma$ième centile pour $s2$ soit 12785 hl alors on a les trois classes

```
Classe 1   : CANADA    NEDERLAND
Classe 2   : BELGIQUE  ITALIE   SUISSE USA RFA
Classe 3   : UK
```

7. Algorithme de la loi hypergéométrique

Enoncé

Soit X la *v.a.* "loi hypergéométrique" $\mathcal{H}(N, D, n)$: on tire n boules sans remise dans N boules dont D sont "spéciales" ; X est la loi "nombre de boules spéciales". Démontrer que X prend les valeurs entières $k = a, a+1, a+2 \ldots b$ où $a = max(0, D + n - N)$ et $b = min(D, n)$ et que la valeur k a pour probabilité $p_k = C_D^k.C_{N-D}^{n-k}/C_N^n$. Expliciter les valeurs de k et p_k pour $N = 6$, $D = 3$ et $n = 2$.

Donner, en respectant la syntaxe algorithmique du cours, un algorithme qui calcule la moyenne de X pour n, N et D donnés ; on supposera connue la fonction $C(n, p)$ qui calcule C_n^p ; calculer au passage la somme des probabilités, la moyenne, la l'écart-type et le coefficient de variation de X.

Solution

Si on tire n boules sans remise parmi N boules, il est possible de n'obtenir aucune boule spéciale : 0 est donc une plus petite valeur possible pour a. Toutefois, supposons qu'on tire 10 boules parmi 30 dont 25 sont spéciales ;

comme il n'y a que 5 boules non spéciales, on voit qu'on aura au minimum 5 boules spéciales. Dans le cas général, si $n + D$ est plus grand que N alors on a au moins $n + D - N$ boules spéciales. Finalement, a se calcule donc comme $max(0, D + n - N)$.

De la même façon, clairement si on tire n boules et qu'il y a au maximum D boules spéciales, on ne peut en obtenir plus de D donc D est une plus grande valeur possible pour b. Par contre puisqu'on tire n boules on ne peut obtenir plus de n boules spéciales et donc finalement b est $min(D, n)$.

Le tirage se faisant sans remise en comptant le nombre de boules spéciales obtenues, il est sans ordre et sans répétition donc on peut utiliser des combinaisons pour modéliser les tirages. Tirer x boules parmi y fournit sans ordre et sans répétition fournit C_y^x combinaisons. Obtenir k boules spéciales parmi D implique d'avoir $n - k$ boules non spéciales parmi $N - D$. Tous tirages confondus on tire n boules parmi N donc la probabilité d'obtenir k boules spéciales est :

$$p_k = \frac{C_D^k . C_{N-D}^{n-k}}{C_N^n}$$

Si $N = 6$, $D = 3$ et $n = 2$ alors $a = max(0, D+n-N) = max(0, 3+2-6) = 0$ et $b = min(D, n) = min(3, 2) = 2$.

X prend donc les valeurs 0, 1 et 2.

Leurs probabilités respectives sont

$$\begin{aligned} p_0 &= C_3^0 C_3^2 / C_6^2 = 3/15, \\ p_1 &= C_3^1 C_3^1 / C_6^2 = 9/15, \\ p_2 &= C_3^2 C_3^0 / C_6^2 = 3/15. \end{aligned}$$

Pour le calcul des valeurs, des probabilités, on pourra utiliser l'algorithme de la page suivante où la fonction *cb* correspond aux coefficients du binome : $cb(x, y) = C_y^x$.

```
Algorithme du calcul de H(N,D,n)

# calcul des bornes des valeurs de X

  borne_a ← D+n-N
  si 0 > borne_a alors
     borne_a ← 0
  finsi
  borne_b ← D
  si n < borne_b alors
     borne_b ← n
  finsi

# valeurs de X, somme et somme des carrés

  u  ← 0       # nombre de valeurs
  s  ← 0       # somme des valeurs
  sc ← 0       # somme des carrés des valeurs
  sp ← 0       # somme des probabilités
  de ← cb(nt,N) # dénominateur constant pour k
  di ← N-D     # différence  constante pour k

  pour k de a à b
       p  ← cb(k,D).cb(n-k,di) / de
       u  ← u  + 1
       s  ← s  + k*p
       sc ← sc + k*k*p
       sp ← sp + p
  fin pour k


# moyenne, variance, etc.

  m    ← s/u
  v    ← sc/u - m*m
  e    ← racine(v)
  cdv  ← 100*e/m
```

8. Algorithmes de m, σ et cdv

Enoncé

Soit X un tableau de n valeurs notées $X[1]$, $X[2]$... $X[n]$. Donner un algorithme du calcul de m_X, σ_X et $cdv_X = \sigma_X / m_X$.

Après avoir trouvé au moins deux méthodes pour calculer σ_X quelle est la meilleure façon de calculer σ_X ?

Soit D un tableau de données dont les n lignes sont repérées par un premier indice i et les p colonnes sont repérées par un second indice j : $D[i, j]$ désigne donc la valeur à l'intersection de la ligne i et de la colonne j.

On admettra que le tableau D est structuré de la façon suivante : les colonnes 1 à t de D correspondent à des variables quantitatives et la colonne $t + 1$ contient les codes d'une variable qualitative avec q codes notés $1, 2...q$.

Sachant que le tableau D est déjà constitué, que les valeurs n, p, t, q sont toutes définies et valides, on veut calculer la moyenne $moy[k, z]$, l'écart-type $sig[k, z]$ et le coefficient de variation $cdv[k, z]$ de chacune des sous-populations (ici correspondant au code k) pour chaque variable z. On supposera les tableaux moy, sig et cdv déjà déclarés. On utilisera un tableau $eff[k]$ que l'on remplira avec l'effectif de la population k.

Voici un exemple de tel tableau de données :

QT1	QT2	QT3		QL
1	1	1	|	1
2	2	4	|	1
1	3	25	|	2
3	4	9	|	1
1	5	16	|	2

a) Compléter le tableau des résultats à 0,1 près correspondant aux données présentées.

```
TABLEAU DES RESULTATS
   n = 5 lignes ; p = 3 colonnes ; q = 2 sous-populations

pop.  1  eff[1]  =  3

         moy[1,1] =    2.0 sig[1,1] =    0.8 cdv[1,1] =    0.4
         moy[1,2] =    2.3 sig[1,2] =    1.2 cdv[1,2] =    0.5
         moy[1,3] =    4.7 sig[1,3] =    3.3 cdv[1,3] =    0.7

pop.  2  eff[2]  =  ????

         moy[2,1] =    ??? sig[2,1] =    0.0 cdv[2,1] =    0.0
         moy[2,2] =    4.0 sig[2,2] =    ??? cdv[2,2] =    0.3
         moy[2,3] =   20.5 sig[2,3] =    4.5 cdv[2,3] =    ???
```

b) Ecrire un algorithme qui calcule ces résultats. Vous n'utiliserez aucun autre tableau que eff, moy, sig et cdv. On ne demande aucun affichage.

Solution

Pour calculer la moyenne du vecteur X, on divise la somme des éléments de X par le nombre d'éléments de X. Une simple boucle `POUR` suffit pour calculer la somme, soit l'algorithme

```
# algorithme du calcul de la moyenne de X
# (nbv est le nombre de valeurs de X)

somme <- 0
pour idv de 1 a nbv
     somme <- somme + X[ idv ]
fin_pour idv

moyenne <- somme / nbv
```

Pour calculer l'écart-type, il faut d'abord calculer la variance et c'est là que les ennuis commencent. Nous connaissons deux formules : la première dit qu'on fait la différence entre la moyenne des carrés et le carré de la moyenne, la seconde calcule la moyenne du carré des différences à la moyenne.

Un premier algorithme du calcul de V est donc

```
# un algorithme du calcul de la variance de X
# on connait deja la moyenne m de X

sommeCar <- 0
pour idv de 1 a nbv
     sommeCar <- sommeCar + X[ idv ] * X[ idv ]
fin_pour idv

moyCar  <- sommeCar / nbv
variance <- moyCar - m*m
```

Le second est

```
# un autre algorithme du calcul de la variance de X
# on connait deja la moyenne m de X

sommeCarDif <- 0
pour idv de 1 a nbv
     dif         <- X[ idv ] - m
     sommeCarDif <- sommeCarDif + dif*dif
fin_pour idv

variance <- sommeCarDif / nbv
```

Savoir quel algorithme est le meilleur est difficile à dire : ainsi le premier algorithme est sans doute plus rapide à exécuter car on pourrait regrouper le calcul de `somme` et de `sommeCar` en une seule même boucle `pour`. Par contre l'algorithme 2 est être meilleur en termes de précision si les valeurs de X sont élévées et peu différentes et c'est donc le seul à programmer pour des calculs "robustes".

Prenons par exemple les quatre valeurs $a-3$, $a-1$, $a+1$ et $a+3$. On devrait trouver $m=a$ et $V=5$ (de tête selon l'algorithme 2, de façon moins facile avec l'algorithme 1). Si a est grand, disons par exemple $a=10^8$ alors l'algorithme 1 risque d'arrondir les calculs et de ne pas fournir les bonnes valeurs. En voici la démonstration sous *Excel* :

	A	B	C	D	E
1	a	1000000000			
2	X	999999997	999999999	1000000001	1000000003
3					
4	Méthode 1				
5	X	999999997	999999999	1000000001	1000000003
6	carré	1E+18	1E+18	1E+18	1E+18
7				moyenne	1000000000
8				**variance**	**0**
9					
10	Méthode 2				
11	X	999999997	999999999	1000000001	1000000003
12	diff	3	1	-1	-3
13	carré	9	1	1	9
14				moyenne	1000000000
15				**variance**	**5**
16					
17	fonctions Excel			moyenne	1000000000
18				var.p	0
19					

On remarquera au passage que la fonction `var.p` ne calcule pas le bonne valeur de la variance. *Rstat* fait les mêmes erreurs d'arrondi pour les grands nombres :

```
x <- 10**8+c(-3,-1,1,3)
m <- sum(x)/4
sum(x*x)/4 - m*m
4
sum( (x-m)**2 ) / 4
5
```

Par contre la fonction implémentée en *Rstat* pour calculer `var` est correcte :

```
x <- 10**8+c(-3,-1,1,3)
n <- length(x)
var(x)*(n-1)/n
5
```

Rappelons au passage que la variance calculée en standard par *Rstat* (de même que par *Excel* avec la fonction `var`) est la variance estimée d'une population dont X est un échantillon. Cette variance estimée utilise $n-1$ au lieu de n au niveau du dénominateur et elle est donc toujours plus grande que "notre variance qui est la variance exacte de l'échantillon. On l'obtient en multipliant par par $n/(n-1)$ notre variance.

Après calculs, les valeurs numériques demandées sont

```
pop.  1  eff[1]  =  3

         moy[1,1] =    2.0 sig[1,1] =    0.8 cdv[1,1] =    0.4
         moy[1,2] =    2.3 sig[1,2] =    1.2 cdv[1,2] =    0.5
         moy[1,3] =    4.7 sig[1,3] =    3.3 cdv[1,3] =    0.7

pop.  2  eff[2]  =  2

         moy[2,1] =    1.0 sig[2,1] =    0.0 cdv[2,1] =    0.0
         moy[2,2] =    4.0 sig[2,2] =    1.0 cdv[2,2] =    0.3
         moy[2,3] =   20.5 sig[2,3] =    4.5 cdv[2,3] =    0.2
```

Ainsi `moy[1,1]` est la moyenne des 3 valeurs 1, 2 et 3 situées en colonne 1 du tableau, lignes 1, 2 et 4 car pour ces lignes la dernière colonne vaut 1.

L'algorithme de calcul est à peine plus compliqué que les précédents. Il faut juste ajouter un indice de colonne et une détection de la modalité...

Un algorithme possible est

```
# initialisation des tableaux

pour ind de1a q
  eff[ind] ← 0
  pour jnd de1a t
     moy[ind,jnd]  ← 0
     sig[ind,jnd]  ← 0
  finpour jnd de1a t
finpour ind de1a q

# calcul des effectifs, des sommes et de leur carre

pour lig de1a n
  pour col de1a t
       ind          ← D[lig,t+1]
       si col = 1 alors  eff[ind] ← eff[ind] + 1 finsi
       valr         ← D[lig,col]
       moy[ind,col] ← moy[ind,col] + valr
       sig[ind,col] ← sig[ind,col] + valr*valr
  finpour col de1a t
finpour lig de1a n

# calcul des moyennes etc.

pour ind de1a q
  pour jnd de1a t
     moy[ind,jnd]  ← moy[ind,jnd]  / eff[ind]
     sig[ind,jnd]  ← sig[ind,jnd]  / eff[ind]
     sig[ind,jnd]  ← sig[ind,jnd]  - moy[ind,jnd]*moy[ind,jnd]
     sig[ind,jnd]  ← racine( sig[ind,jnd]  )
     si moy[ind,jnd]<> 0
        alors cdv[ind,jnd]  ← abs( sig[ind,jnd] ) / moy[ind,jnd]
        sinon cdv[ind,jnd]  ← -1
     finsi moy[ind,jnd]<> 0
  finpour jnd de1a t
finpour col de1a t
```