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.

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.

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...

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 $p(B) = 0,5$.

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 ?

6. Obtenir un as aux cartes

Enoncé

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

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 ?

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 $n = 17$.

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 ?

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 ?

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 ?

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 ?

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.

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.

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 $(1+x)^{n_1+n_2} = (1+x)^{n_1}.(1+x)^{n_2}$

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.

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.

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 $X/\sigma(X) - m(X)$.

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 $V(\mathsf{U})$.

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.

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 ?

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 $\mathcal{B}(n, p)$.

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 ?

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 ?

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" ?

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.

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.*

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}.

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.

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 $Rstat$?

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 ?

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...

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* ?

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 $\mathsf{T}=a.\mathsf{D}+b$.

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

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 ?

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.

Indication : On fera apparaitre les variances et la covariance.

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
```

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.

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.

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.

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	??

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

5. Transitivité de la relation $\mathbf{Y = aX + b}$

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 ?

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.

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.

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 $\alpha = 5$ %.

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.

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.

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

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 ?

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 ?

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 $\Sigma\ a_i t_i = 165$, $\Sigma\ a_i t_i^2 = 1495$, $\Sigma\ b_i t_i = 201$, $\Sigma\ b_i t_i^2 = 1671$.

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.

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.

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
```

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
```

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* ?

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*.

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.

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.

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 ?

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*

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.

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.