Gilles HUNAULT

2000

Probabilités et Statistiques

(Travaux Dirigés)

Université d'Angers

T.D. 1 : Probabilités élémentaires

1. Rappels de la définition d'une probabilité
2. Calcul de $p(A \cap \overline{B}) \cup (B \cap \overline{A})$ en fonction de $\alpha = p(A \cap B)$.
3. Détermination de A, B, p_1, p_2 tels que A et B sont indépendants pour p_1 mais pas pour p_2.
4. Etude de l'obtention d'un as lorsqu'on tire n cartes.
5. Etude des $p_i = C_n^i 0.2^i 0.8^{n-i}$ pour i de 1 à n.
6. Comptage des fichiers en $ghFs$.
7. Comparaison des chances de gain au tiercé avec n chevaux et au quarté avec $n+1$ chevaux.
8. Formule de $Bayes$ pour les variables typées en $ghProlog$.

1. Définition de la notion de probabilité

Enoncé

Qu'est-ce qu'une probabilité ?

Solution

Une probabilité sur E est une **fonction** définie sur $\mathcal{P}(E)$, à valeurs dans $[0,1]$ et **additive** pour des ensembles disjoints. Si (E_i) est un système complet d'évènements de E, donner chacun des $p(E_i)$ suffit à définir p. On doit alors avoir $p(E_i) > 0$ et $\sum p(E_i) = 1$.

Par exemple, pour i de 1 à n, la valeur $p_i = p(E_i) = i/(\ n(n+1)(n+3)\ )$ ne définit pas une probabilité car pour $n = 2$, $p_1 + p_2 = 3/10 < 1$.

2. Calculs 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.

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 de $A = A \cap E = A \cap (B \sqcup \overline{B}) = (A \cap B) \sqcup (A \cap \overline{B})$ on déduit $p(A \cap \overline{B}) = p(A) - p(A \cap B)$ et de façon symétrique $p(B \cap \overline{A}) = p(B) - p(B \cap A)$ donc $p(\ (A \cap \overline{B}) \cup (B \cap \overline{A})\ ) = p(A \cap \overline{B}) + p(B \cap \overline{A}) = p(A) + p(B) - 2p(A \cap B)$ car $A \cap \overline{B}$ et $B \cap \overline{A}$ sont disjoints.

Avec des valeurs numériques, la probabilité α cherchée doit donc vérifier $p = p(A) + p(B) - 2p(\alpha)$ ce qui impose des conditions sur α pour avoir $0 \leq p \leq 1$. Ainsi, pour $p(A) = 0,2$, $p(B) = 0,5$, on en déduit, que $-0,15 \leq \alpha \leq 0,35$. Mais $0 \leq \alpha \leq 1$ et $\alpha \leq p(A)$ donc, toutes inégalités confondues, $0 \leq \alpha \leq 0.2$

3. Indépendance relative à une probabilité

Enoncé

Montrer que A et B sont indépendants fait référence à une probabilité, que A et B peuvent être dépendants pour p_1, indépendants pour p_2.

Solution

Prenons le jet d'un dé avec A ="pair" et $B = \{5, 6\}$. Puisque $A \cap B = \{6\}$, A et B ne sont pas incompatibles. Soit p_1 la probabilité équirépartie, alors $p_1(A \cap B) = 1/6$ et comme $p_1(A) = 1/2$, $p_1(B) = 1/3$, $p_1(A \cap B) = p_1(A).p_1(B) = 1/3$ donc A et B sont indépendants pour p_1.

Définissons maintenant p_2 par $p_2(6) = 1/12$, $p_2(2) = p_2(4) = p_2(5) = 1/6$. Alors $p_2(A \cap B) = 1/12$ et $p_2(A) = 5/12$, $p_2(B) = 3/12$, $p_2(A \cap B) \neq p_2(A).p_1(B) = 1/3$ donc A et B ne sont pas indépendants pour p_2.

4. Tirer un as...

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 au moins 1 as si on tire deux cartes", il y a ambiguité 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).

5. Sont-ce des valeurs de probabilité ?

Enoncé

Les n nombres $C_n^i a^i b^{n-i}$ pour i de 1 à n sont-ils des valeurs de probabilité ? pour un système complet ?

Solution

Si on connait la formule du binome $(a+b)^n = \sum_{i=0}^{n} C_n^i a^i b^{n-i}$, les p_i correspondent aux termes pour i de 1 à n seulement. Ce sont donc bien des nombres positifs inférieurs à 1 mais leur somme fait seulement $1 - 0.8^n$ donc ces nombres ne définissent pas une probabilité.

6. Nombres de 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.

7. Comptons, vive les boules !

ENONCÉ

On dispose de n boules numérotées de 1 à n, nommées $b(i)$, i variant de 1 à n. On dispose aussi de N boules de couleurs nommées $c(j,k)$. Il y a p couleurs différentes et $q(j)$ boules de couleur j, j variant de 1 à p. $c(j,k)$ est la $k$ième boule de couleur j, k variant de 1 à $q(j)$.

a) 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.

b) Au tiercé, une course met en présence n chevaux (classiquement une à deux dizaines, mais parfois moins). Chaque cheval comporte un numéro de 1 à n (on omet ici la notion de "non-partant"). A l'arrivée, on ne tient compte que des numéros des trois premiers chevaux. On dit qu'on a gagné "dans l'ordre" si on parié (et si on a acheté le ticket correspondant) sur ces trois numéros exactement dans l'ordre d'arrivée. On dit qu'on a gagné "dans le désordre" si on a parié sur ces trois numéros mais dans un ordre quelconque (et sauf l'ordre exact).

   Quel type de boules peut-on utiliser pour structurer les événements associés aux résultats du tiercé ? Quelle est la probabilité de gagner dans l'ordre et dans le désordre pour n=17 chevaux ? Reprendre avec n=17 chevaux au quarté (on s'intéresse alors aux quatre premiers chevaux).

On ne demande pour les deux questions suivantes aucun calcul numérique explicite (même s'ils sont effectués "au brouillon").

c) 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 ?

d) Vaut-il mieux tenter de gagner (dans l'ordre, dans le désordre), au tiercé avec n chevaux ou au quarté avec $n + 1$ chevaux ?

SOLUTION

Le loto utilise 49 boules numérotées de 1 à 49. On utilisera donc $b(49)$. 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 $0,715.10^{-}7$.

Pour le tiercé dans le désordre, on peut utiliser la même structuration mais avec des triplets (3-uplets) où les $b(n)$ 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 5 / 4080. 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 $1/(n.(n-1).(n-2))$. 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.

Jouer avec 8 boules et 60 numéros diminuerait les chances de gagner car les coefficients du binome croissent très rapidement. De même, rajouter un cheval ne fait que diminuer la probabilité de gagner car on multiplie par 1/(n-3).

Numériquement, le rapport des chances est de 1/183 au loto si on passe de 49 et 6 à 60 et 8, pour le tiercé ce rapport est de 1/18 dans l'ordre et de 1/4.5 dans le désordre si on passe de 17 à 18 chevaux.

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

8. Typage en *ghProlog*

Enoncé

En *ghProlog* (langage qui n'existe pas encore), une variable est soit libre, soit liée, soit indécise. On dit alors qu'elle est typée. Dans un programme de ce langage, la probabilité qu'une variable soit typée comme variable libre est de 0.45, la probabilité qu'une variable soit typée comme variable liée est de 0.45, la probabilité qu'une variable soit typée comme variable indécise est de 0.1. La probabilité qu'une variable soit correctement typée (par le programmeur) sachant qu'il s'agit d'une variable libre est de 0.90. La probabilité qu'une variable soit correctement typée (par le programmeur) sachant qu'il s'agit

d'une variable liée est de 0.90. La probabilité qu'une variable soit correctement typée (par le programmeur) sachant qu'il s'agit d'une variable indécise est de 0.10.

Quelle est la probabilité qu'une variable soit liée sachant qu'elle est correctement typée ?

SOLUTION

Soit X l'évènement correspondant au typage d'une variable, soient A, B et C les évènements associés respectivement au typage libre, lié et indécis. Comme il n'y a pas d'autre typage possible : $X = A \cup B \cup C$ et cette union est disjointe. On sait par l'énoncé que $p(A) = 0{,}45$, $p(B) = 0{,}45$ et $p(C) = 0{,}10$. Soit maintenant Y l'évènement correspondant à un typage correct. L'énoncé dit aussi que $p(Y|A) = 0,9$, $p(Y|B) = 0,9$ et $p(Y|C) = 0,1$. La question posée est celle du calcul de $p(B|Y)$.

Décomposons Y en l'union disjointe $(Y \cap A) \cup (Y \cap B) \cup (Y \cap C)$. Comme $p(Y \cap A) = p(Y|A).p(A)$, $p(Y \cap B) = p(Y|B).p(B)$ et $p(Y \cap C) = p(Y|C).p(C)$, on en déduit $p(Y \cap B) = 0{,}9\text{x}0{,}45 = 0{,}405$. De même la valeur de $p(Y)$ est 0,9x0,45 + 0,9x0,45 + 0,1x0,1 donc $p(Y) = 0{,}82$.

Finalement, $p(B|Y) = 0{,}405/0{,}820$ soit $p(B|Y) = 0{,}494$.

T.D. 2 : Dénombrements et Variables alétaoires

1. Probabilités conditionnelles avec 2 dés.
2. Probabilités de Fiabilité d'un réseau téléphonique.
3. Probabilités conditionnelles chez un programmeur
4. Formule de Bayes "pour les jeunes"
5. Comparaison de probabilités et formule de *Poincaré* généralisée
6. Calcul du nombre d'applications, injections...
7. Démonstration de $C_{2n}^{n} = \sum_{k=0}^{n} (C_x^k)^2$ via $C_{n_1+n_2}^{k} = \sum_{k_1+k_2=k} C_{n_1}^{k_1} C_{n_2}^{k_2}$.

1. 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 troisième 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 $(2/3)^2.1/2 + (1/3)^2.1/2 = 5/18$.

* 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.

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".

2. Fiabilité d'un réseau téléphonique

Enoncé

Une compagnie de messagerie téléphonique assure que son réseau en fibre optique est si fiable que lorsqu'on compose correctement le numéro, on a 19 chances sur 20 d'obtenir la tonalité chez le correspondant désiré.

a) quelle est la probabilité d'obtenir pour la première fois la tonalité chez le correspondant désiré au bout de 3 essais au plus si à chaque fois on compose le bon numéro ?

b) toutefois, on estime que pour le premier appel, on a 1 chance sur 10 que l'utilisateur fasse une fausse numérotation, que pour la deuxième fois il y ait 1 chance sur 100 de fausse numérotation et que pour la troisième fois il y ait 1 chance sur 1000 de fausse numérotation. Quelle est la probabilité d'obtenir le correspondant au bout du fil en 3 essais au plus compte-tenu de la fiabilité du réseau ?

Solution

Dire "3 essais au plus" (évènement X) signifie "1 essai exactement" (disons A) ou "2 essais exactement" (B) ou "3 essais exactement" (C) et ce, de façon disjointe (exclusive).

Soit S_i l'évènement "*on a la tonalité chez le bon correspondant à l'essai* i (dont la probabilité, constante, vaut $p = 19/20$ pour tout i), alors $A = S_1$, $B = \overline{S_1} \cap S_2$, $C = \overline{S_2} \cap S_3$. On peut donc écrire

$$\begin{aligned} p(X) &= p(A \sqcup B \sqcup C) \\ &= p(A) + p(B) + p(C) \\ &= p + (1-p)p + (1-p)^2 p \\ &= 19/20 + 1/20.19/20 + (1/20)^2.19/20 \\ &= 7999/8000 \simeq 99.98\ \% \end{aligned}$$

Il y a un piège dans l'énoncé ! Obtenir la tonalité chez le correspondant, ce n'est pas obtenir le correspondant. Notons B_i "on réalise une bonne numérotation au i-ième coup". On va essayer de calculer la probabilité de l'évènement "on obtient la tonalité chez le correspondant" seulement. On obtient cette tonalité si on compose le bon numéro et si le réseau est fiable. L'évènement A doit alors s'érire $A = S_1 \cap B_1$ et B devient $B = (\overline{S_1} \cap B_1 \sqcup \overline{B_1}) \cap B_2 \cap S_2$. Enfin, C devient $Y \cap B_3 \cap S_3$ où $Y = (\overline{S_1} \cap B_1 \sqcup \overline{B_1}) \cap (\overline{S_1} \cap B_2 \sqcup \overline{B_2})$. Numériquement, $p(A) = p(1-q_1) = 19/20.9/10$, $p(B) = ((1-p)(1-q_1) + q_1)(1-q_2)p = (1/20.9/10 + 1/10).99/100.19/20$, $p(C) = (1/20.9/10 + 1/10).(1/20.99/100 + 1/100).999/1000.19/20$. Le total de ces trois probabilités donne 99.96 % soit un résultat très peu différent du précédent.

3. Probabilités conditionnelles chez un programmeur

Enoncé

Un programmeur en assembleur estime qu'il a une probabilité p d'avoir mis un cycle de trop dans un calcul de décalage de registres, calcul qui se décompose en 8 sous-programmes. Ce programmeur a relu en détail les 7 premiers premiers sous-programmes et il n'a pas trouvé de cycle en trop. Quelle est la probabilité $f(p)$ que le cycle de trop soit dans le 8-ième sous-programme ?

Solution

Notons S_i l'évènement *"l'erreur est dans le sous-programme numéro i* où i varie de 1 à 8. Le nombre $f(p)$ demandé correspond à la probabilité conditionnelle $p(S_8|_T)$ si T désigne l'évènement $\bigcap_{j=1}^{j=7} \overline{S}_j$. Comme l'énoncé ne précise rien, nous supposerons les sous-programmes équi-probables.

Les sous-programmes définissent un système complet d'évènements, donc

$$p(T) = 1 - p(\bigsqcup_{j=1}^{j=7} S_j) = 1 - 7.(p/8)$$

d'où $f(p) = (p/8)/(1 - 7p/8)$ soit $f(p) = p/(8 - 7p)$.

4. 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

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

6.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.

5. Formule de *Sylvester*

Enoncé

Soient $(E, \mathcal{T}, p)$ un espace probabilisé, A, B, C des sous-ensembles de E. On note $\overline{X}$ le complémentaire de X dans E et on suppose donnés n sous-ensembles de E notés A_i pour i de 1 à n.

a) Comparer à l'aide de $\leq$ les nombres $p(A \cap B)$ et $p(A) - p(\overline{B})$.

b) Comparer à l'aide de $\leq$ les nombres $p(A \cap B \cap C)$

et $1 - p(\overline{A}) - p(\overline{B}) - p(\overline{C})$.

c) Comparer à l'aide de $\leq$ les nombres $p(\bigcap_{i=1}^{n} A_i)$ et $1 - \sum_{i=1}^{n} p(\overline{A}_i)$.

d) Peut-on en déduire $p(\bigcup_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} p(A_i)$?

Solution

Par définition d'une probabilité p, on a $p(X) \leq 1$ pour tout X. Prenant $X = A \cup B$ et avec la formule classique $p(A \cup B) = p(A) + p(B) - p(A \cap B)$ on en déduit aisément $p(A) + p(B) - p(A \cap B) \leq 1$ soit, en passant $p(A \cap B)$ à droite et en remplaçant $p(B) - 1$ par $-p(\overline{B})$:

$$p(A) - p(\overline{B}) \leq p(A \cap B)$$

Cette formule s'écrit aussi $1 - p(\overline{A}) - p(\overline{B}) \leq p(A \cap B)$.

Si maintenant on écrit $p(A \cap B \cap C)$ sous la forme $p(\ (A \cap B) \cap C\ )$, avec le résultat précédent, on a $p(A \cap B) - p(\overline{C}) \leq p(A \cap B \cap C)$ et en réapplicant la même formule : $p(A) - p(\overline{B}) - p(\overline{C}) \leq p(A \cap B \cap C)$ soit finalement, en remplaçant $p(A)$ par $1 - p(\overline{A})$:

$$1 - p(\overline{A}) - p(\overline{B}) - p(\overline{C}) \leq p(A \cap B \cap C)$$

Puisque $1 - p(\overline{A_1}) - p(\overline{A_2}) \leq p(A_1 \cap A_2)$, supposant $1 - \sum_{i=1}^{n-1} p(\overline{A}_i) \leq p(\bigcap_{i=1}^{n-1} A_i)$

$$\begin{aligned}
\text{alors } p(\bigcap_{i=1}^{n} A_i) &= p((\bigcap_{i=1}^{n-1} A_i) \cap A_n) \\
&\geq p(\bigcap_{i=1}^{n-1} A_i) - p(\overline{A_n}) \\
&\geq 1 - \sum_{i=1}^{n-1} p(\overline{A}_i) - p(\overline{A_n}) \\
&= 1 - \sum_{i=1}^{n} p(\overline{A}_i)
\end{aligned}$$

La relation

$$1 - \sum_{i=1}^{n} p(\overline{A}_i) \leq p(\bigcap_{i=1}^{n} A_i)$$

est donc démontrée par récurence. On ne peut bien sûr pas en déduire $p(\bigcup_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} p(A_i)$ puisque c'est la relation inverse qui est vraie déjà à l'ordre deux : $p(A \cup B) = p(A) + p(B) - p(A \cap B) \leq p(A) + p(A)$.

On montre par récurence en écrivant $\bigcup_{i=1}^{n} A_i = \left(\bigcup_{i=1}^{n-1} A_i\right) \cup A_n$ que

$$p(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} p(A_i)$$

La formule dite de *Sylvester* est

$$\left| \bigcup_{i=1}^{n} A_i \right| = \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1 < \ldots < i_k} \left| \bigcap A_{i_k} \right|$$

où $|X|$ désigne le cardinal de l'ensemble X. Elle se démontre aussi par récurence. On en déduit la formule de *Poincaré* généralisée:

$$p\left(\bigcup_{i=1}^{n} A_i \right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1 < \ldots < i_k} p\left(\bigcap A_{i_k} \right)$$

Enoncé

Soit E un ensemble à n éléments et F un ensemble p éléments. Combien y a-t-il d'applications de E dans F ? d'injections, de surjections, de bijections

? Comment seraient les produits cartésiens correspondants pour n et p de 1 à 10 ?

Solution

Il y a bien sûr p^n applications dont $A(p,n) = p!/(p-n)!$ injections , $n!$ bijections et $\sum_{k=0}^{p}(-1)^k C_p^k (p-k)^n$ surjections.

Le nombre de surjections $surj(n,p)$ vérifie la relation

$$surj(n,p) = p.(\ surj(n-1,p) + surj(n-1,p-1)\ )$$

Voici quelques exemples numériques de valeurs.

```
Nombre de fonctions de E dans F avec card(E) = n et card(F) = p

   n       p    appli.   injec.   surjec.  bijections
   5       1         1        0         1           0
   5       2        32        0        30           0
   5       3       243        0       150           0
   5       4      1024        0       240           0
   5       5      3125      120       120         120
   5       6      7776      720         0           0
   5       7     16807     2520         0           0
   5       8     32768     6720         0           0

   n       p    appli.   injec.   surjec.  bijections
   1       5         5        5         0           0
   2       5        25       20         0           0
   3       5       125       60         0           0
   4       5       625      120         0           0
   5       5      3125      120       120         120
   6       5     15625        0      1800           0
   7       5     78125        0     16800           0
   8       5    390625        0    126000           0


  Applications
```

n	1	2	3	4	5
1	1	2	3	4	5
2	1	4	9	16	25
3	1	8	27	64	125
4	1	16	81	256	625
5	1	32	243	1024	3125

Injections

n	1	2	3	4	5
1	1	2	3	4	5
2	0	2	6	12	20
3	0	0	6	24	60
4	0	0	0	24	120
5	0	0	0	0	120

Surjections

n	1	2	3	4	5
1	1	0	0	0	0
2	1	2	0	0	0
3	1	6	6	0	0
4	1	14	36	24	0
5	1	30	150	240	120

Bijections

n	1	2	3	4	5
1	1	0	0	0	0
2	0	2	0	0	0
3	0	0	6	0	0
4	0	0	0	24	0
5	0	0	0	0	120

6. Un calcul combinatoire

Enoncé

Montrer que $C_{2n}^n = \sum_{k=0}^{n} (C_\alpha^k)^2$ pour α bien choisi dépendant de n.

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 $k_1 + k_2 = k$. Le coefficient de x^k est donc $C_{n_1+n_2}^k = \sum_{k_1+k_2=k} C_{n_1}^{k_1} C_{n_2}^{k_2}$. 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 $C_{2n}^n = \sum_{k=0}^{n} (C_n^k)^2$.

7. En passant... vite !

Enoncé

Comment calculer $p(n) = a.p(n-1) + b$ en temps constant, $p(0) = \beta$ étant donné ?

Solution

D'abord, il faut vérifier qu'à chaque fois $p(n)$ est une probabilité, c'est à dire que $p(n)$ est positif ou nul, inférieur à 1. Se poserait aussi la question de vérifier que la somme des $p(n)$ fait 1. Si on admet que ces vérifications ont été faites, on ne doit pas programmer un calcul itératif direct comme

```
p <-- beta
```

```
pour k de1a n
   p <-- p * a + b
fin pour k
```

qui effectue le calcul en temps proportionnel à n mais passer par la formule close. Rappelons la méthode : on associe à $p_n = p(n)$ une suite w_n géométrique, de même raison a que p_n et décalée de p_n par une constante : $w_n = p_n + \alpha$. Remplaçant p_n et p_{n-1} par leur expression en fonction de w_n et w_{n-1} dans $p_n = a.p_{n-1} + b$, on en déduit que α vaut $b/(a-1)$. Puisque (w_n) est géométrique, $w_n = a^n w_0 = a^n(p_0 + \alpha) = a^n(\beta + \alpha)$.

On en déduit donc $p_n = a^n(\beta + \alpha) - \alpha$. Le calcul exact de a^n doit pouvoir se faire via les fonctions *log* et *exp*, d'où les instructions de calcul

```
# a,b, beta et n sont connus
alpha <--  b/(a-1)
puiss <--  exp( n*log(a) )
p     <--  (beta+alpha)*puiss - alpha
```

8. Question de vocabuaire...

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.

T.D. 3 : Variables aléatoires et Corrélation

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

1. Non additivité de la variance

Enoncé

Peut-on inventer une v.a. X avec 2 valeurs, une v.a. Y avec 3 valeurs et montrer facilement que $V(X+Y) \neq V(X)+V(Y)$?

Peut-on facilement trouver un cas où $V(X+Y) = V(X)+V(Y)$?

Solution

On ne peut pas prendre des valeurs x_1 et x_2 pour X avec des probabilités respectives $p_{X,1}$ et $p_{X,2}$, des valeurs y_1, y_2 et y_3 Y avec comme probabilité respective $p_{Y,1}$, $p_{Y,2}$ et $p_{Y,3}$ car ne sachant pas comment X et Y sont construites (c'est à dire sur quels évènements elles reposent), on ne peut pas définir $X+Y$.

Si Y est constante, alors $V(Y) = 0$ et $V(X+Y) = V(X)+V(Y) = V(X)$.

2. 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, le 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 discernables. Ainsi $(1,2)$ et $(2,1)$ sont deux évènements élementaires, alors qu'on ne les identifier que si par exemple les dés sont de couleur différente. Le système complet d'évènements est alors 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 $x \leq y$. Mais alors on aurait moins de couples (combien ?) et ils ne seraient plus équiprobables.

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 est d'écrire un programme pour gérer tous les calculs. La partie remplissage du tableau des données, noté svd (comme série de valeurs), peut se faire par l'algorithme suivant où col_S désigne la colonne des valeurs de S, col_P celle de P etc. Une implémentation simple consiste à prendre col_S=1, col_P=2 etc.

On trouvera en annexe un programme complet écrit en *tcl* qui effectue tous les calculs.

```
i <-- 1
v <-- 0
tant que i <= n
  j <-- 1
  tant que j <= n
      v <-- v + 1
      sdv[v,col_S] <-- (i + j)
      sdv[v,col_P] <-- (i * j)
      sdv[v,col_M] <-- (i * j)*(i + j)
      sdv[v,col_D] <-- abs(i - j)
      j <-- j + 1
  fin tant que sur j
  i <-- i + 1
fin tant que sur i
```

Ce programme permet de gérer aisément deux dés à n faces. L'énoncé demandait d'étudier le cas $n = 3$, ce qui fournit les résultats suivants.

```
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
```

On pourra comparer avec 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
```

3. 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 deux façons 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 (voir plus bas) $V(X) = M_{X^2} - m_X^2 \geq 0$ 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.

4. Loi de *Bernoulli* généralisée

Enoncé

Soit U=$b(x,y,p)$ la loi de Bernoulli généralisée, qui prend les valeurs y et x avec les probabilités respectives p et $1-p$ avec $x \leq y$. 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

$V(aX+b) = a^2V(X)$, $m(\mathsf{U}) = (y-x)m(\mathsf{T}) + x = (y-x)p + x$ et $V(\mathsf{U}) = (y-x)^2V(\mathsf{T}) = (y-x)^2p(1-p)$.

5. Centrage et Réduction

Enoncé

Soit X une v.a. ; construire une v.a. Y liée linéairement à X telle que $m(Y) = 0$. On la nomme la variable centrée issue de X. Construire une v.a. Z liée linéairement à X telle que $\sigma(Z) = 1$. 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 nomme Z la variable centrée réduite issué de X. Comparer T avec la variable $X - m(X)/\sigma(X)$ et à la variable $X/\sigma(X) - m(X)$.

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 $m(X/\sigma(X) - m(X)) = m(X)(1/\sigma - 1)$.

6. 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 :

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

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

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

Comparer ma, mg, mq et mh 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 $ma((x_i)) = \frac{1}{n}\sum_{i=1}^{n} x_i$.

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

Leur moyenne quadratique $mq((x_i)) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$.

Leur moyenne harmonique $mh((x_i)) = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$.

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

De même $mg((x_i)) = \sqrt[n]{c^n} = c$, $mq((x_i)) = \sqrt{nc^2/n} = c$, $mh((x_i)) = n/(n/c) = c$.

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

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

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

$mg(x, y) = \sqrt{8 + 2} = \sqrt{16} = 4$ $mh(x, y) = 1/((1/2) + (1/8)) = 8/5 = 1.6$
$mq(x, y = \sqrt{(2^2 + 8^2)/2} = \sqrt{34} = 5.83$

Donc $mh < mg < ma < mq$.

Annexe

```
#########################################################
#                                                       #
# deuxdes.tcl : programme de calcul des v.a.            #
#               S,P,M,D et leur caractéristiques        #
#               pour deux dés à n faces                 #
#                                                       #
#########################################################
#
#            auteur : Gilles HUNAULT
#            email  : gilles.hunault@univ-angers.fr
#            web    : http://www.info.univ-angers.fr/pub/gh/gh.html
#

proc copies { a b } {

 # renvoie b copies de a

 set c ""
 set i 1
 while { $i <= $b } {
   append c $a
   set i [incr i ]
 } ; # fin tant que sur i

 return $c

} ; # fin de proc copies

proc mve { ns } {

 ## affiche les valeurs puis moyenne, variance, écart-type

 global n sdv nom

 puts "\n étude de la variable : $nom($ns,1) = $nom($ns,2)\n"
```

```
# 1. affichage en produit cartésien

# 1.1 affichage de la ligne de label

set i 1
set  ls "    $nom($ns,1)  |"
while { $i <= $n } {
  set f [format "%5d" $i]
  append ls "$f"
  set i [incr i ]
} ; # fin tant que sur i
puts "$ls"
puts "  [copies "-" 5]+[copies "-" 32]"

# 1.1 affichage des lignes de valeurs

set i 1
set v 0
while { $i <= $n } {
  set  j 1
  set  ls ""
  set  f [format "%5d" $i]
  append ls "$f  |"
  while { $j <= $n } {
      set v [incr v]
      set f [format "%5d" [expr int($sdv($v,$ns))] ]
      append ls "$f"
      set j  [incr j]
  } ; # fin tant que sur j
  puts "$ls"
  set i [incr i ]
} ; # fin tant que sur i
```

```
# 2. affichage résumé par valeur

# 2.1 initialisation des effectifs via le max en nom(.,3)

set im  $nom($ns,3)
set i 0
while { $i <= $im } {
  set eff($i) 0
  set i [incr i ]
} ; # fin tant que sur i

# 2.2.1  remplissage des effectifs

set nc  [expr $n*$n]
set i   1
while { $i <= $nc } {
  set es [expr int($sdv($i,$ns))]
  set eff($es) [expr 1 + $eff($es)]
  set i [incr i ]
} ; # fin tant que sur i

# 2.2.2 comptage des valeurs

set  i   0
set  nv  0
set  sde 0
while { $i <= $im } {
  if { $eff($i) > 0 } {
     set nv  [incr nv]
     set sde [expr $sde + $eff($i)]
  } ; # fin de si
  set i [incr i ]
} ; # fin tant que sur i
```

```
# 2.3 affichage de ces valeurs

# 2.3.1 affichage de la ligne de label

set i  0
set ls "\n    $nom($ns,1)  |"
while { $i <= $im } {
  if { $eff($i) > 0 } {
     set f [format "%5d" $i]
     append ls "$f"
  } ; # fin de si
  set i [incr i ]
} ; # fin tant que sur i
puts "$ls"
puts "  [copies "-" 5]+[copies "-" [expr 2+5*$nv]]"

# 2.3.2 affichage des ligne des effectifs

set i  0
set s  0.0
set sc 0.0
set ls "[format "%4d" $nc]p  |"
while { $i <= $im } {
  set  effi $eff($i)
  if { $effi > 0 } {
     set f  [format "%5d" $effi ]
     set s  [expr $s   + 1.0*$effi*$i]
     set sc [expr $sc  + 1.0*$effi*$i*$i]
     append ls "$f"
  } ; # fin de si
  set i [incr i ]
} ; # fin tant que sur i
puts "$ls    total : $sde"
```

```
 # 2.4 indicateurs numériques

 set m [expr $s/$nc]
 set v [expr ($sc/$nc)-$m*$m]
 set e [expr sqrt($v)]
 set c [expr 100*$e/$m]

 set fm [format "%9.3f" $m]
 set fv [format "%9.3f" $v]
 set fe [format "%9.3f" $e]
 set fc [format "%9.3f" $c]

 puts "\n variable $nom($ns,1) moyenne    : \
      $fm   (somme des valeurs :            \
      [format "%10.0f" $s])"

 puts "            variance   : \
      $fv   (somme des carrés des valeurs : \
      [format "%10.0f" $sc])"

 puts "            écart-type : $fe"

 puts "            cdv        : $fc \
     \% (coefficient de variation)"

} ; # fin de proc mve

proc corr { ia ib } {

  ## calcule le coefficient de corrélation
  ## linéaire entre deux séries de n valeurs mises
  ## dans le tableau (global) sdv
  ## en colonnes ia et ib

  global n sdv nom
```

```
  if { $ia == $ib } { set cor 1 } else {

     set cor 0
     set nc [expr $n*$n]
     set v   1
     set sa  0
     set sca 0
     set sb  0
     set scb 0
     set sab 0

     while { $v <= $nc } {
       set  xa  $sdv($v,$ia)
       set  xb  $sdv($v,$ib)
       set  xab $xa*$xb
       set  sa  [expr $sa  + $xa]
       set  sca [expr $sca + $xa*$xa]
       set  sb  [expr $sb  + $xb]
       set  scb [expr $scb + $xb*$xb]
       set  sab [expr $sab + $xab]
       set  v   [incr v ]
     } ; # fin tant que sur i

     set ma  [expr $sa/$nc]
     set mca [expr $sca/$nc]
     set mb  [expr $sb/$nc]
     set mcb [expr $scb/$nc]
     set mab [expr $sab/$nc]
     set va  [expr $mca-$ma*$ma]
     set vb  [expr $mcb-$mb*$mb]
     set cov [expr $mab-$ma*$mb]
     set cor [expr $cov/(sqrt($va*$vb))]

  } ; # fin si

  return $cor

} ; # fin de proc corr
```

```
#########################################################
##                                                     ##
##  Programme principal                                ##
##                                                     ##
##                                                     ##
#########################################################

 set n 6 ; #### par paramètre : set n $argv

 puts " Etude de 2 dés avec chacun $n faces "

 set nom(1,1) "S" ;  set nom(1,2) "somme des valeurs"
 set nom(1,3) [expr $n+$n]
 set nom(2,1) "P" ;  set nom(2,2) "produits des valeurs"
 set nom(2,3) [expr $n*$n]
 set nom(3,1) "M" ;  set nom(3,2) "somme*produit des valeurs "
 set nom(3,3) [expr ($n+$n)*($n*$n)]
 set nom(4,1) "D" ;  set nom(4,2) "valeur absolue de la différence"
 set nom(4,3) [expr $n-1]

 set i 1
 set v 0
 while { $i <= $n } {
   set  j 1
   while { $j <= $n } {
       set v [incr v]
       set sdv($v,1) [expr 1.0*($i + $j)]
       set sdv($v,2) [expr 1.0*($i * $j)]
       set sdv($v,3) [expr 1.0*(($i * $j)*($i + $j))]
       set sdv($v,4) [expr 1.0*(abs($i - $j))]
       set j  [incr j]
   } ; # fin tant que sur j
   set i [incr i ]
 } ; # fin tant que sur i

 puts " ... On passe en revue les [expr $n*$n] cas    possibles."
```

```
# affichage par variable

set nx 4
set ix 1
while { $ix <= $nx } {
   set tmp [mve $ix]
   set ix [incr ix ]
} ; # fin tant que sur ix

# affichage par couple de variables

puts "\n Matrice de corrélation \n"
set ix 1
while { $ix <= $nx } {
   set ls "   "
   set jx 1
   while { $jx <= $ix } {
       set tc [corr $ix $jx]
       set fc [format "%10.4f" $tc]
       append ls "$fc"
       set jx [incr jx ]
   } ; # fin tant que sur jx
   puts $ls
   set ix [incr ix ]
} ; # fin tant que sur ix

puts "\n -- fin des calculs "
```

T.D. 4 : Lois classiques et v.a. continues

1. Conditions sur n sachant $m \geq 10$ 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. Salles de cinéma
6. Etude de la loi trapézoidale $\mathcal{T}(a,\alpha,h,\beta,b)$
7. Addition de deux lois uniformes continues sur $[0,1]$

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

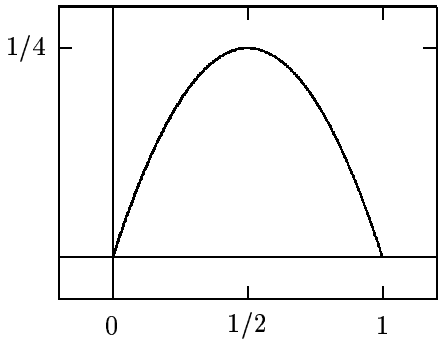
Enoncé

Montrer que $m \geq 10 \Rightarrow n \geq 40$ 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 $[0,1]$ montre que $f(x) \leq 1/4$.

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



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

2. Loi de u erreurs dans un livre de v pages

Enoncé

Un livre de 1000 pages contient 1500 erreurs. 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

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

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 le T.D. 8 comment réaliser une approximation de cette probabilité en utilisant la fonction de répartition de la loi normale $\mathcal{N}(0,1)$.

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 $p(|S| \leq 15)$.

On verra, là encore dans le T.D. 8, 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. Salles de cinéma

Enoncé

Un cinéma comporte 2 salles de cinéma de n places chacune. On admettra que les choix des spectateurs sont indépendants, c'est à dire que si N personnes se présentent, un spectateur a 1 chance sur 2 d'aller dans la première salle et 1 chance sur 2 d'aller dans la deuxième salle. Comment choisir n pour que la probabilité qu'un spectateur ne puisse voir le film qu'il a choisi soit inférieur à ε ? Application : $N = 1000$, $\varepsilon = 0.001$.

Solution

Désignons par S_i l'évènement "le spectateur i veut aller dans la salle 1". Clairement, $S_i = b(1/2)$. Soit S le nombre de spectateurs ayant choisi la salle 1. Alors $S = \sum_{i=1}^{N} S_i = \mathcal{B}(N, 1/2)$. Soit de la même façon T le nombre de spectateurs ayant choisi la salle 2. Alors $T = N - S$. L'évènement "un spectateur ne peut pas voir le film qu'il a choisi" correspond aux évènements "$S > n$" ou "$T > n$" si chaque salle comporte n places. On peut rendre symétrique l'évènement "$S > n \sqcup T > n$" en l'écrivant "$S - N/2 > n - N/2 \sqcup N/2 - T < N/2 - n$" soit encore "$S - N/2 > n - N/2 \sqcup S - N/2 < N/2 - n$" c'est à dire "$|S - N/2| > n - N/2$" bien sûr dans le cas où $N \leq 2n$.

Donc, de deux choses l'une ! Soit $N > 2n$ auquel cas notre évènement est certain et sa probabilité est 1 (par exemple 2000 personnes se présentent pour 2 fois 200 places donc forcément "des spectateurs ne peuvent pas voir le film qu'ils (elles) ont choisi".

Sinon, là encore nos connaissances actuelles ne nous permettent pas de résoudre la question posée, à savoir calculer n pour que $p(|S - N/2| > n - N/2| \leq \varepsilon)$. Signalons seulement qu'avec "la bonne approximation", pour $N = 1000$, $\varepsilon = 0.001$, on trouve $n \geq 540{,}8$ ce qui est un peu plus que $N/2$=500.

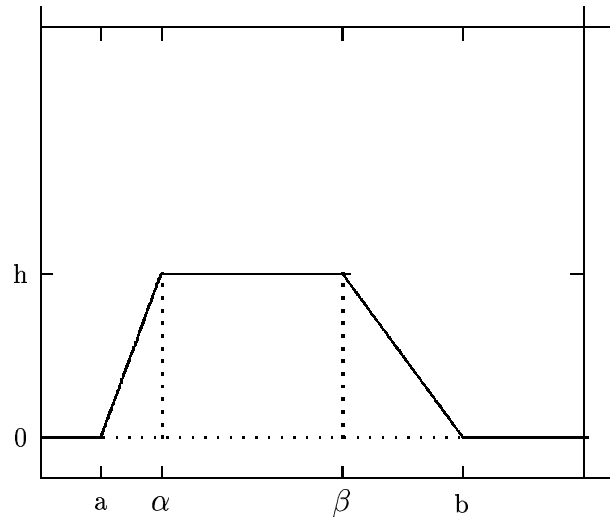
6. Etude la loi trapèze

Enoncé

Une variable aléatoire continue trapézoidale notée $\mathcal{T}(a, \alpha, h, \beta, b)$ est composée d'un segment droite oblique joignant $f(a) = 0$ à $f(\alpha) = h$ puis d'un segment de droite horizontal entre $x = \alpha$ et $x = \beta$ à la hauteur h puis d'un segment droite oblique joignant $f(\beta) = h$ à $f(b) = 0$. Donner dans le cas général l'équation des droites du trapèze et la valeur de la moyenne de $T(a, \alpha, h, \beta, b)$. *Application* : on prend $a = 1$, $\alpha = 2$, $\beta = 5$, $b = 7$. Donnez la valeur numérique de h et de la moyenne. Notant $1_{[u,v]}$ la fonction caractéristique de l'intervalle $[u, v]$ c'est à la dire la fonction qui vaut 1 pour $x \in [u, v]$ et 0 ailleurs, écrire la densite de la variable trapézoidale comme combinaison linéaire de trois fonction caractéristiques bien choisies.

Solution

Commençons par donner une représentation graphique de la densité de $\mathcal{T}(a, \alpha, h, \beta, b)$.



La partie entre a et α est un segment de droite qui s'annule en $x = a$ donc $D_1(x) = K_1.(x - a)$ comme $D_1(\alpha)$ vaut h on en déduit $K_1 = h/(\alpha - a)$. De même, la partie entre β et b est un segment de droite qui s'annule en $x = b$ donc $D_3(x) = K_3.(x - a)$ comme $D_3(\beta)$ vaut h on en déduit $K_3 = h/(\beta - b)$. Enfin, la partie entre α et β est constante et vaut h donc $D_2(x) = h$. On peut résumer ces informations avec les fonctions caractéristiques d'ensemble :

$$f(x) = K_1.1_{[a,\alpha]} + h.1_{[\alpha,\beta]} + K_3.1_{[\beta,b]}$$

La moyenne m vaut $\int_0^\alpha IdD_3 + \int_\alpha^\beta IdD_2 + \int_\beta^b IdD_3$ soit encore

$$\frac{h}{\alpha - a}\left[\frac{x^3}{3} - a\frac{x^2}{2}\right]_a^\alpha + h\left[\frac{x^2}{2}\right]_\alpha^\beta \frac{h}{\beta - b}\left[\frac{x^3}{3} - b\frac{x^2}{2}\right]_\beta^b$$

d'où finalement

$$m = \frac{h}{\alpha - a}\left(\frac{\alpha^3}{3} - a\frac{\alpha^2}{2} - \frac{a^3}{2} + \frac{a^3}{3}\right) + h\left(\frac{\beta^2}{2} - \frac{\alpha^2}{2}\right) + \frac{h}{b - a}\left(\frac{b^3}{3} - \frac{b^3}{2} - \frac{\beta^3}{3} + b\frac{\beta^2}{2}\right)$$

Numériquement, si a=1, α=2, β=5 et b=7 alors pour avoir une densité il faut (géométriquement) $h(\alpha - a)/2 + h(\beta - \alpha) + h(b - \beta)/2$=1 soit $h(9/2)$=1 donc $h = 2/9$. On trouve alors m=34/9 soit à peu près 3,778.

7. Addition de deux lois uniformes continues

Enoncé

Soient X et Y deux lois uniformes continues sur [0,1] et indépendantes. Quelle est la loi de $Z = X + Y$?

Solution

Soit f_X la densité de X et f_Y la densité de Y. Puisque X et Y sont indépendantes, la densité g de $Z = X + Y$ est définie par

$$g(s) = \int_{-\infty}^{+\infty} f_1(x) f_2(s-x) dx$$

qui est l'équivalent de la décomposition

$$p(\text{"}Z = s\text{"}) = \sum_{x=0}^{s} p(\text{"}X = x\text{"}) p(\text{"}Y = s - x\text{"})$$

pour la somme de deux variables aléatoires discrètes indépendantes.

Puisque f_1 est nulle hors de $[0,1]$, $g(s)$ se réduit à $g(s) = \int_0^1 f_2(s-x)dx$. Le changement de variable $x = s - u$ permet de remplacer $g(s)$ par $g(s) = \int_{s-1}^{s} f_2(u)du$ et il reste à distinguer différents cas pour s avant de conclure.

Si $s < 0$ alors les bornes d'intégration sont $s-1$ et 0 donc $g(s) = 0$ ce qui est "évident" car la somme de deux fonctions positives ne peut donner de valeur négative. De même, pour $s > 2$, les bornes sont toutes deux supérieures à 1 et f_2 étant nulle hors $[0,1]$, $g(s) = 0$, résultat là aussi "évident" car la somme de deux fonctions inférieures à 1 ne peut donner de valeur supérieure à 2.

Reste à envisager $0 \leq s \leq 2$. Séparons en deux cas, par rapport à 1.

Si $0 \leq s \leq 1$ alors

$$g(s) = \int_{s-1}^{s} f_2(u)du = \int_0^s 1du = s$$

Si $1 \leq s \leq 2$ alors

$$g(s) = \int_{s-1}^{s} f_2(u)du = \int_{s-1}^{1} 1du = 1 - (s-1) = 2 - s$$

Avec les notations du cours, Z est donc la loi triangle $\mathcal{T}([0,2])$.

On laisse au lecteur le soin de démontrer que la somme Z de n lois uniformes continues sur [0,1] indépendantes de densité f_n définie par

$$f_n(x) = \int_{-\infty}^{+\infty} f_1(x-y) f_{n-1}(y) dy = \int_{x-1}^{x} f_{n-1}(y) dy$$

vaut, par récurrence

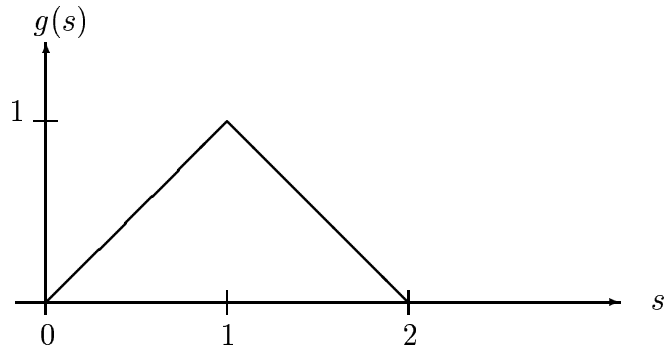
$$f_n(x) = \frac{1}{(n-1)!} \left[x^{n-1} - C_n^1 (x-1)^{n-1} + C_n^2 (x-2)^{n-1} + ... \right]$$

soit encore

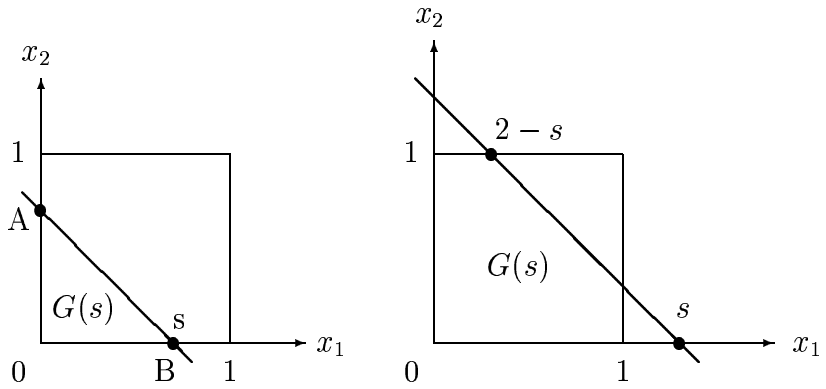
$$f_n(x) = \frac{1}{(n-1)!} \sum C_n^k (x-k)^{n-1}$$

où la sommation est étendue aussi longtemps que les termes x,$x-1$,$x-2$... sont positifs, c'est à dire jusqu'au terme en $(x-p)^{n-1}$ sur l'intervalle $[p, p+1]$ avec bien sûr $f_n(x) = 0$ pour $x \geq n$.

La représentation graphique de $\mathcal{T}([0,2])$ est



Une représentation graphique de la fonction de réparition G et un peu de calcul géométrique élémentaire aideront à comprendre les calculs (puisque $g = G'$) :



Pour la figure de gauche, s est inférieur à 1 et $G(s)$ correspond à la surface du triangle isocèle OAB où A et B sont sur la droite $x_1 + x_2 = s$. A a pour coordonnées $(0, s)$ et B $(s, 0)$. Donc $G(s) = s^2$ d'où $g(s) = s$.

Pour la figure de droite, $G(s)$ est la surface du trapèze obtenu en otant au carre $[0, 1] \times [0, 1]$ le triangle isocèle de coté $2 - s$. Donc $G(s) = 1 - (2 - s)^2/2$ d'où $g(s) = 2 - s$.Cqfd.

T.D. 5 : Statistiques Descriptives

1. Types de variables (QT,QL,QX...) en Statistiques Descriptives
2. Calculs concrets de QT : temps de transit
3. Calculs concrets de QL : bande passante
4. Valeurs de a et b pour $|\rho| = 1$
5. Programme de test si d est une distance

1. Types de variables (QT,QL,QX...)

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*,
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 sommation 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.

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

Solution

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 %

Le coefficient de corrélation liénaire ρ vaut ici 1 ce qui indique une liaison linéaire. Utilisant les formules $a = \rho.\sigma_Y/\sigma_X$ et $b = m_y - a.m_X$ pour la liaison $Y = a.X + b$ avec Y =D et X =R, on obtient D = 10,01.T-4,75. L'ordre d'affichage est ici indifférent puisqu'on a le même coefficient de variation.

3. 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.

4. 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 $a = \rho(X,Y).\sigma(Y)/\sigma(X) = (m(X).m(Y) - m(XY))/d$ et $b = m(Y) - a.m(X)$
$= (m(X).m(XY) - m(Y)mt(X,2))/d$ 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 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 F : $moy(XY) = a.moy(X^2) + b.moy(X)$. X et Y étant donnés, (E,F) est un système de deux équations linéaires en les deux inconnuées a et b dont la résolution donne les valeurs indiquées avec d. Pour revenir à $a = \rho(X,Y).\sigma(Y)/\sigma(X)$, il suffit de remarquer que $d = V(X) = \sigma(X)^2$ et que $m(X).m(Y) - m(XY) = cov(X,Y)$ et puisque $\rho(X,Y) = cov(X,Y)/\sigma(Y)/\sigma(X)$, le tour est joué!

5. Programme de test si d est une distance

Enoncé

Soit $X = (X[i,j])$ est un tableau de données où les n individus i sont en ligne et les variables j sont en colonne, on note $d(j_1, j_2)$ la valeur d'une fonction définie sur JxJ où J représente l'ensemble des colonnes. Par exemple $d(j_1, j_2) = \rho((j_1, j_2)$. On suppose que les $d(j_1, j_2)$ sont données par un un tableau nommé tabdist tel que tabdist[j,k]=$d(j, k$. Ecrire un programme Awk qui teste si tabdist peut être considéré comme une matrice triangulaire inféreure de distances.

Solution

```
# tstdist.awk : teste si le fichier peut être
#               considéré comme une matrice triangulaire
#               inféreure de distances

BEGIN { fs = "tstdist.sor"
        print "\n print " Etude de " ARGV[1]\n"
} # fin du  Begin

# lecture : on transfère les données dans le tableau tabdist

(FNR>1) {
  j = FNR-1 ;
  nomelt[j] = $1
  for (i=2;i<=NF;i++) {     k = i-1
      tabdist[j,k] = $i ; tabdist[k,j] = $i
      if ($i<0) {
         if (nbneg==0) {
            nbneg = 1 ; ilig = j ; jcol = k
         } # fin si nbneg = 0
      } # fin si $i <0
  } # fin pour i
} # fin de lecture après la ligne 1
```

```
END{

  if (nbneg>0) {
     print
     print "L'élément en ligne " ilig " colonne " jcol " est négatif"
     print "puisqu'il vaut " tabdist[ilig,jcol] " ; votre fichier " FILENAME
     print "ne peut donc PAS être considéré comme le fichier d'une matrice "
     print "triangulaire inférieure."
     print
    exit
  } # fin de si

  nbelt = FNR - 1

  print " test de positivité OK "

  # 1. Vérification de d(x,x)=0, soit 0 sur la diagonale

  ok1 = 1
  i   = 1
  while (i<=nbelt) {
    # print i " : " tabdist[i,i] " - "
    if (tabdist[i,i]!=0) {
    print
    print "L'élément sur la diagonale en position " i " est non nul"
    print "puisqu'il vaut " tabdist[i,i] " ; votre fichier " FILENAME
    print "ne peut donc PAS être considéré comme le fichier d'une matrice "
    print "triangulaire inférieure de distances."
    print

       i   = nbelt + 1
       ok1 = 0
    } # fin si
    i++
  } # fin tant que sur i
```

```
  if (ok1==1) {
     print " test de diagonale  OK "
     # 2. Vérification de l'inégalité triangulaire
     ok2 = 1 ;
     i   = 1
      while (i<=nbelt) {
        j = 1
        while (j<=nbelt) {
          j++ ;
          k = 1
          while (k<=nbelt) {
            ## test de l'inégalité proprement dite
            x = tabdist[i,j] ; y = tabdist[i,k] ; z = tabdist[k,j]
            if (x>y+z) {
                ok2 = 0
print
print "    Les éléments x = " x " ligne " i " colonne " j
print "                 y = " y " ligne " i " colonne " k
print "                 z = " z " ligne " k " colonne " j
print "    ne vérifient pas l'inégalité triangulaire puisque "
print "    " x " > " y+z " = " y " + " z " soit x > y + z "
print "    c'est à dire d(i,j) > d(i,k) + d(k,j) "
print "    pour i = " i", j = " j " et k " k"."
print " "
print "    Votre fichier " FILENAME
print "    ne peut donc PAS être considéré comme le fichier d'une matrice "
print "    triangulaire inférieure de distances."
print
               i = nbelt+1 ;
               j = nbelt+1 ;
               k = nbelt+1
            } # fin de si sur (x>y+z)
            k++
          } # fin pour k
        } # fin pour j
        i++
      } # fin pour i
```

```
  } # fin si ok1==1


  if (ok2==1) {
print " test triangulaire  OK "
print " "
print " D'après les test effectués, il semble bien que "
print "    votre fichier " FILENAME
print "    PEUT être considéré comme le fichier d'une matrice "
print "    triangulaire inférieure de distances."
print
  } # fin de si

  print " "
  print "  -- fin de tstdist "
  print " "

} # fin du END
```

T.D. 6 : Statistiques Descriptives

1. Asg de QT : dossier VINS
2. Asg de QL : dossier ELF88QL
3. Calculs concrets de χ^2 : pièces de fonderie
4. Approximations $\mathcal{B}$ et $\mathcal{P}$: filtrage de substrats
5. Distance ou pas ?

1. Analyse Statistique 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.

Nous fournissons également les sommes suivantes :

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

Solution

Nous donnons ici l'affichage "intelligent" des résultats tels que nous le donne notre programme de statistiques AsgQT.Prg (écrit sous Dbase). Nous encourageons le lecteur (et la lectrice !) à vérifier "à la main" ces résultats à partir des sommes fournies.

```
ETUDE de la base  L:VINS

Caractéristiques de la base

Il y a 18  lignes et 9 colonnes (au sens de Dbase).

Analyse par variable

Tri par ordre décroissant de moyenne
```

Champ	Nom du champ	Moyenne m	Ecart-Type s	Cdv s/m	Min	Max
1	NUM	non numérique ; son type est : N				
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

Tri par ordre décroissant de coefficient de variation

Champ	Nom du champ	Moyenne m	Ecart-Type s	Cdv s/m	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

Analyse par couple de variables

Matrice des corrélations

	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	0.9693	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	-0.0246	0.7469

Liste des meilleures corrélations linéaires
par ordre décroissant de rho.

On prend comme borne de corrélation sûre la valeur 0,9
et comme borne de corrélation possible la valeur 0,6.

*	0.969	UK	RFA
*	0.948	CANA	RFA
*	0.942	UK	BELG
*	0.926	CANA	UK
	0.894	USA	UK

```
0.870 NEDE  BELG
0.870 USA   BELG
0.869 RFA   BELG
0.848 USA   RFA
0.814 CANA  BELG
0.747 CANA  USA
0.717 USA   ITAL
0.700 ITAL  RFA
0.700 UK    NEDE
0.691 UK    ITAL
0.680 USA   NEDE
0.659 CANA  ITAL

Valeurs des formules linéaires pour les corrélation sûres.

 Corrélation    0.969  :  UK    =  0.513  *  RFA   +  3903.587
 Corrélation    0.969  :  RFA   =  1.831  *  UK    -  5921.349
 Corrélation    0.948  :  CANA  =  0.185  *  RFA   -   932.386
 Corrélation    0.948  :  RFA   =  4.857  *  CANA  +  6596.411
 Corrélation    0.942  :  UK    =  2.225  *  BELG  -  2322.591
 Corrélation    0.942  :  BELG  =  0.398  *  UK    +  1772.722
 Corrélation    0.926  :  CANA  =  0.341  *  UK    -  2064.845
 Corrélation    0.926  :  UK    =  2.511  *  CANA  +  7233.245
```

2. Analyse Statistique 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.

```
Sexe            : 0 Homme        1 Femme
Niveau d'étude  : 0 Non réponse  1 Primaire ( niveau CEP )
  2 Secondaire ( niveau BEPC ) 3 Secondaire ( niveau BAC ) 4 Supérieur
Régionalité     : 0 NR, 1 faible, 2 moyenne 3 forte 4 très forte
Usage de la langue : 0 NR, 1 peu fréquent, 2 commun, 3 très particulier
```

Les résultats informatiques sont

```
SEXE        0       35 ( 35.35 %)        1       64 ( 64.65 %)
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 %)
```

Présentez-les comme convenu et essayez de les commenter.

Solution

```
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 pas mal de gens de niveau bepc (30 %) ou bac (21 %).

3. 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.

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

La somme des n_i vaut 248.

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

La variance estimée 4,103 soit un écart-type de 2,026.

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). Venant regrouper les deux dernières classes, on dispose alors de 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 et 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.

5. Distance ou pas ?

Enoncé

L'appplication $\rho : (X, Y) \rightarrow \rho(X, Y)$ est une application qui à un couple de variables quantitatives associe un coefficient. Est-ce une distance ? Idem pour son carré ρ^2.

Soient a, b et c trois vecteurs de même longueur dont tous les éléments sont positifs ou nuls. Si la somme des éléments de c est non nulle, on appelle χ^2 entre a et b en base c la quantité $d_c(a, b)$ définie par $\Sigma(a_i - b_i)^2/ci$. La fonction $(a, b) \rightarrow d_b(a, b)$ est-elle une distance ?

Soient p et q deux variables qualitatives binaires c'est à dire qu'elles ne prennent que les valeurs 0 et 1. On appelle coefficient de ressemblance de Russel-Rao entre p et q la quantité $c_1(p, q) = 1_p1_q/(1_p1_q+1_p0_q+0_p1_q)$ où x_py_q désigne le nombre de fois où on trouve simultanément x pour p et y pour q. De même, le coefficient de Sokal-Michener entre p et q est $c_2(p, q) = (1_p1_q+0_p0_q)/n$ où n est la longueur commune des deux variables. Les fonctions $(p, q) \rightarrow c_1(p, q)$, $(p, q) \rightarrow 1 - c_1(p, q)$ et $(p, q) \rightarrow 1 - c_2(p, q)$ sont-elles des distances ?

Solution

Une distance est une fonction positive. Or $\rho(X, -X)$ vaut -1. Donc ρ ne peut être une distance. Une distance vérifie $d(X, X) = 0$ donc puisque $\rho^2(X, X) = 1$, ρ^2 ne peut pas être une distance. Par contre $1 - \rho^2$ est un candidat intéressant.

La fonction $(a, b) \rightarrow d_b(a, b)$ n'est pas une distance car ce n'est pas une fonction symétrique, puisque pour des vecteurs de longueur 1 $(a - b)^2/b$ est différent de $(a - b)^2/a$. Par contre, on laisse au lecteur le soin de démontrer que pour tout vecteur c donné, $(a, b) \rightarrow d_c(a, b)$ est une distance.

La fonction $(p, q) \rightarrow c_1(p, q)$ n'est pas une distance car $c_1(p, p) = 1$. La fonction $(p, q) \rightarrow 1 - c_1(p, q)$ n'est pas définie partout : par exemple si p ne contient que des 1, $c_1(p, p)$ vaut 0/0. Sinon, elle est symétrique, nulle sur la diagonale. Pour l'inégalité triangulaire, c'est une autre histoire.

$(p, q) \rightarrow 1 - c_2(p, q)$ est une distance car si $c_2(x, y) = (1_x 0_y + 0_x 1_y)/n$ alors $c_2(x, x) = (1_x 0_x + 0_x 1_x)/n = 0/n = 0$ et $c_2(y, x) = (1_y 0_x + 0_y 1_x)/n = (1_x 0_y + 0_x 1_y)/n = c_2(x, y)$. De plus, si $x = 1, y = 0$ alors $x = 1, z = 0$ vaut 1 et $y = 0, z = 1$ vaut 1. De même, si $x = 0, y = 1$ alors $x = 0, z = 1$ vaut 1 et $y = 1, z = 0$ vaut 1 donc $c_2(x, y) = (1_x 0_y + 0_x 1_y)/n$ est inférieur ou égal à $(1_x 0_z + 0_y 1_z)/n + (1_z 0_x + 0_z 1_y)/n$ qu'on peut écrire en $(1_x 0_z + 1_z 0_x)/n + (0_y 1_z + 0_z 1_y)/n$ soit encore $c_2(x, z) + c_2(y, z)$, CQFD.

T.D. 7 : Corrélation de rangs, Compararaisons...

1. Corrélation des rangs de Kendall

2. Algorithmes de comparaison de moyennes et pourcentages

3. Comparaison de moyennes : durées de tri

4. Coefficient de corrélation des rangs

5. Nombre d'appels sur une "hotline"

6. Calculs sur la loi de Mendel

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

1. Corrélation des rangs de Spearman et Kendall

Enoncé

Calculer le 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

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 donnera 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.

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

Solution

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

```
sdc ← 0
pour i de1a n
     eps ← A[i]-B[i]
     sdc ← sdc + epseps
```

```
finpour
denom ← n(n n-1)
rhos  ← 1 -6sdc/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 de somme 7. ρ_K vaut donc $-1 + 4.7/6.5$ 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... Un algorithme possible de calcul avec tri des valeurs de A et B est `/* kendall : on commence par le calcul des inversions */`

```
   r = 0
   pour j de 1 a  n-1
      x = b.j
      pour k de j+1 a n
         si b.k > x alors
            r = r + 1
         finsi
      finpour sur k
   finpour sur j
   rk = (4*r/(n*(n-1))) -1
   écrire " Kendall rk = " rk

/* spearman : calcul direct */

   s = 0
   pour j de 1 a n
      x = b.j
      y = a.j
      s = s + (x-y)*(x-y)
   finpour
   d = n*(n*n-1)
```

```
écrire " n*(n*n-1)  " d
rs = 1 - 6*s/d
écrire "Spearman s = " s " et rs = "  rs
```

2. Algorithmes de comparaisons

Enoncé

Donner l'algorithme du calcul de comparaison de pourcentages et du calcul de comparaison de moyennes pour des variances connues. On ne supposera pas la fonction racine existante.

Solution

```
/* compourc.rex : comparaison de pourcentages */


   ia ← ...
   na ← ...
   ib ← ...
   nb ← ...

   pa ← ia/na
   pb ← ib/nb
   p  ←(ia+ib)/(na+nb)
   df ← pa-pb
   si df <0 alors
      df ← -1*df
   finsi
   q ← p*(1-p)*(1/na+1/nb)
   r ← racine(q)

   écrire "ia " ia " na " na " pa " pa
   écrire "ib " ib " nb " nb " pb " pb
   écrire "ii " ia+ib " nn " na+nb " p " p
```

```
   écrire "df " df " r2 " q " r " r " r*r " r*r
   écrire "eps" df/r

   écrire " résultat du test : " df*racine(10)/r



/* sous-programme  */

   racine:
    arg x
    si x> 1 alors rr ← 1
            else rr ← 0.9
    do i ← 1 to 10
       rr ← (x/(2*rr)) + rr/2
    fin
   renvoyer rr

/* compmoy.rex : comparaison de moyennes */

   n ← ...

   a.0 ← ...
   a.1 ← ...
   ...
   a.n ← ...

   na ← 0
   nb ← 0
   sa ← 0
   sb ← 0
   ca ← 0
   cb ← 0
   do i ← 0 to n
      x.i ← i
      na ← na + a.i
      nb ← nb + b.i
      sa ← sa + x.i*a.i
```

```
   sb ← sb + x.i*b.i
   ca ← ca + x.i*x.i*a.i
   cb ← cb + x.i*x.i*b.i
fin

ma ← sa/na
mb ← sb/nb
va ← (ca/na) - (ma*ma)
vb ← (cb/nb) - (mb*mb)
ea ← racine(va)
eb ← racine(vb)
df ← ma - mb
si df < 0 alors df ← -df
d ← (va/na)+(vb/nb)
r ← racine(d)
delta ← df/r

écrire " A  na = " na " ma = " ma " va = " va " ea = " ea
écrire " B  nb = " nb " mb = " mb " vb = " vb " eb = " eb
écrire " |ma-mb| " df " r = racine( " d " ) = " r
écrire " delta = " delta
```

3. Comparaison de moyennes entre ordinateurs

Enoncé

On dispose de deux ordinateurs (A) et (B). Des simulations de taille de fichiers pour des calculs de tris en mémoire ont amenés aux valeurs suivantes pour les fichiers temporaires

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.

Solution

Tous calculs effectués, on trouve $n_a = 22$, $m_a = 7.5$, $v_a = 11.7045455$ et $n_b = 29$, $m_b = 6.93103448$, $v_b = 9.5814507$ d'où $|m_a - m_b| = 0.56896552$, $r = \sqrt{0.862419647} = 0.928665519$, $\delta = 0.612670018$. 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. Coefficient 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.

Solution

Soient $R = r_i$ et $S = s_i$ les rangs pour n objets de 1 à n. Alors $m(R) = m(S) = (n+1)/2$ et $\sigma(R) = \sigma(S) = \sqrt{(n^2-1)/12}$ car R et S correspondent à la variable uniforme discrète $\mathcal{UD}(n)$. Posons $\varepsilon_i = r_i - s_i$. Comme $\Sigma\ (r_i - s_i)^2 = \Sigma\ r_i^2 + \Sigma\ s_i^2 - 2\Sigma\ r_i s_i$, on en déduit que $\Sigma\ \varepsilon_i^2 = 2n(n+1)(2n+1)/6 - 2\Sigma\ r_i s_i$. D'autre part, $\rho = cov/\sigma_X\sigma_Y = \Sigma\ r_i s_i/n - m_X m_Y$ donc $\Sigma\ r_i s_i = n\rho\sigma_X\sigma_Y - n m_X m_Y$. Reportant cette valeur dans l'équation précédente, on obtient $\Sigma\ \varepsilon_i^2 = n(n+1)(2n+1)/3 - 2\,(\,\rho n(n^2-1)/12 - n(n+1)^2/4\,)$. Développant ces calculs, avec $n(n+1)/12$ en facteur, on a $\Sigma\ \varepsilon_i^2 = n(n+1)/12\,[\,4(2n+1) - 6(n+1)] - n\rho(n^2-1)/6\,]$, 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 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 : $S = R$ donc $\rho(R,S) = 1$. Par contre, si les choix sont inverses : $S = n+1-R$ donc $\rho(R,S) = -1$, la démonstration directe avec $b_i = n+1-a_i$ étant plus longue...

Si on a les mêmes rangs, alors a_i=b_i donc il y a $n-1$ inversions puis $n-2$ puis... jusqu'à 1. Alors S vaut n(n-1)/2 d'où 4/D vaut 2 et donc $\rho_K = 1$. Par contre si les choix sont inverses il n'y a aucune inversion donc $\rho_K = 1$.

5. 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 soit une moyenne de 6 jours avec un écart-type de 1.2083. Peut-on comparer l'utilisation de la ligne ?

Solution

Bien sûr qu'on peut comparer ! 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.

6. Calculs sur la loi de Mendel

Enoncé

Le brillant généticien Mendel avait introduit un schéma de répartition 9/3/3/1 pour décrire la couleur dominante et la couleur récessive de certaines espèces florales. Des élèves de Deug Biologie à l'Université d'Angers décident de vérifier cela sur un exemple. Pour cela, ils prennent des fleurs de balsamine avec un croisement blanc/pourpre. Ils obtiennent les valeurs suivantes

couleur	pourpre	rose	blanc/lavande	blanc
nb. fleurs	1790	547	548	213

le schéma de Mendel est-il vérifié ?

D'autres élèves penchent plutôt pour une équirépartition. Leurs données sont

couleur	pourpre	rose	blanc/lavande et blanc
nb. fleurs	1065	326	457

Peut-on leur donner raison ? ou doit-on encore prendre le schéma de Mendel ?

Solution

Le total général est 3098 et la répartition théorique de Mendel donnerait comme effectifs 1743, 581, 581 et 194 (ou 193). Le calcul du χ^2 donne 7 alors que le χ^2 théorique maximal autorisé à 5 % est 7.81. Le schéma est donc vérifié.

Pour le deuxième jeu de valeur, le total général est 1829 et la répartition théorique de Mendel donnerait comme effectifs 1029, 343 et 457 (343+114). Le calcul du χ^2 donne 2.89 alors que le χ^2 théorique maximal autorisé à 5 % est 5.99 Le schéma est donc encore vérifié. Par contre l'équirépartition donnerait 610 comme effectif moyen d'où un χ^2 de 520 qui est inacceptable.

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

Enoncé

Une enquête récente a fourni les données suivantes relatives au croisement lieu de vente (Grand Magasin sauf Fnac, Fnac, Vente par correspondance, Librairie Générale) et famille de livres (notée par souci de confidentialité famille A, B, C, D, E et F) :

	GM	Fnac	Vpc	Lib.G
A	1	0	0	0
B	11	4	3	6
C	11	1	7	0
D	6	4	11	4
E	7	12	11	10
F	1	1	1	3

Y a-t-il un lien entre lieu de vente et famille de livre ?

La vente de tels livres est-elle équirépartie par lieu ?

Solution

L'utilisation du programme Awk vu en cours donne comme résultat :

```
6 lignes et 4 colonnes
    1     0     0     0  +     1
   11     4     3     6  +    24
   11     1     7     0  +    19
    6     4    11     4  +    25
    7    12    11    10  +    40
    1     1     1     3  +     6
++++++++++++++++++++++++++++++++++
   37    22    33    23  +   115

    0.322     0.191     0.287     0.200  +     1.000
    7.722     4.591     6.887     4.800  +    24.000
    6.113     3.635     5.452     3.800  +    19.000
    8.043     4.783     7.174     5.000  +    25.000
   12.870     7.652    11.478     8.000  +    40.000
    1.930     1.148     1.722     1.200  +     6.000
++++++++++++++++++++++++++++++++++++++++++++++++++++++
   37.000    22.000    33.000    23.000  +     0.000

 chi-deux 28.151
```

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 qu'il y a liaison, la plus forte contribution étant 3.91, fournie par la ligne 3, colonne 1.

Puisque le total général est 115, s'il y avait équirépartition de la vente des livres par lieu, chaque lieu aurait 28,75 livres. Conservant cette valeur pour les TH, le χ^2 d'ajustement est alors 5.73 alors que le χ^2 théorique maximal est 7.81 pour 3 ddl. Une déformation "raisonnable" des valeurs pour avoir des nombres théoriques entiers pourrait être de prendre comme répartition 30 28 29 28 ce qui fournit un χ^2 de 4.63 ; là encore, on peut accepter l'hypothèse d'équirépartition.

T.D. 8 : Approximations et Algorithmes

1. Intervalle de confiance et de variabilité
2. Intervalle de confiance à 5 %
3. Etalonnage de fréquence d'un spectrophotomètre
4. Approximation de $\mathcal{B}(n,p)$ par $\mathcal{N}(0,1)$
5. Approximation de $\mathcal{P}(\lambda)$ par $\mathcal{N}(0,1)$
6. Approximation de $\mathcal{B}(n,p)$ par $\mathcal{P}(\lambda)$
7. Reprise des exercices *montre*, *concentrateur* et *cinéma*
8. Algorithme du calcul de m, σ pour une matrice de données
9. Algorithme de $H(N,D,n)$ (exam. mai 97)

1. 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 $\alpha\%$ l'intervalle centré $I_m = [m - t\sigma, m + t\sigma]$ où t correspond $p(U > \alpha)$. On appelle intervalle de variabilité l'intervalle centré $I_V = [m - tV, m + tV]$ et intervalle de sûreté $I_s = [m - t\sigma/\sqrt{n}, m + t\sigma\sqrt{n}]$

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 %.

La matrice des corrélations est

```
        X        Y        Z

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

Et on on trouve donc que $Y = 2X + 1$ et donne des km par min, $Z = 64.071 * X - 716.159$ (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 $[8, 4, 38.6]$

2. Intervalle de confiance à 5 %

Enoncé

Au risque de $\alpha = 5\%$, la valeur de t pour l'intervalle de confiance est 1.96 ; comparer ces intervalles de confiance pour $\mathcal{B}(12; 0.3)$ et $\mathcal{P}(\lambda)$ pour λ bien choisi.

Solution

Le λ bien est celui tel que les deux lois aient la même moyenne, d'où $\lambda = np = 12 * 0.3 = 3.6$. Pour la loi binomiale $\sigma = \sqrt{np(1-p)} = \sqrt{2.52} = 1.59$ alors que pour la loi de Poisson, $\sigma = \sqrt{\lambda} = \sqrt{3.6} = 1.90$. Pour la loi binomiale, $t\sigma/\sqrt{n} = 1.96 * 1.59/\sqrt{13} = 0.8629$ donc l'intervalle est $[2.7371, 4.4629]$. Pour la loi de Poisson, on suppose que la loi est tronquée aux 13 premières valeurs, d'où $t\sigma/\sqrt{n} = 1.96 * 1.9/\sqrt{13} = 1.0314$, $I = [2.5686, 4.6314]$.

3. Etalonnage de fréquence

Enoncé

Soient A et B les mesures d'étalonnage 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 ces résultats à l'aide du test des rapport des variances $R = V_A/V_B$ où le seuil maximal est donné dans la table de *Snedecor*.

Solution

Pour nos deux séries de 10 valeurs, on trouve $m_A = 147$, $m_B = 197$, $V_A = 208.67$, $V_B = 188.89$. R vaut donc 1.10 or le seuil R_{max} lu dans la table au seuil de α =5 pour 9 ddl. On considére donc qu'il n'y pas de différence significative entre nos séries de valeurs...

4. Approximation de $\mathcal{B}(n,p)$ par $\mathcal{N}(0,1)$

Enoncé

Soit $X = \mathcal{B}(15, 0.3)$. Effectuer un calcul direct de $p(X \in [3,6])$. Calculer cette même probabilité de façon approchée en utilisant la loi normale.

Solution

La valeur exacte est $\Sigma_{k=3}^{6}\, C_{15}^{k} 0.3^k (1-0.3)^{15-k}$ soit (par exemple avec $Maple$) 0.7419. On a bien sûr $m_X = 4.5$, $V_X = 3.15$, $\sigma_X = 1.7748$. Si on utilise la loi normale, on assimile$[\![3,6]\!]$ à $[2.5, 6.5]$ soit $U \in [(2.5-m_X)/\sigma_X, (6.5+m_X)/\sigma_X]$ c'est à dire $U \in [-1.126, +1.126]$. Utilisant la fonction de répartition F_U de la loi normale $U = \mathcal{N}(0,1)$ on trouve $p(U \in [-a,a]) = 2F_u(a) - 1$ et pour $a = 1.126$, $F_U(0.8699]$ donc $p = 0.7398$.

5. Approximation de $\mathcal{P}(\lambda)$ par $\mathcal{N}(0,1)$

Enoncé

Soit $X = \mathcal{P}(20)$. Effectuer un calcul direct de $p(X \leq 10)$.

Calculer cette même probabilité de façon approchée en utilisant la loi normale.

Solution

Le calcul direct donne $\displaystyle\sum_{k=0}^{10} e^{-20} 20^k / k!$ soit 0.0165.

On a $m_X = 20 = V_X$,$\sigma_X = 4.4721$. Si on assimile $X \leq 10$ à $X \leq 10.5$ alors

$p = p(\ (X-m)/\sigma \leq (10.5-20)/4.4721\ )$ vaut $F_U(-2.1243)$ soit encore $1 - F_U(2.1243) = 1 - 0.09832 = 0.0168$

6. Approximation de $\mathcal{B}(n,p)$ par $\mathcal{P}(\lambda)$

Enoncé

Un livre de 1000 pages contient 1500 erreurs. Donner une approximation de la probabilité qu'une page contienne moins de 2 erreurs en utilisant la loi normale. Donner la valeur exacte de cet événement.

Calculer la probabilité de cet événement si on remplace la loi binomiale sous-jacente par une loi de Poisson bien choisie.

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 peuprès 0.55774.

L'approximation gaussienne dit que

$$p(X < 2) = p\left(\frac{X-m}{\sigma} < \frac{2-np}{\sqrt{np(1-p)}}\right) \text{ vaut } F_U\left(\frac{1.5-np}{\sqrt{np(1-p)}}\right)$$

soit ici $F_U(0) = 0.5$. L'erreur commise entre le 0.5 d'approximation et les 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, $p(X = 0) + p(X = 1)$ devient $e^{-\lambda}\lambda^0/0! + e^{-\lambda}\lambda^1/1!$ soit, puisque $e^{-\lambda} = e^{-1.5} = 0.22313$, 0.22313+0.33469 ce qui aboutit à la valeur 0.55782 qui elle, est très proche du résultat exact 0.55774.

7. Reprise : *montre*, *concentrateur* et ***cinéma***

Enoncé

Une montre fait une erreur d'au plus 30 secondes par jour. Calculer la probabilité que l'erreur soit inférieure à 15 minutes au bout d'un an.

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 ?

Un cinéma comporte 2 salles de n places. Si N personnes se présentent, en admettant que les choix des spectateurs sont indépendants, chaque spectateur a une chance sur 2 d'aller dans la prmière salle. Comment choisir n pour que la probabilité qu'un spectateur ne puisse pas voir le film qu'il (elle) a choisi soit inférieure à ε ? Application numérique : $N = 1000$, $\varepsilon = 0.01$.

Solution

Pour la montre, 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) : $S = \sum_{i=1}^{365} X_i$. La moyenne de S est donc $\mu = 0$, $\sigma = 1/\sqrt{12}$. Le théorème de limite centrale permet d'assimiler à

$$\frac{S - n\mu}{\sigma\sqrt{n}} = S\sqrt{\frac{12}{365}} \text{ à } \mathcal{N}(0,1)$$

$p(|S| \leq 15)$ est donc assimilable à $p\left(|U| \leq 15 \leq \sqrt{15/365} \right)$ soit $p(|U| \leq 2.72)$. Cette probabilité vaut donc $2F_U(2.72) = -1 = 2 * 0.99674 - 1$ soit à peu près 0.99.

Pour le concentrateur, à 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 $\mathcal{B}(n,p) = \mathcal{B}(1000, 0.0417)$ La saturation correspond à l'évènement $Y > 50$. Sa proba-

bilité exacte est

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

et elle vaut 0.08413259 (avec *Maple* par exemple).

Si au lieu de la calculer directement on effectue une approximation par la loi normale, puisque $n = 1000 > 36$, $np = 41.6 > 5$ et $np(16p) = 39.869 > 5$, on remplace

$$p(Y > a) = p\left(\frac{Y-m}{\sigma} > \frac{a-np}{\sqrt{np(1-p)}}\right) \text{ par } \int_a^{+\infty} \frac{1}{\sqrt{2\pi}e^{-t^2/2}}dt$$

soit, pour $a = 50$, $p = 1 - F_U(1.319)$=1-0.9066 donc 0.0934

Pour les salles de cinéma, à partir de l'évènement S_i "le spectateur i veut aller dans la salle 1" modélisé par $b(1/2)$, pour S le nombre de spectateurs ayant choisi la salle 1. et T le nombre de spectateurs ayant choisi la salle 2 on a $T = N - S$ et $S = \sum_{i=1}^{N} S_i = \mathcal{B}(N, 1/2)$ de moyenne $\mu = 1/2$, de variance $V = 1/4$ et d'écart-type $\sigma = 1/2$. L'évènement "un spectateur ne peut pas voir le film qu'il a choisi" correspond aux évènements "$S > n$" ou "$T > n$" si chaque salle comporte n places ce qui s'écrit encore "$|S-N/2| > n-N/2$" au cas où $N \leq 2n$. On veut donc calculer n pour que $p(|S-N/2| > n-N/2 \leq \varepsilon)$. Utilisant l'approximation par la loi normale, et remplaçant $|S - N\mu| > \alpha$ par $|\mathcal{N}(0,1)|/\sigma\sqrt{N} > \alpha$ on calcule $p\Big(\ |\mathcal{N}(0,1)| > (n - N/2).2/\sqrt{N}\ \Big)$. On trouve $p(|U| > (2n-N)/\sqrt{N}) = 2\Big(\ 1 - F_U((2n-N)/\sqrt{N})\ \Big) < \varepsilon$. Pour l'application numérique : $p < 0.01$ donne $F_U(y) > 0.995 = F_U(2.58)$ ce qui implique $y \geq 2.58$, soit $(2n-N)/\sqrt{N} \geq 2.58$ c'est à dire finalement $n \geq (\sqrt{N}2.58 + N)/2$ donc $n \geq 540,8$ pour $N = 1000$, $\sqrt{N} = 31.62$.

8. Algorithme de m, σ

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 σ_X/m_X.

Soit D un tableau de données dont les n lignes sont repérées par un premier 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 les colonnes 1 à t de D correspondent à des variables quantitatives et que 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 la 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.

a) Si les données sont

```
1       1        1    |   1
2       2        4    |   1
1       3       25    |   2
3       4        9    |   1
1       5       16    |   2
```

compléter le tableau suivant des résultats à 0,1 près

```
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 Les valeurs numériques 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
```

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
```

9. Algorithme de $H(N, D, n)$ (exam. mai 97)

Enoncé

Soit X la *v.a.* "loi hypergéométrique" $\mathcal{H}(N, D, n)$: on tire n valeurs sans remise dans N boules dont D sont défectueux ; X est la loi "nombre de défectueux". X prend les valeurs entières $k = a, a+1, a+2$ ... b où $a = max(0, D+n-N)$ et $b = min(D, n)$. 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. On calculera au passage la somme des probabilités, la moyenne, la variance, l'écart-type et le coefficient de variation de X.

Solution

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. Leur probabilité respective est $p_0 = C_3^0 C_3^2 / C_6^2 = 3/15$, $p_1 = C_3^1 C_3^1 / C_6^2 = 9/15$ et $p_2 = C_3^2 C_3^0 / C_6^2 = 3/15$.

Pour le calcul des valeurs, des probabilités, on pourra utiliser l'algorithme suivant

```
Algorithme du calcul de H(N,D,n)

# calcul des bornes des valeurs de X
# nt est mis pour n, NB pour N

  a ← D+nt-NB
  si 0 > a alors
     a ← 0
  finsi
  b ← D
  si nt < b alors
     b ← nt
```

```
  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 ← c(nt,N) # dénominateur constant pour k
  di ← N-D     # différence  constante pour k
  pour k de a à b
       p  ← c(k,D).c(nt-k,di) / de
       u  ← u  + 1
       s  ← s  + k*p
       sc ← sc + k*k*p
       sp ← sp + p
  fin pour k

# moyenne, variance

  m    ← s/u
  v    ← sc/u - m*m
  e    ← racine(v)
  cdv  ← 100*e/m
```