STATISTIQUE COMPUTATIONNELLE – 3

VECTEURS ALÉATOIRES

3.1

Cas de plusieurs variables aléatoires

Il est fréquent que nous voulons étudier les variations simultanées de deux ou plusieurs grandeurs statistiques. Dans ce cas il faut utiliser une *loi jointe* caractérisée par fréquences jointes (ou une densité jointe dans le cas des variables continues) et où l'espérance est un vecteur et la variance une matrice. Précisons ce point.

Soient $X_1,\ldots,X_q$ des v.a. et supposons que chaque v.a. peut prendre plusieurs valeurs, par exemple la v.a. X_j prend ses valeurs dans l'ensemble $B_j = \{x_{j1},\ldots x_{jn_j}\}$; $j = 1,\ldots,q$. Pour traiter ces v.a. en même temps, on construit un *vecteur aléatoire* $\mathbf{X} = [X_1,\ldots,X_q]^\top$. La *probabilité jointe* est donnée par

$$P(\mathbf{X} = \mathbf{x}) = P\left(X_1 = x_{1i_1},\ldots,X_q = x_{qi_q}\right); \text{ avec } x_{ji_j} \in B_j$$

EXEMPLE 3.0.2 *Considérons deux dés et soient deux v.a. X qui représentent le maximum de deux faces et Y qui est leur somme. On a $P(x,y) = P(X = x, Y = y)$. Par exemple $p(3,4) = P(X = 3, Y = 4) = \frac{2}{36}$.*

En utilisant la probabilité jointe on peut définir pour chaque v.a. sa *probabilité marginale* selon la formule

$$P(X_j = x_{ji}) = P(X_1 \in B_1,\ldots,X_j = x_{ji},\ldots,X_q \in B_q)$$

qui donne la probabilité marginale de la v.a. X_j. Le second membre de cette relation doit être compris comme étant la somme des probabilités pour chaque combinaison des valeurs des v.a. $X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_q$. Ainsi dans le cas des v.a. continue la densité marginale de la v.a. X_j s'écrit

$$f_{X_j}(x_j) = \int_{X_1=-\infty}^{+\infty} \cdots \int_{X_{j-1}=-\infty}^{+\infty} \int_{X_{j+1}=-\infty}^{+\infty} \cdots \int_{X_q=-\infty}^{+\infty} f(x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_q)\,dx\cdots dx_{j-1}dx_{j+1}\cdots dx_q$$

EXEMPLE 3.0.3 *(suite)* $p_Y(4) = p(X=2,Y=4)+p(X=3,Y=4) = \frac{3}{36}$.

Si on construit une table à deux entrées et on place en lignes les valeur de X et en colonnes les valeurs de Y et dans les cases de la table les valeurs $P(X=x_i,Y=y_j)$ correspondantes, les probabilités marginales de X sont les somme des lignes et les probabilités marginales de Y sont les sommes des colonnes. Elles apparaissent donc aux marges de la table, d'où leur nom des marginales.

3.2

Somme de v.a. indépendantes

Nous allons examiner la distribution de la somme des v.a. indépendantes

V.a. binomiales

Soient $X_1 \sim \mathcal{B}(n_1,p), X_2 \sim \mathcal{B}(n_2,p)$ deux v.a. indépendantes. Alors on a $X_1+X_2 \sim \mathcal{B}(n_1+n_2,p)$.

V.a. de Poisson

Soient $X_1 \sim \mathcal{P}(\lambda_1), X_2 \sim \mathcal{P}(\lambda_2)$ deux v.a. indépendantes. Alors on a $X_1+X_2 \sim \mathcal{P}(\lambda_1+\lambda_2)$.

V.a. normales

Soient $X_1 \sim \mathcal{N}(\mu_1,\sigma_1^2), X_2 \sim \mathcal{N}(\mu_2,\sigma_2^2)$ deux v.a. indépendantes. Alors on a $X_1+X_2 \sim \mathcal{N}(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$.

V.a. quelconques

Soient X et Y deux v.a. indépendantes, continue de densités f_X et f_Y. La densité de la v.a. $X+Y$ est

$$f_{X+Y}(z) = \int_{x=-\infty}^{+\infty} f_X(x)\,f_Y(z-x)\,dx = \int_{y=-\infty}^{+\infty} f_X(z-x)\,f_Y(x)\,dy = f_X * f_Y$$

c-à-d. la convolution de densités.

3.3

Grandeurs statistiques d'un vecteur aléatoire

L'espérance mathématique du vecteur aléatoire $\mathbf{X}$ est le vecteur

$$E(\mathbf{X}) = [E(X_1),\cdots,E(X_q)]^\top$$

Pour une v.a. X la variance est. $V(X) = E\,(X - E(X))^2$. Si on passe à un vecteur aléatoire nous obtenons, à cause du fait que la variance est le carré de l'espérance mathématique du scalaire $[X - E(X)]$, la matrice $E\left[(\mathbf{X} - E(\mathbf{X})) \cdot (\mathbf{X} - E(\mathbf{X}))^\top\right]$, de dimension $(q \times q)$, qui peut être vu comme le carré de l'espérance mathématique du vecteur $[\mathbf{X} - E(\mathbf{X})]$.

DÉFINITION 3.0.5 *On appelle matrice de variance-covariance du vecteur aléatoire* $\mathbf{X}$ *, la matrice*

$$C(\mathbf{X}) = E\left[(\mathbf{X} - E(\mathbf{X})) \cdot (\mathbf{X} - E(\mathbf{X}))^\top\right]$$

de terme général

$$C_{ij} = C(X_i, X_j) = E\left[(X_i - E(X_i)) \cdot (X_j - E(X_j))\right] \;;\; i, j = 1, \ldots, q$$

La quantité C_{ij} *s'appelle la covariance des v.a.* X_i *et* X_j*.*

Remarquons que

- la matrice $C(\mathbf{X})$ est symétrique
- $C_{ii} = C(X_i, X_i) = E\left[(X_i - E(X_i)) \cdot (X_i - E(X_i))\right] = E\,(X_i - E(X_i))^2 = V(X_i)$; $i = 1, \ldots, q$ c'est-à-dire que les termes diagonaux de la matrice de variance-covariance sont les variances $V(X_i)$ des coordonnées X_i du vecteur aléatoire $\mathbf{X}$.

Il est possible aussi, comme dans le cas des v.a., de définir une transformation du vecteur aléatoire $\mathbf{X}$. Nous allons nous limiter aux transformations linéaires, c'est-à-dire aux transformations qui peuvent s'exprimer à l'aide d'une matrice. Nous avons ainsi que si $\mathbf{X}$ est un vecteur aléatoire composé de q v.a. $X_1, X_2, \ldots, X_q$, alors $\mathbf{Y} = A \cdot \mathbf{X}$ est aussi un vecteur aléatoire, avec A matrice carrée $(q \times q)$ de terme général (a_{ij}) .L'espérance mathématique de $\mathbf{Y}$ est :

$$E(\mathbf{Y}) = A \cdot E(\mathbf{X})$$

La matrice de variance-covariance $C(\mathbf{Y})$ a comme terme général :

$$\begin{aligned} C(Y_i, Y_j) &= E\left[(Y_i - E(Y_i)) \cdot (Y_j - E(Y_j))\right] \\ &= \textstyle\sum_{k,l} a_{ik}\, a_{jl}\, E\left[(X_i - E(X_i)) \cdot (X_j - E(X_j))\right] \\ &= \textstyle\sum_{k,l} a_{ik}\, a_{jl}\, C(X_i, X_j) \end{aligned}$$

ou, sous forme matricielle :

$$C(\mathbf{Y}) = A \cdot C(\mathbf{X}) \cdot A^\top$$

Dans le cas de deux v.a. X et Y on pour la covariance

$$\mu_{11} = cov(X, Y) = E\left[(X - E(X))\;(Y - E(Y))\right]$$

s'appelle *covariance*. Nous avons la relation :

$$cov(X, Y) = E\,(X\,Y)\; - E\,(X) \cdot E(Y)$$

Il s'ensuit que si les v.a. sont indépendantes ($\Rightarrow E\,(X\,Y)\; = E\,(X) \cdot E(Y)$), alors $cov(X, Y) = 0$, mais la réciproque n'est pas, en général, vraie.

Dans le cas de deux v.a. nous avons aussi la notion de la corrélation dont la définition est la suivante :

DÉFINITION 3.0.6 *Soit un couple de v.a.* (X,Y) *La quantité* $\sigma^2_{X_1}$

$$\rho_{XY} = \frac{E\left[(X-E(X))\ (Y-E(Y))\right]}{\sigma(X)\cdot\sigma(Y)}$$

s'appelle coefficient de corrélation.

S'il n'y a pas d'ambigu¨é, on note simplement par ρ le coefficient de corrélation. Nous avons aussi la relation

$$\rho = \frac{cov(X,Y)}{\sigma(X)\cdot\sigma(Y)}$$

Si $X,\ Y$ sont indépendantes, alors $\rho = 0$ et on appelle les v.a. $X,\ Y$ non corrélées. Le contraire, en général, n'est pas vrai, i.e. $\rho = 0$ n'implique pas que $X,\ Y$ soient indépendantes.

3.4 Exercices

EXERCICE 3.1 *X et Y sont deux v.a. avec les probabilités jointes* $p(X,Y)$ *suivantes :*

$p(7,1) = p(3,3) = p(2,6) = 0$

$p(2,1) = p(3,5) = p(7,3) = p(7,6) = \frac{1}{20}$

$p(3,1) = p(7,3) = \frac{2}{20}$

$p(2,3) = \frac{3}{20},\ p(3,6) = \frac{4}{20},\ p(2,5) = \frac{5}{20}$

Calculer les probabilités marginales $p_X,\ p_Y$.

EXERCICE 3.2 *La temperature dans un four est une v.a. X avec densité*

$$f(x) = 11\,(1-x)^{10}\ ;\ \textit{pour}\ 0 \le x \le 1$$

Le four a un arrêt automatique de fonctionnement dès que la température dépasse la valeur $t_0 \in]0,1[$. *On voudrait que le four s'arrete avec une probabilité de* 10^{-22}. *Calculer la valeur de* t_0.

EXERCICE 3.3 *Une machine a 7 composants identiques qui fonctionnent indépendamment avec durée de vie des v.a.* $X_1,\ldots,X_7$ *respectivement. Notons par* $f(x)$ *la densité commune et* $F(x)$ *la fonction de repartition commune.*

(1) Admettons que la machine peut tomber en panne selon les quatre façons suivantes (qui sont exclusives) :

(a) si tous les composants sont HS ;

(b) si au moins un de ses composants est HS ;

(c) si un seul de composants est HS.

(d) s'il reste un seul composant en fonctionnement

Calculer dans chaque cas la probabilité que la machine ait fonctionné au moins 5 heures.

(2) Trouver la loi que chaque v.a. $X_i; i = 1,\ldots,7$ *suit et calculer l'espérance mathématique que la machine a fonctionné au moins 5 heures dans le cas (a) ci-dessus.*

STATISTIQUE COMPUTATIONNELLE – 4

ÉCHANTILLONNAGE

4.1
Introduction

Une étude statistique est une étude des caractéristiques (c-à-d. des grandeurs statistiques, moyenne, variance, moments, ...) des mesures de différents variables effectuées sur un ensemble d'objets qui forment une population.

Si à notre disposition on a les mesures sur la totalité de la population, on emploiera des outils de la statistique descriptive. Il s'agit en fait d'une situation qu'on rencontre rarement dans la réalité. Soit parce que la population est de taille très grande ou elle est dispersée aux quatre coins du monde, soit parce qu'il est très coûteux d'obtenir des mesures, soit ... . Dans ce cas on va se contenter d'un sous-ensemble de la population, l'*échantillon*, sur lequel nous allons faire l'étude des grandeurs statistiques et que nous allons, par la suite, extrapoler à la population entière. Cette extrapolation, qui part des résultats issus du particulier – un échantillon – pour conclure sur le général – la population – est l'*inférence statistique*. Son statut épistémologique est le même que celui de l'inférence en physique qui part des observations des résultats des expériences pour établir les lois de la nature. Sauf qu'en statistique les résultats des expériences sont de nature aléatoire, ce qui n'est pas le cas en physique, et on est, par conséquent, obligé en statistique de travailler dans un univers stochastique. Ce qui fait que l'inférence statistique est probable – et non pas certaine, comme en physique – et elle est donc associée à un risque quantifié de se tromper.

4.2

Principes de l'échantillonnage

Le caractère stochastique de la population implique que nous ne pouvons pas construire l'échantillon n'importe comment. Il faut d'une part faire des hypothèses et, d'autre part, respecter des contraintes.

Les hypothèses sont au nombre de deux, à savoir

- la population est considérée comme infinie ;
- la valeur de la mesure d'une variable d'un objet est résultat aléatoire et donc la variable qui représente cette mesure est une variable aléatoire qui suit une loi de probabilités ;
- les mesures de différents objets sont indépendantes les unes des autres, c-à-d. les résultats des mesures sur un objet ne conditionnent pas les résultats de ces mêmes mesures sur un autre objet. On dit dans ce cas que les v.a. qui représentent les mesures sont *indépendantes* ;
- toutes les mesures se sont effectuées en même temps. Bien que ceci est rarement le cas, on suppose que le temps n'influe pas sur les résultats des mesures. Par exemple si on fait un sondage d'opinion, on suppose qu'un individu interrogé à 9h00 du matin donnera les mêmes réponses que s'il était interrogé à 16h00 de l'après-midi du même jour. Cette hypothèse exprime le fait que les variables aléatoires qui représentent les mesures suivent la même loi de probabilité et avec les mêmes grandeurs statistiques. On dit dans ce cas que les v.a. sont *identiquement distribuées* (i.d.).

Il faut que ces hypothèses soient rigoureusement respectées si nous voulons que les résultats de l'inférence statistique soient pertinents.

4.3

Statistique inférentielle : objectifs

L'objectif de la statistique inférentielle est d'identifier les lois qu'elles suivent les v.a. de la population en examinant un échantillon de cette population à l'aide de différents types de méthodes.

Les lois de probabilités que suivent les v.a. sont

- soit totalement inconnues, et dans ce cas nous avons à résoudre de problèmes de la statistique inférentielle non-paramétrique ;
- soit partiellement connues, par exemple nous connaissons la loi mais pas son espérance et/ou sa variance. Dans ce cas on a de problèmes de la statistique inférentielle paramétrique.

4.4

Échantillonnage : méthodes

Il y a deux grandes catégories de méthodes

- méthodes empiriques, utilisées surtout par les instituts de sondage et qui sont fondées en particulier sur la technique des quotas ;
- méthodes aléatoires fondées sur le tirage aléatoire. On peut envisager
  - un tirage aléatoire sans remise. Tous les objets de la population ont la même probabilité d'être prélevés. Les v.a. associées aux objets sont indépendantes et identiquement distribuées.
  - un tirage stratifié. Les objets de la population sont partagés en sous-populations, que l'on appelle strates, et qui sont homogènes par rapport à un critère défini. Le tirage dans chaque strate se fait de façon aléatoire et le nombre d'objets prélevés dans chaque strate est proportionnel à l'effectif de la strate.
  - un tirage par grappe. On partage de façon aléatoire la population en sous-populations, les grappes, et on fait des tirages aléatoires dans chacune de grappes. Chaque grappe ne doit pas être homgène mais les différentes grappes doivent être assez ressemblantes du point de vue composition.

Dans la suite on se placera uniquement dans le cas du tirage aléatoire simple sans remise.

4.5

Les méthodes d'échantillonnage

Quelle que soit la technique d'échantillonnage utilisée, le contenu de l'échantillon prélevé varie d'un tirage à l'autre. Donc résultat d'un tirage est aléatoire.

Il y a deux façons différentes de modéliser cet aléa

(1) L'échantillon prélevé consiste en n réalisations $X_1, \ldots, X_n$ de la v.a. X.

(2) On associe au premier individu tiré une variable aléatoire de même loi que X. Elle vaut, selon le sondage. On fait de même pour les $n-1$ autres individus. $(X_1, \ldots, X_n)$, où X_i est la valeur de X pour le i-ième objet tiré, est un vecteur de v.a. i.i.d. de même loi que X. Un tirage correspond à une seule réalisation de celui-ci : $(x_1, \ldots, x_n) = (X_1(\omega), \ldots, X_n(\omega))$. Le vecteur $[X_1, \ldots, X_n]^\top$ est appelé échantillon aléatoire.

Le problème de l'estimation est d'étudier une statistique, c-à-d. une v.a. définie comme une fonction de l'échantillon aléatoire

$$S = f(X_1, \ldots, X_n)$$

Si $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$, la réalisation de S est donnée par $s = f(s_1, \ldots, s_n)$.

Les statistiques qu'on utilise le plus souvent sont la moyenne empirique de l'échantillon, la variance empirique, la covariance empirique.

4.6

Distributions d'échantillonnage

On s'intéresse à la caractéristique X d'une population avec X=v.a. On pose $E(X) = m$, $V(X) = s^2$.

On note $(X_1, \ldots, X_n)$ l'échantillon aléatoire associé à un tirage aléatoire simple de n individus de cette population et une réalisation de celui ci $(x_1, \ldots, x_n)$ c-à-d. un tirage particulier.

La définition de la *moyenne empirique* est

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

La loi de $\overline{X}_n$ est en général inconnue. On suppose que $\overline{X}_n$ est un estimateur de la moyenne μ_X de la population.

L'estimateur de la moyenne est

- sans biais : $E\left(\overline{X}_n\right) = \mu_X$;
- asymptotiquement efficace : $V\left(\overline{X}_n\right) = \frac{\sigma_X^2}{n}$;
- fortement convergent : $\overline{X}_n \rightarrow_{.ps} \mu_X$;
- approximativement gaussien lorsque n est grand. Dans ce cas on peut appliquer le TCL et on aura

$$\sqrt{n}\frac{\overline{X}_n - \mu_X}{\sigma_X} \rightarrow_l \mathcal{N}(0,1) \text{ ou } \sqrt{n}\left(\overline{X}_n - \mu_X\right) \rightarrow_l \mathcal{N}(0,\sigma_X)$$

La définition de la variance empirique est

$$S_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X}_n\right)^2$$

S_n^2 est un estimateur de la variance σ_X^2 de la population. Parfois on utilise la formule

$$S_{b,n}^2 = \frac{1}{n}\sum_{k=1}^{n}\left(X_k - \overline{X}_n\right)^2$$

qui fournit une estimateur biaisé de σ_X^2.

L'estimateur de la variance est

- sans biais : $E\left(S_n^2\right) = \sigma_X^2$;
- fortement convergent : $S_n^2 \rightarrow_{.ps} \sigma_X^2$;

La variance de cet estimateur est donnée par

$$V\left(S_n^2\right) = \frac{\mu^4 - \sigma_X^4}{n}$$

4.7
Exercices

EXERCICE 4.1 *Soit $\{X_n\}$ suite de v.a.i.i.d., qui représentent les valeurs d'une caractéristique d'une population mesurée dans un échantillon de taille n.*

- *La moyenne de l'échantillon est calculée par :*

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

- *La variance de l'échantillon est calculée par :*

$$S_1^2 = \sum_{k=1}^{n} \frac{\left(X_k - \overline{X}\right)^2}{n}$$

ou bien par la variance corrigée (sans biais) :

$$S^2 = \sum_{k=1}^{n} \frac{\left(X_k - \overline{X}\right)^2}{n-1}$$

- *L'écart-type de l'échantillon est donné par S ou S_1.*

On prélève un échantillon de taille 4 dont les valeurs sont 2, 3, 3.5 et 1.5.

(1) Calculer les caractéristiques de l'échantillon

(2) Calcul des distribution d'échantillonnage (lois suivies par les caractéristiques des lois).

(3) Calculer l'espérance et la variance de la moyenne expérimentale.

(4) Montrer que :

$$\frac{\overline{X} - \mu}{\sigma}\sqrt{n}$$

suit une loi $\mathcal{N}(0,1)$. Justifier votre réponse.

(5) Calculer l'espérance des v.a. S^2 et S_1^2.

EXERCICE 4.2 *Soit $(X_k)_{k=1,\ldots,n}$ un échantillon v.a. de loi uniforme dans $[0,1]$. Le programme Scilab* `lgn.sce` *utilise la fonction grand pour créer cet échantillon. Ensuite on calcule la somme cumulative des X_k que l'on stocke dans le vecteur $\mathbf{Y} = [Y_1, \ldots, Y_n]$ avec $Y_k = \sum_{i=1}^{k} X_i$, et on trace le diagramme (k, Y_k).*

Tester le programme pour différentes valeurs de n (par exemple $n = 100, 1000, 100000, 1000000$). Est-ce que les résultats sont en accord avec la loi des grands nombres ?

EXERCICE 4.3 *On considère un échantillon $(X_1, \ldots, X_n)$ où les variables aléatoires X_k sont i.i.d de*

même loi que $\sqrt{12p}\left(\frac{1}{p}\sum_{k=1}^{p} U_k - \frac{1}{2}\right)$ avec U_k v.a.i.i.d. de loi uniforme $[0,1]$.

Utiliser le programme `tcl.sce` *pour tracer l'histogramme de n_c classes de $X_k; k \leq n$. Qu'observez-vous pour $p = 1, p = 12$ n_c petit et n_c grand ? Pourriez-vous expliquer vos observations en utilisant le TCL ?*

N.B. On prendra $n = 1000$.