Rappels de probabilité
(extrait d'un poly d'O. Gaudoin, Ensimag)

Annexe A : Bases de probabilités pour la statistique

Cette annexe énonce quelques résultats de base du calcul des probabilités utiles pour la statistique. Les notions sont présentées sans aucune démonstration. Les détails ont été vus dans le cours de Probabilités Appliquées du premier semestre, ou seront développés dans les cours de probabilités de deuxième année.

7.1 Variables aléatoires réelles

7.1.1 Loi de probabilité d'une variable aléatoire

Mathématiquement, une variable aléatoire est définie comme une application mesurable. On se contentera ici de la conception intuitive suivante.

Une **variable aléatoire** est une grandeur dépendant du résultat d'une expérience aléatoire, c'est-à-dire non prévisible à l'avance avec certitude. Par exemple, on peut dire que la durée de vie d'une ampoule électrique ou le résultat du lancer d'un dé sont des variables aléatoires. Pour une expérience donnée, ces grandeurs prendront une valeur donnée, appelée réalisation de la variable aléatoire. Si on recommence l'expérience, on obtiendra une réalisation différente de la même variable aléatoire.

On ne s'intéresse ici qu'aux **variables aléatoires réelles**, c'est-à-dire à valeurs dans $\mathbb{R}$ ou un sous-ensemble de $\mathbb{R}$. On note traditionnellement une variable aléatoire par une lettre majuscule (X) et sa réalisation par une lettre minuscule (x).

Le calcul des probabilités va permettre de calculer des grandeurs comme la durée de vie moyenne d'une ampoule ou la probabilité d'obtenir un 6 en lançant le dé. Ces grandeurs sont déterminées par la **loi de probabilité** de ces variables aléatoires.

Il y a plusieurs moyens de caractériser la loi de probabilité d'une variable aléatoire. La plus simple est la fonction de répartition.

On appelle **fonction de répartition** de la variable aléatoire X la fonction

$$\begin{array}{rccl} F_X: & \mathbb{R} & \to & [0,1] \\ & x & \mapsto & F_X(x) = P(X \leq x) \end{array}$$

F_X est croissante, continue à droite, telle que $\lim_{x \to -\infty} F_X(x) = 0$ et $\lim_{x \to +\infty} F_X(x) = 1$. Elle permet de calculer la probabilité que X appartienne à n'importe quel intervalle de $\mathbb{R}$:

$$\forall (a,b) \in \mathbb{R}^2,\ a < b,\ P(a < X \leq b) = F_X(b) - F_X(a)$$

Les variables aléatoires peuvent être classées selon le type d'ensemble dans lequel elles prennent leurs valeurs. Dans la pratique, on ne s'intéressera qu'à deux catégories : les variables aléatoires discrètes et les variables aléatoires continues (ou à densité).

7.1.2 Variables aléatoires discrètes et continues

Une **variable aléatoire** X est dite **discrète (v.a.d.)** si et seulement si elle est à valeurs dans un ensemble E fini ou dénombrable. On peut noter $E = \{x_1, x_2, ...\}$.

Exemples :

- Face obtenue lors du lancer d'un dé : $E = \{1, 2, 3, 4, 5, 6\}$.
- Nombre de bugs dans un programme : $E = \mathbb{N}$.

La loi de probabilité d'une v.a.d. X est entièrement déterminée par les probabilités élémentaires $P(X = x_i)$, $\forall x_i \in E$.

La fonction de répartition de X est alors $F_X(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i)$.

Une **variable aléatoire** X est dite **continue (v.a.c.)** si et seulement si sa fonction de répartition F_X est continue et presque partout dérivable. Sa dérivée f_X est alors appelée densité de probabilité de X, ou plus simplement **densité** de X. Une v.a.c. est forcément à valeurs dans un ensemble non dénombrable.

Exemples :

- Appel de la fonction Random d'une calculatrice : $E = [0, 1]$.
- Durée de bon fonctionnement d'un système : $E = \mathbb{R}^+$.

On a alors $\forall (a,b) \in \mathbb{R}^2,\ a < b,\ P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)\,dx$.

Plus généralement, $\forall B \subset \mathbb{R}, P(X \in B) = \int_B f_X(x)\,dx$. Donc la densité détermine entièrement la loi de probabilité de X.

f_X est une fonction positive telle que $\int_{-\infty}^{+\infty} f_X(x)\,dx = P(X \in \mathbb{R}) = 1$.

Connaissant la loi de X, on est souvent amenés à déterminer celle de $Y = \varphi(X)$. Quand X est discrète, il suffit d'écrire $P(Y = y) = P(\varphi(X) = y)$. Si φ est inversible, on obtient $P(Y = y) = P(X = \varphi^{-1}(y))$. Quand X est continue, on commence par déterminer la fonction de répartition de Y en écrivant $F_Y(y) = P(Y \leq y) = P(\varphi(X) \leq y)$, puis on en déduit sa densité par dérivation. Quand φ est inversible, on obtient la **formule du changement de variable** :

$$f_Y(y) = \frac{1}{|\varphi'(\varphi^{-1}(y))|} f_X\left(\varphi^{-1}(y)\right)$$

Remarque : Il existe des lois de probabilité de variables aléatoires réelles qui ne sont ni discrètes ni continues. Par exemple, si X est la durée de bon fonctionnement d'un système

qui a une probabilité non nulle p d'être en panne à l'instant initial, on a $\lim_{x \to 0^-} F_X(x) = 0$ (une durée ne peut pas être négative) et $F_X(0) = P(X \leq 0) = P(X = 0) = p$. Par conséquent F_X n'est pas continue en 0. La loi de X ne peut donc pas être continue, et elle n'est pas non plus discrète puisqu'elle est à valeurs dans $\mathbb{R}^+$. Ce type de variable aléatoire ne sera pas étudié dans ce cours.

7.1.3 Moments et quantiles d'une variable aléatoire réelle

Si X est une variable aléatoire discrète, son **espérance mathématique** est définie par :

$$E[X] = \sum_{x_i \in E} x_i P(X = x_i)$$

Si X est une variable aléatoire continue, son espérance mathématique est définie par :

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x)\, dx$$

Concrètement, $E[X]$ est ce qu'on s'attend à trouver comme moyenne des résultats obtenus si on répète l'expérience un grand nombre de fois. Par exemple, si on lance une pièce de monnaie 10 fois, on s'attend à trouver en moyenne 5 piles.

Plus généralement, on peut s'intéresser à l'espérance mathématique d'une fonction de X :

- Si X est une v.a.d., $E[\varphi(X)] = \sum_{x_i \in E} \varphi(x_i) P(X = x_i)$.
- Si X est une v.a.c., $E[\varphi(X)] = \int_{-\infty}^{+\infty} \varphi(x) f_X(x)\, dx$.

Ce résultat permet de calculer l'espérance de $\varphi(X)$ sans avoir à déterminer entièrement sa loi.

Deux espérances de ce type sont particulièrement utiles :

- Si X est une v.a.d., sa **fonction génératrice** est définie par $G_X(z) = E\left[z^X\right] = \sum_{x_i \in E} z^{x_i} P(X = x_i)$.
- Si X est une v.a.c., sa **fonction caractéristique** est définie par $\phi_X(t) = E\left[e^{itX}\right] = \int_{-\infty}^{+\infty} e^{itx} f_X(x)\, dx$.

Au même titre que la fonction de répartition et la densité, les fonctions génératrices et caractéristiques définissent entièrement les lois de probabilité concernées.

Soit k un entier naturel quelconque. Le **moment d'ordre** k de X est $E\left[X^k\right]$ et le **moment centré d'ordre** k est $E\left[(X - E(X))^k\right]$.

De tous les moments, le plus important est le moment centré d'ordre 2, appelé aussi **variance**. La variance de X est $Var[X] = E\left[(X - E(X))^2\right]$, qui se calcule plus facilement sous la forme $Var[X] = E[X^2] - [E[X]]^2$.

L'**écart-type** de X est $\sigma[X] = \sqrt{Var[X]}$.

La variance et l'écart-type sont des indicateurs de la dispersion de X : plus la variance de X est petite, plus les réalisations de X seront concentrées autour de son espérance.

Le **coefficient de variation** de X est $CV[X] = \dfrac{\sigma[X]}{E[X]}$. C'est également un indicateur de dispersion, dont l'avantage est d'être sans dimension. Il permet de comparer les dispersions de variables aléatoires d'ordres de grandeur différents ou exprimées dans des unités différentes. En pratique, on considère que, quand $CV[X]$ est inférieur à 15%, l'espérance peut être considérée comme un bon résumé de la loi.

Soit $p \in]0,1[$. Le **quantile d'ordre** p (ou **p-quantile**) de la loi de X est tout réel q_p vérifiant $P(X < q_p) \leq p \leq P(X \leq q_p)$.

- Si F est continue et strictement croissante (donc inversible), on a simplement $P(X < q_p) = P(X \leq q_p) = F_X(q_p) = p$, d'où $q_p = F_X^{-1}(p)$.
- Si F_X est constante égale à p sur un intervalle $[a,b]$, n'importe quel réel de $[a,b]$ est un quantile d'ordre p. En général, on choisit de prendre le milieu de l'intervalle : $q_p = \dfrac{a+b}{2}$.
- Si F_X est discontinue en q et telle que $\lim\limits_{x \to q^-} F_X(x) < p \leq F_X(q)$, alors $q_p = q$.

Les tables fournies donnent les quantiles les plus usuels des lois normale, du chi-deux, de Student et de Fisher-Snedecor.

Covariance

Soient X, Y 2 variables aléatoires. La covariance de X et de Y :

$$Cov(X,Y) = E(XY) - (EX)(EY)$$

$$= E\left((X - EX)(Y - EY)\right)$$

7.3 Lois de probabilité usuelles

Les tables de lois de probabilité fournies donnent notamment, pour les lois les plus usuelles, les probabilités élémentaires ou la densité, l'espérance, la variance, et la fonction génératrice ou la fonction caractéristique. On présente dans cette section quelques propriétés supplémentaires de quelques unes de ces lois.

7.3.1 Loi binomiale

Une variable aléatoire K est de loi binomiale $\mathcal{B}(n,p)$ si et seulement si elle est à valeurs dans $\{0,1,...,n\}$ et $P(K=k) = C_n^k p^k (1-p)^{n-k}$.

Le nombre de fois où, en n expériences identiques et indépendantes, un évènement de probabilité p s'est produit, est une variable aléatoire de loi $\mathcal{B}(n,p)$.

La loi de Bernoulli $\mathcal{B}(p)$ est la loi $\mathcal{B}(1,p)$.

Si $X_1,\ldots,X_n$ sont indépendantes et de même loi $\mathcal{B}(m,p)$, alors $\sum_{i=1}^{n} X_i$ est de loi $\mathcal{B}(nm,p)$. En particulier, la somme de n v.a. indépendantes et de même loi $\mathcal{B}(p)$ est de loi $\mathcal{B}(n,p)$.

7.3.2 Loi géométrique

Une variable aléatoire K est de loi géométrique $\mathcal{G}(p)$ si et seulement si elle est à valeurs dans $\mathbb{N}^*$ et $P(K=k) = p\,(1-p)^{k-1}$.

Dans une suite d'expériences identiques et indépendantes, le nombre d'expériences nécessaires pour que se produise pour la première fois un évènement de probabilité p, est une variable aléatoire de loi $\mathcal{G}(p)$.

Si $X_1,\ldots,X_n$ sont indépendantes et de même loi $\mathcal{G}(p)$, alors $\sum_{i=1}^{n} X_i$ est de loi binomiale négative $\mathcal{BN}(n,p)$.

7.3.3 Loi de Poisson

Une variable aléatoire K est de loi de Poisson $\mathcal{P}(\lambda)$ si et seulement si elle est à valeurs dans $\mathbb{N}$ et $P(K=k) = e^{-\lambda}\dfrac{\lambda^k}{k!}$.

Pour $n \geq 50$ et $p \leq 0.1$, la loi binomiale $\mathcal{B}(n,p)$ peut être approchée par la loi de Poisson $\mathcal{P}(np)$. On dit que la loi de Poisson est la loi des évènements rares : loi du nombre de fois où un événement de probabilité très faible se produit au cours d'un très grand nombre d'expériences identiques et indépendantes.

Si $X_1, \ldots, X_n$ sont indépendantes et de même loi $\mathcal{P}(\lambda)$, alors $\sum_{i=1}^{n} X_i$ est de loi $\mathcal{P}(n\lambda)$.

7.3.4 Loi exponentielle

Une variable aléatoire X est de loi exponentielle $exp(\lambda)$ si et seulement si elle est à valeurs dans $\mathbb{R}^+$ et $f_X(x) = \lambda\, e^{-\lambda x}$.

La loi exponentielle est dite sans mémoire : $\forall (t,x) \in \mathbb{R}^{+^2}, P(X > t+x | X > t) = P(X > x)$.

Si $X_1, \ldots, X_n$ sont indépendantes et de même loi $exp(\lambda)$, alors $\sum_{i=1}^{n} X_i$ est de loi gamma $G(n,\lambda)$.

7.3.5 Loi gamma et loi du chi-2

Une variable aléatoire X est de loi gamma $G(a,\lambda)$ si et seulement si elle est à valeurs dans $\mathbb{R}^+$ et $f_X(x) = \dfrac{\lambda^a}{\Gamma(a)} e^{-\lambda x} x^{a-1}$. Les propriétés de la fonction gamma sont rappelées sur les tables.

La loi $G(1,\lambda)$ est la loi $exp(\lambda)$.

La loi $G\left(\dfrac{n}{2}, \dfrac{1}{2}\right)$ est appelée loi du chi-2 à n degrés de liberté, notée χ^2_n.

Si X est de loi $G(a,\lambda)$ et α est un réel strictement positif, alors αX est de loi $G\left(a, \dfrac{\lambda}{\alpha}\right)$.

Si X et Y sont des variables aléatoires indépendantes de lois respectives $G(\alpha,\lambda)$ et $G(\beta,\lambda)$, alors $X+Y$ est de loi $G(\alpha+\beta,\lambda)$. En particulier, si X et Y sont indépendantes et de lois respectives χ^2_n et χ^2_m, alors $X+Y$ est de loi χ^2_{n+m}.

7.3.6 Loi normale

Une variable aléatoire X est de loi normale $\mathcal{N}(m,\sigma^2)$ si et seulement si elle est à valeurs dans $\mathbb{R}$ et $f_X(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(x-m)^2}{2\sigma^2}}$.

Si X est de loi $\mathcal{N}(m, \sigma^2)$, alors $aX+b$ est de loi $\mathcal{N}(am+b, a^2\sigma^2)$. En particulier, $\dfrac{X-m}{\sigma}$ est de loi $\mathcal{N}(0,1)$.

$P(X \in [m-\sigma, m+\sigma]) = 68.3\%$. $\quad P(X \in [m-2\sigma, m+2\sigma]) = 95.4\%$.
$P(X \in [m-3\sigma, m+3\sigma]) = 99.7\%$.

Si X est de loi $\mathcal{N}(0,1)$, alors X^2 est de loi χ^2_1.

Si X et Y sont des variables aléatoires indépendantes de lois respectives $\mathcal{N}(m_1, \sigma_1^2)$ et $\mathcal{N}(m_2, \sigma_2^2)$, alors $aX + bY$ est de loi $\mathcal{N}\left(am_1 + bm_2, a^2\sigma_1^2 + b^2\sigma_2^2\right)$.

7.3.7 Lois de Student et de Fisher-Snedecor

Soit U une variable aléatoire de loi $\mathcal{N}(0,1)$ et X une variable aléatoire de loi χ^2_n. Si U et X sont indépendantes, alors $\sqrt{n}\,\dfrac{U}{\sqrt{X}}$ est de loi de Student à n degrés de liberté $St(n)$.

Soit X une variable aléatoire de loi χ^2_n et Y une variable aléatoire de loi χ^2_m. Si X et Y sont indépendantes, alors $\dfrac{mX}{nY}$ est de loi de Fisher-Snedecor $F(n,m)$.

Ces deux définitions entrainent que si T est de loi $St(n)$, alors T^2 est de loi $F(1,n)$.

Les lois de Student et de Fisher-Snedecor sont toujours utilisées par l'intermédiaire de tables ou à l'aide d'un logiciel de statistique. Il n'est donc pas nécessaire de donner l'expression de leur densité.