Rappel (très) informel de probabilité et statistique

September 17, 2015

1 Variable aléatoire (V.A.)

- Une variable X qui peut prendre différentes valeurs parmi un ensemble $\mathcal{X}$, mais va prendre certaines valeurs plus souvent que d'autres.
- Les valeurs sont donc plus ou moins probables. À chaque valeur est associée une "probabilité".

1.1 Variable aléatoire scalaire discrète (ou catégorique)

Ex: V.A. à valeur entière $\in \mathbb{N}$.

Ex: X est le résultat du lancé d'un dé à 6 faces. $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$
Si le dé n'est pas pipé, les probabilités associées à chaque valeur sont:

- $P(X = 1) = 1/6$
- $P(X = 2) = 1/6$
- ...
- $P(X = 6) = 1/6$

Ceci définit la loi de probabilité (ou distribution). Elle est le plus souvent représentée, comme ci-dessus par une **fonction de (masse) de probabilité** *(proability mass function p.m.f)*.

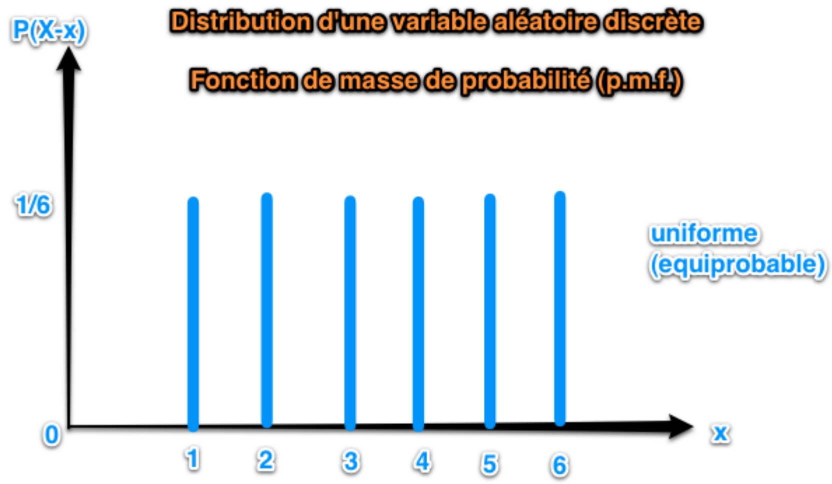
Notations équivalentes:

- $P(X = x)$ "probabilité que la variable aléatoire X prenne la valeur x"
- $P_X(x)$

Cette distribution peut aussi être représentée dans une table de probabilité:

x	1	2	3	4	5	6
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6

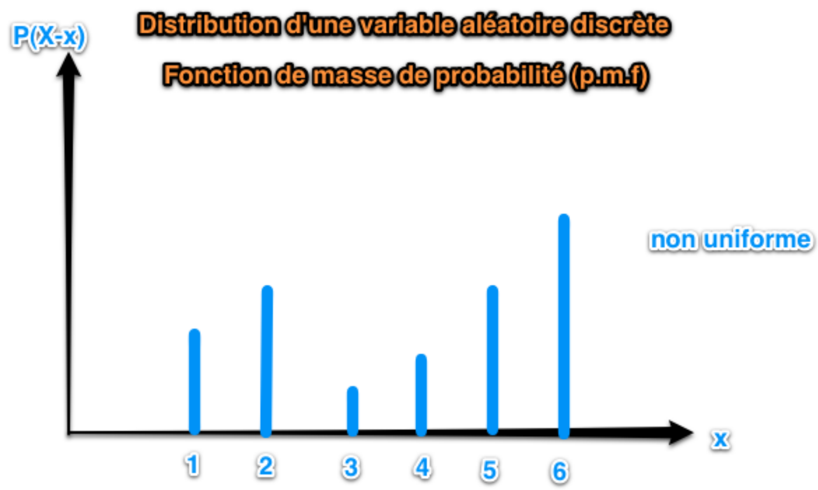
Ou bien par un graphique...



Pour un dé non pipé, les valeurs sont ***équiprobables***. Cela correspond à une **distribution uniforme** sur $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$.

Si le dé était pipé ou déséquilibré, on pourrait avoir des probabilités différentes pour chaque valeur discrète, ex:

x	1	2	3	4	5	6
$P(X = x)$	0.15	0.20	0.05	0.10	0.20	0.30



Propriétés de la fonction de (masse) de probabilité:

Si X peut prendre des valeurs dans l'ensemble $\mathcal{X}$, alors la p.m.f vérifie les propriétés suivantes:

- $\forall x \in \mathcal{X}, \quad P(X = x) \geq 0$
- $\left(\sum_{x \in \mathcal{X}} P(X = x)\right) = 1$

En utilisant l'autre notation:

- $\forall x \in \mathcal{X}, \quad P_X(x) \geq 0$
- $\left(\sum_{x \in \mathcal{X}} P_X(x)\right) = 1$

Une conséquence est également que les probabilités sont toujours ≤ 1.

Probabilité que X prenne une valeur dans un sous-ensemble:

La proabilité que X prenne une valeur dans un sous-ensemble $\mathcal{C} \subset \mathcal{X}$:

$$P(X \in \mathcal{C}) = \left(\sum_{x \in \mathcal{C}} P(X = x)\right) = \sum_{x \in \mathcal{C}} P_X(x)$$

Ex: probabilité que la valeur tirée avec un dé non pipé soit dans $\{2, 4, 6\}$ est de $\frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = 0.5 = 50\%$

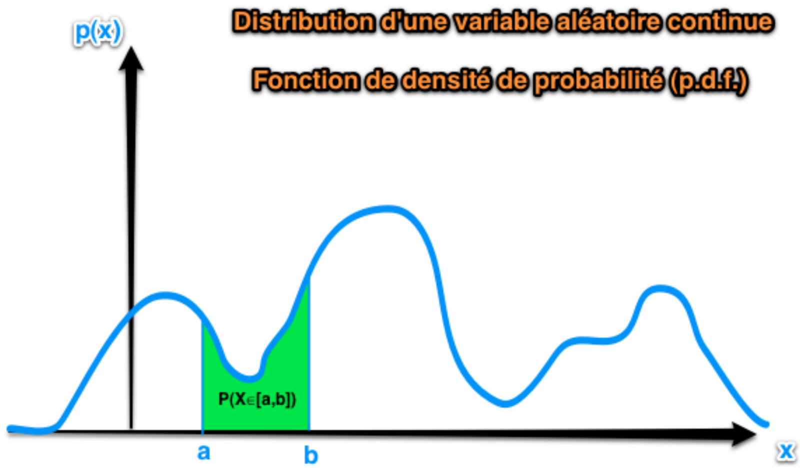
1.2 Variable aléatoire scalaire continue

Ex: V.A. dont les valeurs sont réelles $\in \mathbb{R}$

La loi de probabilité va souvent être représentée par une **fonction de densité de probabilité** *(probability density function p.d.f.)*.

Notée: $p_X(x) = p(x)$.

Ex:



Propriétés de la fonction de densité de probabilité:

Si X peut prendre des valeurs dans l'ensemble continu $\mathcal{X}$ (ex: $\mathbb{R}$), la p.d.f. p_X vérifie les propriétés:

- $\forall x \in \mathcal{X}, \quad p_X(x) \geq 0$
- $\left(\int_{\mathcal{X}} p_X(x)\,\mathrm{d}x\right) = 1$ "L'aire (surface) en dessous de la courbe de p.d.f." vaut 1.

Une conséquence est que conrairement aux (masses de) probabilités, **les densités de probabilités peuvent être plus grande que 1**!

Probabilité que X prenne une valeur dans un sous-ensemble:

La proabilité que X prenne une valeur dans un sous-ensemble $\mathcal{C} \subset \mathcal{X}$ est

$$P(X \in \mathcal{C}) = \int_{x \in \mathcal{C}} p_X(x) \mathrm{d}x$$

Là aussi cela correspond à la "l'aire (surface) en dessous de la courbe de p.d.f."dans la région $\mathcal{C}$.

Remarquez que la formule est similaire au cas discret, mais où on a remplacé la somme par une intégrale (et la p.m.f. par une p.d.f.).

- Remarque générale: pour les V.A. continues, pour un point précis x la probabilité $P(X = x)$ est en principe nulle, et ça n'a donc pas beaucoup de sens d'écrire $P(X = x)$. Ce qu'on veut peut-être exprimer est plutôt $P(X \in [x - \epsilon, x + \epsilon])$ qui a davantage de sens.

- Remarque plus avancée: en réalité certaines V.A. peuvent être des mélanges de distributions discrètes et continues, auquel cas on aura des masses de probabilité non-nulles pour certaines valeurs spécifiques (donc un $P(X = x)$ non-nul pour certaines valeurs, contrairement à la remarque précédente). En terme de densité de probabilité cela peut s'exprimer par un delta de Dirac (qui correspond à une densité infinie en ces points).

Règle informelle

Pour passer d'une formule utilisant une p.m.f d'une variable discrète, à la formule équivalente utilisant une p.d.f. d'une variable continue:

- remplacer les sommes par des intégrales

2 Espérance et Variance

2.1 Espérance

L'espérance d'une V.A. X, c'est "la moyenne" sur une infinité de tirage.

Soit un ensemble obtenu en effecuant n "tirages" d'une certaine distribution correspondant à une variable aléatoire X: $D = \{x^{(1)}, \ldots, x^{(n)}\}$

$$\mathbb{E}[X] = \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

Pour une V.A. discrète (catégorique):

$$\mathbb{E}[X] = \sum_{x\in\mathcal{X}} P_X(x)x$$

Pour une V.A. continue:

$$\mathbb{E}[X] = \int_{x\in\mathcal{X}} p_X(x)x \,\mathrm{d}x$$

Espérance d'une fonction

$$\mathbb{E}\left[f(X)\right] = \sum_{x\in\mathcal{X}} P_X(x)f(x)$$

ou pour une V.A. continue:

$$\mathbb{E}\left[f(X)\right] = \int_{x\in\mathcal{X}} p_X(x)f(x)$$

Linéarité ex:

$$\mathbb{E}\left[aX + bf(X) + cY + d\right] = a\mathbb{E}[X] + b\mathbb{E}[f(X)] + c\mathbb{E}[Y] + d$$

où X et$Y$ sont des $V.A.$ et a, b, c, d sont des scalaires.

2.2 Variance et écart type

Variance: *"la moyenne des carrés de l'écart à la moyenne"*

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2\end{aligned}$$

Écart type $\sigma_X = \sqrt{\mathrm{Var}[X]}$: racine carrée de la variance.

3 Cas "multivarié"

On peut le voir comme plusieurs (ex: d) variables aléatoires scalaires: $X_1, \dots, X_d$
Ou comme une seule variable qui est un Vecteur Aléatoire $X = (X_1, \dots, X_d)$.

3.1 Loi de probabilité jointe $P(X) = P(X_1, \dots, X_d)$

$P_X(x)$ mais cette fois ci $x \in \mathbb{R}^d$.

$$\begin{aligned}P_X(x) &= P(X = x) \\ &= P(X = (x_1, \dots, x_d)) \\ &= P(X_1 = x_1 \,\mathrm{ET}\, X_2 = x_2 \,\mathrm{ET}\, \dots \,\mathrm{ET}\, X_d = x_d) \\ &= P(X_1 = x_1,\, X_2 = x_2,\, \dots,\, X_d = x_d)\end{aligned}$$

Pour simplifier pour la suite, on va considérer $d = 2$, donc $X = (X_1, X_2)$

3.2 Loi marginale $P(X_1)$

Marginalization:

$$P(X_1 = x_1) = \sum_{x_2 \in \mathcal{X}_2} P(X_1 = x_1, X_2 = x_2)$$

où on lit la somme sur x_2 comme étant une somme sur toutes les valeurs possibles que peut prendre x_2

ce qu'on pourra écrire plus succintement comme:

$$P(x_1) = \sum_{x_2} P(x_1, x_2)$$

Avec plus que 2 variables cela devient:

$$P(x_1) = \sum_{x_2, \dots, x_d} P(x_1, x_2, \dots, x_d)$$

où la somme signifie la somme sur toutes les configurations possibles de valeurs de $x_2, \dots, x_d$:

$$
\begin{aligned}
P(x_1) &= \sum_{x_2,\ldots,x_d} P(x_1, x_2, \ldots, x_d) \\
&= \sum_{x_2 \in \mathcal{X}_2} \sum_{x_3 \in \mathcal{X}_3} \cdots \sum_{x_d \in \mathcal{X}_d} P(x_1, x_2, \ldots, x_d)
\end{aligned}
$$

3.3 Loi conditionnelle $P(X_1|X_2)$

"X_1 sachant X_2"

$$
P(X_1|X_2) = \frac{P(X_1, X_2)}{P(X_2)}
$$

$$
\begin{aligned}
P(X_1 = x_1|X_2 = x_2) &= \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = x_2)} \\
&= \frac{P(X_1 = x_1, X_2 = x_2)}{\sum_{x'} P(X_1 = x', X_2 = x_2)}
\end{aligned}
$$

ou bien plus succintement:

$$
\begin{aligned}
P(x_1|x_2) &= \frac{P(x_1, x_2)}{P(x_2)} \\
&= \frac{P(x_1, x_2)}{\sum_{x'} P(x', x_2)}
\end{aligned}
$$

Avec plusieurs variables cela devient:

$$
\begin{aligned}
P(x_1|x_2, \ldots, x_d) &= \frac{P(x_1, x_2, \ldots, x_d)}{P(x_2, \ldots, x_d)} \\
&= \frac{P(x_1, x_2, \ldots, x_d)}{\sum_{x'} P(x', x_2, \ldots, x_d)}
\end{aligned}
$$

3.4 Indépendance

Deux V.A. X_1 et X_2 sont **indépendantes** si et seulement si

$$
P(X_1, X_2) = P(X_1)P(X_2)
$$

Ce qui est équivalent à dire:

$$
P(X_1|X_2) = P(X_1)
$$

ou encore

$$P(X_2|X_1) = P(X_2)$$

"si deux V.A. sont indépendantes, connaître la valeur de l'une ne donne aucune information sur l'autre".

3.5 Règle de Bayes

Comment "intervertir" une probabilité conditionnelle

$$P(X_1|X_2) = \frac{P(X_2|X_1)P(X_1)}{P(X_2)}$$

Ce qui, écrit au long, signifie:

$$\begin{aligned} P(X_1 = x_1|X_2 = x_2) &= \frac{P(X_2 = x_2|X_1 = x_1)P(X_1 = x_1)}{P(X_2 = x_2)} \\ &= \frac{P(X_2 = x_2|X_1 = x_1)P(X_1 = x_1)}{\sum_{x'} P(X_2 = x_2, X_1 = x')} \\ &= \frac{P(X_2 = x_2|X_1 = x_1)P(X_1 = x_1)}{\sum_{x'} P(X_2 = x_2|X_1 = x')P(X_1 = x')} \end{aligned}$$

On peut résumer cela en écrivant:

$$P(X_1|X_2) \propto P(X_2|X_1)P(X_1)$$

où $\propto$ signifie "proportionnel à" et où le coefficient de proportionalité (le dénominateur dans les expressions ci-dessus) assure que les probabilités somment à 1.

Avec plusieurs variables, cela devient:

$$P(X_1|X_2, \ldots, X_d) \propto P(X_2, \ldots, X_d|X_1)P(X_1)$$

c.a.d.

$$\begin{aligned} P(X_1|X_2, \ldots, X_d) &= \frac{P(X_2, \ldots, X_d|X_1)P(X_1)}{P(X_2, \ldots, X_d)} \\ &= \frac{P(X_2, \ldots, X_d|X_1)P(X_1)}{\sum_{x'} P(X_1 = x', X_2, \ldots, X_d)} \end{aligned}$$

3.6 Décomposition générale

On peut toujours décomposer n'importe quelle distribution jointe en un produit de conditionnelles comme ceci:

$$P(X_1, X_2, \ldots, X_d) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)P(X_4|X_2, X_1, X_3) \ldots P(X_d|X_{d-1}, \ldots, X_1)$$

On pourrait tout aussi bien choisir un ordre différent pour les variabls (c'est arbitraire).

Cas où les variables sont indépendantes:

Dans le cas où toutes les V.A. sont **indépendantes**, l'expression ci-desus se simplifie en:

$$P(X_1, X_2, \ldots, X_d) = P(X_1)P(X_2)P(X_3) \ldots P(X_d)$$

3.7 Espérance et Covariance

$$\mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])$$

Covariance entre 2 variables:

$$\mathrm{Cov}[X_1, X_2] = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])\right]$$

Remarques:

$$\mathrm{Var}[X_1] = \mathrm{Cov}[X_1, X_1]$$

Matrice de covariance $C = \Sigma$ est telle que

$$C_{ij} = \mathrm{Cov}[X_i, X_j]$$

3.7.1 Espérance (moyenne) "empirique" et covariance "empirique":

Soit un ensemble obtenu en effecuant n "tirages" d'une certaine distribution correspondant à une variable aléatoire X de dimension d: $D = \{x^{(1)}, \ldots, x^{(n)}\}$. Chaque $x^{(t)}$ est ici considéré comme un vecteur colonne de dimension d.

Moyenne ("espérance empirique"):

$$\mu = \frac{1}{n} \sum_{t=1}^{n} x^{(t)}$$

Matrice de covariance empirique: $\Sigma = C$ telle que:

$$C_{ij} = \frac{1}{n} \sum_{t=1}^{n} \left(x_i^{(t)} - \mu_i\right)\left(x_j^{(t)} - \mu_j\right)$$

Autre manière de la calculer: une moyenne de matrices (chacune obtenues par un produit externe):

$$C = \frac{1}{n} \sum_{t=1}^{n} \left(x^{(t)} - \mu \right) \left(x^{(t)} - \mu \right)^T$$

3.8 Covariance, corrélation, indépendance

La corrélation est une version normalisée de la covariance:

$$\mathrm{Corr}[X_1, X_2] = \frac{\mathrm{Cov}[X_1, X_2]}{\sqrt{\mathrm{Var}[X_1]\mathrm{Var}[X_2]}}$$

On dit que deux V.A. sont **décorrélées** si leur corrélation (donc leur covariance) est 0.

$$X_1 \text{ et } X_2 \text{ décorrellées} \iff \mathrm{Cov}[X_1, X_2] = 0 \iff \mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]$$

$$\text{indépendance} \quad \Longrightarrow \quad \text{décorrélation}$$

Mais en général:

$$\text{décorrélation} \quad \not\Longrightarrow \quad \text{indépendance}$$

(sauf dans certains cas particuliers, ex: V.A. Gaussiennes).
L'indépendance est une propriété **beaucoup** plus forte que la décorrélation.

Rappel: indépendance signifie $P(X_1, X_2) = P(X_1)P(X_2)$

4 Quantités issues de la théorie de l'information

4.1 Entropie d'une distribution (ou d'une V.A.)

Entropie (mesure d'***"incertitude"***) d'une V.A. X ou de sa distribution p_X:

$$H(X) = H(p_X) = \mathbb{E}_{p_X}\left[-\log p_X(X)\right]$$

Unité d'entropie selon la base du logarithme: base 2 = "bit"; base e="nat".
Pour les V.A. continues on l'appelle plus précisément *entropie différentielle*.

4.2 Entropie croisée (entre 2 distributions)

$$H(p, q) = \mathbb{E}_p[-\log q]$$

donc dans le cas discret (fonctions de mase de probabilités):

$$H(P,Q) = -\sum_{x \in \mathcal{X}} P(X=x) \log Q(X=x)$$

et dans le cas continu (fonctions de densité de probsbilité):

$$H(p,q) = -\int_{x \in \mathcal{X}} p(x) \log q(x) \, \mathrm{d}x$$

4.3 Divergence de Kullback-Leibler

Un genre de ***"distance"*** entre 2 distributions p et q (mais pas symmétrique).

$$\begin{aligned} KL(p \| q) &= \mathbb{E}_p \left[\log \frac{p}{q} \right] \\ &= \mathbb{E}_p \left[\log p - \log q \right] \\ &= H(p,q) - H(p) \end{aligned}$$

donc dans le cas discret (fonctions de mase de probabilités):

$$KL(P \| Q) = \sum_{x \in \mathcal{X}} P(X=x) \log \frac{P(X=x)}{Q(X=x)}$$

et le cas continu (fonctions de densité de probsbilité)

$$KL(p \| q) = \int_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$$

La KL divergence est toujours positive ou nulle.
Et **elle est nulle si et seulement si** $p = q$.

4.4 Information mutuelle

C'est une mesure de ***dépendance*** statistique entre deux V.A.:

$$I(X_1, X_2) = KL(P(X_1, X_2) \parallel P(X_1)P(X_2))$$

Donc l'information mutuelle est nulle si et seulement si $P(X_1, X_2) = P(X_1)P(X_2)$ c.a.d. lorsque X_1 et X_2 sont indépendantes.