Analyse, Classification et Indexation des données (ACID)

Rappels de probabilités et de statistiques

Akka Zemmari

LaBRI, Université de Bordeaux

2024 - 2025

Pourquoi les probabilités

L'utilisation des probabilités et statistiques remonte à très loin ...

- premiers concepts apparus dans l'antiquité : jeux de hasard, etc
- les activités humaines sont affectées par le hasard :
  - les accidents,
  - le temps qu'il fait,
  - tirage au sort dans le sport,
  - étude des épédimies,
  - etc

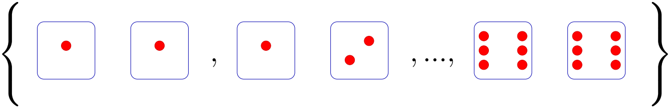
Définitions

- *Expérience aléatoire* : expérience où le hasard intervient rendant le résultat imprévisible.
- *Événement* : assertion relative au résultat d'une expérience aléatoire.
- La *probabilité* est une évaluation du caractère probable d'un événement.
- Le mot *probable* signifie : qui peut se produire dans l'expérience aléatoire.

Exemple

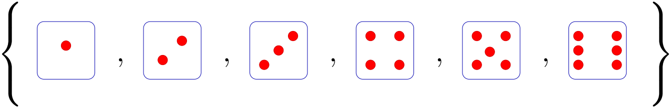
On lance deux dés :

$\Omega = D \times D =$



avec

$D =$



L'espace Ω contient $6 \times 6 = 36$ éléments.

Distribution uniforme (cas discret) :

Tous les éléments de Ω sont de même probabilité $\frac{1}{36}$

$$\Longleftrightarrow$$

Tous les éléments de D sont de même probabilité $\frac{1}{6}$.

Définitions. Un *espace probabilisé discret* est un couple $(\Omega, \mathbb{P}r)$, où Ω est un ensemble non vide au plus dénombrable[1] et $\mathbb{P}r$ une application de Ω dans $[0, 1]$ telle que :

$$\sum_{\omega \in \Omega} \mathbb{P}r(\omega) = 1.$$

[1] Voir le définition au tableau.

- Ω : *espace des événements élémentaires*
- $A \subset \Omega$: *événement*
- $\mathbb{P}r$: une *loi* (ou *distribution*) *de probabilité sur* Ω
- On prolonge $\mathbb{P}r$ sur $\mathcal{P}(\Omega)$ par :

$$\mathbb{P}r(A) = \sum_{\omega \in A} \mathbb{P}r(\omega), \quad \forall A \subset \Omega$$

$\mathbb{P}r(A)$: *probabilité de A*

Proposition.

- $\mathbb{P}r(\emptyset) = 0$
- $\mathbb{P}r(\bar{A}) = 1 - \mathbb{P}r(A)$
- $\mathbb{P}r(\cup_{i \in I} A_i) = \sum_{i \in I} \mathbb{P}r(A_i)$, pour toute famille au plus dénombrable $A_i, i \in I$ d'éléments de $\mathcal{P}(\Omega)$ 2-à-2 disjoints.

Probabilités conditionnelles

Définitions. Soit $(\Omega, \mathbb{P}r)$ un espace probabilisé discret et soit A un événement de probabilité non nulle. On définit sur $\mathcal{P}(\Omega)$, l'application $\mathbb{P}r(./A)$ à valeurs dans $[0,1]$ par :

$$\mathbb{P}r(B/A) = \frac{\mathbb{P}r(A \cap B)}{\mathbb{P}r(A)}, \quad \forall B \in \mathcal{P}(\Omega).$$

On appelle $\mathbb{P}r(B/A)$ *probabilité conditionnelle de B sachant A* (ou *probabilité de B sachant A*).

Probabilités conditionnelles

Nous avons alors :

$$\mathbb{P}r(A \cap B) = \mathbb{P}r(B/A)\mathbb{P}r(A).$$

Nous disons alors que l'on a *effectué un conditionnement par A.*

Proposition. (Loi des probabilités totales)
Soit $A_1, ..., A_n$ une partition de Ω. Si chacun de ces ensembles est de probabilité non nulle, alors :

$$\mathbb{P}r(B) = \sum_{i=1}^{n} \mathbb{P}r(A_i)\mathbb{P}r(B/A_i), \ \forall B \in \mathcal{P}(\Omega).$$

Exemple

Exemple . Le jeu du Monty Hall s'énonce comme suit :
Le jeu oppose un présentateur à un candidat (le joueur). Ce joueur est placé devant trois portes fermées. Derrière l'une d'elles se trouve une voiture et derrière chacune des deux autres se trouve une chèvre. Il doit tout d'abord désigner une porte. Puis le présentateur doit ouvrir une porte qui n'est ni celle choisie par le candidat, ni celle cachant la voiture (le présentateur sait quelle est la bonne porte dès le début). Le candidat a alors le droit ou bien d'ouvrir la porte qu'il a choisie initialement, ou bien d'ouvrir la troisième porte.
Quelle est la meilleure stratégie à adopter pour augmenter la probabilité de gain de la voiture

Indépendance

Définitions.

- Deux événements A et B sont dit *indépendants*, si :

$$\mathbb{P}r(A \cap B) = \mathbb{P}r(A)\mathbb{P}r(B).$$

Théorème de Bayes

Soit l'événement B causé par l'un des événements $A_1, A_2, ..., A_n$, tous de probabilité non nulle.
On connaît les probabilités $\mathbb{P}r(A_i)$ de ces derniers événements et aussi les probabilités conditionnelles $\mathbb{P}r(B/A_i)$.
Comment trouver les probabilités des causes sachant que B s'est produit, i.e. les probabilités $\mathbb{P}r(A_i/B)$?
Théorème (Formule de Bayes). Soit $A_1, ..., A_n$ une partition de Ω telle que $\mathbb{P}r(A_i) \neq 0, \forall i = 1, ..., n$.
Soit B un événement de probabilité non nulle. Nous avons :

$$\mathbb{P}r(A_i/B) = \frac{\mathbb{P}r(A_i)\mathbb{P}r(B/A_i)}{\sum_{j=1}^{n} \mathbb{P}r(A_j)\mathbb{P}r(B/A_j)}, \quad i = 1, 2, ..., n$$

Il s'agit de l'un des outils les plus importants dans ce qui va suivre dans cette UE ...

Théorème de Bayes

En particulier, pour deux événements A et B de probabilité non nulle, nous avons :

$$\mathbb{P}r(A/B) = \frac{\mathbb{P}r(A)\mathbb{P}r(B/A)}{\mathbb{P}r(B)}.$$

- $\mathbb{P}r(A)$: *probabilité a priori* de A
- $\mathbb{P}r(A/B)$: *probabilité a posteriori* de A

Théorème de Bayes

Exemple (F. Dress). Dans un jeu télévisé, un candidat doit choisir une question de repêchage en tirant au hasard parmi 3 papiers. Il y a :

- une question facile avec 3 chances sur 4 de donner la réponse correcte,
- une question moyenne avec 2 chances sur 5 de donner la réponse correcte, et
- une question difficile avec 1 chance sur 5 de donner la réponse correcte.

Sachant que le candidat a donné la réponse correcte, quelle est la probabilité pour qu'il s'agisse de la question facile ?

Théorème de Bayes

Solution : Désignons par F, M et D les événements tirage de la question facile, de la question moyenne et de la question difficile respectivement.
Soit, par ailleurs, l'événement "réponse correcte" noté par C.
D'après la formule de Bayes, nous avons :

$$Pr(F/C) = \frac{Pr(F)Pr(C/F)}{Pr(C/F)Pr(F) + Pr(C/M)Pr(M) + Pr(C/D)Pr(D)},$$

ce qui vaut :

$$\frac{\frac{1}{3}\cdot\frac{3}{4}}{\frac{1}{3}\cdot\frac{3}{4} + \frac{1}{3}\cdot\frac{2}{5} + \frac{1}{3}\cdot\frac{1}{5}} = \frac{5}{9}.$$

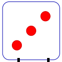
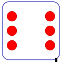
Variables aléatoires

Définition. Soit $(\Omega, \mathbb{P}r)$ un espace probabilisé discret et soit Ω' un ensemble non vide au plus dénombrable. Une *variable aléatoire* (v.a.) X à valeurs dans Ω' est une application de Ω dans Ω'. Nous prenons souvent pour Ω' un sous-ensemble de $\mathbb{N}$ (v.a. Discrète) ou de $\mathbb{R}$ (v.a. continue).
On pourra munir Ω' d'une loi de probabilité $\mathbb{P}r_X$ en posant, pour tout $\omega' \in \Omega'$:

$$\mathbb{P}r_X(\omega') = \mathbb{P}r(X^{-1}(\{\omega'\}))$$

Variables aléatoires

Exemple (cas discret (D)).

On lance deux dés. Soit $S(\omega)$ la v.a. qui est la somme des points des deux dés, par exemple $S($  ,  $) = 3 + 6 = 9$.
Le tableau ci-dessous caractérise la loi de probabilité pour la v.a. S par rapport à chacune des distributions (uniforme et biaisée):

s	2	3	4	5	6	7	8	9	10	11	12
$\mathbb{P}r_{11}(S = s)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
$\mathbb{P}r_{22}(S = s)$	$\frac{4}{64}$	$\frac{4}{64}$	$\frac{5}{64}$	$\frac{6}{64}$	$\frac{7}{64}$	$\frac{12}{64}$	$\frac{7}{64}$	$\frac{6}{64}$	$\frac{5}{64}$	$\frac{4}{64}$	$\frac{4}{64}$

Variables aléatoires

Exemple (cas continu(C)).

Une procédure génère les réels X et Y uniformément et indépendamment dans l'intervalle $[a, b]$, où $a < b$. On s'intéresse alors au maximum des deux valeurs.
Quelle est la probabilité que la valeur retournée soit $\leq z \in [a, b]$?
Quelle est la probabilité qu'elle appartienne à l'intervalle $[c, d] \subset [a, b]$?
Soit $Z = \max\{X, Y\}$ et $z \in [a, b]$.

$$\begin{aligned}\mathbb{P}r\,(Z \leq z) &= \mathbb{P}r\,(X \leq z \text{ et } Y \leq z) = \mathbb{P}r\,(X \leq z) \times \mathbb{P}r\,(Y \leq z) \\ &= \frac{(z-a)^2}{(b-a)^2}.\end{aligned}$$

On en déduit :

$$\mathbb{P}r\,(c < Z \leq d) = \frac{(d-a)^2}{(b-a)^2} - \frac{(c-a)^2}{(b-a)^2}.$$

Fonction de Répartition (CDF)

Soit $\mathbb{P}r$ une mesure de probabilité. Sa *fonction de répartition* (f.r.) F est définie par :

$$F(t) = \mathbb{P}r\left(]-\infty, t]\right), \quad \forall t \in \mathbb{R}.$$

Densité (PDF)

Soit F la f.r. de la distribution $\mathbb{P}r$. S'il existe une fonction f telle que :

$$F(t) = \int_{-\infty}^{t} f(x)dx, \quad \forall t \in \mathbb{R},$$

alors cette fonction f s'appelle *densité* de $\mathbb{P}r$.
La f.r. étant croissante, si la densité f existe, son intégrale de $-\infty$ à t, doit valoir $F(t)$. Donc, on peut définir la densité f comme la dérivée de de la f.r., si cette dernière en admet une.

Variables aléatoires

Propriétés des PDF et CDF :

Cas	CDF F	PDF f	Propriétés
(D)	$F(t) = \sum_{x_i \leq t} \mathbb{P}r\,(X = x_i)$	$f(x_i) = \mathbb{P}r\,(X = x_i)$	$0 \leq f(x_i) \leq 1$ et $\sum_i f(x_i) = 1$
(C)	$F(t) = \int_{-\infty}^{t} f\,(x)\,dx$	$f(x) = \frac{dF}{dx}$	$f(x) \geq 0$ et $\int_{-\infty}^{+\infty} f(x)dx = 1$

Espérance Mathématique et Variance

Une variable aléatoire admet un certain nombre de valeurs typiques. Nous considérons dans la suite les v.a. à valeurs dans $\mathbb{R}$.
Définition. En arithmétique usuelle la valeur moyenne de n nombres est définie comme leur somme divisée par n. En calcul des probabilités, l'*espérance* d'une v.a. est définie comme la somme des valeurs prises pondérées par les probabilités respectives.

Cas	$\mathbb{E}(X)$	$\mathbb{E}(g(X))$
(D)	$\sum_{i=1}^{n} x_i \mathbb{P}r(X = x_i)$	$\sum_{i=1}^{n} g(x_i) \mathbb{P}r(X = x_i)$
(C)	$\int_{-\infty}^{+\infty} x f(x)\, dx$	$\int_{-\infty}^{+\infty} g(x) f(x)\, dx$

Espérance Mathématique et Variance

Exemple (D).

Revenons au cas de la v.a. S, désignant la somme des deux dés, et calculons son espérance pour chacune des distributions $\mathbb{P}r_{11}$ et $\mathbb{P}r_{22}$:

- $\mathbb{E}_{11}(S) = 2.\frac{1}{36} + 3.\frac{2}{36} + ... + 12.\frac{1}{36} \;\; = 7.$
- $\mathbb{E}_{22}(S) = 2.\frac{4}{64} + 3.\frac{4}{64} + 4.\frac{5}{64} + ... + 12.\frac{4}{64} \;\; = 7.$

Espérance Mathématique et Variance

Définitions. Un paramètre important qui vient tout de suite après l'espérance est la variance. Elle mesure la dispersion d'une v.a. Si X est une v.a. définie sur $(\Omega, \mathbb{P}r)$, sa *variance* est définie par :

$$\mathbb{V}ar(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$$

lorsque celle-ci existe.

Espérance Mathématique et Variance

Définition. La racine carrée de la variance est appelée *écart-type* et est notée par σ :

$$\sigma(X) = \sqrt{\mathbb{V}ar(X)}.$$

Proposition. Nous avons :

$$\mathbb{V}ar(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

Proposition. Si X et Y sont deux v.a. indépendantes admettant chacune une variance, alors la v.a. $X + Y$ admet une variance qui est la somme des deux variances.

Distributions importantes

Type	**Distribution**	**PDF**	$\mathbb{E}(X)$	$\mathbb{V}ar(X)$
(D)	$X \sim \mathcal{B}(n,p)$	$\mathbb{P}r(X=k) = \binom{n}{k} p^k q^{n-k}$	np	npq
(D)	$X \sim \mathcal{P}(\lambda)$	$\mathbb{P}r(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$	λ	λ
(C)	$X \sim \mathcal{U}(a,b)$	$f(x) = \frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
(C)	$X \sim \mathcal{E}(\lambda)$	$f(x) = \lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
(C)	$X \sim \mathcal{N}(\mu,\sigma)$	$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ	σ

Inférences Statistiques

- Le calcul des probabilités permet, à partir d'un modèle théorique, d'associer à un événement une probabilité qui mesure la fréquence de son apparition dans une suite d'épreuves identiques.
- Les techniques statistiques vont nous aider à reconstruire le modèle ou nous autoriser à nous poser un certain nombre de questions sur le modèle à partir d'un bon nombre d'observations (un échantillon) de réalisations de l'événement étudié.
- Ces inférences sont fondées sur des critères tels que la vraisemblance des valeurs observées, l'absence de biais ou la convergence des prévisions suggérées par l'échantillon lorsque la taille de celui-ci augmente.

Quelques Estimateurs Standards

Estimateur standard de l'espérance.

Soit $(x_1, \dots, x_n)$ un échantillon d'une v.a. possédant une espérance inconnue μ. Son estimateur standard est la moyenne de l'échantillon:

$$\hat{\mu}_n(x_1, \dots, x_n) \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Dans la suite, pour alléger la notation, nous supprimons la référence de l'estimateur à l'indice n et aux arguments $x_1, \dots, x_n$, lorsqu'il n'y a pas de risque de confusion.

Estimateur standard d'une distribution de probabilité finie.

Soit $(x_1, \ldots, x_n)$ un échantillon d'une v.a. X prenant une des valeurs $\alpha_1, \ldots, \alpha_k$ avec les probabilités inconnues $p_1, \ldots, p_k$ respectivement. L'estimateur standard du vecteur $p = (p_1, \ldots, p_k)$ est défini par :

$$\hat{p}_i = \frac{1}{n} \sum_{j=1}^{n} 1_{\{\alpha_i\}}(x_j), \quad i = 1, \ldots, k.$$

Autrement dit, l'estimateur standard de chaque p_i est la fréquence relative d'occurrences de α_i dans l'échantillon.

Estimateur standard de la variance.

Soit X une v.a. admettant une espérance μ et une variance V, toutes deux inconnues. Soit $(x_1, ..., x_n)$ un échantillon de X. L'estimateur standard joint $(\hat{\mu}, \hat{V})$ est donné par :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{et} \quad \hat{V} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.$$

Estimateur standard de l'écart-type.

Il est donnée par :

$$\hat{\sigma} = \sqrt{\hat{V}},$$

où $\hat{V}$ est l'estimateur standard de V.