Notions de probabilités et de statistiques

1. Généralités

Événement. Notons Ω l'ensemble des *événements élémentaires* pouvant résulter d'un phénomène aléatoire. Un *événement* ω est un sous-ensemble de Ω constitué d'un ou plusieurs événements élémentaires.

On note $P(\omega)$ la *probabilité* que ω se réalise. On a $0 \leq P(\omega) \leq 1$ et, en particulier, $P(\Omega) = 1$.

Négation. La probabilité que ω ne se réalise pas est $\boxed{P(\Omega \setminus \omega) = 1 - P(\omega)}$.

En particulier, $P(\emptyset) = 0$.

Et. On note $P(\omega_1 \cap \omega_2)$ la probabilité que les événements ω_1 et ω_2 se réalisent tous deux.

Si ω_1 et ω_2 sont des événements exclusifs (c.-à-d. incompatibles), $\boxed{P(\omega_1 \cap \omega_2) = 0}$.

Ou. On note $P(\omega_1 \cup \omega_2)$ la probabilité que l'événement ω_1 ou[1] l'événement ω_2 se réalise. On a

$$P(\omega_1 \cup \omega_2) = P(\omega_1) + P(\omega_2) - P(\omega_1 \cap \omega_2).$$

En particulier, *si ω_1 et ω_2 sont exclusifs*, $\boxed{P(\omega_1 \cup \omega_2) = P(\omega_1) + P(\omega_2)}$.

Probabilité conditionnelle. La probabilité conditionnelle que ω_1 se réalise, sachant que ω_2 s'est réalisé, est

$$\boxed{P(\omega_1 | \omega_2) = \frac{P(\omega_1 \cap \omega_2)}{P(\omega_2)}}.$$

Si ω_1 et ω_2 sont indépendants, $P(\omega_1 | \omega_2) = P(\omega_1)$, donc

$$\boxed{P(\omega_1 \cap \omega_2) = P(\omega_1) P(\omega_2)}.$$

2. Variable aléatoire, loi de probabilité

Variable aléatoire discrète. On considère la variable aléatoire X qui prend des valeurs x_i, où $i \in \mathbf{N}$.

La *moyenne* de la loi de probabilité P suivie par X est

$$\boxed{M(X) = \sum_i x_i P(X = x_i)}$$

et sa *variance* est

$$\boxed{V(X) = \sum_i [x_i - M(X)]^2 P(X = x_i)}.$$

1. En l'absence de précision, en mathématiques comme en français, le « ou » est inclusif, c.-à-d. que $\omega_1 \cup \omega_2$ recouvre, entre autres, le cas où *et* ω_1 *et* ω_2 se réalisent.

Variable aléatoire continue. On considère la variable aléatoire X qui prend des valeurs x dans $\mathbf{R}$. La probabilité que X soit dans $[x, x+\mathrm{d}x]$ est

$$\boxed{P(X \in [x, x+\mathrm{d}x]) = D(x)\,\mathrm{d}x},$$

où $D(x)$ est la *densité de probabilité*. On appelle *fonction de répartition* la probabilité $P(X \leq x)$. On a

$$\boxed{P(X \leq x) = \int_{x'=-\infty}^{x} D(x')\,\mathrm{d}x'}.$$

La *moyenne* de la loi de probabilité suivie par X est

$$\boxed{M(X) = \int_{x=-\infty}^{+\infty} x \cdot D(x)\,\mathrm{d}x}$$

et sa *variance* est

$$\boxed{V(X) = \int_{x=-\infty}^{+\infty} [x - M(X)]^2 D(x)\,\mathrm{d}x}.$$

Que la variable aléatoire soit discrète ou continue, on a

$$\boxed{V(X) = M(X^2) - M^2(X)}.$$

On appelle *écart-type* de X la quantité

$$\boxed{S(X) = \sqrt{V(X)}}.$$

On utilise aussi souvent la *médiane* $m(X)$ définie par

$$\boxed{P[X \leq m(X)] = 1/2}$$

(c.-à-d. qu'il y a autant de chances d'obtenir une valeur au-dessus de $m(X)$ qu'en dessous).

3. Loi binômiale

On considère n événements ω_i indépendants de probabilité p chacun. Soit X la variable aléatoire discrète égale au nombre $k \in [0 \ldots n]$ d'événements qui se sont réalisés parmi les n. En notant $k!$ la factorielle de k définie par

$$0! = 1 \quad \text{et} \quad k! = 1 \times 2 \times 3 \times \cdots \times (k-1) \times k,$$

on a

$$\boxed{P(X = k) = C_n^k p^k (1-p)^{n-k}} \quad (\textit{loi binômiale}),$$

où

$$\boxed{C_n^k = \frac{n!}{k!(n-k)!}}.$$

En utilisant la formule du binôme,

$$(a+b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k},$$

avec $a = p$ et $b = 1 - p$, on obtient bien

$$\sum_{k=0}^{n} P(X = k) = 1.$$

On a également

$$\boxed{M(X) = n{\cdot}p}$$

et

$$\boxed{V(X) = n{\cdot}p{\cdot}(1-p)}.$$

Convergence vers la loi normale. En utilisant la formule de Stirling,

$$k! \approx \sqrt{2\pi}\, k^{k+1/2} \mathrm{e}^{-k},$$

on peut montrer que, pour $n{\cdot}p{\cdot}(1-p)$ suffisamment grand,

$$\boxed{P(X=k) \approx G_{\mu,\sigma^2}(k)},$$

où

$$\boxed{\mu = M(X)},$$

$$\boxed{\sigma^2 = V(X)}$$

et

$$\boxed{G_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]}$$

est la *loi normale* (ou *gaussienne*) de paramètres μ et σ^2.

Preuve sommaire.

$$\begin{aligned}
P\left(X=k\right) &\approx \frac{\sqrt{2\pi}\exp\left[\left(n+\frac{1}{2}\right)\ln n - n\right]}{\sqrt{2\pi}\exp\left[\left(k+\frac{1}{2}\right)\ln k - k\right]\cdot\sqrt{2\pi}\exp\left[\left(n-k+\frac{1}{2}\right)\ln\left(n-k\right)-\left(n-k\right)\right]} p^k \left(1-p\right)^{n-k} \\
&= \frac{\exp\left[\left(n+\frac{1}{2}\right)\ln n - \left(k+\frac{1}{2}\right)\ln k - \left(n-k+\frac{1}{2}\right)\ln\left(n-k\right)\right] n^{\frac{1}{2}} p^{k+\frac{1}{2}} \left(1-p\right)^{n-k+\frac{1}{2}}}{\sqrt{2\pi}\sqrt{n{\cdot}p{\cdot}\left(1-p\right)}} \\
&= \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\left(n+\frac{1}{2}\right)\ln n - \left(k+\frac{1}{2}\right)\ln k - \left(n-k+\frac{1}{2}\right)\ln\left(n-k\right)\right. \\
&\qquad \left. +\frac{1}{2}\ln n + \left(k+\frac{1}{2}\right)\ln p + \left(n-k+\frac{1}{2}\right)\ln\left(1-p\right)\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\left(k+\frac{1}{2}\right)\ln\left(\frac{k}{np}\right) - \left(n-k+\frac{1}{2}\right)\ln\left(\frac{n-k}{n{\cdot}\left(1-p\right)}\right)\right].
\end{aligned}$$

Posons $k = \mu + \delta$. Le terme entre crochets vaut

$$\begin{aligned}
&-\left(k+\frac{1}{2}\right)\ln\left(\frac{k}{np}\right) - \left(n-k+\frac{1}{2}\right)\ln\left(\frac{n-k}{n{\cdot}\left(1-p\right)}\right) \\
&\quad = -\left(\mu+\delta+\frac{1}{2}\right)\ln\left(1+\frac{\delta}{\mu}\right) - \left(n-\mu-\delta+\frac{1}{2}\right)\ln\left(1-\frac{\delta}{n-\mu}\right) \\
&\quad = -\left(\mu+\delta+\frac{1}{2}\right)\left(\frac{\delta}{\mu}-\frac{\delta^2}{2\mu^2}+\cdots\right) - \left((n-\mu)-\delta+\frac{1}{2}\right)\left(-\frac{\delta}{n-\mu}-\frac{\delta^2}{2(n-\mu)^2}+\cdots\right) \\
&\quad \approx -\frac{\delta^2}{2}\left(\frac{1}{\mu}+\frac{1}{n-\mu}\right) = -\frac{\delta^2}{2\sigma^2},
\end{aligned}$$

d'où $P(X=k) \approx G_{\mu,\sigma^2}(k)$.

La probabilité que X prenne une valeur k comprise entre k_1 et k_2 est

$$P(X \in [k_1 \ldots k_2]) \approx \sum_{k=k_1}^{k_2} G_{\mu,\sigma^2}(k),$$

puisque les événements $k = k_1$, $k = k_1 + 1$, ..., $k = k_2 - 1$ et $k = k_2$ sont incompatibles.

$$P(X \in [k_1 \dots k_2]) \approx \sum_{k=k_1}^{k_2} G_{\mu,\sigma^2}(k) \overbrace{[(k+1/2) - (k-1/2)]}^{\Delta k=1}$$
$$\approx \int_{x=k_1-1/2}^{k_2+1/2} G_{\mu,\sigma^2}(x)\,\mathrm{d}x = \int_{x=-\infty}^{k_2+1/2} G_{\mu,\sigma^2}(x)\,\mathrm{d}x - \int_{x=-\infty}^{k_1-1/2} G_{\mu,\sigma^2}(x)\,\mathrm{d}x.$$

4. Loi normale

La loi normale est une approximation de la loi binômiale valable quand $n{\cdot}p{\cdot}(1-p)$ est grand. Elle est aussi très souvent suivie par des variables aléatoires continues (c.-à-d. que leur densité de probabilité est de la forme $D(x) = G_{\mu,\sigma^2}(x)$). Ceci est dû notamment au théorème[2] suivant (en version simplifiée) :

Théorème de la limite centrale. Si les variables aléatoires Y_i, $i \in [1 \dots n]$, sont indépendantes et suivent la même loi, de moyenne $M(Y_i) = M_Y$ et de variance $V(Y_i) = V_Y$ communes, la moyenne $M(Z)$ de la variable $Z = \frac{1}{n}\sum_{i=1}^{n} Y_i$ tend à suivre la loi normale de paramètres $\mu = M_Y$ et $\sigma^2 = V_Y/n$ quand n tend vers l'infini.

Plus généralement, la somme de nombreuses variables aléatoires, même si celles-ci ne suivent pas la même loi, sera très souvent distribuée selon une loi à peu près gaussienne.

Propriétés de la loi normale. On considère une variable aléatoire continue X distribuée selon la loi normale de paramètres μ et σ^2.

On peut vérifier que l'on a bien

$$P(X \in]-\infty, +\infty[) = \int_{x=-\infty}^{+\infty} G_{\mu,\sigma^2}(x)\,\mathrm{d}x = 1,$$

$$\boxed{M(X) = \int_{x=-\infty}^{+\infty} x{\cdot}G_{\mu,\sigma^2}(x)\,\mathrm{d}x = \mu}$$

et

$$\boxed{V(X) = \int_{x=-\infty}^{+\infty} (x-\mu)^2 G_{\mu,\sigma^2}(x)\,\mathrm{d}x = \sigma^2}.$$

La loi normale est symétrique autour de sa moyenne (qui est également son pic et sa médiane). La largeur à mi-hauteur du pic est

$$l_{\mathrm{mh}} = 2\sqrt{2\ln 2}\,\sigma.$$

En faisant le changement de variable $\boxed{y = (x-\mu)/\sigma}$, on remarque aussi que la fonction de répartition vaut

$$\boxed{P(X \le x) = \int_{x'=-\infty}^{x} G_{\mu,\sigma^2}(x')\,\mathrm{d}x' = \int_{y'=-\infty}^{y} G_{0,1}(y')\,\mathrm{d}y' = F(y)}$$

($G_{0,1}$ est la *loi normale centrée réduite*). $F(y)$ n'est pas calculable analytiquement, mais on peut prendre sa valeur dans une table. $F(y)$ a les propriétés suivantes :

$$F(-\infty) = 0\,;$$
$$P(X \le \mu) = F(0) = 1/2\,;$$
$$F(+\infty) = 1\,;$$
$$F(-y) = 1 - F(y).$$

2. La convergence de la loi binômiale vers la loi normale peut d'ailleurs se déduire de ce théorème en prenant pour Y_i la variable aléatoire qui vaut 1 si ω_i est réalisé et 0 sinon. On a $M(Y_i) = p \times 1 + (1-p) \times 0 = p$ et $V(Y_i) = p \times (1-p)^2 + (1-p) \times p^2 = p{\cdot}(1-p)$. En posant $X = nZ = \sum_{i=1}^{n} Y_i$, c.-à-d. le nombre d'événements qui se réalisent, on obtient bien que X tend vers $G_{n\cdot p,n\cdot p\cdot(1-p)}$.

Les intervalles de confiance suivants sont particulièrement utilisés :

$$\boxed{P(X \in [\mu - \sigma, \mu + \sigma]) = \int_{x=\mu-\sigma}^{\mu+\sigma} G_{\mu,\sigma^2}(x)\,\mathrm{d}x = F(1) - F(-1) = 68{,}3\,\%} ;$$

$$\boxed{P(X \in [\mu - 2\sigma, \mu + 2\sigma]) = \int_{x=\mu-2\sigma}^{\mu+2\sigma} G_{\mu,\sigma^2}(x)\,\mathrm{d}x = F(2) - F(-2) = 95{,}4\,\%} ;$$

$$\boxed{P(X \in [\mu - 3\sigma, \mu + 3\sigma]) = \int_{x=\mu-3\sigma}^{\mu+3\sigma} G_{\mu,\sigma^2}(x)\,\mathrm{d}x = F(3) - F(-3) = 99{,}7\,\%}$$

(la probabilité d'obtenir une valeur à plus de trois écarts-types de la moyenne, dans les « ailes » de la gaussienne, est donc très faible).

5. Statistiques

La loi suivie par une variable aléatoire est généralement inconnue. Même si sa forme est connue (p. ex. normale), ses paramètres (p. ex. μ et σ^2) le sont rarement. On doit les estimer à partir d'un *échantillon* de mesures $\xi_i, i \in [1 \dots n]$. Un cas fréquent est celui où l'on effectue n mesures indépendantes d'une même quantité de la même façon (il faut notamment que l'état de l'appareil n'ait pas évolué au cours de l'expérience et qu'aucune mesure ne soit influencée par les précédentes). Les mesures ξ_i suivent donc la même loi de probabilité.

Ces mesures sont souvent regroupées dans des intervalles distincts $[\zeta_j^-, \zeta_j^+]$ de centre ζ_j ($j \in [1 \dots k]$). On note n_j le nombre de mesures d'un intervalle j quelconque. Une estimation $\widehat{P}_j$ de la probabilité P_j que X soit dans l'intervalle $[\zeta_j^-, \zeta_j^+]$ est alors fournie par la *fréquence* observée

$$\boxed{\widehat{P}_j = n_j / n}.$$

Il est préférable que l'intervalle ne soit ni trop petit (car il contiendrait peu de mesures et une variation de n_j de quelques unités se traduirait par une variation importante de $\widehat{P}_j$) ni trop grand (car la loi de probabilité risquerait de changer significativement entre ses deux extrémités).

La moyenne $\widehat{M}$ de l'échantillon,

$$\boxed{\widehat{M} = \frac{\sum_{i=1}^{n} \xi_i}{n} \approx \sum_{j=1}^{k} \zeta_j \widehat{P}_j},$$

est une estimation de la moyenne M de la loi suivie par X.

La variance $\widehat{V}$ de l'échantillon,

$$\boxed{\widehat{V} = \frac{\sum_{i=1}^{n} (\xi_i - \widehat{M})^2}{n} \approx \sum_{j=1}^{k} (\zeta_j - \widehat{M})^2 \widehat{P}_j},$$

est une estimation de la variance V de la loi, mais, comme $\widehat{V}$ utilise $\widehat{M}$ au lieu de M, $\widehat{V}$ sous-estime légèrement V. Une meilleure estimation de V est

$$\boxed{\widehat{V}' = \frac{\sum_{i=1}^{n} (\xi_i - \widehat{M})^2}{n-1}}.$$

Pour calculer la médiane $\widehat{m}$ de l'échantillon, il faut ordonner les ξ_i. Notons ξ_i^* les éléments ξ_i rangés par ordre croissant.

$$\boxed{\widehat{m} = \begin{cases} \frac{\xi_{n'}^* + \xi_{n'+1}^*}{2} & \text{si } n \text{ est pair, avec } n = 2n' \text{ et} \\ \xi_{n'+1}^* & \text{si } n \text{ est impair, avec } n = 2n' + 1. \end{cases}}$$

La médiane $\widehat{m}$ de l'échantillon est une estimation de la médiane m de la loi. L'intérêt de la médiane par rapport à la moyenne est qu'elle est moins sensible aux mesures aberrantes, mais elle est plus difficile à calculer et à manipuler.

Lorsque le nombre n de mesures augmente, les quantités $\widehat{P}_j$, $\widehat{M}$, $\widehat{V}$ (ou $\widehat{V}'$) et $\widehat{m}$ convergent respectivement vers P_j, M, V et m (*loi empirique des grands nombres*).

En particulier, d'après le théorème de la limite centrale, quand n est grand, la distribution des valeurs de $\widehat{M}$ tend vers une loi normale de paramètres $\mu = M$ et $\sigma^2 = V/n \approx \widehat{V}'/n$, et ce quelle que soit la loi suivie par X (il suffit que les mesures suivent la même loi et qu'elles soient indépendantes). L'écart entre la valeur mesurée, $\widehat{M}$, et la valeur recherchée, M (c.-à-d. l'*erreur* sur M) décroît donc comme $1/\sqrt{n}$. En définissant l'*incertitude statistique* Δ_M sur M par

$$\boxed{P(M \in [\widehat{M} - \Delta_M, \widehat{M} + \Delta_M]) = 95{,}4\,\%},$$

on a

$$\boxed{\Delta_M \approx 2\sqrt{\widehat{V}'/n}}.$$