1 Distributions binomiale, poissonienne et gaussienne

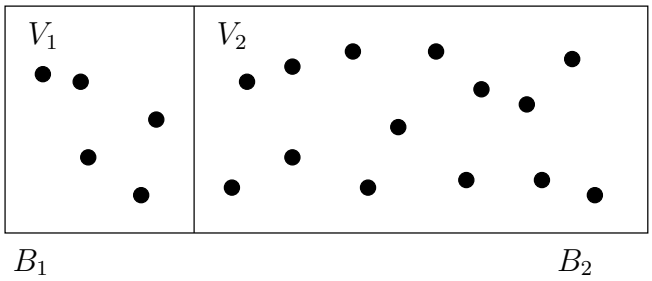
1.1 Rappels théoriques

Voir aussi le cours et les TDs d'OM4 pour de bases théoriques sur la distribution binomiale. Ci-dessous il y a des petits rappels (définition de la distribution binomiale, exemples physiques, valeurs théoriques pour moyenne et variance et aussi pour certaines limites qui correspondent aux distributions poissonienne et gaussienne).

Ceux qui connaissent bien le cours d'OM4 peuvent directement passer aux exercices numériques qui se trouvent sur la page 4 de ce document.

1. Distribution binomiale

On considère une variable aléatoire discrète ξ avec deux valeurs $\xi = 0$ ou $\xi = 1$ possibles et avec les probabilités $P(\xi = 1) = p$ et $P(\xi = 0) = 1 - p$ où $p \in [0, 1]$ est la probabilité individuelle d'obtenir la valeur de 1 (et $1 - p$ est la probabilité d'obtenir 0). Comme exemple on peut s'imaginer une pièce truquée (si $p \neq \frac{1}{2}$) où "pile" correspond à la valeur de $\xi = 1$ et "face" à la valeur de $\xi = 0$. Un autre exemple est un réservoir d'un grand nombre de boules rouges ou noires et la valeur de p est le rapport du nombre de boules rouges et du nombre total de boules dans le réservoir. Dans ce cas la variable aléatoire correspond au tirage d'une boule au hasard du reservoir où les résultats "boule rouge" (probilité p) ou "boule noire" (probabilité $1 - p$) correpondent soit à "$\xi = 1$" soit à "$\xi = 0$" respectivement.



Encore un autre exemple similaire est un ensemble de molécules d'un gaz reparties dans deux boîtes B_1 et B_2 de volumes V_1 et V_2. Dans ce cas les valeurs $\xi = 1$ et $\xi = 0$ de la variable aléatoire correpondent à la situation qu'**une molécule spécifique** se trouve soit dans B_1 (si $\xi = 1$) soit dans B_2 (si $\xi = 0$) où $p = V_1/(V_1 + V_2)$ est la probabilité que cette molécule spécifique se trouve dans la boîte B_1 (et $1 - p$ est la probabilité d'être dans la boîte B_2).

Soit N un entier positif et $\xi_1, \ldots, \xi_N$ un jeu de N variables aléatoires indépendantes et avec les même valeurs et probabilités que ξ (donc $\forall j = 1, \ldots, N \quad : \quad P(\xi_j = 1) = p$ et $P(\xi_j = 0) = 1 - p$). Ces variables aléatoires représentent N tirages aléatoires indépendantes des deux valeurs 0 et 1. Soit ζ la variable aléatoire définie par la somme $\zeta \equiv \xi_1 + \ldots + \xi_N$ ayant des valeurs dans $\{0, 1, \ldots, N\}$. Cette somme correspond au nombre de fois où on a obtenu la valeur de $\xi_j = 1$. Par exemple on lance N fois la pièce truquée et la variable aléatoire ζ fournit le nombre n des fois qu'on a obtenu "pile" ou $n \in \{0, 1, \ldots, N\}$.

Dans ce cas on peut montrer (voir cours et TD d'OM4) que pour tout $n \in \{0, 1, \ldots, N\}$ la

probabilité que $\zeta = n$ est donneée par la *distribution binomiale* :

$$P(n = \zeta) = \binom{N}{n}\ p^n\, q^{N-n}, \quad p + q = 1 \tag{1}$$

avec le coefficient binomial défini par

$$\binom{N}{n} = \frac{N!}{n!\,(N-n)!} = \frac{N(N-1)\cdot\ldots\cdot(N-n+1)}{1\cdot 2\cdot\ldots\cdot n}. \tag{2}$$

Une méthode facile de calculer (dans un code numérique) tous les coefficients binomiaux (pour N donné et tout $n = 0, 1, \ldots, N$) est la relation de récurrence :

$$\binom{N}{n} = \binom{N}{n-1}\frac{N-n+1}{n} \quad \text{avec} \quad \binom{N}{0} = 1\ . \tag{3}$$

Cette récurrence est efficace et évite le problème de factoriels trop grands si N est assez grand. Pour cette raison on conseille d'éviter la formule des factoriels dans des codes numériques et c'est mieux d'utiliser cette récurrence.

Pour comprendre la distribution binomiale on mentionne brièvement (voir cours/TD d'OM4 pour plus de détails) que le coefficient binomiale représente le nombre de sous-ensembles différents de n éléments parmi N éléments au total et le facteur $p^n q^{N-n}$ est la probabilité que pour **un sous-ensemble spécifique** toutes les valeurs de ces n éléments sont 1 (et 0 pour les tous autres $N - n$ éléments).

En utilisant le binôme de Newton, on vérifie facilement que la normalisation de (1) est bonne :

$$\sum_{n=0}^{N} P(n = \xi) = (p + q)^N = 1. \tag{4}$$

Pour évaluer la moyenne $\equiv \langle\zeta\rangle$ on remplace $n\,p^n = p\frac{\partial}{\partial p}p^n$ (et on traite temporairement q =const. comme une variable différente et indépendante de p), ce qui donne :

$$\langle\zeta\rangle = \sum_{n=0}^{N} n\ P(n = \xi) = p\frac{\partial}{\partial p}\sum_{n=0}^{N}\binom{N}{n}\ p^n\, q^{N-n} = p\frac{\partial}{\partial p}(p + q)^N = N\,p. \tag{5}$$

Ici, il faut remplacer $q = 1-p$ **après** avoir effectué la dérivée. D'une façon similaire, on trouve

$$\left\langle\zeta^2\right\rangle = p\frac{\partial}{\partial p}p\frac{\partial}{\partial p}(p + q)^N = p\frac{\partial}{\partial p}\left(pN(p + q)^{N-1}\right) = pN + N(N-1)p^2 \tag{6}$$

et pour la variance, on obtient

$$\mathrm{Var}(\zeta) \equiv \left\langle\zeta^2\right\rangle - \langle\zeta\rangle^2 = N\,p\,(1-p). \tag{7}$$

2. Distribution poissonienne

La distribution poissonienne est obtenue à partir de la distribution binomiale en prenant la limite $p \to 0$, $N \to \infty$ de façon que $a = N\,p$ reste constant. Dans ce cas on trouve pour $n \ll N$ (ou plus précisement : n =const. et $N \to \infty$)

$$\begin{aligned} P(\zeta = n) &= \lim_{N\to\infty} \left[\frac{1}{n!}\, N(N-1)\cdot \ldots \cdot (N-n+1) \left(\frac{a}{N}\right)^n \left(1-\frac{a}{N}\right)^{N-n} \right] \\ &= \frac{a^n}{n!} \underbrace{\lim_{N\to\infty} \left(\prod_{j=0}^{n-1} \frac{N-j}{N} \right)}_{=1} \underbrace{\lim_{N\to\infty} \left(1-\frac{a}{N}\right)^{-n}}_{=1} \underbrace{\lim_{N\to\infty} \left(1-\frac{a}{N}\right)^{N}}_{=e^{-a}} \quad \Rightarrow \\ P(\zeta = n) &= \frac{a^n}{n!}\, e^{-a} \quad \text{car} \quad \lim_{N\to\infty} \left(1+\frac{x}{N}\right)^N = e^x \quad (\text{avec } x = -a) \quad . \end{aligned} \tag{8}$$

(Voir par exemple cours OM4 et probablement aussi OM1 pour la dernière limite classique liée à l'exponentiel.) Dans un code numérique on peut utiliser la récurrence :

$$P(n) = \frac{a}{n}\, P(n-1) \quad \text{avec} \quad P(0) = e^{-a} \tag{9}$$

qui ne nécessite qu'un seul calcul de e^{-a} et pas de calculs explicites de puissances ni de factoriels (seulement de multiplications, divisions etc.).

3. Distribution gaussienne

On peut également considérer une autre limite définie par p =constante et $N \to \infty$. Dans ce cas, un calcul plus long montre que la distribution binomiale est proche d'une distribution gaussienne :

$$P(\zeta = n) \simeq \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(n-\langle\zeta\rangle)^2/\,(2\sigma^2)} \tag{10}$$

avec

$$\langle\zeta\rangle = N\,p \quad , \quad \sigma^2 = \mathrm{Var}(\zeta) = N\,p\,(1-p) \tag{11}$$

et en supposant que $|n - \langle\zeta\rangle\,|/\sigma$ =const. dans la limite $N \to \infty$, c'est-à-dire on ne considère que le domaine autour de la moyenne $\langle\zeta\rangle = Np$ avec un nombre fini (mais potentiellement assez grand) d'écart-types $\sigma = \sqrt{Np(1-p)}$ pour la différence $|n - \langle\zeta\rangle\,|$.

1.2 Exercices numériques

Dans les exercices ci-dessous choisir des valeurs raisonables pour $p \in [0,1]$, N et R (par exemple $p = 0.3$, $N = 100$, $R = 100$ et plus tard on peut augmenter $R = 500$, 1000 etc.).

1. Distribution binomiale

a) En utilisant la fonction `rand` écrire une fonction python qui simule les variables aléatoires ξ_j, c'est-à-dire qui fournit avec probabilité p la valeur 1 et avec probabilité $1-p$ la valeur 0 ou p est un paramètre de cette fonction. Par exemple on peut appeler cette fonction `tirage1(p)`. Ensuite écrire une autre fonction avec deux paramètres p et N (par exemple `tirageN(p,N)`) qui simule la variable aléatoire $\zeta = \xi_1 + \ldots + \xi_N$ en calculant la somme de N appels de `tirage1(p)`. Comme premier test appeler cette fonction 5 fois et afficher les résultats.

b) Écrire un code python pour appeler R fois la fonction `tirageN` et stocker les valeurs obtenues dans un vecteur (ou une liste python) appelé(e) `N1`.

c) Tracer l'histogramme $F(n)$ du nombre de séries qui contiennent n fois le nombre 1 (en utilisant les valeurs dans le vecteur `N1`). Pour cela on peut utiliser la fonction `plt.hist` de Python (utiliser les options "`density=True`" pour assurer la bonne normalisation et `bins=...` pour choisir le nombre de boites de l'histogramme ; conseil : utiliser `bins=max(N1)-min(N1)` pour avoir une largeur de boîte égale à 1).

d) Pour comprendre ce que fait la fonction `plt.hist` calculer dans un code (ou une fonction à part) soi-même la distribution du nombre de 1 et comparer son graphe avec l'histogramme. Explicitement : calculer un tableau/liste `compter` ou `compter[n]` contiendra pour tout $n \in \{0, 1, \ldots, N\}$ le nombre de fois que la valeur n apparaît dans le tableau/liste `N1`.

e) Remplir un autre vecteur avec les valeurs théoriques (1) pour tout $n \in \{0, 1, \ldots, N\}$. Pour cela on conseille d'utiliser la récurrence (3). Comparer (graphiquement) le résultat avec l'histogramme obtenu avant pour différentes valeurs de R (faire attention à une normalisation cohérente entre les deux, soit normalisé par 1 soit par R mais cela pour les deux).

f) Calculer numériquement la moyenne et la variance de la distribution (en utilisant les données du tableau `N1` avec des sommes etc.) et les comparer aux valeurs théoriques.

2. Distributions poissonienne et gaussienne

Le but est maintenant de comparer (graphiquement) les distributions poissonienne et gaussienne théoriques à la distribution binomiale pour différentes valeurs des paramètres p et N.

a) Comparer la distribution binomiale (1) pour $p \ll 1$ et N grand avec la distribution poissonienne (8) de paramètre $a = N\,p$.

b) Comparer la distribution binomiale (1) (par exemple pour $p = 0.3$) avec la distribution gaussienne (10) (avec les valeurs théoriques pour $\langle\zeta\rangle$ et $\mathrm{Var}(\zeta)$).

Indications : Remplir pour les deux cas des vecteurs/listes avec les valeurs théoriques. Pour les plots on pourra utiliser la commande python : `plt.xlim([nmin,nmax])` (avant `plt.plot(...)`) pour définir des valeurs minimale et maximale raisonables pour les plots (pour éviter/réduire les zones où $P(n) \approx 0$).