université PARIS-SACLAY | FACULTÉ DES SCIENCES D'ORSAY

Exercices

Guillermo Durand

Exercice 1 (Loi hypergéométrique)
Soit $N \geq 1$ et $1 \leq n \leq N$. On tire n boules simultanément dans une urne qui contient N boules numérotées de 1 à N.

1. Donner un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ qui modélise l'expérience.
2. Soit $k, \ell \in \{1, \ldots, N\}$. On note $E_{k,\ell}$ l'événement "on a tiré la boule k et la boule ℓ" et, pour simplifier les notations, on pose $E_k = E_{k,k}$. Déterminer $\mathbb{P}(E_k)$ et $\mathbb{P}(E_{k,\ell})$ si $k \neq \ell$ (et $n \geq 2$).
3. E_k et E_ℓ sont-ils indépendants pour $k \neq \ell$?
4. Soit $G \in \{1, \ldots, N\}$. On note X la variable aléatoire qui donne le nombre d'éléments de $\{1, \ldots, G\}$ qui ont été tirés. Donner la loi de X. On note cette loi $\mathcal{H}(N, G, n)$ et on l'appelle la loi hypergéométrique de paramètres N, G et n.
5. (a) Écrire X en fonction des événements E_k (à l'aide de fonctions indicatrices).
   (b) En déduire $\mathbb{E}[X]$.
   (c) En déduire aussi $\mathbb{V}(X)$.
6. Montrer que $\mathcal{H}(N, G, n) = \mathcal{H}(N, n, G)$.
7. On a donc défini X comme le nombre d'éléments de $\{1, \ldots, G\}$ tirés en effectuant n tirages simultanés sans remplacement dans l'urne. On définit Y comme le nombre d'éléments de $\{1, \ldots, G\}$ tirés en effectuant n tirages consécutifs avec remplacement dans l'urne. Donner la loi de Y, et comparer entre elles $\mathbb{E}[X]$ et $\mathbb{E}[Y]$, ainsi que $\mathbb{V}(X)$ et $\mathbb{V}(Y)$.
8. Soit $X_1, \ldots, X_{n_X}$ un n_X-échantillon de loi commune $\mathcal{B}(p_X)$, $p_X \in [0,1]$, et $Y_1, \ldots, Y_{n_Y}$ un n_Y-échantillon de loi commune $\mathcal{B}(p_Y)$, $p_Y \in [0,1]$. Les deux échantillons sont indépendants. On pose $N_{1X} = \sum_{i=1}^{n_X} X_i$, $N_{1Y} = \sum_{i=1}^{n_Y} Y_i$ les nombres respectifs de succès dans chaque échantillon, et $N_{1.} = N_{1X} + N_{1Y}$ le nombre total de succès.
   (a) Donner les lois de N_{1X} et N_{1Y}.
   (b) Dans cette question, on suppose que $p_X = p_Y$. Soit $u \in \{0, \ldots, n_X + n_Y\}$. Donner la loi de N_{1X} conditionnellement à l'événement "$N_{1.} = u$". C'est-à-dire, donner $\mathbb{P}(N_{1X} = k | N_{1.} = u)$ pour tout entier $k \in \mathbb{N}$.
   (c) En déduire un test ϕ_u de niveau $\alpha \in [0,1]$ de $H_0 : p_X = p_Y$ contre $H_1 : p_X \neq p_Y$ conditionnellement à l'événement "$N_{1.} = u$". C'est-à-dire, tel que $\mathbb{P}_{H_0}(\phi_u = 1 | N_{1.} = u) \leq \alpha$.
   (d) En déduire un test ϕ de niveau $\alpha \in [0,1]$ de $H_0 : p_X = p_Y$ contre $H_1 : p_X \neq p_Y$ inconditionnel. C'est-à-dire, tel que $\mathbb{P}_{H_0}(\phi = 1) \leq \alpha$.

Solution:

1. $\Omega = \{\text{parties à } n \text{ éléments de } [\![1, N]\!]\}$, $\mathcal{A} = \mathcal{P}(\Omega)$, $\mathbb{P} : A \mapsto \frac{|A|}{|\Omega|} = \frac{|A|}{\binom{N}{n}}$ la mesure uniforme.

2.

$$\begin{aligned}\mathbb{P}(E_k) &= \mathbb{P}\left(\{\text{parties à } n \text{ éléments de } [\![1, N]\!] \text{ contenant } k\}\right)\\ &= \frac{|\{\text{parties à } n-1 \text{ éléments de } [\![1, N]\!] \setminus \{k\}\}|}{\begin{pmatrix} N \\ n \end{pmatrix}}\\ &= \frac{\begin{pmatrix} N-1 \\ n-1 \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}\\ &= \frac{(N-1)!n!(N-n)!}{(n-1)!(N-n)!N!}\\ &= \frac{n}{N}.\end{aligned}$$

$$\begin{aligned}\mathbb{P}(E_{k,\ell}) &= \mathbb{P}\left(\{\text{parties à } n \text{ éléments de } [\![1, N]\!] \text{ contenant } k \text{ et } \ell\}\right)\\ &= \frac{|\{\text{parties à } n-2 \text{ éléments de } [\![1, N]\!] \setminus \{k, \ell\}\}|}{\begin{pmatrix} N \\ n \end{pmatrix}}\\ &= \frac{\begin{pmatrix} N-2 \\ n-2 \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}\\ &= \frac{(N-2)!n!(N-n)!}{(n-2)!(N-n)!N!}\\ &= \frac{n(n-1)}{N(N-1)}.\end{aligned}$$

3. On a $\mathbb{P}(E_k \cap E_\ell) = \mathbb{P}(E_{k,\ell}) = \frac{n(n-1)}{N(N-1)}$ tandis que $\mathbb{P}(E_k)\,\mathbb{P}(E_\ell) = \frac{n^2}{N^2}$. On a égalité si et seulement si $\frac{n-1}{N-1} = \frac{n}{N}$ soit si et seulement si $\frac{n-1}{n} = \frac{N-1}{N}$ soit si et seulement si $1 - \frac{1}{n} = 1 - \frac{1}{N}$ soit si et seulement si $n = N$. En conclusion, E_k et E_ℓ sont indépendants si et seulement si $n = N$ (et on remarque que dans ce cas ils sont tous les deux presque sûrs).

4. X est donc la fonction qui à une partie A de Ω associe $|A \cap [\![1, G]\!]|$. Vu que $|A| = n$, on a nécessairement $|A \cap [\![1, G]\!]| \leq G \wedge n$. De plus $|A \cap [\![1, G]\!]| = G - |A^{\mathsf{c}} \cap [\![1, G]\!]| \geq G - |A^{\mathsf{c}}| = G + n - N$ donc $X(\Omega) \subset [\![0 \vee (G + n - N), G \wedge n]\!]$. Réciproquement, pour $k \in [\![0 \vee (G + n - N), G \wedge n]\!]$, on a $X([\![1, k]\!] \cup [\![G + 1, G + n - k]\!]) = k$ (en remarquant que $[\![1, 0]\!] = [\![G+1, G]\!] = \varnothing$ pour les éventuels cas $k = 0$ et $k = n$). Donc $X(\Omega) = [\![0 \vee (G + n - N), G \wedge n]\!]$.

Soit $k \in [\![0 \vee (G+n-N), G \wedge n]\!]$.

$$\begin{aligned}
\mathbb{P}(X=k) &= \frac{|\{\text{parties à } k \text{ éléments de } [\![1,G]\!] \text{ et à } n-k \text{ éléments de } [\![G+1,N]\!]\}\ |}{\binom{N}{n}} \\
&= \frac{|\{\text{parties à } k \text{ éléments de } [\![1,G]\!]\}||\{\text{parties à } n-k \text{ éléments de } [\![G+1,N]\!]\}\ |}{\binom{N}{n}} \\
&= \frac{\binom{G}{k}\binom{N-G}{n-k}}{\binom{N}{n}}.
\end{aligned}$$

5. (a) $X = \sum_{k=1}^{G} \mathbb{1}_{E_k}$

(b) Par linéarité de l'espérance,

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}\left[\sum_{k=1}^{G} \mathbb{1}_{E_k}\right] \\
&= \sum_{k=1}^{G} \mathbb{E}[\mathbb{1}_{E_k}] \\
&= \sum_{k=1}^{G} \mathbb{P}(E_k) \\
&= G\frac{n}{N}.
\end{aligned}$$

(c) $\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \frac{G^2n^2}{N^2}$

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \sum_{1\leq k,\ell\leq G} \mathbb{E}[\mathbb{1}_{E_k}\mathbb{1}_{E_\ell}] \\
&= \sum_{1\leq k,\ell\leq G} \mathbb{E}[\mathbb{1}_{E_k\cap E_\ell}] \\
&= \sum_{1\leq k,\ell\leq G} \mathbb{E}\left[\mathbb{1}_{E_{k,\ell}}\right] \\
&= \sum_{\substack{1\leq k,\ell\leq G\\ k\neq\ell}} \mathbb{E}\left[\mathbb{1}_{E_{k,\ell}}\right] + \sum_{k=1}^{G} \mathbb{E}[\mathbb{1}_{E_k}] \\
&= \sum_{\substack{1\leq k,\ell\leq G\\ k\neq\ell}} \mathbb{P}(E_{k,\ell}) + \frac{Gn}{N} \\
&= G(G-1)\frac{n(n-1)}{N(N-1)} + \frac{Gn}{N},
\end{aligned}$$

donc

$$\begin{aligned}
\mathbb{V}(X) &= \frac{G(G-1)n(n-1)}{N(N-1)} + \frac{Gn}{N} - \frac{G^2n^2}{N^2} \\
&= \frac{Gn}{N}\left(\frac{(G-1)(n-1)}{N-1} + 1 - \frac{Gn}{N}\right) \\
&= \frac{Gn}{N}\frac{N(G-1)(n-1) + N(N-1) - Gn(N-1)}{N(N-1)} \\
&= \frac{Gn}{N}\frac{NGn - nN - GN + N + N^2 - N - GnN + Gn}{N(N-1)} \\
&= \frac{Gn}{N}\frac{-nN - GN + N^2 + Gn}{N(N-1)} \\
&= \frac{Gn}{N}\frac{G(n-N) + N(N-n)}{N(N-1)} \\
&= \frac{Gn}{N}\frac{(N-G)(N-n)}{N(N-1)} \\
&= \frac{Gn}{N}\left(1 - \frac{G}{N}\right)\frac{N-n}{N-1}.
\end{aligned}$$

6. Soit $\tilde{X}$ de loi $\mathcal{H}(N, n, G)$. D'après la question 3, le support de $\tilde{X}$ est le même que celui de X soit $[\![0 \vee (G+n-N), G \wedge n]\!]$, vu que les rôles de G et n sont symétriques dedans. De plus, toujours d'après cette question, pour tout $k \in [\![0 \vee (G+n-N), G \wedge n]\!]$, on a $\mathbb{P}\left(\tilde{X} = k\right) = \frac{\binom{n}{k}\binom{N-n}{G-k}}{\binom{N}{G}}$. Or,

$$\begin{aligned}
\frac{\binom{n}{k}\binom{N-n}{G-k}}{\binom{N}{G}} &= \frac{n!(N-n)!G!(N-G)!}{k!(n-k)!(G-k)!(N-n-G+k)!N!} \\
&= \frac{G}{k!(G-k)!}\frac{(N-n)!}{(G-k)!(N-n-G+k)!}\frac{n!(N-n)!}{N!} \\
&= \frac{\dbinom{G}{k}\dbinom{N-G}{n-k}}{\dbinom{N}{n}} \\
&= \mathbb{P}(X = k),
\end{aligned}$$

donc X et $\tilde{X}$ ont bien la même loi : $\mathcal{H}(N, G, n) = \mathcal{H}(N, n, G)$.

7. $Y \sim \mathcal{B}\left(n, \frac{G}{N}\right)$, donc $\mathbb{E}[Y] = n\frac{G}{N} = \mathbb{E}[X]$, mais $\mathbb{V}(Y) = n\frac{G}{N}\left(1 - \frac{G}{N}\right) \geq \mathbb{V}(X)$.

8. (a) Par indépendance, $N_{1X} \sim \mathcal{B}(n_X, p_X)$ et $N_{1Y} \sim \mathcal{B}(n_Y, p_Y)$.

(b) Soit $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{P}\left(N_{1X}=k|N_{1\cdot}=u\right) &= \frac{\mathbb{P}\left(\{N_{1X}=k\}\cap\{N_{1\cdot}=u\}\right)}{\mathbb{P}\left(N_{1\cdot}=u\right)} \\ &= \frac{\mathbb{P}\left(\{N_{1X}=k\}\cap\{N_{1Y}=u-k\}\right)}{\mathbb{P}\left(N_{1X}+N_{1X}=u\right)} \\ &= \frac{\mathbb{P}\left(N_{1X}=k\right)\mathbb{P}\left(N_{1Y}=u-k\right)}{\mathbb{P}\left(N_{1X}+N_{1X}=u\right)}.\end{aligned}$$

Le numérateur est nul sauf si $0 \leq k \leq n_X$ et $0 \leq u-k \leq n_Y$ soit $0 \leq k \leq n_X$ et $u-n_Y \leq k \leq u$ soit $k \in [\![0 \vee (u-n_Y), n_X \wedge u]\!]$.

Pour le dénominateur, on note que, par indépendance et vu que $p_X = p_Y$, $N_{1X} + N_{1X} \sim \mathcal{B}(n_X+n_Y, p_X)$.

On a donc $\mathbb{P}\left(N_{1X}=k|N_{1\cdot}=u\right) = 0$ pour $k \notin [\![0 \vee (u-n_Y), n_X \wedge u]\!]$ et, pour $k \in [\![0 \vee (u-n_Y), n_X \wedge u]\!]$,

$$\begin{aligned}\mathbb{P}\left(N_{1X}=k|N_{1\cdot}=u\right) &= \frac{\mathbb{P}\left(N_{1X}=k\right)\mathbb{P}\left(N_{1Y}=u-k\right)}{\mathbb{P}\left(N_{1X}+N_{1X}=u\right)} \\ &= \frac{\binom{n_X}{k}p_X^k(1-p_X)^{n_X-k}\binom{n_Y}{u-k}p_X^{u-k}(1-p_X)^{n_Y-u+k}}{\binom{n_X+n_Y}{u}p_X^u(1-p_X)^{n_X+n_Y-u}} \\ &= \frac{\binom{n_X}{k}\binom{n_Y}{u-k}}{\binom{n_X+n_Y}{u}} = \frac{\binom{n_X}{k}\binom{n_X+n_Y-n_X}{u-k}}{\binom{n_X+n_Y}{u}}.\end{aligned}$$

On reconnaît la loi $\mathcal{H}(n_X+n_Y, n_X, u)$.

(c) Conditionnellement à $N_{1\cdot} = u$ et si H_0 est vraie, d'après la question précédente, N_{1X} suit la loi $\mathcal{H}(n_X+n_Y, n_X, u)$ d'espérance $\frac{un_X}{n_X+n_Y}$ et de variance $\frac{un_Xn_Y(n_X+n_Y-u)}{(n_X+n_Y)^2(n_X+n_Y-1)}$.

On peut choisir de rejeter si $\left|N_{1X} - \frac{un_X}{n_X+n_Y}\right| \geq \varepsilon_u$, c'est-à-dire que $\phi_u = \mathbb{1}_{\left\{\left|N_{1X}-\frac{un_X}{n_X+n_Y}\right|\geq\varepsilon_u\right\}}$, avec ε_u à calibrer pour assurer que $\mathbb{P}\left(\phi_u=1|N_{1\cdot}=u\right) \leq \alpha$ sous H_0. Pour cela on peut utiliser par exemple l'inégalité de Bienaymé-Tchebychev : sous H_0, on a

$$\begin{aligned}\mathbb{P}\left(\phi_u=1|N_{1\cdot}=u\right) &= \mathbb{P}\left(\left|N_{1X}-\frac{un_X}{n_X+n_Y}\right| \geq \varepsilon_u|N_{1\cdot}=u\right) \\ &\leq \frac{1}{\varepsilon_u^2}\frac{un_Xn_Y(n_X+n_Y-u)}{(n_X+n_Y)^2(n_X+n_Y-1)}.\end{aligned}$$

Il suffit donc de choisir $\varepsilon_u = \sqrt{\frac{un_Xn_Y(n_X+n_Y-u)}{\alpha(n_X+n_Y)^2(n_X+n_Y-1)}}$.

Alternativement, on peut choisir de rejeter si $N_{1X} < c_{1,u}$ ou si $N_{1X} > c_{2,u}$ avec $c_{1,u}$ et $c_{2,u}$ choisis tels que $\mathbb{P}_{H_0}(N_{1X} < c_{1,u} | N_{1\cdot} = u) \leq \frac{\alpha}{2}$ et $\mathbb{P}_{H_0}(N_{1X} > c_{2,u} | N_{1\cdot} = u) \leq \frac{\alpha}{2}$. On note F_u la fonction de répartition de la loi $\mathcal{H}(n_X + n_Y, u, n_X)$ et F_u^- sa fonction limite à gauche. Les deux contraintes sur $c_{1,u}$ et $c_{2,u}$ se réécrivent donc $F_u^-(c_{1,u}) \leq \frac{\alpha}{2}$ et $F_u(c_{2,u}) \geq 1 - \frac{\alpha}{2}$. Pour maximiser la puissance, on est donc libre de choisir $c_{1,u} = \sup\{c : F_u^-(c) \leq \frac{\alpha}{2}\}$ et $c_{2,u} = \inf\{c : F_u(c) \geq 1 - \frac{\alpha}{2}\}$. $c_{2,u}$ s'appelle le quantile de la loi $\mathcal{H}(n_X + n_Y, u, n_X)$ à l'ordre $1 - \frac{\alpha}{2}$. On peut démontrer que ce sup et cet inf existent bien, et sont respectivement un max et un min, grâce aux propriétés caractéristiques de la fonction de répartition (croissante, càdlàg, de limites 0 en $-\infty$ et 1 en ∞).

(d) On pose $\phi = \phi_{N_{1\cdot}}$. En effet,

$$
\begin{aligned}
\mathbb{P}_{H_0}(\phi = 1) &= \mathbb{P}_{H_0}(\phi_{N_{1\cdot}} = 1) \\
&= \sum_{u=0}^{n_X+n_Y} \mathbb{P}_{H_0}(\phi_{N_{1\cdot}} = 1 \cap N_{1\cdot} = u) \\
&= \sum_{u=0}^{n_X+n_Y} \mathbb{P}_{H_0}(\phi_u = 1 \cap N_{1\cdot} = u) \\
&= \sum_{u=0}^{n_X+n_Y} \mathbb{P}_{H_0}(\phi_u = 1 | N_{1\cdot} = u)\, \mathbb{P}_{H_0}(N_{1\cdot} = u) \\
&\leq \alpha \sum_{u=0}^{n_X+n_Y} \mathbb{P}_{H_0}(N_{1\cdot} = u) \\
&= \alpha.
\end{aligned}
$$