

# Les urnes de Polya

Pierre Grihon(\*)

L'expérience des urnes de Polya est intéressante pour au moins quatre raisons :

- Elle est très simple à décrire,
- On peut en étudier les premiers résultats avec très peu de connaissances en probabilité,
- On peut la simuler facilement et conjecturer des résultats rapidement,
- Elle offre des développements et des applications étonnants.

On peut commencer à l'étudier en seconde, prolonger en première et terminale.

Les notations adoptées sont celles de terminale, mais on peut les adapter pour les niveaux précédents.

Ce sujet a été proposé dans le cadre d'un atelier Math en Jeans à des élèves de seconde et de première. Une partie de cet article est issue de leur travail.



**George (György) Pólya** est né à Budapest (Hongrie) le 13 décembre 1887 et mort à Palo Alto (États-Unis) le 7 septembre 1985. En construisant ce modèle d'urne, sa motivation était, semble-t-il, d'étudier les phénomènes de contagion lors d'une épidémie.

## 1. L'expérience de l'urne de Polya

On considère une urne contenant au départ  $a$  boules blanches et  $b$  boules rouges. On tire une boule de l'urne et on la remet avec une autre boule de la même couleur ; on a donc maintenant  $a + b + 1$  boules. On recommence un certain nombre de fois l'opération qui consiste à tirer une boule de l'urne et à la remettre avec une autre boule de la même couleur. On cherche à déterminer l'évolution de la proportion de boules blanches dans l'urne.

On note  $X_n$  le nombre de boules blanches dans l'urne au bout de  $n$  tirages.

On a donc  $X_0 = a$ .

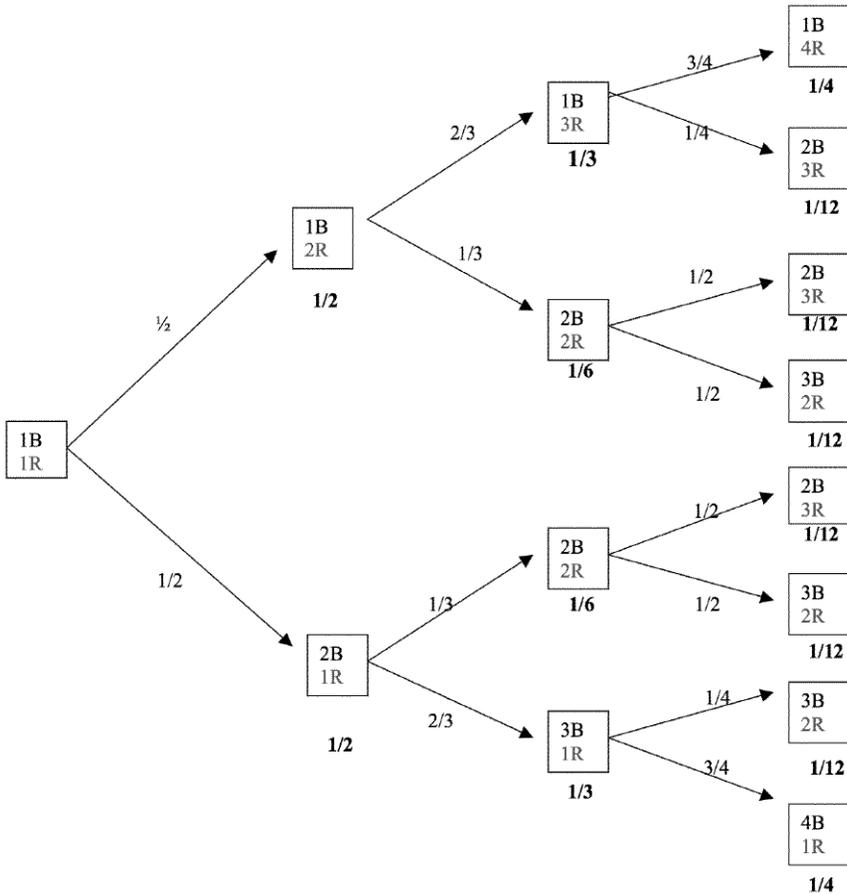
**a) Commençons par le cas  $a = b = 1$ .**

On a immédiatement la loi de  $X_1$  :  $P(X_1 = 1) = P(X_1 = 2) = \frac{1}{2}$ .

À l'aide d'un arbre, on peut trouver les lois de  $X_2$  et  $X_3$ .

---

(\*) Lycée Montaigne-BORDEAUX.



On voit donc que  $X_2$  et  $X_3$  suivent des lois uniformes respectivement sur  $\{1, 2, 3\}$  et  $\{1, 2, 3, 4\}$ .

On peut donc conjecturer puis démontrer par récurrence que  $X_n$  suit une loi uniforme sur  $\{1, 2, \dots, n + 1\}$ .

Si on suppose que

$$\forall k \in \{1, \dots, n + 1\}, P(X_n = k) = \frac{1}{n + 1},$$

par la formule des probabilités totales avec le système complet d'événements  $(X_n = k), k \in \{1, \dots, n + 1\}$ , on a :

$$\forall k \in \{1, \dots, n + 2\}, P(X_{n+1} = k) = \sum_{i=1}^{n+1} P_{[X_n=i]}(X_{n+1} = k)P(X_n = i).$$

Dans cette somme, seuls deux termes ne sont pas nuls, d'où :

$$\begin{aligned}
 P(X_{n+1} = k) &= \frac{1}{n+1} \left( \underset{\substack{\uparrow \\ \text{on tire une rouge}}}{P_{[X_n=k]}(X_{n+1} = k)} + \underset{\substack{\uparrow \\ \text{on tire une blanche}}}{P_{[X_n=k-1]}(X_{n+1} = k)} \right) \\
 &= \frac{1}{n+1} \left( \frac{n+1-k}{n+2} + \frac{k}{n+2} \right) = \frac{1}{n+2}.
 \end{aligned}$$

**Cela signifie que après le  $n$ -ième tirage, toutes les configurations sont équiprobables.**

Donc, si on renouvelle cette expérience (faire  $n$  tirages successifs) plusieurs fois, on aura des résultats très différents, uniformément répartis sur  $\{1, 2, \dots, n+1\}$ .

Cela pourrait faire penser à une « fluctuation d'échantillonnage ».

Mais cela n'a rien à voir avec cette notion.

Si je lance 100 fois une pièce de monnaie équilibrée, je vais avoir une fréquence de

Piles proche de  $\frac{1}{2}$ , et ceci à chaque renouvellement de la série de lancers mais avec

une petite fluctuation autour de  $\frac{1}{2}$ .

Dans l'expérience de Polya, à chaque série de  $n$  tirages on aura des résultats différents mais qui ne fluctuent pas autour d'une valeur unique : en fait on échantillonne une loi uniforme sur  $\{1, \dots, n+1\}$ .

**Avant de se lancer dans une série de tirages, on peut se demander quelle est la probabilité de l'événement  $B_{n+1} = \ll \text{tirer une boule blanche au } (n+1)\text{-ième tirage} \gg$ .**

Avec la formule des probabilités totales, cela donne :

$$P(B_{n+1}) = \sum_{k=1}^{n+1} P_{[X_n=k]}(B_{n+1}) P(X_n = k) = \sum_{k=1}^{n+1} \frac{k}{n+2} \frac{1}{n+1} = \frac{1}{2}.$$

On a toujours *au départ* une chance sur deux de tirer une blanche au  $(n+1)$ -ième tirage (Polya parle de chance « *a priori* »)

Ce résultat peut ne pas surprendre car ayant égalité au départ, il y a symétrie des couleurs et donc égalité des probabilités :  $P(B_{n+1}) = P(R_{n+1})$  et évidemment

$$P(B_{n+1}) + P(R_{n+1}) = 1.$$

Mais cette notion de symétrie a ses limites : comme on l'a vu, malgré cette symétrie, la répartition des couleurs après le  $n$ -ième tirage n'est pas équilibrée.

Si on demande à quelqu'un n'ayant jamais étudié ce problème, l'intuition qu'il a de la proportion atteinte au bout d'un grand nombre de tirages, sa réponse est

fréquemment  $\frac{1}{2}$  : il lui semble naturel de penser qu'il n'y a aucune raison qu'une couleur l'emporte sur l'autre. Nous verrons plus loin ce qu'il en est...

**b) Généralisation**

On suppose maintenant que l'on a au départ  $a$  boules blanches et  $b$  boules rouges.

La loi de  $X_n$  est un peu plus compliquée et dépasse le niveau du lycée.

On peut la trouver par un raisonnement direct de dénombrement.

$X_n$  peut prendre les valeurs  $\{a, a+1, \dots, a+n\}$ .

En notant  $B_k = \ll \text{tirer une blanche au } k\text{-ième tirage} \gg$  et  $R_k = \ll \text{tirer une rouge au } k\text{-ième tirage} \gg$ , on a :

$$P(B_1 \cap \dots \cap B_k \cap R_{k+1} \cap \dots \cap R_n) = \frac{a}{a+b} \dots \frac{a+k-1}{a+b+k-1} \frac{b}{a+b+k} \dots \frac{b+n-k-1}{a+b+n-1}.$$

Si l'ordre d'apparition des couleurs change, le numérateur et le dénominateur sont

« globalement invariants », donc toujours les mêmes. Il y a  $\binom{n}{k}$  façons de faire les

$n$  tirages en tenant compte de l'ordre des couleurs, tous de même probabilité donc :

pour  $1 \leq k \leq n-1$ ,

$$P(X_n = a+k) = \binom{n}{k} \frac{a \dots (a+k-1) b \dots (b+n-k-1)}{(a+b) \dots (a+b+n-1)},$$

ce qui donne finalement pour  $0 \leq k \leq n$  :

$$P(X_n = a+k) = \binom{n}{k} \frac{ab}{a+b} \frac{\binom{a+b}{a}}{a+k} \frac{\binom{n}{k}}{\binom{a+b+n-1}{a+k}}$$

en remarquant que cette formule est vraie pour les cas extrêmes  $k=0$  (que des rouges) et  $k=n$  (que des blanches).

On peut démontrer que  $P(B_{n+1}) = \frac{a}{a+b}$  (à partir de la loi de  $X_n$ , mais c'est un peu technique).

**2. Simulations**

Je propose deux simulations :

- une première avec un algorithme adaptable sur n'importe quelle calculatrice,
- une deuxième avec un tableur permettant de visualiser l'évolution de la proportion.

On peut utiliser ces deux simulations de deux manières :

- pour étudier la composition de l'urne au bout d'un petit nombre de tirages (particulièrement en seconde),
- pour étudier la composition de l'urne au bout d'un grand nombre de tirages et voir s'il y a « convergence ».

C'est la deuxième qui est privilégiée ici.

**a) Programmation des tirages**

Je donne l'algorithme en français. Il a pour tâche de simuler une réalisation de  $n$  tirages et de retourner la suite  $(Y_1, \dots, Y_n)$  des proportions de blanches dans l'urne.

**Début**

*Entrer  $a$  et  $b$  ;*

*Entrer le nombre  $n$  de tirages ;*

*Pour  $k$  variant de 1 à  $n$  faire*

*Début*

*Si  $\text{aléa} \leq a/(a+b)$  alors  $a$  devient  $a+1$  sinon  $b$  devient  $b+1$  ; afficher  $a/(a+b)$  ;*

*Fin ;*

**Fin.**

Chaque fois que l'on exécute ce programme avec  $n = 500$  (par exemple), on effectue les 500 premiers tirages d'une suite infinie de tirages notée  $\omega$  (appelée réalisation) et on observe que la suite des proportions obtenues  $(Y_1(\omega), \dots, Y_n(\omega))$  se stabilise autour d'une valeur  $Y(\omega)$ .

Cela permet de conjecturer que la suite  $(Y_n(\omega))$  converge vers  $Y(\omega)$  et, comme 500 est grand, on peut penser que  $Y_{500}(\omega)$  est proche de  $Y(\omega)$ .

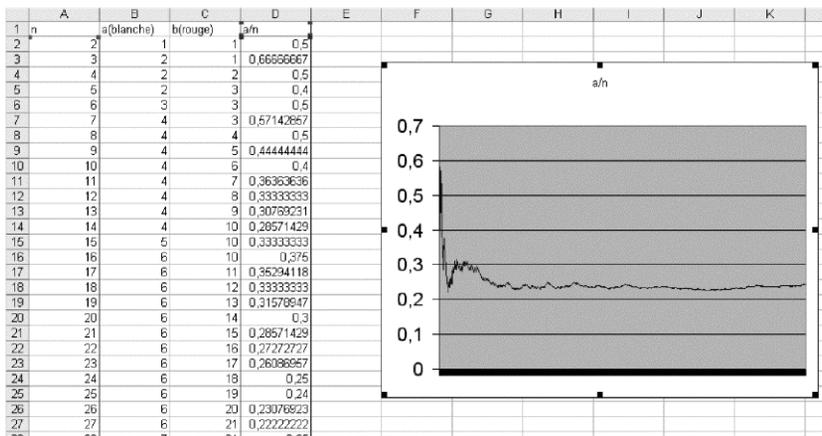
Si on exécute une deuxième fois ce programme, on effectue les 500 premiers tirages d'une autre réalisation  $\omega'$  ; on obtient une autre suite de proportions qui semble également converger vers une proportion limite  $Y(\omega')$  en général différente de la première  $Y(\omega)$ .

$Y(\omega)$  dépend donc de  $\omega$  et de ce fait  $Y$  est une variable aléatoire.

Si maintenant on exécute  $k$  fois dans un même programme cet algorithme pour  $n = 500$  (par exemple), on aura  $k$  réalisations  $\omega_1, \dots, \omega_k$  et on observera une stabilisation de la proportion sur un nombre différent à chaque fois. Avec les valeurs  $Y_{500}(\omega_1), \dots, Y_{500}(\omega_k)$  on a une base de données et on peut faire des statistiques permettant de voir comment se répartit la variable  $Y_{500}$ , ce qui donne une idée de la répartition de  $Y$ .

**b) Avec un tableur**

On peut très facilement simuler une suite de tirages et produire une courbe donnant l'évolution de la proportion de blanches.



fx =SI(ALEA()<D2;B2+1;B2)

	A	B	C	D
1	n	a(blanche)	b(rouge)	a/n
2		2	1	0,5
3		3	2	0,66666667
4		4	3	0,75

On peut réaliser une nouvelle série de tirages avec les mêmes données initiales en appuyant sur la touche F9 du clavier.

On observe, comme avec l’algorithme du a), que lors de chaque série de tirages la proportion se stabilise, mais vers un nombre différent à chaque fois là-aussi.

L’avantage ici est la visualisation en courbe mais on ne peut pas faire de statistiques.

En observant ces courbes, les élèves se sont demandés comment se « répartissaient » ces valeurs « limites » (le « bout de la courbe »).

Ils ont donc eu recours à la simulation du a) qui leur permet de refaire cette même expérience *k* fois et de stocker les résultats en vue de les analyser.

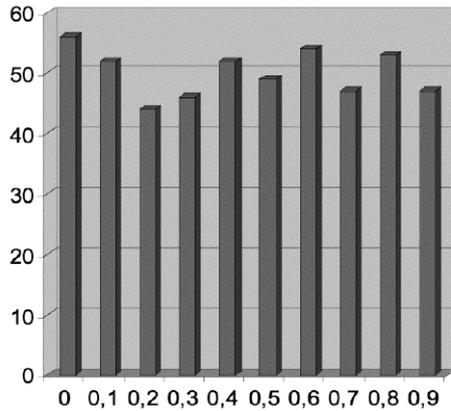
Ces questions de limite seront précisées un peu plus loin.

### c) Statistiques

Ils ont donc obtenu une base de données qu’ils ont représentée par un diagramme en bâtons.

En abscisse la proportion obtenue rangée en classe d’amplitude 0,1.

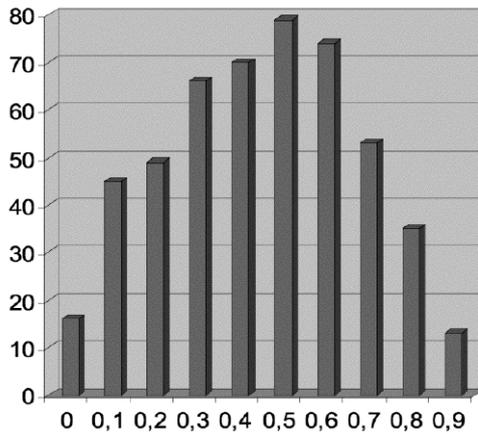
En ordonnée, le nombre de suites ayant donné une « limite » comprise entre 0 et 0,1, entre 0,1 et 0,2, etc.



Il semble évident que l'on a une répartition très uniforme des « limites ».

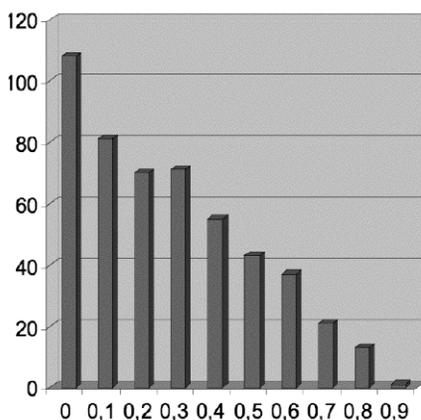
On peut refaire ce travail en changeant le nombre de boules au départ, par exemple en prenant 2 blanches et 2 rouges.

L'intuition pourrait laisser penser que cela ne change rien sur la répartition des « limites »...



En fait cela change tout !

Et avec 1 blanche et 2 rouges au départ :



### 3. Convergence

On a pu voir dans le paragraphe précédent que si l'on fait un très grand nombre de tirages successifs dans l'urne la proportion de boules blanches se stabilise autour d'une valeur  $p$  totalement imprévisible.

On peut donc penser qu'en un sens à préciser la suite des proportions  $\frac{X_n}{n+2}$  converge.

Posons  $Y_n = \frac{X_n}{n+2}$ .

Il s'agit d'une suite de variables aléatoires et il y a plusieurs types de convergences d'une telle suite.

Les simulations du paragraphe 2) mettent en évidence une convergence vers une variable aléatoire  $Y$ , convergence que l'on définit comme presque sûre.

Précisément, cela signifie que l'ensemble des  $\omega$  pour les quels  $Y_n(\omega) \rightarrow Y(\omega)$  est de probabilité 1.

$\omega$  représente ici une suite infinie de tirages.

La convergence est qualifiée de « presque » sûre dans le sens suivant : il existe des suites de tirages qui ne donneront pas une convergence vers une valeur unique, mais l'ensemble de ces suites est de probabilité nulle, ce qui fait que expérimentalement on ne les verra pas...

Pour un exemple d'une telle suite, voir en annexe.

On illustre ici la différence entre événement impossible et événement de probabilité nulle.

La convergence presque sûre est une convergence du type convergence simple de fonctions.

Ce résultat n'est pas évident et sa démonstration utilise des théorèmes qui dépassent largement le niveau de cet article...

Les statistiques du 2c) permettent d'avoir une petite idée de la loi de la variable limite  $Y$ .

Dans le cas  $a = b = 1$ , il semble que cette loi soit uniforme mais pas sur un ensemble discret comme  $Y_n$ , mais sur l'intervalle  $[0,1]$ .

On peut établir de manière élémentaire cette loi limite.

Pour cela, on va établir que la suite de fonctions de répartition  $F_{Y_n}$  converge simplement vers la fonction de répartition de la loi uniforme sur  $[0,1]$ .

Il est clair que pour  $x < 0$ ,  $F_{Y_n}(x) = 0 \xrightarrow{n \rightarrow +\infty} 0 = F_Y(x)$

et que pour  $x > 1$ ,  $F_{Y_n}(x) = 1 \xrightarrow{n \rightarrow +\infty} 1 = F_Y(x)$ .

Soit  $0 \leq x \leq 1$  et  $n \in \mathbb{N}^*$ . Il existe  $k \in \{1, \dots, n\}$  tel que

$$\frac{k-1}{n} \leq x < \frac{k}{n} \Leftrightarrow nx < k \leq nx + 1.$$

On a alors :

$$P(Y_n \leq x) = P\left(Y_n \leq \frac{k-1}{n}\right) = \sum_{i=0}^{k-1} P\left(Y_n = \frac{i}{n+2}\right) = \frac{k-1}{n+2}.$$

On déduit de l'encadrement de  $k$  l'encadrement  $\frac{nx-1}{n+2} < F_{Y_n}(x) \leq \frac{nx}{n+2}$  qui donne

$$\lim_{n \rightarrow +\infty} F_{Y_n}(x) = x.$$

On a donc tous les éléments pour affirmer qu'à la « limite » la loi de  $Y_n$  est une loi uniforme à densité sur  $[0,1]$ .

On parle alors de convergence en loi. Celle-ci est moins forte que la convergence presque sûre, puisqu'il s'agit de la convergence de la « loi » et non de la variable elle-même. Comme on l'a dit plus haut, dans notre cas, il y a en fait convergence presque sûre, laquelle entraîne la convergence en loi, mais ce résultat est plus difficile à établir.

Plus généralement, on peut établir que pour  $a$  et  $b$  quelconques, la suite de variables  $(Y_n)$  converge presque sûrement vers une variable dont la densité est de la forme

$$f(x) = C x^{a-1} (1-x)^{b-1} \text{ pour } 0 \leq x \leq 1, \text{ et nulle ailleurs (on parle de loi } \beta \text{).}$$

Pour  $a = b = 2$ , on a une fonction du second degré sur  $0 \leq x \leq 1$  dont la courbe apparaît approximativement sur le diagramme du 2c) et pour  $a = 1, b = 2$ , on a bien une droite.

#### 4. Applications

Le processus des urnes de Polya a en particulier des applications en économie comme le montre l'extrait suivant de l'article « *Quoi de neuf du côté des marchés* » par Nathalie Moureau, Hélène Tordjman (Paris 13), Annick Vignes (Paris 12) :

Ce processus fournit une base théorique susceptible d'expliquer un grand nombre de phénomènes allant de la contagion des opinions sur les marchés

financiers (Orléan (1990), Kirman (1993)) aux modèles de compétition technologique (Arthur (1989)) et à la localisation d'une industrie. Le point commun entre ces différentes problématiques réside dans l'existence d'effets cumulatifs au fur et à mesure du déroulement du processus dus à la présence de rétroactions positives. Kirman propose ainsi par exemple un modèle proche de celui de Polya pour analyser le fonctionnement du marché financier. Il modélise les comportements numériques des agents sur le marché, ceux-ci étant convertis à l'opinion d'autrui au gré de rencontres aléatoires. Mais il introduit en plus une probabilité de conversion automatique des agents d'un groupe d'opinion à l'autre, et évite ainsi que le système ne se bloque dans un état particulier (lock-in) comme c'était le cas dans le processus de Polya. Les deux types de modèle présentent des propriétés dynamiques différentes : tandis que l'un converge vers un état absorbant, l'autre est caractérisé par une instabilité et des oscillations infinies. L'introduction d'une probabilité de conversion spontanée indépendante de toute influence extérieure empêche le système d'atteindre un état absorbant ; celui-ci oscille alors indéfiniment entre deux états. Kirman fournit ainsi une modélisation possible des turbulences que connaissent les marchés financiers et explique l'apparition de bulles spéculatives par le comportement mimétique des agents.

**Annexe :** Exemple de suites de tirages ne convergeant pas.

*On a au départ une boule blanche et une boule rouge, on tire 1 blanche, puis 1 rouge, puis 2 fois de suite une blanche, puis 2 fois de suite une rouge, puis 4 fois de suite une blanche, puis 4 fois de suite une rouge, et ainsi de suite en doublant indéfiniment. Le nombre de blanches prend les valeurs 1, 2, 4, 8, 16, ..., le nombre de rouges aussi, mais avec un décalage. Le rapport nombre de blanches/nombre de rouges oscille donc entre 1 et 2. Et la proportion de boules blanches oscille indéfiniment entre 1/2 et 2/3.*

### **Bibliographie :**

• **Deux livres généraux** (il n'y est pas question de Polya mais on y trouve les notions utilisées dans cet article) :

*Mathématiques TOUT-EN-UN, ECE Première et deuxième année, DUNOD.*

Il n'y a pas que des probabilités, mais celles-ci sont traitées rigoureusement au niveau Bac+1 et +2.

On y trouvera aussi un cours sur l'estimation et des algorithmes pour simuler les lois usuelles.

*L3 Mathématiques Appliquées, A. Yger et J. A. Weil (Eds) Éditeur PEARSON.*

Ce livre contient un cours de probabilité niveau L3, très clair, avec des simulations là-aussi.

• **Un article où Polya évoque son modèle d'urne (en français) :**

*Sur quelques points de la théorie des probabilités, Annales de l'I.H.P, tome 1, n° 2 (1930), p. 117-161.*

[http://www.numdam.org/numdam-bin/item?id=AIHP\\_1930\\_\\_1\\_2\\_117\\_0](http://www.numdam.org/numdam-bin/item?id=AIHP_1930__1_2_117_0)