Lois de probabilité (2/3)

Anita Burgun

Contenu des cours

- *Loi binomiale*
- *Loi de Poisson*
- Loi hypergéométrique
- *Loi normale*
- *Loi du chi2*
- *Loi de Student*

Loi hypergéométrique

- La loi du tirage exhaustif
- Puce à ADN avec des gènes annotés fonctionnellement
- Annotations les plus représentatives

Définition

- Soit une population de N individus parmi lesquels une proportion p (donc Np individus) possède un caractère. On prélève un échantillon de n individus parmi cette population (le tirage pouvant s'effectuer d'un seul coup ou au fur et à mesure mais sans remise). Soit X le nombre aléatoire d'individus de l'échantillon possédant la propriété considérée.
- X suit une loi hypergéométrique

Définition

- X suit une loi hypergéométrique

Nb de groupes de k individus possédant la propriété

Nb de groupes de (n-k) individus ne possédant pas la propriété

$$P(X = k) = \frac{C_{Np}^{k}\ C_{N-Np}^{n-k}}{C_{N}^{n}}$$

Nb d'échantillons possibles

X ~ $\mathcal{H}$ (N, n, p)

Calcul de p(X=k) *cas possibles* C_N^n

on a tiré k boules gagnantes

parmi Np C_{Np}^k

on a tiré (n - k) boules perdantes

parmi N(1 - p) $C_{N(1-p)}^{n-k}$

$$P(X = k) = \frac{C_{Np}^k \ C_{N-Np}^{n-k}}{C_N^n}$$

Espérance

L'espérance d'une v.a. suivant une loi hypergéométrique est la même que dans le cas d'une loi binomiale

$$E(X) = np$$

Variance

$$V(X) = \frac{N-n}{N-1} np(1-p)$$

Illustration

- Dans une urne de 20 boules comprenant des boules rouges et des boules blanches, on effectue un tirage sans remise de 8 boules. La probabilité, à 10^{-4} près de tirer 3 boules rouges est de 35,76% et celle de tirer 4 boules rouges est de 19,87%. Quel est le nombre de boules rouges dans l'urne?

Illustration

- Dans une urne de 20 boules comprenant des boules rouges et des boules blanches, on effectue un tirage sans remise de 8 boules. La probabilité, à 10^{-4} près de tirer 3 boules rouges est de 35,76% et celle de tirer 4 boules rouges est de 19,87%. Quel est le nombre de boules rouges dans l’urne?
- Tirage sans remise de n=8 boules dans une urne de N=20 boules.
- Soit R le nombre de boules rouges, et p=R/N leur proportion
- Soit X le nb de boules rouges parmi les n qui sont tirées
- X suit une loi hypergéométrique de paramètres N=20, n=8 et p

Illustration

- X suit une loi hypergéométrique de paramètres N=20, n=8 et p = R/N
- On connait p(X=3) et p(X=4)

$$P(X=3)=\frac{C_R^3\ C_{20-R}^5}{C_{20}^8}$$

$$P(X=k)=\frac{C_{Np}^k\ C_{N-Np}^{n-k}}{C_N^n}$$

$$P(X=4)=\frac{C_R^4\ C_{20-R}^4}{C_{20}^8}$$

Illustration

- X suit une loi hypergéométrique de paramètres N=20, n=8 et p = R/N
- On a p(X=3) et p(X=4)

$$P(X = k) = \frac{C_{Np}^{k}\ C_{N-Np}^{n-k}}{C_N^n}$$

$$\frac{p(X=3)}{p(X=4)} = \frac{4!}{3!}\frac{(R-4)!}{(R-3)!}\ \ \frac{4!}{5!}\frac{(16-R)!}{(15-R)!} = \frac{4}{5}\frac{(16-R)}{(R-3)}$$

$$R = \frac{15p_3 + 64p_4}{5p_3 + 4p_4} = 7$$

Illustration

- On a une boîte de N comprimés constituée de N_1=Np comprimés défectueux et N_2=N(1-p)= Nq comprimés corrects.
- Donc p=N_1/N
- On veut vérifier le taux de comprimés défectueux. On prélève un échantillon de taille n sans remise.
- Soit X la variable aléatoire discrète « nombre de comprimés défectueux » dans l'échantillon.
- X suit une loi hypergéométrique de paramètres N, n, p
- La probabilité de succès est modifiée d'un tirage à l'autre
- Cf 2e tirage
  - Np cp défectueux et Nq-1 corrects si on a tiré un correct avant
  - Np-1 défectueux et Nq corrects si on a tiré un cp défectueux avant

Illustration

- On fait un sondage dans une population de 1 million de personnes pour rechercher la caractère C. On fait un tirage au sort de 100 personnes
- Dans la population on considère que la répartition de C est de 20%.
- Quelle est la probabilité qu'on ait dans le sondage 10 individus ayant C

$$P(X = k) = \frac{C_{Np}^{k}\ C_{N-Np}^{n-k}}{C_{N}^{n}}$$

$$P(X = 10) = \frac{C_{200.000}^{10}\ C_{1.000.000-200.000}^{90}}{C_{1.000.000}^{100}}$$

Hypergéométrique -> binomiale

quand $N \to \infty$,

$\{X \sim \mathcal{H}(N, n, p)\}$ peut être approximée par

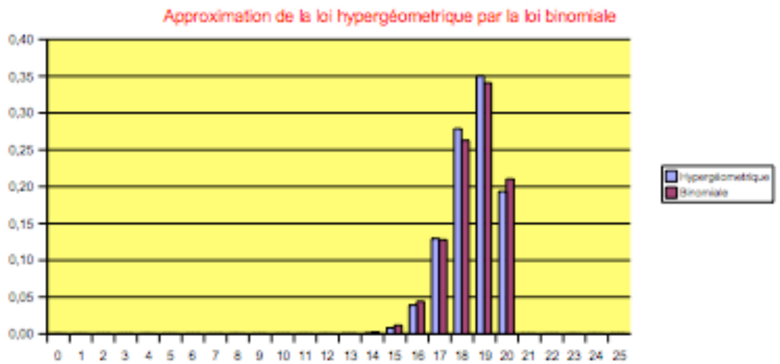
$\{X \sim \mathcal{B}(n, p)\}$

- On ignore le fait que le sondage soit un tirage sans remise
- En pratique quand $n/N < 1/10$ (d'après Saporta, Schwartz)

Approximation binomiale de le loi hypergéometrique:

	N 200	n 20	Rapport des variances
	s 185	p 0,93	VH/VB 0,9
x	P(X=x) Hypergéometrique	P(X=x) Binomiale	H-B Différence
0	0,00000E+000	3,17121E-023	-3,17121E-023
1	0,00000E+000	7,82232E-021	-7,82232E-021
2	0,00000E+000	9,16515E-019	-9,16515E-019
3	0,00000E+000	6,78221E-017	-6,78221E-017
4	0,00000E+000	3,55501E-015	-3,55501E-015
5	1,05978E-018	1,40304E-013	-1,40303E-013
6	4,76902E-016	4,32605E-012	-4,32558E-012
7	8,53655E-014	1,06709E-010	-1,06624E-010
8	8,23066E-012	2,13863E-009	-2,13040E-009
9	4,85609E-010	3,51686E-008	-3,46830E-008
10	1,88028E-008	4,77121E-007	-4,58318E-007
11	4,98558E-007	5,34954E-006	-4,85098E-006
12	9,29455E-006	4,94832E-005	-4,01887E-005
13	1,23689E-004	3,75565E-004	-2,51876E-004
14	1,18192E-003	2,31598E-003	-1,13407E-003
15	8,08431E-003	1,14255E-002	-3,34121E-003
16	3,90436E-002	4,40359E-002	-4,99231E-003
17	1,29380E-001	1,27790E-001	1,58928E-003
18	2,78664E-001	2,62680E-001	1,59837E-002
19	3,49901E-001	3,41023E-001	8,87782E-003
20	1,93612E-001	2,10298E-001	-1,66858E-002
21	0,00000E+000	0,00000E+000	0,00000E+000
22	0,00000E+000	0,00000E+000	0,00000E+000
23	0,00000E+000	0,00000E+000	0,00000E+000
24	0,00000E+000	0,00000E+000	0,00000E+000
25	0,00000E+000	0,00000E+000	0,00000E+000
Différence maximale en valeur absolue			1,66858E-002

N= 200, n=20
p= 0.93



Tiré de:

http://www-timc.imag.fr/Cecile.Amblard/Enseignement/STA230/sortieTP3.pdf

Contenu des cours

- *Loi binomiale*
- *Loi hypergéométrique*
- Loi de Poisson
- *Loi normale*
- *Loi du chi2*
- *Loi de Student*

Loi de Poisson

- La loi des évènements rares
- Survenue d’un accident lors des examens radiologiques
- Effets secondaires des médicaments (pharmacovigilance)

Notion de processus de Poisson

- Evènements rares
- Proba qu'accident sur un intervalle de temps dt proportionnel à durée

  λ = nb moyen d'accidents par unité de temps
- Proba d'observer 2 accidents au cours de dt est très faible par rapport à celle d'en observer 1
- Le nombre d'accidents survenant dans un intervalle de temps donné est indépendant du fait qu'on ait observé beaucoup, ou peu d'accidents dans un autre intervalle de temps précédent.

Loi de Poisson

- Une v.a. X suit une loi de Poisson de paramètre réel positif λ, notée
  $X \sim \mathcal{P}(\lambda)$
  si elle suit:

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

C'est une loi de probabilité

- Rappel

$$e^u = 1 + u + \frac{u^2}{2!} + \frac{u^3}{3!} + \ldots. + \frac{u^n}{n!} + \ldots = \sum_{k=0}^{+\infty} \frac{u^k}{k!}$$

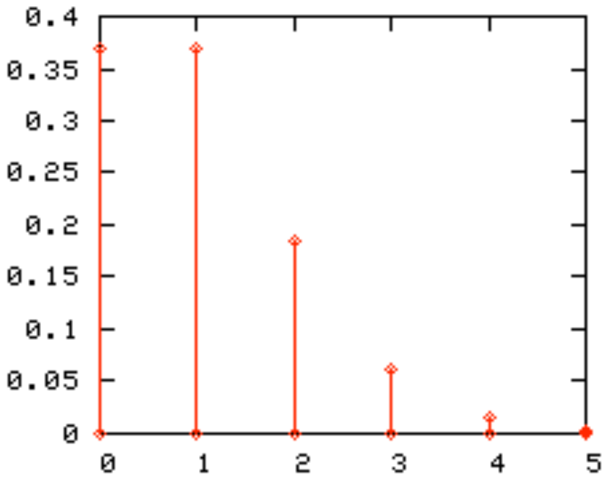
- Loi de probabilité

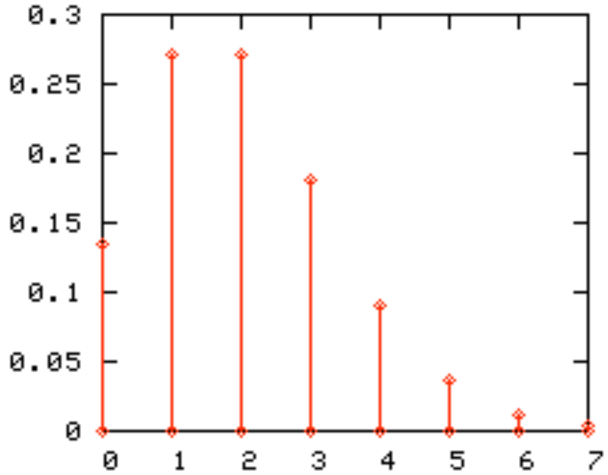
$$\sum_{k=0}^{+\infty} p_k = 1$$

$$\sum_{k=0}^{+\infty} p_k = e^{-\lambda} \sum_{k=0}^{+\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1$$

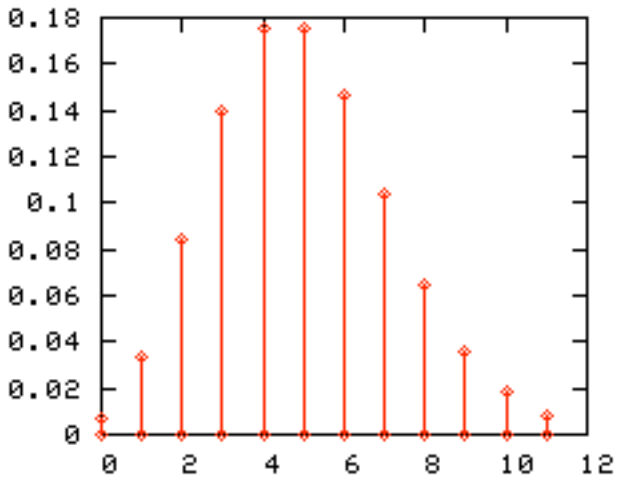
Représentation graphique

$\lambda=1$





$\lambda=2$



$\lambda=5$

Espérance, variance

$$E(X) = \lambda$$

$$V(X) = \lambda$$

Montrons que E(X) = λ

Rappel : $$e^{\lambda} = \sum_{k=0}^{+\infty} \frac{\lambda^k}{k!}$$

$$E(X) = \sum_{k=0}^{+\infty} k \ e^{-\lambda} \ \frac{\lambda^k}{k!} \qquad = e^{-\lambda} \sum_{k=0}^{+\infty} k \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=1}^{+\infty} \frac{\lambda^k}{(k-1)!}$$

$$E(X) = e^{-\lambda} \sum_{k=1}^{+\infty} \lambda \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{k=1}^{+\infty} \frac{\lambda^{k-1}}{(k-1)!}$$

Si on pose m = k-1, on a : $$\sum_{k=1}^{+\infty} \frac{\lambda^{k-1}}{(k-1)!} = \sum_{m=0}^{+\infty} \frac{\lambda^m}{m!} = e^{\lambda}$$

Donc : $$E(X) = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

Remarques utiles

- Lorsque la v.a. X suit une loi de Poisson

$$p_0 = P(X = 0) = e^{-\lambda}$$

(probabilité qu'il n'y ait aucun accident)

$$p_{1+} = P(X > 0) = 1 - e^{-\lambda}$$

(probabilité qu'il y ait au moins un accident)

Remarques utiles (2)

- Si $X_1 \sim \mathcal{P}(\lambda_1)$ et $X_2 \sim \mathcal{P}(\lambda_2)$ avec X_1 et X_2 deux v.a. indépendantes alors X, somme de X_1 et X_2 suit une loi de Poisson
  $X \sim \mathcal{P}(\lambda_1 + \lambda_2)$
- Un traitement produit 2 types d'accidents, les accidents cutanés et les accidents digestifs

Binomiale -> Poisson: illustration

- Un liquide contient 10^5 bactéries par litre. On en prélève 1 mm^3. Quelle est la probabilité que ce prélèvement ne contienne aucune bactérie? Contienne 1, 2, 3 bactéries?
- Quelle loi régit le nombre de bactéries par mm^3 ?

Binomiale -> Poisson: illustration

- *Un liquide contient 10^5 bactéries par litre. On en prélève 1 mm^3. Quelle est la probabilité que ce prélèvement ne contienne aucune bactérie? Contienne 1, 2, 3 bactéries?*
- La probabilité pour qu'une bactérie donnée présente dans le liquide soit dans le mm^3 prélevé est, si les bactéries sont réparties au hasard dans le liquide, c'est à dire si elles ont autant de chances d'être dans chacun des 10^6 mm^3 du litre,

$$\frac{1}{10^6}$$

Binomiale -> Poisson: illustration

- *Un liquide contient 10^5 bactéries par litre. On en prélève 1 mm^3. Quelle est la probabilité que ce prélèvement ne contienne aucune bactérie? Contienne 1, 2, 3 bactéries?*
- *La probabilité pour qu'une bactérie donnée soit dans le mm^3 prélevé est $1/10^6$*
- Nous avons 10^5 bactéries au total
- Donc le nombre de bactéries contenues dans le prélèvement est régi par une loi binomiale de paramètres

$$p = 10^{-6} \qquad n = 10^5$$

Binomiale -> Poisson: illustration

- *La probabilité pour qu'une bactérie donnée soit dans le mm^3 prélevé est $1/10^6$. Nous avons 10^5 bactéries au total. Donc le nombre de bactéries contenues dans le prélèvement est régi par une* <u>*loi binomiale de paramètres*</u> $p = 10^{-6}$ $\quad n = 10^5$

$$P_0 = (1 - p)^n = (1 - 10^{-6})^{10^5} = 0.90$$

$$P_1 = 10^{-6} \frac{10^5}{1} P_0 = 0.09$$

$$P_2 = 0.0045$$

$$P_3 = 0.00015$$

Par récurrence,
en confondant $1-10^{-6}$
avec 1 et en négligeant
1 ou 2 devant 10^5

Binomiale -> Poisson: illustration

- *La probabilité pour qu'une bactérie donnée soit dans le mm³ prélevé est 1/10⁶. Nous avons 10⁵ bactéries au total. Donc le nombre de bactéries contenues dans le prélèvement est régi par une loi binomiale de paramètres* $p = 10^{-6}$ $n = 10^{5}$

P_k décroit très vite quand k augmente

L'espérance du nb de bactéries contenues dans le prélèvement est np= 0.1

La variance np(1-p) est pratiquement égale à np (donc E).

Approximation binomiale-> Poisson

- La loi binomiale $\mathcal{B}(n; p)$ se confond quand n grand et p petit avec une loi de Poisson de paramètre $\lambda = np$.

- Elle modélise donc les expériences de Bernoulli avec une très faible probabilité de succès, mais avec un grand nombre d'essais (np quasiment constant).
- Si p est proche de 0, alors (1-p) proche de 1, donc np=npq. On vérifie bien E(X)=V(X)

- On considère $p<0.1$ <u>et</u> $n>50$

(source: Lazar, Schwartz)

On s’intéresse à une maladie génétique M qui touche environ 1 individu sur 1 000 000. On s’intéresse au nombre de naissances d’enfants atteints dans une région qui a 100 000 naissances par an

- A- La probabilité pour que sur un an, on n’observe aucun enfant atteint de la maladie M est $(0{,}999999)^{100000}$
- P(non atteint)= q= 1-p= 0,999999
- P(aucun enfant atteint)= q^{100000}
- A est vrai

On s'intéresse à une maladie génétique M qui touche environ 1 individu sur 1 000 000. On s'intéresse au nombre de naissances d'enfants atteints dans une région qui a 100 000 naissances par an

- B- La loi de Poisson permet d'évaluer que la probabilité pour que, sur 100 000 naissances, on n'observe aucun enfant atteint de la maladie M est environ de 0,9
- Loi de Poisson applicable: la loi binomiale $\mathcal{B}$(n; p) se confond quand n grand et p petit avec une loi de Poisson de paramètre λ= np.
- λ= 0.1
- B vrai

$$p_0 = P(X = 0) = e^{-\lambda}$$

On s’intéresse à une maladie génétique M qui touche environ 1 individu sur 1 000 000. On s’intéresse au nombre de naissances d’enfants atteints dans une région qui a 100 000 naissances par an

- C La loi de Poisson permet d’évaluer que la probabilité pour qu’il y ait, sur 100 000 naissances, exactement un enfant atteint de la maladie M est de 0,09
- k=1
- C vrai

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

On s’intéresse à une maladie génétique M qui touche environ 1 individu sur 1 000 000. On s’intéresse au nombre de naissances d’enfants atteints dans une région qui a 100 000 naissances par an

- D La loi de Poisson permet d’évaluer que la probabilité pour qu’il y ait, sur 100 000 naissances, au moins un enfant atteint de la maladie M est de 0,1
- D vrai

$$p_{1+} = P(X > 0) = 1 - e^{-\lambda}$$

On s’intéresse à une maladie génétique M qui touche environ 1 individu sur 1 000 000. On s’intéresse au nombre de naissances d’enfants atteints dans une région qui a 100 000 naissances par an

- A, B, C, D vrai

Sachant que l'incidence du diabète dans la population générale est de 2%, quelle est, dans une école de 200 élèves, la probabilité pour qu'aucun élève ne soit diabétique

n=200

p=0,02

On peut utiliser une loi de Poisson quand p<0,1 <u>et</u> n>50

$$p(X=0)=e^{-\lambda}$$

$$p=0{,}018$$

Sachant que l'incidence du diabète dans la population générale est de 2%, quelle est, dans une école de 200 élèves, la probabilité pour qu'un élève soit diabétique

Loi de Poisson

$$p(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

$$p(X = 1) = e^{-4} \frac{4}{1!}$$

$$p(X = 1) = 0{,}073$$

Sachant que l'incidence du diabète dans la population générale est de 2%, quelle est, dans une école de 200 élèves, la probabilité pour que plus de 2 élèves soient diabétiques

Loi de Poisson

$$p(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

$$p(X > 2) = 1 - p(X \leq 2) = 1 - (p(X = 0) + p(X = 1) + p(X = 2))$$

$$p(X = 0) = 0{,}018$$

$$p(X = 1) = 0{,}073$$

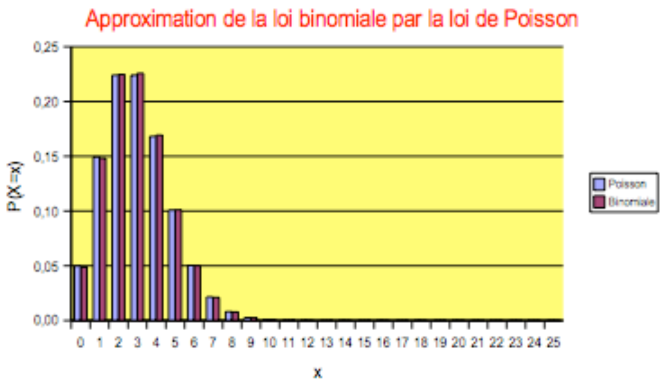
$$p(X = 2) = 0{,}144$$

$$p(X > 2) = 0{,}765$$

Approximation poissonnienne de la loi binomiale:

	N 200		Rapport des variances
	P 0,015	Lambda 3	VH/VB 0,99

x	P(X=x) Poisson	P(X=x) Binomiale	P-B Différence
0	4,97871E-002	4,86683E-002	1,11878E-003
1	1,49361E-001	1,48228E-001	1,13291E-003
2	2,24042E-001	2,24600E-001	-5,57924E-004
3	2,24042E-001	2,25740E-001	-1,69802E-003
4	1,68031E-001	1,69305E-001	-1,27352E-003
5	1,00819E-001	1,01067E-001	-2,48462E-004
6	5,04094E-002	5,00206E-002	3,88801E-004
7	2,16040E-002	2,11109E-002	4,93087E-004
8	8,10151E-003	7,75586E-003	3,45651E-004
9	2,70050E-003	2,51967E-003	1,80833E-004
10	8,10151E-004	7,32879E-004	7,72724E-005
11	2,20950E-004	1,92774E-004	2,81765E-005
12	5,52376E-005	4,62364E-005	9,00121E-006
13	1,27471E-005	1,01825E-005	2,56466E-006
14	2,73153E-006	2,07120E-006	6,60330E-007
15	5,46306E-007	3,91110E-007	1,55196E-007
16	1,02432E-007	6,88661E-008	3,35663E-008
17	1,80763E-008	1,13509E-008	6,72542E-009
18	3,01272E-009	1,75737E-009	1,25535E-009
19	4,75692E-010	2,56351E-010	2,19341E-010
20	7,13538E-011	3,53296E-011	3,60241E-011
21	1,01934E-011	4,61156E-012	5,58184E-012
22	1,39001E-012	5,71390E-013	8,18619E-013
23	1,81306E-013	6,73411E-014	1,13964E-013
24	2,26632E-014	7,56305E-015	1,51001E-014
25	2,71958E-015	8,10821E-016	1,90876E-015
Différence maximale en valeur absolue			1,69802E-003



$n = 200$

$p = 0.015$

$\lambda = 3$

Tiré de:

http://www-timc.imag.fr/Cecile.Amblard/Enseignement/STA230/sortieTP3.pdf