Lois de probabilité
Variables discrètes

Anita Burgun

Cours PCEM 1 (2006-2007)

Problème posé

- Le problème posé en statistique:
  - On s’intéresse à une population
  - On extrait un échantillon
  - On se demande quelle sera la composition de l’échantillon (pourcentage théorique vs pourcentage observé)
  - Calcul des probabilités : lois

Contenu des cours

- Loi binomiale
- Loi hypergéométrique
- Loi de Poisson
- Loi normale
- Loi du chi2 (définition)
- Loi T de Student (définition)

Contenu des cours

- Loi binomiale
- *Loi hypergéométrique*
- *Loi de Poisson*
- *Loi normale*
- *Loi du chi2 (définition)*
- *Loi T de Student (définition)*

Rappel

- Une v.a. X est discrète si l'ensemble des réalisations possibles x_1, x_2,..x_n pour cette variable est fini ou dénombrable.
- Famille de 4 enfants dont les deux parents sont porteurs d'un gène d'une maladie héréditaire. Nombre d'enfants atteints. $X=0,1,2,3,4$
- Nombre d'étudiants ayant les yeux bleus en PCEM1. $X=0,1,2,\ldots\ldots,n$

Définition

- A chacune des réalisations x_i de la v.a. X est associée une probabilité P ($X=x_i$) = p_i

$$\forall i, 0 \leq p_i \leq 1$$

- L'ensemble des couples (x_i , p_i) forme une loi de probabilité.

$$\sum_{i=0}^{n} p_i = 1$$

Loi de Bernoulli

- Soit une expérience aléatoire ayant deux résultats possibles :

  le succès (probabilité p)

  ou l'échec (q=1-p)
- Famille de 4 enfants dont les deux parents sont porteurs d'un gène d'une maladie héréditaire (récessive).
- La v.a. « avoir la maladie m » suit une loi de Bernoulli de paramètre p égal à 0.25
- $E(X) = p$ $\qquad\qquad$ $V(X)= p(1-p)$

Loi binomiale

- On répète n fois dans des conditions identiques une expérience aléatoire dont l'issue se traduit par l'apparition ou la non apparition d'un événement A de probabilité p, le résultat de chaque expérience étant indépendant des résultats précédents.
- Soit X le nombre d'apparitions de l'événement A parmi ces n expériences ($0 \leq X \leq n$).
- X suit une loi binomiale de paramètres n et p.

$$\mathcal{B}(n; p)$$

Loi binomiale

- *On répète n fois dans des conditions identiques une expérience aléatoire dont l'issue se traduit par l'apparition ou la non apparition d'un événement A de probabilité p, le résultat de chaque expérience étant indépendant des résultats précédents.*
- *Soit X le nombre d'apparitions de l'événement A parmi ces n expériences ($0 \leq X \leq n$).*
- *X suit une loi binomiale de paramètres n et p.*
- A chaque expérience on peut associer une variable de Bernoulli

Loi binomiale (def 2)

- La somme de n v.a. de Bernoulli Y_i indépendantes et de même paramètre p est une v.a. discrète qui suit une loi binomiale de paramètres n et p

$$X = \sum_{i=1}^{n} Y_i$$

Loi binomiale

Proba d'obtenir
k succès

$$p(X = k) = C_n^k \; p^k (1 - p)^{n-k}$$

Nombre de manières
de choisir k X_i parmi n

Proba d'obtenir
(n-k) échecs

Loi binomiale

$$p(X = k) = C_n^k \; p^k (1 - p)^{n-k}$$

$$P_k = p(X = k) = \frac{n!}{k!\,(n-k)!} \; p^k (1 - p)^{n-k}$$

Loi BINOMIALE

$$p(X = k) = C_n^k \; p^k (1 - p)^{n-k}$$

Loi BINOMIALE parce que ses différents termes sont les termes du développement de $(p+(1-p))^n$ selon la formule du binôme de Newton

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

$$(p + (1 - p))^n = \sum p_k = 1$$

$$P_0 = (1-p)^n$$

$$P_{k+1} = \frac{n!}{(k+1)!(n-(k+1))!} p^{k+1}(1-p)^{n-k-1}$$

$$P_{k+1} = \frac{n-k}{k+1} \frac{n!}{k!(n-k)!} \frac{p}{(1-p)} p^k (1-p)^{n-k}$$

$$P_{k+1} = \frac{p}{(1-p)} \frac{n-k}{k+1} P_k$$

Pour aider dans les calculs

$$(k+1)! = (k+1) \times k!$$

$$k! = \frac{(k+1)!}{k+1}$$

$$\frac{1}{k!} = \frac{k+1}{(k+1)!}$$

Espérance, variance

- La somme de n v.a. de Bernoulli Y_i indépendantes et de même paramètre p est une v.a. discrète qui suit une loi binomiale de paramètres n et p

$$X = \sum_{i=1}^{n} Y_i$$

Espérance

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

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

$$E(X) = \sum E\,(Y_i)$$

$$E(X) = np$$

Variance

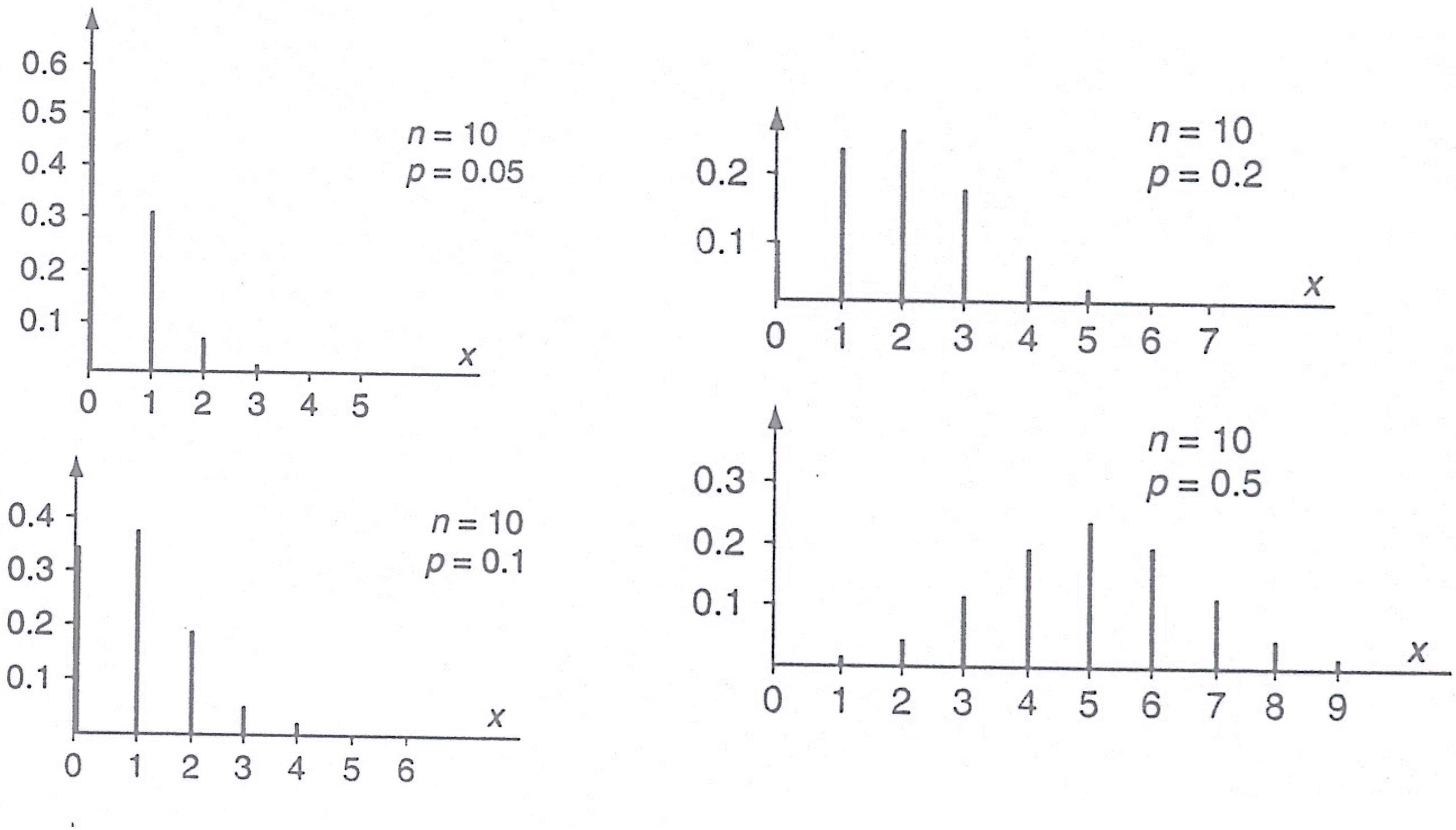
$$V(X_1 + X_2) = V(X_1) + V(X_2)$$

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

$$V(X) = \sum \; V\,(Y_i\,)$$

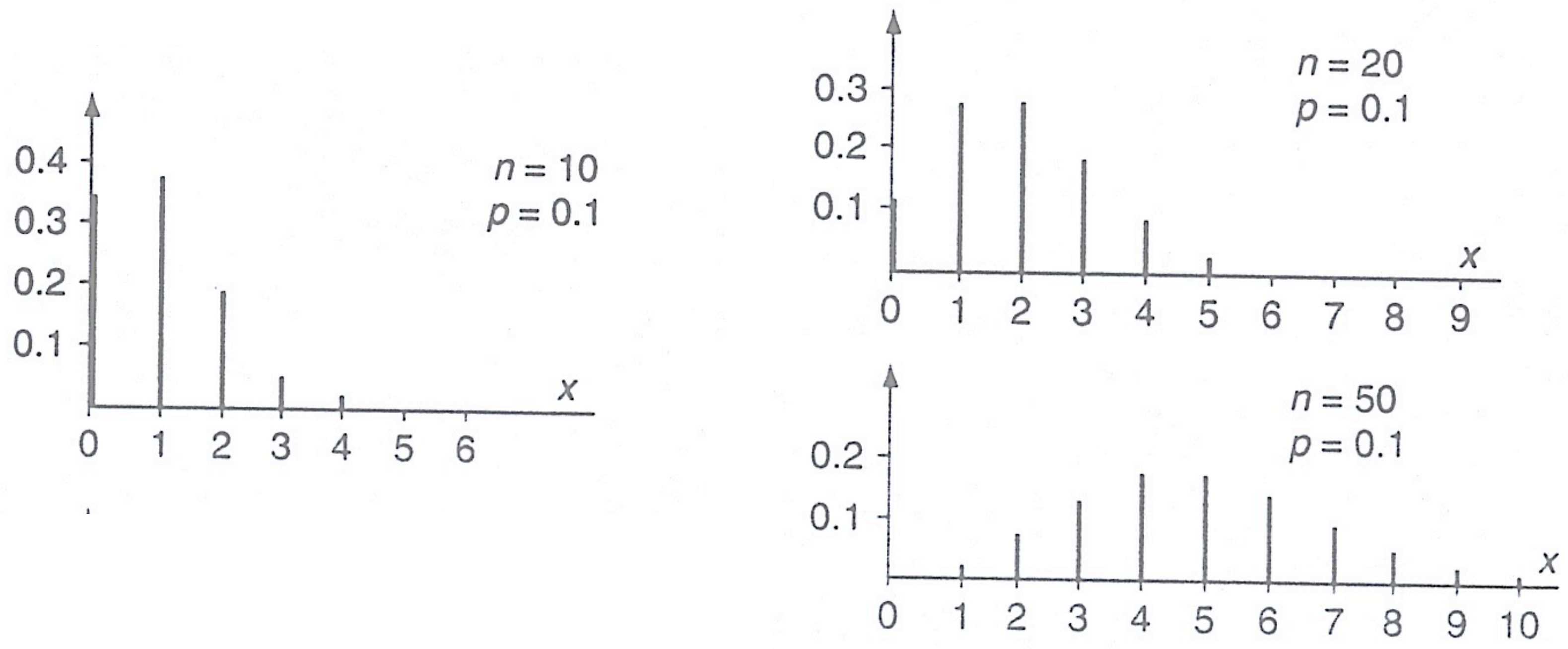
$$V(X) = n\;p\;(1\text{-}p)$$

Représentation graphique (1)



Tiré de G Saporta, Probabilités, analyse de données et statistique Technip

Représentation graphique (2)



Tiré de G Saporta, Probabilités, analyse de données et statistique Technip

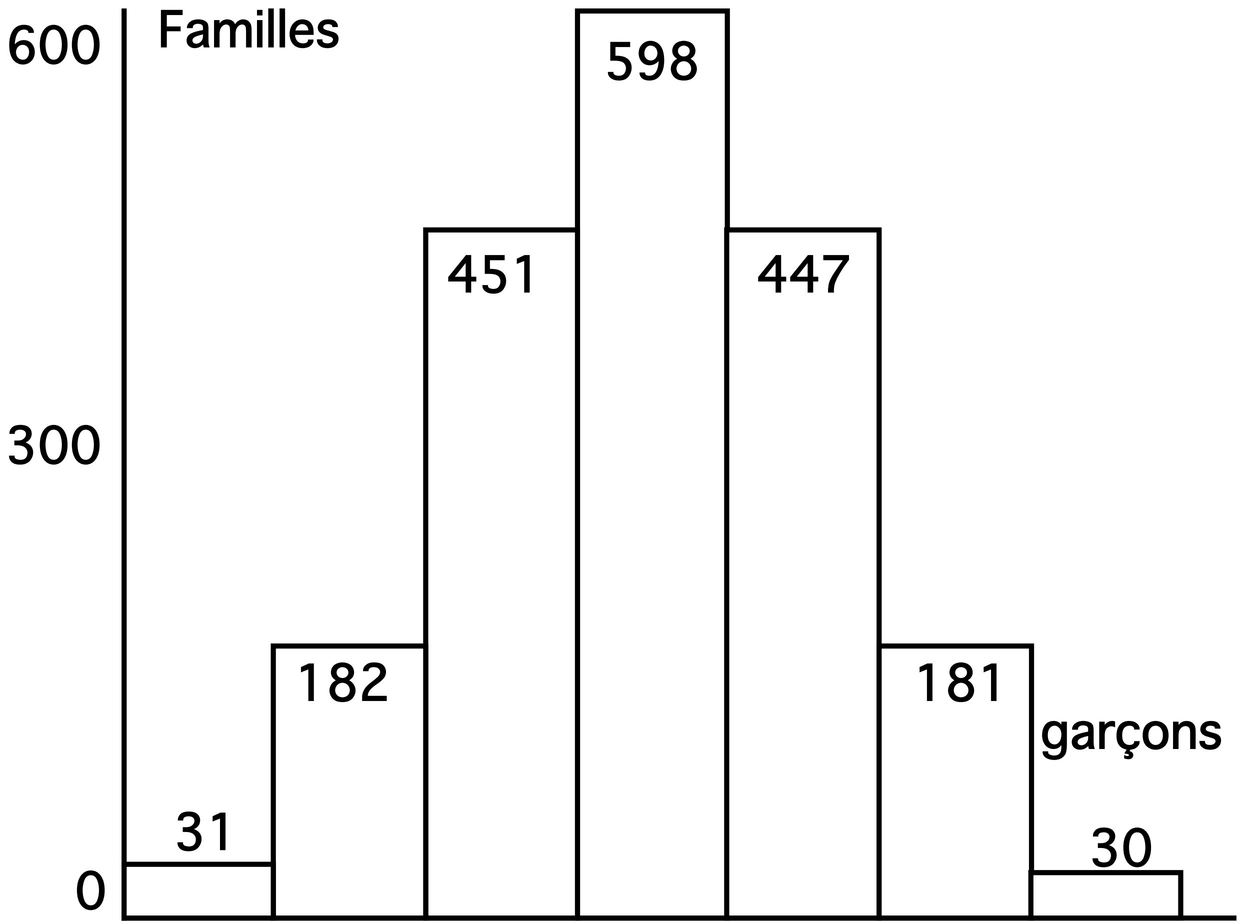
Remarques utiles

- Dans certaines circonstances, la loi binomiale peut être approximée soit par une loi de Poisson (p petit) soit par une loi normale.
- La somme de 2 v.a. binomiales indépendantes et de même paramètre p est une v.a. binomiale

$X_1 + X_2 \qquad \mathcal{B}\ (n_1+n_2;\ p)$

Illustration

- Répartition des sexes à la naissance: p=1/2
- Famille de 6 enfantsloi binomiale n=6 et p=0.5
    - $(1\text{-}p)^6 = 1/64 = 1{,}6\%$ pour 0 garçon
    - $6p(1\text{-}p)^5 = 6/64 = 9{,}4\%$ pour 1 garçon
    - $15p^2(1\text{-}p)^4 = 15/64 = 23{,}4\%$ pour 2 garçons
    - $20p^3(1\text{-}p)^3 = 20/64 = 31{,}2\%$ pour 3 garçons
    - $15p^4(1\text{-}p)^2 = 15/64 = 23{,}4\%$ pour 4 garçons
    - $6p^5(1\text{-}p) = 6/64 = 9{,}4\%$ pour 5 garçons
    - $p^6 = 1/64 = 1{,}6\%$ pour 6 garçons



D'après Lamotte (total: 1920 familles)

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 x individus possédant la propriété

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

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

Nb d'échantillons possibles

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

Espérance, variance

$$E(X) = np$$

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

Illustration

- Un étudiant passant un oral de mathématiques choisit dans une urne 5 enveloppes contenant chacune une question à tirer. Il y a dans l'urne, au total, 12 questions de statistique et 8 questions d'algèbre. On appelle X le nombre de questions de statistique qu'il a tirées. E(X)? V(X)?

Illustration

- Un étudiant passant un oral de mathématiques choisit dans une urne 5 enveloppes contenant chacune une question à tirer. Il y a dans l'urne, au total, 12 questions de statistique et 8 questions d'algèbre. On appelle X le nombre de questions de statistique qu'il a tirées. E(X)? V(X)?

$$E(X) = 5 \times 0.6 = 3$$

$$V(X) = np(1-p)\frac{N-n}{N-1} = 5 \times 0.6 \times 0.4 \times \frac{15}{19} = 0.9474$$

Hypergéométrique -> binomiale

quand $N \rightarrow \infty$,

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

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

- En pratique quand $n/N < 1/10$ (d'après Saporta, Schwartz)

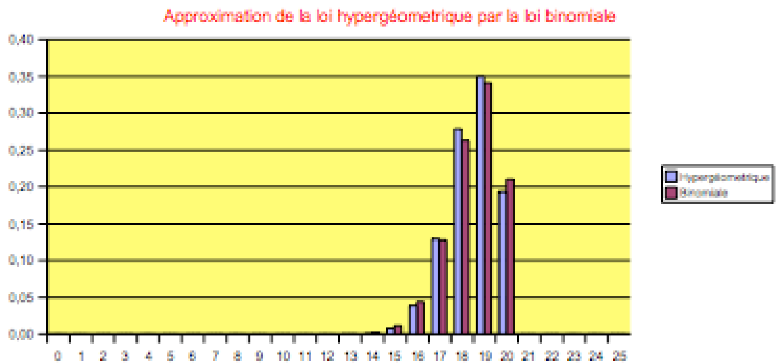
Approximation binomiale de le loi hypergéometrique:

N		n		Rapport des variances VH/VB
	200		20	
s		p		
	185		0,93	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

Lois de probabilité (2/3)

Anita Burgun

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

- Il faut que 3 hypothèses soient vérifiées
- Proba qu'accident sur 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 ~ $\mathcal{P}$ (λ)

  si elle suit:

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

C’est une loi de probabilité

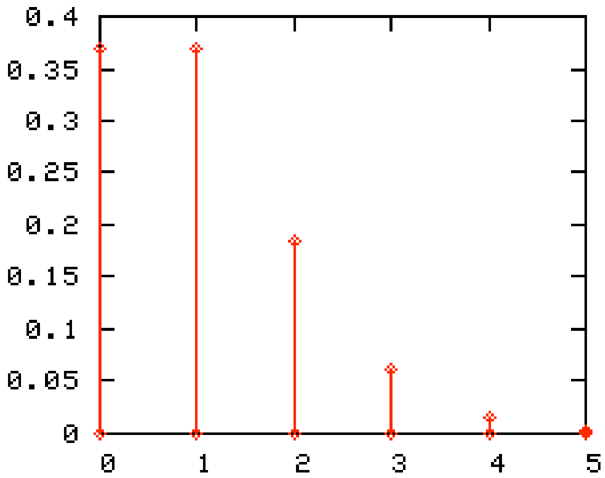
- Rappel $$e^u = 1 + u + \frac{u^2}{2!} + \frac{u^3}{3!} + ... + \frac{u^n}{n!} + ... = \sum_{k=0}^{+\infty} \frac{u^k}{k!}$$

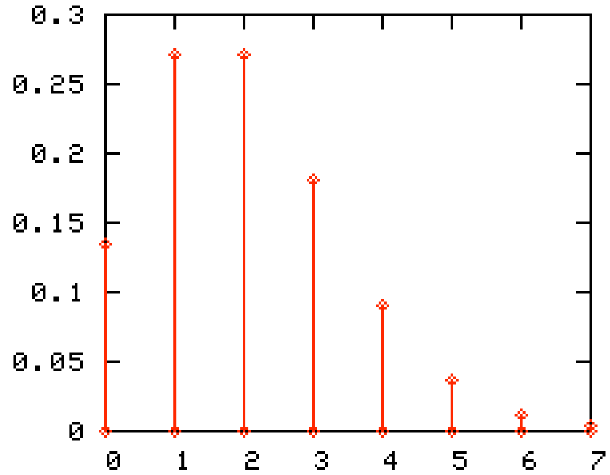
- Loi de probabilité $$\sum_{k=0}^{+\infty} p_k = 1$$

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

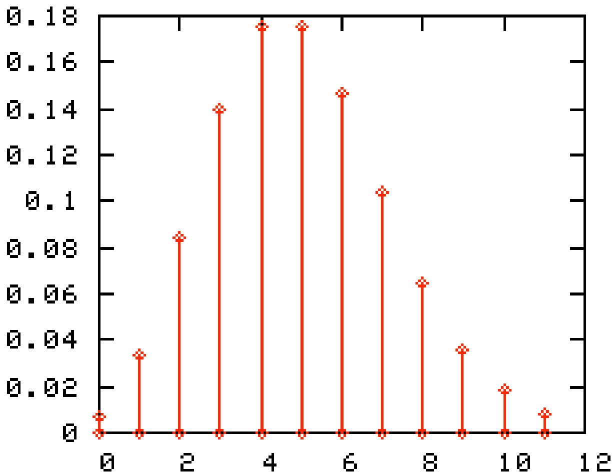
Représentation graphique

λ=1





λ=2



λ=5

Espérance, variance

$$E(X) = \lambda$$

$$V(X) = \lambda$$

Montrons que E(X) = λ

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

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

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

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?

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, 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 <u>loi binomiale de paramètres</u>

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

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

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

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

$$P_2 = 10^{-6}\frac{10^5}{2}P_1 = \frac{P_1}{20} = 0.0045$$

$$P_3 = 10^{-6}\frac{10^5}{3}P_2 = \frac{P_2}{30} = 0.00015$$

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

Binomiale -> Poisson: illustration

- *La probabilité pour qu'une bactérie donnée soit dans le 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>* $p = 10^{-6}$ $n = 10^{5}$

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 $\lambda = np$.

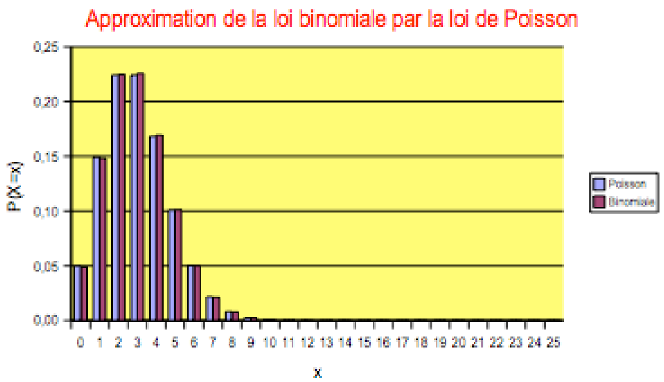
- Elle modélise donc les expériences de Bernoulli avec une très faible probabilité de succès, mais avec un grand nombre d'essais (np quasiment constant).

- On considère $p<0.1$ et $n>50$

(source: Lazar, Schwartz)

Approximation poissonnienne de la loi binomiale:

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

λ = 3

Tiré de:

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