Statistique et probabilités

Alessandro Torcini et Andreas Honecker

LPTM

Université de Cergy-Pontoise

La loi de Bernoulli

La loi de Bernoulli décrit une situation ou seulement deux résultats sont possibles, comment à pile ou face
On s'est mis d'accord d'appeler les deux résultat 0 (pile) et 1 (face). La loi de Bernoulli est défini par

$$p(1) = p\,, \qquad p(0) = 1 - p := q \qquad \text{avec} \qquad p \in [0, 1]$$

Jouer à pile ou face avec une pièce de monnaie parfaitement équilibrée satisfait à

$$p = q = \frac{1}{2}$$

.
La situation devient plus interessante quand on répète l'expérience.

Pile ou Face

Supposons qu'on lance une pièce de monnaie équilibrée N fois et cherche la probabilité de trouver n faces dans la suite.

Pour $N = 2$ on a (p = pile, f = face) avec $p(p) = p(f) = \frac{1}{2}$:

$$\begin{aligned} p(n=0) &= p(\mathrm{pp}) = p(\mathrm{p})\,p(\mathrm{p}) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4} \\ p(n=1) &= p(\mathrm{pf}) + p(\mathrm{fp}) = 2\,p(\mathrm{p})\,p(\mathrm{f}) = 2\cdot\frac{1}{4} = \frac{1}{2} \\ p(n=2) &= p(\mathrm{ff}) = p(\mathrm{f})\,p(\mathrm{f}) = \frac{1}{4} \end{aligned}$$

et pour $N = 3$ on trouve

$$\begin{aligned} p(n=0) &= p(\mathrm{ppp}) = p(\mathrm{p})^3 = \left(\frac{1}{2}\right)^2 = \frac{1}{8} \\ p(n=1) &= p(\mathrm{ppf}) + p(\mathrm{pfp}) + p(\mathrm{fpp}) = 3\,p(\mathrm{p})^2\,p(\mathrm{f}) = 3\cdot\frac{1}{8} = \frac{3}{8} \\ p(n=2) &= p(\mathrm{ffp}) + p(\mathrm{fpf}) + p(\mathrm{pff}) = 3\,p(\mathrm{f})^2\,p(\mathrm{p}) = 3\cdot\frac{1}{8} = \frac{3}{8} \\ p(n=3) &= p(\mathrm{fff}) = p(\mathrm{f})^3 = \frac{1}{8}\,. \end{aligned}$$

Le point central est que $p(x,y) = p(x)\cdot p(y)$, $p(x,y,x) = p(x)\cdot p(y)\cdot p(z)$ soit que les lancements de pièces de monnaie sont vraiment indépendants.

Processus de Bernoulli

Pour N lancements d'une pièce de monnaie pas équilibrée , (avec une pièce truquée)
donc on a une variable aléatoire de Bernoulli, avec $p(0) = q$, $p(1) = p$ et $p + q = 1$
Pour $N = 2$ on trouve :

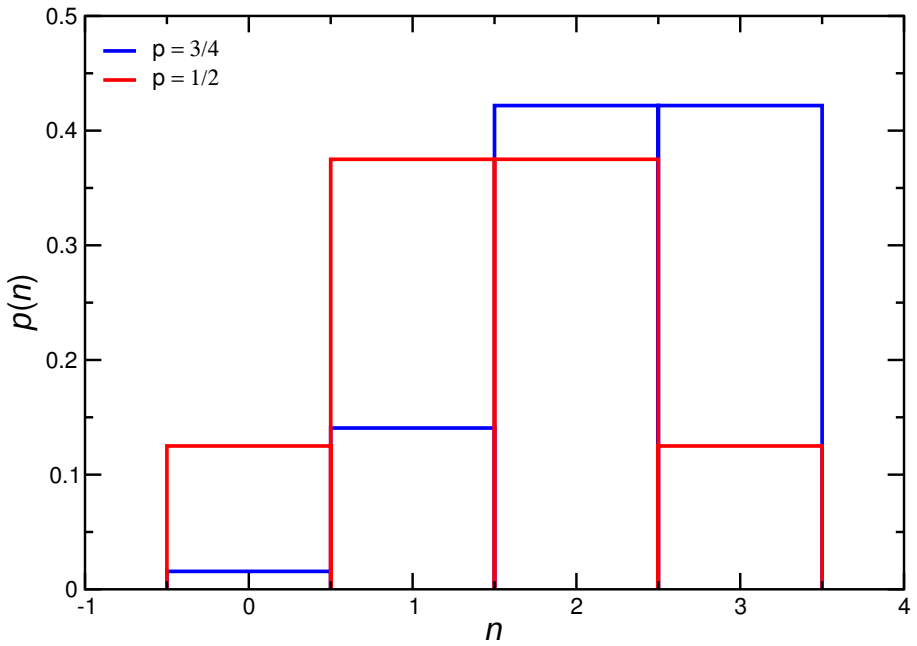
$$\begin{aligned} p(n=0) &= p(00) = p(0)\,p(0) = q^2 \\ p(n=1) &= p(01) + p(10) = 2\,p(1)\,p(0) = 2\,p\,q \\ p(n=2) &= p(11) = p(1)\,p(1) = p^2 \end{aligned}$$

et pour $N = 3$ vous trouvez

$$\begin{aligned} p(n=0) &= p(000) = p(0)^3 = q^3 \\ p(n=1) &= p(001) + p(010) + p(100) = 3\,p(0)^2\,p(1) = 3\,p\,q^2 \\ p(n=2) &= p(110) + p(101) + p(011) = 3\,p(1)^2\,p(0) = 3\,p^2\,q \\ p(n=3) &= p(111) = p(1)^3 = p^3\,. \end{aligned}$$

Encore une fois $p(x,y) = p(x) \cdot p(y)$, $p(x,y,x) = p(x) \cdot p(y) \cdot p(z)$

Processus de Bernoulli



1. Histogramme de la probabilité $p(n)$ de trouver une somme n de $N = 3$ variables aléatoires de Bernoulli.
2. La ligne rouge ($p = 1/2$) correspond à une pièce équilibré et indique donc la probabilité de trouver n faces parmi trois pièces.
3. La ligne bleue montre le résultat avec $p = 3/4$, soit $p > q = 1/4$.
4. Évidemment, quand p est plus grand, il est plus probable de trouver une somme grande (et moins probable de trouver une somme petite).

Loi binomiale

Les résultats précédents peuvent être réécrites pour $N = 3$ comment :

$$(p+q)^3 = p^3 + 3p^2q + 3q^2p + q^3 = p(n=3) + p(n=2) + p(n=1) + p(n=0) \equiv 1$$

ou p est la probabilité de avoir un 1 dans la suite et q de observer un 0 et $p(n)$ est la probabilité de trouver une somme n pour la somme de $N = 3$ variables aléatoires de Bernoulli.

Avez-vous reconnu le système derrière les résultats ? C'est la formule du binôme de Newton

$$(p+q)^N = \sum_{n=0}^{N} \binom{N}{n} p^n \, q^{N-n} \qquad \text{avec} \qquad \binom{N}{n} = \frac{N!}{n!\,(N-n)!}$$

ou $\binom{N}{n}$ est le coefficient binomial.

Par conséquent, si on pose $p + q = 1$, on arrive à la probabilité de trouver une somme n pour la somme N variables aléatoires de Bernoulli

$$p(n) = \binom{N}{n} p^n \, q^{N-n} \qquad \text{pour} \qquad 0 \leq n \leq N$$

ceci c'est la loi binomiale

Les coefficients binomiaux

Nous avons donc besoin des réalisation des coefficients binomiaux. On peut utiliser la formule $\binom{N}{n} = \frac{N!}{n!\,(N-n)!}$ et la factorielle que nous connaissons déjà.
Mais les valeurs de la factorielle deviennent assez vite très grande lors que le coefficient binomial reste un entier pas trop grande. Vous pouvez éviter des problèmes provoqués par des résultats intermédiaires grandes si vous utilisez la formule de Pascal

$$\binom{N}{n} = \binom{N-1}{n-1} + \binom{N-1}{n}$$

avec les conditions initiales $\binom{N}{0} = \binom{N}{N} = 1$

```
def binomialC(N,n):
  if n<0 or n>N:                                  # on ne sait jamais
    return 0
  if n==0 or n==N:                                # conditions initiales
    return 1
  return binomialC(N-1,n-1) + binomialC(N-1,n)    # formule de Pascal
```

Pour les valeurs de N grandes, cette solution pour le coefficient binomial devient lent.

Les coefficients binomiaux

Donc, je vous donne encore une autre solution. On utilisant la récursivité

$$\binom{N}{n} = \prod_{i=1}^{n} \frac{N-i+1}{i} = \frac{N-n+1}{n}\binom{N}{n-1}$$

pour $n > 0$.

Par conséquent, nous avons une deuxième réalisation pour le coefficient binomial en Python :

```
def binomialC(N,k):
  if k>N/2:                                  # utiliser symmetrie
     k = N-k
  if k<0:                                    # on ne sait jamais
    return 0
  if k==0:                                   # condition initiale
    return 1
  return ((N-k+1)*binomialC(N,k-1))/k        # recursivite
```

Cette solution est beaucoup plus vite de la précédente sur le coût d'un résultat intermédiaire qui est un facteur n plus grande (en Python, ça ne fait rien, mais pour des autres langues ceci peut être important).

Moyenne en statistique

On a une série statistique des mesures des valuers d'une variable aleatoire x,

$$\{x_1, x_2, x_3, \ldots, x_N\}$$

avec M valeurs possibles $\{X_1, X_2, \ldots, X_M\}$ et leur fréquences d'occurrence sont

$$\{f_1, f_2, f_3, \ldots, f_M\} = \{\frac{n_1}{N}, \frac{n_2}{N}, \frac{n_3}{N}, \ldots, \frac{n_M}{N}\}$$

ou $\sum_i f_i = 1$ et donc la moyenne est

$$M(x) = X_1 f_1 + X_2 f_2 + X_3 f_3 + \cdots + X_M f_M = \sum_{i=1}^{M} X_i f_i \equiv \frac{\sum_{j=1}^{N} x_j}{N}$$

Pour une variable X avec une distribution de probabilités $p(X)$ la moyenne est

$$M(x) = \int_0^1 dX \;\; X \;\; p(X) \qquad \text{avec} \qquad X \in [0, 1]$$

Variance en statistique

La deuxième caractérisation importante d'une variable aléatoire est sa dispersion autour de la valeur moyenne $M(x)$. On peut l'estimer en calculant un "moment d'inertie" de la variable par rapport à cette valeur moyenne. Par exemple à partir de N mesures $\{x_j\}$, on calculera

$$\sigma^2(x) = \frac{1}{N}\sum_{j=1}^{N}(x_j - M_x)^2 = \frac{1}{N}\sum_{j=1}^{N} x_j^2 - M(x)^2 = M(x^2) - M(x)^2$$

or

$$\sigma^2(x) = \sum_{i=1}^{M} X_i^2 f_M - (\sum_{i=1}^{M} X_i f_M)^2$$

σ^2 est la "variance" de la variable aléatoire.

1. Sa racine carrée, σ est son écart-type. Cette quantité caractérise la dispersion de la variable autour de la moyenne.
2. Lorsque la variance est nulle, la densité de probabilité est concentrée en un point et la variable aléatoire est connue avec certitude.
3. La notion de variance est fondamentale dans un bon nombre d'applications en particulier en traitement du signal.

Calculer moyenne et variance

```
import numpy as np                  # importer le module comme "np"
N=10000000          # repetitions
moyenne = 0.
variance=0.
i = 0
while i<N:
 r = np.random.random_integers(0,100)   # Nombres aleatoires 0 <= r < 100
 moyenne = moyenne + r                  # avec Distribution plate
 variance = variance + r*r
 i += 1

moyenne=moyenne/N
variance=variance/N - moyenne*moyenne
ecart=np.sqrt(variance)

print "moyenne", moyenne, " variance",  variance, "écart-type", ecart
```

Moyenne et variance pour la binomiale

Pour la loi binomiale $\boxed{p(k) = \binom{N}{k} p^k \, q^{N-k}}$ on trouve

$$M(k) = \sum_{k=0}^{N} k \, p(k) = \sum_{k=0}^{N} k \binom{N}{k} p^k \, q^{N-k} = N \, p \, ,$$

$$\sigma^2(k) = \left(\sum_{k=0}^{N} k^2 \binom{N}{k} p^k \, q^{N-k} \right) - \langle k \rangle^2 = N \, p \, q \, .$$

Démonstration

La première dérivée de l'identité $(p+q)^N = \sum_{k=0}^{N} \binom{N}{k} p^k \, q^{N-k}$ par rapport à p (q fixe) donne :

$$N(p+q)^{N-1} = \sum_{n=0}^{N} \binom{N}{k} k \, p^{k-1} \, q^{N-k}$$

en multipliant les deux côtés pour p on obtient

$$Np(p+q)^{N-1} = \sum_{n=0}^{N} \binom{N}{k} k \, p^k \, q^{N-k} \equiv M(k)$$

en notant que $p + q = 1$ le résultat pour la moyenne est trouvé.

La variance pour la loi binomiale

La dérivée seconde de l'identité $(p+q)^N = \sum_{k=0}^{N} \binom{N}{k} p^k\, q^{N-k}$ par rapport à p (q fixe) donne :

$$N(N-1)(p+q)^{N-2} = \sum_{n=0}^{N} \binom{N}{k} k\,(k-1)\, p^{k-2}\, q^{N-k}$$

en multipliant les deux côtés pour p^2 on obtient

$$N(N-1)p^2 = \sum_{n=0}^{N} \binom{N}{k} k^2\, p^k\, q^{N-k} - \sum_{n=0}^{N} \binom{N}{k} k\, p^k\, q^{N-k} \equiv M(k^2) - M(k)$$

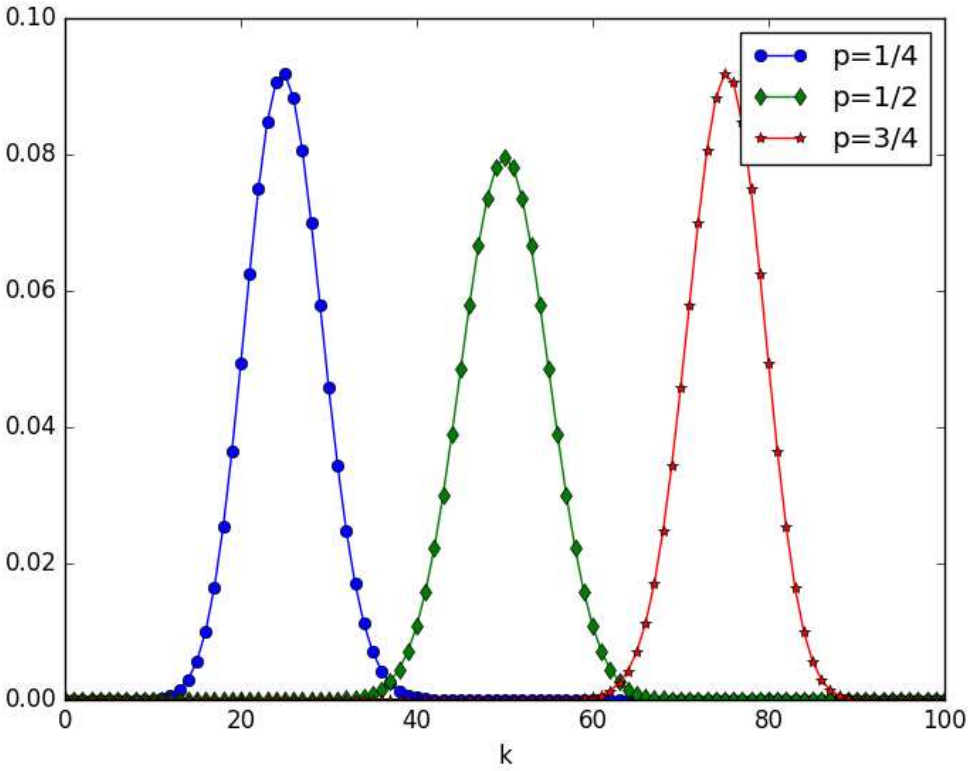
en rappellant que $M(k) = Np$

$$\sigma^2(k) = M(k^2) - M(k)^2 = N(N-1)p^2 + Np - (Np)^2 = Np(1-p) = Npq$$

Tracer la loi binomiale

la loi binomiale

$$p(k) = \binom{N}{k} p^k \, q^{N-k} \qquad \text{pour} \qquad 0 \leq k \leq N$$



IDLE

Théorème de la limite centrale

Le Théorème de la limite centrale

Soit $\mathbf{X} = \{x1, x2, \ldots, x_i \ldots\}$ une suite de variables aléatoires réelles définies sur le même espace de probabilité, indépendante et avec la même distribution de probabilité. Supposons que la moyenne μ et l'écart type σ de $\mathbf{X}$ existent et soient finis avec $\sigma \neq 0$.

Considérons la moyenne (somme) de N variables aléatoires x_i

$$S_N = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Alors

1. La moyenne de S_N est μ ;
2. Son écart type vaut $\sigma_0 = \frac{\sigma}{\sqrt{N}} \to 0$ pour $N \to \infty$
3. De plus, quand n est assez grand, la loi normale $\mathcal{N}(\mu, \sigma_0 = \sigma/\sqrt{N})$ est une bonne approximation de la distribution de probabilités de $y = S_N$

$$\mathcal{N}(\mu, \sigma_0)(y) = \frac{1}{\sigma_0 \sqrt{2\pi}} \mathrm{e}^{-\frac{(y-\mu)^2}{2\,\sigma_0^2}},$$

Loi normale ou de Gauss

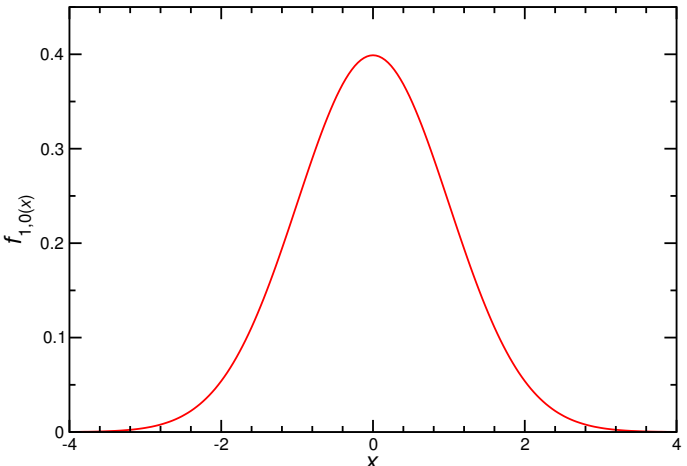
La loi normale est caractérisé par la distribution de probabilité suivante :

$$\mathcal{N}(\mu, \sigma_0) = \frac{1}{\sigma_0 \sqrt{2\pi}} \, \mathrm{e}^{-\frac{(y-\mu)^2}{2\,\sigma_0^2}} \, ,$$

Le facteur $\frac{1}{\sigma_0 \sqrt{2\pi}}$ assure la normalisation de probabilité

$$\int_{-\infty}^{\infty} \mathrm{d}y \, f_{\sigma_0,\mu}(y) = 1 \, .$$

(Noter que dans la limite $N \to \infty$ il faut remplacer les sommes par les integrales).



La figure montre la forme du loi normale qui est vraiment très similaire à la loi binomiale pour N grande. Cette similarité n'est pas une coïncidence : selon le théorème de la limite centrale on trouve toujours une loi normale si on considère une somme d'un grand nombre de variables aléatoires indépendantes.

Loi normale ou de Gauss

Je cite encore les résultats pour la moyenne et la variance pour une variable normale y :

$$M(y) = \int_{-\infty}^{\infty} \mathrm{d}y\, y\, \mathcal{N}_{\sigma_0,\mu}(y) = \mu\,,$$

$$\sigma^2(y) = \int_{-\infty}^{\infty} \mathrm{d}y\, y^2\, \mathcal{N}_{\sigma_0,\mu}(y) - M(y)^2 = \sigma_0^2\,.$$

Vous voyez que μ est la moyenne et σ_0^2 la variance de la loi normale.