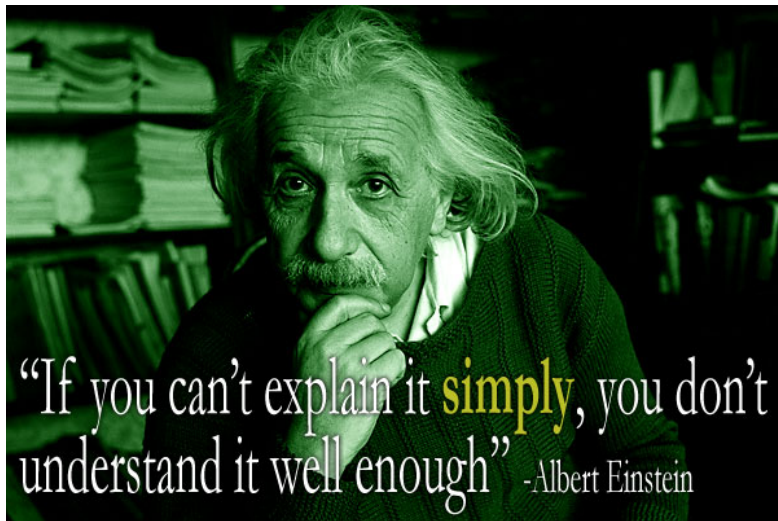


# Simulation sur ordinateur

A. Buys

Département d'Informatique  
Université de Mons





- 1 Introduction
- 2 Génération de nombres aléatoires suivant une loi uniforme
- 3 Tests des séries générées
- 4 Génération de nombres aléatoires suivant des lois non uniformes
- 5 Divers

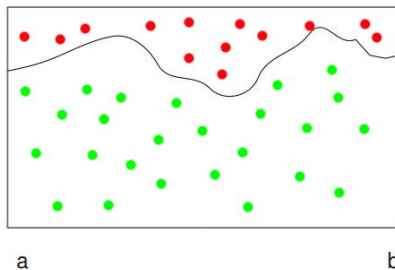
# Introduction

- Qu'est-ce qu'un nombre aléatoire ?
- Qu'est-ce qu'une variable aléatoire ?
- Qu'est-ce qu'une séquence de nombres aléatoires ?
  - uniformité (toutes les valeurs sont équiprobables)
  - indépendance (dimensions)
- A quoi servent des nombres pseudo-aléatoires ?
  - Problèmes non déterministes
  - Problèmes avec trop de variables

Référence: Donald E. Knuth "The Art of Computer programming" (Addison Wesley), Vol. 2.

■ (...)

■ Calcul de  $\int_a^b f(x)dx$



(rarement à 2 dimensions, plutôt 6-7)

## ■ (...)

- Physique (interactions entre particules, efficacité d'un détecteur, etc)
- Echantillonnage (sondage, ...)
- Programmation (tests de programme)
- Génération de mots de passe
- Prise de décision (avec facteurs non quantifiables)
- Jeux (début de partie)

## ■ A quoi reconnaît-on une séquence de nombres “aléatoires” ?

- Quelle est la séquence la plus probable ?

1 1 1 1 1 1 1 1 1

1 4 6 9 8 7 2 3 5

1 2 3 4 5 6 7 8 9



- On veut une séquence qui ait les mêmes caractéristiques qu'une séquence réellement aléatoire

0 4 6 9 8 7 6 3 5

0 0 4 4 6 6 9 9 8 8 7 7 6 6 3 3 5 5

→ notion de dimension



- Comment obtenir des nombres “aléatoires” ?
    - Fichier de nombres aléatoires (Roulette à Monte-Carlo)
      - Inconvénient: toujours les mêmes, lent
    - Machine du lotto (périphérique qui génère des nombres aléatoires)
      - Inconvénient: jamais les mêmes, lent
    - Nombres pseudo-aléatoires
      - obtenus à partir d'un nombre “semence” de façon déterministe
- même semence → même séquence  
semence différente → séquence différente



# Génération de nombres aléatoires suivant une loi uniforme

- Middle-square (Von Neuman, 1946)  
... historique
- Congruence linéaire
- Autres méthodes (parfois plus modernes)

## Middle-square (Von Neuman, 1946)

- On part d'un nombre de 10 chiffres (par exemple)
- $(7548962587)^2 = 56986836139925732569$
- Dangereux car possibles dégénérescences
  - $xxxxx00000 \longrightarrow xxxxx00000$
  - On peut obtenir des centaines de milliers de nombres qui satisfont une batterie de tests avant dégénérescence
  - Pour les curieux ...  $(3792)^2 = 14379264$   
 $3792 = 3\ 79\ 2^4$

## Congruence linéaire (Lehmer, 1948)

- On raisonne sur des entiers, mais si  $0 \leq X_i < m$  alors  $0 \leq X_i/m < 1$  (loi uniforme dans  $[0,1[$ )
- De façon générale, on considère une fonction  $y = f(x)$  sur des entiers  $f(X_0) = X_1 ; f(X_1) = X_2 ; \dots$
- Soit  $0 \leq X_0 \leq m$ ,  $0 \leq f(x) = y < m$ , alors on a une période maximale égale à  $m$  (après avoir généré  $m$  valeurs, on va forcément retomber sur l'une d'elles).
- Soit  $\lambda$  la période,  $\mu$ , l'amorce ( $1 \leq \lambda \leq m$ ,  $0 \leq \mu < m$ ,  $\mu + \lambda \leq m$ )  
 $X_0 X_1 X_2 \dots X_{\mu-1} X_\mu \dots X_{\mu+\lambda-1} X_{\mu+\lambda} \quad (X_\mu = X_{\mu+\lambda})$   
 $\forall i, j \text{ tel que } (i \neq j) \wedge (i < \mu + \lambda) \wedge (j < \mu + \lambda) : X_i \neq X_j$
- Comment déterminer  $\mu$  et  $\lambda$  pour une série quelconque ?  
 On va rechercher la plus petite valeur de  $n$  telle que  $X_n = X_{2n}$



## A démontrer

$\exists n$  tel que  $X_n = X_{2n}$  et  $\mu \leq n \leq \mu + \lambda$

$$X_r = X_n (r \geq n) \iff \begin{cases} (r - n) \text{ multiple de } \lambda \\ n \geq \mu \end{cases}$$

$\Downarrow$

$$X_{2n} = X_n \iff \begin{cases} n \text{ multiple de } \lambda \\ n \geq \mu \end{cases}$$

$\Downarrow$

Le plus petit multiple  $n$  de  $\lambda$  ( $n \geq \mu$ ) est tel que  $X_n = X_{2n}$

Or, entre  $\mu$  et  $\mu + \lambda$ , il existe un multiple de  $\lambda$ . Si on part de  $X_0$ , c'est le premier qu'on va rencontrer (pas d'égalité avant  $\mu$ ).

- On a donc  $n$  tel que  $X_n = X_{2n}$ , il faut maintenant trouver  $\mu, \lambda$ .

$$X_0 \dots X_{\mu-1} X_\mu \dots X_n \dots X_{2n}$$

On va comparer  $X_i$  et  $X_{n+i}$  ( $i \geq 0$ ),  $X_0$  avec  $X_n$ ,  $X_1$  avec  $X_{n+1}$ , etc.

La première égalité donne  $i = \mu$ .

- On pourrait partir de  $X_n = X_{2n}$ ,  $X_{n-1} = X_{2n-1}$ , etc, mais on ne connaît pas  $f^{-1}$ .
- Si un des  $X_{n+i}$  vaut  $X_n$ , alors  $\lambda = i$ , sinon,  $\lambda = n$
- On peut aussi partir de  $X_n$  pour trouver  $X_{n+\lambda}$

- Fonction généralement utilisée

$$f(x) = (ax + c) \bmod m \quad (\text{congruence } \underline{\text{linéaire}})$$

$m$  : module

$a$  : multiplicateur

$c$  : incrément

$X_0$  : valeur de départ (semence)

- Méthode congruentielle “mixte”. Si  $c = 0$ , méthode multiplicative.

## A démontrer

*Si  $a$  et  $m$  sont premiers entre eux, alors  $\mu = 0$ .*

(Par l'absurde)

Si  $\mu \neq 0$ ,  $\exists X_\mu$  précédé de  $X_{\mu-1}$  et  $\exists \alpha$  tel qu'on trouve dans la séquence ...  $X_\alpha X_\mu$  ... avec  $X_\alpha \neq X_{\mu-1}$ .

On va montrer qu'on ne peut pas avoir un nombre avec deux prédécesseurs différents si  $a$  et  $m$  sont premiers entre eux.



- Soient  $r, s$  tels que  
 $(a X_{r-1} + c) \bmod m = (a X_{s-1} + c) \bmod m$  et  $X_{r-1} \neq X_{s-1}$

$$\Downarrow$$
$$a (X_{r-1} - X_{s-1}) = \alpha m$$

Or,  $-m + 1 \leq (X_{r-1} - X_{s-1}) \leq m - 1$   
... mais doit être multiple de  $m$  si  $\alpha \neq 0$ . CQFD.

- Il suffit alors de regarder  $X_0, X_1, \dots$  pour déterminer  $\lambda$ .
- Exemple:  $a = c = 7, m = 10, X_0 = 7 \implies 7, 6, 9, 0, 7, 6, 9, 0, \dots$

## A démontrer

$$X_{k+n} = \left(a^k X_n + \frac{a^k - 1}{a - 1} c\right) \bmod m$$

(Par récurrence)

- $k = 1$  évident
- $$\begin{aligned} X_{k+n+1} &= a \left(a^k X_n + \frac{a^k - 1}{a - 1} c\right) + c \bmod m \\ &= a^{k+1} X_n + \left(\frac{a^{k+1} - a}{a - 1} + 1\right) c \bmod m \\ &= a^{k+1} X_n + \left(\frac{a^{k+1} - a + a - 1}{a - 1}\right) c \bmod m \end{aligned}$$

CQFD.

$$X_{k+n} = (a' X_n + c') \bmod m \quad \text{avec}$$

$$\begin{cases} a' = a^k \\ c' = \frac{a^k - 1}{a - 1} c \end{cases}$$



## ■ Choix du module $m$

- $m$  limite la période qu'on peut obtenir

Exemple extrême : simulation de pile ou face avec  $m=2$ .

- On a intérêt à choisir  $m$  grand
- Rapidité de calcul

$X_{n+1} = (aX_n + c) \bmod m$  ; On choisit  $0 \leq a, c \leq m - 1$

- Lors du calcul, le plus grand nombre est de  $(m - 1)(m - 1) + m - 1 = m^2 - m < m^2$
- On prend  $m = 2^e$  où  $e$  est le nombre de bits d'un mot d'ordinateur
- $m^2$  est représenté par  $2e$  bits et la division par  $m$  est un décalage de  $e$  positions.  
 $\implies$  Pas de division à effectuer, maximalisation de la période.

## A démontrer

Soit  $d$  qui divise  $m$ ,  $Y_i = X_i \bmod d$ , alors  $Y_{n+1} = (a'Y_n + c') \bmod d$   
avec  $a' = a \bmod d$ ,  $c' = c \bmod d$

$$X_{n+1} = aX_n + c - qm$$



$$Y_{n+1} = ((a' + \alpha d)(Y_n + \beta d) + c' + \gamma d) \bmod d$$

$$Y_{n+1} = (a'Y_n + c' + (... )d) \bmod d$$

- Pour  $m = 2^e$ ,  $d = 2$ , le dernier bit des  $X_i$  a une période de 2 (pas bon pour générer un pile ou face)
- Si 4 divise  $m$ , période de 4, par exemple 00,01,10,11,...
- Caractère aléatoire moins bon pour les bits moins significatifs.
- Peu important pour générer une loi uniforme.
- Alternatives :  $m = 2^e - 1$ , plus grand premier  $< 2^e$ .

## ■ Choix de $a$ et $c$

### Théorème

(Hull et Dobel, 1962)

La congruence définie par  $m$ ,  $c$ ,  $a$  et  $X_0$  est de période  $m$  si et seulement si

- $c$  est premier avec  $m$
- $b = a - 1$  est multiple de  $p$ ,  $\forall p$  premier diviseur de  $m$
- $b$  est multiple de 4 si 4 divise  $m$

- Assure une période maximum mais pas forcément une bonne génération aléatoire ( $a = 1, c = 1$ )

- Congruence multiplicative ( $c = 0$ )

$$X_{k+n} = a^k X_n \mod m$$

- Si  $X_n$  est multiple d'un diviseur de  $m$ , tous les nombres générés le sont.
- On prend  $X_0$  premier avec  $m$  et on essaie de conserver des nombres premiers avec  $m$ .
- La période maximum sera  $\varphi(m)$  (fonction d'Euler).

- En général  $m = p_1^{e_1} p_2^{e_2} \dots p_n^{e_n}$
- Soit  $m = p^e$ , si  $a$  est un multiple de  $p$ ,  
 $X_{n+e} = a^e X_n \bmod p^e = 0$  et la période vaut 1.
- Soit  $a$  premier avec  $p$ , la période est le plus petit  $\lambda$  tel que  
 $a^\lambda X_n \bmod p^e = X_{n+\lambda} = X_n \implies a^\lambda = 1 \bmod p^e$
- $\lambda$  est l'ordre de  $a$  modulo  $p^e$
- Un élément  $a$  pour lequel cet ordre est le plus élevé est un "élément primitif"
- On choisit alors  $a$  comme étant un des éléments primitifs.
- La période maximale  $\lambda(m)$  (pour  $c = 0$ ) est obtenue avec  $X_0$  premier avec  $m$  et est le PPCM des  $\lambda_i$  relatifs aux  $p_i^{e_i}$ .

## ■ Reste à trouver les éléments primitifs ...

### Théorème

(Carmichael, 1910)

*Le nombre  $a$  est un élément primitif modulo  $p^e$  si et seulement si*

*i)  $p^e = 2$ ,  $a$  est impair; ou  $p^e = 4$ ,  $a \bmod 4 = 3$ ; ou  $p^e = 8$ ,  $a \bmod 8 = 3, 5, 7$ ;  
ou  $p = 2, e \geq 4$ ,  $a \bmod 8 = 3, 5$*

*ii)  $p$  est impair,  $e = 1$ ,  $a \not\equiv 0 \bmod p$ , et  $a^{(p-1)/q} \not\equiv 1 \bmod p$  pour aucun diviseur premier  $q$  de  $p-1$*

*iii)  $p$  est impair,  $e > 1$ ,  $a$  satisfait (ii) et  $a^{p-1} \not\equiv 1 \bmod p^2$*

## Autres méthodes

- Comment obtenir une période  $> m$  ?

On peut par exemple partir de  $X_0, X_1$  et

$$(X_{n+1} = aX_n + bX_{n-1} + c) \mod m$$

( $m^2$  doublets différents au maximum, période maximale :  $m^2 + 1$  )

- $c = 0, a = b = 1$  “Mauvais exemple”
- $(X_{n+1} = X_{n-24} + X_{n-55}) \mod m \quad (n \geq 55)$
- Méthodes quadratiques
- Fibonnaci avec décalages de registre
- “Mersenne Twister” (Makoto Matsumoto et Takuji Nishimura, 1997)
  - Congruence linéaire (produit matriciel) et décalages de registre
  - basé sur  $m = 2^{19937} - 1$  (nombre de Mersenne)
  - “uniformément distribué sur un grand nombre de dimensions (623 pour les nombres de 32 bits)”
  - Pas très bon pour de la cryptographie.

- Nombres remarquables ( $\pi$ ,  $e$ , ...)

$e = 2.71828182845904523536028747135266249...$

$\pi = 3.14159265358979323846264338327950288...$

"The Quest for Pi", David H. Bailey, Jonathan M. Borwein, Peter B. Borwein and Simon Plouffe,  
June 25, 1996, Mathematical Intelligencer, vol. 19 , no. 1 (Jan. 1997), pp. 50–57  
<http://crd.lbl.gov/~dhbailey/dhbpapers/pi-quest.pdf>

- Section suivante : comment juger du caractère pseudo-aléatoire d'une séquence de nombres ?
  - Et si on vous demandait de générer des nombres à la main ?
  - $\pi = 3.14159$ 2653589793238462643383279...

# Tests des séries générées

- Rappel d'analyse
- Rappels de théorie des probabilités (et statistique)
- Test du  $\chi^2$
- Test de Kolmogorov-Smirnov
- Test du gap
- Test du poker
- Test du collectionneur de coupons
- Test de permutation
- Méthode des “runs”
- Test du maximum



## Rappel d'analyse

- Soit une fonction  $F$  de  $\mathbb{R}^n$  dans  $\mathbb{R}^m$  :

$$F : \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \rightarrow \begin{pmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_m(x_1, \dots, x_n) \end{pmatrix}$$

$$\text{alors } J_F = \frac{\partial(f_1, \dots, f_m)}{\partial(x_1, \dots, x_n)} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

est la matrice jacobienne de  $F$ .

- Si  $m = n$ , on peut calculer son déterminant  $\det J_F$  ("Jacobien"), dont la valeur absolue intervient lors d'un changement de variables dans les intégrales multiples.

- Soit par exemple une fonction  $F$  de  $\mathbb{R}^2$  dans  $\mathbb{R}^2$  :

$$F : \begin{pmatrix} u \\ v \end{pmatrix} \rightarrow F(u, v) = \begin{pmatrix} x = f_1(u, v) \\ y = f_2(u, v) \end{pmatrix}$$

alors

$$\int \int g(x, y) dx dy = \int \int g(F(u, v)) |\det J_F| du dv$$

- Soit

$$\begin{cases} x = r \cos \theta \\ y = r \sin \theta \end{cases} \rightarrow J_F = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{pmatrix}$$

et  $\det J_F = r$ .

■ Donc

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) dx dy = \int_0^{+\infty} \int_0^{2\pi} g(F(r, \theta)) r dr d\theta$$

$$\begin{aligned} \left( \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx \right)^2 &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}} dx dy \\ &= \int_0^{+\infty} \int_0^{2\pi} e^{-\frac{r^2}{2}} r dr d\theta \\ &= 2\pi \int_0^{+\infty} e^{-\frac{r^2}{2}} r dr \\ &= 2\pi \left[ -e^{-\frac{r^2}{2}} \right]_0^{+\infty} = 2\pi \end{aligned}$$



## Rappels de théorie des probabilités



■ Loi binomiale  $P_X(k) = \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}$   $E(X) = np, \sigma^2(X) = np(1-p)$

■ Loi de Poisson  $P_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}$   $E(X) = \sigma^2(X) = \lambda$

■ Loi normale  $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$

■ Loi exponentielle  $f(x) = \lambda e^{-\lambda x}$   $E(X) = \frac{1}{\lambda}, \sigma^2(X) = \frac{1}{\lambda^2}, P(x \geq t) = e^{-\lambda t}$

- Somme de variables aléatoires indépendantes (convolution des densités)



$$f_{X+Y}(z) = \int_{-\infty}^{+\infty} f_X(z-y) f_Y(y) dy$$

- Changement de variables :  $Y = g(X)$




$$f_Y(y) = \int_{-\infty}^{+\infty} \delta(y - g(x)) f_X(x) dx$$

avec la distribution (fonctionnelle)  $\delta$  de Dirac :

$$\int_{-\infty}^{+\infty} \delta(x - y) f(x) dx = f(y)$$



- Au sens des fonctionnelles,  $\delta(g(x)) = \sum_i \frac{\delta(x-x_i)}{|g'(x_i)|}$  où  $g(x_i) = 0$  et  $g'(x_i) \neq 0$ .
- Par exemple,  $X^2$  avec  $X$  de loi  $N(0,1)$  

$$\begin{aligned}
 f_{X^2}(x) &= \int_{-\infty}^{+\infty} dy \, \delta(x - y^2) \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \\
 &= \int_{-\infty}^{+\infty} dy \, \frac{1}{2|y|} \{ \delta(y + \sqrt{x}) + \delta(y - \sqrt{x}) \} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \\
 &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{x}} e^{-x/2} \\
 &= \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2} \quad (x > 0)
 \end{aligned}$$

$\chi^2$  à 1 degré de liberté.

- Autre exemple,  $W(t)$  densité de probabilité de  $X^2 + Y^2$  avec  $X, Y$  de lois  $N(0,1)$  indépendantes

$$W(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy e^{-x^2/2} e^{-y^2/2} \delta(t - x^2 - y^2)$$

$$\delta(t - x^2 - y^2) = \frac{1}{2\sqrt{t - y^2}} \left\{ \delta(x - \sqrt{t - y^2}) + \delta(x + \sqrt{t - y^2}) \right\}$$

$$\begin{aligned} W(t) &= \frac{1}{2\pi} \int_{-\sqrt{t}}^{+\sqrt{t}} dy \frac{e^{-y^2/2}}{2\sqrt{t - y^2}} \int_{-\infty}^{+\infty} dx e^{-x^2/2} \{ \dots \} \\ &= \frac{1}{2\pi} \int_{-\sqrt{t}}^{+\sqrt{t}} dy \frac{e^{-y^2/2}}{2\sqrt{t - y^2}} (e^{-(t-y^2)/2} + e^{-(t-y^2)/2}) \\ &= \frac{1}{2\pi} e^{-t/2} \int_{-\sqrt{t}}^{+\sqrt{t}} \frac{dy}{\sqrt{t - y^2}} \end{aligned}$$

$$\begin{aligned}
 W(t) &= \frac{1}{2\pi} e^{-t/2} \int_{-\sqrt{t}}^{+\sqrt{t}} \frac{dy}{\sqrt{t-y^2}} \\
 &= \frac{1}{2\pi} e^{-t/2} \int_{-1}^{+1} \frac{1}{\sqrt{1-(\frac{y}{\sqrt{t}})^2}} d(\frac{y}{\sqrt{t}}) \\
 &= \frac{1}{2\pi} e^{-t/2} \int_{-1}^{+1} \frac{dx}{\sqrt{1-x^2}} \\
 &= \frac{1}{\pi} e^{-t/2} \int_0^{+1} \frac{dx}{\sqrt{1-x^2}} \\
 &= \frac{1}{\pi} e^{-t/2} \int_0^{+1} d(\arcsin x) = \frac{1}{\pi} e^{-t/2} \frac{\pi}{2} \\
 &= \frac{1}{2} e^{-t/2}
 \end{aligned}$$

$\chi^2$  à 2 degrés de liberté.

- On peut généraliser pour  $n$  variables aléatoires indépendantes de lois  $N(0,1)$

$$f_n(x) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-x/2} \Theta(x)$$

où  $\Theta$  est la fonction de Heaviside ( $\Theta(x) = 0$  si  $x < 0$ ,  $1$  si  $x \geq 0$ ).

(au sens des fonctionnelles,  $\Theta'(x) = \delta(x)$ )

- Pour  $\chi_n^2$ ,  $E(X) = n$ ,  $\sigma^2(X) = 2n$



## ■ Probabilité conditionnelle

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Par exemple, pour la loi exponentielle  $P(x \geq t) = e^{-\lambda t}$  et

$$\begin{aligned} P(x \geq t + h | x \geq t) &= \frac{P(x \geq t + h \cap x \geq t)}{P(x \geq t)} \\ &= \frac{P(x \geq t + h)}{P(x \geq t)} \\ &= \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} = e^{-\lambda h} \\ &= P(x \geq h) \end{aligned}$$

(Loi dite “sans mémoire”)

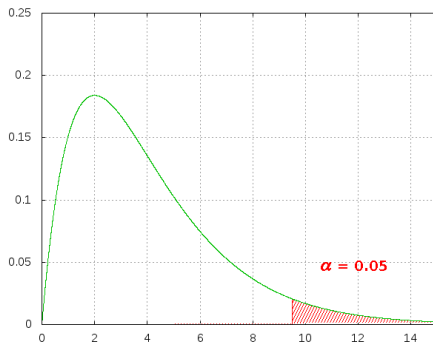
## ■ Théorème de Bayes ("théorème sur la probabilité des causes")

Soient  $\{B_j\}$  un ensemble d'événements tels que la somme de leurs probabilités vaut 1,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$$

En pratique, les probabilités "a priori"  $P(B_j)$  ne sont pas toujours connues.

- Tests d'hypothèse (version courte et adaptée à ce qui suit)
  - Le but est de déterminer si une hypothèse  $H_0$  (hypothèse “nulle”) est raisonnable en fonction d'une observation  $O$  par rapport à une hypothèse alternative  $H_1$  (Exemple : les nombres générés sont-ils distribués uniformément ou pas ?).
  - On ne va pas essayer de calculer la probabilité de  $H_0$  (probabilité a priori inconnue) mais déterminer si la probabilité conditionnelle par rapport à  $H_0$  de ce qu'on observe  $P(O|H_0)$  dépasse un certain seuil fixé. C'est ce qu'on appelle l'erreur de première espèce  $\alpha$  qui est la probabilité de rejeter une hypothèse  $H_0$  alors qu'elle est vraie (l'erreur de seconde espèce  $\beta$  à l'inverse est la probabilité d'accepter  $H_0$  fausse).
  - Si on considère la distribution théorique d'une grandeur observée (sur base de l'hypothèse  $H_0$ ), on obtient par exemple la courbe représentée sur la figure suivante :

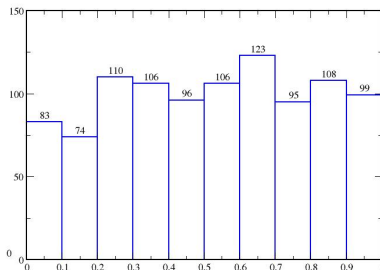


- On peut alors rejeter l'hypothèse  $H_0$  si notre observation se trouve dans la zone d'exclusion (hachuré sur la figure). NB: dans l'exemple, on a choisi une grandeur toujours positive et le test est unilatéral.

## Test du $\chi^2$

- On crée un histogramme ( $r$  intervalles) en comptant le nombre de valeurs générées dans chaque intervalle ( $n_i$ ).
- On s'attend à avoir “à peu près” le même nombre de points dans chaque intervalle (probabilités  $p_i$  égales).

Rand (awk) 1000 random numbers



## ■ On construit

$$K_r = \sum_{i=1}^r \frac{(n_i - (\sum_{j=1}^r n_j)p_i)^2}{(\sum_{j=1}^r n_j)p_i} = \sum_{i=1}^r \left( \frac{n_i - Np_i}{\sqrt{Np_i}} \right)^2$$

- Dans l'exemple,  $r = 10$ ,  $N = \sum_{j=1}^r n_j = 1000$ ,  $p_i = 0.1$ ,  $K_r = 17.72$

- La distribution des effectifs  $n_i$  suit une loi multinomiale :

$$P(X_1 = n_1, \dots, X_r = n_r) = \frac{N!}{n_1! \dots n_r!} p_1^{n_1} \dots p_r^{n_r} \text{ si } \sum_{j=1}^r n_j = N, 0 \text{ sinon.}$$

- $\frac{X_i - Np_i}{\sqrt{Np_i}} \xrightarrow[N \rightarrow \infty]{} N(0, \sqrt{1 - p_i})$ ,  $\sum_{i=1}^r \left( \frac{X_i - Np_i}{\sqrt{Np_i}} \right)^2 \xrightarrow[loi]{} \chi_{r-1}^2$

- Les  $Np_i$  sont des variances poissonniennes.

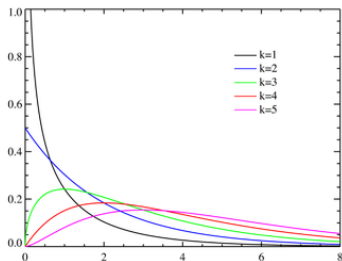
- On obtient une loi de  $\chi^2$  à  $r - 1$  degrés de liberté (on a une contrainte sur  $\sum_{j=1}^r n_j$ ).

- On se donne une probabilité  $\alpha$  telle que l'hypothèse examinée est rejetée si elle a une probabilité  $P \{ \chi^2 \geq \text{valeur obtenue} \} < \alpha$  de se produire (voir table).

$k \setminus \alpha$	0.10	0.05	0.025	0.01	0.001
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.125
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588
11	17.275	19.675	21.920	24.725	31.264
12	18.549	21.026	23.337	26.217	32.910
13	19.812	22.362	24.736	27.688	34.528
14	21.064	23.685	26.119	29.141	36.123

Figure: Valeurs critiques  $\chi^2_\alpha$  du  $\chi^2$  pour une probabilité  $\alpha$  et par nombre  $k$  de degrés de liberté

- On n'a pas de méthode pour décider à coup sûr si l'hypothèse doit être rejetée. On a juste une probabilité.



- On pourrait multiplier les expériences pour voir si on reproduit la courbe.

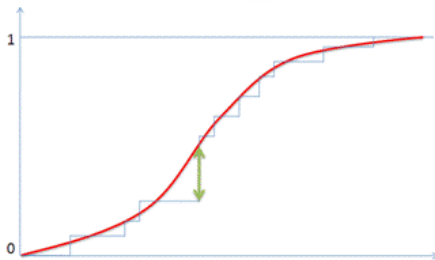


## Test de Kolmogorov-Smirnov

- On compare les fonctions de répartition des distributions

$$F(x) = \int_{-\infty}^x f(y) dy = P\{X \leq x\}$$

- On compare la fonction de répartition théorique  $F(x)$  à la fonction de répartition expérimentale  $F_n(x) = \frac{\text{nombre de valeurs} \leq x}{n}$ .
- De façon générale,  $F(-\infty) = 0$ ,  $F(+\infty) = 1$ .
- $F(x)$  doit être continue.



- Pour une loi uniforme  $[0, 1[$ ,  $F(x \leq 0) = 0$ ,  $F(x \geq 1) = 1$ ,  
 $F(0 \leq x \leq 1) = x$ .
- Soit  $D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$

## Théorème

(Kolmogorov-Smirnov)

$$P \{ \sqrt{n} D_n < x \} \xrightarrow{n \rightarrow \infty} K(x) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 n^2} \quad (x > 0)$$

- Comme précédemment, on utilisera une valeur critique  $\alpha$ :

$$P \{ D_n > D_\alpha \} = \alpha$$

ou bien, de façon équivalente,

$$P \{ \sqrt{n} D_n > \sqrt{n} D_\alpha \} = \alpha$$

$n \setminus \alpha$	0.20	0.15	0.10	0.05	0.01
3	.565	.597	.642	.708	.828
4	.494	.525	.564	.624	.733
5	.446	.474	.510	.565	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.410	.490
15	.266	.283	.304	.338	.404
20	.231	.246	.264	.294	.356
25	.210	.220	.240	.270	.320
30	.190	.200	.220	.240	.290
35	.180	.190	.210	.230	.270
$\geq 35$	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Figure: Valeurs critiques  $D_\alpha$  de KS pour une probabilité  $\alpha$  et pour une taille  $n$  d'échantillon

- Contrairement au  $\chi^2$ , on ne fait pas de classes (intervalles) et on ne perd pas d'information.
- Le calcul est aussi plus gourmand (mémoire, classement).
- Dans la littérature, on trouve des variantes

$$K_n^+ = \sqrt{n} \sup_{x \in \mathbb{R}} (F_n(x) - F(x))$$

$$K_n^- = \sqrt{n} \sup_{x \in \mathbb{R}} (F(x) - F_n(x))$$

$D_\alpha \rightarrow K_\alpha$  (et table correspondante)

Avec cette formulation, il existe des algorithmes qui évitent de devoir classer les données.

$n \setminus \alpha$	0.25	0.05	0.01
3	0.7539	1.1017	1.3589
4	0.7642	1.1304	1.3777
5	0.7674	1.1392	1.4024
10	0.7845	1.1658	1.4440
15	0.7926	1.1773	1.4606
20	0.7975	1.1839	1.4698
30	0.8036	1.1916	1.4801
$> 30$	$0.8326 - 1/(6\sqrt{n})$	$1.2239 - 1/(6\sqrt{n})$	$1.5174 - 1/(6\sqrt{n})$

**Figure:** Valeurs critiques  $K_\alpha$  de  $K_n^+/K_n^-$  pour une probabilité  $\alpha$  et pour une taille  $n$  d'échantillon

## Test du gap



- On se donne  $0 \leq a < b \leq 1$  et on génère  $u_1, u_2, u_3, u_4, u_5, u_6, u_7, \dots, u_n$
- On marque ceux qui tombent dans  $[a, b]$ , probabilité  $p = b - a$ .
- On s'intéresse aux distances entre deux nombres marqués.
- $(u_j, u_{j+1}, \dots, u_{j+r})$  avec  $u_{j+r} \in [a, b]$  et  $u_j, u_{j+1}, \dots, u_{j+r-1} \notin [a, b]$  est de longueur  $r$ .
- Soit  $l_r$  la probabilité d'avoir une séquence de longueur  $r$ .

$$l_0 = p$$

$$l_1 = (1 - p)p$$

...

$$l_i = (1 - p)^i p$$

$$l_{>i} = (1 - p)^{i+1}$$

- On compte les nombres d'occurrences de chacune des longueurs et on compare aux valeurs attendues  $Np, Np(1 - p), \dots$  (où  $N$  est le nombre de gaps  $\approx np$ )  $\implies \chi^2$ .
- Pour un temps de calcul optimal, choisir  $a$  ou  $b = 0$  ou  $1$ , par exemple  $(a = 0, b = \frac{1}{2})$  ou  $(a = \frac{1}{2}, b = 1)$ .

## Test du poker

- On joue avec  $k = 5$  dés à  $d = 6$  faces.
  - 5 faces identiques : poker
  - 4 faces identiques : carré
  - $3 + 2$  faces identiques : full
  - 3 faces identiques : brelan
  - $2 + 2$  faces identiques : double paire
  - 2 faces identiques : paire
- On divise  $[0, 1[$  en  $d = 6$  intervalles, 5 nombres dans le même intervalle donnent un poker, etc ... et, pour simplifier
  - Poker  $\rightarrow$  une seule case
  - Carré ou full  $\rightarrow$  2 cases différentes
  - Brelan ou double paire  $\rightarrow$  3 cases différentes
  - Paire  $\rightarrow$  4 cases différentes
  - Rien (ici)  $\rightarrow$  5 cases différentes
- Soit  $r$  le nombre de cases différentes, on peut calculer la probabilité  $P_r$  d'une de ces configurations.

Le nombre de manières de constituer  $r$  paquets avec  $k$  nombres est le nombre de Stirling

$$\left\{ \begin{matrix} k \\ r \end{matrix} \right\} = \left\{ \begin{matrix} k-1 \\ r-1 \end{matrix} \right\} + r \left\{ \begin{matrix} k-1 \\ r \end{matrix} \right\}$$

avec

$$\left\{ \begin{matrix} k \\ 1 \end{matrix} \right\} = \left\{ \begin{matrix} k \\ k \end{matrix} \right\} = 1$$

$k \backslash r$	1	2	3	4	5
1	1				
2	1	1			
3	1	3	1		
4	1	7	6	1	
5	1	15	25	10	1

- Jusque là, on distingue les objets mais pas les paquets (indiscernables). Il faut encore affecter les intervalles aux paquets.  
Par exemple, on a une paire, on vient de distinguer qu'elle était formée du dé  $n^{\circ}1$  et du dé  $n^{\circ}5$  mais on ne sait pas encore si c'est une paire d'as ou de valets.





On multiplie alors par le nombre de façons d'affecter les paquets aux intervalles et on divise par le nombre total de manières de répartir les  $k$  nombres dans les intervalles :

$$P_r = \frac{\left\{ \begin{matrix} k \\ r \end{matrix} \right\} d(d-1)\dots(d-r+1)}{d^k} \quad (r \leq d)$$

- On va donc générer  $n$   $k$ -tuples, chaque fois déterminer les nombres de cases différentes et faire un comptage  $\Rightarrow \chi^2$ . Au besoin ( $P_1$  est petit), on peut regrouper des classes.

## Test du collectionneur de coupons

- On a un dé à  $d = 6$  faces.
- On va lancer le dé tant qu'on n'a pas obtenu les  $d$  faces au moins une fois chacune.
- $S_r$  : probabilité de devoir lancer le dé  $r$  fois ( $r < d \rightarrow S_r = 0$ ).
  - Après  $r$  lancers, la probabilité d'avoir au moins  $d$  faces différentes vaut

$$p_r = \left\{ \begin{array}{c} r \\ d \end{array} \right\} \frac{d!}{d^r}$$

et la probabilité  $q_r$  de ne pas encore avoir toutes les faces vaut

$$q_r = \left( 1 - \left\{ \begin{array}{c} r \\ d \end{array} \right\} \frac{d!}{d^r} \right)$$

- Alors, si on note que l'événement "pas toutes les faces après  $r$  jets" est un sous-ensemble de "pas toutes les faces après  $r - 1$  jets"

$$S_r = q_{r-1} - q_r = \left\{ \begin{array}{c} r \\ d \end{array} \right\} \frac{d!}{d^r} - \left\{ \begin{array}{c} r-1 \\ d \end{array} \right\} \frac{d!}{d^{r-1}}$$

Et par la formule de récurrence donnée plus haut

$$S_r = \frac{d!}{d^r} \left( \left\{ \begin{matrix} r \\ d \end{matrix} \right\} - d \left\{ \begin{matrix} r-1 \\ d \end{matrix} \right\} \right) = \frac{d!}{d^r} \left\{ \begin{matrix} r-1 \\ d-1 \end{matrix} \right\}$$

- Si on s'arrête à un nombre de jets  $t > r$ , on peut alors former un dernier intervalle "pas toutes les faces après  $t$  jets" avec pour probabilité

$$1 - \left\{ \begin{matrix} t-1 \\ d \end{matrix} \right\} \frac{d!}{d^{t-1}}$$

- La comparaison avec les comptages conduit à un test de  $\chi^2$ .

## Test de permutation

- Pour un  $k$ -tuple  $X_1, X_2, \dots, X_k$ , on va regarder l'ordre relatif des nombres par rapport à une séquence triée (on suppose qu'on n'a pas deux fois le même nombre dans le  $k$ -tuple).
- On a  $k!$  permutations possibles et on s'attend à ce qu'elles soient également représentées.
- Exemple ( $k=3$ ):

$$\begin{pmatrix} 0.1 & 0.3 & 0.2 \\ 0.4 & 0.5 & 0.6 \\ 0.05 & 0.01 & 0.99 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 3 & 2 \\ 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} \quad \text{⏏}$$

- On veut un compteur pour chacune des  $k!$  permutations.
- A chaque permutation 1 5 6 3 4 2, on va faire correspondre un nombre  $0 \leq f \leq 6!$  (de façon générale  $k!$ )
- On numérote les positions de 0 à 5 (0 à  $k-1$ ):

$$\begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 5 & 6 & 3 & 4 & 2 \end{array}$$

0	1	2	3	4	5	
1	5	6	3	4	2	0
2	5	6	3	4	1	0
4	5	6	3	2	1	3
4	5	6	3	2	1	0
6	5	4	3	2	1	1
6	5	4	3	2	1	0

- On commence par noter la position du plus petit et on le met de côté.
- On l'échange avec le dernier et on recommence. Notons

$$0 \leq b_1 = 0 \leq k - 1$$

$$0 \leq b_2 = 0 \leq k - 2$$

$$0 \leq b_3 = 3 \leq k - 3$$

$$0 \leq b_4 = 0 \quad \dots$$

$$0 \leq b_5 = 1$$

$$0 \leq b_6 = 0$$

- A partir des  $b_i$ , on peut reconstruire la permutation d'origine.
- On a une base variable dans laquelle on peut représenter  $0 \leq f \leq k!$  :

$$b_1 \rightarrow \star 1$$

$$b_2 \rightarrow \star k$$

$$b_3 \rightarrow \star k(k-1)$$

$$b_4 \rightarrow \star k(k-1)(k-2)$$

...

$$b_k \rightarrow \star k! \quad b_k = 0$$

$$f = k! b_k + k(k-1) \dots 3 b_{k-1} + k(k-1) \dots 4 b_{k-2} + \dots + k(k-1) b_3 + k b_2 + b_1$$

- Comment calculer  $f$  intelligemment ?

- Méthode analogue à Horner pour les polynômes:  

$$6 \star (5 \star (4 \star (3 \star (2 \star b_6 + b_5) + b_4) + b_3) + b_2) + b_1$$
- On devrait avoir  $b_6$  d'abord au lieu de  $b_1$ , On va changer l'ordre des poids dans la base et calculer (on peut montrer que la méthode est équivalente):  $(((((b_1 \star 5 + b_2) \star 4 + b_3) \star 3 + b_4) \star 2 + b_5) \star 1) + b_6$
- Algorithme :
  - 1  $r \leftarrow k; f \leftarrow 0$
  - 2 Trouver le minimum des  $\{X_1, \dots, X_r\} \rightarrow X_s$   
 $f \leftarrow f \star r + s - 1$  (car les positions sont numérotées de 0 à  $k - 1$ )
  - 3  $X_r \leftrightarrow X_s$
  - 4  $r \leftarrow r - 1$
  - 5 Si  $r > 1$ , aller en (2)
- On peut évidemment utiliser le maximum au lieu du minimum.
- On génère  $n$   $k$ -tuples et pour chacune des  $k!$  permutations, on compare le compteur avec la valeur attendue  $\implies \chi^2$ .

## Méthode des “runs” (séries croissantes)

- 1 2 9 8 5 3 6 7 0 4
- On va étudier la fréquence des runs de chaque longueur (dans l'exemple 2 de 3, 1 de 2, 2 de 1).
- Problème : la fréquence des runs successifs n'est pas indépendante. Un run long a de bonnes chances d'être suivi d'un run plus court.

$$\chi^2 = \sum_{i=1}^{r_{\max}} \frac{(r_i - r_i^{th})^2}{\sigma^2(r_i^{th})}$$

n'est pas applicable et doit être généralisé par

$$\chi^2 = \sum_{i,j=1}^{r_{\max}} (r_i - r_i^{th})(r_j - r_j^{th}) a_{ij}$$

où  $(a_{ij})$  est l'inverse de la matrice de covariance.



- Soit  $Z_{pi} = 1$  si la position  $i$  est le début d'un run de longueur  $\geq p$ , 0 sinon.

- 1 2 9 8 5 3 6 7 0 4

$$1 = Z_{11} = Z_{21} = Z_{31} = Z_{14} = Z_{15} = Z_{16} = Z_{26} = Z_{36} = Z_{19} = Z_{29}$$

- Soit  $n$  la longueur totale de la séquence générée, le nombre de runs de longueur  $\geq p$  est donné par ( $Z_{pi} = 0$  pour  $i > n - p + 1$ )

$$R'_p = Z_{p1} + Z_{p2} + \dots + Z_{pn} = Z_{p1} + Z_{p2} + \dots + Z_{p,n-p+1}$$

- Le nombre de runs de longueur  $p$  est  $R_p = R'_p - R'_{p+1}$
- Les nombres moyens pour toutes les permutations possibles de la séquence sont  $\overline{R_p} = \frac{1}{n!} \sum_{(perm)} R_p^{(perm)}$  et  $\overline{R'_p} = \frac{1}{n!} \sum_{(perm)} R'^{(perm)}_p$ .

$$\overline{R'_p} = \frac{1}{n!} \sum_{(perm)} \left[ \sum_{i=1}^n Z_{pi}^{(perm)} \right] = \frac{1}{n!} \sum_{i=1}^n \{ \text{nombre de permutations pour lesquelles commence en } i \text{ un run de longueur } \geq p \}$$



- Considérons les  $p + 1$  nombres suivants

$$\begin{array}{ccccccc} & & & & & & \\ \cdot & > & = & < & < & < & ? \\ \cdot & i-1 & i & \cdot & \cdot & i+p-1 & \cdot \end{array}$$

- Cas ( $i \neq 1$ )

Pour  $1 \leq i \leq n - p + 1$ , on a

$$\{\dots\} = \frac{n!}{(p+1)!(n-p-1)!} p = \frac{n!p}{(p+1)!}$$

(l'ordre relatif dans le run est sans importance)

$(i - 1)$  ne peut pas être le plus petit des  $p + 1$ ,  $p$  possibilités)

Ce terme est indépendant de  $i$  et apparaît  $n - p$  fois

- Cas ( $i = 1$ )

$$\{\dots\} = \frac{n!}{p!}$$

- En rassemblant les 2 cas ( $i = 1, i \neq 1$ )

$$\overline{R'_p} = \frac{1}{n!} \left[ \frac{n!}{p!} + \frac{(n-p)p n!}{(p+1)!} \right] = \frac{(n+1)p}{(p+1)!} - \frac{p-1}{p!}$$

et

$$\overline{R_p} = \overline{R'_p} - \overline{R'_{p+1}}$$

- Le calcul de la covariance est un peu plus long mais suit le même principe

$$\text{cov}(R_p, R_q) = \overline{(R_p - \overline{R_p})(R_q - \overline{R_q})}$$

$$\text{cov}(R_p, R_q) = \text{cov}(R_p, R'_q) - \text{cov}(R_p, R'_{q+1})$$

$$\text{cov}(R_p, R'_q) = \text{cov}(R'_p, R'_q) - \text{cov}(R'_{p+1}, R'_q)$$

$$\text{cov}(R'_p, R'_q) = \frac{1}{n!} \sum_{(perm)} \left[ \sum_{i,j=1}^n Z_{pi}^{(perm)} Z_{qj}^{(perm)} \right] - \overline{R'_p} \overline{R'_q}$$

$$\text{cov}(R'_p, R'_q) = \frac{1}{n!} \sum_{i,j=1}^n \{ \text{nombre de permutations pour lesquelles commencent en } i \text{ un run de longueur } \geq p \text{ et en } j \text{ un de longueur } \geq q \} - \overline{R'_p} \overline{R'_q}$$

et après avoir distingué les différents cas où les runs sont contigus ( $j = i + p$ ) ou disjoints, on obtient :

$$\text{cov}(R'_p, R'_q) = \begin{cases} \overline{R'_t} + f(p, q, n) & (p + q \leq n) \\ \overline{R'_t} - \overline{R'_p} \overline{R'_q} & (p + q > n) \end{cases}$$

où  $t = \max(p, q)$ ,  $s = p + q$  et

$$\begin{aligned} f(p, q, n) &= (n+1) \left( \frac{s(1-pq) + pq}{(p+1)!(q+1)!} - \frac{2s}{(s+1)!} \right) + 2 \left( \frac{s-1}{s!} \right) \\ &+ \frac{(s^2 - s - 2)pq - s^2 - p^2q^2 + 1}{(p+1)!(q+1)!} \end{aligned}$$

- Si des classes sont trop peu remplies, on peut les grouper (par exemple, 1, 2, 3, 4, 5,  $\geq 6$ )
- Neutraliser le nombre qui suit un run ? (indépendance ?)

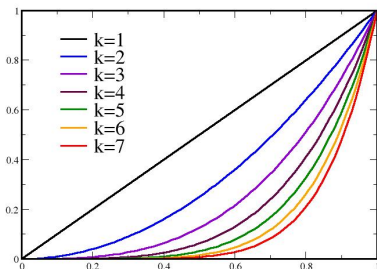
## Test du maximum

- Pour un  $k$ -tuple  $X_1, X_2, \dots, X_k$ , soit  $Y = \max(X_1, X_2, \dots, X_k)$

$$\{Y \leq y\} = \{x_1 \leq y\} \wedge \{x_2 \leq y\} \wedge \dots \wedge \{x_k \leq y\}$$

$$P\{Y \leq y\} = P\{x_1 \leq y\} P\{x_2 \leq y\} \dots P\{x_k \leq y\} = y^k$$

(loi uniforme)



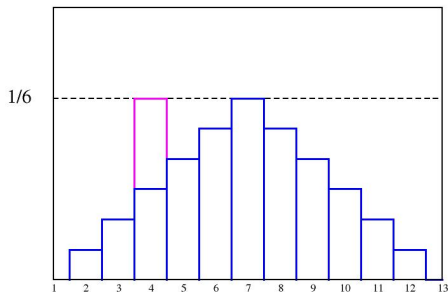
- Ceci permet de faire un test de Kolmogorov-Smirnov

# Génération de nombres aléatoires suivant des lois non uniformes

## A. Loi discrète

### ■ Exemple : jet de 2 dés

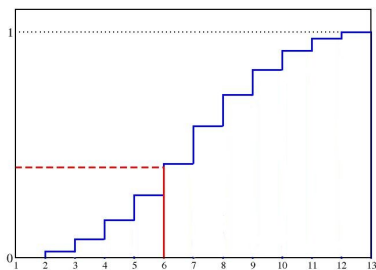
$$2 \rightarrow \frac{1}{36}, 3 \rightarrow \frac{2}{36}, 4 \rightarrow \frac{3}{36}, \dots, 12 \rightarrow \frac{1}{36}.$$



## Acceptance - Rejet

- On génère des couples de valeurs dans  $[2, 13[$  (abscisse) et  $[0, \frac{1}{6}[$  (ordonnée).  
 $x, y$  dans  $[0, 1[$   
 $x' = 2. + 11. \star x$   
 $y' = 0.166666667 \star y$
- Si un point est sous le diagramme, on prend la partie entière de son abscisse, sinon on le rejette.
- C'est le rapport des surfaces qui compte.
- Peu efficace si on a des variations brusques.

## Fonction de répartition



- Escalier à “11 marches”
- On génère  $y$  dans  $[0, 1[$  (ordonnée)
- La probabilité de se trouver en face d’une contre-marche est égale à la probabilité de l’abscisse correspondant.
- $P\{X \leq 2\} = \frac{1}{36}, P\{X \leq 3\} = \frac{3}{36}, P\{X \leq 4\} = \frac{6}{36}, P\{X \leq 5\} = \frac{10}{36}$   
 $P\{X \leq 6\} = \frac{15}{36}, P\{X \leq 7\} = \frac{21}{36}, \dots$
- Pas de rejet mais recherche dans une table (dichotomique).



## Méthode des aliases (Walker, 1974)

- On construit les deux tableaux  $P_j, Y_j$  suivants ( $j = 0, k-1; k = 16$ ) :

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$P_j$	0	0	$\frac{4}{9}$	$\frac{8}{9}$	1	$\frac{7}{9}$	1	1	1	$\frac{7}{9}$	$\frac{7}{9}$	$\frac{8}{9}$	$\frac{4}{9}$	0	0	0
$Y_j$	5	9	7	4	*	6	*	*	*	8	4	7	10	6	7	8

- On génère  $U$  dans  $[0, 1[$ ,  $K = \lfloor kU \rfloor$ ,  $V = kU - K$ .
- Soit  $N$  le nombre à générer

1  $j \leftarrow K$

2 Si  $V \leq P_j$  alors  $N \leftarrow j$ , sinon  $N \leftarrow Y_j$

- Si  $j = 0, 1, 13, 14, 15$ ,  $V \geq P_j = 0$ , 1 cas sur 16

- $N = 5, 9 \rightarrow \frac{1}{16} + \frac{1}{16} \frac{7}{9} = \frac{1}{9}$

- $N = 7 \rightarrow \frac{1}{16} \left( \frac{5}{9} + 1 + \frac{1}{9} + 1 \right) = \frac{1}{16} \frac{24}{9} = \frac{1}{6}$

- $N = 2, 12 \rightarrow \frac{1}{16} \frac{4}{9} = \frac{1}{36}$

- $N = 3 \rightarrow \frac{1}{16} \frac{8}{9} = \frac{2}{36}$

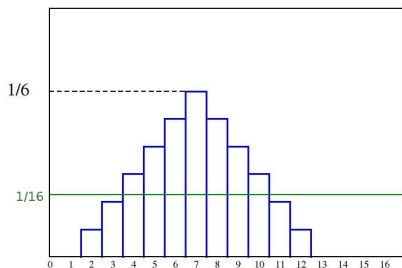
- $N = 4 \rightarrow \frac{1}{16} \left( \frac{1}{9} + 1 + \frac{2}{9} \right) = \frac{1}{16} \frac{12}{9} = \frac{3}{36}$

- $N = 6, 8 \rightarrow \frac{1}{16} \left( \frac{2}{9} + 1 + 1 \right) = \frac{1}{16} \frac{20}{9} = \frac{5}{36}$

- $N = 10 \rightarrow \frac{1}{16} \left( \frac{7}{9} + \frac{5}{9} \right) = \frac{1}{16} \frac{12}{9} = \frac{3}{36}$

- $N = 11 \rightarrow \frac{1}{16} \frac{8}{9} = \frac{2}{36}$

- Miracle ? On égalise la courbe en déplaçant ce qui est au-dessus en dessous.



- Par exemple, on est trop bas pour 3, on prend un morceau du 4.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$P_j$	0	0	$\frac{4}{9}$	$\frac{8}{9}$	1	$\frac{7}{9}$	1	1	1	$\frac{7}{9}$	$\frac{7}{9}$	$\frac{8}{9}$	$\frac{4}{9}$	0	0	0
$Y_j$	5	9	7	4	*	6	*	*	*	8	4	7	10	6	7	8

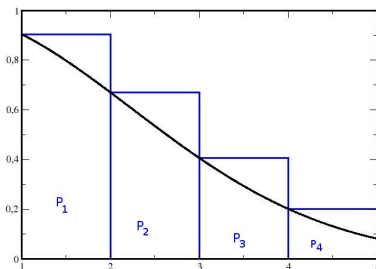
- On s'arrange pour n'avoir que 2 morceaux par colonne (2 tableaux).
- On part d'une colonne sous la barre et on égalise avec une colonne au-dessus, même si on prend de trop, et ainsi de suite.

# Génération de nombres aléatoires suivant des lois non uniformes

## B. Loi continue

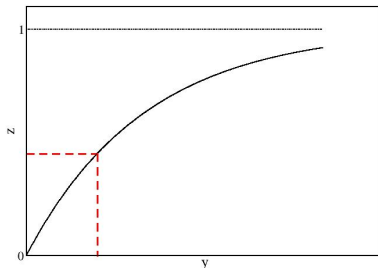
### Acceptance - Rejet

- Même principe que pour une loi discrète.
- Si la loi est définie sur  $(0, +\infty)$ , peu efficace pour la queue de la distribution.



- $P_1 + P_2 + P_3 + P_4 = 1$
- On génère un abscisse en fonction des  $P_i$  (on suppose connaître l'intégrale sur chaque intervalle), ce qui donne un intervalle.
- Dans cet intervalle, on génère des points jusqu'au moment où on en accepte un.
- A priori, les taux d'acceptance dépendent de l'intervalle.
- A n'utiliser que si on ne peut pas faire autrement.

## Inversion de la fonction de répartition



- $F(y) = \int_{-\infty}^y f(x) dx = \int_0^y f(x) dx$  est strictement croissante.
- On génère  $z$  dans  $[0, 1[$  et on a alors  $z = F(y) \implies y = F^{-1}(z)$ .  
En effet,  $F(y) = P\{Y \leq y\} = P\{Z \leq z\} = z$

- On peut encore le voir autrement. Soit  $F'_X(x) = f_X(x)$ ,  $H'_Y(y) = h_Y(y)$  et  $y = G(x)$  ( $G$  strictement monotone  $\rightarrow (x - G^{-1}(y))$  a 1 ! racine).

$$h_Y(y) = \int dx f_X(x) \delta(y - G(x)) = \int dx f_X(x) \frac{\delta(x - G^{-1}(y))}{|G'(x)|_{x=G^{-1}(y)}}$$

$$\text{Or } G(G^{-1}(y)) = y \implies G'(x)|_{x=G^{-1}(y)} \frac{d}{dy} G^{-1}(y) = 1$$

$$h_Y(y) = \int dx f_X(x) \delta(x - G^{-1}(y)) \frac{d}{dy} G^{-1}(y) = f_X(G^{-1}(y)) \frac{d}{dy} G^{-1}(y)$$

Soit maintenant  $f_X(x) = 1$  dans  $[0, 1[$ , alors  $h_Y(y) = (G^{-1})'(y)$  et  $H_Y(y) = G^{-1}(y)$ .

## Exemple : loi exponentielle

- $f(x) = \alpha e^{-\alpha x}$

$$F(y) = \int_0^y f(x) dx = -e^{-\alpha x} \Big|_0^y = 1 - e^{-\alpha y} = z$$

$$-\alpha y = \ln(1 - z) \implies y = -\frac{1}{\alpha} \ln(1 - z)$$

- On peut donc prendre  $y = -\frac{1}{\alpha} \ln z$  ( $z \neq 0$ ).

## Lien avec la loi de Poisson

- $P_k = \frac{\lambda^k}{k!} e^{-\lambda}$
- On veut étudier le nombre d'appels téléphoniques (ou le nombre de voitures qui passent dans la rue) dans un intervalle de temps  $]0, t]$ .
- On va faire les hypothèses suivantes :
  - 1 Le nombre d'appels dans  $]t, t + h]$  est indépendant de  $]0, t]$  et ne dépend que de  $h$ .
  - 2 La probabilité d'avoir un appel dans  $]t, t + h]$  est de la forme  $\alpha h + O(h)$ .
  - 3 La probabilité d'avoir plusieurs appels dans  $]t, t + h]$  est de la forme  $O(h)$  (rare).

## ■ En résumé

- 1 appel :  $\alpha h + O(h)$
- $\geq 2$  appels :  $O(h)$
- 0 appel :  $1 - \alpha h + O(h)$

- Soit  $P_n(t + h)$  la probabilité d'avoir  $n$  appels dans  $]0, t + h]$ , on peut la décomposer en

- $n - 1$  appels dans  $]0, t]$  et 1 appel dans  $]t, t + h]$  (1)
- $n$  appels dans  $]0, t]$  et 0 appel dans  $]t, t + h]$  (2)
- $k$  appels dans  $]0, t]$  et  $n - k \geq 2$  appels dans  $]t, t + h]$  (3)
- $P(1) = P_{n-1}(t) \cdot [\alpha h + O(h)]$
- $P(2) = P_n(t) \cdot [1 - \alpha h + O(h)]$
- $P(3) = P_k(t) \cdot O(h)$

$$P_n(t + h) = \alpha h P_{n-1}(t) + (1 - \alpha h) P_n(t) + O(h)$$

$$P_n(t + h) - P_n(t) = \alpha h (P_{n-1}(t) - P_n(t)) + O(h)$$

$$P'_n(t) = \alpha (P_{n-1}(t) - P_n(t))$$

- Pour  $n = 0$ ,  $P_0(t + h) - P_0(t) = -\alpha h P_0(t) + O(h)$  et  
 $P'_0(t) = -\alpha P_0(t) \implies P_0(t) = Ce^{-\alpha t}$



- On peut continuer pour  $n = 1, \dots$

$$P_1'(t) = \alpha (P_0(t) - P_1(t)) = \alpha C e^{-\alpha t} - \alpha P_1(t)$$

$$P_1(t) = C\alpha t e^{-\alpha t}$$

et par récurrence

$$P_n(t) = \frac{(\alpha t)^n}{n!} e^{-\alpha t}$$

$$C = \sum_0^{\infty} P_n(t) = 1$$

- Probabilité d'un intervalle de temps  $t$  entre deux appels (ou plus généralement temps d'attente, cf. hypothèse 1 !)

$$P(T > t) = P_0(t) = e^{-\alpha t}$$

$$F(t) = P(T \leq t) = 1 - e^{-\alpha t}$$

$$f(t) = F'(t) = \alpha e^{-\alpha t}$$

- On va générer des temps  $t_i$  suivant une loi exponentielle et on choisit  $t_\alpha$  de telle façon que  $\alpha t_\alpha = \lambda$ .
- On s'arrête lorsque  $\sum_{i=1}^n t_i > t_\alpha$ , alors  $k = n - 1$  est distribué suivant une loi de Poisson.

$$t_i = -\frac{1}{\alpha} \ln z_i$$

$$\sum_{i=1}^{n-1} t_i \leq t_\alpha < \sum_{i=1}^n t_i$$

$$-\frac{1}{\alpha} \sum_{i=1}^{n-1} \ln z_i \leq t_\alpha < -\frac{1}{\alpha} \sum_{i=1}^n \ln z_i$$

$$\sum_{i=1}^n \ln z_i < -\alpha t_\alpha \leq \sum_{i=1}^{n-1} \ln z_i$$

$$\prod_{i=1}^n z_i < e^{-\alpha t_\alpha} \leq \prod_{i=1}^{n-1} z_i$$

- Ceci revient à calculer  $\epsilon = e^{-\alpha t_\alpha}$  et à générer des  $z_i$  suivant une loi uniforme tant que  $\prod_{i=1}^n z_i > \epsilon$ .

- 1  $Z \leftarrow 1; N \leftarrow 0$
- 2  $Z \leftarrow Z \star z_i; N \leftarrow N + 1$
- 3 Si  $Z \geq \epsilon$ , aller en (2)
- 4  $k \leftarrow N - 1$



## Loi normale - méthode approchée

- Centrée, réduite :  $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ ,  $F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$
- $F(x)$  n'a pas d'expression analytique.
- $\left(\sum_{i=1}^{12} z_i\right) - 6$ , avec  $z_i$  uniformes dans  $[0, 1[$
- $\sigma^2(z) = \int_0^1 \left(x - \frac{1}{2}\right)^2 dx = \frac{1}{12}$

## Loi normale - méthode polaire

- 1  $U_1, U_2$  uniformes dans  $[0, 1[$ ,  $V_1 = 2U_1 - 1$  et  $V_2 = 2U_2 - 1$  uniformes sur  $[-1, +1[$ .
- 2  $S \leftarrow V_1^2 + V_2^2$
- 3 Si  $S > 1$ , aller en (1)
- 4  $X_1 \leftarrow V_1 \sqrt{-\frac{2 \ln S}{S}}, X_2 \leftarrow V_2 \sqrt{-\frac{2 \ln S}{S}}$

$(V_1, V_2)$  uniforme sur le carré.

$$S = R^2$$

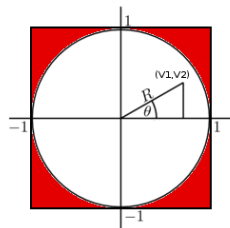
$$V_1 = R \cos \theta$$

$$V_2 = R \sin \theta$$

$$R' = \sqrt{-2 \ln S}$$

$$X_1 = \frac{V_1}{R} \sqrt{-2 \ln S} = R' \cos \theta$$

$$X_2 = \frac{V_2}{R} \sqrt{-2 \ln S} = R' \sin \theta$$



- Loi de probabilité de  $R'$

$$P(R' \leq r) = P(\sqrt{-2 \ln S} \leq r) = P(S \geq e^{-r^2/2}) = 1 - e^{-r^2/2}$$

- D'où la densité  $f_{R'}(r) = r e^{-r^2/2}$

$$\begin{aligned} P(X_1 \leq x_1 \wedge X_2 \leq x_2) &= \int_{r \cos \theta \leq x_1 \wedge r \sin \theta \leq x_2} d\theta dr \frac{1}{2\pi} r e^{-r^2/2} \\ &= \frac{1}{2\pi} \int_{X_1 \leq x_1 \wedge X_2 \leq x_2} dX_1 dX_2 e^{-\frac{X_1^2 + X_2^2}{2}} \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_1} dX_1 e^{-\frac{X_1^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_2} dX_2 e^{-\frac{X_2^2}{2}} \end{aligned}$$

- $X_1$  et  $X_2$  sont bien indépendantes et de lois  $N(0, 1)$ .

## Méthode paire-impair

- Soit

$$f(x) = \begin{cases} C e^{-h(x)} & \text{si } a \leq x \leq b \\ 0 & \text{sinon} \end{cases}$$

avec  $0 \leq h(x) \leq 1$  si  $a \leq x \leq b$ .

- 1  $U$  uniforme dans  $[0, 1[$ ,  $X = a + (b - a)U$  uniforme sur  $[a, b[$ .
- 2  $V_0 \leftarrow h(X)$ .
- 3 Générer  $V_1, V_2, V_3, \dots$  uniformes dans  $[0, 1[$  aussi longtemps que  $V_{i-1} \geq V_i$ .
- 4 Si  $i$  est pair, retourner en (1), sinon, accepter  $X$ .

$$P(i) = P\{(V_0 \geq V_1 \geq \dots \geq V_{i-1}) \wedge (V_{i-1} < V_i)\}$$

$$P\{V_0 \geq V_1 \geq \dots \geq V_{i-1}\} = \frac{[h(X)]^{i-1}}{(i-1)!}$$

- Les  $V_i$  sont indépendants et tous plus petits que  $V_0$ .
- Un seul ordre convient.

$$P(i) = \frac{[h(X)]^{i-1}}{(i-1)!} - \frac{[h(X)]^i}{i!}$$

- La probabilité de s'arrêter à une valeur de  $i$  impaire vaut

$$\begin{aligned} \sum_{i=1,3,5,\dots}^{\infty} \frac{[h(X)]^{i-1}}{(i-1)!} - \frac{[h(X)]^i}{i!} &= \sum_{j=0}^{\infty} (-1)^j \frac{[h(X)]^j}{j!} \\ &= \sum_{j=0}^{\infty} \frac{[-h(X)]^j}{j!} = e^{-h(X)} \end{aligned}$$

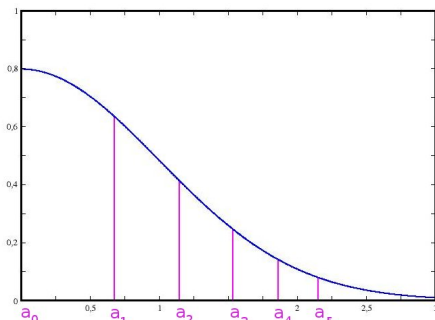
- Combien de nombres à générer en moyenne ?

$$\begin{aligned} \bar{k} = \sum_{k=1}^{\infty} k P(k) &= \sum_{k=1}^{\infty} k \left\{ \frac{[h(X)]^{k-1}}{(k-1)!} - \frac{[h(X)]^k}{k!} \right\} \\ &= \sum_{k=0}^{\infty} (k+1) \frac{[h(X)]^k}{k!} - \sum_{k=1}^{\infty} k \frac{[h(X)]^k}{k!} \\ &= \sum_{k=0}^{\infty} \frac{[h(X)]^k}{k!} + 0 = e^{h(X)} \leq e \end{aligned}$$

## Exemple : loi normale

- On va se contenter la partie droite de la distribution  $f(x) = \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}}$  (le signe peut être tiré au sort par la suite).
- Pour  $1 \leq j \leq n+1$  ("n + 1 bits de précision"), on a une table des  $d_j = a_j - a_{j-1}$  avec  $a_j$  qui est donné par

$$\int_{a_j}^{+\infty} f(x) dx = \frac{1}{2^j}$$





$j$	$d_j$	$a_j$	$j$	$d_j$	$a_j$
1	0.674489750	0.674489750	16	0.155349717	4.324919039
2	0.475859630	1.150349380	17	0.150409384	4.475328423
3	0.383771164	1.534120544	18	0.145902577	4.621231000
4	0.328611323	1.862731867	19	0.141770033	4.763001033
5	0.291142827	2.153874694	20	0.137963174	4.900964207
6	0.263684322	2.417559016	21	0.134441762	5.035405969
7	0.242508452	2.660067468	22	0.131172150	5.166578119
8	0.225567444	2.885634912	23	0.128125965	5.294704084
9	0.211634166	3.097269078	24	0.125279090	5.419983174
10	0.199924267	3.297193345	25	0.122610883	5.542594057
11	0.189910758	3.487104103	26	0.120103560	5.662697617
12	0.181225181	3.668329284	27	0.117741707	5.780439324
13	0.173601400	3.841930684	28	0.115511892	5.895951216
14	0.166841909	4.008772593	29	0.113402349	6.009353565
15	0.160796729	4.169569322	30	0.111402720	6.120756285

- Soit, pour  $a_j \leq t \leq a_{j+1}$ ,  $h(t) = \frac{t^2 - a_j^2}{2}$
- $h(t) = \frac{t^2 - a_j^2}{2}$  est-elle une bonne fonction pour la méthode paire-impair, i.e.  $h(t) \leq 1$  sur tout l'intervalle ?
- Il faut montrer que sur tout l'intervalle  $t^2 - a_j^2 \leq 2$  ou encore que  $a_{j+1}^2 - a_j^2 \leq 2$
- Considérons la fonction  $m(x) = e^{\frac{x^2}{2}} \int_x^{+\infty} e^{-\frac{t^2}{2}} dt$

$$\begin{aligned}
 \int_x^{+\infty} \frac{e^{-\frac{t^2}{2}}}{t^2} dt &= -\frac{1}{t} e^{-\frac{t^2}{2}} \Big|_x^{+\infty} - \int_x^{+\infty} \frac{-t e^{-\frac{t^2}{2}}}{-t} dt \\
 &= \frac{e^{-\frac{x^2}{2}}}{x} - \int_x^{+\infty} e^{-\frac{t^2}{2}} dt \\
 e^{\frac{x^2}{2}} \int_x^{+\infty} \frac{e^{-\frac{t^2}{2}}}{t^2} dt &= \frac{1}{x} - m(x) \\
 \text{Donc } m(x) &= \frac{1}{x} - e^{\frac{x^2}{2}} \int_x^{+\infty} \frac{e^{-\frac{t^2}{2}}}{t^2} dt < \frac{1}{x}
 \end{aligned}$$

- $m(x)$  est une fonction décroissante :

$$m'(x) = x m(x) + e^{\frac{x^2}{2}} \left[ -e^{-\frac{t^2}{2}} \right]_{-\infty}^x = x m(x) - 1 < 0$$

- Soit alors  $y = \sqrt{x^2 + 2 \ln 2} > x$ .

$$\sqrt{\frac{2}{\pi}} \int_y^{+\infty} e^{-\frac{t^2}{2}} dt = \sqrt{\frac{2}{\pi}} e^{-\frac{y^2}{2}} m(y) = \frac{1}{2} \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}} m(y) < \frac{1}{2} \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}} m(x)$$

- Posons  $x = a_j$ .

$$\sqrt{\frac{2}{\pi}} \int_y^{+\infty} e^{-\frac{t^2}{2}} dt < \frac{1}{2} \sqrt{\frac{2}{\pi}} e^{-\frac{a_j^2}{2}} m(a_j) = \frac{1}{2} \frac{1}{2^j} = \sqrt{\frac{2}{\pi}} \int_{a_{j+1}}^{+\infty} e^{-\frac{t^2}{2}} dt$$

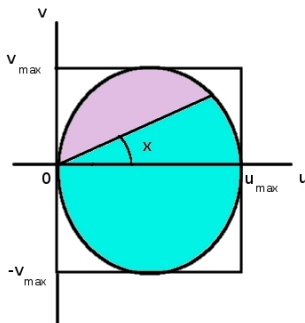
- Finalement  $y > a_{j+1} \implies a_j^2 + 2 \ln 2 > a_{j+1}^2 \implies a_{j+1}^2 - a_j^2 < 2 \ln 2 < 2$ .

## Algorithme

- 1 Générer un nombre aléatoire suivant la loi uniforme  
 $U = (b_0, b_1, b_2, \dots, b_n)$  (en binaire).
- 2  $B \leftarrow b_0, j \leftarrow 1, a \leftarrow 0$ .
- 3 Si  $b_j = 1, a \leftarrow a + d_j, j \leftarrow j + 1$ . Si  $j < n + 1$ , répéter (3).
- 4 (On arrive en  $j$  avec une probabilité  $2^{-j}$ ).  
On va générer  $x$  dans  $[a_{j-1}, a_j[$  avec  
 $h(x) = x^2 - a^2 = y^2 + a y \quad (y = x - a)$ .
- 5 Générer  $Y$  dans  $[0, d_j[, V \leftarrow (\frac{1}{2}Y + a)Y$   
(On peut prendre  $Y = d_j.(b_{j+1}, \dots, b_n)$ )
- 6 Générer  $U$  dans  $[0, 1[$ . Si  $U > V$ , aller en (7), sinon (on continue la séquence) régénérer  $V$  dans  $[0, 1[$ .  
Si  $U < V$  (cas pair), aller en (5), sinon répéter (6).
- 7  $X \leftarrow a + Y$ . Si  $B = 1, X \leftarrow -X$ .

## Méthode de Kinderman et Monahan (1976)

- On se donne le rectangle suivant :  $0 \leq u \leq u_{\max}$ ,  $-v_{\max} \leq v \leq v_{\max}$ .
- Soit  $x$  le coefficient angulaire du segment  $(0,0) \rightarrow (u, v)$ .
- A l'intérieur du rectangle, on définit une courbe telle qu'on rejette les points qui tombent en dehors.
- On veut que les points acceptés soient tels que  $x$  est distribué suivant la loi  $f(x)$ .



$\int_{-\infty}^x f(t)dt = \text{rapport des surfaces (surface turquoise/surface colorée)}.$

## A démontrer

La courbe voulue est définie par  $u^2 = f(\frac{v}{u})$  ou  $u = \sqrt{f(\frac{v}{u})}$ .

- Dénominateur :  $u > 0$ ,  $u^2 \leq f(\frac{v}{u})$ .
- Soit  $v = tu$ ,  $dv = udt$ ,

$$\begin{aligned} \int_{\substack{u > 0 \\ u^2 \leq f(t)}} du \, dv &= \int_{\dots} u \, du \, dt = \int_{-\infty}^{+\infty} dt \int_0^{\sqrt{f(t)}} u \, du \\ &= \int_{-\infty}^{+\infty} dt \left[ \frac{u^2}{2} \right]_0^{\sqrt{f(t)}} = \int_{-\infty}^{+\infty} dt \frac{f(t)}{2} = \frac{1}{2} \end{aligned}$$

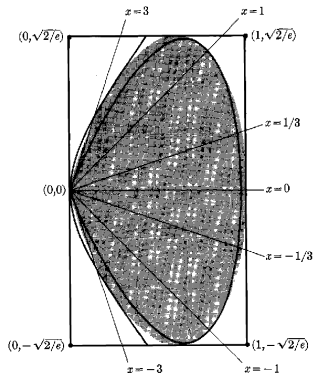
- Numérateur :  $u > 0$ ,  $u^2 \leq f(\frac{v}{u})$ ,  $\frac{v}{u} \leq x$ .

$$\int_{\substack{u > 0 \\ u^2 \leq f(t) \\ t \leq x}} du \, dv = \int_{-\infty}^x dt \int_0^{\sqrt{f(t)}} u \, du = \frac{1}{2} \int_{-\infty}^x f(t) dt$$

CQFD (NB: Encore valable même si  $f$  n'est pas normalisée).

## Exemple : loi normale

$$f(t) = e^{-\frac{t^2}{2}}, \quad \frac{v^2}{u^2} \leq -4 \ln u.$$



Soit  $u_{max} = 1$ ,

$$2v \frac{\partial v}{\partial u} = -8u \ln u - 4 \frac{u^2}{u} = 0$$

$$\ln u = -\frac{1}{2}$$

$$u = e^{-\frac{1}{2}}$$

$$v^2 = -4u^2 \ln u = \frac{2}{e}$$

$$v_{max} = \sqrt{\frac{2}{e}}$$

- Ceci nécessite chaque fois le calcul d'un logarithme.
- On va encadrer la courbe avec des courbes plus simples et on ne devra faire le calcul que si on génère un point entre les deux.

$$\begin{aligned}
 \forall x, e^x &\geq 1+x \\
 \forall c, e^{-1+cu} \geq cu &\Rightarrow \ln c + \ln u \leq cu - 1 \\
 &\Rightarrow \ln c - cu + 1 \leq -\ln u \\
 \forall c, e^{-1+\frac{1}{cu}} \geq \frac{1}{cu} &\Rightarrow -1 + \frac{1}{cu} \geq -\ln c - \ln u \\
 &\Rightarrow -\ln u \leq \ln c + \frac{1}{cu} - 1 \\
 \ln c - cu + 1 \leq -\ln u &\leq \ln c + \frac{1}{cu} - 1
 \end{aligned}$$

On peut donc

- 1 Accepter systématiquement  $(u, v)$  si  $\frac{1}{4}\left(\frac{v}{u}\right)^2 \leq \ln c - cu + 1$
- 2 Refuser systématiquement  $(u, v)$  si  $\frac{1}{4}\left(\frac{v}{u}\right)^2 \geq \ln c' + \frac{1}{c'u} - 1$



- Il reste à déterminer  $c$  pour avoir une surface la plus grande possible et  $c'$  ... la plus petite possible.

### Cas (1)

$$\left(\frac{v}{u}\right)^2 \leq 4(1 + \ln c) - 4cu$$

$$|v| \leq |u| \sqrt{4(1 + \ln c) - 4cu} = u\sqrt{a - bu}$$

avec  $b = 4c$  et  $a = 4(1 + \ln c)$ .

- Quand  $v = 0$  (les bornes de  $u$  sont atteintes),  $u = \frac{a}{b}$  ou 0
- Il faut donc maximiser  $R = 2 \int_0^{\frac{a}{b}} du \int_0^{u\sqrt{a-bu}} dv = 2 \int_0^{\frac{a}{b}} u\sqrt{a-bu} du$   
Soit  $t = \sqrt{a-bu}$ ,  $u = \frac{a-t^2}{b}$ ,  $-\frac{1}{2}b(a-bu)^{-\frac{1}{2}} du = dt \rightarrow du = -\frac{2}{b}t dt$ .

$$R = -\frac{4}{b^2} \int_{\sqrt{a}}^0 t^2(a-t^2)dt = -\frac{4a}{b^2} \left[ \frac{t^3}{3} \right]_{\sqrt{a}}^0 + \frac{4}{b^2} \left[ \frac{t^5}{5} \right]_{\sqrt{a}}^0$$

$$R = \frac{4a^{\frac{5}{2}}}{3b^2} - \frac{4a^{\frac{5}{2}}}{5b^2} = \frac{8a^{\frac{5}{2}}}{15b^2} = \dots \frac{(1 + \ln c)^{\frac{5}{2}}}{c^2} \quad \text{à maximiser.}$$

- On dérive par rapport à  $c$  :

$$R' = \dots \frac{5(1 + \ln c)^{\frac{3}{2}}}{2c^3} - \frac{2(1 + \ln c)^{\frac{5}{2}}}{c^3} = \frac{(1 + \ln c)^{\frac{3}{2}}}{c^3} \left[ \frac{5}{2} - 2(1 + \ln c) \right]$$

- Comme  $(1 + \ln c)$  n'est pas un maximum ( $R = 0$ ), il reste le second facteur et  $(1 + \ln c) = \frac{5}{4} \implies \ln c = \frac{1}{4}$  et  $c = e^{\frac{1}{4}}$ .
- Acceptance inconditionnelle :  $(\frac{v}{u})^2 \leq 5 - 4u e^{\frac{1}{4}}$

## Cas (2)

- La même intégration ne peut pas être faite de façon analytique. Des tests ont montré que la meilleure valeur de  $c'$  est approximativement  $e^{1.35}$
- Rejet inconditionnel :  $(\frac{v}{u})^2 \geq \frac{4 e^{-1.35}}{u} + 1.4$

## Rectangles et coins (Marsaglia, 1964)

“Rectangle-wedge-tail method” (Loi normale)

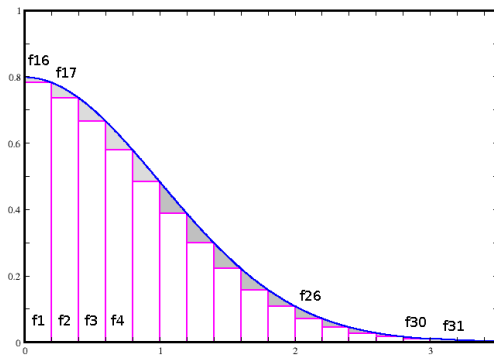
- Comme pour la méthode “paire-impair” appliquée à la loi normale (somme de fonctions sur des intervalles donnés par les  $a_j$ ), on va considérer la distribution voulue comme une somme pondérée de fonctions

$$f(x) = \sum_i p_i f_i(x) \quad \text{avec} \quad \sum_i p_i = 1$$

- On va encore se contenter la partie droite de la distribution

$$f(x) = \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}}$$

- Les  $f_i$  “difficiles” auront une probabilité  $p_i$  faible.
- Les  $\{f_i, i = 1, 15\}$  sont des rectangles  $\rightarrow$  distributions uniformes.

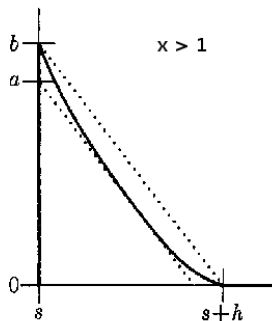
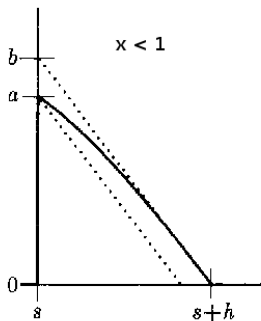


$$p_j = \frac{1}{5} f\left(\frac{j}{5}\right) = \sqrt{\frac{2}{25\pi}} e^{-\frac{j^2}{50}} \quad \text{pour } 1 \leq j \leq 15, \quad \sum_{i=1}^{15} p_i = 0.9183$$

- On peut générer  $U$  suivant une loi uniforme sur  $[0, 1[$  et prendre

$$X = \frac{1}{5} U + S \quad \text{avec} \quad S = \frac{j-1}{5}$$

- Les  $\{f_i, i = 16, 30\}$  sont des "coins".

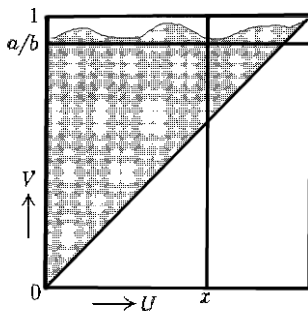


- On va entourer la courbe par 2 droites
  - Pour  $x > 1$  ( $i = 21, 30$ , concavité vers le haut), on joint les deux points extrêmes et on prend la tangente parallèle.
  - Pour  $x < 1$ , ( $i = 16, 20$ , concavité vers le bas), on prend la tangente au point de droite et sa parallèle qui passe par l'autre point.
- On a alors

$$a - b \frac{(x-s)}{h} \leq f(x) \leq b - b \frac{(x-s)}{h}$$

$$\frac{a}{b} \leq f^*(x) = \frac{1}{b}f(x) + \frac{(x-s)}{h} \leq 1$$

- On peut générer  $U$  et  $V$  suivant une loi uniforme sur  $[0, 1[$  avec  $V > U$  (sinon, on les échange).
- Si  $V \leq \frac{a}{b}$ , on accepte  $U$ , sinon, on teste si  $V > U + \frac{1}{b}f(s + hU)$ . Si c'est le cas, on accepte  $U$ , sinon, on recommence. A la fin,  $X \leftarrow s + hU$ .



Vérification : Probabilité d'avoir  $X \leq s + hx$  avec  $0 \leq x \leq 1$  (rapport de la surface à gauche de  $U = x$  avec la surface totale) :

$$\frac{\int_0^x \frac{1}{b} f(s + hu) du}{\int_0^1 \frac{1}{b} f(s + hu) du} = \frac{\int_s^{s+hx} f(t) dt}{\int_s^{s+h} f(t) dt} = \int_s^{s+hx} f(t) dt$$

- $f_{31}$  est la queue de la distribution (environ une fois sur 370).

## Cas général

- Soit  $f(x)$  une densité de probabilité suivant laquelle on génère des nombres aléatoires. Soit  $g(x)$  une densité de probabilité suivant laquelle il est facile de générer des nombres aléatoires.  
Soit  $f(x) \leq c g(x)$  avec  $c$  le plus petit possible ( $c \geq 1$ ).
- L'algorithme suivant va générer des nombres selon  $f(x)$ :
  - 1 Générer  $X$  selon  $g(x)$ .
  - 2 Générer  $U$  uniforme dans  $[0, 1[$ . Si  $\frac{f(X)}{c g(X)} > U$ , accepter  $X$ , sinon retourner en (1).
- Probabilité de rejeter  $X$  quel qu'il soit :

$$\int_{-\infty}^{+\infty} dx g(x) \left[ 1 - \frac{f(x)}{c g(x)} \right] = 1 - \frac{1}{c}$$



- Probabilité d'accepter  $X$  quel qu'il soit :  $\frac{1}{c}$
- Probabilité d'accepter  $X$  généré suivant  $g(x)$  (2) :  $\frac{f(x)}{c g(x)}$  (probabilité conditionnelle).
- Densité de probabilité de  $X$  :

$$\frac{g(X) \cdot \frac{f(X)}{c g(X)}}{\frac{1}{c}} = f(x)$$

- Probabilité de générer  $X$  (densité)
- Probabilité d'accepter en (2)

Soient

$$f(t) = \frac{e^{-\frac{t^2}{2}}}{\int_3^{+\infty} e^{-\frac{u^2}{2}} du} \quad \text{et} \quad g(t) = a t e^{-\frac{t^2}{2}}$$

On doit déterminer  $a$  et  $c$ .

$$\int_3^{+\infty} g(t) dt = 1 = a \int_3^{+\infty} t e^{-\frac{t^2}{2}} dt = \frac{a}{2} \int_9^{+\infty} e^{-\frac{t^2}{2}} dt^2 = a \left[ -e^{-\frac{x}{2}} \right]_9^{+\infty}$$

- D'où on tire :  $a = e^{\frac{9}{2}}$ ,  $g(t) = t e^{\frac{9-t^2}{2}}$ .
- Fonction de répartition  $G(x)$  de  $g(x)$ :

$$\begin{aligned} G(x) = \int_3^x t e^{\frac{9-t^2}{2}} dt &= \frac{1}{2} \int_9^{x^2} e^{\frac{9-t^2}{2}} dt^2 = \frac{1}{2} e^{\frac{9}{2}} \int_9^{x^2} e^{-\frac{t}{2}} dt \\ &= e^{\frac{9}{2}} \left[ -e^{-\frac{x^2}{2}} + e^{-\frac{9}{2}} \right] = 1 - e^{\frac{9-x^2}{2}} = y \end{aligned}$$

- On génère alors  $z = 1 - y = e^{\frac{9-x^2}{2}}$  dans  $[0, 1[$  et on trouve  $x = \sqrt{9 - 2 \ln z}$ .

$$\begin{aligned} \frac{f(t)}{c g(t)} &= \frac{e^{-\frac{t^2}{2}}}{c \left[ \int_3^{+\infty} e^{-\frac{u^2}{2}} du \right] t e^{\frac{9-t^2}{2}}} = \frac{e^{-\frac{9}{2}}}{c t \int_3^{+\infty} e^{-\frac{u^2}{2}} du} \leq 1 \\ c &\geq \frac{e^{-\frac{9}{2}}}{t \int_3^{+\infty} e^{-\frac{u^2}{2}} du} \Rightarrow \text{On peut prendre } c = \frac{e^{-\frac{9}{2}}}{3 \int_3^{+\infty} e^{-\frac{u^2}{2}} du} \end{aligned}$$

On accepte alors  $X$  si  $U < \frac{f(X)}{c g(X)} = \frac{3}{X}$

## Loi exponentielle - minimisation aléatoire

- On a déjà vu une méthode (inversion de la fonction de répartition), mais on voudrait éviter le logarithme.
- Soit  $Q[k] = \frac{\ln 2}{1!} + \frac{(\ln 2)^2}{2!} + \frac{(\ln 2)^3}{3!} + \dots + \frac{(\ln 2)^k}{k!}$  ( $k \geq 1$ ).  
( $Q[k] \rightarrow 1$  quand  $k \rightarrow +\infty$ )
- En pratique, on s'arrête quand  $Q[k] > 1 - 2^{1-n}$  pour des mots sur  $n$  bits.

### Algorithme

- 1 Générer un nombre aléatoire suivant la loi uniforme  
 $V = (b_0, b_1, b_2, \dots, b_n)$  (en binaire). Soit  $j$  le premier bit dans l'état 1  
( $P\{j = J\} = \frac{1}{2^{J+1}}$ ). Soit alors  $U = (b_{j+1}, \dots, b_n)$ .
- 2 Si  $U < \ln 2$ , alors  $X \leftarrow \mu(j \ln 2 + U)$  et on s'arrête là.
- 3 Trouver la plus petite valeur de  $k$  telle que  $Q[k] > U$ , générer  $k$  nombres aléatoires  $U_1, U_2, \dots, U_k$ .  $V \leftarrow \min(U_1, U_2, \dots, U_k)$ .
- 4  $X \leftarrow \mu(j + V) \ln 2$

■ Calcul de la fonction de répartition de  $X$

$$P\{X \leq X_0\} = (\ln 2)P\{\mu(j \ln 2 + U) \leq X_0\} \quad (2)$$

$$+ \sum_{k=2}^{+\infty} \frac{(\ln 2)^k}{k!} P\{\mu(j + V) \ln 2 \leq X_0\} \quad (4)$$

$$= (\ln 2)P\left\{j + \frac{U}{\ln 2} \leq \frac{X_0}{\mu \ln 2}\right\} \\ + \sum_{k=2}^{+\infty} \frac{(\ln 2)^k}{k!} P\left\{(j + V) \leq \frac{X_0}{\mu \ln 2}\right\}$$

■ Soit  $\bar{X} = \left\lceil \frac{X_0}{\mu \ln 2} \right\rceil$ . Dans (2),  $\frac{U}{\ln 2} < 1$ . Si  $j \leq \bar{X} + \left(\frac{X_0}{\mu \ln 2} - \bar{X}\right) - \frac{U}{\ln 2} \Rightarrow j \leq \bar{X}$

$$P\left\{j + \frac{U}{\ln 2} \leq \frac{X_0}{\mu \ln 2}\right\} = P\{j < \bar{X}\} \\ + P\{j = \bar{X}\} P\left\{\frac{U}{\ln 2} \leq \left(\frac{X_0}{\mu \ln 2} - \bar{X}\right)\right\}$$

$$\begin{aligned}
 P \left\{ j + \frac{U}{\ln 2} \leq \frac{X_0}{\mu \ln 2} \right\} &= \left( \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \dots + \frac{1}{2^{\bar{X}}} \right) \\
 &+ \frac{1}{2^{\bar{X}+1}} \left( \frac{X_0}{\mu \ln 2} - \bar{X} \right) \\
 &= \frac{\frac{1}{2^{\bar{X}+1}} - \frac{1}{2}}{\frac{1}{2} - 1} + \frac{1}{2^{\bar{X}+1}} \left( \frac{X_0}{\mu \ln 2} - \bar{X} \right) \\
 &= \left( 1 - \frac{1}{2^{\bar{X}}} \right) + \frac{1}{2^{\bar{X}+1}} \left( \frac{X_0}{\mu \ln 2} - \bar{X} \right)
 \end{aligned}$$

■ De la même manière,

$$\begin{aligned}
 P \left\{ (j + V) \leq \frac{X_0}{\mu \ln 2} \right\} &= P \{ j < \bar{X} \} \\
 &+ P \{ j = \bar{X} \} P \left\{ V \leq \left( \frac{X_0}{\mu \ln 2} - \bar{X} \right) \right\}
 \end{aligned}$$

$$P\left\{(j+V) \leq \frac{X_0}{\mu \ln 2}\right\} = \left(1 - \frac{1}{2^{\bar{X}}}\right) + \frac{1}{2^{\bar{X}+1}} P\left\{V \leq \frac{X_0}{\mu \ln 2} - \bar{X}\right\}$$

$$\begin{aligned} P\left\{V \leq \frac{X_0}{\mu \ln 2} - \bar{X}\right\} &= P\left\{\min(U_1, U_2, \dots, U_k) \leq \frac{X_0}{\mu \ln 2} - \bar{X}\right\} \\ &= 1 - P\left\{\min(U_1, U_2, \dots, U_k) > \frac{X_0}{\mu \ln 2} - \bar{X}\right\} \\ &= 1 - \prod_{i=1}^k P\left\{U_i > \frac{X_0}{\mu \ln 2} - \bar{X}\right\} \\ &= 1 - \left[1 - \frac{X_0}{\mu \ln 2} + \bar{X}\right]^k \end{aligned}$$

$$\begin{aligned}
P\{X \leq X_0\} &= \ln 2 \left(1 - \frac{1}{2^{\bar{X}}}\right) + \left(1 - \frac{1}{2^{\bar{X}}}\right) \sum_{k=2}^{+\infty} \frac{(\ln 2)^k}{k!} \\
&+ \frac{1}{2^{\bar{X}+1}} \left\{ \left( \frac{X_0}{\mu \ln 2} - \bar{X} \right) \ln 2 + \sum_{k=2}^{+\infty} \frac{(\ln 2)^k}{k!} \right. \\
&- \left. \sum_{k=2}^{+\infty} \frac{\left[ \ln 2 (1 + \bar{X}) - \frac{X_0}{\mu} \right]^k}{k!} \right\} \\
&= \left(1 - \frac{1}{2^{\bar{X}}}\right) + \frac{1}{2^{\bar{X}+1}} \left\{ \frac{X_0}{\mu} - \bar{X} \ln 2 + 1 - \ln 2 \right. \\
&- \left. e^{\left[ \ln 2 (1 + \bar{X}) - \frac{X_0}{\mu} \right]} + 1 + \ln 2 (1 + \bar{X}) - \frac{X_0}{\mu} \right\} \\
&= \left(1 - \frac{1}{2^{\bar{X}}}\right) + \frac{1}{2^{\bar{X}+1}} \left\{ 2 - 2^{1+\bar{X}} e^{-\frac{X_0}{\mu}} \right\} = 1 - e^{-\frac{X_0}{\mu}}
\end{aligned}$$

# Divers

- Autres lois continues
- Permutations
- Echantillonnage
- Simulations par événements discrets (DES)



## Autres lois continues

- Nous avons vu des méthodes pour générer des nombres suivant une loi uniforme, non uniforme et discrète ou continue (normale, exponentielle). Il y en a évidemment d'autres.

$$F(x) = \frac{1}{\Gamma(a)} \int_0^x t^{a-1} e^{-t} dt \quad (x \geq 0) \quad ; \quad \Gamma(a) = \int_0^{+\infty} t^{a-1} e^{-t} dt$$
$$\Gamma(n) = (n-1)!$$

- Cas particuliers :  $a = 1$  (exponentielle),  $a = \frac{1}{2}$  (2 fois un  $\chi_1^2$ ) et  $a = k$  correspond à une somme de  $k$  exponentielles indépendantes de moyenne 1.
- Si  $k$  est grand, générer des exponentielles est lourd mais il y a des méthodes plus efficaces (cfr. référence).

## Permutations

- Soient  $X_1, X_2, \dots, X_n$  un ensemble d'éléments à mélanger. On pourrait s'inspirer du test de permutation et générer  $n$  valeurs dans  $[0, 1[$  et donner à chaque  $X_i$  la position de la  $i^{\text{ème}}$  valeur par ordre croissant ou décroissant dans la séquence générée.

Algorithme d'une autre méthode (plus simple)

- 1  $j \leftarrow n$ .
- 2 Générer  $U$  dans  $[0, 1[$ .
- 3  $k \leftarrow \lfloor jU + 1 \rfloor$ ,  $X_k \leftrightarrow X_j$ .
- 4  $j \leftarrow j - 1$ . Si  $j > 1$ , aller en (2).

## Echantillonnage

- On veut extraire  $n$  éléments d'un fichier qui en contient  $N$  en le parcourant une seule fois. Chaque élément a une probabilité  $\frac{n}{N}$  d'être pris.
- Solution : le  $(t + 1)^{\text{ème}}$  élément est choisi avec la probabilité  $\frac{n-m}{N-t}$  si on a déjà  $m$  éléments.  
Sur toutes les manières de choisir  $n$  parmi  $N$  éléments avec  $m$  parmi les  $t$  premiers, il y en a

$$\begin{aligned} & \binom{N-t-1}{n-m-1} / \binom{N-t}{n-m} \\ = & \frac{(N-t-1)!(N-t-n+m)!(n-m)!}{(N-t)!(N-t-1-n+m+1)!(n-m-1)!} \\ = & \frac{n-m}{N-t} \end{aligned}$$

avec le  $(t + 1)^{\text{ème}}$  élément choisi.

## Simulations par événements discrets (DES)

- Système à événements discrets = Système dont l'état change de façon discontinue en un certain nombre d'instants qui peuvent être aléatoires.
  - Guichets d'un bureau de poste, chaînes de montage
  - Pas le système circulatoire sanguin du postier
- La DES traite de la modélisation de ces systèmes.
- Dans la simulation (en fonction des objectifs), certains éléments seront pris en compte, d'autres pas. On rencontre des éléments de types suivants:
  - Des clients (clients du bureau de poste, automobiles)
  - Des ressources (éléments du système qui offrent un service, les postiers, les ouvriers)
  - Des éléments de contrôle (heure d'ouverture, pause de midi)
  - Des opérations par ou sur les entités (traitement d'un client par un postier, peinture des portes)
  - Des unités de stockage (les files d'attente, la réserve de rétroviseurs)
  - Des unités de transport (tapis roulant)
  - ...

- Différentes approches existent, par
  - activité (intervalle de temps pendant lequel l'état d'une ressource ne change pas)
  - événement (instant précis de l'état de changement d'une ressource)
  - processus (succession d'un nombre fini d'états d'une ressource)
- Gestion des événements
  - Dans l'ordre chronologique
  - Horloge propre par entité ou ordonnanceur (échéancier)
  - Liste d'événements à simuler
  - Lois de probabilité (arrivées de clients poissonniennes ? temps de traitement par le postier gaussien ?)
- Gestion des files d'attente
  - Probablement FIFO pour les clients
  - Peut-être LIFO pour les pièces d'automobile
- Utilisation des résultats
  - Régime stationnaire - régime transitoire (e.g. réaction à une perturbation)

- Ordonnancement - structures de données
  - Files de priorité
  - Représentation par tas
  - Temps de recherche, d'insertion et de suppression courts
    - Arbres binaires
    - Arbres binaires "équilibrés"
    - Arbres binaires "gauchers" (leftist binary trees/leftist heaps)