MAT4681 - Statistique pour les sciences

Arthur Charpentier

07 - Générer du hasard

été 2022

Générer du hasard I

A Million Random Digits with 100,000 Normal Deviates, 1955

★☆☆☆☆ **Errors Throughout**
Reviewed in the United States on December 30, 2013

They sure don't come up with random numbers like they used to. If you look closely, you will note that every tenth digit or so is just a repeat of the last digit and every hundredth or so is a just the same digit repeated three times. How sloppy!

A sampling of this "work":

Page 36 - Line 6 - 15 characters in should be 5, not 4.
Page 99 - Line 18 - first three characters should be "453" not "345".
Page 145 - Line 2 - 7th and 19th characters transposed.
Page 190 - Whole line of numbers omitted betwen 6th and 7th lines.
Pages 210 and 211 - Two sections appear quasi-randomized, instead of randomized.

★★★☆☆ **Predictable**
Reviewed in the United States on October 30, 2013

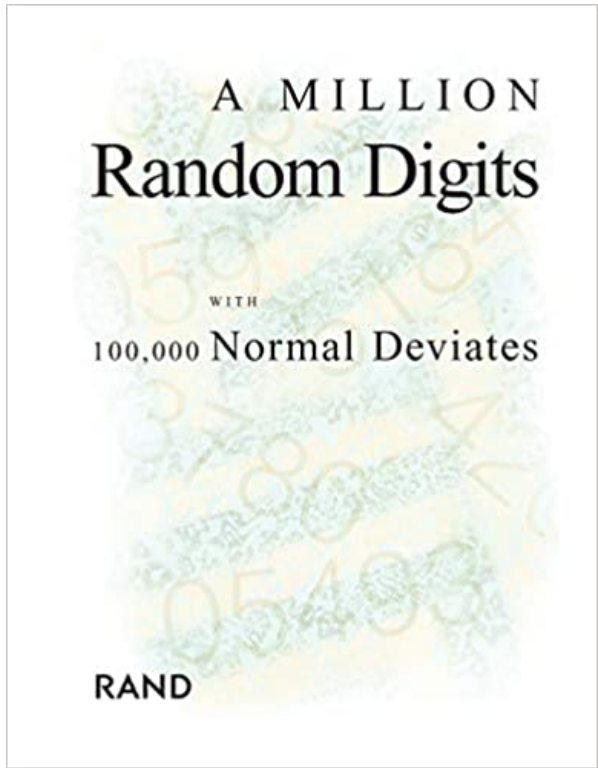
It seemed like about 10% of the time I was able to predict which number was next. It was still better than Life of Pi which, aside from being irrational, included no estimations of Pi at all.

★☆☆☆☆ **Not really random**
Reviewed in the United States on September 26, 2012

I bought two copies of this book. I find that the first copy perfectly predicts what the numbers will be in the second copy. I feel cheated.

★★★★★ **great yet lacking....**
Reviewed in the United States on May 1, 2013

So far, it's great, but what I REALLY want is an audio version that I can listen to at the beach or the gym. Maybe narrated by Morgan Freeman???



Générer du hasard II

A Million Random Digits with 100,000 Normal Deviates, 1955

★☆☆☆☆ **Errors Throughout**
Reviewed in the United States on December 30, 2013

They sure don't come up with random numbers like they used to. If you look closely, you will note that every tenth digit or so is just a repeat of the last digit and every hundredth or so is a just the same digit repeated three times. How sloppy!

A sampling of this "work":

Page 36 - Line 6 - 15 characters in should be 5, not 4.
Page 99 - Line 18 - first three characters should be "453" not "345".
Page 145 - Line 2 - 7th and 19th characters transposed.
Page 190 - Whole line of numbers omitted betwen 6th and 7th lines.
Pages 210 and 211 - Two sections appear quasi-randomized, instead of randomized.

★★★☆☆ **Predictable**
Reviewed in the United States on October 30, 2013

It seemed like about 10% of the time I was able to predict which number was next. It was still better than Life of Pi which, aside from being irrational, included no estimations of Pi at all.

★☆☆☆☆ **Not really random**
Reviewed in the United States on September 26, 2012

I bought two copies of this book. I find that the first copy perfectly predicts what the numbers will be in the second copy. I feel cheated.

★★★★★ **great yet lacking....**
Reviewed in the United States on May 1, 2013

So far, it's great, but what I REALLY want is an audio version that I can listen to at the beach or the gym. Maybe narrated by Morgan Freeman???

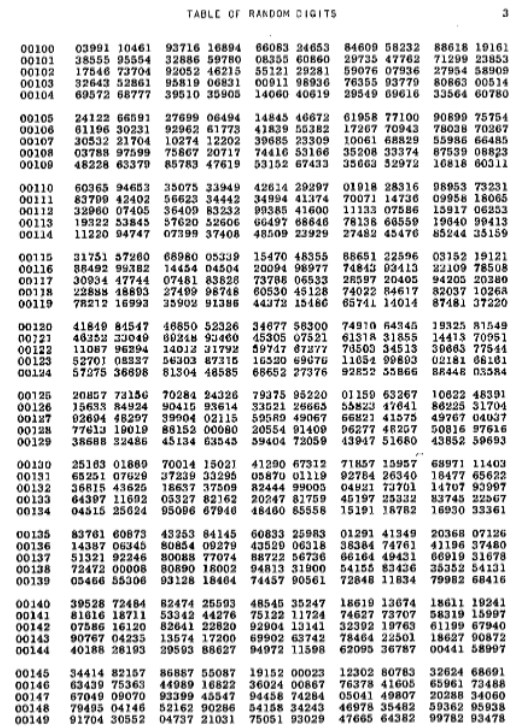
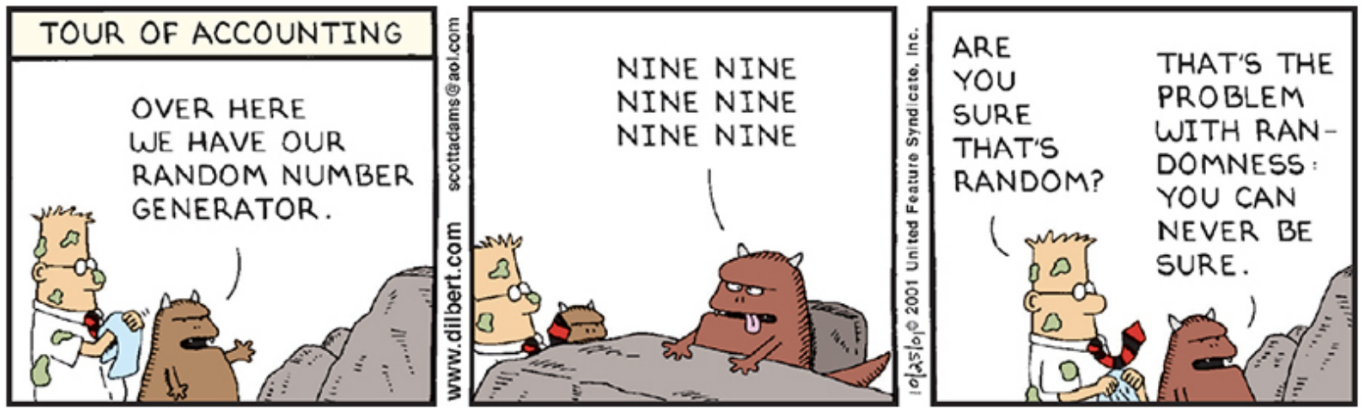


TABLE OF RANDOM DIGITS 3

00100	03991	10461	93716	16894	66083	24653	84609	58232	88618	19161
00101	38555	95554	32886	59780	08355	60860	29735	47762	71299	23853
00102	17546	73704	92052	46215	55121	29281	59076	07936	27954	58909
00103	32643	52861	95819	06831	00911	98936	76355	93779	80863	00514
00104	69572	68777	39510	35905	14060	40619	29549	69616	33564	60780
00105	24122	66591	27699	06494	14845	46672	61958	77100	90899	75754
00106	61196	30231	92962	61773	41839	55382	17267	70943	78038	70267
00107	30532	21704	10274	12202	39685	23309	10061	68829	55986	66485
00108	03788	97599	75867	20717	74416	53166	35208	33374	87539	08823
00109	48228	63379	85783	47619	53152	67433	35663	52972	16818	60311
00110	60365	94653	35075	33949	42614	29297	01918	28316	98953	73231
00111	83799	42402	56623	34442	34994	41374	70071	14736	09958	18065
00112	32960	07405	36409	83232	99385	41600	11133	07586	15917	06253
00113	19322	53845	57620	52606	66497	68646	78136	66559	19640	99413
00114	11220	94747	07399	37408	48509	23929	27482	45476	85244	35159
00115	31751	57260	68980	05339	15470	48355	88651	22596	03152	19121
00116	88492	99382	14454	04504	20094	98977	74843	93413	22109	78508
00117	30934	47744	07481	83828	73786	06533	28597	20405	94205	20380
00118	22888	48893	27499	98748	60530	45128	74022	84617	82037	10268
00119	78212	16993	35902	91386	44372	15486	65741	14014	87481	37220
00120	41849	84547	46850	52326	34677	58300	74910	64345	19325	81549
00121	48252	33049	60248	93460	45305	07521	61518	31855	14413	70951
00122	11087	96294	14012	31792	59747	87277	70503	34513	39663	77544
00123	52701	08337	56303	87315	16520	69676	11654	99893	02181	66161
00124	57275	36898	81304	48585	68652	27376	92852	55866	88448	03584
00125	20857	73156	70284	24326	79375	95220	01159	63267	10622	48391
00126	15633	84924	90415	93614	33521	26665	55823	47641	86225	31704
00127	92694	48297	39904	02115	59589	49067	66821	41575	49767	04037
00128	77613	19019	88152	00080	20554	91409	96277	48257	50816	97616
00129	38688	32486	45134	63545	59404	72059	43947	51680	43852	59693
00130	25163	01869	70014	15021	41290	67312	71857	15957	68971	11403
00131	65251	07629	37239	33295	05870	01119	92784	26340	16477	65622
00132	36815	43625	18637	37509	82444	99005	04921	73701	14707	93997
00133	64397	11692	05327	82162	20247	81759	45197	25332	83745	22567
00134	04515	25024	95096	67946	48460	85558	15191	18782	16930	33361
00135	83761	60873	43253	84145	60833	25983	01291	41349	20368	07126
00136	14387	06345	80854	09279	43529	06318	38384	74761	41196	37480
00137	51321	92246	80088	77074	88722	56736	66164	49431	66919	31678
00138	72472	00008	80890	18002	94813	31900	54155	83436	35352	54131
00139	05466	55306	93128	18404	74457	90561	72848	11834	79982	68416
00140	39528	72484	82474	25593	48545	35247	18619	13674	18611	19241
00141	81616	18711	53342	44276	75122	11724	74627	73707	58319	15997
00142	07586	16120	82641	22820	92904	13141	32392	19763	61199	67940
00143	90767	04235	13574	17200	69902	63742	78464	22501	18627	90872
00144	40188	28193	29593	88627	94972	11598	62095	36787	00441	58997
00145	34414	82157	86887	55087	19152	00023	12302	80783	32624	68691
00146	63439	75363	44989	16822	36024	00867	76378	41605	65961	73488
00147	67049	09070	93399	45547	94458	74284	05041	49807	20288	34060
00148	79495	04146	52162	90286	54158	34243	46978	35482	59362	95938
00149	91704	30552	04737	21031	75051	93029	47665	64382	99782	93478

Générer du hasard III

voir aussi Philipps (2020)



(dessin via https://dilbert.com/strip/2001-10-25)

"*A random sequence is a vague notion... in which each term is unpredictable to the uninitiated and whose digits pass a certain number of tests traditional with statisticians...*" Derrick Lehmer, cité par Knuth (1997) (voir aussi L'Ecuyer (2017) pour une perspective historique)

Générer du hasard IV

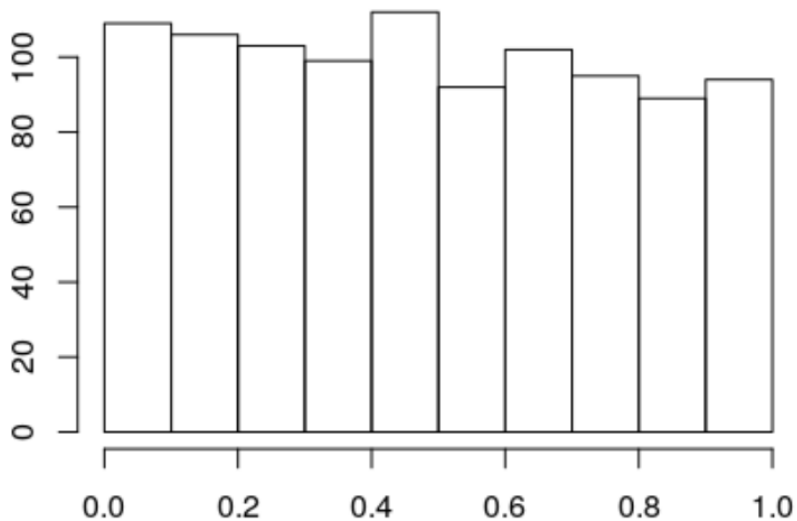
Heuristiquement, une suite (u_i) semble aléatoire si

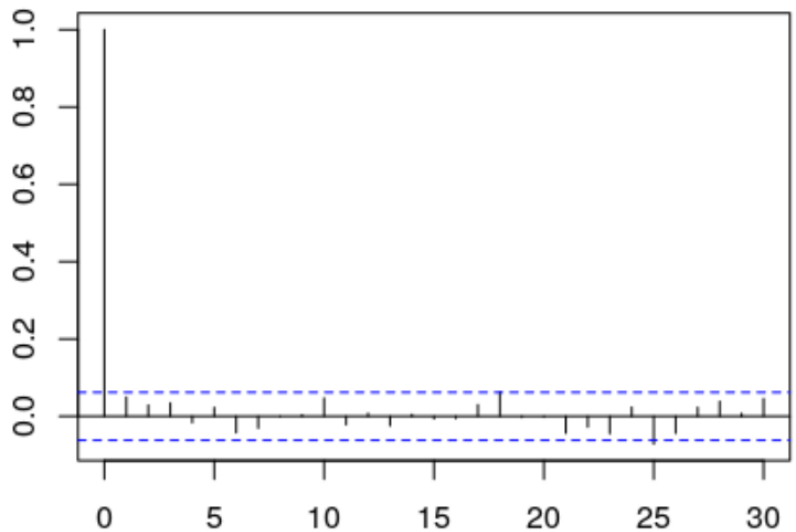
1. les tirages sont uniformes sur $[0,1]$: $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{u_i \in (a,b)} = b - a$

avec $b > a$,

2. les appels sont indépendants: pour $b > a$ et $d > c$.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{u_i \in (a,b), u_{i+k} \in (c,d)} = (b-a)(d-c), \ \forall k \in \mathbb{N}$$





Générer du hasard V

- approche physique

Utiliser des tirages de pièces, de dés, de roulette, etc

PFPPFPPPFF $\rightarrow$ $2^9 \cdot 1 + 2^8 \cdot 0 + 2^7 \cdot 1 + 2^6 \cdot 1 + 2^5 \cdot 0 + 2^4 \cdot 1 + 2^3 \cdot 1 + 2^2 \cdot 1 + 2^1 \cdot 0 + 2^0 \cdot 0 = 732$

PPPPPPFPPF $\rightarrow$ $2^9 \cdot 1 + 2^8 \cdot 1 + 2^7 \cdot 1 + 2^6 \cdot 1 + 2^5 \cdot 1 + 2^4 \cdot 1 + 2^3 \cdot 0 + 2^2 \cdot 1 + 2^1 \cdot 1 + 2^0 \cdot 0 = 1014$

- Utiliser des techniques mathématiques

von Neumann (1949), "middle-square method",

$8653^2 = 74874409 \rightarrow 8744^2 = 76457536 \rightarrow 4575^2 = 20930625$

$9306^2 = 86601636 \rightarrow 6016^2 = 36192256 \rightarrow 1922^2 = 3694084$

$6940^2 = 48163600 \rightarrow 1636^2 = 2676496 \rightarrow 6764^2 = 45751696$

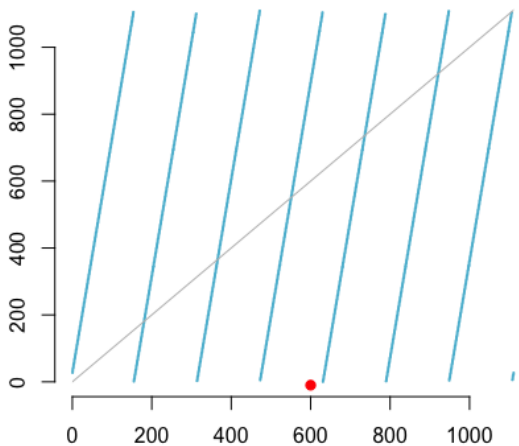
or use some linear congruential generator, from Lehmer (1951)

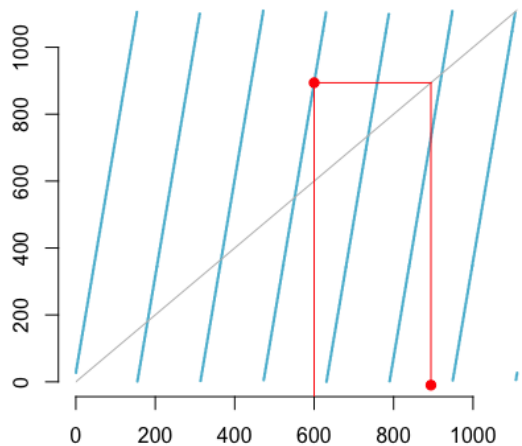
Générer du hasard VI

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = f(x_{n-1}) = (ax_{n-1} + c) \text{ modulo } m$$

$a = 7$, $c = 27$ and $m = 1111$, with $x_0 = 600$, so that $x_1 = 894$



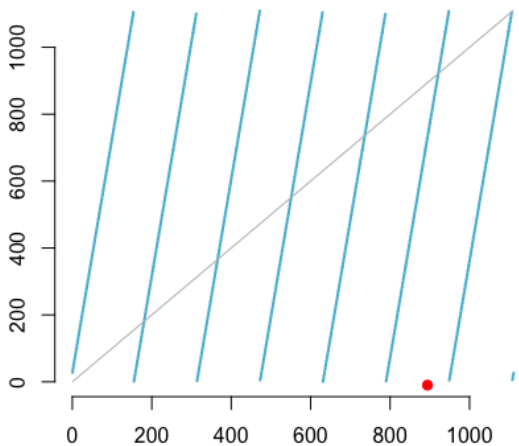


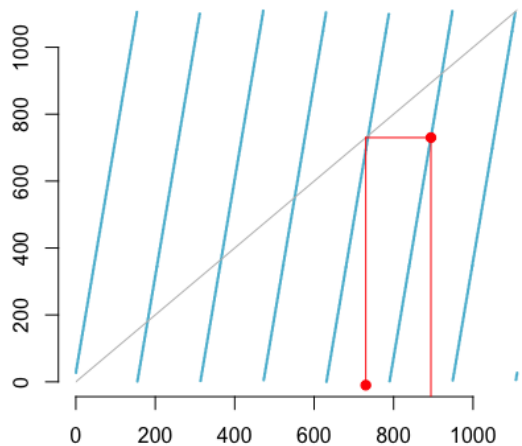
Générer du hasard VII

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 7$, $c = 27$ and $m = 1111$, with $x_1 = 894$, so that $x_2 = 730$



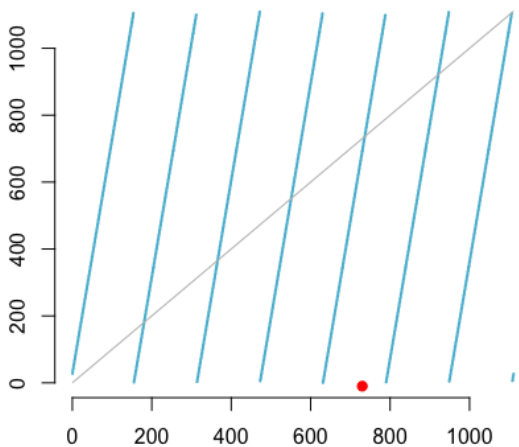


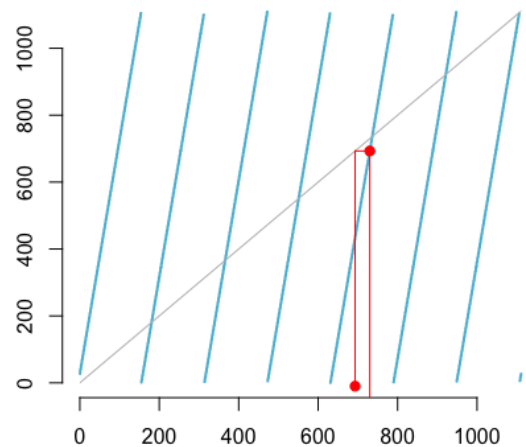
Générer du hasard VIII

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 7$, $c = 27$ and $m = 1111$, with $x_2 = 730$, so that $x_3 = 693$



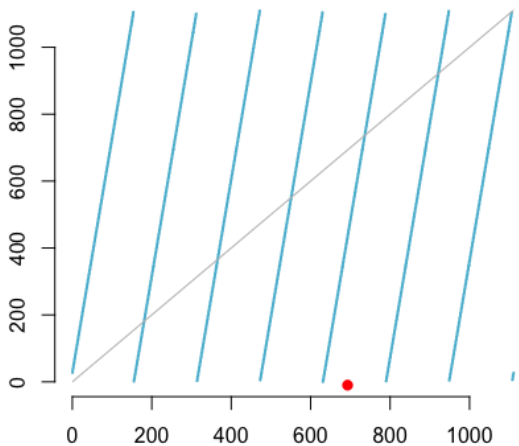


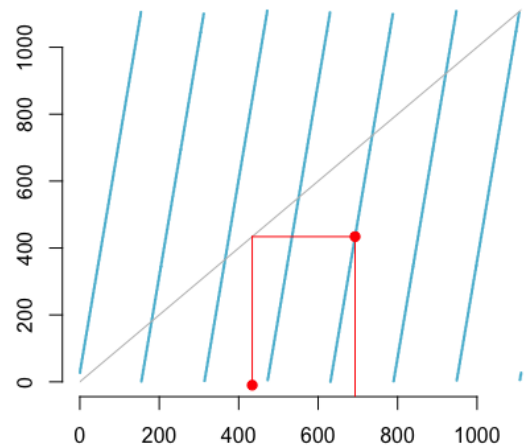
Générer du hasard IX

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 7$, $c = 27$ and $m = 1111$, with $x_3 = 693$, so that $x_4 = 434$



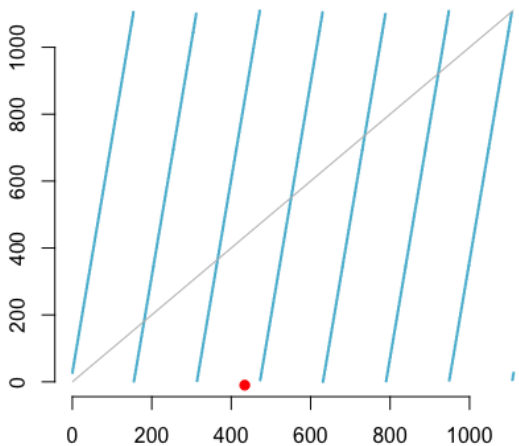


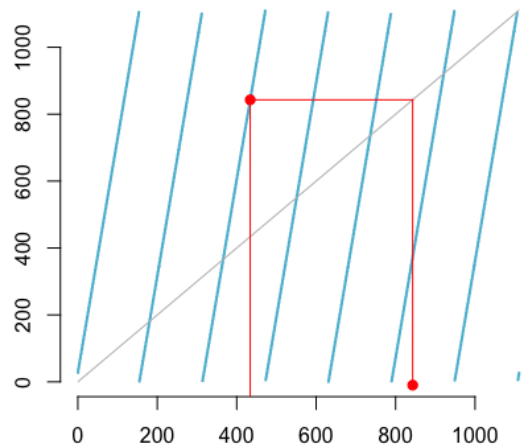
Générer du hasard X

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 7$, $c = 27$ and $m = 1111$, with $x_4 = 434$, so that $x_5 = 843$



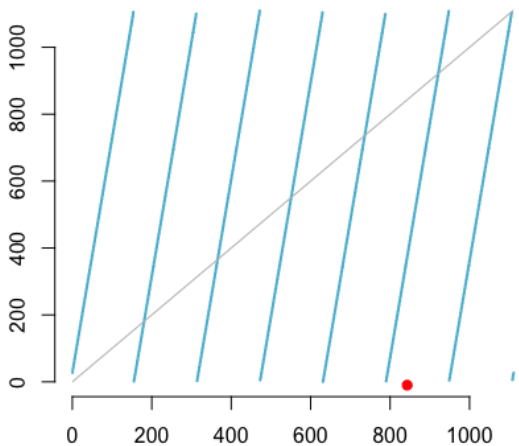


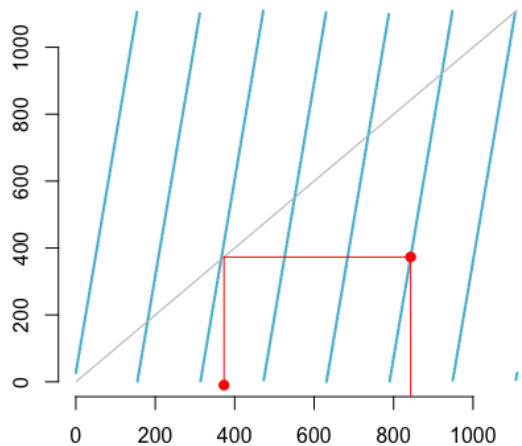
Générer du hasard XI

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 7$, $c = 27$ and $m = 1111$, with $x_5 = 843$, so that $x_6 = 373$



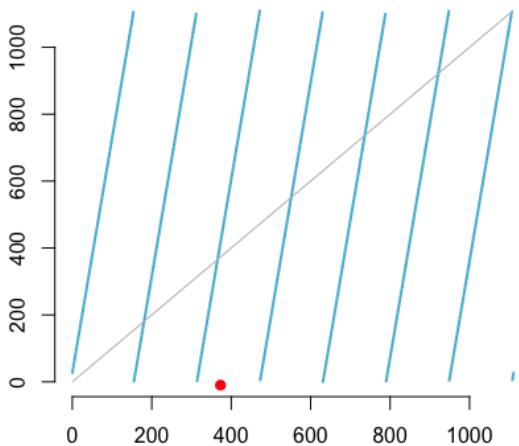


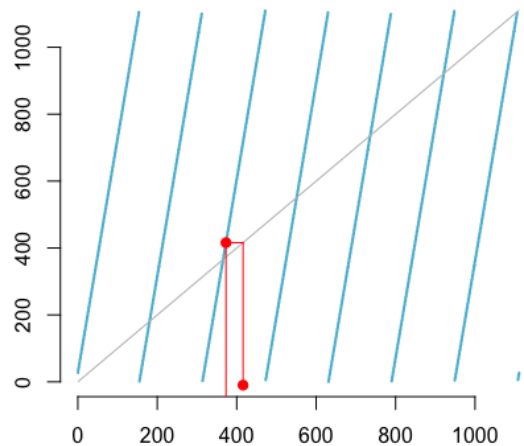
Générer du hasard XII

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 7$, $c = 27$ and $m = 1111$, with $x_6 = 373$, so that $x_7 = 416$



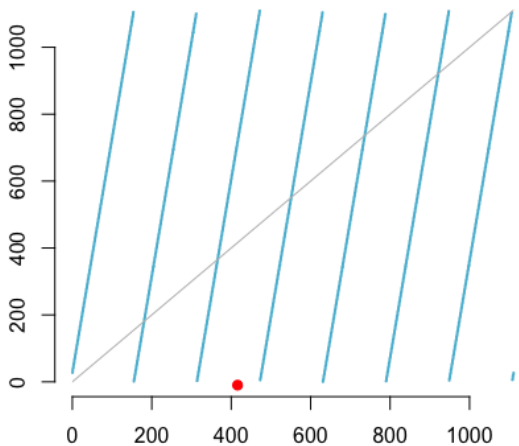


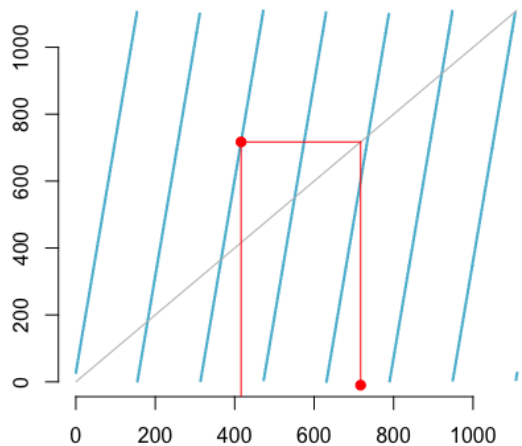
Générer du hasard XIII

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 7$, $c = 27$ and $m = 1111$, with $x_7 = 416$, so that $x_8 = 717$



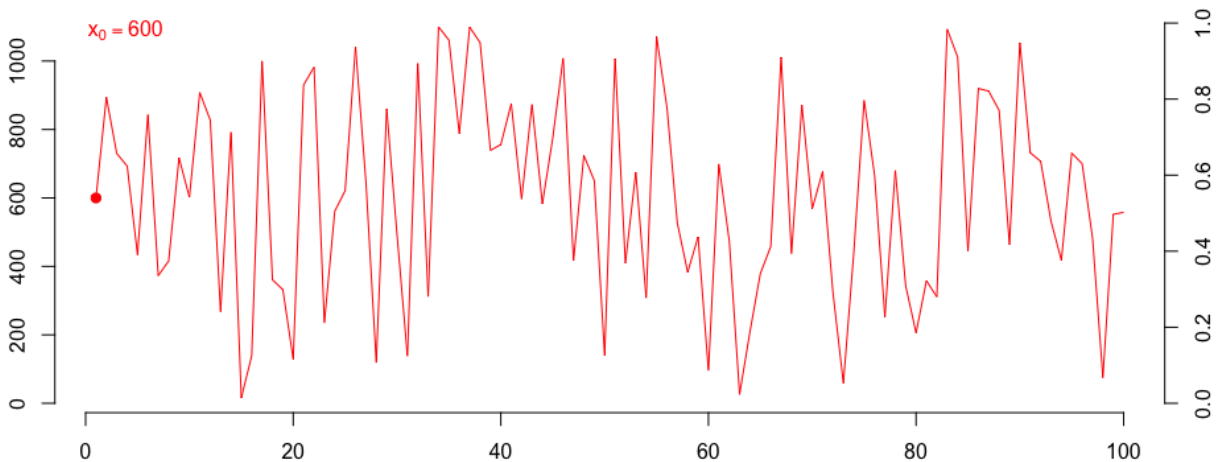


Générer du hasard XIV

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 7$, $c = 27$ and $m = 1111$, with $x_0 = 600$, see $x_1, x_2, \cdots, x_{100}$

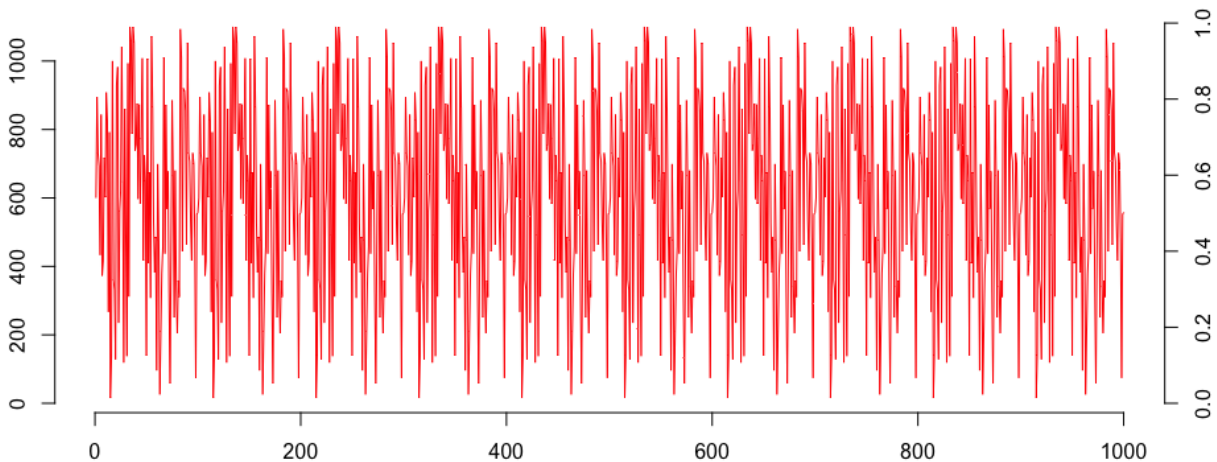


Générer du hasard XV

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 7$, $c = 27$ and $m = 1111$, with $x_0 = 600$, see $x_1, x_2, \cdots, x_{1000}$

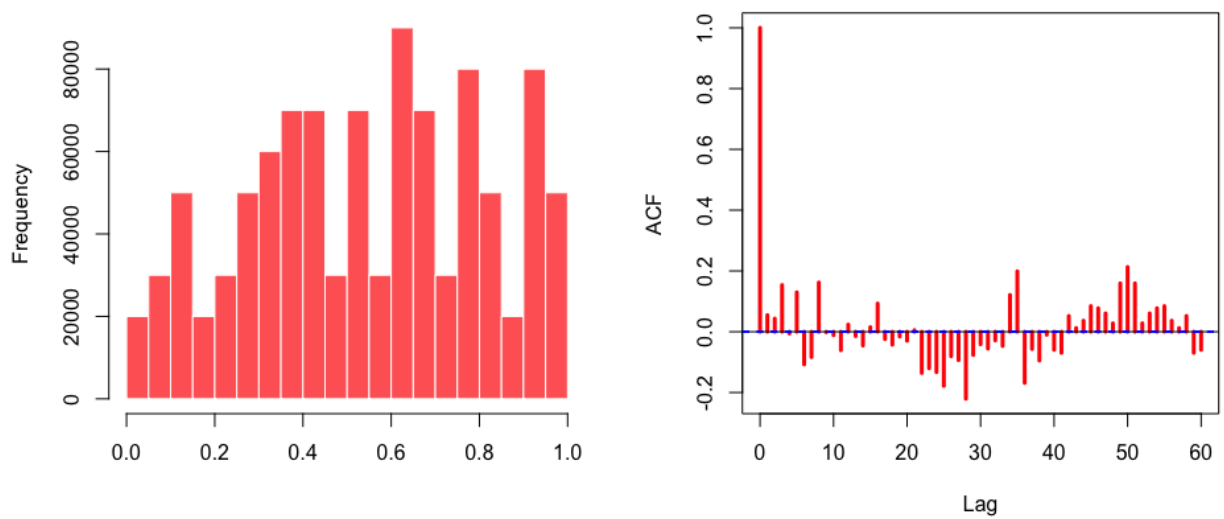


Générer du hasard XVI

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$,

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 7$, $c = 27$ and $m = 1111$, with $x_0 = 600$, distribution of x_i's and autocorrelation

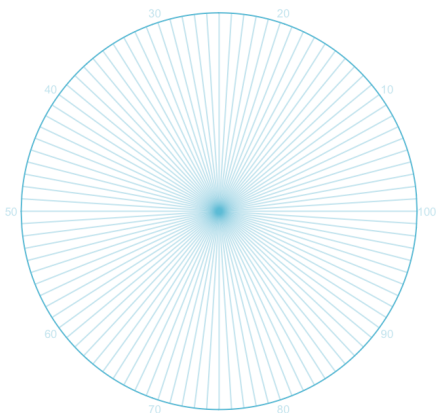


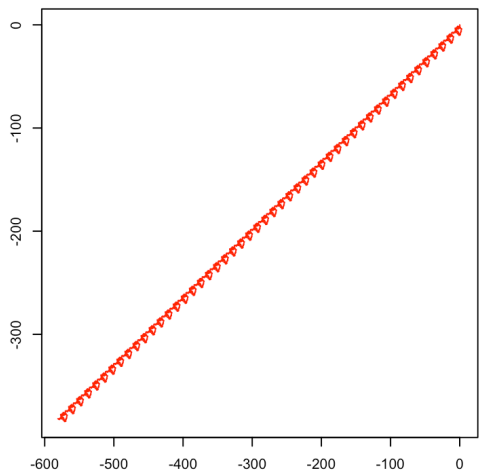
Générer du hasard XVII

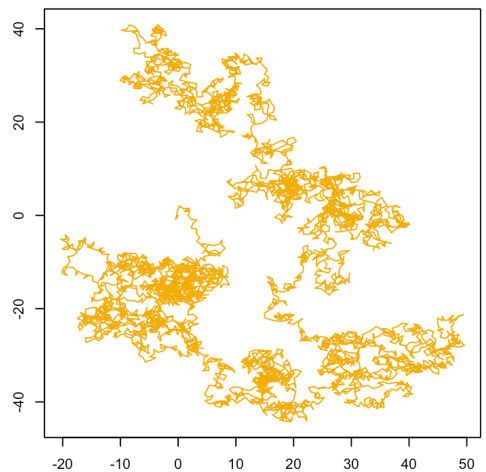
Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 7$, $c = 27$ and $m = 1111$, or $a = 74$, $c = 75$ and $m = 2^{16} + 1 = 65537$ (see ZX81),





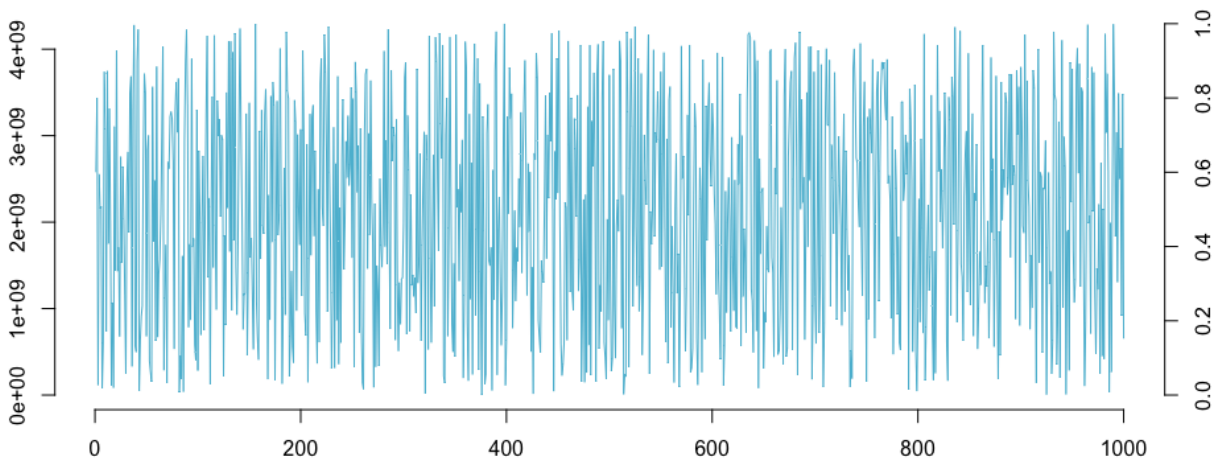


Générer du hasard XVIII

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 1013904223$, $c = 1664525$ and $m = 2^{32}$, with $x_0 = 2576980378$,

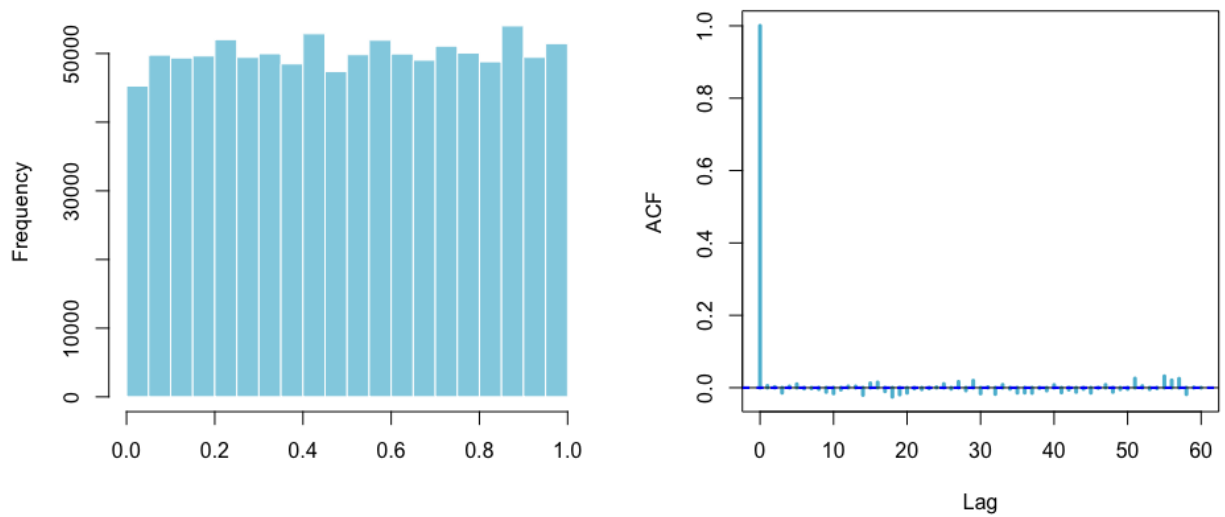


Générer du hasard XIX

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 1013904223$, $c = 1664525$ and $m = 2^{32}$, with $x_0 = 2576980378$,

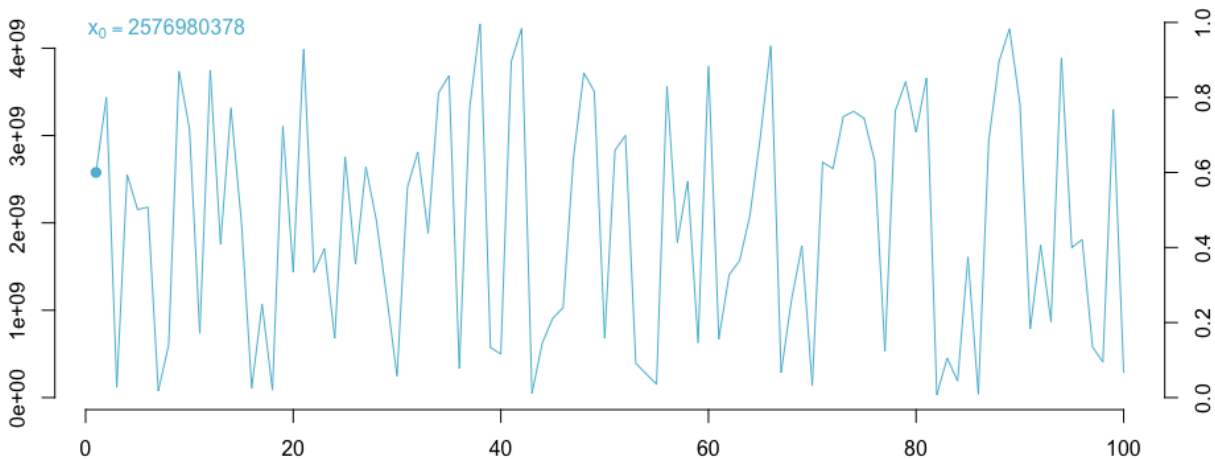


Générer du hasard XX

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 1013904223$, $c = 1664525$ and $m = 2^{32}$, with $x_0 = 2576980378$

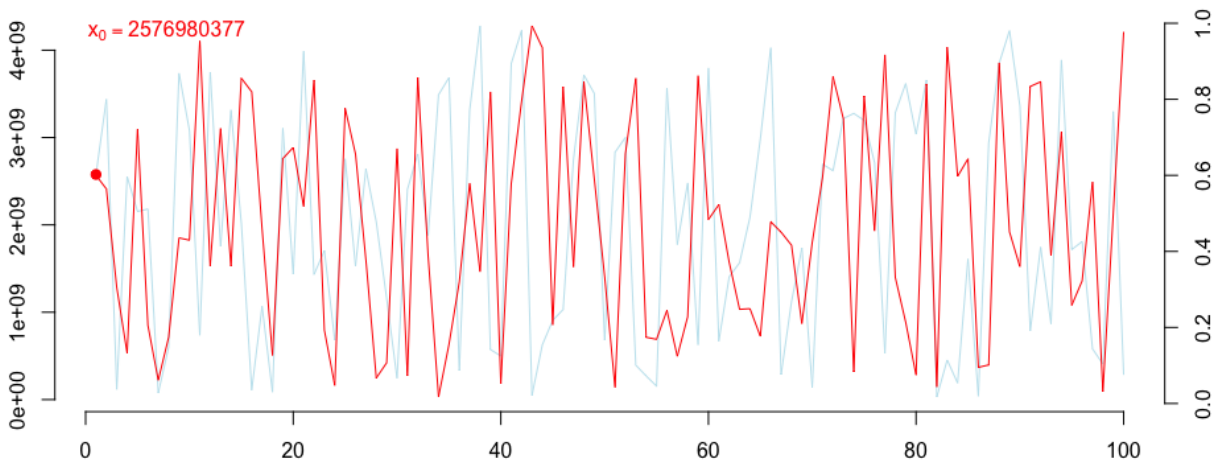


Générer du hasard XXI

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 1013904223$, $c = 1664525$ and $m = 2^{32}$, with $x_0 = 2576980377$

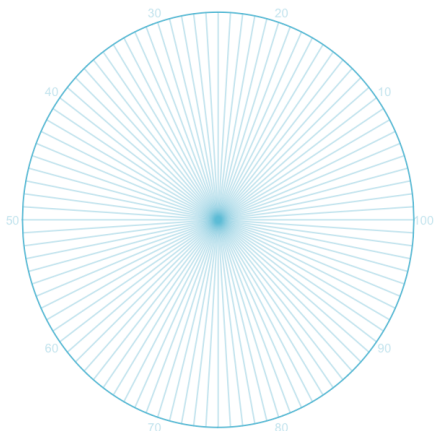


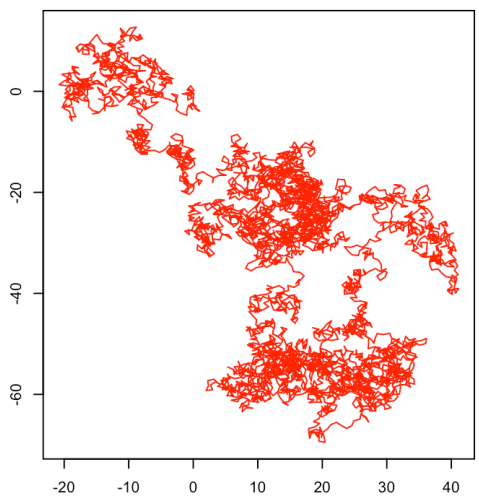
Générer du hasard XXII

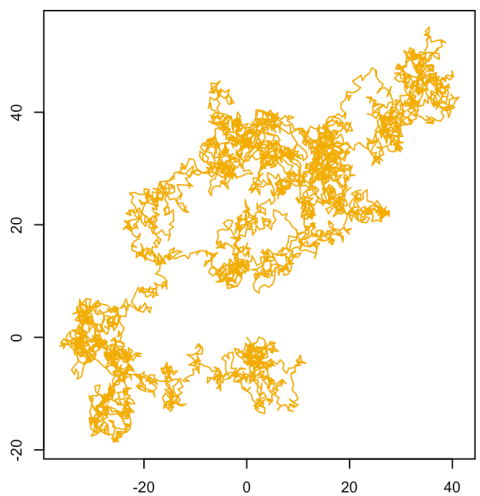
Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$

$a = 1013904223$, $c = 1664525$ and $m = 2^{32}$,





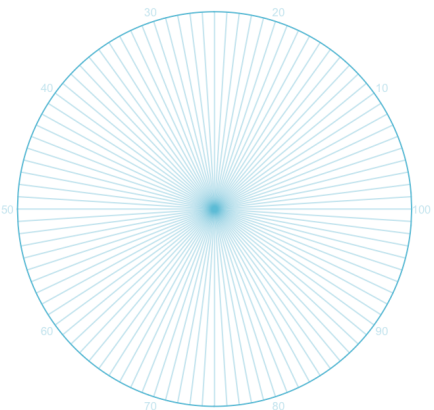


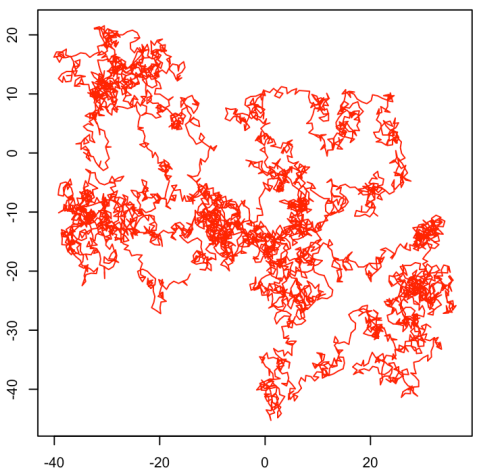
Générer du hasard XXIII

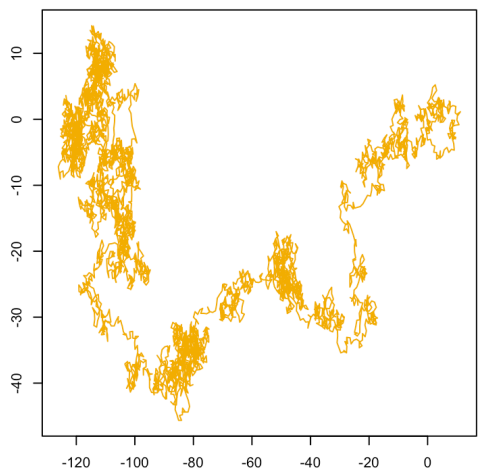
See also the use of π's digits

3.14159265358979323846264338327950288419716939937510582097494459230
81640628620899862803482534211706798214808651328230664709384460955 05

3.14159265358979323846264338327950288419716939937510582097494459230
81640628620899862803482534211706798214808651328230664709384460955 05







Générer du hasard XXIV

A lot of connexions with chaos theory
"*when the present determines the future, but the approximate present does not approximately determine the future*" Edward Lorenz, cited in May (1976)

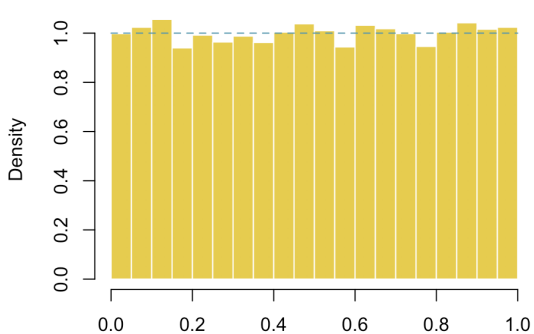
Monte Carlo I

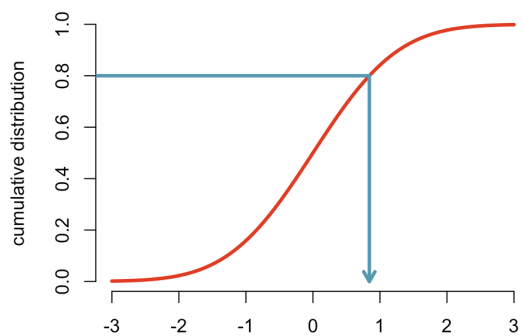
Probability Inverse Transform

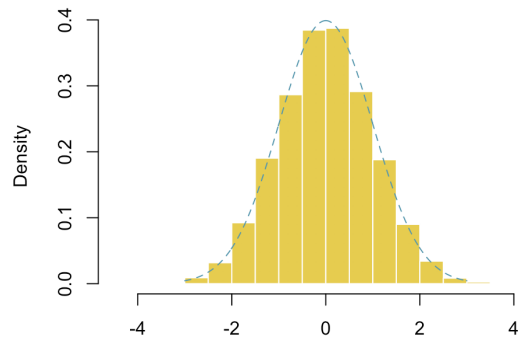
Let $F : \mathbb{R} \to [0, 1]$ denote a cumulative distribution function. If $U \sim \mathcal{U}([0, 1])$, then $X = F^{-1}(U)$ has cumulative distribution function F, where $F^{-1}(u) = \inf\{x : F(x) \geq u\}$.

Proof: Let $x \in \mathbb{R}$, $\mathbb{P}[X \leq x]$ is equal to

$$\mathbb{P}[F^{-1}(U) \leq x] = \mathbb{P}[F(F^{-1}(U)) \leq F(x)] = \mathbb{P}[U \leq F(x)] = F(x)$$







Monte Carlo II

Law of large numbers - Central limit theorem

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \xrightarrow{\mathbb{P}} \mu = \mathbb{E}_{\mathbb{P}}[X], \text{ i.e. } \forall \varepsilon, \lim_{n\to\infty} \mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| < \varepsilon\right] = 1$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \xrightarrow{\text{a.s.}} \mu = \mathbb{E}_{\mathbb{P}}[X], \text{ i.e. } \forall \varepsilon, \mathbb{P}\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right] = 1$$

$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$ where $\mu = \mathbb{E}_{\mathbb{P}}[X]$ and $\sigma^2 = \mathbb{E}_{\mathbb{P}}\left[(X-\mu)^2\right]$

Interestingly, the central limit theorem is valid also in higher dimension

$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = \mathbb{E}_{\mathbb{P}}[\boldsymbol{X}]$ and $\boldsymbol{\Sigma} = \mathbb{E}_{\mathbb{P}}\left[(\boldsymbol{X}-\boldsymbol{\mu})x(\boldsymbol{X}-\boldsymbol{\mu}\right.$

(speed of convergence does not depend on the dimension)

Monte Carlo III

If (U_i) are uniform independent random variables, if $g : \mathbb{R} \to \mathbb{R}$,

$$\frac{1}{n}\sum_{i=1}^{n} g(U_i) \xrightarrow{\mathbb{P}} \mathbb{E}_{\mathbb{P}}[g(U)] = \int_0^1 g(u)du, \text{ where } U \sim \mathcal{U}([0,1]).$$

If $(\boldsymbol{U}_i)$ are uniform independent random vectors, with $\boldsymbol{U}_i = (U_{i,1}, \cdots, U_{i,d})$, if $g : \mathbb{R}^d \to \mathbb{R}$,

$$\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{U}_i) \xrightarrow{\mathbb{P}} \mathbb{E}_{\mathbb{P}}[g(\boldsymbol{U})] = \int_{[0,1]^d} g(\boldsymbol{u})d\boldsymbol{u}, \text{ where } \boldsymbol{U} \sim \mathcal{U}([0,1]^d).$$

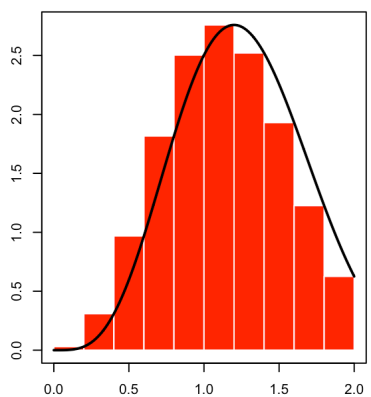
Note that either in dimension 1 or in dimension d, the speed of convergence is $O(n^{-1/2})$.

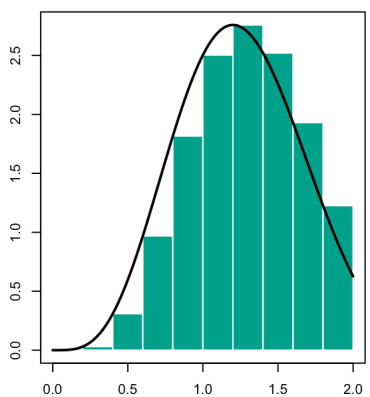
Monte Carlo IV

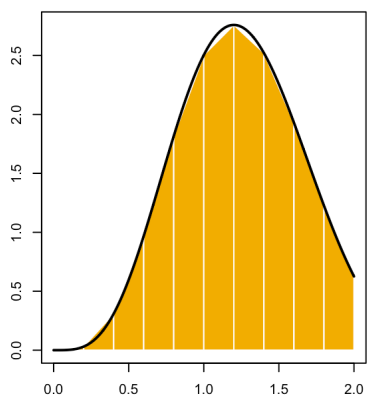
Example: integral approximation, error bounded in $O(n^{-1})$, $O(n^{-1})$ or $O(n^{-2})$

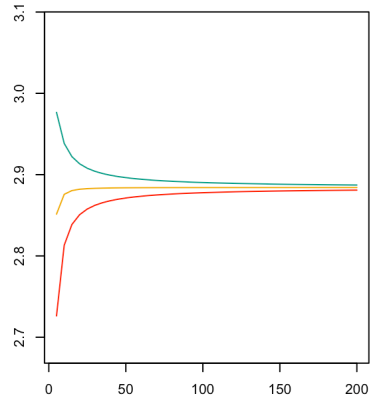
$$\int_0^2 h(y)dy = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{2}{n} \times h\left((i-1)\frac{2}{n}\right) = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{2}{n} \times h\left(i\frac{2}{n}\right)$$

$$\int_0^2 h(y)dy = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{2}{n} \times \frac{1}{2}\left(h\left((i-1)\frac{2}{n}\right) + h\left(i\frac{2}{n}\right)\right)$$



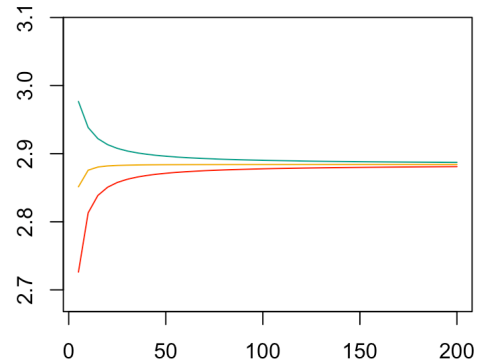






Monte Carlo V

$$\left| \int_0^2 h(y)dy - \sum_{i=1}^{n} \frac{2}{n} \times h\left(i\frac{2}{n}\right) \right| \leq \frac{2^2}{n} \|f'\|_\infty$$

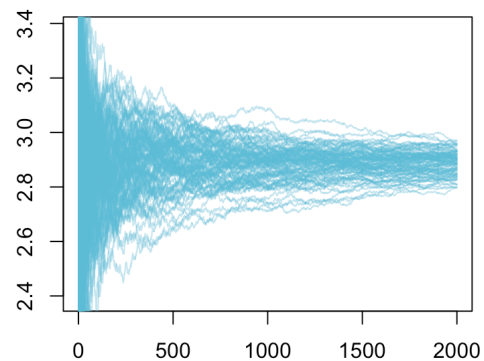


Using the law of large numbers

$$\int_0^2 h(y)dy = \int_0^1 2h(2u)du \lim_{n\to\infty} \frac{2}{n} \sum_{i=1}^{n} h(2u_i)$$

Concergence is slow,

$$\mathbb{P}\left[\left| \int_0^2 h(y)dy - \frac{2}{n} \sum_{i=1}^{n} h(2u_i) \right| \leq \frac{1.96\sigma}{\sqrt{n}} \right] = 95\%$$



which is $O(n^{-1/2})$, with a given probability.

How to create randomness?

Linear Congruential Method

Given $a, b, m \in \mathbb{N}_*$ and $x_0 \in \{0, 1, \cdots, m\}$, define

$$x_{i+1} = (ax_i + b) \text{ modulo } m,$$

and set $u_i = x_i / m$.

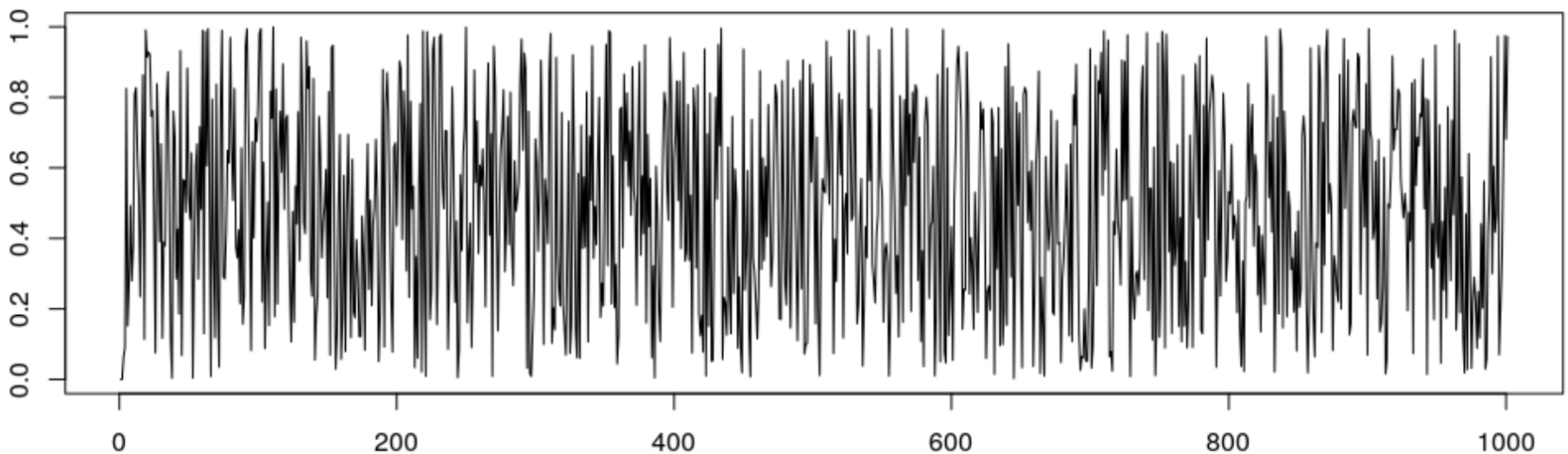
```
> a = 13; b = 43; m = 100; x = 77; u = rep(NA, 40)
> for (i in 1:40) {x = (a * x + b) %% m
+      u[i] = x / m }
> u
 [1] 0.44 0.15 0.38 0.37 0.24 0.55 0.58 0.97 0.04 0.95
[11] 0.78 0.57 0.84 0.35 0.98 0.17 0.64 0.75 0.18 0.77
[21] 0.44 0.15 0.38 0.37 0.24 0.55 0.58 0.97 0.04 0.95
```

Problem: not all values in $\{0, \cdots, m-1\}$ are obtained, and there is a cycle here.

How to create randomness?

Solution: (very) large values for m and choose properly a and b.

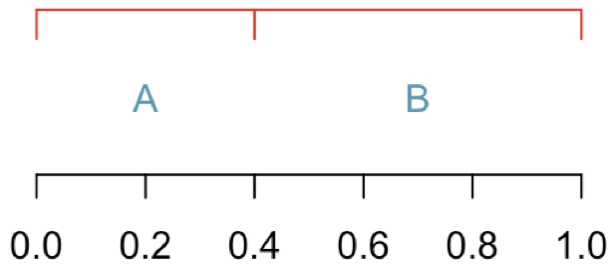
E.g. $m = 2^{32} - 1$, $a = 16807$ $(= 7^5)$ and $b = 0$ (used in Matlab).



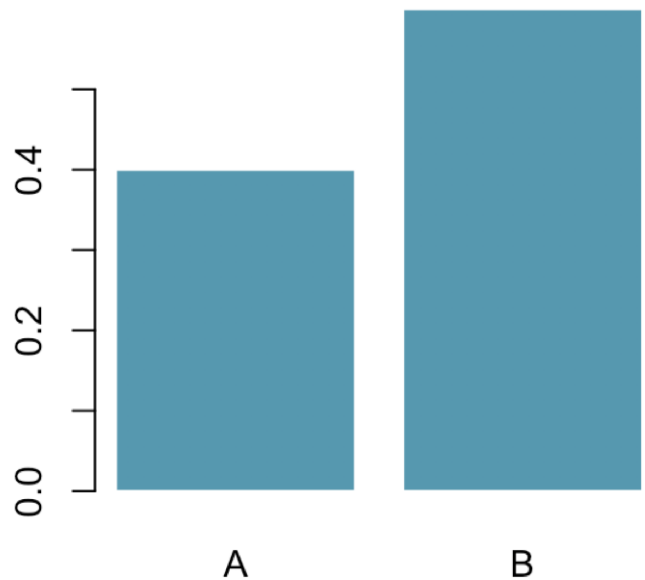
Génération de loi binomiale

Soit $p \in (0,1)$, $U \sim \mathcal{U}_{[0,1]}$

$$X = \begin{cases} 1 \text{ si } U < p \\ 0 \text{ si } U \geq p \end{cases}$$




```
> p = 0.4
> n = 1e7
> U1 = runif(n)
> Z = (U1<p)*1
> barplot(table(Z)/n)
```



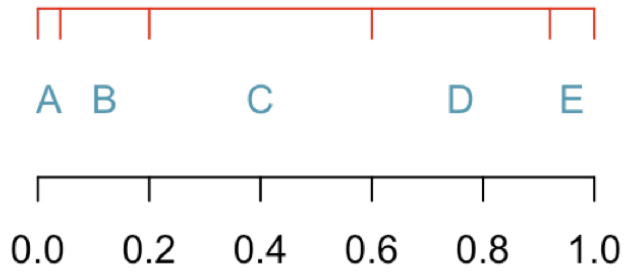

```
> Z = sample(0:1, size=n,
    replace=TRUE,
  prob=c(1-p,p))
> table(Z)
Z
      0       1
6000144 3999856
```

Génération de loi multinomiale

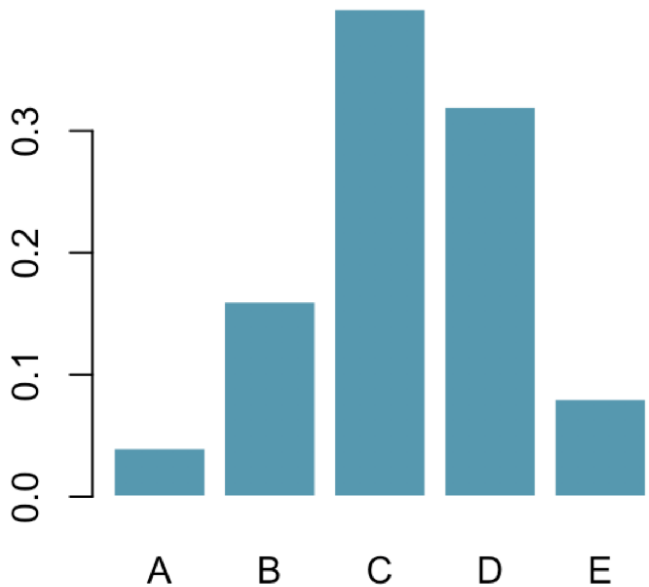
Soit $\boldsymbol{p}$ un vecteur de probabilité,

$$\overline{p}_1 = 0 \text{ et } \overline{p}_{j+1} = \sum_{i=1}^{j} p_i$$

$$U \in [\overline{p}_j, \overline{p}_{j+1}) \Longrightarrow X = j + 1$$



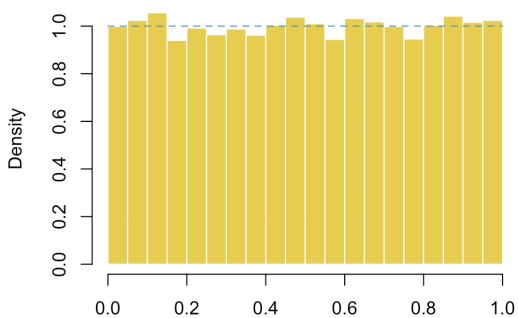

```
> p = c(0.04, 0.16, 0.40,
    0.32, 0.08)
> cumsum(p)
[1] 0.04 0.20 0.60 0.92 1.00
> n = 1e7
> U1 = runif()
> Z = cut(U1, breaks = c(0,
    cumsum(p)), labels =
    LETTERS[1:5])
> barplot(table(Z)/n)
```

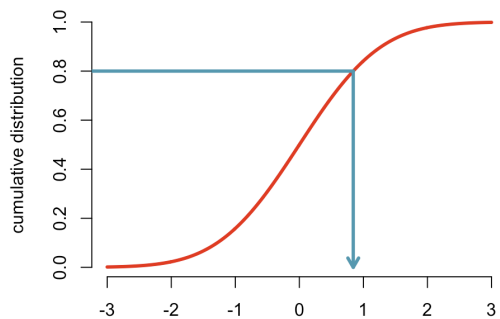


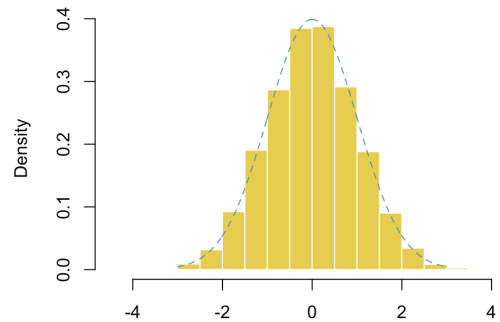
Inversion de la Fonction de Répartition

Inversion de fonction de répartition, inverse method sampling

Soit F une fonction de répartition, si $U \sim \mathcal{U}([0,1])$, $X = F^{-1}(U)$ a pour fonction de répartition F.







F^{-1} est simplement la fonction quantile,

```
> Q = function(u) qnorm(u, 0, 1)
```

Inversion de la Fonction de Répartition

```
> U = runif(100)
 [1] 0.27 0.37 0.57 0.91 0.20 0.90 0.94 0.66 0.63 0.06
[11] 0.21 0.18 0.69 0.38 0.77 0.50 0.72 0.99 0.38 0.78
[21] 0.93 0.21 0.65 0.13 0.27 0.39 0.01 0.38 0.87 0.34
[31] 0.48 0.60 0.49 0.19 0.83 0.67 0.79 0.11 0.72 0.41
[41] 0.82 0.65 0.78 0.55 0.53 0.79 0.02 0.48 0.73 0.69
[51] 0.48 0.86 0.44 0.24 0.07 0.10 0.32 0.52 0.66 0.41
[61] 0.91 0.29 0.46 0.33 0.65 0.26 0.48 0.77 0.08 0.88
[71] 0.34 0.84 0.35 0.33 0.48 0.89 0.86 0.39 0.78 0.96
```

```
> Q(U)
 [1] -0.63 -0.33  0.18  1.33 -0.84  1.27  1.60  0.41
 [9]  0.33 -1.54 -0.82 -0.93  0.49 -0.29  0.74 -0.01
[17]  0.58  2.40 -0.31  0.76  1.51 -0.80  0.39 -1.15
[25] -0.62 -0.29 -2.21 -0.30  1.12 -0.41 -0.04  0.25
[33] -0.02 -0.89  0.94  0.44  0.82 -1.24  0.59 -0.22
[41]  0.92  0.38  0.78  0.13  0.07  0.80 -1.99 -0.06
[49]  0.62  0.50 -0.06  1.09 -0.16 -0.69 -1.47 -1.28
[57] -0.48  0.05  0.42 -0.24  1.36 -0.54 -0.10 -0.43
[65]  0.39 -0.65 -0.05  0.73 -1.38  1.15 -0.41  0.99
```

Inversion de la Fonction de Répartition Empirique

Given a sample $\{x_1, \cdots, x_n\}$ i.i.d. from F,

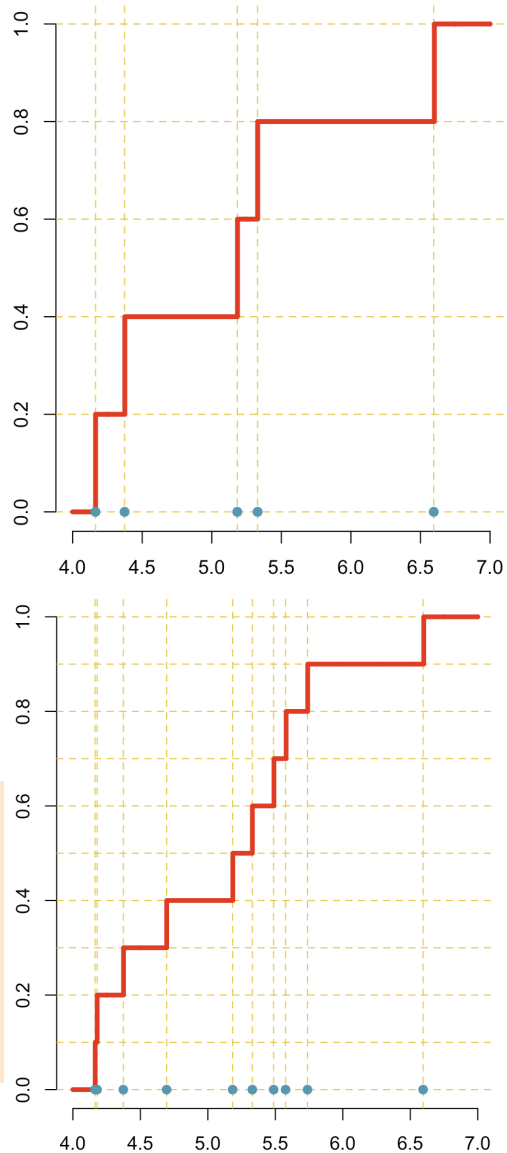
$$F(x) = \mathbb{P}[X \leq x],$$

the empirical cumulative distribution function is

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i \leq x), \ x \in \mathbb{R}$$

Glivenko-Cantelli: $\hat{F}_n \to F$ as $n \to \infty$.

```
1 > Finvemp = function(u,x) sort(x)[
      ceiling(u*length(x))]
2 >  Qinv = Vectorize(function(u){
3    Finvemp(u,x)
4    })
```



Inversion de la Fonction de Répartition Empirique

The inverse method with $\widehat{F}_n$ simply means resampling within $\{x_1, \cdots, x_n\}$ with equal probabilities $1/n$ (or *with replacement*)

```
> x
[1] 4.164 4.374 5.184 5.330 6.595
> Qemp(U)
 [1] 6.60 6.60 6.60 5.33 4.37 5.33 5.33 4.16 6.60 5.33
[11] 4.37 4.37 4.37 6.60 5.33 5.18 5.33 5.18 6.60 5.18
[21] 5.18 4.37 6.60 4.37 4.16 6.60 4.16 6.60 5.33 4.16
[31] 4.16 6.60 4.37 4.37 5.33 5.18 5.18 5.18 5.33 5.33
[41] 4.37 5.18 5.33 5.18 4.37 5.18 5.18 5.18 5.33 5.18
[51] 5.33 4.37 4.37 4.16 5.18 5.18 5.18 5.18 4.16 5.18
[61] 4.37 4.16 4.16 4.16 6.60 4.37 4.37 5.33 5.18 4.16
[71] 5.33 4.16 6.60 5.18 4.16 4.16 5.18 4.16 5.18 4.16
```

called bootstrapping

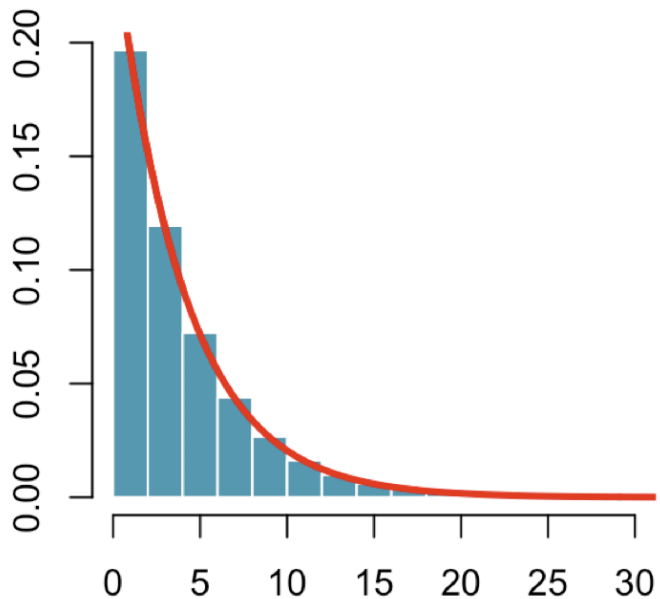
```
> sample(x, size = 80, replace = TRUE)
```

Génération de loi Exponentielle

$F(x) = \mathbb{P}[X < x] = 1 - e^{-ax}$ pour $x \geq 0$. On veut $1 - e^{-aq} = u$, i.e. $e^{-aq} = 1 - u$, $aq = -\log(1-u)$

$$F^{-1}(u) = \frac{-1}{a}\log(1-u), \text{ pour } u \in [0,1].$$

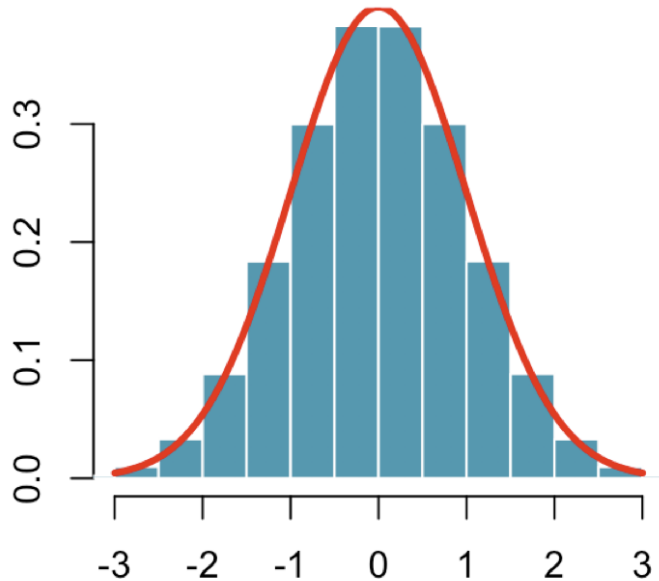
```
1 > a = 1/4
2 > U1 = runif(1e7)
3 > Z = -log(1-U1)/a
4 > hist(Z, probability=TRUE,
    xlim=c(0,30))
5 > curve(dexp(x,a), add=TRUE)
```



Génération de loi Gaussienne $\mathcal{N}(0,1)$

Si $U_1, U_2 \sim \mathcal{U}_{[0,1]}$, indépendantes, $R = \sqrt{-2\log(U_1)}$ et $\Theta = 2\pi U_2$, alors $(X_1, X_2) = (R\cos\Theta, R\sin\Theta)$ est un couple de variables $\mathcal{N}(0,1)$ indépendantes

```
> U1 = runif(1e7)
> U2 = runif(1e7)
> R = sqrt(-2*log(U1))
> Theta = 2*pi*U2
> Z = R*cos(Theta)
> hist(Z, proba=TRUE)
> curve(dnorm(x,0,1),add=TRUE)
```

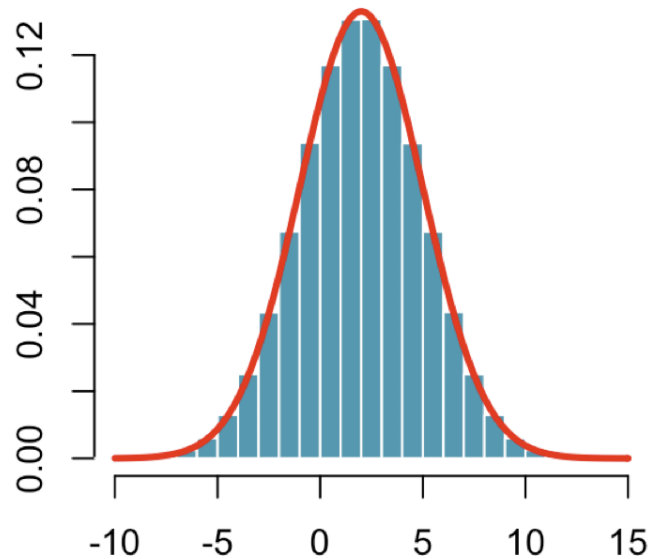


On parle de la méthode de Box-Muller.

Génération de loi Gaussienne $\mathcal{N}(\mu, \sigma^2)$

Si $Z \sim \mathcal{N}(0,1)$, $X = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$

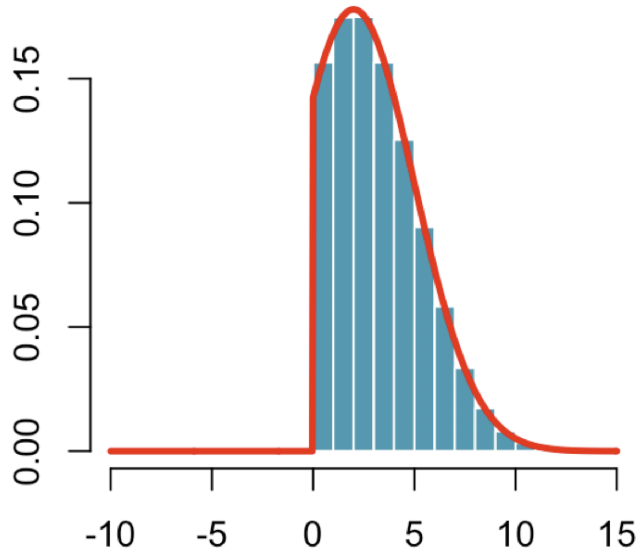
```
> U1 = runif(1e7)
> U2 = runif(1e7)
> R = sqrt(-2*log(U1))
> Theta = 2*pi*U2
> Z = R*cos(Theta)
> X = 2+3*Z
> hist(X, proba=TRUE)
> curve(dnorm(x,2,3),add=TRUE)
```



Génération de loi Gaussienne $\mathcal{N}(\mu, \sigma^2)$ censurée

On veut simuler X conditionnellement à $X > 0$

```
> U1 = runif(1e7)
> U2 = runif(1e7)
> R = sqrt(-2*log(U1))
> Theta = 2*pi*U2
> Z = R*cos(Theta)
> X = 2+3*Z
> X = X[X>0]
> hist(X, proba=TRUE)
```

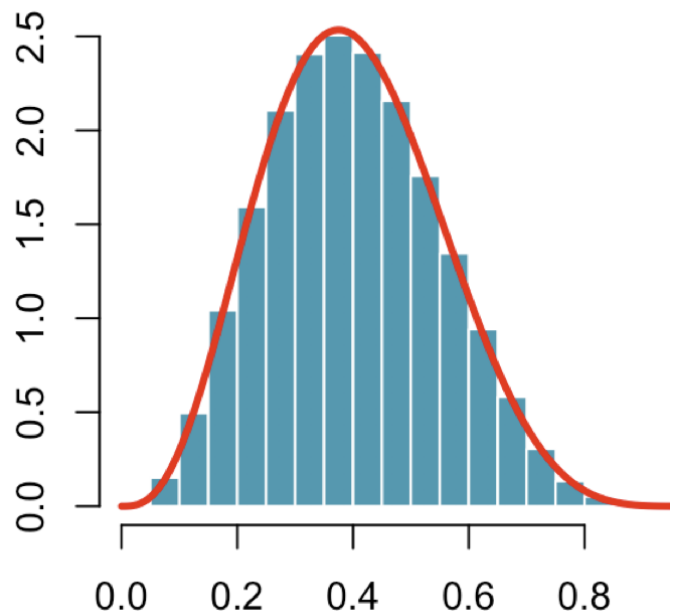


Génération de loi Gamma

C'est plus compliqué...

On peut utiliser les fonctions R pour les lois usuelles, `runif`, `rbinom`, `rpois`, `rexp`, `rnorm`, `rlnorm`, `rgamma`, etc.

```
> a = 4
> b = 6
> Z = rgamma(1e7, a, b)
> hist(Z, probability=TRUE)
> curve(dgamma(x,a,b))
```



Notion de mélange

Sommme de fonctions de répartition

Si F_1 et F_2 sont deux fonctions de répartition, si $p_1, p_2 \in (0,1)$, avec $p_1 + p_2 = 1$

$$F : x \mapsto p_1 F_1(x) + p_2 F_2(x)$$

est également une fonction de répartition

Mélange de lois

Si $X_1 \sim F_1$ et $X_2 \sim F_2$,

$$X = \mathbf{1}(Z = 1)X_1 + \mathbf{1}(Z = 2)X_2 = \begin{cases} X_1 \text{ si } Z = 1 \text{ (proba } p_1) \\ X_2 \text{ si } Z = 2 \text{ (proba } p_2) \end{cases}$$

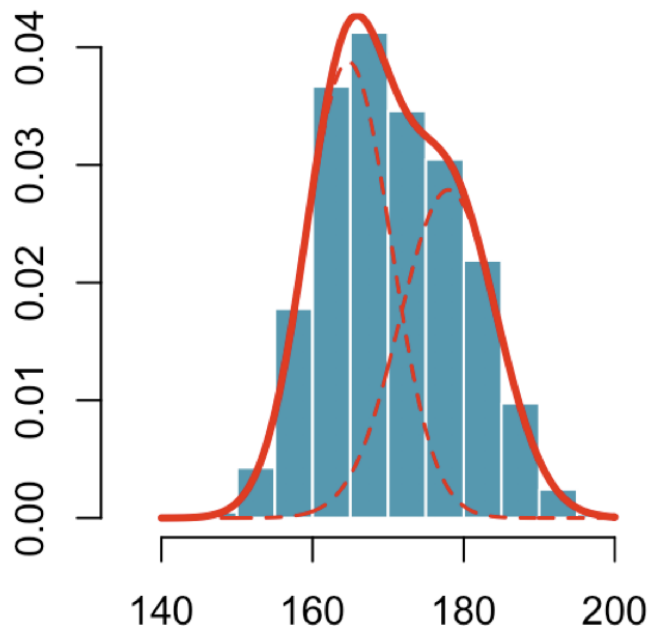
a pour loi $F(x) = p_1 F_1(x) + p_2 F_2(x)$.

Génération d'une loi mélange

Soit $I \sim \mathcal{B}(p)$, $p \in (0,1)$

$$X \sim \begin{cases} \mathcal{N}(\mu_1, \sigma_1^2) \text{ si } I = 1 \\ \mathcal{N}(\mu_2, \sigma_2^2) \text{ si } I = 2 \end{cases}$$

```
1 > m1 = 178
2 > m2 - 164
3 > s1 = 6.44
4 > s2 = 5.66
5 > p  = 0.45
6 > I = sample(1:2,size = 1e6,
      prob = c(p,1-p), replace=
      TRUE)
7 > Z = rnorm(1e6,m1,s1)*(I==1)+
      rnorm(1e6,m2,s2)*(I==2)
8 > hist(Z, proba=TRUE)
```



Génération d'un vecteur Gaussien ★★★

Décomposition de Cholesky

Soit $\boldsymbol{\Sigma}$ une matrice de variance-covariance (matrice symétrique définie positive), il existe une matrice triangulaire inférieure $\boldsymbol{L}$ telle que $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$,

$$\boldsymbol{L} = \begin{bmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix}$$

Il existe une unique matrice triangulaire inférieure dont les termes de la diagonale sont positifs telle que $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$.

Génération d'un vecteur Gaussien ★★★

Example:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & -6 & 8 & 2 \\ -6 & 10 & -15 & -3 \\ 8 & -15 & 26 & -1 \\ 2 & -3 & -1 & 62 \end{bmatrix} \text{ et } \boldsymbol{L} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ -3 & 1 & 0 & 0 \\ 4 & -3 & 1 & 0 \\ 1 & 0 & -5 & 6 \end{bmatrix}$$

```
> S = matrix(c(4,-6,8,2,-6,10,-15,-3,8,-15,26,
    -1,2,-3,-1,62),4,4)
> chol(S)
     [,1] [,2] [,3] [,4]
[1,]    2   -3    4    1
[2,]    0    1   -3    0
[3,]    0    0    1   -5
[4,]    0    0    0    6
```

Génération d'un vecteur Gaussien ★★★

Vecteur Gaussien, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ et $\mathcal{N}(\mathbf{0}, \mathbb{I})$

$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ si et seulement si $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}^\top \boldsymbol{Z}$ où $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$ et $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$.

$$\begin{cases} X_1 &= \mu_1 + L_{11} Z_1 \\ X_2 &= \mu_2 + L_{21} Z_1 + L_{22} Z_2 \\ X_3 &= \mu_3 + L_{31} Z_1 + L_{32} Z_2 + L_{33} Z_3 \\ \vdots & \\ X_d &= \mu_d + L_{d1} Z_1 + L_{d2} Z_2 + \cdots + L_{d(d-1)} Z_{d-1} + L_{dd} Z_d \end{cases}$$