Générateurs pseudo-aléatoires et tests statistiques

Maple, Matlab et tous les autres langages de programmation contiennent une commande `rand` qui permet d'obtenir des nombres "au hasard". Ses utilisations sont très nombreuses et variées : de la cryptographie au calcul d'intégrales multiples, en passant bien sûr par la modélisation stochastique. Pour simuler n'importe quels processus aléatoires, il suffit d'être capable de générer une suite de variables aléatoire indépendantes uniformes sur l'intervalle $[0, 1]$. On appelle "générateur pseudo-aléatoire" une procédure qui fait cela. Bien sûr, ces nombres ne sont pas vraiment aléatoires : ils sont produits par un ordinateur, c'est-à-dire une machine strictement déterministe. Dès lors, quand peut-on dire qu'un générateur est "bon" ? Quand il fournit des nombres qui ont les *mêmes propriétés statistiques* que des variables aléatoires indépendantes uniformes sur $[0, 1]$.

1 Quelques idées plus ou moins bonnes de générateurs

- Programmer des fonctions Maple permettant d'obtenir les n premiers termes des suites définies par les égalités suivantes :
  - $a_n = ne \mod 1$;
  - $b_n = n\frac{1+\sqrt{5}}{2} \mod 1$;
  - $c_n = \log n \mod 1$;
  - $l_n = \frac{L_n}{2^{31}}$, où $L_n = 65539L_{n-1} \mod 2^{31}$ et L_1 est fixé selon votre choix ;
  - $m_n = \frac{M_n}{2^{31}-1}$, où $M_n = 16807M_{n-1} \mod \left(2^{31} - 1\right)$ et M_1 est fixé selon votre choix ;

  Définir (et comprendre) également les procédures `gen_p` et `gen_q` suivantes :
  ```
  > gen_p :=proc(n) local i, M ; M=2∧30 ; [seq(rand(M)()/M, i=1..n)] ; end ;
  > gen_q :=n -> map(x->evalf(sin(Pi/2*x)), gen_p(n)) ;
  ```
  Que pensez-vous de ces différents candidats générateurs ?
- l_n a été à partir des années soixante un générateur de référence très largement utilisé sous le nom RANDU. Son succès provenait de sa vitesse d'exécution : l'arithmétique 32 bits se faisait "naturellement" modulo 2^{31}, et l'écriture $65539 = 2^{16} + 3$ permettait des calculs très rapides. Il reste aujourd'hui comme le modèle de la fausse bonne idée...
  Pour comprendre pourquoi, représenter sur un graphe 3D les points de coordonnées (l_n, l_{n+1}, l_{n+2}) ; on pourra utiliser la fonction `pointplot3d` préalablement chargée par la commande `with(plots)`. Que remarque-t-on ?
  Pour comprendre ce phénomène, on peut remarquer que $L_{n+2} = 6L_{n+1} + 9L_n \mod 2^{31}$.
- Par contre m_n, introduit par Park & Miller à la fin des années 80, est aujourd'hui recommandé comme "Standart Minimal".

2 Tests et intervalles de confiance

2.1 Estimateur consistant d'une probabilité

Certains utilisateurs mettent en doute la qualité du générateur `p` : ils pensent en effet que celui-ci génère trop de valeurs proches de la moyenne 0.5. On va donc s'intéresser à la probabilité que notre générateur pseudo-aléatoire renvoie un nombre compris entre 0.4 et 0.6.

Pour cela, on modélise la situation de la façon suivante : on suppose que les valeurs renvoyées par notre générateurs forment une suite de variables aléatoires $(X_i)_{i\in\mathbb{N}^*}$ indépendantes à valeur dans l'intervalle $[0, 1]$. On suppose que l'on observe les réalisations de $X_1, \ldots, X_n$ grâce à n appels au générateur.

Soit p la probabilité que X_i appartienne à l'intervalle $[0.4, 0.6]$. Il vient immédiatement à l'idée d'estimer p par la fréquence empirique

$$F_n = \frac{N_n}{n}, \text{ avec } N_n = \#\big\{i \in \{1, \ldots, n\} : X_i \in [0.4, 0.6]\big\}.$$

Notons $Z_i = \mathbb{1}_{X_i\in[0.4,0.6]}$ la variable aléatoire qui vaut 1 si le nombre renvoyé est dans l'intervalle visé et 0 sinon. Z_i suit une loi de Bernoulli de paramètre p inconnu, et donc $N_n = \sum_{i=1}^n Z_i$ suit une loi binomiale $\mathcal{B}in(n, p)$. Par conséquent, la valeur moyenne attendue pour N_n est $\mathbb{E}_p[N_n] = np$, et on s'attend à observer autour de cette valeur des fluctuations de l'ordre de $\sqrt{\operatorname{Var}_p N_n} = \sqrt{np(1-p)}$. Autrement dit, F_n est bien un estimateur *consistant* de

p : $\mathbb{E}_p[F_n] = \frac{1}{n}\mathbb{E}_p[N_n] = p$. La loi forte des grands nombre nous assure que presque sûrement, quand n grandit, F_n tend vers p. Cela justifie la première réponse au problème : utiliser F_n pour estimer p est une bonne idée (on peut montrer que c'est la meilleure idée possible !). Pour autant, cette réponse n'est pas totalement satisfaisante.

2.2 Fluctuations autour de la moyenne

Car il est très important de se demander quelle est la qualité de cette estimation, c'est-à-dire de combien F_n diffère "normalement" de p pour une valeur donnée de n. Malheureusement, les distributions binomiales ne sont pas faciles à manipuler, ni de façon théorique ni de façon pratique. Pour résoudre cette difficulté, on a recours à une *approximation* asymptotique (donc pas parfaitement maîtrisée) mais très utile en pratique, ou bien à des *inégalités* (parfaitement sûres, mais souvent pessimistes).

2.2.1 Estimation asymptotique des flucuations

Le *Théorème Central Limite* (TCL) affirme que la suite de variables aléatoires $\left(\frac{N_n - np}{\sqrt{np(1-p)}}\right)_n$ converge en loi vers une variable Y de loi normale centrée réduite $\mathcal{N}(0,1)$: cela signifie exactement que pour tout $a < b$ on a $\mathbb{P}_p\left(a \leqslant \frac{N_n - np}{\sqrt{np(1-p)}} \leqslant b\right) \to \mathbb{P}_{\mathcal{N}(0,1)}(a \leqslant Y \leqslant b)$ quand n tend vers l'infini. Ainsi, pour tout $\epsilon > 0$ on a

$$\mathbb{P}_p(|F_n - p| > \epsilon) = \mathbb{P}_p\left(\left|\frac{N_n - np}{\sqrt{np(1-p)}}\right| > \epsilon\sqrt{\frac{n}{p(1-p)}}\right) \approx \mathbb{P}_{\mathcal{N}(0,1)}\left(|Y| > \epsilon\sqrt{\frac{n}{p(1-p)}}\right).$$

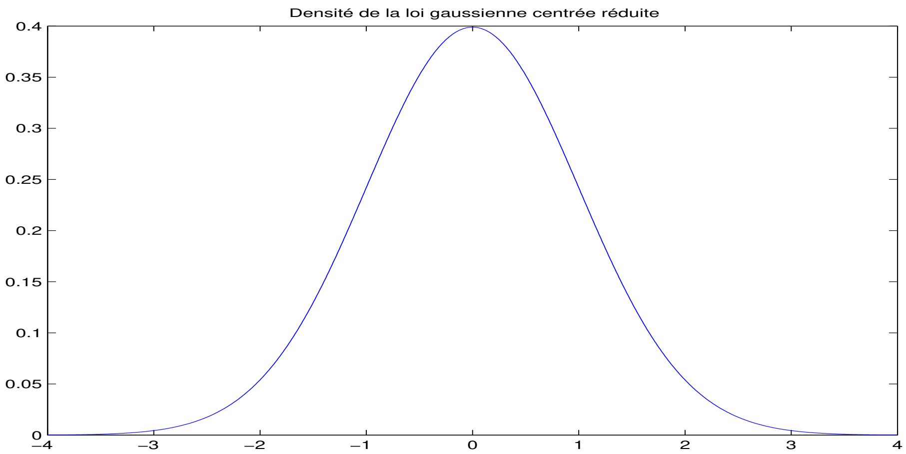
L'avantage, c'est que les probabilités pour une variable normale centrée réduite sont facilement accessibles. Notons u_α le réel positif tel que $\mathbb{P}_{\mathcal{N}(0,1)}(|Y| > \alpha) = \alpha$. On peut retenir que $u_{5\%} = 1.96$; les valeurs de u_α sont présentes dans toutes les tables de statistiques, et Maple permet de les retrouver numériquement.

Rappel : la densité de la loi normale centrée réduite est :

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

On peut évaluer numériquement la probabilité que Y soit compris entre a et b par la command Maple :

```
> evalf(int(f, a..b))
```



Expérience :	pour $n = 500$ et $p = 0.2$, calculer la probabilité que F_n soit inférieur à 0.17 ou supérieur à 0.23 d'abord exactement (avec les probabilités de loi binomiale), et de façon approchée avec le TCL.

2.2.2 Inégalités de concentration non asymptotiques

– L'*inégalité de Bienaymé-Chebyshev* (très simple à montrer, au programme commun) affirme que pour tout $n \in \mathbb{N}^*$ et pour tout $\epsilon > 0$:

$$\mathbb{P}_p(|F_n - p| > \epsilon) \leqslant \frac{\mathrm{Var}_p[F_n]}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2}.$$

– Une autre inégalité (qui n'est pas au programme commun de l'agrégation) attribuée à Höffding, valable seulement pour les variables aléatoires à valeur dans l'intervalle $[0,1]$, est plus précise quand n est grand :

$$\mathbb{P}_p(|F_n - p| > \epsilon) \leqslant 2e^{-2n\epsilon^2}.$$

On peut remarquer que la borne de Höffding a l'avantage (et l'inconvénient) d'être uniforme en p.

Expérience : que donnent ces inégalités pour pour $n = 500$, $p = 0.2$ et $\epsilon = 0.03$?

Toutes les inégalités de flucuation précédentes peuvent se lire de deux façons :
- probabiliste : "la probabilité que F soit loin de p est de l'ordre de ..."
- statistique : "la probabilité que p soit loin de F est de l'ordre de ..."

Nous allons voir que la première lecture permet de faire des *tests*, alors que la deuxième conduit à la construction d'*intervalles de confiance.*

2.3 Test statistique

Grâce à la maîtrise des fluctuations de F_n autour de son espérance p, on peut *tester l'hypothèse* $p = 0.2$. Le principe est simple :
- on suppose que $p = 0.2$, et on prédit grâce à cela où F_n va très probablement tomber ;
- on regarde si les observations sont conformes aux attentes ; sinon, on remet en cause notre hypothèse.

Remarquons tout de suite qu'on ne peut rien décider de façon certaine à partir des observations $X_1, \dots, X_n$: même si, par exemple, elles tombent toutes dans l'intervalle $[0.4, 0.6]$, on ne peut être *sûr* à 100% que le générateur n'est pas bon, car la probabilité de cet évènement est non nulle même pour un générateur parfait ! Toutefois, un tel évènement est très improbable : c'est cela que nous allons utiliser pour notre test.

Plus précisément, appelons H_0 l'hypothèse (dite *conservatrice* ou *nulle*) : "$p = 0.2$" et H_1 l'hypothèse (dite *alternative*) $p \neq 0.2$. On fixe un *risque* α, qui est la probabilité maximale que l'on s'autorise de rejeter H_0 alors qu'elle est vraie ; typiquement, on choisit $\alpha = 5\%$. Sous H_0, les inégalités de fluctuation précédentes permettent de prédire où tombe F_n avec très grande probabilité. On peut donc construire une *zone de rejet* ZR_n telle que $P_{H_0}[F_n \in ZR_n] \leqslant \alpha$. En fait, il existe de très nombreuses zones de rejet possibles : on choisit celle qui est la plus logique compte tenu de l'hypothèse alternative. Ici, si $p \neq 0.2$ on aura tendance à observer des valeurs de F_n significativement inférieures ou supérieure à 0.2 : on construit donc une zone de rejet de la forme $[0, 0.2 - \epsilon[\cup]0.2 + \epsilon, 1]$.
- En utilisant l'inégalité de Bienaymé-Chebyshev, on trouve
$$ZR_n^C = \left[0, 0.2 - \sqrt{\frac{0.2(1-0.2)}{n\alpha}}\right[\bigcup \left]0.2 + \sqrt{\frac{0.2(1-0.2)}{n\alpha}}, 1\right].$$
- En utilisant l'inégalité de Höffding, on trouve
$$ZR_n^H = \left[0, 0.2 - \sqrt{\frac{-\log\alpha}{2n}}\right[\bigcup \left]0.2 + \sqrt{\frac{-\log\alpha}{2n}}, 1\right].$$
- En utilisant le TCL, on trouve
$$ZR_n^a = \left[0, 0.2 - u_\alpha\sqrt{\frac{0.2\,(1-0.2)}{n}}\right[\bigcup \left]0.2 + u_\alpha\sqrt{\frac{0.2\,(1-0.2)}{n}}, 1\right].$$

Expérience : faire le test précédent au risque $\alpha = 5\%$ pour quelques générateurs de votre choix. Grâce à un générateur "sûr", calculer leur niveau réel, c'est-à-dire la vraie probabilité de rejet sous H_0.

2.4 Intervalle de confiance

D'après ce qui précède, le générateur `gen_q` est visiblement mauvais : on peut admettre qu'il renvoie des valeurs indépendantes les unes des autres, mais pas selon une loi uniforme sur l'intervalle $[0, 1]$. Le problème se pose alors d'*estimer* la probabilité p qu'il renvoie un nombre compris entre 0.4 et 0.6. Il est de nature statistique : nous ne connaissons pas p, nous voulons nous en faire une idée grâce à F. La réponse la plus satisfaisante du statisticien à ce problème est la notion d'*intervalle de confiance.* Elle suppose à nouveau de choisir un *niveau de risque* α, complémentaire du *niveau de confiance* $1 - \alpha$. On construit à partir des observations $X_1, \dots, X_n$ un intervalle $I_{n,\alpha} = I_{n,\alpha}(X_1, \dots, X_n)$ (qui est donc aléatoire) contenant p avec probabilité au moins $1 - \alpha$ *quelle que soit la valeur de* p : on dit que $I_{n,\alpha}$ est un intervalle de confiance de risque α pour p si
$$\forall p \in [0, 1], \mathbb{P}_p[p \in I_{n,\alpha}] \geqslant 1 - \alpha.$$

2.4.1 Intervalle de confiance non asymptotique

– De l'inégalité d'Höffding on déduit pour $\epsilon = \sqrt{\frac{-\log\alpha}{2n}}$:

$$\mathbb{P}\left[p \notin \left[F_n - \sqrt{\frac{-\log\alpha}{2n}}, F_n + \sqrt{\frac{-\log\alpha}{2n}}\right]\right] \leqslant \alpha.$$

On aboutit ainsi à l'intervalle de confiance :

$$I^H_{n,\alpha} = \left[F_n - \sqrt{\frac{-\log\alpha}{2n}}, F_n + \sqrt{\frac{-\log\alpha}{2n}}\right].$$

– De l'inégalité de Bienaymé-Chebyshev, on déduit pour $\epsilon = \sqrt{\frac{p(1-p)}{n\alpha}}$:

$$\mathbb{P}\left[p \notin \left[F_n - \sqrt{\frac{p(1-p)}{n\alpha}}, F_n + \sqrt{\frac{p(1-p)}{n\alpha}}\right]\right] \leqslant \alpha.$$

Pourtant, $\left[F_n - \sqrt{\frac{p(1-p)}{n\alpha}}, F_n + \sqrt{\frac{p(1-p)}{n\alpha}}\right]$ n'est pas un intervalle de confiance pour p : il dépend de p, qui est justement inconnu ! Toutefois, en remarquant que $p\,(1-p) \leqslant \frac{1}{4}$ pour tout $p \in [0,1]$, on aboutit à l'intervalle de confiance :

$$I^C_{n,\alpha} = \left[F_n - \sqrt{\frac{1}{4n\alpha}}, F_n + \sqrt{\frac{1}{4n\alpha}}\right].$$

2.4.2 Intervalle de confiance asymptotique

Grâce au TCL on peut écrire pour $\epsilon = u_\alpha\sqrt{\frac{p(1-p)}{n}}$:

$$\mathbb{P}\left(\notin \left[F_n - u_\alpha\sqrt{\frac{p\,(1-p)}{n}}, F_n + u_\alpha\sqrt{\frac{p\,(1-p)}{n}}\right]\right) \approx \alpha.$$

Là aussi, la borne dépend de p : il faut soit utiliser l'inégalité $p\,(1-p) \leqslant \frac{1}{4}$, soit (on n'est plus à une approximation près) remplacer p par F_n. On obtient donc l'*intervalle de confiance asymptotique* :

$$I^a_{n,\alpha} = \left[F_n - u_\alpha\sqrt{\frac{F_n\,(1-F_n)}{n}}, F_n + u_\alpha\sqrt{\frac{F_n\,(1-F_n)}{n}}\right].$$

Expérience :	construire les trois intervalles de confiance de niveau 95% pour la probabilité qu'un générateur pseudo-aléatoire renvoie un nombre compris entre 0.4 et 0.6. Répétez l'expérience de nombreuses fois pour estimer les niveaux réels sous l'hyptothèse que le générateur est bon (c'est-à-dire les vraies probabilités que les intervalles ne contiennent pas la valeur 0.2).

Exercice supplémentaire pour numériciens	soit I l'intersection des deux ellipses d'équation $x^2 + 2y^2 < 1$ et $2x^2 + y^2 < 1$, et A l'aire de I. Il est difficile de calculer A, même numériquement. Par contre, il est facile d'estimer A par la méthode de Monte Carlo : un point tiré au hasard uniformément dans le carré $[-1,1] \times [-1,1]$ a probabilité $A/4$ d'être dans I. En répétant l'expérience un grand nombre de fois, on peut donc construire une bonne estimation de A. Construire un intervalle de confiance pour A d'amplitude 2.10^{-2} et de niveau de confiance au moins 95%. Faire l'expérience. Comment améliorer le calcul ici ?

3 Test et intervalle de confiance sur la moyenne

Testons maintenant la moyenne des valeurs $(X_k)_k$ renvoyées par le générateur. On vérifiera d'abord que, si le générateur est bon, l'espérance de X_k est $\mu = \frac{1}{2}$ et sa variance est $\frac{1}{12}$. On définit la moyenne empirique des n premières observations : $\bar{X}_n = \frac{X_1+\ldots+X_n}{n}$. Sa loi est difficile à décrire exactement, mais on sait que son espérance est $\frac{1}{2}$ et sa variance $\frac{1}{12n}$. On peut donc, de façon analogue à ci-dessus, écrire des inégalités de déviation pour $\bar{X}_n$, et en déduire un test de l'hypothèse "$\mu = \frac{1}{2}$", et un intervalle de confiance pour μ.

Expérience :	montrer une inégalité de déviation pour $\bar{X}_n$ (parmi les trois suggérées ci-dessus), et mettre en œuvre le test sur μ qui en découle.

4 Distribution uniforme : test du χ^2

Le test précédent permet de vérifier que la probabilité pour notre générateur de renvoyer un nombre dans chaque intervalle $[a, b]$ est bien $b - a$. Mais exécuter plusieurs de ces tests à la fois ? Par exemple, comment tester que la probabilité de tomber dans l'intervalle $\left[\frac{j-1}{10}, \frac{j}{10}\right[$ est bien 0.1 pour tous les $j \in \{1, \ldots, 10\}$ à la fois ?

On peut penser à pratiquer un "test composé", en pratiquant une série de 10 tests successifs : c'est une très mauvaise idée (pourquoi ?). Il faut un test multinomial, qui compare la répartition empirique à la répartition attendue *dans son ensemble.*

Pour $j \in \{1, \ldots, 10\}$, notons p^j la probabilité que le générateur renvoie une valeur de l'intervalle $\left[\frac{j-1}{10}, \frac{j}{10}\right[$, et N_n^j le nombre de valeurs X_i tombées dans cet intervalle :

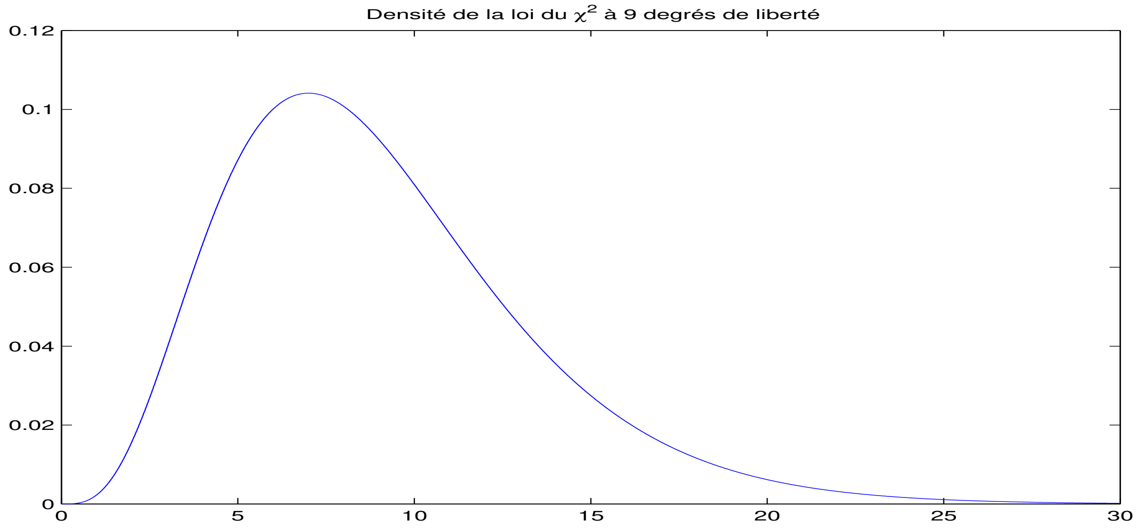
$$N_n^j = \#\left\{i \in \{1, \ldots, n\} : X_i \in \left[\frac{j-1}{10}, \frac{j}{10}\right[\right\}.$$

Le vecteur $\vec{N_n} = (N_n^j)_{1\leqslant j\leqslant 10}$ suit une loi multinomiale de paramètre $\vec{p} = (p^j)_{1\leqslant j\leqslant 10}$. Son espérance, par linéarité, vaut $(np^j)_{1\leqslant j\leqslant 10}$; mais bien entendu, les N_n^j ne sont pas indépendantes : leur somme vaut n. On veut tester l'hypothèse H_0 "$(p^j)_{1\leqslant j\leqslant 10} = \left(\frac{1}{10}, \ldots, \frac{1}{10}\right)$" contre l'hypothèse H_1 : "$\exists j \in \{1, \ldots, 10\}$ tq $p_j \neq \frac{1}{10}$". Comme précédemment, on va la rejeter si l'observation $\vec{N_n}$ est "trop loin" de la valeur attendue $\left(\frac{1}{10}, \ldots, \frac{1}{10}\right)$. On a donc besoin d'une notion de distance sur $\mathbb{R}^{10}$: on en connaît plusieurs, mais elle ne sont pas vraiment satisfaisantes (pourquoi ?). La "bonne" est la *(pseudo-)distance du* χ^2, définie pour $\vec{x}, \vec{y} \in \mathbb{R}^{10}$ par :

$$\chi^2(\vec{x}, \vec{y}) = \sum_{j=1}^{10} \frac{(x_j - y_j)^2}{x_j}.$$

Bien remarquer qu'elle n'est pas symétrique !

Ce qui justifie l'utilisation (et le nom) de cette distance, c'est le résultat statistique suivant : *quand n tend vers l'infini, la variable aléatoire $Z_n = \chi^2\left(n\vec{p}, \vec{N_n}\right)$ converge en loi vers une variable $\tilde{Z}$ qui suit une loi du χ^2 à* $10 - 1 = 9$ *degrés de liberté*. Grâce à une table statistique, on peut donc dire quelles sont les valeurs "acceptables" pour Z_n, et quelles sont les valeurs anormalement élevées. Ci-dessous est représenté le graphe de la densité de $\tilde{Z}$, ainsi que quelques valeurs de $P(\tilde{Z} \leqslant u)$ pour différents u :



u	2.088	3.325	4.168	8.343	14.684	16.92	21.67	27.88
$\mathbb{P}(\tilde{Z} \leqslant u)$	.01	0.05	0.1	0.5	0.9	0.95	0.99	0.999

D'autre part, sous H_1 la loi des grands nombres montrer que Z_n tend vers l'infini : on construit donc le test en rejetant les trop grandes valeurs de Z_n.

Expérience : Mettre en pratique le test du χ^2 présenté ici. Essayer avec d'autres découpages que $\bigcup_{i=1}^{10}\left[\frac{i-1}{10}, \frac{i}{10}\right[$.

Expérience : Proposer et mettre en œuvre un test du χ^2 pour les couples de deux valeurs successives renvoyées par un générateur pseudo-aléatoire. Qu'est-ce que de cela permet de tester en plus de l'uniforme répartition ?

5 Expérience supplémentaire : temps d'attente

Parmi les tests classiques des générateurs pseudo-aléatoire figure aussi le suivant : soit T le nombre d'appels qu'il faut faire au générateur avant de tomber dans l'intervalle $[0.4, 0.6]$. Si le générateur est bon, T est une variable aléatoire de loi géométrique $\mathcal{G}(p = 0.2)$, c'est-à-dire que pour tout entier positif k on a $\mathbb{P}(T = k) = p\,(1-p)^{k-1}$. Son espérance est $\mathbb{E}[T] = \frac{1}{p}$ et sa variance $\mathrm{Var}[T] = \frac{1-p}{p^2}$.

Expérience : Proposer et mettre en œuvre un test sur la moyenne de T. S'il est négatif, proposer un intervalle de confiance asymptotique pour cette moyenne.