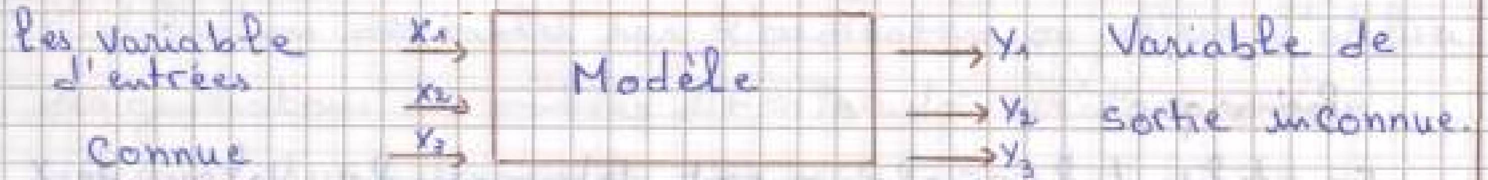
Chapitre 1 : Nombres aleataire et pseudo-aleatoire

1.1. Introduction :

Soit X une variable aleataire et F_X sa fonction de répartition $x_1, x_2, \dots, x_n$ les realisation de X. En simulant un système on a très souvent besoin de ces réalisations.



graphe 1 : Shéma representatif de la simulation.

Propriété :

si X est une variable aleataire de fonction de répartition F_X alors, $Y = F_X(x)$ suit une loi Uniforme.

Preuve :

si X est continue alors $F(x)$ est croissante et F_X^{-1} existe.

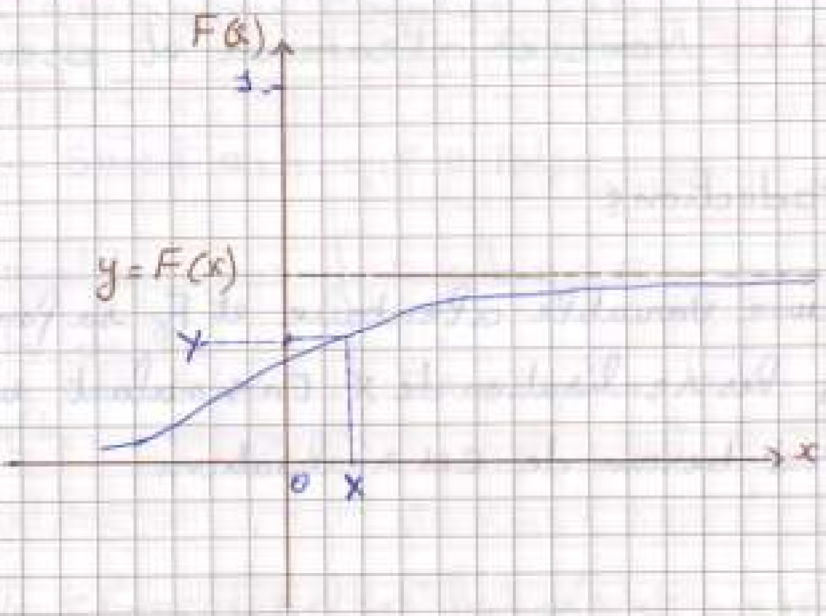
$$\text{Soit } y \in [0,1],\ F_Y(y) = P(Y \leq y) = P\left[F_X(x) \leq y\right]$$

$$= P\left(X \leq F_X^{-1}(y)\right)$$

$$= F_X\left(F_X^{-1}(y)\right) = y$$

$$\Rightarrow \forall y \in [0,1],\ F_Y(y) = y \Rightarrow Y = F_X(x) \rightsquigarrow U[0,1]$$

$$\Rightarrow y \rightsquigarrow U_{[0,1]} \Rightarrow F_X(x) \rightsquigarrow U_{[0,1]}$$



<u>Remarque</u>:

En utilisant le graphe on peut déterminer les valeurs de X à partir des valeurs de Y.

* Cette propriété permet de ramener la génération des réalisations de la variable aléatoire X de Fonction de répartition $F_X(x)$ à la génération des réalisations de la V.a $U_{[0,1]}$.

Le problème est donc de générer des nombres aléatoires Uniforme sur $[0, 1]$.

1.2. La génération des nombres aléatoires (ou au hasard) et les tables.

A l'origine, on utilise des tirages des boules dans des urnes numéroté de 0 à 9, des cartes ou Dés pour obtenir des tables de nombre aléatoires.

Ceci étant insuffisant, on a Construit des machines générant des nombres au hasard.

Exemple 01:

Les machines servant au tirage des numéros gagnant de la Loterie nationale.

Exemple 02:

Kendall et Babington-Smith, on obtenue 100.000 chiffres à partir d'un disque tournant divisé en secteurs multiples 10 et éclairé à intervalles régulier. Vu la lourdeur de l'utilisation des tables sur l'ordinateur on préfère d'utiliser des générateurs de nombres dit « Pseudo-aléatoire » qui génerant des nombres à la demande.

1.3. Génération des nombres pseudo-aléatoires:

Le terme " Pseudo " vien du fait qu'on construit par un procédé déterministe une suite de nombres aléatoire

$x_n = f(x_{n-1}, x_{n-2}, \dots, x_0)$

Si le nombre x_0 est donné (Initialisation) toute la suite $x_n, n \in \mathbb{N}$ est déterminée, mais on demande à la suite d'avoir l'aspect d'une suite de nombres au hasard.

Supposons que N soit la plus grande valeur que peut atteindre x_n, le problème est arrivée à obtenir des suites telle que:

1- Les x_n sont réparties uniformement sur $[0, N]$
2- Les x_n sont indépendants
3- L'algorithme de génération soit plus rapide possible.

1.3.1 : Les générateurs congruentiels :

Rappel : Soient $a, b, q, r \in \mathbb{N}$

$$a \equiv r \mod b \Leftrightarrow \begin{cases} a = bq + r \\ r = 0, 1, \dots, b-1 \end{cases}$$

r : est le reste de la division a/b

Méthode congruentielle linéaire et mixte :

$$\begin{cases} X_{n+1} = a X_n + C \pmod{m} & n \geq 0 \\ X_0 \text{ donné} \end{cases}$$

où a, C, m et X_0 sont des entiers tellque :

m : modulo $m > 0$

a : multiplicateur $0 \leq a < m$

c : incrément $0 \leq c \leq m$.

X_0 : la valeur de dépard $0 \leq x_0 < m$.

+ Dans le cas où $c = 0$: on obtient la méthode congruentielle multiplicative ou lineaire.

+ Dans le cas où $c \neq 0$: on obtient la méthode congruentielle mixte.

Exemple :

Supposons les constantes suivantes : $a = 3$, $C = 0$, $m = 5$ & $X_0 = 4$

le générateur des nombres pseudo-aléatoires est suivant :

$$\begin{cases} x_{n+1} = 3 x_n \mod 5 & n \geq 0 \\ x_0 = 4 \end{cases}$$

on calcule la suite des nombre pseudo-aléatoire

$x_1 = 3x_0 = 12 \mod 5 \Rightarrow x_1 = 2$

$x_2 = 3x_1 \mod 5 \Rightarrow x_2 = 1$.

$x_3 = 3x_2 \mod 5 \Rightarrow x_3 = 3$

$$\Rightarrow \begin{cases} x_0 = 4 \\ x_1 = 2 \\ x_2 = 1 \\ x_3 = 3 \end{cases}$$

Aucune valeur ne doit dépasser $m = 5$.

La suite de nombre pseudo-aléatoire (n.p.a) est $(4, 2, 1, 3, 4, 2, 1, 4)$

Remarque:

Pour obtenir une suite de nombres aléatoire entre 0 et 1 il suffit de prendre $u_n = \frac{x_n}{m}$, $n \geq 0$

Contre exemple:

Supposons les paramètres suivants: $a = c = x_0 = 1$ et $m = 3850$

le générateur:

$$x_{n+1} = x_n + 1 \mod 3850$$

la suite de n.p.a est $(1, 2, 3, 4, 5, \ldots)$

ce générateur n'a aucun intérêt car la suite générer n'a pas l'aspect du hasard. ce n'est pas un bon générateur bien que le paramètre est très élevé.

Remarque 02 :

Donnons la formule générale de x_n en fonction de x_0

$$x_n = a x_{n-1} + c \pmod m$$
$$= a(a x_{n-2} + c) + c \pmod m$$
$$= a^2 x_{n-2} + ac + c \pmod m$$
$$= a^2(a x_{n-3} + c) + ac + c \pmod m$$
$$= a^3 x_{n-3} + a^2 c + ac + c \pmod m$$
$$= a^3 x_{n-3} + c(a^2 + a^1 + a^0)$$
$$\vdots$$
$$x_n = a^n x_0 + c \sum_{i=0}^{n-1} a^i \pmod m$$

$$\boxed{x_n = a^n x_0 + c \,\frac{1 - a^n}{1 - a} \mod m.}$$

1 - 3 - 2 - Problème de paramètre :

Le problème est le choix des 4 paramètre a, c, m et x_0 pour que la suite des nombres générésant de bonne propriétés statistique (satisfie 1, 2, 3 de paragraphe 3).
Les paramètres « m » et « a » sont les plus importants, les paramètres « a » et « c » sont choisés pour que la période soit de longueur « m ». Certain auteurs affirme que la réalisation d'un bonne générateur exige de paramètre « m » être grand mais ceci n'est pas une condition nécessaire (voir Contre exemple)

En général, on dit qu'un générateur est bon si il engendre une bonne suite de nombre aléatoire c-a-d qu'il satisfait les propriétés statistique 1, 2, 3 de paragraphe 03.

1-4- Tests de générateurs des nombres pseudo-aléatoires:

Pour que les suites de nombres générés possèdent un comportement aléatoire sur [0, 1], il faut tester ce comportement selon deux critères:

a- Uniformité: Peut-on considérer les résultats d'une génération comme des réalisations d'une variable aléatoire Uniforme $U_{[0,1]}$?

b- Indépendance: Peut-on considérer les U_n comme des réalisations de variable aléatoire indépendantes?

1-4-1- Tests de l'uniformité de la distribution.

Comparaison d'une loi empirique (obtenue à partir des nombres générés entre [0, 1]) avec la loi théorique (loi uniforme [0, 1]).

Donc il s'agit de tester l'hypothèse H_0 selon laquelle les observations formées par le générateur sont bien réparties uniformément entre 0 et 1.

1-4-1-1: Test de Khi-deux :

Soient :

n : le nombre d'observation (le nombre pseudo-aléatoire générer)

k : le nombre de classe de l'intervalle $[0,1]$.

n_i : les effectifs observer dans la ième classe.

$np_i = n \times \frac{1}{k}$: l'effectifs théorique de chaque classe (loi uniforme)

p_i : la probabilité associe à la classe i.

on calcule le test statistique :

$$\boxed{\chi^2 = \frac{k}{n} \sum_{n=1}^{k} \left(n_i - \frac{n}{k}\right)^2 = \frac{1}{np_i} \sum_{i=1}^{k} \left(n_i - \frac{n}{k}\right)^2}$$

qui suit approximativement sous l'hypothèse H_0, une loi de Khi-deux à $(k-1)$ degrés de liberté. (χ^2_{k-1}).

On fixe un seuil de signification α et on rejette l'hypothèse H_0 si $\chi^2 >$ valeur critique (lue sur la table du $\chi^2_{(k-1)}$)

Remarque :

- Pour que ce test ait un sens il faut que $np_i \geq 5$ c'est à dire $n \geq 5k$.

- Si $k > 30$, on utilise l'approximation de Fisher de la loi de χ^2 par la loi normale.

- Pour le calcule de k, on peut utiliser les approximations suivantes:

* Approximation de Mann et Wald : $K = (4n)^{2/5}$
* Approximation de Sturges : $K = 1 + (3,3) \log n$.

Exemple:

Soit la suite des nombres aléatoire suivante sur [0,1]:

0,99 - 0,01 - 0,03 - 0,05 - 0,89 - 0,91 - 0,19 - 0,33 - 0,36 - 0,45 - 0,80 - 0,70 - 0,75 - 0,20 - 0,31 - 0,23 - 0,65 - 0,35 - 0,43 - 0,42 - 0,62 - 0,79 - 0,11 - 0,15 - 0,21.

- La taille de l'echantillon $n = 25$
- Le nombre de classe on le prend $K = 4$
- l'hypothèse H_0 : " les obervation suivent une loi $U_{[0,1]}$ "
- Le seuil de signification $\alpha = 0,05$

La réponse:

Classe	[0 - 0,25[	[0,25 - 0,5[	[0,5 - 0,75[	[0,75 - 1[	Total
n_i	9	7	3	6	25
$\frac{n}{k} = np_i$	6,25	6,25	6,25	6,25	25
$(n_i - \frac{n}{k})^2$	7,5625	0,5625	10,5675	0,0625	18,75.

$$\chi^2 = \frac{k}{n} \sum_{i=1}^{k} \left(n_i - \frac{n}{k}\right)^2 = \frac{4 \times 18,75}{25} = 3$$

Pour $\alpha = 0,05$, la valeur critique lue sur la table χ^2_3 est $\chi^2_3(0,05) = 7,815$ on a $\chi^2 = 3 < 7,815$

Decision: on accepte l'hypothèse H_0

Donc les observations suivent la loi uniforme $[0,1]$.

1.4.1.2. Test de kolmogorove - smirnov:

Ce Test est plus puissant que le précedent car c'est celui pour lequel le risque d'accepter H_0 est plus faible.

La procedure est la suivante:

1. on tire un échantillon de n observations à l'aide du générateur.

2. on classe les observations en ordre croissant.

3. on compare la fonction de répartition empirique $F_n(x)$ calculé à partir de ces n nombre pseudo-aléatoire avec la fonction de répartition théorique $F(X)$ de la loi $U_{[0,1]}$.

on calcule la statistique:

$$D = \mathrm{Max}\,|F_n(x) - F(x)| = \max D(X_i)$$

où

$$F_n(x) = \frac{\text{nombre d'observation} \leq X}{\text{taille de l'echantillon } n} \quad \text{et } F(x) = x$$

On fixe un seuil de signification α et on compare cet écart D à des valeurs critiques particuliers qu'on note D_n (où n est la taille de l'échantillon) obtenus à partir de la table de Kolmogorov-Smirnov.

La décision sera :

+ Accepter H_0, si $D < D_n$: c-à-d le générateur est bon.
+ Rejeter H_0, sinon : c-à-d le générateur n'a pas les qualités requises.

Exemple :

Soit un échantillon de taille n = 4 d'une population

$X_i = 0,34 - 0,49 - 0,56 - 0,70$.

Tester l'hypothèse selon laquelle la distribution de la population suit une loi uniforme sur $[0,1]$, en prenant un seuil de signification $\alpha = 0,2$, $\forall x \in [0,1]$,

$F(x) = x$ loi théorique.

X_i	0,34	0,49	0,56	0,70
$F_n(X_i)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{3}{4}$	1
$F(X_i)$	0,34	0,49	0,56	0,70
$D(X_i)$	0,09	0,01	0,19	0,30

On déduit le test statistique de KS : $D = \max D(X_i) = 0,30$

Pour $\alpha = 0,2$, la valeur critique lue sur la table de K-S est $D_4 = 0,494$

Comme $D = 0,30 < D_4 = 0,494$ on accepte l'hypothèse H_0

1-4-1-3- Test de l'histogramme :

Soit n nombre pseudo-aléatoire générer sur $[0,1]$
soit k le nombre de classe sur $[0,1]$

n_i : l'effectif observé dans la classe i

On divise l'intervalle $[0,1]$ en k classes et on trace l'histogramme de la serie $(x_i, k\frac{n_i}{n})$ et si l'uniformité est parfaite, on aura alors $n_i \simeq \frac{n}{k} \Rightarrow k\frac{n_i}{n} \simeq 1$

Ce teste n'est utilisé si l'uniformité était parfaite.

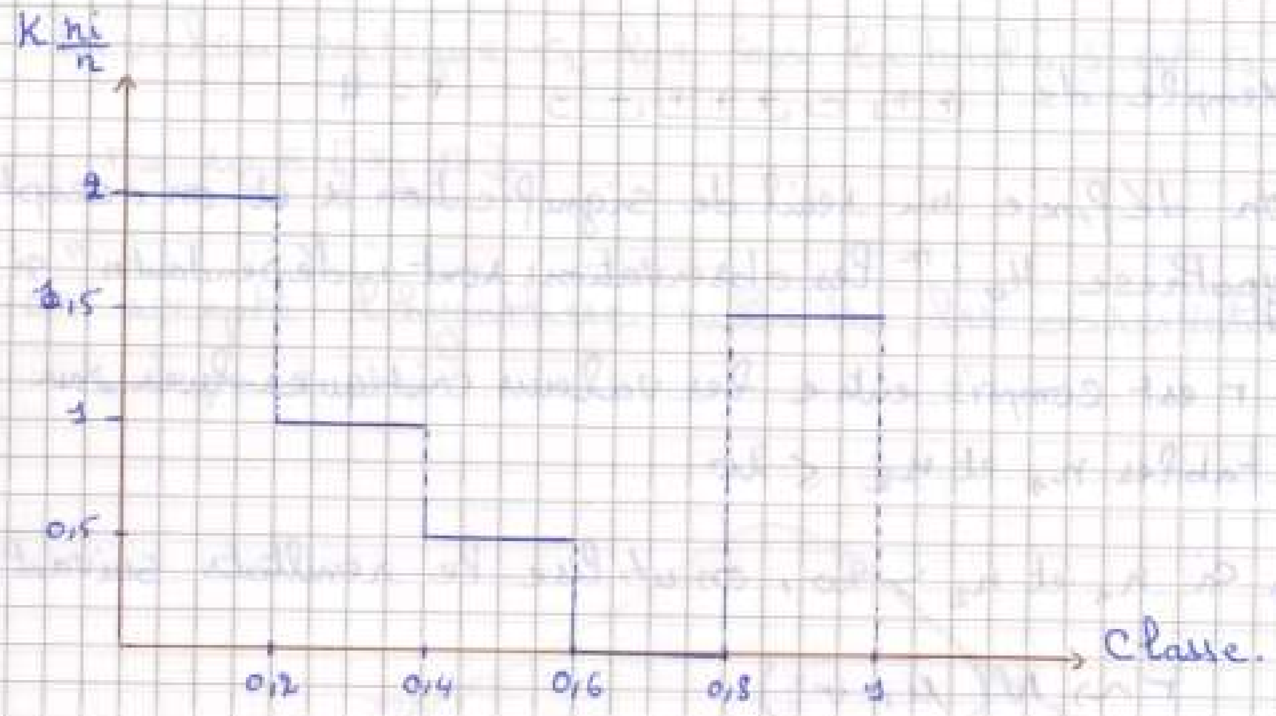
Exemple :

Soit un échantillon de la taille $n = 10$ (exemple du test χ^2) on utilise l'approximation de Man et Wald, on trouve :

$k = 4,37 \simeq 5$

$k = (4.10)^{2/5} = 4,37.$

Classe	$[0 - 0,2[$	$[0,2 - 0,4[$	$[0,4 - 0,6[$	$[0,6 - 0,8[$	$[0,8 - 1[$
n_i	4	2	1	0	3
$k\frac{n_i}{n}$	2	1	0,5	0	1,5



<u>Remarque:</u>

D'aprés l'histogramme, on ne peut pas décider avec ce test il faut faire d'autre tests.

1-4-2- Test de l'indépendance (test de sequences ou « runs » test.

Ce test est utilisé pour tester si les observations sont mutuellement indépendantes.

- on considère les données tell qu'ils sont collectées l'echantillon doit être diviser en deux classes.

soient n_1 et n_2 le nombre d'observations dans la classe 1 et 2 respectivement.

On registre (+) si l'observation est de la classe 1 et un (-) si l'observation est de la classe 2

- On compte le nombre de sequences homogènes de (+) et (-)

Exemple : ds : $\underbrace{++}\ \underbrace{-}\ \underbrace{+++}\ \underbrace{--}$ $\qquad r = 4$

On définie un seuil de signification α et on accepte l'hypothèse H_0 : " les observations sont indépendantes " si :

* r est compris entre les valeurs critiques lues sur les tables n_1 et $n_2 \leq 20$

* Si n_1 et $n_2 > 20$, on utilise le resultats suivant :

$$r \leadsto N(\mu, \sigma^2)$$

où $\quad \mu = \dfrac{2n_1n_2}{n} + 1 \quad$ et $\quad \sigma^2 = \dfrac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{n^2(n-1)}$

<u>Exemple</u> :

- On définie la classe 1 (+) par les nbres $\leq 0{,}5$
- On définie la classe 2 (−) par les nbres $> 0{,}5$

<u>Réponse</u>.

$n = 25$

$\left.\begin{array}{l} n_1 = 16 \\ n_2 = 9 \end{array}\right\} \; n_1 + n_2 = n$

$\overset{-}{0{,}99} - \overset{+}{0{,}01} - \overset{+}{0{,}03} - \overset{+}{0{,}05} - \overset{-}{0{,}89} - \overset{-}{0{,}91} - \overset{+}{0{,}19} - \overset{+}{0{,}33} - \overset{+}{0{,}36} -$
$\overset{+}{0{,}45} - \overset{-}{0{,}80} - \overset{-}{0{,}70} - \overset{-}{0{,}75} - \overset{+}{0{,}20} - \overset{+}{0{,}31} - \overset{+}{0{,}23} - \overset{-}{0{,}65} - \overset{+}{0{,}35}$
$\overset{+}{0{,}43} - \overset{+}{0{,}42} - \overset{-}{0{,}62} - \overset{-}{0{,}79} - \overset{+}{0{,}11} - \overset{+}{0{,}15} - \overset{+}{0{,}21}$

r = nombre de séquence = 10

test bilateral pour un seuil de signification $\alpha = 0{,}05$ la valeur critique r_1 lue sur la table 4 est $r_1 = 7$

La valeur critique r_2 lue sur la table 5 est $r_2 = 18$

$r = 10 \in [7, 18]$

on accepte l'hypothèse nulle H_0, les observations sont independantes.