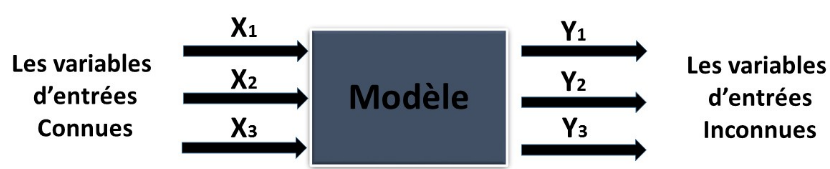
Chapitre 1 : Nombres aléatoires et pseudo-aléatoires

1 Introduction

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



Graphe 1 : Schéma représentatif de la simulation

Propriété

Si X est une variable aléatoire de fonction de répartition F_X, alors $Y = F_X(X)$ suit une loi uniforme.

Preuve détaillée

1. **Variable aléatoire continue :** Supposons que X soit une variable aléatoire continue avec une fonction de répartition cumulative (CDF) $F_X(x)$. La fonction $F_X(x)$ est croissante et continue sur son domaine.

2. **Transformation avec la CDF :** On définit une nouvelle variable aléatoire Y telle que :

$$Y = F_X(X)$$

3. **Calcul de la probabilité de Y :** Nous voulons montrer que Y suit une distribution uniforme sur l'intervalle $[0, 1]$. Pour ce faire, nous devons prouver que :

$$P(Y \leq y) = y \quad \text{pour tout} \quad y \in [0, 1]$$

4. **Utilisation de la CDF inverse :** Soit $y \in [0, 1]$. Nous avons :

$$P(Y \leq y) = P(F_X(X) \leq y)$$

Comme F_X est croissante, $F_X(X) \leq y$ si et seulement si $X \leq F_X^{-1}(y)$, où F_X^{-1} est la fonction inverse de la CDF. Donc :

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

5. **Application de la CDF :** Par définition de la CDF, nous avons :

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

Comme F_X est continue et croissante, $F_X(F_X^{-1}(y)) = y$. Donc :

$$P(Y \leq y) = y$$

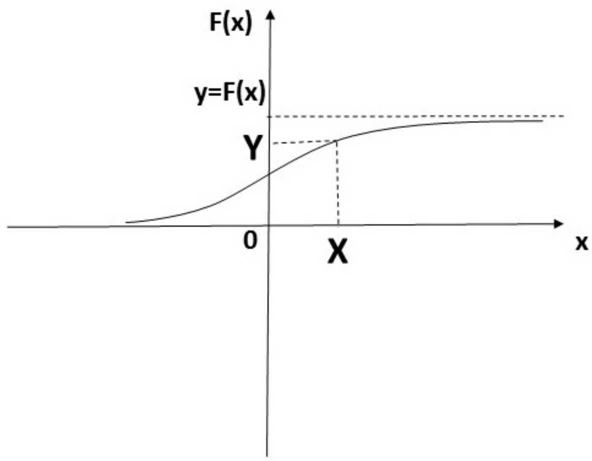
6. **Conclusion :** Comme $P(Y \leq y) = y$ pour tout $y \in [0, 1]$, cela signifie que Y suit une distribution uniforme sur $[0, 1]$.

Résumé de la Preuve

$$\begin{aligned} \forall y \ \in \ [0,1], \quad & P(Y \leq y) = P(F_X(X) \leq y) \\ = \ & P(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y \end{aligned}$$

Ce qui prouve que :

$$Y = F_X(X) \sim U[0, 1]$$



Remarque

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 fonctions de répartitions $F(x)$ à la génération des réalisations de la V.a $U[0,1]$. Le problème est donc de générer des nombres aléatoires Uniforme sur $[0,1]$.

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

À l'origine, on utilise des tirages des boules dans des urnes numérotées de 0 à 9, des cartes ou des dés pour obtenir des tables de nombres aléatoires. Ceci étant insuffisant, on a construit des machines générant des nombres au hasard.

Exemple 01

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

Exemple 02

Kendall et Babington-Smith, on obtient 100.000 chiffres à partir d'un disque tournant divisé en secteurs multiples de 10 et éclairé à intervalles réguliers. Vu la lourdeur de l'utilisation des tables sur l'ordinateur, on préfère utiliser des générateurs de nombres dits "Pseudo-aléatoires" qui génèrent des nombres à la demande.

3 Génération des nombres pseudo-aléatoires

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

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

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 d'obtenir des suites telles que :

1. Les x_n sont répartis uniformément sur $[0, N]$
2. Les x_n sont indépendants

3. L'algorithme de génération soit le plus rapide possible.

3.1 Les générateurs congruentiels

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

$$a \equiv r \pmod{b} \Longleftrightarrow \begin{cases} a = bq + r \\ r = 0, 1, \ldots, b-1 \end{cases}$$

r est le reste de la division a/b.
Méthode congruentielle linéaire et mixte :

$$X_{n+1} = aX_n + c \pmod{m}, \quad n \geq 0$$

$$X_0 \text{ donné}$$

où a, c, m et X_0 sont des entiers tels que :
- m : module $m > 0$
- a : multiplicateur $0 \leq a < m$
- c : incrément $0 \leq c < m$
- X_0 : la valeur de départ $0 \leq X_0 < m$.

– Dans le cas où $c = 0$, on obtient la méthode congruentielle multiplicative ou linéaire.
– 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 :

$$X_{n+1} = 3X_n \pmod 5, \quad n \geq 0$$

$$X_0 = 4$$

On calcule la suite des nombres pseudo-aléatoires

$$x_{n+1} = 3x_n \mod 5 \Rightarrow x_1 = 2$$

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

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

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

Aucune valeur ne doit dépasser $m = 5$.

La suite de nombres pseudo-aléatoires (n.p.a) est (4, 2, 1, 3, 4, 2, 1, 3, 4)

Remarque : Pour obtenir une suite de nombres aléatoires entre 0 et 1 il suffit de prendre $U_n = \frac{x_n}{m}$, $n \geq 0$.

Contre exemple : Supposons les paramètres suivants : $a = 4$, $x_0 = 1$ et $m = 3850$

$$x_{n+1} = x_n + 4 \mod 3850$$

La suite de n.p.a est (1, 5, 9, 13, ...)

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

Remarque 2 :

Donnons la formule générale de x_n en fonction de x_0 :

$$x_n = ax_{n-1} + c \mod m$$

$$= a(ax_{n-2} + c) + c \mod m$$

$$= a^2x_{n-2} + ac + c \mod m$$

$$= a^2(ax_{n-3} + c) + ac + c \mod m$$

$$= a^3x_{n-3} + a^2c + ac + c \mod m$$

$$= a^3x_{n-3} + c(a^2 + a + a^0)$$

$$x_n = a^nx_0 + c\sum_{i=0}^{n-1} a^i \mod m$$

$$x_n = a^nx_0 + c\frac{1-a^n}{1-a} \mod m$$

3.2 Problème de paramètre

Le problème est le choix des 4 paramètres a, c, m et x_0 pour que la suite des nombres générés ait de bonnes propriétés statistiques (satisfie 1, 2, 3 de paragraphe 3). Les paramètres a et m sont les plus importants, les paramètres c et x_0 sont choisis pour que la période soit de longueur m. Certains auteurs affirment que la réalisation d'un bon générateur exige des paramètres m soit étre grand mais ceci n'est pas une condition nécessaire (Voir Contre exemple).

En général, on dit qu'un générateur est bon s'il engendre une bonne suite de nombres aléatoires, c'est-à-dire qu'il satisfait les propriétés statistiques 1, 2, 3 du paragraphe 3.

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_i comme des réalisations de variables aléatoires indépendantes ?

4.1 Tests de l'uniformité de la distribution

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

Donc, il s'agit de tests d'hypothèse H_0, selon laquelle les observations faites par le générateur sont bien réparties uniformément entre 0 et 1.

4.1.1 Test de Khi-deux

Soient :

- n : le nombre d'observations (le nombre pseudo-aléatoire généré)
- K : le nombre de classes de l'intervalle [0,1]
- n_i : les effectifs observés dans la i-ème classe.
- np_i : $n \times p_i$ les effectifs théoriques de chaque classe (loi uniforme)
- p_i : la probabilité associée à la classe i.

On calcule le test statistique :

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

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 significativité et on rejette l'hypothèse H_0 si $\chi^2 >$ Valeur critique (lue sur la table de χ^2_{K-1}).

Remarques :

– 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 calcul de K, on peut utiliser les approximations suivantes :

– Approximation de Mann et Wald : $K = (4n)^{3/5} \textit{ou bien} \sqrt{n}$
– Approximation de Sturges : $K = 1 + (3,3)\log n$

Exemple :

Soit la suite des nombres aléatoires 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'échantillon $n = 25$
Hypothèse H_0 : Les observations suivent une loi $U_{[0,1]}$
Le seuil de signification $\alpha = 0.05$

La réponse détaillée :

Détermination du nombre de classes K :

La méthode de Mann-Whitney-Wald suggère que le nombre de classes K peut être déterminé par :

$$K = \sqrt{n}$$

Pour $n = 25$:

$$K = \sqrt{25} = 5$$

Nous diviserons donc l'intervalle $[0, 1]$ en 5 classes d'égale longueur :

$$[0, 0.2[, [0.2, 0.4[, [0.4, 0.6[, [0.6, 0.8[, [0.8, 1]$$

1. **Classes et Comptage :**

Nous allons distribuer les 25 observations dans les 5 classes définies :

Classe	Observations
$[0, 0.2[$	$0.01, 0.03, 0.05, 0.11, 0.15, 0.19$
$[0.2, 0.4[$	$0.20, 0.21, 0.23, 0.31, 0.33, 0.35, 0.36$
$[0.4, 0.6[$	$0.42, 0.43, 0.45$
$[0.6, 0.8[$	$0.62, 0.65, 0.70, 0.75, 0.79$
$[0.8, 1]$	$0.80, 0.89, 0.91, 0.99$

2. **Comptage des observations dans chaque classe :**

Classe	$[0, 0.2[$	$[0.2, 0.4[$	$[0.4, 0.6[$	$[0.6, 0.8[$	$[0.8, 1]$
n_i	6	7	3	5	4

3. **Calcul des effectifs théoriques** $\frac{n}{K} = np_i$:

Classe	$[0, 0.2[$	$[0.2, 0.4[$	$[0.4, 0.6[$	$[0.6, 0.8[$	$[0.8, 1]$	Total
n_i	6	7	3	5	4	25
$\frac{n}{K} = np_i$	5	5	5	5	5	25
$(n_i - \frac{n}{K})^2$	1	4	4	0	1	10

4. **Calcul de la statistique de Khi-deux** :

$$\chi^2 = \frac{K}{n} \sum_{i=1}^{K} (n_i - \frac{n}{K})^2 = \frac{5}{25} \times 10 = 2$$

5. **Décision** :

Pour $\alpha = 0.05$, la valeur critique lue sur la table χ^2_4 (degrés de liberté $K - 1$) est $\chi^2_4(0.05) = 9.488$, on a $\chi^2 = 2 < 9.488$.

Par conséquent, on accepte l'hypothèse H_0.

Conclusion :

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

4.1.2 Test de Kolmogorov-Smirnov

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

La procédure 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 $F_n(x)$ calculée à partir de ces n nombres pseudo-aléatoires avec la fonction de répartition théorique $F(x)$ de la loi $U[0, 1]$.

On calcule la statistique :

$$D = \max |F_n(x) - F(x)| = \max D(x_i)$$

où $F_n(x) = \frac{\text{nombre d'observations} \leq X}{\text{taille de l'échantillon } n}$ et $F(x) = x$

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

La décision sera : - Accepter H_0, si $D < D_n$ c'est-à-dire le générateur est bon. - Rejeter H_0, sinon c'est-à-dire 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 $\alpha = 0,05$.

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 Kolmogorov-Smirnov : $D = \max D(X_i) = 0,30$

Pour $\alpha = 0,05$, la valeur critique lue sur la table de K-S est $D_4 = 0,624$.

Comme $D = 0,30 < D_4 = 0,624$, on accepte l'hypothèse H_0.

4.1.3 Test de l'histogramme

Soit n nombre pseudo-aléatoire généré sur $[0,1]$, soit K le nombre de classes 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 série $(x_i, K\frac{n_i}{n})$ et si l'uniformité est parfaite, on aura alors $n_i \approx \frac{n}{K} \Rightarrow K\frac{n_i}{n} \approx 1$.

Ce test n'est utilisé que si l'uniformité est parfaite.

Exemple :

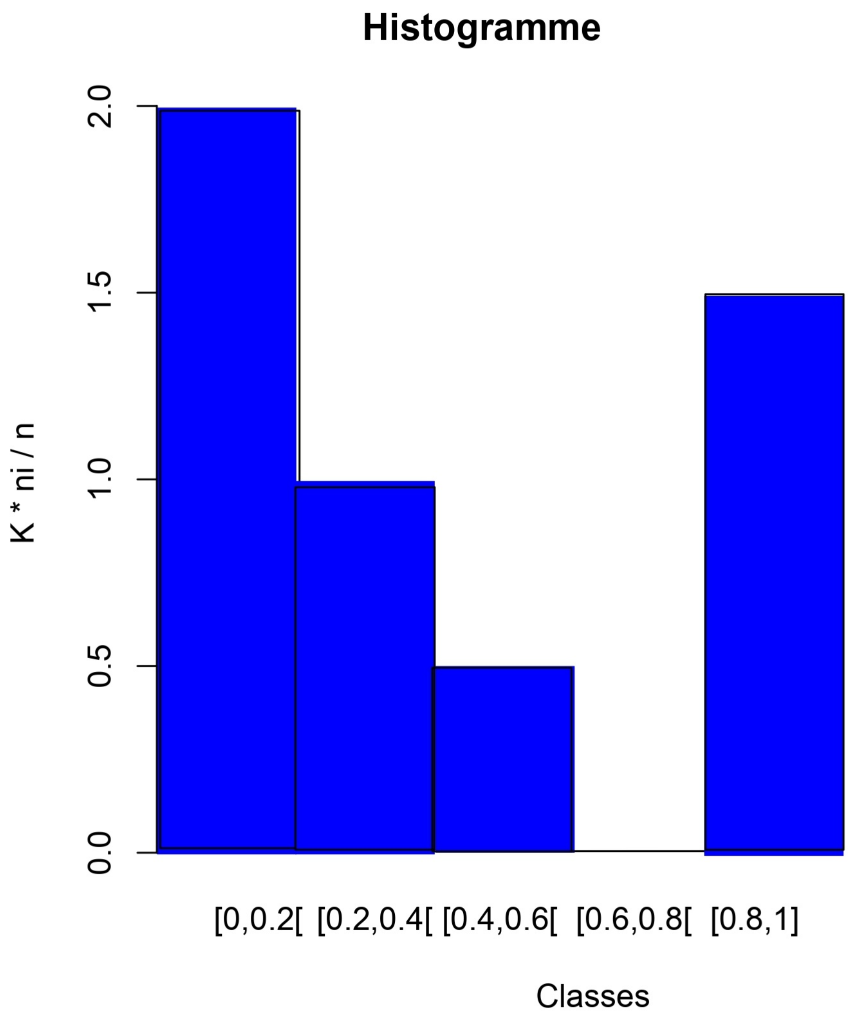
Soit un échantillon de taille $n = 10$ (exemple du test χ^2), on utilise l'approximation de Mann et Wald, on trouve :

$$K = 4,37 \approx 5$$

$$K = (4n)^{\frac{3}{5}} = 4,37$$

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

TAB. 1 – Tableau des classes pour le test de l'histogramme



Remarque : D'après l'histogramme, on ne peut pas décider avec ce test. Il faut faire d'autres tests.

4.2 Test de l'indépendance (test de séquences ou rans test)

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

- On considère les données telles qu'elles sont collectées.

L'échantillon doit être divisé en deux classes. Savoir n_1 et n_2 le nombre d'observations dans la classe 1 et 2 respectivement.

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

- On compte le nombre de séquences homogènes de $(+)$ et $(-)$.

Exemple :

(+ +)(-)(+ + +) (- -) r = 4

On définit 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 résultat suivant :

$$F_{n_1 n_2} \sim N(\mu, \sigma^2)$$

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

Exemple :

– On définit la classe 1 (+) par les nombres $\leq 0,5$

– On définit la classe 2 (-) par les nombres $> 0,5$

Réponse :

$n = 25$
$n_1 = 16$
$n_2 = 9$
$n_1 + n_2 = n$

$$0,99^-, 0,01^+, 0,03^+, 0,05^+, 0,89^-, 0,91^-, 0,19^+, 0,33^+, 0,36^+, 0,45^+$$
$$0,80^-, 0,70^-, 0,75^-, 0,20^+, 0,31^+, 0,23^+, 0,65^-, 0.35^+, 0,43^+, 0,42^+$$
$$0,62^-, 0,79^-, 0,11^+, 0,15^+, 0,21^+$$

r = nombre de séquences = 10
Test bilatéral pour un seuil de signification $\alpha = 0,05$
La valeur critique r_1 lue sur la table 1 est $r_1 = 7$
La valeur critique r_2, lue sur la table 2 est $r_2 = 18$
$r = 10 \in [7, 18]$
On accepte l'hypothèse nulle H_0, les observations sont indépendantes.