UNIVERSITÉ DE VERSAILLES ST-QUENTIN-EN-YVELINES

université PARIS-SACLAY

LES SIMULATIONS : APPLICATIONS A L'ECONOMETRIE

Réalisé par Wassim CHAFEI et Julie LAVAINE, étudiants en L3MIASHS à l'UVSQ au cours de l'année universitaire 2019-2020.
Enseignant-chercheur tuteur de ce projet : Fabrice BARTHELEMY.

Table des matières

Résumé explicatif 1

Illustration de la loi des grands nombres 2

Théoriquement, la méthode par inversion 2

Simulation d'une variable de Bernouilli 3

Exemple Scilab 3

Exemple Excel 3

Simulation d'une loi exponentielle 4

Exemple Scilab 5

Illustration du théorème central limite 6

Théoriquement, le théorème central limite 6

Applications – Approximation de la loi binomiale par la loi normale 7

Applications – Approximation de la loi de poisson par la loi normale 7

Montrons comment la loi se concentre autour de l'espérance lorsque n augmente. 8

Cadre d'analyse – Exemple avec une pièce de monnaie à deux faces : « Pile » et « Face » 9

Estimation d'une loi d'un estimateur en économétrie 10

Théoriquement, le test de Student 10

Théoriquement, le test de Fisher 11

Applications – Tester les contraintes 12

Applications – Test de nullité du R^2 12

Etapes à suivre afin de simuler une loi à partir du test de Fisher 13

Le modèle de régression linéaire simple : $y = ax + b + \varepsilon$ 13

Objectif de ce modèle 13

Principe de l'ajustement des moindres carrés 14

Les estimateurs des moindres carrés a et b 14

Exemple Excel – Application sur le prix des logements à Paris 15

En fonction de la taille de l'échantillon – Test de Student 16

Simulations 18

Exemple avec $n = 10$ 21

Conclusion 22

Bibliographie 23

Résumé explicatif

"L'économétrie est un outil d'analyse quantitative, permettant de vérifier l'existence de certaines relations entre des phénomènes économiques et de mesurer concrètement ces relations sur la base d'observations de faits réels"

ÉRIC DOR, "ECONOMETRIE"

L'économétrie est née dans les années 30, et est devenue un support très important pour de multiples domaines. Elle mobilise un ensemble important de méthodes numériques et statistiques et requiert d'outils informatiques de calculs et de services de gestion de données.

Généralement, les données utilisées par cette discipline sont des séries chronologiques de périodicité annuelle, trimestrielle ou mensuelle. Plusieurs outils informatiques permettent d'effectuer les calculs nécessaires, tels que : APACHE (Logiciel d'économétrie), EAS (Econometric Analysis System), SAS (Statistical Analysis System), SPSS (Statistical Package for the Social Science), MGVM (Module de Gestion de Vecteurs et de Matrices), Excel…

Depuis longtemps, les moyens mathématiques mis à disposition n'étaient pas suffisants, ou encore inadéquats, contraignant les chercheurs à déformer les modèles de la réalité. L'avancé technologique a permis de réaliser d'énormes progrès. L'analyse de la régression linéaire, l'analyse des tableaux, la programmation linéaire et non linéaire témoigne cette évolution. Ces méthodes de simulations seront nommées "simulations économétriques" et sont également appelées "modèles" ; elles permettent d'imiter le monde réel en employant des méthodes informatiques.

C'est grâce à l'arrivé des ordinateurs à grande échelle que les études réalistes des modèles de systèmes économiques et complexes ont été plus simples à simuler. Les unités informatiques imitent les fonctions de l'objet et permettent de représenter au mieux la fidélité du modèle. Il est possible d'effectuer des expériences sur tous types de modèle, néanmoins, il est impossible de les réaliser dans le cas d'un système réel. La statistique envahit pratiquement tous les domaines d'application, aucun n'en est exclu ; elle permet d'explorer et d'analyser des bases de données de plus en plus volumineuses notamment grâce à l'apparition du big data et du data mining.

Les domaines de l'automobile, de l'immobilier, du commerce, de l'industrie, de la santé, de la sécurité illustrent certains domaines d'application de ces simulations. Actuellement, nous traversons une crise sanitaire sans précédent et les épidémiologistes usent aussi des simulations. En effet, l'épidémiologiste effectue toutes les recherches visant à mieux comprendre et maîtriser les mécanismes de propagation des maladies contagieuses, les facteurs déclenchants, la fréquence des crises, l'évolution du mal jusqu'à l'état de pandémie. Il permet aussi de mettre au point les mesures sanitaires pour prévenir les maladies, les contrôler et les enrayer.

Illustration de la loi des grands nombres

La « loi des grands nombres » est l'un des théorèmes fondamentaux de la théorie des probabilités, formalisée au XVIIe siècle lors de la découverte de nouveaux langages mathématiques. Cette loi indique que lorsque l'on fait un tirage à caractère aléatoire dans une série de grande taille, plus on augmente la taille de l'échantillon, plus les caractéristiques statistiques de l'échantillon se rapprocheront des caractéristiques de la population étudiée. L'espérance est présentée comme une moyenne lors de l'application de cette loi. Elle permet d'interpréter la probabilité comme une fréquence de réalisation, un des domaines d'application de cette loi est les sondages. En effet, le principe est d'interroger un grand nombre d'individus pour connaître l'orientation de l'opinion de la population entière. Un autre domaine d'application de cette loi est pour le monde assurantiel où la loi permet aux assureurs de déterminer les probabilités que les sinistres dont ils sont garants se réaliseront ou non.

Plusieurs théorèmes expriment cette loi, pour différents types de convergence en théorie des probabilités. La loi faible des grands nombres met en évidence une convergence en probabilité, tandis que la loi forte des grands nombres donne une convergence presque sûre.

Théoriquement, la méthode par inversion

Pour illustrer le principe des simulations au travers de la loi des grands nombres, nous allons utiliser la méthode par inversion de la fonction de répartition appliqué à une loi uniforme.

Soit X une variable aléatoire réelle de fonction de répartition F définie par :

$$F(x) = P(X \leq x)\,, \forall x \in R$$

Il nous suffit de définir son inverse généralisé :

$$F^{-1}(u) = \inf\{x | F(x) \geq u\}\,, si\ u \in [0,1]$$

Pour réaliser cette simulation, nous allons vérifier la propriété suivante. Si une variable aléatoire réelle U suit une loi uniforme sur $[0,1]$, alors $F^{-1}(U)$ a la même loi que la variable aléatoire réelle X et a donc pour fonction de répartition F. En effet, on obtient :

$$P(X \leq u) = P(F^{-1}(U) \leq x) = P\big(U \leq F(x)\big) = F(x)$$

Ainsi, si on calcule son inverse généralisé, on peut simuler la variable aléatoire réelle X à partir de celle de U.

Nous allons aborder deux cas de simulation ayant recours à la méthode par inversion de la fonction de répartition, la simulation d'une variable de Bernoulli et la simulation d'une loi exponentielle.

Simulation d'une variable de Bernouilli

Soit une variable de Bernoulli de paramètre p avec $\mathrm{X} \in \{0, 1\}$, et de loi :

$$\mathrm{p} = \mathrm{P}(\mathrm{X} = 1) = 1 - \mathrm{P}(\mathrm{X} = 0)$$

Soit une variable aléatoire réelle $\mathrm{U} \sim \mathrm{U}([0, 1])$ tirée au hasard, si $\mathrm{U} < (1 - \mathrm{p})$ on retient la valeur $\mathrm{X} = 0$ sinon on retient $\mathrm{X} = 1$.

Exemple Scilab

Nous pouvons simuler le principe de la loi de Bernoulli grâce un exemple sur le logiciel Scilab.

Le programme suivant permet de représenter le principe d'une loi de Bernoulli :

```
//Simulation loi de Bernoulli
//Supposons que la piece est équilibré et que la proba de réussir est de 1/2

p = 1/2 //Probalité de réussir
u = rand()
if u < p then //Si le nombre est compris en 0 et la valeur p
    disp("Pile")//Réussite
else
    disp("Face")//Echec
end
```

Avec ces valeurs :

p	0.5	p	0.5
u	0.411	u	0.189

On obtient le résultat suivant : `"Pile"`

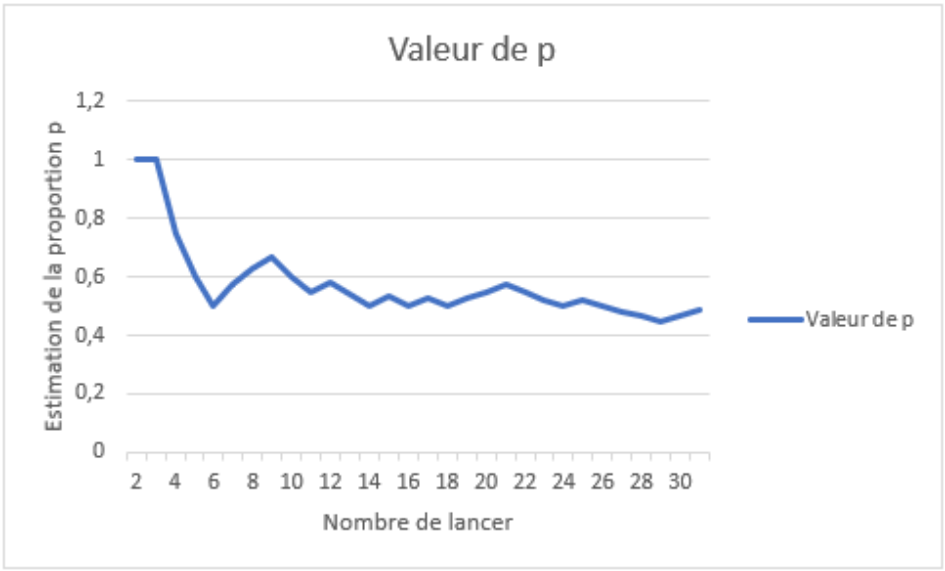
La valeur de la variable u étant une variable aléatoire réelle, on obtiendra « Face » lorsque celle-ci sera au-dessus de la variable p, c'est-à-dire lorsque $u > p$.

Exemple Excel

Modélisons ce problème sur Excel. Nous allons définir la probabilité p la proportion de réussite d'obtenir $pile$ et n le nombre de lancers. Nous allons ensuite calculer la probabilité de réussite ou d'échec selon la commande :

$$SI(ALEA < p; 1; 0)$$

On calcule la somme puis la moyenne de ces réalisations. On obtient le graphe suivant :



On remarque que lorsque l'on simule aléatoirement les valeurs dans notre tableau Excel, la courbe a tendance à se rapprocher de la valeur $p = 0.5$ avec $n = 30$. Pour avoir une meilleure précision on peut augmenter le nombre de lancers n.

(Cf. : Cadre d'analyse - Exemple avec n lancers d'une pièce de monnaie à deux faces : « Pile » et « Face » - Page)

Simulation d'une loi exponentielle

Soit $X \sim \mathrm{Exp}(\lambda)$, suit une loi exponentielle de paramètre $\lambda > 0$, alors $F(x) = 0$ si $x < 0$ et on a $F(x) = e^{-\lambda x}$ si $x \geq 0$. On en déduit que, pour $\forall\, u \in [0,1]$:

$$F^{-1}(u) = -\frac{log(1-u)}{\lambda}$$

Notons $(1-u)$ comme la variable aléatoire réelle U, U étant uniforme sur l'intervalle $[0,1]$, si $U \sim U[0,1]$ alors,

$$X = -\frac{\log(U)}{\lambda}$$

X simule une loi exponentielle de paramètre λ.

Exemple Scilab

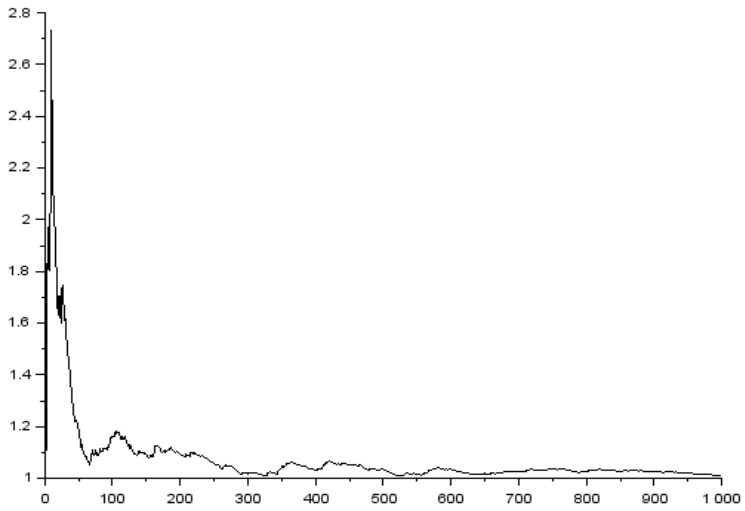
Grace au logiciel Scilab nous pouvons facilement tracer la courbe logarithmique avec des valeurs générées aléatoirement. Pour cela on effectue le code suivant :

```
clf;
plot2d(cumsum(-log(rand(1,1000))./[1:1000]);

//Log-rand fournit un 1000echantillon de variables exponentielles de parametre 1
//sur le principe de la méthode d'inversion de la fonction de répartition
//Cumsum permet de stocker les sommes cumulées des réalisations

//La commande représente toujours la moyenne empirique d'echantillons de lois ex
ponentielles de parametre 1 en fonction du nombre de réalisations dans l'echanti
llon
```

En effectuant cette commande, le calcul fait directement référence au procédé de simulations des lois exponentielles se basant sur la méthode d'inversion de la fonction de répartition. On obtient les graphes suivants :



Voici un autre graphe généré aléatoirement :

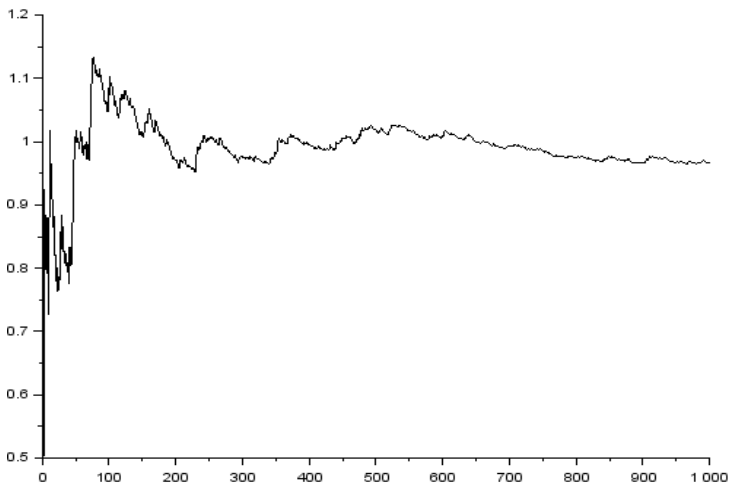


Illustration du théorème central limite

Le théorème central limite est un pilier dans le domaine des statistiques. Bien que connu depuis 1733, pour le cas particulier où les variables suivent la loi de Bernoulli de paramètre $p = 0.5$ dans les travaux de De Moivre. Ce théorème obtient sa première démonstration en 1809 par Laplace.

Il explique pourquoi les moyennes d'un grand nombre d'échantillons suivent une loi normale, et ce même si ces échantillons suivent individuellement une autre loi de probabilité. On dit qu'il établit la convergence en loi de la somme d'une suite de variables aléatoires vers la loi normale. Ainsi, on obtient une courbe en cloche, appelée "de Gauss", affirmant que toute somme de variables aléatoires indépendantes tend dans certains cas vers une variable aléatoire gaussienne. La variable aléatoire peut être discrète ou continue. Ce théorème permet de calculer la précision exacte d'un résultat obtenu par simulation. Ce qui permet d'approximer des ensembles de données avec des distributions inconnues comme étant normales.

Le théorème central limite exige une variance finie, c'est la seule hypothèse exigée sur la loi de probabilité suivie par les variables aléatoires. Par ce principe, le théorème s'applique à un grand nombre de lois de probabilité offrant une explication sur l'omniprésence de la loi normale. En effet, on peut facilement remarquer que dans la majorité des cas, c'est l'addition d'un nombre important de petites perturbations aléatoires qui implique un phénomène. Un exemple très simple peut être imagé par un environnement très sec, où la pluie s'est faite rare ; les petites perturbations aléatoires sont les gouttes de pluie et le phénomène créé est un torrent d'eau qui ne peut pas s'infiltrer dans le sol, conséquence directe de la présence massive de gouttes de pluie.

Théoriquement, le théorème central limite

Soit $X_1, X_2, X_3, \dots, X_n$ un ensemble de n variables aléatoires indépendantes et identiquement distribuées. Chaque X_i a une distribution de probabilité arbitraire $P(X_1, X_2, X_3, \dots, X_n)$ d'espérance μ_i et une variance ${\sigma_i}^2$. La loi de probabilité de la moyenne centrée réduite converge vers la loi normale centrée réduite lorsque $n \to +\infty$. C'est-à-dire que lorsque l'effectif n est suffisamment grand, la variable aléatoire somme telle que $S_n = X_1, + X_2 + X_3 + \cdots + X_n$ est approximativement de loi $\mathcal{N}(n\mu, n\sigma^2)$ et la variable aléatoire $\overline{X_n} = \frac{S_n}{n}$ est approximativement de loi $\mathcal{N}(\mu, \frac{\sigma^2}{n})$,

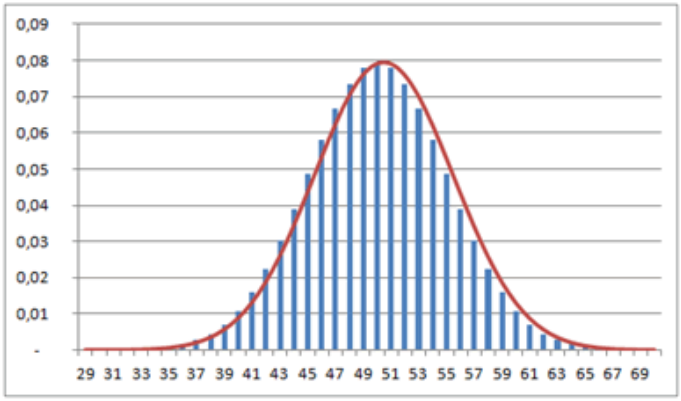
$$\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} = Z_n \sim \mathcal{N}(0,1)$$

Applications – Approximation de la loi binomiale par la loi normale

La variable binomiale est bien une somme de variables indépendantes de Bernoulli. On sait qu'une loi $B(n,p)$ a pour espérance np et pour variance $np(1-p)$. Donc,

$$B(n,p) \sim \mathcal{N}\left(np, \sqrt{np(1-p)}\right)$$

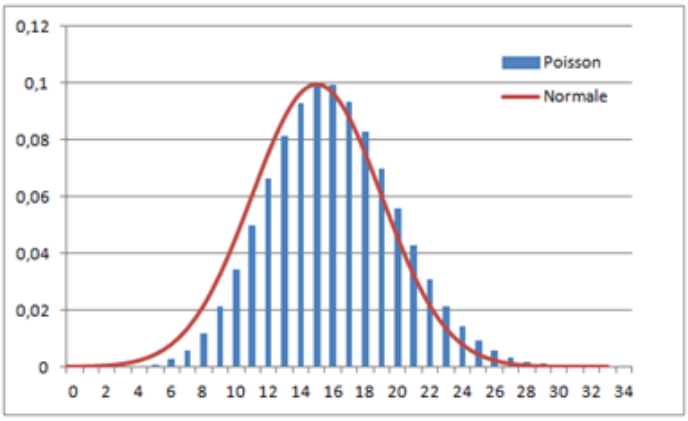
Par exemple, on peut avoir : $B(100,0.5) \sim \mathcal{N}(50,5)$. Cet exemple est illustré ci-dessous, les valeurs prises par la loi binomiale paraissent en bâtons et celles de la loi normale figurent sous forme de courbe. L'échantillon étant important, de taille $n = 100$, on constate une forte similitude entre les deux distributions.



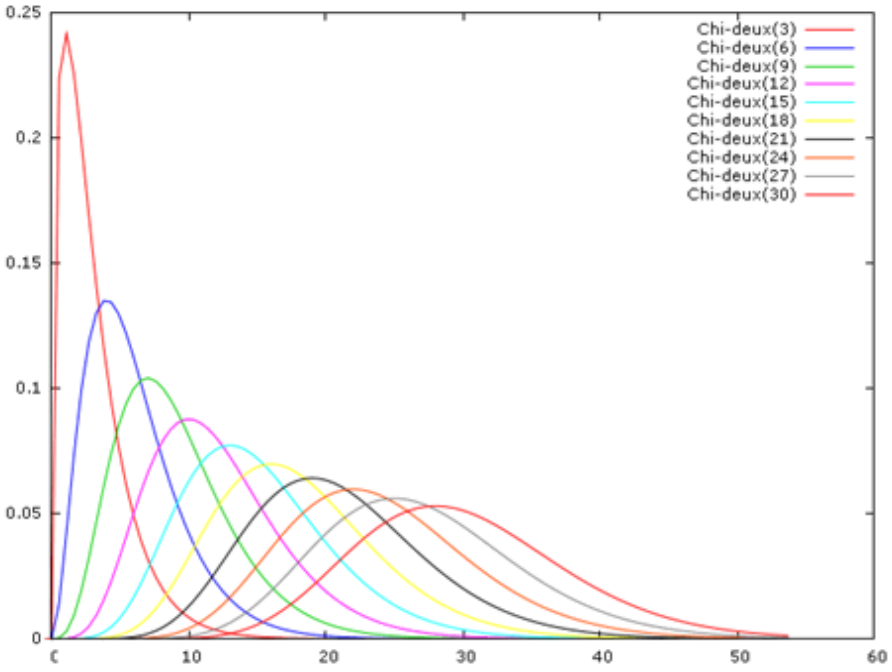
Applications – Approximation de la loi de poisson par la loi normale

Lorsqu'une variable aléatoire X suit une loi de Poisson de paramètre λ on note : $X \sim P(\lambda)$.

Voici la comparaison entre la distribution de la loi de Poisson : $P(16) \sim \mathcal{N}(16,4)$. Rappelons que c'est à partir d'une valeur de paramètre située autour de 18 que la loi de Poisson est approchosée par une loi normale, ce qui explique le décalage visible sur la figure suivante :



En effet, lorsque $n \to +\infty$ les courbes ressemblent de plus en plus à celle de la loi normale.



Montrons comment la loi se concentre autour de l'espérance lorsque n augmente.

La moyenne empirique des variables aléatoires $X_1, X_2, X_3, \dots, X_n$ est la variable aléatoire

$$\overline{X_n} = \frac{X_1, + \cdots + X_n}{n}$$

On sait d'ores et déjà que la moyenne empirique a pour espérance m et pour variance $\frac{\sigma^2}{n}$.

Ainsi, plus n est grand, moins cette variable aléatoire varie. A la limite, quand n tend vers l'infini, elle se concentre sur son espérance, m.

Démontrons-le :

Quand n est grand $\overline{X_n}$ est proche de m avec une forte probabilité. Autrement dit,

$$\forall\, \varepsilon > 0, \lim_{n\to\infty} P(|\overline{X_n} - m| > \varepsilon) = 0$$

En effet, d'après l'inégalité de Tchebychev,

$$P(|\overline{X_n} - m| > \varepsilon) \leq \frac{Var(\overline{X_n})}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2 n} \to 0$$

On dit que $\overline{X_n}$ converge en probabilité vers m.

Cadre d'analyse – Exemple avec une pièce de monnaie à deux faces : « Pile » et « Face »

On a dans cet exemple, un joueur qui effectue n lancers d'une pièce de monnaie à deux faces. On sait que la probabilité de faire « Pile » est $p_{pile} = 0{,}5$.
$X \sim B(0{,}5)$ est une variable aléatoire telle que :

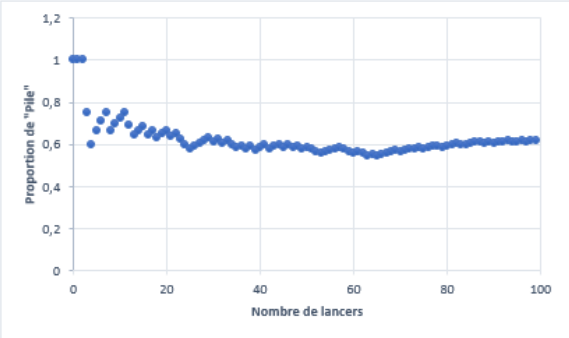
$$\begin{cases} X = 1 \; quand \; on \; fait \; "Pile" \\ X = 0 \; quand \; on \; fait \; "Face" \end{cases}$$

Avec Excel, nous simulons ces lancers. Ainsi, nous pouvons représenter :

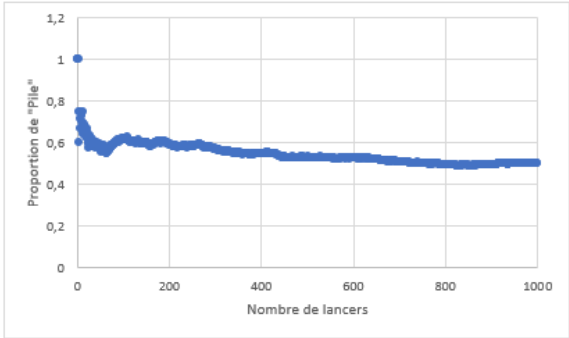
Lancers	"Pile" ou "Face"		
n	Réalisations	Nombre de "Pile"	Proportion de "Pile"
0	1	1	1
1	1	2	1
2	0	2	0,666666667
3	0	2	0,5
...	...	...	...
9999	1	4991	0,4991
10000	0	4991	0,499050095

Ici, $\frac{1}{n}\sum_{i=1}^{n} X_i$ est représentée dans le tableau ci-dessus par la colonne Proportion de « Pile ». On remarque que $\frac{1}{n}\sum_{i=1}^{n} X_i$ tend en probabilité vers $0{,}5$.
Voici trois graphiques permettant de schématiser les lancers effectués et la proportion de « Pile » obtenue.

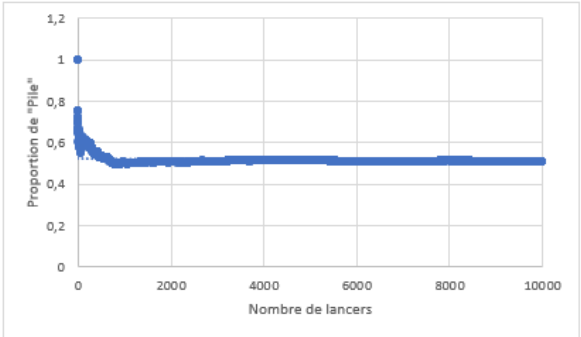
Pour n = 100
Valeur estimée : **0,53**



Pour n = 1000
Valeur estimée : **0,512487512**



Pour n = 10 000
Valeur estimée : **0,5076**



On sait que l'espérance d'une variable aléatoire de Bernoulli est $m = p$.

On peut donc affirmer que $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ converge en probabilité vers m.

Plus n est grand, plus cela est facilement observable sur un graphique.

Estimation d'une loi d'un estimateur en économétrie

Le dernier test du document de cours. Là-encore, proposer moi en texte votre méthodologie pour les simulations. L'idée sera de comparer vos lois simulées et les lois théoriques correspondantes. Cela va dépendre de la taille d'échantillon car ils modifient les degrés de liberté.

Théoriquement, le test de Student

Le test-t de Student est un test statistique permettant de comparer les moyennes de deux groupes d'échantillons. Il s'agit donc de savoir si les moyennes des deux groupes sont significativement différentes d'un point de vue statistique.

On veut tester si un coefficient est égal ou non à une valeur donnée :

$$\begin{cases} H_0 : a = b \\ H_1 : a \neq b \end{cases}$$

Avec b étant une constante donnée. En économétrie il est important de retenir le cas particulier lorsque $b = 0$.

D'une part, si le modèle du test de Student indique un coefficient significatif (c'est à dire à la décision de H_1) alors on peut admettre que la variable correspondante est « intéressante » dans le modèle. Dans ce cas, elle permet d'expliquer le modèle.

D'autre part, si le modèle du test de Student indique un coefficient non significatif (c'est à dire à la décision H_0) alors on peut admettre ce coefficient comme étant nul, ainsi la variable est dite « inintéressante » pour le modèle. Dans ce cas, cela ne nous permet pas d'expliquer mieux le modèle.

Lorsqu'il existe une forte colinéarité entre les variables, la seconde conclusion n'est pas toujours juste. Pour savoir si la variable est négligeable dans le modèle, nous allons construire un nouveau modèle sans cette variable.

Ici nous allons considérer un test bilatéral. Tout d'abord nous allons considérer la distance : $(\hat{a} - a)$

Nous allons donc considérer la statistique : $\dfrac{(\hat{a} - a)}{\sigma_{\hat{a}}}$

Où $\sigma_{\hat{a}}$ est l'écart type ou erreur-type de l'estimateur $\hat{a}$. Cela suit une loi normale centrée réduite.

En utilisant la définition de la loi de student comme rapport d'une variable normale centrée réduite et de la racine d'une loi de Khi-deux, on obtient :

$$T = \frac{\hat{a} - a}{\sigma_{\hat{a}}\sqrt{(X'X)^{-1}}}$$

D'après le résultat que l'on a vu sur le lien entre Student et Fisher, un test bilatéral sur un paramètre est équivalent à un test de Fisher, la Fisher est donc le résultat d'une Student au carré : $t^2 = F$

On obtient alors :

$$t^2 = F = \frac{(\hat{a} - a)^2}{\sigma^2_{\hat{a}}} \sim F(1, n - p - 1)$$

Théoriquement, le test de Fisher

Le test de Fisher est basé sur la comparaison des résidus dans le modèle contraint et le modèle non contraint. On souhaite estimer un modèle économétrique linéaire en introduisant des informations sur les paramètres prenant la forme de contraintes linéaires. Nous allons tester si certaines de ces relations entre les paramètres sont bien acceptées par les données. On va montrer les étapes à suivre qui vont nous conduire à obtenir un estimateur différent de celui des moindres carrés ordinaires, appelé estimateur des moindres carrés contraints. Nous allons également introduire un test très utilisé permettant de tester des contraintes linéaire, le test de Fisher et voir comment le mettre en œuvre.

Le but de ce test est de comparer la qualité du modèle avec et sans variables explicatives c'est à dire avec uniquement la constante. Si cette différence est significative, cela veut dire qu'au moins une variable explicative permet d'améliorer la qualité du modèle.

Lorsque que nous réalisons des tests de contraintes sur les paramètres, nous avons deux choix, le test de Student ou le test de Fisher. Ces tests de contraintes sont basés sur la distribution de l'estimateur des MCO et sur la variance estimée des aléas. La différence entre ces deux tests sera au niveau du nombre de contraintes testé. Les tests d'une contrainte sont des tests de Student, en revanche, les tests de plusieurs contraintes sont des tests de Fisher.

Pour écrire la statistique de ce modèle dans le cas Gaussien sous hypothèse nulle :

$$f = \frac{\frac{SCRc - SCRnc}{k}}{\sigma^2 n}$$

Avec $f \sim (k, n - p - 1)$, k étant le nombre de contraintes, n la taille de l'échantillon et p le nombre de variables explicatives.

En fonction de la somme des carrée des résidus, on sait que :

$$\sigma^2 n = \frac{SRCnc}{(n - p - 1)}$$

On peut alors encore écrire f sous la forme : (1)

$$f = \frac{\frac{SCRc - SCRnc}{k}}{\frac{SCRnc}{n - p - 1}}$$

Où SCRc et SCRnc représentent respectivement la somme des carrés des résidus du modèle ayant pour variable explicative la constante uniquement et la somme des carrés des résidus du modèle

complet. Alors f suit une loi de Fischer car le numérateur le dénominateur suivent des lois du Chi-deux.

Nous voulons dans notre cas précis, réaliser le test de nullité du R^2, pour cela on montre que f peut encore s'écrire : (2)

$$f = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-p-1}}$$

Avec R^2, le coefficient de détermination du modèle complet. Ce test permet de choisir le meilleur modèle ou encore de juger de la pertinence de l'ajout ou du retrait de variables explicatives.

Applications – Tester les contraintes

Nous allons mettre en œuvre le test de Fisher d'un ensemble de contraintes $\mathrm{H_o} = 0$.

Tout d'abord, on commence par estimer le modèle avec et sans contraintes comme nous l'avons vu précédemment. On récupère par la suite la somme des carrés des résidus SCR ainsi que le nombre de degrés de liberté et le nombre de contraintes. Une fois que nous avons calculé notre SCR, on calcule alors notre statistique de test f (1). On compare ensuite ce résultat au fractile d'ordre $(1-\alpha)$ de loi $f\sim(k, n-p-1)$ pour savoir si on rejette ou non l'hypothèse H_o.

Dans le premier cas si : $f < (1-a)$, on décide de rejeter H_o. En effet, la somme des carrés des résidus estimés sous contraintes est totalement différente de celle des carrés des résidus estimés sans contrainte, on ne peut donc accepter que H_o est vraie.

Dans le second cas si : $f > (1-a)$, on décide d'accepter l'hypothèse H_o.

Applications – Test de nullité du R²

Pour le test de nullité, on cherche à tester l'hypothèse de l'égalité de plusieurs coefficients pour H_o.

Les étapes son similaire au test précèdent, on utilise le test de Fisher, il suffit d'estimer le modèle non contraint, on calcule la somme SCR des carrés des résidus estimés puis on estime le modèle contraint et enfin on calcule la somme SCRc des carrés des résidus estimés. On obtient bien alors le modèle f (2). On gardera les mêmes conditions pour les zones de rejets.

Etapes à suivre afin de simuler une loi à partir du test de Fisher

On souhaite estimer un modèle économétrique linéaire en introduisant des informations sur les paramètres prenant la forme de contraintes linéaires. Nous allons tester si certaines de ces relations entre les paramètres sont bien acceptées par les données.

Montrons les étapes à suivre permettant de nous conduire à un estimateur différent de celui des moindres carrés ordinaires, appelé estimateur des moindres carrés contraints. Nous allons également introduire un test très utilisé permettant de tester des contraintes linéaires, le test de Fisher et voir comment le mettre en œuvre.

Le modèle de régression linéaire simple : $y = ax + b + \varepsilon$

Soit le modèle $y_i = ax_i + b + \varepsilon_i$ et $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$

On souhaite générer ce modèle.

Dans notre cas, les paramètres a et b représentent respectivement la pente et la constante. Ces deux paramètres doivent être connus, ainsi que σ^2.

La variable y_i doit être étudiée à x_i donné, y_i et x_i sont respectivement des variables endogène (fixée par le modèle) et exogène (fixée en dehors du modèle). Y est aléatoire par l'intermédiaire de ε et X est non aléatoire.

La composante stochastique du modèle est ε_i, aussi appelée erreur du modèle telle que $\varepsilon_i \sim N(0, \sigma_\varepsilon)$ est l'hypothèse de normalité.

On a :

$$E(\varepsilon_i) = 0 \qquad V(\varepsilon_i) = \sigma_\varepsilon^2 \qquad COV(x_i, \varepsilon_i) = 0 \qquad COV(\varepsilon_i, \varepsilon_j) = 0$$

Un modèle seulement stochastique permet d'avoir différentes valeurs de y_i pour une même valeur de x_i. La composante stochastique du modèle correspond à l'espérance de y_i sachant x_i.

Objectif de ce modèle

Lorsque ce modèle est appliqué sur un échantillon de taille n, notre but est de déterminer les valeurs des paramètres a et b. En effet, nous souhaitons obtenir une bonne estimation, ainsi nous pourrons approcher le nuage de points.

Pour ce faire, nous avons besoin d'utiliser le critère des moindres carrés afin de minimiser la somme des carrés des erreurs présentent entre les valeurs de Y et celles prédites par le modèle de prédiction.

Principe de l'ajustement des moindres carrés

L'estimateur des moindres carrés ordinaires des paramètres a et b doit répondre à la minimisation de :

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (-ax_i - b + y_i)^2 \quad \Leftrightarrow \quad \begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases}$$

Ainsi, nous avons :

$$\begin{cases} \sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i^2 - b \sum_{i=1}^{n} x_i = 0 \\ \bar{y} - a\bar{x} - b = 0 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^{n} x_i \varepsilon_i = 0 \\ \sum_{i=1}^{n} \varepsilon_i = 0 \end{cases} \Leftrightarrow \begin{cases} \hat{a} \\ \hat{b} \end{cases}$$

Avec $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ et $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Les estimateurs des moindres carrés $\hat{a}$ et $\hat{b}$

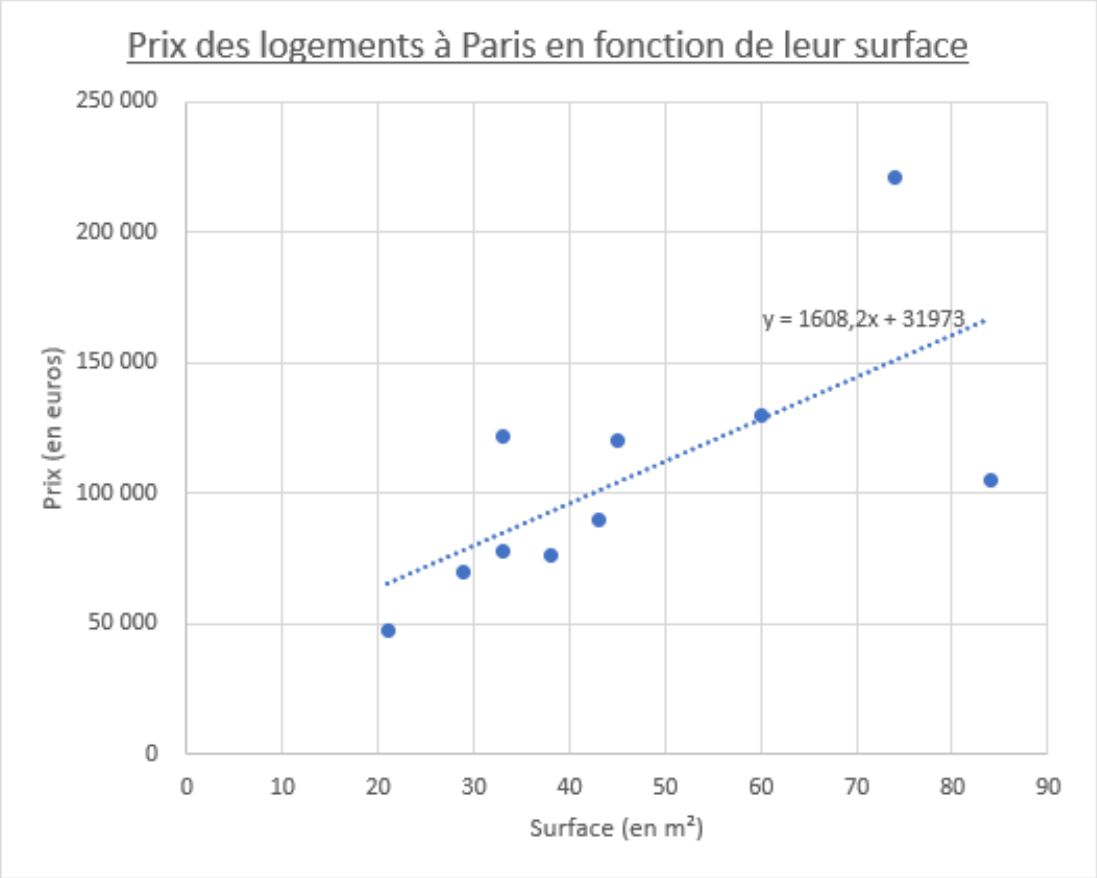
On sait que le point moyen $(\bar{x}, \bar{y})$ vérifie l'équation de droite estimée : $\bar{y} = \hat{a}\bar{x} + \hat{b}$. Ce point est le centre de gravité du nuage de points.

Estimateur	Espérance	Variance
$\boldsymbol{\hat{a} = \frac{Cov_{emp}(x,y)}{Var_{emp}(x)} = \frac{\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - \bar{x}^2}}$	$Esp(\hat{a}) = a$	$Var(\hat{a}) = \sigma_{\hat{a}}^2 = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$
$\boldsymbol{\hat{b} = \bar{y} - \hat{a}\bar{x}}$	$Esp(\hat{b}) = b$	$Var(\hat{b}) = \sigma_{\hat{b}}^2 = \sigma_\varepsilon^2 [\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}]$

Exemple Excel – Application sur le prix des logements à Paris

Obs	Surface	Prix	Surface²	Prix²	Surf*Prix
1	84	105 189	7 056	11 064 725 721	8835876
2	33	77 748	1 089	6 044 751 504	2565684
3	21	47 259	441	2 233 413 081	992439
4	45	120 434	2 025	14 504 348 356	5419530
5	33	121 959	1 089	14 873 997 681	4024647
6	60	129 581	3 600	16 791 235 561	7774860
7	43	89 944	1 849	8 089 923 136	3867592
8	38	76 224	1 444	5 810 098 176	2896512
9	74	221 051	5 476	48 863 544 601	16357774
10	29	70 126	841	4 917 655 876	2033654
Somme	**460**	**1 059 515**	**24 910**	**133 193 693 693**	**54768568**
Taille n : 10	**46**	**105 952**	**2491**	**13319369369**	**5476856,8**

Pente	**â = 1608,2**
Ordonnée à l'origine	**b^ = 31972,7**
Le modèle est : ^y_i =	**1608,2 *x_i + 31972,7**



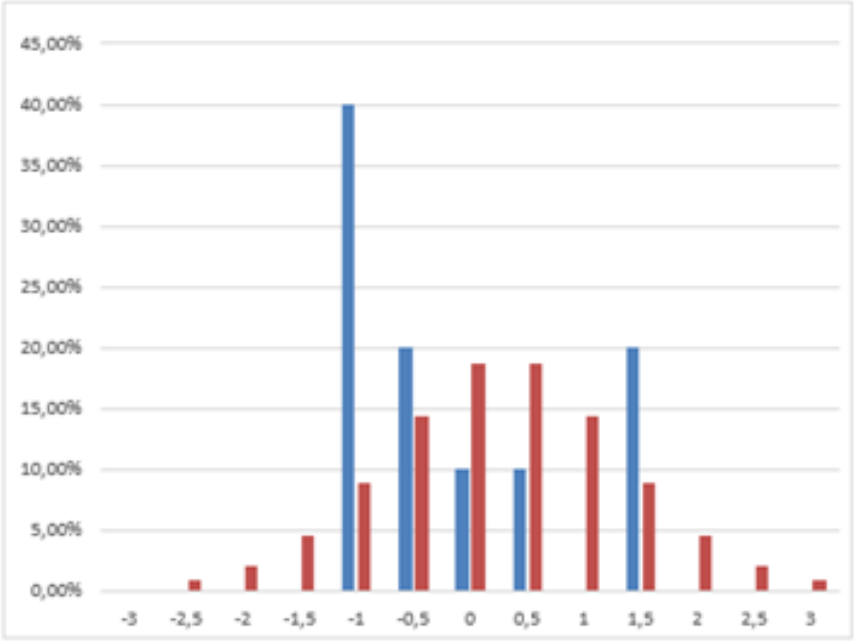
En fonction de la taille de l'échantillon – Test de Student

Nous avons généré de manière aléatoire n valeurs issues d'une Gaussienne sur Excel. Sur ces valeurs, nous les analysons sur des intervalles donnés. Les valeurs générées par la Gaussienne sont globalement situées dans l'intervalle -3 et 3.

Ainsi, nous pouvons comparée les valeurs simulées par la Gaussienne à celles théoriques données par la Student. Nous avons représenté dans la dernière colonne, la différence distinguée entre la loi théorique et celle simulée.

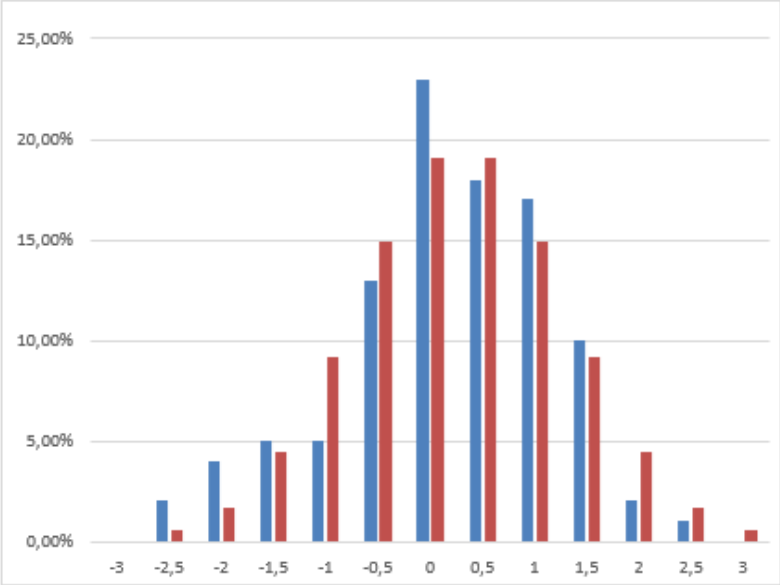
$Avec\ n = 10$

Intervalles	Simulation cumulé	Simulé	Proportion simulation	Proportion théorique student	Théorique cumulée student	Comparaison
-3	0	0	0,00%	0,00%	0,0067	0,00%
-2,5	0	0	0,00%	0,91%	0,0157	0,91%
-2	0	0	0,00%	2,10%	0,0367	2,10%
-1,5	0	0	0,00%	4,56%	0,0823	4,56%
-1	4	4	40,00%	8,82%	0,1704	-31,18%
-0,5	6	2	20,00%	14,35%	0,3139	-5,65%
0	7	1	10,00%	18,61%	0,5000	8,61%
0,5	8	1	10,00%	18,61%	0,6861	8,61%
1	8	0	0,00%	14,35%	0,8296	14,35%
1,5	10	2	20,00%	8,82%	0,9177	-11,18%
2	10	0	0,00%	4,56%	0,9633	4,56%
2,5	10	0	0,00%	2,10%	0,9843	2,10%
3	10	0	0,00%	0,91%	0,9933	0,91%



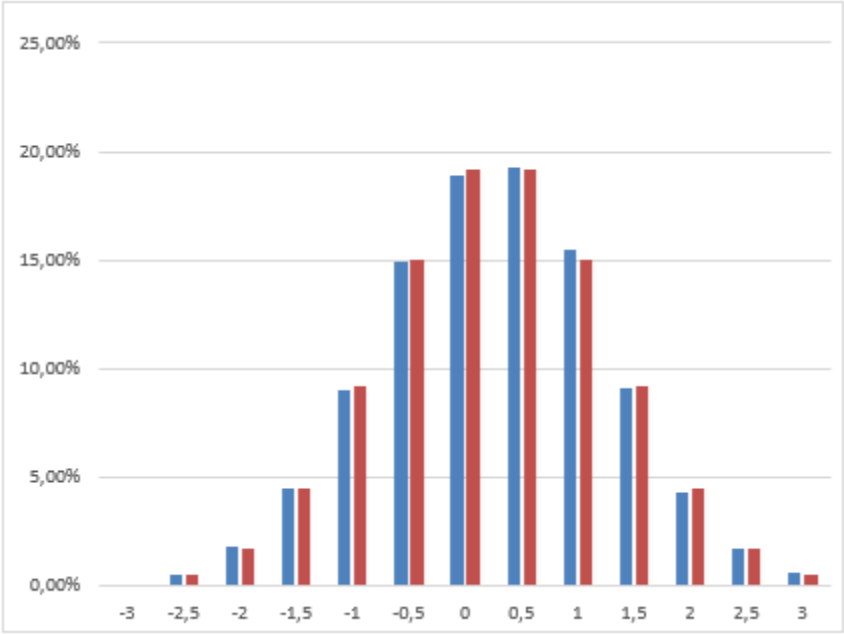
Avec $n = 100$

Intervalles	Simulation cumulé	Simulé	Proportion simulation	Proportion théorique student	Théorique cumulée student	Comparaison
-3	0	0	0,00%	0,00%	0,0017	0,00%
-2,5	2	2	2,00%	0,53%	0,0070	-1,47%
-2	6	4	4,00%	1,71%	0,0241	-2,29%
-1,5	11	5	5,00%	4,43%	0,0684	-0,57%
-1	16	5	5,00%	9,15%	0,1599	**4,15%**
-0,5	29	13	13,00%	14,92%	0,3091	1,92%
0	52	23	23,00%	19,09%	0,5000	**-3,91%**
0,5	70	18	18,00%	19,09%	0,6909	1,09%
1	87	17	17,00%	14,92%	0,8401	-2,08%
1,5	97	10	10,00%	9,15%	0,9316	-0,85%
2	99	2	2,00%	4,43%	0,9759	2,43%
2,5	100	1	1,00%	1,71%	0,9930	0,71%
3	100	0	0,00%	0,53%	0,9983	0,53%



Avec $n = 10\,000$

Intervalles	Simulation cumulé	Simulé	Proportion simulation	Proportion théorique student	Théorique cumulée student	Comparaison
-3	11	0	0,00%	0,00%	0,0014	0,00%
-2,5	61	50	0,50%	0,49%	0,0062	-0,01%
-2	237	176	1,76%	1,65%	0,0228	-0,10%
-1,5	682	445	4,44%	4,41%	0,0668	-0,04%
-1	1585	903	9,02%	9,18%	0,1587	0,17%
-0,5	3075	1490	14,88%	14,99%	0,3085	0,11%
0	4971	1896	18,93%	19,15%	0,5000	0,21%
0,5	6898	1927	19,24%	19,15%	0,6915	-0,10%
1	8450	1552	15,50%	14,99%	0,8413	**-0,51%**
1,5	9355	905	9,04%	9,18%	0,9332	0,15%
2	9778	423	4,22%	4,41%	0,9772	**0,18%**
2,5	9945	167	1,67%	1,65%	0,9938	-0,01%
3	10000	55	0,55%	0,49%	0,9986	-0,06%



Conclusion de la simulation

Lorsque le nombre de réalisation n augmente, on remarque que la loi simulée prend rapidement la forme de la loi théorique. En effet, on peut l'observer aisément sur les graphiques et dans la colonne « Comparaison » des tableaux, plus spécifiquement sur les cases de couleur verte. On pourra rajouter également que si l'on rajoute la loi théorique d'une gaussienne, celle-ci sera égal à la loi théorique de la Student.

Simulations

La loi de $\hat{a}$, l'estimateur de a, ne dépend ni de σ_ε^2 ni de b, par contre elle dépend de la taille n de l'échantillon. Ainsi, il s'agit soit d'une loi $St(n-2)$, soit d'une loi $F(1, n-2)$. En effet, il y a deux degrés de liberté car on estime deux paramètres a et b afin d'obtenir Y. Pour connaître le nombre de variables à générer, il faut connaître la taille n de l'échantillon. Ainsi, nous pourrons générer des valeurs de Y. Théorquement, on ne trouve aucun lien entre les Y et les X.

Nous prendrons pour commencer comme taille d'échantillon : $n = 10$ puis nous comparerons avec $n = 100$.

Le modèle est de la forme : $y = ax + b + \varepsilon$ où $\varepsilon \sim N(0, \sigma_\varepsilon)$

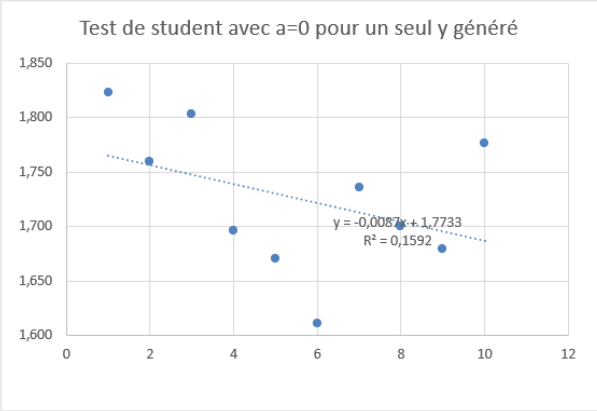
Ainsi, si on a $a = 0,\ b = 1$ et $\sigma_\varepsilon^2 = 1$, alors le modèle sera de la forme : $y = 1 + \varepsilon$.

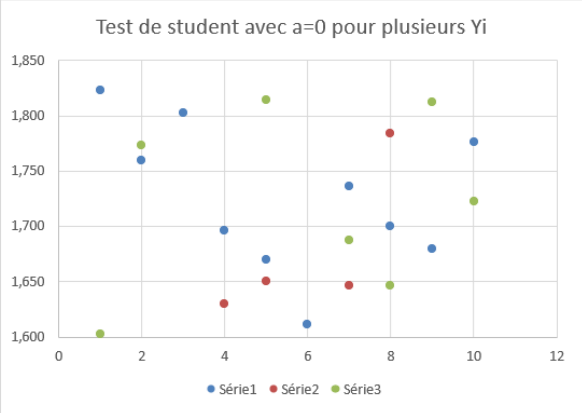
On utilisera principalement $a = 0$, en effet, la loi de la statistique de test est la loi en supposant H_0 vraie, c'est l'hypothèse de normalité.

Les estimations vont être différentes à chaque fois même si les paramètres restent inchangés.

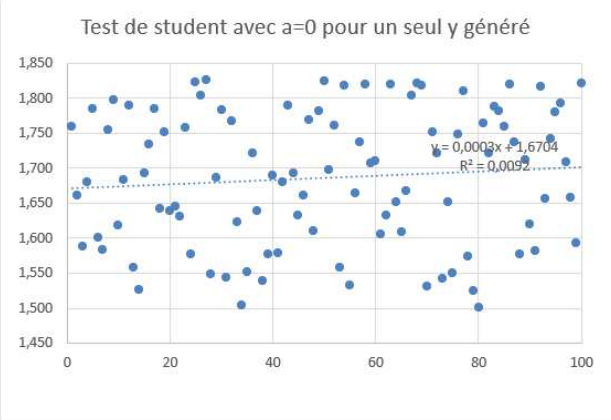
Avec $a = 0$, $b = 1$ *et* $\sigma_\varepsilon^2 = 1$.

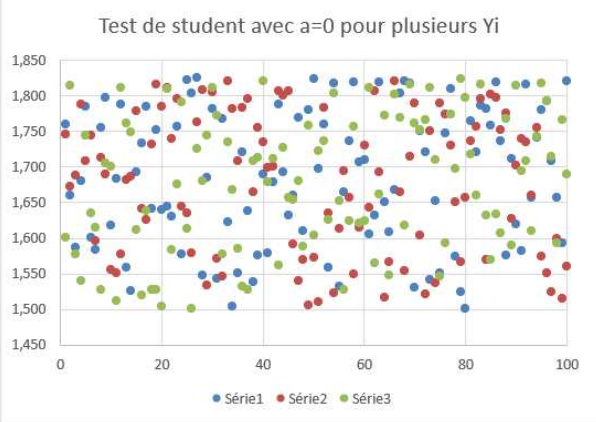
Pour $n = 10$





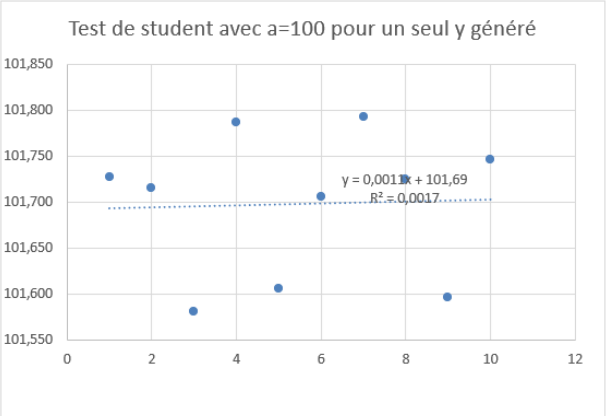
Pour $n = 100$

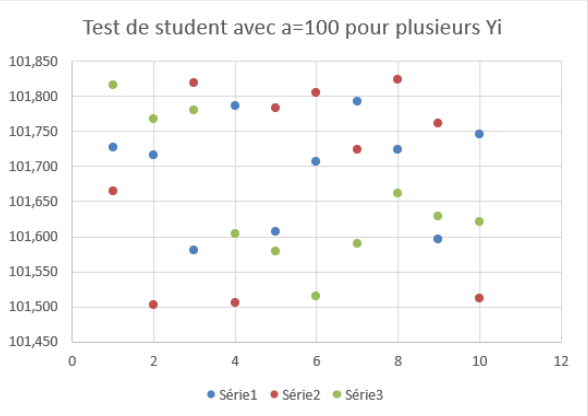




Avec $a = 100$, $b = 1$ *et* $\sigma_\varepsilon^2 = 1$.

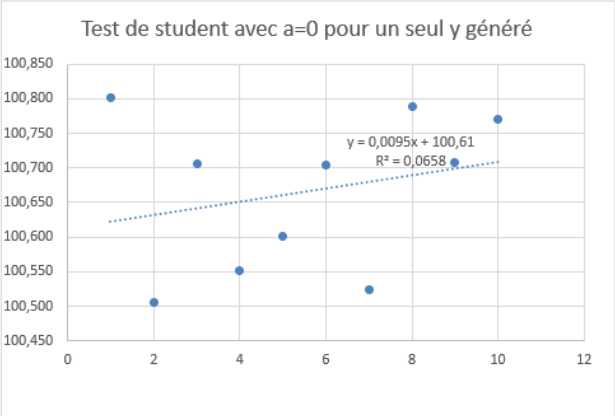
Pour $n = 10$

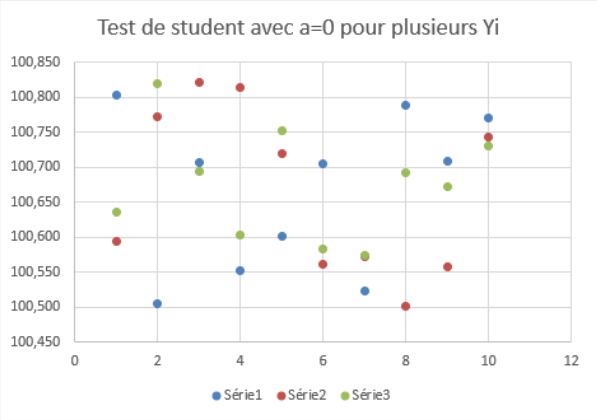




$Avec\ a = 0,\ b = 100\ et\ \sigma_\varepsilon^2 = 1.$

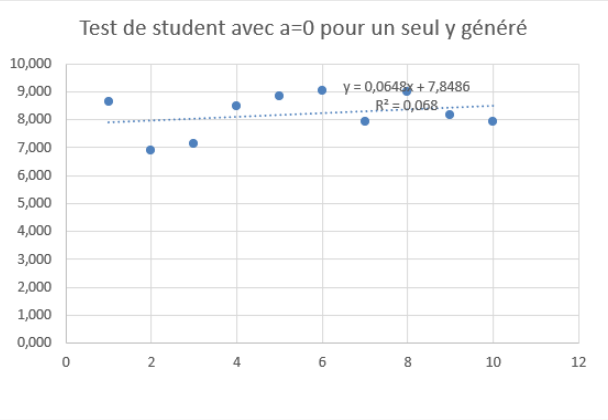
Pour $n = 10$

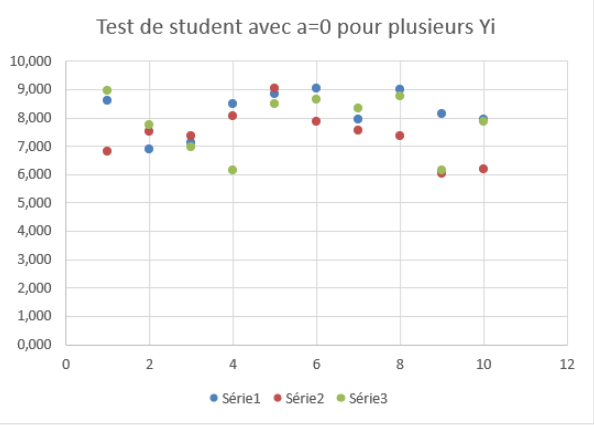




$Avec\ a = 0,\ b = 1\ et\ \sigma_\varepsilon^2 = 100.$

Pour $n = 10$





On remarque sur les nuages de points que ceux-ci ne sont pas du tout corrélés étant donné que dans cette exemple la valeur de a vaut 0.

En partant des Yi générés précédemment, nous pouvons calculer le R^2 de chaque Yi.

R2	0,05835	0,178	0,046	0,307	0,028	0,067	0,004	0,038	0,004	0,041

Ainsi, on remarque que la statistique de test d'une Student ne sera jamais nulle.

Exemple avec $n = 10$

Taille de l'échantillon : n = 10

Le modèle est de la forme : y = a * x + b + eps

Avec : —> On a :

b = 1 y = a * x + 1 + eps

sigma² = 1

La loi de la statistique de test est la loi en supposant **H0 vraie (hypothèse de normalité)**

Ainsi : a = 0

On a donc : **y = b + eps** eps de loi N(0,sigma)

Les x_i sont iid et suivent une loi $U[0,1]$.

Obs	x	eps	y	x²	y²	x * y	ŷ	ŷ²	(y-ŷ)²	(x-xb)²
1	0,439	0,655	1,655	0,1926	2,738	0,7261	0,0056	0,0000	2,7192	0,0087
2	0,042	0,539	1,539	0,0018	2,369	0,0648	0,0005	0,0000	2,3675	2,5337
3	0,925	0,398	1,398	0,8550	1,955	1,2929	0,0117	0,0001	1,9225	0,5278
4	0,099	0,392	1,392	0,0098	1,938	0,1375	0,0013	0,0000	1,9345	6,7245
5	0,248	0,730	1,730	0,0615	2,991	0,4290	0,0031	0,0000	2,9806	0,1008
6	0,082	0,550	1,550	0,0067	2,402	0,1267	0,0010	0,0000	2,3988	0,0067
7	0,101	0,699	1,699	0,0102	2,886	0,1717	0,0013	0,0000	2,8821	0,0102
8	0,631	0,734	1,734	0,3985	3,007	1,0947	0,0080	0,0001	2,9793	0,3985
9	0,286	0,841	1,841	0,0818	3,390	0,5264	0,0036	0,0000	3,3766	0,0818
10	0,603	0,801	1,801	0,3634	3,242	1,0855	0,0077	0,0001	3,2149	0,3634

		Sur n :
Somme des x =	3,455226	0,346
Somme des y =	16,33849	1,634
Somme des x² =	1,981157	0,198
Somme des y² =	26,91928	2,692
Somme des x*y =	5,655315	0,566
Somme des (x-xb)² =	10,7561	

Ainsi :

â = 0,0127

Calculons e.â :

SCR = 26,7760

De plus : (e.epshat)² = 3,3470

Ainsi : **e.â** = 0,3049

Calculon maintenant le t de Student :

t = 0,0416 où t suit une loi St(8)

Conclusion

Les moyens de simulation ont fortement changé la tendance actuelle de la recherche économétrique, en effet elle fait apparaître un changement de perspective par rapport à celle qui prévalait jusqu'au début des années 1970. L'économétrie visait à estimer les modèles élaborés par la théorie économique sans les remettre en cause, actuellement, ce sont souvent les problèmes soulevés par la pratique de l'économétrie qui conduisent les théoriciens à concevoir des modèles susceptibles de tenir compte de ces observations. Grace à ces méthodes mathématiques et de simulation économétriques basé sur certains modèle, les simulations permettent d'effectuer des simulations sur le monde réel que nous pouvons comparer aux modèles théoriques. On peut réaliser des expériences qui nous permettent au mieux de nous rapprocher du modèle du monde réelle. Nous avons étudié à travers notre projet l'importance de ces moyens de calculs au travers de certains domaines. Nous pouvons approfondir cette étude avec d'autres méthode de simulation, la méthode de Monte-Carlo permet de calculer une valeur numérique en générant des procédés aléatoires, cela est très utilisée par les statisticiens pour des simulation en finance comme également par les physiciens pour des simulation probabiliste sur des particules.

Bibliographie

ERIC DOR, Econométrie.
MARCEL G. DAGENAIS, Revue d'économie politique, Vol. 78, No. 3, DOMAINES ET MÉTHODES DES SCIENCES ÉCONOMIQUES (MAI-JUIN 1968), pp. 532-548.
POUPA J.C., Quelle informatique pour l'économétrie?, Économie rurale. No. 157, 1983. pp. 97-100.
J. DUESENBURRY, "The Methodological Basics of Economic Theory", dans The Review of Economics et Statistics, Nov. 1954, pp. 361-363.
MATTHIEU KOWALSKI, "La loi des grands nombres et le théorème de la limite centrale", Cours Supelec, 2008-2009.