Simulation, bootstrap

Gilbert Saporta

Conservatoire National des Arts et Métiers, Paris

gilbert.saporta@cnam.fr

http://cedric.cnam.fr/~saporta

Octobre 2012

- Il est souvent impossible d'obtenir des expressions exactes pour les distributions de statistiques de test ou d'estimateurs, car les calculs sont trop complexes.
- Les méthodes de simulation et de rééchantillonnage permettent de substituer à une étude théorique, une démarche expérimentale où les lois exactes sont approchées par des répartitions empiriques.
- La simulation aléatoire consiste à engendrer sur ordinateur des échantillons artificiels, à effectuer pour chacun de ces échantillons les calculs nécessaires, qui sont ensuite synthétisés.

D'où vient le nom de Monte Carlo?

- Travaux menés à Los Alamos
- N.Metropolis donna le nom de Monte Carlo car l'oncle de son co-auteur Stan Ulam « had an uncle who would borrow money from relatives because he "just had to go to Monte Carlo." »
- Nicholas Metropolis et Stanislas Ulam, The Monte Carlo Method , *Journal of the American Statistical Association,* vol. 44, n° 247, 1949, p. 335-341
- Nicholas Metropolis, « The Beginning of the Monte Carlo Method », Los Alamos Science, n° 15, 1987, p. 125-130

I Génération de variables aléatoires

- Toutes les méthodes reposent sur la génération de variables uniformes.

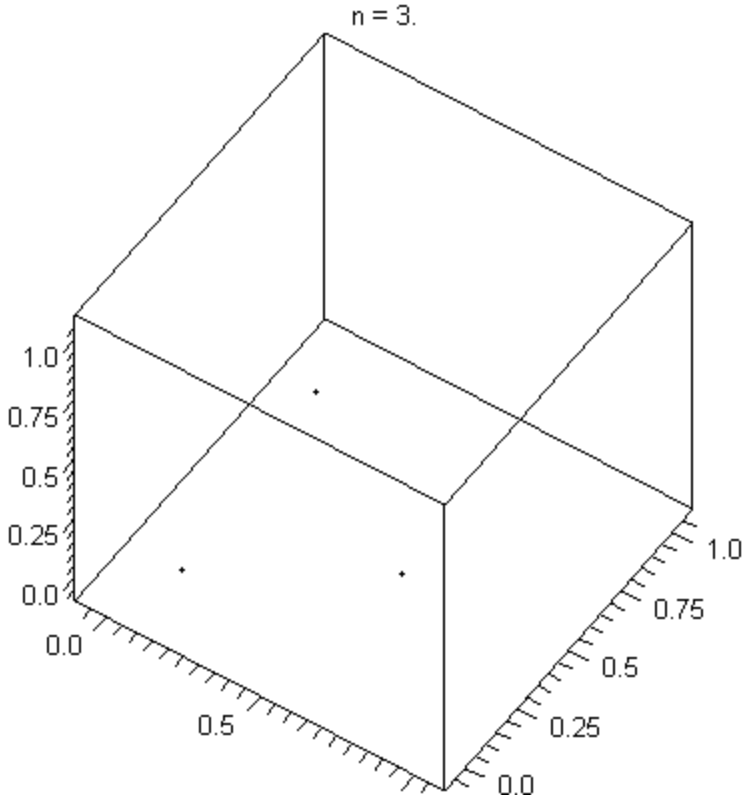
- I.1 Génération de variables uniformes R sur [0 ;1]
  - Pour mémoire : procédés physiques (roue de loterie) incompatibles avec l'informatique et la nécessité de disposer très rapidement de grands échantillons.
  - Algorithmes de génération de valeurs comprises entre 0 et 1 : déterministe, nombres pseudo-aléatoires.

- Un bon algorithme doit pouvoir réaliser des suites très grandes de nombres qui ont en apparence toutes les propriétés d'un n-échantillon de variables indépendantes et identiquement distribuées.
- **Méthode de Lehmer :**

$$r_{i+1} = a r_i \ (m)$$

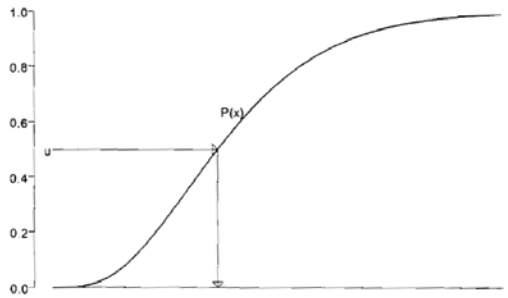
  - choix classiques: $a = 7^5 = 16807$

    ou $a = 2^{16}+3 = 65539$ avec $m = 2^{31}-1$
- Tests d'ajustement et d'indépendance.


n = 3.
1.0
0.75
0.5
0.25
0.0
0.0
0.5
0.0
0.25
0.5
0.75
1.0

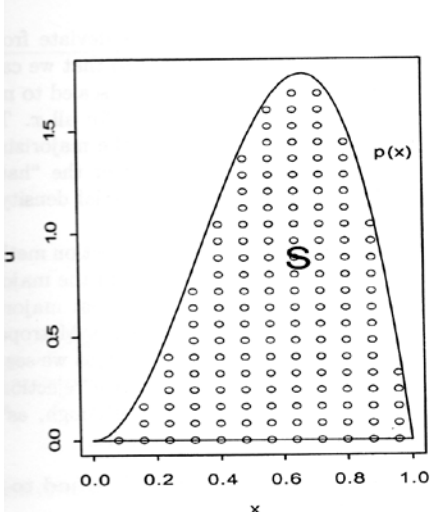
I.2 Simulation de variables de loi connue

- I.2.1 Méthodes générales
  - Anamorphose
    - Théorème:
      - Si X de loi F, alors F(X) uniforme sur [0;1]
      - $F^{-1}(R)$ est de distribution F



Exemple: loi exponentielle
F(x)= 1-exp(-cx)
X=-ln(1-R)/c ou X=-ln(R)/c

- Méthode du rejet (Von Neumann)
  - X à support borné. On prendra 0<X<1
  - Soit m tel que f(x)<m
  - Tirer R uniforme entre 0 et 1 et V uniforme entre 0 et m
  - Si V<f(R) alors R=X . sinon R est rejeté et on recommence



  - Suppose f facilement calculable. Taux de rejet souvent élevé

John von Neumann
1903-1957

- J.Dieudonné:

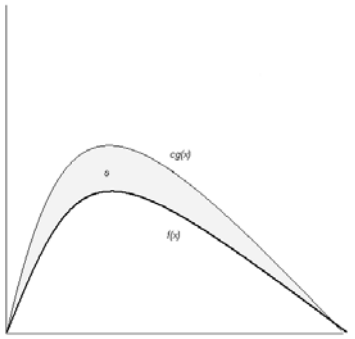
« Le dernier des grands Mathématiciens »

- Analyse fonctionnelle
- Théorie des jeux
- Informatique

Rejet amélioré

- Utiliser une autre fonction de densité g facilement simulable, telle que $cg(x) \geq f(x)$
- On génére alors un couple (y ; r) de réalisations indépendantes de Y de densité g, et de R uniforme.
- Si $$R < \frac{f(y)}{cg(y)}$$

  y est accepté comme réalisation de X de densité f(x). Sinon on rejette la valeur y et on recommence. Si X est à support borné, on prendra par exemple pour Y une loi triangulaire.
- S surface comprise entre les deux courbes, taux de valeurs acceptées = 1/(S+1)



I.3 Méthodes spécifiques

- Bernoulli : tirer R, si R<p alors X=1, sinon X=0
- Loi de Gauss
  - Théorème central limite: tirer 12 valeurs de R, $(R_1+R_2+..+R_{12})-6$ suit approximativement une N(0;1)
  - Box-Muller
  - Bray-Marsaglia

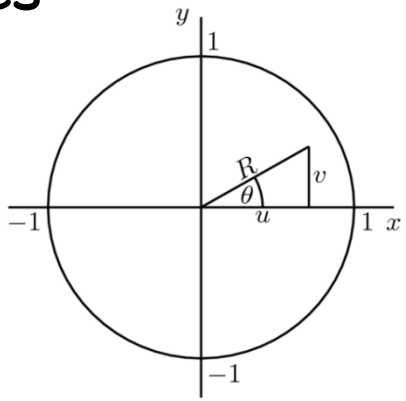
Méthode de Box-Muller

Soit U_1 et U_2 deux variables uniformes sur [0;1] indépendantes, alors:

$$\sqrt{-2\ln(U_1)}\cos(2\pi U_2) \text{ et } \sqrt{-2\ln(U_1)}\sin(2\pi U_2)$$

suivent des lois N(0;1) indépendantes

Forme polaire pour une $N_2(0;I)$



$$R^2 = u^2 + v^2$$
$$\cos\theta = \frac{u}{R}$$
$$\sin\theta = \frac{v}{R}$$

R^2 suit un χ^2 à deux ddl, $R^2/2$ une exponentielle
θ suit une loi uniforme sur $[0;2\pi]$

Variante de Bray-Marsaglia

Méthode de rejet

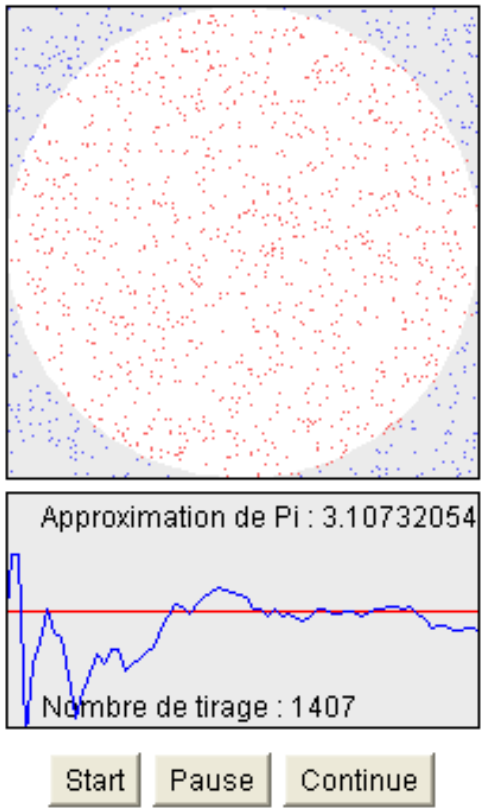
- Engendrer r_1 et r_2 uniformes indépendantes sur [0;1]
- $U_1=2\ r_1-1$ $U_2=2\ r_2-1$ uniformes indépendantes sur [-1;1]
- On rejette U_1 et U_2 si $U_1^2+U_2^2>1$
- Il reste une distribution uniforme dans le cercle

$$\sqrt{\frac{-2\ln R^2}{R^2}}U_1 \quad \text{et} \quad \sqrt{\frac{-2\ln R^2}{R^2}}U_2$$

sont deux réalisations indépendantes d'une N(0;1)

estimation de π

- http://www-sop.inria.fr/mefisto/java/tutorial1/node15.html#SECTION00033120000000000000



I.4 Simulation d'un vecteur aléatoire gaussien

- Deux méthodes voisines:
  - ACP sur matrice de variance covariance Σ, les composantes principales sont des combinaisons linéaires gaussiennes et indépendantes que l'on simule
  - Transformation de Mahalanobis. On simule des N(0;1) et on effectue ensuite la transformation inverse

II Applications

II.1 Simulation de fonctions de variables aléatoires

- $Y = f(X_1, X_2, .. X_p)$ fonction de variables de lois connues. En général difficile de donner la loi (densité ou fonction de répartition) de Y même dans des cas simples.
- Si les X_i sont indépendantes, pour obtenir un échantillon artificiel de Y : il suffit de générer indépendemment une valeur de chaque variable, de calculer f et de recommencer

exemple

- Calcul du volume Y de réserves récupérables d'un champ pétrolifère

$$GRV \times \emptyset \times N/G \times Sh / FVF \times RF$$

Input Parameter
Gross Rock Volume
Porosity
Net-to-Gross
Hydrocarbon Saturation
Formation Volume Factor
Recovery Factor

Benmore, R., Cooper, M., Wells, B., "Stochastic Modelling for Reducing Risk in Prospect Evaluation", AAPG Annual Convention, Houston, Tx., March 1995

Table 1: Drillable Prospect Reservoir Parameter Table

Parameter	Lowside	Most Likely	High Side
Gross Rock Volume(Acre-feet)	7600	70150	246400
Porosity (%)	14	18	24
Net to Gross(%)	26	40	58
Hydrocarbon Saturation(%)	15	48	78
Formation Volume Factor(RB/STB)	1.38	1.25	1.12
Recover Factor(%)	10	15	30

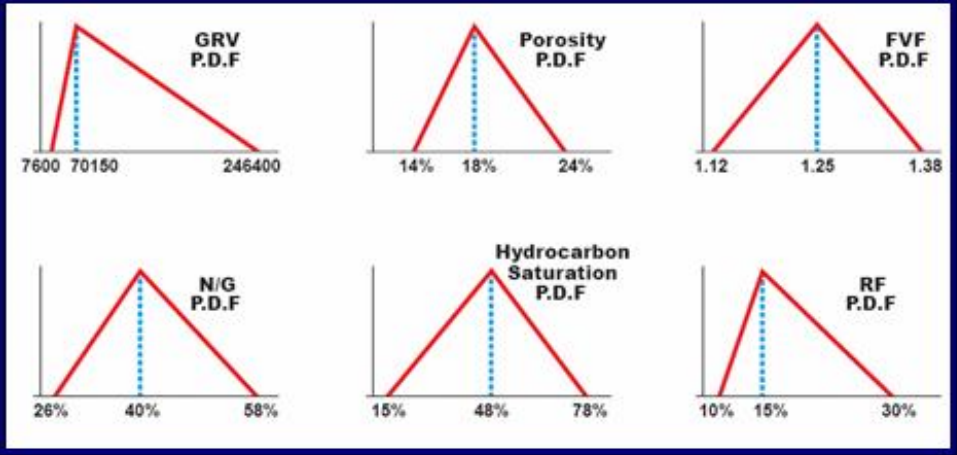


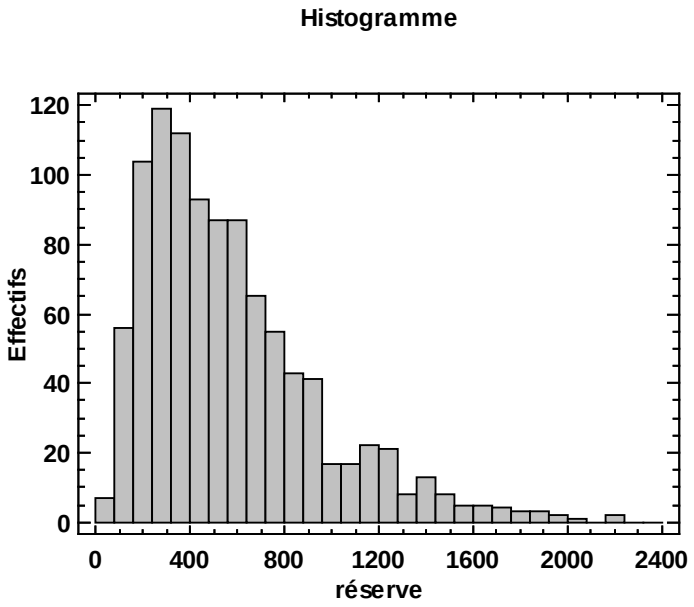
Figure 1: Reservoir Parameter Triangular Probability Distributions

En simulant 1000 valeurs de Y on trouve :

	GRV	P	NG	HS	RF	FVF	réserve
1	56023.9	0.220741	0.438213	0.392735	0.156242	1.25748	264.446553239
2	53210.2	0.185953	0.517904	0.321516	0.136371	1.22415	183.542787967
3	57240.3	0.177645	0.314504	0.428436	0.214454	1.20505	243.835032975
4	53223.9	0.211714	0.486297	0.533004	0.208118	1.15873	524.584824817
5	122606	0.209305	0.361224	0.41779	0.132169	1.22393	418.214406652
6	47329.3	0.217159	0.362596	0.572333	0.150839	1.32884	242.114423664
7	122608	0.194195	0.381732	0.579309	0.162393	1.24922	684.469454126
8	158408	0.188349	0.369238	0.454667	0.202875	1.23301	824.142265986
9	42778.8	0.20021	0.380509	0.454249	0.258159	1.30324	293.248733921
10	83575.9	0.177836	0.467661	0.498827	0.11335	1.36766	287.359281489
11	94690.1	0.167059	0.43176	0.525784	0.140194	1.16142	433.475294646
12	99284.4	0.193857	0.437648	0.487003	0.229221	1.28554	731.455477248
13	133230	0.157045	0.389395	0.43234	0.169033	1.23166	483.417750142
14	50915.9	0.214575	0.438096	0.631195	0.193767	1.22652	477.277064534
15	78891.4	0.215187	0.475378	0.396659	0.215775	1.22246	565.026157446
16	45502.6	0.185391	0.464125	0.713851	0.21344	1.24103	480.685394599
17	36992	0.174435	0.413387	0.731486	0.171583	1.29885	257.76238037
18	72899.3	0.198735	0.388209	0.407771	0.131901	1.25163	241.686157553
19	59862.6	0.183378	0.499169	0.511464	0.258229	1.24303	582.222369921
20	64359.2	0.209574	0.334156	0.594514	0.16226	1.1274	385.649551393
21	150726	0.221734	0.340023	0.514681	0.22389	1.23626	1059.23366471
22	56267.1	0.222894	0.407025	0.325565	0.218587	1.33905	271.293407714
23	44221.7	0.174319	0.33451	0.443093	0.23785	1.20402	225.711459468
24	75912.3	0.173226	0.50562	0.497355	0.183724	1.23615	491.485465007
25	114745	0.197879	0.382554	0.519608	0.151806	1.22386	559.833973096
26	68931.6	0.193423	0.350218	0.339404	0.110057	1.25121	139.402113324
27	175014	0.170844	0.346957	0.446705	0.236098	1.25071	874.791743545
28	123338	0.178268	0.393756	0.262381	0.20889	1.35647	349.814105985
29	171858	0.164691	0.403785	0.420791	0.135892	1.20713	541.372375165
30	87432.6	0.187101	0.479574	0.685097	0.190057	1.24378	821.29182188
31	62137.8	0.182718	0.459289	0.282873	0.169886	1.15275	217.388820594
32	42147.9	0.188025	0.4204	0.504728	0.176073	1.20686	245.328223731
33	98833.5	0.163312	0.389239	0.505146	0.130344	1.34396	307.794064537
34	71813.6	0.1799	0.323942	0.482909	0.145919	1.31905	223.573850643
35	97160.4	0.190687	0.505394	0.524558	0.268818	1.23724	1067.18170191
36	26391.2	0.185931	0.393326	0.529285	0.161417	1.26188	130.6725599

Statistiques résumées pour réserve

Moyenne	**571.188**
Ecart-type	369.078
Coef. de variation	64.6158%
Minimum	28.5575
Maximum	2220.11
Etendue	2191.55
Asymétrie std.	16.6978
Aplatissement std.	11.7676



Quantiles pour réserve

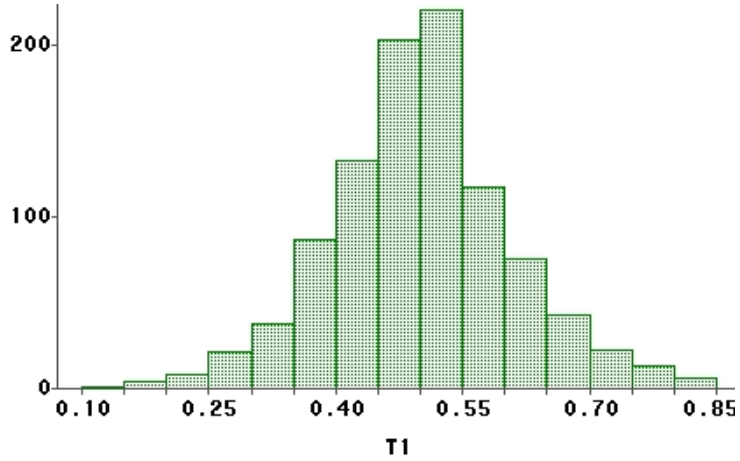
	Quantiles
1.0%	87.6372
5.0%	148.66
10.0%	189.576
25.0%	303.302
50.0%	483.626
75.0%	749.64
90.0%	1117.42
95.0%	1302.86
99.0%	1766.73

II.2 Distributions d'échantillonnage de statistiques complexes

• Distribution approchée de statistiques complexes, et même de statistiques simples quand la population a une distribution peu maniable.
• Répéter N fois la simulation d'un n-échantillon de X pour obtenir N valeurs de la statistique d'intérêt T

• Exemple : U variable aléatoire uniforme sur [0 ; q] le milieu de l'étendue d'un n-échantillon est un estimateur de q/2 noté T.
• Quelle est la distribution et la variance de T pour n=5 ?

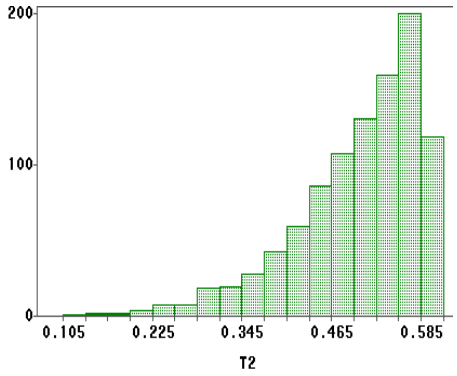
• En prenant q=1, on simule N=1000 échantillons de 5 valeurs d'une loi uniforme, on calcule à chaque fois T , d'où sa distribution empirique.



$\bar{t} = 0.5003752 \qquad s = 0.1106459$

Remarque : l'estimateur sans biais de variance minimale est:

$$\frac{n+1}{2n}\sup(X_1;...;X_n)$$



II.3 Calcul d'intégrales (méthode de Monte Carlo)

- **Première méthode :** $I = \int_0^1 g(t)dt = E(g(U))$

on simule *n* valeurs de *U*

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} g(u_i)$$

- **Deuxième méthode:** fonction d'importance

T variable sur [0 ;1] de densité p(t)

$$I = \int_0^1 \frac{g(t)}{p(t)} p(t)dt = E\left(\frac{g(T)}{p(T)}\right) \qquad \hat{I} = \frac{1}{n}\sum_{i=1}^{n} \frac{g(t_i)}{p(t_i)}$$

III Méthodes de rééchantillonnage

- Facile d'obtenir les distributions d'échantillonnage d'estimateurs dans le cas classique où l'on dispose d'un modèle parametrique f(x ;θ).
- En l'absence de modèle réaliste, comment simuler une distribution inconnue ?
- **Réponse :** simuler une distribution proche. La meilleure en l'absence d'information, n'est autre que la distribution empirique.
- **Rééchantillonnage :** tirer au hasard des observations dans l'échantillon dont on dispose.
  - Bootstrap : tirages avec remise
  - Jack-knife : tirages sans remise.

Le bootstrap (B.Efron)

- Soit une variable X de loi F inconnue ; on dispose d'un échantillon $(x_1,x_2,..,x_n)$ et on veut étudier par exemple la distribution d'un estimateur T d'un certain paramètre θ, calculer sa variance, en donner un intervalle de confiance.

- Tirage iid dans la loi empirique F_n

  Tirage avec remise de n parmi n

- Si le nombre de réplications B tend vers l'infini , la moyenne de toutes les estimations bootstrap converge vers l'estimateur du maximum de vraisemblance empirique (c'est à dire utilisant la loi F_n) et permet ainsi d'estimer sa variance. En pratique on se contente de quelques centaines de tirages au plus.

- En anglais, le ***bootstrapping*** fait référence aux aventures du baron de Münchhausen, lequel est censé s'être sorti d'un marécage où il était embourbé rien qu'en se tirant par les bottes et se propulsant ainsi dans les airs. Les ***bootstraps*** sont les anneaux, en cuir ou en tissu, cousus sur le rebord des bottes et dans lesquels on passe les doigts pour s'aider à les enfiler. (Wikipedia)

« To pull oneself up with his one booststrap »

- Thinking the unthinkable (Efron)

intervalles de confiance bootstrap

- Méthode des percentiles : repérer les quantiles dans la distribution des *B* valeurs.

Approximation normale : moyenne et l'écart-type des *B* réalisations puis intervalle à $\pm$ 1.96 écart-types.

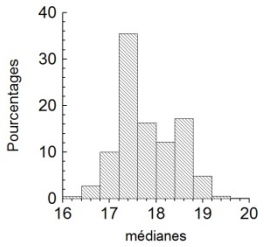
(Vérifier la normalité approximative de la distribution des *B* valeurs).

Exemple :

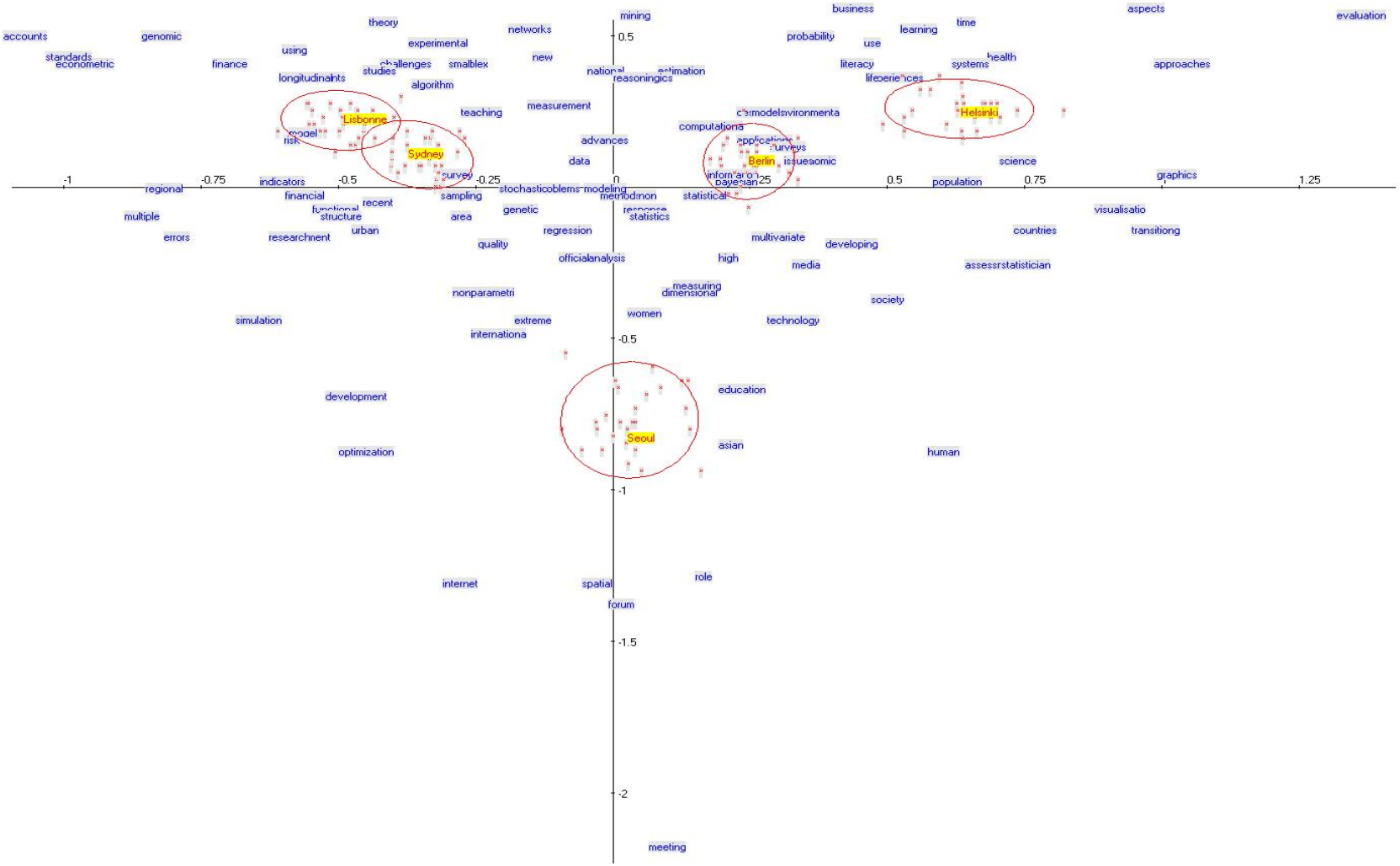
Médiane du taux de taxe d'habitation de 100 communes françaises.
médiane =17.625 *B*=1000

- Intervalle de confiance à 95 % des percentiles [16.70 18.92]
en prenant respectivement la 25ème et la 975ème valeur ordonnée.
- Intervalle avec approximation normale [16.55 19.02]

Moyenne = 17.7872
Médiane = 17.625
Ecart-type = 0.630658
Minimum = 15.87
Maximum = 19.39



Stabilité de résultats en analyse factorielle



Logiciel DTM http://www.dtmvic.com

- Méthode très générale qui permet de répondre à des problèmes jusque là quasi impossibles à résoudre comme l'étude de la variabilité de résultats d'analyses factorielles (valeurs propres, vecteurs propres etc.) ou l'estimation de variance dans des sondages complexes.
- Si la taille n de l'échantillon de départ est faible, en général sous-estimation de la variabilité : les intervalles de confiance auront tendance à être trop petits (couverture insuffisante).

  - En effet le rééchantillonnage ne permet pas par définition d'engendrer des valeurs autres que celles déjà observées, ce qui peut être génant pour des variables numériques, mais l'est moins pour des variables qualitatives où en général, toutes les modalités sont observées, au moins marginalement.
  - Le bootstrap est une méthode d'étude de la variabilité intrinsèque à un échantillon.

Références

- site internet de Pierre Lécuyer (http://www.iro.umontreal.ca/~lecuyer).
- J.Gentle: *Random Number Generation and Monte Carlo Methods*, Springer, 2003