Epreuve de modélisation (Option Proba/Stats)
Exemple de simualution et d'analyse de phénomènes aléatoires. Fluctuations d'échantillons.

Le problème de la simulation des lois classiques à partir d'un générateur de variables uniformes a été vu dans les séances précédentes dans une approche "à la main". Par ailleurs vous disposez sous scilab ou sous matlab avec le complément stixbox des routines nécessaires pour simuler directement pas mal de lois classiques. Quelques compléments possibles sur le sujet sélectionnés pour vous en français que nous vous invitons à consulter :

- Le chapitre 1 du très bon cours NOISE de C. Robert (Université Paris X)[1] et le tp associé
- Le petit texte de Floriant Malrieu (Université de Rennes 1) sur la simulation de variables aléatoires (avec des exercices)[2]
- Le tp de Gilles Stoltz (ENS Paris)[3]

Dans ce tp, nous cherchons plutôt à travailler autour de la notion d'indépendance, d'échantillon et de fluctuations aléatoires (cf expérience 1) que nous observerons sous différents points de vue et en travaillant ici sur l'utilisation de la théorie des probabilités et des statistiques pour la modélisation et l'analyse de situations réelles (cf expériences 3 et 4).

Expérience 1.

(1) Construire la fonction `function X=rndber(p,N)` simulant un N échantillon de va de Bernoulli de paramètre $p \in [0,1]$. On prendra pour X un vecteur ligne.

(2) Exécuter pour $N = 10, 100, 1000, 10000$ la commande : `plot(cumsum(rndber(0.5,N)) ;`. Comparer les échelles en x et y et expliquer ce qui se passe (on rappelle l'inégalité de Hoeffding :

$$P(S_n/n - E(S_1) > \epsilon) \leq \exp(-2n\epsilon^2)$$

si $S_n = \sum_{i=1}^n X_i$ avec X_i i.i.d. telles que $0 \leq X_i \leq 1$).

(3) Exécuter pour $N = 10, 100, 1000, 10000$ la commande :
`plot(cumsum(rndber(0.5,N)-0.5*(1 :1 :N)))` ;
Comparer les échelles en x et y et expliquer ce qui se passe.

(4) Que se passe-t-il si l'on reproduit ce genre d'expériences avec des lois de Cauchy (utiliser la méthode d'inversion de la fonction de répartition) ?

Expérience 2. En utilisant la méthode du rejet,

(1) Ecrire une fonction permettant de tirer N points uniformément dans l'intérieur de l'ellipse d'équation $ax^2 + by^2 < 1$ avec a et b (libres) positifs. Construire un programme permettant de calculer l'aire de l'ellipse par la méthode de Monte Carlo avec une précision donnée construite sur un intervalle de confiance à 95%. Faire tourner votre programme pour l'ellipse d'équation $x^2 + 2y^2 < 1$;

(2) Modifier le programme précédent pour calculer l'aire de l'intersection de l'ellipse d'équation $x^2 + 2y^2 < 1$ et de l'ellipse d'équation $2x^2 + y^2 < 1$. Déterminer le plus petit carré contenant l'intersection et en déduire une méthode plus rapide que la précédente.

Expérience 3. *Un joueur conteste le "hasard programmé" des jeux de grattage* Lire ici l'article du Monde du Mardi 30.01.06, ici ceux du journal *20 minutes* parus sur le sujet Vendredi 03.02.06 et ici un article du Monde du 31.05.06.

De quoi s'agit-il ? La Française des jeux propose depuis plusieurs années des jeux de "grattage" sous la forme d'un ticket d'un prix de 2 ou 3 euros et comportant plusieurs zones à gratter. En les grattant, le joueur peut gagner différentes sommes qui sont regroupées en un lot indivisible de quelques euros jusqu'a plusieurs milliers d'euros.

[1] http ://www.ceremade.dauphine.fr/~xian/Noise/CChap1.pdf
[2] http ://name.math.univ-rennes1.fr/florent.malrieu/AGREG/simulation.pdf
[3] http ://www.dma.ens.fr/~stoltz/TPs/TP3.pdf

Nous nous intéresserons ici à un jeu lancé en Mars 2000, Vegas, qui permet de gagner jusqu'a 40000 euros (voir le site ici). Comment sont fabriqués les tickets ? Ils sont imprimés par blocs de 500000, et les différents lots, prédéfinis à l'avance (voir leur ventilation ici) sont répartis aléatoirement dans les 500000 tickets. Les tickets sont ensuite regroupés par carnets (ou bandes) de 50 tickets et c'est sous la forme de ces carnets qu'ils arrivent chez les détaillants. Les tickets de Vegas sont vendus 3 euros pièce.

(1) A partir du fichier de ventilation, construire une matrice G contenant dans la premiere colonne les nombres de lots et dans la deuxième, les montants (organiser par ordre de gains croissants) où télécharger la ici. Calculer l'espérance du gain sur un ticket et sa variance (commentaires ?). Simuler l'accumulation des gains d'un joueur fictif sur 10, 100, 1000, 10000, 100000 et 1000000 de tickets (faire la même chose en retirant le prix de chaque ticket). On pourra utiliser une version vectorielle `rvmulti` de `rmulti` téléchargeable ici.

(2) On veut étudier ici les gains d'un joueur modéré sur un an et mieux comprendre le jeu en prenant le point de vue du vécu du joueur sur cette durée. On suppose qu'il gratte 100 tickets par an (soit environ 2 par semaine). Calculer la loi du lot maximal obtenu par le joueur ainsi que sa fonction de répartition (mettre en x les codes 0, 1, 2, etc correspondant aux différents niveaux de lots plutôt que les valeurs réelles des lots pour éviter que tous les petits lots soient "écrasés" au même endroit). Montrer que le joueur n'a presque aucune chance de voir passer un lot de 500 euros ou plus mais qu'il a *toutes les chances* d'avoir obtenu un lot de 20 euros ou plus (quantifier cela précisément).

(3) On se concentre maintenant sur les 99.5% des joueurs qui ne verront passer que des lots au plus de 200 euros. Calculer pour eux leur espérance de gain et ainsi que la variance. Quelle est la proportion des mises qui est redistribuée à cette masse de joueurs ? Construire un intervalle de confiance bilatère sur bilan total (gains+mises) sur un an de niveau de confiance asymptotique de 95% en vous basant sur l'approximation du tcl. Faire la même pour les 84.7% des joueurs qui ne verront pas passer de lots supérieurs à 100 euros ? Combien auront-ils perdus en moyenne ?

(4) En consultant le site du jeu Vegas, on y apprend ici qu'en 2004, 204 heureux joueurs ont remporté la somme de 40 000 euros. Comment évaluer à partir de cette information le nombre de tickets "Vegas" vendus par la Française des jeux cette année là et des sommes perdues par les joueurs ?[4]

(5) Venons-en à l'objet du titre de ce texte. En fait, ce qui n'est pas précisé dans les articles, c'est que M. Riblet, pour en avoir le coeur net, a acheté 100 *carnets complets* de tickets Vegas sur ses propres deniers et les a tous grattés. Il a noté alors plusieurs faits troublants concernant la somme des petits lots sur les tickets mais surtout et c'est le point que nous retiendrons ici, qu'aucun carnet de contenait plus d'un lot supérieur ou égal à 20 euros. Plus précisement, 1/3 de ces carnets ne contenait aucun lot de plus de 20 euros et 2/3 exactement un. Il prétend que la Francaise des Jeux ne distribue pas aléatoirement les lots sur les tickets mais opère une homogénisation sur chaque carnet pour répartir les lots intéressants sur un maximum de carnets. Ce que la FdJ dément en argumentant sur le fait que l'échantillon de 100 carnets est trop petit pour pouvoir fournir des conclusions signiticatives. Elle assure qu'il existe bien des carnets contenant plusieurs lots à plus de 20 euros mais que Riblet n'en a pas eu car son échantillon n'est pas assez grand. Les probabilités et les statistiques peuvent-elles dire quelque chose dans ce débat ?

   (a) Calculer, si les lots sont distribuées parfaitement au hasard sur les tickets et dans les carnets (on appellera cela l'hypothèse H_0), la probabilité d'avoir 0, 1 et au moins 2 lots à plus de 20 euros dans un carnet de 50 et comparer sur un diagramme en barres (voir commande `bar` sur matlab) avec les fréquences observées par M. Riblet (qu'en pensez-vous ?).

   (b) Proposer au moins une façon de montrer avec un risque inférieur *garanti* à 10^{-6} que H_0 doit être rejetée.

   (c) Imaginons donc que la FdJ utilise un procédé pour homogénéiser les carnets mais qu'elle distribue cependant parfaitement au hasard les carnets chez les détaillants. Peut-on donner une majoration au niveau 1% de la probabilité p qu'un carnet contienne au moins 2 lots à plus de 20 euros ?

[4] La FdJ vend au total sur tous ses jeux environ 2 milliards de tickets à gratter par an.

Il s'agit de construire un intervalle de confiance unilatère (ici de la forme $[0\ b[$) en utilisant la méthode générale pour pouvoir garantir les niveaux (cf BIC-DOK p235). Soit $\alpha \in]0,1[$, et pour tout $0 \leq p \leq 1$, $k_p \doteq \sup\{\ k \in \mathbb{N} \mid P_p(N \geq k_p) \geq 1-\alpha\}$ où N est le nombre de carnets contenant plus de 2 lots supérieurs ou égaux à 20 euros dans un échantillon de 100 carnets. Montrer que $P_p(\ p \in R_N) \geq 1-\alpha$ où $R_n = \{\ p \in [0,1] \mid k_p \leq n\}$ puis vérifier que $R_0 = [0, p_0[$ où $(1-p_0)^{100} = \alpha$. En déduire la borne demandée au niveau $\alpha = 1\%$.[5]

Expérience 4. Avant le premier tour de la dernière présidentielle, les instituts de sondages ont effectué des sondages pour de nombreux commanditaires, qui ont été publiés dans la presse et ont été souvent présentés comme un moyen pour mesurer les évolutions de l'opinion durant la campagne et notamment l'impact des interventions des candidats dans les médias. Consulter ici des informations sur le sujet, en particulier les sondages de la Sofres.

Plaçons nous dans une situation idéale où l'on tire au hasard n votants qui nous disent la vérité sur leur intention de vote. On suppose que l'on a r candidats $c_1, \cdots, c_r$ et que les pourcentages de voix exprimées recueillis au final par chacun des candidat sont $p_1, \cdots, p_r$.

(1) Construire une fonction `s=sondage(p,n)` simulant les réponses de n personnes tirées au hasard dans la population des votants : `p` est le vecteur contenant les proportions des votes pour chaque candidat et `s` est le vecteur contenant les proportions pour chaque candidat dans les personnes interrogées. Simuler 50 sondages différents sur 700 personnes et comparer les différents résultats des sondages par un graphique adéquat en prenant pour p les valeurs données par les résultats officiels donnés en annexe. Comparer avec les résultats des sondages Sofres fournis en annexe. Quelque chose ne saute-t-il pas aux yeux ?

(2) Refaire l'expérience précédente mais afficher pour chaque nouveau sondage k une moyenne escomptée avec les précédents : $\tilde{s}_k = \lambda s_k + (1-\lambda)s_{k-1}$. Expliquer l'effet produit en fonction de λ et pourquoi cette technique donne des résultats plus *présentables*. Expliquer pourquoi il est illusoire de prétendre suivre les évolutions des intentions de vote au jour le jour et la contradiction dans laquelle se trouvent les instituts de sondage.

(3) Construire pour les trois premiers candidats des intervalles de confiance de niveau de confiance $1-\alpha$ avec $\alpha = 0.05$ (en utilisant la méthode classique par approximation par le Tcl) et afficher les résultats en utilisant la fonction `errorbar`. N'est-il pas étonnant qu'aucun sondeur n'est pronostiqué l'élimination de Jospin au premier tour ?

(4) Faire un test du chi2 d'adéquation entre le dernier sondage proposé par la Sofres et les résultats officiels (utiliser la fonction `pchisq` ou `qchisq`).

Expérience 5. On veut tester ici la qualité de l'approximation du tcl sur les binomiales.

(1) Soit $(X_i)_{i\geq 1}$ une suite i.i.d. de loi B_p et l'on note $S_n = \sum_{i=1}^n X_i$. En utilisant l'approximation du tcl, calculer une valeur $t_\alpha(n,p)$ tel que

$$P(|\overline{X}_n - p| > t_\alpha) \simeq \alpha$$

où $\overline{X}_n$ est la moyenne empirique (on prendra la valeur $\alpha = 0.05$).

(2) Tracer pour différentes valeurs de p entre 0.1 et 0.5, la fonction $n \to P(|\overline{X}_n - p| > t_\alpha(n,p))$ pour n entre 1 et 1000 en utilisant la fonction `pbinom` (on pourra tracer également la droite d'équation $y = 0.06$).

(3) On dit souvent que l'approximation du tcl est valide si $np(1-p) \geq 30$. Qu'en pensez-vous ?

[5] Questions subsidiaires : En prenant la valeur de p_0 trouvée, peut-on imaginer pourquoi M Riblet trouve les proportions 1/3 et 2/3 ? Comment évaluer l'espérance de gain pour un joueur informé qui n'acheterait des tickets que dans des carnets où aucun lot de plus de 20 euros n'a encore été trouvé ? Puis celle d'un joueur qui sans le savoir n'acherait que les tickets résiduels ? A suivre...