TP 1
Comment estimer (et écouter) les fluctuations de variables aléatoires

Simulation, moyenne géométrique, estimateurs, sonorisation et Matlab

Partie 1

Le principe de la simulation par tirage aléatoire

Le but de cet exercice est d'expliquer le principe des simulations que nous allons faire dans les séances à venir. Elles consistent à remplacer une variable aléatoire par une série de nombres ayant la même densité de probabilité.

Considérons une variable aléatoire X définie par sa densité de probabilité $f(x)$. Si on arrive à générer une série de nombres $s = \{x_i\}$ ayant un histogramme qui ressemble à $f(x)$ on pourra remplacer X par s dans toutes les opérations à effectuer sur X. Nous allons illustrer cette méthode par quelques exemples.

1 D'après la notice de Matlab la fonction $randn$ permet de générer une suite de nombres ayant une distribution normale réduite (c'est-à-dire de moyenne nulle et d'écart-type 1). Pour vérifier si c'est bien vrai on va tracer l'histogramme correspondant et on va calculer la moyenne et l'écart-type. On pourra utiliser le code Matlab ci-dessous.

```
% Histogramme d'une variable normale simulée
n=1000 ; A1=randn(n,1) ;
subplot(2,1,1);
hold on
X=-10:0.2:10 ; hist(A1,X)
h = findobj(gca,'Type','patch');
set(h,'FaceColor','r','EdgeColor','w')
A2=randn(n,1) ;
A2=4+A2/3 ; hist(A2,X)
h = findobj(gca,'Type','patch');
set(h,'EdgeColor','w')
hold off
m1=mean(A1)
sig1=std(A1)
m2=mean(A2)
sig2=std(A2)
```

Les 2 commandes relatives à h ont pour but de mettre un liseré blanc entre 2 marches de l'histogramme

et de le colorer en rouge.

Quel est en $\%$ l'écart entre m1,sig1, m2,sig2 et les valeurs théoriques correspondantes?

2 Simuler des opérations sur des variables aléatoires.

A titre d'exemple d'opération, on additionne les variables $A1, A2$ et on veut voir si la somme a bien les caractéristiques prédites par la théorie des probabilités.

```
% Addition des variables aleatoires A1 et A2
subplot(2,1,2)
B=A1+A2 ; hist(B,X)
h = findobj(gca,'Type','patch');
set(h,'FaceColor','r','EdgeColor','w')
mb=mean(B)
sigb=std(B)
sigth=sqrt(1+(1/3)^2 )
```

Pouvez-vous vérifier la règle selon laquelle: "La somme de deux variables gaussiennes indépendantes $N(m_a, \sigma_a)$ et $N(m_b, \sigma_b)$ est elle-même une variable gaussienne $N(m_a + m_b, \sqrt{\sigma_a^2 + \sigma_b^2}$", Quel est en $\%$ l'écart par rapport à la valeur attendue?

Une variable aléatoire a-t-elle toujours une moyenne?

Dans l'optique précédente où on remplace toute variable aléatoire par un échantillon de nombres on peut bien sûr toujours calculer la moyenne de cette série de nombres. Pourtant il arrive que cette moyenne n'a pas de signification réelle du fait qu'elle change de façon dramatique lorsqu'on passe d'un échantillon à un autre. Ce symptôme suggère que la variable aléatoire probabiliste sous-jacente a une moyenne infinie ce qui veut dire qu'en fait sa moyenne n'existe pas du fait que l'intégrale $m = \int_{-\infty}^{\infty} xf(x)dx$ *diverge. Ce cas, illustré dans l'exercice qui suit, n'est nullement une curiosité mathématique mais trouve au contraire une illustration très concrète dans les distributions de revenu.*

1 Simulation d'une variable aléatoire sans moyenne

On part d'une variable aléatoire U ayant une distribution constante sur $(0, 1)$ et zéro ailleurs. Une telle distribution est simulée par la fonction rand(.). A partir de U on définit une nouvelle variable qui est son inverse $A = 1/U$. Puis on calcule la moyenne et la valeur maximum de A et on construit l'histogramme de sa fonction densité. Ces étapes sont réalisées dans le script ci-dessous.

```
% Variable aleatoire sans moyenne n=50 ; U=rand(n,1) ;
A=1./U ;
mx=max(A)
b=mx/20 ; X=0:b:mx ; hist(A,X)
h = findobj(gca,'Type','patch');
set(h,'FaceColor','r','EdgeColor','w')
m=mean(A)
```

Relancez une dizaine de fois ce script et faites un tableau des moyennes et des valeurs maxima obtenues. Calculez la moyenne m et l'écart-type σ des 10 moyennes. Calculez aussi le coefficient de variation $CV = \sigma/m$ que vous exprimerez en %. Ce coefficient permet d'estimer la variabilité d'une grandeur. S'il est inférieur à 5% on dira que la grandeur est relativement bien définie, s'il est supérieur à 50% ou 100% on dira qu'elle est très fluctuante et par conséquent mal définie.

2 Calcul de l'espérance de la variable aléatoire A

Calculer l'espérance de $A = 1/U$ en utilisant la formule:

$$E[g(U)] = \int_{-\infty}^{\infty} g(x) f_U(x) dx$$

où $f_U(x)$ est la fonction densité de la variable U, c'est-à-dire dans le cas présent la variable uniforme générée par rand.
Conclusion?

Que donne ce calcul pour $A = 1/\sqrt{U}$?

3 Refaites pour $A = 1/\sqrt{U}$ la simulation faite en (1). Construisez à nouveau un tableau avec 10 valeurs (si nécessaire augmentez n) et calculez m, σ, CV. A quoi vous attendez-vous? Est-ce que la simulation confirme votre intuition?

4 Lien avec les distributions de revenu

a) Expliquez pourquoi et comment le cas considéré dans la question 2 se rattache à la distribution individuelle des revenus ou des richesses dans un pays. Quelle conséquence en résulte-t-il pour le revenu moyen?

b) Pour caractériser l'inégalité des revenus on a souvent recours à un estimateur simple qui est la part revenant aux 0.1% (ou aux 1%) revenus les plus élevés. Cette part qu'on peut noter $s(0.1)$ s'obtient par le calcul suivant.
Supposons qu'on ait $n = 10,000$ revenus.
(i) On les classe du plus grand au plus petit.
(ii) On fait le total des 10 revenus les plus élevés soit $S(0.1)$, ainsi que le total de tous les revenus soit S.
(iii) On calcule le rapport: $s(0.1) = S(0.1)/S$.
Que vaudrait $s(0.1)$ pour une distribution uniforme de revenus c'est-à-dire une distribution parfaitement égalitaire?
En prenant un échantillon de taille $n = 10,000$ calculez $s(0.1)$, d'une part pour une distribution normale de moyenne 4 et d'écart-type 1 et d'autre part pour la distribution de $A = 1/U$ de la question 2. Pour classer les nombres du plus grand au plus petit on pourra utiliser la fonction "sort(A,'descend')" de Matlab.

c) Pour situer les résultats qu'on vient d'obtenir par rapport à des données réelles on pourra utiliser les chiffres du tableau 1. Que vous suggère cette comparaison?

Table 1 Part du revenu total revenant à la fraction constituée des 0.1% revenus les plus élevés

Pays	1950	1970	1998
Etats-Unis	3.4%	2.1%	6.0%
Grande-Bretagne	3.4%	1.4%	3.4%
France	2.6%	2.2%	2.2%
Japon	1.8%	2.0%	1.8%

Notes: Ces statistiques sont basées sur les revenus par ménage (avant impôt et ne comprenant pas les revenus du capital) tels que déclarés à l'administration fiscale. *Source: Piketty (T.), Saez (E.) 2003: Income inequality in the United States 1913-1998. Quarterly Journal of Economics, Volume 118, p. 1-39.*

5 Utilisation de la moyenne géométrique

Revenons à la question (1). On a vu que la moyenne est mal définie parce que trop fluctuante. Ne

peut-on pas définir une autre grandeur qui serait plus stable et pourrait donc jouer le même rôle que la moyenne?
C'est en effet possible.
Il y a plusieurs possibilités. L'une d'elle est ce qu'on appelle la *moyenne géométrique* g et qui est en fait essentiellement la moyenne des logarithmes. Plus précisément[1]:

$$\text{Moyenne géométrique de X:} \quad g = \exp[E(\ln(X)]$$

Pourquoi prendre la moyenne des logarithmes? Précédemment, ce qui nous gênait était le fait que la variable A prenait des valeurs pouvant être très grandes. En prenant leur logarithme on rend ces valeurs bien plus petites; ainsi 10^6 va devenir $6 \times 2.3 = 13.8$.

Réutilisez le script de la question (1) mais en définissant cette fois A par A=log(1/U). A la fin du script on remplacera bien sûr m=mean(A) par g=exp[mean(A)].
Construire à nouveau un tableau correspondant à 10 itérations et calculer les moyenne, écart-type et coefficient de variation. Dans quel proportion le CV de g est-il plus petit que celui de m dans la question (1).

6 Les énormes fluctuations de la moyenne statistique venaient de ce que la moyenne probabiliste divergeait. Puisque les fluctuations de g sont bien plus modérées on s'attend à ce que la moyenne géométrique ne diverge plus. Est-ce bien le cas?
En utilisant la formule donnée dans la question (2), calculez $g = \exp[E(\ln(1/U)]$. Dans ce calcul on aura besoin de la primitive de $\ln x$ qui est $x \ln x - x$.
Est-ce que la valeur de g que vous obtenez est en accord avec celle fournie par la simulation?

7 **Non-existence de la variance** En utilisant à nouveau la formule donnée en (2), calculez la variance de $A = 1/\sqrt{U}$. Que constatez-vous et que vous suggère ce résultat? Pour tester votre intuition, relancez une dizaine de fois le script ci-dessus en incluant une instruction donnant l'écart-type. Est-ce que votre intuition est confirmée?

Il est important de réaliser qu'il y a des distributions n'ayant pas de variance car dans ces cas la plupart des théorèmes de la théorie des probabilités (tels que loi des grands nombres ou théorème de la limite centrale) ne s'appliquent pas.

Notion d'estimateur statistique

Une des difficultés majeures d'un cours de probabilité est de bien comprendre la relation entre probabilités et statistiques c'est-à-dire la relation entre la théorie des probabilités et le monde réel. C'est par la notion d'estimateur statistique d'une grandeur probabiliste que se fait le lien entre les deux domaines. Dans cet exercice on illustre ce lien en cherchant les estimateurs de la moyenne et de la variance.

A Estimateur de l'espérance

a) Dans la suite X désigne une variable aléatoire continue d'espérance m et de variance $D = \sigma^2$. Rappelons que l'expression de l'espérance est:

$$m = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

[1] Notez qu'après avoir pris le logarithme, il faut pour obtenir un ordre de grandeur raisonnable faire l'opération inverse en prenant l'exponentielle.

où $f(x)$ est la fonction densité de probabilité de X. Calculer $E(X)$ pour une variable ayant une densité constante sur $(0,1)$ et zéro ailleurs.

b) On fait une expérience E consistant en deux observations qui donnent les valeurs x_1, x_2 de X. Par exemple, si X est la période d'un pendule les nombres x_1, x_2 seront deux mesures de cette période. On veut définir un estimateur qui, à partir de ces deux nombres, permette de trouver m. Une expression possible est:

$$\widehat{m} = \frac{x_1 + x_2}{2}$$

Comment peut-on savoir si cet estimateur est correct ou non?

On va répéter l'expérience E un grand nombre de fois. Les différentes valeurs prises par x_1, x_2 formeront deux variables aléatoires X_1, X_2 ayant la même distribution de probabilité que X, si bien que l'estimateur $\widehat{m}$ deviendra lui-même une variable aléatoire

$$\widehat{M} = \frac{X_1 + X_2}{2}$$

On dira que l'estimateur $\widehat{M}$ est non biaisé si: $E(\widehat{M}) = m$. Trouver la réponse à cette question est un problème de théorie des probabilités mais dans cette séance, au lieu de faire une démonstration mathématique, nous allons simuler les variables X_1, X_2 en faisant des tirages de nombres au hasard comme dans les questions précédentes.

1 Simulation pour la moyenne

En décomposant l'instruction suivante en étapes successives expliquer l'opération qu'elle réalise.

mt=mean(mean(rand(2,5),1))

Expliquer en particulier le rôle du "1".

Quelle est la valeur de la moyenne m prévue par la théorie des probabilités? Si on remplace le nombre d'itérations initialement égal à 5 par un nombre plus grand, voit-on mt se rapprocher de m. Faites un tableau portant sur une dizaine d'observations et donnant en % l'écart $(m - mt)/m$.

B Les deux estimateurs de la variance.

2 Que fait la commande std de MATLAB?

La définition probabiliste de la variance est: Var $= \sigma^2 = E[(X - m)^2]$.

On suppose que X prend les valeurs 1 2 3. Matlab peut vous calculer l'écart-type de ces 3 nombres de deux façons:

a=std([1 2 3],0) ou b=std([1 2 3],1). Que trouve-t-on?

Chercher dans le help de Matlab à quelle formules différentes correspondent ces deux options? Cela pose naturellement la question de savoir laquelle de ces deux formules est la "bonne"? Pour le savoir on fait l'expérience suivante.

3 Quel est l'estimateur non biaisé?

On suppose que X a une distribution de probabilité uniforme sur l'intervalle $(0,1)$. Montrer que sa variance vaut $\sigma^2 = 1/12$. On voudrait donc que, le calcul statistique converge vers la même valeur.

4 Simulation pour la variance

On va faire $p = 5$ tirages de $n = 10$ nombres par les commandes ci-dessous.

sa=mean(std(rand(10,5),0,1))

sa=mean(std(rand(10,5),1,1))

De même que précédemment expliquer ce que font ces instructions; en particulier expliquer le rôle des 0 et des 1.

Si on remplace le nombre d'itérations initialement égal à 5 par des nombres plus grands voit-on sa et sb se rapprocher de $\sigma = 1/\sqrt{12}$?
Faites un tableau portant sur une dizaine d'observations et donnant en % les écarts par rapport à la valeur exacte.
Quelle est à votre avis l'estimateur non biaisé, c'est-à-dire celui qui se rapproche le mieux de la valeur théorique?

5 Complément: le cas où n est petit

Recommencer les mêmes tests relatifs a l'écart-type pour $n = 2, 3$.
Le résultat vous étonne-t-il? Pouvez-vous l'expliquer?