Emmanuel Jeandel

MÉTHODES STATISTIQUES EN INFORMATIQUE

TABLE DES MATIÈRES

INTRODUCTION

Voici une liste de livres dont je m'inspire pour le cours

Chapitre 1

ESTIMATEURS

Les statistiques comme on les examinera dans ce cours sont appelées *statistiques inférentielles* : On cherche à déterminer des informations sur une réalité à partir d'observations.

Définition 1.1 (*Population*)

Une population est un ensemble d'individus S. Une donnée est une variable aléatoire X sur S.

Le but est d'obtenir de l'information sur X, comme sa distribution, sa moyenne, etc. Pour ce faire, on va utiliser un échantillon :

Définition 1.2 (*Échantillon*)

Un échantillon $\overline{S} = (S_1, \ldots, S_n)$ est une variable aléatoire tel que $S_i \in S$ et pour tout individu s, $P[s \in S] > 0$. (C'est une façon de choisir n individus, pas forcément distincts, de sorte que tout individu ait une chance d'être choisi) On note n la taille de l'échantillon. On note X_i la valeur de X en S_i.

Il y a principalement deux façons d'échantillonner (mais on en verra d'autres en TD) :

Sans remise On choisit n individus distincts, uniformément parmi tous les individus possibles.

Avec remise On choisit, pour chaque i, uniformément parmi tous les individus possibles.

On se place dans toute la suite sous la deuxième hypothèse. On appelle un tel échantillon un échantillon *aléatoire*.

Définition 1.3 (*Échantillon aléatoire*)

Un échantillon aléatoire $\overline{X} = (X_1, \ldots, X_n)$ est une variable aléatoire tel que tous les X_i sont indépendants et de même distribution que X

1.1 Espérance

Dans cette partie, on cherche à estimer la moyenne de la population, c'est à dire $E(X) = \mu$. Mais que veut dire estimer ?

Définition 1.4 (*Estimateur*)

Un estimateur pour une valeur θ est une fonction $\hat{\theta}(X_1, X_2 \ldots X_n)$ d'un échantillon.

Par exemple $\hat{\theta}(X_1, X_2) = 3 \log X_2 + \pi - 1664 \cos X_1$ est un estinateur de μ, mais il n'a pas l'air très bon. $\hat{\theta}(X_1, X_2) = \frac{X_1 + X_2}{2}$ semble meilleur.

Définition 1.5 (*Estimateur non-biaisé*)

Un estimateur $\hat{\theta}$ pour une valeur θ est non-biaisé si $E(\hat{\theta}) = \theta$

Théorème 1.1

$\hat{X} = \frac{X_1+...X_n}{n}$ est un estimateur non biaisé de μ.

Preuve : Par linéarité de l'espérance :

$$E\left(\frac{X_1 + \dots X_n}{n}\right) = \frac{E(X_1) + \dots E(X_n)}{n}) = \frac{nE(X)}{n} = E(X)$$

■

Si on dispose d'un échantillon de taille n, on peut obtenir plusieurs estimateurs non biaisés de μ : On peut faire la moyenne des n termes, ou simplement la moyenne des 3 premiers termes. Le premier semble meilleur. Comment le mesurer ?

Définition 1.6 (*Qualité d'un estimateur*)

La qualité d'un estimateur non-biaisé $\hat{\theta}$ est $Var(\hat{\theta})$.

Théorème 1.2

Si $\hat{X}$ est la moyenne sur un échantillon de taille n, alors $Var(\hat{X}) = \frac{Var(X)}{n}$.

(Preuve à faire)

La moyenne sur 3 échantillons est donc toujours meilleure que la moyenne sur 2. On peut montrer qu'en un certain sens, la meilleure façon d'estimer l'espérance à partir d'un échantillon de taille n est de faire la moyenne des n termes. Mais ce n'est pas toujours vrai, on le verra en TD.

1.2 Variance

Essayons maintenant d'estimer la variance.

Théorème 1.3

$\hat{\theta} = \frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2$ est un estimateur biaisé

Preuve : Par linéarité de l'espérance :

$$E(\hat{\theta}) = E(X^2) - E\left(\left(\frac{\sum X_i}{n}\right)^2\right)$$

On développe le deuxième terme. On tombe sur n termes de la forme $E(X_i^2)$ (qui valent donc $E(X^2)$) et $n(n-1)$ termes de la forme $E(X_iX_j)$ (qui, par indépendance, valent $E(X_i)E(X_j) = E(X)^2$).

D'où

$$\begin{aligned} E(\hat{\theta}) &= E(X^2) - \frac{1}{n^2}\left(nE(X^2) + n(n-1)E(X)^2\right) \\ &= E(X^2) - \frac{1}{n}\left(E(X^2) + (n-1)E(X)^2\right) \\ &= \frac{n-1}{n}\left(E(X^2) - E(X)^2\right) \\ &= \frac{n-1}{n}Var(X) \end{aligned}$$

A part si $Var(X)$ est nul, $\hat{\theta}$ est donc biaisé. ■

On sous-évalue donc la variance avec cette formule. La bonne formule à utiliser est donc

$$\begin{aligned} \hat{\theta}_{corr} &= \frac{n}{n-1}\left(\frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2\right) \\ &= \frac{\sum (X_i - \hat{X})^2}{n-1} \end{aligned}$$

Mais est-ce que ce facteur $n/(n-1)$ est si important ? On voit bien que si n est très grand, il ne change pas grand chose. On dit que (la famille d'estimateurs) est *asymptotiquement non-biaisée*.

Supposons qu'on utilise quand même la formule naturelle pour estimer la variance. Comment mesurer son efficacité ?

Définition 1.7 (*Erreur quadratique moyenne*)

L'erreur quadratique moyenne d'un estimateur $\hat{\theta}$ d'une quantité θ est

$$MSE(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right)$$

Notons si on développe que

$$\begin{aligned} E\left((\hat{\theta} - \theta)^2\right) &= E(\hat{\theta}^2) + \theta^2 - 2\theta E(\hat{\theta}) \\ &= E(\hat{\theta}^2 - E(\hat{\theta})^2) + \theta^2 - 2\theta E(\hat{\theta}) + E(\hat{\theta})^2 \\ &= Var(\hat{\theta}) + (\theta - E(\hat{\theta}))^2 \\ &= \text{variance} + \text{biais}^2 \end{aligned}$$

Cette quantité est la bonne pour mesure la précision d'un estimateur. Notons qu'elle revient à la formule de la qualité pour un estimateur non biaisé. On a donc un compromis à faire.

1.3 Confiance des estimateurs

On va maintenant développer dans cette section les arguments probabilistes qui permettent de justifier que ces estimateurs apportent effectivement de l'information sur ce qu'on cherche

1.3.1 Markov

Retour sur l'inégalité de Markov, vu dans le cours de MPI.

Proposition 1.4 (*Markov*)

$P(Y \geq \lambda)) \leq \frac{E(Y)}{\lambda}$

et appliquons là à $Y = (\hat{\theta} - \theta)^2$ (en remplaçant λ par λ^2 au passage)

$$P\left|\hat{\theta} - \theta| > \lambda\right) \leq \frac{MSE(\hat{\theta})}{\lambda^2}$$

(Dans le cas où $\hat{\theta}$ est nonbiaisée, on retombe exactement sur Chebyshev)

Si l'erreur quadratique moyenne est petite, on a donc de grandes chances que $\hat{\theta}$ soit très proche de la valeur de θ qu'on souhaite estimer.

En particulier, regardons le cas de la moyenne $\hat{X}_n$ calculée sur un échantillon de taille n. On a alors

$$P\left(|\hat{X}_n - \mu| > \lambda\right) \leq \frac{Var(X)}{n\lambda^2}$$

En particulier $\hat{X}_n$ "converge" vers μ :

Définition 1.8

Un(e suite d')estimateur(s) $\hat{\theta}_n$ pour une donnée θ est convergent(e) si pour tout ϵ :

$$P\left(|\theta_n - \theta| > \epsilon\right) \rightarrow_{n\rightarrow\infty} 0$$

Théorème 1.5

La moyenne $\hat{X}_n$ d'un échantillon de taille n est convergente (sous réserve que la variance ne soit pas infinie)

♦ **Exemple**

On veut savoir quel est le pourcentage de personnes en France qui pensent que Lyon va remporter la Ligue 1 de football. Notons p ce pourcentage. On note X la variable aléatoire qui vaut 0 si la personne pense que non, et 1 sinon. X est donc une variable de Bernoulli de paramètre p.

Supposons qu'on interroge 2000 personnes. Alors

$$P\left(|\hat{X}_n - p| > 0.05\right) \leq \frac{Var(X)}{2000 \times .05 \times .05}$$

$Var(X) = p(1-p) \leq 1/4$. Donc :

$$P\left(|\hat{X}_n - p| > 0.05\right) \leq \frac{1}{2000 \times .05 \times .05 \times 4} = \frac{1}{20}$$

On a donc au moins $19/20 = 95\%$ de chance que le résultat du sondage sur 2000 personnes nous donne une approximation à 5% de ce paramètre p. Il est très important de noter que tout ceci *ne dépend pas* de la taille totale de la population.

1.3.2 Intervalles de confiance

L'observation de l'exemple précédent se généralise et donne lieu à la notion d'intervalle de confiance :

Définition 1.9

Un intervalle de confiance de niveau α pour un paramètre θ est un intervalle I tel que

$$P(\theta \in I) \geq \alpha$$

Attention : θ est une constante (inconnue), pas une variable. La probabilité porte sur I, pas sur θ. Si par exemple on a un intervalle I de niveau 95%, cela signifie intuitivement que si on renouvelle 100 fois l'expérience, l'intervalle sera bon dans 95 des cas.

Comment calculer cet intervalle ? Plusieurs méthodes. La première est d'utiliser des astuces propres au paramètre qu'on calcule. C'est d'ailleurs ce qu'on a fait : On a utilisé le fait qu'on simulait une loi de Bernoulli pour dire que la variance était inférieure à $1/4$ et utiliser Chebyshev.

♦ **Exemple**

Supposons par exemple qu'on a sondé 2000 personnes et qu'on obtient comme résultat une moyenne de .3. On veut calculer l'intervalle de confiance à précision 99%.

Rappelons la formule :

$$P\left(|\hat{X}_n - p| > \lambda\right) \leq \frac{Var(X)}{n\lambda^2}$$

On a dit que $Var(X) \leq 1/4$, donc on résout $\frac{1}{4\lambda^2 \times 2000} = .01$ ce qui donne $\lambda = \sqrt{\frac{1}{4 \times 2000 \times .01}} = 0.12$. On en déduit que $p = 0.3 \pm 0.12(99\%)$

Dans l'exemple ci-dessus, on peut même faire mieux, en utilisant Chernoff.

Théorème 1.6 (*Chernoff*)

$$P(|\hat{X}_n - p| \geq \lambda) \leq 2e^{-2n\lambda^2}$$

♦ **Exemple**

On continue avec le même exemple, mais on va améliorer l'intervalle de confiance. On cherche donc λ tel que $2e^{-2n\lambda^2} \leq .01$. Comme c'est l'exponentiel d'un nombre négatif, le pire cas, c'est quand p est le plus grand possible, c'est à dire quand p vaut 1. Donc on résoud $2e^{-2n\lambda^2} = .01$ (on rappelle que $n = 2000$, et on obtient $\lambda = \sqrt{\frac{\log .005}{-2 \times 2000}} = 0.036$. On en déduit donc que $p = .3 \pm 0.036(99\%)$.

Que faire dans le cas général, où on ne sait rien sur la loi de X ? On utilise le théorème centrale limite :

Théorème 1.7 (*Théorème central limite*)

Alors $\sqrt{n}\frac{\hat{X_n}-\mu}{\sqrt{Var(X)}} \to Z$ où Z est une variable aléatoire de loi normale centrée réduite.

En particulier quand $n \geq 30$, on peut considérer en première approximation qu'il y a égalité.

Comment utiliser ce théorème ? Supposons qu'on cherche un intervalle de confiance avec probabilité 99%.

On cherche donc λ

$$P(|\hat{X_n} - \mu| \leq \lambda) = .99$$

En utilisant le théorème, on peut transformer ça en :

$$P(|Z\sqrt{Var(X)}/\sqrt{n}| \leq \lambda) = .99$$

ou encore

$$P\left(|Z| \leq \lambda\sqrt{n}/\sqrt{Var(X)}\right) = .99$$

On regarde maintenant le dessin de la courbe de la Gaussienne et on s'aperçoit que 99% de la courbe est concentrée sur des valeurs de $|Z| \leq 2.58$. On en déduit donc qu'un intervalle de confiance valide est donné par

$$\hat{X_n} \pm 2.58\sqrt{Var(X)}/\sqrt{n}(99\%)$$

Notez qu'on ne connait pas en général $Var(X)$! En pratique, quand n est suffisament grand, on peut remplacer $Var(X)$ par l'estimation de $Var(X)$ qu'on a calculé précédemment (pour que le calcul soit correct, il faudrait plutôt utiliser la *loi de Student*).

♦ **Exemple**

Dans notre cas, on peut dire que $Var(X) \leq 1/4$, donc $\sqrt{Var(X)} \leq 1/2$. On obtient donc un intervalle de confiance de $.3 \pm 2.58/\sqrt{2000} = .3 \pm 0.06(99\%)$. Le résultat est moins bon que la borne de Chernoff, mais valable pour n'importe quelle distribution de départ.

Pour finir, voici une table indicative donnant les constantes qu'il faut prendre pour mener l'étude pour d'autres intervalles de confiance que 99%.

Le tableau donne λ et α tel que $P(|Z| \leq \lambda) = \alpha$ (les valeurs de λ sont toutes arrondies par excès).

α	λ
80%	1.29
90%	1.65
95%	1.96
98%	2.33
99%	2.58

Exercices

(1 - 1) *(Loi uniforme)*

Un professeur un peu TOCé note chaque jour le numéro du bus qu'il prend pour aller à la fac. Son observation pendant deux mois, sur une quarantaine de jours ouvrable donne le tableau suivant :

61, 31, 46, 62, 61, 103, 19, 92, 77, 39, 126, 27, 103, 2, 101, 65, 40, 130, 107, 28, 42, 108, 87, 63, 121, 21, 122, 2, 66, 2, 56, 119, 109, 36, 126, 15, 102, 31, 14, 118

La moyenne est de 67, la variance de 1635.

On cherche à donner une estimation du nombre de bus qui desservent la ligne du professeur.

Pour cela, on suppose dans toute la suite du problème que la distribution est uniforme sur l'intervalle $[0, N]$, et on cherche à estimer N La distribution X vérifie donc $P[X = i] = \frac{1}{N+1}$ si $0 \leq i \leq N$

Q 1) Donner une estimation de la moyenne de X, et donner un intervalle de confiance à 95% (le coefficient à utiliser est 1.96).

Q 2) Calculer $E(X)$ et $Var(X)$ en fonction de N. En déduire une estimation de N.

Il existe une deuxième façon d'estimer N. On note $\hat{Y}$ le plus grand nombre observé sur un échantillon de taille n (ici $n = 40$).

Q 3) Calculer $P[\hat{Y} \geq k]$. En déduire $E(\hat{Y})$. On utilisera la formule suivante, valable quand m est grand. $1^p + 2^p + 3^p + \cdots + m^p \simeq \frac{(m+1)^{p+1}}{p+1}$

Q 4) En déduire comment estimer N en fonction de $E(\hat{Y})$. Application sur l'exemple.

On cherche maintenant à calculer la variance de $\hat{Y}$. Pour cela on va estimer $E(\hat{Y}^2)$ et utiliser $Var(\hat{Y}) = E(\hat{Y}^2) - E(\hat{Y})^2$. Pour vous aider, le professeur vous donne la formule :

$$\sum_k P[\hat{Y}^2 \geq k] = \sum_i (2i+1) P[\hat{Y}^2 \geq i^2]$$

Q 5) Chercher d'où vient la formule

Q 6) Calculer $E(\hat{Y}^2)$. En déduire $Var(\hat{Y})$.

Q 7) Montrer que l'estimateur que vous avez obtenu est meilleur que l'estimateur précédent.

Note : les données ont été tirées aléatoirement entre 0 et 131.

Note : Cet exercice est connu sous le nom de "problème des chars allemands". Voir Ruggles et Brodie : *An Empirical Approach to Economic Intelligence in World War II*.

(1 - 2) *(Moyenne vs Moyenne)*

On se donne un échantillon (X_1, X_2) de taille 2 d'une population de distribution X. On décide d'estimer l'espérance de X avec la formule $\hat{X} = aX_1 + bX_2$.

Q 1) Quelles sont les conditions sur a et b pour que l'estimateur soit non biaisé ?

Q 2) Quel est le meilleur choix de a et b ?

(Note : le résultat se généralise aux échantillons de taille n)

Q 3) La médiane sur un échantillon de taille 3 est-elle un estimateur biaisé ?

(1 - 3) *(Prédiction)*

Un test a été mené sur 100 disques dur de même marque et produits dans la même usine. Le premier a crashé au bout de 5 jours, et le dernier au bout de 5 ans.

On se donne un 101ème disque dur. Montrer qu'avec probabilité au moins 98% il crashera entre 5 jours et 5 ans. Pour cela :

- Modéliser la situation. On ne se donnera *aucune* propriété particulière sur la distribution de la durée de vie d'un disque dur.
- Prouver le résultat.

(1 - 4) *(Marque et recapture)*

Pour estimer le nombre lions dans la savane, les biologistes procèdent ainsi : Ils choisissent un échantillon de 100 lions distincts et leur marquent le nez de peinture rouge. Une semaine plus tard, ils choisissent aléatoirement 100 lions (pas forcément distincts), et comptent combien ont le nez rouge. On suppose qu'il y en a n.

Q 1) Modéliser le problème et en déduire une estimation du nombre de lions.

Chapitre 2

ESTIMATIONS DE PARAMÈTRES

Dans ce chapitre, on cherche à utiliser un échantillon pour donner une estimation de certains paramètres de la population, plus généraux que la moyenne et la variance calculée précédemment.

Pour guider le cours, on s'intéressera en particulier à une variable X de loi uniforme sur l'intervalle $[a, b]$, et à comment estimer a et b.

2.1 Méthode des moments

Définition 2.1

Le moment d'ordre k de la population est $\mu_k = E(X^k)$. Le moment d'ordre k de l'échantillon est $\hat{\mu}_k = \frac{\sum X_i^k}{n}$

Théorème 2.1

$\hat{\mu}_k$ est un estimateur non biaisé pour μ_k. De plus $Var(\hat{\mu}_k) = \frac{Var(\mu_k)}{n}$.

La méthode des moments consiste à trouver les paramètres $\theta_1 \dots \theta_n$ en les exprimant en fonction des quantités μ_k puis en utilisant $\hat{\mu}_k$ pour les estimer.

♦ **Exemple**

Dans l'exemple, on calcule

$$\mu_1 = \frac{a+b}{2}$$

$$\mu_2 = \frac{a^2+ab+b^2}{3}$$

On en déduit donc

$$3\mu_2 = a^2 + a(2\mu_1 - a) + (2\mu_1 - a)^2 = a^2 - 2\mu_1 a + 4\mu_1^2 = (a-\mu_1)^2 + 3\mu_1^2$$

D'où

$$a = \mu_1 - \sqrt{3\mu_2 - 3\mu_1^2} = E(X) - \sqrt{3Var(X)}$$

La méthode des moments nous dit donc d'estimer a en faisant

$$\hat{a} = \hat{\mu}_1 - \sqrt{3\hat{\mu}_2 - 3\hat{\mu}_1^2}.$$

C'est un estimateur *biaisé*. Il est en effet difficile de trouver un estimateur non biaisé pour $\sqrt{Var(X)}$ (on a un estimateur non biaisé pour $Var(X)$, mais pas pour sa racine carrée !) Dans l'exemple cependant, l'estimateur $\hat{a}$ est consistent.

♦ **Exemple**

On cherche à estimer le paramètre λ d'une loi exponentielle de paramètre λ.
On sait d'après le cours que $\mu_1 = \frac{1}{\lambda}$.
Donc on obtient $\lambda = \frac{1}{\hat{\mu}_1}$ comme estimateur. Il est également biaisé, mais convergent.

2.2 Maximum de vraisemblance

L'idée du maximum de vraisemblance est de chercher la valeur du paramètre qui maximise la probabilité d'avoir observé ce qu'on a vraiment observé.

♦ **Exemple**

Reprenons l'exemple, mais cette fois en discret : on cherche la distribution X entre $[a, b]$, mais cette fois discrète : $P[X = i] = \frac{1}{b-a+1}$.
Supposons qu'on ait vu les entiers $35, 17$ et 128.
La probabilité d'avoir vu ces 3 entiers est :

- $\left(\frac{1}{b-a+1}\right)^3$ si $a \leq 17$ et $b \geq 128$
- 0 sinon

On cherche maintenant a et b qui maximise la probabilité. Il est clair qu'il faut prendre b le plus petit possible et a le plus grand possible, donc le maximum de vraisemblance est $a = 17$ et $b = 128$. Plus généralement, pour cette distribution, le maximum de vraisemblance pour a est de prendre le minimum des valeurs observées.

♦ **Exemple**

Essayons sur une variable continue. On cherche le paramètre λ d'un loi exponentielle, qui vérifie donc $P(X \leq t) = 1 - e^{-\lambda t}$
Supposons qu'on voie $35, 17$ et 128.
La probabilité qu'on ait vu ces trois nombres est nulle ! En effet pour une loi continue, $P(X = t) = 0$.
Pour calculer le maximum de vraisemblance, il faut donc regarder la densité de X en chacun de ces points.
La densité de X en t est la dérivée de $F(X \leq t)$ et vaut donc $\lambda e^{-\lambda t}$
La densité *jointe* sur les trois observations est donc $\lambda^3 e^{-17\lambda} e^{-35\lambda t_2} e^{-128\lambda} = \lambda^3 e^{-180\lambda}$
On cherche donc λ qui maximise cete quantité. Pour faire ça il faut voir ça comme une fonction de λ et dériver. Mais en général (et c'est le cas ici), au lieu de dériver $f(\lambda)$, on va regarder son logarithme, et dériver le log

$$\ln f(\lambda) = 3 \ln \lambda - 180\lambda$$

Sa dérivée est

$$\frac{3}{\lambda} - 180$$

Le maximum de vraisemblance est donc obtenu en $\lambda = \frac{3}{180}$.
Plus généralement on trouve, pour la loi exponentielle, un maximum de vraisemblance en $\frac{1}{E(X)}$.
Tout comme pour la méthode des moments, l'estimateur est biaisé.

2.3 Intervalles de confiance pour la méthode des moments

Théorème 2.2 (*Méthode delta*)

Soit f une fonction raisonnable.
Soit $\hat{X}_n = \frac{X_1+...X_n}{n}$ et $\theta = E(X) = E(\hat{X}_n)$.
Alors $\sqrt{n}\frac{f(\hat{X}_n)-f(\theta)}{\sqrt{VarXf'(\theta)^2}}$ converge vers une v.a. de loi normale centrée réduite. (si la dérivée est non nulle).

Autrement dit, on obtient un intervalle de confiance correct de niveau 99% pour $f(\hat{X}_n)$ en prenant $f(\hat{X}_n) \pm 2.58\sqrt{\frac{Var(X)f'(E(X))^2}{n}}$

♦ **Exemple**

Pour le paramètre de la loi exponentielle, on avait donc $f(x) = 1/x$. On en déduit donc un intervalle de confiance à 99% en prenant

$$\frac{1}{\hat{X}} \pm 2.58\sqrt{\frac{\hat{V}}{n\hat{X}^4}}$$

où $\hat{V}$ désigne l'estimateur de la variance.

Comment faire si on a une fonction de deux variables ? La méthode peut se généraliser, mais dépasse les compétences d'étudiants en L3.

2.4 Méthodes plus élaborées

2.4.1 Espacement maximal

La technique est basée sur le résultat intuitif suivant :

Proposition 2.3

Soit X_1, X_2, X_3 un échantillon de taille 3 d'une population (dont la distribution est continue).
Alors

$$P(X_3 \leq X_1 \wedge X_3 \leq X_2) = \frac{1}{3}$$
$$P(X_1 \leq X_3 \leq X_2) = \frac{1}{3}$$
$$P(X_3 \geq X_1 \wedge X_3 \geq X_2) = \frac{1}{3}$$

Ce qui se généralise évidemment.
On en déduit en particulier le résultat suivant :

Théorème 2.4

$$P[\min_{i\leq n} X_i \leq X_{n+1} \leq \max_{i\leq n} X_i] = \frac{2}{n+1}$$

En particulier si on a un échantillon de taille 101, on peut prédire avec confiance 98% (plus exactement $99/101$) que le 101-ème élément a une valeur comprise entre les valeurs observées précédemment. On peut donc utiliser ce résultat pour prédire les futures valeurs.

On note $X^(n)$ le n-ème plus petit élément parmi $X_1 \dots X_n$. On prendra comme convention $X^{(0)} = -\infty$ et $X^{n+1} = +\infty$

Le résultat nous dit donc :

Proposition 2.5

Soit Y de même loi que tous les X_i (et donc que X). Alors

$$P[X^{(0)} \leq Y \leq X^{(1)}] = P[X^{(1)} \leq Y \leq X^{(2)}] = \cdots = P[X^{(n)} \leq Y \leq X^{(n+1)}]$$

On cherchera donc à trouver le paramètre qui s'arrange pour que tous ces nombres soient les plus proches les uns des autres.

Voici la définition formelle du paramètre.

Définition 2.2

On note $P_i = P[X^{(i)} \leq X \leq X^{(i+1)}]$. L'estimateur d'espacement maximal est celui qui maximise $\sum_i \ln P_i$.

L'idée est que pour maximiser cette quantité, sachant que $\sum_i P_i = 1$, c'est qu'ils soient tous égaux.

♦ **Exemple**

On prend l'exemple d'une distribution uniforme sur $[0, b]$.

$$P_0 = \frac{X^{(1)}}{b}$$

$$P_i = \frac{X^{(i+1)} - X^{(i)}}{b}$$

$$P_n = \frac{b - X^{(n)}}{b}$$

Donc

$$\sum_i \ln P_i = \ln X^{(1)} + \sum_i \ln \left(X^{(i+1)} - X^{(i)} \right) + \ln(b - X^{(n)}) - (n+1) \ln b$$

On cherche le b qui maximise. En fait il n'y a que deux termes qui dépendent de b, les deux derniers. La dérivée vaut

$$\frac{1}{b - X^{(n)}} - \frac{n+1}{b}$$

Qui s'annule (c'est clairement un maximum) pour

$$b = \frac{n+1}{n} X^{(n)}$$

On retrouve un résultat vu en TD.

Exercices

(2 - 1) *(Intervalles de confiance)*

On lance six fois un dé non truqué.

Q 1) Montrer que l'évènement le moins fréquent a plus d'une chance sur 100 de se produire.

Q 2) Que peut-on en conclure pour l'intervalle de confiance à 99% ?

Q 3) En supposant que la pièce est tombée six fois consécutivement sur face, déterminer les intervalles de confiance en suivant les 3 méthodes vues en cours (Markov/Chebyshev, Chernoff, et la méthode générale). Que peut-on remarquer ?

(2 - 2) *(Pièces et dés)*

On lance une pièce truquée, qu'on assimile à une variable de Bernouilli de paramètre p, 100 fois. On observe 60 fois pile et 40 fois face.

Q 1) Estimer p en utilisant la méthode des moments. Donner un intervalle de confiance à 95% avec la méthode de votre choix.

Q 2) Estimer p avec la méthode du maximum de vraisemblance.

On lance un dé à 3 faces. On note p_1, p_2, p_3 les probabilités que le dé tombe sur 1, 2 et 3. Notons que $p_1 + p_2 = p_3$ (il n'y a donc que deux variables à estimer). En lançant 100 fois le dé, on obtient le résultat suivant :

1	20
2	30
3	50

Q 3) Estimer p_1, p_2, p_3 en utilisant la méthode des moments.

Q 4) Estimer p_1, p_2, p_3 en utilisant la méthode du maximum de vraisemblance.

On considère maintenant une distribution "tente" sur l'intervalle $[0, 1]$ dont la densité est

$$f(x) = \begin{cases} \frac{2x}{c} & x < c \\ \frac{2(1-x)}{1-c} & x \geq c \end{cases}$$

Q 5) Représenter $f(x)$

Q 6) Donner une expression de $P[X \leq t] = \int_0^t f(x)dx$. Vérifier que $P[X \leq 1] = 1$. Essayer de la représenter quand $c = 1/3$.

On observe les données 0.2 et 0.4.

Q 7) Estimer c en utilisant la méthode des moments. Qu'observe-t-on ?

Q 8) Estimer c en utilisant le maximum de vraisemblance.

(2 - 3) *(Marque et recapture)*

Pour estimer le nombre lions dans la savane, les biologistes procèdent ainsi : Ils choisissent un échantillon de 100 lions distincts et leur marquent le nez de peinture rouge. Une semaine plus tard, ils choisissent aléatoirement 100 lions (pas forcément distincts), et comptent combien ont le nez rouge. On suppose qu'il y en a n.

Q 1) Modéliser le problème et en déduire une estimation du nombre de lions.

(2 - 4) *(Détection passive)*

Quand un paquet va d'une machine à l'autre sur l'Internet, il passe par plusieurs routeurs. Chaque paquet est initialisé avec une valeur, appelée TTL, qui vaut 128 pour les machines Windows et 64 pour une machine Linux. A chaque fois que le paquet franchit un routeur, le champ TTL est décrémenté, et le paquet disparaît si le TTL vaut 0. Ce mécanisme est utilisé pour éviter qu'un paquet mal routé ne tourne en rond infiniment sur l'Internet.

On estime que le nombre moyen de routeurs traversés par un paquet est modélisé correctement par une variable Gaussienne de moyenne 15 et de variance 16. On note X cette v.a.

Un serveur sur Internet note, à chaque fois qu'elle reçoit un paquet, le TTL du paquet en question.

Q 1) Expliquer comment utiliser cette information pour obtenir une estimation du pourcentage de machines sous Linux. On utilisera la méthode des moments.

(2 - 5) *(Données censurées)* On suppose que la durée de vie d'un disque dur suit une loi exponentielle de paramètre $\lambda : P[X \geq t] = e^{-\lambda t}$. La durée de vie moyenne est donc $\frac{1}{\lambda}$. On suppose que t est exprimé en jours.

Q 1) On se donne deux disques durs. Le premier est mort au bout de 5 jours, le deuxième au bout de 10 jours. Estimer λ en utilisant le maximum de vraisemblance. (On pourra utiliser le changement de variable $x = e^{-\lambda}$ pour se simplifier la vie)

Q 2) En fait, la mesure n'est pas très précise, on sait juste que le premier est mort entre le début du 5e jour et la fin du 5e jour, de même pour le deuxième. A votre avis, quelle doit être la valeur de λ ? Estimer λ en utilisant le maximum de vraisemblance. (Même remarque). Vérifiez que le résultat est conforme à votre prévision.

Q 3) On se donne un échantillon de 101 disques durs. Sur une période de 365 jours, un seul est mort, et ce au bout de 200 jours. Estimer λ en utilisant le maximum de vraisemblance.

Chapitre 3

TESTS STATISTIQUES

Un test cherche si une hypothèse, appelée hypothèse nulle et souvent notée H_0, est crédible ou non, en la comparant à une hypothèse H_1. Un test de valeur α est tel que, si H_0 est vrai, alors la probabilité de rejeter H_0 si H_0 est vrai est au plus α.

Si α est suffisament petit, l'idée est de se dire que si jamais le test rejette H_0, c'est parce que H_0 est vrai : il est plus raisonnable de penser qu'on se trompe dans H_0 plutôt que de penser qu'on vient d'assister à un évenement de probabilité α petit.

Si jamais le test réussit, on peut juste conclure qu'il n'a pas rejeté H_0, pas que H_0 est vraie.

Il y a donc deux types d'erreur dans un test : rejeter l'hypothèse nulle alors qu'elle est satisfaite (faux positif), ou ne pas rejeter l'hypothèse alors qu'elle est fausse (faux négatif).

3.1 Introduction aux tests

3.1.1 Exemple

Commençons par tester si une pièce est non biaisée. L'hypothèse nulle est donc $H_0 : p = 1/2$ (vs $H_1 : p \neq 1/2$).

Supposons qu'on fasse un test avec un échantillon de taille 100. On cherche un test qui réussit au niveau 5%. Cela veut dire que si l'échantillon échoue au test, c'est soit que l'hypothèse est fausse, soit que l'échantillon fait partie des 5% d'échantillons qui ne passe pas le test.

Un test typique qu'on peut faire est de regarder la moyenne $\hat{X}_n$ calculée sur l'échantillon. Comme $n = 100$ est très grand, on sait que avec probabilité 95%,

$$\hat{X}_n \in 0.5 \pm 1.96\sqrt{Var(X)}/\sqrt{n}$$

Ici, l'hypothèse H_0 nous permet de connaître entièrement X, et donc en particulier sa variance vaut $1/4$. Donc en remplaçant n par 100, on s'aperçoit que

$$\hat{X}_{100} \in 0.5 \pm 0.1$$

Donc si on observe un nombre de piles qui n'est pas entre 40 et 60, on considèrera que l'hypothèse est fausse au niveau 5%.

Il est à noter que c'est exactement la même démarche que pour les intervalles de confiances, sauf qu'on n'estime pas $E(X)$ à partir de $\hat{X}_n$, mais le contraire.

3.1.2 Valeur p (p-values)

A retenir *Valeur p*

La valeur p représente la probabilité de rejeter à tort l'hypothèse nulle (faux positif).

Dans l'exemple ci-dessus, supposons avoir obtenu 35 piles, on a donc

$$|\hat{X}_{100} - 0.5| = 0.15$$

Cela correspond à un écart, au niveau de la loi normale de

$$0.15 \times \sqrt{n}/\sqrt{Var(X)} = 3$$

Si on regarde les tables de la loi normale, la probabilité de s'éloigner à plus de 3 de l'origine est 0.004. Si on obtient 35 piles, la valeur p du test est donc 4‰.

3.1.3 Puissance statistique

A retenir *Puissance d'un test*

La puissance d'un test est la probabilité de rejeter l'hypothèse nulle sachant qu'elle est fausse.

Toujours sur le même exemple, on va calculer la puissance d'un test à $p = 0.65$. C'est la probabilité de rejeter l'hypothèse nulle alors qu'en réalité la pièce est biaisée avec $p = 0.65$.

La probabilité de rejeter est la probabilité d'observer moins de 40 piles, ou plus que 60 piles. La probabilité moins de 40 piles est très faible, donc on va la considérer nulle dans la suite. Pour calculer la probabilité d'observer plus que 60 piles, on va utiliser le fait que $n = 100$ est très grand.

Si on note N le nombre de piles observés, il suit à peu près une loi normale de moyenne 65 et de variance $100 * .65 * (1 - .65) = 22.75$. On cherche donc la probabilité que $N > 60$, soit la probabilité que $(N - 65)/\sqrt{22.75} > -5/\sqrt{22.75} \sim 1.05$.

On regarde donc sur une table la probabilité qu'une variable Z de loi normale centrée réduite soit supérieure à -1.05, et cette probabilité vaut 0.85.

La puissance du test à $p = 0.65$ est donc 0.85.

3.2 Test du χ^2

Le test du χ^2 est un test statistique très connue et très utilisé, qu'on utilise dans deux contextes un peu différents.

Définition 3.1

La somme du carré de n variables aléatoires normales centrées réduites indépendentes est une variable aléatoire de loi du χ^2 avec n degrés de libertés.

Le principe est le suivant. On considère une expérience qui peut avoir k résultats possibles, de probabilité $p_1 \dots p_k$ respectivement.

Supposons qu'on respecte l'expérience n fois. On s'attend à observer np_1 fois le premier résultat, np_2 fois le deuxième, etc.

On va regarder à quel point le résultat est différent de la théorie en regardant

$$\sum_i \frac{(N_i - np_i)^2}{np_i}$$

où N_i représente le nombre de fois où le résultat i est sorti.

Avant d'en dire plus, regardons le cas très simple d'une pièce truquée qui tombe avec probabilité $p = 1/3$ sur pile Il y a deux résultats possibles pour l'expérience : pile et face. Si on note N_p le nombre de piles, on regarde donc

$$\begin{aligned} \frac{(N_p-n/3)^2}{n/3} + \frac{(N_f-2n/3)^2}{2n/3} &= \frac{(N_p-n/3)^2}{n/3} + \frac{(N_p-n/3)^2}{2n/3} \\ &= \frac{(N_p-n/3)^2}{2n/9} \\ &= (\frac{N_p}{n} - 1/3)^2 * \frac{n}{2/9} \\ &= (\frac{N_p}{n} - 1/3)^2 * \frac{n}{2/9} \end{aligned}$$

Par le théorème centrale limite ($2/9 = (2/3) \times (1/3)$ est la variance), c'est quand n est très grand, le carré d'une v.a. de loi normale centrée réduite, donc une v.a. de loi du χ^2 avec 1 degré de liberté.

Théorème 3.1

$$\sum_i \frac{(N_i - np_i)^2}{np_i}$$

suit une loi du χ^2 avec $k-1$ degrés de liberté quand n est suffisament grand, et quand, pour chaque résultat possible $np_i \geq 5$.

Le test du χ^2 fonctionne alors ainsi : On regarde dans une table pour quelle valeur x la probabilité qu'une variable de loi χ^2 avec $j-1$ degrés de liberté soit supérieure à x faut 5%. On calcule alors la

quantité du théorème, et on rejette l'hypothèse si la valeur est supérieure à x.

A retenir

Pour appliquer la méthode du χ^2, calculer la somme sur toutes les classes i de $\frac{(O_i - E_i)^2}{E_i}$ où O_i est l'effectif observé et E_i l'effectif théorique.

Le nombre de degrés de liberté est en général le nombre de classes moins un (mais voir plus loin !)

Le test réussit alors avec valeur α si, dans la table du χ^2 , la valeur indiquée est plus faible que ce qu'on vient de trouver.

Voici une table du χ^2 :

d	0.05	0.01	0.005
1	3.84	6.64	7.88
2	6	9.21	10.6
3	7.82	11.35	12.84
4	9.5	13.28	14.86
5	11.07	15.09	16.75
6	12.6	16.81	18.55
7	14.07	18.48	20.28

3.2.1 Test d'ajustement

Donner trois exemples d'utilisation du test.

♦ **Exemple**

Revenons d'abord à la pièce non truquée et supposons qu'on observe 35 piles et 65 faces.

Ecrivons le *tableau de contigence* :

n	0	1
observé	65	35
théorique	50	50

On calcule donc :

$$\frac{(65-50)^2}{50} + \frac{(35-50)^2}{50} = 9$$

9 est plus grand que 3.84, donc on rejette l'hypothèse nulle avec valeur 5% (même 5‰), et on en conclut que la pièce est biaisée.

♦ **Exemple**

On lance un dé 120 fois et on obtient les résultats suivants :

n	1	2	3	4	5	6
observé	25	21	25	18	15	16
théorique	20	20	20	20	20	20

On calcule et on obtient 5.55. Si on regarde dans la table, on s'aperçoit que le test n'est pas rejeté à 5%.

♦ **Exemple**

L'hypothèse nulle est que le nombre de buts dans un match de football en Ligue 1 suit une loi de Poisson de paramètre 2.5. On obtient le tableau de contigence suivant :

buts	0	1	2	3	4	5	6	7	8	9
marqués	27	71	114	78	55	21	6	3	4	1
théorie	31.19	77.98	97.48	81.23	50.77	25.38	10.58	3.78	1.18	0.33

Avant de faire quoi que ce soit, il faut réunir des classes, puisque toutes les classes doivent avoir au moins un effectif théorique de 5. Il suffit dans ce cas de regrouper les 3 dernières classes :

buts	0	1	2	3	4	5	6	7+
marqués	27	71	114	78	55	21	6	8
théorie	31.19	77.98	97.48	81.23	50.77	25.38	10.58	5.29

On calcule alors et on obtient 8.59. On compare avec la table du χ^2 avec 7 degrés de liberté (puisqu'on a 8 possibilités), et on voit qu'on ne rejette pas l'hypothèse.

Le test d'ajustement peut également s'utiliser lorsqu'on ne connaît pas certains paramètres. Dans ce cas on diminue les degrés de liberté de 1 pour chaque paramètre qu'on doit estimer.

Par exemple, on suppose avoir les données suivantes :

n	1	4	5	6	7	8	9	10
T[n]	1	2	22	175	479	455	58	12

On veut savoir si les données proviennent d'une distribution Gaussienne. L'hypothèse nulle est donc qu'il s'agit d'une distribution Gaussienne. Une telle distribution est spécifiée par deux paramètres (moyenne et écart type). On ne les connaît pas, donc on va les estimer à partir de l'échantillon. Cela diminue de deux le nombre de degrés de liberté et on arrive à 8-2-1 = 5 degrés. On effectuerait donc le test du χ^2 avec 5 degrés de liberté.

A retenir

S'il est nécessaire dans le test d'estimer k paramètres, le nombre de degrés de liberté diminue de k.

Le nombre de degrés de liberté est donc nb classes - paramètres - 1.

3.2.2 Test d'indépendance

Le test du χ^2 permet aussi de tester l'indépendance de deux quantités.

♦ **Exemple**

Regardons l'exemple suivant, purement fictif. On veut savoir si le fait de venir d'IUT est un gage de réussite dans l'enseignement supérieur.

On observe le résultat suivant :

	IUT	pas IUT
a son L3	12	25
n'a pas son L3	6	17

Supposons que les deux données soient indépendantes. L'effectif total est de 60.

Notons p_A la probabilité qu'un étudiant ait son L3, et p_B la probabilité qu'il vienne d'IUT.

Le tableau théorique, s'il y a indépendance, est donc :

	IUT	pas IUT
a son L3	60 $p_A p_B$	60 $p_A(1-p_B)$
n'a pas son L3	60 $(1-p_A)p_B$	60 $(1-p_A)(1-p_B)$

On ne connaît pas p_A et p_B mais on peut les estimer en fonction des données. On estime donc $p_A = 37/60$. et $p_B = 18/60$. Cela nous donne le tableau théorique suivant :

	IUT	pas IUT
a son L3	11.1	25.9
n'a pas son L3	6.9	16.1

Le test du ξ^2 nous donne un résultat de 0.27. Remarquons qu'il faut faire un test avec un seul degré de liberté, puisqu'on part d'un découpage en 4, mais qu'on a du estimer 2 paramètres. Si on regarde la table, on ne rejette pas l'hypothèse d'indépendance.

♦ **Exemple**

On veut savoir si votre équipe préférée joue mieux en 4-4-2 qu'en 4-3-3.

	4-4-2	4-3-3
Victoire	40	21
Match Nul	10	12
Défaite	7	10

Pour faciliter les choses, on va calculer les sommes par ligne et par colonne :

	4-4-2	4-3-3	Total
Victoire	40	21	61
Match Nul	10	12	22
Défaite	7	10	17
Total	57	43	100

On obtient alors le tableau théorique suivant :

	4-4-2	4-3-3
Victoire	34.77	26.23
Match Nul	12.54	9.46
Défaite	9.69	7.31

Pour information le premier chiffre vaut $\frac{61*57}{100}$.

A retenir

Pour obtenir le coefficient théorique en (i, j), on calcule la somme de la ligne i multiplié par la somme de la colonne j divisé par la somme totale.

Le test du ξ^2 donne 4.76. Combien a-t-on de degré de libertés ? On a fixe 3 paramètres (la proba de 4-4-2, la proba de Victoire, la proba de Match Nul, les autres s'en déduisant), donc on obtient

6 - 3 - 1 = 2 degrés de liberté.

A retenir

En règle générale, le nombre de degrés de liberté dans un test du χ^2 d'indépendance est $(i-1)(j-1)$ s'il y a i possibilités pour les lignes, et j pour les colonnes.

On peut maintenant regarder dans la table, et 4.76 est plus petit que 6, donc on ne rejette pas l'hypothèse d'indépendance.

3.3 Kolmogorov-Smirnov

La méthode KS est une méthode qui a l'avantage de marcher sans aucune hypothèse. L'inconvénient est qu'elle nécessite de connaître les valeurs d'une certaine loi, qui est très délicate à obtenir en pratique (contrairement à χ^2 ou la loi Normale).

La méthode est principalement utile pour des variables continues. Soit par exemple à tester l'hypothèse "Les données suivent une loi exponentielle de paramètre 2" et supposons que les données soient : 1.6, 2.92, 0.02, 0.12.

L'idée est de construire deux courbes :

- La courbe théorique $P[X \leq t]$
- La courbe pratique : On approche la loi par une loi discrète qui vaut 0 entre 0 et 0.02, 1/4 entre 0.02 et 0.12, 2/4 entre 0.12 et 1.6, etc.

On regarde alors la différence maximum entre ces deux courbes.

On peut prouver que la variable aléatoire D de cette différence, lorsque l'hypothèse est la bonne, a la même distribution quelle que soit la distribution de X.

Elle est cependant très difficile à calculer.

Voilà la table de la valeur critique de D (Si D dépasse cette valeur, l'hypothèse est rejetée) suivant n :

n	0.05	0.01
1	.975	.995
2	.842	.929
3	.708	.828
4	.624	.733
5	.565	.669
6	.521	.618
7	.486	.577
$\sim \infty$	$1.36/\sqrt{n}$	$1.63/\sqrt{n}$

♦ **Exemple**

On va utiliser le test de KS pour vérifier que le générateur random de python donne les bonnes valeurs.

On le teste sur 4 valeurs et il nous donne : 0.4, .78, .41, .18.

Dessinons les deux courbes au tableau. On voit que le maximum est de $.75 - .41 = .34$. On regarde dans la table en face de $n = 4$, et on s'aperçoit que tout va bien, l'hypothèse n'est pas rejetée.

Exercices

(3 - 1) *(Partiel 2011)*

On effectue une étude sur le développement de 80 villes, pour lesquelles on a constaté que l'extension à partir du centre s'effectuait dans une direction géographique privilégiée. Les informations sont détaillées dans le tableau ci-dessous :

Direction	N	NE	E	SE	S	SO	O	NO
Nombre de villes	7	6	4	6	10	16	18	13

Ces données sont-elles compatibles avec l'hypothèse que l'extension d'une ville a les mêmes chances de se faire dans n'importe quelle direction ?

(3 - 2) *(Pièce biaisée)* On lance un million (1000000) de fois une pièce et elle tombe 501553 fois sur pile. Est-elle biaisée ?

Q 1) Formuler l'hypothèse nulle.

Q 2) Effectuez un test de risque 1% en utilisant un intervalle de confiance.

Q 3) Effectuez le test du χ^2 de risque 1%.

Q 4) Concluez

(3 - 3) *(Loi exponentielle)* On observe 50 disques dur et on obtient les durées de vie suivantes, exprimées en années :

0.06	0.21	0.41	0.49	0.55	0.55	0.92	0.95	0.98	1.04	1.09	1.20	
1.31	1.37	1.46	1.47	1.61	1.61	1.63	1.64	1.86	1.86	1.99	2.06	2.10
2.11	2.18	2.94	3.00	3.10	3.16	3.83	4.07	4.45	4.63	4.90	5.09	
5.94	6.09	6.26	6.34	6.64	7.40	7.72	7.75	8.22	9.19	9.55	10.58	15.90

On cherche à savoir si la durée de vie est bien modélisée par une variable exponentielle. (Aide : la somme de tous les nombres est 181.46)

Q 1) Formulez l'hypothèse nulle

Q 2) Effectuez le test du χ^2 de risque 1%

Q 3) Concluez

(3 - 4) *(Homogénéité)*

Le restaurant de l'université achetait sa nourriture chez le fournisseur A jusqu'au jour où il constata, sur 100 steaks, que 40 contenaient de la viande de cheval. Il décida donc de changer pour le fournisseur B, et constata que 8 steaks sur les 50 achetés contenaient de la viande de cheval.

Q 1) Au vu de ces valeurs, peut-on dire que les fournisseurs A et B n'achètent pas leur viande au même endroit ?

(3 - 5) *(random)* On lance 100 fois la fonction `random` de python, et on obtient les résultats suivants :

0.18	0.95	0.24	0.88	0.77	0.10	0.26	0.91	0.20	0.16
0.86	0.31	0.16	0.83	0.91	0.99	0.23	0.54	0.81	0.43
0.98	0.36	0.21	0.93	0.40	0.80	0.97	0.54	0.89	0.44
0.21	0.88	0.36	0.74	0.46	0.16	0.67	0.29	0.65	0.84
0.76	0.11	0.80	0.81	0.16	0.46	0.36	0.34	0.08	0.25
0.68	0.86	0.64	0.86	0.84	0.72	0.98	0.36	0.96	0.69
0.67	0.82	0.64	0.83	0.76	0.32	0.50	0.23	0.42	0.05
0.21	0.80	0.68	0.05	0.48	0.15	0.62	0.59	0.22	0.99
0.01	0.69	0.10	0.09	0.67	0.39	0.23	0.01	0.24	0.69
0.36	0.84	0.06	0.58	0.81	0.79	0.19	0.90	0.07	0.02

Si jamais on trie les résultats on obtient :

0.01	0.01	0.02	0.05	0.05	0.06	0.07	0.08	0.09	0.10
0.10	0.11	0.15	0.16	0.16	0.16	0.16	0.18	0.19	0.20
0.21	0.21	0.21	0.22	0.23	0.23	0.23	0.24	0.24	0.25
0.26	0.29	0.31	0.32	0.34	0.36	0.36	0.36	0.36	0.36
0.39	0.40	0.42	0.43	0.44	0.46	0.46	0.48	0.50	0.54
0.54	0.58	0.59	0.62	0.64	0.64	0.65	0.67	0.67	0.67
0.68	0.68	0.69	0.69	0.69	0.72	0.74	0.76	0.76	0.77
0.79	0.80	0.80	0.80	0.81	0.81	0.81	0.82	0.83	0.83
0.84	0.84	0.84	0.86	0.86	0.86	0.88	0.88	0.89	0.90
0.91	0.91	0.93	0.95	0.96	0.97	0.98	0.98	0.99	0.99

La moyenne est de .519 et la moyenne de la somme des carrés de .3631

Peut-on considérer qu'elle simule bien la loi uniforme ?

Q 1) Essayez tous les tests possibles, et conclure.

STATISTIQUES BAYESIENNES

La majorité du cours jusqu'à présent s'intéresse aux statistiques dites fréquentistes. S'y oppose une deuxième branche des statistiques, appelée statistiques bayesiennes. Leur nom vient de l'utilisation prépondérante de la loi de Bayes :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Dans ce formalisme, on considère les probabilités de façon très différente, comme une plausibilité, et utilise la notion d'estimation *a priori*.

Supposons que le but soit de trouver un paramètre λ. L'idée est de se fixer *a priori* une distribution L sur λ, l'idée étant que $P[L = t]$ traduit à quel point on pense que $\lambda = t$ (Cette distribution a priori est subjective !).

On note $X|L = t$ la distribution de la variable X si jamais le paramètre est t.

A chaque observation, on met à jour la distribution L en suivant la loi de Bayes. Formellement, si on observe le résultat k, on change L en L' ainsi : (A joue le rôle de $L = t$, et B de $X = k$) :

$$P\left[L' = t\right] = \frac{P\left[X = k|L = t\right] P\left[L = t\right]}{P\left[X = k\right]}$$

ou encore :

$$P\left[L' = t\right] = \frac{P\left[X = k|L = t\right] P\left[L = t\right]}{\sum_t P\left[X = k|L = t\right] P\left[L = t\right]}$$

PARTIE 1 - Un premier exemple

On dispose d'une pièce dont on sait qu'il s'agit soit d'une pièce non truquée, soit d'une pièce conçue pour tomber trois fois plus souvent sur pile (donc pile a une probabilité 3/4 d'arriver).

On lance 4 fois la pièce et on observe 3 piles et 1 face. On veut savoir quelle est la probabilité que la pièce soit truquée.

On voit la pièce comme une v.a. X de loi de Bernoulli de paramètre p, on sait que p vaut $1/2$ ou $3/4$. Au départ, on ne sait rien sur la pièce, donc $P[normal] = 1/2$ et $P[truque] = 1/2$.

On a de plus $P(face|normal) = 1/2$ et $P(face|truque) = 1/4$ et de même avec 1.

Q 1) Quelle est la probabilité d'observer un pile ?

Q 2) On observe un pile. Mettez à jour les probabilités pour $normal$ et $truque$:

$$P(normal') = \frac{P(pile|normal)P(normal)}{P(pile)}$$

Q 3) Recalculer la probabilité d'observer un pile.

Q 4) On observe un face. Mettez à jour les probabilités.

Q 5) On observe un pile. Mettez à jour les probabilités.

Q 6) On observe un pile. Mettez à jour les probabilités.

Q 7) Conclusion ?

PARTIE 2 - Un deuxième exemple

Les étudiants veulent savoir si les notes de cours seront autorisées à l'examen. Lorsqu'ils demandent à l'enseignant, celui-ci répond de la façon suivante :

- Il commence par lancer une pièce (non truquée). Si elle tombe sur pile, il répond la vérité (oui s'il les notes sont autorisées, non si elles ne le sont pas)
- Si la pièce est tombé sur face, il répondu oui ou non aléatoirement avec probabilité 1/2.

On interroge 3 fois l'enseignant et il répond deux fois oui et une fois non.

Q 1) Reprenez la même étude que précédemment.

PARTIE 3 - Un exemple en continu

On reprend le premier exemple, mais cette fois-ci, on n'a aucun apriori sur la pièce, donc on part avec un apriori d'une distribution uniforme sur p : $P[p \leq t] = t$.

On tombe 3 fois sur pile puis une fois sur face.

Q 1) Reprenez la même étude que précédemment. Attention, c'est plus difficile.

Q 2) Donner un intervalle de crédibilité à 95% pour p.

PARTIE 4 - Les chars allemands

On observe quatre numéros de chars allemands 4,17,4 et 32. Les numéros sont supposés uniformes dans l'intervalle $[0, N]$.

Pour cet exercice, il serait plus judicieux de prendre N comme un paramètre continu plutôt que discret. Cependant, on peut s'en sortir en utilisant l'approximation suivante :

$$\frac{1}{k^m} + \frac{1}{(k+1)^m} + \cdots \sim \frac{1}{(m-1)k^{m-1}}$$

Q 1) Reprenez la même étude que précédemment. Que prendre comme distribution sur N a priori ? Calculez la distribution a posteriori, et sa moyenne.

Q 2) On généralise à n chars allemands et on note Y le plus grnad nombre observé. On note N la distribution a posteriori. Calculer $E(N)$.

Chapitre 4

RÉGRESSION LINÉAIRE

On cherche dans cette partie à essayer de comprendre le lien qui existe entre des variables aléatoires X et Y. On a traité précédemment le cas où X et Y sont indépendantes, et on regarde ici l'autre cas extrême, lorsque Y est une fonction de X.

Il faut d'abord se rendre compte que la question sous cette forme est mal posée.

Si on se donne un échantillon (X_i, Y_i), il est toujours possible de faire passer une courbe par tous les points (X_i, Y_i), mais ce n'est pas forcément la bonne solution. Si par exemple, les points sont $(1,1), (2,2), (3,3.1), (4,4), (5,5), (6,6), (7,7)$, il faudrait un polynôme de degré 7 pour passer pour tous les points, alors qu'il semble plus naturel de penser que le troisième point est faux (erreur de mesure ? bruit ?) et que le bon modèle est une droite. C'est un problème classique appelé surapprentissage (ou sur-ajustement). On va en général se contenter de fonctions f très simples, et en particulier de fonctions *linéaires*, du type $f(x) = ax + b$.

Si jamais il n'y a pas de terme d'erreurs, il n'y a pas besoin de faire des stats pour répondre à la question, on résout le système d'équations $aX_1 + b = Y_1, aX_2 + b = Y_2$ et on a gagné.

En général, la situation est plus compliquée, et on la modélise par

$$Y = aX + b + \epsilon$$

où ϵ est un terme d'erreur, qui peut venir d'erreur de mesure, d'un terme de bruit, ou d'un phénomène inexpliqué.

4.1 Hypothèses

Pour pouvoir retrouver les paramètres a et b à partir de l'échantillon, nous avons besoin d'hypothèses sur ϵ, qui traduisent le fait que ϵ est "indépendant" de X.

Première condition - Exogénéité La moyenne de ϵ, à X fixé, ne dépend pas de X.

$$E(\epsilon|X) = c$$

où c est une constante. Mais que désigne le terme $E(\epsilon|X)$?

Définition 4.1

Si A et B sont deux variables aléatoires, $A|B = t$ est la variable aléatoire C_t telle que

$$P(C_t = k) = P(A = k|B = t)$$

. On note en particulier $E(A|B = t)$ son espérance.

Définition 4.2

$E(A|B)$ est une variable aléatoire D telle que $D = f(B)$ où $f(t) = E(A|B = t)$

♦ **Exemple**

Considérons une population de 5 personnes, ayant respectivement comme poids et taille : $(70, 180), (70, 180), (82, 180), (60, 170), (50, 170)$. On choisit une personne au hasard dans la population et on note T sa taille et P son poids.
Alors $E(P|T = 180) = 74$, $E(P|T = 170) = 55$. La variable $E(P|T)$ est alors une variable qui vaut 74 avec probabilité $3/5$ et 55 avec probabilité $2/5$.

Deuxième condition, $E(\epsilon) = 0$ Si l'éxogénéité est satistfaite, cela revient à dire $c = 0$, où c est la constante ci-dessus. Autrement dit le terme d'erreur s'annule en moyenne. Si on fait une régression linéaire, on peut toujours ajuster le terme a pour que ce soit le cas.

Homoscédasticité

$$\text{Var}\,(\epsilon|X) = d$$

où d est une constante.

Autrement dit, ces trois conditions signifient que moyenne et variance de ϵ ne "dépendent" pas de la valeur de X.

En particulier

$$E(Y|X) = E(aX + b + \epsilon|X) = aE(X|X) + b = aX + b$$

et on peut voir que $E(\epsilon X) = 0$

4.2 Regression linéaire

Théorème 4.1

Sous les hypothèses ci-dessus, a et b sont les coefficients qui minimisent : $E[(Y - (pX + q))^2]$

Preuve :

$$E[(Y - (pX + q))^2] = E[(aX + b + \epsilon - pX - q)^2] = E[(aX + b - pX - q)^2] + E(\epsilon^2)$$

(en utilisant le fait que $E(\epsilon) = E(\epsilon X) = 0$). Le terme en $E(\epsilon)^2 = \text{Var}\ \epsilon$ ne nous intéresse pas puisqu'il ne dépend pas de p et q.

On développe le carré :

$$(a - p)^2 E(X^2) + 2(a - p)(b - q)E(X) + (b - q)^2$$

Et il faut montrer que c'est minimal quand $a = p$ et $b = q$, ce qui est clair puisque dans ce cas, ce terme vaut 0, et que ce terme est non nul (et positif, saut dans le cas $E(X^2) = 0$, mais qui signifie que X est constant égal à 0) sinon. ■

Si on a un échantillon, on peut esayer de déterminer a et b en cherchant l'équivalent "échantillon" de l'espérance.

Théorème 4.2

On définit les estimateurs $\hat{a}$ et $\hat{b}$ comme les nombres qui minimisent

$$\frac{1}{n}\sum_i (Y_i - \hat{a}X_i - \hat{b})^2$$

Comment peut-on les trouver ? Si on dérive par rapport à $\hat{a}$, on obtient :

$$\sum_i X_i(Y_i - \hat{a}X_i - \hat{b}) = 0 = \sum_i X_iY_i - \hat{a}\sum X_i^2 - \hat{b}\sum_i X_i$$

Si on dérive par rapport à $\hat{b}$, on obtient :

$$\sum_i (Y_i - \hat{a}X_i - \hat{b}) = 0 = \sum_i Y_i - \hat{a}\sum X_i - n\hat{b}$$

D'où on peut sortir $\hat{a}$ (On multiplie la deuxième équation par $\sum_i X_i$ et la première par n) :

$$\hat{a} = \frac{\sum Y_i \sum X_i - n\sum X_iY_i}{(\sum X_i)^2 - n\sum X_i^2}$$

où encore

$$\hat{a} = \frac{\widehat{coVar}\,(\overline{X},\overline{Y})}{\widehat{\mathrm{Var}}\,\overline{X}}$$

avec

$$\widehat{coVar}\,(\overline{X},\overline{Y}) = \frac{\sum_i X_iY_i}{n} - \frac{\sum_i X_i}{n}\frac{\sum_i Y_i}{n}$$

(c'est $E(XY) - E(X)E(Y)$ mais calculé sur l'échantillon)

$$\widehat{\mathrm{Var}}\,(\overline{X}) = \frac{\sum_i X_i^2}{n} - \left(\frac{\sum_i X_i}{n}\right)^2$$

Pour $\hat{b}$, le plus simple est d'utiliser

$$\hat{b} = \hat{Y} - \hat{a}\hat{X}$$

Théorème 4.3

$\hat{a}$ et $\hat{b}$ sont des estimateurs non biaisés de a et b.

♦ **Exemple**

0	1	2	3	4	5	6	7	8	9
-0.3	1.6	6.0	6.5	8.4	10.0	13.3	14.6	17.7	17.9

On trouve $\hat{a} = 2.07$ et $\hat{b} = 0.25$ (Note : la taille de l'échantillon ici est trop petit, donc on est loin de la vraie valeur, l'échantillon ayant été généré avec $a = 2$ et $b = 1$).

Comment obtient-on un intervalle de confiance ?

Théorème 4.4

$$\text{Var}\,(\hat{a}) = \frac{\text{Var}\,\epsilon \sum X_i^2}{n \sum (X_i - \hat{X})^2}$$

$$\text{Var}\,(\hat{b}) = \frac{\text{Var}\,\epsilon}{\sum (X_i - \hat{X})^2}$$

De plus, quand la taille de l'échantillon tend vers l'infini, $\hat{a}$ et $\hat{b}$ ont approximativement une distribution normale.

Mais il faut connaire $\text{Var}\,\epsilon$.

Théorème 4.5

$$S = \frac{1}{n-2} \sum (Y_i - g(X_i))^2$$

est un estimateur non biaisé de $\text{Var}\,\epsilon$.

On peut ainsi obtenir des écarts de confiance sur a et b, en pluggant les deux formules ensembles, ce qui est possible si n est suffisament grand.

4.3 Prédiction

Supposons avoir la droite, et une valeur x. Comment puis-je l'utiliser pour prédire la valeur y attendue ? Clairement la formule à utiliser est $y = \hat{a}x + \hat{b}$, qui est un estimateur sans biais pour $E(Y|X = x)$.

Théorème 4.6

Si on veut estimer la valeur moyenne de Y pour $X = x$:

$$\text{Var}\,(E(Y) - \hat{a} - \hat{b}x | X = x) = (\text{Var}\,\epsilon) \left(\frac{1}{n} + \frac{(x - \hat{X})^2}{\sum (X_i - \hat{X})^2} \right)$$

Si on veut donner une valeur de Y :

$$\text{Var}\,(Y - \hat{a} - \hat{b}x | X = x) = (\text{Var}\,\epsilon) \left(1 + \frac{1}{n} + \frac{(x - \hat{X})^2}{\sum (X_i - \hat{X})^2} \right)$$

Dans les deux cas, si n est suffisament grand, on peut supposer que l'estimateur suit approximativement une loi normale.

4.4 Détermination

Le coefficient de détermination est un coefficient qui indique à quel point la fonction $g(x) = ax+b$ trouvée colle aux données.

L'idée est de regarder

$$\sum (Y_i - \hat{Y})^2$$

Notons que ceci peut se réécrire :

$$\sum (Y_i - \hat{Y})^2 = \sum (Y_i - g(X_i))^2 + \sum (g(X_i) - \hat{Y})^2 + 2 \sum_i (Y_i - g(X_i))(g(X_i) - \hat{Y})$$

Il se trouve que le dernier terme est nul,si on a estimé a et b avec les formules ci-dessus.

$$\sum (Y_i - \hat{Y})^2 = \sum (Y_i - g(X_i))^2 + \sum (g(X_i) - \hat{Y})^2$$

Le premier terme est naturel, il explique la déviation de $g(X_i)$ par rapport à sa moyenne, et s'appelle *variance expliquée*. Le coefficient de *détermination* est le rapport entre la variance expliquée et la variance totale. C'est une mesure de la proportion de la variation qui peut s'expliquer par la variation naturelle de la varigble X.

Exercices

(4 - 1) *(Disques durs)*

On cherche à savoir s'il y a une relation entre le prix d'un disque dur SSD et sa capacité.

Une étude menée sur un site spécialisé donne les résultats suivants (sur 197 disques durs)

- Prix moyen : 212 euros
- Prix carré moyen : 66953
- Capacité moyenne : 213 Gb
- Capacité carrée moyenne : 67589
- Produit capacité.prix moyen : 64638

On note P et C les variables prix et capacité

Q 1) Trouver a et b tel que $P \sim aC + b$ et un intervalle de confiance sur a et b.

Q 2) Est-ce que les paramètres collent aux données ?

Q 3) Quel est la prévision du prix moyen d'un disque dur de 500 Gb ?

Note : certaines des questions nécessitent de connaître $\sum_i (Y_i - aX_i - b)^2$ (l'écart entre la valeur réelle et la valeur "calculée"). On peut montrer que cette formule est égale à

$$n\left(\widehat{Var}(Y) - \frac{(\widehat{CoVar}(X,Y))^2}{\widehat{Var}(X)}\right)$$

En particulier le coefficient de détermination peut aussi s'écrire :

$$\frac{(\widehat{coVar}(X,Y))^2}{\widehat{Var(X)}\widehat{Var(Y)}}$$

(4 - 2) *(CPU)*

La loi de Grosch (1965) suggère que si le prix d'un processeur double, sa vitesse quadruple.

Une étude a été menée à partir d'un site spécialisé. Elle donne les résultats suivants :

- Prix moyen : 159
- Prix carré moyen : 55528
- Vitesse moyenne : 4529
- Vitesse carré moyenne : 27511777
- Produit vitesse.prix moyen : 1107815
- (Log prix) moyen : 4.793
- (Log Vitesse) moyen : 8.265
- (Log prix)^2 moyen : 23.4
- (Log Vitesse)^2 moyen : 68.62
- Produit (Log Vitesse)(Log prix) moyen : 39.94

(La vitesse est obtenue à partir d'un benchmark)

Q 1) Trouver a et b tel que $P \sim aV^b$

Q 2) Est-ce que les paramètres collent aux données ?

Q 3) Donner une estimation du prix d'un processeur de vitesse 5000.

(4 - 3) *(Révision)*

On dispose d'une population correspondant à une variable aléatoire (discrète) uniforme sur l'intervalle $[N, N + 100]$, et on cherche à obtenir N.

On observe l'échantillon suivant de taille 4 de la population : 89, 47, 35, 46.

Q 1) En utilisant la méthode des moments, déterminer un intervalle de confiance à 95% pour N. Attention : 4 est un petit nombre.

Q 2) Trouver N par la méthode du maximum de vraisemblance.

(4 - 4) *(Révision 2)*

On lance une pièce 100 fois et on veut savoir s'il s'agit d'une pièce truquée de loi de Bernoulli de paramètre 1/4.

Q 1) Formuler l'hypothèse nulle.

Q 2) On note N le nombre de piles observés. Que doit vérifier N pour que l'échantillon passe un test de la moyenne de valeur 5% ?

Q 3) On note N le nombre de piles observés. Que doit vérifier N pour que l'échantillon passe un test du χ^2 de valeur 5% ?

Q 4) Que constate-t-on ?

(4 - 5) *(Révision 3)*

On dispose de quatre pièces : deux pièces non truquées, une pièce truquée correspondant à une loi de Bernoulli de paramètre 1/4, et une pièce truquée correspodnant à une loi de Bernoulli de paramètre 3/4.

On choisit une pièce parmi les quatre au hasard.

Q 1) On lance 5 fois la pièce et elle tombe 3 fois sur pile et une fois sur face. Quelle est la probabilité que la pièce soit non truquée ?

(4 - 6) *(Révision 4)*

Q 1) On tire un nombre X suivant une loi uniforme sur $[0, 1]$. Calculer la loi de $D = \max(X, 1 - X)$. Trouver λ tel que $P[D > \lambda] \leq 0.05$

Q 2) On tire deux nombres suivant une loi uniforme sur $[0, 1]$. On appelle X_1 le plus petit des deux nombres et X_2 le plus grand des deux nombres. On regarde ensuite le nombre $D = \max(X_1, |X_1 - 1/2|, |X_2 - 1/2|, 1 - X_2)$. Calculer la loi de D, et trouver λ tel que $P[D > \lambda] \leq 0.05$.

Aide : faire 3 cas : $X_2 < 1/2, X_1 > 1/2$ et $X_1 < 1/2 < X_2$.

Q 3) Quel est le rapport avec le cours de statistiques ?

Annexe A

TABLES

Attention, il s'agit de valeurs arrondies.

A.1 Loi Normales

Valeur critique de la loi normale centrée réduite

$P[\|Z\| \leq \lambda]$	λ
80%	1.282
90%	1.645
95%	1.960
98%	2.326
99%	2.576
99.5%	2.807
99.9%	3.291

A.2 Loi du χ^2

Le chiffre en position (d, α) représente la valeur critique λ tel que $P[X > \lambda] \leq \alpha$ pour X une variable du χ^2 de degrés de liberté d.

d	0.2	0.1	.05	.02	.01	.005	.001
1	1.642	2.706	3.841	5.412	6.635	7.879	10.828
2	3.219	4.605	5.991	7.824	9.210	10.597	13.816
3	4.642	6.251	7.815	9.837	11.345	12.838	16.266
4	5.989	7.779	9.488	11.668	13.277	14.860	18.467
5	7.289	9.236	11.070	13.388	15.086	16.750	20.515
6	8.558	10.645	12.592	15.033	16.812	18.548	22.458
7	9.803	12.017	14.067	16.622	18.475	20.278	24.322
8	11.030	13.362	15.507	18.168	20.090	21.955	26.124
9	12.242	14.684	16.919	19.679	21.666	23.589	27.877

A.3 Kolmogorov-Smirnov

Le chiffre en position (n, α) représente la valeur critique λ tel que $P[D > \lambda] \leq \alpha$ pour D une variable suivant une loi de Kolmogorov-Smirnov de paramètre n.

n	0.05	0.01
1	.975	.995
2	.842	.929
3	.708	.828
4	.624	.733
5	.565	.669
6	.521	.618
7	.486	.577
$\sim \infty$	$1.36/\sqrt{n}$	$1.63/\sqrt{n}$

Annexe B

DEVOIR

Ce devoir doit être rendu avant le ... , en version électronique (PDF seulement) à l'adresse `emmanuel.jeandel@loria.fr` *ou délivrée en mains propres à E. Jeandel. Plusieurs questions du devoir nécessitent d'écrire un programme. Celui-ci devra être envoyé exclusivement par mail pour la même date.*

Le devoir doit être effectué par groupe de 1 *à* 2 *personnes, chaque personne appartenant à exactement un groupe. Chaque devoir contiendra dans une première partie au moins un paragraphe expliquant comment la répartition du travail s'est effectuée entre les différents membres d'un même groupe. Cette partie n'est facultative que pour les groupes d'une personne et sera prise en compte dans la notation.*

Pour tracer et rendre les différentes courbes, on pourra utiliser gnuplot. En ajoutant la ligne `use terminal png` *puis la ligne* `set output "toto.png"` *en haut d'un fichier gnuplot, le résultat ira dans le fichier* `toto.png`*.*

PARTIE 1 - Fin des TDs

Q 1) On se donne un échantillon $X_1 \dots X_n$ de taille n d'une population X suivant une loi uniforme sur $[0, N]$. On note Y le maximum de $X_1 \dots X_n$ et $Z = \frac{n+1}{n} Y$. Calculer $E(Y)$ et $\mathrm{Var}\, Y$ et en déduire $E(Z)$ et $\mathrm{Var}\, Z$. Montrer que Z est un meilleur estimateur de N que 2 fois la moyenne de l'échantillon.

On considère maintenant une distribution "tente" sur l'intervalle $[0, 1]$ dont la densité est

$$f(x) = \begin{cases} \frac{2x}{c} & x < c \\ \frac{2(1-x)}{1-c} & x \geq c \end{cases}$$

Q 2) Représenter $f(x)$

Q 3) Donner une expression de $P[X \leq t] = \int_0^t f(x)dx$. Vérifier que $P[X \leq 1] = 1$. Représenter le résultat quand $c = 1/3$.

On observe les données 0.2 et 0.4.

Q 4) Estimer c en utilisant la méthode des moments. Qu'observe-t-on ?

Q 5) Estimer c en utilisant le maximum de vraisemblance.

PARTIE 2 - Stratification et méthode des quotas

On cherche à estimer la taille de la population. La variance de la taille peut être élevée dans la population, mais on s'aperçoit en pratique qu'elle elle est plus faible si on se restreint à la population masculine (resp. féminine).

On se propose de modéliser le phénomène de la façon suivante. On note F (resp. H) la variable aléatoire "taille" au sein de la population féminine (resp. masculine), et soit X une variable aléatoire de Bernoulli de paramètre $p = 0.513$ (représentant la proportion de femmes). Les trois variables X, F, H sont indépendantes.

La v.a. représentant la taille dans la population est donc

$$T = XF + (1 - X)H$$

Q 1) Exprimer $\mu_T = E(T)$ en fonction de p et de $\mu_F = E(F)$ et $\mu_H = E(H)$.

Q 2) Montrer que $\mathrm{Var}\, T = p\mathrm{Var}\, F + (1-p)\mathrm{Var}\, H + p(1-p)(\mu_F - \mu_H)^2$. (Aide : Calculer $E(T^2)$, puis utiliser $\mathrm{Var}\, T = E(T^2) - E(T)^2$. De plus, comme X vaut 0 ou 1, $X^2 = X$).

On dispose maintenant de plusieurs moyens de sonder la population :

Q 3) On choisit 1000 personnes au hasard, et on calcule la moyenne $\hat{T}_{1000} = \frac{T_1 + \cdots + T_{1000}}{1000}$. Exprimer la variance de $\hat{T}_{1000}$.

Q 4) On choisit 513 femmes au hasard, et on calcule la moyenne $\hat{F}_{513}$ et on choisit 487 hommes au hasard, et on calcule la moyenne $\hat{H}_{487}$ puis on calcule $\hat{T} = \frac{513\hat{F}_{513} + 487\hat{H}_{487}}{1000}$. Montrer que $\hat{T}$ est un estimateur non biaisé de $E(T)$. Calculer sa variance. En déduire que cette méthode est *toujours* meilleure que la première.

Note : Pour que cette méthode soit applicable, il est nécessaire de connaître le coefficient p exactement. Sinon on fausse les résultats. C'est le problème avec la méthode des quotas, comme utilisée dans les sondages pour les élections. Si on choisit par exemple de diviser la population entre hommes et femmes pour le sondage, il faut, pour connaître p, connaître la proportion des femmes parmi les personnes qui vont voter, et non pas la proportion totale dans toute la population.

Une troisième méthode consiste à choisir 1000 personnes, puis à calculer la moyenne $\hat{F}$ des tailles sur les femmes, et la moyenne des $\hat{H}$ des tailles sur les hommes, puis à faire comme précédemment $\hat{T} = .513\hat{F} + 487\hat{H}$.

La difficulté pour analyser est maintenant que $\hat{F}$ et $\hat{H}$ ne sont pas indépendantes. Par exemple, si on suppose que les tailles sont toujours entières (exprimées en cm) alors la probabilité que $\hat{F} = \frac{517}{3} = 172,33\ldots$ est non nulle (c'est le cas par exemple s'il y a seulement 3 femmes sondées sur l'échantillon de 1000 personnes, et 2 d'entre elles mesurent 1.72m et la troisième 1.73m) et $\hat{H} = \frac{517}{3}$ est tout aussi possible, mais on ne peut pas avoir les deux simultanément (La seule possibilité pour avoir cette valeur de $\hat{F}$ est d'avoir interrogé un nombre de femmes multiple de 3, et de même pour les hommes, ce qui est impossible avec un total de 1000 personnes).

Ce calcul vous est donc épargné.

PARTIE 3 - Problème du bandit à plusieurs bras

Le problème du bandit à plusieurs bras est un problème fondamental en statistiques. Imaginons un site web spécialisé dans la critique d'albums et EPs. Pour s'assurer un revenu lui permettant de continuer à maintenir la page web, le webmestre installe une publicité. A chaque fois que l'utilisateur clique sur la publicité, le webmestre gagne 1 euro.

Le webmestre a le choix entre plusieurs publicités : l'une vers `www.m3tal.org`, un site de ventes d'album de metal, l'une vers `www.heap-hop.org`, site spécialisé dans le hip-hop, et l'une vers `www.geazz.org`, site spécialisé dans les vinyls de jazz.

La meilleure solution pour le webmestre est de choisir la publicité qui plairait au plus de personnes. Il ne connaît malheureusement pas cette donnée, et il peut juste se contenter de l'estimer en fonction de ce qu'il a observé.

Le problème se généralise et se formalise ainsi. On se donne $p_1 \dots p_k$, la probabilité de succès des choix $1 \dots k$. Les probabilites $p_1 \dots p_k$ sont inconnues.

A chaque étape, en fonction de ce qu'il s'est passé précédemment, le webmestre choisit la publicité i, et l'utilisateur va cliquer (et donc le webmestre va obtenir un euro) avec probabilité p_i.

En moyenne, le meilleur choix poiur le webmestre est donc de mettre la publicité qui correspond à la plus grande valeur de p_i, qu'il ne connait pas. Au bout de n étapes de temps, le webmestre gagnerait en moyenne $n \max_{i \leq k} p_i$ en suivant cette politique.

Il existe plusieurs algorithmes qui permettent de s'approcher de cette valeur. Pour les comparer, on regarde le "regret moyen", qui est la différence entre le gain maximal potentiel ($n \max_{i \leq k} p_i$) et le gain de la stratégie.

L'objectif de ce problème est de programmer différentes stratégies pour les comparer, et de montrer théoriquement que l'une d'entre elles (appelée UCB1 dans la littérature) obtient de très bons résultats.

Stratégies

Stratégie UCB1

La première stratégie, dont on prouvera plus loin qu'elle est très bonne en théorie, fonctionne ainsi. A chaque étape, garder en mémoire pour chaque i, combien de fois la publicité i à été placée (on note ce nombre n_i), combien de fois on a cliqué sur la publicité i (on note ce nombre $x_i \leq n_i$) et combien de fois l'expérience a été réalisée (noté n).

La stratégie choisit alors la publicité i qui maximise $x_i/n_i + \sqrt{\frac{2 \ln n}{n_i}}$.

Pour initialiser la stratégie, on commence par placer une fois chacune des publicités.

ϵ-gloutonne

Cette stratégie probabiliste choisit, avec probabilité $1 - \epsilon$, de placer la publicité qui a été cliquée le plus souvent (en moyenne), et avec probabilité ϵ une publicité choisie au hasard uniformément.

Méthode de Thompson

L'idée vient de l'approche Bayesienne des statistiques qu'on évoquera plus tard en cours. L'approche est similaire à la méthode du maximum de vraisemblance. Pour chaque i, considérons la fonction $f_i(p) = p^{x_i}(1-p)^{n_i-x_i}$ (en reprenant les notations précédentes). Dans le principe du maximum de vraisemblance, on chercherait p qui maximise cette expression. Ici, on voit ça comme une *distribution de probabilité* sur p. On regarde donc :

$$g_i(p) = \frac{(n_i+1)!}{(x_i)!(n_i-x_i)!} p^{x_i}(1-p)^{n_i-x_i}$$

(le coefficient devant g_i est juste là pour que ça soit une probabilité, donc que ça somme à 1 lorsqu'on regarde toutes les valeurs possibles pour p) et la variable aléatoire G_i dont la densité est g_i.

L'idée de la méthode de Thompson est alors la suivante. A chaque étape :

- Pour chaque i, tirer p_i suivant la densité g_i
- Choisir le i qui maximise p_i.

Pour choisir p_i suivant la densité p_i, on utilisera le fichier `Beta.java` qui vous est fourni. Il faut appeler la fonction `samplebeta` avec comme paramètres $1 + x_i$ et $1 + n_i - x_i$.

Méthode par poursuite

A chaque étape, chaque publicité a une probabilité p_i d'être placée. Au début, chaque publicité a exactement une chance sur k d'être choisie.

La probabilité évolue alors de la façon suivante : Si à un moment donné, c'est la publicité j qui est celle qui a le plus grand ratio succès/placés (x_i/n_i), alors on change tous les coefficients p_i de la façon suivante :

$$p'_i := \begin{cases} p_i + \beta(1-p_i) & \text{si } i = j \\ p_i - \beta p_i & \text{sinon} \end{cases}$$

où β est un facteur d'apprentissage qu'on choisira égal à 0.2.

Implémentation

Q 1) Implémenter les trois méthodes et comparer leur efficacité. Pour cela, on examinera quel est le gain du webmestre en suivant chacune des 4 méthodes dans le cas où il y a 3 publicités, l'une de paramètre $p_1 = 0.3$, l'autre de paramètre $p_2 = 0.5$ et la troisième de paramètre $p_3 = 0.55$. On testera les 4 méthodes pour un nombre de clics de 4 à 1000 et on produira un graphique en gnuplot permettant de comparer les résultats. Pour la deuxième méthode, on choisira plusieurs valeurs de ϵ.

Note : Pour avoir une idée du genre de courbes qu'on souhaite obtenir, on pourra regarder l'article http://www.cs.mcgill.ca/~vkules/bandits.pdf.

Efficacité de UCB1

On va montrer ici que la première méthode, en moyenne, va toujours choisir la bonne publicité à placer. Plus exactement, on montre qu'elle ne choisira la mauvaise publicité qu'un nombre logarithmique de fois sur un nombre de clics de taille n.

Pour simplifier les choses, on va se contenter du cas où on doit choisir entre deux publicités qu'on appelle X et Y.

Commençons par décrire l'algorithme :

```
def facteur(places, succes, total):
  return succes/places + sqrt(2 ln total/places)

#initialisation
x = pubx()
y = puby()
nx = 1
ny = 1

POUR i allant de 3 à n
  si facteur(nx,x,i) >= facteur(ny,y,i):
     # on place x
     x = x + pubx()
     nx = nx + 1
  sinon:
     # on place y
     y = y + puby()
     ny = ny + 1
```

`pubx` est une routine qui place la publicité X, et renvoie 1 si l'utilisateur a cliqué et 0 sinon. C'est donc techniquement une fonction qui renvoie 1 avec probabilité p_X.

On note X_i (resp. Y_i) le résultat de la fonction `pubx` (resp. `puby`) la $i$ème fois qu'on l'a lancé. C'est donc une variable de Bernoulli de paramètre p_X (resp. p_Y).

Pour analyser le problème, on va supposer que la première publicité est plus intéressante, c'est à dire $p_X > p_Y$.

On note enfin $\bar{X}_n = X_1 + \cdots + X_n$ (et idem pour Y). $\bar{X}_n$ est donc le nombre de succès obtenus si on place n fois la publicité X. $\bar{X}_{nx}$ correspond donc à la variable `x` du programme.

Q 2) On se place à l'étape de temps i. On suppose qu'on a placé pour l'instant nx fois la publicité X et ny fois la publicité Y ($nx + ny = i$). Montrer qu'on va placer la publicité Y si et seulement si $\texttt{facteur}(nx, \bar{X}_{nx}, i) < \texttt{facteur}(ny, \bar{Y}_{ny}, i)$

Q 3) En déduire que si on a placé la publicité Y à l'instant i, il existe $k \leq i$ et $k' \leq i$ tel que $\texttt{facteur}(k, \bar{X}_k, i) < \texttt{facteur}(k', \bar{Y}_{k'}, i)$

Q 4) Justifier que le nombre de fois où on a placé la publicité Y entre les instants 3 et n est inférieur à

$$\sum_i \sum_{k \leq i} \sum_{k' \leq i} P\left(\texttt{facteur}(k, \bar{X}_k, i) < \texttt{facteur}(k', \bar{Y}_{k'}, i)\right)$$

(Aide : Introduire des variables intermédaires T_i, avec $T_i = 1$ si on place la publicité Y à l'instant i, et $T_i = 0$ sinon et utiliser la linéarité de l'espérance)

Pour la preuve, on a besoin d'être un peu plus subtil. Soit l une constante, qu'on fixera plus tard.

Q 5) Justifier que le nombre de fois où on a placé la publicité Y entre les instants 3 et n est inférieur à

$$l + \sum_i \sum_{k \leq i} \sum_{l \leq k' \leq i} P\left(\texttt{facteur}(k, \bar{X}_k, i) < \texttt{facteur}(k', \bar{Y}_{k'}, i)\right)$$

(Aide : compter indépendamment les l premières fois où a placé la publicité Y (représentées par la constante l) et les fois suivantes.

Il reste à évaluer ce terme. Revenons à l'expression de `facteur` :

```
def facteur(places, succes, total):
  return succes/places + sqrt(2 ln total/places)
```

Q 6) Montrer que si $\texttt{facteur}(k, \bar{X}_k, i) < \texttt{facteur}(k', \bar{Y}_{k'}, i)$ alors l'une des trois conditions suivante est vérifiée :

- $\dfrac{\bar{X}_k}{k} \leq p_X - \sqrt{2 \ln i/k}$
- $\dfrac{\bar{Y}_{k'}}{k'} \geq p_Y + \sqrt{2 \ln i/k'}$
- $p_X < p_Y + 2\sqrt{2 \ln i/k'}$

Q 7) Montrer en utilisant Chernoff que la probabilité que la première condition se réalise (pour k et k' donné) est inférieur à $2/i^4$, et de même pour la deuxième.

Q 8) Montrer que si on prend $l = \frac{8 \ln i}{(p_X - p_Y)^2}$ alors la troisième condition n'est jamais possible (on rappelle que $k' \geq l$)

Q 9) En déduire que le nombre moyen de fois où on a placé la deuxième publicité entre les instants 3 et n est inférieur à

$$\frac{8 \ln i}{(p_X - p_Y)^2} + 1 + \sum_i \sum_{k \leq i} \sum_{1 \leq k' \leq i} \frac{4}{s^4}$$

(Aide : le terme "+1" vient uniquement du fait que le terme à gauche n'est pas forcément entier)

On peut démontrer que la grosse somme vaut au plus $2\pi^2/3 \leq 7$, donc le nombre de fois où on se trompe de bras est logarithmique en i (et inversement proportionnel à la différence entre $p_X et p_Y$).

Note : cette preuve formelle est inspirée de Auer,Cesa-Bianchi,Fischer *Finite-time analysis of the Multiarmed Bandit Problem*.