TP L3 ISIMA : Estimation

L'objectif de ce TP est d'estimer par une statistique une mesure probabiliste. Par exemple, en mesurant la fréquence de réalisation d'une valeur sur un dé, estimer la probabilité de réalisation de cette valeur. La difficulté principale est de déterminer la quantité d'échantillons à tester afin d'obtenir une estimation suffisante par rapport à une problématique donnée.

Introduction

---

1 Approche graphique

Soit X une variable aléatoire de support $\{0;1\}$, telle que $p(X=1)=0.1$. On définit $f(n)$ la fréquence de réalisation du 1 dans une suite de n réalisations de X.

1. Représenter plusieurs courbes $|f(n)-P(X=1)|$
2. Représenter plusi eurs courbes $\frac{|f(n)-P(X=1)|}{P(X=1)}$.
3. Construisez une équation approximative de la vitesse de convergence dans les deux cas précéents.

2 Création des outils

L'inégalité de Bienaymé Tchebychev affirme que si une variable aléatoire U admet une espérance $E(U)$ et un écart type fini σ, alors $\forall\alpha\in\mathbb{R}_+^*\ P(|U-E(U)|\geq\alpha)\leq\frac{\sigma^2}{\alpha^2}$.

1. D'après ce théorème, si on désire estimer $P(X=1)$ à 10^{-3} près, avec une fiabilité de la réponse d'au moins 99%, quelle doit être la valeur de 'n' ?
2. Vérifiez votre résultat de façon expérimentale.

On appelle intervalle de prédiction d'une variable aléatoire X, au seuil de confiance (ou avec une fiabilité) $1-\alpha$ le plus petit intervalle I vérifiant $P(X\in I)\geq 1-\alpha$. On peut démontrer (via une approximation par une loi normale) qu'asymptotiquement dans le cas d'une estimation de proportion,

$$I=\left[\frac{1}{n+c^2}\left(nf+\frac{c^2}{2}-c\sqrt{nf(1-f)+\frac{c^2}{4}}\right)\ ;\ \frac{1}{n+c^2}\left(nf+\frac{c^2}{2}+c\sqrt{nf(1-f)+\frac{c^2}{4}}\right)\right]$$

avec f la fréquence observée de l'évènement, n le nombre d'expériences de Bernoulli, et c tel que pour U une variable aléatoire de loi normale centrée réduite, $P(U\leq c)=1-\frac{\alpha}{2}$.

Dans la suite, on suppose que l'asymptote est atteinte dès que

$$\begin{cases} n\geq 30 \\ n\times f\geq 5 \\ n\times(1-f)\geq 5 \end{cases}$$

Dans l'éventualité où on réalise une estimation d'un paramètre Θ par une valeur ponctuelle θ, avec une fiabilité $1-\alpha$ la précision est la plus petite valeur η vérifiant $P(|\Theta-\theta|\leq\eta)\geq\alpha$. On dispose alors d'une estimation par fourchette : $P(\theta-\eta\leq\Theta\leq\theta+\eta)\geq\alpha$

Loi de proba inversenormale et Python

```
from scipy.stats import norm                          # Librairie de statistiques
print(norm.ppf(0.975, loc = 0, scale = 1))            # Affiche u tel que P(X <= u) = 97,5%
                                                      # pour X suivant une loi normale
                                                      # d'espérance 0 et d'écart type 1
print(norm.cdf(1.96, loc = 0, scale = 1))             # Affiche P(X <= 1.96)
                                                      # pour X suivant une loi normale
                                                      # d'espérance 0 et d'écart type 1
```

3. Sur les cas traités précédemment expliquer, en prenant appuis sur les courbes, ce que sont la précision et la fiabilité.

4. Ecrire une fonction prenant en arguments la précision η ainsi que la fiabilité $1-\alpha$ et renvoyant la valeur de n minimale permettant d'obtenir le résultat désiré.

Après avoir copié le modules 'boite_noire' dans votre répertoire de travail, importez-le. Il contient une classe 'quota' qui simule la réponse (oui/non) d'un individu tiré au hasard. Il s'utilise comme suit :

Utilisation du générateur 'quota'

```
from boite_noire import quota
generateur = quota()                    # création du générateur
for k in range(100):                    # on va utiliser 100 fois ce générateur
    print(generateur())                 # pour générer des données ('oui' / 'non')
```

5. Ecrivez une fonction python prenant en entrée une précision et une fiabilité, et qui renvoie une estimation de la probabilité d'obtenir 'oui'.

Méthode des quotas / conditionnement

---

3 Découverte de la méthode

1. L'idée sous-jacente à la méthode des quotas est tirée de la formule définissant l'incertitude I. Représenter la largeur de l'intervalle I en fonction de f, tous les autres paramètres étant fixés.
2. En déduire que ceraines estimations sont plus simples à réaliser que d'autres.
3. En reprenant les courbes et en modifiant la valeur de la probabilité de réalisation du 1, mettre en évidence graphiquement ce que vous avez observé par l'examen de la formule.

La méthode des quotas consiste à essayer de se ramener à plusieurs estimations 'simples' plutôt que d'essayer d'établir une estimation 'compliquée'. Pour cela, on réalise une partition où les individus d'une classe doivent avoir des réponses les plus homogènes possibles.

4 Mise en place

Le générateur *quota* peut être applelé avec un paramètre '$generateur(categorie\_visible = True)$' et renvoie alors un couple : (catégorie, réponse), où catégorie est un élément de {A, B, C} et réponse est dans {oui ; non}. Cette façon d'utiliser le générateur correspond à choisir au hasard un individu, puis mesurer sa catégorie ainsi que sa réponse. Le générateur peut également être appelé avec un paramètre '$generateur(etat = xxx)$' où $xxx \in \{'A', 'B, 'C'\}$ ce qui simule la réponse d'un individu tiré au hasard dans la catégorie xxx.

4. Estimer la probabilité pour un individu d'être de chacune des catégories pour une fiabilité et une incertitude données
5. Estimer pour une précision et une fiabilité données $\left| \begin{array}{l} P_{\text{catégorie}=A}(\text{réponse} = 1) \\ P_{\text{catégorie}=B}(\text{réponse} = 1) \\ P_{\text{catégorie}=C}(\text{réponse} = 1) \end{array} \right.$
6. Quelles sont la précision et la fiabilité de l'estimation de P(répnse = oui) sachant que :
   - si $P(A)$, $P(B)$, $P(C)$, $P_A(reponse = 1)$, $P_B(reponse = 1)$ et $P_C(reponse = 1)$ sont connues avec une même fiabilité $1-\alpha$
   - avec des précisions respectives θ_A, θ_B, θ_C, $\theta_{rep|A}$, $\theta_{rep|B}$, $\theta_{rep|C}$
7. En déduire une estimation de P(réponse = oui), pour une fiabilité et une précision données.

5 Mise en oeuvre

8. Ecrire une fonction qui prend en entrée N le nombre d'individus que vous pouvez interroger et qui vous renvoie une estimation de P(réponse = oui). La difficulté ici est de choisir le nombre de 'A', 'B' et 'C' à interroger afin d'obtenir la meilleure estimation possible (à N fixé, et N asez grand, de façon à ce qu'en interogeant au hasard des individus dans la population, on ait interrogé au moins 30 individus dans chaque classe). Pour vous aider, voici un plan possible (mais pas parfait) pour élaborer votre algorithme :

— Votre fonction accepte deux entrées :

N :	nombre total d'individus que l'on peut interroger
$fiabilite$:	la fiabilité de l'estimation à renvoyer

et renvoie

ϕ :	une estimation de P(réponse $= 1$)
θ :	la fiabilité de l'estimation ϕ

— Commencer par interroger des individus au hasard dans la population jusqu'à obtenir 30 individus au minimum dans chaque catégorie (expliquer pourquoi cette phase n'est parfois ni possible, ni nécessaire, mais ... tant pis !)

— Pour chaque individu que vous pouvez encore interroger (vous n'avez pas encore atteint N), déterminer quel couple (catégorie xxx, réponse) engendre la plus grande incertitude sur votre estimation θ,

— Le couple étant déterminé, entre P(catégorie = xxx) et $P_{xxx}(reponse = 1)$ déterminer celui qui engendre la plus grande incertitude sur θ

— Si c'est la catégorie, interroger un individu au hasard dans la population, et mettre à jour les probabilités des catégories ainsi que tout ce qui en découle

— Si c'est la probabilité conditionnelle, interroger au hasard un individu de la catégorie 'xxx' et mettre à jour $P_{xxx}(reponse = 1)$ ainsi que tout ce qui en découle.

— recommencer jusqu'à avoir interrogé N individus, et renvoyer l'estimation ainsi que l'incertitude

9. Comparer l'incertitude obtenue par la méthode sans quota et par la méthode avec quotas.
10. Représenter graphiquement la valeur de l'estimation en fonction de N, en utilisant la méthode des quotas, et sans l'utiliser

6 Pour aller plus loin

Il serait interressant de construire les catégories 'à la volée' afin de construire des classes au fur et à mesure du sondage qui soient adaptées au critère que l'on désire estimer... ce qui correspond à un partionnement dynamique de l'espace des individus.

Jeu de cartes

7 Estimations de probabilité

Le but de cette partie est d'estimer la probabilité de réalisation de certaines mains de 5 cartes tirées aléatoirement dans un jeu de 52 cartes, ce qui nous mettra sur la voie de la réalisation d'un jeu de Poker, et d'un point de vue plus général, à la prise de décision en univers incertain.
Estimer la probabilité de réalisation des situations suivantes. Chacune des estimations doit correspondre à une fonction Python, prenant en entrée une précision et une fiabilité, renvoyant l'estimation et affichant le nombre de réalisations utilisées.

1. Obtenir une main contenant 4 as
2. Obtenir un carré
3. Obtenir au moins 2 as
4. Obtenir au moins 2 as, dont l'as de coeur
5. Obtenir au plus 2 as
6. Obtenir uniquement du coeur
7. Obtenir au moins 2 coeurs et 1 as
8. Obtenir uniquement des figures
9. Obtenir au moins 2 cartes rouges et pas de figure
10. Obtenir au moins une carte rouge, dans l'éventualité où les trois premières cartes deéjà regardées sont rouges.
11. Obtenir une grande suite (les valeurs des cartes se suivent, l'as est avant le 2 ou au dessus du roi)
12. Obtenir une grande suite de valeurs dans le même bois
13. En fonction de leur probabilité de réalisation, attribuer une note 'honnête' sur 100 à chacune des combinaisons suivantes : paire, double paire, brelan, full, carré, bois, couleur , suite couleur, suite bois.

8 Des probabilités au comptage

Usuellement, on utilise des calculs de dénombrement pour calculer des probabilités. Par exemple dans le cas d'un univers fini, où les évenements sont équiprobables, on calcule $P(A) = \frac{|A|}{|U|}$. Il est possible de retourner cette approche : connaissant p et $|U|$ on peut calculer $|A|$. En général on ne connait qu'une estimation de p (et parfois un intervalle de vraissemblance), et on recherche une estimation du dénombrement correspondant.

14. On se place dans le cas d'un univers équiprobable, on connaît une estimation de $p(A)$ notée θ avec une précision η et une fiabilité $1 - \alpha$. Retranscrire ces informations sur une estimation de $|A|$
15. Reprendre les situations précédentes afin d'obtenir les dénombrements correspondants.
16. Confrontez vos résultats expérimentaux aux résultats exacts (papier/crayon ! !)

9 Un début de jeu de Poker

Le jeu de Poker auquel nous allons nous intéresser se conforme aux règles suivantes :

— Le jeu se joue à deux joueurs

— Chaque tour, le premier joueur est celui qui ne était pas à la partie précédente

— Deux cartes sont distribuées à chacun des joueurs

— Trois cartes sont posées faces visibles au centre de la table

— Chaque joueur estime sa main, qui est constituée par au maximum 5 cartes prises parmi celles qu'il a dans la main et celles qui sont sur la table. Le premier joueur dépose une mise sur la table ou abandonne. L'autre joueur doit alors pour continuer à jouer déposer la même mise.

— Si aucun des joueurs n'a abandonné, une nouvelle carte est distribuée à chaque joueur, et retour à l'étape précédente jusqu'à un maximum de 4 cartes dans la main de chaque joueur.

— Si un joueur abandonne, l'autre gagne la totalité de la mise ; si aucun des joueurs n'a abandonné jusqu'à distribution des quatre cartes à chacun des joueurs, alors les jeux sont abattus, et le gagnant est celui qui a la combinaison la plus forte (les combinaisons possibles sont celles vues dans un exercice précédent).

17. Estimer l'espérance de la note d'une main de 5 cartes
18. Une main contient 'p' cartes connues, et q cartes inconnues, estimer l'espérance de la note de votre main (la fonction prend en entrée les p cartes connues ainsi que q).
19. Simuler une partie, où l'ordin ateur est le deuxième joueur et décide de suivre lorsque son estimation de l'espérance de la note de son jeu est supérieure à son estimation de la note du premier joueur.