Devoir surveillé n°5

Samedi 8 février (8h - 11h)

La calculatrice est *autorisée*.

Notations et rappels

- Si X est une variable aléatoire admettant une espérance ou une variance, elles sont notées respectivement $E(X)$ et $V(X)$.
- On rappelle le résultat suivant : si X et Y sont deux variables aléatoires à densité indépendantes de densités respectives f_X et f_Y, alors $Z = X + Y$ est une variable aléatoire à densité, et une densité de Z est donnée par :
$$f_Z : x \mapsto \int_{-\infty}^{+\infty} f_X(t) f_Y(x-t)\,\mathrm{d}t$$
- Les programmes en Python pourront utiliser, sans qu'il soit nécessaire de les importer, les fonctions `random` et `randint`. On rappelle que `random()` renvoie un nombre de l'intervalle $]0,1[$ de loi uniforme sur cet intervalle, et que `randint(i, j)` renvoie un entier de l'intervalle $[\![i,j]\!]$ avec équiprobabilité. Les différents appels de ces fonctions sont supposés donner des résultats indépendants.

Partie I. La Loi bêta

Dans toute cette partie I, a et b désignent deux entiers naturels non nuls.

1. Densité, espérance et variance

1. On pose $B(a,b) = \displaystyle\int_0^1 x^{a-1}(1-x)^{b-1}\,\mathrm{d}x$.

(a) Montrer l'égalité : $B(a+1,b) + B(a,b+1) = B(a,b)$.

(b) À l'aide d'une intégration par parties, établir : $B(a,b+1) = \dfrac{b}{a}\,B(a+1,b)$

(c) En déduire la formule : $B(a+1,b) = \dfrac{a}{a+b}\,B(a,b)$.

(d) Montrer par récurrence : $B(a,b) = \dfrac{(a-1)!\,(b-1)!}{(a+b-1)!}$.

2. On définit la fonction f sur $\mathbb{R}$ par $f(x) = \begin{cases} \dfrac{1}{B(a,b)}\,x^{a-1}(1-x)^{b-1} & \text{si } x \in]0,1[\\ 0 & \text{sinon.} \end{cases}$

Montrer que f est une densité de probabilité sur $\mathbb{R}$.

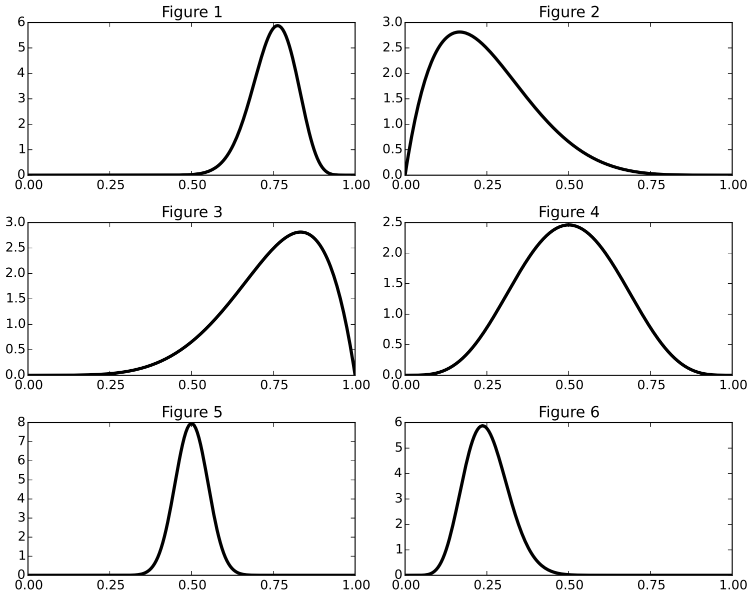
La loi de probabilité associée est appelée *loi bêta de paramètres a et b*, que l'on note $\beta(a,b)$ dans toute la suite.

3. Soit X une variable aléatoire suivant la loi $\beta(a,b)$.

(a) Montrer que $E(X) = \dfrac{a}{a+b}$.

(b) Montrer que $V(X) = \dfrac{ab}{(a+b)^2(a+b+1)}$.

(c) Les six graphiques qui suivent représentent des densités des lois bêta de paramètres $(5,5)$, $(50,50)$, $(2,6)$, $(6,2)$, $(30,10)$, $(10,30)$.
En interprétant ce que représentent l'espérance et la variance, identifier les paramètres qui correspondent à chaque graphique.



4. Soit F_X la fonction de répartition de X.

(a) Pour un réel $x \in [0,1]$, écrire $F_X(x)$ à l'aide d'une intégrale.
Quelle est la valeur de $F_X(x)$ lorsque $x < 0$ et lorsque $x > 1$?

(b) Écrire en Python une fonction `fact(n)` qui prend en argument un entier naturel n et qui renvoie $n!$.

(c) En utilisant la méthode des rectangles, que l'on détaillera, écrire une fonction `F(a, b, x)` qui renvoie une valeur approchée de $F_X(x)$ pour un argument $x \in [0,1]$. On pourra par exemple subdiviser l'intervalle d'intégration en 1000 intervalles.

2. La loi Gamma

Dans la suite, λ désigne un réel strictement positif, et n un entier naturel non nul.

On considère $X_1, X_2, \ldots, X_n$ des variables aléatoires indépendantes suivant toutes la loi exponentielle de paramètre λ. On définit la variable aléatoire S_n par $S_n = \sum_{k=1}^{n} X_k$.

5. Montrer, par récurrence, qu'une densité de S_n est donnée par :

$$f_{S_n}(x) = \begin{cases} \dfrac{\lambda^n}{(n-1)!}\, x^{n-1}\, \mathrm{e}^{-\lambda x} & \text{si } x > 0 \\ 0 & \text{sinon} \end{cases}$$

On appelle la loi associée *loi Gamma de paramètres n et λ*, que l'on note $\Gamma(n,\lambda)$.

6. On pose $I(n,\lambda) = \displaystyle\int_0^{+\infty} x^n \,\mathrm{e}^{-\lambda x}\,\mathrm{d}x$.

En utilisant la question précédente, montrer que $I(n,\lambda) = \dfrac{n!}{\lambda^{n+1}}$.

7. Simulation de la loi Gamma.

(a) Montrer que si une variable aléatoire U suit la loi uniforme sur $]0,1[$, alors la variable

$$X = -\frac{1}{\lambda}\ln(1-U)$$

suit la loi exponentielle de paramètre λ.

(b) Écrire une fonction `gamma(n, lam)` d'arguments `n` et `lam` qui simule une variable aléatoire suivant la loi Γ de paramètres `n` et `lam`.

3. Simulation de la loi bêta

Soit λ un réel strictement positif.

Soit X et Y deux variables aléatoires à valeurs dans $]0,+\infty[$, indépendantes, et qui suivent respectivement la loi $\Gamma(a,\lambda)$ et $\Gamma(b,\lambda)$. On rappelle que a et b sont deux entiers strictement positifs.

8. On pose $U = \ln X$ et $V = -\ln Y$.

(a) Montrer que U est une variable à densité, et qu'une densité de U est donnée sur $\mathbb{R}$ par :

$$f_U(x) = \frac{\lambda^a}{(a-1)!}\exp\left(ax - \lambda\,\mathrm{e}^{x}\right)$$

(b) Montrer que V est une variable à densité, et qu'une densité de V est donnée sur $\mathbb{R}$ par :

$$f_V(x) = \frac{\lambda^b}{(b-1)!}\exp\left(-bx - \lambda\,\mathrm{e}^{-x}\right)$$

(c) On pose $W = U + V$. Montrer qu'une densité de W est donnée sur $\mathbb{R}$ par :

$$f_W(x) = \frac{\lambda^{a+b}}{(a-1)!(b-1)!}\,\mathrm{e}^{-bx}\int_{-\infty}^{+\infty}\exp\left[(a+b)t - \lambda\left(1+\mathrm{e}^{-x}\right)\mathrm{e}^{t}\right]\mathrm{d}t$$

(d) En effectuant le changement de variable $u = \mathrm{e}^t$, puis en utilisant la question 6, établir :

$$f_W(x) = \frac{1}{B(a,b)}\,\frac{\mathrm{e}^{-bx}}{(1+\mathrm{e}^{-x})^{a+b}}$$

9. En déduire que la variable aléatoire $Z = \dfrac{X}{Y}$ est à densité, et qu'une densité de Z est donnée par :

$$f_Z(x) = \begin{cases} \dfrac{1}{B(a,b)}\,\dfrac{x^{a-1}}{(1+x)^{a+b}} & \text{si } x > 0 \\ 0 & \text{sinon} \end{cases}$$

10. On définit sur $]0,+\infty[$ la fonction g par : $\forall x > 0,\ g(x) = \dfrac{x}{1+x}$.

(a) Montrer que g réalise une bijection de $]0,+\infty[$ dans $]0,1[$, et donner une expression de g^{-1}.

(b) On pose $T = g(Z)$. En notant F_Z et F_T les fonctions de répartition de Z et T, établir que pour tout $x \in]0,1[$, $F_T(x) = F_Z\left(g^{-1}(x)\right)$.

(c) Conclure que T suit la loi bêta de paramètres a et b.

On remarque que $T = \dfrac{X}{X+Y}$.

11. Application : simulation de la loi bêta.

Écrire une fonction `beta(a, b)` d'arguments `a` et `b` qui simule une variable aléatoire suivant la loi bêta de paramètres `a` et `b`.

Partie II. Introduction à l'inférence bayésienne

Présentation du problème : en biologie, on est souvent amené à estimer la proportion d'une population vérifiant une propriété donnée à partir d'un échantillon de cette population. On prend comme exemple la mesure d'un taux de germination d'un ensemble de graines. On plante n graines ($n \geqslant 1$) et on note, pour tout $k \in [\![1,n]\!]$, Z_k la variable aléatoire de Bernoulli de paramètre p, qui prend la valeur 1 si la k-ème graine germe, et 0 sinon. On suppose que les variables $Z_1,\ldots,Z_n$ sont indépendantes.

On note $Y_n = \displaystyle\sum_{k=1}^{n} Z_k$.

1. L'approche classique

Dans la méthode *classique*, ou *fréquentiste*, le paramètre p, même s'il est inconnu, est un réel fixé de l'intervalle $]0,1[$.

On note $F_n = \dfrac{1}{n} Y_n$ et $\sigma = \sqrt{p(1-p)}$.

12. (a) Quelle est la loi de Y_n ?

Que représente F_n ?

Calculer l'espérance et la variance de F_n.

(b) En utilisant le théorème central limite, que peut-on dire concernant la suite de variables aléatoires $(F_n)_{n\in\mathbb{N}^*}$?

On note Φ la fonction de répartition de la loi normale centrée et réduite. On rappelle que Φ réalise une bijection de $\mathbb{R}$ dans $]0,1[$. La bijection réciproque Φ^{-1} est appelée *fonction des quantiles*.

On se donne un réel α de l'intervalle $]0,1[$.

13. (a) On pose $c_\alpha = \Phi^{-1}\left(1-\dfrac{\alpha}{2}\right)$. Établir :

$$\lim_{n\to+\infty} P\left(F_n - c_\alpha \frac{\sigma}{\sqrt{n}} \leqslant p \leqslant F_n + c_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

(b) Montrer que $\sigma \leqslant \dfrac{1}{2}$.

(c) En déduire que si n est grand, on peut considérer que l'on a :

$$P\left(F_n - \frac{c_\alpha}{2\sqrt{n}} \leqslant p \leqslant F_n + \frac{c_\alpha}{2\sqrt{n}}\right) \geqslant 1-\alpha$$

On dit alors qu'on estime p par *intervalle de confiance au risque α*.

14. Un exemple : on plante 100 graines et on constate au bout de plusieurs jours que 60 d'entre elles ont germé. En utilisant la question précédente, donner une estimation par intervalle de confiance au risque 5% du taux de germination.

On donne la valeur : $\Phi(2) \simeq 0,975$.

15. On plante n graine et on représente les résultats dans une liste L de n booléens : on note `true` pour chaque graine qui germe, et `false` pour chaque graine qui ne germe pas.

Écrire une fonction Python `IC95(L)` qui prend en argument la liste L et qui renvoie sous forme de couple un intervalle de confiance au risque 5% du taux de germination.

2. L'approche bayésienne

Dans la méthode dite *bayésienne*, on ne considère plus que p est fixé : c'est une variable aléatoire à valeurs dans $[0,1]$. On la note X.

Avant de commencer les expériences, n'ayant aucune information sur p, on choisit *a priori* de poser que X suit une loi uniforme.

On considère à nouveau la variable Y_n définie à la partie précédente.

Dans toute la suite de cette partie, r est un entier fixé dans l'intervalle $[\![0,n]\!]$.

16. On suppose pour commencer que X suit une loi uniforme sur l'ensemble

$$E_N = \left\{0, \frac{1}{N}, \frac{2}{N}, \dots, \frac{N-1}{N}, 1\right\} = \left\{\frac{k}{N},\ k\in[\![0,N]\!]\right\}$$

avec N un entier naturel non nul.

À titre d'exemple, dans le cas $N = 100$, on considère que la probabilité inconnue p prend *a priori* les valeurs 0%, 1%,…,99%, 100% avec équiprobabilité.

(a) Soit $k \in [\![0,N]\!]$, reconnaitre la loi conditionnelle de Y_n sachant l'événement $\left(X = \dfrac{k}{N}\right)$.

(b) Écrire une fonction Python `simulY(N, n)` qui prend comme arguments les entiers N et n, simule X, puis renvoie une simulation de Y.

(c) Montrer : $P(Y_n = r) = \dfrac{1}{N+1}\dbinom{n}{r}\displaystyle\sum_{k=0}^{N}\left(\frac{k}{N}\right)^r\left(1-\frac{k}{N}\right)^{n-r}$.

(d) Montrer que la loi conditionnelle de X sachant $(Y_n = r)$ est donnée par :

$$\forall i \in [\![0, N]\!], \quad P_{(Y_n=r)}\left(X = \frac{i}{N}\right) = \frac{\left(\frac{i}{N}\right)^r\left(1-\frac{i}{N}\right)^{n-r}}{\sum\limits_{k=0}^{N}\left(\frac{k}{N}\right)^r\left(1-\frac{k}{N}\right)^{n-r}}.$$

17. Soit x un réel de l'intervalle $]0,1[$ tel que Nx soit entier.

(a) Établir : $P_{(Y_n=r)}(X \leqslant x) = \dfrac{\frac{1}{N}\sum\limits_{i=0}^{Nx}\left(\frac{i}{N}\right)^r\left(1-\frac{i}{N}\right)^{n-r}}{\frac{1}{N}\sum\limits_{k=0}^{N}\left(\frac{k}{N}\right)^r\left(1-\frac{k}{N}\right)^{n-r}}$.

On note C_N le numérateur et D_N le dénominateur de cette fraction.

(b) Montrer que : $\lim\limits_{N\to+\infty} D_N = B(r+1, n+1-r)$ (voir la partie I.1. pour la notation).

(c) Après avoir établi : $C_N = x^{n+1} \times \dfrac{1}{Nx}\displaystyle\sum_{i=0}^{Nx}\left(\frac{i}{Nx}\right)^r\left(\frac{1}{x}-\frac{i}{Nx}\right)^{n-r}$, montrer :

$$\lim_{N\to+\infty} C_N = x^{n+1}\int_0^1 t^r\left(\frac{1}{x}-t\right)^{n-r}\,\mathrm{d}t$$

Conclure par un changement de variable : $\lim\limits_{N\to+\infty} C_N = \displaystyle\int_0^x u^r(1-u)^{n-r}\,\mathrm{d}u$.

On admet que le calcul de cette limite reste valable que Nx soit un entier ou non.

18. Rappeler la definition de la convergence en loi d'une suite de variables aléatoires.
Justifier alors l'affirmation : sachant $(Y_n = r)$, la variable X converge en loi vers une variable aléatoire suivant la loi bêta $\beta(r+1, n+1-r)$ lorsque N tend vers $+\infty$.

19. Compte tenu de ce résultat, on propose dans la suite de modéliser la loi de X par une loi bêta de paramètres $r+1$ et $n+1-r$.

(a) Soit F_X la fonction de répartition de X. Établir que F_X réalise une bijection de $[0,1]$ dans $[0,1]$.

(b) Pour tout $\alpha \in]0,1[$, on note $u_\alpha = F_X^{-1}(\alpha)$.
Montrer que :

$$P\left(u_{\frac{\alpha}{2}} \leqslant X \leqslant u_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

On obtient ainsi une estimation de X (donc de p) par intervalle de confiance au risque α.

(c) Justifier que la fonction suivante renvoie une valeur approchée à 10^{-2} de $F_X^{-1}(x)$ pour tout argument $x \in [0,1]$.

```
def Fmoins1(n, r, x):
    y = 0
    while F(r+1, n+1-r, y) <= x:  # Voir 4(c)
        y += 1/100
    return y
```

(d) Les deux appels `Fmoins1(60, 40, 0.025)` et `Fmoins1(60,40,0.975)` renvoie respectivement 0.51 et 0.70.
Quel intervalle de confiance au risque 5% obtient-on pour p ?
Comparer avec l'intervalle de confiance obtenu par la méthode classique à la question 14.