BE de Probabilités/Statistique n° 1
Variables discrètes et continues

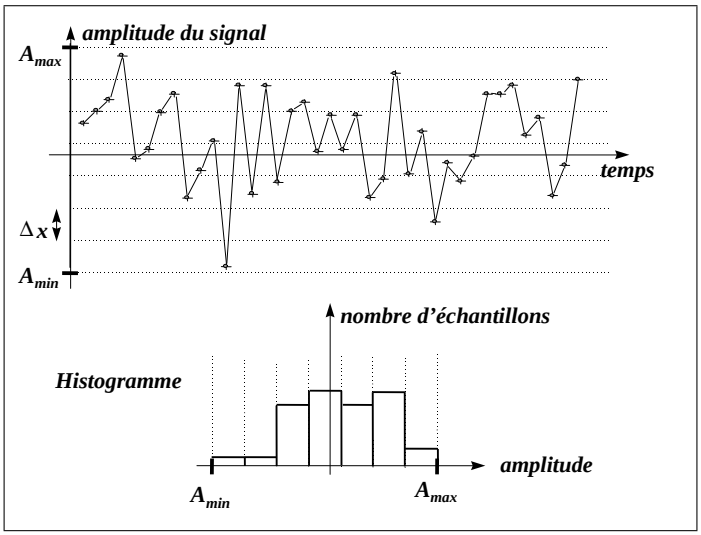
1 Variables continues

1.1 Histogramme et densité de probabilité

La probabilité pour que la variable (v.a.) continue X soit comprise entre x et $x + \Delta x$ peut être approchée pour Δx "petit" par $p(x)\Delta x$, $p(x)$ étant la densité de probabilité de la v.a. X. A partir de N réalisations de la v.a. X notées x_i, on peut donc estimer $p(x)\Delta x$ par le rapport entre le nombre de points x_i situés dans l'intervalle $[x, x + \Delta x[$ et le nombre de points total N. Etant donné N réalisations d'une v.a. X, on peut donc estimer $p(x)$ de la façon suivante : on définit des classes en divisant l'axe des y en intervalles de même largeur Δx. Notons N_x le nombre d'observations de X appartenant à la classe $[x, x + \Delta x[$ de largeur Δx et N le nombre total d'observations de X. La densité de probabilité s'approche alors de la façon suivante :

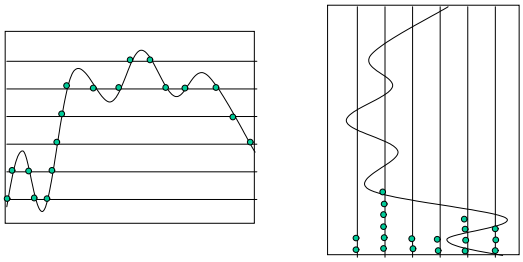
$$\forall x \in C,\ p(x) \approx \widehat{p}(x) = \frac{N_x}{N\Delta x}$$

N_x représente l'histogramme du signal (voir figure suivante). $\frac{N_x}{N\Delta x}$ est alors l'histogramme "normalisé" de X. Notons que la valeur de Δx est déterminée par la dynamique du signal et le nombre d'intervalles (ou nombre de classes).



Histogramme d'un signal échantillonné

On peut illustrer l'opération précédente de la façon suivante : considérons une fonction représentée graphiquement sur un boulier. Sur ce boulier sont placées des billes qui représentent plus ou moins finement la fonction : les tiges peuvent être assez rapprochées et les billes assez petites pour donner une impression de continuité selon les deux axes de coordonnées.



Si on place ce boulier dans le plan vertical et que ensuite, on lui fait subir une rotation de 90° dans ce plan vertical, les billes vont glisser le long des fils et on aura le nombre de billes sur chaque fil, image (grossière bien sûr) de la densité de probabilité. C'est exactement l'opération que l'on réalise lorsqu'on fait l'histogramme d'amplitude. C'est ce que nous nous proposons de faire dans ce TP : utiliser l'histogramme pour estimer la densité de probabilité d'une variable aléatoire.

1.2 Fonction de répartition

La fonction de répartition F de la v.a. X est définie de la façon suivante :

$$F(x) = P\left[X < x\right] = \int_{-\infty}^{x} p(u)du$$

Pour estimer $F(x)$, il suffit de calculer la somme cumulée de l'histogramme (normalisé) c'est-à-dire compter le nombre de réalisations de x_i inférieures à x et de diviser par le nombre total de points N. On a alors l'expression suivante :

$$F(x) \approx \sum \widehat{p}(x)\Delta x = \sum \frac{N_x}{N}$$

les sommations étant effectuées sur les classes d'amplitude inférieures à x.

1.3 Travail à réaliser

Générer à l'aide de la fonction `randn` un vecteur gaussien x de N points (prendre $N = 1000$), de moyenne 1 et de variance 0.1. Afficher l'histogramme de ce signal (commande `hist`) avec $N_C = 50$ classes. Pour vérifier le bon comportement de l'histogramme, on veut superposer la densité de probabilité théorique. Pour cela, définir un vecteur l contenant N_C points entre $\min(x)$ et $\max(x)$.Calculer alors la densité théorique aux points l à l'aide de la fonction `normpdf`. **Attention** : pour pouvoir comparer ces deux courbes, il faut penser à normaliser l'histogramme.

Visualiser l'effet sur l'histogramme de l'augmentation :

- du nombre de classes N_C pour un nombre de points N fixe (prendre $N = 5000$ et N_C variant de 10 à 100 avec un pas de 10) ;
- du nombre de points N pour un nombre de classes N_C fixe (prendre $N_C = 50$ et N variant de 1000 à 4000 avec un pas de 200).

Tracer l'histogramme d'un signal de loi uniforme sur l'intervalle $[1, 2]$ et comparer cet histogramme avec la densité de probabilité correspondante.

Générer un signal binaire de N points tel que $x_n = 1$ ou $x_n = -1$ (on peut utiliser les fonctions "`sign`" et "`randn`" ou la méthode de genération de bits vue précédemment). Visualiser son histogramme et expliquer. On perturbe ce signal par un *bruit* gaussien. Pour cela, générer un signal gaussien b de moyenne nulle et de variance $\sigma^2 = 0.1$, et l'ajouter au signal binaire x. Observer l'évolution de l'histogramme lorsque σ^2 varie. Commenter. Même question avec un bruit de loi uniforme sur l'intervalle $[-1; 1]$.

Générer une sinusoïde de fréquence $f = 0.125$ avec $N = 1000$ points, c'est-à-dire

$$x(n) = \sin(2\pi fn) \text{ pour } n = 0, ..., N-1$$

Observer l'histogramme de cette sinusoïde et commenter. Même question avec une fréquence $f = 0.1277$. Commentaires.

Bruiter comme précédemment cette sinusoïde et tracer l'histogramme du signal obtenu. Augmenter la variance. Que constate-t-on ?

2 Variables discrètes

2.1 Fréquence relative

Lorsque le nombre de résultats d'expérience est fini (égal à N_R), la probabilité d'un évènement A se définit généralement comme

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

où $|A|$ représente le nombre d'éléments de l'évènement A, c'est-à-dire que les évènements élémentaires sont considérés comme équiprobables. Ainsi, la probabilité $P(A)$ se calcule par dénombrement. En pratique, pour estimer le probabilité $P(A)$, on génère N évènements, et on compte le nombre N_{CF} de cas favorables parmi ces évènements. On peut alors approcher $P(A)$ par :

$$P(A) \approx \frac{N_{CF}}{N}$$

2.2 Travail à réaliser

2.2.1 ex. 1 : variables de Bernoulli

Une variable aléatoire discrète X qui prend les valeurs 0 et 1 avec les probabilités

$$\begin{aligned} P[X = 0] &= 1 - p \\ P[X = 1] &= p \end{aligned}$$

est appelée *variable de Bernoulli de paramètre* p (avec $0 < p < 1$). On a alors $E[X] = p$ et $var(X) = p(1-p)$. Générer une séquence de variables de Bernoulli à l'aide de la fonction `rand`. Vérifier les résultats théoriques concernant la moyenne et la variance de X grâce aux commandes `mean` et `var`.

2.2.2 ex. 2 : variables binômiales

La somme de n variables de Bernoulli de paramètre p est appelée *variable binômiale* $\mathcal{B}(n, p)$. On a alors $E[X] = np$ et $var(X) = np(1-p)$. Cette loi permet de modéliser le tirage *avec remise* de n boules dans une urne contenant des boules blanches et des boules noires en proportion p et $1-p$. La probabilité d'avoir k boules blanches est alors donnée par :

$$P[X = k] = C_n^k p^k (1-p)^{n-k} \qquad k \in \{0, ..., n\}$$

A l'aide de la question précédente, générer des variables binômiales. Vérifier alors les résultats théoriques (moyenne, variance, probabilités).

2.2.3 ex. 3 : générateur de nombres aléatoires

Un générateur de nombres aléatoires produit des séquences de triplets (indépendants) d'entiers équiprobables (i, j, k) avec $i, j, k \in \{1, \ldots, 10\}$. On veut estimer par simulations :

1. la probabilité que ce générateur produise un triplet de la forme (i, i, i) ;
2. la probabilité qu'au moins un triplet de la forme (i, i, i) apparaisse dans une séquence de 10 triplets ;
3. la probabilité d'avoir exactement 4 triplets de la forme (i, i, i) dans une séquence de 20 triplets.

Générer sous Matlab de telles séquences de triplets, et estimer les probabilités ci-dessus en calculant la fréquence relative (on pourra utiliser les fonctions `unidrnd`, `find`, `mean`, `sum`). Comparer avec les probabilités théoriques (respectivement : 0.01, 0.0956, $4.125e^{-5}$). Pour les questions 2 et 3, on utilisera la fonction `hist` pour pouvoir compter le nombre de triplets équiprobables dans des séries de 10 ou 20 triplets

Observer la précision des probabilités estimées en fonction du nombre de tirages.

3 Compléments : génération de variables discrètes

Dans un grand nombre d'applications, on peut être amené à générer des variables discrètes qui ne suivent pas une loi standard, comme la loi uniforme ou la loi de Poisson.

On cherche dans cet exercice à créer une fonction `variables_discretes` permettant de renvoyer une matrice $N \times M$ d'échantillons générés indépendamment suivant une loi définie par

$$P[X = x_k] = p_k \ , \quad k = 1, \ldots, K$$

où les valeurs $(x_k)_{k=1,\ldots,K}$ et les probabilités $(p_k)_{k=1,\ldots,K}$ sont spécifiées par l'utilisateur. Les variables d'entrée de la fonction `variables_discretes` sont donc :

- le vecteur $x = [x_1, \ldots, x_K]^T$ des valeurs prises par la variable aléatoire X ;
- le vecteur $p = [p_1, \ldots, p_K]^T$ des probabilités associées aux valeurs $(x_k)_{k=1,\ldots,K}$;
- les dimensions N et M de la matrice à générer.

Une méthode consiste à créer un vecteur $y = (y_1, y_2, \ldots, y_{NM})$ de longueur NM, où les y_i sont tirés suivant une loi uniforme sur $[0; 1]$ (variables continues). A partir du vecteur y, on génère le vecteur des réalisations $X = (X_1, X_2, \ldots, X_{NM})$ en posant

$$X_i = x_k \ \text{ si } \ y_i \in [P_{k-1}; P_k[,$$

où $P_k = \sum_{l=1}^{k} p_l$ pour $k = 1, \ldots, K$ et $P_0 = 0$.

1. Vérifier que pour tout i, on a bien $P[X_i = x_k] = p_k$ (dans le cas où on a bien $\sum_{l=1}^{K} p_l = 1$).
2. Dans le cas où $\sum_{l=1}^{K} p_l \neq 1$, c'est-à-dire que l'utilisateur n'indique pas les probabilités absolues, mais des "poids" p_k (supposés toujours positifs) aux différentes valeurs x_k, comment peut-on procéder ?
3. Créer la fonction `variables_discretes` de variables d'entrée x, p, N et M da la façon suivante :
   - Construire le vecteur $P = (P_0, P_1, \ldots, P_K)$ défini précédemment (on pourra utiliser la fonction `cumsum`).
   - Générer le vecteur y, et construire à partir de y et de P le vecteur X. Construire alors la matrice souhaitée.
   - Retourner alors cette matrice, ainsi que la moyenne et la variance de la loi discrète utilisée.
4. Créer un fichier `test.m` permettant de tester la fonction ainsi créer sur des lois discrètes standard (comme la loi uniforme, la loi binômiale, ou la loi de Poisson) en précisant la méthode ainsi que les paramètres utilisés pour valider la fonction.