Annales de compilation

VIAL Sébastien

11 septembre 2024

A) Modélisation de densités de probabilité (6 pts)
Soit les 3 distributions $H1(x)$, $H2(x)$ et $H3(x)$, toutes composées de 100 événements, représentant les occurrences pour x compris entre 0 et 31 :

$$H1(x) = \{3, 2, 3, 2, 4, 3, 2, 4, 2, 2, 3, 4, 2, 3, 4, 3, 5, 3, 4, 4, 3, 3, 5, 3, 3, 3, 4, 3, 3, 3, 2, 3\}$$

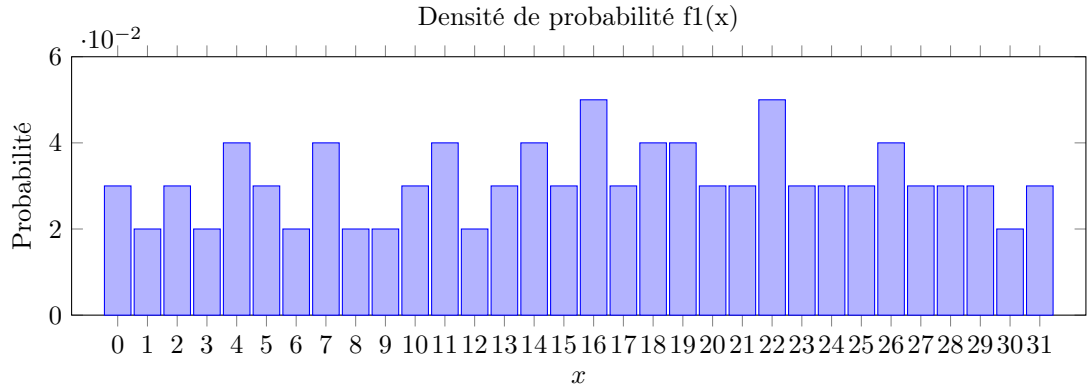
$$H2(x) = \{0, 0, 0, 0, 0, 0, 1, 2, 3, 6, 18, 40, 18, 6, 3, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$$

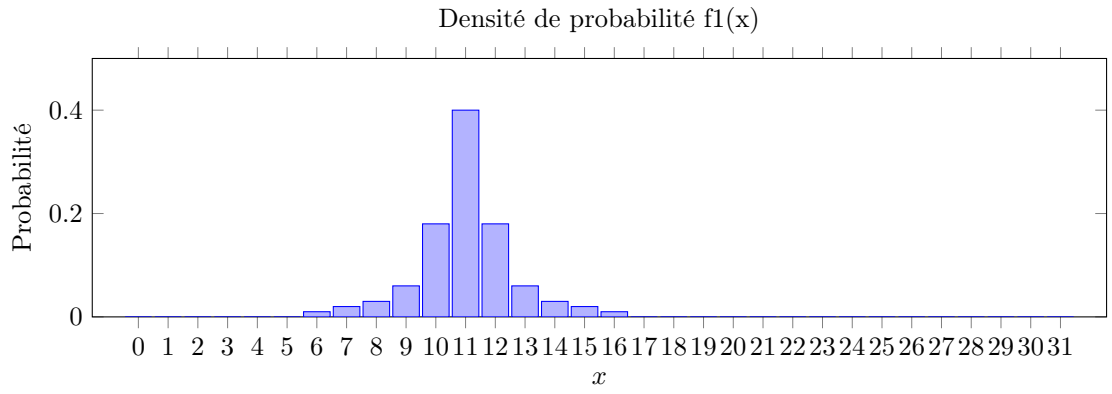
$$H3(x) = \{0, 0, 0, 0, 0, 1, 4, 6, 10, 17, 20, 17, 10, 6, 4, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$$

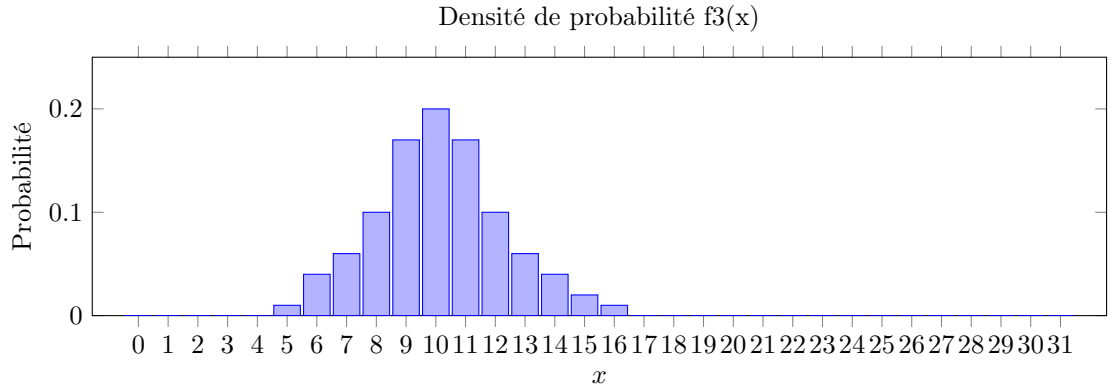
1. À partir de ces 3 distributions, calculer et tracer les densités de probabilité (ddp) correspondantes $f1(x)$, $f2(x)$ et $f3(x)$.
2. À partir de la fonction générique $f(x) = C\exp(-K|x-\mu|^{\alpha})$ en déduire la valeur de α la plus pertinente pour chacune des ddp $f1(x)$, $f2(x)$ et $f3(x)$. Pour rappel, $\alpha = 0$: distribution uniforme ; $\alpha = 1$: distribution Laplacienne ; $\alpha = 2$: distribution Gaussienne.
3. Approcher la ddp $f2(x)$ par une distribution Gaussienne en calculant $moy2$ et $sigm2$, correspondant à la valeur moyenne et l'écart type (donner la formule obtenue). Tracer cette distribution sur la courbe représentant $f2(x)$.
4. Approcher la ddp $f3(x)$ par une distribution Gaussienne en calculant $moy3$ et $sigm3$, correspondant à la valeur moyenne et l'écart type (donner la formule obtenue). Tracer cette distribution sur la courbe représentant $f3(x)$.
5. Qu'en déduisez-vous ? Comment mesurer la distance entre une ddp et la distribution Gaussienne qui l'approche ?

1.

$$f_1(x) = \{0.03, 0.02, 0.03, 0.02, 0.04, 0.03, 0.02, 0.04, 0.02, 0.02, 0.03, 0.04, 0.02, 0.03, 0.04,$$
$$0.03, 0.05, 0.03, 0.04, 0.04, 0.03, 0.03, 0.05, 0.03, 0.03, 0.03, 0.04, 0.03, 0.03, 0.03, 0.02, 0.03\}$$
$$f_2(x) = \{0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.01, 0.02, 0.03, 0.06, 0.18, 0.40, 0.18, 0.06, 0.03, 0.02,$$
$$0.01, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00\}$$
$$f_3(x) = \{0.00, 0.00, 0.00, 0.00, 0.00, 0.01, 0.04, 0.06, 0.10, 0.17, 0.20, 0.17, 0.10, 0.06, 0.04,$$
$$0.02, 0.01, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00\}$$







2. — **Uniforme :** La distribution uniforme est caractérisée par une probabilité constante pour toutes les valeurs possibles de la variable aléatoire dans un certain intervalle. La fonction de densité de probabilité pour une distribution uniforme continue sur l'intervalle [a, b] est définie comme suit : $f(x) = \frac{1}{b-a}$ pour $a \leq x \leq b$, et $f(x) = 0$ sinon. Par exemple, un dé à six faces non truqué peut être modélisé par une distribution uniforme sur l'ensemble $\{1, 2, 3, 4, 5, 6\}$.
   — **Laplacienne :** La distribution laplacienne est également connue sous le nom de distribution à double exponentielle. Elle est caractérisée par sa forme en cloche avec des queues lourdes. La fonction de densité de probabilité pour une distribution laplacienne avec moyenne μ et écart-type β est donnée par : $f(x; \mu, \beta) = \frac{1}{2\beta} \exp\left(-\frac{|x-\mu|}{\beta}\right)$. La distribution laplacienne est souvent utilisée en traitement du signal et en statistique bayésienne.
   — **Gaussienne (ou Normale) :** La distribution gaussienne, également appelée distribution normale, est l'une des distributions les plus couramment rencontrées. Elle est caractérisée par une cloche symétrique autour de sa moyenne. La fonction de densité de probabilité pour une distribution normale avec moyenne μ et écart-type σ est donnée par : $f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. La distribution normale apparaît naturellement dans de nombreux phénomènes du monde réel en

raison du théorème central limite.

Pour la courbe f_1, on va choisir $\alpha = 0$ car on ne remarque pas de pic. Pour la courbe f_2, $\alpha = 1$ car on remarque un pic "brutal". Pour la courbe f_3, nous choisirons $\alpha = 2$.

3. Pour la distribution $f2(x)$, nous approchons la densité de probabilité par une distribution gaussienne. Les paramètres de cette distribution gaussienne, la moyenne (notée $moy2$) et l'écart type (noté $sigm2$), sont calculés comme suit :

   La moyenne est calculée par la formule :

$$moy2 = (\sum_{i=0}^{31} x_i \times f2(x_i))/$$

où x_i sont les valeurs de x (de 0 à 31) et $f2(x_i)$ sont les probabilités associées à ces valeurs.

L'écart type est calculé en utilisant la formule :

$$sigm2 = \sqrt{\sum_{i=0}^{31}(x_i - moy2)^2 \times f2(x_i)}$$

Après calcul, nous obtenons :

— Moyenne (notée $moy2$) : 11.0
— Écart type (noté $sigm2$) : 1.59

Ces paramètres nous permettent de définir la distribution gaussienne approchée pour $f2(x)$ comme suit :

La loi de Gauss ou loi normale $\mathcal{N}(\mu, \sigma^2)$

C'est la loi de probabilité fondamentale de la statistique, en raison du Théorème central limite.
La variable aléatoire X suit une loi de Gauss, ou loi normale, notée $\mathcal{N}(\mu, \sigma^2)$ si elle a pour densité définie sur $\mathbb{R}$:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

$E(X) = \mu$
$Var(X) = \sigma^2$

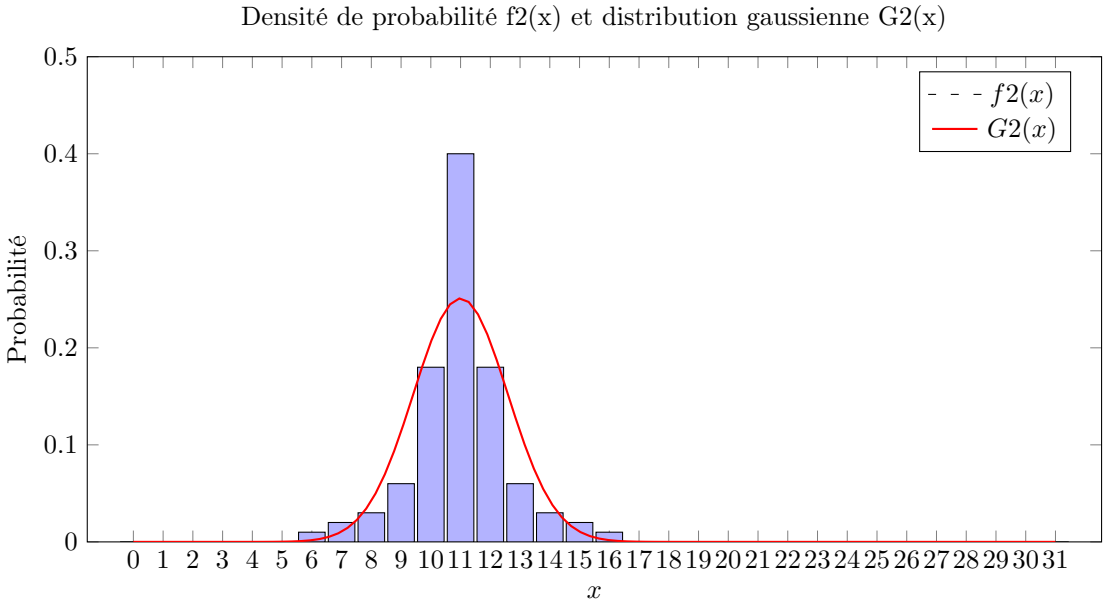
Pour $\mu = 0$ et $\sigma^2 = 1$, on parle de loi de Gauss centrée réduite $\mathcal{N}(0, 1)$, et sa fonction de répartition est

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz.$$

$F_X(x) =$

$$G2(x) = \frac{1}{1.59\sqrt{2\pi}} \exp\left(-\frac{(x-11)^2}{2 \times 1.59^2}\right)$$

où $G2(x)$ est la densité de probabilité de la distribution gaussienne approchée pour $f2(x)$.

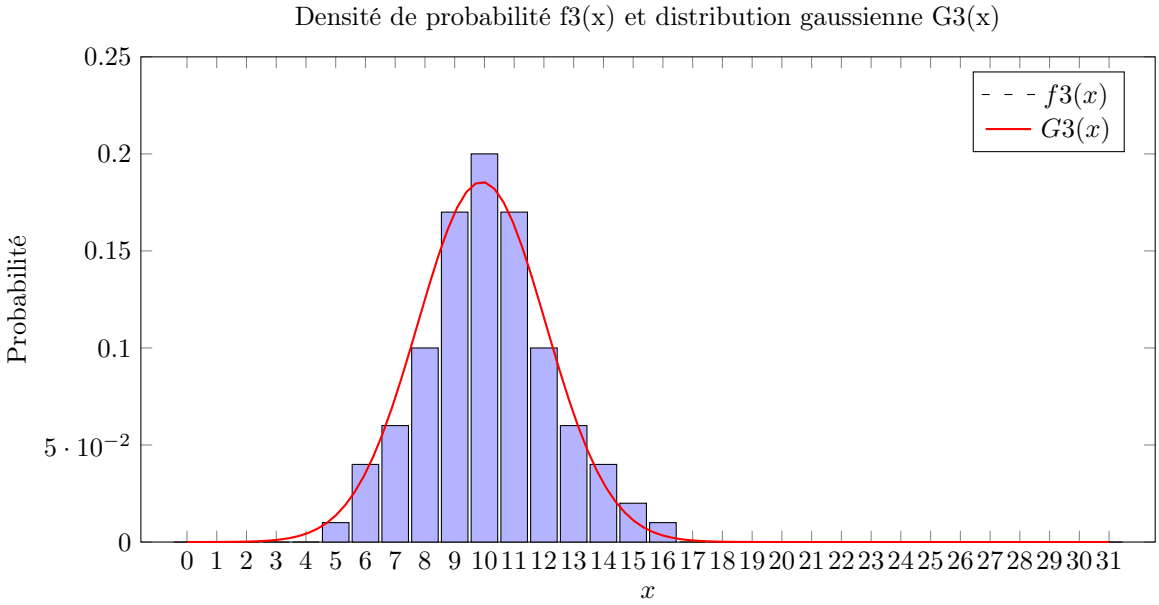


4. Pour la distribution $f3(x)$, nous avons approché la densité de probabilité par une distribution gaussienne. Les paramètres calculés de cette distribution gaussienne sont :
   - Moyenne (notée $moy3$) : 9.91
   - Écart type (noté $sigm3$) : 2.15

   Ces paramètres nous permettent de définir la distribution gaussienne approchée pour $f3(x)$ comme suit :

$$G3(x) = \frac{1}{2.15\sqrt{2\pi}} \exp\left(-\frac{(x-9.91)^2}{2 \times 2.15^2}\right)$$

   où $G3(x)$ est la densité de probabilité de la distribution gaussienne approchée pour $f3(x)$.



5. Pour mesurer la distance ou la différence entre la densité de probabilité (ddp) empirique et la distribution gaussienne approximative, plusieurs méthodes peuvent être utilisées :

   (a) **Erreur Quadratique Moyenne (EQM)** : Elle est calculée comme la somme des carrés des différences entre les valeurs de probabilité observées et celles prédites par la distribution gaussienne,

divisée par le nombre de points.

$$EQM = \frac{1}{n}\sum_{i=0}^{n-1}(f(x_i) - G(x_i))^2$$

où $f(x_i)$ est la valeur de la ddp observée et $G(x_i)$ est la valeur de la distribution gaussienne pour chaque x_i.

(b) **Distance de Kullback-Leibler (Divergence KL)** : Cette mesure quantifie la quantité d'information perdue lorsqu'on utilise la distribution gaussienne pour approximer la distribution réelle.

$$D_{KL}(f||G) = \sum_i f(x_i)\log\left(\frac{f(x_i)}{G(x_i)}\right)$$

où $f(x_i)$ est la probabilité dans la ddp empirique et $G(x_i)$ dans la distribution gaussienne.

(c) **Test de Kolmogorov-Smirnov** : Ce test non paramétrique mesure la distance maximale entre les fonctions de répartition cumulée des deux distributions.

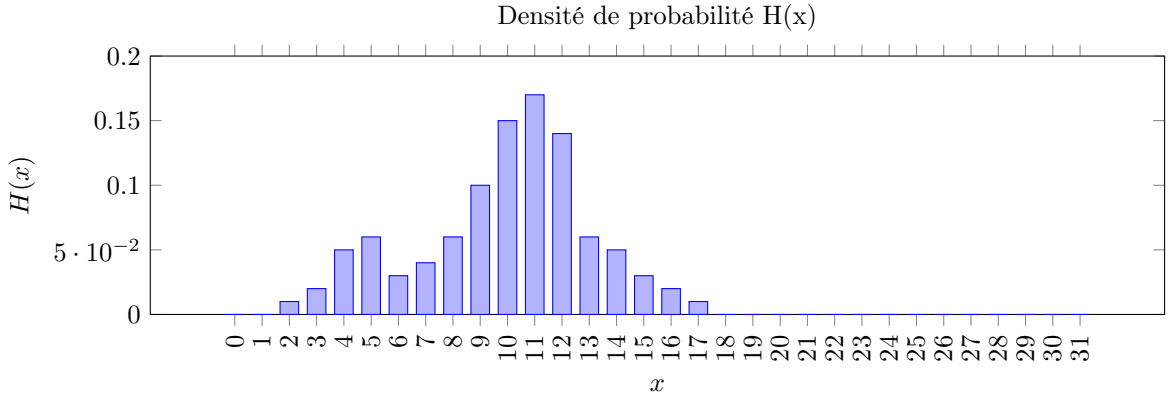
Chaque méthode a ses propres avantages et convient mieux à différents types d'analyses. Par exemple, l'EQM est simple et intuitive, mais peut être excessivement influencée par des valeurs aberrantes. La divergence KL fournit une mesure basée sur l'information, et le test de Kolmogorov-Smirnov est utile pour tester l'adéquation de la distribution sans supposer la normalité.

Mélange de 2 Gaussiennes (4 pts)
Soit la densité de probabilité (ddp) suivante $f(x)$ représentant les probabilités pour x compris entre 0 et 31 :

$$H(x) = \{0, 0, 0.01, 0.02, 0.05, 0.06, 0.03, 0.04, 0.06, 0.1, 0.15, 0.17, 0.14, 0.06,$$
$$0.05, 0.03, 0.02, 0.01, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$$

1. Tracer $f(x)$.
2. En considérant que $f(x)$ est un mélange de 2 gaussiennes, indiquer la valeur des 2 modes (valeurs maximales) et proposer une valeur de seuil séparant les 2 modes.
3. Pour chacun des modes, calculer la valeur moyenne et l'écart type μ_1, σ_1, μ_2, σ_2.
4. Par rapport aux probabilités par mode (β_1 et β_2), proposer un modèle de mélange de 2 Gaussiennes - paramètres à introduire μ_1, σ_1, μ_2, σ_2 et β_1 (sachant que $\beta_2 = 1 - \beta_1$).

1.



2. Les deux pics se trouvent respectivement à $x = 5$ et $x = 11$. Afin de séparer ces deux pics, nous cherchons la probabilité la plus basse entre ces deux pics, le "creux". Ici ça sera $x = 6$.
3. **Calculs pour le premier mode :**

La moyenne μ_1 est la moyenne pondérée des valeurs de x :

$$\mu_1 = \frac{\sum_{i=0}^{6} x_i \times H(x_i)}{\sum_{i=0}^{6} H(x_i)}$$

L'écart type σ_1 est la racine carrée de la moyenne pondérée des carrés des écarts de chaque valeur de x par rapport à μ_1 :

$$\sigma_1 = \sqrt{\frac{\sum_{i=0}^{6} (x_i - \mu_1)^2 \times H(x_i)}{\sum_{i=0}^{6} H(x_i)}}$$

Calculs pour le second mode :

La moyenne μ_2 est calculée de façon similaire au premier mode :

$$\mu_2 = \frac{\sum_{i=7}^{31} x_i \times H(x_i)}{\sum_{i=7}^{31} H(x_i)}$$

L'écart type σ_2 est également calculé de façon similaire :

$$\sigma_2 = \sqrt{\frac{\sum_{i=7}^{31} (x_i - \mu_2)^2 \times H(x_i)}{\sum_{i=7}^{31} H(x_i)}}$$

En effectuant ces calculs, nous obtenons les résultats suivants pour les deux modes :

— Pour le premier mode :
  — Moyenne (notée μ_1) : 4.47
  — Écart type (noté σ_1) : 1.09
— Pour le second mode :
  — Moyenne (notée μ_2) : 11.00
  — Écart type (noté σ_2) : 2.17

Ces valeurs seront utilisées pour modéliser chaque mode comme une distribution gaussienne distincte.

Densité de probabilité $H(x)$ avec deux gaussiennes superposées