**Exercice 1 - Variable continu (40%)**

Soit $x$ un variable continue (pour example un mesure des revenu en kilo€ des ménages françaises), avec $x \in [0, 100]$. En Figure 1 vous trouvez l'histogramme qui représente la fonction de densité (pdf) de $x$ ci-dessous (important, cette histogramme divise l'échantillon en tranches inégales [1]) :

$$f(x) \quad = \quad \begin{cases} c & \text{si } 0 \leq x < \beta \\ a & \text{si } \beta \leq x \leq 100 \\ 0 & \text{autrement} \end{cases}$$

En plus, on connais que $\beta = 25$ et que dois être toujours vérifié que $c \;=\; 2 \cdot a$. On utilisant ces importantes informations, répondez à les questions suivantes :

1. Si on a un échantillon d'observations de $x$, quelle hypothèse est utilisée pour construire l'histogramme (et le graph de $f(x)$) de $x$ ?
2. Utilisez les donnes de l'exercice et les propriété de la pdf pour démontrer que $a = 0.008$ et donc $c = ?$.
3. Avec les valeurs de $a$ et $c$ dans votre pdf, calculez la moyenne arithmétique ($\mu_1(x)$) et la moyenne géométrique ($\mu_0(x)$).
4. Calculez la variance $\sigma^2(x)$.
5. Calculez la médiane et le quintile inférieur et supérieur.
6. Quelle est le plus petite valeur modale ?
7. Calculez, écrivez-là et représentez la fonction de répartition (cdf) $F(x)$.
8. ★ Dans cette exercice on travaille avec la Courbe de Lorenz, [2] que nous décris la distribution des revenu. Pour tous les niveaux $p$ que représente un tranche du échantillon (avec $p \in [0, 1]$) la Courbe de Lorenz nous donne la parte des revenu $L$ (avec $L \in [0, 1]$) possédé par la tranche $p$ de l'échantillon. Pour exemple, Pareto a noté que à son temps en Italie le 20% de la population la *plus riche* possédé le 80% des revenu totales (ainsi la règle du "

1. Ici un vite explication : `http ://en.wikipedia.org/wiki/Histogram` .
2. Voir `http ://en.wikipedia.org/wiki/Lorenz_curve`.

80-20 "[3]). Dans cette exercice, detérminez la relation $x = F^{-1}(p)$ et vérifiez que (utilisez la soluction de la question 7) :

$$\int_0^1 F^{-1}(p)\ dp \ = \ \mu_1(x).$$

9. ★ Calculez le valeur de la Courbe de Lorenz pour le quintile supérieure ($p = 0.8$), sachant que :

$$L(0.8) \ = \ \frac{\int_0^{0.8} F^{-1}(p)\ dp}{\mu_1(x)}.$$

Est-ce que la règle de Pareto (80-20) est vrai aussi avec votre distribution (et donc dois être vrai que le 80% de la population la *plus pauvre* possédé le 20% de la richesse) ? Et $L(1) =?$. Est-ce que le on peux utiliser le quintile supérieur divisé par le valeur maximum de $x$ comme un mesure de l'inégalité que vous permet de comparer le résultat avec la règle de Pareto ? Pourquoi ? Et si la distribution de la $x$ etait uniforme ?

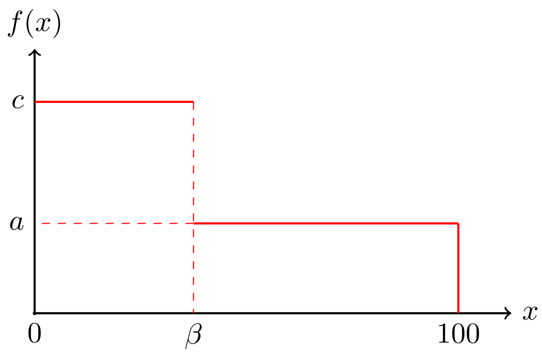


FIGURE 1 – Exercice 1 : l'histogramme que réprésente $f(x)$.

**Exercice 2 - pdf et cdf (10%)**

En utilisant le graphique de la pdf de $x$ ($f(x)$, en rouge) représenté en Figure 2, représenter dans un autre graphique la cdf de $x$, $F(x)$. Sur votre graphique représentez les points $a$, $b$, $c$ et ensuite la mode globale.
Ensuite, représentez aussi la pdf et la cdf de la densité suivante :

$$f(x) \quad = \quad \begin{cases} 0 & \text{si } x < c \\ 0.2 & \text{si } c \leq x \leq b \\ 0 & \text{si} b < x < a \\ N(a, \sigma^2) & \text{si} x \geq a. \end{cases}$$

3. Voir pour exemple : `http ://en.wikipedia.org/wiki/Pareto_principle`.

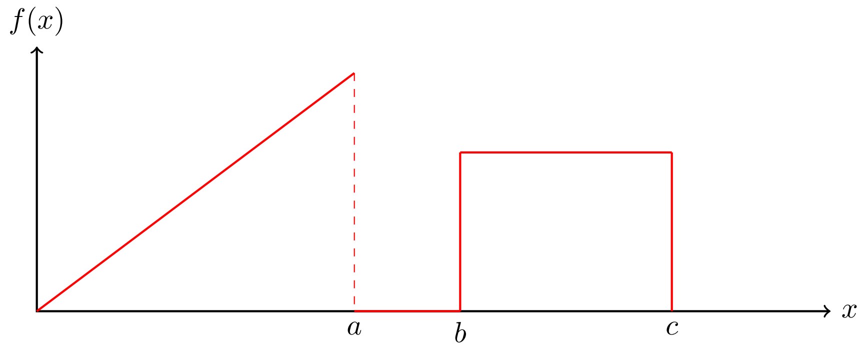


Figure 2 – Exercice 2 : $f(x)$ en rouge.

**Exercice 3 - Statistique bivariée (50%)**

Vous avez un échantillon de 10 observations de revenu en kilo€ ($y$) et de ans de université après le bac ($x$). Chaque observation correspond à un individu enqueté. Les donnés sont :

| Observ. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 30 | 30 | 50 | 70 | 50 | 50 | 70 | 30 | 70 | 50 |
| $x$ | 2 | 1 | 2 | 3 | 2 | 1 | 2 | 1 | 3 | 3 |

On considère $y$ et $x$ comme variables discrètes et donc on utilise les définition de moyenne et variance pour variables discrètes. Ici, $y$ a seulement 3 valeur possibles : $y \in \{30, 50, 70\}$ ; autant que $x$ : $x \in \{1, 2, 3\}$. Il est nécessaire d'étudier les pages 58/61 pour bien comprendre comme travailler avec de sous populations. Ensuit, étudiez les définitions de covariance et corrélation et répondez à les questions suivantes.

1. Compilez le tableau suivant avec les effectifs approprié :

| | | $x$ | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | Tot |
| $y$ | 30 | $n_{30,1}$ | $n_{30,2}$ | $n_{30,3}$ | $n_{30}$ |
| | 50 | $n_{50,1}$ | $n_{50,2}$ | $n_{50,3}$ | $n_{50}$ |
| | 70 | $n_{70,1}$ | $n_{70,2}$ | $n_{70,3}$ | $n_{70}$ |
| | Tot | $n_1$ | $n_2$ | $n_3$ | $n$ |

2. En utilisant cette donnes, calculez un système de poids pour les revenu $y$, *conditionnelles* à un niveau d'études après le bac $x$. Ensuite , représentez cette poids dans un tableau similaire a lequel de la question précédente. Vérifiez que , pour chaque colonne, les conditionnes sur les poids ($\sum_i w_{i,j} = 1$ et $w_{i,j} \geq 0$) sont satisfaites.
3. Calculez la moyenne arithmétique de $y$ conditionnelle a chaque valeur de $x$ , $\mu_1(y|x)$. Pour exemple, $\mu_1(y|\text{un an après bac}) = \mu_1(y|1)$. Vérifiez que $\mu_1(y|1) = 36.66$, $\mu_1(y|2) = 50$ et $\mu_1(y|3) = 63.33$
4. Calculez la moyenne arithmétique de $y$, $\mu_1(y)$ et montrer que peut être écrit comme $\mu_1(y) = \alpha\mu_1(y|1) + \beta\mu_1(y|2) + \gamma\mu_1(y|3)$ et spécifiez les poids $\alpha, \beta, \gamma$ nécessaires pour obtenir le résultat. Ensuite, démontre votre résultat avec les valeurs.
5. Calculez l'écart type de $y$ conditionnelle a chaque valeur de $x$ , $\sigma(y|x)$. Pour exemple, $\sigma(y|\text{un an après bac}) = \sigma(y|1)$. Vérifiez que $\sigma(y|1) = 11.54$, $\sigma(y|2) = 16.33$ et $\sigma(y|3) = 11.54$. Ensuite, calculez la décomposition de la variance.

6. Calculez $\mu_1(x)$, et ensuite $\sigma(x)$. En utilisant les définitions corrects, calculez la covariance $\text{Cov}(y, x)$ et le coefficient de corrélation linaire $\rho(y, x)$ et montrez que $\text{Cov}(y, x) = 8.88$ et $\rho(y, x) = 0.66$.
7. Représentez dans l'espace en Figure 3 tous les observations de l'échantillon avec des points. Si certaines couples de $(y, x)$ sont observé plusieurs fois, représentez-les avec des points plus marqués. Ensuite, représentez dans le même espace avec des points rouge les trois niveaux de les moyennes conditionnelles $\mu_1(y|1)$, $\mu_1(y|2)$ et $\mu_1(y|3)$ en correspondence aux niveaux de $x$.
8. ★ En Figure 3, croyez vous que on peux représenter la relation entre moyenne conditionnelle de $y$ et $x$ avec un modèle linaire $\mu_1(y|x) = \alpha + \beta \cdot x$ ? Justifiez (*Hint :* vérifiez pour exemple que la pante de cette droite est constante entre différent observations). Cette modèle est le modèle de régression linaire, que impose une relation linaire entre les moyenne conditionnelles et la variable $x$.
9. Calculez les paramètres de cette droite en utilisant deux points ou elle passe. Ensuite, vérifiez que $\beta = \frac{\text{Cov}(y,x)}{\sigma^2(x)} = 13.33$ et que $\alpha = \mu_1(y) - \beta \cdot \mu_1(x) = 23.33$. Représentez la courbe $\mu_1(y|x) = \alpha + \beta \cdot x$ sur le graphique en Fig 3.
10. ★ $\frac{d\mu_1(y|x)}{dx}$ =? Que est le valeur de $\mu_1(y|1.5)$ ? Utilisez le modèle linaire pour le calculer. En prévision, de combien de euro vont augmenter les revenu des individu qui ont choisit l'option bac+3 ?
11. Créez des nouvelles variables $y'$ et $x'$ dans votre tableaux initiale des donnes, ou $y' = \frac{y}{\sigma(y)}$ et $x' = \frac{x}{\sigma(x)}$. Sans faire des calcules, en utilisant les règles de la variance démontrez que $\sigma(x') = \sigma(y') = 1$, ou $\sigma$ est l'écart type.
12. Calculez le valeur de $\beta'$ en un modèle pour la moyenne arithmétique conditionnelle de $y'$ linaire $\mu_1(y'|x') = \alpha' + \beta' \cdot x'$, en utilisant $\beta' = \frac{\text{Cov}(y',x')}{\sigma^2(x')}$ et démontrez que $\beta' = \rho(y, x)$.
13. ★ Si la vraie relation entre $\mu_1(y'|x')$ et $x'$ n'est pas linaire mais quadratique, expliquez pourquoi il serait assez probable que $\beta' \approx 0$.

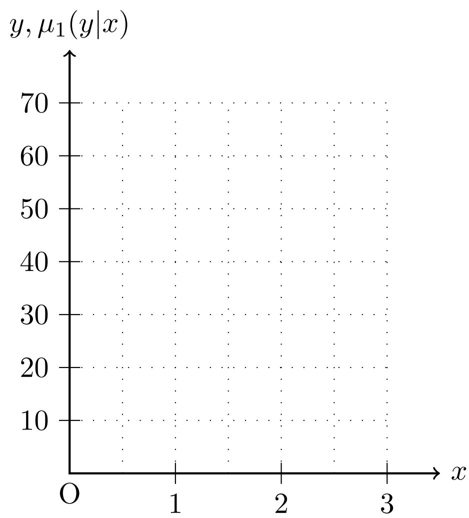


FIGURE 3 – Exercice 3 : Observations, moyennes conditionnelles et regréssion.