Université de Strasbourg — Année 2010/2011
L3 - Statistique Mathématique
TD : Myriam Maumy-Bertrand — mmaumy@math.unistra.fr
Cours : Ségolen Geffray — geffray@math.unistra.fr

---

**TD 0 : Pour ceux qui n'auraient jamais entendu parler de variable aléatoire**

---

**Exercice 1.**
Soit $X$ une variable aléatoire de domaine de définition $X(\Omega) = \{0; 1; 2; 3; 4\}$ et dont la fonction de densité est donnée pour $x \in X(\Omega)$ par

$$f_X(x) = \frac{2x+1}{25}\,.$$

1. Déterminer puis tracer la fonction de répartition $F_X$ de $X$.
2. Calculer les probabilités suivantes : $\mathbb{P}[X = 4]$, $\mathbb{P}[X \geq -2]$, $\mathbb{P}[X \leqslant 1]$ et $\mathbb{P}[2 \leqslant X < 4]$.
3. Déterminer les réels $a$ vérifiant $\mathbb{P}[X \leqslant a] \leqslant 1/5$ puis $\mathbb{P}[X \geqslant a] \leqslant 1/5$.
4. Déterminer la médiane de la variable $X$.

**Exercice 2.**
Soit $X$ une variable aléatoire de fonction de répartition $F_X$ tracée sur la figure 1.

1. Quel est l'ensemble des valeurs prises par $X$ ?
2. Quelle est la fonction de densité de $X$ ?
3. Calculer $\mathbb{P}[X \leqslant -2]$, $\mathbb{P}[X > 6]$, $\mathbb{P}[X \leqslant 3]$, $\mathbb{P}[X \leqslant 4]$, $\mathbb{P}[X < 2]$ et $\mathbb{P}[1 \leqslant X \leqslant 4]$.
4. Déterminer les réels $a$ vérifiant $\mathbb{P}[X \leqslant a] \geqslant 1/4$ puis $\mathbb{P}[X \geqslant a] \geqslant 1/4$.
5. Déterminer la médiane de la variable $X$.

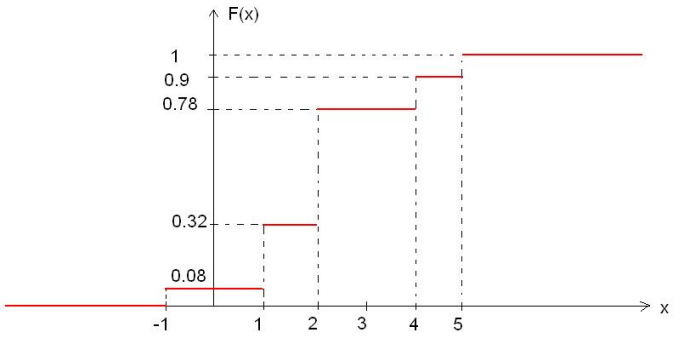



Figure 1 – Fonction de répartition de $X$

**Exercice 3.**
Soit une variable aléatoire $X$ définie sur $[0; 2]$ régie par la densité suivante : $f_X(x) = 3(4-x^2)/16$.

1. Quel est le domaine de définition $X(\Omega)$ de $X$ ?

2. Tracer la fonction de densité $f_X$ de $X$.
3. Déterminer puis tracer la fonction de répartition $F_X$ de $X$.
4. Calculer les probabilités suivantes : $\mathbb{P}[X = 0,5]$, $\mathbb{P}[X \leqslant -0,5]$, $\mathbb{P}[X > 2]$ et $\mathbb{P}[0,2 \leqslant X < 0,5]$.
5. Déterminer la médiane de la variable $X$.
6. Déterminer l'intervalle interquartile de la variable $X$.
7. Déterminer le coefficient d'asymétrie ainsi que le coefficient d'aplatissement de la loi de variable $X$.

**Exercice 4.**
Soit une variable aléatoire $X$ régie par la densité suivante pour $k \neq 0$ :

$$f_X(x) = \begin{cases} \dfrac{x}{k} & \text{pour } 0 \leqslant x \leqslant 10 \\ \dfrac{20-x}{k} & \text{pour } 10 < x \leqslant 20 \\ 0 & \text{ailleurs} \end{cases}$$

1. Déterminer la valeur de la constante $k$ pour s'assurer que la fonction $f_X$ soit bien une densité.
2. Tracer la densité obtenue.
3. Déterminer la fonction de répartition.
4. Déterminer les probabilités suivantes : $\mathbb{P}[X \leqslant 0]$, $\mathbb{P}[1 \leqslant X \leqslant 2]$, $\mathbb{P}[15 \leqslant X \leqslant 18]$, $\mathbb{P}[5 \leqslant X \leqslant 15]$, $\mathbb{P}[X \geqslant 19]$, $\mathbb{P}[X \geqslant 9]$ et $\mathbb{P}[X \geqslant 29]$.
5. Déterminer les réels $a$ tel que $\mathbb{P}[X \leqslant a]$=1/4 puis $\mathbb{P}[X \leqslant a]$=3/4.
6. Déterminer la médiane de la variable $X$.
7. Déterminer l'intervalle interquartile de la variable $X$.
8. Déterminer le coefficient d'asymétrie ainsi que le coefficient d'aplatissement de la loi de variable $X$.

## TD 1 : Modèle dominé

**Exercice 5.**
Pour chacun des modèles suivants, décrire l'ensemble des paramètres et l'ensemble de définition de $X$, les ensembles "non-pathologiques" de mesure nulle, dire si le modèle est dominé par une des mesures introduites en cours (exhiber une mesure dominante) et, lorsque la fonction de répartition est absolument continue, donner sa densité.

1. $F_\theta(x) = \begin{cases} 1 - e^{-x^\theta} & \text{si } x \geqslant 0 \\ 0 & \text{sinon} \end{cases} \qquad (\theta > 0)\,;$
2. $F_{a,\,b}(x) = \begin{cases} 0 & \text{si } x < a \\ \dfrac{x-a}{b-a} & \text{si } a \leqslant x < b \\ 1 & \text{si } b \leqslant x \end{cases} \qquad (a < b)\,;$
3. $F_\theta(x) = \begin{cases} 0 & \text{si } x < 0 \\ \dfrac{1}{2}(1 - e^{-\theta x}) & \text{si } 0 \leqslant x < 1/\theta \\ 1 - \dfrac{1}{2}e^{-\theta x} & \text{si } 1/\theta \leqslant x \end{cases} \qquad (\theta > 0)\,;$
4. $F_{\alpha,\,\lambda}(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 - e^{-\lambda x^\alpha} & \text{si } 0 \leqslant x < 1/\lambda^{1/\alpha} \\ 1 & \text{si } 1/\lambda^{1/\alpha} \leqslant x \end{cases} \qquad (\alpha > 0,\, \lambda > 0)\,;$
5. $F_\theta(x) = \begin{cases} \displaystyle\sum_{k=0}^{\lfloor x \rfloor} \frac{\theta^k}{k!} e^{-\theta} & \text{si } x \geqslant 0 \\ 0 & \text{sinon} \end{cases} \qquad (\theta > 0)\,;$
6. $F_\theta$ est la fonction de répartition de $X = (Y - \theta)I(Y > \theta)$ où $Y$ suit une loi normale centrée réduite et où $\theta \in \mathbb{R}$ ;
7. $F_\theta$ est la fonction de répartition de $X = I(Y \leqslant 1) + YI(Y > 1)$ où $Y$ suit une loi uniforme sur $[0;\, 1 + \theta]$, $\theta > 0$.

**Exercice 6.**
(d'après le partiel de juin 2010)
Pour chacun des modèles suivants, dire si le modèle est dominé (auquel cas exhiber une mesure dominante), donner le support de la loi de $X$ noté $X(\Omega)$, donner l'espace des paramètres noté $\Theta$, donner les ensembles "non-pathologiques" de mesure nulle et dire si le modèle est régulier :

1. $\mathbb{P}_{a,\lambda}[X = x] = e^{-\lambda} \dfrac{\lambda^{x-a}}{(x-a)!}$ pour $x \in \{a, a+1, ...\}$ avec $\lambda > 0$ et $a \in \mathbb{N}$.
2. $F_{\alpha,\,\lambda}(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 - e^{-\alpha - \lambda x} & \text{si } x \geqslant 0 \end{cases}$ avec $\alpha, \lambda > 0$.

3. $F_a(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } 0 \leqslant x < a \\ \dfrac{(a-1)x - a}{a-2} & \text{si } a \leqslant x < 2 \\ 1 & \text{si } x \geqslant 2 \end{cases}$ avec $1,5 \leqslant a < 2$.

## TD 2 : Estimation d'un ou plusieurs paramètres

**Exercice 7.**
L'ancien ingénieur d'usine de l'entreprise Product a noté, en se basant sur une évaluation de plusieurs années, que 2% de la production de l'usine est défectueuse. Un nouvel ingénieur est embauché et met en œuvre une politique de fabrication plus stricte en vue d'améliorer la qualité de la production. Ce nouvel ingénieur tire successivement 200 pièces dans la nouvelle production et contrôle leur qualité. Il trouve 3 pièces défectueuses.

1. On suppose que le fait d'observer une pièce défectueuse n'influe pas sur la qualité des autres pièces. Comment traduire statistiquement la situation ?
2. Les chiffres obtenus indiquent-ils un changement de la qualité de la production ?

**Exercice 8.**
On admet ici que la durée de vie d'un pneu peut être représentée par une variable aléatoire $X$ de loi exponentielle de paramètre $\lambda$. Le fabricant de pneus de la marque Michel affirme que la durée de vie de ses pneus (en km) est une variable aléatoire $X$ de loi exponentielle de paramètre $\lambda = 0,0007$. Un statisticien embauché pour une mission d'audit fait réaliser une étude sur 100 pneus en banc d'essai. Il obtient $\sum_{i=1}^{100} x_i = 159992$km.

1. Estimer $\theta = \mathbb{P}[X \geqslant 2000]$ ponctuellement et par intervalle de confiance (en prenant le niveau de confiance de 95%).
2. Un contrat spécifie que la durée de vie minimum des pneus doit atteindre 1800km dans 95% des cas. Est-ce que le fabricant peut rentrer dans les critères du contrat en l'état actuel des choses ?

**Exercice 9.**
(d'après le partiel de janvier 2011)
L'entreprise Granulex distribue un aliment pour chat dans un contenant métallique dont le poids après remplissage, représenté par une variable aléatoire $X$, est calibré à une valeur nominale de 350g. On prélève un échantillon de 100 contenants afin de vérifier la calibration de la chaîne de production. On obtient $\sum_{i=1}^{100} X_i = 3600$g et $\sum_{i=1}^{100} X_i^2 = 129\,700\,000\text{g}^2$. Des études antérieures de la chaîne de production avaient déjà montré que le poids est distribué selon une loi normale.

1. Proposer un estimateur des paramètres de la loi normale en justifiant votre choix. Déterminer les estimations associées.
2. Estimer la probabilité qu'un contenant choisi au hasard de la production ait un poids compris entre 340g et 360g.
3. Estimer la probabilité que le contenant le plus lourd parmi 100 contenants prélevés soit inférieur à 360g.
4. Estimer la probabilité que le contenant le plus léger parmi 100 contenants prélevés soit supérieur à 330g.
5. L'entreprise met en production 1000 contenants. Estimer le nombre de contenants de poids inférieur à 330g.

NB : les résultats sont à exprimer en fonction de $\Phi$ la fonction de répartition de la loi normale centrée réduite.

**Exercice 10.**
Lors d'une campagne de dépistage du VHC, 100 travailleurs manuels se présentent pour effectuer un test de dépistage. Trois tests sont positifs. Sachant que 1% de la population française est porteuse du VHC, peut-on considérer que la population des travailleurs manuels est plus à risque que la moyenne ?

**Exercice 11.**
Un laboratoire pharmaceutique produit des comprimés qu'il commercialise sous la forme de boites de 40 comprimés. On note $p$ la proportion de comprimés défectueux dans la production du laboratoire. Introduisons la variable aléatoire $X$ représentant le nombre de comprimés défectueux dans une boîte prise au hasard de la production. Soit $(X_1, ..., X_n)$ un échantillon i.i.d. distribué comme $X$.

1. Quelle est la loi de $X$ ?
2. Quel est le support de la loi de $X$ noté $X(\Omega)$ ? Quel est l'ensemble des paramètres noté $\Theta$ ?
3. Déterminer l'espérance et la variance de $X$.
4. Déterminer l'information de Fisher du modèle probabiliste (à une observation).
5. Déterminer $\widetilde{p}_n$ l'estimateur de $p$ obtenu par la méthode des moments. Est-il biaisé ? Est-il fortement consistant ? Etudier son comportement asymptotique.
6. Déterminer $\widehat{p}_n$ l'estimateur du maximum de vraisemblance de $p$. Est-il biaisé ? Est-il fortement consistant ? Etudier son comportement asymptotique. Le modèle est-il régulier ?
7. Proposer un estimateur consistant de $\mathbb{P}[X = 0]$. Etudier son comportement asymptotique. En déduire un intervalle de confiance symétrique asymptotique au niveau de confiance 95% pour $\mathbb{P}[X = 0]$.

**Exercice 12.**
Un assureur s'intéresse aux sinistres que peuvent subir ses clients afin d'établir sa politique d'assurance. Soit $X$ la variable aléatoire représentant le montant annuel (exprimé en milliers d'euros) des sinistres que subit un client pris au hasard. On suppose que $X$ suit la loi de Pareto dont la distribution admet la densité suivante par rapport à la mesure de Lebesgue sur $\mathbb{R}$

$$f_\theta(x) = \frac{\theta}{x^{\theta+1}} I(x \geqslant 1).$$

où $\theta$ est un paramètre inconnu que l'on veut estimer. On dispose pour cela d'un échantillon i.i.d. $(X_1, ...., X_n)$ distribué comme $X$.

1. Quel est le support de la loi de $X$ noté $X(\Omega)$ ? Quel est l'espace naturel des paramètres noté $\Theta$ ?
2. Calculer l'espérance de $X$ et en déduire pourquoi on ne peut pas appliquer la méthode des moments pour estimer $\theta$ ?
3. Déterminer l'information de Fisher du modèle probabiliste (à une observation) lorsque c'est possible.
4. Déterminer $\widehat{\theta}_n$ l'estimateur du maximum de vraisemblance de $\theta$. Est-il biaisé ? Est-il fortement consistant ? Etudier son comportement asymptotique. Le modèle est-il régulier ?

5. Proposer un estimateur consistant de $\mathbb{P}[X > 1000]$. Etudier son comportement asymptotique. En déduire un intervalle de confiance symétrique asymptotique au niveau de confiance 95% pour $\mathbb{P}[X > 1000]$.

**Exercice 13.**
(d'après le partiel de janvier 2011)
On souhaite estimer le ratio de la durée de vie moyenne des fumeurs par rapport à la durée de vie moyenne des non-fumeurs :

$$\theta := \frac{\mathbb{E}[X]}{\mathbb{E}[Y]}$$

en notant $X$ la variable aléatoire représentant la durée de vie des fumeurs et $Y$ la variable aléatoire représentant la durée de vie des non-fumeurs.
Soit $(X_1, ..., X_{n_1})$ un échantillon i.i.d. distribué comme $X$ et soit $(Y_1, ..., Y_{n_2})$ un échantillon i.i.d. distribué comme $Y$, indépendant de $(X_1, ..., X_{n_1})$.

1. Proposer un estimateur $\widehat{\theta}_{n_1,n_2} = T(X_1, ..., X_{n_1}, Y_1, ..., Y_{n_2})$ de $\theta$ et justifier votre démarche.
2. On rappelle que la delta-méthode affirme que si

   $$\sqrt{n}\left((U_n, V_n) - (u, v)\right) \xrightarrow{\mathcal{D}} (U, V)$$

   alors, pour toute fonction $\Psi$ continûment différentiable (au moins en $(u, v)$),

   $$\sqrt{n}\left(\Psi(U_n, V_n) - \Psi(u, v)\right) \xrightarrow{\mathcal{D}} D\Psi(u, v).\begin{pmatrix} U \\ V \end{pmatrix}$$

   Déterminer la loi asymptotique de $\widehat{\theta}_{n_1,n_2}$ lorsque $n_1 = n_2 = n$.
3. En déduire la construction d'un intervalle de confiance bilatéral symétrique asymptotique pour $\theta$ de niveau de confiance $(1 - \alpha)$.

**Exercice 14.**
Afin de mieux dimensionner la capacité d'un serveur informatique, un ingénieur réalise une étude sur la durée de connexion des clients au serveur. On admet que la durée (exprimée en minutes) de connexion d'un client au serveur peut être représentée par une variable aléatoire $X$ positive de loi exponentielle de paramètre $\lambda > 0$ à estimer. Pour cela, l'ingénieur constitue un échantillon $(X_1, ...., X_n)$ de variables aléatoires i.i.d. distribuées comme la variable parente $X$. On rappelle que la fonction de répartition de la loi exponentielle est donnée pour $x > 0$ par

$$F_\lambda(x) = 1 - e^{-\lambda x}.$$

1. Déterminer $\widehat{\lambda}_n$ un estimateur de $\lambda$.
2. Etudier les propriétés et le comportement asymptotique de $\widehat{\lambda}_n$.
3. Déterminer un intervalle de confiance symétrique asymptotique pour $\lambda$ de niveau de confiance 95%.
4. Proposer un estimateur consistant de la probabilité qu'un client se connecte au serveur pour une durée comprise entre 15 et 20 minutes. Etablir son comportement asymptotique. En déduire un intervalle de confiance asymptotique de niveau de confiance 95% pour la probabilité qu'un client se connecte au serveur pour une durée comprise entre 15 et 20 minutes.

5. Estimer, parmi 1000 clients pris au hasard, le nombre moyen de clients se connectant au serveur pour une durée inférieure à 15 minutes. Etablir la consistance et le comportement asymptotique de l'estimateur choisi. En déduire un intervalle de confiance asymptotique de niveau de confiance 95% pour le nombre de clients se connectant au serveur pour une durée inférieure à 15 minutes parmi 1000 clients pris au hasard.

**Exercice 15.**
L'entreprise Simtech produit des tubes de verre de haute qualité qu'elle vend à l'entreprise Gescom sous forme de lots de 100 tubes de verre. L'ingénieur d'usine de l'entreprise Simtech s'intéresse à la proportion de tubes défectueux issus de cette production. Il dispose des données suivantes obtenues sur 200 lots :

| nombre de tubes défectueux | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| effectif | 98 | 60 | 22 | 16 | 2 | 2 |

1. Estimer la proportion du nombre de tubes défectueux dans l'ensemble de la production.
2. Estimer par intervalle la proportion de tubes défectueux dans l'ensemble de la production au niveau de confiance 95%.
3. Le qualiticien de l'entreprise Gescom utilise le plan de contrôle suivant à la réception de chaque lot. Prélever au hasard 5 tubes de verre. S'il y a, dans cet échantillon, 1 tube de verre (ou plus) défectueux, le lot est refusé et retourné à l'entreprise Simtech sans plus d'inspection. Estimer la probabilité qu'un lot soit refusé avec ce plan de contrôle.

**Exercice 16.**
Le directeur d'un hôpital étudie l'arrivée des patients dans le service d'urgence en pédiatrie ouvert 24h sur 24h. Il compte le nombre de patients se présentant aux urgences en pédiatrie par heure, et ce, pendant 24h. Il obtient les résultats suivants :
9 6 3 5 4 5 4 6 3 2 2 10 7 8 6 4 4 6 8 6 3 4 1 9
On admet que le nombre de patients se présentant aux urgences en pédiatrie par heure suit une loi de Poisson.

1. Proposer une estimation ponctuelle du nombre moyen de patients se présentant aux urgences en pédiatrie par heure.
2. Proposer une estimation par intervalle du nombre moyen de patients se présentant aux urgences en pédiatrie par heure au niveau de confiance 95%.
3. Estimer la probabilité qu'au cours d'une heure plus de 6 patients se présentent aux urgences de pédiatrie.
4. Estimer la probabilité qu'au cours d'une heure un nombre inférieur ou égal à 4 de patients se présentent aux urgences de pédiatrie.

**Exercice 17.**
Un hydrologue souhaite étudier le débit moyen annuel d'un fleuve donné en une région précise de façon à pouvoir établir des recommandations en ce qui concerne le pompage destiné à l'irrigation des cultures. Il dispose pour cela des volumes écoulés annuels entre le 1er janvier 1990 et le 31 décembre 2009 exprimés en $m^3$/an.

1. Pourquoi ne peut-on pas utiliser de résultats asymptotiques ?
2. L'hydrologue pense que le volume écoulé au cours d'une année donnée n'influe pas sur le volume écoulé l'année suivante. Il pense également que le fleuve n'a pas subi de changement de régime au cours de la période d'observation. Par ailleurs, un statisticien effectue un test validant la normalité des données. Qu'est-ce que cela implique pour les données ?

3. Le statisticien propose des estimateurs de la moyenne et de la variance du volume écoulé annuel du fleuve en cette région et en détermine la loi jointe exacte. Que fait-il ?
4. Le statisticien propose un intervalle de confiance exact au niveau de confiance 95% pour la moyenne du volume écoulé annuel du fleuve en cette région. Que fait-il ?
5. L'hydrologue fixe la règle suivante. La quantité d'eau pompée doit être telle que le volume écoulé restant ne doit pas être inférieur aux trois quarts de la moyenne observée sans pompage, et ce, avec une probabilité de 95%. Comment le statisticien peut-il estimer la quantité d'eau autorisée au pompage ?

**Exercice 18.**
Un barrage doit être construit dans une région montagneuse de façon à réguler le débit d'une rivière alimentée par la fonte des neiges et par les pluies. Un hydrologue souhaite déterminer la hauteur de la digue de façon à ce que les terres cultivées en aval ne soient pas inondées. Il dispose du relevé des maxima annuels du fleuve à l'endroit de la future construction entre le 1er janvier 1970 et le 31 décembre 2009 exprimés en $m$.

1. L'hydrologue trace l'histogramme des données (Figure 2), la boîte à moustaches correspondante (Figure 3) et estime les quantités suivantes :
   - $\widehat{m}$=2.55,
   - $\widehat{\sigma}^2$=2.67,
   - $\widehat{\gamma}_1$=0.94,
   - $\widehat{\gamma}_2$=3.32,

   en notant $m$ la moyenne, $\sigma^2$ la variance, $\gamma_1$ le coefficient d'asymétrie et $\gamma_2$ le coefficient d'aplatissement. Le statisticien lui fait remarquer que la normalité des données est peu

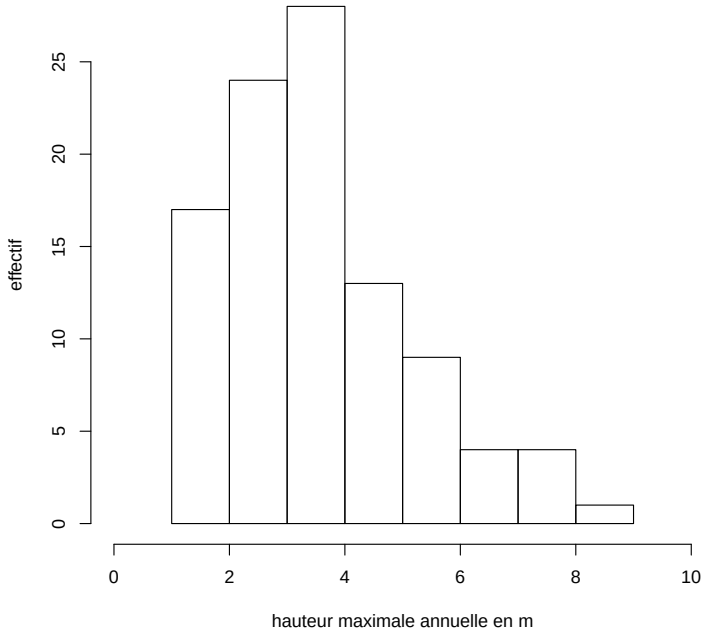



FIGURE 2 – Histogramme des données de maxima annuels

plausible. Pourquoi ?

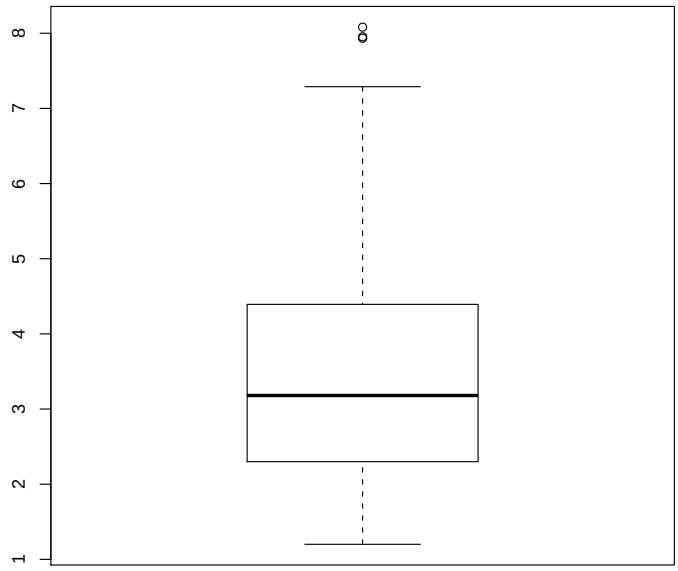



FIGURE 3 – Boîte à moustaches des données de maxima annuels

2. L'hydrologue pense que le niveau maximal de la rivière observé une année donnée n'influe pas sur le niveau maximal de la rivière observé l'année suivante. Il pense également que le fleuve n'a pas subi de changement de régime au cours de la période d'observation. Par ailleurs, un statisticien effectue un test invalidant la normalité des données. Qu'est-ce que cela implique pour les données ?
3. Le statisticien propose un estimateur respectivement de la moyenne et de la variance du maximum annuel de la rivière en cette région et en détermine la loi jointe asymptotique. Que fait-il ?
4. Le statisticien propose un intervalle de confiance asymptotique au niveau de confiance 95% pour la moyenne du niveau maximal annuel de la rivière en cette région. Que fait-il ?
5. A l'intention de l'hydrologue, le statisticien illustre la convergence de l'intervalle de confiance asymptotique avec des simulations. Pour cela, il génère 10000 échantillons i.i.d. de taille $n = 30; 50; 100; 200$ de loi non gaussienne. Il détermine à partir des 10000 intervalles de confiance obtenus la probabilité de couverture et la largeur moyenne de l'intervalle de confiance en fixant le niveau de confiance à 95%. Il obtient les résultats suivants :

| $n$ | probabilité de couverture | largeur moyenne |
|---|---|---|
| 30 | 0,933 | 12,97 |
| 50 | 0,937 | 10,08 |
| 100 | 0,944 | 7,18 |
| 200 | 0,949 | 5,08 |

   Quel commentaire fait-il à l'hydrologue ?
6. L'hydrologue souhaite déterminer la hauteur de la digue ayant une probabilité de 1/10000 d'être dépassée au cours d'un an. Le statisticien explique alors qu'il est nécessaire d'ajuster

une loi paramétrique afin de pouvoir estimer la hauteur de la digue. Le statisticien propose d'ajuster une loi Gamma notée $\Gamma(a,b)$ avec $a,b>$ ont la fonction de répartition est donnée pour $x>0$ par :

$$F_{a,b}(x)=\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}I(x>0)$$

et effectue un test validant cet ajustement. Il estime alors les paramètres de la loi Gamma et détermine leur comportement asymptotique. Que fait-il ?

7. Le statisticien en déduit alors une estimation de la hauteur de la digue satisfaisant le critère imposé par l'hydrologue. Que fait-il ?

**Exercice 19.**
(d'après le partiel de janvier 2011)
Soit $(X_1,...,X_n)$ un échantillon de variables aléatoires i.i.d. de variable parente $X$ dont la distribution admet la densité suivante par rapport à la mesure de Lebesgue sur $\mathbb{R}$ :

$$f_\mu(x)=\frac{1}{2}\exp\left(-|x-\mu|\right)\quad \text{avec } \mu\in\mathbb{R}.$$

1. Vérifier qu'il s'agit bien d'une densité de probabilité.
2. Déterminer l'espérance, la variance et la médiane de $X$.
3. Déterminer l'information de Fisher du modèle probabiliste (à une observation).
4. Le modèle appartient-il à la famille exponentielle ?
5. Déterminer $\widetilde{\mu}_n$ l'estimateur de $\mu$ obtenu par la méthode des moments. Est-il biaisé ? Etablir son comportement asymptotique.
6. Déterminer $\widehat{\mu}_n$ l'estimateur de $\mu$ obtenu par la méthode du maximum de vraisemblance.

**Exercice 20.**
(d'après le partiel de juin 2010)
Soit $(X_1,...,X_n)$ un échantillon de variables aléatoires i.i.d. de variable parente $X$ dont la distribution admet la densité suivante par rapport à la mesure de Lebesgue sur $\mathbb{R}$

$$f_\theta(x)=\frac{2x}{\theta^2}I(0\leqslant x\leqslant\theta)\quad \text{avec } \theta>0.$$

1. Vérifier qu'il s'agit bien d'une densité de probabilité.
2. Déterminer la fonction de répartition associée ainsi que l'espérance et la variance de $X$.
3. Déterminer l'information de Fisher du modèle probabiliste (à une observation).
4. Déterminer $\widetilde{\theta}_n$ l'estimateur de $\theta$ obtenu par la méthode des moments.
5. Déterminer $\widehat{\theta}_n$ l'estimateur de $\theta$ obtenu par la méthode du maximum de vraisemblance.
6. Déterminer le biais, la variance et le risque quadratique de $\widetilde{\theta}_n$ et $\widehat{\theta}_n$. Comparer $\widetilde{\theta}_n$ et $\widehat{\theta}_n$.
7. La borne de Fréchet-Darmois-Cramer-Rao s'applique-t-elle ici ? Pourquoi ?
8. Etudier le comportement asymptotique de $\widetilde{\theta}_n$ et de $\widehat{\theta}_n$.

**Exercice 21.**
Soient $X_1,...,X_n$ des variables aléatoires i.i.d. de variable parente $X$. On suppose que la distribution de $X$ admet la densité suivante par rapport à la mesure de Lebesgue sur $\mathbb{R}$

$$f_\theta(x)=(1+\theta)I(0\leqslant x\leqslant 1/2)+(1-\theta)I(1/2\leqslant x\leqslant 1).$$

où $\theta$ est un paramètre inconnu que l'on veut estimer.

1. Quel est le support de la loi de $X$ noté $X(\Omega)$ ? Quel est l'espace naturel des paramètres noté $\Theta$ ?
2. Déterminer l'espérance et la variance de $X$. Exhiber une statistique exhaustive pour $\theta$.
3. Déterminer l'information de Fisher du modèle probabiliste (à une observation) lorsque c'est possible.
4. Déterminer $\widetilde{\theta}_n$ l'estimateur de $\theta$ obtenu par la méthode des moments. Est-il biaisé ? Est-il fortement consistant ? Etudier son comportement asymptotique.
5. Déterminer $\widehat{\theta}_n$ l'estimateur du maximum de vraisemblance de $\theta$. Est-il biaisé ? Est-il fortement consistant ? Etudier son comportement asymptotique. Le modèle est-il régulier ?
6. Comparer $\widetilde{\theta}_n$ et $\widehat{\theta}_n$.
7. Proposer un estimateur consistant de $\mathbb{P}[0 \leqslant X \leqslant 1/4]$. Etudier son comportement asymptotique. En déduire un intervalle de confiance asymptotique au niveau de confiance 95% pour $\mathbb{P}[0 \leqslant X \leqslant 1/4]$.

**Exercice 22.**
Soit $X$ une variable aléatoire dont la distribution admet la densité suivante par rapport à la mesure de Lebesgue sur $\mathbb{R}$ pour $\lambda, \mu > 0$

$$f_{\mu,\lambda}(x) = \frac{\sqrt{\lambda}}{\sqrt{2\pi x^3}} \exp\left(-\lambda \frac{(x-\mu)^2}{2\mu^2 x}\right) I(x > 0).$$

Afin d'estimer les paramètres inconnus $\mu$ et $\lambda$, on dispose d'un échantillon i.i.d. $X_1, ...., X_n$ distribué comme $X$.

1. Quel est le support de la loi de $X$ noté $X(\Omega)$ ? Quel est l'espace naturel du paramètre bidimensionnel $\theta = (\mu, \lambda)$ noté $\Theta$ ?
2. Montrer que le modèle appartient à la famille exponentielle. En déduire l'espérance et la variance de $X$. Exhiber une statistique exhaustive pour $\theta$.
3. Déterminer l'information de Fisher du modèle probabiliste (à une observation) lorsque c'est possible.
4. Déterminer $\widetilde{\theta}_n$ l'estimateur de $\theta$ obtenu par la méthode des moments.
5. Déterminer $\widehat{\theta}_n$ l'estimateur du maximum de vraisemblance de $\theta$. Est-il fortement consistant ? Etudier son comportement asymptotique. Le modèle est-il régulier ?
6. En déduire un estimateur consistant de $\mathbb{E}[X]$ de variance minimale. Etudier son comportement asymptotique. En déduire un intervalle de confiance asymptotique au niveau de confiance 95% pour $\mathbb{E}[X]$.

## TD 3 : Autour de la notion d'espérance conditionnelle

**Exercice 23.**
Soit $(X, Y)$ un couple aléatoire à densité

$$(x, y) \mapsto \begin{cases} \dfrac{x}{2} + \dfrac{3y}{2} & \text{si } 0 \leqslant x, y \leqslant 1 \\ 0 & \text{sinon} \end{cases}$$

et soient $(x, y) \in [0; 1]$.
Calculer les moments conditionnels $\mathbb{E}(Y \mid X = x)$ et $\mathbb{E}(X \mid Y = y)$.

**Exercice 24.**
Soit $(X, Y)$ un couple aléatoire tel que

- $Y$ admet la densité $y \mapsto y\, e^{-y}\, I(y > 0)$ par rapport à la mesure de Lebesgue ;
- Il existe $\theta \in \mathbb{R}$ tel que pour tout $y > 0$, conditionnellement à $\{Y = y\}$, $X$ suit une loi uniforme sur $[\theta - y; \theta + y]$.

1. Déterminer la loi marginale de $X$.
2. $X$ et $Y$ sont-elles indépendantes ?
3. Calculer $\mathrm{Cov}(X, Y)$.

**Exercice 25.**
Soit $(X, Y)$ un couple aléatoire à densité $(x, y) \mapsto k\, I(0 < x < y < \theta)$ où $\theta > 0$ et $k$ est une constante réelle strictement positive.

1. Exprimer la constante $k$ en fonction de $\theta$.
2. Pour $y \in ]0; \theta[$ fixé, déterminer la loi conditionnelle de $X$ sachant $\{Y = y\}$.
3. Pour $x \in ]0; \theta[$ fixé, déterminer la loi conditionnelle de $Y$ sachant $\{X = x\}$.
4. $X$ et $Y$ sont-elles indépendantes ?
5. Calculer $\mathrm{Cov}(X, Y)$.

**Exercice 26.**
Soit $(X, Y)$ un couple aléatoire tel que

- $Y$ admet la densité $y \mapsto \dfrac{1}{\sqrt{2\pi y}}\, e^{-y/2}\, I(y > 0)$ par rapport à la mesure de Lebesgue ;
- Pour tout $y > 0$, conditionnellement à $\{Y = y\}$, $X$ a une loi à densité $x \mapsto \sqrt{\dfrac{y}{2\pi}}\, e^{-yx^2/2}$, $x \in \mathbb{R}$.

1. Déterminer la loi marginale de $X$ et montrer que $X$ n'admet pas d'espérance.
2. Pour $y > 0$ fixé, calculer $\varphi(y) = \mathbb{E}(X \mid Y = y)$, puis calculer $\mathbb{E}(\varphi(Y))$. Conclure que l'existence de $\mathbb{E}(\varphi(Y))$ n'implique pas celle de $\mathbb{E}(X)$.

**Exercice 27.**
Soit $(X, Y, Z)$ un triplet aléatoire tel que

- $X$ suit la loi uniforme sur $[0;\,1]$.
- Pour tout $x \in [0;\,1]$, conditionnellement à $\{X = x\}$, $Y$ a une loi à densité

$$y \mapsto (y-x)e^{-(y-x)}\, I(y > x).$$

- Pour tous $x \in [0;\,1]$ et $y > 0$, conditionnellement à $\{X = x,\, Y = y\}$, $Z$ a une loi à densité

$$z \mapsto (y-x)e^{-z(y-x)} I(z > 0).$$

Déterminer la loi jointe de $(X,\, Y,\, Z)$ ainsi que les lois marginales de $(X,\, Y)$, $(X,\, Z)$, $(Y,\, Z)$, $Y$ et $Z$.

**Exercice 28.**
Soit $(U, V)$ un couple de variables aléatoires indépendantes, toutes deux de loi uniforme sur $[0; 1]$. On définit la variable aléatoire discrète $N$ par $N = I(U \geqslant V)$.

1. Pour tout $n \in N(\Omega)$, déterminer la loi conditionnelle de $U$ sachant $\{N = n\}$.
2. En déduire, pour tout $n \in N(\Omega)$, le moment conditionnel $\mathbb{E}(U|N = n)$ puis l'espérance conditionnelle $\mathbb{E}(U|N)$.

**Exercice 29.**
Soit $(X_1, \ldots,\, X_n)$ un échantillon de variables aléatoires indépendantes et de même loi à densité $f_X$ par rapport à la mesure de Lebesgue, de fonction de répartition $F_X$ et d'espérance finie. On pose $Y = \max(X_1, \ldots,\, X_n)$.

1. Déterminer la fonction de répartition jointe de $(X_1, Y)$ notée

$$F_{(X_1,\,Y)}(x,\, y) = \mathbb{P}(X_1 \leqslant x,\, Y \leqslant y)$$

   et la fonction de répartition de $Y$ notée $F_Y$.
2. Dans la suite, on fixe $y \in Y(\Omega)$. Déterminer la fonction de répartition conditionnelle $F_{X_1\,|\,Y=y}(x) = \mathbb{P}(X_1 \leqslant x \,|\, Y = y)$ (remarquer que $Y$ est à densité par rapport à la mesure de Lebesgue).
3. En déduire que la loi conditionnelle de $X_1$ sachant $\{Y = y\}$ est la somme d'une mesure à densité par rapport à la mesure de Lebesgue et d'une mesure à densité par rapport à une masse de Dirac.
4. En déduire le moment conditionnel $\mathbb{E}(X_1 \,|\, Y = y)$.

**Exercice 30.**
Soient $X$ et $Y$ deux variables aléatoires indépendantes de loi de Bernoulli de paramètre $p \in ]0, 1[$. On pose $Z = I(X + Y = 0)$ et on note $\mathcal{G}$ la tribu engendrée par $Z$.

1. Déterminer $\mathbb{E}(X|\mathcal{G})$ et $\mathbb{E}(Y|\mathcal{G})$.
2. Les variables aléatoires $\mathbb{E}(X|\mathcal{G})$ et $\mathbb{E}(Y|\mathcal{G})$ sont-elles indépendantes ?

## TD 4 : Le cas particulier des vecteurs gaussiens

**Exercice 31.**
Soit $X$ une variable aléatoire gaussienne centrée réduite. On pose

$$Y = \begin{cases} X & \text{si } |X| < 1 \\ -X & \text{si } |X| \geqslant 1 \end{cases}$$

Montrer que $Y$ est une variable aléatoire gaussienne mais que le vecteur $(X, Y)$ n'est pas gaussien.

**Exercice 32.**
Soit $X$ une variable aléatoire gaussienne centrée réduite. Soit $\varepsilon$ une variable aléatoire centrée à valeurs dans $\{-1; 1\}$ indépendante de $X$. Montrer que $X$ et $\varepsilon X$ sont gaussiennes, orthogonales dans $L^2(\Omega)$ (l'espace des variables aléatoires de carré intégrable) mais pas indépendantes. En déduire que $(X, \varepsilon X)$ n'est pas un vecteur gaussien.

**Exercice 33.**

1. Parmi les matrices suivantes, lesquelles peuvent être la matrice de variance d'un vecteur aléatoire $\mathbf{U}$ de $\mathbb{R}^2$ ?

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} -1 & -1/2 \\ -1/2 & -1 \end{pmatrix} \begin{pmatrix} 1 & -1/2 \\ -1/2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 1/2 \\ 1/3 & 1 \end{pmatrix}$$

   Dans la suite, on notera $\mathbf{V}$ les matrices répondant à ce critère et on supposera que $\mathbf{U}$ suit la loi $\mathcal{N}_2(0_2, \mathbf{V})$.
2. Pour chaque matrice $\mathbf{V}$ retenue, déterminer les valeurs propres $(\lambda_1, \lambda_2)$ ainsi que les vecteurs propres associés $(\mathbf{v_1}, \mathbf{v_2})$.
3. Donner la loi jointe de $(\mathbf{v_1'}.\mathbf{U}, \mathbf{v_2'}.\mathbf{U})$.

**Exercice 34.**
Soit $(X, Y, Z)$ un vecteur gaussien dont l'espérance et la variance sont données respectivement par

$$\begin{pmatrix} 7 \\ 0 \\ 1 \end{pmatrix} \qquad \text{et} \qquad \begin{pmatrix} 10 & -1 & 4 \\ -1 & 1 & -1 \\ 4 & -1 & 2 \end{pmatrix}$$

Montrer que $(X, Y, Z)$ appartient presque-sûrement à un hyperplan de $\mathbb{R}^3$ que l'on déterminera.

**Exercice 35.**
Soit $(X, Y)$ un vecteur gaussien centré de matrice de variance l'identité. Calculer $\mathbb{E}[\max(X, Y)]$.

**Exercice 36.**
Soit $(X, Y)$ un vecteur gaussien centré de matrice de variance $\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$.

1. Calculer $\mathbb{E}[X|Y]$. En déduire la loi de $\mathbb{E}[X|Y]$.
2. Calculer $\mathbb{E}[X|Y-X]$. En déduire la loi de $\mathbb{E}[X|Y-X]$.

**Exercice 37.**

Soit $(X,Y,Z)$ un vecteur gaussien centré de matrice de variance $\begin{pmatrix}1 & 1 & 1\\ 1 & 5 & -1\\ 1 & -1 & 2\end{pmatrix}$.

1. Quelle est la loi de chacune des variables aléatoires $X$, $Y$ et $Z$ ?
2. Montrer que $(X-Y, Y+Z)$ est un vecteur gaussien.
3. Déterminer la loi de $U = X+Y+Z$.
4. Déterminer l'ensemble des variables $\eta_{a,b,c} = aX+bY+cZ$ indépendantes de $U$.
5. Quelle est la loi du vecteur $(Z,X,Y)$ ?

**Exercice 38.**

Soit $(X,Y,Z)$ un vecteur gaussien dont l'espérance et la variance sont données respectivement par

$$\begin{pmatrix}4\\1\\3\end{pmatrix} \qquad \text{et} \qquad \begin{pmatrix}4 & 2 & 2\\ 2 & 3 & 0\\ 2 & 0 & 2\end{pmatrix}$$

1. Calculer $\mathbb{E}[Y|X,Z]$ et $\mathbb{E}[XY|Z]$.
2. Calculer $\mathrm{Var}(Y|X,Z)$, $\mathrm{Var}(Y|X)$ et $\mathrm{Var}(Y|Z)$.
3. Donner la loi de $X$.
4. Donner la loi de $X$ sachant $Z$.
5. Donner la loi de $Z$ sachant $(X,Y)$.
6. Donner la loi de $X$ sachant $2Y+Z$.
7. Ces lois sont-elles absolument continues par rapport à la mesure de Lebesgue ? Si oui, donner leur densité.

**Exercice 39.**

Soient $U_1$, $U_2$ et $U_3$ des variables aléatoires indépendantes normales centrées et réduites.

1. Quelle est la loi du vecteur
$$\begin{pmatrix}U_1 - 2U_2\\ U_1+U_2+U_3\\ U_2-U_3\end{pmatrix} ?$$
2. Quelle est la loi de la variable $X = \dfrac{1}{3}(U_1+U_2+U_3)^2 + \dfrac{1}{2}(U_2-U_3)^2$ ?
3. Quelle est la loi de la variable $Y = \dfrac{2}{3}\dfrac{(U_1+U_2+U_3)^2}{(U_2-U_3)^2}$ ?
4. Quelle est la loi de la variable $Z = \sqrt{\dfrac{2}{3}}\dfrac{U_1+U_2+U_3}{|U_2-U_3|}$ ?

**Exercice 40.**

Soit $\begin{pmatrix}X\\Y\\Z\end{pmatrix}$ un vecteur aléatoire de loi $\mathcal{N}_3\left(\begin{pmatrix}2\\1\\5\end{pmatrix}, \begin{pmatrix}4 & 1 & 2\\ 1 & 9 & -1\\ 2 & -1 & 3\end{pmatrix}\right)$.

1. Calculer $\mathbb{E}[X - Z|Y - X]$. En déduire la loi de $\mathbb{E}[X - Z|Y - X]$.
2. Déterminer l'ensemble des variables $\eta_{a,b,c} = aX + bY + cZ$ avec $a, b, c \in \mathbb{R}$ indépendante de $U = Y + Z - X$.
3. Déterminer la loi de $\begin{pmatrix} Y - X \\ Z - X \end{pmatrix}$.
4. Déterminer la loi de $\begin{pmatrix} Y \\ X \\ Z \end{pmatrix}$.

**Exercice 41.**
(d'après le partiel de janvier 2011)

1. Les vecteurs suivants sont-ils gaussiens ?

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \mathcal{N}_3 \left( \begin{pmatrix} 2 \\ 1 \\ 5 \end{pmatrix}, \begin{pmatrix} 6 & -3 & -2 \\ -3 & 3 & 2 \\ -2 & -2 & 3 \end{pmatrix} \right) ; \quad \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \mathcal{N}_3 \left( \begin{pmatrix} 2 \\ 1 \\ 5 \end{pmatrix}, \begin{pmatrix} 4 & 1 & 2 \\ 1 & 9 & -1 \\ 2 & -1 & -3 \end{pmatrix} \right) ;$$
$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \mathcal{N}_3 \left( \begin{pmatrix} 2 \\ 1 \\ 5 \end{pmatrix}, \begin{pmatrix} 6 & -3 & -2 \\ -3 & 3 & 2 \\ -2 & 2 & 3 \end{pmatrix} \right) .$$

2. Soit $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ un vecteur aléatoire de loi $\mathcal{N}_3 \left( \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 6 & -3 & 0 \\ -3 & 3 & 2 \\ 0 & 2 & 3 \end{pmatrix} \right)$.
Calculer $\mathbb{E}[X - Z|Y - X]$. En déduire la loi de $\mathbb{E}[X - Z|Y - X]$.

## TD 5 : Corrélation linéaire et régression linéaire

**Exercice 42.**
Soient $X$ et $Y$ deux variables aléatoires représentant des quantités que l'on suspecte être associées. Notons $\rho(X,Y)$ le coefficient de corrélation linéaire de Bravais-Pearson entre $X$ et $Y$ défini par

$$\rho(X,Y) = \frac{\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Cet indicateur permet de quantifier le degré d'association linéaire entre $X$ et $Y$. Soient $(X_i, Y_i)$ pour $i = 1, ..., n$ des couples i.i.d. distribués comme $(X, Y)$.

1. Proposer un estimateur de $\rho(X,Y)$ basé sur l'échantillon des $(X_i, Y_i)_{i=1,...,n}$.
2. Démontrer la consistance forte de l'estimateur choisi et établir son comportement asymptotique. En déduire un intervalle de confiance pour $\rho(X,Y)$ au niveau de confiance 95%.
3. Pour chacun des quatre échantillons ci-dessous, tracer le nuage de points puis calculer une estimation de $\rho(X,Y)$. Commenter les résultats obtenus.

| données A | | données B | | données C | | données D | |
|---|---|---|---|---|---|---|---|
| $x$ | $y$ | $x$ | $y$ | $x$ | $y$ | $x$ | $y$ |
| 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 14 | 9.96 | 14 | 8.10 | 14 | 8.84 | 8 | 7.04 |
| 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 4 | 4.26 | 4 | 3.10 | 4 | 5.39 | 19 | 12.50 |
| 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |

**Exercice 43.**
Soient $X$ et $Y$ deux variables aléatoires représentant des quantités que l'on suspecte être associées. Plus particulièrement, on se demande s'il existe une influence de $X$ sur $Y$. Soient $(X_i, Y_i)$ pour $i = 1, ..., n$ des couples i.i.d. distribués comme $(X, Y)$.

1. (a) On souhaite ajuster un modèle de régression linéaire de la forme $Y = \beta_0 + \beta_1 X + \varepsilon$ où $\varepsilon$ est une variable aléatoire centrée indépendante de $X$ et de variance $\sigma^2$. Comment interpréter ce modèle ? Que valent $\mathbb{E}[Y|X]$ et $\mathrm{Var}(Y|X)$ ?

   (b) Parmi les quatre nuages de points ci-dessous, quelle doit être la forme du nuage de points formé par les données obtenues lors de la réalisation d'une expérience pour qu'il soit raisonnable d'ajuster le modèle

   $$Y = \beta_0 + \beta_1 X + \varepsilon$$

   où $\varepsilon$ est une variable aléatoire centrée indépendante de $X$ et de variance $\sigma^2$ ?

(c) Déterminer $\widehat{\beta}_0$ et $\widehat{\beta}_1$ les estimateurs de $\beta_0$ et $\beta_1$ respectivement obtenus par la méthode des moindres carrés i.e. satisfaisant le problème d'optimisation suivant :

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \text{argmin}_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

(d) Déterminer le biais, la variance et la covariance de $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sachant $X_1, ..., X_n$.

(e) Soit $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$ l'écart résiduel entre $Y_i$ la variable observée et $\widehat{Y}_i$ la valeur prédite par le modèle. Déterminer la moyenne et la variance empiriques des $\widehat{\varepsilon}_i$ pour $i = 1, ..., n$.

(f) Proposer un estimateur sans biais de $\sigma^2$ sachant $X_1, ..., X_n$.

2. (a) On souhaite maintenant ajuster un modèle de régression linéaire de la forme

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

où $\varepsilon$ est une variable aléatoire centrée indépendante de $(X, X^2)$ et de variance $\sigma^2$. Comment interpréter ce modèle ? Que valent $\mathbb{E}[Y|X]$ et $\text{Var}(Y|X)$ ?

(b) Parmi les quatre nuages de points ci-dessous, quelle doit être la forme du nuage de points formé par les données obtenues lors de la réalisation d'une expérience pour qu'il soit raisonnable d'ajuster le modèle $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$ où $\varepsilon$ est une variable aléatoire centrée indépendante de $(X, X^2)$ et de variance $\sigma^2$ ?

(c) Déterminer $\widehat{\beta}_0$, $\widehat{\beta}_1$ et $\widehat{\beta}_2$ les estimateurs de $\beta_0$, $\beta_1$ et $\beta_2$ respectivement obtenus par la méthode des moindres carrés i.e. satisfaisant le problème d'optimisation suivant :

$$(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2) = \text{argmin}_{(\beta_0,\beta_1,\beta_2)\in\mathbb{R}^3} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2)^2.$$

(d) Déterminer le biais et la variance de $\widehat{\beta}_0$, $\widehat{\beta}_1$ et $\widehat{\beta}_2$ sachant $X_1, ..., X_n$.

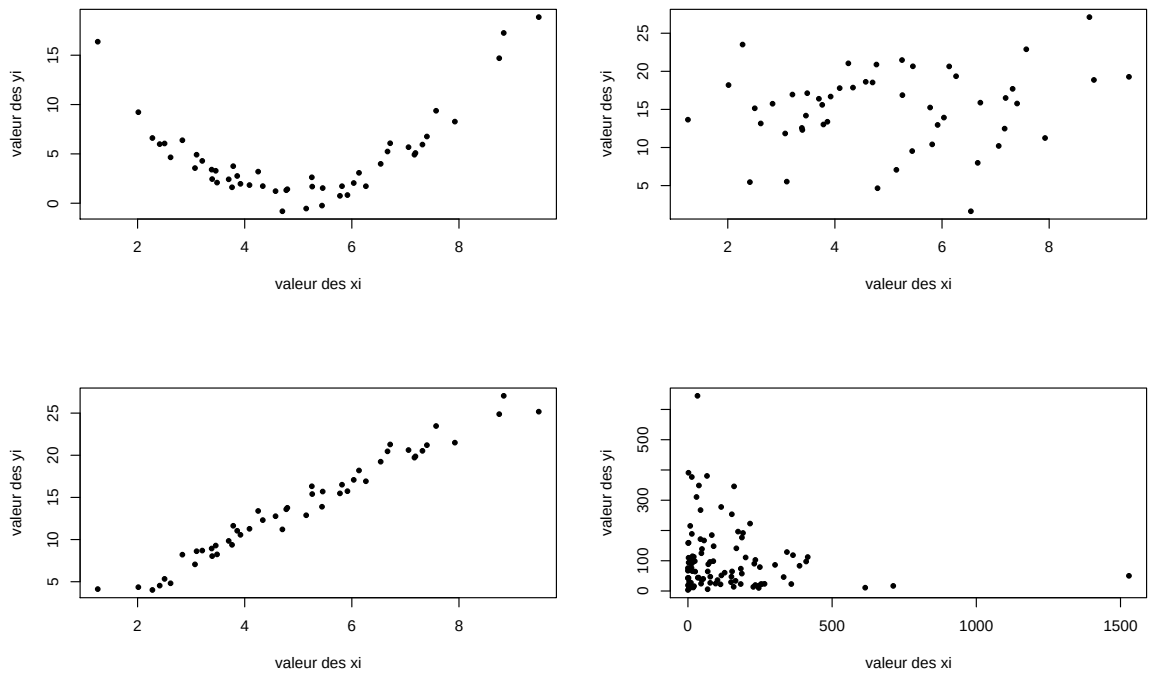
(e) Soit $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$ l'écart résiduel entre $Y_i$ la variable observée et $\widehat{Y}_i$ la valeur prédite par le modèle. Déterminer la moyenne et la variance empiriques des $\widehat{\varepsilon}_i$ pour $i = 1, ..., n$.

(f) Proposer un estimateur sans biais de $\sigma^2$ sachant $X_1, ..., X_n$.

**Exercice 44.**
(d'après le partiel de janvier 2011)
Soient $X$ et $Y$ des variables aléatoires représentant des quantités que l'on suspecte être associées. Plus particulièrement, on se demande s'il existe une influence de $\log(X)$ sur $Y$. Soient $(X_i, Y_i)$ pour $i = 1, ..., n$ des triplets i.i.d. distribués comme $(X, Y)$.

1. On souhaite ajuster un modèle de régression linéaire de la forme $Y = \beta_0 + \beta_1 \log(X) + \varepsilon$ où $\varepsilon$ est une variable aléatoire centrée indépendante de $X$ et de variance $\sigma^2$. Comment interpréter ce modèle ? Que valent $\mathbb{E}[Y|X]$ et $\text{Var}(Y|X)$ ?
2. Déterminer $\widehat{\beta}_0$ et $\widehat{\beta}_1$ les estimateurs de $\beta_0$ et $\beta_1$ respectivement obtenus par la méthode des moindres carrés i.e. satisfaisant le problème d'optimisation suivant :

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \text{argmin}_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \log(X_i))^2.$$




**Exercice 45.**
(d'après le partiel de juin 2010)
Soient $U$, $V$ et $Y$ des variables aléatoires représentant des quantités que l'on suspecte être associées. Plus particulièrement, on se demande s'il existe une influence de $U$ et $V$ sur $Y$. Soient $(U_i, V_i, Y_i)$ pour $i = 1, ..., n$ des triplets i.i.d. distribués comme $(U, V, Y)$.

1. On souhaite ajuster un modèle de régression linéaire de la forme $Y = \beta_0 + \beta_1 U + \beta_2 V + \varepsilon$ où $\varepsilon$ est une variable aléatoire centrée indépendante de $(U, V)$ et de variance $\sigma^2$. Comment interpréter ce modèle ? Que valent $\mathbb{E}[Y|U, V]$ et $\mathrm{Var}(Y|U, V)$ ?
2. Déterminer $\widehat{\beta}_0$, $\widehat{\beta}_1$ et $\widehat{\beta}_2$ les estimateurs de $\beta_0$, $\beta_1$ et $\beta_2$ respectivement obtenus par la méthode des moindres carrés i.e. satisfaisant le problème d'optimisation suivant :

$$(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2) = \mathrm{argmin}_{(\beta_0,\beta_1,\beta_2)\in\mathbb{R}^3} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 U_i - \beta_2 V_i)^2.$$

**Exercice 46.**
Soient $X$ et $Y$ deux variables aléatoires représentant des quantités que l'on suspecte être associées. Plus particulièrement, on se demande s'il existe une influence de $X$ sur $Y$. Soient $(X_i, Y_i)$ pour $i = 1, ..., n$ des couples i.i.d. distribués comme $(X, Y)$.

1. On souhaite ajuster un modèle de régression linéaire de la forme $Y = \beta_0 + \beta_1 X + \varepsilon$ où $\varepsilon$ est une variable aléatoire indépendante de $X$, de loi $\mathcal{N}(0, \sigma^2)$. Comment interpréter ce modèle ? Que valent $\mathbb{E}[Y|X]$ et $\mathrm{Var}(Y|X)$ ?
2. Déterminer la loi de $\widehat{\beta}_0$ et $\widehat{\beta}_1$ les estimateurs des moindres carrés de $\beta_0$ et $\beta_1$ respectivement.
3. Construire deux intervalles de confiance au niveau de confiance 95% l'un pour $\beta_0$ et l'autre pour $\beta_1$.

4. Soit $x_0$ une nouvelle réalisation de $X$. Utiliser le modèle pour prédire ponctuellement et par intervalle la valeur de $Y$ associée à cette réalisation de $X$.
5. On souhaite utiliser ce modèle de régression linéaire gaussien pour étudier l'influence du niveau de précipitations (en mm) sur le nombre d'accidents journaliers sur une portion d'autoroute. Qu'en pensez-vous ?