ENSIMAG 1ère année Mai 2014

# Principes et Méthodes Statistiques

Durée : 3 heures. Tous documents autorisés.

Les deux parties sont indépendantes.

Les résultats vus en cours ou en TD peuvent être utilisés sans être redémontrés.

Il sera grandement tenu compte de la qualité de la rédaction (présentation et justification des réponses) dans la notation.

Barème indicatif - Partie 1 : 11 pts, Partie 2 : 9 pts.

## Première partie

En médecine, le risque de thrombose (formation d'un caillot dans le réseau veineux des membres inférieurs) d'un adulte peut être évalué en mesurant son taux de D-dimères par dosage sanguin. La variable d'intérêt, notée $X$, est le logarithme de ce taux. $X$ est une variable aléatoire de loi normale de moyenne $m$ et de variance connue $\sigma^2 = 0.09$. On considère que $m$ ne peut prendre que deux valeurs : $m = -1$ pour les individus n'ayant pas de risque de thrombose, et $m = 0$ pour les individus ayant un risque de thrombose.

Pour un patient donné, on mesure la réalisation $x$ de $X$. On souhaite, au vu de $x$, se prononcer sur les deux hypothèses "le patient a un risque de thrombose" et "le patient n'a pas de risque de thrombose".

1. Le docteur A pense que le plus important est de ne pas inquiéter ses patients à tort.
   (a) Expliquer pourquoi cela le conduit à tester $H_0$ : "$m = -1$" contre $H_1$ : "$m = 0$".
   (b) Construire la région critique de ce test.
   (c) A partir de quelle valeur de $x$ peut-on conclure que le patient a un risque de thrombose, au seuil $\alpha = 5\%$ puis $\alpha = 1\%$ ?
   (d) Calculer la puissance du test. Donner la probabilité que le docteur A détecte correctement un patient à risque pour $\alpha = 5\%$ puis $\alpha = 1\%$.

2. Le docteur B préfère inquiéter un patient à tort plutôt que de ne pas l'avertir d'un risque réel.
   (a) Donner les hypothèses du test mis en place par le docteur B et construire sa région critique.

(b) A partir de quelle valeur de $x$ peut-on conclure que le patient n'a pas de risque de thrombose, au seuil $\alpha = 5\%$ puis $\alpha = 1\%$ ?

3. Montrer que, selon la valeur de $\alpha$, il peut ou non exister des valeurs de $x$ pour lesquelles le docteur A conclura que le patient a un risque de thrombose alors que le docteur B conclura qu'il n'en a pas. Etudier les cas $\alpha = 5\%$ et $\alpha = 1\%$.

4. Il y a un mois, $n = 9$ personnes ont été diagnostiquées comme présentant un risque de thrombose. Depuis, elles ont suivi un régime alimentaire particulier. Aujourd'hui, on mesure à nouveau le logarithme de leur taux de D-dimères, noté $Y$. $Y$ est de loi normale de moyenne $\mu$. Dans un premier temps, on suppose que $Y$ est de variance connue $\sigma^2 = 0.09$. Donc les observations forment un échantillon de $n$ variables aléatoires $Y_1, \ldots, Y_n$ indépendantes et de même loi $\mathcal{N}(\mu, 0.09)$. La moyenne empirique observée est $\bar{y}_n = -0.24$. On souhaite déterminer si le régime alimentaire a permis à ces patients d'éliminer le risque de thrombose.

   Ecrire le problème sous forme de test d'hypothèses simples, construire sa région critique, calculer la p-valeur et conclure.

5. Dans ce groupe de $n = 9$ patients, l'écart-type estimé est $s'_n = 0.37$. Cela remet-il en cause l'hypothèse que $\sigma^2 = 0.09$ ?

## Deuxième partie

On observe des réalisations $x_1, \ldots, x_n$ de variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi de probabilité à valeurs dans $\mathbb{R}^+$, définie par la densité :

$$f(x) = \sqrt{\frac{\lambda}{2\pi x^3}}\, e^{-\frac{\lambda(x-m)^2}{2m^2 x}}$$

où $m$ et $\lambda$ sont deux paramètres dans $\mathbb{R}^{+*}$.

1. Calculer les estimateurs de maximum de vraisemblance de $m$ et $\lambda$. Montrer qu'on peut les exprimer à l'aide de $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ et de $\overline{1/X}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{X_i}$.

2. On admet que l'espérance et la variance de cette loi de probabilité sont respectivement $m$ et $\frac{m^3}{\lambda}$. En déduire les estimateurs des moments de $m$ et $\lambda$. Montrer que l'estimateur de $m$ est sans biais.

On suppose à partir de maintenant que $\lambda = 1$.

3. Calculer la quantité d'information sur $m$ apportée par les observations. En déduire que l'estimateur de $m$ obtenu est sans biais et de variance minimale.

4. On suppose maintenant que la taille de l'échantillon est suffisamment grande pour que l'on puisse approcher la loi de $\bar{X}_n$ par une loi normale, en utilisant le théorème central-limite. Construire un test asymptotique de $H_0$ : "$m \leq m_0$" contre $H_1$ : "$m > m_0$".

5. Pour $n = 100$, à quelle condition sur la moyenne empirique des observations conclura-t-on que $m > 2$ au seuil 5% ?

# Statistique 2013-2014

## Première partie

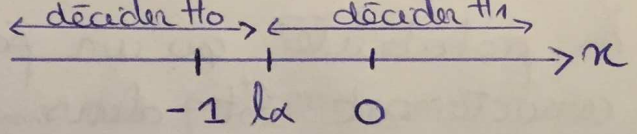
**1)** On a $X \sim \mathcal{N}(m, \sigma^2)$ avec $\begin{cases} m \in \{0, -1\} \\ \sigma^2 = 0{,}09 \end{cases}$

a) On choisit $H_0$ de façon à ne pas se tromper lorsqu'on rejette $H_0$.

Ici le docteur A cherche à ne pas se tromper lorsqu'il dit que le patient est malade.

D'où $H_0$ = "Le patient n'a pas de risque" = "$m = -1$"

b) On cherche $l_\alpha$ de façon à se retrouver dans le schéma suivant




avec la probabilité de rejeter $H_0$ à tort

$$P(x > l_\alpha \,;\, m = -1) = \alpha$$

- Sous $H_0$, $\frac{x+1}{\sigma} = \frac{x-m}{\sigma} \sim \mathcal{N}(0,1)$

Déc lors, $P\left(\frac{x+1}{\sigma} > \frac{l_\alpha + 1}{\sigma}\right) = 1 - \Phi\left(\frac{l_\alpha + 1}{\sigma}\right) = \alpha$

Ainsi, $\frac{l_\alpha + 1}{\sigma} = u_{2\alpha}$ où $u_\alpha$ est le $(1-\frac{\alpha}{2})$-quantile de $\mathcal{N}(0,1)$

$\Rightarrow$ La zone de rejet est donc $W_\alpha = \{x \in \mathbb{R} / x > \sigma u_{2\alpha} - 1\}$

c) On a $\sigma = 0{,}3$ et $u_{2\alpha} = \begin{cases} 1{,}65 \text{ si } \alpha = 5\% \\ 2{,}33 \text{ si } \alpha = 1\% \end{cases}$

Ainsi, on décide que le patient a un risque de thrombose

pour $x > -0{,}505$ si $\alpha = 5\%$

$x > -0{,}301$ si $\alpha = 1\%$

d) La puissance du test est $\beta = P(x \in W_\alpha ; m = 0)$

Sous l'hypothèse $m = 0$, $\frac{x}{\sigma} \sim \mathcal{N}(0,1)$

Posons $\mu = \frac{1}{\sigma}$ le paramètre de décentrage.

On a $\beta = P(x > l_\alpha) = P\left(\frac{x}{\sigma} > u_{2\alpha} - \mu\right)$

D'où $\beta = 1 - \Phi(u_{2\alpha} - \mu) = \Phi(\mu - u_{2\alpha})$

La probabilité qu'un patient soit détecté malade correctement est alors

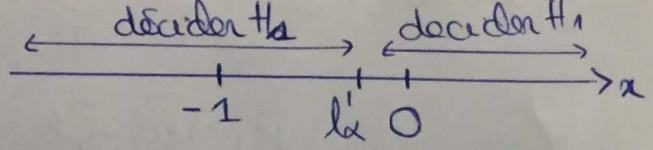
$\rightarrow \mu = \frac{10}{3} \simeq 3{,}3$

$$\beta = \begin{cases} 0{,}9505 & \text{si } \alpha = 5\% \\ 0{,}834 & \text{si } \alpha = 1\% \end{cases}$$

---

② Désormais on a $H_0$ : "$m = 0$" et $H_1$ : "$m = -1$"

Dans la situation suivante :




de façon à ce que $P(x < l'_\alpha ; m = 0) = \alpha$

Or sous $m = 0$, $\frac{x}{\sigma} \sim \mathcal{N}(0,1)$ et

$$P\left(\frac{x}{\sigma} < \frac{l'_\alpha}{\sigma}\right) = \Phi\left(\frac{l'_\alpha}{\sigma}\right) = \alpha = 1 - \Phi\left(-\frac{l_\alpha}{\sigma}\right)$$

Dès lors, $\frac{-l'_\alpha}{\sigma} = u_{2\alpha}$ et la région critique

est donc $W'_\alpha = \{x \in \mathbb{R} \,/\, x < -\sigma u_{2\alpha}\}$

b) On décide que le patient n'a pas de réponse ③
de thrompose pour $x < -0{,}49$ $\alpha = 5\%$

car $u_{1-2\alpha} = \begin{cases} 1{,}65 & \text{pour } \alpha = 5\% \\ 2{,}32 & \text{pour } \alpha = 1\% \end{cases}$ $x < -0{,}69$ $\alpha = 1\%$

---

③ On a $W_\alpha \cap W'_\alpha \subset ]-1, 0[$ et peut être non vide.
Dans ce cas, un patient aura une réponse différente selon le docteur A ou B.

Cas $\alpha = 5\%$ $W_\alpha \cap W_\alpha' = [-0{,}505 ; -0{,}49]$

Dans cette zone le patient aura des réponses différentes selon le docteur

Cas $\alpha = 1\%$ $W_\alpha \cap W_\alpha' = \emptyset$ $\Rightarrow$ Les patients auront des réponses identiques des deux docteurs

---

④ On ne veut pas se tromper lorsqu'on dit que le régime n'a pas fonctionné.

Ainsi $H_0$ : "$\mu = -1$" sous $H_1$ : "$\mu = 0$"

D'après fisher $\left(\frac{\overline{Y_n} - \mu}{\sigma}\right)\sqrt{n} \sim \mathcal{N}(0,1)$.

Ainsi sous $H_0$, $\frac{\sqrt{n}}{\sigma}\left(\overline{Y_n} + 1\right) \sim \mathcal{N}(0,1)$ et la

région critique est clairement

$$W_\alpha = \left\{ y \in \mathbb{R}^n / \overline{y_n} > \frac{\sigma u_{2\alpha}}{\sqrt{n}} - 1 \right\}$$

La p-valeur est $\alpha_y$ tq $\overline{y_n} = \frac{\sigma U_{2\alpha_y}}{\sqrt{n}} - 1$

soit $u_{2\alpha_y} = \frac{\sqrt{n}}{\sigma}(\overline{y_n}+1) = 7{,}6 \Rightarrow \alpha_y \simeq 10^{-10}$

La p-valeur étant très petite, on a aucune résitation à rejeter $H_0$

↳ Ce régime n'a pas fonctionné

---

⑤ On rappelle qu'un bon estimateur de $\sigma^2$ est $S_n'^2$. Or ici $S_n'^2 = 0{,}1369$ soit un écart relatif de 52% par rapport à $\sigma^2$. Ceci peut remettre en cause les résultats.

---

Deuxième partie $(X_1, \dots X_n)$ de loi $f : x \mapsto \sqrt{\frac{\lambda}{2\pi x^3}} e^{-\frac{\lambda(x-m)^2}{2m^2 x}} \geq 0$

① On a $\mathcal{L}(X_1, \dots X_n ; (\lambda, m)) = \prod_{i=1}^{n} f_{X_i}(x_i) = \left(\sqrt{\frac{\lambda}{2\pi}}\right)^n \frac{1}{\prod_{i=1}^{n} x_i^3} e^{-\frac{\lambda}{2m^2}\sum_{i=1}^{n}\frac{(x_i-m)^2}{x_i}}$

① d'où $\ln \mathcal{L}(X_1, \dots X_n ; (\lambda, m)) = \ln\left(\sqrt{\frac{\lambda}{2\pi}}\right)^n - \sum_{i=1}^{n} 3\ln(x_i) - \frac{\lambda}{2m^2}\sum_{i=1}^{n}\frac{(x_i-m)^2}{x_i}$

et $\frac{\partial}{\partial \lambda}\left(\ln \mathcal{L}(\dots)\right) = \frac{n}{2\lambda} - \frac{1}{2m^2}\sum_{i=1}^{n}\frac{(x_i-m)^2}{x_i} = 0 \Leftrightarrow \hat{\lambda}_n = \frac{m^2}{\frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-m)^2}{x_i}}$

finalement $\hat{\lambda}_n = \frac{1}{\left(\frac{\overline{x_n}}{m^2} - \frac{2}{m} + \overline{\left(\frac{1}{x_n}\right)}\right)}$

De plus $\frac{\partial}{\partial m}(\ln \mathcal{L})(\dots) = \frac{\partial}{m^2}\left(\frac{1}{m}\sum_{i=1}^{n}\frac{(x_i-m)^2}{x_i} + \sum_{i=1}^{n}\frac{(x_i-m)}{x_i}\right) = 0$

$\Longleftrightarrow \quad \hat{m}_n = \overline{x_n}$

(après développement)

---

② * On sait que $E[X_i] = m$. Dès lors l'estimateur des moments de m est $\tilde{m}_n = \overline{x_n} = \hat{m}_n$ qui est sans biais (cours)

↳ En effet $E(\tilde{m}_n) = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} m = m.$

* Un bon estimateur de $\sigma^2$ est $S_n'^2 = \frac{n}{n-1} S_n^2$

Or on sait que $\sigma^2 = \frac{m^3}{\lambda}$.

Dès lors l'EMM de $\lambda$ est $\tilde{\lambda}_n = \frac{m^3}{S_n'^2}$

---

③ L'information de fisher est

* $I_n(m) = -E\left[\frac{\partial^2}{\partial m^2} \ln \mathcal{L}(X_1, \dots X_n; (1, m))\right]$

$$= -E\left[\frac{2n}{m^3} - \frac{3n}{m^4}\overline{X_n}\right]$$

$$= n\left(\frac{3}{m^4}E[\overline{X_n}] - \frac{2}{m^3}\right)$$

$$= \frac{n}{m^3}$$

* $\hat{m}_n$ étant sans biais, son efficacité est $Eff(\hat{m}_n) = \frac{1}{I_n(m)\,var(\hat{m}_n)}$

Or $var(\hat{m}_n) = \sum_{i=1}^{n}\frac{var(X_i)}{n^2} = \frac{m^3}{n}$

L'efficacité est donc $\underline{Eff(\hat{m}_n) = 1}$

$\Rightarrow \hat{m}_n$ est un ESBVM

---

④ TCL : $\frac{\sqrt{n}}{\sigma[X]}(\overline{X}_n - E[X]) \sim \mathcal{N}(0,1)$

Ici $E[X] = m$ et $\sigma[X] = m^3$

On construit une zone $W_\alpha = \{x \in \mathbb{R}^n / \overline{x}_n > l_\alpha\}$

tq $P(X \in W_\alpha ; m \leqslant m_0) = \alpha$

Sous $H_0$ ; $\frac{\sqrt{n}}{m_0^3}(\overline{X}_n - m_0) \leqslant \frac{\sqrt{n}}{m^3}(\overline{X}_n - m)$

Donc $P\left(\frac{\sqrt{n}}{m_0^3}(\overline{X}_n - m_0) > \frac{\sqrt{n}}{m_0^3}(l_\alpha - m_0)\right) \leqslant 1 - \Phi\left(\frac{\sqrt{n}}{m_0^3}(l_\alpha - m_0)\right)$

Nécessairement $\frac{\sqrt{n}}{m_0^3}(l_\alpha - m_0) = u_{2\alpha}$ , la zone critique est alors $\underline{W = \{x \in \mathbb{R}^n / \overline{x}_n > \frac{m_0^3}{\sqrt{n}} u_{2\alpha} + m_0\}}$

---

⑤ On cherche $\overline{x}_n$ tq $x \in W$ avec $m_0 = 2$, $n = 100$, $\alpha = 5\%$

On trouve $\underline{\overline{x}_n > 3,316}$