TD 4 : Tests et IC

Préambule :
On rappelle que :

1. $\chi^2(n)$ désigne la loi d'une somme des carrés de n variables aléatoires $\mathcal{N}(0,1)$ indépendantes :
   Si $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, alors :
   $$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n).$$
2. **On admet** que dans un modèle gaussien i.i.d. $\mathcal{N}(m,\sigma^2)$, l'estimateur non biaisé de la variance $\hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ vérifie :
   $$(n-1)\frac{\hat{S}_n^2}{\sigma^2} \sim \chi^2(n-1) \text{ et } \hat{S}_n^2 \coprod \bar{X}.$$
3. $\mathrm{St}(n)$ désigne la loi du quotient entre une variable aléatoire $\mathcal{N}(0,1)$ et la racine d'une $\chi^2(n)$ normalisée par le nombre de degrés de liberté, indépendantes :
   si $Z \sim \mathcal{N}(0,1), U \sim \chi^2(n), Z \coprod U$:
   $$\frac{Z}{\sqrt{\frac{U}{n}}} \sim \mathrm{St}(n).$$
4. **On rappelle** que dans un modèle gaussien $\mathcal{N}(m,\sigma^2)$, on a
   $$\sqrt{n}\frac{\bar{X}-m}{\hat{S}_n} \sim \mathrm{St}(n-1).$$
5. Si Y est issu d'une loi discrète donnée par $\forall k \in [\![1,K]\!], \mathbb{P}(Y=k) = p_k$, que $Y_1, \cdots, Y_N$ est un échantillon i.i.d. de même loi que Y, et qu'on note $O_k = \sum_{i=1}^N \mathbb{1}_{Y_i=k}$, et $E_k = Np_k$, alors :
   $$\sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} \xrightarrow{\mathcal{L}} \chi^2(K-1) \text{ (quand N tend vers l'infini).}$$
6. Si (X,Y) désigne un couple de variables aléatoires discrètes indépendantes, à valeurs dans $[\![1,K_1]\!]\times[\![1,K_2]\!]$, que $(X_1,Y_1),\cdots,(X_N,Y_N)$ est un échantillon i.i.d. de même loi que (X,Y), et qu'on note $O_{k_1,k_2} = \sum_{i=1}^N \mathbb{1}_{X_i=k_1,Y_i=k_2}$, et $M_{k_1} = \sum_k O_{k_1,k}$ et $L_{k_2} = \sum_k O_{k,k_2}$, et $E_{k_1,k_2} = \frac{M_{k_1}L_{k_2}}{N}$ on a :
   $$\sum_{k_1,k_2=1}^{K_1,K_2} \frac{\left(O_{k_1,k_2} - E_{k_1,k_2}\right)^2}{E_{k_1,k_2}} \xrightarrow{\mathcal{L}} \chi^2\Big((K_1-1)(K_2-1)\Big).$$
7. La loi de Fisher $F(n_1,n_2)$ désigne un rapport de deux χ^2 à n_1, n_2 degrés de libertés, renormalisés par le nombre de degrés de liberté : Si $U_1 \sim \chi^2(n_1)$ et $U_2 \sim \chi^2(n_2)$, alors :
   $$\frac{\frac{U_1}{n_1}}{\frac{U_2}{n_2}} \sim F(n_1,n_2).$$
8. On rappelle qu'une combinaison affine de variables aléatoires de loi normales indépendantes reste de loi normale.
9. On rappelle le TCL : Si $(X_i)_{i\in\mathbb{N}}$ est une suite de variables aléatoires i.i.d. d'espérance m et de variance σ^2, alors
   $$\sqrt{n}\frac{\bar{X}-m}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$
10. On rappelle enfin que grâce au lemme de Slutsky, on peut remplacer l'écart-type dans la convergence précédente par une estimation consistante de ce dernier.

Exercice 1 ***IC et test bilatéraux vs unilatéraux***
On mesure la concentration en nitrates dans 30 *points d'un cours d'eau. On modélise ces mesures comme des réalisations i.i.d. d'une loi normale d'espérance* m *inconnue et de variance* σ^2 *inconnue. Les mesures ont donné* $\bar{x} = 46.7mg.L^{-1}$ *et* $\hat{s}^2 = 81.3mg^2.L^{-2}$.

1. *Donner un intervalle de confiance pour* m*, au degré de confiance* 99%
2. *Donner la p-valeur du test*
$$H_0 : m = 50 \ mg.L^{-1}$$
$$H_1 : m \neq 50 \ mg.L^{-1}$$
3. *Le test précédent est il rejeté au niveau* $\alpha = 0.05$ *?*
4. *Le taux ne doit surtout pas dépasser* $50mg.L^{-1}$*, sous peine de hauts risques sanitaires. Expliquer pourquoi on choisira de tester*
$$H_0 : m \geq 50 \ mg.L^{-1}$$
$$H_1 : m < 50 \ mg.L^{-1}$$
plutôt que dans l'autre sens.
5. *Donner la p-valeur du test précédent.*
6. *Ce test est il rejeté au niveau* $\alpha = 0.05$ *? et au niveau* $\alpha = 0.01$ *?*
7. *Donner un intervalle de confiance unilatéral à droite (de la forme* $]-\infty, A]$*) pour* m*, au degré de confiance* 99%.
8. *Aurait-on pu utiliser cet intervalle de confiance pour effectuer le test au niveau* 1%

Exercice 2 ***Test sur la variance et la moyenne***
On vous demande de tester la qualité de vis devant correspondre aux caractéristiques suivantes :

- *Longueur :* $40mm$
- *Diamètre :* $3mm$
- *Diamètre de la tête :* $6.5mm$

Le fabriquant annonce que l'écart-type de toutes les mesures est de $0.5mm$*. Vous procédez à* 25 *mesures* (L_i, d_i, D_i) *de la longueur, diamètre, et diamètre de la tête, et vous obtenez*
$$\bar{L} = 40.2, \bar{d} = 3.1, \bar{D} = 6.4, \hat{S}_{L,n} = 0.54, \hat{S}_{d,n} = 0.25, \hat{S}_{D,n} = 0.60$$
On modélise les données comme issues de lois normales indépendantes.

1. *Tester (au niveau* 5%*) pour chaque dimension, si la variance est celle donnée par le fabriquant, et donner les* $p-valeurs$ *associées.*
2. *Tester (au niveau* 5%*) si les dimensions sont celles données par le fabriquant, donner les* $p-valeurs$ *associées.*
3. *Donner les intervalles de confiance bilatéraux pour la variance, au degré de confiance de* 95%.
4. *Donner les intervalles de confiance unilatéraux à droite pour la variance (de la forme* $[0, A]$*), au degré de confiance de* 95%.
5. *Refaire le test (au niveau* 5%*) sur le diamètre en considérant la variance connue (et en prenant celle donnée par le fabriquant).*
6. *Comment pourriez-vous corriger le fait d'avoir voulu faire plusieur tests simultanéments pour contrôler la probabilité qu'au moins un de ces tests soit rejeté à tord ? Est-ce pertinent dans ce cadre là ?*

Exercice 3 ***Statistiques de tests***
Soient $X_1, \cdots, X_{n_1} \overset{i.i.d.}{\sim} \mathcal{N}(m_1, \sigma^2)$*, et* $Y_1, \cdots, Y_{n_2} \overset{i.i.d.}{\sim} \mathcal{N}(m_2, \sigma^2)$*, des échantillons indépendants de même variance* σ^2*. On note* $\bar{X}$*,* $\bar{Y}$ *les moyennes empiriques associées, et* $\hat{S}^2_{X,n_1}$*,*$\hat{S}^2_{Y,n_2}$ *les estimateurs sans biais de la variances associés.*

1. *En admettant le point* 2 *et* 3 *du préambule, montrer le point* 4 *du préambule.*
2. *En utilisant le point* 1 *du préambule, donner la loi d'une somme de* χ^2 *indépendantes, à respectivement* n_1 *et* n_2 *degrés de liberté.*
3. *En déduire la loi de* $\frac{(n_1-1)\hat{S}^2_{X,n_1}+(n_2-1)\hat{S}^2_{Y,n_2}}{\sigma^2}$.
4. *Calculer la loi de* $\sqrt{n'}\frac{\bar{X}-m_1-(\bar{Y}-m_2)}{\sigma}$*, où* $n' = \frac{1}{\frac{1}{n_1}+\frac{1}{n_2}}$.
5. *En utilisant le point* 3 *du préambule, et les deux dernières questions, montrer que* $U = \sqrt{n'}\frac{\bar{X}-m_1-(\bar{Y}-m_2)}{\sqrt{\frac{(n_1-1)\hat{S}^2_{X,n_1}+(n_2-1)\hat{S}^2_{Y,n_2}}{n_1+n_2-2}}}$ *vérifie*
$$U \sim \mathrm{St}(n_1 + n_2 - 2).$$

Exercice 4 ***Test de comparaison de variances et de moyenne***
On a mesuré les dimensions d'une tumeur chez des souris traitées ou non avec une substance antitumorale. On modélise les données comme issues de lois normales indépendantes d'espérances respectives m_1 et m_2, et de variances respectives σ_1^2 et σ_2^2. On a obtenu les résultats suivants :
Souris témoins : $n_1 = 20, \bar{X} = 7,075cm^2, \hat{S}_{X,20} = 0,576cm^2$.
Souris traitées : $n_2 = 18, \bar{Y} = 5,850cm^2, \hat{S}_{X,20} = 0,614cm^2$.

1. *Montrer que si $\sigma_1^2 = \sigma_2^2$, alors on a*
$$W := \frac{\hat{S}_X^2}{\hat{S}_Y^2} \sim F(n_1 - 1, n_2 - 1).$$
2. *Tester (au niveau 5%, et en donnant la p-valeur) :*
$$H_0 : \sigma_1 = \sigma_2$$
$$H_1 : \sigma_1 \neq \sigma_2$$
3. *Tester (au niveau 5%, et en donnant la p-valeur), en supposant l'égalité des variances, et en utilisant l'exercice précédent :*
$$H_0 : m_1 = m_2$$
$$H_1 : m_1 \neq m_2$$
4. *Tester (au niveau 5%, et en donnant la p-valeur), en supposant l'égalité des variances, et en utilisant l'exercice précédent :*
$$H_0 : m_1 \geq m_2$$
$$H_1 : m_1 > m_2$$

Exercice 5 ***Test de comparaison de variances et de moyenne 2***
Le pH (degré d'acidité) a été mesuré dans deux types de solutions chimiques A et B. Dans la solution A, six mesures ont été faites, avec un pH moyen de $7,52$ et un écart-type estimé de $0,024$. Dans la solution B, cinq mesures ont été faites, avec un pH moyen de $7,49$ et un écart-type estimé de $0,032$. On modélise les données comme issues de lois normales indépendantes. Déterminer si, au niveau $0,05$, les deux solutions ont des pH différents (donner la p-valeur).

Exercice 6 ***Test sur la variance***
On veut évaluer la performance de deux appareils de mesures de pression, à haute-pression. Pour cela, on travaille a pression contante et contrôlée de $m_0 = 100$ bars, et on prend 20 mesures avec chaque appareil, et calcule leur estimateur de la variance à espérance connue, $\hat{V}_1^2$ et $\hat{V}_2^2$ (définies par $\hat{V}^2 := \frac{1}{n}\sum_{i=1}^n (X_i - m_0)^2$.)
On modélise les mesures faites avec l'appareil i comme des réalisations indépendantes de lois normales d'espérance m_0 et de variance σ_i^2. Les observations ont donné $\hat{v}_1^2 = 0.0009$ et $\hat{v}_2^2 = 0.0019$.

1. *Calculer la p-valeur du test :*
$$H_0 : \sigma_1^2 \geq \sigma_2^2$$
$$H_1 : \sigma_1^2 < \sigma_2^2$$
2. *Donner aussi la p-valeur du test :*
$$H_0 : \sigma_1^2 \leq \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2$$