SEMAINE 13

Exercice 75. D'après le lemme 4.30, il suffit de montrer que $G_0(T(X_1, \dots, X_n)) \sim U[0,1]$ sous H_0. Or, sous H_0, G_0 est la fonction de distribution de $T(X_1, \dots, X_n)$, supposée continue. Lorsque Z est une variable aléatoire avec fonction de distribution continue G (cf. l'exercice 2, série 2)

$$\mathbb{P}(G(Z) \geq u) = \mathbb{P}(Z \geq G^{-1}(u)) = 1 - G(G^{-1}(u)) = 1 - u, \qquad u \in]0,1[,$$

où $G^{-1}(u) = \inf\{t : G(t) \geq u\}$ et les deux dernières égalités découlent de la continuité de G. Ainsi $G_0(T(X_1, \dots, X_n)) \sim U[0,1]$ et $1 - G_0(T(X_1, \dots, X_n)) \sim U[0,1]$ si H_0 est vraie.

Exercice 76. (i) Il faut résoudre le problème suivant :

$$\min U - L \qquad \text{t.q.} \qquad \Phi(U) - \Phi(L) \geq 1 - \alpha \qquad (U, L \in \mathbb{R}).$$

Puisque Φ est une fonction croissante, la contrainte peut s'écrire $U \geq \Phi^{-1}(1-\alpha+\Phi(L))$. Pour un L donné, il faut choisir le U le plus petit qui satisfait la contrainte. Ainsi, notre problème se réduit à trouver

$$\min g(L) = \Phi^{-1}(1 - \alpha + \Phi(L)) - L, \qquad L \in \mathbb{R}.$$

Notons cependant que $\Phi(L) \leq \Phi(U) - 1 + \alpha < \alpha$ et le domaine de g est $]-\infty, \Phi^{-1}(\alpha)[$. De plus, $g(L) \to \infty$ lorsque $L \to -\infty$ ou lorsque $L \to \Phi^{-1}(\alpha)$, et $g \geq 0$. Le minimum de g sera donc atteint à un point intérieur du domaine de g. Celle-ci est dérivable par le théorème de la fonction inverse (car Φ est strictement croissante et continûment dérivable).

La dérivée de g s'annule si et seulement si

$$1 = \frac{\Phi'(L)}{\Phi'(\Phi^{-1}(1 - \alpha + \Phi(L)))} = \frac{\Phi'(L)}{\Phi'(U)} = \frac{\exp(-L^2/2)}{\exp(-U^2/2)},$$

c'est-à-dire lorsque $L = \pm U$. Or, Φ est croissante et $\Phi(U) - \Phi(L) = 1 - \alpha > 0$, donc forcément $L < U$. On a donc $L = -U$ et par symétrie $\Phi(U) = 1 - \Phi(L)$, donc

$$1 - \alpha = \Phi(U) - \Phi(L) = 1 - 2\Phi(L) \Longrightarrow \Phi(L) = \frac{\alpha}{2} \Longrightarrow \Phi(U) = 1 - \frac{\alpha}{2}.$$

Le but de la discussion ci-dessus était de montrer qu'il s'agit d'un minimum sans devoir calculer la dérivée seconde de g . À noter qu'il est facile dans ce cas de montrer que $g''(\Phi^{-1}(\alpha/2)) > 0$ et donc qu'il s'agit bel et bien d'un minimum.

Remarque. Le choix $L = \Phi^{-1}(\alpha)$ correspond à $U = \infty$ et donne l'intervalle de confiance unilatéral à gauche. Le choix $L = -\infty$ correspond à $U = \Phi^{-1}(1-\alpha)$ et donne l'intervalle unilatéral à droite.

(ii) Comme dans l'exemple 5.3, posons $Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma \sim N(0,1)$ et remarquons que

$$\mathbb{P}[A_n \leq \mu \leq B_n] = \mathbb{P}(\overline{X}_n - B_n \leq \overline{X}_n - \mu \leq \overline{X}_n - A_n) = \mathbb{P}\left[\frac{\sqrt{n}}{\sigma}(\overline{X}_n - B_n) \leq Z_n \leq \frac{\sqrt{n}}{\sigma}(\overline{X}_n - A_n)\right].$$

Il faut minimiser $B_n - A_n$, ce qui équivaut à minimiser $(\sqrt{n}/\sigma)(\overline{X}_n - A_n) - (\sqrt{n}/\sigma)(\overline{X}_n - B_n)$, mais sous la contrainte que cette probabilité soit au moins $1-\alpha$. Par la partie (i), la solution est

$$\left[\frac{\sqrt{n}}{\sigma}(\overline{X}_n - B_n), \frac{\sqrt{n}}{\sigma}(\overline{X}_n - A_n)\right] = [z_{\alpha/2}, z_{1-\alpha/2}] = [-z_{1-\alpha/2}, z_{1-\alpha/2}].$$

Ainsi, la solution de notre problème est

$$[A_n, B_n] = \left[\overline{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

qui est donc l'intervalle de confiance basé sur $\overline{X}_n$ de seuil (supérieure ou égale à) $(1-\alpha)$ ayant la plus petite longueur.

(iii) Le même résultat est valable lorsque Z suit une loi ayant une densité symétrique f, qui est strictement décroissante sur $\mathbb{R}_+$. C'est-à-dire, le résultat est valable si
— pour chaque $x \in \mathbb{R}$, $f(x) = f(-x)$;
— pour chaque $0 < x < y$, $f(x) > f(y)$.
Par exemple, ceci est bien le cas si $Z \sim \mathbf{t}_k$ pour $k > 0$. Ainsi, même si la variance σ^2 est inconnue, en la remplaçant par l'estimateur S^2, on obtiendra l'intervalle de confiance ayant la plus petite longueur.
Remarque. Sous ces conditions, on peut montrer que l'intervalle $[L, U]$ est l'ensemble (mesurable) F ayant la mesure de Lebesgue la plus petite et tel que $\mathbb{P}(F \ni Z) \geq 1-\alpha$. Il est donc inutile de chercher (par exemple) une union d'intervalles.

Exercice 77. (i) D'après l'exercice 74 (avec $n = m$), la variable aléatoire

$$T = \frac{\sqrt{\frac{n^2}{n+n}}\left(\overline{X} - \mu_X - \overline{Y} + \mu_Y\right)}{\sqrt{\frac{1}{n+n-2}[(n-1)S_X^2 + (n-1)S_Y^2]}} = \frac{\sqrt{n}\left(\overline{X} - \overline{Y} - \theta\right)}{\sqrt{S_X^2 + S_Y^2}} = \frac{\overline{X} - \overline{Y} - \theta}{\sqrt{\frac{1}{n}(S_X^2 + S_Y^2)}}$$

suit une loi $\mathbf{t}_{2n-2}$ pour chaque $\theta \in \mathbb{R}$. (Parce que $S_X^2 = S_{X-c}^2$ pour chaque constante $c \in \mathbb{R}$.) Ainsi, $T = g(X_1, \ldots, X_n, Y_1, \ldots, Y_n, \theta)$ est un pivot (la continuité par rapport à θ est évidente). À partir de là, on n'a qu'à faire les manipulations habituelles :

$$\begin{aligned}
1-\alpha &= \mathbb{P}_\theta\left[t_{2n-2,\alpha/2} \leq \frac{\bar{X} - \bar{Y} - \theta}{\sqrt{\frac{1}{n}(S_X^2 + S_Y^2)}} \leq t_{2n-2,1-\alpha/2}\right] \\
&= \mathbb{P}_\theta\left[t_{2n-2,\alpha/2}\sqrt{\frac{1}{n}(S_X^2 + S_Y^2)} \leq \bar{X} - \bar{Y} - \theta \leq t_{2n-2,1-\alpha/2}\sqrt{\frac{1}{n}(S_X^2 + S_Y^2)}\right] \\
&= \mathbb{P}_\theta\left[\bar{X} - \bar{Y} - t_{2n-2,1-\alpha/2}\sqrt{\frac{1}{n}(S_X^2 + S_Y^2)} \leq \theta \leq \bar{X} - \bar{Y} - t_{2n-2,\alpha/2}\sqrt{\frac{1}{n}(S_X^2 + S_Y^2)}\right] \\
&= \mathbb{P}_\theta\left[\bar{X} - \bar{Y} - t_{2n-2,1-\alpha/2}\sqrt{\frac{1}{n}(S_X^2 + S_Y^2)} \leq \theta \leq \bar{X} - \bar{Y} + t_{2n-2,1-\alpha/2}\sqrt{\frac{1}{n}(S_X^2 + S_Y^2)}\right].
\end{aligned}$$

On conclut que $\left[\bar{X} - \bar{Y} - t_{2n-2,1-\alpha/2}\sqrt{\frac{1}{n}(S_X^2 + S_Y^2)}, \bar{X} - \bar{Y} + t_{2n-2,1-\alpha/2}\sqrt{\frac{1}{n}(S_X^2 + S_Y^2)}\right]$ est un intervalle de confiance pour $\theta = \mu_X - \mu_Y$ avec un seuil $1-\alpha$.

Remarque. On peut définir $Z_i = X_i - Y_i \overset{\text{iid}}{\sim} N(\theta, 2\sigma^2)$ (ce qui par ailleurs aurait été plus compliqué si $m \neq n$, c'est-à-dire si le nombre de X_i n'était pas égal au nombre de Y_i) de sorte que

$$\frac{\overline{X} - \overline{Y} - \theta}{\sqrt{\frac{1}{n}S_Z^2}} = \sqrt{n}\frac{\overline{Z} - \theta}{\sqrt{S_Z^2}} = \frac{\sqrt{\frac{n}{2\sigma^2}}(\overline{Z} - \theta)}{\sqrt{\frac{1}{n-1}}\sqrt{\frac{(n-1)S_Z^2}{2\sigma^2}}} \sim \mathbf{t}_{n-1}, \qquad S_Z^2 = \sum_{i=1}^{n}(Z_i - \overline{Z})^2,$$

car le nominateur suit une loi $N(0,1)$ et le dénominateur est $V/\sqrt{n-1}$ où $V \sim \chi^2_{n-1}$, et les deux sont indépendantes. Ainsi on obtient l'intervalle de confiance

$$\left[\bar{X} - \bar{Y} - t_{n-1,1-\alpha/2}\sqrt{\frac{1}{n}S_Z^2}, \bar{X} - \bar{Y} + t_{n-1,1-\alpha/2}\sqrt{\frac{1}{n}S_Z^2}\right].$$

Cet intervalle sera probablement plus grand que celui d'avant : puisque $S_Z^2 \to 2\sigma^2$ et $S_X^2 + S_Y^2 \to 2\sigma^2$, on s'attend à ce que les deux aient une taille similaire (ils ont en tous cas la même espérance et la même variance). Or pour β fixé, la fonction $k \mapsto t_{k,\beta}$ est décroissante. Notre deuxième intervalle aura donc tendance à être plus grand, puisqu'on utilise t_{n-1} au lieu de t_{2n-2}. Intuitivement, on a utilisé n données (les différences $X_i - Y_i$) au lieu d'en utiliser $2n$.

En revanche, le premier intervalle est moins général que le deuxième : ce dernier suppose uniquement que les différences Z_i sont iid, alors que dans le premier cas on a supposé que toutes les X_i sont indépendantes de toutes les Y_i, une supposition plus forte. Dans le cas apparié (exercice 6, série 10), on ne peut utiliser que le deuxième intervalle !

(ii) En utilisant le résultat de la partie a), on obtient l'intervalle de confiance à 95% :

$$\left[\overline{X}_1 - \overline{X}_2 - t_{28,0.975}\sqrt{\frac{1}{15}(S_1^2 + S_2^2)}, \overline{X}_1 - \overline{X}_2 + t_{28,0.975}\sqrt{\frac{1}{15}(S_1^2 + S_2^2)}\right] \approx [0.30, 2.10].$$

On peut donc conclure que le temps de l'efficacité de M_1 est meilleur que celui de M_2 par 18–126 minutes avec un seuil de confiance 95%.

Exercice 78. Soient $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ avec $\mu \in \mathbb{R}$ et $\sigma^2 > 0$ inconnus. Supposons qu'on aimerait trouver un intervalle de confiance pour μ (donc σ^2 est un paramètre de nuisance). D'après l'exemple 5.7, on sait que

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \to N(0,1), \qquad \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Or ici les X_i sont normales ; on connaît donc la distribution exacte de $T_n = \sqrt{n}(\overline{X}_n - \mu)/S_n$: d'après le théorème 2.9, elle est $\mathbf{t}_{n-1}$. L'énoncé est donc démontré.

Exercice 79. Montrons tout d'abord le résultat dans l'indice.
Soient F_n et F les fonctions de répartition de Z_n et Z respectivement. Pour tout $\epsilon > 0$,

$$0 \leq \mathbb{P}(Z_n = a) \leq \mathbb{P}(a - \epsilon < Z_n \leq a) = F_n(a) - F_n(a - \epsilon) \to F(a) - F(a - \epsilon),$$

lorsque $n \to \infty$, puisque F est continue. Ainsi

$$\limsup_{n\to\infty} \mathbb{P}(Z_n = a) \leq F(a) - F(a - \epsilon) \to 0, \qquad \epsilon \to 0,$$

car F est continue. Ceci prouve le résultat cherché.
Vérifions le cas bilatéral :

$$\begin{aligned}
&\mathbb{P}[\hat{\theta}_n - z_{1-\alpha/2}\hat{J}_n^{-1/2} \leq \theta \leq \hat{\theta}_n + z_{1-\alpha/2}\hat{J}_n^{-1/2}] \\
=&\mathbb{P}[-z_{1-\alpha/2} \leq \hat{J}_n^{1/2}(\hat{\theta}_n - \theta) \leq z_{1-\alpha/2}] \\
=&\mathbb{P}[z_{\alpha/2} \leq \hat{J}_n^{1/2}(\hat{\theta}_n - \theta) \leq z_{1-\alpha/2}] \\
=&F_n(z_{1-\alpha/2}) - F_n(z_{\alpha/2}) + \mathbb{P}[\hat{J}_n^{1/2}(\hat{\theta}_n - \theta) = z_{\alpha/2}],
\end{aligned}$$

où F_n est la fonction de répartition de $\hat{J}_n^{1/2}(\hat{\theta}_n - \theta)$ et où on a utilisé le fait que $z_{\alpha/2} = -z_{1-\alpha/2}$. Par la proposition 5.8 (p. 123), on sait que $F_n(x) \to \Phi(x)$ pour chaque $x \in \mathbb{R}$, où $\Phi(x)$ est la fonction de répartition de $N(0,1)$. De plus, par la proposition ci-dessus

$$\mathbb{P}[\hat{J}_n^{1/2}(\hat{\theta}_n - \theta) = z_{\alpha/2}] \to 0, \qquad n \to \infty.$$

Donc

$$\mathbb{P}[\hat{\theta}_n - z_{1-\alpha/2}\hat{J}_n^{-1/2} \leq \theta \leq \hat{\theta}_n + z_{1-\alpha/2}\hat{J}_n^{-1/2}] \to \Phi(z_{1-\alpha/2}) - \Phi(z_{\alpha/2}) = 1 - \alpha/2 - \alpha/2 = 1 - \alpha.$$

Il s'en suit que $[\hat{\theta}_n - z_{1-\alpha/2}\hat{J}_n^{-1/2}, \hat{\theta}_n + z_{1-\alpha/2}\hat{J}_n^{-1/2}]$ est un intervalle de confiance approximatif avec seuil $1 - \alpha$.
De la même façon, on trouve que

$$\mathbb{P}[\hat{\theta}_n - z_{1-\alpha}\hat{J}_n^{-1/2} \leq \theta] = \mathbb{P}[\hat{J}_n^{1/2}(\hat{\theta}_n - \theta) \leq z_{1-\alpha}] = F_n(z_{1-\alpha}) \to \Phi(z_{1-\alpha}) = 1 - \alpha$$

et que

$$\begin{aligned}
\mathbb{P}[\theta \leq \hat{\theta}_n + z_{1-\alpha}\hat{J}_n^{-1/2}] &= 1 - \mathbb{P}[\theta > \hat{\theta}_n + z_{1-\alpha}\hat{J}_n^{-1/2}] \\
&= 1 - \mathbb{P}[\hat{J}_n^{1/2}(\hat{\theta}_n - \theta) < z_\alpha] \\
&= 1 - F_n(z_\alpha) + \mathbb{P}[\hat{J}_n^{1/2}(\hat{\theta}_n - \theta) = z_\alpha] \\
&\to 1 - \Phi(z_\alpha) = 1 - \alpha.
\end{aligned}$$

Donc, $[\hat{\theta} - z_{1-\alpha}\hat{J}_n^{-1/2}, +\infty]$ et $[-\infty, \hat{\theta} + z_{1-\alpha}\hat{J}_n^{-1/2}]$ sont les intervalles de confiance unilatéraux approximatifs avec seuil $1 - \alpha$.