

Série 11

Solution 45. Soient $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ et $p \in (0, 1)$, alors par hypothèse la quantile empirique $Y_{(\lceil np \rceil)} \xrightarrow{P} y_p$ quand $n \rightarrow \infty$, où y_p dénote le quantile théorique correspondant.

- a) Quand $n \rightarrow \infty$, on a alors $M = Y_{(\lceil n/2 \rceil)} \xrightarrow{P} y_{1/2}$, où $F(y_{1/2}) = 1/2 = \Phi(0) = \Phi(\frac{(y_{1/2} - \mu)}{\sigma})$, ce qui donne $y_{1/2} = \mu$.

De même façon, les limites des quartiles empiriques sont les solutions aux équations

$$F(y_{3/4}) = \frac{3}{4} = \Phi\left(\frac{y_{3/4} - \mu}{\sigma}\right), \quad F(y_{1/4}) = \frac{1}{4} = \Phi\left(\frac{y_{1/4} - \mu}{\sigma}\right),$$

c'est-à-dire $y_{3/4} = \mu + \sigma \times \Phi^{-1}(\frac{3}{4}) = \mu + 0.6745\sigma$, $y_{1/4} = \mu + \sigma \times \Phi^{-1}(\frac{1}{4}) = \mu - 0.6745\sigma$.

Puisque $Y_{(\lceil 3n/4 \rceil)} \xrightarrow{P} y_{3/4}$, on a $Y_{(\lceil 3n/4 \rceil)} \xrightarrow{D} y_{3/4}$, et ainsi le lemme de Slutsky (3ème ligne du théorème 115) donne

$$Y_{(\lceil 3n/4 \rceil)} - Y_{(\lceil n/4 \rceil)} \xrightarrow{D} (\mu + 0.6745\sigma) - (\mu - 0.6745\sigma) \approx 1.35\sigma.$$

Mais puisque $Y_{(\lceil 3n/4 \rceil)} - Y_{(\lceil n/4 \rceil)}$ converge en loi vers une constante, elle converge aussi en probabilité par la 1ère ligne du théorème, et

$$\text{IQR} = Y_{(\lceil 3n/4 \rceil)} - Y_{(\lceil n/4 \rceil)} \xrightarrow{P} 1.35\sigma, \quad n \rightarrow \infty.$$

- b) Les limites de M et de IQR peuvent être considérés comme des paramètres du modèle puisqu'elles sont des fonctions fixes de la fonction de répartition.
- c) Soient $Q_{\text{high}} = y_{3/4} + 1.5 \times \text{IQR} \approx \mu + 2.7\sigma$, $Y \sim \mathcal{N}(\mu, \sigma^2)$ et $Z \sim \mathcal{N}(0, 1)$, alors par symétrie la proportion des données gaussiennes à l'extérieur des moustaches d'un boxplot est

$$\begin{aligned} 2\Pr(Y > Q_{\text{high}}) &= 2\{1 - \Pr(Y \leq Q_{\text{high}})\} \\ &= 2\{1 - \Pr(\mu + \sigma Z \leq \mu + 2.7\sigma)\} \\ &= 2\{1 - \Pr(Z \leq 2.7)\} \\ &= 2\{1 - \Phi(2.7)\} \\ &= 0.007. \end{aligned}$$

- d) Ici on a $y_{1/2} = \lambda^{-1} \ln 2$, $y_{1/4} = \lambda^{-1} \ln(4/3)$, $y_{3/4} = \lambda^{-1} \ln 4$, $\text{IQR} = \lambda^{-1} \ln 3$, et

$$Q_{\text{high}} = y_{3/4} + 1.5 \times \text{IQR} = \lambda^{-1}(\ln 4 + 1.5 \ln 3) \approx 3\lambda^{-1},$$

mais

$$Q_{\text{low}} = y_{1/4} - 1.5 \times \text{IQR} = \lambda^{-1}(\ln(4/3) - 1.5 \ln 3) \approx -1.36\lambda^{-1} < 0.$$

Ainsi la probabilité que $Y \notin (Q_{\text{low}}, Q_{\text{high}})$ est $\Pr(Y > Q_{\text{high}}) \approx \exp(-3) \approx 0.05$.

Solution 46. On va mesurer une variable sur n individus indépendants, dont n_1 vont recevoir un traitement, T , et $n_2 = n - n_1$ un placebo, P . Supposons que la variance d'un individu qui reçoit le placebo est σ_P^2 et que celle d'un individu qui reçoit le traitement est $\sigma_T^2 = \rho^2 \sigma_P^2$. Comment choisir n_1 pour minimiser la variance de la différence des moyennes,

$$D = \bar{X}_T - \bar{X}_P = \frac{1}{n_T} \sum_{j=1}^{n_T} X_{T,j} - \frac{1}{n_P} \sum_{j=1}^{n_P} X_{P,j}?$$

- Soit $X_j \stackrel{\text{iid}}{\sim} (\mu + \delta, \sigma_T^2)$ pour les individus traités et soit $X_j \stackrel{\text{iid}}{\sim} (\mu, \sigma_P^2)$ pour les autres. Si on écrit $n_T = nt$, $n_P = n(1 - t)$, $t \in (0, 1)$, on obtient

$$\begin{aligned} E(D) &= E(\bar{X}_T) - E(\bar{X}_P) = (\mu + \delta) - \mu = \delta, \\ \text{var}(D) &= \text{var}(\bar{X}_T) + \text{var}(\bar{X}_P) = \sigma_T^2/n_T + \sigma_P^2/n_P = n^{-1}\sigma_P^2\{\rho^2/t + 1/(1-t)\}, \end{aligned}$$

et on minimise la variance par rapport à t quand $\rho^2/t^2 = 1/(1-t)^2$, c'est-à-dire $t = \rho/(1 + \rho)$. Ainsi on devrait prendre $n_T = \rho n/(1 + \rho)$, $n_P = n/(1 + \rho)$, alors $\text{var}(D) \approx (1 + \rho)^2 \sigma_P^2/n$.

Mettre $\rho = 1$ donne $\sigma_P^2 = \sigma_T^2$ on retrouve le résultat de l'exemple 132 du classe.

Avec $n = 200$ et $\rho = 2$, on a $n_T = n\rho/(1 + \rho) \approx 133$ et donc $n_P \approx 67$.

Solution 47.

- a) Suivant les arguments du cours, on voit que si les plus petites des $n\alpha + 1$ des n valeurs sont corrompues par l'addition de chiffres grands et négatifs, $\bar{y}_\alpha$ peut partir vers $-\infty$. Donc le point de rupture est $\lim_{n \rightarrow \infty} (n\alpha + 1)/n = \alpha$, ou $100\alpha\%$. Pour $\alpha = 0$ et 0.5 on retrouve la moyenne arithmétique et la médiane, dont les points de ruptures sont 0% et 50% , en accord avec le cours.
- b) Puisque le coefficient de la corrélation dépend de la moyenne, son point de rupture devrait être 0% , mais il doit rester dans l'intervalle $r_{xy} \in [-1, 1]$. Dans ce cas il serait plus approprié de modifier la définition de façon que 'la statistique tends vers les limites de son domaine', ici ± 1 plutôt que $\pm \infty$.

Solution 48. a) Les tirages sont dépendants. En effet, si n unités sont sélectionnées par hasard et sans remplacement à partir d'une population finie de taille N , le nombre total possible d'échantillons est $\binom{N}{n}$, et la probabilité de sélectionner un de ces échantillons est $1/\binom{N}{n}$, car ils sont équi-probables. Si les n unités sélectionnées dans l'échantillon sont $x_1, \dots, x_n$ et $I_j = I(\text{individu } j \text{ est sélectionné})$, alors la probabilité de leur sélection est

$$\begin{aligned} \Pr(I_1 = \dots = I_n = 1) &= \Pr(I_1 = 1) \times \prod_{j=2}^n \Pr(I_j = 1 \mid I_1 = \dots = I_{j-1} = 1) \\ &= \frac{n}{N} \prod_{j=2}^n \left(\frac{n-j+1}{N-j+1} \right) = \frac{n!(N-n)!}{N!} = \frac{1}{\binom{N}{n}}. \end{aligned}$$

- b) Soit $\Pr_i(j)$ la probabilité de sélection de x_i lors du j ème tirage, alors,

$$\begin{aligned} &\text{— } E[I_j^2] = E[I_j] = 1 \times \Pr(I_j = 1) = \frac{n}{N}, \text{ pour } j = 1, \dots, N, \\ &\text{— } \text{var}(I_j) = \frac{n}{N} - \left(\frac{n}{N}\right)^2 = \frac{n}{N}\left(1 - \frac{n}{N}\right), \text{ pour } j = 1, \dots, N, \\ &\text{— } E(I_i I_j) = 1 \times \Pr(I_i = 1, I_j = 1) = \frac{n(n-1)}{N(N-1)}, \text{ et} \\ &\text{— } \text{cov}(I_j, I_i) = E(I_i I_j) - E(I_i)E(I_j) = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = -\frac{n(N-n)}{N^2(N-1)}. \end{aligned}$$

La covariance est négative car la sélection dans un échantillon de l'individu i sans remplacement rend moins probable la sélection de l'individu j . La corrélation est de $-1/(N-1)$, qui devient minime si N est très grand, car l'échantillonnage avec et sans remise sont alors presque identiques.

- c) Par linéarité de l'espérance on a

$$E(\bar{Y}) = \frac{1}{n} \sum_{i=1}^N x_i E(I_i) = \frac{1}{n} \sum_{i=1}^N x_i \frac{n}{N} = \frac{1}{N} \sum_{i=1}^N x_i = \bar{x}.$$

Pour trouver la variance on note que puisque $\bar{x}$ est constante et $\sum_{j=1}^N I_j = n$,

$$\begin{aligned}\text{var}(\bar{Y}) &= \text{var}(\bar{Y} - \bar{x}) \\ &= \text{var} \left\{ n^{-1} \sum_{j=1}^N I_j (x_j - \bar{x}) \right\} \\ &= \frac{1}{n^2} \sum_{j=1}^N (x_j - \bar{x})^2 \text{var}(I_j) + \frac{1}{n^2} \sum_{i=1}^N \sum_{j \neq i}^N (x_i - \bar{x})(x_j - \bar{x}) \text{cov}(I_i, I_j).\end{aligned}$$

Mais puisque $\sum_{j=1}^N (x_j - \bar{x}) = 0$, on a

$$\sum_{i=1}^N (x_i - \bar{x}) \sum_{j \neq i}^N (x_j - \bar{x}) = \sum_{i=1}^N (x_i - \bar{x}) \times -(x_i - \bar{x}) = - \sum_{i=1}^N (x_i - \bar{x})^2 = -(N-1)s^2.$$

Ainsi

$$\begin{aligned}\text{var}(\bar{Y}) &= \frac{(N-1)s^2}{n^2} \{ \text{var}(I_j) - \text{cov}(I_i, I_j) \} \\ &= \frac{(N-1)s^2}{n^2} \frac{n(N-n)}{N^2} \left(1 - \frac{1}{N-1} \right) \\ &= \frac{N-n}{nN} s^2 \\ &= (1-f)n^{-1}s^2\end{aligned}$$

comme annoncé. Si $n \approx N$, alors $1-f \approx 0$, car si mon échantillon approche la population, $\text{var}(\bar{Y}) \approx 0$.