T. D. n II. Intervalles de confiance

Exercice n°1.

Le gérant d'une pisciculture souhaite connaître la taille moyenne de la population de poissons d'élevage dans ses bassins. Avec un échantillon de $n = 64$ individus, il obtient une taille moyenne de $32\,cm$. D'autres études sur des piscicultures similaires donnent un écart-type de $12\,cm$.

1. Donner les intervalles de confiance à 90%, 95% et 99% de la moyenne. Que remarque-t-on?

Exercice n°2. Cas d'une proportion

Soit X une variable aléatoire de Bernoulli telle que $P(X = 1) = p$ et $P(X = 0) = 1 - p$ où $0 < p < 1$.

1. Tracer la distribution de X, calculer $E(X)$ et $V(X)$.
2. Soit $(X_1, \cdots, X_n)$ un échantillon théorique de taille n de la variable X. Soit $\hat{P}_n = \frac{X_1+...+X_n}{n}$ un estimateur de p. Déterminer $E\left(\hat{P}_n\right)$ et $V\left(\hat{P}_n\right)$. Qu'en déduire?
3. Dans un échantillon de $n = 100$ poissons de la pisciculture précédente on a détecter 46 individus porteurs d'un parasite. Donner les IC_{90}, IC_{95} et IC_{99} de la proportion d'individus parasités dans la population.

Exercice n°3. Cas de petits échantillons

Une machine remplit des sacs automatiquement. On mesure 5 sacs et on obtient : $2.8\,kg$,$3\,kg$,$2.7\,kg$,$3.3\,kg$,$3.1\,kg$.

1. Donner l'IC_{95} du poids moyen des sacs.

Corrections

Correction exercice n°1.

Ici, l'écart type de la population est donné : $\sigma = 12\,cm$. Soit T, la variable taille. Pour évaluer l'intervalle de confiance de l'espérance μ de la population, nous allons supposer que $T \rightsquigarrow \mathcal{N}\left(\mu, \sigma^2\right)$. Soit $\overline{T} = \frac{1}{n}\sum_{i=1}^{n} T_i$, l'estimateur de cette moyenne. Le Théorème Central Limite stipule que si l'on dispose de n variables aléatoires $i.i.d.$ $\{X_1, \cdots, X_n\}$ de même loi mère qu'une variable $X \rightsquigarrow \mathcal{N}\left(\mu, \sigma^2\right)$, alors la variable $S_n = \sum_{k=1}^{n} X_k$ est telle que $S_n \rightsquigarrow \mathcal{N}\left(n\mu, n\sigma^2\right)$. Dans notre cas, la variable aléatoire $\overline{T}$ suit donc une loi normale de même espérance que la variable T et de variance dépendante de la taille n de l'échantillon :

$$\overline{T} \rightsquigarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Nous savons que la variable centrée-réduite Z suit une loi normale de moyenne nulle et de variance unité :

$$Z = \frac{\overline{T} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow \mathcal{N}(0,1).$$

Soit

$$P\left(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}\right) = 1 - \alpha,$$

la probabilité que la variable Z soit comprise dans l'intervalle $\left[z_{\alpha/2}; z_{1-\alpha/2}\right]$, symétrique autour de 0, où $z_{\alpha/2}$ et $z_{1-\alpha/2}$ sont les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la gaussienne centrée-réduite. Cette probabilité dépend du risque α de ne pas appartenir à cet intervalle, risque choisi arbitrairement petit. On déduit de l'égalité précédente que :

$$P\left(\overline{T} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \mu \leq \overline{T} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) = 1 - \alpha.$$

Cette relation permet de borner la valeur de μ entre deux valeurs

$$m_l = \overline{T} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \text{ et } m_r = \overline{T} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} = \overline{T} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$$

qui formeront les bornes de l'intervalle de confiance $IC_{1-\alpha} = [m_l;\, m_r]$. Ces bornes qui dépendent de la variable aléatoire $\overline{T}$ sont donc aléatoires et leur valeur variera en fonction de la valeur $\overline{t}_{obs}$ obtenue à partir d'une réalisation de l'échantillon. En fait, l'énoncé nous indique que $\overline{t}_{obs} = 32\,cm$. Si l'on choisit $\alpha = 0.05$, alors $z_{0.025} = -1.96$ et $z_{0.975} = 1.96$ (par symétrie de la gaussienne) et la probabilité pour que μ appartienne à l'intervalle $\left[32 - \frac{12}{\sqrt{64}} \times 1.96;\, 32 + \frac{12}{\sqrt{64}} \times 1.96\right] = [29.06;\, 34.94]$ est de 0.95. Nous venons de construire l'$IC_{0.95}$. Si l'on fait varier la valeur de α alors les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ varieront également. Pour $\alpha = 0.1$ alors $z_{0.05} = -1.64$, $z_{0.95} = 1.64$ et l'intervalle de confiance devient

$$IC_{0.9} = \left[32 - \frac{12}{\sqrt{64}} \times 1.64;\, 32 + \frac{12}{\sqrt{64}} \times 1.64\right] = [29.54;\, 34.46].$$

Pour $\alpha = 0.01$ alors $z_{0.005} = -2.57$, $z_{0.995} = 2.57$ et l'intervalle de confiance devient

$$IC_{0.99} = \left[32 - \frac{12}{\sqrt{64}} \times 2.57;\, 32 + \frac{12}{\sqrt{64}} \times 2.57\right] = [28.14;\, 35.85].$$

Si l'on note $LIC_{1-\alpha}$ la longueur de l'$IC_{1-\alpha}$, on se rend compte que $LIC_{0.99} > LIC_{0.95} > LIC_{0.90}$: l'intervalle de confiance est d'autant plus grand que le risque α diminue.

Correction exercice n°2.

On montre facilement que

$$E(X) = \sum_{x \in \mathbb{N}} xP(X = x) = 0 \times P(X = 0) + 1 \times P(X = 1) = p.$$

De même,

$$V(X) = E\left(X^2\right) - E^2(X) = p - p^2 = p \times (1 - p).$$

Le tracé de cette loi discrète est laissé à votre discrétion.

Soit l'estimateur de la quantité p donné par

$$\hat{P}_n = \frac{X_1 + \cdots + X_n}{n}.$$

Son espérance et sa variance se calcule grâce au fait que l'on suppose que l'échantillon est $i.i.d$ et tel que $X_i \rightsquigarrow \mathcal{B}(p)$. On a donc

$$E\left(\hat{P}_n\right) = \frac{1}{n}\sum_{k=1}^{n} E(X_i) = p,$$
$$V\left(\hat{P}_n\right) = \frac{1}{n^2}\sum_{k=1}^{n} V(X_i) = \frac{p(1-p)}{n}.$$

On en déduit que $\hat{P}_n$ est un estimateur sans biais et convergent. Le théorème de Moivre-Laplace stipule que, pour n assez grand, si une variable $X_n \rightsquigarrow \mathcal{B}(n,p)$ ($E(X_n) = np$, $V(X_n) = np(1-p)$) alors la variable

$$Z_n = \frac{X_n - np}{\sqrt{np(1-p)}} \rightsquigarrow \mathcal{N}(0,1).$$

Ici, si on suppose que l'échantillon est de grande taille (en pratique $n \geq 30$) alors l'estimateur est une variable aléatoire gaussienne telle que

$$\hat{P}_n \rightsquigarrow \mathcal{N}\left(p, \frac{p(1-p)}{n}\right).$$

Ici $n = 100$, nous sommes dans un cas d'application du résultat précédent. Considérons la variable centrée-réduite

$$Z = \frac{\hat{P}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow \mathcal{N}(0,1).$$

Soit

$$P\left(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}\right) = 1 - \alpha,$$

la probabilité que la variable Z soit comprise dans l'intervalle $\left[z_{\alpha/2}; z_{1-\alpha/2}\right]$, symétrique autour de 0, où $z_{\alpha/2}$ et $z_{1-\alpha/2}$ sont les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la gaussienne centrée-réduite. Cette probabilité dépend du risque α de ne pas appartenir à cet intervalle, risque choisi arbitrairement petit. On déduit de l'égalité précédente que :

$$P\left(\hat{P}_n - \sqrt{\frac{p(1-p)}{n}} z_{1-\alpha/2} \leq p \leq \hat{P}_n - \sqrt{\frac{p(1-p)}{n}} z_{\alpha/2}\right) = 1 - \alpha.$$

Cette relation permet de borner la valeur de p entre deux valeurs

$$p_l = \hat{P}_n - \sqrt{\frac{p(1-p)}{n}} z_{1-\alpha/2}$$

et

$$p_r = \hat{P}_n - \sqrt{\frac{p(1-p)}{n}} z_{\alpha/2},$$

qui formeront les bornes de l'intervalle de confiance $IC_{1-\alpha} = [p_l; p_r]$. Le problème est que les bornes de cet intervalle dépendent du paramètre p que l'on cherche à borner. On va en fait substituer la valeur de p et celle de $\hat{P}_n$ dans les bornes en utilisant la valeur $p_{obs} = \frac{46}{100}$ obtenue avec l'échantillon. On peut montrer qu'asymptotiquement cette approximation est admissible. On obtient ainsi les trois intervalles :

$$\begin{cases} IC_{0.90} = \left[p_{obs} - \sqrt{\frac{p_{obs}(1-p_{obs})}{n}} z_{0.95};\ p_{obs} - \sqrt{\frac{p_{obs}(1-p_{obs})}{n}} z_{0.05}\right] = [0.378;\ 0.541] \\ IC_{0.95} = \left[p_{obs} - \sqrt{\frac{p_{obs}(1-p_{obs})}{n}} z_{0.975};\ p_{obs} - \sqrt{\frac{p_{obs}(1-p_{obs})}{n}} z_{0.025}\right] = [0.362;\ 0.557] \\ IC_{0.99} = \left[p_{obs} - \sqrt{\frac{p_{obs}(1-p_{obs})}{n}} z_{0.995};\ p_{obs} - \sqrt{\frac{p_{obs}(1-p_{obs})}{n}} z_{0.005}\right] = [0.331;\ 0.588] \end{cases}.$$

De la même manière que dans l'exercice précédent, plus α augmente, plus l'intervalle de confiance diminue.

Correction exercice n°3.

On a ici un échantillon de taille $n = 5$ et aucun renseignement sur les paramètres populationnels. Nous allons supposer que la variable X: poids des sacs (kg) est une variable aléatoire gaussienne telle que

$$X \rightsquigarrow \mathcal{N}\left(\mu, \sigma^2\right).$$

Les paramètres μ et σ sont inconnus ici. On doit les estimer avec un échantillon $\{X_1, \cdots, X_n\}$ de variables supposées $i.i.d$ de même loi mère que X. Soit les estimateurs suivant :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,\ S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right),^2$$

ceux de μ et σ^2 (version sans biais et empirique) respectivement. Soit la variable aléatoire

$$T = \frac{Y}{\sqrt{\frac{U}{n}}}$$

où $Y \rightsquigarrow \mathcal{N}(0,1)$ est gaussienne centrée-réduite et $U \rightsquigarrow \chi_n^2$ suit une loi du Chi2 à n degrés de liberté, Y et U indépendante. La variable T suit dans ce cas une loi de Student à n degrés de liberté

$$T \rightsquigarrow \mathcal{T}_n.$$

Or on sait d'autre part que la variable $\frac{n-1}{\sigma^2}S_{n-1}^2 = \frac{n}{\sigma^2}S_n^2$ est distribuée selon une loi du χ_{n-1}^2 à $n-1$ degrés de liberté. Rappelons que si $X_i \sim \mathcal{N}(0,1)$ alors $X_1^2 + \cdots + X_n^2 \sim \chi_n^2$. Donc si l'on considère la variable

$$Z = \frac{\overline{X} - \mu}{\sqrt{\frac{S_{n-1}^2}{n}}} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S_{n-1}^2}{\sigma^2} \times \frac{1}{n-1}}}$$

elle correspond bien au rapport d'une gaussienne centrée réduite sur la racine carrée d'une variable qui suit une loi χ_{n-1}^2 divisée par son nombre de dégrés de liberté $n-1$. Si l'on considère également la variable

$$Z' = \frac{\overline{X} - \mu}{\sqrt{\frac{S_n^2}{n-1}}} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{nS_n^2}{\sigma^2} \times \frac{1}{n-1}}}$$

elle correspond également à une loi du χ_{n-1}^2.On en déduit que :

$$Z \rightsquigarrow \mathcal{T}_{n-1}.$$

Soit

$$P\left(t_{n-1;\alpha/2} \leq Z \leq t_{n-1;1-\alpha/2}\right) = 1 - \alpha,$$

la probabilité que la variable Z soit comprise dans l'intervalle $\left[t_{n-1;\alpha/2}; t_{n-1;1-\alpha/2}\right]$, symétrique autour de 0, où $t_{n-1;\alpha/2}$ et $t_{n-1;1-\alpha/2}$ sont les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi de Student à $n-1$ degrés de liberté. Cette probabilité dépend du risque α de ne pas appartenir à cet intervalle, risque choisi arbitrairement petit. On déduit de l'égalité précédente que :

$$P\left(\overline{X} - \frac{S_{n-1}}{\sqrt{n}}t_{n-1;1-\alpha/2} \leq \mu \leq \overline{X} - \frac{S_{n-1}}{\sqrt{n}}t_{n-1;\alpha/2}\right) = 1 - \alpha.$$

Cette relation permet de borner la valeur de μ entre deux valeurs $m_l = \overline{X} - \frac{S_{n-1}}{\sqrt{n}}t_{n-1;1-\alpha/2}$ et $m_r = \overline{X} - \frac{S_{n-1}}{\sqrt{n}}t_{n-1;\alpha/2}$, qui formeront les bornes de l'intervalle de confiance $IC_{1-\alpha} = [m_l;\ m_r]$. Ces bornes qui dépendent de la variable aléatoire $\overline{X}$ sont donc aléatoires et leur valeur variera en fonction de la réalisation de l'échantillon. On calcule $\overline{x}_{obs} = 2.98\,kg$ et $4 \times s_{obs}^2 = 0.285$. Si l'on choisit $\alpha = 0.05$, alors $t_{4;0.025} = -2.77$ et $t_{4;0.975} = 2.77$ (par symétrie de la loi de Student) et la probabilité pour que μ appartienne à l'intervalle

$$\left[2.98 - \sqrt{\frac{0.285}{4\times 5}} \times 2.77;\ 2.98 + \sqrt{\frac{0.285}{4\times 5}} \times 2.77\right] = [2.78;\ 3.18]$$

est de 0.95. Nous venons de construire l'$IC_{0.95}$ pour de petits échantillons, quand la moyenne et la variance de la population sont inconnues et doivent être estimées.