DELEGUE Hugo - TANIEL Rémi GIS2A3

Compte rendu TP2 : Intervalles de Confiance

Introduction

Nous avons un jeu de données représentant le poids à la naissance des nouveau-nés d'une maternité en 1 année, la valeur est indiquée en gramme. Ces données sont censées suivre une loi normale $N(m, \sigma)$.
L'objectif du TP est donc de construire des Intervalles de Confiance (IC)

Question 1

On se décide de calculer la moyenne empirique $\overline{X}$ ainsi que la variance empirique S^2 du poids à la naissance pour l'ensemble du jeu de donnée. Pour rappel :

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X_n})^2$$

Nous obtenons les résultats suivants :

$\overline{X_n} = 3125,6$

$S_n^2 = 191413,1$

Question 2

On se décide de créer une fonction paramétrée par notre jeu de donnée en entier n renvoyant un échantillon aléatoire de taille n $(X_1, X_2, ..., X_n)$ provenant de notre jeu de donnée, ainsi que la moyenne empirique $\overline{X_n}$ et la variance empirique S_n^2 de cet échantillon.

On obtient les résultats suivants selon la valeur de $n = 30, 60$ et 80 :

n = 30

$\overline{X_{30}} = 2981.17$

$S_{30}^2 = 170166.54$

n = 60

$\overline{X_{60}} = 3179.23$
$S_{60}^2 = \ 195895.78$

n = 80

$\overline{X_{80}} = 3087.4$
$S_{80}^2 = 193804.5$

Question 3

Nous créons maintenant une fonction qui à partir d'un vecteur de données normales, calcule un intervalle de confiance de niveau de risque α pour la moyenne m en supposant la variance σ connue.

Pour rappel $X_i \sim N(m, \sigma)$ donc

$$E(\overline{X_n}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}E(\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} \times n \times m = m$$

$$V(\overline{X_n}) = V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}V(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} \times n \times \sigma = \frac{\sigma}{n}$$

On en déduit que

$$\overline{X_n} \sim N(m, \frac{\sigma}{n})$$

De ce fait, nous pouvons déterminer la statistique de test suivant :

$$\frac{\overline{X_n} - E(\overline{X_n})}{V(\overline{X_n})} = \frac{\overline{X_n} - m}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

et

$$P(-u_{1-\frac{\alpha}{2}} \leq \frac{\overline{X_n} - m}{\frac{\sigma}{\sqrt{n}}} \leq u_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(\overline{X_n} - \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}} \leq m \leq \overline{X_n} + \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

Notre intervalle de confiance de niveau de risque α est :

$$IC \ = \ [\overline{X_n} - \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}} ; \overline{X_n} + \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}]$$

Pour les 3 échantillons aléatoires générés précédemment, nous obtenons :

n = 30

$IC = [2976.14 \ ; \ 2986.03]$

n = 60

$IC = [3175.68 \ ; \ 3182.73]$

n = 80

$IC = [3084.32 \ ; \ 3090.44]$

Question 4

Le but de ce question est le même que la question précédente, c'est à dire de construire un intervalle de confiance de niveau de risque α pour la moyenne m mais en supposant la variance σ inconnue. On utilisera donc une statistique de test similaire à la question précédente en remplaçant sigma par la variance corrigée que l'on nommera S_n

$$\frac{\overline{X_n}-m}{S_n\frac{1}{\sqrt{n}}}$$

Pour un n très grand, on peut supposer que notre statistique de test suit toujours une loi normal:

$$\frac{\overline{X_n}-m}{S_n\frac{1}{\sqrt{n}}} \sim N(0,1)$$

On construit donc notre intervalle de confiance:

$$P(-u_{1-\frac{\alpha}{2}} \leq \frac{\overline{X_n}-m}{S_n\frac{1}{\sqrt{n}}} \leq u_{1-\frac{\alpha}{2}}) = 1-\alpha$$

$$P(\overline{X_n} - \overline{S_n\frac{1}{\sqrt{n}}u}_{1-\frac{\alpha}{2}} \leq m \leq \overline{X_n} + \overline{S_n\frac{1}{\sqrt{n}}u}_{1-\frac{\alpha}{2}}) = 1-\alpha$$

Notre intervalle de confiance de niveau de risque α est :

$$IC = [\overline{X_n} - \overline{S_n\frac{1}{\sqrt{n}}u}_{1-\frac{\alpha}{2}} ; \overline{X_n} + \overline{S_n\frac{1}{\sqrt{n}}u}_{1-\frac{\alpha}{2}}]$$

Pour les 3 échantillons aléatoires générés précédemment, nous obtenons :

n = 30

$IC = [2642.78 \ ; \ 3308.81]$

n = 60

$IC = [2981.07 \ ; \ 3374.18]$

n = 80

$IC = [2939.71 \ ; \ 3233.28]$

Cependant, on peut également dire que cette statistique de test suit une loi de student à n-1 degré de liberté :

$$\frac{\overline{X_n}-m}{S_n\frac{1}{\sqrt{n}}} \sim \tau_{n-1}$$

On pose

$$P(-t_{1-\frac{\alpha}{2},n-1} \leq \frac{\overline{X_n}-m}{S_n\frac{1}{\sqrt{n}}} \leq t_{1-\frac{\alpha}{2},n-1}) = 1-\alpha$$

$$P(\overline{X_n} - \overline{S_n}\frac{1}{\sqrt{n}}t_{1-\frac{\alpha}{2},n-1} \leq m \leq \overline{X_n} + \overline{S_n}\frac{1}{\sqrt{n}}t_{1-\frac{\alpha}{2},n-1}) = 1-\alpha$$

Notre intervalle de confiance de niveau de risque α :

$IC = [\overline{X_n} - \overline{S_n}\frac{1}{\sqrt{n}}t_{1-\frac{\alpha}{2},n-1} ; \overline{X_n} + \overline{S_n}\frac{1}{\sqrt{n}}t_{1-\frac{\alpha}{2},n-1}]$

Pour les 3 échantillons aléatoires générés précédemment, nous obtenons :

n = 30

$IC = [2639.83\ ;\ 3311.66]$

n = 60

$IC = [2980.22\ ;\ 3375.01]$

n = 80

$IC = [2939.24\ ;\ 3233.74]$

On effectue ensuite la même procédure grâce à la fonction t.test de R, on obtient les résultats suivant:

n = 30

$IC = [2824.50\ ;\ 3137.83]$

n = 60

$IC = [3063.93\ ;\ 3294.53]$

n = 80

$IC = [2988.81\ ;\ 3185.99]$

On remarque que les intervalles obtenues grâce à la fonction t.test sont plus courts que ceux obtenus manuellement.

Question 5

Cette fois nous devons déterminer un intervalle de confiance à 95% pour une proportion de carton de lait défectueux, dans notre cas lors de la livraison, sur les 197 cartons de lait, seuls 177 respectent les normes d'hygiène de la maternité.

On pose $\widehat{p} = \frac{20}{197}$

On utilise la statistique de test suivante :

$\frac{\widehat{p}-p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \sim N(0, 1)$

On pose

$P(-u_{1-\frac{\alpha}{2}} \leq \frac{\widehat{p}-p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \leq u_{1-\frac{\alpha}{2}}) = 1 - \alpha$

$P(\widehat{p} - \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}u_{1-\frac{\alpha}{2}} \leq p \leq \widehat{p} + \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}u_{1-\frac{\alpha}{2}}) = 1 - \alpha$

Notre intervalle de confiance de niveau de risque α est donc le suivant :

$$IC = [\widehat{p} - \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} u_{1-\frac{\alpha}{2}} ; \widehat{p} + \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} u_{1-\frac{\alpha}{2}}]$$

Pour $\alpha = 0,05$ (intervalle de confiance à 95%), on obtient :

$IC = [0.1002\ ;\ 0.1029]$

On effectue ensuite la même démarche en utilisant cette fois la fonction binom.test de R:

Pour $\alpha = 0,05$ (intervalle de confiance à 95%), on obtient :

$IC' = [0.0631\ ;\ 0.1524]$

On remarque que l’intervalle de confiance obtenu par la fonction binom.test est beaucoup plus large que celui obtenu manuellement.