Université Grenoble Alpes
L2/STA351

TP4 : Intervalles de confiance

Objectifs : Savoir calculer un intervalle de confiance pour une moyenne μ ou une probabilité p et comprendre ce qu'est le niveau de confiance de l'intervalle.

Exercice 1

Dans cet exercice, on suppose que X est une variable gaussienne, dont on connait la variance σ^2 mais pas la moyenne μ.

1. Simuler un échantillon gaussien de X de taille $n = 50$, de moyenne $\mu = 1$ (`mu <- 1`) et d'ecart-type $\sigma = 1$. Calculer sa moyenne empirique `xbar`.
2. On suppose maintenant ne pas connaître la valeur de μ et on va calculer un intervalle de confiance pour μ de niveau de confiance $1 - \alpha$. Fixer α à la valeur 5% (`alpha <- 0.05`).
3. Calculer la borne inférieure et la borne supérieure de l'intervalle de confiance pour μ de niveau 1- α en supposant $\sigma = 1$ connu et en se servant de la fonction quantile pour la loi normale `qnorm`. Appeler `borneInf` et `borneSup` ces deux bornes, et les afficher avec `c(borneInf, borneSup)`
4. Calculer les bornes de l'intervalle de confiance pour μ de niveau 90% et 99%. Commenter la largeur de l'intervalle en fonction de la confiance. Cela vous parait-il logique ?
5. Un intervalle de confiance n'inclut pas toujours la vraie valeur. Pour le vérifier, nous allons simuler 100 échantillons et calculer leurs intervalles de confiance.
   (a) Simuler $N = 100$ échantillons gaussiens de taille $n = 50$ de X de moyenne $\mu = 1$ et d'ecart-type $\sigma = 1$, en les stockant dans une matrice `X100`.
   ```
   X100<-matrix(rnorm(N*n,mu,sig), N,n) # N echantillons de taille n
   ```
   (b) Calculer les moyennes empiriques de chaque échantillon et les stocker dans un vecteur `x100bar` avec la fonction `rowMeans`.
   (c) Calculer les bornes inférieures et supérieures des 100 intervalles de confiance de μ de niveau 95% (en supposant la variance $\sigma = 1$ connue). Les appeler respectivement `borneInf100` et `borneSup100`.
   (d) Tracer les 100 intervalles de confiance ainsi que la vraie valeur :
   ```
   CI100<-rbind(borneInf100, borneSup100) #100 intervalles de confiance
   matplot(CI100, rbind(1:100, 1:100), type='l', lty=1, ylab="") #graphique
   des IC
   abline(v=mu) #vraie valeur
   ```
   (e) Compter le nombre de fois où μ est plus grand que la borne supérieure :
   ```
   which(mu>borneSup100) # retourne les indices pour lesquels mu dépasse borneSup100
   length(which(mu>borneSup100)) # retourne le nombre d'indices
   ```
   (f) Compter le nombre de fois où μ est plus petit que la borne inférieure.

(g) En déduire le nombre de fois où l'intervalle de confiance contient la vraie valeur μ.

(h) Commenter.

6. On va maintenant étudier la précision d'un intervalle de confiance en fonction de la taille d'un échantillon.

   (a) Simuler un échantillon gaussien de X de taille $n = 1000$ de moyenne $\mu = 0$ et d'écart-type $\sigma = 1$ et l'affecter à `x`.

   (b) Calculer les bornes de l'intervalle de confiance de μ au niveau 95% quand on utilise seulement les 100 premières valeurs de l'échantillon simulé.X : `x100<-x[1:100]`.

   (c) Calculer les bornes de l'intervalle de confiance de μ au niveau 95% quand on utilise seulement les 500 premières valeurs de `x`.

   (d) Calculer les bornes de l'intervalle de confiance de μ au niveau 95% quand on utilise toutes les valeurs de `x`.

   (e) Tracer sur un même graphique les bornes des intervalles de confiance, ainsi que la vraie valeur. Commenter la largeur de l'intervalle en fonction de la taille de l'échantillon. Cela vous parait-il logique ?

Exercice 2

On utilise la base de données `ferretti.csv`, fournie par le Professeur Gilbert Ferretti. Elle contient 43 patients avec une tumeur du poumon et les variables

— `height` : taille de la tumeur en mm
— `diameter` : diamètre de la tumeur en mm
— `density` : densité de la tumeur `negative` (si moins dense que l'eau), `null` (si aussi dense que l'eau), `positive` (si plus dense que l'eau)
— `invasive` : si la tumeur est invasive ou non (`yes` ou `no`)

1. Charger la base de données `ferretti.csv` et affecter à la variable `H` la taille de la tumeur.
2. On peut calculer l'intervalle de confiance pour la moyenne de `H`, notée μ_H, en supposant sa variance inconnue en utilisant la fonction `t.test`

   ```
   t.test(H)$conf.int
   ```

   On peut changer le niveau de confiance avec le paramètre `conf.level` :

   ```
   t.test(H,conf.level=0.9)$conf.int
   t.test(H,conf.level=0.99)$conf.int
   ```

3. Comparer avec les résultats obtenus en calculant vous-meme l'intervalle de confiance pour $\alpha = 0.1$ par exemple.
4. On s'intéresse maintenant à la variable diamètre de la tumeur `diameter`. On appellera `D` les données de cette variable. Calculer les statistiques descriptives de ces nouvelles données.
5. On suppose que `D` est un échantillon gaussien. Calculer les bornes inférieure et supérieure de l'intervalle de confiance de μ_D de niveau 90%, 95%, 99% avec la fonction `t.test`. Commenter.
6. Calculer les intervalles de confiance au niveau 95% de μ_D chez les individus ayant une tumeur non-invasive, puis chez les individus avec une tumeur invasive. Commenter.

Exercice 3

On étudie l'influence des rayons X sur la spermatogenèse de Bombyx mori. Les mâles ont été exposés à des radiations le jour 2 et 4 de leur état larvaire. Ces mâles ont été accouplés à des femelles non exposées et le nombre d'oeufs fertiles pondus par femmes a été compté. Parmi un total de 5636 oeufs, 4998 étaient fertiles. Dans un groupe contrôle de males non exposés, parmi un total de 6221 oeufs, 5834 étaient fertiles.

1. Rentrer les données

   ```
   x <- 4998; n <- 5636
   ```

2. Quelle loi proposez vous pour modéliser cet évènement ? Proposer une approximation pour cette loi.

3. On veut calculer un intervalle de confiance de niveau 95% pour la proportion d'oeufs fertiles après exposition aux radiations des males. On peut utiliser une approximation normale pour ce calcul :

   ```
   Fexpo <- x/n
   sig <- sqrt(Fexpo*(1-Fexpo))
   Fexpo - qnorm(0.975)*sig/sqrt(n)
   Fexpo + qnorm(0.975)*sig/sqrt(n)
   ```

4. On peut aussi calculer cet intervalle de façon exacte avec la loi binomiale. On l'obtient automatiquement dans R avec la commande suivante :

   ```
   prop.test(x=x, n=n)$conf.int
   ```

   Comparer les deux intervalles et commenter

5. Calculer un intervalle de confiance de niveau 95% pour la proportion d'oeufs fertiles chez les males controles.

6. Est ce que les deux intervalles se chevauchent ? Qu'en déduisez vous sur l'influence de l'exposition à la radiation sur la fertilité ?

Exercice 4

La concentration fluorescente d'une solution a été mesurée 90 fois. A partir de ces mesures, la moyenne empirique a été calculée et vaut 4.38 mg/k et l'écart-type exact vaut 0.08 mg/l.

1. Calculer l'intervalle de confiance de la vraie concentration de la solution, à un niveau de confiance de 95% puis de 99%.

2. Pour quelle taille n d'échantillon obtient-ton comme intervalle calculé de niveau de confiance de 95% un intervalle centré en 4.38 et de précision ± 0.01 ?

3. Pour l'échantillon dont on dispose ici et qui est de taille $n = 90$: pour quel niveau de confiance l'intervalle calculé serait-il centré en 4.38 avec une précision de ± 0.01 ?