Intervalles de confiance et tests statistiques

Semaine 5

Table des matières

Objectifs

Les objectifs sont de comprendre les concepts d'intervalles de confiance et de tests d'hypothèse. Pour quelques jeux de données correspondant à des applications spécifiques, on construira un test et on interprétera les résultats. Les travaux dirigés introduiront la notion de niveau de confiance, de significativité d'un test, de p-valeur et le calcul de p-valeur.

Exercice 1

On considère un échantillon de taille n de loi normale de moyenne inconnue θ et de variance supposée connue $\sigma^2 = 1$.

Question 1. Vraisemblance

Calculer la fonction de vraisemblance du paramètre θ. Montrer que l'estimateur du maximum de vraisemblance de θ est la moyenne empirique.

Solution.

$L_n(\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^n exp(-\frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2)$

En appliquant le logarithme, on obtient :

$$l(\theta, x) = -\frac{n}{2} ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2$$

On annule la dérivée selon θ pour trouver :

$$\hat{\theta}_{MV} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Question 2.

On suppose que $n = 20$ et que l'échantillon provient du tirage aléatoire suivant

```
set.seed(42)
x <- rnorm(n = 20, mean = 2.4)
```

Réécrire la log-vraisemblance à l'aide de la moyenne empirique $\bar{x}_n$ et la variance empirique s_n^2.

Indication : faire apparaitre $\bar{x}_n$ dans la somme des carrés en l'ajoutant et le retranchant.

Tracer la log-vraisemblance pour des valeurs de θ entre 0 et 5.

Solution.

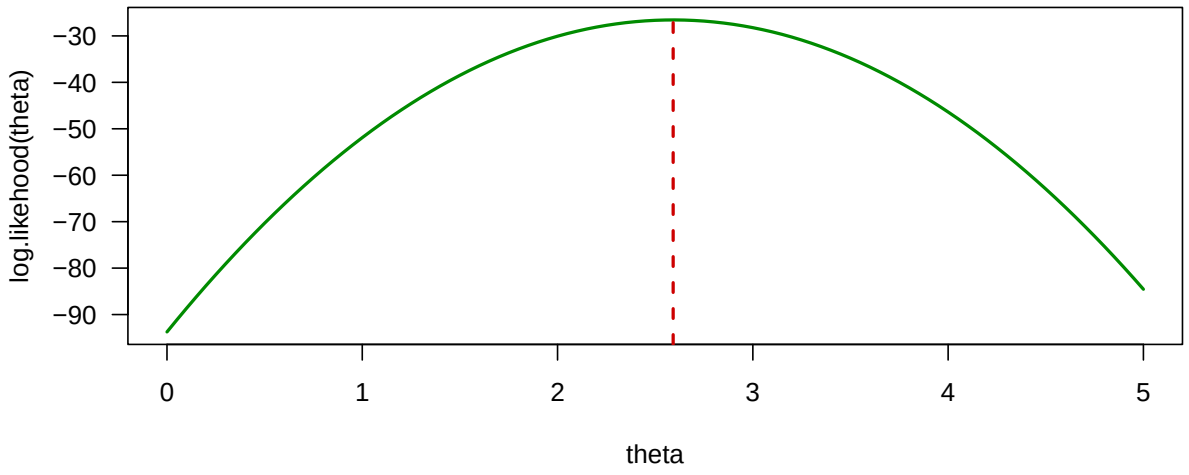
$$l_n(\theta, x) = -\frac{n}{2}\left[\ln(2\pi) + s_n^2(x) + (\overline{x_n} - \theta)^2\right]$$

```
# initialisation de  n à la longueur de l'échantillon x
n = length(x)
m= mean(x)
# définition de la moyenne empirique
x.bar = mean(x)
sn.2 = mean((x-x.bar)^2)

# Création de la fonction de log vraisemblance
log.likehood = function(theta){
  -n/2*(log(2*pi) + sn.2/2 + (x.bar-theta)**2)
}

# Tracé de la log vraisemblance et de la droite verticale y = m
curve(log.likehood, xname = "theta",
      col = "green4", lwd = 2,
      from = 0, to = 5, las = 1)

abline(v = m,
       col = "red3", lwd = 2, lty = 2)
```



Question 3. Pivot

On note $\overline{X}_n$ la moyenne empirique de l'échantillon. Démontrer que la loi de la variable aléatoire

$$Z = \sqrt{n}(\overline{X}_n - \theta)$$

est la loi normale $N(0, 1)$. En déduire un pivot pour le paramètre θ.

Solution.

De la même mannière que dans l'exercice 3 du CTD 3, on a $Z \sim N(0,1)$

Question 4. Intervalle de confiance.

À l'aide de la fonction `qnorm`, calculer les quantiles correspondant aux probabilités $\alpha/2 = 2.5\%$ et $1 - \alpha/2 = 97.5\%$. En déduire un intervalle (a, b) tel que

$$\mathbb{P}(a < Z < b) = 1 - \alpha\,.$$

Solution.

```
c(a = qnorm(2.5/100), b = qnorm(1-2.5/100))
```

```
##         a         b
## -1.959964  1.959964
```

Question 5. Intervalle de confiance

Montrer que l'intervalle

$$\text{IC} = \left(\overline{X}_n - \frac{1.96}{\sqrt{n}}, \overline{X}_n + \frac{1.96}{\sqrt{n}}\right)$$

est un intervalle de confiance au seuil $1 - \alpha = 95\%$ pour θ. Existe-il d'autres intervalles de confiance de même seuil ?

Solution. Comme $Z \sim N(0,1)$ on a :

$$\mathbb{P}\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) \leq Z \leq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

En isolant θ on en déduit que IC est un intervalle de confiance au seuil $1 - \alpha = 95\%$ pour θ. On pourrait choisir d'autre intervalles de confiance en répartissant différemment les 5% par rapport à la courbe.

Question 6. Couverture de l'IC.

On répète 10000 fois l'échantillonnage de la population effectué dans la question 2. Vérifier que la vraie valeur, $\theta = 2.4$, se trouve dans l'IC pour environ 95% des echantillons ainsi créés.

Solution. Il s'agit d'une expérience de type Monte Carlo. Compléter le code suivant.

```
theta = 2.4
n = 20

## définition de la fonction intervalle_IC
intervalle_IC <- function()
{
  x.rep <- rnorm(n, mean = theta)
  (mean(x.rep) +qnorm(2.5/100)/sqrt(n) <= theta && theta <= mean(x.rep) +qnorm(1-2.5/100)/sqrt(n))
}

## On répète 10000 fois
m = 10000
couverture.ic <- replicate(m, intervalle_IC())

## Arrondi de la moyenne à trois décimales
round(mean(couverture.ic), 3)
```

```
## [1] 0.947
```

Question 7. Test

Pour un échantillon de taille $n = 20$ dont la variance théorique est supposée égale à 1, on observe que $\overline{x}_n = 2.9$. A-t-on suffisamment de données pour affirmer avec un risque d'erreur inférieur à 5% que cet échantillon n'est pas issu de la loi normale $N(m = 2.4, 1)$?

Solution. Calculer l'IC au niveau de confiance 95% en admettant que la variance est égale à 1. Conclure.

Comme $m = 2.4 \notin [2.46; 3.34]$ on peut affirmer avec un risque d'erreur inférieur à 5% que cet échantillon n'est pas issu de la loi normale $N(m = 2.4, 1)$

Question 8. Valeur p (ou p-valeur)

En supposant que les échantillons sont issus de la loi $N(2.4, 1)$, calculer

$$p = \mathbb{P}(|\overline{X}_n - 2.4| \geq 0.5)$$

Comment interpréter cette valeur ? Rappel : on observe que $\overline{x}_n = 2.9 = 2.4 + 0.5$.

Solution.

```
2*pnorm(-sqrt(n)/2)
```

```
## [1] 0.02534732
```

la valeur 2.9 est peu probable quand H_0 est vrai donc on en concult que l'on rejette H_0 et que $m \neq 2.4$

Problème

On considère un échantillon de taille $n = 30$ représentant le temps d'attente en minutes du premier but dans un match de football du championnat de France, saison 2021-2022 (données réelles).

```
x <- c(42, 99, 30 , 28,  69,  39,  17,   9,  21,  39,  49,  28,  46,  26,  27,   2,  19, 50, 117,   9,
```

On suppose que les données sont issues de variables aléatoires indépendantes $(X_1, \ldots, X_n)$ de loi exponentielle de taux inconnu $\theta > 0$.

Question 1

Que représente $1/\theta$? Proposer une estimation du paramètre θ par la méthode des moments, puis par maximum de vraisemblance.

Solution. Rechercher la réponse sur le papier, puis calculer la valeur des estimations à l'aide d'un code R (créer un bout de code avec le bouton +c). méthode des moments : $E[X] = \frac{1}{\theta}$ donc $\theta = \frac{1}{E[X]}$ et $\hat{\theta} = \frac{1}{\overline{X}_n}$

maximum de vraisemblance : $L_n(\theta, x_i) = \prod_{i=1}^n \theta e^{-\theta x_i}$ donc $l(\theta, x_i) = \sum_{i=1}^n (ln\theta - \theta x_i) = nln\theta - \theta \sum_{i=1}^n x_i$

donc $\frac{dl(\theta,x_i)}{d\theta} = \frac{n}{\theta} - \sum_{i=1}^n x_i = 0 <=> \theta = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\overline{X}}$

```
#calcul de l'estimation de theta
1/mean(x)
```

```
## [1] 0.02982107
```

Question 2

Soit n variables aléatoires indépendantes, $(X_1, \ldots, X_n)$, issues de la loi exponentielle de paramètre θ. On pose

$$S_n = \sum_{i=1}^{n} X_i \,.$$

Montrer que la loi de la variable $Z_n = \theta S_n$ est la loi Gamma$(n, 1)$.

Solution. Rappel : la loi de la variable aléatoire S_n est la loi Gamma(n, θ). $F(Z_n < x) = F(S_n < \frac{x}{\theta})$

donc $f_{Z_n}(x) = \frac{1}{\theta} f_{S_n}(\frac{x}{\theta}) = \frac{x^{n-1}}{\Gamma(n)} e^{-x}$

donc Z_n est de loi Gamma(n,1)

Question 3

En déduire un pivot et un intervalle de confiance de niveau $1 - \alpha = 95\%$ pour le paramètre θ

$$\text{IC} = \left(\frac{q_{\alpha/2}}{S_n}, \frac{q_{1-\alpha/2}}{S_n} \right)$$

Calculer l'intervalle de confiance pour l'échantillon de données $\mathbf{x}$.

$P(a < Z_n < b) = 0,95 <=> P(a < \theta S_n < b) = 0,95 <=> P(\frac{a}{S_n} < \theta < \frac{b}{S_n}) = 0,95$ avec a et b les quantille de loi gamma(n,1)

Solution. Compléter le code suivant.

```
# calcul de Sn
sn = sum(x)

# Calcul de l'intervalle de confiance
CI = c(qgamma(0.025, 30)/sn, qgamma(0.975, 30)/sn)
names(CI) <- c("inf", "sup")
round(CI, 3)
```

```
##   inf   sup
## 0.020 0.041
```

Question 4

Calculer un intervalle de confiance de niveau $1-\alpha = 95\%$ pour le paramètre θ en effectuant une approximation gaussienne de la somme des observations.

Solution. Calculer l'intervalle de confiance à l'aide d'un code R (créer un bout de code avec le bouton +c). intervalle de confiance :

$a < \sqrt{n}\frac{\bar{X_n} - E[X]}{\sqrt{Var[X]}} < b <=> a < \sqrt{n}\frac{\frac{1}{n}S_n - \frac{1}{\theta}}{\frac{1}{\theta}} < b <=> (\frac{a}{\sqrt{n}} + 1)\frac{n}{S_n} < \theta < (\frac{b}{\sqrt{n}} + 1)\frac{n}{S_n}$ avec a et b les quantille de loi N(0,1)

```
born_inf <- 30/sn*(qnorm(0.025)/sqrt(30)+1)
born_sup <- 30/sn*(qnorm(0.975)/sqrt(30)+1)
CI = c(born_inf,born_sup)
names(CI) <- c("inf", "sup")
round(CI, 3)
```

```
##   inf   sup
## 0.019 0.040
```

Question 5. Méthode de bootstrap.

Cette question suggère de calculer un intervalle de confiance sans supposer connus les résultats sur la loi gamma, à l'aide d'une méthode de Monte Carlo (bootstrap).

Le bootstrap consiste à tirer avec remise n observations dans l'échantillon de départ (de taille n) pour obtenir m sous-échantillons sur lesquels on estime le paramètre d'intérêt. Comme dans une méthode de Monte Carlo, on peut calculer l'intervalle de confiance à partir de la loi empirique de l'estimateur.

Commmenter le code et proposer un intervalle de confiance (bilatéral) de niveau 95% à l'aide de ce code.

Solution.

```
# algorithme de bootstrap
bootestim <- function()
{
  # création d'un échantillon de taille n tiré avec remise dans x
  x.rep <- x[sample(n, replace = TRUE)]

  # calcul de l'estimation de theta
  1/mean(x.rep)
}

# reproduit l'experience 10000 fois
theta.boot <- replicate(10000, bootestim())
```

Utiliser la fonction quantile pour calculer l'intervalle de confiance à partir de l'échantillon simulé par l'algorithme de bootstrap.

```
CI = c(quantile(theta.boot, 0.025),quantile(theta.boot, 0.975))
round(CI, 3)
```

```
##  2.5% 97.5%
## 0.019 0.037
```

Question 6

Un passionné de football affirme qu'en moyenne dans un match de championnat de France, le premier but est observé après la 25eme minute.

À quelle hypothèse alternative cette affirmation correspond-elle ? Quelle hypothèse nulle choisir ? Tester l'hypothèse nulle et donner la p-valeur du test. Que conclure : avec les observations, l'hypothèse nulle est-elle rejetée au seuil 5 % ?

Solution.
On peut calculer la p-valeur à partir de la loi Gamma et de l'observation s_n. On notera qu'une fonction pivot n'est pas nécessaire. Expliquer les choix effectués. On va tester H0 $\theta = 1/25$ contre H1 $\theta < \frac{1}{25}$

```
## H1 : theta < 1/25 car mean(x) > 25 donc 1/mean(x) < 1/25
## H0 : theta = 1/25
##
#calcul probabilité que Zn > sn/25 avec Zn suivant une loi Gamma(n, 1)
pgamma(sn/25, 30, rate = 1, lower = FALSE)
```

```
## [1] 0.04001276
```

On trouve une valeur inférieur à 0.05 donc on rejette H0.

Question 7. Intervalle de confiance et test sur l'espérance de la loi exponentielle.

On note m l'espérance de la loi exponentielle. On considère désormais que m est le paramètre inconnu à estimer. Reprendre les questions précédentes (3-6) afin de

- proposer une estimation de m,
- proposer un intervalle de confiance à 95% pour m ; on pourra comparer deux méthodes : IC exact, méthode de bootstrap,
- tester l'hypothèse nulle $H_0 : m = 25$ (décider de la nature du test : unilatéral ou bilatéral, calculer une valeur p pour le test, est elle différente de celle obtenue question 6).

Solution. Réutiliser les codes précédents.

```
##Estimation de m
m = mean(x)
round(m, 3)
```

```
## [1] 33.533
```

```
## IC exact à 95%
alpha = 0.05
sn = sum(x)
n = length(x)

q1 = qgamma(1 - alpha/2, shape = n)
q2 = qgamma(alpha/2, shape = n)

IC_m = c(sn / q1, sn / q2)
names(IC_m) = c("inf", "sup")
round(IC_m, 3)
```

```
##    inf    sup
## 24.154 49.701
```

```
# Fonction bootstrap pour m
bootestim_m <- function() {
  x.rep = x[sample(n, replace = TRUE)]
  mean(x.rep)
}

# Simulations bootstrap
m.boot = replicate(10000, bootestim_m())

# IC bootstrap à 95%
IC_boot_m = quantile(m.boot, c(0.025, 0.975))
round(IC_boot_m, 3)
```

```
##   2.5%  97.5%
## 25.500 43.067
```

```
#Test H0
theta0 = 1 / 25
p_value_m = pgamma(sn/25, shape = n, rate = 1, lower.tail = FALSE)
round(p_value_m, 4)
```

```
## [1] 0.04
```

La p valeur est la même qu'à la question 6 car m = 25 est la même chose que $\theta = 1/25$