Université de Strasbourg
UFR de Mathématiques et d'Informatique
Han-Ping LI

Année 2020/2021
Statistique L3
Etude de cas

Fiche 3 Intervalles de confiance

Exercice 1 : la taille du lobe frontal des crabes

On s'intéresse à la taille du lobe frontal des crabes (Leptograpsus variegatus). On prend 200 valeurs de la variable la taille du lobe frontal "FL" du fichier "crabs" de la librairie "MASS" comme population. On note μ et σ^2 la moyenne théorique et la variance théorique de la variable FL. Après une étude statistique préalable (test de la normalité), on suppose raisonnablement que la population admet une loi normale $\mathcal{N}(\mu, \sigma^2)$ avec les deux paramètres inconnus.

1. `install.packages("MASS"); library(MASS)`

2. Sauvegarder ces 200 valeurs de "FL" du fichier "crabs" (sans "e") dans une variable nommée **Population**.

3. Générer une réalisation $(x_1...x_n)$ d'un échantillon $(X_1...X_n)$ avec n=28 , puis construire l'intervalle de confiance de μ de niveau de confiance $1-\alpha = 95\%$. Contient-il la valeur de μ=`mean(Population)` ?

4. Générer 100 réalisations d'un échantillon $(X_1...X_n)$ avec n=28 , puis construire 100 réalisations de l'intervalle de confiance de μ de niveau de confiance $1-\alpha = 95\%$. Quel est le pourcentage des réalisations qui ne contiennent pas le paramètre $\mu = 15.583$?

5. Générer une réalisation $(x_1...x_n)$ d'un échantillon $(X_1...X_n)$ avec n=15 , puis construire l'intervalle de confiance de σ^2 de niveau de confiance $1-\alpha = 90\%$. Contient-il la valeur de $\sigma^2 =$ `var(Population)` ?

6. Générer 100 réalisations d'un échantillon $(X_1...X_n)$ avec n=15 , puis construire l'intervalle de confiance de σ^2 de niveau de confiance $1-\alpha = 90\%$. Quel est le pourcentage des réalisations qui ne contiennent pas le paramètre $\sigma^2 = 12.2173$?

7. Comparer les résultat avec ceux des autres étudiants.

Exercice 1' : l'indice Standard and Poors 500 pour les économistes :

On s'intéresse aux rendements journaliers de bourse de l'indice Standard and Poors 500 de 1990 à 1999. On prend 2780 valeurs du fichier "SP500" de la librairie "MASS" comme population. On note μ et σ^2 la moyenne théorique et la variance théorique du rendement. Après une étude statistique préalable (test de la normalité), on suppose raisonnablement que la population admet une loi normale $\mathcal{N}(\mu, \sigma^2)$ avec les deux paramètres inconnus (*mais cette approximation n'est pas de bonne qualité, ce qui implique que le coefficient de confiance n'est pas très précis*).

1. `install.packages("MASS"); library(MASS)`

2. Sauvegarder ces $N = 2780$ valeurs du rendement du fichier "SP500" dans une variable nommée **Population**.

3. Générer une réalisation $(x_1...x_n)$ d'un échantillon $(X_1...X_n)$ avec n=28 , puis construire l'intervalle de confiance de μ de niveau de confiance $1 - \alpha = 95\%$. Contient-il la valeur de $\mu = mean(Population)$?

4. Générer 100 réalisations d'un échantillon $(X_1...X_n)$ avec n=28 , puis construire 100 réalisations de l'intervalle de confiance de μ de niveau de confiance $1 - \alpha = 95\%$. Quel est le pourcentage des réalisations qui ne contiennent pas le paramètre $\mu = 0.04575267$?

5. Générer une réalisation $(x_1...x_n)$ d'un échantillon $(X_1...X_n)$ avec n=15 , puis construire l'intervalle de confiance de σ^2 de niveau de confiance $1 - \alpha = 90\%$. Contient-il la valeur de $\sigma^2 = var(Population)$?

6. Générer 100 réalisations d'un échantillon $(X_1...X_n)$ avec n=15 , puis construire l'intervalle de confiance de σ^2 de niveau de confiance $1 - \alpha = 90\%$. Quel est le pourcentage des réalisations qui ne contiennent pas le paramètre $\sigma^2 = 0.8982233$?

7. Comparer les résultat avec ceux des autres étudiants.

Exercice 2 : loi gaussienne

On simule $M = 100$ réalisations d'un échantillon $(X_1, \ldots, X_{20})$ (n=20) de la loi $\mathcal{N}(-10, 1.5^2)$. On n'est pas censé connaitre les deux paramètres $\mu = -10$ et $\sigma^2 = 1.5^2$. On sauvegarde 100 réalisations de la statistique $\overline{X}$ ainsi que celles de S_c^2.

1. Déterminer 100 réalisations de l'intervalle de confiance de μ au niveau de confiance de **90**%, basée sur ces 100 réalisations. Quel est le pourcentage des réalisations qui ne contiennent pas le paramètre $\mu = -10$? Commenter vos résultats.

2. Comment utiliser R pour représenter ces 100 réalisations de l'intervalle de confiance su un même graphe ?

3. Déterminer 100 réalisations de l'intervalle de confiance de σ^2 au niveau de confiance de 95%, basée sur ces 100 réalisations. Quel est le pourcentage des réalisations qui ne contiennent pas le paramètre $\sigma^2 = 1.5^2$?

4. Ayant un niveau de confiance 95% ($\alpha = 0.05$), les deux quantiles suivants

$$\mathbb{P}\Big(\frac{(n-1)S_c^2}{\sigma^2} < C_1\Big) = \mathbb{P}\Big(\frac{(n-1)S_c^2}{\sigma^2} > C_2\Big) = \frac{\alpha}{2}.$$

   fournissent un choix simple à calculer, et proche du choix optimal. Sauvegarder les longueurs des 100 réalisations des IC en c).

5. Pour le même niveau de confiance 95%, on définit dans cette question

$$\begin{aligned}\mathbb{P}\Big(\frac{(n-1)S_c^2}{\sigma^2} < C_1\Big) &= 0.04461\\ \mathbb{P}\Big(\frac{(n-1)S_c^2}{\sigma^2} > C_2\Big) &= 0.00539.\end{aligned}$$

   On sait qu'ils sont biaisés. Montrer qu'ils ont systématiques les longueurs plus courtes que celles en 4).

6. En remplaçant $\mu = -10$ et $\sigma^2 = 1.5^2$ par d'autres valeurs, refaire les questions de 1 à 4.

Exercice 3 : illustration du Théorème central limite

On simule une réalisation de N=M*n = 50000 (M=1000, n=50) variables aléatoires de la loi de Beta avec paramètre $a = 1/2$ et $b = 1/2$ par

```
M=1000; n=50; N=M*n; x=rbeta(N, 1/2,1/2).
```

1. Tracer son histogramme basé sur ces N réalisations, en utilisant :

```
mx=min(x) ; Mx=max(x)
brk=seq(mx,Mx,length=101)
hist(x,probability = TRUE,col ="lightblue",breaks=brk)
```

2. Superposer le graphique de la densité théorique B(0.5,0.5) :
   `curve(dbeta(x,0.5,0.5),col="red",lwd =2,add=TRUE)`

   On admet que son espérance mathématique sa variance théorique sont

$$\mathbf{E}(X) = \frac{b}{(a+b)} = 1/2 \text{ et } \mathbf{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)} = 1/8.$$

   La réalisation $x = (x_1, \ldots, x_{50000})$ peut être considérée comme M=1000 réalisations de l'échantillon $(X_1, \ldots, X_n)$ avec $n = 50$:

```
Echant=matrix(x, nrow=1000, byrow=TRUE)
```

3. On utilise $M = 1000$ réalisations d'un échantillon de taille $n = 50$ pour obtenir et Sauvegarder ces M réalisations de la variable aléatoire $Z = \dfrac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$. Tracer l'histogramme Z.

```
z = numeric(M);
 for (k in 1:M){
     z[k] = sqrt(n)*(mean(Echant[k,])-0.5 )/(1/sqrt(8)) }
 hist( z, proba=T, breaks=0.1*(-55:55), xlim=c(-5,5),
     main=" loi de Beta(0.5,0.5)" )
```

4. Superposer la densité de la loi gaussienne

```
curve( dnorm( x, mean=0, sd=1),col="red",lwd = 2,add=TRUE)
```

Commenter vos résultats.

Exercice 4 : Intervalle de confiance pour π (π = la proportion de largeur d'anneaux $\leqslant 0,89$)

On s'intéresse à la largeur des anneaux de croissance des arbres. On prend le fichier "treering" avec 7980 comme étant le cardinal de la population. On note π la proportion de largeurs qui sont inférieures ou égales à 0,89 (unité) et on souhaite construire un intervalle de confiance de π de niveau de confiance de 95%.

1. Convertir ensuite les valeurs des "treering" en 1 ou 0 selon que les valeurs des largeurs $\leqslant 0,89$ (unité) ou non. Sauvegarder ces 7980 valeurs converties dans un vecteur nommé **Population**.

2. Générer une réalisation $(x_1...x_n)$ d'un échantillon $(X_1, ..., X_n)$ de taille $n = 65$. On note $s = \sum_{i=1}^{n} x_i$ la réalisation de la variable aléatoire $S = \sum_{i=1}^{n} X_i$.

3. Appliquer Théorème Central Limite pour construire un intervalle de confiance de niveau 95% du paramètre π. Contient-il le paramètre π ?

4. Comparer avec le résultat par la méthode de Wald (asymptotique) :

```
install.packages("binom"); library(binom)

binom.confint(x=s, n=65, conf.level = 0.95, methods = "asymptotic")
```

5. Déterminer l'intervalle par la méthode exacte :

```
binom.confint(x=s, n=65, conf.level = 0.95, methods = "exact")
```

6. Comparer avec les deux bornes obtenues en utilisant la loi de Beta :

```
(BInf=qbeta(0.025, s, 65-s+1));
(BSup=qbeta(1-0.025, s+1, 65-s))
```

Que constatez-vous ?

7. Comparer avec le résultat par la méthode de Bayes :
   binom.confint(x=s, n=65, conf.level = 0.95, methods = "bayes")

8. Mêmes questions avec $M = 100$ réalisations.

Exercice 5 : loi de Bernoulli avec la correction de continuité de Yates

On simule $M = 3000$ réalisations d'un échantillon $(X_1, \ldots, X_{15})$ (n=15) de la loi de Bernoulli $\mathcal{B}(1, p)$ avec $p = 0.45$. On sauvegarde les M réalisations des statistiques $S = \sum_{i=1}^{n} X_i$ et $\overline{X}$.

1. Déterminer, par la méthode exacte, M réalisations de l'intervalle de confiance de p au niveau de confiance de 95%, basée sur ces M réalisations. Quel est le pourcentage des réalisations qui ne contiennent pas le paramètre $p = 0.45$? Que valent les moyennes de M réalisations de la bornes inférieures et les bornes supérieures ?

   ```
   install.packages("binom"); library(binom)
   ```

   On sauvegarde la sortie de la commande :
   ```
   IC = binom.confint(x=s, n=15, conf.level = 0.95, methods ="exact")
   ```

   On peut récupérer une réalisation de la borne inférieure et de la borne supérieure respectivement par
   ```
   IC[5]; IC[6]
   ```

   ou simplement
   ```
   (BInf[i] = qbeta(0.025, s, 15-s+1));
   (BSup[i] = qbeta(1-0.025, s+1, 15-s))
   ```

2. Normalement, n=15 est trop petit pour pouvoir appliquer le Théorème Central Limite. Appliquer le quand même pour voir ce qui se passe. Quel est le pourcentage des réalisations qui ne contiennent pas le paramètre $p = 0.45$? Que valent les moyennes de M réalisations de la bornes inférieures et les bornes supérieures ?

   ou appliquer la méthode de Wald :

   ```
   binom.confint(x=s, n=15, conf.level = 0.95, methods = "asymptotic")
   ```

   Remarque 1 : Si on note la fonction de répartition de la loi exacte $\mathcal{L}_B = B(15, 0.45)$ par $F_B(x)$ et celle de la loi normale $\mathcal{L}_N = \mathcal{N}\Big(np, np(1-p)\Big)$ par $F_N(x)$, alors la distance de Kolmogorov entre les deux lois (qui mesure la qualité de l'approximation) est définie par :

$$d(\mathcal{L}_B, \mathcal{L}_N) = \sup_{x \in \mathbb{R}} |F_B(x) - F_N(x)|.$$

3. Calculer la distance de Kolmogorov de l'approximation. Montrer que cette distance telle qu'elle est définie est plus grand que 0.10.

   Remarque 2 : pour une variable aléatoire discrète X, on constate que la probabilité pour un entier k

   $$\mathbf{P}(X = k) = \mathbf{P}(k - \frac{1}{2} < X \leq k + \frac{1}{2}).$$

   Par conséquent, sa fonction de répartition vérifie :

   $$F(k) = \mathbf{P}(X \leq k) = \mathbf{P}(X \leq k + \frac{1}{2}) \ \ \forall\, k \in \mathbb{N}$$

   devrait être approchée, non pas par la fonction de répartition normale $F_N(k)$, mais plutôt par $F_N(k + \frac{1}{2})$. C'est ce que l'on appelle **la correction de continuité de Yates** :

   $$F_B(k) \approx F_N(k + \frac{1}{2}) \ \ \forall\, k \in \mathbb{N}.$$

4. Calculer la distance de Kolmogorov de l'approximation avec la correction de continuité de Yates. Montrer qu'avec cette correction de continuité de Yates, la distance est plus petite que 0.004.

Exercice 6 : loi de Bernoulli

On simule $M = 50$ réalisations d'un échantillon $(X_1, \ldots, X_{1000})$ $(n = 1000)$ de la loi de Bernoulli $\mathcal{B}(1, p)$ avec $p = 0.005$. On sauvegarde les M réalisations des statistiques $S = \sum_{i=1}^{n} X_i$. et $\overline{X}$.

1. Déterminer, par la méthode exacte, M réalisations de l'intervalle de confiance de p au niveau de confiance de 95%. Quel est le pourcentage des réalisations qui ne contiennent pas le paramètre $p = 0.005$? Que valent la moyenne de M réalisations de la bornes inférieures et celle des bornes supérieures ?

   `(BInf[i] = qbeta(0.025, s, 1000-s+1))` ;
   `(BSup[i] = qbeta(1-0.025, s+1, 1000-s))`

2. Comment utiliser R pour représenter ces M réalisations de l'intervalle de confiance su un même graphe ?
3. Déterminer, par la méthode asymptotique, M réalisations de l'intervalle de confiance de p au niveau de confiance de 95%. Quel est le pourcentage des réalisations qui ne contiennent pas le paramètre p ? Que valent les moyennes de M réalisations de la bornes inférieures et les bornes supérieures ? Que constatez-vous ?

4. Comment utiliser R pour représenter ces 50 réalisations de l'intervalle de confiance su un même graphe ?

5. Calculer la distance de Kolmogorov de l'approximation. Montrer qu'avec cette correction de continuité de Yates, la distance est plus grande que 0.0287. Commenter vos résultats.

Exercice 7 : modélisation du nombre de particules émises

On souhaite modéliser le nombre de particules alpha émises par Américium-241 (une source radioactive) pour une durée de temps donnée. Un jeu de données est fourni dans l'édition de 1988 de «Statistiques mathématiques et analyse des données» de John Rice.

https ://www.r-bloggers.com/checking-the-goodness-of-fit-of-the-poisson-distribution-in-r-for-alpha-decay-by-americium-241/

Les nombres d'émissions alpha d'américium-241 :

0	1	2	3	4	5	6	7	8	9
1	4	13	28	56	105	126	146	164	161
10	11	12	13	14	15	16	17	18	19
123	101	74	53	23	15	9	3	1	1

Il y avait 1 207 intervalles de temps dans cette étude, chacun d'une durée de 10 secondes. Le nombre de particules alpha émises dans chaque intervalle de temps a été enregistré ; ce nombre variait de 0 à 19. La deuxième colonne du tableau de données suivant montre les comptes observés pour chaque nombre d'émissions alpha. Par exemple, il y avait 1 intervalle avec 0 émission, 4 intervalles avec 1 émission et 13 intervalles avec 2 émissions. c.f.

1. Que vaut le nombre total des particules alpha émises par Américium-241 ? Que vaut $\hat{\lambda}$= la moyenne observée ?

2. Si on modélise le nombre de particules alpha émises par Américium-241 par une loi de Poisson avec $\lambda = \hat{\lambda}$, déterminer les effectifs observés et les effectifs théoriques.

3. Tracer un diagramme à bâtons à barres accolées des effectifs observés et des effectifs théoriques.

4. Tracer un diagramme à bâtons à barres accolées des proportions observés et des probabilités.

5. Déterminer, par la méthode exacte (respectivement asymptotique, de Bayes), l'intervalle de confiance de λ au niveau de confiance de 90%, basée sur le jeu de données.

Exercice 8 : loi de Poisson

On simule $M = 1000$ réalisations d'un échantillon $(X_1, \ldots, X_{15})$ (n=15) de la loi de Poisson $\mathcal{P}(\lambda)$ avec $\lambda = 1,5$. On sauvegarde M réalisations de la statistique $\overline{X}$.

1. Déterminer, par la méthode exacte (respectivement wald (asymptotique)), M réalisations de l'intervalle de confiance de λ au niveau de confiance de 95%. Quels sont les pourcentages des réalisations qui ne contiennent pas le paramètre $\lambda = 1,5$?

   ```
   install.packages("DescTools"); library("DescTools")
   n = 15 ; lambda = 1.5@
     BInf.e=c();  BSup.e=c();  BInf.a=c();  BSup.a=c()
    for (i in 1:M) {
        echant=rpois(n, lambda)
        s = sum(echant)
        IC.e=PoissonCI(s, n=15, conf.level=0.95, sides ="two.sided",
              method="exact")
        IC.a=PoissonCI(s, n=15, conf.level=0.95, sides ="two.sided",
              method="wald")
        BInf.e[i] = IC.e[2]; BSup.e[i] = IC.e[3]
        BInf.a[i] = IC.a[2]; BSup.a[i] = IC.a[3]  }
   ```

2. Comparer la moyenne de bornes inférieures obtenues par deux méthodes. Même question pour les bornes supérieures. Commenter vos résultats.

3. La méthode de Wald (asymptotique) est-elle efficace ? Pourquoi ?

Exercice 9 : loi de Poisson

On simule $M = 500$ réalisations d'un échantillon $(X_1, \ldots, X_{80})$ $(n = 80)$ de la loi de Poisson $\mathcal{P}(\lambda)$ avec $\lambda = 0.05$. On sauvegarde M réalisations de la statistique $\overline{X}$.

1. Déterminer, par la méthode exacte (respectivement la méthode de wald (asymptotique)), M réalisations de l'intervalle de confiance de λ au niveau de confiance de 95%, basée sur ces M réalisations. Quel est le pourcentage des réalisations qui ne contiennent pas le paramètre $\lambda = 0.05$? Commenter vos résultats.

   b) Comparer la moyenne de bornes inférieures obtenues par deux méthodes. Même question pour les bornes supérieures. Commenter vos résultats.

2. La méthode de Wald (asymptotique) est-elle efficace ? Pourquoi ?