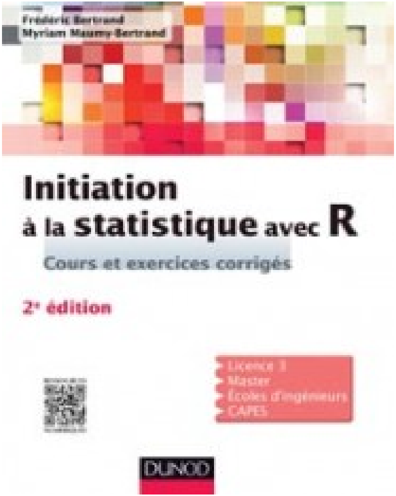
Statistique : étude de cas. Intervalles de confiance

Myriam Maumy-Bertrand

IRMA, UMR 7501, Université de Strasbourg

Vendredi 06 octobre 2017

Ce chapitre s'appuie essentiellement sur le livre suivant :



Sommaire

Il est souvent plus réaliste et plus intéressant de fournir un renseignement du type

$$\theta_1 < \theta < \theta_2$$

plutôt que d'écrire sèchement

$$\widehat{\theta}_n = c.$$

Définition

Fournir un tel intervalle $]\theta_1;\theta_2[$ s'appelle donner une estimation par intervalle de confiance de θ ou une estimation ensembliste de θ.

Sommaire

La méthode des intervalles de confiance est la suivante :
soit $\widehat{\theta}_n$ un estimateur de θ dont nous connaissons la loi de probabilité pour chaque valeur de θ.

Définition

Étant donné une valeur θ_0 du paramètre θ, nous déterminons un **intervalle de probabilité bilatéral de niveau** $(1-\alpha)$ *pour l'estimateur $\widehat{\theta}_n$, c'est-à-dire deux bornes θ_1^n et θ_2^n telles que*

$$\mathbb{P}\left(\theta_1^n < \widehat{\theta}_n < \theta_2^n | \theta = \theta_0\right) \geqslant 1 - \alpha.$$

Remarques

1. Ces deux bornes dépendent évidemment de la valeur θ_0.
2. Nous pourrions également construire des **intervalles unilatéraux** pour lesquels $\theta_1^n = -\infty$ ou $\theta_2^n = +\infty$.
3. Nous choisissons dans la plupart des cas un intervalle de probabilité à risque symétrique $\alpha/2$ et $\alpha/2$.

Nous adoptons la règle de décision suivante.
Soit $\widehat{\theta}_n(x_1, \ldots, x_n) = \widehat{\theta}_n(obs)$ la valeur observée de $\widehat{\theta}_n$:

- si $\widehat{\theta}_n(obs) \in]\theta_1^n; \theta_2^n[$, nous conservons θ_0 comme valeur possible du paramètre θ ;
- si $\widehat{\theta}_n(obs) \notin]\theta_1^n; \theta_2^n[$, nous éliminons θ_0.

Nous répétons cette opération pour toutes les valeurs de θ.

Sommaire

Définition

Nous appelons **intervalle de confiance de niveau de confiance** $(1-\alpha)$ **(coefficient de confiance) du paramètre** θ *tout intervalle* $]\theta_1;\theta_2[$ *tel que :*$\mathbb{P}\left(\theta \in]\theta_1;\theta_2[\right) = 1-\alpha$ *pour* $\alpha \in [0;1]$ *fixé.*

Propriétés

1. $]\theta_1; \theta_2[$ est un intervalle aléatoire car il dépend de l'estimateur $\widehat{\theta}_n$.
2. $]\theta_1; \theta_2[$ s'obtient par :

$$\left\{ \begin{array}{lcl} \theta_1 & = & (\theta_2^n)^{-1}\,(\widehat{\theta}_n(obs)) \\ \theta_2 & = & (\theta_1^n)^{-1}\,(\widehat{\theta}_n(obs)) \end{array} \right. .$$

3. Si nous augmentons le niveau $1-\alpha$, nous augmentons la longueur de l'intervalle de probabilité.

Remarque

Si la taille de l'échantillon notée n augmente, comme l'estimateur $\widehat{\theta}_n$ est supposé convergent, la variance de l'estimateur notée $\mathrm{Var}\left(\widehat{\theta}_n\right)$ diminue et par conséquent l'intervalle $]\theta_1; \theta_2[$ diminue également.

Sommaire

$\widehat{\mu}_n$ est le meilleur estimateur de la moyenne μ_{pop} et $\widehat{\mu}_n$ suit une loi normale $\mathcal{N}\left(\mu_{pop}; \dfrac{\sigma^2_{pop}}{n}\right)$.

Définition

L'intervalle de probabilité de $\widehat{\mu}_n$ à $1-\alpha$ est :

$$\mu_{pop} - u_{1-(\alpha/2)}\frac{\sigma_{pop}}{\sqrt{n}} < \widehat{\mu}_n < \mu_{pop} + u_{1-(\alpha/2)}\frac{\sigma_{pop}}{\sqrt{n}},$$

où u_p est le quantile d'ordre p pour la loi gaussienne centrée et réduite.

Définition

L'intervalle de confiance de $\widehat{\mu}_n$ à $1-\alpha$ est :

$$\widehat{\mu}_n(x_1,\ldots,x_n) - u_{1-(\alpha/2)}\frac{\sigma_{pop}}{\sqrt{n}} < \mu < \widehat{\mu}_n(x_1,\ldots,x_n) + u_{1-(\alpha/2)}\frac{\sigma_{pop}}{\sqrt{n}},$$

où $u_{1-(\alpha/2)} \simeq 1,96$ si $1-\alpha = 0,95$.

Remarque

Pour obtenir le quantile d'ordre $0,975$ de la loi normale centrée et réduite, sous `R`, vous tapez la ligne de commande suivante :

```
>qnorm(0.975)
[1] 1.959964
```

Nous utilisons le fait que la variable aléatoire $T_{n-1} = \sqrt{n-1}\,\dfrac{\widehat{\mu}_n - \mu}{S_n}$ suit une loi de Student à $(n-1)$ degrés de liberté.

Définition

L'intervalle de probabilité pour T_{n-1} à $1-\alpha$ est :

$$-t_{n-1;1-(\alpha/2)} < \sqrt{n-1}\,\frac{\widehat{\mu}_n - \mu}{S_n} < t_{n-1;1-(\alpha/2)},$$

où $t_{n-1;1-(\alpha/2)}$ est le quantile d'ordre $1-(\alpha/2)$ pour la loi de Student à $(n-1)$ degrés de liberté.

Définition

L'intervalle de confiance pour μ à $1-\alpha$ est :

$$\widehat{\mu}_n(obs) - t_{n-1;1-(\alpha/2)}\frac{S_n(obs)}{\sqrt{n-1}} < \mu < \widehat{\mu}_n(obs) + t_{n-1;1-(\alpha/2)}\frac{S_n(obs)}{\sqrt{n-1}}$$

ou bien

$$\widehat{\mu}_n(obs) - t_{n-1;1-(\alpha/2)}\frac{S_{n,c}(obs)}{\sqrt{n}} < \mu < \widehat{\mu}_n(obs) + t_{n-1;1-(\alpha/2)}\frac{S_{n,c}(obs)}{\sqrt{n}},$$

où $t_{n-1;1-(\alpha/2)}$ est le quantile d'ordre $1-(\alpha/2)$ pour la loi de Student à $(n-1)$ degrés de liberté.

Remarques

1. Pour obtenir le quantile d'une loi de Student, sous R, vous tapez la ligne de commande suivante :
   ```
   >qt(0.975,n-1)
   ```
   où la quantité $n-1$ est remplacée par la valeur adéquate.
2. Le théorème de la limite centrée a pour conséquence que les intervalles précédents sont valables pour estimer μ d'une loi quelconque lorsque la taille n de l'échantillon est assez grande.

Exemple : L'airbag, d'après l'examen de février 2014

L'airbag (ou coussin gonflable) est un système de sécurité de plus en plus souvent installé dans les automobiles. Son gonflement est assuré par un dispositif pyrotechnique dont les caractéristiques sont la moyenne et l'écart-type du délai entre la mise à feu et l'explosion. Lors de l'étude d'un certain dispositif d'allumage, les résultats des mesures qui proviennent d'une loi normale, effectués sur 30 exemplaires, ont été (en millisecondes) les suivants :

28,0	28,0	31,0	31,0	32	33,0	32,5	29,0	30,5	31,0
28,5	27,5	32,0	29,5	28	26,0	30,0	31,0	32,5	33,0
27,5	29,0	30,0	28,5	27	25,0	31,5	33,0	34,5	29,0

L'airbag, d'après l'examen de février 2014 : suite

Calculer l'intervalle de confiance à 95% de la moyenne du délai si nous connaissons l'écart-type de la population de référence et qu'il est égal à 2.
Nous commençons par donner une estimation ponctuelle de la moyenne du délai μ :

```
>gonflable<-c(28,28,31,31,32,33,32.5,29,30.5,31,
28.5,27.5,32,29.5,28,26,30,31,32.5,33,27.5,29,
30,28.5,27,25,31.5,33,34.5,29)
>mean(gonflable)
[1] 29.96667
```

Maintenant appliquons la formule du cours, à savoir celle qui donne un intervalle de confiance pour une moyenne μ lorsque la variance σ^2 est connue (ici elle vaut 4).

L'airbag, d'après l'examen de février 2014 : suite

Pour cela, il faut vérifier au préalable que les données suivent une loi normale. Réalisons donc un test de normalité de Shapiro-Wilk.

```
>shapiro.test(gonflable)
Shapiro-Wilk normality test
data: gonflable
W = 0.9796, p-value = 0.8149
```

La p-valeur (0,8149) du test de Shapiro-Wilk étant strictement supérieure à $\alpha = 5\%$, le test n'est pas significatif. Vous conservez donc l'hypothèse nulle H_0 du test de Shapiro-Wilk. Le risque d'erreur associé à cette décision est un risque de deuxième espèce β. Vous ne pouvez pas l'évaluer dans le cas d'un test de Shapiro-Wilk.

L'airbag, d'après l'examen de février 2014 : suite

Maintenant nous pouvons calculer avec `R` l'intervalle de confiance cherché (nous rappelons que nous connaissons la variance σ^2 de la population).

```
>mean(gonflable)-qnorm(0.975)*2/sqrt(30)
[1] 29.25099
>mean(gonflable)+qnorm(0.975)*2/sqrt(30)
[1] 30.68234
```

L'intervalle de confiance à 95% de la moyenne du délai, en millisecondes, est égal à :

$$]29,25099; 30,68234[.$$

Exemple : Les plantes marines

Un biologiste étudie un type d'algue qui attaque les plantes marines. La toxine contenue dans cette algue est obtenue sous forme d'une solution organique. Il mesure la quantité de toxine par gramme de solution. Il a obtenu les neuf mesures suivantes, exprimées en milligrammes :

$$1,2; 0,8; 0,6; 1,1; 1,2; 0,9; 1,5; 0,9; 1,0.$$

Nous supposons que ces mesures sont les réalisations de variables aléatoires indépendantes et identiquement distribuées suivant une loi normale d'espérance μ et d'écart-type σ.

Les plantes marines : suite

Donnons une estimation ponctuelle de la moyenne de la quantité de toxine par gramme de solution :

$$\begin{aligned}\widehat{\mu}_9(obs) &= \frac{1}{9}\Big(1,2+0,8+0,6+1,1+1,2+0,9+1,5 \\ &\qquad +0,9+1,0\Big) \\ &= 1,022222 \; mg.\end{aligned}$$

Les plantes marines : suite

Donnons une estimation ponctuelle de la variance de la quantité de toxine :

$$S_{9,c}^2(obs) = (0,2635231\ mg)^2.$$

Les plantes marines : suite

Déterminons un intervalle de confiance à 95% pour la moyenne de la quantité de toxine par gramme de solution. La moyenne μ et la variance étant inconnues, l'intervalle de confiance à 95% pour la moyenne μ s'obtient avec la formule suivante :

$$\begin{aligned} \widehat{\mu}_9(obs) - t_{8;0,975}\frac{S_{9,c}(obs)}{\sqrt{9}} < & \quad \mu \quad < \widehat{\mu}_9(obs) + t_{8;0,975}\frac{S_{9,c}(obs)}{\sqrt{9}} \\ 0,820\ mg < & \quad \mu \quad < 1,225\ mg \end{aligned}$$

où $t_{8;0,975} = 2,3060$ est le quantile d'ordre $0,975$ pour la loi de Student à huit degrés de liberté $t(8)$.

Les plantes marines : fin

Les commandes sous R qui donnent les résultats ci-dessus sont les suivantes :

```
> toxine<-c(1.2,0.8,0.6,1.1,1.2,0.9,1.5,0.9,1)
> mean(toxine)
[1] 1.022222
> sd(toxine)
[1] 0.263523
> mean(toxine)-qt(0.975,8)*sd(toxine)/sqrt(9)
[1] 0.8196604
> mean(toxine)+qt(0.975,8)*sd(toxine)/sqrt(9)
[1] 1.224784
```

Sommaire

Nous utilisons le fait que

1. $\widehat{\sigma^2}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \mu)^2$ est le meilleur estimateur de la variance σ^2 lorsque la moyenne μ est connue,
2. $\dfrac{n\widehat{\sigma^2}_n}{\sigma^2}$ suit une loi du Khi-deux à n degrés de liberté comme la somme de n carrés de loi $\mathcal{N}(0;1)$ indépendantes.

Définition

k_1 et k_2 sont les bornes de l'intervalle de probabilité pour $\dfrac{n\widehat{\sigma^2}_n}{\sigma^2}$ si :

$$\mathbb{P}\left(k_1 < \frac{n\widehat{\sigma^2}_n}{\sigma^2} < k_2\right) = 1 - \alpha.$$

Remarque

Par exemple, nous pouvons prendre la borne k_1 égale au quantile d'ordre $\alpha/2$ pour la loi du Khi-deux à n degrés de liberté et la borne k_2 égale au quantile d'ordre $1 - \alpha/2$ pour la loi du Khi-deux à n degrés de liberté.

Définition

L'intervalle de confiance pour σ^2 à $1-\alpha$ est égal à :

$$\frac{n\widehat{\sigma^2}_n(x_1,\ldots,x_n)}{k_2} < \sigma^2 < \frac{n\widehat{\sigma^2}_n(x_1,\ldots,x_n)}{k_1}.$$

Remarque

Par exemple, nous pouvons prendre la borne k_1 égale au quantile d'ordre $\alpha/2$ pour la loi du Khi-deux à n degrés de liberté et la borne k_2 égale au quantile d'ordre $1-\alpha/2$ pour la loi du Khi-deux à n degrés de liberté.

Nous utilisons le fait que

1. $S_n^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)^2$,
2. $\dfrac{nS_n^2}{\sigma^2}$ suit une loi du Khi-deux à $(n-1)$ degrés de liberté comme la somme de $(n-1)$ carrés de loi $\mathcal{N}(0;1)$ indépendantes.

Définition

l_1 et l_2 sont les bornes de l'intervalle de probabilité pour $\frac{nS_n^2}{\sigma^2}$ si :

$$\mathbb{P}\left(l_1 < \frac{nS_n^2}{\sigma^2} < l_2\right) = 1 - \alpha.$$

Définition

L'intervalle de confiance pour σ^2 à $1 - \alpha$ est égal à :

$$\frac{nS_n^2(x_1, \ldots, x_n)}{l_2} < \sigma^2 < \frac{nS_n^2(x_1, \ldots, x_n)}{l_1},$$

où l_1 et l_2 sont les quantiles respectivement à $\alpha/2$ et à $1 - (\alpha/2)$ du loi du Khi-deux à $(n - 1)$ degrés de liberté.

Définition

l_1 et l_2 sont les bornes de l'intervalle de probabilité pour $(n-1)S_{n,c}^2/\sigma^2$ si :

$$\mathbb{P}\left(l_1 < \frac{(n-1)S_{n,c}^2}{\sigma^2} < l_2\right) = 1 - \alpha.$$

Définition

L'intervalle de confiance à $1-\alpha$ pour la variance σ^2 est égal à :

$$\frac{(n-1)S^2_{n,c}(x_1,\ldots,x_n)}{l_2} < \sigma^2 < \frac{(n-1)S^2_{n,c}(x_1,\ldots,x_n)}{l_1},$$

où l_1 et l_2 sont les quantiles définis dans la diapositive précédente.

Remarque

Pour obtenir le quantile l_1 d'une loi du Khi-deux, sous R, vous tapez la ligne de commande suivante :

```
>qchisq(0.025,n-1)
```

où la quantité $n-1$ est remplacée par la valeur adéquate. De même pour l_2 en remplaçant $0,025$ par $0,975$.

Exemple : Les plantes marines

Un biologiste étudie un type d'algue qui attaque les plantes marines. La toxine contenue dans cette algue est obtenue sous forme d'une solution organique. Il mesure la quantité de toxine par gramme de solution. Il a obtenu les neuf mesures suivantes, exprimées en milligrammes :

$$1,2; 0,8; 0,6; 1,1; 1,2; 0,9; 1,5; 0,9; 1,0.$$

Nous supposons que ces mesures sont les réalisations de variables aléatoires indépendantes et identiquement distribuées suivant une loi normale d'espérance μ et d'écart-type σ_{pop}.

Les plantes marines : suite

Donnons une estimation ponctuelle de la moyenne de la quantité de toxine par gramme de solution :

$$\widehat{\mu}_9(obs) = 1,022222 \ mg.$$

Donnons une estimation ponctuelle de la variance de la quantité de toxine à l'aide de l'estimateur corrigé de la variance $S^2_{9,c}$:

$$S^2_{9,c}(obs) = 0,06944444 \ mg^2$$

Les plantes marines : suite

Déterminons un intervalle de confiance à 95% pour la variance σ^2_{pop} de la quantité de toxine par gramme de solution. La moyenne μ et la variance σ^2_{pop} étant inconnues, l'intervalle de confiance à 95% pour la variance σ^2_{pop} s'obtient avec la formule suivante :

$$\begin{array}{rcl} \dfrac{8S^2_{9,c}(obs)}{I_2} < & \sigma^2_{pop} & < \dfrac{8S^2_{9,c}(obs)}{I_1} \cdot \\ 0,032\ mg^2 < & \sigma^2_{pop} & < 0,255\ mg^2 \end{array}$$

où $I_1 = 2,180$ est le quantile d'ordre $0,025$ pour une loi du Khi-deux $\chi^2(8)$ et $I_2 = 17,535$ est le quantile d'ordre $0,975$ pour une loi du Khi-deux $\chi^2(8)$.

Les plantes marines : fin

Les commandes sous R qui donnent les résultats ci-dessus sont les suivantes :

```
> mean(toxine)
[1] 1.022222
> var(toxine)
[1] 0.06944444
> 8*var(toxine)/qchisq(0.975,8)
[1] 0.03168349
> 8*var(toxine)/qchisq(0.025,8)
[1] 0.2548735
```

Sommaire

Nous souhaitons construire un intervalle de confiance pour une proportion π_A d'individus de la population qui possèdent un certain caractère A. Pour estimer π_A, nous allons nous servir de l'estimateur $\widehat{\pi}_{n,A}$, qui a été défini auparavant, à partir du moment où nous faisons l'hypothèse que le tirage est aléatoire avec remise, ce qui correspond à une population infinie.
D'autre part, nous pouvons montrer que $n\widehat{\pi}_{n,A}$ suit une loi binomiale de paramètres n et π_A. À partir de ce résultat, nous pouvons construire un intervalle de confiance pour π_A.

Les trois méthodes pour construire un intervalle de confiance pour une proportion π_A que nous rencontrerons le plus souvent sont :

1. la méthode exacte ou encore appelée la méthode de Clopper-Pearson qui maintenant est réalisable avec par exemple le logiciel `R` ;
2. la méthode du score ou encore appelée la méthode de Wilson ou encore la méthode de l'ellipse ;
3. la méthode asymptotique ou encore appelée la méthode de Wald.

Des études statistiques ont montré que, parmi ces trois méthodes, la méthode à privilégier est la méthode du score. Néanmoins, la méthode de Wald reste très utilisée et présentée dans de nombreux manuels alors qu'elle ne permet généralement pas d'obtenir des intervalles de confiance de qualité convenable.

L'intervalle de confiance pour la proportion π_A au niveau de confiance $(1-\alpha)$ est égal à

$$\frac{\widehat{\pi}_{n,A}(obs) + \frac{1}{2n}u^2_{1-\alpha/2} - u_{1-\alpha/2} \times \sqrt{\frac{\widehat{\pi}_{n,A}(obs)(1-\widehat{\pi}_{n,A}(obs))}{n} + \frac{u^2_{1-\alpha/2}}{4n^2}}}{1+\frac{1}{n}u^2_{1-\alpha/2}} < \pi_A$$

$$< \frac{\widehat{\pi}_{n,A}(obs) + \frac{1}{2n}u^2_{1-(\alpha/2)} + u_{1-(\alpha/2)} \times \sqrt{\frac{\widehat{\pi}_{n,A}(obs)(1-\widehat{\pi}_{n,A}(obs))}{n} + \frac{u^2_{1-(\alpha/2)}}{4n^2}}}{1+\frac{1}{n}u^2_{1-(\alpha/2)}},$$

où $u_{1-(\alpha/2)}$ est le quantile de la loi normale centrée et réduite.

L'intervalle de confiance pour la proportion π_A au niveau de confiance $(1-\alpha)$ est égal à

$$\widehat{\pi}_{n,A}(obs) - u_{1-(\alpha/2)} \times \sqrt{\frac{\widehat{\pi}_{n,A}(obs)(1-\widehat{\pi}_{n,A}(obs))}{n}} < \pi_A$$
$$< \widehat{\pi}_{n,A}(obs) + u_{1-(\alpha/2)} \times \sqrt{\frac{\widehat{\pi}_{n,A}(obs)(1-\widehat{\pi}_{n,A}(obs))}{n}},$$

où $u_{1-\alpha/2}$ est le quantile de la loi normale centrée et réduite.

Remarques

1. Il existe pour ces deux intervalles une formule avec une correction de continuité de Yates pour tenir compte du passage d'une loi discrète à une loi continue. Donc, parfois, nous pouvons constater un léger écart entre le calcul de cette formule si nous le réalisons à la main et le calcul donné par un logiciel de statistique. Donc une recommandation que nous pouvons faire : regarder ce qui est programmé dans le manuel du logiciel qui est utilisé.

Remarques

2. Pour établir l'intervalle de confiance par la méthode de Wald, l'approximation de la loi binomiale par la loi normale a été utilisée. Il faut se rappeler que pour utiliser cette approximation, certaines conditions doivent être remplies.
3. Enfin, il manque la formule mathématique de l'intervalle de confiance par la méthode de Clopper-Pearson. Elle n'a pas été présentée ici car le calcul est un peu plus compliqué que les deux autres. Néanmoins elle peut être retrouvée facilement sur internet.

Exemple : La bactérie Brucella abortus

Dans le cas d'une contamination d'un grand cheptel bovin par la bactérie *Brucella abortus*, un vétérinaire observe 53 avortements pour 134 vaches gestantes.

Quelle proportion d'avortements peut-il prédire dans le cheptel au seuil de confiance égal à 95% ?

Au sujet de la bactérie *Brucella abortus*, nous renvoyons le lecteur au rapport de Juillet 2015 rédigé par l'Anses et téléchargeable à l'adresse suivante :

`https://www.anses.fr/fr/system/files/SANT2014sa0218Ra.pdf`

La bactérie Brucella abortus : suite et fin

L'estimation ponctuelle de la proportion d'avortements est égale à : 53/134.

L'intervalle de confiance asymptotique à 95% pour la proportion d'avortements est égal à :

$$]0,3127336; 0,4783111[$$

Les commandes sous R pour obtenir les résultats ci-dessus sont les suivantes :

```
> pe<-53/134
> pe-qnorm(0.975)*sqrt((pe*(1-pe))/134)
[1] 0.3127336
> pe+qnorm(0.975)*sqrt((pe*(1-pe))/134)
[1] 0.4783111
```