Statistiques

Intervalles de confiance

Julian Tugaut

Télécom Saint-Étienne

Sommaire

Plan

Préliminaires - 1

Les estimations ponctuelles fournies dans les deux chapitres précédents sont totalement dépendantes de l'échantillon prélevé au hasard. En effet, estimer la moyenne μ par $\overline{X_n}$, c'est faire la "meilleure estimation possible". La meilleure estimation **ponctuelle** possible du moins.

Préliminaires - 2

La variable aléatoire $\overline{X_n}$ n'étant pas une constante, selon l'aléa sous-jacent ω, on obtiendra ainsi un résultat donné ou un autre. En tout cas, il est très improbable que l'on ait $\overline{X_n}(\omega) = \mu$. Pire, si le modèle paramétrique sous-jacent est à densité, alors la probabilité que $\overline{X_n}$ soit exactement égal à μ est 0.

Préliminaires - 2

La variable aléatoire $\overline{X_n}$ n'étant pas une constante, selon l'aléa sous-jacent ω, on obtiendra ainsi un résultat donné ou un autre. En tout cas, il est très improbable que l'on ait $\overline{X_n}(\omega) = \mu$. Pire, si le modèle paramétrique sous-jacent est à densité, alors la probabilité que $\overline{X_n}$ soit exactement égal à μ est 0.
Le même phénomène a lieu avec S_n^2 et avec $\widetilde{S_n^2}$.

Philosophie - 1

En général, on ne peut pas évaluer exactement l'erreur que l'on commet en faisant une estimation ponctuelle. Néanmoins, en étudiant la loi de variabilité de cette erreur, on peut calculer une marge d'erreur pour laquelle il existe une certitude suffisante qu'elle ne sera pas dépassée.

Philosophie - 1

En général, on ne peut pas évaluer exactement l'erreur que l'on commet en faisant une estimation ponctuelle. Néanmoins, en étudiant la loi de variabilité de cette erreur, on peut calculer une marge d'erreur pour laquelle il existe une certitude suffisante qu'elle ne sera pas dépassée.
Typiquement, on se donne un risque α que l'on est prêt à accepter. Le risque en question dépend des usages que l'on en fait. Il peut s'agir de 10%, 5%, 1% voire 0.5% selon les sciences dans lesquelles on applique les intervalles de confiance en question.

Philosophie - 2

Toute la difficulté est maintenant de fournir un intervalle de $\mathbb{R}$ dans lequel le paramètre à estimer appartient avec une probabilité d'au moins $1-\alpha$. Cela peut être pour une estimation de la moyenne, de la variance...

Philosophie - 3

Il convient de noter que l'intervalle n'a aucune raison d'être unique. Également, plus le risque est faible, plus l'intervalle est large. Enfin, les intervalles de confiance que l'on va construire n'ont aucune raison d'être symétriques par rapport à l'estimateur sur lequel on se fonde. Il est cependant vrai que si l'estimateur est sans biais comme c'est le cas pour la moyenne dans le modèle gaussien, alors on prendra plutôt un intervalle symétrique. Néanmoins, nous verrons que dans le cas de la variance (même dans le modèle gaussien), l'intervalle est construit de manière très différente.

Philosophie - 4

Le degré de confiance à savoir $1 - \alpha$ est aussi appelé le niveau de confiance ou le seuil de confiance.

Philosophie - 4

Le degré de confiance à savoir $1-\alpha$ est aussi appelé le niveau de confiance ou le seuil de confiance.
Pour la moyenne μ, en notant par γ le degré de confiance et par d la marge d'erreur, on doit avoir

$$\gamma = \mathbb{P}\left(\left|\overline{X}_n - \mu\right| \leq d\right) ,$$

où $\overline{X}_n$ est la moyenne d'échantillon. C'est à partir de cette équation de base que sont déterminées toutes les quantités qui entrent en jeu dans un intervalle de confiance ; pour la moyenne.

Définitions - 1

Définition : Intervalle de confiance

Soit X une variable aléatoire suivant une loi qui dépend d'un paramètre inconnu $\theta^* \in \Theta \subset \mathbb{R}$ de la loi parente. Les intervalles de confiance de θ^* au seuil de risque $\alpha \in]0;1[$ sont des intervalles

$$\mathcal{I}_\alpha := [a(X_1, \cdots, X_n, \alpha); b(X_1, \cdots, X_n, \alpha)] \,,$$

construits d'une façon telle qu'*a priori* une proportion $1 - \alpha$ de ces intervalles contienne la valeur de θ^*.

Définitions - 2

Définition

La quantité $1-\alpha$ est ici appelée niveau de confiance de l'intervalle.

Définitions - 2

Définition

La quantité $1-\alpha$ est ici appelée niveau de confiance de l'intervalle.

Remarque

Il est également nécessaire que l'intervalle $\mathcal{I}_\alpha$ ne dépende pas d'autres paramètres inconnus. En effet, il ne serait pas raisonnable de construire un intervalle de confiance pour un paramètre inconnu en utilisant pour cela un paramètre que l'on ne connaît même pas.

Plan

1. Préliminaires
2. Cas gaussien
   - Estimation de μ si σ^2 est connue
   - Estimation de μ si σ^2 est inconnue
   - Estimation de σ^2 si μ est connue
   - Estimation de σ^2 si μ est inconnue
3. Intervalles de confiance asymptotiques
4. Cas exponentiel

Introduction du cas gaussien

Dans cette section, on suppose que l'on dispose d'un modèle paramétrique gaussien. En d'autres termes, les variables aléatoires sous-jacentes suivent une loi normale de paramètres μ et σ^2.

Introduction du cas gaussien

Dans cette section, on suppose que l'on dispose d'un modèle paramétrique gaussien. En d'autres termes, les variables aléatoires sous-jacentes suivent une loi normale de paramètres μ et σ^2. L'objectif est ainsi d'estimer par intervalle de confiance la moyenne μ et la variance σ^2.

Intervalle de la moyenne en connaissant la variance - 1

On commence par estimer la moyenne μ. Pour ce faire, on fait l'hypothèse peu raisonnable (mais qui fournit un cadre simple pour s'entraîner à construire des intervalles de confiance) que la variance σ^2 est connue.

Intervalle de la moyenne en connaissant la variance - 1

On commence par estimer la moyenne μ. Pour ce faire, on fait l'hypothèse peu raisonnable (mais qui fournit un cadre simple pour s'entraîner à construire des intervalles de confiance) que la variance σ^2 est connue.
On sait alors que la variable aléatoire $\overline{X_n}$ suit une loi normale de paramètres μ et $\frac{\sigma^2}{n}$. De fait, la variable aléatoire $Y_n := \frac{\sqrt{n}}{\sigma}\left(\overline{X_n} - \mu\right)$ suit la loi normale centrée réduite.

Intervalle de la moyenne en connaissant la variance - 2

En notant $q_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite, il vient

$$\begin{aligned}
\mathbb{P}\left(|Y_n| \leq q_{1-\frac{\alpha}{2}}\right) &= \mathbb{P}\left(-q_{1-\frac{\alpha}{2}} \leq Y_n \leq q_{1-\frac{\alpha}{2}}\right) \\
&= \Phi\left(q_{1-\frac{\alpha}{2}}\right) - \Phi(-q_{1-\frac{\alpha}{2}}) \\
&= \left(1-\frac{\alpha}{2}\right) - \left(1-\left(1-\frac{\alpha}{2}\right)\right) \\
&= 1-\alpha\,.
\end{aligned}$$

Intervalle de la moyenne en connaissant la variance - 3

Ainsi, on a

$$\mathbb{P}\left(-q_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}}{\sigma}\left(\overline{X_n} - \mu\right) \leq q_{1-\frac{\alpha}{2}}\right) = 1 - \alpha\,,$$

ce qui donne

$$\mathbb{P}\left(\mu - \frac{\sigma}{\sqrt{n}}q_{1-\frac{\alpha}{2}} \leq \overline{X_n} \leq \mu + \frac{\sigma}{\sqrt{n}}q_{1-\frac{\alpha}{2}}\right) = 1 - \alpha\,.$$

Intervalle de la moyenne en connaissant la variance - 4

On en déduit

$$\mathbb{P}\left(\mu \in \mathcal{I}_\alpha(n)\right) = 1 - \alpha\,,$$

où

$$\mathcal{I}_\alpha(n) := \left[\overline{X_n} - \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\,; \overline{X_n} + \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right]\,,$$

est l'intervalle de confiance de μ de risque α.

Intervalle de la moyenne sans connaître la variance - 1

Dans un cas plus général, on ne dispose pas de la variance. En effet, comment pourrait-on connaître la variance alors qu'on ignore la moyenne ?

Intervalle de la moyenne sans connaître la variance - 1

Dans un cas plus général, on ne dispose pas de la variance. En effet, comment pourrait-on connaître la variance alors qu'on ignore la moyenne ?
Dans ce cas, on estime σ^2 par la variance d'échantillon corrigée à savoir la variable aléatoire $\widetilde{S_n^2} := \frac{1}{n-1} \sum_{i=1}^n \left(X_i - \overline{X_n}\right)^2$.

Intervalle de la moyenne sans connaître la variance - 2

On va maintenant se ramener à une loi connue qui ne dépend ni de σ^2 ni de μ. On sait, pour commencer que la variable aléatoire $Y_n := \frac{\sqrt{n}}{\sigma}\left(\overline{X_n} - \mu\right)$ suit la loi normale centrée réduite, voir le paragraphe précédent. On sait également que $Z_n := \frac{n-1}{\sigma^2}\widetilde{S_n^2}$ suit la loi du Khi-deux à $n-1$ degrés de liberté. Puis, en admettant que Z_n et Y_n sont indépendantes (ce qui se démontre avec le Théorème de Cochran), on en déduit que la variable aléatoire

$$T_n := \frac{Y_n}{\sqrt{\frac{Z_n}{n-1}}} = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\sqrt{\widetilde{S_n^2}}}$$

suit la loi de Student à $n-1$ degrés de liberté.

Intervalle de la moyenne sans connaître la variance - 3

Le nouvel intervalle de confiance est donc

$$\mathcal{I}_\alpha(n) := \left[\overline{X_n} - \frac{\widetilde{S_n}}{\sqrt{n}} t_{1-\frac{\alpha}{2}} ; \overline{X_n} + \frac{\widetilde{S_n}}{\sqrt{n}} t_{1-\frac{\alpha}{2}}\right] ,$$

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

Intervalle de la variance en connaissant la moyenne - 1

Ici, on cherche à estimer σ^2. On suppose que l'on connaît la moyenne μ.

Intervalle de la variance en connaissant la moyenne - 1

Ici, on cherche à estimer σ^2. On suppose que l'on connaît la moyenne μ.
On sait alors que les variables aléatoires $\frac{X_i-\mu}{\sigma}$ suivent une loi normale centrée réduite. De plus, elles sont mutuellement indépendantes. De fait,

$$\chi^2 := \sum_{k=1}^{n} \left(\frac{X_k - \mu}{\sigma} \right)^2$$

suit une loi du Khi-deux à n degrés de liberté.

Intervalle de la variance en connaissant la moyenne - 2

Par conséquent,

$$\mathbb{P}\left(k_1 \leq \chi^2 \leq k_2\right) = (1 - \frac{\alpha}{2}) - \frac{\alpha}{2} = 1 - \alpha\,,$$

où k_1 et k_2 sont respectivement les quantiles d'ordre $1 - \frac{\alpha}{2}$ et $\frac{\alpha}{2}$ de la loi du khi-deux à n degrés de liberté.

Intervalle de la variance en connaissant la moyenne - 3

On en déduit

$$\mathbb{P}\left(\sigma^2 \in \mathcal{I}_\alpha\right) = 1 - \alpha\,,$$

où l'intervalle de confiance de la variance $\mathcal{I}_\alpha$ au risque α est

$$\mathcal{I}_\alpha := \left[\frac{1}{k_2}\sum_{i=1}^{n}(X_i-\mu)^2; \frac{1}{k_1}\sum_{i=1}^{n}(X_i-\mu)^2\right]\,.$$

Intervalle de la variance en connaissant la moyenne - 4

Remarque

Il peut sembler curieux à première lecture que l'on ne divise pas par n et on pourrait croire que l'intervalle de confiance ne devient pas meilleur pour peu que n augmente. Il faut en fait se rappeler que k_1 comme k_2 dépendent de n vu qu'ils sont des quantiles de la loi du khi-deux à n degrés de liberté.

Intervalle de la variance sans connaître la moyenne - 1

On suppose maintenant que la moyenne est également inconnue. Il est naturel de remplacer cette moyenne μ par son estimation ponctuelle $\overline{X_n}$. Néanmoins, on perd alors un degré de liberté d'après le Théorème de Cochran.

Intervalle de la variance sans connaître la moyenne - 2

Ainsi, un intervalle de confiance de la variance $\mathcal{I}_\alpha$ au risque α est désormais :

$$\mathcal{I}_\alpha := \left[\frac{1}{k_2'}\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2 ; \frac{1}{k_1'}\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2\right] ,$$

où k_1' et k_2' sont respectivement lees quantiles d'ordre $1 - \frac{\alpha}{2}$ et $\frac{\alpha}{2}$ de la loi du khi-deux à $n-1$ degrés de liberté.

Plan

Préliminaires - 1

Précédemment, on a établi des intervalles de confiance lorsque les variances sont supposées égales ou même lorsqu'elles sont connues. On peut se demander ce qu'il se passe lorsqu'elles ne sont ni connues ni supposées égales. Néanmoins, dans ce cas de figure, l'intervalle de confiance sera asymptotique, c'est-à-dire que la probabilité pour que $\mu_1 - \mu_2$ soit dans ledit intervalle tend vers $1 - \alpha$ quand N et M tendent vers l'infini. Nous avons choisi de ne pas détailler ce cas.

Préliminaires - 2

À la place, nous allons présenter les intervalles de confiance asymptotiques dans des cas simples. D'abord, nous étudierons le cas de la loi de Poisson puis nous nous intéresserons à la loi de Bernoulli pour la proportion d'échantillon.

Loi de Poisson - 1

Supposons que les variables aléatoires soient indépendantes et identiquement distribuées suivant une loi de Poisson de paramètre $\theta > 0$ inconnu. On sait, d'après le Théorème central de la limite que $Y_n := \sqrt{n}(\overline{X_n} - \theta)$ converge en loi vers la loi normale centrée et de variance $\theta > 0$.

Loi de Poisson - 2

Or, $\overline{X_n}$ estime le paramètre de variance θ. En effet, $\overline{X_n}$ converge presque sûrement vers θ d'après la loi forte des grands nombres. Puis, cela implique que $\frac{1}{\sqrt{\overline{X_n}}}$ converge presque sûrement et donc en loi vers $\frac{1}{\sqrt{\theta}}$. Puis, comme Y_n converge en loi vers la loi normale centrée de variance θ, il s'ensuit que $\frac{Y_n}{\sqrt{\overline{X_n}}}$ converge en loi vers la gaussienne centrée réduite, d'après le lemme de Slutsky.

Loi de Poisson - 3

On obtient ainsi un intervalle de confiance asymptotique :

$$\mathcal{I}_\alpha := \left[\overline{X_n} - q_{1-\frac{\alpha}{2}} \frac{\sqrt{\overline{X_n}}}{\sqrt{n}} ; \overline{X_n} + q_{1-\frac{\alpha}{2}} \frac{\sqrt{\overline{X_n}}}{\sqrt{n}}\right] ,$$

où $q_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite.

Loi de Bernoulli - 1

On s'intéresse ici à la proportion d'échantillon. En d'autres termes, les variables aléatoires $(X_n)_n$ sont indépendantes et identiquement distribuées suivant la loi de Bernoulli de paramètre p.

Loi de Bernoulli - 1

On s'intéresse ici à la proportion d'échantillon. En d'autres termes, les variables aléatoires $(X_n)_n$ sont indépendantes et identiquement distribuées suivant la loi de Bernoulli de paramètre p.
Alors, $n\overline{X_n}$ suit une loi binomiale de paramètres n et p donc l'espérance de $\overline{X_n}$ est p tandis que sa variance est $\frac{p(1-p)}{n}$.

Loi de Bernoulli - 2

En utilisant le Théorème central de la limite, on aboutit à l'intervalle de confiance asymptotique suivant au seuil de risque α :

$$\mathcal{I}_\alpha := \left[\overline{X_n} - q_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X_n}(1-\overline{X_n})}{n}} ; \overline{X_n} + q_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X_n}(1-\overline{X_n})}{n}}\right] ,$$

où $q_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite.

Loi de Bernoulli - 3

Cet intervalle est dû à la méthode de Wald. Il existe d'autres méthodes (lesquelles présentent certains avantages) comme la méthode du score de Wilson. On peut également appliquer ce que l'on appelle la correction de continuité de Yates.

Loi de Bernoulli - 3

Cet intervalle est dû à la méthode de Wald. Il existe d'autres méthodes (lesquelles présentent certains avantages) comme la méthode du score de Wilson. On peut également appliquer ce que l'on appelle la correction de continuité de Yates.
Néanmoins, ici, nous ne présentons que les bases des statistiques. Il faut donc simplement être conscient que cela existe.

Plan

Préliminaires

Ici, on présente le cas où les variables aléatoires sous-jacentes suivent la loi exponentielle de paramètre $\theta > 0$.

Préliminaires

Ici, on présente le cas où les variables aléatoires sous-jacentes suivent la loi exponentielle de paramètre $\theta > 0$.
On a donc $\mathbb{E}[X_i] = \frac{1}{\theta}$ et $\mathrm{Var}[X_i] = \frac{1}{\theta^2}$. Ainsi :

$$\mathbb{E}[\overline{X_n}] = \frac{1}{\theta} \quad \text{et } \mathrm{Var}[\overline{X_n}] = \frac{1}{n\theta^2} \,.$$

On va ici chercher un intervalle de confiance pour θ.

Intervalles asymptotiques - 1

On commence avec un intervalle asymptotique. On applique le théorème central de la limite et l'on aboutit à la convergence en loi de $\frac{\sqrt{n}\left(\overline{X_n}-\frac{1}{\theta}\right)}{\frac{1}{\theta}} = \sqrt{n}(\theta\overline{X_n} - 1)$ vers la loi normale centrée réduite.

Intervalles asymptotiques - 2

De fait, un intervalle de confiance asymptotique immédiat de θ est

$$\mathcal{I}_\alpha := \left[\frac{1}{\overline{X_n}} - \frac{1}{\sqrt{n}} \frac{q_{1-\frac{\alpha}{2}}}{\overline{X_n}} ; \frac{1}{\overline{X_n}} + \frac{1}{\sqrt{n}} \frac{q_{1-\frac{\alpha}{2}}}{\overline{X_n}} \right] .$$

Intervalles asymptotiques - 2

De fait, un intervalle de confiance asymptotique immédiat de θ est

$$\mathcal{I}_\alpha := \left[\frac{1}{\overline{X_n}} - \frac{1}{\sqrt{n}} \frac{q_{1-\frac{\alpha}{2}}}{\overline{X_n}} ; \frac{1}{\overline{X_n}} + \frac{1}{\sqrt{n}} \frac{q_{1-\frac{\alpha}{2}}}{\overline{X_n}} \right] .$$

Néanmoins, il est essentiel de bien comprendre que cet intervalle n'a d'intérêt que pour des grandes valeurs de n.

Intervalles non asymptotiques - 1

On rappelle que la somme de n variables aléatoires indépendantes et suivant une même loi exponentielle suit la loi d'Erlang, laquelle est une instance particulière de la loi Gamma.
Ainsi, $n\theta\overline{X_n}$ suit la loi d'Erlang de paramètres n et 1.

Intervalles non asymptotiques - 2

Il s'ensuit qu'un intervalle de confiance (non asymptotique) du paramètre θ est

$$\mathcal{I}_\alpha := \left[\frac{q_{\frac{\alpha}{2}}}{n\overline{X_n}} ; \frac{q_{1-\frac{\alpha}{2}}}{n\overline{X_n}} \right] ,$$

où $q_{\frac{\alpha}{2}}$ et $q_{1-\frac{\alpha}{2}}$ sont les quantiles d'ordre respectifs $\frac{\alpha}{2}$ et $1 - \frac{\alpha}{2}$ de la loi d'Erlang de paramètres n et 1.