Chapitre 4 : Intervalles de confiance

- Problématique et définition
- Intervalle de confiance pour la moyenne de la loi normale
- Intervalle de confiance pour la variance de la loi normale
- Intervalle de confiance pour une proportion

Problématique et définition

Objectif : plutôt qu'estimer un paramètre $\theta \in \mathbf{R}$ par une unique valeur $\hat{\theta}_n$ (estimation ponctuelle), proposer un ensemble de valeurs vraisemblables pour θ (estimation ensembliste).

Problématique et définition

Objectif : plutôt qu'estimer un paramètre $\theta \in \mathbf{R}$ par une unique valeur $\hat{\theta}_n$ (estimation ponctuelle), proposer un ensemble de valeurs vraisemblables pour θ (estimation ensembliste).

Un **intervalle de confiance** de **seuil** $\alpha \in [0, 1]$ pour un paramètre θ, est un intervalle aléatoire I tel que

$$P(\theta \in I) = 1 - \alpha$$

Problématique et définition

Objectif : plutôt qu'estimer un paramètre $\theta \in \mathbf{R}$ par une unique valeur $\hat{\theta}_n$ (estimation ponctuelle), proposer un ensemble de valeurs vraisemblables pour θ (estimation ensembliste).

Un **intervalle de confiance** de **seuil** $\alpha \in [0, 1]$ pour un paramètre θ, est un intervalle aléatoire I tel que

$$P(\theta \in I) = 1 - \alpha$$

α est la probabilité que l'on se trompe en affirmant que $\theta \in I$.

Valeurs usuelles de α : 10%, 5%, 1%, etc.

Interprétation

$I = [Z_1, Z_2]$. Z_1 et Z_2 sont des variables aléatoires.

$$P(Z_1 \leq \theta \leq Z_2) = 1 - \alpha$$

Interprétation

$I = [Z_1, Z_2]$. Z_1 et Z_2 sont des variables aléatoires.

$$P(Z_1 \leq \theta \leq Z_2) = 1 - \alpha$$

Soient z_1 et z_2 les réalisations de Z_1 et Z_2 pour une expérience donnée.

θ est ou n'est pas dans l'intervalle déterministe $[z_1, z_2]$.

Interprétation

$I = [Z_1, Z_2]$. Z_1 et Z_2 sont des variables aléatoires.

$$P(Z_1 \leq \theta \leq Z_2) = 1 - \alpha$$

Soient z_1 et z_2 les réalisations de Z_1 et Z_2 pour une expérience donnée.

θ est ou n'est pas dans l'intervalle déterministe $[z_1, z_2]$.

$\Rightarrow$ On ne peut pas dire : "θ a 95% de chances d'être compris entre 61 et 66".

Interprétation

$I = [Z_1, Z_2]$. Z_1 et Z_2 sont des variables aléatoires.

$$P(Z_1 \leq \theta \leq Z_2) = 1 - \alpha$$

Soient z_1 et z_2 les réalisations de Z_1 et Z_2 pour une expérience donnée.

θ est ou n'est pas dans l'intervalle déterministe $[z_1, z_2]$.

$\Rightarrow$ On ne peut pas dire : "θ a 95% de chances d'être compris entre 61 et 66".

Si on recommence 100 fois l'expérience, on aura 100 réalisations du couple (Z_1, Z_2), et donc 100 intervalles de confiance différents.

Si $\alpha = 5\%$, en moyenne θ sera dans 95 de ces intervalles.

Interprétation

$I = [Z_1, Z_2]$. Z_1 et Z_2 sont des variables aléatoires.

$$P(Z_1 \leq \theta \leq Z_2) = 1 - \alpha$$

Soient z_1 et z_2 les réalisations de Z_1 et Z_2 pour une expérience donnée.

θ est ou n'est pas dans l'intervalle déterministe $[z_1, z_2]$.

$\Rightarrow$ On ne peut pas dire : "θ a 95% de chances d'être compris entre 61 et 66".

Si on recommence 100 fois l'expérience, on aura 100 réalisations du couple (Z_1, Z_2), et donc 100 intervalles de confiance différents.

Si $\alpha = 5\%$, en moyenne θ sera dans 95 de ces intervalles.

$\Rightarrow$ On peut dire : "on a une confiance de 95% dans le fait que θ soit compris entre 61 et 66".

Comment trouver un intervalle de confiance ?

Idée naturelle : proposer $I = \left[\hat{\theta}_n - \varepsilon, \hat{\theta}_n + \varepsilon\right]$.

Comment trouver un intervalle de confiance ?

Idée naturelle : proposer $I = \left[\hat{\theta}_n - \varepsilon, \hat{\theta}_n + \varepsilon\right]$.

Alors il faut déterminer ε de sorte que :

$$P(\theta \in I) = P(\hat{\theta}_n - \varepsilon \leq \theta \leq \hat{\theta}_n + \varepsilon) = P(|\hat{\theta}_n - \theta| \leq \varepsilon) = 1 - \alpha$$

Comment trouver un intervalle de confiance ?

Idée naturelle : proposer $I = \left[\hat{\theta}_n - \varepsilon, \hat{\theta}_n + \varepsilon\right]$.

Alors il faut déterminer ε de sorte que :

$$P(\theta \in I) = P(\hat{\theta}_n - \varepsilon \leq \theta \leq \hat{\theta}_n + \varepsilon) = P(|\hat{\theta}_n - \theta| \leq \varepsilon) = 1 - \alpha$$

Cette démarche ne peut aboutir que si la loi de probabilité de $\hat{\theta}_n - \theta$ ne dépend pas de θ.

Comment trouver un intervalle de confiance ?

Idée naturelle : proposer $I = \left[\hat{\theta}_n - \varepsilon, \hat{\theta}_n + \varepsilon\right]$.

Alors il faut déterminer ε de sorte que :

$$P(\theta \in I) = P(\hat{\theta}_n - \varepsilon \leq \theta \leq \hat{\theta}_n + \varepsilon) = P(|\hat{\theta}_n - \theta| \leq \varepsilon) = 1 - \alpha$$

Cette démarche ne peut aboutir que si la loi de probabilité de $\hat{\theta}_n - \theta$ ne dépend pas de θ.

La méthode la plus efficace consiste à chercher une **fonction pivotale**, c'est à dire une variable aléatoire fonction à la fois du paramètre θ et des observations $X_1, \ldots, X_n$, dont la loi de probabilité ne dépende pas de θ.

Théorème de Fisher

Si $X_1, \ldots, X_n$ sont indépendantes et de même loi $\mathcal{N}\left(m, \sigma^2\right)$, alors on a :

- $\sum_{i=1}^{n} X_i$ est de loi $\mathcal{N}\left(nm, n\sigma^2\right)$.
- $\bar{X}_n$ est de loi $\mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$.
- $\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - m)^2$ est de loi χ^2_n.
- $\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \frac{nS_n^2}{\sigma^2}$ est de loi χ^2_{n-1}.
- $\bar{X}_n$ et S_n^2 sont indépendantes.
- $\sqrt{n}\,\frac{\bar{X}_n - m}{S'_n} = \sqrt{n-1}\,\frac{\bar{X}_n - m}{S_n}$ est de loi de Student $St(n-1)$.

Intervalle de confiance pour la moyenne de la loi normale

$X_1, \ldots, X_n$ indépendantes et de même loi normale $\mathcal{N}\left(m, \sigma^2\right)$.

L'ESBVM de m est $\bar{X}_n$.

$\Rightarrow$ on cherche un IC pour m de la forme $\left[\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon\right]$.

Intervalle de confiance pour la moyenne de la loi normale

$X_1, \ldots, X_n$ indépendantes et de même loi normale $\mathcal{N}\left(m, \sigma^2\right)$.

L'ESBVM de m est $\bar{X}_n$.

$\Rightarrow$ on cherche un IC pour m de la forme $\left[\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon\right]$.

Un intervalle de confiance de seuil α pour m est :

$$\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}}\, u_\alpha, \bar{X}_n + \frac{\sigma}{\sqrt{n}}\, u_\alpha\right]$$

Intervalle de confiance pour la moyenne de la loi normale

$X_1, \ldots, X_n$ indépendantes et de même loi normale $\mathcal{N}\left(m, \sigma^2\right)$.

L'ESBVM de m est $\bar{X}_n$.

$\Rightarrow$ on cherche un IC pour m de la forme $\left[\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon\right]$.

Un intervalle de confiance de seuil α pour m est :

$$\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}}\, u_\alpha, \bar{X}_n + \frac{\sigma}{\sqrt{n}}\, u_\alpha\right]$$

Problème : cet intervalle n'est utilisable que si on connaît la valeur de σ.

Intervalle de confiance pour la moyenne de la loi normale

$X_1, \ldots, X_n$ indépendantes et de même loi normale $\mathcal{N}\left(m, \sigma^2\right)$.

L'ESBVM de m est $\bar{X}_n$.

$\Rightarrow$ on cherche un IC pour m de la forme $\left[\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon\right]$.

Un intervalle de confiance de seuil α pour m est :

$$\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}}\, u_\alpha, \bar{X}_n + \frac{\sigma}{\sqrt{n}}\, u_\alpha\right]$$

Problème : cet intervalle n'est utilisable que si on connaît la valeur de σ.

Idée : remplacer σ par un estimateur, par exemple S'_n.

Intervalle de confiance pour la moyenne de la loi normale

Un intervalle de confiance de seuil α pour m est :

$$\left[\bar{X}_n - \frac{S'_n}{\sqrt{n}}\, t_{n-1,\alpha}, \bar{X}_n + \frac{S'_n}{\sqrt{n}}\, t_{n-1,\alpha}\right]$$

Exemple des niveaux de bruit : $n = 20$, $\bar{x}_n = 64.2$ et $s'_n = 5.15$.
Pour $\alpha = 5\%$, la table de la loi de Student donne $t_{19,0.05} = 2.093$.
On en déduit qu'un intervalle de confiance de seuil 5% pour le niveau de bruit moyen est [61.8, 66.7].

Interprétation : La meilleure estimation possible du niveau de bruit moyen est 64.2 db. De plus, on a une confiance de 95% dans le fait que ce niveau de bruit moyen est compris entre 61.8 db et 66.7 db.

Script R : calcul brut

```
>   bruit <- c(54.8, 55.4, 57.7, 59.6, 60.1, 61.2, 62.0,
   63.1, 63.5, 64.2, 65.2, 65.4, 65.9, 66.0, 67.6, 68.1,
   69.5, 70.6, 71.5, 73.4)
>   n <- length(bruit)
>   alpha <- 0.05
>   mean(bruit)-sd(bruit)*qt(1-alpha/2,n-1)/sqrt(n)
[1] 61.82992
>   mean(bruit)+sd(bruit)*qt(1-alpha/2,n-1)/sqrt(n)
[1] 66.65008
```

Script R : à l'aide de la commande t.test

```
>   t.test(bruit, conf.level=0.95)

        One Sample t-test

data: bruit
t = 55.7889, df = 19, p-value <  2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 61.82992 66.65008
sample estimates:
mean of x
    64.24
```

Largeur de l'intervalle de confiance

La largeur de l'IC est $2\frac{S'_n}{\sqrt{n}}\, t_{n-1,\alpha}$.

C'est une fonction décroissante en n comme en α :

- plus on a d'observations, plus on a d'informations, donc plus l'incertitude sur le paramètre diminue et plus l'intervalle de confiance est étroit.
- plus α est petit, moins on veut prendre de risques de se tromper en disant que m est dans l'intervalle, donc plus on aura tendance à prendre des intervalles larges.

En pratique, un intervalle de confiance trop large n'a aucun intérêt, donc il faut parfois accepter un risque d'erreur relativement fort pour obtenir un intervalle de confiance utilisable.

Fonction pivotale

Le point-clé pour trouver l'IC est l'utilisation de la variable aléatoire $\sqrt{n}\,\dfrac{\bar{X}_n - m}{S'_n}$.

C'est une fonction des observations $X_1, \ldots, X_n$ et du paramètre m, dont la loi de probabilité $(St(n-1))$ ne dépend pas des paramètres du modèle m et σ^2.

$\Rightarrow$ C'est ce qu'on a appelé une **fonction pivotale**.

A partir de maintenant, c'est ce que nous utiliserons pour construire des intervalles de confiance.

Intervalle de confiance pour la variance de la loi normale

L'ESBVM de σ^2 est S'^2_n.

Fonction pivotale : $\dfrac{nS_n^2}{\sigma^2}$ est de loi χ^2_{n-1}.

Intervalle de confiance pour la variance de la loi normale

L'ESBVM de σ^2 est S'^2_n.

Fonction pivotale : $\dfrac{nS_n^2}{\sigma^2}$ est de loi χ^2_{n-1}.

Un intervalle de confiance de seuil α pour σ^2 est :

$$\left[\frac{nS_n^2}{z_{n-1,\alpha/2}}, \frac{nS_n^2}{z_{n-1,1-\alpha/2}}\right] = \left[\frac{(n-1)S'^2_n}{z_{n-1,\alpha/2}}, \frac{(n-1)S'^2_n}{z_{n-1,1-\alpha/2}}\right]$$

Intervalle de confiance pour la variance de la loi normale

L'ESBVM de σ^2 est S'^2_n.

Fonction pivotale : $\frac{nS_n^2}{\sigma^2}$ est de loi χ^2_{n-1}.

Un intervalle de confiance de seuil α pour σ^2 est :

$$\left[\frac{nS_n^2}{z_{n-1,\alpha/2}}, \frac{nS_n^2}{z_{n-1,1-\alpha/2}}\right] = \left[\frac{(n-1)S'^2_n}{z_{n-1,\alpha/2}}, \frac{(n-1)S'^2_n}{z_{n-1,1-\alpha/2}}\right]$$

Exemple des niveaux de bruit : $n = 20$ et $s'^2_n = 26.5$. $z_{19,0.025} = 32.85$ et $z_{19,0.975} = 8.91$. On en déduit qu'un intervalle de confiance de seuil 5% pour la variance du niveau de bruit est [15.3, 56.6].

Remarque : L'IC est de la forme $\left[\varepsilon_1 S'^2_n, \varepsilon_2 S'^2_n\right]$ avec $\varepsilon_1 < 1$ et $\varepsilon_2 > 1$ et non pas de la forme $\left[S'^2_n - \varepsilon, S'^2_n + \varepsilon\right]$.

Intervalle de confiance pour une proportion

Détermination d'un intervalle de confiance pour le paramètre p de la loi de Bernoulli, au vu d'un échantillon $X_1, \ldots, X_n$ de cette loi.

On a montré que l'ESBVM de p est $\hat{p}_n = \bar{X}_n$.

Exemple : Sondages

- Une élection oppose deux candidats A et B.
- Un institut de sondage interroge 800 personnes sur leurs intentions de vote. 420 déclarent voter pour A et 380 pour B.
- Estimer le résultat de l'élection, c'est estimer le pourcentage p de voix qu'obtiendra le candidat A le jour de l'élection.
- *Hypothèses* : les n personnes interrogées ont des votes indépendants et la probabilité qu'une personne choisie au hasard vote pour A est p.

Sondages

$$x_i = \begin{cases} 1 & \text{si la } i^{\text{ème}} \text{ personne interrogée déclare voter pour A} \\ 0 & \text{sinon} \end{cases}$$

Alors X_i est bien de loi $\mathcal{B}(p)$ et les X_i sont indépendantes.

L'ESBVM de p est $\hat{p}_n = \bar{x}_n = 420/800 = 52.5\%$.

L'institut de sondage estime donc que le candidat A va gagner l'élection.

Pour évaluer l'incertitude portant sur cette estimation, A demande un intervalle de confiance de seuil 5% pour p.

Intervalle de confiance exact

Fonction pivotale = fonction des X_i et de p dont la loi ne dépend pas de p.

On sait que $T = \sum_{i=1}^{n} X_i = n\hat{p}_n$ est de loi binomiale $\mathcal{B}(n, p)$, mais cela ne permet pas d'en déduire une fonction pivotale simple.

Intervalle de confiance exact

Fonction pivotale = fonction des X_i et de p dont la loi ne dépend pas de p.

On sait que $T = \sum_{i=1}^{n} X_i = n\hat{p}_n$ est de loi binomiale $\mathcal{B}(n, p)$, mais cela ne permet pas d'en déduire une fonction pivotale simple.

Un intervalle de confiance exact de seuil α pour p est :

$$\left[\frac{1}{1 + \frac{n-T+1}{T} f_{2(n-T+1),2T,\alpha/2}}, \frac{1}{1 + \frac{n-T}{T+1} f_{2(n-T),2(T+1),1-\alpha/2}} \right]$$

où les $f_{\nu_1,\nu_2,\alpha}$ sont des quantiles de la loi de Fisher-Snedecor.

Intervalle de confiance asymptotique

Théorème Central-Limite :

$$\sqrt{n}\,\frac{\bar{X}_n - E(X)}{\sqrt{Var(X)}} = \sqrt{n}\,\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} = \frac{T - np}{\sqrt{np(1-p)}}$$

est approximativement de loi $\mathcal{N}(0,1)$, ce qui fournit la fonction pivotale cherchée.

Intervalle de confiance asymptotique

Théorème Central-Limite :

$$\sqrt{n}\,\frac{\bar{X}_n - E(X)}{\sqrt{Var(X)}} = \sqrt{n}\,\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} = \frac{T - np}{\sqrt{np(1-p)}}$$

est approximativement de loi $\mathcal{N}(0,1)$, ce qui fournit la fonction pivotale cherchée.

Un intervalle de confiance asymptotique de seuil α pour p est :

$$\left[\hat{p}_n - u_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + u_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right]$$

Exemple du sondage

$n = 800$, $t = 420$ et $\hat{p}_n = 52.5\%$.

$f_{762,840,0.025} = 1.1486$, $f_{760,842,0.975} = 0.8702$ et $u_{0.05} = 1.96$.

IC exact = [0.4897, 0.5601]. IC asymptotique = [0.4904, 0.5596].

```
>   binom.test(t,n,conf.level=1-alpha)

        Exact binomial test

data: t and n
number of successes = 420, number of trials = 800, p-value =
0.1679
alternative hypothesis: true probability of success is not
equal to 0.5
95 percent confidence interval:
  0.4897328 0.5600823
sample estimates:
probability of success
                 0.525
```

Exemple du sondage, suite

On conclut que l'on a une confiance de 95% dans le fait que le pourcentage de voix qu'obtiendra le candidat A sera compris entre 49% et 56%.

Exemple du sondage, suite

On conclut que l'on a une confiance de 95% dans le fait que le pourcentage de voix qu'obtiendra le candidat A sera compris entre 49% et 56%.

Problème : cet IC n'est pas entièrement situé au-dessus de 50%.

$\Rightarrow$ il semble donc possible que, malgré l'estimation de 52.5%, le candidat A soit battu.

Exemple du sondage, suite

On conclut que l'on a une confiance de 95% dans le fait que le pourcentage de voix qu'obtiendra le candidat A sera compris entre 49% et 56%.

Problème : cet IC n'est pas entièrement situé au-dessus de 50%.

$\Rightarrow$ il semble donc possible que, malgré l'estimation de 52.5%, le candidat A soit battu.

$\Rightarrow$ ce qui importe dans cette situation, ce n'est pas vraiment d'estimer p, mais de déterminer si on peut admettre avec une confiance raisonnable que p est supérieur à 50%.

$\Rightarrow$ tests d'hypothèses.

Exemple du sondage, fin

Autre possibilité : déterminer à quelle condition l'intervalle de confiance pour p sera entièrement au-dessus des 50%.

$\Rightarrow$ il faut réduire la largeur de l'IC asymptotique $2u_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$.

- Diminuer u_α, c'est augmenter α, donc augmenter la probabilité de se tromper en affirmant que le candidat est élu.
- Augmenter n, c'est augmenter le nombre de personnes interrogées.

Exemple du sondage, fin

Autre possibilité : déterminer à quelle condition l'intervalle de confiance pour p sera entièrement au-dessus des 50%.

$\Rightarrow$ il faut réduire la largeur de l'IC asymptotique $2u_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$.

- Diminuer u_α, c'est augmenter α, donc augmenter la probabilité de se tromper en affirmant que le candidat est élu.
- Augmenter n, c'est augmenter le nombre de personnes interrogées.

$\forall p \in [0,1], p(1-p) \leq 1/4$, donc $2u_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \leq \frac{u_\alpha}{\sqrt{n}}$.

$\Rightarrow$ si on détermine n tel que $\frac{u_\alpha}{\sqrt{n}} < l$, c'est à dire $n > \frac{u_\alpha^2}{l^2}$, on est sûr que la largeur de l'intervalle de confiance sera inférieure à l.

Grenoble INP ensimag

Exemple du sondage, fin

Pour $\alpha = 5\%$ et $n = 800$, $\dfrac{u_\alpha}{\sqrt{n}} = \dfrac{1.96}{\sqrt{800}} \approx 7\%$.

Exemple du sondage, fin

Pour $\alpha = 5\%$ et $n = 800$, $\dfrac{u_\alpha}{\sqrt{n}} = \dfrac{1.96}{\sqrt{800}} \approx 7\%$.

Pour une précision inférieure à 1%, il faudra interroger au moins $\dfrac{u_\alpha^2}{l^2} = \dfrac{1.96^2}{0.01^2} = 38416$ personnes.

Exemple du sondage, fin

Pour $\alpha = 5\%$ et $n = 800$, $\dfrac{u_\alpha}{\sqrt{n}} = \dfrac{1.96}{\sqrt{800}} \approx 7\%$.

Pour une précision inférieure à 1%, il faudra interroger au moins $\dfrac{u_\alpha^2}{l^2} = \dfrac{1.96^2}{0.01^2} = 38416$ personnes.

Conclusions :

- Toujours tenir compte du nombre de personnes interrogées pour interpréter les résultats d'un sondage.
- Se méfier des conclusions péremptoires données à partir d'un échantillon de 1000 personnes.