STT 1000 - STATISTIQUES

ARTHUR CHARPENTIER

Intervalle de Confiance

Estimation ponctuelle : $\widehat{\theta}(\boldsymbol{y})$ est une simple valeur numérique

Intervalle de Confiance

Soit $\boldsymbol{Y}$ un échantillon aléatoire de variables i.i.d. de loi f_θ. Un intervalle de confiance de niveau $1-\alpha$ pour le paramètre θ est un intervalle (aléatoire) $\left[\widehat{a}(\boldsymbol{Y}),\widehat{b}(\boldsymbol{Y})\right]$ tel que

$$\mathbb{P}\Big[\theta\in\left[\widehat{a}(\boldsymbol{Y}),\widehat{b}(\boldsymbol{Y})\right]\Big]=1-\alpha$$

Classiquement, α vaut 10%, 5% ou 1%.

Plus α est petit, plus l'intervalle sera grand

Intervalle de Confiance

Intervalle de Confiance

Soit $\boldsymbol{Y}$ un échantillon aléatoire de variables i.i.d. de loi f_θ. Un intervalle de confiance unilatéral à droite de niveau $1-\alpha$ pour le paramètre θ est un intervalle (aléatoire) $\big[-\infty, \widehat{b}(\boldsymbol{Y})\big]$ tel que

$$\mathbb{P}\Big[\theta \in \big(-\infty, \widehat{b}(\boldsymbol{Y})\big]\Big] = 1-\alpha$$

et un intervalle de confiance unilatéral à gauche de niveau $1-\alpha$ pour le paramètre θ est un intervalle (aléatoire) $\big[\widehat{a}(\boldsymbol{Y}), +\infty\big]$ tel que

$$\mathbb{P}\Big[\theta \in \big[\widehat{a}(\boldsymbol{Y}), +\infty\big)\Big] = 1-\alpha$$

Intervalle de Confiance pour un échantillon Gaussien

Soit $\{y_1, \cdots, y_n\}$ un échantillon i.i.d. de loi $\mathcal{N}(\mu, \sigma_0^2)$, où σ_0^2 est supposé connu.

$\widehat{\mu}(\boldsymbol{Y}) = \overline{Y}$, alores $\widehat{\mu}(\boldsymbol{Y}) = \mathcal{N}\left(\mu, \dfrac{\sigma_0^2}{n}\right)$.

Posons $Z = \dfrac{\widehat{\mu}(\boldsymbol{Y}) - \mu}{\sigma_0/\sqrt{n}}$, $Z \sim \mathcal{N}(0, 1)$.

L'intervalle de confiance bilatéral pour μ de niveau $1 - \alpha$ est

$$\left[\widehat{\mu}(\boldsymbol{Y}) - u_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \widehat{\mu}(\boldsymbol{Y}) + u_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right]$$

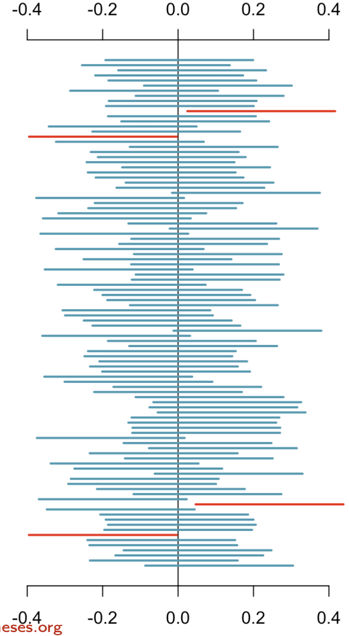
où $u_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$.

Intervalle de condiance de seuil α ?

Échantillon $\mathcal{N}(0,1)$ de taille n,

$$IC = \left[\widehat{\mu}(\boldsymbol{Y}) \pm u_{\alpha/2}\frac{1}{\sqrt{n}}\right]$$

```
alpha = .05
set.seed(1)
n=100
IC = matrix(NA,100,2)
for(s in 1:100){
  x = rnorm(100,0,1)
  m = mean(x)
  IC[s,1] = m-qnorm(1-alpha/2)
    *1/sqrt(n)
  IC[s,2] = m+qnorm(1-alpha/2)
    *1/sqrt(n)
}
idx=which((IC[,1]<0)&(IC
    [,2]>0))
```

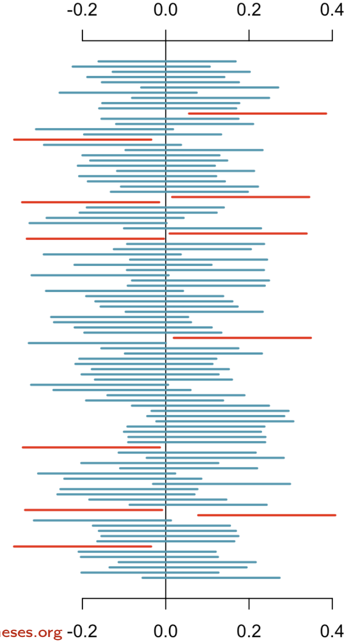


Intervalle de condiance de seuil α ?

Échantillon $\mathcal{N}(0,1)$ de taille n,

$$IC = \left[\widehat{\mu}(\boldsymbol{Y}) \pm u_{\alpha/2}\frac{1}{\sqrt{n}}\right]$$

```
alpha = .1
set.seed(1)
n=100
IC = matrix(NA,100,2)
for(s in 1:100){
  x = rnorm(100,0,1)
  m = mean(x)
  IC[s,1] = m-qnorm(1-alpha/2)*1/sqrt(n)
  IC[s,2] = m+qnorm(1-alpha/2)*1/sqrt(n)
}
idx=which((IC[,1]<0)&(IC[,2]>0))
```

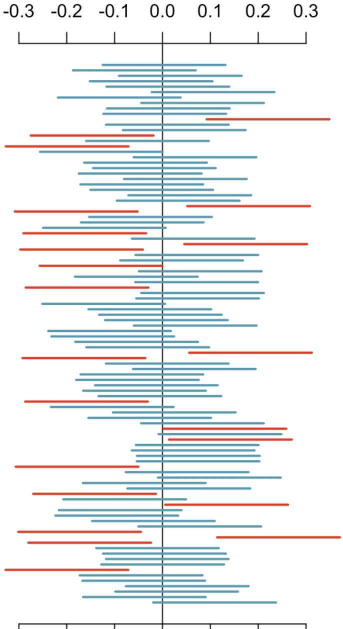


Intervalle de condiance de seuil α ?

Échantillon $\mathcal{N}(0,1)$ de taille n,

$$IC = \left[\widehat{\mu}(\boldsymbol{Y}) \pm u_{\alpha/2}\frac{1}{\sqrt{n}}\right]$$

```
alpha = .2
set.seed(1)
n=100
IC = matrix(NA,100,2)
for(s in 1:100){
  x = rnorm(100,0,1)
  m = mean(x)
  IC[s,1] = m-qnorm(1-alpha/2)
    *1/sqrt(n)
  IC[s,2] = m+qnorm(1-alpha/2)
    *1/sqrt(n)
}
idx=which((IC[,1]<0)&(IC
    [,2]>0))
```



Intervalle de Confiance dans le cas Gaussien

Exercice 1 On a observé les 5 notes suivant, supposées suivre une loi $\mathcal{N}(\mu, 0.04)$. Donner un intervalle de confiance à 90% pour μ.

```
c(3.4, 3.7, 3.9, 3.6, 3.75)
```

(3.5,3.8)

Intervalle de Confiance pour une proportion

Soit X le nombre de cas favorable, avec n tirages de variables de Bernoulli de probabilité p. Alors $X \sim \mathcal{B}(n, p)$,

$$F(k; p) = \mathbb{P}[X \leq k] = \sum_{i=1}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$

$$\overline{F}(k; p) = \mathbb{P}[X \geq k] = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

$$\frac{\partial \overline{F}(k; p)}{\partial p} = \sum_{i=k}^{n} \binom{n}{i} i p^{i-1} (1-p)^{n-i} - \sum_{i=k}^{n-1} \binom{n}{i} (n-i) p^i (1-p)^{n-i-1}$$

$$= n \left[\sum_{i=k}^{n} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i} - \sum_{i=k}^{n-1} \binom{n-1}{i} p^i (1-p)^{n-i-1} \right]$$

$$= k \binom{n}{k} p^{k-1} (1-p)^{n-k} > 0$$

Intervalle de Confiance pour une proportion

On reconnaît des lois Beta,

$$\frac{\partial \overline{F}(k;p)}{\partial p} = k\binom{n}{k}p^{k-1}(1-p)^{n-k} : \text{ loi } \mathcal{B}(k, n-k+1)$$

$$\frac{\partial F(k;p)}{\partial p} = k\binom{n}{k}p^{k}(1-p)^{n-k-1} : \text{ loi } \mathcal{B}(k+1, n-k)$$

Aussi, si on écrit $\mathbb{P}[p^- \leq p \leq p^+] = 1-\alpha$,

$$\begin{cases} p^+ \text{ sera le quantile de niveau } 1-\alpha/2 \text{ de la loi Beta } \mathcal{B}(k+1, n-k) \\ p^- \text{ sera le quantile de niveau } \alpha/2 \text{ de la loi Beta } \mathcal{B}(k, n-k+1) \end{cases}$$

Intervalle de Confiance pour une proportion

Exercice 2: avant une élection opposant deux candidats A et B, on a effectué un sondage auprès de 100 personnes : 55 personnes se prononcent en faveur du candidat A. Estimez p (la proportion d'intention de votes en faveur de A) par intervalle de confiance

$$\begin{cases} p^+ \text{ sera le quantile de niveau } 1-\alpha/2 \text{ de la loi Beta } \mathcal{B}(k+1, n-k) \\ p^- \text{ sera le quantile de niveau } \alpha/2 \text{ de la loi Beta } \mathcal{B}(k, n-k+1) \end{cases}$$

```
> qbeta(0.975, 55+1,100-55)
[1] 0.6496798
> qbeta(0.025, 55,100-55-1)
[1] 0.4573165
```

Intervalle de Confiance Asymptotique

Soit $\{y_1, \cdots, y_n\}$ un échantillon i.i.d. de loi $\mathcal{P}(\lambda)$.
$\widehat{\lambda}(\boldsymbol{Y}) = \overline{Y}$, alors $\widehat{\lambda}(\boldsymbol{Y}) \approx \mathcal{N}\left(\lambda, \dfrac{\lambda}{n}\right)$.
L'intervalle de confiance bilatéral pour de niveau $1-\alpha$ est

$$\left[\widehat{\lambda}(\boldsymbol{Y}) - u_{\alpha/2}\frac{\sqrt{\widehat{\lambda}(\boldsymbol{Y})}}{\sqrt{n}}, \widehat{\lambda}(\boldsymbol{Y}) + u_{\alpha/2}\frac{\sqrt{\widehat{\lambda}(\boldsymbol{Y})}}{\sqrt{n}}\right]$$

où $u_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$.
... mais si on veut faire les calculs proprement, ils sont un peu plus compliqué. En effet

Intervalle de Confiance Asymptotique

changer x en y

Pour un niveau $1-\alpha$, on a

$$P\left(-u_{\alpha/2} \leq \frac{\bar{X}_n - \lambda}{\sqrt{\frac{\lambda}{n}}} \leq u_{\alpha/2}\right) \simeq 1-\alpha$$

que l'on peut aussi écrire

$$P\left(\frac{[\bar{X}_n - \lambda]^2}{\frac{\lambda}{n}} \leq u_{\alpha/2}^2\right) \simeq 1-\alpha$$

ou encore

$$\mathbb{P}\left(\lambda^2 - \lambda\left(2\bar{X}_n + \frac{z_{\frac{1+\gamma}{2}}^2}{n}\right) + \bar{X}_n^2 \leq 0\right) \simeq 1-\alpha$$

on va alors résoudre cette équation de degré 2,

Intervalle de Confiance Asymptotique

$$\Delta = \left(2\overline{y} + \frac{u_{\alpha/2}}{n}\right)^2 - 4\overline{y}^2 = 4\frac{\overline{y}u_{\alpha/2}^2}{n} + \frac{u_{\alpha/2}^4}{n^2} > 0$$

donc le polynôme est négatif lorsque λ est entre les deux racines

$$P\left(\bar{X}_n + \frac{z^2_{\frac{1+\gamma}{2}}}{2n} - \sqrt{\frac{\bar{X}_n z^2_{\frac{1+\gamma}{2}}}{n} + \frac{z^4_{\frac{1+\gamma}{2}}}{4n^2}} < \lambda < \bar{X}_n + \frac{z^2_{\frac{1+\gamma}{2}}}{2n} + \sqrt{\frac{\bar{X}_n z^2_{\frac{1+\gamma}{2}}}{n} + \frac{z^4_{\frac{1+\gamma}{2}}}{4n^2}}\right.$$

(on retrouve l'expression précédante en négligeant le terme en n^2

Intervalle de Confiance Asymptotique

Soit $\{y_1, \cdots, y_n\}$ un échantillon i.i.d. de loi $\mathcal{B}(p)$.
$\widehat{p}(\boldsymbol{Y}) = \overline{Y}$, alors $\widehat{p}(\boldsymbol{Y}) \approx \mathcal{N}\left(p, \dfrac{p(1-p)}{n}\right)$.
L'intervalle de confiance bilatéral pour p de niveau $1-\alpha$ est

$$\left[\widehat{p}(\boldsymbol{Y}) - u_{\alpha/2}\frac{\sqrt{\widehat{p}(\boldsymbol{Y})(1-\widehat{p}(\boldsymbol{Y}))}}{\sqrt{n}}, \widehat{p}(\boldsymbol{Y}) + u_{\alpha/2}\frac{\sqrt{\widehat{p}(\boldsymbol{Y})(1-\widehat{p}(\boldsymbol{Y}))}}{\sqrt{n}}\right]$$

où $u_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$.
Là encore, des calculs plus rigoureux pourraient être faits

$$\frac{\hat{p} + \frac{(1{,}96)^2}{2n}}{1 + \frac{(1{,}96)^2}{n}} \pm \frac{1}{1 + \frac{(1{,}96)^2}{n}}\sqrt{\frac{(1,96)^2}{n}\hat{p}(1-\hat{p}) + \frac{(1,96)^4}{4n^2}}$$

(on oubliera rapidement cette formule)

Intervalle de Confiance Asymptotique

Exercice 2: avant une élection opposant deux candidats A et B, on a effectué un sondage auprès de 100 personnes : 55 personnes se prononcent en faveur du candidat A. Estimez p (la proportion d'intention de votes en faveur de A) par intervalle de confiance

• approximation Gaussienne

L'intervalle de confiance bilatéral pour p de niveau $1-\alpha$ est

$$\left[\widehat{p}(\boldsymbol{Y}) - u_{\alpha/2}\frac{\sqrt{\widehat{p}(\boldsymbol{Y})(1-\widehat{p}(\boldsymbol{Y}))}}{\sqrt{n}}, \widehat{p}(\boldsymbol{Y}) + u_{\alpha/2}\frac{\sqrt{\widehat{p}(\boldsymbol{Y})(1-\widehat{p}(\boldsymbol{Y}))}}{\sqrt{n}}\right]$$

où $u_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$.

```
> alpha = 5\100
> u = qnorm(c(alpha/2,1-alpha/2))
> p = 55/100
> p + u*sqrt(p*(1-p)/100)
[1] 0.452493 0.647507
```

Intervalle de Confiance Asymptotique

```
> prop.test(x = 55, n = 100, conf.level=0.95, correct
    = FALSE)

  1-sample proportions test without continuity
    correction

data:  55 out of 100, null probability 0.5
X-squared = 1, df = 1, p-value = 0.3173
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4524460 0.6438546
sample estimates:
   p
0.55
```

Intervalle de Confiance Asymptotique

```
> library(Hmisc)
> binconf(x=55, n=100)
 PointEst     Lower     Upper
     0.55 0.452446 0.6438546
> library(prevalence)
> propCI(x = 55, n = 100)
   x   n    p        method level     lower     upper
1 55 100 0.55 agresti.coull  0.95 0.4524288 0.6438718
2 55 100 0.55         exact  0.95 0.4472802 0.6496798
3 55 100 0.55      jeffreys  0.95 0.4522290 0.6449231
4 55 100 0.55          wald  0.95 0.4524930 0.6475070
5 55 100 0.55        wilson  0.95 0.4524460 0.6438546
```

Intervalle de Confiance avec 2 échantillons

$$\mathbb{P}\left(-1,96 \leq \frac{\left(\hat{P}_1 - \hat{P}_2\right) - (p_1 - p_2)}{\sqrt{\frac{\hat{P}_1\left(1-\hat{P}_1\right)}{n_1} + \frac{\hat{P}_2\left(1-\hat{P}_2\right)}{n_2}}} \leq 1,96\right) \simeq 0,95 \ ,$$

On peut alors montrer que

$$IC_{0,95}\left(p_1 - p_2\right) = \left(\widehat{p}_1 - \widehat{p}_2\right) \pm 1,96\sqrt{\frac{\widehat{p}_1\left(1-\widehat{p}_1\right)}{n_1} + \frac{\widehat{p}_2\left(1-\widehat{p}_2\right)}{n_2}}$$

Exemple

EXEMPLE 3

Dans le cadre de l'*Enquête sur les dépenses des ménages 2011*, Statistique Canada a établi que les 1 574 ménages québécois de l'échantillon dépensaient en moyenne 1 807 $ par année au restaurant avec un écart type corrigé de 556 $. Construire un intervalle de confiance au niveau de confiance de 90 % permettant d'estimer le montant annuel moyen des dépenses au restaurant pour l'ensemble des ménages du Québec.

Sources: Statistique Canada. *Tableau 203-0021*, CANSIM.
Statistique Canada. *Guide de l'utilisateur, Enquête sur les dépenses des ménages 2011*, février 2013.

(via Simard (2015))

On a observé $\{x_1, \cdots, x_n\}$, avec $n = 1574$, où x_i est la dépense de l'individu i au restorant. On sait que $\overline{x} = 1807$ et $\widehat{\sigma} = 556$.

$$\mu \in \left[\overline{x} - u_{95\%}\frac{\widehat{\sigma}}{\sqrt{n}}; \overline{x} + u_{95\%}\frac{\widehat{\sigma}}{\sqrt{n}}\right]$$

soit

$$\mu \in \left[1807 \pm 1.645\frac{556}{\sqrt{1574}}\right] = \left[1807 \pm 23\right] = \left[1784; 1830\right]$$

Exemple

EXEMPLE

Le problème suivant est inspiré des résultats d'un sondage publié dans *Le Journal de Québec* du 11 mars 2012.

Les deux solitudes s'éloignent

Il y a vraiment deux Canada en un. Le sondage Léger Marketing publié aujourd'hui montre à quel point les Québécois sont distincts des autres Canadiens.

- D'une part, les Québécois sont proportionnellement plus nombreux que les Canadiens à être d'avis que les choses vont mal au Canada (71 % contre 43 %) et à être favorables au droit à l'avortement (85 % contre 66 %).
- D'autre part, ils sont, toujours en proportion, moins nombreux que les Canadiens à se dire favorables : à l'extraction du pétrole des sables bitumineux (36 % contre 63 %) ; à la mise en valeur de la monarchie (9 % contre 36 %) ; au financement accru de l'armée canadienne (19 % contre 37 %).

Méthodologie

Ce sondage a été réalisé du 28 février au 5 mars 2012 par Léger Marketing. Les résultats reposent sur 2 509 entrevues téléphoniques : 1 001 au Québec et 1 508 dans le reste du Canada. La marge d'erreur est d'au plus 3,1 % pour l'échantillon québécois et d'au plus 2,5 % pour l'échantillon hors Québec, et cela, 19 fois sur 20.

(via Simard (2015))

Exercice: Donner un intervalle de confiance (au niveau de 95%) du pourcentage des Québécois qui sont d'avis que les choses vont mal au Canada

Exemple

71 % des 1001 Québécois interrogés sont de cet avis,

donc $n = 1001$ et $\widehat{p} = 71\%$.
$n = 1001$ et $\widehat{p} = 71\%$, l'intervalle de confiance à 95% pour p est

$$\left[\widehat{p} \pm 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right] = \left[71 \pm 1.96\sqrt{\frac{71 \times 29}{1001}}\right] = [71 \pm 2.7] \text{ en } \%$$

Note: le document mentionne $\pm 3.1\%$, qui correspond au pire écart, c'est à dire lorsque $p \sim 50\%$. En effet

$$1.96 \max_{p \in [0,1]} \left\{\sqrt{\frac{p(1-p)}{n}}\right\} = 1.96\sqrt{\frac{50 \times 50}{1001}} \sim 3.907\%$$