Statistiques pour les sciences (MAT-4681)

Arthur Charpentier

10 - Intervalle de confiance

été 2022

Intervalle de Confiance

Comme auparavant, $Y_1, \cdots, Y_n$ sont des copies indépendantes d'une variable aléatoire Y dont la densité est paramétré par un paramétre réel ($\theta \in \Theta \subset \mathbb{R}$) ou vectoriel ($\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$), i.e. $\{Y_1, \cdots, Y_n\} \sim f_\theta \in \mathcal{F}$ où $\mathcal{F}$ est la famille de lois.

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%.

Intervalle de Confiance

Intervalle de Confiance unilatéral

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

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

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

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

Intervalle de Confiance $\mathcal{N}(\mu, \sigma_0^2)$

L'idée de base dans le modèle Gaussien (puis binomial ou Poisson) est que, si

$$Z = \frac{Y - \mu}{\sigma_0} \sim \mathcal{N}(0, 1)$$

comme $\mu = Y - Z\sigma_0$ et que $-Z \in [\Phi^{-1}(a); \Phi^{-1}(1-a)]$ avec une probabilité $1 - 2a$,

$$\mu \in \left[Y + \Phi^{-1}(a)\sigma_0; Y + \Phi^{-1}(1-a)\sigma_0\right] \text{ avec probabilité } 1 - 2a$$

ou, $-Z \in [-\infty; \Phi^{-1}(1-a)]$ avec une probabilité $1 - a$,

$$\mu \in \left(-\infty; Y + \Phi^{-1}(1-a)\sigma_0\right] \text{ avec probabilité } 1 - a$$

Intervalle de Confiance $\mathcal{N}(\mu, \sigma_0^2)$

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.

$$\widehat{\mu}(\boldsymbol{Y}) = \overline{Y} \text{ et } \widehat{\mu}(\boldsymbol{Y}) \sim \mathcal{N}\left(\mu, \frac{\sigma_0^2}{n}\right)$$

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

$$\mathbb{P}\Big(Z \in \big[\Phi^{-1}(\alpha/2), \Phi^{-1}(1-\alpha/2)\big]\Big) = \mathbb{P}\big(Z \in \big[-u_{1-\alpha/2}, u_{1-\alpha/2}\big]\big) = 1-\alpha$$

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

$$\left[\widehat{\mu}(\boldsymbol{Y}) - u_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \widehat{\mu}(\boldsymbol{Y}) + u_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right]$$

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

Intervalle de Confiance $\mathcal{N}(\mu, \sigma_0^2)$

Intervalle de Confiance pour μ, $Y_i \sim \mathcal{N}(\mu, \sigma_0^2)$

Soit $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ un échantillon aléatoire tiré de variables i.i.d. de loi $\mathcal{N}(\mu, \sigma_0^2)$. L'intervalle de confiance bilatéral pour μ de niveau $1 - \alpha$ est

$$\left[\overline{y} - u_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \overline{y} + u_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right]$$

où $u_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$. Soit, au niveau $\alpha = 5\%$

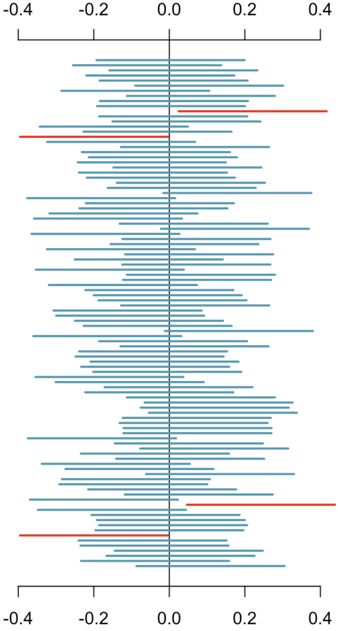
$$\left[\overline{y} - 1.96\frac{\sigma_0}{\sqrt{n}}, \overline{y} + 1.96\frac{\sigma_0}{\sqrt{n}}\right].$$

Intervalle de confiance de seuil α ?

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

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

```
> 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))
```

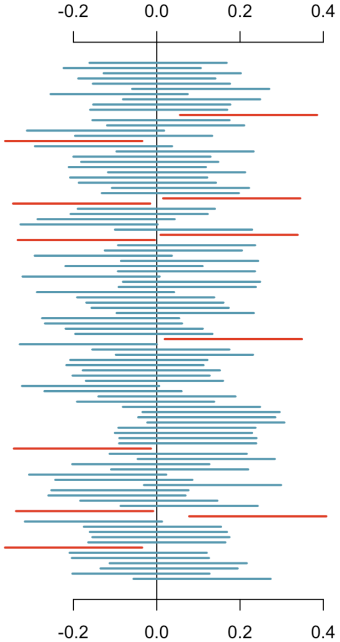


Intervalle de confiance de seuil α ?

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

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

```
> alpha = .10
> 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))
```

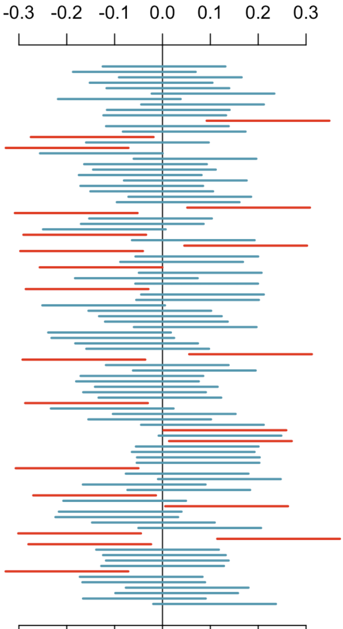


Intervalle de confiance de seuil α ?

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

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

```
> alpha = .20
> 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))
```

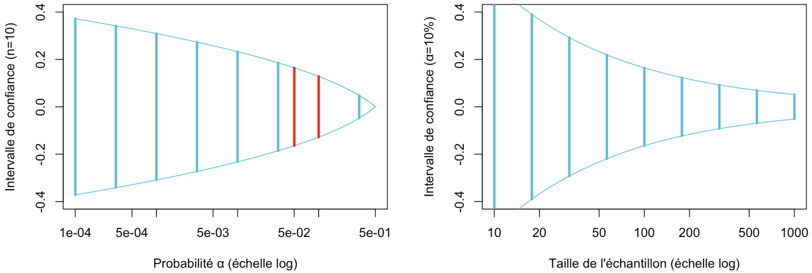


Intervalle de confiance de seuil α ?

De manière générale, l'intervalle de confiance est de la forme

$$IC_\alpha = \left[\widehat{\mu}(\boldsymbol{Y}) \pm u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \text{ de longueur } \ell = 2u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

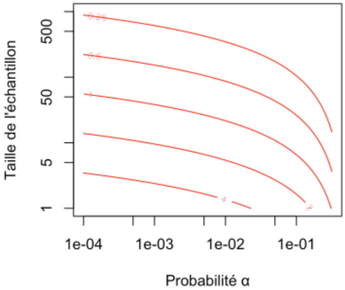
- ℓ est d'autant plus grand que α est petit
- ℓ est d'autant plus grand que n est petit

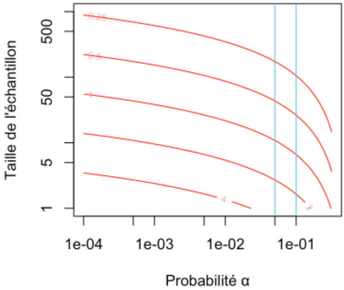


Intervalle de confiance de seuil α ?

Courbes donnant la même taille pour l'intervalle de confiance,

$$n = \frac{4u_{1-\alpha/2}^2\sigma}{\ell^2}$$





$$\ell = 2u_{10\%/2}\frac{\sigma}{\sqrt{n}} = 2u_{5\%/2}\frac{\sigma}{\sqrt{1.4198 \cdot n}}$$

Intervalle de Confiance $\mathcal{N}(\mu, \sigma_0^2)$

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.

```
> y = c(3.4, 3.7, 3.9, 3.6, 3.75)
```

L'intervalle de confiance, pour μ est

$$IC_{10\%} = \left[\overline{y} \pm 1.64 \frac{\sqrt{0.04}}{\sqrt{5}}\right] = [3.523; 3.817]$$

```
> mean(y)+c(-1.64,1.64)*sqrt(.04)/sqrt(5)
[1] 3.523314 3.816686
```

Intervalle de Confiance $\mathcal{N}(\mu, \sigma^2)$

Pour l'instant, on supposait σ_0 connue.
Pour rappel, si $Y_1, \cdots, Y_n$ est une collection de variables indépendantes $\mathcal{N}(\mu, \sigma^2)$. Si

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \text{ et } S_n{}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2.$$

Alors

$$T = \sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \sim \mathcal{S}t(n-1).$$

Comme

$$\mathbb{P}\Big(T \in \big[F_{n-1}^{-1}(\alpha/2), F_{n-1}^{-1}(1-\alpha/2)\big]\Big) = \mathbb{P}\big(Z \in [-t_{n-1,\alpha/2}, t_{n-1,\alpha/2}]\big)$$

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

$$\left[\overline{Y} - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}, \overline{Y} + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}\right]$$

Intervalle de Confiance $\mathcal{N}(\mu, \sigma^2)$

Intervalle de Confiance pour μ, $Y_i \sim \mathcal{N}(\mu, \sigma^2)$

Soit $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ un échantillon aléatoire tiré de variables i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$. L'intervalle de confiance bilatéral pour μ de niveau $1 - \alpha$ est

$$\left[\overline{y} - t_{n-1,\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}, \overline{y} + t_{n-1,\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}\right]$$

où $t_{n-1,\alpha/2} = F_{n-1}^{-1}(1 - \alpha/2)$. Si $n > 100$

$$\left[\overline{y} - u_{1-\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}, \overline{y} + u_{1-\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}\right]$$

Intervalle de Confiance $\mathcal{N}(\mu,\sigma^2)$

Exercice 1' On a observé les 5 notes suivant, supposées suivre une loi $\mathcal{N}(\mu,\sigma^2)$. Donner un intervalle de confiance à 90% pour μ.

```
> y = c(3.4, 3.7, 3.9, 3.6, 3.75)
> qt(.95,length(y)-1)
[1] 2.131847
> var(y)
[1] 0.0345
```

L'intervalle de confiance, pour μ est

$$IC_{10\%} = \left[\overline{y} \pm 2.13\frac{\sqrt{0.0345}}{\sqrt{5}}\right] = [3.493; 3.847]$$

```
> mean(y)+qt(c(.05,.95), 4)*sd(y)/sqrt(5)
[1] 3.492916 3.847084
```

Intervalle de Confiance $\mathcal{N}(\mu, \sigma^2)$

On peut aussi utiliser (comme on le verra plus tard sur les tests de moyenne)

```
> t.test(y, conf.level = 0.9)

90 percent confidence interval:
 3.492916 3.847084
sample estimates:
mean of x
     3.67
```

qui donne exactement la même chose que

```
> mean(y)+qt(c(.05,.95), 4)*sd(y)/sqrt(5)
[1] 3.492916 3.847084
```

Intervalle de Confiance, 2 échantillons $\mathcal{N}(\mu, \sigma_0^2)$

Considérons deux échantillons indépendants,

$$\begin{cases} x_1, \cdots, x_m, \;\; X_i \sim \mathcal{N}(\mu_x, \sigma_{0,x}^2), \text{ où } \sigma_{0,x} \text{ est connu} \\ y_1, \cdots, y_n, \;\; Y_i \sim \mathcal{N}(\mu_y, \sigma_{0,y}^2), \text{ où } \sigma_{0,y} \text{ est connu} \end{cases}$$

On veut un intervalle de confiance pour $\delta = \mu_x - \mu_y$.

Comme $X_i \sim \mathcal{N}(\mu_x, \sigma_{0,x}^2)$, $\overline{X} \sim \mathcal{N}\left(\mu_x, \dfrac{\sigma_{0,x}^2}{m}\right)$

Comme $Y_i \sim \mathcal{N}(\mu_y, \sigma_{0,y}^2)$, $\overline{Y} \sim \mathcal{N}\left(\mu_y, \dfrac{\sigma_{0,y}^2}{n}\right)$ avec $\overline{X} \perp\!\!\!\perp \overline{Y}$,

$$\Delta = \overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_x - \mu_y, \frac{\sigma_{0,x}^2}{m} + \frac{\sigma_{0,y}^2}{n}\right)$$

que l'on va centrer, et réduire.

Intervalle de Confiance, 2 échantillons $\mathcal{N}(\mu, \sigma_0^2)$

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\dfrac{\sigma_{0,x}^2}{m} + \dfrac{\sigma_{0,y}^2}{n}}} \sim \mathcal{N}(0,1)$$

d'où un intervalle de confiance pour $\delta = \mu_x - \mu_y$ de la forme

$$\left[(\overline{x} - \overline{y}) - u_{1-\alpha/2}\sqrt{\frac{\sigma_{0,x}^2}{m} + \frac{\sigma_{0,y}^2}{n}}, \overline{y} + u_{1-\alpha/2}\sqrt{\frac{\sigma_{0,x}^2}{m} + \frac{\sigma_{0,y}^2}{n}} \right]$$

Intervalle de Confiance, 2 échantillons $\mathcal{N}(\mu, \sigma_0^2)$

Intervalle de Confiance pour $\mu_x - \mu_y$, $\mathcal{N}(\mu_*, \sigma^2)$

Soient $\boldsymbol{x} = \{x_1, \cdots, x_m\}$ et $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ deux échantillons aléatoires indépendants tiré de variables i.i.d. de loi $\mathcal{N}(\mu_x, \sigma_{0,x}^2)$ et $\mathcal{N}(\mu_y, \sigma_{0,y}^2)$ respectivement. L'intervalle de confiance bilatéral pour $\delta = \mu_x - \mu_y$ de niveau $1 - \alpha$ est

$$\left[(\overline{x} - \overline{y}) \pm u_{1-\alpha/2}\sqrt{\frac{\sigma_{0,x}^2}{m} + \frac{\sigma_{0,y}^2}{n}}\right]$$

Intervalle de Confiance, 2 échantillons $\mathcal{N}(\mu, \sigma^2)$

Si les variances sont inconnues, mais égales

Intervalle de Confiance pour $\mu_x - \mu_y$, $\mathcal{N}(\mu_*, \sigma^2)$

Soient $\boldsymbol{x} = \{x_1, \cdots, x_m\}$ et $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ deux échantillons aléatoires indépendants tiré de variables i.i.d. de loi $\mathcal{N}(\mu_x, \sigma^2)$ et $\mathcal{N}(\mu_y, \sigma^2)$ respectivement. L'intervalle de confiance bilatéral pour $\delta = \mu_x - \mu_y$ de niveau $1 - \alpha$ est

$$\left[(\overline{x} - \overline{y}) \pm t_{m+n-2,\alpha/2}\widehat{\sigma}\sqrt{\frac{1}{m} + \frac{1}{n}}\right]$$

où $\widehat{\sigma} = \sqrt{\dfrac{(m-1)\widehat{\sigma}_x^2 + (n-1)\widehat{\sigma}_y^2}{m+n-2}}$.

Intervalle de Confiance, 2 échantillons $\mathcal{N}(\mu, \sigma^2)$

Si les variances sont inconnues, et différentes

Intervalle de Confiance pour $\mu_x - \mu_y$, $\mathcal{N}(\mu_*, \sigma^2)$

Soient $\boldsymbol{x} = \{x_1, \cdots, x_m\}$ et $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ deux échantillons aléatoires de loi $\mathcal{N}(\mu_x, \sigma_x^2)$ et $\mathcal{N}(\mu_y, \sigma_y^2)$ respectivement. L'intervalle de confiance bilatéral pour $\delta = \mu_x - \mu_y$ de niveau $1 - \alpha$ est

$$\left[(\overline{x} - \overline{y}) \pm t_{\nu,\alpha/2} \sqrt{\frac{\widehat{\sigma}_x^2}{m} + \frac{\widehat{\sigma}_y^2}{n}} \right]$$

où $\nu = \dfrac{\left(\dfrac{\widehat{\sigma_x}^2}{m} + \dfrac{\widehat{\sigma_y}^2}{n} \right)^2}{\dfrac{1}{m-1}\left(\dfrac{\widehat{\sigma_x}^2}{m} \right)^2 + \dfrac{1}{n-1}\left(\dfrac{\widehat{\sigma_y}^2}{n} \right)^2}$

Intervalle de Confiance, 2 échantillons $\mathcal{N}(\mu, \sigma^2)$

```
> x = Davis$height[Davis$sex == "F"]
> y = Davis$height[Davis$sex == "M"]
> m = length(x)
> n = length(y)
> sx2 = var(x)
> sy2 = var(y)
```

On peut tenter d'avoir un intervalle de confiance pour Δ, différence entre la taille moyenne (en cm) des hommes et des femmes. Et montrer que

$$\mathbb{P}\Big(\Delta \in [11.58; 15.01]\Big) = 95\%.$$

Intervalle de Confiance, 2 échantillons $\mathcal{N}(\mu, \sigma^2)$

```
> x = Davis$height[Davis$sex == "F"]
> y = Davis$height[Davis$sex == "M"]
> t.test(y,x)

t = 15.28, df = 174.29, p-value < 2.2e-16
alternative hypothesis: true difference in means is
    not equal to 0
95 percent confidence interval:
 11.57949 15.01467
sample estimates:
mean of x mean of y
 178.0114  164.7143
```

```
> (nu = (sx2/m+sy2/n)^2/((sx2/m)^2/(m-1)+(sy2/n)^2/(n
    -1)))
[1] 174.2935
> (mean(y)-mean(x)) + qt(c(.025,.975),df = nu) * sqrt(
    sx2/m+sy2/n)
[1] 11.57949 15.01467
```

Intervalle de Confiance pour la moyenne μ

Pour résumer (rapidement)

- si $X \sim \mathcal{N}(\mu, \sigma_0^2)$ avec σ_0^2 connue,

  comme $\sqrt{n}\dfrac{\overline{X} - \mu}{\sigma} \sim \mathcal{N}(0,1)$, $\mu \in \left[\overline{x} \pm u_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right]$
- si $X \sim \mathcal{N}(\mu, \sigma^2)$ avec σ^2 inconnue,

  comme $\sqrt{n}\dfrac{\overline{X} - \mu}{s} \sim \mathcal{S}td(n-1)$, $\mu \in \left[\overline{x} \pm t_{n-1,1-\alpha/2}\dfrac{s}{\sqrt{n}}\right]$

mais on a aussi un intervalle de confiance approché (asymptotique)

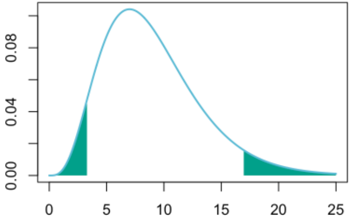
- si n est grand, et si s estime (correctement) $\text{Var}(X)$,

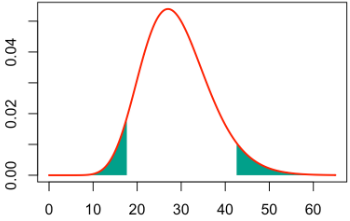
  comme $\sqrt{n}\dfrac{\overline{X} - \mu}{s} \approx \mathcal{N}(0,1)$, $\mu \in \left[\overline{x} \pm u_{1-\alpha/2}\dfrac{s}{\sqrt{n}}\right]$

Intervalle de Confiance pour la variance

Pour rappel, si $X_1, \cdots, X_n$ sont indépendantes, de loi $\mathcal{N}(\mu, \sigma^2)$,

$$\text{si } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \text{ alors } \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$





$$\mathbb{P}\left(F_{n-1}^{-1}(\alpha/2) \leq \frac{(n-1)S^2}{\sigma^2} \leq F_{n-1}^{-1}(1-\alpha/2)\right) = 1-\alpha$$

où F_ν désigne la fonction de répartition de la loi $\chi^2(\nu)$.

Intervalle de Confiance pour la variance

Intervalle de Confiance pour σ^2, $\mathcal{N}(\mu,\sigma^2)$

Soit $\boldsymbol{x} = \{x_1,\cdots,x_n\}$ de loi $\mathcal{N}(\mu,\sigma^2)$. L'intervalle de confiance bilatéral pour σ^2 de niveau $1-\alpha$ est

$$\left[\frac{(n-1)\widehat{\sigma}^2}{F_{n-1}^{-1}(1-\alpha/2)};\frac{(n-1)\widehat{\sigma}^2}{F_{n-1}^{-1}(\alpha/2)}\right]$$

Intervalle de Confiance pour σ, $\mathcal{N}(\mu,\sigma^2)$

Soit $\boldsymbol{x} = \{x_1,\cdots,x_n\}$ de loi $\mathcal{N}(\mu,\sigma^2)$. L'intervalle de confiance bilatéral pour σ^2 de niveau $1-\alpha$ est

$$\left[\sqrt{\frac{(n-1)}{F_{n-1}^{-1}(1-\alpha/2)}}\cdot\widehat{\sigma};\sqrt{\frac{(n-1)}{F_{n-1}^{-1}(\alpha/2)}}\cdot\widehat{\sigma}\right]$$

Intervalle de Confiance pour un rapport de variances ★★★

Intervalle de Confiance pour σ_x^2/σ_y^2, $\mathcal{N}(\mu,\sigma^2)$

Soient $\boldsymbol{x}=\{x_1,\cdots,x_m\}$ et $\boldsymbol{y}=\{y_1,\cdots,y_n\}$ deux échantillons aléatoires de loi $\mathcal{N}(\mu_x,\sigma_x^2)$ et $\mathcal{N}(\mu_y,\sigma_y^2)$ respectivement. L'intervalle de confiance bilatéral pour $r=\sigma_x^2/\sigma_y^2$ de niveau $1-\alpha$ est

$$\left[F_{n-1,m-1}^{-1}(\alpha/2)\cdot\frac{\widehat{\sigma}_x^2}{\widehat{\sigma}_y^2};F_{n-1,m-1}^{-1}(1-\alpha/2)\cdot\frac{\widehat{\sigma}_x^2}{\widehat{\sigma}_y^2}\right]$$

où $F_{a,b}$ est la fonction de répartition de la loi de Fisher $\mathcal{F}(a,b)$.

Intervalle de Confiance pour un rapport de variances ★★★

```
> x = c(9.1,12.5,10.2,9.5,7.3,5.6,10.1,13.0,12.8,9.0,
  7.9,7.7)
> y = c(11.6,21.0,20.9,7.1,15.9,15.6,17.9,10.3,16.5,
  17.4,15.7,17.1,13.5,12.7,19.0)
> var(x)/var(y)
[1] 0.359796
> qf(c(.025,.975),length(y)-1,length(x)-1)
[1] 0.3231446 3.3588102
> qf(c(.025,.975),length(y)-1,length(x)-1)*var(x)/var(
  y)
[1] 0.1162661 1.2084866
```

Aussi, l'estimation de $\text{Var}[X]/\text{Var}[Y]$ est 0.36 et

$$\mathbb{P}\left(\frac{\text{Var}[X]}{\text{Var}[Y]} \in [0.116; 1.208]\right) = 95\%$$

Intervalle de Confiance pour un rapport de variances ★★★

```
> x = Davis$height[Davis$sex == "F"]
> y = Davis$height[Davis$sex == "M"]
> var.test(x,y)

  F test to compare two variances

data:  x and y
F = 0.77203, num df = 111, denom df = 87,
p-value = 0.1979
alternative hypothesis: true ratio of variances is not
     equal to 1
95 percent confidence interval:
 0.5153698 1.1452526
```

```
> var(x)/var(y)
[1] 0.7720278
> qf(c(.025,.975),n-1,m-1)*var(x)/var(y)
[1] 0.5153698 1.1452526
```

Intervalle de Confiance pour une proportion

Soient $Y_1, Y_2, \cdots, Y_n$ des variables $\mathcal{B}(p)$ indépendantes.

Soit $S = Y_1 + Y_2 + \cdots + Y_n = \sum_{i=1}^{n} Y_i \sim \mathcal{B}(n,p)$.

La méthode de Clopper & Pearson consiste à chercher p^- et p^+, tels que $\mathbb{P}[p^- \leq p \leq p^+] = 1 - \alpha$, quel que soit n.

Intervalle de Confiance pour une proportion

Soient $Y_1, Y_2, \cdots, Y_n$ des variables $\mathcal{B}(p)$ indépendantes.

Soit $S = Y_1 + Y_2 + \cdots + Y_n = \sum_{i=1}^{n} Y_i$.

Si n est suffisamment grand, on peut invoquer le théorème central limite,

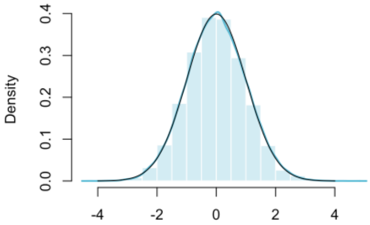
$$\frac{S - np}{\sqrt{np(1-p)}} = \sqrt{n}\frac{\overline{Y} - p}{\sqrt{p(1-p)}} \approx \mathcal{N}(0,1)$$

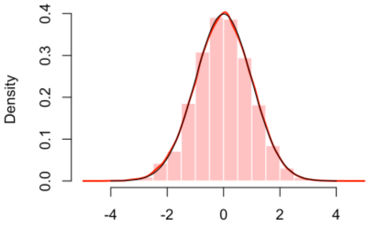
mais aussi

$$\sqrt{n}\frac{S - np}{\sqrt{S(n-S)}} = \sqrt{n}\frac{\overline{Y} - p}{\sqrt{\overline{Y}(1-\overline{Y})}} \approx \mathcal{N}(0,1)$$

Intervalle de Confiance pour une proportion

```
> n = 256
> p = .4
> S1 = S2 = rep(NA, 10000)
> for(i in 1:10000){
+     y = sample(0:1, size = n, prob = c(1-p,p),
    replace = TRUE)
+     S1[i] = sqrt(n)*(mean(y)-p)/(sqrt(p*(1-p)))
+     S2[i] = sqrt(n)*(mean(y)-p)/(sqrt(mean(y)*(1-mean
    (y))))
+  }
```





Intervalle de Confiance pour une proportion

Comme $Z = \sqrt{n}\dfrac{\overline{Y} - p}{\sqrt{\overline{Y}(1-\overline{Y})}} \approx \mathcal{N}(0,1)$, on peut obtenir facilement un intervalle de confiance pour p

Intervalle de Confiance pour p, $\mathcal{B}(p)$ - Wald

Soit $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ un échantillon, réalisation de variables indépendantes X_i de loi $\mathcal{B}(p)$. L'intervalle de confiance bilatéral pour p de niveau $1-\alpha$ est

$$\hat{p} \pm u_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} = \overline{x}$$

On parler d'approche de Wald.

Valide si $n \geq 50$, $n\hat{p} \geq 10$ et $n(1-\hat{p}) \geq 10$.

Intervalle de Confiance pour une loi binomiale

Exercice 1: 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

```
> 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 pour une loi binomiale

```
> 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
```

Nous reviendrons sur ces différentes approches dans la partie 12 sur les proportions.

Intervalle de Confiance pour une loi géométrique ★★★

Intervalle de Confiance pour p, $\mathcal{G}(p)$

Soit $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ un échantillon, réalisation de variables indépendantes X_i de loi $\mathcal{G}(p)$. L'intervalle de confiance bi-latéral pour p de niveau $1-\alpha$ est

$$\left[\widehat{p} \pm u_{1-\alpha/2}\widehat{p}\sqrt{\frac{1-\widehat{p}}{n}}\right] \text{ où } \widehat{p} = \frac{1}{\overline{y}}.$$

alors que celui pour p^{-1} (correspondant à l'espérance de Y) est

$$\left[\overline{y} - u_{1-\alpha/2}\sqrt{\frac{\overline{y}(\overline{y}-1)}{n}}; \overline{y} + u_{1-\alpha/2}\sqrt{\frac{\overline{y}(\overline{y}-1)}{n}}\right]$$

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

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

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

$$1.96 \max_{p \in [0,1]} \left\{\sqrt{\frac{p(1-p)}{n}}\right\} = 1.96\sqrt{\frac{50 \times 50}{1001}} \sim 3.0907\%$$

Exemple I

On dispose des données suivantes correspondant à des durées d'attente. Quel serait l'intervalle de confiance de la durée moyenne d'attente θ, à 95% ?

```
> y = c(0.76, 1.18, 0.15, 0.14, 0.44, 2.89, 1.23,
   0.54, 0.96, 0.15, 1.39, 0.76, 1.24, 4.42, 1.05,
   1.04, 1.88, 0.65, 0.34, 0.59)
```

1. En supposant les données Gaussiennes,

```
> t.test(y)

95 percent confidence interval:
 0.6130457 1.5669543
```

aussi $\mathbb{P}(\theta \in [0.613; 1.567]) = 95\%$.

Exemple II

2. On peut supposer données exponentielles, $Y_i \sim \mathcal{E}(\lambda)$, et $\theta = \lambda^{-1}$. D'après le théorème central limite

$$Z = \sqrt{n}\frac{\overline{Y} - \theta}{\sqrt{\theta^2}} = \sqrt{n}\frac{\overline{Y} - \theta}{\theta} =\sim \mathcal{N}(0,1)$$

et donc, comme auparavant,

$$\mathbb{P}\left(-u_{1-\alpha/2} \leq \sqrt{n}\frac{\overline{Y} - \theta}{\theta} \leq u_{1-\alpha/2}\right) = 1 - \alpha$$

qui s'inverse en

$$\mathbb{P}\left(\frac{\overline{Y}}{1 + u_{1-\alpha/2}/\sqrt{n}} \leq \theta \leq \frac{\overline{Y}}{1 - u_{1-\alpha/2}/\sqrt{n}}\right) = 1 - \alpha$$

Exemple III

Intervalle de Confiance pour λ^{-1}, $\mathcal{E}(\lambda)$

Soit $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ un échantillon, réalisation de variables indépendantes X_i de loi $\mathcal{E}(\lambda)$. L'intervalle de confiance bi-latéral pour λ^{-1} (correspondant à la moyenne) de niveau $1-\alpha$ est

$$\left[\frac{\overline{y}}{1+u_{1-\alpha/2}/\sqrt{n}}; \frac{\overline{y}}{1-u_{1-\alpha/2}/\sqrt{n}}\right]$$

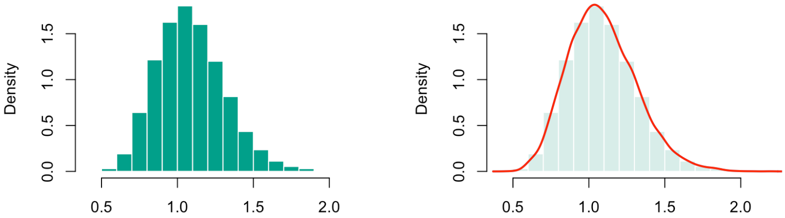
```
> mean(y)/(1+qnorm(c(.975,.025))/sqrt(length(y)))
[1] 0.7578595 1.9404039
```

aussi $\mathbb{P}(\theta \in [0.758; 1.940]) = 95\%$.

Exemple IV

On peut tenter du rééchantillonage

```
> ybar = rep(NA,10000)
> for(i in 1:10000) ybar[i] = mean(sample(y,size=
    length(y),replace=TRUE))
```



Les quantiles empiriques sont

```
> quantile(ybar,c(.025,.975))
    2.5%     97.5%
0.706000 1.577012
```

aussi $\mathbb{P}(\theta \in [0.706; 1.577]) = 95\%$.

Tolérance ★★★

Il existe une notion très proche de l'intervalle de confiance, ou de prévision, parfois utilisé dans certains contextes (en ingénierie)

Intervalle de tolérence

Soit $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ un échantillon, réalisation de variables indépendantes X_i. L'intervalle un tolérance de taux de couverture q et de niveau de confiance α doit contenir une proportion q des observations avec probabilité $1 - \alpha$.

```
> install.package("tolerance")
```

voir Young (2011) par exemple.

Note il existe aussi la notion d'intervalle de crédibilité en statistique bayésienne

Intervalle de confiance

EXEMPLE

> Dans une usine, une machine est réglée de telle sorte que le poids du produit qu'elle verse dans un contenant est distribué selon une loi normale. Un échantillon aléatoire de 10 contenants prélevé dans la production d'une journée donne un poids moyen $\overline{x} = 200{,}1$ g et un écart type corrigé $s = 2{,}5$ g. Estimer, à partir de cet échantillon, le poids moyen par contenant pour l'ensemble de la production de la journée. Le niveau de confiance est fixé à 95 %.

(via Simard (2015))

Soit X le poids du produit. On suppose que $X \sim \mathcal{N}(\mu, \sigma^2)$. On dispose d'un échantillon de taille $n = 10$, et on sait que $\overline{x} = 200.1$ g, et $s = 2.5$ g.

Comme on a seulement 10 observations, on utilise l'intervalle de confiance de Student, pour μ

$$\left[\overline{x} \pm t_{9,97.5\%} \frac{s}{\sqrt{10}}\right] = [198.31; 201.89]$$

```
> 200.1 + qt(c(.025,.975),10-1)*2.5/sqrt(10)
[1] 198.3116 201.8884
```

Intervalle de confiance

EXEMPLE

Dans une usine, une machine est réglée de telle sorte que le poids du produit qu'elle verse dans un contenant est distribué selon une loi normale. Un échantillon aléatoire de 10 contenants prélevé dans la production d'une journée donne un poids moyen $\overline{x} = 200,1$ g et un écart type corrigé $s = 2,5$ g. Estimer, à partir de cet échantillon, le poids moyen par contenant pour l'ensemble de la production de la journée. Le niveau de confiance est fixé à 95 %.

(via Simard (2015))

$$\left[\overline{x} \pm t_{9,97.5\%}\frac{s}{\sqrt{10}}\right] = [198.31; 201.89]$$

```
> 200.1 + qt(c(.025,.975),10-1)*2.5/sqrt(10)
[1] 198.3116 201.8884
```

Il y a 95% de chances que le poids moyen par contenant de la production se situe entre 198.3 g et 201.9 g.

Intervalle de confiance

EXEMPLE

Quelle taille minimale d'échantillon faudrait-il prendre pour estimer la moyenne d'âge des étudiants d'une université avec une marge d'erreur d'au plus 1,5 an et un niveau de confiance de 95 %, si des études antérieures ont donné un écart type σ de 5,7 ans pour la population ?

(via Simard (2015))

A priori, n est grand, donc l'intervalle de confiance à 95% sera

$$\left[\overline{x} \pm u_{97.5\%}\frac{s}{\sqrt{n}}\right] = \left[\overline{x} \pm 1.96\frac{s}{\sqrt{n}}\right] = [\overline{x} \pm \text{ marge d'erreur}]$$

donc

$$\text{marge d'erreur} = 1.5 = 1.96\frac{s}{\sqrt{n}} = 1.96\frac{5.7}{\sqrt{n}}$$

donc

$$\sqrt{n} = \frac{1.96 \times 5.7}{1.5} \text{ ou } n = \frac{1.96^2 \times 5.7^2}{1.5^2} = 55.5$$

Il faut un échantillon de $n = 56$ étudiants, pour obtenir une marge d'erreur d'au plus 1.5 an dans l'estimation de l'âge moyen.

Intervalle de confiance

EXEMPLE

Quelle taille minimale d'échantillon faudrait-il prendre pour estimer la moyenne d'âge des étudiants d'une université avec une marge d'erreur d'au plus 1,5 an et un niveau de confiance de 95 %, si des études antérieures ont donné un écart type σ de 5,7 ans pour la population ?

(via Simard (2015))

On notera que $n = 56$ n'est pas *si* grand, et que l'intervalle de confiance sera plutôt de la forme

$$\left[\overline{x} \pm t_{n-1,97.5\%}\frac{s}{\sqrt{n}}\right]$$

```
> qnorm(.975)
[1] 1.959964
> qt(.975,55)
[1] 2.004045
```

$$n = \frac{2^2 \times 5.7^2}{1.5^2} = 57.76$$

donc il faut un échantillon de $n = 58$ étudiants...

Intervalle de confiance

EXEMPLE

Quelle devrait être la taille de l'échantillon à prélever si l'on désire estimer le pourcentage des électeurs qui appuient le parti A avec une marge d'erreur inférieure à 2 %, au niveau de confiance de 95 % ?

(via Simard (2015))

L'intervalle de confiance à 95% est

$$\left[\overline{x} \pm u_{97.5\%}\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] = [\overline{x} \pm \text{ marge d'erreur}]$$

soit

$$\text{marge d'erreur} = 0.03 = 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq 1.96\frac{\sqrt{0.5(1-0.5)}}{\sqrt{n}}$$

donc

$$\sqrt{n} \leq \frac{1.96 \times \sqrt{0.5 \times 0.5}}{0.03} \text{ ou } n \leq \frac{1.96^2 \times 0.2 \times 0.8}{0.03^2}$$

Il faut un échantillon de $n \leq 1068$ personnes

Intervalle de confiance

> b) Quelle taille devrait avoir l'échantillon si, *a priori*, on estime à environ 20 % le pourcentage de personnes favorables au projet ?

(via Simard (2015))

L'intervalle de confiance à 95% est

$$\left[\overline{x} \pm u_{97.5\%}\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] = [\overline{x} \pm \text{ marge d'erreur}]$$

et ici $p \approx 20\%$

$$\text{marge d'erreur} = 0.03 = 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

donc

$$\sqrt{n} = \frac{1.96 \times \sqrt{0.2 \times 0.8}}{0.03} \text{ ou } n = \frac{1.96^2 \times 0.2 \times 0.8}{0.03^2}$$

Il faut un échantillon de $n = 683$ personnes