

Statistiques pour les sciences (MAT-4681)

Arthur Charpentier

10 - Intervalle de confiance

été 2022

Intervalle de Confiance

Comme auparavant, Y_1, \dots, Y_n sont des copies indépendantes d'une variable aléatoire Y dont la densité est paramétré par un paramètre réel ($\theta \in \Theta \subset \mathbb{R}$) ou vectoriel ($\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$), i.e.

$\{Y_1, \dots, Y_n\} \sim f_\theta \in \mathcal{F}$ où \mathcal{F} est la famille de lois.

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

Intervalle de Confiance

Soit \mathbf{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) $[\hat{a}(\mathbf{Y}), \hat{b}(\mathbf{Y})]$ tel que

$$\mathbb{P}[\theta \in [\hat{a}(\mathbf{Y}), \hat{b}(\mathbf{Y})]] = 1 - \alpha$$

Classiquement, α vaut 10%, 5% ou 1%.

Intervalle de Confiance

Intervalle de Confiance unilatéral

Soit \mathbf{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) $[-\infty, \hat{b}(\mathbf{Y})]$ tel que

$$\mathbb{P}[\theta \leq \hat{b}(\mathbf{Y})] = \mathbb{P}[\theta \in (-\infty, \hat{b}(\mathbf{Y})]] = 1 - \alpha$$

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

$$\mathbb{P}[\theta \geq \hat{a}(\mathbf{Y})] = \mathbb{P}[\theta \in [\hat{a}(\mathbf{Y}), +\infty)] = 1 - \alpha$$

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 $1 - 2a$,

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

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

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

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

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

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

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

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 $\mathbf{y} = \{y_1, \dots, y_n\}$ 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

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

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

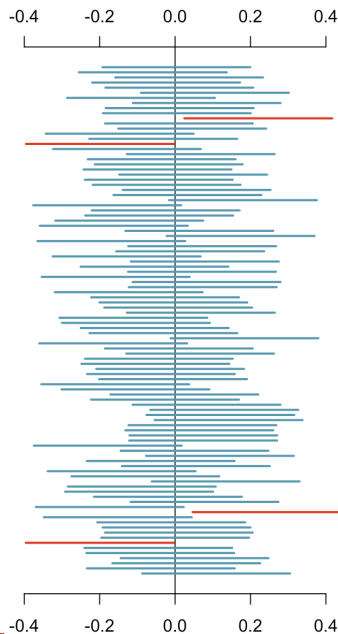
$$\left[\bar{y} - 1.96 \frac{\sigma_0}{\sqrt{n}}, \bar{y} + 1.96 \frac{\sigma_0}{\sqrt{n}} \right].$$

Intervalle de confiance de seuil α ?

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

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

```
1 > alpha = .05
2 > set.seed(1)
3 > n = 100
4 > IC = matrix(NA,100,2)
5 > for(s in 1:100){
6 +   x = rnorm(100,0,1)
7 +   m = mean(x)
8 +   IC[s,1] = m-qnorm(1-alpha
9 +             /2)*1/sqrt(n)
10 +  IC[s,2] = m+qnorm(1-alpha
11 +                    /2)*1/sqrt(n)
12 + }
13 > idx = which((IC[,1]<0)&(IC
14 +               [,2]>0))
```

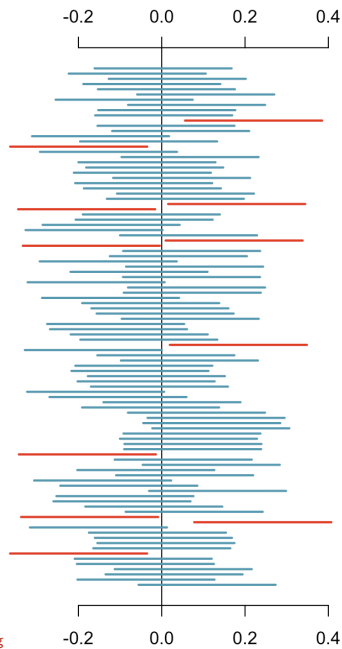


Intervalle de confiance de seuil α ?

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

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

```
1 > alpha = .10
2 > set.seed(1)
3 > n = 100
4 > IC = matrix(NA,100,2)
5 > for(s in 1:100){
6 +   x = rnorm(100,0,1)
7 +   m = mean(x)
8 +   IC[s,1] = m-qnorm(1-alpha
9 +     /2)*1/sqrt(n)
9 +   IC[s,2] = m+qnorm(1-alpha
10 +     /2)*1/sqrt(n)
10 + }
11 > idx = which((IC[,1]<0)&(IC
    [,2]>0))
```

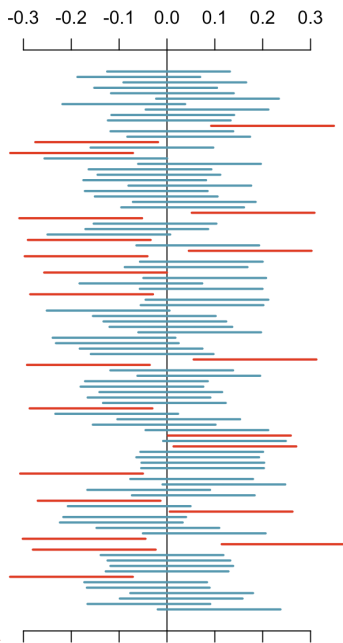


Intervalle de confiance de seuil α ?

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

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

```
1 > alpha = .20
2 > set.seed(1)
3 > n = 100
4 > IC = matrix(NA,100,2)
5 > for(s in 1:100){
6 +   x = rnorm(100,0,1)
7 +   m = mean(x)
8 +   IC[s,1] = m-qnorm(1-alpha
9 +     /2)*1/sqrt(n)
10 +   IC[s,2] = m+qnorm(1-alpha
11 +     /2)*1/sqrt(n)
12 }
13 > idx = which((IC[,1]<0)&(IC
14   [,2]>0))
```

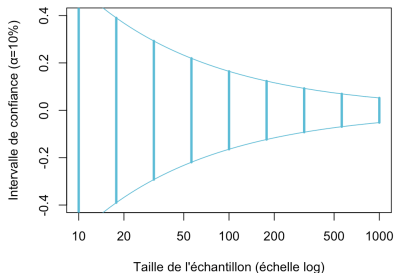
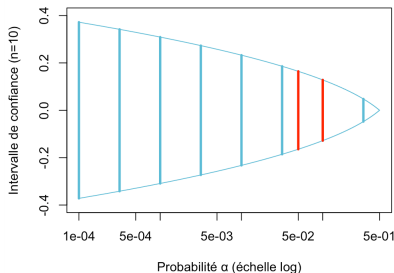


Intervalle de confiance de seuil α ?

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

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

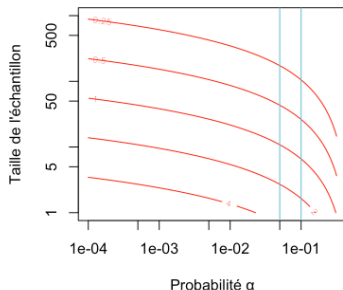
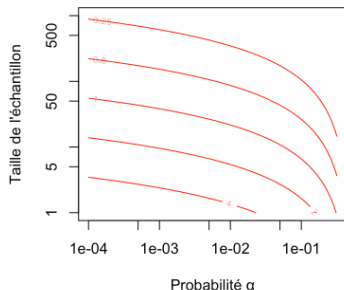
- ▶ ℓ est d'autant plus grand que α est petit
- ▶ ℓ est d'autant plus grand que n est petit



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 .

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

L'intervalle de confiance, pour μ est

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

```
1 > mean(y)+c(-1.64,1.64)*sqrt(.04)/sqrt(5)
2 [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, \dots, Y_n est une collection de variables indépendantes $\mathcal{N}(\mu, \sigma^2)$. Si

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

Alors

$$T = \sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \sim St(n-1).$$

Comme

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

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

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

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

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

Soit $\mathbf{y} = \{y_1, \dots, y_n\}$ 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

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

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

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

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 .

```
1 > y = c(3.4, 3.7, 3.9, 3.6, 3.75)
2 > qt(.95, length(y)-1)
3 [1] 2.131847
4 > var(y)
5 [1] 0.0345
```

L'intervalle de confiance, pour μ est

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

```
1 > mean(y)+qt(c(.05, .95), 4)*sd(y)/sqrt(5)
2 [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)

```
1 > t.test(y, conf.level = 0.9)
2
3 90 percent confidence interval:
4  3.492916 3.847084
5 sample estimates:
6 mean of x
7      3.67
```

qui donne exactement la même chose que

```
1 > mean(y)+qt(c(.05,.95), 4)*sd(y)/sqrt(5)
2 [1] 3.492916 3.847084
```


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

Considérons deux échantillons indépendants,

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

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

$$\text{Comme } X_i \sim \mathcal{N}(\mu_x, \sigma_{0,x}^2), \quad \bar{X} \sim \mathcal{N}\left(\mu_x, \frac{\sigma_{0,x}^2}{m}\right)$$

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

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

que l'on va centrer, et réduire.

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

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

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

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

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 $\mathbf{x} = \{x_1, \dots, x_m\}$ et $\mathbf{y} = \{y_1, \dots, y_n\}$ 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

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

RAJOUTER LES CODES R !!!

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 $\mathbf{x} = \{x_1, \dots, x_m\}$ et $\mathbf{y} = \{y_1, \dots, y_n\}$ 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

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

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

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 $\mathbf{x} = \{x_1, \dots, x_m\}$ et $\mathbf{y} = \{y_1, \dots, y_n\}$ 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

$$\left[(\bar{x} - \bar{y}) \pm t_{\nu, \alpha/2} \sqrt{\frac{\hat{\sigma}_x^2}{m} + \frac{\hat{\sigma}_y^2}{n}} \right]$$
$$\text{où } \nu = \frac{\left(\frac{\hat{\sigma}_x^2}{m} + \frac{\hat{\sigma}_y^2}{n} \right)^2}{\frac{1}{m-1} \left(\frac{\hat{\sigma}_x^2}{m} \right)^2 + \frac{1}{n-1} \left(\frac{\hat{\sigma}_y^2}{n} \right)^2}$$

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

```
1 > x = Davis$height[Davis$sex == "F"]
2 > y = Davis$height[Davis$sex == "M"]
3 > m = length(x)
4 > n = length(y)
5 > sx2 = var(x)
6 > 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

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

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

```
1 > x = Davis$height[Davis$sex == "F"]
2 > y = Davis$height[Davis$sex == "M"]
3 > t.test(y,x)
4
5 t = 15.28, df = 174.29, p-value < 2.2e-16
6 alternative hypothesis: true difference in means is
   not equal to 0
7 95 percent confidence interval:
8  11.57949 15.01467
9 sample estimates:
10 mean of x mean of y
11 178.0114 164.7143
```

```
1 > (nu = (sx2/m+sy2/n)^2/((sx2/m)^2/(m-1)+(sy2/n)^2/(n
   -1)))
2 [1] 174.2935
3 > (mean(y)-mean(x)) + qt(c(.025,.975),df = nu) * sqrt(
   sx2/m+sy2/n)
4 [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,

$$\text{comme } \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \sim \mathcal{N}(0, 1), \mu \in \left[\bar{x} \pm u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

- ▶ si $X \sim \mathcal{N}(\mu, \sigma^2)$ avec σ^2 inconnue,

$$\text{comme } \sqrt{n} \frac{\bar{X} - \mu}{s} \sim \text{Std}(n-1), \mu \in \left[\bar{x} \pm t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}} \right]$$

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

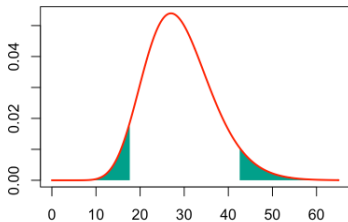
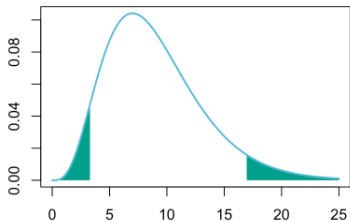
- ▶ si n est grand, et si s estime (correctement) $\text{Var}(X)$,

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

Intervalle de Confiance pour la variance

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

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



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

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

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

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

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

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

Intervalle de Confiance pour un rapport de variances ★★★

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

Soient $\mathbf{x} = \{x_1, \dots, x_m\}$ et $\mathbf{y} = \{y_1, \dots, y_n\}$ 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

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

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 ★★★

```
1 > 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)  
2 > 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)  
3 > var(x)/var(y)  
4 [1] 0.359796  
5 > qf(c(.025,.975),length(y)-1,length(x)-1)  
6 [1] 0.3231446 3.3588102  
7 > qf(c(.025,.975),length(y)-1,length(x)-1)*var(x)/var(  
    y)  
8 [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 ★★★

```
1 > x = Davis$height[Davis$sex == "F"]
2 > y = Davis$height[Davis$sex == "M"]
3 > var.test(x,y)
4
5 F test to compare two variances
6
7 data: x and y
8 F = 0.77203, num df = 111, denom df = 87,
9 p-value = 0.1979
10 alternative hypothesis: true ratio of variances is not
    equal to 1
11 95 percent confidence interval:
12 0.5153698 1.1452526
```

```
1 > var(x)/var(y)
2 [1] 0.7720278
3 > qf(c(.025,.975),n-1,m-1)*var(x)/var(y)
4 [1] 0.5153698 1.1452526
```

Intervalle de Confiance pour une proportion

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

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

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, \dots, Y_n des variables $\mathcal{B}(p)$ indépendantes.

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

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

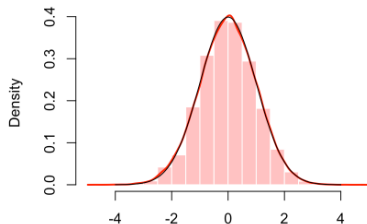
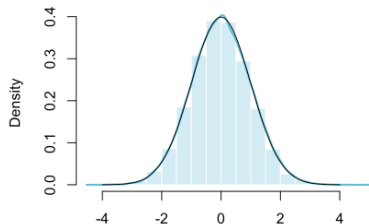
$$\frac{S - np}{\sqrt{np(1-p)}} = \sqrt{n} \frac{\bar{Y} - p}{\sqrt{p(1-p)}} \approx \mathcal{N}(0, 1)$$

mais aussi

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

Intervalle de Confiance pour une proportion

```
1 > n = 256
2 > p = .4
3 > S1 = S2 = rep(NA, 10000)
4 > for(i in 1:10000){
5 +   y = sample(0:1, size = n, prob = c(1-p,p),
6 +     replace = TRUE)
7 +   S1[i] = sqrt(n)*(mean(y)-p)/(sqrt(p*(1-p)))
8 +   S2[i] = sqrt(n)*(mean(y)-p)/(sqrt(mean(y)*(1-mean
  (y))))
9 + }
```



Intervalle de Confiance pour une proportion

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

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

Soit $\mathbf{x} = \{x_1, \dots, x_n\}$ 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

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

On parle 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

```
1 > prop.test(x = 55, n = 100, conf.level=0.95, correct
  = FALSE)
2
3 1-sample proportions test without continuity
  correction
4
5 data: 55 out of 100, null probability 0.5
6 X-squared = 1, df = 1, p-value = 0.3173
7 alternative hypothesis: true p is not equal to 0.5
8 95 percent confidence interval:
9 0.4524460 0.6438546
10 sample estimates:
11 p
12 0.55
```

Intervalle de Confiance pour une loi binomiale

```
1 > library(Hmisc)
2 > binconf(x=55, n=100)
3   PointEst      Lower      Upper
4     0.55 0.452446 0.6438546
5 > library(prevalence)
6 > propCI(x = 55, n = 100)
7   x    n    p      method level      lower      upper
8 1 55 100 0.55  agresti.coull  0.95 0.4524288 0.6438718
9 2 55 100 0.55      exact    0.95 0.4472802 0.6496798
10 3 55 100 0.55   jeffreys    0.95 0.4522290 0.6449231
11 4 55 100 0.55     wald    0.95 0.4524930 0.6475070
12 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 $\mathbf{x} = \{x_1, \dots, x_n\}$ un échantillon, réalisation de variables indépendantes X_i de loi $\mathcal{G}(p)$. L'intervalle de confiance bilatéral pour p de niveau $1 - \alpha$ est

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

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

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

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 $\{x_1, \dots, x_n\}$, avec $n = 1574$, où x_i est la dépense de l'individu i au restaurant. On sait que $\bar{x} = 1807$ et $\hat{\sigma} = 556$.

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

soit

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

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 $\hat{p} = 71\%$.

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

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

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

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

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% ?

```
1 > 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**,

```
1 > t.test(y)  
2  
3 95 percent confidence interval:  
4  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

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

et donc, comme auparavant,

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

qui s'inverse en

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

Exemple III

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

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

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

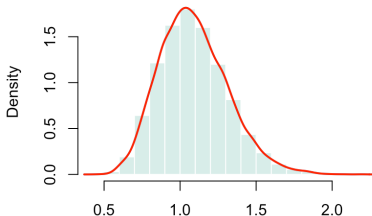
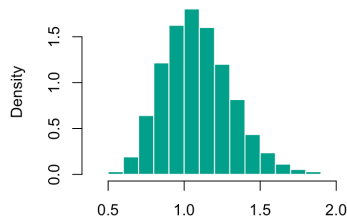
```
1 > mean(y)/(1+qnorm(c(.975,.025))/sqrt(length(y)))  
2 [1] 0.7578595 1.9404039
```

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

Exemple IV

On peut tenter du rééchantillonnage

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



Les quantiles empiriques sont

```
1 > quantile(ybar,c(.025,.975))
2   2.5%   97.5%
3 0.706000 1.577012
```

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

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

Intervalle de tolérance

Soit $\mathbf{x} = \{x_1, \dots, x_n\}$ 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 $1 - \alpha$.

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

voir [Young \(2011\)](#) par exemple.

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