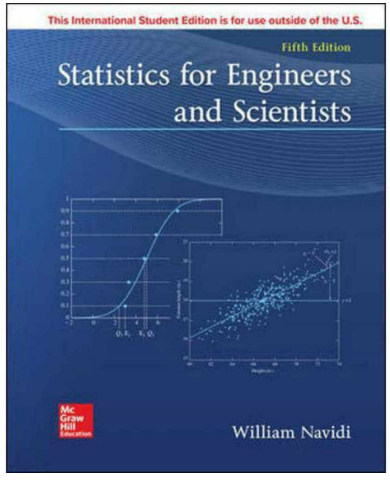
Introduction aux statistiques pour l'Ingénieur
Estimation par intervalle

Stéphane Canu
asi.insa-rouen.fr/enseignants/~scanu
scanu@insa-rouen.fr

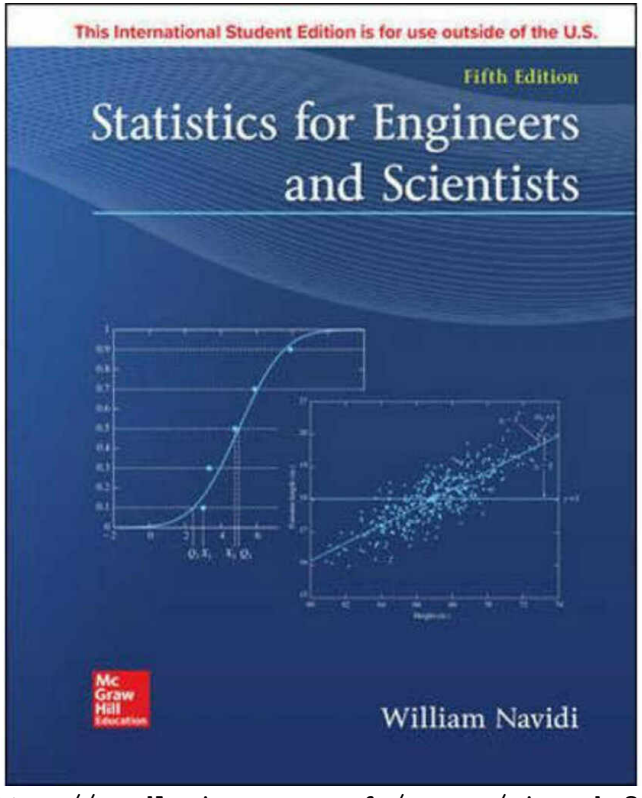


ITI 3, INSA Rouen Normandie, mars 2022

Lecture road map

1. Rappels

2. Estimation par intervalle
   - La loi du χ^2
     - Définition
     - Propriétés et approximation
   - La loi de Student
     - Définition
     - Propriétés et approximation



https://moodle.insa-rouen.fr/course/view.php?id=93

Estimateur et vraisemblance

Le modèle : X de loi parente de distribution $f_\theta(x)$
Échantillon i.i.d.

$$(X_1, \ldots, X_n)$$

Vraisemblance :

$$L(\theta, X_1, \ldots, X_n) = \prod_{i=1}^{n} f_\theta(X_i)$$

Log vraisemblance :

$$\ell(\theta, X_1, \ldots, X_n) = log\ L(\theta, X_1, \ldots, X_n) = \sum_{i=1}^{n} log\ f_\theta(X_i)$$

Estimateur du max de vraisemblance

$$\widehat{\theta}_{MV} = \arg\max_{\theta} \ell(\theta, X_1, \ldots, X_n) \quad \Leftrightarrow \quad \left.\frac{\partial \ell(\theta, X_1, \ldots, X_n)}{\partial \theta}\right|_{\theta=\widehat{\theta}_{MV}} = 0$$

Propriété des estimateurs max de vraisemblance

Si

- la probabilité ou de la densité dépend d'un paramètre θ
- le support de la probabilité ou de la densité ne dépend pas de θ
- si I existe, est inversible et quelques autres conditions plus techniques

Asymptotiquement sans biais

$$\widehat{\theta}_{MV} \underset{n\to\infty}{\longrightarrow} \theta^\star$$

Asymptotiquement efficace

$$Var(\widehat{\theta}_{MV}) \underset{n\to\infty}{\longrightarrow} I_n^{-1}(\theta^\star)$$

Asymptotiquement normal

$$\sqrt{n}(\widehat{\theta}_{MV} - \theta^\star) \underset{n\to\infty}{\longrightarrow} \mathcal{N}\left(0, I_1^{-1}(\theta^\star)\right)$$

On a pour l'information de Fisher : $I_n(\theta^\star) = nI_1(\theta^\star)$

Propriété des estimateurs max de vraisemblance

Asymptotiquement normal

$$\sqrt{n}(\widehat{\theta}_{MV} - \theta^\star) \underset{n\to\infty}{\longrightarrow} \mathcal{N}\left(0, I_1^{-1}(\theta^\star)\right)$$

Approximation normale, si

- n est grand,

la loi de $\widehat{\theta}_{MV}$ peut être approchée par

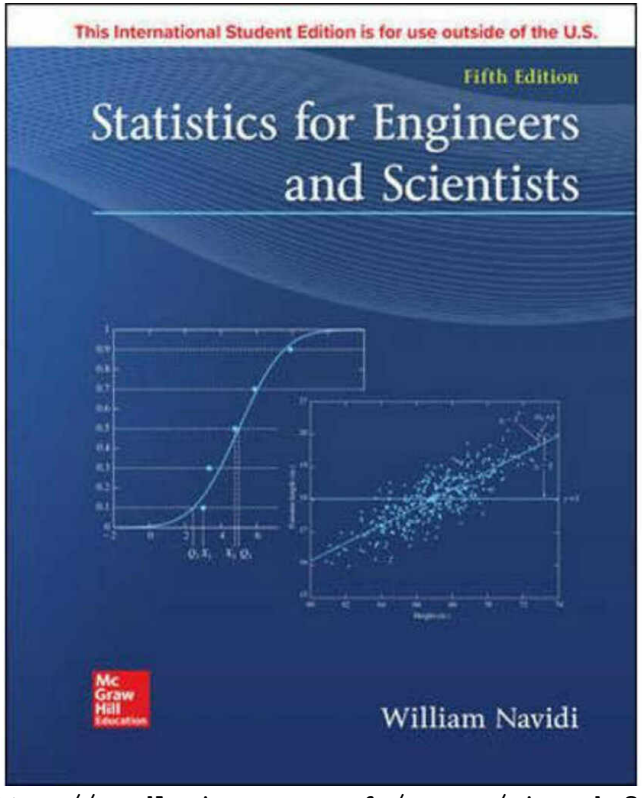
$$\mathcal{N}\left(\theta^\star, I_n^{-1}(\theta^\star)\right)$$

I_n étant l'information de Fisher

Lecture road map

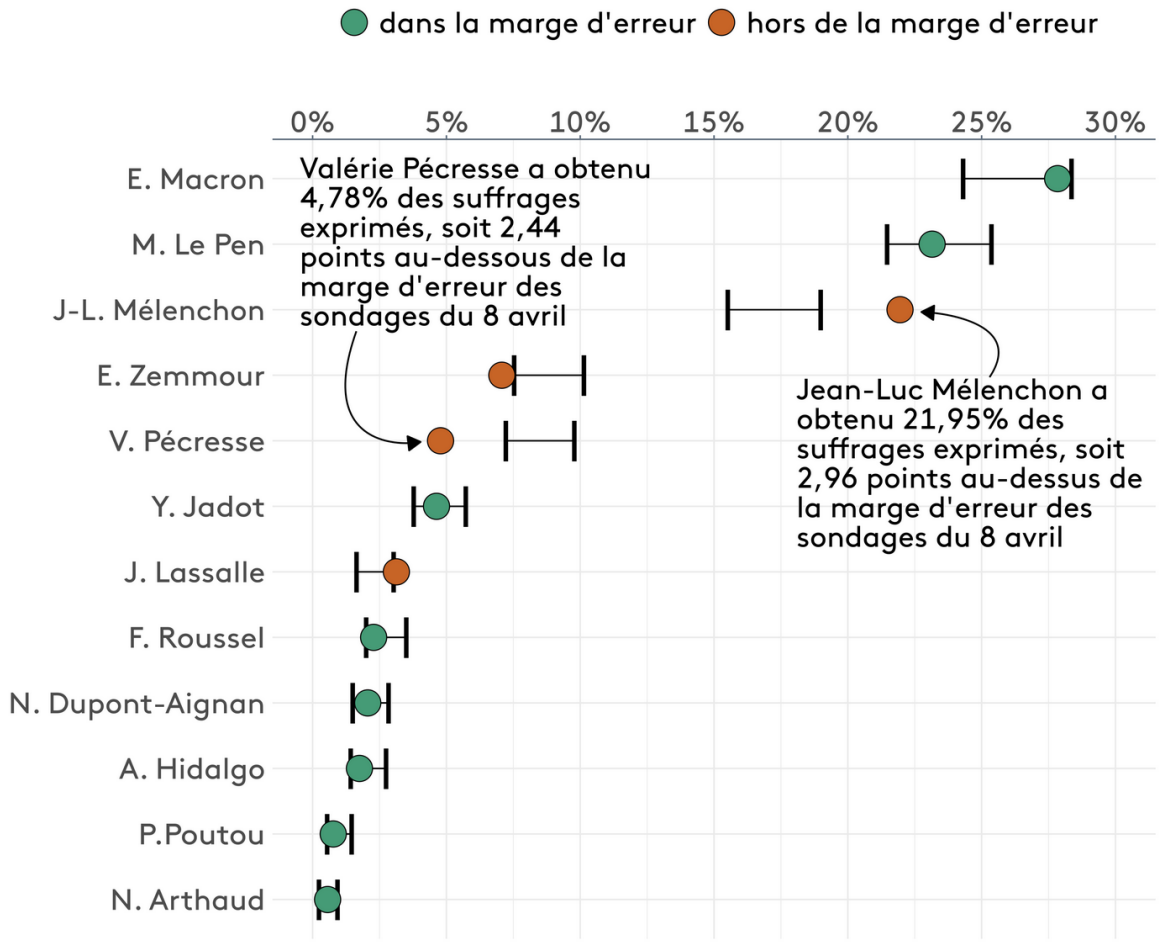
1. Rappels

2. Estimation par intervalle
   - La loi du χ^2
     - Définition
     - Propriétés et approximation
   - La loi de Student
     - Définition
     - Propriétés et approximation



https://moodle.insa-rouen.fr/course/view.php?id=93

Estimation par intervalle



Liste des sondages utilisés : Opinion Way (Les Echos), Ifop (Paris Match), Elabe (Bfm), Ipsos (Le Monde), Harris (Challenges), Ipsos (Le Parisien, Franceinfo)

Sources : NspPolls, ministère de l'Intérieur - Crédits : franceinfo

Principe de l'estimation par intervalle

Au lieu de chercher un seul estimateur (comme $\widehat{\theta}_{MV}(x_1, \ldots, X_n)$),
On en cherche deux estimateurs

$$\widehat{B}_i(X_1, \ldots, X_n) \qquad \text{et} \qquad \widehat{B}_s(X_1, \ldots, X_n)$$

de sorte que avec grande probabilité

$$\widehat{B}_i(X_1, \ldots, X_n) \leqslant \theta^\star \leqslant \widehat{B}_s(X_1, \ldots, X_n)$$

Objectif : un intervalle de confiance

- est plus « réaliste » qu'une seule valeur (*cf les sondages)*
- *aide à prendre une décision*
- *permet la détection d'observations aberrantes*

C'est ce qu'on appelle

Intervalle de confiance

Qu'est-ce qu'un « bon » intervalle de confiance ?

Intervalle de confiance
Intervalle de prédiction

$$\underset{\text{estimation}}{\widehat{\theta}} \quad \pm \quad \underset{\text{valeur critique}}{f_{\alpha/2}} \quad \times \quad \underset{\text{variabilité}}{\mathrm{Var}(\widehat{\theta})}$$

Estimation de l'espérance σ^2 connue (loi normale)

$$\widehat{\theta}_{MV}(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \sim \quad \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

On a donc

$$\mathbb{P}\left(-1,96 \quad \leqslant \quad \frac{\widehat{\theta}_{MV}-\mu}{\frac{\sigma}{\sqrt{n}}} \quad \leqslant \quad 1,96\right) = 0,95$$

soit

$$\mathbb{P}\Big(\underbrace{\widehat{\theta}_{MV} - 1,96\,\frac{\sigma}{\sqrt{n}}}_{\widehat{B}_i(X_1,\ldots,X_n)} \quad \leqslant \quad \mu \quad \leqslant \quad \underbrace{\widehat{\theta}_{MV} + 1,96\,\frac{\sigma}{\sqrt{n}}}_{\widehat{B}_s(X_1,\ldots,X_n)}\Big) = 0,95$$

Que peut-on en déduire ?

Si l'on répète de nombreuses fois l'enquête, on obtiendra des intervalles différents qui contiendront, dans 95% des cas, la valeur du paramètre μ.

Estimateurs max de vraisemblance (n grand)

Soit $\widehat{\theta}_{MV}$ l'estimateur max de vraisemblance d'un paramètre θ.
Approximation normale, si n est grand,
la loi de $\widehat{\theta}_{MV}$ peut être approchée par

$$\mathcal{N}\left(\theta^\star, I_n^{-1}(\theta^\star)\right)$$

I_n étant l'information de Fisher

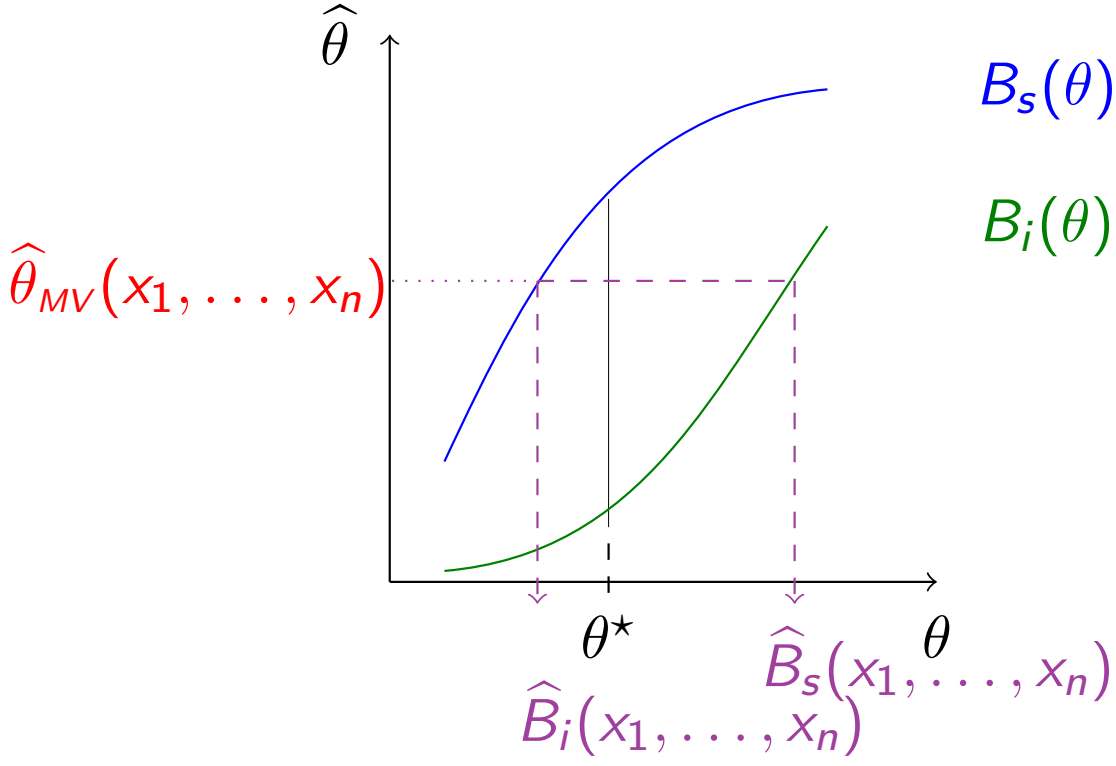
$$\mathbb{P}\Big(\underbrace{\widehat{\theta}_{MV} - 1,96\sqrt{I_n^{-1}(\theta^\star)}}_{\widehat{B}_i(X_1,\ldots,X_n)} \leqslant \theta^\star \leqslant \underbrace{\widehat{\theta}_{MV} + 1,96\sqrt{I_n^{-1}(\theta^\star)}}_{\widehat{B}_s(X_1,\ldots,X_n)}\Big) \approx 0,95$$

$$\mathbb{P}\Big(\underbrace{\widehat{\theta}_{MV} + u_{\alpha/2}\sqrt{I_n^{-1}(\theta^\star)}}_{\widehat{B}_i(X_1,\ldots,X_n)} \leqslant \theta^\star \leqslant \underbrace{\widehat{\theta}_{MV} + u_{1-\alpha/2}\sqrt{I_n^{-1}(\theta^\star)}}_{\widehat{B}_s(X_1,\ldots,X_n)}\Big) \approx 1-\alpha$$

Interprétation graphique

On veut $$\mathbb{P}\Big(\underbrace{\widehat{\theta}_{MV} - 1,96\,\frac{\sigma}{\sqrt{n}}}_{\widehat{B}_i(X_1,\ldots,X_n)} \leqslant \mu \leqslant \underbrace{\widehat{\theta}_{MV} + 1,96\,\frac{\sigma}{\sqrt{n}}}_{\widehat{B}_s(X_1,\ldots,X_n)}\Big) = 0,95$$

On a $$\mathbb{P}\Big(\underbrace{\mu - 1,96\,\frac{\sigma}{\sqrt{n}}}_{B_i(\theta)} \leqslant \widehat{\theta}_{MV} \leqslant \underbrace{\mu + 1,96\,\frac{\sigma}{\sqrt{n}}}_{B_s(\theta)}\Big) = 0,95$$



Principe général : fonction pivotale

Plan

La loi du χ^2

Soit $Y \sim \mathcal{N}(0,1)$ une variable aléatoire normale centrée réduite. Soit $Y_1, Y_2, ..., Y_n$ un échantillon de n réalisation i.i.d. de cette variable aléatoire.

Definition (La loi du χ^2)

On appelle loi du χ^2 à n degrés de libertés la loi de la variable aléatoire Z_n

$$Z_n = \sum_{i=1}^{n} Y_i^2$$

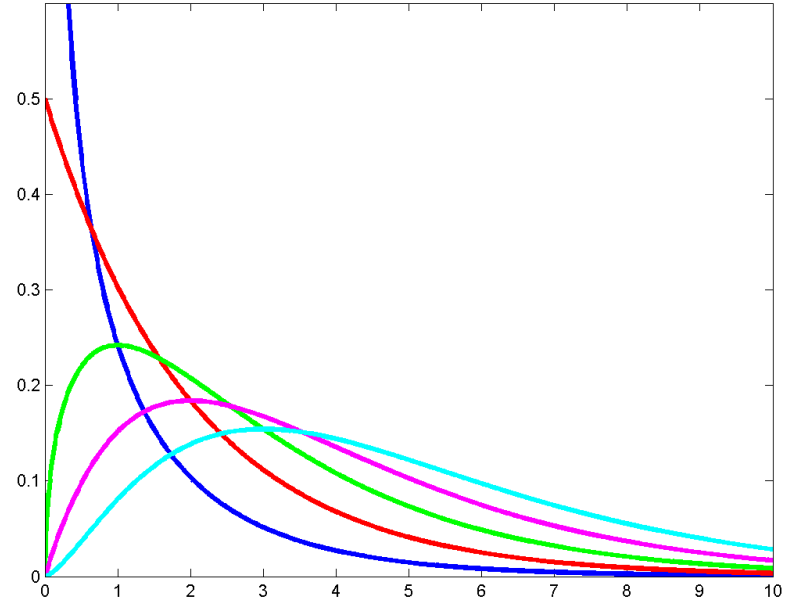


Figure: Exemples de loi du chi 2 pour 1 (bleu), 2 (rouge), 3 (vert), 4 (violet) et 5 (bleu ciel) degrés de liberté

Propriétés et approximation

$$\mathbb{E}(Z_n) = n \qquad \text{Var}(Z_n) = 2n \qquad \text{mode}(Z_n) = n - 2$$

En effet :

$$\mathbb{E}(Z_n) = \mathbb{E}(\sum_{i=1}^{n} Y_i^2) = n\mathbb{E}(Y^2) = nV(Y) = n$$

Lorsque le nombre de degrés de libertés est important on peut utiliser une approximation asymptotique de la loi du chi 2. Les plus utilisées sont les approximations de Paul Levy et de Fisher :

$$\text{Levy : } \frac{Z_n - n}{\sqrt{2n}} \to \mathcal{N}(0,1) \qquad \text{Fisher : } \sqrt{2Z_n} - \sqrt{2n-1} \to \mathcal{N}(0,1)$$

Il existe aussi une loi du chi 2 dite décentrée. C'est la somme de carrés de carrés d'une variable gaussienne non centrée. Nous ne l'utiliserons pas dans ce cours.

si $X_1, X_2, ..., X_n$ est un échantillon de n réalisation i.i.d. d'une variable aléatoire normale $\mathcal{N}(\mu, \sigma^2)$ d'espérance μ et le variance σ^2.

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2_n$$

puisque $Y_i = \frac{X_i - \mu}{\sigma}$ suit une loi normale centrée réduite.

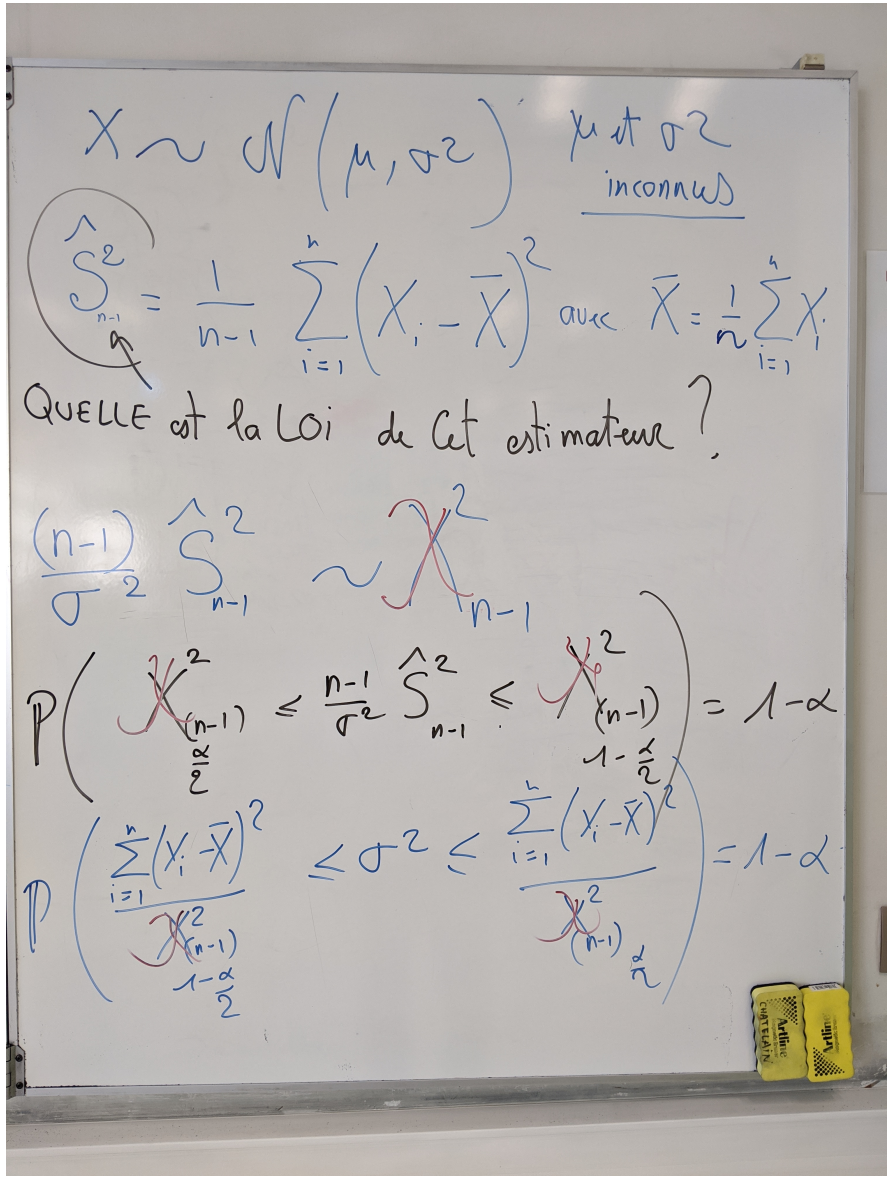
Il est moins évident de montrer que : $\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$

Lorsque l'on remplace le paramètre μ par son estimation $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ on perd un degré de liberté. En effet on a la décomposition suivante :

$$\underbrace{\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}}_{\chi^2_n} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} + \underbrace{n\frac{(\overline{X} - \mu)^2}{\sigma^2}}_{\chi^2_1}$$

Qui permet de conclure en invoquant le théorème de Cochran sur l'additivité des degrés de liberté.

L'estimation de la variance d'une loi normale



Plan

La loi de Student : définition

- Soit $N \sim \mathcal{N}(0,1)$ une variable aléatoire normale centrée réduite.

- Soit X_n la variable aléatoire distribuée suivant une loi du χ^2 à n ddl
  - C'est le cas par exemple, si $N_1, N_2, ..., N_n$ un échantillon de n réalisation i.i.d. une variable aléatoire normale centrée réduite quand $X_n = \sum_{i=1}^{n} N_i^2$

- supposons que N et X_n sont indépendantes $(i.e.\ cov(Y, X_n) = 0)$

Definition (La loi de student)

On appelle loi de student à n degrés de libertés la loi de la variable aléatoire T_n

$$T_n = \frac{N}{\sqrt{\frac{X_n}{n}}} \qquad N \sim \mathcal{N}(0,1) \qquad X_n \sim \chi^2_n$$

La loi de Student : $T_n = \frac{N}{\sqrt{\frac{X_n}{n}}}$

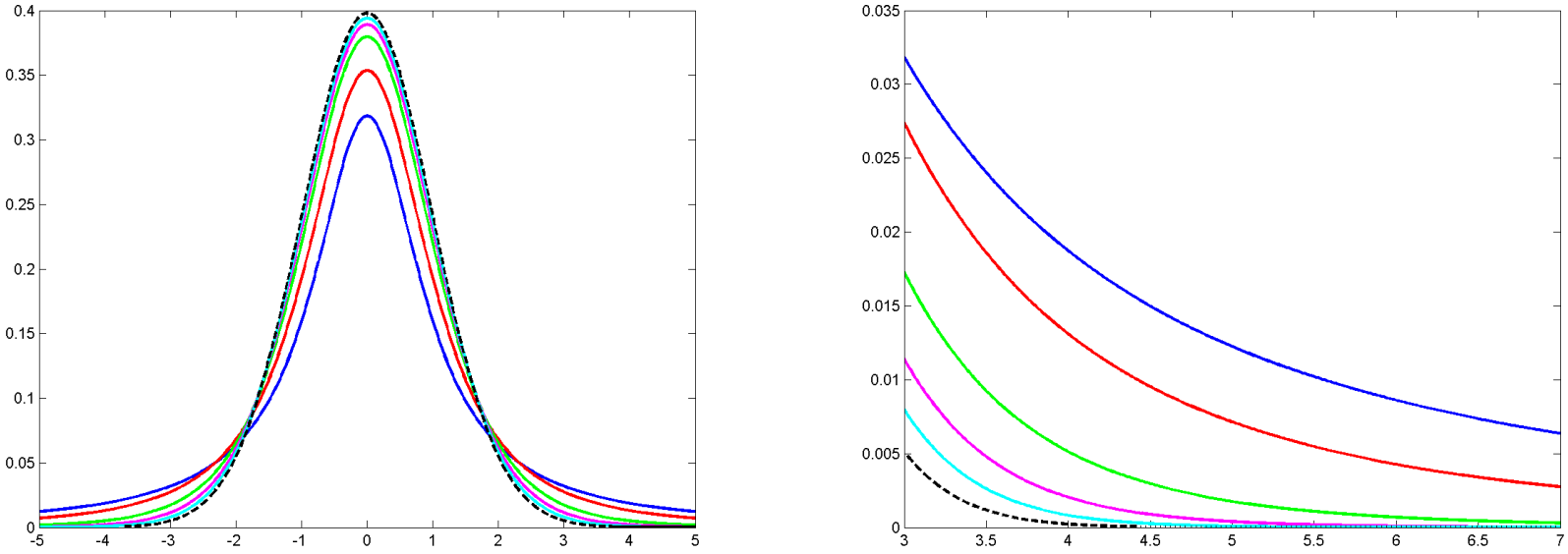


Figure: Exemples de loi de student pour 1 (bleu), 2 (rouge), 5 (vert), 10 (violet) et 20 (bleu ciel) degrés de liberté. La courbe en pointillés noir est la courbe de Gauss donnée comme référence. La figure de droite montre un zoom sur la « queue » de la distribution.

Loi de Student et loi normale

$$T_n \xrightarrow[n\to+\infty]{} \mathcal{N}(0,1)$$

Propriétés et approximation

- Publiée pour la première fois en 1908 par William Sealy Gosset qui travaillait chez Guinness (la brasserie de Dublin). Pour des raisons commerciales, il a du utiliser le pseudonyme de Student, qui restera attaché à cette loi.

- tend vers une loi normale n $>$ 30

- attention la différence est plus importante dans les « queue » de la distribution :
  - $N \sim \mathcal{N}(0,1)$: $\mathbb{P}(N > 2) = 0,023$ `p1 = 1-cdf('norm',2,0,1)`
  - $T \sim \mathcal{T}_1$: $\mathbb{P}(T > 2) = 0,148$ `p2 = 1-cdf('t',2,1)`
  - $T \sim \mathcal{T}_2$: $\mathbb{P}(T > 2) = 0,092$ `p2 = 1-cdf('t',2,2)`
  - $T \sim \mathcal{T}_{10}$: $\mathbb{P}(T > 2) = 0,038$ `p2 = 1-cdf('t',2,10)`

$$U \sim \mathcal{N}(0,\sigma^2) \qquad N = \frac{U}{\sigma} \sim \mathcal{N}(0,1) \qquad T = \frac{N}{\widehat{\sigma}} = \frac{N}{\sqrt{\frac{N_1^2+N_2^2}{2}}} \sim \mathcal{T}_2$$

Intervalle sur la moyenne d'une loi normale

Intervalle sur l'estimation dune proportion

D'après l'approximation normale de la binomiale $\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$,
et donc

$$\mathbb{P}\left(u_{\alpha/2} \leqslant \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leqslant -u_{\alpha/2}\right) \approx 1-\alpha$$

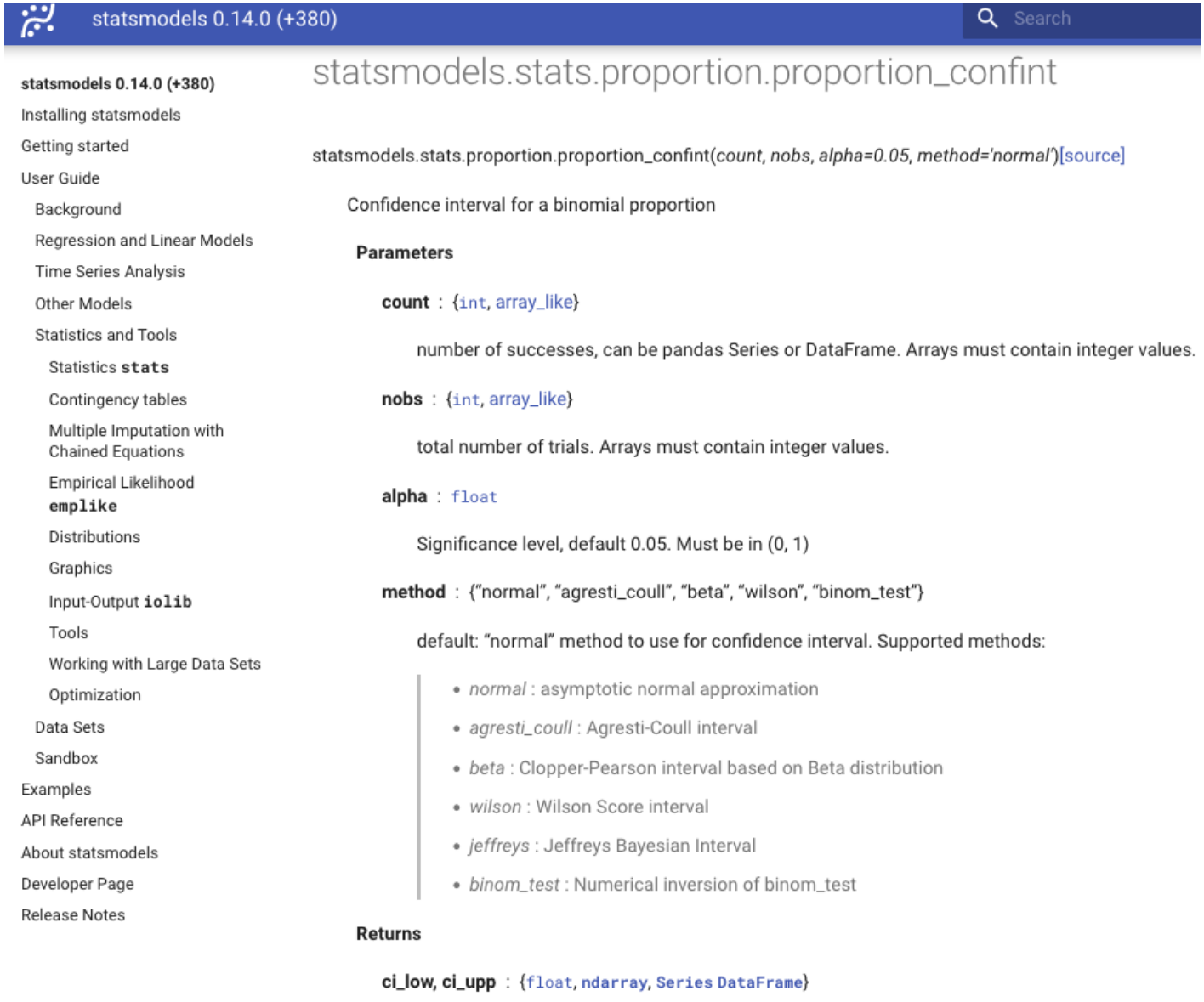
ce qui suggère d'utiliser comme fonction pivotale

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

qui donne comme bornes (Wilson score interval):

$$\frac{\hat{p} + \frac{u_{\alpha/2}^2}{2n}}{1+\frac{u_{\alpha/2}^2}{n}} \pm \frac{u_{\alpha/2}}{1+\frac{u_{\alpha/2}^2}{n}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{u_{\alpha/2}^2}{4n^2}}$$

Intervalle sur l'estimation dune proportion



statsmodels 0.14.0 (+380)

Search

statsmodels.stats.proportion.proportion_confint

statsmodels.stats.proportion.proportion_confint(*count*, *nobs*, *alpha=0.05*, *method='normal'*)[source]

Confidence interval for a binomial proportion

Parameters

count : {int, array_like}

number of successes, can be pandas Series or DataFrame. Arrays must contain integer values.

nobs : {int, array_like}

total number of trials. Arrays must contain integer values.

alpha : float

Significance level, default 0.05. Must be in (0, 1)

method : {"normal", "agresti_coull", "beta", "wilson", "binom_test"}

default: "normal" method to use for confidence interval. Supported methods:

- *normal* : asymptotic normal approximation
- *agresti_coull* : Agresti-Coull interval
- *beta* : Clopper-Pearson interval based on Beta distribution
- *wilson* : Wilson Score interval
- *jeffreys* : Jeffreys Bayesian Interval
- *binom_test* : Numerical inversion of binom_test

Returns

ci_low, ci_upp : {float, ndarray, Series DataFrame}