MAP 311 - Aléatoire

Leçon 8

2016-2017

Méthode de Monte-Carlo

- On cherche à évaluer une intégrale multi-dimensionnelle :

$$\theta = \int_{\mathbb{R}^d} g(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}$$

où $f : \mathbb{R}^d \to \mathbb{R}$ est une densité de probabilité et $g : \mathbb{R}^d \to \mathbb{R}$.

- Soit $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ une suite de v.a. à valeurs dans $\mathbb{R}^d$ indépendantes de loi de densité f. On pose :

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{X}_i)$$

- Par la loi des grands nombres, on sait que $\hat{\theta}_n \overset{n\to\infty}{\longrightarrow} \theta$ avec probabilité 1.
- On sait aussi que

$$\mathbb{E}\big((\hat{\theta}_n - \theta)^2\big) = \frac{1}{n} \mathrm{Var}\big(g(\boldsymbol{X}_1)\big)$$

ce qui montre que l'erreur $\hat{\theta}_n - \theta$ est d'ordre $1/\sqrt{n}$.

- Enfin, par le TCL, on sait que

$$\sqrt{n}\big(\hat{\theta}_n - \theta\big) \overset{n\to\infty}{\longrightarrow} \mathcal{N}\big(0, \mathrm{Var}\big(g(\boldsymbol{X}_1)\big)\big)$$

Méthode de Monte-Carlo

- Interprétation statistique :
  - $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ est un échantillon de loi de densité f,
  - $\hat{\theta}_n$ est un estimateur non-biaisé de $\theta = \mathbb{E}\big(g(\boldsymbol{X}_1)\big)$,
  - $\hat{\theta}_n$ est asymptotiquement normal, de variance asymptotique $\mathrm{Var}\big(g(\boldsymbol{X}_1)\big)$.
- On veut plus qu'un estimateur. On cherche un intervalle de confiance, c'est-à-dire un intervalle $[\hat{a}_n, \hat{b}_n]$ construit à partir de l'échantillon tel que

$$\mathbb{P}\big(\theta \in [\hat{a}_n, \hat{b}_n]\big) \geq 1 - \alpha \text{ ou } = 1 - \alpha$$

  pour un niveau de confiance $1 - \alpha$ donné.
  Ici les bornes $\hat{a}_n$ et $\hat{b}_n$ de l'intervalle ne dépendent que de l'échantillon $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$.

Intervalles de confiance

Definition

Soit $(\mathcal{X}, \mathcal{A}, \boldsymbol{\mathcal{P}})$ un modèle statistique, avec $\boldsymbol{\mathcal{P}} = \{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$.
Soit $g : \Theta \to \mathbb{R}$. Soit $\alpha \in]0, 1[$. On dit qu'un intervalle $\hat{I} = [\hat{a}_n, \hat{b}_n]$ qui s'exprime en fonction d'un n-échantillon $\boldsymbol{X}$ est un intervalle de confiance pour $g(\boldsymbol{\theta})$ de niveau $1 - \alpha$ si pour tout $\boldsymbol{\theta} \in \Theta$:

$$\mathbb{P}_{\boldsymbol{\theta}}\big(g(\boldsymbol{\theta}) \in \hat{I}\big) = 1 - \alpha\,.$$

Lorsque pour tout $\boldsymbol{\theta} \in \Theta$, on a $\mathbb{P}_{\boldsymbol{\theta}}(g(\boldsymbol{\theta}) \in \hat{I}) \geq 1 - \alpha$, on parle d'intervalle de confiance de niveau $1 - \alpha$ par excès.

L'intervalle de confiance $\hat{I}$ est donc aléatoire dans le sens où ses bornes $\hat{a}_n, \hat{b}_n$ dépendent de l'échantillon $\boldsymbol{X}$.

Exemple : L'autonomie d'une batterie de téléphone portable peut être représentée par une loi $\mathcal{E}(\lambda)$, pour un certain $\lambda > 0$.
Pour $n = 10$ batteries, on observe $\sum_{i=1}^{n} X_i = 313$h. On souhaite estimer λ.

• Estimateurs des moments.

On a $\mathbb{E}_\lambda[X_1] = 1/\lambda$, donc un estimateur possible est

$$\check{\lambda}_n^{(1)} = \frac{n}{\sum_{i=1}^{n} X_i}$$

On a $\mathbb{E}_\lambda[X_1^2] = 2/\lambda^2$, donc un estimateur possible est

$$\check{\lambda}_n^{(2)} = \left(\frac{2n}{\sum_{i=1}^{n} X_i^2}\right)^{1/2}$$

• EMV. La vraisemblance est

$$p_n\big((x_1, \ldots, x_n), \lambda\big) = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \mathbf{1}_{[0,+\infty[}(x_i)$$

$1/\overline{x}_n$ est l'unique maximum de $\lambda \mapsto p_n((x_1, \ldots, x_n), \lambda)$. Donc l'EMV est

$$\check{\lambda}_n = \frac{n}{\sum_{i=1}^{n} X_i}$$

• Etudions les propriétés de $\check{\lambda}_n = \frac{n}{\sum_{i=1}^n X_i}$.
$\sum_{i=1}^n X_i$ suit une loi $\Gamma(n, \lambda)$. Donc

$$\begin{aligned}
\mathbb{E}_\lambda(\check{\lambda}_n) &= \int_0^\infty \frac{\lambda^n}{(n-1)!} e^{-\lambda y} y^{n-1} \frac{n}{y} dy = \frac{n}{n-1}\lambda \\
\mathrm{RQM}_\lambda(\check{\lambda}_n) &= \mathbb{E}_\lambda(\check{\lambda}_n^2) - 2\lambda\mathbb{E}_\lambda(\check{\lambda}_n) + \lambda^2 \\
&= \lambda^2 \frac{n^2}{(n-1)(n-2)} - 2\lambda^2 \frac{n}{n-1} + \lambda^2 = \frac{n+2}{n-1}\frac{\lambda^2}{n-2}
\end{aligned}$$

L'EMV $\check{\lambda}_n$ est donc biaisé. On peut en déduire l'estimateur sans biais

$$\hat{\lambda}_n = \frac{n-1}{\sum_{i=1}^n X_i}$$

qui satisfait

$$\mathrm{RQM}_\lambda(\hat{\lambda}_n) = \mathrm{Var}_\lambda(\hat{\lambda}_n) = \frac{\lambda^2}{n-2}$$

En conclusion, une "bonne" estimation de λ est donnée par $\hat{\lambda}_n = 0{,}0288$.

• Intervalle de confiance exact.

On sait que $\lambda(n-1)/\hat{\lambda}_n = \lambda \sum_{i=1}^n X_i$ suit la loi $\Gamma(n,1)$ sous $\mathbb{P}_\lambda$.

On peut trouver deux nombres $a_n, b_n \in [0,+\infty]$ tels que

$$\frac{1}{(n-1)!}\int_{a_n}^{b_n} x^{n-1}e^{-x}dx = 0{,}95\,.$$

On a alors

$$\begin{aligned}\mathbb{P}_\lambda\Big(\lambda \in \Big[\frac{\hat{\lambda}_n a_n}{n-1}, \frac{\hat{\lambda}_n b_n}{n-1}\Big]\Big) &= \mathbb{P}_\lambda\Big(\frac{\lambda(n-1)}{\hat{\lambda}_n} \in [a_n, b_n]\Big) \\ &= \frac{1}{(n-1)!}\int_{a_n}^{b_n} x^{n-1}e^{-x}dx = 0{,}95\end{aligned}$$

$\hookrightarrow \left[\frac{\hat{\lambda}_n a_n}{n-1}, \frac{\hat{\lambda}_n b_n}{n-1}\right]$ est un intervalle de confiance exact au niveau 95%.

A.N. : $a_{10} = 0$, $b_{10} = 14{,}4$, $\hat{I} = [0, 0{,}046]$.

A.N. : $a_{10} = 4{,}7$, $b_{10} = +\infty$, $\hat{I} = [0{,}015, +\infty[$.

A.N. : $a_{10} = 4{,}1$, $b_{10} = 15{,}8$, $\hat{I} = [0{,}013, 0{,}051]$.

Quantile

Definition

Soit F une fonction de répartition. Pour $r \in]0,1[$, on appelle quantile (ou fractile) d'ordre r de la loi le nombre

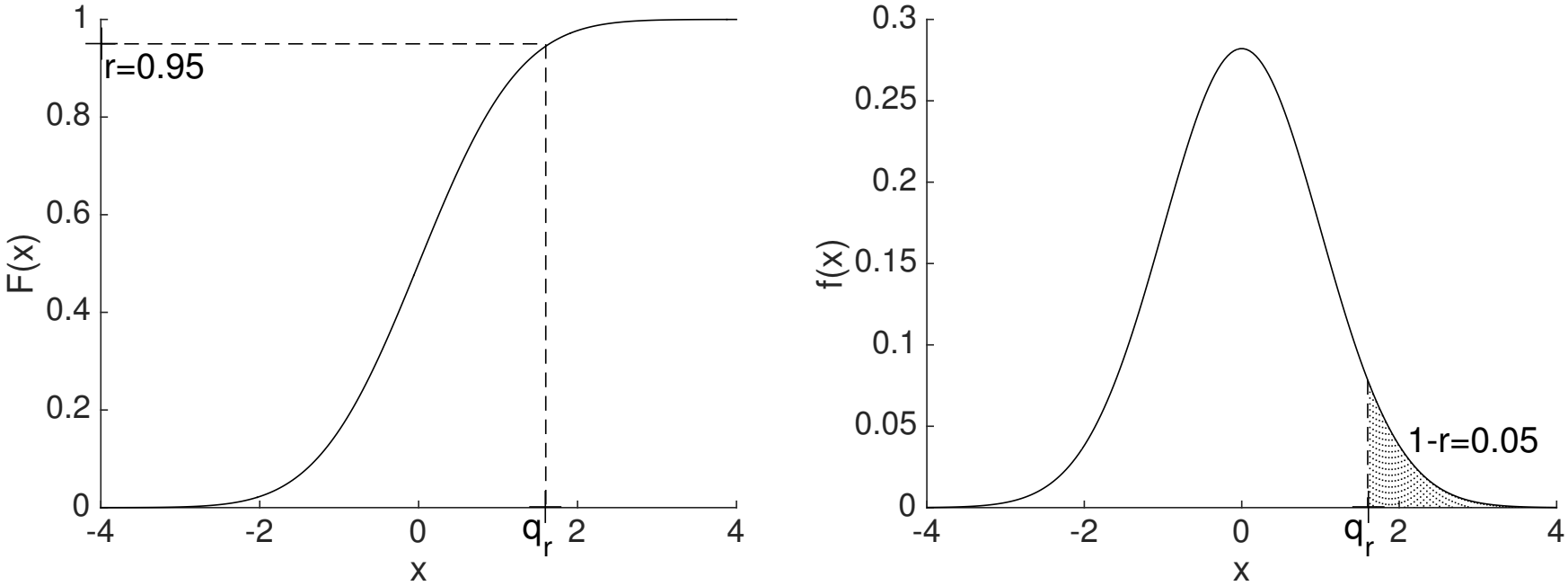
$$q_r = \inf \left\{ x \in \mathbb{R},\ F(x) \geq r \right\}.$$

Lorsque la fonction de répartition F est continue et strictement croissante, elle est inversible d'inverse F^{-1} et pour tout $r \in]0,1[$, on a $q_r = F^{-1}(r)$.

Exemples :

- La médiane est le quantile d'ordre $1/2$: Une v.a. réelle a autant de chances d'être plus petite ou plus grande que la médiane.
- Le premier quartile est le quantile d'ordre $1/4$: Une v.a. réelle a une chance sur quatre d'être plus petite que le premier quartile.

Quantile



Evaluation d'un quantile pour la loi $\mathcal{N}(0,1)$. Ici $r = 0{,}95$ et $q_{0,95} = 1{,}65$.

- Utilisation d'une table.
- Utilisation d'un logiciel.

x	0,00	0,010	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Table de valeurs de la fonction de répartition $\Phi(x)$ de la loi $\mathcal{N}(0,1)$ pour $x \geq 0$.
On a $\Phi(x) = 1 - \Phi(-x)$ pour $x \leq 0$.

Relations : Soit $Z \sim \mathcal{N}(0,1)$ de fonction de répartition Φ.

- $\Phi(x) = 1 - \Phi(-x)$ car

$$\begin{aligned}\mathbb{P}(Z \leq x) &= 1 - \mathbb{P}(Z \geq x) \\ &= 1 - \mathbb{P}(-Z \geq x) \text{ car } Z \text{ et } -Z \text{ ont même loi} \\ &= 1 - \mathbb{P}(Z \leq -x)\end{aligned}$$

- $\mathbb{P}(|Z| \leq x) = 1 - \alpha$ ssi $\mathbb{P}(Z \leq x) = 1 - \alpha/2$ car

$$\begin{aligned}\mathbb{P}(|Z| \leq x) &= \mathbb{P}(Z \leq x) - \mathbb{P}(Z \leq -x) \\ &= \mathbb{P}(Z \leq x) - \mathbb{P}(Z \geq x) \\ &= \mathbb{P}(Z \leq x) - \big(1 - \mathbb{P}(Z \leq x)\big) \\ &= 2\mathbb{P}(Z \leq x) - 1\end{aligned}$$

Intervalles exacts pour le modèle gaussien

Dans le cas des vecteurs gaussiens, nous allons obtenir des résultats non-asymptotiques (valables pout tout n).
Exemple : On souhaite connaître la durée de vie d'ampoules produites par une usine. Cette durée est de moyenne μ inconnue. On veut estimer μ.
On mesure les durées de vie (en heures) de n ampoules : données $x_1, \ldots, x_n$.
On suppose que la loi inconnue est gaussienne.
Deux cas de figures : soit σ est connu, soit σ est inconnu.

- On suppose σ connu.

On sait que la moyenne empirique $\overline{X}_n = \frac{X_1+\cdots+X_n}{n}$ est de loi $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.
Autrement dit, $\sqrt{n}\frac{\overline{X}_n-\mu}{\sigma}$ suit une loi $\mathcal{N}(0,1)$:

$$\mathbb{P}\left(\sqrt{n}\left|\frac{\overline{X}_n-\mu}{\sigma}\right| \leq a\right) = \mathbb{P}(|Z| \leq a)$$

où $Z \sim \mathcal{N}(0,1)$ et $a \geq 0$.
Comme $\mathbb{P}(Z \leq 1{,}96) = 0{,}975$, on a $\mathbb{P}(|Z| \leq 1{,}96) = 0{,}95$.
Résultat : L'intervalle de confiance de la moyenne μ au niveau de confiance 0,95 est l'intervalle aléatoire

$$\hat{I} = \left[\overline{X}_n - 1{,}96\frac{\sigma}{\sqrt{n}}\,, \overline{X}_n + 1{,}96\frac{\sigma}{\sqrt{n}}\right]$$

On a bien

$$\mathbb{P}(\mu \in \hat{I}) = \mathbb{P}\left(\sqrt{n}\left|\frac{\overline{X}_n-\mu}{\sigma}\right| \leq 1{,}96\right) = \mathbb{P}(|Z| \leq 1{,}96) = 0{,}95$$

Résultat : L'intervalle de confiance de la moyenne μ au niveau de confiance 0,95 (resp. 0,9) est l'intervalle aléatoire

$$\hat{I} = \left[\overline{X}_n - 1{,}96(\text{resp. } 1{,}645)\frac{\sigma}{\sqrt{n}}, \overline{X}_n + 1{,}96(\text{resp. } 1{,}645)\frac{\sigma}{\sqrt{n}}\right]$$

Remarques :
- A niveau de confiance fixé, l'intervalle diminue lorsque n augmente.
- A n fixé, l'intervalle augmente lorsque le niveau de confiance augmente.

• Que faire si σ^2 est inconnu ?
$\hookrightarrow$ il faut remplacer σ^2 inconnu par une estimation, et tenir compte de l'erreur d'estimation.

Résultats exacts sur le modèle gaussien

Proposition : Soit $(X_i)_{i=1,\dots,n}$ un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$, avec $n \geq 2$. Les v.a. réelles $\overline{X}_n$ et V_n définies par

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j, \qquad V_n = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$$

sont indépendantes pour tout n. De plus, pour tout n, on a :

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1)\,,$$

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sqrt{V_n}} \sim T_{n-1}\,, \quad \text{(loi de Student)}$$

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

$$(n-1)\frac{V_n}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \sim \chi^2_{n-1}\,.$$

Lois utiles :

- Loi χ^2_n : loi de la somme des carrés de n v.a. gaussiennes centrées réduites. Loi de densité :

$$f_{\chi^2_n}(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \exp\big(-\frac{x}{2}\big) x^{\frac{n}{2}-1} \mathbf{1}_{[0,\infty[}(x)$$

  En particulier $\mathbb{E}(X) = n$ et $\mathrm{Var}(X) = 2n$ si $X \sim \chi^2_n$.

- Loi T_n (Student) : loi de $Z/\sqrt{U/n}$ où Z est une variable aléatoire de loi gaussienne centrée réduite et U est une variable indépendante de Z de loi de χ^2_n. Loi de densité :

$$f_{T_n}(t) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

  En particulier $\mathbb{E}(T) = 0$ pour $n \geq 2$ et $\mathrm{Var}(T) = n/(n-2)$ pour $n \geq 3$ si $T \sim T_n$.
  Lorsque $n \to +\infty$, la loi T_n converge vers la la loi $\mathcal{N}(0,1)$.

- Estimation de la moyenne, variance σ^2 inconnue.

On sait que $\sqrt{n}\frac{\overline{X}_n-\mu}{\sqrt{V_n}} \sim T_{n-1}$.
Soit un niveau de confiance $1-\alpha$ donné.
Le quantile $t_{1-\alpha/2}$ d'ordre $1-\alpha/2$ de la loi T_{n-1} est tel que, si $Z \sim T_{n-1}$, alors $\mathbb{P}(|Z| \leq t_{1-\alpha/2}) = 1-\alpha$. Donc :

$$\mathbb{P}\left(\sqrt{n}\frac{\overline{X}_n-\mu}{\sqrt{V_n}} \in [-t_{1-\alpha/2}, t_{1-\alpha/2}]\right) = \mathbb{P}(|Z| \leq t_{1-\alpha/2}) = 1-\alpha\,.$$

En réécrivant l'événement en question, on obtient l'intervalle de confiance $\hat{I}$ pour μ au niveau $1-\alpha$:

$$\hat{I} = \left[\overline{X}_n - \frac{\sqrt{V_n}t_{1-\alpha/2}}{\sqrt{n}}, \overline{X}_n + \frac{\sqrt{V_n}t_{1-\alpha/2}}{\sqrt{n}}\right].$$

Il est tel que $\mathbb{P}(\mu \in \hat{I}) = 1-\alpha$.

Exemple : On mesure les durées de vie (en heures) de $n = 10$ ampoules. On obtient :

$$1864\,,\quad 1934\,,\quad 2033\,,\quad 1890\,,\quad 1997\,,\quad 1974\,,\quad 1837\,,\quad 1903\,,\quad 2009\,,\quad 1950\,.$$

On cherche la durée de vie moyenne μ d'une ampoule, en supposant que la loi décrivant cette durée de vie est une gaussienne. La moyenne empirique et la variance empirique non-biaisée sont :

$$\overline{X}_{10} \simeq 1939\,, \qquad V_{10} \simeq 4244\,.$$

Un intervalle de confiance pour la moyenne au niveau 95% est donc :

$$\hat{I} = \left[\overline{X}_{10} - \frac{\sqrt{V_{10}}\, t_{0,975}}{\sqrt{10}}, \overline{X}_{10} + \frac{\sqrt{V_{10}}\, t_{0,975}}{\sqrt{10}}\right],$$

où $t_{0,975}$ est tel que $\mathbb{P}(Z \leq t_{0,975}) = 0{,}975$, avec $Z \sim T_9$. En utilisant une table de la loi T_9, on trouve $t_{0,975} \simeq 2{,}26$, et on obtient ainsi l'intervalle de confiance $\hat{I} = [1892, 1986]$ pour la valeur inconnue μ au niveau 0,95.

n \ p	0,40	0,25	0,10	0,05	0,025	0,01	0,005	0,0005
1	0,324920	1,000000	3,077684	6,313752	12,70620	31,82052	63,65674	636,6192
2	0,288675	0,816497	1,885618	2,919986	4,30265	6,96456	9,92484	31,5991
3	0,276671	0,764892	1,637744	2,353363	3,18245	4,54070	5,84091	12,9240
4	0,270722	0,740697	1,533206	2,131847	2,77645	3,74695	4,60409	8,6103
5	0,267181	0,726687	1,475884	2,015048	2,57058	3,36493	4,03214	6,8688
6	0,264835	0,717558	1,439756	1,943180	2,44691	3,14267	3,70743	5,9588
7	0,263167	0,711142	1,414924	1,894579	2,36462	2,99795	3,49948	5,4079
8	0,261921	0,706387	1,396815	1,859548	2,30600	2,89646	3,35539	5,0413
9	0,260955	0,702722	1,383029	1,833113	2,26216	2,82144	3,24984	4,7809
10	0,260185	0,699812	1,372184	1,812461	2,22814	2,76377	3,16927	4,5869
11	0,259556	0,697445	1,363430	1,795885	2,20099	2,71808	3,10581	4,4370
12	0,259033	0,695483	1,356217	1,782288	2,17881	2,68100	3,05454	4,3178
13	0,258591	0,693829	1,350171	1,770933	2,16037	2,65031	3,01228	4,2208
14	0,258213	0,692417	1,345030	1,761310	2,14479	2,62449	2,97684	4,1405
15	0,257885	0,691197	1,340606	1,753050	2,13145	2,60248	2,94671	4,0728
16	0,257599	0,690132	1,336757	1,745884	2,11991	2,58349	2,92078	4,0150
17	0,257347	0,689195	1,333379	1,739607	2,10982	2,56693	2,89823	3,9651
18	0,257123	0,688364	1,330391	1,734064	2,10092	2,55238	2,87844	3,9216
19	0,256923	0,687621	1,327728	1,729133	2,09302	2,53948	2,86093	3,8834
20	0,256743	0,686954	1,325341	1,724718	2,08596	2,52798	2,84534	3,8495
21	0,256580	0,686352	1,323188	1,720743	2,07961	2,51765	2,83136	3,8193
22	0,256432	0,685805	1,321237	1,717144	2,07387	2,50832	2,81876	3,7921
23	0,256297	0,685306	1,319460	1,713872	2,06866	2,49987	2,80734	3,7676
24	0,256173	0,684850	1,317836	1,710882	2,06390	2,49216	2,79694	3,7454
25	0,256060	0,684430	1,316345	1,708141	2,05954	2,48511	2,78744	3,7251
26	0,255955	0,684043	1,314972	1,705618	2,05553	2,47863	2,77871	3,7066
27	0,255858	0,683685	1,313703	1,703288	2,05183	2,47266	2,77068	3,6896
28	0,255768	0,683353	1,312527	1,701131	2,04841	2,46714	2,76326	3,6739
29	0,255684	0,683044	1,311434	1,699127	2,04523	2,46202	2,75639	3,6594
30	0,255605	0,682756	1,310415	1,697261	2,04227	2,45726	2,75000	3,6460
$+\infty$	0,253347	0,674490	1,281552	1,644854	1,95996	2,32635	2,57583	3,2905

Table de valeurs x pour lesquelles $\mathbb{P}(T_n \geq x) = p$.

- Estimation de la variance, moyenne connue.

Lorsque la moyenne μ est connue, l'estimateur empirique de la variance :

$$V_n^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\,,$$

est sans biais, convergent, et nV_n^*/σ^2 suit la loi χ_n^2.
On peut donc affirmer que :

$$\mathbb{P}\left(\frac{nV_n^*}{\sigma^2} \in [t_{1-\alpha,l}, t_{1-\alpha,r}]\right) = 1-\alpha\,,$$

où $t_{1-\alpha,l}$ est le quantile d'ordre $\alpha/2$ de la loi χ_n^2 et $t_{1-\alpha,r}$ est le quantile d'ordre $1-\alpha/2$ de la loi χ_n^2.
En réécrivant l'événement en question, on trouve un intervalle de confiance pour la variance au niveau $1-\alpha$:

$$\hat{I} = \left[\frac{nV_n^*}{t_{1-\alpha,r}}, \frac{nV_n^*}{t_{1-\alpha,l}}\right].$$

Exemple : Imaginons que le fabricant ait fait des mesures extensives, et qu'il indique sur la boîte la durée de vie moyenne : $\mu = 1920$. On recherche la variance σ^2 inconnue de la loi de durée de vie. On a :

$$V_{10}^* = \frac{1}{10}\sum_{i=1}^{10}(X_i - 1920)^2 \simeq 4184\,.$$

Un intervalle de confiance au niveau $1-\alpha = 95\%$ de la variance à partir de l'échantillon de 10 ampoules observées est :

$$\hat{I} = \left[\frac{10V_{10}^*}{t_{0,95,r}}, \frac{10V_{10}^*}{t_{0,95,l}}\right]\,.$$

Si $Z \sim \chi^2_{10}$, alors $\mathbb{P}(Z < t_{0,95,l}) = 0{,}025$ et $\mathbb{P}(Z < t_{0,95,r}) = 0{,}975$ pour $t_{0,95,l} = 3{,}25$ et $t_{0,95,r} = 20{,}48$. On obtient alors que $\hat{I} = [2043,\ 12875]$ est un intervalle de confiance pour σ^2 au niveau 0,95.

- Estimation de la variance, moyenne inconnue.

Lorsque la moyenne μ est inconnue, l'estimateur empirique non-biaisé de la variance :

$$V_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \, .$$

est sans biais, convergent, et $(n-1)V_n/\sigma^2$ suit la loi χ^2_{n-1}.

On choisit $t_{1-\alpha,l}$ et $t_{1-\alpha,r}$ tels que, si $Z \sim \chi^2_{n-1}$, alors
$\mathbb{P}(t_{1-\alpha,l} \leq Z \leq t_{1-\alpha,r}) = 1 - \alpha$.
Un intervalle de confiance au niveau $1-\alpha$ de la variance est donc :

$$\hat{I} = \left[\frac{(n-1)V_n}{t_{1-\alpha,r}}, \frac{(n-1)V_n}{t_{1-\alpha,l}} \right] \, .$$

Exemple : L'estimateur empirique non-biaisé de la variance est :

$$V_{10} = \frac{1}{9}\sum_{i=1}^{10}(X_i - \overline{X}_{10})^2 \simeq 4244\,.$$

Un intervalle de confiance au niveau $1-\alpha = 95\%$ de la variance est :

$$\hat{I} = \left[\frac{9V_{10}}{t_{0,95,r}}, \frac{9V_{10}}{t_{0,95,l}}\right]\,.$$

Si $Z \sim \chi^2_9$, alors $\mathbb{P}(Z < t_{0,95,l}) = 0{,}025$ et $\mathbb{P}(Z < t_{0,95,r}) = 0{,}975$ pour $t_{0,95,l} = 2{,}70$ et $t_{0,95,r} = 19{,}02$. On obtient alors que $\hat{I} = [2008,\ 14147]$ est un intervalle de confiance pour σ^2 au niveau 0,95.
Remarque : l'intervalle est un peu plus grand que dans le cas où on connaît l'espérance, car il y a plus d'incertitude !

Intervalles de confiance, cas non-gaussien

Deux stratégies :

1. On trouve un estimateur dont on connaît la loi suffisamment bien pour tout n pour pouvoir trouver ses quantiles. Cf. exemple des batteries ou exemple des gaussiennes.
2. On trouve un estimateur dont la loi est trop compliquée pour n arbitraire mais qui est asymptotiquement normal. On a n assez grand pour pouvoir utiliser l'approximation normale (pratique usuelle : $n \geq 30$).

On va examiner deux exemples :
- les sondages,
- la méthode de Monte Carlo.

Sondage

A la veille d'une élection on effectue un sondage afin de déterminer la proportion $\theta \in [0, 1]$ de votes pour le candidat C.
Le sondage porte sur $n = 2500$ individus choisis au hasard dans le corps électoral.
On note $X_i = 1$ si le i^{eme} individu interrogé vote pour C, $X_i = 0$ sinon.
Les X_i sont des réalisations de v.a. i.i.d. de loi de Bernoulli de paramètre θ.
On cherche à estimer θ par :

$$\hat{\theta}_n = \frac{X_1 + \cdots + X_n}{n}.$$

Cet estimateur est sans biais, convergent, asymptotiquement normal de variance asymptotique $\theta(1 - \theta)$.

Le sondage donne 1300 intentions de votes pour C, et 1200 pour son adversaire. Donc $\hat{\theta}_n = 0{,}52$.
Comment quantifier la confiance en cette prédiction ? On a :

$$\mathbb{P}\left(\sqrt{n}|\hat{\theta}_n - \theta| \leq a\sqrt{\theta(1-\theta)}\right) \simeq \mathbb{P}(|Z| \leq a)\,,$$

où $Z \sim \mathcal{N}(0,1)$.
On a alors :

$$\mathbb{P}\left(\sqrt{n}|\hat{\theta}_n - \theta| \leq 1{,}96\sqrt{\theta(1-\theta)}\right) \simeq 0{,}95\,.$$

En réécrivant l'événement en question, on obtient :

$$\mathbb{P}\big(\theta \in \hat{I}\big) = 0{,}95 \text{ pour } \hat{I} = \left[\hat{\theta}_n - \frac{1{,}96\sqrt{\theta(1-\theta)}}{\sqrt{n}}, \hat{\theta}_n + \frac{1{,}96\sqrt{\theta(1-\theta)}}{\sqrt{n}}\right]$$

Cet intervalle dépend malencontreusement du paramètre θ inconnu !

Méthode conservative. La fonction $\sqrt{\theta(1-\theta)}$ est majorée par sa valeur maximale $1/2$, de sorte qu'en remplaçant le facteur $\sqrt{\theta(1-\theta)}$ par $1/2$ dans les seuils précédents, on ne fait qu'augmenter notre quasi-certitude. Donc :

Proposition (Intervalle de confiance pour l'estimation de θ)

Dès que n est assez grand ($n\theta$ et $n(1-\theta) \geq 10$ en pratique) :

$$\theta \in \left[\hat{\theta}_n - \frac{0{,}98 \text{ (resp. } 1{,}29)}{\sqrt{n}} \, , \, \hat{\theta}_n + \frac{0{,}98 \text{ (resp. } 1{,}29)}{\sqrt{n}}\right] ,$$

avec le niveau de confiance de 95% *(resp.* 99%*).*

Dans notre exemple, l'intervalle de confiance à 95% pour θ est $[0{,}50, 0{,}54]$. Les instituts de sondage annoncent leurs résultats ainsi (sous forme de fourchette).

Méthode asymptotique. On a la convergence en loi :

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)}} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)$$

et $\hat{\theta}_n \overset{n\to\infty}{\longrightarrow} \theta$ en probabilité. D'après le théorème de Slutsky,

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)$$

On a donc

$$\mathbb{P}\left(\sqrt{n}\left|\frac{\hat{\theta}_n - \theta}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}\right| \leq 1{,}96\right) \overset{n\to\infty}{\longrightarrow} 0{,}95$$

et

$$\left[\hat{\theta}_n - \frac{1{,}96\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}}\ ,\ \hat{\theta}_n + \frac{1{,}96\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}}\right]$$

est un intervalle asymptotique pour θ au niveau 0,95.

Les "échantillons représentatifs"

Il y a en France $p = 51\%$ de femmes.
La proposition d'hommes (resp. femmes) votant pour C est θ_{h}, resp. θ_{f}.
La proportion de la population votant pour C est $\theta = p\theta_{\mathrm{f}} + (1-p)\theta_{\mathrm{h}}$.
Deux méthodes de sondage pour estimer θ :
1) on sonde $n = 1000$ personnes prises au hasard. Le RQM de l'estimateur empirique $\hat{\theta}_n$ est $\frac{\theta(1-\theta)}{n}$.
2) on sonde $n_{\mathrm{f}} = pn = 510$ femmes et $n_{\mathrm{h}} = (1-p)n = 490$ hommes. Un nouvel estimateur de θ est

$$\check{\theta}_n = p\hat{\theta}_{\mathrm{f},pn} + (1-p)\hat{\theta}_{\mathrm{h},(1-p)n}$$

dont le RQM est

$$\begin{aligned} \mathrm{Var}(\check{\theta}_n) &= p^2 \frac{\theta_{\mathrm{f}}(1-\theta_{\mathrm{f}})}{pn} + (1-p)^2 \frac{\theta_{\mathrm{h}}(1-\theta_{\mathrm{h}})}{(1-p)n} \\ &= \frac{p\theta_{\mathrm{f}}(1-\theta_{\mathrm{f}}) + (1-p)\theta_{\mathrm{h}}(1-\theta_{\mathrm{h}})}{n} \end{aligned}$$

Or $\phi(x) = x(1-x)$ est concave, donc, par Jensen

$$n\mathrm{Var}(\hat{\theta}_n) = \phi(p\theta_{\mathrm{f}} + (1-p)\theta_{\mathrm{h}}) \geq p\phi(\theta_{\mathrm{f}}) + (1-p)\phi(\theta_{\mathrm{h}}) = n\mathrm{Var}(\check{\theta}_n)$$

Calcul d'intégrale par la méthode de Monte-Carlo

Soit $g : \mathbb{R}^d \to \mathbb{R}$ une fonction et f une densité de probabilité. On cherche à estimer $\theta = \int_{\mathbb{R}^d} g(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}$.
θ s'écrit comme :

$$\theta = \mathbb{E}(g(\boldsymbol{X})) \text{ avec } \boldsymbol{X} \text{ v.a. de densité } f \, .$$

Si $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ sont indépendantes et de densité f, alors

$$\hat{\theta}_n = \frac{1}{n} \left[g(\boldsymbol{X}_1) + \cdots + g(\boldsymbol{X}_n) \right]$$

est un estimateur de θ, sans-biais, convergent, asymptotiquement normal lorsque $\mathbb{E}(g(\boldsymbol{X})^2) < +\infty$, avec la variance asymptotique :

$$\sigma^2 = \mathbb{E}(g(\boldsymbol{X})^2) - \mathbb{E}(g(\boldsymbol{X}))^2 = \int_{\mathbb{R}^d} g(\boldsymbol{x})^2 f(\boldsymbol{x}) d\boldsymbol{x} - \theta^2$$

Si σ est connu, on obtient un intervalle de confiance pour θ à la quasi-certitude 95% de la forme :

$$\left[\hat{\theta}_n - \frac{1{,}96\sigma}{\sqrt{n}} \ , \ \hat{\theta}_n + \frac{1{,}96\sigma}{\sqrt{n}}\right] .$$

Si σ est inconnu (situation habituelle), il faut estimer sa valeur elle aussi. Deux méthodes :

Méthode conservative : Si g est bornée, alors $\sigma \leq \|g\|_\infty$. Donc

$$\mathbb{P}\left(\theta \in \left[\hat{\theta}_n - \frac{1{,}96\|g\|_\infty}{\sqrt{n}} \ , \ \hat{\theta}_n + \frac{1{,}96\|g\|_\infty}{\sqrt{n}}\right]\right) \geq 0{,}95$$

Méthode asymptotique : On définit :

$$\hat{\sigma}_n = \left(\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i)^2 - \hat{\theta}_n^2\right)^{1/2} ,$$

qui vérifie $\hat{\sigma}_n \to \sigma$ quand $n \to +\infty$ d'après la loi des grands nombres, et on utilise alors l'intervalle de confiance asymptotique :

$$\left[\hat{\theta}_n - \frac{1{,}96\hat{\sigma}_n}{\sqrt{n}} \ , \ \hat{\theta}_n + \frac{1{,}96\hat{\sigma}_n}{\sqrt{n}}\right] .$$

Tests

Un jeu de pile ou face avec deux joueurs : vous et votre meilleur(e) ami(e). Pile : vous lui donnez un euro. Face : il(elle) vous donne un euro.

Vous jouez 500 fois. Vous donnez 400 euros et vous recevez 100 euros. C'est louche... La pièce est-t-elle truquée ?

Tests

Comment répondre ?

1. 1er choix : Décider que la pièce est truquée et accuser votre ami(e) d'avoir truqué le jeu.
   Attention, il y a un risque : perdre votre ami.
2. 2ème choix : Décider que l'on peut accepter que le jeu est équilibré et que vous n'avez pas eu de chance.
   Attention, il y a un risque : être pris pour une andouille.

Le modèle statistique

On considère $\boldsymbol{X} = (X_1, \ldots, X_n)$ un n-échantillon du modèle statistique $\boldsymbol{\mathcal{P}} = \{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$.
Soit $H_0 \subset \Theta$ donné par le contexte.
On souhaite décider, au vu des observations, si l'on peut accepter l'hypothèse :

$$H_0 : \quad \boldsymbol{\theta} \in H_0$$

ou au contraire refuser cette hypothèse :

$$H_1 : \quad \boldsymbol{\theta} \notin H_0$$

Exemple : le modèle statistique est $\boldsymbol{\mathcal{P}} = \{\mathcal{B}(p), p \in [0,1]\}$. On veut tester l'hypothèse $H_0 = \{1/2\}$ à partir d'un échantillon à valeurs dans $\{P, F\}^n$ de taille $n = 500$.

Tests

Deux manières de se tromper :
Rejeter H_0 (et accepter H_1), alors que H_0 est vraie (risque de première espèce).
Accepter H_0 alors que H_0 est fausse (risque de seconde espèce).

Décision \ Réalité	H_0 vraie	H_1 vraie
H_0 acceptée	correct	risque de seconde espèce
H_1 acceptée	risque de première espèce	correct

Règle de décision et région critique

- On observe l'échantillon et on souhaite tester si $\boldsymbol{\theta} \in H_0$ ou pas.
- Un test est déterminé par sa région critique W qui constitue un sous-ensemble de l'ensemble $\mathcal{X}^n$ des valeurs possibles de $\boldsymbol{X}$.
- La règle de décision du test associé à W est la suivante.

Lorsqu'on observe $\boldsymbol{x} = (x_1, \ldots, x_n)$,
- si $\boldsymbol{x} \in W$, alors on rejette H_0 et on accepte H_1 i.e. on décide que $\boldsymbol{\theta} \in H_1$,
- si $\boldsymbol{x} \notin W$, alors on accepte H_0 et on rejette H_1 i.e. on décide que $\boldsymbol{\theta} \in H_0$.

Exemple : on peut construire un test basé sur la région critique

$$W = \left\{ \boldsymbol{x} \in \{P, F\}^n \, , \, \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_P(x_i) - \frac{1}{2} \right| > \epsilon \right\}$$

pour un certain ϵ. Cela veut dire qu'on rejette l'hypothèse H_0 (la pièce est non truquée) si la proportion empirique de Pile s'écarte "trop" de $1/2$. Comment choisir ϵ ?