ECHANTILLONNAGE, ESTIMATION

I Echantillonnage, estimation ponctuelle

<u>exemple d'introduction</u> : on considère une population de N individus et l'on s'attache à un caractère X v.a. de moyenne m, de variance σ^2.

Pour étudier X, on prélève un échantillon de n individus et l'on note les caractères observés $x_1, \dots, x_n$: réalisations de n v.a. $X_1, \dots, X_n$.

- Si le tirage se fait <u>avec remise</u> : $X_1, \dots, X_n$ sont iid. Donc $\mathbb{E}(X_i) = m$, $var(X_i) = \sigma^2$.

  estimation de m : $\hat{m}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

  estimation ponctuelle : $\frac{1}{n}\sum_{i=1}^{n} x_i$.

  $\mathbb{E}(\hat{m}_n) = m$, $var(\hat{m}_n) = \frac{\sigma^2}{n}$ → erreur d'échantillonnage.

  $\hat{m}_n$: estimateur de m sans biais.

2 modèles : • 1 expérience globale $Y : \Omega \to \mathbb{R}^n$, $\omega \mapsto (X_1(\omega), \dots, X_n(\omega))$

• n expériences séparées : $Z : \Omega^n \to \mathbb{R}^n$, $(\omega_1, \dots, \omega_n) \mapsto (X(\omega_1), \dots, X(\omega_n))$

Si $\pi_i : \Omega^n \to \Omega$, $(\omega_1, \dots, \omega_n) \mapsto \omega_i$, $Z = (X \circ \pi_1, \dots, X \circ \pi_n)$. En posant $X_i = X \circ \pi_i$, $Z \equiv Y$.

- Si le tirage se fait <u>sans remise</u> : les $X_1, \dots, X_n$ ne sont plus indépendants

  $\mathbb{E}(\hat{m}_n) = m$ mais $var(\hat{m}_n) = \frac{N-n}{N-1}\frac{\sigma^2}{n} \underset{N \text{ grand}}{\approx} (1-f)\frac{\sigma^2}{n}$, $f = \frac{n}{N}$

  Plus f est grand, plus $var(\hat{m}_n)$ sera petit, meilleure sera l'estimation. Mais ce cas est plus difficile à étudier que le précédent.

<u>Problème</u> : bien souvent σ est inconnu et on aura besoin d'estimer σ pour prévoir l'erreur d'échantillonnage.

Dans la suite on ne considérera que des tirages <u>avec remise</u> (non exhaustifs)

II Estimateur.

1) <u>Définition</u> : Considérons une population suivant une loi dépendant d'un paramètre $\theta \in \Theta$ inconnu. Un estimateur de θ est une fonction d'un échantillon iid $(X_1, \dots, X_n)$

$\hat{\theta}_n = f(X_1, \dots, X_n)$. (v.a.)

Une estimation ponctuelle de θ est une valeur observée $f(x_1, \dots, x_n)$.

$\hat{\theta}_n$ estimateur sans biais si $\mathbb{E}(\hat{\theta}_n) = \theta$. Sinon, le biais est $B(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$.

<u>Objectif</u> : construire des estimateurs, sans biais si possible, convergeant vers θ et le plus rapidement possible.

<u>Exemples</u> : rappels : si $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $V = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

pour une population telle que $\mathbb{E}(X) = m$, $var(X) = \sigma^2$, $\mathbb{E}(X-m)^i = m'_i$.

$$\begin{cases} \mathbb{E}(\bar{X}) = m, \quad var(\bar{X}) = \frac{\sigma^2}{n} \\ \mathbb{E}(V) = \frac{n-1}{n}\sigma^2, \quad var(V) \sim \frac{m'_4 - \sigma^2}{n} \\ cov(\bar{X}, V) = \frac{n-1}{n^2} m'_3 \end{cases}$$

a) <u>estimation de la moyenne m</u> : $\boxed{\hat{m}_n = \frac{1}{n}\sum_{i=1}^{n} X_i}$, $\mathbb{E}(\hat{m}_n) = m$ → sans biais.

b) <u>estimation de la variance σ^2</u> :

<u>1^{er} cas</u> : m connue. $\boxed{\widehat{\sigma_n^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2}$, $\mathbb{E}(\widehat{\sigma_n^2}) = \sigma^2$ → sans biais

<u>2^e cas</u> : m inconnue. * $\boxed{\widehat{s_n^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{m}_n)^2}$, $\mathbb{E}(\widehat{s_n^2}) = (1 - \frac{1}{n})\sigma^2$

→ biais $B(\widehat{s_n^2}) = -\frac{\sigma^2}{n} \xrightarrow[n \to \infty]{} 0$.

asymptotiquement sans biais

* $\boxed{\widehat{S_n^2} = \frac{n}{n-1}\widehat{s_n^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{m}_n)^2}$, $\mathbb{E}(\widehat{S_n^2}) = \sigma^2$ → sans biais.

c) <u>estimation d'une proportion</u> : Soit $p = \frac{N_A}{N}$ la proportion d'individus possédant une propriété A. On prélève n individus, on en dénombre n_A possédant la propriété A

→ proportion observée $\frac{n_A}{n}$ (estimation ponctuelle).

En fait p est la moyenne de la loi de Bernoulli $\mathcal{B}(p)$. Soit $X_i = 1_{\{i^e \text{ individu a la prop } A\}}$
On obtient ainsi l'estimateur $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\mathbb{E}(\hat{p}_n) = \mathbb{E}(X) = \mathbb{P}(A) = p$ → sans biais.

$\hat{n}_A = n\hat{p}_n$ suit la loi binomiale $\mathcal{B}(n,p)$.

2) Précision d'un estimateur

- Soit $\hat{\theta}_n$ un estimateur sans biais de θ. On mesurera la précision de $\hat{\theta}_n$ à l'aide de la variance $var(\hat{\theta}_n) = \mathbb{E}[(\hat{\theta}_n - \theta)^2]$.
Si $\tilde{\theta}_{1,n}$ et $\tilde{\theta}_{2,n}$ sont deux tels estimateurs, $\hat{\theta}_{1,n}$ est plus précis que $\hat{\theta}_{2,n}$ si $var\,\hat{\theta}_{1,n} < var\,\hat{\theta}_{2,n}$

- Cas général (avec biais) : on mesure la précision de $\hat{\theta}_n$ à l'aide de la fonction de risque $R(\theta) = B(\hat{\theta}_n)^2 + var(\hat{\theta}_n) = \mathbb{E}[(\hat{\theta}_n - \theta)^2]$. La précision de $\hat{\theta}_n$ est $\frac{1}{R(\theta)}$.

<u>Définition</u> :
- Un estimateur sans biais de variance minimale est dit efficace.
- $\hat{\theta}_n$ est un estimateur convergent si $\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta$. (ou "correct")
- $\hat{\theta}_n$ est un estimateur absolument correct s'il est correct et sans biais.

<u>Proposition</u> : si $\hat{\theta}_n$ vérifie $\left\{\begin{array}{l}\mathbb{E}(\hat{\theta}_n) \to \theta \\ \text{ou} \\ \text{sans biais}\end{array}\right.$ et $var(\hat{\theta}_n) \to 0$ alors $\hat{\theta}_n$ est convergent vers θ.

(dém : avec Bienaymé-Tchebychev)

<u>Exemples</u> : 1) $\hat{m}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$: sans biais, $var(\hat{m}_n) = \frac{\sigma^2}{n} \to 0$ donc $\hat{m}_n \xrightarrow[n\to\infty]{P} m$ (absolument correct)

2) $\widehat{\sigma_n^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$: sans biais et $var(\widehat{\sigma_n^2}) = \frac{1}{n^2}\sum_{i=1}^{n} var[(X_i - m)^2]$

$$= \frac{1}{n}\left\{\mathbb{E}[(X_1-m)^4] - \left[\mathbb{E}[(X_1-m)^2]\right]^2\right\}$$

$$= \frac{m'_4 - \sigma^4}{n} \xrightarrow[n\to\infty]{} 0 \quad \text{donc } \widehat{\sigma_n^2} \xrightarrow[n\to\infty]{P} \sigma^2.$$
(absolument correct)

3) $\widehat{s_n^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{m}_n)^2$ → avec biais.

$\widehat{S_n^2} = \frac{n}{n-1}\widehat{s_n^2}$: sans biais

$$var(\widehat{s_n^2}) = \frac{m'_4 - \sigma^4}{n} - 2\frac{m'_4 - 2\sigma^4}{n^2} + \frac{m'_4 - 3\sigma^4}{n^3}$$

$$= \frac{(n-1)^2}{n^3}\left[m'_4 - \frac{n-3}{n-1}\sigma^4\right]$$

besoin de connaître m

$$var(\widehat{S_n^2}) = \frac{1}{n}\left[m'_4 - \frac{n-3}{n-1}\sigma^4\right] \geq var(\widehat{\sigma_n^2}) = \frac{m'_4 - \sigma^4}{n}$$

$$= \left(\frac{n}{n-1}\right)^2 var(\widehat{s_n^2}) \geq var(\widehat{s_n^2})$$

↳ avec biais

$\widehat{S_n^2}$ a une plus grande variance que $\widehat{s_n^2}$ mais $\widehat{s_n^2}$ a un biais.

3) Estimateur du maximum de vraisemblance

<u>Définition</u> : vraisemblance d'un échantillon : $L(x_1,\dots,x_n;\theta) = P_\theta(X_1 = x_1,\dots,X_n = x_n) = \prod_{i=1}^{n}\mathbb{P}(X = x_i)$
ou $\prod_{i=1}^{n} f_X(x_i;\theta)$.

On cherche le ou les θ qui maximisent $\theta \mapsto L(x_1,\dots,x_n;\theta)$.
Ceci se fait en général en résolvant l'équation $\frac{\partial}{\partial\theta}\ln L(x_1,\dots,x_n;\theta) = 0$.

<u>exemples</u> : 1) <u>loi binomiale $\mathcal{B}(\nu,\theta)$</u> : $L(x_1,\dots,x_n;\theta) = \prod_{i=1}^{n} C_\nu^{x_i}\ \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n\nu - \sum_{i=1}^{n} x_i}$

$$\frac{\partial}{\partial\theta}\ln L(x_1,\dots,x_n;\theta) = \frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{n\nu - \sum_{i=1}^{n} x_i}{1-\theta} = 0 \Rightarrow \theta = \frac{\sum_{i=1}^{n} x_i}{n\nu}$$

$\rightarrow \boxed{\hat{\theta}_{MDV} = \frac{\bar{X}}{\nu}}$, $\mathbb{E}(\hat{\theta}_{MDV}) = \theta$, $var(\hat{\theta}_{MDV}) = \frac{\theta(1-\theta)}{n\nu}$, $\hat{\theta}_{MDV}$ absolument correct.

2) <u>loi de Poisson $\mathcal{P}(\theta)$</u> : $L(x_1,\dots,x_n;\theta) = \dfrac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\theta}$

$$\frac{\partial}{\partial\theta} \ln L(x_1,\dots,x_n;\theta) = \frac{\sum_{i=1}^n x_i}{\theta} - n = 0 \Rightarrow \theta = \frac{1}{n}\sum_{i=1}^n x_i .$$

$\rightarrow \boxed{\hat{\theta}_{MDV} = \overline{X}}$, $\quad \mathbb{E}(\hat{\theta}_{MDV}) = \theta$, $\operatorname{var}(\hat{\theta}_{MDV}) = \frac{\theta}{n}$, $\quad \hat{\theta}_{MDV}$ absolument correct.

3) <u>loi exponentielle $\mathcal{E}(\theta)$</u> : $L(x_1,\dots,x_n;\theta) = \theta^n e^{-\theta\sum_{i=1}^n x_i}$

$$\frac{\partial}{\partial\theta} \ln L(x_1,\dots,x_n;\theta) = \frac{n}{\theta} - \sum_{i=1}^n x_i = 0 \Rightarrow \theta = \frac{n}{\sum_{i=1}^n x_i} .$$

$\rightarrow \boxed{\hat{\theta}_{MDV} = \frac{1}{\overline{X}}}$ $\quad$ densité de $\hat{\theta}_{MDV}$: $f_{\hat{\theta}_{MDV}}(t) = \dfrac{(n\theta)^n}{(n-1)!}\dfrac{e^{-\frac{n\theta}{t}}}{t^{n+1}}$

$$\Rightarrow \mathbb{E}(\hat{\theta}_{MDV}) = \frac{n}{n-1}\theta, \quad \mathbb{E}(\hat{\theta}^2_{MDV}) = \frac{n^2}{(n-1)(n-2)}\theta^2$$

$\operatorname{var}(\hat{\theta}_{MDV}) = \dfrac{n^2}{(n-1)^2(n-2)}\theta^2$, $\hat{\theta}_{MDV}$ correct mais avec biais.

4) <u>loi normale $N(\theta,\sigma^2)$</u> : $L(x_1,\dots,x_n;\theta) = \dfrac{\exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - 2\theta\sum_{i=1}^n x_i + n\theta^2\right)\right]}{(\sqrt{2\pi}\sigma)^n}$

$$\frac{\partial}{\partial\theta} \ln L(x_1,\dots,x_n;\theta) = \frac{1}{\sigma^2}\left[\sum_{i=1}^n x_i - n\theta\right] = 0 \Rightarrow \theta = \frac{1}{n}\sum_{i=1}^n x_i .$$

$\boxed{\hat{\theta}_{MDV} = \overline{X}}$ $\quad \mathbb{E}(\hat{\theta}_{MDV}) = \theta$, $\operatorname{var}(\hat{\theta}_{MDV}) = \frac{\sigma^2}{n}$ $\quad \hat{\theta}_{MDV}$ absolument correct

5) <u>loi normale $N(m,\theta)$</u> : $L(x_1,\dots,x_n;\theta) = \dfrac{\exp\left[-\frac{1}{2\theta}\sum_{i=1}^n (x_i - m)^2\right]}{(\sqrt{2\pi\theta})^n}$

$$\frac{\partial}{\partial\theta} \ln L(x_1,\dots,x_n;\theta) = \frac{1}{2\theta^2}\sum_{i=1}^n (x_i-m)^2 - \frac{n}{2\theta} = 0 \Rightarrow \theta = \frac{1}{n}\sum_{i=1}^n (x_i - m)^2$$

$\rightarrow \boxed{\hat{\theta}_{MDV} = \frac{1}{n}\sum_{i=1}^n (X_i - m)^2}$, $\quad \mathbb{E}(\hat{\theta}_{MDV}) = \theta$, $\operatorname{var}(\hat{\theta}_{MDV}) = \frac{m_4' - \sigma^4}{n} = \frac{2\theta^2}{n}$, abs. correct.

<u>III. Estimation par intervalle</u>

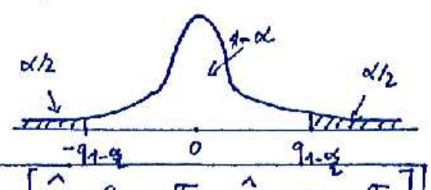
<u>Définition</u> : Un intervalle de confiance de niveau de confiance $1-\alpha$ ($\alpha \in]0,1[$) pour un paramètre θ est un intervalle aléatoire $[Z_1, Z_2]$ (Z_1, Z_2 : v.a.) tel que $\mathbb{P}(\theta \in [Z_1, Z_2]) = 1-\alpha$. α est le risque que $\theta \notin [Z_1, Z_2]$.

1) <u>Intervalle de confiance pour la moyenne m d'une population $N(m,\sigma^2)$</u>

a) <u>cas σ^2 connue</u> : $\hat{m}_n = \frac{1}{n}\sum_{i=1}^n X_i : N(m, \frac{\sigma^2}{n})$ donc $N = \dfrac{\hat{m}_n - m}{\sigma/\sqrt{n}} : N(0,1)$.

Soit $q_{1-\frac{\alpha}{2}}$ le $(1-\frac{\alpha}{2})$-quantile de la loi $N(0,1)$: $\mathbb{P}(N \leq q_{1-\frac{\alpha}{2}}) = 1-\frac{\alpha}{2} \Rightarrow \mathbb{P}(|N| \leq q_{1-\frac{\alpha}{2}}) = 1-\alpha$

On a alors
$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(\left|\frac{\hat{m}_n - m}{\sigma/\sqrt{n}}\right| \leq q_{1-\frac{\alpha}{2}}\right) \\ &= \mathbb{P}\left(m \in \left[\hat{m}_n - q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \hat{m}_n + q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]\right) \\ &= \mathbb{P}(m \in I_m) \end{aligned}$$



d'où un intervalle de confiance pour m de niveau $1-\alpha$: $\boxed{I_m = \left[\hat{m}_n - q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \hat{m}_n + q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]}$

Longueur de $I_m = 2q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$.

- Plus on demande un niveau de confiance élevé (α plus petit), plus long(I_m) sera grande (intervalle moins précis)
- Plus on demande une précision fine (long(I_m) petite), plus α sera grand (risque plus important)
- On peut alors jouer sur n (augmenter la taille de l'échantillon) : si on veut une précision de $\pm\varepsilon$ (i.e. long(I_n) $= 2\varepsilon$), choisir $n \geq \left(\frac{\sigma}{\varepsilon} q_{1-\frac{\alpha}{2}}\right)^2$.

quelques valeurs numériques : $\alpha = 0.01$ (niveau 99 %) $1-\frac{\alpha}{2} = 0.995$ $q_{0.995} = 2.57$
$\alpha = 0.05$ (niveau 95 %) $1-\frac{\alpha}{2} = 0.975$ $q_{0.975} = 1.96$
$\alpha = 0.1$ (niveau 90 %) $1-\frac{\alpha}{2} = 0.95$ $q_{0.95} = 1.64$

b) cas σ^2 inconnue : on utilise $\widehat{s_n^2} = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{m}_n)^2$ ou $\widehat{S_n^2} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \widehat{m}_n)^2$.

$$T_n \equiv \frac{\widehat{m}_n - m}{\sqrt{\widehat{s_n^2}/(n-1)}} = \frac{\widehat{m}_n - m}{\sqrt{\widehat{S_n^2}/n}} : \text{Student } \mathcal{T}(n-1) \quad \text{(même type de courbe que } N(0,1))$$

Soit $t_{1-\frac{\alpha}{2}}(n-1)$ le $(1-\frac{\alpha}{2})$-quantile de $\mathcal{T}(n-1)$: $\mathbb{P}(T_n \leq t_{1-\frac{\alpha}{2}}(n-1)) = 1-\frac{\alpha}{2}$
$\Rightarrow \mathbb{P}(|T_n| \leq t_{1-\frac{\alpha}{2}}(n-1)) = 1-\alpha$.

D'où un intervalle de confiance pour m de niveau $1-\alpha$:

$$\boxed{I_m = \left[\widehat{m}_n - t_{1-\frac{\alpha}{2}}(n-1)\sqrt{\frac{\widehat{s_n^2}}{n-1}} \;,\; \widehat{m}_n + t_{1-\frac{\alpha}{2}}(n-1)\sqrt{\frac{\widehat{s_n^2}}{n-1}}\right]}$$

Pour n "grand" ($n \geq 30$) : $T(n-1) \simeq N(0,1)$ et $t_{1-\frac{\alpha}{2}}(n-1) \approx q_{1-\frac{\alpha}{2}}$ → petite erreur supplémentaire

remarque : pour des échantillons sans remise il faut rajouter un coefficient d'exhaustivité $\sqrt{1-f}$ ($f = \frac{n}{N}$ proportion de l'échantillon) devant σ et $\sqrt{\widehat{s_n^2}}$.

2) Cas d'une population inconnue

a) si n est "grand" ($n \geq 30$) : approximation normale $\frac{\widehat{m}_n - m}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} N(0,1)$

→ même type d'intervalle de confiance (avec erreur supplémentaire due à l'approximation)

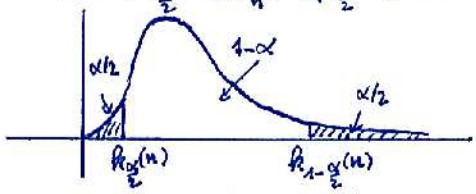
b) si $n < 30$: rien !

3) Intervalle de confiance pour la variance σ^2 d'une population $N(m, \sigma^2)$.

a) cas m connue : $\widehat{\sigma_n^2} = \frac{1}{n}\sum_{i=1}^n (X_i - m)^2$, $\quad K_n \equiv \frac{\widehat{\sigma_n^2}}{\sigma^2/n} = \sum_{i=1}^n \left(\frac{X_i - m}{\sigma}\right)^2 : \chi^2(n)$

Soit $k_{\frac{\alpha}{2}}(n)$ et $k_{1-\frac{\alpha}{2}}(n)$ les $\frac{\alpha}{2}$ et $(1-\frac{\alpha}{2})$-quantiles de $\chi^2(n)$: $\mathbb{P}(K_n \leq k_{\frac{\alpha}{2}}(n)) = \frac{\alpha}{2}$
$\mathbb{P}(K_n \leq k_{1-\frac{\alpha}{2}}(n)) = 1-\frac{\alpha}{2}$
$\Rightarrow \mathbb{P}(k_{\frac{\alpha}{2}}(n) \leq K_n \leq k_{1-\frac{\alpha}{2}}(n)) = 1-\alpha$.

On a alors
$$1-\alpha = \mathbb{P}\left(\frac{\widehat{\sigma_n^2}}{\sigma^2/n} \in \left[k_{\frac{\alpha}{2}}(n), k_{1-\frac{\alpha}{2}}(n)\right]\right)$$
$$= \mathbb{P}\left(\sigma^2 \in \left[\frac{n\widehat{\sigma_n^2}}{k_{1-\frac{\alpha}{2}}(n)}, \frac{n\widehat{\sigma_n^2}}{k_{\frac{\alpha}{2}}(n)}\right]\right)$$
$$= \mathbb{P}(\sigma^2 \in I_{\sigma^2})$$



d'où un intervalle de confiance pour σ^2 de niveau $1-\alpha$: $\boxed{I_{\sigma^2} = \left[\frac{n\widehat{\sigma_n^2}}{k_{1-\frac{\alpha}{2}}(n)}, \frac{n\widehat{\sigma_n^2}}{k_{\frac{\alpha}{2}}(n)}\right]}$

b) cas m inconnue : on utilise $\widehat{s_n^2} = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{m}_n)^2$ ou $\widehat{S_n^2} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \widehat{m}_n)^2$

Dans ce cas $K_{n-1} \equiv \frac{\widehat{s_n^2}}{\sigma^2/n} = \frac{\widehat{S_n^2}}{\sigma^2/(n-1)} = \sum_{i=1}^n \left(\frac{X_i - \widehat{m}_n}{\sigma}\right)^2 : \chi^2(n-1)$

$$\boxed{I_{\sigma^2} = \left[\frac{n\widehat{s_n^2}}{k_{1-\frac{\alpha}{2}}(n-1)}, \frac{n\widehat{s_n^2}}{k_{\frac{\alpha}{2}}(n-1)}\right]}$$

4) Estimation d'une proportion

Soit $p = \frac{N_A}{N}$ la proportion d'individus possédant un caractère particulier A.

$\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ où $X_i = \mathbb{1}_{\{i^e \text{ individu a la prop. } A\}} \to$ Bernoulli $B(p)$, $n\hat{p}_n : B(n,p)$.

$\mathbb{E}(\hat{p}_n) = p$ et $\mathrm{var}(\hat{p}_n) = \frac{p(1-p)}{n}$.

Lorsque n est "grand", $\quad \frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \approx N(0,1)$.

On trouverait donc ainsi approximativement un intervalle de confiance pour p

$$I_p = \left[\hat{p}_n - q_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}},\ \hat{p}_n + q_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \right], \quad q_{1-\frac{\alpha}{2}} : (1-\tfrac{\alpha}{2})\text{-quantile de } N(0,1).$$

Malheureusement p est inconnue ! Exutoire :

a) <u>estimer p par $\hat{p}_n$</u> : $\quad I_p = \left[\hat{p}_n - q_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}},\ \hat{p}_n + q_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \right]$

$\to$ très approximatif !

b) <u>intervalle de confiance par excès</u> : $\quad p(1-p) \leq \frac{1}{4} \Rightarrow \sqrt{\frac{p(1-p)}{n}} \leq \frac{1}{2\sqrt{n}}$

$$I_p = \left[\hat{p}_n - \frac{q_{1-\frac{\alpha}{2}}}{2\sqrt{n}},\ \hat{p}_n - \frac{q_{1-\frac{\alpha}{2}}}{2\sqrt{n}} \right] \text{ et } \mathbb{P}(p \in I_p) \geq 1-\alpha.$$

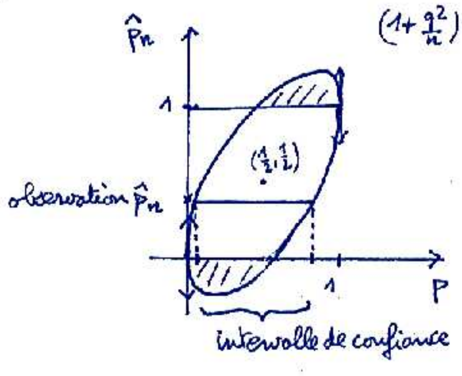
intervalle un peu plus large (moins précis)

c) <u>méthode de l'ellipse (valable aussi pour les petits échantillons)</u> :

$$p \in \left[\hat{p}_n - q\sqrt{\frac{p(1-p)}{n}},\ \hat{p}_n + q\sqrt{\frac{p(1-p)}{n}} \right] \Leftrightarrow |p - \hat{p}_n| \leq q\sqrt{\frac{p(1-p)}{n}} \quad (q \text{ nb donné}; \text{ si } n \text{ grand } q = q_{1-\frac{\alpha}{2}}.)$$

$$\Leftrightarrow (p, \hat{p}_n) \in \text{Ellipse } (x-y)^2 - \frac{q^2}{n} x(1-x) \leq 0$$

$$\left(1+\frac{q^2}{n}\right)x^2 - 2xy + y^2 - \frac{q^2}{n}x \leq 0$$



d) <u>Utilisation de la fonction arcsinus</u> : $\quad$ Soit $Z = 2\sqrt{n}\left[\arcsin\sqrt{\hat{p}_n} - \arcsin\sqrt{p}\right]$.

$$\begin{aligned}
\mathbb{P}(Z \leq z) &= \mathbb{P}\left(\hat{p}_n \leq \sin^2(\arcsin\sqrt{p} + z/2\sqrt{n})\right) = \mathbb{P}\left(\hat{p}_n \leq \left[\sqrt{p}\cos(z/2\sqrt{n}) + \sqrt{1-p}\sin(z/2\sqrt{n})\right]^2\right) \\
&= \mathbb{P}\left(\hat{p}_n \leq p\cos^2(z/2\sqrt{n}) + (1-p)\sin^2(z/2\sqrt{n}) + \sqrt{p(1-p)}\sin(z/\sqrt{n})\right) \\
&= \mathbb{P}\Bigg(\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \leq \underbrace{\sqrt{n}\,\frac{(1-2p)}{\sqrt{p(1-p)}}\sin^2\left(\frac{z}{2\sqrt{n}}\right) + \sqrt{n}\sin\frac{z}{\sqrt{n}}}_{\sim z}\Bigg) \xrightarrow[n\to\infty]{\mathcal{L}} \mathbb{P}(N < z) = \Phi(z).
\end{aligned}$$

d'où $I_{\arcsin\sqrt{p}} = \left[\arcsin\sqrt{\hat{p}_n} - \frac{q_{1-\frac{\alpha}{2}}}{2\sqrt{n}},\ \arcsin\sqrt{\hat{p}_n} + \frac{q_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right]$.

<u>Bibliographie</u> : Saporta : Probabilités, analyse des données et statistique 1990
Droesbeke : Eléments de statistique 1991
Fourgeaud et Fuchs : Statistique 1967
Aide-mémoire statistique 1995
Polycope INSA d'Anguenot-Fanchon et al.

Complément : estimation par intervalle directe à partir d'une seule observation.

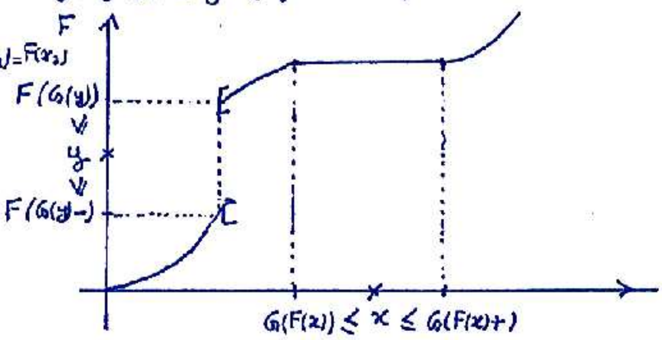
Théorème : Soit F_θ la f.r. de X : $F_\theta(x) = \mathbb{P}_\theta(X \leq x)$, $F_\theta^-(x) = F_\theta(x-) = \mathbb{P}_\theta(X < x)$.
On suppose que $\forall x$, $\theta \in (\theta_0, \theta_1) \mapsto F_\theta(x)$ est continue strictement décroissante (resp. croissante) ainsi que $\theta \mapsto F_\theta^-(x)$, et que $\lim\limits_{\substack{\theta \to \theta_0 \\ \theta_1}} F_\theta(x) = \begin{smallmatrix} 1 \\ 0 \end{smallmatrix}$ (resp. $\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}$)
Soit $\alpha \in]0,1[$. On a :

$$\forall x,\ \exists!\, \theta_\alpha^+(x) \in (\theta_0, \theta_1) \ /\ F_{\theta_\alpha^+(x)}(x) = \alpha \quad \text{et} \quad \exists!\, \theta_{1-\alpha}^-(x) \in (\theta_0, \theta_1) \ /\ F^-_{\theta^-_{1-\alpha}(x)}(x) = 1-\alpha .$$

Alors $(\theta_0, \theta_\alpha^+(x)]$, $[\theta_{1-\alpha}^-(x), \theta_1)$ et $[\theta^-_{1-\frac{\alpha}{2}}(x), \theta^+_{\frac{\alpha}{2}}(x)]$ sont des intervalles de confiance de θ de niveau $\geq 1-\alpha$.

Démonstration : 1) $\mathbb{P}\{\theta \in (\theta_0, \theta_\alpha^+(X)]\} = \mathbb{P}(F_\theta(X) \geq \alpha)$.

Soit G_θ l'inverse continue à gauche de F_θ : $G_\theta(y) = \inf\{x : F_\theta(x) \geq y\}$.
On a $F_\theta(x) \geq y \iff G_\theta(y) \leq x$ et $F_\theta^-(G_\theta(y)) \leq y$. (cf dessin)



$\left\{ \begin{array}{l} \text{si } F|_{[x_1,x_2]} = C^{te},\ \forall x \in [x_1,x_2]\ F(x) = F(x_1) = F(x_2) \\ \Rightarrow G(F(x)) = x_1 \text{ et } G(F(x)) = x_2 \\ \text{si } G|_{[y_1,y_2]} = C^{te},\ \forall y \in [y_1,y_2],\ G(y) = G(y_1) = G(y_2) \\ \Rightarrow F(G(y)-) = y_1 \text{ et } F(G(y)) = y_2 . \end{array} \right.$

D'où $\mathbb{P}\{\theta \in (\theta_0, \theta_\alpha^+(X)]\} = \mathbb{P}(X \geq G_\theta(\alpha)) = 1 - \mathbb{P}(X < G_\theta(\alpha)) = 1 - F_\theta^-(G_\theta(\alpha)) \geq 1-\alpha$.

2) le 2e cas se traite avec $Y = -X$, $\varphi = -\theta$, $G_\varphi(x) = \mathbb{P}_\varphi(Y \leq x) = \mathbb{P}_{-\varphi}(X \geq -x) = 1 - F_\theta^-(-x)$.

3) $\mathbb{P}(\theta \in [\theta^-_{1-\frac{\alpha}{2}}(X), \theta^+_{\frac{\alpha}{2}}(X)]) = \mathbb{P}(A \cap B)$ avec $A = \{\theta \in [\theta^-_{1-\frac{\alpha}{2}}(X), \theta_1)\}$, $B = \{\theta \in (\theta_0, \theta^+_{\frac{\alpha}{2}}(X)]\}$
$= \mathbb{P}(A) + \mathbb{P}(B) - \underbrace{\mathbb{P}(A \cup B)}_{\leq 1}$ $\qquad \mathbb{P}(A), \mathbb{P}(B) \geq 1 - \frac{\alpha}{2}$
$\geq 1-\alpha$. □

Exemples : 1) Loi binomiale $\mathcal{B}(n, \theta)$. $\qquad F_\theta(k) = \sum\limits_{i=0}^{k} C_n^i \theta^i (1-\theta)^{n-i}$

$\theta \in [0,1] \mapsto F_\theta(k)$ est continue strictement décroissante :
$$\frac{d}{d\theta} F_\theta(k) = \sum_{i=0}^{k} C_n^i \left[i\theta^{i-1}(1-\theta)^{n-i} - (n-i)\theta^i (1-\theta)^{n-i-1} \right] = n! \left[\sum_{i=1}^{k} \frac{\theta^{i-1}}{(i-1)!} \frac{(1-\theta)^{n-i}}{(n-i)!} - \sum_{i=0}^{k} \frac{\theta^i}{i!} \frac{(1-\theta)^{n-i-1}}{(n-i-1)!} \right]$$
$$= -\frac{n!}{k!(n-k-1)!} \theta^k (1-\theta)^{n-k-1} < 0 .$$
Equation $F_\theta(k) = \alpha$ difficile à résoudre !

2) Loi de Poisson $\mathcal{P}(\theta)$. $\qquad F_\theta(n) = \sum\limits_{i=0}^{n} e^{-\theta} \frac{\theta^i}{i!} = 1 - \underbrace{\int_0^\theta \frac{x^n}{n!} e^{-x} dx}_{F_{\chi^2_{(2n+2)}}(2\theta)}$ (lien Poisson - Gamma) $\to$ f.r. χ^2.

$\theta \in [0, +\infty[\mapsto F_\theta(n)$ continue strict. décroissante.

$F_\theta(n) = \alpha \iff \int_0^\theta \frac{x^n}{n!} e^{-x} dx = 1-\alpha \iff \theta = \theta_\alpha^+(n) = \frac{1}{2} \chi^2_{1-\alpha}(2n+2)$.

$F_\theta^-(n) = 1-\alpha \iff F_\theta(n-1) = 1-\alpha \iff \theta = \theta_\alpha^-(n) = \frac{1}{2} \chi^2_\alpha(2n)$.

D'où des intervalles de confiance pour θ : $[0, \frac{1}{2}\chi^2_{1-\alpha}(2X+2)]$, $[\frac{1}{2}\chi^2_\alpha(2X), +\infty[$,
$[\frac{1}{2}\chi^2_{\frac{\alpha}{2}}(2X), \frac{1}{2}\chi^2_{1-\frac{\alpha}{2}}(2X+2)]$ de niveau $\geq 1-\alpha$.

3) <u>Loi géométrique $\mathcal{G}(\theta)$</u>. $F_\theta(n) = \sum_{i=1}^{n} \theta(1-\theta)^{i-1} = 1-(1-\theta)^n$

$\theta \in [0,1] \mapsto F_\theta(n)$ continue strictement croissante.

$F_\theta(n) = \alpha \iff \theta = \theta_\alpha^+(n) = 1-(1-\alpha)^{\frac{1}{n}}$

$F_\theta^-(n) = 1-\alpha \iff F_\theta(n-1) = 1-\alpha \iff \theta = \theta_{1-\alpha}^-(n) = 1-\alpha^{\frac{1}{n-1}}$

D'où les intervalles de confiance : $[0, 1-\alpha^{\frac{1}{X-1}}]$, $[1-(1-\alpha)^{\frac{1}{X}}, 1]$, $[1-(1-\frac{\alpha}{2})^{\frac{1}{X}}, 1-(\frac{\alpha}{2})^{\frac{1}{X-1}}]$.

4) <u>Loi exponentielle $\mathcal{E}(\theta)$</u>. $F_\theta(x) = 1-e^{-\theta x}$

$\theta \in [0,+\infty[\mapsto F_\theta(x)$ continue strict. croissante.

$F_\theta(x) = \alpha \iff \theta = \theta_\alpha^+(x) = -\frac{1}{x}\ln(1-\alpha)$

D'où les intervalles de confiance : $[0, -\frac{1}{X}\ln\alpha]$, $[-\frac{1}{X}\ln(1-\alpha), +\infty[$, $[-\frac{1}{X}\ln(1-\frac{\alpha}{2}), -\frac{1}{X}\ln\frac{\alpha}{2}]$.

5) <u>Loi normale $N(\theta,\sigma^2)$</u>. $F_\theta(x) = \int_{-\infty}^{x-\theta} e^{-\frac{y^2}{2\sigma^2}} \frac{dy}{\sqrt{2\pi}\,\sigma} = \int_{-\infty}^{\frac{x-\theta}{\sigma}} e^{-\frac{y^2}{2}} \frac{dy}{\sqrt{2\pi}}$

$\theta \in \mathbb{R} \mapsto F_\theta(x)$ continue strict. décroissante. Equation

$F_\theta(x) = \alpha \iff \Phi(\frac{x-\theta}{\sigma}) = \alpha \iff \theta = \theta_\alpha^+(x) = x - \sigma q_\alpha = x + \sigma q_{1-\alpha}$

D'où les intervalles de confiance : $]-\infty, X+\sigma q_{1-\alpha}]$, $[X-\sigma q_{1-\alpha}, +\infty[$, $[X-\sigma q_{1-\frac{\alpha}{2}}, X+\sigma q_{1-\frac{\alpha}{2}}]$

<u>Condition suffisante de monotonie de $\theta \mapsto F_\theta(x)$</u> (rapport de vraisemblance monotone)

> Soit f_θ une densité >0 de X par rapport à une mesure μ ne dépendant pas de θ, et F_θ : f.r. de X. On suppose que $\forall \theta, \theta'$, $\theta < \theta' \Rightarrow \left(y \mapsto \frac{f_{\theta'}(y)}{f_\theta(y)}\right)$ est strict. croissante (resp. décroissante).
> Alors $\forall x$ tel que $\forall \theta$, $0 < F_\theta(x) < 1$, $\theta \mapsto F_\theta(x)$ est strictement décroissante (resp. croissante).

<u>démonstration</u> : Soit $\theta < \theta'$, $a = F_\theta(x)$, $b = \frac{f_{\theta'}(x)}{f_\theta(x)}$.

$$f_{\theta'}(y) - b f_\theta(y) = f_\theta(y)\left(\frac{f_{\theta'}(y)}{f_\theta(y)} - \frac{f_{\theta'}(x)}{f_\theta(x)}\right) \begin{cases} > 0 & \text{si } y > x \quad (\text{resp. } y < x) \\ < 0 & \text{si } y < x \quad (\text{resp. } y > x) \end{cases}$$

$$\begin{aligned} F_{\theta'}(x) - F_\theta(x) &= (F_{\theta'}(x) - bF_\theta(x)) + (b-1)F_\theta(x) \\ &= \int_{-\infty}^{x} [f_{\theta'}(y) - bf_\theta(y)]\,d\mu(y) - a\int_{-\infty}^{+\infty} [f_{\theta'}(y) - bf_{\theta'}(y)]\,d\mu(y) \\ &= \underbrace{(1-a)}_{>0} \underbrace{\int_{-\infty}^{x} [f_{\theta'}(y) - bf_\theta(y)]\,d\mu(y)}_{\substack{<0 \\ (\text{resp} >0)}} - \underbrace{a}_{>0} \underbrace{\int_{x}^{+\infty} [f_{\theta'}(y) - bf_{\theta'}(y)]\,d\mu(y)}_{\substack{>0 \\ (\text{resp} <0)}} \end{aligned}$$

< 0 (resp. > 0). $\square$