Lycée Lakanal

2023 - 2024

Cours – Statistiques inférentielles

A Objectifs de la statistique inférentielle

On souhaite étudier une ***population*** Ω (éventuellement infinie), pour laquelle on ne peut mener d'enquête exhaustive. On prélève alors un ***échantillon*** (sous-population) de cette population, sur laquelle on effectue l'analyse souhaitée. On dispose donc d'une information partielle, et entachée d'erreur (appelée ***erreur d'échantillonnage***).

- ***Que dire sur la population à partir de l'analyse de l'échantillon ?***
- ***Quelle erreur commet-on ?***

C'est à ces questions que répond la ***statistique inférentielle*** .

On dispose de deux grandes catégories de méthodes :

- La ***théorie de l'estimation*** , dont l'objectif est d'estimer un ou plusieurs paramètres par une valeur précise (***estimation ponctuelle***), ou un intervalle de valeurs possibles (***estimation par intervalles***).

- La ***théorie des tests*** , dont l'objectif est de confronter une hypothèse théorique sur les paramètres (théoriques) du modèle choisi pour représenter le phénomène étudié, avec la réalité des observations.

B Définitions basiques

Données

A partir d'un n – échantillon (échantillon de taille n), les ***données*** sont les valeurs $x_1, ..., x_n$, considérées comme des réalisations de n variables aléatoires $X_1, ..., X_n$, qui sont définies plus précisément ci-dessous, où l'on tente d'expliquer ce qu'est un *modèle statistique* .

Modèle statistique

On effectue une ***modélisation*** : on se donne des lois *a priori* suivies par les VAR $X_1, ..., X_n$.

Ces lois ne sont pas entièrement déterminées, mais appartiennent à une famille $\left(\mu_\theta\right)_{\theta \in \Theta}$ de lois définies sur une même partie χ de $\mathbb{R}$, et dépendent d'un paramètre $\theta \in \Theta$, scalaire ou vectoriel, *voir la rubrique suivante pour plus de précisions sur les paramètres*…

Le plus souvent, on considère que les $\left(X_k\right)_{k \in [\![1, n]\!]}$ sont indépendantes et identiquement distribuées (ie indépendantes et de même loi).

Bref, pour tout entier $n \in \mathbb{N}^*$, on dit alors que $\left(X_1, ..., X_n\right)$ est (pour tout $\theta \in \Theta$), un

n – échantillon indépendant identiquement distribué ***i.i.d.***).

Paramètres

Les ***paramètres du modèle*** sont les paramètres des lois de probabilité $\left(\mu_\theta\right)_{\theta \in \Theta}$ des VAR $X_1, ..., X_n$ de l'échantillon. Le paramètre, scalaire ou vectoriel, est en général noté θ.

Le paramètre θ est supposé défini sur un ouvert Θ de $\mathbb{R}$, ou de $\mathbb{R}^k$ s'il y a plusieurs paramètres.

Exemples

- $\theta = p$ pour une loi $\mathcal{B}(p)$,
- $\theta = \left(\mu, \sigma^2\right)$ pour une loi normale $\mathcal{N}\left(\mu, \sigma^2\right)$.

On limitera alors cette année la problématique à l'estimation d'un réel de la forme $g(\theta)$, où g désigne une application de $\Theta \subset \mathbb{R}^k$ dans $\mathbb{R}$.

Le but est d'***estimer***, *i.e.* de *localiser* $g(\theta)$ uniquement à partir des données d'un échantillon $x_1, ..., x_n$ de réalisations des VAR $X_1, ..., X_n$ obtenues après avoir observé n fois le phénomène que l'on veut observer dans des conditions identiques et indépendantes.

Estimateur

Un ***estimateur*** de $g(\theta)$ ou encore une ***statistique*** T_n ou T est une fonction des VAR $X_1, ..., X_n$ associées à l'échantillon : $T_n = \varphi_n(X_1, ..., X_n)$.

Un estimateur (ou une statistique) est donc une VAR.

Cette VAR est destinée à fournir une « estimation » (au sens naïf du terme) du paramètre θ.

Un estimateur est donc une **VAR**. On la note parfois aussi θ_n ou θ.

On dit aussi que la suite $(\theta_n)_{n \in \mathbb{N}}$ est **un** estimateur de θ.

Exemple La moyenne empirique $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ sera très fréquemment pris comme estimateur de l'espérance lorsque μ_θ admet une espérance pour tout $\theta \in \Theta$.

Estimation

Une ***estimation*** est une réalisation (valeur) de la VAR « estimateur » : $t_n = \varphi_n(x_1, ..., x_n)$ est appelée **estimation** du paramètre θ. Cette estimation ne dépend que de l'échantillon observé.

On la note θ, mézoci souvent θ, de même que l'estimateur, ce qui est une atrocité, mais k'est ben pratik kan même…

On distingue comme annoncé un peu plus haut des ***estimations ponctuelles*** (« $\theta = ...$ ») et des ***estimations par intervalles*** (« $\theta \in I_\theta \subset \mathbb{R}$ »).

Qualités d'un estimateur

Nous allons dans ce qui suit essayer de dégager QUELQUES propriétés universellement reconnues comme gages de qualité pour un(e suite d') estimateur(s). La liste produite est bien loin loin loin loin d'être exhaustive.

C Qualités d'un estimateur

1_ Biais ou non biais

Soit $T_n = \varphi_n\left(X_1, ..., X_n\right)$ un estimateur de $g(\theta)$.

On présente ici des critères de performance permettant de mesurer la qualité d'un estimateur.

Ces critères doivent être fondés sur l'écart entre l'estimateur T_n et la vraie valeur $g(\theta)$:

$T_n - g(\theta)$ est l' ***erreur d'estimation*** . Si l'on suppose que, pour tout $\theta \in \Theta$, T_n admet une espérance pour la probabilité $\mathbb{P}_\theta$, notée $\mathbb{E}_\theta\left(T_n\right)$, on a alors la décomposition :

$$\underbrace{T_n - g(\theta)}_{\text{erreur d'estimation}} = \underbrace{\left[T_n - \mathbb{E}_\theta\left(T_n\right)\right]}_{\text{erreur aléatoire}} + \underbrace{\left[\mathbb{E}_\theta\left(T_n\right) - g(\theta)\right]}_{\text{erreur systématique, ou biais}}.$$

Définition 1

Soit $T_n = \varphi_n\left(X_1, ..., X_n\right)$ un estimateur de $g(\theta)$.

On suppose que, pour tout $\theta \in \Theta$, T_n admet une espérance pour la probabilité $\mathbb{P}_\theta$, cette espérance étant notée $\mathbb{E}_\theta\left(T_n\right)$, ou parfois simplement $\mathbb{E}\left(T_n\right)$.

- On appelle ***biais*** de l'estimateur T_n de $g(\theta)$ l'application b_{T_n} définie par :

$$b_{T_n} : \begin{pmatrix} \Theta & \to & \mathbb{R} \\ \theta & \mapsto & b_{T_n}(\theta) = \mathbb{E}_\theta\left(T_n\right) - g(\theta) \end{pmatrix}.$$

- On appelle ***biais en θ*** de l'estimateur T_n de $g(\theta)$ le ***réel***

$$\mathbb{E}_\theta\left(T_n\right) - g(\theta).$$

Inutile de dire que le « en θ » de la deuxième définition passera souvent à la trappe, et que l'on confondra comme on le fait très régulièrement en Analyse ou en Probabilités la fonction avec la valeur prise par la fonction.

Définition 1 bis

Soit $T_n = \varphi_n\left(X_1, ..., X_n\right)$ un estimateur de $g(\theta)$.

On suppose que, pour tout $\theta \in \Theta$, T_n admet une espérance pour la probabilité $\mathbb{P}_\theta$, cette espérance étant notée $\mathbb{E}_\theta\left(T_n\right)$, ou parfois simplement $\mathbb{E}\left(T_n\right)$.

L'estimateur T_n de $g(\theta)$ est dit ***sans biais*** ou ***non biaisé*** si et seulement si $\forall\ \theta \in \Theta,\ \mathbb{E}_\theta\left(T_n\right) = g(\theta)$, *i.e.* ssi b_{T_n} est l'application nulle sur Θ.

Il est dit ***biaisé*** dans le cas contraire…

2_ Risque quadratique ou écart quadratique moyen (EQM)

Définition 2

Soit $T_n = \varphi_n\left(X_1, ..., X_n\right)$ un estimateur de $g(\theta)$. On suppose maintenant que, pour tout $\theta \in \Theta$, T_n admet un moment d'ordre 2 pour la probabilité $\mathbb{P}_\theta$.

- On appelle ***risque quadratique*** ou ***écart quadratique moyen*** de l'estimateur T_n de $g(\theta)$ l'application r_{T_n} définie par :

$$r_{T_n} : \left\{ \begin{array}{ccc} \Theta & \to & \mathbb{R} \\ \theta & \mapsto & r_{T_n}(\theta) = \mathbb{E}_\theta\left(\left(T_n - g(\theta)\right)^2\right) \end{array} \right. .$$

- On appelle ***risque quadratique en θ*** de l'estimateur T_n de $g(\theta)$ le ***réel*** $\mathbb{E}_\theta\left(\left(T_n - g(\theta)\right)^2\right)$, *même remarque que pour le biais…*

Remarque essentielle

On considère usuellement qu'un estimateur est « ***meilleur*** » qu'un autre lorsque son **risque quadratique est plus faible**. En cas d'égalité, on départage par le biais.

Définition 3

Soit $T_n = \varphi_n\left(X_1, ..., X_n\right)$ un estimateur de $g(\theta)$. On suppose toujours que, pour tout $\theta \in \Theta$, T_n admet un moment d'ordre 2 pour la probabilité $\mathbb{P}_\theta$.

On appelle ***variance*** de l'estimateur T_n l'application $\mathbb{V}$ définie par :

$$\mathbb{V} : \left\{ \begin{array}{ccc} \Theta & \to & \mathbb{R} \\ \theta & \mapsto & \mathbb{V}_\theta\left(T_n\right) = \mathbb{E}_\theta\left(\left(T_n - \mathbb{E}_\theta\left(T_n\right)\right)^2\right) \end{array} \right.$$

Remarques

◦ L'application $\mathbb{P}_\theta$ est une probabilité, donc $\mathbb{E}_\theta$ est une espérance, avec toutes les propriétés qui vont bien, et $\mathbb{V}_\theta$ est une variance, avec toutes les propriétés qui vont bien aussi, en particulier la ***formule de König – Huygens***, qui assure que :

$$\mathbb{V}_\theta\left(T_n\right) = \mathbb{E}_\theta\left(T_n^2\right) - \left(\mathbb{E}_\theta\left(T_n\right)\right)^2 .$$

◦◦ Si $T_n = \varphi_n\left(X_1, ..., X_n\right)$ est un estimateur sans biais de $g(\theta)$, on a bien sûr clairement :

$$r_{T_n}(\theta) = \mathbb{V}_\theta\left(T_n\right) .$$

Mais on peut dire mieux de manière plus générale, et c'est l'objet de la proposition suivante.

Proposition

Soit $T_n = \varphi_n\left(X_1, ..., X_n\right)$ un estimateur de $g(\theta)$. On suppose que, pour tout $\theta \in \Theta$, T_n admet un moment d'ordre 2 pour la probabilité $\mathbb{P}_\theta$. Alors :

$$r_{T_n}(\theta) = \left(b_{T_n}(\theta)\right)^2 + \mathbb{V}_\theta\left(T_n\right),$$

i.e. le risque quadratique d'un estimateur est égal à la somme de sa variance et du carré de son biais.

On retrouve ainsi le fait que $r_{T_n}(\theta) = \mathbb{V}_\theta(T_n)$ *lorsque l'estimateur est non biaisé.*

C ' Qualités d'une suite d'estimateurs

Vous verrez que l'on considère en fait souvent en pratique une suite $(T_n)_{n\in\mathbb{N}}$ d'estimateurs de $g(\theta)$, chaque T_n étant donc une statistique $T_n = \varphi_n(X_1, ..., X_n)$.

1_ Non – biais asymptotique

Définition 2

Une suite $(T_n)_{n\in\mathbb{N}}$ d'estimateurs de $g(\theta)$ est dite ***asymptotiquement sans biais*** si et seulement si :

$$(\forall\ \theta \in \Theta),\ \boxed{\lim_{n\to+\infty} \mathbb{E}_\theta(T_n) = g(\theta)}.$$

On dit aussi que T_n est ***asymptotiquement sans biais*** par abus de langage.

2_ Consistance

On est amené à penser intuitivement que si la taille de l'échantillon augmente, l'information sur le paramètre θ, donc aussi l'estimateur qu'on en choisit, devrait se « rapprocher » de la vraie valeur θ.

La traduction mathématique se fait en termes de **convergence en probabilité** de la suite $(T_n)_{n\in\mathbb{N}}$.

Définition 3

Un(e) (suite d') estimateur(s) $(T_n)_{n\in\mathbb{N}}$ de $g(\theta)$ est dit(e) ***convergent(e)*** ou ***consistant(e)*** ssi $T_n \xrightarrow[n\to+\infty]{\mathbb{P}_\theta} g(\theta)$, *i.e.* si et seulement si : $(\forall\ \theta \in \Theta)$,

$$\forall\ \varepsilon > 0\ ,\ \lim_{n \to +\infty} \mathbb{P}_{\theta}\left(\left| T_n - g(\theta) \right| < \varepsilon \right) = 1$$

ou

$$\forall\ \varepsilon > 0\ ,\ \lim_{n \to +\infty} \mathbb{P}_{\theta}\left(\left| T_n - g(\theta) \right| \geq \varepsilon \right) = 0\ .$$

Proposition 1

Soit $\left(T_n \right)_{n \in \mathbb{N}}$ une suite d'estimateurs de $g(\theta)$, admettant tous une variance.

On suppose que :

- $\forall\ n \in \mathbb{N}$, T_n est non biaisé.
- $\lim_{n \to +\infty} \mathbb{V}\left(T_n - g(\theta) \right) = \lim_{n \to +\infty} \mathbb{V}\left(T_n \right) = 0\ .$

Alors la suite $\left(T_n \right)_{n \in \mathbb{N}}$ est consistante (ou convergente).

Proof C'est une conséquence directe de l'inégalité de Bienaymé – Cebycev :

Proposition 2 ***(hors programme, mais à connaître)***

Soit $\left(T_n \right)_{n \in \mathbb{N}}$ une suite d'estimateurs de $g(\theta)$, admettant tous une variance.

On suppose que :

- $\left(T_n \right)_{n \in \mathbb{N}}$ est asymptotiquement sans biais.
- $\lim_{n \to +\infty} \mathbb{V}\left(T_n - g(\theta) \right) = \lim_{n \to +\infty} \mathbb{V}\left(T_n \right) = 0\ .$

Alors la suite $\left(T_n \right)_{n \in \mathbb{N}}$ est consistante (ou convergente).

D Quelques mots sur l'échantillonnage

On distingue deux types d'échantillons :

les ***échantillons exhaustifs*** et les ***échantillons non exhaustifs*** .

Un échantillon est dit ***exhaustif*** lorsque chaque individu de l'échantillon ne peut être choisi qu'une fois. Il est dit ***non – exhaustif*** dans le cas contraire.

Remarques :

1_ Les échantillons *exhaustifs* jouent en Statistiques, le rôle des tirages sans remise en Probabilités. Les *non – exhaustifs* celui des tirages avec remise.

2_ L'indépendance des VAR $X_1, ..., X_n$ sera bien sûr assurée dans le cas d'échantillons non – exhaustifs, c'est dans ce cas que se situent les n – échantillons *i.i.d.*.

3_ En pratique, on assimilera les échantillons exhaustifs à des échantillons non – exhaustifs lorsque la population parente a un effectif N largement plus grand que l'effectif n de l'échantillon. Plus prosaïquement, lorsque $N > 10\,n$, en l'occurrence et en pratique…

Dernière remarque plus générale, enfin… L' *estimation* dépend bien évidemment de l'échantillon observé. Si l'on change d'échantillon, on obtient une autre estimation. On appelle ce phénomène la ***fluctuation d'échantillonnage*** .

E Estimation ponctuelle d'une espérance

On suppose ici que le paramètre θ étudié sur la population parente est la moyenne (l'espérance) d'une VAR X définie sur la population $\mathcal{P}$. Chaque individu $\omega_1, ..., \omega_n$ de l'échantillon donne une valeur observée $X(\omega_1), ..., X(\omega_n)$. Ces valeurs peuvent être considérées comme les

valeurs de n VAR $X_1, ..., X_n$ définies sur Ω. Si l'échantillon est non – exhaustif ou assimilé comme tel, les VAR $X_1, ..., X_n$ sont indépendantes, de même loi que X et constituent donc un n – échantillon *i.i.d.* de la loi de X. *En fait, c'est bien plus compliqué, il faudrait montrer qu'il existe bien un espace probabilisable sur lequel on pourrait définir n « copies » indépendantes de la VAR X, appelées en vérité « versions », mais cela est bien trop théorique pour que l'on puisse vous titiller avec ces considérations…*

Définition

On appelle ***moyenne empirique*** de X sur le n – échantillon la statistique $Y_n = \overline{X}_n$ définie par : $\boxed{Y_n = \overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k}$.

Proposition 1

Toujours dans la situation, et avec les notations précédentes…

- $Y_n = \overline{X}_n$ est un estimateur non biaisé de $\theta = \mathbb{E}(X)$.
- Si X admet une variance, $Y_n = \overline{X}_n$ est un estimateur consistant (convergent).

Proposition 1'

Si les lois μ_θ admettent une variance, la moyenne empirique est un estimateur sans biais et consistant de l'espérance de μ_θ.

Remarque

La ***moyenne empirique*** $Y_n = \overline{X}_n$ est un **estimateur** de $\theta = \mathbb{E}(X)$.

La ***moyenne observée*** (valeur de $Y_n = \overline{X}_n$ sur un échantillon de n individus donnés) est une **estimation** de $\theta = \mathbb{E}(X)$. Elle est en général notée X_n, ou $\boxed{X_n}$.

II *Estimation par intervalles de confiance*

Z *Introduction*

- L'estimation ponctuelle (précise) a un inconvénient majeur :

  l'échantillon n'est peut être pas représentatif, et la mesure faite peut donner, à cause de la fluctuation d'échantillonnage, une idée fausse de la valeur du paramètre θ.

- ● On préfère donc donner une ***fourchette*** de valeurs possibles du paramètre θ, fourchette où le paramètre θ est situé **avec une probabilité minimale** fixée *a priori* .

Dans toute la suite, U_n et V_n désignent deux statistiques, estimateurs de $g(\theta)$,

$$U_n = \gamma_n(X_1, ..., X_n) \text{ et } V_n = \delta_n(X_1, ..., X_n),$$

où $(X_1, ..., X_n)$ est un n – échantillon *i.i.d.* de la loi μ_θ .

On devrait noter dans la suite $U_n(\theta)$ et $V_n(\theta)$, mais on notera abusivement U_n et V_n afin de ne pas alourdir encore les écritures…

Définition

Soit $\alpha \in]0,1[$.

Le segment $[U_n, V_n]$ est appelé ***intervalle de confiance de $g(\theta)$ au niveau de confiance $1-\alpha$ (ou $(100-100\alpha)\%$)*** si pour tout $\theta \in \Theta$, on a :

$$\mathbb{P}_\theta(g(\theta) \in [U_n, V_n]) \geq 1-\alpha \quad \textit{i.e.} \quad \mathbb{P}_\theta(U_n \leq g(\theta) \leq V_n) \geq 1-\alpha .$$

Remarques

1_ α s'appelle le ***(seuil de) risque*** , et $1-\alpha$ le ***(seuil de) confiance*** .

2_ Si $\left(x_1, ..., x_n \right)$ est la valeur observée de $\left(X_1, ..., X_n \right)$ sur un échantillon, l'intervalle $\left[U_n\left(x_1, ..., x_n \right), V_n\left(x_1, ..., x_n \right) \right]$ est une ***estimation de l'intervalle de confiance.***

3_ Les intervalles de confiance permettent de contrôler une valeur expérimentale, nous voyons apparaître là les prémices de la théorie des tests :

si la valeur expérimentale est située dans l'intervalle de confiance, on conclut à la validité du modèle théorique (on ***accepte*** ce modèle), et sinon, on le ***rejette*** avec un risque inférieur à α.

Définition bis

Soit $\alpha \in]0,1[$.

Le segment $\left[U_n, V_n \right]$ est appelé ***intervalle de confiance de*** $g(\theta)$ ***au niveau de risque*** α ***(ou*** $100\,\alpha\,\%$***)*** si pour tout $\theta \in \Theta$, on a :

$$\boxed{\mathbb{P}_\theta\left(g(\theta) \in \left[U_n, V_n \right] \right) \geq 1-\alpha} \quad \textit{i.e.} \quad \boxed{\mathbb{P}_\theta\left(U_n \leq g(\theta) \leq V_n \right) \geq 1-\alpha}.$$

Définition ter

Soit $\alpha \in]0,1[$.

Le segment $\left[U_n, V_n \right]$ est appelé ***intervalle de confiance de*** $g(\theta)$ ***au niveau de risque*** α ***(ou*** $100\,\alpha\,\%$***)*** si pour tout $\theta \in \Theta$, on a :

$$\boxed{\mathbb{P}_\theta\left(g(\theta) \notin \left[U_n, V_n \right] \right) \leq \alpha}.$$

Grammaire statistique…

Les mots ***« niveau »*** *,* ***« seuil »*** *, et* ***« degré »*** *sont synonymes…*

A *Estimation par intervalles de l'espérance d'une loi normale d'écart – type connu, et d'une moyenne*

Théorème

Soit $\left(X_1, ..., X_n \right)$ un n – échantillon ***i.i.d.*** défini sur un espace probabilisé $\left(\Omega, \mathcal{T}, \mathbb{P} \right)$ de la loi normale $\mathcal{N}\left(m, \sigma^2 \right)$, **où σ^2 est supposé connu**.

Soit $\alpha \in \left[0, 1 \right]$, et t_α l'unique nombre réel tel que $\Phi\left(t_\alpha \right) = 1 - \frac{\alpha}{2}$. Soit enfin

$$Y_n = \frac{1}{n} \sum_{k=1}^{n} X_k$$

la moyenne empirique. Alors l'intervalle :

$$\left[Y_n - \frac{t_\alpha . \sigma}{\sqrt{n}}, Y_n + \frac{t_\alpha . \sigma}{\sqrt{n}} \right]$$

est un intervalle de confiance pour m au seuil de confiance $1 - \alpha$.

Théorème bis ***(version estimation)***

Soit m l'espérance (la moyenne) d'une loi normale **d'écart – type σ connu**.

Soit $\alpha \in \left[0, 1 \right]$, et t_α le nombre réel tel que $\Phi\left(t_\alpha \right) = 1 - \frac{\alpha}{2}$.

Si $Z_n = \overline{X_n}$ désigne la moyenne observée sur un n – échantillon, alors l'intervalle

$\left[Z_n - \frac{t_\alpha . \sigma}{\sqrt{n}}, Z_n + \frac{t_\alpha . \sigma}{\sqrt{n}} \right]$ est une estimation de l'intervalle de confiance pour m au seuil de confiance $1 - \alpha$.

Mini – bestiaire numérique

A titre d'exemples, citons les très classiques valeurs suivantes :

- Si $\alpha = 0{,}1$, alors $t_\alpha \simeq 1{,}645$.
- Si $\alpha = 0{,}05$, alors $t_\alpha \simeq 1{,}96$.
- Si $\alpha = 0{,}01$, alors $t_\alpha \simeq 2{,}325$.

B *Estimation par intervalles du paramètre d'une variable de Bernoulli & d'une proportion (fréquence)*

Théorème

Soit $\left(X_1, ..., X_n \right)$ un n – échantillon ***i.i.d.*** de la loi de Bernoulli de paramètre p .

Soit $\alpha \in \left[0, 1 \right]$, et t_α l'unique nombre réel tel que $\Phi\left(t_\alpha \right) = 1 - \frac{\alpha}{2}$. Soit enfin

$$F_n = \frac{1}{n} \sum_{k=1}^{n} X_k = \frac{S_n}{n}$$

la fréquence empirique. Alors l'intervalle :

$$\left[F_n - \frac{t_\alpha}{2\sqrt{n}}, F_n + \frac{t_\alpha}{2\sqrt{n}} \right]$$

est un intervalle de confiance pour p au seuil de confiance $1 - \alpha$.

Preuve

Par indépendance des $\left(X_k \right)_{k \in [\![1, n]\!]}$, S_n suit la loi binomiale $\mathcal{B}\left(n, p \right)$. Par suite :

$$\mathbb{E}\left(F_n \right) = p, \text{ et } \mathbb{V}\left(F_n \right) = \frac{p\left(1 - p \right)}{n} = \frac{p\,q}{n}, \text{ d'où } \sigma\left(F_n \right) = \sqrt{\frac{p\,q}{n}} .$$

Soit $F_n^* = \dfrac{F_n - p}{\sqrt{\dfrac{p\,q}{n}}} = \dfrac{\sqrt{n}\left(F_n - p \right)}{\sqrt{p\,q}}$.

Le théorème central limite assure que si n est grand ($n > 30$ en pratique), on peut approcher la

loi de la VAR F_n^* par la loi normale centrée – réduite. Ainsi, pour tout réel positif t :

$$\mathbb{P}\left(\left| F_n^* \right| \leq t \right) \simeq \Phi(t) - \Phi(-t) = \Phi(t) - \left(1 - \Phi(t)\right) = 2\,\Phi(t) - 1.$$

Soit t_α l'unique réel tel que $2\,\Phi\left(t_\alpha\right) - 1 = 1 - \alpha$, *i.e.* $t_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$. On a alors :

$$\mathbb{P}\left(\left| F_n^* \right| \leq t_\alpha \right) = 1 - \alpha.$$

On vient de construire un intervalle de dispersion pour la loi normale centrée-réduite, au seuil de confiance $1 - \alpha$.

On a alors :

$$\mathbb{P}\left(\left| F_n^* \right| \leq t_\alpha \right) \simeq 1 - \alpha, \text{ i.e. } \mathbb{P}\left(\left| F_n - p \right| \leq \frac{t_\alpha \cdot \sqrt{p\,q}}{\sqrt{n}} \right) \simeq 1 - \alpha.$$

Malheureusement, on cherche à estimer $\theta = p$, qui n'est donc pas connu, et $\sqrt{p\,q}$ non plus.

Mais, *heureusement*, on sait que :

$$\forall\ p \in \left]0, 1\right[,\ p(1-p) \leq \frac{1}{4}, \text{ donc } \sqrt{p\,q} \leq \frac{1}{2}.$$

Par suite :

$$\left[\left| F_n - p \right| \leq \frac{t_\alpha \cdot \sqrt{p\,q}}{\sqrt{n}} \right] \subset \left[\left| F_n - p \right| \leq \frac{t_\alpha}{2\sqrt{n}} \right],$$

d'où, par isotonie de la probabilité $\mathbb{P}$:

$$\mathbb{P}\left(\left| F_n - p \right| \leq \frac{t_\alpha \cdot \sqrt{p\,q}}{\sqrt{n}} \right) \leq \mathbb{P}\left(\left| F_n - p \right| \leq \frac{t_\alpha}{2\sqrt{n}} \right).$$

Finalement :

$$\mathbb{P}\left(\left| F_n - p \right| \leq \frac{t_\alpha}{2\sqrt{n}} \right) \geq 1 - \alpha.$$

C'est dire très exactement que l'intervalle $\left[F_n - \frac{t_\alpha}{2\sqrt{n}},\ F_n + \frac{t_\alpha}{2\sqrt{n}} \right]$ est un intervalle de confiance pour p au seuil de confiance (supérieur à) $1 - \alpha$. □

<u>Théorème bis</u> ***(version estimation)***

Soit p la proportion d'individus d'une population $\mathcal{P}$ qui possède la propriété Q.

Soit $\alpha \in \left[0, 1\right]$, et t_α le nombre réel tel que $\Phi\left(t_\alpha\right) = 1 - \frac{\alpha}{2}$.

Si F_n désigne la fréquence observée sur le n – échantillon donné de $\mathcal{P}$, alors l'intervalle $\left[F_n - \frac{t_\alpha}{2\sqrt{n}} , F_n + \frac{t_\alpha}{2\sqrt{n}} \right]$ est une estimation de l'intervalle de confiance pour p au seuil de confiance $1 - \alpha$.

On considère une population $\mathcal{P}$, et une propriété logique Q.

On désigne par $p \in]0, 1[$ la proportion d'individus de la population $\mathcal{P}$ vérifiant la propriété Q.

On se donne un n – échantillon $\left(\omega_1, ..., \omega_n \right)$ de la population $\mathcal{P}$, non – exhaustif ou assimilé.

Pour tout entier $k \in \llbracket 1, n \rrbracket$, soit X_k la VAR indicatrice (de Bernoulli) égale à 1 si l'individu ω_k vérifie Q, et égale à 0 sinon. On pose $F_n = \frac{1}{n} \sum_{k=1}^{n} X_k = \frac{S_n}{n}$ (***fréquence empirique***).

Pour tout entier $k \in \llbracket 1 , n \rrbracket$, on a déjà vu que la VAR X_k suit la loi de Bernoulli de paramètre p.