14

Statistiques

Estimations ponctuelle et par intervalle de confiance

Introduction...

Dans le chapitre 7, nous avons fourni des outils d'analyse de données numériques : moyenne, écart-type, médiane, quartiles, tableaux, graphiques... On se plaçait alors avec un certain nombre de données, que l'on analysait dans le but d'en extraire les informations essentielles : on parle de **statistique descriptive**.
Dans le chapitre qui suit, nous allons tenter, à partir de données obtenues sur un échantillon, dans déduire des informations sur la population globale. On parle alors de **statistique inférentielle**. La statistique inférentielle est la statistique des sondages : on interroge un échantillon réduit de la population, et on tente d'estimer alors la tendance générale sur population entière. L'enjeu ici est de préciser la qualité des estimations proposées.

Pour bien démarrer...

1 # Rappeler la loi faible des grands nombres.

2 # Rappeler l'inégalité de Bienaymé-Tchebychev.

3 # Rappeler le théorème central limite.

I Introduction

Plaçons nous dans la peau d'enquêteurs souhaitant réaliser un sondage sur l'avis des français concernant la qualité de l'enseignement privé par rapport à l'enseignement public.
On interroge alors des personnes, choisies de façon la plus aléatoire possible et en veillant à ce que cet échantillon soit le plus représentatif possible de la population française. A la question "Pensez-vous que l'on suit une meilleure scolarité dans un établissement privé que dans un établissement public ?", le sondé répond "oui" ou "non".
On attribue à "oui" la valeur 1 et à "non" la valeur 0.

Petite remarque
Le sondage permet de n'enquêter que sur un groupe restreint d'individus, par manque de temps ou de moyen pour interroger la population entière. Sinon, on organise un référendum.

On note θ la proportion de français votant "oui" et l'on considère donc une variable aléatoire X sur un certain espace probabilisable, telle que $X \hookrightarrow \mathscr{B}(\theta)$. **L'objectif est alors d'estimer au mieux la valeur de θ.**

Pour cela, revenons à notre sondage...
On considère une suite de variables aléatoires $(X_k)_{k\in\mathbb{N}^*}$, toutes de même loi que X, telles que : pour tout $k \in \mathbb{N}^*$, X_k prend la valeur 1 si la k-ième personne interrogée répond "oui" à la question, et 0 sinon.
Pour tout $k \in \mathbb{N}^*$, on notera x_k la réalisation de la variable aléatoire X_k dans le sondage effectué.
Le caractère aléatoire ne repose pas dans la réponse de l'individu choisi, mais bien dans le fait que l'individu fasse partie du panel de sondés ou non. Bien évidemment, on comprend le risque de la méthode : choisir un échantillon qui ne soit pas suffisamment représentatif de la population et qui ne nous fournit donc pas une bonne estimation de l'avis général de la population.
L'objectif est de pouvoir conclure un sondage par une phrase du type :

Petite remarque
Majuscules pour les variables aléatoires, minuscules pour les réalisations.

Trivialement...
On imagine bien que la taille de l'échantillon interrogé a un impact assez fort sur la qualité de la prévision faite.

"la valeur observée sur l'échantillon est *proche* de la valeur théorique de θ avec *tel niveau de confiance* (ou avec telle probabilité)"

L'enjeu mathématique est alors de préciser le "*proche*" et le "*tel niveau de confiance*" de la phrase précédente.

On considère un phénomène aléatoire, modélisé par une variable aléatoire X définie sur un espace probabilisable $(\Omega, \mathcal{A})$, dont la loi appartient à une famille de lois bien précise, dépendant d'un ou deux paramètres réels que l'on cherche à déterminer.

Exemples 1

- **E1** On modélise le phénomène par une variable aléatoire X suivant une loi de Bernoulli, comme dans l'introduction ci-dessus, mais on ne connaît pas le paramètre de cette loi, que l'on note θ et qui est à chercher dans l'ensemble $\Theta = [0; 1]$.
- **E2** On modélise la durée de vie d'une ampoule par une variable aléatoire X suivant une loi exponentielle, dont on ne connaît pas le paramètre, noté θ, qui est à chercher dans l'ensemble $\Theta = \mathbb{R}_*^+$.
- **E3** On modélise le poids des nouveaux-nés par une variable aléatoire X suivant une loi normale, dont on ne connaît pas le couple des paramètres, noté $\theta = (\theta_1, \theta_2)$, qui est à chercher dans l'ensemble $\Theta = \mathbb{R} \times \mathbb{R}_*^+$.

En fait...
En pratique, on commence par les outils de statistique descriptive pour analyser les données et essayer d'imaginer la famille de lois modélisant la situation : normale, exponentielle, Poisson, log-normale, Pareto...
Après avoir fixé une famille de lois, on cherche le ou les paramètres de cette dernière par les méthodes de statistique inférentielle.

On pourra proposer une unique valeur possible de ce paramètre, on parle alors d'**estimation ponctuelle**, ou bien tout un intervalle de valeurs possibles, on parle alors d'**estimation par intervalle de confiance**.

> Dans tout le chapitre, n désignera un entier naturel non nul. On considèrera également une variable aléatoire X définie sur un espace probabilisable $(\Omega, \mathcal{A})$ dont la loi dépend d'un paramètre θ réel prenant ses valeurs dans un ensemble Θ.
> On munit l'espace probabilisable d'une famille de probabilités $\left(\mathbb{P}_\theta\right)_{\theta\in\Theta}$.

Petite remarque
Si la loi dépend de deux paramètres inconnus, on cherchera toujours à estimer l'un puis l'autre... On peut donc considérer dans la suite du cours, que la loi ne dépend que d'un seul paramètre à déterminer.

II Estimation ponctuelle

II.1 Échantillon et estimateur

Définition 1 - Échantillon

Un **n-échantillon de X** est un n-uplet $(X_1, \ldots, X_n)$ de variables aléatoires définies sur $(\Omega, \mathcal{A})$ indépendantes et de même loi que X pour toutes les $\mathbb{P}_\theta$.

Définitions 2 - Estimateur, estimation

Soit $(X_1, \ldots, X_n)$ un n-échantillon de X.

D1# Un **estimateur de θ** est une variable aléatoire sur $(\Omega, \mathcal{A})$ fonction des variables aléatoires $X_1, \ldots, X_n$ et dont l'expression ne fait pas mention de θ.
Autrement dit, une variable aléatoire T_n sur $(\Omega, \mathcal{A})$ est un estimateur de θ lorsqu'il existe une fonction φ définie sur $X_1(\Omega) \times \ldots \times X_n(\Omega)$ et à valeurs dans $\mathbb{R}$ telle que $T_n = \varphi(X_1, \ldots, X_n)$.

D2# Soit $\varphi(X_1, \ldots, X_n)$ est un estimateur de θ. Soit $(x_1, \ldots, x_n) \in X_1(\Omega) \times \ldots \times X_n(\Omega)$. On dit que $\varphi(x_1, \ldots, x_n)$ est une **estimation ponctuelle de θ**.
Autrement dit, une estimation ponctuelle de θ est une réalisation d'un estimateur de θ.

Important !
Un estimateur de θ dépend naturellement de θ (puisqu'il est fonction des X_k dont la loi dépend de θ) ; **mais son expression ne doit pas faire intervenir θ**.

Notation
Si existence, on notera $\mathbb{E}_\theta(T_n)$ et $\mathbb{V}_\theta(T_n)$ l'espérance et la variance de T_n pour $\mathbb{P}_\theta$.

Exemple 2

Soit X une variable aléatoire suivant la loi $\mathscr{B}(\theta)$, avec $\theta \in]0;1[$. On considère $X_1, \ldots, X_n$ des variables aléatoires indépendantes suivant toutes la même loi que X et on note : $T_n = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$.

Justifions que T_n est un estimateur de θ.
On sait que :

- $X_1, \ldots, X_n$ sont indépendantes et suivent toutes la même loi $\mathscr{B}(\theta)$,
- T_n est fonction de $X_1, \ldots, x_n$ dont l'expression ne fait pas apparaître θ.

Par conséquent : T_n est un estimateur de θ.

De même : $\displaystyle\prod_{k=1}^{n} X_k$, $\max(X_1, \ldots, X_n)$, $\min(X_1, \ldots, X_n)$ sont des estimateurs de θ.

En revanche, $\displaystyle\sum_{k=1}^{n} X_k - \theta$ n'est pas un estimateur de θ, car son expression fait intervenir θ.

Petite remarque
Ce n'est pas ce qui manque des estimateurs, il y en a une infinité : la définition est très générale. Mais lesquels sont des *bons* estimateurs ?

II.2 Outils de mesure de la qualité d'un estimateur (HP)

Dans ce paragraphe, nous introduisons des outils, tous devenus hors programme, qui permettent de mesurer la qualité d'un estimateur.
L'idée générale est que l'estimateur T_n soit proche de θ. Pour mesurer l'écart entre T_n et θ, on peut utiliser les quantités suivantes : $T_n - \theta$, $|T_n - \theta|$, $(T_n - \theta)^2$,...
Mais bien évidemment, ce qui nous intéresse le plus est l'écart moyen... D'où les définitions suivantes :

Définitions 3 - Biais, risque quadratique (HP)

Soit T_n un estimateur de θ.

D1# Si T_n admet une espérance, le **biais de l'estimateur T_n**, noté $b_\theta(T_n)$, est le réel défini par :

$$b_\theta(T_n) = \mathbb{E}_\theta(T_n - \theta)$$

D2# Si T_n admet une variance, le **risque quadratique de l'estimateur T_n**, noté $r_\theta(T_n)$, est le réel défini par :

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

Vocabulaire
- Si $b_\theta(T_n) = 0$, on dit que l'estimateur T_n est **sans biais**.
- Si $\lim\limits_{n \to +\infty} b_\theta(T_n) = 0$, on dit que l'estimateur T_n est **asymptotiquement sans biais**.

Propriétés 1 (HP)

Avec les notations précédentes :

P1# Si T_n admet une espérance, alors : $b_\theta(T_n) = \mathbb{E}_\theta(T_n) - \theta$.

P2# Si T_n admet une variance, alors : $r_\theta(T_n) = \mathbb{V}_\theta(T_n) + \left(b_\theta(T_n)\right)^2$.

Petite remarque
Si T_n est sans biais, alors $r_\theta(T_n) = \mathbb{V}_\theta(T_n)$.

⋆ Démonstration :

$\star$

Dans l'idéal, on cherche à obtenir un estimateur dont le biais et le risque quadratique sont les plus proches de 0 possible, l'idéal étant un estimateur sans biais et de variance minimale... Mais celui-ci n'existe pas toujours, ou n'est pas simple à trouver.
De façon générale, pour comparer deux estimateurs :

- s'ils ont même biais, on préférera celui de risque quadratique minimal,
- sinon, on pourra parfois préférer un estimateur biaisé à un estimateur sans biais si son risque quadratique est plus faible que la variance de l'estimateur sans biais.

Confusion d'objets !
On verra parfois la confusion entre estimateur et suite d'estimateurs.

Définition 4 - Estimateur convergent (HP)

Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs de θ.
On dit que $(T_n)_{n\in\mathbb{N}^*}$ est convergente (ou plus simplement que l'estimateur T_n est convergent) lorsque :

$$\forall\theta\in\Theta,\ \forall\varepsilon>0,\ \lim_{n\to+\infty}\mathbb{P}_\theta\big([|T_n-\theta|\geq\varepsilon]\big)=0$$

Autrement dit :
T_n est convergent ssi $T_n \xrightarrow[n\to+\infty]{\mathbb{P}} \theta$

Exemple 3

Soit T_n un estimateur de θ admettant une variance. Montrons que si $\lim\limits_{n\to+\infty} r_\theta(T_n)=0$, alors T_n est convergent.

À retenir...
Cas classique pour établir la convergence d'un estimateur.

Un estimateur sans biais garantit l'absence d'erreur en moyenne, mais peut produire de très mauvaises estimations ponctuelles. Au contraire, un estimateur convergent, biaisé ou non, sera d'autant plus fiable que la taille de l'échantillon est grande. En général, on préfère ces derniers.

II.3 Moyenne et variance empiriques

Dans cette partie, nous allons donner et étudier deux estimateurs usuels et pertinents, l'un pour l'espérance et l'autre pour la variance. Ces estimateurs peuvent ainsi être utilisés pour déterminer les paramètres des lois $\mathscr{B}(p)$, $\mathscr{P}(\lambda)$, $\mathscr{U}\left([\![1;n]\!]\right)$, $\mathscr{E}(\lambda)$, $\mathscr{U}\left([0;b]\right)$, $\mathscr{N}(\mu;\sigma^2)$... puisque chacune des lois fait intervenir espérance et/ou variance en paramètres !

Définition 5 - Moyenne empirique

Soient X une variable aléatoire admettant une espérance et $(X_1,\ldots,X_n)$ un n-échantillon de X.

La **moyenne empirique** de $X_1,\ldots,X_n$ est l'estimateur $\overline{X_n}=\dfrac{1}{n}\sum\limits_{k=1}^{n}X_k$.

Petite remarque
C'est l'estimateur le plus naturel et le plus usuel de l'espérance d'une variable aléatoire !

Propriété 2 - Qualité de la moyenne empirique (HP)

Soient X une variable aléatoire admettant une espérance et une variance ainsi que $(X_1, \dots, X_n)$ un n-échantillon de X.
La moyenne empirique de $X_1, \dots, X_n$ est un estimateur sans biais et convergent de $\mathbb{E}(X)$.

Démonstration :

Exemples 4

E1 Dans le cas du sondage initial, X suivait une loi de Bernoulli de paramètre p inconnu. Puisque $\mathbb{E}(X) = p$, on peut obtenir une estimation du paramètre p de la loi de Bernoulli en considérant la réalisation de la moyenne empirique sur un échantillon suffisamment grand.
On retrouve notre intuition initiale !
Pour estimer $\mathbb{V}(X) = p(1-p)$, on pourrait par exemple proposer $\overline{X_n}(1-\overline{X_n})$...

E2 On suppose maintenant que X suit un loi exponentielle de paramètre λ inconnu.
On peut utiliser la moyenne empirique d'un échantillon pour obtenir une estimation de $\mathbb{E}(X) = \dfrac{1}{\lambda}$, ce qui nous fournit ensuite une estimation de λ...

E3 On suppose maintenant que X suit un loi uniforme sur $[0; b]$, avec b inconnu.
On peut utiliser la moyenne empirique d'un échantillon pour obtenir une estimation de $\mathbb{E}(X) = \dfrac{b}{2}$, ce qui nous fournit ensuite une estimation de b...

Important !
Tout ce qui précède est à adapter dans le cas où l'on cherche à estimer $g(\theta)$, et pas θ (où g est une fonction définie sur Θ).
De façon générale, si $\mathbb{E}(X) = f(\theta)$, où f est bijective sur Θ, alors une estimation de θ sera fournie par une réalisation de $f^{-1}(\overline{X_n})$.

Définition 6 - Variance empirique

Soient X une variable aléatoire admettant une espérance et une variance, ainsi que $(X_1, \dots, X_n)$ un n-échantillon de X.
La **variance empirique** de $X_1, \dots, X_n$ est l'estimateur $V_n = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(X_k - \overline{X_n})^2$.

Exemple 5

Déterminons $\mathbb{E}(V_n)$. Donnons, à partir de V_n, un estimateur V_n' qui soit un estimateur sans biais de $\mathbb{V}(X)$.

☞ Pour info...
De nombreux logiciels informatiques (et des calculatrices) utilisent l'estimateur V_n' (**variance empirique corrigée**) pour la variance plutôt que V_n.

Propriété 3 - Qualité de la variance empirique corrigée (HP)

Soient X une variable aléatoire admettant un moment d'ordre 4 ainsi que $(X_1, \dots, X_n)$ un n-échantillon de X. La variance empirique corrigée de $X_1, \dots, X_n$ est un estimateur sans biais et convergent de $\mathbb{V}(X)$.

⋆ Démonstration :

- Par construction $\mathbb{E}(V'_n) = \mathbb{V}(X)$.
- Et on a :

$$r_\theta(V'_n) = \mathbb{V}(V'_n) = \dots = \frac{1}{n}\mathbb{E}\left((X - \mathbb{E}(X))^4\right) - \frac{n-3}{n(n-1)}\mathbb{V}(X)^2$$

Ainsi : $\lim\limits_{n \to +\infty} r_\theta(V'_n) = 0$ et on conclut sur la convergence en utilisant l'inégalité de Markov et le théorème d'encadrement... ⋆

Petite remarque

Cet estimateur permettrait d'obtenir une estimation du second paramètre d'une loi normale par exemple.

II.4 Estimateur du maximum de vraisemblance

On considère toujours une variable aléatoire suivant une certaine loi, dont un paramètre est inconnu, et un n-échantillon $(X_1, \dots, X_n)$ de X ayant fourni un n-uplet de réalisations $(x_1, \dots, x_n)$.
Une idée assez naturelle consiste à considérer que la valeur de θ qui a permis de générer les observations est celle qui avait la plus grande probabilité de les générer.

Exemple 6

Commençons par un cas simple où $n = 1$. On considère une variable aléatoire X suivant la loi $\mathscr{B}(15, p)$, avec p inconnu. On observe la valeur $x = 5$.
Question : quelle valeur de p maximise $\mathbb{P}([X = 5])$?

Définitions 7 - Vraisemblance

Soit X une variable aléatoire suivant une loi dépendant d'un certain paramètre θ inconnu et $(X_1 \dots, X_n)$ un n-échantillon de X. Soit $(x_1, \dots, x_n) \in X_1(\Omega) \times \dots \times X_n(\Omega)$ un n-uplet d'observations.

D1# Si X est discrète, on appelle **fonction de vraisemblance** pour $(x_1, \dots, x_n)$ la fonction :

$$L_n : \theta \longmapsto \prod_{k=1}^{n} \mathbb{P}_\theta([X_k = x_k])$$

D2# Si X est à densité, de densité f_θ, on appelle **fonction de vraisemblance** pour $(x_1, \dots, x_n)$ la fonction :

$$L_n : \theta \longmapsto \prod_{k=1}^{n} f_\theta(x_k)$$

Important !

Puisqu'il plus commode de dériver une somme qu'un produit (surtout pour les fonctions de deux variables), on travaille souvent sur la **log-vraisemblance**, égale à $\ln \circ L_n$.
Les fonctions vraisemblance et log-vraisemblance ont mêmes variations et mêmes extrema...

L'idée est alors d'attribuer à θ la valeur (si elle existe et si elle est unique) du réel θ_n^* qui maximise la fonction L_n (ce qui équivaut à maximiser la fonction $\ln \circ L_n$).

Définitions 8 - Maximum de vraisemblance

Soit X une variable aléatoire suivant une loi dépendant d'un certain paramètre θ inconnu et $(X_1 \ldots, X_n)$ un n-échantillon de X. Soit $(x_1, \ldots, x_n) \in X_1(\Omega) \times \ldots \times X_n(\Omega)$ un n-uplet d'observations. On note L_n la fonction de vraisemblance associée et on suppose qu'elle admet un maximum sur Θ, atteint en un unique réel.

D1# L'**estimation du maximum de vraisemblance** de θ est le réel θ_n^* défini par : $\theta_n^* = \max\limits_{\theta \in \Theta} \left(L_n(\theta) \right)$.

D2# Si g est la fonction telle que $\theta_n^* = g(x_1, \ldots, x_n)$, alors l'**estimateur du maximum de vraisemblance** de θ est la variable aléatoire $T_n = g(X_1, \ldots, X_n)$.

Sympa !
Le cadre du cours était celui d'un unique paramètre réel θ. La fonction L_n est ainsi une fonction d'une variable.
En revanche, si l'on cherche les deux paramètres d'une loi, L_n sera une fonction de deux variables... Vous voyez les liens ?!

Exemple 7

Déterminons l'estimateur du maximum de vraisemblance du paramètre p d'une loi de Bernoulli.
Pour cela, considérons une variable aléatoire X suivant la loi $\mathscr{B}(p)$, avec $p \in]0;1[$ inconnu. Considérons également, pour tout $n \in \mathbb{N}^*$, $(X_1, \ldots, X_n)$ un n-échantillon de X et $(x_1, \ldots, x_n)$ une réalisation de cet échantillon.
Soit $n \in \mathbb{N}^*$. On a :

$$\forall k \in [\![1;n]\!],\ \forall x_k \in \{0;1\},\ \mathbb{P}([X_k = x_k]) = \left\{ \begin{array}{ll} p & \text{si } x_k = 1 \\ 1-p & \text{si } x_k = 0 \end{array} \right. = p^{x_k}(1-p)^{1-x_k}$$

Ainsi, en notant L_n la fonction de vraisemblance pour $(x_1, \ldots, x_n)$ et $\varphi_n(p) = \ln\left(L_n(p)\right)$, on a, pour tout $p \in]0;1[$:

$$\begin{array}{rcl} L_n(p) & = & \prod\limits_{k=1}^{n} \mathbb{P}([X_k = x_k]) \\ & = & \prod\limits_{k=1}^{n} p^{x_k}(1-p)^{1-x_k} \end{array}$$

Pénible !
On utilise souvent la même lettre pour le paramètre à déterminer et pour la variable de la fonction de vraisemblance...

D'où :

$$\forall p \in]0;1[,\ L_n(p) > 0$$

et, pour tout $p \in]0;1[$:

$$\begin{array}{rcll} \varphi_n(p) & = & \ln\left(L_n(p)\right) & \\ & = & \ln\left(\sum\limits_{k=1}^{n} p^{x_k}(1-p)^{1-x_k}\right) & \forall k \in [\![1;n]\!],\ p^{x_k}(1-p)^{1-x_k} > 0 \\ & = & \sum\limits_{k=1}^{n} \ln\left(p^{x_k}(1-p)^{1-x_k}\right) & \forall k \in [\![1;n]\!],\ p > 0,\ 1-p > 0 \\ & = & \sum\limits_{k=1}^{n} \left(x_k \ln(p) + (1-x_k)\ln(1-p)\right) & \\ & = & \left(\sum\limits_{k=1}^{n} x_k\right)\ln(p) + \left(n - \sum\limits_{k=1}^{n} x_k\right)\ln(1-p) & \text{en notant } s_n = \sum\limits_{k=1}^{n} x_k \\ & = & s_n \ln(p) + (n-s_n)\ln(1-p) & \end{array}$$

La fonction φ_n est donc dérivable sur $]0;1[$ et, pour tout $p \in]0;1[$:

$$\begin{array}{rcl} \varphi_n'(p) & = & s_n \dfrac{1}{p} - (n-s_n)\dfrac{1}{1-p} \\ & = & \dfrac{s_n(1-p) - (n-s_n)p}{p(1-p)} \\ & = & \dfrac{s_n - np}{p(1-p)} \end{array}$$

D'où :

p	0		$\frac{s_n}{n}$		1
$\varphi_n'(p)$	||	$+$	0	$-$	||
φ_n	||	$\nearrow$	$\ldots$	$\searrow$	||

La fonction φ_n admet donc un maximum en $\dfrac{s_n}{n} = \dfrac{1}{n}\sum\limits_{k=1}^{n} x_k$.

Or, par stricte croissance de ln sur $\mathbb{R}_*^+$, la fonction φ_n et L_n ont les mêmes variations sur $]0;1[$.

Par conséquent : la fonction L_n admet un maximum en $\dfrac{1}{n}\sum\limits_{k=1}^{n} x_k$.

Conclusion : l'estimateur du maximum de vraisemblance de p est $\dfrac{1}{n}\sum\limits_{k=1}^{n} X_k = \overline{X_n}$.

Petite remarque
On retrouve la moyenne empirique... Tout ça pour ça ?!

L'estimateur du maximum de vraisemblance et la moyenne empirique (ou la variance empirique) peuvent parfois coïncider... De façon générale, ce n'est pas le cas.

☞ Pour info...
L'estimateur du maximum de vraisemblance est toujours convergent, mais pas nécessairement sans biais.

III Estimation par intervalle de confiance

III.1 Intervalle de confiance (exact)

Définition 9 – Intervalle de confiance (exact)

Soit X une variable aléatoire suivant une loi dépendant d'un certain paramètre θ inconnu. Soient $(U_n)_{n\in\mathbb{N}^*}$ et $(V_n)_{n\in\mathbb{N}^*}$ deux suites d'estimateurs de θ telles que : $\forall n \in \mathbb{N}^*,\ \mathbb{P}_\theta([U_n \leq V_n]) = 1$.
Soit $\alpha \in]0;1[$. On dit que **$[U_n, V_n]$ est un intervalle de confiance de θ au niveau de confiance $1-\alpha$ (ou au risque α)**, lorsque :

$$\mathbb{P}_\theta\big(\left[\theta \in [U_n, V_n]\right]\big) \geq 1-\alpha$$

✘ Attention !
θ n'est pas une variable aléatoire ! En revanche, $[U_n, V_n]$ est un intervalle aléatoire !
On garde en tête :
$[\theta \in [U_n, V_n]]{=}[U_n \leq \theta]\cap[V_n \geq \theta]$
et on voit mieux, sous cette forme, ce qui est aléatoire et ce qui ne l'est pas !

Si u_n et v_n sont des réalisations de U_n et V_n respectivement, il est faux de dire "θ est compris entre u_n et v_n avec une probabilité au moins égale à $1-\alpha$".
En effet, rien d'aléatoire dans θ, u_n et v_n.
On dira : "on a une confiance de $1-\alpha$ dans le fait que θ soit compris entre u_n et v_n".

Exemple 8

On considère une variable aléatoire X suivant une loi $\mathcal{N}(\mu;\sigma^2)$, où m est inconnu et σ^2 connu non nul. On dispose d'un n-échantillon $(X_1,\dots,X_n)$ de X.

1. Donnons la loi de $\overline{X_n}$ et celle de $\overline{X_n}^* = \sqrt{n}\dfrac{\overline{X_n}-m}{\sigma}$.
2. Soit $\alpha \in]0;1[$. Justifions l'existence d'un unique réel t_α tel que $\Phi(t_\alpha) = 1-\dfrac{\alpha}{2}$, où Φ désigne la fonction de répartition d'une variable aléatoire suivant la loi $\mathcal{N}(0;1)$.
3. Déduisons-en que $\left[\overline{X_n} - t_\alpha\dfrac{\sigma}{\sqrt{n}}, \overline{X_n} + t_\alpha\dfrac{\sigma}{\sqrt{n}}\right]$ est un intervalle de confiance de m au niveau de confiance $1-\alpha$.

☞ Rappel...
Si $X_1,\dots,X_n$ sont indépendantes telles que pour tout $k \in [\![1;n]\!]$, $X_k \hookrightarrow \mathcal{N}(m_k,\sigma_k^2)$, alors :

$$\sum_{k=1}^{n} a_k X_k \hookrightarrow \dots\dots\dots\dots\dots$$

Plus l'intervalle de confiance est réduit, plus le risque augmente ; alors que plus l'intervalle de confiance a une grande amplitude, plus le risque est faible. En pratique, il faudra trouver un compromis entre les deux.

☞ Pour info...
En pratique, un risque de $0,05$ est acceptable... on cherchera donc souvent un intervalle de confiance de niveau de confiance $0,95$ (95%).

♣ Méthode 1 ♣ Recherche d'un intervalle de confiance exact pour $\mathbb{E}(X)$ avec l'inégalité de Bienaymé-Tchebychev.

X admet une espérance m inconnue et une variance σ^2. **On considère T_n un estimateur de m tel que $\mathbb{E}(T_n) = m$** (estimateur sans biais). Résumons les étapes de l'obtention de l'intervalle de confiance :

1. Inégalité de Bienaymé-Tchebychev :
$$\forall \varepsilon > 0,\ \mathbb{P}\big(|T_n - m| \geq \varepsilon\big) \leq \frac{\mathbb{V}(T_n)}{\varepsilon^2}$$
2. Si $\mathbb{V}(T_n)$ dépend de m, on majore $\mathbb{V}(T_n)$ par v_n, indépendant de m (sinon, on prend $v_n = \mathbb{V}(T_n)$) et ainsi :
$$\forall \varepsilon > 0,\ \mathbb{P}\big(|T_n - m| \geq \varepsilon\big) \leq \frac{v_n}{\varepsilon^2}$$
3. Par passage à l'évènement contraire, on obtient :
$$\forall \varepsilon > 0,\ \mathbb{P}\big(|T_n - m| < \varepsilon\big) \geq 1 - \frac{v_n}{\varepsilon^2}$$
Autrement dit :
$$\forall \varepsilon > 0,\ \mathbb{P}\big([T_n - \varepsilon < m < T_n + \varepsilon]\big) \geq 1 - \frac{v_n}{\varepsilon^2}$$
4. Et puisque $[T_n - \varepsilon < m < T_n + \varepsilon] \subset [T_n - \varepsilon \leq m \leq T_n + \varepsilon]$, on obtient :
$$\forall \varepsilon > 0,\ \mathbb{P}\big([m \in [T_n - \varepsilon ; T_n + \varepsilon]]\big) \geq 1 - \frac{v_n}{\varepsilon^2}$$
Reste à choisir ε tel que, pour le α donné, on ait $\frac{v_n}{\varepsilon^2} = \alpha$... et ainsi :
$$\mathbb{P}\big([m \in [T_n - \varepsilon ; T_n + \varepsilon]]\big) \geq 1 - \alpha$$

★ Classique ! ★
C'est très classique (et le seul cas qui semble être au programme pour cette partie du cours).
On prendra souvent (toujours ?) la moyenne empirique comme estimateur de la moyenne.

✘ Attention !
L'intervalle de confiance est donné par deux estimateurs, qui ne peuvent pas dépendre du paramètre m recherché ! Autrement dit, l'expression des bornes de l'intervalle de confiance ne doit pas faire apparaître m !

Petite remarque
L'intervalle de confiance proposé est centré en T_n, d'amplitude 2ε.

Exemple 9

On considère une variable aléatoire suivant une loi de Bernoulli de paramètre p inconnu ainsi qu'un n échantillon $(X_1, \ldots, X_n)$ de X. On note $\overline{X_n}$ la moyenne empirique de cet échantillon.

1. Montrons : $\forall \varepsilon > 0,\ \mathbb{P}\big([\overline{X_n} - \varepsilon \leq p \leq \overline{X_n} + \varepsilon]\big) \geq 1 - \dfrac{1}{4n\varepsilon^2}$.
2. Déduisons-en que $\left[\overline{X_n} - \sqrt{\dfrac{5}{n}}, \overline{X_n} + \sqrt{\dfrac{5}{n}}\right]$ est un intervalle de confiance de p au niveau de confiance 95%.
3. Au second tour d'une élection présidentielle, les citoyens ont le choix entre deux candidats A et B. Un institut de sondage réalise un sondage auprès de 1000 personnes, dont 570 affirment vouloir voter pour le candidat A.
Peut-on affirmer, avec un risque d'erreur de 5% que la candidat A sera élu ?

Petite remarque
• En se fixant un risque $\alpha = \frac{1}{4n\varepsilon^2}$, plus ε diminue et plus n doit être grand pour conserver le même risque.
• En prenant $\alpha = \frac{1}{4n\varepsilon^2}$, on voit aussi que, à n fixé, plus ε est petit et plus α est grand... et plus ε est grand, plus α est petit.

III.2 Intervalle de confiance asymptotique

Définition 10 – Intervalle de confiance asymptotique

Soit X une variable aléatoire suivant une loi dépendant d'un certain paramètre θ inconnu. Soient $(U_n)_{n\in\mathbb{N}^*}$ et $(V_n)_{n\in\mathbb{N}^*}$ deux suites d'estimateurs de θ telles que : $\forall n \in \mathbb{N}^*,\ \mathbb{P}_\theta([U_n \leq V_n]) = 1$.
Soit $\alpha \in]0;1[$. On dit que $[U_n, V_n]$ **est un intervalle de confiance asymptotique de θ au niveau de confiance $1-\alpha$ (ou au risque α)**, lorsque :

$$\lim_{n\to+\infty} \mathbb{P}_\theta\big([\theta \in [U_n, V_n]]\big) \geq 1-\alpha$$

♣ **Méthode 2** ♣ **Recherche d'un intervalle de confiance asymptotique pour $\mathbb{E}(X)$ avec le TCL.**
X admet une espérance m inconnue et une variance σ^2 non nulle. **On travaille avec l'estimateur $\overline{X_n}$ de m** (estimateur sans biais). Résumons les étapes de l'obtention de l'intervalle de confiance asymptotique :

1. En posant $\overline{X_n}^* = \sqrt{n}\dfrac{\overline{X_n}-m}{\sigma}$, on a, par le TCL :
$$\overline{X_n}^* \xrightarrow[n\to+\infty]{\mathscr{L}} Z$$
où $Z \hookrightarrow \mathscr{N}(0;1)$.
2. Ainsi :
$$\forall x \geq 0,\ \lim_{n\to+\infty} \mathbb{P}([-x \leq \overline{X_n}^* \leq x]) = \mathbb{P}([-x \leq Z \leq x]) = 2\Phi(x)-1$$
3. Or :
$$\forall x \geq 0,\ \mathbb{P}([-x \leq \overline{X_n}^* \leq x]) = \ldots = \mathbb{P}\left(\left[\overline{X_n} - \frac{\sigma x}{\sqrt{n}} \leq m \leq \overline{X_n} + \frac{\sigma x}{\sqrt{n}}\right]\right)$$
D'où :
$$\lim_{n\to+\infty} \mathbb{P}\left(\left[\overline{X_n} - \frac{\sigma x}{\sqrt{n}} \leq m \leq \overline{X_n} + \frac{\sigma x}{\sqrt{n}}\right]\right) = 2\Phi(x)-1$$
4. Si σ dépend de m, on majore σ par s, indépendant de m (sinon, on prend $s=\sigma$) et ainsi :
$$\left[\overline{X_n} - \frac{\sigma x}{\sqrt{n}} \leq m \leq \overline{X_n} + \frac{\sigma x}{\sqrt{n}}\right] \subset \left[\overline{X_n} - \frac{s x}{\sqrt{n}} \leq m \leq \overline{X_n} + \frac{s x}{\sqrt{n}}\right]$$
D'où, par croissance de $\mathcal{P}$:
$$\lim_{n\to+\infty} \mathbb{P}\left(\left[\overline{X_n} - \frac{s x}{\sqrt{n}} \leq m \leq \overline{X_n} + \frac{s x}{\sqrt{n}}\right]\right) \geq 2\Phi(x)-1$$
5. Reste à choisir x tel que, pour le α donné, on ait $2\Phi(x)-1 = 1-\alpha$.
Mais :
$$\begin{aligned} 2\Phi(x)-1 = 1-\alpha &\iff \Phi(x) = 1-\frac{\alpha}{2} \\ &\iff x = \Phi^{-1}\left(1-\frac{\alpha}{2}\right) \end{aligned}$$
Φ est bijective de $\mathbb{R}$ dans $]0;1[$ et $1-\dfrac{\alpha}{2} \in]0;1[$
En notant $t_\alpha = \Phi^{-1}\left(1-\dfrac{\alpha}{2}\right)$, on obtient :
$$\lim_{n\to+\infty} \mathbb{P}\left(\left[m \in \left[\overline{X_n} - t_\alpha\frac{s}{\sqrt{n}}, \overline{X_n} + t_\alpha\frac{s}{\sqrt{n}}\right]\right]\right) \geq 2\Phi(t_\alpha)-1 = 1-\alpha$$

★ **Classique !** ★
C'est très classique (et le seul cas qui semble être au programme pour cette partie du cours).

✘ **Attention !**
L'intervalle de confiance est donné par deux estimateurs, qui ne peuvent pas dépendre du paramètre m recherché ! Autrement dit, l'expression des bornes de l'intervalle de confiance ne doit pas faire apparaître m !

☞ **Pour info...**
Avec $\alpha = 0,05$, on a : $\Phi^{-1}\left(1-\dfrac{\alpha}{2}\right) \simeq 1,96$.

Petite remarque
L'intervalle de confiance asymptotique proposé est centré en $\overline{X_n}$, d'amplitude $2t_\alpha\dfrac{s}{\sqrt{n}}$.

☞ **Pour info...**
De façon générale, σ est inconnu et on peut alors utiliser un estimateur de l'écart-type dans l'intervalle de confiance asymptotique...

Exemple 10

On considère une variable aléatoire suivant une loi de Bernoulli de paramètre p inconnu ainsi qu'un n échantillon $(X_1, \ldots, X_n)$ de X. On note $\overline{X_n}$ la moyenne empirique de cet échantillon.

1. Justifions que :
$$\forall x \geq 0, \lim_{n \to +\infty} \mathbb{P}\left(\left[-x \leq \sqrt{n}\frac{\overline{X_n} - p}{\sqrt{p(1-p)}} \leq x\right]\right) = \Phi(x) - 1$$
où Φ est la fonction de répartition d'une variable aléatoire suivant la loi $\mathcal{N}(0;1)$.
2. Soit $\alpha \in]0;1[$. Démontrons l'existence d'un unique réel t_α tel que $\Phi(t_\alpha) = 1 - \dfrac{\alpha}{2}$. On donne : $t_{0,05} \simeq 1,96$.
3. Déduisons-en que $\left[\overline{X_n} - \sqrt{\dfrac{1}{n}}, \overline{X_n} + \sqrt{\dfrac{1}{n}}\right]$ est un intervalle de confiance asymptotique de p au niveau de confiance 95%.
4. Quelle taille d'échantillon devons-nous avoir pour que l'intervalle de confiance asymptotique donné ait une amplitude inférieure ou égale à $0,02$?
5. Comparons cet intervalle avec celui obtenu dans l'exemple 9.

Petite remarque

Quand $n = 100$, on obtient un intervalle de confiance de la forme $[\overline{X_{100}} - 0,1; \overline{X_{100}} + 0,1]$ ce qui est relativement grand pour estimer un paramètre $p_i n]0;1[$...