Premiers pas en statistique inférentielle et estimation

MAP - FISA 1

Introduction

Cadre : nous disposons d'une série d'observations provenant d'un modèle inconnu. Le travail du statisticien consiste alors à faire correspondre un modèle probabiliste.

- $\mathcal{P}$ population
- le caractère d'intérêt : information qualitative ou quantitative
- le phénomène suit la loi de probabilité $\{P_\theta; \theta \in \Theta\}$: connue ou pas ? (si inconnue, on cherche à la déterminer)
- le paramètre du modèle à estimer : θ

Exemples de questions :

SONDAGE ALÉATOIRE intention de vote pour un candidat à l'élection présidentielle,

ÉTUDES DE MARCHÉ : dépenses moyennes affectées par les différentes catégories d'une population à un type d'achat, proportion des ménages possèdant une voiture et la distribution de ces véhicules suivant la marque, l'âge ...,

CONTRÔLE DE QUALITÉ : proportion de déchets dans un lot de pièces industrielles, pourcentage d'erreurs commises lors d'un inventaire, ...

1. vocabulaire de base
2. intervalle de confiance en supposant le modèle connu a priori (tests d'adéquation pour déterminer la loi, hors programme)

1. Vocabulaire de base

2. Estimation des paramètres
   - Intervalle de confiance
   - Estimation d'une moyenne
   - Estimation d'une proportion

Mise en oeuvre

$\hookrightarrow$ passe par la réalisation d'un prélèvement d'échantillon (terme générique : sondage) due à l'impossibilité d'évaluer le caractère d'intérêt sur la population totale.

Remarque : Théorie des sondages non détaillée dans ce cours.

- Vocabulaire : $\mathcal{P}$ population, on appelle individu chaque élément de $\mathcal{P}$
- Le tirage d'un individu est dit probabiliste si la probabilité qu'à chaque individu de $\mathcal{P}$ d'être tiré est connue à priori.
- Le tirage est dit au hasard si chaque individu a la même probabilité d'être tiré (Si Card$\mathcal{P}$ = N, alors $\mathbb{P}$(tirer l'individu numéro i)= 1/N). Pour cela on utilise un générateur de nombre aléatoire.

- Un tirage au hasard est dit non exhaustif lorsque l'effectif de $\mathcal{P}$ ne varie pas au cours des tirages (ou très peu) ; la probabilité de chaque individu d'être tiré reste constante (ou est considérée comme telle) : tirage avec remise ou N très grand.
- Dans toute la suite, on se place dans ce cas. Ceci nous permet l'utilisation pour établir la statistique mathématique, de la théorie des probabilités (lois de probabilités, LGN, TCL ...)
- On note C le caractère d'intérêt quantitatif (exemple : dépense moyenne sur un type d'achat, nombre d'articles rejetés dans un lot prélevé, nombre d'erreurs, ...) ou qualitatif - dans ce cas, on transforme les donnés selon un codage predéfini (exemple : O/N pour intention de vote traduit par une variable de Bernouilli (succès = Oui), ...

Un exemple : $\text{marque du véhicule } X = \begin{cases} 1 \text{ si "Peugeot"} \\ 2 \text{ si "Renault"} \\ 3 \text{ si "Citroën"} \\ 4 \text{ si "Volkswagen"} \\ 5 \text{ si "Ford"} \\ 6 \text{ si "Dacia"} \\ 7 \text{ si "Opel"} \\ 8 \text{ si "Nissan"} \\ 9 \text{ si "Toyota"} \\ 10 \text{ si "Audi"} \\ 11 \text{ si "autres"} \end{cases}$

$$\text{âge du véhicule } X = \begin{cases} 1 \text{ si âge} < 1 \text{ an} \\ 2 \text{ si entre 1 et 2 ans} \\ 3 \text{ si entre 2 et 3 ans} \\ 4 \text{ si entre 3 et 4 ans} \\ 5 \text{ si entre 4 et 5 ans} \\ 6 \text{ si entre 5 et 10 ans} \\ 7 \text{ si} > 10 \text{ ans} \end{cases}$$

Cadre probabiliste

- Expérience aléatoire e = tirage au hasard d'un individu de $\mathcal{P}$
  Notons ω_i l'événement élémentaire : "on obtient l'individu numéro i".

- On définit X la variable aléatoire associée au caractère C (réalisation = valeur de C pour l'individu tiré)
  $$X : \begin{array}{ccc} (\Omega, \mathcal{F}, \mathbb{P}) & \rightarrow & E \\ \omega_i & \mapsto & c_i = x_i \end{array}$$

- **Hypothèse** : on connait la loi $\{P_\theta ; \theta \in \Theta\}$ de la variable X mais pas son paramètre θ que l'on souhaite estimer.

- Observation d'un échantillon de X de taille n
- E.a. E = tirage au hasard non exhaustif de n individus de $\mathcal{P}$
- On dit que les v.a.r. $(X_1, ..., X_n)$ forment un échantillon indépendant indentiquement distribué (i.i.d.) de X (n copies indépendantes de X).
  Réalisation de l'échantillon : $(x_1, ..., x_n) = (X_1(\omega), ..., X_n(\omega))$ pour une expérience ω.
- Le problème consiste à estimer θ à partir de $(x_1, ..., x_n)$.
- On parle de modèle statistique pour définir la famille $(\Omega, \mathcal{F}, X, E, (P_\theta; \theta \in \Theta))$.

Exemple :

Expérience aléatoire : lancé d'un dé. On notera la face lue sur le dé (ω). Question : le dé est il pipé ?
Pour la résoudre, on s'intéresse à la probabilité d'obtenir la face 6.

On définit la variable X qui vaut 1 si la face lue est 6, 0 sinon. On répète $n = 10$ fois de façon indépendante l'expérience ; on a donc un échantillon i.i.d. de taille 10 de la variable X. Notons $(x_1, ..., x_{10})$ les observations obtenues, par exemple : (1, 0, 0, 1, 0, 0, 0, 0, 0, 0).

Hypothèse : X suit une loi de Bernouilli de paramètre la probabilité d'obtenir 6 (notons la p).
But : à partir de notre réalisation de l'échantillon, estimer p pour savoir si $p = 1/6$.

Critères

Definition

Soit $(X_1, ..., X_n)$ un échantillon i.i.d. de X, $(x_1, ..., x_n)$ une réalisation de celui-ci. On appelle estimateur d'ordre n de θ, une statistique de $(X_1, ..., X_n)$: $T_n = f(X_1, ..., X_n)$ dont la réalisation observée est $f(x_1, ..., x_n)$.

On appelle

Definition

Une statistique de X est une variable aléatoire $f(X_1, ..., X_n)$ où f est une fonction mesurable de n variables et $(X_1, ..., X_n)$ est un échantillon de X.

Un estimateur est donc une variable aléatoire.

Quel choix pour la fonction f qui nous aide à estimer le paramètre θ ? Elle doit satisfaire les propriétés suivantes :

- T_n est un estimateur convergent de θ lorsque la suite de v.a.r. $(T_n, n \geq 1) \xrightarrow{\mathbb{P}} \theta$ lorsque $n \to \infty$.
- T_n est un estimateur fortement consistant de θ lorsque la suite de v.a.r. $(T_n, n \geq 1) \xrightarrow{p.s.} \theta$ lorsque $n \to \infty$.
- T_n est un estimateur sans biais de θ si $\mathbb{E}(T_n) = \theta$, sinon on note $b_n(\theta) = (\mathbb{E}(T_n) - \theta)$ le biais ;
- ou asymptotiquement sans biais si $\mathbb{E}(T_n) \to \theta$ lorsque $n \to \infty$.

Critères - II

- On a le résultat suivant pour montrer la convergence d'un estimateur :

Theorem

Soit T_n un estimateur de θ.
(i) Si $\mathbb{E}(T_n) \to \theta$ et $Var\,(T_n) \to 0$ lorsque $n \to \infty$, alors T_n est convergent ;
(ii) Si T_n est sans biais et si $Var\,(T_n) \to 0$ lorsque $n \to \infty$, alors T_n est convergent.

- Si T_n et S_n sont deux estimateurs sans biais de θ, et si $\mathrm{Var}\,(T_n) \leq \mathrm{Var}\,(S_n)$, alors on dit que T_n est plus efficace que S_n. Il aura la préférence.

Suite de l'exemple : Jet d'un dé ; $X \sim \mathcal{B}e(p)$, on souhaite estimer p.

Idée naturelle : $T_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i = \bar{X}_n$ (moyenne empirique de l'échantillon).

On calcule $\bar{x_{10}} = 0.2$ dans notre échantillon.

Remarque : si on prélève un autre échantillon, par exemple (1, 0, 0, 1, 0, 0, 0, 0, 0, 1), on obtient $\bar{x_{10}}^{(2)} = 0.3$. L'estimation dépend donc de l'échantillon !

On montre que T_n ainsi choisi est un estimateur sans biais et convergent. Il sera donc l'estimateur naturel choisi.

Moyenne empirique de l'échantillon

Definition

On appelle Moyenne de l'échantillon $(X_1, ..., X_n)$ et on note $\overline{X} : \Omega^n \to \mathbb{R}$ la v.a. définie par : $\overline{X} = \dfrac{X_1 + ... + X_n}{n}$.

Sa réalisation dans l'échantillon est $\overline{x}$.

Proposition 1 : $\mathbb{E}(\overline{X}) = m, \; Var(\overline{X}) = \dfrac{\sigma^2}{n}$.

Proposition 2 : $\overline{X}$ est un estimateur sans biais et convergent de $\theta = \mathbb{E}(X)$. De plus, si $\mathbb{E}(X) \neq 0$, c'est le plus efficace des estimateurs sans biais linéaires (de la forme $a_1X_1 + ... + a_nX_n$).

Variance empirique de l'échantillon

Première définition :

Definition

On appelle Variance empirique de l'échantillon $(X_1, ..., X_n)$ et on note $S^2 : \Omega^n \to \mathbb{R}$ la v.a. définie par : $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Proposition 3 : $\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2$

Proposition 4 : $Var(S^2) = \frac{\mu_4 - \sigma^4}{n} - 2\,\frac{\mu_4 - 2\sigma^2}{n^2} + \frac{\mu_4 - 3\sigma^4}{n^3}$ avec $\mu_n = \mathbb{E}((X-m)^n)$ le moment centré d'ordre n de X.

Seconde définition :

Proposition 5 :
(i) Notons $\tilde{S}^2 : \Omega^n \to \mathbb{R}$ la v.a. définie par :

$$\tilde{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

la variance empirique de l'échantillon $(X_1, ..., X_n)$. C'est un estimateur sans biais et convergent de $\theta = \mathrm{Var}(X)$.

(ii) Si de plus on connaît $\mathbb{E}(X)$, alors $T^2 = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (X_i - \mathbb{E}(X))^2$ est aussi un estimateur sans biais et convergent de $\theta = \mathrm{Var}(X)$; il est plus efficace que $\tilde{S}^2$.

1. Vocabulaire de base

2. Estimation des paramètres
   - Intervalle de confiance
   - Estimation d’une moyenne
   - Estimation d’une proportion

Estimation des paramètres

Cadre d'étude :

Population $\mathcal{P}$ de taille N
Caractère d'intérêt : C
Variable aléatoire associée à C : X de loi $\{P_\theta; \theta \in \Theta\}$
Modèle statistique : $(\Omega, \mathcal{F}, X, E, (P_\theta; \theta \in \Theta))$.

$\Downarrow$ $\Uparrow$ θ ? statistique inférentielle

Prélevement d'un échantillon de taille n
Observation $(x_1, ..., x_n)$ de l'échantillon i.i.d. de X
Détermination de la valeur du paramètre θ dans l'échantillon : $\tilde{\theta}$

Intervalle de confiance

- Rappel : nous avons vu qu'un estimation ponctuelle est fortement dépendante de l'échantillon prélevé.
- But : construire un intervalle qui contienne la valeur exacte de θ pour pouvoir apprécier la précision de l'estimation effectuée.

Definition

Si deux statistiques S_n, T_n sont telles que $\mathbb{P}(S_n \leq \theta \leq T_n) \geq 1 - \alpha$, on appelle $[S_n, T_n]$ un intervalle de confiance du paramètre θ au risque α ($1 - \alpha$ est appelé niveau ou degré de confiance de l'intervalle).

Estimation d'une moyenne

Contexte et notations :

Dans une population $\mathcal{P}$ de taille N
Caractère d'intérêt : C
Variable aléatoire associée à C : X de moyenne inconnue m et de variance σ^2

$$\Downarrow$$

Prélevement d'un échantillon de taille n
Observation $(x_1, ..., x_n)$ de l'échantillon
Moyenne dans l'échantillon : $\overline{x}$
Variance dans 'échantillon : s^2

Cas d'une loi normale

- Hypothèse : $\mathcal{L}(X) = \mathcal{N}(m, \sigma)$
- **Proposition 6** : $\mathcal{L}(\overline{X}) = \mathcal{N}\left(m, \dfrac{\sigma}{\sqrt{n}}\right)$
- Donc : $\mathcal{L}\left(\dfrac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}\right) = \mathcal{N}(0, 1)$
- On distingue deux cas : si la valeur de σ est connue ; et si elle est inconnue, on va l'estimer.

Si σ est connu

Definition

Le quantile d'ordre a d'une v.a.r. T est la valeur z telle que $\mathbb{P}(T < z) = a$.

On note $z = z(\frac{\alpha}{2})$ la valeur telle que $\mathbb{P}\left(T > z\right) = \dfrac{\alpha}{2}$ avec

$$T = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

z est le quantile d'ordre $1 - \frac{\alpha}{2}$ de T.

Ainsi $\mathbb{P}(T < z) = 1 - \dfrac{\alpha}{2}$ donc : $\mathbb{P}(-z < T < z) = 1 - \alpha$

Proposition 7 : Intervalle de confiance de *m* au risque α dans le cas d'une variable *X* de loi normale $\mathcal{N}(m, \sigma)$ avec *m* inconnue et σ connue :

$$\left[\overline{X} - z\frac{\sigma}{\sqrt{n}} ; \overline{X} + z\frac{\sigma}{\sqrt{n}}\right]$$

- Lorsque α diminue, la précision de l'intervalle diminue.
- Lorsque *n* augmente, la précision augmente.

Si σ est inconnu

Il faut l'estimer !

- Rappel : variance empirique de l'échantillon $(X_1, ..., X_n)$
$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ estimateur biaisé et convergent de la variance
$\tilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ estimateur sans biais et convergent de la variance
- On estime donc σ par $\sqrt{\tilde{S}^2}$ ou $\sqrt{S^2\frac{n}{n-1}}$.

- On a besoin du résultat suivant :
  Proposition 8 : Si X suit une loi normale $\mathcal{N}(m, \sigma)$ avec m et σ inconnus, alors $\mathcal{L}\left(\dfrac{\overline{X} - m}{S/\sqrt{n-1}}\right)$ est une loi de Student à $n-1$ degrés de liberté $\mathcal{S}_{n-1}$.
- C'est une nouvelle loi !

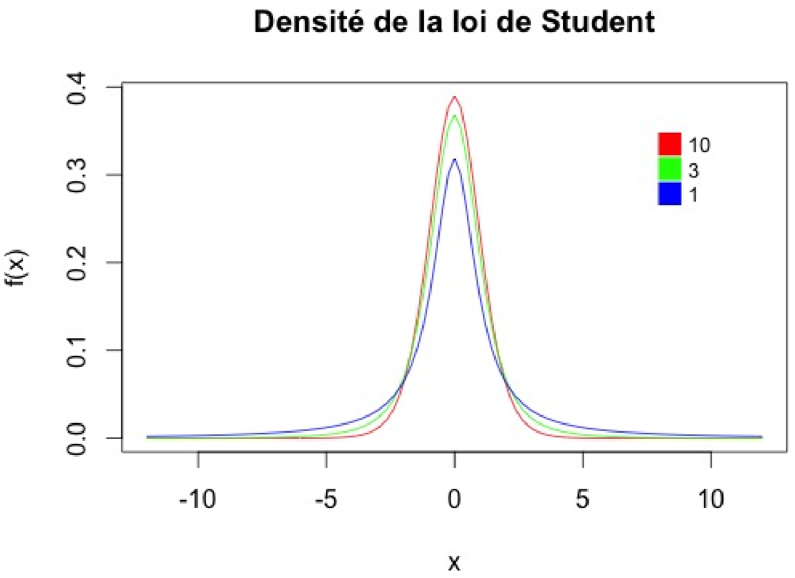
Loi de Student/Student-Fisher

Soit un entier $n > 0$. Une v.a.r. X suit la loi du Student à n degrés de liberté si elle a pour densité de probabilité :

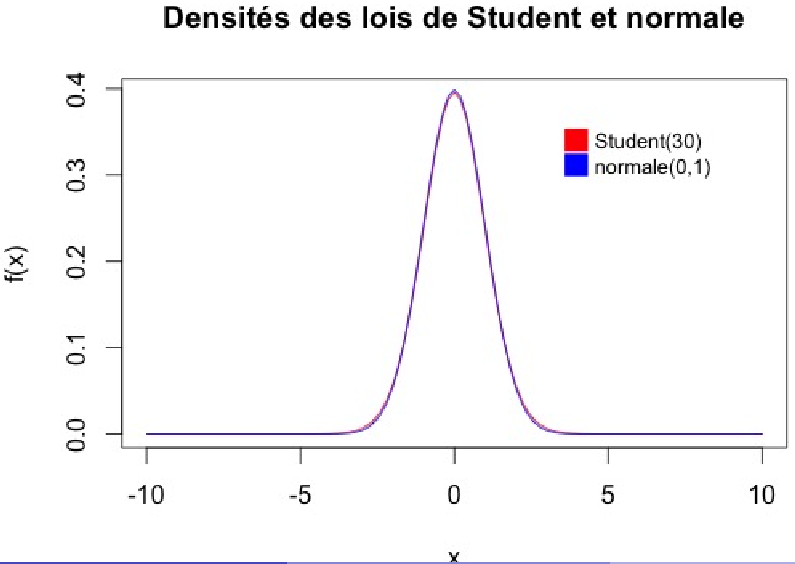
$$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \frac{1}{(1 + \frac{x^2}{n})^{\frac{n+1}{2}}}$$

avec $(a > 0)$ $\Gamma(a) = \int_0^{+\infty} u^{a-1} \exp(-u)\, du$ fonction Gamma

- $\mathcal{L}(X) = \mathcal{S}_n$
- décrite par le statisticien William Gosset (pseudonyme Student) en 1908


Densité de la loi de Student
f(x)
0.0
0.1
0.2
0.3
0.4
-10
-5
0
5
10
x
10
3
1

- Approximation d'une loi de Student à n degrés de liberté par une loi normale centrée réduite si $n > 30$.



- $\mathbb{E}(X) = 0$ pour $n > 1$
- $\sigma_X^2 = \dfrac{n}{n-2}$ pour $n > 2$
- On notera $t_n = t_n(\beta)$ le quantile d'ordre $1 - \frac{\beta}{2}$ de la loi de Student à n degrés de liberté. Pour $0 \leq \beta \leq 1$, soit $t_n(\beta)$ défini par $\mathbb{P}(Y > t_n(\beta)) = \beta$ si Y est une v.a.r. de loi $\mathcal{S}_n$.
- Pourquoi utilise-t-on la loi de Student ? on l'obtient par l'intermédiaire d'une autre loi.

Proposition 9 : Intervalle de confiance de m au risque α dans le cas d'une loi normale $\mathcal{N}(m, \sigma)$ avec m et σ inconnus

$$\left[\overline{X} - t_{n-1}\frac{\tilde{s}}{\sqrt{n}} ; \overline{X} + t_{n-1}\frac{\tilde{s}}{\sqrt{n}}\right]$$

Remarque : Si n grand, $t_n \approx z$.

Cas d'une loi quelconque et des grands échantillons

- Loi de X inconnue $\Rightarrow$ Loi de $\bar{X}$ inconnue a priori
- On peut appliquer le Théorème Central Limite lorsque n est grand
- Pour $n > 30$ (grands échantillons), $\mathcal{L}\left(\dfrac{\bar{X} - m}{\sigma/\sqrt{n}}\right) \approx \mathcal{N}(0, 1)$
- variance estimée
- **Proposition 10** : Intervalle de confiance approché de m au risque α pour de grands échantillons ($n > 30$) :

$$\left[\overline{X} - z\frac{\tilde{s}}{\sqrt{n}}; \overline{X} + z\frac{\tilde{s}}{\sqrt{n}}\right]$$

Application

- On se place dans le cas d'une loi de X normale avec écart-type connu
- En estimant m par $\bar{X}$, on commet l'erreur aléatoire $\left|m - \bar{X}\right|$
- Quelle est la taille de l'échantillon à extraire pour obtenir une précision donnée ?
- On impose : l'erreur ne doit pas dépasser a avec une probabilité au moins égale à $1 - \alpha$
- $\mathbb{P}\left(\left|m - \bar{X}\right| \leq a\right) \geq 1 - \alpha$
- Taille minimale de l'échantillon : $n \geq \left(\frac{z\sigma}{a}\right)^2$

Estimation d'une proportion

Contexte et notations :

Dans une population $\mathcal{P}$ de taille N
on veut estimer la proportion p des individus qui satisfont à une propriété M
Variable aléatoire associée : v.a.r. indicatrice X de loi de Bernouilli $\mathcal{B}e(p)$

$$\Downarrow$$

Prélevement d'un échantillon de taille n
Echantillonnage au hasard non exhaustif $(X_1, ..., X_n)$ de X
Observation $(x_1, ..., x_n)$ de l'échantillon

$\bar{X} = K/n$ v.a.r. moyenne dans l'échantillon, sa réalisation $\bar{x} = k/n$ est la proportion d'individus dans l'échantillon satisfaisant M.

- $\mathbb{E}(X) = p$ donc estimer p revient à estimer la moyenne de X.
- **Proposition 11** :On estime p par la proportion d'individus satisfaisant à M dans l'échantillon extrait : k/n.
- Intervalle de confiance : cela revient à chercher un intervalle de confiance d'une moyenne dans le cas d'une loi quelconque d'écart-type inconnu, d'où :

$$\left[\overline{X} - z\frac{\tilde{s}}{\sqrt{n}}; \overline{X} + z\frac{\tilde{s}}{\sqrt{n}}\right]$$

où α est le risque. On obtient :
Proposition 12 :Intervalle de confiance approché pour n grand de la proportion p :

$$\left[\frac{k}{n} - z\sqrt{\frac{k(n-k)}{n^3}}; \frac{k}{n} + z\sqrt{\frac{k(n-k)}{n^3}}\right]$$

Application

- En estimant p par $\dfrac{K}{n}$, on commet l'erreur aléatoire $\left|p - \frac{K}{n}\right|$
- Quelle est la taille de l'échantillon à extraire pour obtenir une précision donnée ?
- On impose : l'erreur ne doit pas dépasser a avec une probabilité au moins égale à $1 - \alpha$
- $\mathbb{P}\left(\left|p - \dfrac{K}{n}\right| \leq a\right) \geq 1 - \alpha$
- Taille minimale de l'échantillon : $n \geq \frac{z^2}{4\,a^2}$