Cours de Statistiques inférentielles en Master 1 Data Science et Intelligence Artificielle(**DSIA)**

IBRAHIMA SY

*Institut Supérieur Informatique(**ISI**)*

Table des matières

Introduction

Les probabilités sont l'étude du hasard et de l'incertain. Elle permettent de donner un cadre formel et rigoureux aux nombreux phénomènes physique aléatoires.

L'étude des probabilités a connu son essor au **XX**ème siècle lorsque leur application à d'autre domaines des sciences ont été découvert : en physique (mécanique quantique, physique statistique), en biologie (météorologie, génétique des populations), en économie (théorie des jeux, mathématiques financières, assurances), en sociologie (démographie, sondage)... Elle constitue actuellement un champs d'étude très actif.

CHAPITRE UN

L'ECHANTILLONNAGE STATISTIQUE

1.1 Rôle de l'échantillonnage

1.2 Vocabulaire

1.3 Méthode d'Echantillonnage

1.4 Définition et Caractéristiques d'un Echantillon

1.5 Distribution Echantillonnage

1.5.1 La distribution normale comme distribution d'échantillonnage

1.5.2 La distribution du χ^2 (Khi-Deux)

1.5.3 La distribution du t (Student)

Définition 1.5.1. *On appelle variable aléatoire du t , ou v.a. de Student à k degrés de liberté une v.a. continue qui prend des valeurs réelles , et dont la densité est*

$$f_T(t) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}, -\infty < t < +\infty$$

où k est un entier positif . Par la suite, on désignera par $St(k)$ la famille des v.a. du t à k dergés de liberté, de telle sorte que pour une v.a. X , on pourra écrire $X \sim St(k)$

Caractéristiques d'une v.a. du t

Il est possible de prouver que l'espérance et la variance d'une v.a. X du t à m degrés de liberté sont données par : $E(X) = 0$ si $m > 1$ et $Var(X) = \frac{m}{m-2}$ si $m > 2$

Représentation graphique et utilisation des tables

La courbe représentant la densité de la v.a. du t à k degrés de liberté est une courbe en forme de cloche, qui est symétrique par rapport à la droite $t = 0$, tout comme celle de la densité normale centrée réduite. Cependant, la courbe est plus aplatie que celle de la normale centrée réduite, comme en témoigne la figure ci- dessous.

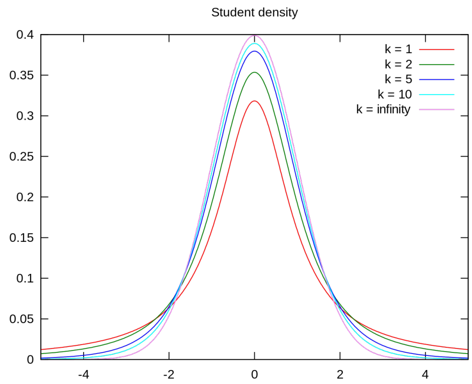


FIGURE 1.1 – Densité de probabilité

Lotrsque k augmente, la courbe de la t converge celle de la normale. On peut même montrer que, lorsque k tend vers l'infini, la densité de la densité de la variable de $St(k)$ converge vers la densité normale centrée réduite. La table 3 de l'annexe donne les les probabilités cumulées de la forme "plus grand ou égal".

On utilise la notation t_α pour désigner une valeur particulière de la v.a. de St à k degrés de liberté telle que la probabilité que la v.a. prenne une valeur plus grande ou égale à cette valeur particulière est égale à α , c'est à dire $\mathbb{P}(t \geq t_{\alpha,m}) = \alpha$.

Propriété 1.5.1. *Soit $X \sim \mathcal{N}(0,1)$ et $Y \sim \chi^2_m$ X et Y etant indépendantes, alors la v.a $X/\sqrt{Y/m} \sim St(m)$*

Utilisation comme distribution d'échantillonnage

Théorème 1.5.1. *La distribution de $(\bar{X} - \mu)/(S/\sqrt{n})$ Soit $(X_1, \ldots, X_n)$ un échantillon aléatoire de taille n tiré d'une population normale de moyenne μ et de variance σ^2 , et soit S^2 la variance de cet échantillon (S son écart type) ; alors la statistique*

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim St(n-1)$$

Exemple 1.5.1. *Pour estimer le montant hebdomadaire moyen dépensé par les familles de 4 personnes pour leur épicerie, on tire un échantillon aléatoire de 25 personnes (chaque personne représentant famille). On suppose que les montants moyens dépensés sont distribués normalement avec une moyenne $\mu = 120\$$ et une variance inconnue. Si la variance de l'échantillon de taille 25 est $s^2 = 36\$$, calculer la probabilité que la moyenne $\bar{X}$ de l'échantillon soit supérieure ou égale à 123\$.*

La distribution de inconnues mais égales $(\bar{X}_1 - \bar{X}_2)$ pour deux populations normales de variances inconnues mais égales

Précedemment, on a vu que pour comparer deux populations par rapport à une certaine carac- téristique, on tire indépendamment un échantillon de taille n_1 dans la première, et un échantillon de taille n_2 dans la deuxième puis on considère la dèistribution de la différence $\bar{X}_1 - \bar{X}_2$ entre les deux moyennes échantillonnales. On va maintenant supposer que les deux les deux popu- lations étudiées sont normales de moyennes μ_1 et μ_2 sont connuues, mais que les variances σ_1^2et σ_2^2 sont inconnues. On va aussi supposer que ces variances inconnues sont égales ($\sigma_1^2 = \sigma_2^2$) . Alors on considère une statistique analogue à $(\bar{X}?\mu)/(S/\sqrt{n})$ mais adaptée aux cas de deux populations. On a le résultat suivant : si S_1^2 et S_2^2 sont les variances de chacun des deux échantillons alors la statistique

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}}$$

suit une distribution du t à $(n_1 + n_2 - 2)$ degrés de de libertés. Autrement dit, si $X_1 \sim N(\mu_1, \sigma_1^2)$ et $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, σ_1^2 et σ_2^2 inconnues mais égales, alors

$$\left[\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}}\right] \sim t(n_1 + n_2 - 2)$$

Exemple :

Reprenons l'exemple concernant la durée de vie des pneus produits par deux manufacturiers M_1 et M_2 . Supposons que pneus de M_1 aient une durée de vie moyenne de 50.000 km avec un écart type inconnu, et que ceux de M_2 aient une durée de vie moyenne de 40.000 km avec un écart type inconnu, ces deux durées de vie étant distribuées normalement, et les écarts types inconnus mais supposés égaux. Pour comparer les pneus de ces deux manufacturiers, on a tiré un échantillon aléatoire de 10 pneus M_1 , et l'on a obtenu un écart type de 6.000 km ; on a tiré un échantillon aléatoire de 15 pneus M_2 , et l'on a obtenu un écart type de 4.000 km. Quelle est la probabilité que la durée de vie moyennes de ces 10 pneus M_1 soit d'au moins 15.000 km de plus que la durée de vie moyennes de ces 15 pneus M_2 ?

1.5.4 La distribution du F (Fisher)

Définition 1.5.2. *On appelle variable aléatoire du F , ou v.a. de Fisher à d_1 et d_2 degrés de liberté une v.a. continue qui prend des valeurs réelles positives, et dont la densité est :*

$$f(x) = \frac{\left(\frac{d_1\,x}{d_1\,x+d_2}\right)^{d_1/2} \left(1 - \frac{d_1\,x}{d_1\,x+d_2}\right)^{d_2/2}}{x\;\mathrm{B}(d_1/2, d_2/2)}$$

où d_1 et d_2 sont des entiers positifs et B est la fonction bêta. Par la suite, on désignera par $F(d_1, d_2)$ la famille des v.a. du F à d_1 et d_2 dergés de liberté, de telle sorte que pour une v.a. X , on pourra écrire $X \sim F(d_1, d_2)$.

Caractéristiques d'une v.a. du F

Il est possible de prouver que l'espérance et la variance d'une v.a. X du F à d_1 et d_2 degrés de liberté sont données par :

$E(X) = \frac{d_2}{d_2-1}, sid_2 > 2$ et

$Var(X) = \frac{2m_2^2(m_1+m_2-2)}{m_1(m_2-2)^2(m_2-4)}, sim_2 > 4$

Représentation graphique et utilisation des tables

Tout comme la courbe représentant la densité de la v.a. du χ^2 , la courbe de la F est non symétrique, et elle est définie seulement pour des valeurs positives de la variable.

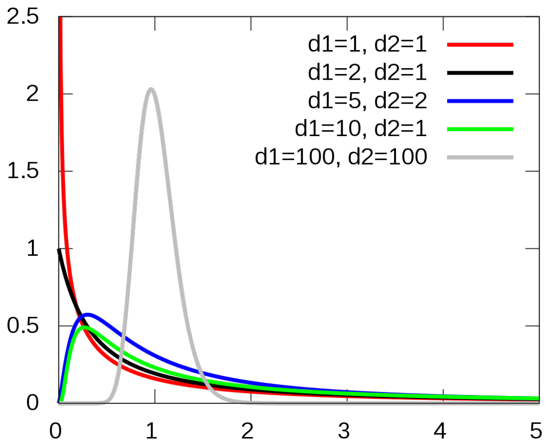


FIGURE 1.2 – Densité de probabilité

Propriété 1.5.2. *Soient X_1 et X_2 deux variables aléatoires indépendantes suivant toutes deux une distribution du χ^2 à m_1 et m_2 degrés de liberté respectivement, alors la v.a. $Y = (X_1/m_1)/(X_2/m_2) \sim F(m_1, m_2)$.*

Utilisation comme distribution d'échantillonnage

Pour comparer deux populations par rapport à une certaine carctéristique, on peut comparer leurs moyennes, on peut aussi vouloir comparer les variances de ces populations. À cette fin, on tire indépendamment un échantillon aélatoire de taille n_1 dans la première population, et un échantillon aélatoire de taille n_2 dans la deuxième. Pour comparer les variances σ_1^2 et σ_2^2 de ces deux populations, on se sert du rapport S_1^2/S_2^2 des variances des deux échantillons tirés. Plus précisément, si les deux populations considérées sont normales, alors la statistique une distribution du F à $(n_1?1)$ et $(n_2?1)$ degrés de liberté. Autrement dit, si $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ alors $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$.

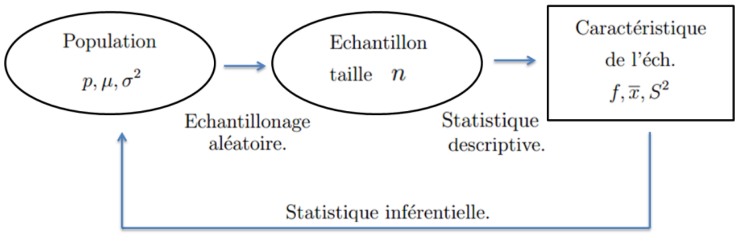
On peut aussi noter que, sous l'hypothèse $\sigma_1^2 = \sigma_2^2$ (les deux populations ont la même variance), la statistique $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ se réduit simplement au quotient S_1^2/S_2^2 des deux variances échantillonnales.

Exemple 1.5.2. *On désire maintenant comparer les variances des durées de vie des pneus des deux manufacturiers M_1 et M_2 . Supposons que la durée de vie des pneus M_1 soit normale de variance $\sigma_1^2 = (3015)^2 km^2$, et que celle des pneus M_2 soit normale de variance $\sigma_2^2 = (4000)^2 km^2$. Si l'on tire indépendamment un échantillon aléatoire de 10 pneus de M_1 et un échantillon aléatoire de 20 pneus de M_2 , calculer la probabilité que la variance S_1^2 du premier échantillon soit au moins 2 fois plus grande que la variance S_2^2 du deuxième échantillon.*

Chapitre Deux

Estimation ponctuelle

L'estimation consiste à donner des valeurs approchées aux paramètres d ?une popula- tion (p, μ, σ^2) ou (proportion, moyenne, variance) à partir des données de l ?échantillon $(f, \bar{x}, S^2)$. On supposera vérifiée l ?hypothèse d ?échantillonnage aléatoire simple. La statistique inférentielle peut se résumer par le schéma suivant :



2.1 Généralités

2.1.1 Estimateur

Soient $(X_1, X_2, \ldots, X_n)$ un n-échantillon aléatoire simple issu d ?une variable aléatoire X (discrète ou continue) et θ un paramètre associé à la distribution $\mathbb{P}_\theta(X = x)$ ou $f_\theta(x)$ de X. Un estimateur de θ est une variable aléatoire T fonction des X_i

$$T = f(X_1, X_2, \ldots, X_n)$$

Si on considère n observations $x_1, x_2, \ldots, x_n$, l'estimateur T fournira une estimation de θ notée :

$$\hat{\theta} = f(x_1, x_2, \ldots, x_n)$$

Définition 2.1.1. *(Lois de distribution). Soit $(X_1, \ldots, X_n)$ un échantillon de taille n de la variable aléatoire $X \sim \mathcal{L}(\theta)$ et $(x_1, \ldots, x_n)$ une réalisation. On appelle distribution de la loi $\mathcal{L}(\theta)$ la valeur : $\mathbb{P}_\theta(X = x)$ ou $f_\theta(x)$*

On définit alors la loi de distribution de l'échantillon :

— *Si $\mathcal{L}(\theta)$ est discrète par :*

$$
\begin{aligned}
L(x_1, \dots, x_n, \theta) &= \mathbb{P}(X_1 = x_1, \dots, X_n = x_n, \theta) && (2.1)\\
&= \mathbb{P}_\theta(X_1 = x_1) \times \cdots \times \mathbb{P}_\theta(X_n = x_n) && (2.2)
\end{aligned}
$$

— *Si $\mathcal{L}(\theta)$ est continue :*

$$L(x_1, \dots, x_n, \theta) = f_\theta(x_1) \times \cdots \times f_\theta(x_n) \tag{2.3}$$

On nomme, dans les deux cas, $L(x_1, \dots, x_n, \theta)$ la fonction de vraisemblance du paramètre θ

2.1.2 Exemples élémentaires

$\bar{x}, s^2$ sont des estimations de μ et de σ^2 (resp.). Les variables aléatoires $\overline{X}$ et S^2 , sont les estimateurs de μ et σ^2 (resp.).

Remarque 2.1.1. *Le même paramètre peut-être estimé à l ?aide d ?estimateurs différents.*

Exemple 2.1.1. *Le paramètre λ d ?une loi de Poisson peut-être estimé par $\overline{X}$ et S^2 .*

2.1.3 Qualités d'un estimateur

Estimateur convergent

Soit θ le paramètre à estimer et $T = f(X_1, X_2, \dots, X_n)$ un estimateur. La première qualité d'un estimateur est d ?être **convergent** (consistant : $\lim\limits_{n\to+\infty} T = \theta$). Deux estimateurs convergents ne convergent cependant pas nécessairement à la même vitesse, ceci est lié à la notion de précision d'un estimateur. On mesure généralement la précision d ?un estimateur par l'erreur quadratique moyenne.

Perte quadratique : C'est l ?écart au carré entre le paramètre et son estimateur :

$$l(\theta, T) = (\theta - T)^2$$

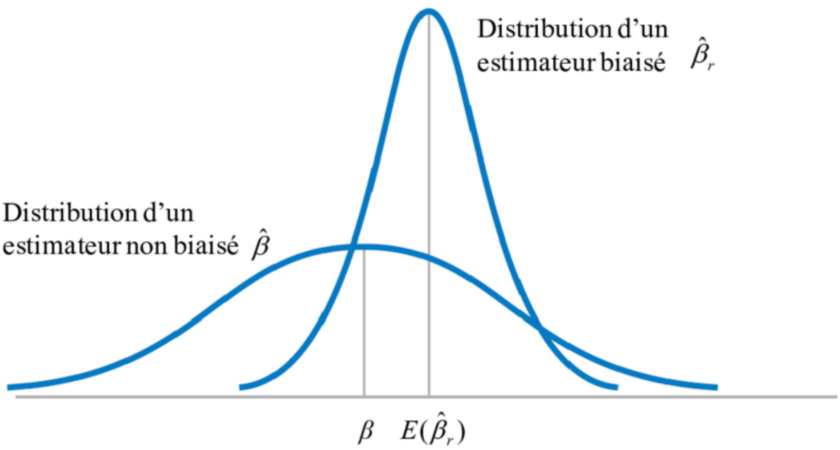
Risque d'un estimateur : C'est la moyenne des pertes :

$$R(T, \theta) = \mathbb{E}(l(\theta, T))$$

$R(T, \theta) = \mathbb{E}(\theta - T)^2$ est le risque quadratique moyen.
Un estimateur T est dit convergent si $\lim\limits_{n\to+\infty} R(T, \theta) = 0$

Estimateur sans biais

La quantité $\mathbb{E}(T) - \theta$ est appelée biais de l'estimateur. T est dit sans biais si $\mathbb{E}(T) = \theta$.



Exemple 2.1.2. $\overline{X}$ *est sans bias pour* μ*, mais* $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ *est biasé pour* σ^2*. C'est pourquoi on utilise* $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Remarque 2.1.2.

$$R(T,\theta) = E\Big((T-\theta)^2\Big) = Var(T) + \Big((\mathbb{E}(T)-\theta)^2\Big)$$

<u>Preuve :</u> *utilse l'astuce de cette décomposition :* $T - \theta = T - \mathbb{E}(T) + \mathbb{E}(T) - \theta$

Remarque 2.1.3. *Si* T *est sans biais alors* $R(T,\theta) = Var(T)$ *et* T *convergent si* $\lim\limits_{n\to+\infty} Var(T) = 0$*. Entre deux estimateurs sans biais, le plus précis est donc celui de variance minimale.*

Meilleur estimateur

Soient T_1 , T_2 deux estimateur de θ. On dit que T_1 est meilleur que T_2 si :

$$R(T_1,\theta) < R(T_2,\theta), \forall\theta$$

Estimateur asymptotiquement sans biais

T est dit asymptotiquement sans biais si :

$$\lim_{n\to+\infty} \mathbb{E}(T) = \theta$$

Pour un échantillon $(X_1, X_2, \ldots, X_n)$ issu de X de loi $f(x,\theta)$, on se contentera de rechercher un estimateur sans biais de variance minimale, ce problème est lié à l'existence de statistiques exhaustives.

2.2 Statistique suffisante (exhaustive)

Dans un problème statistique où figure un paramètre θ inconnu, un échantillon nous apporte certaine information sur ce paramètre. Lorsque l'on résume cet échantillon par une statistique, il s'agit de ne pas perdre cette information. Une statistique qui conserve l'information sera dite suffisante (exhaustive).

2.2.1 Définition d'une statistique exhaustive

Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon et $L(x, \theta)$ sa vraisemblance. Soit θ un paramètre influant sur la loi de probabilité à laquelle sont soumis les X_i . Une statistique T est dite exhaustive (pour le paramètre θ) si la loi conditionnelle de l'échantillon sachant T est indépendante de θ . Cela peut se traduire par la formule suivante :

$$\mathbb{P}(X = x | S(X) = s, \theta) = \mathbb{P}(X = x | T = s)$$

En pratique l'on se sert peu de cette formule pour montrer qu'une statistique est exhaustive et l'on préfère en règle générale utiliser le critère suivant appelé critère de factorisation (parfois aussi appelé critère de Fisher-Neyman) :

Une statistique T est exhaustive si et seulement s'il existe deux fonctions g et h mesurables telles que :

$$L(x, \theta) = g(\theta, T) \times h(x)$$

Exercice 2.2.1. *Soit $X_1, X_2, \ldots, X_n$ issu de $X \sim \mathcal{P}(\theta)$, θ inconnu*
La Statistique $T = \sum_{i=1}^{n} X_i$ est-elle exhaustive pour θ ?

2.2.2 Lois permettant une statistique exhaustive

Soit une variable X dont le domaine de définition ne dépend pas de θ

— Une condition nécessaire et suffisante pour que l'échantillon $(X_1, ..., X_n)$ admette une statistique exhaustive est que la forme de la densité soit : $f(x, \theta) = \exp(a(x) \cdot \alpha(\theta) + b(x) + \beta(\theta))$ (famille exponentielle)

— Si la densité est de cette forme et si de plus l'application $x_i \to \sum_{i=1}^{n} a(x_i)$ est bijective et continûment différentiable pour tout i, alors $T = \sum_{i=1}^{n} a(X_i)$ est une statistique exhaustive particulière.

Remarque 2.2.1. *Ce théorème est un outil très puissant pour la recherche d ?une statistique exhaustive.*

Exercice 2.2.2. *Soit $(X_1, \ldots, X_n)$ un n-echantillon issu de $X \sim \gamma(\theta)$*

$$f(x, \theta) = \frac{1}{\Gamma(\theta)} e^{-x} x^{\theta - 1}, x > 0$$

Determiner une statistique exhaustive pour θ ?

Remarque 2.2.2. *Lorsque le domaine de X dépend de θ, le théorème de Darmois ne s'applique pas, ce qui n'empêche pas de trouver une statistique exhaustive.*

2.2.3 L'information de Fisher

Définition 2.2.1. *On appelle quantité d'information de Fisher $I_n(\theta)$ apportée par un échantillon sur le paramètre θ, la quantité suivante positive ou nulle (si elle existe) : "*

$$I_n(\theta) = \mathbb{E}\Big[\Big(\frac{\partial \ln L(X,\theta)}{\partial\theta}\Big)^2\Big]$$

avec $L(X,\theta) = \prod_{i=1}^{n} f(X_i,\theta)$

Théorème 2.2.1. *Si le domaine de définition de X ne dépend pas de θ alors*

$$I_n(\theta) = -\mathbb{E}\Big[\frac{\partial^2 \ln L(X,\theta)}{\partial\theta^2}\Big]$$

Propriété 2.2.1. *d'additivité de l'information de Fisher : Soit Y une variable aléatoire indépendante de X dont la loi dépend du même paramètre θ. Soit $f(x,\theta), g(y,\theta), h(x,y,\theta)$ les densités de X, Y et du couple (X,Y) (resp.), d'information de Fisher $I_X(\theta)$, $I_Y(\theta)$, $I(\theta)$ (resp.) Si les domaines de définition de X, Y ne dépendent pas de θ alors :*

$$I(\theta) = I_X(\theta) + I_Y(\theta).$$

Consequence : Si le domaine de définition ne dépend pas de θ alors : $I_n(\theta) = nI_1(\theta)$ avec

$$I_1 = \mathbb{E}\left(\frac{\partial^2 \ln f(X,\theta)}{\partial\theta^2}\right)$$

i.e : que chaque observation a la même importance. Ce qui n ?est pas le cas pour la loi uniforme sur $[0,?]$ où la plus grande observation est la plus intéressante.

Exercice 2.2.3. *Soit $X \sim \mathcal{N}(\mu,\sigma^2)$ Calculer l ?information de Fisher apportée par un n-échantillon issu de X sur le paramètre μ.*

2.3 Information de Fisher et suffisance

On montre que l'nformation portée par une statistique est inférieure ou égale à celle apportée par un échantillon. En effet : soit T la statistique de densité $g(t,\theta)$ que l'on substitue à l'échantillon.

$$I_n(\theta) \geq I_T(\theta)$$

Si T est suffisante $I_n(\theta) = I_T(\theta)$
La réciproque est vraie si le domaine de X est indépendant de θ.

2.4 Borne de Freshet-Damois-Cramer-Rao (FDCR)

Le résultat suivant nous indique que la variance d'un estimateur ne peut être inférieur à une certaine borne, qui dépend de la quantité d'information de Fisher apportée par l ?échantillon sur le paramètre θ

Théorème 2.4.1. *Si le domaine de définition de X ne dépend pas de θ alors pour tout estimateur sans biais :*

$$Var(T) \geq \frac{1}{I_n(\theta)}$$

Si T est un estimateur sans biais de $h(\theta)$

$$Var(T) \geq \frac{[h'(\theta)]^2}{I_n(\theta)}$$

2.5 Estimateur efficace

Définition 2.5.1. *Un estimateur T est dit efficace si sa variance est égale à la borne de FDCR.*

Propriété 2.5.1. *Un estimateur sans biais efficace est convergent.*

Preuve : A faire

Remarque 2.5.1. *Un estimateur efficace T est un estimateur sans biais de variance minimale.*

2.6 Estimateur sans biais de variance minimale (MVUE)

Théorème 2.6.1. *S ?il existe un estimateur de θ sans biais, de variance minimale, il est unique presque sûrement.*

2.7 Généralisation (Cas multidimensionnel)

Statistique exhaustive

Soit le modéle statistique paramétrique : $(\mathcal{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$, avec $\theta \in \Theta \subset \mathbb{R}^p$

Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon issu de $X \sim f(x,\theta)$, on suppose que $f(x,\theta)$ appartiennent à la famille exponentielle :

$$f(x,\theta) = c(\theta).h(x).\exp(\sum_{j=1}^{n} \alpha_j(\theta)T_j(x))$$

Si $X(\Omega)$ ne dépend pas de θ

La statistique $(\sum_{i=1}^n T_1(X_i)), (\sum_{i=1}^n T_2(X_i)), \ldots, (\sum_{i=1}^n T_p(X_i))$ est exhaustive pour θ.

Information de Fisher

Soit $X \sim f(x,\theta)$, $\theta \in \Theta \subset \mathbb{R}^p$. On note $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$

On fait les hypothèses de régularité suivante :

(H_1) : $\forall x, \forall\theta, f(x,\theta) > 0$

(H_2) : $grad_\theta f$ existe $\forall\theta$, ie : on peut dériver f par rapport à θ_i , $\forall i \in \{1, \ldots p\}$, $\mathbb{P}_\theta$ *ps*

(H_3) : On peut dériver au moins deux fois $f(x,\theta)$ par rapport à θ_i , $\forall i \in \{1,\dots p\}$, dériver $\int f(x,\theta)dx$ et permuter entre dérivation et intégration.

$grad_\theta f = \left(\frac{\partial f}{\partial \theta_1}, \frac{\partial f}{\partial \theta_2}, \dots, \frac{\partial f}{\partial \theta_p}\right)$ vecteur ligne $(1,p)$. On appelle score $S(x,\theta)$ le vecteur $(1,p)$ défini par $grad_\theta \log f$

Définition 2.7.1. *On appelle information en θ la matrice (p,p) de variance-covariance de $grad_\theta \log f(x,\theta)$*

$$I(\theta) = \mathbb{E}(S^T S)$$

Comme dans le cas réel :
- $\mathbb{E}\Big[grad_\theta \log f(x,\theta)\Big] = 0$
- $\mathbb{E}\Big[\frac{1}{f}\frac{\partial^2 f(X,\theta)}{\partial \theta_i \partial \theta_j}\Big] = 0$, $\forall i,j \in \{1,\dots,p\}$

Sous les hypothèses H_1, H_2, H_3 on obtient, $I(\theta)$: la matrice dont le terme général $I_{ij}(\theta)$,

$$-\mathbb{E}\Big[\frac{\partial^2 f(X,\theta)}{\partial \theta_i \partial \theta_j}\Big]$$

$I(\theta)$ est une matrice définie positive.

Exercice 2.7.1. *Soit $(X_1, X_2, \dots, X_n)$ un n-échantillon issu de $N(\mu,\sigma^2)$, avec μ, σ^2 inconnus.*

Determiner l'information de Fisher $I(\mu,\sigma^2)$

Théorème 2.7.1. *Si T^* est un estimateur sans biais de θ dépendant d ?une statistique exhaustive complète U alors T^* est l'unique estimateur MVUE de θ. En particulier si l ?on dispose déjà d ?un estimateur T sans biais de θ : $T^* = \mathbb{E}[T/U]$.*

2.8 Méthodes de calcul d'un estimateur

Il s'agit de trouver un estimateur de θ, qui maximise la fonction de vraisemblance de l'échantillon. La valeur $\hat{\theta}$ qui maximise $L(x,\theta)$ serait un bon estimateur car elle donne la plus grande probabilité pour cet échantillon.

Cette méthode consiste à résoudre :

$$\sigma(s,i) = \begin{cases} \frac{\partial L(x,\theta)}{\partial \theta} = 0 & \text{pour trouver } \hat{\theta} \\ \frac{\partial^2 L(x,\theta)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}} < 0 & \text{pour assurer l'existence du} \max_{\theta\in\Theta} L(x,\theta) \end{cases}$$

Exercice 2.8.1. *Soit $(X_1, X_2, \dots, X_n)$ un n-échantillon issu de X*

Déterminer l ?estimateur EMV de μ et σ^2 ? .

Propriété 2.8.1. *S'il existe une statistique exhaustive U alors l'EMV en dépend.*

Propriété 2.8.2. *Caractéristiques de l'EMV*

1. *$L(x,\theta)$ n'a aucune raison d'être différentiable en θ.*
2. *L'EMV n'est pas forcement sans biais, donc pas forcement efficace.*
3. *L'EMV n ?est pas unique.*

Propriété 2.8.3. *Si $\hat{\theta}$ est l'etimateur du MV de θ alors $f(\hat{\theta})$ est l'EMV de $f(\theta)$.*

Chapitre Trois

Estimation par intervalle de confiance

Dans le chapitre précédent, à partir d'un échantillon aléatoire, on a cherché à construire de "bons" estimateurs pour le paramètre θ inconnu d'une population ; on a cherché en particulier à obtenir un estimateur T non biaisé, c'est-à-dire un estimateur qui donne en moyenne comme estimation la vraie valeur du paramètre θ inconnu. Cependant même si T est non biaisé, puisqu'il est une variable aléatoire, il est presque certain que, pour une réalisation particulière de l'échantillon, T fournira une estimation qui différera légérement de θ . Ainsi, X est un estimateur non biaisé de μ mais, pour une réalisation particulière de l'échantillon, on ne peut pas affirmer que μ soit exactement égal à $\bar{x}$. Alors puisqu'on ne peut être certain que l'estimation ponctuelle obtenue pour un paramètre est exacte, on apportera un complément d'information en construisant un intervalle autour de l'estimateur, intervalle dans lequel θ aura de grandes chances d'être inclus. Si l'on connaît la distribution de probabilité de l'estimateur utilisé, on pourra calculer la probabilité que cet intervalle aléatoire, appelé **intervalle de confiance**, englobe la vraie valeur du paramétre θ . On sera ainsi en mesure d'exprimer explicitement la marge d'erreur associée à l'utilisation d'un estimateur ponctuel de θ .

3.1 Obejctifs

On veut construire un intervalle aléatoire qui contiendra la valeur du paramètre θ avec une probabilité donnée. Cette probabilité (que l'on prendra assez près de 1) sera désignée par $1-\alpha$, et sera appelée le niveau de confiance de l'intervalle. La probabilité complémentaire α mesure le risque d'erreur de l'intervalle, c'est-á-dire la probabilité que l'intervalle aléatoire ne contienne pas la vraie valeur de θ . D'une façon générale, on doit résoudre une équation de la forme suivante :

$$\mathbb{P}(LI \leq \theta \leq LS) = 1-\alpha \tag{3.1}$$

où $\alpha =$ paramètre à estimer,
$LI =$ limite inférieure de l'intervalle de confiance,
$LS =$ limite supérieure de l'intervalle de confiance,
$1-\alpha =$ niveau de confiance l'intervalle.

De l'équation (3.1) , on déduit un intervalle de la forme $[LI, LS]$ dont les limites LI et LS sont des fonctions de l'échantillon aléatoire $(X_1, \ldots, X_n)$. Cet intervalle sera appelé intervalle de confiance pour θ au niveau $1-\alpha$.

Pour construire cet intervalle de confiance, on commence par définir un intervalle qui contient une fonction $f(T,\theta)$ de T et de θ (où T est estimateur ponctuel pour θ) avec une probabilité $1-\alpha$. On choisit une statistique dont on connaît la distribution de probabilité. Définir cet intervalle pour $f(T, \theta)$ revient à écrire l'équation :

$$\mathbb{P}(k_1 \leq \theta \leq k_2) = 1-\alpha \tag{3.2}$$

où les constantes k_1 et k_2 de (3.2) sont déterminées par l'intermédiaire de la distribution de probabilité de la statistique $f(T, \theta)$. La plupart du temps, le risque d'erreur α est divisé en deux parties égales à $\frac{\alpha}{2}$, et est réparti à chaque extrémité de la distribution de $f(T, \theta)$. Si, par exemple, la statistique $f(T, \theta)$ suit une distribution normale centrée réduite, les constantes k_1 et k_2 seront symétriques et pourront être désignées par $-z_{\frac{\alpha}{2}}$ et $z_{\frac{\alpha}{2}}$ comme l'indique la figure ci dessous. On obtiendra alors un intervalle de confiance bilatéral symétrique.

Exemple 3.1.1. *Un manufacturier de peinture veut estimer le temps moyen de séchage d'une nouvelle peinture d'intérieur qu'il désire mettre sur le marché. Le temps de séchage de cette peinture est une variable aléatoire X qui se distribue selon une loi normale. Supposons de plus qu'il connaisse l'écart type σ de ce temps de séchage (prenons $\sigma = 10$ minutes). Pour estimer μ , le manufacturier peint 25 surfaces de même taille et, pour ces 25 surfaces, il obtient un temps moyen de séchage $\bar{x} = 65$ minutes. Construire un intervalle de confiance au niveau de confiance $1-\alpha = 0.95$ pour la moyenne μ du temps de séchage de cette peinture*

3.2 Intervalles de confiance classiques

Maintenant que l'on a procédure générale utilisée pour construire un intervalle de confiance, on va construire explicitement des intervalles de confiance au niveau $1-\alpha$ pour les paramètres auxquels on s'intéresse le plus souvent dans une étude statistique. Pour construire ces intervalles de confiance, on utilise les distributions d'échantillonnage présentées dans le chapitre précédent

3.2.1 Intervalle de l'espérance μ

Dans la construction d'un intervalle de confiance pour la moyenne μ d'une population, on distingue deux cas : celui où la variance σ^2 de la population est connue, et celui où σ^2 est inconnue.

1. <u>Cas où σ^2 est connue</u>
   Pour estimer le paramètre $\theta = \mu$, on est naturellement porté à utiliser la statistique $T = \overline{X}$, la moyenne de l'échantillon. Si σ^2 est connue, on prend dans l'équation (3.2) la fonction

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

   Si X est normale, ou encore si X suit une distribution quelconque, mais n est assez grand $(n \geq 30)$, alors cette statistique $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ suit soit exactement, soit approximativement une distribution normale centrée réduite. Pour obtenir un intervalle de confiance pour μ au niveau $(1-\alpha)$, on écrit une équation de la forme (3.2) à savoir

$$\mathbb{P}\Big(-u_{\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq u_{\frac{\alpha}{2}}\Big)$$

dans laquelle les valeurs $\pm\frac{z}{2}$ sont lues dans la table de la loi normale centrée réduite. En isolant μ dans cette équation, on obtient une équation de la forme (3.1) à savoir :

$$\mathbb{P}\Big(\overline{X} - \frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}\Big) = 1 - \alpha$$

En résumé, on a donc : L'intervalle de confiance pour μ au niveau $(1?\alpha)$, dans le cas où la variance σ^2 de la population est connue, lorsque cette population est normale, ou encore lorsque la taille n de l'échantillon est assez grand $(n \geq 30)$, est de la forme $IC_{(1-\alpha)}(\mu) = \left[\overline{X} \pm \frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}\right]$ où $z_{\frac{\alpha}{2}}$ est une valeur de la table de la loi normale centrée réduite.

L'exemple 1 illustre cet intervalle de confiance dans le cas d'une population normale, pour un niveau de confance 95% . La manufacturier aurait pu utiliser un intervalle de confiance de la même forme sans faire l'hypothèse de normalité de la population, mais il aurait dû utiliser un échantillon de taille n plus grand que 25 pour s'assurer que l'approximation normale soit suffisament bonne.

2. Cas2 : σ^2 inconnue, population normale ou population quelconque avec n grand

   La plupart du temps, lorsque dans la population, μ est inconnue, σ^2 l'est aussi. Dans ces cas-là, pour estimer le paramètre $\theta = \mu$ par intervalle, on ne peut plus utiliser la statistique $(\overline{X} - \mu)/(\frac{\sigma}{\sqrt{n}})$. On utilise alors dans l'équation (3.2) la fonction :

$$f(T, \theta) = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

   l'écart type S de l'échantillon remplaçant l'écart type σ de la population. Si la population est normale, on a vu au chapitre sur l'échantillonnage que la statistique $(\overline{X} - \mu)/(S/n)$ suit une distribution du t à $n - 1$ degrés de liberté. Dans ce cas, l'équation (3.2) prend la forme :

$$\mathbb{P}\Big(-t_{\frac{\alpha}{2}} \leq \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \leq t_{\frac{\alpha}{2}}\Big)$$

   dans laquelle les valeurs symétriques $\pm t_{\frac{\alpha}{2}}2$ sont lues dans la table de la loi t à $n - 1$ degrés de liberté. Par la suite, en isolant μ dans cette équation, on obtient

$$\mathbb{P}\Big(\overline{X} - \frac{\sigma}{\sqrt{n}}t_{\frac{\alpha}{2}} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}}t_{\frac{\alpha}{2}}\Big) = 1 - \alpha$$

   Si n est assez grand $(n \geq 30)$, la distribution du t peut être approximée par la distribution normale, d'où on peut remplacer dans cette dernière équation les valeurs $\pm t_{\frac{\alpha}{2}}$ par les valeurs $\pm z_{\frac{\alpha}{2}}$. De plus, si la distribution de la population n'est pas normale, lorsque n est assez grand $(n \geq 30)$, le théorème central limite s'applique, et la distribution de la statistique $(\overline{X} - \mu)/(S/n)$ est approximativement normale. En résumé, on a donc : L'intervalle de confiance pour μ au niveau $(1 - \alpha)$, dans le cas où la variance σ^2 de la population est inconnue

   — lorsque cette population est normale, est de la forme $IC_{(1-\alpha)}(\mu) = \left[\overline{X} \pm \frac{\sigma}{\sqrt{n}}t_{\frac{\alpha}{2}}\right]$ où $t_{\frac{\alpha}{2}}$ est une valeur de la distribution du t à $(n - 1)$ degrés de liberté, et

— lorsque la taille n de l'échantillon est assez grande ($n \geq 30$) , que la population soit normale ou non, est de la forme $IC_{(1-\alpha)}(\mu) = \left[\overline{X} \pm \frac{\sigma}{\sqrt{n}} u_{\frac{\alpha}{2}}\right]$ où $u_{\frac{\alpha}{2}}$ est une valeur de la table de la loi normale centrée

Exemple 3.2.1. *Reprenons l'exemple concernant le temps de séchage d'une nouvelle peinture, mais en faisant l'hypothèse plus réaliste que le manufacturier, en plus d'ignorer la moyenne μ , ignore aussi la variance σ^2 . Supposons de nouveau que, pour estimer μ , le manufacturier ait tiré un échantillon aléatoire de taille* 25 *, et qu'il ait obtenu un temps de séchage moyen $\bar{x} = 65$ minutes, avec un écart type $s = 15$ minutes. En supposant de nouveau que le temps de séchage est distribué normalement, estimer μ à l'aide d'un intervalle de confiance au niveau* 95% *.*

3.2.2 Intervalle de confiance pour une proportion p

Souvent, on désire estimer dans une population la proportion p des unités qui possèdent une certaine carctéristique. Dans ce cas, la variable étudiée dans la population est une variable de Bernoulli de paramètre p. Pour construire un intervalle de confiance pour p, on distingue deux cas : n grand ($n \geq 30$), et n petit($n < 30$).

1. Cas1 : n grand

   Si X est une variable de Bernoulli de paramètre p, la moyenne X de l'échantillon (qui exprime la proportion des succès dans n tirages) est un estimateur sans biais de p. Si n est assez grand, on sait que X suit approximativement une distribution normale de moyenne p et de variance pq/n, où $q = 1 - p$. En conséquence, écrire une équation de la forme (3.2) , on va recourir à la statistique

$$f(T, \theta) = \frac{\overline{X} - p}{\sqrt{pq/n}}$$

   On obtient alors l'équation :

$$\mathbb{P}\Big(-u_{\frac{\alpha}{2}} \leq f(T, \theta) = \frac{\overline{X} - p}{\sqrt{pq/n}} \leq u_{\frac{\alpha}{2}}\Big)$$

   dans laquelle la valeur $-u_{\frac{\alpha}{2}}$ est lue dans la table de la normale centrée réduite. En isolant pq/n partiellement p dans cette équation, on obtient

$$\mathbb{P}\Big(\overline{X} - u_{\frac{\alpha}{2}}\sqrt{pq/n} \leq p \leq \overline{X} + u_{\frac{\alpha}{2}}\sqrt{pq/n}\Big) = 1 - \alpha$$

   Malheureusement, dans cette dernière équation, les limites de l'intervalle de con ? ?ance pour p sont données en fonction de la proportion p inconnue. Pour contourner cette difficulté, on remplace dans les expressions impliquant une racine carrée, p par X, et q par $(1 - X)$. En procédant ainsi, on introduit une nouvelle source d'erreur, mais si n est grand, cette erreur est négligeable. On obtient alors l'équation

$$\mathbb{P}\Big(\overline{X} - u_{\frac{\alpha}{2}}\sqrt{\overline{X}(1-\overline{X})/n} \leq p \leq \overline{X} + u_{\frac{\alpha}{2}}\sqrt{\overline{X}(1-\overline{X})/n}\Big) = 1 - \alpha$$

   En résumé, on a donc : L'intervalle de confiance au niveau $(1-\alpha)$ pour le paramètre p d'une population de Bernoulli, lorsque la taille n de l'échantillon est grande ($n \geq 30$), est de la forme où $u_{\frac{\alpha}{2}}$ est une valeur de la distribution normale centrée réduite.

Exemple : On veut estimer la proportion des conducteurs d'automobiles qui portent leur cein- ture de sécurité. Sur un échantillon de 200 conducteurs observés à une intersection, on a noté

qu'il y en avait 130 qui portaient leur ceinture de sécurité. Construitre un intervalle de confiance au niveau 95% pour estimer la vraie proportion p des conducteurs qui portent leur ceinture de sécurité.

2. Cas2 : n petit(utilisation des abaques)

3.2.3 Intervalle de confiance pour une la variance σ^2 ou l'écart type σ d'une population normale

En plus de la moyenne, un autre paramètre que l'on cherche souvent à connaîttre dans une population, c'est sa variance σ^2 ou son écart type σ . On sait que la statistique $T = S^2$, la variance de l'échantillon, est un estimateur sans biais du paramètre $\sigma = \sigma^2$. Pour définir un intervalle de confiance pour σ^2, on utilise la

$$f(T,\theta) = f(S^2,\sigma^2) = \frac{(n-1)S^2}{\sigma^2}$$

qui suit une distribution du χ^2_{n-1} lorsque la population est normale. On peut alors écrire pour cette statistique une équation de la forme (3.2) à savoir

$$\mathbb{P}\Big(\chi^2_{1-\frac{\alpha}{2}} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2}}\Big) = 1-\alpha$$

Dans cette dernière équation, les constantes k_1 et k_2 ont été désignées par $\chi^2_{1-\frac{\alpha}{2}}$ et , $\chi^2_{\frac{\alpha}{2}}$ car ce sont deux valeurs non symétriques de la distribution χ^2_{n-1} : $\chi^2_{1-\frac{\alpha}{2}}$ désigne la valeur à droite de laquelle il y a une probabilité $1-(\frac{\alpha}{2})$ et $\chi^2_{\frac{\alpha}{2}}$ désigne la valeur à droite de laquelle il y a une probabilité $\frac{\alpha}{2}$. Ces valeurs sont représentées dans la figure ci-dessous.

En isolant σ^2 dans l'équation précédente, on obtient

$$\mathbb{P}\Big(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}\Big) = 1-\alpha$$

Pour obtenir un intervalle de confiance pour ? au niveau $(1?\alpha)$, il suffit d'extraire la racine carrée de chaque membre de l'inégalité dont on prend la probabilité dans l'équation précédente. En résumé, on a donc :

L'intervalle de confiance au niveau $(1-\alpha)$ pour la variance σ^2 d'une population normale (de moyenne inconnue) est de la forme

$$\Big[\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} ; \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}\Big]$$

et l'intervalle de confiance au niveau $(1?\alpha)$ pour la variance σ d'une population normale (de moyenne inconnue) est de la forme

$$\left[\sqrt{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}}}; \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}}\right]$$

Exemple

On s'intéresse au temps X nécessaire aux candidats pour répondre à un test écrit exigé pour l'obtention d'un permis de conduire. Pour un échqantillon aléatoire de 25 personnes qui ont passé ce test, on a obtenu un temps moyen de 57 minutes avec un écart type de $6,2$ minutes. Construire un intervalle de confiance au niveau 95% pour l'écart type σ du temps nécessaire pour compléter ce test (en supposant que le temps X suive une distribution normale).