Introduction à la Statistique Inférentielle

Thibault Modeste

Année scolaire 2024-2025

CY Tech - Campus de Pau

Table des matières

Chapitre 1

Modèle et estimation statistique

Contents

La statistique inférentielle est une branche de la statistique qui se concentre sur la déduction de propriétés d'une population à partir d'un échantillon. Contrairement à la statistique descriptive, qui se contente de résumer et d'organiser des données, la statistique inférentielle vise à tirer des conclusions, à faire des prédictions et à prendre des décisions basées sur des données observées.

1.1 Echantillon, statistique et estimateur

1.1.1 Notation

Définition 1.1. *Un modèle statistique pour n observations est un couple $(\mathcal{H}^n, \mathcal{Q})$ où $\mathcal{H}$ est un espace topologique où vit nos observations (par exemple $\mathbb{R}$, $\mathbb{Z}$ ou $\{0,1\}$) et $\mathcal{Q}$ est une famille de mesures de probabilités sur $(\mathcal{H}^n, \mathcal{B}(\mathcal{H}^n))$ de la forme $\mathcal{Q} = (Q_\theta)_{\theta\in\Theta}$ et parmi laquelle on cherche la loi inconnue Q^* régissant nos n observations.*

Remarque 1.2. On se placera généralement dans le cadre de n observations indépendantes et identiquement distribuées (*i.i.d.*). Ainsi, nos mesures auront la forme

$$Q_\theta = \mathbb{P}_\theta^{\otimes n}, \text{ avec } \mathbb{P}_\theta \text{ une mesure de probabilité sur } \mathcal{H}.$$

On note $\mathcal{P}$ la famille $(\mathbb{P}_\theta)_{\theta\in\Theta}$.

Exemple 1.3. 1. On cherche à estimer une proportion inconnue p^*, e.g. proportion de boules rouges dans une urne, résultat d'un sondage, le modèle s'écrit alors dans le cas de n observations indépendantes

$$\mathcal{H} = \{0,1\} \text{ et } \mathcal{P} = (\mathcal{B}(p))_{p\in[0,1]}.$$

2. On observe des réalisations indépendantes de durées de vie d'ampoules électriques, même marque et même modèle. On modélise généralement cette durée de vie par une loi exponentielle $\mathcal{E}(\lambda)$ avec $\lambda > 0$. On rappelle que cette loi est caractérisée par la propriété d'absence de mémoire. Dans ce cas le modèle s'écrit

$$(\mathbb{R}_+^n, (\mathcal{E}(\lambda)^{\otimes n})_{\lambda>0}).$$

3. Un professeur, pour gagner du temps, note aléatoirement et de manière uniforme ses élèves entre 0 et 20. Il n'aime pas mettre des 20/20. On notera par θ^* la note maximale qu'il accepte de mettre. Ici, le modèle statistique est

$$\mathcal{H} = [0, 20] \text{ et } \mathbb{P}_\theta = \mathcal{U}([0, \theta]) \text{ avec } \theta \in [0, 20[.$$

Définition 1.4.
— *Un modèle statistique est dit identifiable si la fonction $\theta \mapsto Q_\theta$ est injective, i.e. si $\theta \neq \nu$ alors $Q_\theta \neq Q_\nu$.*
— *le modèle est dit paramétrique si l'ensemble des paramètres Θ est inclus dans $\mathbb{R}^d$ pour $d \in \mathbb{N}$.*

Dans la suite, on se placera dans ces situations. L'identifiabilité d'un modèle permet de confondre le paramètre θ avec la mesure de probabilité Q_θ associée. Le cadre paramètrique signifie que l'on se place dans un cadre où l'on estime qu'un nombre fini de paramètres.

1.1.2 Estimation d'un paramètre

Comme dit en introduction, le but de la statistique fréquentiste est de déterminer à l'aide des observations un paramètre θ^* inconnu et fixe. Selon le cadre, il est plus pertinent d'estimer non pas θ^* mais $g(\theta^*)$ avec g une fonction. Dans ce cours, on restera dans le cadre simple où g vaut l'identité, i.e. on estime uniquement θ^*. Pour estimer ce paramètre, nous avons à notre disposition

Définition 1.5. *Un échantillon de loi Q_θ est le vecteur aléatoire canonique $(X_1, \ldots, X_n)$ sur $\mathcal{H}^n$,*

$$X_i \colon (x_1, \ldots, x_n) \mapsto x_i.$$

Le vecteur $(X_1, \ldots, X_n)$ a pour loi Q_θ lorsque l'on munit $\mathcal{H}^n$ de la loi Q_θ.

Cette définition peut paraître lourde, et c'est le cas. Mais ce formalisme permettra l'écriture rigoureuse de futur résultat. Avec ce formalisme, si on se place dans le cadre *i.i.d.*, les questions que l'on se posera serait du genre *est-ce que nos observations $(x_1, \ldots, x_n)$ peuvent provenir de $(X_1, \ldots, X_n)$ lorsque l'on munit $\mathcal{H}$ de $\mathbb{P}_\theta$?*

Définition 1.6.
— *Une statistique est une variable aléatoire sur $\mathcal{H}^n$.*
— *Un estimateur de θ^* est une statistique qui s'exprime indépendamment de θ et à valeur dans Θ.*

Exemple 1.7. On se place dans le modèle $(\{0, 1\}^n, (\mathcal{B}(p)^{\otimes n})_{p\in[0,1]})$. La variable $\bar{X}_n := (X_1 + \ldots + X_n)/n$ est un estimateur alors que $Z_n = 0.5p + 0.5\bar{X}_n$ n'est pas un estimateur.

1.2 Propriétés d'un estimateur

1.2.1 Biais d'un estimateur

Pour S une statistique, on notera $\mathbb{E}_\theta[S]$ l'espérance de S lorsque $\mathcal{H}^n$ est muni de Q_θ.

Définition 1.8. *Soit $\hat{\theta}_n$ un estimateur de θ^*. On appelle biais de $\hat{\theta}_n$ la fonction*

$$b_n(\theta) := \mathbb{E}_\theta[\hat{\theta}_n] - \theta.$$

On dit que l'estimateur $\hat{\theta}_n$ est sans biais si pour tout $\theta \in \Theta$, $b_n(\theta) = 0$, et asymptotiquement sans biais si

$$\forall \theta \in \Theta, b_n(\theta) \to 0.$$

Remarque 1.9. La définition se fait pour tout $\theta \in \Theta$ et non uniquement pour le vrai paramètre θ^*. En effet, comme on ne connait pas sa valeur, c'est bien d'avoir un estimateur ayant une bonne propriété quelque soit *le monde dans lequel on est.*

Exemple 1.10.

1. Soit le modèle $(\mathcal{H}^n, (\mathbb{P}_\theta^{\otimes})_{\theta\in\Theta})$, supposons que pour $\theta \in \Theta$,

$$\mathbb{E}_\theta[X_1] = \theta,$$

i.e. le paramètre est le moment d'ordre 1 de la mesure $\mathbb{P}_\theta$, alors l'estimateur $\bar{X}_n$ est sans biais. En effet, par linéarité de l'espérance

$$\mathbb{E}_\theta[\bar{X}_n] = \mathbb{E}_\theta\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta[X_i] = n\theta/n = \theta.$$

Dans la suite, la moyenne empirique fera référence à cet estimateur.

2. Prenons le modèle $(\mathbb{R}^n, (\mathcal{N}(\mu, \sigma^2)^{\otimes n})_{\mu\in\mathbb{R},\sigma^2\in\mathbb{R}_+^*})$. Supposons que l'on souhaite estimer uniquement la variance. Considérons l'estimateur suivant

$$S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

On peut montrer que

$$\mathbb{E}_{(\mu,\sigma^2)}[S_n^2] = \frac{n-1}{n}\sigma^2,$$

donc cet estimateur est un estimateur biaisé de la variance. Pour être plus précis, on sous estime la vrai variance en utilisant S_n^2. Ce fait est assez naturel car la dispersion des X_i n'est pas estimé à partir de la vraie moyenne μ mais avec l'estimation $\bar{X}_n$ qui dépend déjà des X_i. Toutefois, cet estimateur est non biaisé de manière asymptotique. Selon le cadre, on pourra considérer deux estimateurs différentes de la variance, la variance empirique classique S_n^2 et la variance empirique non biaisée

$$S_{n,c}^n = \frac{n}{n-1}S_n^2.$$

Définition 1.11. *Soit $\hat{\theta}_n$ un estimateur, on appelle risque quadratique de $\hat{\theta}_n$ sous Q_θ*

$$\mathcal{R}(\hat{\theta}_n; \theta) = \mathbb{E}_\theta[\|\hat{\theta}_n - \theta\|^2].$$

Le risque quadratique correspond à l'erreur quadratique moyenne lorsque l'on estime θ par $\hat{\theta}_n$. Cette quantité possède une réécriture beaucoup plus simple à calculer.

Proposition 1.12 (décomposition biais/variance)**.** *Soit $\hat{\theta}_n$ un estimateur, on a*

$$\mathcal{R}(\hat{\theta}_n; \theta) = \|\mathbb{E}_\theta[\hat{\theta}_n] - \theta\|^2 + \mathrm{Var}_\theta(\hat{\theta}_n),$$

où la variance d'un vecteur est

$$\mathrm{Var}_\theta(\hat{\theta}_n) = \mathbb{E}_\theta[\|\hat{\theta}_n - \mathbb{E}_\theta[\hat{\theta}_n]\|^2]$$

On remarque donc que pour avoir un risque quadratique faible, il faut avoir un biais faible ainsi qu'une variance faible, i.e. centré autour du paramètre que l'on essaye d'estimer et une incertitude faible autour de l'estimation. Le risque est un bon critère intuitif pour comparer des estimateurs.

Exemple 1.13. 1. Soit $X_1, \ldots, X_7$ un échantillon suivant une loi normale de moyenne μ et de variance σ^2 inconnues. Le problème est d'estimer μ. On considère les deux estimateurs suivants

$$\hat{\mu}_1 = \frac{1}{7}(X_1 + \ldots, X_7) \text{ et } \hat{\mu}_2 = \frac{1}{2}(X_1 - X_3 + 2X_5).$$

On peut vérifier que ces deux estimateurs sont sans biais. Mais y-a-t-il un plus performant ? En calculant la variance de chaque estimateur, on trouve que

$$\hat{\mu}_1 \sim \mathcal{N}(\mu, \sigma^2/7) \text{ et } \hat{\mu}_2 \sim \mathcal{N}(\mu, 3\sigma^2/2).$$

Ainsi $\hat{\mu}_1$ est plus concentré autour de μ que $\hat{\mu}_2$.

2. Dans le modèle du Pile-Face $(\{0,1\}^n, \mathcal{B}(p)^{\otimes n})$ avec $p \in (0,1)$, prenons l'estimateur de la moyenne empirique $\hat{p}_n = \bar{X}_n$ pour estimer le paramètre p^*. D'après le point 1 de l'Exemple 1.10, cet estimateur est sans biais et sa variance vaut par l'indépendance des observations

$$\begin{aligned}\mathrm{Var}_p(\hat{p}_n) = \mathrm{Var}_p\left(\frac{1}{n}\sum_{i=1}^n X_i\right) &= \frac{\mathrm{Var}_p\left(\sum_{i=1}^n X_i\right)}{n^2} \\ &= \frac{\sum_{i=1}^n \mathrm{Var}_p(X_i)}{n^2} = \frac{p(1-p)}{n},\end{aligned}$$

car la variance d'une variable de loi $\mathcal{B}(p)$ est $p(1-p)$. Ainsi par la décomposition biais/variance de l'erreur quadratique, on en déduit que

$$\mathcal{R}(\hat{p}_n; p) = 0^2 + \frac{p(1-p)}{n}.$$

On remarque que l'erreur diminue avec l'augmentation du nombre d'observations.

Définition 1.14. — *On dit que $\hat{\theta}_n^{(1)}$ est préférable à un autre estimateur $\hat{\theta}_n^{(2)}$ si*

$$\forall \theta \in \Theta,\ \mathcal{R}(\hat{\theta}_n^{(1)}; \theta) \leqslant \mathcal{R}(\hat{\theta}_n^{(2)}; \theta).$$

— *On dit qu'un estimateur sans biais $\hat{\theta}_n^{(1)}$ est de variance uniformément minimale parmi les estimateurs sans biais (VUMSB) si l'estimateur est préférable à tout autre estimateur.*

Proposition 1.15. *Soit deux estimateurs $\hat{\theta}_n^{(1)}$ et $\hat{\theta}_n^{(2)}$ VUMSB alors*

$$\forall \theta \in \Theta,\ \hat{\theta}_n^{(1)} = \hat{\theta}_n^{(2)},\ \mathbb{P}_\theta - p.s.$$

1.2.2 Consistance d'un estimateur

Définition 1.16. *On dit qu'un estimateur $\hat{\theta}_n$ est consistance, ou convergeant, si*

$$\forall \theta \in \Theta,\ \hat{\theta}_n \xrightarrow{\mathbb{P}_\theta} \theta,$$

et fortement consistant lorsque

$$\forall \theta \in \Theta,\ \hat{\theta}_n \xrightarrow{\mathbb{P}_\theta - p.s.} \theta.$$

La Loi des Grands Nombres (LGN) est un résultat majeur en Probabilité pour obtenir la forte consistance de nos estimateurs.

Exemple 1.17. Dans le cadre du point 1 de l'Exemple 1.10, la moyenne empirique est fortement consistance d'après la LGN.

Définition 1.18. *On dit qu'un estimateur $\hat{\theta}_n$ est de vitesse $(\nu_n)_{n\in\mathbb{N}}$ avec $(\nu_n)_n$ une suite strictement croissante de réels positifs qui tend vers $+\infty$ si pour tout $\theta \in \Theta$, il existe une loi $l(\theta) \neq \delta_c$ tel que*

$$\nu_n(\hat{\theta}_n - \theta) \rightsquigarrow l(\theta).$$

Cette fois-ci, c'est le Théorème Central Limite (TCL) qui permettra d'obtenir des vitesses de certains estimateurs. Par exemple, d'après le TCL, la vitesse de la moyenne empirique est $(\sqrt{n})_n$. On peut interpréter la vitesse comme le développement asymptotique de l'erreur lorsque l'on estime le paramètre θ par l'estimateur $\hat{\theta}_n$. Par exemple, le TCL nous dit que

$$\sqrt{n}(\bar{X}_n - \mu) \rightsquigarrow \mathcal{N}(0, \sigma^2),$$

où μ est la moyenne que l'on essaye d'estimer et σ^2 la variance que l'on connait. On peut interpréter ce théorème de la manière suivante : lorsqu'on estimer μ par $\bar{X}_n$ pour n grand, la LGN nous dit que cette estimation est *proche* du vrai paramètre, mais que veut dit *proche* ? Ici, l'écart entre l'estimation et le vrai paramètre est de l'ordre de $\varepsilon/\sqrt{n}$ avec $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Proposition 1.19. *Si un estimateur $\hat{\theta}_n$ possède une vitesse alors il est consistant.*

1.2.3 Robustesse

De nombreux autres critères existent pour mesurer la qualité d'un estimateur. Pour finir cette section, nous parlons rapidement d'une notion importante : la robustesse. Elle peut être définie intuitivement comme la faible sensibilité à des déviations des hyptohèses de bases, à des valeurs aberrantes. Nous allons voir en TD un exercice donnant un cadre plus théorique à cette définition.

Exemple 1.20. Considérons l'échantillon suivant : $1, 2, 3, 4, 4$. Supposons qu'il y ait une erreur de saisie, par exemple une mauvaise unité, et que l'on considère l'échantillon $1, 2, 3, 4000, 4$. La moyenne empirique passe de 2.8 à 802, ainsi la moyenne n'est pas robuste face à cette erreur. Alors que la médiane n'est pas embêtée par ce problème.

1.3 Famille d'estimateurs

1.3.1 Estimateur par la méthode des moments

Il s'agit d'une méthode simple et classique qui fournit des estimateurs de façon immédiate lorsque les moments des mesures de probabilité s'écrivent en fonction des paramètres à estimer. Ces estimateurs ne seront pas forcément très performant mais héritera tout de même des propriétés de consistances des moments empiriques. Cette méthode fonctionnera très bien dans un modèle Gaussie, loi de Poisson, loi exponentielle, loi géométrique, loi Bernoulli, loi Gammma,..., car les différents paramètres de ces lois s'écrivent en fonction des moments.

Notation : Étant donné un échantillon $(X_1, \ldots, X_n)$ issu d'un modèle $(\mathcal{H}^n, (\mathbb{P}_\theta^{\otimes n})_\theta)$, on considère les quantités suivantes lorsqu'elles existent

- moment théorique d'ordre p : $m_p(\theta) = \mathbb{E}_\theta[X_1^p]$
- moment théorique centré d'ordre p : $\mu_p(\theta) = \mathbb{E}_\theta[(X_1 - \mathbb{E}_\theta[X_1])^p]$
- moment empirique d'ordre p : $U_p(n) = \frac{1}{n}\sum_{i=1}^n X_i^p$
- moment empirique centré d'ordre p : $W_p(n) = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^p$

Définition 1.21. *On appelle estimateur de $\theta \in \Theta$ par la méthode des moments (EMM) la solution $\hat{\theta}_{MM}$, quand elle existe et est unique, d'un* ***sous-système*** *du système suivant*

$$\begin{cases} U_1(n) = m_1(\hat{\theta}_{MM}) \\ U_2(n) = m_2(\hat{\theta}_{MM}) \\ W_2(n) = \mu_2(\hat{\theta}_{MM}) \\ \vdots \\ W_p(n) = \mu_p(\hat{\theta}_{MM}) \end{cases}.$$

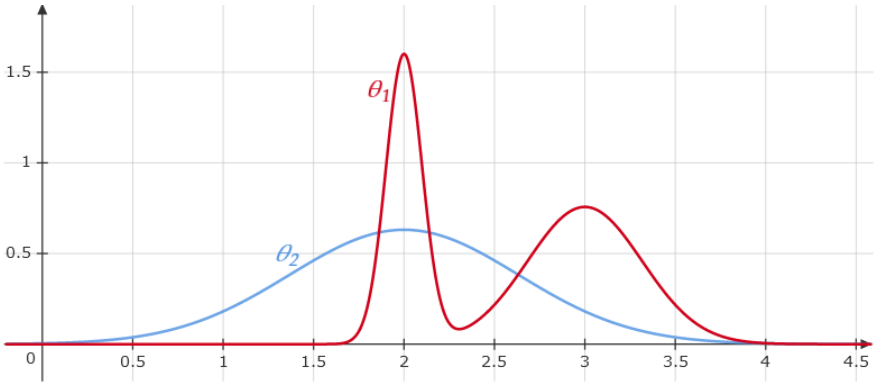
Exemple 1.22.

1. Prenons le modèle $(\mathbb{R}_+^n, (\mathcal{E}(\theta)^{\otimes n})_{\theta>0})$, on a
$$m_1(\theta) = \frac{1}{\theta}.$$
Pour trouver l'estimateur par la méthode des moments, on remplace les moments théoriques (dans cet exemple, il n'y en a qu'un seul) et on remplace les paramètres par les estimateurs, on obtient donc
$$U_1(n) = \frac{1}{\hat{\theta}_{MM}}.$$
Ainsi en remplaçant $U_1(n)$ par sa forme plus usuelle, on obtient $\hat{\theta}_{MM} = 1/\bar{X}_n$.
2. Prenons un modèle avec deux paramètres à estimer, $(\mathbb{R}_+^n, (\gamma(a,b)^{\otimes n})_{a,b>0})$, où $\gamma(a,b)$ est une loi dont la densité vaut
$$\forall x \in \mathbb{R},\ f(x;a,b) = x^{a-1}\frac{b\mathrm{e}^{-bx}}{\Gamma(a)}\mathbb{1}_{x>0}.$$
On peut montrer que
$$\begin{cases} m_1(a,b) = a/b \\ \mu_2(a,b) = a/b^2 \end{cases}.$$
Ainsi en remplaçant les moments théoriques (m_1, μ_2) par leurs versions empiriques $(\bar{X}_n, S_n^2)$, on doit résoudre le système
$$\begin{cases} \bar{X}_n = \hat{a}_{MM}/\hat{b}_{MM} \\ S_n^2 = \hat{a}_{MM}/\hat{b}_{MM}^2 \end{cases}.$$
On trouve après calcul, $\hat{a}_{MM} = \frac{\bar{X}_n^2}{S_n^2}$ et $\hat{b}_{MM} = \frac{\bar{X}_n}{S_n^2}$.

Proposition 1.23. *L'estimateur $\hat{\theta}_{MM}$ est fortement consistant et admet une limite.*

1.3.2 Estimation par la vraisemblance

Prenons le cadre simpliste où $\Theta = \{\theta_1, \theta_2\}$ et où les deux mesures du modèle $\mathbb{P}_{\theta_1}$, $\mathbb{P}_{\theta_2}$ sont à densité. Imaginons que, nous observons après une mesure la valeur 2. Dans ce cas-ci, intuitivement quel paramètre choisit-on ?



Idée : On estime θ^*, ou $\mathbb{P}_{\theta^*}$, par le paramètre associé au modèle rendant le plus crédible nos observations. Dans cet exemple, on choisirait $\hat{\theta} = \theta_1$, car la densité de la mesure $\mathbb{P}_{\theta_1}$ est plus importante en 2. **Attention**, on ne choisit pas le modèle le plus crédible car d'un point de vue fréquentiste, ça n'a pas de sens ! Mais bien le modèle où l'on a le plus de chance de voir notre observation.

Dans la suite, on adoptera une notation commune pour le cas continu et discret. Pour une variable aléatoire X, on notera par $f(x;\theta)$

— pour le cas discret : $f(x;\theta) = \mathbb{P}_\theta(X_1 = x)$

— dans le cas continue : $f(x;\theta)$ est la densité de la mesure $\mathbb{P}_\theta$.

Ainsi pour $\mu = \eta$ (mesure de comptage) ou $\mu = \lambda$ (mesure de Lebesgue), on a pour un borélien A,

$$\mathbb{P}_\theta(X_1 \in A) = \int_A f(x;\theta)\ \mathrm{d}\mu(x).$$

Définition 1.24. *Dans le cas où $Q_\theta = \mathbb{P}_\theta^{\otimes n}$, on appelle **vraisemblance du modèle**, la fonction pour $x_1, \dots, x_n \in \mathcal{H}$,*

$$L(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i;\theta).$$

Exemple 1.25. Prenons $n = 3$ et $\mathcal{H} = \{0, 1\}$ avec $\mathbb{P}_p \sim \mathcal{B}(p)$ pour $p \in \{1/3, 2/3\}$. On observe les résultats $0, 0, 1$, quelle est la vraisemblance de cette observation ? On calcule

$$Q_{1/3}(X_1 = 0, X_2 = 0, X_3 = 1) = \frac{2}{3}\frac{2}{3}\frac{1}{3} = \frac{4}{27} \text{ et } Q_{2/3}(X_1 = 0, X_2 = 0, X_3 = 1) = \frac{2}{27}.$$

Définition 1.26. *Un estimateur du maximum de vraisemblance (EMV) est un estimateur $\hat{\theta}_n$ vérifiant*

$$L_n(X_1, \dots, X_n; \hat{\theta}_n) = \sup_{\theta \in \Theta} L_n(X_1, \dots, X_n; \theta).$$

Remarque 1.27.

— Ni l'existence, ni l'unicité de l'EMV est garanti dans un modèle quelconque. Le maximum peut avoir une forme explicite mais il est parfois nécessaire de recourir à des méthodes d'optimisation numérique pour le déterminer, par exemple avec une descente de gradient.

— En pratique, on considère la log-vraisemblance car il est plus simple d'étudier une somme qu'un produit. Dans la suite, on notera log pour le logarithme népérien.

Exemple 1.28.

1. Dans l'Exemple 1.25, l'estimateur de maximum de vraisemblance vaut $1/3$.
2. Dans le modèle plus général du Pile-Face $(\{0,1\}^n, \mathcal{B}(p)^{\otimes n})$ avec $p \in (0,1)$. Dans ce cas là, la vraisemblance pour $x_1, \dots, x_n \in \{0,1\}$ et $p \in (0,1)$,

$$L_n(x_1, \dots, x_n; p) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}.$$

Cette fonction est plus facile à étudier en passant au logarithme, et fixons les observations et considérons la fonction

$$h(p) = \log L_n(x_1, \dots, x_n; p) = n\bar{x}_n \log p + n(1-\bar{x}_n)\log(1-p),$$

où $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$. L'objectif est de maximiser cette fonction. Dérivons la fonction,

$$h'(p) = \frac{n\bar{x}_n}{p} - \frac{n(1-\bar{x}_n)}{1-p} = \frac{n(\bar{x}_n - p)}{p(1-p)}.$$

Une étude du signe de la dérivée montre qu'un maximum est atteint en $p = \bar{x}_n$. Ainsi, l'EMV de ce modèle est $\hat{p}_{MV} = \bar{X}_n$. On peut vérifier qu'il s'agit aussi d'un estimateur obtenu par la méthode du moment.

Chapitre 2

Vraisemblance et information de Fisher

Contents

Dans le chapitre précédent, nous avons introduit un estimateur intuitivement intéressent, l'estimateur de maximum de vraisemblance. Il est plus difficile à l'oeil nu de déterminer si cet estimateur est performant, voir l'estimateur obtenu dans le point 3 de l'Exemple 1.28. Pour étudier leur performance, nous allons introduire une nouvelle quantité représentant l'information d'un modèle. Nous verrons ensuite si nos estimateurs utilisent bien cette information. On se place ici dans des modèles discret ou à densité.

2.1 Information d'un modèle

2.1.1 Divergence de Kullback-Leibler

Un outil important pour introduire l'information d'un modèle est la divergence de Kullback-Leibler.

Définition 2.1. *Supposons que* $\log L_n(\cdot;\alpha) \in \mathrm{L}^1(\mathbb{P}_\theta)$ *pour tout* $\alpha, \theta \in \Theta$. *La divergence de Kullback-Leibler entre les mesures* $\mathbb{P}_\alpha$ *et* $\mathbb{P}_\theta$ *est définie par*

$$\kappa_n(\alpha;\theta) = -\mathbb{E}_\theta\left[\log \frac{L_n(X_1;\alpha)}{L_n(X_1;\theta)}\right].$$

L'information de Kullback-Leibler est une mesure de dissimilarité entre deux mesures de probabilités. Si la quantité $\kappa(\alpha,\theta)$ est grande alors les vraisemblances du *monde* α et du *monde* θ sont très différentes. Ainsi les observations obtenues dans ces *mondes* ne se ressemblent pas, il est donc peu probable d'estimer α par θ avec la méthode du maximum de vraisemblance.

Exemple 2.2. Reprenons le modèle du Pile-Face de l'Exemple 1.28. Pour $\alpha, \theta \in (0,1)$,

$$\begin{aligned}\kappa_n(\alpha,\theta) &= -\mathbb{E}_\theta[n\bar{X}_n \log\frac{\alpha}{\theta} + n(1-\bar{X}_n)\log\frac{1-\alpha}{1-\theta}] \\ &= n\theta\log\frac{\alpha}{\theta} + n(1-\theta)\log\frac{1-\alpha}{1-\theta}.\end{aligned}$$

On peut vérifier que $\kappa_n(\alpha,\theta) = 0$ si et seulement si $\alpha = \theta$.

Proposition 2.3. *Pour $\alpha, \theta \in \Theta$, on a $\kappa_n(\alpha, \theta) \geqslant 0$ et si le modèle est identifiable alors*

$$\kappa_n(\alpha, \theta) = 0 \iff \alpha = \theta.$$

2.1.2 Information de Fisher

La divergence de Kullback-Leibler permet de quantifier la différence entre la vraisemblance du *monde* Q_α et le *monde* Q_θ. Plus ces vraisemblances sont différentes, plus les observations provenant de ces *mondes* sont différentes. Ainsi, si le vrai *monde* est Q_θ, il est peu probable d'observer un échantillon maximisant la vraisemblance $L_n(\cdot; \alpha)$. Une quantité utile est donc de savoir la vitesse de changement de $\theta \mapsto Q_\theta$. Pour $\theta \in \Theta$, on voudrait déterminer les variations de la fonction $K : \alpha \mapsto \kappa_n(\alpha, \theta)$. Sous des conditions de régularité, la formule de Taylor-Young nous donne que

$$\kappa_n(\alpha, \theta) = K''(\theta)(\alpha - \theta)^2 + o(\alpha - \theta)^2.$$

Dérivons la fonction K en s'autorisant toutes les opérations,

Définition 2.4. *Supposons que Θ est une ouvert et que $\nabla \log L_n(\cdot; \theta) \in \mathrm{L}^2(Q_\theta)$ pour chaque $\theta \in \Theta$. L'information de Fisher est la quantité*

$$\begin{aligned} I_n(\theta) &= \mathrm{Var}_\theta(\nabla \log L_n(X_1, \dots, X_n; \theta)) \\ &= (\mathrm{cov}_\theta(\partial_i \log L_n(X_1, \dots, X_n; \theta), \partial_j \log L_n(X_1, \dots, X_n; \theta)))_{1 \leqslant i,j \leqslant d} \end{aligned}$$

L'information de Fisher est donc une matrice permettant de mesure la courbure de la divergence de KL. Elle précise donc sa capacité à discriminer des mesures de probabilité. Dans le cas $d = 1$, une grande valeur de $I_n(\theta)$ traduit une variation importante de la divergence et donc une variation importante de la nature des mesures $(Q_\theta)_{\theta \in \Theta}$.

Exemple 2.5. Pour le cadre du Pile-Face, nous avons pour $x_1, \dots, x_n \in \{0, 1\}$,

$$L_n(x_1, \dots, x_n; p) = p^{n\bar{x}_n}(1 - p)^{n(1-\bar{x}_n)}.$$

Donc en passant au log et en dérivant

$$\nabla \log L_n(x_1, \dots, x_n; p) = \frac{n\bar{x}_n}{p} - \frac{n(1 - \bar{x}_n)}{1 - p} = \frac{n}{1 - p} + \frac{n}{p(1 - p)}\bar{x}_n.$$

On rappelle que la variance d'une variable $X \sim \mathcal{B}(p)$ est $p(1 - p)$, donc

$$I_n(\theta) = \frac{n^2}{p^2(1 - p)^2}\mathrm{Var}_\theta(\bar{X}_n) = \frac{n}{p(1 - p)},$$

car les observations $X_1, \dots, X_n$ sont indépendantes. Dans ce modèle, l'incertitude est faible pour p proche de 0 ou 1.

Proposition 2.6. *Soit I l'information de Fisher du modèle à une observation $(\mathcal{H}, (\mathbb{P}_\theta)_\theta)$, l'information de Fisher du même modèle avec n observations i.i.d. est*

$$I_n(\theta) = nI(\theta).$$

2.1.3 Modèle régulier

Définition 2.7. *Un modèle statistique est dit régulier si les propriétés suivantes sont vérifiées pour chaque $\theta \in \Theta$,*

1. *son information de Fisher est inversible, i.e. $I_n(\theta) \in \mathrm{GL}_d(\mathbb{R})$*
2. $\mathbb{E}_\Theta[\nabla \log L_n(X_1, \dots, X_n; \theta)] = 0$

3. $I_n(\theta) = -\mathbb{E}_\theta[\mathrm{Hess}(\log L_n(X_1, \dots, X_n; \theta))]$.

La plupart des modèles que l'on va considérer seront réguliers. En effet, nous travaillerons généralement avec des distributions provenant d'une famille de lois assez grandes.

Définition 2.8. *Un modèle $(\mathbb{R}, (\mathbb{P}_\theta)_\theta)$ est un modèle de la famille exponentielle si la densité par rapport à la mesure η ou λ est de la forme pour $x \in \mathbb{R}$ et $\theta \in \Theta$,*

$$f(x;\theta) = c(\theta)h(x)\exp(\sum_{i=1}^{m} \alpha_i(\theta)T_i(x)),$$

avec c, h, α_i et T_i des fonctions mesurables.

Cette famille comprend le modèle Bernoulli, Gaussien, Exponentielle, Gamma, Binomiale, Poisson... Par contre, la loi uniforme ne fait pas partie de cette famille !

Théorème 2.9. *Si la fonction $\theta \mapsto (\alpha_1(\theta), \dots, \alpha_m(\theta))$ est injective et deux fois différentiables, $(x \mapsto T_i(x))_{i=1}^m$ sont affinement indépendants et dans $\mathrm{L}^2(\mathbb{P}_\theta)$, alors le modèle de la famille exponentielle est régulier.*

On ne montrera jamais qu'un modèle régulier.

2.2 Propriétés remarquables de l'EMV

Nous voyons dans cette partie plusieurs théorèmes assez généraux sont la qualité des EMV. Cette section est hors-programme. On n'utilisera jamais les résultats dans cette partie pour montrer la consistance ou la vitesse de cet estimateur.

Théorème 2.10. *Soit $(\mathcal{H}^n, (\mathbb{P}_\theta^{\otimes n})_{\theta\in\Theta})$ un modèle identifiable avec Θ compact. Si l'EMV existe, il est consistant sous les conditions suivantes*

1. *pour tout $x \in \mathcal{H}$, $\log L_n(x;\cdot)$ est continue sur Θ ;*
2. *pour tout $\theta \in \Theta$, il existe un voisinage V de θ vérifiant*

$$\forall x \in \mathcal{H}, \forall \alpha \in V, \exists H \in \mathrm{L}^1(\mathbb{P}_\theta),\ |\log L_n(x;\alpha)| \leqslant H(x)$$

Les conditions pour vérifier ce théorème sont assez lourdes et souvent pas vérifiées en pratique, par exemple Θ compact est trop restrictif. Il est plus simple de vérifier à la main lorsque l'on a une forme explicite de l'EMV.

Théorème 2.11. *Supposons que $(\mathcal{H}^n, (\mathbb{P}_\theta^{\otimes n})_{\theta\in\Theta})$ un modèle régulier et que pour tout $\theta \in \Theta$, il existe un voisinage V de θ vérifiant*

$$\forall x \in \mathcal{H}, \forall \alpha \in V, \exists H \in \mathrm{L}^1(\mathbb{P}_\theta),\ |\log L_n(x;\alpha)| \leqslant H(x),$$

alors si l'EMV est consistant alors pour tout $\theta \in \Theta$,

$$\sqrt{n}(\hat{\theta}_{MV} - \theta) \rightsquigarrow \mathcal{N}(0, I(\theta)^{-1})$$

Attention, la normalité asymptotique de l'EMV n'est pas toujours vérifiée. Prenons l'exemple suivant basé sur un modèle non régulier.

Exemple 2.12. Pour le modèle uniforme $(\mathbb{R}_+^n, (\mathcal{U}([0,\theta])^{\otimes n})_{\theta>0})$, nous avons montré en TD que l'EMV était

$$\hat{\theta}_{MV} = \max(X_1, \dots, X_n).$$

Pour $t \in \mathbb{R}$ et $\theta > 0$, on a

$$Q_\theta(n(\hat{\theta}_{MV} - \theta) \leqslant t) = \left(1 + \frac{t}{n\theta}^{\ n}\right),$$

si $t \in [-n\theta, 0]$ et 0 sinon. Ainsi, l'EMV vérifie

$$n(\hat{\theta}_{MV} - \theta) \rightsquigarrow -\mathcal{E}(1/\theta).$$

2.3 Borne de Cramer-Rao

On se place dans le cas où $\Theta \subset \mathbb{R}$. On rappelle que la mesure μ fait référence soit à la mesure de comptage η soit la mesure de Lebesgue λ.

Définition 2.13. *L'estimateur $\hat{\theta}_n$ est dit régulier si*

$$\int_{\mathcal{H}^n} \hat{\theta}_n(x_1, \ldots, x_n) \partial_\theta L_n(x_1, \ldots, x_n; \theta) \, \mathrm{d}\mu = \partial_\theta \int_{\mathcal{H}^n} \hat{\theta}_n(x_1, \ldots, x_n) L_n(x_1, \ldots, x_n; \theta) \, \mathrm{d}\mu.$$

Le théorème suivant est le résultat le plus important de ce chapitre. Il permet de relier l'information de Fisher et l'erreur quadratique vue dans le chapitre précédent.

Théorème 2.14. *Pour un estimateur régulier et sans biais, on a pour $\theta \in \Theta$,*

$$\mathcal{R}(\hat{\theta}_n; \theta) \geqslant I_n(\theta)^{-1}.$$

Le minorant $I_n(\theta)^{-1}$ s'appelle *borne de Cramer-Rao*. Ainsi, l'erreur quadratique d'un estimateur ne pas être plus faible que cette borne. On voit que si l'information de Fisher est importante alors cette borne est très faible. Ce qui permet d'avoir potentiellement des estimateurs avec une faible erreur quadratique. Inversement, si l'information de Fisher est faible, il ne peut pas exister d'estimateur d'estimateur sans biais ayant une erreur quadratique négligeable.

Définition 2.15. *Un estimateur sans biais est dit efficace s'il atteint la borne de Cramer-Rao.*

Exemple 2.16. Dans le cas du modèle Pile-Face, nous avons déjà calculé l'erreur quadratique de la moyenne empirique, Exemple 1.13 et l'information de Fisher du modèle, Exemple 2.5. On avait trouvé pour $p \in (0, 1)$,

$$\mathcal{R}(\bar{X}_n; p) = \frac{p(1-p)}{n} \text{ et } I_n(p) = \frac{n}{p(1-p)}.$$

Comme la moyenne empirique est sans biais, on a bien que cet estimateur est efficace.

Chapitre 3

Intervalle de confiance et test statistique

Contents

Nous avons vu dans les chapitres précédents des manières d'estimer un paramètre inconnu θ^*. Malheureusement, il est quasi impossible que nos estimations $\hat{\theta}_n$ soient parfaitement égales à θ^*. L'idée de ce chapitre est d'introduire des estimateurs essayant de prendre en compte cette incertitude.

3.1 Principe général d'intervalle de confiance

Soit $(\mathcal{H}^n, (Q_\theta)_{\theta\in\Theta})$ un modèle statistique.

Définition 3.1. *Soit $\alpha \in (0,1)$, un intervalle de confiance pour θ de niveau de confiance (resp. confiance par excès) $1-\alpha$ est une statistique I à valeur dans les intervalles de $\mathbb{R}$ telle que pour chaque $\theta \in \Theta$,*

$$Q_\theta(I \ni \theta) = 1-\alpha \ (resp. \geqslant 1-\alpha).$$

Attention, il faut interpréter cette propriété comme la probabilité que notre intervalle de confiance contienne le paramètre θ sous Q_θ est de $1-\alpha$, et non que la probabilité que notre paramètre soit dedans est de $1-\alpha$. Ces formulations sont équivalentes, mais la première fait porter l'aléatoire sur I alors que la seconde sur θ. On rappelle que θ **n'est pas** aléatoire.

Pour essayer de construire de tel intervalle, il nous faut connaître deux quantités,
— le comportement approximative de l'estimateur $\hat{\theta}_n$;

— les quantiles d'ordre de ce comportement.

Définition 3.2. *Soit F une fonction de répartition d'une mesure de probabilité ν sur $\mathbb{R}$. On appelle quantile d'ordre $r \in (0,1)$,*

$$q_r := \inf\{t \in \mathbb{R} \mid F(t) \geqslant r\},$$

noté aussi $F^{\leftarrow}(r)$.

Remarque 3.3.
1. Si F est continue alors $F(q_r) = r$.
2. Si F est en plus strictement croissante alors $q_r = F^{-1}(r)$.
3. Dans le cas où ν est à densité par rapport à la mesure de Lebesgue, le quantile d'ordre s'interprète comme le plus petit t tel que l'aire sous la courbe entre $-\infty$ et t vaut r.
4. Dans le cas où la densité est pair, l'aire entre $-\infty$ et q_{1-r} est égale à l'aire entre $-q_{1-r/2}$ et $q_{1-r/2}$. C'est le cas, par exemple, de la densité de la distribution gaussienne.

3.1.1 Intervalle de confiance dans le cadre gaussien

On se place dans le modèle $(\mathbb{R}^n, (\mathcal{N}(m,\sigma^2)^{\otimes n})_{m,\sigma^2})$.

Estimation de m lorsque σ^2 est connu

Pour estimer la moyenne dans le cas gaussien, nous avons vu que la moyenne empirique était un bon estimateur. Sous la loi $Q_m = \mathbb{P}_m^{\otimes n}$, on a $\bar{X}_n \sim \mathcal{N}(m, \sigma^2/n)$. En effet, si X et Y sont deux variables gaussiennes indépendantes centrées en m_X et m_Y et de variance σ_X^2 et σ_Y^2 alors leur somme est encore une variable gaussienne centrée en $m_X + m_Y$ et de variance $\sigma_X^2 + \sigma_Y^2$.

Donc on connait exactement le comportement de l'estimateur $\bar{X}_n$ et les quantiles de la loi normale sont très bien connus. En normalisant la moyenne empirique pour se ramener à une gaussienne centrée réduite,

$$\sqrt{n} \times \frac{\bar{X}_n - m}{\sigma} \sim \mathcal{N}(0,1) \text{ sous } Q_m.$$

Ainsi pour $m \in \mathbb{R}$, on a en notant $q_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la distribution gaussienne centrée réduite, on a

$$Q_m\left(\sqrt{n}\left|\frac{\bar{X}_n - m}{\sigma}\right| \leqslant q_{1-\alpha/2}\right) = 1-\alpha.$$

Donc sous Q_m, la probabilité que que $\sqrt{n}\frac{\bar{X}_n - m}{\sigma} \in [-q_{1-\alpha/2}, q_{1-\alpha/2}]$ est de $1-\alpha$. Il suffit donc pour trouver un intervalle de confiance *d'inverser* le précédent intervalle, i.e. isoler le paramètre m et faire rentrer le reste dans l'intervalle.

$$\begin{aligned}
\sqrt{n}\frac{\bar{X}_n - m}{\sigma} \in [-q_{1-\alpha/2}, q_{1-\alpha/2}] &\Leftrightarrow \bar{X}_n - m \in \left[-\frac{\sigma q_{1-\alpha/2}}{\sqrt{n}}, +\frac{\sigma q_{1-\alpha/2}}{\sqrt{n}}\right] \\
&\Leftrightarrow -m \in \left[-\bar{X}_n - \frac{\sigma q_{1-\alpha/2}}{\sqrt{n}}, -\bar{X}_n + \frac{\sigma q_{1-\alpha/2}}{\sqrt{n}}\right] \\
&\Leftrightarrow m \in \left[\bar{X}_n - \frac{\sigma q_{1-\alpha/2}}{\sqrt{n}}, \bar{X}_n + \frac{\sigma q_{1-\alpha/2}}{\sqrt{n}}\right]
\end{aligned}$$

Pour la dernière équivalence, lorsque l'on multiplie par -1, il faut bien penser à changer l'ordre des bornes car en multipliant par un réel négatif, on doit changer le sens des inégalités !

Un intervalle de confiance possible pour le paramètre inconnu m est donc $\left[\bar{X}_n - \frac{\sigma q_{1-\alpha/2}}{\sqrt{n}}, \bar{X}_n + \frac{\sigma q_{1-\alpha/2}}{\sqrt{n}}\right]$. Sous Q_m, cet intervalle a une probabilité $1-\alpha$ de contenir le paramètre m. Analysons un peu la taille de cet intervalle. Lorsque le nombre d'observations augmente, la taille de l'intervalle diminue, i.e. avec plus d'observations, les estimations du paramètre sont plus précises. Inversement, si la variance σ^2 du phénomène augmente, alors l'estimation devient plus vague.

Estimation de m lorsque σ^2 est inconnu

Dans ce cas, on ne peut pas considérer le même intervalle, car il contiendrait un terme inconnu. Nous allons donc le remplacer par une estimation. On prendra l'estimateur non biaisé

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2,$$

car il permettra d'avoir les résultats suivants

Théorème 3.4 (de Fisher)**.** *Dans le modèle* $(\mathbb{R}^n, (\mathcal{N}(m,\sigma^2)^{\otimes n})_{m,\sigma^2})$*, on a les résultats suivants,*

1. $\sqrt{n}\frac{\bar{X}_n - m}{\sigma} \sim \mathcal{N}(0,1)$ *;*
2. $n\frac{S_n^2}{\sigma} \sim \chi^2(n-1)$ *;*
3. $\sqrt{n}\frac{\bar{X}_n - m}{\sqrt{S_n^2}} \sim \mathcal{T}(n-1)$*.*

Le point 1 a déjà été détaillé dans le paragraphe précédent. Le point 2 permet d'obtenir un intervalle de confiance lorsque l'on essaye d'estimer la variance. Le point 3 sert à estimer le paramètre m lorsque σ nous est inconnus. La loi de Student est aussi une distribution centrée et symétrique, ainsi en notant par $t_{1-\alpha/2}^{(n-1)}$ le quantile d'ordre $1-\alpha/2$ de $\mathcal{T}(n-1)$, on a

$$Q_m\left(\sqrt{n}\left|\frac{\bar{X}_n - m}{S_n^2}\right| \leqslant t_{1-\alpha/2}^{(n-1)}\right) = 1-\alpha.$$

Ainsi, en faisant les mêmes calculs que dans le paragraphe précédent, un intervalle de confiance de niveau de confiance $1-\alpha$ du paramètre m est

$$\left[\bar{X}_n - \frac{S_n^2 t_{1-\alpha/2}^{(n-1)}}{\sqrt{n}}, \bar{X}_n + \frac{S_n^2 t_{1-\alpha/2}^{(n-1)}}{\sqrt{n}}\right].$$

À noter qu'il s'agit aussi d'un IC dans le cas où σ est connu !

3.1.2 Intervalle de confiance asymptotique

Définition 3.5. *Pour* $\alpha \in (0,1)$ *, on appelle intervalle de confiance asymptotique de niveau de confiance* $1-\alpha$*, la suite d'intervalle de confiance* $I_n \subset \Theta$ *vérifiant*

$$\forall \theta \in \Theta,\ \lim_{n\to+\infty} Q_\theta(I_n(X_1,\dots,X_n) \ni \theta) = 1-\alpha.$$

Situation typique lorsque l'on ne connait que le comportement de l'estimateur de manière asymptotique, par exemple l'EMV sous des hypothèses de régularité ou la moyenne empirique avec le TCL. On rappelle plusieurs définitions équivalentes de la convergence en loi.

Définition 3.6. *Soit* $(X_n)_n$ *une suite de variables aléatoires et* X *une variable aléatoire. On dit que la suite converge en loi vers la variable* X *si l'un des points suivants (équivalents) est vérifié*

- *pour toutes fonctions* f *continues bornées,*

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$$

- *En notant par* $(F_n)_n$ *et* F *les fonctions de répartitions de ces variables, pour tous points de continuité* t *de* F*, on a* $F_n(t) \to F(t)$
- *pour tous boréliens* A *avec* $\mathbb{P}(X \in \partial A) = 0$*,*

$$\lim \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A).$$

Attention, la convergence en loi est une convergence sur les distributions des variables aléatoires et pas vraiment sur la variable en tant que telle. On rappelle aussi que cette convergence ne possède pas les mêmes propriétés de linéarité que les convergences classiques vues durant votre jeunesse. On interprète la convergence en loi comme le fait que le phénomène ayant comme distribution $\mathbb{P}_{X_n}$ ressemble énormément au phénomène régit par $\mathbb{P}_X$.

Exemple 3.7. Cette fois-ci considérons le modèle uniforme $(\mathbb{R}_+^n, (\mathcal{U}([0,\theta])^{\otimes n})_{\theta>0})$. Pour $\theta > 0$, on rappelle que l'espérance de cette loi est $\theta/2$. D'après le TCL, on a

$$\sqrt{n}\frac{\bar{X}_n - \theta/2}{\sqrt{\sigma^2}} \rightsquigarrow \mathcal{N}(0,1),$$

où $\sigma^2 = \frac{\theta^2}{12}$. Ainsi pour $n >> 1$, on a

$$Q_\theta\left(\sqrt{12n}\frac{\bar{X}_n - \theta/2}{\theta} \in [-q_{1-\alpha/2}, +q_{1-\alpha/2}]\right) \approx 1-\alpha.$$

Il faut donc *inverser* l'intervalle pour isoler le paramètre

$$\begin{aligned}
\sqrt{12n}\frac{\bar{X}_n - \theta/2}{\theta} \in [-q_{1-\alpha/2}, +q_{1-\alpha/2}] &\Leftrightarrow \frac{\bar{X}_n}{\theta} \in \left[\frac{1}{2} \pm \frac{q_{1-\alpha/2}}{\sqrt{12n}}\right] \\
&\Leftrightarrow \frac{\bar{X}_n}{\theta} \in \left[\max\left(0, \frac{1}{2} - \frac{q_{1-\alpha/2}}{2\sqrt{3n}}\right), \frac{1}{2} + \frac{q_{1-\alpha/2}}{2\sqrt{3n}}\right] \\
&\Leftrightarrow \theta \in \left[\frac{1}{\bar{X}_n}\frac{2\sqrt{3n}}{\sqrt{3n}+q_{1-\alpha/2}}, \frac{1}{\bar{X}_n \max\left(0, \frac{2\sqrt{3n}}{\sqrt{3n}-q_{1-\alpha/2}}\right)}\right].
\end{aligned}$$

Ainsi, la probabilité que cet intervalle contient le paramètre sous Q_θ vaut asymptotiquement $1-\alpha$. On aurait pu simplifier les calculs en utilisant le lemme de Stlusky. En effet, comme $\theta/\bar{X}_n \to 1$, on a par ce lemme

$$\frac{\theta}{\bar{X}_n}\sqrt{12n}\frac{\bar{X}_n - \theta/2}{\theta} \rightsquigarrow \mathcal{N}(0,1) \times 1.$$

Avec des calculs similaires que dans la Section 3.1.1, on a l'intervalle est un IC asymptotique de niveau $1-\alpha$,

$$\left[\bar{X}_n - \frac{\bar{X}_n q_{1-\alpha/2}}{2\sqrt{3n}}, \bar{X}_n + \frac{\bar{X}_n q_{1-\alpha/2}}{2\sqrt{3n}}\right].$$

Théoriquement, cet intervalle est moins précis car le lemme de Slutsky rajoute des approximations.

Dans un cadre plus général, lorsque l'on essaye d'estimer la moyenne, la méthode la plus simple est d'utiliser le TCL pour construire un intervalle de confiance asymptotique. En effet,

$$\sqrt{n}\frac{\bar{X}_n - m}{\sqrt{S_n^2}} \rightsquigarrow \mathcal{N}(0,1),$$

donc par la définition de la convergence en loi, on a

$$\lim_{n\to+\infty} Q_m\left(\sqrt{n}\frac{\bar{X}_n - m}{\sqrt{S_n^2}} \in [-q_{1-\alpha/2}, +q_{1-\alpha/2}]\right) = 1-\alpha.$$

Donc en *inversant* l'intervalle, on obtient l'IC asymptotique suivant

$$\left[\bar{X}_n - \frac{\sqrt{S_n^2}q_{1-\alpha/2}}{\sqrt{n}}, \bar{X}_n + \frac{\sqrt{S_n^2}q_{1-\alpha/2}}{\sqrt{n}}\right].$$

Remarque 3.8. Nous avons introduit dans cette section uniquement des intervalles de confiance dit bilatéral, c'est à dire que l'on a utilisé au départ pour construire notre intervalle de confiance l'intervalle $[\pm q_{1-\alpha/2}]$. On aurait pu utiliser $]-\infty, q_{1-\alpha}]$ ou $[-q_{1-\alpha}, +\infty[$. On aurait alors eu des intervalles de confiance complètement différent que l'on caractérise d'unilatéral. On verra dans la suite quel type d'intervalle est utile en fonction du problème que l'on considère.

3.2 Tests statistiques

3.2.1 Principe général des tests

Le but est de confronter deux hypothèses, l'une dite hypothèse nulle (H_0) et l'autre dite alternative (H_1), contradictoire avec (H_0). À l'issu d'un test,

— soit on rejette (H_0) ;

— soit on accepte (H_0), **on préférera la formulation *ne pas rejeter* (H_0).**

Il existe ainsi 4 cas possibles

Décision \ Réalité	(H_0) est vraie	(H_0) est fausse
non rejet	bonne décision	erreur de 2^e espèce
rejet	erreur de 1^{ere} espèce	bonne décision

Exemple 3.9. On testera en exercice généralement des hypothèses de la forme

— (H_0) : "le médicament est efficace" *Vs* (H_1) : "le médicament n'est pas efficace" ;

— (H_0) : "le dé est truqué" *Vs* (H_1) : "le dé n'est pas truqué" ;

— (H_0) : "$\theta = 0$" *Vs* (H_1) : "$\theta < 0$".

À noter que (H_1) **n'est pas forcément** le complémentaire de l'hypothèse nulle.

Remarque 3.10. On verra dans la suite qu'il y a une asymétrie entre les deux hypothèses. L'hypothèse nulle sera favorisée dans le sens qu'elle sera considéré comme vraie jusqu'à preuve du contraire. C'est le même principe que pour la justice avec la présomption d'innocence. On met généralement dans (H_1) un fait que l'on essaye de découvrir. Un autre fait à retenir est l'utilisation de la formulation ***non rejet*** au lieu de ***acceptation***. C'est un peu le même principe qu'en physique, nos observations ne peuvent pas prouver une théorie mais peut la réfuter ou ne pas la réfuter. Observer un stylo tombé ne prouve pas la théorie de la gravité de Newton mais permet juste de ne pas la rejeter car la chute de ce stylo est en accord avec les prédictions de la théorie newtonienne. Dans le 3^e point de l'exemple précédent, le test permettra de vérifier si notre estimation du paramètre θ avec $\hat{\theta}$ est compatible avec le fait que $\theta = 0$, mais en aucun cas prouvera que le vrai paramètre est nul !

Définition 3.11.

1. *Un test pur est une statistique T définie sur un modèle $(\mathcal{H}^n, (Q_\theta)_{\theta\in\Theta})$ à valeur dans $\{0,1\}$. L'hypothèse (H_0) sera rejetée si $T(x) = 1$ pour $x \in \mathcal{H}^n$ et pas rejeter dans le cas contraire. L'ensemble $\{x \in \mathcal{H}^n \mid T(x) = 1\}$ s'appelle zone de rejet.*
2. *Un test stochastique est une statistique $T\colon \mathcal{H}^n \to [0,1]$ tel que pour une observation x, on rejette l'hypothèse nulle avec probabilité $T(x)$.*

Dans la suite, on ne considérera que les tests purs. Ainsi, nos tests auront tous la forme $T(x) = \mathbb{1}_{x\in\mathcal{R}}$. L'ensemble $\mathcal{R}$ sera appelé zone de rejet. Si notre observation tombe dans cette zone, on conclura que l'observation est en désaccord (ou est peu probable) avec l'hypothèse nulle. De manière générale, nos tests auront la forme

$$(H_0) : "Q_{\theta^*} \in \mathcal{P}_0" \ Vs \ (H_1) : "Q_{\theta^*} \in \mathcal{P}_1",$$

avec $\mathcal{P}_0, \mathcal{P}_1 \subset \{Q_\theta,\ \theta \in \Theta\}$ et $\mathcal{P}_0 \cap \mathcal{P}_1 = \emptyset$. Avec les mains, on peut reformuler ces tests par la question de savoir si notre monde appartient à $\mathcal{P}_0$ ou $\mathcal{P}_1$. Lorsque ces deux ensembles sont des singletons, on dira que le test est simple et composé dans le cas contraire.

Définition 3.12. *1. Le risque de première espèce d'un test pur est l'application*

$$Q_\theta \in \mathcal{P}_0 \mapsto \mathbb{E}_\theta[T] = Q_\theta(T = 1) \in [0, 1],$$

i.e. la probabilité que notre test rejette à tord dans le cas où l'on se situe dans Q_θ. Lorsque la borne supérieure de cette fonction vaut α, on dira que le risque de 1^{ere} espèce maximale vaut α ou que le test est de niveau α.

2. Le risque de deuxième espèce est l'application

$$Q_\theta \in \mathcal{P}_1 \mapsto 1 - \mathbb{E}_\theta[T] = Q_\theta(T = 0),$$

i.e. la probabilité de ne pas rejeter alors qu'on aurait dû rejeter. En pratique, on considère plutôt l'application

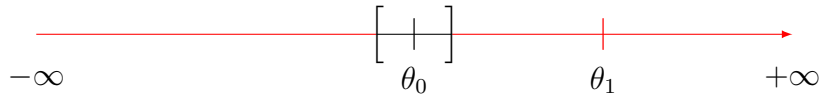
$$Q_\theta \in \mathcal{P}_1 \mapsto \mathbb{E}_\theta[T] = Q_\theta(T = 1),$$

et on appelle puissance sa borne inférieure.

Un bon test doit un niveau faible et une puissance importante. Malheureusement, on ne peut pas optimiser les deux en même. Lorsque l'on voudra faire baisser le niveau, on diminuera automatiquement la puissance. En effet, pour avoir un niveau faible, il faut pouvoir facilement de ne pas rejeter, ce qui est incompatible avec une puissance importante.

Exemple 3.13. Prenons le modèle $(\mathbb{R}^n, (\mathcal{N}(\theta, 1)^{\otimes n})_{\theta\in\mathbb{R}})$, pour $\theta_0 \in \mathbb{R}$, étudions les hypothèses

$$(H_0) : "\theta^* = \theta_0" \ Vs \ (H_1) : "\theta^* \neq \theta_0".$$



Posons la zone de rejet,

$$\mathcal{R}_n := \{x \in \mathbb{R}^n \mid |\sqrt{n}(\bar{x}_n - \theta_0)| \geqslant q_{1-\alpha/2}\} = \{x \in \mathbb{R}^n \mid \bar{x}_n \in \mathbb{R} \setminus [\theta_0 \pm q_{1-\alpha_2}/\sqrt{n}]\}.$$

Pour $n >> 1$, on a $\theta_1 \in \mathcal{R}_n$ et posons le test $T(X_1, \dots, X_n) = \mathbb{1}_{(X_1,\dots,X_n)\in\mathcal{R}_n}$. Ainsi, si notre observation $(x_1, \dots, x_n)$ tombe dans $\mathcal{R}_n$, on rejette sinon on ne rejette pas. Est-ce que ce choix fonctionne ?

Pour le modèle Gaussien, on connait le comportement de la moyenne empirique, on a

$$\begin{aligned} Q_{\theta_0}(T(X_1, \dots, X_n) = 0) &= Q_{\theta_0}\left(|\sqrt{n}(\bar{X}_n - \theta_0)| \leqslant q_{1-\alpha/2}\right) \\ &= Q_{\theta_0}(-q_{1-\alpha/2} \leqslant \sqrt{n}(\bar{X}_n - \theta_0) \leqslant q_{1-\alpha/2}) \\ &= 1 - \alpha/2 - \alpha/2 = 1 - \alpha, \end{aligned}$$

car sous (H_0), on a $\sqrt{n}(\bar{X}_n - \theta_0) \sim \mathcal{N}(0, 1)$. Ainsi, si le vrai paramètre (inconnu) vaut bien θ_0, la probabilité de rejeter à tord est α. On voit aussi que si $\theta_1 \approx \theta_0$, θ_1 ne sera pas dans la zone de rejet. Raison de plus pour dire, ne pas rejeter à la place d'accepter.

On remarque que le test construit dans cet exemple s'inspire énormément de la forme de l'intervalle de confiance trouvé pour l'estimation de la moyenne lorsque σ^2 est connu. En effet, lorsque l'on possède un intervalle de confiance de niveau α, on peut construire des tests lorsque l'hypothèse nulle est un singleton.

Exemple 3.14. Plaçons nous dans le cas où l'on confronte les deux hypothèses suivantes

$$(H_0) : "\theta^* = \theta_0" \ Vs \ (H_1) : "\theta^* \neq \theta_0".$$

Supposons que l'on possède un intervalle de confiance $I(X_1, \dots, X_n)$, i.e. pour tout $\theta \in \Theta$,

$$Q_\theta(I(X_1, \dots, X_n) \ni \theta) = 1 - \alpha.$$

On peut alors vérifier que le test suivant

$$T(X_1, \dots, X_n) = \begin{cases} 0 & \text{si } \theta_0 \in I(X_1, \dots, X_n) \\ 1 & \text{sinon} \end{cases},$$

est bien un test de niveau α car la probabilité de rejeter à tord lorsque (H_0) est vraie est de α.

Définition 3.15. *Un test T de niveau α est dit sans biais si sa puissance est supérieur à α,*

$$\forall Q_\theta \in \mathcal{P}_0, \ \mathbb{E}_\theta[T] \geqslant \alpha.$$

3.2.2 Tests asymptotiques

Comme pour les intervalles de confiance, il n'est pas forcément possible de construire un test avec un niveau exacte à n fixé. On introduit alors la notion de niveau asymptotique d'un test.

Définition 3.16. *Un test T_n est asymptotiquement de niveau α si*

$$\sup_{Q_\theta \in \mathcal{P}_0} \lim_{n \to +\infty} \mathbb{E}_\theta[T_n] = \alpha.$$

Attention, on ne peut pas forcément intervertir l'ordre la borne supérieure et de la limite.

Définition 3.17. *Un test sera dit convergeant si pour tout $Q_\theta \in \mathcal{P}_1$,*

$$\lim_{n \to +\infty} \mathbb{E}_\theta[T_n] = 1.$$

Cette propriété signifie que si l'hypothèse (H_1) est vraie, alors de manière asymptotique, on est sûr de rejeter. Ca peut être le cas d'un test utilisant une statistique ayant un bon comportement sous (H_0) et qui explose sous (H_1), voir Section 3.4.2.

Remarque 3.18. L'Exemple 3.14 peut s'adapter facilement au cas d'intervalle de confiance asymptotique pour construire un test asymptotique. Cette vérification est laissée en exercice au lecteur.

Exemple 3.19. Concentrons dans cette exemple sur le cadre Pile-Face, nous allons montrer que le test asymptotique construit à l'aide du TCL est convergent. Prenons le modèle déjà présenté dans l'Exemple 1.25, et confrontons les hypothèses

$$(H_0) : "p^* = 1/2" \ Vs \ (H_1) : "p^* \neq 1/2".$$

On peut vérifier à l'aide du TCL que

$$2\sqrt{n}(\bar{X}_n - 1/2) \rightsquigarrow \mathcal{N}(0, 1),$$

sous (H_0) car $\mathrm{Var}(X_1) = 1/4$. Donc par la définition de la convergence en loi,

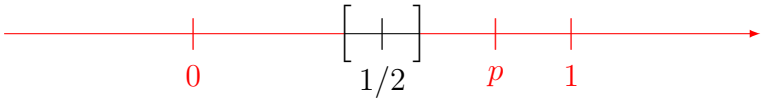
$$Q_{1/2}(2\sqrt{n}|\bar{X}_n - 1/2| \geqslant q_{1-\alpha/2}) \to 1 - \alpha.$$

Avec la zone de rejet,

$$\mathcal{R}_n = \{x \in \mathbb{R}^n \mid 2\sqrt{n}|\bar{x}_n - 1/2| \geqslant q_{1-\alpha/2}\},$$

on construit un test asymptotiquement de niveau α en prenant $T(X_1, \dots, X_n) = \mathbb{1}_{(X_1,\dots,X_n) \in \mathcal{R}_n}$. Maintenant pour étudier la convergence, il faut se placer dans (H_1), prenons $p \neq 1/2$, alors on a dans le cas des tests purs

$$\mathbb{E}_p[T_n] = Q_p(T_n = 1) = Q_p(2\sqrt{n}|\bar{X}_n - 1/2| \geqslant q_{1-\alpha/2}) = Q_p\left(\bar{X}_n \in \left[1/2 \pm \frac{\sqrt{q_{1-\alpha/2}}}{2\sqrt{n}}\right]\right)$$



L'intervalle autour de $1/2$ se rétrécit lorsque $n \to +\infty$, donc à partir d'un certains rang, p sort de cet intervalle. De plus, par la LGN, la moyenne empirique converge vers p, donc pour n grand, on sait que

$$\bar{X}_n \notin \left[1/2 \pm \frac{\sqrt{q_{1-\alpha/2}}}{2\sqrt{n}}\right],$$

d'où la convergence du test.

3.3 Mise en pratique d'un test pur

3.3.1 Test bilatéral

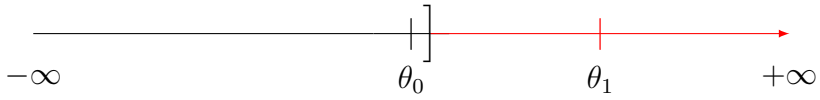
Il s'agit de la forme des tests vus juste avant. La zone de rejet construite à l'aide des intervalles de confiance se situe des deux côtés de la valeur que l'on teste. On trouve de tel test pour des confrontations de la forme

$$(H_0) : "\theta^* = \theta_0" \ Vs \ (H_1) : "\theta^* \neq \theta_0".$$

3.3.2 Test unilatéral

Dans certains cas, l'alternative est différente de $(H_1) : "\theta^* \neq \theta_0"$. Il se peut que l'on essaye de savoir si le vrai paramètre est plus grand ou plus petit qu'une certaine valeur.

$$(H_0) : "\theta^* = \theta_0" \ Vs \ (H_1) : \begin{cases} "\theta^* < \theta_0" \\ "\theta^* > \theta_0" \\ "\theta^* = \theta_1" \end{cases}.$$



La zone de rejet dépendra de l'alternative. Si l'alternative est *à droite* de l'hypothèse nulle alors on mettra la zone de rejet à droite. Dans ce schéma la zone de rejet sera

$$\mathcal{R} = \{x \in \mathbb{R}^n \mid \bar{x}_n \geqslant \theta_0 + \delta_n\}.$$

Exemple 3.20. Le temps de réaction X d'un médicament administré à une souris suit une loi normale centrée en 19 (minutes). On suppose que l'on connait $\sigma^2 = 1$. On expérimente un nouveau produit pour lequel on observe les temps de réaction suivants : $15, 14, 21, 12, 17, 19, 18$. La réaction est-elle plus rapide avec le nouveau produit ?

1. **Formulation du problème**
   On va mettre en hypothèse nulle le fait que le nouveau produit n'est pas plus efficace. Pour simplifier la calibration du niveau du test, on va réduire l'hypothèse nulle à un singleton. On considère alors la confrontation suivante

   $$(H_0) : "m = 19" \ Vs \ (H_1) : "m < 19".$$

   Avec un tel test, on ne peut pas différentier si le produit agit de manière identique ou plus lentement. Ici, on souhaite se prémunir en priorité du risque de déclarer à tord que le produit est plus rapide. On a envie de rejeter l'hypothèse nulle si $\bar{X}_n$ est très à gauche de 19.

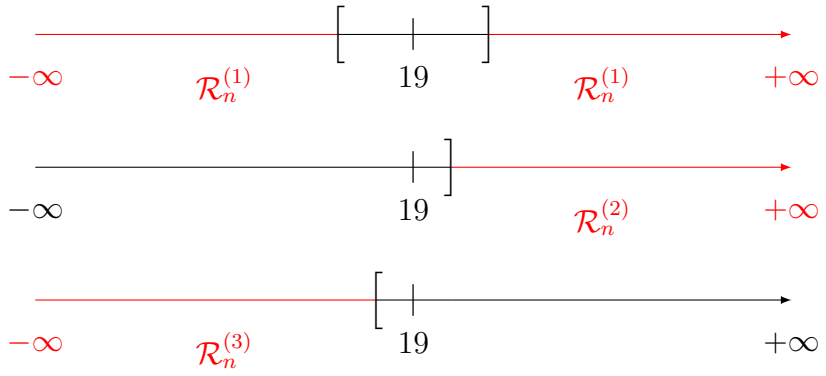
2. **Choix du seuil**
   On choisit $\alpha \in (0,1)$. Généralement, on prendra $\alpha = 0.05 = 5\%$.
3. **Choix de la statistique du test et de la zone de rejet**
   Cette partie est direction liée l'étape 1 sur la formulation du problème. Il y a trois formes naturelles de test basées sur les trois égalités suivantes

$$Q_{19}(\sqrt{n}|\bar{X}_n - 19| \leqslant q_{1-\alpha/2}) = 1-\alpha \tag{3.1}$$
$$Q_{19}(\sqrt{n}(\bar{X}_n - 19) \leqslant q_{1-\alpha}) = 1-\alpha \tag{3.2}$$
$$Q_{19}(\sqrt{n}(\bar{X}_n - 19) \geqslant -q_{1-\alpha}) = 1-\alpha \tag{3.3}$$

   Pour retrouver ces résultats, on rappelle que $\sqrt{n}(\bar{X}_n - 19) \sim \mathcal{N}(0,1)$ sous (H_0). Il faut maintenant déterminer laquelle de ces égalités nous allons utiliser en regardant les zones de rejet associées

$$\begin{cases} Q_{19}\left(\bar{X}_n \in \left[19 \pm \frac{q_{1-\alpha/2}}{\sqrt{n}}\right]\right) = 1-\alpha & \rightsquigarrow \mathcal{R}_n^{(1)} = \{x \in \mathbb{R}^n \mid \bar{x}_n \notin [19 \pm q_{1-\alpha/2}/\sqrt{n}]\} \\ Q_{19}\left(\bar{X}_n \in \left]-\infty, 19 + \frac{q_{1-\alpha}}{\sqrt{n}}\right]\right) = 1-\alpha & \rightsquigarrow \mathcal{R}_n^{(2)} = \{x \in \mathbb{R}^n \mid \bar{x}_n \notin]-\infty, 19 + q_{1-\alpha}/\sqrt{n}]\} \\ Q_{19}\left(\bar{X}_n \in \left[19 - \frac{q_{1-\alpha}}{\sqrt{n}}, +\infty\right[\right) = 1-\alpha & \rightsquigarrow \mathcal{R}_n^{(3)} = \{x \in \mathbb{R}^n \mid \bar{x}_n \notin [19 - q_{1-\alpha}/\sqrt{n}, +\infty[\} \end{cases}.$$

   Ces trois zones de rejet permettent bien de construire un test de niveau $1-\alpha$ mais les tests ne sont pas tous pertinent de la même manière au vue de l'hypothèse alternative. Visualisons ces zones de rejet



   Intuitivement, au vue de l'hypothèse alternative (H_1) : "$m < 19$", la meilleur zone de rejet est $\mathcal{R}_n^{(3)}$. Sous (H_1), on a $m < 19$ donc

$$\sqrt{n}(\bar{X}_n - 19) \sim \mathcal{N}(\underbrace{\sqrt{n}(m-19)}_{\to -\infty}, 1),$$

   donc sous (H_1), on peut montrer que

$$\begin{cases} Q_m(T_3(X_1,\dots,X_n) = 0) < Q_m(T_1(X_1,\dots,X_n) = 0) \\ Q_m(T_3(X_1,\dots,X_n) = 0) < Q_m(T_2(X_1,\dots,X_n) = 0) \end{cases},$$

   où l'on note T_i le test associé à la zone de rejet $\mathcal{R}_n^{(i)}$. Ainsi, le test 3 possède une meilleur puissance, i.e. si (H_1) est vraie, le test 3 a moins de chance de ne pas rejeter à tord.
4. **Décision**
   On a observé $\bar{X}_n = 16$ et $n = 8$. On choisit de prendre un seuil $\alpha = 5\%$, donc $q_{1-\alpha} = 1.645$, donc on rejette si la moyenne empirique est inférieur à

$$19 - 1.645/\sqrt{8} = 18.418.$$

5. **Conclusion**
   On rejette l'hypothèse nulle en faveur de l'hypothèse alternative, donc le nouveau produit est plus efficace que l'ancien produit.

3.3.3 La p-valeur

Lors de la conclusion, on rejette ou non. Mais peut on quantifier *à quel point* on rejette ? Plus précisément, si on change le niveau de confiance, comment évolue la réponse ? Pour $\alpha \in (0,1)$, le test de niveau α avait la forme

$$T_\alpha(x_1, \dots, x_n) = \mathbb{1}_{x \in \mathcal{R}(\alpha)},$$

où $\mathcal{R}(\alpha)$ était la zone de rejet. Dans nos exemples, ces régions étaient croissantes en α,

$$\forall \alpha \leqslant \beta,\ \mathcal{R}(\alpha) \subset \mathcal{R}(\beta),$$

i.e. lorsque α diminue, la zone de rejet aussi.

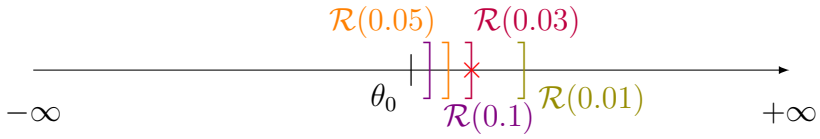
Un test avec un très faible niveau a forcément une petite zone de rejet, car on souhaite à tout prix éviter de rejeter à tord, donc on prend la stratégie de rejeter difficilement.

Définition 3.21. *Soit $x = (x_1, \dots, x_n)$ une observation, on appelle p-valeur, p-value en anglais, de l'observation la quantité*

$$p(x) = \sup\ \{\alpha \mid x \notin \mathcal{R}(\alpha)\} = \inf\ \{\alpha \mid x \in \mathcal{R}(\alpha)\},$$

i.e. la valeur α correspondant à la plus petite zone de rejet contenant notre observation.

Soit x une observation et $p(x)$ sa p-valeur, si on fait un test au niveau α,
- on rejette si $p(x) < \alpha$;
- on ne rejette pas si $p(x) > \alpha$;
- convention à choisir si $p(x) = \alpha$ (mais cas impossible si on manipule des quantités continues)



Ainsi, si $p(x) = 0.03$, on rejette si $\alpha = 0.05$ mais on ne rejette pas si $\alpha = 0.01$. Cette quantité me donne plus d'information que rejeter ou ne pas rejeter. La convention autour de cette quantité est la suivante
- $p \leqslant 0.01$: très forte présomption contre (H_0)
- $0.01 < p \leqslant 0.05$: forte présomption contre (H_0)
- $0.05 < p \leqslant 0.1$: faible présomption contre (H_0)
- $p > 0.1$: pas de présomption contre (H_0)

On peut interpréter la p-valeur comme la probabilité $p(x)$ d'observer un événement plus contradictoire que l'observation x lorsque l'hypothèse (H_0) est vraie. **Attention**, il **ne faut surtout pas** interpréter la p-valeur comme la probabilité que l'hypothèse nulle soit vraie !

3.4 Tests du χ^2

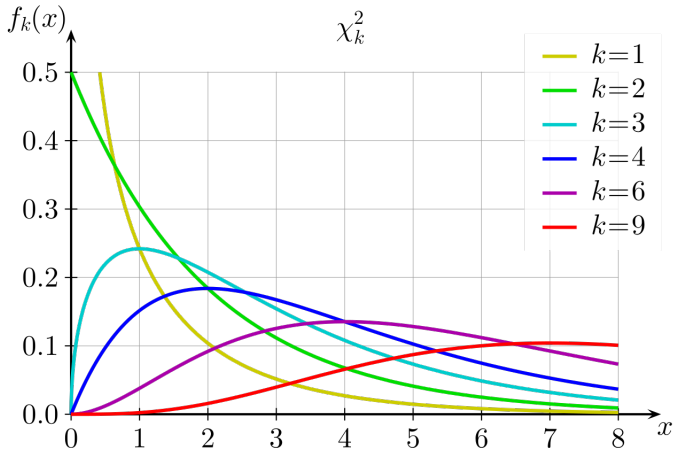
Nous allons voir dans cette dernière section plusieurs tests basés sur la distribution du χ^2.

3.4.1 Distribution du χ^2

Définition 3.22. *Pour $d \in \mathbb{N}^*$, soit $Z_1, \dots, Z_d$, d variables gaussiennes centrées réduites indépendantes. On dit que X suit une loi du χ^2 à d degrés de liberté si X a la même loi que*

$$Z_1^2 + \dots + Z_d^2.$$

Plus le degré de liberté d est grand, plus la distribution a tendance *à charger* les grandes valeurs. Ces distributions sont des distributions possédant une densité par rapport à la mesure de Lebesgue.



Une autre manière d'observer ce phénomène est de regarder les quantiles d'ordre. Fixonx $\alpha = 5\%$, on a

$$\begin{cases} q_{1-\alpha}^{(d=1)} = 3.84 & \text{i.e. } \mathbb{P}(X \leqslant 3.84) = 0.95 \text{ si } X \sim \chi^2(1) \\ q_{1-\alpha}^{(d=3)} = 7.81 & \text{i.e. } \mathbb{P}(X \leqslant 7.81) = 0.95 \text{ si } X \sim \chi^2(3) \\ q_{1-\alpha}^{(d=10)} = 18.31 & \text{i.e. } \mathbb{P}(X \leqslant 18.31) = 0.95 \text{ si } X \sim \chi^2(10) \end{cases}$$

3.4.2 Test d'adéquation à une loi discrète

On se place dans le cas où notre phénomène est discret fini, i.e.

$$\mathcal{H} = \{a_1, \dots, a_r\},$$

et on considère l'ensemble des lois possibles suivant $\mathcal{P} = \{\sum_{i=1}^r p_i \delta_{a_i} \mid \sum p_i = 1,\ p_i > 0\}$. On se demande maintenant si pour $\pi = (p_1, \dots, p_r) \in]0,1[^r$,

$$(H_0) : \text{"}\mathbb{P}_\theta = \sum_{i=1}^r p_i \delta_{a_i}\text{"} \text{ Vs } (H_1) : \text{"}\mathbb{P}_\theta \neq \sum_{i=1}^r p_i \delta_{a_i}\text{"},$$

i.e. est-ce que notre phénomène est régi par $\sum_{i=1}^r p_i \delta_{a_i}$? Par exemple, avec le contexte du dé, on peut se demander si notre dé est truqué, donc $p_i = 1/6$. Pour tester cette hypothèse, on dispose d'observations i.i.d. $(X_1, \dots, X_n)$. Notons par N_j le nombre de fois où a_j a été observé. On sait que d'après la LGN, sous (H_0),

$$f_j = \hat{p}_j = \frac{N_j}{n} \overset{p.s.}{\to} p_j,$$

i.e. la fréquence empirique d'apparition de a_j tend vers la probabilité théorique. On va donc comparer le vecteur empirique $(\hat{p}_1, \dots, \hat{p}_r)$ au vecteur π. Sous (H_0), ces deux vecteurs doivent être proche. Pour quantifier la notion de proche, nous allons utiliser la statistique suivante, ressemblant à une distance,

$$D_n = \sum_{i=1}^r \frac{(p_j - \hat{p}_j)^2}{p_j}.$$

Ainsi, sous (H_0), cette quantité doit tendre vers 0 et donc être très petit. Pour calibrer la notion de *petit* ou de *grand*, nous allons utiliser le théorème suivant

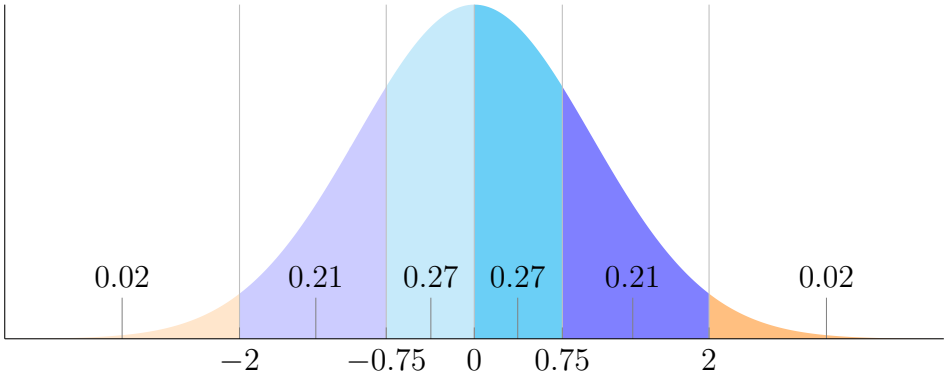
Théorème 3.23. *Dans le cas où les observations $(X_1, \dots, X_n)$ sont i.i.d., nous avons*

1. *sous (H_0), $U_n = nD_n \rightsquigarrow \chi^2(r-1)$;*
2. *sous (H_1), $U_n \overset{p.s.}{\to} +\infty$.*

Ainsi, c'est la loi du $\chi^2(r-1)$ qui va nous servir pour la calibrer la notion de grand. Notons $q_{1-\alpha}^{(r-1)}$ le quantile d'ordre $1-\alpha$ de $\chi^2(r-1)$, si U_n est plus petit que $q_{1-\alpha}^{(r-1)}$, on va considérer que notre observation de U_n est en accord avec (H_0), on ne va donc pas rejeter. Alors que si U_n est plus grand que ce quantile, on considère qu'il est peu probable d'observer ceci sous (H_0), on rejette alors l'hypothèse nulle.

Remarque 3.24.

1. Le deuxième point nous assure que si on est sous (H_1) alors forcément on rejettera l'hypothèse nulle lorsque n sera très grand.
2. On effectue une approximation ! Cela fonctionne bien à condition que $np_j \geqslant 5$, où p_j est la probabilité théorique. Si ce n'est pas le cas, il faut fusionner des classes.
3. On peut adapter ce test à des lois non discrètes infinies, en faisant des groupes. Par exemple, si on veut tester si $\mathbb{P}_\theta$ est une loi normale $\mathcal{N}(0,1)$, on découpe $\mathbb{R}$ en un nombre fini de cases.



Par exemple, ici on a découpé $\mathbb{R}$ en 6 parties de taille différentes.

4. Pour déterminer la p-valeur de ce test, on cherche le plus petit $\alpha \in (0,1)$ tel que $U_n > q_{1-\alpha}^{(r-1)}$. Par la continuité et la stricte croissance de la fonction de répartition des lois du χ^2, on cherche α vérifiant $U_n = q_{1-\alpha}^{(r-1)}$. Par exemple, si on observe $U_{obs} = 7.96$, la p-valeur dépendra du nombre de degré de liberté de notre distribution :
   — si $r = 4$ et donc $r-1 = 3$, la p-valeur est inférieur à 5%, on a donc une forte présomption contre (H_0) ;
   — si $r = 17$ et donc $r - 1 = 16$, la p-valeur vaut 95%, donc aucune présomption contre (H_0).
   Cette exemple s'interprète comme 7.96 est une grande valeur lorsque $r = 4$ mais petit si $r = 17$.

3.4.3 Extension à une famille de distributions

Avec le point 3 de la Remarque 3.24, on peut confronter les alternatives

$$(H_0) : \text{"}\mathbb{P}_\theta = \mathcal{G}(1/2)\text{"} \text{ Vs } (H_1) : \text{"}\mathbb{P}_\theta \neq \mathcal{G}(1/2)\text{"}.$$

On peut par exemple diviser l'ensemble $\mathbb{N}^*$ en 8 sous-ensembles, mais cela dépend du nombre d'observations (cf point 2). Mais maintenant, si l'on souhaite confronter les alternatives

$$(H_0) : \text{"}\mathbb{P}_\theta \text{ est une loi géométrique"} \text{ Vs } (H_1) : \text{"}\mathbb{P}_\theta \text{ n'est pas une loi géométrique"},$$

on ne peut plus utiliser directement l'astuce précédente. En effet, dans ce cas, quelles sont les probabilités théoriques que l'on doit utiliser ? L'idée de cette extension est d'estimer le paramètre

p de la loi géométrique à l'aide de l'EMV $\hat{p}_{MV}$ puis tester si $\mathbb{P}_\theta$ est la distribution $\mathcal{G}(\hat{p}_{MV})$. De manière plus générale, notons pour $j \in \{1, \dots, r\}$, $\hat{\pi}_j$ la probabilité de la classe j pour la distribution $\mathbb{P}_{\hat{\theta}_{MV}}$. On considère alors la statistique suivante comparant les probabilités *empirico-théoriques* $(\hat{\pi}_j)_{j=1}^r$ et les probabilités empiriques $(\hat{p}_j)_{j=1}^r$

$$D_n = \sum_{j=1}^{r} \frac{(\hat{\pi}_j - \hat{p}_j)^2}{\hat{\pi}_j}.$$

Alors en notant q le nombre de paramètres estimés à l'aide de l'EMV, nous avons une généralisation du Théorème 3.23.

Théorème 3.25. *Dans le cas où les observations $(X_1, \dots, X_n)$ sont i.i.d., nous avons*

1. *sous (H_0), $U_n = nD_n \rightsquigarrow \chi^2(r-1-q)$;*
2. *sous (H_1), $U_n \overset{p.s.}{\to} +\infty$.*

En remplaçant, les probabilités théoriques $(p_j)_{j=1}^r$ par des probabilités *empirico-théoriques* $(\hat{\pi}_j)_{j=1}^r$, la variable D_n peut plus facilement être *petite* car ces probabilités ont été construites avec l'EMV pour *coller* aux données, et donc aux probabilités empiriques $(\hat{p}_j)_{j=1}^r$. Pour prendre en compte ce sur-ajustement, on diminue le degré de liberté de la distribution du χ^2 pour qu'il soit *plus facile d'être grand.*

3.4.4 Test d'indépendance du χ^2

On considère maintenant des vecteurs i.i.d. $(Y_1, Z_1), \dots, (Y_n, Z_n)$ de loi commune $\mathcal{L}(Y, Z)$ à valeurs dans $\{a_1, \dots, a_r\} \times \{b_1, \dots, b_s\}$. On note la loi inconnue du couple par $p = (p_{i,j})_{1 \leqslant i \leqslant r, 1 \leqslant \leqslant s}$ ainsi que les lois marginales

$$p_{i,\bullet} = \sum_{j=1}^{s} p_{i,j} \text{ et } p_{\bullet,j} = \sum_{i=1}^{r} p_{i,j}.$$

Nous allons tester

$$(H_0) : \text{"les deux phénomènes sont indépendants" Vs } (H_1) : (H_0)^c.$$

On rappelle que les deux phénomènes sont indépendants si et seulement si $p_{i,j} = p_{\bullet,j} p_{i,\bullet}$. Ainsi théoriquement, si il y a bien indépendance alors

$$f_{i,j} \approx f_{i,\bullet} f_{\bullet,j}, \tag{3.4}$$

où $f_{i,j}$ est la fréquence empirique d'apparition de (a_i, b_j), et $f_{\bullet,j}$ et $f_{i,\bullet}$ sont les fréquences d'apparition de b_j et a_i. Nous allons utiliser une statistique proche des précédentes pour quantifier l'approximation de l'équation (3.4),

$$D_n = \sum_{j=1}^{s} \sum_{i=1}^{r} \frac{(f_{i,j} - f_{i,\bullet} f_{\bullet,j})^2}{f_{i,\bullet} f_{\bullet,j}}.$$

Théorème 3.26. *Dans le cas où les observations $(Y_1, Z_1), \dots, (Y_n, Z_n)$ sont i.i.d., nous avons*

1. *sous (H_0), $U_n = nD_n \rightsquigarrow \chi^2((r-1)(s-1))$;*
2. *sous (H_1), $U_n \overset{p.s.}{\to} +\infty$.*

Ainsi la procédure du test est identique que pour les tests d'adéquation, si U_n est *petit* alors on ne rejette pas sinon on rejette. La distribution qui cette fois va calibrer la notion de grand est la loi $\chi^2((r-1)(s-1))$.