

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Mohammed Seddik Ben Yahia - Jijel
Faculté des Sciences Exactes et Informatique
Département de Mathématiques



Mémoire de fin d'études

Présenté pour l'obtention du diplôme de

Master

Spécialité : Mathématiques fondamentales.

Option : Probabilités et Statistique.

Thème

Utilité de la prédiction de satisfaction dans la planification expérimentale

Présenté par :

* Boutelala Amina

Devant le jury composé de :

Madi Meriem M.A.A Université de Jijel

Président

Djeridi Zohra M.C.B Université de Jijel

Encadreur

Sellami Nawel M.A.A Université de Jijel

Examineur

Promotion 2020/2021

♡ *Remerciements* ♡

Tout d'abord, je remercie **Allah** le tout puissant et le miséricordieux, pour me donner la santé, la volonté et la patience pour mener à terme mon formation de master.

C'est avec un grand honneur que je remercie mon enseignante et promoteur Madame **Djeridi Zohra** pour me conduire pour la réalisation de ce travail, pour ces précieux conseils et ces encouragements.

Je tiens à remercier, Mme **Madi Meriem** , d'avoir accepté la présidence du jury de mon mémoire, qu'il trouve ici toutes mes expressions respectueuses. Je remercies également, Mme **Sellami Nawel** , de m'avoir fait l'honneur de faire partie des membres du jury et d'examiner mon mémoire de fin d'étude .
je tiens à vous remercier.

Pour finir, je souhaite exprimer mes remerciements à : ma famille surtout **Mes parents** pour leurs intérêts envers mon travail et pour leurs aide ;

Dédicaces

À mes chers parents **Ali & Manouba**

Qui m'ont fourni les meilleures conditions pour terminer mes études, en témoignage de grand amour, tendresse et encouragement. Qui dieu m'offre la chance d'être à la hauteur de leur attentes.

Vous avez guidé mes premiers pas, vous m'avez toujours servi de modèle et vous rester toute ma vie. Vous avez fait de moi ce que je suis aujourd'hui.

Je vous dédie ce modeste travail qui n'est d'autre que le fruit vos nobles sacrifices.

À mes chères sœurs **Aya & Djennat**

Pour leurs amour, leurs serviabilité et pour la tendre affectation qu'ils m'ont toujours témoigné.

À mes frères **Amine & Zakaria & Ishake**

Qui n'ont cessé d'être pour moi des exemples de persévérance, de courage et de générosité.

À Mon fiancé (mon prince) **Amran**

Pour ta compréhension, ta confiance, t'amour, ta fidélité, ta patience et ta tendresse. Tu m'as toujours soutenu et réconforté, tu es et tu resteras toujours ma source d'encouragement.

À tous ma famille **Boutelala et Channibe.**

À tous mes chères ami(e)s

À Tous ceux qui ont contribué de loin ou de près à réaliser ce projet
Je vous dédie ce travail avec tous mes vœux de bonheur, de santé et de réussite.

AMINA

Table des matières

1	Prédiction dans les essais clinique de phase II	9
1.1	Essais cliniques	10
1.1.1	Différentes phases d'un essai clinique	11
1.1.2	Modèles statistiques dans les essais cliniques	12
1.2	Tests d'hypothèses paramétriques	12
1.2.1	Test d'une hypothèse simple avec alternative simple	13
1.2.2	Test du rapport de vraisemblance simple	16
1.2.3	Tests d'hypothèses multiples	19
1.3	Approche Bayésienne	20
1.3.1	Relation entre l'approche classique et bayésienne	20
1.3.2	Inférence bayésienne	23
1.3.3	Distribution a priori	24
1.4	Loi prédictive	26
2	Indice et prédiction de satisfaction	30
2.1	Introduction	30
2.1.1	La contrainte	30
2.1.2	Les insatisfactions	31
2.2	Statistique classique	31
2.2.1	Choix du modèle	31
2.2.2	Présentation des indices améliorés	32
2.3	Approche Hybride Fréquentiste- Bayésienne	33
2.3.1	Indice de satisfaction	34

2.3.2	Prediction de satisfaction	35
2.3.3	Prédiction de satisfaction pour le modèle exponentielle	36
3	Simulation	39
3.1	Méthode de Monte Carlo	39
3.1.1	Principe de la méthode	39
3.1.2	Contexte bayésien	40
3.2	Simulation du modèle exponentielle par Monte Carlo	41
	Conclusion Générale	41
	ANNEXE A	44
	ANNEXE B	49
	Résumé	50

Table des figures

1.1	Type d'erreur dans les tests hypothèses	13
1.2	La P-valeur dans un test unilatéral	15
1.3	Les étapes de base conduisant à la densité prédictive.	27
3.1	Prédiction de satisfaction dans le modèle exponentielle pour différentes valeurs de θ_0 , lorsque $\alpha = 0.05$; $k = 5$; $n = 5$	42
3.2	Prédiction de satisfaction dans le modèle exponentielle pour différentes valeurs de θ_0 , lorsque $\alpha = 0.05$; $k = 10$; $n = 5$	42
3.3	Prédiction de satisfaction dans le modèle exponentielle pour différentes valeurs de θ_0 , lorsque $\alpha = 0.05$; $k = 10$; $n = 10$	42
3.4	Prédiction de satisfaction dans le modèle exponentielle pour différentes valeurs de θ_0 , lorsque $\alpha = 0.1$; $k = 5$; $n = 5$	43
3.5	Prédiction de satisfaction dans le modèle exponentielle pour différentes valeurs de θ_0 , lorsque $\alpha = 0.1$; $k = 10$; $n = 5$	43
3.6	Prédiction de satisfaction dans le modèle exponentielle pour différentes valeurs de θ_0 , lorsque $\alpha = 0.1$; $k = 10$; $n = 10$	43

Introduction Générale

Les essais cliniques sont une étape nécessaire à la mise en service d'un médicament sur le marché. Ils permettent de déterminer les populations pour lesquelles le médicament est le plus efficace et les conditions optimales de son utilisation.([18],[10])

La mise en marché d'un médicament passe par plusieurs études différentes (ou phase cliniques) avec chacune objectifs différents. Dans notre cas nous nous intéressons à une étude de phase II qui consiste à estimer l'efficacité du médicament sous l'étude.

L'approche bayésienne apporte une grande souplesse à la méthodologie statistique des essais expérimentaux. Un des atouts majeurs du paradigme bayésien est la facilité avec laquelle on peut faire des prédictions sur les observations futures. En particulier, nous nous intéressons à l'utilisation de cette approche dans la prévision dans le contexte d'essais expérimentaux en raison du rôle essentiel que joue la probabilité prédictive dans la conception et la surveillance des essais (fiabilité des systèmes, médecine, biologie, écologie, . . .).([8],[13],[17])

Les procédures prédictives bayésiennes ont largement contribué à l'inférence et à l'analyse des données. Dans cette perspective, les probabilités prédictives bayésiennes constituent un outil particulièrement utile pour communiquer avec le chercheur. Ils leur donnent une méthode très attrayante pour répondre à des questions essentielles telles que : "Compte tenu des données actuelles, quelle est la probabilité que le résultat final soit en quelque sorte concluant, ou au contraire non concluant ?" Cette question est inconditionnelle en ce qu'elle nécessite toutes les valeurs possibles des paramètres. Alors que la méthode fréquentiste traditionnelle ne répond pas à ces questions, les probabilités prédictives leur donnent des réponses directes et naturelles. En particulier, les procédures prédictives peuvent être utilisées pour illustrer les effets de la planification d'une expérience avec un très petit échantillon et pour aider à la décision d'abandonner précocement une expérience.([8],[9])

L'objectif de ce mémoire est de démontrer l'importance de l'indice de satisfaction et sa prédiction bayésienne dans les essais expérimentaux. Par exemple pour définir des règles d'arrêt en terme d'efficacité dans la planification des essais cliniques de phase II, qui servent à contrôler la procédure de prendre de bonnes décisions.

Pour démontrer l'efficacité et la flexibilité des designs Bayésiennes, on calcul cet indice et sa prédiction dans le cas hybride (fréquentist, bayésienne) pour le modèle exponentielle en utilisant la distribution a priori conjuguée pour le paramètre inconnu avec une étude de simulation.

Ce mémoire est réparti en trois chapitres :

Le premier chapitre résume la prédiction dans les essais clinique de phase II, nous avons touché les essais cliniques, tests d'hypothèses paramétriques, l'approche bayésienne et la loi prédictive.

Le deuxième chapitre décrit l'aspect fréquentiste de l'indice de satisfaction et sa prédiction bayésienne, en appliquant cette approche sur le modèle exponentielle.

Dans le troisième chapitre, nous avons illustré notre point de vue par des applications numériques en choisissant la méthode de Monte Carlo pour la simulation de la prédiction de satisfaction.

Chapitre **1**

Prédiction dans les essais clinique de phase II

Introduction

Ce chapitre traite des aspects biostatistiques de la conception et l'analyse de la phase II des essais cliniques. Bien que la nature des essais cliniques de phase II varie considérablement selon les domaines thérapeutique et les institutions de recherche, ces essais sont généralement des études à petite échelle destinés à nous aider à décider de poursuivre l'évaluation clinique de thérapie expérimentale dans d'autres essais à plus grande échelle. Le développement de ces approches est l'objet de ce chapitre. La petite taille de l'échantillon et les fonctions décisionnelles sont des caractéristique nécessaires dans l'analyse des essais cliniques de phase II et exigent des approches statistique spéciales : soit fréquentistes ou bayésienne.

1.1 Essais cliniques

Définition 1.1.1 *Les essais cliniques sont des études médicales statistiques effectuées chez l'homme pour prouver la validité d'un nouveau traitement. La réalisation d'essais cliniques est une étape essentielle du développement de nouveaux médicaments. Un essai clinique permet d'étudier et d'assurer des propriétés de molécules afin de pouvoir en faire, par la suite, un usage sécurisé et responsable. Ces molécules présenteront, peut être, un nouveau progrès thérapeutique pour des centaines de milliers de patients et pour le corps médical. Les essais cliniques sont le plus souvent pratiqués en recherche médicale ou biomédicale, mais d'autres disciplines cliniques, telle la psychologie, mènent des travaux de recherche dans le but d'évaluer- ou habituellement de comparer - des interventions. Les essais cliniques sont envisagés sous l'angle de la recherche biomédicale et surtout des essais pharmaceutiques. Les recherches qui réalisent des essais cliniques poursuivent divers objectifs de différentes façons. De tels essais peuvent comprendre*

des questions indirectement reliées à une thérapie (rapport coût - efficacité, méthodisme des médicaments, ... etc). ([3],[4],[7])

1.1.1 Différentes phases d'un essai clinique

phase 0 (Études pré-clinique)

Elle consiste en l'étude de la molécule, sa structure, son effet sur les cellules, son effet sur l'animal au niveau comportemental et biologique, l'étude des organes-cibles. partir de ces études on détermine la dose maximale tolérée qui représente la dose maximale que l'animal de laboratoire peut tolérer, la dose sans effet observable et la dose sans effet toxique observable.

phase I

Une étude de phase I est le préliminaire à l'étude d'efficacité d'un médicament. Elle a lieu après la phase pré-clinique. Il s'agit d'évaluer la tolérance et l'absence d'effets indésirables chez des sujets le plus souvent volontaires sains, rémunérés pour cela. Parfois ces essais peuvent être proposés pour des patients en impasse thérapeutique, pour lesquels le traitement étudié représente la seule chance de survie.

Cette phase permet également d'étudier la cinétique et le métabolisme chez l'homme de la substance étudiée.

Les groupes étudiés sont le plus souvent de petite taille (20 à 80 participants).

Phase II

La phase II ou étude pilote consiste à déterminer la dose optimale du médicament et ses éventuels effets indésirables. Population éligible : malades (souvent < 500). Elle est subdivisée en deux phases : phases (II A) et (II B).

La phase (II A) estime l'efficacité de la molécule sur un nombre limité (de 100 à 200) de malades, alors que la phase (II B) détermine la dose thérapeutique de la molécule sur une plus grande échelle (de 100 à plus de 300 malades).

phase III

La phase III ou «étude pivot» est l'étude comparative d'efficacité proprement dite. Elle compare le traitement soit avec un placebo, soit avec un traitement de référence. Les groupes

sont de taille importante, souvent plusieurs milliers de participants. Il s'agit de programmes extrêmement onéreux, dont le financement peut être public ou privé (compagnies pharmaceutiques). Compte tenu des enjeux financiers, certaines dérives éthiques ont été dénoncées.

phase IV

La phase IV (ou post-marketing) est le suivi long terme d'un traitement alors que le traitement est autorisé sur le marché. Elle doit permettre de dépister des effets secondaires rares ou des complications tardives. Cette phase est la charge des laboratoires.

1.1.2 Modèles statistiques dans les essais cliniques

Un modèle statistique est une description mathématique approximative du mécanisme qui a généré les observations, que l'on suppose être un processus stochastique et non un processus déterministe. Il s'exprime généralement à l'aide d'une famille de distribution et d'hypothèses sur les variables aléatoires Y_1, \dots, Y_n .

Les distributions standard Binomiale, Normale, Poisson et exponentielle ont des dérivations des modèles de probabilités simples. La distribution Binomiale est motivée par le comptage des résultats échangeables, et la distribution Normale s'applique à une variable aléatoire qui est la somme de plusieurs termes échangeables et indépendantes.[4]

Les distributions de Poisson et exponentielles apparaissent comme le nombre de comptes et les temps d'attente, respectivement, pour modéliser les événements se produisant interchangeablement à tous les intervalles de temps ; c'est à dire indépendamment dans le temps, avec un taux d'occurrence constant.([4]) Dans l'annexe A, nous donnons certaines lois de probabilité utilisées dans les essais cliniques à travers la fonction de densité, l'espérance(E) et la variance(V).

1.2 Tests d'hypothèses paramétriques

Dans l'approche paramétrique, la plus générale, un test statistique consiste de décider d'accepter ou de rejeter une hypothèse spécifiant que θ appartient à un ensemble de valeurs Θ_0 . Cette hypothèse de référence est appelée hypothèse nulle et est notée H_0 . Au contraire on définit l'hypothèse alternative, notée H_1 , pour laquelle θ appartient à $\Theta_1 = \Theta - \Theta_0$ où $\Theta - \Theta_0$ dénote le complémentaire de Θ_0 par rapport à Θ . En bref on identifiera cette situation en écrivant que l'on teste :

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1,$$

Suivant la nature de Θ_0 et de Θ_1 on distinguera trois cas :

i) Hypothèse nulle simple et alternative simple où $\Theta = \{\theta_0, \theta_1\}$:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta = \theta_1$$

ii) Hypothèse nulle simple et alternative multiple :

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0$$

iii) Hypothèse multiple et alternative multiple :

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1$$

Pour une hypothèse nulle ou une hypothèse alternative multiple il est sous entendu qu'il y a plusieurs valeurs possibles de θ .

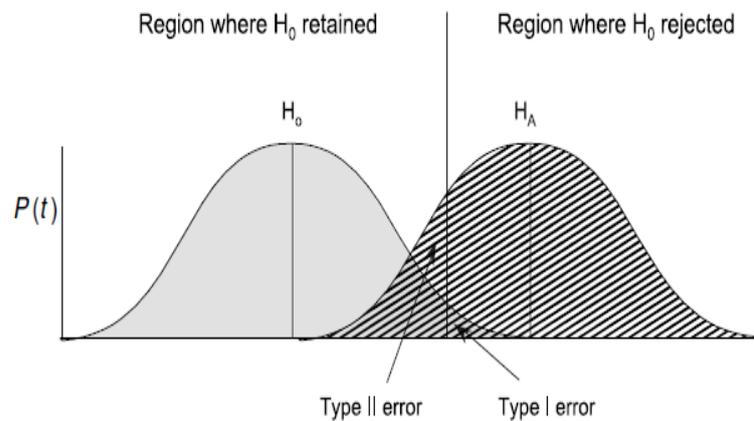


FIGURE 1.1 – Type d'erreur dans les tests hypothèses

1.2.1 Test d'une hypothèse simple avec alternative simple

L'espace paramétrique Θ ne comprend donc que deux valeurs θ_0 et θ_1 , la valeur θ_0 étant la valeur spécifiée tester, i.e. :

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta = \theta_1.$$

Un test pour H_0 est une règle de décision fondée sur la valeur réalisée t d'une statistique T appelée statistique de test. Sauf exception la statistique T sera à valeurs dans \mathbb{R} , nous le supposerons implicitement. La règle est comme suit :

Si $t \in A$ (une partie de \mathbb{R}) on accepte H_0 ,

Si $t \in \bar{A} = R$ (partie complémentaire) on rejette H_0 .

La région A , qui est généralement un intervalle, sera appelée région d'acceptation et R région de rejet.

Une telle règle de décision recèle deux types d'erreur possibles du fait que la vraie valeur du paramètre est inconnue :

Définition 1.2.1 *On appelle risque de première espèce la valeur α telle que :*

$$\alpha = P_{\theta_0}(T \in R),$$

c'est à dire la probabilité de rejeter H_0 alors quelle est vraie.

Il est usuel de noter cette probabilité $P(T \in R | H_0)$ même s'il ne s'agit pas là d'une probabilité conditionnelle et, dorénavant, nous adopterons cette notation. Le risque de première espèce est aussi appelé en bref risque erreur de type I.

Définition 1.2.2 *On appelle risque de deuxième espèce la valeur β telle que :*

$$\beta = P_{\theta_1}(T \in A),$$

c'est à dire la probabilité d'accepter H_0 alors que H_1 est vraie.

Ici également on notera cette probabilité $P(T \in A | H_1)$ et on parlera de risque β ou erreur du type II.

Définition 1.2.3 *On appelle puissance d'un test la probabilité de rejeter H_0 alors quelle est effectivement fautive soit, dans les notations précédentes :*

$$P(T \in R | H_1) = 1 - \beta.$$

La puissance, qui est la capacité à détecter qu'une hypothèse nulle est fautive.

Nous pouvons maintenant clairement exprimer notre exigence.

Définition 1.2.4 *On dit qu'un test est sans biais si sa puissance est supérieure ou égale à son risque α , soit :*

$$P(T \in R \mid H_1) \geq P(T \in R \mid H_0).$$

Définition 1.2.5 Un test ϕ^* est dit UPP—uniformément le plus puissant—de niveau α si :

i) ϕ^* est de niveau α :

$$\alpha = \sup_{\theta \in \Theta_0} \alpha_{\theta}(\phi^*)$$

ii) $1 - \beta_{\theta}(\phi^*) \geq 1 - \beta_{\theta}(\phi) \quad \forall \phi, \forall \theta \in \Theta_1.$ ([16])

La P-valeur

La décision d'accepter ou de refuser une hypothèse est sujette au choix du risque de première espèce α . Afin d'éviter ce choix on peut recourir, et c'est ce que font les logiciels, la notion de P-valeur pour simplement rendre compte du résultat d'un test. La P-valeur est la probabilité que, sous H_0 , la statistique de test prenne une valeur au moins aussi extrême que celle qui a été observée.

D'une façon générale, la P-valeur permet de déterminer si l'on rejette un niveau donné (à condition toutefois que, dans le cas bilatéral, la zone de rejet soit partagée en risque $\alpha/2$ équitablement sur chaque extrémité, ce qui est l'usage courant). Si la P-valeur est inférieure au risque α on rejette H_0 sinon on l'accepte. Comme autre façon de voir les choses on peut dire que plus la P-valeur est faible plus l'hypothèse nulle est suspecte. Ainsi l'indication des P-valeurs dans les logiciels a rendu obsolète l'usage des tables pour le praticien.

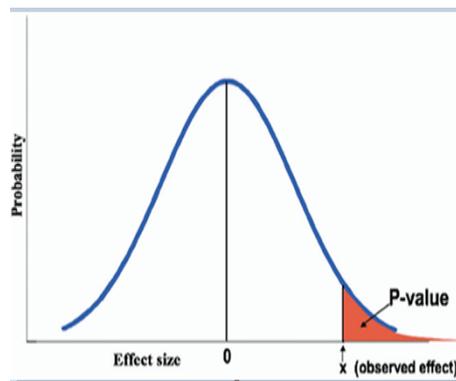


FIGURE 1.2 – La P-valeur dans un test unilatéral

Définition 1.2.6 La p-value associée à un test est le niveau de signification α le plus petit pour lequel l'hypothèse nulle est rejetée.

Dans le cas générale pour les hypothèses nulles ponctuelles (voir [16]) la p -value est une statistique admettant une loi uniforme sous l'hypothèse nulle ; pose alors un problème du choix de statistiques appropriée, comme d'ailleurs pour le test introduit dans la définition ci-dessus. En réalité, si un test de région critique R_α est disponible pour tout niveau de signification α et si ces régions sont imbriquées (c'est à dire si $R_\alpha \subset R_\beta$ pour $\beta > \alpha$), la procédure

$$p(x) = \inf\{\alpha; x \in R_\alpha\}$$

est distribuée selon une loi uniforme si $\mathbb{E}_{\theta_0}[\mathbb{I}_{R_\alpha}(x)] = \alpha$. Dans l'éventualité de plusieurs tests donnant des réponses opposées, nous suggérons d'utiliser la loi du rapport de vraisemblance sous l'hypothèse nulle, si cette dernière est ponctuelle. (voir section suivant)

Exemple 1.2.1 Si, pour $x \sim N(\theta, 1)$, on teste $H_0 : \theta = 0$ contre $H_1 : \theta \neq 0$, il n'existe pas de test UPP. Un test UPP sans biais au niveau $\alpha = 0.05$ est $\phi(x) = \begin{cases} 1 & \text{si } |x| \leq 1.96 \\ 0 & \text{sinon.} \end{cases}$

Puisque la région critique (qui est la région de rejet pour H_0) du test UPP sans biais est $\{|x| > k\}$, une p -value usuelle est

$$\begin{aligned} p(x) &= \inf\{\alpha; |x| > k_\alpha\} \\ &= P^x(|X| > |x|), X \sim N(0, 1) \\ &= 1 - \Phi(|x|) + \Phi(|x|) = 2[1 - \Phi(|x|)] \end{aligned}$$

où Φ est la distribution de la loi $N(0, 1)$

Par conséquent, si $x = 1.68$, $p(x) = 0.10$ et, $x = 1.96$, $p(x) = 0.05$.

1.2.2 Test du rapport de vraisemblance simple

Les rapports de vraisemblance décrivent également les qualités intrinsèques du test. Ils sont indépendants de la prévalence de la maladie et sont un bon indice de la «valeur diagnostique» d'un test.

A-Propriété d'optimalité

Définition 1.2.7 Soit un échantillon aléatoire (X_1, X_2, \dots, X_n) dont la loi mère appartient une famille paramétrique de densités (ou fonctions de probabilité) $\{f(x; \theta), \theta \in \Theta\}$ où $\Theta \subseteq \mathbb{R}^k$.

On appelle fonction de vraisemblance de θ pour une réalisation donnée (x_1, x_2, \dots, x_n) de l'échantillon,

la fonction de θ :

$$L(\theta; x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n, \theta) = \prod_{i=1}^n f(x_i, \theta)$$

Nous restons dans le cadre d'une hypothèse nulle et d'une hypothèse alternative simples, soit $\Theta = \{\theta_0, \theta_1\}$. On supposera dans cette section que le support de la densité $f(x; \theta)$ ne dépend pas de θ .

Définition 1.2.8 On appelle test du rapport de vraisemblance (RV) de l'hypothèse $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ au niveau α , le test défini par la région de rejet de la forme :

$$\frac{L(\theta_0; x_1, x_2, \dots, x_n)}{L(\theta_1; x_1, x_2, \dots, x_n)} < k_\alpha$$

où k_α est une valeur (positive) déterminée en fonction du risque de première espèce α .

Ce test a une certaine logique intuitive puisqu'il conduit à rejeter la valeur spécifiée θ_0 lorsqu'elle est moins vraisemblable que la valeur alternative θ_1 , car k_α (dont on admettra l'existence) se trouvera, en fait, être plus petit que 1 pour garantir un risque α faible. Notons que le rapport des deux vraisemblances (que l'on nomme rapport de vraisemblance) est bien une statistique puisque θ_0 et θ_1 sont donnés.

Exemple 1.2.2 Soit un échantillon aléatoire X_1, X_2, \dots, X_n issu d'une loi géométrique de paramètre inconnu p dont nous rappelons la fonction de probabilité :

$$f(x; p) = p(1-p)^x; x = 0, 1, 2, \dots$$

On a $\sum_{i=1}^n X_i$ suit une loi binomiale négative de paramètres n et p .

En effet $X_1 \sim \text{Géométrique}(p)$ est le nombre d'échecs avant le premier succès. Les v.a : X_1 et X_2 étant indépendantes, $X_1 + X_2$ peut être vue comme le nombre d'échecs avant le deuxième succès et ainsi de suite $X_1 + X_2 + \dots + X_n$ peut-être vue comme le nombre d'échecs avant le n -ième succès et ainsi $\sum_{i=1}^n X_i \sim BN(n, p)$.

Soit à tester :

$$H_0 : p = \frac{1}{3} \text{ vs. } H_1 : p = \frac{2}{3}$$

—La fonction de vraisemblance de p est $L(p; x_1, x_2, \dots, x_n) = p^n (1-p)^{\sum_{i=1}^n x_i}$, n étant connu.

Pour ce test le RV est :

$$\frac{L(\frac{1}{3})}{L(\frac{2}{3})} = \frac{(\frac{1}{3})^n (\frac{2}{3})^{\sum_{i=1}^n x_i}}{(\frac{2}{3})^n (\frac{1}{3})^{\sum_{i=1}^n x_i}} = (\frac{1}{2})^n 2^{\sum_{i=1}^n x_i}$$

et $\frac{L(\frac{1}{3})}{L(\frac{2}{3})} < k_\alpha$ équivaut à $\sum_{i=1}^n x_i < k_\alpha$. Donc la région critique pour le test fondé sur le rapport de vraisemblance est de la forme $\sum_{i=1}^n x_i < k_\alpha$

Si $n = 4$ et $\alpha = 0.05$ on a $T = \sum_{i=1}^n X_i \sim BN(4, \frac{1}{3})$ sous H_0 . Alors $P(T = 0) = C_3^0(\frac{1}{3})^4(\frac{2}{3})^0 = 0.0123$,

$P(T = 1) = C_4^1(\frac{1}{3})^4(\frac{2}{3})^1 = 0.0329$ et $P(T = 2) = C_5^2(\frac{1}{3})^4(\frac{2}{3})^2 = 0.0549$. On a donc un test conservateur de niveau 0.05 en rejetant H_0 si $\sum_{i=1}^n x_i \leq 1$.

La puissance du test est donnée par $P(T = 1/p = \frac{2}{3}) = (\frac{2}{3})^4 + 4(\frac{2}{3})^4(\frac{1}{3}) = 0.461$.

B-Cas d'un paramètre de dimension 1

Définition 1.2.9 On dit que T est une statistique exhaustive pour $\theta \in \Theta \subseteq \mathbb{R}^k$ si la loi conditionnelle de (X_1, X_2, \dots, X_n) sachant T ne dépend pas de θ .

Telle que $T = t(X_1, X_2, \dots, X_n)$, $X = (X_1, X_2, \dots, X_n)$, $P_\theta(X/t(X))$ n'est pas une fonction du paramètre θ .

Remarque 1.2.1 1. Le fait que la loi conditionnelle de X , sachant T , ne dépende pas de θ signifie que T contient toute information sur le paramètre inconnu θ .

2. En pratique, il est très difficile de répondre à la question s'il existe une statistique exhaustive ou non en utilisant cette définition. Mais, ce qui est plus ennuyeux c'est que cette définition ne donne aucune méthode pour construire des statistiques exhaustives. Il est donc très important d'avoir un critère simple qui permettrait de trouver des statistiques exhaustives.

Exemple 1.2.3 1. Soit $X \sim P(\theta)$. On a

$$f(x_1, \dots, x_n, \theta) = \frac{\exp(-n\theta)\theta^{\sum_{j=1}^n x_j}}{\prod_{j=1}^n (x_j!)}$$

et donc $T(X_1, \dots, X_n) = \sum_{j=1}^n X_j$ est bien une statistique exhaustive.

2. Le cas de la loi de Gauss $N(\mu, \sigma^2)$ où le paramètre de dimension 2, (μ, σ^2) , est inconnu.

On a :

$$f(x_1, \dots, x_n; \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right\}.$$

En développant $(x_i - \mu)^2$ et en regroupant les termes du produit on obtient :

$$f(x_1, \dots, x_n; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} \left(\frac{\sum_{i=1}^n x_i^2}{\sigma^2} - \frac{2\mu \sum_{i=1}^n x_i}{\sigma^2} + \frac{n\mu^2}{\sigma^2}\right)\right\}.$$

Le couple : $T(X_1, \dots, X_n) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ est une statistique exhaustive.

1.2.3 Tests d'hypothèses multiples

Risques, puissance et optimalité

Lorsque l'une des hypothèses H_0 ou H_1 est multiple les définitions de la section 1.2.1 doivent être revues. En effet si dans une hypothèse où plusieurs valeurs du paramètre sont possibles il n'y a plus de risque unique. Ainsi une expression telle que $P(T \in R \mid H_0)$ n'a pas de sens si H_0 est multiple. ([16])

Plaçons-nous dans le cas le plus général où l'on souhaite tester :

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1,$$

où $\Theta_1 = \Theta - \Theta_0$ est le complémentaire de Θ_0 par rapport à Θ . Comme précédemment, de la façon la plus générale, un test est défini par une région $A = \bar{R} \subset \mathbb{R}^n$ d'acceptation de l'hypothèse nulle H_0 . Nous supposons ici, comme cela se trouve en pratique, que cette région se réduit un intervalle A de \mathbb{R} pour une statistique de test T . Alors la règle de décision consiste à accepter H_0 si la valeur réalisée t de T appartient à A et rejeter H_0 sinon. Si H_0 est multiple le risque de première espèce $P(T \in R_\alpha)$ dépend de θ appartenant à Θ_0 . Le niveau du test est alors défini comme le risque maximal que l'on encourt à rejeter H_0 alors quelle serait fausse.

Définition 1.2.10 Soit $H_0 : \theta \in \Theta_0$ une hypothèse nulle multiple et $\alpha(\theta)$ le risque de première espèce pour la valeur $\theta \in \Theta_0$. On appelle niveau du test (ou seuil du test) la valeur telle que :

$$\alpha = \sup_{\theta \in \Theta_0} \alpha(\theta)$$

De même si l'hypothèse alternative $H_1 : \theta \in \Theta_1$ est multiple le risque de deuxième espèce est une fonction $\beta(\theta)$ ainsi que la puissance. On définit alors la fonction puissance du test :

$$h(\theta) = 1 - \beta(\theta) = P_\theta(T \in R)$$

définie pour tout $\theta \in \Theta_1$.

Définition 1.2.11 On dit qu'un test est sans biais si sa fonction puissance reste supérieure ou égale son niveau α , soit :

$$P_\theta(T \in R) \geq \alpha \text{ pour tout } \theta \in \Theta_1$$

En d'autres termes, la probabilité de rejeter H_0 si elle est fausse, quelle que soit la valeur de θ dans Θ_1 , est toujours plus élevée que la probabilité de la rejeter si elle est vraie, quelle que soit alors la valeur de θ dans Θ_0 .

1.3 Approche Bayésienne

L'approche bayésienne peut être présentée comme une généralisation de l'approche classique. Les paramètres ne sont plus des valeurs inconnues fixées, mais des variables aléatoires. Nous considérons que l'objet principal de la statistique est de tirer, au vu d'observations d'un phénomène aléatoire, une inférence au sujet de la loi générant ces observations ; afin, soit d'analyser un phénomène passé, soit de prévoir un événement futur, en nous attachant tout particulièrement aux aspects décisionnels de l'inférence bayésienne. (Thèse Djeridi)

1.3.1 Relation entre l'approche classique et bayésienne

Approche statistique classique

Dans l'approche fréquentiste, les données observées sont considérées comme des observations de variables aléatoires. Elles servent alors à faire porter l'inférence sur les paramètres Θ ayant dirigé leur mécanisme de génération. Autrement dit, l'information provenant des données observées est l'unique source d'information.

Dans cette approche, l'inversion est flagrante dans la notion de vraisemblance. En effet, on écrit

$$L(\theta; x) = f(x/\theta) = \prod f_i(x).$$

En considérant $L(\theta; x)$ comme une fonction de θ , on la normalise (quand cela est possible) pour en faire une fonction de densité sur x et on l'utilise dans l'estimation de θ . Par exemple, en estimation, on cherche la valeur $\hat{\theta}_{MV}$ qui maximise $L(\theta; x)$, c'est l'estimation au sens du maximum de vraisemblance :

$$\hat{\theta}_{MV} = \arg \max_{\theta \in \Theta} \{L(\theta; x)\}.$$

On utilise $L(\theta; x)$ comme si elle était une fonction de densité de probabilité conditionnelle aux observations x . Cette inversion est purement formelle alors que, dans l'approche bayésienne, comme nous le verrons un peu plus tard, cette inversion se fait d'une manière plus satisfaisante par la règle de Bayes : Si B et E sont des événements tels que $P(E) \neq 0$;

$$P(B/E) = \frac{P(E/B)P(A)}{P(E/B)P(B) + P(E/B^c)P(B^c)} = \frac{P(E/B)P(B)}{P(E)}$$

où B^c désigne le complémentaire de B .

Approche bayésienne

Dans la section précédente, nous avons considéré le cas où le paramètre inconnu θ est non aléatoire. Dans l'approche bayésienne, l'idée de base consiste traiter le paramètre inconnu θ comme une variable aléatoire admettant une densité de probabilité $p(\theta)$ qui s'appelle densité a priori.

L'objectif est donc d'utiliser cette information supplémentaire. Sachant que l'information contenue dans les observations x est contenue dans $f(x/\theta)$ et l'information a priori sur θ dans $p(\theta)$, on peut utiliser la règle de Bayes pour combiner ces deux types d'informations en définissant la densité a posteriori par :

$$p(\theta/x) = \frac{f(x/\theta)p(\theta)}{\int_{\Theta} f(x/\theta)p(\theta)d\theta}.$$

Qui contiendra donc toutes les informations sur θ .

On remarque que l'inversion de cause effet est ici beaucoup plus naturelle. Elle se fait d'une manière cohérente, car l'état de connaissance a priori sur θ traduite par la densité a priori $p(\theta)$ est transformé, après les observations x , en état de connaissance a posteriori par la densité a posteriori $p(\theta/x)$.

Remarquons que la densité a posteriori peut s'écrire

$$p(\theta/x) \propto f(x/\theta)p(\theta)$$

Modèle bayésien

On appelle modèle bayésien la donnée d'un modèle paramétrique, $f(x/\theta)$, et d'une densité a priori $p(\theta)$ sur les paramètres.

Étant donné la densité des observations $f(x/\theta)$ et la densité a priori $p(\theta)$ on peut construire :

i) La densité jointe de (θ, x) ,

$$\phi(\theta, x) = f(x/\theta)p(\theta);$$

ii) La densité marginale de x ,

$$m(x) = \int_{\Theta} \phi(\theta, x)d\theta = \int_{\Theta} f(x/\theta)p(\theta)d\theta;$$

et

iii) La densité a posteriori de θ , obtenue par formulation de Bayes,

$$p(\theta/x) = \frac{f(x/\theta)p(\theta)}{\int_{\Theta} f(x/\theta)p(\theta)d\theta}$$

$$= \frac{f(x/\theta)p(\theta)}{m(x)} \\ \propto f(x/\theta)p(\theta).$$

Autrement dit, la densité a posteriori représente une actualisation de l'information a priori au vu de l'information apportée par les observations.

Exemple 1.3.1 Si $X \sim P(\theta)$ (c-à-d, $f(x/\theta) = \frac{e^{-\theta} \theta^x}{x!}$) avec θ est inconnue $\theta \geq 0$ considérons la loi a priori de θ de type $G(a, b)$, $a > 0$, $b > 0$.

$$p(\theta) = \frac{b^a e^{-b\theta}}{\Gamma(a)} \theta^{a-1}, \theta > 0.$$

$$\Gamma(a) = \int_0^{+\infty} x^{a-1} e^{-x} dx, a > 0$$

où Γ est la fonction Gamma Calculons l'a posteriori $p(\theta/x)$:

$$f(x/\theta) = \prod_{i=1}^n f(x_i, \theta) \\ = \prod_{i=1}^n \frac{e^{-\theta} \theta^{x_i}}{x_i!} \\ = \frac{e^{-n\theta} \theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

et on a :

$$\phi(\theta, x) = f(\theta/x)p(\theta)$$

Alors :

$$\phi(\theta, x) = \frac{e^{-n\theta} \theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \cdot \frac{b^a e^{-b\theta}}{\Gamma(a)} \theta^{a-1} \\ = \frac{b^a e^{-(n+b)\theta} \cdot \theta^{\sum_{i=1}^n x_i + a - 1}}{\Gamma(a) \prod_{i=1}^n x_i!}$$

Donc :

$$m(x) = \int_{\Theta} \frac{b^a e^{-(n+b)\theta} \cdot \theta^{\sum_{i=1}^n x_i + a - 1}}{\Gamma(a) \prod_{i=1}^n x_i!} d\theta \\ = \frac{b^a}{\Gamma(a) \prod_{i=1}^n x_i!} \int_{\Theta} e^{-(n+b)\theta} \cdot \theta^{\sum_{i=1}^n x_i + a - 1} d\theta$$

On pose $x = (n + b)\theta \implies dx = (n + b)d\theta \implies d\theta = \frac{dx}{n + b}$

Donc :

$$m(x) = \frac{b^a}{\prod_{i=1}^n x_i! \Gamma(a) (n + b)^{\sum_{i=1}^n x_i + a}} \int_0^{+\infty} e^{-x} x^{\sum_{i=1}^n x_i + a - 1} dx$$

On pose $c = \sum_{i=1}^n x_i + a$

Alors :

$$m(x) = \frac{b^a \Gamma(\sum_{i=1}^n x_i + a)}{\prod_{i=1}^n x_i! \Gamma(a) (n+b)^{\sum_{i=1}^n x_i + a}}$$

Donc :

$$p(\theta/x) = \frac{e^{-(n+b)\theta} \theta^{\sum_{i=1}^n x_i + a - 1} (n+b)^{\sum_{i=1}^n x_i + a}}{\Gamma(\sum_{i=1}^n x_i + a)}$$

Par conséquent :

$$\theta/x \sim G(\sum_{i=1}^n x_i + a, n+b).$$

Relation entre la p-valeur et la probabilité a posteriori

Il existe une relation approximative entre la *p-valeur* et la probabilité a posteriori de H_0 . La décision de rejeter l'hypothèse nulle en fonction d'un seuil de significativité donné est équivalent à rejeter l'hypothèse nulle lorsque la probabilité a posteriori de H_0 est inférieure à un seuil d'évidence donné.

Cependant, le seuil classique de 5% utilisé en statistique fréquentiste correspond souvent à un seuil d'évidence plutôt faible dans le cas bayésien.

1.3.2 Inférence bayésienne

Une inférence bayésienne est fondée sur la détermination rigoureuse de trois facteurs :

1. La loi des observations $f(x/\theta)$.
2. La distribution a priori des paramètres $p(\theta)$.
3. La perte (le coût) associé aux décisions $l(\theta, \delta(x))$.

Notons que les critiques fréquentes de l'approche bayésienne se font souvent sur le point 2, alors que, conceptuellement, les points 1 et 3 se trouvent sur le même rang et que le point 3, qui est un point commun entre l'approche classique et l'approche bayésienne, est encore plus subtil.

Le choix d'une fonction de perte dépend de l'application concernée. Lorsque le contexte ne conduit pas la détermination de la fonction d'utilité qui permet d'en déduire la fonction de

coût, on peut avoir recours des fonctions de coûts classiques, qui sont la fois simples et bien étudiées.

1.3.3 Distribution a priori

L'aspect de l'analyse bayésienne, qui est à la fois le plus délicat et le plus critiqué, est certainement celui du choix de la loi a priori des paramètres. Le choix des lois a priori est une étape fondamentale dans l'analyse bayésienne. On entend par information a priori sur le paramètre θ toute information disponible sur θ en dehors de celle apportée par les observations. L'information a priori sur θ est entachée d'incertitude (si ce n'était pas le cas, le paramètre θ serait connu avec certitude et on n'aurait pas à l'estimer). Il est naturel de modéliser cette information a priori au travers d'une loi de probabilité appelée loi a priori, notée $p(\theta)$ sur Θ . Il existe plusieurs types de loi a priori :

- **Loi a priori conjuguée**

Une des difficultés de l'approche bayésienne est le calcul de la loi a posteriori. Ce calcul sera facilité lorsque la loi a priori et la loi a posteriori sont de la même famille. Dans ce cas, on parle de lois a priori conjuguées. L'approche a priori conjuguée, introduite par Raiffa et Schlaifer (1961), peut être justifié partiellement par un raisonnement d'invariance. En fait, quand l'observation de $x \sim f(x|\theta)$ modifie $p(\theta)$ en $p(\theta|x)$, l'information transmise par x sur θ est évidemment limitée ; par conséquent, elle ne devrait pas entraîner une modification de toute la structure de $p(\theta)$, mais simplement de ses paramètres. En d'autres termes, la modification résultant de l'observation de x devrait être de dimension finie (Robert, 2006).

Définition 1.3.1 *Une famille F de distributions de probabilité sur Θ est dite conjuguée pour une vraisemblance $f(x|\theta)$ si, pour toute loi a priori $p(\theta) \in F$, la distribution a posteriori $p(\cdot|x)$ appartient également à F .*

Remarque 1.3.1 *Le tableau suivant donne quelques lois a priori conjuguées naturelles pour quelques familles exponentielles usuelles.*

$f(x \theta)$	$p(\theta)$	$p(\theta x)$
Normale $N(\theta, \sigma^2)$	Normale $N(\mu, \tau^2)$	$N(\varrho(\sigma^2\mu + \tau^2x), \varrho\sigma^2\tau^2)$ $\varrho^{-1} = \sigma^2 + \tau^2$
Poisson $P(\theta)$	Gamma $G(a, b)$	$G(a + x, b + 1)$
Gamma $G(\nu, \theta)$	Gamma $G(a, b)$	$G(a + \nu, b + x)$
Binomiale $Bin(n, \theta)$	Bêta $Beta(a, b)$	$Beta(a + x, b + n - x)$
Binomiale Négative $Neg(m, \theta)$	Bêta $Beta(a, b)$	$Beta(a + m, b + x)$
Normale $N(\mu, 1/\theta)$	Gamma $G(a, b)$	$G(a + 0.5, b + (\mu - x)^2/2)$

• **Loi a priori non-informative**

Ces distributions sont conçues dans le but de faire de l'analyse bayésienne lorsqu'il y a absence d'information a priori sur le paramètre d'intérêt ou dans le cas où il est difficile de traduire en terme de loi a priori l'information disponible sur les paramètres en une loi de probabilité.

Ce choix repose sur l'équiprobabilité des valeurs possibles du paramètre θ sur son domaine de définition.

Dans le cas où l'espace des paramètres est discret et fini, par exemple, $\theta = (\theta_1, \theta_2, \dots, \theta_n)$. Alors, la distribution a priori non informative est

$$p(\theta_i) = \frac{1}{n}, \forall i = \overline{1, n}$$

Aussi, dans le cas où nous avons un espace du paramètre continue et borné, comme par exemple $\Theta = [a, b]$, $\infty < a < b < \infty$, alors la distribution a priori est uniforme

$$p(\theta) = \frac{1}{b - a}; a < \theta < b$$

donc il s'agit d'une distribution non informative pour θ .

Nous pouvons guider le lecteur vers les références suivantes pour voir les techniques (les plus utilisées) dans la construction de la loi a priori non informative (Thèse Djeridi, le choix bayésienne).

- **Loi a priori impropres**

Lorsque le paramètre θ peut être traité comme une variable aléatoire avec une distribution de probabilité p connue, nous avons vu que le théorème de Bayes est la base de l'inférence bayésienne, car il donne la distribution a posteriori.

Cependant, dans de nombreux cas, la distribution a priori est déterminée par des critères subjectifs ou théoriques qui conduisent à une mesure σ -finie sur l'espace des paramètres Θ , c'est à dire, une mesure p telle que :

$$\int_{\Theta} p(\theta) d\theta = +\infty$$

On dit que la distribution a priori est impropre (ou généralisée).

La difficulté pratique dans l'utilisation des lois impropres est de vérifier la condition d'intégrabilité

$$\int_{\Theta} f(x|\theta)p(\theta)d\theta.$$

L'estimateur de Bayes généralisé de θ qui vérifie la condition d'intégrabilité est la moyenne de la loi a posteriori.

1.4 Loi prédictive

Dans la pratique, nous sommes intéressés par les données de l'expérience future F et on veut utiliser nos reconnaissances sur le paramètre inconnu pour prévoir les probabilités du future échantillon, et ce, est le fond de l'approche bayésienne.

Le concept des problèmes de prédiction est de trouver la distribution de probabilité d'observation y d'un échantillon future F en donnant l'échantillon x de l'expérience informative F .

Définition 1.4.1 *Soit la fonction de densité prédictive pour une expérience futur F de fonction de densité*

$$\{p(y/\theta) : \theta \in \Theta\} \text{ sur } \mathcal{Y}$$

Une expérience informative E de classe de fonction de densité

$$\{p(x/\theta) : \theta \in \Theta\} \text{ sur } \mathcal{X}$$

Et une fonction de densité $p(\theta)$ sur Θ , la fonction de densité prédictive $p(y/x)$ pour F est définie par :

$$p(y/x) = \int p(y/\theta, x)p(\theta/x)d\theta$$

$$= \frac{\int p(y/\theta)p(\theta)p(x/\theta)d\theta}{\int p(\theta)p(x/\theta)d\theta}$$

Remarque 1.4.1 Figure (1.3) donne les différentes distributions utilisées dans le paradigme bayésien, (B) la posteriori et (A) la loi prédictive.

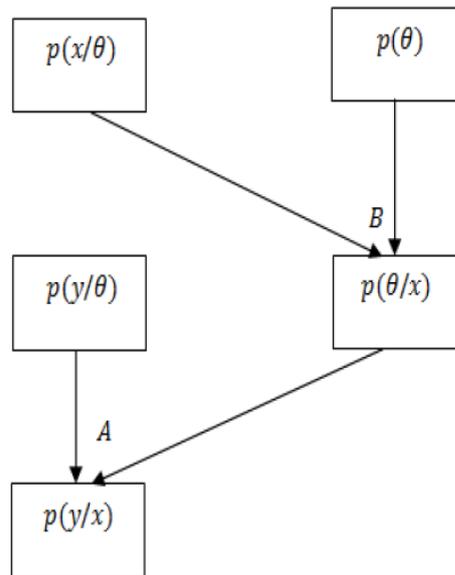


FIGURE 1.3 – Les étapes de base conduisant à la densité prédictive.

Exemple 1.4.1 (loi exponentielle)

a) La vraisemblance :

Considérons un n -échantillon sa fonction de vraisemblance est définie pour tout $\theta \in [0, +\infty[$ et pour tout $x = (x_1, \dots, x_n) \in \{0, 1, \dots, n\}$ par :

La fonction de densité : $f(x) = \theta e^{-\theta x} 1_{x>0}$

b) La loi a priori :

Dans ce cas on utilise la loi a priori $G(a, b)$ donnée par :

$$p(\theta) = \frac{b^a e^{-b\theta}}{\Gamma(a)} \theta^{a-1}, \theta > 0,$$

telle que : $\Gamma(a) = \int_0^{+\infty} x^{a-1} e^{-x} dx$, $a > 0$, est la fonction Gamma.

c) La loi a posteriori :

On a :

$$p(\theta/x) = \frac{\phi(\theta, x)}{m(x)}$$

$$\begin{aligned}
\bullet \phi(\theta, x) &= f(x/\theta)p(\theta) \\
&= \theta^n e^{-\theta \sum_{i=1}^n x_i} \cdot \frac{b^a e^{-b\theta}}{\Gamma(a)} \theta^{a-1} \\
&= \frac{\theta^{n+a-1} b^a e^{-\theta(b+\sum_{i=1}^n x_i)}}{\Gamma(a)}
\end{aligned}$$

$$\begin{aligned}
\bullet m(x) &= \int_{\Theta} \frac{\theta^{n+a-1} b^a e^{-\theta(b+\sum_{i=1}^n x_i)}}{\Gamma(a)} d\theta \\
&= \frac{b^a}{\Gamma(a)} \int_{\Theta} \theta^{n+a-1} e^{-\theta(b+\sum_{i=1}^n x_i)} d\theta
\end{aligned}$$

Posons : $x = \theta(b + \sum_{i=1}^n x_i) \implies \theta = \frac{x}{b + \sum_{i=1}^n x_i}$

Alors : $dx = (b + \sum_{i=1}^n x_i) d\theta \implies d\theta = \frac{dx}{(b + \sum_{i=1}^n x_i)}$

Donc :

$$\begin{aligned}
m(x) &= \frac{b^a}{\Gamma(a)(b + \sum_{i=1}^n x_i)^{n+a}} \int_0^{+\infty} x^{n+a-1} e^{-x} dx \\
&= \frac{b^a \Gamma(a+n)}{\Gamma(a)(b + \sum_{i=1}^n x_i)^{n+a}}
\end{aligned}$$

Donc la posteriori est :

$$\begin{aligned}
p(\theta/x) &= \frac{\theta^{n+a-1} b^a e^{-\theta(b+\sum_{i=1}^n x_i)}}{\Gamma(a)} / \frac{b^a \Gamma(a+n)}{\Gamma(a)(b + \sum_{i=1}^n x_i)^{n+a}} \\
&= \frac{\theta^{n+a-1} b^a e^{-\theta(b+\sum_{i=1}^n x_i)}}{\Gamma(a)} \times \frac{\Gamma(a)(b + \sum_{i=1}^n x_i)^{n+a}}{b^a \Gamma(a+n)} \\
&= \frac{\theta^{n+a-1} e^{-\theta(b+\sum_{i=1}^n x_i)} (b + \sum_{i=1}^n x_i)^{n+a}}{\Gamma(a+n)}
\end{aligned}$$

Posons : $g = a + n$ et $h = b + \sum_{i=1}^n x_i$

$$p(\theta/x) = \frac{\theta^{g-1} h^g}{\Gamma(g)} e^{-\theta h}$$

Par conséquent : $\theta/x \sim G(a+n; b + \sum_{i=1}^n x_i)$.

d) La loi prédictive :

Prenons un futur échantillon (y_1, \dots, y_{n_2}) iid de même $\exp(\theta) \implies y = \sum_{i=1}^{n_2} y_i \sim G(n_2, \theta)$ D'où la loi prédictive bayésienne :

$$\begin{aligned}
p(y/x) &= \int p(y/\theta, x) \pi(\theta/x) d\theta \\
&= \int \frac{\theta^{n_2} y^{n_2-1} e^{-\theta y}}{\Gamma(n_2)} \cdot \frac{\theta^{n_1+a-1} h^g e^{-\theta h}}{\Gamma(g)} d\theta \\
&= \frac{y^{n_2-1} h^g}{\Gamma(n_2) \Gamma(g)} \int \theta^{n_2+g-1} e^{-\theta(y+h)} d\theta
\end{aligned}$$

Posons $z = \theta(y + h) \Rightarrow dz = (y + h)d\theta$, et $G = n_2 + g$.

Alors :

$$\begin{aligned} p(y/x) &= \frac{y^{n_2-1}h^g}{\Gamma(n_2)\Gamma(g)} \int \left(\frac{z}{y+h}\right)^{G-1} e^{-z} \frac{dz}{y+h} \\ &= \frac{y^{n_2-1}h^g}{\Gamma(n_2)\Gamma(g)(y+h)^G} \Gamma(G) \\ &= \frac{h^g y^{n_2-1}}{B(n_2, g)(y+h)^G} \end{aligned}$$

Par conséquent : $p(y/x) \sim \text{InBeta}(n_2, g, h)$ (voir l'annexe A).

Le tableau suivant donne les distributions prédictives les plus utilisées

$p(x \theta)$ $p(y \theta)$ $p(y x)$	$p(\theta)$ $p(\theta x)$
1) Binomiale $\text{Bin}(n, \theta)$ $\text{Bin}(N, \theta)$ $\text{BetaBin}(N, g+x, h+n-x)$	$\text{Beta}(g, h)$ $\text{Beta}(g+x, h+n-x)$
2) Poisson $P(k \theta)$ $P(K \theta)$ $\text{NeBin}(g+x, \frac{K}{K+(h+k)})$	$G(g, h)$ $G(g+x, h+k)$
3) Gamma $G(k, \theta)$ $G(K, \theta)$ $\text{InBeta}(K, g+k, h+x)$	$G(g, h)$ $G(g+k, h+x)$
4) Multinomiale $\text{Mu}(n, \theta)$ $\text{Mu}(N, \theta)$ $\text{DiMu}(N, g+x, h+n-x)$	$\text{Di}(g, h)$ $\text{Di}(g+x, h+n-x)$

Indice et prédiction de satisfaction

2.1 Introduction

L'approche bayésienne apporte une grande souplesse à la méthodologie statistique des essais expérimentaux. Un des atouts majeurs du paradigme bayésien est la facilité avec laquelle on peut faire des prédictions sur les observations futures. En particulier, nous nous intéressons à l'utilisation de cette approche dans la prévision dans le contexte d'essais expérimentaux en raison du rôle essentiel que joue la probabilité prédictive dans la conception et la surveillance des essais (fiabilité des systèmes, médecine, biologie, écologie, . . .). ([8],[13]).

Le secteur d'application de la statistique nous est apparu caractérisé par une contrainte est des insatisfaction.

2.1.1 La contrainte

Elle est d'ordre légal et financier, puisque les essais clinique, par exemple dans la procédure d'autorisation de mise sur le marché d'un médicament(A.M.M.), comportant une phase très codifié de tests qui doivent avoir conclu, "significativement" à l'efficacité et à non-nocivité du produit.

Ces tests sont onéreux, et il y a souvent intérêt à ne les mettre en oeuvre que si on peut raisonnablement prédire qu'ils conduiront à des conclusions significatives, d'où le besoin "d'une première phase" d'essais, plus légère et moins encadrée. On peut donc dans une situation où :
– L'usage de la théorie classique des tests est imposé, la dissymétrie "hypothèse nulle/l'hypothèse alternative" étant sous entendue par un désir affirmé de pouvoir conclure en faveur de l'hypothèse alternative et

– L'utilisateur souhaite qu'on lui fournisse un indicateur préliminaire sur ses chances de voir se réaliser son désir ;

Alors on a besoin d'utiliser un fondement bayésien pour lui satisfaire. ([13])

2.1.2 Les insatisfactions

Elle tiennent à l'inadaptation des critères qui lui sont fournis par la "statistique fréquentiste" qui lui a été souvent présentée. Par exemple, "la fonction puissance" d'un test, la "fonction de risque" d'un estimateur lui parlent peu car la variable en est le paramètre, qui reste inconnu. Il privilégie des indicateurs qui, en fonction de "(qu'on a résultats expérimentaux obtenus) lui dirons à quel point la satisfaction est satisfaisant.

C'est ainsi que, quand il pratique un test, il ne se contente pas de dire si le résultat est significatif au niveau α , mais dans ce cas aime bien dire jusqu'à quelle valeur on pourrait descendre le niveau, c'est-à-dire augmenter la vérité du test en gardant un résultat significatif. Cette pratique est tout à fait courante mais se contente souvent d'être une indication "supplémentaire" comme la "*P - valeur*". "prédictive" indiquée dans la section(1.4) chapitre 1.

2.2 Statistique classique

Dans cette section, on formule le contexte du problème en donnant le modèle statistique et la prédiction de l'indice de satisfaction en utilisant les tests d'hypothèses classiques.

2.2.1 Choix du modèle

Soit $(P_\theta)_{\theta \in \Theta}$, une famille de probabilités sur un espace d'observation Ω , à tester l'hypothèse nulle H_0 contre l'hypothèse alternative H_1 . Dans la conception asymétrique classique des problèmes de test, une telle situation correspond en général, chez l'expérimentateur, à un désir de mettre en évidence un résultat significatif, c'est-à-dire pouvoir conclure au rejet de l'hypothèse nulle.

Pour adoptée une procédure de test fréquentiste, du niveau α , conduisant à partitionner Ω en une région de non rejet $A = \bar{R}^\alpha$ est une région de rejet R_1^α , un indice de satisfaction particulièrement simple, est la fonction indicatrice R_1^α , on est satisfait si le résultat est significatif au niveau α , mécontent sinon, par conséquent la fonction puissance peut s'interpréter alors comme la donnée, pour tout θ , de la satisfaction moyenne relativement à P_θ .

Mais très souvent les utilisateur souhaitent plutôt, face à un résultat qui leur paraît susceptible de conduire au rejet de H_0 , connaître quel est son degré de signification, c'est-à-dire savoir jusqu'à quel niveau le résultat apparaîtra significatif.

2.2.2 Présentation des indices améliorés

En utilisant le fait que toute technique de test raisonnable conduit à une famille de régions de non-rejet $\bar{R}^{(\alpha)}$ décroissant (au sens de l'inclusion) quand α croît, (c'est-à-dire, quand nos précaution s'atténuent), on utilisera alors un nouveau indice de satisfaction due à Merabet ([13]), noté $\phi^{(\alpha)}$ et est défini par :

$$\phi^{(\alpha)}(x) = \begin{cases} 0 & \text{si } x \in \bar{R}^{(\alpha)} \\ 1 - \sup_{x \in \bar{R}^{(\beta)}} \beta & \text{sinon.} \end{cases}$$

$\sup_{x \in \bar{R}^{(\beta)}}(\beta)$ est ce que les praticiens notent le p associée à x et est appelé la p -valeur ; et plus ce p est faible, plus le praticien considère que son résultat est significatif.

Sous des condition de régularité du test UUP, on a, pour tout α , $1 - \alpha = \inf_{\theta \in \Theta_0} P_{\theta}(\bar{R}^{(\alpha)})$; donc si $x \notin R_0^{(\alpha)}$, il vient donc

$$1 - \phi^{(\alpha)}(x) = \inf_{x \in \bar{R}^{(\beta)}} \inf_{\theta \in \Theta_{\theta}} P(\bar{R}^{(\beta)})$$

Il en est en particulier ainsi dans le cas usuel où, Θ et Ω étant des espaces topologiques, il existe trois fonctions réelles continues, ψ , ξ et g , définies respectivement sur Θ , sur Ω et sur $]0, 1[$, et un nombre réel t_0 vérifiant les propriétés suivantes :

- $\Theta_0 = \{\theta; \psi(\theta) \leq t_0\}$;
- pour tout θ , la loi de la variable aléatoire ξ sous P_{α} ne dépend que de $\psi(\theta)$ et est diffuse ;
- la famille des lois de ξ est stochastiquement croissante avec $\psi(\theta)$ au sens que, si $\psi(\theta_1) \leq \psi(\theta_2)$, on a, pour tout x ,

$$P_{\theta_1}([\xi \leq x]) \geq P_{\theta_2}([\xi \leq x]),$$

- pour tout α ,

$$\bar{R}^{\alpha} = \{x; \xi(x) \leq g(\alpha)\}.$$

Si on appelle frontière entre H_0 et H_1 , l'ensemble des θ tels que $\psi(\theta) = t_0$, $g(\alpha)$ est alors $(1 - \alpha)$ -fractile de la loi de ξ sur la frontière et l'indice de satisfaction $\phi^{(\alpha)}(x)$ n'est autre que

la valeur en $\xi(x)$ de la fonction la répartition de cette loi.

Dans la logique de l'introduction d'un indice de satisfaction, il est naturel de proposer de quantifier la valeur de la procédure de test, par la fonction de satisfaction moyenne, qui à tout $\theta \in \Theta_1$ associe $\int \phi^{(\alpha)} dP_\theta$, au lieu par la fonction puissance.

Remarque 2.2.1 *Cette valeur est particulièrement pertinente pour le statisticien, car elle indique comment la restriction à $\bar{R}^{(\alpha)}$ de la loi de ξ évolue quand θ s'éloigne de la frontière, à l'intérieur de Θ_1 . Elle est certes moins facile à expliciter analytiquement que la puissance, mais elle est accessible aux moyens de calcul dont dispose en général un laboratoire de statistique appliquée.*

2.3 Approche Hybride Fréquentiste- Bayésienne

Précision le contexte expérimental qui consiste en deux expérimentations successives, de résultats $x_1 \in \Omega_1$ et $x_2 \in \Omega_2$, qui sont en général menées indépendamment. Leurs lois dépendent, dans le cadre d'un modèle bien établi, d'un paramètre $\theta \in \Theta$; Seul x_2 sert à fonder la conclusion officielle de l'étude et détermine la satisfaction de l'utilisateur qu'on notera $\phi(x_2)$. Mais il est utile sur la base du résultat x_1 de la première phase, de prévoir ce que sera la satisfaction à l'issue de la seconde phase. Dans cette étude, cette prévision est effectuée dans un contexte bayésien, c'est à dire fondée sur le choix d'une probabilité a priori sur Θ . On note :

- $P_{\Theta \times \Omega_1 \times \Omega_2}$: probabilité sur $\Theta \times \Omega_1 \times \Omega_2$,
- P_Θ : probabilité a priori sur Θ ,
- $P_\Theta^{x_1}$: probabilité a posteriori sur Θ , fondée sur le résultat de la première phase,
- $P_{\Omega_2}^\theta$: loi d'échantillonnage de la deuxième phase.
- $P_{\Omega_2}^{x_1}$: probabilité sur Ω_2 , conditionnée par le résultat de la première phase x_1 .

(à ne pas confondre, bien sûr, avec $P_{\Omega_2}^{\theta, x_1}$ qui, puisque à θ fixé, les deux expérimentation sont indépendantes, n'est autre que $P_{\Omega_2}^\theta$).

2.3.1 Indice de satisfaction

On considère la satisfaction apportée par la seconde phase de l'expérimentation et on la prédit à l'aide de la première, on fait les calculs avec deux phase indépendantes mais c'est généralisable (en particulier la deuxième phase peut intégrer la première $x_2 = (x_1, x_2)$).

Nous nous plaçons dans le cadre où le statisticien "souhaite" observer un résultat significatif, c'est à dire rejeter une hypothèse nulle Θ_0 . Sa "satisfaction" sera donc plus grande en cas de rejet, et même en général d'autant plus grande que l'observation qui conduit à ce rejet est plus significative. Dans ce cas, on peut définir deux types d'indices de satisfaction :

A) Prédiction de indice fruste

Etant fixé α , soit un test de niveau α défini par la région critique $R_2^{(\alpha)}$. Un premier indice de satisfaction défini par Grouin (1995)([15]), est défini par :

$$\phi(x_2) = 1_{R_2^{(\alpha)}}(x_2);$$

Remarque 2.3.1 *Le défaut de cet indices ci-dessus est qu'il exprime une satisfaction en "tout ou rien".*

B) Prédiction de indice amélioré

Il est plus intéressant de prendre en compte jusqu'à quel niveau le résultat apparaîtra toujours significatif. On utilise donc un nouvel indice de satisfaction défini par la formule suivant :

$$\phi(x_2) = \begin{cases} 0 & \text{si } x_2 \in \bar{R}_2^{(\alpha)} \\ 1 - \inf\{\beta; x_2 \in R_2^{(\beta)}\} & \text{si } x_2 \in R_2^{(\alpha)} \end{cases}$$

Une situation standard est celle où il existe une application $\psi(\Theta \rightarrow R)$ telle que

$$\Theta_0 = \{\theta; \psi(\theta) \leq t_0\}$$

et où il existe aussi $\xi(\Omega_2 \rightarrow R)$ et $g(]0, 1[\rightarrow R)$ tels que

$$\bar{R}_2^{(\alpha)} = \{x_2; \xi(x_2) \leq g(\alpha)\}.$$

On suppose de plus que la loi de ξ sous $P_{\Omega_2}^\theta$ ne dépend que de $\psi(\theta)$ (on la note $Q_{\psi(\theta)}$) et que la famille des lois Q_t est stochastiquement croissante, dans le sens où ξ a de plus en plus tendance

à prendre de grandes valeurs lorsque $\psi(\theta)$ devient de plus en plus grand.

Soit alors G_t la fonction de répartition de Q_t . Il est clair que :

$$\phi(x_2) = \begin{cases} 0 & \text{si } x_2 \notin R_2^{(\alpha)} \\ G_{t_0}(x_2) & \text{si } x_2 \in R_2^{(\alpha)} \end{cases}$$

où G_{t^*} s'interprète comme la fonction de répartition "à la frontière" de ξ . La prévision est alors donnée par :

$$\begin{aligned} \pi(x_1) &= \int_{R_2^{(\alpha)}} \phi(x_2) P_{\Omega_2}^{x_1}(dx_2) \\ &= \int_{\Theta} [\phi(x_2) P_{\Omega_2}^{\theta}(dx_2)] P_{\Theta}^{x_1} d\theta \\ &= \int_{\Theta} \left[\int_{g(\alpha)}^{\infty} G_{t_0}(dt) dG_{\psi(\theta)}(x) \right] P_{\Theta}^{x_1}(d\theta). \end{aligned}$$

2.3.2 Prediction de satisfaction

Rappelons que, selon la terminologie bayésienne usuelle, on appellera probabilité prédictive la probabilité $P_{\Omega_1 \times \Omega_2}$ sur l'espace des résultats complets (Chapitre 1, section 1.4). Ce qui va nous servir par la suite est la probabilité sur Ω_2 qui s'en déduit par conditionnement par x_1 . On trouve comme prédiction, sur le vu de première phase, d'un résultat significatif lors de la seconde phase. C'est en la valeur $P_{\Omega_2}^{x_1}(R_2^{(\alpha)})$ où $R_2^{(\alpha)}$, la région de rejet du test opéré du résultat de la seconde phase. C'est ce qu'on se trouve ici pratiquer à la fois de la statistique classique et de la statistique bayésienne.

Tout indice de satisfaction relatif à la seconde phase est associé par un indice de prévision qui est l'espérance mathématique relativement à $P_{\Omega_2}^{x_1}$. Autrement dit ce qui va intéresser l'utilisateur potentiel de l'indicateur de satisfaction $\phi(x_2)$, c'est, à l'issue de la première phase, la prévision de sa valeur moyenne connaissant x_1 .

L'indicateur de prédiction est défini comme suit :

$$\pi(x_1) = \int_{\Omega_2} \phi(x_2) P_{\Omega_2}^{x_1}(dx_2), \quad (2.1)$$

où $\phi(x_2)$ est l'indicateur de satisfaction, ou encore :

$$\pi(x_1) = \int_{\Theta} \left(\int_{\Omega_2} \phi(x_2) P_{\Omega_2}^{\theta}(dx_2) \right) P_{\Theta}^{x_1}(d\theta). \quad (2.2)$$

Considérons le cas où on dispose de densités relativement à des mesures μ, ν_1, ν_2 sur Θ, Ω_1 et Ω_2 respectivement, celle de l'a priori P_Θ étant notée g et celles des probabilités d'échantillonnage $P_{\Omega_1}^\theta$, et $P_{\Omega_2}^\theta$ étant notées respectivement $f_1(\cdot/\theta)$ et $f_2(\cdot/\theta)$.

Les expressions (2.1) et (2.2) deviennent alors :

$$\pi(x_1) = \frac{\int_{\Omega_2} \phi(x_2) [\int_{\Theta} f_1(x_1|\theta) f_2(x_2|\theta) g(\theta) \mu(d\theta)] \nu_2(dx_2)}{\int_{\Theta} f_1(x_1|\theta) g(\theta) \mu(d\theta)},$$

et

$$\pi(x_1) = \frac{\int_{\Theta} [\int_{\Omega_2} \phi(x_2) f_2(x_2|\theta) \nu_2(dx_2)] f_1(x_1|\theta) g(\theta) \mu(d\theta)}{\int_{\Theta} f_1(x_1|\theta) g(\theta) \mu(d\theta)}.$$

à x_1 fixé, La prévision est alors

$$\pi(x_1) = P_{\Omega_2}^{x_1}(R_2^{(\alpha)}) = \int_{\Theta} P_{\Omega_2}^\theta(R_2^{(\alpha)}) P_{\Theta}^{x_1}(d\theta),$$

où $P_{\Omega_2}^\theta(R_2^{(\alpha)})$ est la valeur en θ de la puissance du test.

La prédiction de satisfaction de l'indice amélioré est donnée par :

$$\begin{aligned} \pi(x_1) &= \int_{R_2^{(\alpha)}} \phi(x_2) P_{\Omega_2}^{x_1}(dx_2) \\ &= \int_{\Theta} \left(\int_{R_2^{(\alpha)}} \phi(x_2) P_{\Omega_2}^\theta(dx_2) \right) P_{\Theta}^{x_1}(d\theta). \end{aligned}$$

On remarque que $\int_{R_2^{(\alpha)}} \phi(x_2) P_{\Omega_2}^\theta(dx_2)$ généralise la puissance du test dans la logique de l'indice de satisfaction proposé.

2.3.3 Prédiction de satisfaction pour le modèle exponentielle

Dû à l'importance de la modélisation de la fonction de survie par la loi exponentielle, où le paramètre θ , est le taux d'hasard ou la force de mortalité, on se propose de calculer, explicitement ou numériquement selon le cas l'indice et la prédiction de satisfaction dans modèle exponentiel, on choisit la loi a priori conjuguée pour le paramètre inconnu, θ , pour un test, au seuil, α , où l'hypothèse nulle est du type

$$\theta \leq \theta_0$$

Et conformément aux notations précédentes, on a donc :

$$\Theta_1 = \{\theta; \psi(\theta) > \theta_0\}$$

et il y a rejet de l'hypothèse H_0 , si le résultat expérimental, soit x_2 , vérifie

$$\xi(x_2) > g(\alpha)$$

où $g(\alpha)$ est le $(1 - \alpha)$ -fractile de la loi de ξ quand $\psi(\theta) = \theta_0$.

On effectue des observations indépendantes et de même loi, Le premier résultat est une suite X_1^1, \dots, X_k^1 de k observation et le second est aussi une suite X_1^2, \dots, X_n^2 de n observation. Tous les calculs sont fondés, pour des raisons d'exhaustivité, sur $x_1 = \sum_{i=1}^k X_i^1 \sim G(k, \theta_0)$.

et $x_2 = \sum_{i=1}^n X_i^2 \sim G(n, \theta_0)$. Dans tout ce qui suit, on a les notations simplifiées suivantes :

$p(\theta)$: densité de loi priori de θ ,

$p(x_2|x_1)$: densité de loi prédictive conditionnelle de x_2 sachant x_1 ,

F_{θ_0} : fonction de répartition de la loi de x pour la valeur θ_0 du paramètre.

Cas de la loi exponentielle

La fonction de densité : $f(x) = \theta e^{-\theta x} 1_{x>0}$; d'où $x_1 = \sum_{i=1}^k x_i^1 \sim G(k, \theta_0)$.

Dans ce cas, nous utilisons la loi a priori $G(a, b)$ donnée par :

$$p(\theta) = \frac{b^a e^{-b\theta}}{\Gamma(a)} \theta^{a-1}, \theta > 0,$$

telle que : $\Gamma(a) = \int_0^{+\infty} x^{a-1} e^{-x} dx$, $a > 0$ et $b > 0$ est la fonction Gamma.

Donc l'posterio est :

$$\begin{aligned} p(\theta/x_1) &= \frac{\theta^{k+a-1} b^a e^{-\theta(b+\sum_{i=1}^k x_i)}}{\Gamma(a)} / \frac{b^a \Gamma(a+k)}{\Gamma(a)(b+\sum_{i=1}^k x_i)^{a+k}} \\ &= \frac{\theta^{k+a-1} b^a e^{-\theta(b+\sum_{i=1}^k x_i)}}{\Gamma(a)} \times \frac{\Gamma(a)(b+\sum_{i=1}^k x_i)^{a+k}}{b^a \Gamma(a+k)} \\ &= \frac{\theta^{k+a-1} e^{-\theta(b+\sum_{i=1}^k x_i)} (b+\sum_{i=1}^k x_i)^{a+k}}{\Gamma(a+k)} \end{aligned}$$

Posons : $g = a + k$ et $h = b + \sum_{i=1}^k x_i = b + x_1$

$$p(\theta/x_1) = \frac{\theta^{g-1} h^g}{\Gamma(g)} e^{-\theta h}, \theta > 0$$

Par conséquent : $\theta/x \sim G(a+k; b+x_1)$, on aura :

L'indice de satisfaction est définie pour la deuxième phase par ;

$$\phi(x_2) = \begin{cases} 0 & \text{si } x_2 < q_0 \\ F(x_2) = \frac{\int_0^{x_2} t^{ba-1} e^{-t}}{\Gamma(a)} dt & \text{si } x_2 \geq q_0, \end{cases}$$

où $F(x_2)$ est la fonction de répartition de Gamma,
et q_0 est défini par

$$F(q_0) = 1 - \alpha,$$

et la loi prédictive du deuxième résultat x_2 sachant le premier x_1 est :

$$\begin{aligned} p(x_2|x_1) &= \int p(x_2|\theta, x_1)p(\theta|x_1)d\theta \\ &= \int \frac{\theta^n x_2^{n-1} e^{-\theta x_2}}{\Gamma(n)} \cdot \frac{\theta^{k+a-1} h^g e^{-\theta h}}{\Gamma(g)} d\theta \\ &= \frac{x_2^{n-1} h^g}{\Gamma(n)\Gamma(g)} \int \theta^{n+g-1} e^{-\theta(x_2+h)} d\theta \end{aligned}$$

Posons $z = \theta(x_2 + h) \Rightarrow dz = (x_2 + h)d\theta$, et $G = n + g = n + a + k$

$$\begin{aligned} p(x_2|x_1) &= \frac{x_2^{n-1} h^g}{\Gamma(n)\Gamma(g)} \int \left(\frac{z}{x_2 + h}\right)^{G-1} e^{-z} \frac{dz}{x_2 + h} \\ &= \frac{x_2^{n-1} h^g}{\Gamma(n)\Gamma(g)(x_2 + h)^G} \Gamma(G) \\ &= \frac{h^g x_2^{n-1}}{B(n, g)(x_2 + h)^G} \end{aligned}$$

Par conséquent : $p(x_2/x_1) \sim InBeta(n, g, h)$

Finalement la prédiction de satisfaction est donnée par :

$$\begin{aligned} \pi(x_1) &= \int_{q_0}^{+\infty} F_{\theta_0}(x_2)p(x_2|x_1)dx_1 \\ &= \int_{q_0}^{+\infty} F_{\theta_0}(x_2) \left[\frac{h^g x_2^{n-1}}{B(n, g)(x_2 + h)^G} \right] dx_1. \end{aligned}$$

qui peut être évaluée numériquement pour le chapitre 3.

Simulation

Introduction

Dans ce chapitre, nous allons simuler la prédiction de satisfaction, pour le modèle exponentielle. Pour ce faire, il faut utiliser les méthodes de simulations. Dans ce cas, nous avons choisit la méthode de Monte Carlo pour pouvoir simuler la loi prédictive Inverse Beta.

3.1 Méthode de Monte Carlo

La méthode de Monte Carlo, initialement élaborée par Nicholas Metropolis et Stanislaw Ulam, a été utilisée en 1940 pour le projet Manhattan au laboratoires de Los Alamos.[12] Elle est appliquée en 1966 à l'étude du transport électronique par Kurosawa. Quand il s'agit des systèmes complexes, il s'avère que l'application des méthodes analytiques devient très compliquée, d'où la nécessité d'appliquer la méthode de Monte-Carlo. Une des phases les plus importantes de cette méthode est la génération de nombres aléatoires.

Il existe des méthodes numériques d'intégration et d'optimisation qui peuvent être intéressantes pour résoudre les problèmes d'implémentation pratique de l'analyse bayésienne (voir Fletcher, 1987). Une autre alternative est donnée par les méthodes de simulation.

3.1.1 Principe de la méthode

soit $\theta \in \mathbb{R}$ de densité $p(\theta)$. On souhaite calculer $E(g(\theta))$ où $g(\theta)$ est un réel.

Le principe de la méthode de simulation de Monte-Carlo est de simuler une suite de v.a.

$\theta_1, \theta_2, \dots, \theta_L$ i.i.d. selon la loi de θ puis approcher $E(g(\theta))$ à l'aide de la loi forte des grands nombres par

$$\frac{1}{L} \sum_{i=1}^L g(\theta_i), \text{ lorsque } L \rightarrow \infty.$$

3.1.2 Contexte bayésien

On veut calculer $E(g(\theta|x))$, où g est une fonction d'intérêt.

On a

$$E(g(\theta|x)) = \int_{\Theta} g(\theta) p(\theta|x) d\theta$$

où $p(\theta|x) = \frac{f(x|\theta)p(\theta)}{f(x)}$ et $f(x) = \int_{\Theta} f(x|\theta)p(\theta)d\theta$,

D'où,

$$E(g(\theta|x)) = \frac{\int_{\Theta} g(\theta) f(x|\theta) p(\theta|x) d\theta}{\int_{\Theta} f(x|\theta) p(\theta) d\theta},$$

Ce qui revient à approcher l'intégrale $\int_{\Theta} g(\theta) f(x|\theta) p(\theta|x) d\theta$ par la méthode de simulation de Monte Carlo.

– On peut simuler $\theta_1, \theta_2, \dots, \theta_L$ selon la fonction d'importance $p(\theta)$. L'intégrale peut alors être approchée par :

$$\frac{1}{L} \sum_{i=1}^L g(\theta_i) f(x|\theta_i), \text{ lorsque } L \rightarrow \infty.$$

– On peut simuler $\theta_1, \theta_2, \dots, \theta_L$ selon la fonction d'importance $f(x|\theta)p(\theta|x)$ L'intégrale peut alors être approchée par :

$$\frac{1}{L} \sum_{i=1}^L g(\theta_i), \text{ lorsque } L \rightarrow \infty.$$

– On peut simuler $\theta_1, \theta_2, \dots, \theta_L$ selon la fonction d'importance $p(\theta|x)$. La quantité $E(g(\theta|x))$ peut alors être approchée, à l'aide de la loi forte des grands nombres, par :

$$\frac{1}{L} \sum_{i=1}^L g(\theta_i), \text{ lorsque } L \rightarrow \infty.$$

De plus, si la variance a posteriori $Var(g(\theta|x))$ est finie le Théorème Central-Limite (TCL) s'applique à la moyenne $E(g(\theta|x)) \simeq \frac{1}{L} \sum_{i=1}^L g(\theta_i)$ qui est asymptotiquement normale de variance $\frac{Var(g(\theta|x))}{L}$. Des régions de confiance peuvent alors être construites à partir de cette approximation. L'ordre de grandeur est de $\frac{1}{\sqrt{L}}$ quelque soit la dimension du problème, contrairement aux méthodes numériques.

3.2 Simulation du modèle exponentielle par Monte Carlo

Pour effectuer le calcul de $\pi(x_1)$ en utilisant la méthode de Monte Carlo, nous utilisons un changement de variable pour le réécrire sous la forme suivante :

$$\pi(x_1) = [1 - \Phi(q_0, n, g)] \int_0^{+\infty} (F(hz)) \frac{\varphi(z, n, g)}{[1 - \Phi(q_0, n, g)]} 1_{[q_0, +\infty[}(z)$$

où $z = \frac{x_2}{h}$, $F(q_0) = 1 - \alpha$, $g = a + k$, Φ est la fonction de répartition de la loi Inverse Beta, φ est la densité de probabilité de Inverse Beta, on a $\frac{\varphi(z, n, g)}{1 - \Phi(q_0)} 1_{[q_0, +\infty[}(z)$ est la densité de probabilité Q à partir de la fonction de distribution cumulative de la distribution de $InvBeta(n, g)$ conditionnée par l'événement $[q_0, +\infty[$

La méthode de Monte Carlo consiste alors à approcher $\pi(x_1)$ par :

$$[1 - \Phi(q_0, n, g)] \left[\frac{1}{N} \sum_{i=1}^N F(hz_i) \right]$$

Telle que les z_i sont N réalisations de la probabilité Q . Le tirage de z_i se fait par la fonction "betapr" qui utilise les package "Rcpp" et "ExtraDistr".

Le programme sous R est donné dans l'ANNEXE B.

Représentation des résultats et discussion

Nous trouvons, ci-dessus, les courbes représentatives de la prédiction de satisfaction, $\pi(x_1)$, en fonction des observations $x_1 = \sum_{i=1}^k X_i^1$. Nous considérons trois cas où $\theta_0 = 4 ; 5 ; 6$, et pour $\alpha = 0.05; 0.1$. D'autre part nous avons pris $n = 5; 10$, $k = 5; 10$, Pour la loi a priori conjuguée $G(a = 1; b = 2)$.

discussion

On peut constater dans les toutes les figure (3.1),..., (3.6), les affirmations suivantes :

- Lorsque θ_0 augmente, la prédiction de satisfaction augmente, ce qui dû au fait que la loi exponentielle est une fonction croissante selon θ .
- On peut remarquer, aussi, que lorsque la taille du premier échantillon augmente, la prédiction de satisfaction diminue, car la loi prédictive sera plus dégénérée.

En fin, lorsque l'erreur α augmente la prédiction de satisfaction augmente, car la p – valeur sera plus grande.

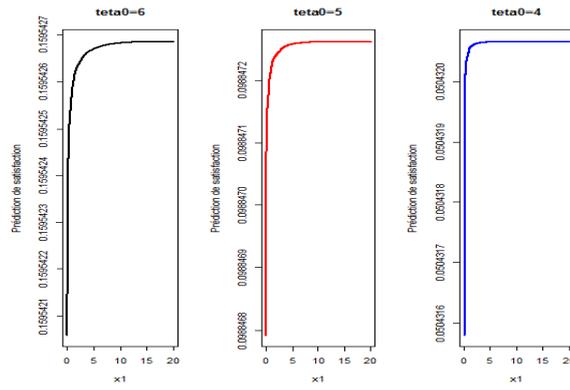


FIGURE 3.1 – Prédiction de satisfaction dans le modèle exponentielle pour différentes valeurs de θ_0 , lorsque $\alpha = 0.05$; $k = 5$; $n = 5$

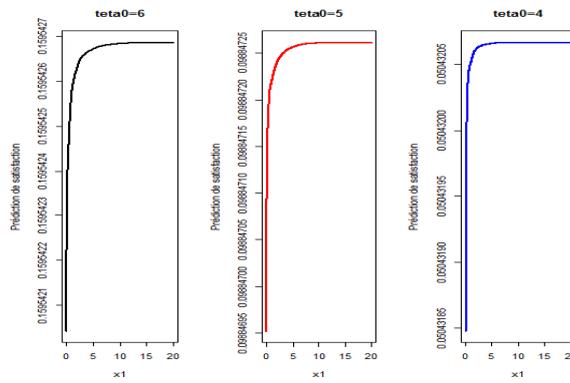


FIGURE 3.2 – Prédiction de satisfaction dans le modèle exponentielle pour différentes valeurs de θ_0 , lorsque $\alpha = 0.05$; $k = 10$; $n = 5$

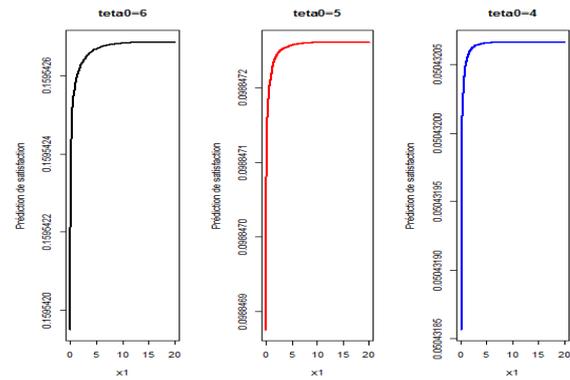


FIGURE 3.3 – Prédiction de satisfaction dans le modèle exponentielle pour différentes valeurs de θ_0 , lorsque $\alpha = 0.05$; $k = 10$; $n = 10$

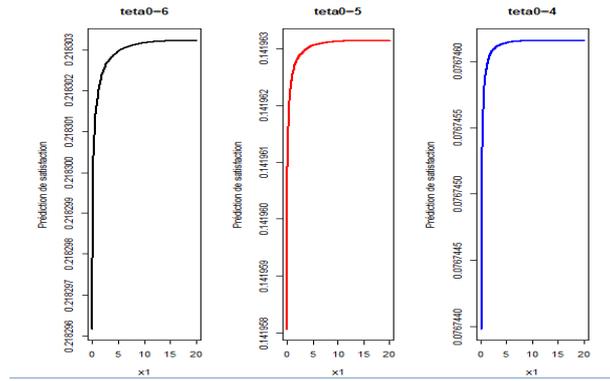


FIGURE 3.4 – Prédiction de satisfaction dans le modèle exponentielle pour différentes valeurs de θ_0 , lorsque $\alpha = 0.1$; $k = 5$; $n = 5$

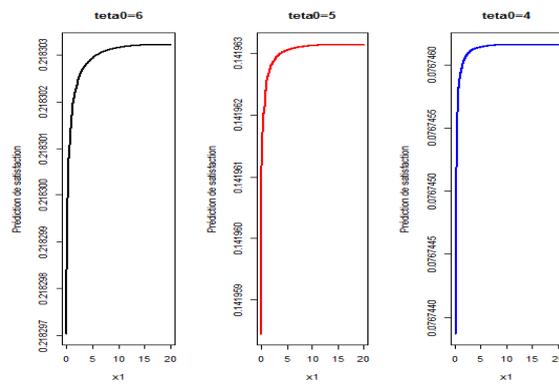


FIGURE 3.5 – Prédiction de satisfaction dans le modèle exponentielle pour différentes valeurs de θ_0 , lorsque $\alpha = 0.1$; $k = 10$; $n = 5$

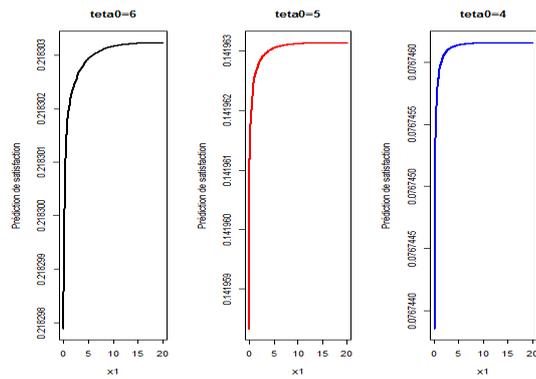


FIGURE 3.6 – Prédiction de satisfaction dans le modèle exponentielle pour différentes valeurs de θ_0 , lorsque $\alpha = 0.1$; $k = 10$; $n = 10$

Conclusion Générale

Ce travail consiste à appliquer une approche hybride (fréquentiste–bayésienne), en employant la méthode de Monte–Carlo pour calculer la prédiction de satisfaction lors d’une analyse intermédiaire dans un essai clinique de phase *II*. Le modèle choisit était le modèle exponentielle puisqu’il joue un rôle très important dans l’étude de la fonction de survie des patients.

La procédure utilisée est d’adapté une approche prédictive bayésienne pour conclure l’efficacité d’un traitement expérimentale, en se basant sur la *p – valeur* d’un test statistique fréquentiste.

Comme nous l’avons vu, la prédiction de satisfaction s’avère être un concept complémentaire, à celui de la puissance et nous recommandant systématiquement son usage pour la planification des essais. Le calcul numérique et les résultats de simulation ont confirmé la prédiction de satisfaction est une mesure concluant après le traitement des données comme le suggère la définition de la *p – valeur*.

ANNEXE A

• Binomiale et Bernoulli

Une variable binomiale discrète Y est la distribution d'échantillonnage du nombre total de succès dans n essais bernoulli indépendants, chacun avec une probabilité de succès θ . La probabilité $\theta^y(1 - \theta)^{n-y}$ donne la probabilité d'une séquence spécifique de $(n - y)$ échecs et y succès, il existe C_n^y de telles séquences. Ainsi $Y \sim \text{Bin}(n, \theta)$ représente une distribution binomiale avec des propriétés :

$$P(y|n, \theta) = C_n^y \theta^y (1 - \theta)^{n-y}; y = 0, 1, \dots, n,$$

$$E(Y|n, \theta) = n\theta,$$

$$V(Y|n, \theta) = n\theta(1 - \theta).$$

La binomiale avec $n = 1$ est simplement une distribution de bernoulli, notée $Y \sim \text{Bern}(\theta)$.

• Binomiale négative

En probabilité et en statistiques, la loi binomiale négative est une distribution de probabilité discrète. Elle décrit la situation suivante : une expérience consiste en une série de tirages indépendants, donnant un « succès » avec probabilité p (constante durant toute l'expérience) et un « échec » avec une probabilité complémentaire. Cette expérience se poursuit jusqu'à l'obtention d'un nombre donné n de succès. La variable aléatoire représentant le nombre d'échecs (avant l'obtention du nombre donné n de succès) suit alors une loi binomiale négative. Ses paramètres sont n , le nombre de succès attendus, et p , la probabilité d'un succès.

La loi se généralise à deux paramètres r et p , où r peut prendre des valeurs réelles strictement

positives. Cette généralisation est aussi connue sous le nom de loi de Pólya, en l'honneur du mathématicien George Pólya.

$$P(Y = n) = C_{n-1}^{r-1} p^r (1-p)^{n-r},$$

$$E(Y|r, p) = \frac{r(1-p)}{p},$$

$$V(Y|r, p) = \frac{r(1-p)}{p^2}.$$

• Poisson

Supposons qu'il existe un grand nombre d'opportunités pour qu'un événement se produise, mais que la chance qu'un événement particulier se produise est très faible. Alors le nombre total d'événements occurrents peuvent souvent être représentés par une variable aléatoire discrète Y , telle que $Y \sim P(\theta)$ représente la distribution de Poisson avec les propriétés :

$$P(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}; y = 1, 2, 3, \dots, n,$$

$$E(Y|\theta) = \theta,$$

$$V(Y|\theta) = \theta.$$

Dans de nombreuses applications, il apparaîtra comme un nombre total d'événements se produisant dans une période de temps T , où les événements se produisant à un taux inconnu λ par unité de temps, auquel cas la valeur attendue de Y est $\theta = \lambda T$.

• Géométrique

La loi Géométrique est une loi de probabilité discret qui modélise l'observation du nombre d'épreuves de Bernoulli identiques et indépendantes devant se succéder pour espérer un premier succès. $Y \sim \text{Géométrique}(p)$ représente une distribution de Géométrique avec les propriétés :

$$P(y|p) = pq^{n-1} \qquad p \in [0, 1], q = 1 - p,$$

$$E(Y|p) = \frac{1}{p},$$

$$V(Y|p) = \frac{q}{p^2}.$$

- **Bêta**

La distribution Bêta forme une classe flexible et mathématiquement pratique pour les quantités situant entre 0 et 1, et peut donc être utilisée comme la distribution a priori pour les proportions inconnues (voir section 1.3). $Y \sim \text{Bêta}(a, b)$ représente une distribution avec les propriétés :

$$P(y|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1} (1-y)^{b-1}; y \in (0, 1),$$
$$E(Y|a, b) = \frac{a}{a+b},$$
$$V(Y|a, b) = \frac{ab}{(a+b)^2(a+b+1)}.$$

- **Bêta prime**

En théorie des probabilités et on statistique, la loi Bêta prime (également connue sous les noms loi Bêta II ou loi Bêta du second type) est une loi de probabilité continue définie dont le support est $]0, \infty[$ et dépendant de deux paramètres de forme. Si une variable aléatoire Y suit une loi Bêta prime, on notera $Y \sim \beta'(a, b)$ représente une distribution avec les propriétés :

$$P(y|a, b) = \frac{y^{a-1}(1+y)^{-a-b}}{B(a, b)},$$
$$E(Y|a, b) = \frac{a}{b-1} \quad b > 1,$$
$$V(Y|a, b) = \frac{a(a+b-1)}{(b-2)(b-1)^2} \quad b > 2.$$

- **Gamma**

La distribution gamma forme une classe flexible et mathématiquement pratique pour quantités contraintes d'être positive. $Y \sim G(a, b)$ représente une distribution gamma avec les propriétés :

$$P(y|a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}; y \in [0, \infty[,$$
$$E(Y|a, b) = \frac{a}{b},$$
$$V(Y|a, b) = \frac{a}{b^2}.$$

Les cas particuliers incluent la distribution $G(1, b)$, qui est exponentielle avec la moyenne $\frac{1}{b}$, et $G(\frac{1}{2}v, \frac{1}{2})$, qui est la distribution χ_v^2 avec v degrés de liberté. Un élément utile de la théorie de la distribution est que si Y_1, \dots, Y_n sont des variables aléatoires i.i.d. de la loi $N[\theta, \sigma^2]$ de moyenne \bar{Y} et une variance $S^2 = \sum_i (Y_i - \bar{Y})^2/n$, puis $\sum_i (Y_i - \theta)^2/\sigma^2 \sim \chi_n^2$, et $nS^2/\sigma^2 \sim \chi_{n-1}^2$. Une justification est que la distribution gamma se "conjugue" à la Poisson famille. Cependant, comme pour les données binaires, dans la plupart des applications, il est beaucoup plus flexible et pratique de transformer la quantité d'intérêt d'un taux (défini sur $]0, \infty[$ en un log-rate (défini sur $]-\infty, +\infty[$, puis utiliser l'approximation normale.

• Normale

La distribution de probabilité normale (gaussienne) est fondamentale pour la plupart des analyses statistiques. $Y \sim N(\mu, \sigma)$ représente une distribution normale avec propriétés :

$$P(y|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right),$$

$$E(Y|\mu, \sigma) = \mu \qquad -\infty \leq \mu \leq +\infty,$$

$$V(Y|\mu, \sigma) = \sigma^2 \qquad \sigma^2 > 0.$$

Remarque 3.2.1 *Le lecteur peut consulter les références suivantes pour plus de distributions de probabilités ([1], [7], [16]).*

ANNEXE B

Programme

```
> teta0 = 0.5; a = 1; b = 2; k = 5; n = 5; g = a + n
####Estimation de  $p_i(x)$ ####
> pix = function(x, teta0){
alpha = 0.05
q = qgamma(1 - alpha, n, teta0)
q
x1 = rexp(k, teta0)
g = a + k
h = b + sum(x1)
N = 50
V = runif(N, pbetapr(q, n, g, 1), 1)
Z = qbetapr(V, n, g, 1)
px = (1 - pbetapr(q, n, g, 1)) * mean(pgamma(Z * h, n, teta0))
px
}
> density = pix(5, 5)
> density
> par(mfccol = c(1, 3))
> x = seq(0, 20, by = 0.01)
> density = 0
> teta0 = 6
```

```
> for(i in 1 : length(x)){  
# print(i)  
density[i] = pix(x[i], teta0)}  
> plot(x, sort(density),type = "l",xlab = "x1",ylab ="Prédiction de satisfaction",main =  
"teta0 = 6")  
> density = 0  
> teta0 = 5  
> for(i in 1 : length(x)){  
density[i] = pix(x[i], teta0)}  
> plot(x, sort(density),type = "l",xlab = "x1",ylab ="Prédiction de satisfaction",main =  
"teta0 = 5")  
> density = 0  
> teta0 = 4  
> for(i in 1 : length(x)){  
density[i] = pix(x[i], teta0)}  
> plot(x, sort(density),type = "l",xlab = "x1",ylab ="Prédiction de satisfaction",main =  
"teta0 = 4")
```

Résumé

L'objet de ce mémoire, qui a été motivée par des problèmes méthodologiques qui se posent dans le contexte des essais expérimentaux, est l'application d'une approche hybride bayésienne-fréquentiste utilisant la prédiction bayésienne de l'indice de satisfaction pour le modèle exponentielle, ce qui est souvent réalisé dans le cas de protocole d'essais cliniques.

Pour une analyse à deux étapes portant sur l'hypothèse de prédiction, telle que la surveillance de la futilité d'un traitement expérimental avec des résultats retardés, la prédiction de satisfaction prend en compte la quantité de données restant à observer dans un essai clinique et offre la possibilité d'intégrer des informations supplémentaires via la loi a priori.

Abstract

The aim of this memory, which was motivated by methodological problems that arise in the context of experimental trials, is to applicat a hybrid Bayesian-frequentist approach using Bayesian prediction of the index of satisfaction for the exponential model, which is often done in the case of clinical trial protocol.

For a two steps analysis of the prediction hypothesis, such as futility monitoring of the experimental treatment with delayed results, the prediction of satisfaction takes into account the amount of data remaining to be observed in a clinical trial and offers the possibility of integrating additional information via a priori distribution.

Bibliographie

- [1] **Aitchison, J.**, Dunsmore, I.R.(1975). Statistical Prediction Analysis. New York : Com-
pridge University Press. ISBN : 978-0-521-20692-1.
- [2] **Berger, J.**, 1985. Statistical decision theory and bayesian analysis. Second édition.
Springer-Verlag, New York.
- [3] **Berry, D.A.**, (2011) Bayesian Adaptive Methods for Clinical Trials, Miami Beach, FL,
August 2.
- [4] **Biswas, A., Sujay, D., Jason, P., and Mark, R.**, Statistical Advances in the Biomedical
Science : Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics.
- [5] **Bouvenot, G.**, and **Vray, M.**, (2006). Essais Clinique (théorie, pratique et critique)
Médecine Sciences Publication. Paris, Lavoisier SAS ; ISBN : 978-2-257-14224-5.
- [6] **Christopher, J.**, Group Sequential Methods with Application to Clinical Trials. London
New York. University of Bath, UK.
- [7] **David J, Spiegelhalter** and **Keith R, Abrams.**, and **Jonathan P. Myles.**, Bayesian
Approaches to Clinical Trials and Health-Care Evaluation.
- [8] **Djeridi, Z.**, (2019). Choix de l'a priori en statistique Bayésienne dans les essais
expérimentaux, Doctorat.
- [9] **Gerry P, Quinn.**, and **Michael J, Keough.**, Experimental Design and Data Analysis
for Biologists.
- [10] **Guosheng, Y., Nan Chen., and Jack Lee, J.**, Phase II trial design with Bayesian
adaptive randomization and predictive probability.
- [11] **Kelly, K.**, and **Halabi. S.**, (2010). Oncology Clinical Trials. Demos Medical Publishing,
LLC, New York.

- [12] **Kruschke, K.**, Doing Bayesian Data Analysis : A with R and BUGS.
- [13] **Merabet, H.**, 2004. Prevision de satisfaction pour la methodologie des essais preliminaires, Doctorat.
- [14] **Nathan, M.**, (2016). Etude de consistance et applications du modèle Poisson-Gamma : modélisation d'une dynamique de recrutement multicentrique, Doctorat.
- [15] **Lecoutre, B., Derzko, G ., and Grouin, J. M.**, (1995). Bayesian Predictive Approach for Inference About Proportion, *Statistics in Medecine* 14 : 1057-1063.
- [16] **Philippe Tassi.**, METHODES STATISTIQUES., 2 ÉDITION., Ed. ÉCONOMICA., (1989)., Paris.
- [17] **Robert, C. P.**, (2006). Le choix bayésien : principes et pratique. Springer-Verlag France, Paris.
- [18] **Saliby,E**, Descriptive sampling : an improvement over latin hypercube sampling. In Winter simulation Conference (1997), pp. 230-233.
- [19] **Scott M, Berry., Bradley P, Carlin., and J,Jack Lee.** Bayesian Adaptive Methods for Clinical Trials.
- [20] **Shao, Y., Mukhi, V., and Goldberg, J. D.**, (2008). A hybrid Bayesian-Frequentist Approach to Evaluate Clinical Trial Designs for Tests of and Non-Inferiority, *Statistics in Medicine* 27 :504-519.
- [21] **Simone, B.**, "Les Essais clinique : pourquoi sont-ils aujourd'hui critiqués", pour la science, juillet 2009.
- [22] **Zidek, J., Tanner, T., and Julian J.** Bayessian Data Analysis : Third Edition.