M1 IAD
Notes du cours RFIDEC (2)

Jean-Yves Jaffray

8 octobre 2007

1 La statistique inférentielle

1.1 Introduction

La *statistique inférentielle* a pour hypothèse de base que les observations (les données) $(x_1, x_2, \cdots, x_k, \cdots, x_n)$ ne sont qu'une réalisation d'une variable aléatoire multi-dimensionnelle, $X = (X_1, X_2, \cdots, X_k, \cdots, X_n)$ appelée *échantillon empirique* ; cette hypothèse nous permet d'étudier les propriétés des échantillons dans le cadre du modèle probabiliste.
On suppose de plus en général que les n variables X_k ont même loi et sont mutuellement indépendantes, ce qui entraîne l'existence de relations simples entre les caractéristiques de la loi commune aux X_k et celles des lois de variables fonctions de l'échantillon, comme la moyenne empirique $\bar{X}$.
En statistique inférentielle, l'incertitude porte sur la véritable loi suivie par une variable X_0 *dont est tiré l'échantillon*, ce qui veut dire que chacune des variables X_k suit la même loi que X_0. On sait seulement que cette loi appartient à un certain ensemble de lois. Lorsque cet ensemble peut être décrit comme une famille de lois se distinguant les unes des autres par la valeur d'un (ou de plusieurs) paramètre(s), on est dans le cadre de la *statistique paramétrique* : par exemple, l'échantillon peut être tiré d'une loi normale $\mathcal{N}(m, 1)$ de variance connue et d'espérance m inconnue mais localisée dans un intervalle $[m_0, m_1]$.
On peut alors s'interroger sur la vraie valeur du paramètre m dans $[m_0, m_1]$; c'est un problème d'*estimation ponctuelle*.
On peut aussi se demander si la vraie valeur de m est plutôt une valeur m_0 qu'une autre valeur m_1, ou bien le contraire. C'est un *test d'hypothèse*. D'autres catégories de tests aident à répondre à la question de l'indépendance de deux variables (*tests d'indépendance*), ou à décider si l'on peut considérer ou non que la loi de X_0 est bien une loi donnée (*tests d'ajustement*) ou encore si deux échantillons distincts sont bien tirés de la même loi (*tests de comparaison de moyennes ou de variances*) ; etc. Notons que

dans ces exemples, il faut trancher entre une hypothèse bien spécifique et l'hypothèse contraire, qui ne l'est pas ; on est typiquement dans le domaine de la *statistique non-paramétrique*.
Pour résoudre tous ces problèmes, deux approches distinctes sont couramment utilisées : celle de la *statistique classique* et celle de la *statistique bayésienne*. Leur différence essentielle est dans le traitement des paramètres : en statistique classique un paramètre a une valeur fixe mais inconnue ; en statistique bayésienne, c'est une variable sur laquelle on a une information fluctuante, exprimée sous forme d'une loi de probabilité sur l'espace des valeurs possibles de ce paramètre.
Nous allons commencer par décrire l'approche classique dans les problèmes d'estimation ponctuelle puis de tests.

2 La statistique inférentielle

2.1 Estimation ponctuelle en statistique classique

2.1.1 Exemple introductif : estimation d'une moyenne

Supposons que la famille de lois $\mathcal{P}$ à laquelle appartient la vraie loi de la variable X_0 dont est tiré l'échantillon dépend d'un paramètre θ, que la valeur de ce paramètre n'est autre que l'espérance mathématique de la loi correspondante et enfin que toutes ces lois ont même variance σ^2 :
si les lois de $\mathcal{P} = \{P_\theta \,:\, \theta \in \Theta\}$ ont des densités $\{p_\theta \,:\, \theta \in \Theta\}$ et que l'on note $E_\theta(X_0)$ l'espérance de X_0 si sa loi est P_θ, on a donc :

$$\forall \theta \in \Theta,\ E_\theta(X_0) = \int x_0\, p_\theta(x_0)\, dx_0 = \theta.$$

Quelle est la valeur vraie θ_0? On l'ignore, mais une indication sur sa valeur peut nous être donnée par la moyenne de l'échantillon, $\bar{x}$. En effet, des propriétés, vues précédemment, de la variable moyenne empirique $\bar{X}$, dont $\bar{x}$ est une réalisation, il résulte que

$$\forall \theta \in \Theta,\ E_\theta(\bar{X}) = \theta$$

et que (nous écrivons $\bar{X}^{(n)}$ au lieu de $\bar{X}$ quand n varie)

$$\forall \theta \in \Theta,\ V_\theta(\bar{X}^{(n)}) = E_\theta(\bar{X}^{(n)} - \theta)^2 = \frac{\sigma^2}{n}.$$

On est alors justifié à penser que si n est assez grand, la valeur $\bar{x}$ observée sera le plus souvent proche de la valeur vraie θ_0 du paramètre puisque :
i) en moyenne, elle vaudra θ_0 ; et
ii) en moyenne, l'écart quadratique entre sa valeur et θ_0 sera faible car il tend vers 0 comme $\frac{1}{n}$.
On dit que $\bar{X}^{(n)}$ est un *estimateur sans biais et convergent* de θ.

Les définitions précises de ces propriétés sont données plus loin, mais il nous faut d'abord définir quelques termes.

2.1.2 Définitions

Fonction de vraisemblance Soit $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ la famille de lois de la variable X_0 dont est tiré l'échantillon. La loi de l'échantillon, variable à n dimensions, est alors elle-même paramétrée par θ.
Dans le *cas discret*, ces lois sont caractérisées par les probabilités élémentaires

$$L(x, \theta) = L(x_1, x_2, \cdots, x_k, \cdots, x_n, \theta) = \prod_{k=1}^{n} P_\theta(x_k),\ \theta \in \Theta\ ,$$

et, dans le *cas continu*, par les densités

$$L(x, \theta) = L(x_1, x_2, \cdots, x_k, \cdots, x_n, \theta) = \prod_{k=1}^{n} p_\theta(x_k),\ \theta \in \Theta\,;$$

Dans les deux cas $L(x, \theta)$ est appelée la *vraisemblance de l'échantillon*.

Statistique On appelle *statistique* toute fonction $T = f(X_1, X_2, \cdots, X_k, \cdots, X_n) = f(X)$ de l'échantillon empirique.

Estimateur Un *estimateur* est simplement une statistique susceptible d'être utilisée pour estimer la valeur d'une caractéristique de la loi de X_0 en raison de ses propriétés spécifiques. La caractéristique à estimer coïncidera souvent avec le paramètre ou une fonction du paramètre de la famille de lois possibles de X_0.

2.1.3 Propriétés des estimateurs

Il parait souhaitable que l'estimateur d'un paramètre possède les propriétés suivantes :
- un estimateur T est *sans biais* lorsque sa moyenne (= son espérance mathématique) est égale à la valeur vraie (quelle qu'elle soit) du paramètre :

$$\forall \theta \in \Theta,\ E_\theta(T) = \theta,$$

("en moyenne, on ne se trompe pas").
- un estimateur (plus précisément une suite d'estimateurs) $T_n = f_n(X^{(n)})$ est *convergent* lorsque, quelle que soit la valeur vraie de θ, il entraîne une erreur quadratique moyenne (sur la valeur vraie du paramètre) tendant vers zéro lorsque la taille, n, de l'échantillon $X^{(n)}$ tend vers l'infini :

$$\forall \theta \in \Theta,\ lim_{n \to\ \infty} E_\theta[(T_n - \theta)^2] = 0,$$

("si l'on recevait autant d'information que l'on voulait, on ne se tromperait pas sur la valeur de θ").

La comparaison avec d'autres estimateurs utilise les concepts suivants :
- un estimateur $T_n = f_n(X^{(n)})$ est *asymptotiquement efficace* lorsqu'il n'existe pas d'autre estimateur T_n^* satisfaisant

$$limsup_{n\to\infty} \frac{E_\theta[(T_n-\theta)^2]}{E_\theta[(T_n^*-\theta)^2]} > 1,\ \forall\theta\in\Theta$$

("aucun autre estimateur ne donne une erreur quadratique plus faible dans les grands échantillons").
- un estimateur T est *sans biais de variance minimum* quand il est sans biais et que sa variance, qui vaut alors $Var(T) = E_\theta[(T-\theta)^2]$, est inférieure à celle de tout autre estimateur sans biais. C'est une propriété plus forte que la précédente, qui ne disait quelque chose que pour n grand.

Exhaustivité Pour que l'observation d'une variable aléatoire puisse nous apporter de de l'information concernant un paramètre θ, il est nécessaire que sa loi dépende de θ. C'est le cas pour un échantillon empirique X de vraisemblance $L(x,\theta)$. Supposons qu'au lieu du vecteur observé $x = (x_1,\cdots,x_k,\cdots,x_n)$ l'on se contente de retenir la valeur $t = f(x)$ prise par une certaine statistique $T = f(X)$; en général la loi de T dépend de θ et T apporte de l'information sur ce paramètre ; mais en apporte-t-elle autant que l'échantillon ? oui, si conditionnellement à la connaissance la valeur t de T, la loi de X ne dépend plus de θ ; or par définition des probabilités conditionnelles, c'est, dans le cas discret,

$$P_\theta(x/t) = \frac{L(x,\theta)}{P_{T,\theta}(t)} \text{ où } P_{T,\theta} \text{ est la loi de } T\,;$$

$P_\theta(x/t)$ est alors une fonction de x seul (car $t = f(x)$) et donc $L(x,\theta)$ est de la forme

$$L(x,\theta) = g(t,\theta)h(x)$$

On obtiendrait la même factorisation, mais portant sur des densités, dans le cas continu.

Par définition, l'estimateur T est dit *exhaustif* de l'information concernant le paramètre θ contenu dans l'échantillon lorsque la vraisemblance se factorise selon la forme ci-dessus.

Exemple Loi de Poisson $\mathcal{P}(\lambda)$ de paramètre λ inconnu
La vraisemblance est

$$L(k_1,\cdots,k_i,\cdots,k_n,\theta) = \prod_{i=1}^n \frac{\exp\{-\lambda\}.\lambda^{k_i}}{k_i!} = \exp\{-n\lambda\}\frac{\lambda\sum_{i=1}^n k_i}{\prod_{i=1}^n k_i!}$$

La variable $S = \sum_{i=1}^n X_i$, qui suit elle-même une loi $\mathcal{P}(n\lambda)$, de probabilité élémentaire $g(s,n\lambda) = \exp\{-n\lambda\}\frac{(n\lambda)^s}{s!}$ est un estimateur exhaustif de λ car

$$L(k_1,\cdots,k_i,\cdots,k_n,\theta) = g(s,n\lambda)\frac{s!}{n^s\prod_{i=1}^n k_i!}.$$

Estimateurs du maximum de vraisemblance L'estimateur du *maximum de vraisemblance* $T = f(X)$ est défini par :

$$x \longmapsto t = f(x) = argmax_{\theta \in \Theta} L(x, \theta).$$

C'est donc la valeur de θ pour laquelle la probabilité d'observer x était la plus grande. Sous des conditions adéquates (concavité et dérivabilité de L par rapport à θ), c'est la solution de l'équation

$$\frac{\partial L(x, \theta)}{\partial \theta} = 0.$$

Ceci se généralise au cas de paramètres multi-dimensionnels.

Exemple i) Lois normales d'écart-type connu, $\sigma = 1$. Un seul paramètre, l'espérance m. La vraisemblance (ici la densité de probabilité de l'échantillon, puisque l'on est dans le cas continu) est :

$$L(x, m) = \prod_{i=1}^{n} [\frac{1}{\sqrt{2\pi}} \exp \{-\frac{1}{2}(x_i - m)^2\}].$$

Remarquons que

$$\frac{\partial L(x, m)}{\partial m} = 0 \Longleftrightarrow \frac{\partial \ln L(x, m)}{\partial m} = 0,$$

où

$$\ln L(x, m) = -\frac{n}{2} \ln 2\pi - \frac{1}{2} \sum_{i=1}^{n} (x_i - m)^2;$$

donc

$$\frac{\partial \ln L(x, m)}{\partial m} = 0 \Longleftrightarrow \sum_{i=1}^{n} (x_i - m) = 0 \Longleftrightarrow m = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}.$$

L'estimateur cherché est donc $\overline{X}$.

ii) La famille de toutes les lois normales a deux paramètres : m et σ^2 (la variance). Le logarithme de la vraisemblance est :

$$\ln L(x, m, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2$$

les conditions du premier ordre pour un maximum de $\ln L(x, m, \sigma^2)$ sont que ses dérivées partielles doivent être nulles :

$$\frac{\partial L(x, m, \sigma^2)}{\partial m} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - m) = 0$$

$$\frac{\partial L(x,m,\sigma^2)}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - m)^2 = 0 \ ;$$

on en déduit :

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \ ; \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s^2.$$

Les estimateurs cherchés sont $\overline{X}$ pour m et $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ pour σ^2.

Les estimateurs du maximum de vraisemblance ont les propriétés suivantes (sous certaines conditions de régularité) : ils sont convergents, asymptotiquement efficaces et, de plus, sont des fonctions des estimateurs exhaustifs.

2.1.4 Intervalles de confiance

L'estimateur T d'un paramètre θ lui attribue, au vu des observations x, une valeur $\hat{\theta}$; il est important d'avoir une idée de la précision de cette estimation ; c'est ce à quoi servent les intervalles de confiance :
Un *intervalle de confiance* de niveau $1-\alpha$ (on dit aussi : à $100(1-\alpha)\%$) est un intervalle aléatoire (car ses bornes sont des fonctions de T), $]a(T), b(T)[$ ayant une probabilité $1-\alpha$ de contenir la valeur vraie du paramètre (quelle qu'elle soit), donc tel que :

$$\forall \theta \in \Theta,\ P_\theta(]a(T), b(T)[\ni \theta) = 1 - \alpha$$

(Noter que c'est θ qui est fixe et les bornes de l'intervalle qui sont aléatoires).

On construit généralement des intervalles de confiance de la façon suivante :

Pour chaque valeur de θ, on choisit des nombres $B(\theta)$ et $A(\theta)$ tels que

$$P_\theta(T \in]B(\theta), A(\theta)[) = 1 - \alpha.$$

On les prend le plus souvent de manière à avoir
$P_\theta(T \leq B(\theta)) = P_\theta(T \geq A(\theta)) = \frac{\alpha}{2}$.
On prend alors pour a et b les fonctions inverses de A et B, d'où :

$$]a(t), b(t)[\ni \theta \Leftrightarrow t \in]B(\theta), A(\theta)[$$

et donc,

$$\forall \theta,\ P_\theta(]a(T), b(T)[\ni \theta) = 1 - \alpha$$

(noter que pour chaque θ, l'événement $\{t :]a(t), b(t)[\ni \theta\}$ est différent).

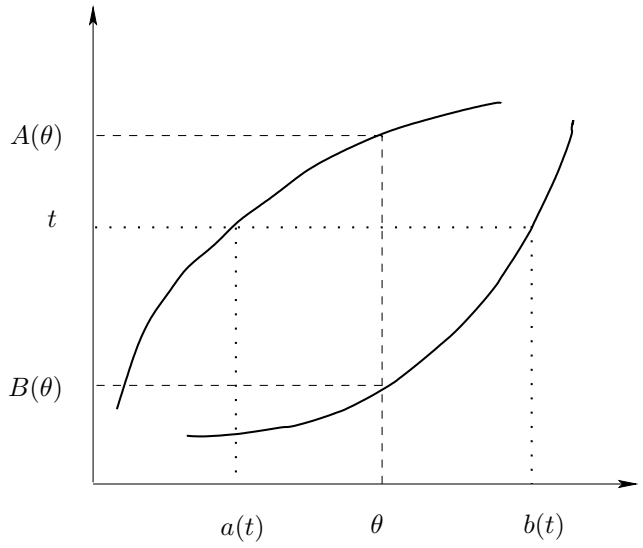

FIG. 1 – Intervalles de confiance

Exemple Lois normales d'écart-type connu, σ .
Le paramètre est l'espérance m ; son estimateur $\overline{X}$ suit une loi normale $\mathcal{N}(m, \frac{\sigma}{\sqrt{n}})$; $Y = \frac{\overline{X}-m}{\sigma/\sqrt{n}}$ suit la loi normale centrée réduite et il y a donc, quel que soit m, une probabilité $1-\alpha$ que

$$m - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \overline{X} < m + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

où $u_{\frac{\alpha}{2}}$ est défini par $P(Y < u_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$
d'où l'intervalle de confiance

$$\overline{X} - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < m < \overline{X} + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

2.2 Tests d'hypothèses en statistique paramétrique classique

L'ensemble Θ des valeurs du paramètre θ est partitionné en deux sous-ensembles Θ_0 et Θ_1.
On appelle *h*ypothèses les assertions $H_0 =$ "$\theta \in \Theta_0$" et $H_1 =$ "$\theta \in \Theta_1$".
Une hypothèse H_i est dite *s*imple lorsque Θ_i est un singleton ; elle est dite *m*ultiple sinon.

Un *t*est entre deux hypothèses H_0 et H_1 est une règle de décision, δ, basée sur les observations :
L'ensemble des décisions possibles est $\mathcal{D} = \{d_0, d_1\}$, avec $d_0 =$ "*accepter* H_0" et $d_1 =$ "*accepter* H_1" $=$ "*rejeter* H_0" .

Un test *déterministe* est donc une application $x \mapsto \delta(x)$ de $\mathbb{R}^n$ dans $\mathcal{D}$. Il est donc caractérisable par sa *région critique* $W = \{x \in \mathbb{R}^n : \delta(x) = d_1\}$; la région complémentaire $\mathbb{R}^n \backslash W$ étant la région *d'acceptation*. On notera T_W un tel test.

On est en fait amené à utiliser aussi des tests *aléatoires*, où $\delta(x)$ est un nombre de [0, 1], la probabilité d'accepter H_1 si x est observé. Il faut alors distinguer la région critique W, où H_1 est acceptée (avec probabilité 1), la région d'acceptation A où c'est H_0 qui est acceptée (avec probabilité 1) et la zone complémentaire de $W \cup A$, où le choix est aléatoire ; en général, cette dernière zone correspondra à la frontière ente W et A et la probabilité d'accepter H_1 y aura une valeur constante.

Tout test peut amener à accepter une hypothèse alors que c'est l'autre qui est vraie. On appelle *erreur de première espèce* le fait de rejeter l'hypothèse H_0 alors qu'elle est vraie et *erreur de deuxième espèce* le fait de rejeter l'autre hypothèse, H_1, alors que celle-ci est vraie.
Les coûts associés aux erreurs sont décrits par une fonction de perte w ; $w(d, \theta)$ est le coût de prendre la décision d quand la valeur vraie du paramètre est θ : w est donnée par le tableau suivant :

valeur vraie de θ	$d_0 =$ "*accepter* H_0"	$d_1 =$ "*accepter* H_1"
$\theta \in \Theta_0$	0	$w(d_1, \theta)$
$\theta \in \Theta_1$	$w(d_0, \theta)$	0

2.2.1 Tests entre hypothèses simples

$\Theta_0 = \{\theta_0\}$; $\Theta_1 = \{\theta_1\}$.

Etant donné un test entre hypothèses simples :
- on appelle *risque de première espèce* et note α, la probabilité de commettre l'erreur de première espèce ; on a donc dans le cas continu (formule analogue dans le cas discret) :

$\alpha = P(x \in W/\theta = \theta_0) = \int_W L(x, \theta)dx$, si c'est un test déterministe ;
$\alpha = \int_W L(x, \theta_0)dx + \int_{[W \cup A]^c} \delta(x)L(x, \theta_0)dx$, si c'est un test aléatoire ;

- et on appelle *risque de deuxième espèce* et note β, la probabilité de commettre l'erreur de deuxième espèce ; d'où, toujours dans le cas continu :

$\beta = P(x \in A/\theta = \theta_1) = \int_A L(x, \theta)dx$, si c'est un test déterministe ;
$\beta = \int_A L(x, \theta_1)dx + \int_{[W \cup A]^c} [1 - \delta(x)]L(x, \theta_1)dx$, si c'est un test aléatoire ;

Il est clair qu' α et β varient en sens inverse l'un de l'autre ; un test doit toujours réaliser un compromis entre les deux risques. Souvent, H_0 désigne une hypothèse privilégiée (par exemple, que la machine que l'on contrôle ne s'est

pas déréglée), vérifiée jusqu'à présent et que l'on n'aimerait pas abandonner à tort. On impose alors un *seuil* α_0 - valeur que α ne doit pas dépasser - et cherche un test minimisant β sous cette contrainte ; minimiser β, c'est maximiser $\eta = 1 - \beta$, que l'on appelle la *puissance* du test.

Un résultat célèbre, connu sous le nom de **lemme de** NEYMAN **et** PEARSON dit qu'*il existe toujours un test (aléatoire) le plus puissant de seuil donné* α_0 et que c'est un *test du rapport de vraisemblance*, c-à-d de la forme

$$\frac{L(x,\theta_0)}{L(x,\theta_1)} > k \Rightarrow x \in A \text{ (accepter } H_0); \frac{L(x,\theta_0)}{L(x,\theta_1)} < k \Rightarrow x \in W \text{ (rejeter } H_0);$$
$$\frac{L(x,\theta_0)}{L(x,\theta_1)} = k \Rightarrow \delta(x) = \rho \text{ (accepter } H_0 \text{ avec probabilité } 1-\rho \text{ et } H_1 \text{ avec probabilité } \rho),$$

les nombres k et ρ étant déterminés de façon unique par la relation $\alpha = \alpha_0$.

Exemple Test de moyenne d'une loi normale d'écart-type connu, σ .
$H_0 : X_0$ suit la loi $\mathcal{N}(m_0, \sigma)$; $H_1 : X_0$ suit la loi $\mathcal{N}(m_1, \sigma)$.

Le rapport de vraisemblance s'exprime en fonction de la statistique exhaustive $\bar{x}$:

$$\frac{L(x,\theta_0)}{L(x,\theta_1)} = exp(-\frac{n}{2\sigma^2}[(\bar{x} - m_0)^2 - (\bar{x} - m_1)^2])$$

d'où, $\frac{L(x,\theta_0)}{L(x,\theta_1)} > k$ est équivalent à $(\bar{x} - m_0)^2 - (\bar{x} - m_1)^2 = [m_1 - m_0][2\bar{x} - m_1 - m_0] < k'$ et donc à $\bar{x} < k''$ si $m_1 > m_0$ et $\bar{x} > k''$ si $m_1 < m_0$
(pour les valeurs adéquates de k' et k''.)
On rejette donc H_0 lorsque $\bar{x}$ est grand dans le cas où $m_1 > m_0$ et lorsqu'il est trop petit dans le cas où $m_1 < m_0$.

2.2.2 Tests entre hypothèses multiples

Il n'y a pas de résultat général simple ; voyons cependant le cas où H_0 est simple et H_1 est multiple :

$$H_0 : \theta = \theta_0 ; H_1 : \theta > \theta_1$$

Pour chaque test, la probabilité d'erreur de seconde espèce et la puissance varient avec $\theta \in \Theta_1$; un test donné peut donc être plus puissant qu'un autre test de même seuil pour une valeur θ_1' et moins puissant pour une autre valeur θ_1''.
LEHMANN a montré que si le rapport de vraisemblance est une fonction monotone d'une statistique donnée, alors il existe un test *uniformément le plus puissant* (UPP), c-à-d tel que, pour tout $\theta - 1 \in \Theta_1$, sa puissance $\eta(\theta_1)$ est supérieure ou égale à celle de tous les autres tests de même seuil.

2.3 Les tests d'ajustement

Les *tests d'ajustement* ont pour issue l'acceptation ou le rejet de l'hypothèse que l'échantillon observé est tiré d'une certaine loi. L'hypothèse alternative ne précise pas de quelle autre loi il aurait pu être tiré. Un test d'ajustement est donc un exemple de *test non-paramétrique*.

Très souvent, la loi testée a été sélectionnée, lors d'une étape précédente, au sein d'une famille de lois dépendant de paramètres par estimation de de ces paramètres.

On rencontre fréquemment le cas où l'échantillon provient de n tirages aléatoires (avec remise) dans une population qui se répartit en k *classes (= catégories)* (exemple : les six classes associées au lancer d'un dé).

2.3.1 loi multinomiale et loi du χ^2

Supposons d'abord que l'on connaisse la proportion d'individus de la population appartenant à chaque classe et donc la probabilité p_l, à chaque tirage, que l'individu tiré appartienne à la classe l ($l = 1, .., k$). Notons N_l la variable (aléatoire) qui a pour valeur n_l, nombre d'individus tirés qui appartiennent à la classe l.
La loi du k-uple $(N_1, .., N_l, .., N_k)$ est une *loi multinomiale de paramètres* $(p_1, .., p_l, .., p_k)$, dont les probabilités élémentaires sont données par

$$p(n_1, .., n_l, .., n_k) = P(N_1 = n_1, .., N_l = n_l, .., N_k = n_k) =$$

$$\frac{n!}{n_1! \times .. \times n_l! \times .. \times n_k!}\ p_1^{n_1} \times .. \times p_l^{n_l} \times .. \times p_k^{n_k}$$

Cette loi généralise la loi binomiale qui correspond au cas $k = 2$.

On démontre que la suite de variables $D^2_{(n)} = \sum_{l=1}^{k} \frac{(N_l - n.p_l)^2}{n.p_l}$ tend en loi, lorsque $n \to \infty$, vers une loi du χ^2_{k-1} (loi du "chi deux" à $(k-1)$ degrés de liberté).
La loi du χ^2_r est la loi de la somme des carrés de r variables indépendantes et de même loi, la loi normale centrée réduite $\mathcal{N}(0, 1)$; son espérance vaut r et sa variance $2r$. Sa densité en $x > 0$ vaut

$$g(y) = \frac{1}{2^{p/2}\Gamma(p/2)} \exp\left(\frac{-y^2}{2}\right)(y^2)^{p/2-1}.$$

2.3.2 Test d'ajustement du χ^2

Ce test est fondé sur l'idée suivante : si le k-uple observé $(n_1, .., n_l, .., n_k)$ est bien tiré selon la loi binomiale de paramètres $(p_1, .., p_l, .., p_k)$, alors la

valeur résultante de d^2, réalisation de $D^2_{(n)}$ avait au départ peu de chance d'avoir une valeur élevée ; si donc la valeur observée est élevée, l'hypothèse faite est peu plausible ; plus précisément, on utilise le résultat asymptotique précédent ; étant donné un seuil α, par exemple $\alpha = 0.1$, on lit dans une table la valeur d^2_α telle que $Pr(\chi^2_{k-1} > d^2_\alpha) = \alpha$; si $d^2 < d^2_\alpha$, on accepte l'hypothèse que la loi de $(N_1, .., N_l, .., N_k)$ est bien celle que l'on a supposée ; on rejette cette hypothèse dans le cas contraire. Le choix de la valeur de α reste un peu arbitraire ; on prend souvent $\alpha = 0.05$; on peut le prendre d'autant plus petit que la taille de l'échantillon est grande et que le coût d'erreur de première espèce (rejeter à tort l'hypothèse) est plus élevé.

2.3.3 Exemple des dés

Reprenons l'exemple des deux dés lancés 360 fois.

l	2	3	4	5	6	7	8	9	10	11	12
n_l	7	25	24	47	58	52	49	34	37	16	11
np_l	10	20	30	40	50	60	50	40	30	20	10
$(n_l - np_l)^2$	9	25	36	49	64	64	1	36	49	16	1
$\frac{(n_l - np_l)^2}{np_l}$	0.9	1.25	1.2	1.225	1.28	1.067	0.02	0.9	1.633	0.8	0.1

D'où $d^2 = 10.375$. Pour $\alpha = 0.60$ on trouve déjà $d^2_\alpha = 10.473 > d^2$. L'hypothèse devrait donc être acceptée pour tout choix de α inférieur ou égal à 0.60 . En fait, sous cette hypothèse, il y a plus de 40 chances sur 100 d'obtenir un écart d^2 au moins aussi grand que celui observé ; l'écart observé ne peut suffire à soupçonner que l'hypothèse est fausse.

2.4 Test d'indépendance du χ^2

Nous avons introduit comme indice de dépendance d'un tableau de contingence la quantité

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - \frac{n_{i.} n_{.j}}{n})^2}{\frac{n_{i.} n_{.j}}{n}}$$

Cette caractéristique suit approximativement une loi de χ^2_k, c-à-d comme nous l'avons vu, la loi de la somme des carrés de k variables indépendantes et de loi $\mathcal{N}(0, 1)$. Le nombre k, *nombre de degrés de liberté*, correspond au nombre de paramètres indépendants, c-à-d au nombre d'éléments du tableau auxquels on peut attribuer des valeurs arbitraires une fois les marges du tableau fixées. Dans un tableau à r lignes et s colonnes, il y a ainsi $k = (r - 1)(s - 1)$ degrés de liberté.
Le test est alors le suivant : étant donné un seuil α, par exemple $\alpha = 0.1$, on lit dans une table la valeur x^2_α telle que $Pr(\chi^2_k > x^2_\alpha) = \alpha$; si $\chi^2 > x^2_\alpha$, on rejette l'hypothèse d'indépendance des deux variables du tableau (= on

considère les variables comme liées) ; on accepte cette hypothèse dans le cas contraire.

Exemple du médicament Reprenons l'exemple médical donné précédemment. Dans le cas du tableau 3×4 (2 médicaments et le placebo), il y a $k = 6$ degrés de liberté. Pour $\alpha = 0.05$, $x_6^2 = 12.6927$; comme on trouve $\chi^2 = 17.50$, on rejette l'hypothèse d'indépendance.
Dans le cas du tableau 2×4 (2 médicaments), il y a $k = 3$ degrés de liberté. Pour $\alpha = 0.05$, $x_3^2 = 7.815$; comme on trouve $\chi^2 = 1.48$, on accepte l'hypothèse d'indépendance : les deux médicaments ont en fait des effets très voisins.

2.5 Tests de corrélation des rangs de Spearman et de Kendall

Dans le cas de deux variables ordinales nous avons introduit deux indices de corrélation, le coefficient de Spearman :

$$r_S = 1- \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n-1)^2},$$

où $d_i = r_i - t_i$ est la différence des rangs d'un même objet i,
et le coefficient de Kendall

$$\tau = \frac{2S}{n(n-1)},$$

où S est la différence entre le nombre de classements concordants et de classements discordants dans l'ensemble de tous les couples d'objets.
Les distributions des variables aléatoires R_S à valeurs r_S et T à valeurs τ ont été tabulées et l'on considérera les classements comme indépendants : dans le test de Spearman, lorsque $\mid r_S \mid < k$; dans le test de Kendall, lorsque $\mid \tau \mid < t$; les classements seront considérés comme corrélés et concordants pour, respectivement, $R_S > k$ et $T > k$; enfin comme corrélés et discordants pour, respectivement, $R_S < k$ et $T < k$.

Exemple des oenologues Reprenons l'exemple des deux oenologues, où pour les classement des douze vins de table nous avons trouvé un coefficient de Spearman $r_S = 0.60$ et un coefficient de Kendall $\tau = 0.36$. Avec $\alpha = 0.05$, on trouve que $Pr(\mid R_S \mid > 0.648) = \alpha$ et que $Pr(\mid T \mid > 0.49) = \alpha$. Puisque $r_S = 0.60 < 0.648$ et $\tau = 0.36 < 0.49$ les deux tests concluent à l'indépendance.

3 La statistique bayésienne

La spécificité du modèle statistique bayésien (par rapport au modèle statistique "classique") tient au fait que tout élément d'incertitude y est évalué sous une forme probabiliste.

Exemple introductif Une personne de votre connaissance vous propose de jouer à *Pile* ou *Face* en pariant sur *Pile* : gain de 10 euros si *Pile*, perte de 10 euros si *Face*, avec une pièce de monnaie qu'elle sort de sa poche, où vous savez qu'elle n'avait que deux pièces, que vous avez déjà eu la possibilité de manipuler ; vous avez pu les jeter mille fois chacune, si bien que, vous fondant sur les fréquences de *Pile* et de *Face* observées, vous estimez qu'avec la première pièce la probabilité de *Pile* est $P_1(Pile) = 0.6$ alors qu'avec la deuxième pièce elle est $P_2(Pile) = 0.45$ (les deux pièces sont toutes deux tordues, mais dans des sens opposés). Quelle est votre espérance de gain si vous acceptez de jouer ?
Si vous considérez qu'il y a une probabilité π_1 pour que la pièce utilisée soit la première pièce et donc $\pi_2 = 1 - \pi_1$ pour que ce soit la deuxième, votre espérance de gain est :
$[\pi_1 \times 0.6 + \pi_2 \times 0.45] \times 10 + [\pi_1 \times 0.4 + \pi_2 \times 0.55] \times (-10) = 2\pi_1 - \pi_2 = 3\pi_1 - 1$
et est donc positive si et seulement si $\pi_1 > \frac{1}{3}$. Ceci peut être une incitation à accepter de jouer dans ce cas et refuser de jouer sinon.
Mais d'où viennent π_1 et π_2 ? Il n'y a pas de données fréquentistes permettant de les estimer. Peut-être pensez-vous que cette personne va choisir une des pièces au hasard : $\pi_1 = \pi_2 = \frac{1}{2}$; ou peut-être que c'est un ami, qui souhaite vous faire gagner et va choisir plutôt la première pièce ; ou au contraire un ennemi qui choisira plutôt la seconde ; est-ce suffisant pour entrainer que vous devez mettre des probabilités bien définies satisfaisant $\pi_1 > \pi_2$ dans le premier cas, $\pi_1 < \pi_2$ dans le second ? La théorie des probabilités subjectives affirme que oui.

3.1 La théorie des probabilités subjectives

Cette théorie, due à De Finetti, part de l'idée que *si vous pensez qu'un événement A a une probabilité π* , alors :
i) vous devez accepter de parier sur l'événement A si le gain net (= gain brut - mise) possible G et la mise M vous offrent une espérance de gain positive ou nulle :
$\pi \times G + (1 - \pi) \times (-M) \geq 0 \Longleftrightarrow G \geq \frac{1-\pi}{\pi} \times M \Longleftrightarrow \frac{G}{M} \geq \frac{1-\pi}{\pi}$; et
ii) vous devez accepter de recevoir (en vous plaçant en position de bookmaker) tout pari sur l'événement A pour lequel votre perte nette possible G (si le parieur gagne) et votre recette certaine, la mise M, vous offrent une espérance de gain positive ou nulle :
$\pi \times (-G) + (1 - \pi) \times M \geq 0 \Longleftrightarrow G \leq \frac{1-\pi}{\pi} \times M \Longleftrightarrow \frac{G}{M} \leq \frac{1-\pi}{\pi}$.
La seule valeur du rapport $\frac{G}{M}$ pour laquelle vous acceptez aussi bien de parier que de prendre le pari vérifie donc $\frac{G}{M} = \frac{1-\pi}{\pi}$ et permet de retrouver la valeur de la probabilité π que vous accordez à l'événement A, qui est donnée par $\pi = \frac{M}{M+G}$.
De Finetti fait alors l'hypothèse que pour tout événement A il existe une valeur-limite du rapport $\frac{G}{M}$ pour laquelle vous êtes indifférent entre parier et

ne pas parier sur A ; par définition, $\pi(A) = \frac{M}{M+G}$ est la *probabilité subjective* que vous accordez à A.
Il est clair que $0 \leq \pi(A) \leq 1$. Ce qui est moins évident, c'est que les probabilités subjectives soient additives.
De Finetti présente l'argument suivant en faveur de cette propriété :
Supposons qu'il existe des événements incompatibles A et B pour lesquels vos probabilités subjectives satisfont : $\pi(A)+\pi(B) > \pi(A \cup B)$; par exemple : $\pi(A) = \pi(B) = \frac{1}{4}$ *et* $\pi(A \cup B) = \frac{4}{10}$. Quelqu'un peut alors vous proposer de, simultanément,
- prendre un pari où il mise 18 (euros) pour un gain net de 22 si $(A \cup B)$ se réalise ;
- parier vous-même avec lui sur A en misant 10 pour un gain net de 31 si A se réalise et sur B avec même mise, 10, et même gain, 31, cette fois si c'est B qui se réalise.

Vous accepterez ces trois propositions puisque $\frac{22}{18} < \frac{3}{2} = \frac{1-\frac{4}{10}}{\frac{4}{10}}$ et $\frac{31}{10} > 3 = \frac{1-\frac{1}{4}}{\frac{1}{4}}$; or, vous allez subir une perte de 1 euro quoi qu'il arrive, comme le montre le tableau ci-dessous :

	A	B	$(A \cup B)^c$
pari accepté	-22	-22	+18
pari sur A	31	-10	-10
pari sur B	-10	31	-10
gain algébrique	-1	-1	-1

Il serait aussi facile d'exhiber trois propositions amenant également à une perte certaine dans le cas où il y aurait eu sur-additivité, par exemple : $\pi(A) = \pi(B) = \frac{1}{4}$ *et* $\pi(A \cup B) = \frac{6}{10}$ (trouvez trois paris adéquats).

D'où la conclusion de De Finetti : *Un décideur qui ne se comporte pas en toute situation de choix comme si il attribuait des probabilités (subjectives) à tous les événements n'est pas rationnel, car il est alors possible à un manipulateur de le placer dans des situations de paris (multiples) où il perdra de l'argent à coup sûr.*

3.2 La formalisation bayésienne

Dans le modèle bayésien (du nom du probabiliste anglais Bayes) les événements sont des parties d'un ensemble produit $\mathcal{X} \times \Theta$, où :
- $\mathcal{X}$ est l'espace des observations, x ; x est le plus souvent un échantillon ;
- Θ est l'espace des paramètres, θ, caractéristiques concrètes ou abstraites intervenant dans le problème et sur lesquelles les observations apportent de l'information.

La famille des événements est dotée d'une loi de probabilité, Π ; les couples (x, θ) sont donc les réalisations d'un couple de variables aléatoires $(X, \widetilde{\theta})$.

- **dans le cas discret**, la loi Π est déterminée par les probabilités élémentaires $\pi(x,\theta) = \Pi(X = x, \widetilde{\theta} = \theta)$;
on peut en dériver par sommation les lois marginales de X et $\widetilde{\theta}$:
$\pi(x) = \Pi(X = x) = \sum_{\theta\in\Theta} \pi(x,\theta)$ $\quad$ $\pi(\theta) = \Pi(\widetilde{\theta} = \theta)) = \sum_{x\in\mathcal{X}} \pi(x,\theta)$
et les lois conditionnelles de X si θ et $\widetilde{\theta}$ si x :
$\pi(x/\theta) = \Pi(X = x/\widetilde{\theta} = \theta) = \frac{\pi(x,\theta)}{\pi(\theta)}$ $\quad$ $\pi(\theta/x) = \Pi(\widetilde{\theta} = \theta/X = x) = \frac{\pi(x,\theta)}{\pi(x)}$;
en fait, les données primitives pourront aussi bien être constituées de la loi marginale de $\widetilde{\theta}$ et des lois conditionnelles de X si θ, pour tout θ, comme dans l'exemple donné ci-après.
- **dans le cas continu**, la loi Π est déterminée par la densité (de probabilité) jointe $\pi(x,\theta)$, dont on peut dériver, par intégration, les densités, notées également $\pi(x)$ et $\pi(\theta)$ des lois marginales de X et $\widetilde{\theta}$:

$$\pi(x) = \int_\Theta \pi(x,\theta)d\theta \quad \pi(\theta) = \int_\mathcal{X} \pi(x,\theta)dx$$

ainsi que celles des lois conditionnelles de X si θ et $\widetilde{\theta}$ si x :
$\pi(x/\theta) = \frac{\pi(x,\theta)}{\pi(\theta)}$ $\quad$ $\pi(\theta/x) = \frac{\pi(x,\theta)}{\pi(x)}$.
La loi conditionnelle de X si θ représente la loi qu'aurait l'échantillon si θ était la valeur vraie du paramètre ; c'est donc la *(fonction de) vraisemblance* de l'échantillon, que nous avons déjà rencontrée en statistique classique, et pour laquelle on utilise le plus souvent l'une des deux notations suivantes (dans le cas continu, où c'est une densité, comme dans le cas discret, où c'est une probabilité élémentaire) :

$$L_\theta(x) = L(x,\theta) = \pi(x/\theta)$$

["L" est l'initiale de "likelihood"=vraisemblance en anglais]
Nous supposerons toujours que l'on observe un *échantillon indépendant identiquement distribué* en abrégé *échantillon i.i.d.*, ce qui signifie que l'observation $x = (x_1, .., x_i, .., x_n)$ est la réalisation d'une variable n-dimensionnelle $X = (X_1, .., X_i, .., X_n)$ dont les composantes sont mutuellement indépendantes à valeur donnée du paramètre θ :
$L_\theta(x) = \prod_{i=1}^n Pr(X_i = x_i/\theta)$ (cas discret), $L_\theta(x) = \prod_{i=1}^n p(x_i/\theta)$ (L_θ, p densités, cas continu).

La loi marginale de $\widetilde{\theta}$ décrit l'idée que l'on se fait de cette variable avant observation : on l'appelle loi *a priori* du paramètre ; en revanche, la loi conditionnelle de $\widetilde{\theta}$ si x exprime ce que l'on pense de cette même variable après avoir observé x : on l'appelle loi *a posteriori* du paramètre.
La relation entre les lois a priori et a posteriori du paramètre est fournie directement par la formule de Bayes dans le cas discret :

$$\pi(\theta/x) = \frac{L_\theta(x)\times\pi(\theta)}{\pi(x)} = \frac{L_\theta(x)\times\pi(\theta)}{\sum_{\theta\in\Theta} L_\theta(x)\times\pi(\theta)} ;$$

dans le cas continu, on a la formule analogue pour les densités :

$$\pi(\theta/x) = \frac{L_\theta(x)\times\pi(\theta)}{\pi(x)} = \frac{L_\theta(x)\times\pi(\theta)}{\int_\Theta L_\theta(x)\times\pi(\theta)\ d\theta}\ .$$

exemple introductif (suite) L'ensemble des paramètres n'a que deux éléments : $\Theta = \{\theta_1, \theta_2\}$: $\theta_1 =$ "*biais pour Pile*" ; $\theta_2 =$ "*biais pour Face*". On prend pour probabilités a priori : $\pi(\theta_1) = \frac{2}{3};\ \pi(\theta_2) = \frac{1}{3}$.
Si l'on observe les résultats de 5 lancers successifs de la pièce, l'espace des observations est $\mathcal{X} = \{Pile,\ Face\}^5$; supposons que ces résultats sont indépendants et qu'à chaque lancer :
$Pr_{\theta_1}(Pile) = 0.6$, d'où $Pr_{\theta_1}(Face) = 0.4$ et
$Pr_{\theta_2}(Pile) = 0.45$, d'où $Pr_{\theta_1}(Face) = 0.55$.
Si l'on observe, par exemple, $x = (Face, Face, Face, Pile, Face)$ les vraisemblances seront : $L_{\theta_1}(x) = [0.6] \times [0,4]^4$ et $L_{\theta_2}(x) = [0.45] \times [0.55]^4$.
D'où la loi a posteriori :

$$\pi(\theta_1/x) = \frac{L_{\theta_1}(x)\times\pi(\theta_1)}{\pi(x)} = \frac{[0.6]\times[0.4]^4\times\frac{2}{3}}{\pi(x)} = \frac{10\,240}{10^6\times\pi(x)},$$

$$\pi(\theta_2/x) = \frac{L_{\theta_2}(x)\times\pi(\theta_2)}{\pi(x)}\ \frac{[0.45]\times[0.55]^4\times\frac{1}{3}}{\pi(x)} = \frac{13\,725}{10^6\times\pi(x)},$$

avec $\pi(x) = L_{\theta_1}(x) \times \pi(\theta_1) + L_{\theta_2}(x) \times \pi(\theta_2) = \frac{23\,965}{10^6}$,
ce qui donne : $\pi(\theta_1/x) = 0.428\ ;\ \pi(\theta_2/x) = 0.572$.
Les observations ont fait basculer les croyances initiales concernant la valeur du paramètre en faveur de θ_2.

3.3 Le modèle décisionnel bayésien

Le modèle probabiliste précédent n'est qu'une partie d'un modèle décisionnel, où sont pris en compte les coûts pouvant résulter des décisions choisies au vu de l'information. Le concept de base est celui de *fonction de coût* : étant donné un ensemble de décision possibles, $\mathcal{D}$, la fonction de coût est une application $w : \mathcal{D} \times \Theta \mapsto \mathbb{R}$, où $w(d, \theta)$ est le coût résultant de la décision d lorsque la valeur du paramètre est θ.

Exemples 1) *Problème de classification* :
$\mathcal{D} = \{d_1, .., d_k, .., d_n\}$ et $\Theta = \{\theta_1, .., \theta_j, .., \theta_n\}$ sont associés à un ensemble de n classes ; d_k signifie "l'individu observé est rangé dans la classe k" alors que le paramètre vaut θ_j lorsque la vraie classe de l'individu est la classe j ; une fonction de coût
$w(d_k, \theta_j) = 1,\ si\ k \neq j,\ \ = 0\ si\ k = j$, correspond à une pénalité constante pour toute erreur de classification.
2) *Diagnostic médical* : c'est aussi un problème de classification, mais les

coûts d'erreur de diagnostic peuvent être très différents ; $\Theta = \{\theta_1, \theta_2\}$; $\theta_1 =$ *maladie grave* ; $\theta_2 =$ *maladie bénigne* ; on aurait alors :
$w(d_1, \theta_1) = w(d_2, \theta_2) = 0$; $w(d_2, \theta_1) >> w(d_1, \theta_2) > 0$,
car mal soigner une maladie grave parce qu'on l'a fait une erreur de diagnostic coûte beaucoup plus, médicalement et humainement, que mal soigner une maladie bénigne parce que l'on a fait l'erreur inverse.
3) *Estimation* : $\mathcal{D} = \Theta = \mathbb{R}$. Ici, se tromper sur la valeur du paramètre, c-à-d décider que c'est d alors que la valeur vraie est θ, entraine un coût croissant avec l'importance de l'erreur : $w(d, \theta) =\mid d - \theta \mid$ (coût égal à l'écart absolu) ; $w(d, \theta) = (d - \theta)^2$ (coût égal à l'écart quadratique).

Critère de décision et fonctions de risque On cherche à minimiser la perte moyenne, c-à-d l'espérance mathématique de la perte.
En l'absence d'observation, il faut choisir une décision d dans $\mathcal{D}$ qui minimise, dans le *cas discret* :

$$W(d) = \sum_{\theta \in \Theta} \pi(\theta) w(d, \theta).$$

$W(d)$ est appelé le *risque a priori*.
En cas d'observation, la décision prise va pouvoir dépendre de l'observation, réalisation x de la variable X ; on doit donc choisir une règle de décision $\delta : \mathcal{X} \mapsto \mathcal{D}$; si x est observé, la décision prise est $d = \delta(x)$.
La perte entrainée par la règle δ sera $w(\delta(x), \theta)$ lorsque x sera observé et que θ sera la valeur du paramètre, ce qui, *dans le cas discret*, arrivera avec la probabilité $\pi(x, \theta)$; d'où une espérance de perte

$$r(\delta) = \sum_{x \in \mathcal{X}} \sum_{\theta \in \Theta} \pi(x, \theta).w(\delta(x), \theta)$$

$r(\delta)$ est appelée le *risque bayésien*. L'optimisation du risque bayésien, c-à-d la détermination de la règle de décision optimale, est facilitée par la remarque que $r(\delta)$ peut encore écrire

$$r(\delta) = \sum_{x \in \mathcal{X}} \sum_{\theta \in \Theta} \pi(x).\pi(\theta/x).w(\delta(x), \theta) = \sum_{x \in \mathcal{X}} \pi(x) \sum_{\theta \in \Theta} \pi(\theta/x).w(\delta(x), \theta)$$

La règle de décision optimale s'obtient donc en minimisant par rapport à d, pour chaque x de $\mathcal{X}$, le *risque a posteriori si x*

$$W(d/x) = \sum_{\theta \in \Theta} \pi(\theta/x).w(d, \theta),$$

et en prenant $\delta(x) = d^*$, où d^* est la meilleure décision trouvée.
Mieux, en pratique, on n'a besoin de connaitre la règle de décision optimale que lorsque certaines options doivent être prises avant observation (choix

de l'expérience, de la taille de l'échantillon, etc..) ; dans les autre cas, on peut attendre l'observation et se contenter d'optimiser le risque a posteriori $W(d/x_0)$ pour l'observation x_0 recueillie.
En statistique classique, où il n'y a pas de probabilités sur Θ, la fonction suivante, R, simplement appelée *(fonction de) risque*, joue un rôle important :

$$R_\theta(\delta) = \sum_{x\in\mathcal{X}} \pi(x/\theta).w(\delta(x),\theta) = \sum_{x\in\mathcal{X}} L_\theta(x).w(\delta(x),\theta).$$

Dans le cas continu, les expressions analogues font intervenir des densités de probabilités au lieu de probabilités élémentaires.
risque a priori : $W(d) = \int_\Theta w(d,\theta).\pi(\theta)\ d\theta$.
risque a posteriori si x : $W(d/x) = \int_\Theta w(d,\theta).\pi(\theta/x)\ d\theta$
risque bayésien : $r(\delta) = \int_\mathcal{X}\int_\Theta w(\delta(x),\theta).\pi(x,\theta)\ dx\ d\theta$
risque : $R_\theta(\delta) = \int_\mathcal{X} w(\delta(x),\theta).L_\theta(x)\ dx$.
On rencontre aussi des cas mixtes où l'une des deux variables X, $\widetilde{\theta}$ est discrète et l'autre continue ; l'adaptation des expressions des diverses fonctions de risque est évidente.

3.4 L'exhaustivité en statistique bayésienne

3.4.1 Statistique

Une *statistique*, T, est une fonction de l'échantillon X ; elle prend une valeur $t = f(x)$ lorsque l'échantillon a la valeur x, ce qu'on note $T = f(X)$; T est donc elle-même aléatoire, sa loi de probabilité dérivant de celle de X par $Pr(T = t) = Pr(X = f^{-1}(t))$ (cas discret) ou $Pr(T \in I) = Pr(X = f^{-1}(I))$ (cas continu).
Exemples de statistiques : la *moyenne (empirique) de l'échantillon*, $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$; la *variance (empirique)de l'échantillon* $S^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$.

3.4.2 Statistique exhaustive

Une statistique $T = f(X)$ est *exhaustive* lorsque la loi a posteriori du paramètre θ ne dépend de la valeur de x que par l'intermédiaire de t :

$$f(x) = f(x') \Longrightarrow \pi(\theta/x) = \pi(\theta/x'),$$

ce qui signifie que toute l'information que l'observation x peut apporter sur le paramètre est contenue dans son résumé t.
Le résultat suivant, dit *théorème de factorisation*, permet de caractériser l'exhaustivité :
$T = f(X)$ est exhaustive si et seulement si la vraisemblance se factorise sous la forme : $L_\theta(x) = g(f(x),\theta).h(x)$. On retrouve donc la propriété qui sert de définition de l'exhaustivité en statistique classique

3.4.3 Familles conjuguées de distributions

Lorsqu'il existe une statistique exhaustive T de dimension indépendante de la taille n de l'échantillon (par exemple, uni-dimensionnelle comme $\overline{X}$ ou bi-dimensionnelle comme $(\overline{X}, S^2)$), on a la propriété suivante :
A la famille des lois de probabilités $p_\theta(t), \theta \in \Theta$ de T [donc à celles des $L_\theta(x),\ \theta \in \Theta$ de X], on peut associer une famille de lois dite *conjuguée* de la première, telle que si la loi a priori $\pi(\theta)$ appartient à cette famille, alors toute loi a posteriori $\pi(\theta/x)$ lui appartient également.
Ceci est particulièrement intéressant lorsque cette famille conjuguée, tout en ne dépendant elle-même que d'un ou deux paramètres, contient une approximation acceptable de la véritable loi a priori.
Exemples de familles conjuguées :

X tiré d'une loi :	lois conjuguées :
de Poisson	*Gamma*
normale	*normales*
uniforme	*de Pareto*

La formule de Bayes permet de vérifier si deux familles sont effectivement conjuguées et de calculer la valeur des paramètres de $\pi(\theta/x)$ connaissant celle des paramètres de $\pi(\theta)$.

3.5 Tests d'hypothèses en statistique bayésienne

3.5.1 Tests d'hypothèses

Comme en statistique classique, un *test* entre deux hypothèses est un problème de décision statistique, avec $\mathcal{D} = \{d_0, d_1\}$, où $d_0 =$ "*accepter* H_0" et $d_1 =$ "*accepter* H_1", associé à une fonction de perte du type suivant :

valeur vraie de θ	$d_0 =$ "*accepter* H_0"	$d_1 =$ "*accepter* H_1"
$\theta \in \Theta_0$	0	$w(d_1, \theta)$
$\theta \in \Theta_1$	$w(d_0, \theta)$	0

Contrairement au cadre classique, *il suffit, en statistique bayésienne, de considérer des tests déterministes.*

Remarquons qu'un test d'hypothèses peut être vu comme un problème de classification où il n'y aurait que deux classes.

3.5.2 Tests entre hypothèses simples

$\Theta_0 = \{\theta_0\}$; $\Theta_1 = \{\theta_1\}$.
La meilleure décision a posteriori, après observation de x, est celle qui minimise le risque a posteriori

$$W(d/x) = \pi(\theta_0/x).w(d, \theta_0) + \pi(\theta_1/x).w(d, \theta_1).$$

D'où,
d_0 est préférable ou équivalent à $d_1 \Longleftrightarrow \pi(\theta_1/x).w(d_0,\theta_1) \leq \pi(\theta_0/x).w(d_1,\theta_0)$.
Par la formule de Bayes, il vient :
$\pi(\theta_0/x) = \frac{L_{\theta_0}(x).\pi(\theta_0)}{\pi(x)}$, $\pi(\theta_1/x) = \frac{L_{\theta_1}(x).\pi(\theta_1)}{\pi(x)}$;

d'où, d_0 est préférable ou équivalent à $d_1 \Longleftrightarrow \frac{L_{\theta_0}(x)}{L_{\theta_1}(x)} \geq \frac{\pi(\theta_1).w(d_0,\theta_1)}{\pi(\theta_0).w(d_1,\theta_0)}$.

L'hypothèse H_0 est d'autant plus facilement acceptée que le second membre de l'inégalité est plus petit, donc que sa probabilité a priori est plus élevée et que le coût de l'accepter à tort est plus faible (relativement au coût de l'autre type d'erreur).

3.5.3 Tests entre hypothèses multiples

On procède encore à la comparaison des risques a posteriori ; il n'y a pas de résultat général simple.

3.6 Estimation ponctuelle en statistique bayésienne

3.6.1 Estimation ponctuelle

Les problèmes d'estimation ponctuelle sont ceux où $\mathcal{D} = \Theta$, c-à-d où l'on cherche à estimer la valeur du paramètre.
Comme nous l'avons vu, dans le cas uni-dimensionnel, $\Theta \subset \mathbb{R}$, la valeur de la perte est souvent prise égale :
- soit à celle de l'écart absolu entre valeur estimée et valeur vraie du paramètre : $w(d,\theta) = \mid d - \theta \mid$,
- soit à celle du carré de cet écart : $w(d,\theta) = (d-\theta)^2$.

Dans le cas pluri-dimensionnel, $\Theta \subset \mathbb{R}^n$, on prend généralement
$w(d,\theta) = {}^{\tau}(d-\theta)A(d-\theta)$,
où A est une matrice (n,n) définie positive.

3.6.2 Cas de la perte quadratique et d'un paramètre réel

C'est une propriété générale que la variance est l'écart quadratique moyen minimum ; d'où, *dans le cas discret* :
i) en l'absence d'information, le risque a priori

$$W(d) = \sum_{\theta \in \Theta} \pi(\theta).(d-\theta)^2 = E((d-\widetilde{\theta})^2)$$

est minimum pour $d^* = E(\widetilde{\theta})$, espérance a priori de $\widetilde{\theta}$ et vaut $Var\widetilde{\theta}$;
ii) après observation de w, le risque a posteriori

$$W(d/x) = \sum_{\theta \in \Theta} \pi(\theta/x).(d-\theta)^2 = E((d-\widetilde{\theta})^2/x)$$

est minimum pour $d^* = E(\widetilde{\theta}/x)$, espérance a posteriori de $\widetilde{\theta}$, et vaut $Var(\widetilde{\theta}/x)$.
Dans le cas continu, les résultats sont les mêmes ; les expressions du risque a priori et a posteriori y sont respectivement :

$$W(d) = E((d-\widetilde{\theta})^2) = \int_\Theta (d-\theta)^2.\pi(\theta)\ d\theta \text{ et}$$

$$W(d/x) = E((d-\widetilde{\theta})^2/x) = \int_\Theta (d-\theta)^2.\pi(\theta/x)\ d\theta\ .$$

3.7 Retour à la statistique classique

La statistique classique offre un modèle moins riche que la statistique bayésienne ; cependant les fonctions de risque y ont un sens et permettent d'introduire un concept important : l'admissibilité.

3.7.1 Fonctions de risque

Comme nous l'avons vu, à toute règle de décision possible, δ, on peut associer le *risque*, c-à-d la perte moyenne, $R_\theta(\delta)$, qu'elle entraîne pour chaque valeur du paramètre θ ; son expression est, *dans le cas discret*,

$$R_\theta(\delta) = \sum_{x \in \mathcal{X}} L_\theta(x).w(\delta(x), \theta)$$

et, *dans le cas continu*,

$$R_\theta(\delta) = \int_{\mathcal{X}} w(\delta(x), \theta).L_\theta(x)\ dx.$$

3.7.2 Admissibilité

On peut alors introduire l'ordre partiel suivant sur les règles de décision :
La règle de décision δ *domine* la règle de décision δ' lorsque
$R_\theta(\delta) \leq R_\theta(\delta'), \forall \theta \in \Theta$ et $\exists \theta_0 \in \Theta$ tel que $R_\theta(\delta) < R_\theta(\delta')$.

Une règle de décision est dite *admissible* lorqu'aucune autre règle ne la domine ; si, de plus, toute autre règle est dominée par une admissible, on dit que l'ensemble des admissibles est *complet*.
Lorsque l'ensemble des admissibles est complet il semble naturel d'écarter a priori toute autre règle et de choisir une règle dans cet ensemble.

3.8 Comparaison entre statistique bayésienne et statistique classique sur un exemple

Le problème de la punaise Lorsqu'on lance une punaise et qu'elle retombe sur une table, elle peut s'immobiliser de deux façons :
- sur le dos, la pointe vers le $H(aut) \leftrightarrow$ événement $X_0 = H$;
- de travers, la pointe vers le $B(as) \leftrightarrow$ événement $X_0 = B$.

Vous avez la possibilité de lancer n fois la punaise et observer à chaque fois le résultat. Au vu de vos observations, qu'allez-vous prédire concernant le résultat du $(n+1)^{ème}$ lancer ? A quelle cote seriez-vous prêt à parier sur chacun des résultats possibles ?

L'approche de la statistique classique Le résultat de chaque lancer $i \in \{1, 2, .., n\}$ est la réalisation d'une variable aléatoire X_i, bivalente avec $\mathcal{X}_i = \{H, B\}$, de paramètre θ, indépendant de i :

$$Pr(X_i = H) = \theta \ ; \ Pr(X_i = B) = 1 - \theta.$$

On suppose qu'*il existe une valeur vraie, mais inconnue, θ_0 de θ* ; tout ce que l'on sait, c'est que $\theta \in \Theta = [0, 1]$.
On admet en outre que les résultats d'un lancer ne sont pas influencés par les résultats des autres lancers ; autrement dit, que *les variables X_i*, composantes de l'échantillon (= variable aléatoire à n dimensions) $X = (X_1, .., X_i, .., X_n)$ *sont indépendantes dans leur ensemble.*
X est donc un échantillon i.i.d. et sa loi, qui dépend de θ, a pour probabilité d'un événement élémentaire (c'est une suite de n H ou B ; par exemple, pour $n = 10$, $BBHBHHHBHH$), c-à-d pour vraisemblance :

$$L_\theta(x) = p_\theta(x_1, .., x_i, .., x_n) = \theta^h.(1-\theta)^b,$$

$$\text{avec } h = \#\{i \ : \ x_i = H\} \text{ et } b = n - h = \#\{i \ : \ x_i = B\}$$

[N.B. : ce n'est pas la probabilité d'avoir h des x_i égaux à H et les b autres égaux à B, qui est égale à $\binom{n}{h}.\theta^h.(1-\theta)^b$]

Le logarithme de la vraisemblance est $\ln L(x, \theta) = h \ln \theta + b \ln(1 - \theta)$.
Comme

$$\frac{\partial \ln L(x,\theta)}{\partial \theta} = \frac{h}{\theta} - \frac{b}{1-\theta} = 0 \Longleftrightarrow \theta = \frac{h}{h+b} = \frac{h}{n},$$

l'estimateur du maximum de vraisemblance, T, a pour valeur $t = \frac{h}{n}$; c'est donc une variable $T = \frac{\widetilde{h}}{n}$, fonction de l'échantillon par l'intermédiaire de la variable $\widetilde{h} = \#\{i \ : \ X_i = H\}$. T est exhaustif, convergent, sans biais et de variance minimum dans cette classe.

La prédiction pour le $(n+1)^{ème}$ lancer est alors que :

$$Pr(X_{n+1} = H) = t = \frac{h}{n} \; ; \; Pr(X_{n+1} = B) = 1 - t = \frac{b}{n}.$$

On serait prêt à parier sur H si gain net G et mise M satisfont

$$G.Pr(X_{n+1} = H) - M.Pr(X_{n+1} = B) \geq 0 \Leftrightarrow G\frac{h}{n} - M\frac{b}{n} \geq 0 \Leftrightarrow \frac{G}{M} \geq \frac{b}{h}.$$

L'approche de la statistique bayésienne La statistique bayésienne ne considère initialement que les événements liés aux $n+1$ lancers (il n'y a pas encore de paramètre) et suppose qu'il existe une probabilité subjective sur *tous* ces événements :
par exemple, l'événement $A =$ "H au $2^{ème}$ lancer ou B au $5^{ème}$ lancer" a une probabilité $P(A)$.
On notera $p(x_1, .., x_i, .., x_{n+1})$ pour $P(X_1 = x_1, .., X_i = x_i, .., X_{n+1} = x_{n+1})$; etc..
L'idée de départ est que *les résultats des lancers ne sont pas indépendants* et que c'est bien pour cela que les résultats des n premières observations sont capables de nous informer sur le résultat du $(n+1)^{ème}$!

En revanche, on fait l'hypothèse que toutes les permutations d'une suite donnée de $(n+1)$ H ou B (par exemple, pour $n = 4$, $BBHBH$) sont équiprobables (mais les probabilités des suites qui n'en sont pas des permutations seront en général différentes) ; on aura donc
$p(B, B, H, B, H) = p(B, B, B, H, H) = p(H, H, B, B, B) = ...$

Sous cette hypothèse dite d'*échangeabilité*, un théorème, dû à De Finetti, dit qu'il existe un espace Λ ("espace des paramètres"), une densité de probabilité $\phi(\lambda)$ sur cet espace et des lois $p_\lambda(.),\ \lambda \in \Lambda$ tels que :

$$p(x_1, .., x_i, .., x_{n+1}) = \int_{\lambda \in \Lambda} [\prod_{i=1}^{n+1} p_\lambda(x_i)].\phi(\lambda)\ d(\lambda)\ ;$$

Comme $p_\lambda(x_i)$ ne prend que deux valeurs et qu'elles ne dépendent que de λ, on peut écrire que

$$p_\lambda(H) = f(\lambda);\ p_\lambda(B) = 1 - f(\lambda)\ ;$$

le changement de paramètre $\theta = f(\lambda)$ et le changement de variable correspondant dans l'intégrale, où la densité devient $\pi(\theta) = \phi(f^{-1}(\theta)).\phi'(\theta)$, permet d'obtenir une expression plus simple

$$p(x_1, .., x_i, .., x_{n+1}) = \int_{\theta \in \Theta} [\prod_{i=1}^{n+1} p_\theta(x_i)].\pi(\theta)\ d(\theta),$$

avec

$$p_\theta(H) = \theta;\ p_\theta(B) = 1 - \theta.$$

Autrement dit, tout se passe comme si :

i) on était dans un espace produit $\mathcal{X} \times \Theta$, sur lequel existait une loi jointe dont les marginales étaient d'une part la loi du couple (X, X_{n+1}), où X désigne le n-échantillon $(X_1, .., X_i, .., X_n)$, et d'autre part la loi a priori du paramètre $\widetilde{\theta}$;
et que, de plus,

ii) conditionnellement à chaque valeur θ du paramètre, les composantes X_i de (X, X_{n+1}) étaient indépendantes et de même loi.

On retrouve bien le formalisme bayésien standard.

La loi a posteriori du paramètre, ayant observé $x = (x_1, .., x_i, .., x_n)$, a donc pour densité

$$\pi(\theta/x) = \frac{L_\theta(x).\pi(\theta)}{p(x)}$$

où l'on a posé
$L_\theta(x) = p_\theta(x) = \prod_{i=1}^n p_\theta(x_i)$ et $p(x) = p(x_1, .., x_i, .., x_n)$.

La prédiction du résultat du $(n+1)^{ème}$ lancer est donnée par la probabilité a posteriori de X_{n+1} sachant x :

$$p(x_{n+1}/x) = \frac{p(x, x_{n+1})}{p(x)} = \frac{1}{p(x)} \int_{\theta \in \Theta} p_\theta(x, x_{n+1}).\pi(\theta)\ d(\theta) =$$

$$\frac{1}{p(x)} \int_{\theta \in \Theta} p_\theta(x).p_\theta(x_{n+1}).\pi(\theta)\ d(\theta) = \int_{\theta \in \Theta} p_\theta(x_{n+1}).\frac{L_\theta(x).\pi(\theta)}{p(x)}\ d(\theta) =$$

$$\int_{\theta \in \Theta} p_\theta(x_{n+1}).\pi(\theta/x)\ d(\theta),$$

soit encore

$$p(H/x) = \int_{\theta \in \Theta} \theta.\pi(\theta/x)\ d(\theta)\ ;\ p(B/x) = \int_{\theta \in \Theta} (1-\theta).\pi(\theta/x)\ d(\theta)\ ;$$

ce sont, respectivement, les espérances a posteriori de $\widetilde{\theta}$ et $[1 - \widetilde{\theta}]$.

cas particulier La prédiction du résultat du $(n+1)^{ème}$ lancer dépend de x et de la loi a priori du paramètre θ d'une façon que l'on peut préciser lorsque l'on choisit cette loi a priori dans la famille des lois Bêta.

La famille des lois Bêta. Une variable Y suit une loi $B(p,q)$ lorsqu'elle a pour support $[0,1]$ et pour densité de probabilité

$$B(y \mid p,q) = \frac{\Gamma(p+q)}{\Gamma(p).\Gamma(q)} y^{p-1}(1-y)^{q-1}$$

où $p>0$, $q>0$ et $\Gamma(r) = \int_0^\infty x^{r-1} exp(-z)\ dz$.
L'espérance d'une loi $B(p,q)$ est $\frac{p}{p+q}$

Supposons que la loi a priori π du paramètre θ suive une loi $B(\alpha_h, \alpha_b)$. On observe $x = (x_1, .., x_i, .., x_n)$ avec $h = \#\{i \ : \ x_i = H\}$ et $b = n - h = \#\{i \ : \ x_i = B\}$.
La loi a posteriori de θ si x a alors pour densité

$$\pi(\theta/x) = \frac{L_\theta(x).\pi(\theta)}{p(x)} = \frac{[\theta^h.(1-\theta)^b].[\frac{\Gamma(\alpha_h+\alpha_b)}{\Gamma(\alpha_h).\Gamma(\alpha_b)}\theta^{\alpha_h-1}(1-\theta)^{\alpha_b-1}]}{p(x)}$$

$$\alpha\ \theta^{\alpha_h+h-1}.(1-\theta)^{\alpha_b+b-1}\ ;$$

c'est donc une loi $B(\alpha_h+h, \alpha_b+b)$.
Les lois Bêta forment donc une famille de lois conjuguée de la loi de l'échantillon.

Les prédictions de $p(H/x)$ et $p(B/x)$ au $(n+1)^{ème}$ coup étant les espérances a posteriori de $\widetilde{\theta}$ et $[1-\widetilde{\theta}]$ valent :

$$p(H/x) = \frac{\alpha_h + h}{\alpha_h + \alpha_b + n} \text{ et } p(B/x) = \frac{\alpha_b + b}{\alpha_h + \alpha_b + n}.$$

Si l'on compare avec les prédictions de la statistique classique,

$$p(H/x) = \frac{h}{n} \text{ et } p(B/x) = \frac{b}{n},$$

on voit (en supposant α_h et α_b entiers) que les croyances a priori sur la valeur du paramètre équivalent à une observation antérieure à celle de l'échantillon de α_h issues H et α_b issues B.
On peut encore remarquer que lorsque $\alpha_h = \alpha_b = 1$, la loi $B(1,1)$ est la loi uniforme et que les prédictions $p(H/x)$=$\frac{1+h}{2+n}$ et $p(B/x)$=$\frac{1+b}{2+n}$ sont proches des prédictions classiques.

paris On sera prêt à parier sur H si le gain net G et la mise M satisfont $G.p(H/x) - M.p(B/x) \geq 0$
Dans le cas général ceci équivaut à

$$G \int_{\theta\in\Theta} \theta.\pi(\theta/x)\ d(\theta)\ \geq M \int_{\theta\in\Theta} (1-\theta).\pi(\theta/x)\ d(\theta)\ ;$$

et, dans le cas particulier d'une loi a priori $B(\alpha_h, \alpha_b)$ à

$$G\frac{\alpha_h + h}{\alpha_h + \alpha_b + n} \geq M\frac{\alpha_b + b}{\alpha_h + \alpha_b + n} \Leftrightarrow \frac{G}{M} \geq \frac{\alpha_b + b}{\alpha_h + h}.$$