Statistique Bayésienne

Éléments de Culture Générale

Lionel RIOU FRANÇA

INSERM U669

Mai 2009

Sommaire

1. Approche Bayésienne
   - a priori, Vraisemblance, a posteriori
   - L'Analyse Bayésienne
2. Emploi pour l'Estimation
   - Introduction
   - Méthode MCMC
3. Exemples
   - Estimation d'une Proportion
   - Modèles à Effets Aléatoires
   - Méta-analyse
   - Données Manquantes
4. Aspects Techniques
   - Convergence
   - Ajustement du Modèle

Inserm U669 Institut national de la santé et de la recherche médicale

Sommaire

5 Emploi Épistémologique
- Épistémologie
- Inférence : Approche de Fisher (1890-1962)
- Inférence : Approche de Neyman (1894-1981) et Pearson (1895-1980)
- Inférence : Approche de Jeffreys (1891-1989)

6 Vers une statistique Bayésienne ?
- Méthodes Bayésiennes
- Problèmes avec la p-value
- Critique Bayésienne du fréquentisme
- Digression : sélection de variables
- Conclusion

L'Approche Bayésienne

Comment l'expérience doit changer notre opinion sur un paramètre θ ? Cela suppose :

- Une opinion *a priori* sur les différentes valeurs plausibles du paramètre avant l'expérience,
- L'information issue de l'expérience ($\rightarrow$ *Vraisemblance*),

L'intégration de ces deux informations nous donne :

- L'opinion *a posteriori* sur θ.

La Vraisemblance

- L'inférence statistique cherche à apprendre sur la valeur dans la population d'un paramètre θ à partir d'un échantillon de données y.
- Inférence traditionnelle : une fois y connu, $P(y|\theta)$ (*fonction de vraisemblance*) quantifie à quel point certaines valeurs de θ sont compatibles avec les valeurs observées y.
- $P(y|\theta)$ résume toute l'information que les données fournissent sur le paramètre θ.

Exemple I

On veut connaître la probabilité θ de tomber sur face pour une pièce donnée. On procède à 10 lancers, on obtient l'échantillon $\{1, 1, 1, 0, 1, 1, 0, 0, 1, 1\}$, soit 7 « face » pour 3 « pile ». La probabilité d'obtenir ce résultat s'exprime en fonction de θ par :

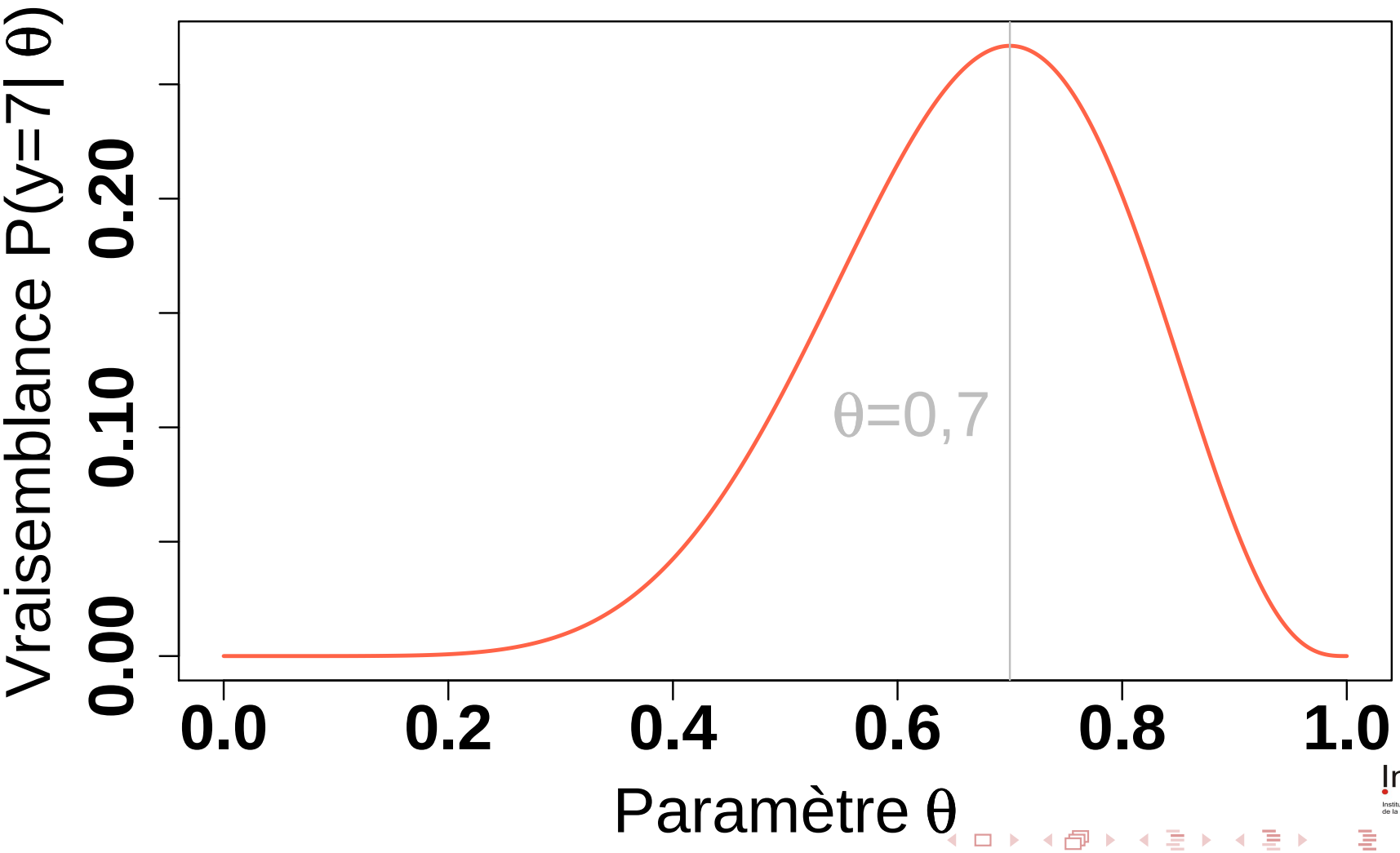
$$P(y = k) = C_n^k \theta^k (1 - \theta)^{n-k}$$

Soit ici :

$$P(y = 7) = 120 \cdot \theta^7 (1 - \theta)^3$$

On vérifie que la vraisemblance est maximale pour $\theta = 0{,}7$.

Exemple II


Vraisemblance P(y=7| θ)
0.00
0.10
0.20
θ=0,7
0.0
0.2
0.4
0.6
0.8
1.0
Paramètre θ

La Probabilité

Soit x un événement qui peut, ou pas, survenir. On peut noter $P(x)$ la probabilité de survenue de l'événement, mais il n'y a pas de consensus sur la définition de cette notion :

- Approche « fréquentiste » : $P(x)$ est la proportion de fois où x survient si on répète l'expérience un nombre infini de fois. Exemple : probabilité d'obtenir 5 fois « face » en lançant une pièce 10 fois.
- Approche « subjectiviste » : $P(x)$ reflète les croyances que l'on a sur la survenue de x. Exemple : probabilité pour qu'il y ait de la vie extraterrestre.

Quelle que soit l'approche, les propriétés mathématiques des probabilités sont les mêmes.

Qu'Attendre de l'Analyse Bayésienne ?[1]

- *Flexibilité* : Possibilité de s'adapter à chaque situation,
- *Efficience* : Emploi de toute l'information pertinente,
- *Utilité* : Fournit des informations permettant la prise de décision,
- *Éthique* : Prise en compte de l'information préexistante.

L'approche consiste à actualiser les croyances a priori sur un événement à la lumière des nouvelles informations disponibles pour obtenir une description quantitative des connaissances actuelles.

[1] Adapté de Keith Abrams, University of Leicester, dans son « Introduction to Bayesian Methods & WinBUGS ».

Théorème de Bayes

$$P(B|A) = \frac{1}{P(A)} P(A|B)P(B)$$

En appliquant à un paramètre θ et des données y ($P(y) = \text{Constante}$) :

$$P(\theta|y) = \frac{1}{P(y)} P(y|\theta)P(\theta)$$

En appliquant à une hypothèse nulle H_0 sur θ :

$$P(H_0|y) = \frac{1}{P(y)} P(y|H_0)P(H_0)$$

Principe de Base de l'Inférence Bayésienne :

La transformation de l'opinion a priori à partir des données (la fonction de vraisemblance) en une opinion a posteriori se fait à l'aide du théorème de Bayes :

$$P(\theta|y) \propto P(y|\theta)P(\theta)$$

A Posteriori $\propto$ Vraisemblance * A Priori

Exemple I

On veut connaître la probabilité θ de tomber sur face pour une pièce donnée. Nous pouvons avoir deux croyances a priori :

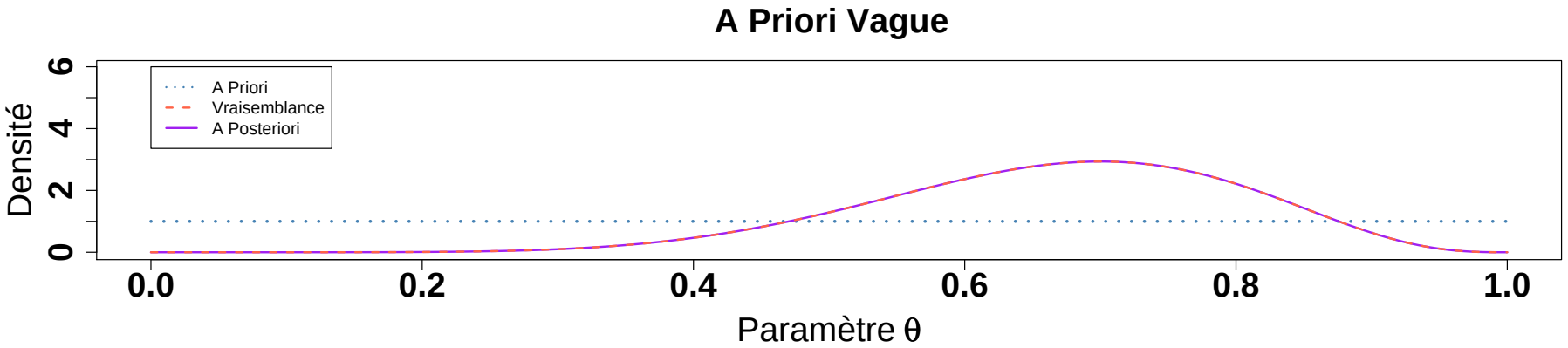
1. Toutes les valeurs de θ sont équiprobables (a priori non-informatif) : $P(\theta) = 1 \ \forall \ \theta$,
2. Il y a 50 % de chances que θ soit compris entre 45 % et 55 % (a priori subjectif).

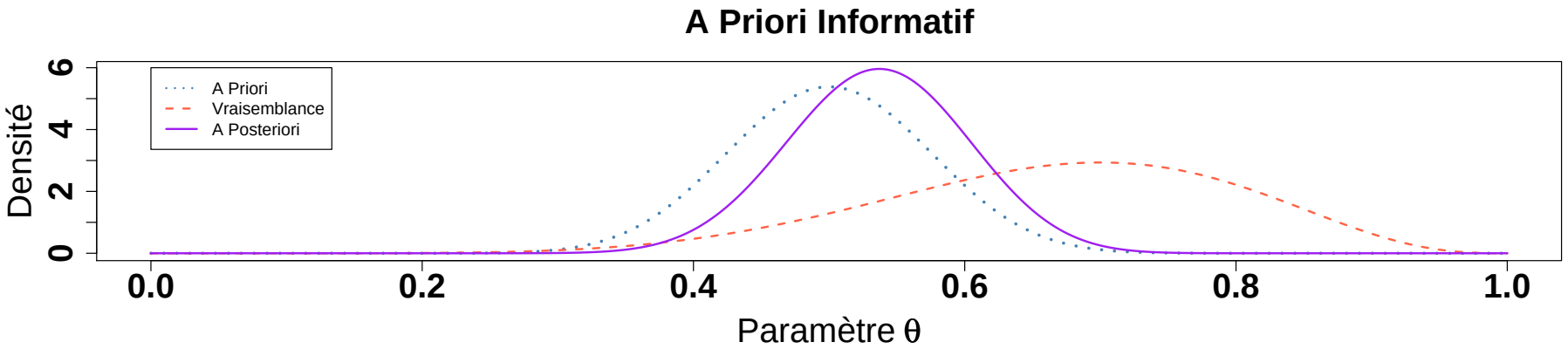
On procède à 10 lancers, on obtient l'échantillon $\{1, 1, 1, 0, 1, 1, 0, 0, 1, 1\}$, soit 7 « face » pour 3 « pile ».

Vraisemblance : $P(y|\theta) \propto \theta^7(1-\theta)^3$.

Exemple II

Avec le théorème de Bayes, on obtient la distribution a priori de θ.





Relation entre Croyances et Données

La distribution a posteriori du paramètre :

- Sera proche de la vraisemblance si l'a priori est non-informatif (ou si l'on dispose de beaucoup de données),
- Sera proche de la distribution a priori si elle est informative (ou si l'on dispose de peu de données).

En pratique :

- Avec une taille d'échantillon usuelle, $P(\theta|y) \approx P(y|\theta)$,
- Il est possible de choisir $P(\theta)$ tel que $P(\theta|y) \propto P(y|\theta) \rightarrow$ estimation Bayésienne $=$ estimation classique.

Distributions a priori

- Non-informative/vague. $P(\theta)$ est non-informative si son impact sur la distribution a posteriori $P(\theta|y)$ est minimal. Graphiquement, $P(\theta)$ sera plus plate que la vraisemblance $P(y|\theta)$, et dominée par celle-ci. Exemples : $\mathrm{Beta}(1,1)$, $N(\mu,\infty)\ldots$
- Informative. Un a priori informatif n'est pas dominé par la vraisemblance et impacte $P(\theta|y)$. Exemple : l'a posteriori d'une étude précédente peut être utilisée comme $P(\theta)$ pour l'étude courante.
- Conjuguée. Un a priori est dit conjugué à une famille de distributions si $P(\theta)$ et $P(\theta|y)$ appartiennent à la même famille.

Inserm U669 Institut national de la santé et de la recherche médicale

Solution Analytique du Problème I

Elle est basée sur la loi de probabilité bêta, dont la fonction de densité est :

$$f(x, a, b) \propto x^{a-1}(1-x)^{b-1}$$

Avec $0 \leq x \leq 1$.
On montre que si l'a priori d'une proportion θ suit une loi beta(a,b) et qu'on observe k événements sur n individus, alors la distribution a posteriori de θ suit une loi bêta(a+k,b+n-k) :

- A priori $P(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$,
- Vraisemblance $P(y|\theta) \propto \theta^{k}(1-\theta)^{n-k}$,
- A posteriori $P(\theta|y) \propto \theta^{a+k-1}(1-\theta)^{b+n-k-1}$.

Solution Analytique du Problème II

La loi bêta est conjuguée avec une vraisemblance binomiale.

Retour à l'exemple :

- $P(\theta) \propto \theta^{1-1}(1-\theta)^{1-1} = 1$ – loi bêta(1,1),
- $P(y|\theta) \propto \theta^{7}(1-\theta)^{3}$,
- $P(\theta|y) \propto \theta^{8-1}(1-\theta)^{4-1}$ – loi bêta(8,4).

N.B. Dans `R`, la densité de la loi bêta(a,b) pour une valeur x est donnée par `dbeta(x,a,b)`.

Que Faire de l'Inférence ?

L'estimation Bayésienne fournit une distribution de probabilité a posteriori $P(\theta|y)$. Elle peut servir :

- En tant que telle, pour représenter graphiquement les connaissances sur θ a l'issue de l'analyse,
- Pour fournir une estimation ponctuelle de θ (moyenne, médiane ou mode de $P(\theta|y)$,
- Pour fournir un *Intervalle de Crédibilité*, tel que θ soit dans l'intervalle $[a; b]$ avec une probabilité $(1 - \alpha)$ %,
- Pour donner directement des estimations de probabilité : $P(\theta > 0)$, par exemple.

Estimation de Paramètres

On a vu qu'il est possible de choisir l'a priori $P(\theta)$ tel que l'a posteriori $P(\theta|y)$ soit distribué comme la vraisemblance $P(y|\theta)$, et donc d'obtenir des estimations Bayésiennes qui reproduisent les résultats d'une estimation classique.

L'estimation Bayésienne peut donc être considérée, pragmatiquement, comme une méthode d'estimation parmi d'autres, compatible avec la statistique classique, et adaptée aux cas :

- Des modèles complexes (flexibilité),
- Des petits échantillons (ne dépend pas de propriétés asymptotiques).

Inserm U669 Institut national de la santé et de la recherche médicale

Quelques Procédures d'Estimation Classiques

En statistique classique, il existe plusieurs manières d'estimer les paramètres d'un modèle, les plus connues étant :

- Les moindres carrés ordinaires (OLS),
- Le maximum de vraisemblance (EMV),

Estimation par les Moindres Carrés I

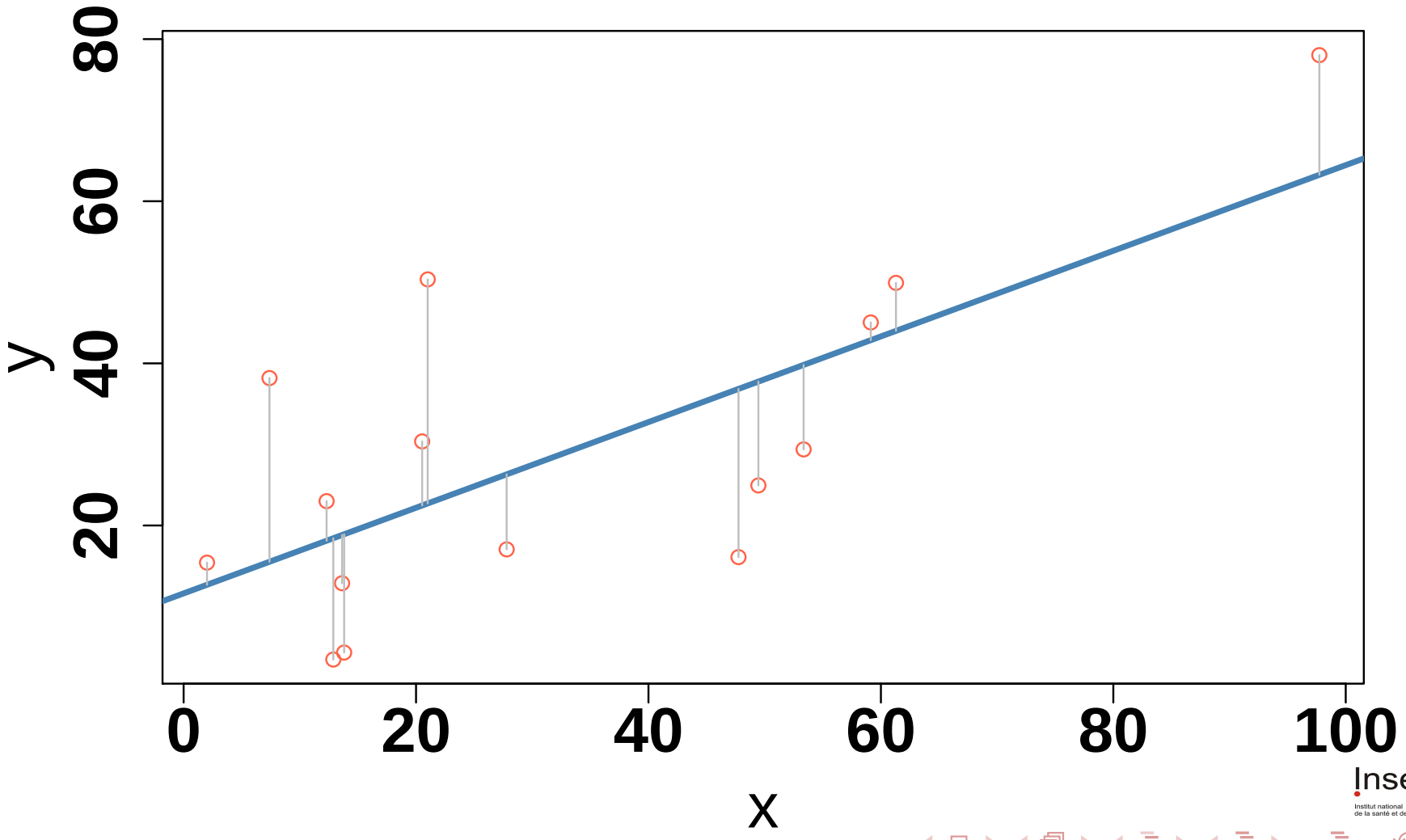
C'est la méthode employée pour le modèle de régression linéaire :

$$y_i = \beta X_i + \epsilon_i$$

On cherche l'estimation $\hat{\beta}$ qui minimise le carré de la distance entre les observations y et les prédictions $\hat{y}$. La solution est donnée par la formule matricielle :

$$\hat{\beta} = \left(X'X\right)^{-1} X'y$$

Estimation par les Moindres Carrés II



Estimation par Maximum de Vraisemblance I

La *vraisemblance* d'un modèle est égale à la probabilité d'observer les données pour des paramètres du modèle et des prédicteurs donnés.
Dans une régression linéaire, on a

$$y_i = \beta X_i + \epsilon_i$$

Avec $\epsilon \hookrightarrow N(0, \sigma^2)$, où $N()$ désigne la loi normale.
Il faut donc estimer β, mais également σ^2, la variance des termes d'erreur ϵ.

La vraisemblance est donnée par :

$$P(y|X, \beta, \sigma) = \prod_i N\left(y_i|X_i\beta, \sigma^2\right)$$

Estimation par Maximum de Vraisemblance II

Plus généralement, en considérant le vecteur des paramètres $\theta = \{\beta, \sigma\}$,

$$P(y|X,\theta) = \prod_i P\left(y_i|X_i,\theta\right)$$

L'EMV $\hat{\theta}$ des paramètres θ est la valeur de ces paramètres qui maximise la vraisemblance.

Pour la régression linéaire, $\hat{\beta}_{OLS} = \hat{\beta}_{EMV}$ mais $\hat{\sigma}_{OLS} \neq \hat{\sigma}_{EMV}$. On préfère l'estimateur OLS. Pour un modèle linéaire généralisé, la méthode OLS ne s'applique pas.

Estimation Bayésienne

L'estimation Bayésienne demande de calculer des intégrations parfois compliquées pour obtenir les distributions a posteriori. Elle repose donc sur des simulations, nommées MCMC (Markov Chain Monte Carlo).

Simulation de Monte Carlo I

La simulation de Monte Carlo vise à estimer des intégrales (ou tout autre calcul compliqué) par simulation plutôt que par analyse algébrique.

Revenons à l'exemple de la pièce de monnaie. On procède à 10 tirages, on obtient 7 « face » pour 3 « pile ». Quelle était la probabilité d'obtenir au moins 7 face si la pièce est équilibrée ?

Méthode algébrique
$P(y \geq 7) = P(y = 7) + P(y = 8) + P(y = 9) + P(y = 10)$. Ces probabilités se calculent à l'aide de la distribution binomiale :

Simulation de Monte Carlo II

$$
\begin{aligned}
P(y \geq 7) &= C_{10}^{7} \left(\frac{1}{2}\right)^{7} \left(\frac{1}{2}\right)^{3} + C_{10}^{8} \left(\frac{1}{2}\right)^{8} \left(\frac{1}{2}\right)^{2} + \\
&\quad C_{10}^{9} \left(\frac{1}{2}\right)^{9} \left(\frac{1}{2}\right)^{1} + C_{10}^{10} \left(\frac{1}{2}\right)^{10} \left(\frac{1}{2}\right)^{0} \\
&= \frac{1}{2^{10}}(120 + 45 + 10 + 1) \\
&\approx 0.1719
\end{aligned}
$$

Simulation de Monte Carlo III

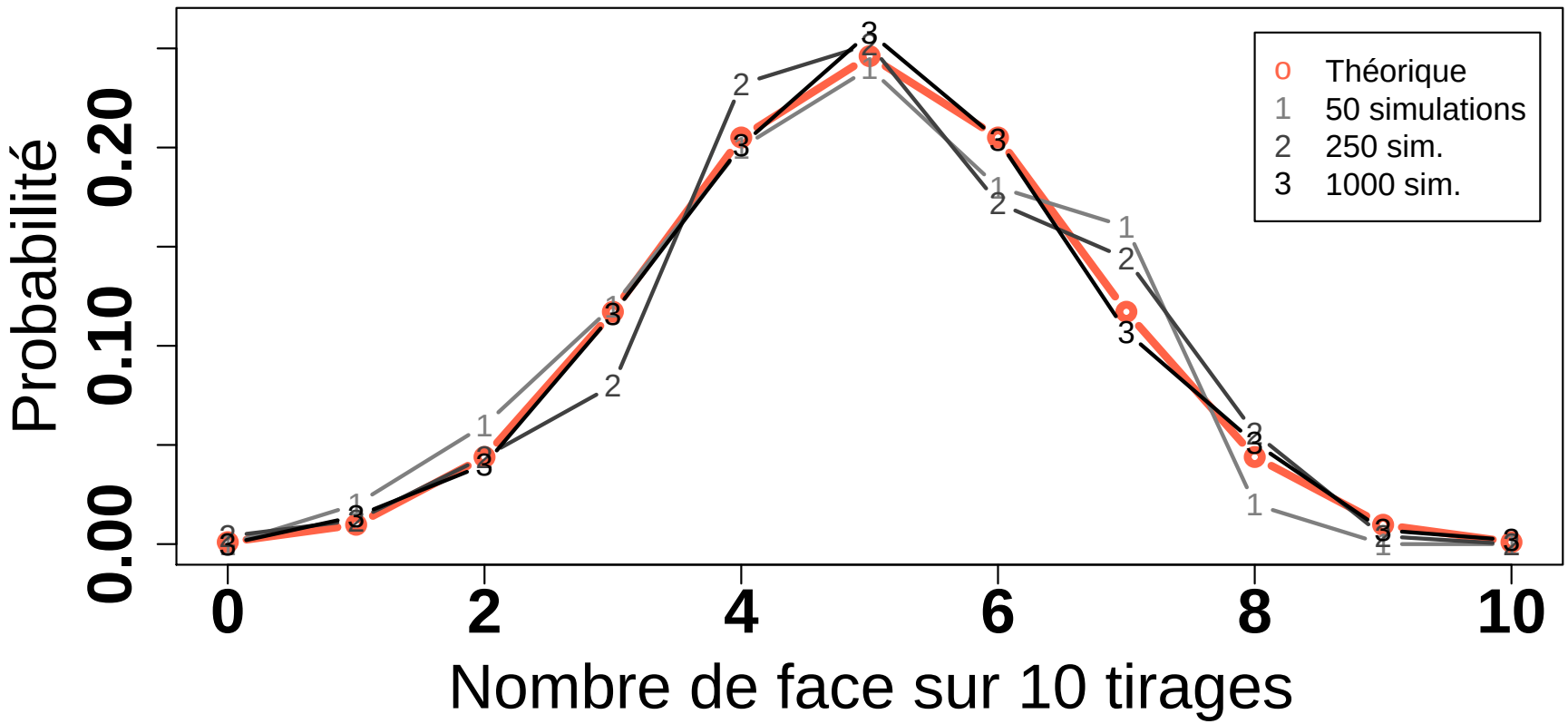
Simulation de Monte Carlo Si l'on répétait un nombre infini de fois 10 tirages de pièce et que l'on comptait le nombre de fois où l'on obtient au moins 7 face, la proportion de réalisations de cet événement serait égale à sa probabilité théorique. En pratique, on peut procéder à 10 000 simulations de 10 tirages, et voir combien de fois on obtient au moins 7 face. Dans R, cela se fait très facilement :

```
> tirages <- rbinom(10000,10,0.5)
> mean(tirages>6)

[1] 0.1801
```

Simulation de Monte Carlo IV

En agmentant le nombre de tirages, on s'approche de plus en plus de la distribution de probabilité théorique :

Simulation de Monte Carlo V



Simulation de Monte Carlo VI

Plus le nombre de simulations augmente, plus il y a convergence vers la distribution théorique. Ainsi, pour 50 simulations, la probabilité de tirer au moins 7 face (qui est théoriquement de 0,17) est de 0,18, elle passe à 0,20 pour 250 simulations et à 0,17 pour 1 000 simulations.

L'Échantillonneur de Gibbs[2]

Principe : simuler des réalisations de paramètres θ au temps t à partir de leurs valeurs au temps $t-1$ (chaîne de Markov).

1. On commence par fournir des valeurs initiales $\theta^{(0)} = \left\{\theta_1^{(0)}, \ldots, \theta_k^{(0)}\right\}$
2. On simule $\theta^{(n+1)}$, une réalisation de la loi $P(\theta|y)$, à partir de $P\left(y|\theta^{(n)}\right) P\left(\theta^{(n)}\right)$:
   1. tirage de θ_1^{n+1} à partir de $\{\theta_1^n, \ldots, \theta_k^n\}$,
   2. tirage de θ_2^{n+1} à partir de $\left\{\theta_1^{n+1}, \theta_2^n, \ldots, \theta_k^n\right\}$,
   3. ...
3. Selon la théorie des chaînes de Markov, il y aura à terme convergence vers une distribution d'équilibre.

[2] Ce n'est pas la seule méthode MCMC possible, mais c'est celle employée par BUGS (Bayesian inference Using Gibbs Sampling).

Estimation avec BUGS

BUGS = Logiciel de simulation Bayésienne MCMC. Il suffit de fournir :

- Les données y,
- Le modèle :
  - Lois a priori $P(\theta)$,
  - Vraisemblance $P(y|\theta)$,
- Des valeurs initiales,
- Le nombre de chaînes et le nombre d'itérations souhaités.

Il existe de nombreuses versions de BUGS, et plusieurs interfaces entre R et BUGS.

Estimation d'une proportion I

On procède à 10 lancers, on obtient l'échantillon $\{1, 1, 1, 0, 1, 1, 0, 0, 1, 1\}$, on veut estimer la probabilité de tomber sur face.

Analyse classique

```
> y <- c(1,1,1,0,1,1,0,0,1,1)
> binom.test(table(y)[2:1])

Exact binomial test

data:  table(y)[2:1]
number of successes = 7, number of trials = 10, p-value = 0.3438
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3475 0.9333
sample estimates:
probability of success
                   0.7
```

Estimation d'une proportion II

Analyse Bayésienne Nous allons utiliser BUGS via le programme WinBUGS et `R2WinBUGS` comme interface R.

```
> library(R2WinBUGS)
> options(digits=4)
```

Nous allons ensuite spécifier le modèle :

```
> modele <- function()
+  {
+   nbface <- sum(y[]) # Nombre de réalisations du coté face
+   nbface ~ dbin(p,n) # Vraisemblance (binomiale)
+   p ~ dbeta(1,1) # A priori (bêta, non-informatif)
+  }
> write.model(modele,"modele1.bug")
```

Puis les valeurs initiales pour p, qui seront tirées à partir de la distribution a priori :

```
> inits <- function() {p=rbeta(1,1,1)}
```

Estimation d'une proportion III

Puis les paramètres dont on veut suivre l'évolution au cours de la simulation :

```
> parametres <- "p"
```

Puis enfin les données à passer au modèle :

```
> n <- length(y)
> data <- list("y","n")
```

On peut alors lancer la simulation, sur 3 chaînes et avec 1 000 itérations :

```
> simulation <- bugs(data, inits, parametres, "modele1.bug",
                     n.chains=3, n.iter=1000)

> print(simulation,digits.summary=3)
```

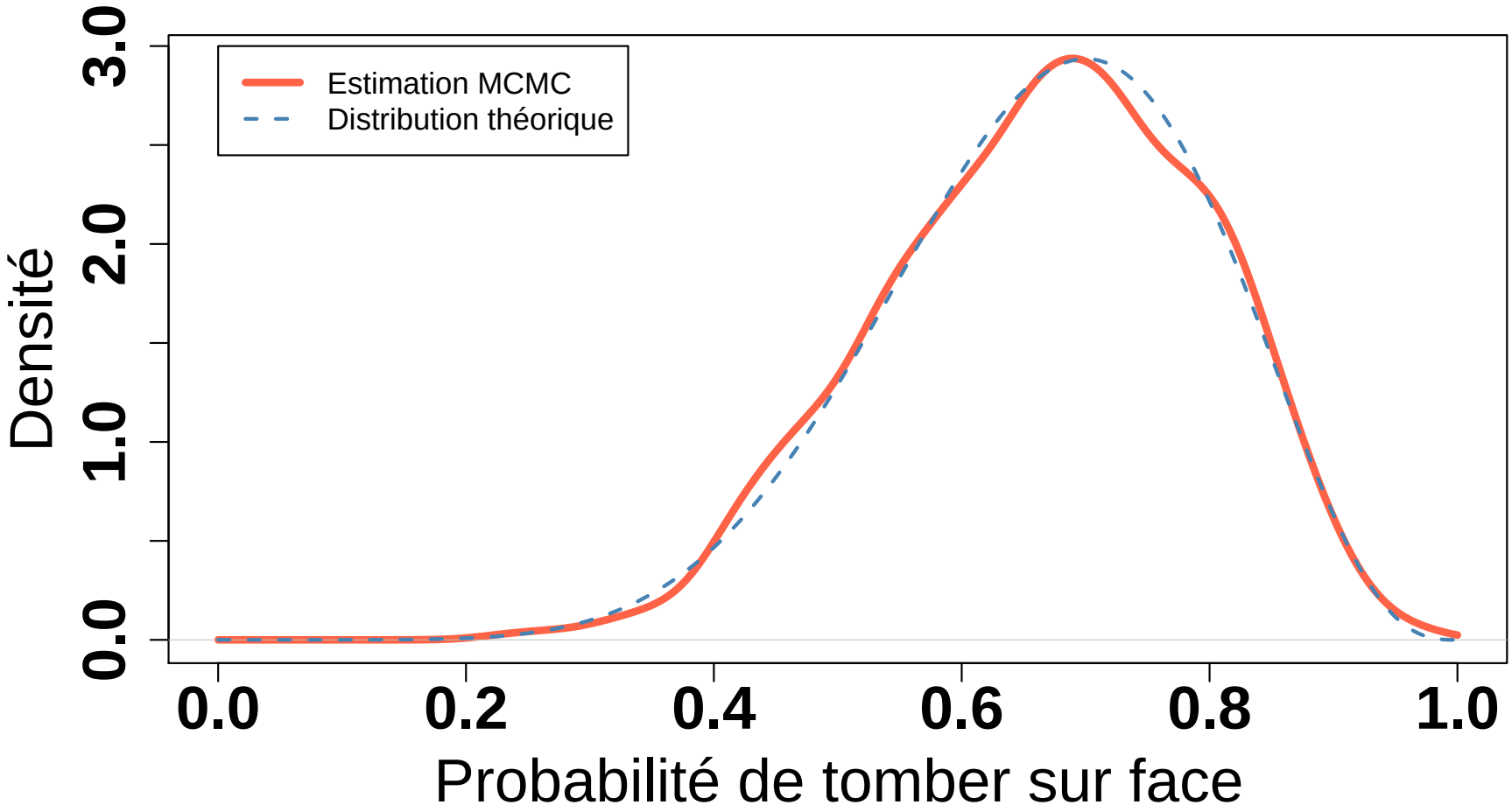
Estimation d'une proportion IV

```
Inference for Bugs model at "modele1.bug", fit using WinBUGS,
 3 chains, each with 1000 iterations (first 500 discarded)
 n.sims = 1500 iterations saved
          mean    sd  2.5%   25%   50%   75% 97.5%  Rhat n.eff
p        0.666 0.130 0.407 0.577 0.674 0.766 0.884 1.000  1500
deviance 3.509 1.215 2.643 2.728 3.066 3.779 6.669 1.001  1500

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = 0.8 and DIC = 4.3
DIC is an estimate of expected predictive error (lower deviance is better).
```

Estimation d'une proportion V

Par défaut, les estimations ne portent que sur la seconde moitié des estimations (pour ne pas être pollué par les premières simulations, où la convergence n'est pas atteinte). La probabilité de tomber sur face est estimée à 0,66 et son intervalle de crédibilité à 95 % est de [0,407 ; 0,884]. On est proches des valeurs théoriques. Avec un a priori non-informatif, la distribution a posteriori de p suit une loi bêta de paramètres 8 et 4, et donc de moyenne $8/12 = 2/3$. On vérifie que la distribution estimée par MCMC est bien proche de la distribution théorique :

Estimation d'une proportion VI



Quel Estimateur pour une Fréquence ?

L'analyse Bayésienne et l'analyse classique ne produisent pas le même estimateur d'une fréquence. Si n est la taille d'échantillon et k le nombre d'événements observés :

- Estimation classique : $\hat{p} = \frac{k}{n}$. Il s'agit d'un estimateur non biaisé.
- Estimation Bayésienne dite « objective » (a priori vague) : $\hat{p} = \frac{k+1}{n+2}$. Cet estimateur est biaisé[3].

[3] Mais plus crédible. Supposons qu'en lançant 10 fois une pièce, on tombe 0 fois sur « face ». L'estimation classique de la probabilité de tomber sur « face » est 0 %, l'estimation Bayésienne 8 %. L'impossibilité pour l'estimateur Bayésien de valoir 0 est aussi utile si l'on veut travailler sur des log-probabilités. Bien sûr, si $n \longrightarrow \infty$ alors les deux estimateurs sont comparables.

Analyse d'un Essai Multicentrique[4]

Un essai clinique sur 8 centres a comparé une crème antiseptique à un Placebo. Le critère de jugement est la guérison d'une infection. Nous cherchons à estimer l'effet de la Crème par rapport à Placebo.

		Réponse	
Centre	Traitement	Succès	Échec
1	Crème	11	25
	Placebo	10	27
2	Crème	16	4
	Placebo	22	10
3	Crème	14	5
	Placebo	7	12
4	Crème	2	14
	Placebo	1	16
5	Crème	6	11
	Placebo	0	12
6	Crème	1	10
	Placebo	0	10
7	Crème	1	4
	Placebo	1	8
8	Crème	4	2
	Placebo	6	1

[4] Données tirées de Agresti A. Categorical Data Analysis. 2002, Wiley

Modèle 1 : Hétérogénéité Entre les Centres

Certains centres peuvent avoir une plus grande probabilité de guérison, quel que soit le traitement employé.
Soient r_i^T le nombre de succès dans le groupe T (Placebo ou Crème) dans le centre i, on peut écrire :

- $r_i^T \hookrightarrow \text{Binomial}\left(p_i^T, n_i^T\right)$,
- $\text{logit}\left(p_i^P\right) = \alpha - \beta/2 + u_i$,
- $\text{logit}\left(p_i^C\right) = \alpha + \beta/2 + u_i$,
- $u_i \hookrightarrow \text{Normal}\left(0, \sigma_u^2\right)$.

Dans ce modèle, l'OR de la Crème versus Placebo est constant et égal à e^β, et l'hétérogénéité (dans la probabilité de succès) entre les centres est mesurée par σ_u^2.

Modèle 1 : Estimation I

Le modèle BUGS va s'écrire :

```
> modele <- function() {
+  for( i in 1 : N ) {
+   rp[i] ~ dbin(pp[i],np[i]) # Vraisemblance pour Placebo
+   rc[i] ~ dbin(pc[i],nc[i]) # Vraisemblance pour Crème
+   logit(pp[i]) <- alpha - beta/2 + u[i] # Modèle Placebo
+   logit(pc[i]) <- alpha + beta/2 + u[i] # Modèle Crème
+   u[i] ~ dnorm(0.0, tau) # Effet aléatoire
+   }
+   alpha ~ dnorm(0.0,1.0E-6) # A priori vague pour alpha
+   beta ~ dnorm(0.0,1.0E-6) # A priori vague pour beta
+   tau ~ dgamma(0.001,0.001) # A priori vague pour la précision
+   sigmau <- 1/sqrt(tau)
+  }
> write.model(modele,"modeleEA1.bug")
```

Inserm U669 Institut national de la santé et de la recherche médicale

Modèle 1 : Estimation II

Nous procédons ensuite à l'estimation, cette fois sur 3 chaînes, avec 100 000 itérations (1 000 suffisaient) chacune, en conservant une itération sur 150.

L'effet (supposé homogène) de la crème est estimé à 0,762, avec un écart-type de 0,301. Dans l'analyse (classique) d'origine, Agresti l'estimait à 0,739, avec un écart-type de 0,300.

Les sorties R sont les suivantes :

Modèle 1 : Estimation III

```
> print(simulation,digits.summary=3)

Inference for Bugs model at "modeleEA1.bug", fit using WinBUGS,
 3 chains, each with 1e+05 iterations (first 50000 discarded), n.thin = 150
 n.sims = 1002 iterations saved
           mean    sd   2.5%    25%    50%    75%  97.5%  Rhat n.eff
alpha    -0.858 0.685 -2.318 -1.266 -0.865 -0.422  0.460 1.001  1000
beta      0.762 0.301  0.192  0.548  0.770  0.966  1.334 1.001  1000
u[1]     -0.086 0.713 -1.482 -0.518 -0.102  0.340  1.503 1.000  1000
u[2]      1.910 0.752  0.518  1.444  1.904  2.340  3.460 1.000  1000
u[3]      1.026 0.743 -0.311  0.540  1.010  1.470  2.495 1.001  1000
u[4]     -1.406 0.798 -2.931 -1.932 -1.387 -0.888  0.066 1.003  1000
u[5]     -0.590 0.792 -2.203 -1.116 -0.563 -0.076  0.982 1.000  1000
u[6]     -1.882 0.999 -4.307 -2.422 -1.820 -1.228 -0.190 1.002  1000
u[7]     -0.809 0.890 -2.634 -1.382 -0.802 -0.242  0.914 1.001  1000
u[8]      1.885 0.893  0.269  1.300  1.860  2.452  3.742 1.001  1000
sigmau    1.746 0.640  0.877  1.291  1.636  2.000  3.293 1.000  1000
deviance 57.102 4.105 50.500 54.232 56.600 59.470 66.595 1.001  1000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = 8.6 and DIC = 65.7
DIC is an estimate of expected predictive error (lower deviance is better).
```

Inserm U669 Institut national de la santé et de la recherche médicale

Modèle 2 : Hétérogénéité Entre les Centres

Certains centres peuvent avoir une plus grande probabilité de guérison, quel que soit le traitement employé. De plus, l'effet du traitement peut être hétérogène entre les centres.
On complexifie le modèle précédent pour ne pas avoir à faire l'hypothèse d'homogénéité de l'effet de la crème en injectant un effet aléatoire sur β :

- $r_i^T \hookrightarrow \text{Binomial}\left(p_i^T, n_i^T\right)$,
- $\text{logit}\left(p_i^P\right) = \alpha - \left(\beta + b_i\right)/2 + u_i$,
- $\text{logit}\left(p_i^C\right) = \alpha + \left(\beta + b_i\right)/2 + u_i$,
- $u_i \hookrightarrow \text{Normal}\left(0, \sigma_u^2\right)$,
- $b_i \hookrightarrow \text{Normal}\left(0, \sigma_b^2\right)$.

Inserm U669 Institut national de la santé et de la recherche médicale

Modèle 2 : Estimation I

Le modèle BUGS va s'écrire :

```
> modele <- function() {
+  for( i in 1 : N ) {
+   rp[i] ~ dbin(pp[i],np[i]) # Vraisemblance pour Placebo
+   rc[i] ~ dbin(pc[i],nc[i]) # Vraisemblance pour Crème
+   logit(pp[i]) <- alpha - (beta+b[i])/2 + u[i] # Modèle Placebo
+   logit(pc[i]) <- alpha + (beta+b[i])/2 + u[i] # Modèle Crème
+   u[i] ~ dnorm(0.0, tauu) # Effet aléatoire sur constante
+   b[i] ~ dnorm(0.0, taub) # Effet aléatoire sur coefficient
+   }
+   alpha ~ dnorm(0.0,1.0E-6) # A priori vague pour alpha
+   beta ~ dnorm(0.0,1.0E-6) # A priori vague pour beta
+   tauu ~ dgamma(0.001,0.001) # A prioris vagues pour les précisions
+   sigmau <- 1/sqrt(tauu)
+   taub ~ dgamma(0.001,0.001)
+   sigmab <- 1/sqrt(taub)
+  }
> write.model(modele,"modeleEA2.bug")
```

Modèle 2 : Estimation II

Nous procédons ensuite à l'estimation, avec les mêmes paramètres que la précédente.

L'effet (supposé hétérogène) de la crème est estimé à 0,784, avec un écart-type de 0,375. Dans l'analyse (classique) d'origine, Agresti l'estimait à 0,746, avec un écart-type de 0,325.

Les sorties R sont les suivantes :

Modèle 2 : Estimation III

```
> print(simulation,digits.summary=3)

Inference for Bugs model at "modeleEA2.bug", fit using WinBUGS,
 3 chains, each with 1e+05 iterations (first 50000 discarded), n.thin = 150
 n.sims = 1002 iterations saved
          mean    sd   2.5%    25%    50%    75%  97.5%  Rhat n.eff
alpha    -0.870 0.680 -2.236 -1.265 -0.867 -0.448  0.338 1.003  1000
beta      0.784 0.375  0.087  0.538  0.786  0.991  1.600 1.006   310
u[1]     -0.061 0.711 -1.452 -0.489 -0.082  0.354  1.425 1.000  1000
u[2]      1.943 0.711  0.650  1.502  1.932  2.332  3.523 1.001  1000
u[3]      1.055 0.725 -0.331  0.621  1.030  1.483  2.547 1.000  1000
u[4]     -1.398 0.848 -3.063 -1.918 -1.392 -0.849  0.245 1.002  1000
u[5]     -0.659 0.798 -2.122 -1.126 -0.665 -0.136  0.908 1.003  1000
u[6]     -1.832 0.997 -3.885 -2.410 -1.763 -1.162 -0.181 1.000  1000
u[7]     -0.847 0.893 -2.704 -1.403 -0.846 -0.284  0.799 1.000  1000
u[8]      1.874 0.877  0.275  1.301  1.838  2.418  3.632 1.000  1000
b[1]     -0.185 0.404 -1.321 -0.293 -0.061  0.031  0.362 1.003   540
b[2]     -0.057 0.379 -1.036 -0.178 -0.012  0.088  0.679 1.001  1000
b[3]      0.181 0.411 -0.462 -0.026  0.070  0.292  1.306 1.002   820
b[4]      0.009 0.453 -0.945 -0.111  0.007  0.136  1.028 1.007   270
```

Inserm U669 Institut national de la santé et de la recherche médicale

Modèle 2 : Estimation IV

```
b[5]      0.260 0.551 -0.322 -0.025  0.073  0.367  1.826 1.007  1000
b[6]      0.039 0.456 -0.898 -0.106  0.007  0.162  1.192 1.011   220
b[7]     -0.017 0.449 -1.027 -0.131  0.001  0.118  0.899 1.009  1000
b[8]     -0.196 0.541 -1.676 -0.298 -0.044  0.042  0.615 1.008   420
sigmau    1.724 0.639  0.852  1.294  1.616  2.020  3.319 1.001  1000
sigmab    0.383 0.404  0.028  0.096  0.236  0.538  1.432 1.000  1000
deviance 56.010 4.545 48.270 52.870 55.440 58.800 66.130 1.000  1000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = 10.0 and DIC = 66.0
DIC is an estimate of expected predictive error (lower deviance is better).
```

Inserm U669 Institut national de la santé et de la recherche médicale

Méta-Analyse

La méta-analyse est une méthode de synthèse quantitative de données qui permet, à partir de plusieurs essais, d'obtenir un estimateur unique de l'effet d'une intervention.

- Modèle à effets fixes : l'effet de l'intervention est homogène entre les études. L'effet poolé est donné par la formule de Mantel-Haenszel : $\bar{Y} = \frac{\sum_i \frac{Y_i}{V_i}}{\sum_i \frac{1}{V_i}}$, où Y_i est la taille d'effet et V_i sa variance dans l'essai i. La variance de l'effet poolé est $\bar{V} = \frac{1}{\sum_i \frac{1}{V_i}}$.
- Modèle à effets aléatoires : l'effet de l'intervention varie entre les études, selon une loi normale : $Y_i \hookrightarrow N\left(\bar{Y}, \sigma^2_Y\right)$. Un estimateur classique du modèle à effets aléatoires est l'estimateur de DerSimonian-Laird.

Efficacité des Gommes à la Nicotine[5]

Les chewing-gum à la nicotine ont fait l'objet de plusieurs essais cliniques, souvent non concluants. Nous allons méta-analyser 26 d'entre eux, le critère de jugement étant l'arrêt du tabagisme pendant au moins 6 mois.

Nous voulons estimer le log-OR d'arrêt du traitement des patients du groupe traité par rapport à un groupe contrôle.

[5] Données tirées de Everitt BS & Hothorn T. A Handbook of statistical analyses using R. 2006, Chapman & Hall

Les Données I

```
> data("smoking",package="HSAUR")
> smoking
```

	qt	tt	qc	tc
Blondal89	37	92	24	90
Campbell91	21	107	21	105
Fagerstrom82	30	50	23	50
Fee82	23	180	15	172
Garcia89	21	68	5	38
Garvey00	75	405	17	203
Gross95	37	131	6	46
Hall85	18	41	10	36
Hall87	30	71	14	68
Hall96	24	98	28	103
Hjalmarson84	31	106	16	100
Huber88	31	54	11	60
Jarvis82	22	58	9	58
Jensen91	90	211	28	82
Killen84	16	44	6	20
Killen90	129	600	112	617

Les Données II

Malcolm80	6	73	3	121
McGovern92	51	146	40	127
Nakamura90	13	30	5	30
Niaura94	5	84	4	89
Pirie92	75	206	50	211
Puska79	29	116	21	113
Schneider85	9	30	6	30
Tonnesen88	23	60	12	53
Villa99	11	21	10	26
Zelman92	23	58	18	58

qt représente le nombre d'arrêts (« quitters ») dans le groupe traité, *tt* le nombre total de sujets dans le groupe, *qc* et *tc* leurs équivalents dans le groupe contrôle.

Analyse Classique : Modèle à Effets Fixes

Nous allons utiliser les fonctions R contenues dans `rmeta`.

```
> library(rmeta)
> (metaMH <- meta.MH(smoking$tt,smoking$tc,smoking$qt,smoking$qc,
+                    names=rownames(smoking)))

Fixed effects ( Mantel-Haenszel ) Meta-Analysis
Call: meta.MH(ntrt = smoking$tt, nctrl = smoking$tc, ptrt = smoking$qt,
    pctrl = smoking$qc, names = rownames(smoking))
Mantel-Haenszel OR =1.67    95% CI ( 1.47, 1.9 )
Test for heterogeneity: X^2( 25 ) = 34.9 ( p-value 0.09 )
```

L'OR poolé est estimé à 1,67, son intervalle de confiance à 95 % est de [1,47 ; 1,90].

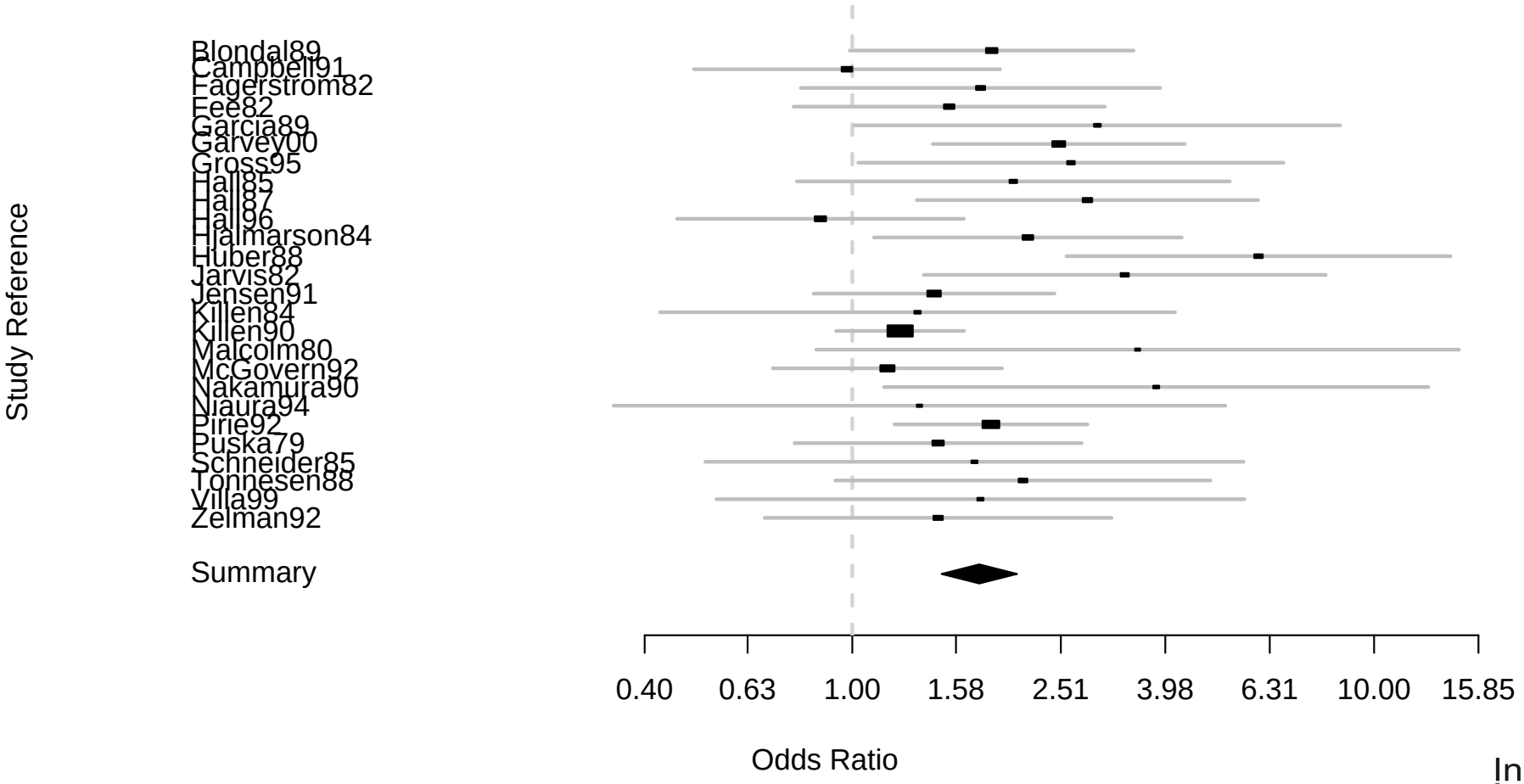
Analyse Classique : Modèle à Effets Aléatoires I

Nous allons utiliser les fonctions R contenues dans `rmeta`.

```
> (metaDSL <- meta.DSL(smoking$tt,smoking$tc,smoking$qt,smoking$qc,
+                      names=rownames(smoking)))

Random effects ( DerSimonian-Laird ) meta-analysis
Call: meta.DSL(ntrt = smoking$tt, nctrl = smoking$tc, ptrt = smoking$qt,
    pctrl = smoking$qc, names = rownames(smoking))
Summary OR= 1.75    95% CI ( 1.48, 2.07 )
Estimated random effects variance: 0.05
```

L'OR poolé est estimé à 1,75, son intervalle de confiance à 95 % est de [1,48 ; 2,08]. L'hétérogénéité est faible, avec une variance de l'effet du traitement inter-essais estimée à 0,05.

Analyse Classique : Modèle à Effets Aléatoires II



Analyse Bayésienne : Modèle à Effets Fixes I

Nous allons considérer le modèle suivant :

- $q_i^T \hookrightarrow \text{Binomial}\left(p_i^T, t_i^T\right)$,
- $\text{logit}\left(p_i^C\right) = \alpha_i$,
- $\text{logit}\left(p_i^T\right) = \alpha_i + \beta$,

Où β est homogène sur tous les centres et représente le log-OR du traitement.

Le programe BUGS correspondant est :

Analyse Bayésienne : Modèle à Effets Fixes II

```
> modele <- function() {
+  for( i in 1 : N ) {
+   qc[i] ~ dbin(pc[i],tc[i]) # Vraisemblance pour Contrôle
+   qt[i] ~ dbin(pt[i],tt[i]) # Vraisemblance pour Traités
+   logit(pc[i]) <- alpha[i]        # Modèle Contrôle
+   logit(pt[i]) <- alpha[i] + beta # Modèle Traités
+   alpha[i] ~ dnorm(0.0,1.0E-6) # A prioris pour les effets Contrôle
+   }
+   or <- exp(beta) # OR estimé
+   beta ~ dnorm(0.0,1.0E-6) # A priori vague pour l'effet du traitement
+  }
> write.model(modele,"modeleMAEF.bug")
```

On procède à 10 000 itérations, en conservant une itération sur 15, et en employant 3 chaînes. Les sorties R sont les suivantes :

Analyse Bayésienne : Modèle à Effets Fixes III

```
> print(simulation,digits.summary=3)

Inference for Bugs model at "modeleMAEF.bug", fit using WinBUGS,
 3 chains, each with 10000 iterations (first 5000 discarded), n.thin = 15
 n.sims = 1002 iterations saved
           mean    sd    2.5%     25%     50%     75%   97.5% Rhat n.eff
beta      0.514 0.069   0.381   0.468   0.514   0.559   0.653    1  1000
or        1.676 0.117   1.463   1.597   1.673   1.749   1.922    1  1000
deviance 294.610 7.537 281.702 289.400 294.250 299.100 311.400    1  1000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = 27.2 and DIC = 321.8
DIC is an estimate of expected predictive error (lower deviance is better).
```

Analyse Bayésienne : Modèle à Effets Fixes IV

En analyse classique, l'OR poolé est estimé à 1,67, son intervalle de confiance à [1,47 ; 1,90].

En analyse Bayésienne, l'OR poolé est estimé à 1,68, son intervalle de crédibilité à [1,46 ; 1,92].

Analyse Bayésienne : Modèle à Effets Aléatoires I

Nous allons considérer le modèle suivant :

- $q_i^T \hookrightarrow \text{Binomial}\left(p_i^T, t_i^T\right)$,
- $\text{logit}\left(p_i^C\right) = \alpha_i$,
- $\text{logit}\left(p_i^T\right) = \alpha_i + \beta_i$,
- $\beta_i \hookrightarrow \text{Normal}\left(\bar{\beta}, \sigma_\beta^2\right)$.

Où β représente le log-OR du traitement, et varie d'un essai à l'autre.

Le programe BUGS correspondant est :

Analyse Bayésienne : Modèle à Effets Aléatoires II

```
> modele <- function() {
+  for( i in 1 : N ) {
+   qc[i] ~ dbin(pc[i],tc[i]) # Vraisemblance pour Contrôle
+   qt[i] ~ dbin(pt[i],tt[i]) # Vraisemblance pour Traités
+   logit(pc[i]) <- alpha[i]            # Modèle Contrôle
+   logit(pt[i]) <- alpha[i] + beta[i] # Modèle Traités
+   beta[i] ~ dnorm(betabar,tau) # Effet aléatoire sur les essais
+   alpha[i] ~ dnorm(0.0,1.0E-6) # A prioris pour les effets Contrôle
+   }
+   or <- exp(betabar) # OR estimé
+   betabar ~ dnorm(0.0,1.0E-6) # A priori vague pour l'effet du traitement
+   tau <- 1/(sigmabeta*sigmabeta)
+   sigmabeta ~ dunif(0,10)     # A priori pour l'écart-type de l'effet aléatoire
+  }
> write.model(modele,"modeleMAEA.bug")
```

On procède à 100 000 itérations, en conservant une itération sur 150, et en employant 3 chaînes. Les sorties R sont les suivantes :

Analyse Bayésienne : Modèle à Effets Aléatoires III

```
> print(simulation,digits.summary=3)

Inference for Bugs model at "modeleMAEA.bug", fit using WinBUGS,
 3 chains, each with 1e+05 iterations (first 50000 discarded), n.thin = 150
 n.sims = 1002 iterations saved
             mean    sd    2.5%     25%     50%     75%   97.5%  Rhat n.eff
betabar     0.582 0.097   0.404   0.517   0.581   0.637   0.784 1.003   770
sigmabeta   0.266 0.117   0.050   0.185   0.258   0.338   0.509 1.026   550
or          1.799 0.177   1.497   1.677   1.787   1.890   2.190 1.002   960
deviance  281.512 9.499 264.100 274.800 281.300 287.775 301.695 1.000  1000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = 36.0 and DIC = 317.5
DIC is an estimate of expected predictive error (lower deviance is better).
```

Analyse Bayésienne : Modèle à Effets Aléatoires IV

En analyse classique, l'OR poolé est estimé à 1,75, son intervalle de confiance à [1,48 ; 2,08], l'hétérogénéité à 0,05.

En analyse Bayésienne, l'OR poolé est estimé à 1,80, son intervalle de crédibilité à [1,50 ; 2,19], l'hétérogénéité (en variance) à $0,266^2 = 0,07$.

Réussite Scolaire en Anglais[6]

Nous allons travailler à partir d'un échantillon de 4 873 étudiants répartis sur 172 écoles d'une étude menée au Royaume Uni. Sont mesurées :

- nlitpre : score à un test de littérature avant l'entrée en classe,
- nlitpost : score après l'entrée,
- fsm : éligible à la gratuité de la cantine,
- gend : sexe (1=masculin, 0=féminin),
- tentry : moment d'entrée en classe (1=printemps, 0=automne)

[6] Cet exemple est issu d'un tutoriel de Carpenter & Kenward disponible sur www.missingdata.org.uk

Le Modèle

Nous voulons prédire le score en anglais après l'entrée des élèves à partir des autres variables, en prenant en compte le fait que les élèves sont regroupés par écoles (identifiées par la variable schn). Pour un élève i dans l'école j :

$$\begin{aligned}
\text{nlitpost}_{ij} &= \beta_{0ij} + \beta_1 \text{nlitpre}_{ij} + \beta_2 \text{gend}_{ij} + \beta_3 \text{fsm}_{ij} + \beta_4 \text{tentry}_{ij} \\
\beta_{0ij} &= \beta_0 + u_j + \epsilon_{ij} \\
u_j &\hookrightarrow \text{Normal}\left(0, \sigma_u^2\right) \\
\epsilon_{ij} &\hookrightarrow \text{Normal}\left(0, \sigma_\epsilon^2\right)
\end{aligned}$$

La variance est donc décomposée entre variance individuelle (σ_ϵ^2) et variance à l'échelle de la classe (σ_u^2).

Inserm U669 Institut national de la santé et de la recherche médicale

Le Modèle : Formulation Alternative

Afin de faciliter la programmation BUGS, le modèle peut s'écrire :

$$\begin{aligned} \text{nlitpost}_i &\hookrightarrow \text{Normal}\left(\mu_i, \sigma^2_\epsilon\right) \\ \mu_i &= \beta_{0j[i]} + \beta_1 \text{nlitpre}_i + \beta_2 \text{gend}_i + \beta_3 \text{fsm}_i + \beta_4 \text{tentry}_i \\ \beta_{0j} &\hookrightarrow \text{Normal}\left(\beta_0, \sigma^2_u\right) \end{aligned}$$

Avec cette notation, $j[i]$ désigne l'école de l'étudiant i.

Analyse Bayésienne : Suppression des Données Manquantes I

Le fichier de données contient des valeurs manquantes. Nous allons les ignorer et faire l'analyse sur les données complètes (3 132 élèves, 171 écoles).

Le programe BUGS correspondant est :

```
> modele <- function() {
+  for( i in 1:NbEl ) { # Régression au niveau des étudiants
+   nlitpost[i] ~ dnorm(mu[i], tau.e ) # Le score a une distribution normale
+   mu[i] <- b.cons[schn[i]] + b.pre*nlitpre[i] + b.gend*gend[i] +
+            b.fsmn*fsmn[i] + b.tentry*tentry[i] # Modèle de régression linéaire
+            # schn[i] identifie l'école de l'élève i.
+   }
+  for (j in 1:NbEc) { # Régression au niveau des écoles
+   b.cons[j] ~ dnorm(mu.cons, tau.u) # Effet de l'école j
+   }
```

Analyse Bayésienne : Suppression des Données Manquantes II

```
+   # A prioris pour les variances au niveau de l'élève et de l'école
+   tau.e ~ dgamma(0.001, 0.001)
+   tau.u ~ dgamma(0.001, 0.001)
+   var.e <- 1/tau.e # Variance au niveau des élèves
+   var.u <- 1/tau.u # Variance au niveau des écoles
+   # A prioris pour les coefficients
+   b.pre ~ dnorm(0.0,1.0E-6)
+   b.gend ~ dnorm(0.0,1.0E-6)
+   b.fsmn ~ dnorm(0.0,1.0E-6)
+   b.tentry ~ dnorm(0.0,1.0E-6)
+   mu.cons ~ dnorm(0.0,1.0E-6)
+   }
> write.model(modele,"modeleEcole.bug")
```

On procède à 5 000 itérations, en conservant une itération sur 7, et en employant 3 chaînes.

Analyse Bayésienne : Suppression des Données Manquantes III

```
> print(simulation,digits.summary=2)

Inference for Bugs model at "modeleEcole.bug", fit using WinBUGS,
 3 chains, each with 5000 iterations (first 2500 discarded), n.thin = 7
 n.sims = 1074 iterations saved
           mean    sd    2.5%     25%     50%     75%   97.5% Rhat n.eff
b.pre      0.70  0.01    0.68    0.69    0.70    0.71    0.73 1.01   550
b.gend    -0.02  0.02   -0.07   -0.04   -0.02   -0.01    0.02 1.01   220
b.fsmn    -0.06  0.03   -0.12   -0.08   -0.06   -0.04    0.00 1.00  1100
b.tentry  -0.53  0.04   -0.60   -0.56   -0.53   -0.51   -0.46 1.01   300
mu.cons    0.14  0.04    0.06    0.11    0.14    0.16    0.22 1.00   730
var.e      0.35  0.01    0.33    0.34    0.35    0.35    0.36 1.00   530
var.u      0.18  0.02    0.14    0.16    0.18    0.19    0.23 1.00  1100
deviance 5563.72 19.90 5525.00 5550.00 5563.00 5576.00 5604.17 1.00   960

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
```

Analyse Bayésienne : Suppression des Données Manquantes IV

```
DIC info (using the rule, pD = Dbar-Dhat)
pD = 149.0 and DIC = 5712.7
DIC is an estimate of expected predictive error (lower deviance is better).
```

Nous allons nous focaliser sur l'effet du sexe. Il est estimé à $\beta_2 = -0,02$ (les garçons ont tendance à avoir des moins bons scores en anglais).

L'intervalle de crédibilité à 95 % de l'effet est de [-0,07 ; 0,02]. Il contient la valeur 0 : il n'y a pas de différence significative entre filles et garçons.

Analyse Bayésienne : Imputation des Données Manquantes I

L'analyse des données complètes, outre le fait qu'elle nous prive de 36 % des données de l'échantillon, peut conduire à des estimations biasées. Il est donc préférable d'imputer les valeurs manquantes.

Ici, seule la variable `nlitpre` contient des valeurs manquantes. Dans l'hypothèse MAR (Missing At Random), le fait que `nlitpre` soit manquant est indépendant de `nlitpre`, conditionnellement à `nlitpost`, `fsmn`, `gend` et `tentry`. On peut donc le prédire par régression, en gardant à l'esprit que les étudiants sont groupés par écoles :

Analyse Bayésienne : Imputation des Données Manquantes II

$$
\begin{aligned}
\text{nlitpre}_i &\hookrightarrow \text{Normal}\left(\mu_i^{imp}, \sigma_\epsilon^{imp^2}\right) \\
\mu_i^{imp} &= \beta_{0j[i]}^{imp} + \beta_1^{imp}\text{nlitpost}_i + \beta_2^{imp}\text{gend}_i + \beta_3^{imp}\text{fsm}_i + \\
&\quad \beta_4^{imp}\text{tentry}_i \\
\beta_{0j}^{imp} &\hookrightarrow \text{Normal}\left(\beta_0^{imp}, \sigma_u^{imp^2}\right)
\end{aligned}
$$

Il suffit de rajouter cette nouvelle régression dans le programe BUGS, et de forcer le sens de l'estimation du modèle d'imputation vers le modèle d'intérêt (commande BUGS cut) :

Analyse Bayésienne : Imputation des Données Manquantes III

```
> modele <- function() {
+  for( i in 1:NbEl ) { # Régression au niveau des étudiants
+ ## Modèle d'intérêt
+   nlitpost[i] ~ dnorm(mu[i], tau.e ) # Le score a une distribution normale
+   mu[i] <- b.cons[schn[i]] + b.pre*nlitpre.cut[i] + b.gend*gend[i] +
+            b.fsmn*fsmn[i] +
+            b.tentry*tentry[i] # Modèle de régression linéaire
+   nlitpre.cut[i]<-cut(nlitpre[i]) # Force le sens de l'estimation
+ ## Modèle d'imputation
+   nlitpre[i]~dnorm(mu.imp[i],tau.imp.e)
+   mu.imp[i] <- b.imp.cons[schn[i]] + b.imp.post*nlitpost[i] +
+                b.imp.gend*gend[i] + b.imp.fsmn*fsmn[i] +
+                b.imp.tentry*tentry[i]
+   }
+  for (j in 1:NbEc) { # Régression au niveau des écoles
+   b.cons[j] ~ dnorm(mu.cons, tau.u)
+   b.imp.cons[j] ~ dnorm(mu.imp.cons, tau.imp.u)
```

Analyse Bayésienne : Imputation des Données Manquantes IV

```
+   }
+  # A prioris pour les variances au niveau de l'élève et de l'école
+ # Modèle d'intérêt
+  tau.e ~ dgamma(0.001, 0.001)
+  tau.u ~ dgamma(0.001, 0.001)
+  var.e <- 1/tau.e # Variance au niveau des élèves
+  var.u <- 1/tau.u # Variance au niveau des écoles
+ # Modèle d'imputation
+  tau.imp.e ~ dgamma(0.001, 0.001)
+  tau.imp.u ~ dgamma(0.001, 0.001)
+  var.imp.e <- 1/tau.imp.e # Variance au niveau des élèves
+  var.imp.u <- 1/tau.imp.u # Variance au niveau des écoles
+  # A prioris pour les coefficients
+ # Modèle d'intérêt
+  b.pre ~ dnorm(0.0,1.0E-6)
+  b.gend ~ dnorm(0.0,1.0E-6)
+  b.fsmn ~ dnorm(0.0,1.0E-6)
```

Inserm U669 Institut national de la santé et de la recherche médicale

Analyse Bayésienne : Imputation des Données Manquantes V

```
+  b.tentry ~ dnorm(0.0,1.0E-6)
+  mu.cons ~ dnorm(0.0,1.0E-6)
+ # Modèle d'imputation
+  b.imp.post ~ dnorm(0.0,1.0E-6)
+  b.imp.gend ~ dnorm(0.0,1.0E-6)
+  b.imp.fsmn ~ dnorm(0.0,1.0E-6)
+  b.imp.tentry ~ dnorm(0.0,1.0E-6)
+  mu.imp.cons ~ dnorm(0.0,1.0E-6)
+  }
> write.model(modele,"modeleEcoleImp.bug")
```

On procède à 5 000 itérations, en conservant une itération sur 7, et en employant 3 chaînes.

Analyse Bayésienne : Imputation des Données Manquantes VI

```
> print(simulation,digits.summary=2)

Inference for Bugs model at "modeleEcoleImp.bug", fit using WinBUGS,
 3 chains, each with 5000 iterations (first 2500 discarded), n.thin = 7
 n.sims = 1074 iterations saved
             mean    sd     2.5%      25%      50%      75%    97.5% Rhat n.eff
b.pre        0.71  0.01     0.69     0.71     0.71     0.72     0.74 1.00  1100
b.gend      -0.05  0.02    -0.09    -0.06    -0.05    -0.04    -0.01 1.00  1100
b.fsmn      -0.09  0.03    -0.15    -0.11    -0.09    -0.07    -0.04 1.00  1100
b.tentry    -0.54  0.03    -0.60    -0.56    -0.54    -0.52    -0.48 1.00   720
mu.cons      0.22  0.04     0.15     0.20     0.22     0.25     0.30 1.01   380
var.e        0.36  0.01     0.34     0.36     0.36     0.37     0.38 1.00  1100
var.u        0.19  0.02     0.15     0.17     0.19     0.20     0.24 1.00   480
deviance 14438.76 81.51 14280.00 14382.50 14440.00 14490.00 14600.00 1.00  1100

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
```

Analyse Bayésienne : Imputation des Données Manquantes VII

```
DIC info (using the rule, pD = Dbar-Dhat)
pD = 1247.2 and DIC = 15685.9
DIC is an estimate of expected predictive error (lower deviance is better).
```

L'effet du sexe est maintenant estimé à $\beta_2 = -0,05$ (contre $\beta_2 = -0,02$ en supprimant les observations manquantes).

L'intervalle de crédibilité à 95 % de l'effet est de [-0,09 ; -0,01]. Il ne contient pas la valeur 0, les filles réussissent significativement mieux que les garçons.

MCMC et Convergence

Deux problèmes avec l'estimation MCMC :

- Si le nombre d'itérations $n \longrightarrow \infty$, on converge bien vers la distribution $P(\theta|y)$. Mais comment savoir si l'on a n assez grand pour obtenir la convergence ?
- Du fait que les itérations soient basées sur une chaîne de Markov, θ^t et θ^{t+1} seront corrélées. Cela impacte sur la vitesse de convergence et sur la précision des estimations.

Critères Quantitatifs de Convergence

La vérification de la convergence repose souvent sur l'étude du comportement de m ($m > 1$) chaînes de Markov indépendantes, sur n itérations :

- Variance intra-chaînes : $W = \frac{1}{m} \sum_{j=1}^{m} s_j^2$, avec $s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\theta_j^{(i)} - \bar{\theta}_j \right)^2$.
- Variance inter-chaînes : $B = \frac{n}{m-1} \sum_{j=1}^{m} \left(\bar{\theta}_j - \bar{\theta} \right)^2$.
- $\hat{R} = \sqrt{\frac{\frac{n-1}{n} W + \frac{1}{n} B}{W}}$ converge vers 1 lorsque $n \to \infty$. Si $\hat{R}$ n'est pas proche de 1, le nombre d'itérations n est trop faible.

Taille d'Échantillon Effective

Du fait de l'autocorrélation des valeurs $\theta_j^{(i)}$ dans chaque chaîne, nous ne pouvons pas considérer que les estimations sont basées sur $m \cdot n$ échantillons indépendants de θ. La taille d'échantillon corrigée de l'autocorrélation est :

$$n_{eff} = m \cdot n \frac{\frac{n-1}{n} W + \frac{1}{n} B}{B}$$

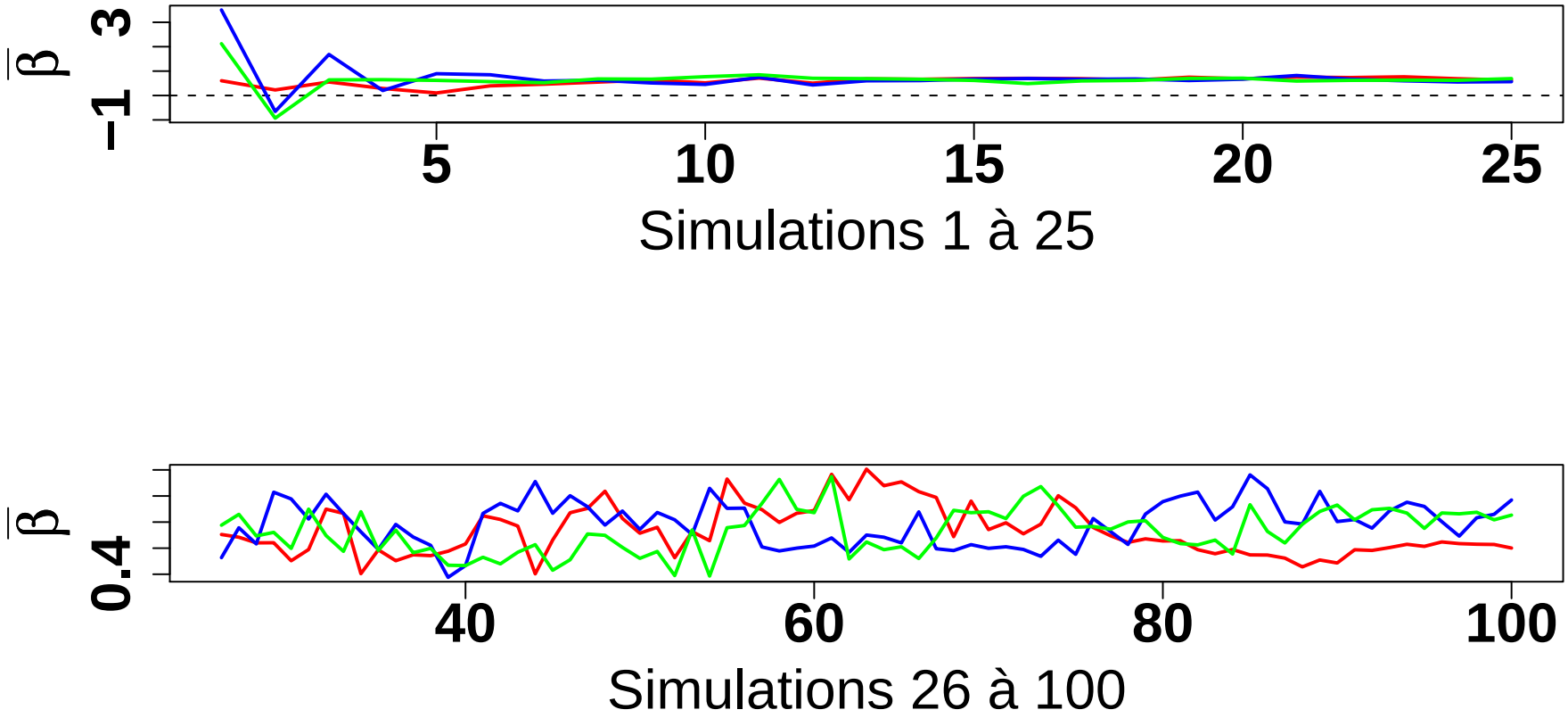
Comme en général m est petit, l'estimation de B est imprécise. `R2WinBUGS` renvoie donc $n_{eff} = \min\left(m \cdot n, n_{eff}\right)$.

Exemple : Méta-Analyse Nicotine I

La vérification graphique des itérations peut également donner une idée de la convergence. Elle peut être étudiée en vérifiant que les chaînes de Markov se mélangent bien entre elles, signe qu'elles ont convergé vers $P(\theta|y)$.

Si l'on s'intéresse à l'évolution de $\bar{\beta}$, l'effet global des gommes à la nicotine dans le modèle de méta-analyse à effets aléatoires sur les 50 premières itérations pour chacune des 3 chaînes de Markov :

Exemple : Méta-Analyse Nicotine II



Exemple : Méta-Analyse Nicotine III

En revenant sur les résultats finaux (100 000 itérations, une itération sur 150 de la seconde moitié de conservées) :

```
Inference for Bugs model at "modeleMAEA.bug", fit using WinBUGS,
 3 chains, each with 1e+05 iterations (first 50000 discarded), n.thin = 150
 n.sims = 1002 iterations saved
             mean    sd    2.5%     25%     50%     75%   97.5%  Rhat n.eff
betabar     0.582 0.097   0.404   0.517   0.581   0.637   0.784 1.003   770
sigmabeta   0.266 0.117   0.050   0.185   0.258   0.338   0.509 1.026   550
or          1.799 0.177   1.497   1.677   1.787   1.890   2.190 1.002   960
deviance  281.512 9.499 264.100 274.800 281.300 287.775 301.695 1.000  1000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = 36.0 and DIC = 317.5
DIC is an estimate of expected predictive error (lower deviance is better).
```

Exemple : Méta-Analyse Nicotine IV

C'est surtout pour σ_{β_i} que la convergence est plus discutable : $\hat{R} = 1.03$. Cette convergence plus longue est liée à une plus importante autocorrélation : $n_{eff} = 550$.

Le Critère DIC I

Pour mesurer à quel point le modèle est ajusté aux données, on peut avoir recours à plusieurs grandeurs :

- Vraisemblance $V = P(y|\theta)$. Plus V est élevée, meilleur est l'ajustement.
- Déviance $D = -2 \log V$. Plus D est faible, meilleur est l'ajustement. En pratique, on travaille sur l'espérance de D : $\bar{D} = E[D(\theta)] = E[-2 \log P(y|\theta)]$.

Le Critère DIC II

Comme $\bar{D}$ va diminuer si l'on augmente le nombre de paramètres du modèle, il faut introduire un terme de pénalité :

- Nombre effectif de paramètres : $pD = \bar{D} - D\left(\bar{\theta}\right)$. Plus pD est important, plus il est facile pour le modèle d'ajuster les données.
- Deviance Information Criterion : $DIC = pD + \bar{D}$. Plus le DIC est faible, plus l'ajustement (pénalisé du nombre de paramètres) est bon.

Le DIC est l'équivalent Bayésien de l'AIC.

Exemple : Méta-Analyse Nicotine I

Dans le modèle à effets fixes, on dispose de 26 essais (donc 26 paramètres α_i), on estime un effet unique β des chewing-gum. Dans des bonnes conditions d'ajustement, on s'attend à $pD = 26 + 1$.

Dans un modèle à effets aléatoires, pD dépend de l'hétérogénéité et donc de σ_β. Si $\sigma_\beta \longrightarrow 0$, on obtient un modèle à effets fixes, si $\sigma_\beta \longrightarrow \infty$, on ne peut pooler les essais, on a un effet β par étude.

Exemple : Méta-Analyse Nicotine II

Si l'on revient aux résultats :

Modèle	pD	DIC
Effets Fixes	27,2	321,8
Effets Aléatoires	36,0	317,5

Le modèle à effets aléatoires procure donc un meilleur ajustement (à comparer avec le test fréquentiste d'hétérogénéité qui conduisait à favoriser le modèle à effets fixes – p=0,09).

Statistique et Science

L'*épistémologie* « étudie de manière critique la méthode scientifique, les formes logiques et modes d'inférence utilisés en science, de même que les principes, concepts fondamentaux, théories et résultats des diverses sciences, et ce, afin de déterminer leur origine logique, leur valeur et leur portée objective » (Nadeau R, 1999).

Selon les Bayésiens, la *statistique* est la science qui s'occupe du degré de preuve contenu dans les observations.

Les Positions de Popper

- Les théories scientifiques ne peuvent être *justifiées*, mais seulement *évaluées* à partir de tests qui cherchent à les mettre à l'épreuve.
- Connaissance scientifique : provisoire, conjecturale, hypothétique. Possibilité de confirmation (provisoire) ou de réfutation (définitive).
- Face à une infinité de théories possibles, on ne peut qu'éliminer les théories qu'on peut démontrer comme fausses, et choisir *rationnellement* entre les autres.

$\Rightarrow$ Une théorie n'est scientifique (et non métaphysique) que si elle permet des prédictions qui peuvent s'avérer par la suite incorrectes (théories « falsifiables »).

Inserm U669 Institut national de la santé et de la recherche médicale

Choisir Rationnellement Entre les Théories[7]

Épistémologie Bayésienne (Jeffreys, De Finetti, Savage) :

- Toute probabilité est un degré de croyance et non une propriété d'un objet,
- Sous des conditions (faibles) de rationalité, tout degré de croyance obéit aux lois de probabilités,
- Les modèles scientifiques prédisent les données, de manière absolue ou probabiliste,
- Il est possible de calculer comment les données modifient les degrés de croyance entre différents modèles (théorème de Bayes).

[7] non falsifiées.

Fisher et la p-value

- On part d'une hypothèse nulle H_0 :
  - La vaccination contre l'hépatite B n'est pas associée à la sclérose en plaques,
  - La prise de vitamine C n'a pas d'effet sur la durée moyenne d'un rhume. . .
- Les données issues de notre échantillon nous donnent-elles des preuves contre l'hypothèse nulle ?
- Calcul de la p-value, la probabilité d'observer une différence au moins aussi importante que celle observée dans l'échantillon, si l'hypothèse nulle est vraie.

Exemple : sexe et salaire aux États-Unis[8]

Sexe	Effectifs	Salaire moyen (USD)	Écart-type
Masculin (1)	$n_1 = 120$	$\bar{x}_1 = 40\,000$	$s_1 = 35\,000$
Féminin (0)	$n_0 = 90$	$\bar{x}_0 = 30\,000$	$s_0 = 30\,000$

- Différence moyenne : $\Delta = \bar{x}_1 - \bar{x}_0 = 10\,000$ USD.
- Écart-type de la différence : $s_\Delta = \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}} = 4\,495$ USD.

$\Rightarrow$ Ces données nous-fournissent-elles des preuves contre l'hypothèse nulle selon laquelle il n'existerait pas d'inégalités salariales entre hommes et femmes ?.

[8] Données fictives.

Test z

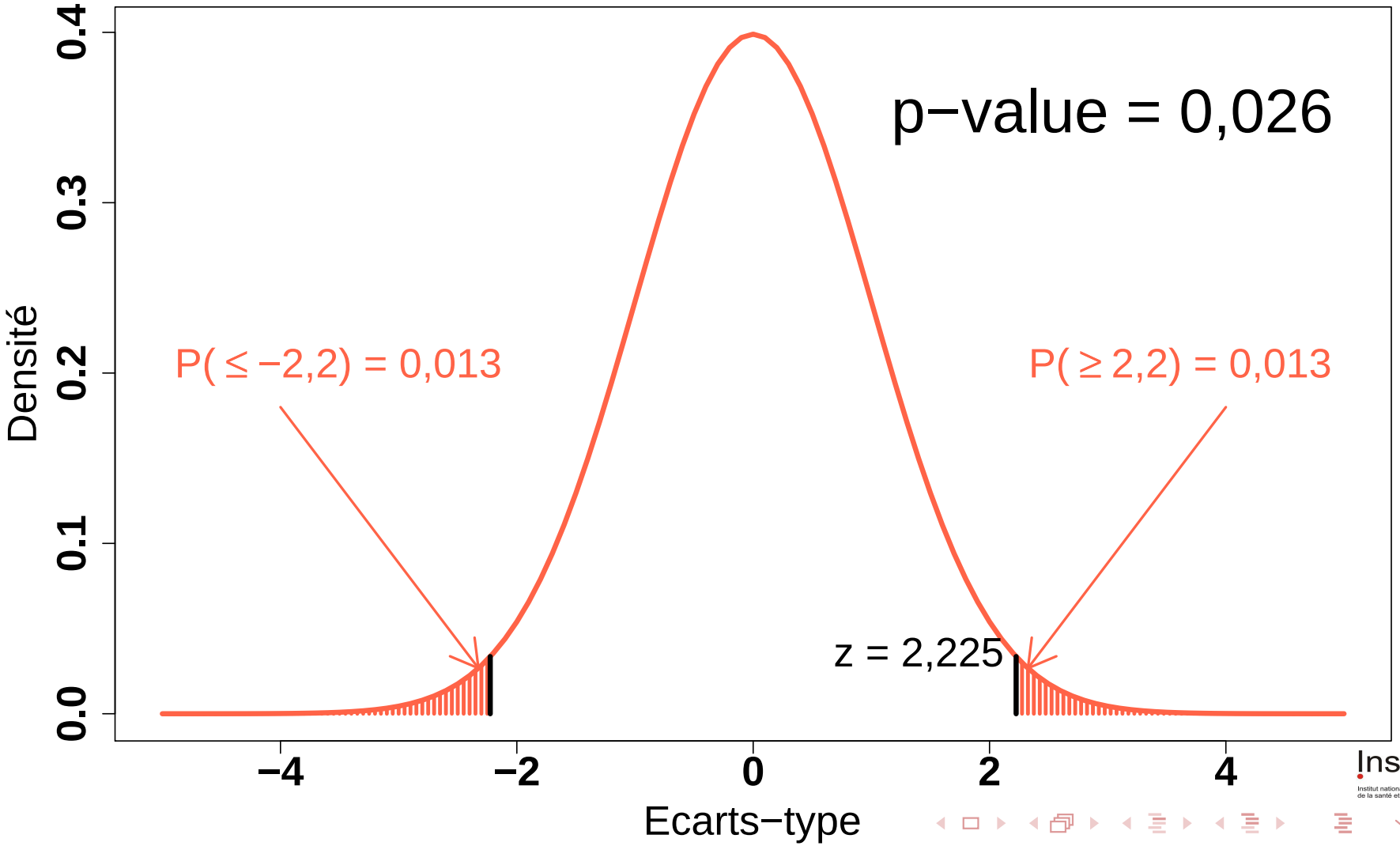
La statistique de test z est :

$$\frac{\Delta}{s_\Delta}$$

z est distribuée selon une loi normale $N(0,1)$.
Dans l'exemple :

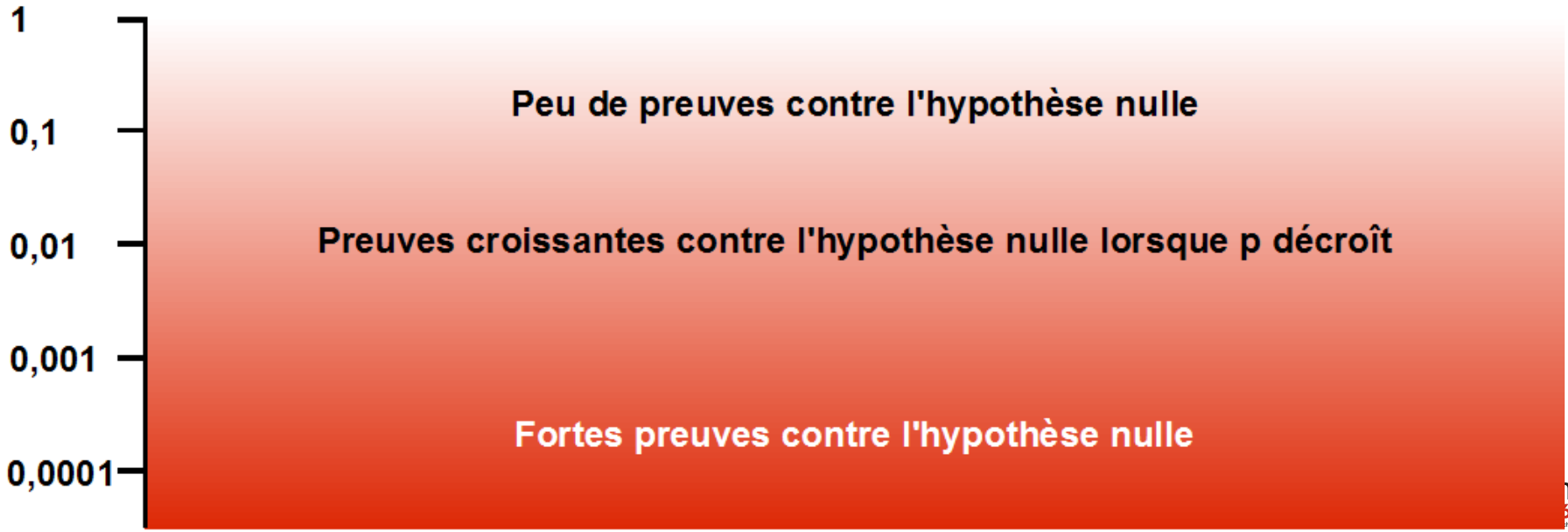
- $z = \frac{10\,000}{4\,495} = 2{,}225$.
- On utilise z pour estimer la p-value, la probabilité d'obtenir une différence d'au moins 2,225, en valeur absolue.

p-value


p−value = 0,026
P(≤ −2,2) = 0,013
P(≥ 2,2) = 0,013
z = 2,225
Densité
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
Ecarts−type

Interprétation Fishérienne de la p-value

Il s'agit d'une interprétation *informelle* : plus la p-value est petite, plus on a de preuves contre H_0.
L'hypothèse nulle n'est jamais prouvée ou établie, mais est possiblement réfutée.
Proposition : considérer les preuves contre H_0 suffisantes dès que $p < 0{,}05$.



Critique de l'approche Fisherienne par Neyman-Pearson

Pour Neyman-Pearson, l'approche Fishérienne viole le principe fréquentiste.

Le Principe Fréquentiste

Dans l'usage répété d'une procédure statistique, la véritable erreur commise en moyenne ne doit pas être supérieure à l'erreur moyenne rapportée.

Ces statisticiens vont donc se placer dans une perspective décisionnelle, et introduisent l'idée d'un choix entre l'hypothèse nulle H_0 et une *hypothèse alternative* H_1.

Le test d'hypothèses selon Neyman-Pearson

	La vérité	
L'expérience	H_0 est vraie	H_0 est fausse
Rejet de H_0	Erreur de type I (α)	Puissance $1-\beta$
Conservation de H_0		Erreur de type II (β)

En fixant à l'avance le risque d'erreur de première espèce (α) et le risque d'erreur de seconde espèce (β), on limite le nombre d'erreurs faites sur le long terme.

La philosophie de Neyman-Pearson

- Approche décisionnelle vs. approche subjective (Fisher),
- Interdiction d'interpréter la p-value !
- Bien qu'il soit impossible de savoir si une hypothèse donnée est vraie ou fausse, on dispose de règles de décision qui permettent d'affirmer que, sur le long terme, on ne se trompera pas souvent.

Critique Fishérienne de l'approche de Neyman-Pearson

- Ne décrit pas l'intensité de la preuve : que z soit égal à 2,0 (p=0,046) ou à 3,2 (p=0,001), on ne rapportera que le risque α.
- Nécessité de spécifier une hypothèse alternative H_1.
- Le calcul de la puissance peut être très compliqué et dépend de paramètres inconnus.

Les critiques de Jeffreys

- Critique de l'approche Fishérienne : La p-value est absurde : une hypothèse nulle vraie peut être rejettée parce qu'elle n'a pas prédit des résultats qui n'ont pas été observés (puisqu'on calcule la probabilité d'observer les données *ou pire*).
- Critique de l'approche Neyman-Pearson : ne décrit pas les variations dans le degré de preuve apporté par les données (même critique que Fisher).

L'approche Bayésienne de Jeffreys

- On garde l'idée d'une hypothèse alternative,
- On accole des probabilités a priori que H_0 et H_1 soient vraies (0,5 pour les deux),
- On définit le *Facteur de Bayes*, un rapport de vraisemblance : $BF(y) = \frac{P(y|\theta_0)}{P(y|\theta_1)}$
- Partie décisionnelle : on rejette H_0 si $BF(y) \leq 1$.
- Quantification de la preuve : à partir du Facteur de Bayes et des probabilités a priori de H_0 et H_1 (égales à 0,5), on calcule la probabilité a posteriori de H_0 comme : $P(H_0|y) = \frac{BF(y)}{1+BF(y)}$.

Les critiques de Fisher, Neyman et Pearson

- Critique de principe : c'est Bayésien.
- Critique spécifique : difficulté du choix des a prioris.

Exemple de synthèse[9]

Imaginons un tirage de n=10 individus $x_1, \ldots, x_{10}$ i.i.d $N(\theta, \sigma^2)$ où σ est connu.
On s'intéresse à l'hypothèse nulle $H_0 : \theta = 0$.
On procède au calcul de $z = \sqrt{n}\frac{\bar{x}}{\sigma}$. Imaginons deux valeurs différentes.

Approche	$z = 2{,}3$	$z = 2{,}9$
Fisher	p = 0,0215	p = 0,0037
Neyman-Pearson	$H_1 : \theta \neq 0$ $\alpha = 0{,}05$, $\beta = ?$ H_0 rejetée au risque 5 %.	
Jeffreys	$H_1 : \theta \neq 0$ $P(H_0) = P(H_1) = 0{,}5$ $P(H_0 \mid x_1, \ldots x_{10}) = 0{,}28$	$P(H_0 \mid x_1, \ldots x_{10}) = 0{,}11$

[9] Tiré de Berger JO. Could Fisher, Jeffreys and Neyman Have Agreed on Testing ? Statistical Science. 2003 ;18(1) :1-32.

Théorème de Bayes et test d'hypothèses

Soient H_0 et H_1 deux hypothèses mutuellement exclusives et exhaustives, et $P(H_0)$ et $P(H_1)$ leurs probabilités a priori.
Les probabilités d'observer les données y sous l'hypothèse (vraisemblance) sont notées $P(y|H_0)$ et $P(y|H_1)$.
La probabilité de l'hypothèse a posteriori H_0 est donc égale à :

$$P(H_0|y) = \frac{P(y|H_0)}{P(y)} P(H_0)$$

Avec $P(y) = P(H_0)\,P(y|H_0) + P(H_1)\,P(y|H_1)$ (probabilité globale de survenue des données).

Formulation sous forme de cotes

Puisque H_1 est l'hypothèse contraire de H_0, $P(H_1) = 1 - P(H_0)$ et $P(y|H_1) = 1 - P(H_0)$.
On peut donc exprimer le théorème de Bayes sous forme de rapport :

$$\frac{P(H_0|y)}{P(H_1|y)} = \frac{P(y|H_0)}{P(y|H_1)} * \frac{P(H_0)}{P(H_1)}$$

En d'autres termes :
cote a posteriori = rapport de vraisemblance * cote a priori.

Le principe de vraisemblance

- Principe de vraisemblance : les vraisemblances de y sous les hypothèses contiennent toute l'information qui peut être extraite des données.
- La statistique Bayésienne obéit au principe de vraisemblance.
- La statistique fréquentiste n'obéit pas au principe de vraisemblance (cf. p-values).

Le facteur de Bayes

Sous le principe de vraisemblance, le facteur de Bayes permet de choisir entre deux modèles :

$$BF = \frac{P(y|H_0)}{P(y|H_1)}$$

Jeffreys proposait les règles suivantes :

BF	**Force de la preuve en faveur de H_0 et contre H_1**
> 100	Décisive
32 - 100	Très forte
10 - 32	Forte
3,2 - 10	Substantielle
1 - 3,2	Ne méritant pas plus que sa mention

Lancer de pièce

Reprenons l'exemple initial : on procède à $n = 10$ tirages d'une pièce de monnaie, on obtient $y = 7$ « face ».
Peut-on en conclure que la pièce est truquée ?

L'approche fréquentiste sera de poser $H_0 : p = 0{,}5$, $H_1 : p \neq 0{,}5$. p est ici la probabilité de tomber sur « face ». Si l'on rejette H_0, on va conclure que la pièce est truquée.
On calcule donc la p-value comme $p = P(y \geq 7) + P(y \leq 3)$:

```
> 2*pbinom(3,10,0.5)
```

```
[1] 0.3438
```

La p-value est ici de 0,34, elle est supérieure à 5 %, on ne peut rejetter H_0 et on ne peut donc conclure.

Lancer de pièce : approche Bayésienne intuitive I

Nous allons poser $P(H_0) = P(H_1) = 0{,}5$.
Calculer la vraisemblance sous H_0 ($p = 0{,}5$) ne pose aucun problème.
Sous H_1, il faut préciser la distribution a priori de p.
Nous allons postuler que seules deux valeurs sont possibles : $p = 0{,}6$ avec une probabilité a priori de 30 % et $p = 0{,}7$ avec une probabilité a priori de 20 %.

	Modèle m	a priori $P(m)$	vraisemblance $P(y\|m)$	loi jointe $P(y \cap m)$	a posteriori $P(m\|y)$
H_0	p=0,5	0,5	0,117	0,059	0,332
H_1	p=0,6	0,3	0,215	0,064	0,366
	p=0,7	0,2	0,267	0,053	0,302
			$P(y)$	0,176	

$$P(y \cap m) = P(m) * P(y|m),\ P(y) = \sum P(y \cap m),\ P(m|y) = P(y \cap m)/P(y).$$

Lancer de pièce : approche Bayésienne intuitive II

```
> p <- c(0.5,0.6,0.7)
> apriori <- c(0.5,0.3,0.2)
> vraisemblance <- dbinom(7,10,p)
> loi.jointe <- apriori*vraisemblance
> p.y <- sum(loi.jointe)
> aposteriori <- loi.jointe/p.y
```

On a donc $P(H_0|y) = 33\ \%$ et donc $P(H_1|y) = 66\ \%$. Avec les distributions a priori (simplistes) choisies, les données observées nous permettent de croire qu'il y a deux chances sur trois que la pièce soit truquée.

Approche Bayésienne plus formelle[10]

Soit $P(p)$ la distribution a priori de p. Généralement, on choisit une distribution symétrique autour de 0,5 : $P(p) = \mathrm{Beta}(a, a)$.
La probabilité a posteriori de H_0 devient :

$$P\left(H_0|y\right) = \frac{P\left(y|H_0\right)}{P\left(H_0\right)P\left(y|H_O\right) + P\left(H_1\right)m_1(y)}P\left(H_0\right)$$

Où $m_1(y)$ est la densité a posteriori de y :

$$m_1(y) = \frac{P(y|p)P(p)}{P(p|y)}$$

Où $P(p|y)$ est la distribution a posteriori de p, $\mathrm{Beta}(a+k, a+n-k)$.

[10] Albert J. Bayesian computation with R.

Application I

On va utiliser une loi a priori $\mathrm{Beta}(10, 10)$ pour p sous H_1, avec $P(H_0)$=50 %.

```
> n <- 10
> y <- 7
> a <- 10
> p <- 0.5
> m1 <- dbinom(y,n,p)*dbeta(p,a,a)/dbeta(p,a+y,a+n-y)
> (aposteriori <- dbinom(y,n,p)/(dbinom(y,n,p)+m1))

[1] 0.4826
```

Application II

La probabilité a posteriori pour que H_0 soit vraie est donc de 48 %.
La probabilité a posteriori que la pièce soit truquée est donc de 52 %.

$P(H_0|y)$ dépend fortment des choix des lois a priori, il est donc conseillé de faire une analyse de sensibilité sur celles-ci ($P(H_0)$, a) avant toute conclusion définitive.

Application sur un échantillon plus large

Imaginons que l'on ait triplé les tirages, on aurait donc $n = 30$ lancers, avec $y = 21$ « face ». La p-value devient :

```
> n <- 30
> y <- 21
> p <- 0.5
> 2*pbinom(n-y,n,p)
```

```
[1] 0.04277
```

La probabilité a posteriori que H_0 soit vraie devient :

```
> a <- 10
> m1 <- dbinom(y,n,p)*dbeta(p,a,a)/dbeta(p,a+y,a+n-y)
> (aposteriori <- dbinom(y,n,p)/(dbinom(y,n,p)+m1))
```

```
[1] 0.2653
```

Écart important entre les analyses !

Le (quasi) consensus actuel

L'usage (l'interprétation) de la p-value dans la recherche biomédicale actuelle pose de nombreux problèmes.

Critiques fréquentistes de la p-value

- La p-value viole le principe fréquentiste.
- La p-value exagère le degré de preuve contre H_0.
- Souvent, la p-value est rapportée dans une perspective Neyman-Pearsonienne en se focalisant sur le risque α mais en oubliant le risque β (et l'hypothèse alternative).

Les critiques de Sterne[11]

Hypothèses de départ :

- H_0 est fausse dans 10 % des cas (cf 300 facteurs de risque identifiés en 1985 pour les cardiopathies coronariennes !).
- La puissance $1 - \beta$ des essais publiés est en moyenne de 50 % (tailles d'échantillon trop petites).
- 1 000 études conduites.

	La vérité	
L'expérience	H_0 est vraie	H_0 est fausse
Rejet de H_0	Erreur de type I (α)	Puissance $1 - \beta$
Conservation de H_0		Erreur de type II (β)
	900	**100**

[11] Sterne JA, Smith GD. Sifting the evidence – what's wrong with significance tests ? BMJ 2001 ;322 :226-30.

Les critiques de Sterne[11]

Hypothèses de départ :

- H_0 est fausse dans 10 % des cas (cf 300 facteurs de risque identifiés en 1985 pour les cardiopathies coronariennes !).
- La puissance $1 - \beta$ des essais publiés est en moyenne de 50 % (tailles d'échantillon trop petites).
- 1 000 études conduites.

	La vérité	
L'expérience	H_0 est vraie	H_0 est fausse
Rejet de H_0	45	Puissance $1 - \beta$
Conservation de H_0	855	Erreur de type II (β)
	900	**100**

[11] Sterne JA, Smith GD. Sifting the evidence – what's wrong with significance tests ? BMJ 2001 ;322 :226-30.

Les critiques de Sterne[11]

Hypothèses de départ :

- H_0 est fausse dans 10 % des cas (cf 300 facteurs de risque identifiés en 1985 pour les cardiopathies coronariennes !).
- La puissance $1 - \beta$ des essais publiés est en moyenne de 50 % (tailles d'échantillon trop petites).
- 1 000 études conduites.

	La vérité	
L'expérience	H_0 est vraie	H_0 est fausse
Rejet de H_0	45	50
Conservation de H_0	855	50
	900	**100**

[11] Sterne JA, Smith GD. Sifting the evidence – what's wrong with significance tests ? BMJ 2001 ;322 :226-30.

Conclusions

	La vérité	
L'expérience	H_0 est vraie	H_0 est fausse
Rejet de H_0	45	50
Conservation de H_0	855	50
	900	**100**

- H_0 rejetée dans $45 + 50 = 95$ cas.
- Sur les hypothèses rejetées au risque $\alpha = 5$ %, **47 %** le sont à tort !

$\Rightarrow$ Près de la moitié des publications scientifiques sont des faux positifs.

Les intervalles de confiance

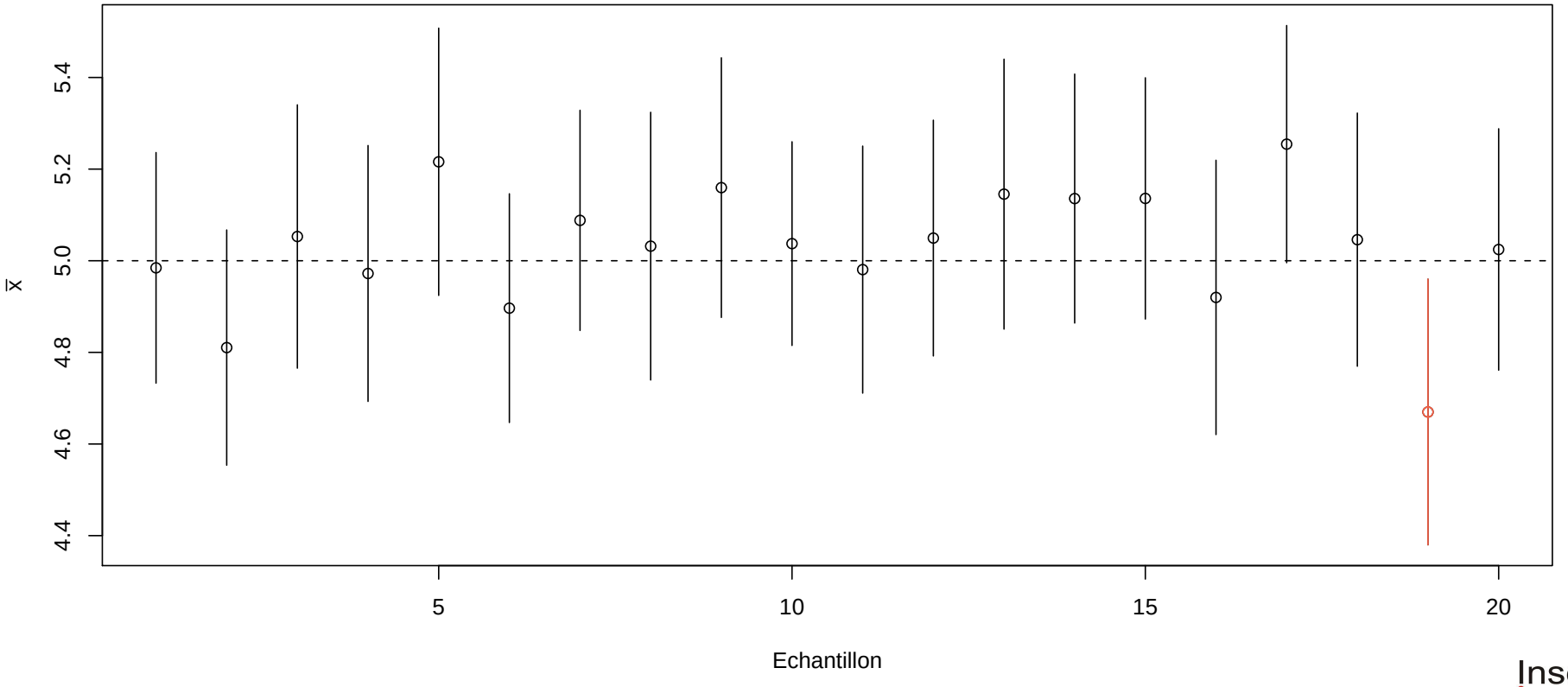
Introduits par Neyman : Si une distribution est normale, alors 95 % des observations seront comprises à $\pm$ 1,96σ de la moyenne μ.

Sur 95 % des échantillons, l'intervalle ($\bar{x} \pm 1{,}96s$) contient la moyenne μ (inconnue) de la population.

L'intervalle est un *intervalle de confiance*.

Illustration graphique

En moyenne, l'un des intervalles de confiance à 95 % de 20 échantillons ne contiendra pas la moyenne μ de la population :



De l'utilité des intervalles de confiance

Étude sur l'efficacité des unités neurovasculaires[12].

- Traitement : unité neurovasculaire vs. unité médicale.
- Patients : large-vessel infarcts vs. lacunar strokes.
- Conclusions : « Stroke units improve the outcome in patients with large-vessel infarcts but not in those with lacunar syndromes. »

Résultats pour mortalité ou institutionnalisation à 1 an :

- Large-vessel : OR=2,8, p=0,01, IC 95 %=(1,3-6,2).
- Lacunar : OR=4,9, p=0,06, IC 95 %=(0,9-25,0).

[12] Evans A et al. Randomized Controlled Study of Stroke Unit Care Versus Stroke Team Care in Different Stroke Subtypes. Stroke. 2002;33 :449-55.

Recommandations de Sterne et Smith

- Employer $\alpha = 0{,}01$ plutôt que 0,05,
- Toujours présenter un intervalle de confiance, à 90 %,
- Penser au risque β et le minimiser,
- Bannir le terme « statistiquement significatif »,
- Interpréter la p-value à la lumière du design de l'étude (biais, facteurs de confusion, analyses de sous-groupes, . . .) et des autres données disponibles.

Goodman et la probabilité de réplication[13]

Probabilité de réplication : probabilité d'obtenir un résultat significatif dans une seconde étude si la vraie valeur du paramètre est égale à celle observée dans la première étude.

p-value initiale	Probabilité de réplication
0,100	0,37
0,050	0,50
0,010	0,73
0,005	0,80
0,001	0,91

[13] Goodman SN. A comment on replication, P-values and evidence. Statistics in Medicine 1992 ;11(7) :875-9.

Les positions de Matthews[14]

1. Les analyses fréquentistes sont facilement mal interprétées,
2. Les méthodes fréquentistes sont arbitraires,
3. Les méthodes fréquentistes exagèrent la significativité,
4. Les méthodes fréquentistes font défaut là où elles sont les plus nécessaires,
5. Les méthodes fréquentistes sont inadaptées à la régulation.

[14] Matthews RAJ. Why should clinicians care about Bayesian methods ? Journal of Statistical Planing and Inference 2001 ;94 :43-58.

Les analyses fréquentistes sont mal interprétées

- On retrouve des erreurs dans la définition de la p-value et de l'intervalle de confiance même dans les manuels de statistique.
- p-value $= P(\geq y|H_0) \neq P(H_0|y)$.
- Question posée à 79 étudiants en statistique[15] : L'intervalle de confiance à 95 % de la proprotion est (0,33 - 0,43). Donc :
  1. La probabilité que la proportion soit comprise entre 0,33 et 0,43 est de 95 %.
  2. Si l'on constitue plusieurs échantillons et que l'on calcule plusieurs intervalles, environ 95 % d'entre eux contiendront la proportion.
  3. Nous sommes sûrs à 95 % que l'intervalle (0,33 - 0,43) contient la proportion.

[15] Albert J. Teaching Baye's rule : a data-oriented approach. 1997.

Les analyses fréquentistes sont mal interprétées

- On retrouve des erreurs dans la définition de la p-value et de l'intervalle de confiance même dans les manuels de statistique.
- p-value $= P(\geq y|H_0) \neq P(H_0|y)$.
- Question posée à 79 étudiants en statistique[15] : L'intervalle de confiance à 95 % de la proprotion est (0,33 - 0,43). Donc :
  1. La probabilité que la proportion soit comprise entre 0,33 et 0,43 est de 95 %.
  2. Si l'on constitue plusieurs échantillons et que l'on calcule plusieurs intervalles, environ 95 % d'entre eux contiendront la proportion.
  3. Nous sommes sûrs à 95 % que l'intervalle (0,33 - 0,43) contient la proportion.
     - Réponse 1 : 25 % des étudiants (Interprétation Bayésienne).

[15] Albert J. Teaching Baye's rule : a data-oriented approach. 1997.

Les analyses fréquentistes sont mal interprétées

- On retrouve des erreurs dans la définition de la p-value et de l'intervalle de confiance même dans les manuels de statistique.
- p-value $= P(\geq y|H_0) \neq P(H_0|y)$.
- Question posée à 79 étudiants en statistique[15] : L'intervalle de confiance à 95 % de la proprotion est (0,33 - 0,43). Donc :
  1. La probabilité que la proportion soit comprise entre 0,33 et 0,43 est de 95 %.
  2. Si l'on constitue plusieurs échantillons et que l'on calcule plusieurs intervalles, environ 95 % d'entre eux contiendront la proportion.
  3. Nous sommes sûrs à 95 % que l'intervalle (0,33 - 0,43) contient la proportion.
     - Réponse 1 : 25 % des étudiants (Interprétation Bayésienne).
     - Réponse 3 : 39 % des étudiants.

[15] Albert J. Teaching Baye's rule : a data-oriented approach. 1997.

Les analyses fréquentistes sont mal interprétées

- On retrouve des erreurs dans la définition de la p-value et de l'intervalle de confiance même dans les manuels de statistique.
- p-value $= P(\geq y|H_0) \neq P(H_0|y)$.
- Question posée à 79 étudiants en statistique[15] : L'intervalle de confiance à 95 % de la proprotion est (0,33 - 0,43). Donc :
  1. La probabilité que la proportion soit comprise entre 0,33 et 0,43 est de 95 %.
  2. Si l'on constitue plusieurs échantillons et que l'on calcule plusieurs intervalles, environ 95 % d'entre eux contiendront la proportion.
  3. Nous sommes sûrs à 95 % que l'intervalle (0,33 - 0,43) contient la proportion.

  - Réponse 2 : 35 % des étudiants !

[15] Albert J. Teaching Baye's rule : a data-oriented approach. 1997.

Les méthodes fréquentistes sont arbitraires

- « By some mistake, $\alpha = 0{,}05$ is often used as a magic cutoff »[16]
- Tests unilatéraux ou bilatéraux ?
- Pour les mêmes données, multiples tests possibles : 13 façons différentes de calculer une p-value pour un tableau de contingence 2*2 !

[16] Harrell FE. Practical Bayesian analysis from a former frequentist. 2000.

La significativité est exagérée

Théorème de Bayes :

$$P(H_0|y) = \left(1 + \frac{1 - P(H_0)}{P(H_0)\, BF}\right)^{-1}$$

Si la statistique de test $z > 1$, le facteur de Bayes est borné par :

$$BF \geq z \exp\left(\frac{1 - z^2}{2}\right)$$

Pour $z = 1{,}96$, $p = 0{,}05$ et $BF \geq 0{,}47$.
$\Rightarrow P(H_0|y)$ n'est égal à 5 % que si $P(H_0) \leq 0{,}1$.
Il faut être a priori convaincu que H_0 a 90 % de chances d'être fausse pour considérer a posteriori que H_0 a 95 % de chances d'être fausse.

Échec maximal sur les questions importantes

À z constant, l'écart entre les p-values et $P(H_0|y)$ augmente avec $P(H_0)$.

$\Rightarrow$ Plus la théorie testée est farfelue, plus la p-value est trompeuse !

Exemple : mon opinion a priori sur l'efficacité de l'homéopathie est qu'elle a 1 % de chances d'être efficace : $P(H_0) = 99$ %.
Un essai clinique randomisé aboutit à une p-value de 5 %.
Mon opinion a posteriori sur la non-efficacité de l'homéopathie est donc $P(H_0|y) \geq 96$ %.

Non-adaptation à la régulation

Accepter un nouveau médicament sur le marché si $p = 5\ \%$ implique qu'on estime a priori qu'il y avait 90 % de chances que le médicament soit efficace. Un a priori neutre $P(H_0) = 50\ \%$ aurait été plus adapté !

Certaines agences demandent deux essais indépendants significatifs pour accepter le médicament, en s'imaginant qu'elles ne laissent ainsi passer que $\frac{1}{20}^2 = \frac{1}{400}$ médicaments inefficaces. Avec un a priori neutre, elles en laissent passer *au moins* $\frac{1}{14}$.

L'hypothèse H_0 d'efficacité nulle est irréaliste : pour la décision médicale, on préférerait tester si le médicament a au moins un effet cliniquement significatif !

Le point de vue économique[17]

Pour un économiste, la rationalité voudrait que l'on adopte le traitement qui ait le plus de chances d'être efficient (i.e. coût-utile).

La méthode fréquentiste, entre ne rien faire et adopter un médicament avec plus d'une chance sur deux d'être efficient ($P(H_1|y) > 50\ \%$), amène à :

- Adopter le médicament si $p < 5\ \%$ pour l'efficacité,
- Continuer à ne rien faire si $p > 5\ \%$.

[17] Claxton K. The irrelevance of inference : a decision-making approach to the stochastic evaluation of health care technologies. J Health Econ. 1999 ;18(3) :341-64 et Claxton K et al. A rational framework for decision making by the national institute for clinical excellence (NICE). Lancet. 2002 ;360(9334) :711-5.

Sélection fréquentiste de variables et p-value

La p-value exagère le degré de preuve $\Rightarrow$ sélectionner des variables dans un modèle de régression multivarié sur sa base aboutit à des modèles surparamétrés.

Si la taille d'échantillon est importante, (presque) tout est significatif.

Solution : Le facteur de Bayes.

Approximation par le BIC

Bayesian Information Criterion :

$$BIC_k = D_k - ddl_k \log n$$

le BIC d'un modèle k est défini comme la déviance pénalisée par la complexité du modèle (nombre de degrés de liberté) et la taille de l'échantillon.

Il existe une relation intéressante entre BIC et facteur de Bayes lorsque l'on compare un modèle j à un modèle k :

$$2 \log BF \approx BIC_k - BIC_j$$

On peut donc appliquer les règles d'interprétation du facteur de Bayes à une différence de BIC.

Relation entre ΔBIC, n, et p-value[18]

Raftery a construit une table de correspondance entre p-value, taille d'échantillon, et force de preuve (définie par le différentiel de BIC) lorsque l'on doit décider s'il faut ou non ajouter un paramètre à un modèle :

		Taille d'échantillon					
Degré de preuve	ΔBIC	30	50	100	1 000	10 000	100 000
Faible	0	0,076	0,053	0,032	0,009	0,002	0,0007
Positif	2	0,028	0,019	0,010	0,003	0,0008	0,0002
Fort	6	0,005	0,003	0,001	0,0003	0,0001	0,00003
Très Fort	10	0,001	0,0005	0,0001	0,00004	0,00001	0,000004

[18] Raftery AE. Bayesian Model Selection in Social Research. Sociological Methodology. 1995 ;25 :111-63.

Comparaison des approches

Statistique Fréquentiste

- La probabilité est une propriété physique : « Probability lives in the world ».
- Décisions = p-value.
- Incertitude = intervalle de confiance.

Statistique Bayésienne

- La probabilité est un concept : « Probability lives in the mind ».
- Décisions = distribution a posteriori.
- Incertitude = l'intervalle de crédibilité.

Inserm U669 Institut national de la santé et de la recherche médicale

Un autre monde est possible

- 1925 : *Statistical Methods for Research Workers* de Fisher : introduction de la p-value.
- Années 1960 : Leonard Savage émet les premières mises en garde sur les dangers de la p-value.
- Années 1980 : James Berger : des résultats peuvent être statistiquement significatifs alors même que les données montrent très peu de preuves sur l'existence d'un effet.
- 1986-1988 : Kenneth Rothman, éditeur de l'*American Journal of Public Health*, décide de ne plus accepter d'articles fondant leurs résultats sur des p-values.
- 1995 : La *British Psychological Society* mandate une commission pour réfléchir à l'exclusion des p-values de ses journaux. Aucune décision n'est prise.

Inserm U669 Institut national de la santé et de la recherche médicale

Un autre monde est possible

« Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding. It is time to pull the plug. »[19]

Ronald Fisher a offert aux scientifiques une machine mathématique pour transformer des balivernes en découvertes capitales, et des coups de bol en financements. Il est temps de la débrancher.

[19] Robert Matthews. Bayesian Critique of Statistics in Health : The Great Health Hoax.