Le paradoxe de Jeffreys-Lindley : pierre dans le jardin des fréquentistes ou épine dans le pied des Bayésiens?

Hommage à Dennis Lindley

- Dennis Lindley (25/7/1923-14/12/2013)
- Mathématicien de formation (Trinity College, Cambridge)
- Grand mentor du renouveau et de l'essor de la statistique bayésienne notamment à « University College » de Londres (1967-1977) où il a créé, non sans difficulté, une équipe dédiée à la stat bayésienne (PhD Adrian Smith)
- « It was though a Jehovah's witness had been elected Pope »
- Parti à la retraite en 1977 à 54 ans pour devenir « Itinerant scholar »
- 2002: Médaille Guy d'or de la Royal Statistical Society
- Discuteur de nombreux articles: Schafer (1982), Aitkin (1991), O'Hagan (1995), Berger, Boukai & Wang (1997), Bernardo (1999)
- A statistical paradox, 1957, Biometrika
- Introduction to Probability & Statistics, 2 vol, 1965, Cambridge UP
- Bayesian Statistics: a review, 1972
- The present position in Bayesian statistics, 1990, Stat Science
- The Philosophy of Science, 2000, JRSS, D

Sommaire

- Introduction
- Position du pb: Ex de Stone
- Rappels sur P-values & Facteur de Bayes
  - ✓ Définitions
  - ✓ Cas discret, normal, uniforme
- Le paradoxe
  - ✓ Effet de n (conditionnement, arrêt optionnel, consistance)
  - ✓ Relation entre BIC et FB
  - ✓ Effet de l'a priori
  - ✓ Légitimité d'une hypothèse nulle
- Mesures d'accommodement
  - ✓ Calibrage des P-values
  - ✓ Choix de l'a priori
- In fine
  - ✓ Retour à l'ex de Stone
  - ✓ Autres approches
  - ✓ Test unilatéral d'hypothèses composites
  - ✓ Discussion & références

Introduction

$$AIC = -2L\left(\hat{\theta}_{ML}\right) + 2p$$

$$BIC = -2L\left(\hat{\theta}_{ML}\right) + p\boxed{\log(n)}$$

La pénalité augmente avec p: OK

Pourquoi la pénalité augmente-t-elle avec n?

Difficile à expliquer intuitivement.

Côté énervant mais salutaire des paradoxes qui révèlent notre méconnaissance ou incompréhension des fondements

Exemple de Stone (1997)

- Exp de physique basée sur n collisions indépendantes produisant 2 types de particules A et B
- Prédiction de la théorie Pr(A)= θ =1/5
- n=527 135 collisions
- y=106 298 de type A
- y-0.2n=+871
- Estimation de θ : 0.2017±0.0005
- IC 95% (0.2006-0.2027)

Ex de Stone/ approximation de la binomiale sous H0

Approche classique asymptotique:

$$\text{Sous } H_0 : Z = \frac{Y - n/5}{0.4\sqrt{n}} \rightarrow \mathcal{N}(0,1)$$

$$\text{Ici } z_{obs} = \frac{527135\text{-}106298}{\sqrt{104\,490\,000}} = \frac{871}{290.4162} = \boxed{2.999}$$

$$\Pr\left[N(0,1) > 2.999\right] = 0.00135$$

P-value (bilatérale) = $\boxed{0.0027}$

L'expérimentateur se réjouit: il a mis en brèche la prédiction du théoricien.

Ex de Stone/rapport de vraisemblance

Approche rapport de vraisemblance: $\boxed{D_{01} = -2L_0 + 2\mathrm{Max}L_1}$

Soit $p = y/n$ fréquence de succès à l'issue de n épreuves

$\boxed{L_0 = \log f(p \mid \theta = \theta_0)}$ $\boxed{L_1 = \log f(p \mid \theta)}$

$$-2L_0 = \log 2\pi - \log n + \log 0.16 + 2.5n\left(p - 0.2\right)^2$$

$$-2L_1 = \log 2\pi - \log n + \log \theta(1-\theta) + n\left(p - \theta\right)^2 / \theta(1-\theta)$$

$$-2\mathrm{Max}L_1 = \log 2\pi - \log n + \ln 0.16$$

$$D_{01} = -2L_0 - (-2\mathrm{Max}L_1) = 2.5n\left(p - 0.2\right)^2 \Rightarrow \boxed{D_{01} = z_{obs}^2}$$

Sous H_0 $Z^2 \sim \chi_1^2$ et $\Pr\left(\chi_1^2 \geq 9\right) = 0.0027$

Test d'hypothèses

Ne pas confondre test d'hypothèses et test de signification

Berger (2003), Hubbard & Bayarri (2003)

Neyman - Pearson: Test d'hypothèses implique deux hypothèses:

une hypothèse nulle H_0 opposée à une hypothèse alternative H_1

Le test basé est défini par une règle de décision :

rejet de $H_0 = I\left(T\left(y\right) \in A\right)$ en faveur d'une alternative H_1

ensemble de rejet $A = \left\{T\left(y\right) \geq t_{1-\alpha}; \alpha = \Pr\left(T\left(y\right) \geq t_{1-\alpha} \mid H_0\right)\right\}$

α : taux d'erreur de 1ère espèce (rejeter H_0 si H_0 vraie)

("size of the test") prédéterminé avant tout examen des données

β : taux d'erreur de 2ème espèce (garder H_0 si H_0 fausse) $\Pr\left(T\left(y\right) < t_{1-\alpha} \mid H_1\right)$

Rapport de vraisemblance pour $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$

Rejet H_0 si $f\left(y \mid \theta_1\right) / f\left(y \mid \theta_0\right) \geq c_\alpha$ constitue un test UMP (test le plus puissant à α fixé)

Ex: contrôle routinier de qualité, mais difficultés d'application en général

Test de signification

Fisher: Test de signification avec P-value (niveau de signification)

P-value= $\Pr\left(T\left(y\right) \geq t_{obs} \mid H_0\right)$ $t_{obs} = T\left(y_{obs}\right)$

$T\left(y\right) \mid H_0 = n\left(\overline{y} - \mu_0\right)^2 / \sigma^2 \sim \chi_1^2$ avec $\overline{y} = p = y/n$ $\mu_0 = \theta_0$ $\sigma^2 = \theta_0\left(1-\theta_0\right)$

Pb d'interprétation: "P value fallacy" (Goodman, 1999; Sellke et al, 2001)

La P-value n'est pas un taux d'erreur de 1ère espèce

La P-value ne donne pas la pté de l'hypothèse nulle

C'est plutôt une mesure de conformité (ou d'étrangeté "surprise index") des données au modèle de base (H_0)

Et même certains ont pu dire qu'il n'y a rien de fréquentiste dans la p-value!

"We essentially agree that there is nothing frequentist about a p value" (Casella & Berger,1987)

"a **handy one - to - one transformation** of the test statistics, one that allows **checking how far on the tail of the null distribution** the observed value of the test statistic is" Hubbard & Bayarri (2003, p 182)

Probabilité des hypothèses et FB

Le théoricien appelle à la rescousse un bayésien

Th de Bayes $\boxed{\frac{\Pr(H_0 \mid y)}{\Pr(H_1 \mid y)}} = \frac{\Pr(Y = y \mid H_0)}{\Pr(Y = y \mid H_1)} \times \frac{\Pr(H_0)}{\Pr(H_1)}$

Facteur de Bayes: $\boxed{B_{01} = \underbrace{\frac{\Pr(H_0 \mid Y = y)}{\Pr(H_1 \mid Y = y)}}_{\text{"posterior odds"}} / \underbrace{\frac{\Pr(H_0)}{\Pr(H_1)}}_{\text{"prior odds"}}}$

("K" de Jeffreys,1939), utilisé indépendamment par Turing à Blechtley Park (Good, 1979)

"Weight of evidence" défini comme $\log_{10} B_{01}$

Probabilité des hypothèses et théorie de la décision

Fonction de coût $\mathrm{L}\left(\theta, \underbrace{a_i}_{décision}\right) = \begin{cases} 0 \text{ si décision correcte: } \theta \in \Theta_i \\ K_i \text{ si décision incorrecte: } \theta \in \Theta_{j \neq i} \end{cases}$

K_0 : coût de choisir H_0 alors qu'elle est fausse (H_1 vraie)

K_1 : coût de choisir H_1 alors qu'elle est fausse (H_0 vraie)

Décision opt (de rejet de H_0) si $\underbrace{\mathbb{E}^{\pi(\theta|\mathbf{y})}\left[\mathrm{L}(\theta, a_1)\right]}_{K_1 \Pr(\mathrm{H}_0|y)} < \underbrace{\mathbb{E}^{\pi(\theta|\mathbf{y})}\left[\mathrm{L}(\theta, a_0)\right]}_{K_0 \Pr(\mathrm{H}_1|y)}$

$$\boxed{\frac{\Pr(\mathrm{H}_0 \mid y)}{\Pr(\mathrm{H}_1 \mid y)} < \frac{K_0}{K_1}} \text{ ou } \mathrm{B}_{01} < \frac{K_0}{K_1} / \frac{\Pr(\mathrm{H}_0)}{\Pr(\mathrm{H}_1)} \text{ (Berger,1985)}$$

Exemple de Stone/Borne inf du FB

Il fait la transition avec le fréquentiste en calculant un facteur de Bayes (rudimentaire) minimum

$$B_{01,\min} = \frac{L(\theta_0)}{L(\hat{\theta}_{ML})} = \frac{L_0}{MaxL_1} = \exp\left(-½ z_0^2\right) = 0.223$$

$$\Pr_{\min}(H_0 \mid y) = \frac{\rho_0 B_{01}}{\rho_0 B_{01} + (1-\rho_0)} = \frac{1}{1 + \frac{(1-\rho_0)}{\rho_0} B_{01}^{-1}}$$

Pour $\rho_0 = \Pr(H_0) = 1/2$

$$\Pr_{\min}(H_0 \mid y) = \frac{1}{1 + (1/B_{01,\min})} = \boxed{0.18}$$

Rappel: P-value=0.0027

Les preuves expérimentales à l'encontre de la théorie s'avèrent moins nettes que précédemment

Vraisemblances sous H0 et H1

$$\Pr(Y = y \mid \mathrm{H}_0) = C_n^y \left(\theta_0\right)^y \left(1 - \theta_0\right)^{n-y}$$

$$\Pr(Y = y \mid \mathrm{H}_1) = \int_0^1 \Pr(Y = y \mid \theta) g\left(\theta\right) d\pi$$

On suppose $\theta \sim \mathrm{Beta}\left(\alpha, \beta\right)$

$$\Pr(Y = y \mid \mathrm{H}_1) = C_n^y B\left(y + \alpha, n - y + \beta\right) / B\left(\alpha, \beta\right)$$

où $B\left(u, v\right) = \int_0^1 t^{u-1} \left(1 - t\right)^{v-1} dt = \Gamma\left(u\right)\Gamma\left(v\right) / \Gamma\left(u + v\right)$

$$\mathrm{B}_{01} = \frac{B\left(\alpha, \beta\right)}{B\left(y + \alpha, n - y + \beta\right)} \left(\theta_0\right)^y \left(1 - \theta_0\right)^{n-y}$$

Ex de Stone/A priori uniforme

Dans le but de le convaincre il choisit un a priori uniforme sur (0,1) à l'instar des pères fondateurs

(**250ème anniversaire du fameux théorème dans 4 jours**)

Avec $\pi \sim \mathrm{U}(0,1)$ $\ln \mathrm{B}_{01} = 2.094 \Rightarrow \mathrm{B}_{01} = \mathbf{8.115}$

$$\Rightarrow \boxed{\Pr(H_0 \mid y) = \frac{B_{01}}{1+B_{01}} = \mathbf{0.89}}$$

Rappel: P-value **= 0.0027**

Le théoricien se dit que c'est presque trop beau pour être vrai.

Cas normal: expression de BF

$$\bar{y} \mid H_0 \sim \mathcal{N}\left(\mu_0, \sigma^2 / n\right)$$

$$f\left(\bar{y} \mid H_1\right) = \int_{-\infty}^{+\infty} f\left(\bar{y} \mid \mu, \sigma^2\right) \pi(\mu)\, d\mu$$

Non défini pour un a priori plat impropre $\pi(\mu)$=Cste

$$\mu \sim \mathcal{N}\left(\mu_0, \tau^2\right) \quad \bar{y} \mid H_1 \sim \mathcal{N}\left(\mu_0, \tau^2 + \sigma^2 / n\right)$$

$$B_{01} = \frac{f\left(\bar{y} \mid H_0\right)}{f\left(\bar{y} \mid H_1\right)} \qquad \boxed{B_{01} = \sqrt{1 + n/\lambda}\, \exp\left(-\frac{1}{2} \frac{z^2}{1 + \lambda / n} \right)}$$

$$z^2 = \frac{n\left(\bar{y} - \mu_0\right)^2}{\sigma^2} \qquad \tau^2 = \sigma^2 / \lambda$$

Calibration du facteur de Bayes

Calibration du facteur de Bayes selon Jeffreys: « measure of evidence against the null »

$K = B_{01}$	$\Pr(H_0 \mid \mathbf{y})^*$	Deciban (dB)	Deviance (ΔD)	Strength of evidence against the null
>1	$>1/2$	>0		0-Null hypothesis supported
1 à $10^{-1/2}$	0.24 à 0.50	0 à -5	0 à 2.3	1-Not worth than a bare mention
$10^{-1/2}$ à 10^{-1}	0.09 à 0.24	-5 à -10	2.3 à 4.6	2-Substantial
10^{-1} à $10^{-3/2}$	0.03 à 0.09	-10 à -15	4.6 à 6.9	3-Strong
$10^{-3/2}$ à 10^{-2}	0.01 à 0.03	-15 à -20	6.9 à 9.2	4-Very strong
$<10^{-2}$	<0.01	<-20	>9.2	5-Decisive

$10^{-1/2} = 0.32$, $10^{-3/2} = 0.0316$

$K = B_{01}$: « grade of decisiveness of evidence » for (1) against (0), Jeffreys (1961) Appendix B

$\Pr(H_0 \mid \mathbf{y})^*$ supposant $\Pr(H_0) = \Pr(H_1) = 1/2$

Deciban: $dB = 10\log_{10} B_{01}$ (Turing A, Good IJ, 1940)

Echelle de déviance: $\Delta D_{01} = D_0 - D_1 = -2\log L_0 / L_1$ (modèle réduit vs modèle complet)

Degré de preuve contre H0

"Rejet de H_0" si $B_{01} \leq b$

$$z^2 \geq \left[\log\left(1+n/\lambda\right)+b'\right]\left(1+\lambda/n\right) \quad \text{où } b' = -2\ln b$$

Si $n >> \lambda$ $\quad 1+n/\lambda \approx n/\lambda \quad 1+\lambda/n \approx 1$

$$\underbrace{z^2 - \log\left(n\right)}_{\Delta BIC_{0-1}} + \log\left(\lambda\right) \geq b'$$

Pour $b = 1/10$ et $\lambda = 1$ ("Unit Prior Information") $b' = 6.9$

Plus n est grand, plus il est difficile de rejeter H_0

FB pour n>>λ

$$B_{01} = \sqrt{1+n/\lambda}\exp\left(-½z^2/\left(1+\lambda/n\right)\right)$$

Si $n >> \lambda$ $\quad 1+n/\lambda \approx n/\lambda \quad 1+\lambda/n \approx 1$

λ tq $\tau^2 = \sigma^2/\lambda$ $\lambda =$ "Nbre d'observations implicites"

$$\boxed{B_{01} \simeq \sqrt{n/\lambda}\exp\left(-½z^2\right) = \sqrt{n/\lambda}B_{01}^{\min}}$$

Numérateur : $f\left(\bar{y} \mid H_0\right) = \dfrac{\sqrt{n}}{\sqrt{2\pi}\sigma}\exp\left(-½z^2\right)$

Dénominateur : $f\left(\bar{y} \mid H_1\right) = \int f\left(\bar{y} \mid \theta\right)\pi\left(\theta\right)d\theta \approx \boxed{\pi\left(\bar{y}\right)}$

-Rôle très important de l'a priori sur le paramètre

Pb en cas de conflit entre a priori et vraisemblance

(risque de sous pondérer les régions de forte vraisemblance $\bar{y}$)

Cas d'un a priori uniforme

Cas considéré par Lindley (1957) puis corrigé par Bartlett (1957)

$$\boxed{\mu \sim \mathrm{U}\left(\mu_0 - ½\delta, \mu_0 + ½\delta\right)} \Rightarrow g\left(\mu\right) = \frac{1}{\delta} I_{\left(\mu_0 - ½\delta \le \mu \le \mu_0 + ½\delta\right)}$$

$$f\left(\bar{y} \mid H_1\right) = \frac{1}{\delta} \int_{\mu_0 - ½\delta}^{\mu_0 + ½\delta} f\left(\bar{y} \mid \mu, \sigma^2\right) d\mu$$

$$f\left(\bar{y} \mid H_1\right) = \frac{1}{\delta} \int_{\mu_0 - ½\delta}^{\mu_0 + ½\delta} \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{n\left(\mu - \bar{y}\right)^2}{2\sigma^2} \right] d\mu$$

$$f\left(\bar{y} \mid H_1\right) = \frac{1}{\delta} \underbrace{\left[\Phi\left(\frac{\sqrt{n}\left(\left[½\delta - \left(\bar{y} - \mu_0\right)\right]\right)}{\sigma} \right) - \Phi\left(\frac{\sqrt{n}\left[-½\delta - \left(\bar{y} - \mu_0\right)\right]}{\sigma} \right) \right]}_{\mathcal{I}}$$

$$Si\ \mu_0 - ½\delta < \bar{y} < \mu_0 + ½\delta \Rightarrow \mathcal{I} \rightarrow_{n \to \infty} 1 \text{ et } \boxed{f\left(\bar{y} \mid H_1\right) \rightarrow 1/\delta}$$

Cas d'un a priori uniforme/suite

$$B_{01} = \frac{\boxed{\delta}}{\sigma}\sqrt{\frac{n}{2\pi}}\exp\left[-\frac{n\left(\bar{y}-\mu_0\right)^2}{2\sigma^2}\right]$$

Grde sensibilité à la valeur de δ

$$\boxed{-2\log B_{01} = z^2 - \log n + \log\left(\sigma^2/\delta^2\right) + \log\left(2\pi\right)}$$

Remarque: Si $\delta^2 = 2\pi\sigma^2$ $-2\log B_{01} = \Delta BIC_{0-1}$

Relation BIC et FB

BIC pris comme approximation de Laplace de

$-2\log\left[m(\mathbf{y})\right]$ où $m(\mathbf{y}) = \int f\left(\mathbf{y} \mid \boldsymbol{\theta}\right)\Pi\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}$

$$\text{BIC} = \boxed{-2L\left(\hat{\boldsymbol{\theta}}_{ML}\right) + p\log(n)} + \boxed{\text{O}(1)} + \text{O}\left(n^{-½}\right)$$

Kass & Raftery, 1995; LeBarbier & Mary-Huard , 2006

avec $\boxed{\text{O}(1) = \log\left|I_1\left(\hat{\boldsymbol{\theta}}_{ML}\right)\right| - p\log\left(2\pi\right) - 2\log\Pi\left(\hat{\boldsymbol{\theta}}_{ML}\right)}$ où $p = \dim(\theta)$

$L\left(\boldsymbol{\theta}\right) = \log\left[f\left(\mathbf{y} \mid \boldsymbol{\theta}\right)\right]$, $\mathbf{I}_1\left(\boldsymbol{\theta}\right) = -n^{-1}E_{\mathbf{y}|\boldsymbol{\theta}}\left(\dfrac{\partial^2 L\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)$ (info Fisher pour une obs)

"The O(1) error does suggest the approximation to be somewhat crude"

(ne décroit pas avec la taille de l'échantillon) (Raftery, 1995)

"BIC cannot replace a fully Bayesian model comparison with prior distributions chosen carefully for a specific problem" (Kuha, 2004)

Relation BIC et FB

Si on prend $\boldsymbol{\theta} \sim \mathcal{N}\left(\hat{\boldsymbol{\theta}}_{ML}, \frac{n}{\lambda}\mathbf{V}\left(\hat{\boldsymbol{\theta}}_{ML}\right)\right)$ où $\underbrace{\mathbf{V}\left(\hat{\boldsymbol{\theta}}_{ML}\right)}_{\text{Var d'échantillonage}} = \left[\mathbf{I}_1\left(\hat{\boldsymbol{\theta}}_{ML}\right)\right]^{-1} / n$

où λ représente un nombre d'observations implicites relatives à l'a priori

alors $\mathrm{BIC} = -2L\left(\hat{\boldsymbol{\theta}}_{ML}\right) + p\log(1 + n/\lambda)$ (Kuha, 2004)

Pour $n \gg \lambda$, on retrouve la formule $\mathrm{BIC} = -2L\left(\hat{\boldsymbol{\theta}}_{ML}\right) + p\log(n/\lambda)$

équivaut à $\Delta\mathrm{BIC}_{01} = \mathrm{BIC}_0 - \mathrm{BIC}_1 = -2\log B_{01} = z^2 - \log(n/\lambda)$ vu précédemment

avec $V\left(\hat{\boldsymbol{\theta}}_{ML}\right) = \sigma^2 / n$

Analogie avec le g-prior de Zellner du moins pour la variance $g\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$

(g identifiable à n/λ comme dans l'interprétation de Marin & Robert, 2007)

Paradoxe/effet de n

Pour $n >> \lambda \quad (\tau^2 = \sigma^2 / \lambda \Rightarrow \lambda = \sigma^2 / \tau^2)$

$$\boxed{\log B_{01} \simeq -½ z^2 + ½ \ln n - ½ \ln \lambda}$$

Que dit le paradoxe ? Conditionnellement aux observations

soit aussi à $|z|$ fixé (p-value fixée*) et pour λ fixé

* "a handy one-to-one transformation of the test statistics" Hubbard & Bayarri (2003)

$n \nearrow \Rightarrow B_{01} \nearrow$ "moins de preuves à l'encontre de H_0 "

$$\boxed{n \to +\infty \Rightarrow B_{01} \to +\infty \Rightarrow \text{acceptation systématique de } H_0}$$

"Diminishing significance of a fixed p-value when sample size is increased" (Good, 1988)

"As the sample size increases in testing precise hypotheses, a given p value provides less and less real evidence against the null" (Berger & Selke, 1987)

"5 % in to-day small sample does not mean the same as 5% in to-morrow large one." (Lindley, 1957)

Que dit Lindley?

Now suppose that the value of $\bar{x}$ is such that, on performing the usual significance test for the mean θ_0 of a normal distribution with known variance, the result is significant at the α percentage point. That is, $\bar{x} = \theta_0 + \lambda_\alpha \sigma/\sqrt{n}$, where λ_α is a number dependent on α only and can be found from tables of the normal distribution function. Inserting this value for $\bar{x}$ in (1) we have the following value for the posterior probability that $\theta = \theta_0$

$$\bar{c} = c e^{-\frac{1}{2}\lambda_\alpha^2} / \{c e^{-\frac{1}{2}\lambda_\alpha^2} + (1-c)\sigma\sqrt{(2\pi/n)}\}. \qquad (2)$$

(Note that $\bar{x} - \theta_0$ tends to zero as n increases so that $\bar{x}$ will lie well within the interval I for sufficiently large n.) From (2) we see that as $n \to \infty$, $\bar{c} \to 1$. It follows that whatever the value of c, a value n can be found, dependent on c and α such that

(i) $\bar{x}$ is significantly different from θ_0 at the α % level;

(ii) the posterior probability that $\theta = \theta_0$ is $(100-\alpha)$ %.

This is the paradox. The usual interpretation of the first result is that there is good reason to believe $\theta \neq \theta_0$; and of the second, that there is good reason to believe $\theta = \theta_0$. The two interpretations are in direct conflict, and the conflict may apparently be made even stronger by remarking that the $(100-\alpha)$ % confidence and fiducial intervals for θ just exclude $\theta = \theta_0$. With $\alpha = 5$ we are 95 % confident that $\theta \neq \theta_0$, but have 95 % belief that $\theta = \theta_0$.

Paradoxe/effet de n/illustration

Berger and Sellke: Testing a Point Null Hypothesis

Table 1. $Pr(H_0 \mid x)$ for Jeffreys-Type Prior

		n						
p	t	1	5	10	20	50	100	1,000
.10	1.645	.42	.44	.47	.56	.65	.72	.89
.05	1.960	.35	.33	.37	.42	.52	.60	.82
.01	2.576	.21	.13	.14	.16	.22	.27	.53
.001	3.291	.086	.026	.024	.026	.034	.045	.124

$\bar{y} \mid H_0 \sim \mathcal{N}\left(\mu_0, \sigma^2 / n\right)$

$\bar{y} \mid H_1 \sim \mathcal{N}\left(\mu, \sigma^2 / n\right)\ \mu \sim \mathcal{N}\left(\mu_0, \sigma^2\right)\ (\lambda{=}1)$

$\Pr\left(H_0\right) = \Pr\left(H_1\right) = 1/2 \quad \Pr\left(H_0 \mid \bar{y}\right) = \left(1 + B_{01}^{-1}\right)^{-1}$

Séquence aléatoire d'expériences H_0 et H_1 en proportions égales.

Dans l'ensemble de celles qui ont une valeur fixée de $|z|$ soit une p-value fixée α

$\Pr\left(H_0\right)$ a posteriori mesure la proportion limite de celles générées sous H_0

Relations BF, m/n0, P values/ Spiegelhalter et al, 2004

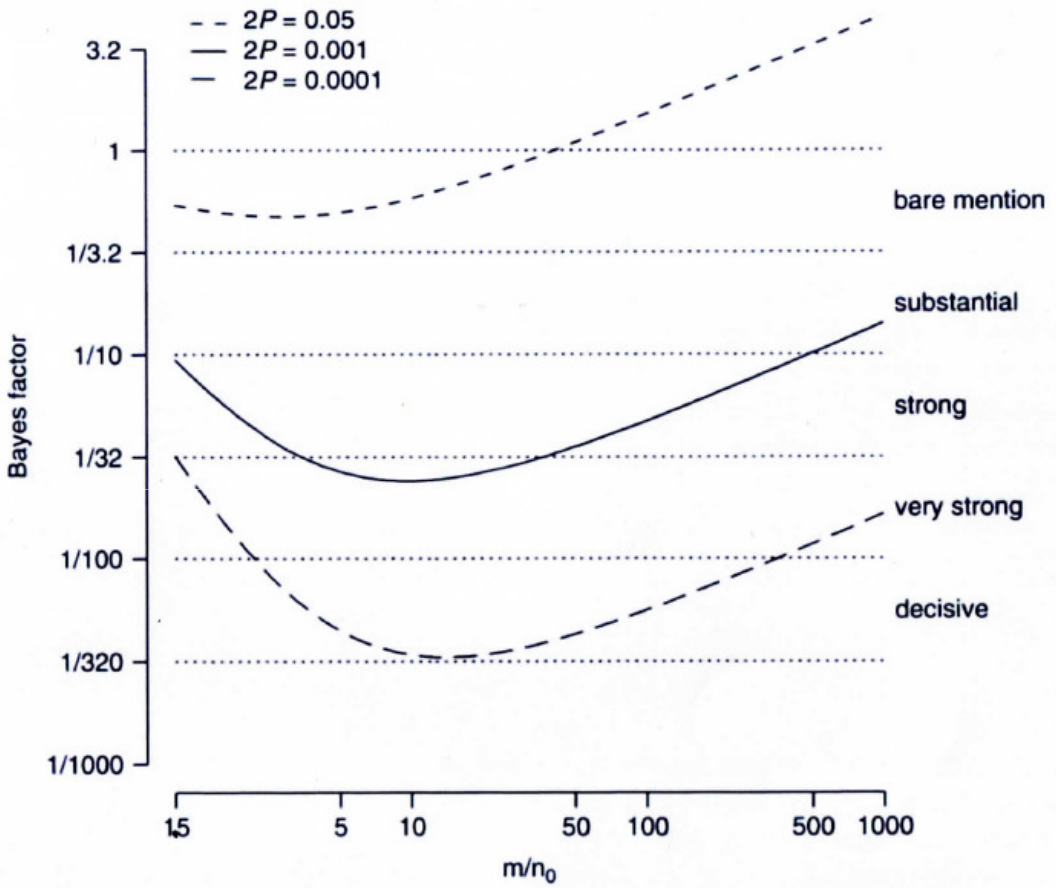


Figure 4.2 Bayes factors for composite normal hypotheses for fixed *P*-values and different m/n_0 ratios, *i.e.* ratio of observed to prior sample size, with areas delineated by Jeffreys' levels of evidence.

Paradoxe lié de Lindley-Scott vs Peto et al

Royall (1986). The effect of sample size of the meaning of significance test. The Am Stat, 40.313-315

Unfortunately, the proper interpretation of significance tests is not as simple as this practice implies. The sample size must also be considered.

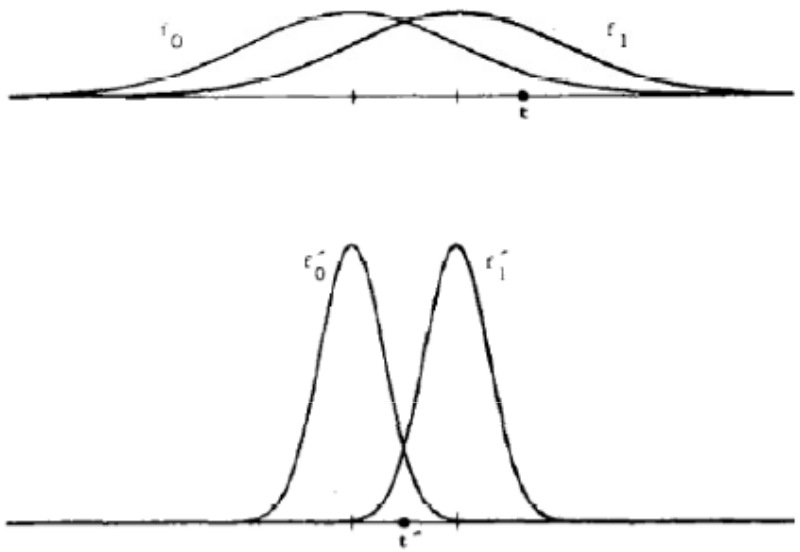
. . . the interpretation to be placed on the phrase 'significant at 5%' depends on the sample size: it is more indicative of the falsity of the null hypothesis with a small sample than with a large one. (Lindley and Scott 1984, p. 3)

Other authors also explained how the meaning of significance tests depends on the sample size. But they reached the opposite conclusion, as follows:

A given *P*-value in a large trial is usually stronger evidence that the treatments really differ than the same *P*-value in a small trial of the same treatments would be. (Peto et al. 1976, p. 593)

Paradoxe de Lindley-Scott vs Peto et al

Royall (2000) On the probability of observing misleading statistical evidence. The Am Stat, 45,760-768



I)Figure du haut : petit échantillon. II) Figure du bas : grand échantillon

Dans les deux cas, **même P-value** (« significance level ») $\Pr(T \geq t) = 0.05$ et **même différence** : $\mathbf{d = \mu_1 - \mu_0}$

En I) $\dfrac{\Pr(T \geq t \mid H_0)}{\Pr(T \geq t \mid H_1)} = \dfrac{0.05}{0.25} = \boxed{\dfrac{1}{5}}$ En II) $\dfrac{\Pr(T \geq t \mid H_0)}{\Pr(T \geq t \mid H_1)} = \dfrac{0.05}{0.95} = \boxed{\dfrac{1}{19}}$

Preuve plus forte contre H0 en II)(GE) qu'en I (PE)

En I) $\dfrac{\Pr(f(t) \mid H_0)}{\Pr(f(t) \mid H_1)} \approx \boxed{\dfrac{1}{3}}$ En II) $\dfrac{\Pr(f(t) \mid H_0)}{\Pr(f(t) \mid H_1)} = \boxed{1}$

Preuve plus forte contre H0 en I)(PE) qu'en II (GE)

Conditionnement par un intervalle

Conditionner par $|Z| = z_{1-\alpha/2}$ ≠ conditionner par $|Z| \geq z_{1-\alpha/2}$

$$A = \{Z : |Z| \geq z_{1-\alpha/2}\} \quad \tilde{B}_{01} = \frac{\Pr(A \mid H_0)}{\Pr(A \mid H_1)}$$

$$\bar{y} \mid H_0 \sim \mathcal{N}\left(\mu_0, \sigma^2 / n\right) \text{ et } \Pr(A \mid H_0) = \alpha$$

$$\Pr(A \mid H_1) = \int \underbrace{\Pr(A \mid H_1, \mu)}_{1-\beta(\alpha,\mu)} \pi(\mu)\, d\mu \quad \text{Puissance intégrée}$$

$\mu \sim \mathcal{N}\left(\mu_0, \sigma^2 / \lambda\right)$ ("skeptical prior", Spiegelhalter et al,1994)

$$\tilde{B}_{01} = \frac{\alpha}{2\left[1 - \Phi\left(\sqrt{\frac{\lambda}{n+\lambda}} z_{1-\alpha/2}\right)\right]}$$

$\lambda / n \to 0 \Rightarrow \tilde{B}_{01} \to \alpha$

$\Pr(H_0 \mid A) \to (1+\alpha^{-1})^{-1} \approx \alpha$ (si α petit) (B&S,1987; Dickey,1977)

$\lambda / n \to \infty \Rightarrow \tilde{B}_{01} \to 1, \ \Pr(H_0 \mid A) \to ½$

Cf aussi Hung et al (1997) pour la distribution de la P-value et DeGroot (1973) pour une interprétation de la p-value en terme de rapport de vraisemblance de H0 vs certaines classes de H_1

Conditionnement par un intervalle/Ex

FB et Probabilité de H0 par conditionnement sur l'intervalle de rejet

	$n/\lambda=4$		$n/\lambda=50$	
α	B(0,1)	Pr(H0/A)	B(0,1)	Pr(H0/A)
0.10	0.216	0.178	0.122	0.109
0.05	0.131	0.116	0.064	0.060
0.01	0.040	0.039	0.014	0.014
0.001	0.0071	0.0070	0.0015	0.0015

Distinction P=P0 vs P<=P0

- "In fact the paradox arises because the significance level argument is based on the area under a curve and the Bayesian argument is based on the ordinate of the curve" Lindley (1957, p 190)
- "What the use of P implies, therefore, is that a hypothesis that may be true can be rejected because it has not predicted observable results that have not occurred" Jeffreys (1961, 3rd ed, sect 7.2 p 385)

Ex de van der Pas (2010)

Table 3: Two different sampling distributions

x	0	1	2	3	4
$H_0(x)$	.75	.14	.04	.037	.033
$H_0'(x)$	.70	.25	.04	.005	.005

We use the test statistic $T(x) = x$. Suppose that $x = 2$ is observed. The corresponding p-values for both hypotheses are:

$$p_0 = P(x \geq 2|H_0) = 0.11 \qquad p_0' = P(x \geq 2|H_0') = 0.05.$$

Que dit Jeffreys?

+Que dit Jeffreys ? (1961, p247-248)

If, however, s is small, so that the exponent can take large values, and $f(\alpha)$ is continuous, the integral of (8) will be nearly $f(a)$, and

$$\frac{P(q \mid aH)}{P(q' \mid aH)} \doteqdot \frac{1}{\sqrt{(2\pi)s f(a)}} \exp\left(-\frac{a^2}{2s^2}\right). \qquad (10)$$

Ici $a = \overline{y}$, $\mu_0 = 0$, $s = \sigma / \sqrt{n}$,

Pour n grand $m(\overline{y} \mid H_1) = \pi(\overline{y}) = f(a)$ dans les notations de Jeffreys soit

$f(a) = \delta^{-1}$ dans le cas uniforme

$f(a) = = \frac{1}{\sqrt{2\pi\tau}} \exp(-½a^2 / \tau^2)$ dans le cas normal

Par le biais de δ ou de τ (ou λ), on peut diminuer à volonté le dénominateur et faire croître BF_{01} (noté K par Jeffreys) et donc $\Pr(H_0 \mid \overline{y})$.

"But it carries with it the consequence that **the critical value of $|a / s|$ increases with n, the increase is very slow since it depends on $\sqrt{\log(n)}$** . The test does not draw the line at a fixed value of $|a / s|$." $K \propto \exp\left(-½\frac{a^2}{s^2}\right)\sqrt{n} = \exp\left[-½\left(\frac{a^2}{s^2} - \log n\right)\right]$

Consistance du FB

$$B_{01} \simeq \sqrt{n/\lambda}\exp(-½z^2) \text{ avec } z = \sqrt{n}\left(\bar{y}-\mu_0\right)/\sigma$$

$$2\log\left(B_{01}\right) = -n\left(\bar{y}-\mu_0\right)^2/\sigma^2 + \log\left(n\right) + Cte$$

$$\boxed{\text{Sous } H_0 : z^2 \sim \chi_1^2 \text{ et } B_{01} \to +\infty, \Pr\left(H_0 \mid y\right) \to 1}$$

$$NP : \Pr\left(|Z| > z_{\alpha/2}\right) = \alpha = \Pr\left(\text{rejet } H_0 \mid H_0\right) > 0$$

$$\text{Sous } H_1 : z^2 \to n\left(\mu-\mu_0\right)^2/\sigma^2 = kn \to +\infty$$

$$2\log\left(B_{01}\right) \to -kn + \log\left(n\right) + Cte \to -\infty$$

$$\boxed{\text{Sous } H_1 \; B_{01} \to 0, \Pr\left(H_0 \mid y\right) \to 0}$$

Consistance du FB mais pas du test d'hypothèses à moins de faire tendre $\alpha \to 0$, qd $n \to \infty$ (calibration de $\alpha(n)$?)

"One could instead argue that the true paradox is that this consistency is overlooked in most commentaries on the Lindley-Jeffreys paradox" Robert (2013)

Echantillonnage avec arrêt optionnel préférentiel

-Intérêt du FB pour se prémunir contre le risque d'un échantillonnage séquentiel avec arrêt optionnel orienté.
"Optional & preferential stopping rule for a foregone conclusion" (Good,1991,1992)

Exemple:
-Echantillonner séquentiellement et arrêter l'analyse dès qu'on a trouvé un résultat significatif (ex rejet de H_0)

-Pté inhérente au principe de la vraisemblance: On n'a pas à prendre en compte la règle d'arrêt (raisonnement conditionnellement aux données observées et non au dispositif expérimental)

Echantillonnage avec arrêt optionnel préférentiel

-Test d'hypothèses NP: Pr(Rejet H_0) croît vite avec le nbre de tests

-$B_{10} \approx C \exp(½z^2) / \sqrt{n} \approx C' \boxed{\log n / \sqrt{n}}$ (loi du logarithme itéré)

-D'un point de vue fréquentiste: $\boxed{\underbrace{\Pr(BF_{10} > k)}_{\Pr(BF_{01}<1/k)} \leq 1/k}$ $H_0 = H$ vraie; $H_1 = H$ fausse

"If you set out to collect data until your posterior probability for a hypothesis which is unknown to you is true has been reduced to 0.01, then 99 times out of 100 you will never make it, no matter how many data, you and your children after you may collect" Edwards et al (1963)

cf Sanborn & Hills (2013) pour plus de détails sur les conditions d'application de la règle

Paradoxe/effet de l'a priori

Pour $n >> \lambda \quad (\tau^2 = \sigma^2 / \lambda \Rightarrow \lambda = \sigma^2 / \tau^2)$

$$\ln B_{01} \simeq -½ z^2 + ½ \ln n - ½ \ln \lambda$$

A z fixé et n fixé

$\tau^2 (\text{ou } \delta) \nearrow \ (\lambda \searrow) \Rightarrow \ln B_{01} \nearrow$ "moins de preuves contre H_0 "

$$\tau^2 (\text{ou } \delta) \to +\infty \text{ (a priori diffus)} \Rightarrow B_{01} \to +\infty \Rightarrow \text{acceptation de } H_0$$

"I thus remained convinced that the richest consequence of Jeffreys' (1939) and Lindley's (1957) exhibitions of this paradox is to highlight **the difficulty in using improper or very vague priors in testing settings**" (Robert, 2013)

"The only assumption that will be questioned is the assignment of a prior distribution of any type, and, in particular, of the chosen form" (Lindley, 1957)

"Being indecisive about the alternative hypothesis we simply should not select it" (Robert, 2013)

Paradoxe/effet du rapport λ/n

Ce qui compte c'est le rapport n/λ de l'information apportée par les données par rapport à celle contenue dans l'a priori

$$-2\log B_{01} \simeq z^2 - \ln n/\lambda$$

A z fixé

1) part des données croissante

$(n/\lambda)\nearrow \Rightarrow -2\log B_{01}\searrow \Rightarrow B_{01}\nearrow$ "moins de preuves contre H_0 "

2)part de l'a priori croissante

$(n/\lambda)\searrow \Rightarrow -2\log B_{01}\nearrow \Rightarrow B_{01}\searrow$ "plus de preuves contre H_0 "

Est-ce vraiment un paradoxe?

- OUI en apparence si l'on s'en tient à des conclusions hâtives, NON en réalité si l'on examine les deux théories qui se réfèrent à des objets et concepts différents.
- Espace d'échantillonnage vs espace paramétrique
- Confrontation des données à une hypothèse nulle (H0) vs comparaison de H0 et d'une alternative particulière (H1)
- Paramètres fixés vs paramètres incertains affectés d'une distribution a priori
- P-value qui n'est pas une mesure de comparaison de l'adéquation de 2 hypothèses mais simplement de l'adéquation de l'une d'elles aux données
- Conditionnement par un intervalle de valeurs possibles vs données observées
- "Tail area" vs rapport de vraisemblance

Pourquoi tant de passion vis-à-vis de ce sujet?

"The overall conclusion is that P values can be highly misleading measures of the evidence provided by the data against the null hypothesis" (Berger & Sellke, 1987)

« Conventional significance tests do not allow the researcher to state evidence for the null. Hence they are not appropriate for competitively testing the null against an alternative" (Rouder et al, 2009)

"Most of the significant P values that fell in the range (0.01-0.05) probably represent P values that were computed from data in which the null hypothesis of no effect was true" (Johnson, 2013)

"One result, I hope, will be that the conventional P value of approximatively 0.05 when testing a simple statistical hypothesis H0, will be correctly interpreted: not as a good reason for rejecting H0, but as a reason for obtaining more evidence provided that the original experiment was worth doing in the first place." (Good, 1987)

"First, we should recognize how unreliable p is, how little information p values give, and how misleading they can be if used to sort results dichotomously" (Cumming, 2008)

Pourquoi tant de passion vis-à-vis de ce sujet?

"I have just stressed the essence of Lindley's paradox is precisely the inability of Bayesian theory to represent the strength of evidence" (Schafer, 1982) (en cause les a priori plats)

"The discussion has also called into question the basic premise of the Jeffreys-Lindley paradox concerning the sagacity of the Bayes factor favoring H0 as n increases as symptomatic of the fallacy of acceptance, the reverse problem plaguing the P-value" (Spanos, 2013)

"What is unreasonable is to regard the sort of 'contradiction' implicit in the Lindley paradox as a reason in itself for regarding P-values as illogical for, to adopt and adapt the rhetoric of Jeffreys" (Senn, 2001)

"So before bashing p-values too much we should be careful because, like democracy to government, p-values may be the worst form of statistical significance calculation except all those other forms that have been tried from time to time" (Leek & Izirrary, 2012)

Pourquoi tant de passion vis-à-vis de ce sujet?

"We argue that this (BF_{01} increasing as $\sqrt{n}$) is an undesirable behavior inconsistent with accepted scientific practice" (Bernardo, 1999)

"Unfortunately from our perspective, the problem (Jeffreys-Lindley paradox) has been studied by Bayesians with an eye of "solving it",..., but we think that this is really a problem without a solution" (Gelman, Shalizi, 2013)

"The main point where we disagree with many Bayesians is that we do not see Bayesian methods as generally useful for giving the posterior probability that a model is true, or the probability for preferring model A over model B, or what so ever" (Gelman, Shalizi, 2013)

"The weight of Lindley's paradoxical result along with related analyses, ...,began to burden proponents of the Bayesian movement who wished to reorient statistical practice to sound principles of inferential evidence" (Lad, 2003)

Statut de l'hypothèse nulle

Critique de la légitimité de l'hypothèse nulle

1) Comparaison de deux races (ou deux variétés)
2) Ségrégation selon les proportions mendeliennes
   Ex Jeffreys : 459 vs 137 H0 ratio : 3-1 ie 447 vs 149
   Idem pour tester un déséquilibre de liaison nul

-Distinction entre « estimation » en (1) et « test d'hypothèses » en (2) (ou décision à prendre)

-Estimation redevable de l'IC (Cumming, 2012) ou de l'inférence a posteriori sur θ (Spiegelhalter et al, 2004)

-H0 : Mise en exergue d'une invariance à laquelle on attache une masse de probabilité en un seul point

-"We are aiming at the best way of progressing not at the unattainable ideal of immediate certainty" Jeffreys (1961, p388)

-"All models are wrong but some are useful" (Box & Draper, 1987) or "without any loss" (Morey et al, 2013)

-"To check whether or not the null model is a suitable proxy for the correct model" (Bernardo, 1999)

Mesures d'accommodement

- Calibration des P-values
- Choix de la distribution des paramètres
- Modifier Pr(H0)/Pr(H1) (Robert, 2013)

Calibration des P-values en fonction de n

Good (1988) propose une règle empirique de standardisation des p-values en référence à un effectif $N = 100$

$$P^* = \min\left(½, P\sqrt{N/100}\right)$$

"I guess that standardized p-values will not become standard before the year 2000" !!

Ex Stone P=0.0027 $\rightarrow$ P* = 0.196

Calibration des P-values en général

Sellke, Bayarri & Berger (2001) proposent deux formules de calibration

$$\boxed{B(p) \leq -ep\log(p)} \quad si\ p < 1/e \text{ (borne inférieure)}$$

$$\boxed{\alpha(p) \geq \frac{1}{1-\frac{1}{ep\log(p)}}}$$

Interp^{on} fréquentiste (erreur de 1ère espèce conditionnelle)

ou bayésienne Pr(H_0 | p)

Borne inférieure trop petite, défavorable à H_0

"can be used as a quick and dirty calibration of a p-value when only H_0 is avalaible" (Berger, 2003)

Ex Stone: p=0.0027 $\rightarrow \alpha(p) = 0.042$

Table 1. Calibration of p Values as Odds (Bayes factors) and Conditional Error Probabilities

p	.2	.1	.05	.01	.005	.001
$B(p)$	.870	.625	.407	.125	.072	.0188
$\alpha(p)$	.465	.385	.289	.111	.067	.0184

Choix des a priori sous H1/Bartlett

Rappel : A priori uniforme $B_{01} = \frac{\delta}{\sigma}\sqrt{\frac{n}{2\pi}}\exp\left(-½z^2\right)$

Se référant à Neyman-Pearson où n est déterminé en fonction de $d = |\mu - \mu_0|$ soit $\sqrt{n} \propto 1/d$

Bartlett (1957) suggère de prendre $\delta/\sigma = A/\sqrt{n}$ $(A = cste)$

$B_{01} = \frac{A}{\sqrt{2\pi}}\exp\left(-½z^2\right)$ ce qui élimine le paradoxe

Choix des a priori sous H1/Smith & Spiegelhalter

Smith et Spiegelhater (1980) suggèrent de prendre

$\mu \sim \mathcal{N}\left(\mu_0, \sigma^2 / n\right)$ ("local alternatives to the null hypotheses)

$$\boxed{B_{01} = \sqrt{2}\exp\left(-\boxed{1/4} z^2\right)} \Rightarrow -2\log B_{01} = 1/2 z^2 - \log 2\left(p_1 - p_0\right)$$

avec ici $p_0 = 0$ et $p_1 = 1$

Disparition du paradoxe mais forte valeur de z pour contrer H_0

ex : Pour $B = 1/19$ ie $(\Pr\ H_0 \mid y)$=0.05, il faut $|z| = 3.93$

Statut de l'a priori dépendant de n "either as a genuine subjective Bayesian analysis,..., or simply as a formal analysis, intended as a theoretical **ad hoc** device..." S&S

Avec un autre a priori et dans le cadre d'un modèle linéaire

$$B_{01} = \exp\left[3/4\left(p_1 - p_0\right) - 1/2 \mathrm{D}_{01}\right] \Rightarrow -2\log B_{01} = \mathrm{D}_{01} - \boxed{3/2}\left(p_1 - p_0\right)$$

Rappel: $\Delta\mathrm{AIC}_{01} = \mathrm{D}_{01} - \boxed{2}\left(p_1 - p_0\right)$

Facteur de Bayes a posteriori

Aitkin (1991) considère sous $H_1, \mu \sim \mathcal{N}\left(\bar{y}, \sigma^2 / n\right)$

"FB a posteriori" $A_{01} = L_0 / \tilde{L}_1$ $\quad \boxed{\tilde{L}_1 = \int f\left(\bar{y} \mid \mu\right) \pi\left(\mu \mid \bar{y}\right) d\mu}$

$\boxed{A_{01} = \sqrt{2} \exp\left(-½ z^2\right)}$ **paradoxe éliminé**

on peut utiliser des a priori "plats" dans le calcul de $\pi\left(\mu \mid \bar{y}\right)$

Plus généralement $\quad -2 \log A_{01} = \Delta D_{01} + \underbrace{\left(p_0 - p_1\right) \boxed{\log 2}}_{\textit{pénalité}}$

$$\Delta D_{01} = -2 \log\left[f(\mathbf{y} \mid \hat{\boldsymbol{\theta}}_{0,ML}) / f(\mathbf{y} \mid \hat{\boldsymbol{\theta}}_{1,ML}) \right]$$

Double utilisation des données (Gelman,Robert & Rousseau,2011)

$$\tilde{L}_1 = \int f\left(\bar{y} \mid \mu\right) \frac{f\left(\bar{y} \mid \mu\right) \pi\left(\mu\right)}{m\left(\bar{y}\right)} d\mu = \frac{m\left(\bar{y}, \bar{y}\right)}{m\left(\bar{y}\right)}$$

Distribution a posteriori du rapport de vraisemblance

Soit $\boxed{\mathrm{LR}(\mu) = \frac{L(\mu_0)}{L(\mu)}}$ où $L(\mu) = f(\bar{y} \mid \mu)$

Aitkin (2005, 2010) propose de calculer $\boxed{\Pr\left[\mathrm{LR}(\mu) < k \mid \bar{y}\right]}$

$k = 0.1,\ 0.01$ ou 0.001 par ex

Il montre que $\Pr\left[\mathrm{LR}(\mu) < 1 \mid \bar{y}\right] = 1 - Pvalue$

La P-value est la probabilité a posteriori que la rapport de vraisemblance considéré comme un fonction du paramètre soit supérieur à 1 (en faveur de H_0 *vs* H_1)

"The posterior distribution of the likelihood is meaningless within a Bayesian perspective" Gelman, Robert &Rousseau (2011)

FB partiels,intrinsèques et fractionnaires

FB partiels $y = \left(\underbrace{y_A}_{apprentissage}, \underbrace{y_B}_{test} \right)$ $\pi_p(\theta) = \pi(\theta \mid y_A)$

FB intrinsèques (Berger, Perrichi, 1996)

FB fractionnaires (O'Hagan, 1995)

$$\boxed{\pi(\theta, b) \propto f(y \mid \theta)^b \, \pi(\theta) \quad 0 \le b \le 1}$$

$BF = FBF^{1/(1-b)}$ ("consistant" si $b \to 0$)

Divers choix de b : $n_{0,\min} / n$, $\log(n) / n$, $\sqrt{n} / n$

$$-2 \log BF_{01} = z^2 + \left[\log(b) \right] / (1-b)$$

$$Si \; b = 1/\sqrt{n} \Rightarrow -2 \log BF_{01} \approx z^2 - \boxed{½} \log(n)$$

Retour à l'exemple de Stone (1997)

Résultats du test H0=0.2 vs H1≠H0

dans l'expérience de Stone@ selon différentes procédures

A priori sous H1	B01	Pr(Ho/y)$
U(0,1)	8.11	0.89
Beta(1,4)	3.99	0.80
U(0.1,0.3)	1.61	0.60
Beta(10,40)	1.17	0.54
FBF*	0.303	0.23
BFmin	0.223	0.18
BF Smith&Spiegelhalter	0.151	0.13
PBF**	0.016	0.016
P-value		(0.0027)#

@ y=106298, n=527135 $P(H0)=1/2

*Prior of fractional Bayes factor b=1/√n=1/726 ** Posterior Bayes Factor

#à titre indicatif et avec les réserves d'usage quant à sa signification

Autres approches

- Tests de sévérité (Mayo &Spanos,2006; Spanos, 2013)
- P-rep (Killeen,2005; Cumming, 2005; Lecoutre et al, 2010) Pté prédictive fiduciaire qu'une réplication donne un résultat de même signe (ou un résultat significatif pour P-srep)
- Bayesian Reference Criterion (Bernardo,1999, Sprenger 2012) (espérance a posteriori d'une fonction d'utilité)
- UMP Bayesian Test (Johnson, 2013ab)
- FB basé sur des fonctions de scores

Test unilatéral d'hypothèses composites

Casella & Berger (1987), Morris (1987)

$H_0 : \mu \le \mu_0$ *vs* $H_1 : \mu > \mu_0$; σ^2 connu

$y_i \sim_{iid} \mathcal{N}\left(\mu, \sigma^2\right)$ $i = 1,..,n$ et $\bar{y} \mid H_k \sim \mathcal{N}\left(\mu, \sigma^2 / n\right)$

$$\pi\left(\mu \mid H_0\right) = \frac{2}{\tau}\phi\left(\frac{\mu - \mu_0}{\tau}\right) I_{(\mu \le \mu_0)}, \pi\left(\mu \mid H_1\right) = \frac{2}{\tau}\phi\left(\frac{\mu - \mu_0}{\tau}\right) I_{(\mu > \mu_0)}$$

$\tau^2 = \sigma^2 / \lambda$ et pour $\Pr(H_0) = ½$,

$$\boxed{\Pr\left(H_0 \mid \bar{y}\right) = \Phi(-Cz)}$$

$C^2 = n / \left(n + \lambda\right)$ $z = \sqrt{n}\left(\bar{y} - \mu_0\right) / \sigma > 0$

Pas de paradoxe

$n \to +\infty \Rightarrow C \to 1 \Rightarrow \Pr\left(H_0 \mid \bar{y}\right) \to \Phi(-z) = \text{P-value}$

Plus λ / n grd, plus C petit, plus $\Pr\left(H_0 \mid \bar{y}\right) > \text{P-value}$

P-values standardisées $\boxed{P^* = \Phi\left(-z^*\right)}$ avec $z^* = Cz$

Test unilatéral/Exemple

Exemple de test unilatéral d'hypothèses composites

$H_0 : \theta \leq ½$ vs $H_1 : \theta > ½$ (Morris, 1987)

	(a)*	(b)	(c)
Données	15/20 (0.75)	115/200 (0.575)	1046/2000 (0.523)
Z	2.03	2.05	2.03
P-value uni	0.021	0.020	0.021
C	0.408	0.816	0.876
Pr(θ<0.5|y)	0.204	0.047	**0.024**

$(n / y, p = y / n)$, $\sigma^2 = (0.5)^2$, $\tau^2 = (0.05)^2$, $\lambda = 100$, $C = \sqrt{n / (n + \lambda)}$ *calculs exacts basés sur la loi binomiale

Discussion

- Le test de H0 ponctuelle vs H1 composite (cas d'école) est en fait plein de chausse-trappes quelle que soit la théorie avec laquelle on l'aborde
- Une P-value donnée n'a pas la même signification selon la taille de l'échantillon
- Le choix de l'a priori sous H1 est déterminant dans l'expression du FB (il fait partie intégrante de H1)
- Bien le calibrer en fonction des connaissances sur le sujet et des objectifs de l'étude

Discussion/suite

- Faut-il élaborer (Berger) ou non (Lindley) une synthèse des théories des trois écoles (Fisher, Neyman-Pearson, Jeffreys) en matière de tests d'hypothèses?
- Si oui, dans quelle direction et sur la base de quels principes?
  - ✓ Analyse classique avec conditionnement par une statistique relative au degré de preuve apportée par les données vis-à-vis de H0/H1 (Berger, Boukai & Wang, 1997; Berger, 2003)
  - ✓ Justification fréquentiste de techniques bayésiennes (Good, 1992; Aitkin, 1995, 2010; Aerts et al, 2004, Johnson, 2013)
  - ✓ Distributions prédictives (Gelman & Shalizi, 2013 ; Lecoutre et al, 2010) +Théorie de la décision (Bernardo, 1999)
- Nécessité d'un consensus minimum dans les résultats et conclusions formulés par les praticiens

Discussion/suite

- Attention aux techniques a priori séduisantes qui manquent de cohérence interne
- Distinction entre estimation et tests d'hypothèses (ou prise de décision)
- Le statut des paradoxes est de créer le trouble et d'obliger à plus de réflexion sur les fondements des diverses théories et c'est bien le cas de celui de Jeffreys-Lindley
- Pour les fans de paradoxes logiques, mathématiques et probabilistes (cf site « mes paradoxes favoris » de Gerville-Reache)

Références

- Aerts M, Claertens G, Hart JD (2004) Bayesian-motivated tests of functions fit and their asymptotic properties. The Annals of Statistics, 32, 2580-2615
- Aitkin M (1991) Posterior Bayes factors. JRSS, 53, 111-142
- Aitkin M (1997) The calibration of P-values, posterior Bayes factors and the AIC from the posterior distribution. Statistics and Computing, 7, 253-261
- Aitkin M (2010) Statistical inference: an integrated Bayesian/Likelihood approach. Chapman & al/CRC monographs on statistics and applied probability
- Aitkin M, Boys RJ, Chadwick T (2005) Bayesian point null hypotheses. Statistics and Computing, 11, 217-230
- Bartlett MS (1957) A comment on DV Lindley's statistical paradox. Biometrika, 44, 533-534
- Berger JO (1985) Statistical theory and Bayesian analysis. Springer
- Berger JO (2003) Could Fisher, Jeffreys and Neyman have agreed on testing. (with discussion), Statistical Science, 18, 1-32
- Berger JO, Sellke T (1987) Testing a point null hypothesis: the irreconcilability of P values and evidence. JASA, 82, 112-122
- Berger JO, Delampady M (1987) Testing precise hypotheses. Statistical Science, 2, 317-352
- Berger JO, Perrichi LR (1996) The intrinsic Bayes factor for model selection and prediction. JASA, 91, 109-122
- Berger JO, Boukai B, Wang Y (1997) Unified frequentist and Bayesian testing of a precise hypothesis (with discussion). Statistical Science, 3, 133-160
- Bernardo JM (1999) Nested hypothesis testing: the Bayesian reference criterion. Bayesian Statistics, 6, 101-130
- Box GEP, Draper NR (1987) Empirical model building and response surfaces, Wiley, 424p
- Casella G, Berger RL (1987) Reconciling Bayesian and frequentist evidence in the one sided tested problem. JASA, 82, 106-111
- Cumming G (2005) Understanding the average probability of replication: Comment on Killeen (2005). Psychological Science. 16, 1002–1004
- Cumming G (2008) Replication and p intervals. Perspective on Psychological Science. 3, 286-300
- Cumming G. (2012). Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. Routledge, New York

Références/suite

- DeGroot MH (1973) Doing what comes naturally: interpreting a tail area as a posterior probability or as a likelihood ratio. JASA, 68, 966-969
- Dickey JM (1977) Is the tail area useful as an approximate Bayes factor? JASA, 72,138-142
- Edwards W, Lindman H, Savage LJ (1963) Bayesian statistical inference for psychological research. Psychological Review, 70, 193-242
- Gelman A, Shalizi CR (2013) Philosophy and the practice of Bayesian statistics. British J of Mathematical and Statistical Psychology, 66, 8-38
- Gelman A, Robert CP, Rousseau J (2013) Inherent difficulties of non-Bayesian likelihood based inference as revealed by an examination of a recent book by Aitkin. arXiv:1012.2184v2. Statistics and Risk Modeling
- Good IJ (1982) Comment on Shafer. JASA, 77, 342-344
- Good IJ (1987) Comment on Berger & Sellke and Casella & Berger. JASA, 82, 125-128
- Good IJ (1991) A comment concerning optional stopping. J of Statistical Compuring and Simulation, 39'191-192
- Good IJ (1992) The Bayes/Non Bayes compromise: a brief review. JASA, 87, 597-606
- Goodman SN (1999) Toward evidence based statistics. 1: the P va lue fallacy. Annals of Internal Medicine, 130, 996-1004
- Hubbard R, Bayrarri MJ (2003) Confusion over measures of evidence (p's) versus errors (α's) in classical testing. The American Statistician, 57, 171-178
- Hung HM, O'Neill RT, Bauer P, Kohne K (1997) The behavior of the P-value when the alternative hypothesis is true. Biometrics, 57, 11-22
- Jeffreys H (1961) Theory of probability (3rd edition) Oxford-Clarendon Press
- Johnson VE (2013) Uniformly most powerful Bayesian tests. Annals of Statistics, 41, 1716-1741
- Johnson VE (2013) Revised standards for statistical evidence. PNAS, 110 (48);19175-19176; doi:10.1073/iti4813110
- Kass RE, Raftery AE (1995) Bayes factors. JASA, 90, 773-795
- Killeen PR (2005) An alternative to null hypothesis tests. Psychological Science, 16, 345-353

Références/suite

- Kuha (2004) AIC and BIC Comparisons of assumptions and performance. Sociological Methods & Research, 33,188-229;doi: 10.1177/0049124103262065.
- Lad F (2003) Appendix: the Jeffreys-Lindley paradox and its relevance to statistical testing. Tech report, Conference on science and democracy, Napoli
- Lebarbier E, Mary-Huard T (2006) BIC: fondements théoriques et interprétation. Journal de la Société Française de Statistique, 147, 39-57
- Lecoutre B, Lecoutre M-P, Poitevineau J (2010) Killeen's probability of replication and predictive probabilities: How to compute, use and interpret them. Psychological Methods, 15, 158-171.
- Leek J, Irizarry R (2012) P-values and hypothesis testing get a bad rap-but we sometimes find it useful. Simply Statistics: http://simplystatistics.tumblr.com/
- Leonard T, Hsu JSJ (1999) Bayesian Methods. Cambridge University Press
- Lindley DV (1957) A statistical paradox. Biometrika, 44, 187-192
- Lindley DV, Scott WF (1984) New Cambridge statistical tables. Cambridge University Press, Cambridge
- Marin JM, Robert CP (2007) Bayesian Core: A Practical Approach to Computational Bayesian Statistics, Springer
- Mayo DG, Spanos A (2006) Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. British J of Philosophy of Science, 57, 323-357
- Morey RD, Rouder JN (2011) Bayes factors approaches for testing interval null hypotheses. Psychological Methods, 16, 406-419
- Morey RD, Romeijn J-W, Rouder JN (2013) The humble Bayesian: Model checking from a fully Bayesian perspective. Bristish J of Mathematical & Statistical Psychology,66,68-75
- Morris CN (1987) Comment on Berger & Sellke and on Casella & Berger. JASA, 82, 131-133
- O'Hagan A (1995) Fractional Bayes factors for model comparison. JRSS B, 57, 99-138
- Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, PcPherson K, Peto J, Smith PG (1976) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I: Introduction and design. British Journal of Cancer, 34, 585-612

Références/suite

- Raftery AE (1995) Bayesian model selection in social research. Sociological Methodology, 25, 111-163
- Robert CP (2013) On the Jeffreys-lindley's paradox. Philosophy of Science (in press)
- Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009) Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review, 16, 225-237
- Royall RM (1986) The effect os sample size on the meaning of significance tests. The American Statistician, 40, 313-315.
- Sandborn AN, Hills TT (2013)The frequentist implications of optional stopping on Bayesian hypothesis tests. Psychonomic Bulletin and Review, in press
- Sellke T, Bayarri MJ, Berger JO (2001) Calibration of p values for testing precise null hypotheses. The American Statistician, 55, 62-71
- Senn S (2001) Two cheers for P-values. J of Epidemiologu & Statistics, 6, 193-204
- Shafer G (1982) On Lindley's paradox (with discussion). JASA, 378, 325-351
- Smith AFM, Spiegelhalter DJ (1980) Bayes factors and choice criteria for linear models. JRSS B, 42, 213-220
- Smith AFM, Spiegelhalter D (1982) Bayes factors for linear and log-linear models with vague prior information. JRSS B, 44, 377-387
- Spanos A (2013) Who should be afraid of the Jeffreys-lindley paradox? Philosophy of Science, 80, 73-93
- Spiegelhalter DJ, Abrams KR, Myles JP (2004) Bayesian approaches to clinical trials and health-care evaluation. J Wiley &Sons.
- Sprenger J (2013) Testing a precise null hypothesis: the case of Lindley's paradox. Philosophy of Science, to appear
- Stone M (1997) Discussion of papers by Dempster and Aitkin. Statistics & Computing, 7, 263-264
- Van der Pas SL (2010) Much ado about the p-value. Thesis, Math Institute, U of Leiden, NL

Annexe: approximation asymptotique du FB dans le cas binomial

B_{01} peut s'exprimer à partir du rapport Dickey-Savage (Leonard & Hsu,1999)

$$\boxed{B_{01} = \frac{\pi_1(\theta_0 \mid y)}{\pi_1(\theta_0)}} \quad \pi_1(\theta),\ \pi_1(\theta \mid y) \text{ a priori et a posteriori de } \theta \text{ sous } H_1$$

$$\pi_1(\theta_0 \mid y) \simeq \frac{1}{\sqrt{2\pi \underbrace{p(1-p)}_{\sigma^2}/n}} \exp\left[-½ \frac{(\theta_0 - p)^2}{\underbrace{p(1-p)/n}_{z^2}} \right] \quad \pi_1(\theta_0) = \frac{1}{\sqrt{2\pi\tau^2}}$$

$$B_{01} \simeq \frac{\tau^2}{\sigma^2/n} \exp(-½z^2) = \sqrt{\frac{n}{\lambda}} \exp(-½z^2) \text{ où } \tau^2 = \sigma^2/\lambda$$

$$-2\log B_{01} \approx z^2 - \log n + \log \lambda$$

Prendre $\tau^2 = \dfrac{1}{2\pi\left[\pi_1(\theta_0)\right]^2}$ soit $\boxed{\lambda = 2\pi p(1-p)\left[\pi_1(\theta_0)\right]^2}$