

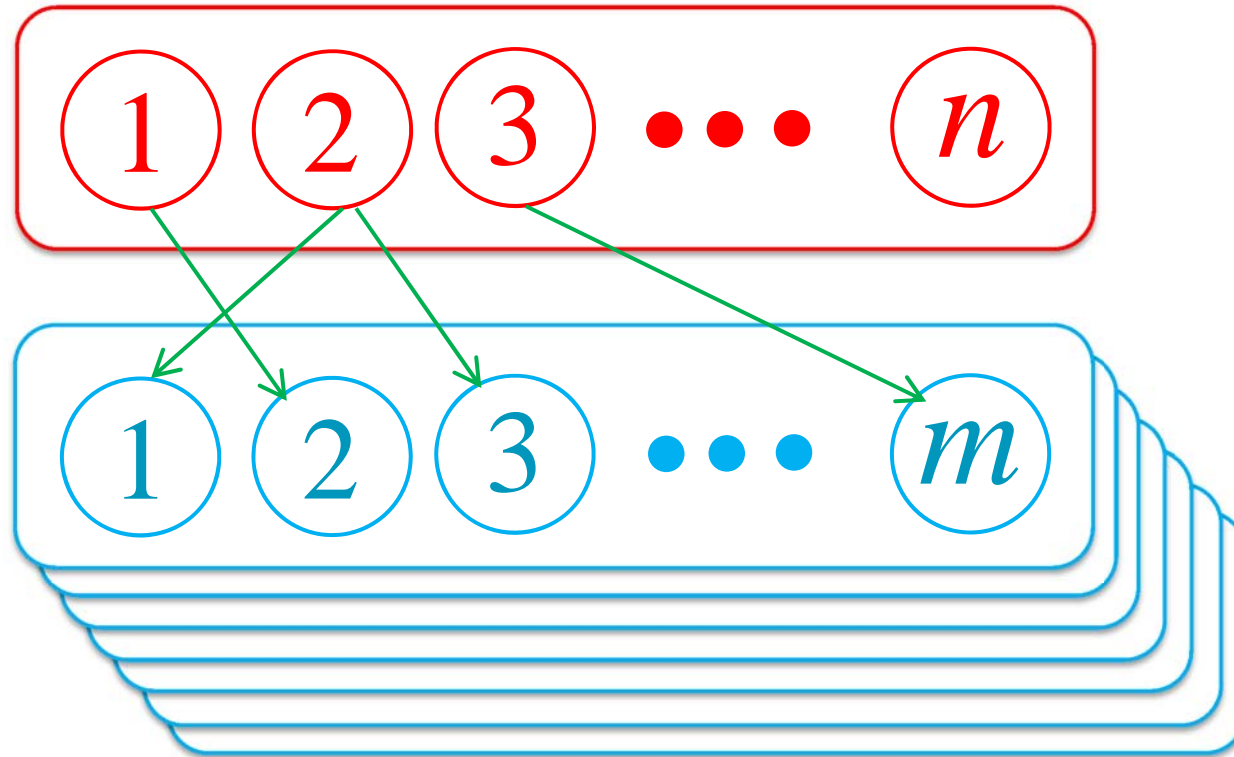
# なぜサンプルサイズを負にすると ベイズが頻度論に変わるのか？



東工大・情報理工

下平英寿

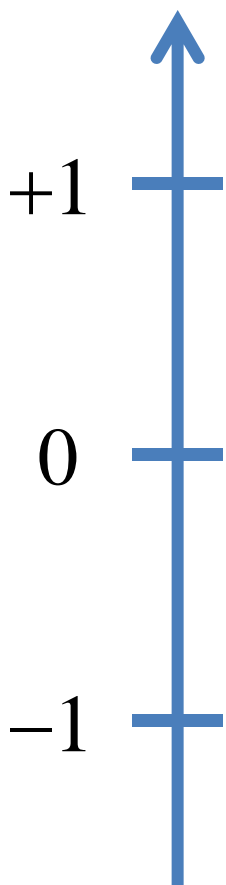
# サンプルサイズを $m$ に変更



$m$ -out-of- $n$  bootstrap

# 負のサンプルサイズ？

$$\sigma^2 = \frac{n}{m} \quad \text{リサンプリングの相対的な分散}$$



$$m = n$$

ブートストラップ法  
Efron (1979)  
Felsenstein (1985)

$$m = \infty$$

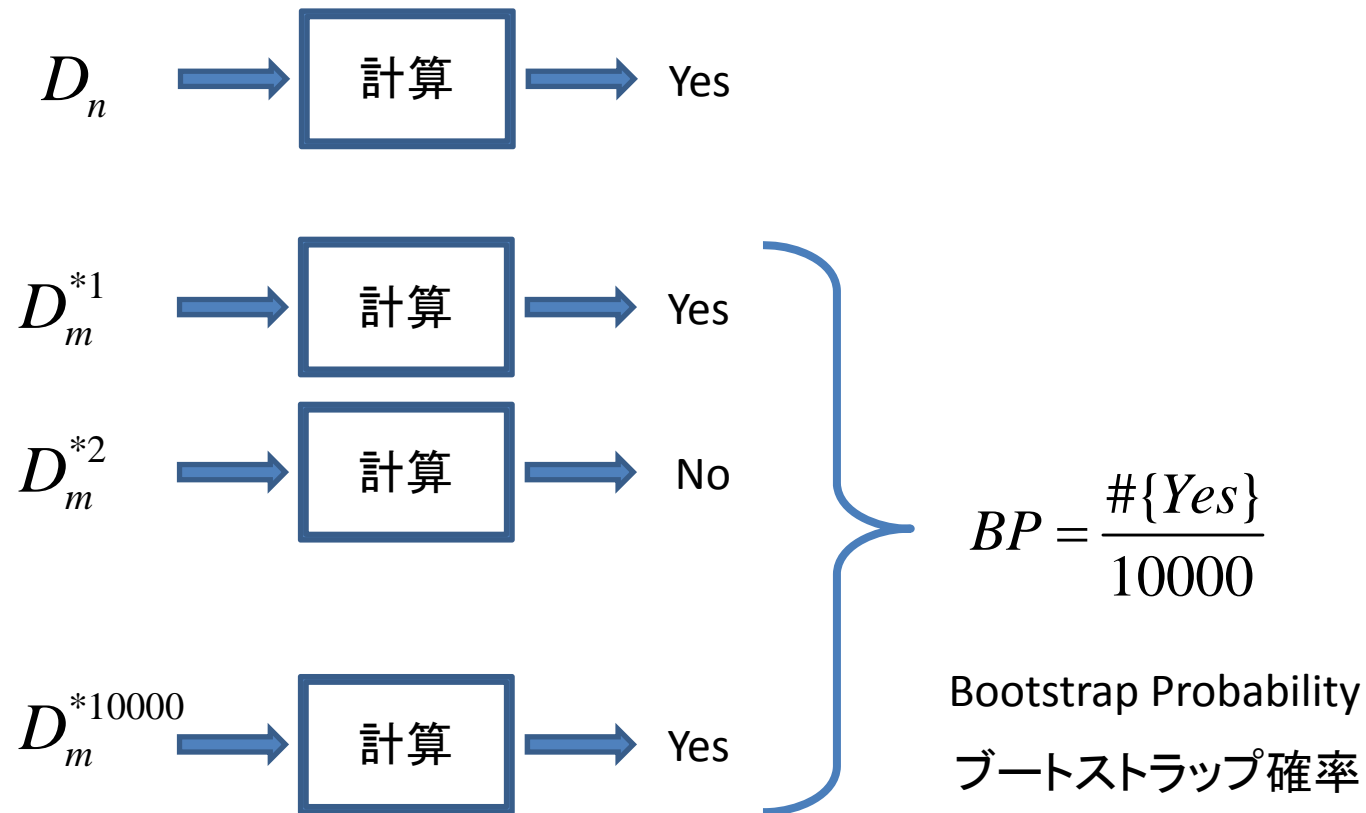
機械学習の能動学習など  
Shimodaira, Kanamori, Aoki, and Mine (in prep)

$$m = -n$$

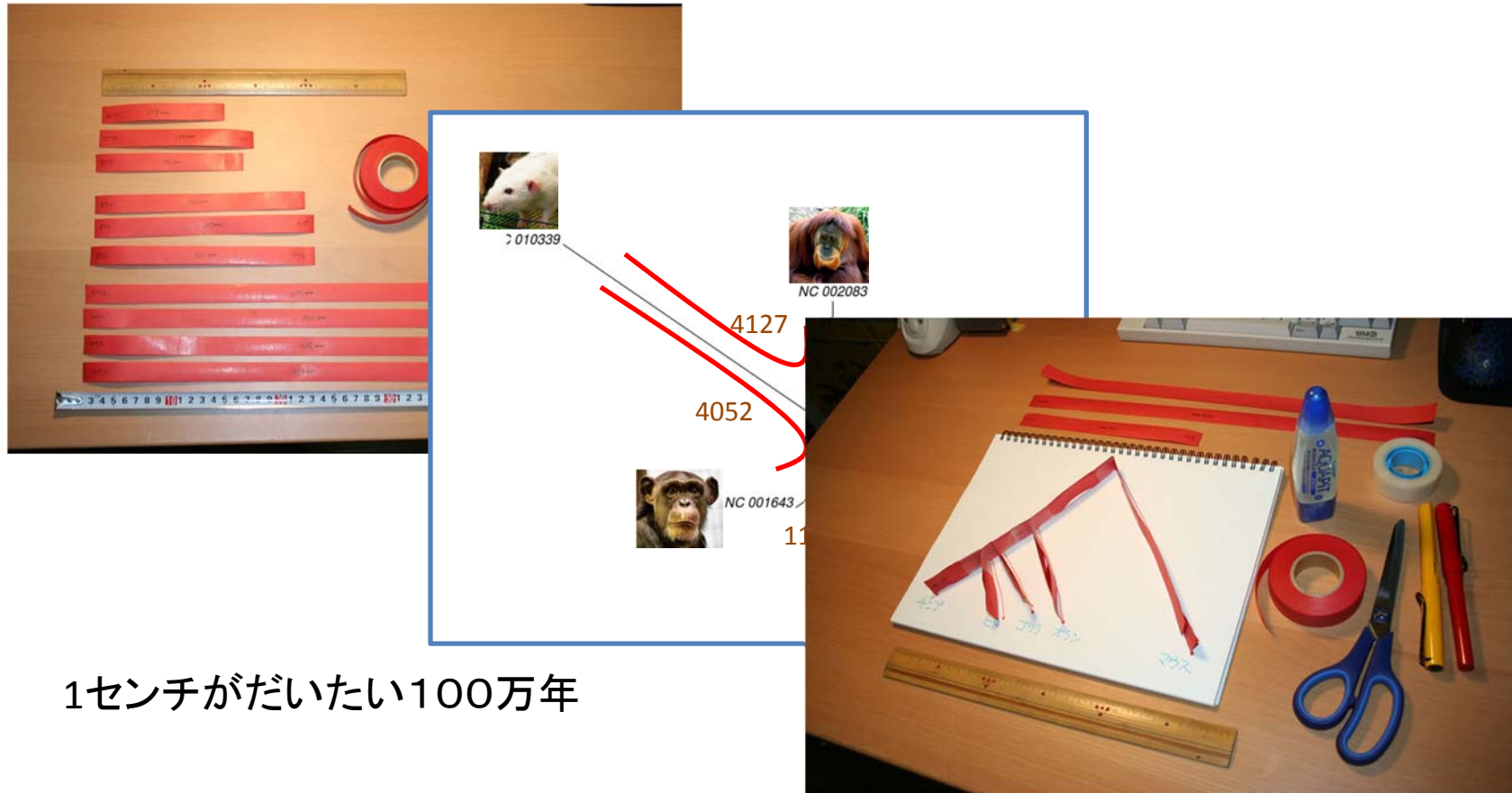
マルチスケール・ブートストラップ法  
Shimodaira (2002, 2004, 2008, 2010)

偽の「発見」  
が多い問題

# カウントだけ使う



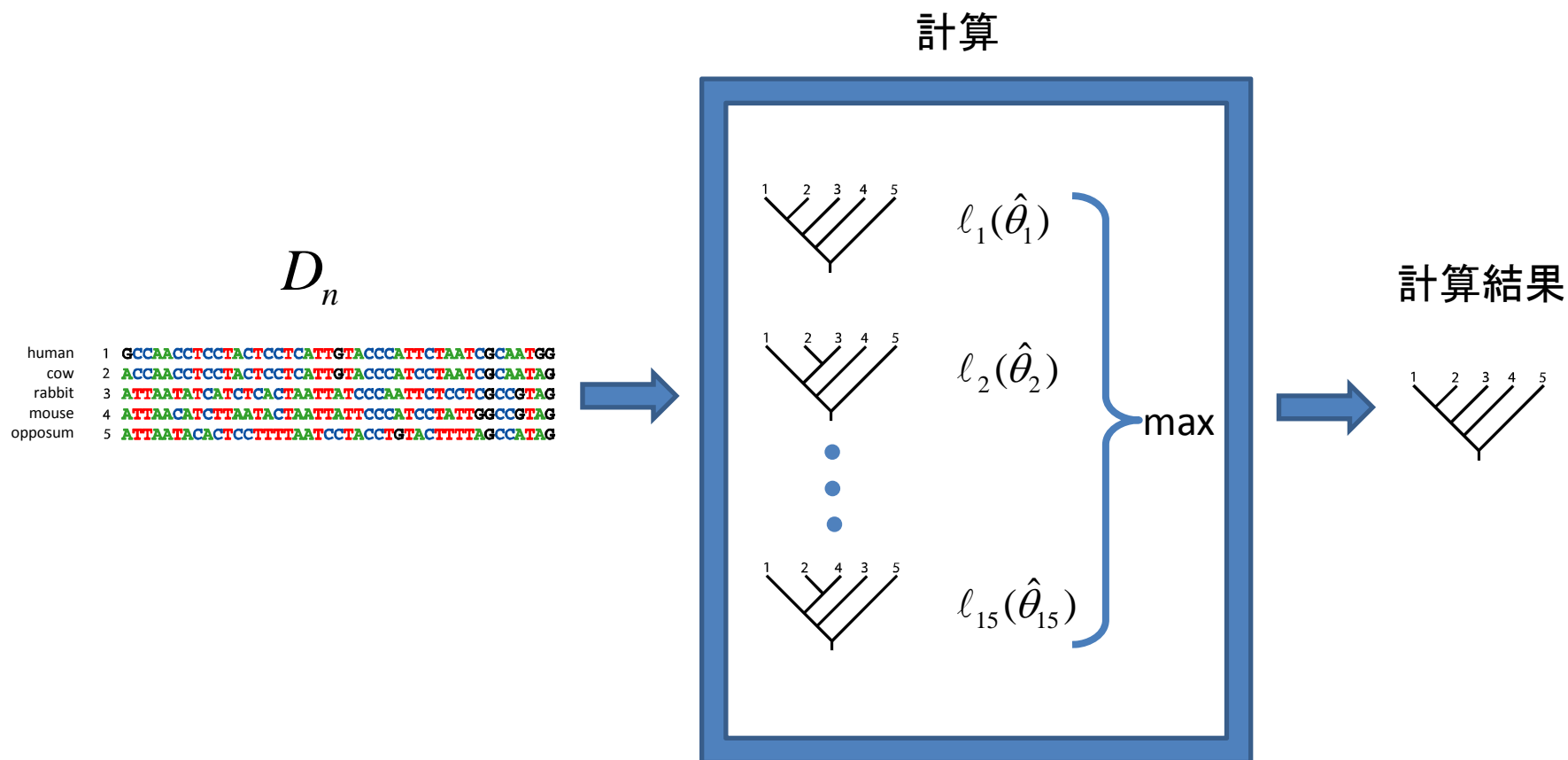
# DNA配列から系統樹を推定



1センチがだいたい100万年

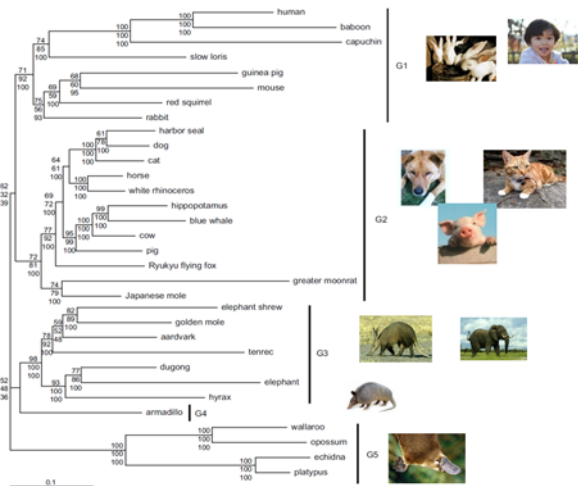
下平@大田区民大学 (2010)

# 「計算」はブラックボックスとみなす



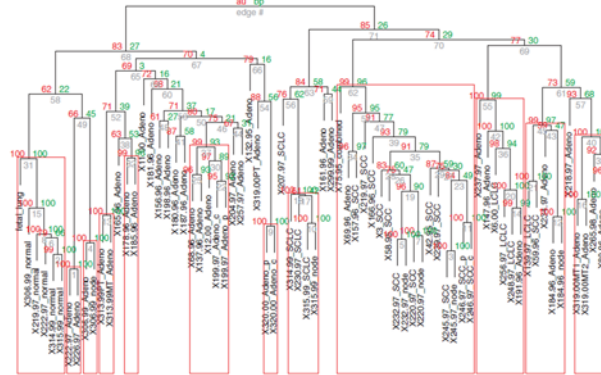
# Universalな方法にしたい

Phylogenetic Tree



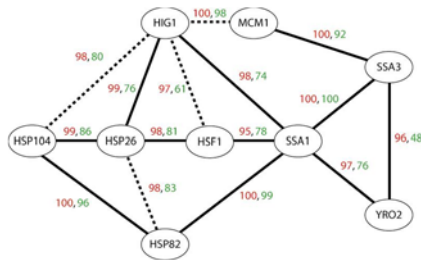
Shimodaira and Hasegawa (2005)

Classifying Lung Tumors from Gene Expressions



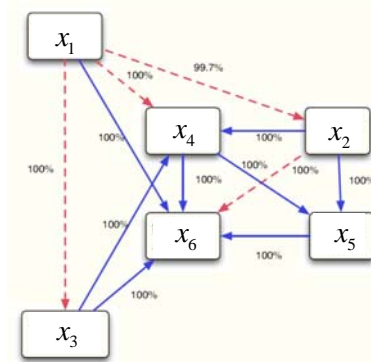
Suzuki and Shimodaira (2006)

Estimating Gene Network



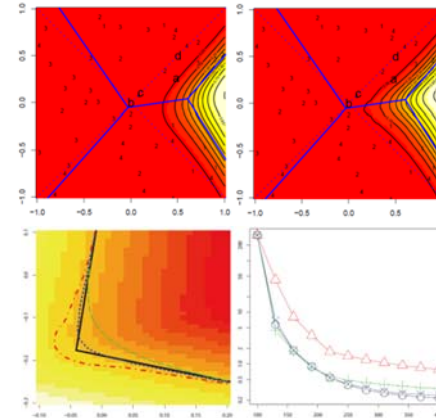
Kamimura et al. (2005)

Causality Inference



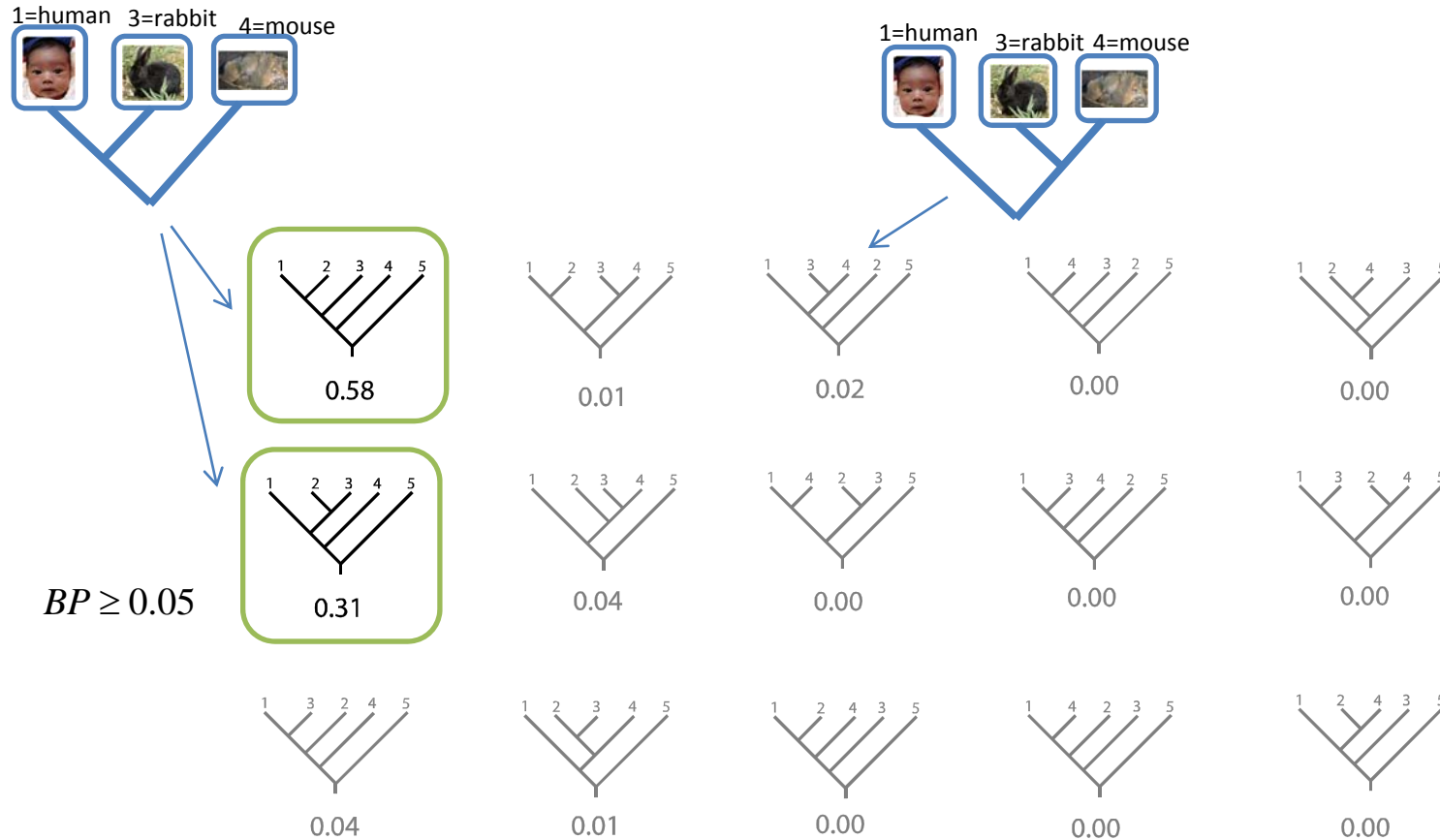
Komatsu, Shimizu and Shimodaira (2010)

Machine Learning



Shimodaira, Kanamori, Aoki and Mine (2010)

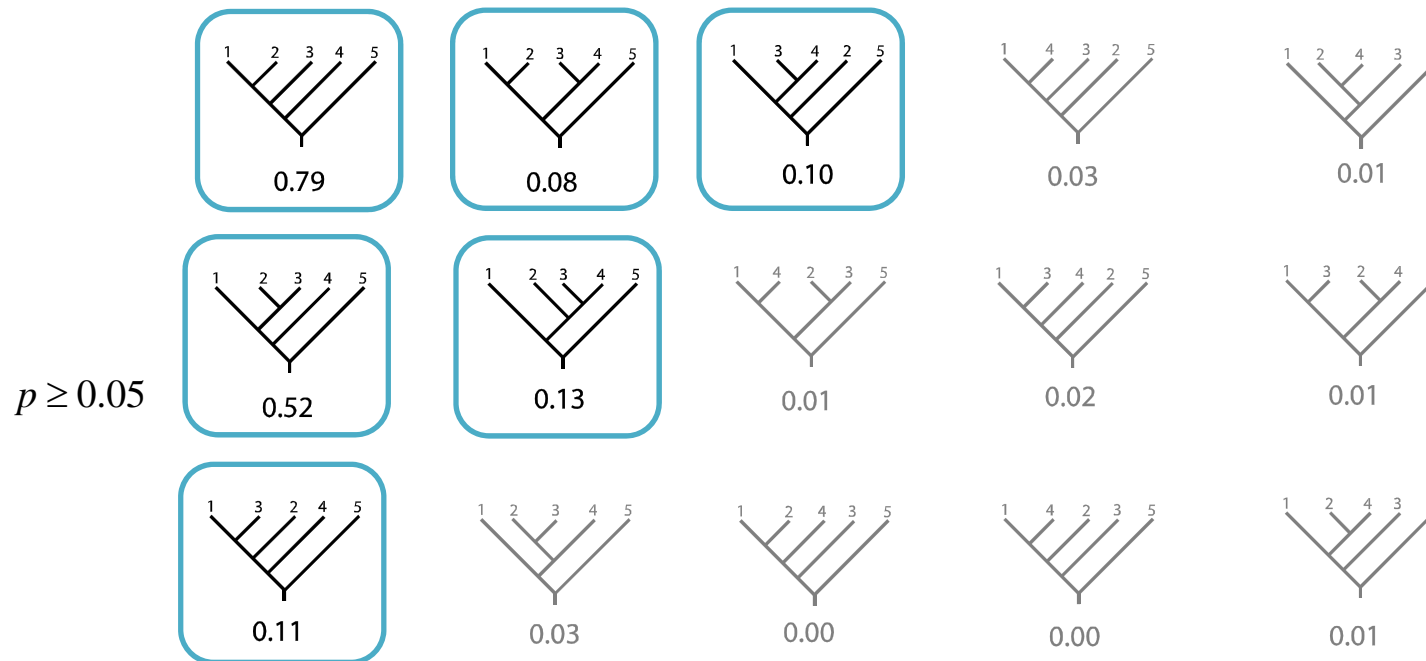
# 「発見」は誤りだった



ブートストラップ確率  $m=n$

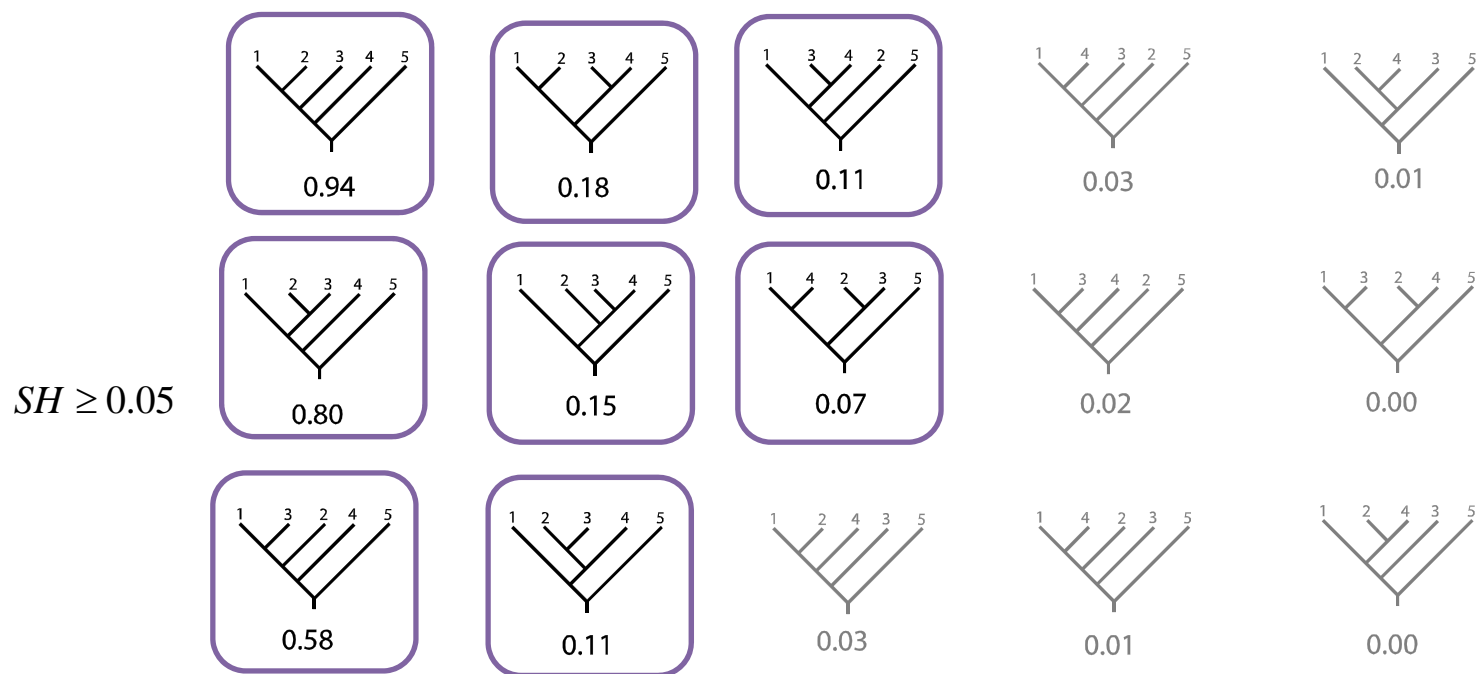


# 「誤った発見」を回避できる



マルチスケール・ブートストラップ  $m=-n$

# 本当の発見を見逃してしまうかも

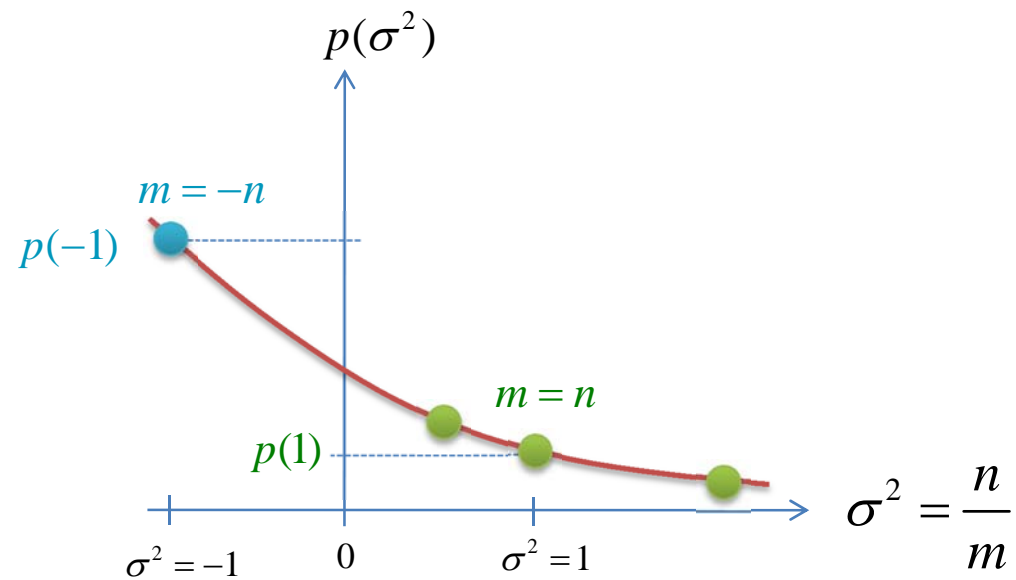


Shimodaira-Hasegawa test (対数尤度の多重比較)

# $m = -n$ へ外挿する

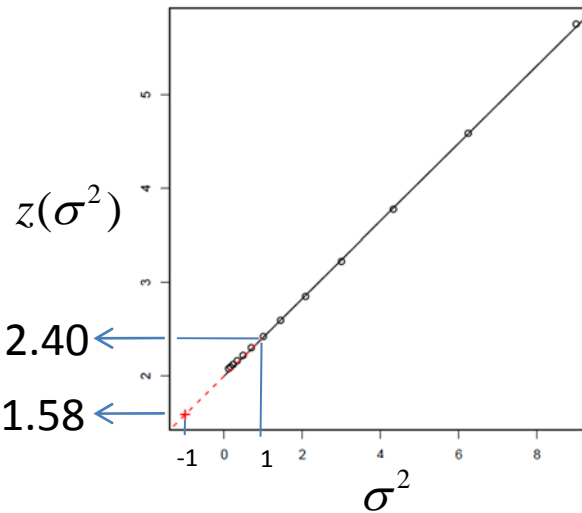
$$p(\sigma^2) \triangleq \Phi \left\{ \sigma \Phi^{-1} \left( \text{BP}(\sigma^2) \right) \right\}$$

正規化ブートストラップ確率 (Shimodaira 2008)



# Bias reduces as $k$ increases

$$\left[ \begin{array}{l} z(\sigma^2) \triangleq -\sigma \Phi^{-1}(BP(\sigma^2)) \\ p(\sigma^2) = \Phi\{-z(\sigma^2)\} \end{array} \right.$$



BP=0.0078 ( $k=1$ )

$p=0.057$  ( $k=2$ )

$p=0.052$  ( $k=3$ )

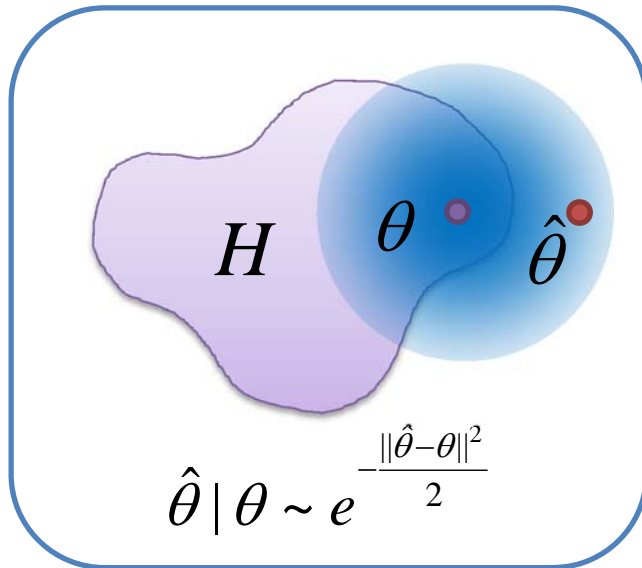
$p=0.05$  (exact)

$$z(\sigma^2) = \beta_0 + \beta_1 \sigma^2 + \beta_2 \sigma^4 + \beta_3 \sigma^6 + \dots$$

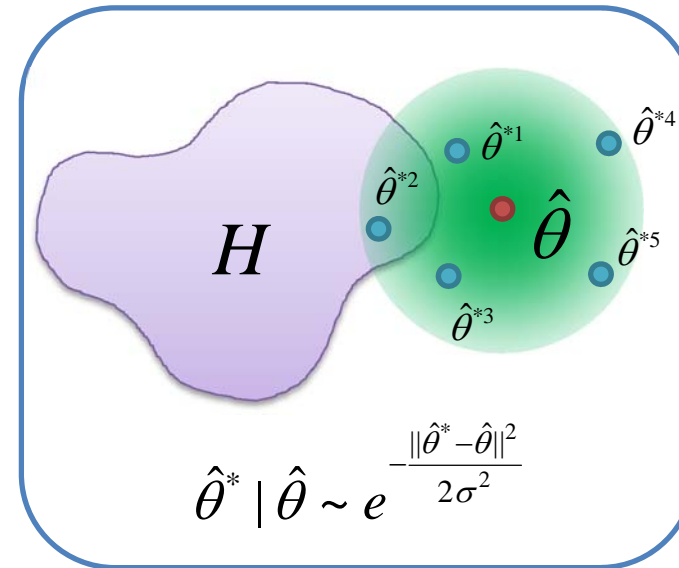
using  $k=2$  terms

# ブートストラップはベイズ

$$\theta \Rightarrow \hat{\theta}$$



$$\hat{\theta} \Rightarrow \hat{\theta}^*$$



$$\theta \Leftarrow \hat{\theta}$$

$$\theta | \hat{\theta} \sim e^{-\frac{\|\theta - \hat{\theta}\|^2}{2}}$$

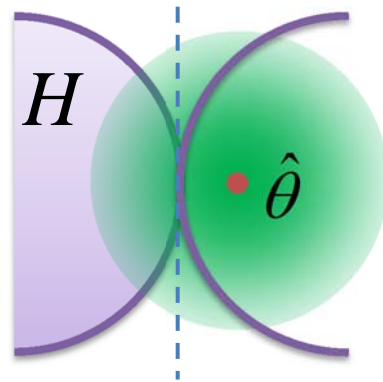


$$\theta \equiv \hat{\theta}^* \text{ with } \sigma^2 = 1$$

Multivariate Normal Model (Efron and Tibshirani 1998)

# 曲率の反転でベイズが頻度論に

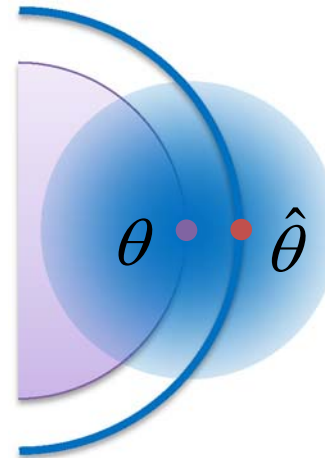
Bayesian



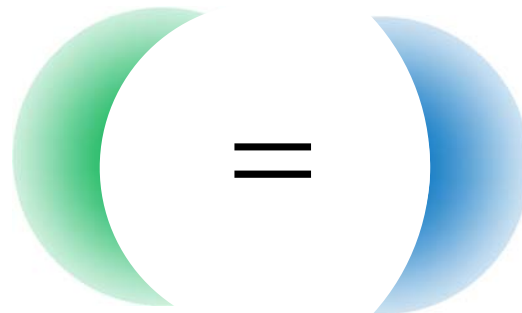
$$BP(1) \approx \Phi[-(\beta_0 + \beta_1)]$$

$\beta_0$ =distance,  $\beta_1$ =curvature

Frequentist

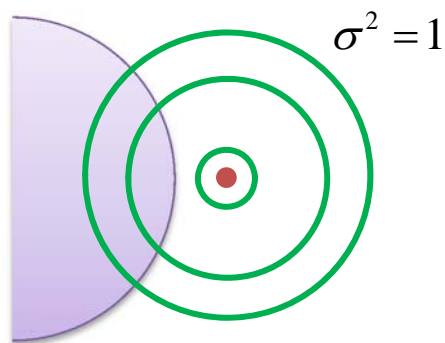


$$p\text{-value} \approx \Phi[-(\beta_0 - \beta_1)]$$



# 負のサンプルサイズで曲率が反転

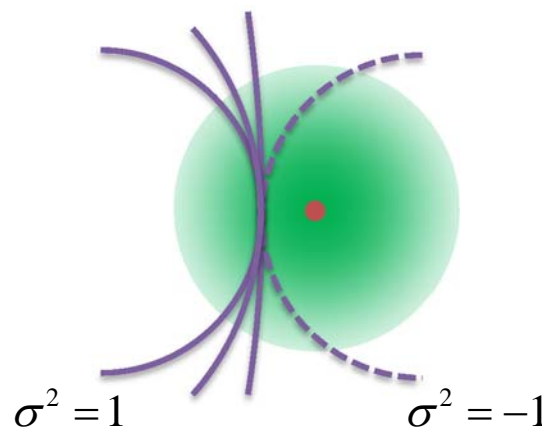
$$BP(\sigma^2)$$



$$BP(1) \approx \Phi[-(\beta_0 + \beta_1)]$$

$$BP(\sigma^2) \approx \Phi\left[-\left(\frac{\beta_0}{\sigma} + \beta_1\sigma\right)\right]$$

$$p(\sigma^2) = \Phi\left\{\sigma \Phi^{-1}\left(BP(\sigma^2)\right)\right\}$$



$$p(1) \approx \Phi[-(\beta_0 + \beta_1)]$$

$$p(\sigma^2) \approx \Phi\left[-(\beta_0 + \beta_1\sigma^2)\right]$$

スケールリング則

# Bridging Bayesian to Frequentist

$$p(\sigma^2) = \Phi \left[ -(\beta_0 + \beta_1 \sigma^2 + \beta_2 \sigma^4 + \beta_3 \sigma^6 + \dots) \right]$$

signed distance      mean curvature

$$\sigma^2 = +1$$

BP(1)  
(Bayesian)

$$p(+1) = \Phi \left[ -(\beta_0 + \beta_1 + \beta_2 + \beta_3 + \dots) \right]$$



$$\sigma^2 = -1$$

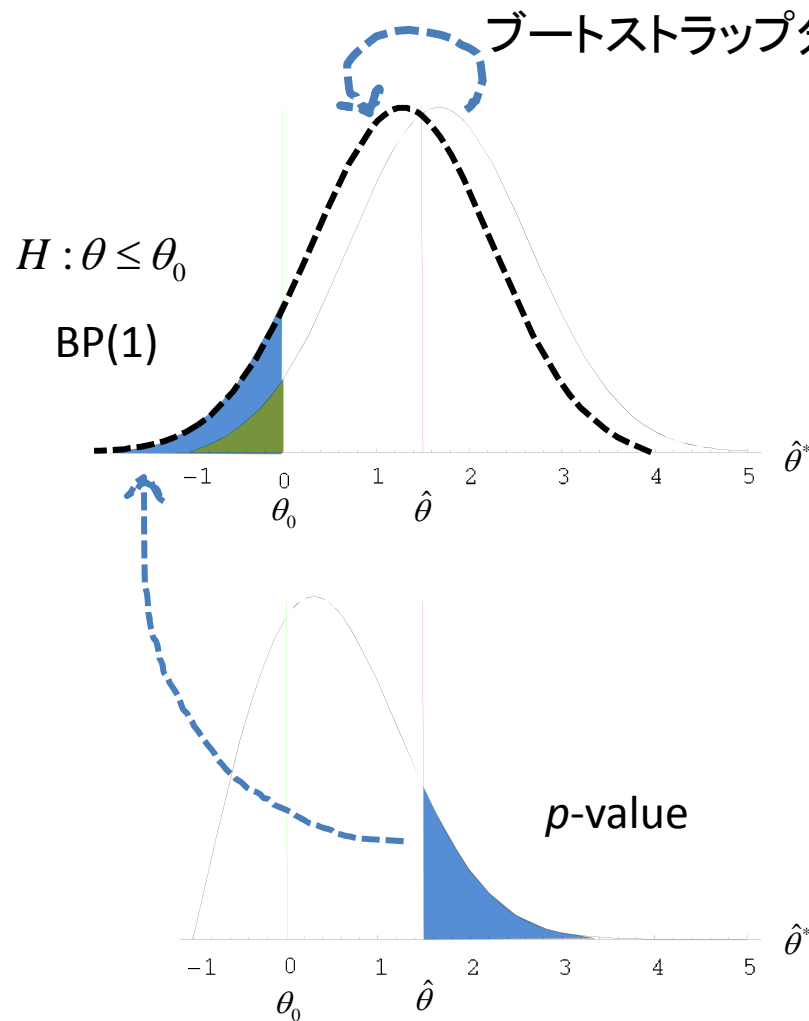
$p$ -value  
(frequentist)

$$p(-1) = \Phi \left[ -(\beta_0 - \beta_1 + \beta_2 - \beta_3 + \dots) \right]$$

Shown for “nearly flat surfaces” in Shimodaira (2008)



# もしパラメータにアクセスすれば簡単



データからブートストラップ分布

帰無仮説からシミュレーション

# $m = -n$ でも2次の精度にならない！

Here we work on the “smooth function model” of Hall (1989)

$$\hat{\theta} = f(\bar{x}_n) \quad \hat{\tau}^2 = g(\bar{x}_n)$$

The percentile confidence interval of Efron

$$BP = \hat{F}(\theta_0) \quad \text{with} \quad \hat{F}(\theta) \triangleq P(\hat{\theta}^* \leq \theta | D_n)$$

$$P\{\theta_0 \in [\hat{F}^{-1}(\alpha), \infty)\} = 1 - \alpha + O(n^{-1/2})$$

The bootstrap- $t$  confidence interval

$$p = 1 - \hat{G}\left(\frac{\hat{\theta} - \theta_0}{\hat{\tau}}\right) \quad \text{with} \quad \hat{G}(t) \triangleq P\left(\frac{\hat{\theta}^* - \hat{\theta}}{\hat{\tau}^*} \leq t | D_n\right)$$

$$P\{\theta_0 \in [\hat{\theta} - \hat{\tau}\hat{G}^{-1}(1 - \alpha), \infty)\} = 1 - \alpha + O(n^{-1})$$

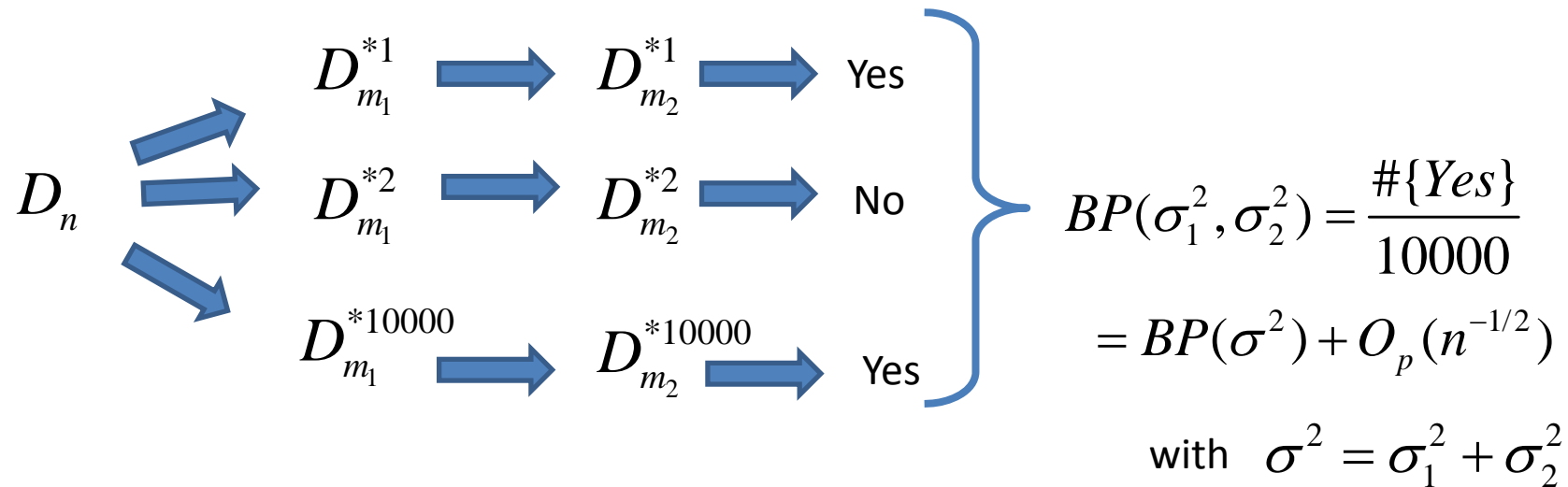
The results of  
Hall (1989)

The confidence interval derived from the multiscale bootstrap

$$p(\sigma^2) = \hat{H}_{\sigma^2}(\theta_0) \quad \text{with} \quad \hat{H}_{\sigma^2}(\theta) \triangleq \Phi\left[\sigma\Phi^{-1}(\hat{F}_{\sigma^2}(\theta))\right] \quad z_\alpha \triangleq \Phi^{-1}(\alpha)$$

$$P\{\theta_0 \in [\hat{H}_{\sigma^2}^{-1}(\alpha), \infty)\} = 1 - \alpha + n^{-1/2}\phi(z_\alpha)\left\{-(1 + \sigma^2)b_1 + z_\alpha^2 A\right\} + O(n^{-1})$$

# Estimating the skewness term $A$



$$-\sigma \Phi^{-1}(BP(\sigma_1^2, \sigma_2^2)) = -\sigma \Phi^{-1}(BP(\sigma^2)) + n^{-1/2} A \sigma_1^2 \sigma_2^2 \sigma^{-4} (\beta_0 - \sigma^2) + O_p(n^{-1})$$

Two-step multiscale bootstrap of Shimodaira (2004)

# もっとアルゴリズムを考えよう

Bias Corrected (BC) method of Efron (1982)

$$BC \triangleq \Phi \left[ \Phi^{-1}(\hat{F}(\theta_0)) - 2 \times \Phi^{-1}(\hat{F}(\hat{\theta})) \right] = p(-1) + O_p(n^{-1})$$

Accelerated Bias Corrected (BCa) method of Efron (1987)

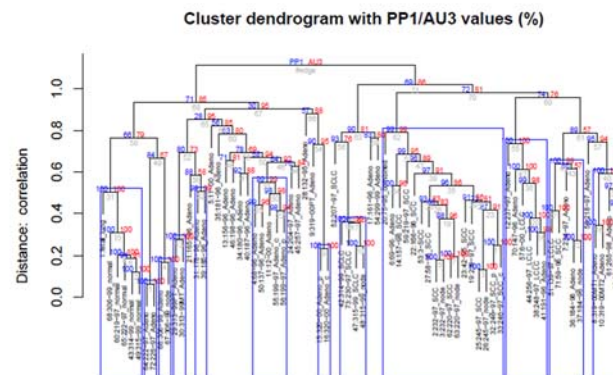
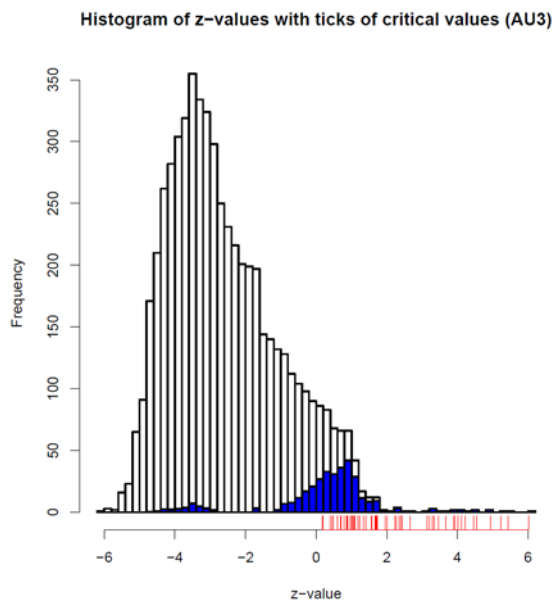
$$BC_a = \text{the two-step multiscale bootstrap} + O_p(n^{-1})$$

The double bootstrap (Hall 1992)

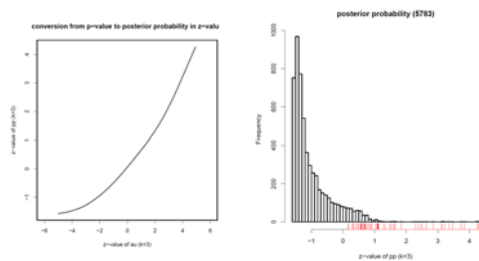
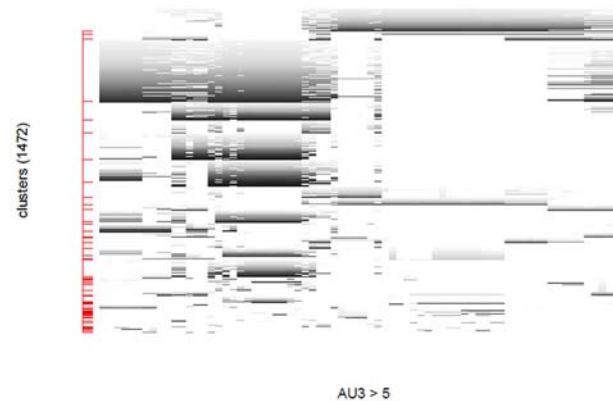
The three-step multiscale bootstrap (Shimodaira 2004)

The error is only  $O(n^{-3/2})$

# たくさんの仮説：FDR, 可視化



PP1 > 95  
Cluster method: average



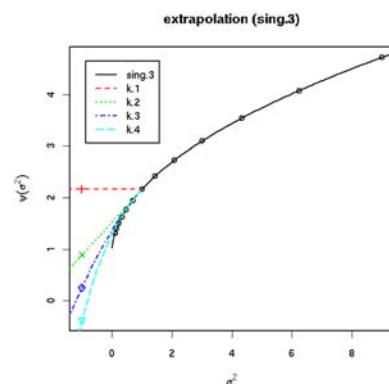
下平英寿 (2009), 統計連合大会, 多数の片側検定に関するFalse Discovery Rate のマルチスケール計算

# 役に立ってるが、やるべき課題が...

たくさんの仮説： かなり深刻

使うときの注意： AICそのまま使うのダメ。パラメタ数の補正は半分にする。

都合の良い仮定が問題： 滑らかな曲面，十分に平坦な曲面，正規分布の仮定， etc.



$$z(\sigma^2) \approx \beta_0 + \beta_1 \sqrt{\sigma^2}$$

そもそも頻度論的なp-値が問題？ ベイズと頻度論の関係， マッチング事前分布