生態学における AICの誤用

粕谷　英一　（九州大学・理・生物）

2014.3

複数のモデルを比べる・片方を選ぶ

検定　　偶然のみによるばらつきとの比較

AIC　　予測の最適化

Bayes factor　モデルへの支持がデータによりどれだけ変化

BIC　　事後確率

モデルを比べる基準がちがうーモデルの「よさ」が異なる

複数のモデルを比べる・片方を選ぶ

検定　　偶然のみによるばらつきとの比較

AIC　　予測の最適化

Bayes factor　モデルへの支持がデータによりどれだけ変化

BIC　　事後確率

モデルを比べる目的がちがうーモデルの「よさ」が異なる

「よさ」の基準がちがえば、
「よいモデル」はちがう

AICは、正しいモデルを選ぶものではないので、正しいモデルを選ばない

AIC＝－２×(最大対数尤度)
＋２×（自由なパラメーター数）

AIC

$$AIC = -2 \times (\text{最大対数尤度}) + 2 \times (\text{自由なパラメーター数})$$

予測の最適化

AICが小さいモデルほど平均的に精度の高い予測

典型的な場合　２つのデータを使う

データセット　説明変数と目的変数
各モデルのAIC算出と予測モデル決定

新しい説明変数のみのデータ
予測モデルで目的変数を計算

AICの使われ方

変わった使い方

ここ数年、生態学会でもよく見かける

地域、対照限定の現象ではない

AIC（赤池情報量規準）が最小のモデル

よいモデル　　真である（あるいは真にとても近い）

原型：単純に最小のモデル　　バリエーション：AICの差の閾値を設ける

2が多い

例.目的変数Yはどんな要因（説明変数）の影響を受けているか

X1,X2,X3

回帰的モデル

どの要因も影響せず　　X1のみ影響　　X2のみ影響　　X3のみ影響

X1とX2が影響　　X1とX3が影響　　X2とX3が影響　　X1とX2とX3が影響

AICが最小のモデル

たとえば「X2とX3が影響」

Yに、X1は影響を与えず、X2とX3が影響を与えていることがわかった

適用が比較的容易

AICが計算できれば、あとは数値の大小比較だけ

対数尤度が計算できて最大化できればいい

結論がはっきりでる（変異型では,出ないこともある）

比較が単純：AICが一番小さいモデルを選べばよい

AICのこういう使い方

AICの目的から外れている一目的外使用

予測の最適化

正しいモデルを選べるのか

すでにわかっていること（例、Shibata(1976)以来、多数）

選べない

正しいモデル

データを生成したモデル

真のモデル

true, generating

分析（計算）

あるモデル（真のモデル、正しいモデル）で所定のサンプルサイズのデータを生成

AIC最小になるモデルを記録

シミュレーション10000回

データ生成したモデル（真、正しい）が選ばれる割合をみる

正しいモデルが候補の中に含まれている［AICの本来の目的では好適条件］

データが多ければ、正しいモデルが非常に高い割合で選ばれる
サンプルサイズが無限大なら、正しいモデルが必ず選ばれる

データ生成するモデル（真、正しい）

比べるモデルたちの中でパラメーター数が最小

サンプル数（n）が無限大のときに解析的な結果が使える

例１．二項分布

Fisherの正確確率検定や２×２分割表のカイ２乗検定のような状況

データ生成（真の、正しい、モデル）

二項分布（確率0.65で1、確率0.35で0）

第１のサブサンプル（n=1000）
第２のサブサンプル（n=1000）

両サブサンプルとも等しい確率で生成

検討するモデル

モデル１　　２つのサブサンプルとも等しい確率
自由なパラメーター数＝1

モデル２　　サブサンプルごとに確率がちがってもいい
自由なパラメーター数＝2

例１．二項分布

同内容だが、こう見てもいい

データ生成（真の、正しい、モデル）

二項分布（確率0.65で1、確率0.35で0）

第１のサブサンプル（n=1000）　説明変数＝0

第２のサブサンプル（n=1000）　説明変数＝1

両サブサンプルとも説明変数に関係なく等しい確率で生成

検討するモデル

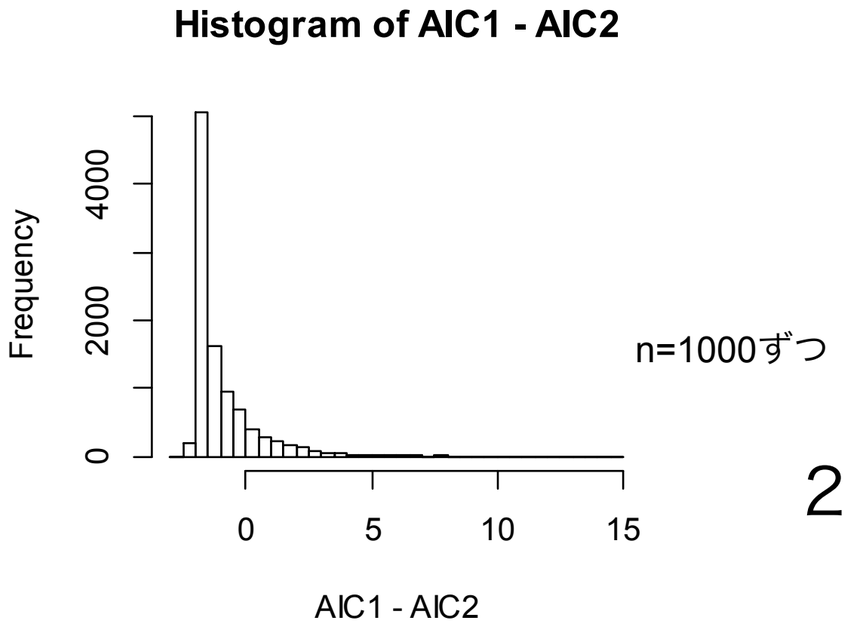
モデル１　説明変数の効果なし　２つのサブサンプルとも等しい確率

自由なパラメーター数＝1

モデル２　説明変数の効果あり　サブサンプルごとに確率がちがってもいい

自由なパラメーター数＝2

例１．二項分布



n=1000ずつ

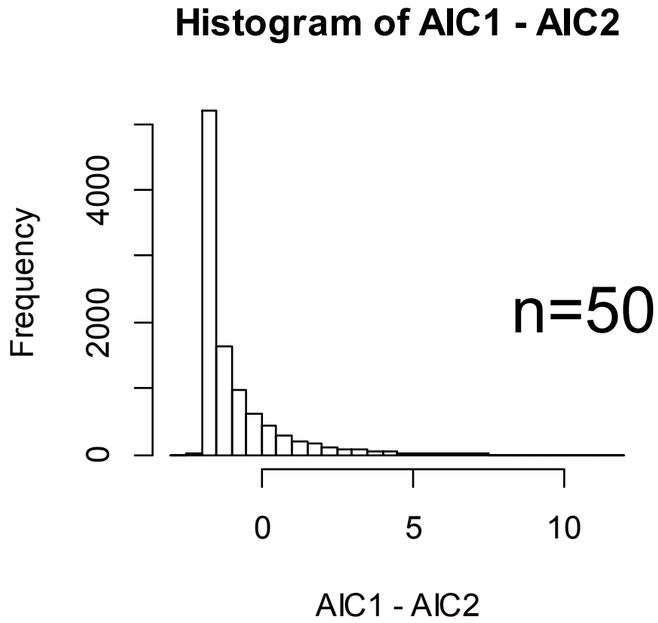
２つのモデルのAICの差のヒストグラム

AIC１がAIC2よりも小さいのが10000回のうち8502回

正しいモデルはモデル1だから、正しいモデルを選ぶ割合は85%

正しくないモデルを選ぶ割合は15%

例１．二項分布



サンプルサイズを増やす

n=50000ずつ

２つのモデルのAICの差のヒストグラム

AIC１がAIC2よりも小さいのが10000回のうち8472回

正しいモデルはモデル1だから、正しいモデルを選ぶ割合は85%
正しくないモデルを選ぶ割合は15%

例２．正規分布

studentのt検定や２水準の一元配置分散分析のような状況

データ生成（真の、正しい、モデル）

目的変数ー正規分布　　説明変数１か0

サンプル（n=2000）

説明変数が1でも0でも目的変数の分布は同じ

検討するモデル

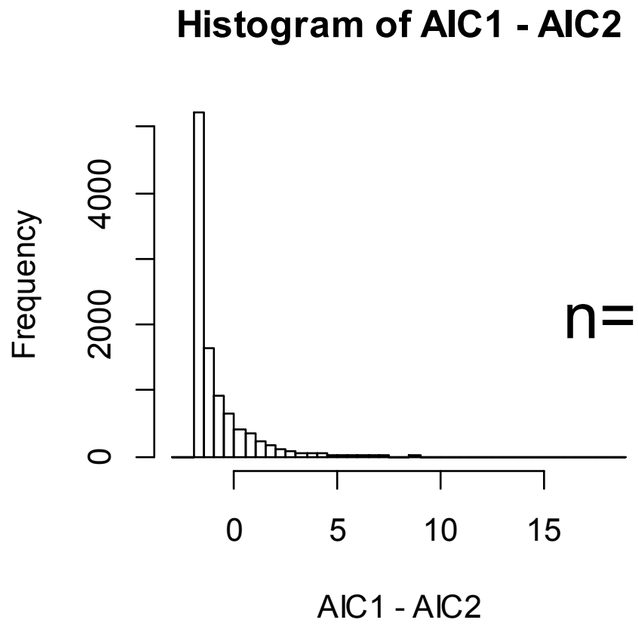
モデル１　　目的変数に説明変数の影響なし

自由なパラメーター数＝2

モデル２　　目的変数に説明変数が影響

自由なパラメーター数＝3

例２．正規分布



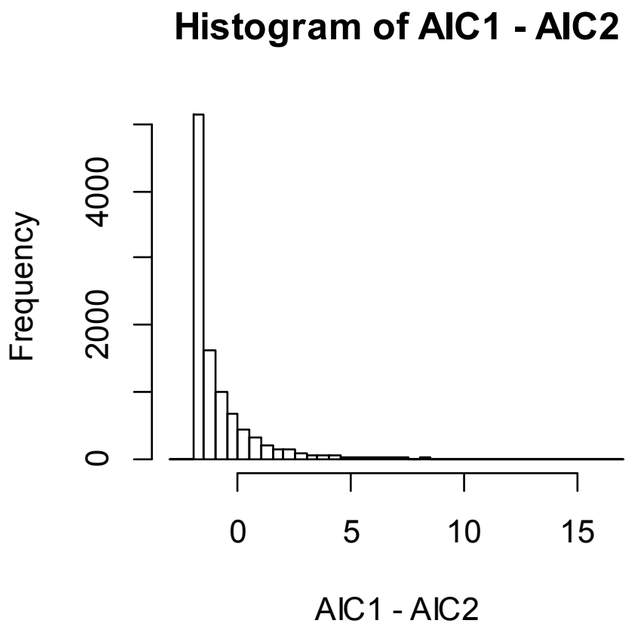
n=2000

２つのモデルのAICの差のヒストグラム

AIC１がAIC2よりも小さいのが10000回のうち8428回

正しいモデルはモデル1だから、正しいモデルを選ぶ割合は84%
正しくないモデルを選ぶ割合は16%

例．正規分布



n=200000

サンプルサイズを増やす

２つのモデルのAICの差のヒストグラム

AIC１がAIC2よりも小さいのが10000回のうち8424回

正しいモデルはモデル1だから、正しいモデルを選ぶ割合は84%

正しくないモデルを選ぶ割合は16%

・かなり大きなサンプルサイズでも正しくないモデルを15~16％選ぶ

・二項分布でも正規分布でも、AICの差のヒストグラムはよく似ている

サンプルサイズが無限大のときは理論的に計算できる

サンプルサイズが無限大のとき、正しいモデルを選ぶ確率はどちらの例でも

８４．２７０％

サンプルサイズが無限大のときは理論的に計算できる
対数尤度比統計量の分布の理論

AIC＝－２×(最大対数尤度)＋２×（自由なパラメーター数）

対数尤度比統計量＝(最大対数尤度)を２倍したものの差

制約を課した（帰無仮説に相当する）モデルが正しい場合、サンプルサイズ無限大のとき、対数尤度比統計量の分布は理論（Wilksの定理）から、自由度１のカイ２乗分布になる。そこで、ここでの正しい（真の）モデルを選ぶ確率は、サンプルサイズが大きくなると0.84270に収束する（いくらでも近づく）。

すると、サンプルサイズが無限大のとき、正しいモデルを選ぶ確率は８４．２７０％

・かなり大きなサンプルサイズでも正しくないモデルを15~16%選ぶ
　　サンプルサイズ無限大なら正しくないモデルを15.7%選ぶ
・二項分布でも正規分布でも、AICの差のヒストグラムはよく似ている

ここで計算に使った状況は単純　　一般性があるのか？

対数尤度比統計量の分布の理論はモデルが複雑でも成り立つ

正しいモデル　vs　正しいモデル＋本当は影響していないパラメーター１つ

に一般的にあてはまる

本当は効果の無い要因も候補として取り上げている場合には、同じことが言える

過去の研究

AICには「正しいモデルを選ぶ」一致性がない

一致性（consistency）

サンプルサイズ（データの量）が大きくなれば、推定値は真の値に近づく

AIC最小のモデルは、サンプルサイズが非常に大きくなれば、いつも正しい（真の）モデルか

ちがう

パラメーターを多く含むモデルを選ぶことが多い

一致性があるのは特殊な場合だけ

Shibata(1976)など

小西・北川『情報量統計学』

検定にも「正しいモデルを選ぶ」一致性がない

帰無仮説が正しい場合でも、有意水準の確率で、
誤って帰無仮説を棄てる

５％水準なら、５％

AICは正しいモデルを選ばない

AICは正しいモデルを選ぶものではないので、
正しいモデルを選ばない

もっとたくさん候補となるモデルがあったらどうなるか

AICを使う場面としてはたぶんより現実的

重回帰の説明変数選択

正規分布　　n=1000　　シミュレーション10000回

データ生成

無相関な正規変数の乱数で目的変数と説明変数を生成

正しい、真のモデル　　定数項（切片）のみ

多数の候補モデル

説明変数が4個　正しいモデルが選ばれる割合＝0.497

"定数"　4971

"x1 + 定数”　927

"x2 + 定数"　916

"x1 + x2 + 定数"　183

"x3 + 定数"　970

"x1 + x3 + 定数"　176

"x2 + x3 + 定数"　187

"x1 + x2 + x3 + 定数"　30

"x4 + 定数"　952

"x1 + x4 + 定数"　184

"x2 + x4 + 定数"　182

"x1 + x2 + x4 + 定数"　46

"x3 + x4 + 定数"　189

"x1 + x3 + x4 + 定数"　47

"x2 + x3 + x4 + 定数"　31

"x1 + x2 + x3 + x4 + 定数"　9

説明変数が5個　正しいモデルが選ばれる割合＝0.418

説明変数が6個　正しいモデルが選ばれる割合＝0.357

説明変数が7個　正しいモデルが選ばれる割合＝0.298

説明変数が8個　正しいモデルが選ばれる割合＝0.247

多数の候補モデル

重回帰の説明変数選択　　正規分布

データ生成のモデル（正しい、真）：どの説明変数も影響していない

正しい（真の）モデルが選ばれる割合は
候補モデルが増えると低下していく

どの説明変数も本当は影響していなくても、説明変数が多数あると、正しいモデルが選ばれる確率はかなり低い

AICの差に閾値を設けたらどうなるか

AICが小さくても差の絶対値が閾値を超えないときは「どのモデルも採用しない」

二項分布の例を使って計算

シミュレーション10000回

$$\mathrm{AIC} = -2 \times (\text{最大対数尤度}) + 2 \times (\text{自由なパラメーター数})$$

同じデータに対しては、
パラメーターを加えたモデルの最大対数尤度≧
パラメーターを加えていないモデルの最大対数尤度
という関係がいつでも成り立つ

n=2000のとき

閾値＝１なら、モデル１を6898回、モデル２を826回、どちらも選ばないのが2276回

閾値＝２なら、モデル１を0回、モデル２を451回、どちらも選ばないのが9549回

閾値が２以上では正しいモデルが選ばれることは皆無

n=20000のとき

閾値＝１なら、モデル１を6844回、モデル２を822回、どちらも選ばないのが2334回

閾値＝２なら、モデル１を0回、モデル２を440回、どちらも選ばないのが8560回

閾値の設定では改善しない

AICの差の閾値が2以上

・モデル１（正しい、真のモデル）はまったく選ばれない

・ 選ばれるときは、正しくないモデル

・2つのモデルが候補で、モデル間のちがいが１つのパラメーターの有無の場合

目的変数の分布をはじめ、モデルの詳細によらず、

サンプルサイズが無限大の場合、パラメーターが１つ少ないモデルが正しいとき、その正しいモデルが選ばれる確率は84.27％

パラメーターが１つ少ないモデルが正しいとき、AICの差の閾値を2以上にすると正しいモデルはまったく選ばれない

「よさ」の基準がちがえば、
「よいモデル」はちがう

AICは、正しいモデルを選ぶものではないので、正しいモデルを選ばない