生物資源モデリング

北門　利英（海洋生物資源学科）

Lecture 6 尤度比検定と遺伝的系群推測への応用

Contents

準備

- 尤度比検定

本題

- 遺伝情報を用いた系群構造解析

補足

- モデル選択

尤度比検定

尤度を利用した統計的推定：2項分布の例

ある魚種の性比を調べるためn尾をサンプリングし，
オスの数を計測した

$$Y_1, Y_2, ..., Y_n \sim (iid)\, Bin(1, p)$$

$$Y = \sum_{i=1}^{n} Y_i \sim Bin(n, p)$$

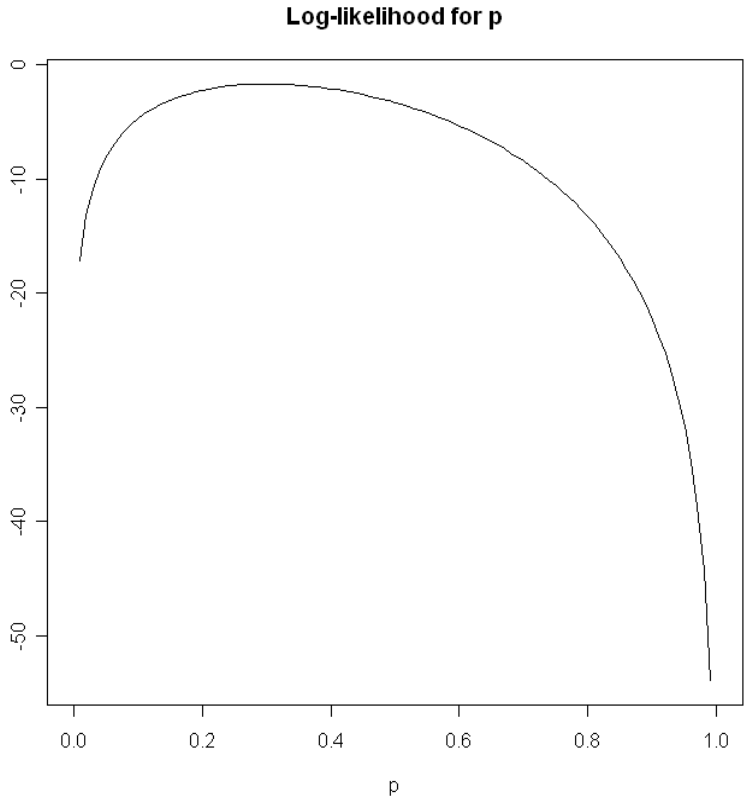
観測値の例

$$n = 20$$

$$Y = \sum_{i=1}^{n} Y_i = 6$$

$$L(p) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}$$

$$l(p) = \log L(p)$$



尤度を利用した検定

従来の知見：p=0.5 (性比一定)

$$L(p) = \frac{n!}{y!(n-y)!} p^{y} (1-p)^{n-y}$$

$$l(p) = \log L(p)$$

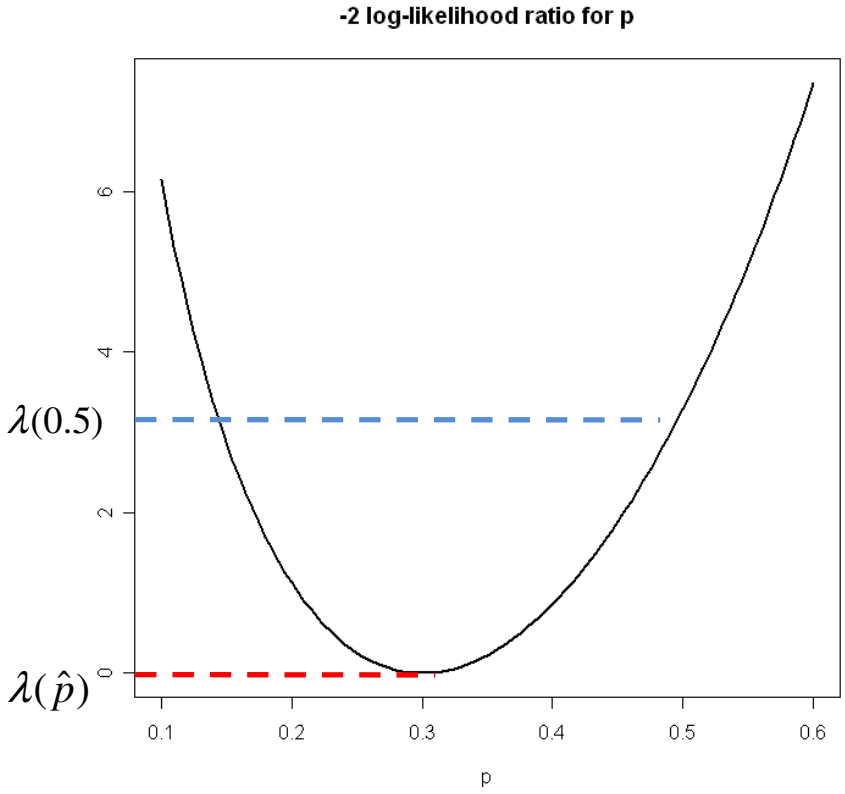
H0: p=0.5 $l(0.5) = \log L(0.5) = -3.30$

H1: p≠0.5 $l(\hat{p}) = \log L(\hat{p}) = -1.65$

$$\lambda(p) = -2\log \frac{L(p)}{L(\hat{p})}$$

H0: p=0.5 $\lambda(0.5) = 3.29$

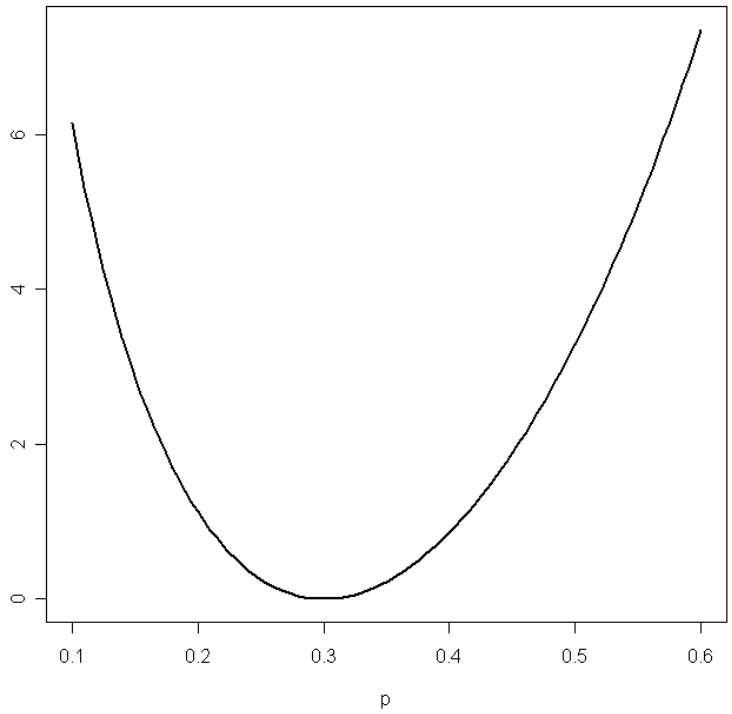
H1: p≠0.5 $\lambda(\hat{p}) = 0$



尤度比検定統計量

Tの値が大きいほど帰無仮説の尤度が相対的に小さい.
すなわち，データが帰無仮説をあまりサポートしない

$$T(p) = -2\log\frac{L(p)}{L(\hat{p})}$$



ではどれくらいTが大きければ帰無仮説が正しくないと判断してよいか？

帰無仮説の下でのTの確率分布を知る必要がある

仮説検定と確率分布

仮説検定法の本質：帰無仮説の下において，「データによる推定結果と帰無仮説の食い違い」を測る統計量の確率分布を知る必要がある

$$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1) \qquad \frac{\bar{X}-\mu}{\sqrt{\hat{\sigma}^2/n}} \sim t(n-1)$$

一般に，パラメータθに対する尤度推測を行う際，
尤度比検定統計量は帰無仮説の下で (漸近的に)
カイ2乗分布に従う

$$T(\theta) = -2\log\frac{L(\theta)}{L(\hat{\theta})} = 2\log L(\hat{\theta}) - 2\log L(\theta) \sim \chi^2(d)$$

ただし，dは対立仮説と帰無仮説のパラメータの数の差

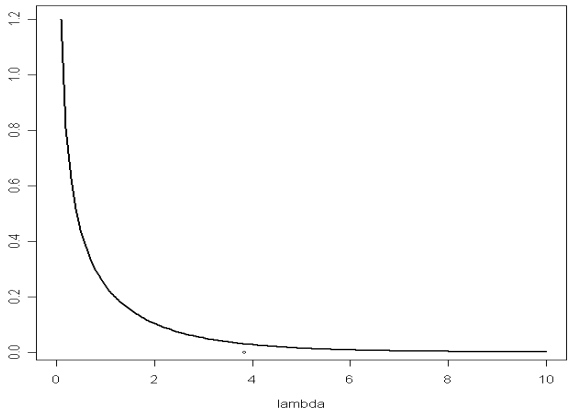
尤度比検定:2項分布の例

H0: p=p0 に対して有意水準0.05で検定

$$T(p) = -2\log\frac{L(p)}{L(\hat{p})}$$

$$T(p_0) = -2[l(p_0) - l(\hat{p})] \sim \chi^2(1)$$

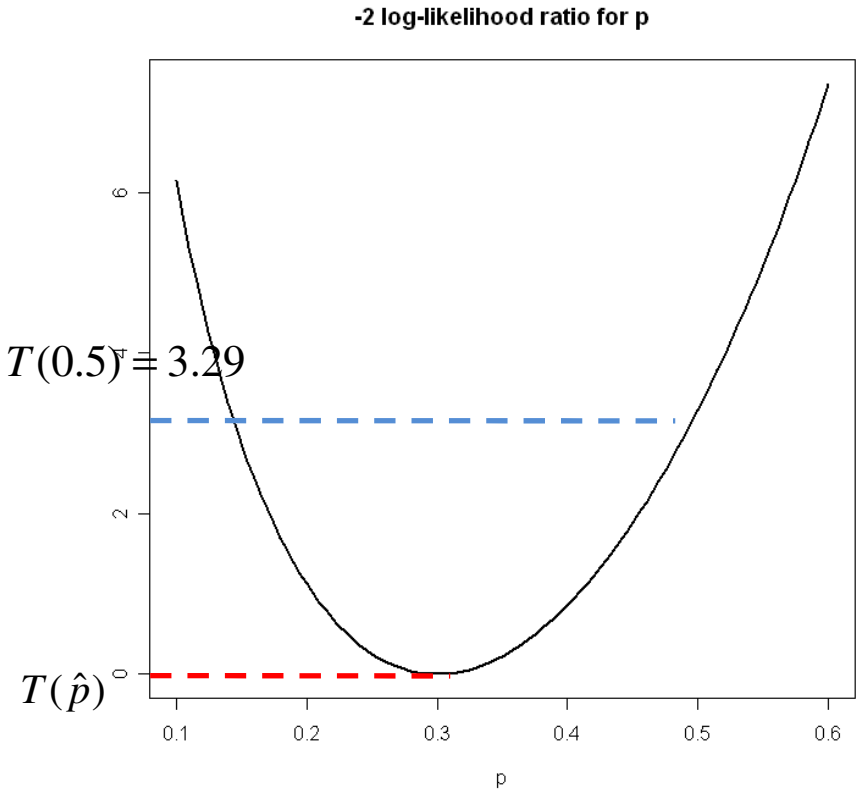
$$\Pr(T(p_0) > \chi^2(0.05;1)) = 0.05$$



したがって

$$T(p_0) > \chi^2(0.05;1) = 3.8415$$

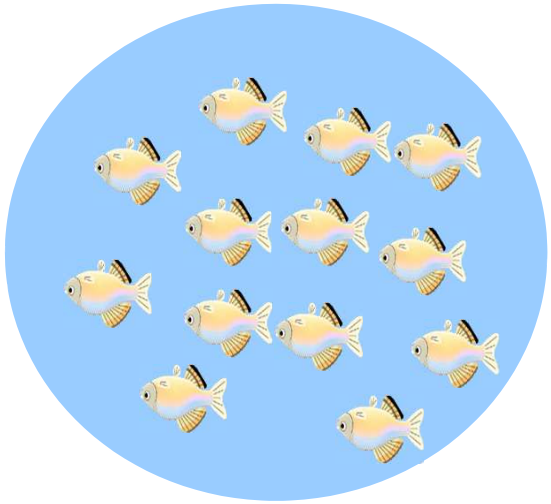
ならば帰無仮説を棄却



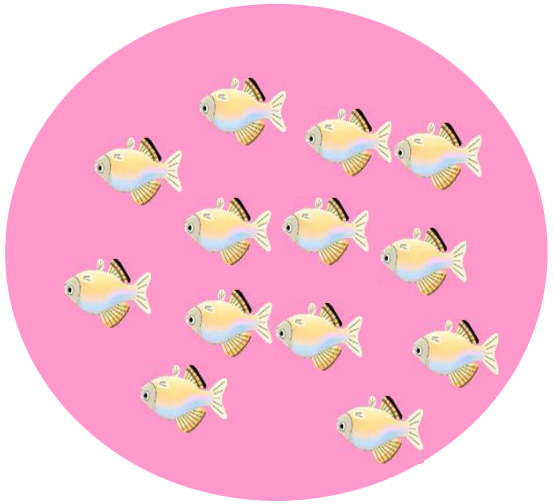
遺伝解析

系群構造

Locality 1
(Sampling Area 1)

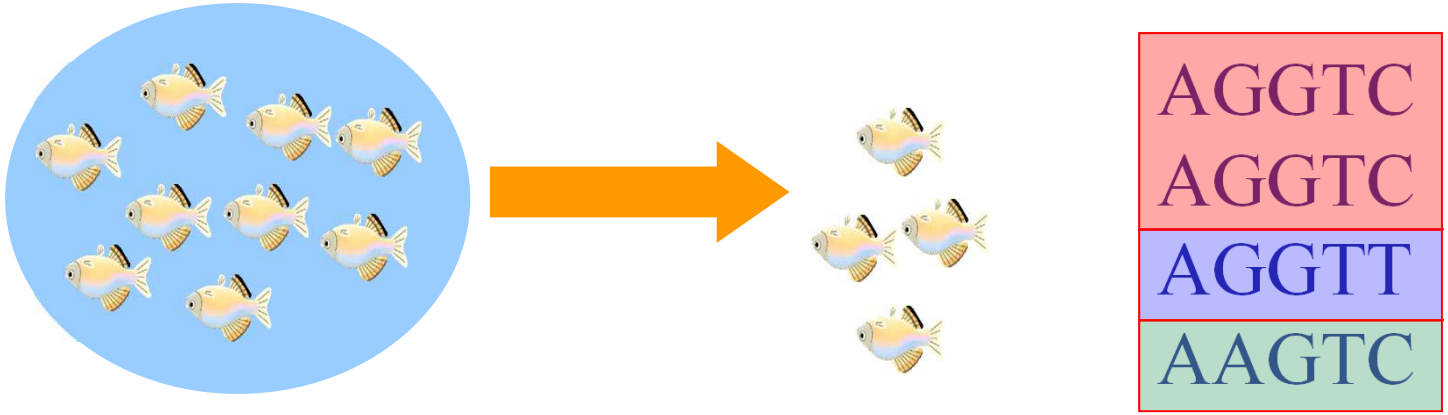


Locality 2
(Sampling Area 2)



同じ集団(系群)か？

遺伝データのサンプリング



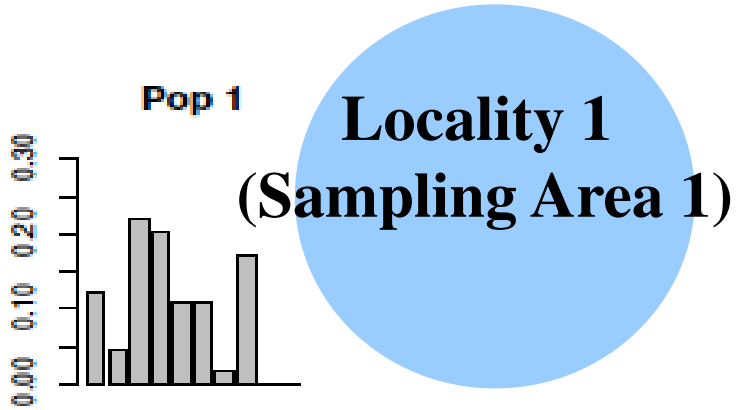
$$n_1 = (n_{11}, \ldots, n_{1J}) \sim Multi(N_1; p_{11}, \ldots, p_{1J})$$

$$(p_{11} + \cdots + p_{1J} = 1)$$

$$f(n_{11}, \ldots, n_{1J}) = \frac{N_1!}{n_{11}! \ldots, n_{1J}!} p_{11}^{n_{11}} \cdots p_{1J}^{n_{1J}}$$

2項分布の一般化

系群推測の観測データ



サンプリングエリア1の真の頻度

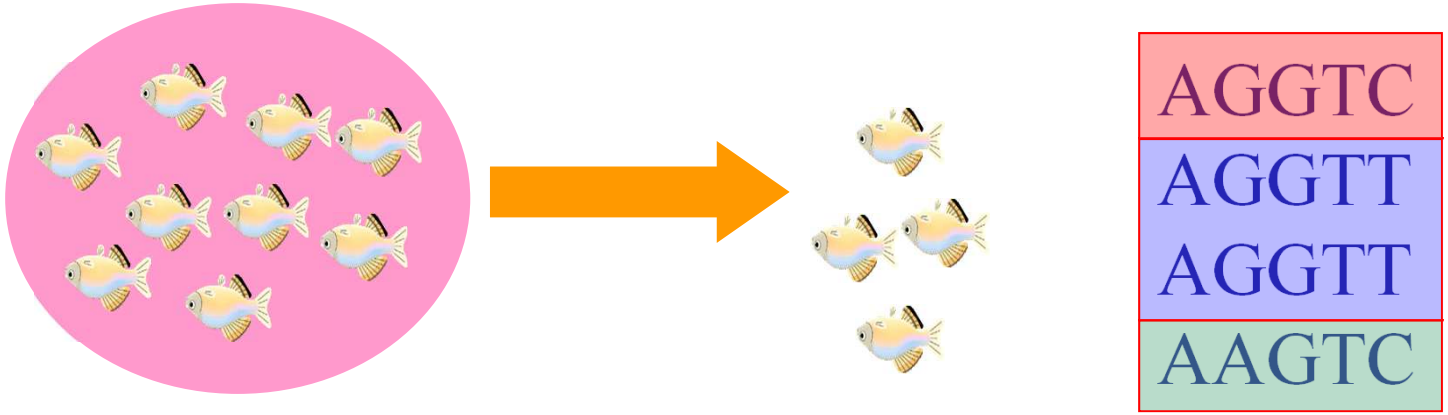
$$p_1 = (p_{11}, \ldots, p_{1J})$$

サンプリングエリア1の観測値

$$n_1 = (n_{11}, \ldots, n_{1J})$$

$$\sim Multi(N_1; p_{11}, \ldots, p_{1J})$$

遺伝データのサンプリング

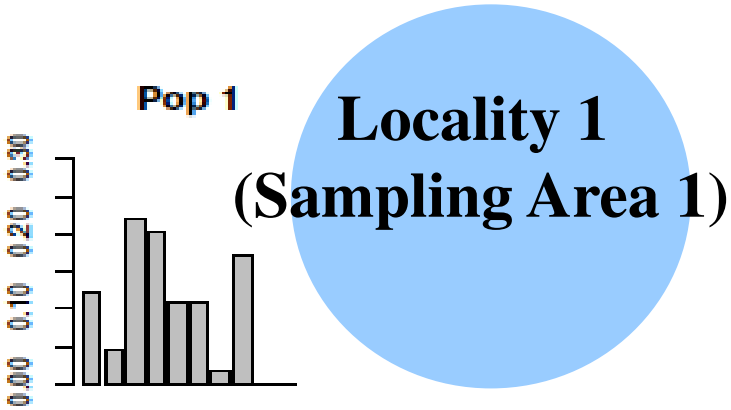


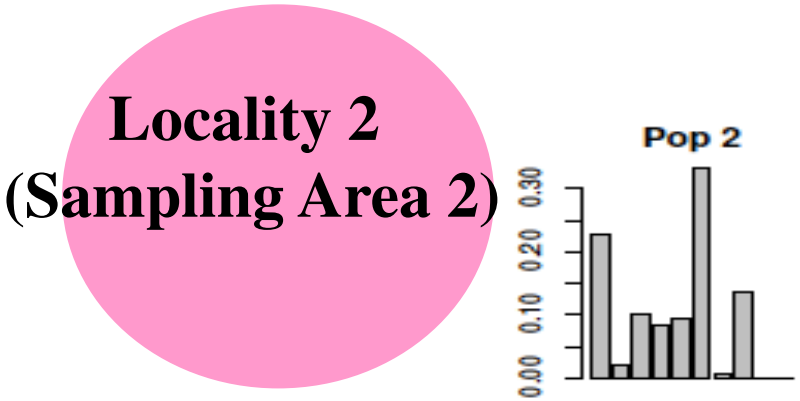
$$n_2 = (n_{21}, \ldots, n_{2J}) \sim Multi(N_2; p_{21}, \ldots, p_{2J})$$

$$(p_{21} + \cdots + p_{2J} = 1)$$

$$f(n_{21}, \ldots, n_{2J}) = \frac{N_2!}{n_{21}! \ldots, n_{2J}!} p_{21}^{n_{21}} \cdots p_{2J}^{n_{2J}}$$

モデルの要約





サンプリングエリア1の真の頻度

$$p_1 = (p_{11}, \ldots, p_{1J})$$

サンプリングエリア1の観測値

$$n_1 = (n_{11}, \ldots, n_{1J})$$

$$\sim Multi(N_1; p_{11}, \ldots, p_{1J})$$

サンプリングエリア2の真の頻度

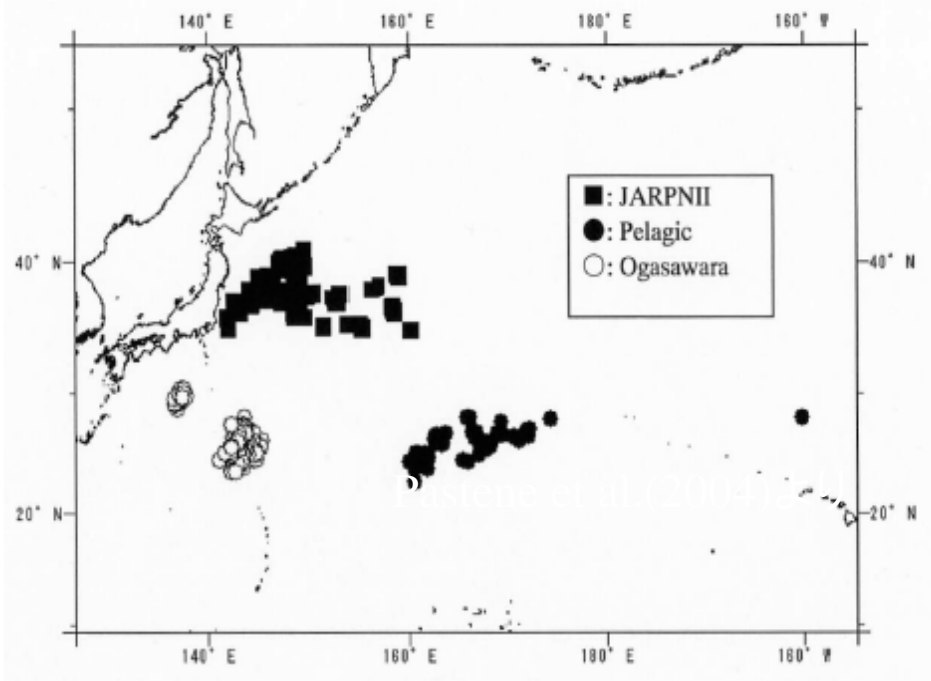
$$p_2 = (p_{21}, \ldots, p_{2J})$$

サンプリングエリア2の観測値

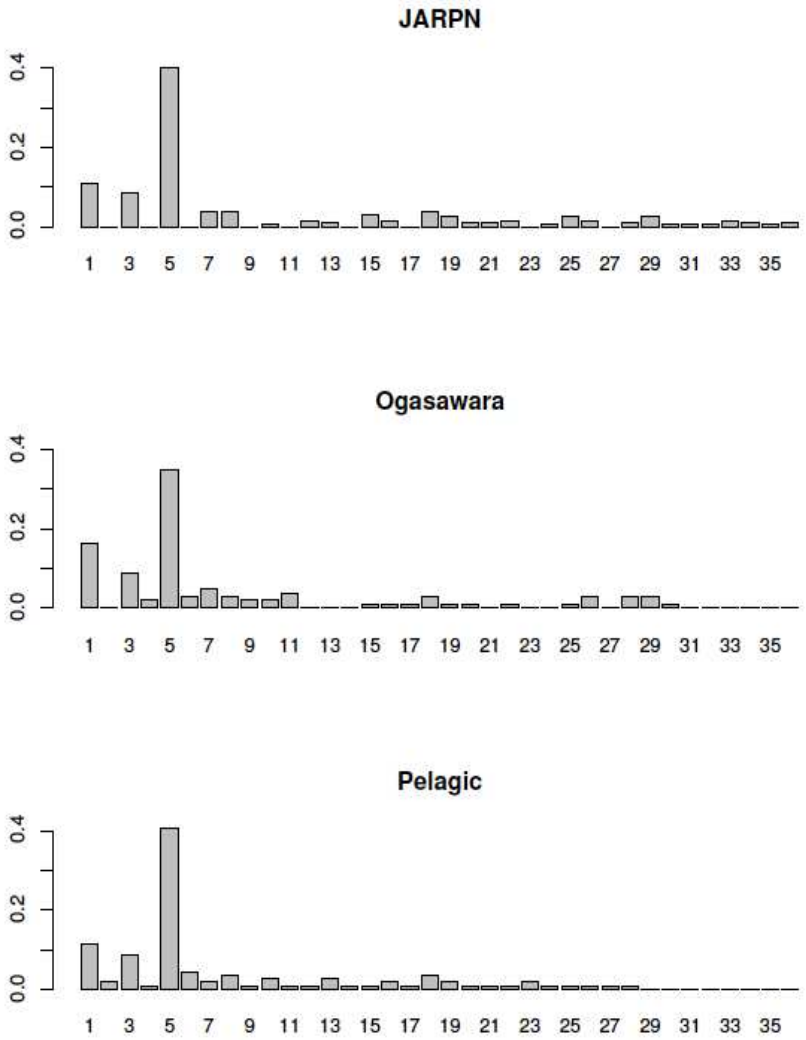
$$n_2 = (n_{21}, \ldots, n_{2J})$$

$$\sim Multi(N_2; p_{21}, \ldots, p_{2J})$$

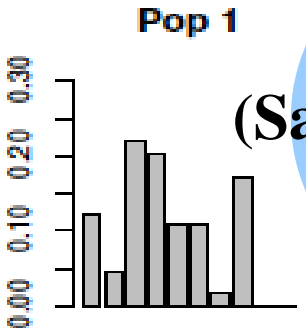
例：北西太平洋ニタリクジラの遺伝データ



ミトコンドリアDNAのハプロタイプ頻度

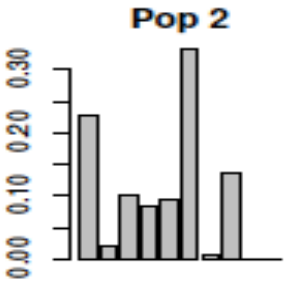


仮説検定による系群の同一性検証



Locality 1
(Sampling Area 1)

Locality 2
(Sampling Area 2)



$$p_1 = (p_{11}, \ldots, p_{1J})$$

$$p_2 = (p_{21}, \ldots, p_{2J})$$

Testing $H_0 : p_1 = p_2 \quad vs \quad H_1 : p_1 \neq p_2$

尤度比検定を利用してみよう！

仮説検定による系群の同一性検証

一般に，パラメータθに対する尤度推測を行う際，
尤度比検定統計量は帰無仮説の下で（漸近的に）
カイ２乗分布に従う

$$T = -2\log\frac{L(\theta)}{L(\hat{\theta})} = -2[l(\theta) - l(\hat{\theta})] \sim \chi^2(d)$$

ただし，dは対立仮説と帰無仮説の未知パラメータ数の差

帰無仮説が正しいとき

カイ２乗分布の上側5%点

$$\Pr(T > \chi^2(0.05; d)) = 0.05$$

よって，次が成り立つとき帰無仮説を棄却する

$$T > \chi^2(0.05; d)$$

尤度比検定統計量

$$T = -2\log\frac{L_1(\hat{\theta}_1)}{L_2(\hat{\theta}_2)} = -2\log\frac{\text{帰無仮説の下での最大尤度}}{\text{対立仮説の下での最大尤度}}$$

$$T = -2\log\frac{\dfrac{N_1!}{n_{11}!\ldots,n_{1J}!}\tilde{p}_1^{\,n_{11}}\cdots\tilde{p}_J^{\,n_{1J}}\ \dfrac{N_2!}{n_{21}!\ldots,n_{2J}!}\tilde{p}_1^{\,n_{21}}\cdots\tilde{p}_J^{\,n_{2J}}}{\dfrac{N_1!}{n_{11}!\ldots,n_{1J}!}\hat{p}_{11}^{\,n_{11}}\cdots\hat{p}_{1J}^{\,n_{1J}}\ \dfrac{N_2!}{n_{21}!\ldots,n_{2J}!}\hat{p}_{21}^{\,n_{21}}\cdots\hat{p}_{2J}^{\,n_{2J}}}$$

$$\tilde{p}_j = \frac{n_{1j}+n_{2j}}{N_1+N_2}$$ 帰無仮説の下での最尤推定量

$$\hat{p}_{1j} = \frac{n_{1j}}{N_1},\ \hat{p}_{2j} = \frac{n_{2j}}{N_2}$$ 対立仮説の下での最尤推定量

χ^2分布の自由度に注意：Ｈ１とＨＯのパラメータ数の差

課題１

2つのローカリティー（たとえば沿岸と沖合）から50匹ずつサンプリングし，それぞれからミトコンドリアDNAの遺伝情報をデータとして得た．ローカリティーAからはハプロタイプ1を持つものが12個体，ハプロタイプ2を持つものが38個体観測され，一方でローカリティーBからはハプロタイプ1を持つものが20個体，ハプロタイプ2を持つものが30個体観測された．

この遺伝情報を基に，2つのローカリティーに生息する集団が等しいかどうかについて，尤度比検定を利用して判断せよ．ただし，2つのローカリティーの真のハプロタイプ頻度が等しいとき，2つのローカリティーに生息する集団はひとつの繁殖単位をなしていると考える．

課題2

仙台湾～東京湾，そして伊勢湾・三河湾の2つのローカリティーからスナメリのミトコンドリアの標本を以下のように得た(吉田, 2003)

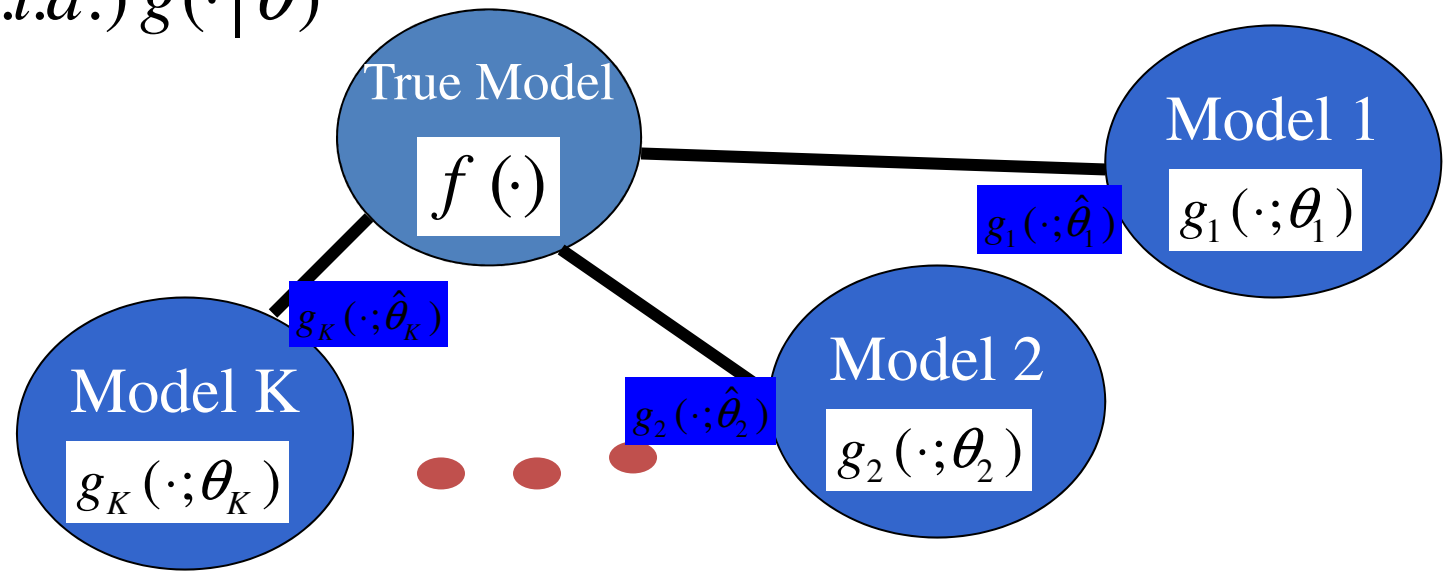
	ハプロタイプ		
	1	2	3
仙台湾～東京湾	7	7	0
伊勢湾・三河湾	0	4	52

このとき，2つのローカリティーの集団が同じ系群かどうか，尤度比検定を利用して判断せよ.

補足：モデル選択

AIC-1

$$Y_1,\ldots,Y_n \sim (i.i.d.)\, g(\cdot \mid \theta)$$



$$D(\hat{f}, g(\cdot \mid \hat{\theta})) = \int \hat{f}(z) \log \hat{f}(z)\, dz - \int \hat{f}(z) \log g(z \mid \hat{\theta})\, dz$$

$$= \text{constant} - \frac{1}{n} \sum_{i=1}^{n} \log g(y_i \mid \hat{\theta})$$

AIC-2

$$D(\hat{f}, g(\cdot|\hat{\theta})) = \int \hat{f}(z) \log \hat{f}(z)\, dz - \int \hat{f}(z) \log g(z|\hat{\theta})\, dz$$

$$= \text{constant} - \sum_{i=1}^{n} \log g(y_i|\hat{\theta})$$

ここで問題：

１．上記はあるデータが測定されたときの距離の推定値である．

２．しかし，この方法だと真の分布 f の推定とパラメータΘの推定に同じデータを2度使っている．このことにより，対数尤度の最大値で真のモデルの相対的距離を測るとバイアスが生じてしまう．．．．

などいろいろ理由があり，またTaylor展開を上手に駆使し漸近正規性を利用することで，上記のバイアス補正が可能

$$AIC = -2 \log L(\hat{\theta}) + 2K \Rightarrow \text{最小化}$$

AIC-3

$$AIC = -2 \log L(\hat{\theta}) + 2K$$

注意：AICは真のモデルと仮定するモデルの相対的距離の推定値である．あくまでも「推定値」なのです．

「推定値」というからには「誤差」があります．

ですから，AICは判断を間違うことだってあるのです．

データから推測，判断をする以上，誤差や誤りがつきものです．