# 一般化線形モデル

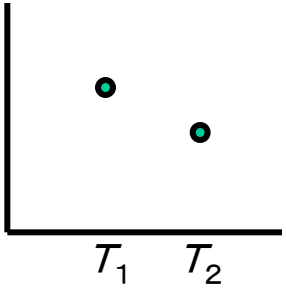
農研機構 農業環境変動研究センター 環境情報基盤研究領域
統計モデル解析ユニット，山村光司

- 一般線型モデルについて
- 一般化線形モデルとは
- （少し寄り道）Pseudo-replication（偽反復）の問題
- 一般化線形モデルの誤用の傾向とその対策（偽反復と関係）
- モデル選択（$R_D$指数）

# 一般線形モデル(General Linear Model)
# 線形モデル(Linear Model)

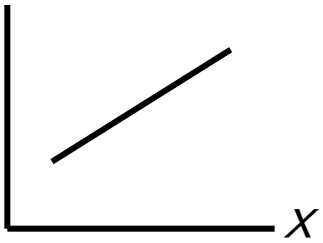
**分散分析** ANOVA

**カテゴリー要因の影響を分析**



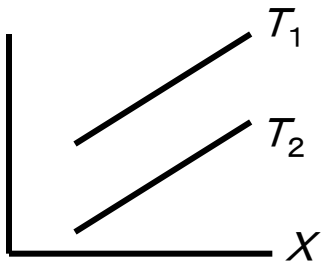

**回帰分析** Regression

**量的要因の影響を分析**




**共分散分析** ANCOVA

**量的要因とカテゴリー要因の影響を分析(一つずつの場合)**




**線型モデル**
Linear model
**一般線形モデル**
General Linear Model

# 単回帰分析

式で書くと

$$y_i = a + bx_i + e_i \quad (i = 1,2,\ldots,n)$$

ただし $e_i$ は等分散の正規分布にしたがう

これは次のようにも書ける

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

→ひとまとめにして書くと

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

# 重回帰分析

**2変量の場合，式で書くと**

$$y_i = a + b_1 x_{i1} + b_2 x_{i2} + e_i \quad (i = 1,2,\ldots,n)$$

**ただし $e_i$ は等分散の分布にしたがう**

**これは次のようにも書ける**

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

**→ひとまとめにして書くと**

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

# 一元配置分散分析

（例）三つの水準を設けて，完全無作為法で
それぞれの処理を2回ずつ反復した場合

$$y_{ij} = T_i + e_{ij} \quad (i = 1,2,3; j = 1,2)$$

ただし$e_{ij}$ は等分散の正規分布にしたがう

全体の平均 $\mu$ をくくり出して表現すると

$$y_{ij} = \mu + T_i + e_{ij} \quad (i = 1,2,3; j = 1,2)$$

これは

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ T_1 \\ T_2 \\ T_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix}$$

→ひとまとめにして書くと

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

以上のすべては次のような式に書けた

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

y：従属変数のベクトル
X：デザイン行列（実験計画などに依存して決まる）
b：パラメーターのベクトル
e：誤差のベクトル（等分散正規分布にしたがう）

これを線型モデル，あるいは一般線型モデルと呼ぶ

同じ式に書けるため，推定・検定の手順は基本的に同じ

→回帰分析と分散分析は区別する必要がない。

→回帰分析と分散分析が混合した分析も可能。

# 分散分析と回帰分析の混合例

四国農試のハスモンヨトウ誘殺実験データ

| 地域 | トラップ番号 | 各月の総誘殺数 | | | |
|---|---|---|---|---|---|
| | | 5月 | 6月 | 7月 | 8月 |
| A | 1 | 10 | 26 | 45 | 356 |
| A | 2 | 8 | 16 | 55 | 341 |
| B | 3 | 16 | 48 | 112 | 874 |

トラップ間で誘殺数の対数値に差があるかどうかを検定（カテゴリー変数，名義変数）→[分散分析的]

月とともに誘殺数が増加するかどうかを検定（量的変数）→[回帰分析的]

```
cat(file="Hasumon.txt",
"Obs Area Trap Month Y
1   A    T1    5    10
2   A    T1    6    26
3   A    T1    7    45
4   A    T1    8    356
5   A    T2    5    8
6   A    T2    6    16
7   A    T2    7    55
8   A    T2    8    341
9   B    T3    5    16
10  B    T3    6    48
11  B    T3    7    112
12  B    T3    8    874
")
```

（↑スペースで区切ること）

```
Hasumon <- read.table("Hasumon.txt", header=TRUE)
Hasumon.lmlog <- lm(log(Y+0.5)~ Trap+Month, data=Hasumon)
```

計算結果がHasumon.lmlogという名前のlmオブジェクトに保存される。

# パラメーター推定値の表示

```
summary(Hasumon.lmlog)
```

サマリー関数からの出力は

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.0042     0.6914  -5.791 0.000409 ***
TrapT2       -0.1324     0.2792  -0.474 0.648123
TrapT3        0.7147     0.2792   2.560 0.033643 *
Month         1.2054     0.1019  11.824  2.4e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
  0.1 ‘ ’ 1
```

この推定値の意味については後で詳しく議論する

# 検定を実行（重要な要因だけをモデルに組み込むため）

## 「分散分析表」の出力（回帰分析も含む「広義の分散分析」）

```
anova(Hasumon.lmlog) # 分散分析表の出力
Analysis of Variance Table
Response: log(Y + 0.5)
           Df  Sum Sq Mean Sq  F value  Pr(>F)
Trap        2  1.6613  0.8306   5.3283 0.03381 *
Month       1 21.7931 21.7931 139.7981 2.4e-06 ***
Residuals   8  1.2471  0.1559
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
  0.1 ‘ ’ 1
```

トラップ効果に関する有意水準 $P = 0.03$（分散分析部分）
→ $P$ 値が 0.05 より小さいので「有意」と判断される。
月の効果に関する有意水準 $P = 2.4e\text{-}06$（回帰分析部分）
→ $P$ 値が 0.05より小さいので「有意」と判断される。

例おわり

# パラメーター推定の原理

- 最尤推定法とは：
  - ✓ 手元の観測データが得られる確率が最大になるようにパラメーターを決定する方法。
  - ✓ この確率$L$を「尤度」とよぶ。

固定効果パラメーター

例：正規分布に従う場合（平均$\mu$, 分散$\sigma^2$）

dispersion パラメーター

この正規分布から 観測値$y$を得たとき，
それが起こる確率（尤度）は

$$L = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

いま平均値$\mu$の推定を考える（分散$\sigma^2$は既知）。

# パラメーター$\mu$の推定 二つの観測値, 3, 5を得たとき, 尤度は

$$L = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(3-\mu)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(5-\mu)^2}{2\sigma^2}\right)$$

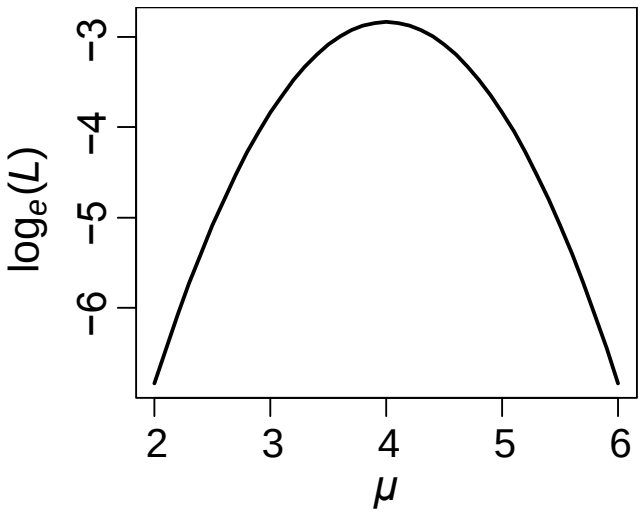
$$\rightarrow \quad L = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(3-\mu)^2 + (5-\mu)^2}{2\sigma^2}\right)$$

尤度の対数を考えると

$$\log(L) = -\frac{(3-\mu)^2 + (5-\mu)^2}{2\sigma^2} - \log(2\pi\sigma^2)$$

尤度$L$を最大化することは対数尤度$\log(L)$を最大化することと同じ

最大点では傾きがゼロ → 最大とする$\mu$は$\log(L)$を$\mu$で微分して0とおくことによって求められる。



# パラメーターμの推定 二つの観測値, 3, 5を得たとき, 尤度は

対数尤度

$$\log(L) = -\frac{(3-\mu)^2 + (5-\mu)^2}{2\sigma^2} - \log(2\pi\sigma^2)$$

μで微分してゼロとおくと $\frac{2(3-\mu) + 2(5-\mu)}{2\sigma^2} = 0$

これを解くと最尤推定値はμ=4 つまり, 平均値と等しい。

今の場合には対数尤度の分子部分を最小化すればよい。二乗和を最小化する方法なので, これは「最小二乗法」である。つまり
**正規分布誤差の場合は「最尤推定法=最小二乗法」。**
線形モデルの場合は最小二乗法のための計算公式がある。

# 最小二乗推定値を簡単に見つける方法

線形モデル $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$

転置行列　逆行列

→ 行列計算で求まる $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

ただし…,
逆行列が決まるためにはXの列が線形独立でなければならない
（列の要素を足しても他の列にはならない）

回帰分析型

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

→OK ！

分散分析型

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ T_1 \\ T_2 \\ T_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix}$$

→ダメ！

# 分散分析型の部分には「制約条件」をつけてやる

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ T_1 \\ T_2 \\ T_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix} \quad y_{ij} = \mu + T_i + e_{ij} \quad (i = 1,2,3; j = 1,2)$$

**いずれにせよ，列の数が一つ減る。**
**線形独立な列の数は「自由度」とよばれる。**
**→このため，分散分析では自由度は一つ減る**

**ゼロ和制約：** $T_1 + T_2 + T_3 = 0$
**（三つの処理の平均が$\mu$である）**
$T_3 = -T_1 - T_2$**で** $T_3$**を置き換える。**

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ T_1 \\ T_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix}$$

**→OK ！ JMP**

**端点制約：たとえば**$T_1 = 0$
**（**$T_1$**の処理が**$\mu$**である）**
**行列の変形は楽。**

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ T_2 \\ T_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix}$$

**→OK ！ SAS, R**
**（S-PLUSも）**

# Rのデフォルト設定で使用されているデザイン行列を見る

```
model.matrix( ~ Trap+Month, data=Hasumon)
```

推定されるパラメーター

| | (Intercept) | TrapT2 | TrapT3 | Month |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 5 |
| 2 | 1 | 0 | 0 | 6 |
| 3 | 1 | 0 | 0 | 7 |
| 4 | 1 | 0 | 0 | 8 |
| 5 | 1 | 1 | 0 | 5 |
| 6 | 1 | 1 | 0 | 6 |
| 7 | 1 | 1 | 0 | 7 |
| 8 | 1 | 1 | 0 | 8 |
| 9 | 1 | 0 | 1 | 5 |
| 10 | 1 | 0 | 1 | 6 |
| 11 | 1 | 0 | 1 | 7 |
| 12 | 1 | 0 | 1 | 8 |

分散分析部分　　回帰分析部分

この端点制約で推定されるパラメーターは四つ

- トラップ1
- トラップ1とトラップ2の差
- トラップ1とトラップ3の差
- 月の効果

端点を変えたい場合には relevel()を用いる。たとえばトラップ3を端点にするときは

```
Hasumon$Trap <- relevel(Hasumon$Trap, ref="T3")
```

# 端点制約とゼロ和制約での推定値の違い

## R の出力(端点制約)

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -4.0042 | 0.6914 | -5.791 | 0.000409 | *** |
| TrapT2 | -0.1324 | 0.2792 | -0.474 | 0.648123 | |
| TrapT3 | 0.7147 | 0.2792 | 2.560 | 0.033643 | * |
| Month | 1.2054 | 0.1019 | 11.824 | 2.4e-06 | *** |

推定されるパラメーターは異なる

## JMP の出力(ゼロ和制約)

| 項 | 推定値 | 標準誤差 | t値 | P |
|---|---|---|---|---|
| 切片 | -3.8100 | 0.6724 | -5.6700 | 0.0005 |
| トラップ[1] | -0.1941 | 0.1612 | -1.2000 | 0.2629 |
| トラップ[2] | -0.3265 | 0.1612 | -2.0300 | 0.0774 |
| 月 | 1.2054 | 0.1019 | 11.8200 | <.0001 |

(ゼロ和制約では切片は回帰分析の場合の切片と同じ)

パラメーターの推定値自体の意味は複雑

# 水準の推定値で議論（有意となった要因だけでよい）

推定された平均値の出力

```
library(lsmeans)
lsmeans(Hasumon.lmlog, ~ Trap)
```

95%信頼限界

```
$`Trap lsmeans`
 Trap    lsmean         SE df lower.CL upper.CL
   T1 3.830642 0.1974146  8 3.375403 4.285880
   T2 3.698289 0.1974146  8 3.243051 4.153528
   T3 4.545382 0.1974146  8 4.090144 5.000621
```

95%信頼限界（2.5%下側信頼限界と2.5%上側信頼限界）

「真の値がこの範囲の外側にあったならば，このようなデータは5%以下の確率でしか生じない」という範囲

→ 真の値はおそらくこの範囲の中にあるだろう，と判断できる。

推定の原理おわり

# 検定の原理

- 三種類の検定
  スコア検定, Wald検定, 尤度比検定。

  統計解析ソフトでデフォルトで出力されるのはWald検定だが,
  尤度比検定の方がカイ二乗近似が良い。

- 尤度比検定の基本コンセプト:
  あるパラメーターを除いて当てはめる。

  →当てはまり具合が悪くなる。つまり, 対数尤度が低下する。

  もしパラメーターが重要だったならあてはまり具合が大きく悪くなるはず。つまり対数尤度が大きく低下するはず

  ⟶ 対数尤度の低下の大きさから検定ができる

  「対数尤度の差」は「尤度の比の対数」なので, これを
  「尤度比検定」や「尤度比カイ二乗検定」とよぶ。

Wald検定やスコア検定では片方の値から外挿して計算するので計算は速いが精度は劣る。

# パラメーターを抜く順番に関して，いくつかの検定の種類

## 要因効果を抜いて尤度比検定を行う際の抜く順序の違い

- **Type I**
  Rでは anova()
  要因間に優劣がある場合に用いる。
  後ろから順番に要因を抜いてゆき，
  そのときの対数尤度の低下から検定する。

- **Type II**
  Rでは drop1() で工夫，あるいは library(car) の Anova()
  An R Companion to Applied Regression
  交互作用を含めた検定の場合に用いる。
  交互作用をまず抜いて，そのときの対数尤度の低下から検定
  主効果の検定の際には，交互作用だけを抜いたモデルから
  その主効果だけを抜いたときの対数尤度の低下から検定する。

- **Type III**
  Rでは library(car) の Anova() でオプション指定
  ある要因の検定の際には，その要因だけを抜いたときの
  対数尤度の低下から検定する。

検定の原理おわり

# この原理を踏まえてデータを再解析する

| 地域 | トラップ番号 | 各月の総誘殺数 | | | |
|---|---|---|---|---|---|
| | | 5月 | 6月 | 7月 | 8月 |
| A | 1 | 10 | 26 | 45 | 356 |
| A | 2 | 8 | 16 | 55 | 341 |
| B | 3 | 16 | 48 | 112 | 874 |

多くの情報を持っている。地域間差, 地域内のトラップ間差,
月に伴う増加が加速的か減速的か, など。
これらをフル活用するべき。

```
Hasumon.lmlogfull <- lm(log(Y+0.5) ~ Area + Trap +
  I(Month) + I(Month^2) +  I(Month^3), data=Hasumon)
anova(Hasumon.lmlogfull)
```

# 後半部分は先週の講義の「多項式回帰」の部分。「I()」でくくるのを忘れないように（inhibit関数）。

# データをフル活用した一般線型モデルの結果

```
Response: log(Y + 0.5)
             Df  Sum Sq Mean Sq  F value    Pr(>F)
Area          1  1.6262  1.6262  57.7540 0.0002702 ***
Trap          1  0.0350  0.0350   1.2442 0.3073317
I(Month)      1 21.7931 21.7931 773.9680 1.427e-07 ***
I(Month^2)    1  0.8851  0.8851  31.4337 0.0013725 **
I(Month^3)    1  0.1931  0.1931   6.8569 0.0396623 *
Residuals     6  0.1689  0.0282
```

3次項は1次項と2次項で説明できない成分を示し，2次項は1次項で説明できない成分を示すため，この部分は Type I が妥当。したがって，anova 関数で大丈夫。**分析の結果，Trapは有意でなかったのでモデルから除く。**

定石にしたがって，有意差が出た成分については平均を計算するかグラフを描く。定量要因についてはグラフを描く。

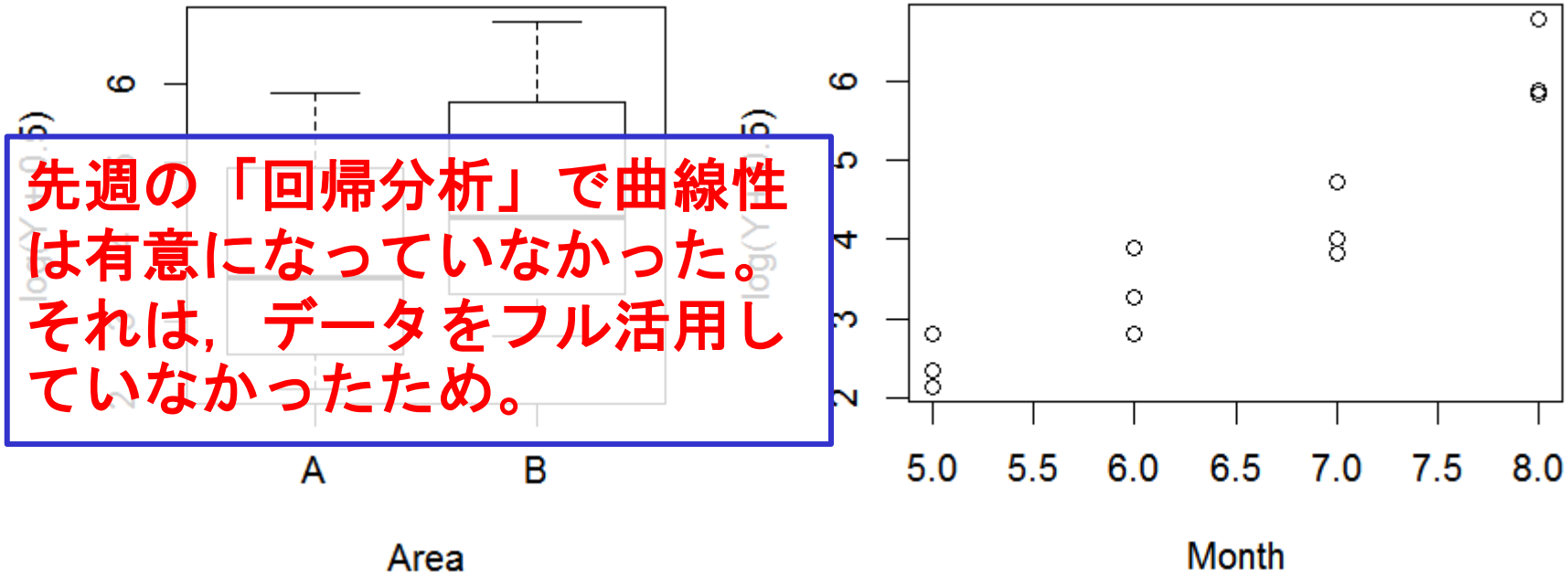
```
plot(log(Y+0.5)~Area,data=Hasumon)
plot(log(Y+0.5)~Month,data=Hasumon)
```

# データをフル活用した一般線型モデルの結果

曲線性の検定

```
Response: log(Y + 0.5)
             Df  Sum Sq Mean Sq  F value    Pr(>F)
Area          1  1.6262  1.6262  57.7540 0.0002702 ***
Trap          1  0.0350  0.0350   1.2442 0.3073317
I(Month)      1 21.7931 21.7931 773.9680 1.427e-07 ***
I(Month^2)    1  0.8851  0.8851  31.4337 0.0013725 **
I(Month^3)    1  0.1931  0.1931   6.8569 0.0396623 *
Residuals     6  0.1689  0.0282
```

先週の「回帰分析」で曲線性は有意になっていなかった。それは，データをフル活用していなかったため。



# 一般化線形モデル(Generalized Linear Model)

Nelder and Wedderburn (1972)

**線型モデル**
Linear model
**一般線形モデル**
General Linear Model

**一般化線形モデル**
Generalized Linear Model

**誤差が等分散正規分布と仮定**

**ロジスティック回帰** Logistic regression

**誤差が二項分布と仮定**

**対数線形モデル loglinear model**

**誤差が多項分布と仮定**

# 一般化線型モデルでは二つの拡張

線型モデル　$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$　e は等分散の正規分布にしたがう

これを言い換えると

(1) 観測値yは期待値$E(\mathbf{y})$の周りに正規分布にしたがって分布
(2) 期待値$E(\mathbf{y})$ に関して$E(\mathbf{y}) = \mathbf{Xb}$

この二つをそれぞれ拡張する

(1) 観測値yは$E(\mathbf{y})$の周りに正規分布で分布する必要はなく,
「指数分布族の分布」にしたがって分布していればよい。
(ポアソン分布, 二項分布, ガンマ分布などを含む)

(2) 期待値$E(\mathbf{y})$に関して$E(\mathbf{y}) = g(\mathbf{Xb})$
$g()$は指数関数やロジスティック関数などの「単調関数」

$g()$の逆関数を**リンク関数**と呼ぶ。これを$f$と記すと$f[E(\mathbf{y})] = \mathbf{Xb}$
期待値$E(\mathbf{y})$をリンク関数で変換したものが線形関数

# 一般化線型モデルの例 1

(1) 観測値yは$E(\mathbf{y})$の周りに二項分布にしたがって分布

(2) 期待値$E(\mathbf{y})$に関して

$$E(\mathbf{y}) = \frac{\exp(\mathbf{Xb})}{1 + \exp(\mathbf{Xb})}$$

**Xbをロジスティック関数で変換したもの**

**ロジスティック関数の逆関数はロジット関数：**
**→ リンク関数がロジット関数**

$$\log_e \left( \frac{E(\mathbf{y})}{1 - E(\mathbf{y})} \right) = \mathbf{Xb}$$

**これはロジスティック回帰**

# 一般化線型モデルの例 ２

(1) 観測値yは$E(\mathrm{y})$の周りにポアソン分布にしたがって分布

(2) 観測値$E(\mathrm{y})$に関して$E(\mathbf{y}) = \exp(\mathbf{Xb})$

$$\log_e[E(\mathbf{y})] = \mathbf{Xb}$$

リンク関数は対数関数

→ これはポアソン回帰の一種

# 最尤推定値を見つける方法は…

誤差に正規分布を仮定する場合（線形モデル）は

行列計算で求まった　$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

正規分布以外の誤差の場合はそう簡単には求まらない

- ✓ 試行錯誤法（ニュートン・ラフソン法，スコア法など）
- ✓ 対数尤度が最大となるパラメーターを探す。

# (計算例)ロジスティック回帰
## (比率の場合，二項分布の場合)

アブラムシと葉の穴との関係についての実験(Crawley, 1993)

「誘導抵抗性」に関する実験。アブラムシが存在すると，木が危険性を察知して有害物質を生産する→後の季節に別の昆虫（イモムシ類）に食われにくくなる。

アブラムシが存在すると，後に葉に穴が出来やすいか否かを調べた。

| 木番号 | アブラムシ | 穴あり葉 | 穴なし葉 | 合計 |
|---|---|---|---|---|
| 1 | 無 | 35 | 1750 | 1785 |
| 1 | 有 | 23 | 1146 | 1169 |
| 2 | 無 | 146 | 1642 | 1788 |
| 2 | 有 | 30 | 333 | 363 |

木の要因とアブラムシの有無の要因の影響を調べる

# Rによるロジスティック回帰

```
cat(file="aphid.txt","
tree aphid   r    non
一番目   無    35   1750
一番目   有    23   1146
二番目   無   146   1642
二番目   有    30    333
")
Aphid <- read.table("aphid.txt", header=TRUE)
Aphid$d <- cbind(Aphid$r,Aphid$non) #(反応数,非反応数)の形
Aphid.Fit1 <- glm(d ~ aphid + tree,
                  family=binomial, data=Aphid)

anova(Aphid.Fit1, test="Chisq") # これは間違い例:
# anova 関数は Type I なので, 二元配置実験の場合は使わないこと。

# car ライブラリの Anova 関数を使うと Type II 検定ができる
library(car)
Anova(Aphid.Fit1)
```

| 木番号 | アブラムシ | 穴あり葉 | 穴なし葉 | 合計 |
|---|---|---|---|---|
| 1 | 無 | 35 | 1750 | 1785 |
| 1 | 有 | 23 | 1146 | 1169 |
| 2 | 無 | 146 | 1642 | 1788 |
| 2 | 有 | 30 | 333 | 363 |

# Typeによる結果の違い

```
anova(Aphid.Fit1, test="Chisq")
           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          3    110.687
aphid       1    6.659          2    104.027  0.009864
tree        1  104.027          1      0.001 < 2.2e-16
```

```
Anova(Aphid.Fit1)
            LR Chisq Df Pr(>Chisq)
aphid          0.003  1    0.9542
tree         104.027  1     <2e-16
```

Type II だと aphid に効果はないが,
Type I だと効果が間違って検出されてしまう !!

# なぜ検定のTypeで結果が異なるのか。
## (層別をきちんと行っているか否かの問題)

```
Aphid.Fit2 <- glm(d ~ tree + aphid, binomial, Aphid)
summary(Aphid.Fit2)
Aphid.Fit3 <- glm(d ~ aphid, binomial, Aphid)
summary(Aphid.Fit3)
```

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.914420   0.148106 -26.430   <2e-16 ***
tree二番目    1.494968   0.158755   9.417   <2e-16 ***
aphid有       0.009518   0.165722   0.057    0.954
```

aphidがプラスに働くと推定(有意ではないが…)

```
             Estimate   Std. Error z value Pr(>|z|)
(Intercept) -2.93068    0.07629  -38.42   <2e-16 ***
aphid有      -0.39815   0.15926   -2.50   0.0124 *
```

aphidがマイナスに働くと推定

不釣合い型実験の場合は層別を無視すると間違った「方向」に推定してしまう場合がある。これは正規分布の場合も普通に生じるが, 二項分布の場合は特に「シンプソンのパラドックス」と呼ばれる。(実はパラドックスでも何でもない。)

# なぜ検定のTypeで結果が異なるのか。
## (層別をきちんと行っているか否かの問題)

```
Aphid.Fit2 <- glm(d ~ tree + aphid, binomial, Aphid)
summary(Aphid.Fit2)
Aphid.Fit3 <- glm(d ~ aphid, binomial, Aphid)
summary(Aphid.Fit3)
```

- 今は Aphid と Tree に優劣がないので，平等に扱わないといけない。

- 注意事項の三つの表現（同一のことを述べている）
  - ✓検定の Type を間違わないこと。
  - ✓階層型モデル群を常に考えておくこと。
  - ✓パラメーターの優劣関係を常に考えておくこと。

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.31420    0.14510 -22.838   <2e-16 ***
tree二番目   1.49068    0.15755   9.461   <2e-16 ***
aphid有      0.09518    0.16572   0.057    0.954
```

aphidがプラスに働くと推定（有意ではないが）

```
            Estimate   Std. Error z value Pr(>|z|)
(Intercept) -2.93068    0.07629  -38.42   <2e-16 ***
aphid有     -0.39815    0.15926   -2.50   0.0124 *
```

aphidがマイナスに働くと推定

不釣合い型実験の場合は層別を無視すると間違った「方向」に推定してしまう場合がある。これは正規分布の場合も普通に生じるが，二項分布の場合は特に「シンプソンのパラドックス」と呼ばれる。（実はパラドックスでも何でもない。）

# 誤差とリンク関数の一般的な使い分け

## (1) 誤差に関して

- 連続的な定量値の場合：正規分布
- カウントデータの場合：ポアソン分布
- 比率データの場合：二項分布

## (2)リンク関数に関して（つまり「期待値の変数変換」に関して）

先ほどの解析ではリンク関数を指定しなかったが，**デフォルトで勝手に変数変換が行われていることに注意する**。デフォルトで用いられるのは以下の canonical link（十分統計量が存在する）

- 正規分布誤差の場合：そのまま：identity link（無変換）
- ポアソン分布の場合：対数関数（対数変換）
- 二項分布の場合：ロジット関数（ロジット変換）
- ガンマ分布の場合：逆数（逆数変換）

# どの変数変換を採用すべきか

いま $f[E(\mathbf{y})] = \mathbf{Xb}$ という式を使っている。つまり，期待値の変数変換値が，要因の足し算で表現できると仮定している。

**足し算の関係になるような変換を行うべき**

掛け算の形で決まる現象の場合→対数変換を行うと「足し算の関係」。
（例）昆虫の4齢幼虫数（$N_4$）は1齢幼虫数（$N_1$）と各段階の生存率（$S_1, S_2, S_3$）の積。

$$N_4 = N_1 \times S_1 \times S_2 \times S_3$$

$S_3$ の変動量が同じでも，$N_1$ の値が２倍になると$N_4$ の変動量は２倍に増幅される。
→ 等分散にはならない。

$\log_e(S_3)$の変動の影響は$\log_e(N_1)$が２倍になっても変わらない。

$$\rightarrow \log_e(N_4) = \log_e(N_1) + \log_e(S_1) + \log_e(S_2) + \log_e(S_3)$$

**掛け算で決まる現象の場合は対数変換が妥当**

# Pseudo-replication の問題

(実験計画における致命的なミス)
(一般化線形モデルの誤用とも関係)

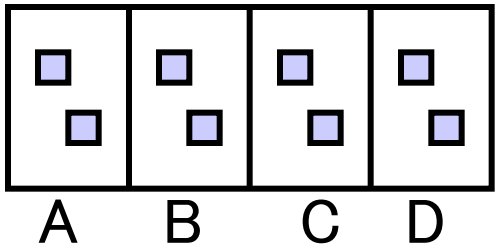
Hurlbert (1984) によって指摘された。

アメリカ統計学会は, この論文が1984年のbiometry分野でもっとも重要な論文であるとして,「スネデカー賞」を与えた。

# Pseudo-replicationの例

四つの水田圃場 (170m$^2$) にそれぞれ異なる施肥処理 (A, B, C, D) を施した。それぞれの圃場に 25 × 25 株の調査区を2ヵ所設定し，それぞれの調査区においてウンカの個体数を計測し，次の結果を得た。

| | 施肥処理 | | | |
|---|---|---|---|---|
| | A | B | C | D |
| 区1 | 10.49 | 15.86 | 7.91 | 0.83 |
| 区2 | 11.22 | 7.12 | 6.41 | 1.12 |




分散を安定化させるため，対数変換を行なった後，分散分析を行なったところ，次表の結果を得た。

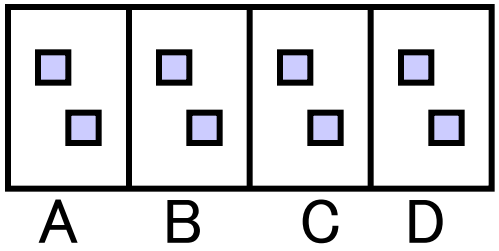
| | 自由度 | 平方和 | 平均平方 | *F* | 確率 |
|---|---|---|---|---|---|
| 施肥処理 | 3 | 7.977 | 2.659 | 27.27 | 0.0040 |
| 誤差 | 4 | 0.390 | 0.0975 | | |
| 全体 | 7 | 8.367 | | | |

この論文は致命的な欠点を含んでいたため，却下された。

# Pseudo-replicationの例

四つの水田圃場（170m$^2$）にそれぞれ異なる施肥処理（A，B，C，D）を施した。それぞれの圃場に 25 × 25 株の調査区を2ヵ所設定し，それぞれの調査区においてウンカの個体数を計測し，次の結果を得た。

| | 施肥処理 | | | |
|---|---|---|---|---|
| | A | B | C | D |
| 区1 | 10.49 | 15.86 | 7.91 | 0.83 |
| 区2 | 11.22 | 7.12 | 6.41 | 1.12 |




（観測値）＝（施肥処理による期待値）
＋（圃場での誤差成分）＋（圃場内の観測誤差成分）

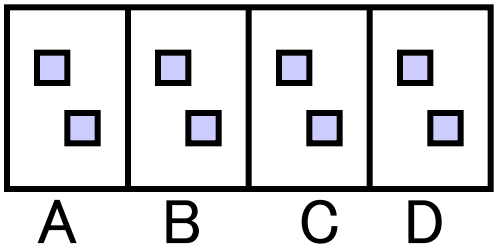
| | 自由度 | 平方和 | 平均平方 | F | 確率 |
|---|---|---|---|---|---|
| 施肥処理 | 3 | 7.977 | 2.659 | | |
| 全体 | 7 | 8.367 | | | |

この部分の差の検定をしてしまっている。

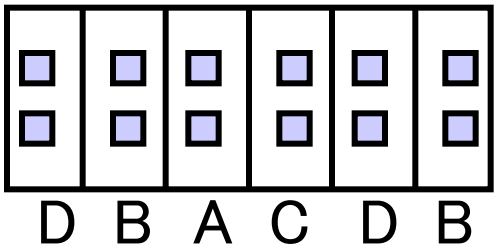
ここで行っているのは「圃場間」の差の検定であり，「施肥処理間」の検定ではない。

この論文は致命的な欠点を含んでいたため，却下された。

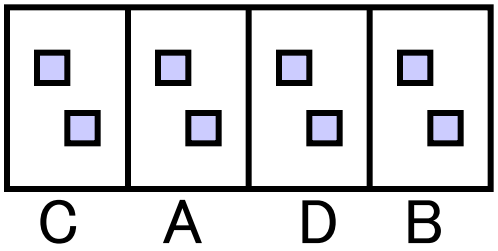
# Pseudo-replicationを避ける




ここで行っているのは「圃場間」の差の検定であり，「施肥処理」間の検定ではない。




区を狭くして，どこかの区で反復を設けてランダマイズする。
ただし，昆虫の場合には区間の移動に注意



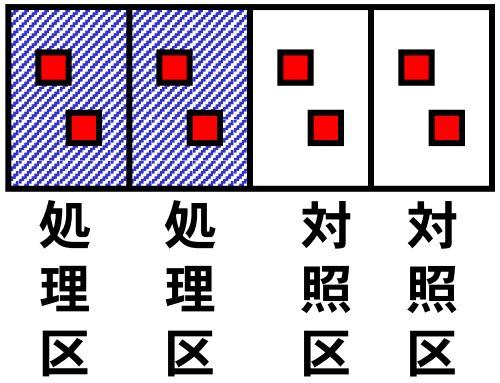

2年目は配置を換えて行い，年をブロックとする分散分析。

もし，実験を繰り返す時間や面積がないならば…

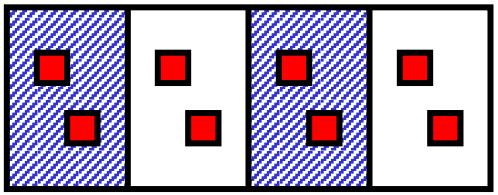
全圃場が精密に管理されているので，「施肥処理」の発現にはいかなる誤差も生じない，と主張する。ただし，レフェリーには通じない可能性が高い。

# 反復の配置に関して

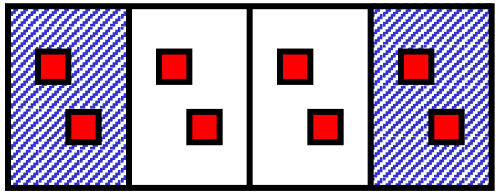
反復の配置はランダマイズされることが想定されるが，実際にはランダマイズは好ましくない。




1/3の確率でこのような配置が生じる。しかし，4圃場で反復数が2に見えるが，環境傾度がある場合には，環境傾度を同じ割合で受けていない。
→したがって反復数は1（反復なし）反復測定数が4 。

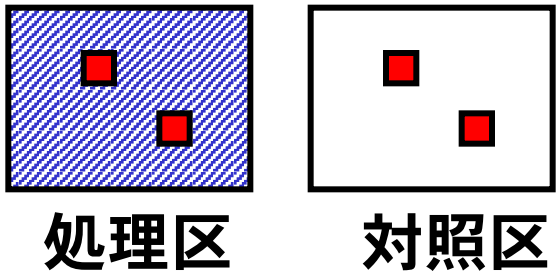


このように処理区と対照区を一定間隔で配置する方がよい。 これは「2ブロックによる乱塊法」や「ペアマッチング」とみなすこともできる。誤差の自由度をかせぐためにブロックやペアを無視した分析も行って，いずれか良い方を採用するとよい。（厳密には正しくはないが，プラクティカルな対処法）。

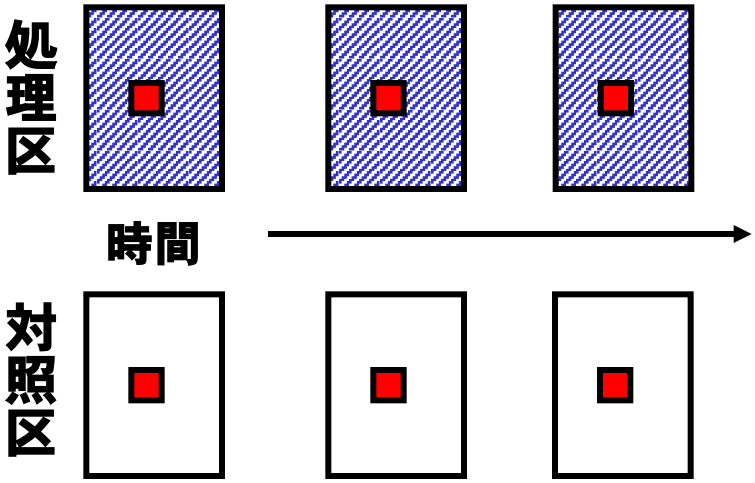


環境傾度が強く疑われる場合にはこれもよいが，対照区が一箇所に固まっているため問題が生じる。また，害虫発生の実験ではエッジ効果があるので問題。

# 時間的な繰り返し測定もよく見られる

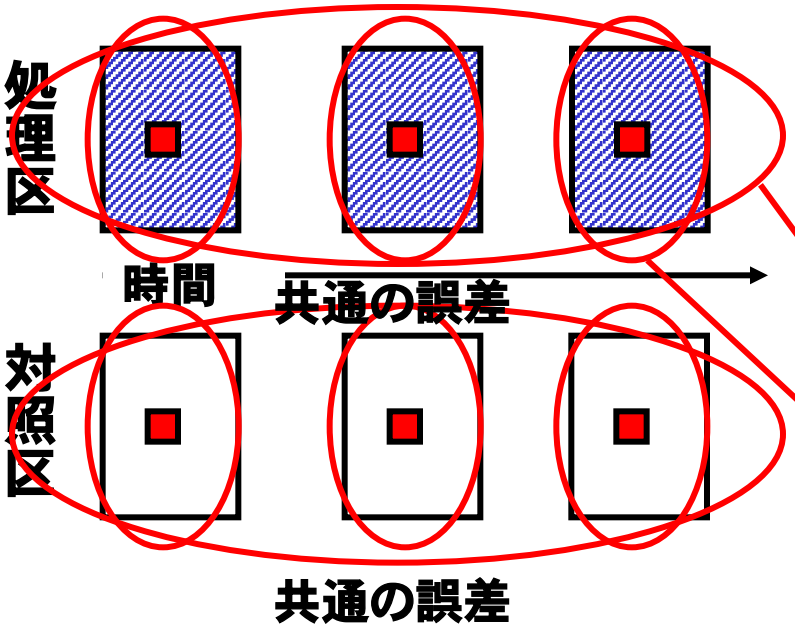



空間的な繰り返し測定




時間的な繰り返し測定

# 繰り返し測定分散分析




ここには2種類の誤差が存在する

- 圃場に固有の誤差（1次誤差）
- 測定時点毎に生じる誤差（2次誤差）

→分割区型分散分析を行う（最低限）

split-plot experiment

ただし，2次誤差間に相関がないと考えてよい場合（やや非現実的）。
最近ではAR(1)構造誤差をなどを入れるのが定石。

# 計算例：有機農法と昆虫数

| 日 | 10 | 16 | 24 | 29 |
|---|---|---|---|---|
| 有機農法 | | | | |
| 反復1 | 50 | 20 | 78 | 40 |
| 反復2 | 5 | 6 | 29 | 11 |
| 慣行農法 | | | | |
| 反復1 | 20 | 8 | 106 | 20 |
| 反復2 | 6 | 1 | 6 | 2 |

```
cat(file="yuuki.txt",
"Obs type repl date y
1   yuuki R1   D1  50
2   yuuki R1   D2  20
3   yuuki R1   D3  78
4   yuuki R1   D4  40
5   yuuki R2   D1  5
6   yuuki R2   D2  6
7   yuuki R2   D3  29
8   yuuki R2   D4  11
9   kanko R3   D1  20
10  kanko R3   D2  8
11  kanko R3   D3  10
12  kanko R3   D4  20
13  kanko R4   D1  6
14  kanko R4   D2  1
15  kanko R4   D3  6
16  kanko R4   D4  2
")
yuuki <-
read.table("yuuki.txt",
header=TRUE)
```

# プログラム例

- 関数 lme を用いる場合

repl 別に誤差が加わる

```
library(nlme)
yuuki.lme <- lme(log(y+0.5) ~ type + date, random = ~1|repl,
data=yuuki)
summary(yuuki.lme);anova(yuuki.lme)
```

- 関数 lmer を用いる場合

repl 別に誤差が加わる

```
library(lme4)
yuuki.lmer <- lmer(log(y+0.5) ~ type + date + (1|repl),
data=yuuki)
summary(yuuki.lmer);anova(yuuki.lmer)
```

- 関数 aov を用いる場合

repl 別に誤差が加わる

```
yuuki.aov <- aov(log(y+0.5) ~ type + date + Error(repl),
data=yuuki)
summary(yuuki.aov)
```

元データ

| 日 | 10 | 16 | 24 | 29 |
|---|---|---|---|---|
| 処理1 | | | | |
| 反復1 | 50 | 20 | 78 | 40 |
| 反復2 | 5 | 6 | 29 | 11 |
| 処理0 | | | | |
| 反復1 | 20 | 8 | 106 | 20 |
| 反復2 | 6 | 1 | 6 | 2 |

# 結果(lme)

有機処理区, 対照区の間では有意差はない。

●推定値

```
Fixed effects: log(y + 0.5) ~ type + date
               Value    Std.Error  DF  t-value   p-value
(Intercept)  2.0789599 0.7243408   9  2.8701408  0.0185
typeyuuki    1.1015544 0.9785684   2  1.1256795  0.3772
dateD2      -0.7702975 0.3497485   9 -2.2024329  0.0551
dateD3       0.3629295 0.3497485   9  1.0376870  0.3265
dateD4      -0.1096460 0.3497485   9 -0.3134995  0.7610
```

●分散分析表

```
            numDF   denDF    F-value p-value
(Intercept)     1      9   26.117142  0.0006
type            1      2    1.267154  0.3772
date            3      9    3.652990  0.0570
```

第1時点と第2時点の間では marginally significant。

四つの時点間で marginally significant。

# 一般化線型モデルの誤用

(平均値自体が変動する,という問題)
pseudo-replication の問題

# (例) ポアソン回帰で個体数データを分析するケース

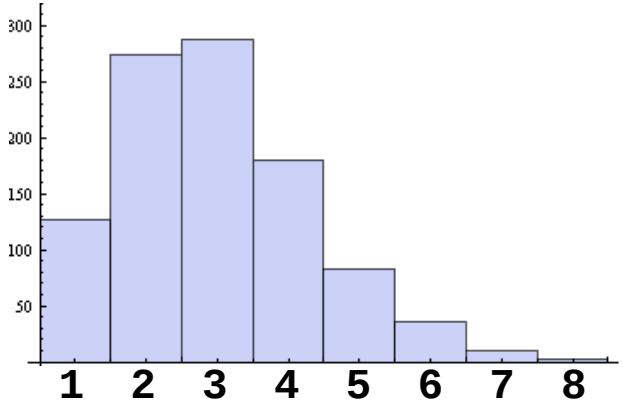
個体数データは 0, 1, 2,... といった「非負の整数」,
これは**計数データ(カウントデータ)**とよばれる

二項分布

各個体が一定の確率で
観測されるならば

さらに, 各個体の観測確率が
十分に小さいならば

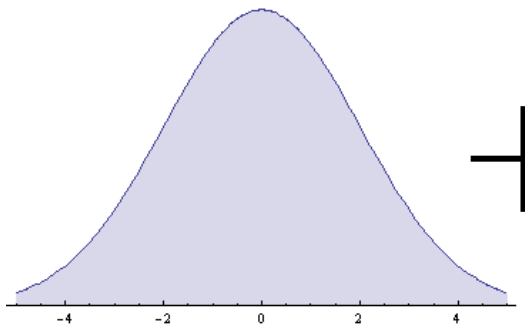
ポアソン分布



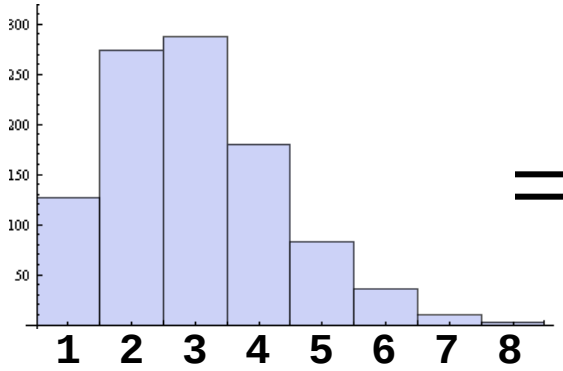

一般的な使い分け:計数データはポアソン分布で解析

しかし, 実は問題はそんなに単純ではない。

# ポアソン分布の期待値自体が変動しているのが普通




＋



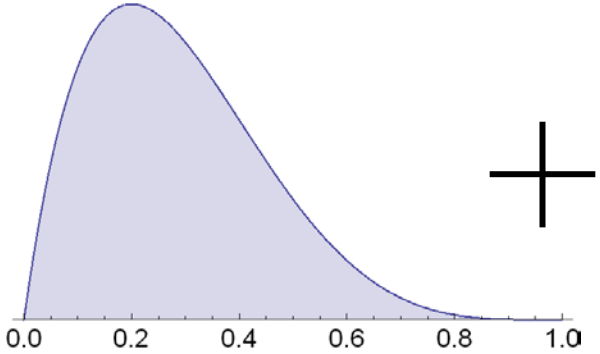

＝

**実際の個体数の変動**

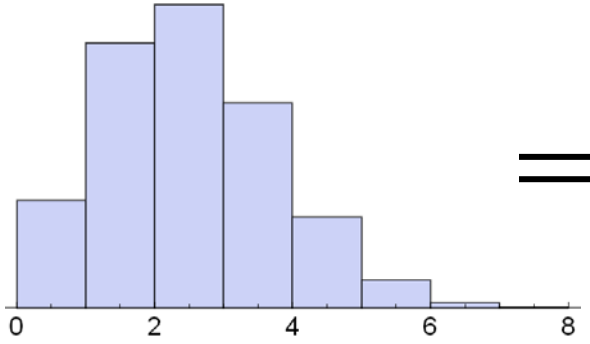
**期待値が要因効果の周りに変動**

**期待値周りの個体数の変動（ポアソン分布）**

# 二項分布の期待値自体も変動している場合がある




＋




＝

**実際の個体数の変動**

***発生率p*が期待値まわりに変動**

***p*周りの発生個体数の変動（二項分布）**

# これらは pseudo-replication 問題の一種

## Hurlbert (1984) の論文(Pseudo-replication を初めて警告した論文)の Table 6

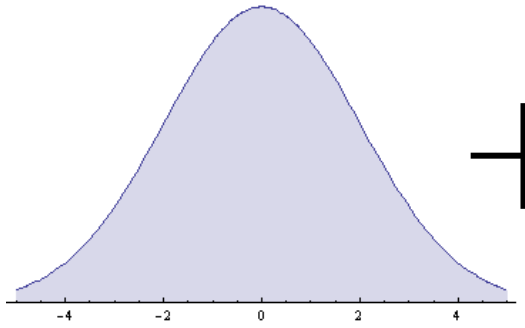


TABLE 6. A hypothetical example of sacrificial pseudoreplication resulting from misuse of chi-square.
*Question:* Does fox predation affect the sex ratio of *Microtus* populations?
*Experimental design:* Establish four 1-ha experimental plots in a large field where foxes hunt; put fox-proof fences around two plots selected at random ($A_1$, $A_2$), keep the other two plots as controls ($B_1$, $B_2$); 1 mo later sample *Microtus* population in each plot.

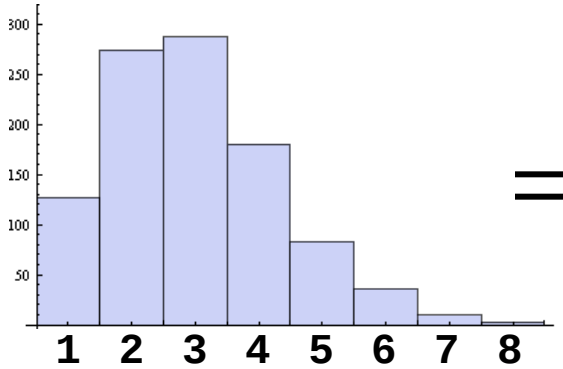
| | Results of sampling | | | | |
|---|---|---|---|---|---|
| | Plot | % males | No. males | No. females | Statistical analysis |
| Foxes | $A_1$ | 63 | 22 | 13 | Test for homogeneity with $\chi^2$<br>Result: $\chi^2 = .019$, $P > .50$<br>So: pool the data (see below) |
| | $A_2$ | 56 | 9 | 7 | |
| No foxes | $B_1$ | 60 | 15 | 10 | Test for homogeneity with $\chi^2$<br>Result: $\chi^2 = 2.06$, $P > .15$<br>So: pool the data (see below) |
| | $B_2$ | 43 | 97 | 130 | |
| | Pooled data | | | | |
| Foxes | $A_1 + A_2$ | 61 | 31 | 20 | Test for homogeneity with $\chi^2$<br>Result: $\chi^2 = 3.91$, $P < .05$<br>*Conclusion:* foxes affect sex ratio |
| No foxes | $B_1 + B_2$ | 44 | 112 | 140 | |

## 二項分布に関するpseudo-replicationの例をあげていた。

# 解決策：ポアソン分布の場合




＋




＝

実際の個体数の変動

期待値が要因効果の周りに変動

期待値周りの個体数の変動（ポアソン分布）

↓

期待値の変動の性質

個体数は掛け算の関係で決まる

→対数スケールでは足し算の関係で決まる

→対数スケールで誤差が足し算の関係になる

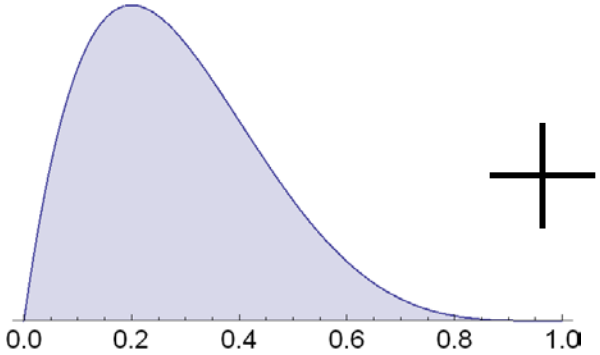
→対数スケールで期待値の変動は正規分布になる

対数レベルの正規分布+真数レベルのポアソン分布 ＝ 実際の個体数の変動

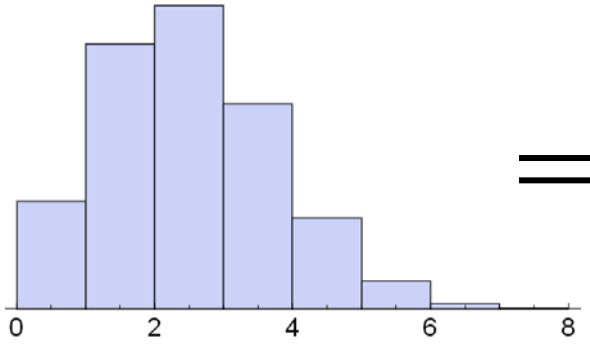
→ これは一般化線形混合モデルの一種

logarithmic Poisson GLMM

# 解決策：二項分布の場合




＋




＝

実際の個体数の変動

発生率$p$が期待値まわりに変動

↓

期待値の変動の性質

$p$周りの発生個体数の変動
（二項分布）

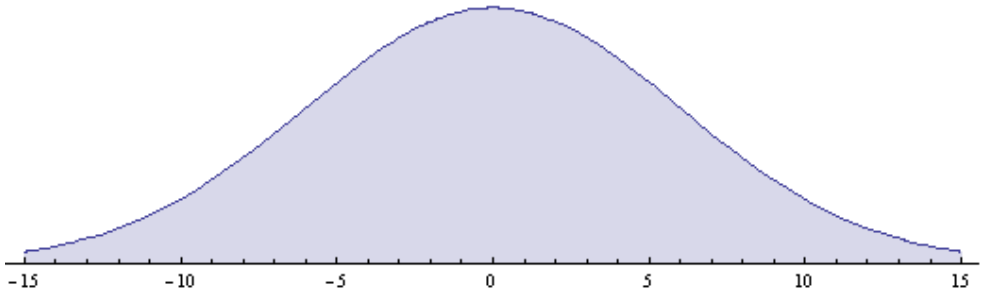
発生確率は二つの量$a, b$の相対比$a/(a+b)$で決まり，
それらの量$a, b$は，それぞれ何らかの要因の掛け算の関係で決まる
→ロジットスケールで期待値の変動は正規分布になる

ロジットスケールの正規分布+真数レベルの二項分布 ＝ 実際の個体数の変動

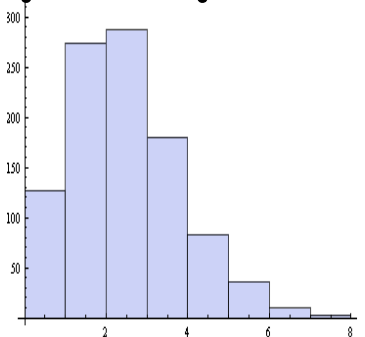
→ これも一般化線形混合モデルの一種 logit binomial GLMM

→ しかし，計算はやや面倒。
（数値積分かラプラス近似）

# 近似的な解決策：ポアソン分布の場合（その1）



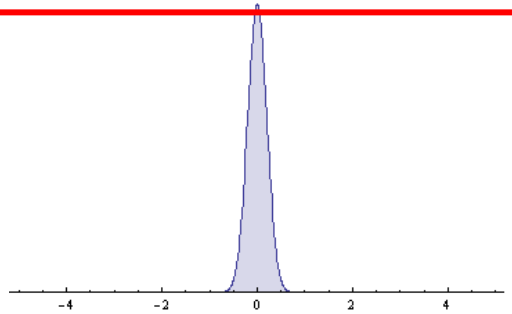

期待値の変動

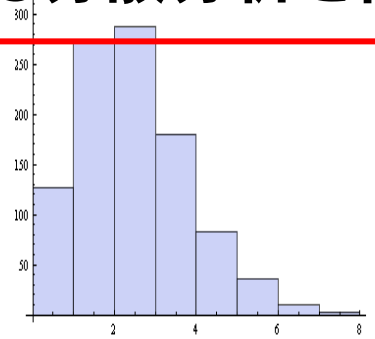



期待値周りの変動

期待値の変動がポアソン分布変動よりもずっと大きいとき
→ ポアソン分布変動を無視できる

**対数変換した個体数$\log_e(x+0.5)$に関する分散分析を行う**

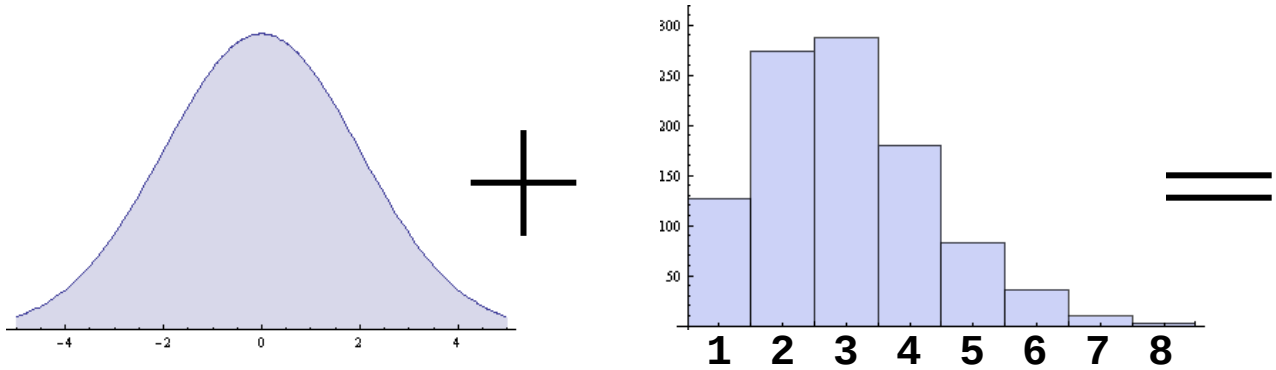



期待値の変動がポアソン分布変動よりもずっと小さいとき
→ 期待値の変動を無視できる

**ポアソン分布と対数リンクを仮定した分析（ポアソン回帰）を行う。**

# 近似的な解決策：ポアソン分布の場合（その2）




実際の個体数の変動

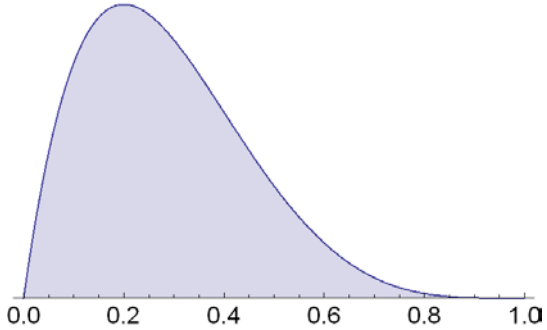
期待値が要因効果の周りに変動　期待値周りの個体数の変動（ポアソン分布）

対数レベルの正規分布+真数レベルのポアソン分布
（一般化線形混合モデル）

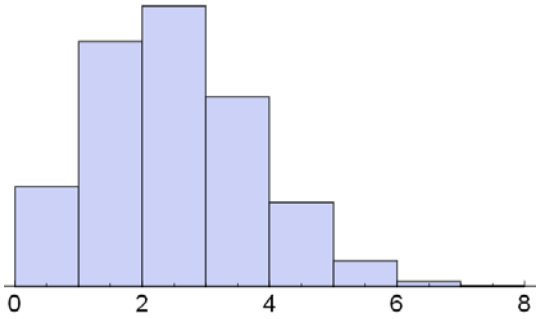
対数レベルの正規分布（等分散）
＝ 真数レベルの対数正規分布（CV一定）
≒ 真数レベルのガンマ分布（CV一定）

真数レベルのガンマ分布 + ポアソン分布 で近似
（$k$一定の負の二項分布，NB2と呼ばれる）

# 近似的な解決策：二項分布の場合




発生率$p$が期待値まわりに変動




$p$周りの発生個体数の変動（二項分布）

期待値の変動が二項分布変動よりもずっと大きいとき
→ 二項分布変動を無視できる

**経験ロジット変換した値に関する分散分析を行う。**

$$f(x) = \log_e\left(\frac{x + 0.5}{n - x + 0.5}\right)$$

期待値の変動が二項分布変動よりもずっと小さいとき
→ 期待値の変動を無視できる

**二項分布とロジットリンクを仮定した分析（ロジスティック回帰）を行う。**

# 実は先ほどのハスモンヨトウのデータも「カウントデータ」。
# →公式どおりに一般化線型モデルを使うとどうなるか?

```
Hasumon.poisson <- glm(Y ~ Trap + Month,
      poisson, data=Hasumon)
summary(Hasumon.poisson)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.61328    0.32371 -20.429   <2e-16 ***
TrapT2      -0.03968    0.06833  -0.581    0.561
TrapT3       0.87661    0.05693  15.399   <2e-16 ***
Month        1.55732    0.04136  37.654   <2e-16 ***
```

→ TrapT3で異常に*P*値が小さくなる

# Rで一般化線形混合モデルで分析(glmer)

```
library(lme4)
Hasumon.glmer <- glmer(Y ~ Trap +
   Month + (1|Obs), family=poisson,
   data=Hasumon)
summary(Hasumon.glmer)
library(car)
Anova(Hasumon.glmer)
```

| Obs | Area | Trap | Month | Y |
|---|---|---|---|---|
| 1 | A | T1 | 5 | 10 |
| 2 | A | T1 | 6 | 26 |
| 3 | A | T1 | 7 | 45 |
| 4 | A | T1 | 8 | 356 |
| 5 | A | T2 | 5 | 8 |
| 6 | A | T2 | 6 | 16 |
| 7 | A | T2 | 7 | 55 |
| 8 | A | T2 | 8 | 341 |
| 9 | B | T3 | 5 | 16 |

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -4.60291 | 0.65492 | -7.028 | 2.09e-12 | *** |
| TrapT2 | -0.10746 | 0.24496 | -0.439 | 0.66090 | |
| TrapT3 | 0.76809 | 0.23515 | 3.266 | 0.00109 | ** |
| Month | 1.28509 | 0.09254 | 13.887 | < 2e-16 | *** |

→ 今の場合は正確だが, 信頼性がある程度確立されている
関数(SASのnlmixed等)で確認しておく方がよいかもしれない。

# SAS で一般化線形混合モデルで分析 (nlmixed)

```
data Hasumon_obs;
input OBS AREA$ TRAP$ MONTH Y; logY = log(Y+0.5);
if TRAP='T1' then do dum1=0; dum2=0; end; * 非線形回帰用のprocedureなので;
if TRAP='T2' then do dum1=1; dum2=0; end; * 自分でデザイン行列を作成する;
if TRAP='T3' then do dum1=0; dum2=1; end;
datalines;
1       A       T1      5       10
2       A       T1      6       26
3       A       T1      7       45
4       A       T1      8       356
5       A       T2      5       8
6       A       T2      6       16
7       A       T2      7       55
8       A       T2      8       341
9       B       T3      5       16
10      B       T3      6       48
11      B       T3      7       112
12      B       T3      8       874
;
proc nlmixed data=Hasumon_obs;
parms b0=1 b1=-1 b2=-1 b3=1 s2=1;
pred = exp(b0 + b1*dum1 + b2*dum2 + b3*MONTH + e1);
model Y ~ poisson(pred);
random e1~normal(0,s2) subject=OBS;
run;
```

## 対数変換後の分散分析

推定値/SE

| | Estimate | Std.Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -4.0042 | 0.6914 | -5.791 | 0.000409 *** |
| TrapT2 | -0.1324 | 0.2792 | -0.474 | 0.648123 |
| TrapT3 | 0.7147 | 0.2792 | 2.560 | 0.033643 * |
| Month | 1.2054 | 0.1019 | 11.824 | 2.4e-06 *** |

## ポアソン回帰

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -6.61328 | 0.32371 | -20.429 | <2e-16 *** |
| TrapT2 | -0.03968 | 0.06833 | -0.581 | 0.561 |
| TrapT3 | 0.87661 | 0.05693 | 15.399 | <2e-16 *** |
| Month | 1.55732 | 0.04136 | 37.654 | <2e-16 *** |

## 一般化線形混合モデル(SAS の proc nlmixed)

| | Value | Std.Error | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | -4.6029 | 0.6715 | -6.85 | <.0001 |
| TrapT2 | -0.1075 | 0.2453 | -0.44 | 0.6698 |
| TrapT3 | 0.7681 | 0.2355 | 3.26 | 0.0076 |
| Month | 1.2851 | 0.09475 | 13.56 | <.0001 |

一般化線型混合モデルの結果は対数変換後の分散分析の結果とだいたい同じだが，ポアソン回帰の結果とは大きく異なる。

# 負の二項分布(NB2)で近似した場合の計算(glm.nb)

```
library(MASS)
Hasumon.nb <- glm.nb(Y ~ Trap + Month, data=Hasumon)
summary(Hasumon.nb)                    # デフォルトで対数リンク。
library(car);Anova(Hasumon.nb)# 推定kは正しく固定される。
```

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -4.60096 | 0.65149 | -7.062 | 1.64e-12 | *** |
| TrapT2 | -0.12371 | 0.23728 | -0.521 | 0.602104 | |
| TrapT3 | 0.76539 | 0.22817 | 3.355 | 0.000795 | *** |
| Month | 1.29168 | 0.09161 | 14.100 | < 2e-16 | *** |

## 一般化線形混合モデル(glmer)

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -4.60291 | 0.65492 | -7.028 | 2.09e-12 | *** |
| TrapT2 | -0.10746 | 0.24496 | -0.439 | 0.66090 | |
| TrapT3 | 0.76809 | 0.23515 | 3.266 | 0.00109 | ** |
| Month | 1.28509 | 0.09254 | 13.887 | < 2e-16 | *** |

このデータでは, GLMMの結果はNB2の結果とほとんど同じ。

# まとめ：一般線型モデル (LM)

- 一般線型モデルは，回帰分析や分散分析を統合するモデル。誤差に正規分布を仮定する。
- 「相加性」と「等分散正規分布」を仮定しているため，適切な変数変換を行う必要がある。（相加性を満たすように変換すれば，中心極限定理などにより，等分散正規分布の仮定も普通は成立する。）
- 分散分析の推定では，制約条件を設けることによりパラメーターを一つに決めている。したがって，用いる制約条件の違いにより，統計解析ソフトにより推定結果は異なる。
- 検定ではパラメーターを抜く順序によって検定結果は異なる。したがって，用いる関数により検定結果は異なる。パラメーターの優劣を判断して適切な順序を選ぶ必要がある

# まとめ：一般化線型モデル (GLM)

- 一般化線型モデルは，誤差を正規分布から指数型分布族に拡張している。
- 期待値に関する変数変換関数はリンク関数と呼ばれ，デフォルトで自動的に変数変換が行われる。
- 期待値に関して「相加性」を仮定しているため，一般線型モデルと同様に，**相加性を満たすような変数変換を選択**しないといけない。
- 「制約条件の問題」と「検定の順序の問題」は一般線型モデルと同じ。したがって，使用する関数により結果は異なる。
- 期待値が変動している際にポアソン回帰を使うと，本来は差がないのに間違って有意差を出してしまうことが多くなる。ロジスティック回帰も同様。→あぶない方向に偏るので要注意。
- 一般化線形混合モデルが妥当である場合が多いと思われる。

# まとめ：一般化線型混合モデル (GLMM)

- パラメーター推定には数値積分やラプラス近似などの複雑な計算が必要。したがって，関数や統計解析ソフトによっては正しい解を出力しない場合もあるので要注意。
- 期待値に関して「相加性」と「等分散正規分布」を仮定しているため，**相加性を満たすような適切な変数変換**を選択する必要がある。
- 「制約条件の問題」と「検定の順序の問題」は一般線型モデルと同じ。したがって，使用する関数により結果は異なる。
- 一般化線形混合モデルの推定が困難な場合は「対数変換後の分散分析」や「経験ロジット変換後の分散分析」が近似として使用可能な場合がある。
- ポアソン一般化線形混合モデルは，$k$が一定の負の二項分布モデルで近似できる。ただし対数リンクを使用する。

# モデル選択とモデル評価

- なぜモデル評価が必要か?
- AIC（赤池情報量基準）の問題点
- $R_D$ 基準の活用

# 検定には意味がない??

- Fisherはモデル選択の便宜的手段として有意性検定を用いた。
- Neyman–Pearsonは真のモデルを選ぶ手段として仮説検定を用いることを提唱した。

- 赤池弘次(1976)はNeyman–Pearson流の検定を次のように批判した。「あるサイコロの正しさを検定するという問題も全く同様で，現実のサイコロで完全に対称なものが存在しえないことは明らかである。このように仮説(帰無仮説)は常に否定される立場にあり，データによる検定結果を待つまでもなく結論は見えている。」

  「有意差なし」という結果は「その効果を検出するには反復数が少なすぎた」という「実験の不備」を示しているにすぎず，Neyman–Pearsonr流の検定にはそもそも意味がない。

# 検定に代わるものが必要

- Fisherの手法「有意性検定によるモデル選択法」は実用的な方法だったが，その意味は不明確であった。モデル選択の手段として，現在では別の手法を用いるべきであろう。
- では，どのような基準で選択するべきか？
- 一般に，モデルは手持ちのデータだけを記述することを目的とするのではなく、何らかの別のデータにも適用できることを暗黙の前提としている。
- こうしたモデルの性質から考えれば、予測力でモデルの妥当性を評価するのが唯一の妥当な評価法だと言える。そうした評価法の一つがAIC（赤池情報量基準）である。

# AIC の考え方

- いま手元のデータによく当てはまるモデルは，次にデータをとったときに，その新しいデータにもうまく当てはまるとは限らない。
- そこで，次にデータをとったときの確率分布が真の確率分布にKullback−Leibler 情報量の尺度でもっとも近くなるようにモデルを選択することを考えて，赤池氏は情報量基準AICを導いた。
- 尤度を$L$とし，モデルに含まれるパラメーター数（切片を含む）を$k$とするとき，AICは次式で定義される。

$$\mathrm{AIC} = -2\log(L) + 2k$$

- 「真のモデル」を選ぶことを目指しているわけではないので，検定のような「論理的矛盾」が生じない。

# AIC の問題点

- AICが根拠としている「予測におけるKullback−Leibler情報量の尺度で測った近さ」の現実的意義が明確ではない。そのため，たとえばAIC=15.2という値が出たときに，この値（15.2）自体には意味はない。AICの使用においては，同じデータのもとで二つ以上のモデルのAICを比較した場合にのみ相対的に意味がある。つまり，AICは量的変数ではなく順序変数でしかない。
- 現在のデータの量や質が十分であるかどうかを判断するためには，「モデル選択」だけでなく「モデル評価」を行うことが極めて重要である。しかし，AICを用いた場合には，これは「相対的な尺度」であるから，「モデル選択」はできても絶対的な「モデル評価」を行うことができない。
- 検定と同様に，データ量が多ければ「もっとも複雑なモデル」が採用されて議論が終了するだけであり，そのモデルの有用性を評価することはできない。

# $R_D$ 指数の提案(Yamamura 2016)

- Kullback–Leibler 情報量の尺度ではなく,「実際に当たる確率」で正しくモデルを評価するべきではないか？ そうすれば, 選択されたモデルが「有用なモデル」かどうかを判断できる。また, 単に「予測力最大」ではなく, コストが少なく「適度の予測力」を持つモデルも選択可能になる。
- 尤度ではなく発生確率で比較するために, ラプラス哲学にしたがって「真のモデル」を飽和モデルなどもっとも複雑なモデルに固定する。(固定しなければ発生確率の比較にはならない。)
- その上で, 予測力の改善割合$R_{\mathrm{pred}}$を考える。将来のデータをすべて知っている「神」が予測した場合に$R_{\mathrm{pred}}$は100%となり, 説明変数をまったく持たない「凡人」が予測した場合に$R_{\mathrm{pred}}$が0%となるように改善割合$R_{\mathrm{pred}}$を定義する。この$R_{\mathrm{pred}}$の推定値として$R_D$指数が導出されている。

# $R_D$ 指数の定義式

- **$R_D$の定義式**

$$R_D = 1 - \frac{l_{\max} - l + k}{l_{\max} - l_{\text{null}} + 1}$$

- ✓ $l$ は候補モデルでの対数尤度
- ✓ $l_{\text{null}}$は切片だけを含むモデル (null model) における対数尤度
- ✓ $l_{\max}$は固定効果パラメーター数とデータ数が等しい最大モデル (maximum model) における対数尤度。
- ✓ ただし, まず「真のモデル」から分散パラメーターを推定し, その分散パラメーター（つまり確率の定義）を固定して, これらの値を計算。
- ✓ 分母の1は切片の数であるため, 多変量の場合は, 分母の1を変量の数に置き換える。

# $R_D$ 指数の計算プログラム

- $R_D$ を計算するための1変量用のR関数（RDcompare）およびSASマクロが以下のサイトにおいてある。論文の著者版原稿もここに置いてある。

  http://cse.naro.affrc.go.jp/yamamura/RD_criterion_program.html

- R用の関数を使えば一般化線型モデルにおいて $R_D$ の計算を自動的に行うことができる。stepAIC関数と同じく，交互作用に関しては，パラメーターの優劣関係を自動的に判定してくれる。要因数が多すぎて誤差の自由度が少ない場合には，あらかじめ分散を推定してから，その分散を固定して指定できる。

- 一方，SAS用の関数を使えば一般化線型混合モデルにおいて $R_D$ の計算を自動的に行うことができる。stepAIC関数と同じく，交互作用に関しては，パラメーターの優劣関係を自動的に判定してくれる。

# $R_D$の計算例(LM)

- $\log_e(x+0.5)$近似を用いた場合(ハスモンヨトウのデータ)

```
source("RDcompare.txt")
RDcompare(log(y+0.5) ~ trap*month, data=MothData)
```

- 出力の一部

```
# RD ranking for the hierarchical family of models #
          RD                                                Model
1   0.90679867                    log(y+0.5) ~ 1 + trap + month
2   0.88850757   log(y+0.5) ~ 1 + trap + month + trap:month
3   0.84623461                       log(y+0.5) ~ 1 + month
4   0.06056406                        log(y+0.5) ~ 1 + trap
5   0.00000000                             log(y+0.5) ~ 1
```

- もっとも予測力が高いモデルは交互作用を無視したモデルである。その予測力の改善割合の推定値は$R_D = 0.90679867$であり, 予測力は十分に高いことが分かる。

# $R_D$の計算例（GLMM）

- R関数は一般化線形モデルまでしか対応していないが，SASマクロは一般化線形混合モデル（GLMM）まで対応している。結果は以下の通り。

| Rank | ModelDF | RD | RSD | Model |
|---|---|---|---|---|
| 1 | 3 | 0.90973 | 0.94266 | Trap Month |
| 2 | 4 | 0.89128 | 0.92354 | Trap Month Trap*Month |
| 3 | 2 | 0.84690 | 0.87755 | Month |
| 4 | 2 | 0.04703 | 0.04873 | Trap |
| 5 | 1 | 0.00000 | 0.00000 | |

一般化線形混合モデルによる結果は$\log_e(x+0.5)$に関する線形モデル分析の結果とほとんど同じ。個体数が大きい場合には，このようにGLMMをLMで近似することができる。

# $R_D$でのモデル選択はAICよりも柔軟性が高い

- $R_D$でのモデルの選び方について
  予測力を問題にする場合であっても，必ずしも$R_D$が最大となるモデルを採択する必要はない。順位が2位以下のモデルであっても，最良モデルと比較してあまり$R_D$が低下しておらず，かつ，利用しやすい性質を持っているモデル（たとえば，容易に測定できる説明変数からなるモデルや，容易に解釈できるモデル）であれば，そちらのモデルを採択するべきである。

- 予測が目的ではなく「変数が生じる主な理由を把握する」のが目的の場合には，$R_D$が大きいモデルよりも，あえて$R_D$が0.8程度のモデルを採用するのが好ましい場合もあるであろう。今の場合は，月だけを用いたモデルで$R_D = 0.85$であり0.8に近いことから，このモデルを「要約モデル」として採用し，「ハスモンヨトウの個体数は主として月によって決まっている」と解釈するのも妥当であろう。

- 是非$R_D$を使ってみてください。

# 私の講義は以上です