

# 統計モデリング入門 2017 (d)

model selection and statistical test  
モデル選択と検定

久保拓弥 kubo@ees.hokudai.ac.jp

京大豊長研の講義 <https://goo.gl/z9yCJY>

2017-11-14

ファイル更新時刻: 2017-11-11 16:02

kubostat2017d (<https://goo.gl/z9yCJY>) 統計モデリング入門 2017 (d) 2017-11-14 1 / 44

もくじ

## 今日のハナシ I

today's example: seed number data, again

- ① 前回と同じ例題: 種子数データ  
植物個体の属性, あるいは実験処理が種子数に影響?
- model selection using AIC
- ② AIC を使ったモデル選択  
badness of fit  
あてはまりの悪さ: deviance
- statistical test
- ③ 統計学的な検定  
and its asymmetricity  
そして, その非対称性
- model selection      statistical test
- ④ モデル選択 と 統計学的な検定  
misunderstanding  
のさまざまな 誤解

kubostat2017d (<https://goo.gl/z9yCJY>) 統計モデリング入門 2017 (d) 2017-11-14 2 / 44

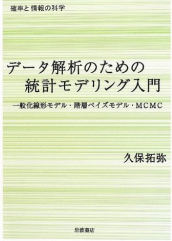
もくじ

## 今日の内容と「統計モデリング入門」との対応

今日はおもに「第4章 GLMのモデル選択」と「第5章 GLMの尤度比検定と検定の非対称性」の内容を説明します。

<http://goo.gl/Ufq2>

- 著者: 久保拓弥
- 出版社: 岩波書店
- 2012-05-18 刊行



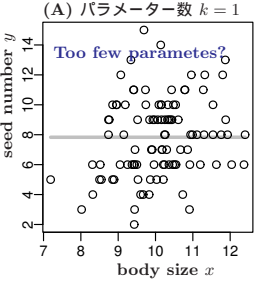
データ解析のための  
統計モデリング入門  
— 一般化線形モデル・隠層ベイズモデル・MCMC —  
久保拓弥

kubostat2017d (<https://goo.gl/z9yCJY>) 統計モデリング入門 2017 (d) 2017-11-14 3 / 44

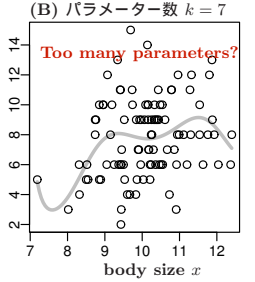
もくじ

## number of parameters パラメーター数 は多くても少なくてもヘン?

(A) パラメーター数  $k = 1$



(B) パラメーター数  $k = 7$



What is the “best?” parameter number  $k$ ?

kubostat2017d (<https://goo.gl/z9yCJY>) 統計モデリング入門 2017 (d) 2017-11-14 4 / 44

前回と同じ例題: 種子数データ      植物個体の属性, あるいは実験処理が種子数に影響?

today's example: seed number data, again

## 1. 前回と同じ例題: 種子数データ

植物個体の属性, あるいは実験処理が種子数に影響?

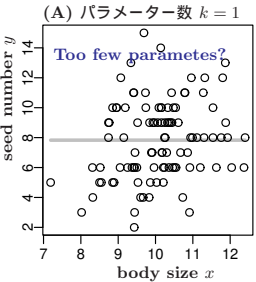
まずはデータの概要を調べる

kubostat2017d (<https://goo.gl/z9yCJY>) 統計モデリング入門 2017 (d) 2017-11-14 5 / 44

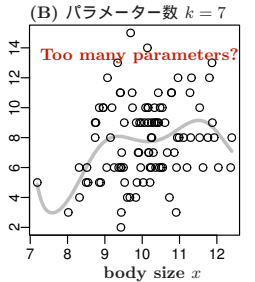
前回と同じ例題: 種子数データ      植物個体の属性, あるいは実験処理が種子数に影響?

## パラメーター数 $k$ は多くても少なくてもヘン?

(A) パラメーター数  $k = 1$



(B) パラメーター数  $k = 7$



“良いモデル” とはなにか?  $k$  も重要なのか?

kubostat2017d (<https://goo.gl/z9yCJY>) 統計モデリング入門 2017 (d) 2017-11-14 6 / 44

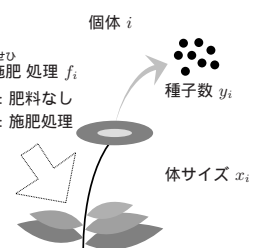
前回と同じ例題: 種子数データ 植物個体の属性, あるいは実験処理が種子数に影響?

body size  $x$  and fertilization  $f$  change seed number  $y$ ?  
 個体サイズと実験処理の効果を調べる例題

response variable seed number  
 • 応答変数: 種子数  $\{y_i\}$

explanatory variable  
 • 説明変数: body size  
 • 体サイズ  $\{x_i\}$   
 fertilization  
 • 施肥処理  $\{f_i\}$

sample size  
 標本数  
 control  
 • 無処理 ( $f_i = C$ ): 50 sample ( $i \in \{1, 2, \dots, 50\}$ )  
 fertilization  
 • 施肥処理 ( $f_i = T$ ): 50 sample ( $i \in \{51, 52, \dots, 100\}$ )



個体  $i$   
 種子数  $y_i$   
 体サイズ  $x_i$   
 施肥処理  $f_i$   
 C: 肥料なし  
 T: 施肥処理

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 7 / 44

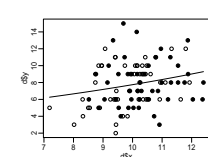
前回と同じ例題: 種子数データ 植物個体の属性, あるいは実験処理が種子数に影響?

a statistical model for this example  
 この例題のための統計モデル

probability distribution Poisson distribution  
 • 確率分布: **ポアソン分布**

linear predictor  
 • 線形予測子:  $\beta_1 + \beta_2 x_i + \beta_3 f_i$

link function log link function  
 • リンク関数: **対数リンク関数**

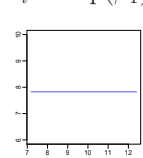


kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 8 / 44

前回と同じ例題: 種子数データ 植物個体の属性, あるいは実験処理が種子数に影響?

4 candidate models  
 4 つの可能なモデル候補: (A) constant  $\lambda$

$\lambda_i = \exp(\beta_1)$



あてはまりの良さを対数尤度 (log likelihood) で評価する

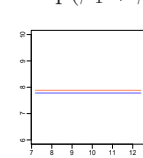
```
> logLik(glm(y ~ 1, data = d, family = poisson))
'log Lik.' -237.64 (df=1)
```

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 9 / 44

前回と同じ例題: 種子数データ 植物個体の属性, あるいは実験処理が種子数に影響?

4 candidate models  
 4 つの可能なモデル候補: (B)  $f$  model

$\lambda_i = \exp(\beta_1 + \beta_3 f_i)$



あてはまりの良さを対数尤度 (log likelihood) で評価する

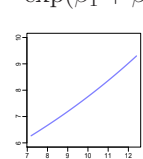
```
> logLik(glm(y ~ f, data = d, family = poisson))
'log Lik.' -237.63 (df=2)
```

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 10 / 44

前回と同じ例題: 種子数データ 植物個体の属性, あるいは実験処理が種子数に影響?

4 candidate models  
 4 つの可能なモデル候補: (C)  $x$  model

$\lambda_i = \exp(\beta_1 + \beta_2 x_i)$



あてはまりの良さを対数尤度 (log likelihood) で評価する

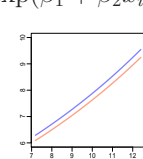
```
> logLik(glm(y ~ x, data = d, family = poisson))
'log Lik.' -235.39 (df=2)
```

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 11 / 44

前回と同じ例題: 種子数データ 植物個体の属性, あるいは実験処理が種子数に影響?

4 candidate models  
 4 つの可能なモデル候補: (D)  $x + f$  model

$\lambda_i = \exp(\beta_1 + \beta_2 x_i + \beta_3 f_i)$



あてはまりの良さを対数尤度 (log likelihood) で評価する

```
> logLik(glm(y ~ x + f, data = d, family = poisson))
'log Lik.' -235.29 (df=3)
```

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 12 / 44

前回と同じ例題: 種子数データ 植物個体の異性, あるいは実験処理が種子数に影響?

$k$  increases  $\rightarrow \log L^*$  increases  
 パラメーター数が多いとあてはまりが良い

(A) constant  $\lambda$  ( $k = 1$ )

(B) f model ( $k = 2$ )

(C) x model ( $k = 2$ )

(D) x + f model ( $k = 3$ )

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 13 / 44

AIC を使ったモデル選択 あてはまりの悪さ: deviance

model selection using AIC  
 2. AIC を使ったモデル選択

badness of fit  
 あてはまりの悪さ: deviance

badness of prediction  
 そして予測の悪さ: AIC

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 14 / 44

AIC を使ったモデル選択 あてはまりの悪さ: deviance

output  
**R の glm() は deviance を出力**

```
> glm(y ~ x + f, data = d, family = poisson)
Call: glm(formula = y ~ x + f, family = poisson, data = d)
Coefficients:
(Intercept)          x          fT
      1.2631      0.0801     -0.0320
Degrees of Freedom: 99 Total (i.e. Null); 97 Residual
Null Deviance:      89.5
Residual Deviance: 84.8      AIC: 477
```

Residual Deviance? Null Deviance? AIC?

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 15 / 44

AIC を使ったモデル選択 あてはまりの悪さ: deviance

deviance  $D = -2 \times \log L^*$

- Maximum log likelihood  $\log L^*$ : goodness of fit
- Deviance  $D = -2 \log L^*$ : badness of fit

model	$k$	$\log L^*$	Deviance $-2 \log L^*$	Residual deviance
constant $\lambda$	1	-237.6	475.3	89.5
f	2	-237.6	475.3	89.5
x	2	-235.4	470.8	85.0
x + f	3	-235.3	470.6	84.8
saturation	100	-192.9	385.8	0.0

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 16 / 44

AIC を使ったモデル選択 あてはまりの悪さ: deviance

Null deviance, Residual deviance, ...

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 17 / 44

AIC を使ったモデル選択 あてはまりの悪さ: deviance

badness of prediction  
 予測の悪さ : AIC =  $-2 \log L^* + 2k$

Look for a model of the smallest AIC  
 AIC 最小のモデルを選ぶ

model	$k$	$\log L^*$	Deviance $-2 \log L^*$	Residual deviance	AIC
constant $\lambda$	1	-237.6	475.3	89.5	477.3
f	2	-237.6	475.3	89.5	479.3
<b>x</b>	<b>2</b>	<b>-235.4</b>	<b>470.8</b>	<b>85.0</b>	<b>474.8</b>
x + f	3	-235.3	470.6	84.8	476.6
saturation	100	-192.9	385.8	0.0	585.8

AIC: A (or Akaike) information criterion

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 18 / 44

AIC を使ったモデル選択      あてはまりの悪さ : deviance

### 統計モデルによる推測 (estimation) って何だっけ?

(人間には見えない) 真の統計モデル  $\beta_1 = 2.08$  のポアソン分布

観測データから推定された constant  $\lambda$   $\hat{\beta}_1 = 2.04$  のポアソン分布

parameter estimation  
パラメーター推定

データをサンプル

推定用の観測データ

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      19 / 44

AIC を使ったモデル選択      あてはまりの悪さ : deviance

### Is it OK? Goodness of fit is evaluated by using the SAME data set ... 推定に使ったデータであてはまりを評価している?

観測データから推定された constant  $\lambda$   $\hat{\beta}_1 = 2.04$  のポアソン分布

推定用の観測データを使ってあてはまりの良さを評価すると最大対数尤度  $\log L^*$  が得られる

パラメーター推定に使ったデータなのであてはまりの良さにバイアスが生じる (過大評価) **biased "goodness of fit"!**

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      20 / 44

AIC を使ったモデル選択      あてはまりの悪さ : deviance

### 重要なこと: 新データがあてはまるかどうか

(人間には見えない) 真の統計モデル  $\beta_1 = 2.08$  のポアソン分布

観測データから推定された constant  $\lambda$   $\hat{\beta}_1 = 2.04$  のポアソン分布

データをサンプル (実際のデータ解析では不可能)

予測の良さ評価用のデータ (200 セット)

評価用のデータにあてはめてみる  
すると平均対数尤度  $E(\log L)$  が得られる

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      21 / 44

AIC を使ったモデル選択      あてはまりの悪さ : deviance

### シミュレーションで予測の良さを調べる

(A) 観測データがひとつ      (B) (A) を何度もくりかえす      (C) バイアス補正

log likelihood

(ひとつの観測データの) 最大対数尤度  $\log L^* = -120.06$

平均対数尤度 (200 セットのデータの平均)  $E(\log L) = -122.96$

推定値  $\hat{\beta}_1 = 2.04$       真の  $\beta_1 = 2.08$

$\beta_1$  の値

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      22 / 44

AIC を使ったモデル選択      あてはまりの悪さ : deviance

### バイアス補正を図示してみる

効果のあるパラメーター追加

無意味なパラメーター追加

最大対数尤度

平均対数尤度

パラメーター数      1      2      2

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      23 / 44

統計学的な検定      そして、その非対称性

### statistical test

## 3. 統計学的な検定

and its asymmetry  
そして、その非対称性

likelihood ratio test  
ここでは 尤度比検定 を紹介

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      24 / 44

統計学的な検定      そして、その非対称性

Although their procedures are similar ... they are totally different!  
モデル選択と検定の手順は途中まで同じ

統計モデルの検定

AICによるモデル選択

解析対象のデータを確定

↓

データを説明できるような統計モデルを設計

(帰無仮説・対立仮説)      ↓      (単純モデル・複雑モデル)

↓      ↓

ネストした統計モデルたちのパラメーターの最尤推定計算

↓      ↓

帰無仮説棄却の危険率を評価      モデル選択規準 AIC の評価

↓      ↓

帰無仮説棄却の可否を判断      予測の良いモデルを選ぶ

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      25 / 44

統計学的な検定      そして、その非対称性

model selection      statistical test

# モデル選択 と 統計学的検定 は

totally different in their objectives

## その目的がぜんぜんちがう

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      26 / 44

統計学的な検定      そして、その非対称性

Objective

目的?

model selection

## モデル選択:

Look for a model of better prediction

### よい予測をするモデルの探索

statistical test      rejection of null hypothesis

## 統計学的検定: 帰無仮説の排除

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      27 / 44

統計学的な検定 (Neyman-Pearson framework)

statistical test



Null hypothesis  
帰無仮説  
 $glm(y \sim 1)$   
is better!

どうでもいい  
... 興味ない...

VS



Alternative hypothesis  
対立仮説  
 $glm(y \sim x)$   
is better!

重要! これを  
主張したい!

非対称性 asymmetry?

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      28 / 44

統計学的な検定 (Neyman-Pearson framework)

statistical test



Null hypothesis  
帰無仮説  
 $glm(y \sim 1)$   
is better!

VS



Alternative hypothesis  
対立仮説  
 $glm(y \sim x)$   
is better!

test! ↓

(if ...)

reject 棄却 ----- support 支持

非対称性 asymmetry?

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      29 / 44

統計学的な検定 (Neyman-Pearson framework)

statistical test



Null hypothesis  
帰無仮説  
 $glm(y \sim 1)$   
is better!

VS



Alternative hypothesis  
対立仮説  
 $glm(y \sim x)$   
is better!

test! ↓

(if ...)

NOT reject ----- Say Nothing!?

非対称性 asymmetry?

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      30 / 44

統計学的な検定 そして、その非対称性

### また同じ例題 The same example, again

個体  $i$

種子数  $y_i$

体サイズ  $x_i$

$D$ : deviance

neglect fertilization treatment (施肥処理は無視!)

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 31 / 44

統計学的な検定 そして、その非対称性

### test statistics 検定統計量 $\Delta D_{1,2}$

difference in deviance  $\Delta D_{1,2} = D_1 - D_2 = 4.51 \approx 4.5$   
 likelihood ratio?  $-\log \frac{L_1^*}{L_2^*} = \log L_1^* - \log L_2^*$

model	$k$	$\log L^*$	Deviance $-2\log L^*$
constant $\lambda$	1	-237.6	$D_1 = 475.3$
$x$	2	-235.4	$D_2 = 470.8$

null hypothesis 帰無仮説   
alternative hypothesis 対立仮説

asymmetry in test Null hypothesis is junk  
**検定の非対称性: 帰無仮説はゴミあつかい**  
 ... yet we are focusing only on null hypothesis  
 .....にもかかわらず、帰無仮説だけをしつこく調べる

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 32 / 44

統計学的な検定 そして、その非対称性

### How to make null model 帰無仮説のつくりかた

Null hypothesis is included in Alt hypothesis  
**対立仮説の中に帰無仮説がある**  
 this is a "nested" model (ネストした関係)

- カウントデータ  $\{y_i\}$  は平均である  $\lambda_i$  のポアソン分布に従う alternative hypothesis
- 対立仮説** の一例:  $\log \lambda_i = \beta_1 + \beta_2 x_i$
- null hypothesis  
 • ネストした **帰無仮説**:  $\log \lambda_i = \beta_1$  (切片だけのモデル)

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 33 / 44

統計学的な検定 そして、その非対称性

### objective null hypothesis rejection 検定の目的: 帰無仮説の棄却

observerd 観察された逸脱度差  $\Delta D_{1,2} = 4.5$  は.....

帰無仮説は	「めったにない差」 (帰無仮説を棄却)	「よくある差」 (棄却できない)
真のモデルである	第一種の過誤 (問題なし)	(問題なし)
真のモデルではない	(問題なし)	第二種の過誤

is ... (Reject)	significant	not significant (Not reject)
TRUE	<b>Type I error</b> (no problem)	(no problem)
NOT true	(no problem)	<b>Type II error</b>

asymmetry in test evaluating only Type-I error  
**検定の非対称性: 第一種の過誤だけに注目**

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 34 / 44

統計学的な検定 そして、その非対称性

### generate $\Delta D_{1,2}$ distribution bootstrap likelihood test $\Delta D_{1,2}$ の分布を生成 : ブートストラップ尤度比検定

Suppose null hypothesis is TRUE!

**帰無仮説が真のモデルであるとして!**  
 帰無仮説が真の統計モデルということにしてしまう (  $\beta_1 = 2.06$  のポアソン分布 )

評価用データに constant  $\lambda$  と  $x$  model をあてはめて逸脱度差  $\Delta D_{1,2}$  の分布を予測

帰無仮説のモデルから新しいデータをたくさん生成する

あてはまりの良さ評価用のデータ (多数)

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 35 / 44

統計学的な検定 そして、その非対称性

### How to generate $\Delta D_{1,2}$ under is TRUE?

```

> d$y.rnd <- rpois(100, lambda = mean(d$y))
> fit1 <- glm(y.rnd ~ 1, data = d, family = poisson)
> fit2 <- glm(y.rnd ~ x, data = d, family = poisson)
> fit1$deviance - fit2$deviance
    
```

- generation of random numbers virtual data  
 • rpois() による ポアソン乱数の生成 (架空データ)
- fitting GLM to the virtual data  
 • 架空データを使って glm() あてはめ

kubostat2017d (https://goo.gl/z9yCJY) 統計モデリング入門 2017 (d) 2017-11-14 36 / 44

統計学的な検定      そして、その非対称性

You must define "rejection region" in advance  
あらかじめ**棄却域**を決めておく

say, 5%?  
たとえば 5% とか?

3500  
2500  
1500  
500  
0

0 5 10 15

NOT significant ←  
→ significant (5%)

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      37 / 44

統計学的な検定      そして、その非対称性

### A random $\Delta D_{1,2}$ generator in R

```

get.dd <- function(d) # データの生成と逸脱度差の評価
{
  n.sample <- nrow(d) # データ数
  y.mean <- mean(d$y) # 標本平均
  d$y.rnd <- rpois(n.sample, lambda = y.mean)
  fit1 <- glm(y.rnd ~ 1, data = d, family = poisson)
  fit2 <- glm(y.rnd ~ x, data = d, family = poisson)
  fit1$deviance - fit2$deviance # 逸脱度の差を返す
}

pb <- function(d, n.bootstrap)
{
  replicate(n.bootstrap, get.dd(d))
}
    
```

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      38 / 44

統計学的な検定      そして、その非対称性

### Generated distribution of $\Delta D_{1,2} = D_1 - D_2$

3500  
2500  
1500  
500  
0

0 5 10 15

observed  $\Delta D_{1,2}$   
観察された逸脱度差

$\Delta D_{1,2} = 4.5$

constant  $\lambda$  と x model の逸脱度の差  $\Delta D_{1,2}$

(R code is in the next page)

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      39 / 44

統計学的な検定      そして、その非対称性

### Probability $\{\Delta D_{1,2} \geq 4.5\} = \frac{38}{1000} = 0.038$

```

> source("pb.R") # reading "pb.R" text file
> dd12 <- pb(d, n.bootstrap = 1000)
> hist(dd12, 100) # to plot histogram
> abline(v = 4.5, lty = 2)
> sum(dd12 >= 4.5)
[1] 38
    
```

so-called "*P*-value" is 0.038.

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      40 / 44

統計学的な検定      そして、その非対称性

null hypothesis 帰無仮説 is rejected

alternative hypothesis 対立仮説 can be accepted.

So we can state that x model is better than constant  $\lambda$ .

個体  $i$       種子数  $y_i$   
体サイズ  $x_i$

seed number  $y_i$   
body size  $x_i$

$D$ : deviance  
x model  $D_2 = 470.8$   
constant  $\lambda$   $D_1 = 475.3$   
帰無仮説

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      41 / 44

統計学的な検定      そして、その非対称性

### In case that $P > 0.05$ ...?

You can conclude **NOTHING!**  
**何も結論できない .....**

You can NOT state that constant  $\lambda$  (Null hypothesis) is better  
 $\lambda$  一定のモデルが良いとは言えない

Null hypothesis is never accepted

asymmetry in stat-test  
検定の非対称性 : 帰無仮説 はけっして受容されない

kubostat2017d (https://goo.gl/z9yCJY)      統計モデリング入門 2017 (d)      2017-11-14      42 / 44

モデル選択 と 統計学的な検定 のさまざまな 誤解

model selection statistical test  
4. モデル選択 と 統計学的な検定  
のさまざまな misunderstanding 誤解

kubostat2017d (<https://goo.gl/e9yCjY>) 統計モデリング入門 2017 (d) 2017-11-14 43 / 44

モデル選択 と 統計学的な検定 のさまざまな 誤解

とりあえず FAQ モデル選択  
<http://hosho.ees.hokudai.ac.jp/~kubo/ce/FaqModelSelection.html>

kubostat2017d (<https://goo.gl/e9yCjY>) 統計モデリング入門 2017 (d) 2017-11-14 44 / 44