

# 災害研究に使えそうな 統計解析手法の入門的解説

2015.4.30

人間・社会対応研究部門

被災地支援研究分野

奥村 誠

Mail:mokmr@m.tohoku.ac.jp

計量行動分析のページ

<http://strep.main.jp> から, 講義情報をたどる

# RPとSP、調査の柔軟性とバイアス

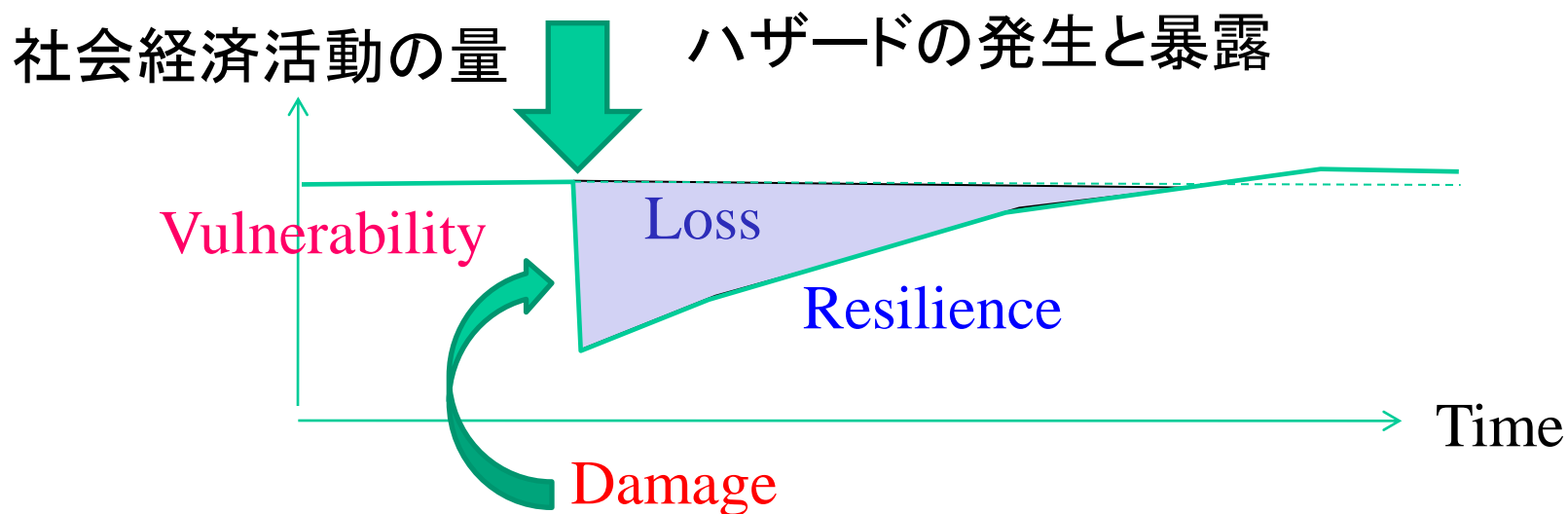
- RP: Revealed Preference 顕示選好(実際の行動)
  - その時、あなたは実際にどう行動しましたか？
  - 経験のない状況に対する行動はわからない
- SP: Stated Preference 表明選好(意向)
  - もし、このような状況になったら、あなたはどうしますか？
  - 現在存在しない状況も、仮想的に設定できる(柔軟性)
  - 仮想的価値評価法CVM(Contingent Valuing Method)
  - 回答と、実際の行動とには大きな差(バイアス)
    - 被験者が、仮想的な状況を理解しづらい
    - 特にメリットに比べ、デメリットの認識がしづらい
    - 調査者の意向を先読みして、好意的回答をする
    - 質問の順序や、言葉遣いが影響を与える
    - 自分の考えより、一般的な道徳規準に合わせた回答

# リスクの認知や対応行動の調査

- 災害のように実経験が少ない事象を扱うため、どうしてもSP(表明選好)に頼りがち
- バイアスの影響が出やすい
  - 「災害への備えをした方がいい」ことはよくわかっているが、実際には「他のことの後回しになって、なかなかできない」という「後ろめたさ」
  - 真偽が問われないアンケート調査で、わざわざ自分の後ろめたい状況を報告する必要なし
  - 実際の自分の状況ではなく、そうあるべき自分の姿を回答してしまう傾向がある
    - 影響を受けそうな直接的な表現を避ける、同じ質問を形を変えて何回か尋ねるなどの工夫が不可欠
    - そのような工夫は、答えにくさにつながり、回答率が減少

# 災害マネジメント論における適応戦略

- Hazard           ハザード: 自然外力の強さ
- Exposure        暴露: 人命, 資産, 土地利用, 活動
- Vulnerability   脆弱性: 社会システムの弱さ
- Resilience      回復力: 回復の速さ



$$\propto \text{Hazard} \times \text{Exposure} \times \text{Vulnerability} \quad 4$$

# 数少ない災害事例(RP)から 政策に役立つ知識・法則性を引き出す

## 脆弱性を小さくするか、回復力を高めるか？

- 要因の政策による変化が、どの程度脆弱性を低減させるかを、客観的・定量的に把握したい(統計手法)

- **脆弱性の定義**(被害／人口・資産): 0-1間の比率

  - 特別な取り扱いが必要(一般化線形モデル)

- 政策操作要因以外にも多くの周辺要因が影響

事例数が少ない、実験はできない

- 周辺要因の値が同じデータを揃えるのは困難

  - **周辺要因の影響を調整**する(傾向スコア法)

# 基本は回帰モデル

- いくつかの変数間に相関関係が存在
- ある変数の値を、別の変数を用いて説明

従属変数、目的変数  
被説明変数

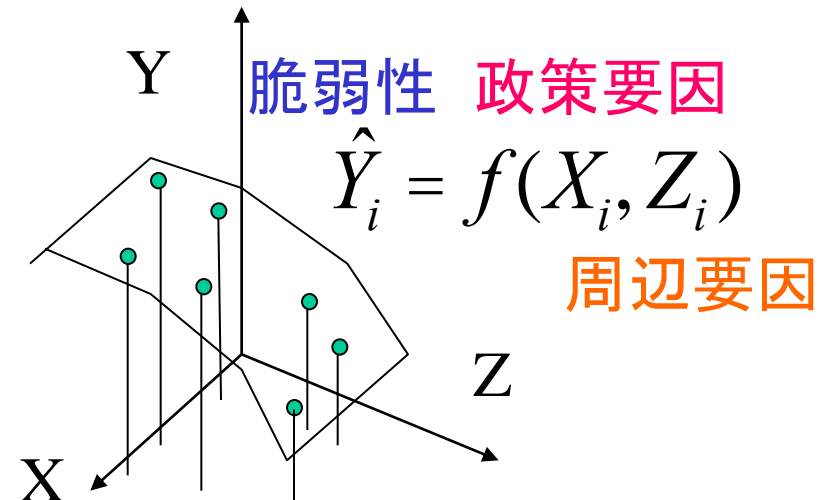
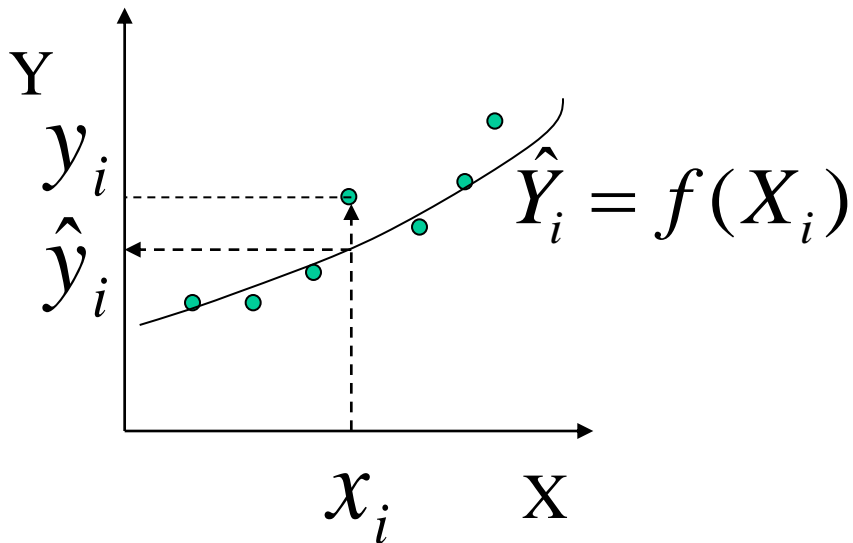
説明式を作成

独立変数、説明変数

変数Y, 実現値 $y_i$

推計値 $y_i = f(x_i)$

変数X, 実現値 $x_i$



通常重回帰式は線形(平面あてはめ) <sup>6</sup>

# Linear Model in R 線形回帰

- Linear Model

- ◆  $y_i = \beta_1 + \beta_2 x_i + \beta_3 f_i + \dots$

- response variable  $\sim$  intercept + slope \* explanatory variable

- ◆  $\text{lm}(y \sim x + f \dots), \text{lm}(y \sim x + f - 1)$  (no intercept)

```
require(graphics)
## Annette Dobson (1990) "An Introduction to Generalized Linear Models".
## Page 9: Plant Weight Data.
ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
group <- gl(2,10,20, labels=c("Ctl","Trt"))
weight <- c(ctl, trt)
lm.D9 <- lm(weight ~ group)
lm.D90 <- lm(weight ~ group - 1) # omitting intercept
anova(lm.D9)
summary(lm.D90)
opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
plot(lm.D9, las = 1)
# Residuals, Fitted, ...
Par(opar)
### less simple examples in "See Also" above
```

# Generalized Linear Models in R

## 一般化線形モデル

- Linear Model

- ◆  $y_i = \beta_1 + \beta_2 x_i + \beta_3 f_i + \dots$

- response variable  $\sim$  intercept + slope \* explanatory variable

- ◆  $\text{lm}(y \sim x + f \dots), \text{lm}(y \sim x + f - 1)$  (no intercept)

- ◆ Generalized Linear Model

- ◆  $f(y_i) = \beta_1 + \beta_2 x_i + \beta_3 f_i + \dots$

- Model & Link function  $\sim$  intercept + slope \* explanatory variable

- ◆  $\text{glm}(y \sim x, \text{data} = d, \text{family} = \text{binomial})$



# Generalized Linear Models

## 一般化線形モデルの種類

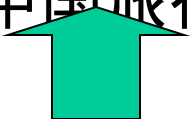
- ◆ Generalized Linear Model  $f(y_i) = \beta_1 + \beta_2 x_i + \beta_3 f_i + \dots$   
glm(y ~ x, data = d, family = **binomial**)
- ◆ Family (Modelled Probability Distribution)
  - ◆ **binomial**(link = “logit”) 2項分布 (規定試行中の発生数)
  - ◆ gaussian(link = “identity”) 正規分布
  - ◆ Gamma(link = “inverse”) ガンマ分布 (正のみ)
  - ◆ inverse.gaussian(link = “1/mu^2”) 逆ガウス分布
  - ◆ poisson(link = “log”) ポアソン分布 (一定時間中の発生回数)
  - ◆ quasi(link = “identity”, variance = “constant”) 正規分布 (不均一)
  - ◆ quasibinomial(link = “logit”) 2項分布 (分散不均一)
  - ◆ quasipoisson(link = “log”) ポアソン分布 (分散不均一)

# ロジットモデルとは (離散的選択のモデル)

- 個人は、採りうる選択肢alternativeを列挙する
- それぞれの選択肢の特徴と費用に基づいて、評価得点utilityをつける
- 評価得点が高いものを選ぶ



中国旅行



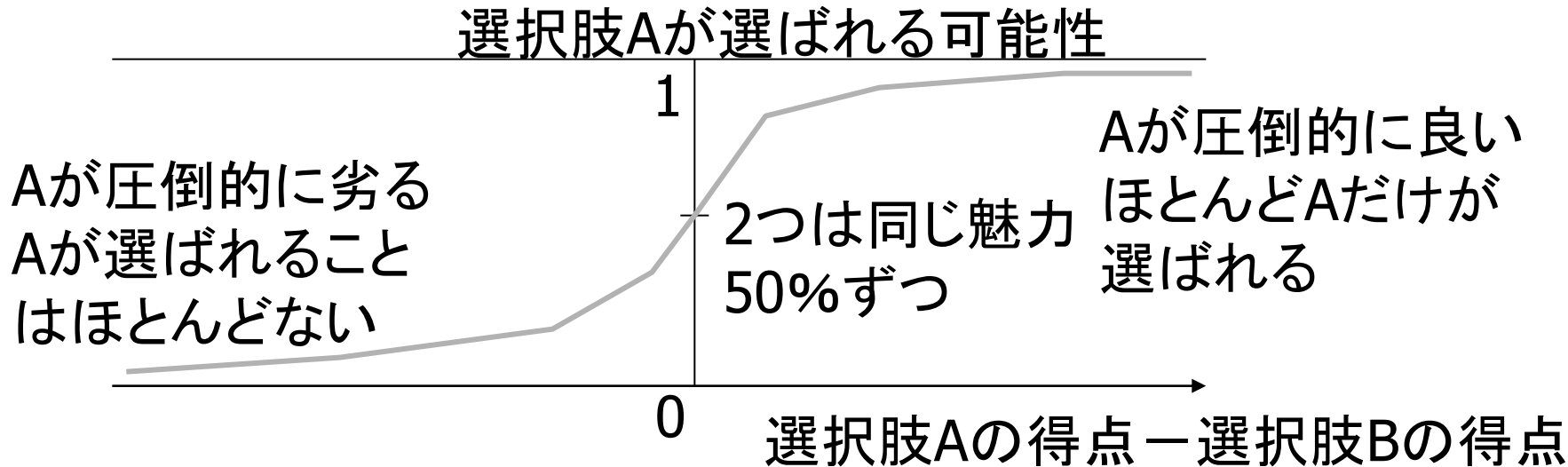
60点 フランス旅行 40点



アメリカ旅行 50点

# 確率的選択：評価点の差と選択率

- 実際には
  - ほとんど評価点と同じときは、どちらも選択される可能性がある
  - 評価点の差が大きいたときは、片方しか選ばれない。



ある事象が発生するかしないかの確率を表現できる

# ロジットモデル(ロジスティック回帰)

- S字型の曲線として,

$$P_1 = \frac{\exp(\eta V_1)}{\exp(\eta V_1) + \exp(\eta V_2)} = \frac{1}{1 + \exp[\eta (V_2 - V_1)]}, \quad P_2 = 1 - P_1$$

という式で表わされる曲線を使うと,

- いろいろな計算が簡単にできる
- 3つ以上の選択肢からの選択も同じ形になる
- 2000年ノーベル経済学賞  
McFadden (1937-) が提案
  - 各自の評価点が安定している部分と確率的に変動する部分の和である場合の選択から理論的に導いた。(ランダム効用モデル)

# Binomial Logistic Model

(occurrence number in given trials)

◆ Binomial Model for the number of survived plant in 8 observations, regressed on plant size and nutrification (p118)

◆ Maximize log-likelihood

$$p(y | N, q) = \binom{N}{y} q^y (1 - q)^{N - y}$$

$$q_i = \text{logistic}(z_i) = \frac{1}{1 + \exp(-z_i)}$$

$$z_i = \log \frac{q_i}{1 - q_i}$$

`glm(cbind(y, N-y) ~ x + f, data = d, family = binomial)`

```
#page 117 plant data
d <- read.csv("data4a.csv")
d$N # number of trials
d$y # number of survived plant
d$x # plant size
d$f # nutrification (treat-control)
plot(d$x, d$y, pch = c(21, 19)[d$f])
```

```
# model p122
fit.all <- glm(cbind(y, N-y) ~ x + f, data=d,
family=binomial)
print(fit.all)
logLik(fit.all)
```

# 東日本大震災における 津波伝承知メディアの減災効果 -地名と津波碑を対象として-

一般化線形モデルの適用例として  
佐藤翔輔先生に  
データをいただきました

津波工学研究室	鹿島 七洋
指導教員	今村 文彦
研究指導教員	佐藤 翔輔

# はじめに

## ー背景ー

我が国には地名、碑文、口承など津波の経験を後世に伝える有形無形の媒体「津波伝承知メディア」が存在する。津波被害軽減効果を目的として生まれる「津波伝承知メディア」であるが、それらが真に津波被害軽減効果を有しているかは定量的には明らかにされていない。

## ー目的ー

本研究では、津波伝承知メディアである津波由来地名と津波碑に着目し、東日本大震災の主な被災地である岩手・宮城・福島における津波由来地名と津波碑を整理・分類し、津波由来地名と津波碑が本大震災において人的被害の軽減に影響を及ぼしたかどうかを明らかにする。



昭和8年大津波碑(岩手県宮古市姉吉地区)





# 研究方法-分析-

3県・地形別の基礎情報

	岩手県	宮城県	福島県	リアス部	平野部
対象町大字数	90	324	136	183	367
人口(人)	164,221	454,582	172,780	261,552	530,031
死者数(人)	4,374	8,743	1,359	7,184	7,292
死亡率(%)	2.66	1.92	0.79	2.75	1.38
津波由来地名	5	33	15	17	36
碑文数	211	68	0	269	1

津波伝承知メディアが  
減災効果を有している  
かどうか明らかにする

## 検討1 津波碑文数と死亡率の相関関係

各町大字の津波碑文数と死亡率から散布図を作成し、傾向を検証。

## 検討2 津波伝承知メディアの有無による平均死亡率の差の検定

県ごと、地形ごとに津波由来地名有無地区、津波碑有無地区それぞれの死亡率を算出し、平均値の差が有意であるかどうか検証。

## 検討3 各町大字の津波最大高を取り入れた重回帰分析による検定

**目的変数**: 死亡率

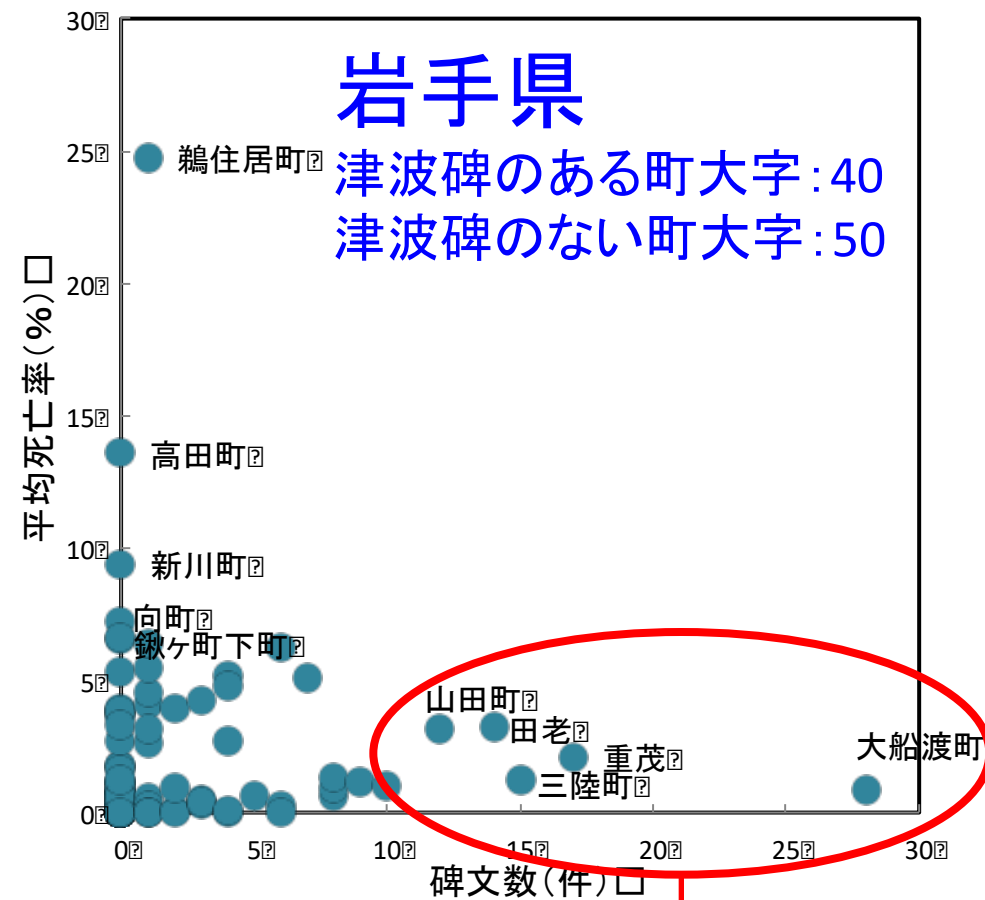
**説明変数**: 最大津波高、津波由来地名有無、津波碑数

として各県・3県・リアス部に対し重回帰分析(強制投入法、ステップワイズ法)を行った。

# 検討1: 各町大字の津波碑文数と死亡率の相関

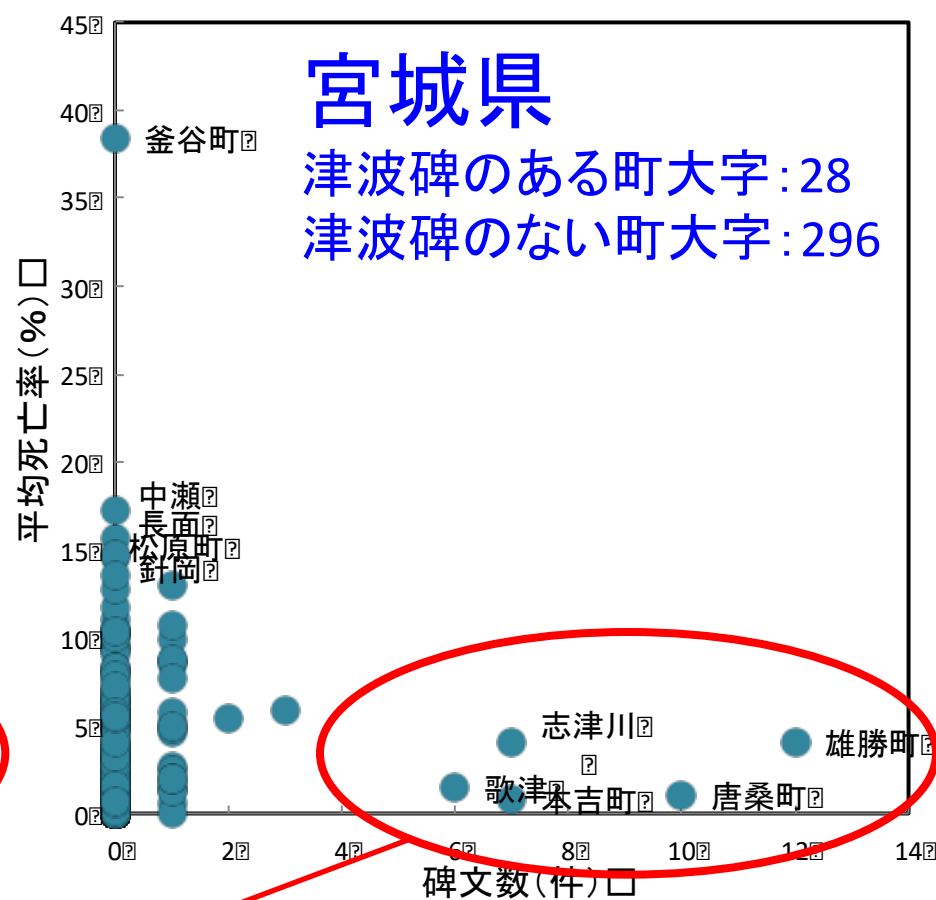
## 岩手県

津波碑のある町大字: 40  
津波碑のない町大字: 50



## 宮城県

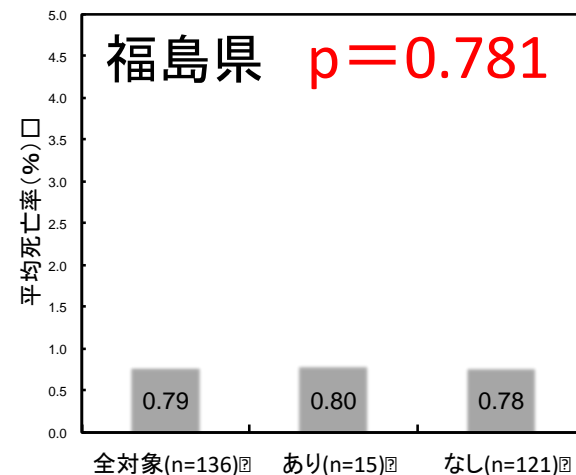
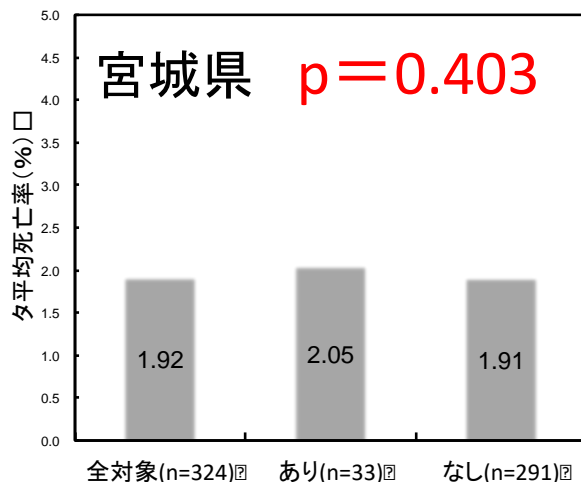
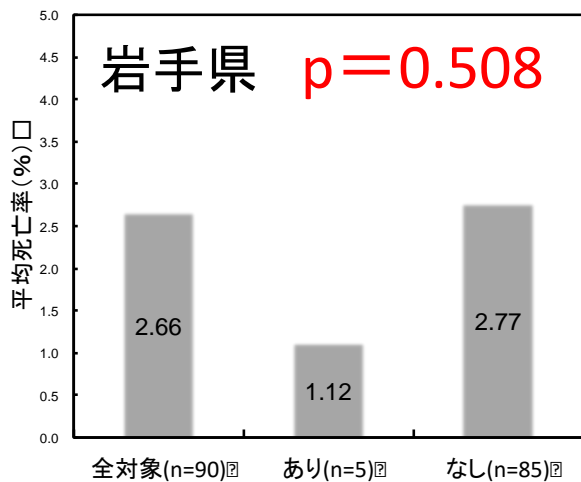
津波碑のある町大字: 28  
津波碑のない町大字: 296



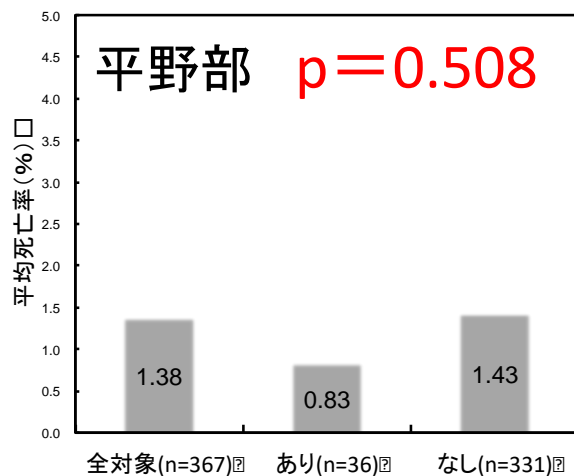
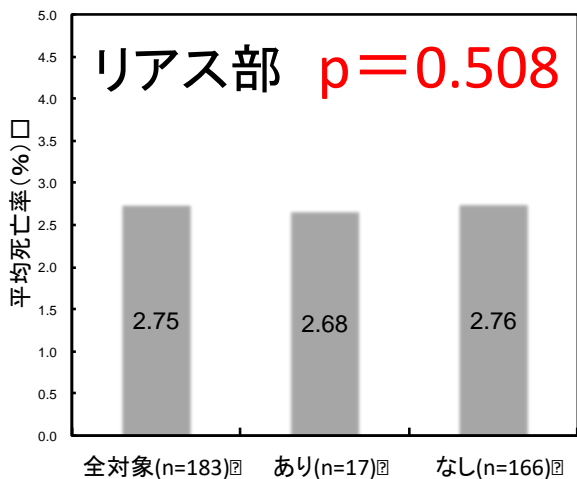
最大津波高はおよそ15~35m ⇔ 死亡率はいずれも5%以下

# 検討2: 平均死亡率の差の検定結果-津波由来地名-

## 県別



## 地形別

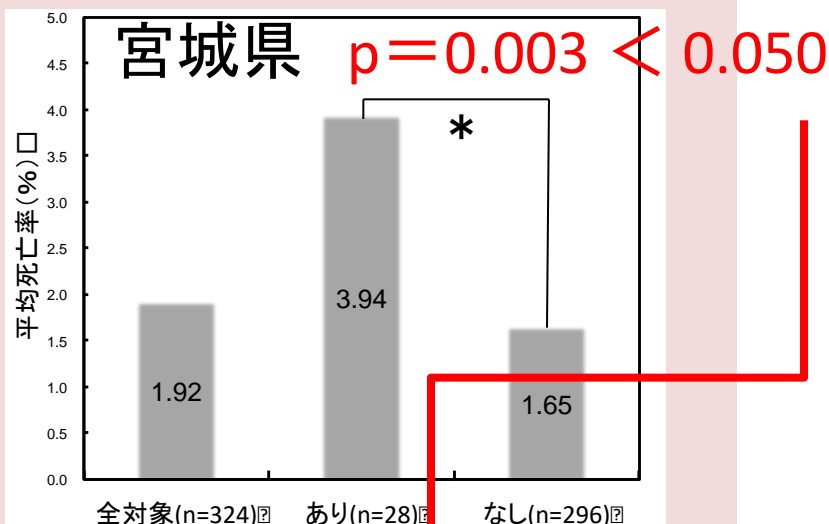
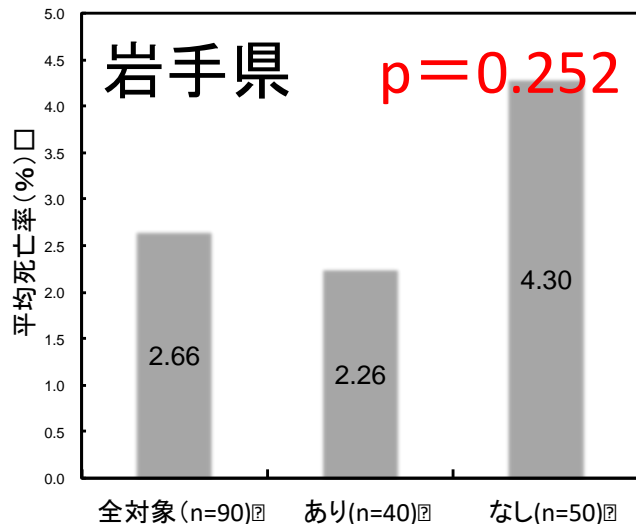


有意性が認められる  
組み合わせなし

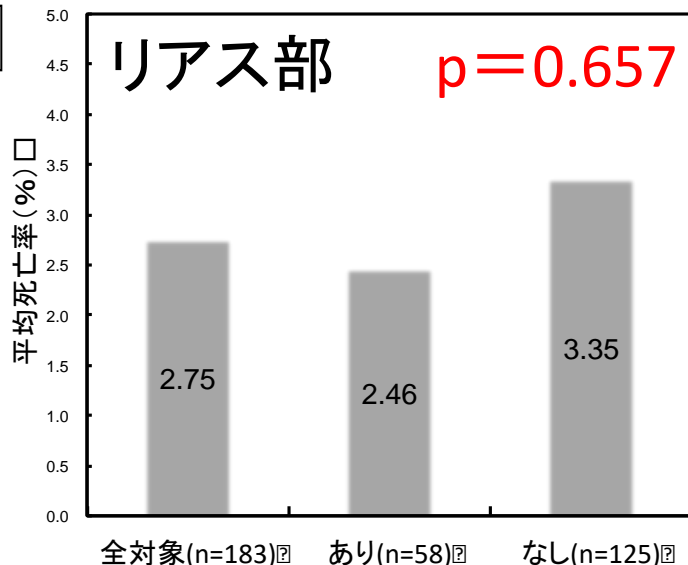
※  $p < 0.05$ で有意と言える

# 検討2: 平均死亡率の差の検定結果-津波碑-

県別



地形別



8つの組み合わせのうち **宮城県、津波碑有無** の組み合わせにのみ有意性が見られるが、死亡率は **碑のある地域 > 碑のない地域**

考察

- ①津波碑が存在する = 過去に津波被害！  
⇒ 今回も津波が襲来 (津波碑効果薄い?)

# 検討3:重回帰分析の結果-津波由来地名・津波碑-

## ○強制投入法

津波由来地名有無、津波碑数の有意性はほとんどの組み合わせで認められなかった( $p = .101 \sim .996$ )が3県で行った津波碑数にのみ負の相関の有意性が見られた(碑文数が増加すると死亡率が低下する)。

強制投入法による重回帰分析結果(3県)

説明変数	標準化されていない係数		標準化係数	t値	有意確率
	B	標準誤差	ベータ		
(定数)	1.561	.278		5.605	.000
最大津波高	.135	.031	.246	4.309	.000
津波碑数	-.211	.087	-.139	-2.433	.015 < .050
津波由来地名	.213	.635	.016	.336	.737

## ○ステップワイズ法

宮城県、3県にのみ適用されたが、津波由来地名有無、碑文数はいずれも除外された(死亡率に対する説明変数にはならなかった)。

# おわりに

## —まとめ—

- ・津波由来地名は減災効果を有していない
- ・津波碑は減災効果を有している
  - ⇒津波碑が存在する地域は防災意識自体が高いと考えられる

## —今後の課題—

- ・他の津波伝承知メディアを統計分析
- ・インタビューやアンケートなどの実施
  - 津波伝承知メディアの認知度等の把握

# 重回帰分析でしていること

Call:  
lm(formula = drate ~ wave + exstone + name)

Residuals:

Min	1Q	Median	3Q	Max
-4.745	-1.958	-1.453	0.700	35.698

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.73025	0.27616	6.265	9.41e-10 ***
wave	0.08876	0.03265	2.719	0.00683 **
exstone	0.19045	0.61058	0.312	0.75526
name	0.16410	0.64200	0.256	0.79838

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.75 on 410 degrees of freedom

Multiple R-squared: 0.03042, Adjusted R-squared: 0.02333

F-statistic: 4.288 on 3 and 410 DF, p-value: 0.005375

死亡率  
(%)

drate

10

0

10

wave

○ 津波碑有り  
□ 津波碑無し

回帰直線の  
切片が違くと  
考える

津波高

# 死亡率の定義に戻ると

死亡率 = 死亡者数 / 居住人口 (本当は昼間人口であるべき)

地域ごとに、居住者の一人一人が、同一の死亡確率にさらされて、たまたまその中のある人数が死亡してしまった

赤玉と白玉が一定の割合で入っている壺から、玉を一つ取り出しす試行を繰り返した場合の、赤玉の出現回数

死亡率がその地域の説明要因のロジット関数として、0-1の間の値で与えられ、それが居住人口の一人一人に試行されて、結果として何人かが死亡した。

二項分布 ロジットリンクの 一般化線形モデル



# 一般化線形モデル(二項分布ロジットリンク)

```
result2 <- glm(cbind(death,pop-death)~wave+exstone+name, family = binomial)
```

```
Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.157152  0.014143 -293.930 < 2e-16 ***
wave         0.034992  0.001332  26.264 < 2e-16 ***
exstone     -0.214065  0.030575  -7.001 2.54e-12 ***
name        -0.156785  0.033607  -4.665 3.08e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
```

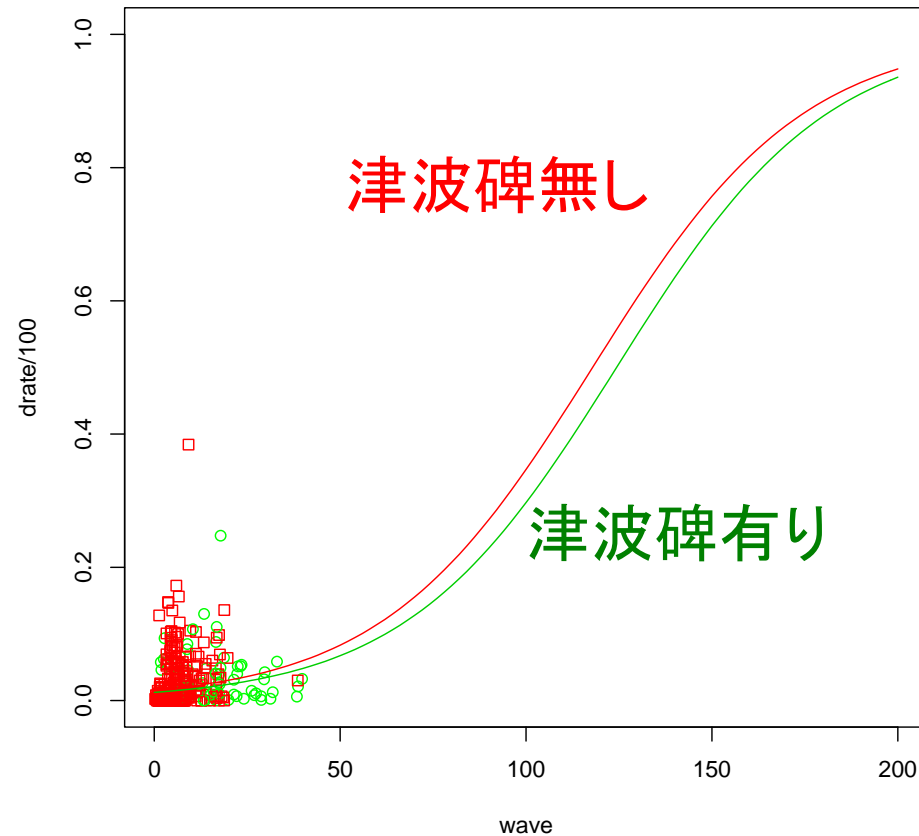
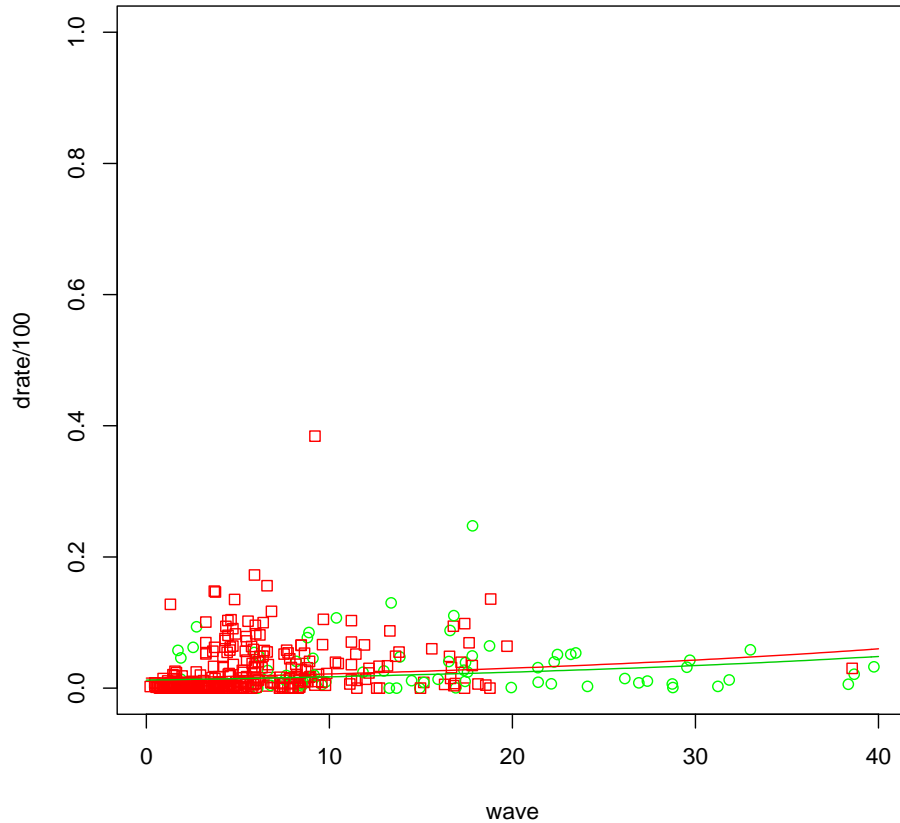
もちろん津波高が最も死亡率に強く影響

津波碑があること、地名が残っていることは、有意に死亡率を低くしている！

```
Null deviance: 21451 on 413 degrees of freedom
Residual deviance: 20274 on 410 degrees of freedom
AIC: 21694
```

```
reg1 <- function(w) 1/(1+exp(4.169587-0.035371 * w))
reg2 <- function(w) 1/(1+exp(4.169857+0.227790 - 0.035371 * w))
plot(wave,drate/100,bg=c(2,3), pch=as.numeric(isstone))
curve(reg1, col=2, add =TRUE)
curve(reg2, col=3, add =TRUE)
```

# 死亡率のS字曲線は少し右にずれた！



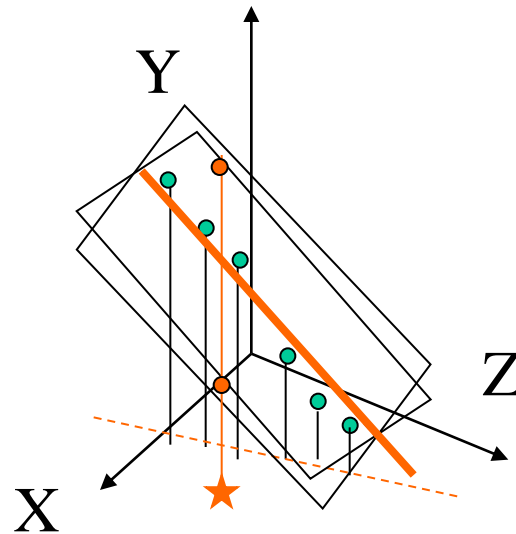
左図の○と□に合うように2つの(左右にずれた)S字曲線を当てはめ  
線形回帰のときとは、効果が逆に出た！

同じ死亡率でも、居住者数が違えば2項分布の確率が異なるため

# 第2の問題：重共線性

- 多数の説明変数の間に相関がある場合
  - 目的変数への効果を一意に分離できない
  - 係数の推計値が安定しない(直感に反する符号を取るなど)

すべての観測値が  
ほぼ一直線上にある



この直線を含むような平面で  
あれば、どの式を使っても当て  
はまりにはほとんど差はない

直線上にない場所のYの予測  
値には大きな差がでる

過去の津波高が高いほど、津波碑が多く残っている

# 今回の津波高と、津波碑の存在

```
result1 <- glm(exstone ~ wave, family=binomial)
```

```
>summary(result1)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.0909	-0.4353	-0.3283	-0.2562	2.5955

Coefficients: Estimate Std. Error z value Pr(>|z|)  
(Intercept) -3.71149 0.31817 -11.665 <2e-16 \*\*\*  
wave 0.21992 0.02617 8.405 <2e-16 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

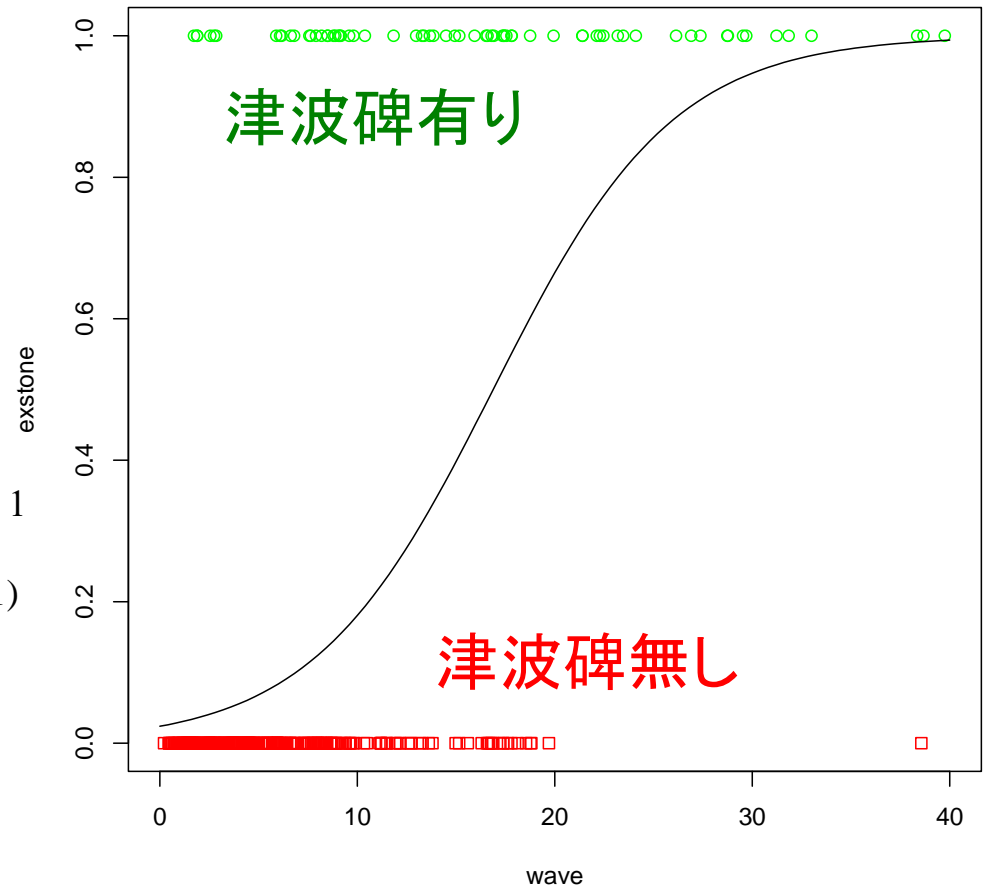
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 369.83 on 413 degrees of freedom

Residual deviance: 253.06 on 412 degrees of freedom

AIC: 257.06

Number of Fisher Scoring iterations: 5

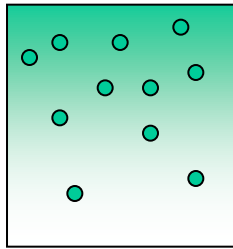


津波碑の存在自体をロジットモデルに当てはめると、  
有意なモデルができる

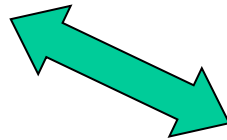
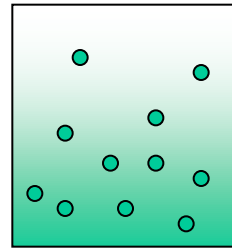
# マッチング法のアイデア

- 他の条件が同じで津波碑があった地域と津波碑がなかった地域の死亡率を比べたい

津波碑有り地域



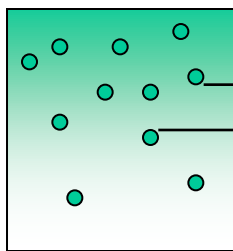
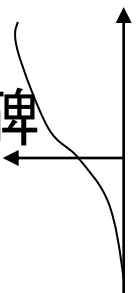
津波碑ない地域



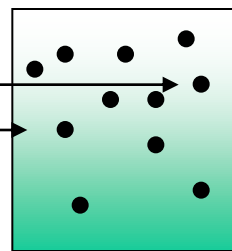
周辺要因（津波高）の違うサンプル間を比較するので、死亡率の違いが周辺要因の違いによるものか、津波碑の存在による影響かがわからない

津波高

津波碑存在確率



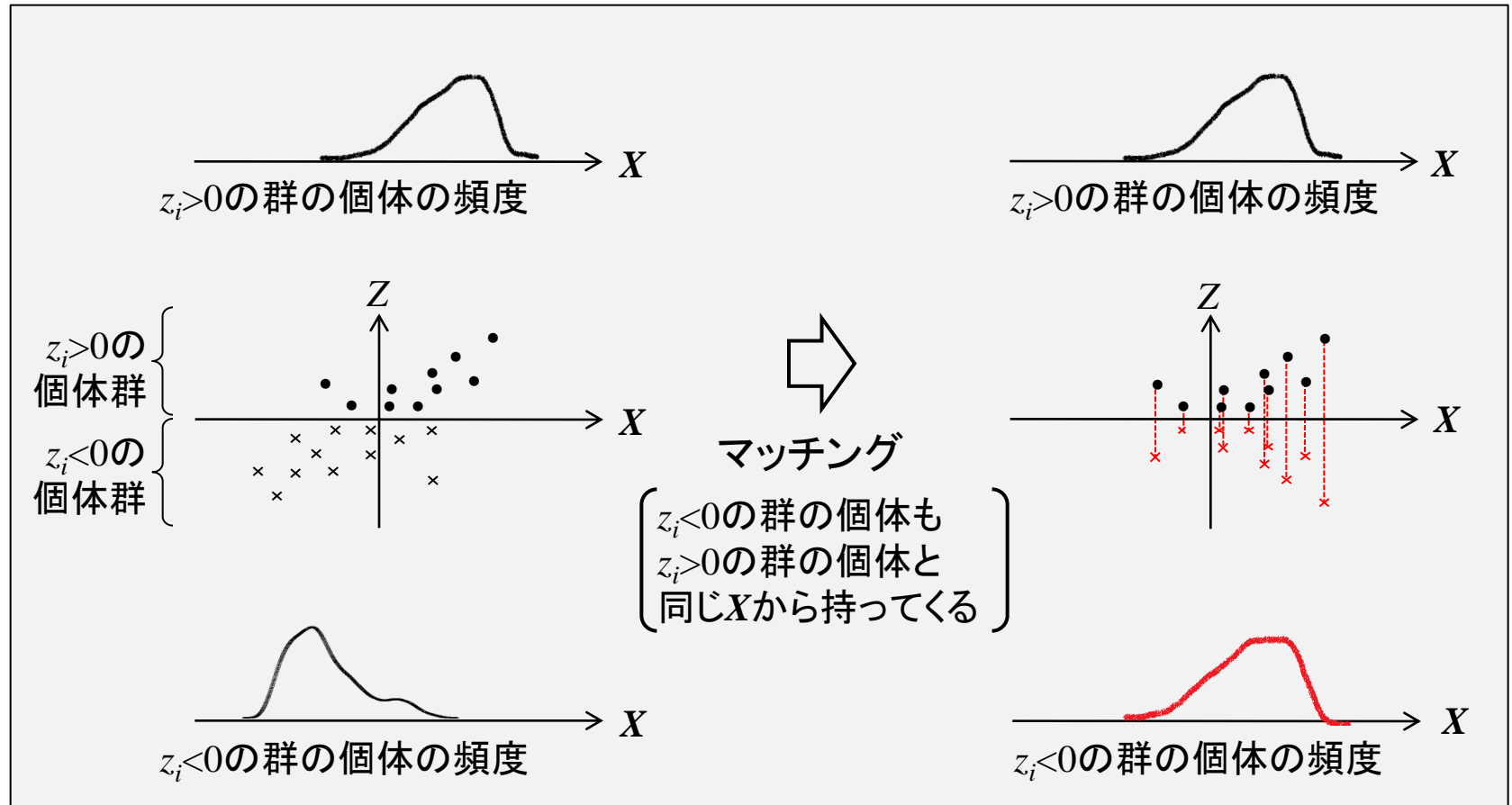
津波碑が有る地域に対して、ない地域の中から最も似た地域を選ぶ



津波碑があってもおかしくない地域で、たまたま津波碑がなかった地域を、比較の相手に持ってくる

# 傾向スコアマッチング(考慮すべき周辺要因が多いとき)<sup>30</sup>

- 傾向スコア  $e_i(x_i)$  (関数) の値に基づき, 比較する個体を選定.



$X$ と $Z$ の相関が消え, 多重共線性が解消される

図2 傾向スコアマッチング

# 傾向スコア

## 傾向スコアの定義

個体  $i$  の着目する変数を  $z_i$ , その他の説明変数の値を  $x_i$  とすると, 個体  $i$  が  $z_i > 0$  の群へ割り当てられる確率  $e_i$  を傾向スコアという ( $0 \leq e_i \leq 1$ ).

$$e_i = p(z_i > 0 | x_i)$$

» ロジスティック回帰モデルにより, 個体  $i$  の傾向スコア  $e_i$  の推定を行う.

$$p(z_i > 0 | x_i) = e_i = \frac{1}{1 + \exp\{-\alpha^t x_i\}}$$

このとき,  $z$  に関する尤度は,

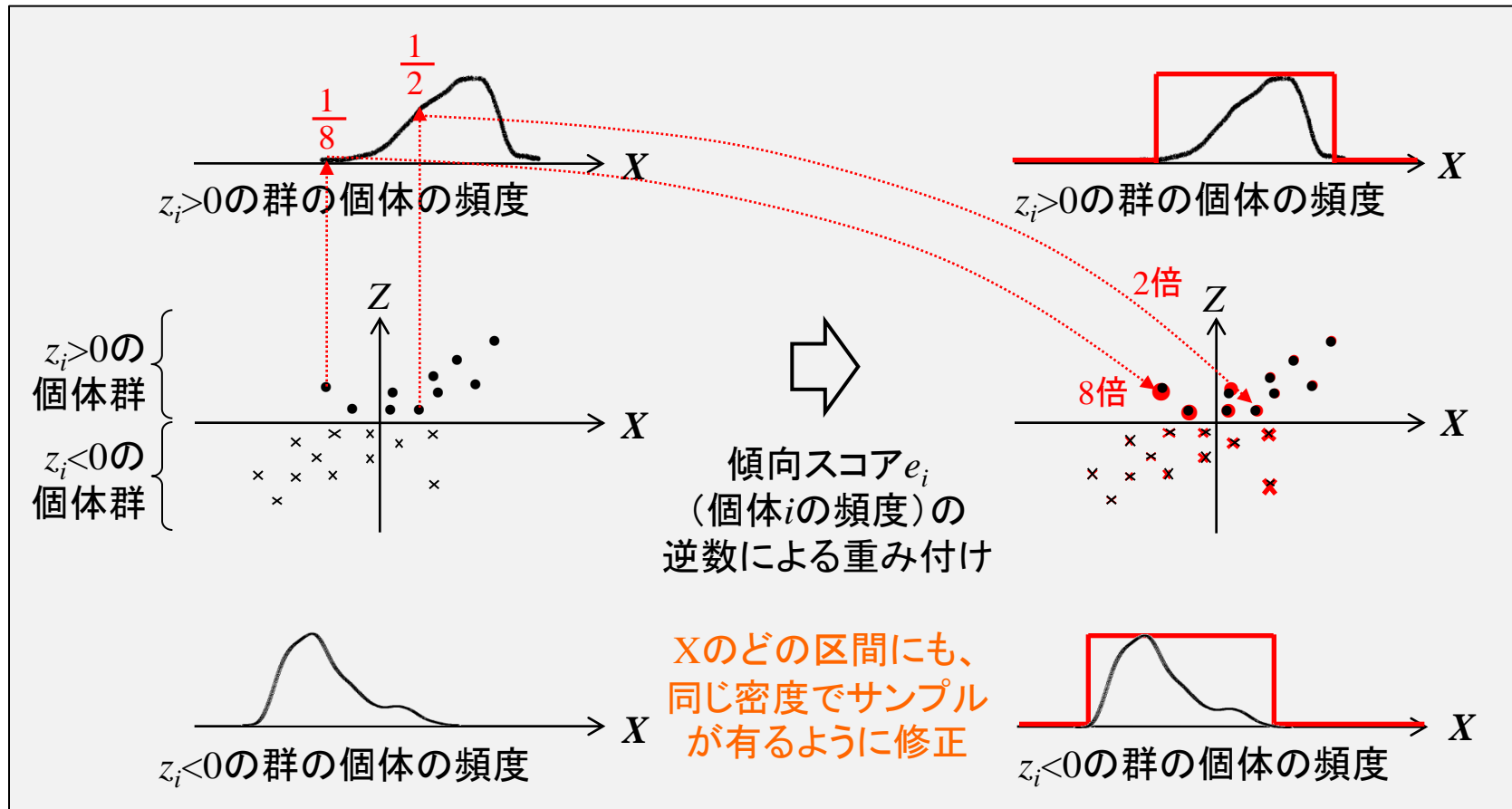
$$\prod_{i=1}^n \left( \frac{1}{1 + \exp\{-\alpha^t x_i\}} \right)^{z_i} \left( 1 - \frac{1}{1 + \exp\{-\alpha^t x_i\}} \right)^{1-z_i}$$

式(3)を最大化する最尤推定値  $\hat{\alpha}$  を用いることで, 個体  $i$  の傾向スコアの推定値は以下のように表される.

$$\hat{e}_i = \frac{1}{1 + \exp\{-\hat{\alpha}^t x_i\}}$$

# 傾向スコアによる重み付け推定法

- 傾向スコア  $e_i$  の逆数を重みとして与え、回帰分析を行う。
- 分布が少ないところに存在する個体数を拡大する。



➡  $X$  と  $Z$  の相関が消え、多重共線性が解消される

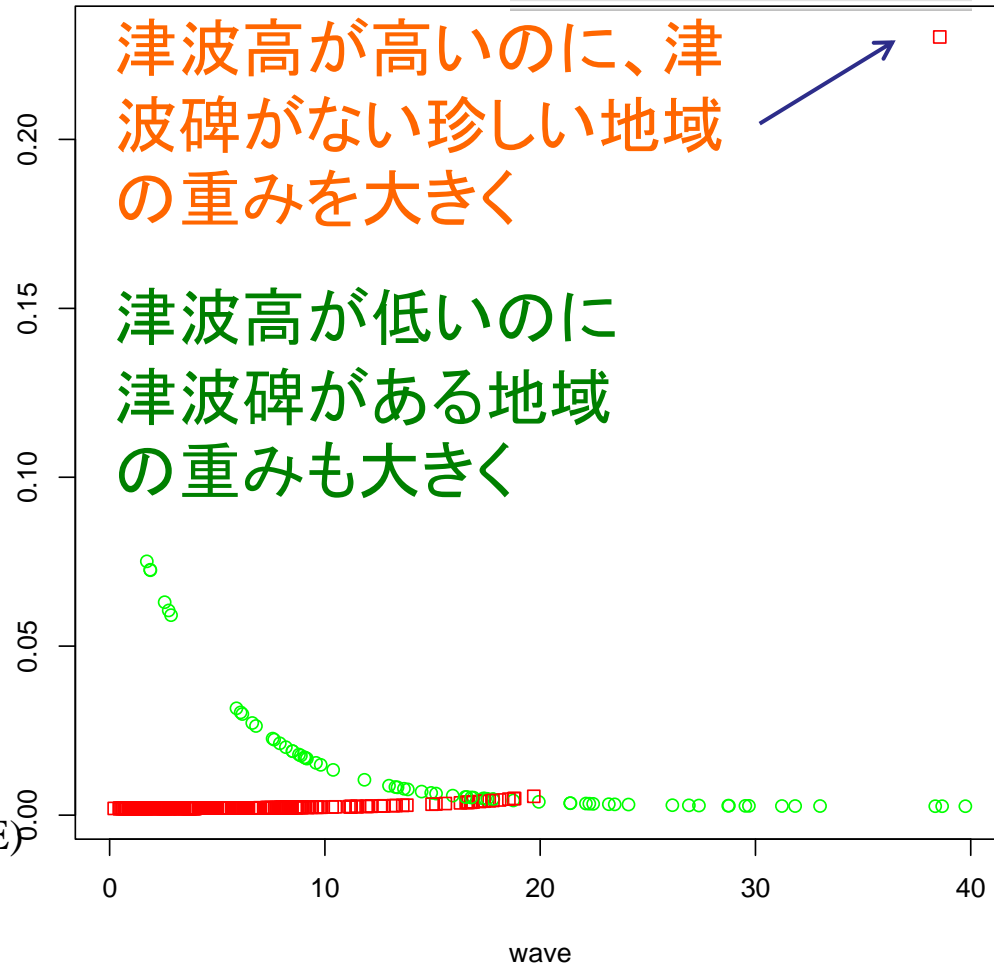
図3 傾向スコアによる重み付け推定法



# 傾向スコア値の算定と逆数の重みの付与

```
result1 <- glm(exstone ~ wave, family=binomial)
summary(result1)
ir=c("red","green")
bg=c(2,3)
plogit <- function(x)
plogit <- 1/(1+exp(3.7115-0.2199*(x)))
plot(wave,exstone, xlim=c(0,40), ylim=c(0,1),
pch=as.numeric(isstone),col=ir[exstone+1])
curve(plogit, xlim=c(0,40), ylim=c(0,1),add=TRUE)
```

```
pstone <- plogit(wave)
cnt1 <- sum(1/pstone * exstone)
cnt2 <- sum( 1/(1-pstone) * (1-exstone))
wgt <- 1/pstone * exstone /cnt1+ 1/(1-pstone)*(1-exstone)/cnt2
plot(wave,wgt, xlim=c(0,40), pch=as.numeric(isstone),col=ir[exstone+1])
```



# 重みを与えた一般化線形モデルの推定

```
result4 <- glm(cbind(death,pop-death)~wave+exstone+name, family = binomial,  
weight = wgt)  
summary(result4)
```

```
Coefficients: Estimate Std. Error z value Pr(>|z|)  
(Intercept) -3.98222  0.24319 -16.375 <2e-16 ***  
wave         0.02229  0.01050  2.123  0.0337 *  
exstone      0.02099  0.25800  0.081  0.9352  
name        -0.22616  0.51646 -0.438  0.6615
```

```
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)  
Null deviance: 83.498 on 413 degrees of freedom  
Residual deviance: 79.204 on 410 degrees of freedom  
AIC: 94.743  
Number of Fisher Scoring iterations: 5
```

高台までの距離や高齢化率などを加える

津波碑の存在も  
地名の存在も、  
統計的に有意で  
はなくなった

今回のデータで  
は、両者の効果  
を分離できるほど、  
十分なサンプル  
がなかった？

# Rに関する情報は

- RjpWiki <http://www.okada.jp.org/RWiki/>

RjpWiki - RjpWiki - Windows Internet Explorer

http://www.okada.jp.org/RWiki/

Google R Wiki

RjpWiki - RjpWiki

**RjpWiki**  
http://www.okada.jp.org/RWiki/?RjpWiki

[ トップ | Tips紹介 | 中級Q&A | 初級Q&A | R掲示板 | 日本語化掲示板 | リンク集 ]  
[ リロード ] [ 新規 | 編集 | 凍結 | 差分 | ファイル添付 ] [ 一覧 | 検索 | 単語検索 | 最終更新 | バックアップ | ヘルプ ]

**RjpWiki** はオープンソースの統計解析システム『R』に関する情報交換を目的とした Wiki です

どなたでも自由にページを追加・編集できます。  
注意！コメント欄への新規投稿は↑の「編集」ではありません！コメント欄の下のお名前:のところです！  
(初めて投稿・既存記事への追加・修正を行なう方はこのページ末の注意\*を御覧下さい)  
ページへのファイル添付については、画像ファイルのみパスワードなしで可能としてあります(ページ上部「画像添付」より)。その他のファイルの添付はパスワードを入力することで可能です(ページ上部「ファイル添付」より)。現在のパスワードは、Rでの round (qt(0.2,df=8),3) の実行結果です。  
スパム書き込みに対処するため、書き込み系の処理に対してパスワードを設けました。ユーザ名の欄には、Rで round (qt(0.2,df=8),3) を実行したときの結果を入力します。パスワード欄には何も入力しないままでOkです。  
Rを起動して、文字がたかさんでいるウィンドウの">"のあとに、round(qt(0.2,df=8),3) をコピーペーストしてEnterキーを押せば、結果が[1] xxxxxx のように表示されます。このxxxxxxの部分をユーザ名として入力します。もし、どうしてもうまくいかなかったら...何度か「キャンセル」ボタンを押してみよう。

**主な内容** (全ての内容を見るには上部のメニューの【一覧】をクリックしてください。)

- 《Rとは》その公式紹介、《Rのインストール》Rを始める気になったら、《R-Online》その前に一寸試してみたかったら
- 《R 2.10.0 の変更予定》《R 2.9.2 の変更予定》《R 2.9.1 の変更予定》《R 2.9.0 の変更予定》《R 2.8.1 の変更予定》  
《R 2.8.0 の変更予定》《R 2.7.2 の変更予定》《R 2.7.1 の変更予定》《R 2.7.0 の変更予定》《R 2.6.0 の変更予定》《R 2.5.1 の変更予定》  
《R 2.5.0 の変更予定》《R 2.4.1 の変更点》《R 2.4.0 の変更点》《R 2.3.0 の変更点》《R 2.2.1 の変更点》  
《R 2.2.0 の新機能・変更》《R 2.1.1 の変更点》《R 2.1.0 の変更点》

本日更新パッケージ

# R のインストール

(現在休止中)

最新の30件

Windows, Mac, Linux(redhat, yellowdog, vine,...), Unix のインストールのコツなどを付け足してください。

2014-10-06

• R掲示板

2014-10-04

• R史

2014-09-30

• Q&A (初級者  
コース)/16

• Q&A (初級者  
コース)/15

2014-09-26

• R本リスト

2014-09-25

• The MakeR Wa  
y

• Rで機械学習

2014-09-21

• Rを使った学術論  
文

2014-09-19

• RでGIS

2014-09-17

• ESS

2014-09-13

• SQLiteMap(SQL  
iteを使ったベク  
タグラフィック地  
図の管理)パッ  
ケージ中のオブ  
ジェクトの一覧

2014-09-04

• SPSSでR

• Rでスポーツ統計

2014-09-02

• QGISでR

2014-08-30

• Rがインポート・エ  
クスポートできる  
データ形式

2014-08-29

• PythonでR

• Recent Deleted

## Windows 版 R のインストール

- Windows Vista での問題点
- 参考リンク

## Mac 版 R のインストール

- Mac OSX
- 参考リンク
- Fink 版 R のインストール
- Mac Ports 版 R のインストール
- バイナリの提供されていないパッケージ

## Linux 版 R のインストール

- Debian GNU/Linux (含む Knoppix)の場合 (インターネット経由でオリジナル版をインストールする場合)
- openSUSE10.2 Linux 版インストール
- Ubuntu Linux の場合:

## Unix 版 R のインストール

- ソースからのコンパイル法 (含む Linux)

## REvolutionR(IntelMKL版R:マルチコア対応:Windows,MacOSX)

- Linuxでソースからのビルド(VineLinuxを例として)
- 興味のある分野のパッケージを丸ごとインストールする

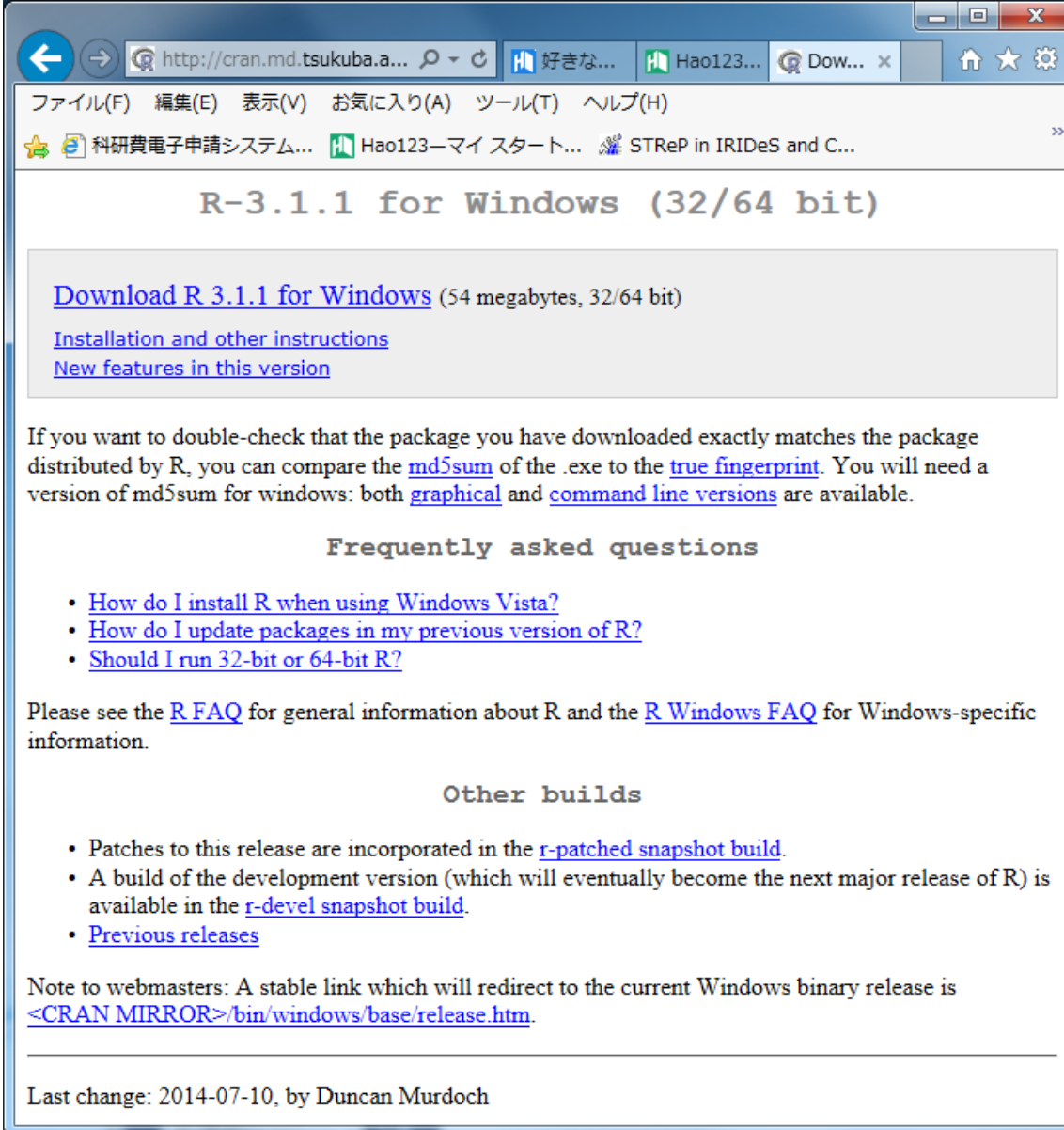
## Windows 版 R のインストール <sup>+</sup>

最新版は[こちら](#)からダウンロードしてください。

## Windows Vista での問題点 <sup>+</sup>

Vistaにインストールして動かないと嘆いている方はここを読んでください。[http://cran.r-project.org/bin/windows/run-under-Windows-Vista\\_003f](http://cran.r-project.org/bin/windows/run-under-Windows-Vista_003f)

# 最新版は3.1.1 (10.10現在)



The screenshot shows a web browser window with the address bar displaying <http://cran.md.tsukuba.a...>. The browser's menu bar includes options like 'ファイル(F)', '編集(E)', '表示(V)', 'お気に入り(A)', 'ツール(T)', and 'ヘルプ(H)'. The page content is for 'R-3.1.1 for Windows (32/64 bit)'. It features a prominent blue link: 'Download R 3.1.1 for Windows (54 megabytes, 32/64 bit)'. Below this are two more blue links: 'Installation and other instructions' and 'New features in this version'. A paragraph of text explains how to verify the downloaded package using md5sum. A section titled 'Frequently asked questions' lists three common queries with blue links. Another paragraph refers to 'R FAQ' and 'R Windows FAQ'. A section titled 'Other builds' lists three additional build options with blue links. At the bottom, a note to webmasters provides a stable link to the current Windows binary release. The footer indicates the last change was on 2014-07-10 by Duncan Murdoch.

R-3.1.1 for Windows (32/64 bit)

[Download R 3.1.1 for Windows](#) (54 megabytes, 32/64 bit)

[Installation and other instructions](#)

[New features in this version](#)

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the [md5sum](#) of the .exe to the [true fingerprint](#). You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.

**Frequently asked questions**

- [How do I install R when using Windows Vista?](#)
- [How do I update packages in my previous version of R?](#)
- [Should I run 32-bit or 64-bit R?](#)

Please see the [R FAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-specific information.

**Other builds**

- Patches to this release are incorporated in the [r-patched snapshot build](#).
- A build of the development version (which will eventually become the next major release of R) is available in the [r-devel snapshot build](#).
- [Previous releases](#)

Note to webmasters: A stable link which will redirect to the current Windows binary release is <http://<CRAN MIRROR>/bin/windows/base/release.htm>.

---

Last change: 2014-07-10, by Duncan Murdoch

災害研での研究において、統計分析を使う場面は少なくない

しかし、災害に関するデータの特徴にあった分析手法が使われているとは言いがたい

例) 被害率曲線の推定

少なくとも他分野で一般化してきている手法を勉強して、恥ずかしくない程度使いこなしたい

- 後期金曜に授業科目を提供しています
- 私のわかる範囲で相談に乗ります
- また、一緒に勉強していきます！

## 後期金曜2限 計量行動分析

- 1 (10/3) 計量行動分析の意義と3つの統計学の考え方 Purpose.ppt
- 2 (10/10) R言語の導入と記述統計学 IntroductionR.ppt
- 3 (10/17) 推測統計学と仮説検定 PointEstimate.ppt
- 4 (10/24) 推測統計学と仮説検定
- 5 (10/31) 回帰分析の記述統計学的方法
- 6 (11/7) 回帰分析の記述統計学的方法 LinearRegresson.ppt
- 7 (11/21) 回帰分析への推測統計学の応用
- 8 (11/28) **ロジットモデル**の誘導 Logit.ppt
- 9 (12/5) 最尤法による非集計ロジットモデルの推定
- 10 (12/12) 因子分析・主成分分析Factor.ppt
- 11 (12/19) 共分散構造モデルの推定 SEM.ppt
- 12 (1/9) **一般化線形モデル**の考え方glm.ppt(1/25改訂)
- 13 (1/16) 一般化線形モデル推定
- 14 (1/23) 課題発表会1
- 15 (1/30) 課題発表会2

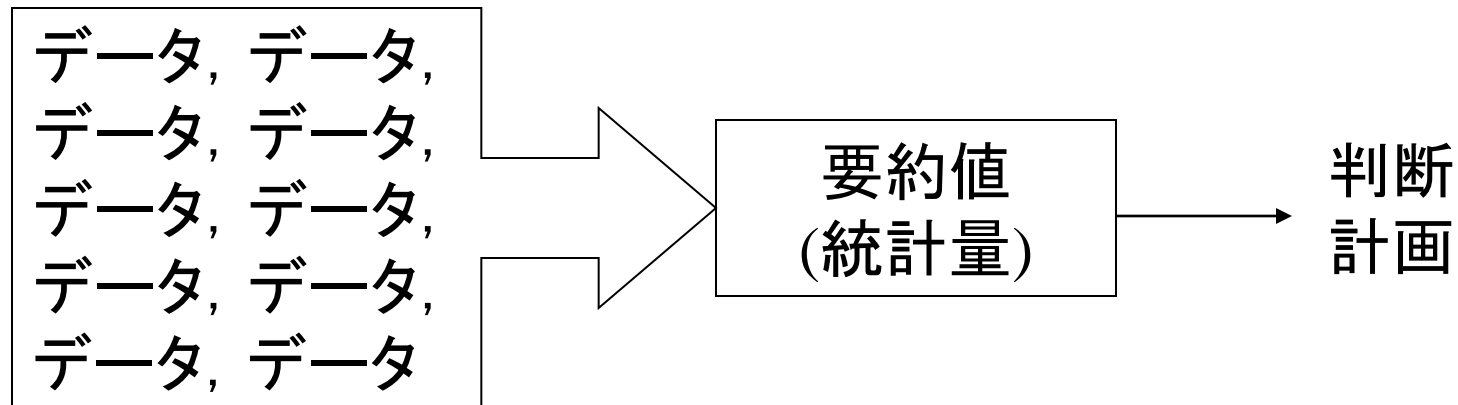
# 統計学 (Statistics) の発展

- 統計学の始まり(紀元前3000年～2300年)
  - 古代エジプト:ピラミッド建設のための基礎調査
  - 古代中国:人口調査
  - 17世紀頃:国勢調査の学問 status(国家)→statistics
- 記述統計学(19世紀末～20世紀初頭)
  - ゴールトン(Francis Galton)、ピアソン(Karl Pearson)
  - データを要約し調査対象の情報を数学的に記述する方法
- 推測統計学(1925年)
  - フィッシャー(Rinald Aylmer Fisher)「研究者のための統計的方法」
  - 標本集団の要約値から母集団の要約値を確率的に推測し、それによって母集団の様子を記述する
- ノンパラメトリック手法
  - 母集団の確率分布を事前に仮定しない方法
- ベイズ統計学
  - 観測値に基づき, 母集団に関する知見を順次修正する



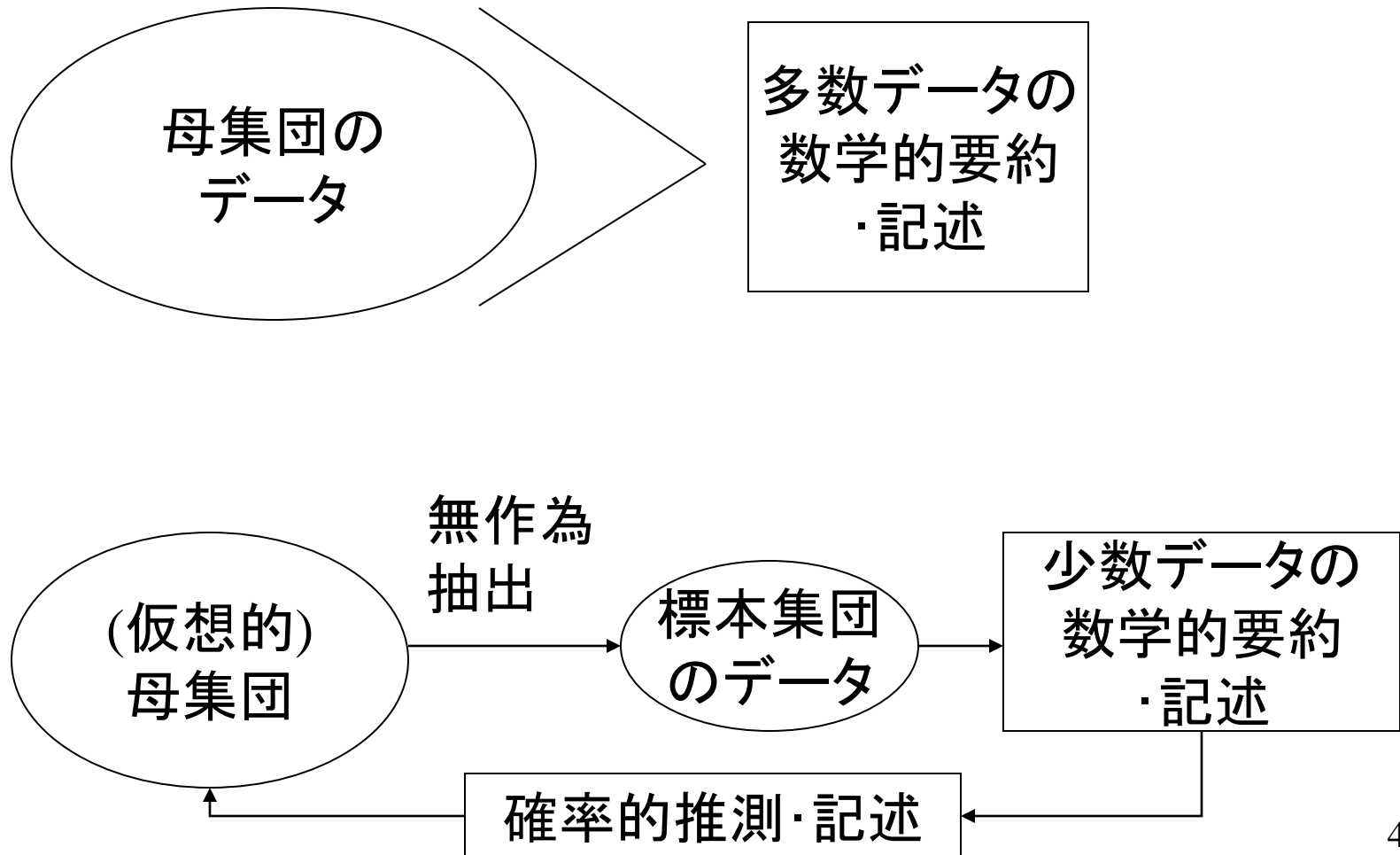
# 統計学の目的

- 沢山のデータを要約し、中に含まれている情報を把握しやすくするための手段
- 例: 学生100人の体重のデータがある.  
その100個の数値持っている情報を簡単に表わしたい

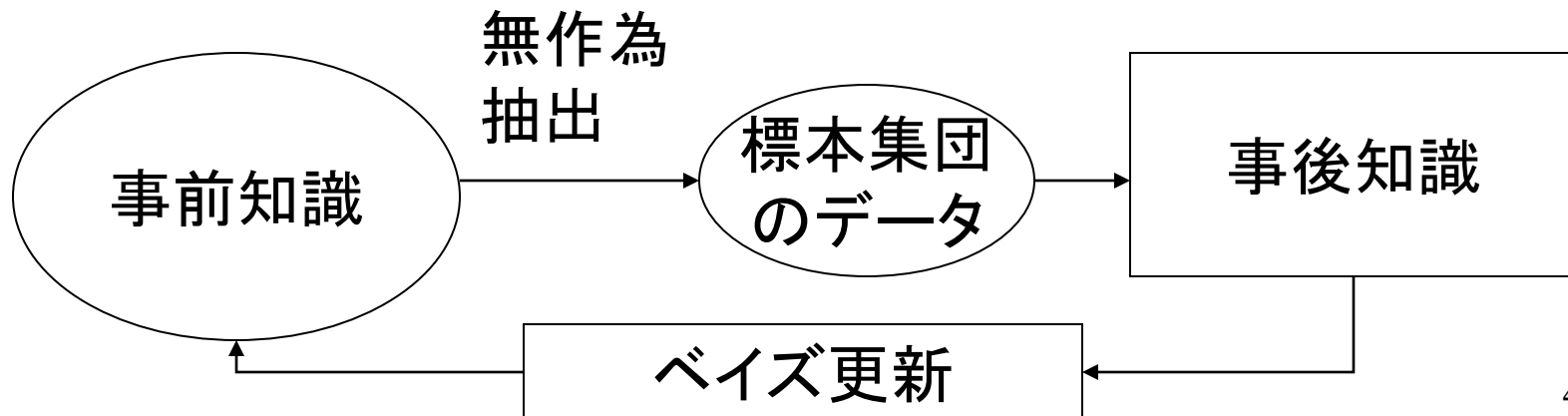
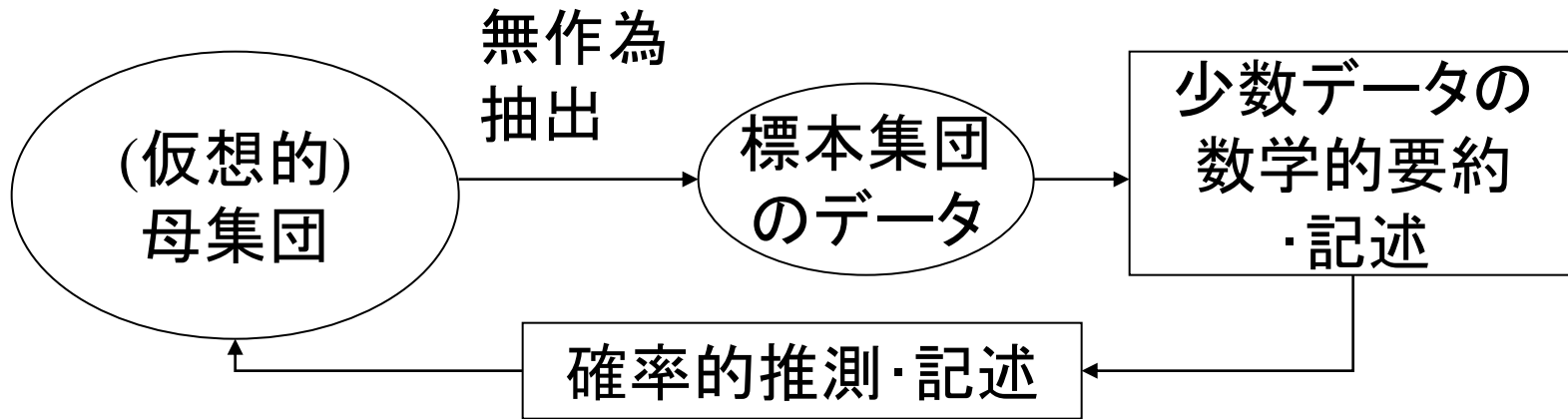


平均値: 「100人の学生の体重はだいたい60kgぐらいである」 41  
+ 標準偏差: 「100人の日本人の体重はだいたい50~70kgである」

# 記述統計学と推測統計学



# 推測統計学とベイズ統計学



# 尤度 (p12)

- ある確率分布でパラメータの値 $\theta$ が決まれば, データ $X$ の値 $x$ についてその値が得られる確率(確率密度)が計算できる.
  - $f(x|\theta)$
  - R上では d確率分布名( $x, \theta$ )の形

#一様分布 (unif)の例

#確率密度関数のグラフ

```
curve(dunif(x,min=0,max=2),xlim=c(-0.5,3),ylim=c(0,1),xlab="y",ylab="probability density")
```

#ある値に対する確率密度の値はdunif関数

```
dunif(0.2, min=0,max=2.0)
```

#分布関数, 累積分布関数: 変数がある値以下を取る確率: punif関数

```
curve(punif(x,min=0,max=2),xlim=c(-0.5,3),ylim=c(0,1),xlab="y",ylab="probability")
```

#分位数(quantile) その値以下を取る確率が $p$ であるような点の値, 分布関数の逆関数

```
qunif(0.75,min=0,max=2.0)
```

#乱数の発生: runif関数, 乱数の個数とパラメータを与える

```
runif(3,min=0, max=2.0)
```

# 尤度 (p12)

- ある確率分布でパラメータの値 $\theta$ が決まれば, データ $X$ の値 $x$ についてその値が得られる確率(確率密度)が計算できる.
  - $f(x|\theta)$
  - $R$ 上では  $d$ 確率分布名( $x, \theta$ )の形
- 逆に, データ $X=x$ が与えられたとき, パラメータの値 $\theta$ に対して, その値 $x$ が得られる確率を尤度: ゆうど(likelihood)という.

# 二項分布の例と尤度関数

- つぼのなかに赤球 $r$ 個，白球 $w$ 個あり，1つ取り出して色を記録して戻すことを $n$ 回繰り返す。
- 赤が出る回数 $Y$ が $y$ を取る確率は，一つの母数 $\phi = r/(r+w)$ を用いると，

$$P(Y = y | f) = {}_n C_y f^y (1 - f)^{n-y} \text{ となる.}$$

- 実際に赤が8回，白が2回でた場合には，そのことが起こる確率は， ${}_{10} C_8 f^8 (1 - f)^2$ で，これを母数 $\phi$ の関数と見なしたものを尤度関数 $L(\phi)$ と呼ぶ。

# 二項分布の例と尤度関数

#二項分布の関数形: Rではdbinom

```
barplot(dbinom(0:10,size=10,prob=0.6),ylab="probability",space=0, names=as.character(0:10), col="white")
```

#赤が8回, 白が2回でた場合の尤度関数 $L(\phi)$

```
Lik <- function(phi) {dbinom(8,size=10,phi)}
```

```
curve(Lik(x), 0, 1)
```

#尤度関数の対数値を対数尤度関数(LogLikelihood)

```
LLik <- function(phi) {log(dbinom(8,size=10,phi))}
```

```
curve(LLik(x), 0.05, 0.95)
```

# 尤度の最大化(最尤推定)

- データがあり, 確率分布の種類は決まっているが, パラメータ(母数)値がわからないとき。
- 得られているデータがもたらされる確率(尤度)が高いパラメータ値だったと考えるのが自然.
- 尤度が最大になるパラメータ値を推定値として使う.
- 赤が8回, 白が2回でた場合の尤度関数

これを母数 $\phi$ で微分すると,

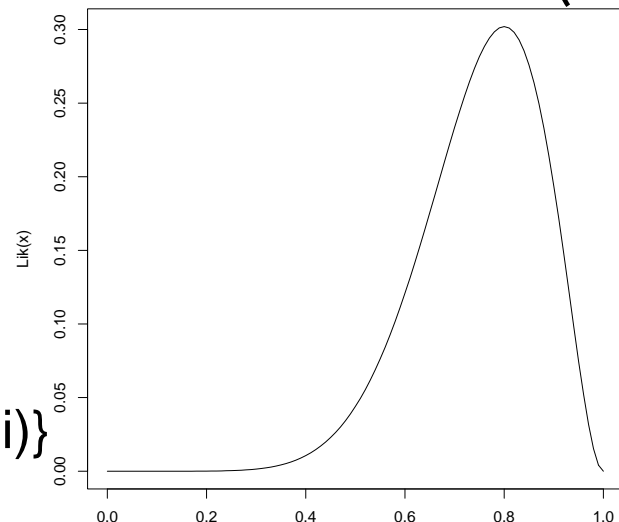
$${}_{10}C_8 f^7 (1-f) \{8(1-f) - 2f\}$$

$$= {}_{10}C_8 f^7 (1-f) (8 - 10f)$$

最大値は $\phi=8/10=0.8$ で取る.

```
Lik <- function(phi) {dbinom(8,size=10,phi)}  
optimize(Lik,c(0,1),maximum=TRUE)
```

$${}_{10}C_8 f^8 (1-f)^2$$





# 対数尤度の最大化(最尤推定)

- 赤が8回, 白が2回でた場合の尤度関数,  ${}_{10}C_8 f^8 (1-f)^2$   
対数尤度関数は,  $\log({}_{10}C_8) + 8\log f + 2\log(1-f)$

これを母数 $\phi$ で微分すると,

$$\frac{8}{f} - \frac{2}{(1-f)} = \frac{8(1-f) - 2f}{f(1-f)} = \frac{8-10f}{f(1-f)}$$

最大値は最後の分子が0になる,

$\phi=8/10=0.8$ で取る.

```
LLik <- function(phi) {log(dbinom(8,size=10,phi))}  
optimize(LLik,c(0.01,0.99),maximum=TRUE)
```

