【統計解析入門②】

獨協医科大学
先端医科学統合研究施設　研究連携・支援センター
西連地　利己

多変量解析

そこで多変量解析

- （完璧ではないが）交絡因子を調整可能
- 既存データを有効活用

多変量解析の例

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2$$

y：平均値の推定値
x_1：A群かB群か；　x_2：年齢
β_1：年齢 x_2 を固定した時の
　　A群とB群の y の違いの大きさ

※β_1とβ_2は、最尤法や最小二乗法などで決定する
　　（統計ソフトが勝手にやってくれる）

多変量解析の一般型

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

$$y = \alpha + \sum_{i=1}^{n} \beta_i x_i$$

従属変数の分布の型で解析手法が変わる

従属変数の型と多変量解析手法

従属変数の型	解析手法
正規分布する 連続量	重回帰分析、分散分析、共分散分析
2値 （0か1） （即ち割合）	ロジスティック回帰分析
稀に起こる回数 （0, 1, 2, ・・・）	ポワソン回帰分析
生存時間 （打ち切りあり）	Cox比例ハザードモデル
ポワソン・ガンマ 混合分布	Tweedie回帰分析

従属変数が正規分布する場合で

- 独立変数が・・・
  - 全て連続量（血圧値など）
    - 重回帰分析
  - 全てカテゴリ（男女や治療方法など）
    - 分散分析
  - 連続量とカテゴリの両方
    - 共分散分析

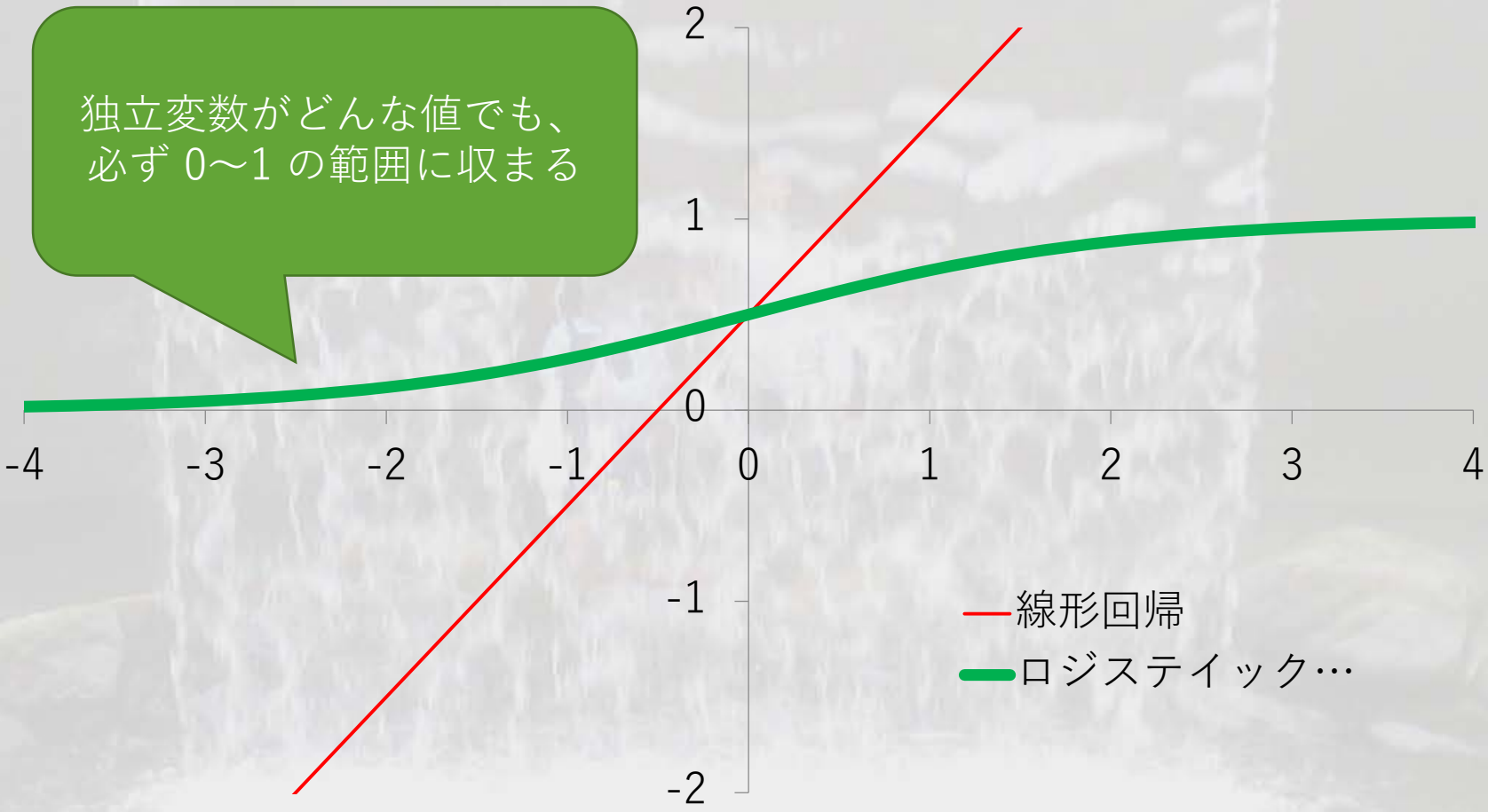
ロジスティック回帰

$$y = ln\left(\frac{p}{1-p}\right) = \alpha + \sum_{i=1}^{n} \beta_i x_i$$

$$p = \frac{1}{1 + e^{-y}}$$

変数 x_i のオッズ比は、$e^{\beta_i} = \exp(\beta_i)$

線形回帰とロジスティック回帰



ちょっとロジスティック回帰分析

y	x1	x2
0	-10	0.351728
0	-9	0.215755
0	-8	0.921308
0	-7	0.230966
0	-6	0.407013
0	-5	0.612901
0	-4	0.451967
1	-3	0.345406
0	-2	0.54416
0	-1	0.715334
0	0	0.242574
1	1	0.608487
1	2	0.473443
1	3	0.086338
1	4	0.915638
0	5	0.765722
1	6	0.37893
1	7	0.76336
1	8	0.88345
1	9	0.928289
1	10	0.521604

解析ソフトSASでのプログラム

```
proc logistic;
    model
        y(event='1') = x1 x2
         / rl
    ;
run;
quit;
```

解析ソフトによる結果表示の例

最尤推定値の分析

パラメータ	自由度	推定値	標準誤差	Wald カイ 2 乗	Pr > ChiSq
Intercept	1			0.3964	0.5290
x1	1	0.5080	0.2122	5.7311	0.0167
x2	1	−2.8711	3.5392	0.6581	0.4172

予測確率と観測データの応		
一致の割合	94.5	Som
不一致の割合	5.5	ガン
タイの割合	0.0	Tau
組	110	c

$\exp(0.5080) \fallingdotseq 1.66$
$\exp(0.5080 - 1.96 \times 0.2122) \fallingdotseq 1.10$
$\exp(0.5080 + 1.96 \times 0.2122) \fallingdotseq 2.52$

オッズ比推定と Wald による信頼区間

効果	単位	推定値	95% 信頼限界	
x1	1.0000	1.662	1.096	2.519
x2	1.0000	0.057	<0.001	58.300

Cox比例ハザードモデル

$$y = ln\left(\frac{h(t|x)}{h_0(t)}\right) = \sum_{i=1}^{n} \beta_i x_i$$

変数 x_i のハザード比は、$e^{\beta_i} = \exp(\beta_i)$

発生率（人年法）

$$\frac{\text{○人}}{\text{41人年}}$$

中途打ち切り例

- 観察期間終了時にエンドポイントに達していない症例

- 観察期間中に追跡が不能になった症例
  - 転居等により消息が不明になった
  - 設定したエンドポイント以外で死亡した
  - etc.

ちょっと比例ハザードモデル

year	y	x1	x2
10	0	-10	0.351728
10	0	-9	0.215755
10	0	-8	0.921308
10	0	-7	0.230966
10	0	-6	0.407013
10	0	-5	0.612901
10	0	-4	0.451967
9	1	-3	0.345406
10	0	-2	0.54416
10	0	-1	0.715334
10	0	0	0.242574
9	1	1	0.608487
8	1	2	0.473443
7	1	3	0.086338
6	1	4	0.915638
10	0	5	0.765722
5	1	6	0.37893
4	1	7	0.76336
3	1	8	0.88345
2	1	9	0.928289
1	1	10	0.521604

解析ソフトSASでのプログラム

```
proc phreg;
    model
        years*y(0) = x1 x2
         /rl
    ;
run;
quit;
```

解析ソフトによる結果表示の例

モデルの適合度統計量		
基準	共変量なし	共変量あり
-2 LOG L	55.916	31.435
AIC	55.916	35.435
SBC	55.916	36.040

包括帰無仮説：BETA=0 の検定			
検定	カイ 2 乗値	自由度	Pr > ChiSq
尤度比	24.4805	2	<.0001
Score	19.0910	2	<.0001
Wald	11.1552	2	0.0038

最尤推定値の分析								
パラメータ	自由度	パラメータ推定値	標準誤差	カイ 2	Pr > ChiSq	ハザード比	95% ハザード比信頼限界	
x1	1	0.51798	0.15776	10.7	0.0010	1.679	1.232	2.287
x2	1	-1.52938	1.49323	1.0490	0.30	0.217	0.01	4.044

コホート研究での例（表１）

(Sasai H, Mayo Clin Proc. 2010 Jan;85:36-40.)

TABLE 1. **Sex-Stratified Baseline Characteristics According to BM**

Baseline characteristic	Men (N=19,926)			
	BMI (kg/m²)			
	<25.0	25.0-29.9	≥30.0	*P* value
No. of participants	14,474	5163	289	
Age (y)	61.5 (9.7)	59.5 (9.6)	58.0 (10.1)	<.001
Fasting blood glucose (mg/dL)	99.1 (10.8)	100.9 (10.8)	99.1 (12.6)	<.001
Nonfasting blood glucose (mg/dL)	111.7 (25.2)	113.5 (25.2)	115.3 (25.2)	.12
Fasting participants (%)	15.1	14.5	14.2	.57
Systolic blood pressure (mm Hg)	135.0 (17.1)	138.9 (16.8)	142.6 (17.8)	<.001
Diastolic blood pressure (mm Hg)	79.7 (10.4)	83.3 (10.4)	86.6 (12.0)	< 001
Antihypertensive medication use (%)	18.6	26.4	33.9	<.001
Total cholesterol (mg/dL)	189.5 (30.9)	201.1 (30.9)	204.9 (34.8)	<.001
HDL cholesterol (mg/dL)	54.1 (15.5)	46.4 (11.6)	42.5 (11.6)	<.001
Triglyceride (mg/dL)	132.9 (79.7)	186.0 (106.3)	212.6 (124.0)	<.001
Lipid medication use (%)	1.1	1.6	3.1	.001
Smoking status (%)				<.001
Never	22.1	26.1	28.4	
Ex-smoker	26.6	32.8	27.3	
Current				
<20 cigarettes/d	17.0	11.0	10.7	
≥20 cigarettes/d	34.3	30.1	33.6	
Alcohol intake (%)				<.001
None	34.9	33.5	42.6	
Occasionally	13.0	16.2	18.3	
Daily				
<60 g/d	46.4	43.7	31.8	
≥60 g/d	5.7	6.6	7.3	

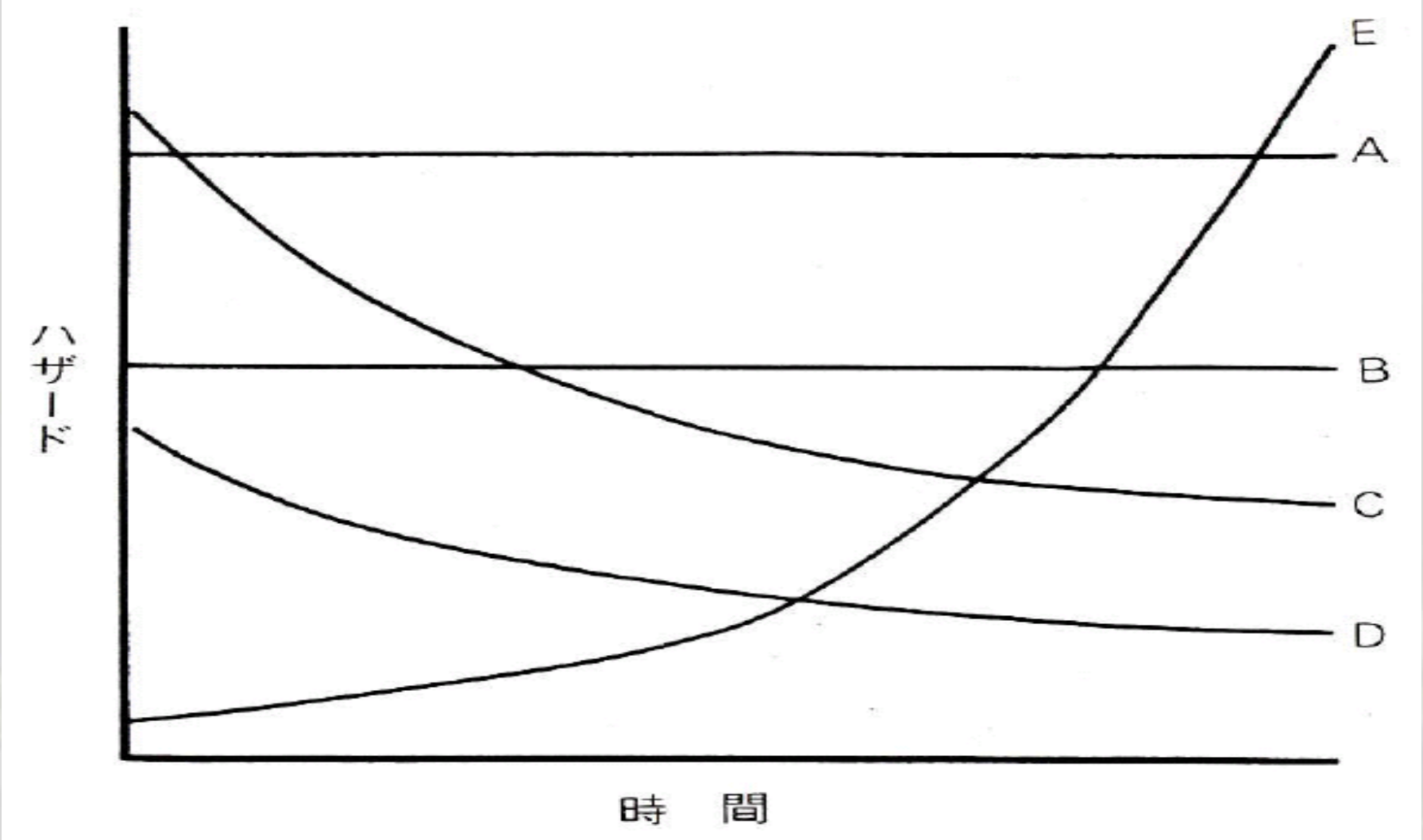
コホート研究での例（表2）

(Sasai H, Mayo Clin Proc. 2010 Jan;85:36-40.)

TABLE 2. **Hazard Ratios (HRs) for Type 2 Diabetes Mellitus According to Body Mass Index Stratified by Sex and Age**

Variable	No. of participants	Person-years	Incidence rates per 1000 person-years	Age-adjusted HRs (95% CI[a])	Multivariate HRs[b] (95% CI[a])
Men (N=19,926)					
Age 40-59 (y)					
BMI (kg/m^2)					
<25.0	4914	26,569	14.9	1.00 (Reference)	1.00 (Reference)
25.0-29.9	2268	11,791	25.3	1.68 (1.44-1.95)	1.42 (1.21-1.67)
≥30.0	147	718	29.2	1.96 (1.26-3.03)	1.40 (0.89-2.20)
Age 60-79 (y)					
BMI (kg/m^2)					
<25.0	9560	47,223	20.5	1.00 (Reference)	1.00 (Reference)
25.0-29.9	2895	14,029	25.7	1.24 (1.10-1.40)	1.13 (0.99-1.29)
≥30.0	142	619	33.9	1.62 (1.05-2.49)	1.26 (0.81-1.96)

proportionality



（浜島信行，多変量解析による臨床研究，名古屋大学出版会，名古屋市，1998）

回帰係数の意味合い

- $\exp(\beta_i)$ は、x_i が 1 増加するあたりの、
  オッズ比またはハザード比

- A群（$x_i = 0$）、B群（$x_i = 1$）なら、
  A群を基準としたときのB群の
  オッズ比またはハザード比

A群とB群とC群の場合

- A群を基準としたときの、B群のオッズ比等
- A群を基準としたときの、C群のオッズ比等

    - ダミー変数を使う

ダミー変数の作成

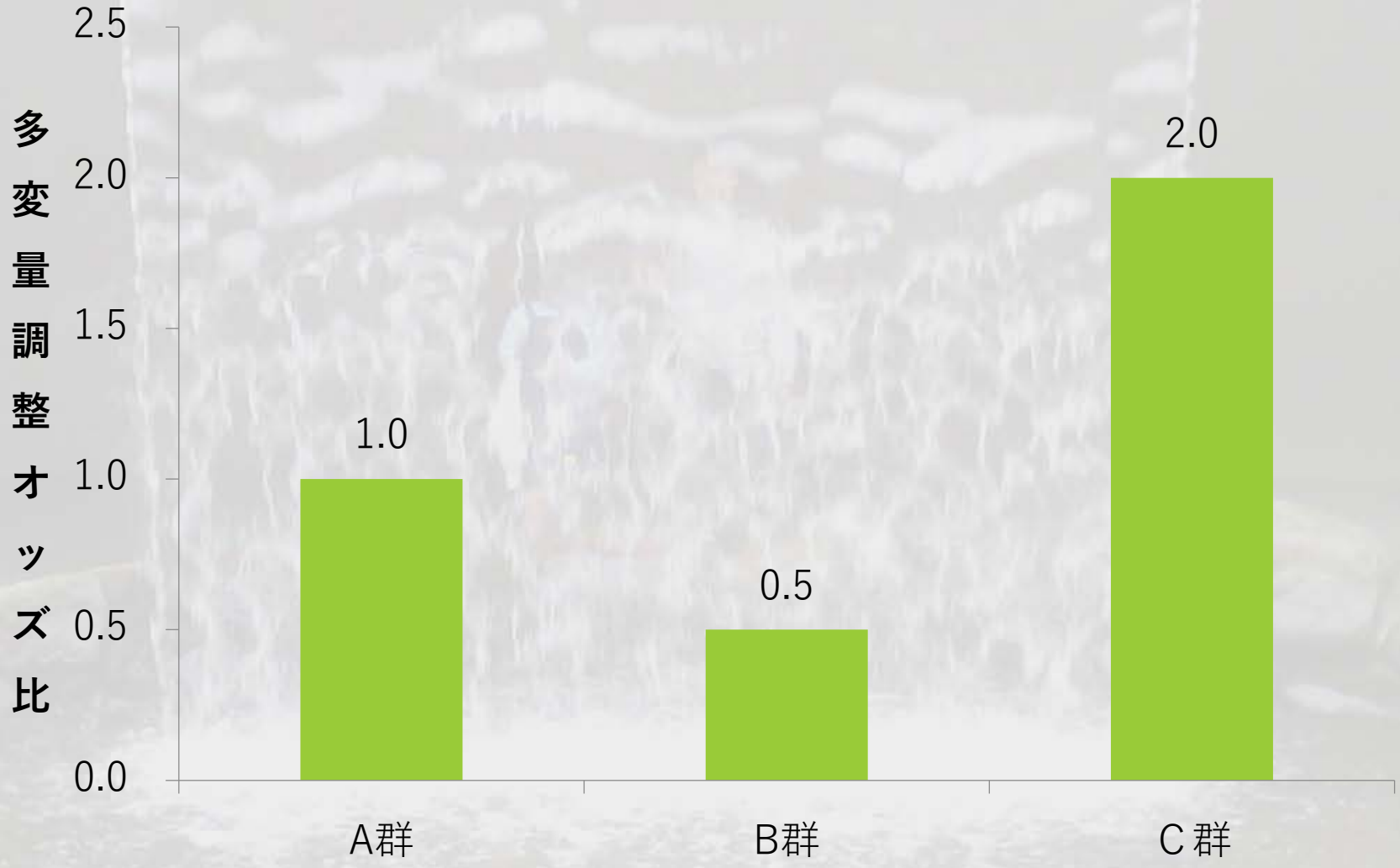
群	D1	D2
A群	0	0
B群	1	0
C群	0	1

$$ln\left(\frac{p}{1-p}\right)$$
$$= \alpha + \beta_{D1}D_1 + \beta_{D2}D_2 + \beta_2 x_2 + \cdots$$

$exp(\beta_{D1})$：B群のオッズ比

$exp(\beta_{D2})$：C群のオッズ比

３群のオッズ比が描ける



まとめ①

- 多変量解析は、
  「他の変数に差が無かったら、
  　この変数はアウトカムに影響しているか？」
  を検討している。
  - RCTでないのに、
    「RCTだったらばこういう結果」という推定をする。

- ただし、変数を多く投入すればするほど、
  大きなサンプルサイズが必要になってくる。

まとめ②

従属変数の型	解析手法
正規分布する 連続量	重回帰分析、分散分析、共分散分析
2値　（0か1） （即ち割合）	ロジスティック回帰分析
稀に起こる回数 （0, 1, 2, ・・・）	ポワソン回帰分析
生存時間 （打ち切りあり）	Cox比例ハザードモデル
ポワソン・ガンマ 混合分布	Tweedie回帰分析