県立広島大学・経営情報学部

経営情報学科

オープンキャンパス2016 模擬講義資料

Prefectural University of Hiroshima Faculty of Management and Information Systems
Department of Management Information Systems

時間や位置によって変わる関係のデータ分析

応用統計学研究室
冨田 哲治

自己紹介

- 名前：　冨田 哲治（とんだ てつじ）
- 専門：　応用統計学
  （統計学を用いた実社会データ分析）
- 担当科目：　統計学，基礎数学，
  ビジネス数理入門など

はじめに

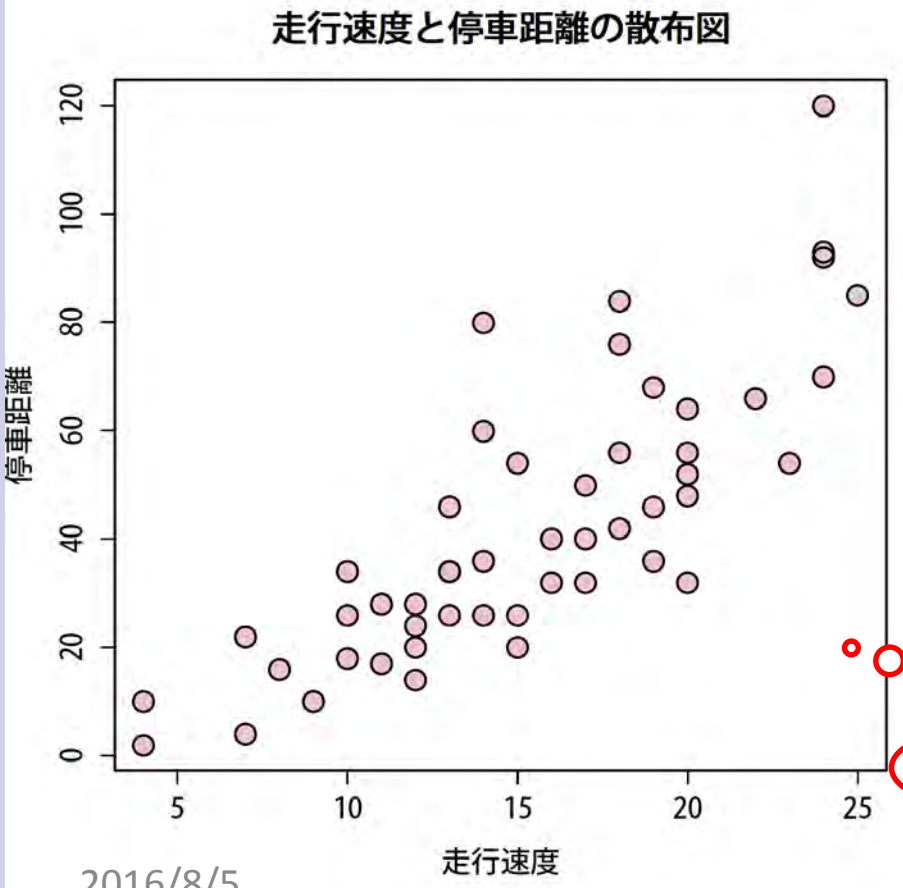
- **2つの変数の間の関係についての分析**
  - 最高気温　と　アイスの売上
  - 走行速度　と　停車距離

車ID	1	2	3	…	50
走行速度	4	4	7	…	25
停車距離	2	10	4	…	85

- **数学Ⅰ「データの分析」で学んだ“散布図”や“相関係数”は関係の分析法の一つ**

関係を可視化・数値化

車ID	1	2	3	…	50
走行速度	4	4	7	…	25
停車距離	2	10	4	…	85



2つの変数の関係は
強い？　弱い？

関係の強さの程度は、見る人に主観によって違ってしまう

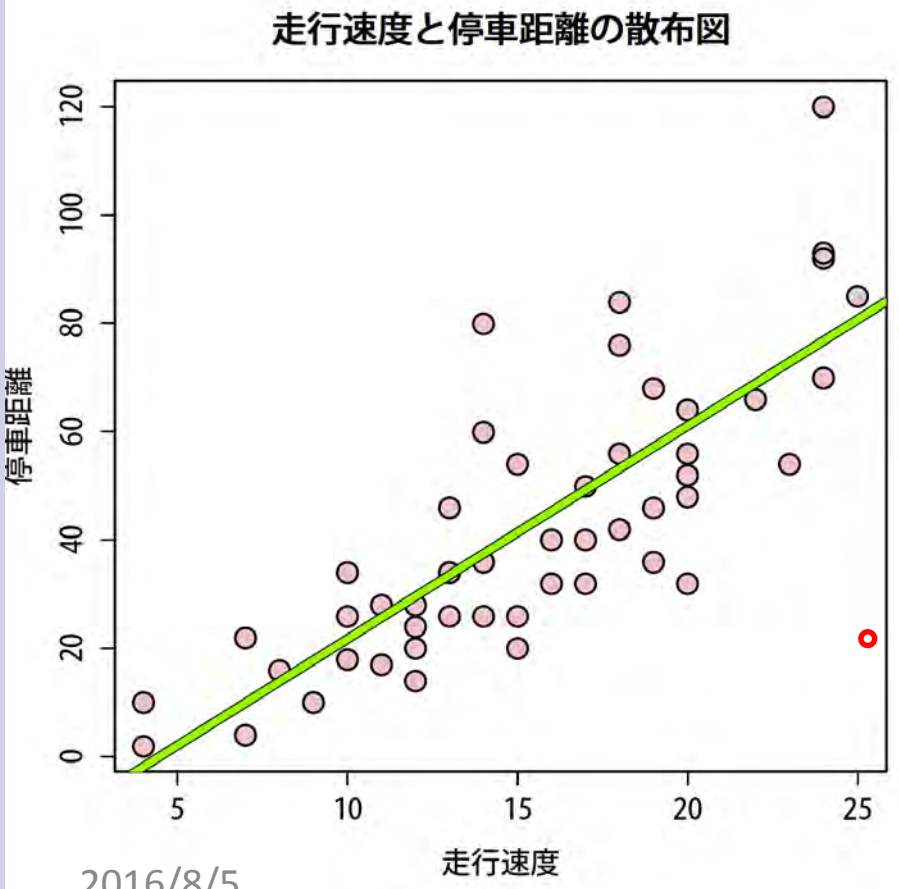
相関係数：$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$

関係の強さを客観的な数値$(-1 \leq r \leq 1)$で要約

相関係数= 0.81

関係を式で要約・・・回帰分析

車ID	1	2	3	…	50
走行速度	4	4	7	…	25
停車距離	2	10	4	…	85



2つの変数はどんな関係？

関係を式で要約することで、関係を定量的につかむ

単回帰：　距離 $= a + b \times$ 速度

関係を直線式（切片・傾き）で要約する方法を単回帰とよぶ

距離 $= -18 + 4 \times$ 速度

補足）重回帰分析

- **単回帰**では“距離”を1つの変数(“速度”)の式で要約した

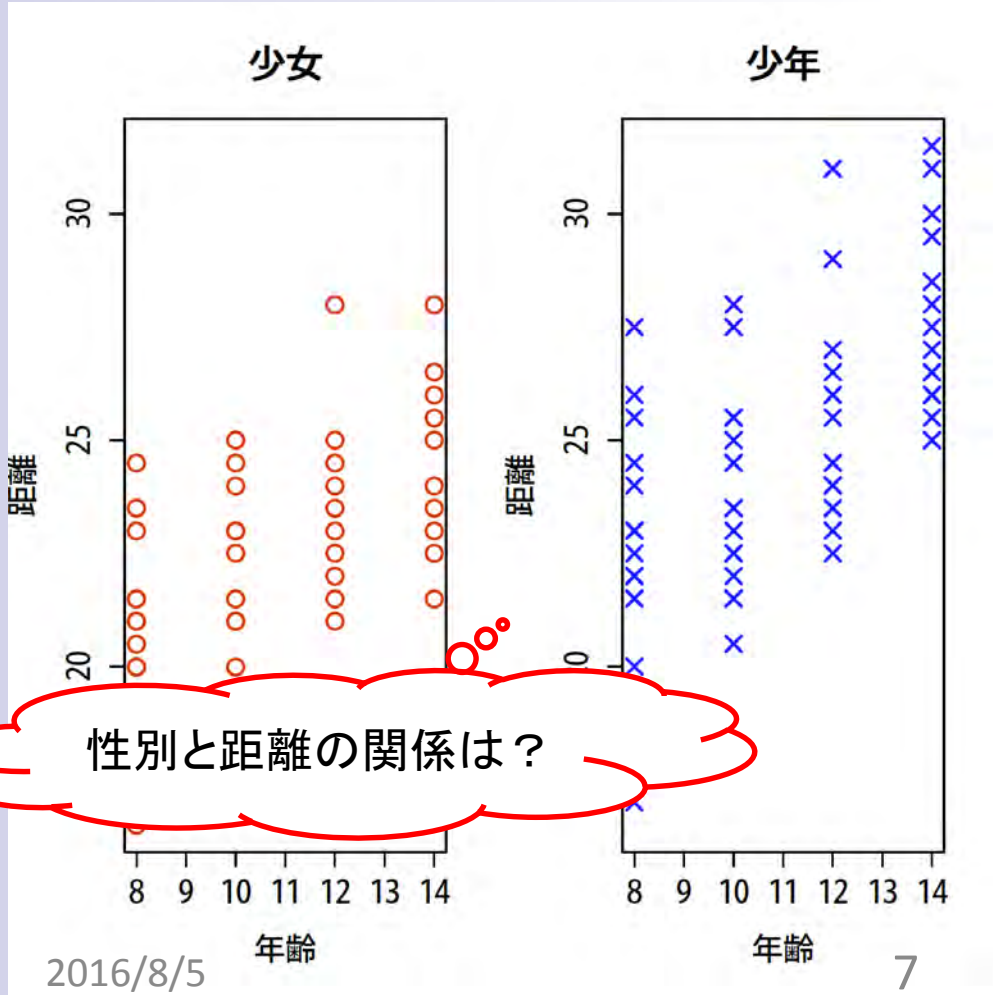
$$\text{距離} = a + b \times \text{速度}$$

- 複数の変数を含む式で要約する分析法は、**重回帰分析**とよばれる

$$\text{距離} = a + b \times \text{速度} + c \times \text{天気} + d \times \text{路面} + \cdots$$

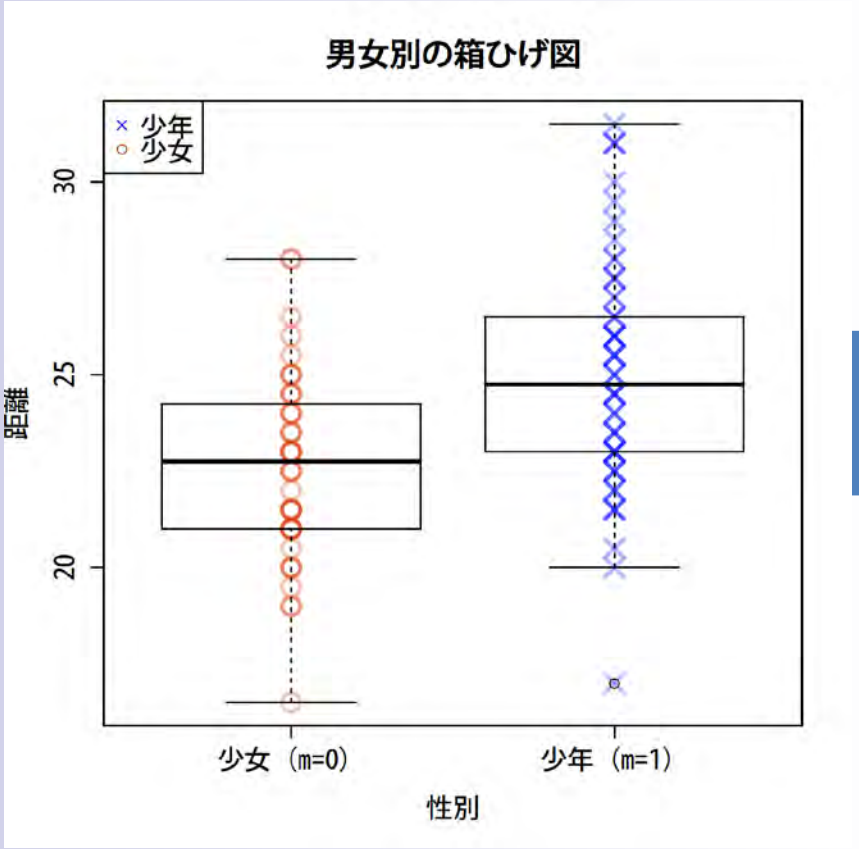
例）子供の成長データ

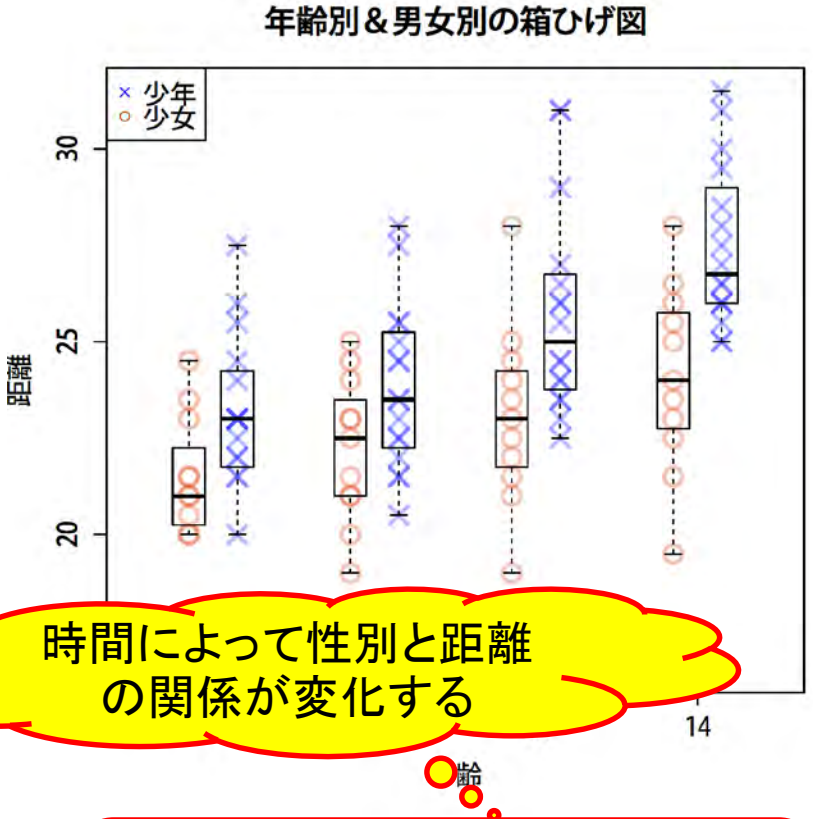
- 少年と少女の成長を8,10,12,14歳の4時点で測定したデータ



性別	ID	基準点（脳下垂体と翼上顎裂）の距離[mm]			
		8歳	10歳	12歳	14歳
少年	M01	26.0	25.0	29.0	31.0
	M02	21.5	22.5	23.0	26.5
	M03	23.0	22.5	24.0	27.5
	M04	25.5	27.5	26.5	27.0
	M05	20.0	23.5	22.5	26.0
	M06	24.5	25.5	27.0	28.5
	M07	22.0	22.0	24.5	26.5
	M08	24.0	21.5	24.5	25.5
	M09	23.0	20.5	31.0	26.0
	M10	27.5	28.0	31.0	31.5
	M11	23.0	23.0	23.5	25.0
	M12	21.5	23.5	24.0	28.0
	M13	17.0	24.5	26.0	29.5
	M14	22.5	25.5	25.5	26.0
	M15	23.0	24.5	26.0	30.0
	M16	22.0	21.5	23.5	25.0
少女	F01	21.0	20.0	21.5	23.0
	F02	21.0	21.5	24.0	25.5
	F03	20.5	24.0	24.5	26.0
	F04	23.5	24.5	25.0	26.5
	F05	21.5	23.0	22.5	23.5
	F06	20.0	21.0	21.0	22.5
	F07	21.5	22.5	23.0	25.0
	F08	23.0	23.0	23.5	24.0
	F09	20.0	21.0	22.0	21.5
	F10	16.5	19.0	19.0	19.5
	F11	24.5	25.0	28.0	28.0

時間によって変わる関係(1/2)





時間によって性別と距離の関係が変化する

$$\text{距離} = a + b \times m = \begin{cases} a & \text{少女} \\ a + b & \text{少年} \end{cases}$$

bは性差
(少年の効果)

$$\text{距離} = a(t) + b(t) \times m = \begin{cases} a(t) & \text{少女} \\ a(t) + b(t) & \text{少年} \end{cases}$$

時間によって変わる関係(2/2)

時間によらず一定の性差

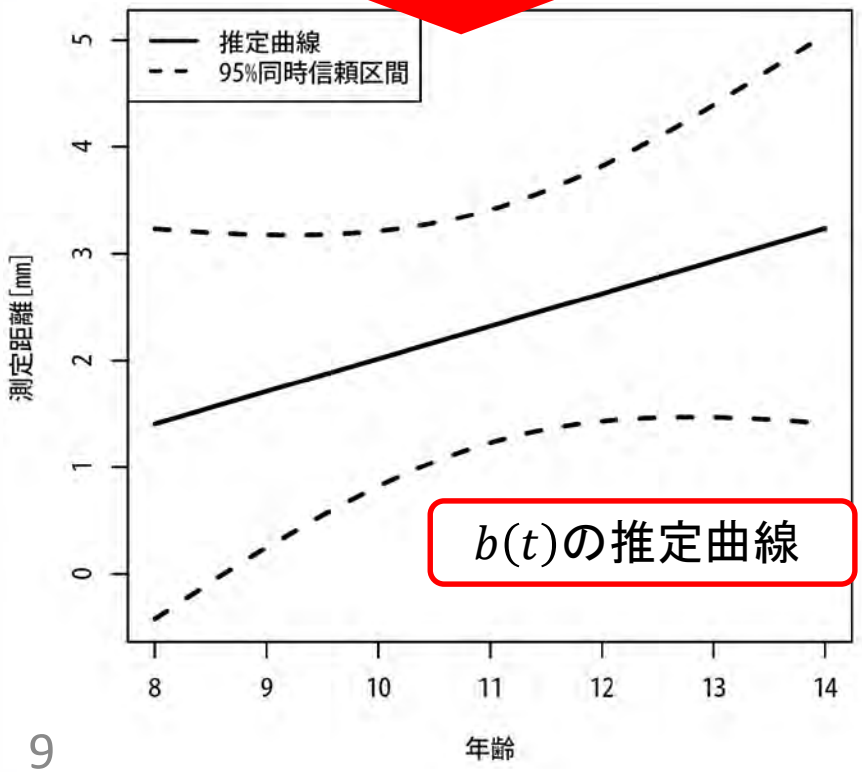
$$距離 = a + b \times m = \begin{cases} a & 少女 \\ a + b & 少年 \end{cases}$$

bの推定値＝2.32

どちらが適切かデータから判断することもできる

時間によって変わる性差

$$距離 = a(t) + b(t) \times m = \begin{cases} a(t) & 少女 \\ a(t) + b(t) & 少年 \end{cases}$$



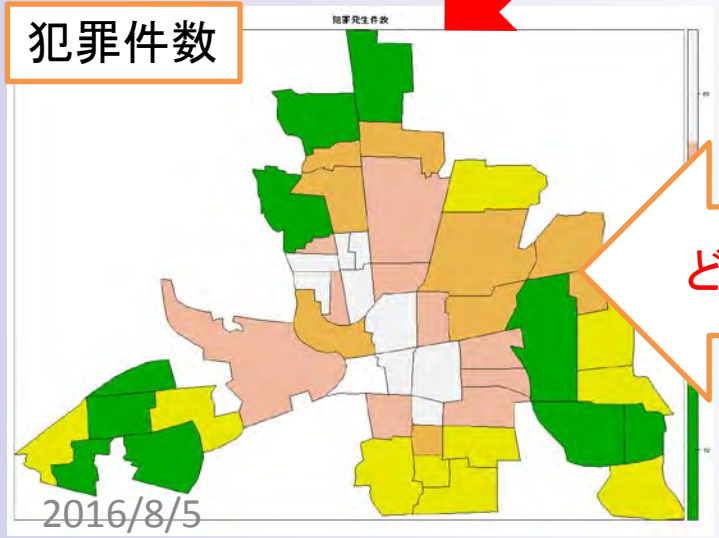
例）コロンバス市の犯罪データ

- 1980年のオハイオ州コロンバス市の49地域における、住居侵入窃盗・車両窃盗の犯罪件数と 平均収入のデータ

	crime	income	x	y
0	15.725980	19.531	8.827218	14.36908
1	18.801754	21.232	8.332658	14.03162
2	30.626781	15.956	9.012265	13.81972
3	32.387760	4.477	8.460801	13.71696
4	50.731510	11.252	9.007982	13.29637
5	26.066658	16.029	9.739926	13.47463

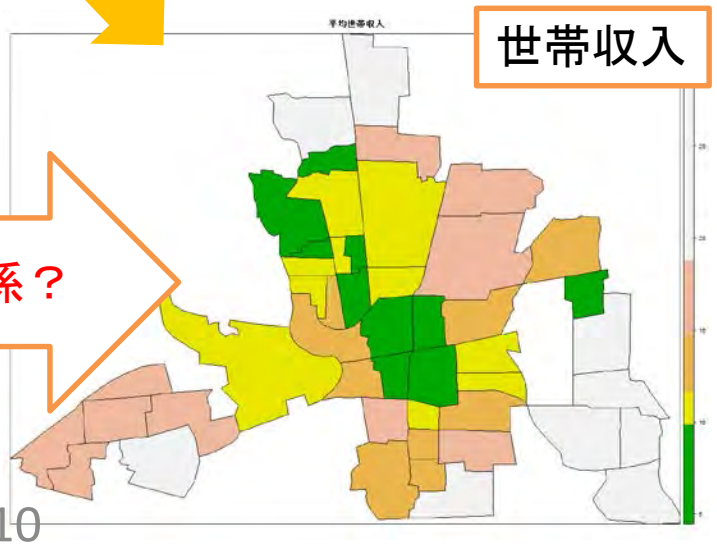
位置情報
(x, y)

犯罪件数



どんな関係？

世帯収入



単回帰・・・地域差を無視した分析

- 犯罪件数と収入の関係を次の直線式で要約（単回帰）

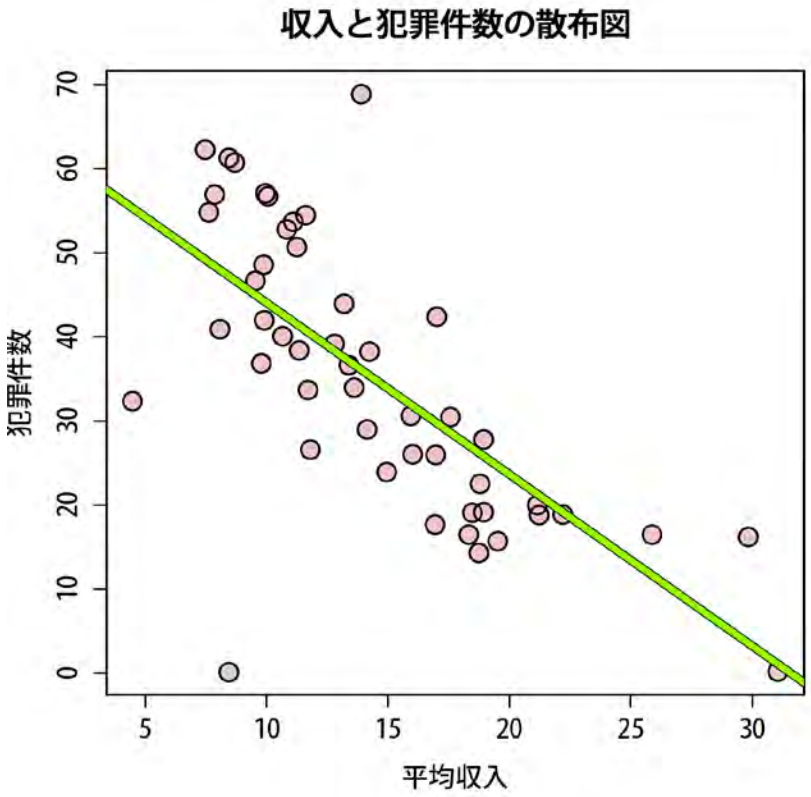
$$\text{犯罪件数} = a + b \times \text{収入}$$

- 犯罪件数と世帯収入の関係は地域によらず同じと仮定した分析

  → 地域差を無視

- もしかしたら、**関係に地域差**があるかもしれない…

  → 位置情報を活用



$$\text{犯罪件数} = 64.4 - 2.0 \times \text{収入}$$

位置によって変わる関係

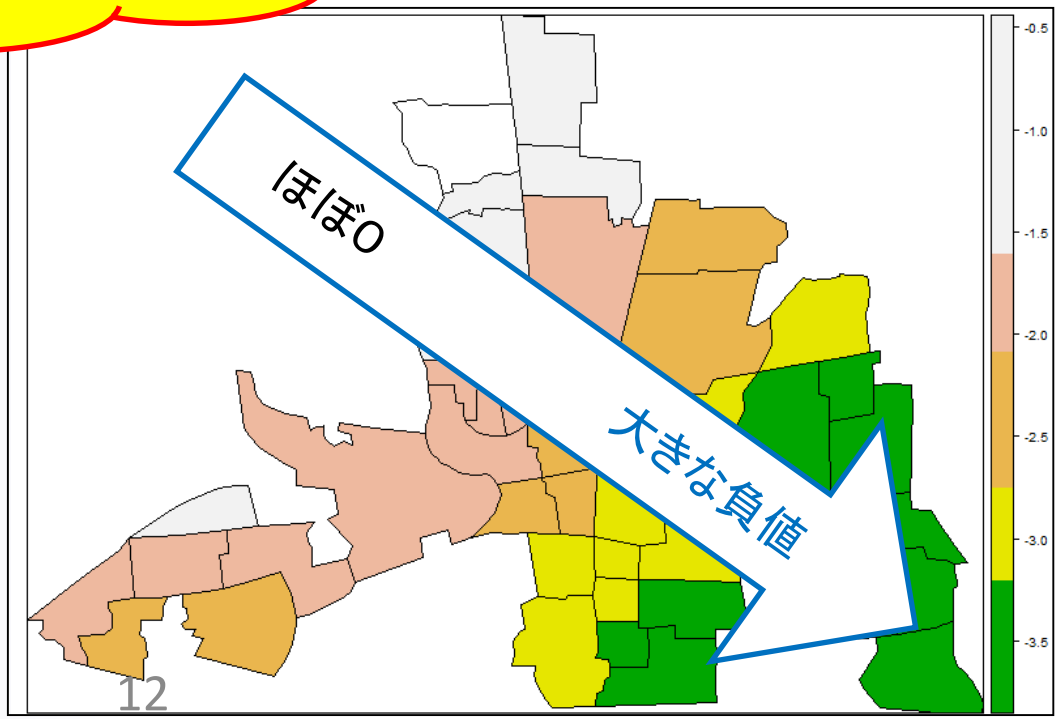
地域差を無視した場合：
犯罪件数 $= a + b \times$ 収入

- **犯罪件数と収入の関係を次式で要約**

$$\text{犯罪件数} = a(x,y) + b(x,y) \times \text{収入}$$

関係が位置(x,y)により異なる
→ 関係に地域差を考える

位置によって変わる収入と犯罪件数の関係$b(x,y)$を地図で可視化



- 北西地域では$b(x,y) \approx 0$
→ 収入と犯罪係数の関係はなさそう

- 南東地域では$b(x,y) < 0$
→ 収入と犯罪件数の関係が大きい

応用事例

広島原爆被爆者コホートデータ分析

- 位置(被爆時所在地)と 時間(年齢)によって変化する
- 放射線被ばくと 健康被害 の関係の分析

データ例

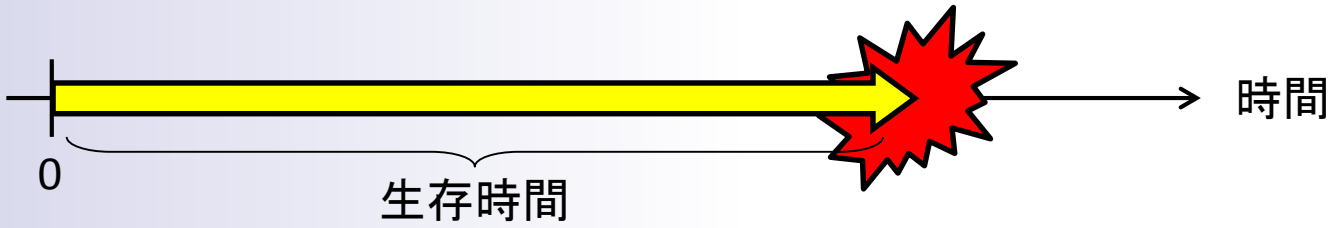
ID	年齢	被爆時年齢	性別	被ばく量	状態	x	y
1	60	10	1	0.5	0	1.0	2.0
2	75	15	1	0.01	1	1.5	2.5
3	45	5	0	0.2	0	0.1	1.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

時間(年齢)　　要因(被爆状況)　　位置(被爆時所在地)

生存時間分析

- 「基準となる時点からあるイベントが発生するまでの時間」を分析
- イベント

  死亡，罹患，故障，転出，出産，…

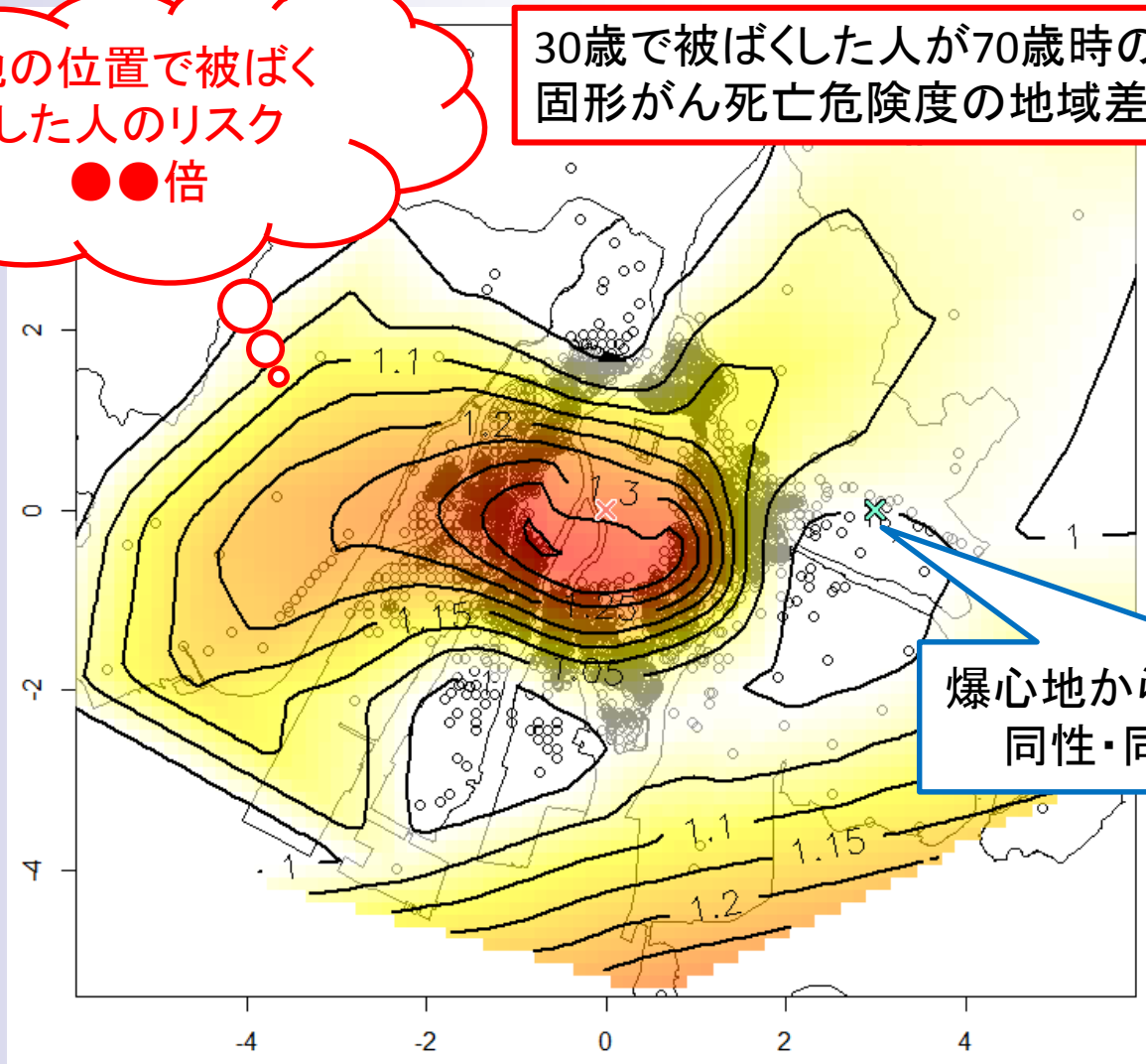


- 生存時間解析は、生存時間と様々な要因との関係を分析する方法

広島原爆被爆者のがん死亡危険度

他の位置で被ばくした人のリスク
●●倍

30歳で被ばくした人が70歳時の
固形がん死亡危険度の地域差地図



爆心地から東へ3kmで被ばくした
同性・同年齢の人に比べて

まとめ

- 時間や位置の情報を活用することで、時間や位置によって変わる関係の分析が可能となる。
- たくさんあるデータ分析法のどれを使うかは、分析目的とデータのもつ情報で決まる。
- データのもつ情報を無駄なく活用できる分析法を選ぶことで、精度の高い分析となる。