2009.09.19
「地域データ分析活用講座
～データをもとに地域と観光を考えてみよう～」
データ分析実習

データを用いて
地域を分析してみよう

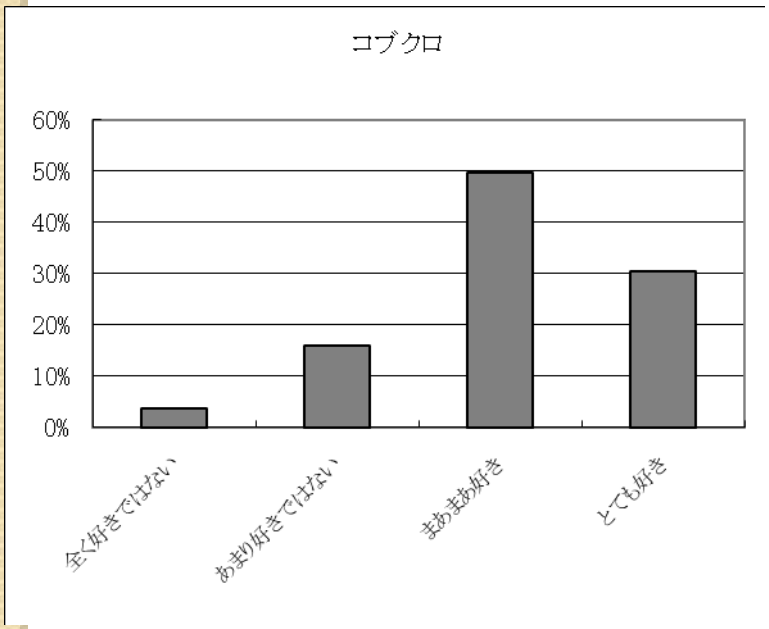
奈良県立大学　地域創造学部
遠 藤　英 樹

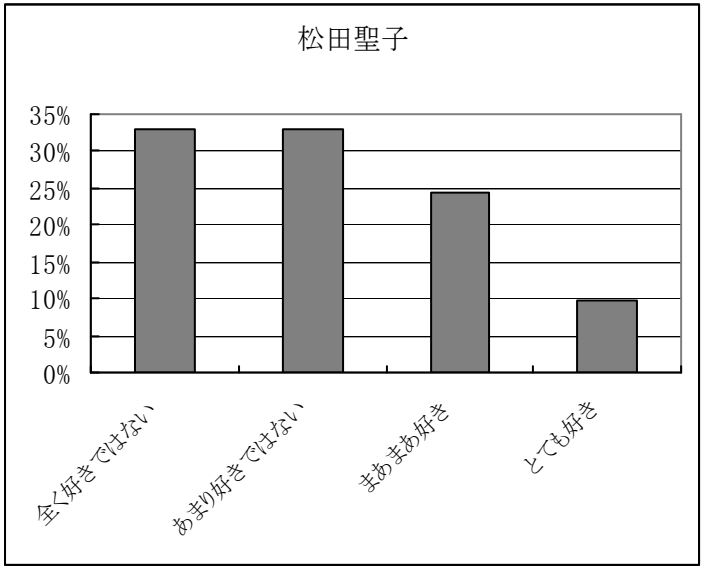
測定する「尺度」の次元

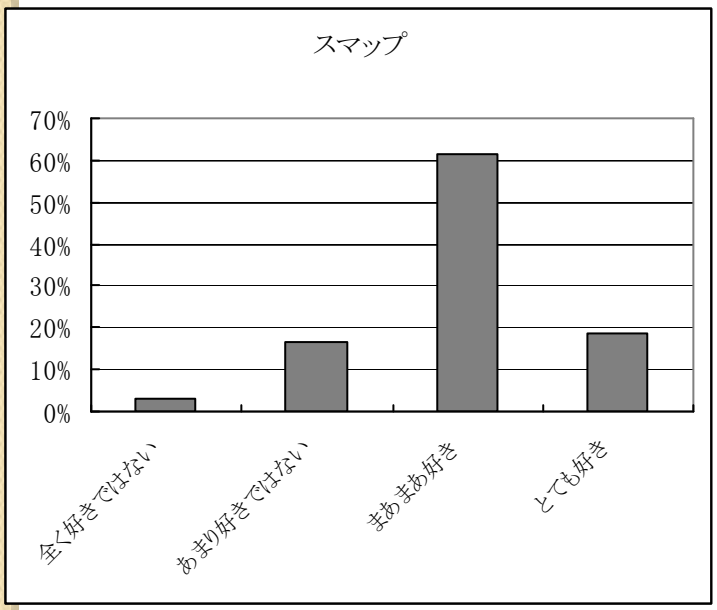
- １）名義尺度
- ２）順序尺度
- ３）間隔尺度
- ４）比例尺度

基礎統計

- 代表値とは何か？
- 散らばりとは何か？


コブクロ
60%
50%
40%
30%
20%
10%
0%
全く好きではない
あまり好きではない
まあまあ好き
とても好き

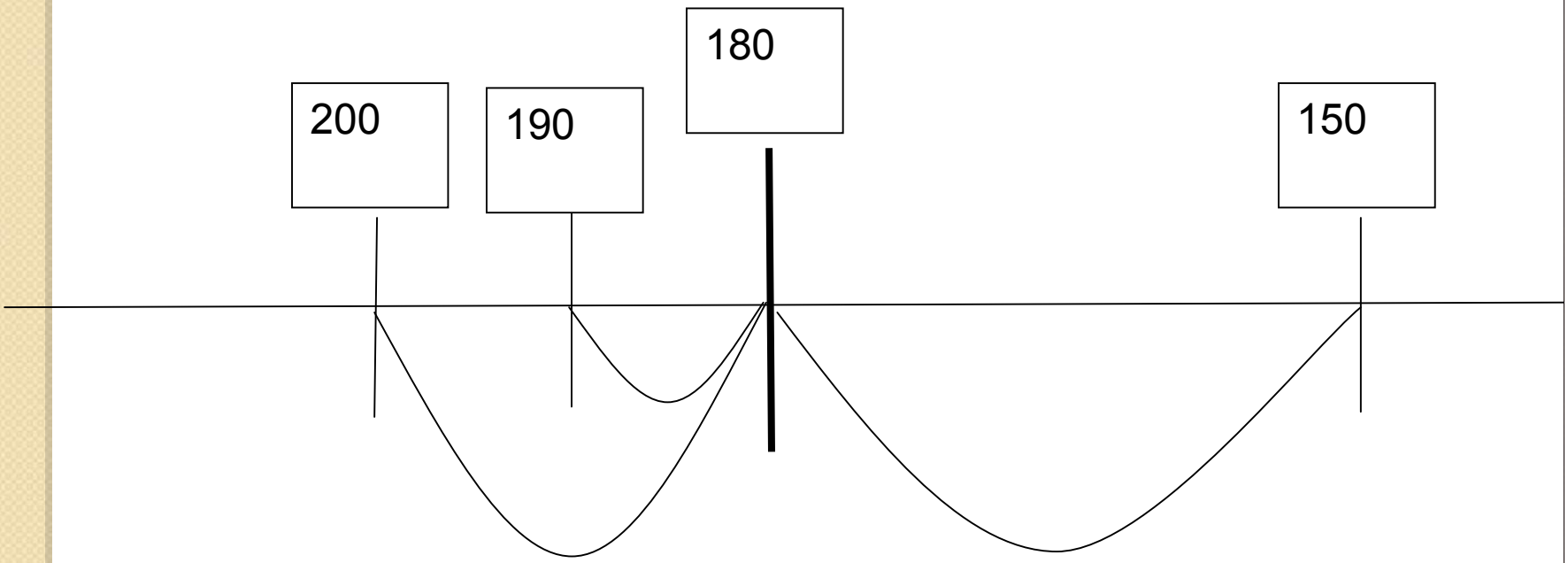

松田聖子
35%
30%
25%
20%
15%
10%
5%
0%
全く好きではない
あまり好きではない
まあまあ好き
とても好き


スマップ
70%
60%
50%
40%
30%
20%
10%
0%
全く好きではない
あまり好きではない
まあまあ好き
とても好き

代表値とは

- １）平均値
- ２）中央値
- ３）最頻値

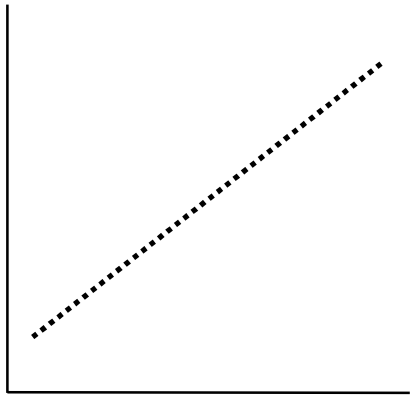
散らばりとは

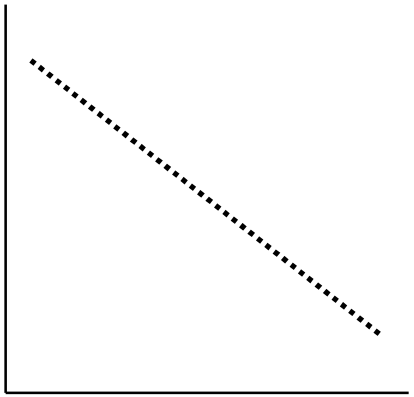


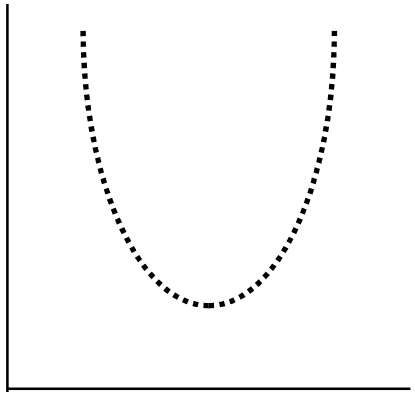
散らばりとは

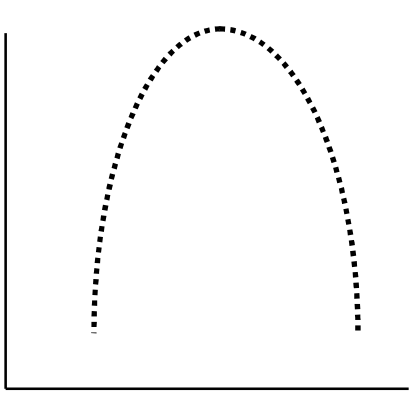
- １）偏差平方和
- ２）分散
  偏差平方和÷個（人）数
- ３）標準偏差
  √分散

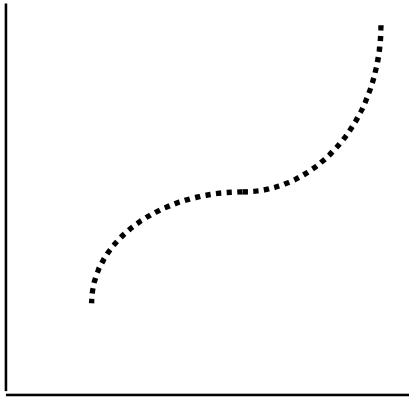
データの世界のいろいろな関係



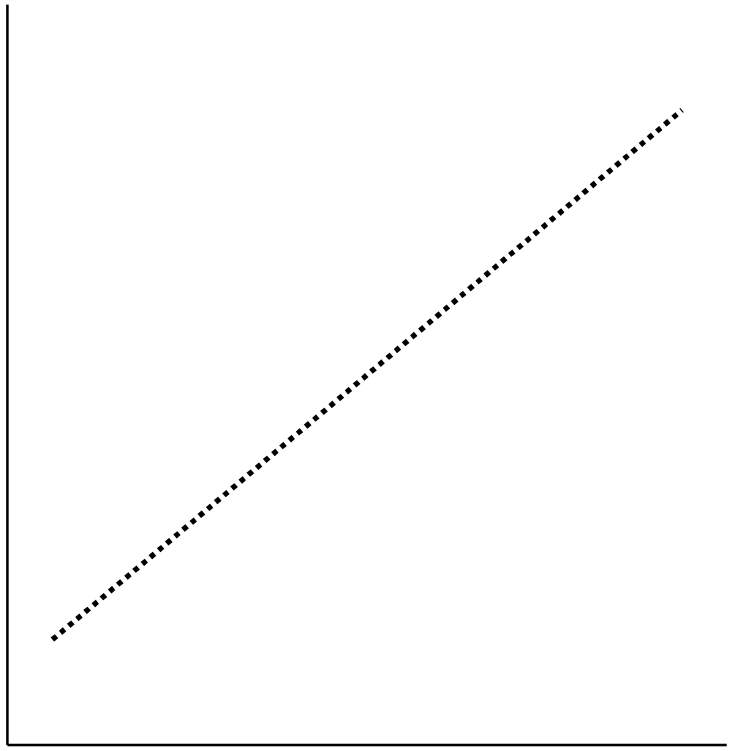


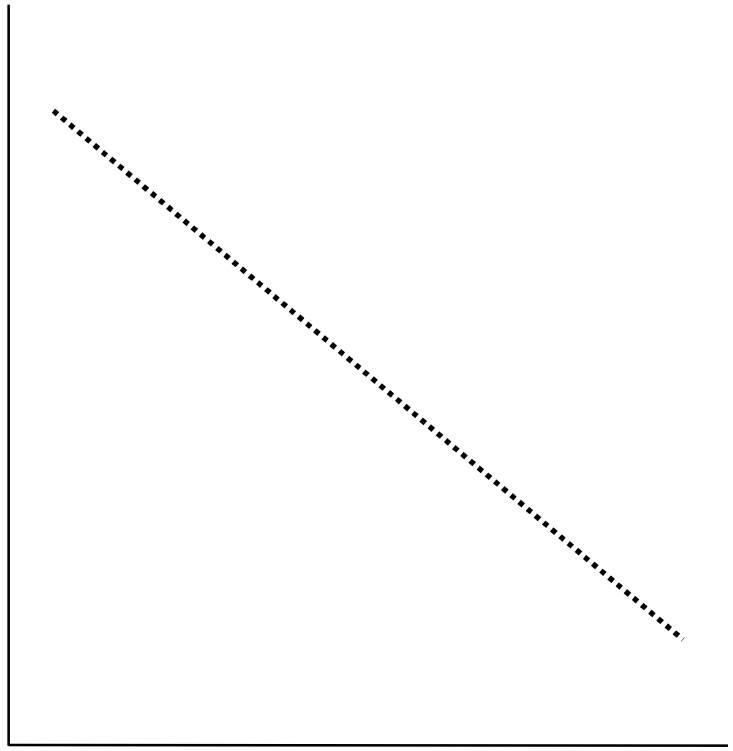






相関係数でわかる関係とは

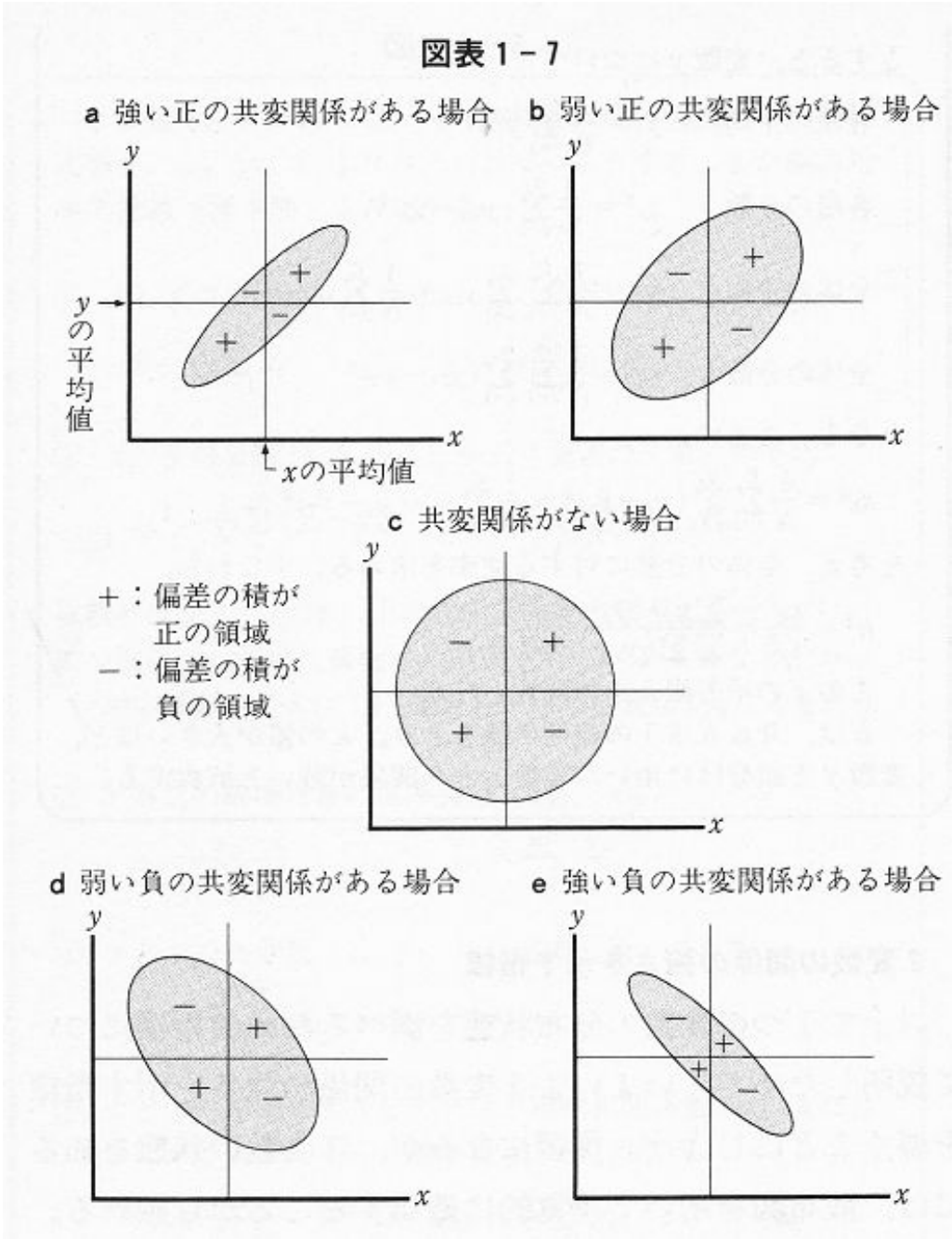




相関係数の数値の見方

- -0.2～0.2　：　ほとんど相関がない
- 0.2～0.4　（-0.2～-0.4）　：　やや相関がある
- 0.4～0.7　（-0.4～-0.7）　：　かなり相関がある
- 0.7～1.0　（-0.7～-1.0）　：　強い相関がある

図表 1－7


a 強い正の共変関係がある場合
b 弱い正の共変関係がある場合
y
x
yの平均値
xの平均値
+
−
c 共変関係がない場合
+：偏差の積が
正の領域
−：偏差の積が
負の領域
d 弱い負の共変関係がある場合
e 強い負の共変関係がある場合

相関係数の計算方法

（xのデーターxの平均値）×（yのデーターyの平均値）
／xの偏差平方和のルート×yの偏差平方和のルート

相関係数が高くなる理由

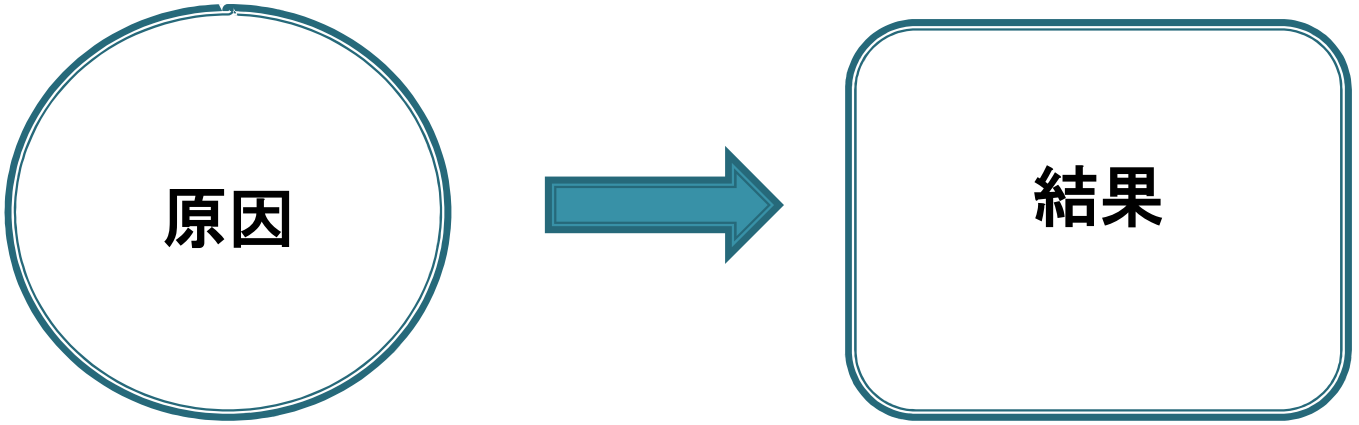
- ①データ間に「原因ー結果」関係が存在する
- ②データ間に「共通したもの」が存在する

データ間に「原因ー結果」関係が存在する

- ある事柄が生じれば、必ず、もう一つ別の事柄が生起するという場合、両者は「原因ー結果」の関係にあると言う。こうした場合、データ間の「相関係数」は高くなると言えよう。
- 【例】「親の身長」と「子どもの身長」

「説明変数（独立変数）」と「被説明変数（従属変数）」

- この場合、原因となるデータを「説明変数（独立変数）」、結果となるデータを「被説明変数（従属変数）」と言う。



データ間に「共通したもの」が存在する

- 　データ（A）とデータ（B）が似たもの同士である場合、両者の間には互いに「共通するもの」があると考えることができる。こうした場合、お互いに「共通するもの」を測っているので、データ間の「相関係数」が高くなると言えよう。
- 【例】「嵐の好き嫌い」と「SMAPの好き嫌い」

多変量解析

- ３つ以上のデータ（変数）の関係の構造（あり方）をとらえる統計的技法

多変量解析の種類

- ①3つ以上のデータ間の「原因ー結果」関係をとらえるためのもの
- （重回帰分析など）
- ②3つ以上のデータを「似たものどうし」にグループ化するためのもの
- （因子分析、クラスター分析など）
- ③上の①と②の組み合わせ
- （共分散構造分析など）

上の統計分析技法のすべて

- ＝平均値と分散をもとにした分析

- 相関係数は、平均値と偏差平方和（分散）が計算式のもとになっている。
- 多変量分析は、相関係数をもとにつくられた分析技法である。
- →もし平均値と分散が使えなければ、意味がない。

平均値と分散が使えないデータ

		名義	順序	間隔	比例
代表値	平均値	×	△	○	○
	中央値	×	○	○	○
	最頻値	○	○	○	○
散らばり	カテゴリーの数	○	○	△	△
	レンジ	×	○	○	○
	偏差平方和	×	△	○	○
	分散	×	△	○	○
	標準偏差	×	△	○	○

クロス集計表とは

- **縦軸（column）と横軸（row）を交差させてつくる表で、2つのデータ間の関係を考えるために用いられる。**

	タイタニック		計
	好き	きらい	
男性	10 20.00%	40 80.00%	50
女性	15 30.00%	35 70.00%	50
計	50	50	100

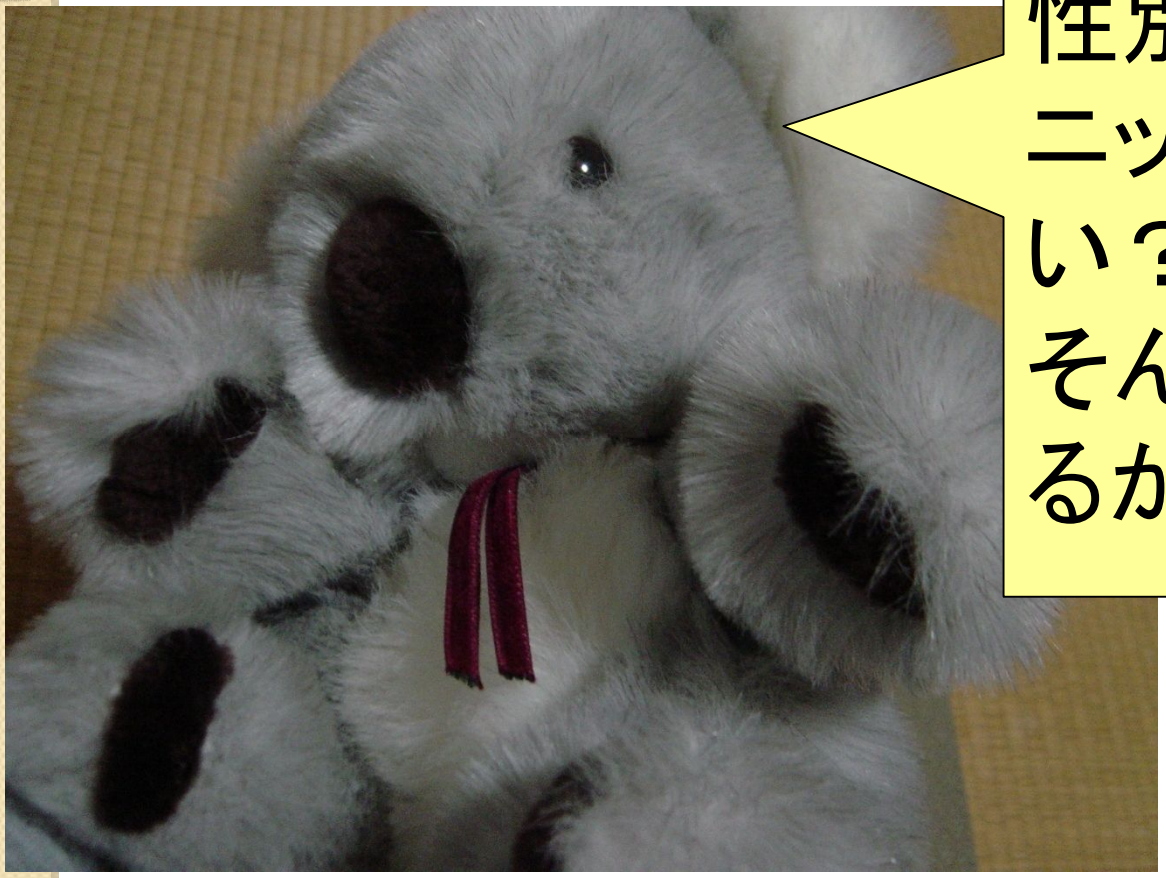
クロス集計表で、でてきた%の違い

- 意味ある違いなのか
- 意味のない違いなのか
- →分からない。
- ＝これを検討するために行う統計がχ^2（カイ二乗）検定である。

もともと証明したかった仮説

性別と、タイタニックの好き嫌いって、関係ある！！！

ライバル仮説（帰無仮説）



性別と、タイタニックの好き嫌い？
そんなの関係あるかよ！！！

カイ二乗検定とは

- ライバル仮説（帰無仮説）が勝つ確率をだす
- ＝ライバル仮説（帰無仮説）の勝つ確率が低ければ低いほど
- →もともと証明したかった仮説（関係あるという仮説）が勝つ確率が高くなる。

ライバル仮説の勝つ確率がどれくらい低ければ、関係あると見なせるのか？

- ライバル仮説が5%以上ならダメ
- ライバル仮説が5%未満なら満足
- ライバル仮説が1%未満ならもっと満足
- ライバル仮説が0.1%未満なら、すごく大満足

第１種の誤りと第２種の誤り

		帰無仮説	
		採択	棄却
帰無仮説	正しい	問題なし	のんびり屋さん （第１種の誤り）
	間違い	あわてんぼさん （第２種の誤り）	問題なし

カイ二乗検定の計算方法

- 「期待値」と「実測値」のズレを基本に計算している。

期待値とは

- 確率どおりでてくるとしたら、生じる値

	タイタニック		計
	好き	きらい	
男性	25	25	50
女性	25	25	50
計	50	50	100

実測値とは

- 実際に生じた値

	タイタニック		計
	好き	きらい	
男性	5	45	50
女性	45	5	50
計	50	50	100

	タイタニック		計
	好き	きらい	
男性	25	25	50
女性	25	25	50
計	50	50	100

	タイタニック		計
	好き	きらい	
男性	5	45	50
女性	45	5	50
計	50	50	100

基本的な考え方

- 実際に生じた値は、期待値（確率どおりの値）と比べて「かたより」がある。
- →ということは、そこに何か意味があるのではないか？
- →そこに意味がある関係があるから、かたよるのだ！！！

もともと証明したかった仮説

- あるデータともう一つのデータとの間には何らかの意味ある関連がある
- →だから、単に確率どおりの値となっているわけではない。ある所はヘンに多かったり、別のところはヘンに少なかったりする。
- ＝期待値と実測値にズレが生じる

ライバル仮説

- あるデータともう一つのデータとの間には何らかの意味ある関連などない
- →だから、単に確率どおりの値となっているにすぎない
- ＝期待値と実測値はズレているわけがない

期待値と実測値のズレ

- 大きい
- ⇒「もともと証明したかった仮説」

- 小さい
- ⇒「ライバル仮説」

カイ二乗値の計算方法

- （実測値－期待値）の二乗÷期待値

- これを一つ一つのマス目でだして、全部たす。

カイ二乗分布表から

- 計算してでてきた、カイ二乗値が生じるのは、何%の確率なのか？

- ⇒カイ二乗分布表というものでチェック！