＜第2回＞

データ解析基礎

統計数理研究所

馬場　康維

baba@ism.ac.jp

データ解析基礎

概　要

データサイエンティストとしてデータ分析を行う際に
知っておくべき統計分析の基礎と手法について学ぶ

1. データの概要を知る
2. 横断的解析と縦断的解析
3. 相関と回帰
4. 重回帰分析
5. 記述・推測・モデル
6. 一般的な注意

参考文献

1. データの概要を知る

データ表現の尺度は何か

データによって適用する手法が異なる

・分類尺度

変数＝性別	値＝(男，女)
変数＝趣味	値＝(スポーツ，読書，…)
	質的データが得られる

・順序尺度

変数＝品質	値＝(秀，優，良)
変数＝等級	値＝(1級，2級，…)
	順序つきの質的データ
	順序つきカテゴリーデータ

1. データの概要を知る

データ表現の尺度は何か

データによって適用する手法が異なる

・間隔尺度

変数＝身長	値＝＊＊＊cm
変数＝体重	値＝＊＊＊kg

量的データが得られる
値の差（間隔）が定義される

・比尺度

変数＝収入/人	値＝＊＊＊円/人

量的データが得られる
比＝分子の数量/分母の数量

1. データの概要を知る

データによって適用する手法が異なる

- 数量から数量を予測
  - 回帰分析
- 数量から群を予測
  - 判別分析
- 調査項目への反応から数を予測
  - 数量化I類
  - ポアソン回帰分析

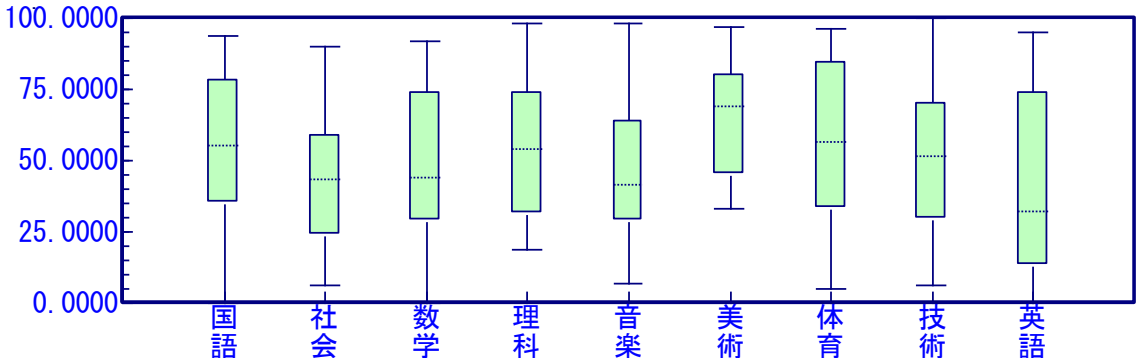
etc.

1. データの概要を知る

データの分布を調べる
・棒グラフを描く
・ヒストグラムを描く
・円グラフを描く
・帯グラフを描く
・箱ひげ図を描く（ボックスプロット）

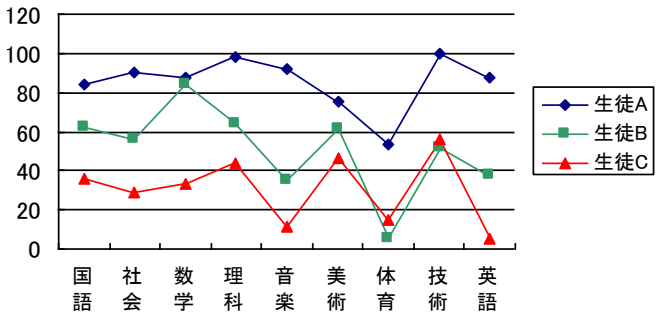
1. データの概要を知る

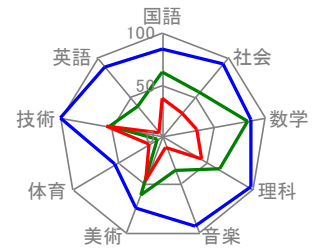
データの分布を調べる
・箱ひげ図を描く（ボックスプロット）



1. データの概要を知る

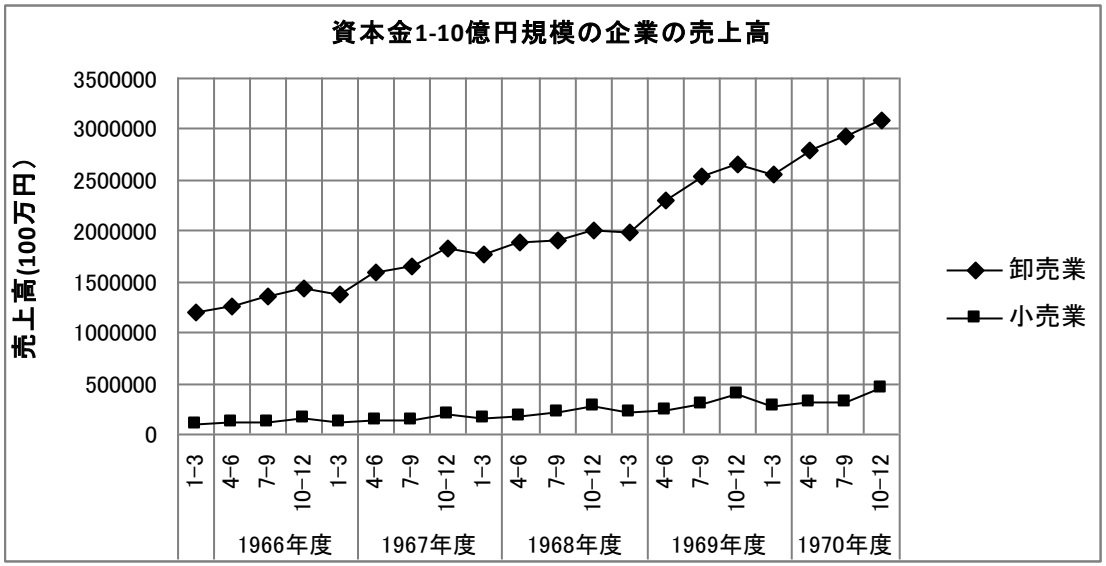
データのプロフィールを見る
・レーダーチャート
・折れ線グラフ





1. データの概要を知る

データのプロフィールを見る
・折れ線グラフ(時系列)

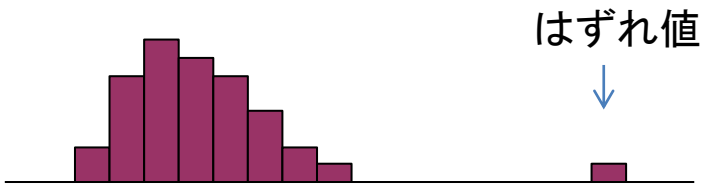


1. データの概要を知る

データの分布を調べる

- 棒グラフを描く
- ヒストグラムを描く
- 円グラフを描く
- 帯グラフを描く
- 箱ひげ図を描く（ボックスプロット）
- レーダーチャートを描く
- 折れ線を描く

データの特徴の抽出
はずれ値の発見
データの入力ミスの発見

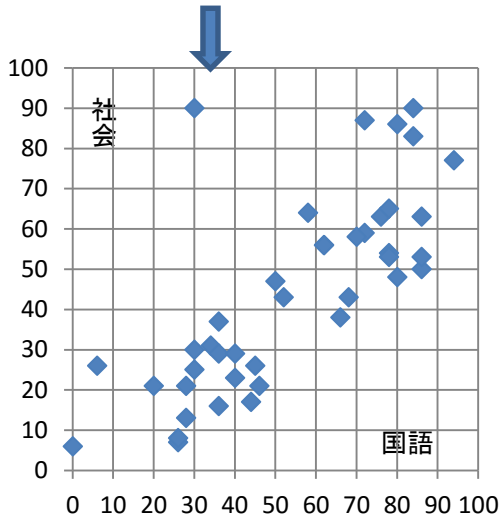


1. データの概要を知る

データの分布を調べる(2変数)
・散布図を描く
・クロス集計表を作る

データの分布を調べる(多変数)
・対散布図を描く
・多重クロス集計表を作る

はずれ値の発見など



国語と社会の点数の散布図

2. 横断的解析と縦断的解析

データ行列

	魚介類	生鮮魚介		鮮　魚	
	金額(円)	金額(円)	数量（ｇ）	金額(円)	数量（ｇ）
北海道	84,340	44,241	36,158	38,406	31,227
東北	84,740	45,199	35,201	40,948	31,721
関東	77,714	43,580	29,131	39,298	26,117
北陸	87,111	49,758	36,874	45,824	33,866
東海	73,875	42,088	27,590	38,493	24,926
近畿	81,673	49,125	32,108	45,137	29,444
中国	75,979	46,276	32,965	42,065	29,920
四国	75,455	44,858	31,483	41,149	28,548
九州	70,464	41,712	31,799	38,853	29,320
沖縄	44,220	25,912	19,828	24,535	18,768

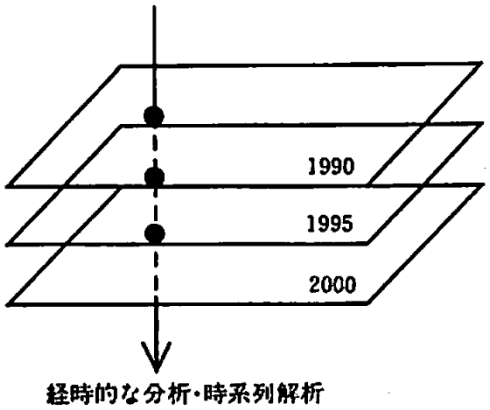
世帯あたり年間消費量および金額（平成24年家計調査）

2. 横断的解析と縦断的解析

年度別のデータ行列を重ねる

・1枚のシートの中での分析
　　横断的分析

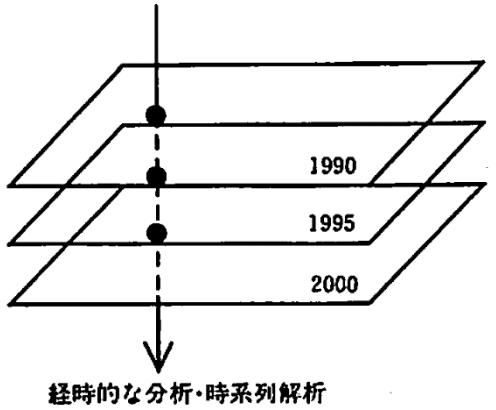
・一つの項目を縦に見ると時系列解析
　　縦断的解析



2. 横断的解析と縦断的解析

縦断的解析
- 多変量解析/多次元解析
  横断的解析
- 時系列解析
  縦断的解析

	魚介類	生鮮魚介		鮮　魚	
	金額(円)	金額(円)	数量（ｇ）	金額(円)	数量（ｇ）
北海道	84,340	44,241	36,158	38,406	31,227
東北	84,740	45,199	35,201	40,948	31,721
関東	77,714	43,580	29,131	39,298	26,117
北陸	87,111	49,758	36,874	45,824	33,866
東海	73,875	42,088	27,590	38,493	24,926
近畿	81,673	49,125	32,108	45,137	29,444
中国	75,979	46,276	32,965	42,065	29,920
四国	75,455	44,858	31,483	41,149	28,548
九州	70,464	41,712	31,799	38,853	29,320
沖縄	44,220	25,912	19,828	24,535	18,768



2. 横断的解析と縦断的解析

斜めに見る解析
・コウホート解析

	20–29歳	30–39歳	40–49歳	50–59歳	60–69歳	70歳以上	全体
第2次(1958年)	14	30	41	51	66	63	35
第3次(1963年)	12	20	40	43	54	58	31
第4次(1968年)	13	21	32	48	56	63	30
第5次(1973年)	10	16	27	35	47	59	25
第6次(1978年)	19	23	36	44	55	69	34
第7次(1983年)	15	21	31	39	56	55	32
第8次(1988年)	15	19	28	43	48	54	31
第9次(1993年)	19	29	27	34	48	59	33
第10次(1998年)	12	17	25	28	43	49	29
第11次(2003年)	7	22	23	31	40	51	30
第12次(2008年)	13	18	23	27	36	41	27

宗教を信じる人の比率(日本人の国民性調査より)

3. 相関と回帰

相関

相関係数

$-1 \le r \le 1$

2変数のn組のデータ

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

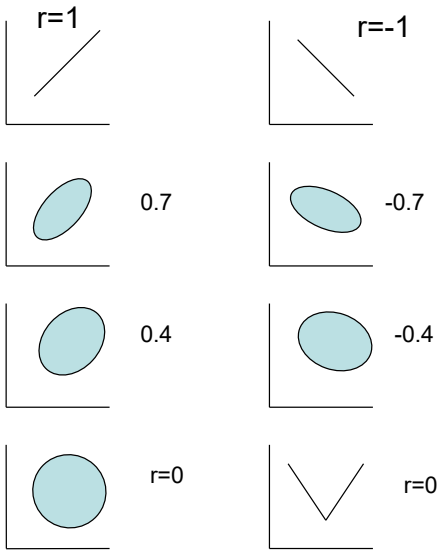
相関係数

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

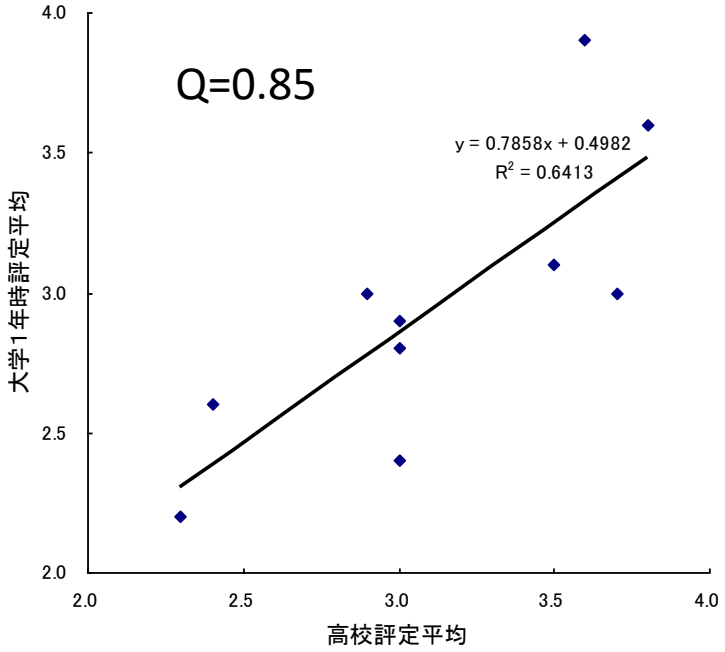
ここで、

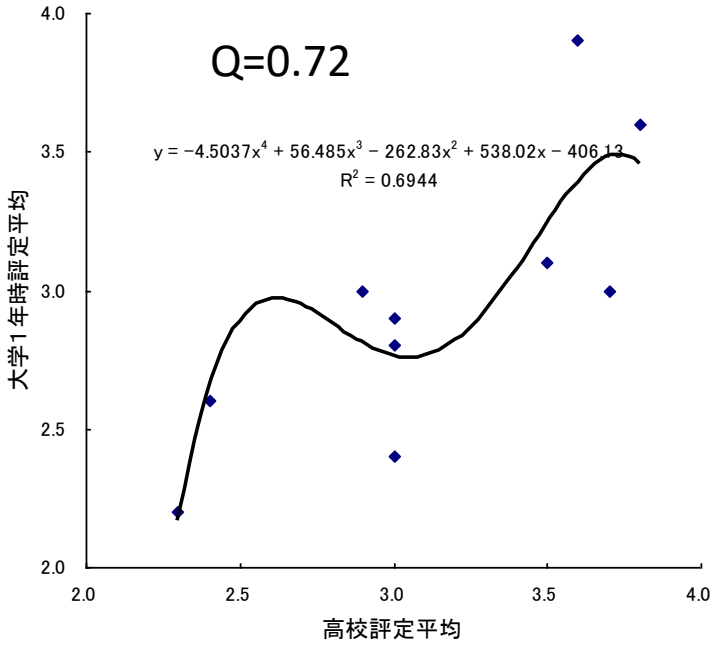
(xの平均) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

(yの平均) $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$



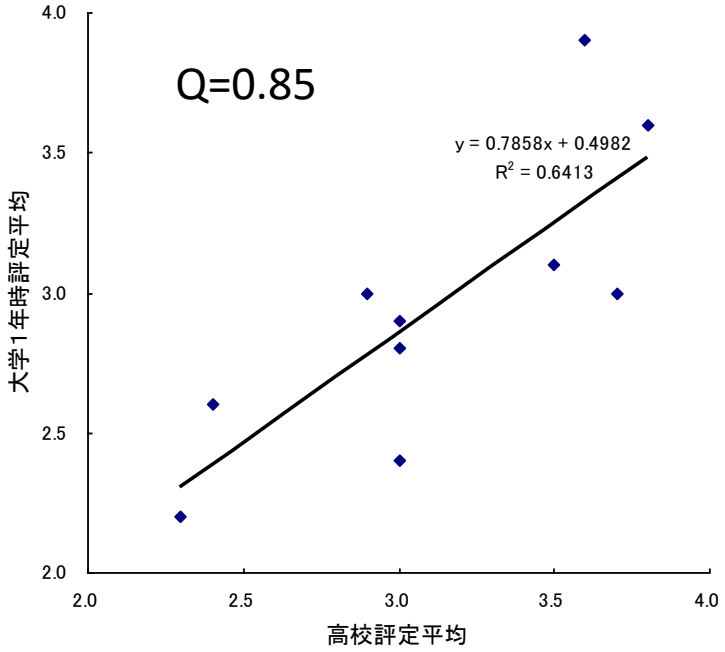
3. 相関と回帰

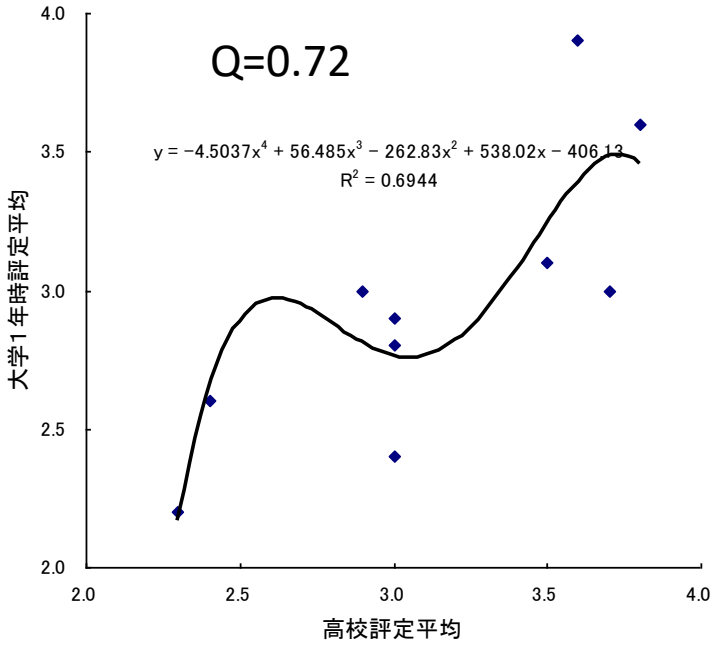




最小2乗法で直線と4次多項式を当てはめた場合
残差平方和(Q) は4次多項式の方が小さくなる

3. 相関と回帰





AIC で評価すると、直線の方が良い
AIC＝赤池情報量規準
AIC＝-2×最大対数尤度 ＋ 2×自由パラメータ数
　　AICの小さなモデルを選ぶ

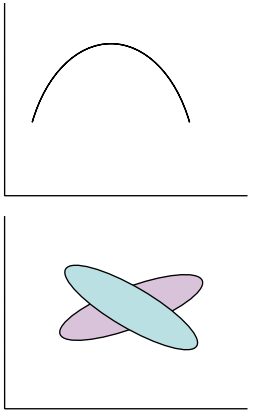
3. 相関と回帰

相関

相関係数が0のとき
　　直線関係はない
　　　しかし
　　直線ではない関係があるかも知れない

グラフを描いてみるのが重要

右図は2次の関係がある場合と傾向の違う2つの層が混じっている場合



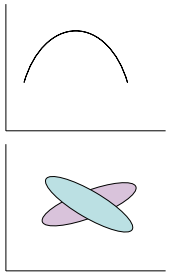
3. 相関と回帰

相関

相関係数が0のとき
　　直線関係はない
　　　しかし
　　直線ではない関係があるかも知れない

グラフを描いてみるのが重要

右図は2次の関係がある場合と傾向の違う2つの層が混じっている場合



二つの層が混じっている場合
　別々に回帰式を作る方が良い場合がある

4. 重回帰分析

説明変数の関数を使って数量を予測する手法

y　　目的変数

$x_1, x_2, \ldots, x_p$　　p 個の説明変数

$\hat{y} = \sum_{j=1}^{p} w_j x_j + w_0$　　目的変数 y の予測式

4. 重回帰分析

説明変数の関数を使って数量を予測する手法

y 目的変数

$x_1, x_2, \ldots, x_p$ p 個の説明変数

$\hat{y} = \sum_{j=1}^{p} w_j x_j + w_0$ 目的変数 y の予測式

・変数を増やせば当てはまりが良くなるが、新しいデータに当てはめたときには必ずしも良くない。
・変数選択が重要になる

4. 重回帰分析

説明変数の関数を使って数量を予測する手法

y 目的変数

$x_1, x_2, ..., x_p$ p 個の説明変数

$\hat{y} = \sum_{j=1}^{p} w_j x_j + w_0$ 目的変数 y の予測式

・変数を増やせば当てはまりが良くなるが、新しいデータに当てはめたときには必ずしも良くない。
・変数選択が重要になる

・AICが変数選択に有効

5. 記述・推測・モデル

データ解析の様々なステージ

・記述的な方法
　　発見的・探索的な方法
・推測の方法
　　推論・推測・予測
・モデル
　　モデルを作ることにより、現象を理解する
　　モデルにより推論を行う

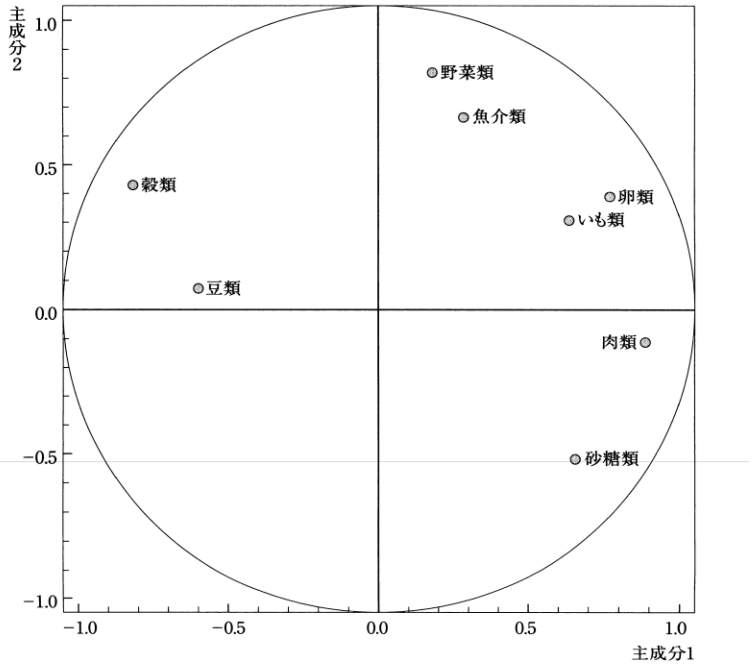
5. 記述・推測・モデル

データ解析の様々な方法

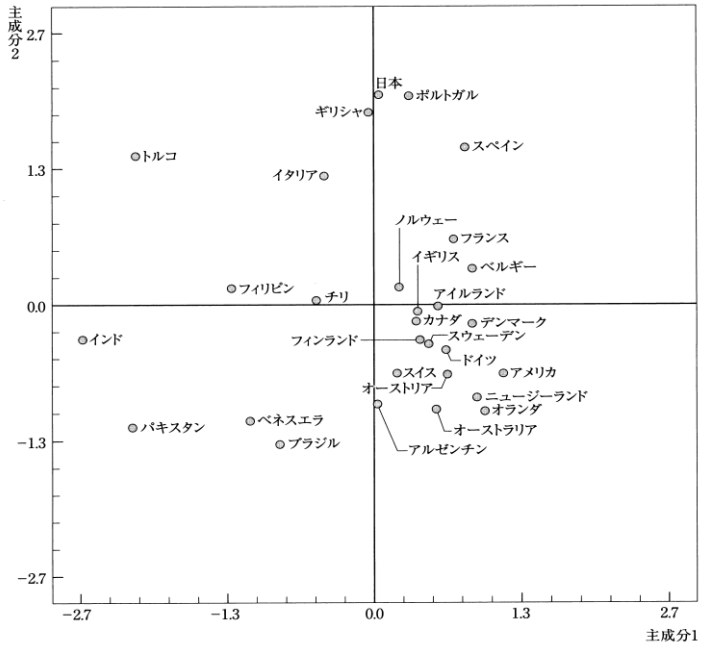
・記述的な方法
　　発見的・探索的な方法
　　主成分分析（量的データに適用）
　　　　相関構造を総合特性値（主成分）により
　　　　見方を変える
　　　　直交変換による座標回転の方法
　　数量化Ⅲ類（質的データに適用）
　　　　コレスポンデンスアナリシス
　　　　双対尺度法

5. 記述・推測・モデル

主成分分析（量的データに適用）



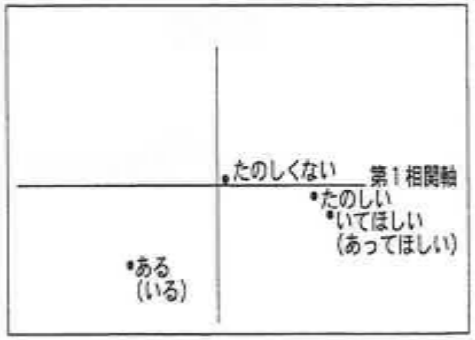
注: データは、8食品項目の1992～1994年平均、一人当たり食糧供給量

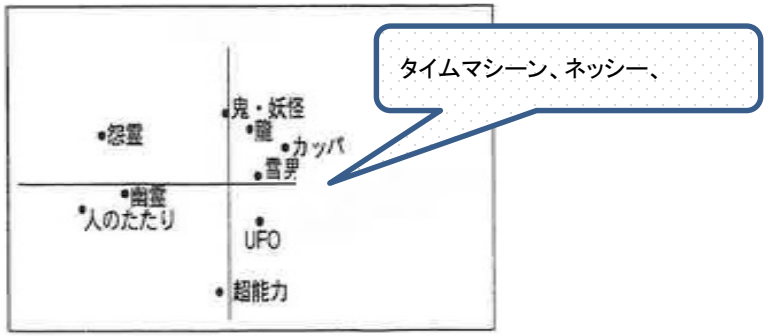


注: データは、8食品項目の1992～1994年平均一人当たり食糧供給量

5. 記述・推測・モデル

数量化Ⅲ類（質的データに適用）





林知己夫他(1979)統計数理研究所研究リポート44より

5. 記述・推測・モデル

データ解析の様々な方法

・推測の方法

推論・推測・予測

回帰分析

多項式回帰（連続的な変数の予測）

重回帰（連続的な変数の予測）

ポアソン回帰（離散的な現象の推測）

5. 記述・推測・モデル

データ解析の様々な方法

・モデル

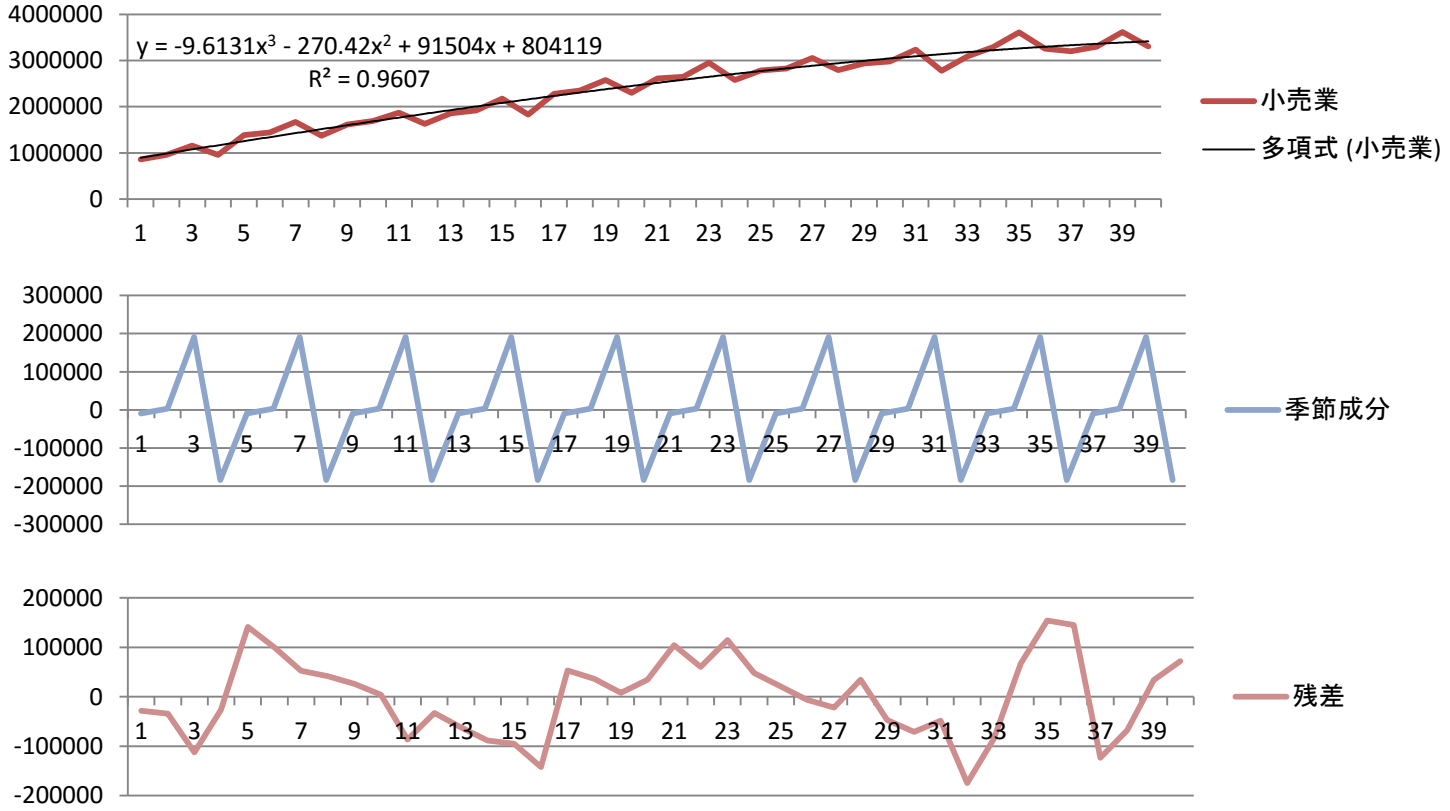
モデルを作ることにより、現象を理解する
モデルにより推論を行う

回帰モデル
一般化線形モデル
因子分析（観測値から因子を抽出）

時系列モデル
時系列の分解
時系列＝トレンド＋季節成分＋偶然変動

5. 記述・推測・モデル

・時系列の分解　小売業売上＝トレンド＋季節成分＋偶然変動



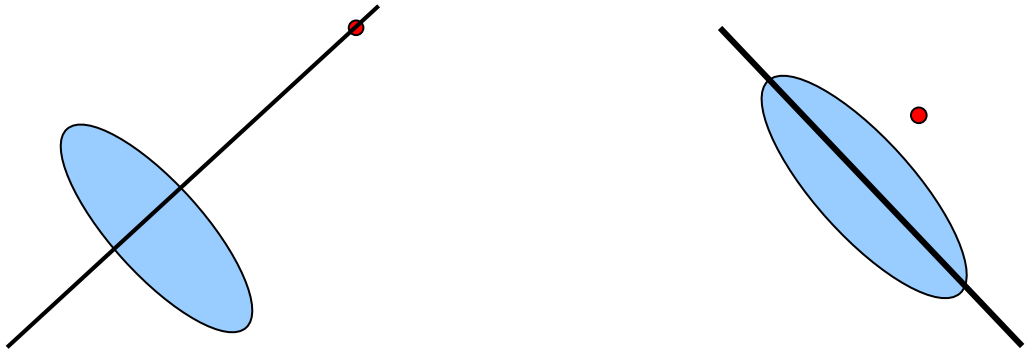
6. 一般的な注意

データ解析の際に注意すること

・データにはずれ値がないか

はずれ値により結果が大きく影響される

はずれ値の吟味が必要



6. 一般的な注意

データ解析の際に注意すること

・データに内部構造がないか

・交絡因子がないか

・欠測値の扱い

・非線形効果の存在

・データの定義

6. 一般的な注意

データ解析の際に注意すること

目的は何か

例：物理実験

直線関係が理論的に想定される
回帰式は、その直線関係の推定
傾きに関心がある

例：体重と身長の計測

直線は身長を与えたときの平均体重の推測
直線からのばらつきに関心がある

どちらも最小2乗法を用いる

参考文献

1. 奥野忠一他，『（改訂版）多変量解析法』（日科技連，1981年）
2. 大隅昇，馬場康維他，『記述的多変量解析法』（日科技連，1994年）
3. 小西貞則『多変量解析入門』（岩波書店，2010年）
4. B.エヴェリット著，石田他訳，『RとS-PLUSによる多変量解析』（丸善出版，2012年）
5. 小西貞則，北川源四郎，『情報量規準』（朝倉書店，2014年）
6. 西里静彦，『行動科学のためのデータ解析ー情報把握に適した方法の利用ー』（培風館，2010年）