

保健統計学実習

第5日目

- 第13回 主成分分析、因子分析、クラスター分析
- 第14回 解析実習まとめ：統計解析ツール R/R studio 解析例
- 第15回 (復習・課題の時間)

滋賀医科大学
NCD疫学研究センター 医療統計学部門

原田 亜紀子
(aharada@belle.shiga-med.ac.jp)

本日の実習内容

1. 主成分分析、因子分析、クラスター分析
2. 実習・まとめ
- 1) まとめ
- 2) 教科書など

講義・演習スケジュール

1 R, EZRの使い方、データセットの読み込み、頻度集計、記述統計, 相関 2 EZRのコード保存, R-studio commander 3 エクセルの基礎(1)	8/29(木)
4 仮説検定の基礎, 2群の比較(t検定, Wilcoxon検定) 5 カイニ乗検定、マクネマー検定 6 調査データ解析(1): 調査票作成、データ入力	8/30(金)
7 重回帰分析 8 ロジスティック回帰、検査データの解析 9 調査データ解析(2): (web調査ツールを使用した)調査票作成、データ入力	9/2(月)
10 分散分析 11 サンプルサイズ 12 調査データ解析(3): 解析用データの作成	9/5(木)
13 主成分分析、因子分析、クラスター分析 14 解析実習・まとめ (復習・課題の時間)	9/6(金)

第13回

1. 主成分分析、因子分析、クラスター分析

主成分分析(Principal Component Analysis)・因子分析(Factor Analysis)

因子分析 (FA)

潜在変数

因子1
文系科目

因子2
理系科目

因子パターン

国語
地理
公民
英語
数学
物理
化学
統計

Unobservable ← Observable

変数と要因の相関を説明する。
 $X_j = a_{j1}f_1 + a_{j2}f_2 + e_j$
x: Observed variable, a: Factor loadings
f: Common factor, e: Uniqueness

主成分分析 (PCA)

主成分

主成分1

主成分2

Weight

国語
地理
公民
英語
数学
物理
化学
統計

Synthesis score ← Observable

複合変数を、変数群の中で可能な限り最大の分散として計算する。
 $Z_j = a_{j1}X_1 + a_{j2}X_2 \quad *a_{j1}^2 + a_{j2}^2 = 1$
z: Principal components, a: Principal component loadings
x: Observed variable

5

主成分分析

第1主成分(PC1)
データの分散が最大となる方向

【寄与率】
この主成分だけで元のデータの何パーセントを説明できるかを示す数値。

第1主成分と直角に交差する方向(直交方向)で分散が最大となる場所を計算する(第2主成分: PC2)。

PC1とPC2で94%説明可能

ID1は理系科目が得意
ID2は両方得意

「寄与率」

	PC1	PC2
寄与率	0.68	0.26
累積寄与率	0.68	0.94

「負荷量」

	PC1	PC2
国語	0.08	0.93
数学	0.96	-0.22
理科	0.98	0.07
社会	0.28	0.90

「主成分得点」

ID	PC1	PC2
1	47	-32
2	51	28

7

主成分分析

データの次元を下げる

- 主成分分析の目的は、測定された変数の集合から、元の変数の変動をできるだけ多くとらえる少数の独立した線形結合(主成分)を導き出すことである。
- 主成分分析は、次元削減の手法であると同時に、探索的データ分析ツールでもある。


主成分分析のアルゴリズム

- 全データの重心(平均値)を算出する。
- 重心から、データの分散が最大となる方向(第1主成分)を算出する。
- 第1主成分と直角(直交)に交わる方向で、分散が最大となる場所を算出する(第2主成分)。
- 直近の主成分と直交する方向で分散が最大となる場所を算出する(第3主成分)。
- データの各次元について、手順④を繰り返す。

6

例題 東京都の自治体データ

- 東京都の自治体の各指標25指標
 - 市町村、世帯あたり人数、年齢15未満比率、年齢65以上比率、転入者対人口比、転出者対人口、昼間人口比、高齢単身世帯比率、
 - 第1次産業従業者数比、第2次産業従業者数比、第3次産業従業者数比、
 - 可住地面積比率、耕地面積、対可住面積比、
 - 課税所得、就業者1人あたり千円、小売業販売額、事業所あたり百万円、小売業販売額、売場面積あたり万円、国民健保一人あたり診療費、円、こみりサイクル率、pct、千人あたり事業所数、千人あたり幼稚園数、千人あたり飲食店数、千人あたり大型小売店数、千人あたり病院数、千人あたり老人ホーム数、千人あたり交通事故発生件数、千人あたり刑法犯認知件数
- 自治体に住みたいかどうかの調査結果を点数化した「人気度」
- データセット:
 - TokyoSTAT_P25.csv



8

5

11

10

12

データ

変数を選択→
「行政CD」、「人気度」を除く変数を選択
データオプション

オプション

解析の指示、主成分得点算出
オプション
主成分の数
4
OK

13

Scree(スクリー) plot & 平行分析

result.prl <- fa.parallel(DF[, -(1:3)], fm="ml")

Parallel Analysis Scree Plots

Parallel analysis suggests that the number of factors = 3 and the number of components = 2

PC Actual Data
PC Simulated Data

- Scree plotは、成分番号に対して固有値をグラフ化したもの。
- 第3成分から、線はほとんど平坦である。
→平坦になる前が目安
→第何成分まで採択するかが目安

第3主成分ないしは第5主成分まで？

15

データセットをみてみよう

市町村	行政CD	人気度
千代田区	13101	160
中央区	13102	125
港区	13103	214
新宿区	13104	147
文京区	13105	151
台東区	13106	67
墨田区	13107	51
江東区	13108	71
品川区	13109	123
目黒区	13110	163
大田区	13111	70
世田谷区	13112	171
渋谷区	13113	166
中野区	13114	107
杉並区	13115	119
豊島区	13116	101
北区	13117	62
荒川区	13118	31
板橋区	13119	54
練馬区	13120	74
足立区	13121	37
葛飾区	13122	35
江戸川区	13123	62
八王子市	13201	31
立川市	13202	54
武蔵野市	13203	95
三鷹市	13204	46
青梅市	13205	12
府中市	13206	44
昭島市	13207	15

... (略) ...

PC1	PC2	PC3	PC4
-15.77124932	8.53288208	-2.05180523	0.917721990
-7.16839787	0.74260754	2.72628731	-1.232065582
-7.30245870	0.01611210	1.91950570	-1.808769402
-4.55689961	-2.11715365	0.61943078	0.060332473
-2.02959854	-1.90450258	-0.34201432	-0.487093675
-2.72989197	-1.40057400	-2.07091636	0.057530572
-0.07253100	-1.20731768	-1.63809521	-1.307611346
0.19857719	-1.17465781	-0.64660146	-0.349826939
-1.29276323	-2.01544656	-0.52657385	-0.904825064
-1.62634895	-2.08562253	0.70129243	0.018556747
0.41380898	-1.09542553	-1.62927471	-1.508014210
-0.80205004	-2.11530868	1.16451525	0.306417725
-5.45777243	-1.58782236	0.40840154	0.696777985
-1.51784041	-2.67210095	0.95037880	0.505816364
-1.43980236	-3.12136829	-0.68138458	0.339685134
-3.38107771	-2.64318215	0.93231287	0.256744821
-0.00400686	-1.87114904	-3.86597193	0.083229747
0.62068500	-1.15398721	-2.47675258	-1.294781301
0.30276468	-1.18658672	-1.34060799	-0.856069872
0.49470165	-1.31491454	0.52167405	-0.530830890
1.19448283	-0.35517311	-2.58586374	-0.705847494
-1.34528366	-0.60067949	-2.10096927	-0.999404176
1.32283573	-0.36842504	0.37016570	-1.729549287
2.33410656	1.63970258	0.54524151	0.425909074
0.60836114	0.06124135	0.41554065	1.958171684
-1.14103487	-0.99124818	0.42900764	1.430747179
0.57597601	-0.30472444	1.41151248	0.009897755
4.81886846	4.93166676	-0.01194936	0.659828420
1.26159989	0.36548562	1.40533303	-0.508603438
2.18112116	1.16410360	-0.50667348	-0.823687308

自治体別の主成分得点(PC1~4)が出力されている

14

EZRでの算出

16

主成分分析:固有ベクトル(第4主成分まで)

Component loadings:

	Comp.1	Comp.2	Comp.3	Comp.4
ごみリサイクル率_pct	0.14313231	0.127942834	0.261682729	0.27452234
可住地面積比率	-0.11383968	-0.257891119	-0.087326397	-0.10087182
課税所得_就業者1人あたり_千円	-0.23531538	-0.003764779	0.128319792	-0.03621981
耕地面積_対可住面積比	0.15650544	0.244937181	0.128000027	0.25558928
高齢単身世帯比率	-0.07399042	-0.218733757	-0.512410750	0.13457580
国民健康一人あたり診療費_円	0.13891741	0.086561773	-0.327457796	-0.01514818
小売業販売額_事業所あたり_百万円	-0.18853341	0.053126273	0.270612089	-0.07899511
小売業販売額_売場面積あたり_万円m2	-0.23711598	-0.087377992	0.069927304	-0.13392322
世帯あたり人数	0.21817823	0.282007358	0.016852864	-0.02088900
千人あたり飲食店数	-0.26230625	0.191112884	-0.023531881	-0.01938072
千人あたり刑法犯認知件数	-0.24883963	0.187829606	-0.097897803	0.06802987
千人あたり交通事故発生件数	-0.22590231	0.277312460	-0.014390255	-0.02436412
千人あたり事業所数	-0.25378255	0.219289176	-0.056588990	-0.01715808
千人あたり大型小売店数	-0.25021531	0.226388573	0.003021881	0.01530467
千人あたり病院数	-0.15137829	0.264148481	-0.261048405	0.18846312
千人あたり幼稚園数	-0.23505208	0.155132892	-0.083435715	-0.06925427
千人あたり老人ホーム数	0.12741348	0.282245218	0.020278499	0.23492013
第1次産業従業者数比	0.14312036	0.166901626	0.250632303	0.07397827
第2次産業従業者数比	0.14036634	0.191454238	-0.128220532	-0.55880673
第3次産業従業者数比	-0.14144537	-0.182633089	0.125130549	0.55641678
昼間人口比_per	-0.24316331	0.243126759	-0.063282831	0.01854164
転出者_対人口比	-0.26794431	-0.117216398	0.117642903	0.02150505
転入者_対人口比	-0.26194219	-0.035917394	0.057305951	-0.01008152
年齢15未満比率	0.17315873	0.262854276	0.107643674	-0.12396277
年齢65以上比率	0.13827649	0.109660617	-0.462833305	0.24480774

17

主成分分析

Importance of components:

	PC1	PC2	PC3	PC4
Standard deviation	3.3971	2.1271	1.4791	1.27554
Proportion of Variance	0.4616	0.1810	0.0875	0.06508
Cumulative Proportion	0.4616	0.6426	0.7301	0.79518

• 固有値 Eigenvalue

- 最初の成分は、常に最も多くの分散を占めたが、最も高い固有値を持つ、次の成分は、できる限り残りの分散の多くを占め、そして以下同様とづく。

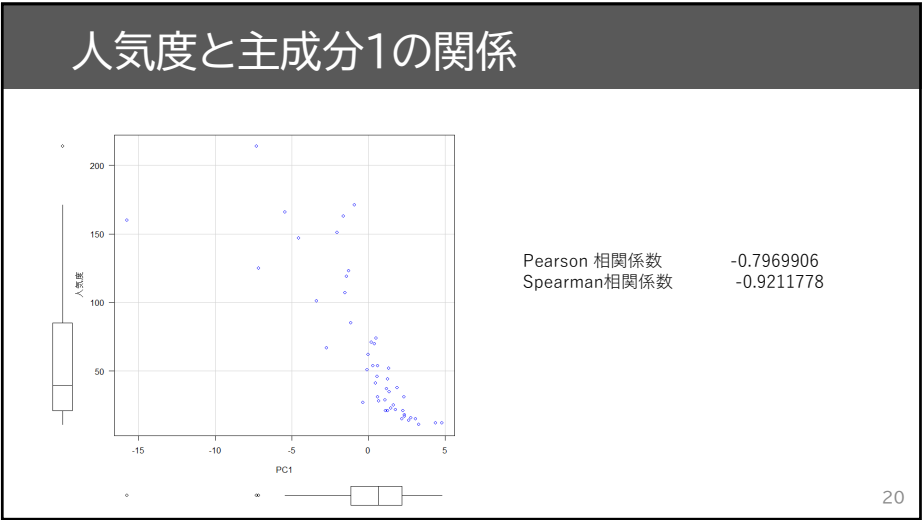
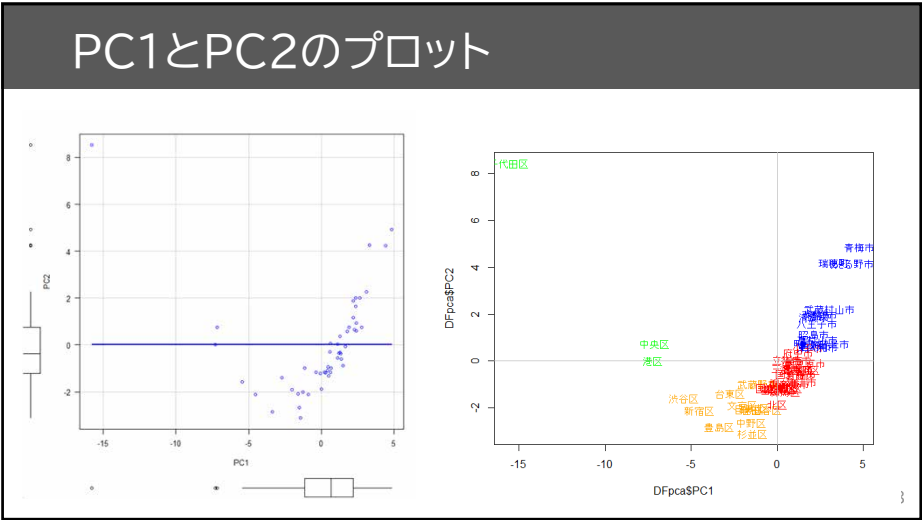
	PC11	PC12	PC13	PC14
Standard deviation	0.53764	0.45391	0.43334	0.36794
Proportion of Variance	0.01156	0.00824	0.00751	0.00541
Cumulative Proportion	0.96718	0.97542	0.98293	0.98835

• 累積 Cumulative

表は、

- 固有値が1以上の成分で累積寄与率が約80%であることを示している

20

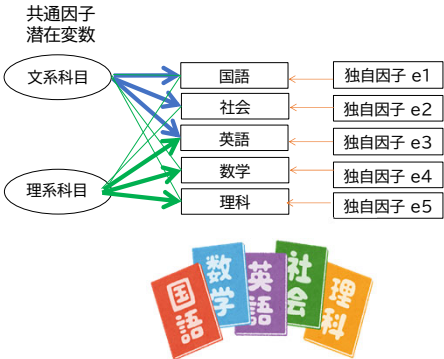


因子分析

21

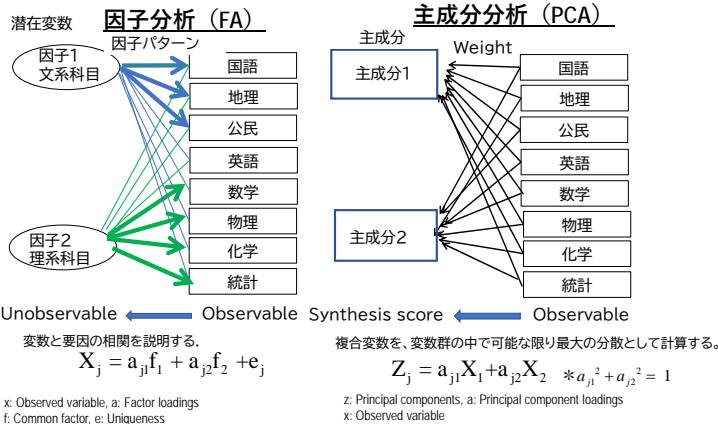
因子分析

- 因子分析は、観測された変数をより少ない数の(観測できない)潜在変数, または因子で記述しようとするものである。
 - 例では“文系科目”“理系科目”
- 因子分析の目的は,
 - ①観測された変数の意味のある解釈を, 観測されない要因の観点から見つける
 - 目に見えない概念
 - ②変数の数を減らすことである。



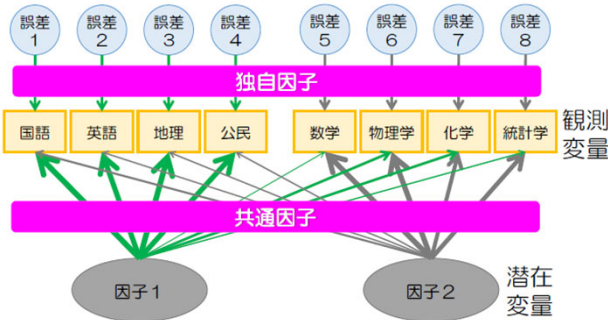
23

主成分分析(Principal Component Analysis)・因子分析(Factor Analysis)



22

共通因子(f)と独自因子(e)



つまり、因子分析とは、独自因子(誤差)をなるべく小さくしながら、因子分析のモデルで説明できる共通因子で表すことができるように、因子分析のモデルは構築される。

24

探索的因子分析→確証(認)的因子分析

■ 探索的因子分析

観測変量に基づいて潜在構造(因子)を探索することが目標。いいかえれば、各因子が影響を与える観測変量を見つけ出し、その影響の大きさから因子に適切な解釈をつけられれば良い。

パス図でもわかるように、探索的因子分析では、因子構造を抽出したとしても、因子に関連のない観測変量に対して、多少の影響を与えるような形式になっている。

潜在構造(因子)が抽出されたのであれば、その仮説のもとで、データがうまく当てはまっているか(つまり因子モデルが妥当であるか)について再検証する必要がある

↓

確証的因子分析

25

因子分析のアルゴリズム

① 相関行列をもとめ、固有値分解する

② 固有値、寄与率から因子数を決定する

③ 因子負荷量の算出：共通因子の影響の強さを示す“因子負荷量”を算出する。

④ 共通因子の名称を決定する。

⑤ 因子スコアの算出

“因子数決定”

“因子の推定法”

【共通性】(Commonality)

- 各観測変数がある因子群でどの程度説明できるかを示す数値。
- 0(全く説明できない)～1(完全に説明できる)の間の値である。
- 1 - 共通性 = 固有の要因の量である。

【要因寄与度】

因子寄与度を観測変数の総数で割ることで、その因子が全体にどれだけ寄与(影響)しているかを見ることができる。

<因子の解釈>

③ 因子軸を回転させる

各観測変数の因子負荷量を散布図グラフにプロットすると、共通因子がそのまま何を指しているのかが分りにくいことが多いので、解釈を容易にするために、各因子の数値が軸に沿うようにグラフの軸を回転させる。

“因子軸の回転法”

27

確証(認)的因子分析

確証的因子分析の目標

確証的因子分析では、予め因子構造を与えたうえで、その妥当性を検討する。

↓

SPSS単体では実行不可能でAMOS(共分散構造分析)を利用する必要がある。

↓

あるいは、JMP15.0以降であれば、共分散構造分析を実行できる。

26

R-commander

28

例題

- ビールの好みに関するマーケティング調査
 - 属性:
 - 性別、年齢(20代、30代、40代、50代)-----231名(欠測データもあり)
 - ビールの嗜好
 - 値段
 - サイズ
 - アルコール度数
 - 評判
 - 色
 - アロマ
 - 味わい

29

因子分析

データ オプション

変数 (3つ以上選択)

アルコール度数
アロマ
サイズ
因子1
因子2
色

部分集合の表現
<全ての有効なケース>

ヘルプ リセット OK キャンセル 適用

「評判」は除いた、5変数

因子分析

データ オプション

因子の回転 因子スコア

☐ なし ☐ なし

☒ バリマックス ☐ バートレットの方法

☐ プロマックス ☒ 回転

回転:バリマックス(直交)
因子得点:回帰

ヘルプ リセット OK キャンセル 適用

「2因子」抽出で

31

[標準メニュー]-[統計量]-[次元解析]-[因子分析]

R コマンドー

ファイル 編集 アクティブデータセット 統計解析 グラフと表 ツール ヘルプ 標準メニュー

データセット: Dataset 編集 表示 保存

スクリプト Rマークダウン

local({
 .FA <-
 factanal("ごみリサイクル率_pct,可住地面積比率,課税所得
 factors=4, rotation="promax", scores="Bartlett", data=
 print(.FA)
 Dataset <- within(Dataset, {
 F4 <- .FA\$scores[,4]
 F3 <- .FA\$scores[,3]
 F2 <- .FA\$scores[,2]
 F1 <- .FA\$scores[,1]
 })
})
Factor Analysis: ごみリサイクル率_pct, 可住地面積比率, 課税所得_就業者1人あたり_千円, 耕地面積_対可住面積比, 高齢単身世帯比率, 国民健
local({
 .FA <-

統計量

要約
グラフ
モデル
分布
ツール
ヘルプ

分割表
平均
比率
分散
ノンパラメトリック検定

次元解析
モデルへの適合

尺度の信頼性...
主成分分析...
因子分析...
検証的因子分析...
クラスター分析

30

スクリーンプロット、固有値

	Factor1	Factor2
SS loadings	2.698	2.531
Proportion Var	0.450	0.422
Cumulative Var	0.450	<u>0.871</u>

Variances

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6

.PC

2.5
2.0
1.5
1.0
0.5
0.0

2.698 2.531 0.450 0.422 0.000 0.000

Red arrow pointing to Comp.3

32

出力

6変数

Call:
factanal(x = アルコール度数 + アロマ + サイズ + 色 + 値段 + 味わい, factors = 2, data = Dataset, scores = "regression", rotation = "varimax")

Uniquenesses:
アルコール度数 0.175
アロマ 0.135
サイズ 0.005
色 0.037
値段 0.288
味わい 0.132

①共通性、独自性

Loadings:
アルコール度数
アロマ
サイズ
色
値段
味わい

Factor1 Factor2
0.930 0.908
0.930 0.993
0.978 0.983
0.929 0.843

②負荷量

SS loadings
Proportion Var
Cumulative Var

Factor1 Factor2
2.698 2.531
0.450 0.422
0.450 0.871

33

補足：因子抽出

- 主因子法
 - 第1因子で説明される全体像を把握
- 最尤法
 - 全体像から傾向ごとに因子を作成してそれぞれを観測変数で説明
- 最小二乗法
 - 最尤法と最小二乗法は誤差の重みづけが異なる
 - 最小二乗法: すべての変数の誤差を同じ重み: 共通性が低い項目の影響も強く受ける
 - 最尤法: 共通性が小さい項目は、重みを小さくして推定

<方針>

- まずひとつの因子にまとめた場合は「主因子法」→それ以外は「最尤法」で実施
- 「最尤法」で実施したいが、各観測変数が正規分布をしていない場合は(対象数が少ない場合なども)「一般化最小二乗法」、それでは不適解になってしまう場合は「重み付けのない最小二乗法」がよい。

35

因子抽出法

- 主因子法 (method=prinit)
 - 第1因子の因子寄与が最大となるように解が得られる(古典的方法)
 - あまり用いられないが、ヘイウッドケースが生じにくい
- 主成分法 (method=principal) * default
 - 各因子の寄与率がなるべく等しくなるように解を求める。
 - 共通性を推定しない
 - 回転を伴わない主成分法の結果は、主成分分析の結果と同じになる
- 最尤法 (method=ml)
 - 解を確率密度により推定する。共分散構造解析でよく利用される。
 - 分布が歪んだデータでも正確な推定ができると言われている
 - サンプル数が大きい時にもっともよい推定ができる方法
- 最小二乗法
 - すべての変数の誤差を同じ重み: 共通性が低い項目の影響も強く受ける
 - SPSSでは重みづけの出力もある

Heywood(ヘイウッド)ケース: 共通性の推定値が1より大きい場合---不適解

- 標本サイズが小さい (もっとデータを集める)。
- データ内に局所的に高い相関が認められる (手法を変更)。
- 因子分析のモデルが適切にデータに当てはまっていない(手法を変更)

34

因子軸の回転

複数の項目に高い因子負荷量がある項目がある状態
→軸を回転させ、単純な構造を得る(解釈しやすいようにする)

直交回転

各因子間の相関が0であること。
通常、バリマックス回転と呼ばれる方法が使われる。

Varimax rotate

Promax rotate

斜交回転

各因子の間に相関がある。
通常、プロマックス回転、直接オブリミンと呼ばれる方法などが使われる。

36

結果

Uniquenesses:

アルコール度数	アロマ	サイズ	色	値段	味わい
0.175	0.135	0.005	0.037	0.288	0.132

各変数の値の変動が因子でどれだけ説明できるか

Uniquenesses (独自性) = 1 - 共通性
取りこぼしの割合(救えなかった情報)

37

因子負荷量
(各変数がどれだけ因子に寄与しているか)

Loadings:

	Factor1	Factor2
アルコール度数		0.908
アロマ	0.930	
サイズ		0.993
色	0.978	
値段		0.843
味わい	0.929	

アロマ、色、味わい
“風味”

アルコール度数
サイズ、値段
“お酒としての実質”

39

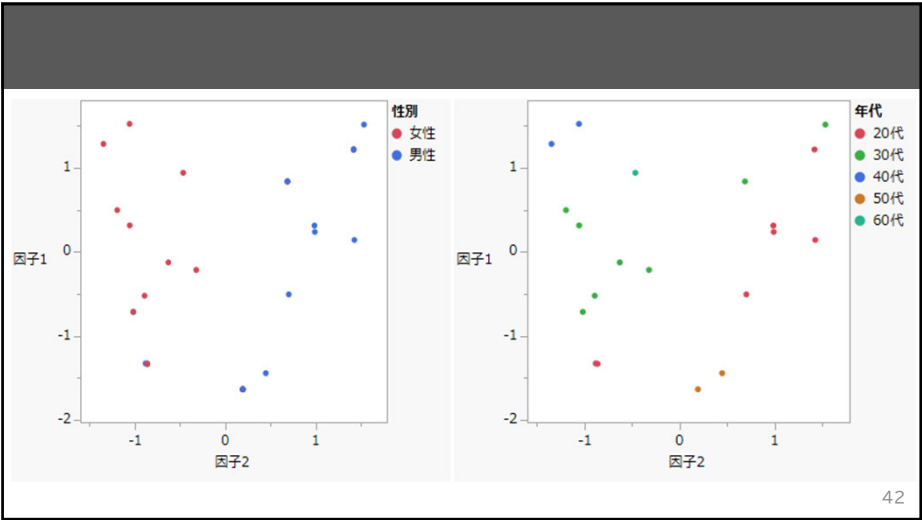
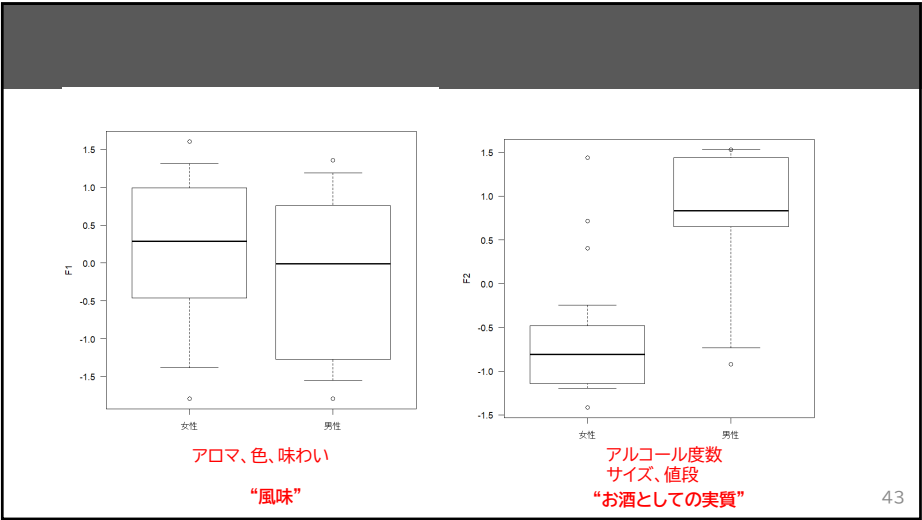
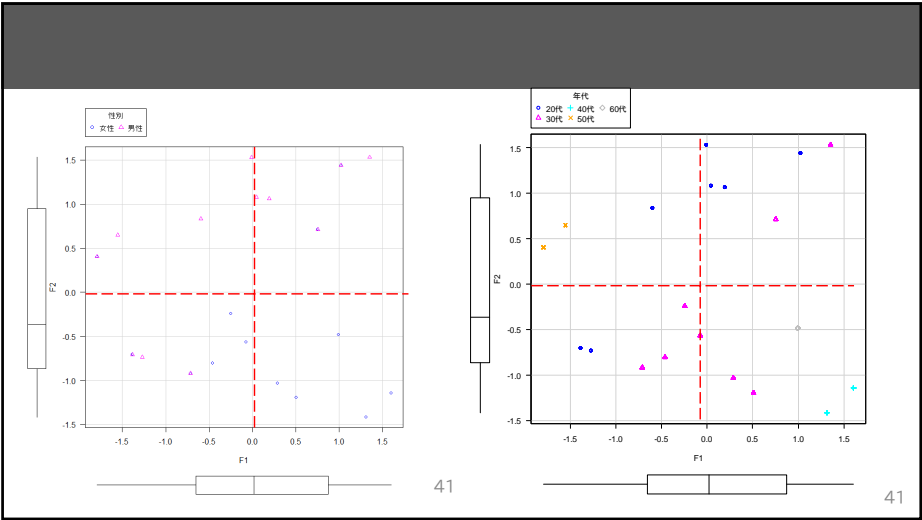
共通因子(f)と独自因子(e)

つまり、因子分析とは、独自因子(誤差)をなるべく小さくしながら、因子分析のモデルで説明できる共通因子で表すことができるように、因子分析のモデルは構築される。

38

因子得点でプロット

40



因子分析ポイント

- 因子抽出法、回転方法など様々な組み合わせが存在
 - ヘイウッドケース
 - 変数の除去 (相関が高い変数)
 - 単一変数で構成されている因子がないか
 - 主因子法の選択
 - 別な回転の選択
- 因子負荷量の小さい変数の対応
- 相関の強さ: 変数の選択、直交、**斜交回転の選択**
 - [再解析の必要性] 様々なパターンを試す必要がある (ある意味正解はない)

クラスター分析

変数、対象者を分類する

45

距離の決め方 いろいろあります

1. **Ward法**:「各クラスターに属するケースの平均値を出し、その平均値から各ケースの差を求め、差を2乗したうえで、全クラスターを合算する」(平方和 指標 E)ものである。この値が最も低いものを融合の対象とする
2. **グループ間平均連結法**:ひとつのクラスター(ケース:A, B, C)ともうひとつのクラスター(ケース:D, E, F)のそれぞれからひとつのケースを選択してできる組み合わせ(AD, AE, AF, BD, BE, BF, CD, CE, CF)の距離を平均し、その値を両クラスター間の距離であるとする。
3. **グループ内平均連結法**:ひとつのクラスター(ケース:A, B, C)ともうひとつのクラスター(ケース:D, E, F)に属するすべてのケースから作る可能性のある2ケースの組み合わせ (AB, AC, AD, AE, AF, BC, BD, BE, BF, CD, CE, CF, DE, DF, EF)の距離を 平均し、その値を両クラスター間の距離であるとする。
4. **最近隣法**:ひとつのクラスター(ケース:A, B, C)ともうひとつのクラスター(ケース:D, E, F)のそれぞれからひとつのケースを選択してできる組み合わせのすべて (A D, AE, AF, BD, BE, BF, CD, CE, CF)の距離のうち最も短いものをもって、両クラスター間の距離であるとする。
5. **最速隣法**:ひとつのクラスター(ケース:A, B, C)ともうひとつのクラスター(ケース:D, E, F)のそれぞれからひとつのケースを選択してできる組み合わせのすべて (AD, AE, AF, BD, BE, BF, CD, CE, CF)の距離のうち最も遠いものをもって、両クラスター間の距離であるとする。
6. **重心法**:ひとつのクラスターについて、ケース間の距離の測定に用いる複数の変数の平均で クラスターの座標を求め、これをそのクラスターの重心とする。クラスターを構成するケース数で重み付けを行ったうえでクラスターの重心間の距離を求め、これが最も短いクラスター群を融合させる。

47

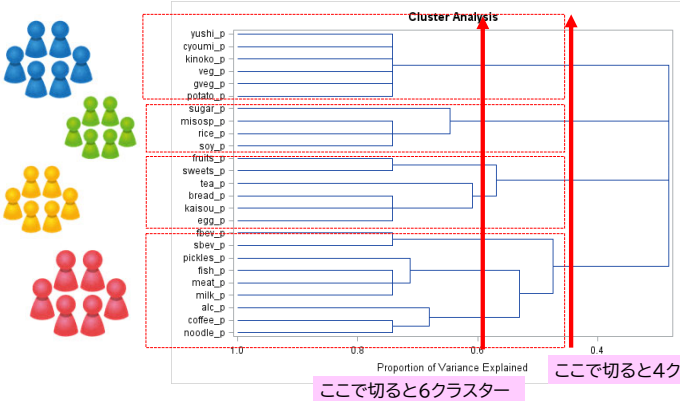
クラスター分析

教師なし分類(クラスタリング)

- 変数のクラスタリング
数値変数の集合を不連続あるいは階層的なクラスタに分割する
- k-means法**による個体のクラスタリング ……あらかじめ分類するクラスター数を決定
1つ以上の量的変数から計算された距離に基づいて、不連続なクラスター分析を行う
- CLUSTER 距離に基づく階層的クラスタリング**
単一連結, 完全連結, 平均連結, ウォード, セントロイド, 密度

46

“変数”のクラスタリング



48

クラスター分析 階層的 対象者を分類

COMPLETE Linkage

Average Linkage

数字は対象者

2つのクラスター間の距離は、1つのクラスターの観測値と他のクラスターの観測値の間の最大距離である。

2つのクラスター間の距離は、各クラスターに1つずつあるオブザベーションのペアの間の平均距離である。

49

階層的クラスタリング

51

例:チョコレート・キャンディーデータ

データセット
Candy Bars.csv
Candy Bars.sav

各社のチョコレートキャンディ(75銘柄)の栄養素含有量のデータから、その特徴によりいくつかのクラスターに分けたい

- 階層的クラスター分析
- K-means法



50

①階層クラスタリング Ward法

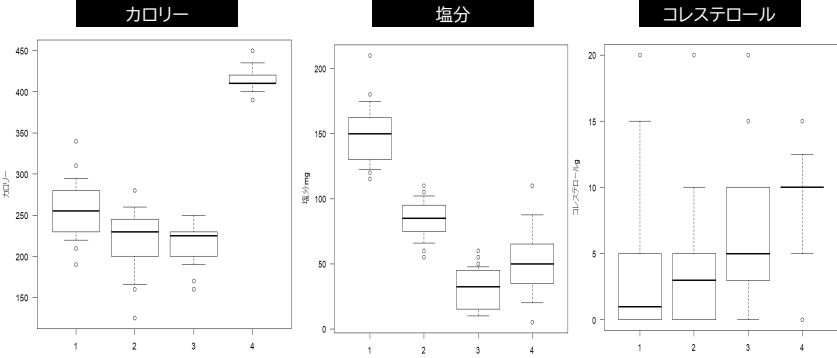
ブランド名	名前	サービング /pkg	オンス /pkg	カロリー	総脂肪(g)	飽和脂肪(g)	コレステロール(g)	塩分(mg)	炭水化物(g)	食物繊維(g)	糖分(g)	タンパク質(g)
M&M/Mars	Snickers Peanut Butter	1	2	310	20	7	5	150	28	1	23	6
Hershey	Cookies 'n' Mint	1	1.55	230	12	6	10	80	27	1	21	4
Hershey	Cadbury Dairy Milk	3.5	5	220	12	8	10	45	24	1	21	3
M&M/Mars	Snickers	3	3.7	170	8	3	5	85	21	1	17	3
Charms	Sugar Daddy	1	1.7	200	2.5	2.5	2	100	43	0	28	1
M&M/Mars	Twix Peanut Butter	1	1.71	260	16	5	5	130	26	2	17	5
Hershey	Twizzler	1	2.2	190	1.5	0	0	150	42	0	24	2
Tobler	Toblerone	1	1.23	190	11	7	5	25	21	0	19	2
Nestle	Crunch	1	1.55	230	12	7	5	60	28	1	23	3
Hershey	Almond Joy	2	3.22	230	13	8	2	85	25	3	17	2
Sherwood	Elana Mint	1	1.6	200	10	6	15	10	29	2	26	2
Hershey	Krackel	1	2.6	390	21	13	10	110	45	1	35	5

4~11列を使用

52

クラスターの要約(クラスター1~4)

INDICES: 1	カロリー	コレステロールg	タンパク質g	塩分mg	食物繊維g	炭水化物g	糖分g
	256.3750	4.5000	4.4375	149.3125	0.9375	31.6875	23.0000
INDICES: 2	カロリー	コレステロールg	タンパク質g	塩分mg	食物繊維g	炭水化物g	糖分g
	220.555556	4.333333	3.296296	84.444444	1.000000	30.333333	22.370370
INDICES: 3	カロリー	コレステロールg	タンパク質g	塩分mg	食物繊維g	炭水化物g	糖分g
	218.4615385	5.8461538	2.5384615	30.8076923	0.9230769	28.6923077	23.3461538
INDICES: 4	カロリー	コレステロールg	タンパク質g	塩分mg	食物繊維g	炭水化物g	糖分g
	415.000000	9.166667	6.666667	52.500000	2.333333	41.000000	31.666667

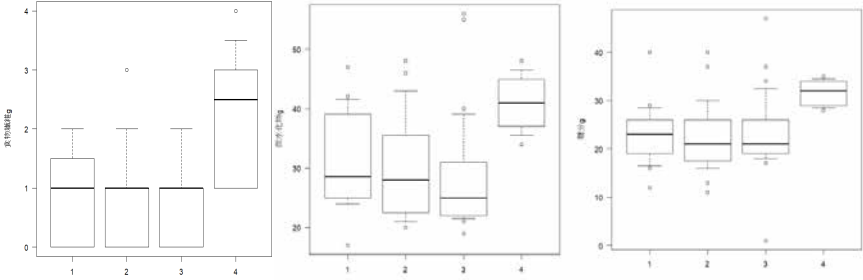


クラスター1~4

Factor	hclus: label	2	3	4	p-value
	n	16	27	26	6
カロリー	256.38 (37.41)	220.56 (37.48)	218.46 (24.77)	415.00 (19.75)	<0.001
コレステロールg	4.50 (6.73)	4.33 (4.80)	5.85 (5.08)	9.17 (4.92)	0.212
タンパク質g	4.44 (2.13)	3.30 (1.88)	2.54 (1.30)	6.67 (2.25)	<0.001
塩分mg	149.31 (24.45)	84.44 (14.83)	30.91 (14.96)	52.50 (34.89)	<0.001
食物繊維g	0.94 (0.77)	1.00 (0.85)	0.92 (0.85)	2.33 (1.21)	0.003
炭水化物g	31.69 (8.46)	30.33 (8.45)	28.69 (9.54)	41.00 (5.44)	0.024
糖分g	23.00 (6.46)	22.37 (6.71)	23.35 (8.29)	31.67 (2.80)	0.040

どのクラスター間で差があるか？
余裕があれば、多重比較で比較してみよう

食物繊維 炭水化物 糖分



K-means法
[標準メニュー]-[統計量]-[次元解析]
-[クラスター分析]-[k-平均クラスタ分析]

R Rコマンダー

ファイル編集アクティブデータセット統計解析グラフと表ツールヘルプ標準メニュー

データセット: Dataset編集表示保存

スクリプトRマークダウン

統計量

要約

クラフ

モザル

分布

ツール

ヘルプ

次元解析

尺度の信頼性...

主成分分析...

因子分析...

検証的因子分析...

クラスター分析

k-平均クラスタ分析...

階層的クラスタ分析...

階層的クラスタリングの要約...

階層的クラスタリングの結果をデータセットに保存...

```
boxdata$stats[,1:4] <- quantile(boxdata$stats[,complete.cases(boxdata$stats[,1:4]) <- quantile(boxdata$stats[,complete.cases(boxdata$outliers <- Dataset[is.na(boxdata$塩分mg) & Dataset$塩分mg>boxdata$stats[5,4]] boxdata$out <- c(boxdata$out, boxdata$outliers) boxdata$group <- c(boxdata$group, rep(4, length(boxdata$outliers))) xplot(boxdata, ylab="塩分mg") remove(boxdata, outliers) remove(boxdata, outliers) #####アクティブデータセットをエクスポートする(Text形式)##### write.table(boxdata, file=paste0("New_遊園地大/講義関連/京都府庁", library(readdata16), pos=20) #####アクティブデータセットをエクスポートする(Stata形式)##### #####アクティブデータセットをエクスポートする(Text形式)#####
```

出力

> boxdata\$outliers <- Dataset[is.na(boxdata\$塩分mg) & Dataset\$塩分mg>boxdata\$stats[5,4]] boxdata\$outliers <- Dataset[is.na(boxdata\$塩分mg) & Dataset\$塩分mg>boxdata\$stats[5,4]]

65

K-means法

	new.x. カロリー	new.x. コレステロールg	new.x. タンパク質g	new.x. 塩分mg	new.x. 食物繊維g	new.x. 炭水化物g	new.x. 糖分g
1	211.7308	5.653846	2.538462	30.61538	0.9230769	28.65385	23.07692
2	259.0526	4.210526	4.210526	141.26316	0.9473684	33.36842	24.78947
3	221.2500	4.583333	3.333333	82.91667	1.0000000	28.87500	21.16667
4	415.0000	9.166667	6.666667	52.50000	2.3333333	41.00000	31.66667

67

データ

オプション

カロリー、コレステロール、塩分、炭水化物、食物繊維、糖分

データ

オプション

変数(1つ以上選択)

塩分mg

食物繊維g

炭水化物g

糖分mg

部分集合の表現

<全ての有効なケース>

オプション

クラスター数: 4

シード初期値の数: 10

最大繰り返し数: 10

クラスを保存する変数: KMeans

☒ クラスのサマリの表示

☒ クラスのイプロット

☒ データセットにクラスを割り当て

オプション選択

クラスター4つ

66

k-means法による

4つのクラスター

カロリー

塩分

68

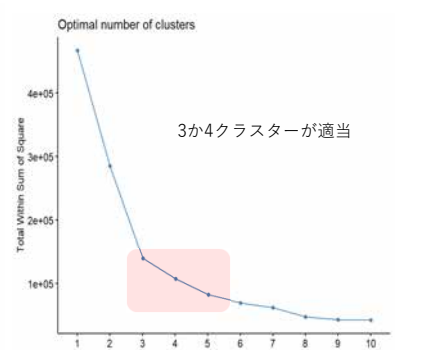
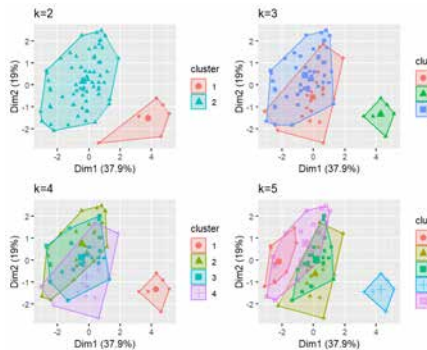
17

チョコバーの分類(k-meansで)

1		2		3		4	
Hershey	Cadbury Dairy Milk	M&M/Mars	Snickers Peanut Butter	Hershey	Rolo	Hershey	Krackel
M&M/Mars	M&Ms Peanut	M&M/Mars	Twix Peanut Butter	Annabelle	Abba-Zabba	Hershey	Mr. Goodbar
Sherrwood	Elena Mocca	Hershey	Twizler	Hershey	Cookies 'n' Mint	Hershey	Golden Collection
Just Born	Super Hot Tamales	Hershey	Twix Caramel	Nestle	Crunch	Hershey	KiKat
Myerson	Big Cherry	Pearson	Peanut Nut Roll	Hershey	Almond Joy	Hershey	Special Dark
Annabelle	U-No (Green)	Brown & Haley	Almond Roca	M&M/Mars	Mars	Hershey	Milk Chocolate Almond
M&M/Mars	Skittles	Leaf	Payday	Hershey	Bar None		
M&M/Mars	Dove	Nestle	Butterfinger	Hershey	Reese's Peanut Butter Cup Crunchy		
Annabelle	U-No (Blue)	Hershey	Reese's Peanut Butter Cup	Hershey	Symphony (Blue)		
Tootsie	Charleston Chew	Nestle	Baby Ruth	Hershey	Mound		
Hershey	Milk Chocolate	M&M/Mars	3 Musketeers	Hershey	Cadbury Roast Almond		
M&M/Mars	M&Ms Plain	Leaf	Whoppers	M&M/Mars	Milky Way Dark		
Hershey	Cadbury Fruit & Nut	Hershey	5th Avenue	Nabisco	Planters Original Peanut Bar		
Hershey	Kisses	Standard	Peanut Butter GooGoo	M&M/Mars	M&Ms Peanut Butter		
Hershey	Symphony (Red)	M&M/Mars	Snickers Munch	Hershey	Reese's Pieces		
Weider	Tiger Sport	M&M/Mars	Milky Way	Hershey	Reese's Nutrageous		
Hershey	Cadbury Caramello	M&M/Mars	Whatchamacallit	M&M/Mars	Snickers		
Tabler	Tablerone	Leaf	Heath	Charms	Sugar Daddy		
Sherrwood	Elena Mint	Annabelle	Big HunK	Bit-O-Honey	Bit-O-Honey		
Nestle	Raisinet			Nestle	100 Grand		
Adams & Bro Cup O Gold				Hershey	Skor		
Hershey	York Peppermint Patty			M&M/Mars	Milky Way Lite		
Tootsie	Jr Mints			Weider	Tiger Milk		
Hershey	Kisses Almond			Annabelle	Look!		
M&M/Mars	M&Ms Almond						
Nestle	Chunky						

69

クラスター2～5



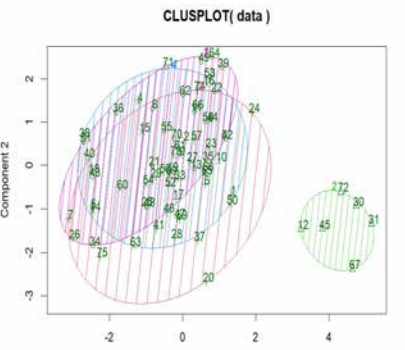
71

Rでの出力例: k-means法による

```
##k-means
km<-kmeans(data,4)
result2 <- km$cluster
write.csv(result2,"result2.csv")

#plot
library(cluster)
clusplot(data, km$cluster, color=TRUE,
shade=TRUE, labels=2, lines=0)





#data出力
table2 <- table(answer, result2)
write.csv(table2,"table2.csv")
```








These two components explain 56.88 % of the point variability.


70

階層(Wards)法とK-means法

	k-means						
	カロリー	コレステロ	タンパク質	塩分mg	食物繊維g	炭水化物g	糖分g
	1 211.73	5.65	2.54	30.62	0.92	28.65	23.08
	2 259.05	4.42	4.21	141.26	0.95	33.37	24.79
	3 221.25	4.58	3.33	82.92	1.00	28.88	21.17
	4 415.00	9.17	6.67	52.50	2.33	41.00	31.67



	階層的						
	カロリー	コレステロ	タンパク質	塩分mg	食物繊維g	炭水化物g	糖分g
	1 256.38	4.50	4.44	149.31	0.94	31.69	23.00
	2 220.56	4.33	3.30	84.44	1.00	30.33	22.37
	3 218.46	5.85	2.54	30.81	0.92	28.69	23.35
	4 415.00	9.17	6.67	52.50	2.33	41.00	31.67



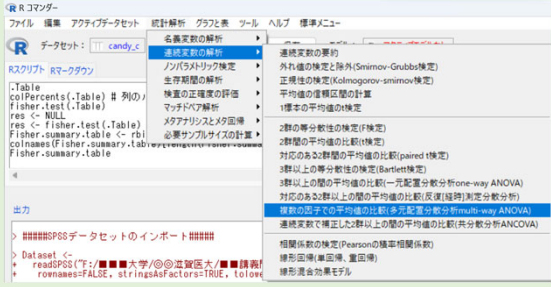
二つの方法での分類の一致度

		K-means			
		1	2	3	4
階層	1	0	16	0	0
	2	2	3	22	0
	3	24	0	2	0
	4	0	0	0	6

73

レポート課題B：因子分析、クラスター分析の例題

<B. Candyデータ>
階層クラスター、k-meansクラスターでの各クラスター別の平均値を比較してみる



「Candy_cluster_out.csv」に
・階層
・k-meansのクラスター情報
・チョコバーの属性
がまとめてある。

階層、k-meansのクラスターごとの特徴を比較してみる
→ ANOVA
箱ひげ図

75

レポート課題A：因子分析、クラスター分析の例題

<A. ビールデータ>
1. 因子分析
・因子得点(A特性,B特性)
性別、年齢別に因子得点の平均を算出
*年齢は、20代、30代、40代以上(40、50、60代をまとめる)の3カテゴリーにすること

2. 階層クラスター分析
クラスター分析により、対象者をビールの好みで区分する。抽出されたクラスターごとの特徴を確認する

3. 性・年齢別での好みのビールについて特徴をまとめる

74

第14回
解析実習 Wrap up

76

教科書など



教科書 （2） 各種手法別

(統計パッケージ関連)

<SPSS>

- 対馬栄輝, 第2版 SPSSで学ぶ医療系データ解析, 東京図書, 2016
- 対馬栄輝, 第2版 SPSSで学ぶ多変量医療系データ解析, 東京図書, 2016

<SAS>

- 臨床評価研究会(ACE)基礎解析分科会 著, 実用SAS 生物統計ハンドブック, サイエントリスト社
- 大橋渉, 統計を知らない人のためのSAS入門 , オーム社, 2012

<R>

笹刈 裕介, 大野 幸子, 橋本 洋平, 石丸 美穂, 超入門！すべての医療従事者のためのRstudioではじめる医療統計, 金芳堂, 2021




<EZ R>

- 神田善伸, EZRでやさしく学ぶ統計学 改訂3版, 中外医学社, 2020
- 新谷歩, みんなの医療統計 12日間で基礎理論とEZ Rを完全マスター！, 講談社, 2016

(疫学研究)

スティープン B. ハリー/スティーブン R. カミングス著, 木原雅子, 木原正博訳, 医学的研究のデザイン 第4版, メディカルサイエンスインターナショナル, 2014

Szklo Moyses, Nieto, F. Javier著アドバンスト分析疫学 369の図表で読み解く疫学的推論の論理と数理, メディカルサイエンスインターナショナル, 2020



教科書 （1） 統計学、生物統計学一般

- (統計学一般)

 - 東京大学教養学部統計学教室, 統計学入門, 東京大学出版会 1991
 - 日本統計学会編, 日本統計学会公式認定 統計検定2級対応 統計学基礎, 東京図書, 2019
 - 江崎貴裕, 分析者のためのデータ解釈学入門, ソシム, 2020
 - 阿部真人, データ分析に必須の知識・考え方 統計学入門, ソシム, 2021

(生物統計学)

 - Altman DG: Practical Statistics for Medical Research, Chapman and Hall, 1991. (佐久間昭監訳: 「医学研究における実用統計学」、サイエントリスト社, 1999)
 - Armitage P and Berry G: Statistical Methods in Medical Research, 3rd ed., Blackwell, 1994. (椿美智子・椿広計共訳: 「医学研究のための統計的方法」、サイエントリスト社, 2001)
 - 丹後俊郎: 新版医学への統計学, 朝倉書店, 1993
 - 浜田知久馬: 学会・論文発表のための統計学新版, 真興交易医書, 2012
 - 中村好一編, 論文を正しく読み書くためのやさしい統計学 改訂第3版, 診断と治療社, 2019

