データマイニング
～お客様の嗜好はこう読み取れ～

諏訪東京理科大学　講師
櫻井　哲朗
sakurai@rs.tus.ac.jp

発表の概要

ここでは

　アンケートからお客様の嗜好を読み取る方法

を解説します

主に次の方法について解説します
- クロス集計
- 多変量解析
  - クラスター分析・因子分析・回帰分析
- テキストマイニング

発表の流れ

- アンケートその1：健康意識に関するアンケート
  - クロス集計による分析
- アンケートその2：店舗利用に関するアンケート
  - クラスター分析による分類
  - 因子分析による把握
  - 回帰分析による予測
- アンケートその3：海物語に関するアンケート
  - 自由記述に関する分析

アンケートその1
健康意識に関するアンケート

今回のデータ分析の目的

アンケートから

お客様の健康志向を読み取り
効果的な景品を探る

データを分析するにあたって

分析するにあたって
・性別が不明な項目は無視しました
・年齢が不明な項目は無視しました
・Q1において⑤と回答した項目は無視しました
・都道府県が不明な項目は無視しました

Q1において
男女間の違い

性別	非常にある	それなりにある	余り興味はない	なし
男性	15.59%	64.97%	15.92%	3.51%
女性	14.24%	69.52%	14.54%	1.69%

p値：0.006248

違いをハッキリさせるために単純化する
興味ある＝「非常にある」または「それなりにある」
興味なし＝「余り興味はない」または「なし」

性別	興味ある	興味ない
男性	80.56%	19.44%
女性	83.76%	16.24%

p値：0.02394

これより、次のことがわかった
・男性と女性の健康への関心は、大きな違いはない
（少しの違いはあるが）

Q1において
年代の違い

	非常にある	それなりにある	余り興味はない	なし
20代以下	22.77%	48.02%	20.79%	8.42%
30代	13.10%	60.29%	22.06%	4.55%
40代	11.95%	68.22%	16.81%	3.02%
50代	15.00%	70.00%	12.50%	2.50%
60代以上	24.91%	67.45%	7.27%	0.36%

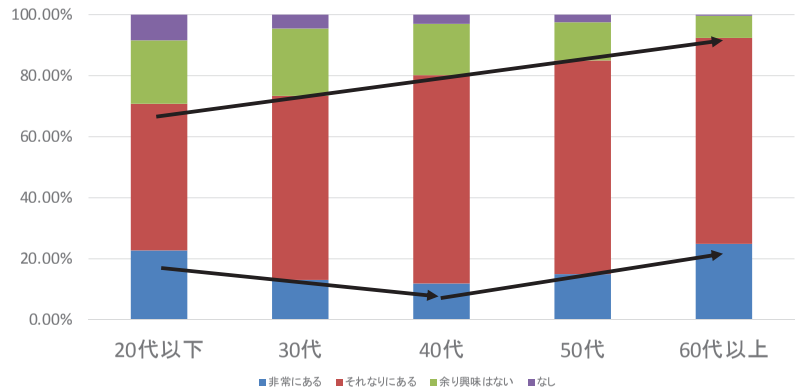
単純化

	興味ある	興味ない
20代以下	70.79%	29.21%
30代	73.40%	26.60%
40代	80.17%	19.83%
50代	85.00%	15.00%
60代以上	92.36%	7.64%

これより、次のことがわかった
・「非常にある」と答える比率は、
　20代から40代で減少し、40代から60代で増加する
・「それにある」と答える比率は、年齢とともに増加する
・「興味ある」、「興味ない」の2分割なら、年齢とともに「興味ある」が増加する

結論：健康への関心は年齢とともに増加する

Q1において
年代の違い



下の矢印は「非常にある」の変化
上の矢印は「非常にある」＋「それなりにある」の変化

Q1において
性別で分けたときの年代の違い

男性	興味ない	興味ある
20代以下	70.73%	29.27%
30代	73.48%	26.52%
40代	79.08%	20.92%
50代	84.59%	15.41%
60代以上	92.48%	7.52%

女性	興味ない	興味ある
20代以下	69.44%	30.56%
30代	73.10%	26.90%
40代	84.23%	15.77%
50代	86.17%	13.83%
60代以上	92.02%	7.98%

これより、次のことがわかった
男性と女性の年代別の健康への関心は多少の違いはあるが

男女ともに健康への関心は年齢とともに増加する

Q1において
地域による違い

ここでは、地域による違いをみるため、とりあえず東京に注目した

	非常にある	それなりにある	余り興味はない	なし
東京都	15.41%	68.10%	14.16%	2.33%
それ以外	15.18%	65.82%	15.80%	3.20%

これより、次のことがわかった
東京都とそれ以外において健康への関心は多少の違いはあるが

健康への関心は東京とそれ以外で違いはない

Q２において
男女による違い

性別	血圧	認知症（ボケ）	体重・体脂肪	喫煙（禁煙）	お酒の量	ストレス
男性	12.37%	6.11%	20.86%	6.94%	5.17%	12.44%
女性	9.76%	8.99%	20.29%	5.71%	2.54%	13.96%

性別	骨粗鬆症	糖尿病	痛風	癌	脳の病気	心臓の病気	その他
男性	1.52%	9.29%	4.35%	9.36%	5.60%	5.98%	4.78%
女性	5.47%	7.41%	1.75%	10.45%	7.43%	6.24%	4.07%

このとき、男性と女性で2%以上異なった項目は次の項目であった
　男性が多い：血圧、お酒の量、糖尿病、痛風
　女性が多い：認知症（ボケ）、骨粗鬆症

これより、次のような仮説を考えることができる
　男性は、お酒の量を気にする傾向が高いため
　生活習慣に関する項目に関心がある
　女性は、具体的な病気に関する項目に関心がある、
　とくに骨粗鬆症に関心があるのは女性がなりやすいためと推測される

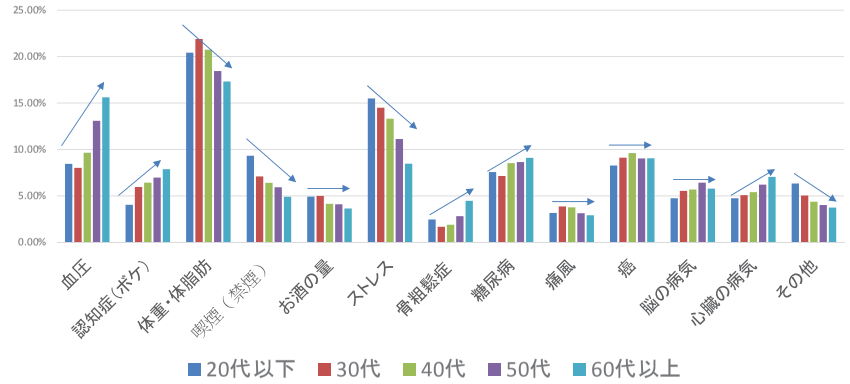
Q２において
年代による違い

	血圧	認知症（ボケ）	体重・体脂肪	喫煙（禁煙）	お酒の量	ストレス
20代以下	8.45%	4.05%	20.42%	9.33%	4.93%	15.49%
30代	8.03%	5.97%	21.90%	7.10%	5.00%	14.50%
40代	9.65%	6.43%	20.75%	6.41%	4.15%	13.31%
50代	13.09%	6.98%	18.45%	5.93%	4.10%	11.13%
60代以上	15.62%	7.88%	17.32%	4.91%	3.65%	8.47%

	骨粗鬆症	糖尿病	痛風	癌	脳の病気	心臓の病気	その他
20代以下	2.46%	7.57%	3.17%	8.27%	4.75%	4.75%	6.34%
30代	1.68%	7.15%	3.87%	9.12%	5.55%	5.09%	5.04%
40代	1.90%	8.53%	3.77%	9.62%	5.69%	5.40%	4.38%
50代	2.82%	8.65%	3.14%	9.04%	6.43%	6.22%	4.02%
60代以上	4.48%	9.10%	2.92%	9.05%	5.79%	7.06%	3.75%

これより、年代とともに3％以上の変化がある項目は次があげられる
　おおよそ年代とともに増加：血圧、認知症
　おおよそ年代とともに減少：体重、喫煙、ストレス

Q２において
年代による違い



主に、増加している項目は病気に関する項目が増加傾向にあるが
しかし、喫煙やストレスに関しては減少傾向
痛風、癌、脳の病気はあまり変化が見られない

Q３において
Q２との関連性

ここでは、Q2の健康について気になることをうけて
Q3の健康のために行っていることとの関連性について調べた

	食事に気を使う	体操、運動	脳の体操（脳トレ）	サプリメントの摂取	睡眠時間	マッサージ	健康グッズを使う	人間ドック・健康診断	特定保健用食品（トクホ）の摂取	ストレス発散	熱中症対策	体を温める
血圧	59.53%	37.43%	13.30%	21.03%	32.49%	15.96%	5.51%	27.23%	12.35%	32.49%	13.17%	5.89%
認知症（ボケ）	53.47%	36.82%	19.10%	24.12%	31.80%	18.25%	7.26%	29.35%	14.73%	37.35%	17.29%	8.75%
体重・体脂肪	54.40%	35.73%	10.55%	20.31%	28.71%	16.14%	6.02%	23.30%	11.51%	30.03%	11.19%	5.34%
喫煙（禁煙）	49.66%	33.00%	12.08%	20.25%	34.23%	16.78%	7.38%	25.73%	11.52%	33.22%	13.87%	6.26%
お酒の量	55.48%	40.86%	13.12%	24.75%	39.87%	19.93%	7.14%	30.07%	13.46%	33.06%	16.11%	7.48%
ストレス	51.78%	33.51%	11.63%	21.13%	34.54%	19.87%	6.01%	23.08%	12.89%	42.27%	14.32%	7.33%
骨粗鬆症	63.94%	39.44%	27.04%	33.80%	39.44%	28.45%	15.77%	36.34%	21.41%	47.61%	27.89%	15.77%
糖尿病	58.84%	34.93%	12.54%	22.39%	33.33%	17.17%	7.66%	29.55%	15.99%	34.93%	14.23%	6.99%
痛風	56.71%	38.82%	15.45%	23.58%	38.62%	20.53%	10.98%	32.93%	19.11%	38.62%	17.89%	8.33%
癌	53.32%	34.63%	14.11%	23.65%	33.71%	17.62%	7.09%	31.81%	14.19%	37.99%	15.87%	8.54%
脳の病気	56.01%	36.54%	18.51%	26.08%	36.06%	21.39%	9.74%	34.01%	18.15%	40.99%	19.71%	10.46%
心臓の病気	58.81%	37.42%	17.98%	23.33%	36.45%	19.44%	9.48%	33.54%	17.13%	41.19%	20.66%	11.18%

横のQ2の質問項目ごとに見ていくとほとんどの項目において
１番に「食事に気を使う」がきており、
２、３番が「体操、運動」、「睡眠時間」、「ストレス発散」となっている

まとめ

Q1：健康への関心において
　・男女間に大きな違いはなく
　・年齢とともに増加する
　・東京とそれ以外でわけたとき、大きな違いはなかった

Q2：健康について気になることの年代における変化
　・喫煙やストレスに関しては減少傾向
　・痛風、癌、脳の病気などの病気に関しては変化が見られない
　・それ以外の項目において増加傾向にある

Q3：健康のために行っていること
　・Q2との関連を調べたところ、ほとんどの項目において
　　健康のために行っていることは１番に「食事に気を使う」がきており、
　　２、３番が「体操、運動」、「睡眠時間」、「ストレス発散」となっている

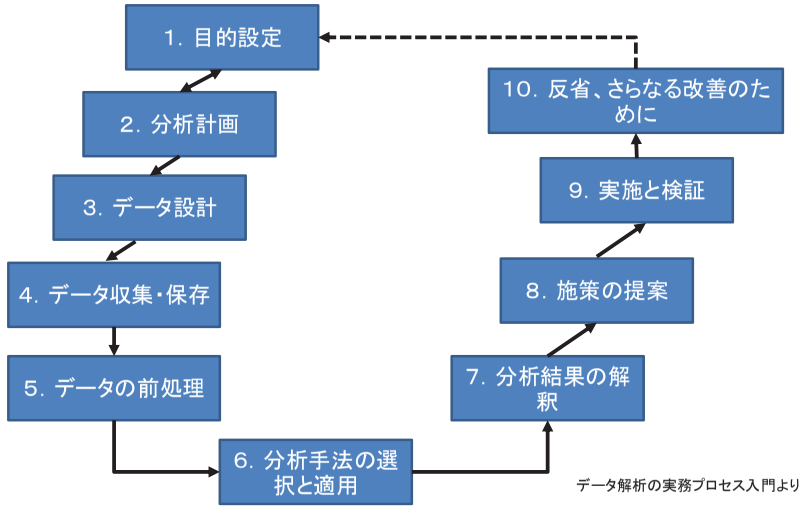
今回のデータ分析の結果

アンケート分析から
健康のために行っていることは
食事・運動・睡眠・ストレスと続くため
サプリメントや運動グッズ、快眠グッズ、
ストレス発散グッズなどの景品の充実が
提案できる

健康への関心は男女差・地域差はないが
年代における差はあるため、ご高齢のお客様が多い
店舗では比較的多め準備したほうがいいかも

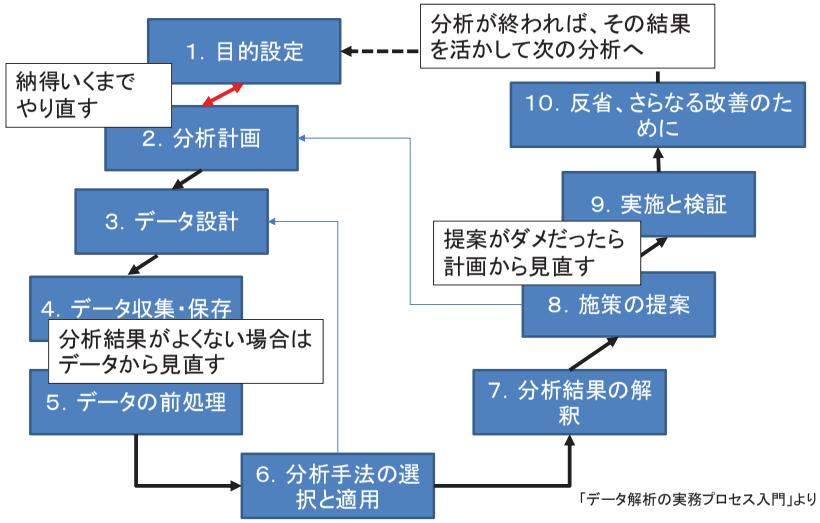
アンケートその2
店舗利用に関するアンケート

データ解析のフロー



データ解析の実務プロセス入門より

データ解析のフロー

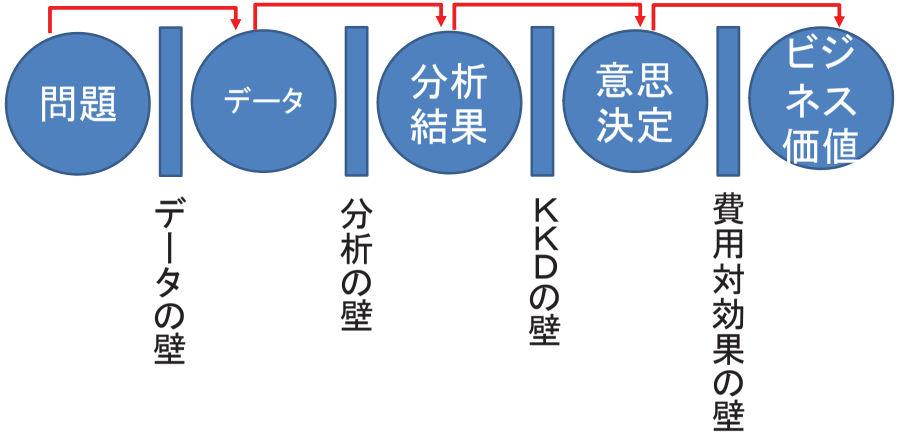


「データ解析の実務プロセス入門」より

データ解析のフロー

「会社を変える分析の力」より

データ分析の４つの壁



＊KKD＝勘と経験と度胸

「会社を変える分析の力」より

今回のデータ分析の目的

アンケートから

お客様の店舗利用に関する実態を調べ、
どのようなお客様がいるのかを調べる

そのため、次のようなアンケート項目を設定した
Q1, Q2：個人のプロファイル
Q3～Q7：店舗利用に関する項目
Q8～Q20：個人の嗜好に関する項目

データを分析するにあたって

アンケートデータ
　質問数：20問、回答数：478人

分析するにあたって
　・未回答の項目がある回答は取り除いた
　←欠損があるデータの取り扱い難しい
　　データ数が少ないときなどは
　　平均値を代入するなどの方法もある

これにより、回答数：376人に減少

遊戯傾向を知る

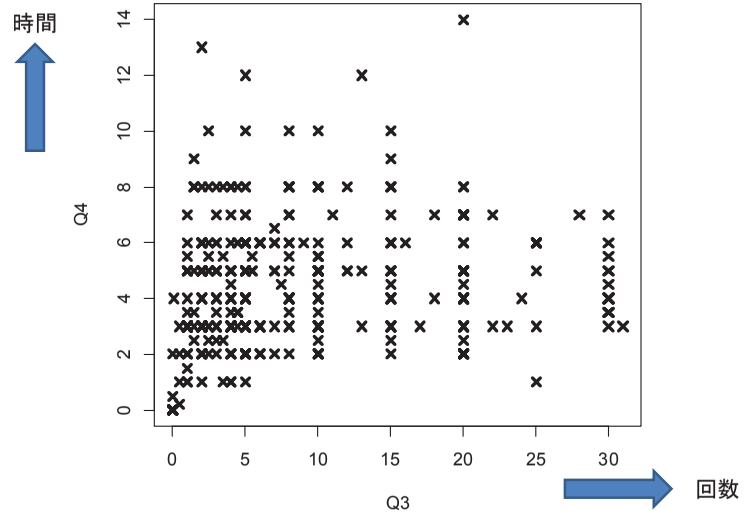
アンケートから遊戯傾向を調べるために

Q3：月の来店回数、Q4：1回の遊戯時間

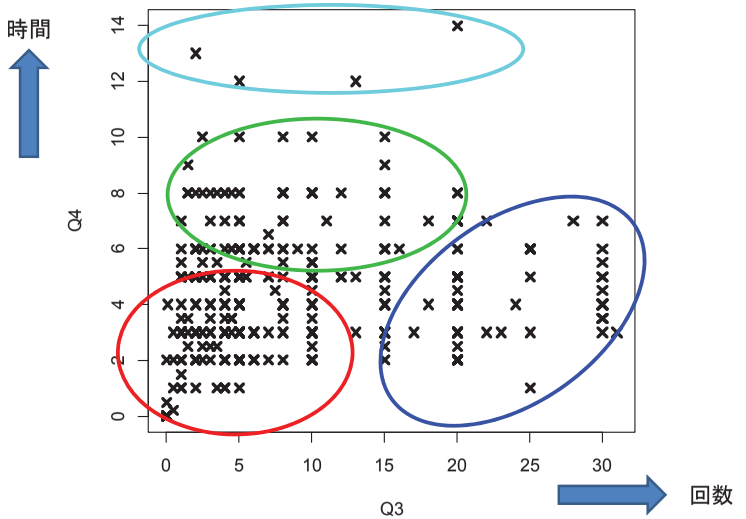
の散布図を作成する

これより、どんなお客さんがいるかを探る
- ・どれくらい来店してくるのか？
- ・どれくらい遊んでくれているのか？
- ・来店回数と時間の関係は？

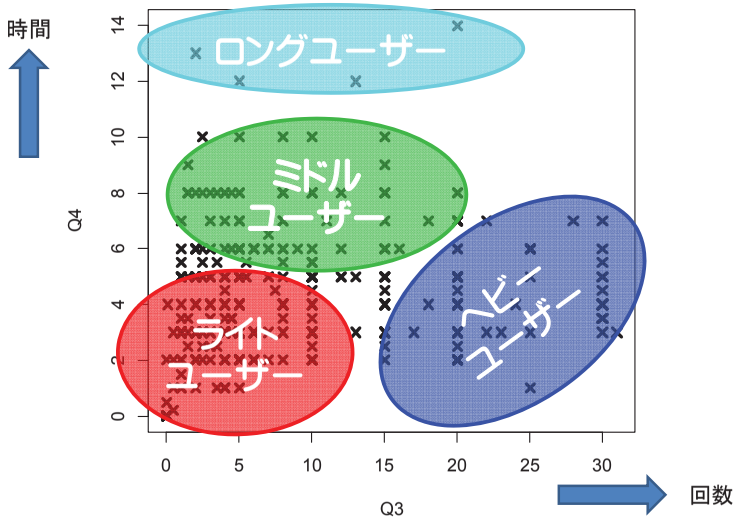
遊戯傾向を知る



遊戯傾向を知る



遊戯傾向を知る



遊戯傾向を知る

このようにグループ分けできそう

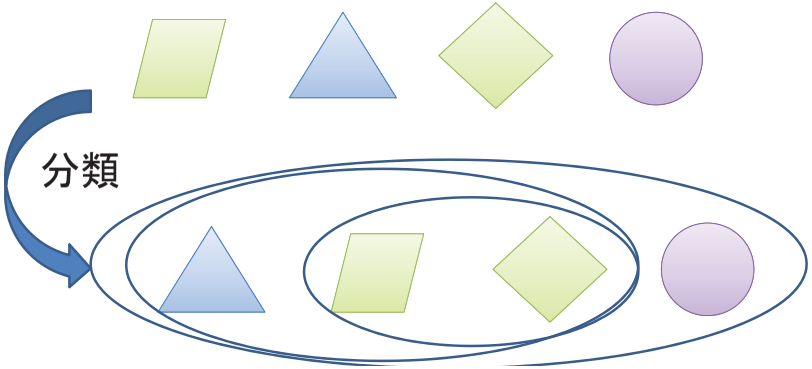
グループ分けを行う分析方法を
　クラスター分析という

クラスター分析では
　階層型と非階層型の2種類の方法があるが
ここでは、構造が把握しやすい階層型で分析してみた

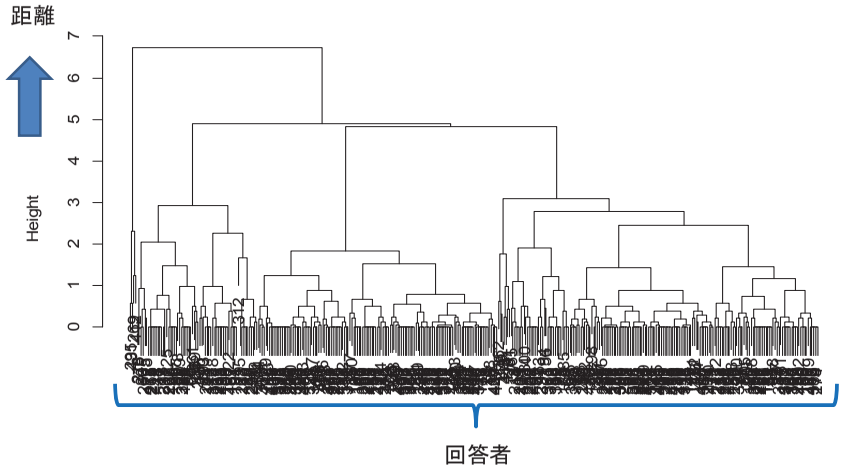
使ったデータはQ3、Q4を基準化したもの
基準化しないとQ3のデータかQ4よりも大きいのでQ3の方向ではかり分類してしまう

クラスター分析とは

クラスター分析とは、
　似ているデータをまとめ
　グループ分けを行う手法である



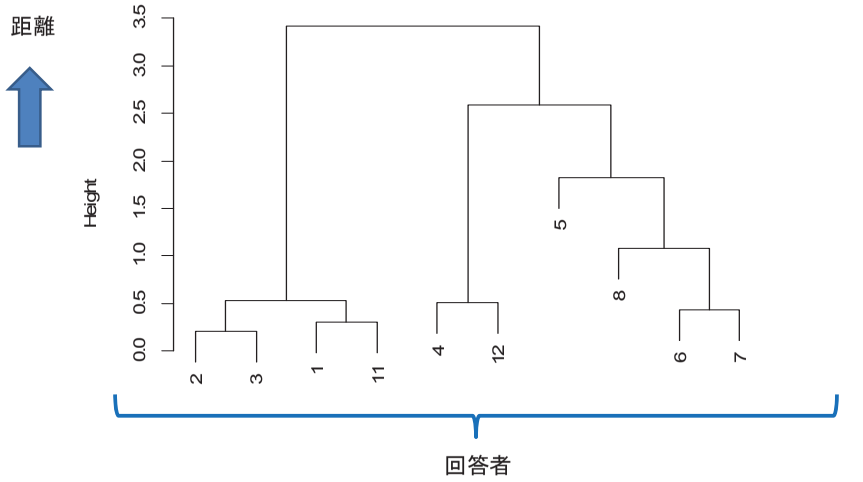
遊戯傾向を知る



＊距離＝類似度＝似ている度合い

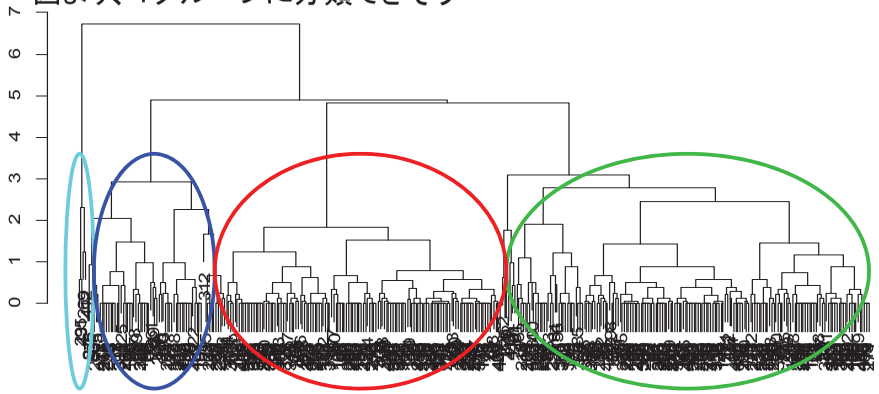
遊戯傾向を知る

先ほどの図ではつぶれてしまっているので、少ないデータで行ってみると



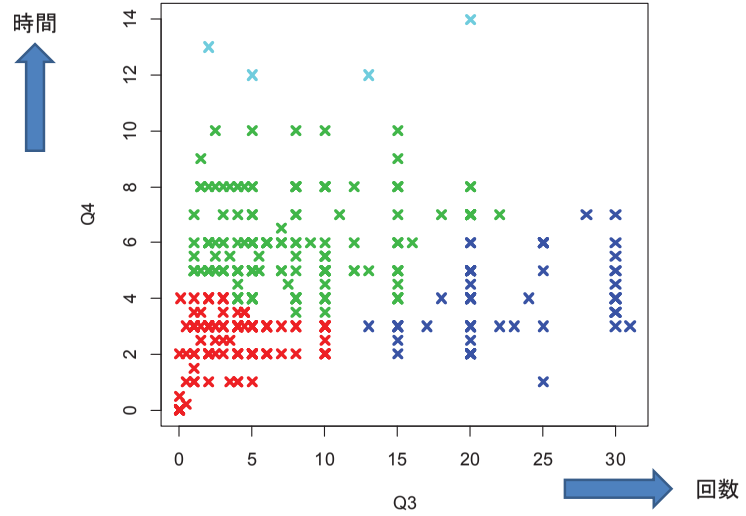
遊戯傾向を知る

もとの図に戻り、グループ分けを行う
図より、4グループに分類できそう



ここでは、グループ数のこちらで決定したがモデル選択規準などによりグループ数を決める方法もある

遊戯傾向を知る



遊戯傾向を知る

グループ分けしたもとでのQ3とQ4の平均など

Q3:回数	ライトユーザー	ミドルユーザー	ヘビーユーザー	ロングユーザー
データ数	131	175	66	4
平均	3.60	7.57	22.52	10.00
標準偏差	2.97	5.12	4.97	8.12

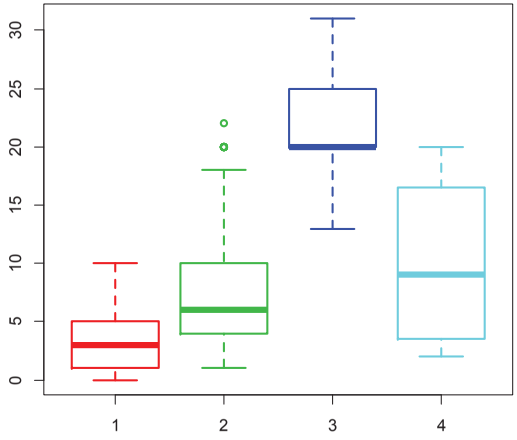
Q4:時間	ライトユーザー	ミドルユーザー	ヘビーユーザー	ロングユーザー
データ数	131	175	66	4
平均	2.37	5.90	3.92	12.75
標準偏差	1.07	1.48	1.41	0.96

平均±2倍の標準偏差の範囲に95%のデータが入る(正規分布のとき)

遊戯傾向を知る

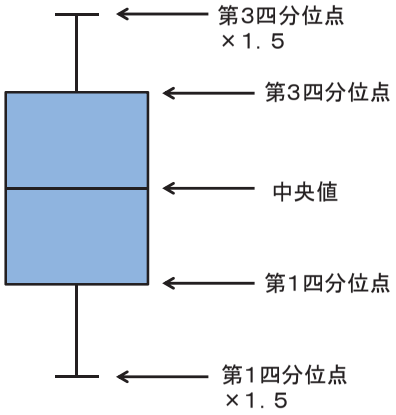
数字だとピンとこなかったらグラフで表してみる

Q3:回数



箱ひげ図

次のような図を箱ひげ図という



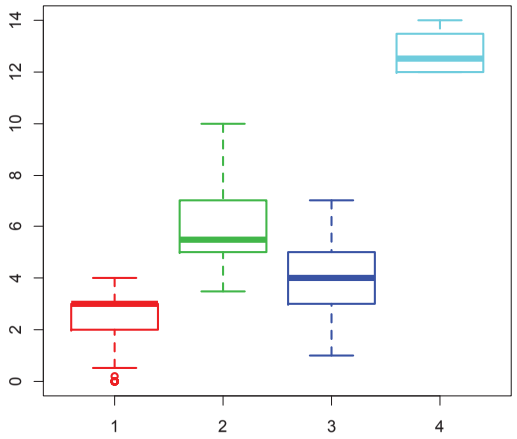
箱ひげ図は
データの分布状況を表した図

箱ひげ図を使うことにより
複数のデータの分布状況を比較できる

遊戯傾向を知る

数字だとピンとこなかったらグラフで表してみる

Q4:時間



遊戯傾向を知る

このようにグループしたら、このグループごとの
「違い」や特徴を探してみよう

たとえば、
ライトユーザーは女性が多くて
ヘビーユーザーは男性が多いのではないか
などなど

遊戯傾向を知る

実際、各タイプ別の男性と女性の違い

Q2:男女	ライトユーザー	ミドルユーザー	ヘビーユーザー	ロングユーザー	合計
男性	84	138	34	4	260
女性	47	37	32	0	116
合計	131	175	66	4	376

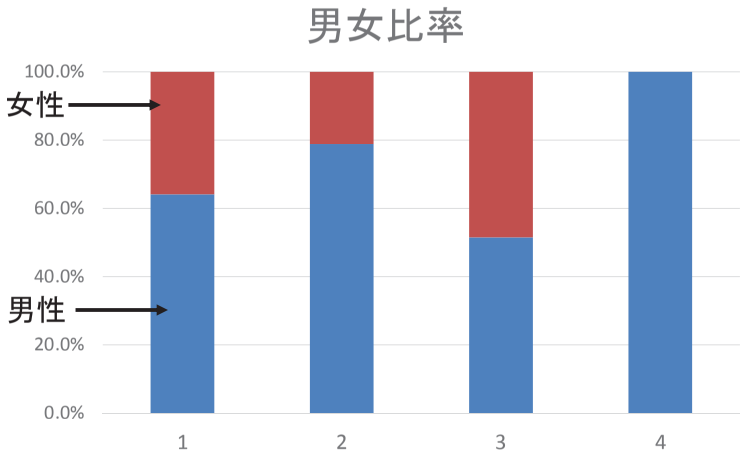
いま知りたいのは、各グループでの男性と女性の割合
が知りたいので、各グループでの合計で割る

比率	ライトユーザー	ミドルユーザー	ヘビーユーザー	ロングユーザー	合計
男性	64.1%	78.9%	51.5%	100.0%	69.1%
女性	35.9%	21.1%	48.5%	0.0%	30.9%
合計	100.0%	100.0%	100.0%	100.0%	100.0%

結論：ヘビーユーザーになるにつれて男女比が同じに
なっていく

遊戯傾向を知る

数字でピンとこなかったらグラフで表してみる



遊戯傾向を知る

他の項目における違い

Q2:年齢	ライトユーザー	ミドルユーザー	ヘビーユーザー	ロングユーザー
平均	36.57	38.93	56.83	35.75
標準偏差	14.04	15.07	15.79	19.55

Q7:お店	ライトユーザー	ミドルユーザー	ヘビーユーザー	ロングユーザー
平均	1.72	2.01	1.42	1.25
標準偏差	1.11	1.41	1.01	0.50

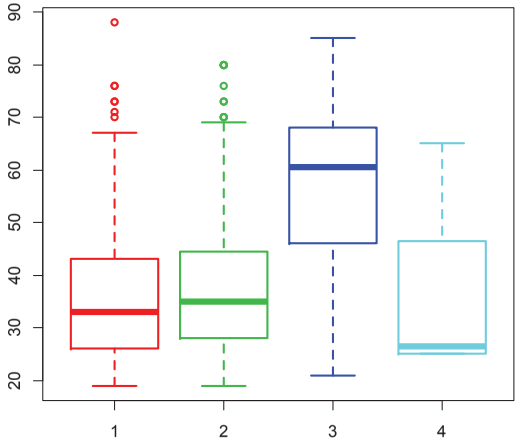
結論:
年齢はヘビーユーザーになるにつれて増加する
利用店舗数は全体を通してあまり変わらず

遊戯傾向を知る

数字でピンとこなかったらグラフで表してみる

Q2:年齢

違いが
ありそう

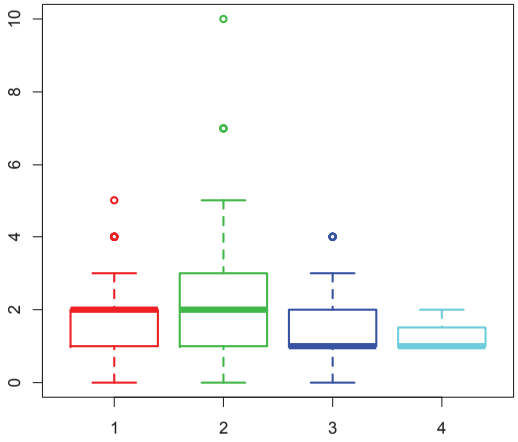


遊戯傾向を知る

数字でピンとこなかったらグラフで表してみる

Q7:お店

あまり
違いが
なさそう



遊戯傾向を知る

アンケートから個人の興味を調べるために

Q8～Q20：興味に関する質問

を使い因子分析を行う

これより、パチンコをする人が潜在的にどのようなことに興味があるかを解釈する

因子分析とはデータの背後に潜む構造を分析する手法

因子分析とは

我々は、複雑な現象を単純な原因（すなわち因子）で理解することがよくある。

たとえば
「K君はO型だからいいかげんである。」
「Lさんは、A型だから真面目である。」
という具合に複雑な人間の性格を「血液型」という1つの因子で単純に説明しようとする。

このように、複雑な現象の構造を単純な因子で分解しようとするのが因子分析である。

因子分析とは

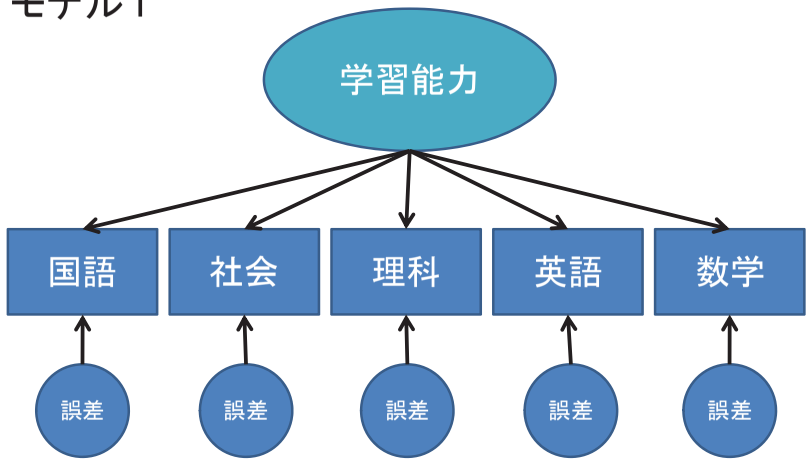
より具体的な的には、次のようなデータを解析する

	国語	社会	理科	英語	数学
Aさん	93	100	89	84	77
Bさん	100	98	89	95	86
Cさん	84	84	99	85	100
Dさん	70	73	92	66	77
Eさん	70	72	89	66	75
Fさん	66	68	95	57	82
Gさん	74	70	96	93	88
Hさん	74	75	95	70	79
Iさん	76	77	92	78	83
Jさん	79	88	100	86	100

因子分析とは

このようなデータに対して次のように構造を考えることができる

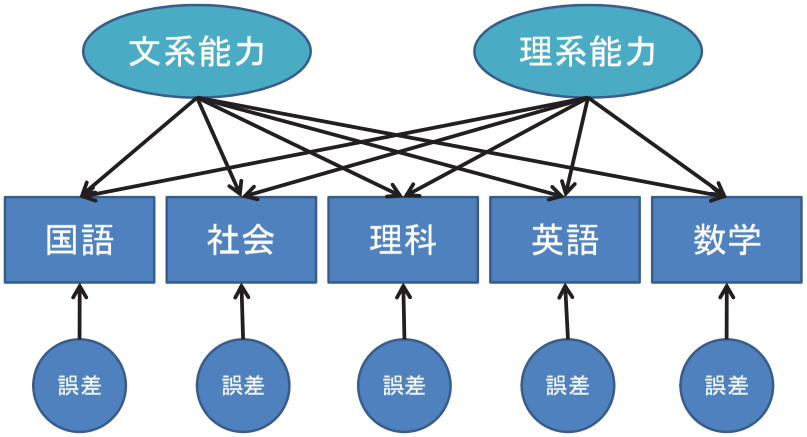
モデル1



因子分析とは

または、次のようにも考えることができる

モデル２



因子分析とは

このようにして5教科テストの得点というデータの構造を
次のように表している。

モデル１では、
学習能力という因子によって5教科テストの得点は
決定されているという構造

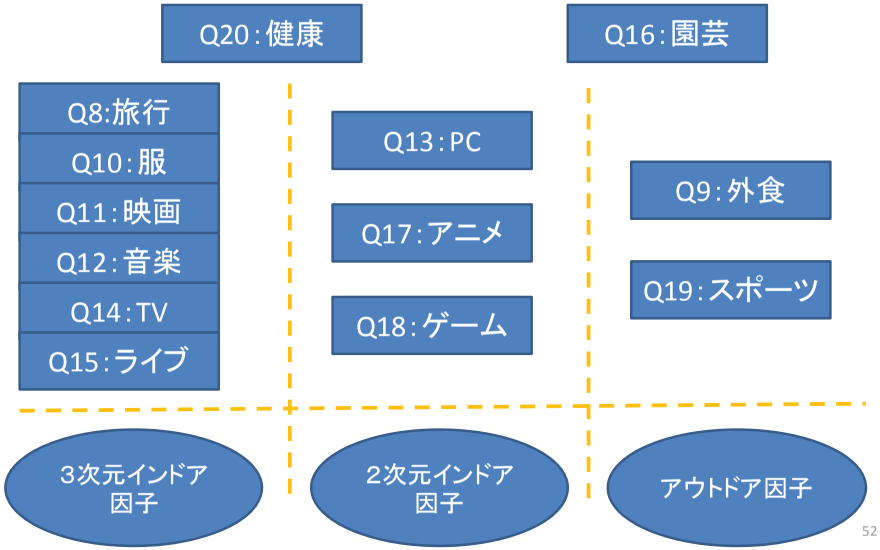
モデル２では、
文系能力と理系能力の２つの因子によって5教科テストの得点は
決定されている構造

遊戯傾向を知る

ここではQ12～Q20の興味があることの質問項目に関して
因子分析を行った

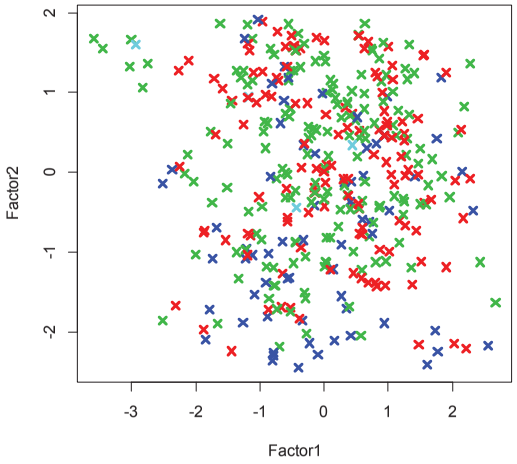
Q8:旅行
Q9:外食
Q10:服
Q11:映画
Q12:音楽
Q13:PC
Q14:TV
Q15:ライブ
Q16:園芸
Q17:アニメ
Q18:ゲーム
Q19:スポーツ
Q20:健康

遊戯傾向を知る



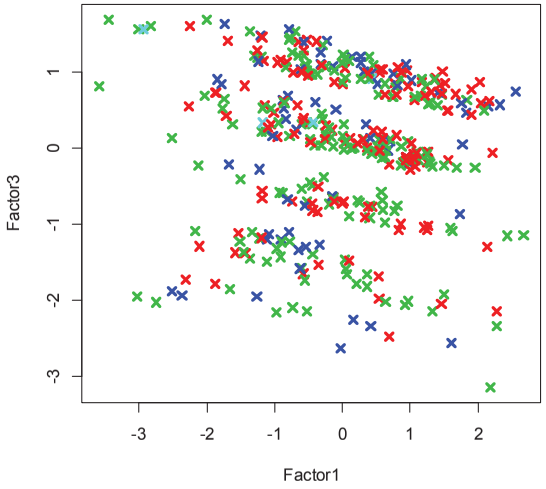
遊戯傾向を知る

因子得点をプロット：因子1と因子2



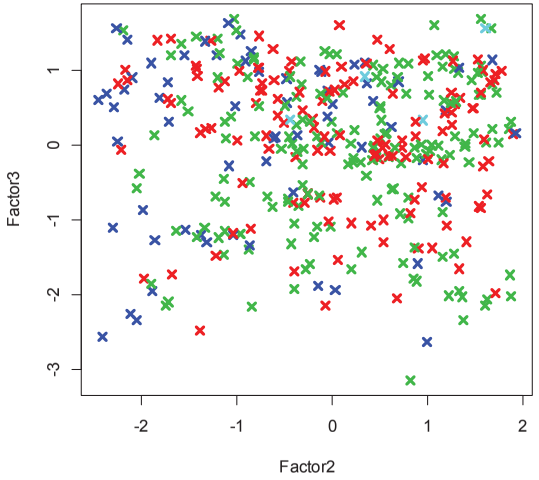
遊戯傾向を知る

因子得点をプロット：因子1と因子3



遊戯傾向を知る

因子得点をプロット：因子2と因子3



遊戯傾向を知る

アンケートから、顧客単価を調べるために

Q4：1回の利用時間
Q5：月の利用回数
Q7：よく行くお店の個数

を使い次を計算する

y＝Q4×Q5／Q7：1店舗での月の利用時間

これを予測する（このとき、まったく利用していないデータを除いた）

回帰分析

回帰分析とは、簡単にいうと
「“原因”と“結果”を結ぶための統計処理」
※実際は、因果関係までの強い関係性はいえない

出力
・目的変数
1店舗での
月の利用時間

モデル

入力
・説明変数

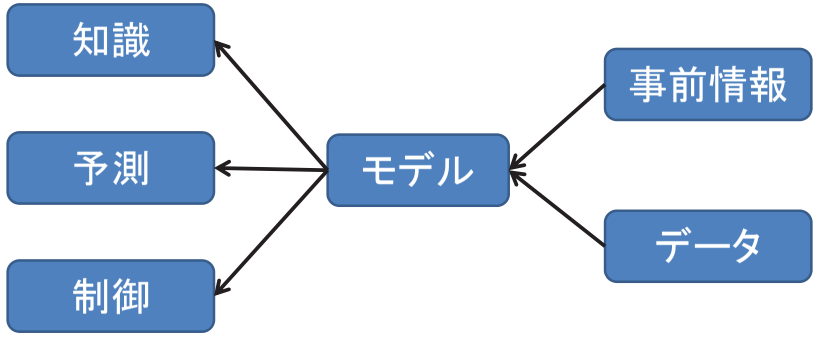
回帰分析の例

たとえば，「百薬の長」といわれるお酒も，飲み過ぎると肝障害をおこし，アルコール性脂肪肝になるといわれています．
その結果，肝硬変となり，「余命はあと何年か？」が気になります

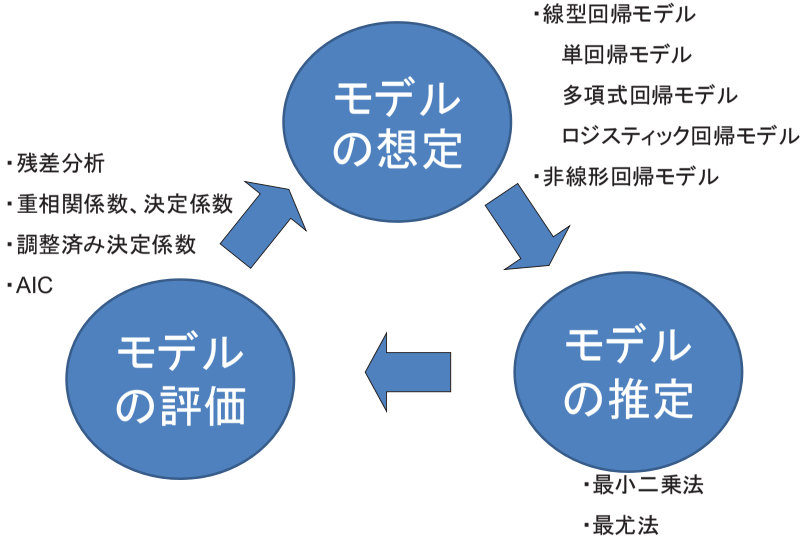
または，お酒を飲み過ぎている人に対して
ただ「お酒を控えるように」と注意してもなかなか聞いてくれません．
そこで，このまま飲み続けた場合の「余命」を教えてあげるで注意をうながす

$$\text{余命} = \beta_0 + \beta_1 \times \text{年齢} + \beta_2 \times \text{血清アルブミン} + \beta_3 \times \text{血清タンパク} + \beta_4 \times \text{血清}\gamma\text{-グロブリン} + \varepsilon$$

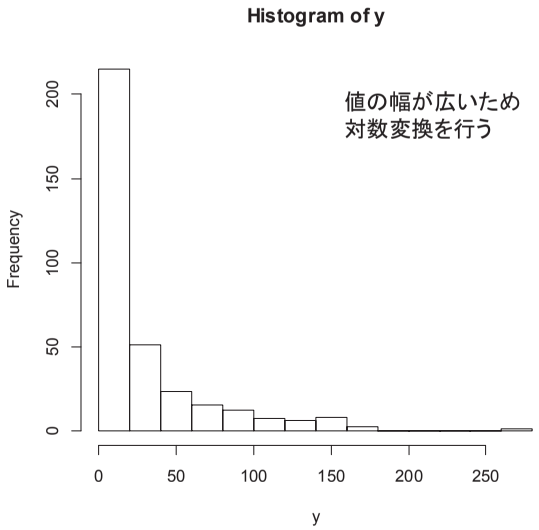
統計的モデリング



統計的モデリング

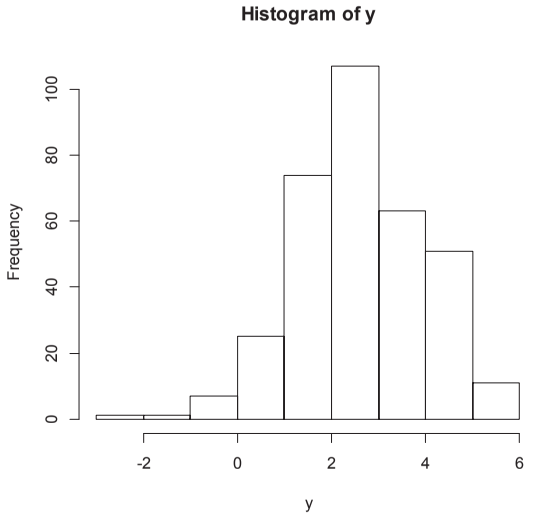


遊戯傾向を知る



値の幅が広いため
対数変換を行う

遊戯傾向を知る



遊戯傾向を知る

このとき、データからまったく利用していない人のデータを除いて解析した

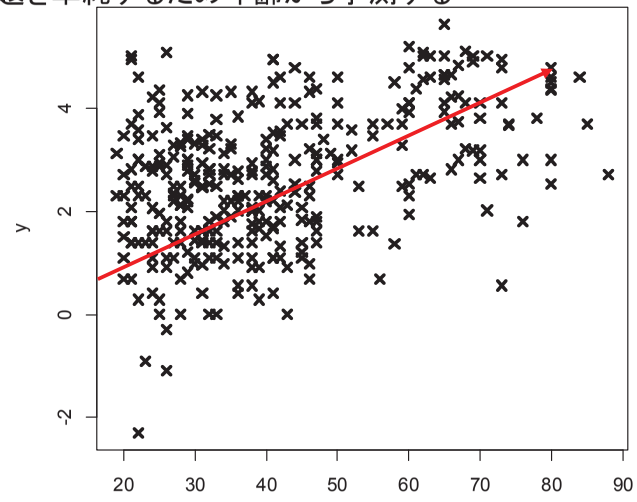
モデルは単純化して次のようなモデルを考える

利用年齢から顧客単価を予測する

このようなモデルが妥当か散布図より確認する

遊戯傾向を知る

問題を単純するため年齢から予測する



遊戯傾向を知る

これより次のように予測できる

	全体			男性だけ			女性だけ		
決定係数	0.1929			0.1201			0.3479		
年齢	予測値	下限	上限	予測値	下限	上限	予測値	下限	上限
20	6.7	5.4	8.2	7.6	5.9	9.7	4.6	3.0	6.9
30	9.5	8.1	11.1	10.3	8.7	12.3	7.1	5.2	9.8
40	13.5	11.9	15.4	14.2	12.2	16.5	11.0	8.6	14.2
50	19.3	16.7	22.2	19.4	15.9	23.6	17.2	13.6	21.7
60	27.4	22.6	33.3	26.5	20.0	35.1	26.6	20.3	34.9

結論：
年齢が増加するごとに次の利用時間は増加傾向にあり
女性だけにしぼったほうが決定係数が増加しているため
あてはまりがよくなっている

今回のデータ分析の結果

アンケート分析から
遊戯傾向として４つのグループに分類できそう
だが、グループごとの嗜好の違いは見いだせなかった
また、遊戯時間は年齢とともに増加傾向
とくに女性の方がブレが少ない

今回のデータ分析の結果

これから
ライトユーザー・ミドルユーザー向けは男性向けに
ヘビーユーザーには男女両方を意識したお店作りを

また一人の人で高時間の遊戯を期待するなら
年代高めのお店作りを心がける
その際の費用は予測値より見積もれるかもしれない

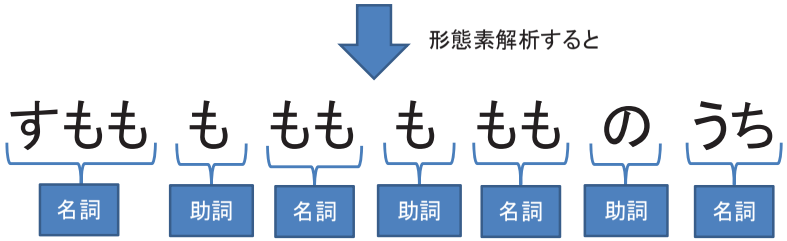
アンケートその3
海物語に関するアンケート

テキストマイニング

テキストマイニングとは、
文章を分析するための方法のこと

文章を形態素と呼ばれる最小の単位に分割して分析する

すももももももものうち



今回のデータ分析の目的

海物語に期待する事の
自由記述のアンケートから

どのようなニーズがあるかを調べる

データを分析するにあたって

アンケートデータ
　質問数：1問ただし自由記述
　回答数：2720人

分析するにあたって
　・年代が不明なの項目は取り除いた

年代別の傾向

特徴的な頻出単語として次の単語が出てきた
「演出」、「信頼」、「当たり」、「シンプル」
次のような文脈で使われている

「ごちゃごちゃした演出は無用。」
「シンプルでいいから魚群の信頼度を上げて！」
「ノーマルリーチでの告知なしの当たりを増やしてほしい。」
「リーチ演出のシンプルなところが海の魅力。」

これらの出現回数を年代別に比較する

年代別の傾向

年代別の傾向

	20代	30代	40代	50代	60代以上
演出	34	62	59	23	5
信頼	17	40	33	19	5
当たり	12	25	32	21	12
シンプル	2	62	6	39	6

各年代の合計で割って比率にする

	20代	30代	40代	50代	60代以上
演出	52%	33%	45%	23%	18%
信頼	26%	21%	25%	19%	18%
当たり	18%	13%	25%	21%	43%
シンプル	3%	33%	5%	38%	21%

年代別の傾向

結論：

20代、40代は似通っていて「演出」に関心がある
30代、50代は似通っていて「演出」と「シンプル」に関心がある
60代以上は「当たり」に監視がある

	20代	30代	40代	50代	60代以上
演出	52%	33%	45%	23%	18%
信頼	26%	21%	25%	19%	18%
当たり	18%	13%	25%	21%	43%
シンプル	3%	33%	5%	38%	21%

今回のデータ分析の結果

これから

２０代～５０代向けならば「演出」を中心に開発を行う。または、より細かく「演出」について追跡調査を行う。

ただし、今回の解析は単純に素データを解析したため、辞書が不十分である。そのため、例えば「継続率アップ」と「継続率上げて」を違うものと認識したりしている。これらの同じものと認識するためにさらなる分析が必要

ご静聴ありがとうございました

参考文献

- 西内 啓：統計学が最強の学問である
  ダイヤモンド社

- あんちべ：データ解析の実務プロセス入門
  森北出版

- 河本 薫：会社を変える分析の力
  講談社

- 石田基広：新米探偵、データ分析に挑む
  SBクリエイティブ

参考文献

マンガでわかる統計学シリーズ
作者: 高橋信,トレンドプロ
出版社/メーカー: オーム社