1

データ分析実践講座

【講義編】明日から使えるうデータ分析

森　裕一
経営学部 経営学科
<mori@mgt.ous.ac.jp>
http://www.mgt.ous.ac.jp/~mori/

2019/11/27
岡山理科大学2019年度SD研修「データ分析について」

統計に対する一般的な感覚

業績アップ
問題特定
儲けが倍増

これを使えば，問題が解決する　＝統計学は錬金術　**世の中にそんなものない**

誤信！
妄信！
過信！

でも，むずかしい　＝統計は数学　**統計の道具が数学**

ソフトさえ使いこなせれば　＝Rは魔法の道具　**出力を解釈するのはあなた**

もう少し進むと

失望…
絶望…
断念…

使ってみたけど，何も解決されない

やっぱり，統計はむずかしい

結局，データで解決なんて，できないじゃん

錬金術になりうる！

公式より背景にある**概念**

それはとても**自然**

この**理解**で魔法の道具に

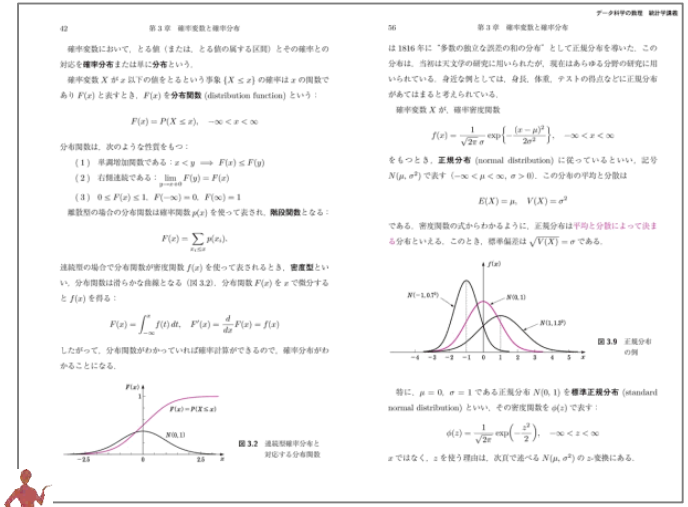
楽器／スポーツ／ダイエット／勉強…… ⇒ **まず手を**つける／そして**本質を理解**して／どれだけ**繰り返す**か ⇒ いっぱい見えてくる

統計は数学か

数学の統計

統計は数学で習うもの　苦?

分散　相関係数　Σ　確率　分布　⇒　計算



社会科の統計

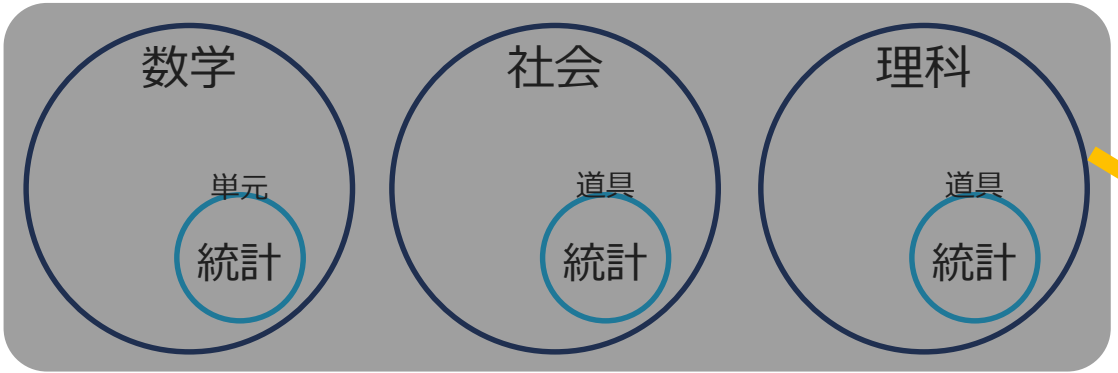
でも地図帳にはいっぱい統計が　楽！

集計　数値で比較　グラフ　GIS　⇒　読み取り

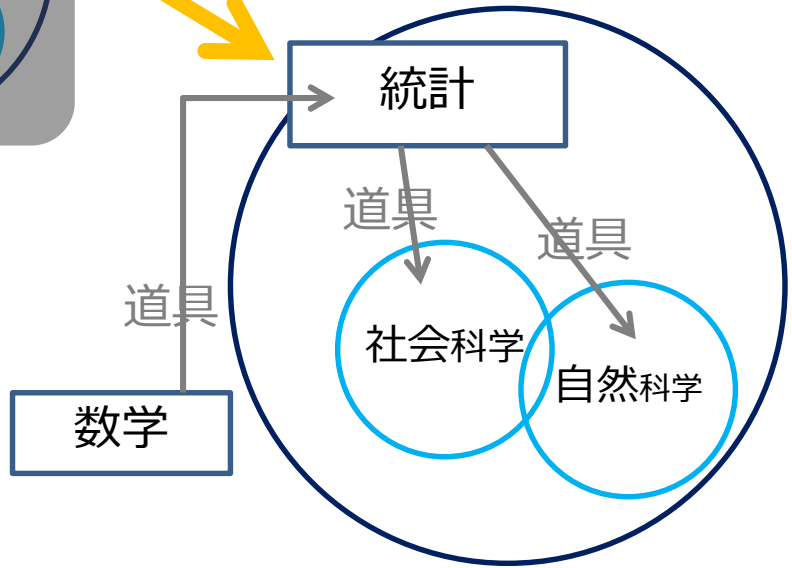
みんな同じ統計！

そういえば 理科でも保健体育でも

統計は分析のツール


数学
単元
統計
社会
道具
統計
理科
道具
統計

データサイエンス
情報技術
データ分析
統計
ビジネス戦略


統計
道具
道具
道具
社会科学
自然科学
数学

データ分析へのアプローチ

いまを知る 記述 ……**現状把握**……→ **問題点**が見えてくる

手立て
- 視覚化する
  - 表
  - グラフ
- 様相を指標で表す
  - 代表値
  - ばらつき具合
- 関連させる

＋ 地図の上に表す
時間推移で見る
層に分ける

過去とそれほど大きく変わらない

現状分析がないとダメ

現在のデータを基にする
未来のデータは手に入らない

視　点
- 中心傾向は？
- ばらつきは？
- 群は？
- 外れているものは？
- 規則性は？
  - 直線関係／一定時間ごとの傾向

× 層別 ＝ 比 較

1. 常に問題意識
   ↓ それに基づき
2. 手元にあるデータで現状を見るクセ
3. 根拠を求める姿勢
   ⊕ 分析結果から判断

将来をみる 予測 ←……現在から判断して**対策**を立てる……

予測通りにならないことは多々ある＝その都度，**軌道修正**

そこで，基礎として

セグメンテーション
ポジショニング
ターゲティング

レコメンデーション

データマイニング
テキストマイニング
ビッグデータ

データの尺度
平均・分散
相関係数
ヒストグラム
箱ひげ図
散布図
クロス集計

正規分布
区間推定
t 検定

χ^2分布
標本調査
市場調査

回帰分析

ロジスティック回帰
コレスポンデンス分析
数量化法
多次元尺度法
決定木

主成分分析
クラスター分析
判別分析
機械学習

要約
視覚化
関連性
予測
分類

層別
比較

公式 より **概念**

どうしてそれができたか
どういう考え方か

その根底は，**自然な考え**
でも，忘れている，あるいは気づいていない
言われれば当たり前のこと

研修内容

- Part1【講義編】　明日使えるデータ分析
  傾向把握と可視化
  相関・回帰・予測
- Part2【実習編】　Excelによる実践演習
  データ分析の準備
  Excelによる講義編の実践
- まとめ

研修内容

- Part1【講義編】　明日使えるデータ分析
  傾向把握と可視化
  相関・回帰・予測
- Part2【実習編】　Excelによる実践演習
  データ分析の準備
  Excelによる講義編の実践
- まとめ

データの種類（尺度水準）

テキスト型

名義尺度

- 対象（の性質）を区別するもの
- 割り振られた数値は単なるコード
- 性別，背番号，地区，商品番号

順序尺度

- 対象の大小・強弱など順序関係を区別するもの
- 割り振られた数値は順序のみ示す（間隔に意味はない）
- マラソンの順位，等級，好き嫌い

質的変数
（カテゴリカルデータ）

間隔尺度

- 対象の量的な大きさを測るもの
- 値は等間隔である（差に意味をもつ＝絶対零点がない）
- 摂氏温度，西暦年号，テストの点

比例尺度

- 対象の量的な大きさを測るもの
- 値に比例関係が成り立つ（比に意味をもつ＝絶対零点が存在）
- 身長(長さ)，体重(重さ)，容量(かさ)

量的変数

【確認１】 変数の尺度は？

各変数の尺度は？

番号	：名義尺度	順序尺度	間隔尺度	比例尺度
性別	：名義尺度	順序尺度	間隔尺度	比例尺度
利き手	：名義尺度	順序尺度	間隔尺度	比例尺度
年齢	：名義尺度	順序尺度	間隔尺度	比例尺度
所持金	：名義尺度	順序尺度	間隔尺度	比例尺度
勉強時間	：名義尺度	順序尺度	間隔尺度	比例尺度
身長	：名義尺度	順序尺度	間隔尺度	比例尺度
評定	：名義尺度	順序尺度	間隔尺度	比例尺度
偏差値	：名義尺度	順序尺度	間隔尺度	比例尺度

番号	性別	利き手	年齢	所持金	勉強時間	身長	評定	偏差値
1	F	L	21	5000	8.5	153.3	A	63
2	M	R	20	3580	2.5	175.0	S	70
3	F	R	19	412	6.5	156.5	C	58
4	M	R	22	879	9.0	168.9	B	60
5	F	L	18	6980	4.0	149.5	A	62
6	F	R	19	18900	3.5	153.5	A	69
7	M	R	20	2100	1.5	171.3	B	59

尺度による処理の違い

質的

量的

連番	氏名	氏名（カタカナ）	性別	電話番号	生年月日	年齢	通勤時間	睡眠時間	購入金額（直近）	購入金額（総額）	会員等級	スマホ
1	小山良昭	オヤマヨシアキ	男	0980893845	1964/8/14	54	30	5	31,500	31,500	B	AU
2	大野凛乃	オオノリノ	女	0761705598	1979/3/6	39	15	4	500	500	PRT	Softbank
3	横尾義久	ヨコオヨシヒサ	男	092988052	1980/9/4	38	50	6	9,000	5,000	B	Docomo
4	大和田明日香	オオワダアスカ	女	0184620181	1989/6/22	29	45	7	6,300	3,500	B	Docomo
5	仲野桃佳	ナカノモモカ	女	0921012553	1964/5/2	54	20	4	3,600	2,000	B	Softbank
6	塚越祐希	ツカゴシユウキ	女	0258091038	1995/12/31	23	20	6	2,700	2,700	PRT	AU
7	上原多紀	ウエハラタキ	[illegible]	[illegible]	1998/10/17	20	30	6	2,000	2,000	B	Docomo
8	大森祥治	オオモリショウジ	[illegible]	[illegible]	1981/3/2	37	25	5	6,300	6,300	B	Softbank
9	早坂美羽	ハヤサカミウ	[illegible]	[illegible]	1985/11/21	33	30	5	9,000	5,000	B	Softbank
10	大内敏行	オウウチトシユキ	男	0834256027	1992/11/19	26	25	4	7,200	4,000	B	Softbank
11	柴崎成光	シバザキナルミツ	男	05926951	1972/11/24	46	30	6	4,500	2,500	B	AU
12	酒井利勝	サカイトシカツ	男	085613974	1963/3/23	55	20	5	2,700	2,700	PRT	AU
13	安藤彰英	アンドウアキヒデ	男	0366209403	1991/6/16	27	35	5	4,500	4,500	B	AU
14	森岡夕菜	モリオカユウナ	女	0898174684	1973/4/29	45	15	[illegible]	[illegible]400	5,400	PRT	ETC
15	田川香帆	タガワカホ	女	0268022773	1983/12/11	35	15	[illegible]	[illegible]8,10[illegible]	8,100	B	ETC
16	大西雅子	オオニシマサコ	女	0861852862	1979/5/23	39	30	[illegible]	[illegible]00	3,000	B	AU
17	入江真一	イリエシンイチ	男	0855914896	1979/11/9	39	15	[illegible]	1,5[illegible]	1,500	PRT	Softbank
18	吉岡正道	ヨシオカマサミチ	男	0186692865	1980/1/4	38	25	6	2,000	2,000	B	Docomo
19	田所章司	タドコロショウジ	男	0610602265	1984/10/23	34	50	7	7,200	7,200	B	Docomo
20	三上清太郎	ミカミセイタロウ	男	0270451847	1995/6/20	23	35	6	3,000	3,000	B	Docomo
21	水口喜一	ミズグチキイチ	男	0959276163	1973/7/26	45	25	5	20,700	20,700	B	AU
22	岡本彦太郎	オカモトヒコタロウ	男	0858009903	1959/7/29	59	30	6	2,500	2,500	B	Docomo
23	三枝瑞貴	サエグサミズキ	女	0142004337	1977/9/19	41	20	5	9,600	9,600	PRT	AU
24	岡清作	オカセイサク	男	0791101949	1963/3/20	55	15	4	7,200	7,200	B	ETC
25	杉本豊子	スギモトトヨコ	女	0769829338	1997/9/16	21	15	4	3,600	2,000	B	Softbank
26	村木和利	ムラキカズトシ	男	0982343563	1962/9/15	56	20	4	7,200	7,200	B	Softbank
27	大井静男	オオイシズオ	男	0245625159	1986/8/21	32	30	6	25,600	25,600	B	Docomo
28	嶋田佳典	シマダヨシノリ	男	0982085233	1981/11/3	37	20	4	6,300	3,500	B	Softbank
29	上杉芳美	ウエスギヨシミ	女	0732913483	1989/6/4	29	30	5	6,300	6,300	B	AU
30	藤木吉男	フジキヨシオ	男	0866768122	1990/11/23	28	35	7	2,500	2,500	B	Docomo
31	村岡幸太郎	ムラオカコウタロウ	男	0858164904	1996/5/9	22	20	4	48,100	48,100	B	Softbank
32	笹川竹志	ササガワタケシ	男	0539026852	1992/1/21	26	30	6	4,500	2,500	B	Docomo
33	長沼翠	ナガヌマミドリ	女	0748816602	1960/9/21	58	20	4	1,500	1,500	PRT	Softbank
34	小森伸子	コモリノブコ	女	0319115190	1987/9/19	31	40	6	7,200	4,000	B	Docomo
35	近藤和幸	コンドウカズユキ	男	0965390741	1998/9/8	20	30	6	4,500	2,500	B	Docomo
36	西野新次郎	ニシノシンジロウ	男	0456028872	1968/7/11	50	30	5	5,000	5,000	B	Softbank
37	中井利吉	ナカイリキチ	男	0748324423	1986/2/26	32	15	4	3,600	3,600	B	Softbank
38	今清吉	コンセイキチ	男	0889769291	1976/5/2	42	40	5	9,000	5,000	B	AU
39	望月和茂	モチヅキカズシゲ	男	0879134368	1987/10/7	31	25	4	9,000	5,000	B	Softbank
40	杉原由香里	スギハラユカリ	女	087601879	1983/3/12	35	25	6	3,600	2,000	B	AU

たとえば，**集計**と**要約**

1変量データ

質的変数

集計，視覚化，カテゴリの結合

量的変数

度数分布表・ヒストグラム・箱ひげ図

基本統計量（代表値（平均等））

散らばり（標準偏差等）

2変量データ

質的vs質的

クロス集計

質的vs量的

層別の要約

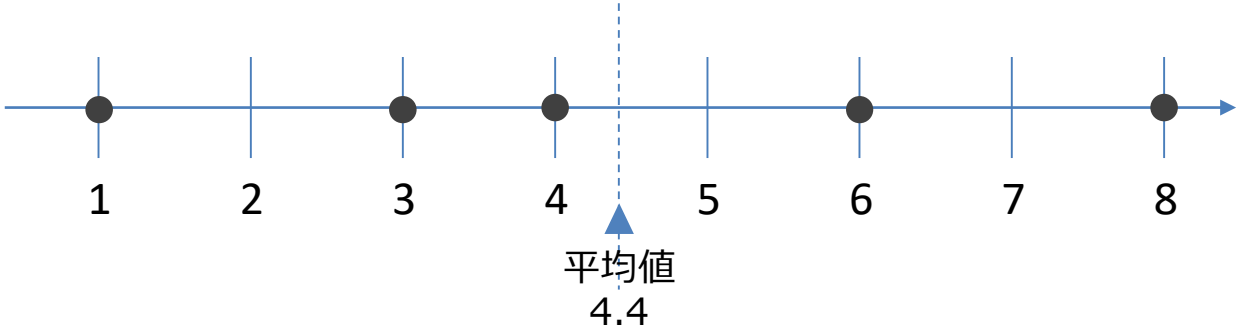
量的vs量的

散布図，相関係数

平 均

平均 ⇒ 説明は・・・不要ですね。

その対象（集団）を**代表する**１つの**指標**です。



$$(1+3+4+6+8)\div 5=4.4$$

Σの話

$x_1 + x_2 + x_3 + \cdots + x_{20}$

を「･･･」を使わず表現してください。

$1 + 3 + 4 + 6 + 8$

を筆算で書くと・・・。

$$\begin{array}{rr} & 1 \\ & 3 \\ & 4 \\ & 6 \\ +) & 8 \\ \hline & \boxed{22} \end{array}$$

$$\begin{array}{rr} & x_1 \\ & x_2 \\ & x_3 \\ & \vdots \\ +) & x_{20} \\ \hline \end{array}$$

たす $x_{1\cdots 20}$

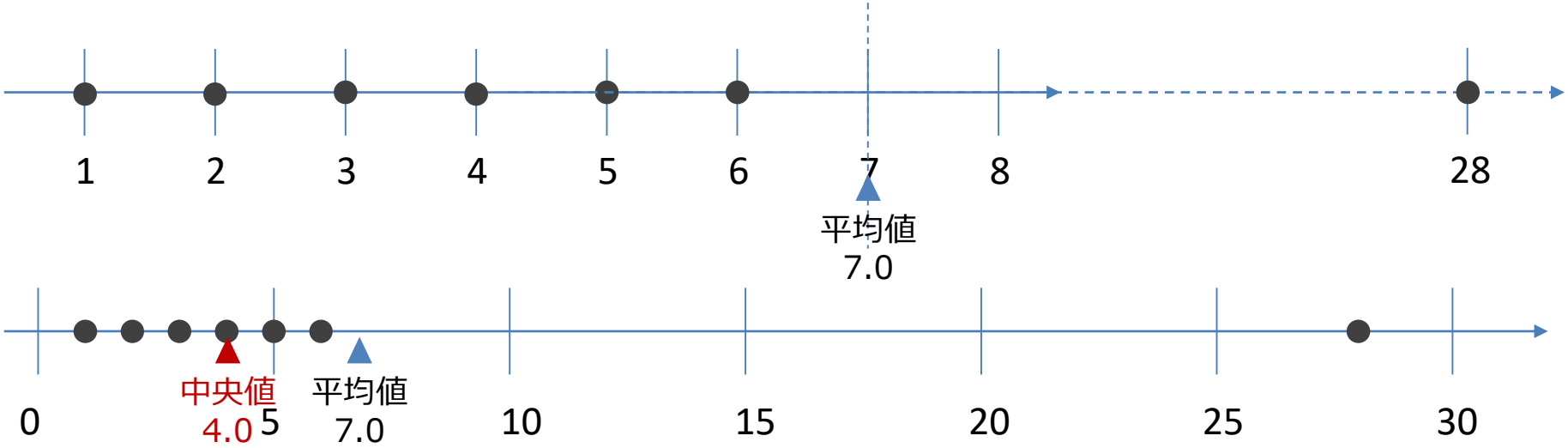
$\mathrm{Sum}_{i=1\sim 20}\ x_i$

$\sum_{i=1}^{20} x_i$

なので，平均は $\frac{1}{n}\sum_{i=1}^{n} x_i$

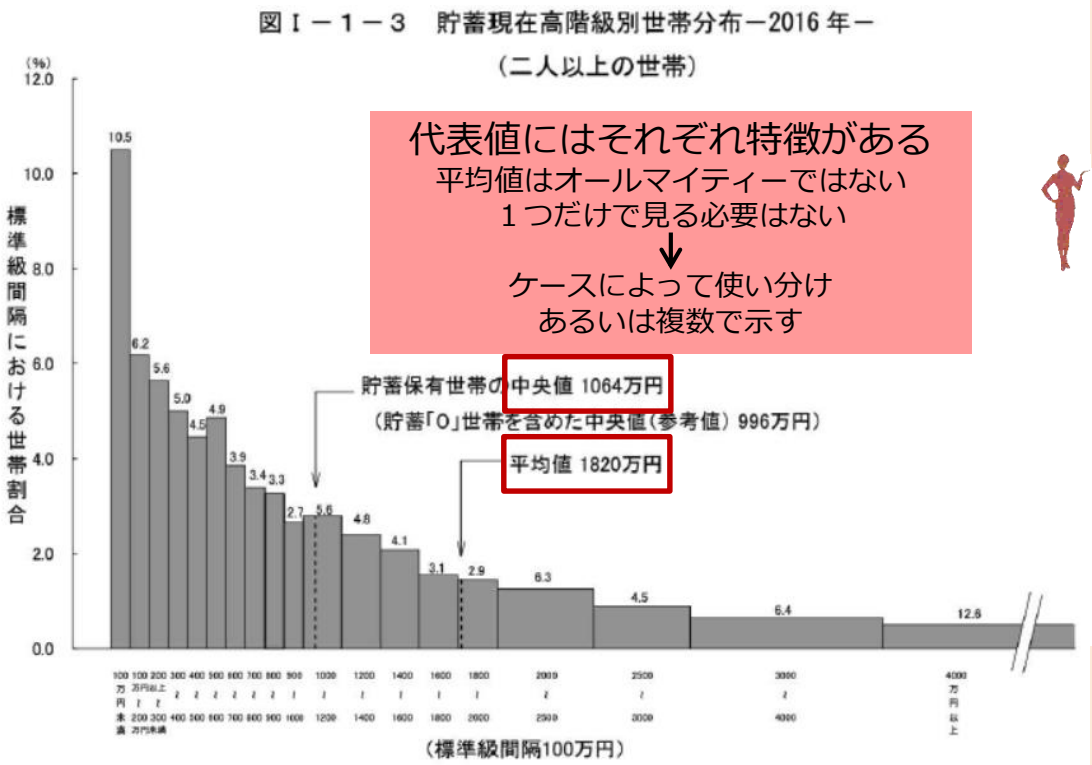
平 均

{1, 2, 3, 4, 5, 6, 28} の代表値は　　**平均**＝7　　… ？



こういう場合は，**中央値**＝4

平均値と中央値



総務省 家計調査報告（貯蓄・負債編） 平成28年度平均結果速報（二人以上の世帯）

代表値にはそれぞれ特徴がある
平均値はオールマイティーではない
１つだけで見る必要はない
↓
ケースによって使い分け
あるいは複数で示す

{1, 2, 3, 4, 5, 6, 28} の代表値は
平均＝7 … ？
こういう場合は，**中央値**＝4

平均値
○ イメージしやすい。
○ 計算が容易である。
○ その性質がさまざまな計算に使われる。
× 外れ値に影響されやすい。
× 右や左に裾野の長い分布の代表値には適さない。

中央値
○ 外れ値に影響を受けにくい。
○ 分布によらず，その前後に半数のデータがある。
× 計算が面倒である（並べ替えが必要）。
× 理論的には平均値ほど扱われない。

↓

平均値の妄信はダメ
||
分布をみて，適している代表値を確認！
（最頻値も）

ばらつきを見ることの大切さ

測定・観察されたデータには，必ずばらつきがある

内閣支持率30%

ま，それぐらいだろう。
新聞社によっても違うし。
全員調べないんだから当然誤差あるよね。

32%から30%で「下がった，下がった」は大げさ。
70%から30%・・・「うーん，これは落ちた」な。

おいしい手作りアンパン

１つ160円。でも大きさにばらつきがあるんだよね。
手作りだから仕方ないか。おっきめのを買おっと。

できるだけ同じ重さになるように機械で作ってるけど，
完全に同じにはできないよね。

視聴率30%

ビデリサーチ社によると，岡山香川地区の調査世帯数は200。
これは，最大７%の誤差を認めていることになる。
具体的に見てみると，視聴率20%の誤差は5.7%，視聴率30%の誤差は6.5%
なので，それぞれの真の視聴率は， 14.3%～25.7%と23.5%～36.5% の間。
えー，じゃ，ひょっとして，20%の番組の方が30%の番組より
多くの世帯が見ている可能性もあるわけだ・・・。

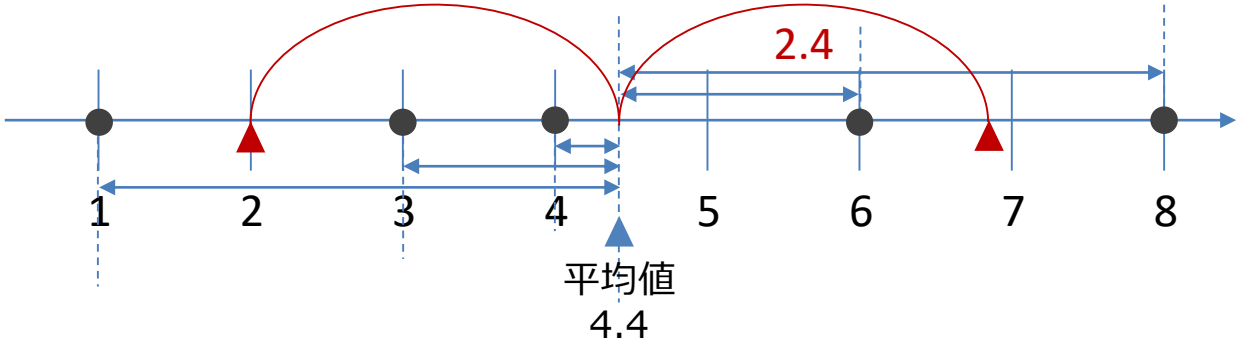
平均だけではダメ
　平均は元の山をならすだけ
　元の山のデコボコまではわからない
誤差の存在

ばらつきを考察 ← 統計

分散（標準偏差）

分散 ⇒ データの散らばり具合を表す指標 ⇒ **標準偏差** = $\sqrt{\text{分散}}$

その対象（集団）の各値が**どれぐらいばらついているか**がわかります。



$$
\begin{array}{rr}
 & (1-4.4)^2 \\
 & (3-4.4)^2 \\
 & (4-4.4)^2 \\
 & (6-4.4)^2 \\
+) & (8-4.4)^2 \\
\hline
 & 29.2
\end{array}
$$

$29.2 \div 5 = 5.84$

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

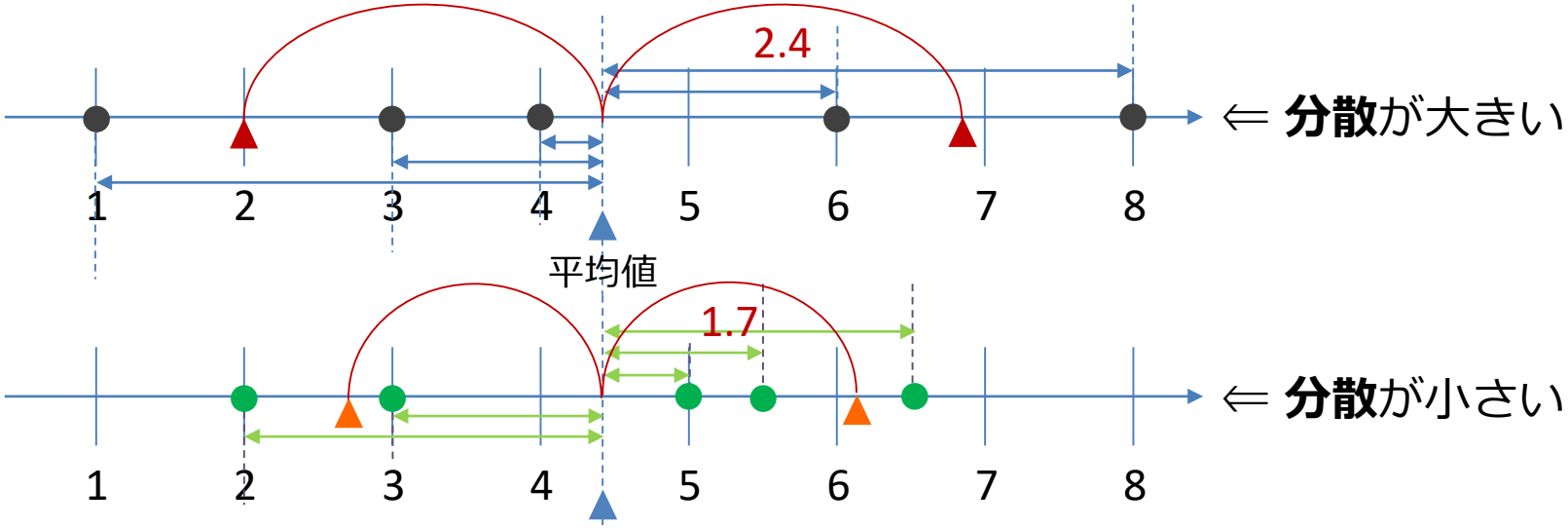
平均からのずれの２乗の１個あたりの値

＝すべての⟷を２乗して，個数で割ったもの

分散（標準偏差）

分散 ⇒ データの散らばり具合を表す指標 ⇒ **標準偏差** = $\sqrt{分散}$

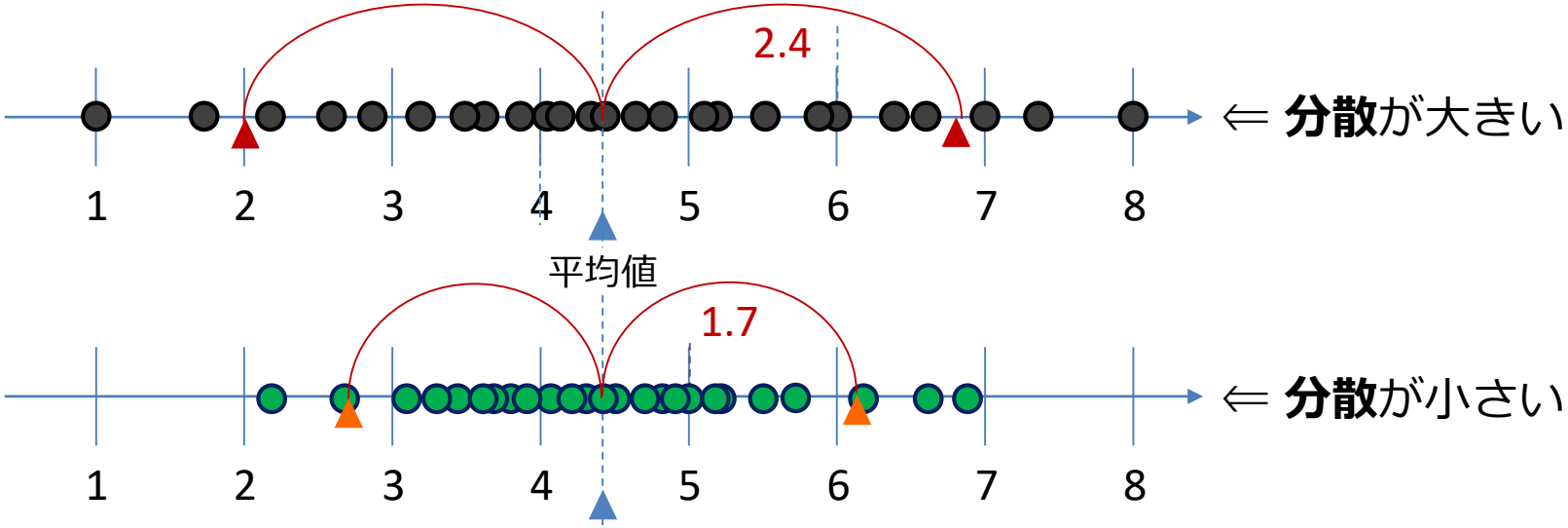
その対象（集団）の各値が**どれぐらいばらついているか**がわかります。



分散（標準偏差）

分散 ⇒ データの散らばり具合を表す指標 ⇒ **標準偏差** = $\sqrt{分散}$

その対象（集団）の各値が**どれぐらいばらついているか**がわかります。



【確認２】 どう使う？どう考える？

４人のボーリングの実力を評価したい。４人の10フレーム×２ゲームのデータをとった。さて？

Frame	Game 1				Game 2			
	A	B	C	D	A	B	C	D
1	5	0	8	7	3	5	8	7
2	4	3	7	6	5	3	10	6
3	4	4	8	6	6	6	8	8
4	6	9	8	8	5	3	9	8
5	3	1	9	7	4	5	9	5
6	7	0	10	9	6	4	8	6
7	3	5	10	7	4	4	7	7
8	6	6	9	7	7	6	10	6
9	3	5	9	6	4	2	10	9
10	0	5	9	9	4	3	9	6

【確認２】 どう使う？どう考える？

４人のボーリングの実力評価 ⇒ 要約すると楽：平均と標準偏差で

Frame	Game 1				Game 2			
	A	B	C	D	A	B	C	D
1	5	0	8	7	3	5	8	7
2	4	3	7	6	5	3	10	6
3	4	4	8	6	6	6	8	8
4	6	9	8	8	5	3	9	8
5	3	1	9	7	4	5	9	5
6	7	0	10	9	6	4	8	6
7	3	5	10	7	4	4	7	7
8	6	6	9	7	7	6	10	6
9	3	5	9	6	4	2	10	9
10	0	5	9	9	4	3	9	6
平　均	4.1	3.8	8.7	7.2	4.8	4.1	8.8	6.8
標準偏差	**1.92**	**2.71**	**0.90**	**1.08**	**1.17**	**1.30**	**0.98**	**1.17**

単純な平均・標準偏差だけで
→　比較や動きが！
||

目 的
× 数値を求めること
○ それを使って**判断**すること

概念・本質
生データのまま(数値の羅列)では判断が困難
なので，**要約**した値で考察。

【確認２】 どう使う？どう考える？

４人のボーリングの実力評価 ⇒ 要約すると楽：平均と標準偏差で

期	2016				2017			
	支店A	支店B	支店C	支店D	支店A	支店B	支店C	支店D
1	122	32	176	158	86	122	176	158
2	104	86	158	140	122	86	212	140
3	104	104	176	140	140	140	176	176
4	140	194	176	176	122	86	194	176
5	86	50	194	158	104	122	194	122
6	158	32	212	194	140	104	176	140
7	86	122	212	158	104	104	158	158
8	140	140	194	158	158	140	212	140
9	86	122	194	140	104	68	212	194
10	32	122	194	194	104	86	194	140
平　均	105.8	100.4	188.6	161.6	118.4	105.8	190.4	154.4
標準偏差	34.577	48.833	16.2	19.387	20.991	23.4	17.636	20.991

単純な平均・標準偏差だけで
→　比較や動きが！
||

目 的
× 数値を求めること
○ それを使って**判断**すること

概念・本質
生データのまま(数値の羅列)では判断が困難
なので，**要約**した値で考察。

店舗別の売り上げ
⇒ 評価や改善指導へ

作業員の能力管理
⇒ 実力アップ

機械の性能チェック
⇒ 保守・効率

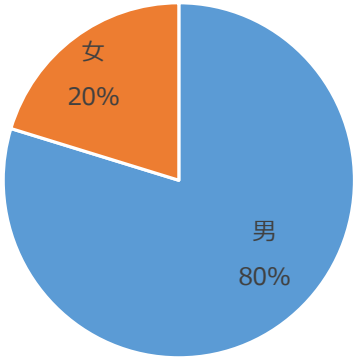
などなど

傾向把握のための可視化

内訳
比較
推移
分布
関係
層別

内訳

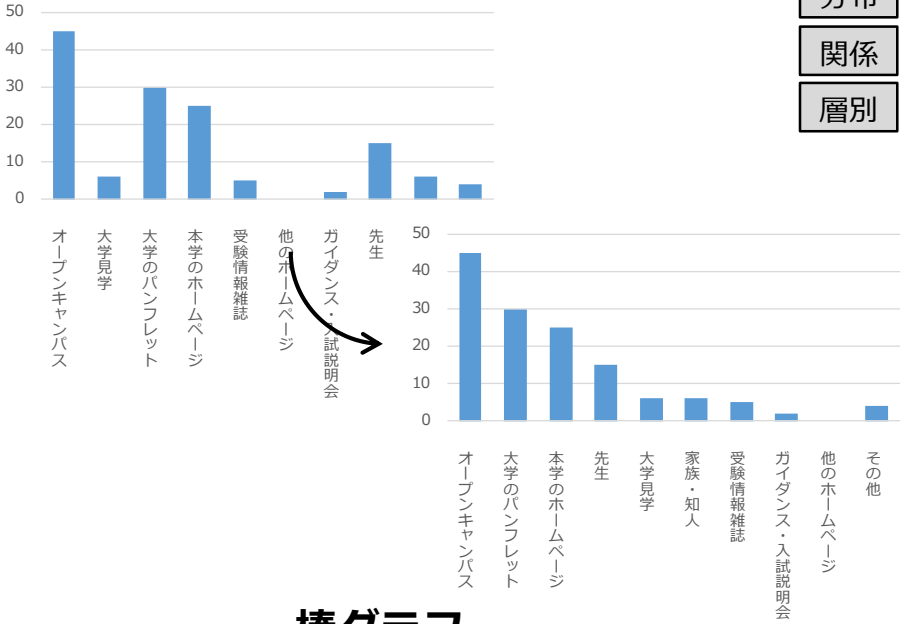
性別	度数
男	110
女	28



円グラフ

比較

参考にした情報源	度数
オープンキャンパス	45
大学見学	6
大学のパンフレット	30
本学のホームページ	25
受験情報雑誌	5
他のホームページ	0
ガイダンス・入試説明会	2
先生	15
家族・知人	6
その他	4
総　計	138



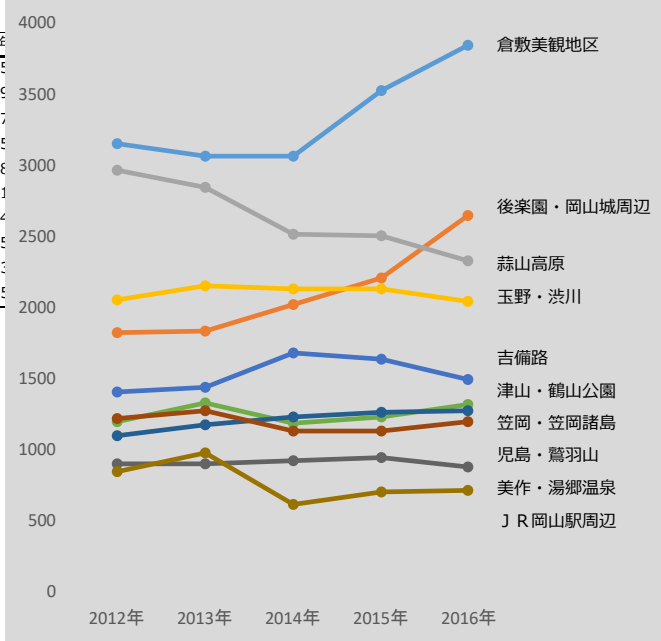
棒グラフ

傾向把握のための可視化

内訳
比較
推移
分布
関係
層別

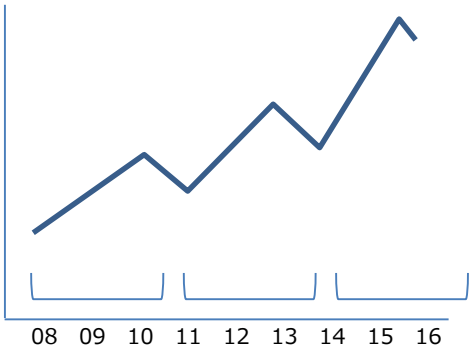
推移

観光地域	2012年	2013年	2014年	2015年	2016年
倉敷美観地区	3151	3069	3065	3534	3845
後楽園・岡山城周辺	1824	1829	2021	2212	2649
蒜山高原	2970	2850	2512	2504	2327
玉野・渋川	2059	2153	2130	2132	2045
吉備路	1401	1433	1675	1636	1498
津山・鶴山公園	1190	1331	1182	1228	1321
笠岡・笠岡諸島	1091	1168	1226	1266	1274
児島・鷲羽山	1217	1268	1125	1126	1195
美作・湯郷温泉	896	903	920	947	873
ＪＲ岡山駅周辺	845	979	610	701	715



折れ線グラフ

時系列データ



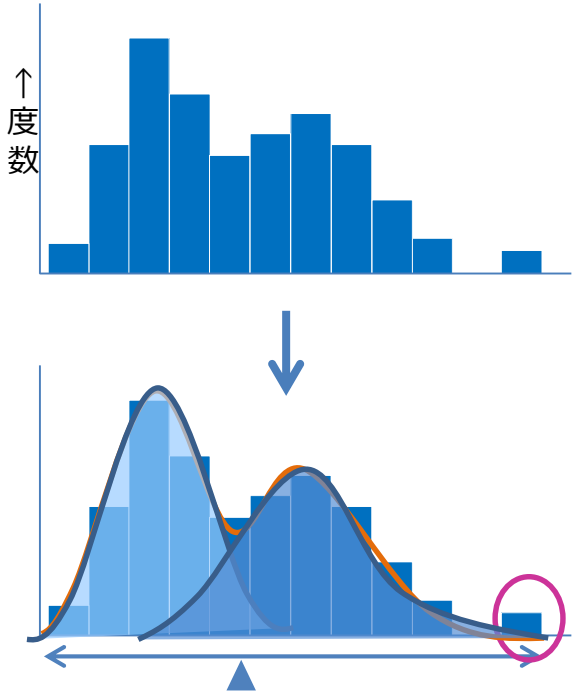
ガタガタしながらも増加傾向

周期性？

傾向把握のための可視化

分布

ヒストグラム



最大・最小

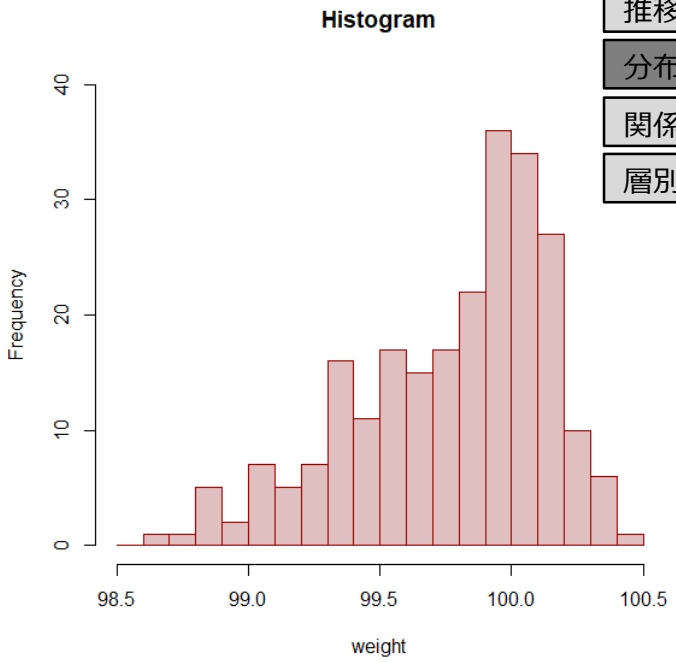
範囲は？

中心傾向は？

とびはなれた観測値は？

山が見える・・・分布のさま
　とんがってる・・・分散小さい
　なだらか　　・・・分散大きい
　左／右 に偏っている！
　２つの山の存在？！

機械加工品・・・完成製品の重量



少し右に偏っている
左にすそ野が長いなぁ・・・・

内訳
比較
推移
分布
関係
層別

傾向把握のための可視化

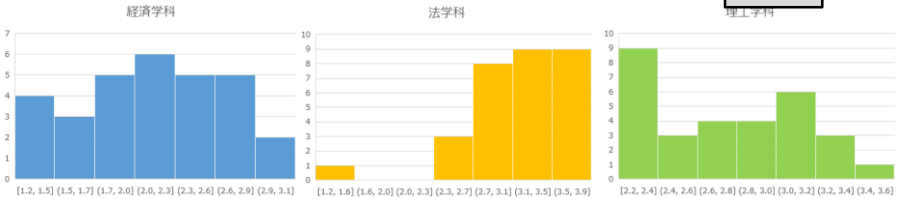
分布

学生番号	学科	GPA
C010	経済学科	3.1
C009	経済学科	3.1
C001	経済学科	2.7
C002	経済学科	2.7
C023	経済学科	2.6
C024	経済学科	2.6
C004	経済学科	2.6
C008	経済学科	2.5
C005	経済学科	2.5
C022	経済学科	2.4
C006	経済学科	2.4
C026	経済学科	2.4
C021	経済学科	2.2
C025	経済学科	2.2
C012	経済学科	2.2
平均		2.2
標準偏差		0.494
C030	経済学科	2.0
C018	経済学科	2.0
C014	経済学科	1.8
C003	経済学科	1.8

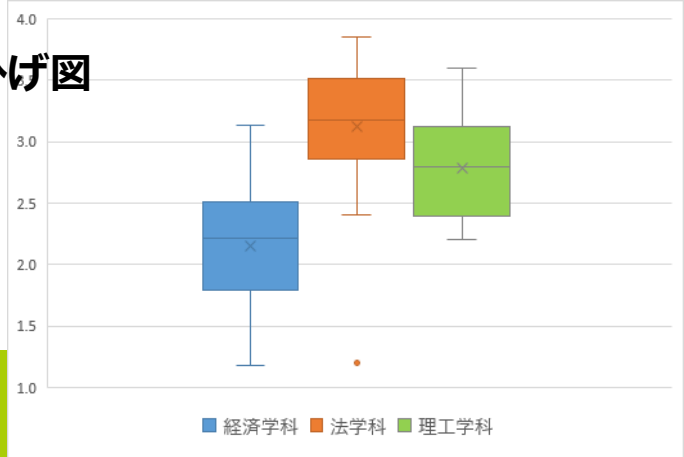
学生番号	学科	GPA
L023	法学科	3.9
L009	法学科	3.7
L016	法学科	3.6
L028	法学科	3.6
L020	法学科	3.6
L005	法学科	3.6
L006	法学科	3.6
L017	法学科	3.5
L026	法学科	3.5
L014	法学科	3.4
L015	法学科	3.3
L010	法学科	3.3
L030	法学科	3.3
L029	法学科	3.2
L011	法学科	3.2
平均		3.1
標準偏差		0.499
L004	法学科	3.0
L008	法学科	3.0
L007	法学科	3.0
L001	法学科	3.0

学生番号	学科	GPA
S008	理工学科	3.6
S007	理工学科	3.4
S014	理工学科	3.3
S019	理工学科	3.3
S023	理工学科	3.2
S011	理工学科	3.2
S026	理工学科	3.2
S004	理工学科	3.1
S021	理工学科	3.1
S029	理工学科	3.1
S017	理工学科	3.0
S015	理工学科	2.9
S001	理工学科	2.9
S005	理工学科	2.9
S028	理工学科	2.8
平均		2.8
標準偏差		0.394
S024	理工学科	2.6
S027	理工学科	2.5
S012	理工学科	2.5
S009	理工学科	2.4

ヒストグラム



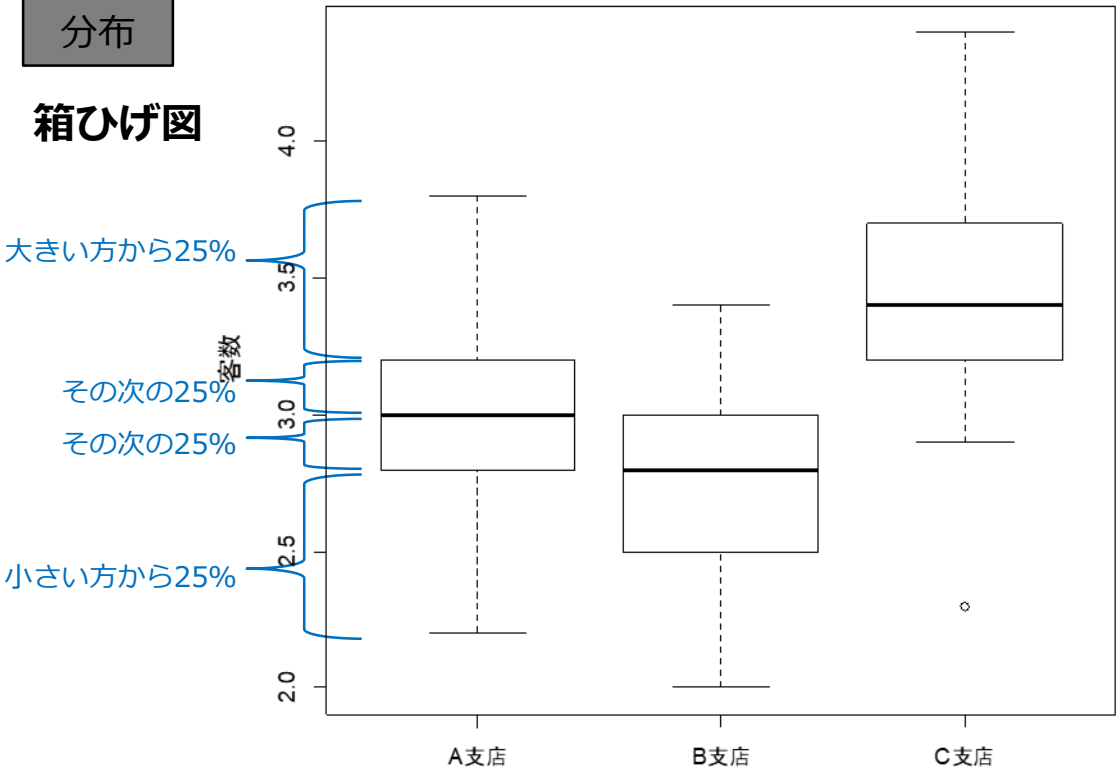
箱ひげ図



傾向把握のための可視化

分布

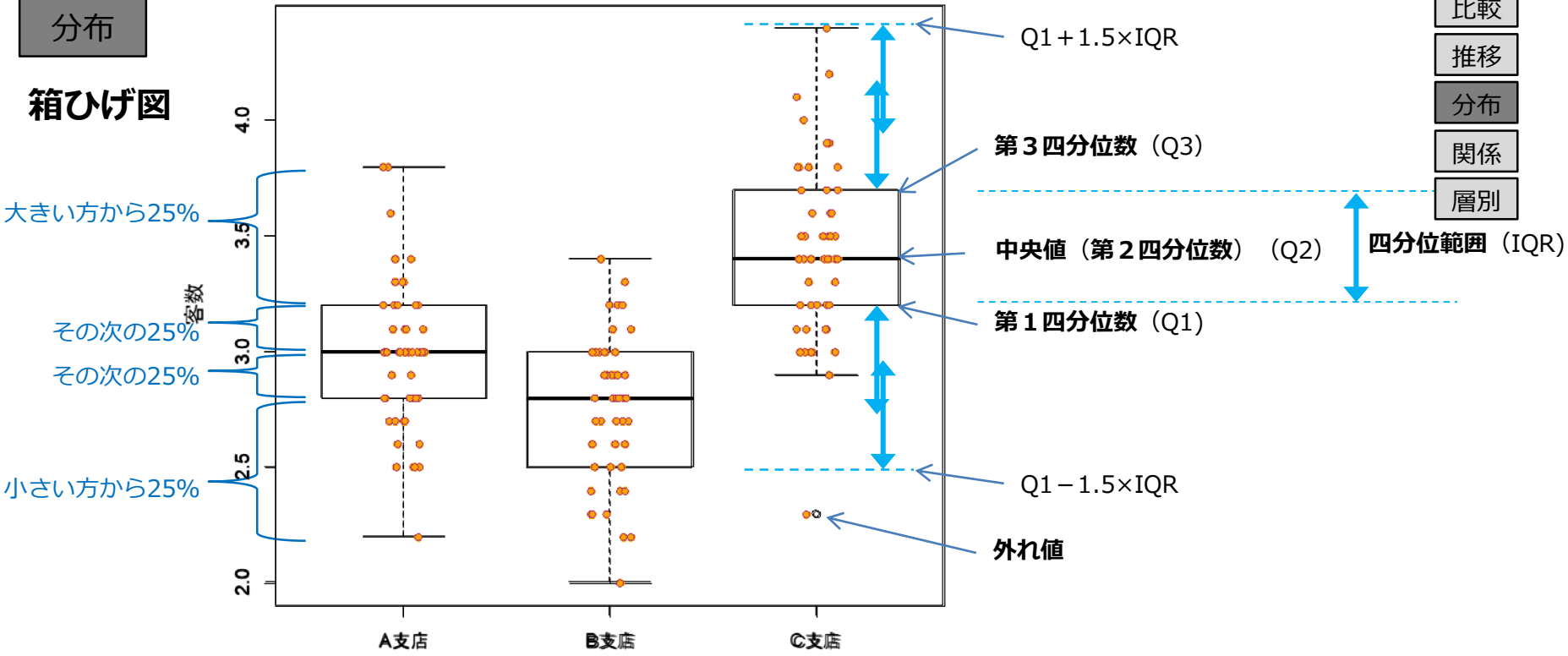
箱ひげ図



傾向把握のための可視化

分布

箱ひげ図



内訳
比較
推移
分布
関係
層別

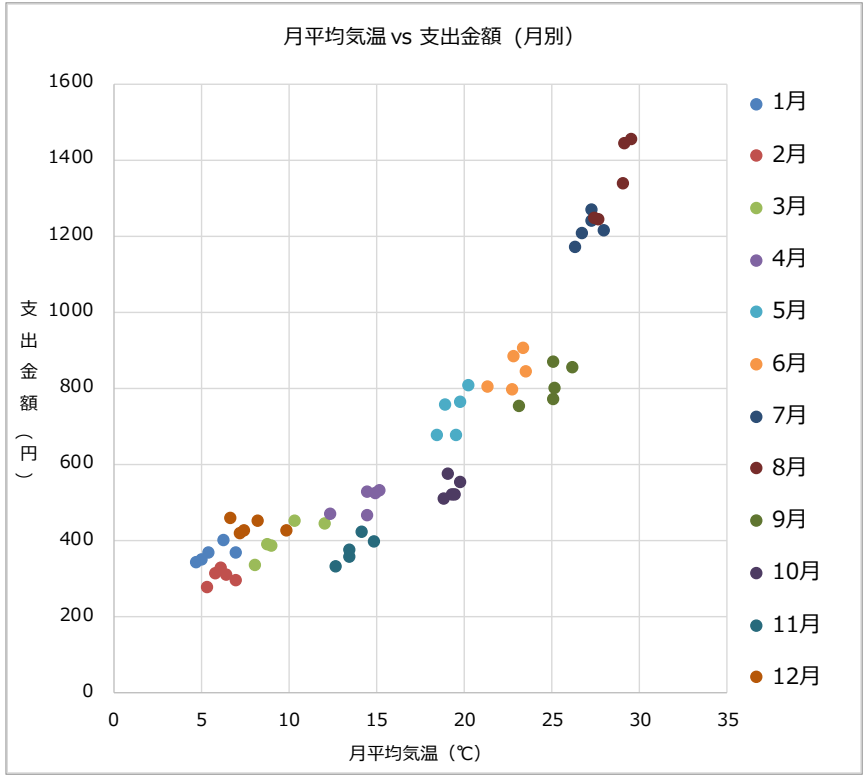
傾向把握のための可視化

関係

年	月	月平均気温（℃）	アイスクリーム支出金額 (円)
2010	1	7.0	362
2010	2	6.5	305
2010	3	9.1	383
2010	4	12.4	464
2010	5	19.0	752
2010	6	23.6	841
2010	7	28.0	1211
2010	8	29.6	1451
2010	9	25.1	864
2010	10	18.9	504
2010	11	13.5	351
2010	12	9.9	423
2011	1	5.1	346
2011	2	7.0	289
2011	3		
2011	4		
2011	5		
2011	6	22.8	791
2011	7	27.3	1265
2011	8	27.5	1241
2011	9	25.1	767

相関係数＝0.905

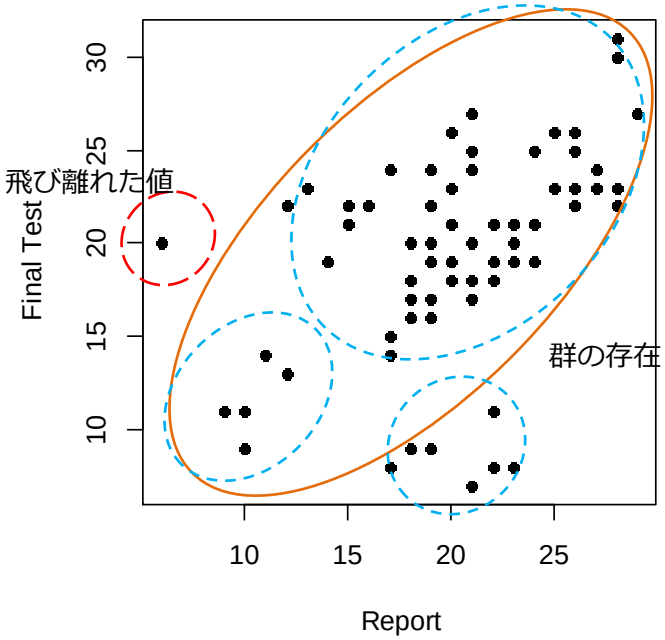
散布図



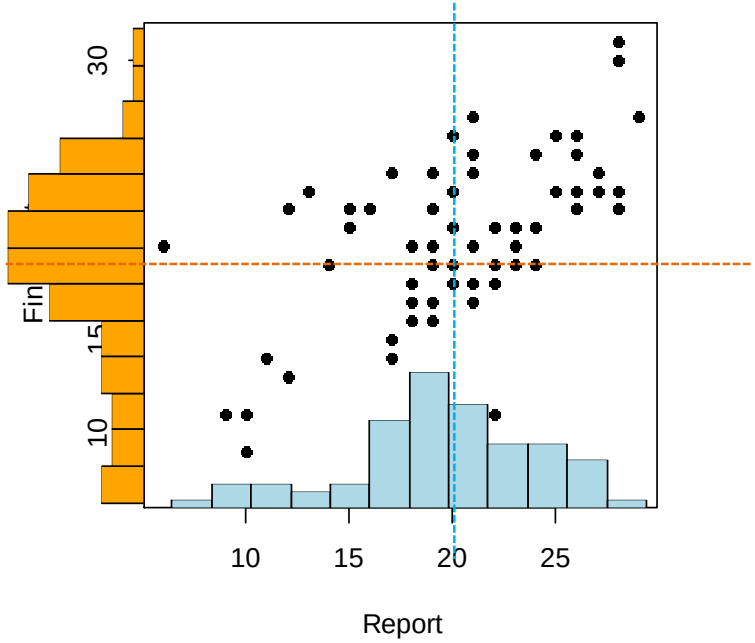
傾向把握のための可視化

関係 **散布図**

右上がり，直線傾向



x，y それぞれで
平均や度数分布が見える？



傾向把握のための可視化

グラフの種類	比較	内訳	推移	分布	相関
面積グラフ	◎				
棒グラフ	◎	○	○	△	△
帯グラフ	○	◎	○		
円グラフ	○	◎			
面グラフ	○	◎	○		
折れ線グラフ	△		◎		
レーダーチャート	◎	○	○		
散布図			○	◎	◎

山本義郎（2005）．グラフの表現術，講談社現代新書.

グラフに表す
→ 様子をつかむ
部分の様子
||
目 的
× 作図すること
○ それを使って**判断**すること

内訳
比較
推移
分布
関係
層別

概念・本質
生データ(数値の羅列)だけ
では判断が困難。
↓
視覚的・図形的な特徴を掴む。
自然な気づきでよい。

どんなデータも視覚化を！
指標値と併用
目的に合ったグラフを

層別で問題解決へ

層別

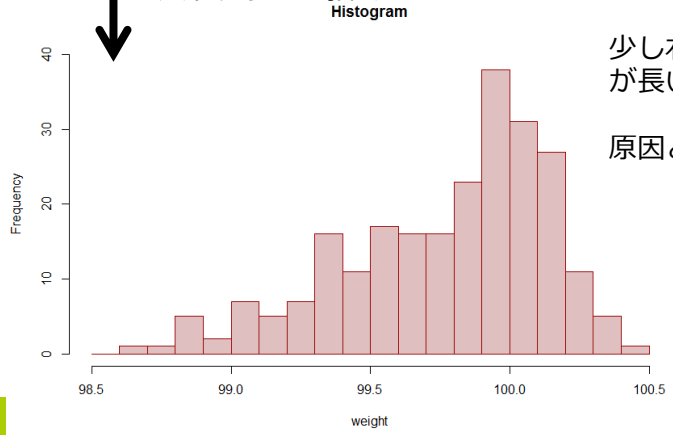
No.	MachineID	Period	Weight
1	A	Day	99.88
2	A	Day	100.09
3	A	Day	99.88
4	A	Day	100.05
5	A	Day	100.12
6	A		
		Night	100.05
98	A	Night	100.30
99	A	Night	100.30
100	A	Night	100.00
101	A	Night	100.05
102	A		
		Day	99.24
165	B	Day	99.54
166	B	Day	99.91
167	B	Day	99.83
168	B		
		Night	99.38
239	B	Night	99.59
240	B	Night	99.56

最近，加工品の完成精度が下がっているとの報告があった。重量にばらつきが出始めているとのこと。２つのマシンを昼夜交代制で 24時間フル稼働させている。原因究明を。

無作為に製品を抜き取って重さを測ってみる。

場合分けができる　⇒　原因究明に役立つかも
　２つのマシン　A　と　B
　昼と夜

ヒストグラムを描くと・・・・



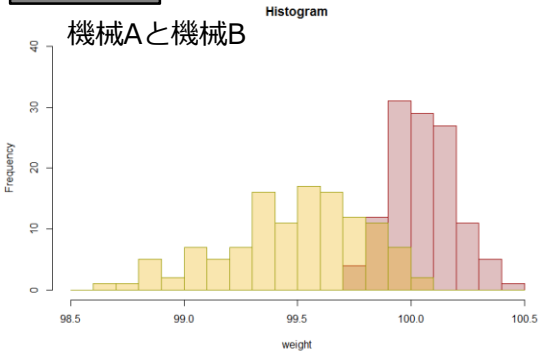
少し右に偏っている／左にすそ野が長いなぁ・・・・

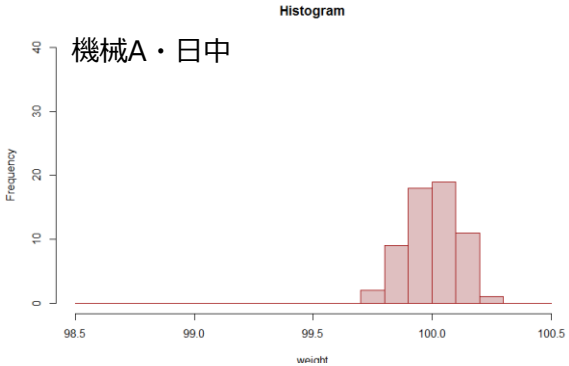
原因として，何が予測されるか？

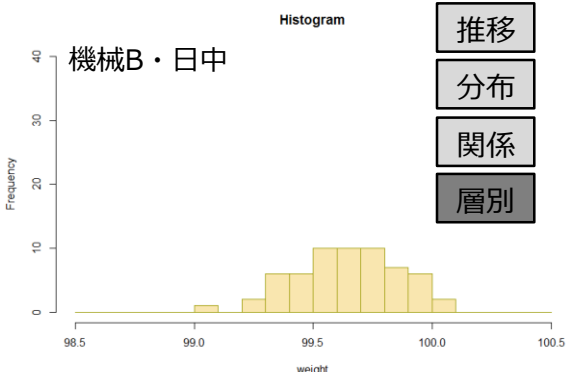
層別で問題解決へ

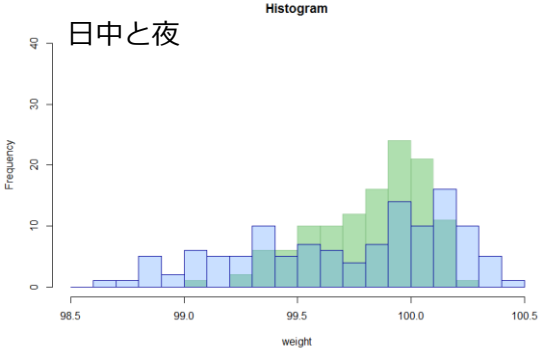
層別

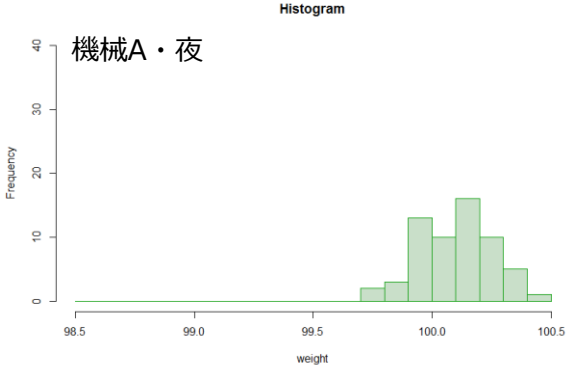
内訳
比較
推移
分布
関係
層別

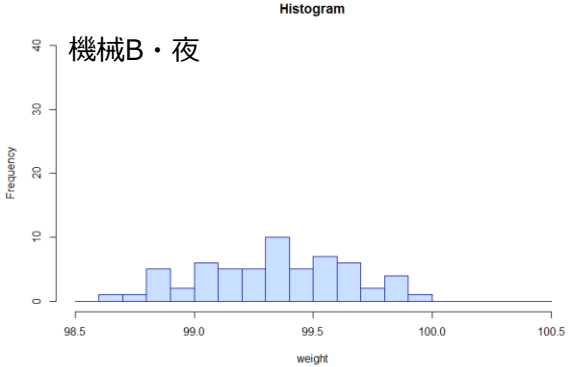












層別分析（クロス集計）

層別

スマホの新製品を売り出した。
このスマホの満足度を評価したい。
購入者にアンケートをとろう！
男女合わせて100人に調査。
さて，ここから・・・？
（右は，性別と満足度を取り出したもの。
他に，年齢，地域，職業，目的，モデル
なども聞いていく。）

性別	満足度	性別	満足度	性別	満足度	性別	満足度
女	4	男	1	男	2	男	2
女	4	女	4	男	3	男	1
女	4	男	4	男	3	女	3
男	1	女	4	女	2	男	1
男	1	男	1	男	2	男	1
男	1	女	4	男	3	男	1
男	1	男	1	女	4	男	1
女	4	女	4	男	2	男	1
女	4	男	1	男	2	女	2
男	1	女	3	男	1	女	2
男	1	女	4	男	1	女	3
男	1	女	4	男	1	女	4
男	2	女	1	男	2	男	2
女	4	女	4	女	3	女	4
男	1	男	1	女	2	女	4
男	1	女	4	女	2	男	3
男	1	男	3	男	1	男	1
女	2	男	2	女	2	女	4
女	1	男	1	男	2	男	1
女	4	男	4	男	1	女	3
女	4	女	4	男	1	女	3
男	1	男	1	男	2	男	2
女	4	女	4	男	2	男	1
女	4	男	3	男	1	男	2
女	1	男	1	男	1	女	4

層別分析（クロス集計）

層別

新商品の満足度（男女100人に調査）

	満 足	少し満足	少し不満足	不満足	計
回答数	39	21	12	28	100

ちょっと不満足が多いけど，
満足が一番多いし，肯定的な人も60人。
なかなかいいんじゃない！

アンケートでは，性別や年齢も聞いている。
性別で分けて集計してみると・・・。

クロス表（分割表）

	満 足	少し満足	少し不満足	不満足	計
男	36	14	6	2	58
女	3	7	6	26	42
計	39	21	12	28	100

女性の支持率が低い。
新商品の受けは女性に悪い。 ⇒原因 男性はどこがよいと思っているのか。
女性にとって，何がよくないのか。 ⇒さらに，年齢などでクロス

さらなる分析ストーリー：男女の差はなぜ？ 使用時間かも ⇒ 使用時間の短い人が男女ともに不満足→使用頻度をあげる工夫へ

単純なクロス集計
→ 要因の特定の資料になる。
||

目 的
△ 全体的な評価を知ること
○ **要因**を見つけること

概念・本質
対象がいくつかの属性に分けられるとき，
その**属性**によって結果が違うかどうかを考察。

【確認３】 どう使う？どう考える？

５つの学科の学力状況を把握したい。どうする？

学生番号	学科	GPA
E027	英文学科	3.3
E019	英文学科	3.3
E026	英文学科	3.2
E003	英文学科	3.1
E023	英文学科	3.0
E010	英文学科	3.0
E007	英文学科	2.9
E005	英文学科	2.9
E002	英文学科	2.9
E022	英文学科	2.9
E025	英文学科	2.8
E021	英文学科	2.8
E006	英文学科	2.7
E018	英文学科	2.7
E013	英文学科	2.6
E015	英文学科	2.6
E011	英文学科	2.6
E017	英文学科	2.6
E030	英文学科	2.5
E028	英文学科	2.5
E004	英文学科	2.5
E009	英文学科	2.5
E014	英文学科	2.4
E001	英文学科	2.3
E016	英文学科	2.2
E024	英文学科	2.1
E012	英文学科	2.1
E020	英文学科	2.0
E008	英文学科	1.8
E029	英文学科	1.7

学生番号	学科	GPA
K023	教育学科	3.6
K010	教育学科	3.2
K021	教育学科	3.1
K007	教育学科	3.1
K017	教育学科	3.1
K008	教育学科	3.1
K030	教育学科	3.0
K019	教育学科	3.0
K006	教育学科	3.0
K012	教育学科	2.9
K025	教育学科	2.8
K026	教育学科	2.8
K018	教育学科	2.7
K013	教育学科	2.7
K027	教育学科	2.7
K020	教育学科	2.7
K002	教育学科	2.6
K022	教育学科	2.6
K005	教育学科	2.6
K004	教育学科	2.6
K003	教育学科	2.5
K011	教育学科	2.4
K014	教育学科	2.3
K016	教育学科	2.1
K028	教育学科	2.0
K015	教育学科	2.0
K029	教育学科	1.9
K024	教育学科	1.8
K009	教育学科	1.8
K001	教育学科	1.4

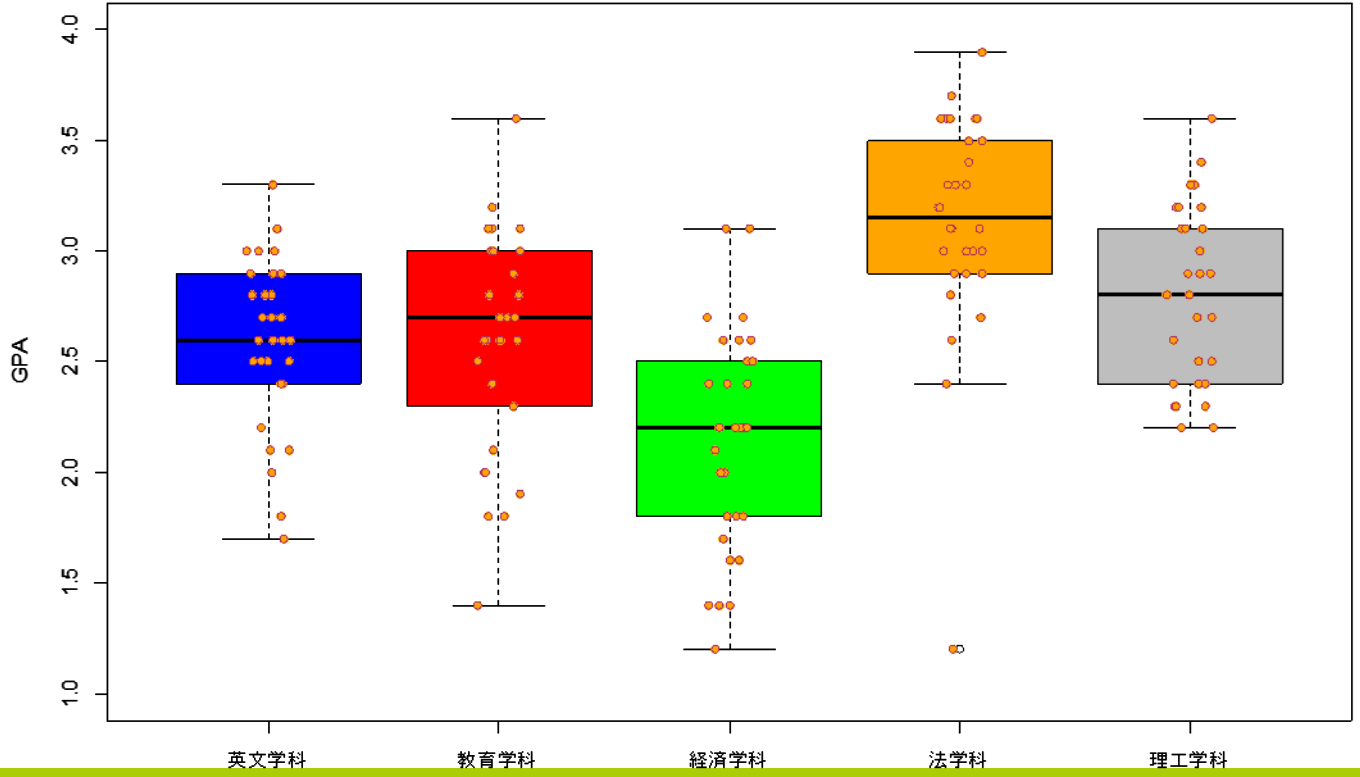
学生番号	学科	GPA
C010	経済学科	3.1
C009	経済学科	3.1
C001	経済学科	2.7
C002	経済学科	2.7
C023	経済学科	2.6
C024	経済学科	2.6
C004	経済学科	2.6
C008	経済学科	2.5
C005	経済学科	2.5
C022	経済学科	2.4
C006	経済学科	2.4
C026	経済学科	2.4
C021	経済学科	2.2
C025	経済学科	2.2
C012	経済学科	2.2
C028	経済学科	2.2
C007	経済学科	2.2
C015	経済学科	2.1
C030	経済学科	2.0
C018	経済学科	2.0
C014	経済学科	1.8
C003	経済学科	1.8
C019	経済学科	1.8
C016	経済学科	1.7
C027	経済学科	1.6
C029	経済学科	1.6
C020	経済学科	1.4
C011	経済学科	1.4
C017	経済学科	1.4
C013	経済学科	1.2

学生番号	学科	GPA
L023	法学科	3.9
L009	法学科	3.7
L016	法学科	3.6
L028	法学科	3.6
L020	法学科	3.6
L005	法学科	3.6
L006	法学科	3.6
L017	法学科	3.5
L026	法学科	3.5
L014	法学科	3.4
L015	法学科	3.3
L010	法学科	3.3
L030	法学科	3.3
L029	法学科	3.2
L011	法学科	3.2
L024	法学科	3.1
L019	法学科	3.1
L012	法学科	3.1
L004	法学科	3.0
L008	法学科	3.0
L007	法学科	3.0
L001	法学科	3.0
L018	法学科	2.9
L003	法学科	2.9
L025	法学科	2.9
L002	法学科	2.8
L013	法学科	2.7
L027	法学科	2.6
L022	法学科	2.4
L021	法学科	1.2

学生番号	学科	GPA
S008	理工学科	3.6
S007	理工学科	3.4
S014	理工学科	3.3
S019	理工学科	3.3
S023	理工学科	3.2
S011	理工学科	3.2
S026	理工学科	3.2
S004	理工学科	3.1
S021	理工学科	3.1
S029	理工学科	3.1
S017	理工学科	3.0
S015	理工学科	2.9
S001	理工学科	2.9
S005	理工学科	2.9
S028	理工学科	2.8
S003	理工学科	2.8
S022	理工学科	2.7
S013	理工学科	2.7
S024	理工学科	2.6
S027	理工学科	2.5
S012	理工学科	2.5
S009	理工学科	2.4
S016	理工学科	2.4
S018	理工学科	2.4
S002	理工学科	2.3
S010	理工学科	2.3
S030	理工学科	2.3
S025	理工学科	2.2
S006	理工学科	2.2
S020	理工学科	2.2

【確認３】 どう使う？どう考える？

５つの学科の学力状況 ⇒ 箱ひげ図で分布の様子を考察



データを箱ひげ図に
→　分布の様子が！
||

目 的
×　作図
○　それを使って**判断**
△　１つの分布
○　複数の分布＝比較

概念・本質
分布の様子を見るのに
４分の１ずつの範囲を
考察してみる。

真ん中の50%は
どのあたりに分布する？
端の方は？

比較に最適！

研修内容

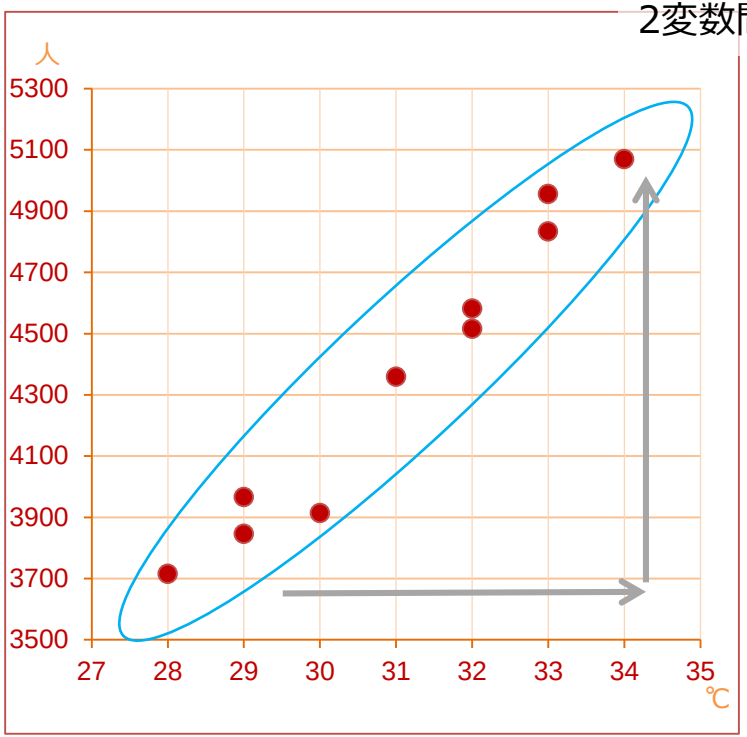
- Part1【講義編】　明日使えるデータ分析
  - 傾向把握と可視化
  - 相関・回帰・予測
- Part2【実習編】　Excelによる実践演習
  - データ分析の準備
  - Excelによる講義編の実践
- まとめ

散布図・相関

最高気温とプールの入場者数

月日	最高気温	入場者数
7月27日	28	3715
7月28日	29	3966
7月29日	29	3846
7月30日	31	4359
7月31日	34	5070
8月1日	32	4582
8月2日	32	4516
8月3日	33	4956
8月4日	33	4833
8月5日	30	3914

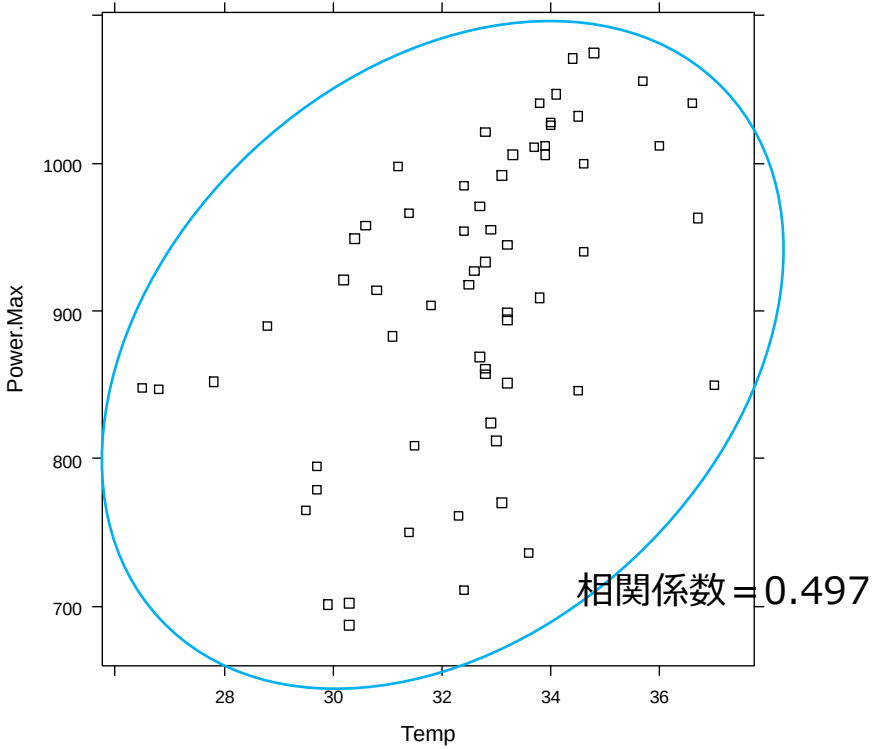
2変数間の関連性を表現



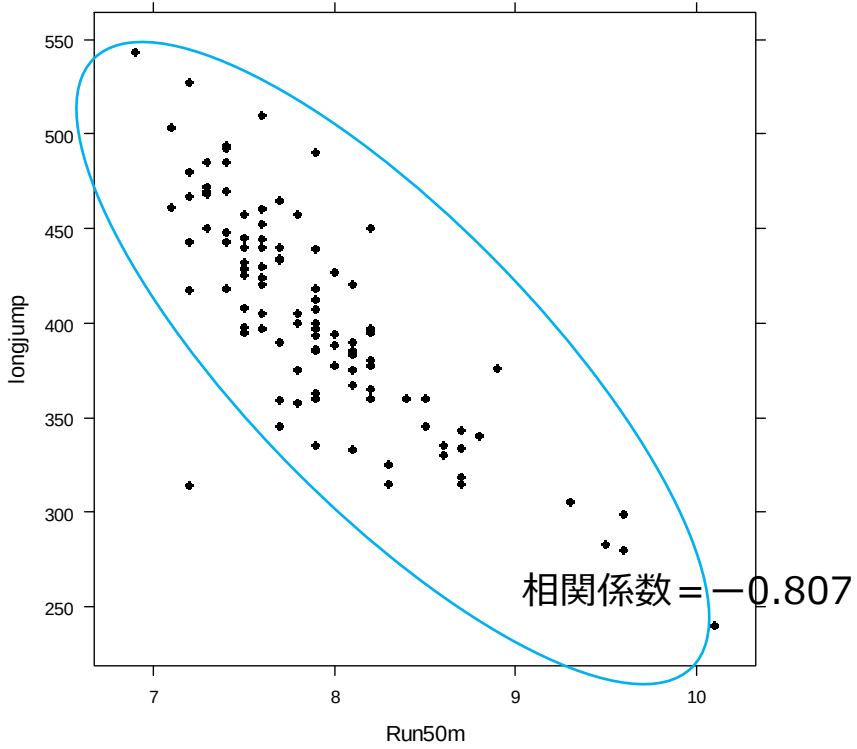
その関連度合い
相関係数＝0.982

散布図・相関

中国地方の最高気温と電力消費量（2017年7～8月）



50m走と走り幅跳びの記録（中学生）



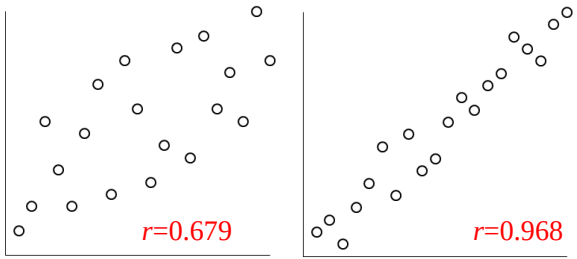
相関係数

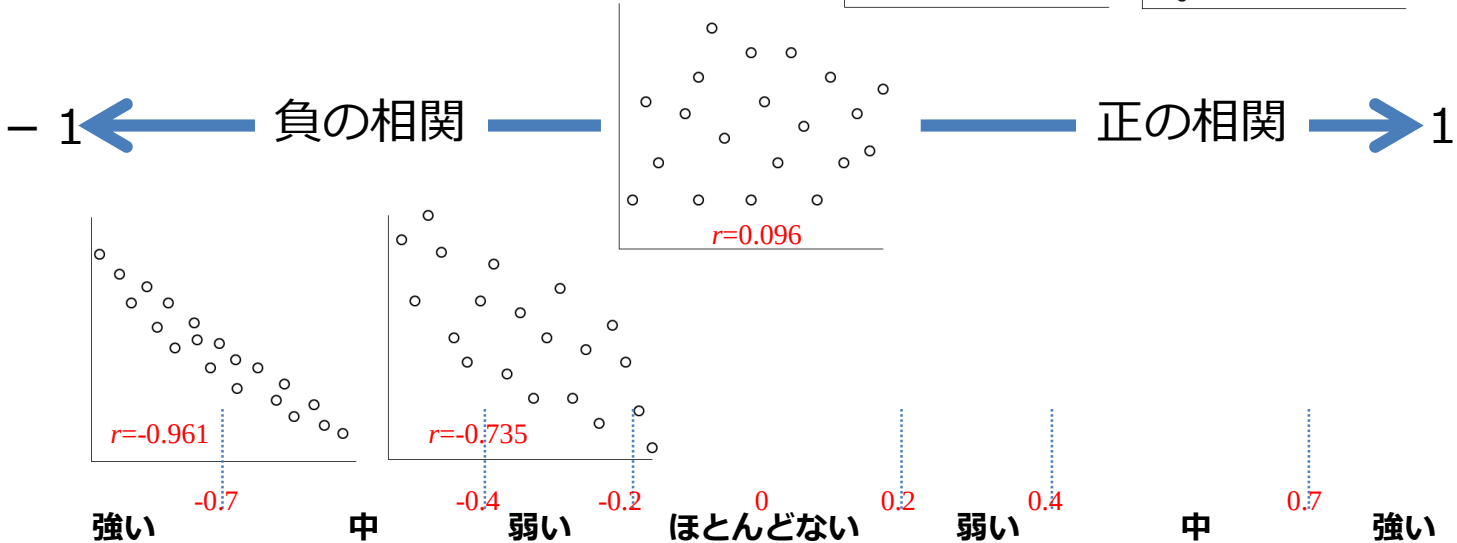
$$r = \frac{s_{xy}}{s_x s_y} = \frac{\text{共分散}}{x\text{の標準偏差} \times y\text{の標準偏差}}$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

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

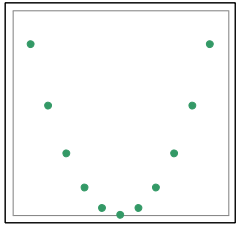
$$s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$



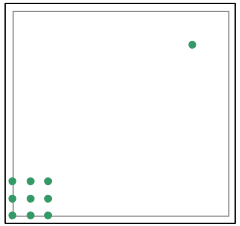


相関係数（留意点）

直線的な関係を示す指標である

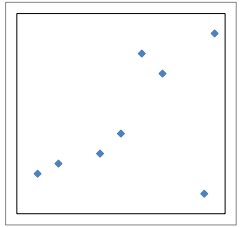


相関係数＝0
しかし，2つの変数には曲線関係がある



相関係数＝0.924
しかし，右上の点に引きずられている

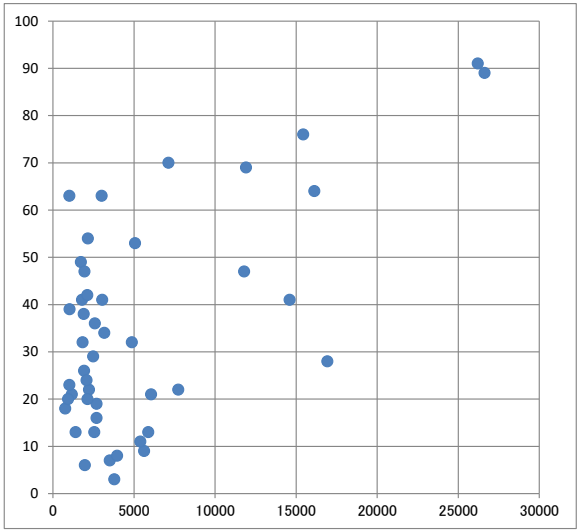
左下の部分のみの相関係数＝0



相関係数＝0.386
しかし，右下の点に引きずられている

右下の１点を除くと相関係数＝0.911

見かけの相関（疑似相関）にだまされるな



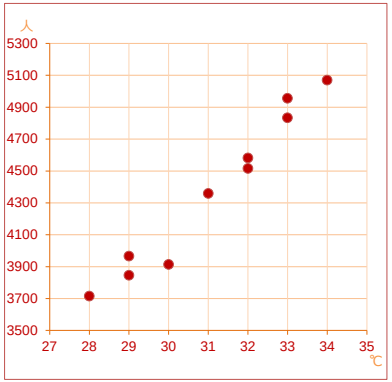
ビールの販売数と水難事故数

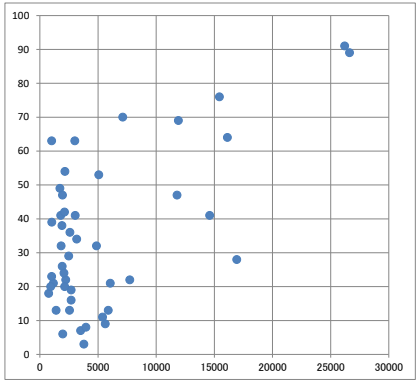
正の相関が見える
||
水難事故を防ぐには，
ビールを売らない。　？？？

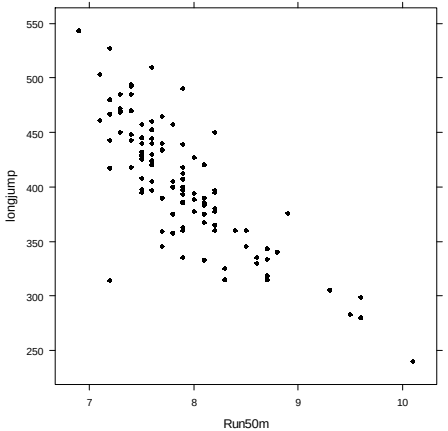
雨の日の交通事故を防ぐには，傘を売らない。

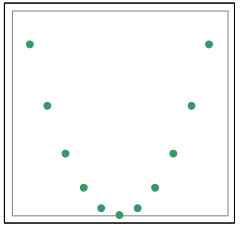
育毛剤を使っている人ほど，10年後にハゲる。

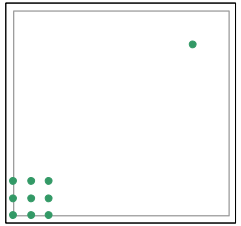
相関係数（留意点）

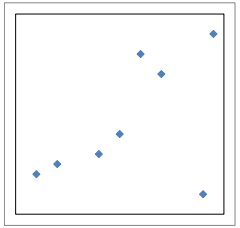












対応する値を散布図に
→　２変数の関係が！
２変数の相関係数を
→　２変数の直線関係が！

目 的
×　作図や計算
○ ２変数の間の**関係**をつかむ

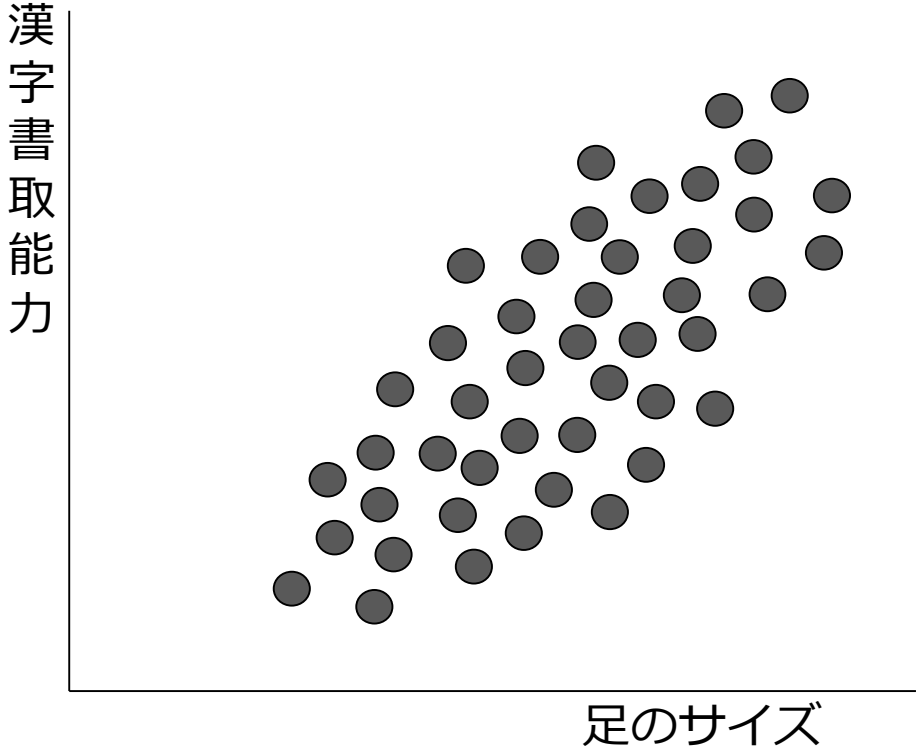
概念・本質
２変数の対応関係をプロットして傾向を考察。
（右上がり／右下がり，
直線関係・・・）

相関係数は関係の指標
ただし，２つの留意点
直線関係しか言っていない
極端な値に引きずられる
必ず，**散布図**と**相関係数**の**両方**を

散布図の考察（留意点）

小学生の足の大きさと漢字書取能力の関係（仮想データ）

足の大きい子ほどかしこい？？



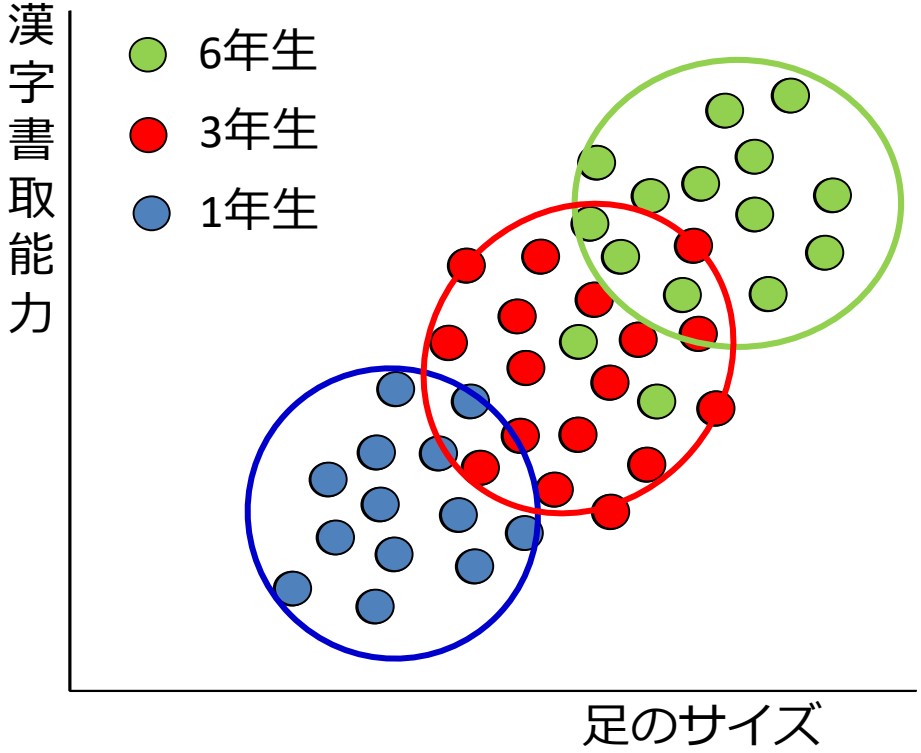
散布図の考察（留意点）

小学生の足の大きさと漢字書取能力の関係（仮想データ）

足の大きい子ほどかしこい？？

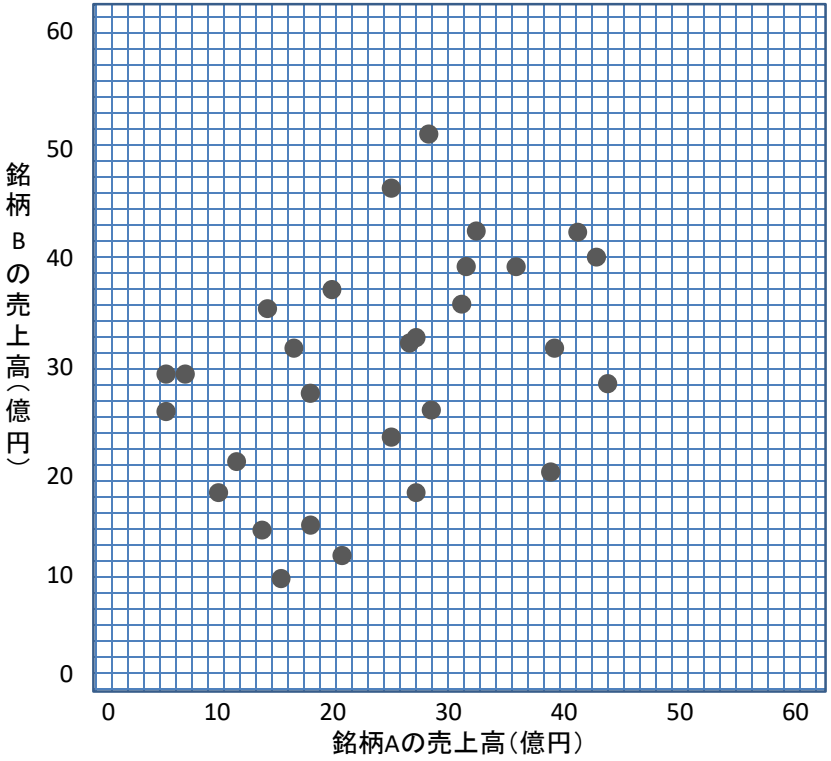
学年で分けてみると

各学年ごとでは2変数間には相関は見られない
小学生＝学年が上がるほど知識が豊富（当然の結果）
↑
学年という第3番目の変数の影響
＝見かけの相関



散布図の考察（留意点）

２つの飲料銘柄の売上高（関西の29都市）



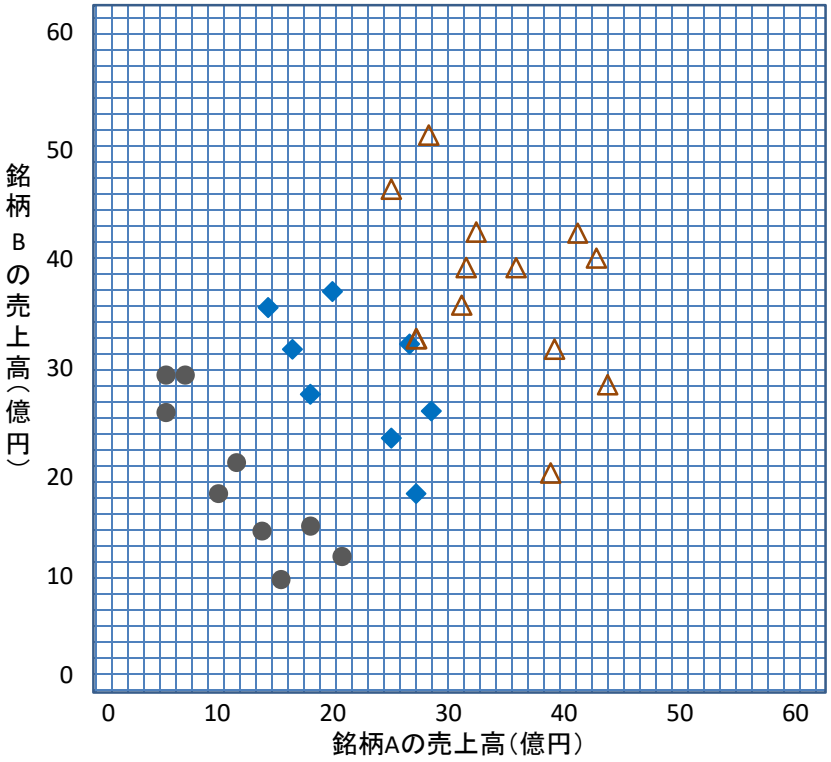
銘柄Aが売れるほど
銘柄Bが売れている。
||
銘柄Aと銘柄Bには
正の相関？？

現場では・・・
両方同時に売れている
感覚はない・・・！！

散布図の考察（留意点）

２つの飲料銘柄の売上高（関西の29都市）

- ● 小都市
- ◆ 中都市
- △ 大都市



銘柄Aが売れると
銘柄Bが売れない。
||
銘柄Aと銘柄Bは競合！
（一方を買えば他方は不要）
||
銘柄Aと銘柄Bには負の相関

対応する値を散布図に
→　２変数の関係が！
||

目 的

○ **グループ**別に描画すると
グループ内，グループ間の
関係もつかめる。

概念・本質

グループ化して同時プロット
すると個別／全体の関係が
わかる。

回帰分析

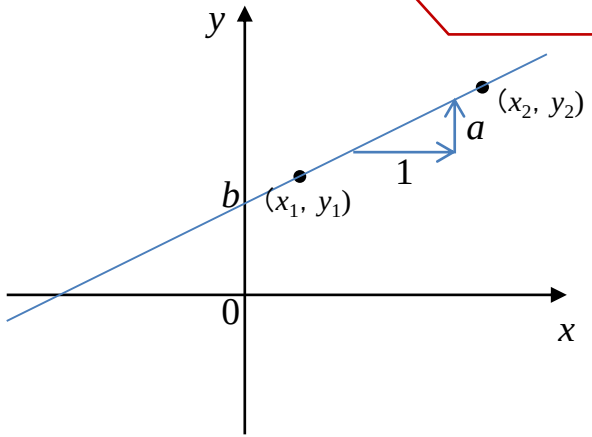
1 次関数　２つの点から式を求める。

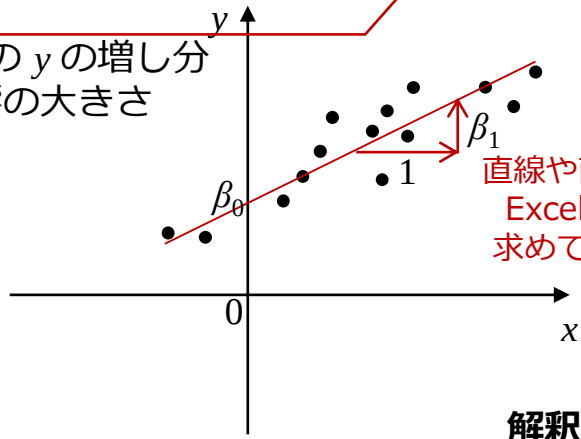
$$y = \underbrace{a}_{\text{傾き}}x + \underbrace{b}_{\text{切片}}$$

$$y = \beta_0 + \beta_1 x$$

x が 0 のときの y の値

x が 1 増えたときの y の増し分
x の y への影響の大きさ





直線や直線の式は
Excel や R が
求めてくれる。

解釈が大事！

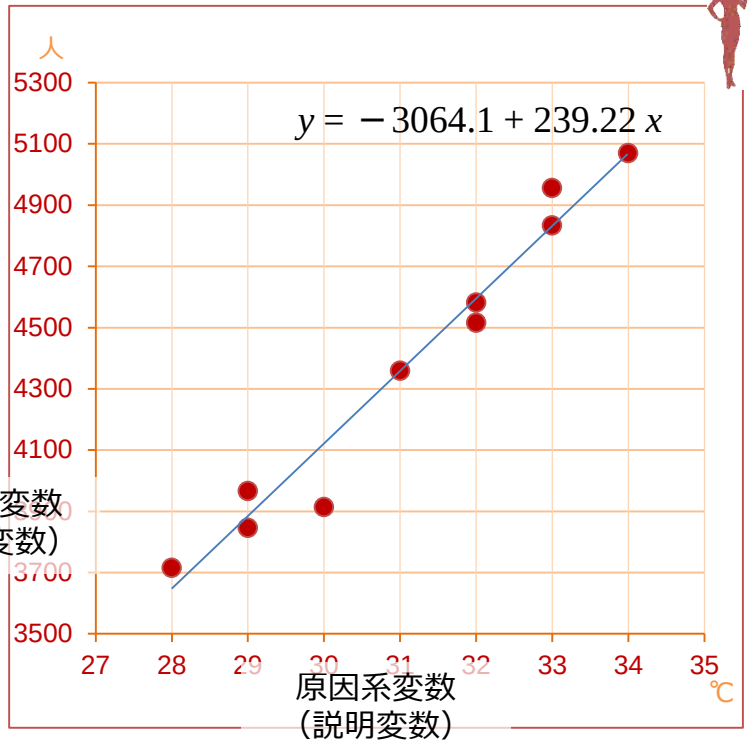
中学校 数学

そういえば，理科の実験！

回帰分析

プールの入場者数予想？

月日	最高気温	入場者数
7月27日	28	3715
7月28日	29	3966
7月29日	29	3846
7月30日	31	4359
7月31日	34	5070
8月1日	32	4582
8月2日	32	4516
8月3日	33	4956
8月4日	33	4833
8月5日	30	3914



結果系変数
（目的変数）

原因系変数
（説明変数）

ほぼ直線
最高気温が1℃上がれば，239人増える
↓
8月6日の予想最高気温が30℃なら，
－3064.1+239.22×30＝4112.5（人）

散布図から回帰直線
→　１つのモデルができた。
→　予測ができる。
||

目 的
△ 数式を求めること
○ 現象の直線傾向を見ること
○ **予測**ができること

概念・本質
x で y を**説明**する。
β_1の値
x が 1 増えると y は・・・
β_1の符号
＋：x は y を増やす方向の影響
－：x は y を減らす方向の影響

重回帰分析

家賃が決まる要因？

	y	x_1	x_2	x_3
Ｎｏ	家賃	面積	築年数	駅からの時間（分）
1	50,000	21.3	12	15
2	78,000	23.5	16	12
3	60,000	30.5	14	20
⋮	⋮	⋮	⋮	⋮

$$y = 50314 + 1120\,x_1 - 456\,x_2 - 995\,x_3$$

家賃 ＝ 50,314 ＋ 1,120×面積 － 456×築年数 － 995×駅からの時間

面積が1 m^{2}広くなると家賃は1,120円高くなる

建物が1年古くなると456円安くなる

駅からの距離が1分遠くなると995円安くなる

係数の値はそのまま評価しないように。

各変数の単位（大きさ）が異なるから。

評価する場合は標準化した値（t 値）を見るようにする。

重回帰分析（予測）

	branch	通行人数 numpass	駅からの時間 minutes	店舗面積 area	駐車台数 parkcar	従業員数 numwork	品　数 kinds	売上高 sales
1	三条	716	16	44	16	7	125	78
2	京都南	2018	30	25	8	3	132	34
3	長岡京	1880	3	68	18	10	110	145
4	生駒	1416	20	30	10	5	70	51
5	高槻	904	10	67	27	10	82	98
6	枚方	1250	2	66	10	10	82	115
7	池田	1039	15	52	15	7	82	75
8	東大阪	2394	1	113	50	20	125	258
9	堺	711	12	30	12	7	102	70
10	八尾	738	10	39	10	7	70	65
11	和歌山	1322	11	60	23	4	72	82
12	宝塚	813	12	34	10	3	97	32
13	西宮	1733	3	96	40	10	145	190
14	西神	1569	5	55	28	10	92	168
15	加古川	1770	6	80	32	8	80	195

新しい店舗を建てたい。最も高い売り上げが期待できる店舗を最初に建てたい。
そのために，既存チェーン店15店舗の売上高と6つの調査観点に関するデータを用いて，3つの候補の売り上げを予測せよ。

↓

	係数	標準誤差	t	P-値
切片	36.873	32.993	1.1176	0.2962
numpass	0.0368	0.0134	2.738	0.0255
minutes	-3.343	1.27	-2.633	0.0301
area	-0.297	0.6726	-0.442	0.6703
parkcar	2.4479	1.0027	2.4413	0.0405
numwork	2.7581	2.2167	1.2442	0.2486
kinds	0.0302	0.2397	0.1261	0.9027

↓

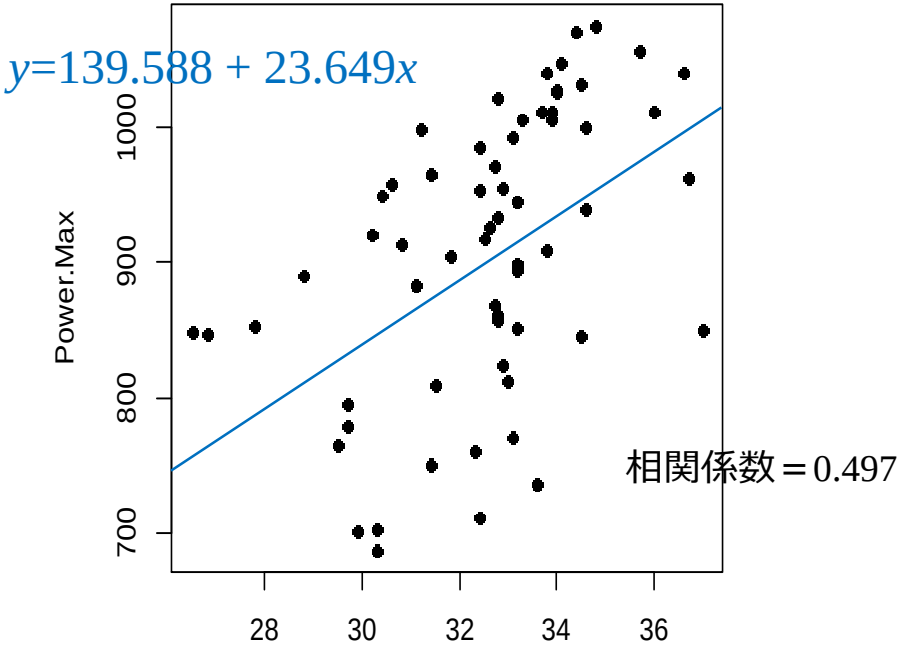
$$y = 36.9 + 0.037\,x_1 - 3.34x_2 - 0.30\,x_3 - 2.45\,x_4 + 2.76\,x_5 + 0.03\,x_6$$

支店名	通行人	駅からの時間	店舗面積	駐車台数	従業員数	品数	
候補1	1956	3	88	42	10	120	206.6
候補2	1300	12	90	45	10	100	158.6
候補3	1423	8	42	36	10	90	168.4

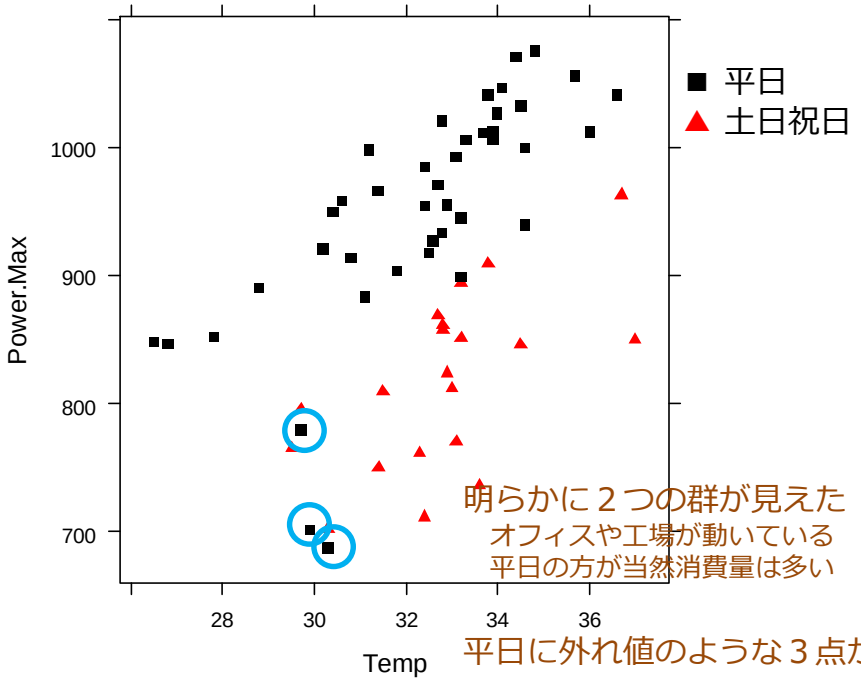
回帰分析でも層別

中国地方の最高気温と電力消費量（2017年7～8月）

最高気温をx，電力消費量をyにして回帰分析を行うと



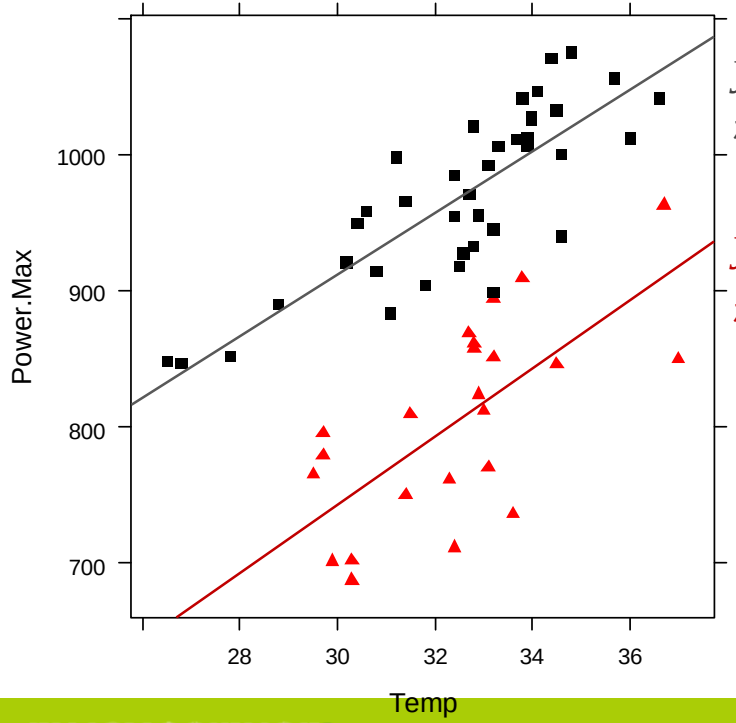
平日と土日祝日は異なるのでは → 曜日で層別



回帰分析でも層別

中国地方の最高気温と電力消費量（2017年7～8月）

お盆（8/14，15，16）も土日祝日に

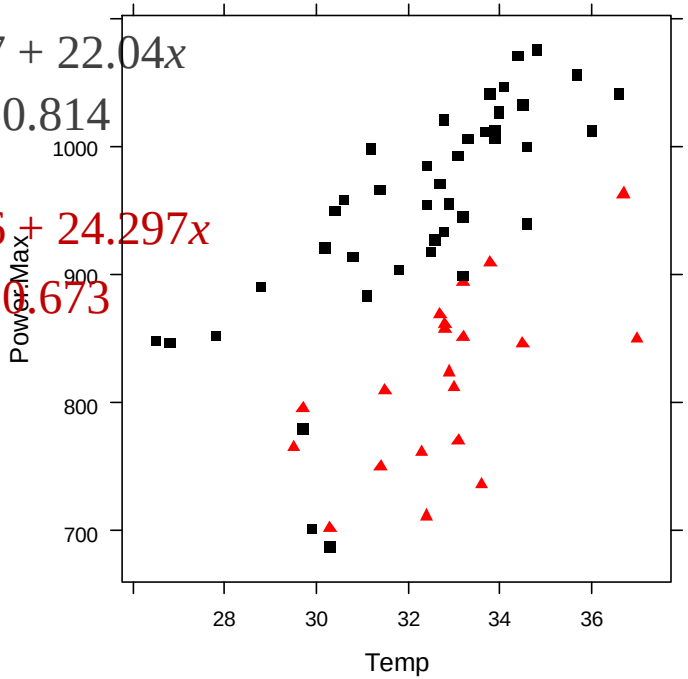


y=252.37 + 22.04x

相関係数=0.814

y=16.046 + 24.297x

相関係数=0.673



相関・回帰・因果関係

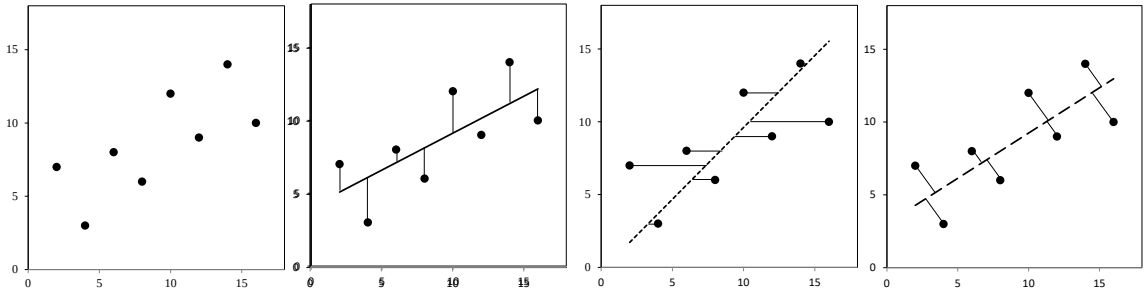
相関と回帰

相関：２つの変量が対等である。

回帰：方向がある。

一方(x)が他方(y)を説明する。

→ 因果の説明につながる。



相関と因果

因果関係があれば，相関はみられる。

相関がみられたからといって
因果関係があるわけではない。

相 関

Y ⇄ X

因 果

Y ← X

疑似相関

Y ⇆ X

↖ ↗

Z