

大学評価・IR
担当者のための

統計解析はじめの一步

藤野友和（福岡女子大学）

大学評価・IR担当者のための初歩的な統計講座

本日の内容

- 1 データの形式と種類（尺度水準）
- 2 分布を調べる（ヒストグラム）
- 3 中心を表す指標（平均値と中央値）
- 4 分布を調べる（箱ひげ図）
- 5 ばらつきを表す指標（分散と標準偏差）
- 6 関連性を調べる（散布図・相関係数）

本日の内容

- 1 データの形式と種類（尺度水準）
- 2 分布を調べる（ヒストグラム）
- 3 中心を表す指標（平均値と中央値）
- 4 分布を調べる（箱ひげ図）
- 5 ばらつきを表す指標（分散と標準偏差）
- 6 関連性を調べる（散布図・相関係数）

データの基本形式

変数								
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

個体

データの種類（尺度水準）

名義尺度

同じものには同じ値（記号）
異なるものには異なる値（記号）



順序尺度

名義尺度 + 順序関係



間隔尺度

値の間隔（差）に意味がある

10℃ → 30℃

- 20℃上昇した！
- ✕ 温度が3倍になった！



比例尺度

間隔尺度 + 値の比に意味がある



変数の種類

質的変数

名義尺度 順序尺度

量的変数

間隔尺度 比例尺度

ワークシート1

スマートフォンをお持ちの方は、
以下の検索で表示されるサイトに入って、
Room nameを入力してください。

socrative student



Room name : DBAA51AA

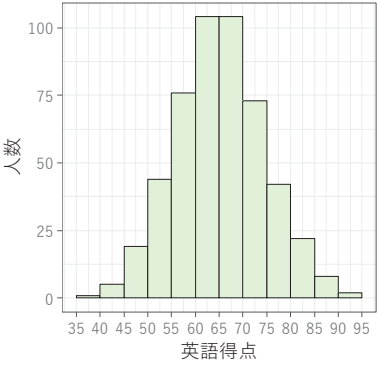
本日の内容

- 1 データの形式と種類（尺度水準）
- 2 分布を調べる（ヒストグラム）
- 3 中心を表す指標（平均値と中央値）
- 4 分布を調べる（箱ひげ図）
- 5 ばらつきを表す指標（分散と標準偏差）
- 6 関連性を調べる（散布図・相関係数）

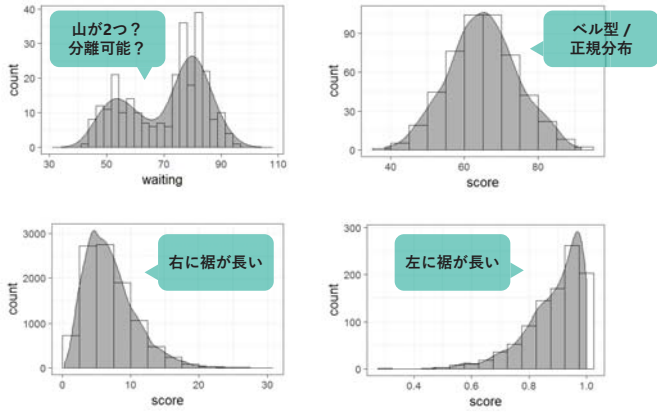
量的変数の値の分布を調べる

度数分布表 & ヒストグラム

階級	階級値	度数
35 ～ 40	37.5	1
40 ～ 45	42.5	5
45 ～ 50	47.5	19
50 ～ 55	52.5	44
55 ～ 60	57.5	76
60 ～ 65	62.5	104
65 ～ 70	67.5	104
70 ～ 75	72.5	73
75 ～ 80	77.5	42
80 ～ 85	82.5	22
85 ～ 90	87.5	8
90 ～ 95	92.5	2



ヒストグラムのチェックポイント



本日の内容

- 1 データの形式と種類（尺度水準）
- 2 分布を調べる（ヒストグラム）
- 3 中心を表す指標（平均値と中央値）
- 4 分布を調べる（箱ひげ図）
- 5 ばらつきを表す指標（分散と標準偏差）
- 6 関連性を調べる（散布図・相関係数）

量的変数の中心を示す指標

平均値

値をすべて足し合わせて、値の個数で割る

7名の勉強時間の平均値 = $\frac{1}{7}(8.5 + 2.5 + 6.5 + 9.0 + 4.0 + 3.5 + 1.5) = 5.1$

中央値

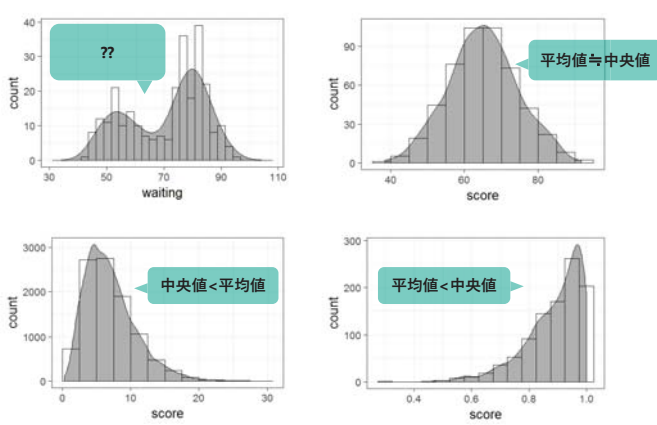
値を昇順にならべたとき、真ん中にくる値

※ 値の個数が偶数個の場合は、真ん中2つの値の平均値

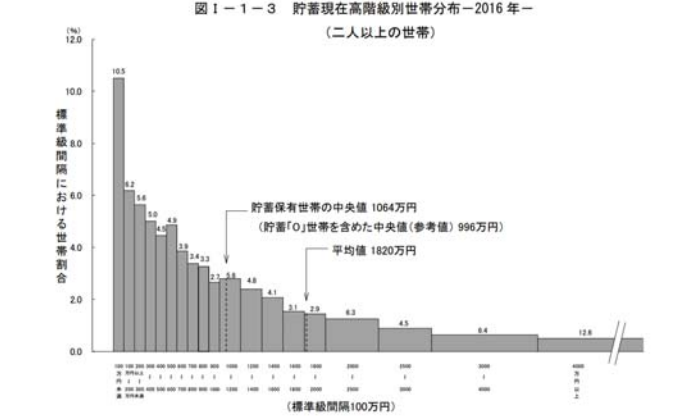
平均値と中央値の性質

	■ 平均値 ■	■ 中央値 ■
○	<ul style="list-style-type: none">計算が容易 / イメージしやすい（平らに均した値）よい性質を持っている性質について多くのことが分かっている	<ul style="list-style-type: none">外れ値に影響を受けにくいデータの分布によらず、常に中央値の上下それぞれに半数のデータを含む
✕	<ul style="list-style-type: none">外れ値に影響を受けやすい右や左に裾の長い分布では代表の値としてふさわしくない	<ul style="list-style-type: none">計算が面倒（並べ替えが必要）理論的には平均値ほど扱われにくい

分布の形と平均値・中央値



日本における貯蓄額の分布

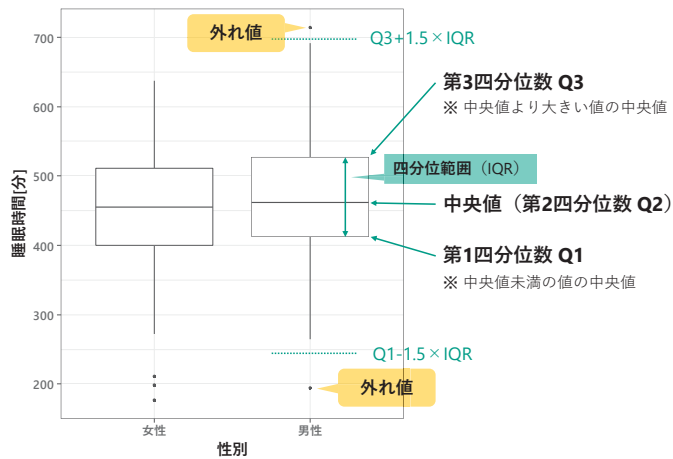


※ 総務省 家計調査報告（貯蓄・負債編）平成28年（2016年）平均結果速報（二人以上の世帯）より

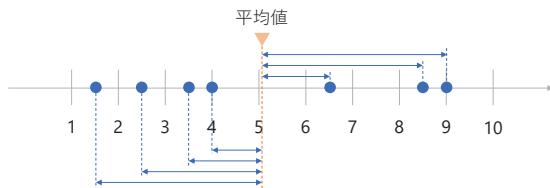
本日の内容

- 1 データの形式と種類（尺度水準）
- 2 分布を調べる（ヒストグラム）
- 3 中心を表す指標（平均値と中央値）
- 4 分布を調べる（箱ひげ図）
- 5 ばらつきを表す指標（分散と標準偏差）
- 6 関連性を調べる（散布図・相関係数）

量的変数の分布を比べる → 箱ひげ図



散らばりの指標



$$\text{標準偏差} = \sqrt{\text{分散}}$$

※ 分散

すべての $\leftarrow \rightarrow$ を2乗して合計した値をデータの個数で割った値

このデータの場合 標準偏差 ≈ 2.7

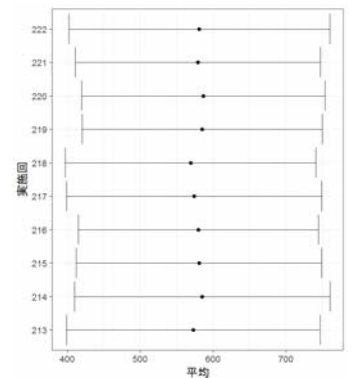
平均値を用いて、 5.1 ± 2.7 などと表記される

本日の内容

- 1 データの形式と種類 (尺度水準)
- 2 分布を調べる (ヒストグラム)
- 3 中心を表す指標 (平均値と中央値)
- 4 分布を調べる (箱ひげ図)
- 5 ばらつきを表す指標 (分散と標準偏差)
- 6 関連性を調べる (散布図・相関係数)

TOEICの得点分布

実施回	平均	標準偏差
222	581.2	179.1
221	579.4	168.4
220	586.8	167.2
219	585	164.8
218	569.2	171.9
217	574.3	175.4
216	580	165
215	580.8	168.5
214	585.2	175.7
213	572.9	174.4

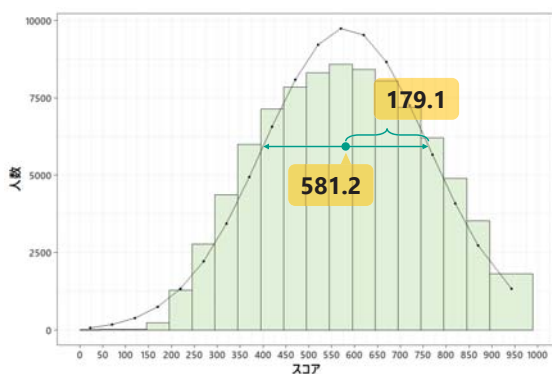


TOEIC 公開テスト 平均スコア・スコア分布一覧より
http://www.iibc-global.org/toEIC/official_data/lr/data_avelist.html

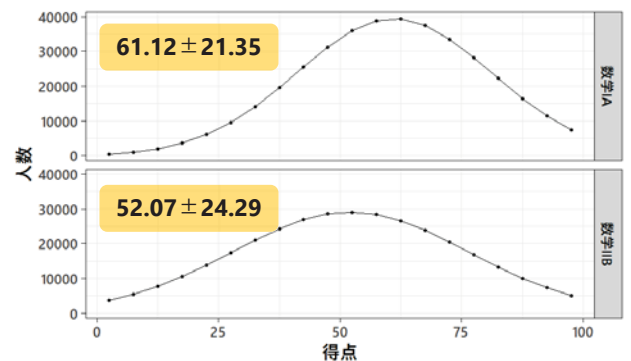
第222回のヒストグラム

581.2 ± 179.1

正規分布であれば
 おおよそ**68%**のデータがこの範囲に入る



H29センター試験

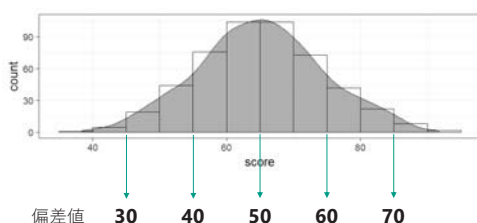


偏差値

ある試験を受験したAさんの偏差値

$$50 + 10 \left(\frac{\text{Aさんの得点} - \text{平均点}}{\text{標準偏差}} \right)$$

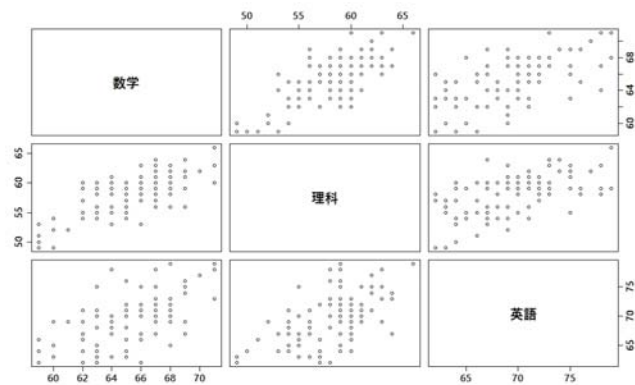
平均点が**50**点、標準偏差が**10**点になるように調整した得点



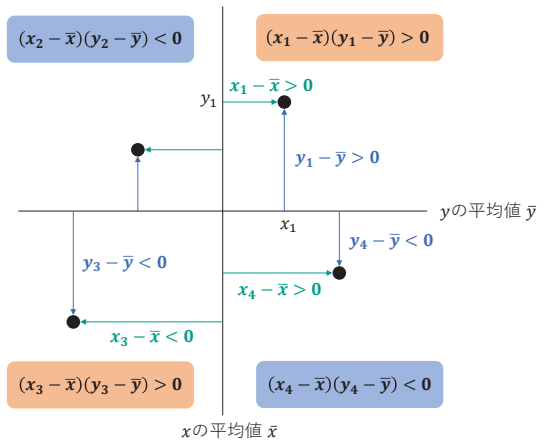
本日の内容

- 1 データの形式と種類 (尺度水準)
- 2 分布を調べる (ヒストグラム)
- 3 中心を表す指標 (平均値と中央値)
- 4 分布を調べる (箱ひげ図)
- 5 ばらつきを表す指標 (分散と標準偏差)
- 6 関連性を調べる (散布図・相関係数)

散布図 2変数の関連を視覚的に捉える

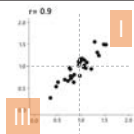


共分散 2変数の関連を数値で捉える



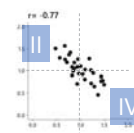
共分散

すべてのデータ点についての $(x_i - \bar{x})(y_i - \bar{y})$ の平均値



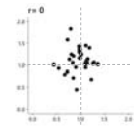
IとIIIにたくさん点があるとき
(右上がりの直線状にデータが分布するとき)

共分散 → +



IIとIVにたくさん点があるとき
(右下がりの直線状にデータが分布するとき)

共分散 → -

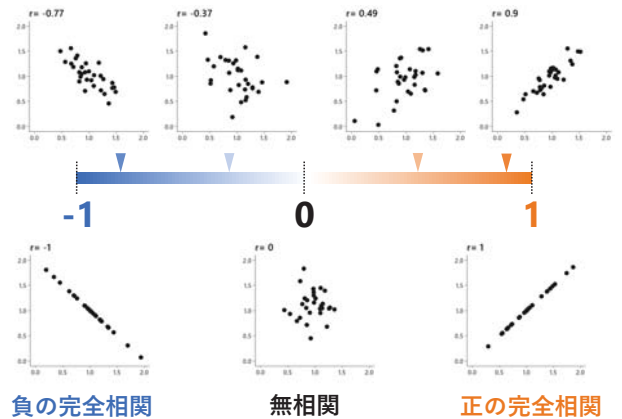


I~IVに同じくらいの点があるとき

共分散 → 0

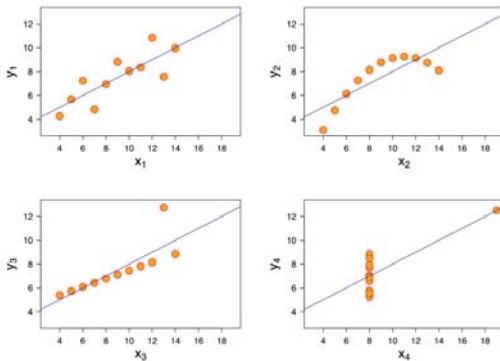
相関係数

共分散をxとyの標準偏差の積で割り算し、-1~1の値をとるよう調整した値



数値だけに頼らない

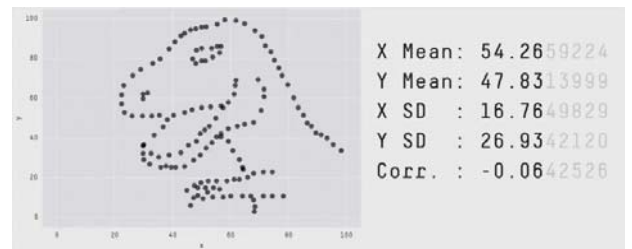
相関係数、平均値、標準偏差すべて同じ！



Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17-21

数値だけに頼らない

相関係数はすべてゼロ！

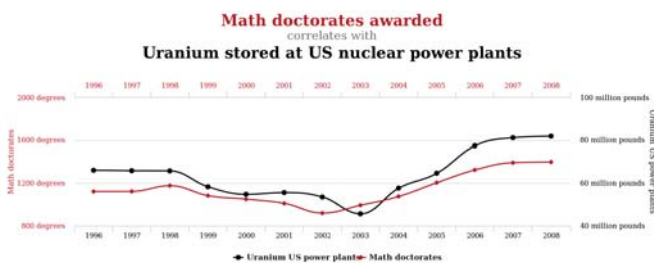


Justin Matejka, George Fitzmaurice (2017)
Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing
CHI 2017 Conference proceedings:
ACM SIGCHI Conference on Human Factors in Computing Systems

見かけ上の相関 / 相関と因果

数学の博士号取得者数 vs アメリカの原発に備蓄されているウランの量

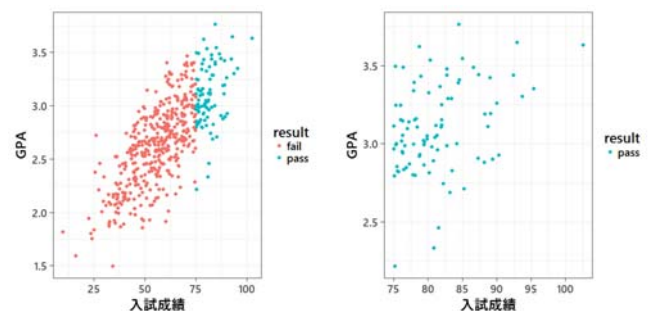
相関係数=0.95



<http://tylervigen.com/spurious-correlations>

選抜効果

入学試験の成績と入学後の成績の相関

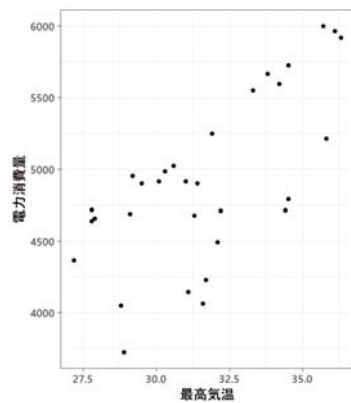


相関係数 = 0.75

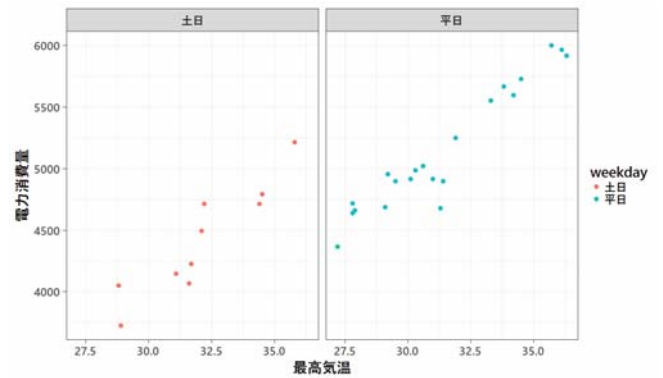
相関係数 = 0.35

層別

東京都の日ごとの最高気温と電力消費量（2010年7月）



層別



まとめ

- 1 データの形式と種類（尺度水準）
- 2 分布を調べる（ヒストグラム）
- 3 中心を表す指標（平均値と中央値）
- 4 分布を調べる（箱ひげ図）
- 5 ばらつきを表す指標（分散と標準偏差）
- 6 関連性を調べる（散布図・相関係数）

本日取り扱っていないこと

- 母集団と標本
- 各種統計グラフ（棒グラフ、円グラフなど）
- 質的変数の取り扱い
- 確率
- 時系列データ