

第7回 回帰分析の考え方 計量経済学(1)

@東京工業大学 倪彬

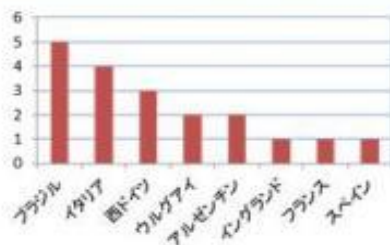
今日の予定

- ▶ 統計学の基礎(の復習)
 - ▶ データを視覚的に表す方法
 - ▶ 記述統計
 - ▶ データの真ん中
 - ▶ データの散らばり
 - ▶ 2種類のデータの関係
- ▶ 回帰分析とは？
 - ▶ 散布図、相関係数から回帰分析へ
 - ▶ 回帰分析と一次関数

データを視覚的に表す

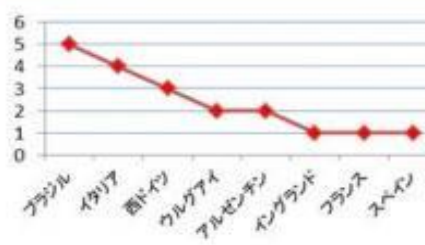
棒グラフ

数の大小を比べるグラフ



折れ線グラフ

時間の経過による変化を比べるグラフ



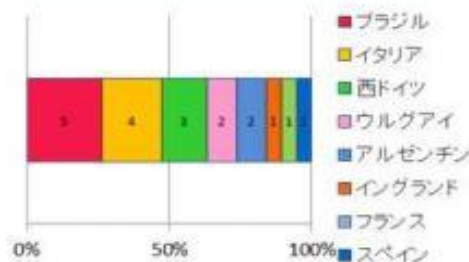
円グラフ

全体に占める構成比(割合)を表すグラフ



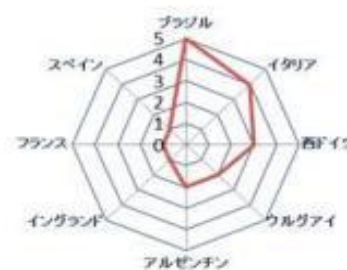
帯グラフ

構成比(割合)を表し、複数のグラフを並べ構成比の変化を比べるグラフ



レーダーチャート

複数の指標をまとめてみるグラフ
項目は5つから6つであることが条件



Active learning(1)

- ▶ 以下の場合、どのグラフで表すのが適切か答えなさい

データ	グラフの種類
これまでの経済の動きを掴むため、GDPや株価の推移を掴みたい	
自分の毎月の飲み会の回数と交遊費との関係が知りたい	
1000人受験した試験の点数の全体的な傾向を掴みたい	

- ▶ 適切なグラフは1つだけとは限らないので、複数答えてもよい

記述統計

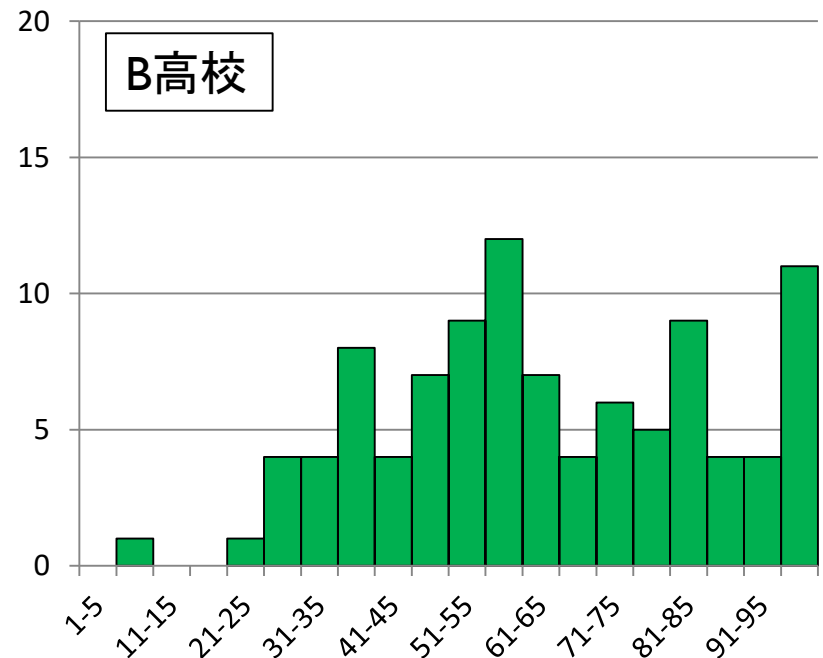
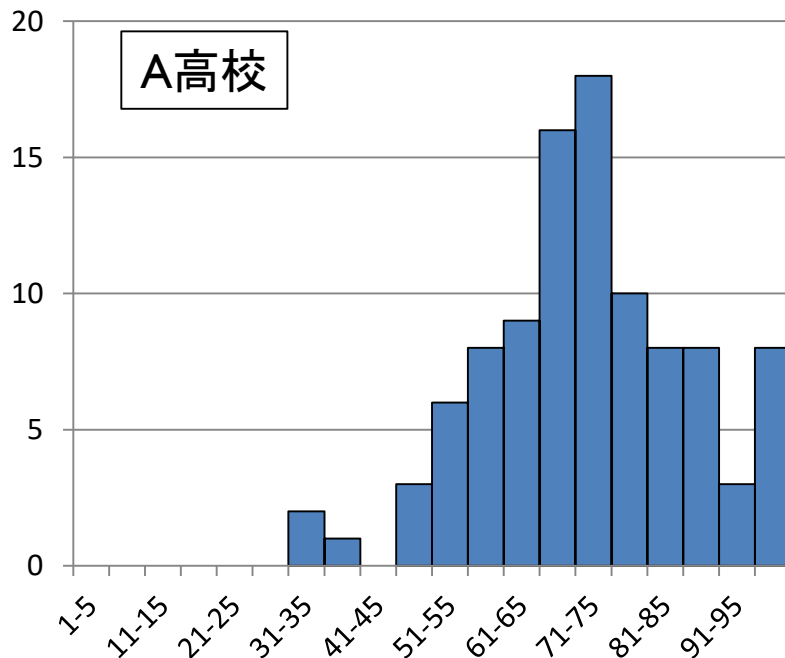
- ▶ データの全体の特徴を様々な視点から数値で確認することも便利

→ 記述統計

- ▶ どのような特徴を見ることが便利か？
 - ▶ そのデータの**代表的な値**(真ん中、一番頻度が高い値、etc.)
 - ▶ そのデータが**どれくらい散らばっているか**

ヒストグラムと記述統計の関係

- ▶ ある高校(A,B,どちらも100人)のテストの点数を比較して特徴を考えたい
 - ▶ その高校の真ん中の学生の点数は？
 - ▶ みんな同じくらいの点数なのか？それとも上下の差が大きいのか？



多数のデータをまとめる：代表値

- ▶ 複数の観測値の代表的な値：代表値
- ▶ どのような考え方で代表するか？
 - ▶ 真ん中：平均値、中央値
 - ▶ 最も頻度が高い値：最頻値

- ▶ 平均値($\bar{X}$) = $\frac{\sum X_i}{n}$ (全ての値を足して、データの数で割る)
 - i : 個体を表す記号 X_i : 各観測値
(X_1 ...1さんのXの値、 X_2 ...2さんのXの値)
 - ▶ これは一番一般的な算術平均
 - ▶ 加重平均や幾何平均などもある

多数のデータをまとめる：代表値

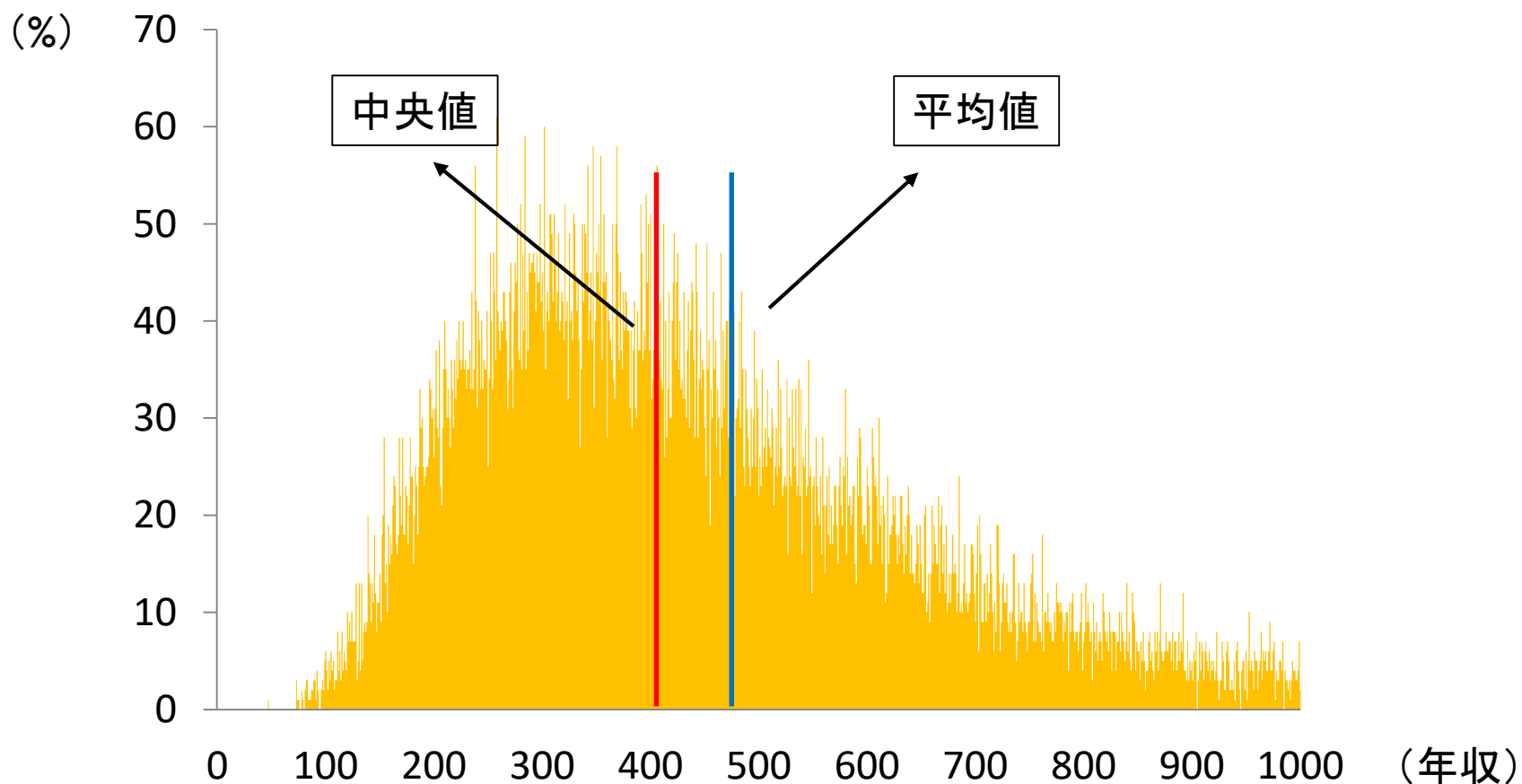
- ▶ 平均値は「データの中心」としては実はイマイチ！
 - ▶ 異常値に引っ張られる
 - ▶ 離散データの場合、計算結果が小数点になると意味不明
- ▶ 中央値：大きい(小さい)順に並べて、全体の中心の人の値
 - ▶ メディアン(median)と言われる場合もある



- ▶ 最頻値：最も多くの人がとっている値
 - ▶ 離散の場合に便利、連続の場合はほぼ使えない
 - ▶ モード(mode)と言われる場合もある

適切な代表値

- ▶ 調べたいデータ次第で適切な代表値は違う



適切な代表値

- ▶ 平均値: 463万円
 - ▶ 中央値: 408万円
 - ▶ 代表的な人の所得がどれくらいなのかを調べたい
 - ▶ 一部の高所得の人(数千万、数億！)のせいで、平均値が過大に計算される
- この場合は平均値よりも中央値の方が望ましい

Active learning(2)

- ▶ 以下のデータについて、平均値、中央値、最頻値を求めなさい。

平均値()

学生10人の履修登録科目数データ				
9	10	10	11	10
30	9	9	10	24

中央値()

最頻値()

- ▶ 上記の3つのうち、代表値として何を提示するのが適切か、自分の意見と共に答えなさい。

多数のデータの特徴：散らばり具合

- ▶ 複数の観測値の散らばり具合を知りたい
 - ▶ 一つの考え方は、中心(平均)からどれくらい離れているか
 - ▶ 各観測値の平均からの差：偏差(へんさ)

偏差＝各観測値－平均値

$$X_i - \bar{X}$$

平均70点の場合		
	点数 X_i	偏差 ($X_i - \bar{X}$)
Aさん	85点	15点
Bさん	40点	-30点
Cさん	75点	5点
Dさん	68点	-2点

散らばり具合

- ▶ 偏差は各観測値(個人ごと)に計算され、そのままではわかりづらい
- 複数の偏差をまとめて、そのグループの散らばり具合を1つの数字にしたい

※ 合計や平均では、答えは必ずゼロになってしまう

$$\frac{\sum (X_i - \bar{X})}{n} = \frac{\sum X_i}{n} - \frac{\sum \bar{X}}{n} = 0$$

偏差の平均

平均

平均(同じ値)の平均
→ 平均

散らばりの指標：分散

- ▶ 解決方法：偏差を「2乗して」平均する：分散

$$\frac{\sum (X_i - \bar{X})^2}{n - 1}$$

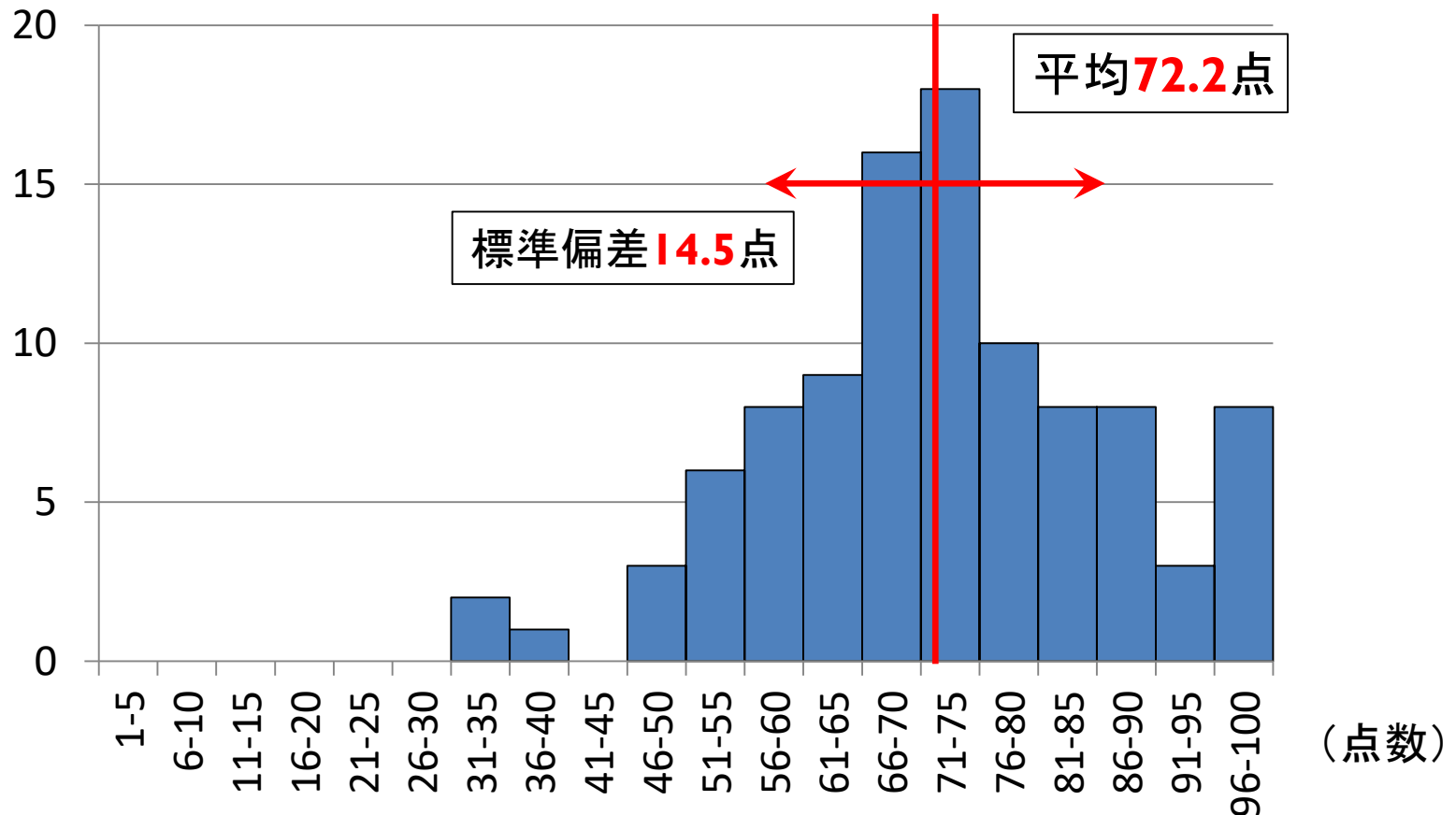
平均のように n で割るのではなく、 $n - 1$ で割る

- ※ 年収のように、桁数が多い場合は、2乗すると膨大な桁数になって使いづらい
- ▶ 桁数を合わせるために、平方根(ルート)をとる
- 標準偏差

$$\sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}}$$

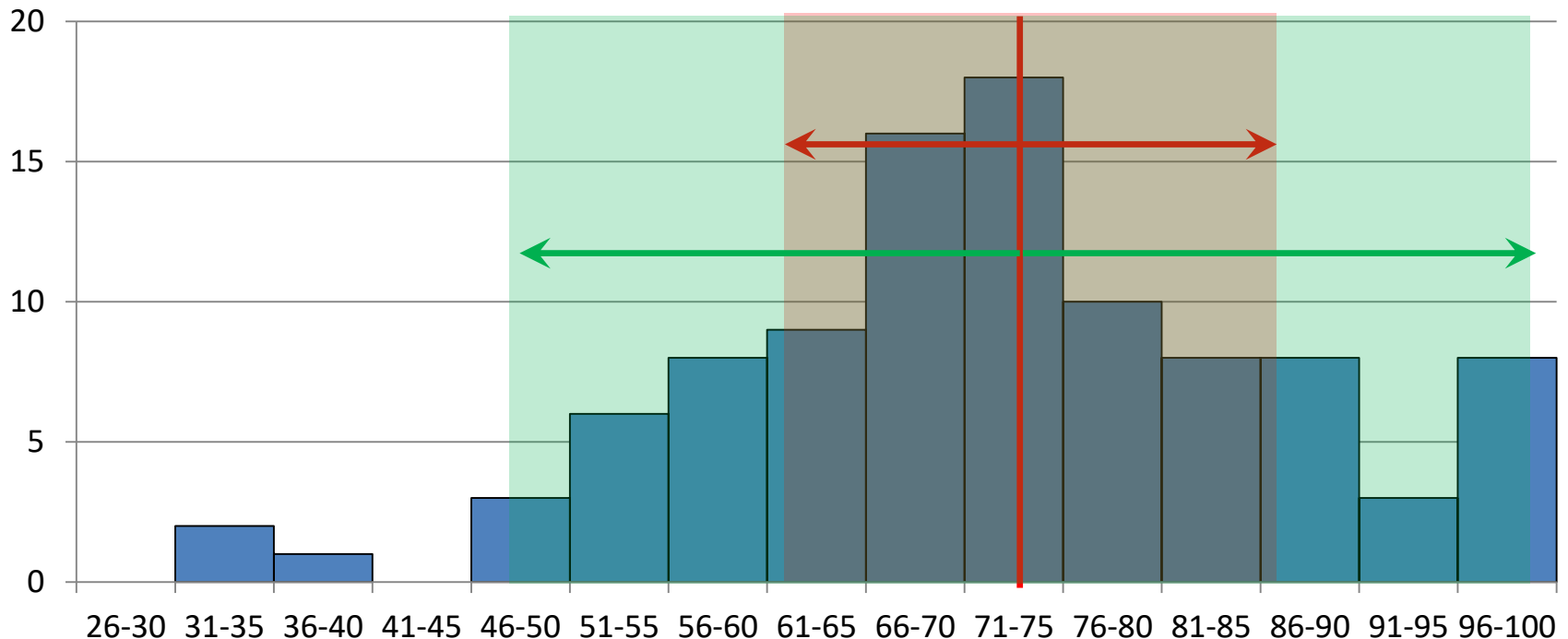
度数分布と平均・標準偏差

- ▶ 平均と分散から、度数分布を数字で補足できる



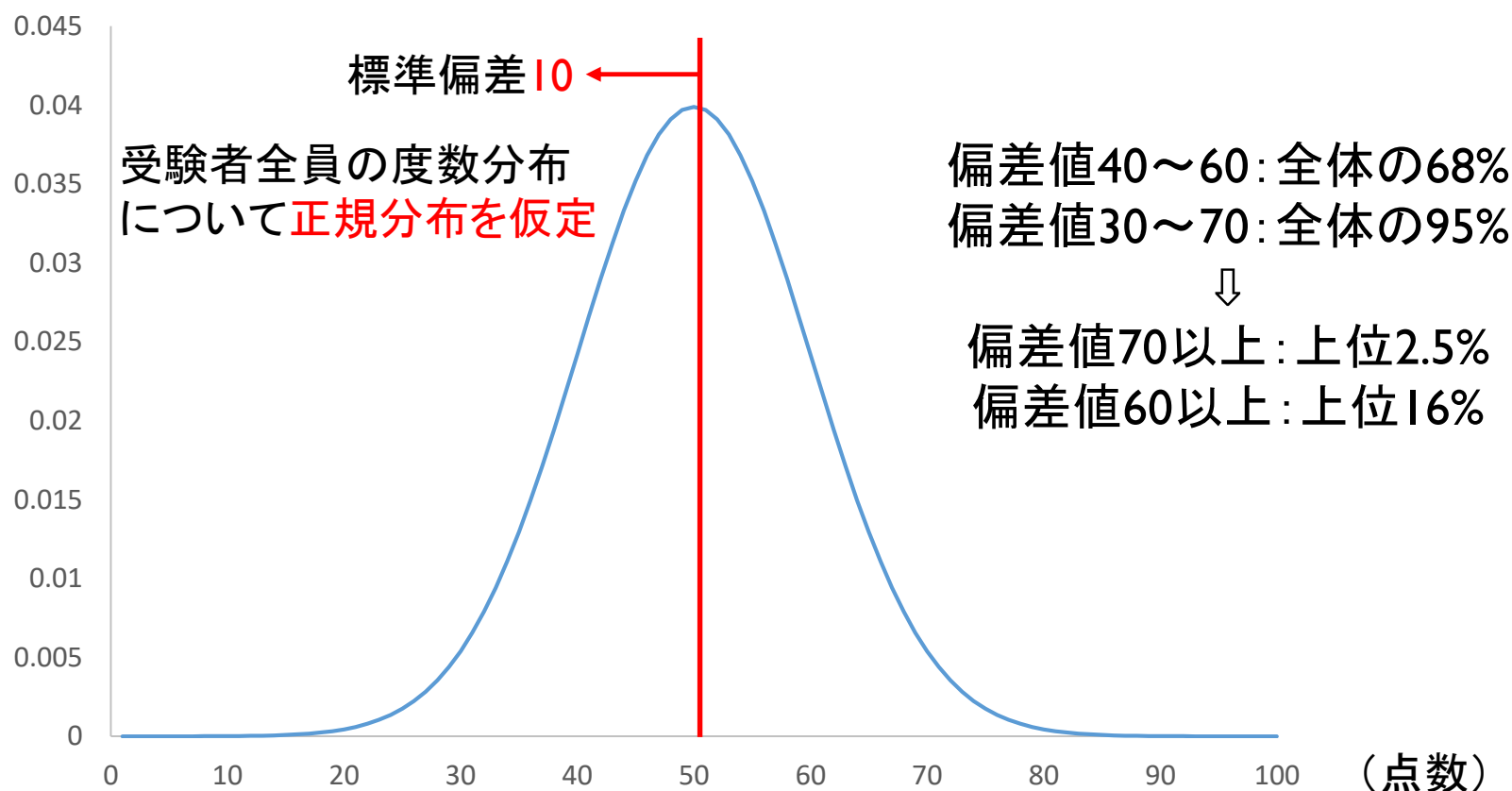
標準偏差の意味

- ▶ 標準偏差はどんな意味・特徴を持つ？
 - ▶ 平均から両側に標準偏差 $\times 1$ 中に、全データのうち68%が含まれる
 - ▶ 平均から両側に標準偏差 $\times 2$ 中に、全データのうちの95%が含まれる
- 標準偏差(分散)が大きいほど、散らばりが大きい



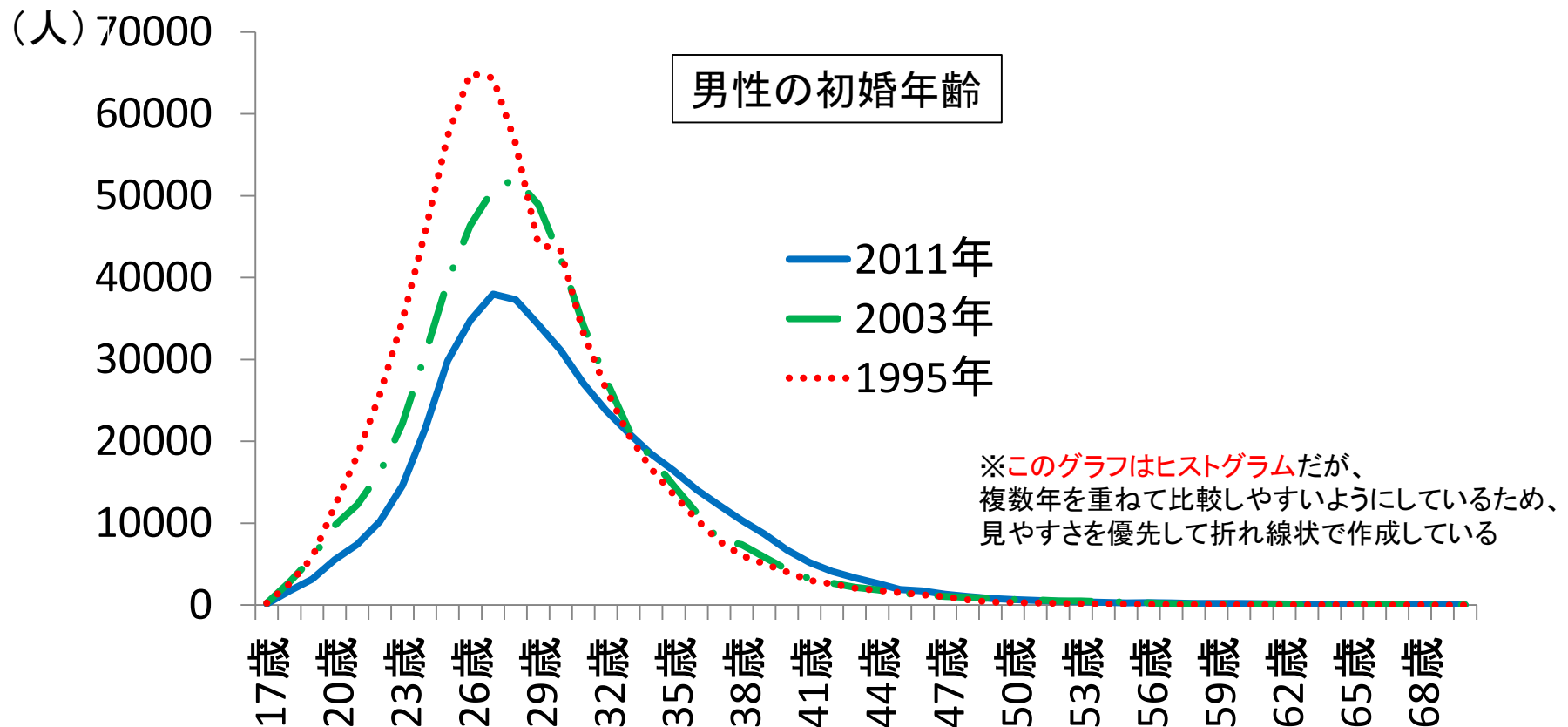
参考：偏差値

- ▶ 偏差値は、度数分布と平均・標準偏差を利用して、自分が受験者全体のどのくらいにいるのかを示す指標

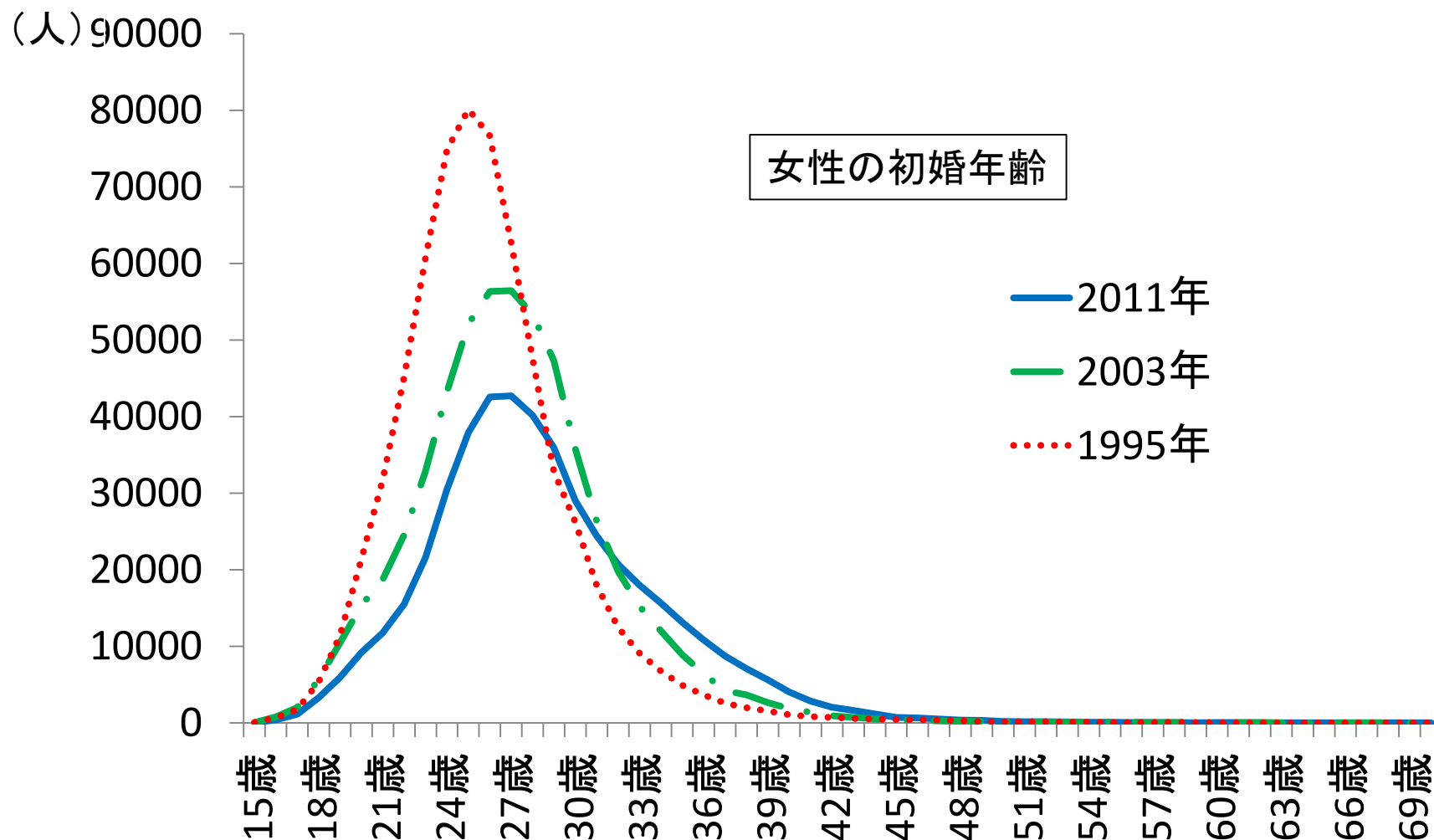


Active learning(3)

- ▶ 日本人の結婚について、「人口動態統計」(厚生労働省)を用い、以下のようなグラフと表を作成した。



Active learning(3)



2変数の関係

- ▶ ここまで見てきたのは、1変数の特徴について
→ より興味があるのは、**2変数の関係**
- ▶ どのようなものが関係ありそう？
 - ▶ 身長と体重
 - ▶ 年収と教育年数(学歴)
- ▶ 例に、「**パン屋さんの人気アンケート**」を使う
 - ▶ 実際のアンケートを使用(少し改変)
 - ▶ 1996年、日本流通新聞による調査

人気アンケート

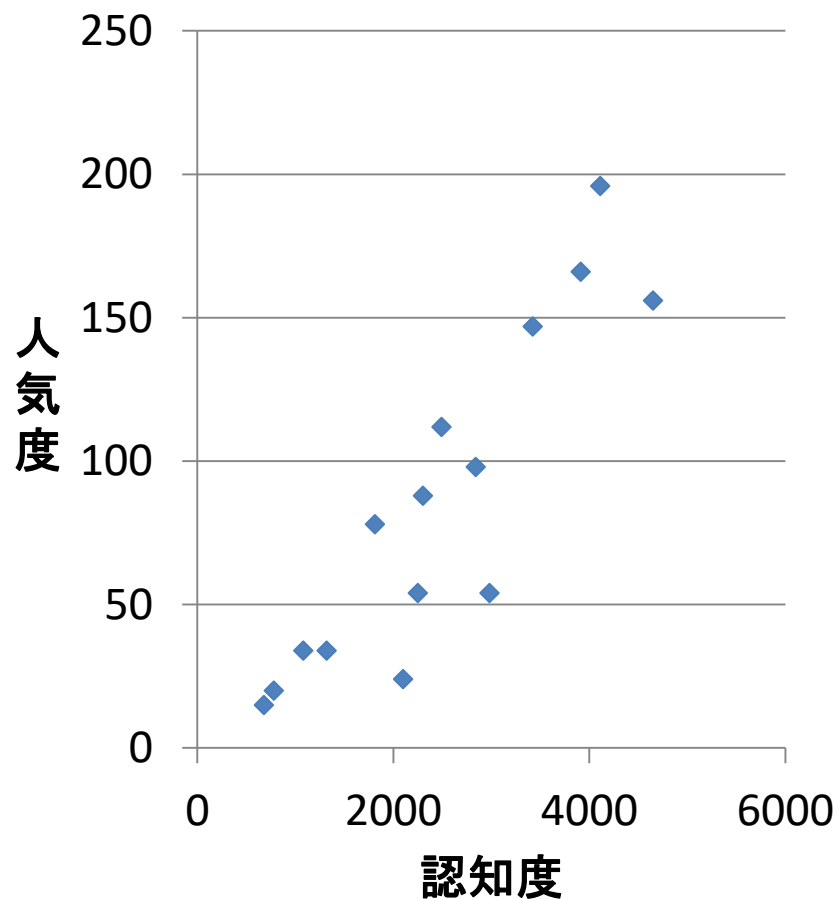
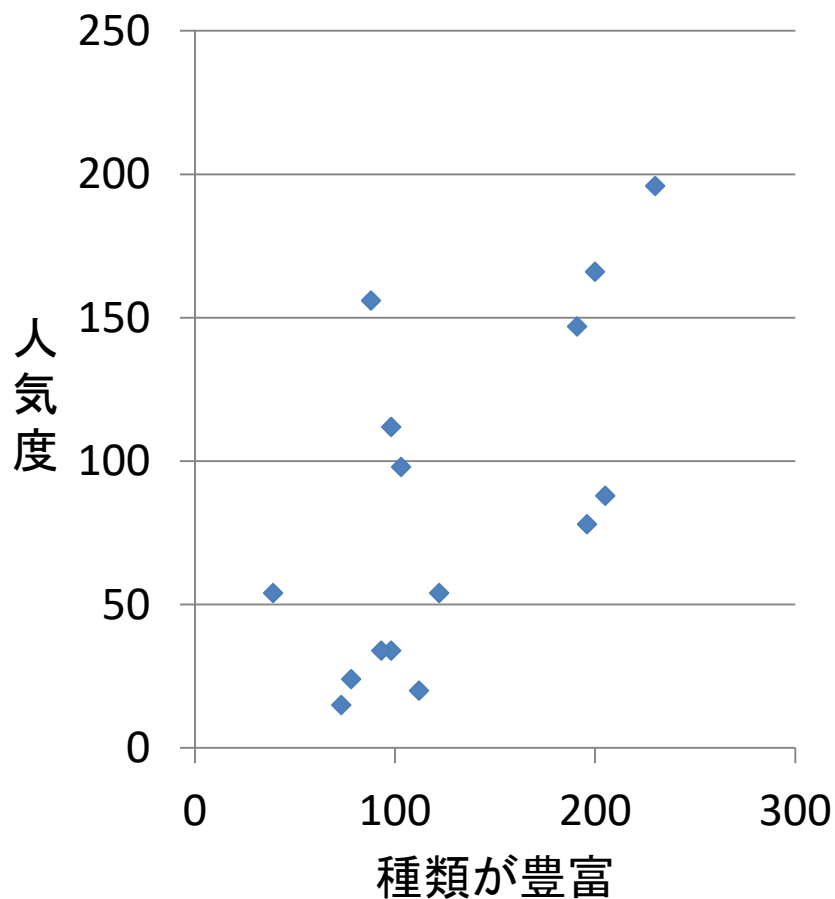
	人気度	認知度	購入経験	食パン・イギリスパンがおいしい	菓子パンがおいしい	種類が豊富
アンデルセン	196	4110	328	88	235	230
サンジェルマン	166	3910	323	108	191	200
木村屋総本店	156	4650	386	44	249	88
ポンバドウル	147	3420	279	122	156	191
フォション	112	2490	166	171	142	98
ドンク	98	2840	225	93	122	103
ヴィ・ド・フランス	88	2300	181	44	176	205
ジョアン	78	1810	127	122	147	196
ホテルオークラ	54	2980	137	230	54	39
ぽるとがる	54	2250	147	93	127	122
ダロワイヨ	34	1320	59	73	166	98
ルノートル	34	1080	73	88	147	93
プチモンド	24	2100	83	39	98	78
ビゴの店	20	780	29	64	93	112
エディアール	15	680	29	73	108	73

人気アンケート

- ▶ 各パン屋さんについて、以下の項目を数値化(人数)
 - ▶ 人気度(そのパン屋さんは好きですか?)
 - ▶ 認知度(そのパン屋を知っている)
 - ▶ 購入経験(買ったことがありますか?)
 - ▶ 食パン・イギリスパンがおいしい
 - ▶ 菓子パンがおいしい
 - ▶ 種類が豊富
 - ▶ どのような要素が高いと、人気だと判断できる？
- **人気度と各項目の関係**が知りたい！

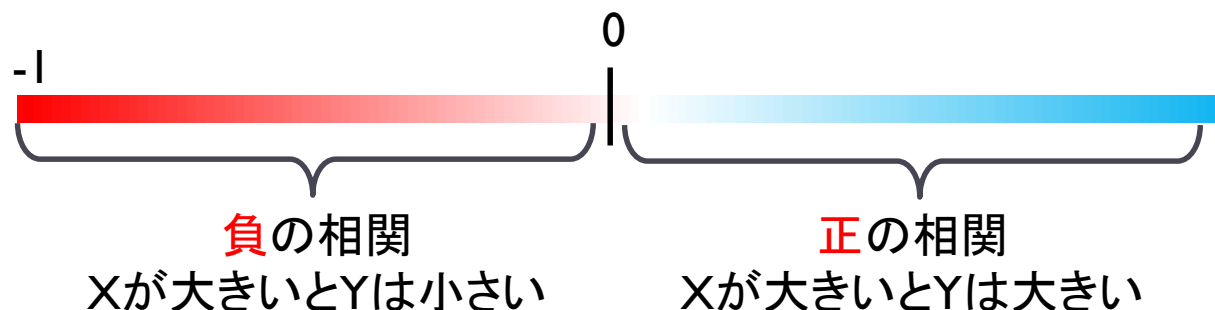
グラフで見てみる：散布図

▶ 散布図を描くとわかりやすい



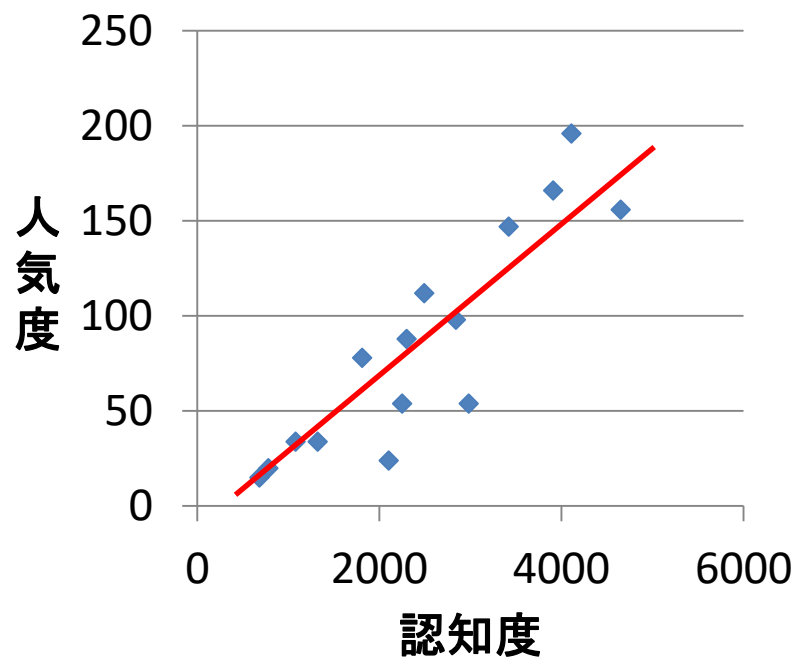
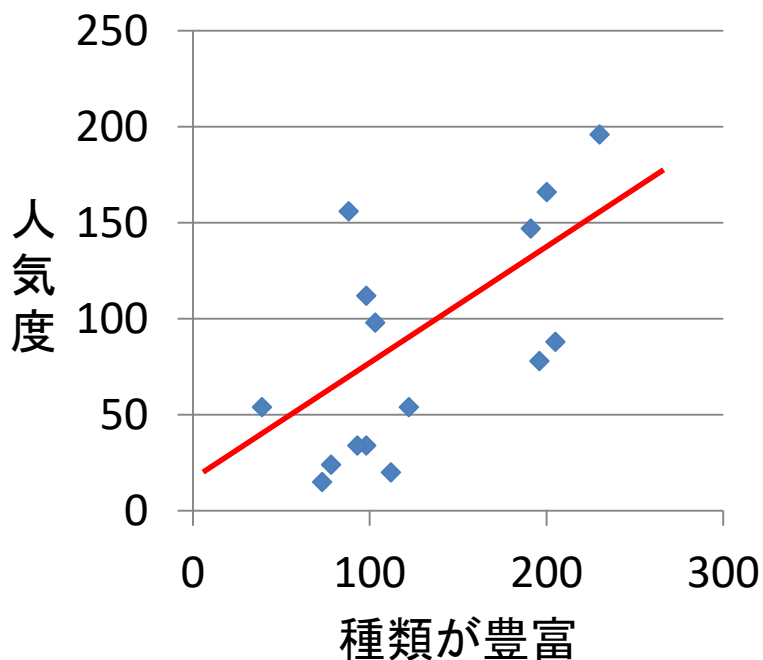
数字で見てもよう：相関係数

- ▶ 相関係数は-1から1までの値をとる
 - ▶ 0よりも大きい場合は**正の相関**、0よりも小さい場合は**負の相関**
- ▶ **絶対値**で数字が大きいほど、関係が強い
 - ▶ 0よりも-0.5、-1の方が、負の相関で関係が強い



散布図と相関係数

- ▶ 散布図ではわからなかった**関係の強さがわかる**



- ▶ 相関係数が高いほど、散らばりが少なく、より**直線の関係に近くなる**

記述統計で議論する際のポイント

- ▶ 一つの値で大小や強弱を議論する時、特に基準値などはない

- ▶ 標準偏差14.5点の場合、散らばっている？まとまっている？わからない
- ▶ 相関係数=0.5の場合、「関係が強い」という人もいれば、「あまり関係ない」という人もいる

→ これだけでは判断しづらい

- ▶ 他のケースなど、比較対象を使うと議論しやすい

- ▶ パン屋の人気アンケートの例

- ▶ 「人気度と購入経験」の相関係数は()、「人気度と種類の豊富さ」の相関係数は()

→ より人気度に重要なのは、()

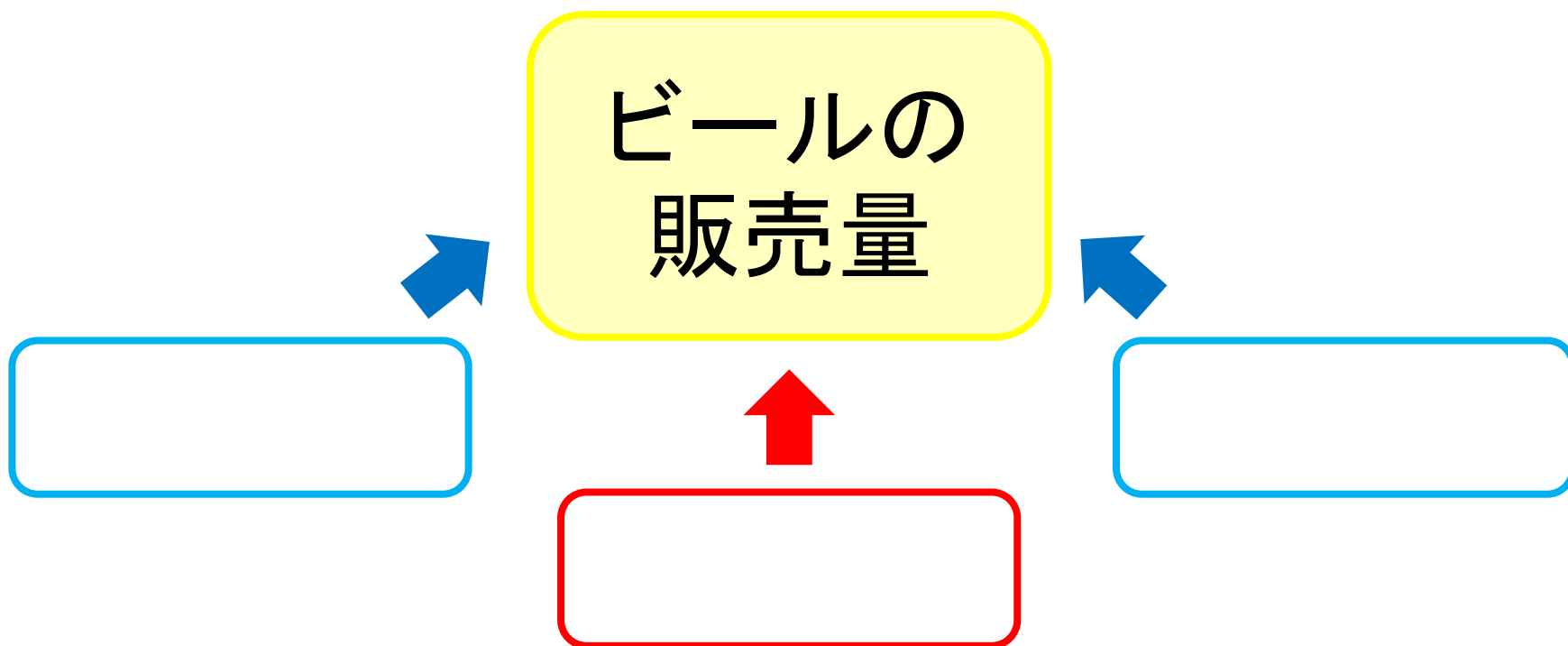
2変数の関係：散布図と相関係数

▶ 散布図と相関係数のメリット・デメリット

	散布図	相関係数
メリット	視覚的に関係がわかりやすい	関係の強さが数字で判断可能
デメリット	関係の強さはわかりづらい	変数間の具体的な関係性(片方が1単位変化したときのもう一方の変化の大きさ)はわからない

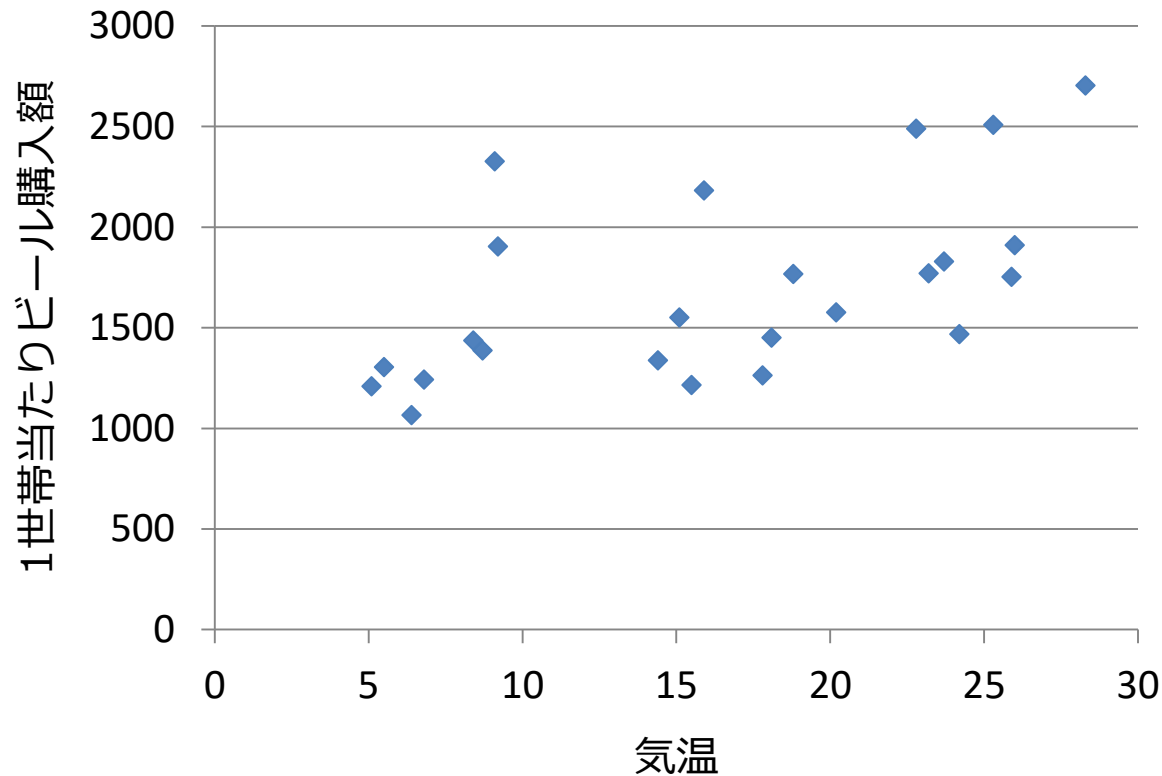
回帰分析へ：より深い2変数の関係

- ▶ より深い2変数の関係を知りたい場合がある
- ▶ 例：ビール会社に入社し、生産量を決める部署に配属



回帰分析へ：より深い2変数の関係

▶ 散布図と相関係数で議論してみよう



相関係数
=0.5645

回帰分析へ：より深い2変数の関係

(散布図、相関係数を使って)
気温とビールの売り上げは**正の相関**にあります！

(相関係数を使って)
価格や**他のビール**との**関係**と同じくらい重要です！

つまり、今年は**気温が高くなりそう**なので、**たくさん生産**しましょう！

正の関係にあるのも重要なのもわかった

じゃあ、何度上がれば
どれだけ販売量増えるの？
具体的な数字出して！



回帰分析へ：より深い2変数の関係

- ▶ ビールの販売量を以下のように分解してみる

ビールの
販売量



変動する要素(価格、気温)に影響されず、確実に売れると見込める**一定の部分**



場所や時期に関わらず、
気温の上下により
増減する販売量の部分



気温以外の要素で決まる部分

回帰分析へ：一次関数

- ▶ この関係を最も簡単な式で表す

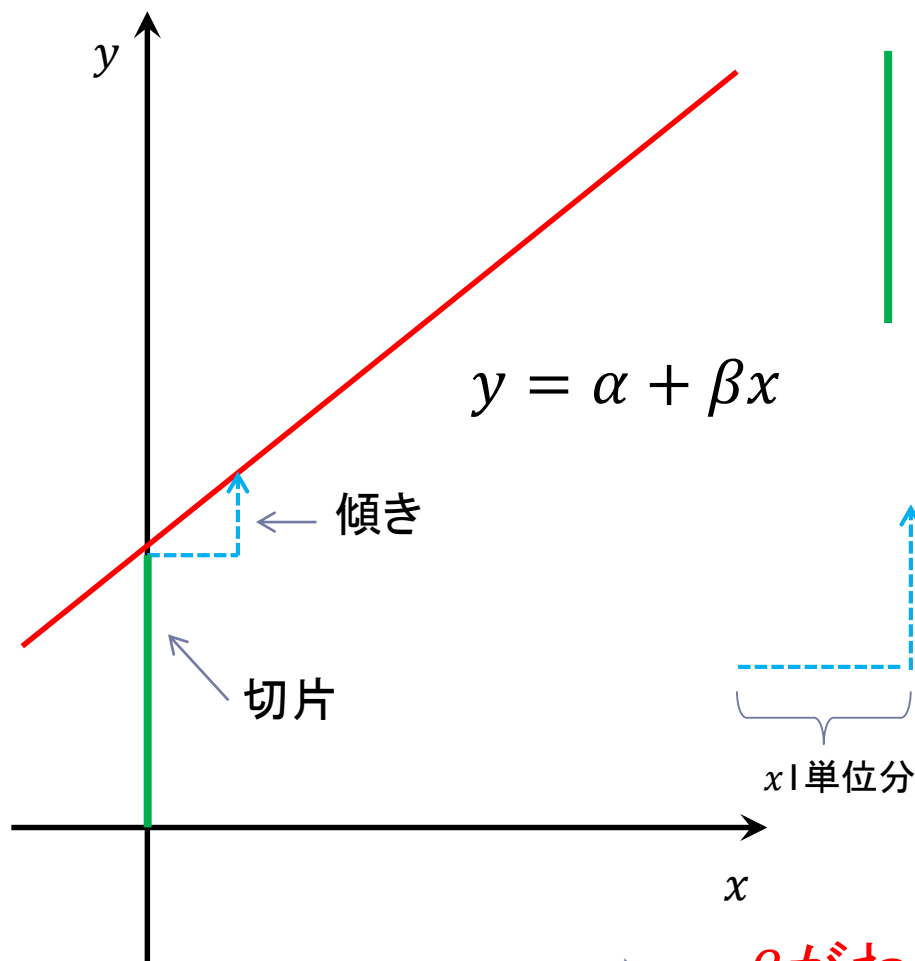
→ 一次関数

x_i から y_i への影響の大きさ(β)
は一定と仮定する


$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- ▶ 各観測値を示す i の部分は省略することがある
- ▶ x, y は変数であり、 α, β は係数(パラメータ)であることを忘れないように
 - ▶ (復習)変数: 主体によって異なる値をとるもの
 - ▶ 係数(パラメータ): x と y の関係における一定の部分、主体に関わらず同じ値

回帰分析へ：一次関数



- ▶ 切片
 - ▶ x の値が0の時の y の値
 - x が0でも存在する一定の部分

- ▶ 傾き
 - ▶ x が1単位増加した時の y の増加分
 - x の値に応じて変化する y の一部分

▶ α, β がわかれば、具体的な直線がひける

回帰分析へ：一次関数の解釈

- ▶ ビールの販売量 y には、定数 α と、 βx 、 ε が影響している
 - ▶ α : 気温や価格が変動しても、一定の α は売れる
 - ▶ βx : 気温 x が1単位(1°C)上がるごとに、 β だけ販売量が増える
 - ▶ ε : 一定の部分や、気温に関わる部分以外の要素によって決まるもの
 - ▶ 例: 価格、時期、都市、好み、他のビールの銘柄...
 - ▶ キリがないので、この部分は全部まとめて誤差扱いする(誤差項)
- ▶ この一次関数では、気温 x がビールの販売量 y の一部を説明していると考える
 - ▶ x : 説明変数
 - ▶ y : 被説明変数

回帰分析へ：一次関数の解釈

- ▶ 一次関数に説明変数と被説明変数を当てはめて、その関係を定量的に分析する方法
- 回帰分析
 - ▶ 回帰:「元のところに帰る」「根源に戻る」という意味であり、「 y の原因 x を考えて分析する」というイメージ
 - ▶ 回帰分析の一次関数(直線)を表す式を、**回帰式**という

相関係数と回帰分析の違い

▶ 相関係数と回帰分析の違い

→ **因果関係**を考えて定式化している

▶ 2変数の関係を見る際に...

- ▶ 相関係数: どちらが x でどちらが y でも関係ない(因果関係は関係ない)
- ▶ 回帰分析: 説明する側が x で説明される側が y (**因果関係を考えて選択**)

▶ 因果関係はどうやって考える？

→ **自分で決める！**

- ▶ 経済理論や、各分野の理論を背景に考えると便利

余談

▶ x と y の名称一覧

x		y	
説明変数	Explanatory variable	被説明変数	Explained variable
独立変数	Independent variable	従属変数	Dependent variable
Regressor		Regressand	
Control variable		Response variable	
Predict variable		Predicted variable	

- ▶ 因果関係を考えて式をたてるが、**回帰分析の結果が必ず因果関係を証明しているわけではない**
 - ▶ 参考にした因果関係も立てた式も、結局はあなたの考え
 - ▶ 計算結果からはあくまで関係性の確認とその大きさがわかるだけで、「(その)**因果関係が本当に存在するよor間違ってるよ**」などの情報は得られない
- ▶ 因果関係については、次回の講義で詳しく扱う予定

Active learning(4)

- ▶ 以下の分析目的に応じて、適切な回帰式をたてなさい。目的が述べられていない場合、自分で目的を想定して、それに適した回帰式をたてなさい。
 1. パン屋の「菓子パンをおいしいと答えた人数」と「人気度」を使って、人気度の要因を探りたい場合
 2. 「気温」と「石油の消費量」
 3. 「賃貸物件の家賃」と「築年数」の関係
 4. (マクロ既履修者)生産関数

練習問題(6)

「ビールの販売量額」(円)と気温(°C)の関係の回帰分析の結果は以下の通りとなった。

$$y = 1125 + 35x$$

5. この結果からどのようなことがわかるか答えなさい。

6. 他にはビールの販売量に影響を与える要素を挙げて下さい。また、追加的な要素を上記のモデルに入れた時に、 x の係数はどう変わるか答えなさい。