統計学

第7回　変数間の関連

期末試験の試験範囲

- 授業範囲全域
  - 統計的常識○×問題
  - 基本統計量に関する穴埋め問題
  - 推定と検定に関する穴埋め問題
  - 相関

期末試験に関する注意事項

- 持込：可
- 持込不可物
  - パソコン・携帯電話・スマートフォン
  - 過去問
- 必要なもの
  - 筆記用具
  - 電卓（ルートの計算ができるもの）
- レポートも忘れずに

授業内容

- 相関係数
- クロス集計表
- 回帰分析
  - 単回帰分析
  - 重回帰分析

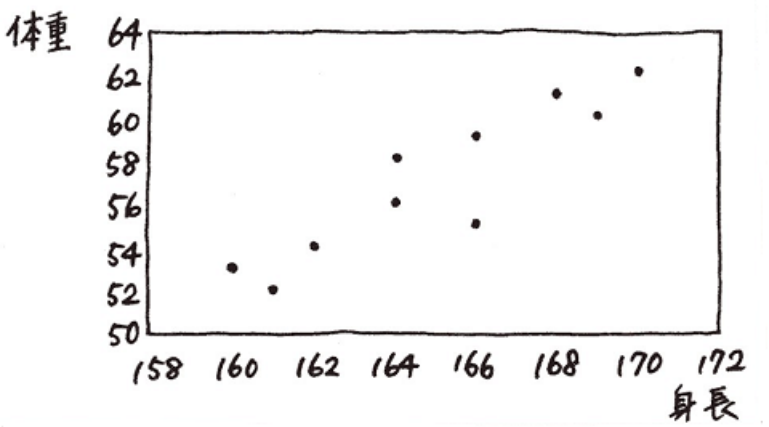
教科書

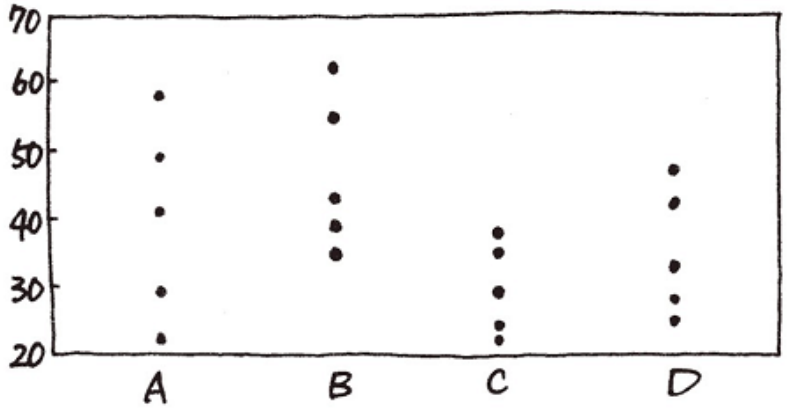
- P214～P228（相関係数・クロス集計表）
- P236～P258（回帰分析）

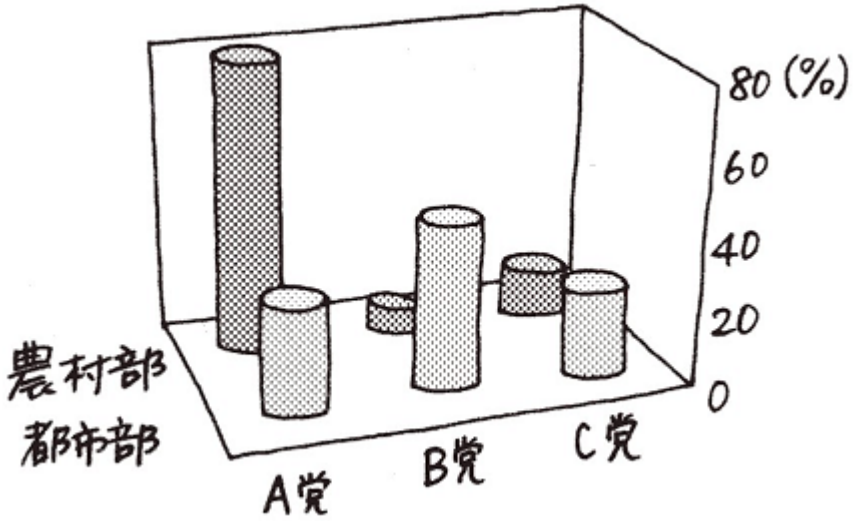
２変数の関連

- 身長が高ければ体重も重い
- 年齢が異なれば好きなビールの銘柄も異なる
- 居住地が異なれば支持政党も異なる
- ……etc

２変数の関連（グラフ）


体重
64
62
60
58
56
54
52
50
158
160
162
164
166
168
170
172
身長


70
60
50
40
30
20
A
B
C
D


80 (%)
60
40
20
0
農村部
都市部
A党
B党
C党

データの型

- 量的データ(数量データ):数値で表すことができるデータ、身長・体重・人数など
  - 連続型
  - 離散型
- 質的データ(カテゴリーデータ):カテゴリで表される定性的なデータ、性別・職種など

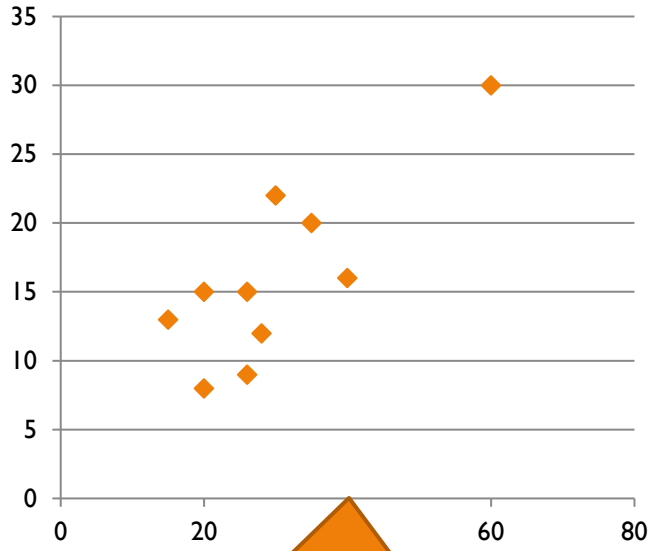
２変数の関連性を測る統計量

- 相関係数（量的データと量的データ）
- 相関比（量的データと質的データ）
- 連関係数（質的データと質的データ）
  - クラメールの連関係数
  - ２×２連関係数

扱うデータの種類によって違う
（質的データか量的データか）

2変数の関連性をみる図表

- 散布図・相関表（量的データと量的データ）
- 層別ヒストグラム・箱ひげ図（量的データと質的データ）
- クロス集計表（質的データと質的データ）



散布図

データの個数 / 性格	性格			
血液型	1	2	3	総計
A	2	1	1	4
AB	1	1		2
B	4	1	1	6
O	5		3	8
総計	12	3	5	20

クロス集計表

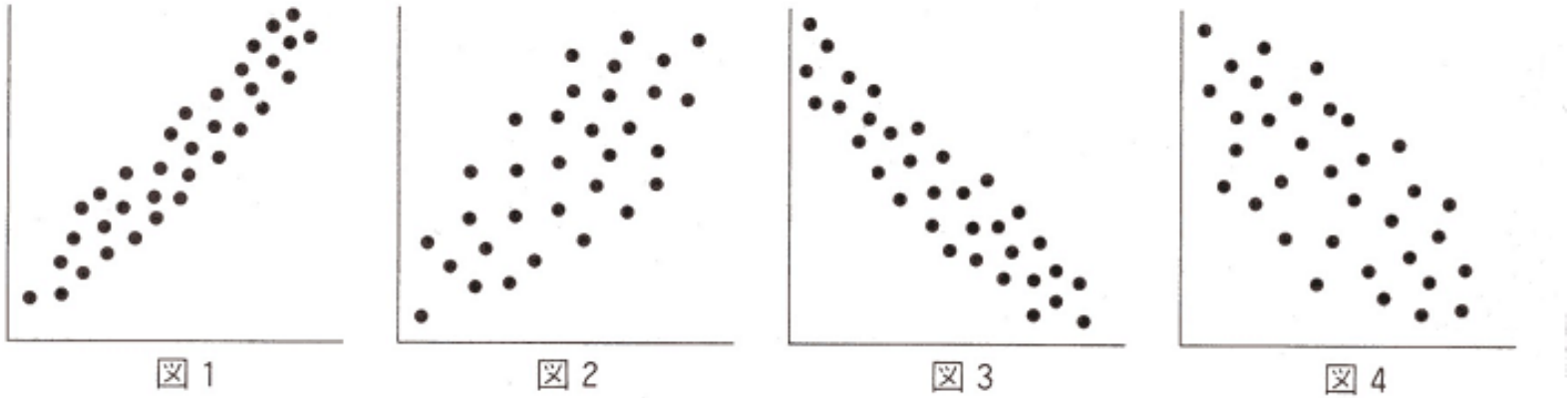
相関関係のパターン

- Aが増えるとBも増える＝正の相関関係
- Aが増えるとBが減る＝負の相関関係

相関係数

散布図

散布図：二つの変数の関係をグラフにプロットしたもの

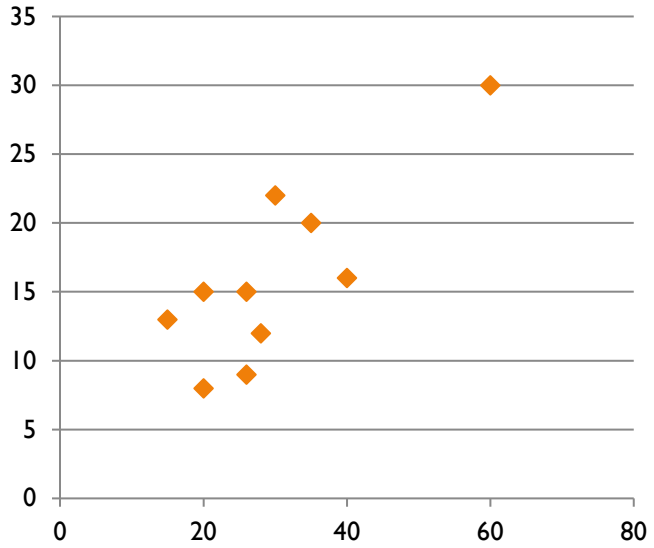


（ほぼ）数量データの場合に使われている

散布図の作り方

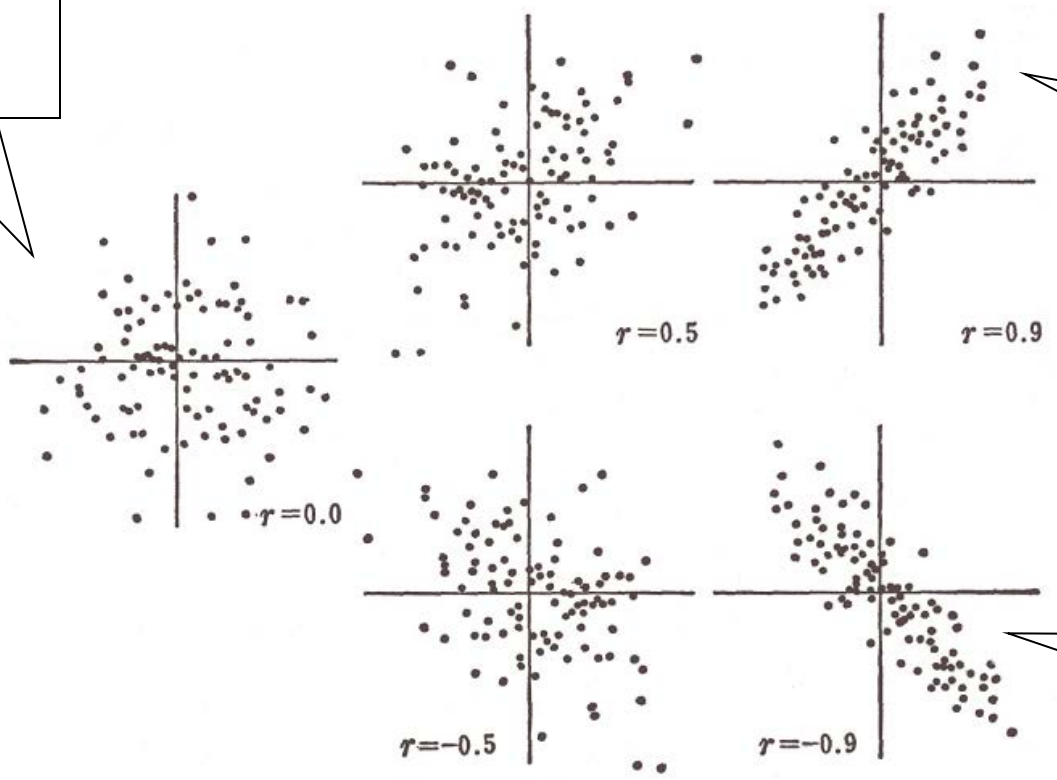
- 縦軸と横軸を描く
- 各変数を横軸と縦軸に割り当てる
  - 原因となる変数を横軸に割り当てる
- 各観測値を一つの点として散布図の中にプロットする
  - 2つの変数の値がその点の座標になる

番号	給料	残業時間
1	30	22
2	26	15
3	40	16
4	28	12
5	26	9
6	35	20
7	20	15
8	60	30
9	15	13
10	20	8



散布図と相関係数

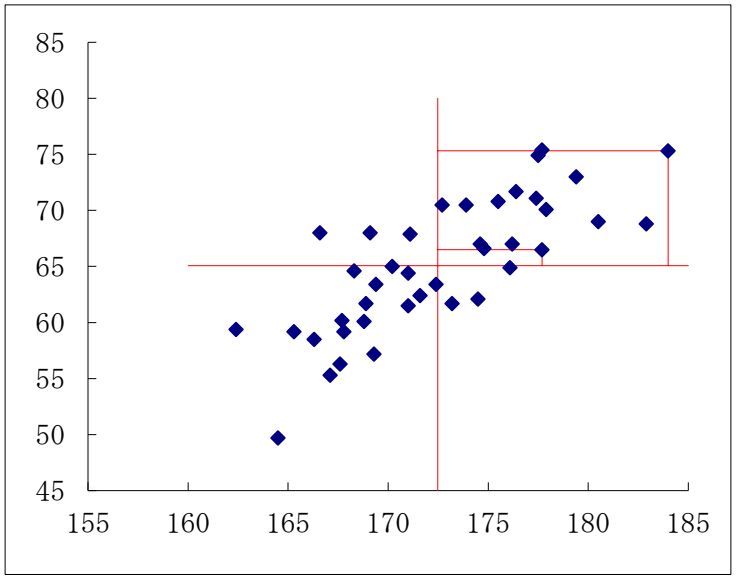
無相関



正の相関

負の相関

関連の度合いを測る

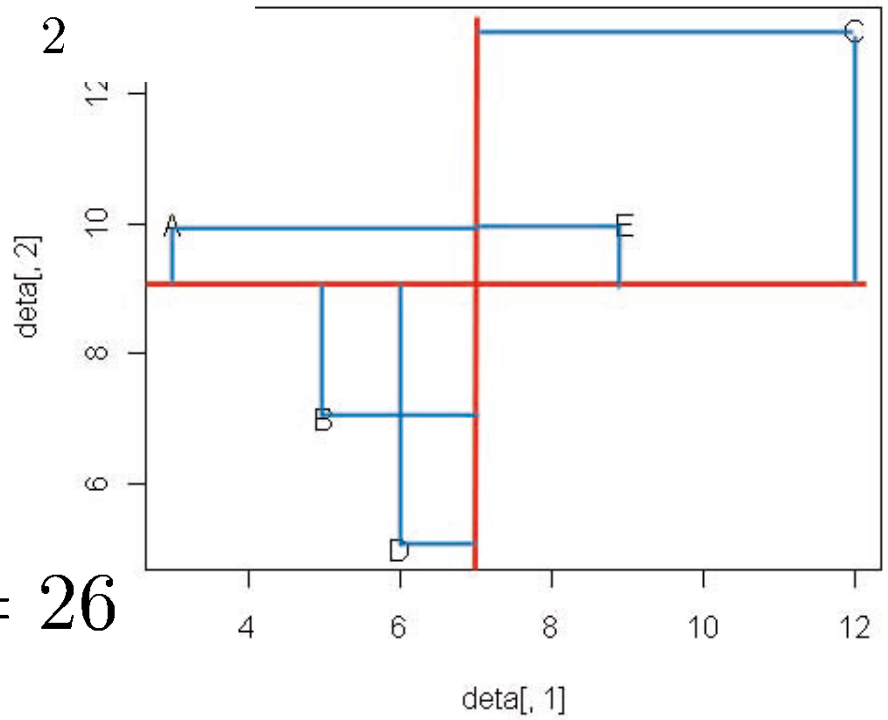


- もし2つの変数に関連性があるならば、散布図の点は各変数の平均で区切られた4つの領域の内「右上と左下」または「右下と左上」に固まるだろう
- 各点から2つの直線に垂線を引いてできる四辺形について、「右上と左下」の総面積と「右下と左上」の総面積の差で関連の度合いを測る

例

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
A	3	10	-4	1	-4
B	5	7	-2	-2	4
C	12	13	5	4	20
D	6	5	-1	-4	4
E	9	10	2	1	2

D	6	5
E	9	10



$$(-4)+4+20+4+2=26$$

共分散

- ２変数データについて、（平均からの）偏差の積の平均を計算したもの
- ２変数の関連の強さを測る

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

共分散の性質

- ２変数の関連の強さ（直線的関連）を測ることができる
- データ数で割るときと、（データ数－１）で割るときがある
  - 後者は「不偏共分散」と呼ばれている
  - EXCELでは「COVARIANCE.P」と「COVARIANCE.S」（2010以降）
- 数値の大きさが測定単位に依存してしまう

例

	m^2	
番号	面積	築後年数
1	17.32	15
2	13.20	12
3	19.80	11
4	17.32	18
5	19.80	6
6	14.85	9
7	27.22	9
8	24.75	8
9	14.85	10
10	13.20	3
11	24.75	6
12	24.75	11
13	32.18	15
14	29.70	5
15	24.75	2
16	34.65	0
17	39.60	1
18	42.08	4
19	32.18	0
20	40.42	9
21	34.65	1
22	38.78	0
23	33.82	0
24	38.78	0
25	42.08	3

	坪	
番号	面積	築後年数
1	5.25	15
2	4.00	12
3	6.00	11
4	5.25	18
5	6.00	6
6	4.50	9
7	8.25	9
8	7.50	8
9	4.50	10
10	4.00	3
11	7.50	6
12	7.50	11
13	9.75	15
14	9.00	5
15	7.50	2
16	10.50	0
17	12.00	1
18	12.75	4
19	9.75	0
20	12.25	9
21	10.50	1
22	11.75	0
23	10.25	0
24	11.75	0
25	12.75	3

- 面積を平方メートルで測った場合、共分散は-29.20
- 面積を坪で測った場合、共分散は-8.85

（単）相関係数

数量データの間に、どんな感じの「直線的傾向」があるかを表す係数

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

共分散を2変数の標準偏差の積で割ったもの
測定単位に依存しない

共分散を標準偏差で割る理由

測定単位に依存しない

- データを標準化して計算
  - データを、(データー平均値)÷標準偏差で加工

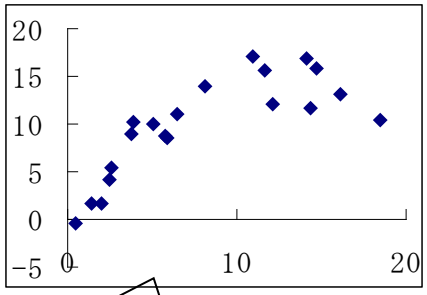
$$
\begin{aligned}
r &= \frac{x \text{ と } y \text{ の共分散}}{x \text{ の標準偏差} \times y \text{ の標準偏差}} \\
&= \frac{(1/n)\,\Sigma_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sqrt{(1/n)\,\Sigma_{i=1}^{n}(x-\bar{x})^2}\sqrt{(1/n)\,\Sigma_{i=1}^{n}(y-\bar{y})^2}} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{x-\bar{x}}{\sqrt{(1/n)\,\Sigma_{i=1}^{n}(x-\bar{x})^2}} \times \frac{y-\bar{y}}{\sqrt{(1/n)\,\Sigma_{i=1}^{n}(y-\bar{y})^2}}
\end{aligned}
$$

相関係数の値の目安

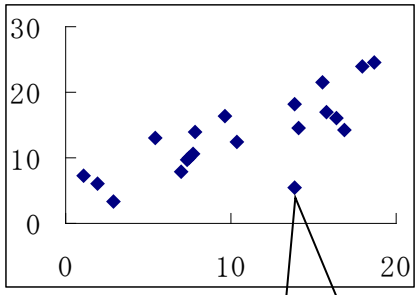
「××以上ならば2変数は強く関連しているといえる」というような統計学的な基準はない

単相関係数の絶対値	細かくいうなら	大雑把に
1.0～0.9	非常に強く関連している	関連している
0.9～0.7	やや強く関連している	
0.7～0.5	やや弱く関連している	
0.5 未満	非常に弱く関連している	関連していない

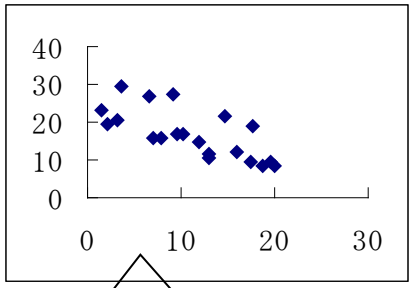
色々な散布図



二次関数っぽい



外れ値が
一つ



２種類のデータが
混じっているような

相関を調べる前に散布図を描いてみる

相関係数は、こういった散布図に弱い
（相関係数は頑健な統計量ではない）

擬似相関

- コウノトリの巣の数と出生数に正の相関
  - コウノトリが子どもを運んでくる？
  - 家の数（世帯数、軒下数）が共通因子
- 携帯メールで絵文字をたくさん使う人はケータイ小説愛読率も高い
  - 「年齢」が隠れた影響因子。若い人ほど絵文字を使い、かつケータイ小説を好む

関連性のないもの同士でも、何らかの共通因子により相関が出てくる

相関係数・相関関係・因果関係

相関係数：直線関係の強さを測る指標

相関関係：変数の一方が変化するにつれ、他の変数が同時に変化する関係

因果関係：変数の一方の変化が、他方の変化を引き起こす、原因と結果の関係

相関係数は相関関係のみ（しかも直線だけ）をあらわしている

因果関係を判断するのは

相関関係があっても、直ちに因果関係があるとはいえない。

因果関係を判断するのは、統計の解析技術ではなく、対象についての固有技術である。

例えば

- 次の食べ物を禁止すべきかどうか検討してみましょう
  - 心筋梗塞で死亡した日本人の95%以上が生前ずっとこの食べ物を口にしていた
  - 強盗や殺人などの凶悪犯の70%以上が犯行前24時間以内にこの食べ物を口にしている
  - 日本人に摂取を禁止すると、精神的なストレス状態が見られることがある
  - 江戸時代以降日本で起こった暴動のほとんどは、この食べ物が原因である

因果関係の３つの基準

- ２つの変数の間に、関係があること
- ２つの変数の間に、時間的な順序関係が存在すること
  - 時間が早い変数が原因である可能性が高い
- ２つの変数の関連が、時間的に先行する他の変数によって説明されないこと
  - 擬似相関
  - 第３変数の存在

変数の種類による相関の計算

- データには量的変数と質的変数がある
- 相関係数は量的変数の時に計算される
  - 質的変数の関連性を測る指標は別にある
- 質的変数の関連性を測る指標として相関係数を用いるとまずいことが起こるので注意
  - 関連が高いものでも相関係数が低めに出る

クロス集計表

- 質的データ同士の関連性をみるために作られる集計表
- EXCELのピボットテーブルでつくれる

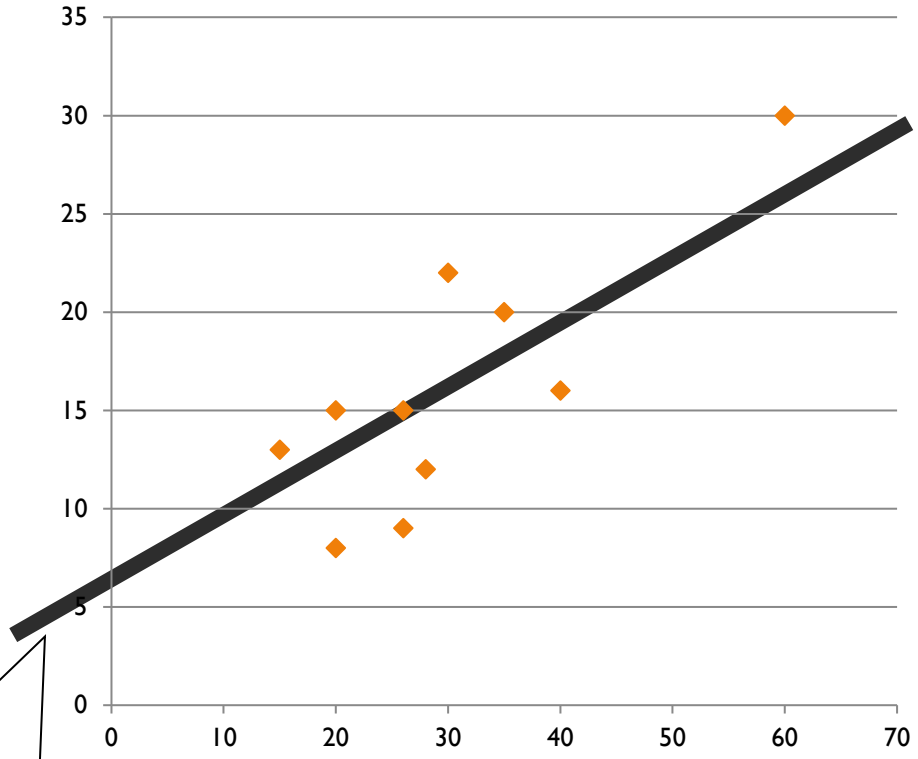
No.	血液型	性格
1	B	2
2	A	1
3	O	1
4	O	1
5	AB	2
6	B	1
7	O	3
8	B	3
9	AB	1
10	A	3
11	A	2
12	O	1
13	B	1
14	O	1
15	O	3
16	O	3
17	O	1
18	A	1
19	B	1
20	B	1

データの個数 / 性格	性格			
血液型	1	2	3	総計
A	2	1	1	4
AB	1	1		2
B	4	1	1	6
O	5		3	8
総計	12	3	5	20

クロス集計表から関連を測る：連関係数

回帰分析

直線の当てはめ



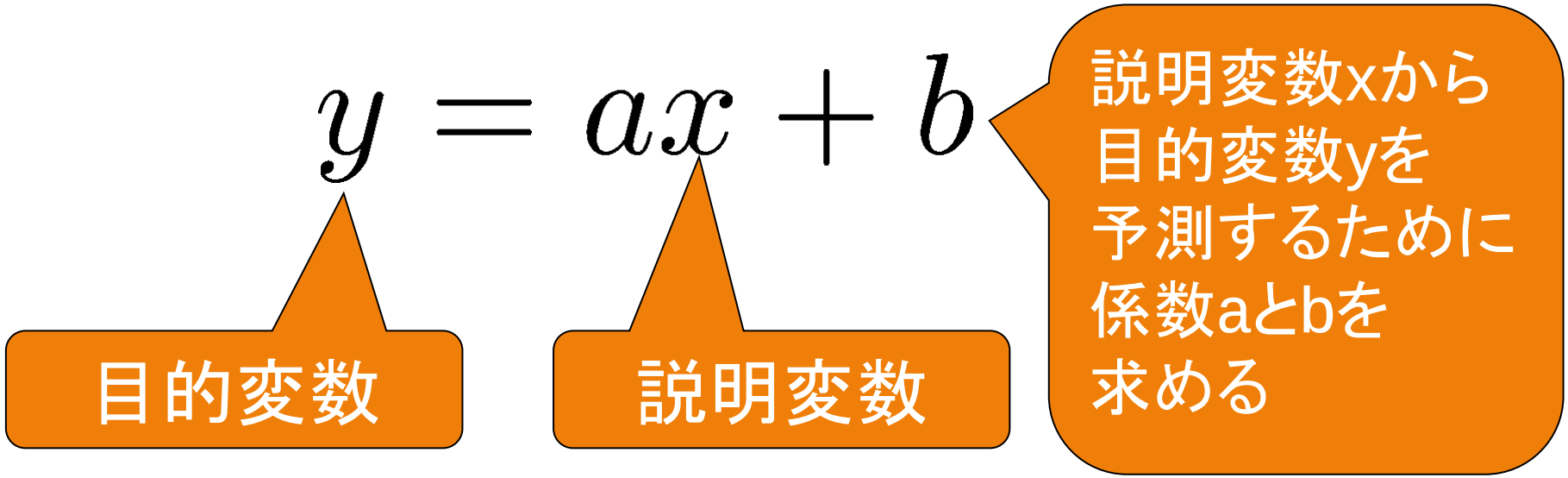
この直線の式を求めるのが
単回帰分析

目的

- 直線的関係のある2つの量的変数について、その直線関係を表す式を求める
- 求めた直線式を用いて、一つの変数から他の変数を推定する

目的変数と説明変数

- 目的変数：回帰式を使って予測する変数
- 説明変数：目的変数を予測する際に使用する変数

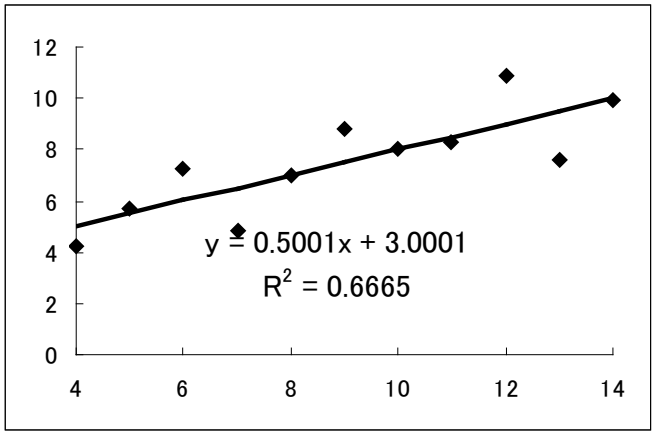


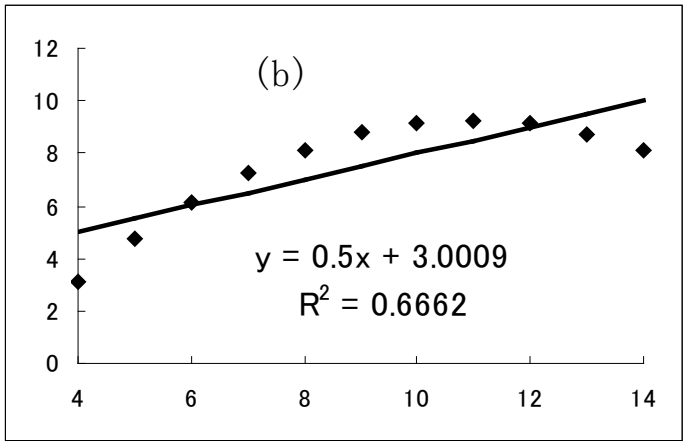
単回帰分析の手順

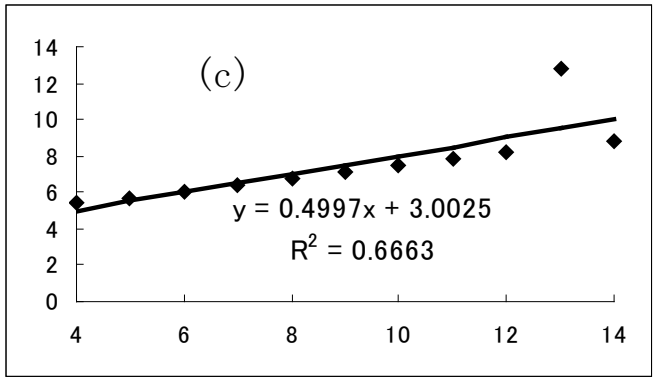
- 回帰式を求める意味があるかどうかを調べる為に、散布図を描いて直線的傾向があるかどうかをみる
- 回帰式の係数（回帰係数）を求める
- 回帰式の精度（寄与率）を求める
- 回帰係数の検定を行う
- 母回帰の推定と予測
- 回帰モデルの妥当性を診断（回帰診断）

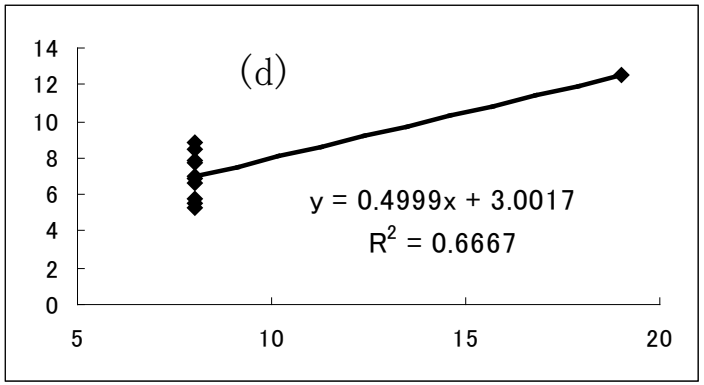
とりあえずは、上3つを行えばよい。

アンスコムの例









ほぼ同じ回帰式が得られているが、データにはかなりの問題がある。

回帰係数を求める

回帰式からの予測値 $\hat{y} = b_0 + b_1 x$ と、実際の値（実測値）y との差 $e = y - \hat{y}$ の 2 乗が、データ全体として小さくなるように、すなわち、

$$S_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y - \hat{y})^2 = \sum_{i=1}^{n} (y - b_0 - b_1 x)^2$$

が最小になるように回帰係数 b_0、b_1 を求める。

$$b_1 = \frac{S_{xy}}{S_{xx}}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

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

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

平方和の分解

$$y_i - \bar{y} = y_i - \hat{y} + \hat{y} - \bar{y} = (\hat{y} - \bar{y}) + e_i$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

総平方和

回帰平方和

残差平方和

$$S_T = S_{yy}、\ S_R = \frac{S_{xy}^2}{S_{xx}}、\ S_e = S_T - S_R$$

寄与率

- 寄与率：総平方和に対する回帰平方和の割合
- 回帰式の精度が高いほど寄与率は1に近づく
- 寄与率が0.5以上なら、回帰式の精度は（経験的に）良いといえる
- 相関係数の2乗は寄与率に等しくなる

$$R^2 = \frac{S_R}{S_T}$$

回帰分析に関する注意

- 回帰分析をする前には必ず、散布図を描きましょう
  - 例え関連性が薄くても、回帰式は出る
  - 直線以外の関連性が見えてくる可能性
  - 外れ値
- 得られた回帰式で予測をするときは「誤差」と「外挿」に注意しましょう
  - 「外挿」：説明変数の範囲の外で回帰式を適用すること
- 寄与率の値に注意
  - 1に近ければ「良い」回帰式

EXCELを用いて

- LINEST関数を用いる
  - 使い方、どこに何が表示されるかに注意
- 分析ツールを用いる

実際の単回帰分析例

- 東京の気温と那覇の気温

レポート４

重回帰分析

- 説明変数が2個以上ある場合において、説明変数から目的変数を予測する式を作る分析手法
- 分析方法は単回帰の時と同じ
  - 寄与率は、説明変数の個数が多くなると高くなるので、調整をする
  - 良い回帰式を作る為に、説明変数の取捨選択を行う

多変量解析とは

データが「量的変数」であるか「質的変数」であるかによって分析方法が若干異なる

- 複数の変数を同時に分析する統計分析手法
- 「目的変数」がある手法とない手法に分かれる
  - 目的変数がある手法
    - 回帰分析・数量化I類・ロジスティック回帰分析
    - 判別分析・数量化2類
    - ニューラルネットワーク（教師あり学習）
  - 目的変数のない手法
    - 主成分分析・数量化3類
    - 因子分析
    - ニューラルネットワーク（教師無し学習）
    - クラスター分析・多次元尺度法・自己組織化マップ

数理統計的な手法と、コンピュータの発達とともに出てきた手法とがある