データ解析の基礎（Ⅰ）

ーその手法はどうして考えるのかー

安田晃

内　容

内　容

内　容

内　容

内　容

データは私たちの周囲にたくさんある（1-1）

- **実験，実習で得ている．**
- 図書館には様々な公的機関のデータ集がいっぱい．
- 会社，団体が公開しているデータもある．
- 日常的にデータを一般化してはいないか．――― △●学の講義開始は毎回5分程度ずれるから，・・・．
- 日常的にデータを活用してはいないか．――― 同じ中古車ならN社よりT社が安い．ならば，T社を買おうか．
- 私たちは何らかのデータによって生活している，といっても過言ではないだろう．

データは私たちの周囲にたくさんある（1-2）

- 実験，実習で得ている．
- **図書館には様々な公的機関のデータ集がいっぱい．**
- 会社，団体が公開しているデータもある．
- 日常的にデータを一般化してはいないか．――― △●学の講義開始は毎回5分程度ずれるから，・・・．
- 日常的にデータを活用してはいないか．――― 同じ中古車ならN社よりT社が安い．ならば，T社を買おうか．
- 私たちは何らかのデータによって生活している，といっても過言ではないだろう．

データは私たちの周囲にたくさんある（1-3）

- 実験，実習で得ている．
- 図書館には様々な公的機関のデータ集がいっぱい．
- **会社，団体が公開しているデータもある．**
- 日常的にデータを一般化してはいないか．――― △●学の講義開始は毎回5分程度ずれるから，・・・．
- 日常的にデータを活用してはいないか．――― 同じ中古車ならN社よりT社が安い．ならば，T社を買おうか．
- 私たちは何らかのデータによって生活している，といっても過言ではないだろう．

データは私たちの周囲にたくさんある（1-4）

- 実験，実習で得ている．
- 図書館には様々な公的機関のデータ集がいっぱい．
- 会社，団体が公開しているデータもある．
- **日常的にデータを一般化してはいないか．**
  ――― △●学の講義開始は毎回5分程度ずれるから，・・・．
- 日常的にデータを活用してはいないか．――― 同じ中古車ならN社よりT社が安い．ならば，T社を買おうか．
- 私たちは何らかのデータによって生活している，といっても過言ではないだろう．

データは私たちの周囲にたくさんある（1-5）

- 実験，実習で得ている．
- 図書館には様々な公的機関のデータ集がいっぱい．
- 会社，団体が公開しているデータもある．
- 日常的にデータを一般化してはいないか．――― △●学の講義開始は毎回5分程度ずれるから，・・・．
- **日常的にデータを活用してはいないか．**
  ――― 同じ中古車ならN社よりT社が安い．ならば，T社を買おうか．
- 私たちは何らかのデータによって生活している，といっても過言ではないだろう．

データは私たちの周囲にたくさんある（1-6）

- 実験，実習で得ている．
- 図書館には様々な公的機関のデータ集がいっぱい．
- 会社，団体が公開しているデータもある．
- 日常的にデータを一般化してはいないか．△●学の講義開始は毎回5分程度ずれるから，・・・．―――
- 日常的にデータを活用してはいないか．同じ中古車ならN社よりT社が安い．ならば，T社を買おうか．―――
- **私たちは何らかのデータによって生活している，といっても過言ではないだろう．**

データは私たちの周囲にたくさんある（2-1）

- **そのデータは私たちに何を言っているのだろうか．**
- そのデータを有用な形で得ることはできないか．
- そのデータから新たな知識が得られないか．
- その知識は科学的に妥当か．
- その知識からデータの背景にある様々な因子は読み取ることが可能か．
- そのデータは私たちに何を言っているのだろうか．

データは私たちの周囲にたくさんある（2-2）

- そのデータは私たちに何を言っているのだろうか．
- **そのデータを有用な形で得ることはできないか．**
- そのデータから新たな知識が得られないか．
- その知識は科学的に妥当か．
- その知識からデータの背景にある様々な因子は読み取ることが可能か．
- そのデータは私たちに何を言っているのだろうか．

データは私たちの周囲にたくさんある（2-3）

- そのデータは私たちに何を言っているのだろうか．
- そのデータを有用な形で得ることはできないか．
- **そのデータから新たな知識が得られないか．**
- その知識は科学的に妥当か．
- その知識からデータの背景にある様々な因子は読み取ることが可能か．
- そのデータは私たちに何を言っているのだろうか．

データは私たちの周囲にたくさんある（2-4）

- そのデータは私たちに何を言っているのだろうか．
- そのデータを有用な形で得ることはできないか．
- そのデータから新たな知識が得られないか．
- **その知識は科学的に妥当か．**
- その知識からデータの背景にある様々な因子は読み取ることが可能か．
- そのデータは私たちに何を言っているのだろうか．

データは私たちの周囲にたくさんある（2-5）

- そのデータは私たちに何を言っているのだろうか．
- そのデータを有用な形で得ることはできないか．
- そのデータから新たな知識が得られないか．
- その知識は科学的に妥当か．
- **その知識からデータの背景にある様々な因子は読み取ることが可能か．**
- そのデータは私たちに何を言っているのだろうか．

データは私たちの周囲にたくさんある（2-6）

- そのデータは私たちに何を言っているのだろうか．
- そのデータを有用な形で得ることはできないか．
- そのデータから新たな知識が得られないか．
- その知識は科学的に妥当か．
- その知識からデータの背景にある様々な因子は読み取ることが可能か．
- **そのデータは私たちに何を言っているのだろうか．**

今までの提言からデータを構成している基礎を学ぶ（1-1）

- **『医療情報学』第3巻には，具体的な教育内容として「医学統計学の基礎」が謳ってある．**
- そもそも統計学とは何か．広い意味では，データが有している意味を科学的観点から学ぶこと．
- そして，その学んだ結果を実際の社会に返すこと．

今までの提言からデータを構成している基礎を学ぶ（1-2）

- 『医療情報学』第3巻には，具体的な教育内容として「医学統計学の基礎」が謳ってある．
- **そもそも統計学とは何か．広い意味では，データが有している意味を科学的観点から学ぶこと．**
- そして，その学んだ結果を実際の社会に返すこと．

今までの提言からデータを構成している基礎を学ぶ（1-3）

- 『医療情報学』第3巻には，具体的な教育内容として「医学統計学の基礎」が謳ってある．
- そもそも統計学とは何か．広い意味では，データが有している意味を科学的観点から学ぶこと．
- **そして，その学んだ結果を実際の社会に返すこと．**

今までの提言からデータを構成している基礎を学ぶ（2-1）

データを取り扱う上で必要なこと

例えば，「平均」

10，5，9の平均は8である．

確かに，$\frac{1}{3}(10+5+9)=\frac{24}{3}=8$

しかし，なぜこのような計算をする？

$\sqrt[3]{10\times5\times9}=\sqrt[3]{450}=7.6630943\cdots$

でもいいだろう．計算が少し面倒くさいが，今ならエクセルで対応できる．

今までの提言からデータを構成している基礎を学ぶ（2-2）

ここで，今まで勉強してきた簡単な数学が役立つ．一般化すれば，

n個のデータ，$x_1, x_2, \cdots, x_n$ があるとき，
平均値 $\bar{x}$ は，

$$\bar{x}=\frac{1}{n}(x_1+x_2+\cdots+x_n)=\frac{1}{n}\sum_{i=1}^{n}x_i$$

と書ける．ここで，

$x_1, x_2 \cdots, x_n$ の代表する値 a を考える．

今までの提言からデータを構成している基礎を学ぶ（2-3）

各値から a までの差を考える．この差をできるだけ小さくした a がきっとあるはずだ．つまり，

$$\min\{(x_1-a)+(x_2-a)+\cdots+(x_n-a)\}$$

ここでは，

$$W=(x_1-a)^2+(x_2-a)^2+\cdots+(x_n-a)^2$$

と考えて，W を最小とする a を考える．
こんなとき最小2乗法というものがあった．
それは，・・・

今までの提言からデータを構成している基礎を学ぶ（2-4）

誤差の2乗を最小するようなパラメータを求めるもの．そこで，

$$\frac{dW}{da}=0$$

を考えればよい．故に，

$$\frac{dW}{da}=2(x_1-a)+2(x_2-a)+\cdots+(x_n-a)=0$$

よって，

$$na=\sum_{i=1}^{n}x_i$$

$$a=\frac{1}{n}\sum_{i=1}^{n}x_i$$

n 個のデータからの誤差の2乗を最小にするようなパラメータは，私たちが考えている平均値の考え方と同じであった．

今までの提言からデータを構成している基礎を学ぶ（2-5）

平均値 $\bar{x}$ はやっぱり，

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

であったが，

$$\bar{x}_G=(x_1x_2\cdots x_n)^{1/n}$$

$$\bar{x}_H=\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}$$

という平均値もあって，

$\bar{x}_G$ を幾何平均，$\bar{x}_H$ を調和平均という．

今までの提言からデータを構成している基礎を学ぶ（3-1）

データを取り扱う上で必要なこと

例えば，「標準偏差」

境港の5つの店で塩鯖の切り身の値段を調査した．
5切れあたり，258，238，298，238，268円だった．

出雲では同じく任意の5店舗で，
5切れあたり，398，298，318，328，358円だった．

今までの提言からデータを構成している基礎を学ぶ（3-2）

平均値は，

$$\bar{x}_{境港}=\frac{1}{5}(258+238+298+238+268)=260$$

$$\bar{x}_{出雲}=\frac{1}{5}(398+298+318+328+358)=340$$

境港の方が出雲より平均値を中心として各データはばらついていないように感じる．そこで，各データから平均値をひいて，絶対値をとって総和し，平均を求めれば，

境港：$\frac{1}{5}\{|258-260|+|238-260|+\cdots+|268-260|\}=18.4$

出雲：$\frac{1}{5}\{|398-340|+|298-340|+\cdots+|358-340|\}=30.4$

今までの提言からデータを構成している基礎を学ぶ（3-3）

これは，出雲の方がデータはばらついていることを示している．このようにして計算した値を平均偏差という．絶対値をとらないと，

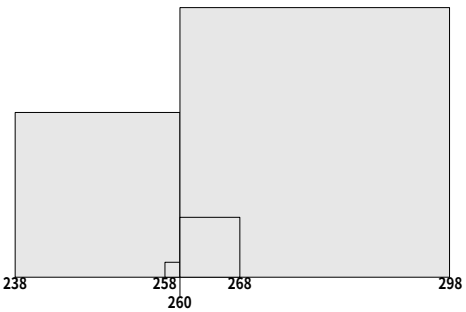
$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})=\frac{1}{n}\left(\sum_{i=1}^{n}x_i-n\bar{x}\right)=\frac{1}{n}\left(\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}x_i\right)=0$$

となってしまい，困る．

平均値の周囲に散らばるデータのばらつきをこのほかの計算式で書けないか．

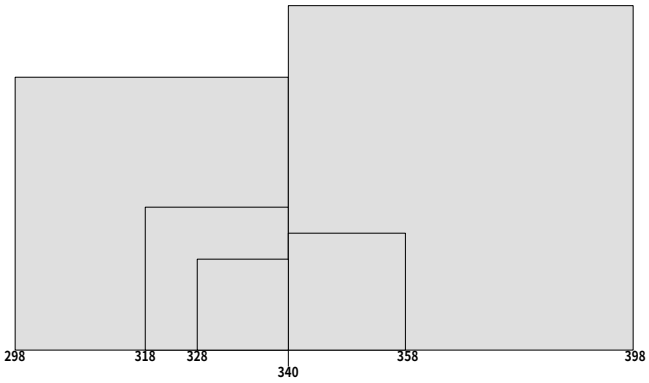
今までの提言からデータを構成している基礎を学ぶ（3-4）

先ほどの塩鯖データをこのように考える．



今までの提言からデータを構成している基礎を学ぶ（3-5）

先ほどの塩鯖データをこのように考える．



今までの提言からデータを構成している基礎を学ぶ（3-6）

先ほどの塩鯖データをこのように考える．

これらの正方形を総和する．一般化すれば，

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

この S を偏差平方和という．字が示すように各データから平均値までを示す偏差の2乗和．つまり正方形の面積の総和．

次に，

今までの提言からデータを構成している基礎を学ぶ（3-7）

正方形の平均値を求めよう．

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

この V を分散という．ばらばらに散らばることであるが，統計の世界では，偏差の2乗和を（算術）平均したもの．式からも明らかなように，平均的な正方形の面積．

次に，

今までの提言からデータを構成している基礎を学ぶ（3-8）

平均的な正方形 V では，例題の単位は円^2．こんな単位は見たことない．そこで，円にするためには

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

この s を標準偏差という．この式によって偏差を具体的に与えられたものの単位で示すことができた．

偏差平方和，分散，標準偏差はデータの散らばりを示す良い指標となっている．

今までの提言からデータを構成している基礎を学ぶ（3-9）

実際に塩鯖データで標準偏差を計算してみると，

$$s_{境港}=\sqrt{\frac{1}{5}(288-260)^2+(238-260)^2+(298-260)^2+(238-260)^2+(268-260)^2}$$
$$=\sqrt{\frac{2480}{5}}=22.27105\cdots$$

$$s_{出雲}=\sqrt{\frac{1}{5}(398-340)^2+(298-340)^2+(318-340)^2+(328-340)^2+(358-340)^2}$$
$$=\sqrt{\frac{6080}{5}}=34.87119\cdots$$

これらのデータからは，境港の方が店舗間格差は小さいだろう．

今までの提言からデータを構成している基礎を学ぶ（4-1）

ここまで学んだ平均値，標準偏差から新たな知識を・・・．

n個のデータ，$x_1, x_2, \cdots, x_n$ があるとき，こんな計算をしてみる．i 番目の新たな関数をx_i^* とおいて，

$$x_i^*=\frac{x_i-\bar{x}}{s}$$

とおけば，

$$\bar{x}^*=\frac{1}{n}\sum_{i=1}^{n}x_i^*=0$$

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

今までの提言からデータを構成している基礎を学ぶ（4-2）

ここまで学んだ平均値，標準偏差から新たな知識を・・・．

(証明)

$$\sum_{i=1}^{n} x_i^* = \frac{1}{s}\left(\sum_{i=1}^{n} x_i - n\bar{x}\right) = \frac{1}{s}\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i\right) = 0$$

よって，

$$\bar{x}^* = 0$$

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

である．

■

今までの提言からデータを構成している基礎を学ぶ（4-3）

ここまで学んだ平均値，標準偏差から新たな知識を・・・．

このような関数を異なった平均，分散を持つデータに使ってみる．例えば，

	数学	社会	理科	平均
Aさん	50	64	66	60
Bくん	60	60	60	60
科目の平均	50	60	60	
科目の標準偏差	10	4	4	

ここで，先ほどの計算をしてみる．

今までの提言からデータを構成している基礎を学ぶ（4-4）

ここまで学んだ平均値，標準偏差から新たな知識を・・・．

	数学	社会	理科	平均
Aさん	$\frac{50-50}{10}=0$	$\frac{64-60}{4}=1$	1.5	0.833
Bくん	1	0	0	0.333
平均	50	60	60	
標準偏差	10	4	4	

同じ平均値のAさん，Bくんに実は違いが見えてきた．

今までの提言からデータを構成している基礎を学ぶ（4-5）

ここまで学んだ平均値，標準偏差から新たな知識を・・・．

ここで，$50+10\frac{x_i-\bar{x}}{s}$ とすれば，

	数学	社会	理科	平均
Aさん	$50+10*\frac{50-50}{10}=50$	60	65	58.33
Bくん	60	50	50	53.33
平均	50	60	60	
標準偏差	10	4	4	

今までの提言からデータを構成している基礎を学ぶ（4-6）

ここまで学んだ平均値，標準偏差から新たな知識を・・・．

このように計算すれば，同じ平均点のふたりだが，Aさんの方がよい成績だと分かる．―― この数字が偏差値である．

	数学	社会	理科	平均
Aさん	$50+10*\frac{50-50}{10}=50$	60	65	58.33
Bくん	60	50	50	53.33
平均	50	60	60	
標準偏差	10	4	4	

今までの提言からデータを構成している基礎を学ぶ（4-7）

ここまで学んだ平均値，標準偏差から新たな知識を・・・．

このように $\frac{個々のデータ－平均値}{標準偏差}$ を標準化という．

これから分かるように，標準化したデータは無名数となる．

そして，標準化したデータは平均が0，標準偏差（あるいは分散）が1である．本当は偏差値は，$50+10\frac{x_i-\bar{x}}{s}$ のような計算をしなくてもいいのだが，実生活に近い形に正規化したものである．

今までの提言からデータを構成している基礎を学ぶ（4-8）

ここまで学んだ平均値，標準偏差から新たな知識を･･･．

偏差値への警鐘･･･

このようなデータがあったとき，

	A	B	C	D	E	平均	標準偏差
統計学概論	50	80	20	40	60	50	20
心理統計学	35	50	45	55	65	50	10
合計	85	130	65	95	125	100	24.5
合計点の順位	4	1	5	3	2		

今までの提言からデータを構成している基礎を学ぶ（4-9）

ここまで学んだ平均値，標準偏差から新たな知識を･･･．

偏差値への警鐘･･･

偏差値は

	A	B	C	D	E
統計学概論	$50+10\times\frac{50-50}{20}=50$	65	35	45	55
心理統計学	$50+10\times\frac{35-50}{10}=35$	50	45	55	65
偏差値の平均	$\frac{1}{2}(50+35)=42.5$	**57.5**	40	50	60
合計点の偏差値	$50+10\times\frac{85-100}{24.5}=44$	62	36	48	60

このことは，何を意味しているのだろう．偏差値は完璧だろうか？

変な差をつける値ではないと証明はできない．

今までの提言からデータを構成している基礎を学ぶ（5-1）

ここまで学んだ平均値，標準偏差から新たな知識を･･･．

変動係数（$C.V.$）

―標準偏差は同じ単位で測定されたデータでのみ意味を持つ．この不便を除くため，単位を含まない次の指標を考える―

$$C.V.=\frac{s}{\bar{x}}$$

あるいは

$$C.V.(\%)=\frac{s}{\bar{x}}*100$$

今までの提言からデータを構成している基礎を学ぶ（5-2）

ここまで学んだ平均値，標準偏差から新たな知識を･･･．

先ほどの塩鯖データから，

$$C.V._{\text{境港}}=\frac{22.271}{260}=8.567\times10^{-2}$$

$$C.V._{\text{出雲}}=\frac{34.871}{340}=1.026\times10^{-1}$$

これから，塩鯖は境港の方が店舗間格差は小さいと思われる．

今までの提言からデータを構成している基礎を学ぶ（6-1）

異なった2変量の関係

こんなデータがあったとする．

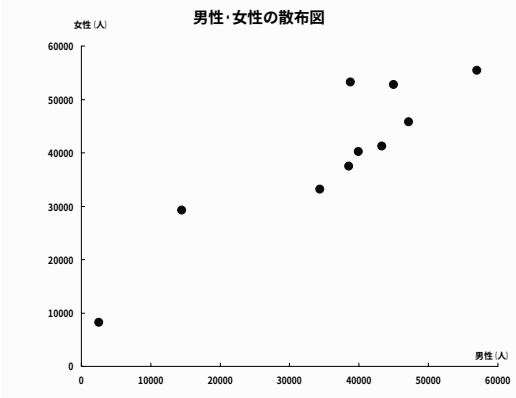
2001年における島根県の各世代の人口は以下のとおりである．

	10歳未満	10～19	20～29	30～39	40～49	50～59	60～69	70～79	80～89	90歳以上
男性	34385	43310	38541	39937	47143	56974	45013	38774	14457	2541
女性	33148	41253	37499	40234	45800	55422	52763	53229	29229	8230

これを，･･･

今までの提言からデータを構成している基礎を学ぶ（6-2）

異なった2変量の関係



今までの提言からデータを構成している基礎を学ぶ（6-3）

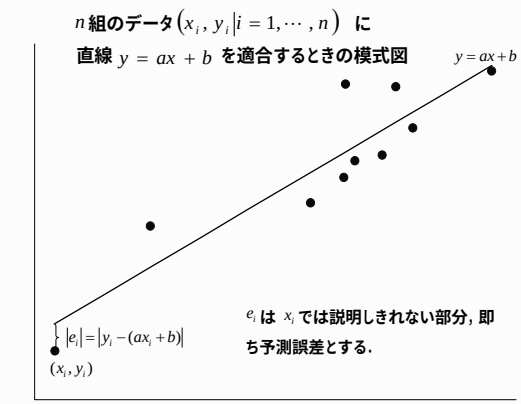
異なった2変量の関係

1. このグラフから，男性を x ，女性を y として，簡単な式，それも直線 $y=ax+b$ で示すことはできないか．
2. この直線は，各点から直線までの距離の2乗和が最小になるように引こう．
3. このとき，各点は直線の上あるいは下にあるかは任意である．

という制約条件をつけよう．一般化して，

今までの提言からデータを構成している基礎を学ぶ（6-4）

異なった2変量の関係



今までの提言からデータを構成している基礎を学ぶ（6-5）

異なった2変量の関係

表にまとめると，

データ番号	x	y	e
1	x_1	y_1	$e_1=y_1-(ax_1+b)$
2	x_2	y_2	$e_2=y_2-(ax_2+b)$
⋮			
n	x_n	y_n	$e_n=y_n-(ax_n+b)$

今までの提言からデータを構成している基礎を学ぶ（6-6）

異なった2変量の関係

ここで私たちは，予測誤差の平方和

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i-(ax_i+b)\}^2$$

を最小とするような a, b を求めよう．

このときに，平均値を求めるときに使った最小2乗法を用いる．

今までの提言からデータを構成している基礎を学ぶ（6-7）

異なった2変量の関係

即ち，未知の定数 a, b を関数として，

$$F(a, b)=\sum_{i=1}^{n} \{y_i-(ax_i+b)\}^2$$

が最小となるように，a, b を求める．ここでは，上の式が最小値のところで微分係数が0となるから，

$$\frac{\partial F(a,b)}{\partial a}=0$$
$$\frac{\partial F(a,b)}{\partial b}=0$$

から a, b を求めればよい．

今までの提言からデータを構成している基礎を学ぶ（6-8）

異なった2変量の関係

$$\frac{\partial F(a,b)}{\partial a}=\sum_{i=1}^{n} 2(y_i-ax_i-b)(-x_i)=0 \qquad (1)$$

$$\frac{\partial F(a,b)}{\partial b}=\sum_{i=1}^{n} 2(y_i-ax_i-b)(-1)=0 \qquad (2)$$

(2)式より

$$\sum_{i=1}^{n} y_i - bn - a\sum_{i=1}^{n} x_i = 0$$

$$b=\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{a}{n}\sum_{i=1}^{n} x_i = \bar{y}-a\bar{x}$$

を得る．

今までの提言からデータを構成している基礎を学ぶ（6-9）

異なった2変量の関係

b を(1)式，(2)式に代入すれば

$$\sum_{i=1}^{n}\{(y_i-\bar{y})-a(x_i-\bar{x})\}=0 \qquad (3)$$

$$\sum_{i=1}^{n}\{(y_i-\bar{y})-a(x_i-\bar{x})\}x_i=0 \qquad (4)$$

{(4)式−(3)式×$\bar{x}$}×1/nを計算すれば

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})-\frac{a}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=0$$

を得る．

今までの提言からデータを構成している基礎を学ぶ（6-10）

異なった2変量の関係

よって，

$$a=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{c(x,y)}{s^2(x)}$$

$c(x,y)$ を x と y の共分散といい，2変量間の分散の指標となっている．今までの計算から，

$$a=\frac{c(x,y)}{s^2(x)},\quad b=\bar{y}-\frac{c(x,y)}{s^2(x)}\bar{x}$$

を得る．

今までの提言からデータを構成している基礎を学ぶ（6-11）

異なった2変量の関係

	10歳未満	10～19	20～29	30～39	40～49	50～59	60～69	70～79	80～89	90歳以上
男性	34385	43310	38541	39937	47143	56974	45013	38774	14457	2541
女性	33148	41253	37499	40234	45800	55422	52763	53229	29229	8230

$$s^2(x)=\frac{1}{10}\{(34385-36107.5)^2+\cdots+(2541-36107.5)^2\}=231449531.3$$

$$s^2(y)=\frac{1}{10}\{(33148-39680.7)^2+\cdots+(8230-39680.7)^2\}=178853618$$

$$c(x,y)=\frac{1}{10}\{(34385-36107.5)\times(33148-39680.7)+\cdots+(2541-36107.5)\times(8230-39680.7)\}=184998755$$

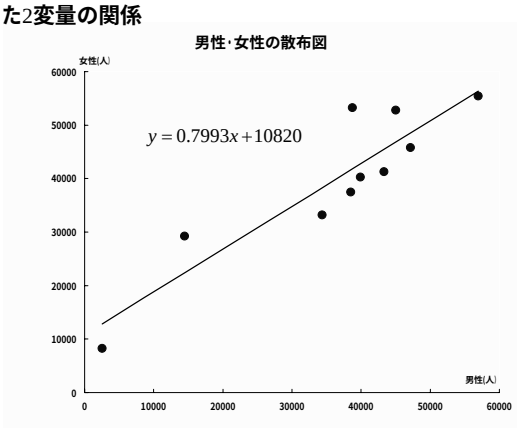
故に，

$$a=\frac{c(x,y)}{s^2(x)}=\frac{184998755}{231449531.3}=0.799305$$

$$b=\bar{y}-\frac{c(x,y)}{s^2(x)}\bar{x}=39680.7-0.799305\times36107.5=10819.80$$

今までの提言からデータを構成している基礎を学ぶ（6-12）

異なった2変量の関係

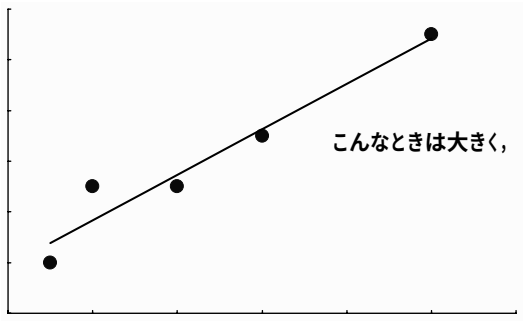


今までの提言からデータを構成している基礎を学ぶ（6-13）

異なった2変量の関係

直線近傍に分布していれば，大きな値に，そうでないときには小さな値となるような計算はないか？

例えば，

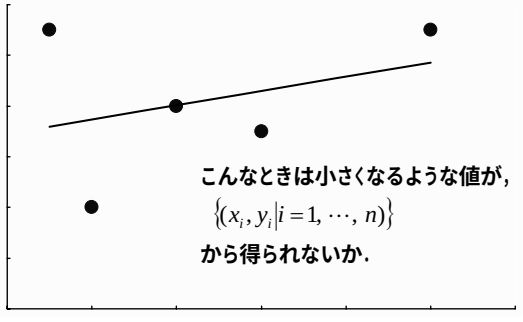


今までの提言からデータを構成している基礎を学ぶ（6-14）

異なった2変量の関係

直線近傍に分布していれば，大きな値に，そうでないときには小さな値となるような計算はないか？

例えば，



今までの提言からデータを構成している基礎を学ぶ（6-15）

異なった2変量の関係

得られる・・・．

$$W=\frac{1}{n}\sum_{i=1}^{n}\{y_i-(ax_i+b)\}^2$$

とおいて，$a=\dfrac{c(x,y)}{s^2(x)}$，$b=\bar{y}-a\bar{x}$ を代入すれば，

$$\begin{aligned}W&=\frac{1}{n}\sum\{(y_i-\bar{y})-a(x_i-\bar{x})\}^2\\&=\frac{1}{n}\sum(y_i-\bar{y})^2-2a\left\{\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})\right\}+a^2\left\{\frac{1}{n}\sum(x_i-\bar{x})^2\right\}\\&=s^2(y)-2\frac{c^2(x,y)}{s^2(x)}+\frac{c^2(x,y)}{s^2(x)}=s^2(y)\left\{1-\frac{c^2(x,y)}{s^2(x)s^2(y)}\right\}\end{aligned}$$

今までの提言からデータを構成している基礎を学ぶ（6-16）

異なった2変量の関係

これより，　$\dfrac{W}{s^2(y)}$

が，0近傍にあれば，直線の近くにデータは集中している．1に近ければ，集中はしていないことが分かるだろう．

このことは，

$$\frac{c(x,y)}{s(x)s(y)}$$

が大きいほど直線近くにデータは集中していることだから，――――

今までの提言からデータを構成している基礎を学ぶ（6-17）

異なった2変量の関係

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

とおいて，この r を x と y の相関係数と呼ぼう．

この相関係数は，2変量がどの程度直線的な傾向を示している指標である．

今までの提言からデータを構成している基礎を学ぶ（6-18）

異なった2変量の関係

実際に計算してみよう．

総理府統計局1999年度小売業関係統計

都道府県	商店数	従業員数（人）	年間販売額（10億円）	人口（千人）
北海道	54396	376654	7117	5683
青森県	18740	94886	1637	1476
⋮	⋮	⋮	⋮	⋮
東京都	128510	813885	17410	12064
⋮	⋮	⋮	⋮	⋮
鳥取県	7634	38826	704	613
島根県	11580	50337	846	762
⋮	⋮	⋮	⋮	⋮
沖縄県	17945	75135	985	1318

今までの提言からデータを構成している基礎を学ぶ（6-19）

異なった2変量の関係

人口とその他のパラメータで考える．計算すれば，

$$s_{\text{商店数}}=23631,\ s_{\text{従業員数}}=156752,$$
$$s_{\text{年間販売額}}=3150,\ s_{\text{人口}}=2490$$

$$c_{\text{人・商}}=57577181,\ c_{\text{人・従}}=387947341,$$
$$c_{\text{人・販}}=7712849$$

今までの提言からデータを構成している基礎を学ぶ（6-20）

異なった2変量の関係

求める相関係数は，

$$r_{\text{人・商}}=0.9785$$

$$r_{\text{人・従}}=0.9940$$

$$r_{\text{人・販}}=0.9834$$

それぞれ1に非常に近い．このことはやはり，小売業は人口に依存した産業であることを示している．

今までの提言からデータを構成している基礎を学ぶ（7-1-1）

今日述べなかったこと

r がどの程度あれば，本当に相関はあるのか．相関係数の範囲は，$-1 \leq r \leq 1$ であることの証明．

相関があるかどうかは，勿論相関係数に依存しているだろうが，他のパラメータで相関の有無を決定してはいないか．

幾何平均，調和平均はどのようなときに使えばいいのか．

今までの提言からデータを構成している基礎を学ぶ（7-1-2）

今日述べなかったこと

r がどの程度あれば，本当に相関はあるのか．

相関があるかどうかは，勿論相関係数に依存しているだろうが，他のパラメータで相関の有無を決定してはいないか．

幾何平均，調和平均はどのようなときに使えばいいのか．

今までの提言からデータを構成している基礎を学ぶ（7-1-3）

今日述べなかったこと

r がどの程度あれば，本当に相関はあるのか．

相関があるかどうかは，勿論相関係数に依存しているだろうが，他のパラメータで相関の有無を決定してはいないか．

幾何平均，調和平均はどのようなときに使えばいいのか．

今までの提言からデータを構成している基礎を学ぶ（7-2-1）

今日述べなかったこと

	変量1	変量2		変量 p	y_i
1	x_{11}	x_{12}		x_{1p}	y_1
2	x_{21}	x_{22}		x_{2p}	y_2
⋮					
n	x_{n1}	x_{n2}		x_{np}	y_n

のようなとき，y はどう表すのか．

のようなとき，相関係数に相当するものがあるか．

今までの提言からデータを構成している基礎を学ぶ（7-2-2）

今日述べなかったこと

	変量1	変量2		変量 p	y_i
1	x_{11}	x_{12}		x_{1p}	y_1
2	x_{21}	x_{22}		x_{2p}	y_2
⋮					
n	x_{n1}	x_{n2}		x_{np}	y_n

のようなとき，y はどう表すのか．

のようなとき，相関係数に相当するものがあるか．

今までの提言からデータを構成している基礎を学ぶ（8）

参 考 文 献

- SASによる回帰分析
- スネデカー･コクラン統計的方法

  これらの本は実データを用いて優しく解説してある．
- 回帰分析（佐和隆光著）

  線形代数を理解しながら読めば目から鱗が落ちる．