情報科学

統計解析(2)

島根大学医学部医療情報学講座
安田 晃

前回の復習 (1)

- 何となくデータを見るだけでは科学性は得られない.
  - 論拠に客観性を加えてほしい.
- 基礎統計量からの情報も有用である.
  - 平均, 偏差平方和, 分散, 標準偏差
- 基礎統計量を関数にした情報も有用である.
  - データの標準化
- 今後はデータの背後に隠れているものを探し出してほしい.

前回の復習 (2)

- n 個のデータ $x_i\ (i=1,\cdots,n)$ の平均値 $\bar{x}$ は,

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

となるが, この $\bar{x}$ は, n 個のデータからの2乗和が最小になる値であった.

- 平均値はこの他に,

幾何平均 $\bar{x}_G=\sqrt[n]{x_1x_2\cdots x_n}$

調和平均 $\bar{x}_H=\dfrac{n}{\dfrac{1}{x_1}+\cdots+\dfrac{1}{x_n}}$

もあった.

前回の復習 (3)

- データの散らばりを, 偏差平方和 S , 分散 V , 標準偏差 s で計ることができた.

$$S=\left(x_1-\bar{x}\right)^2+\cdots+\left(x_n-\bar{x}\right)^2=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$$

$$V=\frac{1}{n}\left\{\left(x_1-\bar{x}\right)^2+\cdots+\left(x_n-\bar{x}\right)^2\right\}=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$$

$$s=\sqrt{\frac{1}{n}\left\{\left(x_1-\bar{x}\right)^2+\cdots+\left(x_n-\bar{x}\right)^2\right\}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}$$

前回の復習 (4)

- 平均値 $\bar{x}$ と標準偏差 s を使って, i 番目のデータを

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

という新たな関数をつくるとき, 平均値と標準偏差はそれぞれ,

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

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

である.

前回の復習 (5)

- このようにデータを変換することを, データの標準化といい, 新たな関数は無名数となる.
- このような変換は, 平均を50点, 標準偏差を10点とする偏差値に見られるが, これだけではない.
- 今後皆さんの周りに溢れてくるデータの山を, 主成分や類似性, 相関など様々な手法で解析するきっかけを述べました.

今日の予定

- 2つの変量の関係を考える(I).
  - (単)回帰式, 相関係数
- 2つの変量の関係を考える(II).
  - 重回帰式, 重相関係数
- 理解と実践
  - 隠れた構造を探るため.

2つの変量の関係を考える(I-1)

- 疑問:本は厚けりゃ値段が高いと感じているが, 本当だろうか.
  - そこで, ジャンルに無関係で値段とページ数を見た.

・・・, 調べた範囲で次のようなデータを得た.

2つの変量の関係を考える(I-2)

タ イ ト ル	価格	ページ数
脳を鍛える大人の計算ドリル	1050	166
ナースのための社会学入門	2520	158
北日本の縄文後期土器編年の研究	16800	228
最新基本地図	2625	198
環境ホルモンのしくみ	1470	163
金属錯体化学	9240	240
数値計算法の基礎と応用	3990	313
テキスト微分積分	2100	194
森林野生動物の調査	3570	287
数の不思議	1470	120
木に学べ	777	251
看護に役立つ数式事典	1020	198
看護系の化学	1427	151
中也ノオト―私と中原中也	1575	126
自分でやりたい人の最新バイク・メンテナンス	1260	143
中国日帰り温泉―広島・岡山・山口・鳥取・島根	1365	175
フィーザー基礎有機化学	2940	422
図解でわかる回帰分析	1480	218
わかりやすい高分子化学	3360	276
心を測る	2940	156

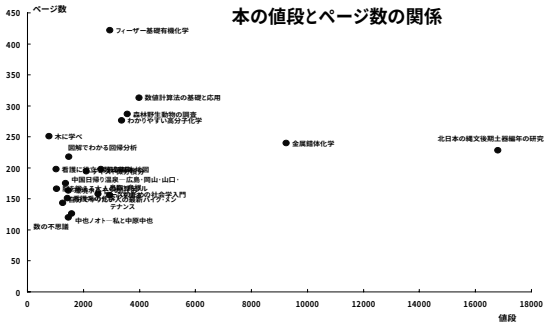
2つの変量の関係を考える(I-3)

値段の安い順に20タイトルを並べ替えれば,

タ イ ト ル	価格	ページ数
木に学べ	777	251
看護に役立つ数式事典	1020	198
脳を鍛える大人の計算ドリル	1050	166
自分でやりたい人の最新バイク・メンテナンス	1260	143
中国日帰り温泉―広島・岡山・山口・鳥取・島根	1365	175
看護系の化学	1427	151
環境ホルモンのしくみ	1470	163
数の不思議	1470	120
図解でわかる回帰分析	1480	218
中也ノオト―私と中原中也	1575	126
テキスト微分積分	2100	194
ナースのための社会学入門	2520	158
最新基本地図	2625	198
フィーザー基礎有機化学	2940	422
心を測る	2940	156
わかりやすい高分子化学	3360	276
森林野生動物の調査	3570	287
数値計算法の基礎と応用	3990	313
金属錯体化学	9240	240
北日本の縄文後期土器編年の研究	16800	228

2つの変量の関係を考える(I-4)

値段が安いものは薄く, 高いものが厚いということは, このデータをプロットすれば, 右肩上がりとなるはずである.



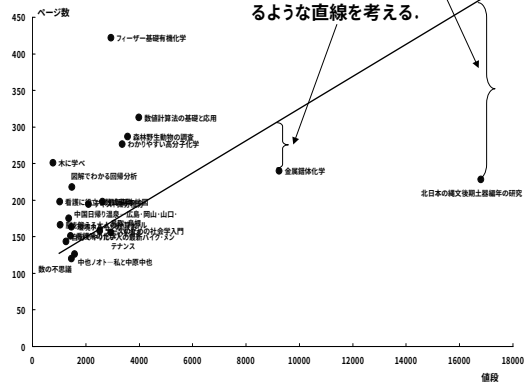
2つの変量の関係を考える(I-5)

- この情報をもとにして, 何らかの直線を引くことはできないか. そこで, いくつかの拘束条件をつける.
  - 求める直線の方程式は $y = ax + b$ とする.
  - 各点が直線から上にあるか, 下にあるかの確率は50%とする.
  - 直線は各点からの距離の2乗和が最小となるように引く.

図に示せば・・・

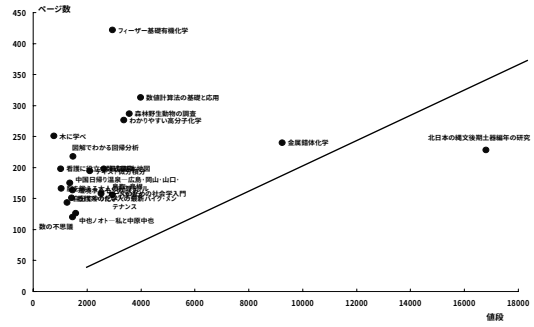
2つの変量の関係を考える（I -6）

この距離（ここでは20個の距離）の2乗和が最小になるような直線を考える．



2つの変量の関係を考える（I -7）

こんな直線でもいいが，2番目の拘束条件を満たしていない．



2つの変量の関係を考える（I -8）

- ここでこの問題は，誤差の2乗和を最小とするような傾きと切片を求める問題に変わった．
  - これには，平均値のところで行った最小2乗法を使おう．一般化して示したい．

2つの変量の関係を考える（I -9）

- 今，n組のデータ$\left(x_i, y_i \middle| i=1,\cdots,n\right)$がある．求める直線の方程式を $y=ax+b$ とするとき，各点と直線までの距離$e_i\ (i=1,\cdots,n)$ は，

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	$e_n=y_n-(ax_n+b)$

である．

2つの変量の関係を考える（I -10）

- ここで，

$$W=e_1^2+\cdots e_n^2=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}\{y_i-(ax_i+b)\}^2$$

とおいて，W が最小となるような a と b を最小2乗法を使って求めよう．これは，W を a と b で偏微分して0とおけばよかった．即ち，

2つの変量の関係を考える（I -11）

$$\frac{\partial W}{\partial a}=-2\sum_{i=1}^{n}x_i\{y_i-(ax_i+b)\}=-2\left\{\sum_{i=1}^{n}x_iy_i-a\sum_{i=1}^{n}x_i^2-b\sum_{i=1}^{n}x_i\right\}=0$$

$$\frac{\partial W}{\partial b}=-2\sum_{i=1}^{n}\{y_i-(ax_i+b)\}=-2\left\{\sum_{i=1}^{n}y_i-a\sum_{i=1}^{n}x_i-nb\right\}=0$$

下の式は

$$b=\frac{1}{n}\left(\sum_i y_i-a\sum_i x_i\right)=\bar{y}-a\bar{x}$$

と書けるので，上の式に代入することによって，

2つの変量の関係を考える（I -12）

$$\sum_i x_i y_i - a\sum_i x_i^2 - n\bar{y}\bar{x} + na\bar{x}^2 = 0$$

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

から，

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

を得る．a の分子を x と y の共分散という．

2つの変量の関係を考える（I -13）

- 今までの一般化を，先ほどの値段-ページ数データに当てはめてみよう．ここでは，値段を x_i ，ページ数を y_i とすれば，

$$\bar{x} = 3148.95,$$

$$\bar{y} = 209.15$$

$$\frac{1}{20}\sum_{i=1}^{20} (x_i - \bar{x})(y_i - \bar{y}) = 62215.56$$

$$\frac{1}{20}\sum_{i=1}^{20} (x_i - \bar{x})^2 = 13070966$$

これより，

2つの変量の関係を考える（I -14）

$$a = \frac{\frac{1}{20}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{20}\sum_i (x_i - \bar{x})^2} = \frac{62215.65}{13070996} = 0.004760 = 4.760\times 10^{-3}$$

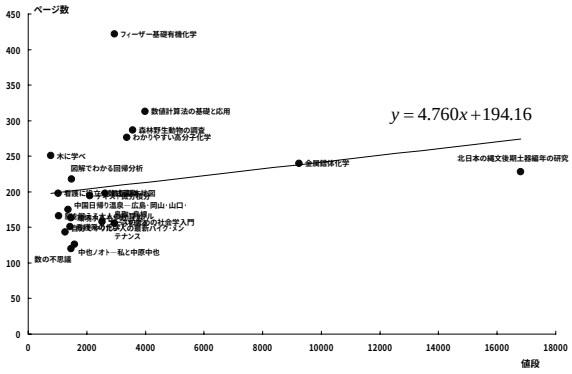
$$b = \bar{y} - a\bar{x} = 209.15 - (4.760\times 10^{-3})\times 3148.95 = 194.16$$

を得る．よって，20冊の本を値段とページ数でプロットした場合，各点からの直線までの距離の2乗和を最小するような直線の方程式は，

$$y = 4.760x + 194.16$$

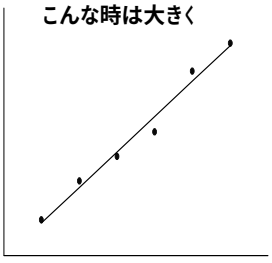
と計算できた．

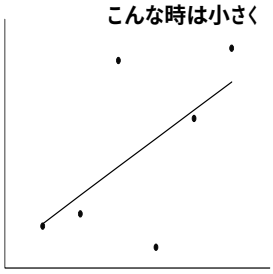
2つの変量の関係を考える（I -15）



2つの変量の関係を考える（I -16）

- ここで，このようにして得られた直線の近くにデータが集まっていると大きな値，離れていると小さな値となるようなものを考えてみよう．即ち，





2つの変量の関係を考える（I -17）

- 結果だけを示そう．データが直線近傍に集中しているとき大きなく，直線から離れているときは小さくなるような値を相関係数 r と呼んで，以下のように計算される．

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

- 先ほどの値段-ページ数データで計算してみれば，

$$\bar{x} = 3148.95,\ s(x) = 3615.38$$
$$\bar{y} = 209.15,\ s(y) = 71.921$$
$$n = 20$$

なので，

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
$$= \frac{\frac{1}{20}\{(777-3148.25)(251-209.15)+\cdots+(16800-3148.25)(228-209.15)}{3615.38\times 71.921}$$
$$= \frac{62215.56}{260021.6} = 0.2393$$

- 先ほどの r の範囲は，

$$-1 \le r \le 1$$

であり，本当に相関があるかどうか検定するためには，

$$t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$$

が，自由度$n-2$のt分布に従うことを利用する．

- 実際に t を計算してみれば，

$$r = 0.2393$$
$$n = 20$$

だから，

$$t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.2393\cdot\sqrt{20-2}}{\sqrt{1-0.2393^2}}$$
$$= \frac{1.0153}{0.9709} = 1.0457$$

- 自由度18で，$t = 1.0457$ の有意水準 α は，

$$\alpha = 0.3095$$

この α が0.05以下であれば，有意に相関があると判断できる．

ここでは，$\alpha > 0.05$ なので，

「本の値段とページ数には相関が認められない」

と判断する．

- 以上より，20個の2変量をプロットした場合の，
  - 各点からの距離の2乗和が最小となる直線を引くことができた．何となれば傾きと切片が求まったから．このような直線を回帰直線といい，求解を直線回帰という．
  - その直線近傍にデータが集中しているときは大きく，離れているときは小さくなるような値，相関係数を計算できた．
  - 相関係数の検定を，相関係数とデータの組数を関数として計算できた．

注意： 但し，直線回帰，相関係数の計算には，平均値や標準偏差が計算できるデータで行わなくてはならない．

- 今までのことを整理して，

n 組のデータ $\left(x_i, y_i \middle| i = 1,\cdots,n\right)$ があるとき，x を独立変数，y を従属変数とした場合，n 個の点から誤差の2乗和を最小となるように引く直線の傾きを a ，切片を b とするとき，

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

であった．

- 今までのことを整理して，

n組のデータ $\left(x_i, y_i \,\middle|\, i=1,\cdots,n\right)$から回帰直線が得られたとき，相関係数 r は，

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

であった． r の範囲は，

$$-1\le r\le 1$$

である．

- ここで， r が有意なものであるかを探るため，

$$t=\frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$$

として，自由度$n-2$ の t 分布から有意水準を求めた．

- 再度検証してみれば，

人口，65歳以上人口は，×10000．
車所有率は人口1万にあたり．

	人口	65歳以上人口	理美容院数	映画館数	車所有率
北海道	565.9	107.4	16694	112	3825
青森県	147.4	29.7	6268	32	3188
岩手県	141.3	31.3	5687	33	3351
宮城県	237.3	42.3	7389	23	3646
秋田県	118.4	28.8	5777	20	3414
山形県	124.1	29.2	5391	27	3574
福島県	211.3	44.3	7175	34	3819
茨城県	299.1	51.2	9022	45	4558
栃木県	201.1	35.6	6438	23	4605
群馬県	203.4	37.9	6322	33	4652
埼玉県	702.9	94.5	14828	51	3431
千葉県	602.4	88.7	12659	83	3467
東京都	1231	200.5	26420	226	2340
神奈川県	868.7	123.8	15245	84	3016
新潟県	246	54.2	8313	28	3418
富山県	112.1	24	3383	25	4074
石川県	118.2	22.7	3592	14	3986
福井県	83	17.5	2632	16	3950
山梨県	89	17.8	2859	13	4053
長野県	221.5	48.7	6197	48	3988
岐阜県	211.1	39.6	6520	26	4157
静岡県	379.3	69	10885	55	3845
愛知県	715.8	107.1	15933	112	4081
三重県	186.2	36.3	5435	41	3898
滋賀県	135.3	22.3	2843	14	3516
京都府	264.1	47.9	6342	30	2797
大阪府	881.6	138.2	20111	119	2491
兵庫県	558.5	97.9	12043	80	2944
奈良県	144.2	24.9	2891	13	3138
和歌山県	106.6	23.2	3877	20	2987
鳥取県	61.3	13.8	2053	14	3212
島根県	76.1	19.4	2665	8	3055
岡山県	195.3	40.5	5576	25	3467
広島県	287.8	54.9	8107	55	3142
山口県	152.4	34.9	4653	27	3298
徳島県	82.2	18.5	3237	14	3430
香川県	102.2	22	3344	11	3350
愛媛県	149.1	32.8	5449	29	2929
高知県	81.3	19.6	3210	15	2878
福岡県	505.1	90	12874	158	3246
佐賀県	87.6	18.3	2411	6	3186
長崎県	151.3	32.4	4803	42	2563
熊本県	185.5	40.6	6482	36	3269
大分県	122.1	27.3	4297	26	3374
宮崎県	116.9	24.8	4148	25	3345
鹿児島県	178.3	41.1	6184	26	3144
沖縄県	132.9	19.3	4681	27	3199

- 人口を独立変数にして，65歳以上人口，映画館数，理美容院数，車所有率を従属変数として見てみれば，

65歳以上人口＝0.1477×人口＋8.515

理美容院数＝19.604×人口＋1977.0

映画館数＝0.1483×人口＋2.7625

車所有率＝-0.5269×人口＋3596.3

- 相関係数も考えてみる．このような時，4変数から2変数をとる組み合わせは，

$${}_5C_2=(5\times4)/(2\times1)=10\ (\text{通り})$$

そこで，以下のように表現しよう．

	人口	65歳以上	理美容院	映画館	車所有
人口	-				
65歳以上	0.9882	-			
理美容院	0.9740	0.9858	-		
映画館	0.8941	0.9268	0.9175	-	
車所有	-0.2573	-0.2696	-0.2107	-0.2636	-

- この相関係数の r と有意水準は，以下の通り．

	人口	65歳以上	理美容院	映画館	車所有
人口	-				
65歳以上	43.215 2.741×10^{-38}	-			
理美容院	28.815 1.208×10^{-30}	39.417 1.550×10^{-36}	-		
映画館	13.392 2.633×10^{-17}	16.558 9.140×10^{-21}	15.472 1.239×10^{-19}	-	
車所有	1.7859 0.0809	1.8778 0.0669	1.44599 0.1551	1.83306 0.07341	-

いずれの変数も人口に大きく関わっていることが，車に関してはそうはいえない．

2つの変量の関係を考える（II-1）

- 注意してほしいのは，今までのような計算は，あくまで平均値や標準偏差が計算できるデータに限られる．次のようなデータではそれができない．

u, v 2人が6つの食べ物を選好する順位をつけた．2人の好みは似ているだろうか．

	納豆	鯖味噌煮	ブイヤベース	ボンゴレ	ラーメン	もんじゃ焼き
u	1	2	6	3	4	5
v	5	6	1	4	3	2

2つの変量の関係を考える（II-2）

- このような時，Aさんの平均は計算できないし，分散や標準偏差も同様である．このようなときには，以下のように順位相関係数 r_Sを計算する．

$$r_S = 1 - \frac{6\sum_{i=1}^{n}(u_i - v_i)^2}{n(n^2-1)}$$

先ほどのデータから計算してみよう．

2つの変量の関係を考える（II-3）

- 計算結果

	納豆 1	鯖味噌煮 2	ブイヤベース 3	ボンゴレ 4	ラーメン 5	もんじゃ焼き 6
u	1	2	6	3	4	5
v	5	6	1	4	3	2
$u_i - v_i$	-4	-4	5	-1	1	3
$(u_i - v_i)^2$	16	16	25	1	1	9

$$\sum_{i=1}^{6}(u_i - v_i)^2 = 16 + 16 + \cdots + 9 = 68$$

より，

2つの変量の関係を考える（II-4）

$$r_S = 1 - \frac{6\sum_{i=1}^{n}(u_i - v_i)^2}{n(n^2-1)}$$
$$= 1 - \frac{6\times 68}{6\times(6^2-1)} = 1 - 1.9429 = -0.9429$$

r_S の範囲は，

$$-1 \le r_S \le 1$$

-1ではまったく逆順に，1ではまったく正順になっている．
順位相関係数は-0.9429なので2人の好みは随分違っているようだ．

2つの変量の関係を考える（II-5）

- r_S を相関係数のように検定できないか．

できる．但し，以下のような近似式である．

$$r = 2\sin\left(\frac{\pi}{6}r_s\right)$$

r で近似されたなら， r の有意性を求めることができる．

2つの変量の関係を考える（II-6）

- 先ほどの選好データでは，

$$r_S = -0.9429$$

だから，
$$r = 2\sin\left(\frac{\pi}{6}r_s\right) = 2\sin(-0.4937)$$
$$= -0.4764$$ 但し，括弧内はラジアン

$$t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{|-0.4764|\sqrt{6-2}}{\sqrt{1-0.4764^2}} = 2.1674$$

自由度4の有意水準は0.0961．近似式では
r はかなり小さくなる傾向がある．

2つの変量の関係を考える（II-7）

- このように質問紙などで得られたデータの2変数の関連を調べるために，順位相関係数で知ることもある．あるいは大小関係の情報だけや分散が大きいもの，簡易な結果をまず求めたいときなど汎用される．
- しかし，ある基準に立脚した2変量の関係性という立場は理解しなければならない．
- 日常的に計算してみるといい． r の近似や有意水準は求めることができないが．（ケータイの）電卓で十分計算できる．

理論と実践

今までのことは基礎統計量の範疇に入るが，理論的な裏付けがあります．そのことを十分理解して，あなたたちが得るであろう，実データを解析してください．

データはこんな感じで終わるのではなく，例えば代表する値とかデータの分散，クロス集計，相関，順位相関などデータの潜在性を引き出す様々な手法がたくさんあります．

それらが必要かどうかは個人に依存する問題です．

来週の予定

- クロス集計表のデータ構造．
- 理論値と実測値の乖離
- 検定統計量
- 連関係数
- 連関の程度