情報科学

統計解析(3)

島根大学医学部医療情報学講座
安田 晃

前回の復習 (1)

- 本の値段とページ数を考えた.
  - 値段が高い本ほど厚かもしれない.
- 表を得た. それをグラフにしてみよう.
  - 高けりゃ厚いと右肩上がりとなるはずだ.
- そのようなプロットになっているか.
  - なっているかもしれない.
- そんな時, この図をある値で表現したい.

前回の復習 (2)

- 一般化して, n 組のデータ $(x_i, y_i \,|\, i = 1, \cdots, n)$ があったとき,

これらのデータの間を, 各点からの距離の2乗和が

最小となるように直線は引けないものか.

前回の復習 (3)

- 一般化して, n 組のデータ $(x_i, y_i \,|\, i = 1, \cdots, n)$ があったとき,

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.

という拘束条件をつけて, 未知数 a と b を求める問題としよう.

前回の復習 (4)

- この問題は,

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

としたとき, W を最小とする傾き a と切片 b を求める問題となった.

この問題は最小2乗法で解を求めればよかった. 即ち,

前回の復習 (5)

- W を a と b で偏微分して0とおけばよかった.

W を a で偏微分するときは b は定数とみなす. b で偏微分するときは a を定数とみなせばよかった.

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

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

である. この連立方程式より,

前回の復習 (6)

- パラメータ a, b は，

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

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

と計算できた．a の分子を x と y の共分散という．

前回の復習 (7)

- このようにもとまった直線の近傍にデータが集まっているときは大きな値，離れているときは小さな値となるような，うまい関係はないか．
  - ある．そんな値を相関係数 (r) と呼んで，

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

と示す．

前回の復習 (8)

- これら回帰直線，相関係数は，平均や標準偏差が計算できるようなデータに限り計算できる．例えば，ようなデータでは計算しない．

	アミノサプリ	燃焼系	凹	アミノ飲料	ポカリ
評価者 I	4	5	3	1	2
評価者 II	4	5	3	2	1

そこで，このような順位データに関しては，以下のような指標を用いよう．

前回の復習 (9)

- 一般化して，

	対象 1	対象 2	…	対象 n
順位 u	$n-2$	1	⋯	7
順位 v	$n-1$	5	⋯	1

であったとき，u と v の順位相関係数 r_S は，

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

である．

今日の予定

- クロス集計表のデータ構造
- 実測値と理論値の乖離
- 検定統計量
- 連関係数
- これで何がわかるか

クロス集計表のデータ構造 (1)

今，次のような質問があったとする．

1．あなたは昨日の夕食にいくら使いましたか (以下，夕食)．
①0円 ②1〜150円 ③151〜250円 ④251〜500円 ⑤501円以上

2．バイトの時給はいくらですか (以下，バイト)．
①〜500円 ②501〜700円 ③701〜900円 ④901〜1100円 ⑤1100円以上

この質問をバイトをしているS大学学生100人に聞いて以下の表を得た．

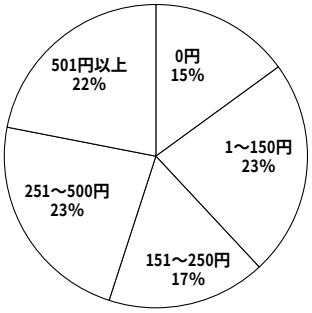
クロス集計表のデータ構造 (2)

学生	夕食	バイト
1	5	4
2	2	1
3	3	3
⋮	⋮	⋮
100	4	5

この表から何が言えるのだろうか．夕食，バイトそれぞれ考えてみよう．

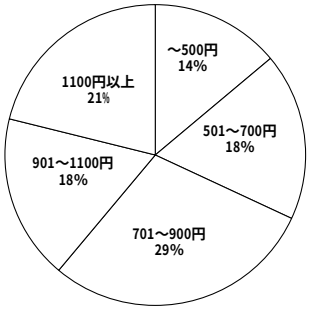
クロス集計表のデータ構造 (3)

夕食に使った金額



クロス集計表のデータ構造 (4)

バイトの時給



クロス集計表のデータ構造 (5)

このグラフで分かったことは，集計した

	夕食	バイト
選択肢1	15	14
選択肢2	23	18
選択肢3	17	29
選択肢4	23	18
選択肢5	22	21
合　計	100	100

（単位は人）

ということと変わりはない．この結果からは，･･･

クロス集計表のデータ構造 (6)

1. **夕食にお金をかけていない人がいた．**
2. 251**円以上かけている人は半数近くいる．**
3. 200**円前後の人は少ない．**
4. **バイトの時給は**900**円を越える人が**4**割いる．**
5. 700**円以下の人は**3**割いる．**
6. 800**円前後の人が割合は最多だ．**
7. ･･･．
8. ･･･．

⋮

クロス集計表のデータ構造 (7)

1. **円グラフを描いて，分布のようなものを確認した．**
2. **円グラフではなく棒グラフでもいいのだが，割合を示すとき円グラフはよく使われている．**
3. **先ほどの結論でほぼ言い尽くした．**
4. **他に考察できない．**

⋮

そんなことはないはずだ！！

見方を変えて，このように考える．

クロス集計表のデータ構造 (8)

1. バイト収入が多い人はいいもの食っているのだろうか.
2. いいもの食っている代償は, バイトによるものか.
3. バイトは何か欲しいものを購入するための手段？
4. この調査した集団では, バイト収入と食費にかける金額はどのようになっているのか.
5. これらのことを数値で示せないものか. 百分率以外に.

⋮

・・・と, 考えてみる.

クロス集計表のデータ構造 (9)

ここで, バイトの選択肢①～⑤までと, 夕食の①～⑤まではどのようになっているのか.

つまり, 高額な (低額な) バイト収入の人は食べるものが豊か (貧しい) のではないか.

括弧内は同順. バイト, 食べるものの交換可能.

つまり, 先週勉強した相関のようなことを考えてみよう.
しかし, 相関のようにこのデータでは平均はとれない.

クロス集計表のデータ構造 (10)

ひとまず, バイトで500円以下の人は, どの程度夕食にかけているのだろうか. 夕食に501円以上かけた人はどの程度収入があるだろうか. そこで, 以下のような表を考える.

		バイト					
		～500	501～700	701～900	901～1100	1100～	合計
夕食	0						
	1～150						
	151～250						
	251～500						
	501～						
	合計						100人

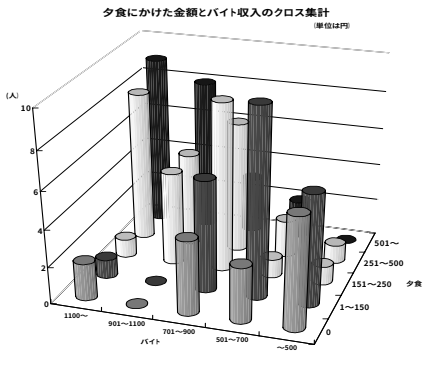
クロス集計表のデータ構造 (11)

このような表をクロス集計表 (あるいは分割表) といって, いくつかの属性変数の観察結果に用いられる. ここでは2次元で示したが, 3次, 4次など高次元のものも考えられる. 実際の数字は,

		バイト					
		～500	501～700	701～900	901～1100	1100～	合計
夕食	0	6	3	4	0	2	15
	1～150	6	10	6	0	1	23
	151～250	1	1	9	5	1	17
	251～500	1	2	7	5	8	23
	501～	0	2	3	8	9	22
	合計	14	18	29	18	21	100人

実測値と理論値の乖離 (1)

この表からどのようなことが分かるだろうか. 分かりにくければ以下のようなグラフにしてもいいだろう.



実測値と理論値の乖離 (2)

このグラフから何が言えるかを考えながら, 以下のことを論理的に見てみよう.

10円玉を2個を同時に100回投げて, 表, 裏の出方を見たところ, 以下のような実測値となった.

	表表	表裏あるいは裏表	裏裏
回数	20	58	22

2個同時に投げて,
表表の出る確率は, $\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}$. この確率は裏裏も同じ.

表裏あるいは裏表の出る確率は, $2\times\left(\frac{1}{2}\times\frac{1}{2}\right)=\frac{1}{2}$.

どうでもいいことだが，

表　　裏

である．

数字が入っていない5円硬貨だけは稲がついているほうが表であるという．

実測値と理論値の乖離（3）

100回投げたので，

表表，裏裏は $\frac{1}{4}\times 100 = 25$ くらい出るだろう．

表裏あるいは裏表は $\frac{1}{2}\times 100 = 50$ くらい出るだろう．

このように，階級に属する度数を理論度数という．

実測値と理論値の乖離（4）

実測値は，

	表表	表裏あるいは裏表	裏裏
回数	20	58	22

理論値は，

	表表	表裏あるいは裏表	裏裏
回数	25	50	25

このように見ると，表裏あるいは裏表が多いような，表表が少ないような気がする．

実測値と理論値の乖離（5）

このようなときに，実測値と理論値の乖離は，以下のように計算する．

$$\frac{(20-25)^2}{25}+\frac{(58-50)^2}{50}+\frac{(22-25)^2}{25}$$
$$=1+1.28+0.36=2.64$$

この2.64という実現値は，自由度3-1の $\chi 2$ 乗分布に従う．
ここで，基準になる有意水準0.05の $\chi 2$ 乗値を見てみれば，

$$\chi^2(0.05)=5.991>2.64$$

である．計算した値が小さいので，表裏の出方は半々であることが分かった．

実測値と理論値の乖離（6）

この考え方を食費−バイトデータに応用してみよう．
今，食費，バイトという属性は独立である（つまり何らの関係もない）と考える．
食費がどこかの選択肢に属する確率は，バイト収入には依存しないし，逆も同じである．

一般化すれば，

実測値と理論値の乖離（7）

		属　性 B				
		B_1	B_2		B_n	合計
属性 A	A_1	f_{11}	f_{12}	$\cdots$	f_{1n}	$f_{1\bullet}=\sum_{j=1}^{n} f_{1j}$
	A_2	f_{21}	f_{22}	$\cdots$	f_{2n}	$f_{2\bullet}$
		$\vdots$	$\vdots$	$\ddots$	$\vdots$	
	A_m	f_{m1}	f_{m2}	$\cdots$	f_{mn}	$f_{m\bullet}$
	合計	$f_{\bullet 1}=\sum_{i=1}^{m} f_{i1}$	$f_{\bullet 2}$	$\cdots$	$f_{\bullet n}$	$N=\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}$

において，

実測値と理論値の乖離 (8)

標本が (A_i, B_j) に属する確率は，

$$P(A_i, B_j) = P(A_i) \cdot P(B_j)$$

で求められる．しかし，$P(A_i), P(B_j)$ は先験的に不明であるから，標本からの推定値を，

$$P(A_i) = \frac{f_{i\bullet}}{N},\ P(B_j) = \frac{f_{\bullet j}}{N}$$

と考える．

実測値と理論値の乖離 (9)

そこで，実測値 f_{ij} に対して標本の (A_i, B_j) に属する理論度数 F_{ij} は，先ほどの10円玉の応用として，

$$F_{ij} = N \cdot P(A_i) \cdot P(B_j) = \frac{1}{N} f_{i\bullet} f_{\bullet j}$$

で求められる．実測値と理論値の乖離は，統計量

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(f_{ij} - F_{ij})^2}{F_{ij}}$$

は，自由度 $(m-1)(n-1)$ の χ 2乗分布に従う．

実測値と理論値の乖離 (10)

実際に夕食-バイトデータを検定してみる．理論値を求めよう．

		バイト					
		～500	501～700	701～900	901～1100	1100～	合計
夕食	0	$\frac{15\times14}{100} = 2.1$	2.7	4.35	2.7	3.15	15
	1～150	3.68	4.14	6.67	4.14	4.83	23
	151～250	2.38	3.06	$\frac{17\times29}{100} = 4.93$	3.06	3.57	17
	251～500	3.22	4.14	6.67	4.14	4.83	23
	501～	3.08	3.96	6.38	3.96	4.62	22
	合計	14	18	29	18	21	100人

実測値と理論値の乖離 (11)

実測値と理論値の乖離の検定は，

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(f_{ij} - F_{ij})^2}{F_{ij}}$$

$$= \frac{(6-2.1)^2}{2.1} + \cdots + \frac{(9-4.62)^2}{4.62} = 50.017$$

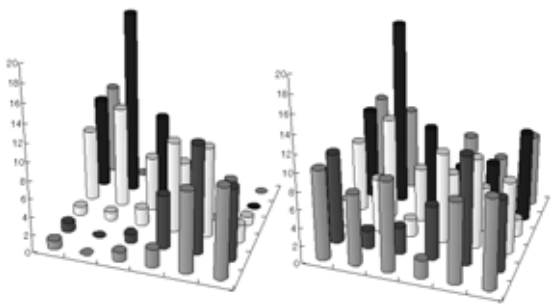
は，自由度 $(5-1)(5-1)$ の χ 2乗分布に従う．

$$\chi^2_{16}(0.05) = 26.30 < 50.017$$

よって，食費とバイト収入は無関係ではないようだ．

実測値と理論値の乖離 (12)

そこで，左のようなときは大きく，右のようなときは小さくなるような値はないものか．



連関係数 (1)

ある．左のグラフのように対角要素近傍にデータが集中した場合高く，右のように対角要素近傍以外にもデータが多く存在する場合低くなるような値をクラメールの連関係数 C_1 と呼び，以下のように計算する．

$$C_1 = \sqrt{\frac{\chi^2}{N \cdot \min(m,n) - 1}}$$

$\min(m,n)$ は行と列の階級数の小さいほうである．

連関係数(2)

ある．左のグラフのように対角要素近傍にデータが集中した場合高く，右のように対角要素近傍以外にもデータが多く存在する場合低くなるような値をクラメールの連関係数C_1と呼び，以下のように計算する．

$$C_1 = \sqrt{\frac{\chi^2}{N \cdot \min(m,n) - 1}}$$

$\min(m,n)$は行と列の階級数の小さいほうである．

連関係数(3)

実際に夕食-バイト収入データのC_1は，

$$C_1 = \sqrt{\frac{\chi^2}{N \cdot \min(m,n) - 1}}$$

$$= \sqrt{\frac{50.017}{100 \times 4}} = 0.3536$$

連関係数(4)

実際に他のデータで計算してみれば，

S大学医学部看護学科でPlayTypeの練習をし，45人から以下のような結果を得た．検定を行ってみよう．

		エラー							
		0	1-2	3-4	5-6	7-8	9-10	11以上	合計
スコア	40以下	2	2	0	0	0	0	1	5
	41-60	3	6	1	0	0	0	0	10
	61-80	5	3	1	3	0	0	1	13
	81-100	0	5	3	1	1	1	1	12
	101-120	1	0	1	0	0	0	0	2
	121以上	0	0	1	0	0	1	0	2
	合計	11	16	7	4	1	2	3	44

連関係数(5)

以下のようである．理論度数は，

		エラー							
		0	1-2	3-4	5-6	7-8	9-10	11以上	合計
スコア	40以下	$\frac{11\times5}{44}=1.25$	1.81	0.80	0.45	0.11	0.23	0.34	5
	41-60	$\frac{11\times10}{44}=2.5$	3.64	1.59	0.91	0.23	0.45	0.68	10
	61-80	3.25	4.23	2.07	1.18	0.30	0.59	0.88	13
	81-100	3	4.36	1.90	1.09	0.27	0.55	0.82	12
	101-120	0.5	0.73	0.32	0.18	0.05	0.09	0.14	2
	121以上	0.5	0.73	0.32	0.18	0.05	0.09	0.14	2
	合計	11	16	7	4	1	2	3	44

連関係数(6)

以下のようである．従って，

$$\chi^2 = \frac{(2-1.25)^2}{1.25} + \cdots + \frac{(0-0.14)^2}{0.14} = 34.65$$

自由度は(6-1)(7-1)=30

$$\chi^2_{30}(0.05) = 43.77 > 34.65$$

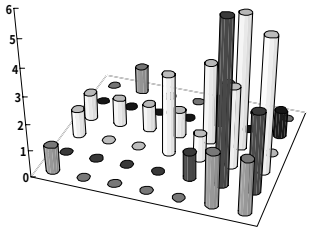
よって，スコアが低い人ほどスコアに対するエラーは多くありそうだったが，そんなことはなかった，ということが数字で示すことができる．

連関係数(7)

そのときの連関係数は，

$$C_1 = \sqrt{\frac{\chi^2}{N \cdot \min(m,n) - 1}} = \sqrt{\frac{34.65}{44 \times 5}} = 0.397$$

グラフにすればこんな感じで，対角要素近傍に集約されていないことがわかる．



これで何がわかるか(1)

このようなクロス集計は，

1. 平均や分散，標準偏差が計算できないデータでも使える．
2. よく考えてみれば，カテゴリに分けた集計票の相関のようなものを計算している．
3. しかし，2つの質問は互いに関連があると思われる質問間で行ってはいけない．
4. 平均や分散，標準偏差が計算できるデータでも一定のカテゴリに分けることで，分析できる．

これで何がわかるか(2)

このようなクロス集計は，

5. 質問紙調査のときに項目間の関係を見たい場合に使えるだろう．
6. それによって隠れた構造がわかるかもしれない．
7. 集計は円グラフでも可能だが，しかし，それは表で示したことと本質的には変わらない．
8. 今後看護の道に進まれるあなたたちには，科学的にデータを見て，更に解析してほしいと思います．今回行った平均，標準偏差，標準化，回帰直線，相関係数，順位相関，クラメールの連関係数などはあくまでプロローグと思ってください．

終わりに

今後看護の道に進まれるあなたたちには，科学的にデータを見て，更に解析してほしいと思います．今回行った平均，標準偏差，標準化，回帰直線，相関係数，順位相関，クラメールの連関係数などはあくまでデータを解析するためのプロローグです．

この後も統計解析というものを忘れないでください．