知能システム演習A

2013.10.24（木）

作成：B4　福田至

分散の基本事項

- 分散はデータが平均値からどれぐらい散らばっているかを示す尺度
- 確率変数Xの平均をμとする．つまり
- $E(X) = \mu$ の時、
  - $\sigma^2 = \mathrm{V}(X) = E[(X-\mu)^2]$ ←分散
  - $\sigma = \sqrt{\mathrm{V}(X)}$ ←標準偏差
- $\mathrm{V}(X) = E(X^2 - 2\mu X + \mu^2)$ 展開
  - $= E(X^2) - 2\mu E(X) + \mu^2$
  - $= E(X^2) - \mu^2$
  - $= E(X^2) - [E(X)]^2$

平均μについて

- 平均μについて:
- Xの確率分布がP(X=x_i) = f(x_i) (I = 1, 2,…)であるとき、
- $E(X)$ = μ = x_1・f(x_1) + x_2・f(x_2) +……+ x_i・f(x_i) +…

  =$\sum_i x_i$・f(x_i)

X＼Y	1	2	計
1	1/4	1/4	1/2
2	1/4	1/4	1/2
計	1/2	1/2	1

共分散とは何か（covariance）

- 2つのデータが，どれだけ関連性・連動性があるかを示す係数.
  - ２つのデータの偏差の積 $(x_i-\bar{x})\,(y_i-\bar{y})$ の
  - 平均 によって求められる
  - $\boldsymbol{C_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}$
  - $= E[(X - E(X))(\mathrm{Y} - \mathrm{E}(\mathrm{Y}))]$
  - $= E[X\,\mathrm{Y} - X\,E(\mathrm{Y}) - \mathrm{Y}\,E(X) + E(X)\,E(\mathrm{Y})]$
    $= E(X\,\mathrm{Y}) - E(X)\,E(\mathrm{Y}) - E(\mathrm{Y})\,E(X) + E[E(X)\,E(\mathrm{Y})]$
    $= \boldsymbol{E(X\,Y) - E(X)\,E(Y)}$

– ex) 数学と国語の点数

	数学	国語
平均点	50 $(\bar{x})$	50 $(\bar{y})$
佐々木くんの点数	80 (x_i)	40 (y_i)
偏差（平均との差）	30 $(x_i - \bar{x})$	-10 $(y_i - \bar{y})$

– 佐々木くんの偏差の積：30＊(-10) ＝ -300

– これを生徒全員について求め、偏差の積の合計を平均にしたものが、数学と国語の共分散になる

例題）

A組の成績

番号	国語	算数
1	25	25
2	30	35
3	40	45
4	50	50
5	51	55
6	60	65
7	70	65
8	80	85

平均	50.8	53.1
分散	315.2	312.1
共分散	307.7	
相関	0.98	

B組の成績

番号	国語	算数
1	25	70
2	30	35
3	40	70
4	50	20
5	51	60
6	60	90
7	70	80
8	80	30

平均	50.8	56.9
分散	315.2	568.4
共分散	-3.9	
相関	-0.009	

例題 4.3 ——(共分散・相関係数)

X, Y の同時分布表が右のように与えられている.

(1) 分散 $V(X)$, $V(Y)$ を求めよ.

(2) 共分散 $C(X, Y)$ を求めよ.

(3) 相関係数 $\rho(X, Y)$ を求めよ.

X \ Y	0	1	2	計
0	0	0.1	0.2	0.3
1	0.1	0.2	0.1	0.4
2	0.2	0.1	0	0.3
計	0.3	0.4	0.3	1

【解答】

(1) $E(X)=0\times0.3+1\times0.4+2\times0.3=1=E(Y)$

$E(X^2)=0^2\times0.3+1^2\times0.4+2^2\times0.3=1.6=E(Y^2)$

これより $V(X)=E(X^2)-\{E(X)\}^2=1.6-1=0.6=V(Y)$

(2) $E(XY)=1\times1\times0.2+1\times2\times0.1+2\times1\times0.1=0.6$

したがって, $C(X, Y)=E(XY)-E(X)E(Y)=0.6-1\times1=-0.4$

(3) $\rho(X, Y)=\dfrac{C(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}}=\dfrac{-0.4}{\sqrt{0.6}\sqrt{0.6}}=-\dfrac{2}{3}=-0.6666\cdots$

- 共分散がプラス⇔片方が増えるともう片方も増える傾向がある（正の相関がある）
- 共分散がマイナス⇔片方が増えるともう片方は減る傾向がある（負の相関がある）
- 共分散がゼロ⇔相関性なし．このとき、2つのデータは独立である[$E(XY) = E(X)E(Y)$]

- 更に、共分散を標準偏差の積で割った値は、2つのデータの相関関数という

- 相関係数とは、2 つの確率変数の相関（類似性の度合い）を示す統計学的指標である。

- 求め方：共分散を標準偏差で割る
  - $\rho(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$

＊相関係数を使う理由＊

- 相関係数は，基準化変量をもとに作られるので，身長をcmで表わされていても、メートルでも，インチでも，計算結果は同じ値をとる.
- 測定単位に依存しない，関連の強さを測る指標となるから.

相関係数の性質

- 最大1，最小−1の値をとる.
- 相関係数の絶対値が1に近い程，相関が強く，-1に近ければ負の相関が強い.
- 相関係数の絶対値が1になるのは，データ点が一直線上に位置するときのみである.
- 相関係数は，直線的な関係の強さをはかるもので，曲線的な関係を調べるのには向いていない.

相関の目安

	正の相関	負の相関
高い相関	$r \geq 0.7$	$r \leq -0.7$
かなりの相関	$0.4 \leq r \leq 0.7$	$-0.4 \geq r \geq -0.7$
低い相関	$0.2 \leq r \leq 0.4$	$-0.2 \geq r \geq -0.4$
ほとんど相関がない	$0 \leq r \leq 0.2$	$-0 \geq r \geq -0.2$

参考資料

- 相関関数:
  - http://ja.wikipedia.org/wiki/%E7%9B%B8%E9%96%A2%E4%BF%82%E6%95%B0
- 共分散:
  - http://ja.wikipedia.org/wiki/%E5%85%B1%E5%88%86%E6%95%A3
- 情報処理・2変数の関係:
  - http://www2.kobe-u.ac.jp/~koba0724/jyohosyori/document/03Corr.pdf
- 共分散と相関関数・2変数の関連を調べる:
  - http://www.ec.kansai-u.ac.jp/user/amatsuo/StatProb2011Spr/CovCorr.pdf
- 共分散と相関関数:
- http://mcn-www.jwu.ac.jp/~yokamoto/ccc/stat/p22covcor/