データ解析基礎
4. 正規分布と相関係数

keyword
- 正規分布
  - 正規分布の性質
  - 偏差値
- 変数間の関係を表す統計量
  - 共分散
  - 相関係数
  - 散布図

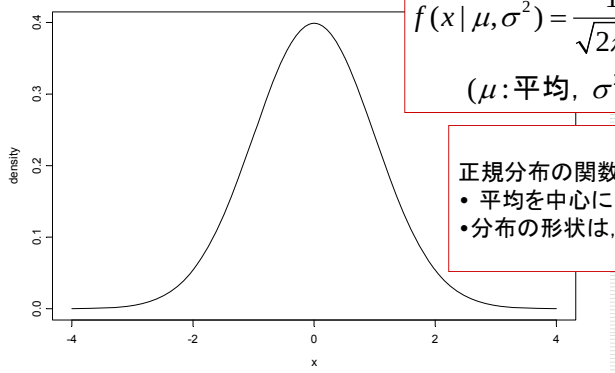
正規分布

- 世の中の多くの現象は，標本数を大きくしていくと，正規分布に近づいていくことが知られている．
- 正規分布
  - データ解析の基礎となる重要な分布
  - 平均と分散によって特徴づけることができる．
    - 平均値：分布の中心を表す値
    - 分散：分布のばらつきを表す値

正規分布

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

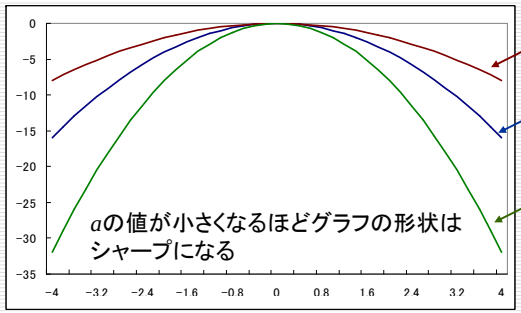
(μ：平均，σ^2：分散)



正規分布の関数（密度関数）の特徴
- 平均を中心にし，左右対称である．
- 分布の形状は，分散によって変化する

正規分布の形状：2次関数の例

$$\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \Rightarrow y = -\frac{1}{a}(x-b)^2 \quad (2\sigma^2 = a,\ \mu = b)$$



$y = -\frac{x^2}{2}\ (a=2)$

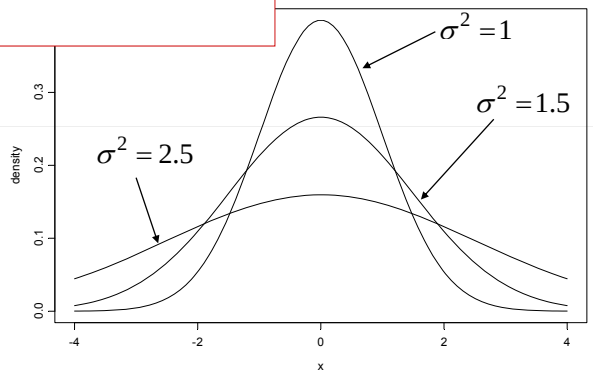
$y = -x^2\ (a=1)$

$y = -2x^2\ (a=\frac{1}{2})$

aの値が小さくなるほどグラフの形状はシャープになる

正規分布の形状

σ^2の値が小さくなるほど，分布の形状はシャープになる



標準正規分布

平均μが分散σ^2である正規分布　　(* $\exp[x]=e^x$)

$$f(x\mid\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$ (μ:平均，σ^2:分散)

について線形変換

$$z=\frac{x-\mu}{\sigma}$$ ← 標準化

をおこなうと，平均が0，分散が1の正規分布となり，

$$f(z\mid 0,1)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{z^2}{2}\right]$$

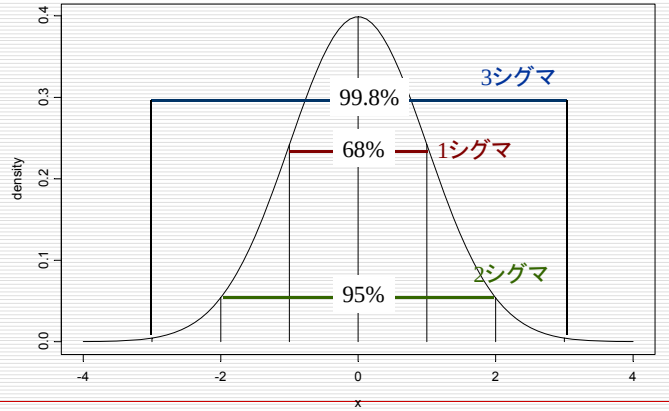
と書くことができる．この正規分布を標準正規分布という．

正規分布と確率
1シグマ，2シグマ，3シグマの法則

- 観測データが正規分布に従う場合，以下ようなの概算を見積もることができる．

$\mu\pm1\times\sigma$の範囲内　⇒
　1シグマ：データ全体の約68%（約2/3）が含まれる．
$\mu\pm2\times\sigma$の範囲内　⇒
　2シグマ：データ全体の約95%（約19/20）が含まれる．
$\mu\pm3\times\sigma$の範囲内　⇒
　3シグマ：データ全体の約99.7%が含まれる．

標準正規分布N(0,1)の密度関数



正規分布と偏差値

- 偏差値の定義
  - 受験者全員の平均点に相当する得点を50に変換し，標準偏差の1倍だけの隔たりを10に換算するような換算法によって算出される指標
- 偏差値 z の算出式

$$z = 50 + 10 \times \frac{x - \bar{x}}{\sigma} \quad (\bar{x}:\ \text{平均点}, \sigma:\ \text{標準偏差})$$

正規分布と偏差値

偏差値	z	確率	順位（100人中）
70	2	97.7%	2
65	1.5	93.3%	7
60	1	84.1%	16
55	0.5	69.1%	31
50	0	50.0%	50
45	-0.5	30.9%	69
40	-1	15.9%	84
35	-1.5	6.7%	93

偏差値65・・・
上位7%
7位（100人）

偏差値50・・・
上位50%
50位（100人）

2変数間の関係を表す統計量

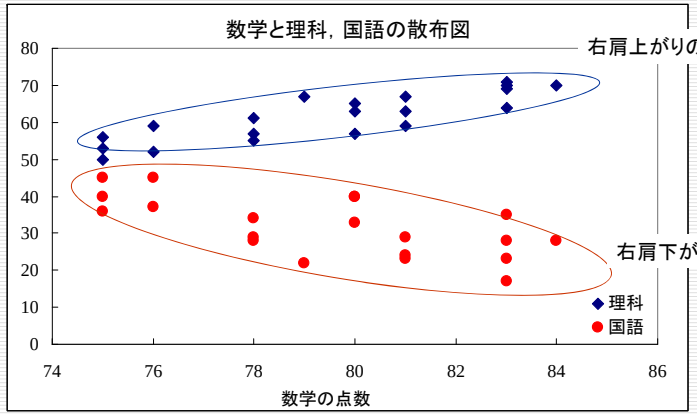
- 共分散
- 相関係数
  - 相関係数の意味
  - 相関係数の定義
- 散布図
  - 視覚的に変量間の関係を見る

相関とは：例題・・・・・

- 3科目（数学，理科，国語）について，試験をしたところ，次の結果であった．このデータから，3教科について，何らかの関係があるか．

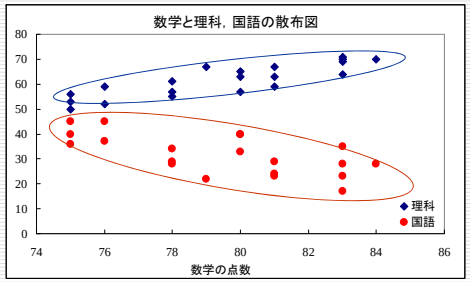
	数学	理科	国語
1	81	59	23
2	83	70	23
3	81	63	24
4	78	57	34
5	80	63	40
6	84	70	28
7	78	55	29
8	76	59	45
9	78	61	28
10	75	53	40
11	79	67	22
12	80	57	40
13	80	65	33
14	75	56	45
15	75	50	36
16	83	64	35
17	83	71	28
18	83	69	17
19	81	67	29
20	76	52	37

相関とは：例題・・・・・

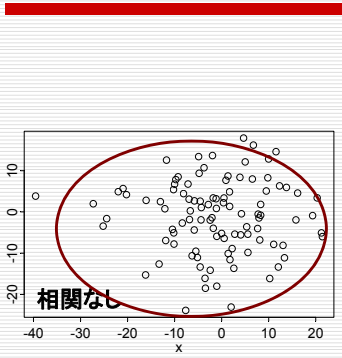


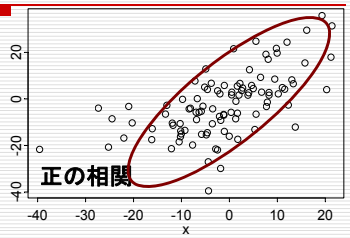
散布図

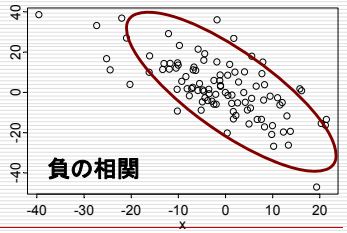
- 2つの変量を，x軸とy軸に割り当て，観測データを座標上の点で表した図を散布図という．
- 2変量間の関係を，視覚的に見ることが出来る．



散布図からみる相関関係







2 変数間の関係を表す量：
相関係数

- 2 変量間(x, y)の関係を測る指標・・・相関係数
- 相関係数 $r(x,y)$ の値：
  - 相関係数の値の範囲：$-1 \leqq r(x,y) \leqq 1$
  - 1 に近いほど正の相関が強い
  - -1 に近いほど負の相関が強い
  - 0 の時，相関がない
- 相関係数は常に因果関係を示すものではない．

2 変数間の関係を表す量

相関係数と関係の強さ

$0 \le |r(x,y)| \le 0.2$ ⇒ほとんど相関がない

$0.2 < |r(x,y)| \le 0.4$ ⇒弱い相関がある

$0.4 < |r(x,y)| \le 0.7$ ⇒比較的強い相関あり

$0.7 < |r(x,y)| \le 1.0$ ⇒強い相関がある

相関係数

- 2 つの変量を(x, y)で表した時，相関係数は以下の式で定義される．
- 共分散の値を，-1から1の範囲内に標準化した数と考えることもできる．

相関係数の定義式

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

共分散：2変量間の関係を表現する量

第i番目の観測値を(x_i, y_i)で表したとき，

$x_i - \bar{x}$: x方向への偏差(deviation)

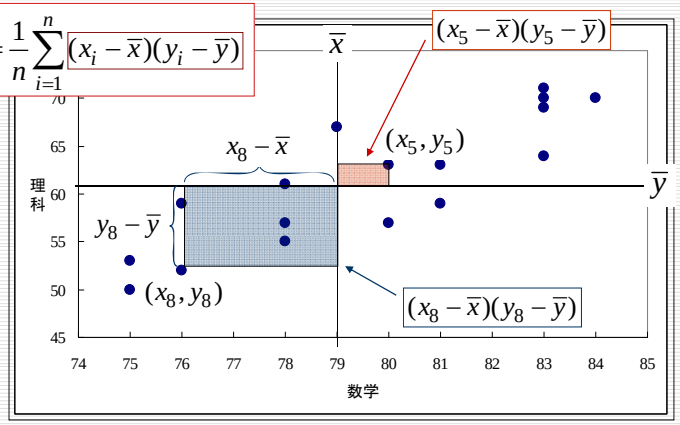
$y_i - \bar{y}$: y方向への偏差(deviation)

という．2つの偏差の積をすべて足して，標本サイズで割ったものを共分散という．

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

共分散の幾何学的意味

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



共分散と分散

分散と共分散

・分散：x の偏差 $(x_i - \bar{x})$ の2乗（xの場合）

・共分散：x の偏差 $(x_i - \bar{x})$ と y 方向への偏差 $(y_i - \bar{y})$ の積

共分散と分散の関係式

$$Var(x) = Cov(x,x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})$$

共分散の大きさを評価することが難しい．

共分散200は大きい？

分散と共分散の値の範囲

$$0 \le Var(x), Var(y) \le \infty, \quad -\infty \le Cov(x,y) \le \infty$$

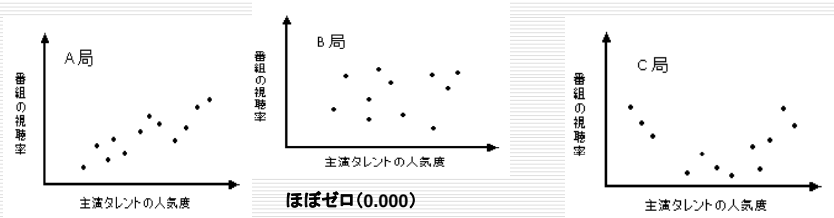
タレントの人気と視聴率の関係は

渡辺久哲「調査データにだまされない法」創元社より

- ある番組分析班が，番組の主演タレントの人気とその番組の視聴率の関係を検討した．
- 分析対象は，A局・B局・C局のある時間帯の番組

タレントの人気と視聴率の関係は

- 3テレビ局について、番組の視聴率とそこに起用したタレントの人気度について相関係数を算出．
- 各局ごとにタレントの人気度（ヨコ軸）と番組の視聴率（タテ軸）でプロットを作成。



ほぼゼロ(0.000)

相関係数　大きい(0.700)　　ほぼゼロ(0.000)

タレントの人気と視聴率の関係は

- タレントの起用は番組の成功を大きく左右する要素であるが，相関係数を見たところB・C局のデータからは相関関係は見られなかった．
- 結論
  「A局のみが、起用したタレントの人気が高いほど視聴率が高く、起用したタレントの人気が低いほど視聴率が低いという傾向が見られる」
- 本当にこの結論でよいのだろうか？

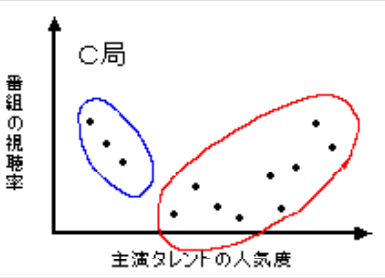
タレントの人気と視聴率の関係は

- A局
  - 相関関係が見られる
- B局
  - 起用しているタレントの人気度と視聴率はほぼ無関係（相関係数もゼロに近い数値）
- C局
  - 相関係数はほぼゼロに近い数値ではあるが，プロットは一風変わってU字型になっている

➡ C局についてはプロットを見ると，相関がないと断言することは出来ない・・・

タレントの人気と視聴率の関係は

- 右半分の群からは，A局と同じタレントの人気度が高いほど番組の視聴率が高いという傾向が読み取れる
- 左半分からは、その逆で人気度の低いタレントでも高い視聴率をとる番組があることが読み取れる



相関係数を見ただけでは分からないことが，散布図から分かることがある

タレントの人気と視聴率の関係は

- なぜ、相関係数をみるだけではわからなかったのか？

⬇

- C局には，i) タレントの人気に依存した番組と ii)依存していない番組の2種類があるために，全体としてはU字型のプロット図になっている．
- 相関係数では，図にしたときの曲線的な関係の大きさをとらえることができない．

例題1：相関係数と散布図

吉田寿夫「本当にわかりやすいすごく大切なことが書いてあるごく初歩の統計の本」北大路書房より

- 以下のデータは，ある女性が12人の男性の積極性と清潔さについて評価したデータとそれぞれの男性に対する好意度に関するデータをまとめたものです．積極性についての評価と好意度および清潔さについての評価と好意度に関して，それぞれの相関係数と散布図を作成し，わかることを述べなさい．

No	1	2	3	4	5	6	7	8	9	10	11	12
積極性	1	6	4	2	4	3	5	4	7	5	2	5
清潔さ	1	3	4	7	6	2	6	6	5	3	6	5
好意度	2	5	6	1	5	4	3	4	7	4	3	4

例題1:回答項目

積極性についての評価

非常に積極的	7
わりと積極的	6
やや積極的	5
どちらともいえない	4
やや消極的	3
わりと消極的	2
非常に消極的	1

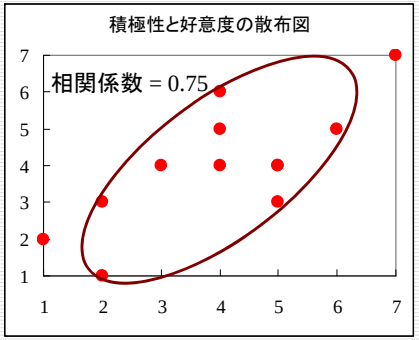
清潔さについての評価

非常に清潔	7
わりと清潔	6
やや清潔	5
どちらともいえない	4
やや不潔	3
わりと不潔	2
非常に不潔	1

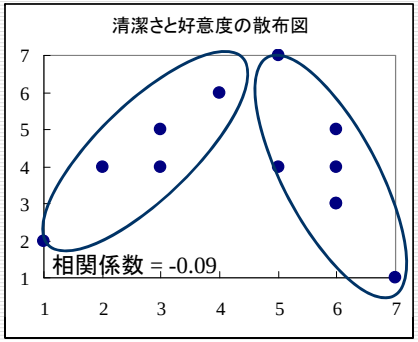
好意度

非常に好き	7
わりと好き	6
やや好き	5
どちらともいえない	4
やや嫌い	3
わりと嫌い	2
非常に嫌い	1

例題1:散布図



積極だと思う男性をより好むという傾向



非常に不潔と思う男性を好まないと同時に，あまりにも清潔な男性もまた好まない

例題2:相関係数と散布図

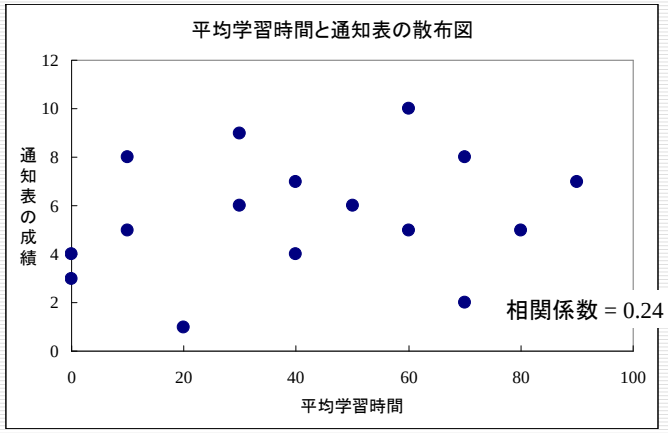
吉田寿夫「本当にわかりやすいすごく大切なことが書いてあるごく初歩の統計の本」北大路書房より

- 中学生の勉強に対する努力量と学業成績の関係の検討する.
- 16人中学2年生について，以下の項目についてデータ得られているとしたとき，平均学習時間と成績の関係について分析をおこなう.
  - 家庭での英語の学習時間
    (1日あたりの平均時間:分)
  - 英語の通知表の成績(10段階評定)
  - 各生徒の知能の高さ(高，低)

例題2:観測データ

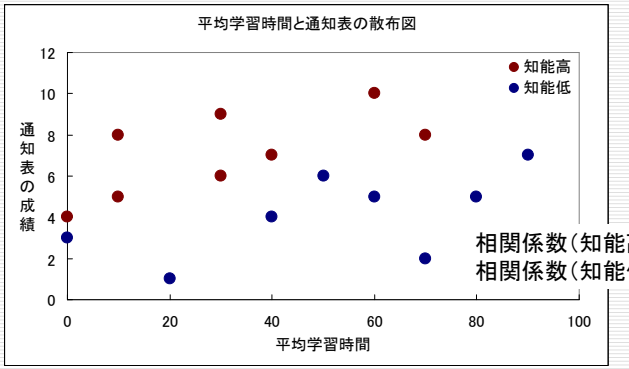
No.	平均学習時間	通知表の成績	知能の高さ
1	20	1	低
2	50	6	低
3	70	8	高
4	80	5	低
5	40	7	高
6	0	3	低
7	90	7	低
8	60	10	高
9	10	5	高
10	30	6	高
11	30	9	高
12	40	4	低
13	0	4	高
14	60	5	低
15	10	8	高
16	70	2	低

例題2：相関係数と散布図（全体）



例題2：相関係数と散布図（層別）

知能の高さがほぼ一定であれば，“英語に関して努力している生徒ほど成績が良い”という正の相関関係が認められる．



例題3：相関係数と散布図

- 以下の表は，売上本数，広告費，キャンペーンの実施について調べたものである．

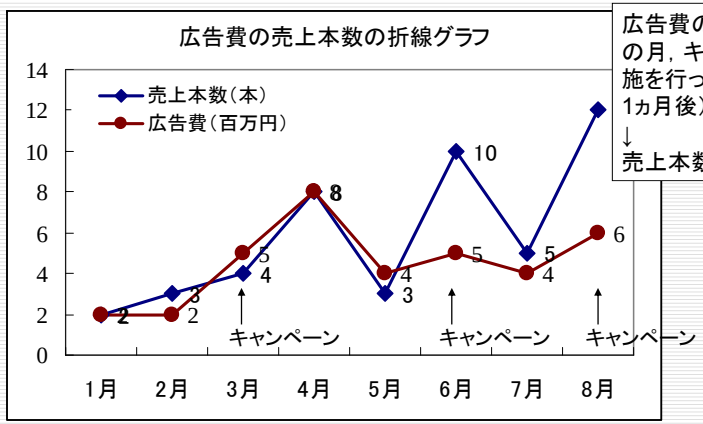
	売上本数（本）	広告費（百万円）	キャンペーンの実施
1月	2	2	無
2月	3	2	無
3月	4	5	有
4月	8	8	無
5月	3	4	無
6月	10	5	有
7月	5	4	無
8月	12	6	有

例題3：相関係数と散布図

- 広告費やキャンペーンの実施が売上に影響を及ぼしているかを調べたい．

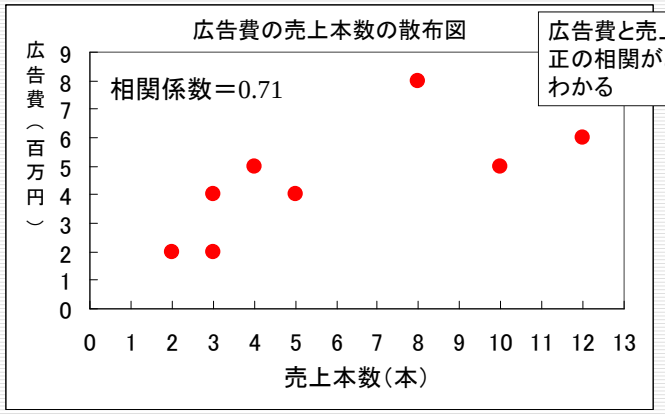
- 相関係数と散布図を活用した解析をおこなう
  - 広告費と売上本数の折線グラフの描画
  - 広告費と売上本数の散布図の描画
  - キャンペーン実施と売上本数の散布図の描画

例題3：折線グラフ



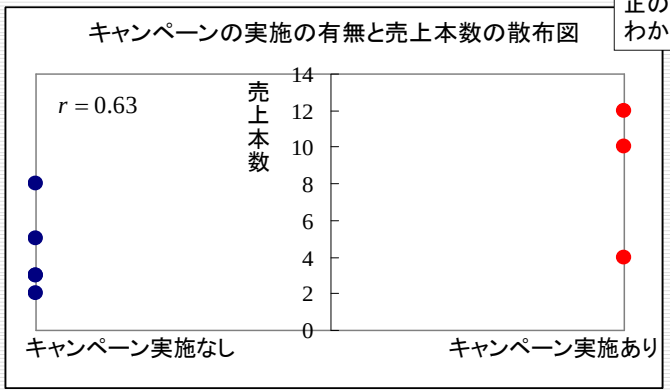
広告費の投入度が大の月，キャンペーン実施を行った月（あるいは1ヵ月後）
↓
売上本数が多くなる

例題3：広告費と売上本数の散布図



広告費と売上本数には正の相関があることがわかる

例3：キャンペーン実施と売上本数の散布図



キャンペーンの実施と売上本数には正の相関があることがわかる

例3：平均値による比較

- キャンペーン実施（あり，なし）別で，売上本数の平均値を計算
- キャンペーン実施あり：12本，10本，4本
  - ➔平均値8.7本
- キャンペーン実施なし：2本，3本，8本，3本，5本
  - ➔平均値4.2本
- 平均値を比較より，キャンペーン実施の効果がうかがえる

例3:相関係数と散布図の活用例

- 広告費やキャンペーンの実施が売上に影響を及ぼしているかを調べる.
- 相関係数と散布図を活用した解析をおこなう
  - 広告費と売上本数の折線グラフの描画
  - 広告費と売上本数の散布図の描画
  - キャンペーン実施と売上本数の散布図の描画
- 広告費とキャンペーン実施は売上に変動を与える要因である！！

まとめ

- 正規分布
  - データ解析の基礎となる重要な分布
  - 平均と分散によって特徴づけることができる.
- 相関係数 $r(x,y)$
  - 2 変量間(x, y)の関係を測る指標
  - $-1 \leqq r(x,y) \leqq 1$
- 散布図
  - 2つの変量を，x軸とy軸に割り当て，観測データを座標上の点で表した図を散布図という.
  - 2変量間の関係を，視覚的に見ることが出来る