基礎・経済統計　4

統計の初歩

0. 今日の目標

- 様々な代表値を理解する
- データの分布，散らばりを把握するためのデータ整理法を理解する
  - 度数分布，ヒストグラム

1.　データの中心の代表値
1．1　平均

- 平均の例
  - 四月の統一地方選で当選した府議（百十二人）と大阪市議（八十九人）の資産が七日、公開された。一九九六年施行の資産公開条例に基づき、土地・建物の面積や課税標準額、預貯金などを報告した。有価証券を除いた土地・建物の課税標準額と預貯金、金銭信託を合計した**平均**額は府議が二千八百五十二万円、市議が四千百九十万円。一億円を超えた府議は九人、市議は八人いた。（2003年10月8日 読売大阪版）

1．2　中央値

- 中央値の例
  - ［東京，2003年9月2日］**主要エコノミスト、欧州中銀の年内再利下げを予想～ロイター調査、次回会合では据え置き観測が大半～**
    - ロイターが主要エコノミスト62人に実施した調査によると、欧州中央銀行 (ECB: European Central Bank) が今年後半に再度の金利引き下げを実施するとの予想が半数以上に達しました。しかし、経済の先行きに楽観的な見方が出ているため、大幅な引き下げの可能性 は縮小しています。．．．．一方、調査に答えたエコノミストのうち25人は、世界の株価が上昇し、米国経済の回復を示すデータが増えていることを踏まえ、金利がすでに2%で底を打ったと予測しています。2003年末時点の金利水準については1．0%から2．0%まで予測の幅があり、**中央値**は1．75%でした。2004年末に関しては0．75%から3．25%に分布しており、**中央値**は2．25%となっています。
- 中央値
  - データを大きさ順に並べて真ん中をとる．しかし，データ数が偶数の場合は，真ん中のデータ2つの平均．
  - 中位数，メディアンとも呼ばれる．
- なぜ例では平均ではなく中央値をとったのか？
  - 平均の場合，極端な予測のエコノミストの影響を受けてしまう．しかし，中央値の場合，極端な予測には影響を受けない．

1．3．最頻値

- 最頻値の例
  - PC Watch読者環境調査2003年7月
    - 今回の調査では、CPUクロックの上昇がめざましく、初めて「～2.4GHz」が最頻値となった。また、HDD、メモリの大容量化傾向も続いている。周辺機器では、マルチフォーマット対応の書き込み/書き換えDVDドライブと、17インチLCDディスプレイが大幅に伸びている。接続環境はADSL/CATVの普及が一服したが、FTTHは引き続きポイントを伸ばしている。
    - **【今回の最頻値】**
      - 所有台数 2台、自作のデスクトップPC/AT互換機
      - OS：Windows XP Professional
      - CPU：Pentium 4
      - クロック：～2.4GHz
      - メモリ：～512MB
      - HDD：～200GB
      - ディスプレイ：17インチCRT、解像度～1,280×1,024ドット
      - インターネット接続：xDSL(フレッツADSL除く)
      - 家庭内LAN：構築している

    （http://pc.watch.impress.co.jp/docs/2003/0806/enq23.htm）
- 最頻値
  - データのなかでもっともよく現れた値．
  - 極端な値の影響を受けづらい
  - 数値化されていないデータでも得ることができる．

1．4　平均，中央値，最頻値

- 平均
  - 極端な値（はずれ値，異常値）の影響大
  - 計算が容易
  - 分散との関わり
- 中央値
  - はずれ値（異常値）の影響を受けない
  - 一度順序で並べなければいけない．多数のデータの場合大変
  - 平均偏差との関わり
- 最頻値
  - はずれ値（異常値）の影響を受けない
  - 最頻データ以外を完全に無視している
- 3つの関係
  - ヒストグラムの山が1つのとき，平均－最頻値≒3×（平均－中央値）＜ピアソンの式＞がほとんどの場合成り立つ．
  - 中央値≒(2 平均＋最頻値)/3，つまり，中央値は平均と最頻値の間を1：2に内分する．

2．散らばりの代表値
2．1　分散，標準偏差，範囲

- 標本分散（標本が調査対象の一部の場合）
  $\frac{(\text{データ}-\text{平均})^2\text{の合計}}{\text{標本数}-1}$
- 分散（標本が調査対象と一致する場合）
  $\frac{(\text{データ}-\text{平均})^2\text{の合計}}{\text{標本数}}$
- （標本）標準偏差　$\sqrt{(\text{標本})\text{分散}}$
- 分散，標準偏差になると経済ニュースにさえも出てこない．これでいいんだろうか？
- 範囲＝最大値ー最小値

2．2．チェビシェフの不等式

- 標準偏差からデータの分布がある程度分かる
- チェビシェフの不等式（どんなデータにも当てはまる）
  - $|\text{データ値}-\text{標本平均}| \geq k \times \text{標本標準偏差}$ を満たすデータの割合は全標本数の $1/k^2$ 以下である．
  - 例えば標本平均から標準偏差の2倍以上離れているデータは全体の1/4以下である．
    - 偏差値70以上は絶対1/4以下しかいない

2．3．その他散らばりの代表値

- 四分位点（しぶんいてん）
  - 第1四分位点
    - 順序どおりに並べ替え，データを4分割する．その4分割点のうち最も小さい値．もし，その分割点にデータがなければ，補間近似する．具体的には$(n+1)/4$ の整数部番目のデータとその次のデータの半分，または，$(n+1)/4$ の小数部を重みにした加重平均
  - 第2四分位点　中央値と同じ
  - 第3四分位点
    - 順序どおりに並べ替え，データを4分割する．その4分割点のうち最も大きい値．もし，その分割点にデータがなければ，補間近似
  - 四分位範囲　第3四分位点－第1四分位点
- 10分位点　データを10分割
- パーセント点（百分位点）　データを百分割

2．4　例の解説

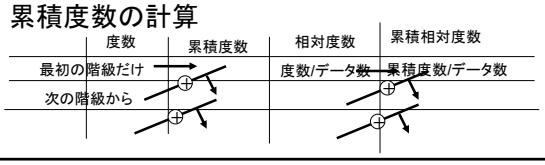
- 例1．1
  - 教科書参照
- 例1．2
  - 中間階層はどこの地域でも似たり寄ったりの収入である．
- 例1．3
  - ピアソンの式から各地域の貯蓄額の最頻値の近似値を求めてみよう．
  - 例　北海道　$\text{最頻値} \cong \text{平均} - 3(\text{平均} - \text{中央値})$
    $= 1007 - 3(1007 - 698) = 80\text{万円}$

3．度数分布表とヒストグラム
3．1　作成法

- 作成法は教科書参照
  - ヒストグラムの高さは度数を表す．
- 階級数の決め方
  - スタージェスの公式　$1 + 3.3\log_{10} n$
    nは標本数（データ数）
    - 例
      データ数が100，範囲が3～57までのデータがあるとする．
      スタージェスの公式から階級数は $3.3\log_{10} 100 + 1 = 3.3 \times 2 + 1 = 7.6$
      階級幅＝範囲/階級数＝7．1，従って，階級幅は5か10．
      幅10を選ぶと階級の下限は，0,10，20，30，40，50で6個の階級
      幅5を選ぶと階級の下限は，0,5,10,･･･，55の12個
- 図1．7の解説
  - 平均ー最頻値≒3×（平均ー中央値）＜ピアソンの式＞が成立していない．このことは，貯蓄の分布のヒストグラムの山が1つではないなど，かなり変形された分布であることを示している．

度数分布表の作り方

- 階級分割　p．13参照
  - 階級値は階級の中点
- データのカウント＝度数の計算
  - 各階級に属するデータ数をカウントする
- 度数分布の記入
- 累積度数の計算



3.2 度数分布表による代表値の計算

- 考え方
  - 階級値のところにデータが度数個集まっている
- 平均

$$\frac{(\text{階級値}\times\text{度数})\text{の合計}}{\text{標本数}}=(\text{階級値}\times\text{相対度数})\text{の合計}$$

  - ただし，階級内平均が与えられている場合は階級値の代わりにそれを使う
- 標本分散

$$\frac{\{(\text{階級値}-\text{平均})^2\times\text{度数}\}\text{の合計}}{\text{標本数}-1}$$

  - 実際の分散≥度数分布表による分散

3.3 度数分布表とパーセント点算出 3.3.1 算出法

- pパーセント点
  - m番目の階級で初めて累積相対度数がp/100を超えるとする.

    この授業では m番目の階級の階級値

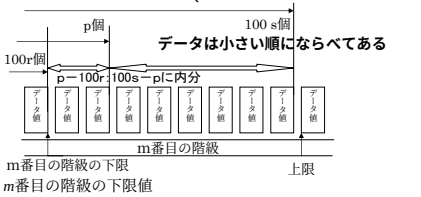
    本当は(教科書の決め方でもある)

$$m\text{番目の階級の下限値}+\left\{\frac{\text{標本数}\times p}{100}-(m-1\text{番目の階級の累積度数})\right\}\times\frac{\text{階級幅}}{m\text{番目の階級の度数}}$$

- 中央値は50%点，第1四分位点は25%点

3．3．2 本当の式の説明

- ある階級に属すデータが階級内に均等分布
  - m-1番目の累積相対度数はr，m番目はs，s-rはm番目の階級の相対度数(データは100個あると仮定)



$$m\text{番目の階級の下限値}+\frac{p/100-(m-1\text{番目の階級の累積相対度数})}{m\text{番目の階級の相対度数}}\times\text{階級幅}$$

3．4 度数分布表による最頻値算出

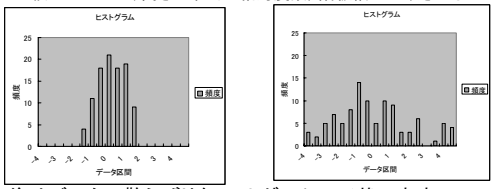
- この授業では最も度数の高かった階級の階級値とする
- 本当は二つの方法がある

$$\text{最大度数の階級の下限値}+\frac{\text{その後の階級の度数}}{\text{前後の階級の度数の合計}}\times\text{階級幅}$$

$$\text{最大度数の階級の下限値}+\frac{\text{最大度数の階級とその直前の階級の度数の差}}{2\times m\text{番目の階級の度数}-\text{前後の階級の度数の合計}}\times\text{階級幅}$$

3．5 データの散らばりとヒストグラム

- 散らばりが小さいデータのヒストグラムと大きいデータのヒストグラム
  - 比較のためには、高さは本当は相対度数，階級幅は一致させる



- 今後はデータの散らばりをヒストグラムの形状で表す

ヒストグラムの作り方

- 度数分布表を棒グラフ
  - ヒストグラム
- 度数分布表を折れ線グラフ
  - 度数多角形
- 累積度数表を棒グラフ
  - 累積ヒストグラム

4.　発展したデータの代表値（1）

- 刈り込み平均
  - 最大値，最小値，およびその周辺の値を平均値の算出からはずす
- 加重平均

$$\sum_{i=1}^{n} w_i x_i \qquad \sum_{i=1}^{n} w_i = 1, \quad w_i \geq 0$$

  - 平均は重み $w_i = 1/n$ となっている加重平均
  - 刈り込み平均は最大値，最小値，およびその周辺の値に関する重みを0にした加重平均

4．発展したデータの代表値（2）

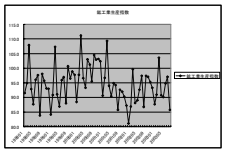
- 幾何平均
  - 成長率などその効果が積によって累積していく指標の平均に利用
  - 考え方
    - 1年目が10%成長，2年目が20%成長．1年あたりの平均成長率は？
      $1.1\times1.2 = a^2$ となるaを求める問題になる
      $a = \sqrt{1.1\times1.2} = 1.149$より年平均14.9%
    - 注意：(10+20)/2%＝15%ではない
  - $\sqrt[n]{\text{データ全部の積}}$
  - 容易に計算する方法：$10^{(\log_{10}\text{データ})\text{の平均}}$

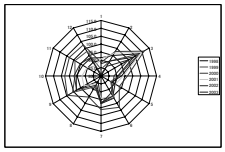
$$\log\text{幾何平均} = \log\sqrt[n]{\text{データ全部の積}}$$
$$= \frac{1}{n}\log(\text{データ全部の積}) = \frac{1}{n}\{\log(\text{データ})\text{の合計}\}$$

    - 度数分布表を使用する場合は，log値の平均を出す

4．2　季節調整
4．2．1　季節変動と季節調整

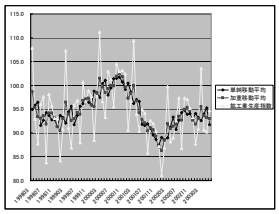
- 季節変動
  - 鉱工業生産の場合1月8月，5月に落ち込む
    - 休みなどのため
  - これでは，景気の指標として鉱工業生産指数を利用することはできない．
- 季節調整
  - 季節変動を取り除く





4．2．2　季節調整と移動平均

- 季節調整
  - 単純移動平均
    - あるデータと前後同一数のデータの平均
    - 例えば前後2期のデータも考えて平均をとる
  - 加重移動平均
    - 単純平均だと前後のデータも同一の比重で考慮
    - 現在のデータに比重をおく
    - 単純平均ではなく加重平均をとる
    - 加重の仕方には色々ある．（教科書は一例）
  - これらは単純な方法で初歩的分析に適する



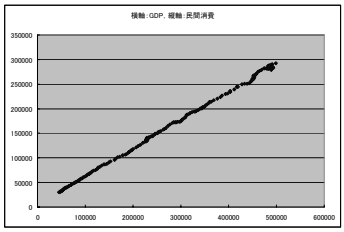
4．3　変動係数

- 散らばりを比較する
  - 平均が同じくらい
    - 分散や範囲，四分位範囲を比較すればよい
  - 平均が異なるデータの比較
    - 平均で標準化した標準偏差＝変動係数 $\dfrac{\text{標本標準偏差}}{\text{標本平均}}$
    - 四分位分散係数
      - 四分位範囲の半分を中央値で割ったもの
        - » 四分位の半分は片側の変動幅を示す
    - 十分位分散係数

$$\frac{\text{第９十分位数}-\text{第１十分位数}}{2\times\text{中央値}}$$

5　変数データの整理
5．1　散布図

- 2変数の関係を視覚的に捉える＝散布図
  - データ値の組み合わせを座標と考えて，データ毎に点をプロットする．（単位10億円）



５．２　標本相関係数（１）

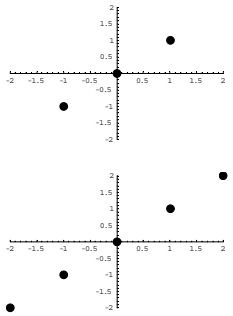
- ２変数の関係の深さを数値化する．
- 共分散

$$\text{標本共分散} = \frac{\left\{\begin{array}{l}(\text{x変数の値} - \text{xの標本平均}) \times (\text{y変数の値} - \text{yの標本平均}) \\ \text{のすべての標本に関する合計}\end{array}\right\}}{\text{n(またはn}-1)} = S_{xy}$$

  - 分散を計算するのにnで割ったかn-1で割ったかによってきめる
  - 共分散の性質
    - xの値がその平均を上回るときにはyもその平均を上回る様な場合，共分散の式の分子は正になる（右上がりの傾向）
    - xの値がその平均を上回るときにはyはその平均を下回る様な場合，共分散の式の分子は負になる（右下がりの傾向）
    - xの値とyの値に関係がない場合は，0になる（無相関）

５．２　標本相関係数（２）

- 共分散の困った点
  - 右の上下の図はともにxとyに線形の関係がある場合の散布図
  - しかし，上の場合の共分散は２／３で，下は２
  - つまり，共分散の絶対値が相関の強さを表していない．
  - xやyの散らばり方が大きい方が共分散は大きくなってしまう．



７．２　標本相関係数（３）

- 共分散をxとyの標準偏差で標準化
  - なぜ分散で標準化しないのか
    - 単位をそろえる．分子は，２乗の単位
    - 分母も2乗の単位でないと困る．すると，x，yともに1乗の単位にならなければならない．
    - 散らばりを表す1乗の単位の代表例が標準偏差
- 標本相関係数

$$r_{xy} = \frac{\text{標本共分散}}{\text{xの標本標準偏差} \times \text{yの標本標準偏差}} = \frac{S_{xy}}{S_x S_y}$$

  - 標本共分散をnで割って求めたら，標本分散（標準偏差）もnで割る
  - N－1で割ったら，分母氏で統一すること，そうしないと相関係数が1を超える場合もある

５．２　標本相関係数（４）

- 計算例（１）
  - {(x,y)}={(1,2),(2,3),(-1,5),(2,2)}の場合
  - 平均の計算

$$\bar{x} = \{1+2+(-1)+2\}/4 = 4/4 = 1$$

$$\bar{y} = (2+3+5+2)/4 = 12/4 = 3$$

  - 分散の計算

$$S_{xx} = \{(1-1)^2 + (2-1)^2 + (-1-1)^2 + (2-1)^2\}/4 = 6/4 = 1.5$$

$$S_{YY} = \{(2-3)^2 + (3-3)^2 + (5-3)^2 + (2-3)^2\}/4 = 1.5$$

裏スライド（１）

- なんで相関係数の式なの？

xデータの平均からの偏差ベクトル

$$\vec{x}^* = (x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_n - \bar{x})$$

yデータの平均からの偏差ベクトル

$$\vec{y}^* = (y_1 - \bar{y}, y_2 - \bar{y}, \cdots, x_n - \bar{y})$$

xの偏差ベクトルとyの偏差ベクトルの内積

$$\vec{x}^* \cdot \vec{y}^* = \sum_{i}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

xの偏差ベクトルの長さ（yも同様）

$$\left|\vec{x}^*\right| = \sqrt{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}$$

裏スライド（２）

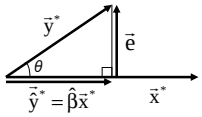
xの偏差ベクトルとyの偏差ベクトルとのなす角をθとすると

$$\cos\theta = \frac{\vec{x}^* \cdot \vec{y}^*}{\left|\vec{x}^*\right|\left|\vec{y}^*\right|} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{XY}}{S_X S_Y} = r_{xy}$$

つまり，相関係数は相関の強さをxの平均からの偏差とyの平均からの偏差のなす角のcos（cosθ）で計っていたのだ．

裏スライド（3）

- $\cos\theta$って？
  $\hat{\vec{y}}^*$を右図の緑のベクトルとすると$\hat{\vec{y}}^* = \hat{\beta}\vec{x}^*$となる$\beta$が存在する．
  $\vec{e}$を右図の青のベクトルとすると，$\vec{y}^* = \hat{\vec{y}}^* + \vec{e}$であり，$\cos\theta = |\hat{\vec{y}}^*| / |\vec{e}|$
  つまり，yの偏差をxの偏差の方向とそれに直交する（無相関）方向に分解したときyの偏差の長さとxの偏差方向の分解成分の長さの比が$\cos\theta$である．つまり，yの偏差のどれだけがx偏差方向の成分で説明できるかを示す



裏スライド（4）

$\vec{e} = (e_1, e_2, \cdots, e_n)$ とおくと $\vec{y}^* = \hat{\beta}\vec{x}^* + \vec{e}$ となるが，
ベクトルの成分表示で書くと

$$(y_1 - \bar{y}, \cdots, y_n - \bar{y}) = \hat{\beta}(x_1 - \bar{x}, \cdots, x_n - \bar{x}) + (e_1, \cdots, e_n)$$

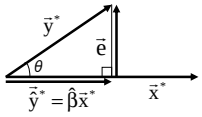
となる．ベクトルの同一性の定義から $y_i - \bar{y} = \hat{\beta}(x_i - \bar{x}) + e_i$
となる．これは，yの偏差を（xの偏差の定数倍＋xと無相関な誤差）と表したということ．さらに，ここから，$\hat{\alpha} = -\hat{\beta}\bar{x} + \bar{y}$ とおくと $y_i = \hat{\beta}x_i + (-\hat{\beta}\bar{x} + \bar{y}) + e_i$

$$y_i = \hat{\beta}x_i + \hat{\alpha} + e_i$$
$$= (x_i\text{の一次式}) + (x_i\text{と無相関, i.e.共分散が0の成分})$$

裏スライド（5）

- だから，$\cos\theta$はyをxの一次式とXと無相関な成分に分解したときにyの偏差のうちどれだけがxの一次式の偏差で表せるかと言うこと
- 三平方の定理

$$|\vec{y}^*|^2 = |\hat{\vec{y}}^*|^2 + |\vec{e}|^2$$
$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\hat{\beta}^2(x_i - \bar{x})^2 + \sum_{i=1}^{n}e_i^2$$

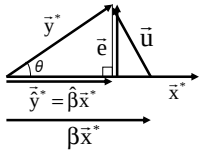


yの変動＝xの一次式の変動＋残差変動
$\cos^2\theta$ ＝xの一次式の変動/yの全変動

裏スライド（6）

$\hat{\beta}$ 以外のβも考えて $\vec{y}^* = \beta\vec{x}^* + \vec{u}$ とする，つまり，成分ごとに考えて，$y_i - \bar{y} = \beta(x_i - \bar{x}) + u_i$ とすると，$|\vec{u}| \geq |\vec{e}|$ となる．すなわち，

$$\sum_{i=1}^{n}u_i^2 = \sum_{i=1}^{n}\{(y_i - \bar{y}) - \beta(x_i - \bar{x})\}^2$$
$$\geq \sum_{i=1}^{n}e_i^2 = \sum_{i=1}^{n}\{(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})\}^2$$



u_i をyの偏差をxの偏差の定数倍で表したときの誤差と考えると，その誤差の二乗和が最小なのは $\beta = \hat{\beta}$ のとき．（最小二乗法）

５．２　標本相関係数（５）

- 計算例（２）
  - 共分散の計算

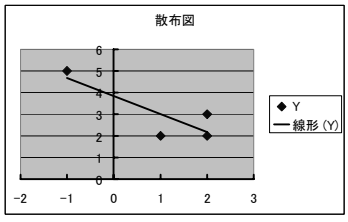
$$S_{XY} = \frac{(1-1)(2-3)+(2-1)(3-3)+(-1-1)(5-3)+(2-1)(2-3)}{4}$$
$$= -5/4 = -1.25$$

  - 相関係数の計算

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}} = \frac{-1.25}{\sqrt{1.5}\sqrt{1.5}} = \frac{-1.25}{1.5} = -0.833$$

５．２　標本相関係数（６）

- 散布図

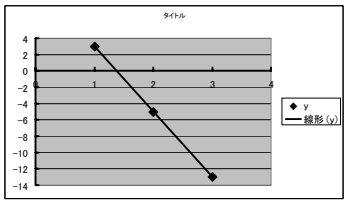


７．２　標本相関係数（7）

- 標本相関係数の性質
  - 符号は相関の方向を表す
    - xが平均より大きかったらyも平均より大きい傾向
      - 符号は正，正の相関を持つと呼ぶ
    - xが平均より大きかったらyは平均より小さい傾向
      - 符号は負，正の相関を持つと呼ぶ
  - 絶対値は相関の強さ，関係の深さを表す
    - 大きいほど関係は深い
    - 0～1の間（絶対値が1の場合はyはxの一次式）
  - 標本相関係数は－1以上＋1以下

例題

- {(x,y)}={(1,3),(2,-5),(3,-13)}のとき，xとyの標本相関係数を求めよ．
  - yがxの1次式なので標本相関係数の絶対値は1
  - yとxが負の相関を示すので，標本相関係数の符号は負
  - よって－1



７．２　標本相関係数（8）

- GDPと消費の例では標本相関係数は0.999
- 相関係数の限界
  - 相関係数は線形関係が完全に成立している場合を基準にしている＜相関係数を計算する前に散布図を書こう＞
    - 非線形関係，例えば，2次式の関係などは表せない．
    - 教科書の例では，相関係数が0なのに非線形の関係がある
  - 見せかけの相関
    - 二つの変数とも，第3の変数と相関を持っていて，その結果，両変数に関係があるように見える場合がある．
    - 例えば，時系列データの場合に，二つの変数が時間とともに増加，または，減少している場合

７．３　標本偏相関係数

- 第3の変数と両方とも相関を持っていて，そのために2つの変数の相関が強いように見える場合の対策
  - 第3の変数（zとする）の影響を取り除いたうえで，二つの変数の相関を調べる．
  - 標本偏相関係数 $r_{xy|z} = \dfrac{r_{xy} - r_{xz} r_{zy}}{\sqrt{\left(1 - r_{xz}^2\right)\left(1 - r_{zy}^2\right)}}$
  - 「子のつく名前は頭がいい」
    ←女子高校の子の付く名前の生徒の比率とその高校の偏差値に相関
    しかし，第3の変数と関係があるだけではないか？
    第3の変数候補，その家庭の所得，親の資産，皇室への関心度
    このような変数との相関が容易には調べられないのをいいことに，さも直接的な関係があるように見せているのでは？
- 分割表は教科書参照のこと