2-1 データを読む

東京大学 数理・情報教育研究センター
2020年5月11日

概要

- 本節では、データを要約したり可視化したりする様々な手法について学ぶことで、グラフや統計情報の読み方を学び、起きている事象の背景や意味合いを理解することを目標とします。

- また、これらの情報を読む上で注意すべきいくつかの点についても学びます。データを正しく読む上では、データがどのような背景から得られたものなのかを正しく理解することも重要です。

本教材の目次

データの種類

- データの種類として大きく分けて「量的変数」と「質的変数」の２つがあります。

- 量的変数
  - 数量で表すことができるもの
    例：長さ、重さ、温度、・・・

- 質的変数
  - 数量で表すことが困難であるもの
    例：性別、職業、既婚／未婚、・・・

データの例

- データの項目はデータによって異なります。
  - 右の例では「緯度」「経度」「地震の深さ」「マグニチュード」「計測地点数」の五つの項目があります。
- 一つのデータに対して一行が割り振られ、データの数だけ行数があります。

＜フィジーの地震のデータ＞

緯度	経度	地震の深さ（km）	マグニチュード	計測地点数
-20.42	181.62	562	4.8	41
-20.62	181.03	650	4.2	15
-26	184.1	42	5.4	43
-17.97	181.66	626	4.1	19
-20.42	181.96	649	4	11
-19.68	184.31	195	4	12
-11.7	166.1	82	4.8	43
-28.11	181.93	194	4.4	15
-28.74	181.74	211	4.7	35
-17.47	179.59	622	4.3	19
-21.44	180.69	583	4.4	13
-12.26	167	249	4.6	16
-18.54	182.11	554	4.4	19
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・

度数分布

- データを適当な範囲で区切ってそれぞれに入るデータ数を表にしたもの
  - 区切る範囲は分析者が指定します。

＜データ＞

緯度	経度	地震の深さ（km）	マグニチュード	計測地点数
-20	182	562	4.8	41
-21	181	650	4.2	15
-26	184	42	5.4	43
-18	182	626	4.1	19
-20	182	649	4	11
-20	184	195	4	12
-12	166	82	4.8	43
-28	182	194	4.4	15
-29	182	211	4.7	35
-17	180	622	4.3	19
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・

＜度数分布＞

マグニチュード	回数
4.0～4.5	377
4.5～5.0	425
5.0～5.5	160
5.5～6.0	33
6.0～6.5	5

データ数が多くても全体の傾向がわかりやすくなります。

※データ数は1000で、マグニチュード４以上の地震のみ。

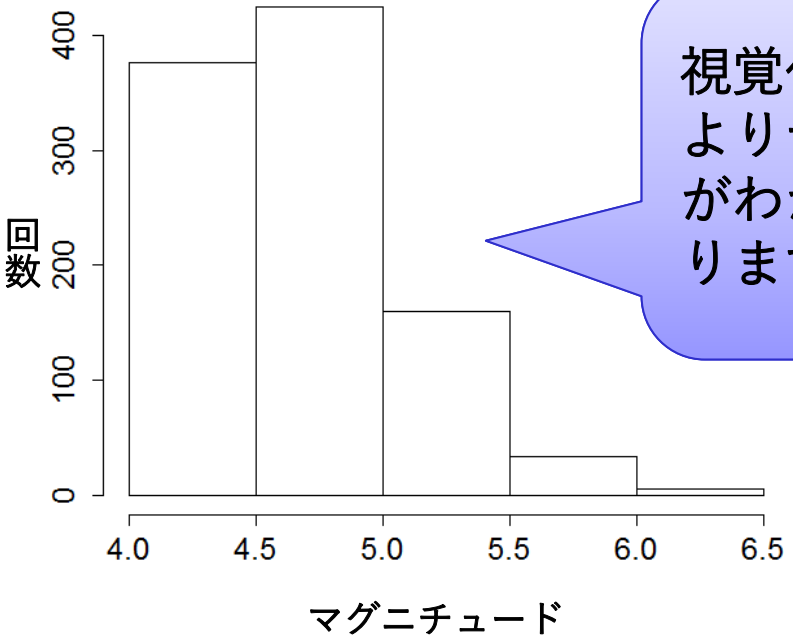
ヒストグラム

- 度数分布の表の値を棒グラフで表し、視覚化したもの

＜度数分布＞

マグニチュード	回数
4.0～4.5	377
4.5～5.0	425
5.0～5.5	160
5.5～6.0	33
6.0～6.5	5

＜ヒストグラム＞



視覚化することでよりデータの傾向がわかりやすくなります。

データを代表する数値：平均

- 「平均」は最もよく使われる代表値です。データを $X_1, \cdots, X_n$と 書くと、平均は$\bar{X}$またはμなどと書き,

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

と計算されます。

- 平均はデータの中心を表していると考えることができます。

＜データ＞
3　7　4　1　8

$$\text{平均値} = \frac{3+7+4+1+8}{5} = 4.6$$

データを代表する数値：中央値

- 「中央値」は、データを小さい順に並べた時に真ん中に来る値です。
- データが偶数個なら真ん中の２つの値を足して２で割った値が中央値になります。

＜データ＞

3　7　4　1　8

↓　並べ替え

1　3　4　7　8

中央値：４

＜データ＞

5　4　7　2

↓　並べ替え

2　4　5　7

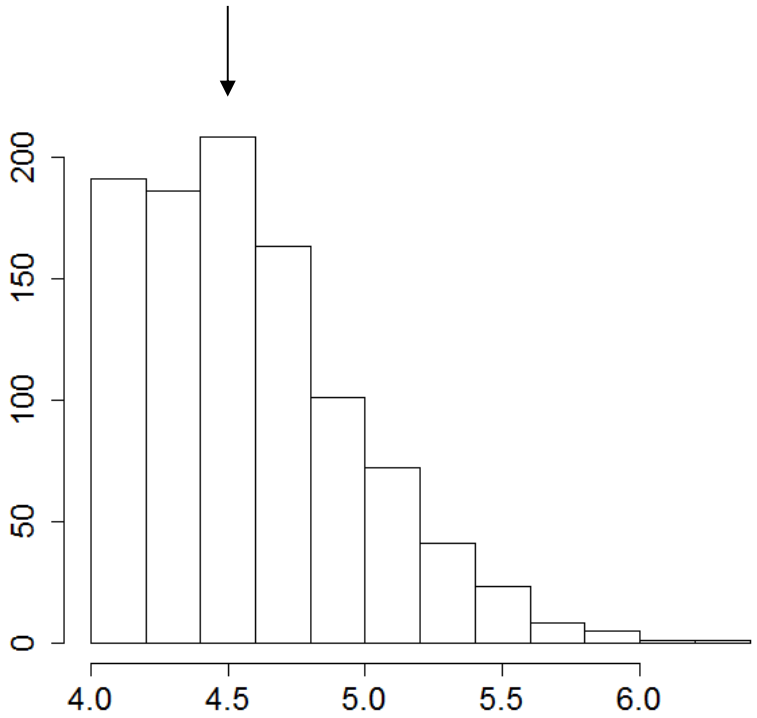
足して２で割る

中央値：４.5

データを代表する数値：最頻値

- 「最頻値」とはデータの中で最も頻繁に現れた値のことです。
- 連続した値のデータでは度数分布の中で最も回数の多い範囲の中央の値を最頻値とします。

最頻値：４．５

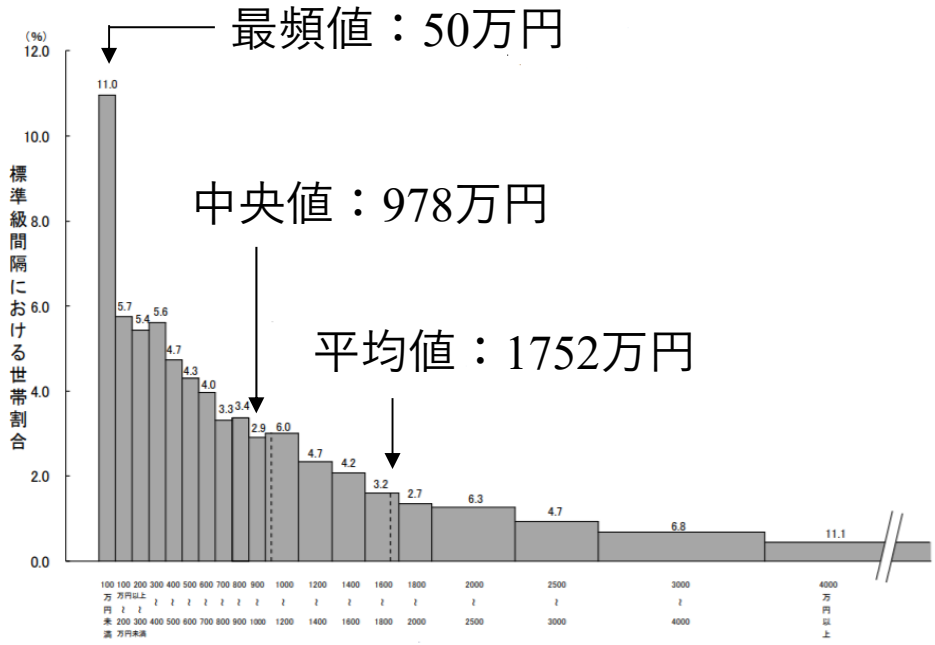


代表値の性質の違い

- ３つの代表値は、実際のデータでは値が異なることも多いです。
- 一部のデータが非常に大きな値となる時、平均値は高くなりやすいです。

- 右図は2018年の全国の二人以上の世帯の貯蓄額のヒストグラムです。

- 一部の裕福な世帯の影響を受け、平均値は中央値よりもかなり高い値になっています。（３分の２の世帯が平均を下回る）



「貯蓄現在高階級別世帯分布（二人以上の世帯）」（総務省統計局）を加工して作成

（https://www.stat.go.jp/data/sav/sokuhou/nen/pdf/2018_gai2.pdf）

データのばらつき（分散、標準偏差）

- データのばらつき度合いを測る指標として、「分散」「標準偏差」「偏差値」などがあります。
- 分散：データを$X_1, \cdots, X_n$として平均を$\bar{X}$とすると、分散σ^2は

$$\sigma^2 = \frac{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n}$$

で与えられます。

  - 各データX_iに対し、$X_i - \bar{X}$の絶対値が大きい時に分散の値が大きくなるので、各データが平均からどの程度離れているかというばらつき度合いを測る指標となります。
- 標準偏差は$\sigma = \sqrt{\sigma^2}$と定義されます。
  - 分散はデータの２乗を計算し、例えば重さ(g)のデータであれば、単位が(g^2)となり、単位が変わってしまいますが、標準偏差の単位は元の単位(g)と同じになります。

データのばらつき（偏差値）

- データを$X_1, \cdots, X_n$として、平均を$\bar{X}$、標準偏差をσとした時、

$$\frac{X_i - \bar{X}}{\sigma} \times 10 + 50$$

  の値を偏差値といいます。
- データが平均値に等しい時（$X_i = \bar{X}$）、偏差値は50となります。
- 偏差値はデータのばらつきを補正した時の各データの位置づけを表していて、おおよそのデータの偏差値は30～70程度の範囲に収まります。

➢ 各偏差値のおおよその目安は右の表のようになります。

偏差値	上位からの割合	偏差値	上位からの割合
70	2.3%	45	69.1%
65	6.7%	40	84.1%
60	15.9%	35	93.3%
55	30.9%	30	97.7%
50	50.0%		

分割表、クロス集計表

- データの２種類の項目について、項目の値のペア毎にデータの個数を数え、表にまとめたものが「分割表」、さらに２種類の項目の値のペア毎（例えば右下表で「文系」と「2組」のペア等）に別の項目（右下表では「点数」）の合計、平均、標準偏差等を集計したものが「クロス集計表」と呼ばれます。

＜データ＞
（クラス・文理別の数学の点数）

文理	クラス	点数
理系	2組	93
理系	１組	48
文系	３組	41
文系	３組	28
理系	３組	75
文系	３組	68
・	・	・
・	・	・
・	・	・

＜分割表＞

	１組	２組	３組	全体
文系	44	39	34	117
理系	38	46	36	120
全体	82	85	70	237

各項目ペアに対してデータの個数を数える

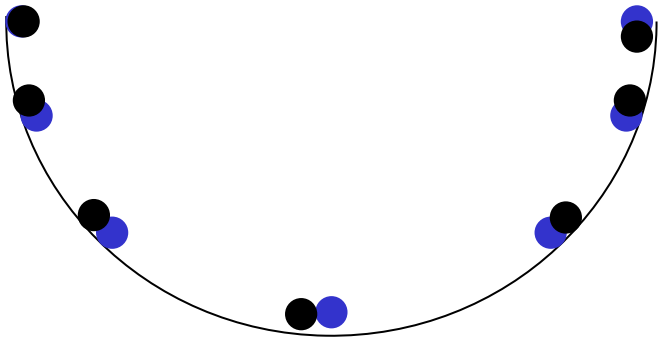
＜クロス集計表＞（点数の平均値を集計）

	１組	２組	３組	全体
文系	38.3	39.8	43.0	40.2
理系	52.4	50.9	57.8	53.5
全体	44.8	45.8	50.6	46.9

各項目ペアでの平均等をそれぞれ計算

観測データに含まれる誤差の扱い

- 物理現象などを観測するとき、観測データに測定誤差がふくまれることがあります。



例：物体の一定時間毎の位置データ

青点：力学から導かれる物体の位置
黒点：実際に測定した物体の位置

「観測値」＝「理論値」＋測定誤差

➢ 測定誤差がランダムだと思って、確率論を使って処理することにより、データの特性をうまく捉えることができます。

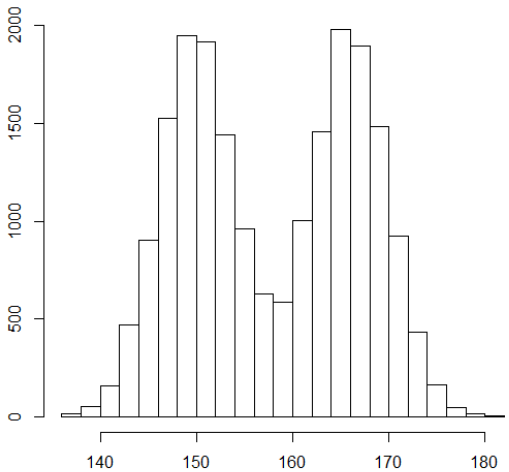
打ち切りや脱落を含むデータ

- データの統計情報を見る時に、打ち切りや脱落により表面上現れないデータがある場合があるので注意が必要です。

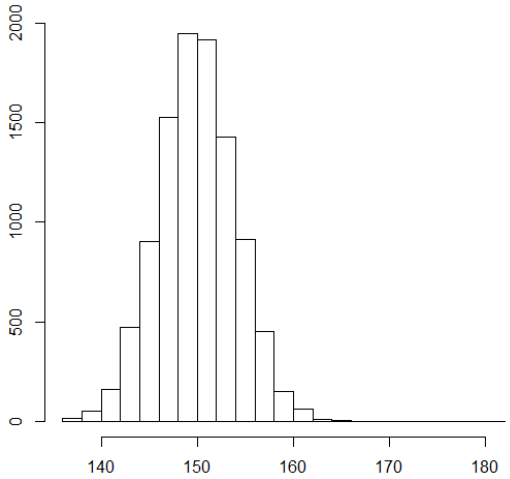
- 例えば、プロ野球選手の平均年俸が4,000万円といった情報を見ると、「野球選手になると収入が高い」と思ってしまいがちですが、背後にはプロ契約を結べなかった選手達が多く存在し、そのような選手たちの情報はデータ上には現れません。

- 背後に脱落したデータがあることを知らないまま分析すると、誤った結論を導くことがあります。

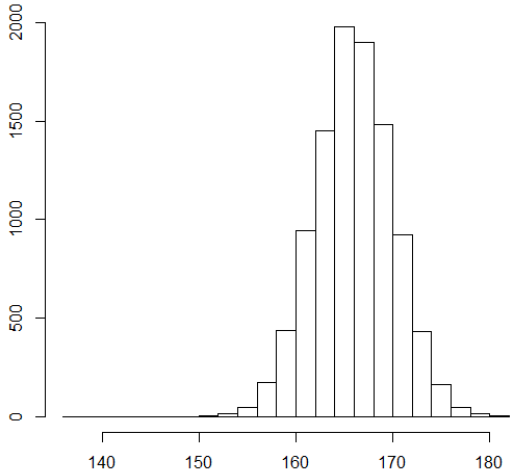
層別の分析が必要なデータ



男女の身長データ



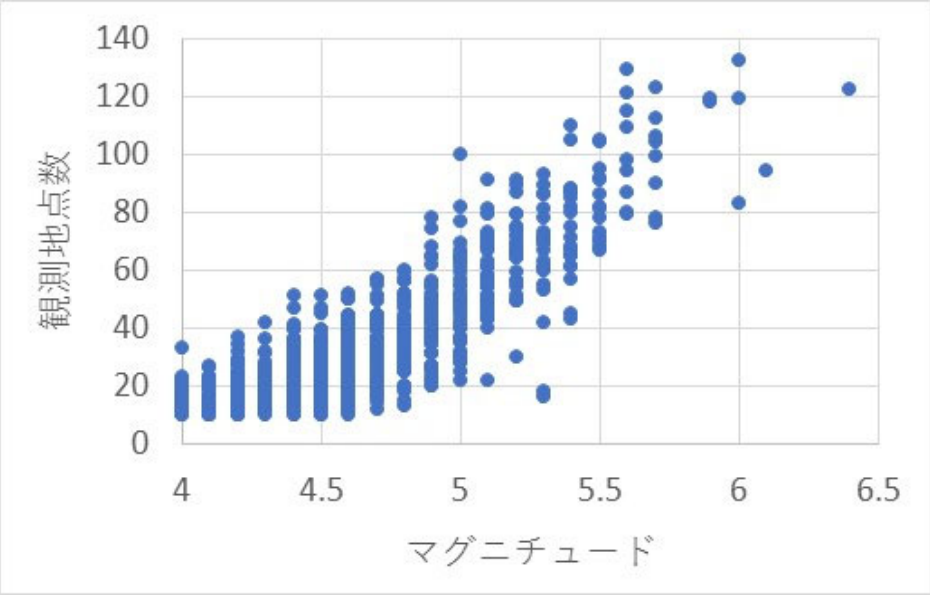
女性の身長データ



男性の身長データ

- 異なる属性のデータが混合したデータにおいては、層別にデータを分けた方が分析しやすいことがあります。
- 上の男女のデータは性別毎に分解することできれいな山のデータになり、分析がしやすいです。

散布図と相関係数



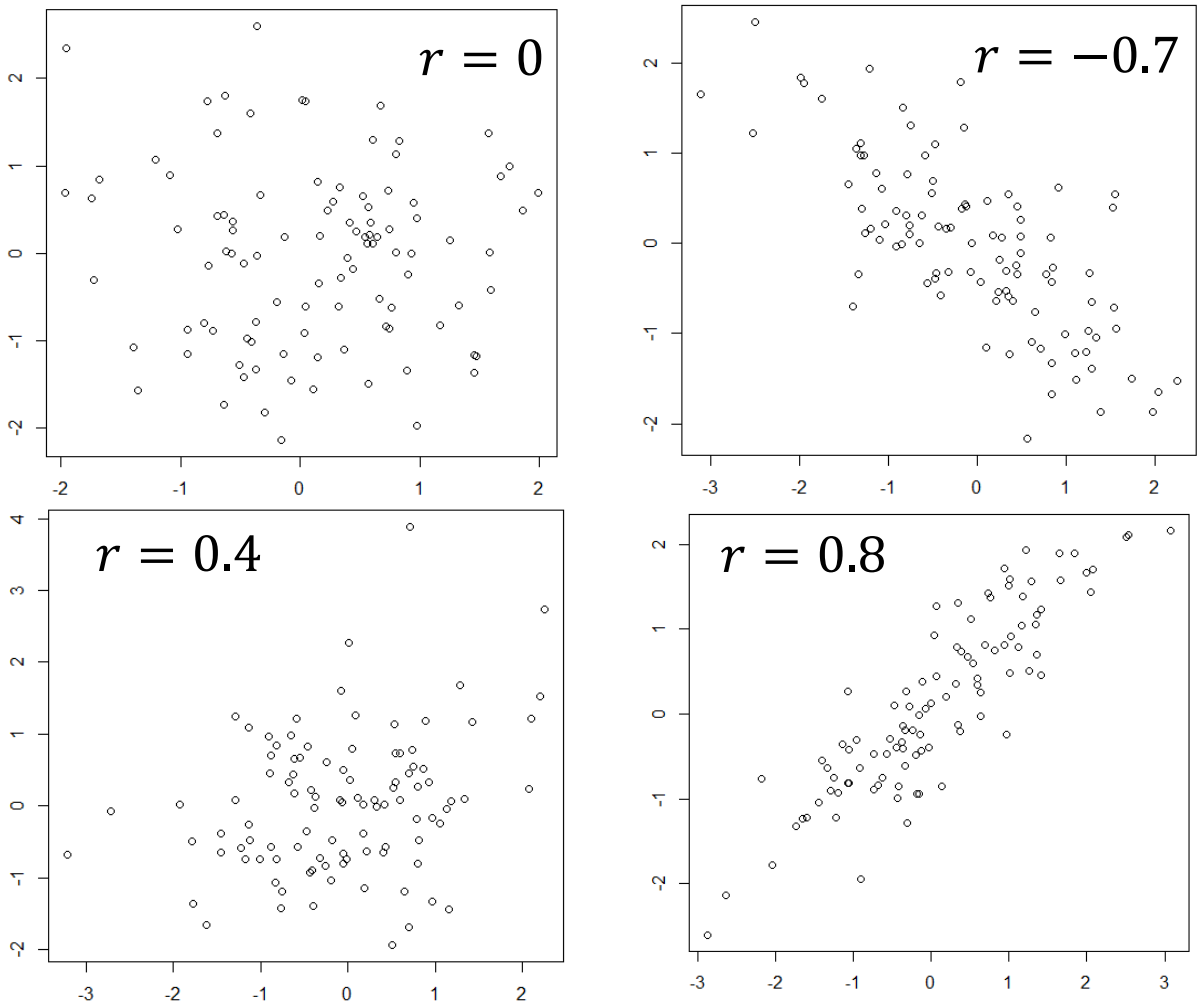
- 散布図
  - 右図のように、データの2種類の項目について2次元にプロットしたものを散布図といいます。
- 相関係数
  - ２種類のデータ$X_1, \cdots, X_n$と$Y_1, \cdots, Y_n$に対して、$X_1, \cdots, X_n$の標準偏差をσ_Xとし、 $Y_1, \cdots, Y_n$の標準偏差をσ_Yとすると、相関係数rは以下で定義されます。

$$r = \frac{(X_1 - \bar{X})(Y_1 - \bar{Y}) + \cdots + (X_n - \bar{X})(Y_n - \bar{Y})}{n\sigma_X\sigma_Y}$$

  - X_i, Y_iが平均から見て同じ方向に動くときにrの値は高くなるので、相関を計算することで連動性を測ることができます。
  - 分母にσ_X, σ_Yがあることで、 $-1 \leq r \leq 1$　となることが保証されます。

相関係数の例

- 相関係数rを変えた時の散布図は以下のようになります。
- rが大きい程散布図は右肩上がりで、-1に近いと右肩下がりになります。



相関行列

- データの全ての項目に対して、任意の 2 種類の相関をマトリックスで表示したものを相関行列といいます。

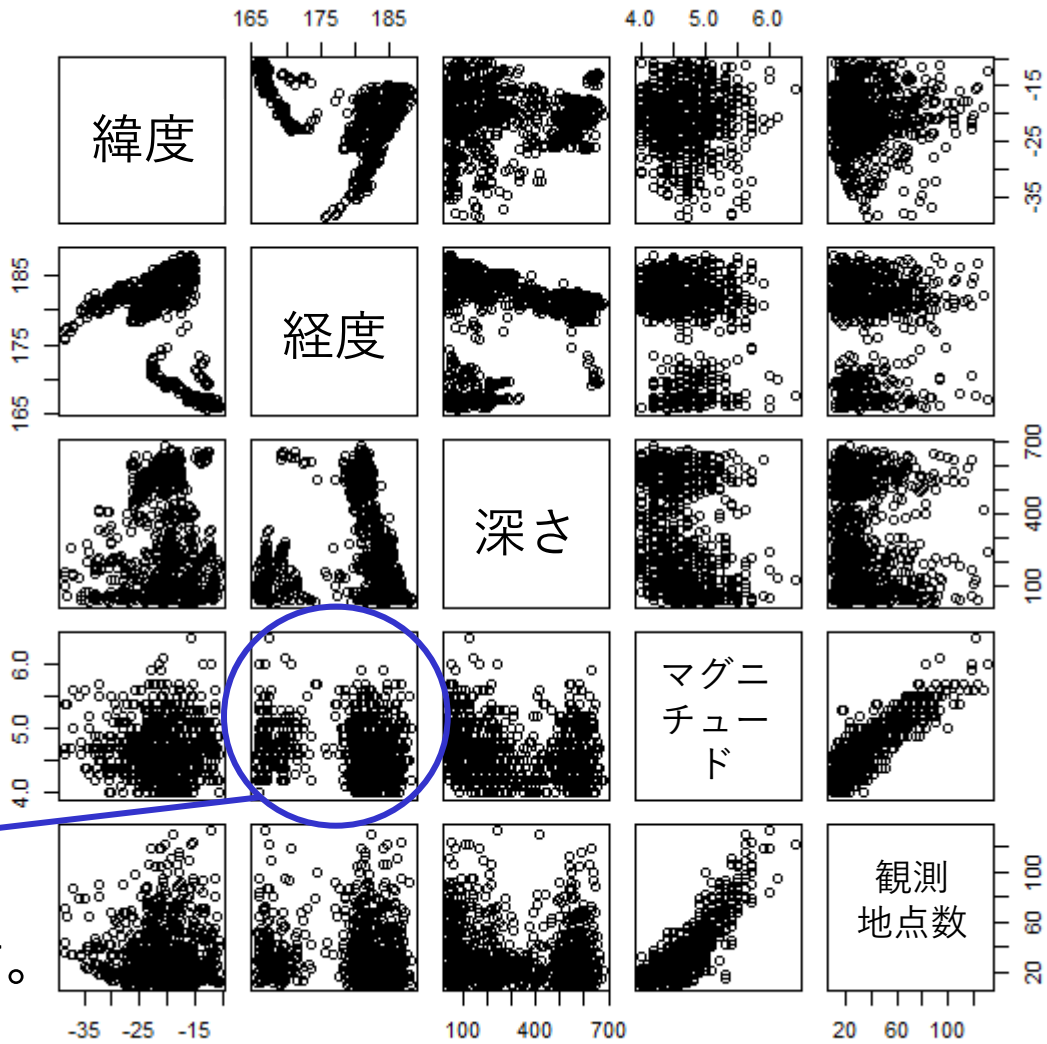
	緯度	経度	地震の深さ（km）	マグニチュード	計測地点数
緯度	1.00	-0.36	0.03	-0.05	-0.00
経度	-0.36	1.00	0.14	-0.17	-0.05
地震の深さ（km）	0.03	0.14	1.00	-0.23	-0.07
マグニチュード	-0.05	-0.17	-0.23	1.00	0.85
計測地点数	-0.00	-0.05	-0.07	0.85	1.00

「経度」と「地震の深さ」の相関が0.14ということになります。

同じ項目の相関は1になります

散布図行列

- データの全ての項目に対して、任意の２種類の散布図をマトリックスで表示したものを散布図行列と呼びます。



「マグニチュード」と「経度」の散布図を表します。

相関と因果

- ある２つの項目の相関係数が高いからといって、その２つの項目に因果関係があると言えるわけではありません。

- 例えば、小学生の「足のサイズ」と「学力」のデータをとると相関係数が正になります。

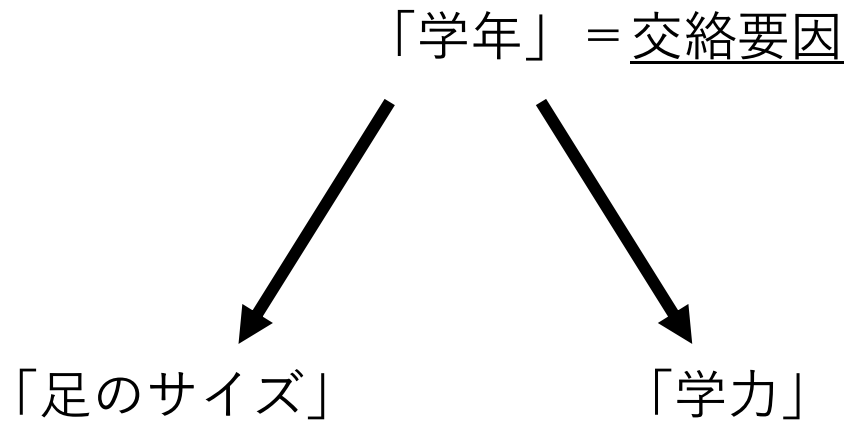
これはどう解釈すればよいでしょうか？

「足が大きくなると学力が高くなる？」

「勉強をすると足も伸びてくる？」

交絡要因と擬似相関

- 「足のサイズ」と「学力」の双方に影響を与える要因として「学年」という要因が考えられます。
  - 学年が高くなると、足のサイズも大きくなり、学力も高くなるので、結果として足のサイズと学力の相関が高くなると解釈できます。
  - このように因果のない２項目の相関が高くなることを擬似相関といい、２項目に影響を与えて相関を高くするような隠れた第三の要因のことを「交絡要因」と呼びます。

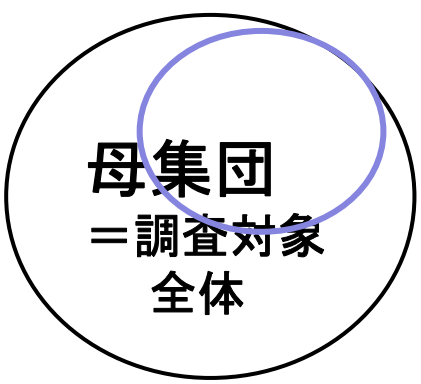


結果として、２項目の相関が高くなります

母集団と標本抽出

- データ解析を行う際には、調査したい対象のデータが全て手に入るとは限りません。一部のデータのみに対して解析を行う場合、（全ては入手できない）調査対象全体を「母集団」と呼び、入手可能な一部のデータを標本といいます。
- 調査対象全体からすべてのデータを集める時、「全数調査」、一部のデータのみを集める時、「標本調査」といいます。

全てのデータを集めるのが「全数調査」



一部のデータのみを集めるのが「標本調査」

例：国民全員のデータを集める国勢調査

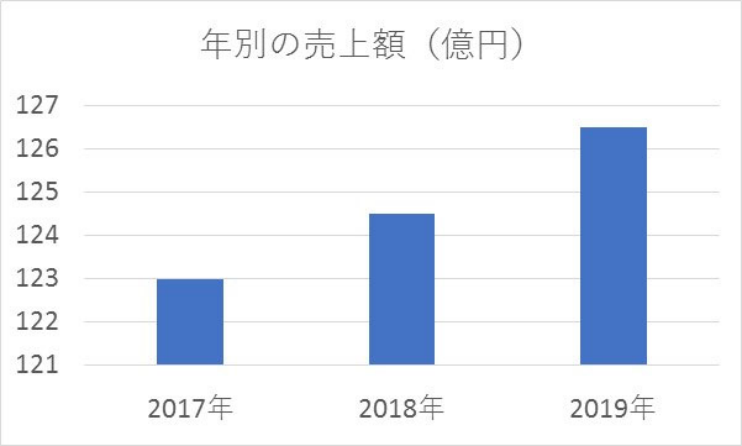
例：ある商品の購入者の一部に行うアンケート調査

標本の抽出方法

- 標本調査を行う場合は、標本のデータから母集団を推測できるように、偏りが生じないような標本の抽出が重要になり、代表的な抽出方法として以下のような方法が挙げられます。
  - <u>無作為抽出</u>
    - 母集団の中からランダムに標本を抽出します。
  - <u>層別抽出</u>
    - 母集団を属性毎にいくつかに分け（例えば性別・年代・業種等）、各層から必要数の標本を抽出することで属性の偏りを避ける方法になります。
  - <u>多段抽出</u>
    - 母集団をいくつかのグループに分け、まずグループをランダムに選びます。選んだグループをさらに小グループに分けてその中からランダムに選ぶことを繰り返し、十分小さくなったら、またその中からランダムにデータを抽出する方法になります。
    - 例えば、まず全国から都道府県をランダムに選んで、その中から地域をランダムに選んで、その地域からランダムに人を選んでデータを抽出するなどがあります。

統計情報の正しい理解

- データに関する数値やグラフを見る際には、誇張表現等に騙されず、統計情報を正しく読み解くことが大事になります。
  - 例えば、既出の「世帯別の貯蓄金額」のデータでは、平均は一部のデータの影響から高めの数値がでており、「データの真ん中」という感覚からはずれたものになっているので、中央値も併せて見ておくのがよいです。
  - また、左下のグラフはある企業の年別の売上額を表し、売上が上がっているように見えますが、軸の目盛りを 0～140に変えた右下図では印象が大きく異なります。このような誇張表現に注意すべきとなります。



軸の目盛りを
変えると・・