1-6-2 統計学の基礎

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。

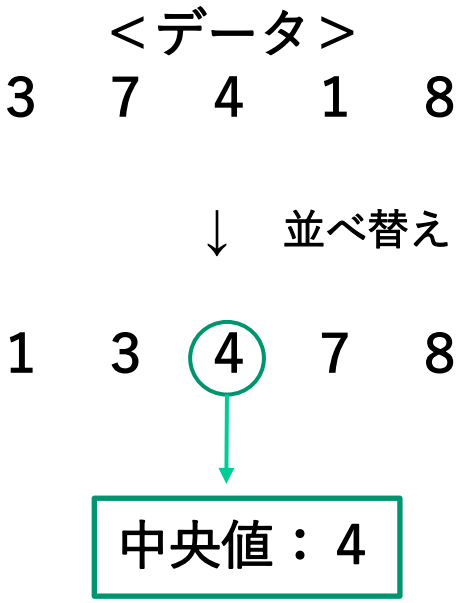
- 平均はデータの中心を表していると考えることができます。

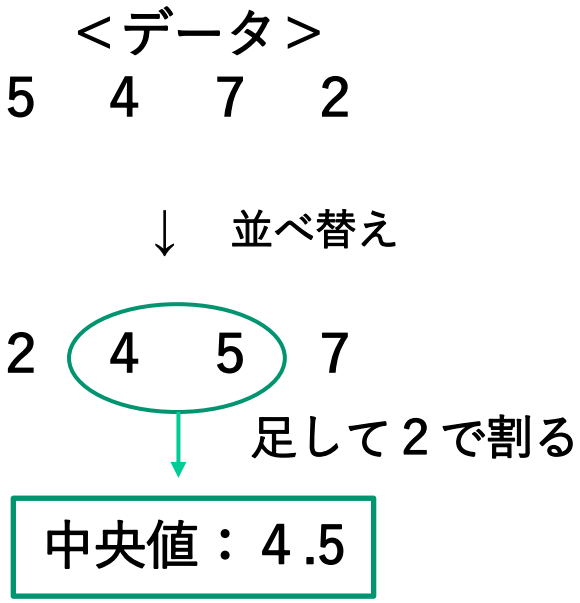
＜データ＞
3　7　4　1　8

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

データを代表する数値：中央値

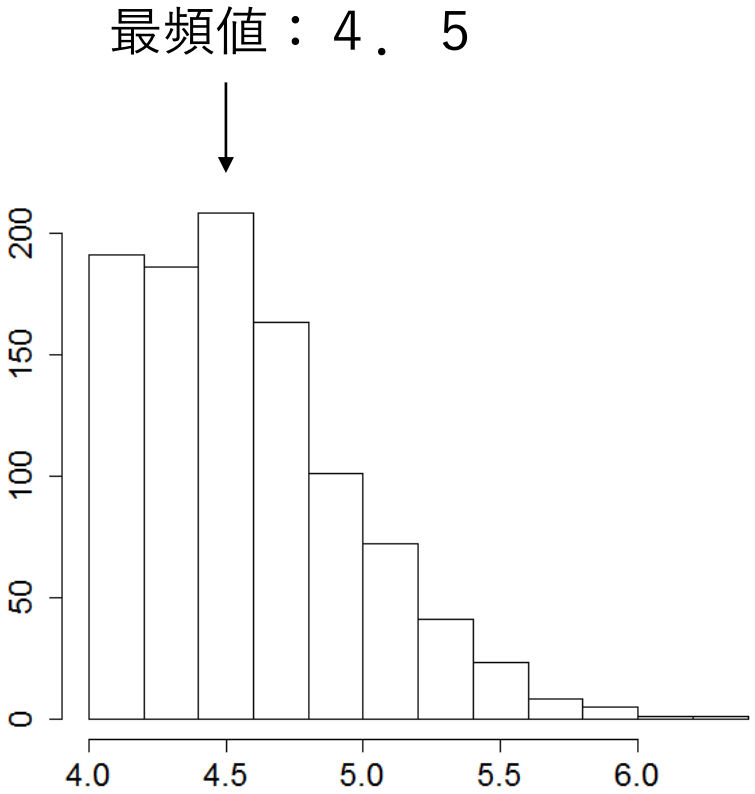
- 「中央値」は、データを小さい順に並べた時に真ん中に来る値です。
- データが偶数個なら真ん中の２つの値を足して２で割った値が中央値になります。





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。

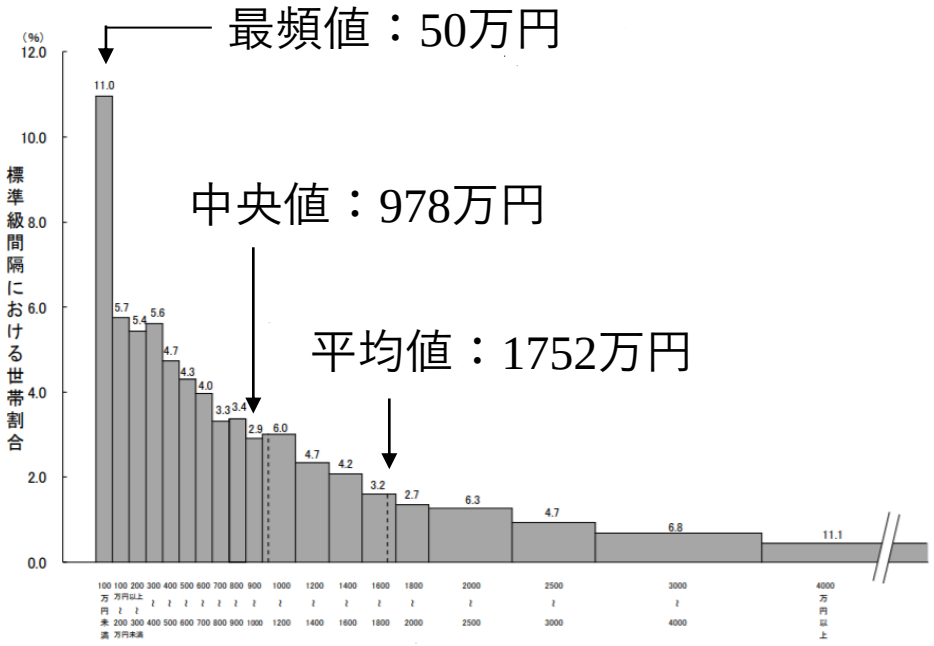


代表値の性質の違い

- 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、

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

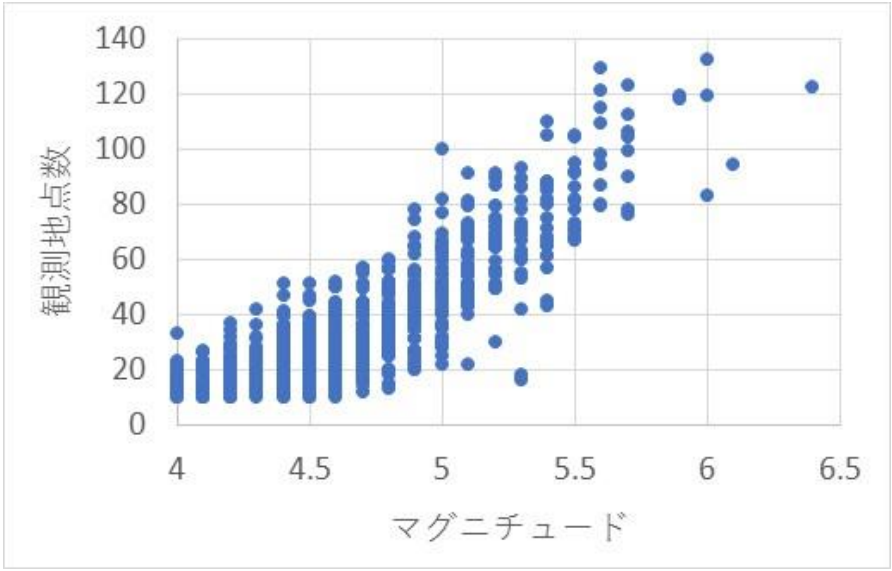
の値を偏差値といいます。

- 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%		

散布図と相関係数



フィジーの地震データ

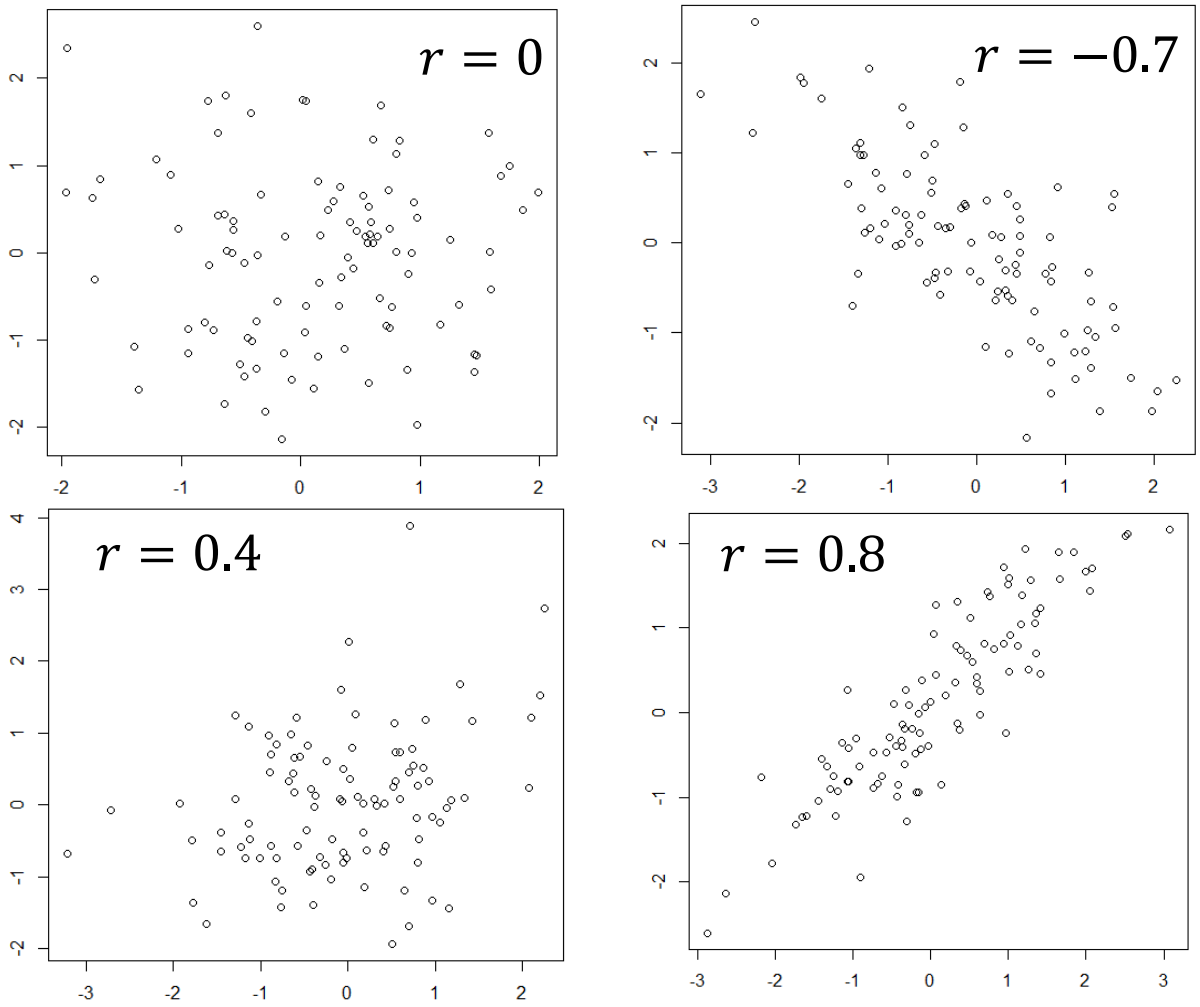
- 散布図
  - 右図のように、データの2種類の項目について2次元にプロットしたものを散布図といいます。
- 相関係数
  - ２種類のデータ$X_1,\cdots,X_n$と$Y_1,\cdots,Y_n$に対して、$X_1,\cdots,X_n$の標準偏差をσ_Xとし、 $Y_1,\cdots,Y_n$の標準偏差をσ_Yとすると、相関係数rは以下で定義されます。

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

  - 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
-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
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・

フィジーの地震データ

※赤字はマイナスのデータ。

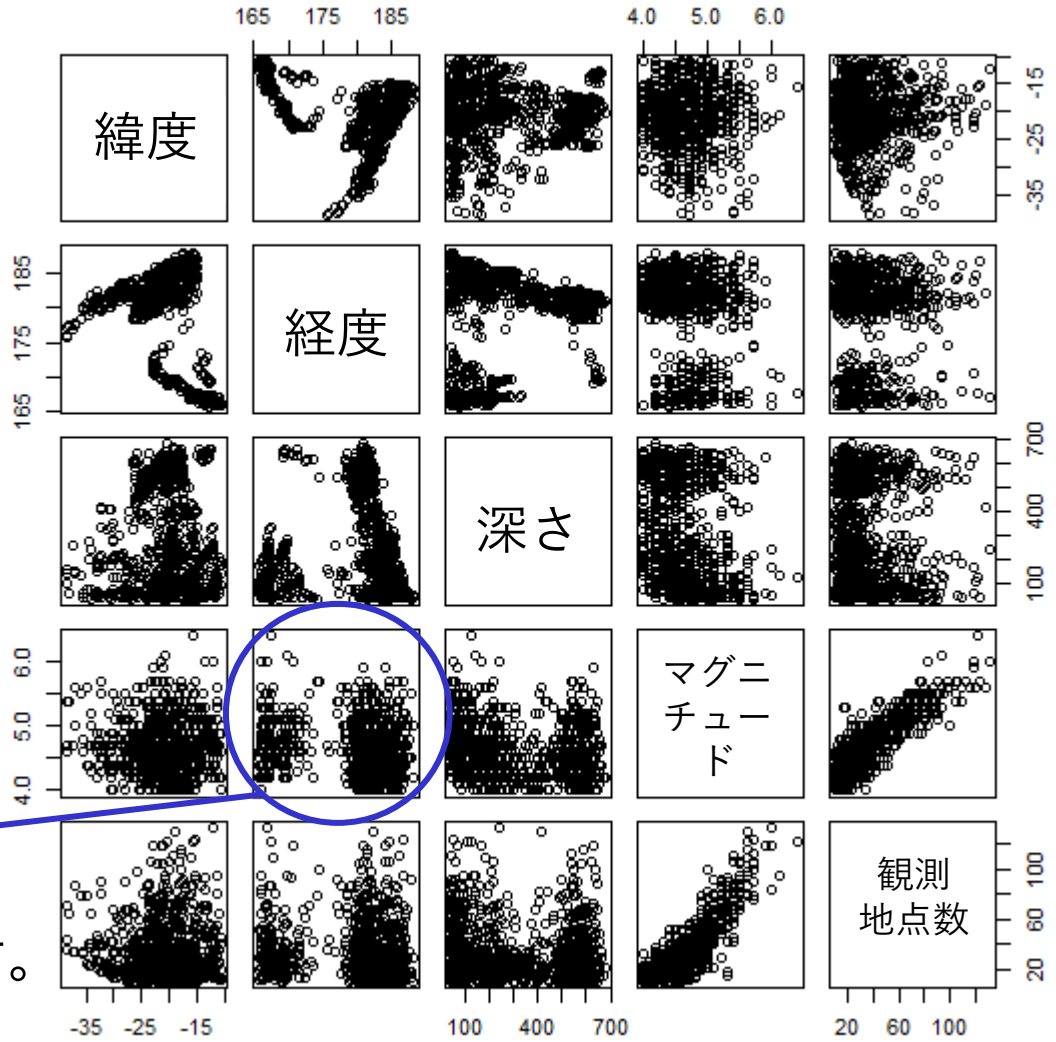
	緯度	経度	地震の深さ（km）	マグニチュード	計測地点数
緯度	1.00	-0.36	0.03	-0.05	-0.00
経度	-0.36	1.00	0.14	-0.17	-0.05
地震の深さ（km）	0.03	0.14	1.00	-0.23	-0.07
マグニチュード	-0.05	-0.17	-0.23	1.00	0.85
計測地点数	-0.00	-0.05	-0.07	0.85	1.00

「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。

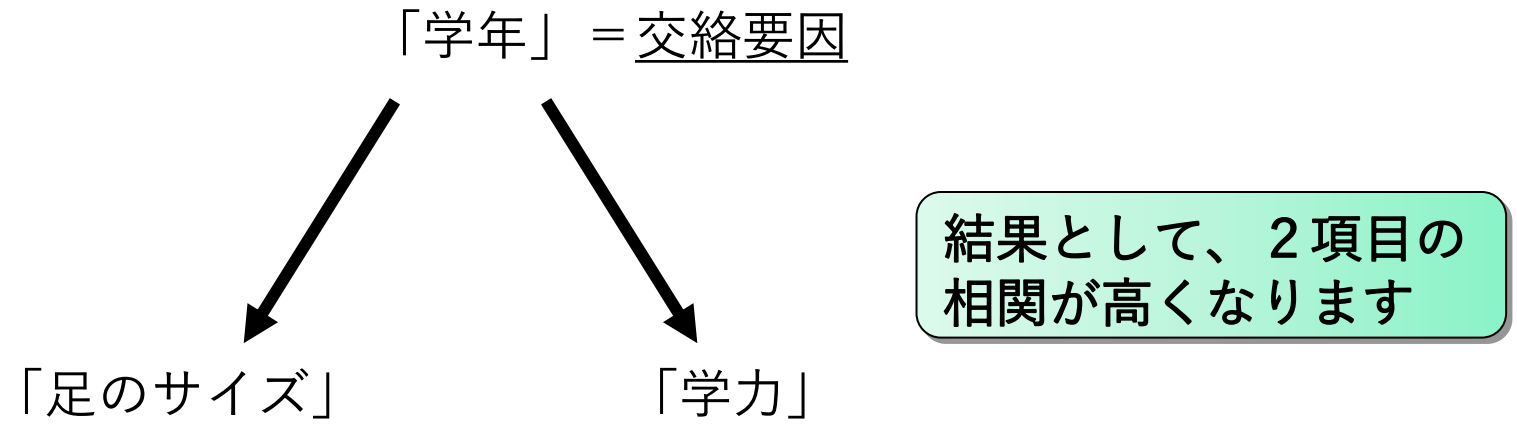
  これはどう解釈すればよいでしょうか？

  「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。



様々な尺度

- 統計学におけるデータは性質に応じて以下の４つの種類（尺度）に分けることができます。

- **名義尺度**・・・互いに区別はできますが順序に意味はない尺度です。
  - 例：氏名、職業、性別

- **順序尺度**・・・順序に意味がありますが、足したり引いたりすることはできない尺度です。
  - 例：大学の成績（優、良、可、不可）、ガンのステージ（I/II/III/IV）
  - 「優＞良＞可＞不可」といったように順序に意味はありますが、「優 ー 良」といった操作はできません。

様々な尺度

- **間隔尺度**・・・数値として表現され、間隔にも意味があり、足したり引いたりできますが、比率（２倍、３倍）には意味がないような尺度です。
  - 例：温度、西暦
  - 「60℃と50℃の差は10℃」「2010年の10年後は2020年」といった足し引きはできますが、「30℃の２倍は60℃」とは言えません。

- **比例尺度**・・・数値として表現され、足したり引いたり、比率を考えることができるような尺度です。
  - 例：身長、体重、金額
  - 「170cmと160cmの差は10cm」「60kgから5kg増えると65kg」「100円の２倍は200円」といった足し引きや比率にはすべて意味があります。

