１．データの代表値

・平均値（mean）…全ての変量をすべて足して、データの大きさで割ったもの。（$\bar{x}$ と表す）

例）ある生徒の学年末考査の点数が次のようなとき、平均値は？

国語→６３点　数学→５１点　地理→７８点　物理→３９点　英語→８５点

$\bar{x}=(63+51+78+39+85)\div 5=316\div 5=63.2$（点）または $\bar{x}=\frac{1}{5}(63+51+78+39+85)=\frac{1}{5}\times 316=63.2$（点）

平均値

変量 x のデータの値が $x_1, x_2, \cdots\cdots, x_n$ であるとき，このデータの平均値 $\bar{x}$ は

$$\bar{x}=\frac{1}{n}(x_1+x_2+\cdots\cdots+x_n)$$

$$(\text{平均値})=\frac{(\text{データの値の総和})}{(\text{データの大きさ(個数)})}$$

・最頻値（mode）…データの中で最も個数の多い値（度数が最も大きい階級値）。

例）下の度数分布表から最頻値を求めよ。

階級値（cm）	度数
１０	5
１４	3
１６	7
１８	1

最も大きい度数は　７
このときの階級値が最頻値なので、
「最頻値は１６」となる。

例）次のデータは 12 人の生徒のハンドボール投げの記録である。最頻値を求めよ。

15　20　13　17　18　21　18　22　15　18　16　17　(m)

データを小さい順に並べると

13　15　15　16　17　17　18　18　18　20　21　22　(m)

よって最頻値は 18m

・中央値（median）…データを値の大きさの順に並べたとき、中央に位置する値。

例）データの個数が奇数個のとき

ある商店の価格を５店舗で調査して，次のデータが得られた。

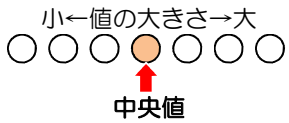
260，　100，　280，　300，　270　（円）

このデータを小さい順に並べると

100，　260，　270，　280，　300　（円）

よって，このデータの中央値は　270（円）

奇数個のとき



例）データの個数が偶数個のとき

８人の生徒の右手の握力を測って，次のデータが得られた。

38，56，43，41，35，49，51，31　（kg）

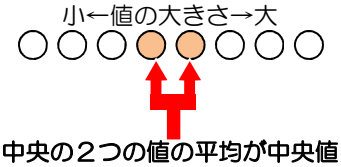
このデータを小さい順に並べると

31，35，38，41，43，49，51，56　（kg）

よって，このデータの中央値は

$\frac{41+43}{2}=\frac{84}{2}=42$　（kg）

偶数個のとき



2．四分位数と箱ひげ図

四分位数（しぶんいすう）とは

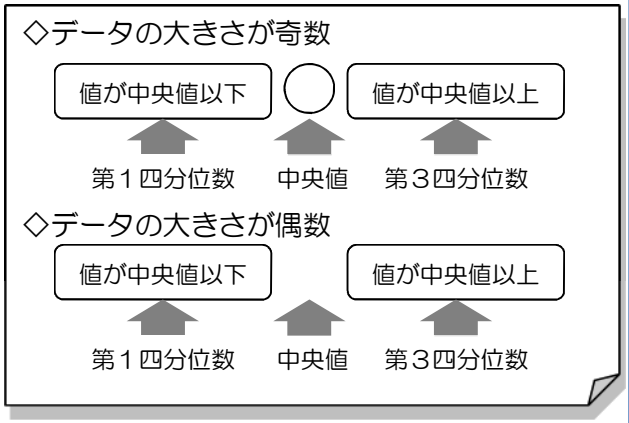
データの値を大きさの順に並べたとき、4等分する位置の値を四分位数という。四分位数は、小さい方から順に第1四分位数、第2四分位数(中央値と同じ値)、第3四分位数といい、順に Q_1、Q_2、Q_3 で表す。

四分位数の求め方

① データを値の大きさの順に並べ、中央値(第2四分位数)を求める。

② ①の中央値を境界としてデータの個数を2等分し、値が中央値以下の下組と値が中央値以上の上組に分ける。ただし、データの大きさが奇数のとき、①の中央値はどちらの組にも含めないものとする。

③ 下組の中央値(第1四分位数)、上組の中央値(第3四分位数)を求める。



四分位範囲とは

第3四分位数 Q_3 と第1四分位数 Q_1 の差 Q_3-Q_1 のこと。この中に中央値周辺に並ぶ約50％のデータが含まれる。よって、四分位範囲は、データの中に極端に飛び離れた値がある場合でも、その影響を受けにくい。また、データの値が中央値の周りに集中しているほど、四分位範囲は小さくなる傾向にある。逆に四分位範囲が大きいほど、データの散らばりが大きいと言える。

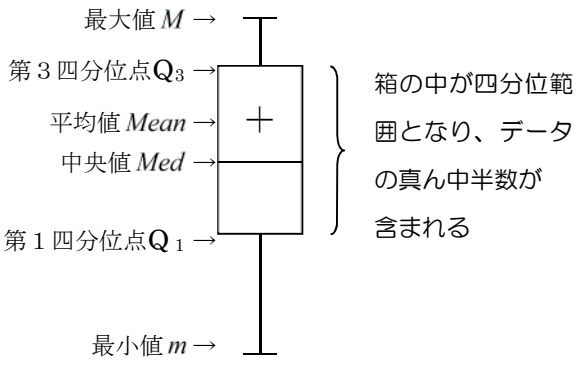
四分位数偏差とは

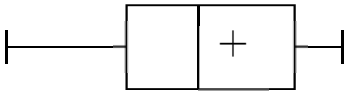
四分位範囲の半分のこと。

箱ひげ図とは

最小値、第１四分位数、中央値(＝第２四分位数)、第３四分位数、

最大値、平均値を「箱」と「線(髭)」を用いて図示したもの。





箱ひげ図は縦にでも横にでも表示することができる。

3. 標準偏差

偏差

データの各値と平均値 $\bar{x}$ との差のこと。$x-\bar{x}$ で表す。偏差の総和は 0 であるので、偏差の平均ももちろん 0。

分散

偏差 $x-\bar{x}$ の 2 乗の平均値のこと。式で表すと $s^2=\frac{1}{n}\left\{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots\cdots+(x_n-\bar{x})^2\right\}$

標準偏差

分散の正の平方根のこと。s で表す。要するに $s=\sqrt{\text{分散}}$ 。

標準偏差が小さくなるほどデータは平均値の周りに集中しており、散らばりの度合いが小さくなる。

逆に標準偏差が大きくなれば散らばりの度合いが大きいといえる(分散も同様である)。

	x	偏差 $x-\bar{x}$	$(x-\bar{x})^2$
α			
β			
γ			
δ			
ε			
ζ			
η			
θ			
ι			
κ			
平均値		分散	
		標準偏差	

分散と平均値の関係式

分散は次のような求め方もできる。

$$(x\text{のデータの分散})=(x^2\text{のデータの平均値})-(x\text{のデータの平均値})^2$$

ただしデータの値が小さくなければ大変なので注意。(上のデータなどでは $74^2=5476$ などとなり大変になる)

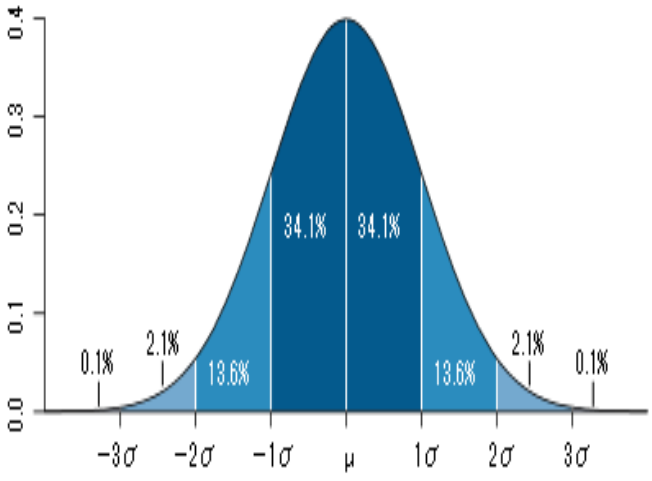
標準偏差と偏差値

四分位範囲(平均値±四分位偏差)にはデータの約 50%が含まれることになります。

一方、平均値 μ からのずれが ±標準偏差 σ 以下の範囲には 68.27%, ±2×標準偏差以下だと 95.45%, さらに ±3×標準偏差 だと 99.73% となります。このことを用いて模擬試験などでは偏差値として数値化し、全体(母集団)との位置関係を示しています。公式は次のようになり

$$(\text{偏差値})=\frac{10\times(\text{得点}-\text{平均点})}{\text{標準偏差}}+50$$

「+50」とするので偏差値 50 が集団の中央(平均点)となるのです。ただしあくまでも分布内の数値であり、確率に関わる数値なので目安として捉えましょう。



4．データの相関

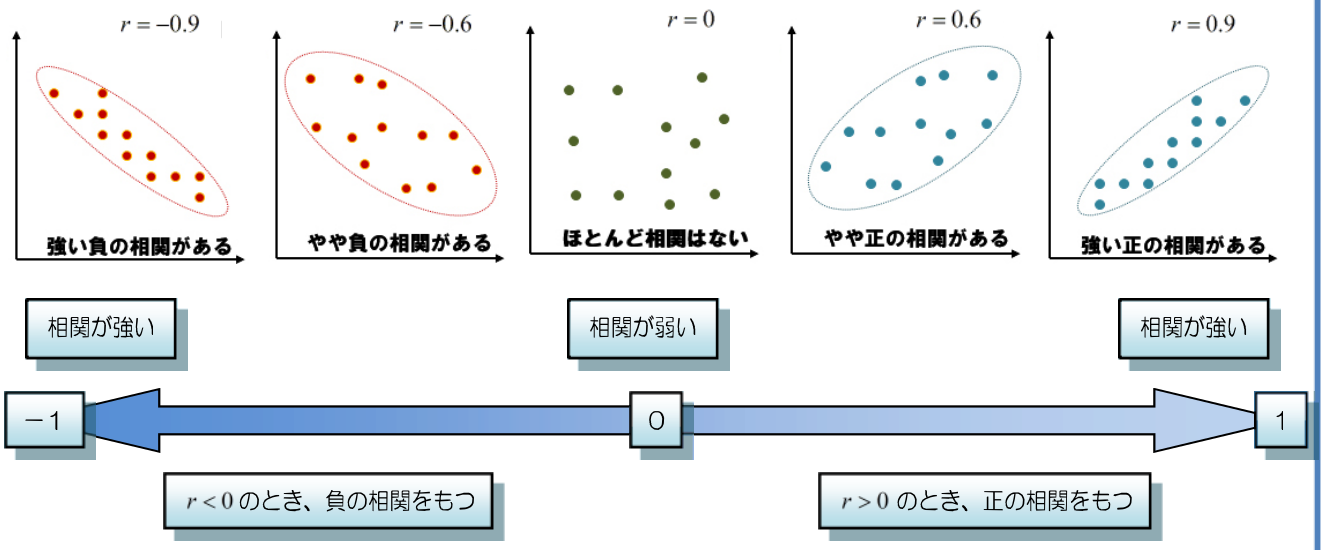
正の相関、負の相関

2つの変量からなるデータにおいて、一方が増加すると他方も増加する傾向が見られるとき、2つの変量には**正の相関**があるという。また、一方が増加すると他方は減少する傾向が見られるとき、2つの変量には**負の相関**があるという。どちらの傾向も見られないときには、**相関がない**または**相関関係がない**という。

2つの変量の間に相関があるとき、散布図における点の分布の様子が1つの直線に接近しているほど**相関が強い**といい、散らばっているほど**相関が弱い**という。

相関係数（r）とは

2変数間にどの程度直線的な関係があるかを数値で表す方法として、**相関係数**を調べる方法がある。



相関係数 r については、$-1 \leqq r \leqq 1$ であることが知られている。また、r が1に近いほど正の相関が強く、-1 に近いほど負の相関が強い。相関がないとき、r は0に近い値を取る。

【参考】

相関係数	相関関係
0	相関がない
0.0～±0.2	ほとんど相関がない
±0.2～±0.4	やや相関がある（低い相関）
±0.4～±0.7	相関がある
±0.7～±0.9	強い相関がある（高い相関）
±0.9～±1.0	きわめて強い相関がある
±1.0	完全な相関

◎相関係数を計算してみよう。

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
①	21	15	−5	-3	15	25	9
②	27	17	1	−1	−1	1	1
③	29	19	3	1	3	9	1
④	23	17	−3	−1	3	9	1
⑤	30	22	4	4	16	16	16
計	130	90	計		36 ～○	60 ～△	28 ～□
平均	26	18	平均		7.2	12	5.6

x と y の共分散　x の分散　y の分散

●計算方法１（合計の値で計算する）

$$r=\frac{○}{\sqrt{△\times□}}=\frac{36}{\sqrt{60\times28}}=\frac{36}{4\sqrt{105}}=0.8783\cdots\fallingdotseq0.88$$

●計算方法２（共分散や分散、標準偏差の値で計算する）

$$r=\frac{(\text{共分散})}{\sqrt{(x\text{の分散})\times(y\text{の分散})}}=\frac{(\text{共分散})}{(x\text{の標準偏差})\times(y\text{の標準偏差})}=\frac{7.2}{\sqrt{12}\times\sqrt{5.6}}=\frac{7.2}{\sqrt{67.2}}=0.8783\cdots\fallingdotseq0.88$$