ネコ仙人と学ぶ

データ分析

2012 年授業力向上研修資料

＜数学Ⅰデータの分析教材プリント＞

学校教育室　下町壽男

＜登場人物＞

ネコ仙人(ネ)

うかれぎつね(ウ)

マドンナうさぎ(マ)

元気ネコ(ゲ゛)

イヌざむらい(イ)

のんきぐま(ノ)

学者うさぎ(ガ゛)

がまんがえる(カ゛マ)

がんばミミズ(ミ)

データの分析の基本について，ネコ仙人とその仲間たちの会話で楽しく学んでいきますよ．頑張ってついてきてください

§１　資料の整理と分析

●統計を学ぶ意義

ネ：人間は一人ひとり個性を持っている．それは人間だけではない．すべてのものには個性がある．でも，そのような個性を持った物なり人なりが集団になったとき，そこに一つの規則や傾向が見えてくるんだ．統計とは，その傾向をどのように表現するかを学ぶものじゃ．

ウ：ふーん．

マ：トーケイって，英語で言うとどうなるの？

ネ：Statistics．State（国）からきているんだ．統治，つまり国を治めるには一人一人の個性を捨象して，全体の傾向をとらえる必要がある．そういうことからうまれた学問なのじゃ．

●代表値と散布度

ネ：今，５人からなるグループ A，B があったとする．数学の試験を行ったところ，以下の表のような結果になった．

A	点数
しもまっち	50
うかれぎつね	50
元気ネコ	45
のんきぐま	55
がまんがえる	50

B	点数
ネコ仙人	100
マドンナうさぎ	10
イヌざむらい	30
学者うさぎ	70
がんばミミズ	40

マ：ちょっと！なんで私が 10 点なの（泣）

ガ：まあ，妥当な例だね．

イ：僕はこんなものよ．

ネ：まあ，仮のはなしじゃ．

し：ネコ仙人さんだけ 100 点というのは・・・

ネ：話を極端にしたかったんじゃ．悪気はない．さあ，いいかい．グループ A とグループ B の平均点はどうなっているかい．

ノ：A は $\dfrac{50+50+45+55+50}{5}=50$

B は $\dfrac{100+10+30+70+40}{5}=50$

どっちも 50 点だね．

ウ：のんきぐま君すごい！

ミ：こんなの俺にもできるよ．

ネ：いいかい，２つのグループの平均はともに 50 点じゃ．この 50 とは，いわばそのグループを代表する値，つまり**「代表値」**の一つじゃ．

ガ：つまり，５人分を１つの値で代表させたってことだね．

ネ：そうじゃ．すると，この「平均」を見る限り，グループ A とグループ B は同じであると言える．しかしどうだい．２つのグループを見てどっちが個性的といえる？

イ：B が個性的だ．A は皆 50 点に近い点数なのに，B は 100 点もあれば 10 点の人もいる．

ネ：ということは，このグループを表す数値としては，代表値である**「平均値」**だけでは足りないということがわかる．そこで，統計（データ分析）では，平均などの「代表値」の他に，分布の散らばり具合を表す「**散布度**」という，もう一つのモノサシを準備するのじゃ．

ガ：「散布度」ってどんな数値なの？

ネ：まず，一番簡単なのは，**「範囲」（レンジ）**と呼ばれるもので，

範囲（レンジ）＝最大値－最小値

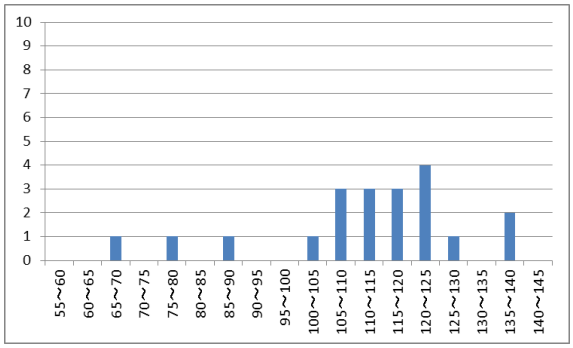
という式で表される．

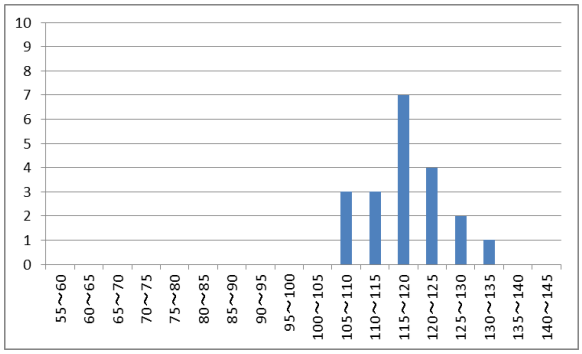
マ：A のレンジは，５５－４５＝１０，B は１００－１０＝９０ということで，大分差がついているわ．

ネ：ではここまでまとめよう．

> 集団の特性は，その集団を代表する数値**「代表値」**と分布の散らばり具合を表す**「散布度」**の２面から見る必要がある．

【§１の練習問題】（表現・意欲関心態度）





1998 年の長野オリンピックのスキージャンプ競技で日本チームは金メダルを取りました．

上の２つのヒストグラムは，原田選手と船木選手のオリンピックまでのいくつかの国際大会での距離の記録をまとめたものです（上：原田選手，下：船木選手）．このことから次のことを考えてみよう．

(1)　この２つのグラフそれぞれについてわかることをグループで話し合ってまとめてみよ．

(2)　もしあなたがジャンプ競技の監督で，原田選手，船木選手のどちらか１人を選ぶとしたら，どちらを選びますか．２人のヒストグラムの特徴を比較して説明しなさい．

参考：階級が，55～60 のように５mの幅があります．これは 55m以上 60m未満と考えてください．

また，平均などを考える場合は，階級の中央値の 57.5m を代表値として用いてください．ヒストグラム，階級，階級値，中央値などは中学校でも学びましたが，§２以降に解説します．尚，この問題は平成 24 年度全国学力調査の問題を参考にしています．

尚，(2)の問題はグループ同士でディベートさせてもよい．

§２　度数分布表と代表値

●度数分布表

ネ：前回は，集団の特性を見るには，平均値のような代表値と，分布のばらつきを見る散布度が必要だという話をした．

マ：前回の分布では，グループ A の散布度が 90 と大きかったから，A の方が B より優れたグループってことじゃないわよね．

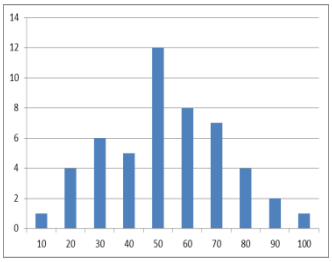
ネ：そうじゃ．例えば，グループ A は皆平均のまわりにいるので，「落ちこぼれがいないグループ」ともいえる．でも，もしこれが，運転免許取得の試験だったら，90 点以上が合格条件とすれば，グループ A からは 1 人の合格者も出ないことになる．一方，グループ B からは合格者が 1 人出ていると見ることができる．何を基準にするかで，考え方が変わってくるのじゃ．

さて，今度は次のように，たくさんの資料を整理してみよう．今，50 人が数学のテストを受けたとして，得点が次のようだったとする．

10,50,20,100,80,80,50,30,30,50,50,60,70,50,50,
50,30,20,70,40,60,40,40,50,70,20,90,30,40,50,
60,70,60,50,20,90,40,50,60,60,30,50,30,70,60,
70,70,80,80,60

このままでは，集団の特性を見極めることができない．そこで，次のような「度数分布表」にまとめることになるのじゃ．

階級	度数	相対度数
10	1	0.02
20	4	0.08
30	6	0.12
40	5	0.1
50	12	0.24
60	8	0.16
70	7	0.14
80	4	0.08
90	2	0.04
100	1	0.02
合計	50	1



10 点，20 点，･･･，100 点を**「階級」**といい，その階級に入っている個数を**「度数」**という．そして，その度数の生じる割合を**相対度数**というのじゃ．このような表を**「度数分布表」**といい，その結果を棒グラフにしたものを**「ヒストグラム」**というのじゃ．

これを見ると，一番頻度が多いのが 50 点であることがわかるし，分布の幅は１００－１０＝９０ということもわかる．

さて，ここで，代表値の３つの例をあげておこう．代表値は§１でも述べたように，このような**統計資料**の**変量**（個々の値）全体を代表する値のことじゃ．

●**中央値（メジアン）と最頻値（モード）**

ネ：平均値以外についての代表値として，**中央値**と**最頻値**がよく利用される．まとめておこう．

【中央値（メジアン）】

資料を大きさの順に並べたとき，中央にくる値のこと．もし資料が偶数個の場合は，中央の２つの値を平均したものを中央値とする．

例　23,35,46,48,56 のときは中央値は 46

23,34,56,67,69,70 のきは $\frac{56+67}{2}=61.5$

【最頻値（モード）】

最も度数の多い階級値のこと．前の度数分布表では 50 のところが最も多いので，モードは 50 となる．

ミ：モードやメジアンはどんなときに使うの？

ネ：例えば，次のような，洋服のサイズごとの売れ行きを度数分布表にしたものを考える．この平均をとってみると次のようになる．

サイズ	度数
6号	60
7号	20
8号	0
9号	100
10号	20

$$\frac{6\times60+7\times20+9\times100+10\times20}{200}=8$$

平均 8 号ということだ．

さて，ここで，洋服屋がこの平均を信じて，8 号サイズの服ばっかりつくったとするとどうなる？

マ：1 着も売れないわ！

ネ：そう．だからこういう場合は，代表値として，どのサイズが一番売れているかという「モード」に注目して考えるべきなんじゃよ．

マ：そっかあ．だから今年の秋のモードなんていうのね．

ネ：メジアンは，例えば，10 以上 20 未満などという，階級の幅があるとき，その階級値として採用する値をとるとき用いられる．

イ：つまり，階級が 10 以上 20 未満だったら階級値はそのメジアンの 15 ということですね．

ネ：もう一つ面白い例をあげよう．例えば，飛行機や，船に乗りながら，何かを観測して，連続的にデータを取ったとする．このとき，船の揺れや人為的な入力ミスなどで異常な値を取ってしまうことがある．今，次のように水温のデータを刻々と取っていたとしよう．

12,13,12,13,14,15,13,12,12,13,122,13,13,‥‥

このとき，11 番目のデータ 122 は明らかに入力ミスだ．しかし，人は見てわかるけれど，コンピュータが制御する場合など，間違えだと判定されずに処理される可能性がある．この 122 という値は**「外れ値」（異常値）**といわれる．この外れ値を取り除くために次のような工夫がある．

まず，最初の数列 12,13,12 からメジアンをとる．次に，1 つずらした 3 つの数列 13,12,13 のメジアンを取る．そうやって，次々データを 1 つずつずらして 3 数のメジアンをとって，新しい数列を作ると，12,13,13,14,14,13,12,13,13,13 と外れ値が取り除かれる．このような手法を**メジアンフィルタ**というのじゃ．

平均値は外れ値の影響を受けやすいが，中央値や最頻値は影響を受けにくい．外れ値の影響を受けにくい性質を**「抵抗性がある」**というのじゃ．

ネ：ではここまでをまとめよう．

【代表値とその性質】

平均値

◎すべてのデータを用いるので，データの持つ情報を有効に活用している．

◎世界中で代表値として普及している

○つねに 1 つだけ存在する

▼外れ値の影響を受けやすい

中央値

◎外れ値の影響を受けにくい

○つねに 1 つだけ存在する

▼データの並べ替えが必要である

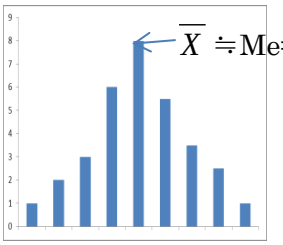
▼個々の数値は代表値に直接反映されないことが多い

最頻値

◎外れ値の影響を受けにくい

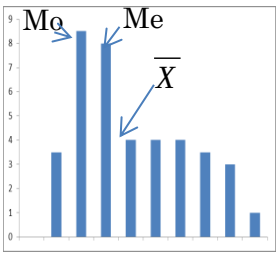
○最頻値は複数ある場合がある

【分布の形と代表値の関係】



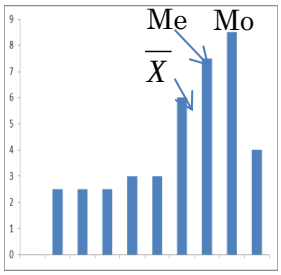
←ほぼ左右対称

平均値≒中央値≒最頻値



←左側に偏っている

最頻値＜中央値＜平均値



←右側に偏っている

平均値＜中央値＜最頻値

【§２の練習問題①】

以下は，あるクラスの生徒４０人の身長の測定値である．ただし単位はｃｍである．

172.1 165.8 165.3 153.3 158.2 167.8 175.2 148.7
155.3 168.4 173.7 151.3 163.4 173.4 154.4 159.4
176.6 158.5 161.3 164.4 175.8 169.4 152.7 168.4
176.6 172.3,167.9,158.7,166.8,154.2,177.5,156.2,
166.4,172.3,173.9,154.2,171.9,163.5,156.5,165.5

(1)　班ごとに，カードにデータを記入し，最大値，最小値，中央値，最頻値を求めよ．**（関心意欲態度）**

(2)　145cm から始まる階級の幅が 5cm の度数分布表を作り，ヒストグラムを作れ．

（知識理解・技能）

(3)　40 人のデータの平均値を求めよ．**（技能）**

（電卓等を利用して良い）

階級	階級値	度数
145～150		
150～155		
155～160		
160～165		
165～170		
170～175		
175～180		

●度数分布表から平均値を求める

ネ：では，度数分布表がつくれるようになったので，この度数分布表から平均値を出す練習をしよう．もちろん，個々の値は，ある幅のある階級に入っているので，正確な平均にはならないが，かなり近い値であることは期待できる．

一つ簡単な例をやってみよう．下図は 10 人のクラスで行った数学のテストの度数分布表である．

階級	階級値	度数	階級値×度数
10～20	15	2	
20～30	25	1	
30～40	35	3	
40～50	45	2	
50～60	55	1	
60～70	65	1	
計		10	

＜がまんがえる君の解答＞

階級	階級値	度数	階級値×度数
10～20	15	2	30
20～30	25	1	25
30～40	35	3	105
40～50	45	2	90
50～60	55	1	55
60～70	65	1	65
計		10	370

$$\text{平均}=\frac{370}{10}=37\text{点}$$

ガマ：計算が面倒だったけどなんとかできたよ．

ガ：だから，コンピュータを利用して「エクセル」などといった表計算ソフトをつかうんだ．

ネ：では，もう少し簡単な**「仮平均から平均を求める方法」**をやってみよう．

例えば，次の数の平均はどのように考えれば簡単になるだろう．

1002 , 1003 , 1005 , 1008 , 1007 , 1005

ミ：平均だから普通は

$$\frac{1002+1003+1005+1008+1007+1005}{6}$$

$$=\frac{6030}{6}=1005$$　だよね．

マ：待って．６つの数は全部千いくつだから，１桁目だけ考えれば早いわ．

$$\frac{2+3+5+8+7+5}{6}=5$$　だから平均は 1005

ゲ：僕は，1005 を基準にしたよ．データの左から 1005 に対して，３少ない，２少ない，３多い，２多い数があるので打ち消しあって０．だから平均は 1005 だ．

ネ：マドンナうさぎさんは，平均を仮に 1000 と置いて過不足分を考え，元気ネコ君は，平均を 1005 と睨んでその過不足を考えたわけだ．

２人とももう立派に「仮平均から平均を求める方法」を使っているね．

ネ：ではやってみよう．

階級	階級値	A:度数	B:仮平均からの差	A×B
10～20	15	2	−20	−40
20～30	25	1	−10	−10
30～40	35	3	**0**	0
40～50	45	2	10	20
50～60	55	1	20	20
60～70	65	1	30	30
計		10	30	20

まず，度数分布表を見て，平均を予想する．

35 のところの頻度が一番多い（モードですね）ので，だいたい 35 点と予想できる．すると，上の表のように，仮平均からの差という項目を作ってやる．今，仮平均を 35 としたので，35 のところは 0 になる．他の部分も埋めていく．そして，それぞれ，度数×（仮平均からの差）を求めるわけだ．

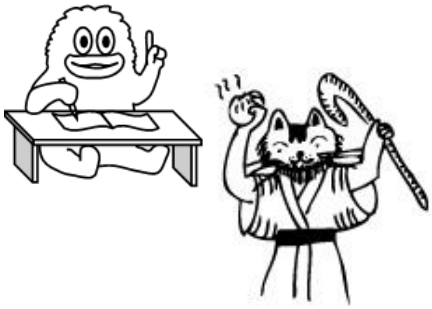
例えば，階級値が 15 のところの-40 という数字は，仮平均から 20 点マイナス（足をひっぱっている）の人が 2 人いるので，合計 40 点のマイナスポイントという意味．

さて，それを全部足し合わせると，＋20 という値になった．これが，10 人総合での仮平均からの上乗せ分なので，1 人に平均すると，$\frac{20}{10}=2$ となり，これを仮平均に足してやると平均が求まるのじゃ．

イ：平均＝３０＋２＝３７点．確かに平均になっている！

【§２の練習問題②】

前ページの§２の練習問題①の平均を，仮平均を用いた方法で求めよ．**（知識理解・技能）**



§３　データの散らばり（散布度）

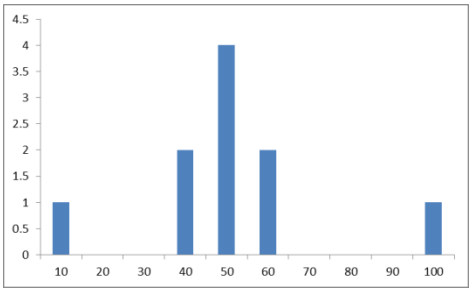
●四分位数と四分位偏差

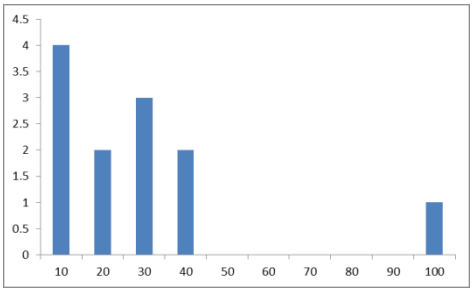
ネ：今回は，データの散らばりについて勉強していこう．

ガ：データの散らばりでは，以前「範囲」（レンジ）を学びましたね．

ミ：確か，最大値－最小値だった．

ネ：そう．でも次のような分布を見ると，明らかに集団の特性は違うけれど，どちらも範囲は同じ値 90 になっていて，ばらつきの違いを数値化できない．



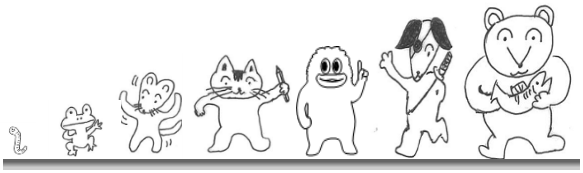


そこで，データ全体を小さい順に並べたものを 4 等分にして，その４つの場所ごとの範囲（レンジ）を調べることで，散らばり具合を明瞭にしていこうという作戦をとるのじゃ．

具体的に説明しよう．

うかれぎつね君，がまんガエル君，のんきぐま君，元気ネコ君，いぬざむらい君，しもまっち君，がんばミミズ君、一列に並んでくれたまえ．

ネ：並んだかい．では皆で背比べをして，小さい順に並び直してくれ．



さあ準備はいいかい．まずおさらいをしよう．

最大値は？

ノ：はい僕の身長です．

ミ：最小値は僕です．

ネ：そう．つまり範囲は「のんきぐま君の身長－がんばミミズ君の身長」だ．じゃあ，中央値は？

ゲ：はい僕です．

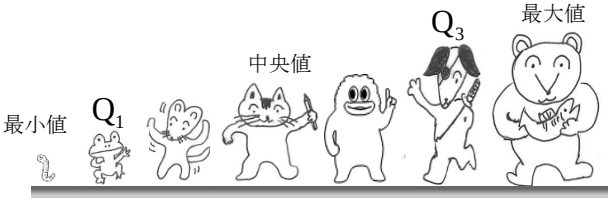
ネ：さあ，ここで，中央値の元気ネコ君より小さい人たちを**「下位のデータ」**，元気ネコ君より大きい人たちを**「上位のデータ」**という．

では，下位のデータ 3 人と，上位のデータ 3 人でそれぞれ中央値を考えてみよう．

ガマ：中央値は僕です．

イ：こっちのデータは僕が中央値です．

ネ：下位データの中央値のがまんガエル君の身長を，**第１四分位数**（Q_1），全体の中央値の元気ネコ君の身長を**第２四分位数**（Q_2），上位データの中央値である，いぬざむらい君の身長を**第３四分位数**（Q_3）というのじゃ．



ネ：ここで，第 3 四分位数から第 1 四分位数を引いたもの，Q_3-Q_1を**四分位範囲**というのじゃ．

この値は，データを大きさの順に並べたときの，中央の 50%のデータの範囲にほぼ等しい．

そして，この四分位範囲の半分を**四分位偏差**というのじゃ．仮に，7 人の身長を順に

10,20,50,60,70,80,120 としよう．

このとき，$Q_1=20$，$Q_2=60$，$Q_3=80$ なので

四分位範囲＝80－20＝60

四分位偏差＝60÷2＝30　となる．

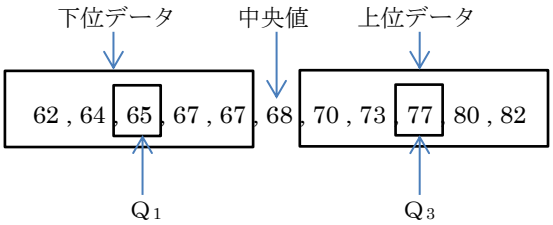
この数の意味は，中央値 60 の±30 の幅の中に，ほぼ 50%のデータが入っているということじゃ．

ガ：60±30 ってことは，データの範囲は 30～90．

すると，うかれぎつね，元気ネコ，しもまっち，いぬざむらいの 4 人が含まれているわけだね．

ネ：では，これから四分位数と四分位偏差を求める練習をしよう．

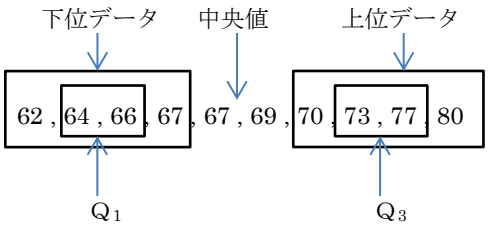
【$4n+3$型】



四分位偏差＝ $\frac{77-65}{2}=6$

68±6　つまり 62～74 の範囲にデータのほぼ 50%が含まれる．

【$4n+2$型】

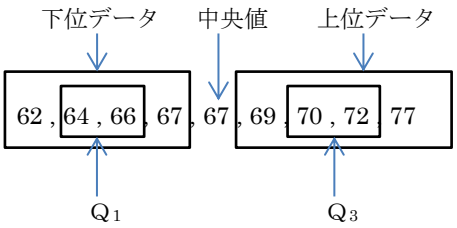


中央値＝ $\frac{67+69}{2}=68$　$Q_1=\frac{64+66}{2}=65$

$Q_3=\frac{73+77}{2}=75$

四分位範囲＝10　四分位偏差＝ 5

【$4n+1$型】

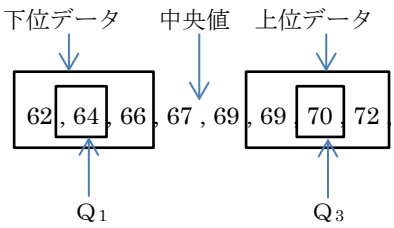


中央値＝67　　$Q_1=\dfrac{64+66}{2}=65$

$Q_3=\dfrac{70+72}{2}=71$

四分位範囲＝6　四分位偏差＝3

【$4n$型】



中央値＝68　　$Q_1=64$

$Q_3=70$

四分位範囲＝6　四分位偏差＝3

【四分位数】

データを小さい順に並べ，全体を四等分に分けることにより，分布の様子をしらべる．

データの中央で区切ったとき

Q_1：**第１四分位数**（下半分のデータの中央値）

Q_2：**第２四分位数**（全体データの中央値）

Q_3：**第３四分位数**（上半分データの中央値）

Q3－Q1：**四分位範囲**

$\dfrac{Q_3-Q_1}{2}$：**四分位偏差**

中央値±四分位偏差　の中に，データのほぼ50%がはいっている．

【§３の練習問題①】

クラスを２つのグループに分ける．仮にクラスの人数を 40 人とすると，20 人のグループが２つできる．

グループごとに，競争で次のことを行わせてみよう．

(1)　誕生日（1 月～12 月）の順に並ぶ
　　（効率的なソーティングの工夫）

(2)　最大値，最小値，中央値，第１四分位数，第３四分位数にあたる数値の生徒を決定する．

(3)　四分位偏差を調べ，中央値±四分位範囲の中に，何人の生徒がいるか調べる．

(1)関心意欲態度・数学的な見方考え方

(2)知識理解・技能

(3)関心意欲態度・知識理解

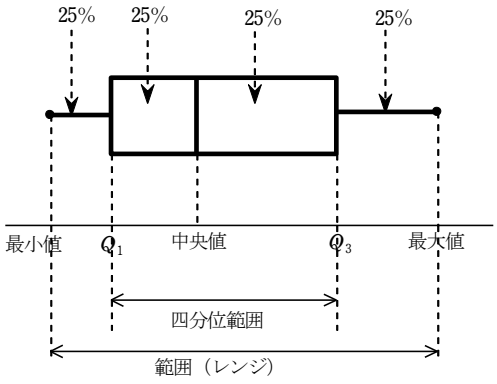
●箱ひげ図

ネ：では，ここまで学んできた，四分位数と，最大値，最小値を合わせて，分布の様子がわかるような図示を考える．**箱ひげ図**とよばれるものじゃ．

ウ：「箱ひげ」なんて，何か下品な名前だなあ．

ネ：分布の中央値から上下 25%のデータが，「箱舟」に入る．そこから外れた上下 25%が「ひげ」にあたるようなイメージじゃ．

だいたいこんな感じじゃ．



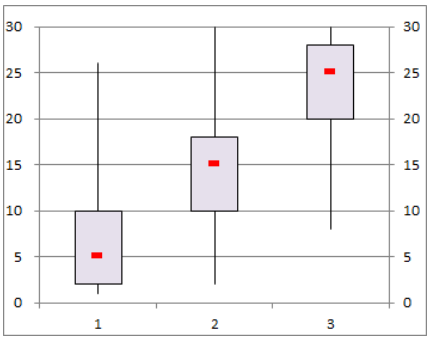
ゲ：この箱ひげ図を見て何がわかるんだろう．

ミ：分布の様子だったらヒストグラムの方がいいじゃん．

ネ：おっ．がんばミミズ君，いいことを言ったね．確かに，分布の様子を詳しく知りたいときはヒストグラムじゃな．

　でも、例えば，ある集団で，小テストを 3 度実施したとする．その時の成績の推移を，度数分布で追いかけるより，下のように箱ひげ図を並べると，直感的なイメージがわくじゃろう．

〈参考図①〉



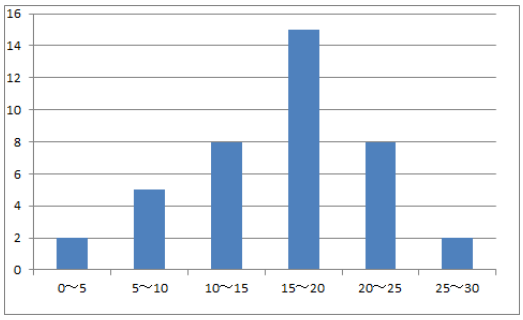
　特に，回数が多い場合ほど，ヒストグラムより箱ひげ図を並べて眺めてみるのも効果的じゃ．

マ：でも私は，まだ「箱ひげ図」のイメージができていないからよくわからないわ．

ネ：多分最初はそうじゃろう．では，いくつか典型的なヒストグラムと，それに対応する箱ひげ図を比べてみよう．

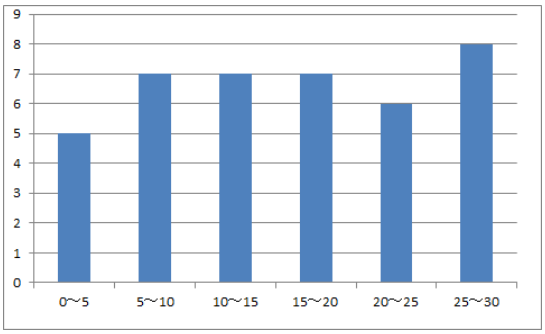
【①正規分布型】

平均値≒中央値≒最頻値型の，左右対称なグラフ．



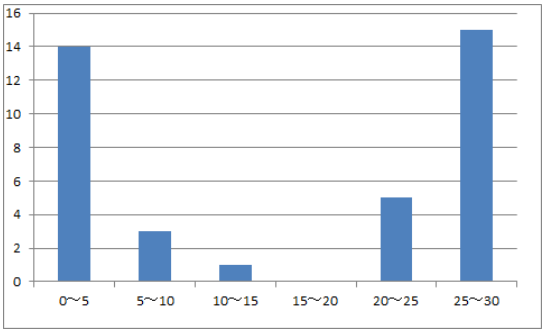
【②一様分布型】

どの階級も同じくらいの度数



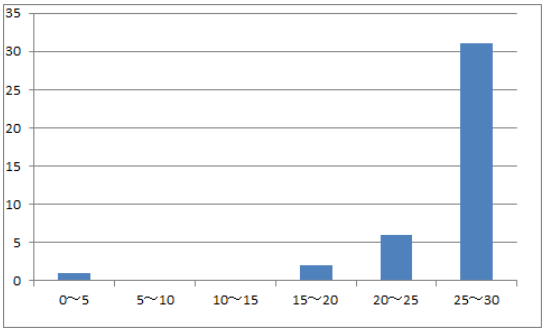
【③双峰分布型】

山が２つに分かれるタイプ

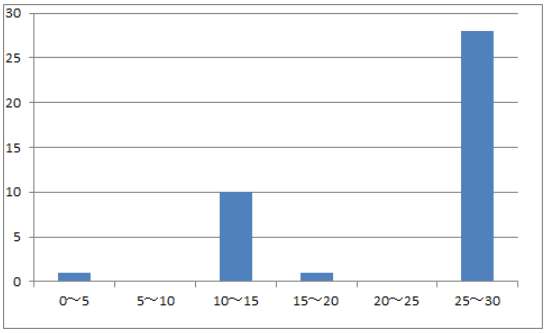


【④歪み型】

山が左右の一方に偏る（外れ値あり）

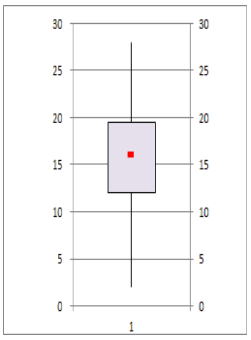


【⑤外れ値を含む双峰型】

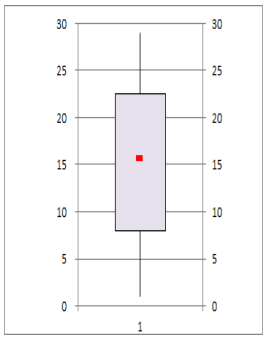


ネ：さあ，ここにあげた①～⑤の特徴的なヒストグラムを，次にあげる箱ひげ図と対応させられれば，君たちも箱ひげ図のプロじゃ！

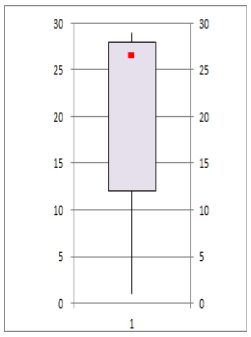
Ⅰ



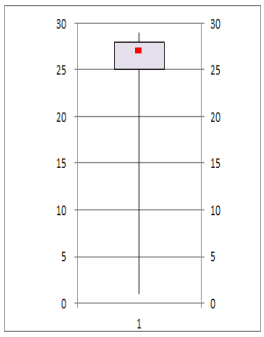
Ⅱ



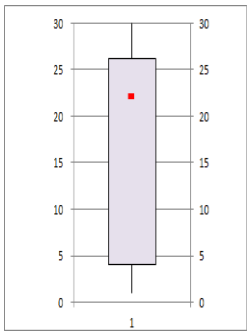
Ⅲ



Ⅳ



Ⅴ



ガ：わかった．図Ⅱは，箱ひげの４つの部分が同じ長さだから，どの階級も同じくらいの度数と考えられる．だから②に対応する．

ノ：なるほど．すると，Ⅰは中央値が分布の中間にあるから，中央で山になっているやつだ．

ガマ：わかった，これは①だ．

ゲ：あとは，特徴的なのはⅣだけど・・・

マ：一番上のひげ（第３四分位点から最大値）が短いわね．ということは・・・そこの度数が少ないってことかしら．

イ：違うよ，箱，ひげ，の４つの部分は全部 25%で同じ人数なんだよ！

マ：あっそうね．すると，上の方に分布が固まっているっていうことね．

ウ：④か⑥だ．どっちだろう．

ガ：④の方が，上に固まっている．だから，箱もつぶれている．つまり，上位 75%は 25 点以上ってことだ．だから④が正解だ！

ミ：最後のひげが長いのは，１個だけ外れたところにあるからだ．

ゲ：じゃあ残りはあと２つ．ⅢとⅤだけど・・・

マ：Ⅴは両方のひげが短い．だから，上と下にいっぱい集まっているってことだわ．だから③よ．

ガ：すると，Ⅲは⑤ってことだな．

ネ：よくやった．もう皆は箱ひげ図と友達じゃ．

【§３の練習問題②】（関心意欲態度）

前頁の＜参考図①＞を見て，集団がどのように推移していったか論ぜよ．

（ディベートをさせてもよい）

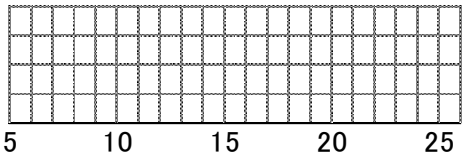
【§３の練習問題③】（知識理解・技能）

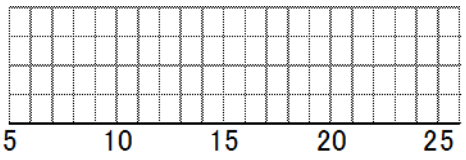
次のデータは 10 人の生徒の教科 AB の得点である．

A　9,15,17,17,18,18,20,20,22,24

B　8,8,9,9,12,12,13,13,21,22

四分位数等を求め A，B の箱ひげ図をかけ．





（箱ひげ図から AB の比較をさせてもよい）

§４　分散と標準偏差

ネ：箱ひげ図はある意味便利だけれど，イメージ先行だった．そこで，今回は，統計資料を，もう少し説得力ある分析を行う指標について学ぶ．

統計資料は「代表値」と「散布度」を組み合わせて表現するという話だった．

イ：代表値には「平均値」「中央値」「最頻値」がありました．

ネ：そう．じゃあ，散布度にはどんなものがあった？

マ：散布度は分布の「散らばり具合」を表す数だったわよね．確か「範囲」がそうだった．

ガ：四分位範囲もある意味散布度だよね．

ネ：そうじゃ．この「散らばり具合」というのをもっと具体的に言うと，「それぞれの変量について，平均からの隔たり具合」ということになる．

それを総合したものとして**「分散」**という重要な散布度を学んでいこう．

ネ：４人からなるＡ，Ｂ２つのグループについて身長を調べた表を見てみよう．

Ａ

ナナ	ヤス	ノブ	シン
162	185	167	164

Ｂ

レイラ	レン	タクミ	ナオキ
164	182	183	178

まず，それぞれの平均を求めてみよう．

イ：Ａの方は，$\dfrac{162+185+167+164}{4}=169.5$

Ｂは，$\dfrac{164+182+183+178}{4}=176.75$　です．

ネ：さて，平均はわかったけれど，ここで，どちらのグループがより「散らばり方」が大きいか考えてみよう．

まず，おのおのの値と，平均との差（偏差）を調べてみよう．

名前	身長	平均からの差
ナナ	162	−7.5
ヤス	185	15.5
ノブ	167	−2.5
シン	164	−5.5

名前	身長	平均からの差
レイラ	164	−12.75
レン	182	5.25
タクミ	183	6.25
ナオキ	178	1.25

マ：面白いわ．Ａはヤス一人がプラスで他の皆はマイナス，Ｂは逆に，レイラが一人マイナスなのね．

ネ：では，このグループの身長について，どちらが平均に対して，より散らばっていると見ることができるだろうか．

ノ：全員の，平均からの散らばり具合を全部足してみればいいんじゃないか．

ネ：それはとても自然な考えだ．つまり，

グループの散らばり度

＝個々の平均からの隔たりの総和

という考えじゃ．だがしかし，そうすると困ったことになる．あったりまえのことなのじゃが，足せばどちらもゼロになってしまうのじゃ．

ノ：そっかあ．ということは，平均からの隔たりが，マイナスであっても，プラスであっても隔たっている量は同じだから，マイナスをプラスに変えて足せばいいじゃん．

ネ：そう．これを，偏差の総和という．つまり，それぞれの平均からの隔たりの絶対値を足していくということじゃ．今，変量を x_k，平均を $\overline{x}$ とすると，偏差の総和は次の式で書ける．

$$\sum_{k=1}^{n}\left|x_k-\overline{x}\right|$$

ミ：ゲッ．Σって何？

ネ：これは，$\left|x_1-\overline{x}\right|+\left|x_2-\overline{x}\right|+\cdots+\left|x_n-\overline{x}\right|$

という意味じゃ．Σはシグマ記号と言って，数学Ｂで数列を習う時に登場する記号じゃが，今覚えておいてもいいじゃろう．

これを，n で割れば，一人分に見なおした値となる．これを**平均偏差**という．

ガ：では計算してみよう．まずＡの方は

$|162-169.5|+|185-169.5|+|167-169.5|+|164-169.5|$

$=31$

１人分にした平均偏差は４で割って　7.75

Ｂの方は

$|164-176.75|+|182-176.75|+\cdots+|178-176.75|$

$=25.5$

平均偏差は，$25.5\div4=6.375$

つまり，Ａの方がバラツキが大きいんだ．

ネ：今のように，個々について，平均からの隔たりを出し，その総和を個数で割ったものを平均偏差という．しかし，この散布度はあまり一般的ではないのじゃ．

統計を数学的に考える場合，絶対値のような処理はあまり応用がきかないのじゃ．後で微分積分というものを習うが，そのとき絶対値は不便なのだ．

そこで，登場するのが「分散」という考え方じゃ．

分散は，平均からの差の絶対値を考える代わりに，平均からの差を２乗するということじゃ．そうすることによるメリットは２つある．

一つは，２乗すれば，平均からの差がマイナスのときもプラスに変えることができること．もう一つは，平均からの隔たりをより大きくして，特徴を際立たせるということだ．

　では，分散を式で定義しておこう．

分散＝ $\dfrac{1}{N}\{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_N-\bar{x})\}$

先ほどのΣの記号を使えば

分散＝ $\dfrac{1}{N}\sum_{k=1}^{N}(x_k-\bar{x})^2$　ともかける．

つまり，分散とは，**「平均からの偏差の２乗の平均」**ということになる．記号で書けば，V(X)とか，s^2,σ^2 などで表される．σはΣの小文字でシグマと読む．で，この分散の平方根，s を**標準偏差**という．標準偏差は，いわば散布度の世界標準じゃ！

ネ：では次の問題をやってみよう．

変量 x が次の 10 個の値を取るとき，平均値と分散を求めよ．5,1,4,3,4,5,5,1,2,5

ウ：ちゃんと度数分布表を作って計算しよう．

X	①度数	②仮平均からの差	①×②
1	2	−2	−4
2	1	−1	−1
3	1	0	0
4	2	1	2
5	4	2	8
		合計	5

仮平均を３にしておくと，仮平均からの差の合計が５だから，平均は $3+\dfrac{5}{10}=3.5$

ゲ：じゃあ分散を計算するよ．やっぱり，表を使った方がわかりやすいね．

X	①度数	②平均からの差の2乗	①×②
1	2	6.25	12.5
2	1	2.25	2.25
3	1	0.25	0.25
4	2	0.25	0.5
5	4	2.25	9
		合計	24.5

分散は，$24.5\div10$ なので，2.45　だ！

ガマ：標準偏差は $\sqrt{2.45}$ なので，電卓で計算すると約 1.5652 だ．

ガ：この標準偏差はどういう意味を持つのだろう．

ネ：前回，箱ひげ図と四分位数の話のとき，

中央値±四分位偏差　の範囲にデータの 50%が入るという話をしたけれど，同じように，

平均値±標準偏差　の範囲に注目しようということじゃ．

マ：平均値±標準偏差の中にどのくらいのデータが入っているの？

ネ：それは，データがどんな分布になっているかによるが，平均値，中央値，最頻値が同じ値のような分布（**正規分布**）のときは，ほぼ 70%が入る

のじゃ．平均と分散をセットにして，様々な統計資料を分析することができるのじゃよ．そういう話は，統計を勉強すると深く学ぶことになる．

今は，標準偏差が大きいほど，散らばり具合が大きく，小さいほど，データは平均のまわりに集中するという傾向をまずは押さえておこう．

●偏差値って何だ

ネ：ここでちょっと余談．偏差値って言葉は聞いたことがあるかい？

ノ：あるある．でもそれってどうやって出すの．

マ：私，数学の模試で０点だったのに，偏差値が30もついてきて嬉しかったわ．

ネ：では，少し寄り道して，偏差値の計算の仕方について教えよう．

ある集団でテストを行ったとき，得点を X とし，それに対応する偏差値を Z とする．全体の平均点を $\overline{X}$，標準偏差を σ とおけば

$Z=\dfrac{X-\overline{X}}{\sigma}\times 10+50$　と表される．これが偏差値だ．

この式の意味を考えよう．

①　$X-\overline{X}$　‥‥平均点を０点にする

②　$\dfrac{X-\overline{X}}{\sigma}$ ‥‥標準偏差を１にする

③　$\dfrac{X-\overline{X}}{\sigma}\times 10$ ‥‥標準偏差を10にする

④　$\dfrac{X-\overline{X}}{\sigma}\times 10+50$ ‥‥平均を50にする

つまり，得点の分布を，平均 50，標準偏差 10 になるように平行移動したり，拡大縮小したものにすぎんのじゃ．このような処理を日本中で行えば，他と比較して自分の位置がどのくらいのところにあるかわかる一つの指標になるというわけじゃが，所詮数字のマジック．こんな値に振り回されることなんかないのじゃよ．

●分散を簡単に求める

ネ：さて，ここで，分散を早く求めるうまい方法を伝授しよう．

ちょっとΣを使った複雑な計算をするが，がまんしくれい．まあここは読みとばしても良い．

平均を m とすると，分散はΣを用いて

$$s^2=\frac{1}{N}\sum_{k=1}^{N}(x_k-m)^2$$

これを次のように変型する．

$$s^2=\frac{1}{N}\sum_{k=1}^{N}({x_k}^2-2mx_k+m^2)$$

$$=\frac{1}{N}\sum_{k=1}^{N}{x_k}^2-2m\cdot\frac{1}{N}\sum_{k=1}^{N}x_k+\frac{1}{N}\sum_{k=1}^{N}m^2$$

ここで，

$$\frac{1}{N}\sum_{k=1}^{N}x_k=\frac{x_1+x_2+\cdots+x_N}{N}=m$$

$$\frac{1}{N}\sum_{k=1}^{N}m^2=\frac{m^2+m^2+\cdots+m^2}{N}=m^2$$

なので，$s^2=\dfrac{1}{N}\sum_{k=1}^{N}{x_k}^2-m^2$　※

つまり，分散とは次のように考えることができる．

分散＝（2乗平均）－(平均)^2

では，これを用いてさっきの問題をやってみよう．

X	X^2	度数	$X^2\times$度数
1	1	2	2
2	4	1	4
3	9	1	9
4	16	2	32
5	25	4	100
		合計	147

上の度数分布表から，２乗平均は，14.7

平均の２乗は，$3.5^2=12.25$

分散は　$14.7-12.25=2.45$　一致した！

【§４の練習問題】（知識理解・技能）

次の変量 x, y について，分散と標準偏差を求めよ．

x	2	3	4	7	8	8	9	9	10	10
y	5	5	5	7	7	7	7	8	9	10

§５　データの相関

●散布図

ネ：これまで，統計データの特徴を調べるために，ヒストグラムや箱ひげ図などを学んできた．また，代表値である平均値，中央値，最頻値，そして，散布度の代表として，分散と標準偏差の話をしてきた．今回は最後のテーマ，**「相関」**について説明していこう．

ミ：相関って何？

ネ：今までは，ある集団のテストの点数とか，身長とか，１つの変量にについて考えてきた．これから考えるのは，２つの変量どうしの関係について考えるんだ．例えば，次の表を見てもらいたい．

	身長	体重
大崎ナナ	162	43
小松奈々	158	46
本城蓮	182	64
高木泰士	185	72
寺島伸夫	167	52
岡崎真一	164	50
芹澤レイラ	164	48
一ノ瀬巧	183	67
藤枝直樹	178	65
上原美里	153	42
早乙女淳子	168	53
高倉京助	180	70
遠藤章司	175	60
川村幸子	146	35

（ＮＡＮＡ／矢沢あい　より引用しました）

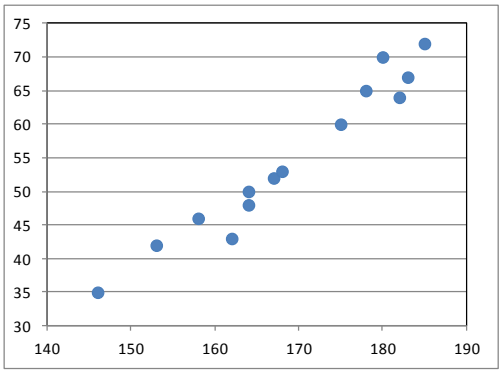
これは，14 人の身長と体重を組にして表したものだ．身長を変量 x，体重を変量 y としたとき，x と y の大小に関係性があるかどうかを調べるための一つのモノサシが**「相関係数」**だ．

まず，相関係数を説明する前に，**散布図**（相関図）について説明しよう．

今，この２つのデータの関係性を見易くするために，横軸に身長，縦軸に体重をとって，一人一人の場所をプロットしてみる．

この図のことを「散布図」という．

さあ，この散布図を見てどんなことがわかるかな．



ノ：14 人を点で表すと，直線的に並んでいる．

ウ：つまりそれは，身長が高いと体重も重いということなんだ．

イ：ときどき例外もあるけどね．

ネ：このように，散布図にすれば，関係性が目につく．このように，右上がりの直線に点が並ぶとき２つの変量には**「強い正の相関がある」**といったりする．

ガ：ってことは，右下がりなら**「負の相関」**ということですね．

ネ：そうじゃ．そのような例はどんなものがあると思う？

ゲ：例えば「年齢」と「美しさ」ってのはどう．年齢が増えれば美しさが減るよ．

ネ：それはどうかな．年齢というのは変量としてきちんと数で定義できるからいいとしても，「美しさ」というのはどうやって数値に表すんだい．それに，「美しさ」というのは主観的なもので，人によって基準が違う．

マ：それに，年をとるほどに美しさを増すってものもあるわ．

ネ：話をもどそう．相関関係とは，２つの変量の関係のことだが，もう少し具体的に言うと，次のようにまとめることができる．

【相関関係】

> ２つの変量について，一方の大小と他方の大小が関係しているどうかを示すもの．

ゲ：さっきの例でいえば，「身長が高ければ体重も重くなる」という関係のことですか．

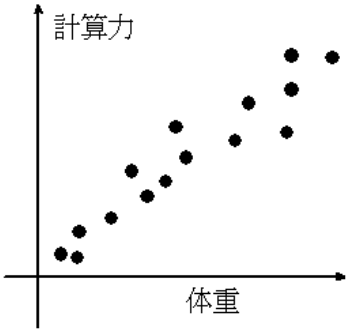
ネ：それは少し違う．確かに，体重は身長の 3 乗に比例しているなんていう説を信じれば，一方が変化するのにともなってもう一方が変化するという関数関係と見ることができるかもしれない．

しかし，普通，相関関係を考えるということは，あくまで２つの変量どうしが似ているかどうかをみるものにすぎないんだ．

だから，相関関係を因果関係（一方が他方の原因になっている）と勘違いしないように気をつけて欲しい．

ミ：なんか難しいなあ．

ネ：例えば，下のデータは，「算数の計算力」と「体重」について散布度に表したものだ．これを見ると，強い正の相関がみられる．



ミ：ええっ．体重が重い方が頭がいいんだ！

ガ：それは違うよ．つまり体重が重いってことは，それだけ年齢が高いってことさ．体重が 10kg だったら，赤ん坊だから当然計算なんかできない．

ネ：その通り．だから「体重」と「計算力」は相関関係はあるけれど，因果関係にはなっていないんだ．どちらの変量にも「年齢」という因子が関わっているので相関が高くなっているということ．

【§５の練習問題①】（関心意欲態度）

身の回りのデータを調べ，変量どうしの関係性を調べレポートせよ．（グループで行ってよい）

これは**「課題学習」**とする．

＜参考＞生徒の作品例

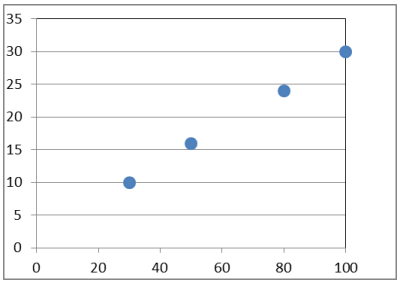
（県のグラフ統計コンクールで特選受賞）

●相関係数

ネ：では，相関係数の話に進む．まあ，公式とかそういうことではなく，気軽に考えてくれ．イメージができればいいんだ．

今度は次のようなデータを考えてみよう．

	テスト1	テスト2
学者うさぎ	100	30
マドンナうさぎ	80	24
元気ネコ	50	16
うかれぎつね	30	10
平均	65	20

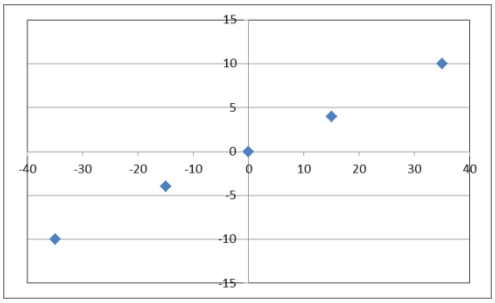


＜テスト１・テスト２散布図＞

テスト１の平均は 65 で，テスト２の平均は 20 だ．平均は違うけれど，正の相関が強いことは，散布図からわかる．では次に，テスト１のデータと，テスト２のデータを比較するかわり，変量から平均を引いたものどうしを比較してみよう．

	テスト1	テスト2	テスト3 テスト1－65	テスト4 テスト2－20
学者うさぎ	100	30	35	10
マドンナうさぎ	80	24	15	4
元気ネコ	50	16	−15	−4
うかれぎつね	30	10	−35	−10
平均	65	20	0	0

つまり $(100,80,50,30)$ と，$(30,24,16,10)$ の相関を調べるかわりに $(35,15,-15,-35)$ と $(10,4,-4,-10)$ の相関を調べるということだ．

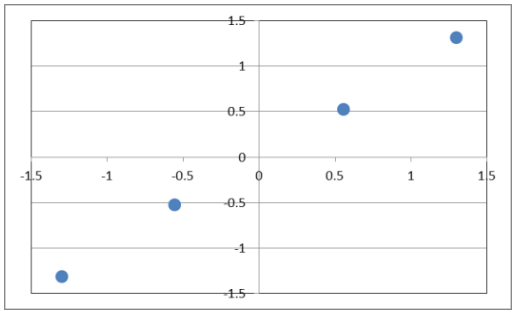


＜テスト３・テスト４散布図＞

図のように，前のグラフを単に平行移動したグラフになるので，相関の強さには変化がなさそうだ．

ところで，テスト３とテスト４の標準偏差を計算すると，26.9 と 7.6 となる．ここで，テスト３とテスト４のデータを，26.9 と 7.6 で割ったデータの散布図を考えてみる．

テスト5	テスト6
1.30	1.31
0.56	0.53
−0.56	−0.53
−1.30	−1.31



＜テスト５・テスト６散布図＞

どうじゃ，やはり相関に変化はなさそうじゃ．

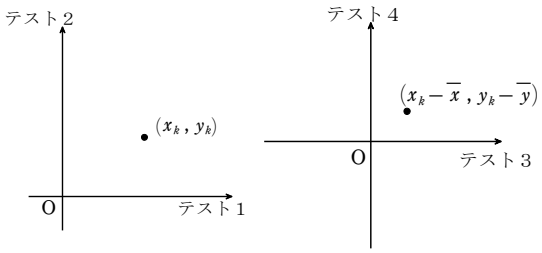
テスト５とテスト６は，テスト１，テスト２をそれぞれ「平均０，標準偏差１」に平行移動し，拡大縮小したものじゃ！

ガ：これって，偏差値のときの考えと同じだ！

ネ：そうじゃ．まとめておこう．

> 変量 x, y の相関を考えることは，各変量を平均０，標準偏差１になるように標準化したデータの相関を考えることと同じである．

ネ：では，これまで述べてきたテスト１とテスト２の相関の具合について，数量化する式を考えてみよう．

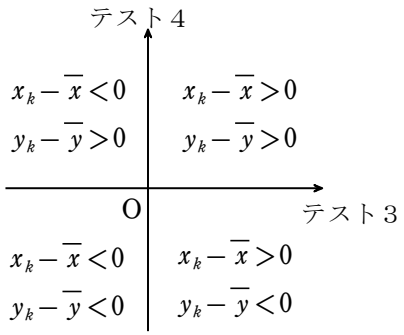


テスト１とテスト２に関するデータは，上の図の点 (x_k, y_k) で表される．

一方，テスト３とテスト４のデータは，平均点分それぞれ平行移動しているので，

$(x_k-\overline{x}, y_k-\overline{y})$ という点に対応している．

このとき，次のことに注意しよう．



上図で，第 1 象限は２つの変量がともに平均を超えている領域，第 2 象限はテスト３の平均は越えず，テスト４の平均は超えている領域，第 3 象限は，２つの変量がともに平均を下回っている領域，第 4 象限は，テスト３の平均は越え，テスト

４の平均が下回っている領域ということになる．

このように，散布図は，２つの変量の平均のところで線を引いて，４分割して考えると分析しやすいんだ．

ここで，あるデータ (x_k, y_k) に対し，平均との差の積　$(x_k-\overline{x})(y_k-\overline{y})$ を考えると，

データが第 1･第 3 象限にあるときは積は正になり，第 2・第 4 象限にあるときは積は負になる．

第 1・第 3 象限にデータがあるということは，正の相関が強いときであり，逆に，データが第 2・第 4 象限にあるときは負の相関が強いときだ．

そこで，すべてのデータに対して，平均との差の積を加えて，その値が大きいほど，データが第 1・第 3 象限にたくさんある，つまり正の相関が強いということがいえる．この考えを利用すると，相関の強さを数量で表すことができるのじゃ．

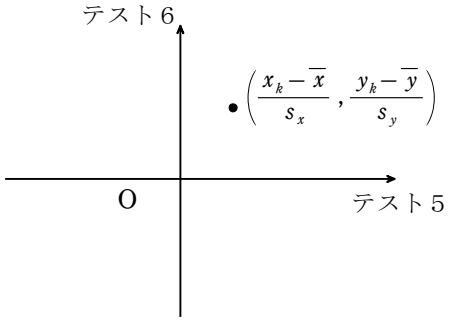
今述べた，平均との差の積の平均

$$\frac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_1-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})\}$$

を**共分散**といって，s_{xy} という文字で表す．

$s_{xy}>0$ ⇒①③にデータがある⇒正の相関が強い
$s_{xy}<0$ ⇒②④にデータがある⇒負の相関が強い

さて，ではこの共分散を，テスト５とテスト６で考えてみよう．



$$\frac{1}{n}\cdot\frac{1}{s_x s_y}\cdot\{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_1-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})\}$$

なんと，$\frac{s_{xy}}{s_x s_y}$　というシンプルな式になった．

この値を**「相関係数」**といい，r で表す．

このように表すことによって，相関の強さを－1 から 1 までの数で標準化できるので，イメージがしやすくなるのじゃ．

ガ：とっても難しいことがわかりました．

ネ：あとで，数学 B でベクトルを学ぶと，相関係数を「内積」という別のアプローチで考えることもできる．それはまたのお楽しみじゃ．

では，一つ簡単な例で相関係数を求めてみよう．

【§５の練習問題】（知識理解・技能）

英語	10	4	7	6	8
国語	6	2	7	4	6

上の表は，ある生徒 5 人の英語と国語の小テストの結果である．英語，国語の得点をそれぞれ変量 x, y として，相関係数 r を求めよ．

x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$
10	6					
4	2					
7	7					
6	4					
8	6					
/	/					

上の表を活用して考えてみよ．

＜ちょっと補足＞ベクトルで説明

ネ：相関係数は，$r=\dfrac{s_{xy}}{s_x s_y}$ という式だったが，

$$s_x=\sqrt{\frac{1}{n}\{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2\}}$$

$$s_y=\sqrt{\frac{1}{n}\{(y_1-\overline{y})^2+(y_2-\overline{y})^2+\cdots+(y_n-\overline{y})^2\}}$$

なので，相関係数は

$$r=\frac{\frac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_1-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})\}}{\sqrt{\frac{1}{n}\{(x_1-\overline{x})^2+\cdots+(x_n-\overline{x})^2\}}\sqrt{\frac{1}{n}\{(y_1-\overline{y})^2+\cdots+(y_n-\overline{y})^2\}}}$$

という恐ろしい式になる．

上の式を更に変形すると

$$r=\frac{\{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_1-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})\}}{\sqrt{\{(x_1-\overline{x})^2+\cdots+(x_n-\overline{x})^2\}}\sqrt{\{(y_1-\overline{y})^2+\cdots+(y_n-\overline{y})^2\}}}$$

とできる．(※)

$$\vec{a}=(x_1-\overline{x}\,,\,x_2-\overline{x}\,,\,x_3-\overline{x}\,,\,\cdots\cdots,\,x_n-\overline{x})$$

$$\vec{b}=(y_1-\overline{y}\,,\,y_2-\overline{y}\,,\,y_3-\overline{y}\,,\,\cdots\cdots,\,y_n-\overline{y})$$

というベクトルを考えると，※式の分子は，２つのベクトルの内積，分母は，２つのベクトルの大きさの積を表しているので，

$$r=\frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|}$$

つまり，２つのベクトルのなす角を θ としたとき，相関係数とは２つのベクトルの $\cos\theta$ を表すものである．

コサインは，２つベクトルの「近さ加減」を表す量なので，相関係数と考えても納得がいく．

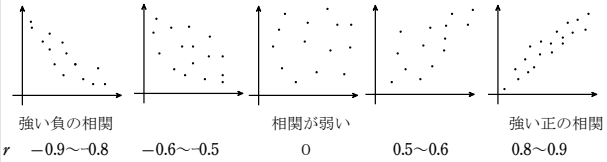
ネ：相関係数はわかったかい．

皆：式が難しかった．

ネ：相関係数の算出はコンピュータ等にやらせればよい．問題は，相関係数の見方じゃ．

散布図と相関係数の関係を図に表しておこう．

これを基に，データの相関の強弱を判断できればよい．



以上でデータ分析の話はおしまいじゃ．これまで学んだことが，君たちが社会に出てから役に立ってくれることを祈っているよ．