1. 分布を把握する（度数分布表・ヒストグラム）

本章の目標

- 度数分布やヒストグラムの必要性やその方法を理解する
- 度数分布やヒストグラムを用いて，分布の様子を調べることができる
- 相対度数や累積相対度数を用いて，異なるグループの分布を比較することができる

Key Words: 階級・度数・相対度数・度数分布・ヒストグラム

1. 度数分布表

- (量的)変数(例：世帯人員数)がとる値の範囲をグループ分けしたそれぞれの区間を**階級**という．
- 階級に含まれる観測値の個数をその階級の**度数**(頻度)といい，階級ごとに度数を整理したものを**度数分布**，その表を**度数分布表**という．
- 各階級の値を代表する値を級の**代表値**または**階級値**とよぶ．
- **相対度数**は，各階級の度数の全体に対する割合をあらわし，

$$\text{階級の度数}/\text{度数の合計}$$

で与えられる．相対度数は，観測値の個数(データに大きさ)が異なる複数のグループの比較を行うときに使われる．度数または相対度数を小さい階級から合計して得られる**累積(相対)度数**も同様に計算できる．

- 例題

次のデータは，某大学のある年度の 50 人の学生の統計学・期末試験(100 点満点)の点数結果である(点数自身の値は小さい順に並べてある):

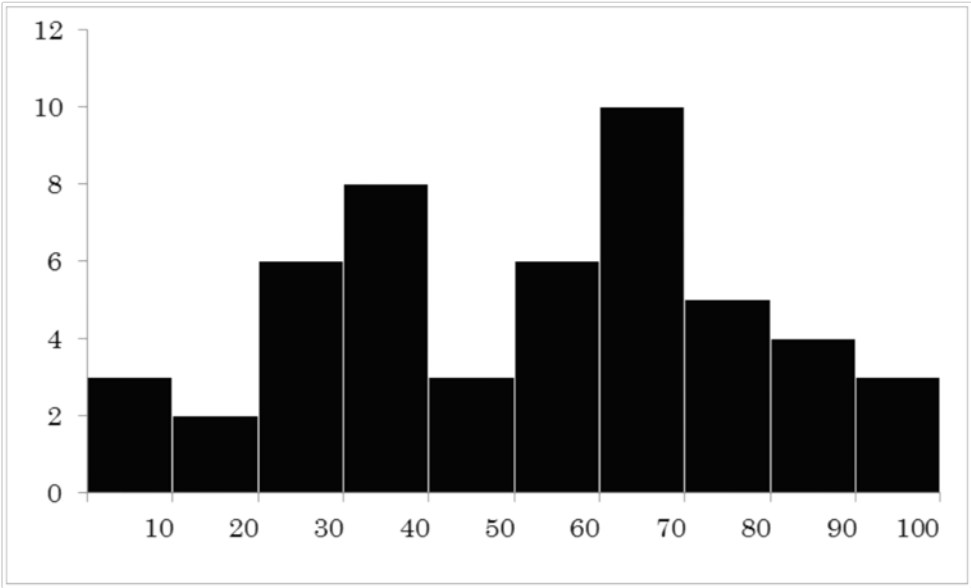
4, 8, 9, 10, 17, 21, 21, 25, 26, 28,29, 31, 33, 33, 36, 36, 36, 37, 39, 42, 43, 44, 51, 51, 53, 54, 58, 59, 61,61, 62, 62, 62, 65, 67, 67, 68, 69, 75, 75, 76, 77, 78, 81, 83, 85, 86, 93, 94, 99.

これに対する度数分布表は次のようになる：

階級 (以上) ~ (未満)	階級値	度数	相対度数	累積度数	累積相対度数
0 ~ 10	5	3	(3/50 =) 0.06	3	0.06
10 ~ 20	15	2	(2/50 =) 0.04	5	0.10
20 ~ 30	25	6	(6/50 =) 0.12	11	0.22
30 ~ 40	35	8	(8/50 =) 0.16	19	0.38
40 ~ 50	45	3	(3/50 =) 0.06	22	0.44
50 ~ 60	55	6	(6/50 =) 0.12	28	0.56
60 ~ 70	65	10	(10/50 =) 0.20	38	0.76
70 ~ 80	75	5	(5/50 =) 0.10	43	0.86
80 ~ 90	85	4	(4/50 =) 0.08	47	0.94
90 ~ 100	95	3	(3/50 =) 0.06	50	1.00
計		50	1.00		

2. ヒストグラム

- 度数分布をグラフ化する方法のひとつにヒストグラムがある．ヒストグラムでは，横軸に変数の値をとり，それぞれの階級の区間上に**面積が度数と比例する**ように長方形を描く．区間の幅が同じときには，長方形の高さは度数に比例する．
- 上記のデータのヒストグラムは次のようになる：



3. まとめ—分布の特徴の把握—

- ヒストグラムを描く目的は，量的変数の分布の特徴を把握すること
- 分布の中心はどのあたりか，散らばりはどの程度の大きさか，全体として左右対称かあるいはどちらの裾が長い分布か，等の特徴を知ることができる．
- これらのことは，ヒストグラムの形状により，代表値がそれのどこにあらわれるか，ということとも関連している．

練習問題

問 1（度数分布表とヒストグラムの解釈）

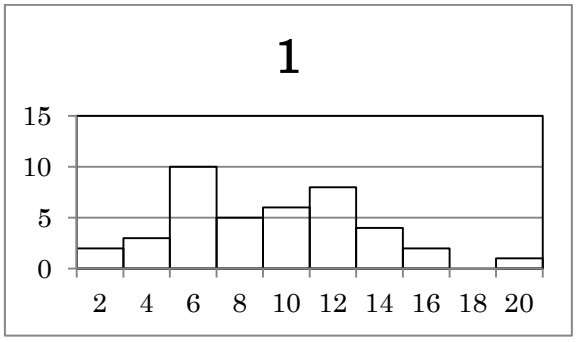
あるクラスで通学時間を調べたところ，次のような度数分布表が得られた．

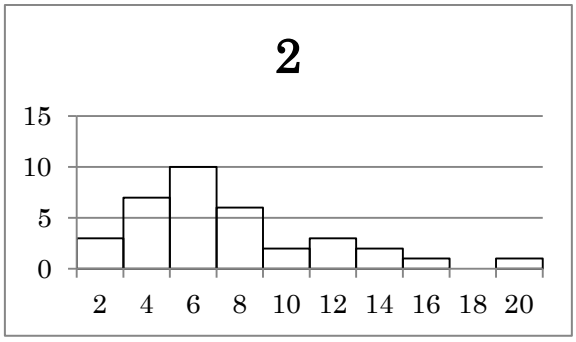
通学時間(分) (以上) ~ (未満)	度数
0 ~ 2	3
2 ~ 4	7
4 ~ 6	10
6 ~ 8	6
8 ~ 10	2
10 ~ 12	3
12 ~ 14	2
14 ~ 16	1
16 ~ 18	0
18 ~ 20	1
計	35

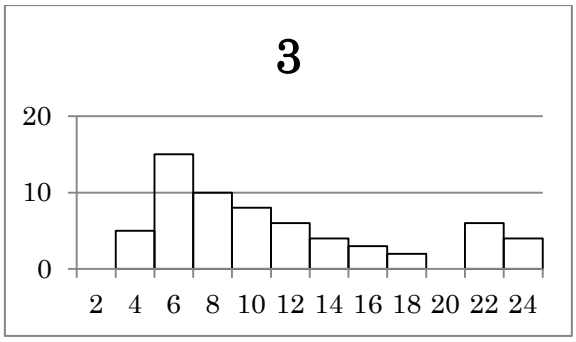
(1) この分布からわかることとして，適切ではない記述を次の 1〜5 のうちから一つ選べ．

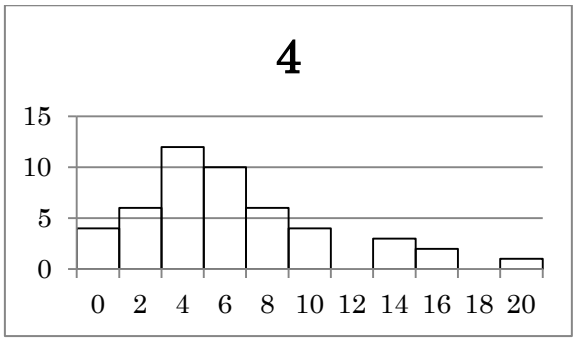
1. 最も度数の高い階級は，4〜6(分)である．
2. 通学時間が 10(分)以上の生徒は 7 人である．
3. 2〜4(分)の階級の相対度数は 0.2 である．
4. 通学時間が 2〜8(分)の生徒の割合は，約 66%である．
5. 半分以上の生徒は，通学時間は 5(分)以下である．

(2)　この度数分布表を使って描かれたヒストグラムとして適切なものを次の 1～4 のうちから一つ選べ.









練習問題の解答

問 1（度数分布表とヒストグラムの解釈）

(1) 解答：**5**

最も度数の多い階級は，4 分以上 6 分未満であるから 1 は正しい．通学時間が 10 分以上の生徒の人数は，10 分以上 12 分未満の階級から下の部分を合計したものであるので 7 人となり 2 は正しい．2 分以上 4 分未満の階級は 7 人，相対度数は 7/35 = 0.2 であり 3 は正しい．通学時間が 2 分以上 8 未満の生徒は 23 人おり，全体の 66%であり，4 は正しい．通学時間が 4 分未満の生徒は 10 人いることは分かるが，5 分以下の生徒の人数は，この度数分布表からは確定できないため，5 は適切ではない．

(2) 解答：**2**

度数分布表の階級幅と度数から対応するヒストグラムを選ぶと 2 が正しい．

2. 分布の中心を知る（代表値）

本章の目標

- 分布の位置をあらわす代表値の意味とその必要性を理解する
- 3 つの代表値の特徴を理解し，適切に用いることができる
- 代表値を用いて分布の様子を説明できる

Key Words: 平均値(mean)・中央値(median)・最頻値(mode)

0. 3 つの代表値

- 量的変数の分布を調べる際には，度数分布表やヒストグラムにあらわすことによって，全体的な特徴をつかむことができた．ここでは，分布の中心的な位置を 1 つの数字のみで代表させることを考える．
- 以下では，分布の中心的な傾向をあらわす値のうち，最も広く用いられている代表値，すなわち，平均値(mean)・中央値(median)・最頻値(mode)を扱う．

1. 平均値

- 平均(値)(mean)は，広く用いられる位置の代表値で，それは，

観測値の合計/観測値の個数

で与えられる．

- 平均は，比較的意味を捉えやすく，(量的変数として)計算も容易である．分布の中心の位置の代表値として用いられることが多い．
- 分布が単峰(ひと山)でほぼ左右対称であるとき，平均は分布の中心の最も観測値の個数が多い位置をあらわす．
- 極端に大きな観測値・小さな観測値(はずれ値)が含まれていると，平均はその影響を強く受け，代表性の解釈には注意が必要となる．

2. 中央値

- 分布の中心をあらわすために，大きさの順に並べ変えたときに真ん中に位置する観測値の値を中央値(中位数・median)という
- 中央値の計算の仕方は、ケース数 Nによって異なる
  ケース数が奇数の場合：$(N+1)/2$ 番目の値
  ケース数が偶数の場合：($N/2$ 番目の値＋$(N/2+1)$番目の値）÷2
- 中央値は，はずれ値の有無にほとんど影響されないという点で，平均とは異なる性質をもつ．

- 例題（ケース数が偶数のときの中央値）

  次は，ある数学のテストを 10 人に対して行った結果である．このテストの中央値を求めよ．

  80, 52, 35, 23, 93, 71, 18, 88, 47, 64

  （解答）

  はじめに点数が小さい順に次のようにデータを並べかえる．

  18, 23, 35, 47, 52, 64, 71, 80, 88, 93

  ケース数は 10 なので、N=10．したがって、中央値は以下のようになる．

  中央値＝（(10/2)番目の値＋(10/2)+1 番目の値）÷2
  ＝（5 番目の値＋6 番目の値）÷2
  ＝（52＋64）÷2　＝58

3. 最頻値

- 最頻値(mode)は，最も頻繁に出現する値を意味している．世帯人員数のように離散変数の場合にはその定義は明確であるが，エネルギー量のような連続変数の場合には同じ値をとることは少ないため，度数分布表を作成し，最も度数の大きな階級の代表値を最頻値とすることが多い．
- データによっては度数の大きな階級が二つ以上出現することがあり，このような場合には最頻値が明確な意味をもたないことがある

4. 平均・中央値・最頻値と単峰なヒストグラムの関係

- 分布が単峰(ひと山)でほぼ左右対称であるとき，**平均・中央値・最頻値は比較的近い値をとる**．
- 所得分布のように単峰で右の裾が長い分布では，**最頻値<中央値<平均**の順になる傾向がある．
- 単峰で左の裾が長い分布では，**平均<中央値<最頻値**の順になる傾向がある．

練習問題

問 1（代表値の計算）

次は，10 人の学生が与えられた時間内に仕上げた課題数を調べたデータである．

5, 5, 5, 10, 10, 10, 10, 15, 20, 50 (単位:題)

このデータに関する記述として，誤っているものを次の 1〜4 のうちから一つ選べ.

1. 中央値は 15(題)である．
2. 平均は 14(題)である．
3. 最頻値は 10(題)である．
4. 最大値は 50(題)である．

問 2（代表値の性質）

代表値の特徴に関する記述として，適切ではないものを次の 1〜4 のうちから一つ選べ.

1. 最大値よりも大きな観測値を一つ加えると，中央値は大きくなる．
2. 最大値よりも大きな観測値を一つ加えると，平均は大きくなる．
3. 左右対称で単峰(ひと山)分布に対して，平均・中央値・最頻値はいずれも近い値となる．
4. データによっては、最頻値が 2 つ以上存在することがある

問 3（代表値の計算 2）

あるクラスで先月のボランティア活動の時間を調べたところ，次のような度数分布表が得られた．この度数分布表からわかることとして，適切でないものを次の 1〜4 のうちから一つ選べ.

時間 (以上) ~ (未満)	度数
0 ~ 2	10
2 ~ 4	16
4 ~ 6	5
6 ~ 8	3
8 ~ 10	1
計	35

1. 中央値は，2 時間以上 4 時間未満である．
2. 最頻値は 3 時間である．
3. この度数分布表から計算される平均は，約 3.2 時間である．
4. 個々の時間から求めた平均は，1.2 時間以上 4.2 時間未満である．

練習問題の解答

問 1（代表値の計算）

解答：**1**

偶数個のデータの中央値を求めると，(10 + 10)/2 = 10 となり，1 が誤りである．

実際，平均は(5+5+5+10+10+10+10+15+20+50)/10 = 14，最頻値は 10，最大値は 50 となり正しい．

問 2（代表値の性質）

解答：**1**

1，2 に関しては，代表値の外れ値に対する依存性を問うている．2, 3, 4 は事実そのもの故，1 が適切ではない．

問 3（代表値の計算 2）

解答：**4**

35 人の真ん中の人数は 18 番目であるから，2 時間以上 4 時間未満の階級にあるので，1 は適切である．最も度数の多い階級は 2 時間以上 4 時間未満なので，最頻値は 3 時間であり，2 は適切である．各階級の値を階級値で置き換えて平均を計算すると，113/35 = 3.23 となり，3 は適切である．各階級ですべて最小の値を取る場合には，平均値よりも約 2 時間小さくなり，最大の値をとる場合には，約 1 時間大きくなるので，個々の時間から求めた平均は，2.2 時間以上 4.2 時間未満となり，適切ではないものは 4 である．

3. 分布の散らばりの指標（分散と標準偏差）

本章の目標

- 個々の観測値の散らばりの程度の概念を理解する
- データの散らばりの程度を数量的に求め，分布の把握やグループを比較することができる

Key Words: 偏差・平均偏差・分散・標準偏差・変動係数

1. 観測値の散らばりの指標

- データの散らばりの程度を数値化する指標を与える．
- まず，各観測値の散らばりを考えるために観測値からデータの平均を引いた差を考える．この値を**偏差**とよぶ：

$$\text{偏差} = \text{観測値} - \text{平均値}.$$

- データの全体の散らばりを考える場合は，偏差の絶対値の平均値，または偏差を平方した値の平均値を考える．特に，前者は，平均偏差，後者は，分散といわれる．

$$\text{平均偏差} = |\text{観測値} - \text{平均値}| \text{の平均値}$$
$$\text{分散} = (\text{観測値} - \text{平均値})^2 \text{の平均値}$$

- 分散の単位は観測値の平方で，平均とは単位が異なって解釈が難しい．そこで，分散の平方根をとり，それを標準偏差とよぶ．

$$\text{標準偏差} = \sqrt{\text{分散}}$$

- **例題**（那覇と札幌の気温のちらばり）

日付	那覇	札幌	那覇の偏差	札幌の偏差	那覇の偏差の絶対値	札幌の偏差の絶対値	那覇の偏差の平方	札幌の偏差の平方
1	29.2	19.4	0.00	0	0	0	0.00	0.09
2	28.7	18.5	-0.50	-1	0.5	1	0.25	1.44
3	26.3	17.1	-2.90	-3	2.9	3	8.41	6.76
・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・
31	30.2	22.0	1.00	2	1	2	1.00	5.29
平均	29.2	19.7	0.00	0.00	0.60	2	0.81	3.31

(解答)

上記の表の最下行と定義から,那覇地区の分散は 0.81,札幌地区の分散は 3.31 である.

これから標準偏差を計算すると那覇地区は，その $\sqrt{0.81}=0.90$ 度，札幌地区は，

$\sqrt{3.31}=1.82$ 度となる．平均偏差は，那覇地区は 0.6 度，札幌地区は 1.6 度である．一部の観測値は見えないが，指標のみで考えると，3 つの指標とも那覇のデータの方が小さく，札幌と比べて気温の散らばりが小さいと考えられる.

2. 変動係数で散らばりを考える

- 散らばりの程度を考える際に平均値の大きさを考慮しないと誤った解釈をする恐れがある．例えば，ある企業の従業者の年収を考えた際に管理職の年収の標準偏差が 450 万円，平均値は 2 千万円，アルバイト・フリーターの年収の標準偏差は 30 万円，平均値は 100 万円とする．このとき，管理職の年収の標準偏差の方がはるかに大きいが，解釈として「管理職の年収のばらつきはアルバイトより大きい」と考えるのは適切とはいえない．管理職の平均年収はアルバイトの 20 倍なのに，標準偏差は 15 倍であるから，ばらつきはかえって小さいと考えることもできる.
- このようなときは，標準偏差を平均値で割った**変動係数**とよばれる値を用いることがある(単位は%であらわすことが多い). この例では,管理職の変動係数は 450/2000 = 0.225，即ち 22.5%．アルバイト・フリーターの変動係数は 30/90 = 0.333，即ち約 33.3%であり，平均値に対するばらつきの程度はアルバイト・フリーターの方が大きいことがわかる．このように散らばりの程度として変動係数を用いることが適切な場合がある.

- **例題**（小学校の登校時間）

  ある地区の小学生の登校時間は平均 10 分，標準偏差 5 分であった．同じ地区の中学生の登校時間は平均 20 分，標準偏差 10 分であった．それぞれの変動係数を求め，それぞれの散らばりの程度を比較せよ.

  (解答)

  このデータでは，それぞれの登校時間を測定しており，標準偏差は 2 倍の違いがある．ただし，平均値が大きく異なるため，変動係数を求めると小学校は 5/10 = 0.50，即ち，50%．中学校は，10/20 = 0.50，即ち，50%となる．このことから平均値の大きさに対しては小学校と中学校で同程度のばらつきであることがわかる.

3. 平均値と分散・標準偏差の性質

- 全員のテストの得点に 5 点を加える場合のように､各ケースの値に a を足す(引く)と、平均値と分散・標準偏差は次のように変化する.

平均値　：もとの平均値＋a
分散　　：変化せず
標準偏差：変化せず

- 全員のテストの得点を 10%増やす（つまり、1.1 倍にする）場合のように、各ケースの値を b 倍すると、平均値と分散・標準偏差は次のように変化する.

平均値　：もとの平均値×b
分散　　：もとの分散×b^2
標準偏差：もとの標準偏差×b

練習問題

問 1（標準偏差の計算）

あるクラスで期末試験の得点から次のような表を得た.

学生	点数	偏差	偏差平方
1	82	13.1	171.61
2	91	22.1	488.41
3	38	-30.9	954.81
・・・	・・・	・・・	・・・
20	69	0.1	0.01
合計	1378	0	5929.80
平均	68.9	0	296.49

このクラスの得点の標準偏差はいくらか. 次の 1～4 のうち最も適切なものを一つ選べ.

1. 5929.80
2. 296.49
3. $\sqrt{296.49} \approx 17.22$
4. この情報のみでは求められない

問 2（代表値と分散にもとづく判断）
つぎの二つのデータはそれぞれ大きさの順に並べてある．a と b で等しいものはどれか．次の 1〜4 のうちから最も適切なものを一つ選べ．

a: 12, 14, 17, 23, 25, 34, 38, 39, 42, 52, 56, 58, 59, 64
b: 27, 29, 32, 38, 40, 49, 53, 54, 57, 67, 71, 73, 74, 79

1. 平均値
2. 中央値
3. 分散
4. すべて異なっている

問 3（平均値と標準偏差の性質）
あるクラスで呼んだ本の冊数を調査したところ，平均 2 冊，標準偏差 1.2 冊であった．その後，入力ミスが見つかり，各人が読んだ本の冊数は，本当はそれぞれ 10 倍の数値であることがわかった(即ち，2 冊と入力された人は，本当は 20 冊読んでいた)．このとき，本当の冊数での平均値と標準偏差の正しい組み合わせを次の 1〜4 のうちから一つ選べ．

1. 平均値: 2 (冊)，標準偏差: 1.2 (冊)
2. 平均値: 2 (冊)，標準偏差: 12 (冊)
3. 平均値: 20 (冊)，標準偏差: 1.2 (冊)
4. 平均値: 20 (冊)，標準偏差: 12 (冊)

練習問題の解答

問 1（標準偏差の計算）

解答：**3**

分散は，定義より，偏差の 2 乗の平均であることから，表の情報より 296.49 であることがわかる．したがって，標準偏差は分散の正の平方根であるから 3 が正解.

問 2（代表値と分散にもとづく判断）

解答：**3**

a の各観測値に 15 を加えたものが b のデータであることが分かる．従って，平均値，中央値は変わるが分散は変わらないので，3 が正解.

問 3（平均値と標準偏差の性質）

解答：**4**

平均値や標準偏差の性質より，各観測値が 10 倍になると平均値，標準偏差共に 10 倍になるため，4 が正解.

4. データをグラフで表現する

本章の目的

- さまざまなグラフの特徴を理解する
- 分析の目的に合わせて，適切にグラフを用いることができる
- 複雑なグラフを解釈することができる
- 箱ヒゲ図の見方とヒストグラムとの関連

Key Words: 幹葉図(幹葉図)・レーダーチャート・積み上げ棒グラフ・箱ヒゲ図

0. グラフ作成の目的

- 統計的な調査を実施すると，数多くの数値が得られるが，この数字のみを眺めていても全体の特徴をつかむことは難しい．データを集計したり，グラフを用いて表現したりすることは，データの中から必要な情報を取り出すための工夫である．
- グラフは，統計データが示す意味を理解したり，説明したりするための有効な手段であるが，データのもつさまざまな特徴の中からある種の特徴に焦点を当てて表現するため，目的に応じてさまざまな統計グラフが存在する．そのため，グラフの特徴を把握し，分析の目的に応じて適切に選択する必要がある．

1. 代表的なグラフの特徴

- 棒グラフ

量の大小を比較する際に用いられるグラフで，棒の高さにより，それぞれのカテゴリの量をあらわしている

- 円グラフ

それぞれのカテゴリの全体に対する割合を表す際に用いられる

- 帯グラフ

円グラフと同様に，全体に対する割合を表すグラフであるが，特に複数のグループや年次的な変化を調べる際に有効である

- 折れ線グラフ

量の時間的な変化の状況を示す際に用いられる

本章では，特に，幹葉図・レーダーチャート・箱ヒゲ図について解説する．

2. 幹葉図

- 幹葉図(幹葉表示)ともいう.
- 幹葉図はデータの大きさ(サンプルサイズ)n が比較的小さい場合に用いられるグラフ表現で，数値データのばらつきを表す際に用いられる.
- 幹葉図の例

次のような，ある数学のテストの 20 人分の成績を考える:

49	71	64	93	80	66	79	58	68	69
80	54	74	75	78	86	85	65	73	86

- この数値のみを見て特徴を見出すことは難しいが，下図のように表すことで，数値のバラツキの様子を把握することができる

```
4|9
5|4 8
6|4 5 6 8 9
7|1 3 4 5 8 9
8|0 0 5 6 6
9|3
```

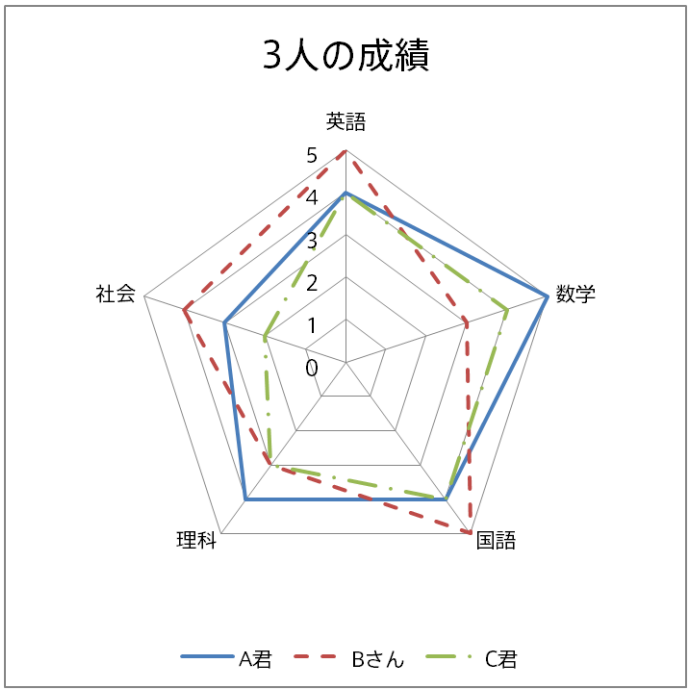
- このグラフ表現では，左側の幹の部分に 10 の位の数値を表示し，葉の部分には観測値の 1 の位の値を右に並べている．コンピューターによる出力では 1 の位の数値は小さい方から順に並べられるが，手書きで作成する場合は観測値が出現する順に記入していく.
- このグラフでは，60 点台，70 点台，80 点台の数値が多く見られ，40 点台，50 点台，90 点台は少ないことがわかると同時に，具体的数値も把握することができる.
- サンプルサイズ n が小さいときには，手描きでも簡単にできるグラフ表現である．ただし，n が大きいときには，複雑になりすぎる．列車の時刻表もある意味で幹葉図と同じような形で構成されている.

3. レーダーチャート

- レーダーチャートは，正多角形上に配置された複数の項目・属性の値をプロットであり，隣り合う値を線で結んで作成されるグラフ．クモの巣グラフとも呼ばれる.
- 同一の項目・属性であれば，複数のレーダーチャートを 1 つのグラフにまとめて表しても良い．それぞれの測定単位は異なっても良い.
- 複数の値をまとめて表現する際に用いられるグラフ表現である．下図は，ある学生の 5 科目の成績をあらわしている．このグラフを見ることで，教科のバランスが判断できる.

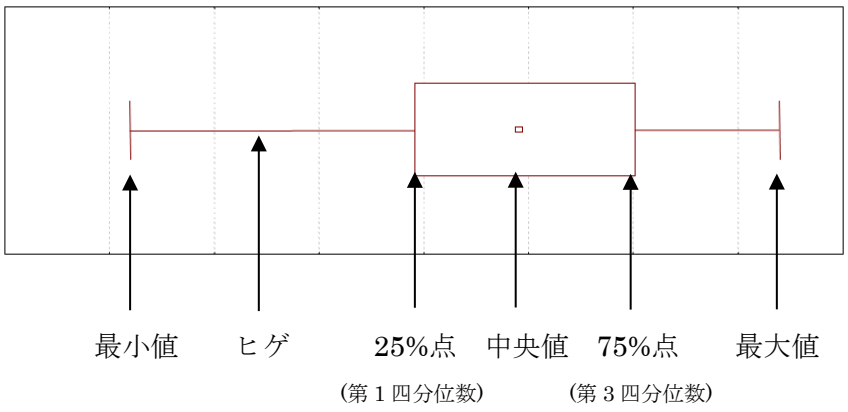
☆レーダーチャートの例

3 人の学生の 5 教科の成績をレーダーチャートで表現すると、以下のようになる。



4. 箱ヒゲ図

- 箱ヒゲ図はヒストグラム同様，データが集中している範囲・バラツキの大きさ・データの値や範囲を指定したとき，そこに全体の何%のデータが含まれるか，分布から調べることができる



- 四分位数

- データの散らばりを表す指標
- 値を小さい順に並べたとき，下からちょうど 25%となる値から 75%となる値までの範囲
  - 下からちょうど 25%となる点：第 1 四分位数
  - 下からちょうど 50%となる点：中央値
  - 下からちょうど 75%となる点：第 3 四分位数
  - 四分位範囲=第 3 四分位数 － 第 1 四分位数
- 四分位範囲の長所
  - 外れ値の影響を受けにくい
- 四分位範囲の短所
  - 計算が面倒
  - ケースを小さい順に並べ直さなければならない
  - ケース数が多くなると，計算すること自体が大変な作業になる

✧ 四分位範囲の解釈
- ばらつきが小さい＝四分位範囲が狭い
- ばらつきが大きい＝四分位範囲が広い

- 分析における注意

✧ 分布の中心やバラツキの程度の比較

✧ 分布の対称性や非対称性のチェック

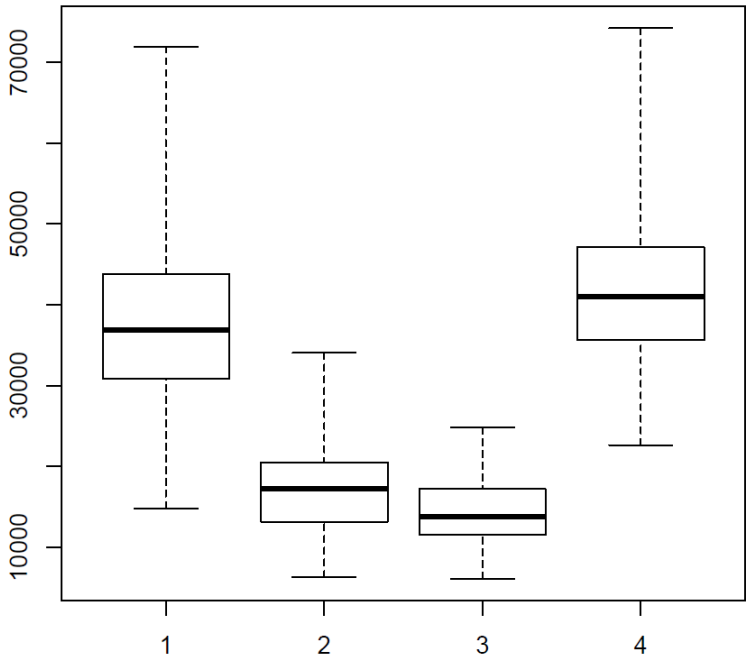
✧ 分布の単峰性・多峰性

＊基本的には単峰分布を想定したグラフである

5. ヒストグラムと箱ヒゲ図の関係

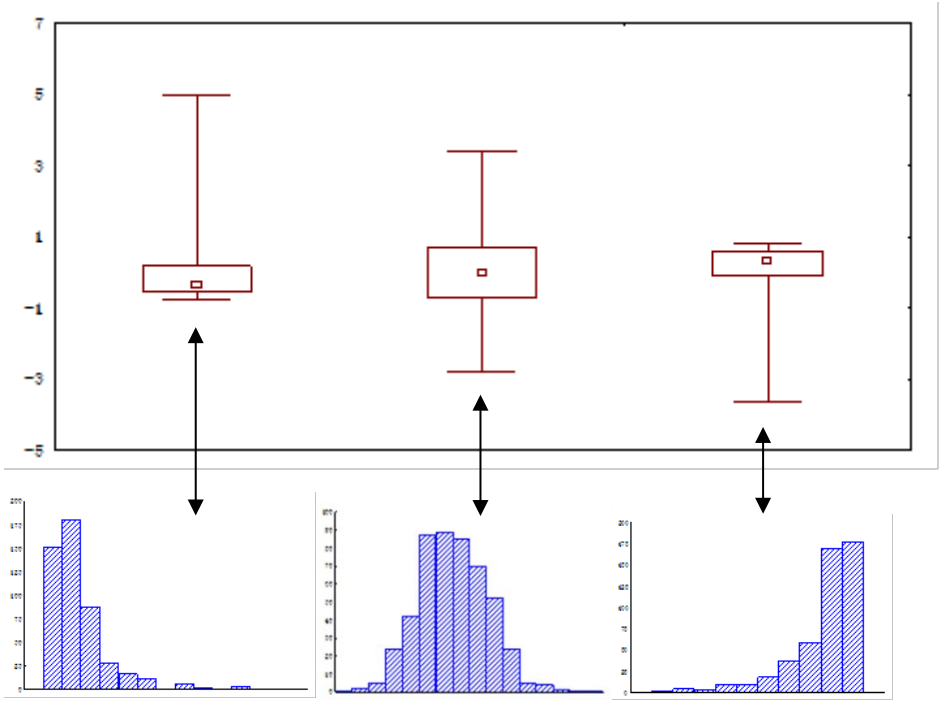
- ヒストグラム
  - 度数分布表（量的データ）をグラフにしたもの
- 箱ヒゲ図
  - 箱と箱からのびる線（ひげ）を使いデータのバラツキを示したグラフ
- 箱ヒゲ図のメリット・デメリット
  - メリット：複数の分布を比較したい場合に有効
  - デメリット：分布の形状，箱・ヒゲの形状・長さには注意を要する
- 勿論，ヒストグラムを複数作り，分布比較も可能であるが，比較する分布が多くなったときは箱ヒゲ図の方がスマート

☆4 社の売上金額の比較



- ヒストグラムと箱ヒゲ図との関係

  上の図が箱ヒゲ図。下の図は各箱ヒゲ図に対応するヒストグラム



6. グラフ表現における注意

- データや表現したい内容に合致した適切なグラフを使う
  - 間違った種類のグラフを選択すると、データの理解ができなかったり、間違った理解をしてしまう可能性がある
- 間違ったグラフの使い方の例
  - 棒グラフと折れ線グラフの間違い
    - 棒グラフで表すべきデータを折れ線グラフで表す
    - 折れ線グラフで表すべきデータを棒グラフで表す
- グラフを作成する際には見づらくならないように注意する
  - 見づらくなるポイントとして
    - 複数の変数の区別がつかない
    - 1つのグラフに多くのデータを表そうとしている
    - グラフの軸などの文字が小さい
    - 色が分かりづらい
      - ✓ モノクロで印刷されるグラフをカラーで作成した際には特に注意

練習問題

問 1（グラフの特徴）

グラフの特徴に関する記述として，適切でないものを次からから一つ選べ.

1. 全体に占める割合をグラフ化する際には，円グラフや帯グラフが用いられる.
2. 積み上げグラフは，カテゴリの割合の年次的な変化を見る際に用いられる.
3. レーダーチャートは，複数の指標のバランスを見る際に用いられる.
4. 折れ線グラフは，ある種の時間的な変化をみる際に用いられる.
5. サンプルサイズが大きいとき，幹葉図を用いると複雑になることがある.

問 2（四分位数の解釈 1）

ある小学校の卒業生を対象に，卒業までに図書館から借りた本の冊数を調査した結果，次のデータを得た.

最小値	1 冊
第 1 四分位数	9 冊
第 2 四分位数	12 冊
平均	18 冊
第 3 四分位数	23 冊
最大値	126 冊

この結果から次の 2 つのことを考えた.

A) 卒業までに半数の児童が 18 冊以上の本を図書館から借りている.
B) 借りた本の冊数は平均よりも少なかった児童が過半数である.

このとき，2 つの考えについて適切な組み合わせは次のうちどれか.

1. A, B 共に正しい.
2. A のみ正しい.
3. B のみ正しい.
4. A, B 共に正しくない.

問 3（四分位数の解釈 2）

ある店舗で顧客 100 人の過去 1 か月間の来店回数を尋ねて，次のような結果が得られた.

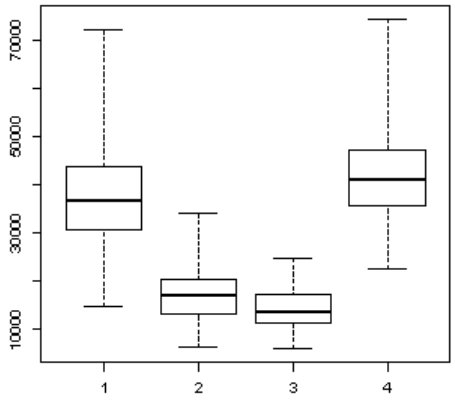
四分位数	第 1 四分位数	第 2 四分位数	第 3 四分位数
来店回数	3	8	16

この表から読み取れることとして，次から最も適切なものを一つ選べ.

1. 半分より多くの顧客の回数は 8 回未満である.
2. 100 回以上来店している顧客はいない.
3. 顧客を来店回数の小さい順で並べ替えたところ，25 番目の人は 3 回来店していた.
4. 顧客を来店回数の小さい順で並べ替えたところ，来店回数が多い上位 20%の人は少なくとも 20 回以上来店している.
5. 表からは上記①～④のことはどれもいえない.

問 4（箱ヒゲ図の解釈）

下図は，ある 4 社の売り上げを箱ヒゲ図により比較したものである．この解釈として，もっとも適切なものを一つ選べ.



1. 最大値が最も大きい箱ヒゲ図は 1 である.
2. 1～4 の中央値が箱ヒゲ図より分かっているので，これから平均値も計算できる.
3. 箱ヒゲ図の箱の面積が最も大きいものは，四分位範囲が最も大きい.
4. 1 と 4 に対応するヒストグラムを考える．このとき，これらの代表値の大きさについて，最頻値<中央値<平均値となる.
5. このように，4 社の売り上げを一度に比較する際，箱ヒゲ図よりヒストグラムで比較するべきである.

練習問題の解答

問 1（グラフの特徴）

解答：**2**

全体に占める割合を調べる際には，円グラフや帯グラフが用いられるので，①は正しい．積み上げ棒グラフは，割合ではなく度数を表しているため，度数の変化をみることはできるが，割合の変化を見るのには適していないため，②は適切ではない．レーダーチャートは，全体のバランスを見るときに用いられるので，③は正しい．折れ線グラフは，時間的な変化を見る際に用いられるので④は正しい．幹葉図は，サンプルサイズが大きいとき複雑になる傾向があるので⑤は正しい．従って，②が答え．

問 2（四分位数の解釈 1）

解答：**3**

A は，半数の児童の借りた本の冊数について考えている．即ち，A では，中央値(第 2 四分位数)の解釈として吟味する必要があり，中央値が 12 冊であるから不適切である．

B は，同様に考え，正しいことが分かる．従って，③が正しい．

問 3（四分位数の解釈 2）

解答：**1**

①は，中央値(第 2 四分位数)に関する問．その定義から①が適切であることが従う．②は，この結果からは断定できない．③は，サンプルサイズが偶数(=100)のため，中央値は，25 番目と 26 番目のデータの平均値で与えられることに注意．26 番目のデータは与えられていないため，表にある 3 になるとは限らない．④は，このデータだけでは分からない．⑤は，①が正しいため不適切．従って，①が答え．

問 4（箱ヒゲ図の解釈）

解答：**4**

最大値が最も大きい箱ヒゲ図は，④であることが分かる．従って，①は不適．一般に，箱ヒゲ図からは，平均値を求めることはできないため，②は不適．四分位範囲の大きさは，箱の面積ではなく箱の長さであるため，③は不適．④は，箱ヒゲ図とヒストグラムとの対応を問うている．その際，一般には，ヒゲの長さで(単峰な)ヒストグラムの裾が対称か非対称かを判断する．１と４は，右に裾が長い(歪度が正ともいう)ことが分かる．このとき，代表値の大きさを考えると，一般には，最頻値<中央値<平均値の順になる(因みに，左に裾が長い(歪度が負ともいう)とき，最頻値>中央値>平均値)．

従って，④が正しい．⑤は，箱ヒゲ図を用いるべき場合と逆のことを述べているため，不適．

5. 観測値の標準化と外れ値

本章の目的

- データの標準化を理解し，単位等が異なる変数間の比較をすることができる
- 外れ値の考え方，客観的な検出方法を理解する

Key Words: 標準化(基準化)・z 値・z スコア・偏差値・外れ値・外れ値の検出

1. 標準化

- 複数のデータセットを比較するとき，平均値や標準偏差が大きく異なると比較することが難しい．また，測定単位が異なる場合も同様の問題が生じる．
- このような場合，データに**標準化(基準化)**とよばれる処理を施し，統一した基準で比較することがある．身近な基準化の例として，成績の偏差値が挙げられる．偏差値は，特別な標準化の例であり，平均値や標準偏差が異なる科目の得点間の比較ができ，現状把握の一つの目安になっている．
- データの標準化とは，各観測値 x_i に対して，次の処理を施す:

$$z_i = (\text{観測値} - \text{平均値}) \div \text{標準偏差}.$$

この処理によって標準化された値(z 値または z スコアという)は，平均値 0，標準偏差 1 で与えられる．

- 試験の偏差値は，下式で与えられる:

$$\{(\text{得点} - \text{得点の平均値}) \div (\text{得点の標準偏差})\} \times 10 + 50$$

成績の場合，z は標準得点とよばれる．
この式により，偏差値は，平均値 50，標準偏差 10 の値をとる．

☆例題

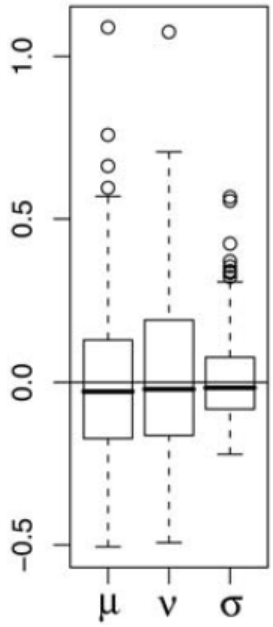
ある学生は，定期試験で，国語の点数が 60 点，社会の点数が 70 点であった．学年全体の結果は，国語は平均 50 点，標準偏差 5 点，社会は平均 50 点，標準偏差 20 点であった．このとき，この学生の国語と社会では，どちらの方が，学年順位が高いと考えられるか.

☆解答

国語と社会では，平均値はともに 50 点であるが，標準偏差が大きく異なっている．このことから，それぞれの値の標準化を行うと，国語は(60-50)/5=2，社会は(70-50)/20=1 となるため，一般に，標準化の値が大きい国語の方が社会よりも学年順位が高いと考えられる．

2. データの外れ値とその検出

- 調査や実験によってデータは得られるが，データの分布を確認せずに平均値や標準偏差を求めることは誤った解釈につながる恐れがある．
- そのために，データが得られたら，ヒストグラムや箱ヒゲ図等の統計グラフを用いて，データ全体の分布を確認することが肝要となる．
- このことにより，複数の分布が混在したデータになっていないか，他の観測値と比べ大きく外れている観測値(**外れ値**)が存在しているか，等を検証する．場合によっては，外れた観測値を除いて計算する等の適切なデータの分析が可能である．
- 箱ヒゲ図は，外れ値を検出するための簡易手法であり，外れ値は，四分位範囲の 1.5 倍よりも外側に離れている値として定義され，「○」等で記される(下図参照).



☆例題

次のデータは，あるクラスの 20 人の登校時間を測った結果である:

56, 24, 32, 19, 33, 60, 31, 23, 22, 87, 45, 47, 12, 28, 7, 12, 43, 32, 101, 26

平均 37.0 分，標準偏差 23.51 分，第 1 四分位数 22.5 分，第 2 四分位数 31.5 分，第 3 四分位数 46.0 分，最小値 7 分，最大値 101 分である．箱ヒゲ図を利用して外れ値の検出を行い，その結果を述べよ．

☆解答

このデータの四分位範囲は，46.0-22.5=23.5 分であり，第 3 四分位数+1.5*23.5=81.25 になるため，箱ヒゲ図を用いると 87 分と 101 分の生徒の登校時間は外れ値と考えられる．

練習**問題**

問 1（標準化の計算）

ある試験の平均値は 54.2 点，標準偏差は 12.3 点であった．このとき，標準化された点数が 0 の学生のもとの点数はいくらか．次の①～④から一つ選べ．

1. 54.2
2. 12.3
3. 12.3/54.2
4. この情報のみでは求められない

問 2（標準化にもとづく解釈）

あるクラスの試験において，以下の人(a～c)を点数で小さい順に並べるとどうなるか．次の 1～4 から最も適切なものを一つ選べ．

a. クラスの平均値と標準偏差で点数を標準化して求めたところ値が 1 となった．
b. 点数がちょうどクラスの点数の第 1 四分位数と一致した
c. 点数がちょうどクラスの点数の平均値と一致した．

なお，今回の試験におけるクラスの点数の分布は，平均値を中心に左右対称なひと山型分布で，平均値と中央値はほぼ一致した．

1. a<b<c の順
2. b<a<c の順
3. b<c<a の順
4. この情報のみでは求められない．

問 3（偏差値の解釈）

A 氏は，今度の期末試験で，国語では 56 点，数学では 45 点であった．なお，国語の平均点は，52.2 点，数学の平均点は，40.4 点，標準偏差はともに 12.1 点であった．このとき，A 氏の国語と数学の偏差値はどちらが大きいか. 次の①～④から最も適切なものを一つ選べ.

1. 国語の偏差値が高い
2. 数学の偏差値が高い
3. 国語と数学の偏差値は一致する
4. この情報のみでは求められない

問 4（データの解釈）

あるクラブで目を閉じて片足立ちして何秒立ち続けられるかの実験を行った．10 人の測定結果(秒)は次の通りであった．

	立ち時間(秒)
	27
	29
	87
	90
	103
	112
	119
	125
	130
	138
平均値	96.0
中央値	107.5

結果は小さい順に並べている．このとき，以下の①～⑤のうちから適切なものを選べ．

I. 平均値がデータの中心と考え，「このクラブの片足立ちの測定の結果，データの中心は 96.0 秒程度と考えられる」とすることが妥当である．

II. 中央値がデータの中心と考え，「このクラブの片足立ちの測定の結果，データの中心は 107.5 秒程度と考えられる」とすることが妥当である．

III. 27 秒と 29 秒は，ほかの観測値と比べ大きく異なることから，値の理由を確認することが望ましい．

1. Ⅰのみ正しい
2. Ⅱのみ正しい
3. Ⅲのみ正しい
4. ⅠとⅢは正しい
5. ⅡとⅢは正しい

練習問題の解答

問 1（標準化の計算）

解答：**1**

標準化の式により，標準化した点数が 0 であるから，実際の点数は平均値と一致することが分かり，①が答え．

問 2（標準化にもとづく解釈）

解答：**3**

c 氏の点数は，与えられた情報により，中央値(第 2 四分位数)と等しいため，b<c の順となる．また，c 氏の点数は，平均値であることから，標準化すると 0 となり，a 氏の点数は標準化すると 1 となり，c<a の順となる．即ち，b<c<a となり，③が答え．

問 3（偏差値の解釈）

解答：**2**

偏差値を比較することは，定義より，標準化の点数を比較することと同値である．特に，ここでは，標準偏差がともに等しいため，それぞれの偏差を比較すればよい．即ち，国語の偏差は，56-52.2=3.8，数学の偏差は，45-40.4=4.6 であるから，数学の偏差値が国語のそれよりも高い．よって，②が答え．

問 4（データの解釈）

解答：**5**

データ全体を考えると数値の小さい方に裾の長い分布になっている．このとき，外れ値(27,29)が混在しているため，平均値は，それに影響を受けるため，平均値を代表値としてとることは不適である．このようなときは，代表値で外れ値に影響を受けない中央値をとることが適切である．また，外れ値を除いて考えることも望ましい．従って，Ⅱと Ⅲが適切であり，⑤が答え．

6. 質的変数の関係を明らかにする

本章の目的

・　クロス集計表を使って、質的変数の間の関係を理解・解釈できるようになる

Key Words：クロス集計表、行と列、セルと周辺度数

1. クロス集計表とは？

- クロス集計表とは？
  - クロス集計表は質的変数の関係を理解するために用いられるもの
  - 2 つの変数のカテゴリーの組み合わせがデータの中でどのくらいの頻度で生じたのかを示したもの
    - ✧　一方の変数のカテゴリー別に、もう一方の度数分布を集計したものと考えることもできる

例：性別と商品満足度の関係についてのクロス集計表

		「この商品に満足していますか？」		
		満足	不満	計
性別	男性	310	90	400
	女性	440	160	600
	計	750	250	1000

  - 男性で商品に満足している人は 310 名
  - 男性で商品に不満な人は 90 名
  - 女性で商品に満足している人は 440 名
  - 男性で商品に不満な人は 160 名
- クロス集計表についての用語
  - クロス集計表の呼び方
    - ✧　「行数×列数」のクロス集計表と呼ぶ

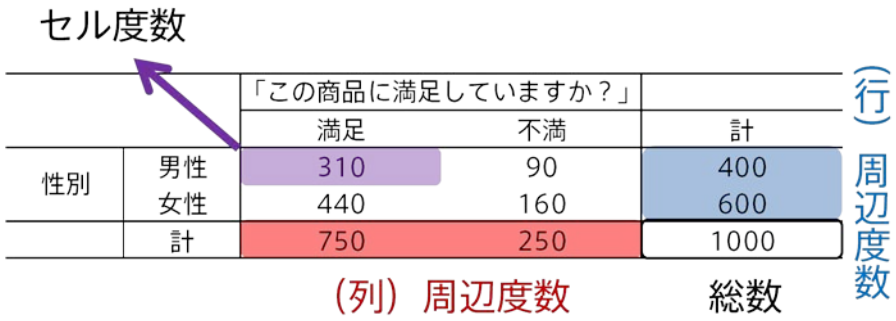
      ⇒例に用いたクロス集計表は「2×2」のクロス集計表
      - 行数：横に来る変数のカテゴリー数
      - 列数：横に来る変数のカテゴリー数

		「この商品に満足していますか？」		
		満足	不満	計
性別	男性	310	90	400
	女性	440	160	600
	計	750	250	1000

行

列

- セルと周辺度数
  - セル度数：各カテゴリーの組合せに対する度数
  - 周辺度数：変数の各カテゴリーの総数



2. クロス集計表を解釈する

- クロス集計表を解釈する
  - 度数の情報だけだとクロス集計表をうまく解釈できないことがある。クロス集計表を解釈するさいには、行ごと／列ごとに割合を求めることが多い。
  - なお、行ごとに割合を求めるのか／列ごとに割合を求めるのか、ということは分析目的に依存する

- 行ごとに割合を求める
  - 各行で合計が 100%になるように割合を求める
  - 列側の変数の各カテゴリーに対し、行ごとの割合を比較することで、変数間の関係を解釈する

例：性別と商品満足度の関係についてのクロス集計表

		「この商品に満足していますか？」					
		満足		不満		計	
性別	男性	77.5	(310)	22.5	(90)	100.0	(400)
	女性	73.3	(440)	26.7	(160)	100.0	(600)
	計	75.0	(750)	25.0	(250)	100.0	(1000)

単位：% （）内の数値は実数

☆男性は約 78%が商品に満足している。一方、女性は約 73%が商品に満足している。
☆男女間で満足している人の割合は大きく違わないので、性別と商品満足度の間の関係はそこまで強くない

＊性別ごとに商品に満足している人の割合を比較している

- 列ごとに割合を求める
  - 各列で合計が 100%になるように割合を求める
  - 行側の変数の各カテゴリーに対し、列ごとの割合を比較することで、変数間の関係を解釈する

例：性別と商品満足度の関係についてのクロス集計表

		「この商品に満足していますか？」					
		満足		不満		計	
性別	男性	41.3	(310)	36.0	(90)	40.0	(400)
	女性	58.7	(440)	64.0	(160)	60.0	(600)
	計	100.0	(750)	100.0	(250)	100.0	(1000)

単位：% ()内の数値は実数

☆商品に満足している人のうち、約 60%は女性である。一方、商品に不満を持っている人のうち、約 64%は女性である。

☆商品に満足している人でも不満を持っている人でも男女比率は大きく変わらないので、性別と商品満足度の間の関係はそこまで強くない

＊商品への満足度に対する回答ごとに男女の割合を比較している

練習問題

問 1（クロス集計表の解釈）

かすみさんの学校では 4 つのクラブのいずれかに所属することになっています。かすみさんは，クラブ活動と好きなテレビ番組の関係を知りたいと思い，所属クラブと最も好きなテレビ番組のジャンルについて調査を行いました。調査の結果を次のようなクロス集計表にまとめました。

	最も好きなジャンル				
所属クラブ	**スポーツ中継**	**歌番組**	**ドラマ**	**バラエティ**	**合計**
サッカー部	15	2	5	8	30
野球部	20	6	6	18	50
合唱部	6	40	30	24	100
演劇部	2	4	8	6	20
合計	43	52	49	56	200

（1）サッカー部に所属する部員で，ドラマを選んだ人の割合の求め方の適切なものを選びなさい。

1． 5/200　　2． 5/30　　3．5/49　　4．5/15

（2）文化系の部（合唱部と演芸部）に所属する部員で，バラエティを選んだ人の割合の求め方の適切なものを選びなさい。

1．24/100+6/20　　2．(24+6)/(100+20)　　3． 24/200+6/200　　4．(24+6)/200

（3） かすみさんは集計した結果から次の①と②を結論としました。

① スポーツ選択した割合は野球部が最も高い
② バラエティを選択した割合は野球部が最も高い

かすみさんの考えた①と②に関して，正しいものを選びなさい。

1. ①と②の両方，正しい
2. ①は正しいが，②は正しくない。
3. ①は正しくないが，②は正しい。
4. ①と②の両方，正しくない

練習問題の解答

問 1（クロス集計表の解釈）

（1）解答：**2**

サッカー部に所属している人は 30 名で、そのうち、ドラマを見ている人は 5 名なので、サッカー部に所属する部員で，ドラマを選んだ人の割合は 2 の「5/30」となる。

（2）解答：**2**

合唱部に所属している人は 100 名、演劇部に所属している人は 20 名なので、文化系の部に所属している人数は 100+20＝120 名となる。一方、合唱部でバラエティを見ている人は 24 名、演劇部でバラエティを見ている人は 6 名なので、文化系の部活でバラエティを見ている人は 24+6＝30 名となる。したがって、文化系の部に所属する部員で，バラエティを選んだ人の割合は 2 の「(24+6)/(100+20)」となる。

（3）解答：**3**

野球部員のうちスポーツを選択したものの割合は 20/50=0.4 だが、これはサッカー部員のうちスポーツを選択したものの割合は 15/30=0.5 よりも小さいので、①は誤り。

一方、野球部員のうちバラエティを選択したものの割合は 18/50=0.36 で、他の部活の部員のものよりも高いので②は正しい。

7. 量的変数の関係を明らかにする

本章の目的

- 量的変数の間の関係を図で表現する事ができるようになる
- 相関係数を使って、量的変数の間の関係を理解・解釈できるようになる

Key Words：散布図、（ピアソンの積率）相関係数

0. 2つの変数の間の関係を明らかにする

- 変数間の関係を分析する理由
  - **関係の性質を知る**

    一方の変数の値が変化すると、他方の変数の値がどう変化するのかを明らかにする

    例：サークル活動に費やす時間が変化すると、平均 GPA は上昇するのか、それとも、減少するのか？
  - **関係の強さを知る**

    一方の変数の値が大きくなると、他方の変数の値がどのくらい大きくなるのかを明らかにする

    例：勉強時間が 1 時間増えると、試験の点数はどのくらい増えるのか？
  - **因果関係を明らかにする**

    2 つの変数のうち、どちらが原因で、どちらが結果かを明らかにする

    ただし、因果関係を統計的に明らかにする方法は本章の範囲を超えるので割愛する
- 本章の方法で明らかにできる関係
  - **正比例の関係**

    一方の変数の値が大きくなると、他方の変数の値も大きくなる

    例：勉強時間が増えると、試験の点数も増える
  - **逆比例の関係**

    一方の変数の値が大きくなると、他方の変数の値は小さくなる

    例：1 日あたりの運動量が増えると、体重は減少する
  - **無関係**

    一方の変数の値が変化しても、他方の変数の値は変化しない

    例：1 食あたりの摂取カロリー量が増えても、試験の得点は変化しない

1. 散布図

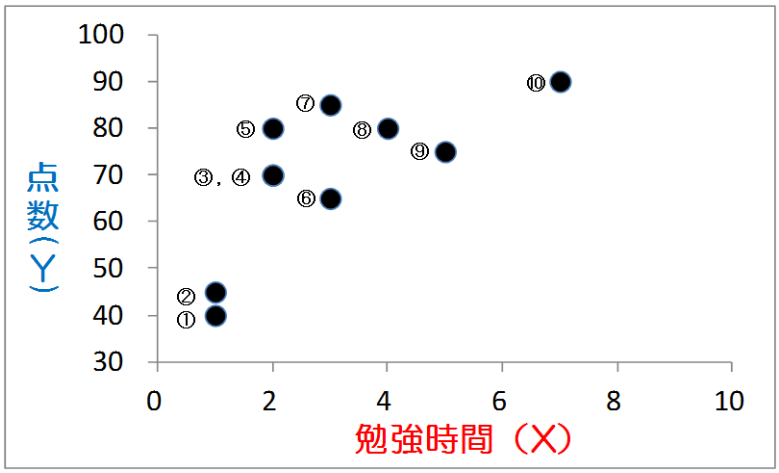
- 2 つの変数の関係を図示する方法
  - すべてのケースの値を XY 平面上に点としてプロットする
  - 点の配置をみることで、変数間の関係を解釈する
    - 正比例：右上がりの方向に点が配置されている
    - 逆比例：右下がりの方向に点が配置されている
    - 無関係：円状に点が配置されている

例：勉強時間とテストの成績の関係

10 人の学生に対し、1 日あたりの勉強時間（X）と統計学の試験の点数（Y）を調査したところ、次のような結果が得られた。

	1 日の勉強強時間	統計学の点数
①	1	45
②	1	40
③	2	70
④	2	70
⑤	2	80
⑥	3	65
⑦	3	85
⑧	4	80
⑨	5	75
⑩	7	90

このデータをもとに、散布図を描くと以下のような形になる。



右上がりの方向に点が配置されているので、勉強時間と点数の間には正比例の関係があることがわかる。

2. ピアソンの積率相関係数

- 関係のモノサシとしての相関係数
  - 散布図だけだと、関係の方向性は分かっても関係の強さはよくわからない
  - 関係の方向性と強さを同時に表現できるモノサシとして、相関係数が必要
- 相関係数の性質
  - 相関係数の定義

$$r_{xy}=\frac{s_{xy}}{s_x s_y}=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}(y_i-\bar{y})^2}}$$

  - 相関係数を理解するポイント

✧ **相関係数の符号**

符号をみることで、関係の方向を理解できる

☆符号が＋：正比例の関係（X の値が増えると、Y の値も増える）

☆符号がー：逆比例の関係（X の値が増えると、Y の値が減る）

✧ **相関係数の数値（絶対値）**

数字の大きさをみることで、関係の強さが分かる

☆数字が 0 であれば、2 つの変数の間に関係はない

☆数字が 1 に近くなるに連れて、関係は強くなり、直線的になる

＊関係の強さとしては、0.55 と－0.55 は同じ

例：勉強時間とテストの成績の関係

先にふれた、勉強時間（X）と試験の点数（Y）のデータに対し、相関係数を計算する。

計算の過程は以下のとおり。

- 分子（共分散 $s_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$）を計算する

	1日の勉強時間（X）	統計学の点数（Y）	X—平均値	Y—平均値	（X—平均値）×（Y—平均値）
1	1	45	1-3=-2	45-70=-25	-2×-25=50
2	1	40	1-3=-2	40-70=-30	-2×-30=60
3	2	70	2-3=-1	70-70=0	-1×0=0
4	2	70	2-3=-1	70-70=0	-1×0=0
5	2	80	2-3=-1	80-70=10	-1×10=-10
6	3	65	3-3=0	65-70=-5	0×-5=0
7	3	85	3-3=0	85-70=15	0×15=0
8	4	80	4-3=1	80-70=10	1×10=10
9	5	75	5-3=2	75-70=5	2×5=10
10	7	90	7-3=4	90-70=20	4×20=80
平均値	3	70			（50+60+0+0-10+0+0+10+10+80）÷10=20
標準偏差	1.8	15.5			

共分散

- 相関係数を計算する

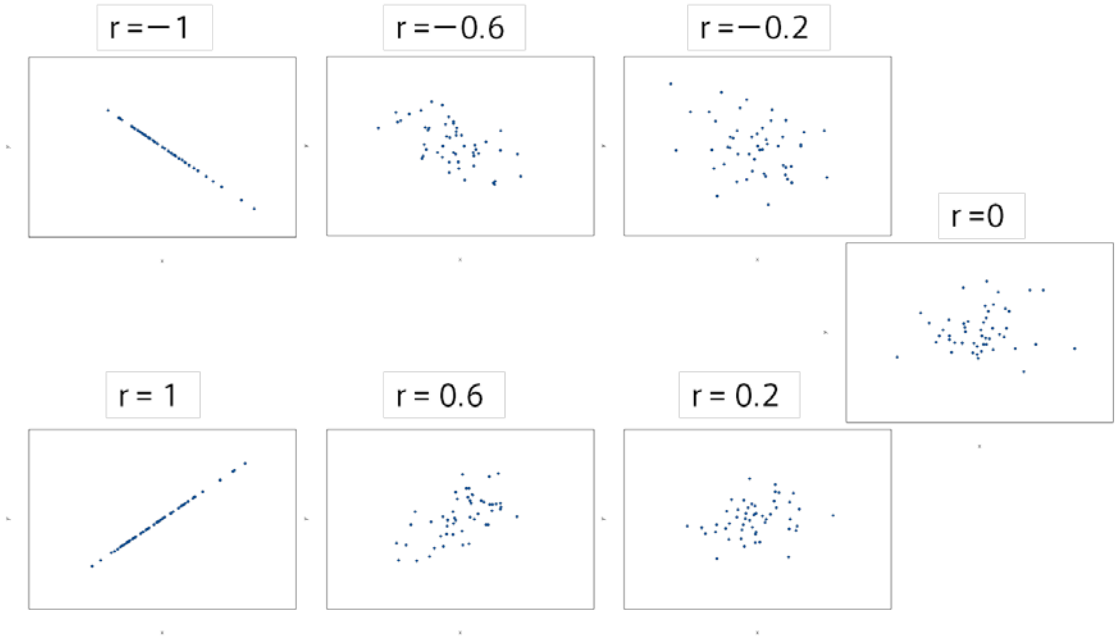
相関係数＝共分散÷（X の標準偏差×Y の標準偏差）

＝20÷（1.8×15.5）＝0.72

- 勉強時間と試験の点数の関係を解釈する
  - 符号が＋　：正比例の関係（勉強時間が増えると、試験の点数もあがる）
  - 絶対値 0.72：1 に近いので、二つの変数の関係は強い

3. 散布図と相関係数の関係

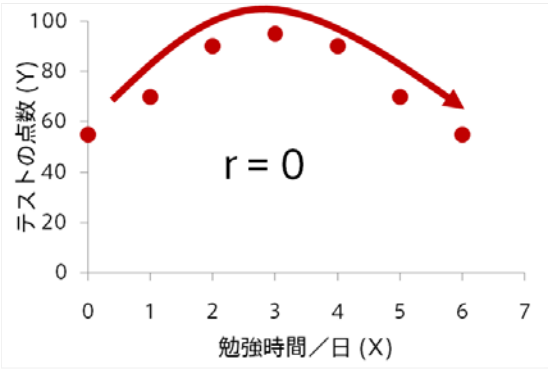
- 各散布図に対応する相関係数を求めると、以下のようになる



4. 相関係数の問題点

- 直線的でない関係を検出できない

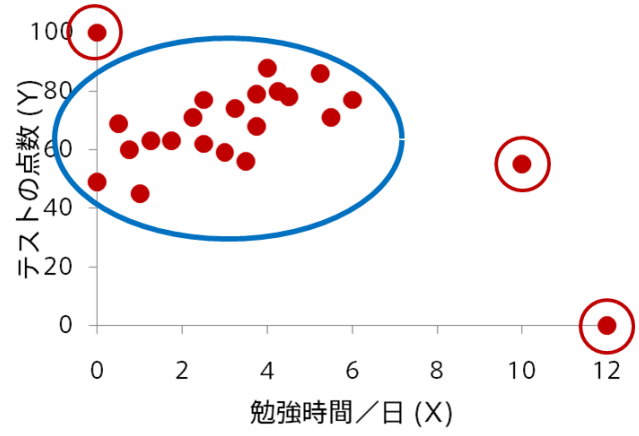
2 つの変数の間の関係が逆 U 字関係のような場合、相関係数の値は 0 になる



- はずれ値に大きな影響を受ける

はずれ値によって相関係数の値は大きな影響を受けてしまう

- 青の楕円で囲まれた部分だけのデータの相関係数は 0.68
- はずれ値を 3 つ追加（赤丸で囲まれたデータ）
- 相関係数を計算すると－0.42 となる

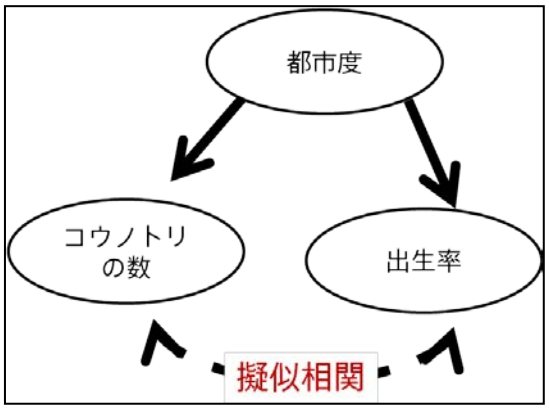


- 擬似相関の可能性

X と Y の相関が第 3 の要因 Z によって生じる可能性がある

例：煙突にコウノトリの巣が数多くある地域では出生率が高い

- 居住地域によって、生じた擬似的関係
  - 出生率：田舎>都会
  - コウノトリの数：田舎>都会



練習問題

問 1（相関係数の性質）

2 つの変量 x と y の相関係数を r とする。このときの記述として**誤っているもの**を、次の 1～5 のうちから一つ選べ。

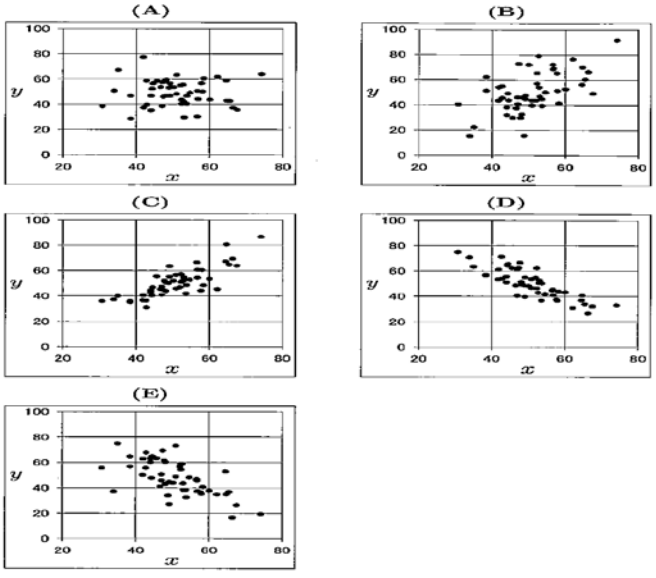
1. x をすべて 2 倍してできる変量 z と変量 y の相関係数は r と等しい。
2. x にすべて 10 を加えてできる変量 z と変量 y の相関係数は r と等しい。
3. r は−1 以上 1 以下の値を必ず取る。
4. 変量 y と変量 x の相関係数は$-r$ となる。
5. 2 つの変量 x と y が右下がりの直線近くに分布しているとき、相関係数 r は−1 に近い値となる。

問 2（相関係数と散布図の関係）

下の 5 つの散布図**(A)～(E)**は、5 種類の変量 x と y の関係を表したものである。

この 5 種類のデータの相関係数についての記述として**誤っているもの**を、次の 1～5 のうちから一つ選べ。

1. **(B)**は正の相関があり、相関係数は正の値をとる。
2. **(A)**の相関係数は他のデータに比べて 0 に近い値をとる。
3. **(B)**よりも**(C)**の方が、相関係数の値は大きい。
4. **(E)**は負の相関があり、相関係数は負の値をとる。
5. **(C)**と**(D)**はいずれも強い相関があり、相関係数の値は他のデータに比べて 1 に近くなる。



練習問題の解答

問 1（相関係数の性質）

解答：**4**

相関係数を求める際に利用する変量の値すべてに特定の値をたしたり、かけたりしても相関係数の値は変わらないので、1 と 2 は正しい。3 と 5 は相関係数の性質そのものなので、正しい。一方、相関係数を求める際に利用した変量の順を入れ替えても相関係数は変わらないので、4 が誤り。

問 2（相関係数と散布図の関係）

解答：**5**

散布図と相関係数の関係より，D は相関は強いが，負の相関のため，相関係数は負の値をとるので、5 は誤り。