2005．7．22

第２単元

データの記述

目 次

第2単元

データの記述

単元のねらい　データに潜む統計的規則性を発見するための最有力な手段はデータのグラフ化であろう．データを数値として眺めていても分からないものが，グラフで表わすことによって顕在化することが多い．グラフ化の道具としてExcelはたいへん手軽なものである．しかし，ただ単にExcelでグラフを描いても不十分なことが多く，それなりに工夫が必要となる．また，その後の統計的な処理のためにはグラフ化だけでなく，データを代表する数字（統計）として表わすことも重要である．このように，得られたデータをグラフ化し，代表する数字で表わすことを データの記述 という．

この単元では，第1単元で学習したように，データを質的データと量的データに分け，それぞれについてデータの記述方法を学習することにする．

1　質的データの記述

1.1　質的データ

(1)　質問調査

あなたが独自の問題意識を持って，ある集団に属する人たちの意識や行動，物の所有などについて調べたいと考えたときには，どのようにデータを集めれば良いだろうか．その人の意識や行動は，本人に尋ねないと分からないので，一人一人に質問をして，その回答を集めなければならない．この方法を 質問調査法 と呼ぶ[1]．

次節以降で，できるだけ具体的な事例を取り上げて，質問調査で得られる質的データの多面的な解析法を説明する．

具体的な事例の一つとして，携帯電話の市場調査担当者が，利用実態を把握するために行った調査を取り上げる．

(2)　携帯電話の利用実態調査

調査の対象者は，全国から無作為に抽出した 20 歳から 59 歳までの男女[2]．

調査票[3]

Q1．あなたは，自分専用の携帯電話またはPHS を持っていますか？(○はひとつ)

1．持っている　　2．持っていない　→Q 10 にお進み下さい．

Q2〜Q9 までは，携帯電話・ＰＨＳをお持ちの方のみがお答え下さい

[1] 質問調査の方法については，[現代統計実務講座」第 6 単元 参照．

[2] 調査協力：　(株) マクロミル．

[3] 調査票には，この他に，性別，年齢，職業など 対象者の属性に関する質問項目を設ける．

Q2．あなたは，普段，携帯電話のどのような機能をお使いですか？（○はいくつでも）

1．通話機能
2．メール機能（文章）
3．メール機能（写真付き）
4．メール機能（動画付き）
5．写真撮影
6．動画撮影
7．インターネット接続サービス
8．地図や位置情報（ＧＰＳ）サービス
9．ゲーム機能
10．その他

Q3．上記のうちで，もっともよくお使いの機能はどれですか？Q2 の番号から，ひとつだけ記入してください．

□

Q4．月々の利用金額は，およそどのくらいですか？（○はひとつ）
この半年くらいのだいたいの平均でお答え下さい．また，携帯電話・ＰＨＳを複数お使いの方は，最もよく使う電話機についてお答え下さい．

1．5000円未満
2．5000円以上１万円未満
3．1万円以上２万円未満
4．2万円以上

Q5-Q8 は省略

Q9．あなたは，今お持ちの携帯電話・ＰＨＳについて，全体として使いやすいとお感じですか？それとも，使いにくいとお感じですか？（○はひとつ）

1．とても使いやすい
2．まあ使いやすい
3．どちらとも言えない
4．やや使いにくい
5．かなり使いにくい

調査票 終わり

(3)　質問形式の分類

ここにあげた質問はいずれも，あらかじめ用意された選択肢の中から回答を選ぶ質問形式である．Q1のように，2つの選択肢から一つを選ぶ形式を **2項選択**，Q2〜Q9のように，3つ以上の選択肢から選ぶ形式を **多項選択** と呼ぶ．選ぶことのできる選択肢の個数は，Q1, Q3, Q4, Q9 は一つだけだが，Q2は複数の選択を認めている．一つだけ選ぶ形式を，**単一回答方式**（Single Answer），複数の選択を **多重回答方式**（Multi Answer） と呼ぶ．それぞれ，SA, MA と省略されることがある．

Q1 は対象者全員に答えてもらうが，Q2以降は，Q1 で「携帯電話・PHS を持っている」人のみに回答してもらう質問になっている．このように質問によって回答する人を条件付けし，一部の人だけに尋ねることもできる．

これらの質問に対する回答は，選択肢の番号として得られるが，番号は，選択肢を識別する記号であり，数値としての大小関係や倍数関係を意味していない．番号の代わりに (a・b・c), (あ・い・う), (イ・ロ・ハ) などの文字を使用しても構わないし，番号や記号をつけないこともありうる．通常は，1から始まる「番号」を使うが，これはデータ処理上，効率的なためである．

つまり，質問から得られるデータ（結果）は数量ではなく，分類を表わしている．これを量的変数と区別して，質的変数と呼ぶ．質的変数を対象とする多変量解析手法である **数量化理論** では，質的変数のことを **アイテム** と呼ぶ．

また，これらの質的変数の取る値（この質問例では，Q1では，「1．持っている」「2．持っていない」，Q2では「1．通話」「2．メール機能（文章）」「3．メール機能（写真付き）」「4．メール機能（動画付き）」など）を **カテゴリー** と呼ぶ．

選択肢の並べ方を見ると，Q4 と Q9 には序列がある．Q4では，若い番号ほど利用金額は安く，Q9では，若い番号ほど使いやすさに関する評価が良い．これらは，逆から並べても構わないが，順序をばらばらにするのは不適当である．このようにカテゴリー（選択肢）に，大小関係や優劣などの序列がある質的変数を，第1単元で学んだように **順序尺度**（Ordinal Scale）と呼ぶ．Q9 のように程度を尋ねる質問では，満足・意向・好みの程度，賛否の度合いなどがある．

その他，頻度，所得，年齢，学歴（教育年数の長さ）などをカテゴリーで尋ねる場合は，順序尺度となる．

これに対して，その他の質問では，選択肢には大小や序列がなく，並べ方は任意である．このような質的変数を 名義尺度（Nominal Scale）と呼ぶ．質的データの記述方法は，各カテゴリー（選択肢）を選んだ人を数えあげ（度数），その度数を回答者の総数で割って構成比率を求めるのが基本である．

本日のまとめ

今日は，質的データの解析の前提となる調査票の事例を見た．受講生は，自分が調査対象になったとき，どのように回答するかを考えてほしい．

明日以降，この調査結果を取り上げて，解析の考え方と方法について解説されるので，質問の形式（単一回答，多重回答，順序尺度，名義尺度）の違いについて十分に理解しておいてほしい．

1.2　一つの質的変数の記述

この節では，名義尺度（Q1, Q2, Q3）と順序尺度（Q4, Q9）別に，構成比率の求め方とグラフ化の方法について説明する．

(1)　構成比率とグラフ表現（名義尺度）

Q1（携帯電話の所有）について，回答の結果は表示1.1 のようにまとめられる．

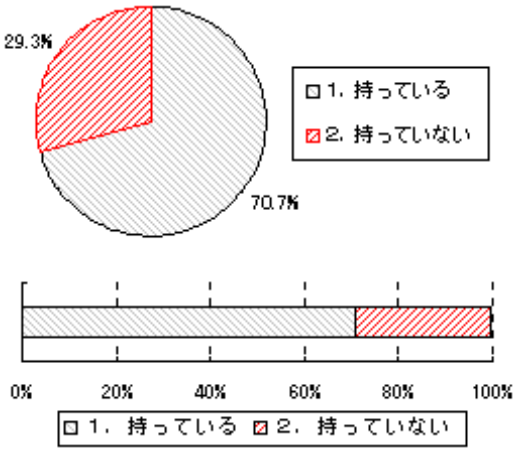
表示1.1: 携帯電話・PHS の所有率

	総数	1．持っている	2．持っていない
人数	2000	1414	586
%	100.0	70.7	29.3

上の行には度数が示され，左に総数が求められている．以下，総数を n で表わすこととし，表示などでは「総数が 2000」を $n = 2000$ と書く．下の行には度数を総数 n で割った構成比率が%で示されている．構成比率の合計は，当然 100% になる．

構成比率をグラフで表現すると，視覚的に特徴をつかむことができる．表示 1.2 は，Q1 の結果を 円グラフ（パイグラフ） と100%積み上げ横棒グラフ（帯グラフ） で表わしたものである．

表示1.2: 携帯電話・PHS の所有率 （n=2000）



$n = 2000$ の中から，携帯電話・PHS を持っている人（1414 人）だけを取り出して，Q2，Q3（利用機能）について回答の結果をまとめたのが，表示1.3 である．

表示1.1 は横にカテゴリーを取ったが，表示1.3 は縦にカテゴリーを取り，左列は度数を，右列は構成比率を表わしている[4]．

[4] ここでは，構成比率を% で表わしている．
表示 1.1 では，表側に % であることを示している．表示1.3 では，個々の値に % が付いている．
前者の式は 分子/分母*100 として，%と求め，後者の式は 分子/分母 として，セルの書式で% 表示をしている．

表示1.3: 利用している機能

（携帯電話・PHS の所有者のみ）

	利用している機能	Q2：多重回答		Q3：単一回答	
		人数	%	人数	%
	総数	1414	100.0%	1414	100.0%
1.	通話	1355	95.8%	444	31.4%
2.	メール（文章）	1261	89.2%	650	46.0%
3.	メール（写真）	569	40.2%	166	11.7%
4.	メール（動画）	81	5.7%	34	2.4%
5.	写真撮影	669	47.3%	8	0.6%
6.	動画撮影	166	11.7%	5	0.4%
7.	インターネット	723	51.1%	97	6.9%
8.	ＧＰＳ	93	6.6%	0	0.0%
9.	ゲーム機能	366	25.9%	6	0.4%
10.	その他	178	12.6%	4	0.3%
	合計	5461	386.2%	1414	100.0%

また，Q2（利用している機能）を左に，Q3（最も利用している機能）を右に取り，２つの質問の結果を一つの表にまとめて示している．

Q3 は単一回答方式であるから，構成比を合わせるとちょうど100 %になる．それに対して，Q2 は一人の人がいくつもの選択肢を選ぶことを許しているので，構成比の合計は100 %を超える．このように構成比の合計が100 %を超える場合は，その理由を数表およびグラフ上に表記するのが良い．ここでは，表頭に 多重回答，単一回答 と記している．

表示1.4 は，Q3 の結果をパイグラフで表わしたものである．

パイグラフは Q3 のようにカテゴリーが多い場合やその構成比が数%以下である場合，その大小は分かりにくい．

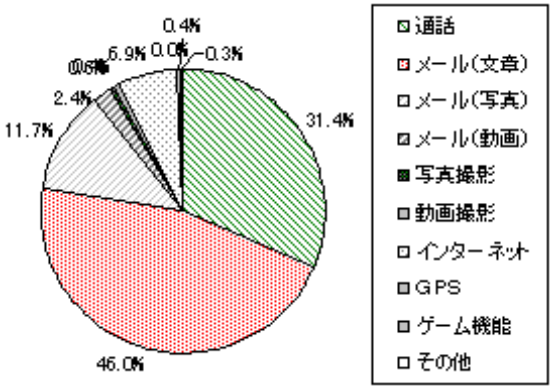
そこで，Q3 の結果を棒グラフで表わしたものが表示1.5 である．

２つのグラフでは横軸の項目の並び順が異なることに注意する．右の棒グラ

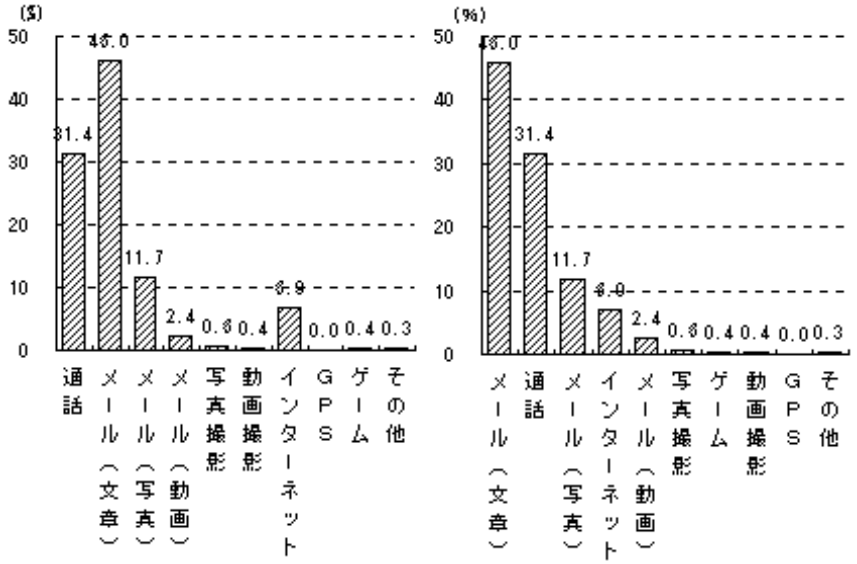
ここでは，% を小数点以下1桁まで表示している．第3 単元で，比率の値の精度について学ぶ．その結果によると，総数が1000 以下であるときには，割合の誤差は数% である．したがって，総数が1000 以下のデータから計算した% は小数点以下を表示しない方が，見る人に誤解を与えないと考えられる．

表示1.4: **最も利用している携帯電話の機能**

携帯電話・PHS **の所有者のみ**（n=1414）



表示1.5: **最も利用している携帯電話の機能**



フは，カテゴリーを回答の多い順に並べ替え（降順），序列を把握しやすくしている．

なお，標本観察のデータの場合は，サンプルの大きさが誤差の大きさと関連

するので，グラフ上にはサンプルサイズ（構成比を算出する分母となった回答者総数 n）を明示する．

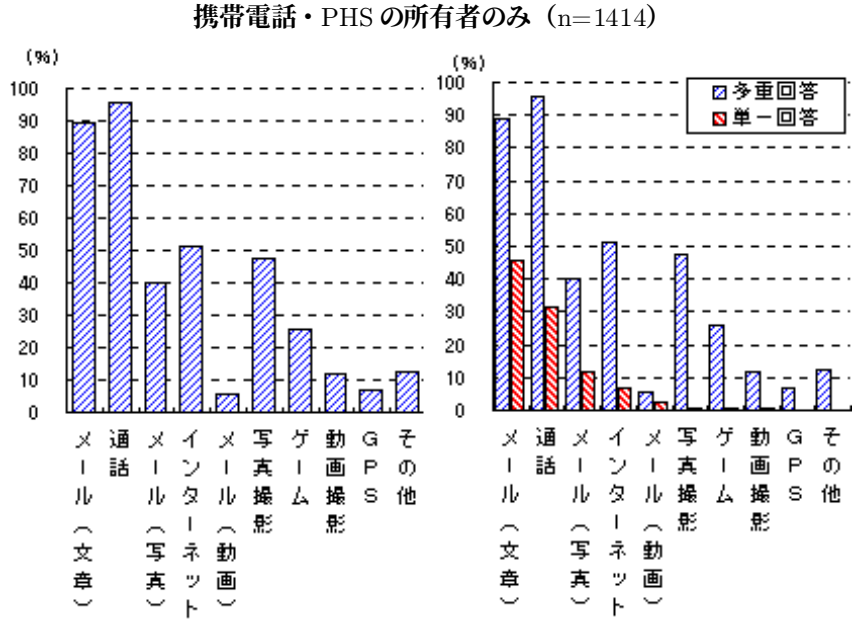
次に多重回答の場合を説明する．

多重回答の結果は，構成比の合計が100 %を超えるため，パイグラフや帯グラフで示すことはできない．

表示1.6 の左はQ2（利用している機能（多重回答））の結果を棒グラフにしたものである．横軸の配列は，Q3（もっとも利用している機能（単一回答））の降順にしてある．

表示1.6の右は，Q2 と Q3 の結果を一つのグラフにまとめ，横軸の配列は Q3 の降順にしてある．このような工夫をすると，両方の情報を比較検討できる．

表示1.6: 利用している携帯電話の機能

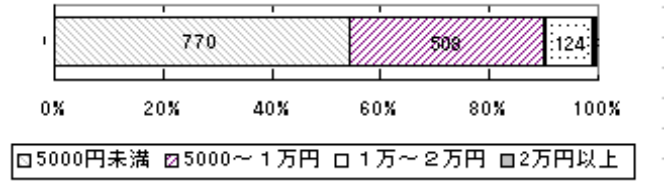


(2) 構成比率とグラフ表現（順序尺度）

順序尺度である Q4，Q9 は，どのようなグラフ表現が適当であろうか．Q1，Q3 と同様に円グラフ，棒グラフを描くこともできるが，順序尺度では一つひとつのカテゴリーの比率だけではなく，累積比率やカテゴリーを合計した比率を把握することも重要になる．このような場合は，表示1.7,1.8 のような100%積み上げグラフ（帯グラフ）が便利である． Q4（利用料金）では，「1万円未満

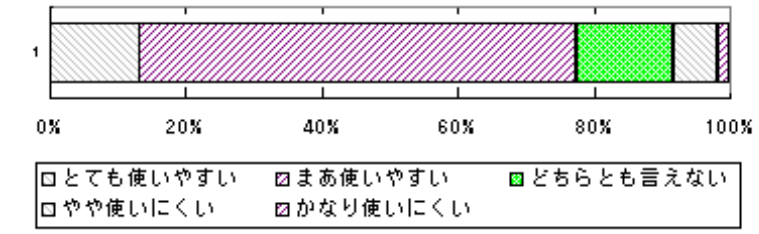
表示1.7: 利用料金（Q4）

総数	5000円未満	5000～１万	１万～２万	2万円以上
1414	770	508	124	12
100.0%	54.5%	35.9%	8.8%	0.8%



表示1.8: 利用している携帯電話の使いやすさ（Q9）

総数	とても使いやすい	まあ使いやすい	どちらとも言えない	やや使いにくい	かなり使いにくい
1414	184	910	201	93	26
100.0%	13.0%	64.4%	14.2%	6.6%	1.8%



（カテゴリー 1 と 2 の合計）」が約90%であることが，Q9（使いやすさ）では，「使いやすいと感じている人の合計（1 と 2 の合計）」が約 3/4 であることが読み取れる．

(3)　尺度の変換

順序尺度は，カテゴリーの順序に意味はあるが，カテゴリーに数値としての意味はないことは先に述べた．例えば，Q4（利用料金）の各カテゴリー幅は一定ではない．また，Q9（使いやすさ）は，測定器で計測されるものではないので，5つのカテゴリーに絶対的な原点はなく，カテゴリー間の距離が等間隔かどうか保証されていない．しかし，ほぼ等間隔であるとみなせる場合（または，カテゴリー間の距離を想定出来る場合）は，カテゴリーに数値（点数）を与えて，量的変数として扱うことがある．例えば，Q9では「とても使いやすい」を5点，「かなり使いにくい」を1点としたり，中間回答の「どちらともいえない」を0点，「とても使いやすい」を2点，「かなり使いにくい」を−2点するなどして，その中間に1点刻みの得点を与えることなどが考えられる．

順序尺度を量的変数に変換して平均値を算出すると，カテゴリーごとの分布情報を集約できるので層別比較や，変数同士の比較などがシンプルになるという利点がある．さらに，平均値と分散から，相関係数 (§4.2) を計算できるので，対応する解析方法が広がる．

名義尺度で2項選択の変数（所有の有無，賛否，男女，生存と死亡，条件該当・非該当などの別など）は，それぞれのカテゴリーに 1，0 （yes と no）の2値を与えることがあり，ダミー変数 と呼ばれる．

3カテゴリー以上の順序尺度で意味的に2つに分類できる場合，適当にカテゴリーを併合して２値とすることもある．例えば，使いやすさを「とても使いやすい」，「まあ使いやすい」をまとめて「使いやすい」とし，それ以外をまとめて「使いやすいとはいえない」とする．どのように併合するかは，回答の内容によるが，ほぼ半々に近くなるようにするのが好ましい場合がある．

(4)　Excelによる入力と単純集計

前項までに取り上げた調査データから，携帯電話を「持っている」と答えた人の中からランダムに抜き取った$n = 410$ の観測データがExcel ファイルに記録されている．

その最初の部分を表示1.9 に示す.

表示1.9: Excel のデータ表

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1						通話	文章メール	写真メール	動画メール	写真撮影	動画撮影	インターネット	地図	ｹﾞｰﾑ	その他	主使用	利用料金	使いやすさ
2	サンプルNo.	SEX	AGE	AREA	JOB	Q2-1	Q2-2	Q2-3	Q2-4	Q2-5	Q2-6	Q2-7	Q2-8	Q2-9	Q2-10	Q3	Q4	Q9
3	1	1	2	8	3	1	1	1		1	1	1		1		2	2	2
4	2	1	2	4	4	1	1	1		1		1		1		2	2	2
5	3	1	4	6	4	1	1			1		1				2	2	2
6	4	2	2	8	9	1			1			1		1		1	1	2
7	5	1	4	5	3	1	1			1						1	2	4

回答の結果は，一人の回答（1サンプル）を横1 行に入力する．はじめに，集まった調査票に連番号（サンプル番号）をふっておく．連番号はアンケート票の到着順などで良い.

サンプル番号，回答者の属性（性別，年齢階層，職業）を入力する．続いて，回答結果を質問の順序どおりに入力できるようにする.

単一回答形式の質問には一つの列を，多重回答形式の質問には選択肢の数だけの列を用意すると，表示1.9 のようなシートが出来上がる.

単一回答方式（Q1，Q3，Q4）については，選択された番号を入力する．多重回答方式（Q2）では，選択された選択肢の列に1 を入力する．選択されなかった選択肢の欄には，空白のままにするか，または 0 を入力する．いずれを取るかは，後の集計に用いるプログラムによって決める．ここでは，空白としている.

通常は入力の前に検票，入力後にデータクリーニングという処理を行う．必須の質問に回答がないサンプル，回答方法に誤まりや回答間の論理的な矛盾があるサンプルの処置方法を決めておき，慎重にチェックした後のサンプル（有効回答）を集計のベースとするのが望ましい.

表示1.9の表が完成したら，表示1.1〜表示1.8 のような集計表を作成する．

集計手続きは，それぞれの番号の出現頻度（度数）を，Excel関数（COUNTIF）を用いて求める．

例えば，男性の回答者の人数を数えるためには，

```
=COUNTIF(B3:B412,1)
```

とする．COUNTIF 関数の最初のパラメータは対象のデータ範囲である．データは3行目から412行目までである．2 番目のパラメータの 1 は男性を表わす．

次に回答総数で除して構成比を計算する．

集計表からグラフを作り，オプション機能を使って，整形を加えると，このテキストに示されているようなグラフが得られる．グラフのオプション機能や整形方法の説明も市販書に譲る．ただし，添付するExcel ファイルには，基本出力として得られるグラフとそれから最終のグラフを導く手順について説明してある．ただし，Excelバージョン2000 についての説明であって，バージョンによって幾分異なる個所のあることをお断りしておく．

■　演習１　　表示1.9 のデータを使って，本文の説明にならい，女性の人数，インターネットの利用者，インターネットを主として利用する人数を求めよ．■

本日のまとめ

今日は，質的変数のデータをグラフ化し，比率を求める方法を学んだ．同じ質的変数でも，単一回答と多重回答で，また，順序に意味のある場合とない場合で，グラフ化するときにいろいろな工夫が必要であった．これらの違いを明確にし，整理することによって，ここで示した多くのグラフを皆さんのお仕事に適切に応用することができるでだろう．

この節の学習を進めながら，演習用データを Excel で整理し，図表化する技術を身につけると良い．応用範囲が広く大変便利である．

1.3　2つの質的変数の関係（1）

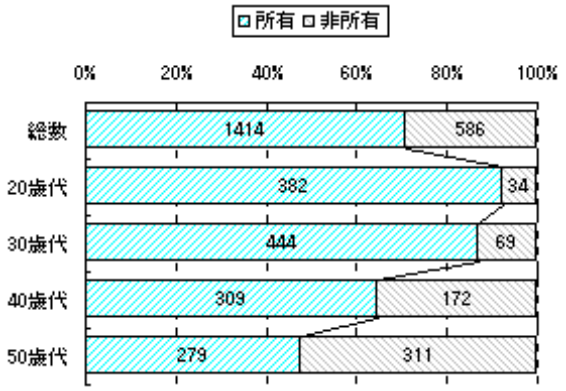
(1)　基本属性との関連

個人の意識や行動，物の所有状況などは，基本的な属性（性別や年齢，未既婚・子供の有無，職業，所得階層など）に影響を受けている場合が少なくない．先の調査の例では，性別や年齢によって携帯電話の所有率や利用機能などが違うことが予想される．そこで，回答者を性別や年齢によって，複数のグループに分け，グループごとの度数と構成比率を求め，差異を考察する．このように，ある変数についてひとかたまりの集団として見てきた度数分布と構成比率を，別の変数を用いて副次的なグループに分け，集計した表を クロス表（連関表） という．

Q1（携帯電話の所有）について年齢層別にみたクロス表は，表示1.10の左のようになった．

表示1.10: 携帯電話・PHSの所有率

	全体	（人）		（%）	
		所有	非所有	所有	非所有
総数	2000	1414	586	70.7	27.1
20歳代	416	382	34	91.8	8.2
30歳代	513	444	69	86.5	13.5
40歳代	481	309	172	64.2	35.8
50歳代	590	279	311	47.3	52.7



年齢層ごとの所有・非所有の人数，構成比率を横の方向に示している．クロス表の左側（表示1.10では 年齢層）を表側，上部（表示1.10では 携帯電話の所有・非所有）を表頭と呼んでいる．日本では上記のように表側を基準にして

横方向に構成比率を表示することが多い.

このクロス表の人数を，グラフ化したのが表示1.10 右のグラフである．グラフの中には人数が表示され，所有の割合は，上の目盛りから読み取ることができる.

表示1.10 から年齢が上がるにつれて，所有率が低くなる様子が一目で理解される.

Q2（利用機能）について，表示1.2(p.7) の回答数を男女別に分けた クロス表を表示1.11 に示す.

表示1.11: 利用している機能（多重回答）

	人数			%		
	全体	男性	女性	全体	男性	女性
総数	1414	761	653	100.0	100.0	100.0
通話	1355	748	607	95.8	98.3	93.0
メール（文章）	1261	646	615	89.2	84.9	94.2
インターネット	723	400	323	51.1	52.6	49.5
写真撮影	669	323	346	47.3	42.4	53.0
メール（写真）	569	253	316	40.2	33.2	48.4
ゲーム	366	211	155	25.9	27.7	23.7
動画撮影	166	76	90	11.7	10.0	13.8
ＧＰＳ	93	50	43	6.6	6.6	6.6
メール（動画）	81	42	39	5.7	5.5	6.0
その他	178	119	59	12.6	15.6	9.0

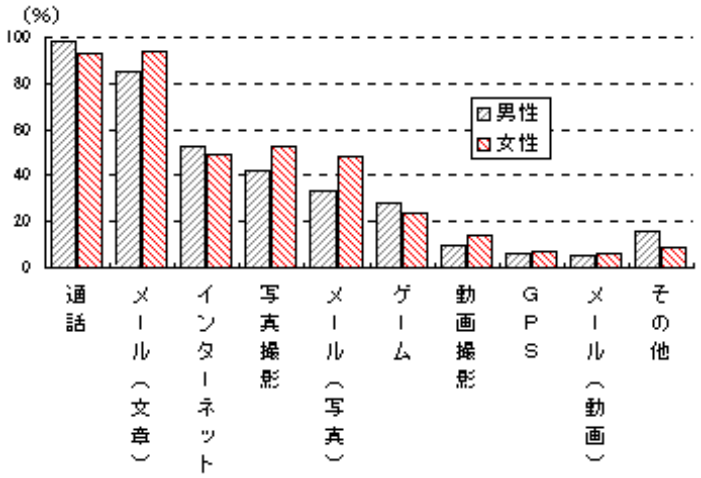
この表の「総数」と「その他」の行を除き，全体の回答数の大きさの順に並べ直し,「その他」の行を含めてグラフ化したのが表示1.12 である.

男女による差異が分かるように，棒を横に並べてある．このようにグラフは視覚的に特徴をとらえやすくするために，データを整理し，表現を工夫することができる.

さて，以上の分析から，年齢や性別が携帯電話の利用状況に影響を与えている要因と分かる．しかし，この逆をいうことはできない．つまり，携帯電話の所有や利用機能という要因が性別に影響を与えているとはいえない．携帯電話の所有や利用状況によって性別が影響される（変化する）ことはあり得ないか

表示1.12: 利用している機能（多重回答）

（携帯電話・PHS の所有者のみ）



らである．言い換えると，携帯電話の所有や利用状況は，結果を示す要因（従属する要因）であり，性・年齢はその違いを説明する要因，または先行する要因といえる．

■　演習２　演習1で用いた調査票のデータについて，本文にあげたもの以外の組合わせについてクロス表とグラフを作成せよ．それから，どのような傾向が掴めるか？　■

調査の対象が法人である場合は，基本属性（業種，従業員や売上，資本金などの規模）によって，クロス表を作ると良い．次の表示 1.13 は，パソコン導入状況を調べ，その結果を業種別に集計したクロス表である．

Q　貴法人では，業務利用のためにパーソナルコンピュータ（パソコン）を導入していますか？

1. １人１台体制である
2. 必要台数のみ導入している
3. 導入しているが必要台数には及んでいない
4. 未導入または導入検討中である

表示 1.13: 業種別パーソナルコンピュータ導入率

回答番号	1	2	3	4		
	1人1台体制	必要台数のみ	必要台数未達	未導入または検討中		
評点	4	3	2	1	計	平均点
建設	23	57	5	0	85	3.21
製造	33	80	14	2	129	3.12
運輸	3	16	0	0	19	3.16
販売	14	67	12	4	97	2.94
医療	15	17	1	1	34	3.35
ｻｰﾋﾞｽ業	11	26	1	0	38	3.26
官公庁	15	5	15	1	36	2.94
他	3	23	5	0	31	2.94
合計	117	291	53	8	469	3.10

質問の選択肢は順序尺度であるので,「1人1台体制」を最高点（4点）に1点刻みで配点し，平均を求めた[5]．全体でみると，評点の平均は3.10であり，業種間の違いは2.94〜3.35 と顕著な差は見られない．

カテゴリーの構成比率を帯グラフ表示したのが表示1.14である．

グラフを見ると，業種によってパターンが大きく異なっていることが分かる．官公庁では，中間が少なく，両極端に別れている．

グラフの中には件数が表示されている．

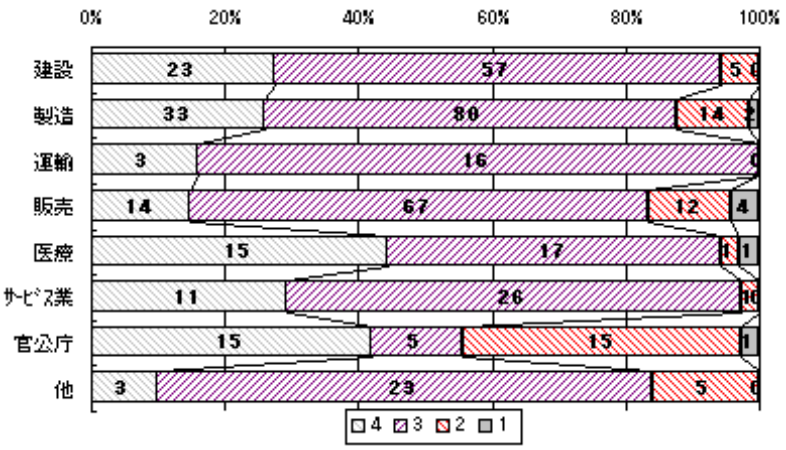
運輸と販売の4点（左端）の割合はほぼ等しい．しかし，件数は3件，14件と大きく異なっている．各業種のサンプル総数の大小がグラフに反映されず，判断を誤る可能性がある．この問題を解決するために工夫されたグラフを表示1.15に示す（市販の統計解析プログラムの一つであるJMP [6]で作成した）．

---

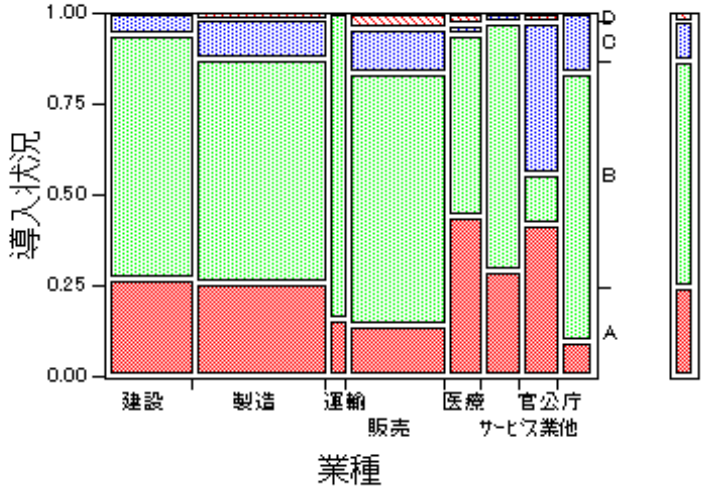
[5] 平均は次の式で計算される．
=SUMPRODUCT(評点の行, 度数の行)/度数の計
評点の行には行番号の前に $ をつける．

[6] アメリカのSAS社が開発したパソコン用統計解析プログラム．SAS Institute Japanから日本語版が購入できる．

表示 1.14: **業種別パソコン導入率**



表示 1.15: **業種別パソコン導入率** (JMP **による**)



ここでは， 4 **つの項目を** A, B, C, D **で表わしている．このグラフでは，構成比は縦に積み上げて表現されるが，構成比を横帯にしたものと本質的な違いは**

ない．このグラフの特徴は柱の太さが企業数に比例している点である．このことにより，業種ごとのウエイトの違いが分かるとともに，積み上げられた棒区画の面積が件数に比例するので，全体の中での各区画の構成比をつかむことができる．

(2)　2つの順序尺度のクロス表

以下の例は，ある電気メーカーが，顧客満足度調査を行った例である．複写機は，紙詰まりがなく，常時きれいなコピー画質を保つために定期メンテナンスが必要である．顧客満足度は，商品そのものの性能に加えて，メンテナンスの良さに対しても大きな影響を受けると考えられる．調査では，複写機を購入してみての総合的な満足度とともに，それに影響を与えていると考えられる要因をいくつか挙げて，それぞれ5段階の順序尺度で質問をした．表示1.16，1.17は，サービスマンの保守状況の説明の分かりやすさと，総合的な満足度の関係をまとめたクロス表とそのグラフである．

表示1.16: 保守説明の分かりやすさと総合満足度のクロス表

	評点	5	4	3	2	1		
評点	総合満足度 / 保守説明のわかりやすさ	満足である	どちらかと言えば満足である	どちらとも言えない	どちらかと言えば不満足である	不満足である	合計	平均
5	満足である	38	25	17	10	0	90	4.0
4	どちらかと言えば満足である	25	35	44	22	6	132	3.4
3	どちらとも言えない	10	26	33	24	8	101	3.1
2	どちらかと言えば不満足である	6	12	25	13	12	68	2.8
1	不満足である	0	3	11	8	7	29	2.3
	合計	79	101	130	77	33	420	3.3

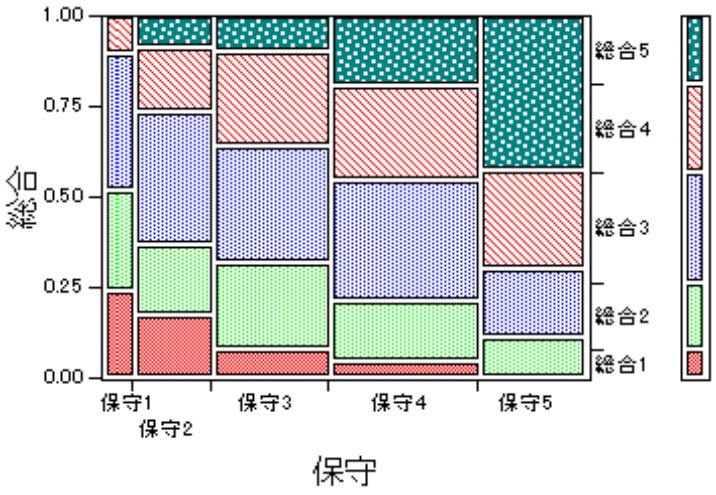
表示1.16 を見よう．表側に説明の分かりやすさ（原因）を，表頭に総合満足度（結果）を取り上げている．

総合満足度の5段階を評点 1〜5 として，平均値が表の右に計算されている．

これを見ると，保守説明の分かりやすさにより，4.0 から 2.3 と変化していることが分かる．

業種別パソコン導入率の変化の場合と同様に，JMP を使ってグラフ化した結果を表示1.17に示す．

表示 1.17: サービスマンの保守説明の分かり易さ別にみた総合満足度



グラフの横軸，縦軸の目盛りは，表示1.16 の表側，表頭にある評点を使って省略している．

表示1.17から，サービスマンの保守状況の説明の分かりやすさと総合的な満足度は，正の相関関係があることが一見して理解される．すなわち，サービスマンの保守状況の説明に対する満足度が，顧客の総合満足度の形成に強い影響をもった要因であるということができる．

また，このように表示することで，2 つの変数の関連が表現されるだけではなく，総合満足度，サービスマンの保守状況ともに，不満をもっている企業が比較的少数に留まっていることが見て取れる．

(3)　グラフの選択

2つの質的変数の関係を表わすいくつかのグラフを説明した．それらのまとめとして，グラフを選択する基準を説明する．

第1単元 §2.1 (2) で質的変数は，カテゴリーの順序に意味のない名義尺度 と，意味のある 順序尺度 に分類した．

グラフは原因と結果の関係を表わすために用いられる．そこで，原因と結果の変数と名義尺度・順序尺度の組合わせによって，どのようにグラフ化したら良いかを考える．

	結果変数	
原因変数	名義尺度	順序尺度
名義尺度	性別/機能 表示 1.12	業種/IT 化度 表示 1.14
順序尺度	年齢階層/所有 表示 1.10	保守満足/総合満足 表示 1.17

結果変数が順序尺度である場合は，積み上げ棒グラフを用いるのが一般的である．

積み上げ棒グラフを用いると，総合満足度が3以上の割合などを容易に読み取ることができる．

結果変数が名義尺度である場合には，携帯電話の機能のように，カテゴリー毎に棒を並べる．名義尺度はカテゴリーを並べる順序に意味がないので，棒の配列順序に工夫する．携帯電話の機能では，割合の多い順としている．このように，順序を工夫することにより，分かりやすいグラフができる．結果変数が名義尺度である場合に，積み上げ棒グラフを用いるのは好ましくない．

原因変数が順序尺度である場合は，配列順序は自動的に定まる．名義尺度の場合は，配列順序に工夫する余地がある．パソコン導入率の場合には，第1，第2，第3 産業の順とか，導入の進んでいる業種から順に並べるなどが考えられる．

データをどのようにグラフ化するかを，いろいろな場合について説明するとそれだけで1冊の本になる．受講生はここに挙げた例を参考にして，工夫してほしい．

本日のまとめ

クロス表の活用は，市場調査や世論調査の分野では欠くことができない．あなたは，クロス表やグラフからどのような情報を読み取り，どのような感想を持ったであろうか．漠然と眺めるのではなく，情報を読み取るという姿勢が大切である．明日の内容も，あなたが分析者であったらどのように解析を進めるか，という気持ちを持って読み進むと興味がわくであろう．

1.4　2つの質的変数の関係（2）

(1)　縦・横のパーセント

ところで，この顧客満足度の例では，サービスマンの保守状況は，総合満足度の形成に影響を与える「原因」系列の変数と位置づけられる．一般に，2つの変数のうち一方を原因，または先行する要因[7]と見なして，その違いが他方に及ぼす影響・効果を見ようとする場合には，表示1.16のように原因と見なす方の変数を表側に置いて横のパーセントをとることが多い．

ところで表示1.10（p.15）では，携帯電話の所有率が年齢層によってどのように異なるかを見たが，このクロス表について，縦の構成比を計算したのが表示1.18である．

この比率は，携帯電話所有者の年齢構成，すなわち，所有者のプロフィール（Profile, 断面図）を意味していると考えて良いであろうか．例えば，携帯電話所有者のうち約40%が40歳以上であると理解して良いであろうか．

このような解析が意味をもつのは，調査対象者の年齢分布が，母集団の年齢分布を代表するときのみである．

表示1.18の年齢別の全体を見ると，年齢分布が第1単元の表示1.4と ややずれており，また，10歳代，60歳以上が含まれていない．

[7] それが真に原因であるかどうかはここで問題としていない．分析の意図が，それを原因ととらえるということだけを意味している．

表示 1.18: 携帯電話の所有率：縦の%表

	全体	所有	非所有	所有率
20歳代	416	382	34	27.0%
30歳代	513	444	69	31.4%
40歳代	481	309	172	21.9%
50歳代	590	279	311	19.7%
総数	2000	1414	586	100.0%

したがって，ここに計算したプロフィールによって判断するのは危険であろう．

(2)　3元クロス表の活用

いままで扱ってきたクロス表は，すべて2つの変数の関係を示すものであった．

現実のデータでは，3つの変数を同時に取り上げて，解析しなければ正しい結論を出せない場合がある．具体的な例について説明する．ただし，数値は仮想値である．

運転免許を所有しているある集団における運転時の事故率を想定する．ある期間における全体での事故率（19.0％）が明らかになったとき，次には，事故率が運転者の持つどのような要因で違うのか，と考えるだろう．そこで，分析が可能な範囲で運転者の特性と事故率とのクロス表を作り，その関連をみることにする．

まず，性別や走行距離が事故率と関連しているのではないかと仮説を持った場合，表示1.19 の2つのクロス表を作成する．走行距離は量的変数であるが，ここでは５万キロ以上と５万キロ未満の2値の質的変数に変換している．

この2つのクロス表は，女性よりも男性で，走行距離が短い者よりも長い者で事故率が高いことを示している．この2つの表から，事故率は走行距離が長い男性で最も高く，走行距離の短い女性では最も低い，と判断して良いだろうか．性別と走行距離の間に，関連はないのであろうか．

性別，走行距離という2つの要因を同時に取り上げて，事故率の変化を見たい．そのためには，男女別の表をさらに走行距離の長短別に分けてクロス表を

表示 1.19: 男女別，走行距離別の事故率

	総数	無事故	事故	事故率
計	1000	810	190	19.0%
男性	500	364	136	27.2%
女性	500	446	54	10.8%
5万km未満	520	482	38	7.3%
5万km以上	480	328	152	31.7%

作り，それを，事故と無事故に分けて，それぞれについて事故率を求めたのが表示1.20の表である．

表示 1.20: 男女別，走行距離×事故の有無の2元表

	総数			無事故			事故			事故率		
5万km	未満	以上	計	未満	以上	計	未満	以上	計	未満	以上	計
男性	104	396	500	94	270	364	10	126	136	9.6%	31.8%	27.2%
女性	416	84	500	388	58	446	28	26	54	6.7%	31.0%	10.8%
計	520	480	1000	482	328	810	38	152	190	7.3%	31.7%	19.0%

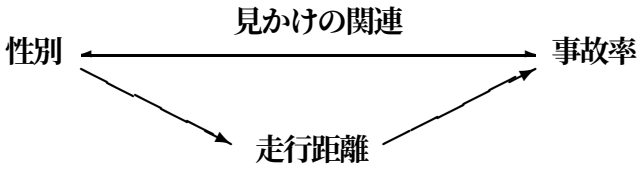
このような表を 3元クロス表 と呼ぶ．この3元クロス表によって，「走行距離別にみた男女の事故率はほぼ同率である」という新しい事実を知ることができる．先の仮説 「事故率は走行距離が長い男性で最も高く，走行距離の短い女性では最も低い」 が誤りであるとともに，性別による事故率の差異は見かけ上のものであり，性別と走行距離の間の関係，すなわち，「男性では走行距離の長い者が多く，女性では逆である」という関係 によって作りだされていることが分かる．

このように，クロス表に追加的な要因（層別因子）を加えることで，見かけの関連が明らかになることがある．

表示1.20の3元クロス表の表わし方を変え，走行距離によってサンプルを分けて，それぞれでの男女別クロス表を作成すると，表示1.22が得られる．

さらに，JMP を使って棒グラフを描くと，表示1.23 が得られる．

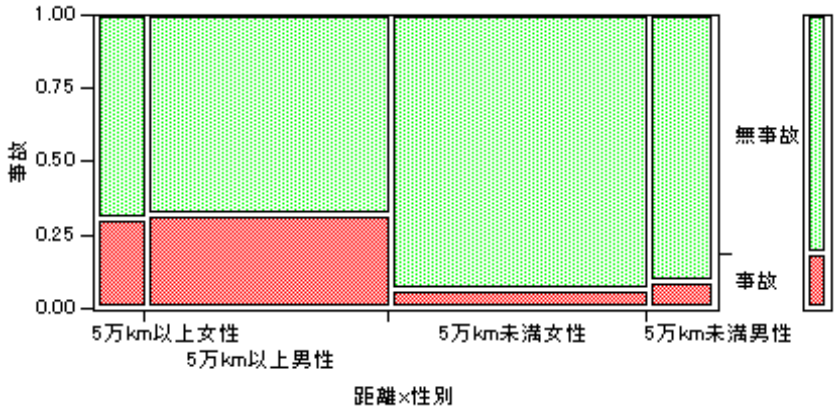
表示 1.21: 見かけの関連



表示 1.22: 走行距離別，男女別を組合わせた事故率

		総数	無事故	事故	事故率
計		1000	810	190	19.0%
5万km未満	男性	104	94	10	9.6%
	女性	416	388	28	6.7%
5万km以上	男性	396	270	126	31.8%
	女性	84	58	26	31.0%

表示 1.23: 走行距離別，男女別を組合わせた事故率（JMP による）



このグラフを見ると，男女で走行距離に大きな違いがあり，走行距離別に事故率を見ると，男女でほとんど差のないことが理解できるであろう．

表示1.22のクロス表（実質上は3元クロス表である）をみると，表示1.19に示されていた性別と事故率の関係が，消されている．このように全体での 2元

表と層別したクロス表（3元クロス表）の間で相反する傾向が示されることがあり，これを シンプソンのパラドックス と呼ぶ.

■　演習3　ある手術の成功率を大学病院と一般病院で比較した結果，次のようになった（仮想データ）.

	手術数	成功数	失敗数	失敗率 (%)
大学病院	200	190	10	5%
一般病院	100	97	3	3%

手術経験が多く，技術が高い大学病院の成功率が高いと予想したが，結果は，大学病院の方が成功率が低い.

この結果をどのように考えたら良いだろうか.　■

(3)　クロス表を作成する方針（集計の計画）

クロス表を作成するにあたって，やみくもに変数同士を掛け合わせるのは有益ではない．分析目的にそった集計計画を立てる必要がある．2元表では，説明されるべき変数に対して，基本属性はじめ，その変数に差異を生み出している要因（原因変数）の候補が掛け合わされる．集計の結果，差異がないことが明らかになった場合も，意味のある情報になる．原因と思われる変数相互で関連性がある場合には，別々にクロス表を作成するのではなく，3元クロス表など同時に考慮される集計表を作ることを考えるのが良い．さらに4元クロス表，それ以上の多元クロス表が考えられるが，あまりに変数が多くなると表のセルごとの数字が小さくなるので，標本観察データに関しては誤差が相対的に大きくなりがちである．構成比率を算出する分母を考えた上で，集計計画を立てておく必要がある.

本日のまとめ

クロス表の活用について，さらに深く学習した.

通常は，原因変数の分類ごとに結果変数の比率を求める．逆にすると，(2) で説明したように，誤った結論を導く危険があるので注意が必要である.

「見かけの相関」と「シンプソンのパラドックス」については，自分のデータを解析するとき，十分に配慮することが大切である．

クロス表に示された結果をどのように解釈するか，その奥行きを感じてもらえたであろうか．

1.5　2つの質的変数の関係（3）

(1)　連関係数

2行2列のクロス表で示された変数間の関連の強さを示す量として，連関係数を用いることがある．

2つの消費財（例えば，液晶テレビ と DVD プレイヤー）の所有の有無を調査した結果も前と同様に，表示1.24のように表わすことができる．

表示1.24: 2つの消費財の所有調査結果

記号				数値例			
	yes	no	計		yes	no	計
yes	a	b	$T_1.$	yes	12	16	28
no	c	d	$T_2.$	no	30	94	124
計	$T._1$	$T._2$	$T..$	計	42	110	152

両方を所有している人数が a，両方とも所有していない人数が d である．

もし，$a:b=c:d$ すなわち，$ad-bc=0$ ならば，両者の連関はないといえる．また，一方を所有していれば他方も所有している割合が多い（これを 正の連関 があるという）のであれば，$ad-bc$ は正の値をとり，逆の場合（負の連関）は負の値をとる．したがって，$ad-bc$ の符号と大きさで連関の方向と度合を見ることができるであろう．

ただし，$ad-bc$ の符号と連関の方向を見ることができるのは，上の例のように2つの質的変数におけるカテゴリーの並び順に意味があるときだけである．

例えば，性別と液晶テレビの所有の有無に関するクロス表では，連関の方向を考えることに意味がない．

連関係数Qは，

$$Q = \frac{ad - bc}{ad + bc} \tag{1.1}$$

と定義される．

Qは-1と$+1$の間にある．

表示1.24のデータの連関係数は，

$$Q = \frac{12 \times 94 - 16 \times 30}{12 \times 94 + 16 \times 30} = \frac{1128 - 480}{1128 + 480} = \frac{648}{1608} = 0.40 \tag{1.2}$$

となる．

■　**演習４**　表示1.24のデータで，1行目と2行目（または1列目と2列目）を入れ替えると，連関係数の符号が逆になることを確かめよ．

1行目（または2行目，または1列目）の件数を何倍しても連関係数の値は変化しないことを確かめよ．

a, b, c, d のいずれかが 0 であるとき，連関係数は $+1$ または -1 となることを確かめよ．■

■　**演習５**　300人の成人にあるメロディを聞かせたとき，メロディの印象についていくつかの項目を回答してもらった．その中から幸福な印象と牧歌的な印象についてクロス表を作ると以下のようになった．

	幸福感がある	幸福感があるとはいえない
牧歌的である	42	17
牧歌的とはいえない	67	174

牧歌的な印象が幸福な印象と関連があると考えられるか．■

(2)　疫学調査におけるクロス表の活用

クロス表は質的データ同士の因果関係を探求しようと考える場合，様々な分野で活用される．その代表的な分野の一つである疫学調査について述べておく．

疫学調査では，疾患の危険因子や原因を探ることが目的になる．肺がんと喫煙の関係が疑われるときには，次のような調査方法でデータを集める．

(i)　60 歳代の男性の集団に対して肺がん検診を行い，同時に喫煙量などを調べる．

(ii)　ある地域で行ったがん検診で肺がんと診断された人に対して，診断されなかった人を比較対照群とし，喫煙量などを 過去に遡って調べる

(iii)　ある地域で健康調査を実施し，喫煙量などのデータをとっておく．喫煙有無などで疾病・死亡率がどのように異なるか，追跡調査 を行う．

(i) の方法は，疾患や危険因子の有無に関係なく，特定の集団を調査の対象とする．ある一時点で調査を行うので,「断面研究」(cross-sectional study) と呼ばれる．この例では，肺がん患者の出現率は低いことが予想されるので，多くのサンプルが必要である．

(ii) は，疾患の有無で集団を設定して，過去に遡って危険因子の有無を調べ，その割合を比較するので，後ろ向き調査 という．専門用語では，症例対照研究 (case control study) と呼ばれる．

(iii) は，危険因子の有無で集団を分け，その後の疾病・死亡率を比較するので，前向き調査 という．専門用語では，コホート研究 (cohort study) と呼ばれる．長期にわたる観察が必要になることもある．

(3)　オッズ比

前項で取り上げた喫煙による肺がん罹患率の増加を例として，喫煙が肺がん罹患率にどれだけ影響を与えるかを定量的に表わす方法を導く．

喫煙と肺がんの関係を調べた断面研究で 10 万人を調査した場合を想定する．調査の結果は，表示 1.25 のクロス表の形にまとめられる．

表示1.25: 2*2 **クロス表**

喫煙	肺がん 有	無	計	肺がん 有	無	計	オッズ
喫煙者	a	b	$T_{1\cdot}$	1500	28500	30000	0.0526
非喫煙者	c	d	$T_{2\cdot}$	1400	68600	70000	0.0204
計	$T_{\cdot 1}$	$T_{\cdot 2}$	$T_{\cdot\cdot}$	2900	97100	100000	

喫煙者，非喫煙者と全体での肺がん罹患率は

$$\text{喫煙者}\quad \frac{a}{T_{1\cdot}} = \frac{1500}{30000} = 0.05$$
$$\text{非喫煙者}\quad \frac{c}{T_{2\cdot}} = \frac{1400}{70000} = 0.02$$
$$\text{全体}\quad \frac{T_{\cdot 1}}{T_{\cdot\cdot}} = \frac{2900}{100000} = 0.029$$

である．

これから，喫煙によって罹患率が $0.05/0.02 = 2.5$ **倍になると考えられる．**

次の比率

$$\text{喫煙者}\quad \frac{a}{b} = \frac{1500}{28500} = 0.0526$$
$$\text{非喫煙者}\quad \frac{c}{d} = \frac{1400}{68600} = 0.0204$$

を用いると，より高度な解析につながるという利点がある．この比を オッズ (Odds) **と呼ぶ**[8]**．この値が，表示**1.25 **のオッズの欄に求められている．**

喫煙者のオッズ 0.0526 **が非喫煙者のオッズ** 0.0204 **の何倍になるかで，喫煙の危険率を表わすことができる．この比**

$$\frac{0.0526}{0.0204} = 2.58$$

を オッズ比 と呼び，肺がんに対する喫煙の影響度の指標とする．

ところで，前項で説明した疫学調査では，疾患と原因の因果関係の究明に主眼があり，母集団全体（例えば，日本人全体，男性全体，60**歳以上の人全体な**

[8] **競馬・競輪などの賭けの世界でも オッズ という言葉が用いられる．これは，本文に説明したオッズとは別物である．**

ど）での罹患率を明らかにすることを目的としない場合が多い．前の節で述べてきたクロス表と，この点で異なっていることに注意しなければならない．

とくに後ろ向き調査では，該当する患者とそうでない者を何人選ぶかは，研究者に委ねられているから，罹患率を求めることに意味はない．

前向き調査の場合も，危険因子の有無で集めるサンプルを任意に決めることが多いので，全体での罹患率の算出はできない．

ところが，このようにサンプルの大きさを決めた場合でも，オッズ比は一定であるという良さを持っているので，要因の結果に与える影響の強さを示す指標としてすぐれている．

下記の例で説明しよう．

まず，(ii) の後向き調査を行い，肺がん患者を200人，患者ではない者をその2倍の400人選んだとする．この場合の，患者・患者以外のそれぞれの喫煙率を表示1.25と同様とすると，表示1.26左のようなデータが得られるであろう．

表示1.26: サンプリングの異なるデータ

	後向き調査				前向き調査			
	肺がん				肺がん			
喫煙	有	無	計	オッズ	有	無	計	オッズ
喫煙者	104	118	222	0.881	10	190	200	0.053
非喫煙者	96	282	378	0.340	4	196	200	0.020
計	200	400	600		14	386	400	

このデータのオッズは，0.881, 0.340 と前の結果と大きく異なるが，その比は2.6 で前と同じ値が得られる．

(iii) の前向き調査として，喫煙者と非喫煙者からそれぞれ200人ずつ選んだとする．この場合，それぞれの肺がん罹患率を表示1.25と同様とすると，表示1.26右のようなデータが得られる．

この場合のオッズ比は 2.6 で，表示1.25と同じである．

以上3つの 2*2 クロス表から求めたオッズ比はすべて同じ値となった．

これは，オッズ比の持つ特長的な性質であって，サンプリング方法が異なっても，同じ結果が得られるという利点がある．

これが，疫学や，医療，薬の効果の評価などにオッズ比が広く用いられる理由である．また，多変量解析の手法であるロジスティック回帰分析の基礎となっている[9]．

(4)　オッズ比と連関係数の関係

表示1.24 のデータを使って，2つの指標の関係を調べる．

$$\text{オッズ比} = \frac{ad}{bc} = \frac{12 \times 94}{16 \times 30} = 2.35$$

$$\text{連関係数}\quad Q = \frac{\text{オッズ比} - 1}{\text{オッズ比} + 1} = 0.40$$

となり，式(1.2)(p.29)と一致する．

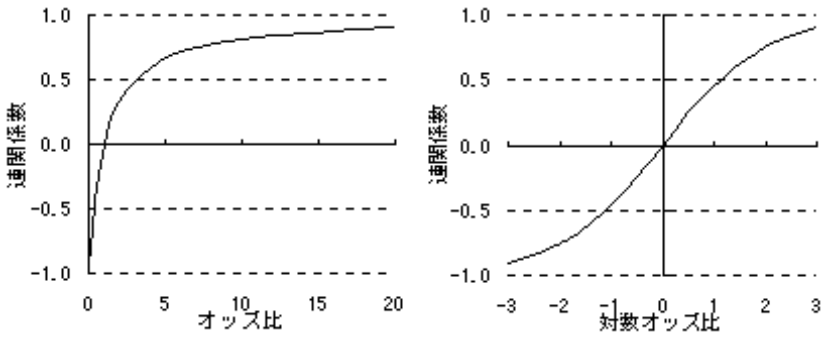
連関係数は -1 と $+1$ の間の値を取り，まったく関連がないときは 0 となる．

それに対して，オッズ比は 0 と ∞ の間の値を取り，割合が等しいときには 1 となる．オッズ比の対数を用いる場合がある．そのときは，$-\infty$ と $+\infty$ の間の値を取り，割合が等しいときには 0 となる．

両者の関係を表示1.27に示す．

[9] ロジスティック回帰分析については,「多変量解析実務講座」の第4単元に詳しく説明されている．

表示 1.27: オッズ比と連関係数の関係



本日のまとめ

クロス表については，前節までにも説明してきた．今日は，2×2 クロス表の別の見方について説明した．

前半では，2 つの質的変数の間の関係の強さを定量的に表わす 連関係数 を学んだ．

後半では，疫学の分野固有の問題を取り上げた．その分野で広く用いられている オッズ と オッズ比 について学んだ．疫学では，原因が結果として表われるまでには長い期間を要する．そのために考え出されたのが，前向き調査 と 後向き調査 であった．

2 量的データの記述（1）

ある特性について調査とか測定を行って得たデータの集まりがあったとき，その集団の構造を端的に表現するようないくつかの代表値を求めることなどを，集団構造の記述 という.

それらの代表値の中で最もよく知られているのが 平均値 である．平均値は広く用いられ，特別に説明するまでもないと思われるかもしれないが，どのような考えで平均値が導かれるかを説明する．そこで用いられる方法は統計の基本的なものである.

2.1 平均値

ここで5人の男性の身長のデータ 169，174，160，165，172cm を考える.

(1) 代表値

一般にn個のデータを，$x_1, x_2, \ldots, x_n$，あるいは，x_i $(i = 1, 2, \ldots, n)$ と書く．この $n = 5$ 人の身長の代表値を仮に a という記号で表わす.

a が代表値の役割を果たすためには，個々の値と代表値の距離が全体として小さくなければならない.

代表値の候補 a を 154cm から 182cm まで順次変えて[1]個々の値との差を求めた結果を表示2.1 に示す.

$x_i - a$ はプラスとマイナスの値を取るので，そのまま合計した値（「和」の行に求められている）で，代表値としての良さを評価することはできない.

そこで，絶対値を取ったり，2乗したりしてプラスの値として合計する（「絶対値の和」と「2乗の和」の行に求められている）.

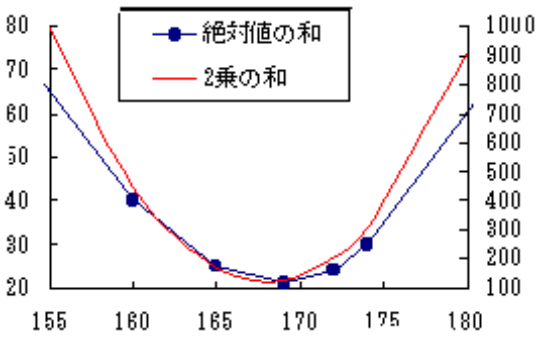
[1] 実際の観測値の5点に，観測値外の3点を追加した.

表示2.1: 代表値 a の選択(計算)

	x-a							
x ＼ a	154	160	165	168	169	172	174	182
169	15	9	4	1	0	−3	−5	−13
174	20	14	9	6	5	2	0	−8
160	6	0	−5	−8	−9	−12	−14	−22
165	11	5	0	−3	−4	−7	−9	−17
172	18	12	7	4	3	0	−2	−10
和	70	40	15	0	−5	−20	−30	−70
絶対値の和	70	40	25	22	**21**	24	30	70
2乗の和	1106	446	171	**126**	131	206	306	1106

横軸に a を，縦軸に「絶対値の和」（左目盛り）と「2乗の和」（右目盛り）を取ったグラフを表示2.2に示す.

表示2.2: 代表値aの選択(グラフ)



絶対値の和のグラフはデータのあるところ（● で表わす）で折れる折れ線となり，169cm で最小となる.

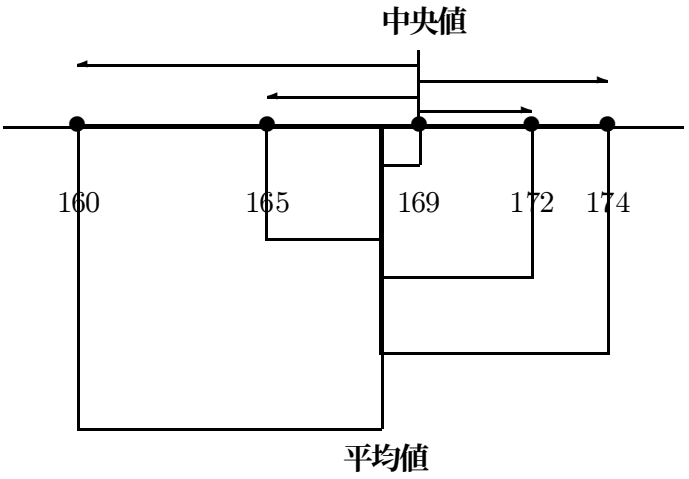
2乗の和のグラフは2次曲線（放物線）となり，168cm で最小となる.

ここに得られた2つの値の性質を調べて見よう.

表示2.3では，5つの身長の位置が黒丸で示されている.

$a = 169$cm **から各点までの距離が上半分に矢線で示されている．最小の身長**

表示2.3: **2つの代表値の性質**



から a **までの距離と，最大の身長から** a **までの距離の和** L **は，** a **をこの範囲内** $(160 \sim 174)$ **で変化させても変化しない．両側から**2 **番目の身長から** a **までの距離の和も同様である．**

したがって，絶対値の合計を最小にするためには，どちらから数えても3 **番目である**169cm **を** a **とすれば良いことが分かる．このような考え方で導かれた代表値を 中央値（またはメジアン，中位数）と呼び，** $\tilde{x}$ **（エックス・波 または エックス・チルド と読む）で表わされる．**

n **が偶数のときは，中央の**2**つの値の平均値が用いられる．**

中央値については，§2.4 **で改めて説明する．**

表示2.3 **の下は，偏差の**2 **乗の和を最小とする** $a = 168$ **と各点までの距離を辺とする正方形を描いたものである．これらの面積の合計を最小とする** a **が**168cm **である．**

$a = 168$cm **としたとき，** $x_i - a$ **は** 1, 6, –8, –3, 4 **で，その合計は** 0 **になる．これから，**168cm **より身長の高い人の身長を低い人に移すと，平らに均す（ナラス）ことができる．すなわち，**168cm **は** 5 **人の身長の平均となっている．**

(2)　平均値

$x_i(i = 1, 2, \ldots, n)$ の合計 T（Total の頭文字）は,

$$T = x_1 + x_2 + \ldots + x_n = \sum_{i=1}^{n} x_i \tag{2.1}$$

で，平均値を $\overline{x}$（エックス・バーと読む）で表わせば,

$$\overline{x} = T/n = \sum_{i=1}^{n} x_i \Big/ n \tag{2.2}$$

となる.

男性5人の身長のデータにこの記号を当てはめると,

$n = 5,\ x_1 = 169,\ x_2 = 174,\ x_3 = 160,\ x_4 = 165,\ x_5 = 172$ で,

$T = \sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5 = 840$ となり,

$\overline{x} = T/n = 840/5 = 168$

が得られる.

$\overline{x}$ で表わしたのは,「x という量に関する平均値」の意味であって，それ以外に y という量があれば，その平均値は $\overline{y}$ で表わされる．例えば，5人の女性の身長のデータを,

$$y_1 = 167,\ y_2 = 150,\ y_3 = 161,\ y_4 = 158,\ y_5 = 164$$

とおけば,

$$\overline{y} = 800/5 = 160$$

である.

(3)　最小2乗法

$(x_i - a)^2$ の和を最小とする代表値が平均値になることは以下のようにして導くことができる.

$(x_i - a)^2$ の和を S で表わす.

$$S = (x_1 - a)^2 + (x_2 - a)^2 + \ldots + (x_n - a)^2 = \sum_{i=1}^{n} (x_i - a)^2 \tag{2.3}$$

式(2.3) **は次のように変形することができる.**

$$\begin{aligned} S &= \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i^2 - 2ax_i + a^2) \\ &= \sum_{i=1}^{n} x_i^2 - 2a\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} a^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2an\overline{x} + na^2 \\ &= \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) + n(a^2 - 2a\overline{x} + \overline{x}^2) \\ &= \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) + n(a - \overline{x})^2 \end{aligned} \tag{2.4}$$

式(2.4) **の第**1 **項には** a **が含まれず一定であり，第**2 **項は必ず正であるから，** S **が最小となるのは第**2 **項が** 0 **となるときである．これから，**

$$a = \overline{x} \tag{2.5}$$

が得られる．すなわち，代表値 a **は平均値** $\overline{x}$ **となる**[2]**.**

この考え方は 最小2 **乗法** (least square method) **と呼ばれ，統計の中で重要な位置を占めるものである.**

個々の値 x_i **と平均値** $\overline{x}$ **との差** $x_i - \overline{x}$ **を 偏差** (deviation) **と呼び，** e_i **で表わすことにする.**

偏差の合計は,

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - n\overline{x} = T - T = 0$$

から，0 **になることが分かる（表示**2.1 **で，** $a = 168$ **の列の和が** 0 **となっていることを確認せよ）.**

---

[2] **式**(2.5) **は，微分を使うともっと簡単に導くことができる．その具体的な方法は** §2.5 **理論的補足で説明する.**

(4)　Excel による計算

平均値は，=AVERAGE（データの範囲）で計算される．

具体的な例は次の節で説明する．

本日のまとめ

第2章では，重要な統計的な指標とその意味について学習する．今日は，データの集まりを定量的な一つの指標で表現する方法として平均値を学んだ．受講生は，平均値が代表値として好ましい性質を持っているかについて理解されたと思う．その上で，Excelの関数を使って，平均値が正しく計算できることも体験されたであろう．

2.2　バラツキの大きさの定量

(1)　平方和

表示2.1のデータについて，5人の身長のバラツキの大きさを考える．

表示2.6から直観的にわかるように，一つの集団のバラツキが大きいということは，偏差 e_i が大きいことである．したがって，表示2.3の正方形の面積の和 $S = \sum_{i=1}^{n} e_i^2$ でバラツキの大きさを測ることが考えられる．

表示2.1のデータについて，平均値，偏差，偏差の2乗，S をExcelで計算した結果を表示2.4に示す．

E列にはB列のセルに入力された式が表示されている．

COUNT関数で n が，SUM関数で 合計 T が求められる．平均は，$\overline{x} = T/n$ =B10/B9 でも求められるが，AVERAGE関数で x から直接求めることができる．

偏差 $e_i = x_i - \overline{x}$ がC列（eの列）に，その2乗がD列に求められている．これらの合計が 合計の行 に求められている．偏差 e_i の合計が0となり，D10 の

表示2.4: 男性の身長の平均と平方和

	A	B	C	D	E
3		x	e	e^2	
4	1	**169**	1	1	
5	2	**174**	6	36	
6	3	**160**	-8	64	
7	4	**165**	-3	9	
8	5	**172**	4	16	
9	個数(n)	5			=COUNT(B4:B8)
10	合計(T)	840	0	126	=SUM(B4:B8)
11	平均(x-bar)	168			=AVERAGE(B4:B8)
12	平方和(S)	126			=DEVSQ(B4:B8)

セルに求められている 偏差の2乗 e_i^2 の合計 126 が 平方和 （偏差平方和 ともいう）で，S で表わされる．

このような過程を経ないでも，Excel には，元のデータから直接平方和を求める関数 DEVSQ が準備されている．この関数を使って求めた S が B12 の値である．

平方和 S は バラツキの総量 を表わすものである．

(2)　平均平方

いま，表示2.5 のような2組のデータを考えよう．

集団I は，前に取り上げた 5 人の男性の身長のデータの集まりで，集団II は別の 9 人の男性の身長のデータの集まりである．この2つの集団をそれぞれ一直線上での分布を描いてみると，表示2.6 のようになる．

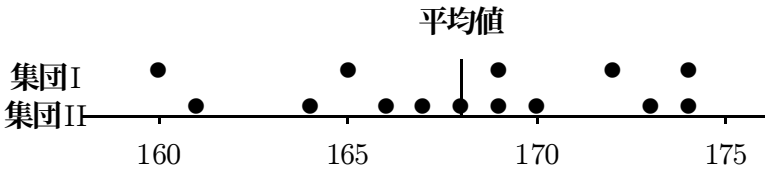
表示2.6 を見ると，2 つの集団は平均値が同じであるにもかかわらず，何か違いがあることに気づくであろう．データの個数の違いの他に，点のバラツキの状態に違いがあるといえよう．このことから，平均値とは別に平均値では表現できない，バラツキ（分布の散布度）の程度を表現する代表値を考える必要のあることが分かる．

表示2.5 の集団I，集団II について平方和 S を求めた．表示2.6 で見た限りで

表示 2.5: 2組の男性の身長のデータの解析

	G	H	I	J
3		集団I	集団II	
4	1	169	161	
5	2	174	164	
6	3	160	166	
7	4	165	167	
8	5	172	168	
9	6		169	
10	7		170	
11	8		173	
12	9		174	
13	個数	5	9	
14	平均	168	168	
15	平方和	126	136	
16	自由度	4	8	=H13-1
17	分散	31.5	17.0	=VAR(H4:H12)
18	標準偏差	5.6	4.1	=STDEV(H4:H12)

表示 2.6: 2つの集団のバラツキの状態



は，集団I の方がバラツキが大きかったにもかかわらず，S の値は集団II の方が大きい．それは集団IIの方が，データの個数が多いからである．そこで，S をデータの個数 n で割って 1個当たりのバラツキ とすれば，データの個数にかかわりなく，バラツキの比較ができる量が得られるであろう．その量は，それぞれ $126/5 = 25.2$ と $136/9 = 15.1$ という値になる．

以上でバラツキの尺度を構成する過程が分かったが，最後の部分について，n で割る代りに $(n-1)$ で割った方が良いことが知られている．$(n-1)$ は 自由度 (Degree of Freedom) と呼ばれ，第3,4 単元で重要な役割を果たす．なぜ $(n-1)$ で割るかについて，厳密な証明は数理統計学の知識が必要なので，専門書にゆ

ずり，§2.5(3) に概念的な説明をする．

このようにして得られたバラツキの尺度を 分散 (Variance) または 平均平方 (Mean Square) と呼び，V または s^2 で表わす．

(3) 標準偏差

平方和と分散の単位は，その計算過程に2乗するという操作が入っているので，もとのデータの単位の2乗になる．例えば，身長のデータ（単位は cm）では，平方和と分散の単位は cm^2 である．バラツキの尺度をもとのデータと同じ単位で考えたいときには，分散の平方根を用いれば良い．それを 標準偏差 (Standard Deviation) と呼び，s で表わす．以上をまとめて定義式に表わし，さらに集団Iの分散と標準偏差を求めると，

$$\text{分散}: V = \frac{S}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{126}{4} = 31.5(cm^2) \tag{2.6}$$

$$\begin{aligned}\text{標準偏差}: s = \sqrt{V} &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ &= \sqrt{\frac{126}{4}} = \sqrt{31.5} = 5.6(cm)\end{aligned} \tag{2.7}$$

となる．

表示2.5では，Excel 関数 =VAR(データの範囲)，=STDEV(データの範囲) を使って，分散と標準偏差を求めている．

これから，平方和は集団II の方が大きかったが，分散，標準偏差は集団IIの方が小さくなっている．

(4) 変動係数

2つの集団の感覚的なバラツキが同じであっても標準偏差が異なることがある．例えば，成人男子の身長のデータの集まりと，小学1年生男子の身長のデータの集まりとでは，それぞれに同程度の大小のバラツキはあるにしても，小学生の方のバラツキは，数量的には大人のそれよりは少ないはずである．それはそもそも集団の平均値が異なることに関係している．そこで，標準偏差 s と平

均値 $\overline{x}$ の比を用いることがある．これを 変動係数 (Coefficient of Variation) または相対標準偏差 といい，$CV = s/\overline{x}$ で表わす．変動係数は % で表わす場合が多い．

例えば，集団Iの身長のデータでは，$CV = 5.6/168 = 0.03 = 3\%$である．

変動係数が役立つもう一つの例をあげよう．

成人男子の身長と体重の平均と標準偏差が次のようであったとする（仮の数値）．

	平均	標準偏差	変動係数
身長	170cm	8cm	4.7%
体重	63kg	7kg	11.1%

身長と体重は単位が異なるので，両者の標準偏差を直接比較することはできない．しかし，変動係数にすると単位がなくなるので，比較が可能となる．体重の変動係数は身長の変動係数の2倍以上であることが分かる．

変動係数が意味を持つのは，x の値が非負で，0 が意味を持つ（比例尺度）の場合のみである．

例えば，摂氏の温度で変化範囲が0℃以上であっても，変動係数は意味を持たない．$\mu = 20$℃，$\sigma = 10$℃ のとき，変動係数は $10/20 = 50\%$ である．アメリカでは温度を華氏で表わすので[3]，華氏に変換すると，平均と標準偏差は $\mu = 32 + 1.8 \times 20 = 68^\circ$ F，$\sigma = 1.8 \times 10 = 18^\circ$ F となり，変動係数は $68/18 = 9\%$ となる．

摂氏は水の氷点を0度，沸騰点を 100度 と決めたもので，特別の意味を持たない．華氏も同様である．したがって，温度変化の変動係数は意味を持たない．それに対して，絶対温度（摂氏温度+273）は 0 度が物理的な意味を持つので，変動係数が意味を持つ．

それに対して，身長 $\mu = 170cm, \sigma = 8cm$ をインチとすると $\mu = 170/2.54$ インチ, $\sigma = 8/2.54$ インチ となる．両者の比である変動係数は変化しない．身長は 比例尺度 だからである．

[3] 摂氏と華氏の間には次の関係がある．

摂氏	-17.8 度	0度	37.8 度	100度	x 度
華氏	0度	32度	100 度	212度	$32 + 1.8x$ 度

本日のまとめ

昨日は，平均値について学習したが，データは，平均以外にもバラツキという情報を持っている．バラツキの程度を測る指標として標準偏差があるが，その意味について，平方和，平均平方と順を追って説明した．

今日は，平方和（バラツキの総量）から平均平方（1 個あたりのバラツキ）を求めるとき，自由度 $n-1$ で割ることを学んだ．自由度は統計の初学者にとって最初の難関であるといわれている．自由度については，§2.5 の補足でさらに詳しく説明するので，十分に理解してほしい．

また，バラツキの程度を比較する方法として変動係数による比較と変動係数の利用方法についても学習した．明日は，平均と標準偏差を使って，個々の値について調べてみよう．

2.3 偏差値と外れ値

(1) 偏差値

平均値と標準偏差が求められたならば，個々の値について検討する．

i 番目の値 x_i が平均値 $\overline{x}$ からどれだけ離れているかを表わす量 $e_i = x_i - \overline{x}$ が偏差であった．

偏差が大きいかどうかは，偏差の値を見ただけでは判断できない．

例えば，$e_3 = x_3 - \overline{x} = 160 - 168 = -8$ で，この人はかなり身長が低い．低さの程度を表わすのに，偏差を標準偏差で割った比が用いられる．この比は 偏差値 と呼ばれる．i 番目の偏差値を z_i とすると，

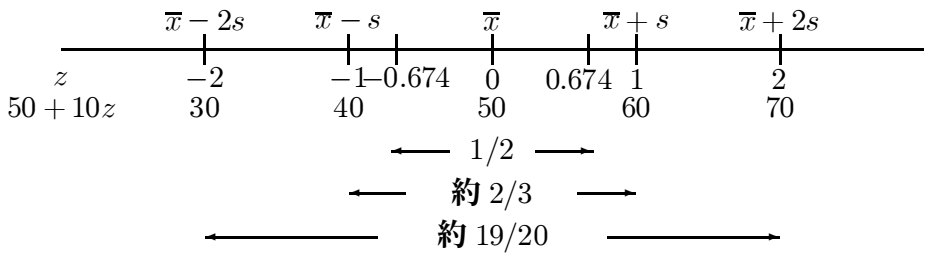
$$z_i = \frac{x_i - \overline{x}}{s} \tag{2.8}$$

$$z_3 = \frac{160 - 168}{5.6} = \frac{-8}{5.6} = -1.43$$

となる．

第1単元 §3.4(3) で説明したように，正規分布に従うとき，偏差値が ± 1.0 の範囲内に入る割合は約 $2/3$ であり，± 2.0 の範囲外に出る割合は約 $1/20$ である．また，第1単元の表示3.11 の左下に 上側確率が $1/4$ となる z が 0.674 であることが示されている．

これらの関係を図にすると次のようになる．



上の図で，z の 50% が含まれる範囲は，$-0.674 < z < 0.674$ である．この数字 0.674 は後に用いられる数値で,「ろくでなし」と記憶しておくと良いであろう．

表示2.7 は，$n = 5 \times 8 = 40$ のデータである．まず，平均値を K4 に，標準偏差を K5 に求める．

表示2.7: 偏差値を求める

	A	B	C	D	E	F	G	H	I	J	K
3	データ										
4	48.0	51.0	50.0	50.0	54.0	55.0	49.5	58.0		平均	50.38
5	50.7	37.0	58.2	48.0	55.0	51.7	51.0	64.0		標準偏差	4.77
6	52.5	51.5	56.2	50.0	48.0	56.0	51.0	52.0			
7	48.0	44.0	46.7	45.0	43.7	48.0	47.0	42.0			
8	47.7	47.7	52.7	49.0	52.2	53.0	50.5	49.5			
9	偏差値										
10	-0.5	0.1	-0.1	-0.1	0.8	1.0	-0.2	1.6			
11	0.1	**-2.8**	1.6	-0.5	1.0	0.3	0.1	**2.9**			
12	0.4	0.2	1.2	-0.1	-0.5	1.2	0.1	0.3			
13	-0.5	-1.3	-0.8	-1.1	-1.4	-0.5	-0.7	-1.8			
14	-0.6	-0.6	0.5	-0.3	0.4	0.6	0.0	-0.2			

ついで，偏差値を求める領域の左上のセル A10 に

```
=(A4-$K$4)/$K$5
```

を入力し，右と下にコピーする．

(2) 外れ値

こうして偏差値を求めたならば，–2〜2 の範囲外のものを拾い出すと，特徴のあるデータを知ることができる．表示2.7 では，–2〜2 の範囲外のものは太字で表わしてある（画面では赤字になっている）[4]．

このようにして，拾い出された特に離れたデータは **外れ値** (Outlier) と呼ばれる．

ここでは $|z| > 2.0$ を拾い出したが，この限界値は状況に応じて適当に決められる．

限界値を 2.0 とすると，データの分布が正規分布であっても，約5%が外れ値と思われてしまう．したがって，n が大きいときには，限界値としてもっと大きな値を用いる．限界値を 2.6 とすると，データの分布が正規分布であるとき，外れ値となる確率は約1%となる．

このようなことを考慮して限界値を決めるべきであるが，通常は 2.5 以上または 3.0 以上が用いられる．

外れ値をたどって見ると，しばしば，測定・転記・入力ミスが発見される．したがって，データをコンピュータに入力したならば，外れ値の有無を調べることは極めて重要であり，データ解析の第一歩として絶対必要な過程である．

外れ値＝異常値 と考え，その観測値を機械的に除外して解析する人がいるが，これは誤りである．

異常と判断するためには，データの背景についての知見にもとづく判断が必要である．また，異常と判断された場合も，それを含む解析と，除いた解析の両方を実行し，両者を比較して総合的な判断をすべきである．

データ解析では，個々のデータから平均や標準偏差を使って集団構造の記述を

[4] 「条件つき書式」を用いる．具体的な方法は §2.5「Excel ヒント 1」参照．

すると共に，逆に，もとの個々のデータに戻って，検討することが大切である．

統計の専門家は「木を見て森を見ない人が多い．統計は森を見るための道具である」という．しかし,「森を見る」ことも大切であるが，もう一度「木」を見直すことも重要である．

なぜ外れ値が得られたかを探索することにより，新しい知見の得られることが少なくない．

§3.4 では，外れ値を発見する便利な方法である「箱ひげ図」について説明する．

本日のまとめ

偏差値について良い印象を持っていない受講生もいたであろう．本日は，偏差値の本来的な意味と使い方について説明を行った．偏差値により，データの集団から大きく外れた値を客観的に選び出すことができる．実際の問題では，平均や標準偏差によって，得られたデータから母集団を類推することも重要であるが，偏差値によって個々のデータの様子を調べることも大いに重要である．多くの問題では，集団から外れた個々観測値が悪さの原因になっており，それらについての処方箋が議論される（例えば，売上げが上がらない営業員をどのようにレベルアップさせるかといった議論や肥満に対する議論など）からである．

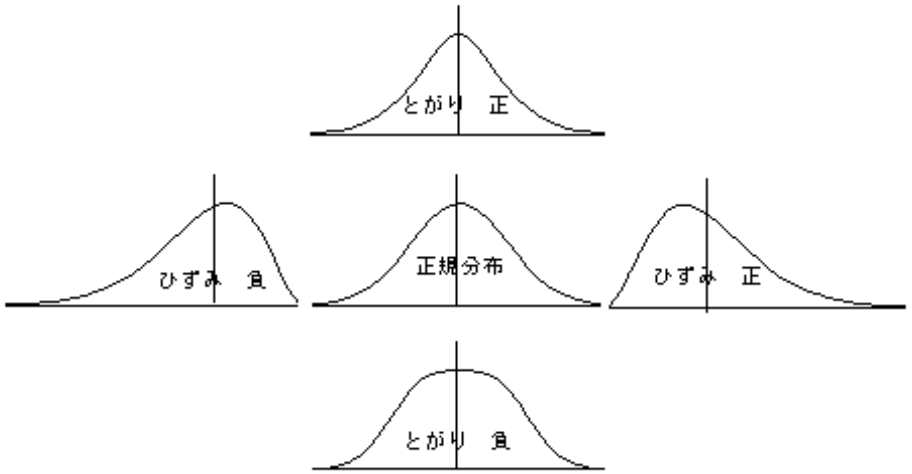
2.4　その他の指標

第1単元の §3 で，標準的な分布として正規分布について説明した．しかし，現実のデータは必ずしも正規分布には従わない．

表示2.8 の中央に正規分布の形を示す．

その左右には，左右非対称の分布を示す．また，正規分布の上には，左右対称であるが，正規分布よりも長い裾を引き，尖った分布が，下には，逆に，両裾が切れ，頭が丸まった分布が示されている．

表示2.8: 分布形とひずみ・とがりの関係



この図に示したように，正規分布から離れた分布があるとき，その分布が正規分布からどの方向に，どの程度離れているかを表わす指標が ひずみ と とがり である.

(1) ひずみ

偏差値の3乗の平均を考える．偏差値の2乗は必ず正になるが，偏差の3乗の符号は偏差の符号のままである．したがって，3乗の和は正負の値を取りうる．分布が左右に対称であれば，正と負がほぼ同数あり，打ち消しあって0に近い値になるであろう．それに対して，右の方（大きい方）に長い裾を引いていると，正の値の大きい偏差値の3乗は極めて大きい値を取り，平均は正の値を取るであろう.

このように考えて導かれた指標は ひずみ (または歪度) (Skewness) と呼ばれ，b_1 で表わされることが多い.

$$b_1 = \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s} \right)^3 \frac{1}{n} = \sum_{i=1}^{n} (x_i - \overline{x})^3 \frac{1}{ns^3} \tag{2.9}$$

(2)　とがり

同様の考えで，正規分布に比べて裾を引いているかどうか（頭がとがっているかどうか）を表わす指標は，式(2.9)で3乗の代わりに4乗すれば良さそうである．ひずみの場合と異なり，偏差値の4乗は負の値は取らない．正規分布の場合，n が大きくなると，4乗の平均は 3 に近づくという性質がある．正規分布と比較するために，3を引いた値を とがり （または尖度）(Kurtosis) と呼び，通常 b_2 で表わされる．

$$b_2 = \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s} \right)^4 \frac{1}{n} - 3 = \sum_{i=1}^{n} (x_i - \overline{x})^4 \frac{1}{ns^4} - 3 \tag{2.10}$$

ひずみ・とがりを求めるExcel関数は SKEW(データの範囲) と KURT(データの範囲) である[5]．この関数の使用例は表示2.9 に示されている．

表示2.9: ひずみ・とがり

	M	N	O	P
3	n	40		=COUNT($A$4:$H$8)
4	平均値	50.38		=AVERAGE($A$4:$H$8)
5	標準偏差	4.77		=STDEV($A$4:$H$8)
6	ひずみ	0.06		=SKEW($A$4:$H$8)
7	とがり	1.77		=KURT($A$4:$H$8)

ひずみ，とがりの絶対値が 1.5 よりも大きいときは，正規分布から外れていると考えて，§3.2 で説明する方法でデータをグラフ化して検討する必要がある．(1.5 は目安である)．

(3)　刈込み平均

n 個の観測値 x_i, $(i = 1, 2, \ldots, n)$ があるとき，それらの代表値として平均値 $\overline{x}$ が広く用いられることは既に述べた．もとの観測値がきれいな場合（正規分

[5] Excel は，ひずみ，とがりの計算に式(2.9)，(2.10) ではなく，改善した式（§2.5 補足で説明）を用いている．

布の場合）には平均値が最も良い代表値であることは数理統計学で証明されている．

しかし，現実の世界ではきれいなデータばかりではない．レフェリーの判断で勝敗が決まるスポーツ，例えば，フィギュアースケートや体操競技では，複数のレフェリーの採点を単純に平均するのではなく，最高点と最低点を除いた $n-2$ 人のレフェリーの採点を平均する．これは，選手と何らかの関係のあるレフェリーが不適切な採点をする可能性があるための「生活の知恵」であろう．

不適切なレフェリーがたくさんいるとすれば（そんなことはないであろうが，仮定の話として），採点を大きさの順に並べ，両側の2人，または3人のレフェリーを除いて平均をとることが考えられる．

このような平均を Trimmed mean (統計学辞典では 刈込み平均 と訳されている）と呼ぶ．

極端な場合，n が奇数であれば中央の採点を，偶数であれば中央の2人の採点の平均を求めることになる．これは 中央値 に他ならない．

刈込み平均は，とんでもない評点をつけたレフェリーがいてもその影響を受けにくい．このように，異常な値の影響を受けにくいことを 頑健性 がある（ロバスト Robust である）という．刈込む割合が増えれば頑健性が向上する．

Excel には，=TRIMMEAN(データの範囲，刈込む割合) という関数が準備されている．

2番目のパラメータは，両側からどれだけの割合を刈込むかを指定する．すなわち，両側からそれぞれ m 個を除いて平均するときは，$2m/n$ とする．例えば，$n=20$ で両側から3個を除くときは，$2\times 3/20 = 0.3$ とする．

(4) 中央値

中央値 $\tilde{x}$ は

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} \qquad n = \text{奇数}$$

$$\tilde{x} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} \qquad n = \text{偶数}$$

と表わすことができる．ここに，$x_{(i)}$ は小さい方から i 番目の観測値を表わす．

表示 2.10: その他の指標

	M	N	O	P
8	刈込み平均(1)	50.37		=TRIMMEAN($A$4:$H$8,2/$N$3)
9	刈込み平均(2)	50.38		=TRIMMEAN($A$4:$H$8,4/$N$3)
10	刈込み平均(3)	50.36		=TRIMMEAN($A$4:$H$8,6/$N$3)
11	中央値	50.25		=MEDIAN($A$4:$H$8)
12	最小値	37.00		=QUARTILE($A$4:$H$8,0)
13	第1四分位値	48.00		=QUARTILE($A$4:$H$8,1)
14	第2四分位値	50.25		=QUARTILE($A$4:$H$8,2)
15	第3四分位値	52.55		=QUARTILE($A$4:$H$8,3)
16	最大値	64.00		=QUARTILE($A$4:$H$8,4)
17	四分位範囲	4.55		=N15-N13
18	四分位範囲/1.35	3.37		=N17/1.35

中央値は，両側のデータの情報を十分に使わないため，精度が落ちる．しかし，中央値は，刈込み平均の刈込み率を最大にした極限であるから，頑健性は極めて高い．したがって，汚いデータの代表値として用いることができる．

もう一つ中央値が役に立つのは，§3 で説明するヒストグラムを描いたとき，左右が対称にならない場合である．

ある企業の平均給与が計算されているとき，その企業の従業員が自分の給与を平均給与と比べると，がっかりする人が多いであろう．

平均給与は少数の高給者によって大きい方に引っ張られるからである．平均給与より高い人の割合は1/3で，低い人の割合は2/3であるということが起こる．

このような場合は，全メンバーを給与の大きさの順に並べ，中心の人の給与，すなわち中央値，と比べれば，自分は上半分か下半分のどちらに属するかを知ることができる．

このように，個々の値がどの辺に位置しているかを見るには中央値や次に説明する四分位値が役に立つ．

しかし，その企業の総人件費を知るためには，平均値が便利である．すなわち，平均給与に人数を掛けると総人件費が求められる．

中央値を求める Excel 関数は =MEDIAN(データの範囲) である．

(5) 四分位値，四分位範囲

前項で，自分の収入を評価するために，中央値 が役立つであろうと書いた．

さらに，自分が 上流，中流の上，中流の下，下流 のどの階層に属しているかを知りたいとしたら，どうしたら良いであろうか．

中央値と同様に考えて，中央値で分けた下半分の中央値と，上半分の中央値が役に立つと思われる．

このように，全体を4つに分割した3つの境界値を 四分位値 または 四分位数 (Quatile) と呼び，小さい方から 第1四分位値，第2四分位値，第3四分位値 という．第2四分位値は中央値と同じである．第1，第3四分位値は 下側 または 上側の四分位値 とも呼ばれる．

上に，四分位値の考え方を書いたが，n が 4 で割り切れないとき，どう決めるかの問題が残る．この処理方法は，市販の統計解析プログラムで異なる．しかし，実務的には問題とならない程度である[6]．

四分位値を求める Excel 関数は =QUATILE(データの範囲，M) である．M = 1, 2, 3 とすると 第1四分位値，第2四分位値，第3四分位値 が求められる．

また，M=0，4 とすると，最小値と最大値が求められる．最小値と最大値を求めるための関数は MIN(データの範囲)，MAX(データの範囲) がある．

バラツキの大きさを表わす量として標準的に用いられるのは標準偏差であった．両側の四分位値の差でバラツキの大きさを表わすこともできる．これを 四分位範囲 と呼ぶ．

もし，データが正規分布に従うとき，四分位範囲を 1.35 で割ると標準偏差に近い値が得られる．中央値が平均値よりも外れ値の影響を受けにくい（頑健性がある）のと同様に，この方法で求めた値は，標準偏差よりも頑健性がある．ここで 1.35 は p.46 で説明した 0.674 「ろくでなし」の2倍である．

表示2.7のデータについて，これまで説明した値を計算した結果を表示2.10 (p.52) に示す．

---

[6] §2.5 の 補足 参照．

■　**演習6**　次の表は，表示の録画時間が120分のビデオテープ100本の実測録画時間から表示録画時間を引いた値（単位 秒）である．

220	192	216	215	204	199	220	218	193	198
207	208	206	216	187	198	203	197	213	220
183	202	194	208	198	209	204	210	201	184
206	204	196	203	198	220	191	195	198	207
208	211	200	201	189	193	201	196	140	212
197	207	198	204	202	215	217	237	196	193
204	213	196	218	198	210	177	203	204	188
198	195	213	203	217	216	183	192	213	203
199	203	212	201	219	181	218	205	196	199
206	204	194	207	204	196	217	218	201	201

§2.5 補足 (7) の手順に従って，外れ値と思われる値を除いたときに，基本統計量がどのように変化するかを観察せよ．

ヒント　データは「第2単元.XLS」のシート「§2.4 演習」に記録されている．

データの下に n，平均値，標準偏差，変動係数，ひずみ，とがり，最小値，最大値を求める．

外れ値を探すために，本文では個々の値の偏差値を計算し，偏差値が ± 2 外を，§2.5 補足 (6) の手順で求めた．この方法の代わりに，$\overline{x} - 2s$, $\overline{x} + 2s$ を計算し，観測値がこの2つの値外のものを探すこともできる．　■

本日のまとめ

今日は，平均値と標準偏差以外に役に立つ統計的な指標とその利用方法を紹介した．多くの場合には，平均値と標準偏差を使って議論すれば事足りるが，分布の様子が左右対称の正規分布に従わない場合には，ひずみやとがりを調べたり，平均値の代わりに中央値を用いたりする方がふさわしいかもしれない．例えば，プロスポーツ選手の年俸についてデータを集めて考えてほしい．課題の本質を捕まえて，統計的な指標を使いこなす術を会得してほしいものである．

2.5 補足

(1) 平均値の導出

本文では，$(x_i - a)^2$ の和を最小とする代表値が平均値になることを表わす式 (2.5)(p.39) を初等的に導いたが，微分法を用いるともっと簡単に求めることができる．

式 (2.3) が極値（最小値 または 最大値）をとる a は，式 (2.3) を a で微分して，0 と置いた方程式を解くことにより求められる．

$$\begin{aligned}\frac{dS}{da} &= \frac{d}{da}\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n} 2(x_i - a)\frac{d}{da}(x_i - a) \\ &= \sum_{i=1}^{n} 2(x_i - a)(-1) = 0 \\ &= -2\sum_{i=1}^{n}(x_i - a) = 0 \qquad (2.11)\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} a = na \\ a &= \sum_{i=1}^{n} x_i \Big/ n = \overline{x}_i \qquad (2.12)\end{aligned}$$

途中で得られる式 (2.11) から偏差 $e_i = x_i - \overline{x}$ の和が 0 になることが分かる．この方法は，回帰式を導くなど広く用いられる．

(2) 平方和の計算

平方和 S は

$$S = \sum_i (x_i - \overline{x})^2 = \sum_i e_i^2$$

で定義される．この式を使って平方和を計算するためには，まず，平均値を計算し，ついで個々の偏差 e_i を求めなければならない．

一般に平均値は割り切れないので端数がつく．端数を適当に四捨五入して偏差を計算して，その2乗を計算すると，計算誤差が蓄積して，正確な値が求められない．

そこで，そろばん・電卓が計算手段であった時代には，上の式を次のように変形した計算式が用いられていた.

$$
\begin{aligned}
S &= \sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= \sum_{i=1}^{n}(x_i^2 - 2x_i\overline{x} + \overline{x}^2) \\
&= \sum_{i=1}^{n} x_i^2 - 2\left(\sum_{i=1}^{n} x_i\right)\overline{x} + n\overline{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2
\end{aligned}
\tag{2.13}
$$

最後の行の第2項は 修正項 または 補正項 (Correction Term) **と呼ばれ，**CT **と略記されることがある.**

(3)　なぜ $(n-1)$ で割るか

偏差の2乗の平均を求めるときに，平方和 S を，データの数 n ではなくて，$(n-1)$ で割るのはなぜであるか．それは第3単元以降で「母集団と標本」を区別するときに，標本の分散（平均平方）は，$(n-1)$で割っておいた方が，母集団の分散の推定値として より 好ましい性質を持つからである．母集団と標本の区別の明らかでなかった昔のテキストや，n が十分大きい場合しか取り扱わない分野のための統計学の本では n で割っていたり，また一部のテキストでは，n で割ったものを標本分散，$(n-1)$ で割ったものを不偏分散として2通り定義したりしている．本講座では，これからは $(n-1)$ で割ったものだけを分散（平均平方）として定義する.

上述のように $(n-1)$ で割る理論的根拠は 第4単元§3 で説明するが，直感的には次のように理解すれば良い．n 個の測定値相互の間には何の制約条件もないから，その代表値としての平均を求めるには合計を n で割れば良い．しかしながら，n 個の偏差（$x_i - \overline{x}.$）には「その和がゼロである」という制約があっ

て，もし $(n-1)$ 個の偏差を与えると，残りの一つは，この条件から自動的に決まるという性質がある．すなわち，独立な偏差の数，あるいは,「自由に決められる」偏差の数は $(n-1)$ であるといえる．それゆえ，これらの偏差の代表値を求めるために，その2乗の平均をとるとき，n ではなく，$(n-1)$ で割るのである．ここでの $(n-1)$ を 自由度 (Degree of Freedom) といい，普通は f で表わす．

関数電卓で標準偏差を求めるキーは2つある．σ_{n-1} または s は 平方和を $n-1$ で割って求める標準偏差が，σ_n または σ は 平方和を n で割って求める標準偏差が得られる．

Excel で分散と標準偏差を求める関数は前に述べたように VAR，STDEV であるが，n を使った分散と標準偏差を求める関数 VARP，STDEVP も準備されている．

たくさんの値が観測されている (n が大きい) ときには，n, $(n-1)$ のいずれを使っても結果は実質的な差がない．市場調査などのように，n が大きいデータを扱う人向けの統計のテキストでは，自由度の説明を省略するために n で割る式だけを説明しているものが見受けられる．しかし，自由度を使うことを原則とするのが良いであろう．

(4) ひずみ，とがり についての補足

分散を求めるとき，平方和を $(n-1)$ で割ったのと同様の理由で，Excel や多くの統計解析ソフトでは，ひずみ，とがりの計算には次の式が用いられている．

$$b_1 = \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s}\right)^3 \frac{n}{(n-1)(n-2)}$$

$$b_2 = \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s}\right)^4 \frac{n(n+1)}{(n-1)(n-2)(n-3)} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

(5) 四分位値についての補足

本文で説明したように，四分位値の定義方法にはいくつかある．その方法は統計解析プログラムによって異なる．以下に Excel ではどう処理しているかを

説明する[7].

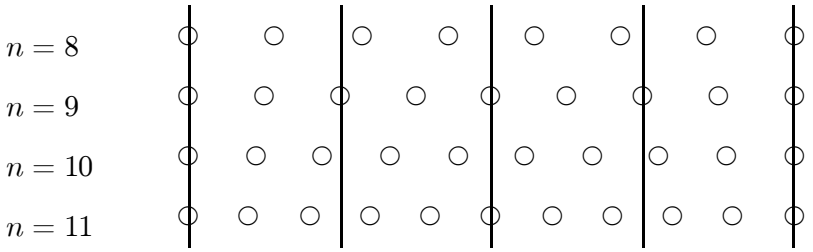
Excel では，四分位値を拡張し，95%位値なども求められるようになっている．それらとの整合性を取るために，小さい方からの割合が p である x は 小さい方からの順位 r が

$$r = p(n-1) + 1$$

である値とする．四分位値は $p = 0.25,\ 0.75$ としたときの値としている．

n が 8, 9, 10, 11 を例として，図で関係を表わす．

表示2.11: Excel の中央値・四分位値



$n = 8$ のときの下側四分位値は $r = 0.25(8-1)+1 = 2.75$ であって，2番目と3番目の加重平均で，3番目に近い．

(6)　数値の丸め

電卓で計算をしていた時代には，計算の途中の値を適当に四捨五入してメモし，次の計算で再度入力した．そのときには，途中の値を小数点以下何桁に丸めるかが大きな問題であった．

Excel では十分な精度で計算し，指定した桁数で四捨五入して表示する．

その結果を次の計算で用いるとき，画面の数値を見て，別のセルに手入力すると誤差が入るの好ましくない．

「= 途中の計算値の入っているセル」という数式を入力すると，丸めの誤差を含まないで以降の計算を続けることができる．

[7] Excel 以外のプログラムを使う場合は，簡単な数値で確認して使うと良いであろう．

このテキストでの計算は 以上の手順を用いているので，テキストに印刷された 途中の値から手で計算すると最後の桁で差の生じることがある．

最終結果をどの桁まで表示するかについては確定したルールはないが，以下に目安を示す．

平均値： 元のデータの桁の一つ下の桁まで求める．ただし，データのばらつきが大きく，n が少ないときは，元のデータの桁で止める．

標準偏差： n が100 以下のときは 有効数字2桁に丸め，100 を超えるときは3桁に丸める．

第4単元で，平均値や標準偏差の推定誤差を考慮した 区間推定 について学ぶ．区間推定の両側の値を考慮して，平均値や標準偏差の表示桁数を適切に決めることができる．

(7) Excel ヒント (1) 外れ値の除外

データの中に外れ値が見られたとき，その値を除外したとき，平均値などの指標がどのように変化するかを知りたくなる．

外れ値のセルを空白にすると，外れ値を除外した指標が自動的に計算される．しかし，もとに戻すためには，改めて値を入力しなければならない．そのために，外れ値を余白にコピーしておくなどの工夫がされる．

このような面倒さを除くためには，以下の裏技が役に立つ．

除外したい値の前に * を付ける．例えば， 120 を *120 とする．このセルは 文字列 と見なされて，計算から除外される．元に戻すためには頭の * を消せば良い．

この方法は次節の度数表やヒストグラムにも利用できる．

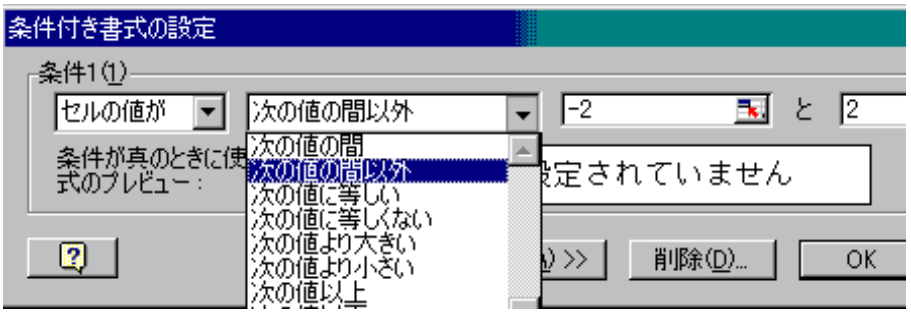
(8) Excel ヒント (2) 条件付き書式

偏差値が –2 と 2 の範囲外のとき，書式（文字のタイプや色）を変更するためには,「条件付き書式」が用いられる．

表示2.7(p.46)の場合，A10:H14 の範囲を選択した後,「トップメニュー」の「書式」,「条件付き書式」を選択する．

「条件つき書式の設定」画面（表示2.12）が現われる．

表示 2.12: 条件つき書式の設定画面



左から2番目の入力セルの右の▼をクリックして，プルダウンメニューから「次の値の間以外」を選択する．その右の2つの入力セルに −2 と 2 を入力する．

右下の「書式(F)」をクリックすると,「セルの書式設定」画面（表示2.13）が現われる．

「スタイル」の中から「太字」を選択する．ついで,「色」の右の▼をクリックして，プルダウンメニューから「赤」を選択する．

他に，字体やサイズを変更したり，下線を引いたりすることができる．

(9)　Excel ヒント (3)　条件付き度数の求め方

(i)　x がある値 a に等しい度数は

=COUNTIF(データの集まり，a の値)

で求められる．

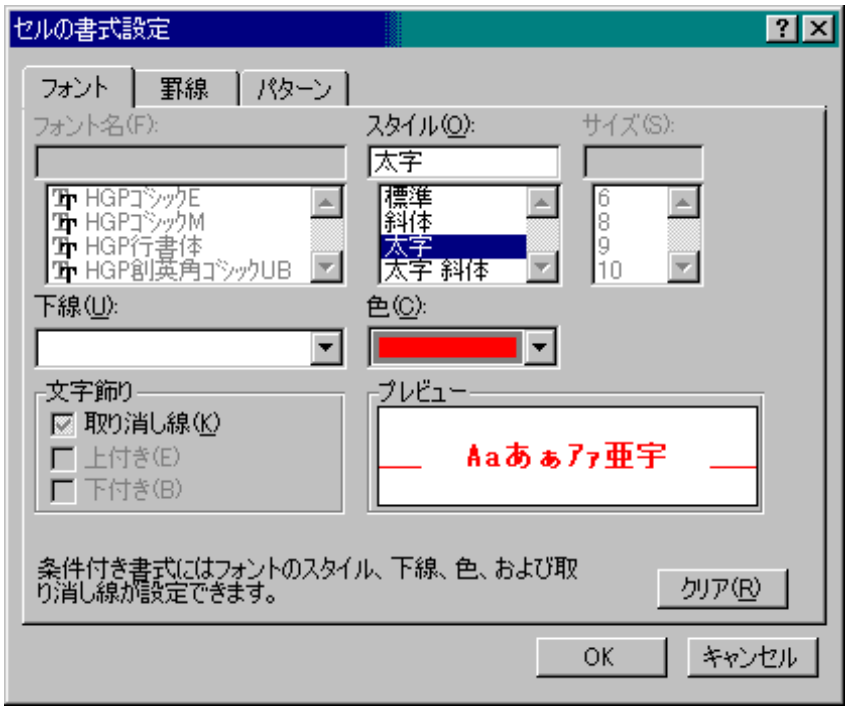
(ii)　a 以下の度数は

=FREQUENCY(データの集まり，a の値)

で求められる．

(iii)　a 以上の度数は

表示 2.13: セルの書式設定画面



=COUNTIF(データの集まり, ">=a の値")

で求められる．例えば，

=COUNTIF(B2：D21, ">=50")

とするか，または >=50 という文字列をどこかのセル（例えば F10）に入力しておき，

=COUNTIF(B2：D21, F10)

としても良い．

>=50 のほかに，<50，=>50，>50，<>50 なども可能である．

=50 または <=50 とすると (i) または (ii) と同じ結果が得られる．

3　量的データの記述（2）

3.1　度数表とヒストグラム (離散量)

(1)　交通事故による死亡者数の統計

ある都市で交通事故により死亡した人の数を，1月から10月まで毎日調査した結果が表示3.1に示されている．

表中の黒く四角で囲まれたところが日曜日を表わしており，斜体網の数字のところが祝日である．表の下に，その月の合計人数と，平均人数が計算されている．

これから，1日の死亡者数はどんな傾向をもっているかを考えよう．

この表をよく見ると，

(a)　正月3日の死亡者数は非常に多かった

とか，

(b)　2月と6月の死亡者数は比較的少ない，

などということに気づくであろう．

また，日曜とか雨の日は死亡者が多いのではないか，というようなことを想像してみることもできるであろう．この点については，§3.1(5) で取り上げる．

(2)　度数分布表

ここでは，そのような詮索はさておいて，1日の死亡者数が0人の日から9人の日まであることに注目し，1月1日から10月末日までの305日のうち，死亡者数が0人の日は何日あったか，1人の日は何日あったか，2人の日は，3人の日は … というように調べてみよう．これには表示3.1で数字0がいくつあるか，1がいくつあるか，...，というように個数を数えれば良い．

個数を一つ一つ数え上げるのは，誤りが入る危険があるので好ましくない．

表示 3.1: 1日に交通事故で死亡した人の数

	A	B	C	D	E	F	G	H	I	J	K
3	2000年	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月
4	1	5	2	3	1	1	4	8	2	2	2
5	2	9	2	3	1	1	2	0	4	4	0
6	3	5	3	1	4	4	1	2	3	4	3
7	4	4	3	2	3	2	2	2	0	2	2
8	5	0	1	1	2	5	1	2	5	1	1
9	6	0	3	2	4	2	1	1	4	1	1
10	7	1	0	4	1	5	3	3	2	1	4
11	8	2	1	1	1	2	2	1	0	1	3
12	9	2	1	1	3	1	3	3	2	6	4
13	10	3	2	3	1	1	0	3	4	3	2
14	11	2	1	4	1	1	2	5	2	1	4
15	12	2	5	4	1	1	0	0	5	4	1
16	13	3	1	1	3	4	1	4	5	3	2
17	14	1	2	2	4	4	6	2	3	1	6
18	15	6	1	3	2	2	2	2	2	2	7
19	16	7	1	3	2	0	1	2	2	3	4
20	17	0	0	4	2	2	2	2	2	2	1
21	18	1	1	3	1	2	0	3	3	2	2
22	19	2	1	3	1	3	2	0	3	0	2
23	20	0	4	3	2	4	3	6	4	1	2
24	21	4	1	5	3	4	5	1	3	0	3
25	22	1	1	1	4	2	1	4	1	1	3
26	23	3	1	1	2	0	0	6	0	3	1
27	24	3	3	0	1	3	3	1	0	6	3
28	25	2	2	7	1	3	3	1	0	1	1
29	26	2	2	7	3	3	4	3	0	2	4
30	27	4	1	1	3	5	0	4	4	3	3
31	28	0	1	5	3	4	1	3	4	2	4
32	29	1	1	0	4	1	3	2	3	2	4
33	30	2		5	7	1	3	9	3	5	1
34	31	4		0		2		3	2		2
35	計	81	48	83	71	75	61	88	77	69	82
36	平均	2.6	1.7	2.7	2.4	2.4	2.0	2.8	2.5	2.3	2.6

Excel には COUNTIF(データの範囲, 値) という関数があるので，それを使って数え上げる．

表示 3.2 に示すように，M列に人数を入力する．N4 のセルに

```
=COUNTIF($B$4:$K$34,M4)
```

と入力する．$B$4:$K$34 は表示 3.1 のデータが記録されている領域である．月によって日数が異なり，下の方に空白のセルがあるが，そこは無視される．この命令によって，死亡人数が0人（M4）の日数が N4 に求められる．

表示3.2: 度数分布表

	M	N	O	P	Q
3		度数	百分率	*=2	*=5
4	0	29	9.5%	**********|	*****||||
5	1	76	24.9%	**************************************	**********|
6	2	70	23.0%	**********************************	**********
7	3	59	19.3%	*****************************|	**********||||
8	4	41	13.4%	********************|	********|
9	5	15	4.9%	*******|	***
10	6	7	2.3%	***|	*||
11	7	5	1.6%	**|	*
12	8	1	0.3%	|	|
13	9	2	0.7%	*	||
14	合計	305	100.0%		

N4 のセルを下にコピーすると，表示3.2 が得られる．

表示3.2では，全調査日数（305 日）のうち，交通事故による死亡者数が，0 人の日は29 日，1 人の日は76 日，…，9 人の日は2 日あったことを示している．すなわち,「交通事故による死亡者数」という変数の取る値 x_i （ここでは0から9までの各値）ごとの出現日数 — これを 度数（または頻度）と呼び，普通は f_i で表わす — が与えられた．このような度数の分布を 度数分布 という．ここでの変数 x は，$0, 1, 2, \ldots$ という整数値のみを取るから 離散量 という．

(3)　相対度数分布

上に得られた度数 f_i を，総度数 $n = 305$ で割った値で表わした分布を 相対度数分布 という．

表示3.2 で，N14 に =SUM(N4:N13) と入力して，総日数を求め，O4 のセルに，=N4/$N14 を入力して，下にコピーする．

これから，交通事故による死亡者数は，1 人，2 人，3 人という日が多く，それぞれ全体の20〜25%あり，死亡者0 人の日は全体の1 割以下で，死亡者が6 人を超える日はきわめて少ないことが分かる．このようにして，表示3.1 の統計表のままでははっきりしなかった事情が，表示3.2 の度数分布または相対度数分布

を作ることによって明らかになったのである.

(4)　ヒストグラム

度数分布表, または相対度数分布表は, そのままでもデータの集団として持つ性質をよく表わすが, これをもっと見易くするために, いろいろな図示法がある.

表示3.2 のP 列には, ＊や | のマークが並んでいる.

P4 のセルには

```
=" "&REPT("*",ROUNDDOWN(N4/2,0))&REPT("|",MOD(N4,2))
```

と入力され, 下にコピーされている.

最初の & は その前の " "(半角のスペース) と その後に続く文字列を連結することを指示する.

REPT 関数は, 最初のパラメータの記号 * を, 2 番目のパラメータの個数 繰り返せという関数である.

ROUNDDOWN(N4/2,0) は, N4 の度数を 2 で割って, 小数点以下の桁数が 0 (整数) になるように切り捨てろという式である.

2番目の REPT 関数の2番目のパラメータの MOD(N4,2) は, 2 で割った余りを求める関数である.

割り算の分母の 2 と MOD 関数の2番目のパラメータの 2 は, 度数の最大値に合わせて適当に調整する. 表頭に, * が何個に対応するかを示しておく.

Q 列に $*=5$ としたときの図を示す[1].

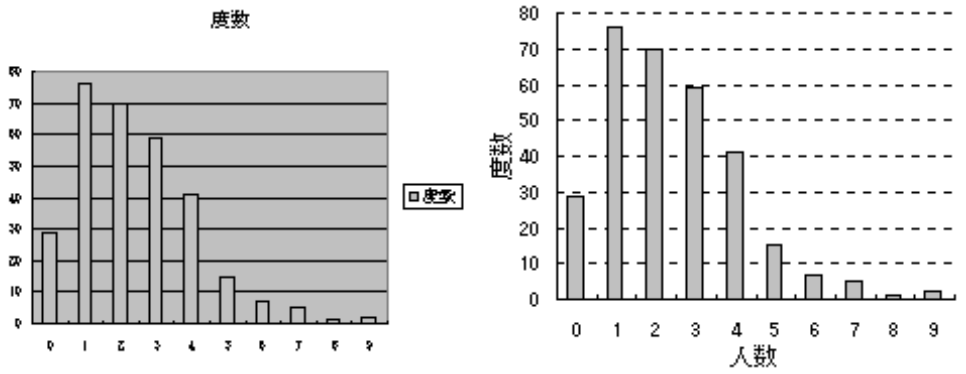
このグラフは, 結果をコンパクトに表示するのに便利である. しかし, グラフらしい出力を得たいときには, Excel のグラフウィザードを使って棒グラフを作成する.

表示3.2の人数と度数の列を反転して,「トップメニュー」から「グラフウィザード」を選択する.「グラフの種類」の画面で,「縦棒」を選択する.「完了」を

[1] | の文字幅を狭くするために, 字体を「MSP ゴシック」に修正する. P は, 文字幅を文字の種類によって変化させる書体であることを示している.

クリックすると，棒グラフ得られる．大きさを調整したグラフを表示3.3の左に示す．

表示3.3: 交通事故による1日の死亡者数の分布（ヒストグラム）



見かけが悪いので，整形（表題を除き，背景色を消し，横軸の目盛線を点線とし，目盛りの値のフォントを大きくする）し，横軸と縦軸に変数名を入力すると，表示3.3の右のグラフが得られる．このように度数分布をグラフ化したものを ヒストグラム（柱状図）という．

(5)　層別ヒストグラム

§1.3, 1.4, 1.5 で，2つの変数の関係を把握することが大切であることを示した．量的な変数でも層別の重要性は変わらない．

交通事故の原因はいろいろあるであろう．道路の混雑も一つの原因と考えられる．平日に比べて土日は交通事故が少ないかもしれない．土日は慣れない人が運転するチャンスが多かったり，通勤や荷物の運送の頻度が少ないため道路が空いていて速度がつきやすくて，死亡事故が多いかもしれない．

また，交通事故の原因として，天候（晴れ，曇り，雨）が考えられる．

このように，交通事故に影響しそうな要因についてデータを集めたならば，それらの要因について層別して平均を求めたり，ヒストグラムを作成することにより，想像したモデルの妥当性を検証することができるであろう．

表示3.1 のデータを，平日，土日，祝日で層別して，度数表を作成すると，表示3.4 の左が得られる[2].

表示3.4: 平日，土日，祝日で層別した1日の死亡者数の度数表

	度数			
	平日	土日	祝日	休日
0	25	4	0	4
1	63	12	1	13
2	51	17	2	19
3	43	15	1	16
4	21	19	1	20
5	5	8	2	10
6	1	5	1	6
7	0	5	0	5
8	0	1	0	1
9	0	2	0	2
日数計	209	88	8	96
人数計	409	298	28	326
平均	1.96	3.39	3.50	3.40

	相対度数			
	平日	土日	祝日	休日
0	0.12	0.05	0.00	0.04
1	0.30	0.14	0.13	0.14
2	0.24	0.19	0.25	0.20
3	0.21	0.17	0.13	0.17
4	0.10	0.22	0.13	0.21
5	0.02	0.09	0.25	0.10
6	0.00	0.06	0.13	0.06
7	0.00	0.06	0.00	0.05
8	0.00	0.01	0.00	0.01
9	0.00	0.02	0.00	0.02
合計	1.00	1.00	1.00	1.00

それぞれの層ごとに，平均死亡者数を計算すると，1.96, 3.39, 3.50 と，平日に比べて，休日（土日＋祝日）の死亡者が1.7倍と多い．各層の日数合計が異なるので，全体が1になるように，右のように相対度数表に変換する．これから，表示3.5 に示す2種類の 層別ヒストグラム が作成される．

いずれのグラフも，群による分布の違いを明確に表わしているとはいえないであろう．

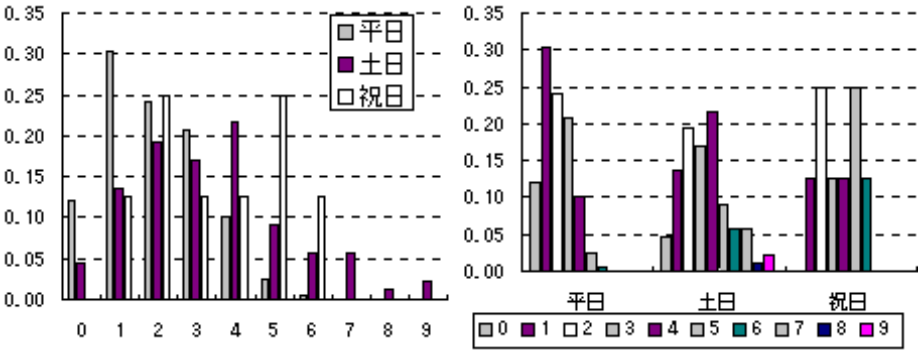
そこで，表示3.6 のように，§1.2 (1) で説明した積み上げグラフ を作成した．

3人以上の死亡者がでる割合は，平日が約 1/3 であるのに対して，休日は約60%に増えることが分かる．

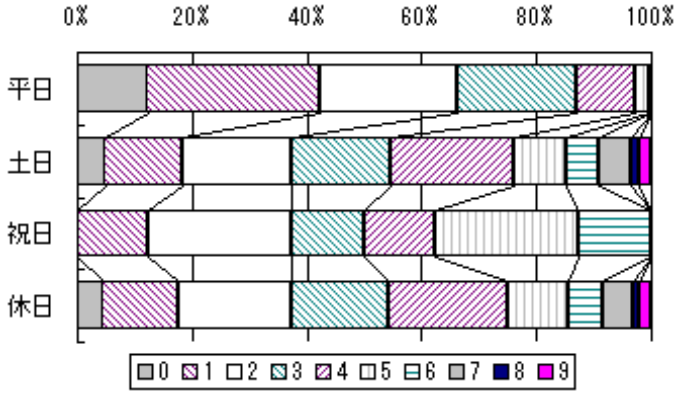
なお，土日と祝日の差が見られないので，両者の合計の列を追加した．

[2] 元の表のデータを層別して度数表を作成する過程は，Excel ファイルのシート「§3.1b」に説明されている．

表示3.5: 層別ヒストグラム



表示3.6: 積み上げグラフ



本日のまとめ

第2章では，データから計算した統計的な指標を紹介したが，第3章では，それをグラフで表現する方法を学習する．データ解析では，統計的な指標とグラフをセットで考察してみると，木と森とを併せて眺めることができるので，誤解なくいろいろなことが分かるであろう．今日は，交通事故による死亡者数の統計を例に，離散量のヒストグラムについて紹介した．明日は，連続量のヒストグラムについて紹介する．

3.2 度数表とヒストグラム (連続量)

(1) データ

前節の交通事故による死亡者数は，整数値のみを取る離散量であった．これに対して，体重や身長，作物の収量，工場製品の歩留り（使用原料の量，または，期待される理論収量に対する製品の量の比率），温度や湿度など，長さ，重さ，その他の量を測って得られる値は 連続量 であり，これらについても度数表とヒストグラムを作成する．

表示3.7には，110人の中学生の体重を，100グラム（0.1kg）単位で測ったデータが示されている．ここで体重54.7kgの人の真の体重は，54.65kgと54.75kgの間にあると考えられる．（体重の分布はもともと連続的なものであるが，ここでの測定器具では0.1kg単位の目盛りまでしか読めなかったので，データは0.1kg単位の離散量であるかのごとく表示されている．）

表示3.7: 110人の中学生の体重（kg）

54.7	61.7	60.0	49.7	50.5	62.7	48.0	58.0	49.0	43.3	54.2
53.7	45.0	48.0	56.0	44.0	49.6	51.5	43.0	53.0	53.0	48.2
64.0	57.2	41.7	42.0	50.0	52.0	49.5	50.0	50.7	49.7	45.0
45.0	57.0	48.0	52.0	62.7	56.7	50.1	52.0	60.5	48.0	53.2
52.0	49.7	59.2	55.0	57.8	44.5	47.7	52.5	46.2	47.0	46.4
48.0	51.0	55.8	50.0	50.0	54.0	55.0	55.0	49.5	58.0	45.6
50.7	37.7	48.5	58.2	48.0	55.0	41.7	51.7	51.0	60.0	54.8
52.5	51.5	50.0	56.2	50.0	48.0	52.0	56.0	51.0	52.0	53.0
48.0	44.0	49.7	46.7	45.0	43.7	47.7	48.0	47.0	41.2	46.5
47.7	47.7	56.0	52.7	49.0	52.2	47.0	53.0	50.5	49.5	54.0

(2) 組分けによる度数表

さて，表示3.7のデータを，前節の方法に従って度数表にまとめようとすると，一番大きい値は64.0で一番小さい値は37.7であり，$64.0 - 37.7 = 26.3$となるから0.1kg単位では，264個の異なる値に対する度数を求めねばならなくなる．しかも，その度数の多くは 0 となる（37.8や37.9に対する度数は 0 である）．これでは，表示3.4よりもかえって複雑になるばかりでなく，度数表の意図する

ところの，集団の特徴を把握しやすくするという目的が達せられない．そこで，0.1kg の代りに，1kg または2kg というようなより大きい単位にデータをまとめて 級（組，階級 ということもある）を作り，それぞれの級に属する個人の度数 — これを以降「級度数」という — を数えることにする．

このデータの最小値と最大値を求めると，37.7, 64.0kg で，範囲は $64.0-37.7=26.3$kg である．1kg 単位に級を作ると 27 の級に分けられることになり，まだ級数が多すぎる．

級の個数について，こうしなければいけないという規則があるわけではない．目安として 10 ないし 15 前後が良いであろう．

いくつかの級の個数で実際に度数表やヒストグラムを作成し，データの分布状況を適切に示していると考えられる級の個数を解析者が選択するのが良いであろう．

ここでは，2kg 単位で級を作ることにすると，級の個数は13個前後となる．ここで，2kg のことを 級間隔 （あるいは，階級幅，区間幅 ということもある）と呼ぶ．

37kg から 2kg 刻みに級の境界値を決めると，級は $37\sim39$, $39\sim41$, …, $63\sim65$ となる．ここで，境界値に等しい値はどちらに入れるかが問題となる．

境界値を上の級に入れると

$$37\le x<39,\ 39\le x<41,\ \ldots,\ 63\le x<65$$

となり，下の級に入れると

$$37<x\le 39,\ 39<x\le 41,\ \ldots,\ 63<x\le 65$$

となる．

品質管理の分野では，前者を用いることが JIS で決められている．

データは0.1kg 単位に丸められているから，実際の級の範囲は

$$37\le x\le 38.9,\ 39\le x\le 40.9,\ \ldots,\ 63\le x\le 64.9$$

となる．

Excel で度数表を作成するときには，このように決められた範囲の上限値 38.9, 40.9, …, 64.9 を使う．

表示3.8: **中学生の体重の度数分布**

	N	O	P	Q	R	S
3	級上限	級代表値	度数	累積度数	累積割合	*=1
4	38.9	37.95	1	1	0.9%	*
5	40.9	39.95	0	1	0.9%	
6	42.9	41.95	4	5	4.5%	****
7	44.9	43.95	6	11	10.0%	******
8	46.9	45.95	9	20	18.2%	*********
9	48.9	47.95	18	38	34.5%	******************
10	50.9	49.95	21	59	53.6%	**********************
11	52.9	51.95	16	75	68.2%	****************
12	54.9	53.95	11	86	78.2%	**********
13	56.9	55.95	10	96	87.3%	**********
14	58.9	57.95	6	102	92.7%	******
15	60.9	59.95	4	106	96.4%	****
16	62.9	61.95	3	109	99.1%	***
17	64.9	63.95	1	110	100.0%	*

表示3.8 **の** N4:N17 **に級上限値を入力する．度数を入力する 領域 （複数のセル）** P4:P17 **を反転し，**

```
=FREQUENCY(B4:L13,N4:N17)
```

を入力する．ここに， B4:L13 **は体重データの記録されている領域である．**

式を入力したならば，シフトキーとコントロールキーを押したままで エンターキーを押す． P4:P17 **に度数表が得られる**[3]**．**

こうして得られるものを 組分けによる度数分布 という．

この表から，体重47.0〜48.9，49.0〜50.9，51.0〜52.9kg **の**3**つの級の度数が多く，これらの中に全体の約** 50% **の人の体重が含まれることが分かる．**

(3) ヒストグラム

表示3.8 **から，交通事故の場合と同様の手順でヒストグラムを作成するが，次の**2**点が異なる．**

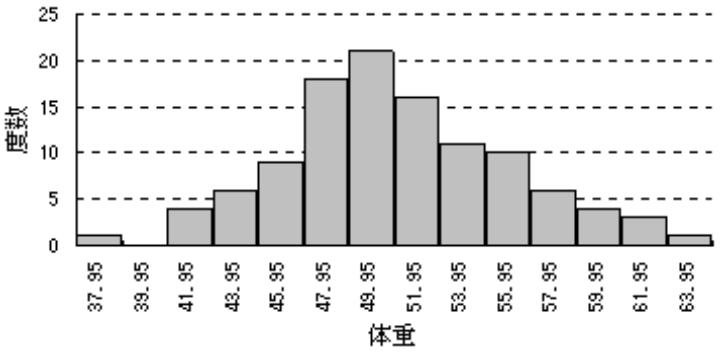
[3] **表示** 3.2 **の度数もこの方法で求めることができる．**

（o）　横軸の目盛りとして，級の範囲の中心の値を用いる．最初の級では，$37.0 \leq x \leq 38.9$ であるから，$(37.0 + 38.9)/2 = 37.95$ とする．これを，級の 代表値 と呼ぶ.

（o）　データが連続であるから，柱の幅をいっぱいに広げ，柱と柱の間を空けない[4].

このような調整をして得られるヒストグラムを表示3.9 に示す.

表示3.9: 110 人の中学生の体重のヒストグラム



(4)　級間隔や級の境界値の決め方

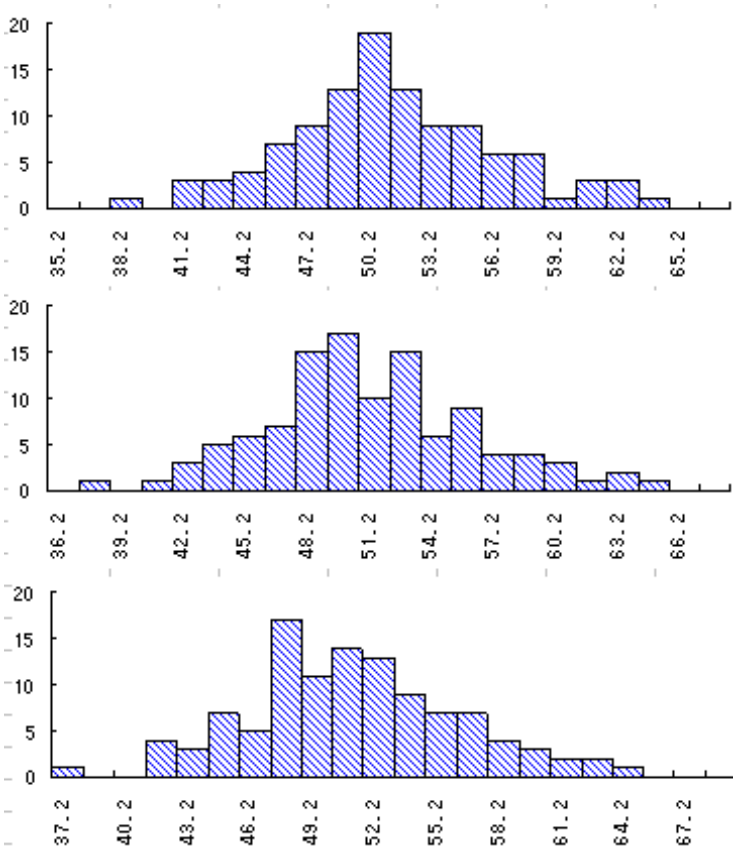
同じデータでも，級間隔や級の境界値を変化させるとヒストグラムの形が変化する.

表示3.7 のデータで，級間隔を 1.5kg とし，最初の上限値を 35.9kg, 36.9kg, 37.9kg と変えてヒストグラムを作成すると，表示3.10 のように形が大きく変化する.

この例が示すように，2 番目のヒストグラムから，2 山があると早合点するのは危険である．ヒストグラムの形に変わった特徴が見られたときは，級間隔や

[4] ヒストグラムが表示されたならば，柱をダブルクリックして「データ系列の書式設定」画面を表示させる．そこで,「オプション」を選択し,「棒の間隔」を 0 に設定する.

表示 3.10: 境界値を変化させたヒストグラム



境界値を変化させて，その特徴が本質的なものかどうかを確認する必要がある．

もし，ヒストグラムで2つの山が見られるときには，異質な2つの集団が混合している可能性がある．

■ 演習 7 演習 6 のデータについて，度数表，ヒストグラム を作成せよ．なお，分布にはどのような特徴が見られるか． ■

本日のまとめ

昨日は，離散量のヒストグラムを紹介した．今日は，中学生の体重の例を用いて，連続量のヒストグラムを紹介した．受講生は，両者の違いについて理解できたであろうか．連続量のヒストグラムでは，級間隔や級の境界値の決め方により，グラフから受ける印象が異なるため，それらを適当に変化させてヒストグラムを眺めてみることが重要であろう．

3.3　いろいろな分布

データをヒストグラムで表わすと，平均値，標準偏差 などの数値だけでは分からない情報が得られる．

(1)　切れた分布と打切り標本

表示3.11のヒストグラムは，規格寸法が62.0mm以上であるパッキングの受入検査値の寸法を調べた結果である．

規格以下はわずか1個である．しかし，規格値のところで分布が切れている．これから，出荷の際全数検査をして，規格以下を除いているのではないかと推察される．

このように，データを取る以前になんらかの事情によって，分布の片端，または，両端が切れてしまっていることがある．これを 切れた分布 と呼ぶ．

また，電球の寿命試験などでは，測定の都合から一定時間（例えば2,000時間）で打切られる．このような試験は 定時打切り試験 という．この場合は，そのときまで寿命のあった個体の真の寿命は分からずじまいになる．このような事情の下で得たデータを 打切り標本 という．現実の場では，このような場合が少なくないので，注意を要する．

表示 3.11: **パッキングの外径分布**

下限	上限	n	度数グラフ(*=2)
61.6	61.799	0	:
61.8	61.999	1	:|
62.0	62.199	19	:*********|
62.2	62.399	30	:***************
62.4	62.599	40	:********************
62.6	62.799	37	:******************|
62.8	62.999	28	:**************
63.0	63.199	22	:**********
63.2	63.399	15	:*******|
63.4	63.599	8	:****
63.6	63.799	0	:

なお，打切り標本から得られたデータの解析には，特別な工夫が必要であり，生存時間解析 や信頼性データ解析 という分野で検討されている．打切りデータの種類については，章末の補足を参照してほしい．

(2) 対数正規分布

表示3.12**の左のデータ** $(n = 40)$ **について，ひずみ と とがりを計算すると** 1.64, 3.35 **となり，正規分布から大きく外れている．ヒストグラムを表示**3.12**の右に示す．**

分布を見ると，大きい方に長く裾を引いている．このようなヒストグラムはいろいろな分野で見られる．

経済分野では，世帯貯蓄額，年収，上場企業の資本金，などが，医学分野では，血液に含まれるコレステロール，血糖，GOT [5]**の含有量などがある．**

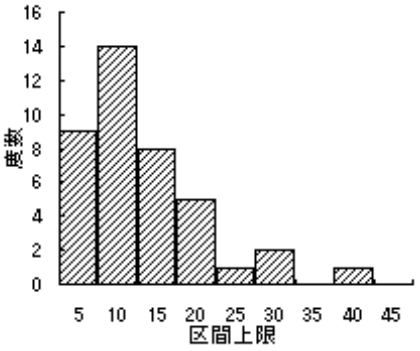
正規分布は，第1 **単元の**§3.4 **で説明した中心極限定理で導かれる．すなわち，観測値** x **にたくさんの変動要因による誤差** $\varepsilon_1, \varepsilon_2, ...$ **が含まれるとき，**

$$x = \mu + \varepsilon_1 + \varepsilon_2 + ... \tag{3.1}$$

[5] **グルタミン酸オキサロ酢酸トランサミラーゼ の略で，この** GOT **が高値の場合，肝疾患が疑われる．最近は** AST**（アスパラギン酸アミノトランスフェラーゼ）と呼ばれることが多い．**

表示 3.12: 対数正規分布

6.71	4.91	2.96	7.42
20.29	10.38	7.13	2.14
15.42	13.00	9.53	6.84
16.28	17.01	37.02	6.43
13.95	11.10	6.61	16.12
3.43	27.91	4.23	6.64
2.06	17.34	10.12	4.13
2.66	8.20	7.63	25.32
8.85	10.44	13.16	5.06
5.30	3.27	12.08	9.76
平均	10.47	ひずみ	1.64
s	7.46	とがり	3.35



と表わすことができ，この x は近似的に正規分布に従うのである．

それに対して，ある期の貯蓄額 x_j は前期の貯蓄額 x_{j-1} の影響を受ける．すなわち，x_j は x_{j-1} に $(1+$増加率$)$ を掛けたもので，増加率が各期によって変化するというモデルが考えられる．このモデルは次の式で表わすことができる．

$$x_j = x_0 \times (1+\varepsilon_1) \times (1+\varepsilon_2) \times \ldots \times (1+\varepsilon_j) \tag{3.2}$$

酒の飲みすぎで肝臓を損なうと血液中のGOTが多くなる．正常でGOT=20の人が酒を飲むと翌日のGOTは2増えて22になるとする．肝臓障害でGOT=80の人が酒を飲んだ翌日のGOTはいくらになるであろうか？　2増えるのではなく，10% 増えて 88 になる．このようなモデルが成立するとき，GOT の分布は大きい方に長い裾を引く．

式(3.2) の両辺の自然対数[6]を取ると，

$$\ln(x_j) = \ln(x_0) + \ln(1+\varepsilon_1) + \ln(1+\varepsilon_2) + \ldots + \ln(1+\varepsilon_j)$$

となる．ここで，ε が小さいとき，$\ln(1+\varepsilon) \approx \varepsilon$ という性質を使うと，

$$\ln(x_j) = \ln(x_0) + \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_j \tag{3.3}$$

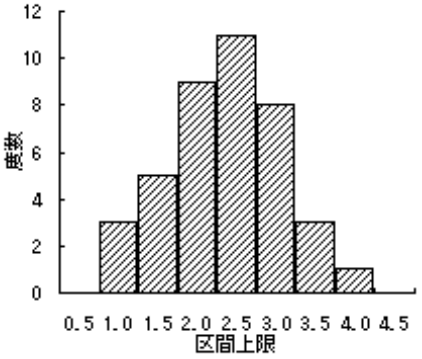
[6] x の自然対数を $\ln(x)$ で表わす．§3.5 補足 参照．

となる．式(3.3)**で，**$\ln(x_j) = y,\ \ln(x_0) = \mu$ **と置き換えれば式**(3.1) **と同じモデルである．つまり，**x **の対数は正規分布に近似的に従うことが分かる．**

表示3.12 **のデータの自然対数を取って，同様の解析をした結果を表示**3.13 **に示す．**

表示 3.13: **対数変換値の分布**

1.90	1.59	1.09	2.00
3.01	2.34	1.96	0.76
2.74	2.57	2.25	1.92
2.79	2.83	3.61	1.86
2.64	2.41	1.89	2.78
1.23	3.33	1.44	1.89
0.72	2.85	2.31	1.42
0.98	2.10	2.03	3.23
2.18	2.35	2.58	1.62
1.67	1.18	2.49	2.28
平均	2.12	ひずみ	-0.11
s	0.70	とがり	-0.37



対数変換によって，ひずみ，とがり が 0 **に近づき，ヒストグラムも正規分布と認められる形になった．**

表示3.12**の最大値** 37.02 **の偏差値は** $(37.02 - 10.47)/7.46 = 3.56$ **で，外れ値であるが，対数変換することにより，偏差値は** $(3.61 - 2.12)/0.70 = 2.13$ **となった．**

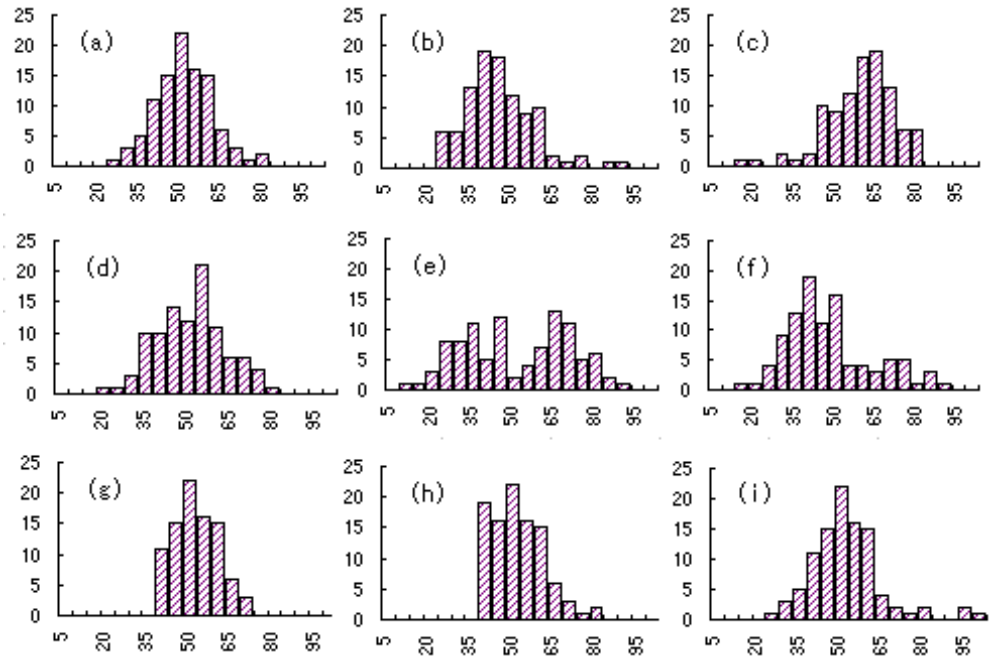
ここでは，対数として 自然対数 を用いたが，常用対数 を用いても，対数変換値，その平均と標準偏差がすべて$1/2.303$ **になるだけで，偏差値の値は変わらない．**

しかし，自然対数を取ると，x **の変動係数** $= 7.46/10.47 = 0.71$ **が** $\ln(x)$ **の標準偏差** 0.70 **にほぼ等しいという関係がある．**

(3)　ヒストグラムの見方

工場の品質管理を例として，ヒストグラムの見方を調べてみよう．工場では，製造工程が管理された状態にある場合は，おおむねデータは正規分布に従うことが知られている．工程が安定な状況にあるかどうかは，収集したデータからヒストグラムを描くことで把握できる．この場合，分布の様子を調べることが目的であるから，標本数は多めに 50 から 200 個くらい収集すると良いだろう．表示3.14は，工程で見られるいろいろなヒストグラムの例である．ヒストグラムを見る場合は，多少のでこぼこを無視して大体の姿に着目すると良い．

表示 3.14: いろいろな形をしたヒストグラム



表示3.14 の9つヒストグラムは次のような場合を想定して，乱数から生成したものである．

a）一般に安定した製造工程．

平均 50**，標準偏差** 10 **の正規分布に従う乱数（以下** a **で表わす）である．**

b）**微量成分の含有率など，ある値以下の値を取らない場合．**

a **から** $b = 20 + (a - 20)^{1.6}/10$ **と変換して求めた．大きい方に裾を引いている．**

c）**純度の高い成分の含有率など，ある値以上の値を取らない場合．**

100 **から** b **を引いて求めた．**c **の分布は，**b **の分布の左右を入れ替えたものになっている．**

d）**平均値がわずかに異なる** 2 **つの分布が半々で混合した場合．**

a **の始めの** 50 **個に** 5 **を加え，残りの** 50 **個から** 5 **を引いて求めた．**

e）**平均値がかなり異なる** 2 **つの分布が半々で混合した場合．**

d **と同様の加工をしたが，加減する量を** 15 **にした．**

f）**平均値が大きい群が** 20% **混ざっている．**

a **の始めの** 20 **個に** 20 **を加え，残りの** 80 **個から** 10 **を引いて求めた．**

g）**規格外（以上，以下の両方）のものを全数選別して取り除いた場合．**

規格の範囲を 35, 70 **とし，**a **の** 35 **以下，**70 **以上の値を除いた．**12 **個が除かれ，** $n = 88$ **となった．**

h）**規格外れのものを手直ししたり，データを偽って報告した場合．**

a **の** 35 **以下の値を** $35 + (35 - x)/2$ **として** 35 **以上に変更した．**

i）**測定誤りがあったり，工程に異常があった場合．**

a **の最初の** 12 **個について，** 60 **を超えるもの（**3 **個）に** 30 **を加えた．**

(4)　ヒストグラムの層別

工程の品質特性や連続量の多くは，正規分布で近似できる．もしも，ヒストグラムを作ったときに分布が正規分布から外れていると見られるときは，中心または，バラツキが異なる母集団からのサンプルが混在していると考えると良いかもしれない．問題解決の基本は，層別に始まり，層別に終わるとまでいわれている．

表示3.14 **の** d, e, f **は**2 **つの異質な集団の集まりである．**2 **つの集団の平均が小さいときは，**d **のように 分布の頭が広くなるが，これから異質集団の集まりであると判断することはむずかしい．平均が大きく離れると，**e **のように明ら**

かな2山になる．現実には，d と e の中間の分布が得られる．また，f のように，2つの集団の大きさが異なるときには，分布の裾野に小さな山が現われる．

このように，複数の異質な集団が混ざっているときには，全体のヒストグラムを見ただけでその実態を掴むことは一般にむずかしい．

このようなことが考えられる場合には，交通事故による死亡者数を，曜日によって層別したように，材料・機械・作業条件・作業者などの異なった処理が混ざっていないか調べ，データを層別してヒストグラムを作成して調べる必要がある．

それでも，分布が正規分布から外れている場合は，その特性の意味を考えて，対数変換や別の分布を当てはめてみると良いだろう．

本日のまとめ

今日は，左右対称な分布である正規分布から得られたヒストグラムを基本として，連続量のヒストグラムの見方についてまとめてみた．取り上げた課題によっては，作成したヒストグラムの様子が異なることも学んだ．例えば，経済データによく見られる大きい値の方向に長く裾をひくような分布からのヒストグラムである．このような分布でも対数変換を行うことで，左右対称化できる場合が多いことを学んだ．明日は，ヒストグラムとは別のグラフ表現について学習する．

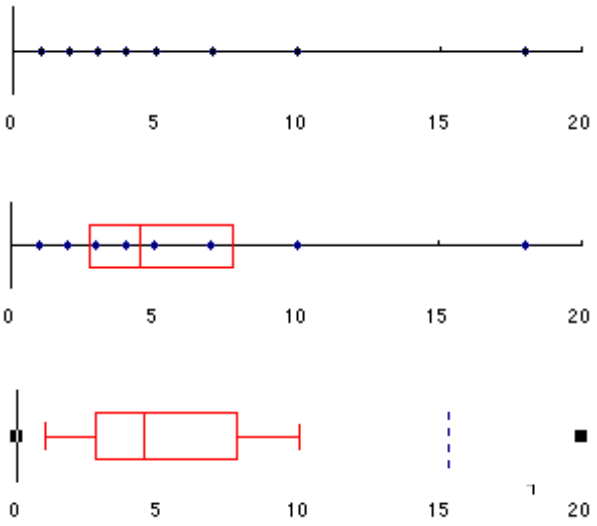
3.4　箱ひげ図とその応用

(1)　箱ひげ図の考え方

$n = 8$ の簡単なデータ $1, 2, 3, 4, 5, 7, 10, 18$ を考える．このデータの ひずみ と とがり は $1.57, 2.51$ で，いずれも 1.5 を超えている

横軸に x をとって観測値をプロットすると，表示3.15の上のグラフが得られる．

表示 3.15: 箱ひげ図



上のグラフに，中央値(4.5) と四分位値$(2.8, 7.8)$ を追加して，箱を描くと，表示3.15の中のグラフが得られる．3つの縦線で観測値が4分される．

ここで，x の最大値18は箱から遠く離れている．このデータのひずみ，とがりが1.5を超えているのは，分布が正規分布から外れているのではなく，$x_8 = 18$ のためかもしれない．

表示3.15 の箱の幅 $W = 7.8 - 2.8 = 5.0$ は四分位範囲である．データが正規分布に従うとき，四分位範囲は 1.35σ の推定値となることは既に述べた．これから，箱の両端から箱の幅 W の 1.5 倍以上離れた値は$\pm(1.35/2 + 1.5 \times 1.35)\sigma = \pm 2.70\sigma$

となり，正規分布がこの範囲外に出る確率は約1%である．

そこで，箱の両端から $1.5W$ 離れた位置に点線を引き，その外側の点は 外れ値 (Outlier) とする．この例では，$7.8+1.5\times 5=15.3$ に点線が引かれ，$x_8=18$ は限界線外であり，外れ値と判断される．

外れ値を除いた最小値と最大値の位置を示し，そこから箱までを結ぶ．外れ値以外の観測点は表示しないようにすると，表示3.15の下のグラフが得られる．

このようにして導かれた図を 箱ひげ図 (Box Wisker Diagram) と呼ぶ．

(2)　箱ひげ図の作り方

表示3.16のデータ $(n=40)$ を用いて，箱ひげ図を作成する手順を示す．計算表と得られた箱ひげ図を表示3.16に示す．

表示 3.16: データと箱ひげ図

	B	C	D	E	F	G	H	I	J
1									
2	データ	48.0	51.0	50.0	50.0	54.0	55.0	49.5	58.0
3		50.7	37.0	66.0	48.0	55.0	51.7	51.0	64.0
4		52.5	51.5	56.2	50.0	48.0	56.0	51.0	52.0
5		48.0	44.0	46.7	45.0	43.7	48.0	47.0	40.0
6		47.7	47.7	52.7	49.0	52.2	53.0	50.5	49.5
7									
8		下側	上側						
9	1	**37.0**	**66.0**						
10	2	**40.0**	**64.0**						
11	3	**43.7**	**58.0**						
12	4	44.0	56.2						
13	四分位値	**48.0**	**52.6**		中央値	**50.3**			
14	限界値	**41.2**	**59.4**		四分位範囲	**4.6**			
15									
16									
17									
18									
19									
20									
21									
22									
23		30	40		50		60		70

(i) **箱ひげ図作成のために必要な数値（中央値，四分位値，大きい方と小さい方の観測値）を求める．**

(ii) **両側の四分位値で箱を描く．**

(iii) **中央値に線を引く．**

(iv) **外れ値かどうかを判断する限界値を次の式で計算する．**

$$\textbf{四分位範囲} = \textbf{上側四分位値} - \textbf{下側四分位値} = 52.6 - 48.0 = 4.6$$
$$\textbf{下側限界値} = \textbf{下側四分位値} - 1.5 \times \textbf{四分位範囲} = 48.0 - 6.8 = 41.2$$
$$\textbf{上側限界値} = \textbf{上側四分位値} + 1.5 \times \textbf{四分位範囲} = 52.6 + 6.8 = 59.4$$

(v) **限界値外の値を外れ値として，その値を箱ひげ図にプロットする．小さい方からの2つの値** 37.0, 40.0 **と大きい方からの2つの値** 66.0, 64.0 **が外れ値となる．**

(vi) **限界値内の最小値** (43.7) **と最大値** (58.0) **まで箱の両側から線** (**ひげ**) **を引く．**

箱ひげ図を見ると，両側に2個ずつの外れ値であることが分かる．

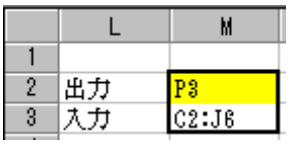
箱ひげ図は，外れ値を見出すのに役立つ他に，分布の対称性（ひずみの有無）や尖り度を見ることができる．中央値が箱の中央付近にあるか，ひげの長さがほぼ同じかで，分布の対称性からの外れ具合が分かる．また，箱の大きさとひげの長さとの比較により，尖った分布か，平坦な分布かを判断できる．

上に述べた手順に従い，表示3.16 **の上の表を作成すると，それから方眼紙の上に箱ひげ図を描くことができる．しかし，現実のデータに対して毎回これを実行するのは煩雑である．そこで，簡単に箱ひげ図を作成する** Excel **の** VBA **マクロを作成した．**

表示3.17 **に示すように，マクロに渡すパラメータを指定する．**

箱ひげ図を描くためには，中央値，四分位値などを計算し，箱ひげ図を構成する直線などの座標値の表を作成しなければならない．最初のパラメータとして，計算表を出力する位置（ここでは，P3**）を入力する．計算表を出力する領**

表示3.17: VBA マクロによる箱ひげ図の作成のためのパラメータ



	L	M
1		
2	出力	P3
3	入力	C2:J6

域は4列で，縦に長い範囲が必要である．マクロの実行によって，前の記録が消去されるので注意する．

その下に，データの記録されているセルの範囲（ここでは，C2:J6）を指定する．データの範囲は，1行，または，1列である必要はなく，複数行・複数列であってもかまわない．また，その範囲内に空白があってもかまわない．

出力位置の入力されているセル（表示3.17では，背景色が付けられている）をクリックしてから，マクロ「箱ひげ図」を実行する．横軸の目盛りを修正すると，表示3.16 下の箱ひげ図が得られる．

(3)　層別箱ひげ図

層別してヒストグラムを作成することの必要性は，すでに述べた．しかし，ヒストグラムは2次元のスペースを取るので，層の個数が多いときには大きな図になってしまうという欠点がある．

それに対して，箱ひげ図は幅が狭いので，複数の層別グラフを一覧で眺めるときに便利である．

前項で説明したマクロは，複数組のデータがあるとき，目盛りを揃えて，複数個の箱ひげ図を描く機能がある．

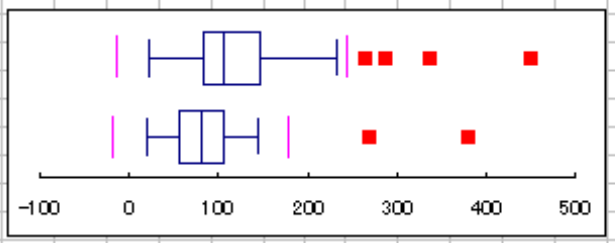
表示3.18の左上に示す2組のデータについて，箱ひげ図を描く．

表示3.18の右上に示すようにパラメータを指定する．

計算表の出力先の右のセル（表示3.18 では Q3）に T を入力すると，縦の箱ひげ図が横に並んだ出力が得られる．

表示3.18: VBA マクロによる箱ひげ図の作成

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
3	×1	19	52	61	40	91		×2	135	232	220	107	97		出力	S2	
4		86	53	107	124	82			76	84	113	336	88		入力	B3:F12	
5		132	59	123	76	27			187	43	92	95	65			I3:M12	
6		85	131	44	125	50			128	119	88	450	168				
7		85	106	72	78	30			38	94	138	264	52				
8		97	32	270	103	68			63	38	162	105	22				
9		81	95	101	56	30			148	146	145	155	98				
10		380	24	145	88	122			82	287	79	213	63				
11		70	111	82	66	82			110	176	111	97	84				
12		115	56	100	34	73			65	87	127	64	107				
13																	
14																	
15																	
16																	
17																	
18																	
19																	
20																	
21																	



■　演習 8　演習 6 のデータについて，箱ひげ図を作成せよ．なお，分布にはどのような特徴が見られるか．■

(4)　対数変換

表示3.18の箱ひげ図を見ると，中央値が箱の左に寄り，外れ値が大きい方に集中している．このようなデータは対数正規分布に従う場合が多い．

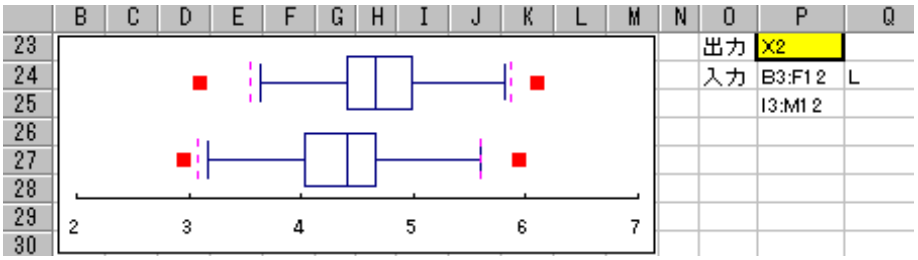
そこで，対数変換してから箱ひげ図を描いてみる．

提供する箱ひげ図のVBA マクロには，対数変換してから箱ひげ図を描く機能が含まれている．

最初の入力データ位置を指定するセルの右のセル（表示3.18 の Q4）に L を入力して，マクロを実行すると表示3.19の箱ひげ図が得られる．

対数変換値の箱ひげ図は左右対称に近くなり，外れ値が少なくなった．

表示 3.19: 対数変換値の箱ひげ図



本日のまとめ

箱ひげ図を初めて学習した受講生もいるであろう．Excel による箱ひげ図の作成も学習した．あまり馴染みのない箱ひげ図であるが，分布からの外れ値の発見や左右対称性の評価などには大きな威力のあることが分かったであろう．第3章では，データの分布の様子を調べる方法として，ヒスグラム，箱ひげ図を紹介した．統計的な指標とこれらを活用することにより，母集団の様子が手にとるように分かるであろう．

3.5　補足

(1)　度数分布による平均値・平方和の計算(1)

そろばんと電卓で平均値や標準偏差を計算していた時代には，n が大きいデータについて，個々の x_i から平均値や標準偏差を計算することは不可能に近かった．古い統計のテキストでは，まず，度数表を作成し，それから平均値や平方和・平均平方・標準偏差を計算する方法が説明されていた．

コンピュータが身近になった現代では，この計算手順を説明する必要はまったくなくなったが，度数表に集計された情報しか得られないときのために，度

数表から平均値と標準偏差を計算する方法を説明する.

まず, 表示3.2 (p.64) **の度数表から, 平均値と標準偏差を求める.**

表示3.2 **から** x **と度数** f **を** A,B **列に転記する.**

表示 3.20: **交通事故による死亡者数の平均値と標準偏差**

	A	B	C	D
1	x	f	fx	(x-平均)^2
2	0	29	0	168.4
3	1	76	76	151.1
4	2	70	140	11.8
5	3	59	177	20.5
6	4	41	164	103.7
7	5	15	75	100.6
8	6	7	42	90.2
9	7	5	35	105.3
10	8	1	8	31.2
11	9	2	18	86.9
12	合計	305	735	869.77
13	平均	2.41	2.41	
14	平方和	869.77		
15	平均平方	2.86		2.86
16	標準偏差	1.69		1.69

平均値は次の式で計算される.

$$\overline{x} = \frac{\sum_{i=1}^{m} f_i x_i}{\sum_{i} f_i}$$

ここで, m **は度数表の行数(この例では** $m = 10$**) である.**

C2 **に** =A2*B2 **を入力して** $f_1 x_1 = 0 \times 29 = 0$ **を計算し, 下にコピーする.** B12 **に** =SUM(B2:B11) **を入力して** f **の合計を求め, 右にコピーする.** fx **の合計** 735 **を** f **の合計** 305 **で割ると** C13 **に平均人数** $\overline{x} = 2.41$ **が得られる.**

平方和は次の式で計算される.

$$S = \sum_{i=1}^{m} f_i (x_i - \overline{x})^2$$

この計算のために，D2 に B2*(A2-C$13)^2 を入力して，$f_1(x_1-\overline{x})^2$ を求め，下にコピーする．D12 に D2:D11 の合計を計算すると，平方和 $S=869.77$ が得られる．S を自由度 $n-1$ で割ると平均平方が，さらにその平方根を取ると標準偏差が得られる．

このようにして，度数表から計算した平均値と標準偏差は，表示3.1 (p. 63) の元データから直接計算した値と完全に一致する．

表示3.20 では，計算の過程を分かりやすくするために，C, D 列の fx, $f(x-\overline{x})^2$ を通して平均と平方和を計算したが，A，B 列から直接 平均や平方和を計算することができる．次に，B13:B14 に入力されている関数を示す．

```
B13:  =SUMPRODUCT(B2:B11,A2:A11)/B12
B14:  =SUMPRODUCT(B2:B11,(A2:A11-$B$13)^2)
```

昔は §2.5 の式(2.13) (p. 56)（修正項 CT を用いる式）に対応する式

$$S=\sum_{i=1}^{m}f_ix_i^2-\frac{\left(\sum_{i=1}^{m}f_ix_i\right)^2}{\sum_{i=1}^{m}f_i}$$

が用いられていた．この方法でも平方和を計算することができるが，計算の誤差が大きくなる危険があるので，すすめられない．

(2)　度数分布による平均値・平方和の計算(2)

次に，表示3.7 (p.69) の体重データの平均値と標準偏差を，表示3.8 の度数表から計算してみよう．このデータは前項の死亡者数と異なり，連続量である．

級代表値を x と見なして，前項とまったく同様の手順で計算した結果を表示3.21 に示す．

このようにして求めた結果を，表示3.7 から直接求めた値と比較する．

度数表にまとめたために，両者の間にはわずかの違いが生じる．

度数表から求めた平均平方は個々の値から求めた平均平方よりも，ほとんどの場合，小さくなる．その差は群の幅（ここでは2.0) によって異なり，平均的

表示 3.21: **中学生の体重の平均値と標準偏差**

級代表値	f	fx	f(x-平均)^2
37.95	1	38.0	171.8
39.95	0	0.0	0.0
41.95	4	167.8	331.9
43.95	6	263.7	303.2
45.95	9	413.6	234.9
47.95	18	863.1	174.0
49.95	21	1049.0	25.8
51.95	16	831.2	12.7
53.95	11	593.5	91.9
55.95	10	559.5	239.2
57.95	6	347.7	284.9
59.95	4	239.8	316.2
61.95	3	185.9	355.8
63.95	1	64.0	166.2
合計	110	5616.5	2708.69
平均	51.1	51.1	
平方和	2708.69		
平均平方	24.85		24.85
標準偏差	4.99		4.99

表示 3.22: **個々のデータからの直接計算と度数表からの計算の差**

	平均値	**平均平方**	**標準偏差**
個々の値から	50.9	25.56	5.06
度数表から	51.1	24.85	4.99

に 幅$^2/12$ **となる．したがって，群の幅が広いとき，度数表から計算した値は小さめになることに注意する必要がある．**

平均値についてはこのような傾向的な差は生じない．

(3) 自然対数

第1**単元** §1.6 (1) **で 対数 について基本的な説明をした．そこでは，** 10 **を底とする対数（常用対数）が説明され，**x **の常用対数は** $\log(x)$ **で表わされた．**

第1単元 §3.4 (3) で正規分布の式の中に e という記号が表われ,「自然対数の底 2.71828 である」と説明された．x の自然対数 (natural logarithm) は，頭文字の n を用いて，$\ln(x)$ で表わされることが多い．

自然対数を用いると，数理的に便利な点が多々あるので，自然対数が理論の世界では広く用いられる．第1単元で取り上げた正規分布や第3単元で取り上げるポアッソン分布の式はその例である．

以下の説明は，自然対数についてもう少し知りたい人のためのものである．

自然対数 $\ln(x)$ と常用対数 $\log(x)$ の間には

$$\ln(x) = 2.303\log(x), \quad \log(x) = \ln(x)/2.303$$

の関係がある．ここに，2.303 は $\ln(10)$ として求められる．

e の 2.71828 は $(1-1/n)^n$ で，$n \to \infty$ とした極限の値である．

x を 1 の前後でわずかに変化させると，$\ln(x) \approx 1 - x$ という性質がある．

この2つの性質は,「第2単元.XLS」の自然対数のシートで確かめることができる．

(4)　打切りデータ

観測された値が，ある値以上であることは分かっても，その具体的な値が分からない場合がある．

例えば，目盛りの最高値が100kg である体重計に 100kg を超える人が乗ったとき，100kg 以上であることは分かっても，本当の体重は分からない．

このような値が得られたとき，どのように取り扱うかは大変むずかしい問題であって，特別な方法が必要となる．その内容は本講座の範囲を超えているので紹介はしない．信頼性や生存時間に関する市販の参考書に目を通してほしい．

このようなデータはいろいろの場で発生し，その発生の仕方も多様である．

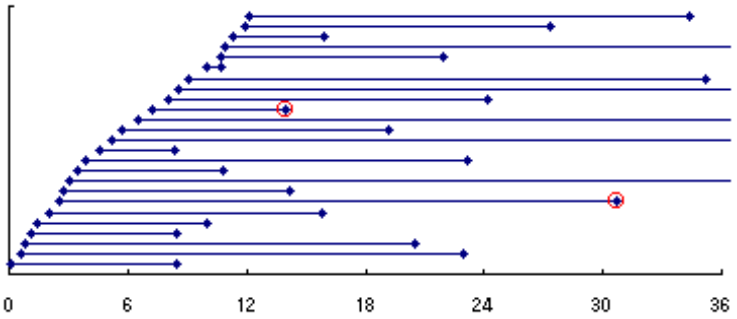
体重の例のように，一定の値以上は値が不明であるというのは，寿命試験で現われる．ある銘柄の蛍光灯の平均寿命を調べるために，10 本の蛍光灯に点灯

し，切れるまでの時間を測定する．寿命は千時間を超えるため1年以上の時間を要してしまう．そこで，電圧を100ボルトよりも高く設定して，実験する．このような試験を高負荷試験という．電圧と寿命との関係は予め詳しく調査をしておいて，高負荷試験の結果を通常の使用条件での寿命に換算する．このような工夫をしても，なお試験期間が長くなるので，例えば，100時間で試験を打切り，その時点でなお点灯している蛍光灯の寿命は100時間以上とする．このような打切り方を 定時打切り という．

このように，一定時間で打切るのではなく，全体の半分の蛍光灯に寿命がきたら打切るという方法もある．このような打ち切り方を 定数打切り という．この方法は，試験の終了時点があらかじめ分からないので，試験計画が立てられないという不便があるので，余り使われない．

上の例は，蛍光灯の点灯開始時点が揃っており，計画が可能である．それに対して，がん手術の終了後の寿命を調べるときは，手術の時点は患者によって異なる．例えば，ある年1月から調査を開始し，その年に手術をした患者について追跡調査をする．調査開始から3年後に調査を打切って，結果をグラフ化すると表示3.23が得られた．

表示 3.23: 手術後の生存月数



横棒の左端が手術時点で，右端が死亡時点である．

試験終了時点で生存している患者が5人いる．これらの患者は手術時点が違

うから，一概に何ヵ月以上ということはできない．したがって 定時打切り とは異なる．

図3.23 では2 本の横線の右端にoのマークがついている．この2 人は，がん手術とはまったく無関係な理由（交通事故，食中毒，など）で亡くなった患者であるから，術後寿命としては，下限しか分からない．

このような場合は ランダム打切り と呼ぶ．

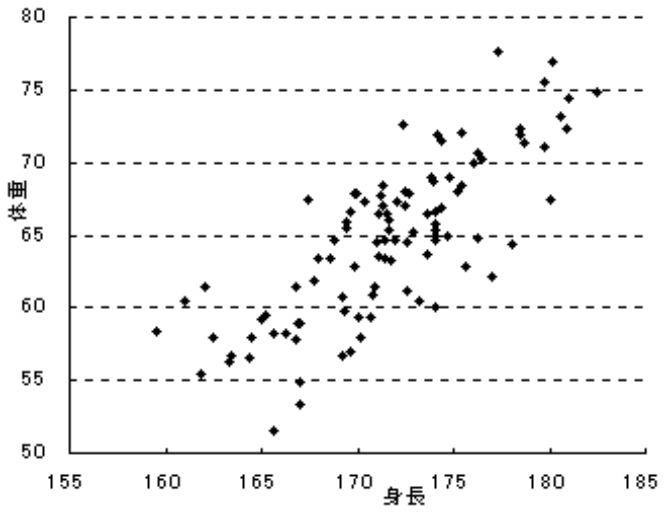
4 相関

4.1 相関・回帰の意味と散布図

(1) 身長と体重（相関の例）

ある人の身長が分かったとしても，それからその人の体重を知ることはできない．同じ身長の人でも太った人，やせた人がいるからである．しかし，身長の高い人は特に太っていなくても意外に体重の重いことはしばしば経験するところである．背が低くて体重の重い人，背が高くて体重の軽い人はもちろんいるが，それらはまれな存在であって，平均的には，あるいは統計的には，身長の高い人は体重が重いということができる．また逆に，体重の重い人は身長が高いということもできる．相関 はこのような関係を取り扱うものである．

表示4.1: 身長と体重の相関



表示4.1 は成人男子100 人について身長と体重を測定した結果をグラフに表わ

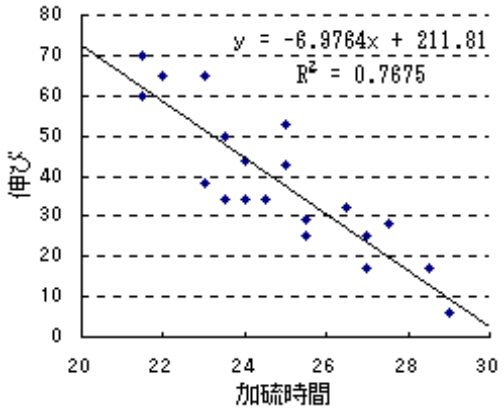
したものである．横軸には身長が，縦軸には体重が取ってある．このようなグラフを 散布図 (Scatter Diagram) と呼ぶ.

同様な例として，父親の身長と男の子の身長の関係，数学の成績と国語の成績との関係などをあげることができる.

(2)　製造条件と製品品質（回帰の例）

表示4.2 は，ゴムの製造条件（加硫時間，ゴムとイオウとを反応させる時間）と製品品質（伸び）の関係を示したものである（元の数値は表示4.3 に示す)．横軸には加硫時間が，縦軸には伸びが取ってある．この図から加硫時間を長くすると伸びが減少することが分かるであろう．また，同じ加硫時間でも製品の伸びは一定でなくバラツキをもっていることにも気づくであろう.

表示4.2: 加硫時間と製品の伸びの関係



この例を，体重，身長の例と比べると，2つの量の間に関係があるという点では同じであるが，その裏にある構造には違いがある.

身長，体重の場合は，身長の高い人は体重が重く，逆に体重の重い人は身長が高いというように，2 つの量が相互に関係し合っていたのに対し，ゴムの例の場合は，2つの変数のうちの一方が原因で，他方がその結果という関係を持っている.

すなわち，加硫時間の長短が製品の伸びに影響したのであって，その逆の関係は考えられない．

ゴムの加硫時間 x を一定に保って加硫処理をした場合，でき上がったゴムの伸びがいつも全く同じというわけにはいかない．原料のゴム品質のバラツキ，加硫時間以外の処理条件（例えば加硫温度のバラツキ）などによって製品の伸び y は，ある平均値のまわりにばらつく．その平均値が x によってどう変わるか，すなわち，ある加硫時間 x のとき伸び y の平均がいくらになるかを表わす式を，y の x に対する **回帰式** という．

例えば，ゴムの伸び y は加硫時間 x の一次式で，

$$y = 211.81 - 6.9764x \tag{4.1}$$

と表わすことができる．

データから回帰式を求めて，2 つの変数の間の関係を解析する方法を **回帰分析** と呼ぶ．この式の求め方と，直線の引き方の簡単な説明は次項で説明する[1]．

このような場合，x を **独立変数（説明変数）**，y を **従属変数（目的変数）** と呼び，グラフを描くときには，独立変数を横軸に，従属変数を縦軸に取るのが普通である．

(3)　正の相関と負の相関

身長と体重の関係のように，身長の高い（大きい）人は体重が重い（大きい）という傾向がある場合は，両者の間に **正の相関** があるという．

逆に，加硫時間と伸びの関係のように，加硫時間が長い（大きい）とき伸びにくい（小さい）という場合は，両者の間に **負の相関** があるという．

(4)　Excel による散布図の描き方

昔は，方眼紙に縦軸と横軸を引き，目盛りをつけて，データ表を見ながら×または○を記入したが，現在では，Excel で簡単に散布図を描くことができる．

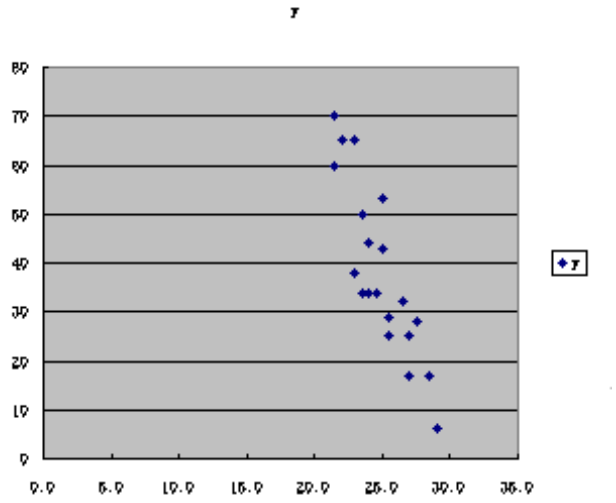
[1] 回帰分析については，「現代統計実務講座」第 5 単元を参照．さらに詳細については「多変量解析実務講座」第 2 単元，第 3 単元参照．

表示4.3の左は，表示4.2のもとになったデータである．

このデータを使って，散布図の描き方を学ぶことにする．

表示4.3: 加硫時間と伸びの関係

	A	B	C
1			
2		加硫時間	伸び
3		x	y
4	1	21.5	60
5	2	21.5	70
6	3	22.0	65
7	4	23.0	38
8	5	23.0	65
9	6	23.5	50
10	7	23.5	34
11	8	24.0	44
12	9	24.0	34
13	10	24.5	34
14	11	25.0	53
15	12	25.0	43
16	13	25.5	25
17	14	25.5	29
18	15	26.5	32
19	16	27.0	25
20	17	27.0	17
21	18	27.5	28
22	19	28.5	17
23	20	29.0	6



表頭を含むデータの範囲 B3:C23 を反転してから,「グラフウィザード」の「散布図」を選択する．

表示4.3の右の散布図が得られる（最初に出力されたグラフの寸法を調整してある）．

このままでは，実用にならないので，整形をする必要がある．整形した結果が前に示した表示4.2 である．

整形した内容は次のとおりである[2]．

[2] 具体的な手順は，添付されたExcel ファイルに説明されている．

- グラフの形が少し横長なので，正方形に近い形にする．
- 背景色を消す．
- 横軸の目盛りの最小値を20，最大値を30，間隔を2に修正する．
- 両軸の目盛りの形式（小数点以下の桁数），文字の大きさを調整する．
- 水平の目盛り線を点線に変更する．
- 表名と凡例を削除する．
- 横軸と縦軸の変数名を追加する．

(5)　回帰式と回帰直線

表示4.2に引かれている右下がりの直線と回帰式は，次の手順で求めることができる．

- 散布図の点をクリックして点を反転させてから，右クリックし「近似曲線の追加」を選択する．
- 近似曲線の追加画面の「種類」から「線形近似」を選択する．
- 「オプション」から「グラフに数式を表示する」と「グラフにR-2乗値を表示する」を選択する．

R-2乗値 の意味は次の節で説明する．

通常回帰式は式(4.1)のように，最初に定数を，次に x の項を並べるが，Excelでは逆の順序になっていることに注意．

本日のまとめ

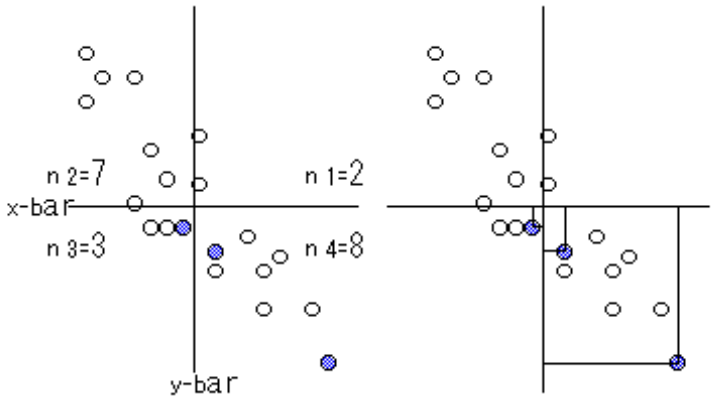
量的な2つの変数間の関係の強さを調べるための方法として，散布図を紹介した．また，散布図をExcelで描画する方法についても示した．2つの変数の関係には，結果と結果の直線的な結びつきを意味する相関関係と，原因と結果の直線的な因果関係を調べる回帰直線について，その概要を紹介した．明日は，相関関係を定量化した指標である，相関係数を紹介する．

4.2　相関係数

(1)　相関の強さの定量評価

表示4.2の散布図に $(\overline{x},\overline{y})$ を通る十字線を引くと表示4.4 の左の図が得られる.

表示4.4: 点の4つの象限への分布



平均値 $\overline{x}$ の左右の点の個数は共に10個である．平均値 $\overline{y}$ の上下の点の個数は9個と 11個でほぼ等しい.

もし，x と y の間に関係がなければ，点は十字線で区切られた4つの領域にほぼ均等に散らばるであろう.

この4つの領域は 象限 と呼ばれ，右上から反時計周りに 第1象限，第2 象限，第3象限，第4象限 と呼ばれる．4つの象限に含まれる点の個数を $n_1,\ n_2,\ n_3,\ n_4$ で表わすことにする．もし点が十字線の上にあるときは，両側に 0.5 ずつ分ける．このデータでは,

$$n_1 = 2,\ n_2 = 7,\ n_3 = 3,\ n_4 = 8$$

である.

この例のように，負の相関がある（x が大きくなると y が小さくなる）場合には，$n_1,\ n_3$ が少なく，$n_2,\ n_4$ が多い傾向がある.

そこで，これらの個数のアンバランスによって 関係の程度を表わすことを考える．

$$\tilde{r} = \frac{(n_1 + n_3) - (n_2 + n_4)}{n_1 + n_2 + n_3 + n_4} = \frac{(2+3) - (7+8)}{2+7+3+8} = \frac{-10}{20} = -0.5 \qquad (4.2)$$

式(4.2)で計算される $\tilde{r}$ は[3]，正の相関 の時には正の値を，負の相関 の時には負の値を取る．

また，相関関係が強く，第1，第3象限（または第2，第4象限）に全部の点が集まったときには± 1.0 となる（式(4.2) で，個数の差を総個数で割っているのは，この性質をもたせるためである）．

すなわち，$\tilde{r}$ によって，相関の方向と強さを表わすことができる．

(2) 相関係数

しかし，表示4.4 の左のグラフの ● の3つの点を考えると，いずれの点も $\tilde{r}$ に対する影響は同じである．

十字線から離れている点は十字線に近い点よりも，直感的に感じられる相関の強さに大きく影響している．そこで，点の個数ではなく，点から十字線に下ろした垂線と十字線でできる四辺形の面積の合計を用いることにする（表示4.4 の右参照）．すなわち，第1，第3象限の面積から第2，第4象限の面積を引いたものを考える．i 番目の点から十字線までの距離は $|x_i - \overline{x}|,\ |y_i - \overline{y}|$ で表わされる．

各象限の点について $(x - \overline{x})(y - \overline{y})$ の符号を調べると次のようになる．

象限	$(x - \overline{x})$	$(y - \overline{y})$	$(x - \overline{x})(y - \overline{y})$
1	+	+	+
2	−	+	−
3	−	−	+
4	+	−	−

これから，$(x_i - \overline{x})(y_i - \overline{y})$ を i について足し合わせると，第1，第3象限の面積から第2，第4象限の面積を引いたものになる．

[3] r は次に説明する相関係数の記号として用いられる．式(4.2) はそれと区別するために，中央値で用いた ～ をつけた．

この面積は，両軸の目盛りの取り方によって変化するので，それぞれの標準偏差 $s_x,\ s_y$ で割って基準化する．すなわち，両軸に $x,\ y$ の偏差値 $z_x,\ z_y$ を取る．

また，平方和を自由度 $n-1$ で割って1個当たりのバラツキを求めたように，$n-1$ で割ることにする．

このような考えで導かれた値を 相関係数 と呼び，r で表わされる．

$$r = \frac{\displaystyle\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right)}{n-1} = \frac{\displaystyle\sum_{i=1}^{n} z_{xi}z_{yi}}{n-1} \tag{4.3}$$
$$= \frac{\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{(n-1)s_x s_y}$$

上の式の分母の標準偏差 $s_x,\ s_y$ は $\sqrt{S_x/(n-1)},\ \sqrt{S_y/(n-1)}$ で求められるので，これを代入すると，

$$r = \frac{\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{S_x S_y}} \tag{4.4}$$

と書くことができる．式(4.4)の分子は x と y の 積和 と呼ばれ，S_{xy} で表わされる．

x の平方和 S_x は x と x の積和と見ることができるので，S_{xx} という記号を用いることがある．この記号を使うと，相関係数は

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \tag{4.5}$$

と表わすことができる．

(3)　相関係数の計算

式(4.5)の定義に従い，Excelで r を求める手順を表示4.5に示す．

計算の手順は次のとおりである．

- $x,\ y$ の平均を求め，偏差 $x-\overline{x},\ y-\overline{y}$ を計算する．
- 平方和 $S_{xx},\ S_{yy}$ を SUMSQ 関数で計算する．

表示4.5: 相関係数の計算（データの途中省略)

	x	y	$x-\bar{x}$	$y-\overline{y}$
1	21.5	60	−3.4	21.6
2	21.5	70	−3.4	31.6
…	…	…	…	…
19	28.5	17	3.7	−21.5
20	29.0	6	4.2	−32.5
平均	24.9	38.5	0.0	0.0
平方和			93.05	5900.95
積和			−649.15	
相関係数	−0.876		−0.876	

- **積和 S_{xy} を SUMPRODUCT 関数で計算する.**
- **式(4.5) で相関係数を計算する.**

以上が，手順を追って相関係数を計算する過程である.

平方和が DEVSQ 関数で直接求められたように，相関係数も CORREL 関数を使って直接求めることができる．すなわち,

=CORREL(x の範囲, y の範囲)

とする.

こうして得られた相関係数 $r=-0.876$ の2乗 $r^2=0.7675$ が表示4.2(p.94) の中で求められていた R-2乗 である．これは，寄与率 または 決定係数 と呼ばれ，2つの変数の関係の強さを表わす値として用いられることが多い[4].

(4) 散布図と相関係数の関係

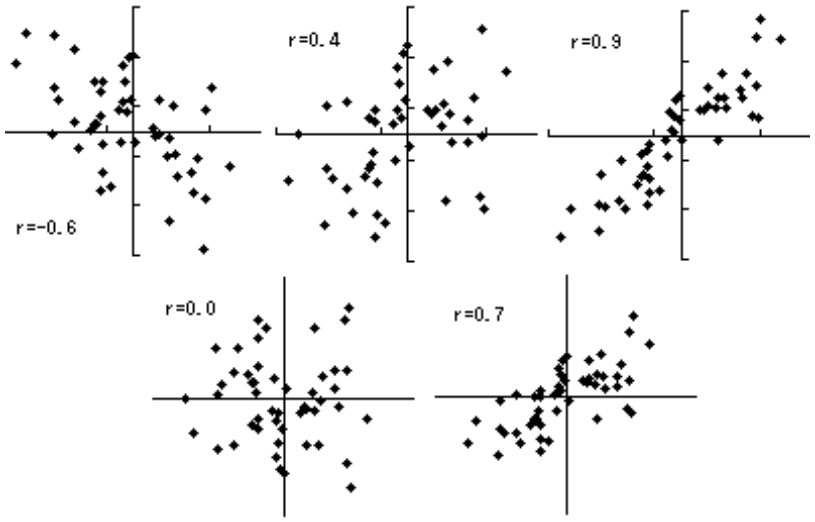
表示4.6 には相関係数が $-0.6,\ 0.0,\ 0.4,\ 0.7,\ 0.9$ の散布図を示す.

表示4.6 から，相関係数の大きさと散布図のパターンのおよその関係は理解できるであろう．しかし，相関係数が 0.6 であるという数値が得られたとしても，その具体的な大きさのイメージがつかめない.

そこで，散布図に楕円を当てはめて，楕円と相関係数の関係を説明する.

[4] 寄与率（決定係数）の詳細については,「現代統計実務講座」第5単元,「多変量解析実務講座」第2単元，第3単元参照.

表示4.6: 異なる相関係数をもつ散布図の例



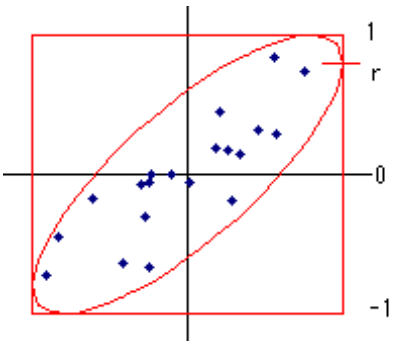
2つの変数が共に正規分布に従い，2つの変数の間に直線関係があるとき，2つの変数の分布を **2** 次元正規分布 という．表示4.7 は母相関係数が 0.8 の母集団からのサンプルの散布図である．この点に楕円を当てはめ，平均を通る2つの軸と楕円に外接する長方形を描き，楕円との接点を求める．平均の座標を 0，長方形の座標を ± 1 とすると，接点の座標として，相関係数の近似値を知ることができる．

逆に，平均値と相関係数が分かっているとき，観測値の大部分 (約 90%) が含まれる近似楕円を次の手順で求めることができる．(表示4.8)

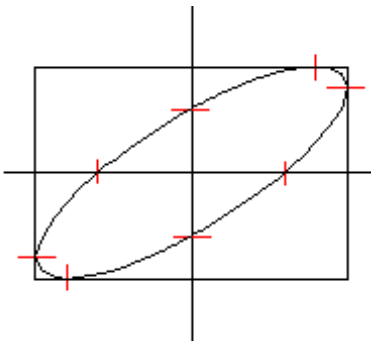
- 平均値 $(\overline{x}, \overline{y})$ を通る十字線を引く
- $\overline{x} \pm 2s_x$, $\overline{y} \pm 2s_y$ の4辺形を描く
- 4辺形の上に楕円の接点の位置 (r) に印をつける (4個所)
- 十字線の上，$\pm\sqrt{1-r^2}$ の位置に印をつける (4 個所) [5]
- 上に求めた8つの点を通る楕円を描く

[5] 楕円と四辺形との接点の位置 (r)，楕円と十字線との交点の位置 $(\pm\sqrt{1-r^2})$ の導き方についての説明は省略する．

表示4.7: 2次元正規分布



表示4.8: 楕円を描く



相関係数の意味を理解するために，次の問題を考える．

ある化学工場で製品品質 y のバラツキが大きく，標準偏差 σ が 10 である．バラツキを小さくする必要がある．y を特性とする特性要因図を作り，y と相関のある要因を探索したところ，水温との相関係数が 0.6 であった．冷凍機と加熱機の設備投資をして水温を年中一定とすると，y の標準偏差 σ はいくらになるであろうか？

σ が 60% 減少し，$\sigma = 4$ になると考えるのは誤りである．正しくは，$\sigma = \sqrt{1-0.6^2}\,\sigma = 8$ となるに過ぎない．これは，表示 4.8で $r = 0.6$ としたときの十字線との切片の位置に相当する．

本日のまとめ

今日は，相関係数の意味と計算方法を学習した．相関係数は，変数間の直線的な結びつきの強さを定量化したものであった．相関係数は，変数間の結びつきを定量的に判断するには，大変便利な指標であるが，変数間のバラツキの様子までは，教えてくれない．このため，データ解析では，必ず散布図と相関係数とを1セットとして並べて見ていくと 誤解が少ないであろう．明日は，出力された散布図や相関係数の見方や考え方について紹介する．

4.3　相関係数，散布図の見方

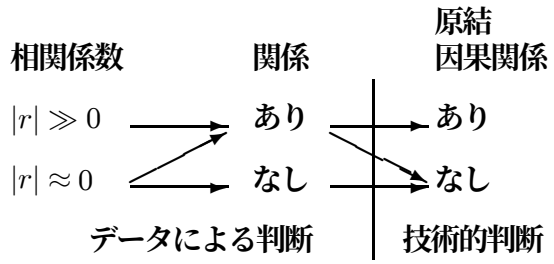
(1)　相関係数と関係の有無

x と y の相関係数が 0 より大きく離れているとき[6]，x と y との間にはなんらかの関係があると考えられる．

それでは，逆に，相関係数が 0 に近いとき，両者には関係がないといえるかというと，必ずしもそうはいえない．

この関係を表示4.9の左の「右上向きの矢印」で示す．

表示4.9: 相関係数，関係，因果関係



相関係数が 0 に近いから，直ちに，2つの変数の間に関係がない（独立である）と判断してはならない．

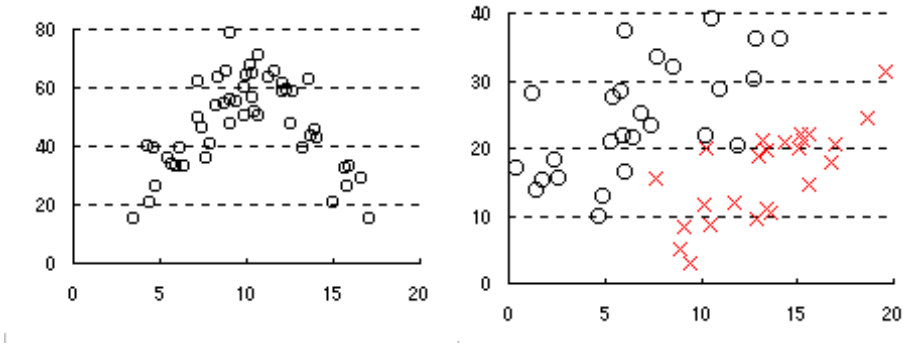
相関係数は，直線関係の強さを測る指標 であって，表示4.10の左のように2つの変数間に密接な関係があっても相関係数はそれ程1に近づかない．曲線関係があるとき相関係数は小さくなる．山または谷が中央にあると相関係数はほとんど 0 になる．

また，異質な集団の集まりであると，全体としては無相関であっても，集団ごとに散布図（層別散布図，§4.4 で説明する）を描くと相関を発見できることがある．表示4.10の右参照．

これらの例が示すように，2変数の関係は，相関係数の値を見るだけでなく，散布図を描いて良く観察することが必要である．そのための手段は §4.4 で説明

[6] r の絶対値がどのくらい大きいとき，確かな結論が出せるかについては，第4単元 §4.4 で取り上げる．

表示4.10: 関係があるが，$|r| \approx 0$ となる場合



する．

(2)　関係と因果関係

因果関係とは，原因と結果の関係である．相関係数や散布図から2つの変数の間に関係のあることが分かっても，直ちにそれが原因と結果の関係であると判断してはならない．

関係があっても因果関係ではない場合のあることを，表示4.9 の右の「右下向きの矢印」に示す．

図の下に示すように，因果関係があるかどうかは 技術的判断が必要である．

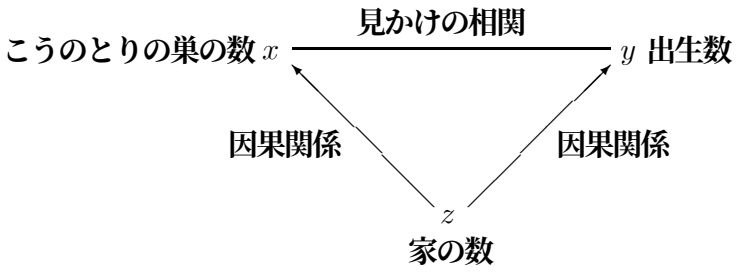
因果関係はないが，相関があるとき 擬似相関 があるという．

(3)　擬似相関の例 (1)

ヨーロッパのある町で，十数年間にわたって，こうのとりの巣の数を調査し，毎年の赤ん坊の出生数との相関を調べたところ，両者の間には明らかに相関のあることが認められた．この事実から「赤ん坊はこうのとりが運んでくる」という話が裏付けられるだろうか．この説が正しくないのは当然であるが，この事実をどう説明したら良いであろうか．

実は，第3の変数，町の大きさまたは家の数（z）がこの数年間に大きくなり，

表示4.11: 擬似相関(1)



それが原因となって，軒先に作られた巣の数（x）が増え，また，世帯数が増えれば必然的に出生数（y）も増えたのである．すなわち，x と z 及び y と z の因果関係によって，x と y の間に見かけの相関が現われたのである．

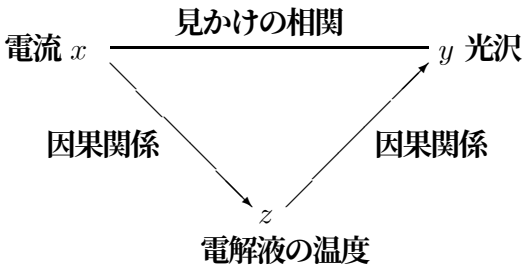
(4)　擬似相関の例(2)

電気メッキ工場で，電流を増やしてメッキ時間を短縮して，生産性を上げることを検討した．ただし，そのためにメッキの品質を損なうことは避けたい．

電流 x とメッキの光沢 y の関係を調べた結果，電流を増やすと光沢が下がるという負の相関が認められた．これから，電流を増やす生産性向上は諦めなければならないだろうか．

電流と光沢の因果関係を技術的に検討し，「電流を増やすと，発熱量が増え，電解液の温度が上昇する．電解液の温度がメッキ面の光沢に影響を与える」というメカニズムが考えられた．この関係を表示4.12に示す．

表示4.12: 擬似相関(2)



電流を増やしても，電解液の温度が上がらないようにすれば，品質に影響を与えずに生産性を向上できるであろう．例えば，電解液を冷却装置に循環させて一定温度に保つことが考えられる．

上にあげた2つの例のように比較的単純な場合ならば，x と y のデータに現われた相関から，直ちに，x と y の因果関係に結びつける過ちはまずおかさないであろうが，実際直面する問題についてはこのような誤解がかなり見受けられるから十分気をつけなければならない．特に時系列データの場合はこの種の誤りを犯すことが多い．

■　演習 9　手書き原稿をコンピュータに入力して講習会のテキストを作成する．2人の講師の原稿とテキストの誤植の関係を見ると，きれいな原稿のテキストの方に誤植が多かった．なぜだろうか？　■

(5)　入学試験の成績と入学後の成績の相関について

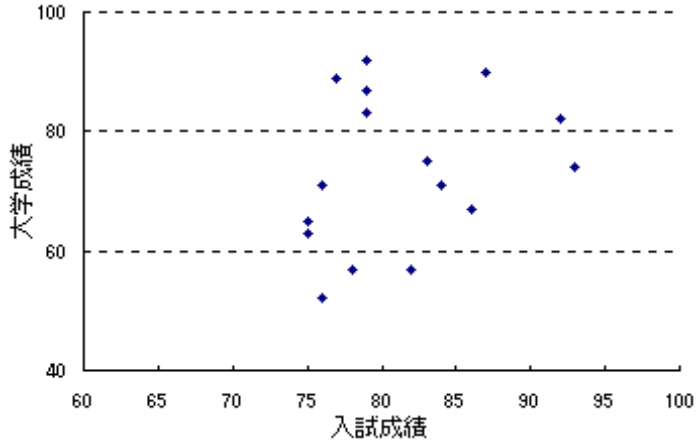
入学試験の成績と入学後の成績とはあまり相関がないという話が，多くの大学でよくいわれている．大学生の両方の成績の散布図を描くと，表示4.13のようになった．

相関係数は 0.350 と小さく，この散布図を見るかぎりにおいては，最初に掲げた話は正しそうである．それは，厳しい受験戦争を勝ち抜いてきた反動の心のゆるみが原因しているのかもしれない．ただこのことのみで，入学試験無用論を掲げるのは大いに問題がある．それは次のような理由からである．

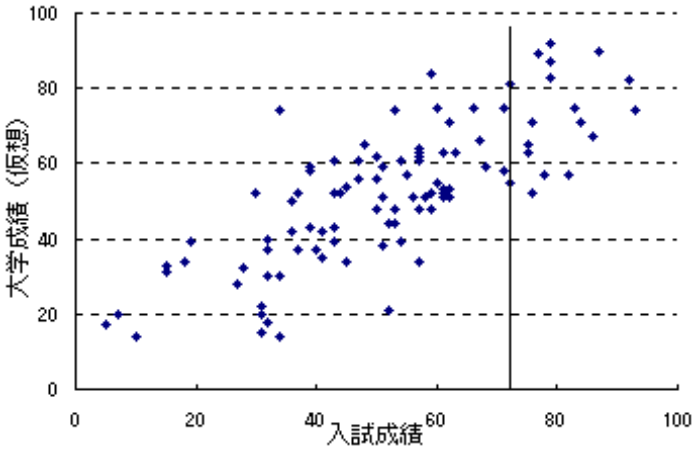
そもそも，入学後の成績が分かるのは入学試験に合格した者，すなわち，合格点（表示4.13 では 75 点）以上についてのみである．もし，不合格であった者が，仮に大学へ入ったとしたときの，大学での成績が分かったとして，両方の成績の散布図を書いたとすれば，表示4.14 となる．

この散布図を見ると，入試の成績と入学後の成績の間には相関がみられるはずである．このようなことは実際には不可能なことであるが，少なくとも次のことがいえる．すなわち，入学試験は不合格者をも対象に行われたものであって，その不合格者を除外したデータで，入学試験の是非を論ずることはできない．

表示 4.13: 入学試験の成績と入学後の成績



表示 4.14: 入学試験の成績と入学後の成績（仮想値）



相関係数をめぐってこの種の誤りがしばしばあるので，くれぐれも注意する必要がある.

■　演習 10　次の（　）に最も適当な言葉を下の用語から選びなさい.

2つの量的なデータをそれぞれ縦軸と横軸に取ってグラフにしたものを（1）

という．

描かれた（1）を眺めると，2つの量的関係をみることができる．

縦軸に取った変量と横軸に取った変量の間に，1つの変量が増えるともう一方の変量が増加する，あるいは減少するという直線的な関係があるとき，2つの変量間には（2）があるという．

（1）で打点が右上がりに集まっているとき，(3）があるという．

逆に打点が右下がりに集まっているとき，(4）があるという．

もし，このような関係が認められない場合には，その2つの変量は（5）であるという．

また，2つの変量の間に，一見（2）があるように見えても論理的なつながりはなく，たまたま2つの変量の両方とそれぞれ（2）がある第3の変量が存在することによって偶然に相関係数が高くなってしまう場合を（6）という．

1．因果関係　2．散布図　3．擬似相関　4．正の相関　5．負の相関

6．相関関係　7．無相関　8．弱い相関　9．強い相関　■

本日のまとめ

データのみからは，2つの変数間に因果関係があるのかは分からない．因果関係があるかどうかは，その問題の背景にある知見が必要であることが強調された．また，第3の変数の影響により，見かけの相関が生じている関係，擬似相関の例を学習した．

同様な例として，都道府県人口と刑法犯罪件数とは正の相関関係があり，都道府県人口と電力消費量の間にも正の相関関係がある．この場合，電力消費量と刑法犯罪件数に正相関があるからといって，電力消費量が犯罪多発と因果関係があるといえるであろうか．

データ解析の結果を鵜呑みにしないで，日ごろから知識を蓄えて，良識的な判断ができるように心掛けてほしい．

4.4 種々の散布図

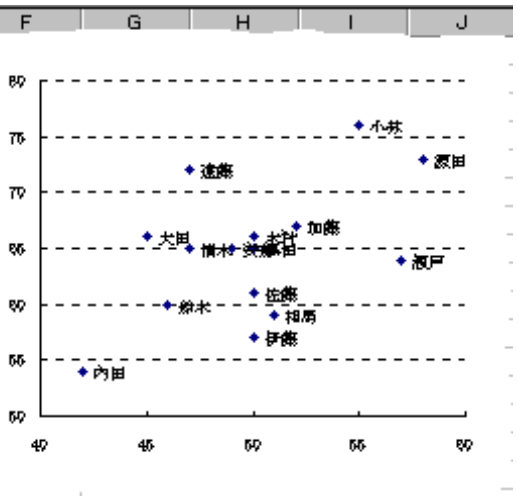
(1) 散布図の点にサンプル名を表示

散布図の点ともとのデータとの繋がりをつけたい場合が多い．すなわち，サンプル名から点の位置を知ったり，点がどのサンプルの値かを知りたい．散布図の点の傍にサンプル名を表示するためのマクロを作成した．

表示4.15に，データと出力を示す．

表示 4.15: サンプル名を追加した散布図

	A	B	C	D	E
3			横跳び	直跳び	背筋力
4		安部	49	65	150
5	伊藤	伊藤	50	57	167
6	内田	内田	42	54	174
7	遠藤	遠藤	47	72	205
8	大田	大田	45	66	155
9		加藤	52	67	140
10		木村	50	66	150
11		黒田	50	65	156
12	源田	源田	58	73	180
13	小林	小林	55	76	135
14		佐藤	50	61	163
15		清水	47	65	105
16	鈴木	鈴木	46	60	143
17	瀬戸	瀬戸	57	64	173
18	相馬	相馬	51	59	140



まず，通常の方法で 横軸に「横跳び」を，縦軸に「垂直跳び」を取って散布図を描く．

グラフをクリックしてから，マクロ「ラベル」を実行する．

「ラベルの範囲？」の問い合わせがでるので，B4:B18 のように，ラベル（サンプル名）の記録されているセルの範囲を入力する．

表示4.15の右の出力が得られる．

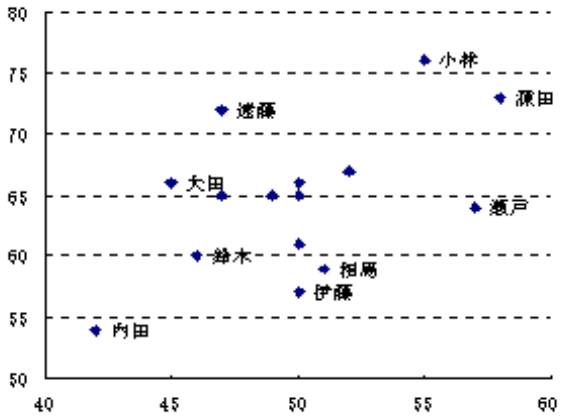
中心付近に点が集まり，サンプル名が重なって見にくい．

一般に散布図の中で，注目されるのは周辺の点である．そこで，周辺の点だけにサンプル名を表示したい．

そのために，表示したいサンプル名の列（表示4.15では列A）を準備する．簡単な方法としては，B列のラベル列をA列にコピーし，サンプル名を表示したくないセルを空白にする．

前と同様にマクロ「ラベル」を実行し，「ラベルの範囲？」として A4:A18 と入力する．表示4.16が得られる．

表示 4.16: 選択したサンプル名を追加した散布図



（注意） データを修正すると，グラフのプロット位置は自動的に変化するが，ラベルのリストを修正しても，グラフのラベルは変更されない．再度マクロを実行する．

(2) 層別散布図

表示4.17の左の3列は，中学高校生の性別・身長・体重データである（下の方は省略）．このままで身長と体重の散布図を描くと右上のグラフが得られる．

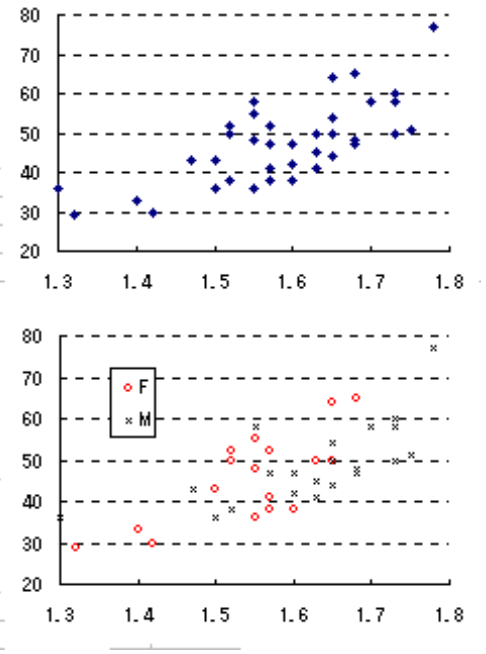
散布図のマークを男女で変えたい．

そのためには，身長はそのままで，体重を男女で列を分ける．この例のように，男女が混ざって入力されている場合は，女の体重を右にずらす．

身長の表頭を空白にし，女男の体重の表頭に性別を入力する．

表示 4.17: 層別散布図

	A	B	C	D	E	F	G	H
1	性別	身長	体重		性別		F	M
2	F	1.50	43		F	1.50	43	
3	F	1.55	55		F	1.55	55	
4	F	1.40	33		F	1.40	33	
5	F	1.68	65		F	1.68	65	
6	F	1.32	29		F	1.32	29	
7	M	1.52	38		F	1.52	50	
8	M	1.55	58		F	1.55	48	
9	M	1.30	36		F	1.42	30	
10	F	1.52	50		F	1.55	36	
11	F	1.55	48		F	1.57	41	
12	F	1.42	30		F	1.65	64	
13	M	1.65	44		F	1.60	38	
14	M	1.60	47		F	1.57	38	
15	M	1.47	43		F	1.57	41	
16	M	1.50	36		F	1.63	50	
17	F	1.55	36		F	1.65	50	
18	F	1.57	41		F	1.52	52	
19	F	1.65	64		F	1.57	52	
20	F	1.60	38		M	1.52		38
21	F	1.57	38		M	1.55		58
22	M	1.60	42		M	1.30		36
23	M	1.63	45		M	1.65		44
24	M	1.65	54		M	1.60		47
25	M	1.63	41		M	1.47		43



表頭を含めて3つの列を選択して，散布図を描くと，列によってマークと色が変わったグラフが得られる．

表示4.17の右下の散布図は，マーク[7]と色を筆者の好みに合わせて修正したものである．

こうして得られた散布図に「近似曲線」を使って，男女別に回帰直線を当てはめることができる．

Excel関数による列の振り分け

並べ替えを行う代わりに，Excel関数を使って性別による列の振り分けを実行

[7] マークを ○，×，△，□，＊ とすると，何番目の層であるかが分かりやすい．

することができる.

元データの記録されているシートの空白列（ここでは O:Q **列）に散布図作成用のデータを生成する.**

表示 4.18: **列の振り分け**

	A	B	C	O	P	Q
1	**性別**	**身長**	**体重**		F	M
2	F	1.50	43	1.50	43	
3	F	1.55	55	1.55	55	
4	F	1.40	33	1.40	33	
5	F	1.68	65	1.68	65	
6	F	1.32	29	1.32	29	
7	M	1.52	38	1.52		38
8	M	1.55	58	1.55		58
...	...	...	...			

P1, Q1 **に** F, M **を入力する.** O2 **に** =B2 **を入力して, 下にコピーする.**

P2 **に** =IF($A2=P$1,$C2,"") **を入力する. この式は,「**2**行目の性別**(A2)**が表頭**(P1)** の**F **に等しいときは, 体重**(C2) **を, 等しくないときは空白**(””)**を表示する」ことを表わす.**

P2 **を** Q2 **にコピーし, さらに下にコピーする.**

O:Q **列から散布図を描くと表示**4.17**の右下の散布図に近いグラフが得られる. この関数で得られた空白は散布図では** 0 **が記録されていると見なされるので,** $y = 0$ **の位置に点が並ぶ. 縦軸の目盛りの範囲を修正すると表示**3.17**の右下の散布図が得られる.**

もし, 縦軸の目盛りの範囲に0 **を含めたいときは, 上の関数の最後の** ”” **を** +999 **のように, 目盛りの範囲を超える値を設定する. 表としては見にくくなるが, 希望の散布図が得られる.**

ここに述べた2**つの方法は,** 3 **つ以上の層に分ける場合に拡張できる.**

本日のまとめ

今日は，Excelの機能を活用して，応用的な散布図の描画方法を紹介した．われわれの課題の多くは，現象を分解して見ると解決の手がかりが得られることが多い．この章で学習した散布図の活用も，現象の分解方法であり，活用してほしいテクニックの一つである．散布図を使って，因果関係の可視化や，隠れた因果関係を層別により発見する喜びをぜひ，体験してほしいものである．

今日で第2単元を終了した．第1単元の第4章と第2単元は，データの特徴をとらえるための手法を取り上げた．手法は大きく次の2つに分けられる．データのグラフ化 と特徴を表わす値の計算である．この分野は 記述統計 と呼ばれる．

第3単元以降は，データの裏にある母集団ついて考える方法を学ぶ．それは推測統計 と呼ばれる分野である．

推測統計を理解し活用するためには，記述統計を十分にマスターしていなければならない．ということで，第2単元の演習問題などで，理解を再確認してほしい．

5 演習解答

5.1 第1章 質的データの記述

演習1 (p.14)

女性の人数は

```
=COUNTIF(B$3:B$412,2)
```

で求められる．データの範囲の行番号の 3 **と** 412 **の前に** $ **マークをつけておくと，男の人数は，この関数を別のセルにコピーしてから，最後の** 2 **を** 1 **に変えるだけで，求めることができる．**

インターネットの利用者を求めるためには，上の関数を別のセルにコピーし，列名の B **を** L **に置換え，最後の** 2 **を** 1 **に修正する．**

```
=COUNTIF(L$3:L$412,1)
```

同様に考えて，インターネットを主として利用する人数は，

```
=COUNTIF(P$3:P$412,7)
```

で求められる．

以上の計算表を表示5.1 **の上半分に示す．**

表示5.1: **度数の計算**

	T	U	V	W	X
2	質問	項目番号	項目	人数	
3	sex	2	女性	184	=COUNTIF(B$3:B$412,2)
4	Q2-7		インターネット	200	=COUNTIF(L$3:L$412,1)
5	Q3	7	インターネット	22	=COUNTIF(P$3:P$412,7)
6					
7	sex	1	男性	226	=COUNTIF(B$3:B$412,U7)
8		2	女性	184	=COUNTIF(B$3:B$412,U8)
9	Q3	7	インターネット	22	=COUNTIF(P$3:P$412,U9)
10		3	写真メール	10	=COUNTIF(P$3:P$412,U10)

表示5.1の下半分に示すように，COUNTIF 関数の最後のパラメータ（項目番号）に数字を書く代わりに，項目番号の列を準備し，対応するセルの名前を書く．

そうすると，項目番号を修正するだけで，対応する人数を簡単に求めることができる．

以上の方法で，各質問ごとに個別に度数表を求めることができるが，表示5.2に示すような表を準備すると，すべての質問に対する度数表が一度に求められ，便利である．

表示5.2: 各質問の回答毎の度数分布

	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR
1						通話	文章メール	写真メール	動画メール	写真撮影	動画撮影	インターネット	地図	ｹﾞｰﾑ	その他	主使用	利用料金	使いやすさ
2	No.	sex	age	area	job	Q2-1	Q2-2	Q2-3	Q2-4	Q2-5	Q2-6	Q2-7	Q2-8	Q2-9	Q2-10	Q3	Q4	Q9
3	1	226	101	9	17	398	361	168	24	187	45	200	24	90	21	171	222	54
4	2	184	103	22	12											201	147	264
5	3		103	175	47											10	38	59
6	4		103	47	67											2	3	26
7	5			81	24											2		7
8	6			32	39											1		
9	7			8	8											22		
10	8			36	66											0		
11	9				52											1		
12	10				50											0		
13	11				28													
14	合計	410	410	410	410											410	410	410

表示5.2の AB3 のセル（sex=1，226 となっている）に

```
=COUNTIF(B$3:B$412,$AA3)
```

を入力する．2番目のパラメータの AA は，No の列に対応する．この列は他の項目についての度数を数えるときにも共通に用いられる．コピーしても AA が変化しないようにするために $ が付けられている．

このセルを下にコピーすると自動的に2番目のパラメータの値が2となり，女

性の人数が求められる.

このセルを右にコピーすると，列名の B が自動的に C に変化するので，年齢階層別の人数が得られる.

このような操作によって，全てのセルに該当する人数を一度に求めることができる．対応する番号のないセルは後に消去する.

テキスト本文には,「調査データが得られたならば，データクリーニングをしなければならない」と書かれている.

データと集計のチェックのために，度数表の下に合計欄を求める．合計はすべて 410 となっており，問題がない．許されない番号が回答されている場合には，合計が小さくなる.

最初の列の No 欄には 1 から 11 までを取っている．11 は職業番号の最大値である.

このような集計表を作成し，注意深く観察することにより，調査データの質を向上することができる.

演習 2 (p.17)

年齢階層別と Q3（主たる用途）のクロス表と棒グラフを作成すると，表示 5.3 が得られる.

グラフの下には年齢階層が示されている．縦軸には，累積割合が取られている．表の行の並び順とは逆である.

年齢階層が上がると「通話」が増え「文章メール」が減少する傾向が明らかである． 3番目以降の項目については，度数が少なく，明らかな傾向は見られない.

COUNTIF 関数では，このような複数の条件を組合わせた件数を求めることができない．そこで，次のような工夫をする.

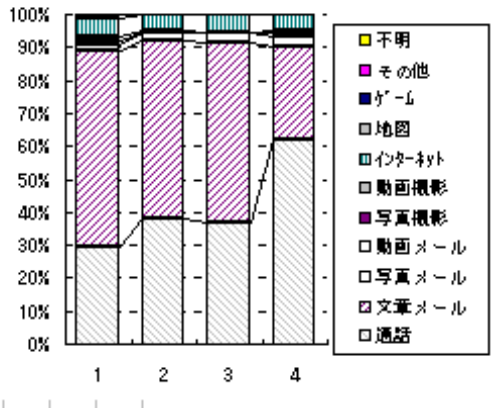
Z5 のセル（年齢階層=1，主用途=1(通話)）には,

```
=COUNT(IF(($P$3:$P$412=$Y5)*($C$3:$C$412=Z$4),1,""))
```

が入力されている.

表示5.3: 年齢階層別，主として使っている機能の度

	X	Y	Z	AA	AB	AC
3			年齢階層			
4			1	2	3	4
5	通話	1	30	39	38	64
6	文章メール	2	60	56	56	29
7	写真メール	3	2	2	3	3
8	動画メール	4	1	0	0	1
9	写真撮影	5	1	1	0	0
10	動画撮影	6	0	0	0	1
11	ｲﾝﾀｰﾈｯﾄ	7	6	5	6	5
12	地図	8	0	0	0	0
13	ｹﾞｰﾑ	9	1	0	0	0
14	その他	10	0	0	0	0
15	不明	11	0	0	0	0
16	合計		101	103	103	103



COUNT 関数の中に IF 関数を含めて，複雑な条件を設定する．

IF 関数は IF(条件，yes の場合，no の場合) の形式を取る．上の関数では，条件が

```
($P$3:$P$412=$Y5)*($C$3:$C$412=Z$4)
```

```
( 主使用     = 1 )*( 年齢階層   = 1 )
```

となっている．ここでは2つの条件がそれぞれ（）で囲まれ，＊記号で繋がっている．＊記号は，2つの条件が共に成立するとき yes，それ以外のとき no となる（and 条件 とも呼ばれる）．

主使用が1で，かつ，年齢階層が1であるとき，yes となり 1 となり，それ以外のとき no となり ””（空白）となる．

COUNT 関数は空白でないセルの数を数えるので，IF 条件で yes となった個数を数えてくれる．

皆さんもこの例にならって，別の組合わせについて解析を試みよ．

上に求めたようなクロス表は Excel の「ピボットテーブル」機能を使って比較的簡単に作成することもできる．

ピボットテーブルは，Excelによる集計の基本機能である．Excelがバージョンアップするときには，機能が強化され，それに伴って使い方が微妙に変化する．

ピボットテーブルの使い方は，市販されているExcelのテキストに詳しく解説されている．

以上の2つの理由で，ここでは説明を省略する．

演習 3 (p.27)

大学病院には一般病院では手におえない重症患者が送り込まれる．したがって，大学病院の手術対象には重症患者が多いと想像される．そこで，重症患者と一般患者に分けて度数表を作成して比較する必要がある．

例えば，次のような結果が得られたとする．

		手術数	成功数	失敗数	(%)
全体	大学病院	200	190	10	5%
	一般病院	100	97	3	3%
重症患者	大学病院	140	130	10	7%
	一般病院	20	18	2	10%
一般患者	大学病院	60	60	0	0%
	一般病院	80	79	1	1%

このように3元クロス表にまとめると，大学病院には重症患者が多いために全体としての失敗割合が多くなるが，重症と一般別に見ると，予想どおり大学病院のほうが失敗率の低いことが分かる．

演習 4 (p.29)

表示5.4の左上に元のクロス表と連関係数 Q が求められている．この計算表を下にコピーして，データの内容を，問題の指定に従って修正する．

Q を観察すると，連関係数には問題に示されている性質のあることが確認される．

表示5.4: 連関係数の計算

元の表

	yes	no	計	Q
yes	12	16	28	
no	30	94	124	
計	42	110	152	0.4

1行2行の入れ替え

	yes	no	計	Q
yes	30	94	124	
no	12	16	28	
計	42	110	152	-0.4

1行目を2倍

	yes	no	計	Q
yes	**24**	**32**	56	
no	30	94	124	
計	54	126	180	0.4

1列目を3倍

	yes	no	計	Q
yes	**36**	16	52	
no	**90**	94	184	
計	126	110	236	0.403

yes|yesを0

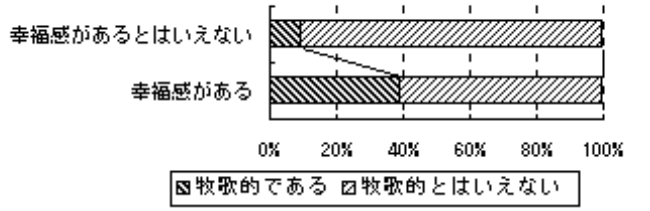
	yes	no	計	Q
yes	**0**	16	16	
no	30	94	124	
計	30	110	140	-1.000

演習5 (p.29)

前問と同様の計算表で，表頭・表側を書き換えて，連関係数を求めると 0.730 **が得られる.**

表示5.5: 連関係数の計算

	幸福感がある	あるとはいえない	計	Q
牧歌的である	42	17	59	
牧歌的とはいえない	67	174	241	
計	109	191	300	0.730



連関係数が 0.730 **とはどのくらいの関連度かを見るためにそのグラフを示す.**

連関係数の計算表の中の数値をいろいろと変えて，連関係数の値とグラフの変化を観察すると，連関係数の値を見てどのくらいの関係の深さかをとらえることができるであろう.

5.2 第2章 量的データの記述(1)

演習6 (p.54)

ヒントに従って，基本統計量と $\overline{x} - 2s$, $\overline{x} + 2s$ を計算すると，表示5.6の「全データ」の列の結果が得られる．とがりが 6.71 と極めて大きい.

データ表で,「条件つき書式」を使って, $\overline{x} - 2s = 178.7$, $\overline{x} + 2s = 226.9$ 外の値を赤の太字で表わすと（Excel ファイル参照）, 140, 177, 237 の3つが見つかる..

表示5.6: 基本統計量の計算

	全データ	削除後
n	100	98
平均	202.83	203.12
標準偏差	12.05	9.77
変動係数	5.94%	4.81%
ひずみ	-1.27	-0.17
とがり	6.71	-0.28
最小値	140	177
最大値	237	220
平均-2s	178.7	183.6
平均+2s	226.9	222.7

異常に外れている 140 と 237 を除く（数値の前に * をつける）と，削除後の列のように数値が変化する．ひずみ，とがりが 0 に近くなった.

5.3 第3章 量的データの記述(2)

演習7 (p.73)

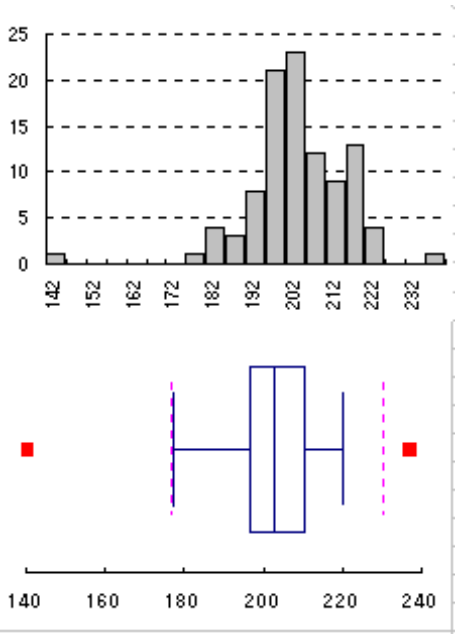
度数表の区間の範囲を表示5.7の「から」,「まで」の列に示すように，指定した．区間の幅は5である．代表値は区間の範囲の中央，例えば，140〜144 の範

囲の代表値は142 とする.

FREQUENCY 関数を用い,「まで」の列を2番目のパラメータとして計算すると，n の列の結果が得られる.

表示5.7: 度数表，ヒストグラム，箱ひげ図の作成

から	まで	代表値	n
140	144	142	1
145	149	147	0
150	154	152	0
155	159	157	0
160	164	162	0
165	169	167	0
170	174	172	0
175	179	177	1
180	184	182	4
185	189	187	3
190	194	192	8
195	199	197	21
200	204	202	23
205	209	207	12
210	214	212	9
215	219	217	13
220	224	222	4
225	229	227	0
230	234	232	0
235	239	237	1
			100



確認のために，n の合計が計算されている.

n から，右上のヒストグラムを作成する.

演習6で $\bar{x} \pm 2s$ の範囲外の3つの値のうち，140 と 237 の2つは他の値と極端に離れていることが分かる．もう一つの値 177 は，裾野に続いている.

本文では「級の個数は10〜15前後が良い」と書かれている．ここでは，級の個数が 20 であって，上の基準よりも多い．すべての最大値と最小値の級をヒストグラム上に示すために，このように級の個数を決めた．実務的には，極端に離れた値があるときは，ある値以上，以下という級を設定することが多い．この例では，170以下の級を作ると，級の個数が15個になる.

演習 8 (p.85)

マクロ「箱ひげ図」を使って，箱ひげ図を作成すると表示5.7の右下の箱ひげ図が得られる.

ヒストグラムで見られた2つの外れ値が箱ひげ図でも外れ値となった.

5.4　第4章　相関

演習 9 (p.107)

編集者は，原稿に目を通し，入力作業を，きれいな原稿は新人に，読みにくい原稿はベテランに割り当てた.

演習 10 (p.108)

2つの量的なデータをそれぞれ縦軸と横軸に取ってグラフにしたものを（1 散布図）という.

描かれた（1 散布図）を眺めると，2つの量的関係をみることができる.

縦軸に取った変量と横軸に取った変量の間に，1つの変量が増えるともう一方の変量が増加する，あるいは減少するという直線的な関係があるとき，2つの変量間には（2 相関関係）があるという.

（1 散布図）で打点が右上がりに集まっているとき,（3 正の相関）があるという.

逆に打点が右下がりに集まっているとき,（4 負の相関）があるという.

もし，このような関係が認められない場合には，その2つの変量は（5 無相関）であるという.

また，2つの変量の間に，一見（2 相関関係）があるように見えても論理的なつながりはなく，たまたま2つの変量の両方とそれぞれ（2 相関関係）がある第3の変量が存在することによって偶然に相関係数が高くなってしまう場合を（6 擬似相関）という.