2005. 7. 22

第１単元

統計とは何か

目 次

第１単元

統計とは何か

単元のねらい　昨今は，ちまたにデータがあふれている．しかし，データは単に数字の束に過ぎない．データから，意味のある情報を取り出すには，統計解析の力が必要である．

情報とは，なんであろうか．広辞苑によると「情」とは”物事に感じておこる心の動き”とある．とすると「情報」とは”心にひびく報せ”ということになる．データを集約して見る人の心をゆさぶるようにする技術，それが統計的方法である．

本講座による学習の目的は，統計的方法の基本の考え方と技術を習得することである．

本単元の §1 〜 §2 は，本講座全般の序説であって，§3 はそれ以降に展開される方法・技術の基礎をなす考え方をここで習得する．また，§4 では時間順に取られたデータ（時系列データ）に対する統計的方法を学習する．

このテキストでは，単に 統計 というときには，統計的に処理された数字 を意味することとする．

1 統計の利用

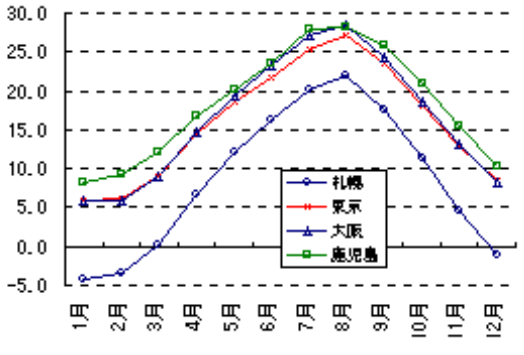
1.1 社会・国民生活と統計

(1) 例1 気温の統計

1月の東京の気温は，平年に比べて1℃低かったなどということが，テレビやラジオでよくいわれるが，おそらくあなたは，比較の基準とされている平年の気温とはなんであろうかと疑問を感じることだろう．同一地点においても気温はたえず変動するものであるから，これを一つの数値で代表させるためには平均値を求めなければならない．そのために各気象測候所では，1日に24回1時間おきに観測値を求め，その平均を出してその日の平均気温とする．それをさらに1ヵ月にわたって，平均したものをもってそれぞれの月の平均気温としている．このようにして出した各月の平均気温を30年間にわたって平均したものが，それぞれの月の平均値である．表示1.1は，札幌，東京，大阪，鹿児島における毎月の気温平年値（1971年から2000年までの30年間の平均）を示している（出典： 総務省「日本の統計」）．

表示1.1: 各地の気温平年値（℃）

	札幌	東京	大阪	鹿児島
1月	−4.1	5.8	5.8	8.3
2月	−3.5	6.1	5.9	9.3
3月	0.1	8.9	9.0	12.1
4月	6.7	14.4	14.8	16.8
5月	12.1	18.7	19.4	20.2
6月	16.3	21.8	23.2	23.6
7月	20.2	25.4	27.2	27.9
8月	22.0	27.1	28.4	28.2
9月	17.6	23.5	24.4	25.8
10月	11.3	18.2	18.7	20.8
11月	4.6	13.0	13.2	15.6
12月	−1.0	8.4	8.3	10.4

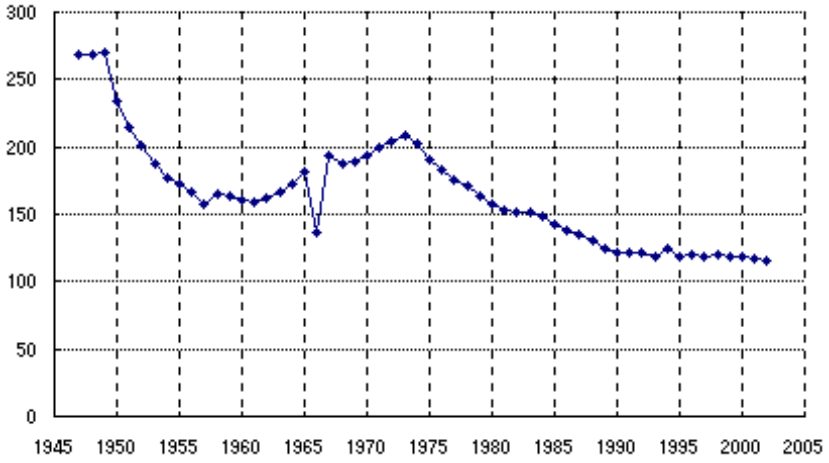


表示1.1の右のようにグラフ化することにより，大阪の気温は，冬季は東京とほぼ同じであるが，夏季は鹿児島に近いというようなことが分かるであろう．

(2)　例２　出生数と出生率の推移

表示1.2は、1947年から2002年までの各年に日本で出生した人数を万人単位で表わしたグラフである（出典：厚生労働省「人口動態統計」）．

表示 1.2: 日本の出生数（万人）

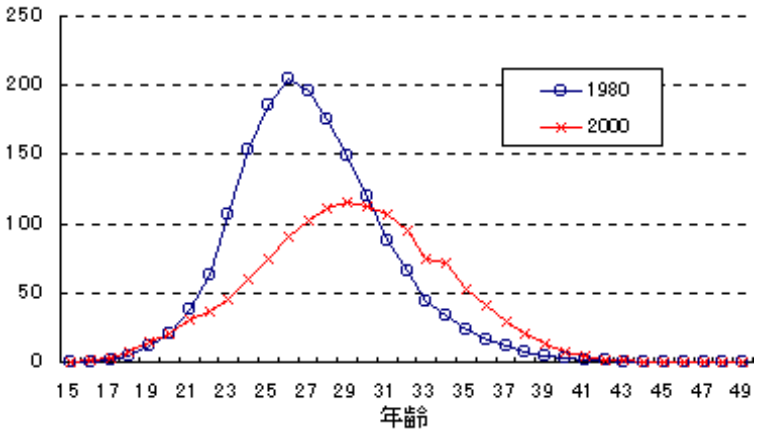


1947年から1949年は，戦争直後における結婚の増加により第1次ベビーブームが起こり，出生数は毎年260万人台と多かったが，1950年頃から出生数は急速に減少し，1960年前後には160万人台になっている．その後は，1966年の丙午（ヒノエウマ）の年を除くと緩やかな上昇傾向にあった．1966年の出生数は，136万人に落ち込んでいるが，その前後の年は出生数が多い．これは，古来日本では，女子が丙午の年に生まれることを嫌う風習があるために，出産を調整した人々が多かったことをうかがわせる．その後，1971〜74年は「第2次ベビーブーム」と呼ばれるが，第1次ベビーブームで生れた人たちが出産適齢期に入り，出生数が増加している．ピークは1973年生れの209万人である．その後出生数は減少していき，2002年には115万人となっている．

ところで，近年の出生数の減少の背景には，女性の出産年齢の変化があることが分かっている．表示1.3は，母親の年齢別にみた女性人口千人当りの出生率である．

表示1.3: 母親の年齢別にみた出生率

（出典：厚生労働省「人口動態統計」）



1980年では，25歳〜27歳が高く，出生率は約200である．その後急速に出生率が低下し，とがった釣り鐘型になっている．2000年では29歳が 112 で最も高く，ピークの年齢が高齢化しているとともに，20歳代の出生率が大幅に低下している．形は，なだらかな山型となり30歳代以上の出生率が，1980年よりもむしろ高いことがわかる．

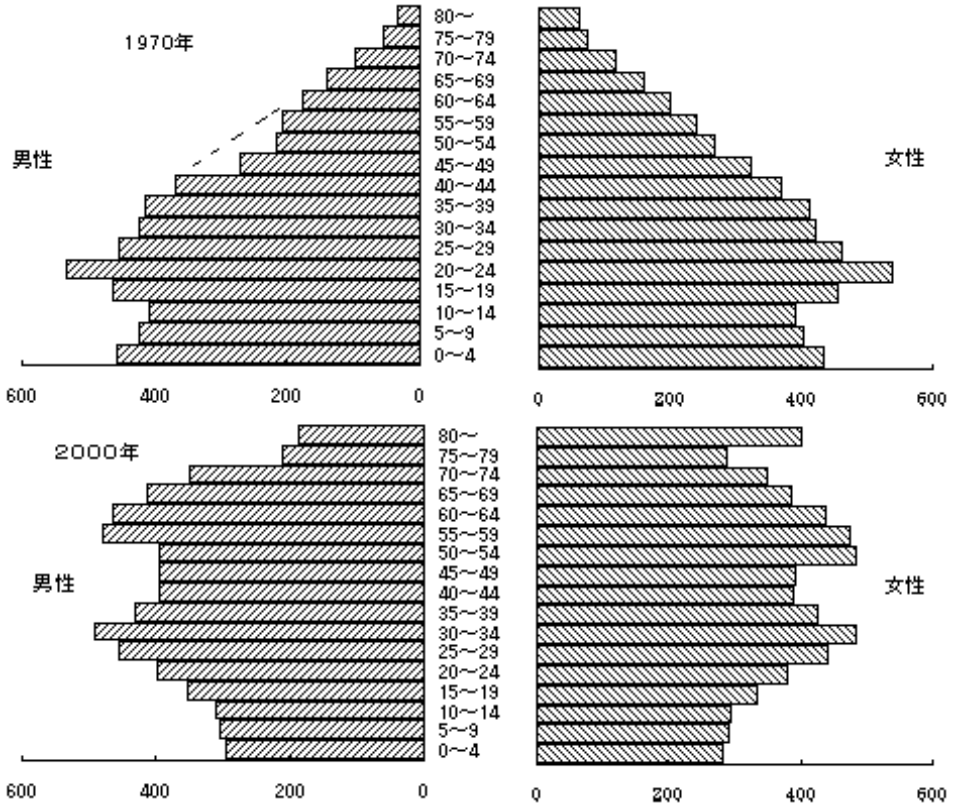
なお，これらの統計は，市町村に届け出られた「出生届」に基づいてまとめられている．

■　演習１　厚生労働省の「人口動態統計」には，死亡，婚姻，離婚 などのデータも含まれている．インターネットなどで調べて，変化の様子を眺め，データの裏にある社会構造などについて考察せよ．■

(3)　例３　人口の年齢別構成

表示1.4は，日本の人口を，男女別，年齢別に分け，男性を左に，女性を右に，年齢を5歳きざみに上にとって，人口の年齢構成をグラフ化したものである（出典： 総務省「国勢調査」）.

表示1.4: 日本の人口の年齢別構成（1970年，2000年）



このようなグラフは 人口ピラミッド と呼ばれることがある.

上は1970年，下は2000年の人口ピラミッドである.

1970年のグラフを見るといろいろな特徴が見られる.

35歳から79歳まで女性はほぼ直線的に変化している．それに対して，45歳から59歳の男性はその上下の変化よりも凹んでいる．これらの人の終戦時の年齢

は，20歳から34歳である．これは，戦争に従軍し多数の人が亡くなったことによる．

20歳から24歳は男女とも飛び出している．これらの人の出生は1946年から1950年であり，表示1.2 の第1次ベビーブームに生まれた人である．

その下の年齢層は，ベビーブームが過ぎて，出生数が減少したため，凹んでいる．

2000年のグラフを1970年のグラフと対応させて見よう．

2000年のグラフは，1970年のグラフを30年だけ上に移動したものに対応している．ただし，老化などによる死亡により対応する年齢層の人口は少なくなっている．

第2次ベビーブームのあと，出生が急速に減少したため，年少人口の減少が顕著に見られる．

最近日本の年金制度の将来が大きな問題となっている．将来の年金制度を考えるとき，例えば2030年の人口分布を予測することが要求される．この予測は極めて困難であるが，表示1.2 の出生数や，表示1.3 の女性のライフスタイルの変化などを考慮して，少しでも良い予測をすることに努力しなければならない．

■ 演習2　ここには，1970年と2000年の5歳きざみの人口ピラミッドを示した．1歳きざみの人口ピラミッドでは，ひのえうま の影響も見られるであろう．さらに遡って，1950年，1920年 などの人口ピラミッドを探して，調べてみよ．

また，国による年齢分布の違いを調べてみるのも興味深いであろう．　■

本日のまとめ

受講生の皆さんの身近なテーマを取り上げて，統計がどのように役立つかを見てもらった．

表示1.1 に皆さんの住んでいる地域の気温を追加プロットしてみると面白い

であろう.

例2, 3 は最近大きく取り上げられている 年金問題 の裏にあるものが，とてつもなく大きいことが分かるであろう.

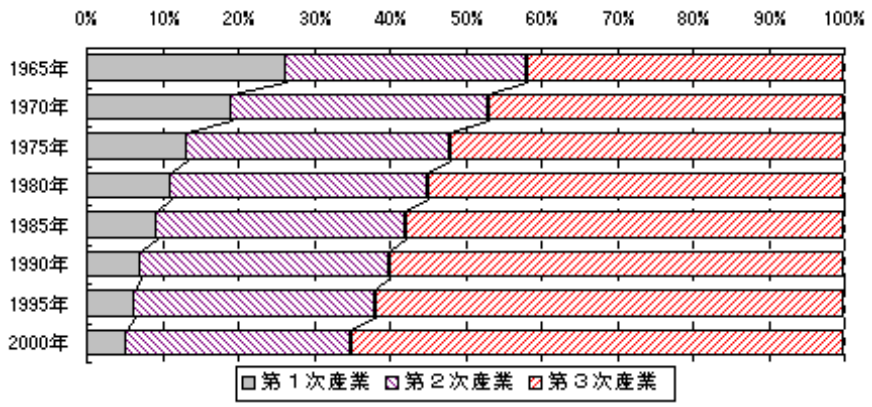
明日は，もう少し視野を広げて，日本の経済変化を見る.

1.2　社会構造・経済動向と統計

(1)　例４　就業構造の変化

ある国の就業構造は，その国の就業者が属する産業別の構成比率（%）で表わされる．産業の種類を，第1次産業，第2次産業，第3次産業[1]の3区分に大別して，日本の現在の就業構造を過去のものと比較したものが表示1.5である（出典: 総務省「就業構造基本調査」）．この図から，1965年から2000年に至る35年の間にわが国の就業構造が著しく変化してきたことが読みとれるであろう.

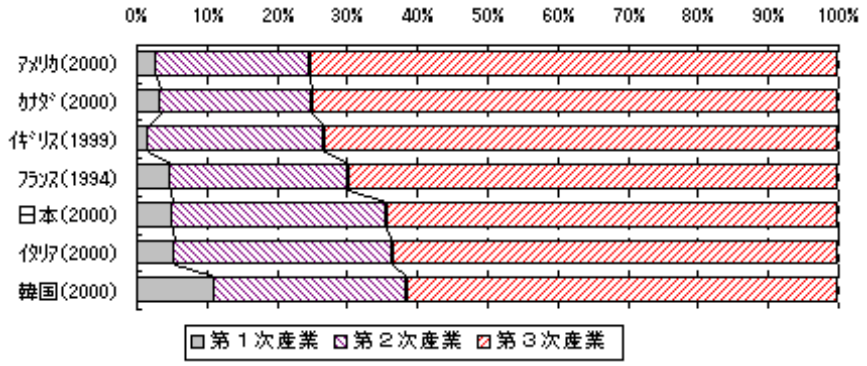
表示1.5: 日本の就業構造の推移



[1] 第1次産業とは，農林水産業のことで，第2次産業とは，鉱業，製造業及び建設業をあわせたものである．第3次産業はそれ以外のもの，すなわち卸売・小売業，運輸・通信業，金融業などの民間サービスと官公サービスとからなる.

まず，第1次産業の割合は急速に減少し，それに比較して，第3次産業の割合が増大している．第3次産業の割合は，1965年以降，年々増大している．このような変化動向は，この期間におけるわが国の第3次産業化の過程をよく反映している．なお，第2次産業の割合がほとんど変化しないのは興味深い．

表示1.6: 就業構造の国際比較



表示1.6は，就業構造の国際比較を示している．

これによると，就業構造の面では日本とイタリアが類似している．イギリス，アメリカ，カナダでは，第1次産業の割合が日本やフランスに比べて低く，韓国では，第1次産業の割合が各国に比べて高くなっている．

(2)　例5　各種の経済指標

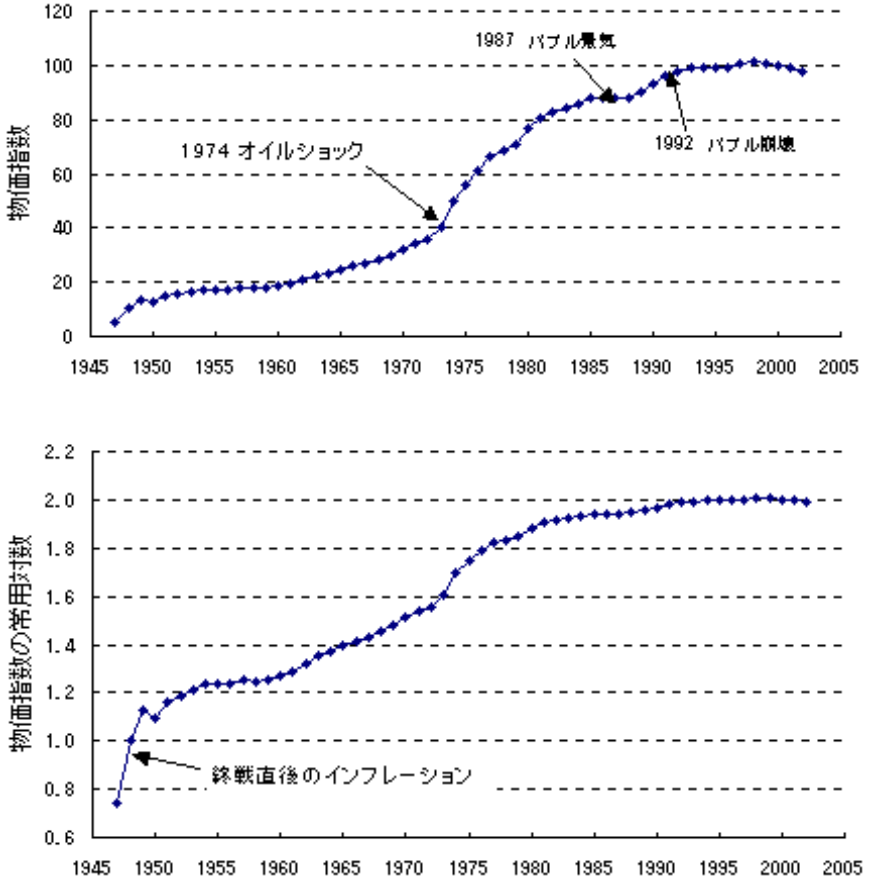
人間の健康状態を診断する基礎として，体温，脈拍，血圧などの測定が必要であるのと同様に，一国の経済動向を診断するためには，国の経済のいろいろな面を表現する経済統計のデータが必要である．それらを一般に 経済指標 と呼んでいる．国民総生産や国民所得などがその例である．その他にも，鉱工業及び農業の生産指数，卸売及び消費者物価指数，雇用指数，賃金指数，鉱工業の出荷指数，製品在庫指数，原材料消費・在庫指数など，きわめて多種類の経済指標が政府の手で定期的（毎月，毎四半期，毎年）に集められている．これらは，いわば経済動向のバロメーターである．このバロメーターによってはじ

めて，国の経済の現実の姿を客観的にとらえることができる．これらの経済指標にもとづいて，政府は適切な経済政策を立てることが可能になるし，民間企業は，一般経済動向の見通しのうえに独自の方針を決めることが可能となる．

例として，戦後から最近までの消費者物価指数の変化を見よう[2]．

横軸に年を，縦軸に2000年を 100 とした物価指数を取ったグラフを表示1.7の上に示す．

表示1.7: 消費者物価指数の変化



1974年のオイルショック後急激に物価が上昇していることがわかる．1985年

[2] 物価指数がどのようにして求められたかは§4.3(3) で詳しく説明する．

頃にはオイルショックの影響が消え，物価が安定したが，1987年 バブル景気が始まると，再度物価の急上昇が見られる．1992年 バブルが崩壊すると，物価は安定し，その後デフレ傾向が生じている．

グラフの左は終戦直後のインフレーションの様子を表わしている．このグラフからは，それほど強烈なインフレーションのようには見えない．

物価指数が 10 から 20 になったとすると物価が2倍になったことを示す．これは，50 から 100 への変化と同じである．このように比率の変化を見たいときには，対数を取るのが良い[3]．

表示1.7 の下は縦軸に物価指数の常用対数を取ったものである．これを見ると，終戦直後の傾斜が大きくなり，インフレーションのすごさが読み取れる．

本日のまとめ

表示1.5 から，日本の就業構造が最近40年間に大きく変化したことが分かり，表示1.6 で日本を世界各国と比較することができた．これが，地方と都市の格差を生じた大きな要因である．

また，表示1.7で戦後60年の物価の動きを大局的に眺めることができた．終戦直後のインフレーションが想像を絶する激しいものであったことは，下の対数のグラフで知ることができる．

なお，対数変換は，データを見る上で極めて有効なものである．対数の予備知識の少ない人は §1.6 の補足で勉強してほしい．

---

[3] **対数の基礎について，§1.6 補足 で説明する．**

1.3　企業経営・技術と統計

(1)　例6　市場調査

メーカーがどんなに優秀な製品を製造しても，消費者がそれを買ってくれなければ企業は成り立たない．保険会社やホテルなどのようにサービスの提供を仕事とする企業の場合も同様である．したがって，企業が成り立っていくためには，市場の実状に関していろいろな知識や情報を収集し，消費者が買いたいと思う製品やサービスを提供し続けなければならない．必要とする情報として，例えば，

(a)　その製品やサービスに対する需要の総量，つまり市場の大きさやその地域分布，またそれらの性別・年齢別など属性による違い

(b)　同種の製品やサービスを提供している他企業との競争上の関係

(c)　消費者の消費習慣や購買習慣

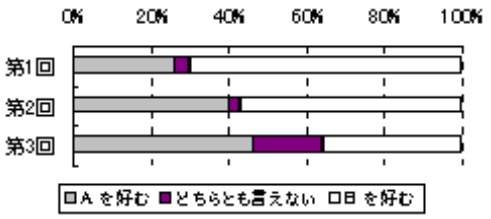
などがあげられる．これらの情報は，日常の企業活動の記録として得られるデータ（例えば，販売数量や来客数など）を分析したり，政府の発表する統計資料を分析したりすることによって得られるものもあるが，それだけでは足りないことが多い．そこで，足りない情報を手に入れるために，消費者や小売店などを対象として，企業が直接あるいは民間の調査専門機関や広告代理店調査部門に委託して，市場調査を実施する必要が生じる．

以下に簡単な例を示そう．表示1.8は，インスタントコーヒーについて，消費者の嗜好を調べるために行われた3回のテスト結果を表わしたものである．

A銘柄のインスタントコーヒーのメーカーは，自社のコーヒーが従来から競争相手であるB銘柄に比べて消費者に受け入れられていないことを，市場調査やその他の情報から知っていた．そこで，その原因がコーヒーの質そのものにあるのかどうかを確かめるために，第1回のテストを実施した．その方法は，市場を構成するある重要都市のインスタントコーヒー使用世帯のうちから300世帯を無作為に選び出し，それらの世帯のそれぞれに，A銘柄コーヒーとB銘柄コーヒーを入れた同形・同大の2つのビンを配布した．ビンには異なる数字のラベルだけが付いていて，消費者には外見による銘柄の識別はできないが，調査

表示1.8: インスタントコーヒー嗜好テスト

	A を好む	どちらとも言えない	B を好む
第1回	26	4	70
第2回	40	3	57
第3回	46	18	36



企画者にはできるようになっていた．これら2つのビンを配布された世帯は，2週間にわたって両コーヒーをしばしば試飲するように依頼され，その後で再訪問した調査員に対して，どちらのビンのコーヒーを好むかの報告を求められた．表示1.8の第1行は，このテストで各々の銘柄を好むと報告した世帯と両者を同程度に好むと報告した世帯のパーセントを示している．

これを見ると，コーヒーの質そのものとしてA銘柄はB銘柄に比べて好まれていないことが明らかである．

そこで，A銘柄コーヒーに改良を加えて試験的に製造されたコーヒーを用いて，6ヵ月後に再びB銘柄との比較のための同様のテストを行った結果が，第2行にあげてある．結果は第1回に比べれば良くなっているが，十分に満足すべきものではなかったので，さらに改良を加えた試験品をもって第3回のテストを実施して，同表の第3行の結果を得た．この最後の結果に満足して，その試験品と同質の製品をA銘柄コーヒーとして市場に出すことを決定したのである．

なお，調査結果の集計やグラフ化は第2単元で，解析(検定・推定)の手法は第3, 4単元で学ぶ[4]．

(2)　例7　品質管理

製品やサービスを提供している企業は，その製品やサービスの品質水準が一定に保たれるように努力している．そのため，製品を製造する際の操業条件やサービスを提供する際の手順などがマニュアルに事細かに決められていること

[4] 市場調査の設計については「現代統計実務講座」第7単元参照．

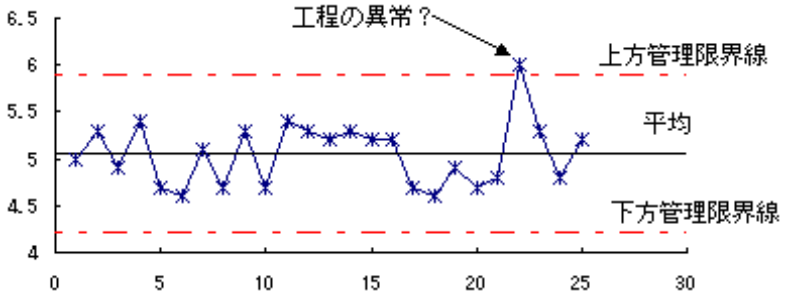
が多い．しかし，そのマニュアル通りに進めても，まったく同一の製品やサービスが提供できるわけではない．また，完全にマニュアルに決められた通りに進められるかというと，やはりそこにもバラツキが生じ，それが最終的に製品やサービスのバラツキを生み出すことにもなる．さらには，予期していなかった突発的な原因によって，製品やサービスがばらつくこともある．つまり，企業が提供する製品やサービスにはバラツキが存在するのが普通であり，そこにはさまざまな原因が存在する．

しかし，例えば，カレー用のスプーンを買いに店に行ったとき，陳列されている同じ種類のスプーンにバラツキを見出すことはまずない．製品の場合，何らかの規格(長さが○○〜○○cm，重さが△△〜△△g など)が決められているのが普通である．規格外のスプーンが製造段階で発見されれば出荷されることはないし，輸送中に破損や変形などしたものは陳列段階で発見され除かれる．このように品質水準を一定に保つためのあらゆる場面での活動が品質管理活動である．日本の企業は，戦後の「安かろう，悪かろう」の時代から品質管理活動を積極的に展開し，現在のように品質面で国際的に評価される製品を製造できるようになった．

製品の製造工程に異常が発生していないかを判断するために現場からデータを取り，それを時間順に並べた折れ線グラフを作る．このようなグラフを品質管理の分野では 管理図 と呼んでいる，管理図は，適用する対象に応じていろいろあるが，ここでは，平均値を用いる例を取り上げ，表示1.9に例を示す．

この図上の多数の点は，スプーン製造工程において連続的に製造されているスプーンを一定時間ごとに無作為に5個取り出し，その重さの平均値を求めてプロットしたものである．横軸の目盛りは測定値の番号を表わし，縦軸の目盛りは重さを表わす．3本の水平線は，過去数週間にわたる同様の測定に基づいて求められたもので，中央の水平線は全体としての平均値に相当する．また，上下2本の水平線はそれぞれ上方管理限界線，下方管理限界線と呼ばれるもので，この範囲内であれば自然のバラツキとして許容できる範囲を示している．1個1個製造されているスプーンの重さはさまざまな原因によってばらついている．し

表示1.9: 管理図



たがって，5個のスプーンから求められる重さの平均値もばらつくことになる．製造工程に異常が発生していなくても平均値はばらつくので，そのバラツキが自然のバラツキとして許容できるのか，あるいは異常が発生していなければ起きないほど，すなわち許容できないほどばらついているのかを判断する必要がある．その判断基準が上下の管理限界線であり[5]，ある時刻の一つの点がこの範囲外に出たならば，製造工程に異常が発生したと判断することになる．

(3)　例８　実験計画

一定の制御された条件のもとで得られる統計を通じて，現象を観察し，それから必要な結論を引き出せるように計画された実験，すなわち統計的実験が科学・技術分野で広く用いられている．その簡単な例としては次のようなものがある．

(a)　新しく作られたかぜ薬Aが，従来からあるかぜ薬Bよりもよく効くかどうかを確かめるために，かぜを引いている者200人を選んでその半数にはAを，残りの半数にはBを与え，この両グループの一定時間後における治癒率を比べて結論を出す．

(b)　$1200m^2$ の田んぼを $100m^2$ ずつ12区画に分け，窒素肥料を 1.0kg, 1.2kg, 1.4kg ずつ施した区画を4区画ずつ用意した．それぞれの区画で稲

[5] 管理限界線の求め方など詳細は「現代統計実務講座」第８単元参照．

を栽培し，窒素肥料の使用量と収穫高との関係を調べる．

このような実験はすべて，研究の対象となっている現象（上の例ではそれぞれ，(a) かぜの治癒率，(b) 稲の収穫高）に対して影響を与えている いろいろな原因のうちから，特定のものを実験因子（上の例ではそれぞれ，(a) かぜ薬の種類，(b) 窒素肥料の使用量）として取り上げ，現象に対するそれらの影響・効果がどうであるかを調べることを目的として計画されている．

上の例では，実験因子を一つだけ取り上げているが，それ以外の条件はできるだけ均一にするように実験を仕組まなければならない．そうしなければ，結果として得られた現象が異なったとしても，それが実験因子による影響なのか，均一にできなかった条件の影響なのか，区別がつかなくなってしまう．

例えば，上の (b) で，稲の収穫高に対する窒素肥料の使用量の影響・効果を調べたい場合，窒素肥料の使用量を変化させることは当然であるが，田んぼの土や水の条件など他の条件は一定に保たなければいけない．そうしなければ，稲の収穫高に差があったとしても，それが窒素肥料の使用量によるものなのか，田んぼの土や水の条件が異なることによるものなのか区別がつかなくなってしまう．反対に，田んぼの土や水の条件が稲の収穫高に影響を与えることが分かっているならば，それらも実験因子として取り上げ，つまり 2 種類以上の実験因子を同時に取り上げて，それらの影響を別々に求められるように仕組むことが望ましい．

このような実験の組み方と，その結果の正しいまとめ方を取り扱う統計的方法を　実験計画法 と呼ぶ[6]．

本日のまとめ

今日は，企業の中で統計がいかに重要な役割を果たしているかを，いくつかの例で示した．個々の例について詳しく説明することは出来なかったが，皆さんの周囲を注意深く見回してほしい．こんなところに! という発見があるであろう．

[6] 実験計画法の詳細は「現代統計実務講座」第 8 単元参照．

明日は，自分が主体となって，身近の問題を取り上げ，データを取ることを考える．

1.4　データを取って見よう

(1)　体験することの意味

データには，あらかじめ数字としてまとめられているものと，自分の手ではじめて作成されるものがある．前者の例として，企業には，製造工程で自動計測されるデータや，仕入れや在庫などの物流データ，販売場面で収集されるPOSデータをはじめとした業務統計がある．また，官公庁や調査機関によって収集された統計資料やデータベースがある．これらのデータは，第3者が作成したいわば既製品といえる．昨今はインターネットなどを使って，既製品のデータを自分のパソコンにダウンロードすることもたやすくできる．

しかし，データを解析する技術を身につけるためには，自分で「計画を立てて」,「データを取る」ところから経験することが是非とも必要である．自分の手で調査や観測，実験などを行うと，データ取得が決して簡単なことではないことが分かるはずである．どのような計画を立てるのか，その重要性も分かるであろう．

また，データの背後にある事実を考察する訓練になり，どのような統計解析をおこなえば良いのか考えることになる．さらに，この体験は，既製品のデータの質を考えるきっかけになるであろう．

以下に身近なところからデータ収集をする事例をあげるので，データの取得を体験してほしい．どのようなデータを取るかは，この例にこだわらず，自由に，あまり気張らずに考えてほしい．

(2)　事例1：健康の自己管理

健康を自己管理するために，血圧，体温，体重などを毎日測定することを考える．

家庭用の血圧計を持っている人は，自分（または家族）の血圧を毎日同じ時刻に，1日に数回測定してみる．血圧には「収縮期（上）」「拡張期（下）」があるが，とりあえず両方を測定する．血圧計のない人は，体重，体温，脈拍 でもかまわない．

測定する時刻によって，これらの値は変化する．毎日の生活，体調には全く変化を感じなくても，血圧の値はかなり変化する．また，続けて複数回測定を繰り返すと，± 10 くらいの差は見られる．

その小さな上下で一喜一憂することはない．それはなぜであろうか．§1.3 例7であげた管理図の考え方を思い出してほしい．異常であると判断するのは，ある一定のバラツキを超えたときであり，このような判断を私たちは日常生活の中で自然に行っていることである．

それでは，健康管理のためには，どのような時点で（例えば，起床直後，食事前，食事直後，食後何時間後，など）測定したら良いであろうか．

また，健康管理の指標としては，何を使ったら良いであろうか．上の血圧，下の血圧を個別に見るのではなく，両者の差の方が良いかもしれない．体重であれば，就寝前の体重と起床直後の体重の差が良いかもしれない．就寝中にトイレに行かなければ，この差は，基礎代謝によるカロリーの消費量，発汗量によるものである．体調が悪く寝汗をかいたときは差が大きくなるであろう．

いろいろな可能性を考えて，しばらく試行を続け，その結果を見て適切な観測時点，管理指標を決めるのが良いであろう．

健康管理のためには，管理指標の測定をするだけでなく，それに影響すると思われる項目，例えば，曜日（出勤日，休日），就寝時刻，飲酒の有無（または飲酒量），など，日々の生活活動についても記録しておくべきであろう．

管理指標について，管理図を作成する．グラフの書き方は横軸に時間を，縦

軸に観測データをとり，折れ線グラフで示す．グラフの書き方は第2単元で詳しく学ぶので，ここでは，自己流で構わない．

管理指標に変化が見られたら，生活活動と関係がないか調べることにより，健康を保つためのヒントが得られるかもしれない．

このような観測を続け，結果をグラフに記入することにより，自分の健康状態に関心をもつようになるという副産物がある．工場における管理図にも，現場の作業者が製造工程や製品品質に関心を持ち，注意深く観察するようになるという効果が期待されている．

(3) 事例2：道路の混雑度

通勤にバスを利用している人は，片道のバスの所要時間を測定してみよう．そのためには，所要時間の定義（例えば，開始時点を自分がバスに乗ったときとするか，バスが動き出したときとするか）きちんと決める必要がある．

道路の混雑状況によって所要時間は変動するから，渋滞状況を左右しているであろう要因（天候，時刻，曜日など）についても記録を忘れないようにする．

道路の渋滞の大きな原因は違法駐車にある．毎日観察できる 駐車禁止道路 があれば，違法駐車の台数の変化を観測するのも面白いであろう．例えば，会社の窓から毎日定時に前の道路の写真を取り，駐車台数を数えるという方法が考えられる．

警察による違反摘発が実施されると，道路面にチョークの跡が残る．チョークの跡が残っている日と残っていない日で，違法駐車台数に違いがあるであろうか？ 調査に先立って，結果を予想することも大切である．

チョークの跡を見ると,「これは危ない」と考える人が多ければ，跡のある日の台数は少なくなるであろう．また,「昨日摘発が実施されたから，今日は大丈夫だろう」と考える人が多ければ，多くなるであろう．

このように，いろいろな考え方を想像して，複数の仮説を設定しておき，データによってどの仮説がより真実に近いかを確かめるという姿勢が必要である．

参考 寺田寅彦随筆集第二巻（岩波文庫）に「電車の混雑度」という表題の

随筆がある．これは，大正11年に書かれたものである．当時は，東京都ではなく東京市で，町には市電が走っていた．

彼は，満員電車でつらい思いをしないためにはどうしたら良いかを考えた．朝の込む時間帯（7:55〜8:40）に，神保町を通る市電の通過時刻と混雑度を観測して表にまとめた．5分間に通過する台数は0〜5台と変化し，混んだ電車の後には空いた電車の来る傾向のあることを見出した[7]．

彼は物理学者の目で，このような現象が生じる理由を追求している．12ページの短編ではあるが，考えさせられる内容がたくさん含まれている．

(4)　データを取るときの注意点

いくつかの例をあげて，データを取るときの注意点について説明した．それらを整理してみよう．

目的を明確にする．何となくデータを取るのではなく，データから何を知りたいのかを十分に考える．

健康管理であれば，自分の健康状況を知り，健康を維持する．

バスの所要時間であれば，所要時間に影響する要因を把握し，その日の状況から所要時間を予想して，自宅を出る時刻を決め，遅刻のないようにしたい．

目的を果たすためには，なにを，いつ，どこで，どのように観測するかを検討する．

これらの項目を決めるためには，データに影響する要因について十分な知識が必要である．健康管理であれば，家庭医学書を開いて収縮期と拡張期の血圧の意味を勉強しなければならない．

得られた知識を整理して，問題の本質を把握するための方法を §2.4，2.5 で取り上げる．

予備知識が不十分であれば，まず，予備調査を実施し，その結果をよく観察して，本調査の計画立案のための知識を吸収する．

---

[7] 観測値とデータがテキストに添付されているExcelファイルに記録されている（名前：寺田寅彦）．

データが得られたならば，グラフに表わして，観察する．

ここには，2 つの事例を取り上げて，考え方を説明したが，受講生の皆さんがやってみるのはこの2つに限らない．皆さんの興味があり，そのバックグラウンドについて知識のある問題を自由に選んで，上にあげた諸点を考慮して，データを取ってほしい．

例えば，次のような問題が考えられる．

- 受信メールの件数，迷惑メールの件数
- 自家用車の走行距離
- 新聞の折り込み広告の枚数
- 家計管理（例えば，毎月の電気・ガス・水使用量）

また，日常業務の中で，データを取り，その結果を活用する立場にある受講生は，上にあげたような点を考慮して，これまでに取ったデータや結果の判断が適切であったかどうかを検討することが，今後の業務の質の向上に貢献するであろう．

本日のまとめ

今日の学習の目的は，テキストを読んだだけでは達成できない．自分のテーマを決め，明日以降 時間をかけてデータを収集してほしい．第 1 単元の §4，第 2 単元でデータのグラフ化を学ぶとき，自分の集めたデータをグラフ化することは，実際問題を解決するための貴重な経験を与えるであろう．

この節で述べた課題に真剣に取り組むことにより得られる知見は，今後の講座の内容の理解に大いに役立つはずである．

1.5　統計の利用とその目的，有用性

(1)　現代生活と統計

19 世紀の英国の有名な物理学者ケルビンは，物事を数値で表現することの大切さについて次のようにいっている．

「あなた方が論じている事柄に関して，みずから測定を行い，結果を数値として表わすことができれば，その事柄についてなんらかの知識を獲得したことになる．しかし，もしそれに関して測定を行うことができず，結果を数値で表わすことができないならば，あなた方の知識は貧弱で，そして不十分なままでいることになろう」．

例えば，自然現象の場合でも，社会現象の場合でも，それを数値で表現する「数量化」は，われわれが物事を客観的に見て，なお具体的に理解するための重要な手段である．

§1.1 からいろいろな例で見てきた．これらの諸例の中で，もし仮に統計がないとすれば，それぞれの問題に対する問題点の把握も，対応もみじめなものになることは，容易に想像される．

(2)　意思決定の指針

§1.1，§1.2 であげた事例は，政府や官公庁がまとめている統計の例である．これらは，統計を通して，起こっている現象を客観的，具体的に把握し，政策立案の基礎とするものである．これらの統計は具体的な問題解決の意図をもって作成されたものではないが，基礎的な資料として幅広く活用される．また，過去から現在にいたるまでの統計が蓄積されているので，その変化を知り，将来起こりうる問題を予測する資料ともなる．

例えば，地球温暖化の問題は，地球上の各地の気温観測によって察知された．私たちの生活が知らず知らずのうちに犯している環境破壊に対して，警鐘をならすことになった．また，日本の年金財源が将来破綻するかもしれないという問題は，出生数と死亡率などの統計をもとに将来の人口推計が行われ，その結果明らかになったことである．これらの統計は，地球温暖化を監視したり，年

金財源の問題を検討する目的で取られてきたものではない．しかし，統計によって，今起こっている現象を浮かび上がらせることになったのである．

§1.3 では，企業が活用している統計の事例をあげた．これらの統計を利用する目的は直接的なものであって，その結果をもとに取るべき行動が決められていた．

市場調査＜例6＞のインスタントコーヒーの嗜好テストでは，消費者の嗜好を調べるために3回の製品テストが行われた．テストの結果を踏まえて製品の改良を続け，3回のテストで自社銘柄コーヒーが競争銘柄コーヒーの評価を上回ったことを確認し，市場に出すことが決定された．

品質管理＜例7＞では，管理図によってスプーンの製造工程が正常であるかどうかを常時監視している．この方法により，万一異常が現われた場合はすばやく察知し，その工程を止めて原因を探し，修復するという対策が取られる．

実験計画＜例8＞のかぜ薬の例は，従来のかぜ薬よりも薬効の高いことを確かめた新薬を市場に出すために用いられた．このような実験で薬効が向上し，副作用の少ない新薬が開発され，その実験結果を添付して厚生労働省に申請して，審査された後に初めて患者に投与される．

また，窒素肥料の使用量と稲の収穫量の関係を見る実験では，最適な肥料使用量が農家に推奨されることになる．

これらの例では，調査や実験がどのような結果を示せばどのような意思決定をするのか，調査や実験を計画する段階であらかじめ決められている．

このように統計が意思決定の指針として活用されている点を強調して，米国の統計学者デミングは「統計は行動のためにある」といっている．この言葉は，現在の社会における統計が，ただ漫然と集められたり，受動的に眺めたりするものではなく，はっきりした問題意識のもとに用いるべきものであると教えているのである．統計を現状の把握や問題の解決のために役立つものとするためには，その利用の目的を明確にし，しかも，具体的に定めて取りかからなければならない．

(3)　問題をはっきりつかむこと

統計利用の目的は，われわれの社会生活や，企業経営や，科学，技術などに関するいろいろな問題を解明する上で役立つような情報を提供することにある．統計という形で真に役立つ情報を提供しうるためには，何よりもまず問題そのものがはっきりとつかまれていなければならない．それには，その問題を具体的に記述する形に書きとめることが必要である．

例えば，

(i)　東京都で夏期に水不足で悩まされないためには，どのくらいの水量を用意すれば良いだろうか．

(ii)　化粧品のメーカーが，その製品の販売を促進するために広告や，その他のマーケテイング活動を行っている．それらの実績を早く知ることによって，常に販売政策を適切に方向づけたい．どうすれば良いか．

(iii)　ある製品のメーカーが品質改善の目的で，従来の製造工程の一部にある変更を行った．そこで，この変更がはたして品質の改善をもたらしたかどうかを知りたい．

(4)　統計としてとらえること

問題がはっきり具体的に述べられたならば，次にその問題を解明するためにはどんな統計が必要であるかを見定めなければならない．

これは与えられた問題を統計の言葉に翻訳する仕事である．例えば，先にあげた例の (i) は，東京都の人口に関する統計や，工場その他の事業所で使う用水量に関する統計を求める必要がある．また，(ii) では銘柄別売上数量や消費者の銘柄評価についてタイムリーな統計を求めることで，問題のありかがわかるであろう．さらに，(iii) は従来のままの工程と，変更の導入された工程のそれぞれから製品の標本をとり出して検査するという方法によって問題に答えることができる．

(5)　役に立つ統計とは

統計から現実に起こっていることを知り，統計をその問題の解決に役立てようとする場合には，統計解析の知識を持っているだけでは，不十分である．その問題領域について専門的な知識を持ち，問題に至っている背景を十分に理解していなければならない．この知識と理解が不十分であれば，統計から正しい情報を得ることはできないし，ひいては間違った意思決定をすることになるだろう．つまり，統計から現実を正しく把握し，問題解決に結びつけるためには，その問題の領域についての専門知識が不可欠である．

例えば，市場調査の場合には，調査する製品について物理的な特性や流通の状況，投下されている広告，および，関連する競争品の状況，消費者の購買行動と嗜好傾向などについて理解していることが必要になる．品質管理のために統計を利用する場合には，製造の現状を知っていることが必須である．

したがって，統計の専門家は，その領域の専門家と力を合わせて問題に当たらなければならない．逆に，統計の専門家に相談するときには，データの背景について詳しく説明し，十分な理解をもってもらうように最大限の努力を払わなければならない．

そのうえで，目的のためにはどのような統計が必要であるかをはっきりさせる．つまり，どのような集団について，何を観察するのかを見極めなければならない．さらに，それらを適切な統計の方法で分析することによって，当初の問題の解決に役立つ情報となる．

統計を役に立つものにできるかどうかは，統計的方法それ自身が決めるものではない．それぞれの問題領域についての専門知識や経験である．このことを理解しておくことは，きわめて重要である．みなさんがこの講座で学ぶさまざまな方法を現実の場面で適用するときには，十分に問題を理解することに努め，その目的のために役立つ統計が何であるかをよく見極めなければならない．その上ではじめて，統計的方法が意義を持つのである．

この節を読んで，前節で各自が取り上げたテーマについての検討が十分であったかを反省し，どのように改善したら良いかを考えてほしい．

本日のまとめ

この節では，統計を利用するときの心構えを学んだ．この講座で学んだことを，実際の問題に応用する際には，その分野の専門知識や経験が必要である．受講生の皆さんは，統計解析をご自身の専門分野として勉強されるのも良いし，統計実務を理解したうえで別の分野の専門性を高めるべく勉強や仕事をされるのも良い．多くの人は後者であろうと思う．その場合には，この講座で学んだ統計実務の知識や技術が，あなたの専門領域での問題の解決のために応用され，力を発揮するだろう．必要なときには，統計解析の専門家と協同して問題にあたることもできる．そのときにもここで学んだことが大いに役に立つはずである．

1.6　補足

(1)　対数

文科系の人々は 対数 というとアレルギーを起こす人が多いであろう．しかし，対数をマスターすると，データの見方が大きく変わる．

バブルが弾ける前は，日本の経済規模は倍々ゲームのように拡大した．最近でもIT産業の成長は目覚しい．

このような場合は，前月（または前年）に比べて いくら増えた ではなく，何パーセント増えた と表現される．このように等比級数的に変化するデータを解析する場合には，対数の知識が大変役に立つ．

また，販売数量と在庫数量，売上高と販売経費，GDPと個人消費支出と物価指数 のように，関連する複数の系列を比較する場合にも，比率の動きが関心の対象であるから，対数は有用である．

2を底とする対数

1, 2, 4, 8, 16, 32, ... **という数値の並び（数列 という）は，左の数の2倍に**

なっている.

4 **は** 2 **を**2 **回掛けた値で,** 8 **は** 2 **を**3**回掛けた値である. これを,**

$$2 = 2^1,\ 4 = 2^2,\ 8 = 2^3,\ 16 = 2^4,\ 32 = 2^5,\ 64 = 2^6,\ 128 = 2^7\ \ldots$$

と表わし, 2**の 冪乗 （べきじょう）という.**

これを左に延ばすためには,「順に 半分 にしていけば良い」ので,

$$\ldots,\ 0.25 = 2^{-2},\ 0.5 = 2^{-1},\ 1 = 2^0,\ 2 = 2^1,\ 4 = 2^2,\ 8 = 2^3$$

となる.

2**の肩の数値を 「2 を底とする対数」といい,**

$$\log_2 8 = 3$$

という式で表わす. 対数は英語で Logarithm **と呼ばれ, 数式では最初の**3 **文字が用いられる.**

ここで,

$4 \times 8 = 32$ **を**

$2^2 \times 2^3 = 2^5$

と書き直す.

下の式の肩の数字, 2, 3, 5 **は上の数値** 4, 8, 32 **の対数となっている.**

左辺の2**の肩の** 2 **と** 3 **を加えると, 右辺の肩の** 5 **が得られる. すなわち,** 32 **の対数**(5) **は** 4 **の対数**(2) **と** 8 **の対数** (3) **の和となっていることが分かる.**

これから, 対数を使うことにより, 乗算が加算に置き換えられることが分かる.

$4^3 = 64$ **である.** $\log_2 4 = 2$ **で** $\log_2 64 = 6$ **であるから,**

$$\log_2 64 = \log_2 4^3 = 3 \log_2 4$$

の関係が成立する．すなわち，4 を 3 乗した 64 の対数は 4 の対数の 3 倍になっている．

以上に説明した性質は，底が 2 であることに限定されず，一般的に成立する．

10 を底とする対数（常用対数）

「10 を底とする対数」は類推によって

$$\log_{10} 0.1 = -1,\ \log_{10} 1 = 0,\ \log_{10} 10 = 1,\ \log_{10} 100 = 2,\ \ldots$$

となる．この対数は広く用いられるので 常用対数 と呼ばれる．

$\sqrt{10} = 3.16$ の常用対数はいくらと定義したら良いであろうか？　$\sqrt{10}$ を 2 乗すると 10 になるから，

$$\left(\sqrt{10}\right)^2 = 10^1, \qquad \sqrt{10} = 10^{0.5}$$

すなわち，3.16 の常用対数は 0.5 とすれば良いであろう．以下，常用対数は底を省略し単に $\log 3.16 = 0.5$ のように書く．また，単に 対数 といえば，常用対数を表わすことにする．

それでは，$\log 2$ はいくらになるであろうか？　$2^3 = 8,\ 2^4 = 16$ であるから，

$$3 \times \log 2 = \log 8 < \log 10 = 1, \quad \log 2 < 1/3$$

$$4 \times \log 2 = \log 16 > \log 10 = 1, \quad \log 2 > 1/4$$

2 の対数は $1/3 = 0.333$ と $1/4 = 0.25$ の間の値を取るであろう．

また，$2^{10} = 1024 \approx 1000$ であるから，

$$10 \log 2 = \log 1024 \approx \log 1000 = 3$$

となり，$\log 2$ は 0.3 より少し大きい値となることが予想される．

$\log 2 = a$ であるとすると，$\log 5$ はいくらになるであろうか？

$2 \times 5 = 10$　　であるから，

$$\log 2 + \log 5 = \log 10 = 1.0$$

が成立しなければならない．したがって，$\log 5 = 1 - a$ **となる．**

このような性質を持つ常用対数は，電卓も電子計算機もなく，ソロバンが唯一の計算手段であった時代には，極めて重要な役割を果たした．1.000 **から** 0.001 **刻みで** 9.999 **まで（詳しい表では** 0.0001 **刻み）の常用対数が表として公刊され，多方面で利用された．**

現在では関数電卓でも，2 **と** log **キーを押すと，**0.30103 **という値が得られる．確かに，上に示した** 0.25 **と** 0.33 **の間の数値で，**0.3 **よりわずかに大きい値である．**

Excel **では，**

```
=LOG(2)
```

を入力すると， 0.30103 **が得られる．**

それでは， 20 **の対数はいくらになるであろうか？** $20 = 2 \times 10$ **であるから，**20 **の対数は** 2 **の対数と** 10 **の対数を加えれば良い．**10 **の対数は** 1 **であるから，**

$$\log 20 = \log 2 + \log 10 = 0.301 + 1 = 1.301$$

となる．このような関係があるから， 1.000 **から** 9.999 **の範囲の対数表があれば，どんなに小さな数，大きな数でも，その対数を求めることができる．**

定率法による固定資産の償却

固定資産の償却方法には 定額法 と 定率法 がある．

定額法では，毎年の償却額が一定である．法定償却年数が10**年のときには，**10 **年間に取得価額の**90%**を均等に償却し，**10**年後の残存価額が取得価額の**1/10**にする．すなわち，**0.90/10 = 0.09 **で毎年の取得価額の** 0.09 **倍を償却する．**

定率法では，毎年の償却額は前年度末の残存価額に一定の係数（償却率）を掛けたものである．

それでは，10**年後の残存価額が取得価額の** 10% **になるためには，償却率はどのようにして決めれば良いであろうか？**

取得価額を X，償却率を a とすると，i 年後の残存価額 Y_i は，

$$Y_1 = (1-a)X,\ Y_2 = (1-a)^2X,\ \ldots,$$
$$Y_i = (1-a)^iX,\ \ldots,\ Y_{10} = (1-a)^{10}X = 0.1X$$

となる．最後の条件が満たされるように，償却率 a を決めれば良い．

$(1-a)^{10} = 0.1$ になるためには，両辺の対数を取り，

$$\log\left((1-a)^{10}\right) = 10\log(1-a) = \log(0.1) = -1$$
$$\log(1-a) = -0.1, \qquad 1-a = 10^{-0.1}$$

であれば良い．

電卓に -0.1 を入力し，10^x キーを押せば，$1-a = 0.794$ が得られる．これから，償却率 a は $a = 1-0.794 = 0.206$ となる．すなわち，前年末の残存価額の $20.6\,\%$だけ償却すれば良い．

Excel では，

```
=1-10^(-0.1)
```

と入力すると，0.206 が直接得られる．

現実には，このようにして計算した償却率を用いるのではなく，国税庁が作成した法定償却率の表に書かれている償却率を用いるのであるが，その値がどのようにして計算されたのかを知っておくのもいつか役に立つことがあるかもしれない．

2 集団の観察と統計的規則性

2.1 集団の観察

(1) 集団の観察

私たちが統計を扱うときには，常になんらかの 集団 を考察しており，そして統計によって，その集団の特徴を数量的に表わす役割を果たしている．この場合に，考察される集団は，ある 共通性 をもつ個体の集合として構成されている．

例えば，§1 の＜例2＞ 出生数（人口動態統計）は「日本に在住している日本人（日本国籍の人）」，＜例3＞ 人口（国勢調査）は「調査時に日本国内に常住している者（外国籍の人を含む）」という共通点をもった人が対象である．また，＜例7＞ 管理図で扱ったデータは，一定の製造工程から特定の期間中に生産されたという共通性をもつ部品の集合である．

このように，ある共通性をもち，一つの集団を構成している個体のことを 集団の構成単位 という．どんな統計を扱う場合でも，その集団の 共通性 は何か，その 構成単位 は何かを明確につかむことが必要である．

(2) 観察の特性

対象とする集団の構成単位が有限個である場合には，人口の総数などのようにその構成単位の総数を表わす統計がよく用いられるが，これを集団の大きさといっている．

集団の大きさを表わす統計のほかに，その集団の内部構造を表わす統計がある．例えば，人口の場合にその構成単位である個人を男女別という特性の面で観察すれば，男女それぞれの人数あるいはそのパーセント (%) という統計が得られる．また，年齢や収入という特性の面で観察すれば，年齢別や収入階層別の人数，あるいはパーセント，さらに，平均年齢や平均収入などの統計が得ら

れる．このように内部構造を表わす統計は，一定の特性を取り上げて，その特性の面における，各構成単位間の違いを観察することにより得られる．

集団の内部構造を表わす統計は，有限個の構成単位からなる集団ばかりでなく，無限個の構成単位からなる集団についても考えられる．例えば，一定の製造工程のもとで生産される部品の集団について，平均直径，平均重量などを考えることができる．

(3)　質的な特性と量的な特性

観察する特性には，質的な特性と量的な特性がある．例えば，下のように，各個人に対して性別，生年月日，血液型，学歴，身長のデータが得られているとしよう．

	性別	生年月日	血液型	学歴	身長 (cm)
荒井	女	1965/4/3	A	高校卒	158.7
井上	男	1970/12/13	A	大学卒	172.3
上田	男	1962/2/3	B	高校卒	176.5
小川	女	1966/8/23	AB	大学院卒	168.3
...	...	...	...	...	...

このデータにおいて，質的な特性は，性別，血液型，学歴で，量的な特性は，生年月日，身長である．質的な特性について得られたデータを 質的データ，量的な特性について得られたデータを 量的データ という．

質的データと量的データの性質をより明確に区別すると，以下の４種類になる．

質的データ	(i)	名義尺度（分類尺度)
	(ii)	順序尺度
量的データ	(iii)	間隔尺度
	(iv)	比例尺度（比尺度）

質的データは，名義尺度と順序尺度に分けられる．名義尺度は，その分類の順序に意味（決まった規則）がないもので，性別（男，女）と血液型（O, A, B, AB）が該当する．順序尺度は分類の順序に意味があるもので，学歴（中学卒，高校卒，大学卒，大学院卒，など）である．

量的データは，間隔尺度と比例尺度に分けられる．間隔尺度はデータの間の差に意味があるが，比率には意味を持たないものである．上の例では生年月日が該当する．例えば，荒井さんの生年月日と井上さんの生年月日の差は意味を持つが，比率は意味を持たない．

一般に，間隔尺度の例としてよく用いられるのが摂氏温度である．摂氏30 ℃は摂氏20 ℃よりも10 ℃熱いという表現は意味があるが，1.5倍熱いという表現は意味がない．

それに対して，比例尺度は，差と率の両方に意味のあるものである．上の例では身長が該当する．量的データの多くは比例尺度である．

量的データなのか，質的データなのか，さらには名義尺度なのか，順序尺度なのかの違いによって，データ解析の方法が異なるので，この区別はたいへん重要である．第2単元では，この区別に沿って，データの記述方法を学習する．

これとは別に，データの統計的な性質から，データを計量値と計数値に分けて考えることがよくある．先の例の身長は，データの表現としては単位が cm **で，小数点以下1桁目まで表示されている．しかし，これはもっと小さな桁まで測定されていたものを小数点以下1桁目にまるめて表示しているのかもしれないし，小数点以下1桁目の数字は目見当で得られているかもしれない．とにかく，より精密に測れる身長計を用いれば，より詳細な値の得られることは確かである．このように元来，連続的な値で得ることのできるデータを計量値という．一方，元来，数を数えて得られるデータを 計数値 という．計量値 は連続的な値，計数値は離散的な値を取るので，それぞれ 連続量，離散量 と呼ばれることもある．**

質的データの場合，得られたデータを分類ごとに集計して，計数値として処理することが多い．例えば，以下のように性別について男，女それぞれの度数を求め，全体に対する割合を求めるといった処理である．

	度数	割合
男	60	40%
女	90	60%
計	150	

この場合，度数のデータは数を数えて得られているので計数値である．また，割合もそれを求める際の分子が度数であるから，計数値として扱われる（割合であっても，食塩水の濃度のように，割合を求める際の分子が計量値の場合，計量値として扱われる）．

計数値なのか，計量値なのかの違いによって，用いられる分布が異なるので，この区別もたいへん重要である．第3単元と第4単元では，計数値と計量値で分けて学習する．

さて，以上のようなデータの種類の別を述べたのは，後述するデータの扱い方や適用する統計解析の方法が異なるからである．

(4)　静態統計と動態統計

2000年の国勢調査によると，日本の総人口は，126,925,843人で，そのうち65歳以上の人口は，22,005,152人となっているが，これらの統計は特定の時点 — 2000年10月1日0時現在 — における人口を表わすものである．

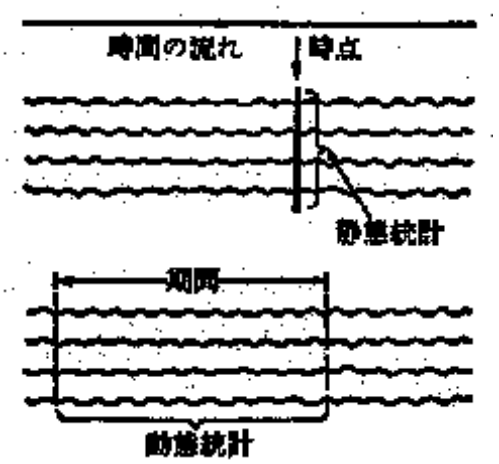
もともと人口は，時々刻々に変化しているものであるから，これを統計としてとらえるためには必ず特定の時点を指定しなければならない．このように，時間とともにその大きさや内部構造の変化している集団を，特定の時点でとらえて，その時点の切断面の大きさや構造を表わすような統計を一般に 静態統計という．

出生数，国民総生産，貿易額，小売販売額などの統計は，時間のとらえ方という点で，人口などのような静態統計とは全く異なっている．特定の時点の国民総生産や貿易額などというものは考えられない．これらの統計は，時点ではなく特定の期間を指定したとき，はじめてその期間中における流れの量を測るものとしての意味をもつものである．このような種類の統計を一般に 動態統計という[1]．

どんな統計でも静態統計か動態統計かのどちらかである．統計を扱う場合に

[1] 企業の会計報告書には「貸借対照表」と「損益計算書」が含まれる．前者は，期末における資産の内容を明らかにしたもので，静態統計に対応する．後者は，その期中における収支をとらえるもので，動態統計に対応する

表示2.1: 静態統計と動態統計



は，それがどちらの種類であるかを明らかにしなければならない．静態統計ならばどんな時点で，また，動態統計ならばどんな期間で規定するかが重要な問題である．統計を利用する側でも，これらの規定をはっきりつかむことが大切である．

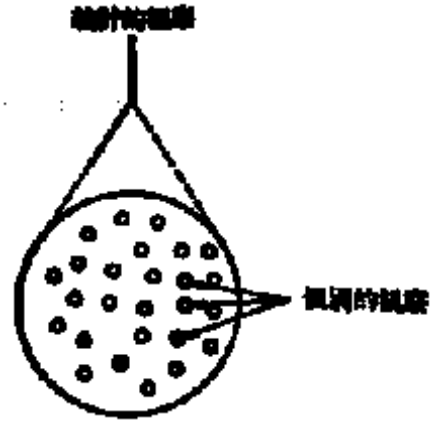
(5) 統計的観察

集団は個体の集合であるから，集団を観察する場合は個体に対する個別の観察にもとづかなければならない．しかし，この場合に，個体に対する個別の観察を集団観察のための単なる手段とみるか，あるいは，個別の観察自体を問題とするかによって，観察の性格は根本的に違ったものとなる．

例えばいま，ある市のすべての食料品店を対象として過去1年間の販売額が調査されたとしよう．もしこの調査で集められた調査票の記入内容を，個々の食料品店に課する税額を決める基礎として税務官が利用するとすれば，それは個体に対する個別観察自体を問題とすることになる．

ところが，もしこの調査の結果を集計して，その市の食料品全体としての販売額総計や，1店あたり平均額を算出したり，あるいはあらかじめ販売額のいく

表示2.2: 個別的観察と統計的観察



つかの階級（例えば750万円未満，750万円以上1,000万円未満，1,000万円以上1,200万円未満など）を作り，それぞれの階級に属する店の数やパーセントを求めて，その数字を他の市や国全体としての対応数値，または，前年の同市の対応数値と比較することによって，なんらかの結論を導こうとするとしよう．この場合には，個々の食料品店に対する個別の観察結果は，集団観察のための1件のデータとして利用されるにすぎないことになる．

後者の例に示されるような利用を目的とする集団観察のことを 統計的観察 という．それに対して前者のような場合には，たとえ，集団を対象とする観察であっても統計的観察ではなく，個別的観察 である．

出生届に記入された子供の生まれた年や性別，母親の年齢が人口動態統計としてまとめられているが，これは統計的観察であり，出生届は個別的観察である．

統計は統計的観察の結果であって，集団について集団としての特徴，すなわち，集団性を記述するものである．

本日のまとめ

統計が，集団の特徴をとらえるものであることを理解するのが§2の学習の中

心である．

§2.1 では，統計はその個体の観察に基づいていることを確認した．個体を観察して得られたデータおよびそれを集計したデータは，その性質によって，質的データ と 量的データ，離散量 と 連続量 に区別できることを学んだ．手近なデータについて，この区分を考えてみると良い．第2単元以降で，データをグラフ化，整理，解析をするとき，データの種類によってその方法が異なるので，非常に大切である．今日の学習で完全に分からなくても，第2 単元に進むと自然に分かるようになるであろう．また，手近な統計が静態統計と動態統計のどちらに該当するのかも考えてみよう．

2.2　全数観察と標本観察

(1)　標本と母集団

§1.2 の＜例4＞就業構造基本調査 の日本のデータは，日本の全世帯の調査によるものでなく，その一部分である約44 万世帯の15 歳以上の人を対象として調査を実施した結果である．それにもかかわらず，政府はこの結果からそれが15 歳以上の国民全体を表わしているものとしてデータを解説している．私たちも同様に，直接の調査対象者そのものの情報としてでなく，むしろそのもととなっている15 歳以上の国民全体についての情報を反映するものとしてこのデータを考察した．一般に，研究の対象となっている集団，すなわちそれについての情報が求められている集団のことを 母集団 といい，それを代表する一部分として実際に観察されている集団のことを 標本（または サンプル）という．

(2)　全数観察と標本観察

母集団のすべての構成単位を実際に観察して，統計を獲得する仕方を 全数観察 または 全数調査 といい，実際には標本だけを観察して，その結果から，母

集団に関する情報を導くような仕方を 標本観察または 標本調査 という.

私たちが取り扱う統計データには，母集団の全数観察の結果よりも，標本観察の結果のほうが多い．§1.3 にあげた例は標本調査として行われるのが普通である.

ところで，標本観察によって，母集団に関する情報を求めるためには，標本が正確な母集団の縮図となっている必要がある.「例7 品質管理」であげたスプーン製造工程の例では，一定時間ごとに取り出したわずか5つのスプーンの重量から，製造工程が異常なく稼動しているかどうかを判断している．もしも計測の担当者が，重量が重そうなスプーンと軽そうなスプーンを除いて検査すれば，製造ラインに異常が発生しても察知することはできない．製造工程の状況を正しく判断するためには，重量をはかるスプーンの取り出し方に恣意が入らないように，選び出す手続きを規程しなければならない.

市場調査で「20歳代女性の1ヵ月の化粧品代がいくらであるか」を聞き取って調べたいとしよう．この目的のために，休日に銀座を歩いている20歳代の女性を対象に調査を行うと，おそらく日本の平均的な女性よりも高い金額となるであろう．それでは,「銀座のように百貨店や高級ブランドの専門店が並ぶショッピング街ではなく，日本の平均的な地域で平均的な女性に対して調査をすれば良い」と考えるかもしれない．しかし，このとき平均的とは，誰がどのように判断できるのであろうか．もしも平均的であることを正しく判断できるのであれば，調査をする必要はないはずである.

このように，標本を選ぶときに選ぶ人の判断が入ってしまうようでは，母集団の正確な縮図となる標本を作ることはできない．そこで，母集団に属しているものや人が どれも 同じ確率で選び出されるように，無作為抽出の手続きが定められている.

■ 演習 3　ある保健所で，健康診断にきた成人男性について，自分の健康状態についてアンケートを取ったところ，次のようなデータを得た.

	30代	40代	50代	60代	70代
自分の健康に自信がある	15	13	28	30	26
自分の健康に自信がない	10	8	14	7	4
合計	25	21	42	37	30
自信がある割合	0.60	0.62	0.67	0.81	0.87

これから，歳を取るほど，自分の健康について自信を持っているといえるか．■

(3) 有限・無限母集団

これらの例で取り上げた標本観察では，考察されている母集団は，一定の時と場所とを限定した有限個の構成単位からなるもので，有限母集団 と呼ばれるものであった．ところが，＜例7＞の品質管理の例や＜例8＞の実験計画で考察されている母集団は，無限個の構成単位からなっている 無限母集団 であると考えられる．

まず，＜例7＞について考えてみよう．ここで実際に観察されたデータは，測定値すなわち標本観察であるが，この場合の母集団はなんであろうか．この問題では，標本観察の結果によって製造工程そのものが安定状態にあるかどうかを判断するわけであるから，標本のデータは製造工程そのもの，あるいは，その製造工程が産み出し得るすべての部品の集団という，「仮説的」な無限母集団に関して情報を提供していることになる．

ここで，有限母集団 と 無限母集団 の違いを説明したが，この違いによって，統計解析の手順を変えなければならない場合は少ない．

本日のまとめ

今日は全数観察と標本観察とを区別して考えることの必要性を学んだ．母集団に関する情報を得ることが目的であり，標本観察はその手段である．皆さんが知っている統計について，母集団であるか標本であるかを明確にしておくことは，非常に重要である．母集団に関する情報を得ることが目的であり，標本観察はそのための手段であることを理解してほしい．

2.3　大数観察における比率の安定性

(1)　硬貨投げの例

1枚の10円硬貨を投げて「表」が出るか「裏」が出るかを観察する実験を考えてみよう．この場合,「表」が出ることも「裏」が出ることも同程度に起こりそうであるから，もし，1枚の硬貨を投げる試行を10回繰り返すならば,「表」と「裏」が5回ずつ現われることを期待したくなる．

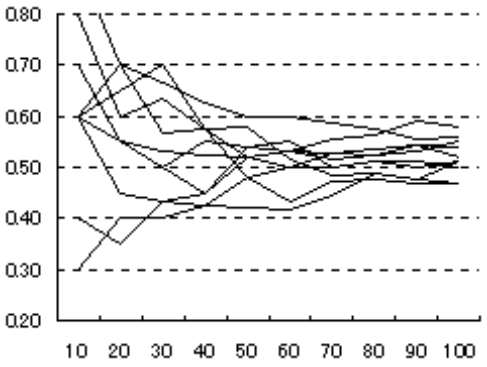
ところがいま，実際に試みてみたところ，10回中「表」が7回現われてしまった．すなわち,「表」の出現比率は，期待した$1/2$とは異なる0.70という値であった．そこで，さらにこの試行を10回だけ追加して，全20回について「表」の出現比率を出してみると $13/20 = 0.65$ となった．さらに10回だけ追加して，全30回についてみると，この比率は $18/30 = 0.60$ となった．このような試行を100回まで行って，途中の40回目，50回目，$\cdots$，90回目まで，及び，100回目までにおける「表」の出現比率を計算し，これらを表示2.3にまとめた．

このような一連の実験を10セット実施し,「表」の出現比率を縦軸に取り，試行回数を横軸に取って図示したものが表示2.4の10本の折れ線である[2]．

表示2.3: 硬貨投げ実験の結果

試行回数	「表」の出現回数	「表」の出現比率
10	8	0.80
20	12	0.60
30	19	0.63
40	23	0.58
50	29	0.58
60	31	0.52
70	34	0.49
80	39	0.49
90	43	0.48
100	51	0.51

表示2.4: 「表」の出現比率



この実験によって次のことが分かる．すなわち，試行回数が少ない間は「表」

[2] 表示2.3，表示2.4の実験は，硬貨を用いた実験結果ではなく，Excelで実行したシミュレーションの結果である．その具体的な方法はExcelシートに説明されている．

の出現比率は，初めに期待された $1/2 = 0.5$ とはかなり隔たった値となるが，試行回数が増すにつれてこの出現比率は，次第に期待された値 0.5 の近辺に安定してくる．試行回数をさらに増して，1,000 回，10,000 回としていけば，この出現比率は，0.5 にますます近いところに安定してくることが想像されるであろう．

ここであなたも，上と同様の実験を試行回数 100 回まで試みて，「表」の出現比率の安定してくる様子を確かめてみると良い．さらに，一つのサイコロを振ったときに 1 の目が出るかどうかを観察する実験を，試行回数 100 回まで試みて，途中，10 回目，20 回目，30 回目，$\cdots$ までの間での 1 の目の出現率を計算し，この出現率が次第に $1/6$ の近くに安定してくる模様を確かめると良い．実験の途中では，硬貨の「表」ばかりが 3 回も 4 回も続けて出たり「1 の目」がなかなか出なかったりして不安になるかもしれないが，実験を続けていくうちに，先に述べたような安定性が次第に認められるようになるはずである．

(2)　統計的規則性

このように，一定の条件のもとで 特定の試行を繰り返すとき，各回の試行結果はまったく偶然にまかされている．それにもかかわらず，多数回の試行を総合してみると，そこに大数観察における比率の安定性が認められる．この事実を，**統計的規則性** という．

硬貨投げの場合には試行の結果は「表」と「裏」との 2 通りしかない．そして，正しく作られた硬貨であれば，そのどちらも均等に起こり得るはずである．それを根拠として試行回数の半数に「表」が出ることが期待される．

またサイコロを振る実験では，「試行の結果として現われる目は，1，2，3，4，5，6 の 6 通りだけであって，そのどれも均等に起こり得るはずだ」という判断から，試行回数の $1/6$ について 1 の目の出現を期待するわけである．

これらの例で代表される射倖（シャコウ）実験（偶然に支配される実験）においては一般に，試行や観察の結果が均等に起こり得ると判断されるいくつかの場合に分類されるために，多数回の試行中，特定の場合の出現する比率を，試行に先立って，近似的ながら予知することができる．しかも，実際に大数観察

を実施してみると，その結果は前の例で見たように，現実の出現比率は予知した数値の近辺に次第に安定してくることを確かめることができる．

このような場合には，試行に先立って大数観察の結果を近似的ながら予知できるであろう．

(3)　出生児の性比

前節で学んだような射倖実験の場合には，試行の結果が均等に起こり得る場合に分類できるために，試行に先立って大数観察における出現比率を予知する根拠がある．しかし，このような事情にない現象でも，一定の条件のもとで多数回の試行や観察を行ってみると，結果として，大数観察における出現比率の安定性が認められることが多い．その一つの顕著な事例として，一国の出生児の男女性比の安定性について見ることにしよう．表示2.5 は わが国の出生届をもとにして出した 1905 年から 2002 年までの期間における，各年の出生率（人口1,000 人に対する出生数）と出生性比（女児100 に対する男児の数）を示している（途中は適当に省略してある）．

表示2.5: 出生率と出生性比

	西暦	出生率	出生性比		西暦	出生率	出生性比
明治38年	1905	31.2	102.7	昭和 49	1974	18.6	106.3
39	**1906**	**29.6**	**108.7**	51	1976	16.3	106.2
40	1907	34.0	102.7	53	1978	14.9	106.0
41	1908	34.7	104.6	55	1980	13.6	106.0
42	1909	34.9	104.1	57	1982	12.8	105.5
...	...	...	...	59	1984	12.5	105.4
昭和38年	1963	17.3	105.7	61	1986	11.4	105.9
39	1964	17.7	107.6	63	1988	10.8	105.6
40	1965	18.6	106.3	平成2年	1990	10.0	105.4
41	**1966**	**13.7**	**107.6**	4	1992	9.8	106.0
42	1967	19.4	105.3	6	1994	10.0	105.6
43	1968	18.6	107.1	8	1996	9.7	105.6
44	1969	18.5	107.2	10	1998	9.6	105.4
45	1970	18.8	107.1	12	2000	9.5	105.8
47	1972	19.3	106.5	14	2002	9.2	105.7

われわれの家庭や親戚，友人の家庭などに限定してみると，男児ばかり，女児ばかりの生れる家庭もあるから，狭い範囲でみると出生児の性比の安定性などということはとうてい認められない事柄であろう．

ところが，これをわが国全体という広い範囲で観察してみると，表示2.5の示すとおり出生児の性比は年によってほとんど変わらない，かなり安定した値を取ることが分かる．

ただし，1906年と1966年は，出生率が低下するとともに，出生性比が高くなる（女児に対して男子の比率が高くなる）というやや特異な値を示している（表示2.5では太字で示す）．また、その前後の年は，逆に出生性比が低くなっている．特に1906年にこの傾向が顕著に現れている．その原因は，この両年が丙午の年に当たるためである．女児が丙午の年に生れることをきらう風習があるので，この年の初めや終りに生れた女児は，その前年，またはその次の年に生れたこととして届け出た人々が多いために，この両年の男児比率が高く，その前後の年の値が低いのであろう．1966年よりも1906年で前後の年を含めて出生性比が異常な値であるのは，出生日の証明が現在よりも緩やかであったためであろう．また，1966年は出生性比よりも出生数の落ち込みが目立つが，これは産児制限の知識の普及が背景にあると思われる．

出生児の大数観察における男女性比の安定性という事実にはじめて気付いたのは，17世紀の英国の人口学者ペッティーである．彼のこの発見は，その他のいろいろな現象の大数観察における統計的規則性を発見する端緒になったものとして有名である．人口の大数観察に現われる統計的規則性のもう一つの例をみよう．

(4)　生命表

ある人がこれからの1年間に死亡するかどうかということはまったく予知することのできない事柄であろう．ところが，人口の大数観察を行ってみると，特定の年齢の男子が1年間に死亡する割合はほぼ安定している．女子についても同様であるが，死亡率は男子の同年齢のものの値とはかなり異なっている．男子についても女子についてもこの死亡率は，年齢によって著しく違った値となる．

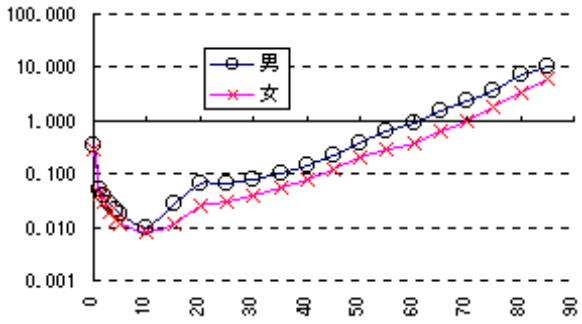
医学の進歩や生活環境の変化によって，死亡率は少しずつ小さくなっているが，その変化は比較的なだらかである．

それらの値を人口の大数観察によって求めておけば，それにもとづいて性別，年齢別の 平均余命，すなわち，特定の年齢の男子や女子が平均してこれから何年間生きのびるだろうかを表わす値を出すことが可能になる．

死亡率や平均余命の数値は，生命保険事業の経営の基礎となる重要なものであるから，厚生労働省では死亡届と国勢調査の結果の人口とによってこれらの数値を求め，それをまとめて表示した 生命表 を公表している．表示2.6 は，2002 年に公表された第19回生命表の一部である．これは2000 年の事実を表わすものである．

表示 2.6: 年齢別・性別死亡率 (%) (2000 年)

年齢	男	女	年齢	男	女
0	0.345	0.298	35	0.099	0.054
1	0.051	0.044	40	0.147	0.078
2	0.038	0.030	45	0.232	0.122
3	0.027	0.020	50	0.392	0.196
4	0.021	0.014	55	0.625	0.279
5	0.018	0.012	60	0.923	0.383
10	0.010	0.008	65	1.498	0.618
15	0.027	0.012	70	2.384	0.999
20	0.063	0.025	75	3.784	1.740
25	0.068	0.031	80	7.401	3.365
30	0.077	0.038	85	10.640	6.316



表示2.6の下は，縦軸に死亡率の対数を取っている．

生命保険会社ではこのような生命表にもとづいて，保険料金を合理的に決めているのである．そのために，必要な計算の原理を研究する数学の分野を 保険数学といい，それを専門としている職業の人々のことを保険会社では，保険計理人 （アクチュアリー）と呼んでいる．

本日のまとめ

一見偶然に見えることであっても，大数観察を行うとある安定した比率で起こる，という統計的規則性は，統計解析の土台となっている大変重要な考え方である．この考え方をよく理解するためには，ここで勧めている硬貨投げやサイコロを振る実験をあなた自身で試みるのが一番の近道である．まず，この単純な実験を行ってみることが大切である．その結果を見ると，統計的規則性という「偶然性の法則」の存在を深く納得できるはずである．

2.4　調査・実験を始める前に(1)

(1)　調査・実験の目的

気象統計では，ある地域の温度の季節変化がとらえられる．ある地方を旅行するとき，服装を決めるときの参考として使う市民がいるかもしれない．

このようなデータは，基本の変化を記述しているが，なぜ変化したかまでは示してくれない．

それに対して，企業で商品の販売を促進するために市場調査をしたり，新製品の開発のために実験をする場合は，販売政策と販売数量の関係や，原料配合と製品の品質の関係を知って，改善に役立てるという基本的な目的がある．

この目的を果たすためには，目的である販売数量や製品品質が何によって，また，どのように変化するかが明らかになるように，調査・実験の計画が立てら

れる.

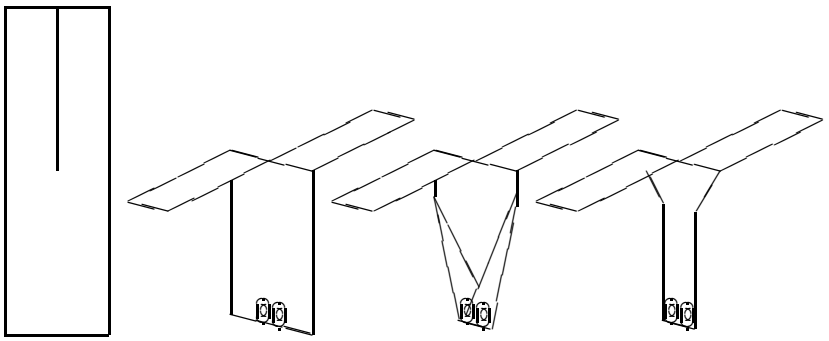
そこでは，どのような項目について調査をしたら良いか，または，どのような条件を変化させて実験したら良いか，を十分に検討して，調査・実験を計画する必要がある.

(2)　特性要因図

適当な大きさの紙に表示2.7(a)のように切れ目を入れ，前後に折り曲げ，重りとしてクリップをつけると，表示2.7(b)のような紙ヘリコプターができ上がる．翼や足の長さや幅，重りのクリップの大きさや個数は自由に変えられる．また，翼や足の形を変化することも許される．(c), (d) のように，足を折り曲げたり，切りとったりしても良い．ただし，2 枚以上の紙を組み合わせることは許されないことにする.

こうして作られた紙ヘリコプターを2メートル位の高さから落とすと，最初は落下し，途中から翼が回転し始め，ゆっくりと下降する.

表示2.7: 紙ヘリコプター



ここで課題は，ある決められた高さから落としたとき，できるだけ滞空時間の長い紙ヘリコプターを設計するための実験をいかに効率良く行うかである.

ここで，目標の滞空時間を 品質管理の分野では特性値，それに影響を与えると考えられる原因項目を 要因 と呼ぶ.

特性値に影響を与えると思われる要因を列挙する．紙ヘリコプターの滞空時

間に対する要因としては

紙の質（厚さ，硬さ，など）

翼（形，幅，長さ，曲げる角度，など）

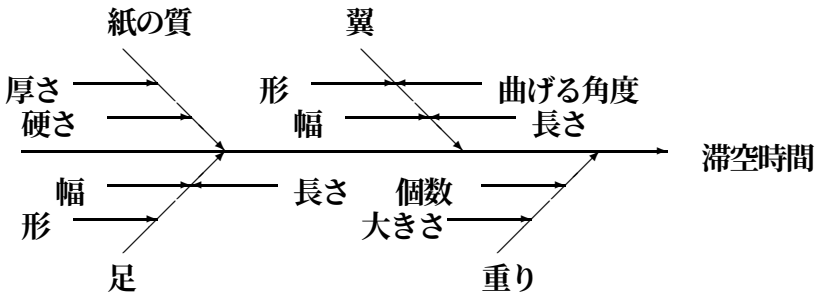
足（形，幅，長さ，など）

重り（クリップの大きさ，数，重量）

などが考えられる．

これらの要因を整理して，表示2.8のような図で表わす．

表示2.8: 特性要因図の例



この図は 特性要因図 と呼ばれ，問題を整理するための道具として品質管理で最も広く用いられているものである．

(3)　因果関連図

特性要因図は，たくさんの要因を洗い出すには有効であるが，要因がなぜ，また，どのような因果の繋がりで特性に影響するかは分からない．

そこで，原因と結果の因果関係がわかるように工夫するのが良いであろう．

例えば，次のように考える．

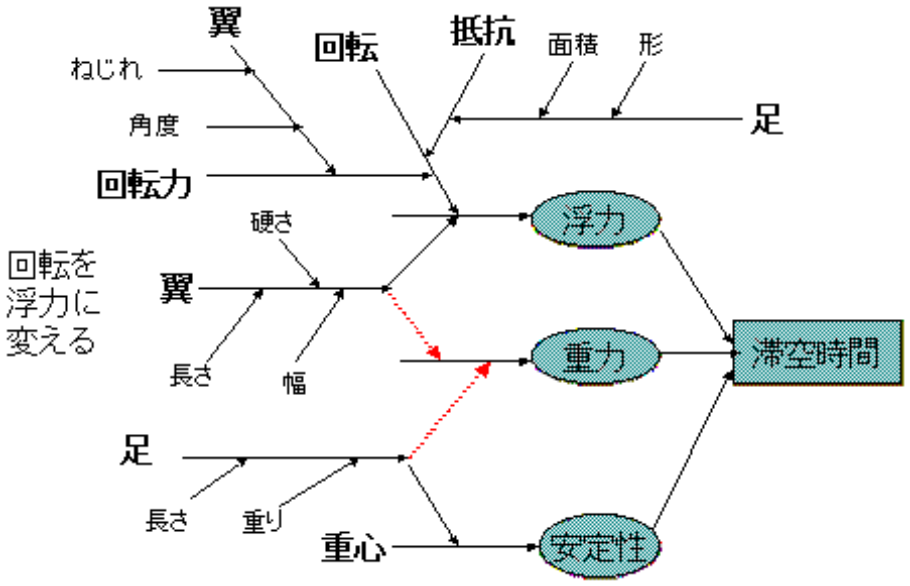
- 紙ヘリコプターが空中に止まるのは，浮力が重力に打ち勝つためである．
- 浮力を増すには，翼が広くてしっかりしており，回転が早く滑らかでなければならない．
- 回転を早くするには，回転力を強く，回転に対する抵抗を低くする．
- 回転力は翼によって生じ，翼のねじれが必要．

- 回転に対する抵抗は足の面積によって決まる.
- 回転を浮力に変えるのは翼の役割.
- 安定して飛ぶかどうかは，浮心と重心の位置関係によって決まる.
- 重心を下げるには，足を長くし，重りを重くする.
- 重りを重くすると，重心は下がるが，重力が増してしまう

…

因果関連図 の作り方はまだ定型化されていないが，例えば，表示2.9のような図が考えられる.

表示2.9: 因果関連図



表示2.9 の左で，翼の大きさは，右上の回転を浮力に変えるという好ましい効果と，右下の重さを増やすという好ましくない効果（点線で表わす）の両方を持っている．このような場合，翼は広ければ広いほど良いのではなく，最適な長さがあると想像される.

同様に，足の長さや重りを増やすと安定性は増すが，重くなるという悪影響（点線で表わす）が生じる.

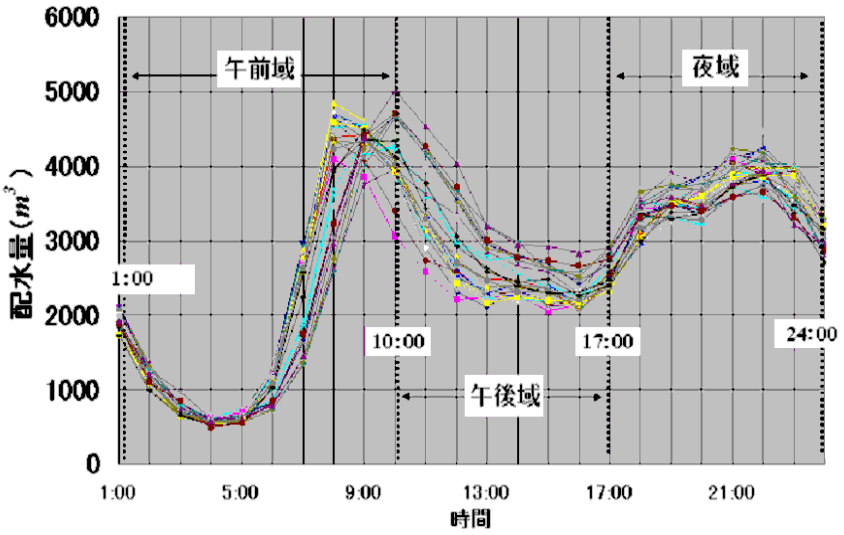
このように，考えを次々に発展させるような因果関連図が望ましい．

(4)　水道水の消費量の予測

ある地方都市の上水道部門が，翌日の水消費量を予測する方法を確立するために，毎日の水消費量について調査をした．

ある1ヵ月間の日々の水消費量の時間的変化を表示2.10 に示す．

表示 2.10: 水の消費量の時間変化

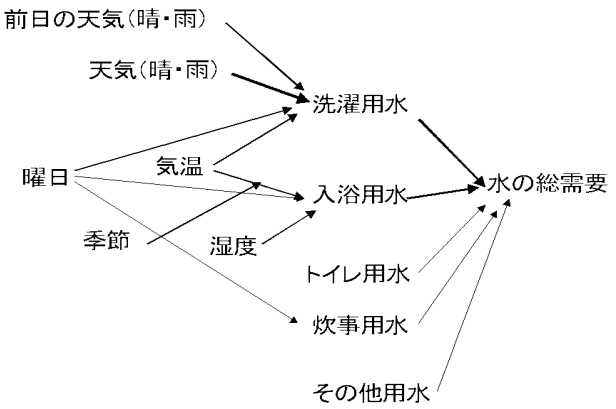


この図から，2 つのピークがあることが分かる．午前のピークは洗濯，夕方から夜のピークは入浴によるものである．

午前のピークには2 つのパターンが混ざっている．8時前後のピークは平日の，10 時前後のピークは土日のパターンである．1日の消費量にはこのように異質なものが含まれている．これを無視して，予測する式を作ってもうまくゆかないであろう．

水の消費量を用途（時間帯）別に分けて，それぞれに対して何が影響するかを考える．これを因果関連図にすると，表示2.11 が得られる．

表示 2.11: 水消費量の因果関連図



洗濯は天候の影響を大きく受ける．雨の翌日の晴れは洗濯日和である．

入浴に対する気温の影響は，夏は暑くて汗を出した日は入浴する．冬は今日は冷えたからゆっくり温まろうかということになるであろう．となると，気温の影響は季節によって異なるのではないか．

入浴に影響を与える因子は，銭湯の客数を調査することによって思わぬ知見が得られるかもしれない．

このようなことを十分に考えて，データを 層別 し[3]，解析を進めることにより，現実に良く合ったモデルが構築できる．

表示2.11の図では，特に重要と思われる因果の繋がりは太線で表わしている．

本日のまとめ

一般にある現象に関して，原因として作用する因子はたくさんあるが，その因果のつながりを皆さんなりに整理して，仮説を立てることは，統計を実際に

[3] **一般にある現象に関して作用する因子はたくさん重なりあっている．しかし，その減少を漫然と眺めているだけでは，情報はあまり得られない．一つの現象を何かの特徴でいくつかの層に分けて調べると，有益な場合がある．この層に分けることを 層別という．**

層別については第2単元 §3.1(5)，§3.3(4) でくわしく学ぶ．

活用する際には欠かせない作業といって良い．そのときに，その問題領域の専門知識や経験は不可欠であり，§1.5 統計の利用とその目的，有用性で述べたことをもう一度確認してほしい．

今日は2つの例で説明した．理解をより深めるために，明日は別な分野の例を取り上げる．

2.5　調査・実験を始める前に(2)

(1)　調査の場合には

企業が調査や実験を行う場合は，その結果として得られた統計をもとに，なんらかの意思決定を行うことが多い．その意思決定にはどのような情報が必要なのかを明確にし，その情報を確実に収集できるように，調査や実験を計画しなければならない．

§1.3 の市場調査＜例6＞では，半年間にわたって行われたインスタントコーヒーの嗜好テストの例を挙げた．第１回の結果から，現行品を改良しなければならないことが明らかになった．このときあなたが製品開発者であったらならば，消費者がA銘柄よりB銘柄を選んだのはなぜかを具体的に理解し，速やかに改良の方針を決めなくてはならない．もしも調査結果からこの行動の方針が決められないのであれば，調査が十分に活かされたとはいえない．開発者は全くの手探りで試行錯誤しなければならなく，これでは調査しなかった場合と同じであるからである．

そこで，調査時に「なぜ，A銘柄よりもB銘柄が良かったのですか」と質問することを考えてみよう.「Bのほうがおいしかったから」と答える人が多いかもしれない．このようなややあいまいな印象を述べた回答であっても，製品開発者が「味や香り」が劣ったのか,「溶けやすさ」に問題があったのか，という視点（仮説）を持っていれば，多少は情報として役立つだろう．しかし，味覚

を改善する手がかりを得るには,「どのようにおいしいのか，まずいのか」をより具体的にするための適切な「枠組み」を定めておく必要がある．つまり，全体的な印象だけではなく，酸味・苦み，飲み口，のどごし，などの評価の視点を用意し，回答してもらう必要がある．もちろん，どのような「視点」,「枠組み」を定めるのかは大変に難しく，このためには，消費者のコーヒーに対する嗜好や識別力について，専門知識が必要になる．

§1.5で述べたように，統計学者デミングは「統計は行動のためにある」といっているが，調査を始める場合には，その後の活用についてあらかじめ見通して，計画を立てておく必要がある．調査や実験の結果から具体的な行動の指針を得るためには，起こっている現象の表層を知るだけでは役に立たない．なぜそのような現象がおこっているのか，その要因（原因）を明らかにすることによって，原因を解決するという行動の指針を決めることができるからである．

(2)　固有技術の重要性

同じ設計図の紙ヘリコプターを，紙の長い辺に沿って切ったものと短い辺に沿って切ったものとの2機を作って，滞空時間を測定して見る．紙ヘリコプターの全長が15cm以上のときは，滞空時間に違いが出てくるであろう．

これは，紙には縦横があり，張り（専門用語では剛度）に大きな差があるためである．表示2.9で，紙の硬さの先に紙の方向を入れてあれば，切る方向を考慮してより滞空時間の長い紙ヘリコプターを設計できるであろう．

この簡単な例で説明したように，調査・実験の計画だけでなく，結果の解析の段階でも，対象に対する固有技術の有無は成果に大きな影響を与える．

固有技術と統計技術を融合することにより，はじめて，統計の有用性を発揮することができる．

数理統計の専門家の中には「データは虚心に見るべきであり，先入観を持ってはいけない」という人がいる．しかし，統計の実務家は経験的な知識を持っており，完全に虚心ということは不可能であろう．

しかし，因果関連図を作成するに際して,「これが唯一正しいモデルである」

という確信を持つのは好ましくない.「これしかない」というモデルではなく，複数の考えられるモデルを考え，現実のデータでどのモデルが妥当であるかを検証するという謙虚な気持ちを持つ必要があるであろう.

自然科学は，このような過程を繰り返して進歩してきたものである.

■　**演習 4**　§1.4 データを取って見よう で取り上げた観測値について因果関連図を作成せよ.（因果関連図を作成する前に，§2.4(3) に示すように，個々の因果関係を列挙することにより，分かりやすい因果関連図が作りやすくなるであろう.）■

本日のまとめ

アンケート調査という全くことなる分野でも，問題を事前に整理することの重要性を理解していただけたであろう.

最後の演習に真剣に取り組んで，問題を整理することが，この章の内容を自分のものとし，将来の実務に役立つためには必須である.

3　統計解析の基礎知識

3.1　期待値・分散・標準偏差 (1)

(1)　期待値

いま，表示3.1 に示された当たりが入っている宝くじを1本持っている．この1本にいくらの価値があると考えたら良いであろうか.

表示3.1: 宝くじの賞金と本数

行番号		賞金 (円)	本数 (本)
1	1等	10,000	10
2	2等	1,000	200
3	3等	100	4,000
4	はずれ	0	5,790
	合計		10,000

全部の宝くじを買い占めたとすると，賞金総額 T は

$$\begin{aligned} T &= 10000 \times 10 + 1000 \times 200 + 100 \times 4000 + 0 \times 5790 \\ &= 100000 + 200000 + 400000 \\ &= 700000\text{円} \end{aligned}$$

となり，宝くじ総数 N は10,000 本であるから，宝くじ1本当たりの賞金額は $\dfrac{T}{N} = \dfrac{700000}{10000} = 70$ 円　となる．抽選前の宝くじは，1等に当たるかもしれないし，紙くずになるかもしれない．しかし，平均的には70円の賞金が期待される．これを,「この宝くじ1本の 期待値 (Expectation) は 70 円である」という.

これを一般式で表わしてみよう．i 等の賞金額を x_i，本数を n_i とする．ここで添え字の i は表示3.1 の行番号に対応し，1 から m （この例では $m = 4$）まで変化する.

総本数 $N = 10,000$**本 と賞金総額** $T = 700,000$**円 は**

$$N = n_1 + n_2 + n_3 + n_4$$
$$T = x_1 n_1 + x_2 n_2 + x_3 n_3 + x_4 n_4$$

として求められるが，これを次のように表わす.

$$N = \sum_{i=1}^{m} n_i$$
$$T = \sum_{i=1}^{m} x_i n_i$$

ここに $\sum_{i=1}^{m}$ **は，その後に続く式**(n_i **または** $n_i x_i$) **の** i **を** $1, 2, \cdots, m$ **と変えて加えることを表わす記号である（**$\sum$ **記号の使い方については** §3.5 **で補足する）.**

1本の宝くじの期待値を $\mathrm{E}[x]$ **で表わすことにすると,**

$$\mathrm{E}[x] = \frac{\sum_{i=1}^{m} x_i n_i}{\sum_{i=1}^{m} n_i} = 70$$

となる．ここで，分母を N **で置き換えると，この式は**

$$\mathrm{E}[x] = \frac{\sum_{i=1}^{m} x_i n_i}{\sum_{i=1}^{m} n_i} = \frac{\sum_{i=1}^{m} x_i n_i}{N} = \sum_{i=1}^{m} \left(x_i \frac{n_i}{N} \right)$$

と書き換えることができる[1]**.**

$\frac{n_i}{N}$ **は，**$i = 1$ **のとき，**$\frac{n_1}{N} = \frac{10}{10000} = 0.001$ **で，**1000**回に**1**回，すなわち，** 0.1% **の割合で**1**等に当選することを示している．**$\frac{n_i}{N}$ **は，**i **等の 確率**

[1] $x_i n_i$ **の合計を** N **で割る代わりに，個々の積** $x_i n_i$ **を** N **で割ってから合計することができる.**

(Probability) と呼ばれる．この確率を π_i で表わすと[2]，$\mathrm{E}[x]$ は

$$\begin{aligned}\mathrm{E}[x] &= \sum_{i=1}^{m} x_i \pi_i \\ &= 10000 \times \frac{10}{10000} + 1000 \times \frac{200}{10000} + 100 \times \frac{4000}{10000} + 0 \times \frac{5790}{10000} \\ &= 10 + 20 + 40 + 0 \\ &= 70\end{aligned}$$

と求めることができる．

確率 π_i を合計すると，

$$\sum_{i=1}^{m} \pi_i = \sum_{i=1}^{m} \frac{n_i}{N} = \frac{\sum_{i=1}^{m} n_i}{N} = \frac{N}{N} = 1$$

となることが分かる．

ここまでに出てきた計算を表にまとめると，表示3.2のようになる．このような計算は Excel が得意とするものである．一度，賞金と本数を入力したら，それ以外の数値が自動的に求められる下のような計算表を作ってみると良い．

表示3.2の上は，計算表に表示された状態を表わし，下は，各セルに記録されている文字，数字，または計算式を表わす．

C7 のセルに，本数の合計を求めるために，=SUM(C3:C6) を入力する．括弧の中は合計する範囲である．

D3 のセルに，賞金 x_1 と本数 n_1 の積を求めるために，=B3*C3 を入力する．これを下の D4:D6 にコピーすると，行番号の 3 が自動的に4, 5, 6 に変更される．

E3 のセルに，1等が当たる確率を求める．=C3/C7 とすれば求められるが，2等以下の当たる確率を求めるために，E4 にコピーすると，行番号が自動的に変化し =C4/C8 になってしまう．分母は自動的変化の対象から外すために，行番号の前に$ をつけ，=C3/C$7 を入力する．これを下にコピーする．

[2] 通常 π は円周率を表わす記号として用いられるが，ここでは確率を表わす記号として用いる．ギリシャ文字の π（パイ） はアルファベットの p に対応している．

表示3.2: **期待値の計算**(1)

	A	B	C	D	E	F
1	等	賞金	本数		確率	
2	i	xi	ni	xini	πi	xiπi
3	1	10,000	10	100,000	0.0010	10
4	2	1,000	200	200,000	0.0200	20
5	3	100	4,000	400,000	0.4000	40
6	はずれ	0	5,790	0	0.5790	0
7	合計		10,000	700,000	1.0000	70

	A	B	C	D	E	F
1	等	賞金	本数		確率	
2	i	xi	ni	xini	πi	xiπi
3	1	10000	10	=B3*C3	=C3/C$7	=B3*E3
4	2	1000	200	=B4*C4	=C4/C$7	=B4*E4
5	3	100	4000	=B5*C5	=C5/C$7	=B5*E5
6	はずれ	0	5790	=B6*C6	=C6/C$7	=B6*E6
7	合計		=SUM(C3:C6)	=SUM(D3:D6)	=SUM(E3:E6)	=SUM(F3:F6)

F 列はD 列と同様の手順で求められる.

C7 の合計を求める式を右にコピーすると，列名が自動的に変化し各列の合計が求められる.

もし，1 等賞金が10 倍の100,000 円で本数が1/10 **の1 本になり，残りの9 本がはずれとなったならば，どうなるであろうか. 表示**3.3 **に期待値の計算結果を示す. これより，期待値は同じであることが分かる.**

表示3.3: **期待値の計算**(2)

	A	B	C	D	E	F
9	等	賞金	本数		確率	
10	i	xi	ni	xini	πi	xiπi
11	1	100,000	1	100,000	0.0001	10
12	2	1,000	200	200,000	0.0200	20
13	3	100	4,000	400,000	0.4000	40
14	はずれ	0	5,799	0	0.5799	0
15	合計		10,000	700,000	1.0000	70

このように，変数 x **の取り得る個々の値** x_i **に対して確率** π_i **が対応づけられ**

るとき，その対応関係を x **の 確率分布**(Probability Distribution)**といい，変数** x **を 確率変数**(Random Variable)**という．このとき，** $\sum_{i=1}^{m} x_i\pi_i$ **を確率変数** x **の期待値** $\mathrm{E}[x]$ **と定義する．**

(2)　分散

表示3.2 **と表示**3.3 **の**2**つの宝くじの違いは，前の宝くじに比べて後の宝くじは当たり外れの差が大きい，すなわちバラツキが大きいということである．これを定量的に表わす方法を考える．期待値に対してどれだけ得をしたか，あるいは損をしたかを 偏差** (Deviation)**という．偏差の平均的大きさを表わすために偏差の期待値を用いることが考えられるので，前の宝くじで偏差の期待値を求めてみよう．** i **等の偏差を** $d_i = x_i - \mathrm{E}[x]$ **とすると，変数** d **の確率分布は，表示**3.4 **の** d_i, π_i **の列となる．したがって，** d **の期待値は** $d_i\pi_i$ **列の合計から分かるように** 0 **になる．**

$$\begin{aligned}\mathrm{E}[d] &= \sum_{i=1}^{m} d_i\pi_i \\ &= 9930 \times 0.001 + 930 \times 0.02 + 30 \times 0.4 + (-70) \times 0.579 = 0\end{aligned}$$

表示3.4: **偏差の期待値の計算**

等	**賞金**	**本数**	**偏差**	**確率**	
i	x_i	n_i	d_i	π_i	$d_i\pi_i$
1	10,000	10	9,930	0.0010	9.93
2	1,000	200	930	0.0200	18.60
3	100	4,000	30	0.4000	12.00
はずれ	0	5,790	−70	0.5790	−40.53
合計		10,000		1.0000	0

後の宝くじでも同様に偏差の期待値は 0 **になる．これは，以下の計算から一般的にいえることである．**

$$\mathrm{E}[d] = \sum_{i=1}^{m} d_i\pi_i = \sum_{i=1}^{m}(x_i - \mathrm{E}[x])\pi_i$$

$$= \sum_{i=1}^{m} x_i \pi_i - \mathrm{E}[x] \sum_{i=1}^{m} \pi_i = \mathrm{E}[x] - \mathrm{E}[x] = 0$$

以上のことから，偏差の期待値ではバラツキの大きさを定量化できないことが分かる．これは，偏差には正負の符号があって，打ち消しあってしまうためである．そこで，偏差の2**乗の期待値を考えることにする**[3]**．これを確率変数** x **の分散**(Variance)**と呼び，**$\mathrm{V}[x]$ **で表わすことにする．分散は次式で定義される．**

$$\mathrm{V}[x] = \mathrm{E}\left[(x - \mathrm{E}[x])^2\right] = \sum_{i=1}^{m} (x_i - \mathrm{E}[x])^2 \pi_i = \sum_{i=1}^{m} d_i^2 \pi_i$$

先の2**つの宝くじについて分散を求めてみると，表示**3.5**のようにそれぞれ，**119,100 **と** 1,019,100 **と，約**8**倍となる．**

表示3.5: **分散の計算**

等	**賞金**	**本数**	**偏差**	**確率**	
i	x_i	n_i	d_i	π_i	$d_i^2 \pi_i$
1	10,000	10	9,930	0.0010	98,605
2	1,000	200	930	0.0200	17,298
3	100	4,000	30	0.4000	360
はずれ	0	5,790	−70	0.5790	2,837
合計		10,000		1.0000	119,100
1	100,000	1	99,930	0.0001	998,600
2	1,000	200	930	0.0200	17,298
3	100	4,000	30	0.4000	360
はずれ	0	5,799	−70	0.5799	2,842
合計		10,000		1.0000	1,019,100

(3) 標準偏差

分散の単位は 円^2 **である．後の宝くじのほうがバラツキが大きいことは分かるが，この単位のままでは実用にならない．もとの単位**(円)**に直すために平方根を取ったものを確率変数** x **の 標準偏差**(Standard Deviation) **と呼び，**$\mathrm{D}[x]$ **で表わすことにする．**

[3] **偏差の絶対値を取って正の値として，その期待値を求めることも考えられるが，本文の方法が望ましい性質を持っていることが，数理統計学から導かれている．**

$$\mathrm{D}[x] = \sqrt{\mathrm{V}[x]}$$

これより，2つの宝くじの標準偏差は，それぞれ，$\sqrt{119100} = 345$円と $\sqrt{1019100} = 1010$円となる．

本日のまとめ

§3から，説明に記号や数式が使われている．このような表現の形式になじんでいる人にとってはスムーズであろうが，とまどう人もあるかもしれない．そのときには自分の手元で電卓やExcelを使って計算したり，あなたなりに分かりやすい言葉で言い換えをしながら，この表現形式になれてほしい．

テキストで例として取り上げられている宝くじを1枚購入すると，裏面に賞金と当たり本数が書かれている．実際の数値を使って平均と標準偏差を計算することは，理解を助けるのに役立つであろう．

ものごとの起こり方を確率変数・確率分布としてとらえる考え方や，期待値と分散・標準偏差の性質は，以降の学習の基本であるので，よく納得できるように学習することが必要である．

今日の学習で理解が不十分な人も，明日 サイコロで 同様の実験をするので，そこで「なるほど，そーか」とひざをたたいてもらえるであろう．

3.2　期待値・分散・標準偏差(2)

(1)　サイコロの目の数の期待値と分散

サイコロを投げて，出た目の数だけの賞金が得られるものとする．賞金の期待値と分散，標準偏差はいくらになるだろうか．

宝くじの場合は，全部の宝くじを買い占めたとして，1枚当たりの賞金額（期待値）を計算することができた．そこでは，宝くじ N 本の内 i 等当選が n_i 本

であるとき， i **等の当選率** $\pi_i = n_i/N$ **を使って，期待値や分散を導いた．**

サイコロは無限回投げられるので，買占めに相当することはできない．そこで，確率 π_i **を直接使って期待値などを導くことにする．**

出た目の数を x **とすると，** x **の取り得る値は** $1, 2, 3, 4, 5, 6$ **で，それぞれの得られる確率** π **は，サイコロが正しく作られているのならば，すべて等しく** $1/6$ **である．**

表示3.6: **サイコロの目の期待値と分散**

i	x_i	π_i	$x_i\pi_i$	d_i	$d_i^2\pi_i$
1	1	1/6	1/6	–2.5	6.25/6
2	2	1/6	2/6	–1.5	2.25/6
3	3	1/6	3/6	–0.5	0.25/6
4	4	1/6	4/6	0.5	0.25/6
5	5	1/6	5/6	1.5	2.25/6
6	6	1/6	6/6	2.5	6.25/6
合計		1.000	21/6	0.0	35/12

表示3.6 **の計算結果から，**

$$\mathrm{E}[x] = \frac{21}{6} = 3.5, \quad \mathrm{V}[x] = \frac{35}{12}, \quad \mathrm{D}[x] = \sqrt{\frac{35}{12}} = 1.708$$

であることが分かる．

■ **演習5　四角の棒の4つの面に，**1,2,3,4 **という数字を書く．この棒を転がして上の面の数字を読む．棒が正しい四角に出来ていれば，**4**つの数字の出る確率はすべて** $1/4$ **になるであろう．**

数字の期待値と標準偏差を計算せよ． ■

(2)　2個のサイコロの目の合計の期待値と分散

2個のサイコロを振ったとき，それらの目の合計を T_2 **とする** $(T_2 = x_1 + x_2)$**．** T_2 **の確率分布を求める．**

$T_2 = 2$ **となるのは** $x_1 = 1,\ x_2 = 1$ **の組合わせの**1**通りである．**

$T_2 = 4$ **は** $(x_1, x_2) = (1,3), (2,2), (3,1)$ **の**3 **通りである．**

この関係を拡張して整理したのが 表示3.7 である.

表示3.7: $x_1,\ x_2$ の組合わせに対する T_2

$x_1 \backslash x_2$	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

これから，T_2 が 2,3,...,12 となる組合わせの個数が求められる.

全部の組合わせは36通りであるから，それぞれの組合わせの個数を36で割ると，T_2 の確率分布が 表示3.8の π_i の列のように求められる.

表示3.8: T_2 の期待値と分散

i	T_{2i}	π_i	$T_{2i}\pi_i$	d_i	$d_i^2\pi_i$
1	2	1/36	2/36	–5.0	25/36
2	3	2/36	6/36	–4.0	32/36
3	4	3/36	12/36	–3.0	27/36
4	5	4/36	20/36	–2.0	16/36
5	6	5/36	30/36	–1.0	5/36
6	7	6/36	42/36	0.0	0/36
7	8	5/36	40/36	1.0	5/36
8	9	4/36	36/36	2.0	16/36
9	10	3/36	30/36	3.0	27/36
10	11	2/36	22/36	4.0	32/36
11	12	1/36	12/36	5.0	25/36
合計		1.0000	7		35/6

期待値と分散，標準偏差を求める．表示3.8に，T_2 の期待値と分散が計算されている．これから，

$$\mathrm{E}[T_2] = 7.0, \quad \mathrm{V}[T_2] = \frac{35}{6}, \quad \mathrm{D}[T_2] = \sqrt{\frac{35}{6}} = 2.415$$

であることが分かる.

ここに得られた T_2 の期待値 7.0 を x の期待値 3.5 と比べてみると，ちょうど2倍になっていることが分かる．これは，極めて常識的な結果である．

それでは，分散・標準偏差ではどのような関係があるであろうか？

T_2 の分散 $35/6$ はx の 分散 $35/12$ の2倍である．分散を分数で表わしているのはこの関係がはっきりと分かるようにするためであった．

標準偏差については2倍になるという関係はない．標準偏差は分散の平方根であるから，標準偏差は $\sqrt{2}$ 倍になる．

■　演習６　前の演習で用いた四角の棒を2回転がして，数字の合計を求めたとき，その期待値と標準偏差はいくらになるか．

この場合も，期待値と分散には，サイコロの目の場合と同様の関係があるかを確かめよ．■

本日のまとめ

昨日のまとめに予告したように，今日は サイコロを使って期待値や分散について学んだ．これは，昨日の復習を兼ねると共に，2個のサイコロの結果は，明日取り上げる分散の加法性の準備である．

今日はテキストのページ数は少ない．演習問題を解くことにより，理解を確実にすることを目指して努力してほしい．

演習問題が短時間で解けてしまった受講生は，四角の棒の数字を 1,2,3,4 の代わりに，1,2,4,8 としたらどうなるか，3回投げたらどうなるか，に挑戦してほしい．

3.3　分散の加法性

(1)　基本公式

前項では，T_2 の確率分布を求め，期待値と分散，標準偏差を求めた．しかし，以下の期待値と分散に関する性質を用いれば確率分布を求めなくても，それぞれの期待値と分散を求めることができる．

【性質1】　x_1 と x_2 を互いに独立に分布する2つの確率変数とするとき，和

$$y = x_1 + x_2$$

の期待値と分散は次式で与えられる．

$$\begin{aligned} \mathrm{E}[y] &= \mathrm{E}[x_1 + x_2] = \mathrm{E}[x_1] + \mathrm{E}[x_2] \\ \mathrm{V}[y] &= \mathrm{V}[x_1 + x_2] = \mathrm{V}[x_1] + \mathrm{V}[x_2] \end{aligned} \tag{3.1}$$

ここで,「x_1 と x_2 を互いに独立に分布する」とは，サイコロの例でいえば，2つのサイコロの目の出方に関係がないことを意味する．2つのサイコロが短いひもで繋がれているなどの理由で両者の目の出方に関係がある場合，独立に分布するとはいわない．

【性質2】　x を確率変数，a を定数とするとき，新しい確率変数

$$y = ax$$

の期待値と標準偏差を考える．例えば，サイコロの目の数が x のとき，$y = 100x$ 円の賞金がもらえるとする．賞金の期待値と標準偏差は，目の数の期待値と標準偏差の100倍になることは自明である．すなわち，

$$\mathrm{E}[y] = \mathrm{E}[ax] = a\mathrm{E}[x], \quad \mathrm{D}[y] = \mathrm{D}[ax] = a\mathrm{D}[x] \tag{3.2}$$

で与えられる．これより，分散 $\mathrm{V}[y]$ は

$$\mathrm{V}[y] = \mathrm{D}[y]^2 = a^2\mathrm{V}[x]$$

となる．

(2) 簡単な応用

和 $y = x_1 + x_2$ の期待値と分散を求める式は前項で説明した．それでは，差 $z = x_1 - x_2$ の期待値と分散はどうなるであろうか?

期待値については，

$$\mathrm{E}[z] = \mathrm{E}[x_1 - x_2] = \mathrm{E}[x_1] - \mathrm{E}[x_2]$$

となることは，直感的に分かるであろう．

それでは，分散はどうなるであろうか?

$$\mathrm{V}[z] = \mathrm{V}[x_1 - x_2] = \mathrm{V}[x_1] - \mathrm{V}[x_2]$$

と考えた方もおられるであろう．

この式が間違いであることは，$\mathrm{V}[x_1] < \mathrm{V}[x_2]$ のとき，差 z の分散が負になることから，理解できる．

正しい式は次のように考えると導き出すことができる．

$$z = x_1 - x_2 = x_1 + (-x_2)$$

と考え，z の分散は x_1 の分散と $-x_2$ の分散の和であるとする．$-x_2$ の分散は，【性質2】で $a = -1$ とすることにより求められ，

$$\mathrm{V}[-x_2] = \mathrm{V}[(-1) \times x_2] = (-1)^2 \mathrm{V}[x_2] = \mathrm{V}[x_2]$$

となる．

これから，次式が導かれる．

$$\mathrm{V}[z] = \mathrm{V}[x_1 - x_2] = \mathrm{V}[x_1] + \mathrm{V}[x_2] \tag{3.3}$$

(3) 一般化公式

上記の【性質1】と【性質2】を繰り返し適用することによって，以下の性質が導かれる．

【性質3】 $x_1,\ x_2,\ \cdots,\ x_n$ **という n 個の確率変数が互いに独立に分布するとき，その1次式，**

$$y = a_1x_1 + a_2x_2 + \cdots + a_nx_n = \sum_{i=1}^{n} a_ix_i$$

で表わされる確率変数 y の期待値と分散は

$$\begin{aligned}
\mathrm{E}[y] &= \mathrm{E}[a_1x_1 + a_2x_2 + \cdots + a_nx_n] \\
&= a_1\mathrm{E}[x_1] + a_2\mathrm{E}[x_2] + \cdots + a_n\mathrm{E}[x_n] = \sum_{i=1}^{n} a_i\mathrm{E}[x_i] \\
\mathrm{V}[y] &= \mathrm{V}[a_1x_1 + a_2x_2 + \cdots + a_nx_n] \\
&= a_1^2\mathrm{V}[x_1] + a_2^2\mathrm{V}[x_2] + \cdots + a_n^2\mathrm{V}[x_n] = \sum_{i=1}^{n} a_i^2\mathrm{V}[x_i] \qquad (3.4)
\end{aligned}$$

で与えられる.

これより，3 個，4 個，6 個のサイコロの目の合計，$T_3,\ T_4,\ T_6$ それぞれの期待値と分散は

$$\mathrm{E}[T_3] = 3 \times 3.5 = 10.5, \quad \mathrm{V}[T_3] = 3 \times \frac{35}{12} = \frac{35}{4}, \quad \mathrm{D}[T_3] = \sqrt{\frac{35}{4}} = 2.958$$
$$\mathrm{E}[T_4] = 4 \times 3.5 = 14.0, \quad \mathrm{V}[T_4] = 4 \times \frac{35}{12} = \frac{35}{3}, \quad \mathrm{D}[T_4] = \sqrt{\frac{35}{3}} = 3.416$$
$$\mathrm{E}[T_6] = 6 \times 3.5 = 21.0, \quad \mathrm{V}[T_6] = 6 \times \frac{35}{12} = \frac{35}{2}, \quad \mathrm{D}[T_6] = \sqrt{\frac{35}{2}} = 4.183$$

となることが分かる.

【性質 1】,【性質 2】,【性質 3】を 誤差法則 という．また，分散 $\mathrm{V}[\ ]$ についての関係を 分散の加法性 という.

(4)　合計と平均の分散・標準偏差

$x_1,\ x_2.\ \cdots,\ x_n$ は，期待値と分散が同じ分布から独立に得られた確率変数とする．合計 $T = x_1 + x_2 + \ \cdots\ + x_n$ と平均 $\overline{x} = \dfrac{T}{n}$ の分散と標準偏差はどうなるのだろうか．x_i の分散を σ^2 で表わすことにする.

合計の分散と標準偏差は,【性質 3】で $a_1 = \cdots = a_n = 1$ とすると,

$$\mathrm{V}[T] = (\underbrace{1^2 + \cdots + 1^2}_{n\text{個}})\sigma^2 = n\sigma^2$$
$$\mathrm{D}[T] = = \sqrt{n}\sigma$$

となる.

平均の分散と標準偏差は,【性質2】で $a = 1/n$ **とすると,**

$$\mathrm{V}[\overline{x}] = \left(\frac{1}{n}\right)^2 \mathrm{V}[T] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\mathrm{D}[\overline{x}] = \frac{\sigma}{\sqrt{n}}$$

となる.

それぞれの標準偏差を求める公式は, 極めて重要であり, 記憶に止めてほしい.

■ **演習7** **男の体重** x **の期待値が** $62\mathrm{kg}$**, 標準偏差が** $8\mathrm{kg}$ **で, 女の体重** y **の期待値が** $47\mathrm{kg}$**, 標準偏差が** $6\mathrm{kg}$ **であると仮定する. 3人の男と2人の女がエレベータに乗った. 合計体重の期待値と標準偏差を求めよ.**

(ヒント) 男の体重を x**, 女の体重を** y **とするとき, 合計** T **を表わす式として,** $T = 3x + 2y$ **と** $T = x_1 + x_2 + x_3 + y_1 + y_2$ **のいずれが適当かを考えよ.** ■

■ **演習8** **紙の厚さの平均が** 10 **ミクロン, 標準偏差が** 1 **ミクロンであるとする. ただし, これは紙1枚1枚の間のバラツキで, 1枚の紙の中でのバラツキは0であるとする. 1枚の紙を4回2つに折り, 16枚(32ページ) とする. これを8折集めて256ページの本が出来上がる. 本の厚さの標準偏差はいくらか.** ■

本日のまとめ

昨日のサイコロの実験で得られた結果を一般化した「分散の加法性」について学んだ.

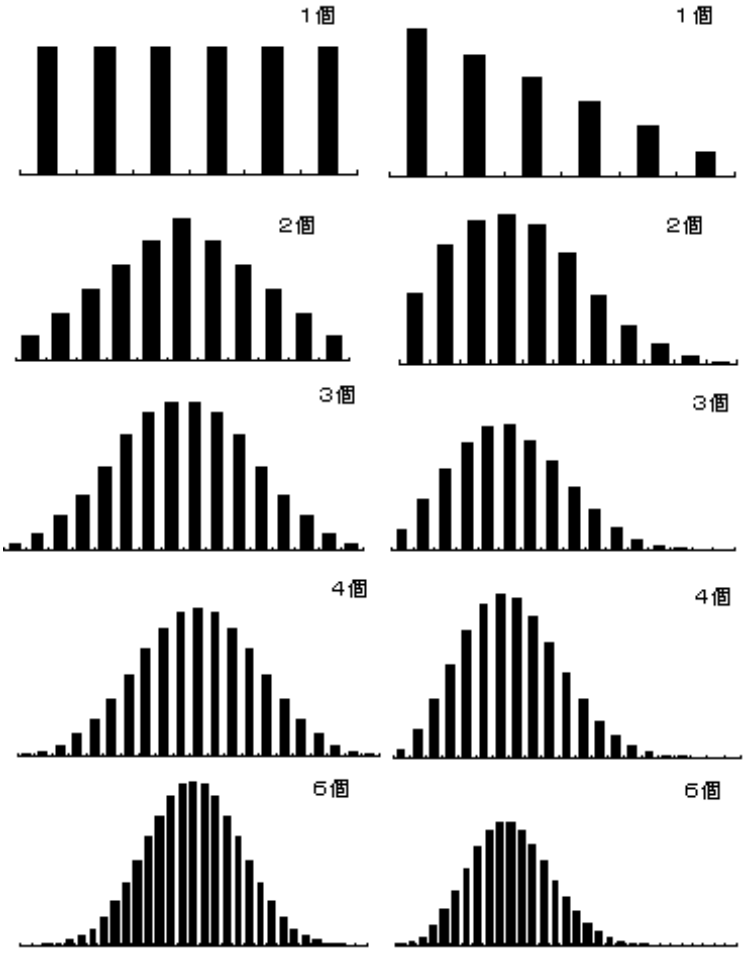
分散の加法性 は今後表われる様々な公式の基礎をなすものであるから, 完全に理解されることが望まれる. 2題の演習を解くことは, 理解を確かめるために役立つであろう. 完全に理解できなくても, 先に進むのに障害はない. 今後いくつかの適用例を見ることにより, この法則を利用することができるようになるであろう.

3.4　中心極限定理と正規分布

(1)　平均値の分布

サイコロの目の合計 $T_2,\ T_3,\ T_4,\ T_6$ をサイコロの個数で割って求めた目の平均を，それぞれ，$\overline{x}_2,\ \overline{x}_3,\ \overline{x}_4,\ \overline{x}_6$ とする．これらの確率分布を求め，グラフ化したものが表示3.9 (左)**である．**

表示3.9: **サイコロの目の平均の分布**



一番上の x の分布は矩形をしているのに対し，$\overline{x}_2$ の分布は三角形をしてい

る．さらに，サイコロの個数が増すにつれて，山が丸みをおび，裾が伸び，つりがね型に近づくことが分かる．ところで，表示3.9 **のグラフの横軸の値として，棒と棒の間にある値は考えられないので，離散量である．また，棒グラフの棒の高さは，各平均の確率である．振るサイコロの数を増やしていくと，**1**や**2**に近い小さい値や**5**や**6 **に近い大きい値の確率は極めて小さくなり，実際問題として無視できるようになる．また，棒と棒の間も狭くなり，連続した面を形作るようになるだろう．サイコロの数を限りなく多くしていくと，分布の極限として連続量の分布として扱うことができる（離散量の分布を近似して連続分布として扱う内容は，第**3 **単元の** §3.2 2 **項分布の正規近似として再び取り上げる）．この極限の連続分布は** 正規分布(Normal Distribution)**，または** ガウス分布 **と呼ばれる．**

(2)　中心極限定理

x **の分布がどんな形をしていても，それをたくさん集めた合計または平均の分布は正規分布に近づくという性質がある．**

例として，$x = 1 \sim 6$ **の確率が** $\frac{6}{21}, \frac{5}{21}, \frac{4}{21}, \frac{3}{21}, \frac{2}{21}, \frac{1}{21}$ **という歪んだサイコロを考えよう．このサイコロの目の数の分布は，表示**3.9 **の右上に示すように左右が非対称の**3 **角形となる．**

このようなサイコロを2 **個投げてその平均を取ると，上から**2 **番目の図のように，山が丸くなる．さらに個数を増やすと，左右対称に近づくことが分かる．**

このように，もとの分布が左右非対称であっても，平均する個数が増えると左右対称な分布，最終的には正規分布に近づく．この性質を 中心極限定理 **という．この性質があるために，正規分布は統計において重要な位置を占める．**

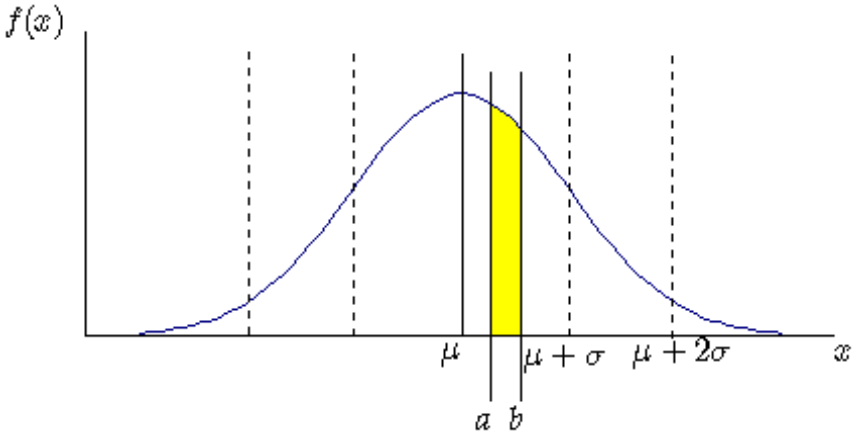
■ **演習9**　Excel **ファイルで，表示**3.9 **のグラフのシートの**B7:B12 **のセルに，**1 **から**6 **までの目が出る確率の比率を入力すると，それに連動してグラフが変化する．表示**3.9 **は** 1,1,1,1,1,1 **または** 6,5,4,3,2,1 **を入力した結果である．**

3 **の目が出ないサイコロ（**1,1,0,1,1,1**）などを入力して，グラフの変化を観察せよ．** ■

(3)　正規分布

確率変数 x の期待値 $\mathrm{E}[x]$ を μ，標準偏差 $\mathrm{D}[x]$ を σ で表わすとき，表示3.10のようなつりがね型の分布を期待値 μ，標準偏差 σ の正規分布という．

表示 3.10: 正規分布の一般型



表示3.10において，横軸の値 x に対する縦軸の値 $f(x)$ は，次の式で与えられる．

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

ここで，π は円周率 $3.14159\cdots$ であり，e は自然対数の底 $2.71828\cdots$ である[4]．

正規分布の期待値 μ と標準偏差 σ の値を与えれば確率は上の式で決まる．μ, σ を正規分布の母数（パラメータ）といい，x が 期待値 μ, 標準偏差 σ の正規分布に従うということを

$$x \sim \mathrm{N}(\mu, \sigma^2)$$

[4] **この式がどのようにして導かれたかについての説明は，実務家にとって必要ではないと思われるので割愛する．データの解析の中で，この式を使って自ら計算することはない．したがって，この式を記憶しなくても問題は生じない．**

という記号で表現することが多い．標準偏差 σ **ではなくその** 2 **乗の分散** σ^2 $(= \mathrm{V}[x])$ **が用いられるのは，数理統計学の分野では分散がバラツキの大きさを表わす基本的な量とされているからである．**

正規分布のように横軸の数値が連続的に無限にあると考えられる連続分布の場合，ある特定の値 を取る確率は 0 **になるので，その代わりにある特定の区間に入る確率を定義する．表示**3.10**でいえば，** x **が区間** $(a,\ b)$ **に入る確率** $\Pr(a < x < b)$ **は，** $x = a$**，** $x = b$ **と曲線** $f(x)$ **で囲まれた部分の面積（塗りつぶされた面積）として与えられる．つまり，縦軸の値** $f(x)$ **は確率そのものを表わしているわけではなく，面積としてはじめて確率を表わすことになるので確率密度関数と呼ぶ．**

正規分布を表わす曲線は次の諸性質をもつ．

1) **期待値** μ **を中心にして左右対称である．**

2) **曲線は期待値** μ **の近傍で高く，両側に行くに従って低くなる．**

3) **期待値** μ **は曲線の位置を定める．期待値** μ **のみ異なる** 2 **つの曲線は，左右に動かすことにより重ねることができる．**

4) **標準偏差** σ **は曲線の形を定める．** σ **が大きければ曲線は扁平に，** σ **が小さければ狭く高くなる．いかなる場合にも** $\mu - \sigma$ **と** $\mu + \sigma$ **における曲線上の点は変曲点となる．すなわち，**2 **つの変曲点の間では曲線は上に凸で，その外側では曲線は上に凹になっている．**

5a) $\mu \pm \sigma$ **の間の正規曲線下の面積は，全面積の約** 68% **（約**2/3**）である．**

5b) $\mu \pm 2\sigma$ **の間の正規曲線下の面積は，全面積の約** 95% **である．**

5c) $\mu \pm 3\sigma$ **の間の正規曲線下の面積は，全面積の約** 99.7% **である．**

x **が期待値** μ **から標準偏差** σ **の何倍離れているのかを表わすのに，偏差値**

$$z = \frac{x - \mu}{\sigma}$$

が用いられる[5]**．**

[5] **一般社会で用いられている 偏差値 は，上の式で計算された値** z **を** 10 **倍して** 50 **を加えたものである．**

なお，z は標準化得点，標準化偏差 などと呼ばれることがある．z は，期待値が 0，分散が $1^2 = 1$ の正規分布

$$z \sim \mathrm{N}(0,\ 1^2)$$

に従う．この分布を 標準正規分布 という．

標準正規分布で z がある値よりも大きい確率を表にしたものが正規分布表である．

基本的な数値を表示3.11 に示す．

表示 3.11: 正規分布表

z	0.5	1.0	1.5	2.0	2.5	3.0
上側確率	0.309	0.159	0.067	0.023	0.006	0.001

上側確率	0.25	0.10	0.05	0.025	0.01	0.005
z	0.674	1.282	**1.645**	**1.960**	2.326	2.576

上の表で，$z = 2.0$ に対する上側確率の値が 0.023 となっている．0.023 は，正規分布で，期待値$+2.0\times$標準偏差 よりも大きい確率である．正規分布は左右対称であるから，期待値$-2.0\times$標準偏差 よりも小さい確率も 0.023 である．したがって，期待値$\pm 2.0\times$標準偏差 の範囲内の確率は $1-2\times 0.023 = 1-0.046 = 0.954 \approx 0.95$ となる．これは，正規分布の性質 5b) に対応している．

下の表で太字で表わされている 1.645 と 1.960 は第3，第4 単元でしばしば出てくる数値である．1.960 は，期待値 $+\ 1.960 \times$ 標準偏差 よりも大きい確率が 0.025 で，期待値 $\pm\ 1.960 \times$ 標準偏差 の範囲外の確率は，$0.025 \times 2 = 0.05$ となる．

表示3.11 はExcel で作成したものである．

z から下側確率を求める関数が `=NORMSDIST(z)` である．上側確率は `=NORMSDIST(-z)` または `=1-NORMSDIST(z)` とすれば求められる．

下側確率から z を求める関数が `=NORMSINV(`下側確率`)` である．上側確率から z を求めたいときは，`=NORMSINV(1-`上側確率`)` とすれば良い．

■ **演習10** 確率変数 x が $\mathrm{N}(10,\ 2^2)$ に従うとき，次の確率を求めなさい．

(1) $x > 12$ (2) $x > 11$ (3) $x < 9$

(4) $9 < x < 12$ (5) $x > 15$ ■

■ **演習11** 演習7 で，総重量が 300kg を越すとブザーがなるエレベータとしたときに，ブザーがなる確率を求めよ． ■

(4) 統計的方法の頑健性

中心極限定理がいかに有力な武器であっても，標本の大きさ n が小さいときには適用できないのではないかと，心配する人もいるかもしれない．いろいろな分布について，n が小さいときの統計的推論の方法が研究されている．しかしながら，正規分布の仮定の下で行われる統計的推論は,「その前提となっている正規分布というモデルに対して頑健性をもつ」という，もう一つの有用な性質が知られている.「頑健性」とは，その理論の前提となっているモデルが少々崩れても，それから導かれる推論があまり大きい影響を受けないことをいう．したがって，もとの母集団が，正規分布をするか否かをあまり気にすることなく，n がかなり小さい場合でも，正規分布から導かれる種々の統計的方法を用いて大きな誤りはないのである（もちろん，必要に応じて対数を取るなど，適当な変数変換をした上で，これらの手法を適用する方が安全な場合は多い.）

本日のまとめ

ひずんだサイコロでもたくさん平均すると正規分布に接近するというExcelによる計算結果とそのグラフを見た人は，驚きを感じ，感動した人もあるであろう.

今後の学習で最も中心的な役割を果たす 正規分布 はこのようにして導かれたものである.

データの分布が，多くの原因によってばらつきの複合したものであるとき，その分布は正規分布に近いことが，中心極限定理によって期待される.

正規分布の形は，期待値と標準偏差によって決まる．これから，ある値以上

（または以下）の確率は正規分布表または，NORMSDIST関数によって簡単に計算される．これは，今後の解析での基本であるから，演習問題で完全にマスターしてほしい．

3.5　補足

(1)　$\sum$ 記号の定義

統計解析の基礎知識を学習するにあたり，今後，$\sum$（シグマ）の記号が頻繁に登場する．$\sum$ 記号の意味は，データの総和を取った値を意味する．ここでは，代表的な$\sum$ 記号の使い方についてまとめることにする．

いま，n 個のデータの合計 T を計算することにする．これは，$T = x_1 + x_2 + \cdots + x_n$ と書き表わすことができる．このとき，$\sum$ 記号を使えば，表記が簡便になる．すなわち

$$\text{基本形：}\quad T = \sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

である．x_i の添え字 "i" を $\sum$ の上下に表示した範囲 1 から n 内で 1 つずつ動かして和を取る．例えば，x_3 から x_6 までの和ならば $\sum_{i=3}^{6} x_i = x_3 + x_4 + x_5 + x_6$ とする．

次に，2つの変数 $x,\ y$ の積の和を $\sum$記号で表現するにはどうしたら良いか．これは，$T = x_1y_1 + x_2y_2 + \cdots + x_ny_n$ であるから

$$\text{応用形}1\text{：}\quad T = \sum_{i=1}^{n} x_iy_i = x_1y_1 + x_2y_2 + \cdots + x_ny_n$$

である．$x,\ y$ の両方に同じ添え字がつくことに注意しよう．

また，1つの変数の和を計算するのであるが，この変数に定数倍する場合の$\sum$記号の使い方を考えよう．$T = ax_1 + ax_2 + \cdots + ax_n = a(x_1 + x_2 + \cdots + x_n)$ であるから，

応用形2：　$T = \sum_{i=1}^{n} ax_i = a(x_1 + x_2 + \cdots + x_n)$

である．

さらに応用例として，添え字が2つつく場合を考えよう．これは，Excel**で表形式で表されているデータの総和を計算するのだと思えば良い．具体的には，以下のようなデータ列の総和を表わす場合の $\sum$ 記号の使い方である．**

$$\begin{array}{cccc} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{array} \qquad T = \left(\begin{array}{c} x_{11} + x_{12} + \cdots + x_{1n} \\ +x_{21} + x_{22} + \cdots + x_{2n} \\ + \cdots \\ +x_{m1} + x_{m2} + \cdots + x_{mn} \end{array} \right)$$

応用形3：

$$\begin{aligned} T &= \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} \\ &= \quad x_{11} + x_{12} + \cdots + x_{1n} \\ &\quad +x_{21} + x_{22} + \cdots + x_{2n} \quad + \cdots \\ &\quad +x_{m1} + x_{m2} + \cdots + x_{mn} \end{aligned}$$

である．

4　時系列データ

4.1　一つの時系列データのグラフ化

毎月の売上高，毎日の株価，製品ロットごとの品質のように，時間的経過に従って得られているデータを 時系列データ と呼ぶ.

時間と共にデータがどのように変化するかをグラフに表わしてみると，種々の傾向を読み取ることができる.

また，複数の系列，例えば，平均株価とドルレート，製造条件の変化と製品品質の変化，などを並べてみると，系列間の関係を見ることができる.

(1)　単純な時系列のグラフ

表示4.1 は あるスーパーでの，最近の2ヵ月間の毎日の売上げ金額（グラフ化の説明のために作成した仮想データ）の一部である.

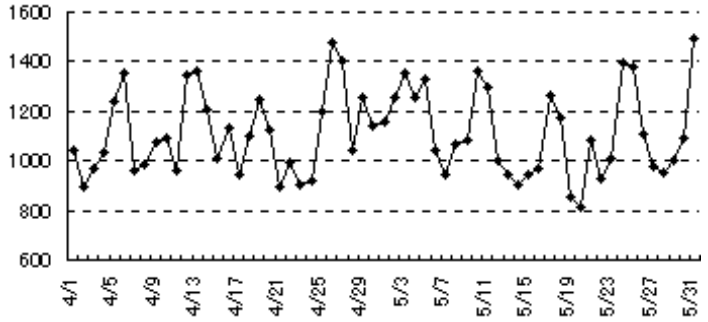
表示4.1: 時系列データ

	A	B	C	D	E	F
3			売上げ	土日祝日	移動平均	平日
4	4/1	火	1040			1040
5	4/2	水	897			897
6	4/3	木	969			969
7	4/4	金	1031		1070	1031
8	4/5	土	1239	1239	1061	
9	4/6	日	1353	1353	1087	
10	4/7	月	958		1105	958
11	4/8	火	982		1095	982
12	4/9	水	1075		1110	1075
13	4/10	木	1095		1111	1095
14	4/11	金	962		1146	962
15	4/12	土	1342	1342	1150	
16	4/13	日	1363	1363	1158	
17	4/14	月	1203		1137	1203
18	4/15	火	1008		1157	1008

D, E, F 列は後に説明する.

このデータのA列とC列から「折れ線グラフ」を描くと表示4.2が得られる[1].

表示4.2: 時系列データのグラフ



売上げ金額は日々変化している. この変化がまったくでたらめ（ランダム）であれば, このデータを取ることには意味がない.

しかし, 変化には何らかの傾向や規則性があるであろう. 例えば, 曜日によって売上げが異なり, また, 天候の悪い日は売上げが少なくなるかもしれない.

工場で毎日生産される製品からいくつかのサンプルを取り出して, その品質を測定すると, 同様のグラフが得られる. ここでは, 製造工程のいろいろな条件, 例えば, 処理温度の変化や, 大気中の湿度が影響するかもしれない.

このように, 毎時, 毎日, 毎月のような時系列データが取られたときには, まず, 表示4.2のようなグラフを描いて, 値の変化に何か規則性がないかを調べることから始める.

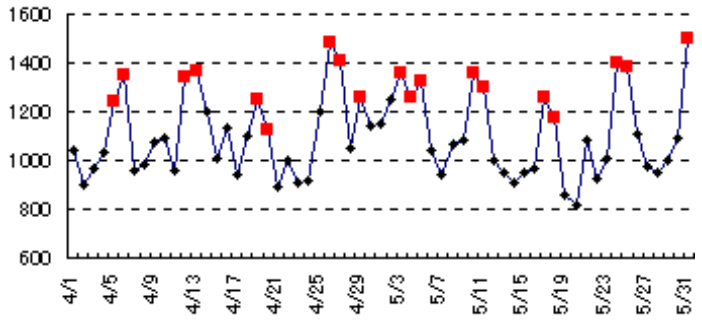
(2) 時系列データの変形

データとグラフを見ると, 土日, 祝日は売上げが多いように見える. そこで, 土日と祝日のデータがグラフで見分けられるように加工する.

[1] 最初に出力されるグラフから, 表示4.2のグラフに整形する過程は, Excelシートで説明する.

そのために，土日と祝日の売上げを D 列にコピーし，C,D 列を一つの折れ線グラフに描く．土日・祝日の折れ線グラフ点を結ぶ線を除き，四角のマークを大きくすると，表示4.3 が得られる．

表示4.3: 土日・祝日のマークを変えた時系列グラフ



Excel 画面では土日・祝日の点はマークの色が赤になっている．

これより，平日の売上げはだいたい1000程度であるが，ゴールデンウィーク期間の平日は若干高く，土日・祝日の売上げは平日に比べて平均的に高いことが分かる．

ここでは，土日・祝日を区別しなかったが，両者を区別したり，土日を分けたりしたいときには，C 列を複数の列に増やし，折れ線グラフの対象の列として指定する．さらに，各列の色やマークを変え，それを凡例に表示する．

■　演習 12　§1.4 の演習で，健康管理のためのデータを観測した受講生は，観測データを時系列グラフに描いてみよ．

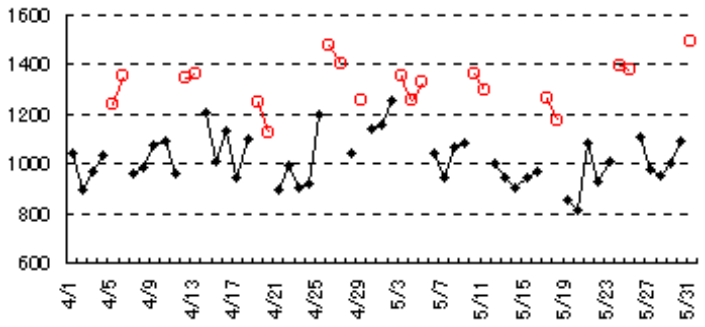
また，出勤日と休日，前日に飲酒をした日，など，健康に影響を与えると思われる生活状況の変化がグラフで観測できないかいろいろ試みよ．　■

(3)　時系列データの別の変形法

表示4.3 は土日・祝日のマークを変えたものであった．このグラフで，土日・祝日と平日とは線で結ばないグラフを作りたい．

Ｆ列に平日のみのデータを取り出す．Ｄ列とＦ列の２つの列を指定して，グラフを描くと表示4.4が得られる．

表示4.4: 平日と土日・祝日のマークを変えた時系列グラフ



ここでは，土日・祝日の点が目立つように○で表わしている．

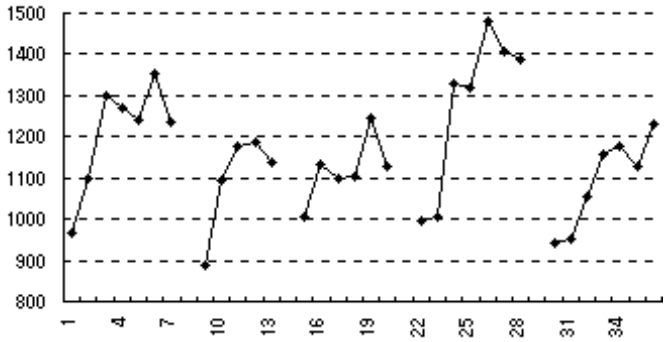
このグラフを表示4.3と比べると，平日とその他の日の違いが見やすくなったことが分かるであろう．

この方法は，一つの工程で数種類の製品を交互に生産している場合，それらを一つのグラフにまとめて表示したい場合にも利用することができる．製品によって，列を変えて折れ線グラフを描くと，色とマークを変えることができる．また，異なる製品のデータを線で結ぶこともなくなる．

また，一つの工程で数種類の製品を交互に生産している場合，同じ製品の違うロットを一つのグラフにまとめて表示したい場合は，違うロット間に1行の空行を挟むと，その間は線で結ばれなくなる．このようにすると，運転を開始した日の製品品質が不安定になるなどの現象をグラフで見ることができる．

次にその例を示す．

表示4.5: ロットの切れ目を明らかにした時系列グラフ



(4)　移動平均

個々の点のバラツキが大きく，点が上下に激しく上下するときは，変化を読み取るのがむずかしくなる．そのような場合は，移動平均が用いられる．

土日と平日では売上げに違いがあるので，その変化を除いて，売上げの動きを見たい．

1日から7日までの平均値，8日から14日までの平均値，... というように，毎週の平均値を計算して，その動きを見ることが考えられる．1週間の中の変化を消して，売上げの長期的な変化を見ることができるが，点の個数は1/7になってしまう．

それに対して，1日から7日までの平均値を4日の値，2日から8日までの平均値を5日の値，... というように，平均値を計算する方法が考えられる．こうすると，点の個数は，両側の3個＋3個=6個減少するだけになる．

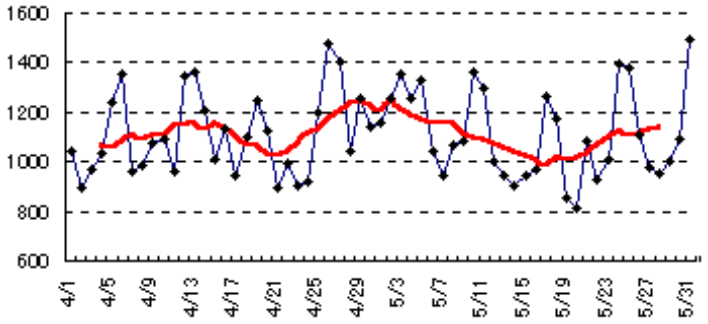
表示4.1(p.76) の E7 のセルに

```
=AVERAGE(C4:C10)
```

と入力すると，C4 から C10 までの7つのセルの平均値が求められる．AVERAGE( ) は () の中のセルの値の平均値を求める関数である．E7 のセルを下にコピーすると，平均するセルの範囲が自動的に変化して，次々と平均値が得られる．

このような平均値を 移動平均値 と呼ぶ.

表示4.6: 時系列データの移動平均グラフ



表示4.6 のように，個々の値と移動平均値を重ねてグラフ化することにより，個々の点の動きと傾向的な変化を同時に見ることができる.

移動平均を取る個数は，データに周期性があるときは，その周期を用いる．例えば，曜日によって変化するときは7日，季節変化のある月次データは12ヵ月の移動平均を取る.

周期が認められないときの個数は次の点を考慮して決める．個数を増やすと点の動きが滑らかになり，長期的な変化が良く見えるようになるが，変化が生じてからそれに気が付くまでの遅れが生じる．自分のデータについて，複数の移動平均グラフを作成して，調査の目的に合う個数を試行錯誤で決めるのが良いであろう．移動平均については，§4.4 で再び取り上げる.

本日のまとめ

時系列データでは，まず，グラフに描き，視覚的に理解することがとても大切である．変化に一定の傾向や規則性があることを発見できることがある．その期間が長ければ 移動平均 を取ってみると良いであろう.

また，§2.4 で学んだ因果関係の考え方を思い出して，仮説を立ててデータを見ることが役に立つ．§1.4 で自分で集めたデータが時系列として観察できる人は，ここで学んだ方法でグラフ化して，眺めてほしい.

4.2　複数の時系列のグラフ

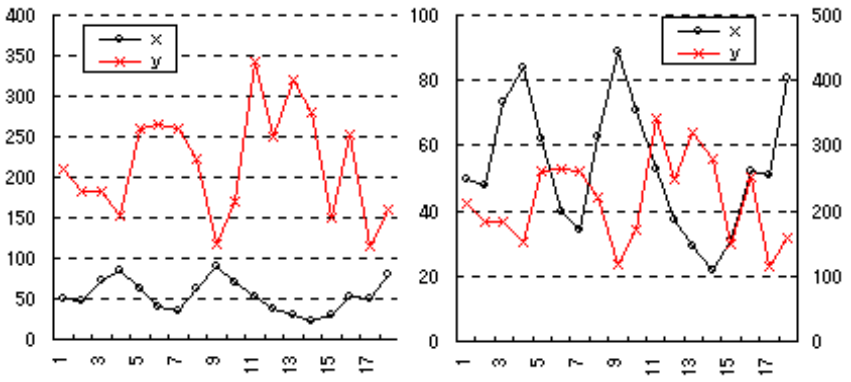
最初に説明したように，ある系列の変化を検討するときは，それに影響していると思われる系列の変化と関連付けながら，傾向と規則性を探索する必要がある．

そのためには，複数の時系列を一つにまとめたグラフを作成する必要がある．

(1)　2変数の時系列のグラフ

2つの時系列の関連を見るためには，2つの系列を一つのグラフに表示するのが良い．

表示4.7:　2つの時系列データのグラフ



値の変化の範囲が大きく異なり，一つの目盛りで表わすと，比較が困難であるときは，表示4.7のように，左右の目盛りを使ったグラフが便利である．

これから，y は x より少し遅れて変化するという傾向が見られる．

(2)　3変数以上の時系列のグラフ

表示4.8の左半分に示すように，3つ以上の時系列データを一つのグラフに集約したいとき，変化の範囲がそれぞれ異なると，表示4.7のように両側の目盛り

だけでは対応できない．また，系列の数が多いと，折れ線が重なり合って，分かりにくい．

このような場合には，グラフ化に工夫が必要である．

表示4.8: 4つの時系列データ

	x1	x2	x3	x4	z1	z2	z3	z4
1	100	42	191	159	15.9	11.5	7.4	4.3
2	94	26	181	165	15.8	11.1	7.0	4.5
3	169	48	217	179	17.1	11.6	8.5	4.9
4	202	28	218	200	17.6	11.2	8.5	5.5
5	136	70	217	157	16.5	12.1	8.5	4.2
6	70	88	204	126	15.4	12.5	7.9	3.3
7	52	30	169	149	15.1	11.2	6.6	4.0
8	139	8	187	189	16.6	10.7	7.3	5.2
9	217	70	244	184	17.9	12.1	9.5	5.0
10	163	44	213	179	17.0	11.5	8.3	4.9
11	109	128	237	119	16.0	13.3	9.2	3.1
12	61	132	223	101	15.2	13.4	8.7	2.6
13	37	62	180	128	14.8	11.9	7.0	3.4
14	16	144	214	80	14.4	13.7	8.3	2.0
15	43	34	168	144	14.9	11.3	6.5	3.9
16	106	22	183	171	16.0	11.0	7.1	4.6
17	103	154	248	104	15.9	13.9	9.7	2.7
平均	107	66	206	149				
標準偏差	58	47	25	35				

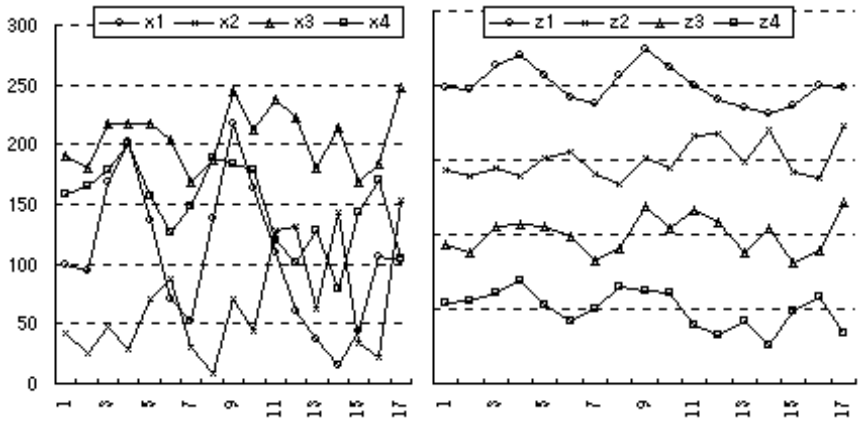
単純に重ねたグラフは表示4.9の左のように，関連を見るのがむずかしい．

それは，各変数の単位が異なり，平均値や標準偏差がまちまちであるからである．変数の単位を除き，変化の範囲を揃えるために，次のようなデータの変換をしてからグラフ化する．

j 番目の系列の i 番目の値を x_{ij}，第 j 番目の時系列の平均値と標準偏差を $\overline{x}_j$, s_j で表わす．また，系列の個数を J とする．

- 表示4.8の下に示すように，まず，各時系列の平均値と標準偏差を計算する．
- x_{ij} から平均値 $\overline{x}_j$ を引いて標準偏差 s_j で割って，基準化した偏差 z を求める．

表示 4.9: 4つの時系列のグラフ



$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}$$

o　グラフが重ならないように，j 番目の系列の z_{ij} に$4(J+1-j)$ を加える．4 は系列間の間隔で，自由に変更することができる．

このような手順で求めたのが，表示4.8の右半分 (z) である．これから表示4.9の右の折れ線グラフが作られる．

このようなグラフから系列間に何らかの傾向が見られるとき，その傾向を統計的に確認するためには，相関係数や回帰分析が有効である．

(3)　縦軸を対数変換した時系列グラフ

販売量の推移(1)

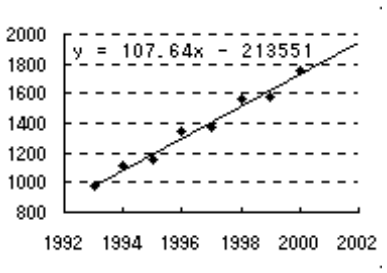
過去8年間の販売量が表示4.10のように求められている．これから，販売量がどのように変化したかを明らかにし，来年以降の予測をしたい．

毎年の増加量はほぼ同程度である．平均増加量を求めるために，Excel グラフに「近似曲線」を追加する[2]

[2] 具体的な方法は 第2単元の §4.1 で説明される．

表示4.10: 販売量の推移(1)

年	y	増加量
1993	976	
1994	1109	133
1995	1156	47
1996	1352	196
1997	1377	25
1998	1560	183
1999	1571	11
2000	1761	190



$$y = 107.64x - 213551$$

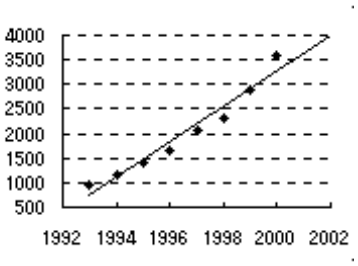
から，x が1増えると（1年経つと）107.64 増えることが分かる．

販売量の推移(2)

表示4.11のようなデータが得られたとき，どのように解析するかを考える．

表示4.11: 販売量の推移(2)

年	y	増加量	前年比
1993	968		
1994	1142	174	1.18
1995	1410	268	1.23
1996	1652	242	1.17
1997	2046	394	1.24
1998	2310	264	1.13
1999	2900	590	1.26
2000	3605	705	1.24



前と同様に対前年差を求めると，一定ではなく，大きく変化する．

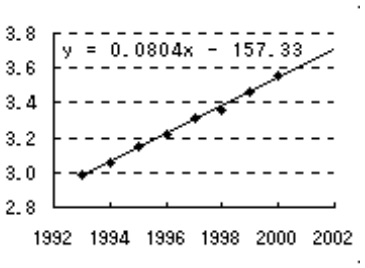
グラフを描くと，点は直線的ではなく，曲がっている．直線を当てはめて将来を予測することはできない．

対前年比を求めて見ると，ほぼ一定であることが分かる．すなわち，成長率が一定に近いと考えられる．このような場合は，y の対数変換が役に立つ．

対数については §1.6 で詳しく説明した．

表示 4.12: 販売量の常用対数の推移

年	y	log(y)	増加量
1993	968	2.986	
1994	1142	3.058	0.072
1995	1410	3.149	0.092
1996	1652	3.218	0.069
1997	2046	3.311	0.093
1998	2310	3.364	0.053
1999	2900	3.462	0.099
2000	3605	3.557	0.095



対数で対前年差を求めると，ほぼ一定となる．そこで，対数を縦軸にとって散布図を描くと，点が直線的に並んでいるので，直線を当てはめる．

この直線によって，将来の予測を求めることができる．

2002年の予測値をグラフから読み取ると 3.7 前後である．この値をもとの値に戻すには `=10^3.7` を入力する．5012 が得られる．

年成長率は，x の係数 0.0804 から `=10^0.0804 = 1.2033` となり，前の年の 1.20倍となることが分かる（成長率は約20%）．

売上げが10倍になるには何年

前のグラフで，売上げが10倍になるのは何年? という問に対する答えは，縦軸の目盛りが1増えるに必要な年数から求められる．上に求めた x の係数 0.0804 から，$1/0.0804 = 12.44$，すなわち，12〜13 年で10倍になることが分かる．

本日のまとめ

今日の学習内容は，昨日の応用である．実際の問題では，ある系列の原因と考えられる因子も，時系列で変化していることは多く，いくつもの系列の変化を同時に観察することが必要となる．

対数変換は極めて有用な手法である．理解が不十分と思われる人は，§1.6 を復習してほしい．

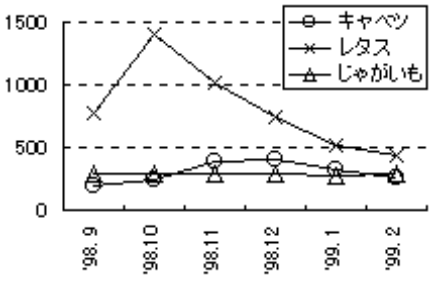
4.3　指数

(1)　小売価格の変動の比較

表示4.13**の左は，**1998**年の**9**月から**1999**年**2**月までの各月における，キャベツ，レタス，じゃがいもの東京での，**1kg**当たりの平均小売価格を示すデータである．**

表示4.13: **東京でのキャベツ，レタス，じゃがいもの小売価格**
— 1998**年**9**月～**1999**年**2**月** — **（単位：**1kg**あたり円）**

	キャベツ	レタス	じゃがいも
'98. 9	201	775	294
'98.10	244	1400	286
'98.11	391	1017	288
'98.12	406	736	283
'99. 1	315	517	278
'99. 2	261	428	284



これを見ると，これらの商品の小売価格は月によって多少とも変動していることが分かる．それで，それぞれの変動の様子を相互に比較したい．表示4.13**のグラフでは比較しにくい．**

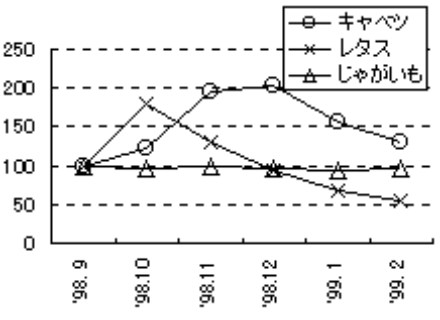
この比較を最も分かりやすくするためには，特定の月を 基準時 として定めそれぞれの商品のその月の価格を基準として，他の月の価格をその基準に対する比率で表わすのが便利である（基準時と比較される各月のことを 比較時 ということもある）．このような比率を100**倍しておけば，基準の値を**100**としたときの相対値が得られていっそう分かりやすくなる．**

9月を基準時に選んで，このようにして計算した相対値が表示**4.14**の左に示してある．**

例えば，10**月のキャベツの価格は**244**円であるから，**9**月の価格**201**円を基準とするこの相対値は，**(244 ÷ 201) × 100 = 121.4**となる．このようにして出し**

表示4.14: 東京でのキャベツ，レタス，じゃがいもの小売価格指数
— 1998年9月〜1999年2月 —（1998年9月=100）

	キャベツ	レタス	じゃがいも
'98. 9	100.0	100.0	100.0
'98.10	121.4	180.6	97.3
'98.11	194.5	131.2	98.0
'98.12	202.0	95.0	96.3
'99. 1	156.7	66.7	94.6
'99. 2	129.9	55.2	96.6



た相対値のことを，それぞれの小売価格の 指数 という．指数を論ずる場合には，何を基準とする指数であるかを必ず明らかにしなければならない．表示4.14右のグラフを見ると，これら商品の間で小売価格の変動の様子がかなり異なっていることが容易につかめるであろう．じゃがいもは安定しているが，レタスとキャベツは季節によって大幅な変動を示すことがはっきり分かる．

(2)　総合指数

§1.2 (2) では，戦後の物価の動きをグラフで観察した．そこで見た物価指数がどのような考えで導かれ，計算されたものかを説明する．

わが国の経済状況は，バブル崩壊が起こった1990年前後を境として大きく変化したといわれる．1960年代からバブル崩壊以前では，「最近は物価が上がっている」とか「わが国の工業生産は近年めざましく上昇している」といわれ，国民生活の水準が豊かになった時代であった．一方，バブル崩壊後では，「わが国の経済はデフレスパイラルで行き先の見えない長いトンネルに入り込んでしまった」とか「消費者の財布の紐が固くて物が売れなくなった」といわれ，経済成長が滞っている時代となった．

バブル崩壊を挟んで対照的な2つの主張は，物価や工業生産を計量的に測るようななんらかの目安がなければ具体的な意味を持たない．例えば，バブル崩

壊前の時代では「わが国の工業生産は近年めざましく上昇している」といわれたが，その陰で，むしろ下降状態にあった産業部門も存在したのである．一方，バブル崩壊後の時代でも「物が売れなくなった」といっても，必ずしも全ての商品が売れなくなったという意味ではない．同じ産業部門にありながら，儲かっている企業もあれば，危機に瀕している企業もあるのである．

物価水準や工業生産量というようなものは，もともと抽象的な概念であってそれを単独の統計系列で表わすことはできない．それらの変化を表わすには，いくつもの統計時系列の変化を総合することが必要である．

例えば，物価水準の動きを表わすには，多数の商品やサービスの価格の変化を総合するような計量化が必要であって，そのような総合的計量化の結果にもとづいてこそ,「物価が上がっている」あるいは「下がっている」ということが,具体的な意味をもってくるわけである．

一般に，いくつもの統計系列（多くの場合，時系列）の変化を総合して，全体としての変化動向を計量的に表現するようなものを，総合指数 というのである．

(3)　物価指数

総合指数の方法は，もともと物価水準の変化を測る手段として考案され，のちに同じ方法が生産量，取引量の動きや経済活動一般の消長などを測定するための手段としても応用されるようになったのである．ここでは，物価水準の変化を測ることを目的とする 物価指数 について学ぶことにしよう．

物価水準といっても，われわれの消費生活の面における物価水準と業者の卸売取引の面での物価水準とでは一般に変化の様相を異にするものである．したがって，どの面における物価水準の変化を測るかによって異なる物価指数を必要とする．消費生活の面での物価水準の変動を測るための物価指数を 消費者物価指数 といい，卸売取引の面における物価水準の変化を測る物価指数を 卸売物価指数 という．わが国では，消費者物価指数としては「総務省統計局消費者物価指数」が，また卸売物価指数としては「日本銀行卸売物価指数」が特に重視されている．そこで，消費者物価指数を例にとって物価指数を作る手順を学

ぶことにしよう.

(4)　品目の選定

消費者物価指数は，われわれの消費生活に用いられるいろいろな商品やサービスの価格の変動を総合して測るためのものであるが，実際にこれを計算するためには，消費生活に現われるいっさいの商品やサービスをもれなく取り上げるということはできない．したがって，特に重要ないくつかの品目（商品のみならず理髪，銭湯のようなサービスをも含む）を選び出し，それらの価格変動を総合するような物価指数を作ろうとするのである．このように，物価指数を作るもととして品目を選び出すことを 品目の選定 という．消費者物価指数の場合と，卸売物価指数の場合とでは選定される品目がずいぶん違うはずであることは容易に理解できるだろう.

「総務省統計局消費者物価指数」では，われわれの消費生活に関係の深い598品目が選定されている．また「日本銀行卸売物価指数」では，卸売取引額の多い，国内971，輸出209，輸入247，合計1,427品目が取り上げられている.

(5)　基準時の指定

単独指数の場合と同様に，総合指数の場合にも基準時を定め，その基準時の数値を100とみるときの相対値として各比較時の指数を出すのである．一般に物価指数の基準時としては，経済状態が正常で物価の安定した時期が選ばれる．基準時として単一の時点を指定することもあるが，1ヵ月，1ヵ年などの長さをもつ期間を選んで，その期間中の平均価格を基準時価格として指定する場合が多い.「総務省統計局消費者物価指数」では，現在は2000年の年間平均価格を基準としている.「日本銀行卸売物価指数」も，1995年の年間平均価格を基準として発表している.

(6)　算式の選択

物価指数は，いくつかの価格系列の変化を総合的に示す指数であるが，いくつかの系列の変化をどのように総合化するかはそう簡単ではない.

仮に，3つの食料品（米，卵，牛乳）の価格を総合化することを考える．

表示4.15: 3**食料品の価格（単位：円）**

		単価				購入金額		指数
品目	単位	基準時	現在	指数	購入量	基準時	現在	
米	円/kg	1300	1400	107.7	30	39000	42000	
卵	円/パック	250	260	104.0	15	3750	3900	
牛乳	円/l	170	175	102.9	20	3400	3500	
合計		1720	1835	**106.7**		46150	49400	**107.0**

単純に，単価の合計を求めて，指数化すると 106.7 **となる．米は**1**キログラム当たり，卵は**1**パック（**10**個）当たり，牛乳は**1**リットル当たりの価格である．単位が異なるから，単純に単価を合計することは意味がない．例えば，日米の価格比較をしようとすると，アメリカでは，米は**1**ポンド当たり，卵の**1**パックは**20**個入り，牛乳はクオート当たりかもしれない．とすると，どちらの国の単位を取るかにより，結果が異なってしまう．**

それぞれの品目の指数を求め，平均すると，

$$\frac{107.7 + 104.0 + 102.9}{3} = 104.9$$

となる．この方法によれば，単位の取り方による違いを除くことができる．

しかし，米，卵，牛乳を同格に扱って指数の平均を取るのは妥当だろうか？標準の4**人家族の**1**ヵ月の平均的な購入量を調査し，表示**4.15**の購入量の欄の値が得られたとする．**

これから，基準時と現在の購入金額合計を計算して，指数を計算すると 107.0 **となる．米の購入金額が多いので，総合指数は米の単独指数に引っ張られる．**

さらに，食生活が変化し，米の消費が減り，パンの消費が増えたときには，どうするかも問題である．そのために，前に述べたように，政府の物価指数の算出では，ときどき，対象品目の入れ替えなどをして，生活実感との乖離を少なくする努力をしている．

この簡単な例から分かるように，適切な総合指数を求めるときには，さまざまな注意が必要である．

本日のまとめ

物価指数については，ほとんどの方が耳にしたことがあるであろう．データを指数化することによって，時系列の変化の様子がつかみやすくなるし，他の変化との比較が容易になる．今日の学習で，指数が身近に感じられるとともに，指数化するむずかしさも知ってもらえたと思う．データの成り立ちについて理解できると，データをみるときには今までとは違った楽しさが感じられるであろう．

4.4　経済時系列の分析

(1)　経済時系列の構成要素

経済時系列の分析では，時系列の変動が3つの構成要素，すなわち，

- 傾向（トレンド）
- 周期変動
- 不規則変動

から成り立っていると考える．

傾向 とはかなり長期間にわたる変化の様相を示すものである．したがって，滑らかな曲線 を当てはめて考える．この中には，経済規模の自然増加による変動と，景気の変動が含まれる．

周期変動 とは，1年，1月，1週間，24時間 などの周期で変化するものである．

例えば，4月と12月は経済活動が活発で，逆に2月と8月は低下すると言われている．また，商店などの売上げは，曜日によって周期的に変化する．

このような変化を総称して 周期変動という．とくに，1年間を周期とする変動については 季節変動 と呼ばれることもある．

以上の2つの変動では説明することのできない残りの部分が不規則変動 である．それは，戦争や地震のような経済活動とは無関係に，突発的原因によって起こる変動と，偶然変動という名で呼ばれる規則性のない変動が含まれている．

平時の不規則変動は，おおむねこの偶然変動と考えて良い．

(2) 時系列分析の目的

経済時系列について前記の構成要素を分離し，解釈することによって次の2つの目的が達せられる．

その一つは，経済変動の将来予測 を容易にすることである．時系列変動の構成要素のうち傾向 及び 周期変動は，その時系列の動きを過去から現在まで支配している統計的規則性が変わらないかぎり，将来にわたって繰り返されるものと考えられるから，いったん分離されたこれらの要素を，将来（統計的規則性自体に変化が生じない程度に近い将来）の時期について合成することによって，かなり良い予測が得られるはずである．不規則変動のうち，戦争や地震のような突発的原因によっておこる経済変動は予測のかぎりではないが，偶然変動については，過去から現在に至る過程で観察された偶然変動の大きさだけ予測に幅をもたせて考えれば良い．

経済時系列分析のもう一つの目的は，経済現象を解明するための計量的分析に役立つ統計データの準備である．そのためには，経済にとっていわば外生的な季節変動及び不規則変動を時系列から除去しておくことが必要である．この作業を 周期変動と不規則変動の調整 といい，これら外生的な変動の除去された時系列を 調整済系列 という．日本銀行統計局及び経済企画庁（現内閣府）では国の重要な月次経済時系列について，原系列とともに調整済系列を公表している．周期変動及び不規則変動が調整されるならば，後には傾向（景気変動を含む）が残るわけである．

(3) 移動平均法

移動平均法 は，もとの系列から周期変動を除き，不規則変動の影響を小さくするために用いられる．

例えば，商店の当日の売上高にその前後各3日間の売上高を加えて7で割ったものは，移動平均である．その平均値の中には，日月火水木金土 がすべて含まれるので，曜日による周期変動は打ち消される．

さらに，雨天による売上の減少 や お祭による売上げの増加 などの不規則変動の影響は，7日分の売上を平均することにより減少する．

上に挙げた例のように，7つの観測値の移動平均を求めるためには，前後各3つの観測値を含めて平均すれば良い．これは7が奇数だからである．

毎月の観測値から12月を周期とする規則変動を除くためには工夫が必要である．

表示 4.16: 小型乗用車の新車登録台数

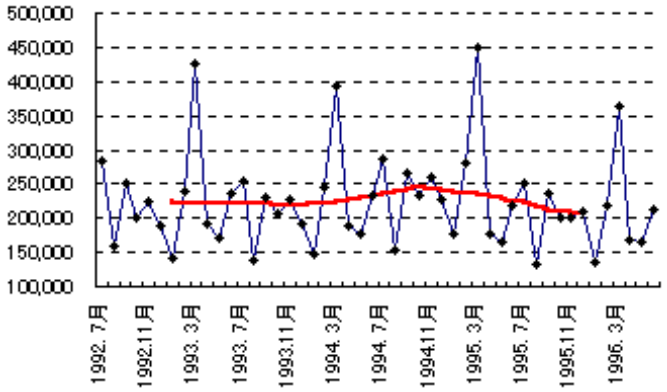
	A	B	C
1			
2	年月	登録台数	移動平均
3	1992. 7月	282,430	
4	1992. 8月	159,193	
5	1992. 9月	250,604	
6	1992.10月	201,527	
7	1992.11月	224,149	
8	1992.12月	188,580	
9	1993. 1月	142,703	224,938
10	1993. 2月	239,273	222,936
11	1993. 3月	425,266	221,216
12	1993. 4月	191,439	220,507
13	1993. 5月	172,113	220,813
14	1993. 6月	236,148	221,060
15	1993. 7月	254,091	221,402
16	1993. 8月	139,496	221,840
17	1993. 9月	229,021	220,753
18	1993.10月	206,077	219,313
19	1993.11月	226,939	219,442
20	1993.12月	191,726	219,543
21	1994. 1月	147,777	220,749
22	1994. 2月	244,695	222,745
23	1994. 3月	393,768	224,898
24	1994. 4月	188,374	227,562
25	1994. 5月	178,271	230,127
26	1994. 6月	232,425	232,997

	A	B	C
25	1994. 5月	178,271	230,127
26	1994. 6月	232,425	232,997
27	1994. 7月	286,747	235,721
28	1994. 8月	154,736	238,465
29	1994. 9月	265,466	242,290
30	1994.10月	233,555	244,222
31	1994.11月	261,033	243,246
32	1994.12月	226,515	242,116
33	1995. 1月	178,359	240,022
34	1995. 2月	279,971	237,560
35	1995. 3月	450,289	235,446
36	1995. 4月	178,222	232,959
37	1995. 5月	164,986	229,099
38	1995. 6月	218,595	225,849
39	1995. 7月	250,338	223,340
40	1995. 8月	132,050	218,930
41	1995. 9月	237,419	212,688
42	1995.10月	201,916	208,620
43	1995.11月	200,033	208,187
44	1995.12月	209,508	207,900
45	1996. 1月	135,142	
46	1996. 2月	217,359	
47	1996. 3月	363,090	
48	1996. 4月	167,780	
49	1996. 5月	165,033	
50	1996. 6月	211,657	

表示4.16は1992年から1996年までの小型乗用車の毎月の新車登録台数を示すデータである．新車登録台数には明らかな季節変動の存在することが予想され

る．そこで12ヵ月の移動平均を取り，季節変動及び不規則変動を除去したデータを求めてみよう．

表示4.17: 小型乗用車の新車登録台数(個々の値と移動平均値)



1993年1月の移動平均は1月を中心として12ヵ月の平均値としたい．1月の前後の５ヵ月（8,9,10,11,12,1,2,3,4,5,6) を取ると1月分足りない．そこで，前年の7月と今年の7月の平均を追加して，12個の平均値を求め，それを1月の移動平均とする．

7	8	9	10	11	12	1	2	3	4	5	6	7
△	○	○	○	○	○	○	○	○	○	○	○	△
1/2	| <	- - - - - - - - - - - -				11個					- - - - - - - - - - - -	> |	1/2

このような移動平均を求めるために，C9 のセルに

```
=SUM(B3/2,B4:B14,B15/2)/12
```

と入力する．SUM関数で，7月の半分，8月から6月，7月の半分を合計する．それを12で割って平均値を求めている．これを下にコピーする．

表示4.17はもとの時系列と移動平均をグラフに示したものである．この移動平均は，傾向（景気変動を含む）と不規則変動の一部の合成されたものと考えることができる．

(4)　季節変動の計算

先にも述べたように，移動平均値は傾向と景気変動と不規則変動の一部の合成されたものと考えることができる．したがって，元の時系列データについて，それぞれの時点における移動平均値からの偏差か，もしくは 移動平均値に対する比を求めることによって，元の時系列データから 傾向 と 景気変動 と 不規則変動の一部 の影響を取り去ったデータを求めることができる．つまり，このように求めた偏差 もしくは 比のデータには，景気変動以外の周期変動（主に季節変動）と 不規則変動の一部 が残ることになる．

移動平均値が全体としてそれほど広い範囲にわたって変動しないときには，それからの偏差を求めるのが良いが，多くの経済時系列にみられるように，移動平均値が対象とする期間にわたって非常に大きく成長しているような場合には，それからの偏差も相対的に大きくなっていくことが多いので，それぞれの実測値の移動平均値に対する比を求める方が良い．

表示4.18は，表示4.16の新車登録台数のそれぞれの時点における移動平均値に対する比を計算した結果である．例えば，1993年1月についてみれば，

$$\frac{142,703}{224,938} = 0.634$$

となる．

表示4.17でも明らかなように，新車の登録台数は，年度末の3月に多くなり，4月，5月には減少する．6月，7月に一時増えるが，毎年新しい年式の車の発表される9月の前にはまた谷がみられる，というかなりはっきりした季節変動を示している．

3年間の季節変動の平均を求め，季節変動の年間平均が 1.00 になるように修正したのが，表示4.18の「季節変動」の値である．

毎年の季節変動と平均した季節変動をグラフ化したのが表示4.19である．

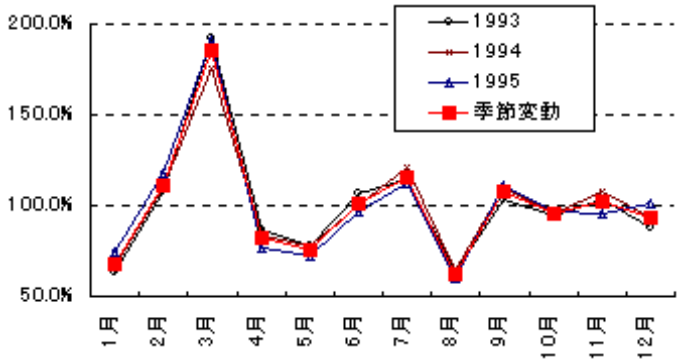
季節変動は毎年ほとんど変化しないことが分かる．

この例から分かるように，3年分の季節変動を求めるためには，その前後に半年ずつを追加した4年分のデータが必要である．

表示 4.18: 新車登録台数の季節変動

	1993	1994	1995	平均	季節変動
1月	63.4%	66.9%	74.3%	68.2%	68.0%
2月	107.3%	109.9%	117.9%	111.7%	111.3%
3月	192.2%	175.1%	191.2%	186.2%	185.6%
4月	86.8%	82.8%	76.5%	82.0%	81.8%
5月	77.9%	77.5%	72.0%	75.8%	75.6%
6月	106.8%	99.8%	96.8%	101.1%	100.8%
7月	114.8%	121.6%	112.1%	116.2%	115.8%
8月	62.9%	64.9%	60.3%	62.7%	62.5%
9月	103.7%	109.6%	111.6%	108.3%	108.0%
10月	94.0%	95.6%	96.8%	95.5%	95.2%
11月	103.4%	107.3%	96.1%	102.3%	101.9%
12月	87.3%	93.6%	100.8%	93.9%	93.6%
平均				100.3%	100.0%

表示 4.19: 小型乗用車の新車登録台数



本日のまとめ

時系列の変動要素を傾向，周期変動，不規則変動に分解するのは，変化について統計的規則性を探索するための考え方である．新車登録台数の例は大変分

かりやすいので，皆さんが業務データなどの時系列データを解析する際にも応用できるであろう．例えば,「ニッパチ（2月と8月）は売上が落ちる」と経験的にいわれていることも，この方法を使って検証することができるであろう．

5 演習解答

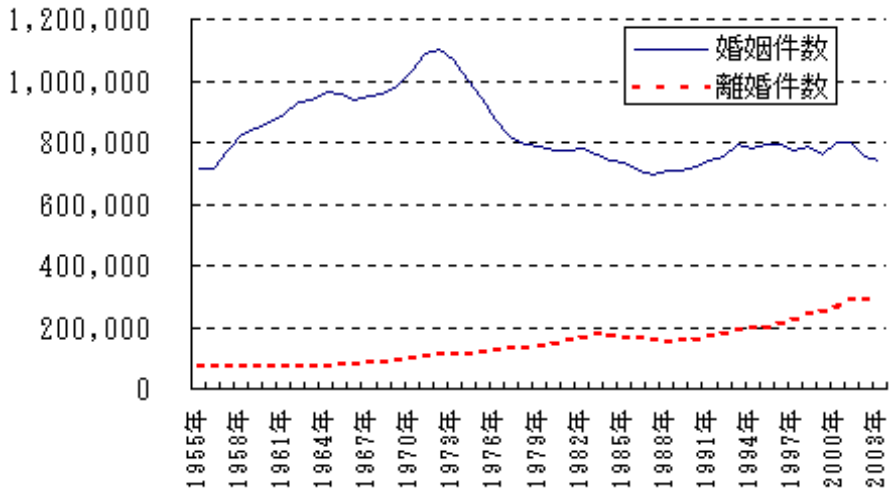
5.1 第1章 統計の利用

演習 1 (p.5)

「人口動態統計」から，1955 年から 2003 年の間の 婚姻，離婚の件数をダウンロードした.

件数を時系列グラフに表わしたのが表示5.1 である.

表示5.1: 層別した時系列グラフ

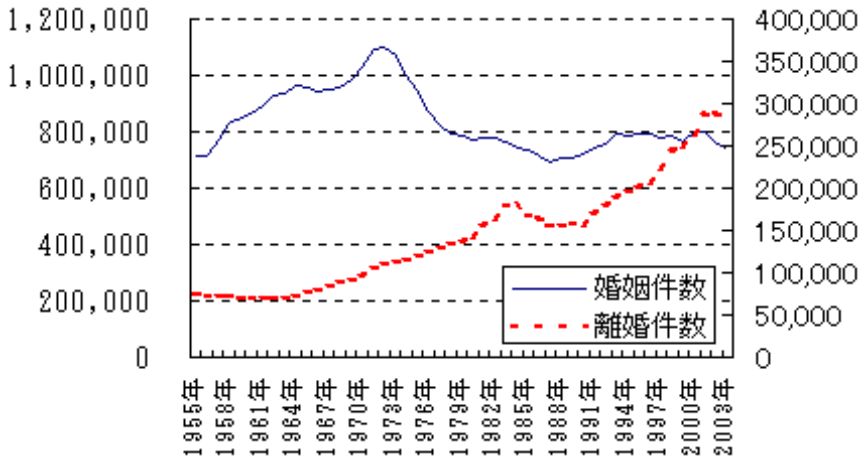


婚姻件数は1970 年から 1974 年が最も多い．団塊世代がこの時期に多く婚姻していると考えられる。

「婚姻件数」の線と「離婚件数」の線の幅がせまくなってきている.

婚姻と離婚の件数の変化範囲に大きな違いがあるので，両者の目盛りを左右に取ってグラフを作り直すと表示5.2 が得られた.

表示5.2: 層別した時系列グラフ



このグラフを眺めると，離婚件数がうなぎ昇りに増加し，50年で約4倍になっていることが分かる．それに対して，婚姻件数は上に述べた大波はあるが，80万件前後で推移しているようである．

演習 2 (p.7)

総務省の2000年国勢調査結果から，1920年と1950年の男女別年齢分布データをダウンロードし，ピラミッドグラフを作成すると，表示5.3が得られた．

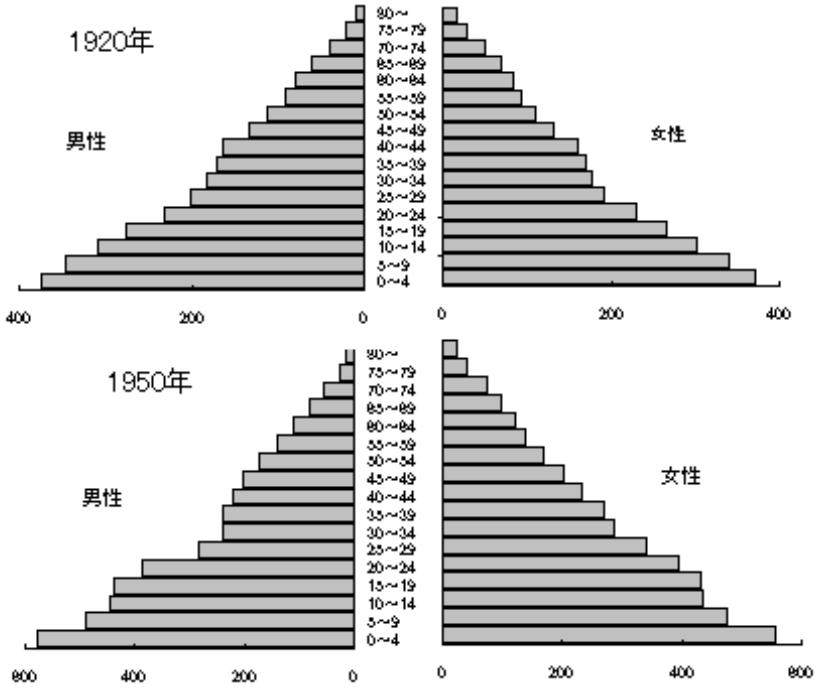
第2次世界大戦の終戦年は，1945 年で，その年に生まれた人は1950年には5歳である．日清戦争，日露戦争の終戦年は，それぞれ，1895, 1905年である．その年に生まれた人は1920年には 25 歳と 15 歳である．

テキスト本文に述べているように，1950年のグラフでは，終戦直前の出生率が減少し，戦後のベビーブームに続いている．また，戦争時に20歳から40歳の男性が女性に比べて少ない．

それに対して，1920年のグラフでは，日清・日露戦争の影響は顕著ではない．第2次世界大戦の影響の大きいことが分かる．

ちなみに，3つの戦争の戦死者は1万，8万，70万人である．この数値には，

表示5.3: 1920年と1950年の人口ピラミッド



非戦闘員を含まない.

演習 3 (p.38)

自覚症状のある人は病院に行くであろう.

自覚症状のない人でも，自分の健康に自信のない人は，病院や保健所などで健康診断を受ける.

健康に自信のある人が健康診断を受診するかどうかは年代によって異なると考えられる.

現役世代では，余裕時間が少なく，自ら健康診断を受診する人はかぎられるであろう．それに対して，退役後の老人は時間の余裕もあり，気軽に健康診断を受診することが考えられる．特に，保健所は費用もわずかであるから，受診率が高いであろう.

上に説明したことから，保健所に来所した人は，保健所管轄地域住民の成人男子全体（母集団）の縮図とはなっているとはいえない．したがって，この調査結果から,「歳をとるほど自分の健康に自信を持っている」という結論を出すのは正しくない.

5.2　第2章　集団の観察と統計的規則性

演習 4 (p.53)

タクシーを拾って外出するとき，空車の拾い易さを考慮して家を出る時間を決めたい．そのために，一定時間（ここでは10 分間）に通過する空車台数を数える調査を行うことにした.

空車の台数は，曜日，時間帯，天候，方向 などによって影響を受けると考えられる.

このような関係を正しくとらえるためには，曜日，時間帯について偏りのない調査をしなければならない．例えば，毎日の帰宅時に，調査をするという ¥bf 安易な 方法では，平日は夜，土日は昼間というように偏った調査になり，正しい判断をすることはできない.

この例が示すように，事前に十分な検討をせずに，思いつきで調査をすることは極めて危険である.

5.3　第3章　統計解析の基礎知識

演習 5 (p.61)

表示3.6 と同様の計算により,

$$\mathrm{E}[x] = 2.5, \quad \mathrm{V}[x] = 1.25, \qquad D[x] = \sqrt{1.25} = 1.118$$

が得られる.

演習 6 (p.63)

表示3.8 **と同様の計算により,**

$$\mathrm{E}[T_2] = 5.0, \quad \mathrm{V}[T_2] = 2.50, \quad D[T_2] = \sqrt{2.50} = 1.5811$$

が得られる.

演習5 **の結果と比較すると，期待値** $\mathrm{E}[x]$**，分散** $\mathrm{V}[x]$ **の**2**倍，標準偏差** $\mathrm{D}[x]$ **の** $\sqrt{2} = 1.414$ **倍になっていることが分かる.**

演習 7 (p.67)

総重量 T **を**

$$T = 3x + 2y \tag{5.1}$$

と表わす.

$$\mathrm{V}[T] = 3^2\mathrm{V}[x] + 2^2\mathrm{V}[y] = 9 \cdot 64 + 4 \cdot 36 = 720 = 26.83^2$$

となる．それに対して,

$$T = x_1 + x_2 + x_3 + y_1 + y_2 \tag{5.2}$$

と考えると,

$$\begin{aligned}\mathrm{V}[T] &= \mathrm{V}[x_1] + \mathrm{V}[x_2] + \mathrm{V}[x_3] + \mathrm{V}[y_1] + \mathrm{V}[y_2] \\ &= 3 \cdot 64 + 2 \cdot 36 = 264 = 16.25^2\end{aligned}$$

となる.

式(5.1) **は**3 **人の男性，**2 **人の女性が同一人物である場合の式である．**3 **人の男性の体重を** $x_1,\ x_2,\ x_3$**，**2 **人の女性の体重を** $y_1,\ y_2$ **とする式**(5.2) **が正しい.**

本文で述べたように，式(5.2) **が成立するのは，**5 **人の体重が独立である場合に限られる．例えば，大学のエレベータにラグビー部の選手が**3 **人乗るという場合は，上の式よりも標準偏差が大きくなる.**

演習 8 (p.67)

1枚の紙の厚さ x の標準偏差は1ミクロンである．1折りの厚さ　y は，同じ厚さの紙が16枚重なったものであるから $y = 16x$ である．その分散は $\mathrm{V}[y] = 16^2 * 1 = 256$ **ミクロンである．本の厚さ T は**

$$T = 16y_1 + 16x_2 + \ldots + 16y_8$$

で表わされ，８つの折の厚さは互いに独立であるから，その分散は，

$$\mathrm{V}[T] = 8 \cdot \mathrm{V}[t] = 2048 = 45.25^2$$

となる．すなわち，本の厚さの標準偏差は 45.25 **ミクロンである．**

演習 9 (p.69)

解答は各自試みよ

演習 10 (p.73)

表示5.4 **は，表示**3.11 **を参考にして，** $\mu,\ \sigma,\ x$ **を入力すると，** $z = (x - \mu)/\sigma$ **を計算し，** z **以下と** z **以上の確率を求める計算表を作成したものである．**

表示5.4**: 正規分布に関する計算表**

	A	B	C	D	E	F	G	H
1		期待値	標準偏差	x	z	以下	以上	答
2	(1)	10	2	12	1.00	0.841	**0.159**	0.159
3	(2)	10	2	11	0.50	0.691	**0.309**	0.309
4	(3)	10	2	9	-0.50	**0.309**	0.691	0.309
5	(4)	10	2	9	-0.50	**0.309**	0.691	
6	(4)	10	2	12	1.00	**0.841**	0.159	0.533
7	(5)	10	2	15	2.50	0.994	**0.006**	0.006
8	E2:	=(D2-B2)/C2						
9	F2:	=NORMSDIST(E2)					H6:	=F6-F5
10	G2:	=1-F2						

計算式は表示5.4 **の下に示されている．**

問題のデータを入力し，さらに，以下，以上の確率から，太字の数値を使って要求された解答を算出した結果を H **列に示す．**

小問(4) の答は，12以下の割合から9以下の割合を引いて求められる．

演習 11 (p.73)

体重合計の期待値 $\mathrm{E}[T]$ は

$$\mathrm{E}[T] = 3 \times 62 + 2 \times 47 = 280$$

である．

体重合計の標準偏差は 演習7 で16.25 と求められている．

これから，ブザーがなる確率は，標準正規分布で

$$z = \frac{300 - 280}{16.25} = 1.231$$

以上となる確率に等しい．その値は，Excel の統計関数 `=1-NORMSDIST(1.231)` を用いて計算され，0.109 となる．

5.4　第4章　時系列データ

演習 12 (p.78)

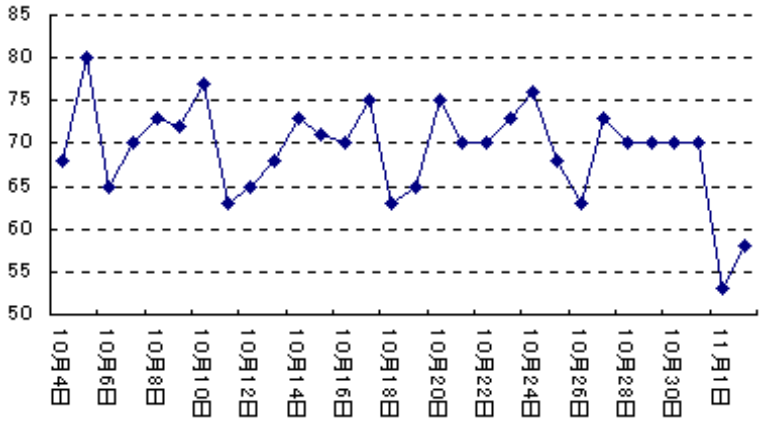
ある人が 10月4日 から 11月2日 までの1ヵ月間にわたり，毎夕定時に脈拍を数えた．実際の数値は，Excel ファイルに示す．

データをそのまま時系列グラフ化すると表示5.5 が得られる．

休日の脈拍は低いように見えるので，平日と休日を分けてプロットしたのが表示5.6 である．10月5日は日曜日であるが，出張したので平日としてプロットした．

この層別した時系列グラフを見ると，この人は平日の脈拍は休日に比べて高くなる．さらに，月曜日から金曜日まで疲労が蓄積するためか，脈拍が右肩上がりの傾向が見られる．

表示5.5: 単純な時系列グラフ



表示5.6: 層別した時系列グラフ