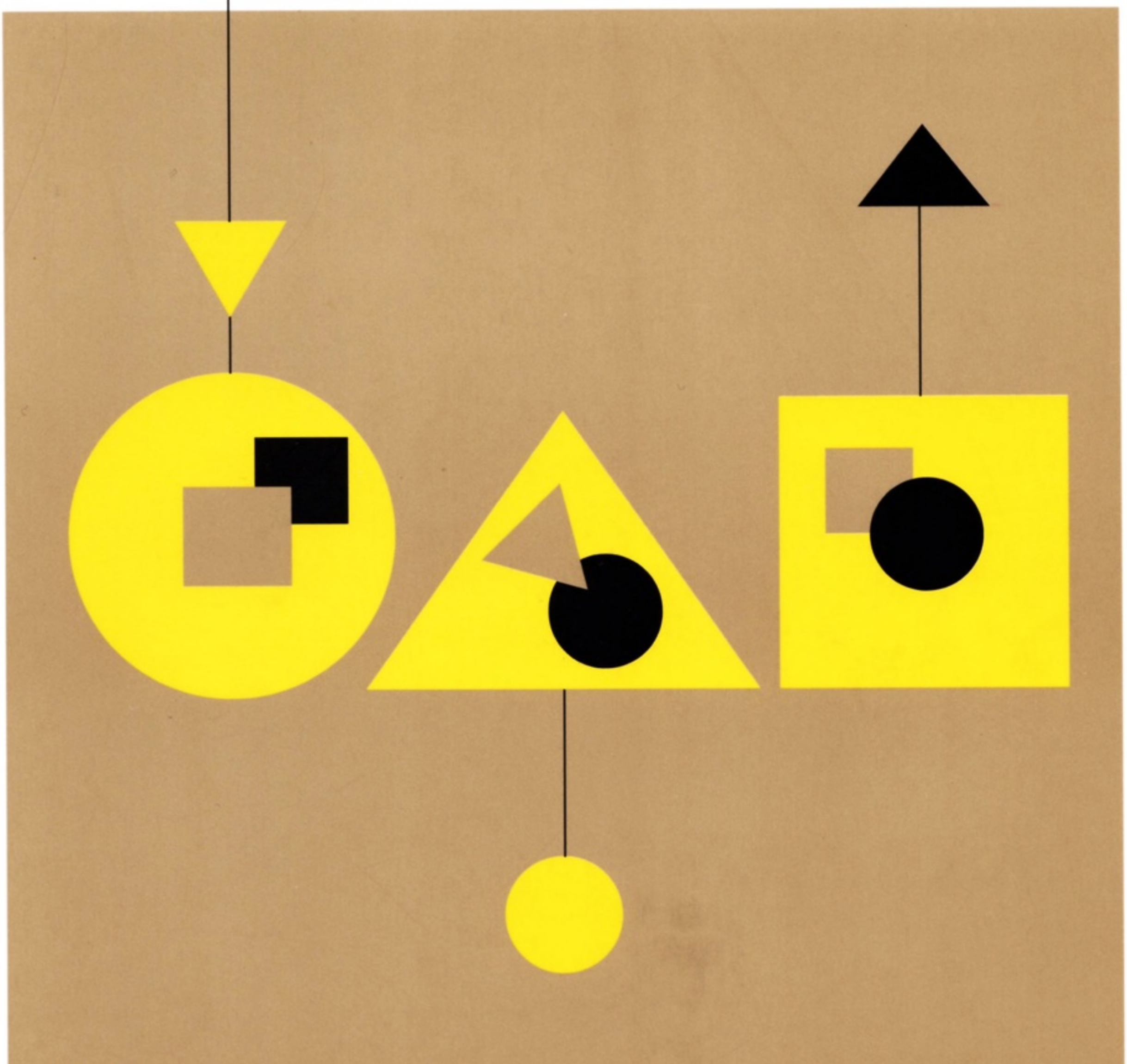
Statistics for Bioscience

川崎医科大学助教授　市原 清志 著

バイオサイエンスの統計学

正しく活用するための実践理論



南江堂

発刊に寄せて

近年のバイオサイエンスの進歩は正に目を見張るものがある．生命現象（バイオ）の巧妙なしくみを追求することはそれ自体，学問（サイエンス）として限りない興味を抱かせるし，そのたくみさを利用した応用科学への発展も留まるところをしらない．しかしながらバイオはそのたくみさの故に，実に複雑なメカニズムから成り立っており，そのすべての理論を解明しさえすれば，表現される結果も当然確定的であるはずなのだが，実際にはブラックボックスに包まれ，あまりにも不確定要素が多い．一方，統計学はこの不確定要素をもっとも理論的に構築したものということができ，これを活用すればよいということは誰しもが考えることであろう．ところが統計学というものが数学といういわばもっとも確定的な論理を基礎としているため，バイオにたずさわっている人達にとってはなかなかなじめないし，ましてやそれを利用しようとしても使いこなせないことが多いのが現実である．数式にあてはめて計算しようにも，どの式を使ったらよいかわからない，その数式の意味をわかろうとして統計理論を勉強していくとますますわけがわからなくなる，かといって数式のない統計の本では文学的な感覚はわかっても実際には使えない，等々といった実践家の悩みはよく聞かされるところである．

そこで統計学の理論的本質に立脚しつつ，不確定要素の多いバイオサイエンスの分野で活用するために企画されたのが本書である．筆者の市原清志先生は医師であり，研究者であり，教育者であるが，何といっても統計学を実践の場から身につけたことがユニークといえる．本書はそういった意味で従来の書とは異なり，多くの野心的な試みがなされている．図式により概念をイメージ化する，例題を豊富にして活用を会得させる，誤用例から正しい認識を把握させる等々である．

はじめにも述べたように，バイオサイエンスは不確定性が多いため，統計学の適切な活用により“みえないものまでみえるようにする（不可視情報の可視化）”という効用がある一方，不適切な利用は“うそのものまで本当にしてしまう（虚偽の歪曲・真実化）”という危険性がある．正に両刃の剣なのである．

願わくば，読者の方々が，本書のユニークな意図をくみとられ，統計学の適切な活用によってバイオサイエンスの発展に寄与される一助となることを祈るものである．

1989年12月

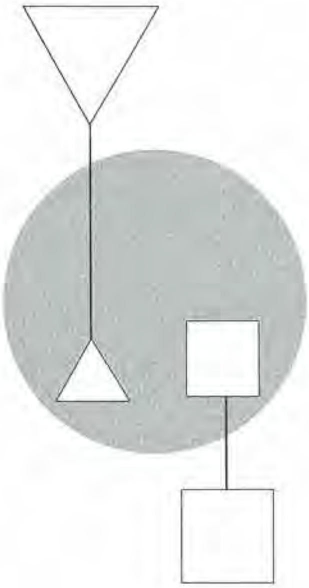
大阪大学医学部教授

宮 井　潔

バイオサイエンスの統計学

正しく活用するための実践理論

川崎医科大学助教授　市原 清志 著



南江堂

発刊に寄せて

近年のバイオサイエンスの進歩は正に目を見張るものがある．生命現象（バイオ）の巧妙なしくみを追求することはそれ自体，学問（サイエンス）として限りない興味を抱かせるし，そのたくみさを利用した応用科学への発展も留まるところをしらない．しかしながらバイオはそのたくみさの故に，実に複雑なメカニズムから成り立っており，そのすべての理論を解明しさえすれば，表現される結果も当然確定的であるはずなのだが，実際にはブラックボックスに包まれ，あまりにも不確定要素が多い．一方，統計学はこの不確定要素をもっとも理論的に構築したものということができ，これを活用すればよいということは誰しもが考えることであろう．ところが統計学というものが数学といういわばもっとも確定的な論理を基礎としているため，バイオにたずさわっている人達にとってはなかなかなじめないし，ましてやそれを利用しようとしても使いこなせないことが多いのが現実である．数式にあてはめて計算しようにも，どの式を使ったらよいかわからない，その数式の意味をわかろうとして統計理論を勉強していくとますますわけがわからなくなる，かといって数式のない統計の本では文学的な感覚はわかっても実際には使えない，等々といった実践家の悩みはよく聞かされるところである．

そこで統計学の理論的本質に立脚しつつ，不確定要素の多いバイオサイエンスの分野で活用するために企画されたのが本書である．筆者の市原清志先生は医師であり，研究者であり，教育者であるが，何といっても統計学を実践の場から身につけたことがユニークといえる．本書はそういった意味で従来の書とは異なり，多くの野心的な試みがなされている．図式により概念をイメージ化する，例題を豊富にして活用を会得させる，誤用例から正しい認識を把握させる等々である．

はじめにも述べたように，バイオサイエンスは不確定性が多いため，統計学の適切な活用により“みえないものまでみえるようにする（不可視情報の可視化）”という効用がある一方，不適切な利用は“うそのものまで本当にしてしまう（虚偽の歪曲・真実化）”という危険性がある．正に両刃の剣なのである．

願わくば，読者の方々が，本書のユニークな意図をくみとられ，統計学の適切な活用によってバイオサイエンスの発展に寄与される一助となることを祈るものである．

1989年12月

大阪大学医学部教授

宮井　潔

はじめに

本書を手にされた方の多くが必要としている“統計”とは，主として検定に関することではなかろうか．しかし，これまでの統計書の展開は，まず確率論にはじまり，分布型の理論，標本抽出の理論，推定論という形がとられ，肝心の検定法の章は，ずっと後にならないと姿を現さない．このため，せっかく向学心に燃えて取り組んでも，そこまで行き着かないうちに挫折してしまうケースが多い．またそれらの序論がわからないと，統計学を真に理解することが困難であるかのような錯覚に陥りがちで，統計は専門家にしか理解できない厚いベールに包まれた存在になってしまっている．

本書では，そんなベールをあっさりはがし，検定法に関する身近な例題と多数の図を通して，統計の原理と使い方が専門外の方にもわかりやすくイメージできるよう，実践的な解説を試みた．

さて，私の専攻は数学ではなく，統計に関しては大学時代ほとんど予備知識がなかった．しかし卒業後，研究のニーズから何冊かの統計書に接した．さらにその後，恩師の宮井教授よりその学習内容を発表する機会が与えられ，いやがうえにもより多くの統計書に取り組まざるをえなくなった．その学習過程はあまりスマートなものでなかったが，次の二つのことがらが重要であった．

一つは，実際に多くの人から，さまざまなデータを持ち込まれ，一緒に考え悩む機会があったこと（多くの実例に触れることの重要性）．

二つめは，問題解決の過程で，多くの統計書に目を通し，共通の概念を自分なりにイメージし，それを図式化したこと（データと数理概念の図式化）であった．

本書は，その過程をそのまま整理したもので，通常のテキストと異なった構成をとった．すなわち実用上重要な検定法からずばり入り，多数の例題の中で，統計学の理論がどのように生きているかを実例に即して解説し，最後に統計学の総論「統計の正しい利用と解釈」を述べる，という形をとった．

また統計学の数理概念のほとんどを図式化し，それをイメージで理解できるよう工夫した．しかし，その形態はこれまでの一般向き統計入門書のように，「数式を省略し，その概念を文学的に表現する」という形をとらなかった．それは，わかったような気になるが，いざ使おうとすると使えないからである．むしろ数式は活用公式として重要となるだけでなく，複雑な概念をコンパクトに語ってくれる．このため，本書では数式をできるだけ多く残し，逆に文章の方をできるだけ省略した．これは，「数学を語るのにあまり多くの言葉を要しない．数式に語らせればよい．図を媒体とし

て，このためには，できるだけ多くの身近なデータとイメージの浮かぶ図式が必要だ」という，本書の基本理念によっている．

最後に強調すべきことは，検定法の使い分けや，結果の解釈法を知らずに使っているため，180度違った結論を導いているケースが多いという事実である．しかし，これまでの多くの本では，純粋な例ばかりが示され，実際上問題になる誤りやすいパターンのデータの取り扱い方については，ほとんど触れられていないのが現状であろう．

そこで本書では，初学者でも誤った使い方をしないよう，最後の総論の章でたくさんの誤用例を図式的に示した．その中でも述べたが，それら誤用例の多くは，次の大切な原則を見失ったために生じたと考えてよい．

「データが何かを語りかけている．統計の正しい利用は，まずそれを感じ取ることから．」

本書が，このような“現実のデータに即した統計の活用”に役立てば幸いである．

終わりに，本書発刊のきっかけをお与えいただくとともに，データを読む視点を日頃いろいろな形で御指導いただいている，宮井　潔教授，稚拙な著者の下絵を巧みにイラスト化していただいた，漫画家北星晃平先生，そして，出版にあたり，絶大なご援助をいただいた南江堂の横井　信氏，中村泰子氏に深謝する．

1989年12月

市 原 清 志

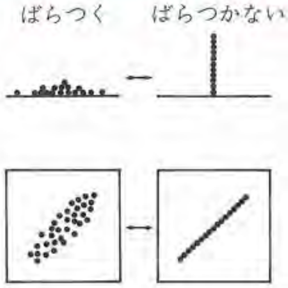
目次

序説

統計学とは

統計学は，ばらつきを伴う情報を，客観的に分析・評価する学問

"ばらつき" とは



◆ あらゆる情報にばらつきはつきもの，我々は日常それを経験的に処理している

ある月の雨の多さ，野菜やくだものの値段など，日常のあらゆる情報には，ばらつき（誤差）がつきものである．したがって我々は，経験と勘または"常識"を通し，ばらつきを考慮してそれらの情報を読んでいる．

たとえば，今年4月の雨日の数が，平年並みかどうかを判断する場合，我々は記憶を頼りに，まず実際の雨日の概数を頭に浮かべる．そして，過去の4月の雨日の割合を漠然と思い起こし，それと比較することで，今年が異常かどうかを判断する．

また，家庭の主婦が八百屋でレタスの値段をみて，今日の値段は買うべき値段か，買い控えるべき値段かを判断する場合も同様で，彼女は無意識のうちに，最近のレタスの値段の変動幅，他の野菜の物価の傾向などの要素を頭に浮かべ，その勘を頼りに総合的な判断を下している．

ただし，これらの判断はかなり大ざっぱなもので，情報を読む人の主観や経験に大きく左右される．

◆ 科学情報では，ばらつきに対してより客観的な評価が要求される

科学の世界でも，あらゆる現象にばらつき（誤差）がつきもので，研究者も同様にそのことを考慮した判断が要求される．

たとえば，研究者が薬物Aと薬物Bをネズミに投与してその作用を比較したとする．いま両者に対する反応にかなりの差が見られても，それが誤差の範囲内のものかどうか容易に判断できない．

また，ある薬物に対する肝障害の症例を何例か集めたとき，肝機能検査値は症例によって，また重症度によってかなり違った動きを示した．しかし，症例数が十分でないと，ただ眺めただけでは，検査値のばらつきにどのような相互関係があるのか浮かび上がってこないし，もちろん障害のパターンを分類することも困難である．

患者＼項目	GOT	GPT	γGTP	ALP	総蛋白	アルブミン	コレステロール
a	23	28	331	320	4.2	2.6	187
b	66	14	142	202	7.0		
c	125	125	152	160	7.6		
d	131	184	48	128	7.4		
e	18	22	131				
f	51	54			4		
g	54	16			6		
h	23	94			5.9	3.	
i	99	2		25	7.2	4.2	170
j	73	153		384	6.0	3.4	186
k	115	26		240	6.8	3.2	103

これらの情報に対する取り扱いは，先の日常的な情報の取り扱いと次の2点が異なる．

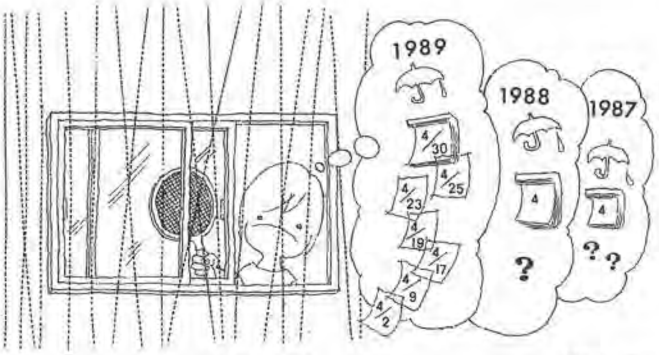
まず第一の相違点は，**判断基準となりうる経験の有無**である．すなわち，日常の情報の場合，経験や勘または常識によって，それがどの程度ばらつくかある程度予想できる（またはそう思い込んでいるだけの場合もあるが…）．これに対して，科学情報では，データ数が不十分か，その現象の全体像がまだ不明なので，ばらつきを経験的に処理できないケースが多い．

第二の相違点は，**判断の客観性**である．すなわち，日常的な情報の取り扱いでは，その判断が大ざっぱでよく，ある程度主観が入っても許容されるのに対し，科学情報では，情報の公表を前提とするので，その評価に客観性が要求される．

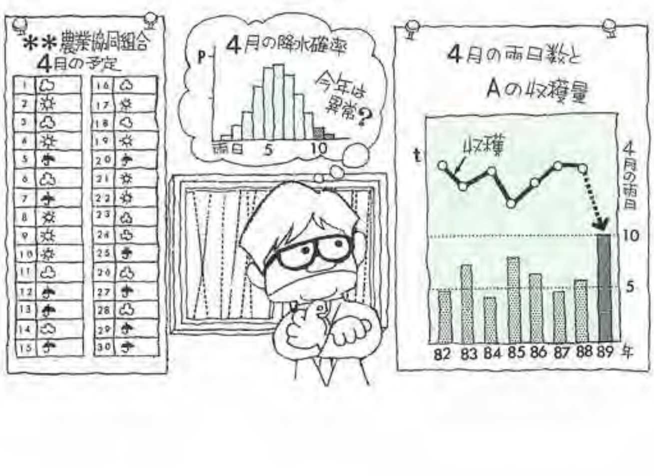
◆ 統計学は，ばらつきを伴う情報の客観的な分析・評価を可能にする

統計学は，このように不確定な情報や，複雑に入り組んだ情報を分析・評価したり，そこから何らかの客観的な判断を導く手段を与える．

たとえば，上の例で，雨日の割合をより科学的な立場で評価しようとすると統計学の助けが必要となる．それを用いると，過去の降水情報に照らして，今年が確率的にどの程度偏っているかを検定したり，雨日数がどの範囲なら妥当かを推定したりできる．



主観的判断



統計的判断

またレタスの値段については，価格の変動を時系列的にとらえ，その棄却限界線を設定できる．さらには，より高度な統計解析手法を用いると，他の物価との関連性や変動の周期性を分析することもできる．

また，ネズミの例では，A，Bの差の生じる確率を計算することにより，それが意味のある差か，偶然の差にすぎないのかを客観的に判定できる．

さらには，薬物性肝障害に対するデータ解析では，その重症度を重回帰分析という手法で予測したり，検査値の相互関係を分析することで重症度を決める因子を見つけだしたりすることも可能となる（因子分析）．さらに，クラスター分析を使うと，類似したパターンを示す症例どうしをくくり出してグループ分けすることもできる．

以上をまとめると，「統計学は，ばらつきを伴う情報を，客観的に分析・評価する学問」ということができる．

統計学の 2 つの機能

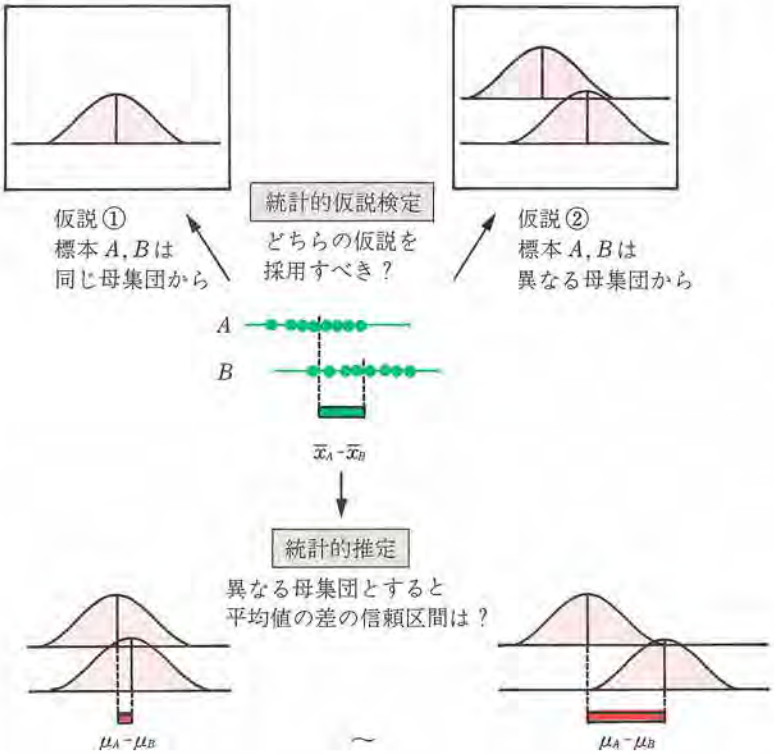
上の事例で紹介したもの以外にも，多数の統計手法があるがその機能は，次の 2 つに大別できる．

機能 1：少数の情報から，全体を推し量る…推論する

これは，部分から全体を推し量る機能である．統計は，全体としての**母集団**からその一部としての**標本**を取り出すことで，少数の情報から効率よく全体について推論することを可能とする．その方法は，大きく検定と推定に分かれる．**検定**とは，標本を基準に母集団の特徴や状態について何らかの仮説を設け，その妥当性を確率論的に検証する方法である．正確には**統計的仮説検定** statistical hypothesis testing と呼ばれる．たとえば，ある 2 つの標本に差異を認めたとき，それらは同一母集団から得られたという仮説と，そうでないという仮説をおき，確率的にどちらがより妥当かを客観的に判定する手段が検定法である．先の 4 月の雨日に関する例では，それが例年どおりという仮説と，そうでないという仮説をおいて比率の検定を行うと，どちらがより妥当かを調べることができる．本書で扱った統計手法の多くはこの検定に属する手法である．

一方，**推定**の方は，標本の特徴（統計量）から，母集団の特徴（母数）を推し量る方法である．正確には**統計的推定** statistical estimation と呼ばれる．たとえば，2 標本の平均値にずれが見られた場合，母集団が違うとすると，その平均値の差はどの範囲かを信頼区間という形で求めたり，x と y の関係を示す回帰直線を求め，任意の x に対する y の予測区間を求めること，などが推定に属する統計手法である．

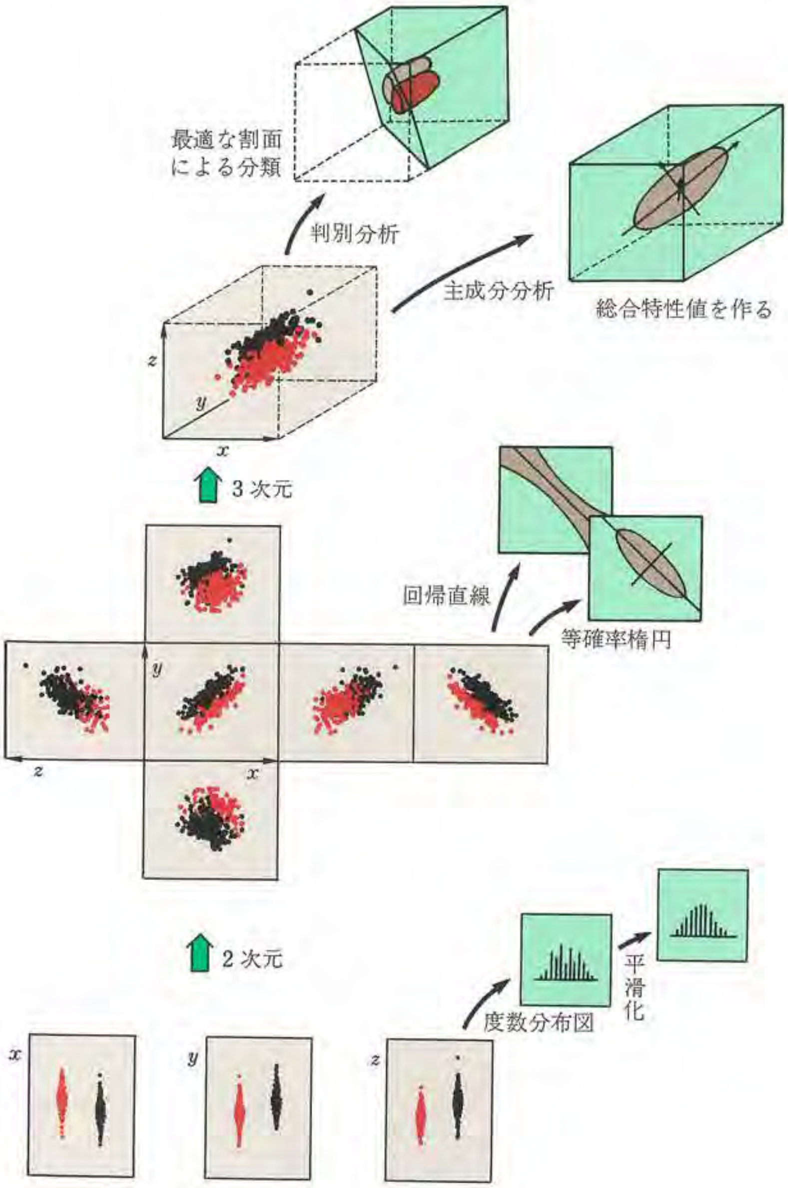
◀検定と推定



統計の機能1：少ない情報から
効率よく全体像
をつかむ

機能 2：情報の基本構造を明らかにする…解析する

これは，標本から得られた情報を数量的に要約したり分類する機能である．これには，平均・標準偏差を求めデータを要約するといった基本的なものから，2 変量の回帰式を求めたり，多次元情報を分類・整理して，その基本構造を明らかにするといっ



統計の機能 2：情報の基本構造を明らかにする

た高度なものまで多数の手法がある．とくに，多次元情報に対するものは，多変量解析 multivariate analysis と呼ばれ，人の頭では考えられない複雑な情報を分析し，その潜在構造を解き明かす手段として注目される．

たとえば前頁の図に示すように，一見複雑に絡み合った情報も，変数を増やして解析するとその相互関係を明らかにでき，またそのプロフィールを取り出すことも可能になる．

すなわち，1次元では，ヒストグラムをつくることにより，集団の分布のプロフィールを明瞭に把握できる．また，2次元的には，回帰分析を用いると，2つの変量間の関係を記述でき，これをもとに新しく生じた現象に対して一定の予測が可能となる．さらに3次元では，集団の関係がより明瞭に浮かび上がるだけでなく，普通とらえにくい変量の相互関係や変動の中心軸を求めたり，個体を最適な条件でグループ分けするための判別式を作成できる．

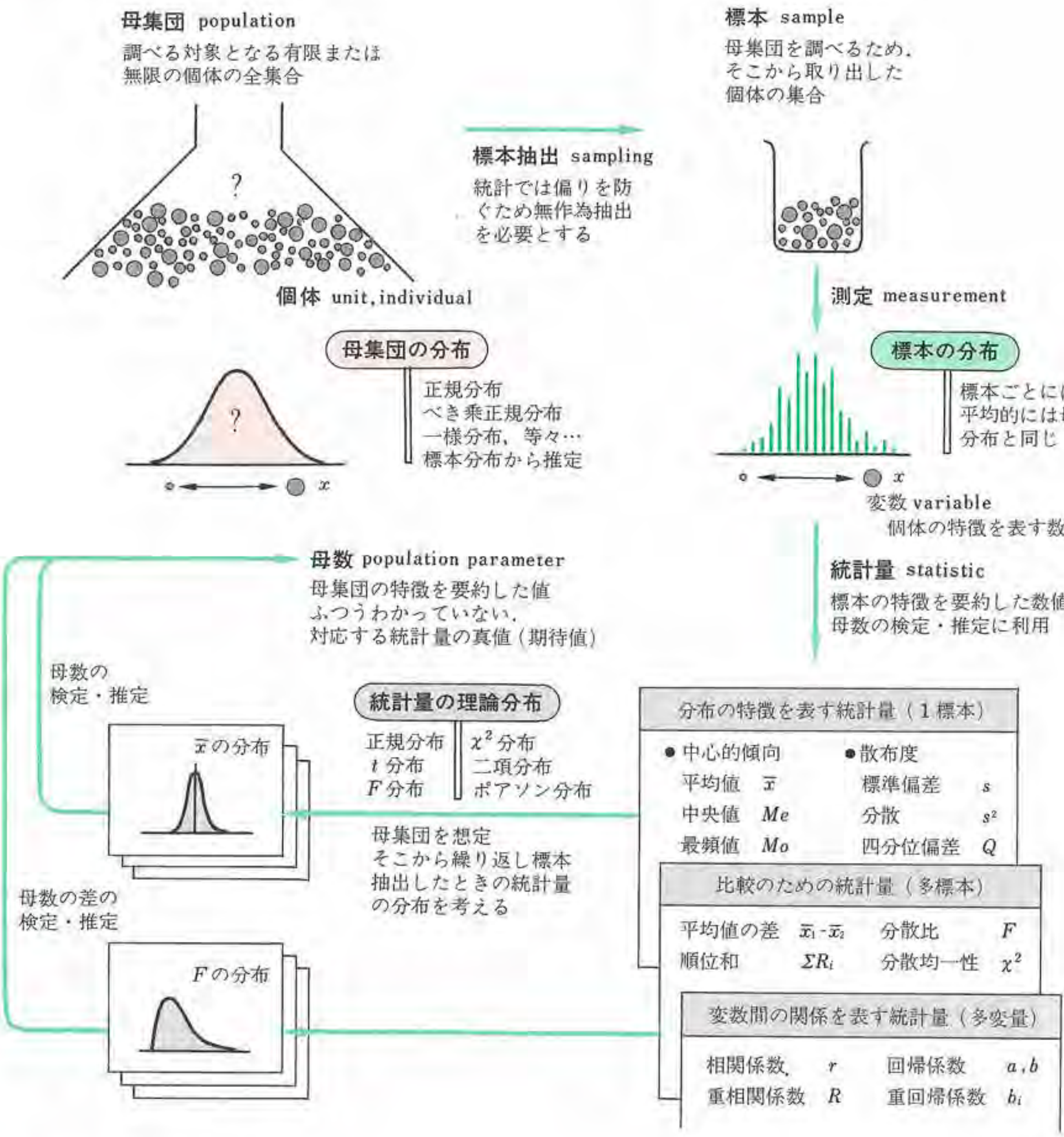
この統計の2つの機能は，決して独立したものでなく，ふつうは一体となって利用される．たとえば，検定では，まずデータを検定しやすい形に要約し，それをもとに一定の推論を行う．また多変量解析では，データの構造を示す関係式を求めると同時に，式の係数の意味を検定によって判定する．

以下本書では，各統計手法を10頁に示すデータ形式にそって分類し，これら2つの基本機能（とくに前者）の使い方を演習形式で解説する．また各部分の理解に必要な統計の基本概念は，実例に添って，その都度図式的に解説する．

統計の基本概念とキーワード

基 礎 知 識

母集団と標本 / 母数と統計量



無作為抽出 random sampling

確率抽出 probability sampling とも呼ばれる．母集団の中のどの個体も等しい確率で選ばれるよう計画された標本抽出法．具体的には，乱数表やコンピュータの乱数機能を利用して乱数を得，その数値に対応した対象個体を順に抽出する．

これに対して，無計画な標本抽出は，非確率抽出 non-random sampling または，便宜抽出 convenience sampling と呼ばれ，思わぬところで偏りを生じることが多い．したがって，厳密な意味で，統計（母集団の特徴記述）処理に用いるべきでない．

測定の尺度 scale of measurement

▶**尺　度** scale

対象とする個体や現象の特性に対して数値を割り振ったり分類するための基準．

● 分類尺度 categorical scale

定義：個体をある定性的な特性（属性）によって，数値（コード）や分類名を割り振って分類する場合，そのような測定（分類）基準を分類尺度と呼ぶ．名義尺度 nominal scale ともいう．

例：背番号による個人分類，コードによる病因分類，男女の分類など．

順序尺度との違い：どの個体の分類も平等で，個体の大小関係が定義されないことが，次の順序尺度と異なる．

● 順序尺度 ordinal scale or ranking scale

定義：個体をある量的な特性の順序関係によって，数値（順位 rank）や分類名を割り振って測定（分類）する場合，そのような測定（分類）基準を順序尺度と呼ぶ．

例：治療効果の判定（著効・有効・無効）や，競技の順位（1位，2位，3位）などは順序尺度による測定にあたる．

間隔尺度との違い：個体間の順序（大小）関係のみが定義され，その距離（間隔）は定義されない点が異なる．数値は離散変量 discrete variate でとびとびの値をとる．

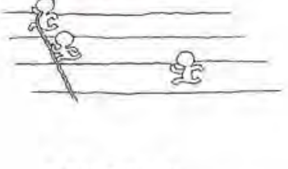
● 間隔尺度 interval scale **と比尺度** ratio scale

定義：ある量的特性を，個体間の順序関係だけでなく，その距離も明確にして測定する場合，その測定基準をさす．数値は，連続変量 continuous variate の形をとることが多い．

例：温度，長さ，重量，時間のような物理的特性の測定など．

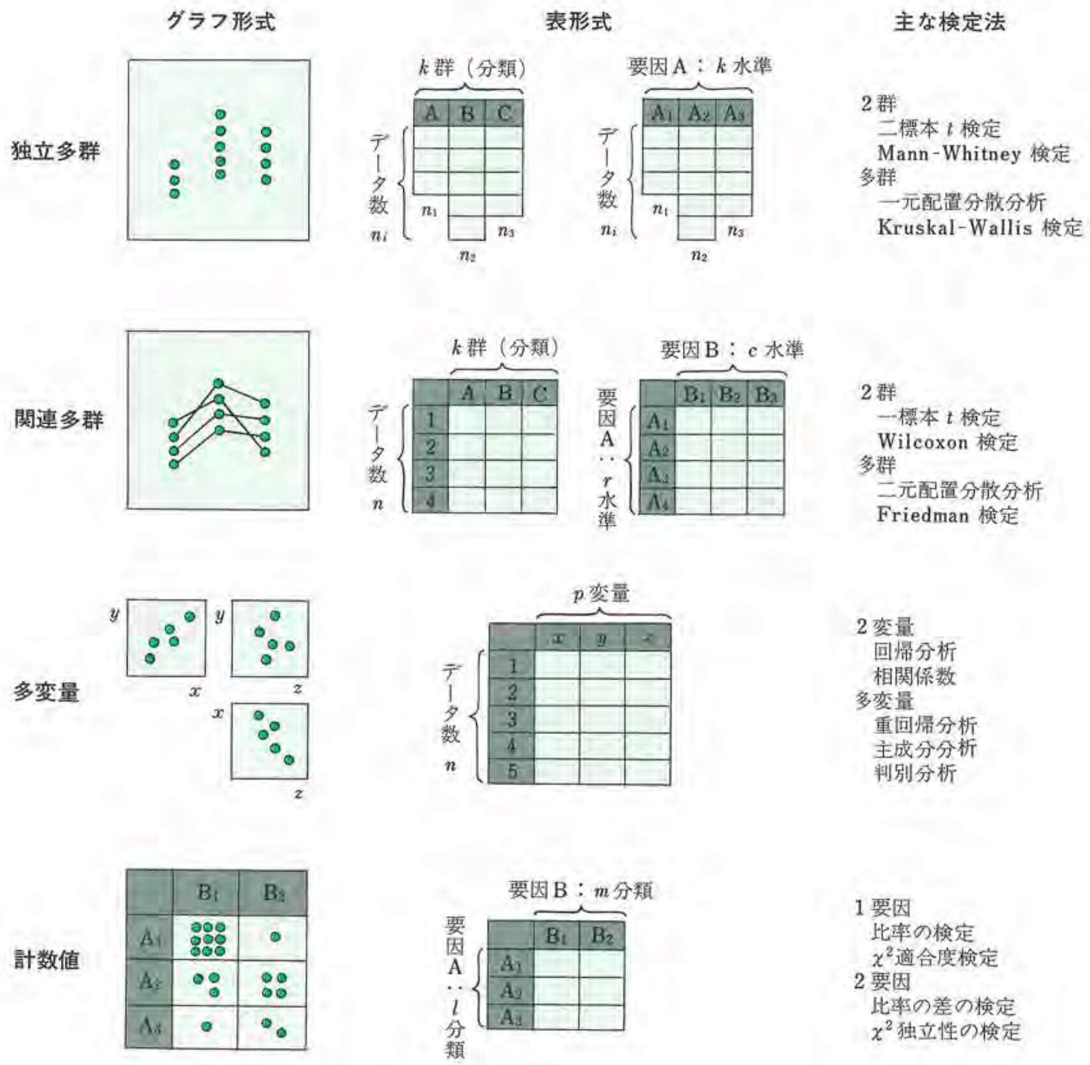
間隔尺度と比尺度の違い：比尺度は，測定値間の距離に加えて，測定値間の比も意味を持つ．たとえば，体重が20 kg から40 kg になるとその比は実感として2倍とわかるが，温度の場合は20℃が40℃になったとしてもその比2は温度が2倍になったことを意味しない（事実同じ温度を摂氏℃から華氏°F に変換するとそれぞれ，68°F と104°F になり，その比は2倍でない）．この意味で，温度は，任意の測定値の比が意味を持たないので間隔尺度であって比尺度でない．これに対して，体重はその比が意味を持つので比尺度にあたる．比が意味を持つかどうかは，別の見方をすると，絶対原点が存在するかどうかにかかっている．すなわち，温度の0℃は水の氷点を任意に0と決めただけで絶対的なものでないが，体重の0 kg は単位をどう変えても，やはり0であることからも，間隔尺度と比尺度の違いがわかる．

両尺度の違いはやや難解であるが，幸い統計学の分野ではその違いを区別する必要がないので，本書では，以下間隔尺度を比尺度も含めたものとして広義に解釈し話を進める．



勝負は順序尺度，
タイムは間隔尺度

統計に用いるデータとその形式



要因と属性 / 水準と分類

要因 factor は調べる数値の変動 variation を説明するための基準．**因子**とも呼ばれる．実験計画では，その意義をいくつかの条件に分けて体系的に調べる．たとえば，化学反応において，温度や pH は反応生成量を変動させる要因にあたる．要因による条件分けは，それが順序尺度的に行われている場合，その各段階を**水準** level と呼ぶ．

属性 attribute は，個体を定性的な条件で分類するための基準．たとえば，性別，居住地域，病因などがそれにあたる．いい換えると，属性は名義尺度的な分類基準にあたり，それによる分画を**カテゴリー（分類）** category または**クラス** class と呼ぶ．

本書では概念を単純化するために，要因を単に条件分けの基準とし，属性と区別せずに用いた．ただし，要因による分け方が順序尺度的な場合には，その分画を水準，名義尺度的な場合には分類またはカテゴリーと呼んだ．なお**群** group は，それを厳密に区別しない場合に，水準や分類に代わる用語として用いた．

パラメトリック検定とノンパラメトリック検定

● **パラメトリック検定** parametric test

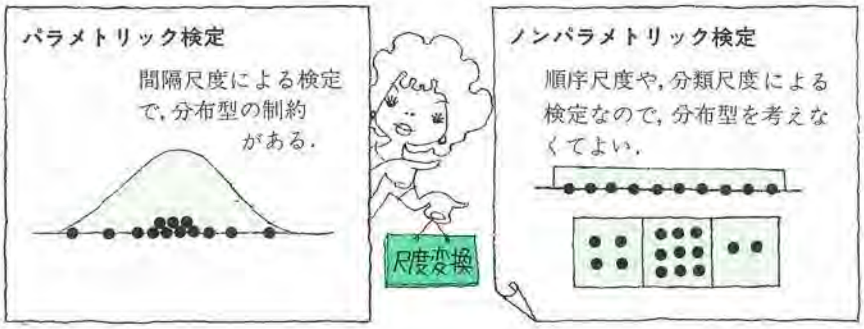
検定法の中で，母集団の分布型に対して一定の仮定をおき，それに基づいて統計的仮説検定を行うものをパラメトリック検定と呼ぶ．たとえば，代表的な一標本または二標本 t 検定，分散分析などがそれにあたり，いずれも母集団が正規分布であることを前提とする．この分布の正規性の仮定は同時に，測定の尺度が間隔尺度でなければならないことを要求する．しかし，少数例の標本しかない場合，正規分布かどうかをはっきりさせにくいので，あまり制限なく利用されることが多い（☞310頁）．

● **ノンパラメトリック検定** non-parametric test

検定法の中で，その適用にあたり母集団の分布型に関して特別の仮定をおく必要がないものをノンパラメトリック検定法または**分布に依存しない検定法** distribution free test と呼ぶ．代表的なものに，Wilcoxon 検定，Mann-Whitney 検定，Friedman 検定，などがある．

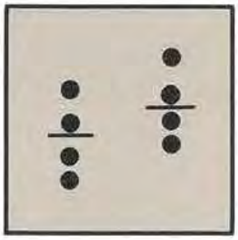
ここで分布型に仮定を置くという概念は，間隔尺度に固有のもので，たとえば順序尺度では，測定の状況によらずその測定値（順位）の分布は一様分布の形をとり，分布型を区別するという概念は存在しない．したがって，ノンパラメトリック検定で分布型の制約がないのは，それがもともと順序尺度や，名義尺度で測定されたデータに対する検定法であるためである．もちろんデータが間隔尺度で測定されていても，それをより低次の尺度に直すので，やはり分布型に依存しない形の検定になる．

なおノンパラメトリック検定法の欠点として，本来パラメトリック検定法が妥当とされる場合に使うと，検定の効率が若干悪くなることがあげられる．しかし，検定法にもよるが，その程度はわずかと考えてよい．



CHAPTER 1

検定の原理

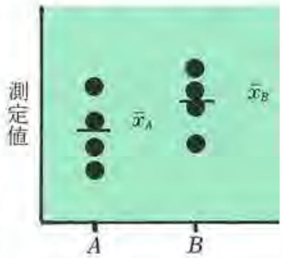


A. 検定法の共通原理

帰無仮説とは...

「差がある」という仮説は，そのままの形では検定できない．

右の図は，2群（A，B）の測定値の比較で，両群の中心位置がずれている．いま，研究者が**直感的にその差を意味のあるものと感じた**とする．しかし，データ数は多くないので，**その差は単なる偶然のずれともいえる．**



検定法は，このような場合に，その**直感に客観性を与える手段**として使われる．

検定にあたって大切なことは，**「差がある」という仮説は，そのままでは検定できない**点である．なぜなら，その差の大きさをあらかじめ特定しておかないと，証明にならないからである．ところが，もともとその大きさがわからないから検定するわけで，論理が噛み合わない．

そこで，**「差がある」という仮説を証明するには，まずその逆の「差がない」という仮説を調べる．**そして，その仮説に何らかの矛盾がみつかれば，もとの「差がある」という仮説を正しいと判断し，逆に，明らかな矛盾がみつからないときは判定を保留する，という形の証明法をとる．

この意味で**検定の論理は反証法**（背理法）にあたり，「差がある」ことを証明するために，その命題を置き換えて「差がないとすると，矛盾する」ことを証明する．

ここで，「差がない」という仮説は，本来無に帰すべきものとして，**「帰無仮説」null hypothesis** と呼び $\mathbf{H_0}$ と略す．また，もとの「差がある」という仮説は，**「対立仮説」alternative hypothesis** と呼び $\mathbf{H_1}$ と略す．

とりあえず帰無仮説を採用して，
差の生じる確率Pを計算

検定のフロー

左図の例の「2群の間に差がある」という仮説の証明は，以下のような手順で行う．

(1) **仮説の設定**：まず「2群に差がない」($A=B$) という仮定 (**帰無仮説**) をおき，本来の「差がある」($A \neq B$) という仮説 (**対立仮説**) はいったん伏せておく．

(2) **統計量の計算**：検定しやすいように，データを一つの数値に要約する．この要約値を**統計量** (正確には，**検定統計量 test statistic**) と呼ぶ．たとえば，この例では2群の平均値 $\bar{x}_A$，$\bar{x}_B$ を求め，その平均値の差 $\bar{x}_A - \bar{x}_B$ を統計量 X とする．この統計量は2群の差を表す数値であれば何でもよく，**検定法の違いは統計量の違い**によっている．

(3) **確率を求める (統計量の偏り度を確率 P で表す)**：帰無仮説から，統計量 X (平均値の差) の期待値は0であるが，実際にはある大きさをもっている．そこで確率的に，その差が十分起こりうるものかどうかを調べる．そのためには，統計量 X の理論分布が必要となるが，実際には統計数値表から統計量 X の生じる確率が求まる．

設　問

差があるのでは？

↓

(1) 仮説の設定

差がない　H_0：帰無仮説
(差がある　H_1：対立仮説)

↓

(2) 統計量 X を求める

検定しやすいように，データを一つの数値に要約．

↓

(3) 確率を求める

H_0 のとき，統計量 X が生じる確率 P を求める．

↓

(4) 判　定

P が有意水準 α より大
→H_0 を棄却できない．(判定を保留)
P が有意水準 α より小
→H_0 を棄却し，H_1 を採用．

(4) **判　定**：統計量 X の生じる確率 P が，有意水準 α よりも大きいとき，その程度の差は十分起こりうるので，帰無仮説を棄却できない (☞次頁)．

逆に，P が有意水準 α より小さいときは，そんな確率の低いことが実際に起こったと考えるよりは，むしろ「差がない」という仮説の方がおかしいと判断する．よって「差がある」とする，対立仮説の方を採用する．

有意水準

level of significance
帰無仮説が正しいのに，それを誤って否定する確率で α で表す．危険率とも呼ぶ．バイオサイエンスの分野では，$\alpha=0.05$ (5%) が用いられる．

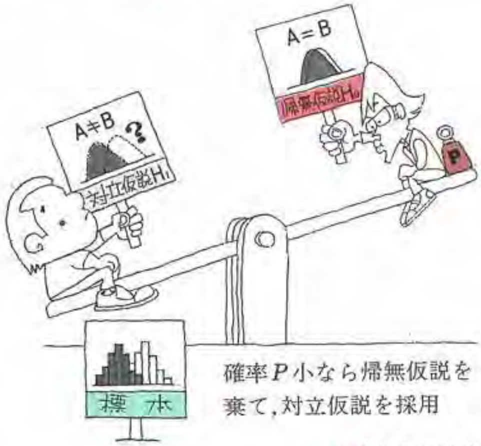
Key Point

検定の基本は，反証の論理

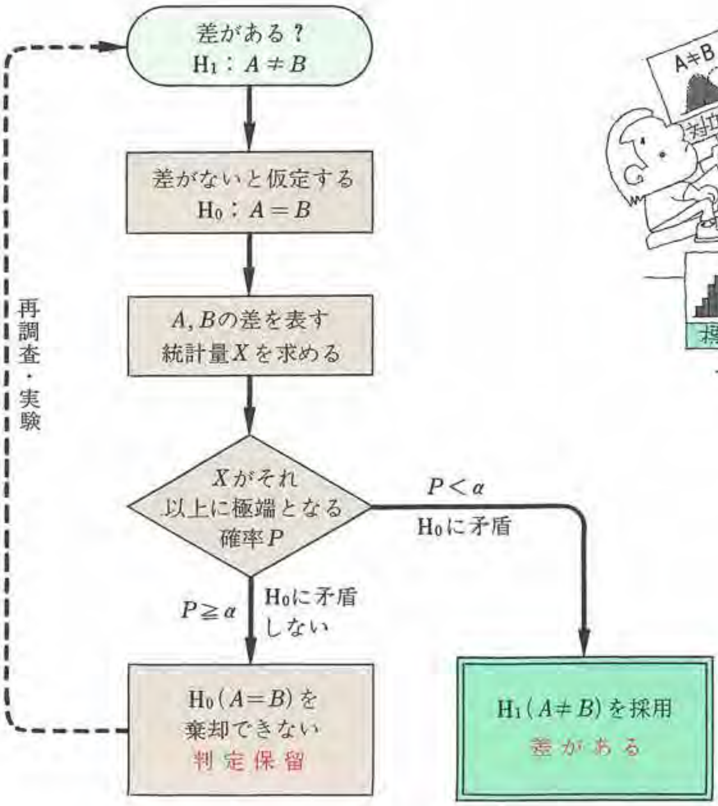
「差がある，$A \neq B$」という仮説の証明 → 「差がない，$A=B$」という仮説が確率論的に矛盾することを証明

確率P大なら帰無仮説を棄てられない

まとめ
検定の原理



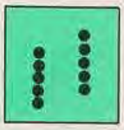
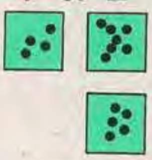
確率P小なら帰無仮説を棄て，対立仮説を採用



▶統計量の生じる確率Pが有意水準αより大きいとき，「差がない」ことを証明したわけではなく，**判定を保留**したことになっている．なぜなら，データ数が増えると検定法の検出力が高まり，より小さな差でも検出できる可能性があるからである．

したがって，この場合の表現は，**「差があるとはいえなかった」「有意な差は見出しえなかった」**とするのが正しい（☞316頁）．

本書で取り扱った検定法一覧

データ形式	間 隔 尺 度	順 序 尺 度	分 類 尺 度
1 標 本 	平均値の検定 $\bar{x} \to z$ 分布の歪度 b_1 尖度 b_2		比率の検定 二項検定 $r \to z$ ポアソン検定 $r \to z$ χ^2 適合度検定 χ^2
関連 2 標本 	一標本 t 検定 $\bar{d} \to t$	Wilcoxon 符号付き順位和検定 $T \to z$	符号検定 $r \to z$
独立 2 標本 	二標本 t 検定 $\bar{x}_1 - \bar{x}_2 \to t$ 等分散の検定 $\frac{s_1^2}{s_2^2} = F$	Mann–Whitney 検定 $U \to z$	2×2 分割表 χ^2 独立性の検定 χ^2 Fisher の直接確率計算法 比率の差の検定 $\hat{p}_1 - \hat{p}_2 \to z$
独立多標本 	一元配置分散分析 $\frac{s_A^2}{s_E^2} = F$ Bartlett 検定 $\frac{M}{C} \to \chi^2$	Kruskal–Wallis 検定 $H \to \chi^2$	$l \times m$ 分割表 χ^2 独立性の検定 χ^2
関連多標本 	二元配置分散分析 $\frac{s_A^2}{s_E^2} = F_A$, $\frac{s_B^2}{s_E^2} = F_B$ Bartlett 検定 $\frac{M}{C} \to \chi^2$	Friedman 検定 $\chi^2_r \to \chi^2$ Kendall の一致係数 W	
2 変 量 	回帰係数の検定 $a \to t$, $b \to t$ 相関係数 r 重相関係数 R	Spearman 順位相関係数 r_S	ϕ 係数 ϕ クラメールの C 係数 C ユールの連関係数 Q
多 変 量 	偏相関係数 $r_{xy \cdot z}$ 重相関係数 R	Kendall の一致係数 W	

検定法名の右に検定統計量を示す．→の右は，その近似値（標準化後）．

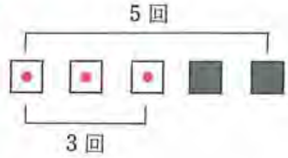
B. 身近な例にみる検定法の原理

比率の検定（二項検定）

test for the proportion

さいころは偏っていると判断してよいか？

比率の検定



例 1-1 さいろころを 5 回投げたところ，1 の目が 3 回でた．
このさいころは偏っていると考えてよいか．

解

さいころの 1 の目のでる確率 p（母比率*1）は1/6で，直感的に 1 の目がですぎと思われる．はたして客観的にもそういえるのか，統計的に検定する．

(1) 仮説の設定：「さいころが偏っている」という仮説を検定したいが，その偏りがどの程度か特定しにくい．そこで，その逆の「さいころは偏っていない」という仮説（帰無仮説：$p=1/6$）をとりあえず採用し，もとの仮説（対立仮説：$p>1/6$）はいったん伏せておく．

(2) 統計量を求める：データを要約して，偏りを表すような統計量を求めるが，この場合単純に **1 の目のでた回数 r を統計量**とすればよい．

帰無仮説が正しいとすると，さいころを投げた回数 $n=5$ に対する，r の期待値は，$np=5\times 1/6=0.83$回となるが，この例では r の実現値は 3 回と比較的大きな値となっている．

(3) 確率を求める（統計量 r を確率で表す）：r が 3 回またはそれ以上*2の極端な値となる確率 P は，次の式により計算できる*3．

検定のフロー

設　問
さいころは偏っている？ 仮説 H_1　$p\neq 1/6$

↓

(1)　仮説の設定
いったんさいころは正しいと仮定する． 帰無仮説 H_0　$p=1/6$ （H_1 は対立仮説として保留）

↓

(2)　統計量を求める
1 の目のでた回数 r を統計量とする． $r=3$

↓

(3)　確率を求める
仮説 H_0 によって 1 の目が 3 回以上でる確率 P を求める．　$P=P_3+P_4+P_5$

↓

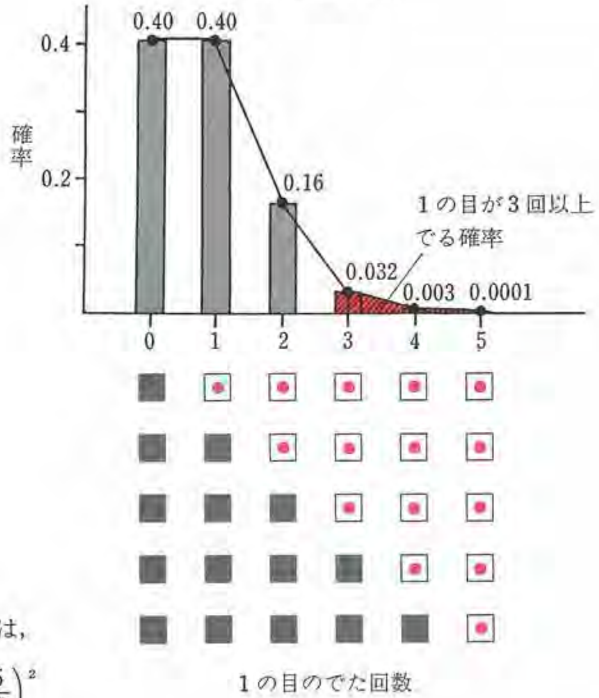
(4)　判　定
$P\geqq 0.05$　仮説 H_0 を棄却できない．（判定保留） $P<0.05$　仮説 H_0 を棄却し，仮説 H_1 を採用．

*1 真の比率を**母比率**と呼ぶ．これに対してある一連の試行で得た比率を**標本比率**と呼ぶ．(☞116頁)
この場合の標本比率は3/5なので，それが母比率を1/6と仮定して十分起こりうる値かどうかを検定する．

*2 以下本書では 3 回**またはそれ以上**を 3 回**以上**と書く．

*3 専門の統計数値表の二項分布表を見ると，n, p, r から，r に対する確率 P が求まる．

1の目のでる回数の理論分布
（試行回数$n=5$）



1の目が5回のうち3回でる確率は，

$$P_3 = {}_5C_3 \times \left(\frac{1}{6}\right)^3 \times \left(\frac{5}{6}\right)^2$$

で，一般に，1の目がn回のうちr回でる確率は，

$$P_r = {}_nC_r \times \left(\frac{1}{6}\right)^r \times \left(\frac{5}{6}\right)^{n-r}$$

である．したがって

1の目が3回でる確率：$P_3 = {}_5C_3 \times \left(\frac{1}{6}\right)^3 \times \left(\frac{5}{6}\right)^2 = 0.032$

1の目が4回でる確率：$P_4 = {}_5C_4 \times \left(\frac{1}{6}\right)^4 \times \left(\frac{5}{6}\right)^1 = 0.003$

1の目が5回でる確率：$P_5 = {}_5C_5 \times \left(\frac{1}{6}\right)^5 \times \left(\frac{5}{6}\right)^0 = 0.0001$

∴　1の目が3回以上でる確率Pは，

$$P = P_3 + P_4 + P_5 = 0.0351 < 0.05$$

(4) 判　定：1の目が3回以上でる確率Pは，有意水準の0.05（5％）より小さい．このようなまれなことが実際に起こったと考えるよりは，むしろ計算の前提となった帰無仮説（1の目のでる確率は1/6）の方がおかしいと考えるのが妥当．したがって，帰無仮説を棄却し，さいころが偏っているとする対立仮説を採用する．

←ある事象の生じる比率がpのとき，n回試行してr回その事象が生じる確率は，

$$P = {}_nC_r \times p^r \times (1-p)^{n-r}$$

で求まる．これを**二項確率**と呼ぶ（☞117頁）．

${}_nC_r$

${}_nC_r$は異なるn個からr個とりだす場合の組合せの数で

$${}_nC_r = \frac{n!}{r!(n-r)!}$$

（例）

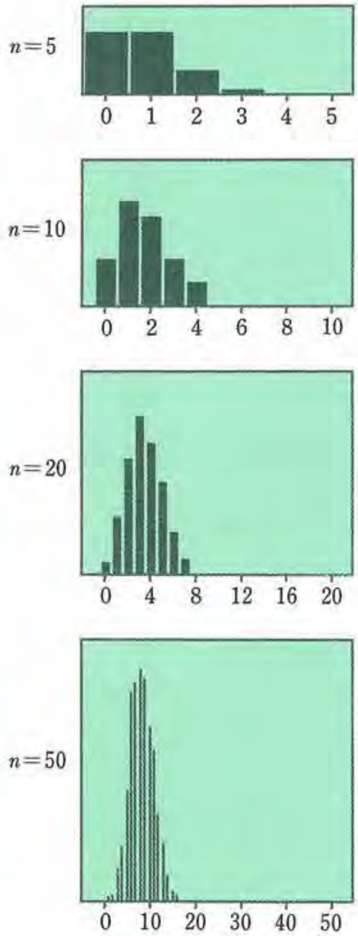
$${}_5C_2 = \frac{5!}{2!3!} = \frac{5 \cdot 4 \cdot \cancel{3!}}{2 \cdot 1 \cdot \cancel{3!}} = 10$$

イメージレッスン――統計量の分布を考える

試行回数 n と二項分布

$\left(p=\dfrac{1}{6}\text{の場合}\right)$

試行回数 n を 5，10，20，50，と増やしてゆくと，1 の目のでる回数の分布は次第に正規分布に近似してゆく（参考☞120頁）．



予 備 知 識

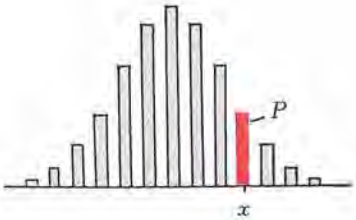
1の目が3回でる確率ではなく，3回以上でる確率とする理由

ある統計量の有意性（期待値からの偏りの程度）を確率で示す場合，(a)その値だけに限定した確率を計算すると，離散変量でかつ「場合の数」が多いと確率が小さくなってしまう．また連続変量の場合には，その区間を限定しないと計算できない．

これに対し，(b) 与えられた統計量がその値以上となる確率（確率密度曲線の裾の面積）として求めると，その確率と統計量の偏りがよく一致する．

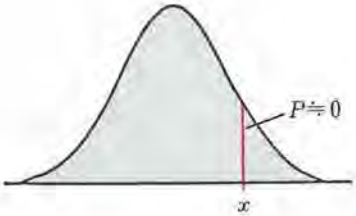
統計量の有意性（偏り）を示す確率

× (a) 統計量が x となる確率



離散変量の場合

統計量のとりうる値（場合の数）が増えるにつれ，x の確率 P は限りなく 0 に近づく（区間確率）．

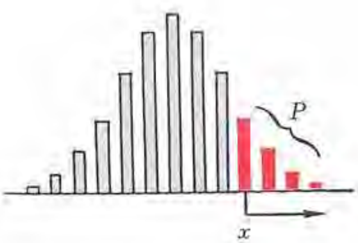


連続変量の場合

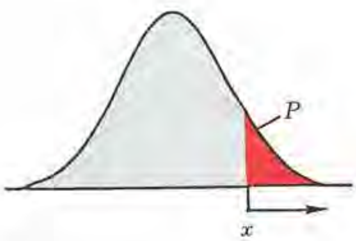
連続変量では区間を指定しないと確率 P は常に 0 となる（点確率）．

検定では，統計量の生じる確率を求める必要があるが，区間確率や点確率だと，それを統計量の偏りの指標にできない．

○ (b) 統計量が x 以上となる確率（すそ側の確率）



離散変量の場合



連続変量の場合

どちらの場合も，統計量が中心から偏るほど，確率(理論分布のすそ側の面積)が小さくなる．したがって，このすそ側の確率は，統計量の偏りの標準的な指標になる．

平均値の検定

test for the mean

あるクラスの身長の分布は，全国水準と違う？

平均値の検定

例 1-2 全国の10歳女子の身長は平均値 μ が140 cm，標準偏差 σ が 5 cm の正規分布をする．いま，あるクラス25人の身長の分布は下図のようになり，その平均値 $\bar{x}$ は，137 cm であった．このクラスの身長は全国水準と違うといえるか．

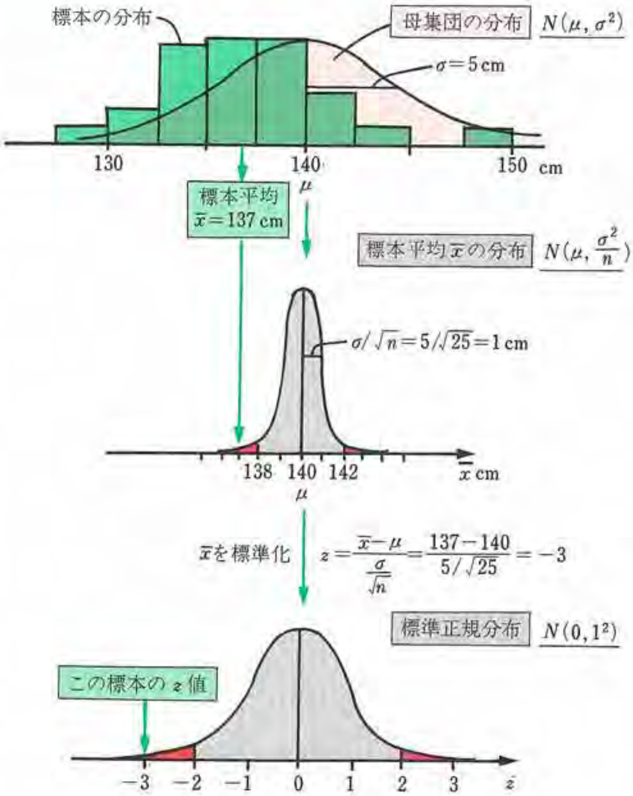
解

右図から，このクラスの身長の分布は，全国水準（母平均 μ）からずれているようにみえる．しかし，客観的にもそういえるか検定する．

(1) 仮説の設定：「全国水準と違う」という仮説は，違いの程度を特定しないと検定できない．そこで，その逆の「全国水準と同じ」という仮説（帰無仮説 H_0：$\mu=140$）をとりあえず採用し，もとの仮説（対立仮説 H_1:$\mu\neq 140$）はいったん伏せておく．

(2) 統計量を求める：データを要約して，偏りを表すような統計量を求める．この場合25人の身長の平均値 $\bar{x}$ を統計量とするのが一般的．

帰無仮説が正しいとすると，$\bar{x}$ の期待値は $\mu=140$ cm であるが，この標本の $\bar{x}$ は137 cm と偏っている．この偏りが帰無仮説のもとで，どの程度の確率で生じるかを求める．



(3) 確率を求める（統計量 $\bar{x}$ の偏り度を確率で表す）：$\bar{x}$ の偏り度を求めるためには，帰無仮説が正しいとしたときの標本平均 $\bar{x}$ の理論分布を知る必要がある．

24頁のイメージレッスンに示すごとく，その分布は，平均値 μ，標準偏差 $\frac{\sigma}{\sqrt{n}}$ の正規分布となる．

$N(\mu, \sigma^2)$

平均値 μ，分散 σ^2 の正規分布を表す．

したがってこの標本平均$\bar{x}$の偏り度を表す確率Pは，$\bar{x}$を

$$z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

と標準化（☞25頁）して，正規分布表から求まる．

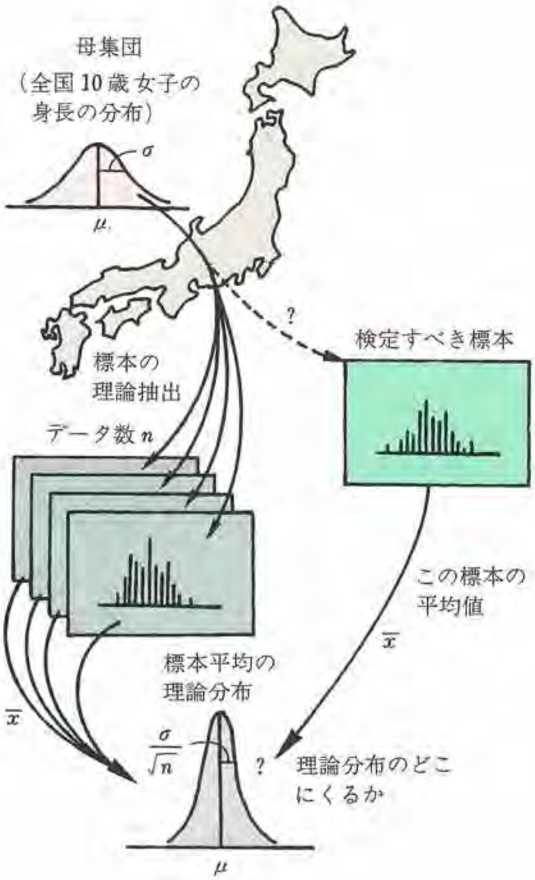
すなわち$z=-3.0$となり，表からPは0.0027と求まる．

⑷ 判　定：

標本平均がその期待値から3 cm以上ずれる確率は，$P=0.0027$と小さい．このようなまれな確率が実際に起こったと考えるよりは，帰無仮説（H_0）が正しくなかったと考える方が妥当．したがって，H_0を棄却して，対立仮説H_1を採用する．

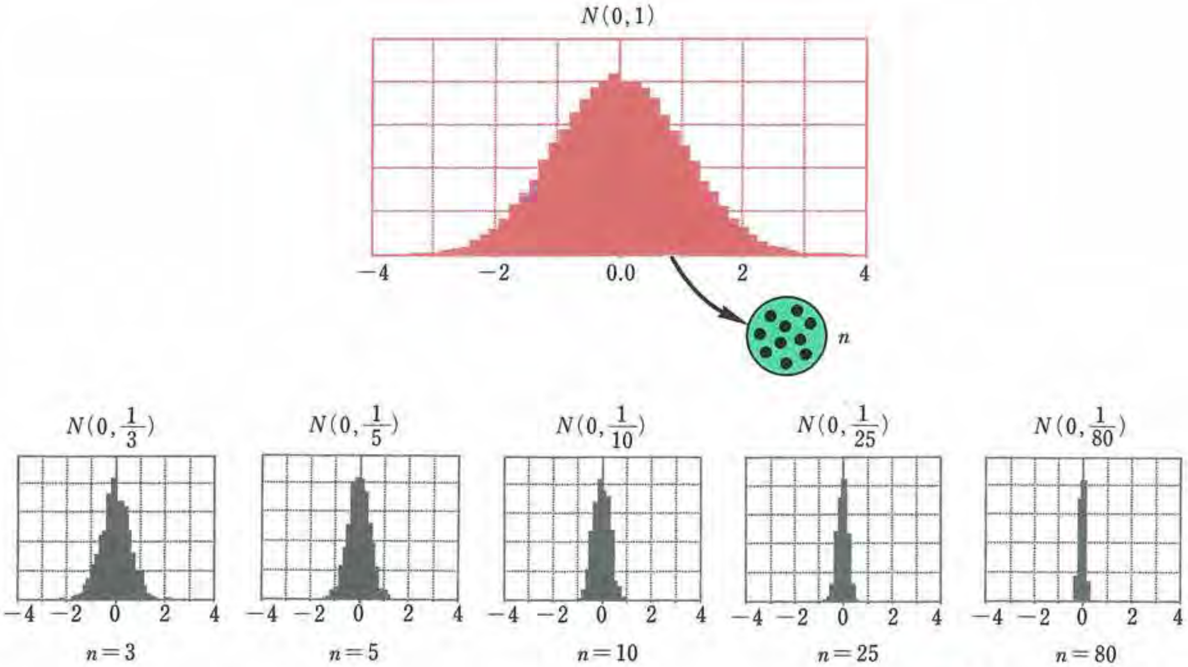
検定の概念

- 母集団から標本平均の理論分布を求める．
- 検定すべき標本の平均値が，その理論分布のどこにくるか．



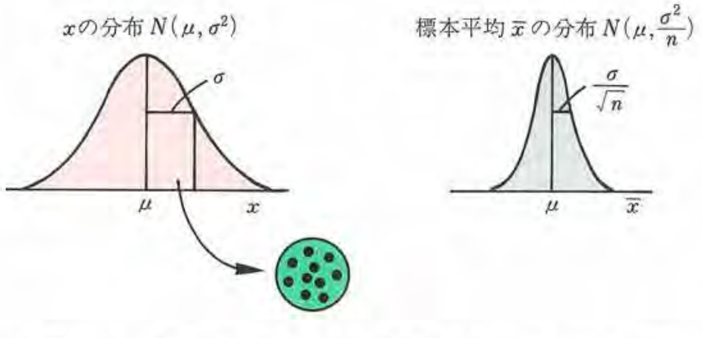
イメージレッスン ── 統計量の分布を考える

標本平均の理論分布



▲平均値 $\mu=0$ 標準偏差 $\sigma=1$ の正規母集団から，データ数 $n=3$，5，10，25，80の標本を取り出し，その平均値 $\bar{x}$ の理論分布を調べた．

Key Point



正規分布 $N(\mu,\sigma^2)$ から，データ数 n の標本をとると，標本平均 $\bar{x}$ は，正規分布 $N(\mu,\frac{\sigma^2}{n})$ に従う．

平均値の標準誤差

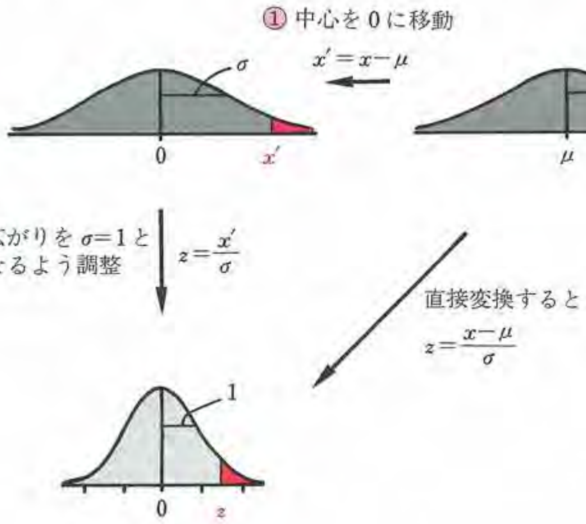
標本平均 $\bar{x}$ の分布の標準偏差 $\sigma/\sqrt{n}$ を**平均値の標準誤差** standard error of mean (SEM) と呼ぶ．

一般にはこれを単に"標準誤差 (SE)"と呼ぶことが多いが，**標準誤差は，統計量の分布の標準偏差**として，他の統計量にも広く用いるので，"平均値の標準誤差"と区別していった方がよい．

データの標準化

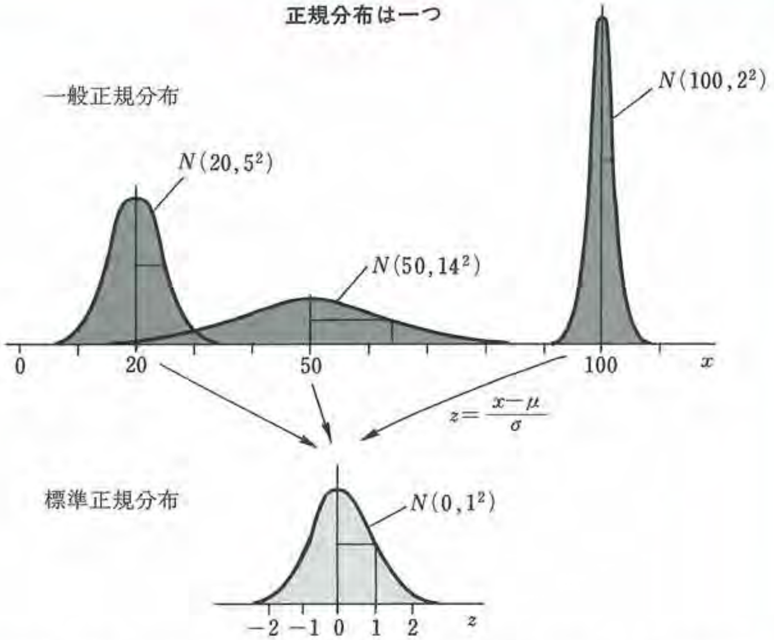
基礎知識

一般正規分布 $N(\mu,\ \sigma^2)$ の値 x を，標準正規分布 $N(0,\ 1^2)$ の値 z へ変換



Key Point

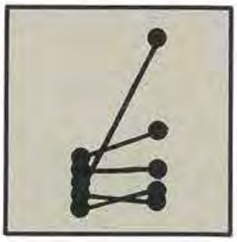
正規分布は一つ



どんな正規分布も，データをその平均値 μ と標準偏差 σ で標準化すれば，単一の分布．

CHAPTER 2

関連 2 群の差の検定



一標本 t 検定（パラメトリック法）

one sample t test; paired t test

検定の手順

“2群間に差がある”といってよいか？

(1) **仮説の設定**：いったん，“差がない”と仮定する（帰無仮説：H_0）．
“差がある”という仮定は対立仮説（H_1）として保留．

(2) **統計量を求める**：n 組のデータにつき各ペアの差 d を求め，その平均値 $\bar{d}$ を統計量とする（$\bar{d}$ に2群の差が要約されている）．

(3) **確率を求める**：$\bar{d}$ が確率的にどの程度偏った値であるかを求める．これには $\bar{d}$ をその標準誤差（$s_d/\sqrt{n}$）でわって標準化すると，その値 t が自由度 $n-1$ の t 分布することを利用する（☞36頁）．

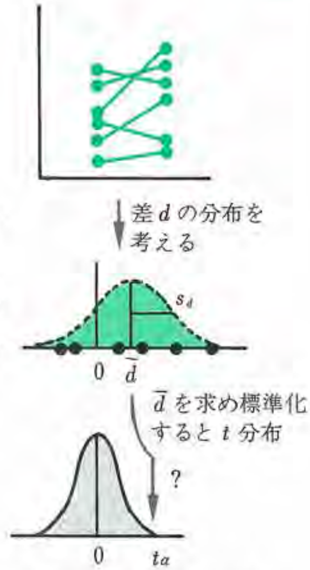
$$\boxed{\bar{d}} \xrightarrow[\text{標準化}]{} \boxed{t=\frac{\bar{d}}{\dfrac{s_d}{\sqrt{n}}}} \quad t\text{ は自由度 } n-1 \text{ の } t \text{ 分布}$$

ここに s_d は差 d の標準偏差で

$$s_d=\sqrt{\frac{\Sigma(d_i-\bar{d})^2}{n-1}}$$

である．この t 値（標準化された $\bar{d}$ 値）の生じる確率は，t 分布表から求まる．

(4) **判　定**：有意水準 α，自由度 $n-1$ の t 値（t_α）*を t 分布表（付表2）から求め，この標本の t 値と比較する．
$|t| \leqq t_\alpha$ のとき，差があるとはいえない（H_0 を棄却できず判定保留）．
$|t| > t_\alpha$ のとき，$P<\alpha$ となり有意差あり（H_0 を棄却し，対立仮説を採用）．

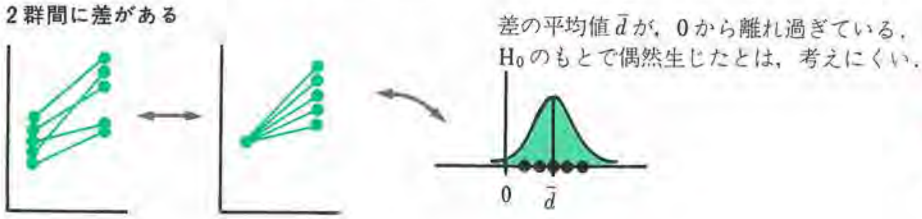


t 分布表から有意性の判定

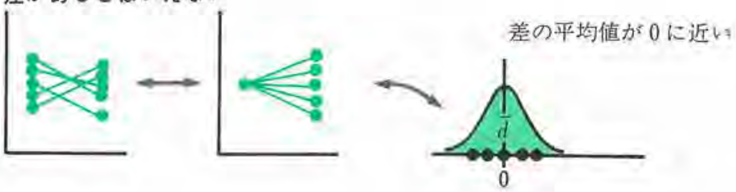
* 自由度df，有意水準 α の t 値を t(df; α) と表す．t_α はその省略型．

検定の概念　　**一標本 t 検定と統計量 $\bar{d}$**（$\bar{d}$：各ペアの差の平均値）

2群間に差がある



差があるとはいえない



▶**差の検出力**
差の分布型が正規分布のとき，Wilcoxon検定や符号検定より差の検出力（検定効率）が高い．しかし，その分布型が非正規分布のときや，飛び離れ点（極端にきわだった差）があるとき，検定の妥当性が失われ，かつ検定効率も低下する．（☞61～65頁）

2 群の差の検定：チェックポイント

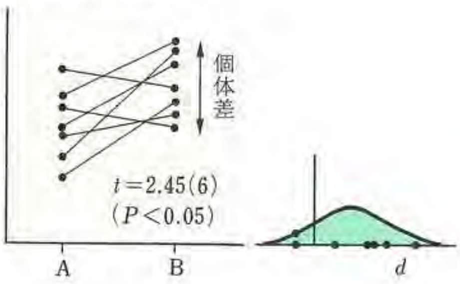
関連 2 群として比較することが有用な場合とは...

*1 longitudinal study, serial study

*2 cross-sectional study

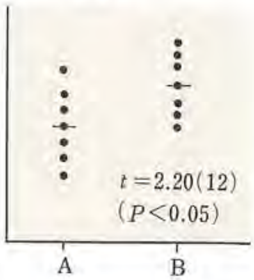
関連 2 群

・同一個体で 2 条件比較
・縦断的研究*1



独立 2 群

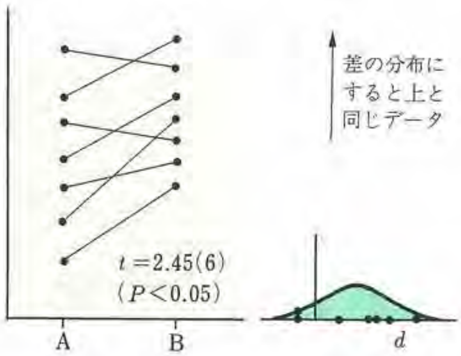
・異なる個体で 2 条件比較
・横断的研究*2

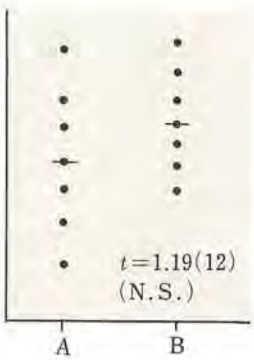


◀上・中・下段とも関連 2 群と独立 2 群の点の配置は，まったく同じにして，それぞれ一標本 t 検定，二標本 t 検定をあてはめ t 値（自由度）とその確率を比較した．上，中段の関連 2 群のデータは，点の位置を個体単位に y 軸にそってずらしただけで，差 d の分布（中央の分布図）にすると同一データ．

（1）　2 条件（A, B）に差はあるが，個体差大

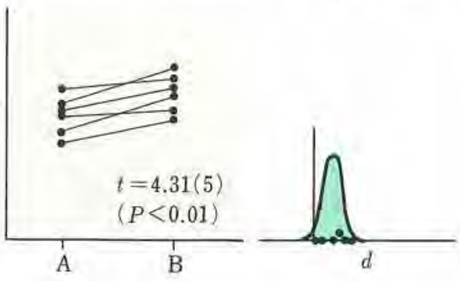
個体差
大

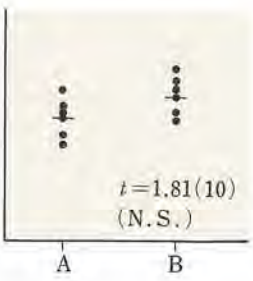




（2）　個体差小さいが，2 条件（A, B）の差もわずか

個体差
小



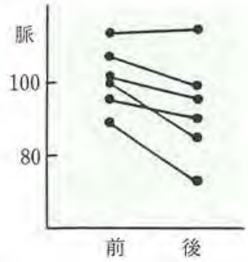


▶ただし……

条件差≫個体差またはデータ数十分大なら，関連 2 群も独立 2 群も実質的には同じ結論に．

一標本 t 検定

例 2-1 6人のバセドウ病患者に自律神経遮断剤（A剤）を投与し，前後の脈拍数を計測した．A剤には効果があるといってよいか．



脈拍(前)	脈拍(後)	d	d^2
98	86	12	144
88	73	15	225
100	95	5	25
96	92	4	16
107	99	8	64
114	116	−2	4
		42	478

検定のフロー

設　問

A剤により脈拍が変化したのでは？
H_1：脈拍(前)≠脈拍(後)

↓

(1) 仮説の設定

いったん，脈拍は変化しないと仮定する．
帰無仮説 H_0：脈拍(前)＝脈拍(後)
設問の H_1 は対立仮説として保留．

↓

(2) 統計量を求める

差 d の平均値 $\bar{d}$ に，2群の差が要約されていると考える．
$\bar{d}=7$

↓

(3) 確率を求める

帰無仮説 H_0 に基づく $\bar{d}$ の理論分布から，$|\bar{d}|\geqq 7$ となる確率 P を求める．
これには，$\bar{d}$ を標準化して，t 分布表（自由度 $n-1$）を利用する．

↓

(4) 判　定

$P\geqq\alpha$　H_0 を棄却できない．
$P<\alpha$　H_0 を棄却し，H_1 を採用．

解

差の平均値 $\bar{d}$，標準偏差 s_d を求める：

$$\bar{d}=\frac{42}{6}=7$$

$$s_d=\sqrt{\frac{\Sigma(d_i-\bar{d})^2}{n-1}}$$

$$=\sqrt{\frac{\Sigma d_i{}^2-(\Sigma d_i)^2/n}{n-1}}$$

$$=6.07$$

◀公式
$\Sigma(d_i-\bar{d})^2$
$=\Sigma d_i{}^2-\dfrac{(\Sigma d_i)^2}{n}$

$\bar{d}$ をその標準誤差 $s_d/\sqrt{n}$ で標準化する：

$$t=\frac{\bar{d}-0}{\frac{s_d}{\sqrt{n}}}=\frac{7}{\frac{6.07}{\sqrt{6}}}=2.82$$

t 分布表から t 値（標準化された $\bar{d}$ 値）の確率を求め判定（☞右頁）．

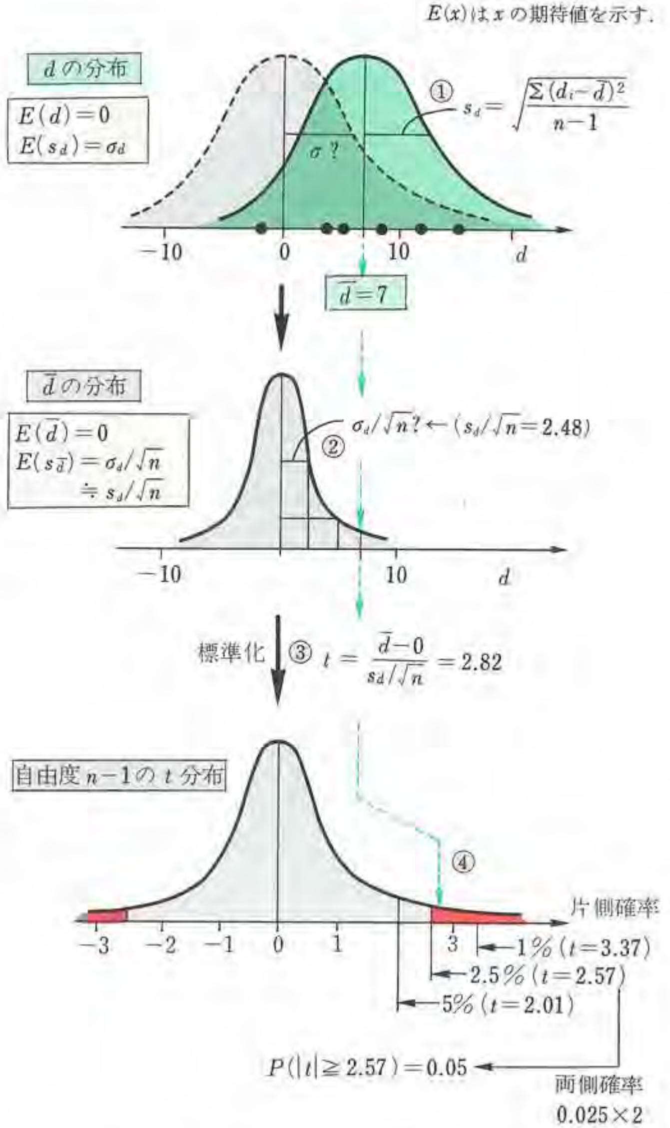
考え方：差がないという仮説（帰無仮説：H_0）をおくと，差の平均値 $\bar{d}$ の期待値は 0，しかしこの標本から実際に求めた $\bar{d}$ は 7．$\bar{d}$ 値が 7 以上（$|\bar{d}| \geqq 7$）となる確率を求め，仮説 H_0 の妥当性を判定する．

① 仮説 H_0 から d の理論分布を想定すると，その中心位置は 0．
しかし，その標準偏差は不明なので，実測値 d の標準偏差 s_d で代用．

② 差の平均値 $\bar{d}$ の理論分布の中心位置は 0．その標準誤差は不明なので s_d の $1/\sqrt{n}$ で代用．

③ この標本の $\bar{d}$ 値が，$\bar{d}$ の理論分布からみて確率的にどの程度偏っているかを知りたい．
これには $\bar{d}$ 値をその標準誤差 $s_d/\sqrt{n}$ で標準化すると，その値 t は自由度 $n-1$ の t 分布に従うことを利用する．

④ この標準化により $\bar{d}=7$ は，$t=2.82$ となり，$|\bar{d}| \geqq 7$ となる確率は $|t| \geqq 2.82$ の確率を求めることと同値．



判　定：t 分布表（付表 2）より，自由度 $n-1$，有意水準の確率 0.05 の t 値は 2.57.

$$\therefore \quad P = P(\underbrace{|\bar{d}| \geqq 7}_{\text{標準化}}) = P(|t| \geqq 2.82) < \underset{\uparrow\ t\text{ 分布表}}{P}(|t| \geqq 2.57) = 0.05$$

$\therefore\ P < 0.05$

このような小さな確率の現象が実際に起こったと考えるよりは，もとの仮説（帰無仮説：H_0）がおかしかったと考える方が妥当．

したがって，H_0 を棄却し，対立仮説 H_1（A 剤投与前後で脈拍が変化したとする仮説）を採用する．

標本の“標準偏差”(s_n)は，母標準偏差σより小さく計算される?!

予 備 知 識

標準偏差は，個々の点xが母集団の平均値μから平均的にどの程度隔たっているかを表し，

$$標準偏差 = \sqrt{\frac{\Sigma(x_i-\mu)^2}{n}} \qquad (i=1, 2, \cdots\cdots, n)$$

と定義される．その母数は，**母標準偏差**[*1]と呼ばれσで表す．

いま，母集団が未知のとき，標本から標準偏差s_nを計算しようとすると，μが不明なので，それを標本平均$\bar{x}$

$$\bar{x} = \frac{\Sigma x_i}{n}$$

で置き換えて，次式のように計算するのが自然である．

$$s_n = \sqrt{\frac{偏差平方和}{データ数}} = \sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n}}$$

この式は一見正しいようだが，実はs_nの期待値は，母標準偏差σより小さく計算されてしまう．

その理由は以下に説明するとおりであるが，結論を先にいうと，この計算上の偏りを是正するには，上式の偏差平方和をnでなく，$n-1$でわって，

$$s = \sqrt{\frac{偏差平方和}{データ数-1}} = \sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}}$$

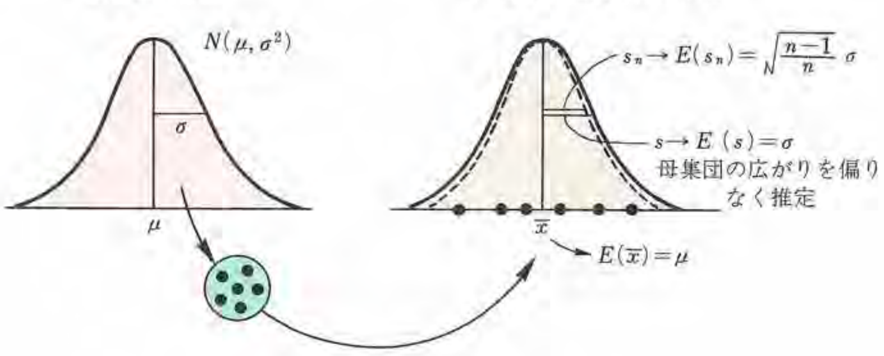
と計算する必要がある．このsは**標本標準偏差**[*2]と呼ばれ，母標準偏差σの偏りない推定値になる．

この意味で，sを2乗した値（標本分散），

$$s^2 = \frac{\Sigma(x_i-\bar{x})^2}{n-1}$$

は，特に**不偏分散**[*3]と呼ばれる．

以下本書では，とくに断らない限り，“標準偏差”をsで表し，標本標準偏差（不偏分散の平方根）を指すものとする．事実，バイオサイエンスの統計では，母集団が未知の場合がほとんどであるから，標準偏差をs_nでなく，sの式で計算する必要がある．ただし，標本のデータ数が大きくなると，$s_n \fallingdotseq s$となるので，両者の区別は重要でなくなる．



[*1] population standard deviation

[*2] sample standard deviation

[*3] unbiased variance

標本標準偏差 s_n が，母標準偏差 σ と一致しない理由

〈概念的説明〉

計算の基礎となる母平均 μ を，標本平均 $\bar{x}$ で置き換えていることによる．すなわち，偏差平方和を計算する基準点は，本来母集団の中心 μ でなければならないが，それを偏差平方和を最小にする標本の中心 $\bar{x}$ に設定するため，s_n は常に σ より小さく計算される．

それを是正し，σ の偏りのない推定値を得るには，偏差平方和 $\Sigma(x_i-\bar{x})^2$ を n でなく，$n-1$ でわるとちょうどよいことが，下記の数学的理論からわかっている．

〈数理的説明〉

$s_n^2=\dfrac{\Sigma\,(x_i-\bar{x})^2}{n}$ と $\sigma^2=\dfrac{\Sigma\,(x_i-\mu)^2}{n}$ の差を調べるため，

s_n^2 の分子を変形すると，

$$\begin{aligned}\Sigma\,(x_i-\bar{x})^2 &= \Sigma\,(x_i-\mu+\mu-\bar{x})^2\\ &= \Sigma\,\{(x_i-\mu)^2-2\,(x_i-\mu)\,(\bar{x}-\mu)+(\bar{x}-\mu)^2\}\\ &= \Sigma\,(x_i-\mu)^2-2\,\boxed{\Sigma\,(x_i-\mu)\,(\bar{x}-\mu)}^{*1}+\Sigma\,(\bar{x}-\mu)^2\\ &= \Sigma\,(x_i-\mu)^2-2\,\boxed{n\,(\bar{x}-\mu)^2}+n\,(\bar{x}-\mu)^2\\ &= \Sigma\,(x_i-\mu)^2-n\,(\bar{x}-\mu)^2\end{aligned}$$

となる．これから，

$$\begin{aligned}s_n^2 &= \frac{\Sigma\,(x_i-\bar{x})^2}{n}=\frac{\Sigma\,(x_i-\mu)^2-n\,(\bar{x}-\mu)^2}{n}\\ &= \frac{\Sigma\,(x_i-\mu)^2}{n}-(\bar{x}-\mu)^2=\sigma^2-(\bar{x}-\mu)^2\end{aligned}$$

より，s_n^2 は，σ^2 より，その差 $(\bar{x}-\mu)^2$ の分だけ小さいことがわかる．

ここで，標本の分散 s_n^2 の期待値を求めると，

$$\begin{aligned}E(s_n^2) &= E\left\{\frac{\Sigma\,(x_i-\bar{x})^2}{n}\right\}\\ &= \sigma^2-\boxed{E\{(\bar{x}-\mu)^2\}}^{*2}\\ &= \sigma^2-\frac{\sigma^2}{n}=\frac{n-1}{n}\,\sigma^2\end{aligned}$$

となり，標本の分散 s_n^2 は，母分散 σ^2 の $(n-1)/n$ になることがわかる．

したがって，偏差平方和を $n-1$ でわると，母分散の偏りない推定値が得られ，これが不偏分散 s^2 にあたる．

*1
$$\begin{aligned}&\Sigma\,(x_i-\mu)\,(\bar{x}-\mu)\\ &=(\bar{x}-\mu)\,\Sigma\,(x_i-\mu)\\ &=(\bar{x}-\mu)\,(\Sigma\,x_i-\Sigma\,\mu)\leftarrow\boxed{\bar{x}=\frac{\Sigma\,x_i}{n}}\\ &=(\bar{x}-\mu)\,(n\bar{x}-n\mu)\\ &=n\,(\bar{x}-\mu)^2\end{aligned}$$

*2 $E\{(\bar{x}-\mu)^2\}=\dfrac{\sigma^2}{n}$

↑

$(\bar{x}-\mu)^2$ の期待値は，（平均値の標準誤差）^{2} に相当

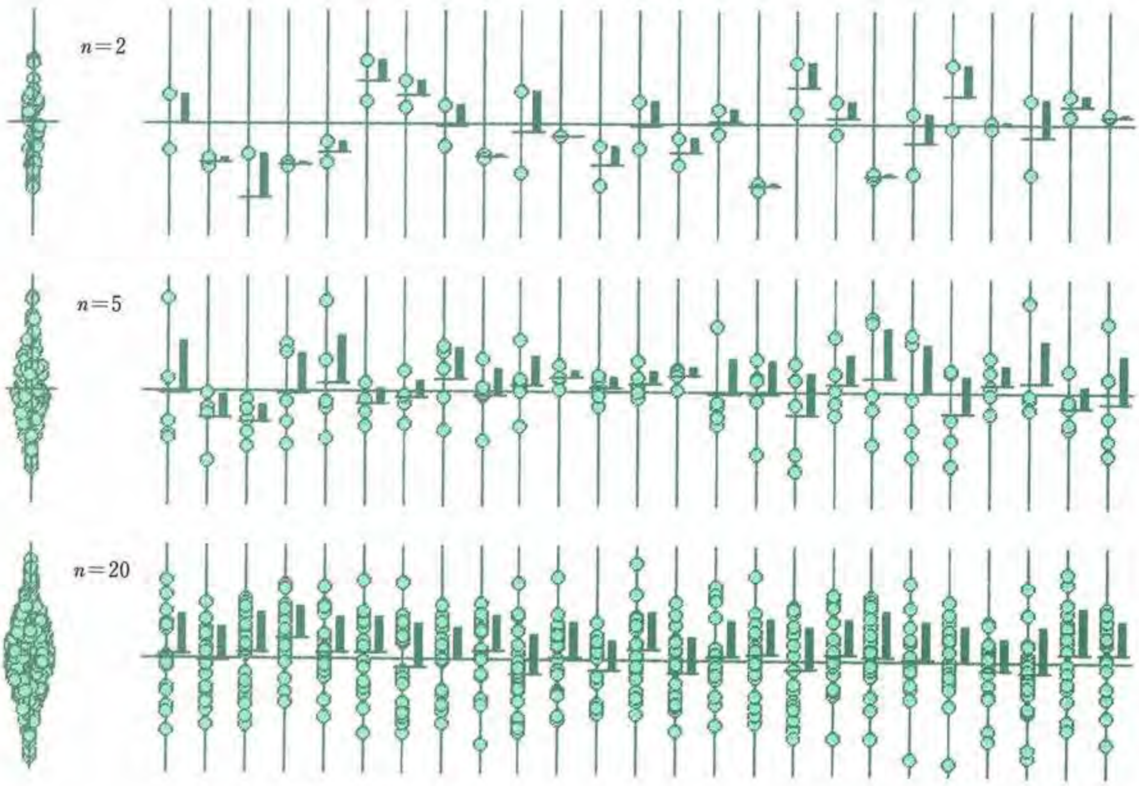
〈シミュレーションによる説明〉

正規乱数を，2，5，20個ずつ発生させ，それぞれの平均値を横線で，標準偏差（s_n または s）を黒の棒で示す．

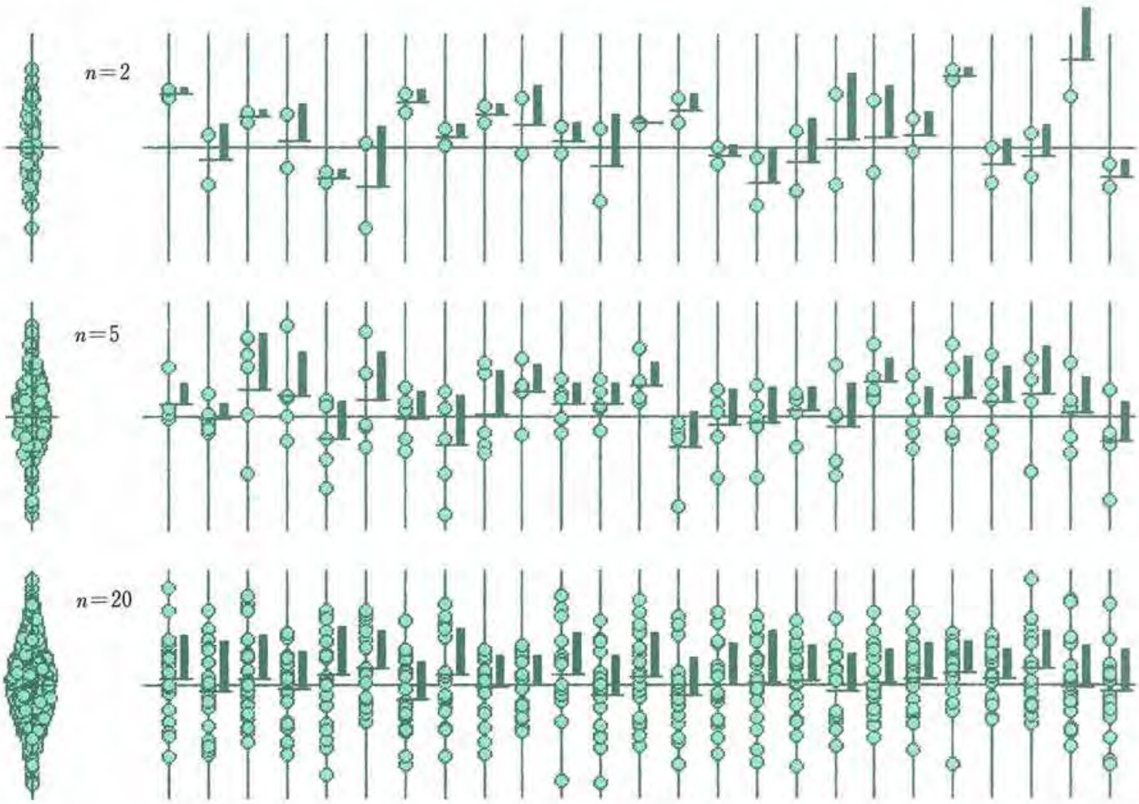
図 a の標準偏差は偏差平方和 $\Sigma(x_i-\bar{x})^2$ を n でわった平方根（s_n），図 b は，それを $n-1$ でわった平方根（s）を示す．

図から明らかなように，s_n と計算した黒棒の平均の長さは，データ数が少ないと短くなるが，s と計算した黒棒の平均の長さは，データ数によらず一定となっている．

(a)　**n 個の標本の，平均値と標準偏差（s_n）のばらつき**



(b) n 個の標本の，平均値と標準偏差（s）のばらつき



t 分布の意味と性質

予 備 知 識

x が正規分布 $N(\mu, \sigma^2)$（図 a）に従うとき，そこから得られたデータ数 n の標本平均 $\bar{x}$ も正規分布（図 b）で，その平均値は μ，標準偏差は $\sigma/\sqrt{n}$ となる．（☞24頁）

ここで，

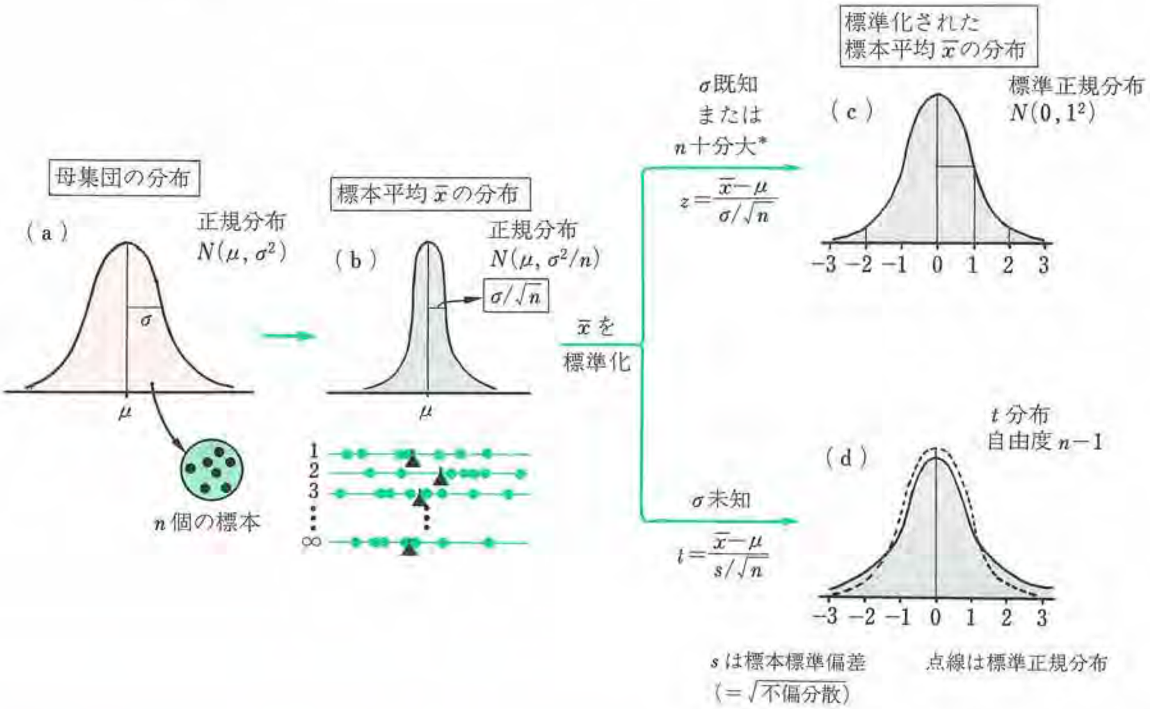
(1) $\bar{x}$ を**母標準偏差 σ を使って標準化**（σ 既知）すると，その値 z は標準正規分布に従う（図 c）

$$z=\frac{\bar{x}-\mu}{\dfrac{\sigma}{\sqrt{n}}}$$

これに対して，

(2) $\bar{x}$ を**標本標準偏差 s を使って標準化**（σ 未知）すると，その値 t は t 分布に従う（図 d）

$$t=\frac{\bar{x}-\mu}{\dfrac{s}{\sqrt{n}}}$$



* n 十分大とはおよそ $n>100$

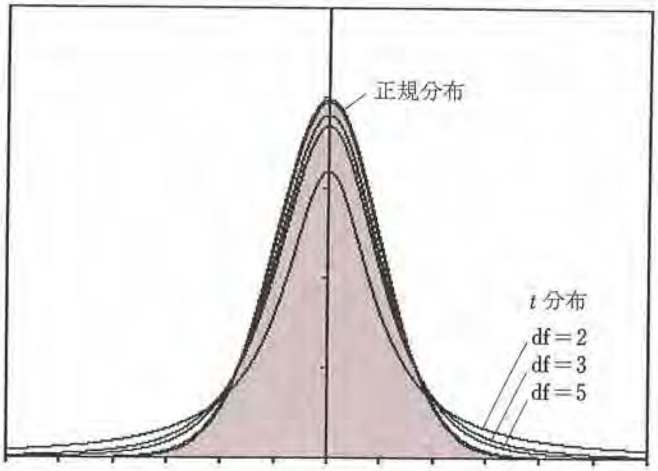
Key Point

- x が正規分布（またはデータ数大）のとき，標本平均 $\bar{x}$ を，母標準偏差 σ を使って標準化すると，標準正規分布．
- 標本平均 $\bar{x}$ を，標本標準偏差 s（曖昧な値）を使って標準化すると t 分布．
- データ数大で，s が正確なら t 分布は正規分布と一致する．

正規分布と t 分布の違い

z と t の違いは，標準化式分母の σ と s の違い．ところが，この違いが分布型に大きな差を生じる．

すなわち，σ は母集団によって決まる定数であるが，s は標本サイズに依存したばらつきやすい値．したがって，**データ数が少ないと，s は大きくばらつき，それに応じて t 値も変動する**（t 分布の両裾が広がる）．逆に，**データ数が増えると**，s は正確となり，母集団の σ に近似するので，**z 値と t 値は一致する**（t 分布の両裾の広がりが小さくなり，正規分布に近似）．



標本のデータ数（自由度）と t 分布の形状

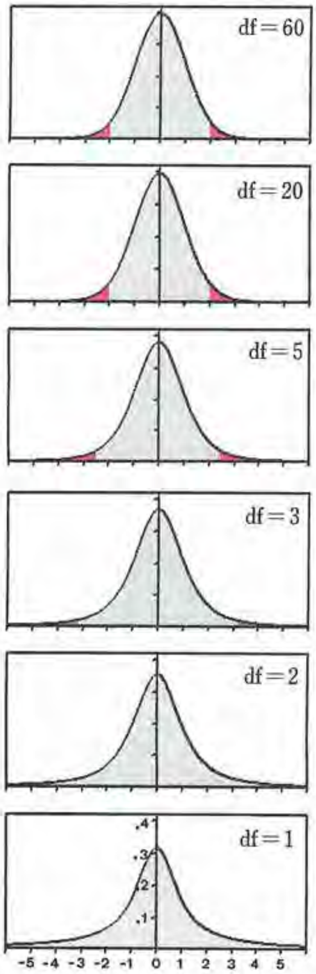
予備知識

データ数が少ないと，t 分布は両裾の広がりの大きな分布を示し，有意点 $P = 0.05$ に対する t 値も大きな値となる（たとえば，自由度 2 のとき $t = 4.30$）．

しかし，データ数が増えると，分布の裾広がりが減り，$P = 0.05$ 値は，

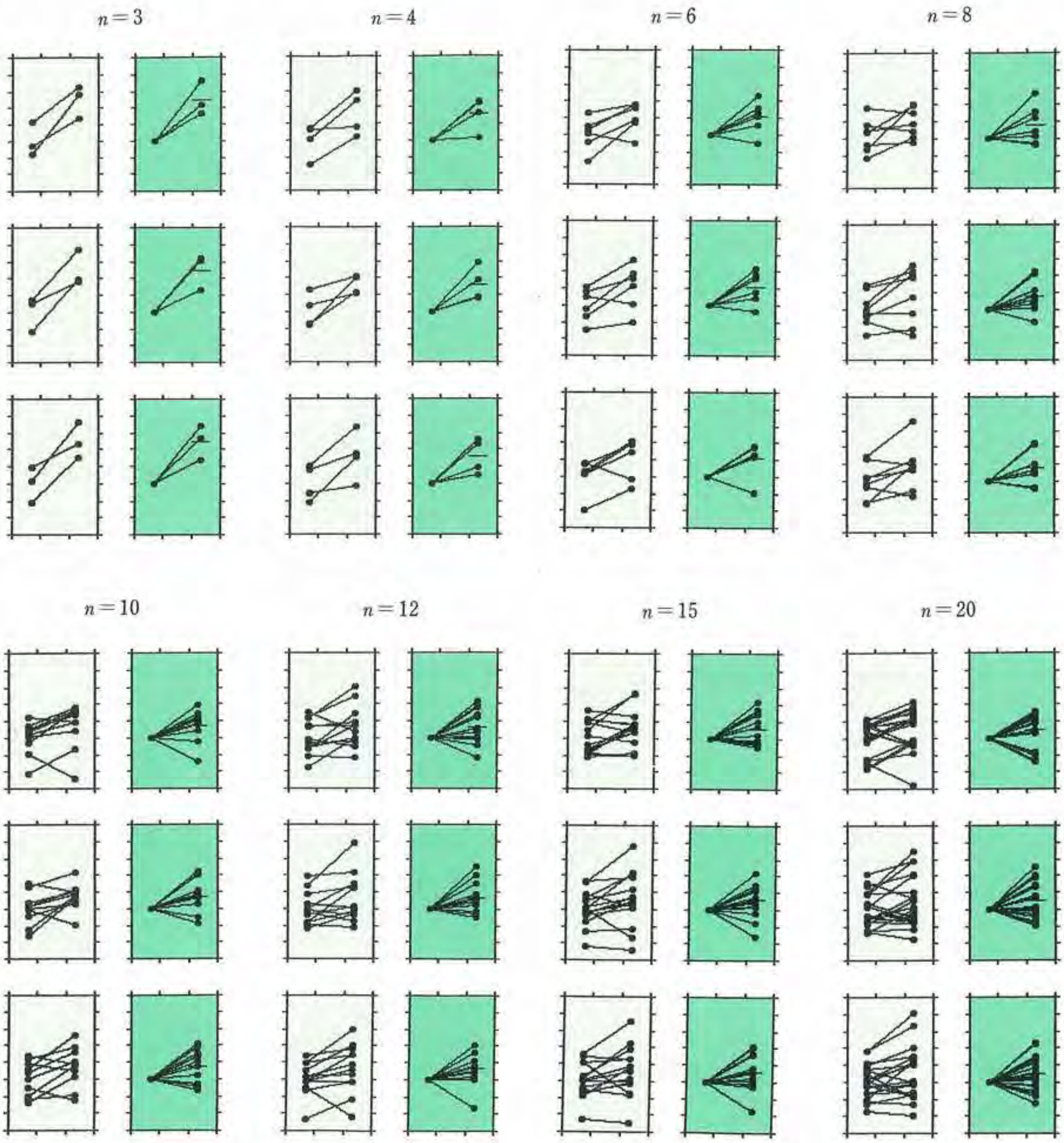
自由度 3 で $t = 3.18$，
5 で $t = 2.57$，
10で $t = 2.23$，
60で $t = 2.00$

と順に小さくなり，最終的に $t = 1.96$ に収束する．



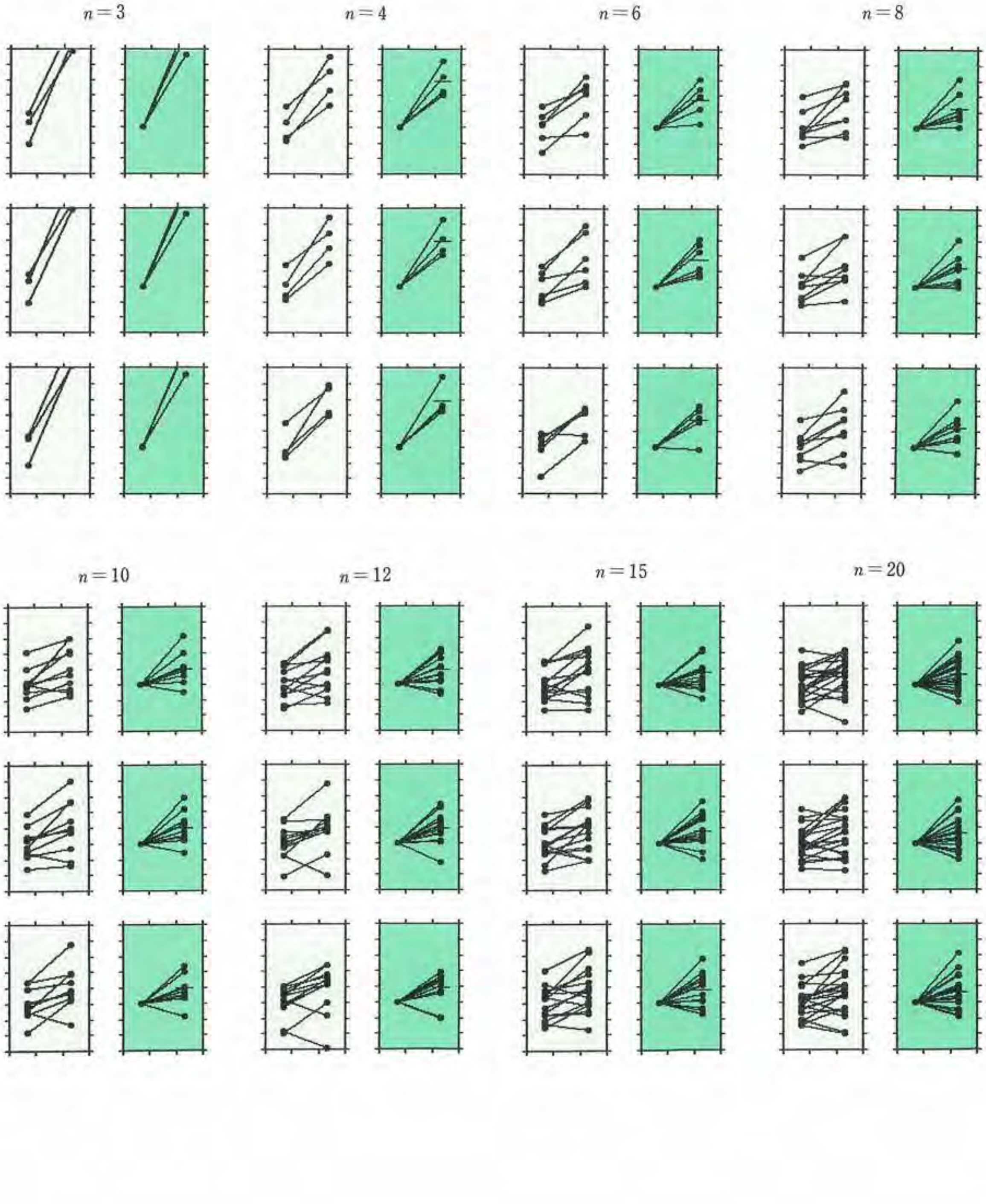
イメージレッスン —— 有意差のめやす

一標本 t 検定　$P = 0.05$



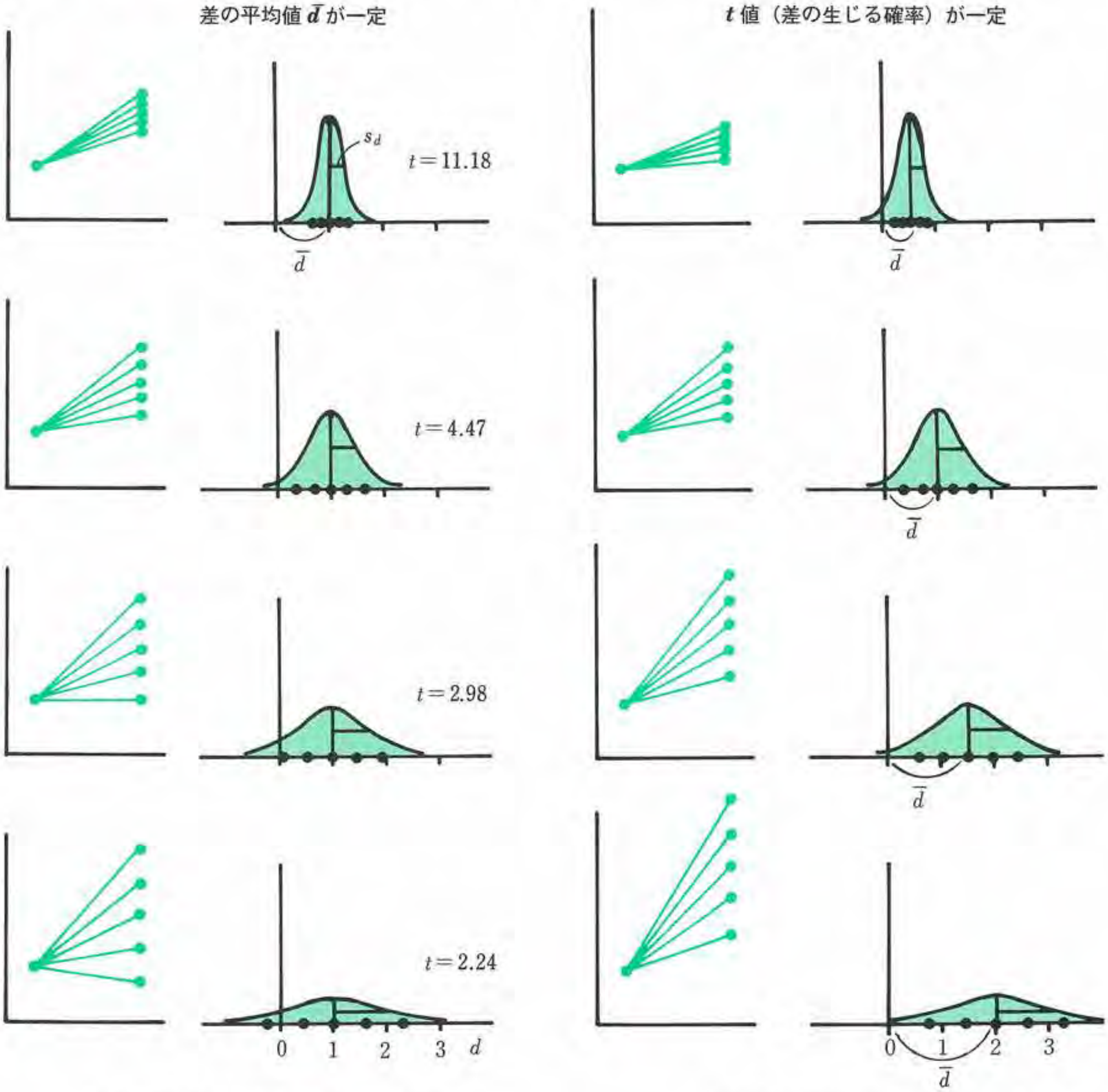
▼一標本 t 検定で $P = 0.05$ および $P = 0.01$ となるようなデータを n 組，正規乱数を使って発生させた．それぞれの n に対して3回発生し，左側が生データ，右側が差の分布を示す．

一標本 t 検定　$P = 0.01$



イメージレッスン

差の分布の広がりと一標本 t 検定における有意差



差の分布の広がりが小さいほど，t 値は大きくなる．

$$t=\frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

$\bar{d}$ 一定

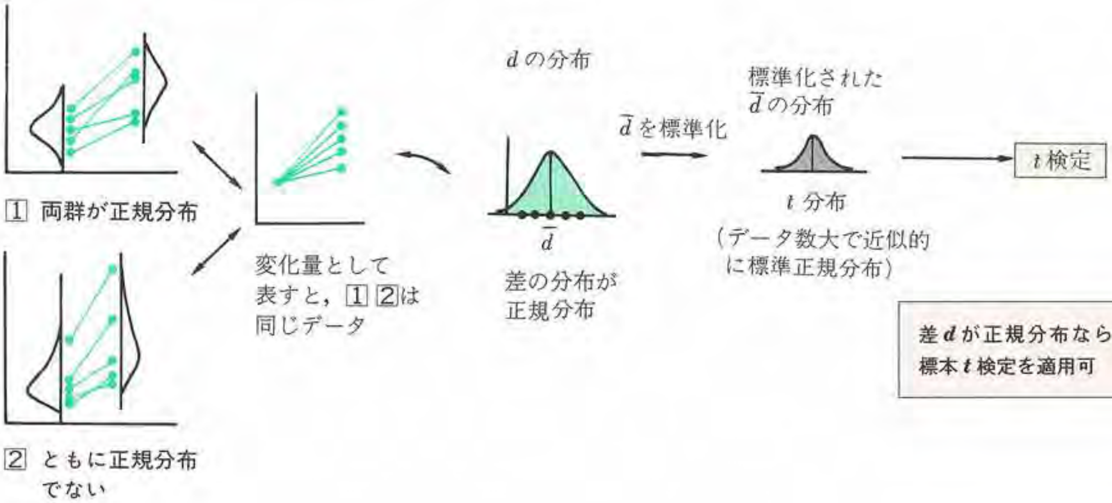
差の分布の広がりが小さいほど，より小さな平均値の差を検出できる．

$$t=\frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

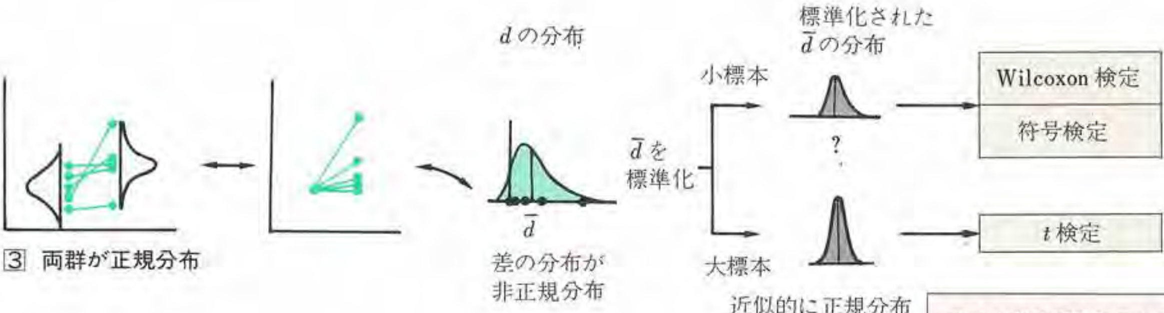
t 一定

イメージレッスン ── 使い分けのパターン認識

一標本 t 検定の成立条件からみた使い分け



差 d が正規分布なら，一標本 t 検定を適用可



差 d が非正規分布のとき，データ数小なら一標本 t 検定は適用困難．データ数大なら $\bar{d}$ の分布は，正規近似でき，一標本 t 検定を適用可．ただし，検定効率はよくないことが多い．

Key Point

一標本 t 検定の妥当性

(1) もとのデータの分布型でなく，差の分布型が使い分けの決め手．

(2) 小標本➡差の分布が正規分布なら，妥当性あり．

大標本➡差の分布型も問わない，しかし，裾広がりの大きな分布では検出力が低下する．

Wilcoxon 検定（ノンパラメトリック法）

Wilcoxon (matched–pairs) signed-ranks test; one sample Wilcoxon test; Wilcoxon test

Wilcoxon：ウイルコクソン

検定の手順

“２群間に差がある”といってよいか？

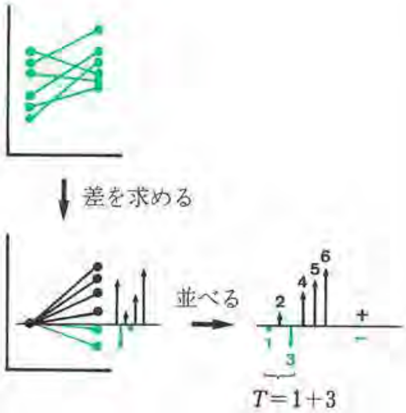
(1)　**仮説の設定**：いったん，“差がない”と仮定する（帰無仮説：H_0）．
“差がある”とする逆の仮定（対立仮説：H_1）は，一時保留．

(2)　**統計量を求める**：

①　n 組のペアにつき，その差 d を求める．

②　符号を無視して，d を**小さい方から順に並べる**．

③　d の符号により，その順位を＋と－に分け，少ない方の符号に属する順位をたし合わせ T とする．
この**符号別順位和** T は，２群の差を表す Wilcoxon 統計量で，差が大きいほど小さくなる．

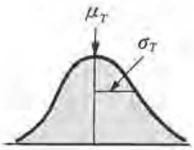


(3)　**確率を求める**：

$n \leqq 25$ のとき[*1] Wilcoxon 検定表（付表６）から，T の有意性を示す確率を求める．

$n > 25$ のとき T は近似的に次の正規分布に従う．

平均値	$\mu_T = \dfrac{n(n+1)}{4}$
標準偏差	$\sigma_T = \sqrt{\dfrac{n(n+1)(2n+1)}{24}}$



したがって，T を標準化し，

$$z = \frac{T-\mu_T}{\sigma_T}$$

から z 値を標準正規分布表（付表１）で調べれば，その確率 P が求まる[*2]．

(4)　**判　定**：有意水準を α とすると，

$P \geqq \alpha$ のとき，H_0 を棄却できず判定保留．

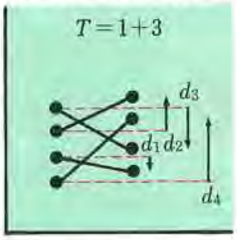
$P < \alpha$ のとき，H_0 を棄却し，対立仮説 H_1 を採用（差があると判定）．

[*1] 付表６では，$n=50$ までの T 値の確率（有意点）を示す．

[*2] 厳密には連続補正のため，
$T < \mu_T$ のとき，
T に0.5を加算．
$T > \mu_T$ のとき，
T から0.5を減算．
しかし，T 値は広い範囲に分布するので補正効果は少ない（☞49頁）

(検定の概念) Wilcoxon 検定と統計量 T（各ペアの差の符号別順位和）

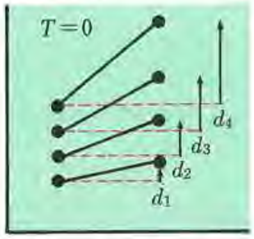
差があるとはいえない



変化の向きと大きさ（差の順位）に偏りなし

少ない方の符号の順位和 T が，2群の差を表す

差がある



変化の向きに偏りあり
逆向きのものがあっても，その大きさはわずか

▶差の検出力

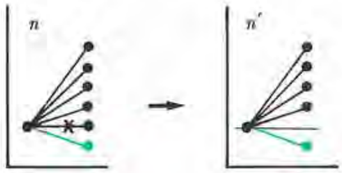
符号検定では差の向きだけを考慮したが，本検定では差の大きさも順位の形で考慮するので，検出力が大きい．

差が正規分布の場合，検出力は一標本 t 検定よりわずかに劣るが，非正規分布の場合，逆に検出力がよくなることが多い（☞61〜65頁）．

統計量 T の求め方：チェックポイント

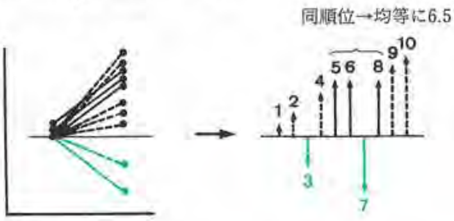
① **差が0となるペアは除外する．** これに応じて，データ数 n を減じる．

n' ペアにつき検定
ただし，0が多すぎると要補正（☞59頁上）



② **同順位があるとき，平均の順位を均等に割り当てる．**
u 番〜v 番まで同順位のとき，平均の順位 $\frac{u+v}{2}$ を割り振る．

たとえば，差 $|d|$ が5〜8番目まで同じ大きさのとき，$\frac{5+8}{2}=6.5$ を4つのペアに均等に割り振る．



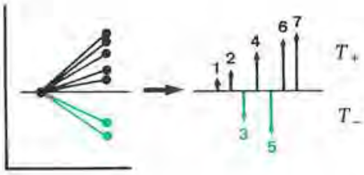
③ **少ない方の符号について，順位をたし合わせた方が楽．**
しかし，プラス符号の順位和を T_+，マイナス符号の順位和を T_- とすると，

$$T_+ + T_- = \frac{n(n+1)}{2}$$

の関係が成立し，逆に求めても相互変換可．

（例）$T_+ = 1+2+4+6+7 = 20$
$T_- = 3+5 = 8$

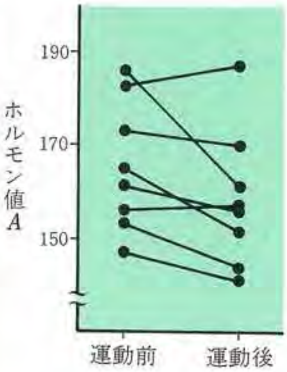
$$\boxed{20+8 = \frac{7(7+1)}{2}}$$



小標本の場合 ($n \leqq 25$)

Wilcoxon 検定 ($n \leqq 25$)

例 2-4 健常者8人を30分間ジョギングさせ，その前後で血中のホルモン値 A を測定した．運動により A 値は変動する，といってよいか．



A(前)	A(後)	① 前後の差 d を求める	② $\|d\|$ の順位*にもとの符号をつける	③ 少ない方の符号をもった順位を抜き出す
182	163	19	7	
169	142	27	8	
173	174	−1	−1	1
143	137	6	4	
158	151	7	5	
156	143	13	6	
176	180	−4	−3	3
165	162	3	2	

* 小さい方から順位をつけた方がわかりやすい．

④ ③の合計を求め，これを前後の差を表す統計量 T とする．$T=1+3=4$

検定のフロー

設 問

運動により，ホルモン値 A が変動したといってよいか？

$H_1 : A_{前値} \neq A_{後値}$

↓

(1) 仮説の設定

いったん，値は変動しなかったと，仮定する．

帰無仮説 $H_0 : A_{前値} = A_{後値}$

(H_1 は対立仮説として保留)

↓

(2) 統計量を求める

①～④により求まる T 値を，群間の差を表す統計量とする．

↓

(3) 確率を求める

H_0 による T 値の理論分布から，この標本の T 値の確率を求める．

($n \leqq 25$ なので Wilcoxon 検定表を参照して求める)

↓

(4) 判 定

有意水準 α に対する T の有意点 T_α と比較

$T > T_\alpha$ なら H_0 を棄却できず，判定保留．

$T \leqq T_\alpha$ なら H_0 を棄却し，H_1 を採用．

解

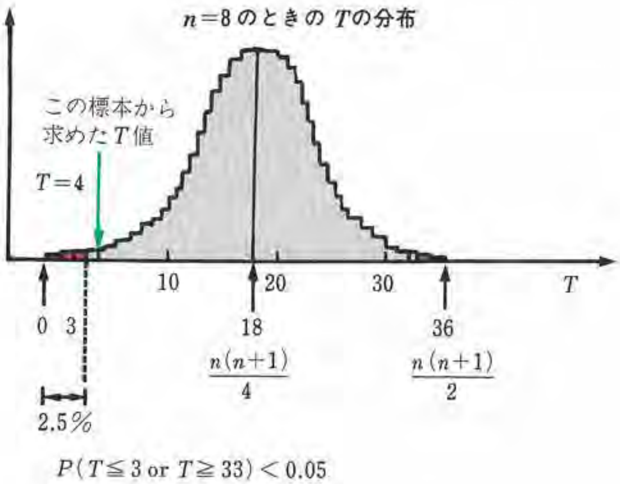
考え方：H_0 が正しいとすると，

> T のとりうる範囲は $0 \sim \dfrac{n(n+1)}{2}$
>
> T の期待値は $\dfrac{n(n+1)}{4}$

判　定：データ数が25以下なので，付表 6 から T の有意点（臨界値）を調べる．表より，両側確率が0.05以下となる T の有意点は $3(n=8)$ この標本の T 値はそれより大きいので，

$P>0.05$

∴　H_0 を棄却できない→差があるとはいえない（判定保留）．



〈Wilcoxon 検定表の読み方〉

Wilcoxon 検定表（付表 6）より，$n=8$ のとき，**両側確率 P が0.05以下となる T の下側の有意点**は 3，上側有意点は次の計算により33．表にはないがその正確な確率は $P=0.038$．

$$\frac{8(8+1)}{2}-3=33$$

◀ 公式
$T_{+}+T_{-}=\dfrac{n(n+1)}{2}$

一方，標本の T 値（下側値）は 4，上側値は

$$\frac{8(8+1)}{2}-4=32$$

したがって，求める両側確率 P は，T が 4 以下または32以上となる確率になる．これを正確に記述すると

$$P=P(T\leqq 4 \text{ or } 32\leqq T)>\underbrace{0.05>P(T\leqq 3 \text{ or } 33\leqq T)}_{\text{付表 6 に示されている } n=8,\ T=3 \text{ の意味}}\ \underbrace{[=0.038]}_{\text{正確な確率}}$$

∴　$P>0.05$

となる．しかし複雑になるし，もともと T の理論分布は左右対称なので以下 $P(T\leqq 4 \text{ or } 32\leqq T)$ を $P(T\leqq 4)_{\text{両側}}$ と略して書く．

そうすると，上の記述は，

$$P=P(T\leqq 4)_{\text{両側}}>0.05>P(T\leqq 3)_{\text{両側}}$$

となる．

大標本の場合（$n>25$）

Wilcoxon 検定（$n>25$）

例 2-5 30人の患者に，時期を変えて2種の利尿剤A，Bを投与し，その効果（尿量）を比較した．両剤の効果に差がある，と判断してよいか．

解

仮説の設定：

帰無仮説 H_0：$尿量_A=尿量_B$

対立仮説 H_1：$尿量_A \neq 尿量_B$

統計量を求める：

① 差が0のデータを除く，データ数を30−3=27とする．

② 同順位のとき平均の順位を用いる．0.1が5回あり，順位1～5の平均 (1+5)/2=3を均等に割り振る．

③ 順位6～8に0.2が3回．平均順位 (6+8)/2=7を均等に割り振る．

⋮

以下同様

⋮

④ 少ない方の符号の順位和

$T=3+3+7+11+14+14+20$
$=72$

がA・B剤の差が表す統計量になっている．

確率を求める*：

⑤ H_0に基づくT値の理論分布は，データ数が25以上なので，近似的に次の正規分布となる．

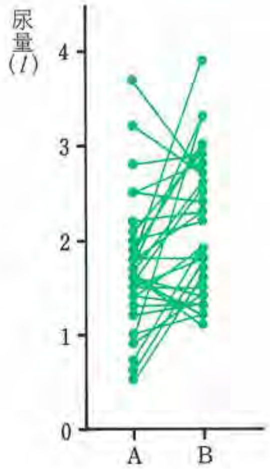
$$\mu_T=\frac{27\times28}{4}=189$$

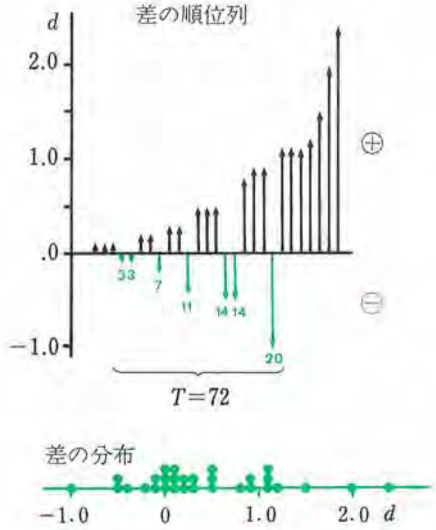
$$\sigma_T=\sqrt{\frac{27\times28\times55}{24}}=41.6$$

患者	A剤による尿量	B剤による尿量	尿量の差 d	$\lvert d\rvert$の順位とdの符号	少ない方の符号の順位
1	1.5	1.5	① 0	×	
2	1.8	1.8	0	×	
3	1.3	1.3	0	×	
4	2.8	2.9	② 0.1	3	
5	1.4	1.3	−0.1	−3	3
6	2.5	2.4	−0.1	−3	3
7	1.2	1.3	0.1	3	
8	2.2	2.3	0.1	3	
9	1.0	1.2	③ 0.2	7	
10	2.0	2.2	0.2	7	
11	1.6	1.4	−0.2	−7	7
12	2.5	2.8	0.3	9.5	
13	1.5	1.8	0.3	9.5	
14	1.6	1.2	−0.4	−11	11
15	1.8	2.3	0.5	14	
16	1.4	1.9	0.5	14	
17	3.2	2.7	−0.5	−14	14
18	1.6	1.1	−0.5	−14	14
19	0.9	1.4	0.5	14	
20	2.1	2.9	0.8	17	
21	1.6	2.5	0.9	18.5	
22	1.7	2.6	0.9	18.5	
23	3.7	2.7	−1.0	−20	20
24	0.5	1.6	1.1	22	
25	1.9	3.0	1.1	22	
26	0.6	1.7	1.1	22	
27	0.7	1.9	1.2	24	
28	1.8	3.3	1.5	25	
29	1.9	3.9	2.0	26	
30	0.9	3.3	2.4	27	

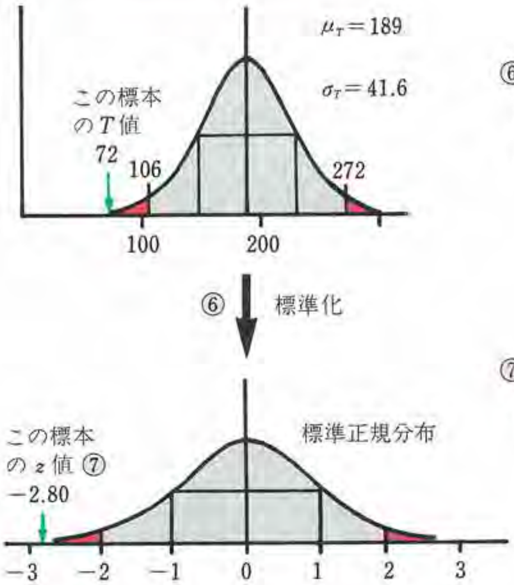
④ 計 $T=72$

* $n>25$のT値の確率はこの例のように正規近似で求まるが，付表6のWilcoxon検定表には，$n=50$までT値の有意点（下側）を示す．





⑤ $n=27$ のときの T の理論分布



⑥ この関係を利用して，この標本の T 値 72 を標準化すると（連続補正のため，T に0.5を加算），

$$z=\frac{(T+0.5)-\mu_T}{\sigma_T}$$

$$=\frac{(72+0.5)-189}{41.6}=-2.80$$

⑦ H_0 のもとで，T 値がそれ以上の極端な値をとる確率 P は，

$$P=P(T\leqq 72)_{\text{両側}}$$
$$=P(|z|\geqq 2.80)^{*1}$$
$$=0.0052<0.01^{*2}$$
$$\therefore\ P<0.01$$

*1 標準正規分布表（両側確率）より．

*2 このままでもよいが，通常はこれに近い**きりのよい値**（round number）で表す．

判　定：確率 P は極端に小さく，H_0 を棄却し，H_1 を採用．
すなわち，利尿剤A，Bの効果に有意な差があると判断する．

イメージレッスン——統計量の分布を考える

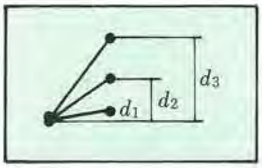
Wilcoxon 検定の統計量 T の理論分布

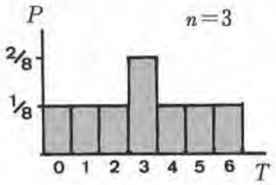
〈データが3組の場合〉

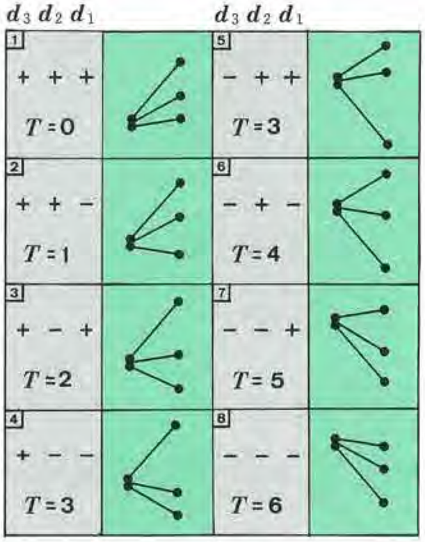
ペアが3組のとき，差の小さいものから順に d_1, d_2, d_3 とすると，各 d の符号は＋，－の2通り．したがって H_0 が正しく，群間に差がないとすると，符号の組合せは $2^3=8$ 通り．

すべての場合について T 値を求めると，$n=3$ に対する T の理論分布が得られる．

1，8番のような極端な場合が分布の両端に位置することに注意．





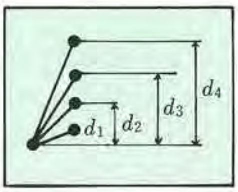


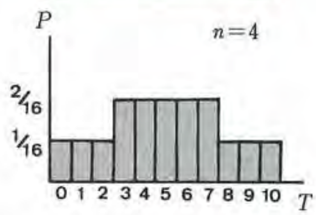
〈データが4組の場合〉

差の小さいものから，d_1, d_2, d_3, d_4 とすると，その符号のとり方は $2^4=16$ 通り．

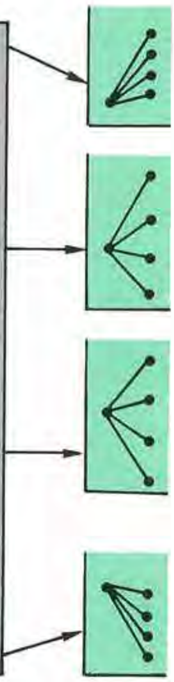
すべての場合について T 値を求めると，$n=4$ に対する，T の理論分布が得られる．

1，16番のような極端な場合が分布の両端に位置することに注意．

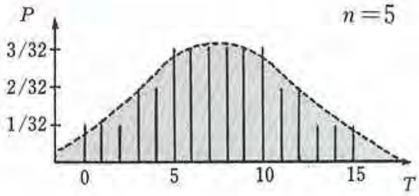




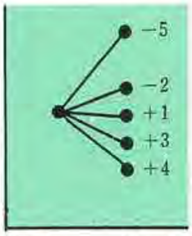
	d_4	d_3	d_2	d_1	T
1	+	+	+	+	0
2	+	+	+	−	1
3	+	+	−	+	2
4	+	+	−	−	3
5	+	−	+	+	3
6	+	−	+	−	4
7	+	−	−	+	5
8	+	−	−	−	6
9	−	+	+	+	4
10	−	+	+	−	5
11	−	+	−	+	6
12	−	+	−	−	7
13	−	−	+	+	7
14	−	−	+	−	8
15	−	−	−	+	9
16	−	−	−	−	10



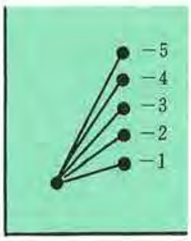
〈データが 5 組の場合〉



H_0 が正しいとすると，5 組の差（d_1～d_5）の符号のとりかたは，全部で $2^5=32$ 通り．



$T_+=8$　　$T_-=7$



$T_+=0$　　$T_-=15$

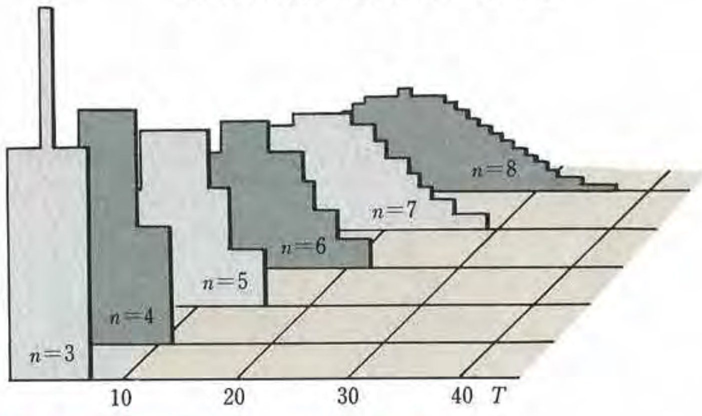
T_+	d_1	d_2	d_3	d_4	d_5	T_-
15	+	+	+	+	+	0
10	+	+	+	+	−	5
11	+	+	+	−	+	4
6	+	+	+	−	−	9
12	+	+	−	+	+	3
7	+	+	−	+	−	8
8	+	+	−	−	+	7
3	+	+	−	−	−	12
13	+	−	+	+	+	2
8	+	−	+	+	−	7
9	+	−	+	−	+	6
4	+	−	+	−	−	11
10	+	−	−	+	+	5
5	+	−	−	+	−	10
6	+	−	−	−	+	9
1	+	−	−	−	−	14
14	−	+	+	+	+	1
9	−	+	+	+	−	6
10	−	+	+	−	+	5
5	−	+	+	−	−	10
11	−	+	−	+	+	4
6	−	+	−	+	−	9
7	−	+	−	−	+	8
2	−	+	−	−	−	13
12	−	−	+	+	+	3
7	−	−	+	+	−	8
8	−	−	+	−	+	7
3	−	−	+	−	−	12
9	−	−	−	+	+	6
4	−	−	−	+	−	11
5	−	−	−	−	+	10
0	−	−	−	−	−	15

計 32 通り

〈一般に，データが n 組の場合〉

差の符号のとりかたは，全部で 2^n 通り，
T のとりうる値の範囲は，$0\sim\dfrac{n(n+1)}{2}$

Wilcoxon 検定の統計量 T の理論分布



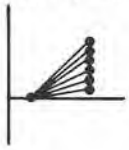
Key Point

データ数が増えると，T 値の理論分布は，正規分布に近似する．

イメージレッスン —— 有意差のめやす

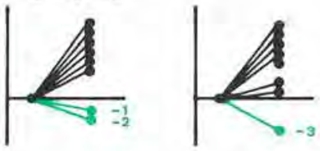
Wilcoxon 検定　$P<0.05$

$n=6$: $T=0$



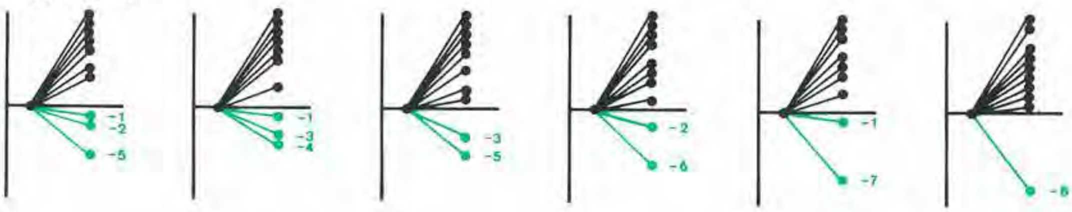
$n=7$: $T=2$

$n=8$: $T=3$



$n=9$: $T=5$

$n=10$: $T=8$



Wilcoxon 検定で差が 0 のデータが多い場合

ちょっと掘り下げて

Wilcoxon の原法では，差が 0 のデータは除外することになっている．しかし右下の図のように，0 が極端に多いとき，その妥当性に疑問が残る．実際に，差が 0 のものも含めて順位をつけるべきとする考え方もある*1.

たとえば，54頁の例 2-5 で差が 0 のものも含めて順位をつけるとすると，差が 0 のペアが 3 つあり，その平均順位 $(1+2+3)/3=2$を各 0 に均等に割り振る．また，＋，－の符号も各 0 に均等に割り振るので，マイナス分として 2×3 の半分の 3 をもとの T 値（72）にたし合わせる．

これにともない，マイナス符号を持つ他の 7 つの順位も 3 ずつ繰り上がる*2ので，0 を含めた T 値（T'）は，連続補正も含めて，

$$T' = 72+0.5+3+3\times7 = 96.5$$

となる．T' を標準化すると，

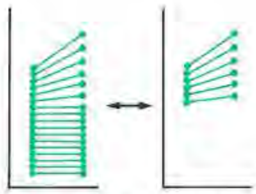
$$\mu_T = \frac{30\times31}{4} = 232.5 \qquad \sigma_T = \sqrt{\frac{30\times31\times61}{24}} = 48.6$$

となるので $z=-2.798$となり，除外しなかったときの $z=-2.800$ とほとんど差がない．

このように，差が 0 のデータ数がこの程度ならはじめから除外しても問題ない．

*1 参考文献 B-3，p 288

*2 もう一度最初から順位を付け直した方が，わかりやすい．



差が 0 のデータを除外すると，同じデータ

同順位（連）が多いときの T 値の補正

ちょっと掘り下げて

同順位（連：tie）が多いとき，正規近似公式で T 値の確率を求めると実際より大きめに計算される．それは σ_T の公式は，同順位がないことを前提として作られているためで，その補正は次のように行う．

例 2-5 について考えると，連の数 t は，

$$\text{差が}\ \left.\begin{array}{l} 0.1\cdots\cdots\cdots t=5 \\ 0.2\cdots\cdots\cdots t=3 \\ 0.3\cdots\cdots\cdots t=2 \\ 0.5\cdots\cdots\cdots t=5 \\ 0.9\cdots\cdots\cdots t=2 \\ 1.1\cdots\cdots\cdots t=3 \end{array}\right\} \quad \begin{array}{l} t=2\ \text{が 2 組} \\ t=3\ \text{が 2 組} \\ t=5\ \text{が 2 組} \end{array}$$

各連に対する補正項は $T_t = t^3-t$

その総和 ΣT_t を求めると，

$$\begin{aligned}\Sigma T_t &= 2\times(2^3-2)+2\times(3^3-3)+2\times(5^3-5) \\ &= 12+48+240 = 300\end{aligned}$$

となる．これを用いて，理論分布の標準偏差 σ_T を次式により補正すると，

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\Sigma T_t}{48}} = \sqrt{\frac{27\times28\times55}{24} - \frac{300}{48}} = 41.55$$

となり，未補正のときの $\sigma_T = 41.62$ とほとんど差がない．補正後の標準化された T 値（z'）は

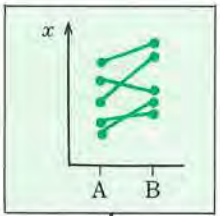
$$z' = \frac{T+0.5-\mu_T}{\sigma_T} = \frac{72+0.5-189}{41.55} = 2.804$$

となり，もとの $z = 2.800$ と大差がない．

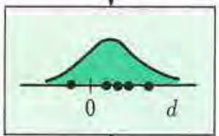
このように，連の数が 5 以下のものがかなりあっても補正の必要はないと考えてよい．

まとめ

関連2群の差の検定の使い分け



差の分布になおす



dは正規分布 → YES → 一標本 t 検定

dは正規分布 → NO → dを変数変換で正規化 → 一標本 t 検定

dは正規分布 → NO → 検定可能なデータ数

検定可能なデータ数 → NO → 不可

検定可能なデータ数 → YES → 両側検定 $n \geqq 6$ → Wilcoxon 検定

検定可能なデータ数 → YES → 両側検定 $n \geqq 6$ → 符号検定

一標本 t 検定

$$t=\frac{\bar{d}}{s_d/\sqrt{n}}$$

t 分布表　自由度 $n-1$

$$s_d=\sqrt{\frac{\Sigma\ (d_i-\bar{d})^2}{n-1}}$$

Wilcoxon 検定

$n \leqq 25$　Wilcoxon 検定表

$n > 25$　正規分布表

$$z=\frac{T-\mu_T}{\sigma_T}$$

$$\mu_T=\frac{n\,(n+1)}{4}$$

$$\sigma_T=\sqrt{\frac{n\,(n+1)\,(2n+1)}{24}}$$

連続補正

$T<\mu_T$　$T \leftarrow T+0.5$

$T>\mu_T$　$T \leftarrow T-0.5$

符号検定

$n \leqq 25$

$$P=\frac{\sum_{k=0}^{r} {}_nC_k}{2^n}$$

$n > 25$

$$z=\frac{r-n/2}{\sqrt{n}/2}$$

連続補正

$r<\frac{n}{2}$　$r \leftarrow r+0.5$

$r>\frac{n}{2}$　$r \leftarrow r-0.5$

Key Point

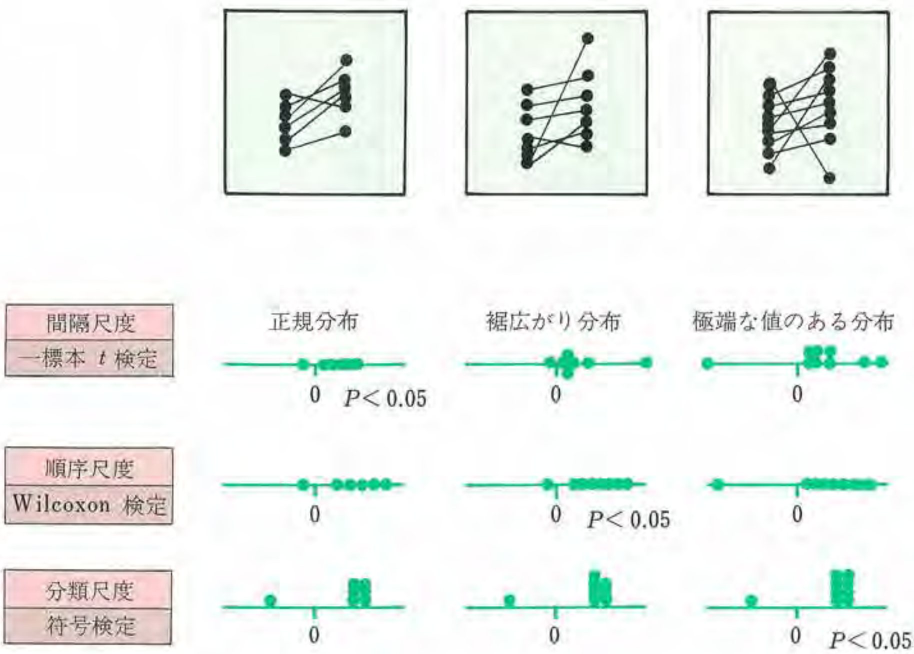
(1) 関連ある2群のデータ，各ペアの差をとれば1群の問題に帰着．
検定法の使い分けは，差の分布型から判断．

(2) データ数が6未満*のとき，ノンパラメトリック法は使えない．

* 片側検定のとき，5未満．

イメージレッスン ──使い分けのパターン認識

データの分布型と測定尺度の変換（関連 2 群の差の検定の場合）



◀上の枠内に示す関連 2 群の関係を，“2 群の差”の点列に直して，下 3 段に示す．

上段の点列は，差をそのまま間隔尺度として一標本 t 検定を，中段の点列は，順序尺度に直して Wilcoxon 検定を，下段の点列は，プラスかマイナス（分類尺度）に直して符号検定を，適用した場合を示す．

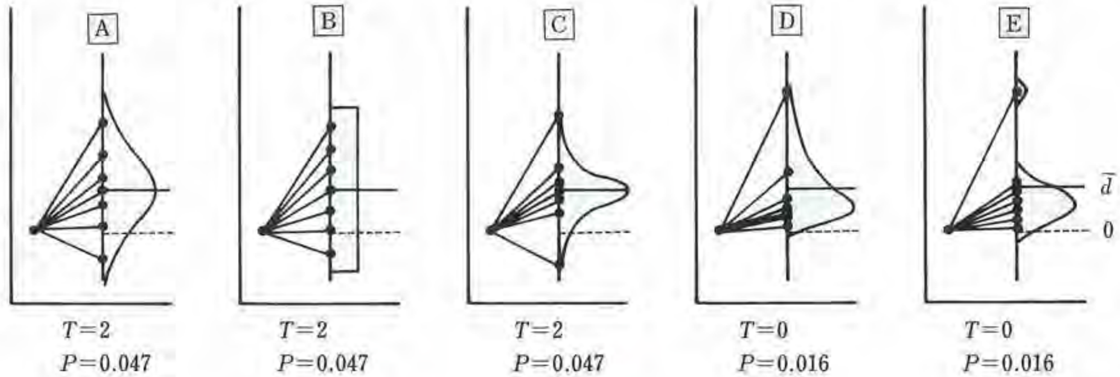
Key Point

検定法によって判定が相違するのは，データを異なる尺度*で利用するため．

* ここでは，それぞれの尺度を生かした検定法で，有意差が生じるような例をあげた．しかし一般には，非正規分布の場合，データを変数変換により正規化し，高次の尺度（間隔尺度）のまま検定すると，差の検出力がもっとも高くなる．これは，低次の尺度を利用すると，確かに分布型に依存しない検定が可能だが，一方で情報のロスが生じるため．

イメージレッスン——使い分けのパターン認識

差の分布型からみた，一標本 t 検定と Wilcoxon 検定の使い分け（少数データの場合）



$t=2.45 \quad P=0.05 \qquad t=\dfrac{\bar{d}}{s_d/\sqrt{n}}$

Ａ～Ｅのいずれも，一標本 t 検定で確率がちょうど0.05となる場合．データ数 n はすべて7で，差の平均値 $\bar{d}$ を0.92，d の標準偏差 s_d を1.0として点の配置を変化させたので，t 値は2.45で一定．

しかし，それぞれの場合に Wilcoxon 検定を適用すると，統計量 T 値は分布型により異なる．

Ａは**正規分布**を想定した場合で，T 値に対する正確な確率は $P=0.047$．この場合は，たまたま Wilcoxon 検定の方が確率が小さくなっているが，理論的には，一標本 t 検定の方がわずかに差の検出力がよい．

Ｂは**一様分布**の場合で，データ数が少ないと，t 検定の方が若干差の検出力がよくなる．しかし理論的には，データ数を限りなく増やしてゆくと，一標本 t 検定と Wilcoxon 検定の検出力は同等になる．

Ｃは左右対称の分布であるが，**正規分布より両裾の広がりが大きい**場合で，理論的には t 検定の方が差の検出力が悪い．これは，点が $\bar{d}$ 付近に集中している割には両裾の点によって s_d（t 値の分母）が大きくなるためで，逆に Wilcoxon 検定では，両裾の点も，順序として扱うのでその影響を受けにくい．

Ｄは**対数正規型の分布**をする場合，Ｅは**分布の片側に大きな飛び離れ点**がある場合で，いずれも Wilcoxon 検定の方が差の検出力が高い．実際にこの例では，T 値は0で，その理論確率は0.016と t 値の確率0.05より小さい値となっている．

これは，分布の片裾が長く尾をひくため，t 検定ではＣの場合同様大多数の点が（$\bar{d}$ 付近）に集中している割には，極端な値によって s_d が相対的に大きくなってしまうため．ただし，64頁の図に示すように，標本数が大きくなると，Ｄのような片側に裾の長い分布では，逆に t 検定の方が検出力がよくなる．

CHAPTER

5

独立多群の差の検定

変動（ばらつき）の表し方

基礎知識

① **偏差 d**：difference

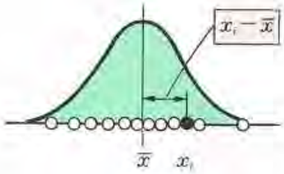
$$d = x - \bar{x}$$

個々のデータの平均値からの偏り
d の合計（Σd_i）は 0 になる．

② **平均偏差 MAD**：mean absolute deviation

$$\mathrm{MAD} = \frac{\sum |x_i - \bar{x}|}{n}$$

偏差の絶対値の平均



③ **偏差平方和 S**：mean square

$$S = \sum (x_i - \bar{x})^2$$

偏差を 2 乗してたし合わせた値．
分散分析の計算の基本統計量になる．

④ **分散 s^2 (σ^2)**：variance

母平均 μ 未知：標本分散*1

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

偏差平方和をデータの自由度でわった値．

*1 正確には不偏分散と呼ぶ．

母平均 μ 既知：母分散

$$\left[\sigma^2 = \frac{\sum (x_i - \mu)^2}{n} \right]$$

⑤ **標準偏差 s (σ)**：standard deviation

母平均 μ 未知：標本標準偏差*2

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

分散の平方根で，一番よく使われるデータの平均偏差指標．

*2 正確には不偏分散の平方根と呼ぶ．

母平均 μ 既知：母標準偏差

$$\left[\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}} \right]$$

⑥ **変動係数 CV**：coefficient of variation

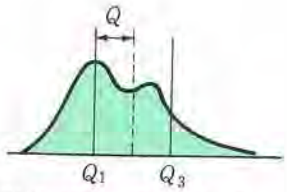
$$\mathrm{CV} = \frac{s}{\bar{x}} \times 100$$

標準偏差を標準化するため，それを平均値でわって百分率で表した値．
適用できるのは，値が常に正の場合に限られる．

⑦ **四分位偏差 Q**：quartile deviation

$$Q = \frac{Q_3 - Q_1}{2}$$

25パーセンタイル値(Q_1)と75パーセンタイル値(Q_3)の差を 2 でわった値*3．標準偏差と違って，分布型の影響を受けない信頼区間が求まる．

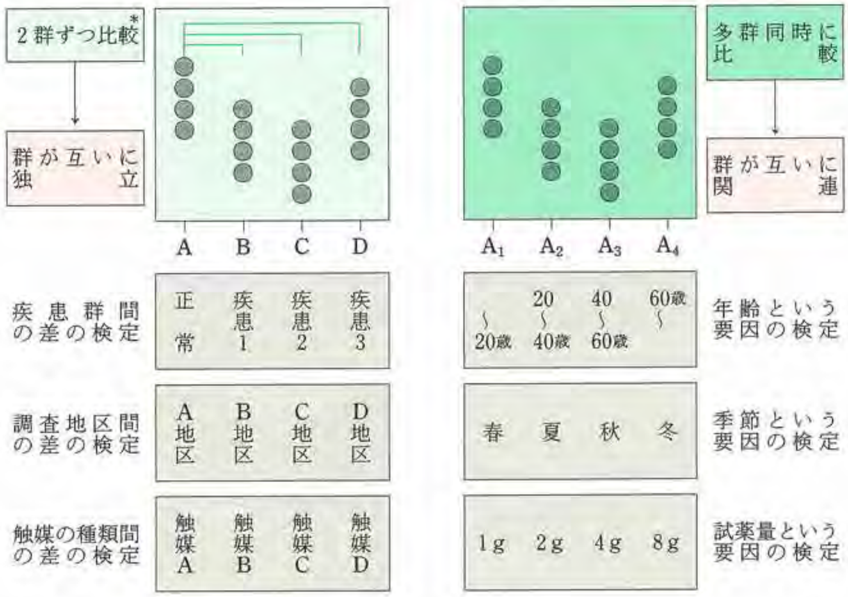


*3 Q_1 は第一·四分位数，Q_3 は第三·四分位数とも呼ばれる．

独立多群を同時に比較する必要がある場合とは

—→相互関連をもった群分類

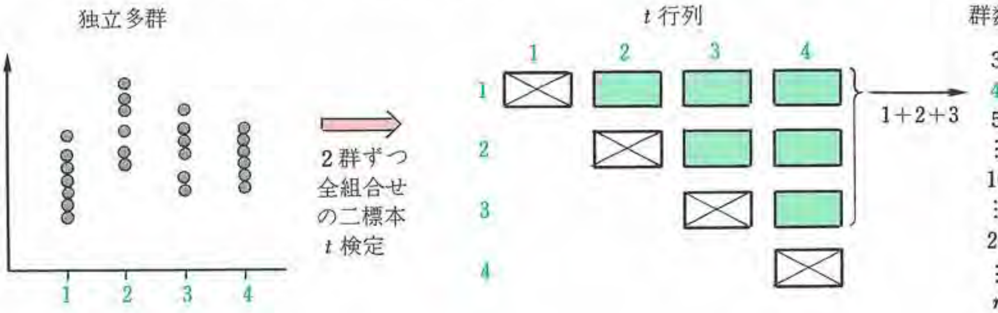
基礎知識



* 群分類が系統的で偏りがなければ，多群同時に比較しうる．たとえば，種差，地域差の有無など．

休憩室

2群の差の検定，すべての組合せは何通り？



$$n \cdots {}_nC_2=\frac{n(n-1)^*}{2}$$

* n 個から2個取り出す組合せ（${}_nC_2$）は1から $n-1$ までの累和と等しい．

$1+2+\cdots+n-1$
$=n(n-1)/2$

(cf. $1+2+\cdots+n$
$=n(n+1)/2$)

一元配置分散分析法（パラメトリック法）

one-way analysis of variance; one-way ANOVA

検定の手順

要因Aにより分類（水準化）された k 群（$A_1, A_2, \cdots\cdots, A_k$）．

要因Aの水準間に差があるといえるか？

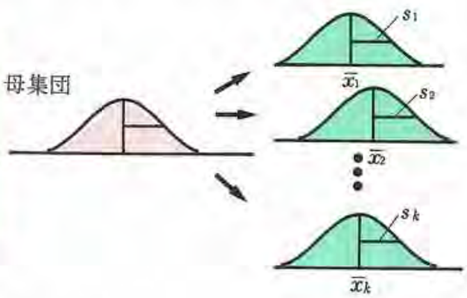
(1) 仮説の設定：

帰無仮説 H_0：各群は同一の母集団からの標本（要因 A の水準間に差はない）*1．

対立仮説 H_1：各群は異なる母集団からの標本（要因 A の水準間に差がある）．

*1 実際には，群間変動（要因 A の水準間変動）は，もともとの測定値の変動（純誤差変動）と同程度と仮定して検定する．

要因A	測定値	データ数	平均値	分散
A_1	○○○	n_1	$\bar{x}_1$	s_1^2
A_2	○○○○	n_2	$\bar{x}_2$	s_2^2
⋮				
A_k	○○○○	n_k	$\bar{x}_k$	s_k^2



(2) 統計量を求める：分散分析表の作成

① 群間変動*2（要因 A による変動）を求める

*2 ここでは要因 A の各水準を一つの群とみなして，水準間変動のかわりに，群間変動ということばを使った．級間変動とも呼ばれる．

- 偏差平方和 S_A を求める．

$$\boxed{S_A=\sum_{i=1}^{k} n_i(\bar{x}_i-\bar{\bar{x}})^2}$$

←群別平均（$\bar{x}_i$）の総平均（$\bar{\bar{x}}$）からの偏差平方和

←平均値の分散は，データ数の分だけ小さくなっているので n_i 倍しておくことに注意！

- 自由度 df_A を求める．

$$df_A=k-1$$

←群数−1

② 群内変動（誤差変動）を求める

- 偏差平方和 S_E を求める．

$$\boxed{S_E=\sum_{i=1}^{k}\underbrace{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}=\sum_{i=1}^{k}\{\underbrace{(n_i-1)\cdot s_i^2}\}}$$

←（データ数−1)·(分散）としてもよい．

↑群ごとに各点(x_{ij})から中心($\bar{x}_i$)までの偏差平方和を求める．

- 自由度 df_E を求める．

$$df_E=\sum_{i=1}^{k}(n_i-1)$$

←各群の（データ数−1）をたし合わせる．

$$=N-k$$

←データ総数（N）から群数（k）をひいても同じ．

③ 分散分析表にまとめて，分散，分散比を求める．

変動要因	偏差平方和	自由度	分散(平均平方)	分散比
群間変動	S_A	$df_A=k-1$	$s_A{}^2=S_A/df_A$	$\boxed{F=\frac{s_A{}^2}{s_E{}^2}}$ *2
群内変動	S_E	$df_E=N-k$	$s_E{}^2=S_E/df_E$	
総変動	S_T*1	$N-1$		

*1 総変動の平方和

$$S_T=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2$$
$$=S_A+S_E$$

*2 分散比 F が群間の差を表す統計量になっている．
帰無仮説が正しければ F の期待値は 1.

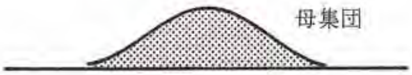
(3) 確率と判定：F 分布表（付表 4）で自由度 df_A, df_E，有意水準 α の F 値(F_α)を求める．

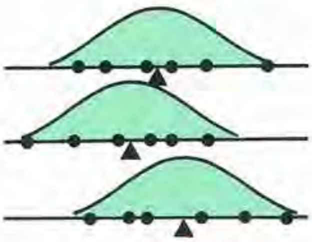
$F \leqq F_\alpha$ ：要因Aの水準間に差があるとはいえない（判定保留）．

$F > F_\alpha$ ：要因Aの水準間に差がある（H_0 を棄却し，H_1 を採用）．

検定の概念

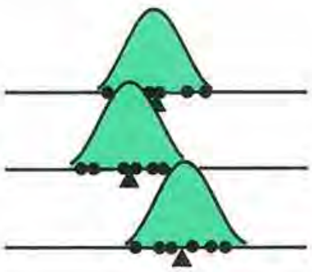
一元配置分散分析と統計量 F（分散比）





差があるとはいえない ⟶ 同一の母集団から得られた標本群と考えて矛盾しない

$$\boxed{F=\frac{\text{群間分散}}{\text{群内分散}}\fallingdotseq 1}$$



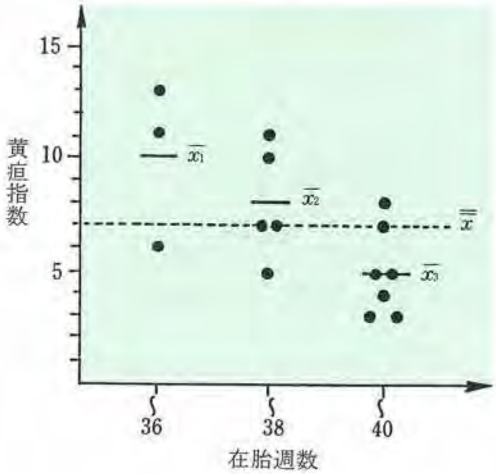
群間に差がある ⟶ 同一の母集団から得られた標本群とは考えにくい

$$\boxed{F=\frac{\text{群間分散}}{\text{群内分散}}\gg 1}$$

一元配置分散分析法

例 5-1 出産までの週数（在胎週数）によって新生児を 3 群に分け，新生児期黄疸の強さを調べたところ次のようなデータを得た．在胎週数によって黄疸の強さに差があると考えてよいか？

	~36	~38	~40	
	13	11	8	群数 $k=3$
	11	10	7	データ総数 $N=15$
	6	7	5	
		7	5	
		5	4	
			3	
			3	
データ数 n_i	3	5	7	総平均 $\bar{\bar{x}}$
平均値 $\bar{x}_i$	10	8	5	7
分散 s_i^2	13	6	$\frac{11}{3}$	



解

仮説の設定：在胎週数間で黄疸の強さに差はない（帰無仮説 H_0）．

⟶ 3 群は同一の母集団から得られた標本．

統計量を求める：仮説が正しければ，群間のばらつきと群内のばらつきは同程度．

⟶群間変動と群内変動の比を求めると，その期待値は 1．この比を統計量とする．

群間変動の偏差平方和：
$$S_A = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2$$
$$= 3\times(10-7)^2 + 5\times(8-7)^2 + 7\times(5-7)^2$$
$$= 3\times 9 + 5\times 1 + 7\times 4 = 60$$

群内変動の偏差平方和：
$$S_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$
$$= \{(13-10)^2 + (11-10)^2 + (6-10)^2\}$$
$$+ \{(11-8)^2 + (10-8)^2 + (7-8)^2\times 2 + (5-8)^2\}$$
$$+ \{(8-5)^2 + (7-5)^2 + (5-5)^2\times 2 + (4-5)^2 + (3-5)^2\times 2\}$$
$$= (9+1+16) + (9+4+2+9) + (9+4+0+1+8) = 72$$

または各群の分散から求めると，

$$S_E = \sum_{i=1}^{k}(n_i - 1)\cdot s_i^2 = 2\times 13 + 4\times 6 + 6\times\frac{11}{3} = 72$$

総変動の偏差平方和：$S_T=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{\bar{x}})^2$
（検算のため）

$$
\begin{aligned}
&=\{(13-7)^2+(11-7)^2+(6-7)^2\}\\
&\quad+\{(11-7)^2+(10-7)^2+(7-7)^2\times 2+(5-7)^2\}\\
&\quad+\{(8-7)^2+(7-7)^2+(5-7)^2\times 2+(4-7)^2\\
&\quad+(3-7)^2\times 2\}\\
&=132
\end{aligned}
$$

分散分析表

変動要因	偏差平方和	自由度	分散	分散比
群間変動	S_A	$df_A=k-1$	$s_A^2=S_A/df_A$	$F=\frac{s_A^2}{s_E^2}$
群内変動	S_E	$df_E=N-k$	$s_E^2=S_E/df_E$	
総 変 動	$S_T=S_A+S_E$	$df_T=N-1$		

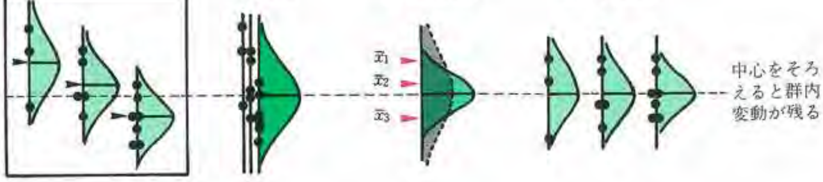
⬇ 数値に置き換える

変動要因	偏差平方和	自由度	分散	分散比
群間変動	60	2	30	$F=5.0$
群内変動	72	12	6	
総 変 動	132	14		

確率と判定：F 分布より自由度 $df_A=2$，$df_E=12$，有意水準 $\alpha=0.05$ の F 値（F_α）を調べると，$F(2,\ 12\ ;\ 0.05)=3.89$

$F>F_{0.05}$ より群間変動は群内変動より有意に大きい（在胎週数によって黄疸の強さに差がある）と判定（$P<0.05$）.

群間変動と群内変動の図解



$$F=\frac{\text{群間分散}}{\text{群内分散}}=\frac{S_A/df_A}{S_E/df_E}$$

この比 F が群間の偏りを表す.

* 平均値の分散はデータ数の分だけ小さくなるので，それを点線のように補正すると，群内変動と対等に比較できる.

▶注意

$$\text{分散}=\frac{\text{偏差平方和}}{\text{自由度}}$$

一元配置分散分析の数理

データの基本構造とその図解

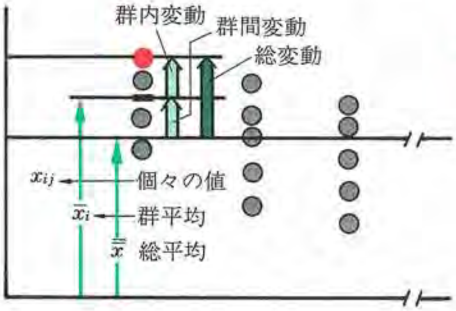
個々の点 x_{ij} を分解すると

$$\boxed{\begin{aligned} x_{ij} &= \bar{\bar{x}} + (\bar{x}_i - \bar{\bar{x}}) + (x_{ij} - \bar{x}_i) \\ &= \mu + \alpha_i + \varepsilon_{ij} \end{aligned}}$$

μ：全データの中心位置.

α_i：群として偏っている部分.

ε_{ij}：もともとのデータのばらつき（純誤差）.



ちょっと掘り下げて

変動の相互関係：数式表現*1

	総変動		群間変動		群内変動
偏差の関係式	$\boxed{x_{ij}-\bar{\bar{x}}}$	$=$	$\boxed{\bar{x}_i-\bar{\bar{x}}}$	$+$	$\boxed{x_{ij}-\bar{x}_i}$
偏差平方和の関係式	$\boxed{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{\bar{x}})^2}$	$=$	$\boxed{\sum_{i=1}^{k} n_i(\bar{x}_i-\bar{\bar{x}})^2}$	$+$	$\boxed{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}$
	S_T		S_A		S_E

*1 偏差平方和の関係式は，偏差の関係式と類似の形をとることに注目.

➡偏差はたし合わせるとどの項も 0 になる.

➡偏差平方和は 0 にならないので，相互に大きさを比較できる.

一元配置分散分析では，この偏差平方和の関係式に基づいて変動を分解し，群間変動と群内変動を比較する.

変動の相互関係：数値例*2（例 5-1 について）

総変動を求める

$$\underset{x_{ij}}{\overset{\text{実データ}}{\begin{bmatrix} 13 & 11 & 8 \\ 11 & 10 & 7 \\ 6 & 7 & 5 \\ & 7 & 5 \\ & 5 & 4 \\ & & 3 \\ & & 3 \end{bmatrix}}} - \underset{\bar{\bar{x}}}{\overset{\text{総平均}}{\begin{bmatrix} 7 & 7 & 7 \\ 7 & 7 & 7 \\ 7 & 7 & 7 \\ & 7 & 7 \\ & 7 & 7 \\ & & 7 \\ & & 7 \end{bmatrix}}} = \underset{x_{ij}-\bar{\bar{x}}}{\overset{\text{総変動}}{\begin{bmatrix} 6 & 4 & 1 \\ 4 & 3 & 0 \\ -1 & 0 & -2 \\ & 0 & -2 \\ & -2 & -3 \\ & & -4 \\ & & -4 \end{bmatrix}}}$$

群内変動を求める

$$\underset{x_{ij}-\bar{\bar{x}}}{\overset{\text{総変動}}{\begin{bmatrix} 6 & 4 & 1 \\ 4 & 3 & 0 \\ -1 & 0 & -2 \\ & 0 & -2 \\ & -2 & -3 \\ & & -4 \\ & & -4 \end{bmatrix}}} - \underset{\bar{x}_i-\bar{\bar{x}}}{\overset{\text{群間変動}}{\begin{bmatrix} 3 & 1 & -2 \\ 3 & 1 & -2 \\ 3 & 1 & -2 \\ & 1 & -2 \\ & 1 & -2 \\ & & -2 \\ & & -2 \end{bmatrix}}} = \underset{x_{ij}-\bar{x}_i}{\overset{\text{群内変動}}{\begin{bmatrix} 3 & 3 & 3 \\ 1 & 2 & 2 \\ -4 & -1 & 0 \\ & -1 & 0 \\ & -3 & -1 \\ & & -2 \\ & & -2 \end{bmatrix}}}$$

*2 [ ] に対する四則演算は対応する各要素についての演算.

□ に対する四則演算はその中の全要素を累和してからの演算であることを示す.

$$\overset{\text{総変動}}{\boxed{\begin{matrix} 6 & 4 & 1 \\ 4 & 3 & 0 \\ -1 & 0 & -2 \\ & 0 & -2 \\ & -2 & -3 \\ & & -4 \\ & & -4 \end{matrix}}} = \overset{\text{群間変動}}{\boxed{\begin{matrix} 3 & 1 & -2 \\ 3 & 1 & -2 \\ 3 & 1 & -2 \\ & 1 & -2 \\ & 1 & -2 \\ & & -2 \\ & & -2 \end{matrix}}} + \overset{\text{群内変動}}{\boxed{\begin{matrix} 3 & 3 & 3 \\ 1 & 2 & 2 \\ -4 & -1 & 0 \\ & -1 & 0 \\ & -3 & -1 \\ & & -2 \\ & & -2 \end{matrix}}}$$

偏差の和　$\sum_i\sum_j(x_{ij}-\bar{\bar{x}})=0$　　$\sum_i n_i(\bar{x}_i-\bar{\bar{x}})=0$　　$\sum_i\sum_j(x_{ij}-\bar{x}_i)=0$

➡たし合わせるといずれも 0 になっている.

$$\boxed{\begin{matrix} 36 & 16 & 1 \\ 16 & 9 & 0 \\ 1 & 0 & 4 \\ & 0 & 4 \\ & 4 & 9 \\ & & 16 \\ & & 16 \end{matrix}} = \boxed{\begin{matrix} 9 & 1 & 4 \\ 9 & 1 & 4 \\ 9 & 1 & 4 \\ & 1 & 4 \\ & 1 & 4 \\ & & 4 \\ & & 4 \end{matrix}} + \boxed{\begin{matrix} 9 & 9 & 9 \\ 1 & 4 & 4 \\ 16 & 1 & 0 \\ & 1 & 0 \\ & 9 & 1 \\ & & 4 \\ & & 4 \end{matrix}}$$

⬅対応する上段枠内の各要素（偏差）を 2 乗して求めた.

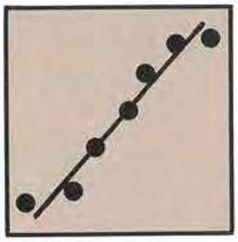
偏差平方和　$\underset{S_T}{\sum_i\sum_j(x_{ij}-\bar{\bar{x}})^2=132}$　　$\underset{S_A}{\sum_i n_i(\bar{x}_i-\bar{\bar{x}})^2=60}$　　$\underset{S_E}{\sum_i\sum_j(x_{ij}-\bar{x}_i)^2=72}$

➡偏差平方和の関係式 $S_T=S_A+S_E$ が成立している.

CHAPTER

7

回帰と相関



回帰と相関の違い

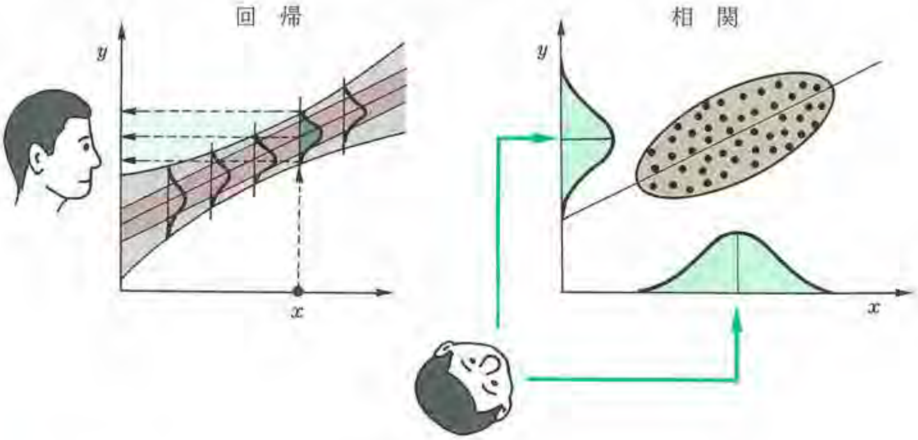
誤差の取り扱い方の違い

2変量x, yの関係を調べるのに，**回帰分析**では一方（たとえばx）を基準にして，他方(y)をそれに関係づける*1．そして両者の関係の強さは，y方向の誤差の大きさによって判断するが，x方向の誤差については考慮しない．この意味で，yの側だけをばらつきある**確率変数***2として取り扱う．

これに対して，**相関分析**では，x, yともにばらつきある確率変数とみなして，その相互関係の強さを調べる．

信頼域の違い

このような誤差の取り扱いの違いのため，回帰分析でいう信頼域は，xを基準にしたyの信頼区間をさし，全体として回帰直線（曲線）のまわりの帯状の領域となる．これに対して相関でいう信頼域は，x, yの分布の中心からの等確率距離*3をさし，平面上楕円の領域となる．



計算目的の違い

回帰と相関の違いを，直線関係についてみると，回帰直線は，xからyを**どのように直線的に関係づけられるか**を示し，相関係数は，xとyの相互関係が**どの程度直線的か**を示す．

このため回帰分析でいう“どのように”はx, yの単位に依存するが，相関分析でいう“どの程度”はx, yの単位に依存しないので異なるデータを同じ基準で比較できる．

予 備 知 識

*1 回帰の基準となる変数を**基準変数**または**独立変数**，回帰される変数を**従属変数**と呼ぶ．

*2 確率変数 random variable その観測値が偶然に左右される変数．

*3 マハラノビス距離と呼ばれる（☞234頁）．

統計計算で使う数式の簡略表現

基礎知識

- $\sum = \sum_{i=1}^{n}$　　データ数を n として $i=1 \sim n$ までの**累和**.
- $S_{xx} = \sum (x_i - \bar{x})^2 \quad = \sum x_i^2 - \dfrac{(\sum x_i)^2}{n}$

x の標本平均からの**偏差平方和**で単に**平方和**（sum of squares）とも呼ぶ．$n-1$ でわると（標本）**分散**になる．

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

x, y の標本平均からの**偏差積和**で単に**積和**（sum of products）とも呼ぶ．$n-1$ でわると（標本）**共分散**になる．

S_{xx}, S_{xy} の右端の項は，平均値を用いない表現で，$\sum x_i, \sum x_i{}^2, \sum x_i y_i$ などの基本要素の和を求めてから計算する方法．コンピュータでは一般にこの方法を用いる．

例：

x	1	2	3	4	5
y	2	2	3	4	4

① そのまま求めると，

$$S_{xx} = \sum (x_i - \bar{x})^2 = (1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2$$
$$= 2^2 + 1^2 + 0 + 1^2 + 2^2 = 10$$
$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = (1-3)(2-3) + (2-3)(2-3) + (3-3)(3-3)$$
$$+ (4-3)(4-3) + (5-3)(4-3)$$
$$= 2+1+0+1+2 = 6$$

② 各要素の和から求めると，

$$\sum x_i = 1+2+3+4+5 = 15$$
$$\sum x_i^2 = 1^2+2^2+3^2+4^2+5^2 = 55$$
$$\sum x_i y_i = 1\times2+2\times2+3\times3+4\times4+5\times4 = 51$$
$$\sum y_i = 2+2+3+4+4 = 15$$

これらを上の公式に代入

$$S_{xx} = \sum x_i{}^2 - \frac{(\sum x_i)^2}{n} = 55 - \frac{15^2}{5} = 10$$
$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 51 - \frac{15\times15}{5} = 6$$

S_{xx}, S_{yy} を使うと，標本の標準偏差 s_x や，相関係数 r などは

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{S_{xx}}{n-1}}$$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

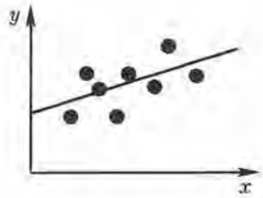
と書け，式が簡潔になる．

A. 回帰直線の求め方とその検定

直線回帰

simple linear regression

回帰直線の求め方



平面上の n 個の点 $(x_i, y_i)\ (i=1, 2, \cdots, n)$ にあてはめた直線を回帰直線 regression line と呼ぶ．その式を $\boldsymbol{y=a+bx}$ とすると，グラフ上，a は直線の y 切片，b は傾きを表す．点の配置にもっともよくあてはまる回帰直線は，次頁の**最小二乗法** least squares method の原理で求めることができ，その係数 a, b は次式で与えられる．

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2} = \frac{\sum x_iy_i-(\sum x_i \cdot \sum y_i)/n}{\sum x_i^2-(\sum x_i)^2/n} \qquad ①$$

$$a = \bar{y}-b\bar{x} = \frac{1}{n}\left(\sum y_i - b\sum x_i\right) \qquad ②$$

ここに，各要素は

$\sum x_i$…………x をそのままたし合わせたもの

$\sum x_i^2$ ………x を 2 乗してたし合わせたもの

$\sum x_i y_i$ ………x と y の積をたし合わせたもの

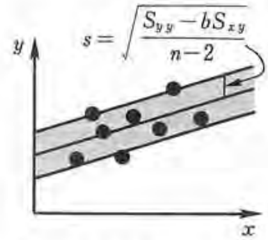
$\sum y_i$…………y をそのままたし合わせたもの

を表す．また S_{xx} は x の偏差平方和，S_{xy} は x, y の偏差積和（☞前頁）を表す．

回帰の適合度

このようにして求まる回帰直線がどの程度データにフィットしているかは，実測値 y_i と推定値 $Y_i\ (=a+bx_i)$ の差（**回帰残差** residual）から

$$s = \sqrt{\frac{\sum (y_i-Y_i)^2}{n-2}} = \sqrt{\frac{S_{yy}-b\times S_{xy}}{n-2}}$$



を計算して判定する．ここに S_{yy} は y の偏差平方和を表す．第 2 の等式の b は，回帰直線の傾きを表す（☞209頁）．

この s は**回帰直線を基準にしたデータの標準偏差***（回帰残差の期待値）に相当し，s^2 は回帰の**残差分散** residual variance と呼ばれる．

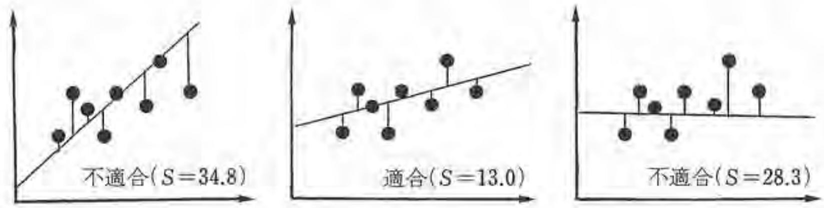
* 以下，**回帰直線からの標準偏差**と呼ぶ．

最小二乗法の原理と回帰直線

ちょっと掘り下げて

n 個の点に対する回帰直線はたくさん引ける．

その中でデータにもっともフィットしたものとして，“各点 (x_i, y_i) から回帰直線までの垂直距離の 2 乗和 S（回帰からの偏差平方和）が最小となる場合”の直線の式を求める．



これを解くには，回帰係数 a, b を未知数として，S を最小にするときの a, b を，次の微分を利用して決める．

今，求める回帰式を $y=a+bx$ として，n 個の点すべてについて，実測値 y_i と回帰直線上の推定値 Y_i との差（回帰残差）y_i-Y_i の平方和を求める．

$$S=\sum(y_i-Y_i)^2 \quad \longleftarrow \quad \boxed{Y_i=a+bx_i}$$
$$=\sum(y_i-a-bx_i)^2$$

この偏差平方和 S を最小にする a, b を求めるには，S を a, b の関数とみて，a, b で偏微分し，その関数 $\delta S/\delta a$, $\delta S/\delta b$ が，それぞれが 0 になる場合を計算すればよい（S の最小値問題）．

ここで，$T=y_i-a-bx_i$ とおいて，a について偏微分すると，

$$\frac{\delta S}{\delta a}=\frac{\delta\sum(y_i-a-bx_i)^2}{\delta a}$$
$$=\frac{\delta\sum T^2}{\delta T}\times\frac{\delta T}{\delta a}$$
$$=2\sum(y_i-a-bx_i)\cdot(-1)$$
$$=-2\sum(y_i-a-bx_i)=0 \qquad ③$$

➡T^2 を T で微分してから，T を a で微分．なお $\sum$ したものを微分するのと，微分してから $\sum$ するのは同じ結果．

$$\frac{\delta\sum T^2}{\delta T}=\sum\frac{\delta T^2}{\delta T}$$

同様に，b について偏微分すると，

$$\frac{\delta S}{\delta b}=\frac{\delta\sum(y_i-a-bx_i)^2}{\delta b}$$
$$=\frac{\delta\sum T^2}{\delta T}\times\frac{\delta T}{\delta b}$$
$$=2\sum(y_i-a-bx_i)\cdot(-x_i)$$
$$=-2\sum(x_iy_i-ax_i-bx_i^2)=0 \qquad ④$$

➡T^2 を T で微分してから，T を b で微分．

式③④を整理すると，

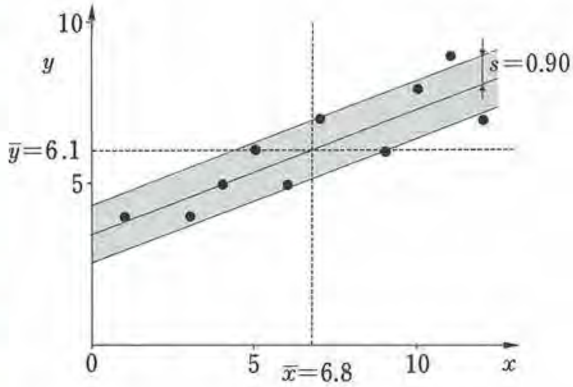
$$\boxed{\begin{aligned} n\cdot a+(\textstyle\sum x_i)b &= \textstyle\sum y_i \\ (\textstyle\sum x_i)a+(\textstyle\sum x_i^2)b &= \textstyle\sum x_iy_i \end{aligned}}$$

となり，これを正規方程式と呼ぶ．これは，a, b に関する 2 元 1 次の連立方程式になっており，その解が前頁の式①②で，a, b の最小二乗推定値にあたる．

回帰直線の計算

例 7-1 (1) 次の10点を通る回帰直線 $y=a+bx$ を，最小二乗法で求めよ．
(2) 回帰直線からの標準偏差 s を求めよ．

x	y	x^2	xy	y^2
1	4	1	4	16
3	4	9	12	16
4	5	16	20	25
5	6	25	30	36
6	5	36	30	25
7	7	49	49	49
9	6	81	54	36
10	8	100	80	64
11	9	121	99	81
12	7	144	84	49
$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum x_i y_i$	$\sum y_i^2$
68	61	582	462	397



解

(1) 回帰係数 a, b を，206頁の式①，②から求める．

$$b=\frac{S_{xy}}{S_{xx}}=\frac{47.2}{119.6}=0.39$$

$$a=\frac{\sum y_i-b\sum x_i}{n}=\frac{61-0.39\times 68}{10}=3.4$$

ここで，S_{xy} は，x, y の偏差積和，S_{xx} は x の偏差平方和で，基本和 $\sum x_i$, $\sum y_i$, $\sum x_i^2$, $\sum x_i y_i$ から求める．

$$S_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y})=\sum x_i y_i-\frac{\sum x_i\sum y_i}{n}=462-\frac{68\times 61}{10}=47.2$$

$$S_{xx}=\sum(x_i-\bar{x})^2=\sum x_i^2-\frac{(\sum x_i)^2}{n}=582-\frac{68^2}{10}=119.6$$

したがって，回帰式は $y=3.4+0.39x$

(2) データの回帰直線からの標準偏差は，

$$s=\sqrt{\frac{\sum(y_i-Y_i)^2}{n-2}}=\sqrt{\frac{S_{yy}-b\cdot S_{xy}}{n-2}}=\sqrt{\frac{24.9-0.39\times 47.2}{10-2}}=\sqrt{\frac{6.49}{8}}=0.90$$

ここに，

$$S_{yy}=\sum y_i^2-\frac{(\sum y_i)^2}{n}=397-\frac{61^2}{10}=24.9$$

回帰直線からの標準偏差（回帰残差の期待値）を求める公式

ちょっと掘り下げて

$$s=\sqrt{\frac{\sum(y_i-Y_i)^2}{n-2}}$$ 回帰からの標準偏差の定義式で，残差分散の平方根に相当

$$=\sqrt{\frac{S_{yy}-b\times S_{xy}}{n-2}}$$ 回帰係数 b を利用した公式

$$=\sqrt{\frac{(1-r^2)S_{yy}}{n-2}}$$ 相関係数 r を利用した公式

ここに $x=x_i$ に対して y_i が実測値，Y_i が回帰直線上の推定値で $Y_i=a+bx_i$，y_i-Y_i は回帰残差．

〈証　明〉

$$\begin{aligned}\sum(y_i-Y_i)^2&=\sum(y_i-a-bx_i)^2 \qquad \longleftarrow \boxed{a=\bar{y}-b\bar{x}}\\&=\sum(y_i-\bar{y}+b\bar{x}-bx_i)^2\\&=\sum\{(y_i-\bar{y})-b(x_i-\bar{x})\}^2\\&=\sum\{(y_i-\bar{y})^2-2b(x_i-\bar{x})(y_i-\bar{y})+b^2(x_i-\bar{x})^2\}.\\&=S_{yy}-2bS_{xy}+b^2S_{xx} \qquad \longleftarrow \boxed{b=\frac{S_{xy}}{S_{xx}}}\\&=S_{yy}-bS_{xy}\end{aligned}$$

さらに，

$$\begin{aligned}&=S_{yy}-\frac{{S_{xy}}^2}{S_{xx}} \qquad \longleftarrow \boxed{r^2=\frac{{S_{xy}}^2}{S_{xx}S_{yy}}}\\&=S_{yy}(1-r^2)\end{aligned}$$

したがって，

$$s=\sqrt{\frac{\sum(y_i-Y_i)^2}{n-2}}=\sqrt{\frac{S_{yy}-S_{xy}\cdot b}{n-2}}=\sqrt{\frac{(1-r^2)S_{yy}}{n-2}}$$

これを用いると，b の標準誤差 s_b（☞212頁）は，x, y の標準偏差 s_x, s_y を使って，

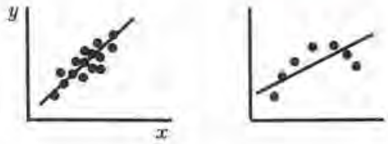
$$s_b=\frac{s}{\sqrt{S_{xx}}}=\sqrt{\frac{1-r^2}{n-2}}\times\frac{s_y}{s_x}$$

と表すこともできる．

回帰直線の計算：チェックポイント

特殊例　　対　策

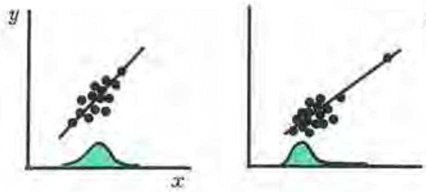
1) 直線関係と考えてよいか？



●y を変数変換して直線化（☞222頁）
●曲線回帰を試みる（整次多項式☞223頁）

2) x の分布に偏りはないか？

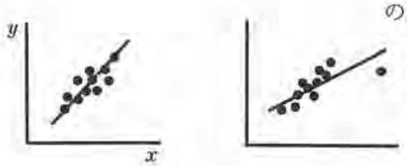
歪度の強い分布，
端の点の影響力大



●できるだけ対称な分布となるよう x を変数変換（☞270頁）

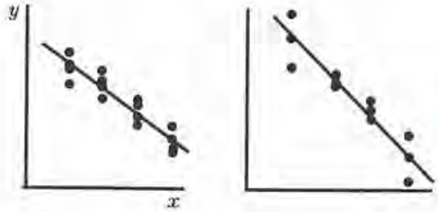
3) 飛び離れ点はないか？

飛び離れ点，y 方向偏差
の 2 乗で影響



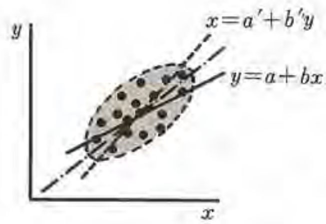
●データの点検，棄却検定（☞284頁）

4) y の分散は，x の値によらず均一とみなせるか？



●重み付き回帰を試みる（☞258頁）
●y の変数変換による場合そのまま曲線回帰（非線形回帰☞223頁）

5) 変数 x, y のとり方は妥当か？



x から y を予測するのか？
y から x を予測するのか？
楕円の長軸を求めるのか？
（☞次頁）

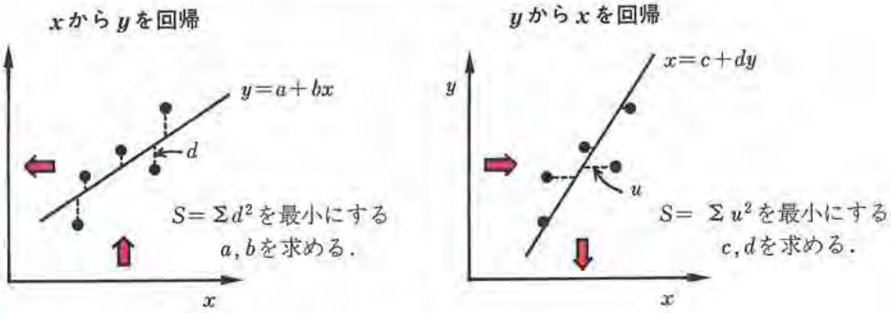
教 訓

回帰の妥当性は，必ずグラフで確認．

予 備 知 識

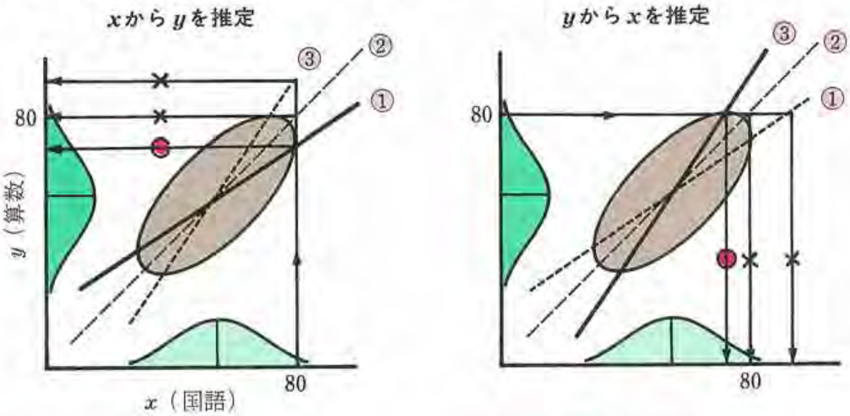
回帰の方向性と求心性

何を独立変数にするかで，回帰式が変わる（回帰の方向性）



回帰直線を 2 本引ける意味（回帰の求心性）

xからyを推定する場合と，yからxを推定する場合の違いは，次のようにx=国語の点数，y=算数の点数と置き換えて考えるとわかりやすい．



◀①回帰直線（xからyを回帰）
②等確率楕円長軸
③回帰直線（yからxを回帰）

〈国語から算数の点を推定〉

いま，よくできる学生で国語が80点のとき，①式を用いると，算数の推定値は80点より低くなるが，③式を用いると逆に80点より高くなってしまう．その中間としての②式は楕円の長軸にあたり，国語点と同じ値を推定する．常識的に考えて“算数点分布の中心に近い側に推定”する①の方式が一番妥当と考えられる．

〈算数から国語の点を推定〉

いま，よくできる学生で算数が80点のとき，③式を用いると，国語の推定値は80点より低くなるが，①式を用いると逆に80点より高くなってしまう．その中間としての②式は楕円の長軸にあたり，算数点と同じ値を推定する．常識的に考えて“国語点分布の中心に近い側に推定”する③の方式が一番妥当と考えられる．

教 訓

回帰式の変形は要注意，
計算前に，どの変数からどの変数を推定するのかよく考えて．

$$y=a+bx$$
$$\not\downarrow^{*}$$
$$x=(y-a)/b$$

この図では，変数xと変数yの相関を0.6～0.7程度としたが，一般に相関が弱いほど，推定値はより安全となるよう中心に向かって偏よる．逆に，相関が強いと，2つの回帰式は一致するので，このような推定の求心性の問題は起こらない．最小二乗法はきわめて合理的にできていると，感心させられる．

* x, yの相関が強いときまたは回帰のフィットが良好なときは，実用上変換しても問題はない．

回帰係数の検定

tests on the regression parameters

回帰の有意性の検定 ($H_0 : \beta=0$ の検定)

β は傾き b の母数

test for significance of the regression; test for regression slope

検定の原理と手順

意味のある回帰かどうかを検定するには，その傾き b が有意に 0 から偏っているかどうかを検定すればよい*．そこで，b の標準誤差を求め，それに照らして，標本から求めた b 値の有意性を判定する．

* $H_0 : \beta=0$
$H_1 : \beta\neq 0$

まず，b の標準誤差は，

$$\boxed{s_b = \frac{s}{\sqrt{S_{xx}}}}$$

で求まる (☞215頁)．ここで s は回帰直線からの標準偏差で，

$$s = \sqrt{\frac{S_{yy} - b \cdot S_{xy}}{n-2}}$$

b 値の偏り度を求めるため，それを s_b でわって標準化する．

$$\boxed{t = \frac{b}{s_b} = \frac{b \cdot \sqrt{S_{xx}}}{s}}$$

この統計量 t は，自由度 $n-2$ の t 分布に従うので，結局，標本の b 値の有意性を判定するには，t 値の生じる確率を t 分布表で調べればよいことになる．

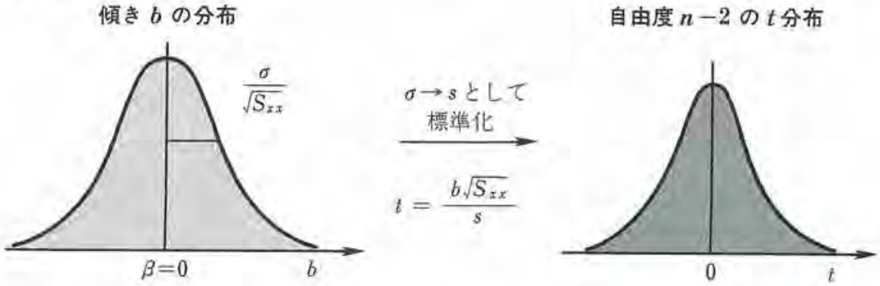
[母回帰係数 β の信頼区間]

標本の傾き b に対する母回帰係数 β の95%信頼区間は，

$$\boxed{\beta = b \pm t_{0.05} \frac{s}{\sqrt{S_{xx}}}}$$

ここに $t_{0.05}$ は，自由度 $n-2$，$P=0.05$の t 値．

上の検定法は，裏返すとこの β の信頼区間が 0 を含むかどうかを調べたことになっている．



σ^2 は残差分散 s^2 の母数

y 切片の検定 (H$_0$: $\alpha=a_0$ の検定) test for regression intercept

α は y 切片 a の母数

(検定の原理と手順)

ある標本から求めた回帰直線の y 切片 a が，一定の値 a_0 から有意に偏っているかどうかを検定する*．このためには，$a-a_0$ を a の標準誤差 s_a で標準化する．

* H$_0$: $\alpha=a_0$
H$_1$: $\alpha\neq a_0$

まず，a の標準誤差は，

$$s_a = s\sqrt{\frac{\sum x_i^2}{nS_{xx}}}$$

で求まる (☞215頁)．ここで，s は回帰直線からの標準偏差で，前頁の回帰の有意性の検定の場合と同様に求める．

a 値の偏り度を求めるため，それを s_a でわって標準化する．

$$t = \frac{a-a_0}{s_a} = \frac{a-a_0}{s}\sqrt{\frac{nS_{xx}}{\sum x_i^2}}$$

この統計量 t は，自由度 $n-2$ の t 分布に従うので，結局，標本の a 値の有意性を判定するには，t 値の生じる確率を t 分布表で調べればよいことになる．

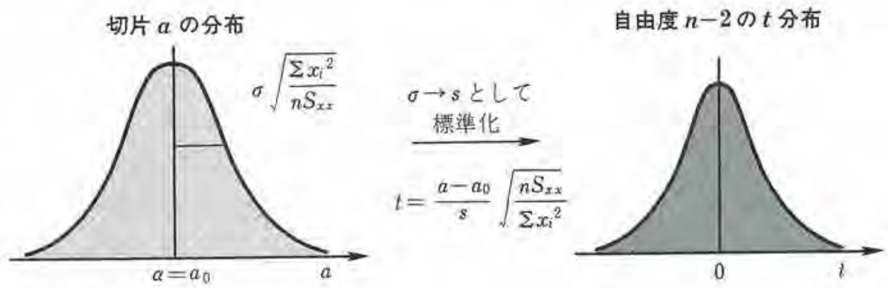
[母回帰係数 α の信頼区間]

標本の傾き a に対する母回帰係数 α の95%信頼区間は，

$$\alpha = a \pm t_{0.05}\, s\sqrt{\frac{\sum x_i^2}{nS_{xx}}}$$

ここに $t_{0.05}$ は，自由度 $n-2$，$P=0.05$の t 値．

上の検定法は，裏返すとこの α の信頼区間が $\alpha=a_0$ を含むかどうかを調べたことになっている．

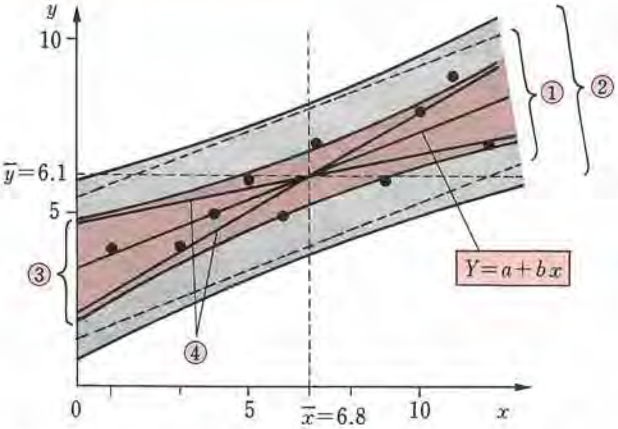


σ^2 は残差分散 s^2 の母数

回帰係数の検定

例 7-2 例 7-1 のデータについて，
(1) a, b の信頼限界を求めよ．またこれから回帰の有意性を判定せよ．
(2) y の推定値の95%信頼区間を求めよ*．

* ☞216頁



▲回帰直線の信頼区間：
①回帰残差の95%信頼区間 [$Y \pm t_{0.05} \cdot s$]
②y の推定値（Y）の95%信頼区間
③Y の平均値の95%信頼区間
④傾き b の95%信頼区間に相当する Y の領域

解

(1) a, b の標準誤差 s_a, s_b は，

$$s_a = s\sqrt{\frac{\sum x_i^2}{nS_{xx}}} = 0.90\sqrt{\frac{582}{10 \times 119.6}} = 0.63$$

$$s_b = \frac{s}{\sqrt{S_{xx}}} = \frac{0.90}{\sqrt{119.6}} = 0.082$$

ここに，s, S_{xx} の計算は例 7-1 のとおりである．

自由度 $n-2=8$，両側確率 0.05 の t 値（$t_{0.05}$）を t 分布表より求めると，$t_{0.05}=2.31$．

これから a の母係数 α の95%信頼区間は，

$$\boxed{a - t_{0.05}\, s_a \leqq \alpha \leqq a + t_{0.05}\, s_a}$$

$$3.4 - 2.31 \times 0.63 \leqq \alpha \leqq 3.4 + 2.31 \times 0.63$$

$$1.9 \leqq \alpha \leqq 4.9$$

また b の母係数 β の95%信頼区間は，

$$\boxed{b - t_{0.05}\, s_b \leqq \beta \leqq b + t_{0.05}\, s_b}$$

$$0.39 - 2.31 \times 0.082 \leqq \beta \leqq 0.39 + 2.31 \times 0.082$$

$$0.20 \leqq \beta \leqq 0.58$$

この β の95%信頼区間は $\beta=0$ を含まないので，回帰は有意と判断できる．逆にいうと，b は有意に 0 から偏っており，その偏りの程度は，

$$t = \frac{b-0}{s_b} = \frac{0.39}{0.082} = 4.75$$

として，自由度10−2 の t 分布表から判定できる（$t_{0.01}=3.355$から，$P<0.01$）

(2) x における，y の推定値（Y）の95%信頼区間は，

$$y = a + bx \pm t_{0.05}\, s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

$$= 3.4 + 0.39 \times x \pm 2.31 \times 0.90\sqrt{1 + \frac{1}{10} + \frac{(x-\bar{x})^2}{119.6}}$$

$$= 3.4 + 0.39 \times x \pm 0.19\sqrt{131.5 + (x-6.8)^2} \quad \longleftarrow \boxed{\bar{x}=6.8}$$

この式は x の関数で，図の②に示すような双曲線型の 2 次曲線を表す．

回帰係数 a, b の標準誤差 s_a, s_b の導入

ちょっと掘り下げて

計算の前提 (1) x を基準にして y を回帰するので，回帰の誤差は y 方向のみについて考える．

(2) x の位置によらず，回帰直線からの標準偏差 s は一様．

傾き b の標準誤差 s_b の期待値を求めるには，それを2乗して，分散の期待値 $Var(b)$ の形で求める．

$$
\begin{aligned}
Var(b) &= Var\left(\frac{S_{xy}}{S_{xx}}\right)\\
&= Var\left[\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\right] \quad \longleftarrow \quad \boxed{\sum(x_i-\bar{x})\bar{y}=0}\ \ \boxed{Var(ky)=k^2Var(y)}\\
&= \left[\frac{1}{\sum(x_i-\bar{x})^2}\right]^2 Var\{\textstyle\sum(x_i-\bar{x})y_i\}\\
&= \left[\frac{1}{\sum(x_i-\bar{x})^2}\right]^2 \textstyle\sum(x_i-\bar{x})^2 Var(y_i) \quad \longleftarrow \quad \boxed{Var(y_i)=\sigma^2}\\
&= \left[\frac{1}{\sum(x_i-\bar{x})^2}\right]^2 \textstyle\sum(x_i-\bar{x})^2\sigma^2\\
&= \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}
$$

➡分子のこの項はなくなる．

➡x を基準にして y を回帰するので x 方向には誤差がないと仮定．したがって，$\sum(x_i-\bar{x})^2$ は定数として2乗してくくり出される．

➡y_i の分散の期待値は i によらず一定と仮定．σ は y の標準偏差 s_y でなく，回帰直線からの標準偏差 s の母数であることに注意．

同様に，y 切片 a の標準誤差 s_a の期待値は，それを2乗して，分散の期待値 $Var(a)$ 形で求める．

$$
\begin{aligned}
Var(a) &= Var(\bar{y}-b\bar{x})\\
&= Var(\bar{y})+\bar{x}^2\times Var(b) \quad \longleftarrow \quad \boxed{\begin{aligned}Var(\bar{y})&=\frac{\sigma^2}{n}\\ Var(b)&=\frac{\sigma^2}{S_{xx}}\end{aligned}}\\
&= \frac{\sigma^2}{n}+\frac{(\sum x_i)^2}{n^2}\times\frac{\sigma^2}{S_{xx}}\\
&= \sigma^2\times\frac{S_{xx}+\frac{(\sum x_i)^2}{n}}{nS_{xx}}\\
&= \sigma^2\times\frac{\sum x_i^2-\frac{(\sum x_i)^2}{n}+\frac{(\sum x_i)^2}{n}}{nS_{xx}}\\
&= \sigma^2\times\frac{\sum x_i^2}{nS_{xx}}
\end{aligned}
$$

回帰直線の信頼区間

ちょっと掘り下げて

点の配置と推定値 Y の信頼区間

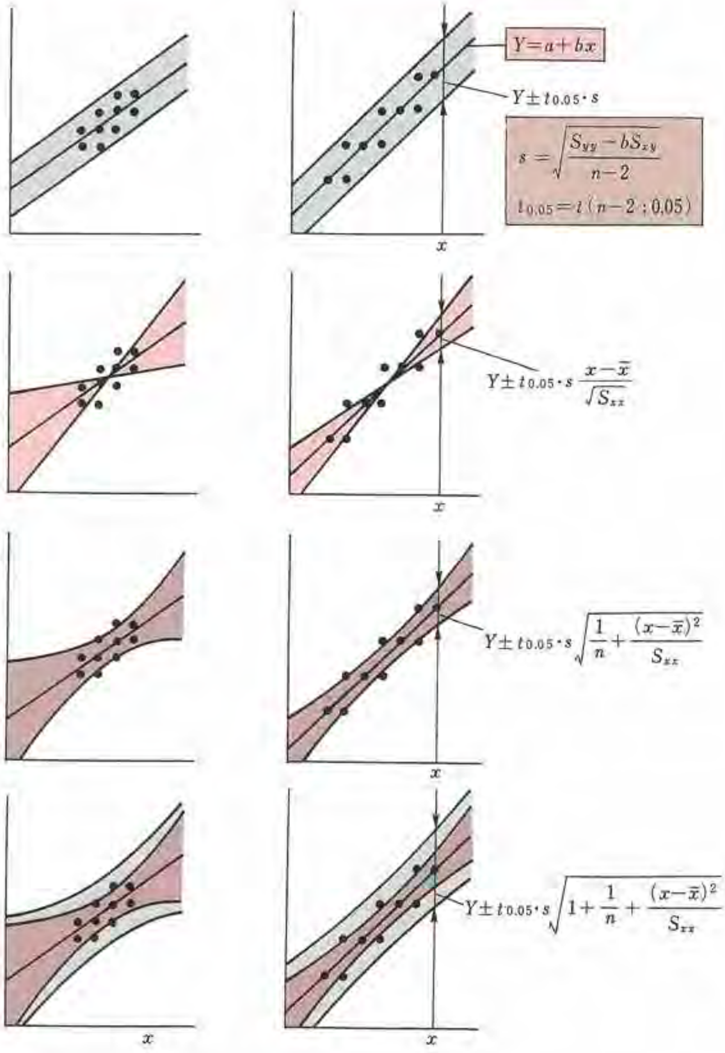
x の分布範囲が狭い　　x の分布範囲が広い

回帰残差の95％信頼区間に相当する Y の領域（回帰係数の誤差を含まない）

傾き b の95％信頼区間に相当する Y の領域（回帰残差を含まない）

Y の平均値の95％信頼区間（回帰係数 a, b の誤差の合成）

個々の Y の95％信頼区間（回帰残差と回帰係数 a, b の誤差の合成）



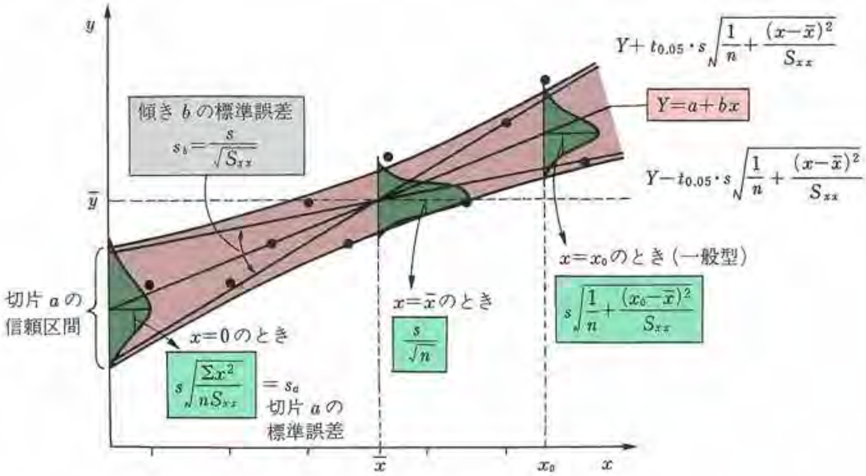
左列と右列のデータは，回帰からの標準偏差 s は同じであるが，基準変数（独立変数）x の分布範囲（S_{xx} に相当）が異なる．その範囲が狭い左列のデータでは推定値 Y の信頼区間が両端で大きくなる．これは主として2段目に示す傾き b の信頼区間が広がるため．数式上，S_{xx}（x の偏差平方和）が分母にあるので S_{xx} が小さくなると，Y の信頼区間が広がる．

教 訓

基準変数 x の範囲が狭いと，回帰はばらつく．

■ x の位置で変わる，推定値 Y の平均値の標準誤差

——回帰係数の標準誤差 s_a，s_b との関係



◀$n=10$点について x から y を直線で回帰した．中央の直線が推定値 $Y=a+bx$ で，その上下の曲線が，Y の平均値の95％信頼区間を示す．

Y の平均値の標準誤差は，回帰係数 a, b の誤差を合成したものに相当し，x の値に依存する．

■ 推定値 Y の標準誤差の導入：分散の形でその期待値を求める．

Y の平均値の誤差分散

（＝回帰係数の誤差分散）

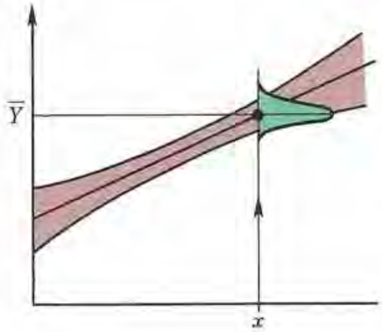
$$
\begin{aligned}
Var(Y) &= Var(a+bx)\\
&= Var(\bar{y}-b\bar{x}+bx)\\
&= Var\{\bar{y}-b(x-\bar{x})\}\\
&= Var(\bar{y})+(x-\bar{x})^2\,Var(b)\\
&= \frac{\sigma^2}{n}+(x-\bar{x})^2\frac{\sigma^2}{S_{xx}}\\
&= \sigma^2\left\{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right\}
\end{aligned}
$$

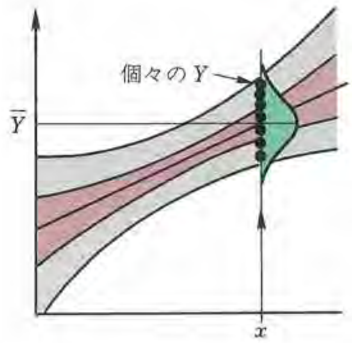
個々の Y の誤差分散

（＝残差分散＋回帰係数の誤差分散）

$$
\begin{aligned}
Var(y-Y) &= Var(y)+(-1)^2\,Var(Y)\\
&= \sigma^2+\sigma^2\left\{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right\}\\
&= \sigma^2\left\{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right\}
\end{aligned}
$$

←**誤差分散** error variance
統計量の分散で，標準誤差の2乗に相当．
$Var(\ )$ は分散の期待値
σ^2 は残差分散の母数





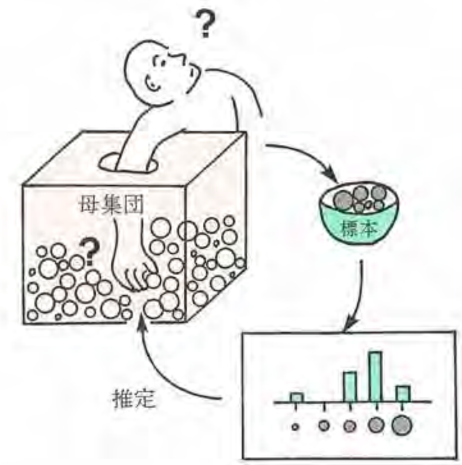
CHAPTER

9

統計の正しい利用と解釈

A 標本の偏り

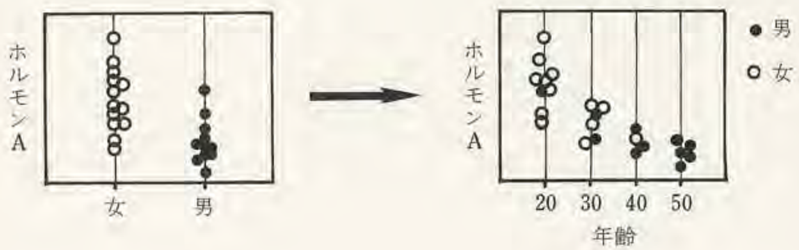
統計の主な目的は少数の標本から，効率よく母集団の情報をつかむことにある．しかし，標本の集め方に偏りが入ると，母集団の実態とはかけ離れた結論を導いてしまう．



統計の機能１：少数の情報から，全体を推し量る（☞4頁）．

1) 調査上の偏り

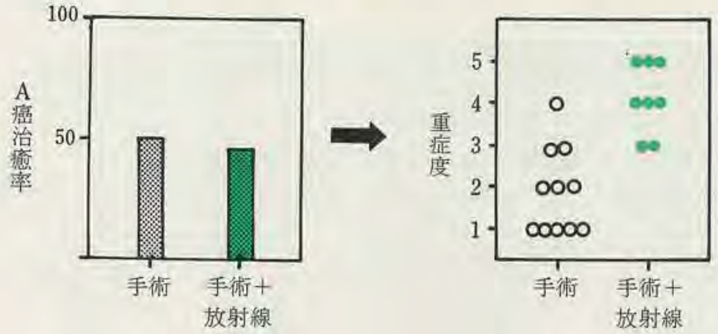
ケーススタディ A_1 ／ 男女差？ 実は年齢差だった！



／問題点と対策／

比較する２群の基本特性に差がないかを常にチェック．この例では年齢・性別で層別化して分析（age-matched, sex-matched analysis）する必要があった．

ケーススタディ A$_2$ ／ 治療効果の違い？　実は症例の重症度が違っていた！

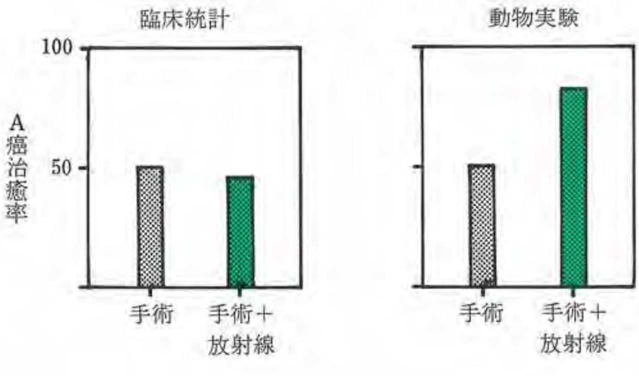


／問題点／

過去のデータを，ただ治療法の違いで調査・分類しただけで，それ以外の要因は考慮されていなかった．このような後ろ向き調査（retrospective study）では，症例のランダム化は困難．

／対　策／

計画して前向きに調査（prospective study）する必要がある．また，純系動物で実験的に確認することも重要．この例では，下図右のように，放射線治療を併用した方がよいという結論になっていたかもしれない．



Key Point

無計画な臨床調査（疫学調査）は要注意！

症例を後からランダム化することは困難．
計画的にランダム化しなければ，正しい調査結果は得られない．

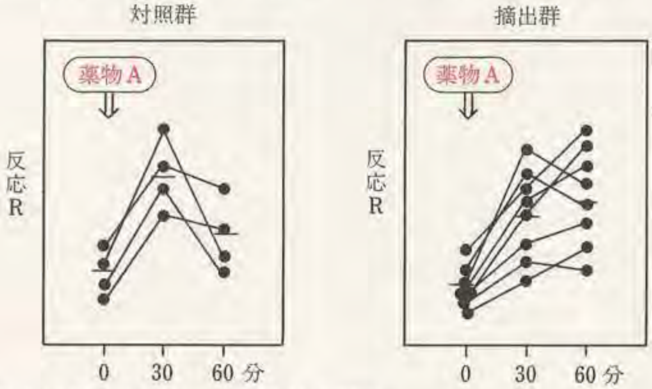
2) 実験上の偏り

ケーススタディ A_3 ／ 対照群にも類似の操作を行ったが…

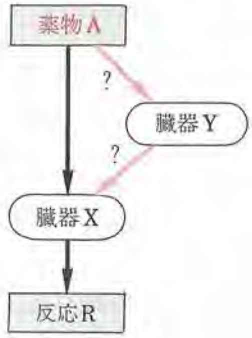
(目的) 薬物 A に対する臓器 X の反応(R)に，臓器 Y が関与するか？

(方法) ウサギ12匹を，臓器 Y 摘出群（**摘出群**：8 匹）と単に模擬手術 sham surgery を行った群（**対照群**：4 匹）に分け反応 R を比較．

(成績)



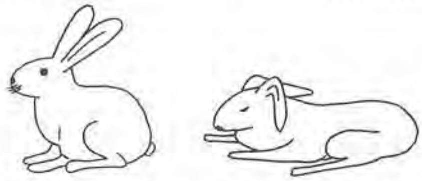
(結論) 薬物 A に対する反応に，臓器 Y が関与．



問題点

臓器 Y を摘出すること自体には，とくに問題ないが，……

		対照(模擬手術)群	臓器 Y 摘出群
n	(匹)	4	8
手術時間	(分)	10±3	40±20
出血量	(m*l*)	1.5±0.5	8.0±4.0
術後発熱例	(匹)	0	5
食事摂取	(g/日)	40±7	25±10
体　重	(kg)	2.5±0.2	2.2±0.5



(1) 手術そのものの影響（出血・感染）が，両群でマッチしていなかった．

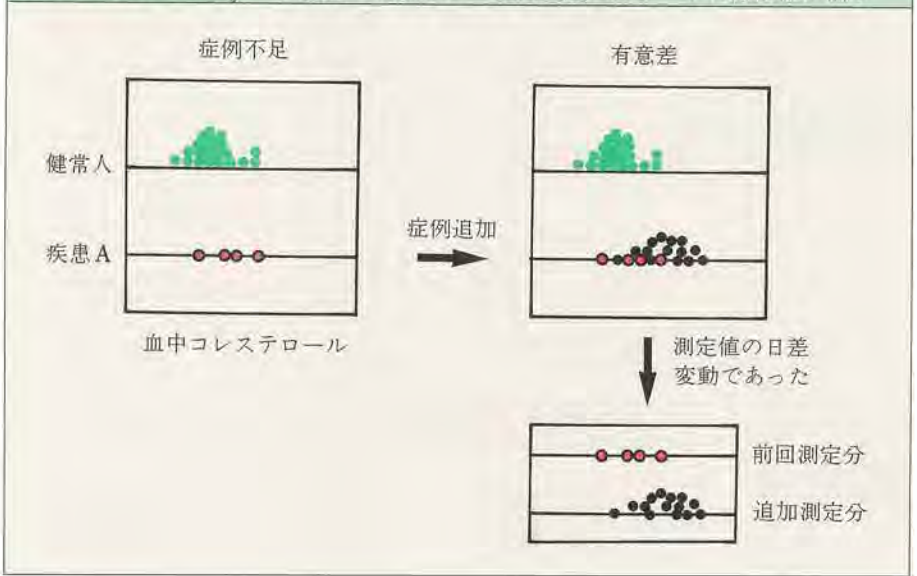
(2) 2 群のデータ数はそろえておく方がよかった．

(3) 有意差ではないが，体重や反応 R の基礎値（0 分の値）にも差がみられた．

/対　策/

対照群との間に不適合（mismatch）が生じないか，さまざまな角度から予備調査しておく．

ケーススタディ A$_4$ ／ 健常群と有意差？　実は追加したデータの測定値に偏り

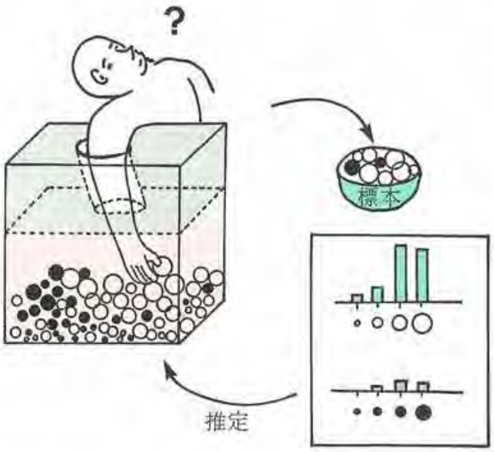


/問題点/

後から，部分的にデータを追加すると，実験条件に差を生じることがある．

/対　策/

計画的な実験でデータの追加を避ける．やむをえないときでも，対照群を同時に追加して，実験条件に偏りがないかを確認．



教訓

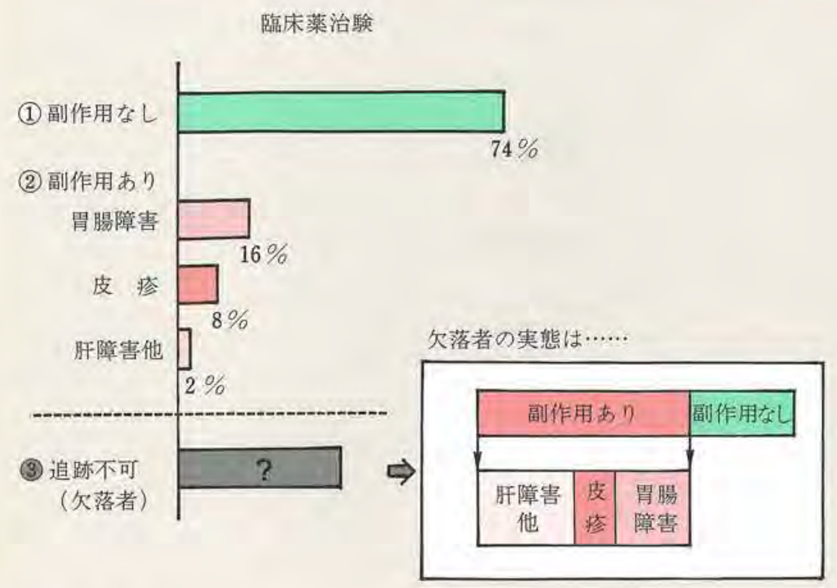
ケースコントロールされていない研究データには危険が一杯．

• 実験による研究の場合，周到な計画で対象をランダム化すれば，偏りをある程度防ぎうる．

• 調査研究の場合，対象のランダム化は，きわめて困難．

3) 欠落値と偏り

ケーススタディ A_5 ／ 薬の効果・副作用調査，欠落データは何を意味する？



副作用のあった人の多くは途中で他院に移り，来なくなっていた．

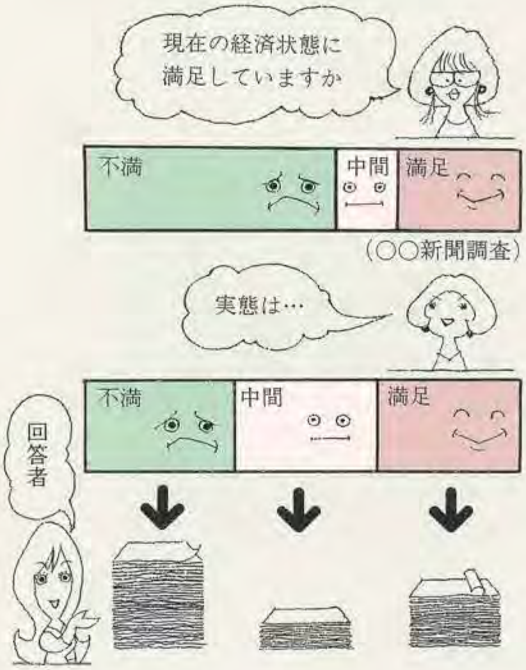
/問題点/

周到な計画で，対象をランダムに選び出しても，欠落値が多ければ偏りを生じうる．

/対　策/

追跡調査には一定の厳密さを必要とする．症例欠落の可能性を十分予測し，何らかの方策を用意しておく．

ケーススタディ A6 ／ ランダム調査？　実は回答者に偏りが！



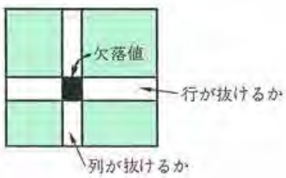
対象者をランダムに抽出した世論調査.
しかし，回答したのは，なんらかの不満を持った人がほとんど.

Key Point

欠落値の 2 つ統計学的欠陥

(1) 検定に利用できるデータ数が行，列単位で減る

部分的に欠損したデータがあると，検定上，その行または列に属す他のデータも同時に無効となる*.



(2) ランダムでなくなる：欠落値に固有の偏り

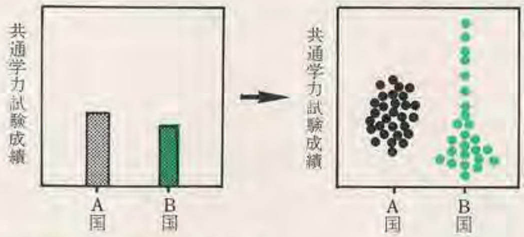
ある一定の偏りを持ったデータが欠落値になることが多い.

* 欠落値の補塡法を利用できることもある（☞200頁）.

B データの表現法……データを丸めることの危険

1) 分布の中心指標としての，平均値

ケーススタディ B₁ ／ 平均点がよいことを，素直に喜んでいいのか？



A国はつぶがそろっている．

B国はできない人も多いが，ずば抜けて優秀な人もいる．

どちらがよいか，平均値だけではわからない．

／問題点／

分布型が同じでなければ，平均値による比較は無意味．

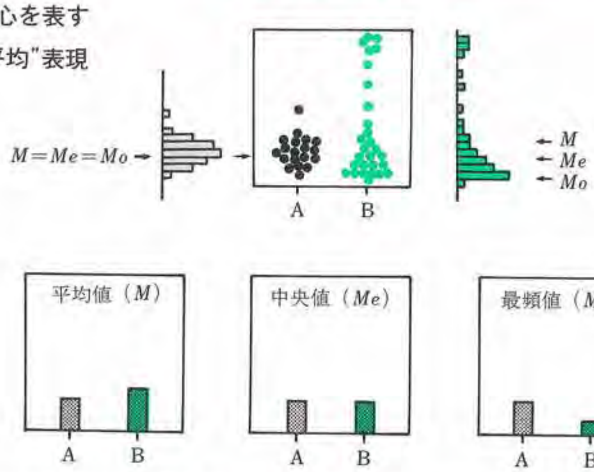
／対　策／

生のデータをありのままに示せば，誤解を生じない．意外な情報を発見できることもある．

Key Point

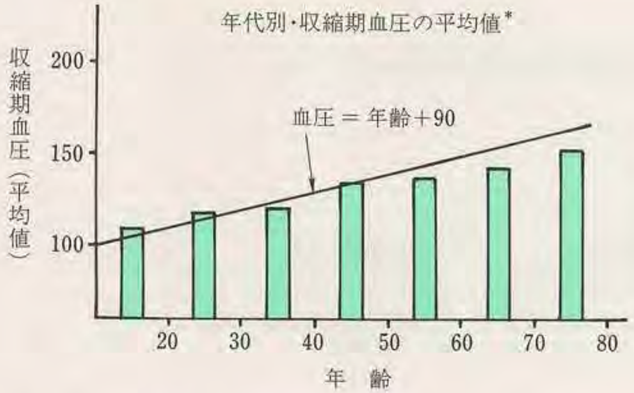
分布の中心を表す

3つの"平均"表現

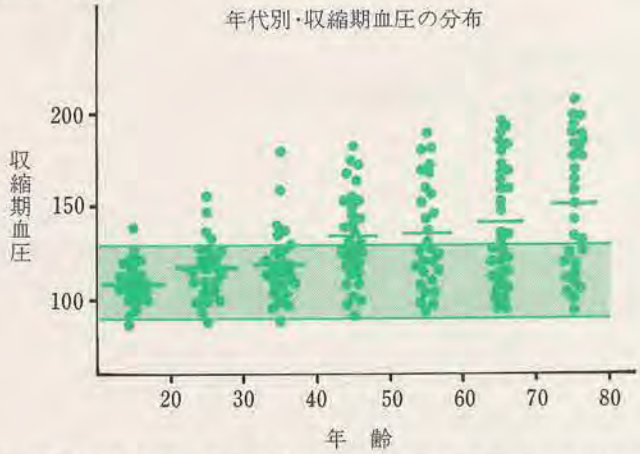


平均値か，中央値か，最頻値かによって，イメージはまったく異なる．

ケーススタディ B_2 “血圧は年齢＋90”のうそ



これをみて，血圧が低め（100前後）のお年寄りが，平均より低いことによく不安をいだく．



しかし，実際に生のデータをみると，多くのお年寄りは，年齢によらず若い頃と同じ100前後の血圧を維持している．

* モデルデータ

問題点

血圧の平均値が変化していくのは，高齢ほど健常でも動脈硬化をきたす層が増加し，その割合に応じて平均値が押し上げられてゆくためにすぎない．

このような誤解の根源は，加齢とともに各年齢層の均一性が失われていくのに，それを無視して，平均値でデータを均一化してしまったことにある．

対　策

(1) 生データの分布を示す．

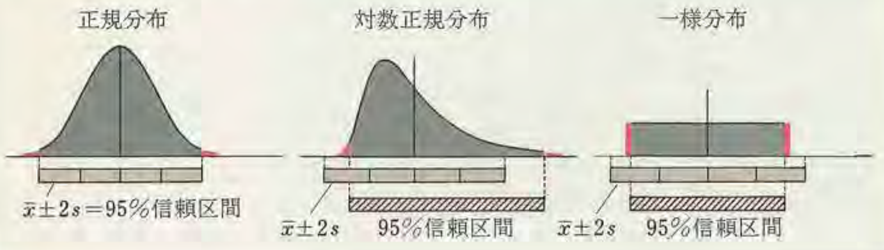
(2) 分布の広がりは均一か？　群内にサブ・グループが存在しないか？

教訓

不均一な集団も，平均値にすると均一化されてしまう．

2) 分布の広がり指標としての，標準偏差

ケーススタディ B$_3$ ／ 平均値± 2 ×標準偏差は，95％信頼区間？



／問題点／

正規分布でなければ，平均値± 2 ×標準偏差*は95％信頼区間として無意味.

／対　策／

標準偏差から信頼区間を求めるには，必ず分布型を確認.

正規分布でないときは，変数変換してその平均値± 2 ×標準偏差を求め，さらにそれを逆変換して 95％信頼区間とする（図中斜線の領域).

* 正規分布の場合，正確には平均値±1.96 標準偏差が 95％信頼区間になる.

チェビシェフの不等式

予 備 知 識

ある分布の平均値を μ，標準偏差を σ とすると，その分布型によらず，$\mu \pm k\sigma$ **の範囲にデータが存在する確率** P は，

$$\boxed{P \geqq 1-\frac{1}{k^2}}$$

これを**チェビシェフの不等式** Chebyshev's inequality と呼ぶ．これから k=1.0, 1.5, 2.0, 3.0 に対する測定値の信頼区間は次のようになる.

$\mu \pm 1.0\,\sigma$ の範囲内に 0 ％以上（$1-1/1.0^2=0$）

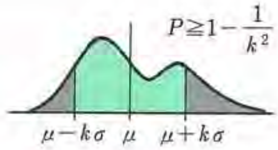
$\mu \pm 1.5\,\sigma$ の範囲内に56％以上（$1-1/1.5^2=0.556$）

$\mu \pm 2.0\,\sigma$ の範囲内に75％以上（$1-1/2.0^2=0.750$）

$\mu \pm 3.0\,\sigma$ の範囲内に89％以上（$1-1/3.0^2=0.889$）

Key Point

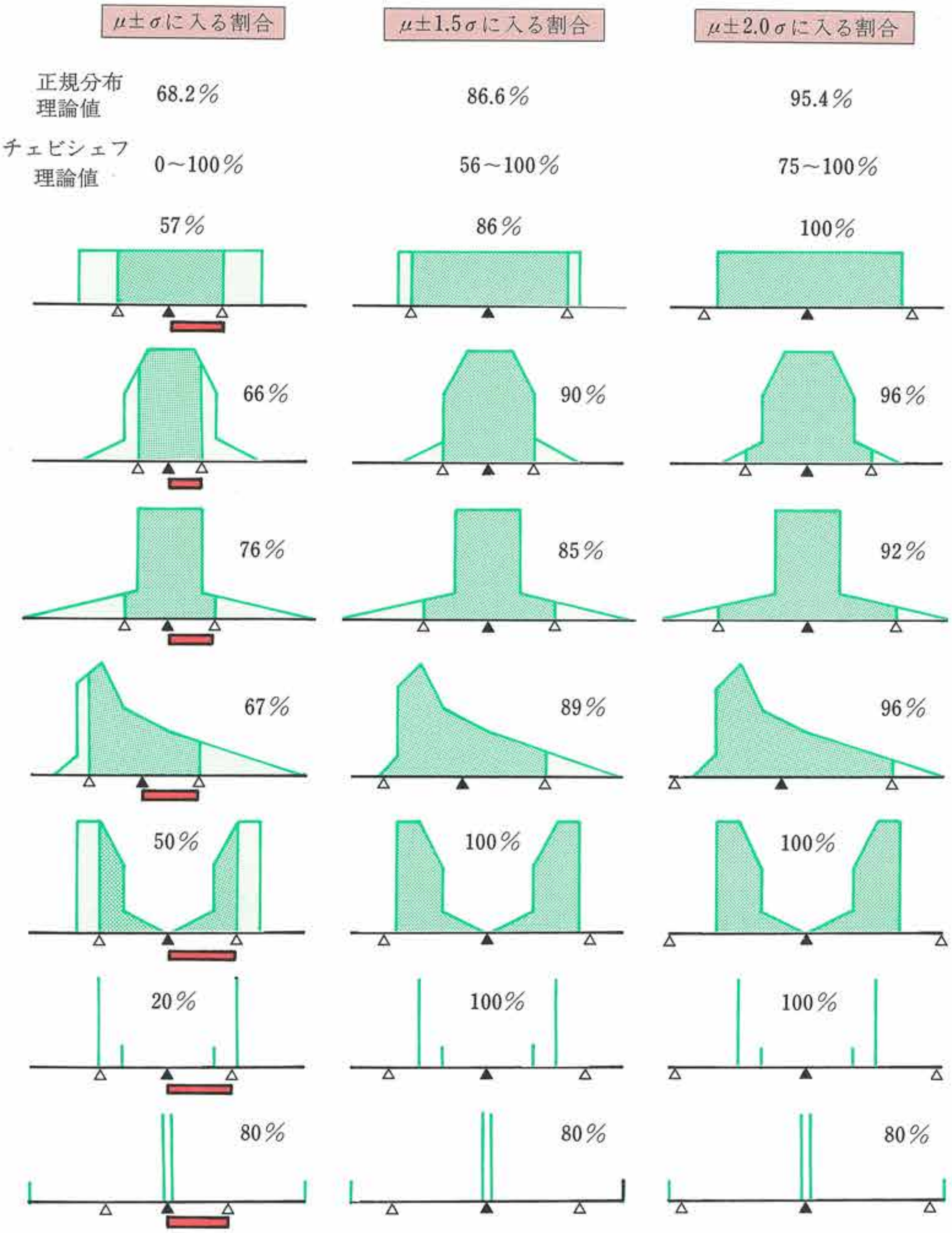
チェビシェフの不等式：
分布型によらず，その標準偏差から測定値の信頼区間を大まかに予測可能.



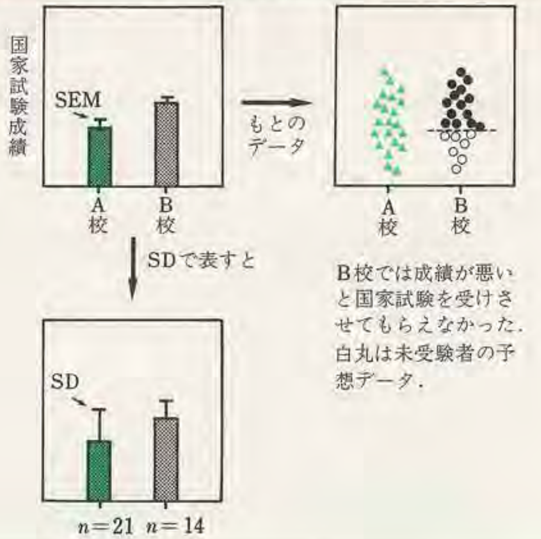
イメージレッスン

さまざまな分布の平均値，標準偏差とその信頼区間

▼横軸の▲は平均値μの位置を示す．分布の標準偏差σは，左端の各分布の下に帯状の範囲として示した．△は塗りつぶされた信頼区間$\mu\pm k\sigma$の上・下限を示す（k=1.0，1.5，2.0）．



ケーススタディ B_4 ／ "標準誤差 (SEM)" が同程度で，平均値にも有意差，しかし？



／問題点／

(1) B校では受験者に操作が加えられていたので，A校と同等に比較できない．

(2) 平均値の標準誤差 SEM を示したが，データ数が示されていない．

／対　策／

(1) 生のデータでみると，分布の偏りが一目瞭然．

(2) 標準偏差 SD を示すか，SEM とデータ数を示す．

SD, SEM の違い

SD は一般に使われる標準偏差 standard deviation の略で，本書の記号 s と同じ．その値は**個々のデータ**が，分布の中心から平均的にどれだけ隔たっているかを示す．SD の期待値はデータ数に依存しない[*1]．

$$\mathrm{SD}=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}\quad(=s)$$

SEM は，一般に，平均値の標準誤差[*2] standard error of mean の意味で使われる略語で，平均値の信頼区間を表すのに使われる．単に SE と略されることも多い．その大きさはデータ数 n に依存し，SD の $1/\sqrt{n}$ 倍になっている．

$$\mathrm{SEM}=\frac{\mathrm{SD}}{\sqrt{n}}=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n\cdot(n-1)}}$$

Key Point

SD　：個々の点の平均的なバラツキ……分布の広がりのめやすを与える．

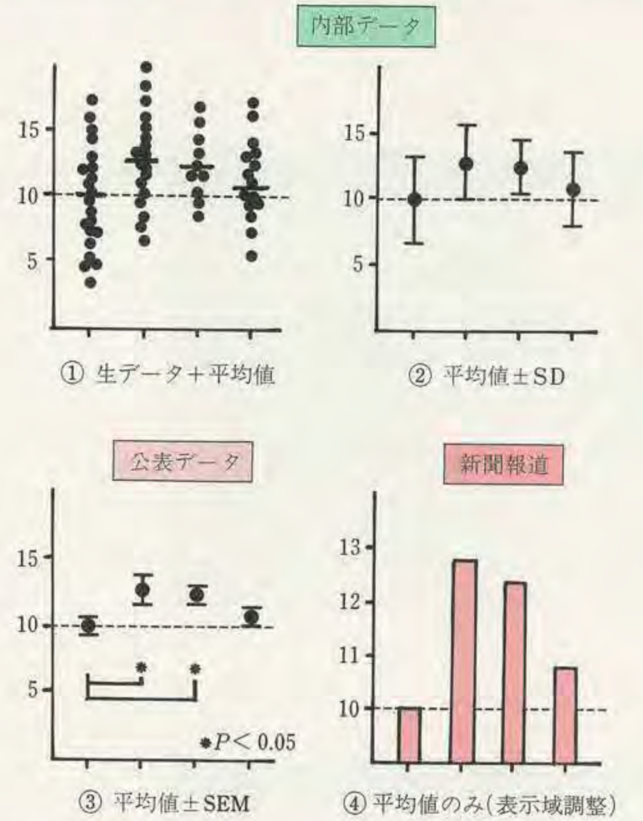
SEM：標本平均の平均的なバラツキ……平均値の信頼区間を与える．

予 備 知 識

*1 ただしその標準誤差 (standard error of SD) はデータ数に依存する．したがって，データ数が増えるほど SD はばらつかず正確に求まる．

*2 標準誤差：統計量の標本分布の標準偏差をさす．統計量は平均値に限らず，標準偏差，回帰係数など何であってもよい．

ケーススタディ B_5 ／ データの表示形式を変える目的は？



/ 問題点 /

データをまるめグラフ表示を工夫すると，情報が単純化され，よりセンセーショナルにアピールできる．しかしその分だけ大切な情報が隠れていくので要注意．

まとめ

データ表示形式のめやす

① **生データ**：インスピレーションを得やすい最善の表示形式．

② **平均値±（2×）SD**：データ構造の単純化のために必要な表示形式．

③ **平均値±SEM**：誤解を生じやすいので，できるだけ避ける．データ数を決して忘れない．

④ **平均値のみ**：絶対にすべきでない．