統計学 2019（心理とビジネスを学ぶ人のための）

［松本治彦］

2019 年 1 月

宇部フロンティア大学出版会

目　次

著者　松本治彦

発行者　宇部フロンティア大学　出版会

発効日　2019.1.31

授業概要

この科目はディプロマポリシーのうち、「柔軟な思考と表現力」を身に付けるために、また「心理学の基礎的思考方法」を学ぶための授業です。

授業ではまず、統計学の歴史について深く学び、統計学や確率の歩んだ道を振り返ってみる。その上で、統計学の様々な用語を理解して統計学を将来の仕事の道具として使いこなせるようにする。統計学は現状把握と予測のためと見られているが、じつは限られたデータを使って全体の因果関係を探る学問です。統計学を通じて得た情報から「ピンとくる」カンを働かせるのに大いに役立たせてください。

到達目標

統計値の科学的意味を的確につかむ。グループ討議を通じて、コミュニケーション能力を発揮する。

成績評価方法

毎回配布する質問・感想カードの内容（20 点)、レポート、グループ討議の態度（30 点)、定期試験（50 点）で総合評価する。

授業外学習

授業計画に沿って、資料の該当単元を熟読してきてください。講義資料を復習して下さい。

関連する科目

心理学統計法、情報処理演習Ⅰ、Ⅱ

受講の心得

毎回の授業内容の疑問点は質問カードに書き込み、次回の復習タイムで理解度を深めるように努力してください。

Ⅰ．はじめに

学生の皆さんは、高校 1，2 年の時に数学Ⅰで「データの分析」について習っていると思います。その時に「度数分布表」「ヒストグラム」「平均値」「代表値」「相対度数」「中央値」「最頻値」「箱ヒゲ図」「偏差」「分散」「標準偏差」「散布図」「相関」「共分散」「相関係数」などの統計用語について説明を受けていると思います。

しかしこの授業ではまず、統計学の歴史について深く学び、統計学や確率の歩んだ道を振り返ってみてください。その上で、統計学の様々な用語を理解して統計学を将来の仕事の道具として使いこなせるようにしてください。統計学は現状把握と予測のためと見られ

ていますが、じつは限られたデータを使って全体の因果関係を探る学問です。統計学を通じて得た情報から「ピンとくる」カンを働かせてみてください。

インターネットが発達した現在、膨大な情報の中から自分の必要な情報を選別し、それを整理する能力が必要です。また統計処理した数値がどのような意味をもつかを判断する能力も必要です。この授業の到達目標は、統計値の科学的な意味を理解することです。そのために、基本的な統計値の意味をしっかりと理解した上で、統計図、統計表の見方を学習する。そうして区間推定や検定を通じてデータの科学的な見方を身につけてください。

＊「テレビのワイドショー、新聞や週刊誌などでもけっこういい加減な結論が取り上げられていることがある。そのような間違った情報や数字に騙されないためにも、最低限の統計学の知識は、誰でも覚えておく必要がある。。

授業計画

第 1 回　　統計学の歴史　その 1
第 2 回　　統計学の歴史　その 2
第 3 回　　統計学の考え方　その 1
第 4 回　　統計学の考え方　その 2
第 5 回　　データの科学的な見方
第 6 回　　ここまでの要点整理（グループ討議とレポート 1 提出準備）
第 7 回　　具体例で統計学を学ぶ　その 1
第 8 回　　具体例で統計学を学ぶ　その 2
第 9 回　　具体例で統計学を学ぶ　その 3
第 10 回　区間推定と検定　その 1
第 11 回　区間推定と検定　その 2
第 12 回　ここまでの要点整理（グループ討議とレポート 2 提出準備
第 13 回　ｔ検定と分散分析
第 14 回　相関と予測
第 15 回　まとめ

Ⅱ．統計学の歴史

Ⅱ－1　古代からパスカル、ライプニッツまで

学問を学ぶときまず、その歴史を知ることは非常に重要です。特に統計学（確率を含む）はその歴史が古く、人類の生活の中で発展してきたことがわかります。その起源はギリシャやローマ時代まで遡ることができます。なかでも確率を扱う場合に重要なランダムネス（無作為性）の考え方は、賭博に通じるもので、古代、おそらく原始時代から存在していたと思います。賭博は人間社会が最初に発明したものの 1 つで、くじで物事を決めることは日常茶飯事であったと思います。このことは化石の中にサイコロのように面によって異なる模様を描いた骨が見つかっていることからも想像できます。

ところでギリシャやローマの時代に、国家（state）の状態（state）を調べることに関心が向けられるようになり、国家の状態を調べることを statistics というようになりました。今、統計学のことを英語で statistics といいますが、語源は古くローマ時代にさかのぼるのです。

さて確率には 2 つの側面があります。それは信念の度合いと安定した頻度です。これをもうすこし詳しく説明すると、１つは確率を人の信念や信頼の度合いとして主観的に解釈する考え方と、もう１つは頻度や対象の性質として客観的に解釈する考え方です。

ところで、パスカル以前にはこの 2 つの側面について考えた人はほとんどいませんでした。そこで、今ではパスカルらが活躍した 1660 年前後の 10 年間が確率の誕生期としているのです。パスカル（Blaise Pascal；1623～1662 年）はフランスの思想家で「人間は考える葦である」とか自然科学者でもあり「パスカルの定理」が有名ですね。現在使われている気圧の単位は彼の名前をとって「パスカル Pa」です。

確率の誕生はパスカルの個人の偉業ではなく、1660 年前後の 10 年間に多くの人たちによってもたらされた出来事です。パスカルのほか、アルノー、ライプニッツ、ホイヘンス、グラント、ヒュッデ、デ・ウィット、ベルヌーイたちがほぼ同時期に、近代的な意味での確率を思いついたようです。

1657 年にホイヘンスが出版物として最初の確率の教科書を執筆しました。その頃パスカルは初めて運任せゲーム以外の問題に、確率的推理を適用して、意志決定理論を考察しました。パスカルの神の存在に関する有名な賭けの議論として「全体」の要約が 1662 年ポール・ロワイヤルによる「論理学」の末尾に掲載されました。、同書では現在、「確率」と実際に呼ばれるものについての数的な測定が初めて書かれています。このころライプニッツが計量的確率を法律問題に適用しました。1660 年代後半には、年金は健全な保健計算に基づくようになりました。また、ロンドンの商人ジョン・グラントが死亡記録に基づく初めての広範な統計推測を行った、

ジョセフ・バトラーの「宗教の類比；1736 年」の中で「確率とはまさに生命の指導原理である」という有名な格言を作り出したのです。

ジャック・ベルヌーイ（1705 年没）の「推測法；1713 年」では、初期の確率論史上最も決定的な概念革新をもたらしました、この著作の主要な数学的功績は極めて重大で、**確率の極限定理**を初めて示したのです。

Ⅱ－2　ベイズ、ラプラス、シール、ガウス、ピアソン、ゴゼット

1763 年に Thomas Bayes の有名な論文 “Bayes の定理”が掲載された本が発刊されました。 Bayes は論理的洞察力に卓越していましたが、解析的技術に秀でていたのは、Laplace（1820）です。ラプラスはベイズの定理に逆確率の原理を取り入れました。また、互いに独立な成分を合成した量の分布に関するすべての特性値－例えば平均、分散など－が、単に各成分の分布における対応する値の和となっているという法則を発見しました。ラプラスの研究の大きな成果は、正規誤差法則の発見です。しかし、この法則は Gauss に負うものとするのがふつうです。ガウスはさらに確率の推定ばかりでなく、他の数量的なパラメーターの推定の問題も提示しました、統計的推定の問題に経験的に接近していったのです。ガウスはさらに、最小二乗法による回帰関数および重回帰関数の系統的なあてはめ法を完成させました。新しい有意性検定に特有な標本分布は、Helmert が初めて提起しました。χ^2 として知られている観測結果と仮説との食い違いの測度は、1900 年に K. Pearson によって再発見されました。

統計量の正確な標本分布の研究は、1908 年の“Student(W.S.Gosset)”の論文“The Probable Error of a Mean”に始まります。一たび問題の本質が示されると、非常に多くの標本抽出の問題が数学的に解決されました。“Student（ゴゼット)”自身は、この論文と後に提出した論文て、正確な標本分布に関する 3 つの問題の解答を出しました。それは分散の推定値の分布、平均を標準偏差の推定値で割った量の分布、および独立な変量の間の相関係数の推定値の分布に関するものです。彼の研究で標本論の“χ^2”及び“t”分布の活用が始まりました。さらに多くの有意性検定の問題が、2 つの分布と z 分布で示されることがわかりました。この研究で、一方では誤差論や数理統計学における伝統的な方法が精密化され、他方ではデータの解釈に必要な計算過程の簡単化が図られたのです。

「ベイズの定理」

ひと言でいうと、条件付き確率です。ある事（A）が起こったという条件のものとでのある事（B）の起こる確率 P（B/A)「これを PB given A と読む」のことを「A を与えた時の B の条件付き確率」という。サイコロ振りを例に説明すると、偶数の目が出た場合（2,4,6 なので、確率は 1/2）のうち、それが 4 以上である確率は 2/3 である。これは、P（B/A）＝P(A∩B)/P（A）なので、P(A∩B)＝1/3、P（A）＝1/2 なので、P（B/A）＝（1/2）÷（1/3）＝2/3 となる。

不一致統計量

我々が平均を非常に大きな標本から求めた時、当然その値は正確であると考える。実際にそのような統計量をいくつか求めて比較すると、それらの間の差は、標本の大きさが大きくなるにしたがってますます小さくなる。さらに標本の大きさが限りなく大きくなれば、その統計量は一般に母集団に特有なパラメーター（母数で定数）を含んだ関数として表される確定値に近づく。そのような統計量をこれらのパラメーターの推定に用いると、ただ1つの関数が定まる。ところが、その統計量が他のパラメーターと等しいとおくと、標本の大きさが限りなく大きくても、その統計量はこの関数と等しくはならない。その統計量はある確定値に収束はするが、その値は誤った値である。そのような統計量は不一致統計量と言われる。

一致統計量

これに反し、一致統計量はすべて、標本が大きくなるにしたがって正しい値に近づく。とにかくそれがある確定値に収束するとすれば、誤った値に収束することはない。

我々がこれから取り扱う最も簡単な場合では、一致統計量は正しい値を与えるようになるばかりでなく、与えられた大きさの標本についての誤差の分布が、正規誤差分布法則、または正規分布に近づきます。

この場合に、誤差の程度はその平方の平均値、すなわち分散と呼ばれる量で表される。我々が問題とするような場合については、大標本となるにしたがって、分散は標本の大きさに反比例して小さくなる。

ここまで一致性の概念を大標本の理論に適した表現で、標本の大きさが限りなく増大した時に要求される性質で定義した。

論理的には、この一致性の概念を次のような規定によって小標本（有限の標本）に対しても厳密に定義できることが重要です。

それには、観測頻度をその期待値で置き換えた時、その統計量を推定しようとするパラメーターに全く等しくなるものを、一致統計量と定義します。

正規分布の中心のようなあるパラメーターを推定するのに、平均または中央値など、幾つかの統計量をみつけることができる。

そしてこれらはいずれも上に定義した意味での一致統計量で、しかもその分散は大標本においては標本の大きさに反比例して小さくなる。しかし、ある定まった大きさの大標本に関して、これら種々の統計量の分散は一般に異なっている。

したがって、標本が大きくなるにつれてその誤差の分布が正規分布に近づくような統計量の中で、最も小さな分散をもつさらに小数の統計量が特に重要になる。

有効統計量

我々は一致統計量という一般的な集まりの中から、特に価値のある一群のものを分離し

て、これを有効統計量と呼んでいる。

例えば、1000 個の観測値からなる大標本から、1 つの有効統計量 A と、分散が A の 2 倍であるようなもう 1 つの一致統計量 B を求めたとする。

すると、B は必要なパラメーター（母数）に対して妥当な推定値には違いないが、その精度は A より劣っている。統計量 B を用いるならば、大きさが 1000 個の標本から求められる統計量 A を用いた場合と同じ精度の推定値を得るためには、大きさが 2000 個の標本が必要である。

この意味で統計量 B は、観測値に含まれている有用な情報の 50%しか役立たない。あるいは、その効率は 50%である。絶対的な意味での”有効“ということばは、効率が 100%の統計量に用いられる。

効率が 100%より小さい統計量も、色々な目的に応じて正しく利用することができる。

例えば、観測結果に複雑な計算を適用するよりは、観測の回数を増すことの方が容易な場合が考えられる。あるいは、当面の問題に答えるためには、ある有効でない統計量で精度は十分な場合もよく起る。

しかし、次のことが見落とされている。それは、有意性検定を正しく行なうには、無作為抽出の誤差に匹敵する誤差が入ってきてはならないことです。

これを調べれば、有意性検定又は適合度検定のあてはめに使う統計量は、一致性だけでなく 100%の効率を持たなければならないことがわかる。どんなデータを検討する場合でも、仮定しうる幾つかの前提について、それが正しいかどうかを検定できることがつねに望まれるので、有効でない統計量の使用に対する制限は非常に重大である。

充足統計量

大標本の場合には、すべての有効統計量は同等になることが示されるので、方法の違いによって不都合の生ずることはほとんどない。しかし、よく例にあがるもので、小標本の場合にこの種の統計量だけは、観測値の提供する有用な情報をすべて含んでいるということである。これは充足統計量というが、小標本の取り扱に際しこういう統計量が存在するならば、それは他のどのような有効統計量よりもすぐれている。例えば、正規分布または Poisson 系列からの標本の算術平均は充足統計量である。算術平均が理論的に重要であるのは、これら 2 つの重要な分布型に対して充足統計量となっているからである。充足統計量が存在するならば、それは最尤法によって求められる。またこの方法をさらに拡張してある特殊な関数関係を利用すれば、もともと充足統計量が存在しない場合でも、補助統計量を用いて充足統計量のもつ利益を得ること、つまり完全な推定を行うことができる。

有効統計量を計算する方法は種々あるとしても、大標本の場合にはどの方法を使うかによって別に矛盾は生じない。しかしどんな場合でも、推定しようとしている母集団のパラメーターと、その推定値として実際に用いる統計量は、はっきりと区別すべきである。また推定のために用いるいろいろな方法の中で、どの方法によってその推定値が実際に求め

られたのかを示すことはもちろん大切である。

拡大された統計学の適用範囲

統計学の実際の適用については、最初は全く異なったもののように思われた問題に対して、同一の数学的解答が次から次へと現れるという意外な事情がなかったならば、必要とされている多種多様な検定に適した方法を与えることはできなかった。

例えば、Helmert が 1875 年に与えた平均からの偏差の平方和の分布は、実は 1900 年に K. Pearson が与えた χ^2 の分布と同等のものである。この分布はまた正規母集団からの標本の分散の分布に関して 1908 年に“Student”によって独立に発見された。フィッシャーは、Poisson（ポアソン）系列からとられた小標本散布指数の分布が、これと同じ分布になっていることを発見した。

上述の 1900 年の K. Pearson の論文には重大な誤りがあった。1921 年までにこの方法で行われていた大部分の適合度検定は、そのために間違ったものとなった。しかし、有効推定値を用いてその誤りを正せば、分布の型はそのままでよく、χ^2 の表を引くときに、1 つの変数から、幾単位かを減らしさえすればよいという事実は、さらに注目すべきことである。

平均の偶然誤差の研究では、1908 年に “Student”が求めた t の分布が、彼がそこで取り扱った場合に限らず、2 つの平均の比較というもっと複雑で、さらに有用な問題にも適用できた。その上この分布は、回帰係数と呼ばれる、広汎な統計量の抽出誤差に関する正確な解にもなった。

級内相関係数、回帰係数の適合度、相関比、あるいは重相関係数などの問題に対する正確な理論的分布の研究で、z の分布と呼ぶ第 3 の分布に何度か到達した。これは Pearson や“Student”によって導入された分布と密接な関係をもち、しかも実はその当然な拡張となっている。このようにして、非常に多くの場合に必要な諸分布を、これら 3 つの主要な組に分類することができた。また、**わずかな表を作りさえすれば、数値に対する要求を満たすことも重要**であった。

Ⅱ－３　1930 年の二つの側面の検討

（研究者のための統計的方法；フィッシャーより一部抜粋）

この部分は、Sir Ronald Aylmer Fisher（1890～1962 年）の著書 Statistical Methods for Research Workers 第 13 版（1958,1963 年）の全訳版を参考にしています。

著者フィッシャーは現代の統計学の開拓者として画期的な数々の業績を打ち立てて、統計学史上に不滅の名を残しています。また農学、遺伝子学などの分野でもその名は広く知られている。

もともとはごく少数の人たちのために制作されたこの本は、長い間に次第に多くの人に広がってきたことは、その計画の中で初めは疑問視されていたに違いない新しい考えのう

ち、少なくとも幾つかは正しかったことを示している（例えば自由度の認識、有意性検定に使う関数の表を作る際に定まった確率水準を用いること、分散分析法、実験を計画する際の無作為化の必要性などである)。

一般論における定理を実際に応用するのは、数学的な証明によって定理を確立することとは別の技術である。応用に際しては、定理の意味をよく理解することが必要であって、数学的な証明を必要としない人たちにでも、定理の応用が役立つ場合は少なくない。

その後の二つの側面に関する研究は、一つは、整合的な信念の理論であり、*F・P・ラムジー*が 1930 年に初めて徹底的に考察した。これは現在「ベイズ主義」と呼ばれているが、トマス・ベイズにはほとんど関係ない。

もう一つは安定した相対頻度の理論を現実世界の予測に適用することである。その理論がサイコロのような人工的な賭博装置の範囲を超えて適用可能かどうかは、世界を変えられるかどうかに大きく依存している。R・A・フィッシャーが同じく 1930 年頃にランダム性を用いた実験計画法を教授して以来、人々は安定した相対頻度の理論を適用し続けてきた。

つまり、人々は自分たちの関わる世界の側面を、できる限りサイコロのような人工的なランダム生成器に似せるように変えている。だが、人々はこれまで、ここで終着点と思われるもの、すなわち二つの異なる推論様式とは折り合いがつかなかった。そして、今後も折り合いがつかない。

臨床医学とエビデンス・ベースト医学

臨床医学とエビデンス・ベースト医学の間の、百出するだけで進展のない討論をみてみよう。

エビデンス・ベースと医学は、過去の症例の頻度やランダム化した試験に基づくものを意味する。一方、臨床医学は、整合的な（coherent）信念の度合いの形成に基づくもの、二元性である。エビデンス・ベースト医学は勝利を収めるだろうが、それはよい帰納的推論故ではない。それは、国民健康保険の必要性と結びついた、ますます高額になる医療技術と薬学の成功ゆえである。医学の基礎を大規模な統計的規則性に置けば、各症例を臨床的に細かく診るよりもはるかに費用が安い。これは、セオドア・ポーターによって非常に的確に研究された状況、すなわち数字への信用は数学の帰納ではなく、民主政治を目指す衝動の帰納であるという状況に類似する。

確率概念がどのようにして現在のような近代的な意味で使用されるようになったのかという問題が論じられている。確率は科学だけでなく政治、経済、日常生活にもあふれている、いまや確率なしでは生活できないほどである。確率の出現は、一人の大物が達成した偉業ではなく、歴史的に起こるべくして起こったのである

確率が出現するための前提条件は臆見である。プロバビリティーは臆見の属性で、普遍的で必然的な知識とは対を成す。臆見が確からしくなるためには、権威者や権威書のお墨付きが必要で、それが証拠と考えられた。今でこそ、証拠は実験や観察で得られたものだ

が、当時は実験・観察は軽視されており、権威による証言が証拠だった。

しかし、ルネサンス期に医師が効果的な治療方法を確かめるために実験や観察を行い、現在で言う証拠を集めた。このとき、証拠の概念が変化し、プロパビリティーの概念も変化した。プロパビリティーは権威が認めるという意味であったが、それだけでなく観察で何度も真実を示すという頻度的な意味へと変化した。

このような成立過程から統計学は様々な分野の問題に対して利用されてきました。しかし、この様々な分野を統一するような数学的理論は構築されなかったのです。この統一について初めて考察したのがフィッシャーさんで、かれは統計学は集団・変動・簡約の計 3 つの研究であると述べた

Ⅲ．統計学の考え方と様々な統計量の説明

Ⅲ-1　統計学で重要な 3 つの項目（集団、変動、簡約）

さてここでは、統計学を 3 つの異なった方面から考察してみる。

1)　集団

個体の集まる研究対象は個々の実験結果ではなく、起こりうる実験結果の集団である。ここでは、平均値や標準誤差（ＳＥ）は集団の何かを知ろうとする指標である。

“Statistics”の語源からすると、統計学は国家に住む人間の集団に関する学問であったと思う。しかし、そこで繰り広ろげられていく方法は、その集団が 1 つの国家に属することとは何の関係もないし、また、人間に限らず、個体の集まり、つまり集団についての学問である。

集団の概念をある観測を限りなく繰り返すものとすれば、その集まりは測定値の集団である。この集団は誤差論の研究分野である。

統計的研究の対象となる集団は、幾つかの点で変動を示す。統計学は変動の学問であると。また、現代の統計学者の目的と昔の統計学者の目的には違いがある。

2)　変動

近代まで、大多数の統計学者の目的は、総数または平均を知ることだけであった。当時、変動は研究の目的ではなく、むしろ平均の価値を減らす厄介な事柄と考えられていた。正規標本で平均の誤差曲線は、既に 1 世紀前からよく知られていた。しかし、標準偏差の誤差曲線は 1915 年まで研究の対象となっていた。小麦の収量から人の知性まで、変動の原因を探る研究は、そこに現れる変動と測定から始めなければならない。

変動の研究から頻度分布の概念に到達する。頻度分布は種々の型があり、集団が分布する級の個数、有限か無限の場合もある。また定量的な変量では、級としての区間は有限と無限小のこともある。最も簡単なのは、出生児の性別のように級が 2 個の場合である。そ

の時、分布はそれらの級が起こる比率だけで決まらない。例えば、出生児の 51%は男で、49%は女であるというような場合である。

各夫婦から生まれる子どもの数のように、変動は不連続であるが級の個数が不確定となることもある。このときの頻度分布は子どもの人数、0,1,2・・・の各に対して記録された頻度を示すことになる。級の個数はその記録の中で最も多くの子どもを持つ家庭が入るようにすれば十分である。子どもの数のように変化しうる量を変量といっているが、その頻度分布は、変量の取りうる各値に対して、その値をとりうる頻度を示す。

身長のように、変量ではその変動範囲にある中間のどんな値もとりうる。その変量は連続的に変化し、頻度分布は変量の関数として次のように表す。

（i）集団の中で、変量がある与えられた値以下になるものの比率を示す。

（ii）この関数を微分する数学的手段て、集団の中で変量がその変動範囲のある無限小の部分に入るものの（無限小の）比率を示す。

頻度分布の考え方は、個数が有限の集団に対しても適用できる。有限の集団は、いくつかの比に分割され、連続的な変動を示すことはない。実際に起こっている原因から生じる可能性の全体を、正確に正しく比率で表されるのは、大抵の場合無限集団だけである。実際の観測結果はそのような可能性の 1 つの標本に過ぎない。無限集団に関して、頻度分布は、集団の中で幾つかの級に属するものの比率を規定する。

（i）Mendel の頻度分布のように合計が１に等しい有限個の比率からなる場合、

（ii）和が１になる無限系列で有限の大きさの比率からなる場合、

（iii）変量の変動範囲を分割した無限小の各部分について、全体に対する比率を示す数学的関数となる場合。

（iii）の場合は頻度曲線によって表現することができる。変量の値は水平軸に沿って記入し、変量の任意の変動範囲に属するものの集団全体に対する比率は、その範囲に対応する水平軸上の線分の上に立つ曲線の下の面積によって表される。

頻度曲線の概念は、連続的変量の無限集団に対してだけ用いられていることに注意すべきである。

変動の研究では、現れた変動の量の測定だけではなく、変動の型、あるいはその形態に関する定性的な問題の研究に到達した。特に重要なのは 2 つ以上の変量の変動を同時に考える場合である。この問題は、主として Galton と Pearson の研究から起こったものであるが、相関という名称で、あるいはもう少し具体的には共変動として一般に知られている。

3)　統計量と簡約

膨大なデータの簡約とは、無用な情報を除外して有用な情報を分離することです。それは母集団から無作為抽出して、いくつかのパラメーター（母数）を使って表す。しかし、実際にはパラメーターを正確に知ることはできないので、推定値を使う。その推定値は、その誤差の大きさと性質を示すことができれば価値は増大する。

多量の観測を行った際には、その結果を簡約するという切実な要求を経験する。どんな人でも、数字で表された膨大なデータの意味を（生データだけで）すべて把握することはできない。

そこで次善の策として、資料の中に含まれている有用な情報のすべてを、比較的少数の数値によって表現しようとする。この要求を、統計学はある程度まで満足させる。ある場合には、1つまたは数個の数値て有用な情報の全部をつかむことが可能である。いかなる場合でも、データがその問題の解決に適切なものであれば、研究者が考えている主要な事項を、簡単な数値の形式に簡約することが可能である。データから得られる個々の事実の数は、ふつう知ろうとする事実の数よりはるかに多い。したがって、実際のデータから得られる情報の多くは無用のものになる。この無用な情報を除外して、そのデータに含まれている有用な情報全部を分離することが、データの簡約に用いられている統計的過程の目的です。

有用な情報と無用な情報との分離は次のように行う。どんな簡単な場合でも、与えられた数値（またはその集まり）に対して、同じ条件のもとで得られた数値全体からなる仮想的な無限母集団を考える。そして手元のデータは、その無限母集団からの無作為標本であると解釈する。この母集団の分布はある種の方法で数学的に規定できる。それはいくつかのパラメーター、つまりその数式の中に現れる“定数”を含んでいる。そのパラメーターは母集団に特有のもので、この値を正確に知れば、その母集団から抽出されたどんな標本についても、そのすべてのものを知ることになる。しかし、我々はパラメーターの値を正確に知ることはできない。実際には、その値の推定値を求めることが可能なだけである。しかも、推定値は多少とも不正確なものとなる。これらの推定値が統計量と呼ばれている。もちろん、観測値から計算されるものである。もしもデータを表現するのに適当な母集団分布の数学的形式を見つけることができたとする。そういして必要なパラメーターに対して、可能な限りで最も良い推定値をデータから求めることができれば、我々はそのデータから利用できる有用な情報をすべて抽出したことになる。

データの簡約は、母集団を規定した上で行うが、その規定が適当かどうかを検定することは特に重要です。このように考えれば、データの簡約の際に起こる問題は便宜上、次の3つの型に分かれる。

（i）規定の問題、これは母集団の分布の数学的な型を選ぶときに起こる。

（ii）ある規定が得られると、推定の問題が生じる。これは、母集団における未知のパラメーターの推定に適した統計量を、標本から計算する方法を選択することを意味している。

（iii）分布の問題は、無作為標本に関するパラメーターの推定値の分布、あるいは母集団の規定が妥当かどうかの検定に用いる他の統計量の分布に関して、その正確な性質を数学的に導く問題を含んでいる。

したがって、データの集まりに対する統計的検討は、論理的には、すべての科学に共通な、**帰納法**と**演繹法**との一般的な交替関係に類似している。1つの仮説を想定してそれを必

要な限り厳密に定義し、演繹的論法によってその論理的帰結をつきとめる。
　その論理的帰結と利用できる観測結果とを比較して、それが演繹的結論と完全に合致すれば、少なくとも、その仮説に適合しない新しい観測結果が得られるまでは、その仮説は正しいことになる。

演繹法は、
　数学における定理や、数式の証明に代表されるように、前提となる定義からの必然的な**論理展開**のみによって、一般的な理論は普遍的な概念の定義がら個別的な概念や具体的な事実を導き出すよう推論を進める方法のことをいう。

帰納法は、
　実験や観察によって得られた**実証的事実**および**経験的な事実**からスタートすることによって、個別的な事例や具体的な事実の方から一般的な理論や普遍的な法則を見つけ出そうとする推論が進められる方法のことをいう。

Ⅲ－２　「把握」・「予測」・「洞察」の統計学

ビジネスと統計学のギャップ

数式が出てくると苦手意識が芽生える。
アプリを使って計算しても、出てきた結果の意味が分からない
自分の仕事に必要な範囲がわからない
　統計学は、多くの学問分野で使われています。しかし、各学問の目的や哲学、扱う研究対象の性質によって、同じ手法でも異なる使われ方をしている。限られた学問分野のみでよく用いられる手法というものも数多く存在している。したがって、経済学部と心理学部で統計学の教科書の内容が大きく違う。逆にそうした違いに踏み込まない統計学の入門書が、現実を抽象化した数式だけを扱う、無味乾燥なものとなってしまうこともある。

ビジネスに必要なのは「人間の気持ちを探る」統計学

統計学は、人間の行動の「因果関係を探る」以外に「現状の把握」、「今後の予測」の２つがあります。

「因果関係を探る」統計学はどのように役立つのか

　マーケティング部門などでは、しばしば予測よりも因果関係を探る方が重要になる。「どのようなプロモーションをすれば商品が売れるのか」「どのような商品を作ればヒットするのか」という因果関係を探る方がビジネスでは重要である。物を買うという行動には、その背後に、どのような原因が存在するのか、という因果関係を探り当てることが重要です。

これは医学や公衆衛生学でも全く同じことが言えます。どうすればその人がより長く健康に生きられるか、という原因を発見するために医学で統計学を使っているのです。

「因果関係を探る統計学に必要な３つの知識

（１）平均値や割合など統計値の意味をしっかりと理解すること

（２）データを点ではなく幅で考えること

（３）「何の値を何ごとに集計すべきか」を考えること

平均値と割合

平均値と割合は本質的にはまったく同じことです。。「量的変数」は「平均値」の形で集計します。これに対して質的変数は「割合」を集計します。量的変数は「量として大きいか小さいか」という情報を示す。質的変数は「大きい小さいということではなく、質が異なるという情報を示す。

割合と平均値は集計方法が全く違う。例えば、200 人に対する調査で 120 人が男性と言うデータが得られたとき、男性の割合が 60%という集計結果が得られたことになる。これを仮に「男性である度合い」という量的変数を考える。この「男性である度合い」調査の結果、自分が男性であると回答した人なら「１」、そうでなければ「０」とする。この平均値は 0.6 となる。これは割合 60%と全く同じ値です。

データの存在する範囲が重要

統計学は、平均値や割合を示す値の次に「おおよそデータはどこからどこまでの範囲に存在しているか」という幅を把握することができます。。

「結果」と「原因」を調べる

統計学で因果関係を探るには、何の値を何ごとに集計するかが重要です。因果関係とは、ある原因によってどのように結果が変わるのか、という関係です。単純に平均値や割合での集計を行うにしても、適切な比較軸という考え方さえ適切であれば因果関係を見ることができます。。

今までの通説に反する新しい発見を行うのに、自分の経験や勘の検証しか行わないのではだめです。データ分析から因果関係を探ること、すなわち最終的にコントロールしたい結果とそれに影響を与える原因を絞り込むのです。この最終的にコントロールしたい結果が**目的変数**（従属変数）、その結果の違いを説明するかもしれない要因を**説明変数**と呼んでいます。

医学では、今回の研究の目的は死亡率だとか、ある病気の発症率だとか、発症率につながるような説明変数（血圧だとか血液検査の値だとか）だと表現する。これも様々なデータが計測される中で最大化すべき、あるいは最小化すべきゴールです。。

ビジネスでは、データ分析を価値につなげようとすればまず、自分のデータから表現できるもののうち、「最大化したり最小化したりすべきゴールとなる項目」が何なのかを考えなければいけない。これが目的変数である。マーケティングなら売上や顧客数を、営業戦略なら成約件数やその合計金額を、調達に関わっていれば在庫破棄率や仕入れ価格、あるいは欠品による損失額などが目的変数にあたる。

逆に、広告の認知率やSNS上での口コミ件数などは目的変数ではなく**単なる出力**である。途中経過で、業種や商品によっては利益と全く関係ないとの状況もありうる。目的変数を左右する「原因の候補」である説明変数が重要になる。

「関係しているか、していないかわからない項目」ほど、あえて説明変数として分析してしまった方が新しい発見に出会える。

「中心極限定理」

多くのデータが正規分布に従うというだけでなく、仮に元のデータが正規分布に従っていなかったとしても、「そのデータの値をいくつか足し合わせたもの」はたいてい正規分布に収束する。このことは中心極限定理と呼ばれ、現代統計学の重要な基本となっている。

「データの値をいくつか足し合わせたもの」が正規分布に従うと、それをさらに「足し合わせたデータの件数」で割ったものである平均値も正規分布に収束する。収束とはデータが増えるにつれて少しずつ近づいていき、無限にデータがあれば完全に一致する、というイメージ。なぜ、このようなことが起きるのか？

この理由のヒントは、ド・モアブルが発見した、コインを何枚か投げてそのうち何枚が表になるか、という確率は、投げる枚数が多くなると正規分布に収束する、ということから考えることができる。

真の値からのズレ方が正規分布に従うのならば、真の値を推定しようとするときは最小二乗法に基づいてデータの平均値を用いることが最良である、というのがガウスの発見である。

この真の値からのズレ方はたった 1 個の原因によって起こるようなものではなく、複数の細かいズレの合計によって生じるものであり、それは正規分布に従う。データ自体のバラつき方を把握したいというのではなく、データの背後にある真の値に興味があるのであれば、平均値を使っておけばよい。

統計学を少しかじった人が混乱する

こうした「元の分布は正規分布ではないが、その平均値は正規分布に従う」という性質が、「現状把握」の統計学と「因果関係を探る」統計学の狭間で、あるいは単純に「わかっている人」と「中途半端にわかっている人」の狭間で、しばしば混乱の原因になる。

元のデータのバラつき方がどうあれ、そこから何十、何百というデータを抜き出して平均値を計算する、という行為を繰り返すと、その繰り返しの数だけ計算された平均値は中

心極限定理に基づいて正規分布に収束する。

この「**元のデータのバラつき方とその代表としての平均値**」という考え方と、「**元のデータのバラつき方とは関係ない、平均値自体のバラつき方**」という考え方を**区別する**ことは、現代統計学の中でも重要である。

混同は、現状把握なのか、因果関係の洞察なのかという目的の違いのほかに、一昔前のデータ数と現在のデータ数の違いも関係している。

いずれにしても、「顧客がどのような集団か」という現状把握ではなく、「ある取り組みによってどれほど売り上げが上がるのか」というような因果関係を洞察しようとする場合、知るべき真の値とは取り組みを行った場合と行わなかった場合の売り上げの差である。

そして、実際に得られるデータはこの真の値に対して様々なズレが加わったものとなる。顧客 1 人ひとりの多様性、というのもそのズレの原因の 1 つだが、顧客自体の売り上げのバラつき方は正規分布らしからぬものでも、その数百人以上のデータから得られた平均値は、大抵の場合正規分布に従う。

標準偏差とデータの範囲

平均値の本質が理解できたら、次に幅でデータを捉える。平均客単価が 3 千円とだけ言われても、「ほとんどの人が 3 千円前後使う」のか、「100 円しか使わない人も 1 万円程度使う人もいる」のかはわからない。これらを適切に区別するためにどのような計算をして、その結果をどのように把握すればいいのか？

データの分散の度合いを表現するから「分散」という、「分散」を感覚的にわかりやすくしたのが「標準偏差」です。標準偏差とは単に「標準的な平均値からの偏り」です。

平均値と標準偏差で現状把握ができるわけ

チェビシェフによってデータのバラつきがどのようなものであれ、「平均値－2SD（標準偏差の 2 倍)」～「平均値＋2SD」までの範囲に必ず全体の 4 分の 3 以上のデータが存在することが証明されている。正規分布に従うデータであればこの「4 分の 3 以上」というボリュームはもっと大きくなり、「平均値±2SD（正確には 1.96SD)」の範囲に 95％のデータが存在する。

標準誤差と仮説検定

統計学では「偶然のバラつきで生じたとは考えにくい差」のことを統計学的有意差あるいは単に有意差と呼ぶ。

現実には、そんなに簡単に有意差は見つけられない

パワーあるいは検出力、標準偏差 2 つ分よりは小さいが現実的な意味があり、そして統計学上有意な差を、最小限のデータからいかに見つけることができるか、すなわち検出力

を大きくできるか、というのが統計学が大事にしているポイントです。

検出力とは「何らかの差が存在しているという仮説が正しいときに、きちんと有意差であると言うことができる確率」です。

二つの過ち

統計学では「何の差もないのに差があるとしてしまう」誤りをαエラー、「本当は差が存在しているのにそれを見逃してしまう」誤りをβエラーと呼んで区別する。

有意水準

統計学の素晴らしいところは、こうした過ちの間で、いかに現実的に正しい判断を行うかが定式化されていることです。この両者の過ちは同時におこらない。百発百中で同じ現象が起こるわけではない。バラつきをもった事象に対して、両方の過ちを同時にゼロにすることはできない。

だから、統計学ではまず、αエラーを犯すリスクをどこまで許容するかを決める。慣例的には 5%、つまり 20 回に 1 回の確率で本当は間違いかもしれない仮説を主張してしまうリスクを想定する。

ただし、より厳密な意思決定が求められる場合には 1%、0.1%といった小さな水準を考えることもあるし、逆に 10%の「αエラー」を許容すると考える場合もある。この 5%なのか 1%なのかというαエラーを許容する水準のことを有意水準と呼ぶ。

検定

与えられた有意水準の範囲内で「βエラー」を最小化する、あるいは検出力を最大化するための方法を考える。

単純に分析に用いるデータを増やすほど検出力は増える。しかし、限られたデータ数でも真実をぼんやり見過ごしてしまわないように、データのバラつき方や、正しいかどうかを判断しようとしている仮説に応じて手法を使い分ける。

このように仮説が正しいと考えられるかどうかを判断するための手法のことを統計学では一般に検定（あるいは統計的仮設検定）と呼ぶ。

二つの過ちの間で、理論上の正しさと現実的な問題の間で、最善の判断は何かを考えられる学問は、統計学しかない。

だから、ありとあらゆる学問分野において理論を実証し、またありとあらゆる失敗の許されない現実的な意思決定を支えるために統計学は用いられている。

「誤差の範囲」とデータの数の関係

日常的に触れる数字に対して、「それは誤差の範囲だ」という表現をする人は多い。たとえば目的地までの移動時間に 50 分かかるのが 45 分で済むのかが「誤差の範囲」だとか、

あるプロジェクトの必要予算が 1 千万円なのか 1100 万円なのかが「誤差の範囲」」だとか。おそらくは「予測値に対して±10%前後は誤差」というざっくりとしたイメージで語られているのではないかと思う。

だが、ある程度本格的に統計学を学んでくると、軽々しく「誤差の範囲」かどうかということが言えなくなる。なぜなら統計学において、「誤差の範囲」とは主観的なイメージで語るのではなく、データの件数やデータのバラつき（つまり分散や標準偏差）をもとにして正確に計算すべきものだからです。

統計学的な意味での「誤差」とは

データの件数が誤差に影響するということを、「日本の高校生に調査した結果、自社の新製品について使ってみたいと回答した人の割合が 75%だった」という調査結果を例にすると、この結果を素直に読めば、日本全体の高校生におけるこの製品の利用意向という「真の値」は 75%、つまり価格などを度外視すれば 4 人中 3 人は新製品をほしがっているという有望な市場が広がっていることである。

しかし、この利用意向 75%の結果は、たった 4 人のうち 3 人だけが「使ってみたい」と回答した場合も、1 千人中の 750 人が「使ってみたい」と回答した場合にも全く同じように成立する。

しかし、直感的に、前者の 4 人から得られた 75%という結果は、後者の 1 千人から得られた 75%という結果よりも信頼できない、と多くの人が感じるはずである。数字上同じ 75%という値であるはずの両者の結果はどう違うのだろうか。

統計学で扱う対象は、すべてが画一的に同じ値や同じ状態を取るものではない。つまり調査対象とする人や物によって値がバラついたり、ある状態を取ったり取らなかったりする。さらに、同じ人物でも日や時間によって値や状態が変わってしまうこともある。

そして限られたデータから求められた平均値や割合は、「たまたま調査対象者に高い値のものが多かった」とか「たまたまある状態を取るものが少なかった」という可能性をはらんだものである。

したがって、今後同じ状況で同じ調査を繰り返したとしても、最終的にどのような結果が得られるかはわからないし。さらに、無限回の調査を行ったとしても、得た値が「真の値」と完全に一致するとも限らない。

ただし、だからといってまったくデタラメな値となるわけでもない。この限られたデータから求めた平均値や割合が「真の値」からどの程度ブレたものになりうるかを示す、それが統計学的な意味での誤差の記述である。

そしてこの「どの程度ブレたものになりうるか」というところでは、データの件数以外にも元のデータのバラつきの大きさが関係する。

データのバラつきが大きいほど、平均値のブレは大きくなる

―＊ブレイクタイム－

ミルクティのおいしい淹れ方

フィッシャーさんといえば、先ほどから説明しているように、現代統計学の開拓者として画期的な数々の業績を上げた研究者ですが、ミルクティのおいしい淹れ方についても統計的な検証を実施したのです。

ミルクが先か、紅茶が先か

1920 年代末のイギリスで、ある婦人がミルクティについて「紅茶を先に入れたミルクティ」か「ミルクを先に入れたミルクティ」かで味が全然違うと答えた。この実験をやったのがフィッシャーさんです。

なぜ、ランダムでなくてはならないのか？

両タイプのミルクティをランダムに飲ませ、どれほど当てられるのかを検証すればよい。これがランダム化比較実験の基本的な考え方です。ミルクティはランダムに飲まされるのだから、見えない場所でミルクティを注がれた場合に、順番を予測することは誰もできない。

「一杯の完璧な紅茶の淹れ方」

フィッシャーさんはさらに「実験計画法」の中で、婦人に実験のやり方をどの程度説明すべきか、何杯のミルクティでテストすべきか、といった詳細を検討し、また想定される婦人の回答結果と「婦人がでたらめに回答してそれだけの正答率が偶然得られる確率」を計算しました。

フィッシャーさんの考えた「科学的に実証するための手順」のうち最も重要なアイディアが、「ランダム化する」ということです。

婦人は出されたミルクティをすべて正確に言い当てました。つまり、彼女がランダムな 5 杯のミルクティを飲んでいたとすれば、偶然すべて当てる確率は 2 の 5 乗分の 1、すなわち 32 分の 1（約 3.1%）、もし 10 杯すべてを当てれば、1024 分の 1（約 0.1%）になる。これほどの確率を示されれば、彼女が何らかの形でミルクティを識別できていると考える方が自然です。

英国王立化学協会が 2003 年に発表した「1 杯の完璧な紅茶の淹れ方」について、「牛乳は紅茶の前に注がれるべきである。なぜなら牛乳蛋白の変性は、牛乳が 75℃になると生じることが確かだからである。もし牛乳がお湯の中に注がれると、それぞれの牛乳滴は牛乳としてのまとまりから外れ、確実に変性が生じる。もしお湯が冷たい牛乳に注がれるならば、このような状況ははるかに起こりにくい。」と。

Ⅲ－3　統計学の 6 つの分野

この部分は「統計学が最強の学問である」より一部抜粋しています。

統計学は数学的な理論に基づくが、それを現実に適用したときには必ずいくつかの仮定や、仮定の扱いに関する現実的な判断が必要になる。この現実的な判断は、分野ごとの哲学、目的、伝統や、扱おうとしているデータの性質によって左右される。

① 実態把握を行う社会調査法
② 原因究明のための疫学・生物統計学
③ 抽象的なものを測定する心理統計学
④ 機械的分類のためのデータマイニング
⑤ 自然言語処理のためのテキストマイニング
⑥ 演繹に関心をよせる計量経済学

ここでは、①、②、および③について説明する。

正確さを追求する社会調査のプロたち

一般に「統計をとる」という表現は、単にデータを集めるという意味で使われる。社会調査に関わる統計家の「平均値やパーセンテージ」に対するこだわりは、「ただの集計」のレベルを大きく超える。ニューディール政策のころに実用化されたサンプリング調査を発展させ、可能な限り偏りなく、求められる誤差の範囲に収まる推定値を最も効率よく得るために、彼らは研究し続けている。

得られるべきデータが測定できなかったことを「欠測」と呼ぶが、社会調査の専門家は可能な限りこの欠測を減らすため調査員を訓練する。また調査方法の改善だけでは対処できない欠測を補完し、推定値の偏りを補正するための様々な手法を考案してきた。こうした統計家の関心は、議論の土台となる正確な数値を推定することにある。

ビジネスの領域では、マーケティング調査に社会調査の専門家がしばしば携わる。

「妥当な判断」を求める疫学・生物統計家

ものや人間以外の生物を対象にする限り、**ランダム化比較実験**は比較的容易である。しかし、倫理や感情によってランダム化が許されない人間対象の領域では、疫学的な方法論を用いる。この両者に共通する考え方は、最終的に結果に与える影響の大きい「原因」を探すことである。逆に言えば、p値に基づき「原因」がちゃんと見つけられるのであれば、推定値の「全国民におけるあてはまり」という社会調査分析の統計家が重視する点についてはそれほどこだわれない傾向にある。

もちろん、仮に「若者だけに限定すると逆に喫煙でも寿命が伸びる」という、結論を覆すレベルの強力な交互作用であれば問題になる。どちらにせよ大きな影響があるなら、とりあえず喫煙率は下げた方がいいんじゃないか？という妥当な判断が下せれば、ある程度それで満足なのである。

そのため生物統計家や疫学者は「国全体からランダムサンプル」という点に関してはほ

とんどこだわりを見せない。「あくまでこの結果は医者という偏った集団のデータですが、こういう関係が見られました」と注釈つきで普通に発表する。また、「他の集団でどうかは厳密にはわかっていませんので、応用する際には注意してください」とか、「今後の課題として別の集団でも同じ関連性が見られるのか確認する必要があります」という文章が、誠実な論文には必ずといっていいほど記述されている。

こうした考え方は、疫学や生物統計学において十分な数の「全体からのサンプル」を得ようとすればとんでもないコストと手間がかかる、という現実的制約が影響している。

ランダム化比較実験

科学は「観察」と「実験」からなる

ポアンカレによると、「観察」とは対象を詳細に見たり測定したりして、そこから何らかの真実を明らかにする行為です。一方、「実験」は、様々に条件を変えたうえで対象を見たり測定したりしてそこから何らかの真実を明らかにする行為です。

ランダム化比較実験という枠組みは「実験とは何か」という考え方を一歩進めたものです。

「誤差」への３つのアプローチ

１つは、実際のデータを全く扱わず、仮説や事例をもとに理論モデルを組み立てること、2 つめは、うまくいった事例のみを結果として報告するやり方、３つめは、フィッシャーが示したランダム化を用いて因果関係を確率的に表現しようとするものです。

「実験計画法」は農場で生まれた

肥料 A/肥料 B と小麦の収穫量の関連性を科学的に分析した。収穫量は水はけ、土地の肥沃さ、日当たり、で左右されるかもしれない。だが、農地を細かい単位に分割し、ランダムに肥料をまき分ければ、平均的な条件をほぼ一致させることができる。

もし、全農地を 40 に分割し、20 地区ずつランダムに肥料 A、B をまいたとし、各地区ごとに五分五分の確率で日当たりの良し悪しが決まるとすれば、肥料 A の地区ばかりが日当たりの良い土地が集中する確率は 2 分の 1 の 20 乗、すなわち 100 万分の 1 という奇跡のような確率となる。また、両グループで日当たりの良い地区の数が全く同じになる確率は 13%、その数の差を±2 まで許容すると、その確率は 57%となる。

「誤り」と決めつけるな

統計学的な裏付けもないのにそれが絶対に正しいと決めつけることと同じくらい、統計学的な裏付けもないのにそれが絶対に誤りだと決めつけることもダメである。

ランダム化はむずかしい

ランダム化とは要するに人間の意志がそこに入り込まないようにすることである。ここで注意しなければいけないのは、人間が「無作為らしく」あるいは「テキトーに」出した数字は、しばしばそれほどランダムではなかったりする。いまなら、エクセルを立ち上げて「＝rand0」とタイプするだけで、簡単にランダムな数値が得られる。

ランダム化の 3 つの限界

世の中には、ランダム化を行うこと自体が不可能な場合、行うことが許されない場合、そして行うこと自体に問題ないが、やると明らかに大損をする場合の 3 つの壁がある。1 つ目の壁は「現実」、2 つ目は「倫理」、3 つめは「感情」と呼ぶこともできる。

「現実」の壁

「現実」の壁は、「絶対的なサンプル数の不足」と「条件が制御不能なこと」である。「1 回きりのチャンス」「数回程度のチャンス」の事象を取り扱う場合、ランダム化は不可能となる。

「倫理」の壁

統計を取り扱う人が共有する倫理のガイドラインです。①ランダム化で人為的に起こる有害な事象、②不公平なことが事前に分かっている事象です。

「感情」の壁

実験の参加自体が思う感情。

「IQ」を生み出した心理統計の専門家

IQ の理解には、心理学者の 100 年にわたる統計手法の研究を調べる必要がある。

「一般知能」の発明

現在の知能研究の基礎を生み出しスピアマンは 1904 年の論文で「イマイチの先行研究」として紹介。「そもそも知能とは何か」という問いには研究者の直感でしか答えていない。スピアマンは、先行研究のいくつか選び、その間の相関を分析した。

相関

相関とは「一方の値が大きいときに他方も大きいか、あるいは一方の値が小さいときに他方も小さいか」という関連性の強さである。ゴルトンは回帰分析を行った際に、「直線の当てはまりがよい状態」と、「平均値への回帰が大きく直線の当てはまりが悪い状態」があることを発見した。この違いを相関という言葉で表し、弟子のピアソンが相関係数という指標の計算方法を考えた。完全な直線で「一方の値が大きいときに他方も大きい」場合は

１、逆に完全な直線で「一方の値が大きいときに他方が小さい」ときはマイナス 1、関連性が全く見られない場合は０となるような指標である。

なお、相関とは「一方の値が大きいときに他方も大きい」という傾向を示しているだけで、「一方の値が大きいから他方も大きい」かどうかという因果関係とは全く別物である。

そうした研究の結果、スピアマンが発見したのは、異なる知能の側面同士がある程度相関しているという結果である。また、それぞれの指標に一定の重みをつけて足し合わせると、全ての指標とよく相関する 1 個の合成変数が作り出せるということがわかった。

彼はこの指標のことを一般知能と呼んだ。

知能を７つに分けた因子知能説

スピアマンの分析方法は、今では**因子分析**と呼ばれている。お互いに相関している複数の値から、それらすべてとよく相関する新しい合成変数を生み出す。この合成変数が因子（factor）と呼ばれ、その因子を抽出する分析だから因子分析という。因子は「知能」などの抽象的な概念を示すと考えられる値であり、これ自体を直接測定することはできない。しかしながら、因子とよく相関する「測定できるもの」は存在する。たとえば、知能であれば、反応速度、記憶力、計算力とか言ったものは測定できる。そして、実際に測定されたものすべてと「よく相関する合成関数」が作り出せるのであれば、それはおそらく知りたかった因子をよく推定しているのではないかと、スピアマンや彼の影響を受けた心理学者は考えた。

なお、因子はスピアマンが考えたように１つだけとは限らない。

1938 年にサーストンの多因子知能説がある。サーストンは様々な知能に関わるテストの結果を因子分析した結果、

① 空間や立体を知覚する空間的知能、②計算能力についての数的知能、③言葉や文章の意味を理解する言語的知能、④判断や反応の速さに繋がる知覚的知能、⑤論理的推論を行う推理的知能、⑥言葉を速く柔軟に使う流暢性知能、⑦暗記力を示す記憶知能

といった７つの知性を示す因子が抽出された。たとえば、①の空間的知能なら、算数の図形問題やパズル、立体的に配置されたブロックを数えるようなテストの結果とほとんどすべての項目とよく相関一方、文章問題や記憶にかかわる問題とはほとんど相関しないというような因子である。

近年の知能研究の中でもこの一般知能と多因子知能かという議論は繰り返されているが、多くの知能検査方法を分析すると「分野ごとではなく検査項目全体と相関する因子」すなわち、一般知能がだいたい全得点の 30%~60%ほどの影響を持つようである。ただし、この一般知能とは一体何か、という点は未だ明確な答えは出せていない。

心理統計の専門家の考え方と手法

知能に限らず、心理統計家は「心」や「精神」といった目に見えない抽象的なものを測定することを目指す。測定することができれば行動や成果や精神疾患との関連性を分析することができるが、そうでなければたとえば「仕事へのモチベーションを左右するのは金銭よりも仕事のやりがいである」といった、単純な仮説すら実証できない。

そこで自分の測定したい「抽象的な概念」が何なのかを定義する。たとえば「仕事のやりがい」を「自分の仕事について社会に対する貢献や正統な社会評価がなされているという実感」と定義すれば、それと関連しそうな質問をいくつも考えることができる。

なお、心理統計家は質問文を自分の思い付きだけで作るようなことはしない。あらかじめ「仕事にやりがいを感じている人」と「そうでない人」にインタビューをして、彼らがどのような言葉で「やりがい」のことを表現するか確認し、先行研究でどのような理論が提唱されているのかを調べたり、同様な心理学的な調査が国内外でなかったかを調べたりしてはじめて質問紙は作られる。

そしてその質問紙は、ふつう本番の前にプレテストにかけられる。微妙に表現を変えたいくつもの質問項目を、数十名程度の人間に回答してもらう。その結果、たとえばほぼ全員「Yes」と答えるだとか、無回答が多いといった、役立たずの質問項目は削除する。

次に、因子分析の結果と照らし合わせて、事前に想定していた因子の構造になるように、複数の因子と相関を持つ項目や、どの因子とも相関しなかったような項目は削除する。さらには回答者が内容を忘れたころにもう一度同じように調査し、答える度にコロコロ回答結果が変わるような質問項目は削除する。

こうして出来上がった質問紙は、科学的な測定を行うための尺度と呼ばれる。因子の構造に基づき算出方法を決めた得点は、測定しようとしていた抽象概念を表しているはずである。あとは、この得点を用いて回帰分析なり何なり、興味のある他の変数とともに分析すればよい。

なお心理統計学の中でも回帰分析はよく用いられるが、それ以外に心理統計家が好みがちな手法の 1 つにパス解析がある。心理的因子を含む変数間の関係性（とその強さ）を、楕円（別に長方形でも構わない）と矢印で示したもの。

高業績な研究者は、そのほとんどがすでに十分に仕事にやりがいを感じており、それ以上にモチベーションを高めたければ、給料や昇進という物質的な報酬を与えた方がよいようだ。

心理統計家は質問紙に命をかける

IQ への結論

ただし、日本で「一般的に用いられている知能テストは、ここで紹介したような意味深い心理学的な検討を経たものではない。

統計学を学ぶことの重要性

統計的手法を使った医学分野の成果について、2つの新聞記事を取り上げてみる。

2016.8.5 日経　AI、がん治療法助言、白血病のタイプ見抜く

膨大な医学論文を学習した人工知能（AI）が、診断が難しい60代の女性患者の白血病を10 分ほどで見抜いて、東京大医科学研究所に適切な治療法を助言、女性の回復に貢献していたことが 4 日、わかった。使われたのは米国のクイズ番組で人間のチャンピオンを破った米 IBM の「ワトソン」。東大は昨年からワトソンを使ったがん診断の研究を始めており、東條教授は「AI が患者の救命に役立ったのは国内初ではないか」と話している。他にもがん患者の診断に役立った例があるという。AI は物事を学習し、考える能力を持つコンピューターのプログラム。チェスや囲碁などで人間に勝つだけでなく、今後は医療への本格的応用が進みそうだ。

女性患者は昨年、血液がんの一種である「急性骨髄性白血病」と診断されて医科研に入院。2 種類の抗がん剤治療を半年続けたが回復が遅く、敗血症などの危険も出た。そこで、がんに関係する女性の遺伝子情報をワトソンに入力すると、急性骨髄性白血病のうち「二次性白血病」というタイプであるとの分析結果が出た。ワトソンは抗がん剤を別のものに変えるよう提案。女性は数カ月で回復して退院し、現在は通院治療を続けているという。**東大と IBM は昨年から、がん研究に関連する約 2 千万件の論文を学習させ、診断に役立てる臨床研究を行っている。**

2016.8.31 日経　受動喫煙　肺がん 1.3 倍、国立がんセンター　リスク評価、因果関係「確実」と指摘

国立がん研究センターは30 日、家庭や職場など人が集まる場所で周りが吸ったたばこの煙にさらされる受動喫煙がある人は、肺がんにかかるリスクが約 1.3 倍に高まるとする研究結果を発表した。同センターはこれまで受動喫煙が招く肺がんのリスク評価を「ほぼ確実」としてきたが、科学的裏付けがとれたとして「確実」に引き上げた。予防対策に生かす。

国立がん研究センターがん対策情報センターの若尾センター長は「日本の受動喫煙対策は世界の中で最低レベルにある。東京五輪を契機に屋内完全禁煙を実施する必要がある」と訴えた。研究グループは受動喫煙と肺がんの関連を示した 420 本の論文の中から、1984年から2013 年に発表された 9 本の論文を選び、たばこを吸わない人が受動喫煙によって肺がんになるリスクを分析した。その結果、受動喫煙のある人はない人より肺がんにかかるリスクが 1.28 倍だった。受動喫煙と肺がんの関係は 80 年代から指摘されていたが、個別の研究では科学的な根拠が無く、リスクを高めるかどうかは確実とは言い切れなかった。複数の論文をそろえて分析したところ、受動喫煙が肺がんのリスクを高めることが確実となった。研究結果を踏まえて、同センターは喫煙、飲酒、食事など 6 項目で予防法を示している「日本人のためのがん予防法」でも現行の「他人のたばこの煙をできるだけ避ける」から「煙を避ける」と修正した。受動喫煙は肺がんだけでなく、循環器疾患や呼吸器疾患

などにも影響する。厚生労働省研究班は受動喫煙が原因で死亡する人は、肺がんや脳卒中などを含めて国内で年間 1 万 5 千人に達するとの推計をまとめている。同センターの片野田統計室長は「遅きに失した感はあるが、対策を急ぐ必要がある」と話す。

Ⅳ．データの科学的な見方

21 世紀は地球規模でものを考える情報の時代。
コンピュータやインターネットは日常生活になくてはならないもの。
コンピュータやインターネットに振り回されずに情報を使いこなすには、統計学が必要。
情報に強い人は、平均値の意味をよく知っている人。
図表が正確に描ける人、それは統計の知識がある人。
基本的な統計の出し方を理解することが必要。
コンピュータのソフトや使用書は日進月歩、古いのは使えない。
統計学の体系は大きくは変わらない（フィッシャー以後)。
人間はしばしば基本的なことを忘れる。
統計学は数学とは同じではないが、共通するのは数式を使って論理・推論を展開する学問。
情報科学の発展で仮説を立証する手段“統計学的検定”はますます重要。

例えば

医療の現場では肝臓ガンの手術を受ける患者に、この手術の 5 年生存率やエタノール注入療法などの内科的治療のメリットを数字で述べ、患者に選択権を与える必要があります。その治療成績などの根拠は、統計による判断が一般には最も客観的です。

また、薬の副作用の説明でも、副作用のあるなしは確率的統計の問題であり、科学的に説明できる人は統計に強い人です。

統計学とは

ある集団の状態を数量的に把握するための方法を統計的手法といい、これらを系統的に集大成したものが統計学です。

統計的手法には

どのような表をつくるのが最も適切か、どのように貯金をするのが最も有利か、など様々な内容が含まれ、日常生活と密着した分野です。

統計学は

生活や仕事のなかでの集団を対象とし、数量によってものごとの特性や規則あるいは法則を見出そうとする学問です。

Ⅳ－1　考え方

（専門の科学には、その科学自体に根ざした独自のものの考え方や目的があります）

1.　ありのままに観察し、正確に数え、数字を通して把握する

計算を「正確」にする、ありのままに観察することが「正確」につながる、面接調査結果から集計を行う場合、実際に面接した人についてだけ結果報告する。

2.　その数字がなにを意味しているのか考える

なにを意味するのか粘り強く考えてみる。１枚の図・表が理解できると一挙に理解が進む。

3.　その数字は真実を示しているのか、偶然の要素はないかどうかをよく吟味する

偶然か否かを検証していく（推測統計学）

4.　使用されている分類や定義が適切かどうか検討し、妥当性を確認する

学会や専門書において用いられている定義・分類・用語に従うことが望ましい（再現性の保証、他調査との比較が可能）

5.　分類された数字に、一定の変化や傾向があるかどうかを考察する

表や図から一定の傾向や変化を読み取ることが重要、これは法則性の発見にもつながる方法

５つの視点に留意し、くり返しこの原点に立ち返ることが統計的なものの見方を身につける方法

待機児童数の推移

- 2001 年に厚生労働省がまとめる保育園の待機児童の算出方法が変化。「通常の交通手段を使って 20～30 分未満で通える施設」に空きがあれば、待機児童とは見なさい項目が追加。兄弟と同じ園に通うために空きを待っていた子どもなどを除外した。
- その結果、2001 年 4 月の待機児童数は 21,201 人と前年よりも 14,000 人も減った。
- この統計マジックは、「前年に始まった新エンゼルプランが掲げた待機児童ゼロに近づける狙いではないか」といわれている

基礎から分かる待機児童（2017.9.10 読売）

- 親が認可保育施設に子どもを入れたいと希望しながら入れない「待機児童」が増えている。国は保育サービスを拡充しているが、共働き家庭が増えたことや待機児童の定義を見直したことなど、3 年連続で前年を上回った。

- 待機児童は、厚労省が毎年、全国の自治体に調査し、結果をまとめて発表している。今年 4 月 1 日時点で全国に 26081 人、昨春より 2528 人増加。
- 昨年までの調査では、認可施設に入れなかった中から、自治体の判断で一定のケースを除外できた。
- ①保護者が育児休業中、②保護者が求職活動を休止、③自治体が独自に助成する認可外施設を利用、④特定の保育所のみを希望の４つにあてはまる場合。これらは、「隠れ待機児童」とも呼ばれ、4 月時点で計 69224 人。待機児童の 2 倍以上。これらのケースが除外されるのは、いずれも保育の必要性が低いとみなされるから。
- しかし、①については、「保育所に入れず、やむおえず、育休を延長している人も多いのに、一律に除外している自治体がある」という不満。そこで同省は、今春から「育休中でも復職の意志がある場合は待機児童に含める」と定義を改めた。定義の全面適用は来春から。
- 待機児童数は、これまでも定義の変更に影響を受けてきた。1990 年代の統計では、認可保育所に入れない子どもを全て待機児童としていたが、自治体の助成する認可外施設が増えてきた 2001 年、そうした施設の利用者などは除外する方針に転換。
- 2001 年；待機児童ゼロ作戦、2008 年；新待機児童ゼロ作戦、2013 年；待機児童解消加速プラン、2015 年；子ども・子育て支援新制度
- 待機児童が増えている最大の理由は、共働きが増えている事。
- 総務省調査で、25～44 歳の女性の就業率は右肩上がり、2011 年の 66.7%から、2016 年には 72.7%に上昇。

<u>GDP 統計大改革始動、14 年かけ米欧の手法に刷新（2017.4.15 日経）</u>

- モノやサービスなどの国内で生み出される付加価値を示す国内総生産（GDP)の見直しが 2017 年度に始まる。IT 産業など複雑な経済の流れを捉えきれなくなったため。14 年間の長い時間をかけて、欧米などの他の先進国のやり方にそろえていく大改革。
- GDP は現在、1 年間に部品などをどのくらい使って生産し、どの程度売れてもうけ(付加価値）がうまれたのかを表にしたものを使っている。新しい方法は、工場や店ごとに仕入れから生産・販売までの流れをより精緻に調べ、付加価値を計算する。現在のやり方では回収率が 4～5 割、分からない部分は仮定の数字をおく。新しい方法では、そこで生み出される全ての付加価値を直接、聞き取り調査する。企業の負担増も。

このように、どの統計も物事の一面でしかない。しかも、作成者の意図が隠されている場合もある。利用者は絶えず行間を読む姿勢が必要です。

＊行間を読む（文章に文字では書かれていない筆者の真意や意向を感じとる。）

＊科学的（論理的、客観的、実証的であるさま）
＊論理的（思考の形式・法則、議論や思考を進める道筋・論法に沿っていること）
＊客観的（個々の主観の恣意（勝手・きままの意）を離れて、普遍妥当性をもっているさま）
＊実証的（思考や推理によるのではなく、経験的な事実をもとにして明らかにされるさま）

Ⅳ－2　統計学の方法

統計学は、記述統計学と推測統計学に分かれる。
記述統計学では、調査や実験で得られた多量のデータをまとめる（解析する）こと、すなわち、度数分布、平均値、分散、相関係数などの指標を用いてデータをまとめることで、データの背後に潜む何らかの特徴を探り、データから最大限の情報を獲得する。

記述統計学では

仮説を立てデータ収集
データ解析（度数分布、平均値、標準偏差など）
データに潜む何らかの特徴を探る
重要なことはデータの誤りをチェックすること（データ解析では必須）

整理段階の転記ミス・パソコン入力時のミス・外れ値（極端に大きいか、極端に小さい）

推測統計学では

データ解析で得られた特徴が本当に科学的立場から受け入れることができるか否か
仮説を立てて推論する
データ解析で得られた特徴
それらに関する全ての集団（母集団）の特徴と一致するかどうか
正規分布・推定・検定

例）；30%の効用があるといわれる新薬が開発されたとする。この薬を 10 人の患者に投与したところ、まったく効き目がなかった。このとき、この薬が同じ病気の患者に 30%の効き目があるという仮説は、はたしてどの程度信頼できるのか。検定結果は 5%の危険率で棄却。30%の患者に効果ありとの前宣伝はきわめて疑わしいものと推測される。

データとは何か

一定のルールに従って測定、あるいは観察された一連の数値、または文字の集合。いつ、何のために、どこで、誰を対象として、どのように収集され、何が記載されているのか、ということがわかると、これらの数値はそれぞれ意味を生じ、データとなる。

データの種類

量的データと質的データがある。

1.　量的データ；「何らかの測定あるいは計測を行ってデータを得る」、身長、体重のように、何らかの単位をもち、数値そのものに意味があり、値の大小を比較できるデータのことを量的データと呼ぶ。

2.　質的データ；単位のないデータ、例えば性、職業、好きな色のようなものが質的データと呼ぶ。

注意

量的データには、比率尺度（比尺度）と間隔尺度がある。間隔尺度は個々の値の間に等間隔が保証されている尺度である。比率尺度は等間隔性に加えてゼロを基点とすることができる尺度である。

質的データには、順位尺度と名義尺度がある。順位尺度は順番で順位だてられるが、個々の値の間に等間隔性が保証されない尺度である。名義尺度は、その順番に意味がないものである。

データの集め方

調査と実験

データの収集方法には、２つの方法がある。実験室等で行う化学、物理あるいは生物実験と、一般地域住民を対象として行うフィールド調査がある。自分で収集した生データを１次データ、既存の資料のように加工されたデータを２次データと呼ぶ。

断面調査、前向き調査、後向き調査

１）断面調査（横断的調査）

ある任意の一時点（あるいは一期間）を設定し、その時点における現況や実態を把握しようとするもの。時間を追った情報を与えてくれるものではない。因果関係にまで議論を広げることは不可能である。

２）前向き調査

時間を追って変化を調べ、因果関係を調べようとする調査研究の代表的なものが前向き調査。観察研究、臨床試験など。前向き調査は事象の時間的関連を調べることができ、その観察も時間を追って自然な状況を追いかけていくため、因果関係を調べるのには理想的な手段である。欠点としては、場合によっては結果がでるのが数十年先になる。

３）後向き調査

結果の有無によるグループ間の比較により、原因の分析に差があるかを調べ

る。現在存在している特定の結果から、時間を遡って過去の原因を探索する。

標本抽出法

国勢調査は、全数調査（悉皆調査；しっかいちょうさ）であるが、普通の調査は仮想する対象全員（母集団）を調査するのではなく、その一部を調査し、母集団の特性を推測する。母集団から取り出された一部を標本と呼び、標本を取り出すことをサンプリングと呼ぶ。また標本を用いて行う調査を標本調査と呼ぶ。統計学は、標本調査で得られた結果から母集団の状況を調べる手法を与える。

一部から全体を推測する場合、その一部が全体の縮図となっていなければ、正しい全体を推測できないことは直感的にも了解できる。標本を選ぶ場合、あるルールに基づいて対象者を選ぶ必要がある。一般によく用いられるものに無作為抽出方法がある。文字どおり何の作為もなくという意味です。対象者が選ばれる確率が、どの人をとってみても等しくなるような抽出方法です。乱数表を利用する。

標本抽出法の追加説明

全数調査、母集団、標本、サンプリング、標本調査、無作為抽出、乱数表、単純無作為抽出法、系統抽出法（母集団の全個体に通し番号を付ける。標本の最初の個体（抽出開始番号）だけは乱数表などでランダムに選ぶ。それ以降の個体は、その数字から始めて一定間隔で順に抽出する)、多段抽出法（たとえば、まず都道府県を無作為に抽出し、次に市区町村を無作為に抽出し、その選ばれた集団単位の中から無作為に標本を選ぶ。)、層化抽出法（たとえば、総合病院で個々の診療科の中から無作為に標本を選ぶ方法)

ところで、断面調査、前向き調査、後ろ向き調査は、標本調査（母集団から取り出された一部が標本、標本を取り出す操作をサンプリングという）なのか？

断面調査は通常、標本調査となる。前向き調査、後ろ向き調査は、無作為抽出した標本からの調査ではない場合が多い。前向き調査は、特定地域住民などの全数調査であることがある。また後ろ向き調査は無作為ではなく、恣意的に選んだ対象者での調査である場合がある。そのため、後ろ向き調査では母集団を特定することが難しい場合があり、推定や検定などを行っても、妥当でないこともありうる。

Ⅴ．具体例で統計学を学ぶ

図表では何も証明できないが、それによってデータの顕著な特性が見やすくなる。図表は、データに適用すべき精密な検定の代わりにはならないが、そのような検定を示唆し、またその検定に基づく結論を説明し得る点で価値がある。

Ⅴ－1　度数分布、分割表、図

(1)　データ表示

4 ステップからなる。

a. データ・リストの作成

b. ヒストグラムの作成

c. 平均と標準偏差の計算

d. 作表・作図

b. ヒストグラムの作成

①　極端値のチェック

極端値は不良値であることが多いので、その極端値が出た原因を個別に追求し、不良値であることがわかったら、それを捨てて先に進む。

②　データの分布形のチェック

大部分、正規分布であることを前提としている。例えば「平均」の計算がそうである。正規分布していないデータから平均を求めても意味がない。そこでヒストグラムをみて、データが正規分布をしているかどうかチェックする必要がある。

(2) 度数分布

表 1　身長のデータ(単位:cm)

155.5	157.5	160.3	172.3	181.6	158.6	175.3	160.5	167.3	170.2
163.0	163.3	162.8	161.9	161.9	158.8	181.5	171.4	167.3	168.9
166.3	165.5	164.5	165.2	168.5	157.8	178.9	170.2	163.8	168.3
167.2	167.8	169.8	168.4	161.8	159.8	180.3	166.5	173.2	177.2
172.3	169.5	170.3	171.6	161.4	160.3	178.5	172.5	166.9	172.9
170.8	172.0	165.5	172.5	168.2	165.4	175.2	172.8	166.7	170.3
173.1	172.6	171.7	176.3	169.9	168.7	167.5	172.4	167.8	169.2
175.4	166.5	166.8	174.3	167.8	166.6	166.8	173.1	164.8	168.2
176.6	173.2	174.2	163.9	170.8	168.9	169.4	173.8	166.6	164.3
162.8	176.5	176.3	173.8	173.5	170.3	169.5	174.5	172.9	170.6
175.2	164.2	178.6	175.2	174.3	173.5	168.6	167.2	164.2	170.3
177.8									

表 2　基本統計量

データ数	111
最大	181.6
最小	155.5
範囲	26.1
平均	169.3

表 3　身長の度数分布表

身長階級	度数	相対度数	累積度数	累積相対度数	階級値
155.5～158.5	3	0.027	3	0.027	157
158.5～161.5	7	0.063	10	0.090	160
161.5～164.5	12	0.108	22	0.198	163
164.5～167.5	19	0.171	41	0.369	166
167.5～170.5	25	0.225	66	0.595	169
170.5～173.5	20	0.180	86	0.775	172
173.5～176.5	15	0.135	101	0.910	175
176.5～179.5	7	0.063	108	0.973	178
179.5～182.5	3	0.027	111	1.000	181
計	111	1.000	107	1.000	

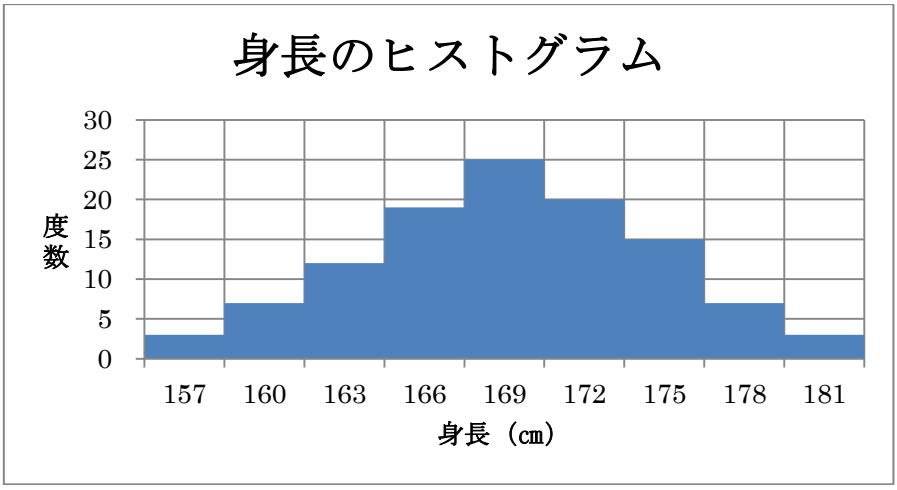


図 1　身長のヒストグラム

度数分布について、身長のデータ（表 1；111 人の男子学生身長データ、表 2；基本統計量）で作成した表 3 を用いて、説明します。

度数分布表（表 3）は、測定データをある一定の間隔（ここでは 3ｃｍ間隔）に区切り、その中にあてはまる人数を記載したものです。測定データを区切った 1 つの層を階級、区

切られた層の数を階級数、階級のまんなかの値を階級値、区切る目安とした間隔を階級の幅という。各階級の測定値の頻度（ここでは何人）を度数という。度数分布表をもとに、これを図にしたものをヒストグラム（頻度分布図）という（図 1)。

ポイント；データをどのように区切るかを検討する

↓

階級数の決定

階級数が少なすぎても、多すぎても集団の性質はわからなくなる。目安としては、10～20 くらいです。階級の幅の決定には、まず集団の最大値と最小値を見つける。上記の身長のデータの場合の最大値と最小値はそれぞれ 181.6 ㎝、155.5 ㎝です。したがって　最大と最小の差（26.1㎝）が範囲となり、これを 10～20 で割った値（この場合、10 で割ると 2.6 ｃｍ）が階級の幅となりますが、実際には階級の幅はその前後の整数 2㎝ あるいは 3㎝ を便宜上用います。この例の場合は、階級の幅を 3㎝ とすると身長の度数分布表は表 2 のようになります。階級数は 9、160㎝ 以下、180㎝ 以上が少なく、167～173㎝ に多くの人が集中していることがわかります。

度数分布を作るときの約束ごと

１）階級幅は一定

２）年齢階級が 0～4、5～9、……..　としてあるとき、階級の幅が 4 歳ではなく、5 歳になる前日までが 0～4 歳の階級に入るので、階級の幅は 5 歳です。

３）階級が 140～145、145～150、….　としてあるときは、145 は“145～150”の階級に入れる。

４）年齢階級が 0、1、2、3、4、5～9、10～14、……としてあるときは、0～4 歳までの間、年齢ごとに著しく変化するため、他の階級と同じように比較できないので、それぞれの年齢で度数を求めて、比較しようとしているのです。

累積度数　→　ある階級以下の度数を合計したもので、最後の階級では、累積度数は測定値データの合計数と等しくなる。

相対度数・累積相対度数　→　度数・累積度数を測定データの合計数で割ったものが、相対度数・累積相対度数で、それぞれ度数・累積度数の百分率に一致する。

問題 1

あるクラスの男女それぞれ 10 人を対象にヘモグロビン濃度 mg/dl を検査し、下記の結果を得ました。

男　15.5, 15.0, 14.0, 14.5, 13.5, 10.0, 16.0, 16.5, 17.0, 15.0

女　10.0, 15.0, 14.5, 12.5, 14.0, 17.5, 8.5, 10.0, 11.0, 14.5

上のデータを使って下の空欄に度数分布表を男女別に作成してください。

(3) 分割表（クロス集計表）

表 4 マスターテーブル

	酢の物を食べた(+)	酢の物を食べなかった(−)	合計
症状あり(+)	52(94.5%)【88.1%】	3(5.5%)【8.3%】	55(100%)
症状なし(−)	7(17.5%)【11.9%】	33(82.5%)【91.7%】	40(100%)
合計	59【100%】	36【100%】	95

()内は行の変数の相対度数、【】内は列の変数の相対度数。

症状(＋)55 人のうち、52 人が酢の物を食べていることから、症状と酢の物の関係は濃厚です。χ^2検定によって統計的有意差をもとめることができます。

事例;ある料亭で懐石料理を食べた人のうち、吐き気・嘔吐・下痢・腹痛を訴える患者が続出した。状況はマスターテーブルの通りである。

縦にある変数、横に別の変数をかいて、それぞれの項を分割集計（クロス集計）したものを分割表（クロス集計表あるいは単にクロス表）という。このように 2 つあるいは 2 つ以上の変数の間の関係をみる場合、分割表を作成するのが普通です。

事例；ある料亭で懐石料理を食べた人のうち、吐き気・嘔吐・下痢・腹痛を訴える患者が続出した。状況はマスターテーブル（表 4）の通りである。当日、懐石料理を食べた人 95 人中、55 人が酢の物を食べた。酢の物を食べた 55 人中、52 人が上記の症状を訴えた。酢の物を食べなかった人 36 のうち 3 人が上記の症状を訴えた。この場合、酢の物が食中毒の原因といえるか。

<u>分割表の作り方</u>

症状はすべて似ているので、症状のあり・なしを 1 つの変数、酢の物を食べた・食べなかったで別の 1 つの変数としてクロス表を作ると、表 4 のようになる。表 4 は 2×2 分割表（四分表）といい、食中毒関係の分野では、マスターテーブルとも呼ばれる。ここでは、2 つのカテゴリー（グループ）に分けたが、症状を重症、軽症、症状なしの 3 つに分けるこ

ともできる。

問題２

男性の肺がん患者 200 名と、肺がんでない健常者 200 名について聞き取り調査を行った。その結果、喫煙歴のある者は、肺がん患者では 170 名、健常者では 120 名であった。また、食事調査で、「毎日必ず朝食をとる」者は肺がん患者で 150 名、健常者で 180 名であった。喫煙歴、朝食と肺がんの関係について下の空欄にそれぞれクロス表を作成してください。

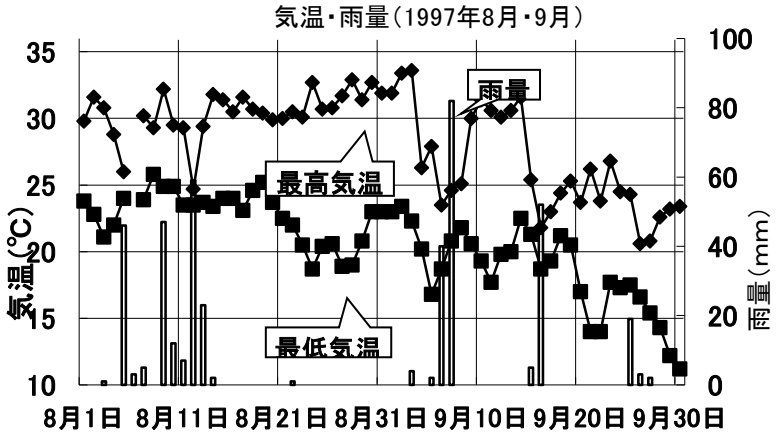


図２　折れ線グラフと棒グラフ（気温と雨量；小野湖）

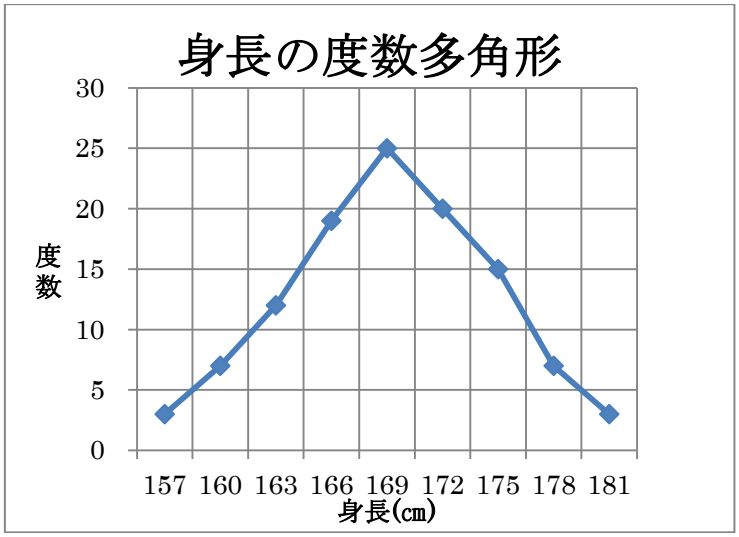


図３　身長の度数多角形

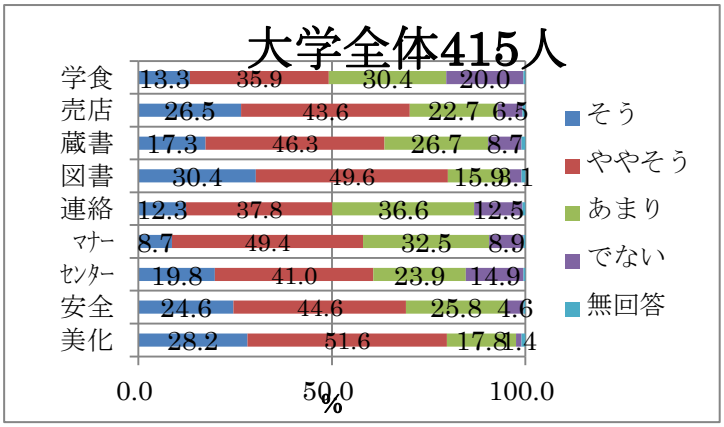


図 4　帯グラフ（学生満足と調査）

表 5　満足度調査表

	美化	安全	ｾﾝﾀｰ	ﾏﾅｰ	連絡	図書	蔵書	売店	学食
そう	28.2	24.6	19.8	8.7	12.3	30.4	17.3	26.5	13.3
ややそう	51.6	44.6	41.0	49.4	37.8	49.6	46.3	43.6	35.9
あまり	17.8	25.8	23.9	32.5	36.6	15.9	26.7	22.7	30.4
でない	1.4	4.6	14.9	8.9	12.5	3.1	8.7	6.5	20.0
無回答	1.0	0.5	0.5	0.5	0.7	1.0	1.0	0.7	0.5

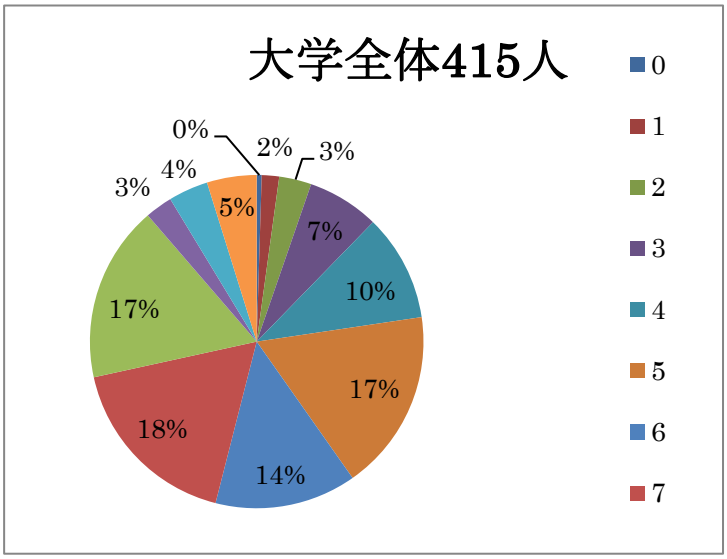


図 5　円グラフ（満足度調査）

表 6　満足度（入学）

表 6　入学したことの満足度

得点	人	%
0	2	0.5
1	7	1.7
2	13	3.1
3	29	7.0
4	43	10.4
5	73	17.6
6	57	13.7
7	73	17.6
8	71	17.1
9	11	2.7
10	16	3.9
無回答	20	4.8
	415	100
平均	5.92	

(4)　図示法

表を図に描くと、その特徴が一層明らかになる。他の人に直観的に理解させるためには図による表現がよい。

①　棒グラフと折れ線グラフ（図 2）

最も一般的な図示法である。作成に留意する点は、

（1）　縦軸、横軸の変数（単位）を明示すること

（2）　図代を明示し、出典、発表年を明記すること

（3）　表題は図では下に書くのが正しい

②　ヒストグラム（頻度分布図）（図 1）

身長、体重、年齢など連続的変数による度数分布表を図に描くときは、ヒストグラムが用いられる。図 1 より、多少の凹凸はあっても左右対称であること、すなわち、両すそ野は低く、中央が高くなっていることが容易に分かる。

③　度数折れ線（度数多角形）

これは、ヒストグラムの長方形の柱の上辺の中点を直線で結んだ折れ線グラフ。図 3 は図 1 と同様に身長のデータを図に表したものです。

④　円グラフ（扇形図表）

いくつかの項目について、相対的な大きさの比較を角度の大きさによって比較するものです。12 時の位置から大きい順に時計の針の動く方向に描く。文字の大きさは一定

の方向に書くと見やすく、また、同心円のなかには対象者・調査数などを記入する。図 5 は、満足度調査です。。

⑤　帯グラフ（帯図表）

これは、円グラフと同様に、いくつかの項目についてその相対的な大きさを比較するときに用いる。帯グラフは、いくつかを並べて変化を見るときに便利である。図 4 は満足度調査です。

⑥　レーダーチャート

各項目間で比較して、一定の傾向が一目で見られるようにしたのが、図に示したレーダーチャートによる表現法である。図 6 は本学の実施した授業評価で一般学生と長期履修学生との違いを比較したものです。

⑦　地図による表現

図 7 はポリオ（小児麻痺）の発生状況です。これを見るとサハラ以南のアフリカ諸国、インド、パキスタンなど南西アジアにポリオが分布していることがわかる。なお、現在ではワクチン接種によってポリオの発生はほとんどなくなっています。

⑧　散布図

散布図は 2 つの変数の関連をみるときによく用いられる。図 8 は萩市の明神池で観測した密度の時間変動です。水面下 0.5mと水面下 3.0mでの違いを示しています。また、図 9 は EC と Ca の関係を示した散布図です。さらに、図の中に回帰直線が描かれています。

図示するときの注意点

①　棒グラフは扱う資料が離散型の変数なので、0 と 1、1 と 2 の間をあけて描く。

②　ヒストグラムで階級の幅とグラフの目盛りは必ずその比を一定とする。

③　棒グラフで長方形を途中でカットし、上辺に値を記入してあるものは、図は直感的な大きさの相違を見るものなので、利用者が間違うおそれがある。この場合、むしろ表だけのほう　　がよい。

図 7　地図で描いた WHO 資料によるポリ関係図

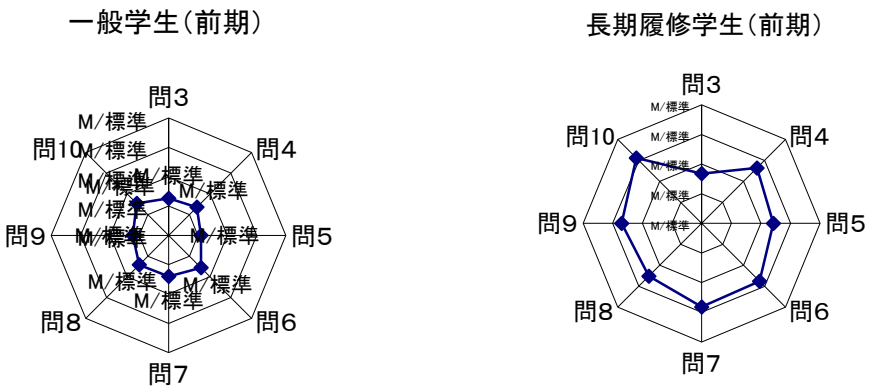


図６　レーダーチャート（一般学生と長期履修学生との授業評価の違い（平成 17 年度））

問１：学生種別、問２：性別、問３：教員の話し方（聞き取りやすさ）、問４：授業の集中度（私語への適切な対処）、問５：理解力への配慮、問６：教員の熱意、問７：理論や専門用語の説明の分かりやすさ、問８：教員の授業準備、問９：授業への興味・関心、問１０：授業の総合評価（有意義度）

各問の評価は５段階でそれぞれ「１」：否定的回答、「２」：やや否定的回答、「３」：中間的回答、「４」：やや肯定的回答、「５」：肯定的回答

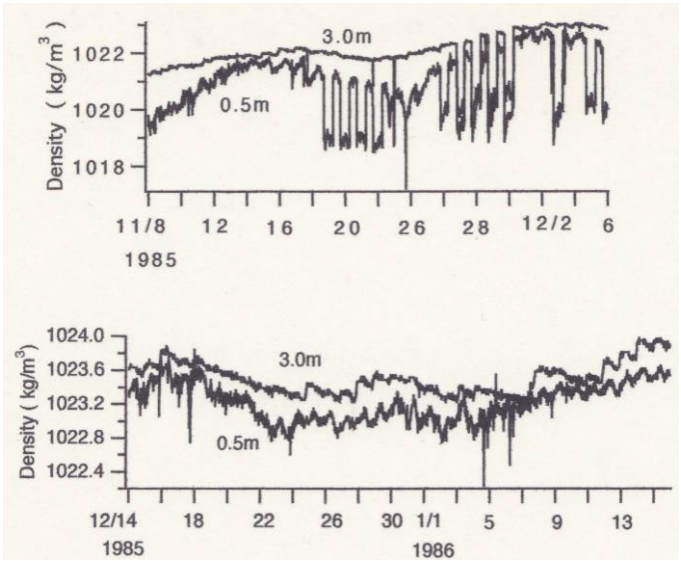


図８　密度の時間変化（明神池 1985 年～1986 年）

ECとCaの関係（1994-1997）

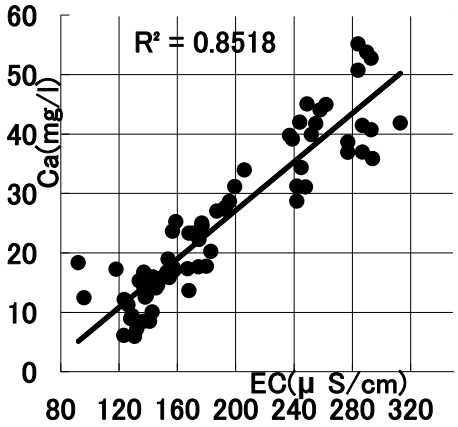


図９　散布図と回帰直線（小野湖の EC と Ca）

平均値、標準偏差、偏差値（Ｖ－2 でも説明しますが）

高校で中間試験、期末試験、実力試験、塾の試験、その成績が気になっていたと思います。成績表をみると、個々の成績と全体の平均、さらに偏差値との関係があまり明確には分からないことが多いと思います。　そこで、この機会に平均値、標準偏差、偏差値の関係を理解してください。

以下に平均値、標準偏差、分散、偏差値の一般式を表してみました。

平均値とは、データの総和をデータの個数で割った値です。式で示すと、N 個のデータ $x_1, x_2, x_3, \ldots\ldots, x_N$ の平均値 $\bar{x}$ は、

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

ここで、Σはｷﾞﾘｼｬ文字“シグマ”であり、和を意味しています。すなわち、i = 1 から N までの x_i の和を意味しています。

平均値は、加算と１回の割り算を含むだけですから、統計の尺度としては便利な値です。しかし、細かい情報は、全部捨ててしまっているのです。細かい情報を知るには、平均値の他に、度数分布とか、分散、標準偏差などを用いることが必要です。

次の式が分散 s^2 の一般式です。データのバラツキの度合いをみる尺度の１つです。

$$SD^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

この式をみると、それぞれのデータから平均値を引いてその二乗した値の総和したものをデータの個数で割った値を意味しています。このようなめんどうな計算をしなくても、それぞれのデータから平均値を引いた値をたして、それをデータの個数で割ればよさそうに思いませんか。しかし、それではマイナスの値とプラスの値が生じて、合計した計算結果がゼロになることもあるのです。それでは、バラツキの度合いをみることができません。

そこで、それぞれのデータについて平均値からの差を求めて、それを二乗することで正の値にして加算する方法をとっているのです。

しかし、分散は元のデータを二乗しているので（点）^{2}となり、もとのデータと単位が合わないのです。バラツキの尺度としては、その平方根をとり、もとのデータと単位をそろえた方が使いやすいことがわかると思います。次式が分散の平方根、すなわち標準偏差なのです。

$$\mathrm{SD} = \sqrt{\frac{1}{\mathrm{N}-1}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{x_i}-\bar{\mathrm{x}})^2}$$

さて、偏差値という言葉をよく耳にしますが、偏差値とは、ある集団における、ある個体のある種の値であるx_iを次のような式で標準化した値y_iなのです。

$$\frac{\mathrm{y_i}-50}{10} = \frac{\mathrm{x_i}-\bar{\mathrm{x}}}{\mathrm{SD}} \qquad \text{すなわち、} \quad \mathrm{y_i} = \frac{10}{\mathrm{SD}}(\mathrm{x_i}-\bar{\mathrm{x}})+50$$

なのです。この式をみると、偏差値y_iはもとの値x_iを平均が 50、標準偏差が 10 になるように標準化した値なのです。

表7 クラスの成績表（得点）

	国語	数学	理科	社会	英語	合計
山田修司	50	35	65	75	55	280
永田和彦	35	65	35	78	89	302
西村雄二	85	95	90	80	91	441
村田力	98	100	75	83	90	446
中村裕子	40	50	50	77	70	287
徳永政治	45	45	45	79	45	259
安部幸三	72	72	70	81	60	355
庄司早苗	60	70	60	85	75	350
高田勇太	65	55	80	87	80	367
平均値	61.11	65.22	63.33	80.56	72.78	343.00
分散	447.11	477.94	312.50	15.03	275.94	4603.00
標準偏差	21.15	21.86	17.68	3.88	16.61	67.85

表8 成績表(偏差値)

	国語	数学	理科	社会	英語	合計
山田修司	44.75	36.18	50.94	35.67	39.30	40.71
永田和彦	37.65	49.90	33.97	43.41	59.77	43.96
西村雄二	61.30	63.62	65.08	48.57	60.97	64.44
村田力	67.45	65.91	56.60	56.31	60.37	65.18
中村裕子	40.02	43.04	42.46	40.83	48.33	41.75
徳永政治	42.38	40.75	39.63	45.99	33.28	37.62
安部幸三	55.15	53.10	53.77	51.15	42.31	51.77
庄司早苗	49.47	52.19	48.11	61.46	51.34	51.03
高田勇太	51.84	45.32	59.43	66.62	54.35	53.54

Ⅴ－2 集団を表す代表値（平均、分散、標準偏差など）

集団を表す代表的な数値を特性値という。先ほど説明した平均値、分散、標準偏差についても再度、説明します。

1）平均

標本からの算術平均はχの文字の上に―をつけて（下記のように）「エックス・バー」と読む。

また、母集団からの平均はμ（ギリシア文字）で「ミュー」と読む

今、A,B,C 3 人の大学生の体重 65kg、80kg、73kg のとき、

$$\bar{x}=\frac{65+80+73}{3}=72.7$$

一般に n 人の体重が測定され、それぞれχ_1、χ_2、χ_3、.....、χ_nとすると、

$\bar{x}$ は

$$\bar{x}=\frac{x_1+x_2+x_3+\cdots\cdots\cdots+x_n}{n}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

Σはギリシア文字シグマの大文字で$\Sigma\chi_i$はχ_iがとりうるすべての値を加えることを意味しています。χ_iはiが1からnまでの間の任意の整数をとる場合の値をさすので、

$$\sum_{i=1}^{n}\chi_i$$

はχ_iがとりうるすべての値を加算する（$\chi_1+\chi_2+\chi_3+....+\chi_n$）とう意味となります。

<u>平均値は「真の値」のよい推定値</u>

平均値は最小二乗法に基づき、観測値に含まれるズレを最も小さくすると考えられる良い推定値です。

今、n 個のデータ$x_1, x_2 \cdots, x_n$が存在しており、真の値を t と置いたときの「真の値からのズレの二乗の合計」 f （t）とすると、

$$\mathbf{f(t)} = \sum_{i=1}^{n}(x_i - t)^2 = \sum_{i=1}^{n} x_i^2 - 2t\sum_{i=1}^{n} x_i + nt^2 = nt^2 - 2n\bar{x}t + \sum_{i=1}^{n} x_i^2$$

この f(t)を最小化する t は

$$f(t) = n(t^2 - 2\bar{x}t + \bar{x}^2 - \bar{x}^2) + \sum_{i=1}^{n} x_i^2 = n(t - \bar{x})^2 - n\bar{x}^2 + \sum_{i=1}^{n} x_i^2$$

この式から f(t)の最小化は t=$\bar{x}$のときになる。なぜなら、この式を微分すると

$f'(t)$ =2nt－2n$\bar{x}$＝2n(t－$\bar{x}$)となり、$f'(t) = 0$とすると、t=$\bar{x}$のときに$f(t)$は最小になる。

なぜ、平均値をバラつきのあるデータの背後にある「真の値」と考えてよいのか？

その答えはガウスが 1809 年に発表した「天体運行論」という論文の中で示している。それは、まず「平均値を使うことが真の値によい推定方法となる条件とは何か」から考えを始めたところにある。そして、その結果、正規分布と呼ばれる法則性を発見したのです。つまり、データのバラつきかたが正規分布に従っているのであれば、最小二乗法が最もよい推定方法であり、その結果、平均値が最もよい推定値となる。

２）分散と標準偏差（SD^2、SD）

度数分布の項で表 4 に示したように、集団の中の個々の値はすべて異なっています。しかも、集団の特性によってその大きさは異なるが、なんらかの変動（**ばらつき**）がみられます。その変動は平均からの隔たりの大きさ（偏差)、言い換えると平均の周囲に標本が密集する程度によってあらわすことができます。

ここで、「ばらつき」とは、集団の中の個々の数値が、一定の基準から離れてその周辺に不規則にちらばって存在（分布）することを意味しています。すなわち、平均値からの隔たりの大きさ「偏差」となります。この「ばらつき」の大きさを示すものとして、分散（variance)、標準偏差（standard deviation）などがあります。計測値の分布の中心（平均値）からの「偏差」を二乗して足し合わせた値を「偏差平方和」と呼びます。分散はその平均値です。標準偏差は分散の平方根。計測値の分布の中心（平均値）からの平均的なゆらぎの幅を表す指標です。

$$\text{標本分散；}\quad SD^2 = \frac{\text{偏差平方和}}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{母分散；}\quad \sigma^2 = \frac{\text{偏差平方和}}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\text{標本標準偏差；SD} = \sqrt{\text{標本分散}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\text{母標準偏差；} \quad \sigma = \sqrt{\text{母分散}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

＊ 一般には、n－1 ではなく n で除する。しかし、n で除するのはデータが母集団全部の場合なので、ここでは、n－1 を採用する。

問題 3

今、大学生 5 人の体重を 70kg、80kg、73kg、60kg、85kg とすると、その平均値、偏差平方和、標本分散および標本標準偏差を下の空欄に式で表してください。

標本集団の特性を表すのに、「(平均値) ± (標準偏差)」という形で表現すると、平均値と標準偏差が一目で見られる。

標準偏差とデータの範囲

平均値の本質が理解できたら、次は点ではなく幅でデータを捉えられるようになろう。さて、平均客単価が 3 千円とだけ言われても、「ほとんどの人が 3 千円前後使う」のか、「100 円しか使わない人も 1 万円程度使う人もいる」のかはわからない。これらを適切に区別するためにどのような計算をして、その結果をどのように把握すればいいのか、というのがここからのテーマです。

「平均値±2SD」の範囲

データの分散の度合いを表現するから「分散」という。

「分散」を感覚的にわかりやすくしたのが「標準偏差」。

標準偏差とは単に標準的な平均値からの偏です。

チェビシェフによってデータのバラつきがどのようなものであれ、「平均値－2SD（標準偏差の 2 倍)」～「平均値＋2SD」までの範囲に必ず全体の 4 分の 3 以上のデータが存在することが証明されている。正規分布に従うデータであればこの「4 分の 3 以上」というボリュームはもっと大きくなり、「平均値±2SD（正確には 1.96SD)」の範囲に 95％のデータが存在する。

異常気象と標準偏差の関係

異常気象、異常高温とは何を基準にしているのか？

- 異常気象とは、一般に過去に経験した現象から大きく外れた現象。人が一生の間にまれにしか経験しない現象です。
- 大雨や強風等の激しい数時間の気象から数ヶ月も続く干ばつ、冷夏などの気候の異常も含まれる。
- 気象庁では､過去 30 年間に観測されなかったような値を観測した場合を異常気象としている。
- 異常高温、異常多雨は世界の天候監視では、次の基準で気温と降水量の異常を判断する。
- 月平均気温の平年差が平年値統計期間（1981 年～2010 年）の標準偏差の 2 倍以上となった場合に異常高温とする。また、月降水量は平年値統計期間における最大値を上回る場合を異常多雨とする。

異常高温と標準偏差の関係

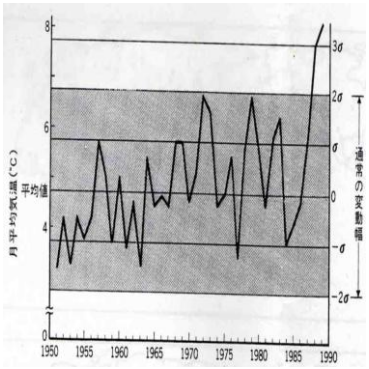


図 10　東京の１月の平均気温

Ⅴ－3　その他の代表値

（1）百分率（%）;

ある調査で A、B 両地区の結核住民検診の受診者数が

A 地区では、対象者　1,800 人のうち 150 人

B 地区では、対象者　8,850 人のうち 500 人　であった。

問題４

両者の百分率を下の空欄に式で表してください。

このように A 地区 8.3%、B 地区 5.6%となって、両者の相対的な大きさがわかります。

（2）重みづけ平均；

表 9　工場別 1 人当たりのたんぱく質摂取量

	A	B	C
たんぱく質量	260	150	200
従業員数	100	300	50

表 9 はある地域の集団での A、B、C それぞれの工場で 1 日 1 人あたりのたんぱく質摂取量です。表 8 の条件の場合、この地域全体としての 1 日 1 人あたりの摂取量を知るには、どのようにしたらよいですか？たとえば、たんぱく質摂取量について求めると、

$$\bar{x} = \frac{260+150+200}{3} = 203.3$$

しかし、被調査員がそれぞれの工場で同数の場合はこの計算でよいが、この場合には異なるので

$$\bar{x} = \frac{260\times100+150\times300+200\times50}{100+300+50} = 180$$

A、B、Cの工場の摂取量とそれぞれの被調査人員の数とが平均値に影響します。これが、この場合には正確な平均を意味します。これを重みづけ平均といいます。一般には、それぞれのグループのある項目の平均を

$\bar{x}_1$、 $\bar{x}_2$、 $\bar{x}_3$、....， $\bar{x}_m$ 、 被調査員を　$n_1,n_2,n_3...,n_m$ とすると、

$$\bar{x} = \frac{n_1\times\bar{x}_1 + n_2\times\bar{x}_2 ..... + n_m\times\bar{x}_m}{n_1 + n_2 + ... + n_m} = \frac{1}{n}\sum_{j=1}^{m} n_j\bar{x}_j$$

ここで、 $n=n_1+n_2+.....+n_m$

<u>空気の平均分子量と重みづけ平均値</u>

空気は色々な気体の混合物なので、空気の分子量は平均分子量として表しています。その平均分子量は混合気体の構成する気体の分子量で重みつき平均を出して表しています。

混合気体の構成は窒素 78%、酸素 21%、アルゴン 1 %です。また、酸素の分子量は 32、窒素の分子量は 28、アルゴンの分子量は 40 なので、空気の分子量は、次のようになります。

平均分子量＝28×0. 78＋32×0. 21＋40×0. 01＝28. 96

（3）中央値（メディアン、median、Me）；

資料を大きさの順に並べたときの中央の測定値で標本数（n）が奇数のときは、(n+1)

/2 番目の測定値、n が偶数であれば、(n/2) 番目と (n/2) +1 番目の測定値の和を 2 で割った値。

問題 5

資料が 2 ,5,6,9,12,1,3 のとき、中央値はどれですか、下の空欄に計算式を立てて求めてください。

問題 6

資料が 2,5,8,6,8,10,15,8 のとき、中央値はどれですか、下の空欄に計算式を立てて求めてください。

（4）モード（mode、最頻値）;

最も度数の多い値を代表させる。

（5）平均偏差（MAD：mean absolute deviation、MD：mean deviation）;

$$\text{平均偏差} = \frac{1}{n}\sum_{i=1}^{n}\left| x_i - \bar{x} \right|$$

（6）変動係数（CV：coefficient of variation ）

長さの単位で標準偏差が 5cm といっても、平均値が 100cm のものと 50cm のものとでは意味が異なる。このように、平均値、標準偏差がともに変化するとき、その変動を変動係数で表すことがある。

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

ただし、平均値がゼロに近いときには変動係数は使うべきではない。

（7）範囲、四分位数、パーセンタイル値（百分位数）、上限、下限、Q_1、Q_3、箱ヒゲ図

今、あるクラスの物理の点数を掲載する（表 10）。

表 10 ある学年の物理学の成績

35	78	95	55	25	78	88	45	46	30	40	65
44	70	40	60	80	90	35	44	55	61	65	70
30	38	78	72	56	44	68	78	53	30	96	78
86	81	89	72	75	74	36	62	56	47	51	61
72	36	81	94	25	10	40	30	20	10	50	70
89	56	23	41	55	70	80	90	12	15	20	60
80	70	60	50	40	30	20	10	56	46	79	36
92	81	75	34	26	38	97	50	60	19	84	74
92	87	73	18	26	38	76	85	65	61	49	56
59	57	32	45	40	60	18	79	70	60	54	30

このデータから次のように最大、最小、平均、範囲、上限、Q_3、Me（Q_2）、Q_1 および下限を求めた。範囲は　範囲＝最大値－最小値である。また、数字の小さい方から数えて全体の 10%に当たる値を 10 パーセンタイル値、全体の 90%に相当する値を 90 パーセンタイル値という。さらに、上限値は全体の 97.5%に当たる値（97.5 パーセンタイル値）、下限値は全体の 2.5%に当たる値（2.5 パーセンタイル値）となります。なお、Me は中央値で 50 パーセンタイル値です。データの度数を 4 つ分するときの値のことを「四分位数」という。小さい方から順に、第 1 四分位数、第 2 四分位数（中央値でもある）、第 3 四分位数となります（表 11）。

表 11　30 代、40 代、50 代、60 代の中性脂肪データのパーセンタイル値を示す。

（単位；mg/dl）

	30 代	40 代	50 代	60 代
上限	400	450	350	280
Q_3	180	300	280	250
Me	100	120	130	115
Q_1	35	40	50	45
下限	20	30	40	35

	30 代	40 代	50 代	60 代
上限－Q3	220	150	70	30
Q_3	180	300	280	250
Me	100	120	130	115
Q_1	35	40	50	45
Q_1－下限	15	10	10	10

	30 代	40 代	50 代	60 代
Q_3	180	300	280	250
Me	100	120	130	115
Q_1	35	40	50	45

最大	97
最小	10
平均	55.8
範囲	87

上限	94.0
Q_3	75.0
Me	56.0
Q_1	38.0
下限	12.0

表 11 は、年代別中性脂肪データからパーセンタイル値を表したものです。
この表から箱ヒゲ図を描いてみました。

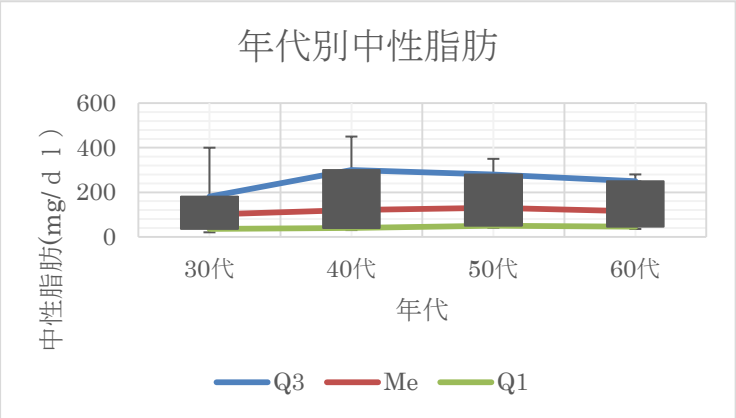


図 11　年代別中性脂肪の箱ヒゲ図

問題 7

2mの規格で作られた鉄棒の長さの平均 200cm、標準偏差 4.0cm、6 個のボールベアリングの長さの平均 1.5cm、標準偏差 0.02cm のとき、下記の空欄にそれぞれの変動係数を求めて、どちらがばらつきが小さいか判断してください。

平均値

平均とは、複数個のデータを代表する一つの統計量である（データの分布が正規分布とみなせることが前提）。統計用語では、「代表値」の一種である。

例えば、平均が 50 であるとすれば 50 がこのデータ集団の代表である。実際に 50 という値のデータがなくても、この付近にデータが密集し、50 から±方向に離れるにつれて分布

がまばらになってゆく。

平均を用いてはならない場合としては、データの尺度が間隔・比率尺度でない場合、データの中に極端値が存在する場合、データの分布が正規分布とみなせない場合である。このときには、名義尺度データあるいは順位尺度データの処理へ進む。

偏差平方和、標準偏差、標準誤差

数値データのばらつきの程度を調べるのに使われる。

偏差平方和（残差平方和ともいう）

偏差を平方した合計である。偏差とは、平均からの隔たりであり、測定値－平均値である。したがって、

偏差＝測定値－平均値、偏差平方＝偏差×偏差、偏差平方和＝Σ（偏差平方）となる。

標準偏差 SD（standard deviation ）

標準偏差とは、データが平均からどの程度ズレているかを表す統計量である。すなわち、標準偏差の値は、データ 1 個分の標準的なズレ幅を示す。統計用語では「散布度」の一種。もし、データの分布が正規分布であれば「平均±標準偏差」の範囲にデータ全体の約 68%がおさまる。例えば、100 個のデータがあり、平均が 50、標準偏差が 10 であるとすると、50±10（つまり、40～60）の範囲にほぼ 68 個のデータがおさまる。

標準偏差の計算式としては、SD＝√｛(偏差平方和）/（N－1)｝＝不偏推定値（unbiased estimate）または、SD＝√（偏差平方和/N）データ処理上は、どちらか一方を一貫して用いるなら支障はない。なお、SD^2 は分散（varinace)。

標準誤差　SE

標準誤差は、標準偏差（不偏推定値の方）を√N で割ったもの。

標準誤差（SE）＝標準偏差の不偏推定値/√N＝√｛偏差平方和/N（N－1)｝

数値データのばらつきの程度を表す場合、1 群のみのときは、(平均値±標準偏差)、2 群以上で平均値の比較を行うときには、(平均値±標準誤差）である。

平均と標準偏差の意味

本来、平均とは真の値のことである。すなわち、データにまったく誤差が加わらないときには、すべてのデータが集中する一点の値を表している。したがって、真の値が一点に定まらないときの平均は意味がない。正規分布は真の値が一点に定まることの証である。また、そのときデータの値は、「データ＝真の値±偶然誤差」として定義することができる。この真の値の推定が平均値、偶然誤差の推定は SD であり、ある観察場面における標準的データは、(平均値±SD）という値をとることを意味している。

平均の推定値、標準偏差の推定値

平均μに対する最良の推定値は、$\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i$　で与えられる$\bar{x}$であり、標準偏差σの最良の推定値としては、$SD^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$　からSDを計算すればよい。これら２つの統計量は、観測値に関する最初の２個の累乗和から計算され、次の点に関して正規分布と特別の関連をもっている。すなわち、母集団分布が正規型とすれば、その分布に関して標本が提供する情報は、全部この２つの統計量に要約されている。しかし、分布が正規型と著しく異なっている時には、この２つの統計量は、ほとんどあるいは全く役に立たないことがある。

平均の分布での標準偏差は標準誤差となる！

平均値に関する統計的な処理の基礎になるのは次の基本的命題である。ある量が分散σ^2の正規分布に従うならば、その量に関する大きさがｎの無作為標本の平均は、分散がσ^2/nの正規分布に従う。もとの分布が正確には正規分布でないときでも、平均の分布は標本の大きさが増すにつれて一般に正規型に近づくという事実があるので、この命題の効用は幾分か増大する。したがって、もとの分布が正規型であるという十分な確証はなくても、平均の分布が正規型に近づかないような例外的な分布ではないと考えられる根拠があれば、この方法を広く適用してもさしつかえない。

このことから、もしも母集団の分散がわかっていれば、与えられた大きさの無作為標本の平均の分散を求めることができて、それによってある定められた値と標本平均との差が有意であるかどうかを検定することができる。その差が標準誤差より何倍も大きければ、それは確かに有意である。標準誤差の２倍をもって有意性の限界にとるのが慣例である。

・統計学は平均のことを「mean」、しかし、エクセルでは「Average」を使っている
　分散＝平均平方和
・標準偏差は通常、測定誤差を表すと考えられている。そのため、測定回数を増やしても、測定値の標準偏差は変化しない特徴がある。
・測定機器の検出限界などを決める時には、標準濃度が 0 のブランクの測定値の「平均値±3SD」を基準にすることが多く、3σと呼ばれる（約 99.7%）。
・標準誤差 standard error SE＝標準偏差をデータ数ｎの平方根で除した値 $SE=\sigma/\sqrt{n}$
標準誤差は通常、母平均値を推定するときの推定誤差と考えられる。そのため、測定値の標準偏差が同じでも、データ数（測定回数）が 4 倍になると、SE は半分になる。つまり、データ数の増加により情報量が増えるので、推定精度が上がったと考えることができる。

Ｖ－4　正規分布

1）正規分布とは

生物現象など自然界で観察される多くの計測値は、何であれ平均値に近いほどその出現率が高く、平均値からその両側に値が遠ざかるにしたがって出現頻度は少なくなる。

このうち、同じものを何度も繰り返し計測し、平均値からのずれ（誤差）の大きさを求め、その出現度数を描いてみると、平均値を中心として左右対称の釣鐘状の分布型になることが多い。

1812 年に数学者ガウスは、この純粋な条件で繰り返し計測したときに、一貫して現れる分布型を発見し、それを正規分布（normal distribution）と名付けた。発見者の名前をとってガウス分布（Gaussian distribution）ともいわれる。

この曲線は誤差曲線とも呼ばれます。その理由は、ある寸法を目標にして何かを作るとき、ちょっとした手のはずみなどで、目標の寸法よりもわずかばかり大きくなってしまったり、逆に小さくなってしまったりする‘誤差’が生じます。この誤差の大きさは、正規分布に従うことが知られているからです。つまり、物を作るときには、人によって、あるいは機械によって、目標より平均して大きめの物を作ったり、あるいは小さめの物を作ったりする‘くせ’があります。この誤差の平均値は 0 でないのが普通ですが、誤差の大きさは、その平均値を中心にして左右対称な正規分布にしたがいます。

正規分布は、その分布に従うあるグループの平均値と標準偏差が分かっていれば、その分布に関する全てが分かります。例えば、

正規分布曲線下の面積は、$-\infty$～$+\infty$　で　1

$\mu-\sigma$ と $\mu+\sigma$ の間の正規分布曲線の面積は全体の約 68％

$\mu-2\sigma$ と $\mu+2\sigma$ の間の正規分布曲線の面積は全体の約 95％

$\mu-3\sigma$ と $\mu+3\sigma$ の間の正規分布曲線の面積は全体の約 99.7％

$\mu-4\sigma$ と $\mu+4\sigma$ の間の正規分布曲線の面積は全体の約 99.99％

という具合となります。「平均値が μ、標準偏差が σ である正規分布」を $N(\mu、\sigma^2)$ と略して記号で表す習慣があります。N は normal distribution の頭文字です。

例）東京オリンピックで優勝した日本女子バレーボールチームの平均身長は 171cm、この当時の女子の平均身長（μ）を 156cm、標準偏差（σ）を 5cm と仮定すると、図 12 から、171cm 以上は図の $\mu+3\sigma$ 以上にあたる。すなわち約 0.15％である。わが国の昭和 20 年代に約 200 万人の出生数があるのですが、女子がその約半分とすると 100 万人です。したがって、171cm 以上は 1,500 人しかいないことになります。その中から運動神経もある程度発達していて、訓練に耐えることができ、なおかつバレーボールが好きな人をさがすのは、大変であったことがわかります（統計学、系統看護学講座より抜粋）。

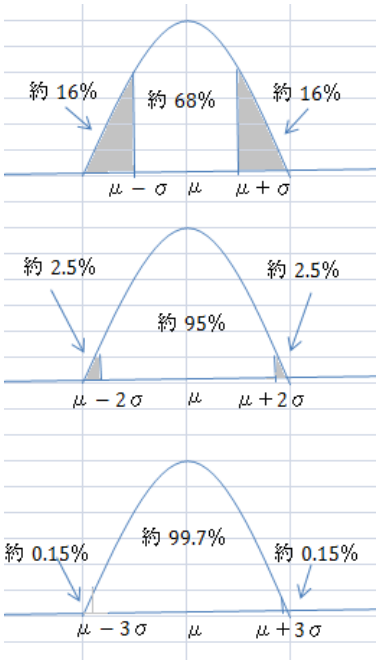


図 12　正規分布曲線と標準偏差

2）正規分布の標準化

先程の例で平均値よりも大きく、175 cmよりも小さい人は何%でしょう。これに答えるには、日本の当時の女子の身長について、細かい数表を作っておく必要があります。また、平均値や標準偏差は、各測定値によって単位が異なっています。例えば、身長は cm、体重は kg、ヘモグロビン濃度は mg/dl などです。これらについて、片っ端から数表を作ることが必要になります。しかし、実際にそのような作業をすることは現実的ではありません。そこで、正規分布に従うものならどんなものにでも適用できる数表があれば便利ですよね。それでは、そのような便利な数表を作るにはどうしたらよいのでしょうか。

それには「正規分布は平均値と標準偏差によって決まる」という性質を利用するのです。つまり、平均値を固定して、標準偏差をものさしにしてばらつきの大きさを表してやれば、数表は 1 つで済むことになります。

次の図を見てください。

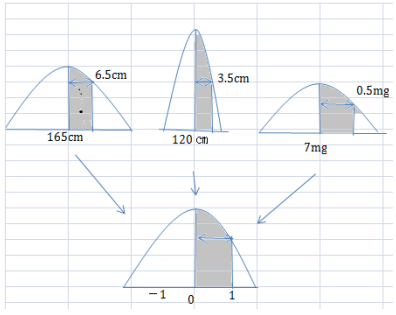


図 13　正規分布の標準化

(a) は青年男子の身長で平均値が 165 ㎝、標準偏差は 6 ㎝、(b) は小学 3 年生の身長で平均値は 124 ㎝、標準偏差は 4 ㎝、(c) は 1 錠の胃腸薬に含まれるパントテン酸カルシウムの量で平均値は 5 ㎎、標準偏差は 0.4 ㎎です。この 3 つの分布は、どれも正規分布なのですが、平均値も違うし標準偏差も異なります。単位も異なります。しかし、ともに正規分布である共通点を利用して、1 つの数表が使えるように工夫できそうです。正規分布であるという共通の点は、平均値の両側に標準偏差だけの幅をとると、その幅の中の面積（図で 2 重斜線の部分）は 0.6826 です。というように、平均値から標準偏差を単位としてある幅をとると、その範囲の面積がどんな正規分布の場合にも等しい値になるのです。そこで、(d) のような正規分布を考えます。この正規分布は平均が 0、標準偏差が 1 です。つまり N（0、1^2）です。この正規分布と他の 3 つの正規分布と比べてみると、例えば、(a) では 165 ㎝のところを 0 とみなし、横軸の目盛を 6 ㎝を単位（1 とする）にして書き直すと、(d) と全く同じになります。つまり、165 ㎝のところが 0 になり、171 ㎝のところが 1 になり、177 ㎝のところが 2 に、159 ㎝のところが－1 になるわけです。

したがって、(d) の正規分布 N（0、1^2）についての詳しい数表があれば、(a) の図形のどの部分の面積もわかることになります。(b) の場合も (c) の場合も全く同じことです。

このように平均値が 0、標準偏差が 1 になるように統計量を考えて、各測定値が平均 0、標準偏差 1 になるような正規分布を作成すると、各測定値の分布上の位置が、比較できて便利です。

問題 9

図 13 で左上の分布で 180 ㎝は下の分布に置き換えると、どのような値になるのですか、答えてください。

このような平均 0、標準偏差が 1 の正規分布を規準正規分布あるいは、標準正規分布といいます。

z を使って、$z = \frac{x-\mu}{\sigma}$

このようにすると、各測定値の単位に関係なく分布上の位置を示したり、検定に利用できます。

z 分布表を使った例）

今、μ ＝156㎝、σ ＝5㎝ の身長分布で 169㎝ 以上の人の割合を求めたいとします。この場合まず、169㎝ に対応する z を求めます。

$z = (169-156)/5 = 2.6$

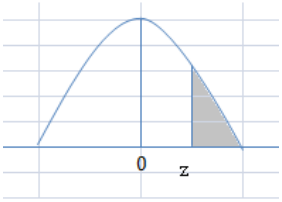
付表1の正規分布表より、$z = 2.6$ より右側の面積は0.0047、したがって、割合は0.47%となります。また、同じμ、σの集団で150cm以下の割合は、

$z = (150-156)/5 = -1.2$

付表1より0.1151となります。したがって、割合は11.51%となります。

＊ 注意；正規分布は左右対称なので、zが負でも面積は± 1.2と同じことになります。

付表１　正規分布表



ｚから外側（色の暗い部分）の面積です。ｚが負の場合も、対称性から同面積です。

z	0.00	0.01	0.02	・・・・	0.08	0.09
0.0	0.5000	0.4960	0.4920		0.4681	0.4641
0.1	0.4602	0.4563	0.4522		0.4286	0.4247
0.2	0.4207	0.4168	0.4129		0.3897	0.3859
0.3	0.3821	0.3783	0.3745		0.3520	0.3483
0.4	0.3446	0.3409	0.3372		0.3156	0.3121
0.5	0.3085	0.3050	0.3015		0.2810	0.2776
0.6	0.2743	0.2709	0.2676		0.2483	0.2451
0.7	0.2420	0.2389	0.2358		0.2177	0.2148
0.8	0.2119	0.2090	0.2061		0.1894	0.1867
0.9	0.1841	0.1884	0.1788		0.1635	0.1611
1.0	0.1587	0.1562	0.1539		0.1401	0.1379
1.1	0.1357	0.1335	0.1314		0.1190	0.1170
1.2	0.1151	0.1131	0.1112		0.1002	0.0985
1.3	0.0968	0.0951	0.0934		0.0838	0.0823
1.4	0.0808	0.0793	0.0778		0.0694	0.0681
1.5	0.0668	0.0655	0.0643		0.0571	0.0559
1.6	0.0548	0.0537	0.0526		0.0465	0.0455
1.7	0.0446	0.0436	0.0427		0.0375	0.0367
1.8	0.0359	0.0351	0.0344		0.0301	0.0294
1.9	0.0287	0.0281	0.0274		0.0239	0.0233
2.0	0.0228	0.0222	0.0217		0.0188	0.0183
2.1	0.0179	0.0174	0.0170		0.0149	0.0143
2.2	0.0139	0.0136	0.0132		0.0113	0.0110
2.3	0.0107	0.0104	0.0102		0.0087	0.0084
2.4	0.0082	0.0080	0.0078		0.0066	0.0064
2.5	0.0062	0.0060	0.0059		0.0049	0.0048
2.6	0.0047	0.0042	0.0044		0.0037	0.0036
2.7	0.0035	0.0034	0.0033		0.0027	0.0026
2.8	0.0026	0.0025	0.0024		0.0020	0.0019
2.9	0.0019	0.0018	0.0018		0.0014	0.0014
3.0	0.0013	0.0013	0.0013		0.0010	0.0016
3.1	0.0010	0.0009	0.0009		0.0007	0.0007
3.2	0.0007	0.0007	0.0006		0.0005	0.0005
3.3	0.0005	0.0005	0.0005		0.0004	0.0003
3.4	0.0003	0.0003	0.0003		0.0003	0.0002
3.5	0.0002	0.0002	0.0002		0.0002	0.0002

3）偏差値から正規化を考えてみる！

偏差値とは、ある集団におけるある個体のある種の測定値 x_i を次のような式で標準化した値 y_i です。$\frac{y_i-50}{10}=\frac{x_i-\bar{x}}{SD}$　すなわち、$y_i=\frac{10}{SD}(x_i-\bar{x})+50$

ここで、SD は標準偏差。偏差値 y_i はもとの測定値 x_i を平均 50、標準偏差 10 になるように標準化したものです。これから考えると基準正規分布は、

$$\frac{z-0}{1}=\frac{x_i-\bar{x}}{SD}\rightarrow\quad\frac{x-\mu}{\sigma}$$

とおけるのです。つまり、みなさんはすでに、正規分の基準化よりも難しい計算を使っているのです。

4）標本平均の分布（中心極限定理）

身長が正規分布に従うことは以前に述べたとおりですが、その身長の集団から n 人を選び出し（抽出)、その標本平均を求めます。この作業を繰り返し行うと、標本平均の集団が出来上がりますが、その集団はどんな集団になるのでしょうか。この答えは重要な定理となっているので、以下に紹介します。

平均値 μ、分散 σ^2 の任意の分布型（どんな分布でもよい）をした母集団から、大きさ n の標本 χ_1、χ_2、χ_3、……、χ_n を選んだとき、標本平均

$$\bar{x}=\frac{x_1+x_2+x_3+\cdots\cdots\cdots+x_n}{n}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

の分布は、n が大きくなると正規分布 N（μ、σ^2/n）に近づく。

この定理は、後述の区間推定や検定で使うので重要です。

なぜ、元のデータが平均値付近になくても、それらを足し合わせた値の集まりは中心（平均値）付近に集まり、そこから左右対称な分布になるのか？

これは、ド・モアブルが見つけた「コインを何枚か投げてそのうち何枚が表になるか」という確率は、投げる枚数が多くなると正規分布に近づくという事実を考えれば理解できる。

問題 10

この正規分布の標準偏差（標準誤差）について答えてください。

ここで、もう一度、平均値、偏差、偏差平方和、分散、標準偏差および標準誤差の関係を整理しておいてください。

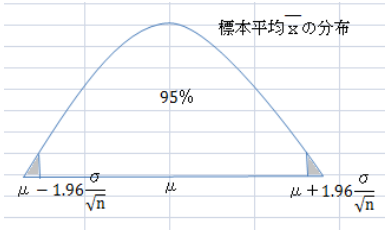


図 14　標本平均の分布

なお、標本平均の分布は n が 25 より大きいときはよく正規分布に近似し、x がどんなにゆがんでいても、n が 50 より大きいと正規分布によく近似することが知られています。

Ⅴ－5　推定と検定

1）推定とは

例えば、インフルエンザにかかった子どもが脳症をおこして死亡する確率（何%とか）は、どれくらいか、知りたいことがあります。その確率はすべての子どもをインフルエンザに罹患させなければわからない。これは、実際には不可能です。

そこで、この場合には全体の一部のみから全体を知ろうとする。これが推定です。

2）母集団と標本

知りたいことは確率だけとは限らない。平均値を知りたいこともある。これらの知りたい値を

母数 parameter といいます。母数を知ろうとする対象が母集団 population、そこから抽出する一部を標本 sample と呼びます。

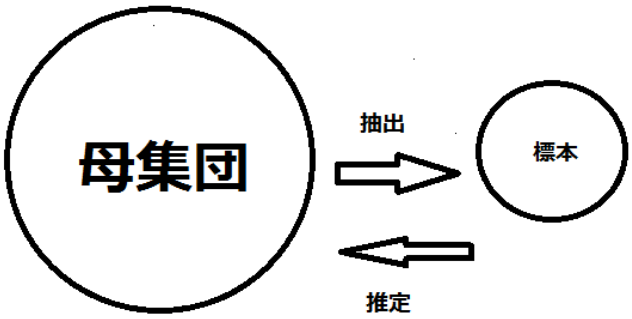


図 15　母集団と標本

3）平均値 μ の区間推定

中心極限定理より、大きさ n の標本を無作為抽出して得られる$\bar{x}$の分布は、平均 μ、標準偏差$\frac{\sigma}{\sqrt{n}}$の正規分布に近似することがわかっている。このことを利用すると、$\bar{x}$が $\mu - 1.96\frac{\sigma}{\sqrt{n}}$

と $\mu + 1.96\frac{\sigma}{\sqrt{n}}$ との間に入る確率は 95%となる。これを式に表すと

$P\left\{\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$ になります。この式を変形して $\bar{x}$ と μ を入れ替える

と、$P\left\{\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$ となります。ここで、σ を SD で置き換えると、

$\bar{x} - 1.96\frac{SD}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{SD}{\sqrt{n}}$、この式が μ の 95%信頼区間となります。

＊正規分布の項では、$\mu \pm 2\sigma$ で約 95%としていましたが、ここではより 95%に近づけるために、 2 →1.96 を用います。

<u>この式の意味</u>

標本平均を求めるため標本を抽出する作業を何回も繰り返して、そのたびごとに信頼区間をつくると、100 回のうち 95 回はその区間に母平均（μ）が含まれているということです。けっして、1 回だけ求めた信頼区間に母平均（μ）が入る確率が 95%という意味ではありません。

<u>問題 11</u>

上記で、標準偏差が SD/√n になっているのはなぜですか、下記の空欄を使って答えてください。

4） t 分布（スチューデント分布ともいう）による区間推定

先ほど、μ の区間推定で標準偏差σのかわりに標本推定値 SD を使用しました。これは、標本が大きいからです。標本の大きさ n が 25 より大きいときは SD を使用して差し支えありません。

しかし、標本の大きさが 5 とか 10 のときには、σを SD で置き換えることはできないので、t 分布を用います。t 分布は平均が 0 の正規分布に似た左右対称の分布で、自由度（df＝n－1）によって曲線が異なります。

$t = \frac{\bar{x} - \mu}{\frac{SD}{\sqrt{n}}}$　自由度は独立な変数の個数でここでは調査対象を n としたとき n－1 となります。df＝5 のときに、t 分布で左右の面積が合計で 5%となる t 値は 2.571 です。n が大きくなるにつれて t 分布の曲線は次第に正規分布に近づき、t ＝2.571 は基準正規分布で左右の面積が合計で 5%ととなる z の値 1.96 に近づく。

df＝5 のとき、μ の 95%信頼区間は　$\overline{x} - 2.571\frac{SD}{\sqrt{n}} < \mu < \overline{x} + 2.571\frac{SD}{\sqrt{n}}$　で一般には

$\overline{x} - t_0 \frac{SD}{\sqrt{n}} < \mu < \overline{x} + t_0 \frac{SD}{\sqrt{n}}$ となります。ここで t_0 は自由度 df と目的とする信頼区間によって決まるもので、t 分布表から求めることができます。例えば、

df＝5 で信頼区間が 95%のときには t_0＝2.571
df＝10 で信頼区間が 95%のときには t_0＝2.228
df＝5 で信頼区間が 99%のときには t_0＝4.032
df＝10 で信頼区間が 99%のときには t_0＝3.168

問題 12

今、10 人の男子学生の身長が、166、167、170、175、173、169、177、171、175、178（cm）とすると、標本平均は 172.1cm、標本標準偏差は 4.15 cmになります。このときに、下記の空欄を利用して t 分布を使ってμに対する 95%の信頼区間を求めください。

＊ここまでの説明で、正規分布でμ＋2σがμ＋1.96σとなること、さらにμ＋1.96SD/√n→μ＋tSD/√nの展開を理解してください。

自由度の説明

正規分布から取り出された 2 つのデータがあるとします。データが 2 つあれば、t 分布を利用して母平均μの区間推定ができます。区間推定をするには、まず 2 つのデータから標本平均を求め、それを使って標本標準偏差 SD を計算するのが第一段階です。しかし、よく考えてみると、SD を計算するに際して、本当の平均値μがわからないので、データから架空の平均値を作り出して、その平均値で後の計算をやっています。本来なら、標本平均が母平均と同じになるように標本を選ばなければなりません。ということは、2 つの標本を選ぶ場合には、2 つめの標本は 1 つめとの平均がちょうど母平均と同じになるように選ばなければならないので、勝手に選ぶことができません。自動的に決まってしまうということです。

すなわち、2 つの標本を選ぶ場合には、自由が許されるのは最初の 1 つだけで、２つめは自由が許されないのが本来の姿ということになります。そういう意味で、n が 2 のときには、自由度は 1 になります。n が 3 以上のときも同じことです。したがって自由度は df＝n－1 になるのです。

自由度という考え方は「データの数から、そのデータで作り出して使用した平均値の数を差し引いたもの」と考えておけば間違いありません。

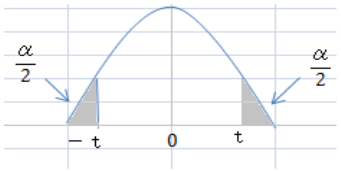
問題 13

18 歳男子 100 人の座高の平均値は 88.8 cm、標準偏差は 3.50 cmであった。μに対する 95%信頼区間を求めてください。

問題 14

上記でμに対する 99%信頼区間を求めてください。

付表２　ｔ分布表



ｔ、－ｔの外側（色の暗い部分）の面積を加えてαとなるが、そのときの自由度dfとｔの値です。

df \ α	0.10	0.05	0.01
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
15	1.753	2.131	2.947
20	1.725	2.086	2.845
25	1.708	2.060	2.787
30	1.697	2.042	2.750
40	1.684	2.021	2.704
50	1.676	2.009	2.678
60	1.671	2.000	2.660
70	1.667	1.994	2.648
80	1.664	1.990	2.639
90	1.662	1.987	2.632
100	1.660	1.984	2.626
120	1.658	1.980	2.617
∞	1.645	1.960	2.576

5）検定

検定とは

「差がある」という仮説を検定する場合、差の程度が不明なため、そのままでは検定できない。そこで考え出されたのが、その逆の「差がない」という仮説を検定して、それに何らかの矛盾が見つかれば、もとの「差がある」という仮説を採用する。逆に、明らかな矛盾がないときにはその判定を保留する。このような論法で検定を行います。

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、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、H_1と略す。

仮説の設定

例えば、2つのグループ (A,B) 間で「差がある」という仮説について検定する場合、まず「2 つのグループ間には差がない；A=B」とう仮説（帰無仮説）を設定し、元々証明したい「差がある；A≠B」という仮説（対立仮説）は、一旦伏せておく。

例題) 具体的な例で学ぶ統計学の項目の最初に紹介した男子学生の集団 111 人の身長データを使って、検定を行ってみます。この集団の標本平均は 169.3cm、標本標準偏差は 5.4cm です。今、この世代の身長の全国平均が 165cm であるとすると、この集団の身長の平均が全国平均と異なるかどうかを検定する。

（この続きは、自分で考えてください）

有意水準（危険率）

実際の検定には、帰無仮説を誤って捨てる確率α（第 1 種の過誤の確率で危険率とも言う）を示し、そのときの検定統計量から確率を出して、

その値がαよりも小さいときに

↓

「有意水準αで統計的に有意である」といい、帰無仮説を棄却する。

検定統計量より得られる**実際の確率をＰ値** (P-value) と呼び、

Ｐ値がα以下となるような統計量の範囲を

棄却域という。

一般に有意水準αには 5% (0.05)、1% (0.01) が用いられる。

ｐ値と信頼区間の意味

例えば「カラスは基本的に黒い」という仮説を主張したいときに、あえてまず考えた。「自説を完全に覆すような仮説」、つまり「カラスが黒いかどうかは半々」というよう

なものを帰無仮説と呼ぶ。主張したいことを「無に帰す」仮説という意味。そして、帰無仮説が成立していると仮定した状態で、実際のデータまたはそれ以上に帰無仮説に反するようなデータが得られる確率をｐ値と呼ぶ。Probability の意味。この場合、カラスが黒いか白いかは半々のときに、100 回連続で黒いカラスが見つかったというような観察結果が得られる確率が、1 兆分の 1 よりも小さいというのが今回の場合のｐ値である。これが小さければ「その帰無仮説はあり得ない」と考える方が自然である。どれくらいｐ値が小さければ「あり得ない」と考えるかという目安として、分野にもよるが、概ね 5%未満、つまり帰無仮説のもとでは 20 回に 1 回程度しか起こらないようなデータが得られたとすれば「あり得ない」と考えるのが慣例です。

なぜ、5%を境目とするか、特に数学的な根拠はないが、統計学者フィッシャーがかつて「p 値を 5%で判断するのが便利だ」と書いたことがきっかけになっているらしいのです。

「黒いカラスが 9 割」「黒いカラスが 8 割」という帰無仮説を唱えてもよいのだが、「完全に台無しにする」仮説以外の帰無仮説についても、どこまでならありえない仮説で、どこからは否定しきれない仮説なのか、という区間を示す。これが信頼区間の真の意味です。信頼区間が平均値±2SE で表すことができる、というのは計算上たまたまそうであるというだけで、本来の意味としては「あり得ない帰無仮説」と「否定しきれない帰無仮説」の境目がどこからどこまで、という範囲を示している。。

統計的判断における 2 種類のエラー（過誤）

第 1 種の過誤（α）→ 有意差があると判断した場合におこる。
本当は、帰無仮説が正しいのに、実際の確率 P がα以下のため、帰無仮説が違っているとして棄却する誤りを第 1 種過誤または、αエラーと呼ぶ。

第 2 種の過誤（β）→ 有意差がない（判定保留）と判断した場合におこる。
本当は帰無仮説が誤っているのに、実際の確率 P がαより大きいため、帰無仮説を棄却しない、すなわち、対立仮説を採用しない誤りを第２種過誤または、βエラーと呼ぶ。
その誤りを起こす確率は、データ数に依存する。

片側検定と両側検定

正規分布やｔ分布を利用する場合、片側、両側検定の区別が問題になります。
両側検定の方が検定が厳密である。

例）正規分布（付表 1）を用いた片側検定で有意水準 0.05 に対するｚは 1.65 です。

これは両側検定の場合の有意水準 0.1 の位置（z の値）に相当する。（両側検定で有意水準 0.05 に対する z は 1.96）
したがって片側検定では、検定統計量の偏りがより少なくても「統計的に有意」となるので、判定が甘くなる。

片側検定は、あらかじめ変化（差）の向きが理論的に片側にだけ起こると想定される場合に行う。例えば、降圧剤の効果を調べる実験で、投与後の血圧の上昇を想定しなくてよい場合に用いる。
両側検定は、処理効果がどちら向きの変化をもたらすかを予想できないときに用いる。通常は、変化の向きを予め予測できないとして、両側検定を用いる。

帰無仮説のもとでの検定統計量の分布の上限、あるいは下限の部分のみの確率を考慮して行われる検定方式を片側検定と呼ぶ。検定の有意水準を α とした場合、上側と下側の有意水準はそれぞれ $\alpha/2$ とするのが普通です。このような検定法は、両側検定（both sided test, two tailed test）と呼ばれている。検定の検出力は片側検定（one sided test, one tailed test）の方が高くなる。しかし、片側検定での対立仮説について、事前に確定的な情報が使用できればよいが、単なる思い込みによるものであれば、検定結果は過大に評価されるものとなる。事前情報が無い場合は、両側検定を行わなければならない。

<u>検定の手順</u>

ある 2 つのグループからそれぞれ抽出した標本について調べたところ、ある統計量に差があり、その差が偶然なのかそれとも意味があるのかを検定する場合の手順を以下に示す。

1）標本の分布などを仮定する（ここでは当然、正規分布になることを考えている）

2）「2 つのグループに差がない」と否定したい仮説「帰無仮説」をたてる。次に「2 つのグループには差がある」の帰無仮説を否定したときに成り立つ仮設「対立仮説」をたてる。

3）偶然と判断する基準「有意水準」を決める（通常は 0.05（5%））。

4）帰無仮説のもとで、2 つのグループの標本の統計量の差を評価するための検定統計量（たとえば、t 値）の理論的な分布と有意水準に基づく棄却域を計算する。

5）標本から求めた検定統計量（t 値など）が棄却域に入っていれば、帰無仮説を棄却して対立仮説を採用する。帰無仮説が棄却できない場合は判断を保留する。

6）対立仮説を採用した場合には、具体的に差の内容を確認する。

【ここで、区間推定と検定を整理してみます】

母集団と標本

集団の中である知りたい値を母数（parameter）という。母数を知ろうとする対象を母集団（population）と呼び、そこから抽出する一部を標本（sample）と呼ぶ。統計学では、その標本から母集団の母数を推定する。

正規分布

身長や知能指数のなどの分布は正規分布に従う。この分布は平均値を中心とした左右対称の釣鐘状の分布です。この分布によると、「平均値±1.96×標準偏差」の範囲にすべてのデータの 95%が存在するなどの性質が分かっています。この性質を利用して推定や検定を行います。

ｔ分布

標本数が少ない時には、正規分布の代わりにｔ分布を使って推定とか検定を行います。

区間推定

「真の母平均は値 A と値 B の間（区間）に“おそらく”含まれているだろう」というような考え方で母平均の推定値をあらわすことができる。

このような考え方で、標本から母集団の母数を推定することを区間推定という。母集団の正規分布が仮定できる場合、”おそらく“の程度を信頼度（信頼係数）といい、一般に利用されるのは 0.95（95%）である。また、設定した信頼度（95%）で母数を含む区間のことを（95%）信頼区間という。

ｔ分布を使った区間推定

データ数がｎ、標本平均が$\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i$、標本標準偏差が$SD=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2}$

のとき、次のように母平均μの取り得る範囲を推定することができます。μの 95%の信頼区間は

$$\bar{x}-t_0\frac{SD}{\sqrt{n}}<\mu<\bar{x}+t_0\frac{SD}{\sqrt{n}}$$

ここで、t_0はｔ分布表より求めることができます。自由度とはｎ－1 です。

検定（平均値の差の検定）

2 つの学校における体重の差やヘモグロビン濃度（貧血検査）の差、あるいは 75g 経口ブドウ糖負荷後 2 時間値の 2 つの地域における差など、2 つの集団における平均値の差の検定を必要とする場合がしばしばある。

「差がある」という仮説を検定する場合、差の程度が不明なため、そのままでは検定できない。そこで考え出されたのが、その逆の「差がない」という仮説を検定して、それに何らかの矛盾が見つかれば、もとの「差がある」という仮説を採用する。逆に、明らかな矛盾がないときにはその判定を保留する。このような論法で検定を行います。

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、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、H_1 と略す。

仮説の設定としては

例えば、2 つのグループ (A,B) 間で「差がある」という仮説について検定する場合、まず「2 つのグループ間には差がない；A=B」とう仮説（帰無仮説）を設定し、元々証明したい「差がある；A≠B」という仮説（対立仮説）は、一旦伏せておく。

実際の検定には、帰無仮説を誤って捨てる確率α（第 1 種の過誤の確率で危険率とも言う）を示し、そのときの検定統計量から確率 P 値を出して、

その値がαよりも小さいときに

↓

「有意水準αで統計的に有意である」といい、帰無仮説を棄却する。

検定統計量より得られる**実際の確率を P 値** (P-value) と呼び、

P 値がα以下となるような統計量の範囲を

棄却域という。

一般に有意水準αには 5% (0.05)、1% (0.01) が用いられる。

t 分布表から t_0 を求めるには、エクセルでは TINV（確率、自由度）で求める。

計算から求める t は、

$$\frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{(n_1-1)SD_1^2+(n_2-1)SD_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

です。

パラメトリック検定とノンパラメトリック検定

パラメトリックとは、「平均・標準偏差・データ数」などのパラメータ（母数）を使った」という意味。例えば、A 校の学生 6 人と B 校の学生 4 人が 400m 競争をして勝負する場合、まず単純にそれぞれの平均タイムで勝負するという方法が考えられる。これは平均タイムを学校全体の様子を代表する値として利用するので、「パラメトリックな勝負」であるとい

える。ただし、学生の中にたまたまオリンピック選手が 1 人まじっていたら、1 人だけ極端なタイム（外れ値）を出して平均に大きく影響してしまう。このような外れ値がある場合、平均などは学校全体の代表値としては不正確になる。

一方、ノンパラメトリックとは、「(平均などの）パラメータを使わない」という意味である。上と同じように 400m 競争を例にとると、タイムを無視してゴール順に 1 位 10 点、2 位 9 点、、、、、10 位 1 点というように得点配分し、A 校と B 校の獲得した得点で勝負を決めるような場合が考えられる。この場合、比較に平均タイムなどを用いていないため、「ノンパラメトリックな勝負」といえる。このような方法の場合、オリンピック選手が大差で勝っても、ふつうの人がぎりぎり勝っても 1 位の 10 点にはかわりはないため、外れ値の影響を排除することができる。ただし、得られた平均得点は学校全体の代表値としては外れ値がない場合の平均タイムに比べて情報量が少なく、両校の差がどの程度あるのかはよくわからない。

検定方法の選択では、パラメトリックとノンパラメトリックのどちらを選ぶかは、標本から予想できる母数が信頼できるかどうかにかかっている。なお、実用上は「標本から母集団が正規分布に従うと予想できるか否か」と考えても差し支えない。

・標本が正規分布をしている場合、平均・標準偏差・データ数の 3 つの値（パラメーター）で、分布の様子を決定できるため、この 3 つのパラメーターを用いた検定を行うことになる（パラメトリック検定）

・量的データであっても、外れ値などが存在し、正規分布が仮定できない（つまり、上述の 3 つのパラメーターから分布の様子が予想できない）場合には、これらのパラメーターを用いない検定を行うことになる（ノンパラメトリック検定)。ノンパラメトリック検定には、いろいろな種類があるが、例えばウィルコクソンの順位和検定のように、数値の順位を用いたものなどがよく利用される。

<u>母平均の検定①</u>

全国 10 歳女子身長の平均と標準偏差は、それぞれ μ ＝140 ㎝、σ ＝5 ㎝の正規分布となる。今、ある 10 歳のクラス 25 人の身長の分布は、正規分布になるものとし、その平均値$\bar{x}$は 137 ㎝である。このクラスの身長は全国水準と違うと言えるのか。

(1) 仮説設定；「全国水準と違う」という仮説は、違いの程度を特定できない。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；μ ＝140 ㎝）を採用し、もとの仮説（対立仮説 H1；$\mu \neq$140 ㎝）を一旦、伏せておく。

(2) 検定統計量を求める；この場合、25 人の身長の平均$\bar{x}$＝137 ㎝を統計検定料とする。

(3) 確率 P を求める；H_0 が正しい場合、標本平均$\bar{x}$は平均値 μ、標準偏差$\frac{\sigma}{\sqrt{n}}$の正規分布に従うことを利用する。 $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{137-140}{\frac{5}{\sqrt{25n}}} = -3$、付表 1 から z=3.0 面積は 0.0013 となる。

両側の面積は 0.0026 となる。つまり、確率 P は 0.26%となる。

(4) 判定；H_0 が起こる確率は 0.26%である。このようなまれな現象が実際に起こった考えるよりは、H_0 が正しくないと考える方が妥当である。したがって、H_0 を棄却して対立仮説 H1 を採用する。すなわち、そのクラスの身長は、全国平均と比べて差があり、それは有意水準 0.05 で統計学的に有意であると判定する。

母平均と検定②

ある看護学校の学生 15 人の身長を計測し、平均$\bar{x}$が 162.0 ㎝、標準偏差 SD が 5.0 ㎝であった。20 歳の全国平均は 157.5 ㎝である。この学生の集団は、有意水準 5%で、一般よりも身長が高いと言えるのか。

(1) 仮説の設定；帰無仮説 H_0；μ ＝157.5 ㎝)、対立仮説 H1；$\mu \neq 157.5$ ㎝

(2) 検定統計量を求める；この場合、15 人の身長の平均$\bar{x}$＝162.0 ㎝を統計検定料とする。

(3) t の値を求める；この場合、t 分布表から有意性の検定を行う。

$$t = \frac{\bar{x} - \mu}{\frac{SD}{\sqrt{n}}} = \frac{162.0 - 157.5}{\frac{5.0}{\sqrt{15}}} \fallingdotseq 3.49$$

自由度 df=n－1=15－1=14 なので、付表 2 より面積が 0.05（5%）の t 値は 2.145 である。

(4) 判定；計算で求めた t ＝3.49 の方が t 分布表より求めた t_0＝2.145 よりも大きい。したがって、帰無仮説は棄却され、この集団の平均身長は全国平均よりも高いと結論付けることができる。ただし、有意水準 0.05 である。

2 つの集団における平均値の差の検定

対応のある場合とない場合とは？

t 検定には対応のあるデータに対する検定と、対応のないデータに対する検定の 2 種類があります。ここで、”対応のある“とか”対応のない“とはどういうことなのでしょうか？

その答えは、同一個体（人でも動物でも機械でもよい）におけるある処置の前と後の状態（前後のある測定値）を比較するのが対応のある場合となります。一方、異なる個体間である測定値を比較するのが対応のない場合となります。

例えば、手術の前後で体温が変化するかどうかを検定するため、5 人の患者の手術直前と直後の体温を測定したとします。この場合、1 人の患者に対して術前の体温と術後の体温の 2 つのデータが対になって得られます。このように 2 つのデータ集団が同一個体から得られる場合を対応があるといいます。

対応のある場合

例題）表のように、手術前後の体温の比較をしてみます。

表 12 対応のあるt検定の方法

患者	術前の体温	術後の体温	差	差の偏差	差の偏差平方
イ	36.7	35.5	−1.2	0.4	0.16
ロ	36.5	35.6	−0.9	0.7	0.49
ハ	36.5	35.6	−0.9	0.7	0.49
ニ	36.4	35.6	−0.8	0.8	0.64
ホ	36.6	35.1	−1.5	0.1	0.01

この場合、表のように同一個体の差を求め、その平均値、偏差、偏差平方、偏差平方和、標準誤差を求める。そして、t ＝差の平均値/差の標準誤差を求めて、t 分布表（付付表 2）の t 値と比較する。危険率 5%で有意差があるかどうか検定してください。下記に、差の平均、差の偏差平方和について、式を立てて、求めてみてください。その後、差の標準誤差について、式立てをしてください。なお、差の標準誤差の答えは 0.281、 t は 2.84 です。ところで、自由度はこの場合、いくつですか？答えてください。

対応のある場合の一般式

先程も説明しましたが、ある集団について、冬と夏の血圧を測定して両者を比較する場合や、同一人の左右の握力を測定して比較する場合、同集団で夏と冬の食事の内容（タンパク質、脂肪、摂取エネルギーなど）を比較する場合なども、2 変数は独立ではなく、対応のあるデータとなります。この場合に用いる t を求めるための一般式を下記に示します。

ここでは、冬の血圧値；x_1,x_2・・・・,x_n、夏の血圧値；y_1,y_2,・・・,y_n、各個人の冬と夏の血圧値の差；$z_1=x_1-y_1,z_2=x_2-y_2$,・・・,$z_n=x_n-y_n$、その平均を$\bar{z}$、標準偏差を SD_z とすると、 t は次のように求めることになります。

$t=\frac{\bar{z}}{\frac{SD_z}{\sqrt{n}}}$ この値と t 分布表より自由度 df から求めた t_0 とを比較する。

対応のない場合

対応のない場合は 2 つのデータ集団の個体は同一ではなく、対応のある場合のように、2 つの測定値の差というものはありません。したがって、2 つの集団それぞれの平均値を比較することになります。

平均値の差の検定

2 つの学校における体重の差のように、2 つの集団における平均値の差の検定を必要と

する場合がしばしば生じる。この場合、通常は 2 つの母集団の平均、標準偏差とも不明なことが多く、基本的には t 検定を行う。ここでは、問題としている変数 x 、 y 、その母標準偏差をそれぞれ σ_x、σ_yとすると、次の条件があります。

（1） x 、 y が独立であること

（2） x 、 y がともに正規分布に従うこと

（3） $\sigma_x = \sigma_y = \sigma$ であること

以上の条件のもとで t 検定を行います。例えば、母集団と標本が次の表のようだとすれば、次式が $df = n_1 + n_2 - 2$ の t 分布に従います。

表 13　対応ない場合の母集団と標本の統計量

	母集団		標本		
	平均	標準偏差	平均	標準偏差	データ数
A グループ	μ_1	σ	$\overline{x}$	SD_x	n_1
B グループ	μ_2	σ	$\overline{y}$	SD_y	n_2

$$t = \frac{\overline{x} - \overline{y} - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)SD_x^2 + (n_2-1)SD_y^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

今、男女 10 人の身長を測定すると、次のようになった。

男子；178、168、170、174、164、171、169、171、170、180

女子；155、162、159、147、162、151、160、151、162、161

男子； $\overline{x} = 171.5$、$SD_x = 4.72$

女子； $\overline{y} = 157.0$、$SD_y = 5.58$

ここで、帰無仮説 H_0；$\mu_1 = \mu_2$、対立仮説 H_1；$\mu_1 \neq \mu_2$ として検定すると t の値は次のようになる。

$$t = \frac{171.5 - 157.0}{\sqrt{\frac{(10-1)4.72^2 + (10-1)5.58^2}{10+10-2}\left(\frac{1}{10}+\frac{1}{10}\right)}} = 6.28$$

となる。大学生では男は女より体格が大きいので、片側検定で行ってみる。 t 分布の自由度 df は 18、有意水準 $\alpha = 0.05$ のとき、片側検定なの付表 2 の α が 0.1 のところを見て、自由度 $df = 18$ でみる。そうすると、 t_0 は 1.734 を読み取ることができる。計算で求めた t 値は 6.28 で t_0 よりも大きい。したがって帰無仮説は棄却され、男の身長は女よりも大きいと結論付けることができる。

演習問題

１．次の資料は、男女学生の体重です。体重が正規分布に従うと仮定して、t 分布により、男女の体重（μ、μ）が等しいという仮説を検定してください。

ポイント：片側検定で行う。有意水準は 0.05 とする。

男：65，58，53，63，70，68，48，62，70，56

女：51，53，47，42，49，51，55，51，55，52

男の平均＝61.3、女の平均＝50.6、男の標準偏差＝7.44、女の標準偏差＝3.89

t＝4.03

２．次の資料は妊娠前と妊娠後の体重です。妊娠前後の体重に差がないという仮説を検定してください。

妊娠前：52.0，51.0，42.0，51.0，42.0，49.0，51.0，44.0，46.1，47.0

妊娠後：61.5，57.3，50.7，60.2，52.5，53.6，57.1，56.0，55.6，54.7

差の平均は-8.41、差の標準偏差は 2.24、ｔ＝-11.87

少ないデータのための t 検定とフィッシャーの正確検定

t 検定の「t」は「test」に由来しているらしい。t 検定の発明者であるウィリアム・ゴセットはオックスフォード大学で化学と数学を専攻し、それに回帰分析や相関係数の発明者であるカール・ピアソンのもとで統計学を学んだ。卒業後の彼の仕事はギネス社で統計学や化学の知識を生かし醸造工程と原材料を改良する事であった。つまり、彼は民間で働く統計家として最も古い時代の人物です。彼の発明した t 検定のことを「student の t 検定」と呼ばれています。これは彼が会社に秘密で研究成果を公表するために student というペンネームを用いたことに由来します。

z 検定と t 検定の基本的な考え方は共通しており、どちらも「平均値の差」が「平均値の差の標準誤差」の何倍になるのか、という値が確率的にどれほどあり得ないかを示す p 値を求める。

理論上、分散の「真の値」とはデータの「真の平均値からのズレの二乗の平均値」です。ただし実際には「真の平均値」はわからないので、「データの平均値」との差の二乗を用いて*計算された推定値を使う*。しかし、データが少なければ少ないほど「データの平均値」は「真の平均値」から離れた値になってしまいがちである。このため、データが少なければ少ないほど、サンプルの分散は「真の分散」より小さめの値になってしまうし、サンプルの分散を用いて計算されたサンプルの標準誤差も当然小さめの値になってしまう。

例えば、成人男性の身長（真の平均値が 170 ㎝）、得られた 3 人のサンプルがそれぞれ 172、174、176 ㎝。平均は 174 ㎝。「真の平均値」からのズレとして分散を考えた場合と、「データの平均値」からのズレとして分散を考えた場合とどうなるのか。「真の平均値」からのズレで計算すると約 18.7.「データの平均値」からのズレで計算すると約 2.7.「データ

の平均値」がデータから計算される以上、データの値とその平均値は独立したものではない。データがたまたま本来の分布の中で大きなのもに偏っていれば当然その平均値も大きなものになるし、逆もまた成立する。したがって、「データの平均値」と「真の平均値」が一致していない限り、「データとそこから求めた平均値」は「データと真の平均値」よりも全体としては必ず近い値になる。それが限られたデータだけで分散を求めるとどうしても値が小さくなりがちであるという理由である。

そこでゴゼットとフィッシャーは「データから求められた分散と、そのデータの数の間にどのような関係があるか」を数学的に整理した。測地学の研究で有名なヘルメルトによって発見され、カール・ピアソンによって名付けられたχ^2分布を用いれば、計算に用いたデータの数ごとに異なる、データから求められた分散が真の分散からどの程度異なるものになるのか、という分布が計算できると明らかにした。χ^2分布というのは、平均値が 0、分散が 1（すなわち標準偏差も 1）の正規分布に従う X という変数を考えた時、この変数の二乗をいくつか足し合わせたものが従う分布である。このχ^2分布は、足し合わせる X 二乗の数（自由度）によって分布の形状が異なる。自由度が無限大の時には正規分布に完全に一致し、数百～数千という大きな自由度では「ほぼ正規分布」と呼んで問題ない。このχ^2分布の自由度ごとに、「平均値の差」が「平均値の差の標準誤差」の何倍以内に収まる確率が何%となるのか、を計算するための分布がχ^2分布である。

<u>データの数が限られた場合は「フィッシャーの正確検定」</u>で

クロス集計でみると、どのセルにもできれば 10、最低でも 5 以上の数字が入る場合は z 検定を行って問題ない、というのが慣例的な目安である。

なお、データ数が極端に少ない場合、フィッシャーの直接確率検定を使う手がある。「直接確率」とは、正規分布への近似ではなく、正確な確率計算を用いて p 値を算出するという意味である。

表 14　データが少数の場合

	主任以上	役職なし	合計
体育会出身	4 人（66.7%）	2 人（33.3%）	6 人
その他	1 人（25%）	3 人（75%）	4 人
合計	5 人（50%）	5 人（50%）	10 人

計算してみると、出世者 5 人中 4 人が体育会出身者となる p 値は 26.2%、ただし体育会の出世率のほうが高い場合のみを考える片側検定の p 値である。

「出世者 5 人中 4 人が体育会出身」という実際に得られた組み合わせが偶然得られる確率は 23.8%、これ以下の確率でしか起こらない状況として、「出世者 5 人中 2 人だけが体育会」と「出世者 5 人中 1 人だけ体育会」を足し合わせ、合計 52.4%（5 人の確率 2.4、4 人 23.8、2 人 23.8、1 人 2.4%）というのが両側検定の p 値である。つまり、たった 10 人のデータ

から、実際に得られたデータあるいはそれ以上に起こりにくいような体育会とその他出身者の出世率の差が偶然に得られる確率は 52.4%、このように 2 回に 1 回以上の確率で偶然生じるような程度の違いであれば、当然「たまたまそうなっただけかもしれない」と疑ってかかったほうがいい。

t 検定について最低限知っておけばよいこと

t 検定とは数十件程度のデータでも正確に z 検定を行えるようにしたものであり、数百～数千件といったデータに対しては t 検定と z 検定の結果はよく一致する。

t 検定は、z 検定と同様に「平均値の差」が「平均値の差の標準誤差」の何倍かを考えてそれがどれほどあり得ないか p 値を求めるものである。

フィッシャーの正確検定は「組み合わせの数」を使って数十件程度のデータでも正確に割合に意味があるのか p 値を求めるものである。

もう一度、検定とは

t 検定は、1 群の平均値について、母集団の平均との比較をするとき、あるいは 2 群の平均値の差について検定するときに使う。なお、2 群の平均値の差の検定の場合、母分散が未知であり、等しいと考えられない場合には、ウェルチの検定を行う。

分散分析は 3 群以上の平均値の差の検定に使う。F 検定は等分散性の検定に使う。これは 2 群の母平均が等しいとみなせるかどうか判断するときに使う。カイ二乗（χ^2）検定は、クロス表（マスターテーブル）、ある食品を摂取した人としない人とで症状を有する人の割合が相違しているかどうかを確かめる（分割表の独立性の検定）ときに使う。この検定は、ほかに適合度、一様性の検定などに使う。なお、観測度数が小さいとき、とくに 4 以下の時は、フィッシャーの直接確率計算法を利用する。

もう一度、パラメトリックな検定（正規分布を仮定している）

t 検定は、1 群の平均値について、母集団の平均との比較をするとき、あるいは 2 群の平均値の差について検定するときに使う。なお、2 群の平均値の差の検定の場合、母分散が未知であり、等しいと考えられない場合には、ウェルチの検定を行う。

分散分析は 3 群以上の平均値の差の検定に使う。F 検定は等分散性の検定に使う。これは 2 群の母平均が等しいとみなせるかどうか判断するときに使う。カイ二乗（χ^2）検定は、クロス表（マスターテーブル）、ある食品を摂取した人としない人とで症状を有する人の割合が相違しているかどうかを確かめる（分割表の独立性の検定）ときに使う。この検定は、ほかに適合度、一様性の検定などに使う。なお、観測度数が小さいとき、とくに 4 以下の時は、フィッシャーの直接確率計算法を利用する。

もう一度、ノンパラメトリックな検定

平均値の差の検定やｔ検定は、正規分布を仮定して成り立つもの（パラメトリック法）。しかし、統計的な検定を行う場合、このよう母部集団の分布を仮定して検定することができる場合がすべてではない。例えば、質問紙調査の回答としてしばしばみられる、「１．とてもそう思う」「２．ややそう思う」「３．あまりそう思わない」「４．まったくそう思わない」といった順位尺度データの場合には、2 つの質問項目の回答に有意な差があるかどうかについて知る方法として、母集団の分布形を仮定しないノンパラメトルック法を用いる。独立 2 標本の代表値の差の検定は、マンホイットニー法（U テスト）、対応のある 2 群の代表値の差の検定はウィルコクソン（の符号付順位和）検定を使う。2 群以上の代表値の差の検定にはクラスカルウォリス検定を使う。抽出した標本の元となる母集団と特定の理論分布が一致するかどうかを検定するにはコルモゴロフスミルノフ法（K－S 法）を使う。このとき、累積相対度数を比較する。また、正規分布からのずれの大きさによって外れ値の有無を判定するには、グラッブス・スミルノフ棄却検定を使う。

Ⅵ　分散分析（Analysis of Variance）

英語の原語を略して通称 ANOVA（アノヴァ）と呼んでいます。分散分析もｔ検定と同様に平均の有意差の検定をします。t 検定は 2 グループの平均しか比べることができないのですが、分散分析は 3 グループ以上の平均も比べることができます。

分散分析の考え方：

ここでは簡単のため、2 条件 A・B の平均の有意差検定について考えてみます。分散分析では、条件 A と条件 B のデータを同一の母集団からの標本とみなし（帰無仮説）、この母集団の真の値の推定値として条件 A と条件 B の「平均の平均」を求める。すなわち、$(X_A+X_B)/2$ です。これを大平均と呼ぶ。分散分析はこの大平均からのズレに注目する。そうすると、

データの大平均からのズレ＝平均の差の影響によって生じたズレ＋偶然の影響によって生じたズレ

分散分析は、このようなズレの分解を全データについて行う。そして、データのズレは、平均の差の影響力なのか、それとも偶然の影響力なのかを比較する。もし、前者の影響力が大きければ、平均の差は文字どおり偶然以上であることになる（有意である）。

Ⅶ　相関・予測

相関係数は 2 つの変量ｘ、ｙの関連の強さを表す指標で－1 から 1 の間の値を取る（負の相関、正の相関、相関なし）。また、両者に直線的相関が想定できるとき、2 変数の関

係は、直線の式で表すことができる。ｙ＝bx＋a　　（この式の場合、ｙが従属変数、ｘが独立変数になる）
この直線を回帰直線、直線の式を回帰式という。また、この式の係数の決定方法が最小２乗法である。

多変量解析

複数個の変量をもつ多変量データを変量間の相互関係を考慮に入れて分析する一連の統計的手法が多変量解析である。医学、看護学分野では、1 人の人間がもつデータは、体格、血圧をはじめ、血圧データ、心理データなど、単に 1 時点をとってみても、非常に多数あり、そのほとんどは相互に関連しあっているといえるもの。したがって、人間に関する医学、看護学データの分析にとって、多変量解析は必要不可欠なもの。
主成分分析は、多くの変量が与えられたとき、それらの変量のもつ情報をできるだけ多く表現できるような合成得点を求める分析方法。例えば、人の「身長」「体重」という 2 つの変量を考えると、身長も体重も大きい人は「体格」がよいとされる。このことから、これら 2 変量をもとに「体格（の良さ、悪さ）」という主成分を求め、「体格」だけで、「身長」「体重」の 2 変量がもつ情報をある程度表現しようとしたもの。
因子分析は、数多くの変量の中に潜む共通の因子を探り出す分析方法。例えば、上記の例で人の「身長」「体重」「胸囲」などの変異から、「体格増大の素因」「肥満傾向」などの基礎として隠れた因子を探り出すもの。典型的なのものには、性格特性の分析がある。
重回帰分析は、ある変数を、いくつかの変数によって予測する式を作成する。
判別分析は、いくつかの変量をもつ個体が、いくつかの母集団のうちのどれかに属しているかを判別する分析方法。変量間の相関関係をもとに、母集団ごとの差が明確に現れるような値が得られるように各変量に重みをつけ、そうして得られた重みづき得点により母集団を判別する。
クラスター分析は個体のもつ変量の類似度をもとに、個体をいくつかの集団（クラスター）にまとめ上げていこうという手法の総称。類似度は、個体間の距離の算出。

相関

相関・予測の分析は条件間の関連を見ます。したがって、各条件のデータどうしは独立であってはならず、対応のあるデータでなければならない。対応のあるデータを入手するには、1 人の被験者または対象者から 2 個以上のデータをとればよい。

例えば、1 被験者を身長と体重である。このとき、データの種類や単位は違っていてもかまわない。相関・予測の分析は、このような対応するデータどうしの間に、次のような 3 点の規則的関係を見出そうとしている。

① 正の相関

いわゆる正比例の関係があるかどうかを見る。すなわち、一方のデータの数値が大きい場合に、それに対応する他方のデータの数値も大きい（小さい場合には小さい）なら、正の相関がある。

② 負の相関

いわゆる逆比例の関係があるかどうかをみる。一方のデータの数値が大きい場合に、それと反対に他方のデータの数値が小さい（小さい場合に大きい）なら、負の相関がある。

③ 予測式

データどうしの相関関係を、一次方程式として表現する。この方程式に一方のデータの任意の数値を代入すれば、他方のデータの数値を予測的に求めることができる。

変数

変数（variable）とは「変化する数値」という意味である。相関・予測の分析では、データを変数とみなす。例えば、サーブの練習本数のデータが一つの変数、成功率のデータがもう一つの変数であるとすると、「2 変数の相関と予測を分析した」という言い方をする。このように、変数という言い方をするのは、相関・予測の分析がデータどうしの対応関係を数学の関数関係になぞらえて分析しようとするからである。

予測と回帰

予測は回帰ともいわれる。回帰（regression）とは「もどす」ことであり、予測関係を逆方向から見たときの言い方である。

このように、予測も回帰も、言っていることは同じであり、見る方向が異なるだけである。「予測」という言い方を多用するが、専門的には「回帰」という言い方が好まれる。

相関係数の計算

変数間の相関・予測が直線的であることをチェックしたら、次に相関係数を計算する。相関係数とは、2 変数の相関の強さと方向を表す統計量である（3 変数以上の場合には重相関係数がある）。この統計量は K.Pearson の考案によるので、特に「ピアソンの相関係数」と呼ぶこともある。

記号

相関係数の記号は“ r ”である。なお、 r はデータの相関係数である。このデータの背後に存在する無限データ集団（母集団）の相関係数を表すときには、 r に相当するギリシャ文字“ ρ ”を用いる。

r の性質

r の値は 0 をはさんで、−1〜＋1 の範囲で変化する。r の値が－の場合は、「負の相関」であり、散布図上のデータは右下がりの直線に収束する傾向を示す。これに対して、r の値が＋の場合は「正の相関」であり、散布図上のデータは右上がりの直線に収束する傾向を示す。完全相関（r ＝−1、r ＝＋1）の場合は、全データが一本の直線上に乗るが、それ以外は、直線への収束には程度の違いがある。なお、r ＝0 は「無相関」であり、データはどのような直線へ収束するきざしもみせない。

相関係数の計算

変数 A と変数 B のデータを、下表のように入手したとする。

表 15　データ・リスト

対象者	変数 A	変数 B
a	1	6
b	2	4
c	3	5
d	4	3
e	5	1

表 16　データ表示(N=5)

	変数 A	変数 B
$\bar{x}$	3.0	3.8
SD	1.4	1.7

① 偏差積和を計算する。

偏差とは、データと平均との差のことです。偏差は各対象者ごとに 2 個できるので、それを掛け合わせて（積）、全被験者について合計した値が「偏差積和」となります。

偏差積和＝（1−3.0）×（6−3.8）＋（2−3.0）×（4−3.8）＋（3−3.0）×（5−3.8）＋（4−3.0）×（3−3.8）＋（5−3.0）×（1−3.8）

＝（−4.4）＋（−0.2）＋（0）＋（−0.8）＋（−5.6）

＝−11.0

ここでは、たまたまマイナスの積が多く出たので、和もマイナスになりました。この「偏差の積」のイメージとしては各データと平均とのズレを面積として表現したものと考えることができます。

①　データ 1 組分の偏差積和に変換する。

上で計算した偏差積和は、変数 A と変数 B の対応するデータ 5 組分の値です。これを 1 組分の値にします。この値が、共分散といいます。

$$\text{共分散}=\text{データ 1 組分の偏差積和}=\frac{\text{偏差積和}}{N}=\frac{-11.0}{5}=-2.2$$

②　下式によって r を計算する。

$$r = \frac{\text{データ1組分の偏差積和（共分散）}}{SD_A \times SD_B} = \frac{SD_{AB}}{SD_A \times SD_B} = \frac{-2.2}{1.4 \times 1.7} = -0.92$$

ここで、相関係数 r ＝－0.92 は二つの情報をもっている。一つは値の大きさ(0.92)が意味する、もう一つは値の符号（マイナス）が意味することです。

r の値の大きさが意味すること

上の r の計算式における分子の「データ 1 組分の偏差積和＝共分散」とは、変数 A と変数 B のデータ 1 組分の偏差で出来る平均的面積です。これに対して、分母の「$SD_A \times SD_B$」は各変数の標準偏差を組みとなしたときに出来る面積です。したがって、データの各組の値のとり方が「$SD_A : SD_B$」と同比率であれば、分子は分母に収束し、「r ＝±1」となります。このとき、散布図上のデータの点は完全に直線上に並ぶ。このように変数 A と変数 B の標準的な座標（SD_A、SD_B）を仮定して、実際のデータの組みがそれにどの程度、近似した値のとり方をするか、という程度を面積比とした値が r です。 r ＝－0.92 とは完全な直線に対する 9 割程度の近似を意味しています。

r の値の符号が意味すること

次に、 r ＝－0.92 はマイナスに出ているが、これは偏差積和がマイナスになっていたからです。マイナスの積を出すデータの組が多かった（①においてデータ 5 組のうち 4 組がマイナスの積を与えた)。このように、マイナスの積が多く出るということは、平均の座標からみて、その左上と右下にデータが集まっていることを意味しています。

さて、検定の結果は、有意でした（F（1,3）＝16.53、p＜.05)。すなわち、 r ＝－0.92 は「真の相関係数 0」＋「偶然のユレ」としてはめったに出現することのない値です。したがって、実質的に相関係数は 0 でないと認めてよいであろう。しかし、この有意性は、その相関が強い相関であることまで保証しない。有意であっても、取るに足らない弱い相関もある。そこで、有意性検定において有意であった場合は、次の相関の強さを判定する必要がある。

判定の基準

相関の強さの判定は経験的であり、一般に、以下の基準を用いている。

注意：相関係数の値と相関の強さの関係

先の判定基準からわかるように、相関係数の値と相関の強さとは直線的関係にない。例えば、「弱い相関」の段階は 0.2～0.4 の 2 ポイントの幅であるが、「中程度の相関」の段階は 0.4～0.7 の 3 ポイントの幅にとってある。したがって、相関係数の値が二倍であるからといって、相関の強さも二倍であると言えない。この点、相関の強さを正確に理解したいなら、説明率または決定係数（coefficient of determination）といわれる指数を計算してお

くほうが、誤ったイメージをもたないですむ。

説明率の計算

説明率とは、一方の変数が他方の変数をどの程度、説明するかを表す統計量です。この説明率と相関の強さとは同義である。専門用語では「決定係数」と呼ぶこともある。　説明率の計算は、下式による。

$$R^2 = \text{説明率（\%）} = r^2 \times 100$$

例えば、r ＝0.35 の説明率は $0.35^2 \times 100 = 12\%$である。r ＝0.35 は r ＝0.70 と比べると相関係数としては二分の一であるが、説明率を計算してみると、実は相関の強さとしてはもっと弱く、四分の一以下であることがわかる（r ＝0.70 の説明率は 49%)。

ちなみに、先の判定基準を説明率に置き換えてみると、強い相関は 50%以上、中程度の相関は 15~50%、弱い相関は 5~15%、ほとんど無相関は 5%以下となる。

e. 予測式の算出

分析の目的

予測の分析の目的は、一方の変数の値から他方の変数の値を予測する方程式を求めることである。この方程式を予測式または回帰式という。予測と回帰のちがいは、たんに見方の違いである。例えば、「変数 A を変数 B から予測する」は「変数 A を変数 B に回帰させる」ことでもある。以下「予測」に統一するが、「回帰」に読み替えてもよい。

予測式の説明

さて、変数 A を変数 B から予測する数式は、一般に、次のような形になる。

A`＝k・B＋y

この予測式は、散布図上では、データが収束すべき直線を表している（この直線を予測直線という)。式中の A`は、変数 A の予測された値であり、必ずしも実際の値と一致しないので、A とは区別して表している。k はいわゆる「傾き」である。k が＋なら予測直線は右上がり、k がーならば予測直線は右下がりになる。専門用語では、「回帰係数」と呼び、変数 A が変数 B に依存する(回帰する)程度とみている。なぜなら、k の値が大きければ大きいほど、変数 A の予測値（A`）は変数 B が 1 ポイント上下するたびに大きく変化するからである。また、このため k を「重み」と表現することもある。重みが大きいことは、変数 A に及ぼす変数 B の影響が重大であることを意味している。予測式の最後の y は定数であり、いわゆる「y 切片」である。すなわち、予測直線がタテ軸を切る点を表している。

3 変数以上の相関・予測の分析

これまで、2 変数の相関を分析してきた。今度は、もっと変数の多い場合、すなわち 3 変数以上の相関を分析してみる。

3 変数以上の相関を分析するには、以下の 2 つのコースがある。

コースⅠ:「相関マトリスクの作成」→　「因子分析」

このコースは、変数間に特定の予測関係が存在しない場合に選択する。基本的には、2 変数ずつを取り出して、一組ずつ相関係数を計算する（相関マトリスクの作成)。また、その結果に基づいて、多くの変数を少数の因子にまとめる「因子分析」へ発展することもできる。

コースⅡ:「重相関係数の計算」→　「回帰分析」

このコースは、変数間に特定の予測関係が存在する場合に選択する。すなわち、全変数のうち、ある 1 個が目的変数、あとの残りがすべて予測変数と最初から決まっていなければならない。このコースを選択すると、まず「重相関係数」を計算し、次に予測式を算出する。ただし、この予測式の算出は「回帰分析」として行われるのが一般的である。すなわち、目的変数の予測に貢献する予測変数と貢献しない予測変数を、複数の候補の中から選択しようとする。

このコースⅠは、まず「相関マトリスク」を作成し、次に「因子分析」へ移行することができる。ただし、因子分析への移行の際には、変数の個数が 10 個以上（できるなら 20 個以上）は、ほしい。まず、相関マトリスクから説明する。

相関マトリスクの作成

相関マトリスクとは「相関行列」という意味であり、多くの相関係数をヨコ（行）とタテ（列）に並べた表である。相関マトリスクを作成するには、いわば全変数の「総当たり表」を作って、この表のなかで「ぶつかる」2 変数どうしの相関係数を一つずつ計算してゆけばよい。例えば、下の表 13 は 4 変数の相関マトリスクの例である。

表 17　4 変数の相関マトリスク（N＝42）

	B	C	D
変数A	.432**	.610**	.279†
変数B		.214	.186
変数C			.378*
†p<.10 *p<.05 **p<.01（両側）			

表中の各相関係数に対しては、個別に有意性検定（両側検定）を行う。そして、その結果を記号を付けて表すこと。

因子分析（Factor Analysis）は、多くの変数の相関関係を、少数の因子に要約するための方法です。ただし、そのような多変数の動きを支配する共通の因子がみつかるかどうかはケース・バイ・ケースであり、探索的です。

重相関と重相関係数の定義

重相関（multiple correlation）とは、変数が 3 個以上の場合で、かつ、この変数間にあらかじめ一定の予測関係が存在する場合の相関関係のことである。すなわち、全ての変数のうち、特定の 1 個が目的変数、あとの残りが予測変数と最初から決まっていなければならない。また、重相関係数とは、そのような 3 変数以上の間の相関関係の強さを表す統計量である。重相関係数の記号には一般に“R”が用いられる。これは 2 変数の場合の相関係数の記号”r“の大文字である。

ａ．重相関係数の計算

R の計算

3 変数 A・B・C があり、予測関数が明確に存在する（大前提）。ここでは、A が目的変数、B・C が予測変数であるとする。

1）データ表示を行う。

表 18　各変数の平均と標準偏差（満点各 20 点、（N＝50）

	条件 A	条件 B	条件 C
$\bar{x}$	14.9	15.7	13.8
SD	7.6	5.2	6.9

2）2 変数ずつ取り出し、相関係数 r を計算する。

ここでは、すでに計算ずみであり、以下のとおりとする。なお、条件 A の全データを変数 A ということにする。以下同様。

変数 A と変数 B の相関係数；$r_{AB}=0.512$

変数 B と変数 C の相関係数；$r_{BC}=0.356$

変数 A と変数 C の相関係数；$r_{AC}=0.434$

3）r の値を下式に代入し、重相関係数 R を計算する。

$$R=\sqrt{\frac{r_{AB}^2+r_{AC}^2-2\times r_{AB}\times r_{BC}\times r_{AC}}{1-r_{BC}^2}}=\sqrt{\frac{0.512^2+0.356^2-2\times 0.512\times 0.356\times 0.434}{1-0.434^2}}$$

$$=0.533$$

重相関係数 R の性質

① R は 0 から 1 までの正の値しかとらない。

重相関は 3 変数以上の対応関係であるため、その方向性をプラス・マイナスだけに限定することはできない。したがって、R は、重相関の強さだけを表す。そのように、R は強さの情報しか持っていないので、論文には R そのままでなく、強さの意味を表す R^2 を掲載するのが通例である。

② 予測関係が変わると R の値も変わる。

例えば、「目的変数 A、予測変数 B・C」としたときの R の値と「目的変数 B、予測変数 A・C」としたときの R の値とは一致しない。この点、変数間の予測関係が事前に確定していないと、なにを求めているのかわからなくなる。

③ 使用上の制約は r と同様である。
各変数間の相関の直線性、データ分布の正規性、目的変数の（各予測変数に対する）等散布性が満たされないと、R は使用できない。

b. 偏相関係数の計算

偏相関の定義

偏相関（partial correlation）とは「部分的相関」という意味である。すなわち、重相関における複数の予測変数のうち、特定の 1 個だけと目的変数との相関関係を表す用語である。例えば、変数 A・B・C があり、A が目的変数、B・C が予測変数であると考えてみる。

このとき、目的変数 A に対する予測変数 B だけの純粋な相関をみるには、この A と B の相関に加わる変数 C の影響力（A と C の相関、B と C の相関）を取り除かなければならない。偏相関とは、そうした他の変数の影響を取り除いたときの、目的変数と予測変数 1 個との部分的相関を意味している。3 変数の場合、偏相関係数は 2 個求めることができる（A と B、および、A と C)。

偏相関係数の記号

偏相関係数の記号は、通常の相関係数の記号 r に特別な添え字を付ける。例えば、“$r_{AB.C}$”のように書く。この添え字「AB.C」の意味は、「変数 A と変数 B の偏相関・変数 C の影響を取り除いたときの」ということです。

計算例

重相関係数を計算したときの例を用いて、さらに偏相関係数を計算してみる。前例では、目的変数 A、予測変数 B・C であった。

① 各 2 変数の相関係数を計算する。
$r_{AB}=0.512$、 $r_{BC}=0.434$、 $r_{AC}=0.356$

② 各相関係数を下式に代入し、偏相関係数を計算する。
まず、変数 A と変数 B の偏相関係数を計算する。

$$r_{AB.C}=\frac{r_{AB}-r_{AC}\times r_{BC}}{\sqrt{1-r_{AC}^2}\times\sqrt{1-r_{BC}^2}}$$

$$=\frac{0.512-0.356\times0.434}{\sqrt{1-0.356^2}\times\sqrt{1-0.434^2}}=0.425$$

次に、同様にして、変数 A と変数 C の偏相関係数を計算する。

$$r_{AC.B}=\frac{r_{AC}-r_{AB}\times r_{BC}}{\sqrt{1-r_{AB}^{2}}\times\sqrt{1-r_{BC}^{2}}}$$

$$=\frac{0.356-0.512\times0.434}{\sqrt{1-0.512^{2}}\times\sqrt{1-0.434^{2}}}=0.173$$

c．予測式の算出と回帰分析

目的

重相関係数と偏相関係数を計算したら、最後は予測式を算出する。3 変数以上の場合に、予測式を求める目的は二つある。１）目的変数を予測すること、および２）予測変数を選別することである。このうち、最初の１）はふつうの目的であり、目的変数の値の変化に関心がある。しかし、一般に研究例は少ない。これに対して、次の２）の目的は、予測変数どうしの有効性の比較に関心がある。すなわち、目的変数を予測するのに、どの予測変数が有効か、どの予測変数が有効でないかをより分ける。この目的をとった場合は実験計画法に似ており、結局「効く要因」の検討になる。3 変数の場合を例にとると、そのときの予測式は、例えば以下のようになる。

A‘＝1.38B＋0.77C＋11.6

上式の傾き（1.39、0.77）が予測変数 B と C の効き具合を表すことになる。ここでは、変数 B のほうが有力である。すなわち、1 ポイント変化したときに目的変数 A の値を上下する度合いが大きい（1.39＞0.77）。

なお、予測変数 B・C が共に 0 のときも、目的変数 A は 11.6 程度の値をとる。これは、なんの説明もできないので、偶然誤差によるユレであるとみなすしかない。かくして、上式は、実験計画法における分散の分解式に類似した意味を持つ（下式）。

データ A‘＝変数 B の効果＋変数 C の効果＋偶然誤差

このように予測式をデータの模型にみたてて、予測変数どうしの重み（傾き）の大きさを比較する方法を回帰分析と呼ぶ。なぜなら、分析の焦点が目的変数でなく予測変数にあるので、視点は「目的変数←予測変数」（予測）という方向でなく、「目的変数→予測変数」（回帰）という方向になるからである。したがって、回帰分析を目的とした場合の予測式は「回帰式」と呼ぶ方が適切であり、専門家も、その呼び方を好む。

参考文献

1）研究者のための統計的方法、R・A・フィッシャー著、遠藤健児・鍋谷清治共訳、森北出版株式会社、1979.

2）確率の出現、イアン・ハッキング著、広田すみれ・森元良太訳、慶應義塾大学出版会、2013.

3）統計学が最強の学問である、西内啓著、ダイヤモンド社、2013.

4）統計学が最強の学区門である[実践編]、西内啓著、ダイヤモンド社、2014.

5）統計学　系統看護学講座（基礎分野）、金森雅夫著、医学書院、2013.

6）宇部フロンティア大学における長期履修学生制度実践と課題、松本治彦著、私学経営（366）、17－25、2005.

7）学生満足度調査に関する一考察、松本治彦、佐藤美幸、網木政江、白石義孝著、宇部フロンティア大学人間社会学部紀要、Vol.1 No.1、宇部フロンティア大学出版会、2010.

8）複数連結汽水湖の水温、塩分の変動、松本治彦、合屋晏秀、李寅鉄、羽田野袈裟善、斎藤隆著、水工学論文集 39、237－242、1995.