統計学 2018（心理とビジネスを学ぶ人のための）

[松本治彦]

その 1（統計学の歴史から正規分布まで）

2018 年 10 月

宇部フロンティア大学

目　次

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ちます。

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はっきする。

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：心理統計学基礎、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。

授業概要：

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る。グループ討議の時間を設定している。

「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　9 月 27 日　統計学の歴史　その 1
第 2 回　10 月 4 日　統計学の歴史　その 2
第 3 回　10 月 11 日　統計学の考え方　その 1
第 4 回　10 月 18 日　統計学の考え方　その 2
第 5 回　10 月 25 日　データの科学的な見方
第 6 回　11 月 1 日　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　11 月 8 日　　具体例で統計学を学ぶ　その 1
第 8 回　11 月 15 日　具体例で統計学を学ぶ　その 2
第 9 回　11 月 22 日　具体例で統計学を学ぶ　その 3
第 10 回　12 月 6 日　区間推定と検定　その 1
第 11 回　12 月 13 日　区間推定と検定　その 2
第 12 回　12 月 20 日　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　1 月 10 日　ｔ検定と分散分析
第 14 回　1 月 17 日　相関と予測
第 15 回　1 月 24 日　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 Pascal1623～1662 年）はフランスの思想家「人間は考える葦である」で自然科学者でもあります「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。そこでは、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、年金が（オランダ）健全な保健計算に基づくようになった。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」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、

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、分散などー 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。Laplace 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ある。Gauss 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。Gauss 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”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。

「ベイズの定理」

ひと言でいうと、条件付き化確率。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

例えば、平均のようなある統計量を、非常に大きな標本から求めた時には、その値は正

確であるとするでしょう。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、そこで用いる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。事実、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、したがって母集団のパラメーターの関数として表されるある確定値に近づく。ゆえに、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すれば、パラメーターのただ 1 つの関数が定まって、その統計量は当然この関数に等しいとおくならば、標本の大きさが限りなく大きい場合ですら、その統計量はこの関数に対して正しい値を与えないことになる。たしかに、それはある確定値に収束はするが、その統計量が使われる観点から見れば、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

これに反して、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いかほどでも近づく。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、すなわち、標本の大きさが限りなく増大した時に要求される性質によって定義した。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ということは重要です。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、一致統計量と定義するのです。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、算術平均または中央値などのように、一般に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。しかしながら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。この語の意味を以下に示す。たとえば、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 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。

たとえば、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。しかしながら、安易ではあるが有効でないこれらの方法を教えるのに少なからず時間を無駄にしているので、他の方を何も学んでいないという学生がよくあるし、また往々にして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あてはめの誤差が入ってきてはならない。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と考えられる。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この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。しかしながら最もよく例に現れるもので、次のような目立った性質のために理論的に重要なある種の統計量がある。その性質とは、小標本の場合においても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て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。たとえば、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。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が、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とはっきりと区別すべきであり、また推定のために用いう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。

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であろう。たとえば、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て。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が、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 の分布と呼ぶことのできる第 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ができるということも同様に重要**であった。さらに広範囲の問題に必要となる数表は、その用例とともに既に刊行されている。

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）

Sir Ronald Aylmer Fisher（1890-1962）の著書　Statistical　Methods for Research Workers 第 13 版（1958,1963）の全訳。著者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、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上記の本は、長い間に次第に増加してきたことは、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（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）。

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ものである、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ということである。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、まず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。確率の誕生はブレーズ・パスカルという個人の偉業によるものとするのが通例だが、

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。いまでこそ、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、プロバビリティーの概念も変化した。プロバ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、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

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。

Ⅲ－1.　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、すなわち社会を構成する生物の集団に限られたものでもないあくまでも、個体の集まり、すなわち集団についての学問である。

集団という概念は生物的あるいは物質的なものだけに限られてはいない。直接測定のようなある観測を限りなく繰り返すものとすれば、その結果の集まりは測定値の集団である。このような集団は誤差論という研究分野に属するものである、

統計的研究の対象となる集団は、ふつうは幾つかの点に関して変動を示している。統計学は変動の学問であるということはまた、現代の統計学者の目的と昔の統計学者の目的との間の差異を強調することにもなる。

Ⅲ－1. 2.　変動

比較的最近に至るまで、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。変動それ自体は研究の目的ではなくても、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。正規標本の平均の誤差曲線は、既に 1 世紀前からよく知られていたが、標準偏差の誤差曲線は 1915 年に至るまでなお研究の対象となっていた。しかし新しい観点からすれば、小麦の収量から人の知性に至るまで、およそ変化する現象について変動の原因を研究するには、そこに現れた変動の検討と測定から始めなければならない。

頻度分布で連続的変化（無限の場合）は変量の変動範囲（集団全体に対する比率）2 つ以上の変量の変動からなる。

変動の研究から直ちに頻度分布の概念に到達する。頻度分布には種々の型があって、集団が分布する級の個数有限の場合も無限の場合もあるし、また定量的な変量の場合には、級としての区間は有限の大きさのことも無限小のこともある。最も簡単なものは、出生児の性別のように級がただ 2 個の場合であって、その時分布はそれらの級が起こる比率だけで規定される。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。

また、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り、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、その頻度分布とは、変量の取りうる各値にたいして、その値をとりうる頻度を示すものである。第 3 の組に属するのは、身長のように、変量がその変動範囲にある中間のどんな値をもとりうる場合である。このとき、その変量は連続的に変化するといい、頻度分布は変量の関数として次の 2 つの方法のどちらかによって表される。(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という数学的手段によって、集団の中で変量がその変動範囲のある無限小の部分にはいるものの (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無限の集団に対しても用いられるが、無限の場合に適用する場合のほうが有用でしかも簡単である。有限の集団はいくつかの限られた比に分割されるだけで、どんな場合でも連続的な変動を示すことはない。さらにまた、実際に起こっている原因から生じる可能性の全体を、正確にしかも正しい比率で表すことができるのは、大抵の場合無限集団だけであって、我々はこの可能性の全体を研究しようとしている。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。無限集団に関しては、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ものであって、(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、または (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がある。(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。よく知られてい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。

変動の研究から、現れた変動の量の測定ということだけではなく、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。

Ⅲ－1. 3. 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ある。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。データから利用で

きる有用な情報すべてを抽出して誤差の大きさと性質を示すことができれば、推定値の価値は増大する。

広汎な観測を行ったことのある人は、その結果を都合の良い分量に簡約するという切実な要求をよく経験するはずです。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我々は、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。これは全く実際的な要求であり、統計学はある程度までそれを満足させる。少なくともある場合には、1 つまたは数個の数値によって有用な情報の全部を与えることが実際に可能であるが、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りさえすれば、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く、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である。そして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ある。

有用な情報と無用な情報との分離は次のように行われる。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、母集団を一応規定した上で、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ある。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や、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、すく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、その仮説は正しいとされる。.

演繹的法は、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

（統計学が最強の学問である「西内啓著 2013.4」を参考して作成）

ビジネスと統計学のギャップはなぜ存在するのか

数式が出てくると読む手が止まる
しかしツールをいきなり触っても何を意味しているのか分からない
それぞれの手法が自分の仕事にどう役立つのかわからない
自分の仕事に必要な範囲がどこまでなのかがわからない

統計学は強力かつ汎用的なツールであるがゆえに、多くの学問分野で使われている。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されることがあるし、その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、ということもしばしばある。また、逆にそうした違いに踏み込まない最大公約数としての統計学としての入門書が、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、人間を「洞察」するための統計学

人間の行動の「因果関係を洞察」する。因果関係の洞察以外の統計学の使用目的には「現状の把握」と「今後の予測」の２つがある。

「洞察」の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洞察のほうが重要になる。「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」という洞察のほうが利益の源泉となる。購買という求める結果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。

これは医学や公衆衛生学においてもまったく同じことが言える。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ことこそが医学で統計学を使う目的。

「洞察」の統計学に必要な３つの知識

（１）平均値や割合など統計指標の本質的な意味の理解

（２）「データを点ではなく幅で捉える」という考え方

（３）「何の値を何ごとに集計すべきか」という考え方

「平均値」の本質が分かれば「割合」もわかる

平均値と割合と言うのは本質的にまったく同じ。「量的変数」は「平均値」の形で集計する。質的変数は「割合」を集計する。量的変数は「量として大きいか小さいか」という情報を示すものである。質的変数は「大きい小さいということではなく、そもそもの質が異なる」という情報を示す。

割合と平均値と言う全く別物の集計方法が存在しているわけではなく。例えば、100 人に対する調査で 60 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と言う値にあるものとする。この平均値は 0.6 となる。これは先ほどの 60%と言う割合と全く同じ値である。

データの存在する「幅」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ための方法を生み出した。

「結果」と「原因」を絞り込め！

何の値を何ごとに集計すべきか、これは統計学を因果関係の洞察に使う上で最も重要な枠組み。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。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ための第一歩を踏み出せる。

せっかく経験と勘に反する新しい発見に出会うためのデータ分析と言う作業をしているのに、自分の経験や勘の検証しかおこなわないのではもったいない。データ分析を因果関係の洞察、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うる原因の候補、という観点で捉える。この最終的にコントロールしたい結果を目的変数（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ぶ。

医学でも同様に、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ある。。

ビジネスにおいても同様に、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機会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るならば、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ところ

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「因果関係の洞察」の統計学の狭間で、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なことだが、混同している。この混同には、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という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。

標準偏差が示す「たいていのデータの範囲」

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、というのがここからのテーマ。

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」である。標準偏差とは単に「標準的な平均値からの偏り」である。

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±SD（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。

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」である。

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」誤りのことをαエラー、一方「本当は差が存在しているのにそれを見逃してしまう」誤りのことを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ある。この両者の過ちはトレードオフ（二つの条件を同時に満たすことはできない）である。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

そうした後に、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が、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、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。そして想定する有意水準において最も検出力の高い検定手法のことを、統計学では最強検定あるいは最強力検定という。

二つの過ちの間で、そして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ある。

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。だが、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。ゆえに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、もし無限回の調査を行ったとすれば得られるのであろう「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も実施したのです。

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」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、どれほどあ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ある。

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（変質）は、牛乳が摂氏 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だけの時間を紅茶の高温に取り囲まれる。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

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。また、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が。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が社会科学を可能にした

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。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。

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。

「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、ただ仮説やこういう事例がありましたという話だけをもとにして理論モデルを組み立てる、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。

「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。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する。

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、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」と決めつけることの愚かさ

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愚か。

ランダム化は意外とむずかしい

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。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は本来何の問題もないはずだが、やると明らかに大損をする場合、という 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」の壁とは、「絶対的なサンプル数の制限」と「条件の制御不可能性」。月へのフライトに限らず「1 回こっきりのチャンス」あるいは、あったとしてもせいぜい数回程度しかチャンスの与えられないもの自体を取り扱うことに対して、ランダム化しようがしまいが統計学は無力である。

「倫理」の壁

統計家たちの間で共有されている倫理的ガイドライン。①ランダム化によって人為的にもたらされる、どれか 1 つまたはすべての介入が明らかに有害である（またはその可能性が高い）場合はダメ、②仮にすべてが有害でなくても、明らかに不公平なレベルで「ものすごくいい」ものと、「それほどでもない」ものが存在していると事前に分かっている場合もダメ。

「感情」の壁

「そういう運次第で自分の運命が左右されるのが何かイヤ」と実験に参加する人が思う

ことを止めることはできない。

「IQ」を生み出した心理統計学

IQ とは何かを理解しようとすれば、心理学者がこの 100 年で積み重ねてきた統計手法について学べばいい。

「一般知能」の発明

現在の知能研究の基礎を生み出した心理統計家であるスピアマンの 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、こうした先行研究で示されていた種々の知能の測定方法をいくつか選び、研究参加者に対して試してみた。そしてそれぞれの「知能を表すはずの指標」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/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。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。

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

スピアマンが行った分析方法は、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。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。

そのためには自分の測定したい「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られる。

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が除外
- 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「育休中でも服飾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。

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

- 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

「大抵の場合に、図表を利用すると手軽にデータの予備的な検討ができる。

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。」

２－１　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　統計データのまとめ方

２－１　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									

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

表 2　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	

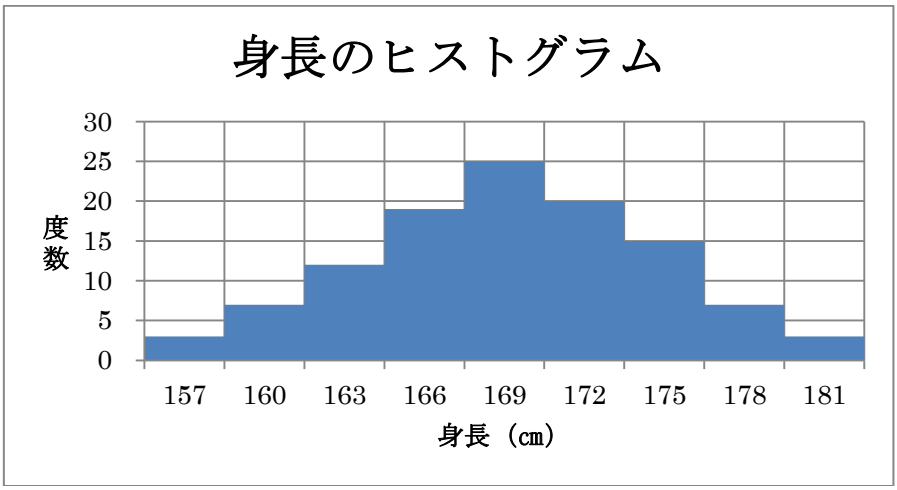


図１　身長のヒストグラム

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

度数分布表（表 2）とは、測定データをある一定の間隔（ここでは 2ｃｍ間隔）に区切り、その中にあてはまる人数を記載したもの。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 15）。

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。階級の幅の決定には、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、最大値と最小値をそれぞれ見つけてください。最大値 181.6 ㎝、最小値 155.5 ㎝　最大と最小の差（26.1cm）を 10～20 で割った値（この場合、10 で割ると 2.6ｃｍ）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 2cm あるいは 3cm を便宜上用います。この例の場合は、階級の幅を 2cm 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、160cm 以下、180cm 以上が少なく、167～173cm 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 歳である。

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ある。

累積度数 → ある階級以下の度数を合計したもの。最期の階級では、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。それぞれ度数・累積度数の百分率に一致する。

問題 1

ある会社の従業員男女それぞれ 10 人を対象に貧血検査（ヘモグロビン濃度 mg/dl）を行い、下記の結果を得た。

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

度数分布表を男女別に作成してください。

２－２ 分割表（クロス集計表）

表 3 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。X 二乗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。

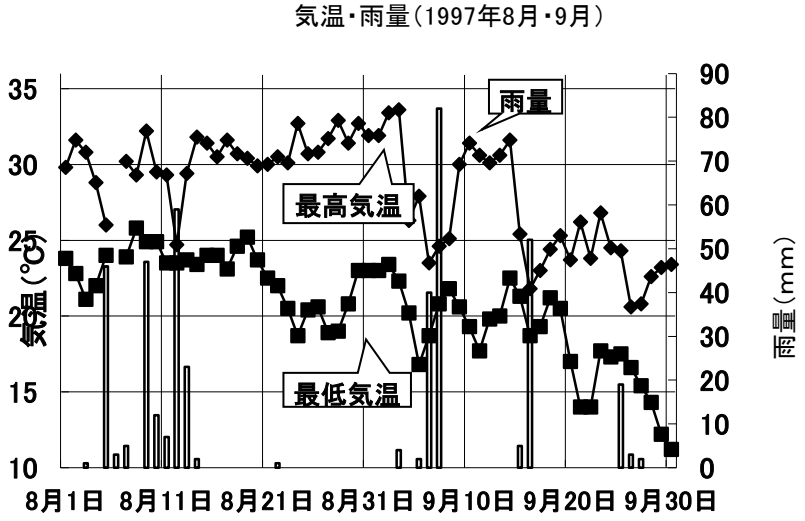
縦にある変数、横に別の変数をかいて、それぞれの項を分割集計（クロス集計）したものを分割表（クロス集計表あるいは単にクロス表という。このように 2 つあるいは 2 つ以上の変数の間の関係をみる場合、分割表を作成するのが普通です。

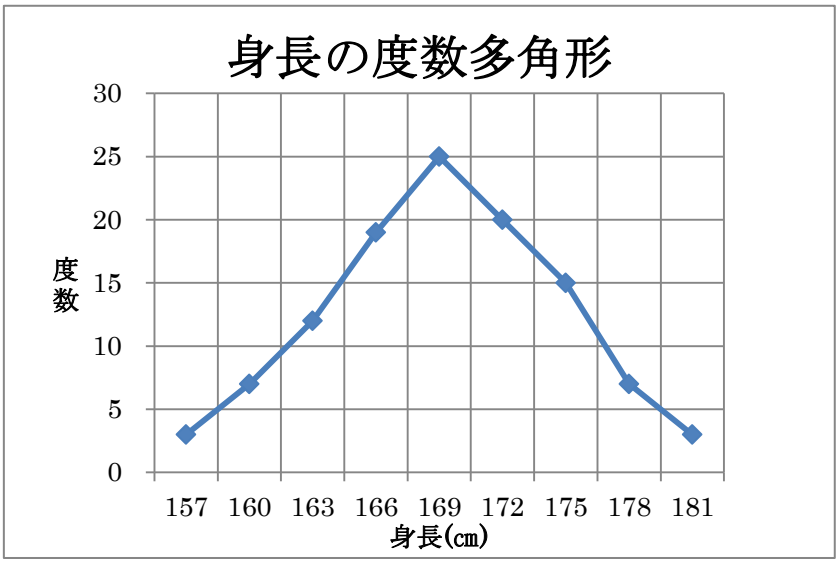
事例;ある料亭で懐石料理を食べた人のうち、吐き気・嘔吐・下痢・腹痛を訴える患者が続出した。状況はマスターテーブルの通りである。当日、懐石料理を食べた人 95 人中、55 人が酢の物を食べた。酢の物を食べた 55 人中、52 人が上記の症状を訴えた。酢の物を食べなかった人 36 のうち 3 人が上記の症状を訴えた。この場合、酢の物が食中毒の原因といえるか。

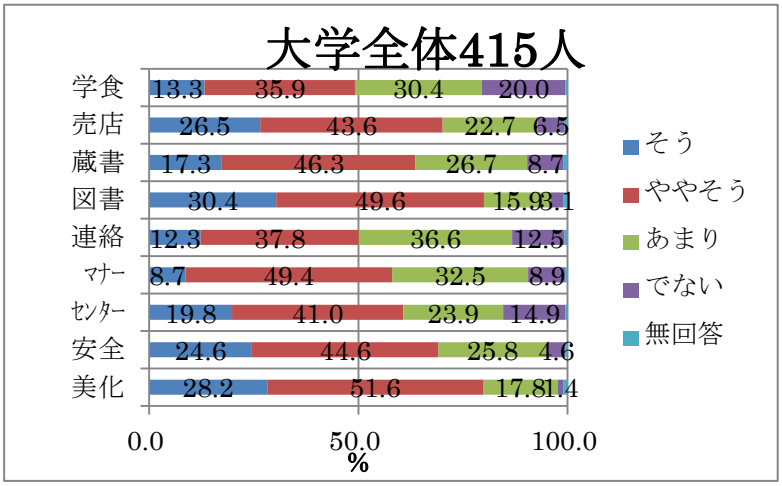
分割表の作り方）症状はすべて似ているので、症状のあり・なしを 1 つの変数、スープを飲んだ・飲まなかったで別の 1 つの変数としてクロス表を作ると、表 5 のようになる。表 5 は 2×2 分割表（四分表）といい、食中毒関係の分野では、マスターテーブルとも呼ばれる。ここでは、2 つのカテゴリー（グループ）に分けたが、症状を重症、軽症、症状なしの 3 つに分けることもできる。

<u>問題２</u>

男性の肺がん患者 100 名と、肺がんでない健常者 100 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 82 名、健常者では 60 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 74 名、健常者で 80 名であった。喫煙歴、朝食と肺がんの関係についてそれぞれクロス表を作成してください。







	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

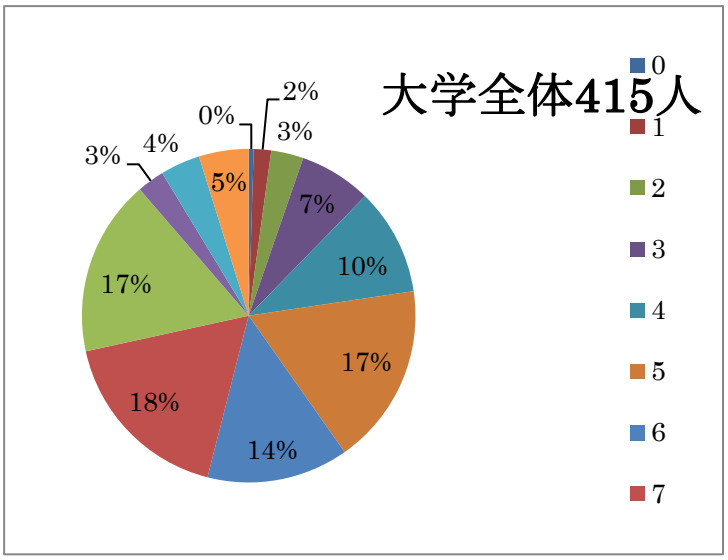


表5　入学したことの満足度

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	

２－３　図示法

表 4 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 14）

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 15）

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 15 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 16 は 1977 年の死亡者約 69 万人のうち、糖尿病との診断名で記載されたもの全てを抽出し、年齢別の百分率分布を示したもの。図 16 をみると、男女の年齢別の分布に違いのあることが分かる。

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 17 は、調査時の健康状態に関する回答の一例。一見して、数量的な違いが把握できる。

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、いくつかを並べて年次的な変化を見るときに便利である。図 18 は死因群別死亡割合の年次比較をしたもの。

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 19 に示したレーダーチャートによる表現法である。図 19 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 20 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 21 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。

図示するときの注意点：

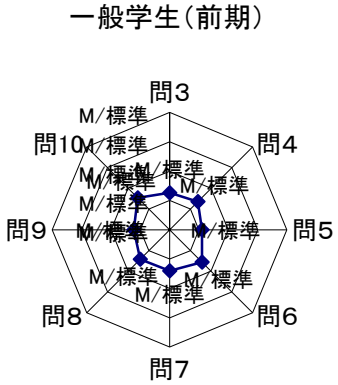
① 棒グラフは扱う資料が離散型の変数なので、0 と 1、1 と 2 の間をあけて描く。

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。

WHO 資料

図 19　レーダーチャートの一例（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））



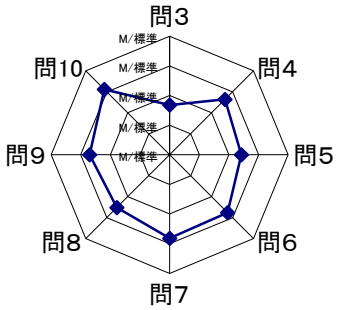
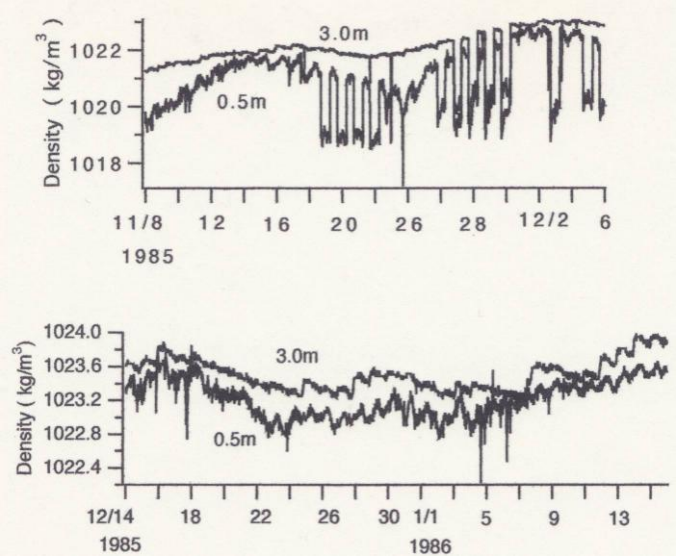
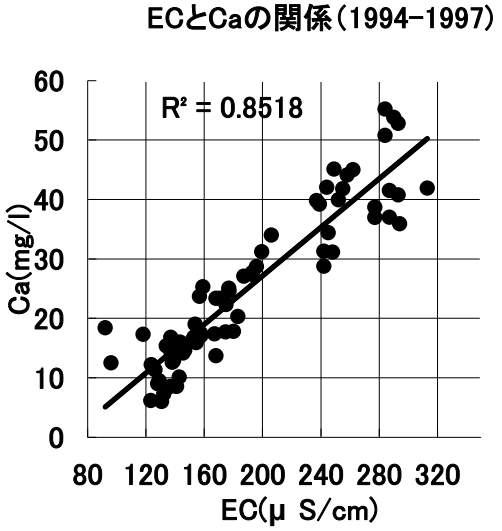


図 21　密度の時間変化（明神池 1985 年～1986 年）



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



平均値、標準偏差、偏差値

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 $x_1, x_2, x_3, ..., x_N$ 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。

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

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。

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}$$

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。

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

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。

成績表(得点)

	国語	数学	理科	社会	英語	合計
青木周三	50	35	65	75	55	280
加藤次郎	35	65	35	78	89	302
斎藤隆	85	95	90	80	91	441
田中芳子	98	100	75	83	90	446
中本真理	40	50	50	77	70	287
平野留美	45	45	45	79	45	259
松川孝二	72	72	70	81	60	355
山下久美	60	70	60	85	75	350
脇信夫	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

成績表(偏差値)

	国語	数学	理科	社会	英語	合計
青木周三	44.75	36.18	50.94	35.67	39.30	40.71
加藤次郎	37.65	49.90	33.97	43.41	59.77	43.96
斎藤隆	61.30	63.62	65.08	48.57	60.97	64.44
田中芳子	67.45	65.91	56.60	56.31	60.37	65.18
中本真理	40.02	43.04	42.46	40.83	48.33	41.75
平野留美	42.38	40.75	39.63	45.99	33.28	37.62
松川孝二	55.15	53.10	53.77	51.15	42.31	51.77
山下久美	49.47	52.19	48.11	61.46	51.34	51.03
脇信夫	51.84	45.32	59.43	66.62	54.35	53.54

２－４　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。

平均値は「真の値」のよい推定値

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$\overline{x}$＝2n(t－$\overline{x}$)となり、$f'(t) = 0$とすると、t=$\overline{x}$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ある。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。

じつは深い「平均値」

「洞察」には中央値よりも平均値を

ちなみに､「現状把握」のための統計学では平均値だけで物事を判断してはいけない、という注意がなされることもある。「現状把握」の統計学では、平均値の代わりに「中央値」「最頻値」の代表値も併せて使うといい、教えている。

「洞察」のための統計学では中央値や最頻値を気にするということはあまりない。

「代表値」をめぐる数学者たちの奮闘

データの平均値を計算して代表値とする考え方は古来から存在していたそうだが､なぜ､そうしてよいのかという根拠が数学的に定式化された歴史はほんの 200 年前から。例えばイギリスの数学者トーマス・シンプソンは 1775 年に「観測器具や感覚器官の不正確さなどが原因で生じる観測誤差を少なくするために、天文学者は普通、複数個のデータをとってその平均値を使う、という方法を用いておりますが、これは未だ一般に受け入れられてはおらず、著名な人の中にも、注意深く取られた 1 個の観測値は平均値と同じ程に信頼できる､という意見の方々もいるようです」それ以外にもボスコヴィッチ、ラプラス、ルジャンドル、ガウスなどの 18 世紀の数学者は、データのバラつきと平均値の関係についての研究に取り組んでいる。当時の科学者たちの多くは天文学に関心を寄せていたが、方位や高度など天体の位置を測定するための道具は、ちょっと手元が揺れたり視線が動いただけでも異なる測定結果が得られてしまうようなものであった。ゆえに、もし測定が完全に正確なものであれば、得られたであろう本当の天体の位置を示す､「真の値」がどこになるかを数学的に突き詰めようと考えた。

ボスコヴィッチは、バラつきを含む複数のデータから「真の値」を計算しようとすれば、データを「真の値」と「真の値からのズレ」に分けなければいけないこと、そして信頼できる「真の値」とは「ズレ」を最小化するようなものであると考えた。

天体よりも少し現実的な例として、ある建物の高さが何 m であるか、目分量で 3 人の人間に答えさせた状況を考えてみよう。ここで、1 人目は 10m だと答え、2 人目は 12m だと答え、3 人目は 13m だと答えた。この 3 人の答えからこの建物の高さを我々が推測するに当たり、誰を信じてもいいし、3 人全員を信じなくても構わない。極端な話、本当は 100m の巨大な建造物のことを全員が低めに見積もった可能性だってある。

だがそれはあまりにも不自然だ。

自然な考え方は、やはりこの建物はだいたい 12m ぐらいではないか、というところに落ち着くだろう。では真の値は 12m ではないか、という推測と、真の値は 100m ではないか、という推測のどこが違うのかというと、一番の違いは 3 人の報告に含まれるズレの大きさである。

仮に真の値が 12m であった場合、1 人目は 2m 小さく見積もり、2 人目はちょうど正確にズレが 0 の見積もりをし、3 人目は 1m 大きく見積もった。3 人合わせて 3m 分のズレが生じたことになる。一方、真の値が 100m であったとすると、1 人目は 90m、2 人目は 88m、3 人目は 87m 小さく見積もっており、合計 265m 分のズレが生じたことになる。3 人合わせて 3m のズレが生じたと考えるのか、265m ものズレが生じたと考えるのか、どちらが自然かといえば、前者だろう。すなわち、信頼できる真の値の推測値とは、その真の値を仮定したときに、得られたデータの真の値からのズレが最小となるものなのではないか、というのがボスコヴィッチの考え方。こうした彼の考え方、つまり、観測値に含まれる真の値からのズレの合計を最小にすると考えられる「信頼できる推測値」とは、じつは中央値である。

だったら、やはり中央値のほうがよい指標じゃないか、と思われるかもしれない。しかし、まず当時問題になったのが計算のための手間である。

2m 小さく見積もった場合も逆に 2m 大きく見積もった場合も同じように「2m のズレ」として考えるということは、数学的に言えば「絶対値を計算する」ということである。こうした考え方はわざわざ絶対値という言葉やそのための数学記号などを用いずとも、ふだん無意識に使いこなしているかもしれないが、数学的な処理においてじつはとても面倒な特徴を持つ。すなわちそれぞれの観測値が真の値より大きいのか小さいのかをあらかじめ場合分けして、プラスマイナスを入れ替えてやらないと数式の処理ができないため、発展的な考察や証明が困難になる。

あるいは単純にデータから中央値を探す、という体験をしてもらってもいい。順番に並べ替える作業はけっこう頭を疲れさせる。一方、足し算と割り算だけで完結する平均値なら、計算の得意な人ならすぐに終えられる。

そのため、大数学者ラプラスも 1795 年ごろまではこうした「絶対値の合計を最小化する」という考え方に基づいた推測の方法を研究していたものの、それ以降はこの課題を放棄していたそうである。絶対値を用いた考え方を行う限り、数式の展開や証明を行うにせよ、実際の推測値を求めるにせよ、あまりにも作業が煩雑になってしまうのだ。こうした「絶対値の煩雑さ」という問題は、ルジャンドルによって、あるいはガウスによって発見された最小二乗法という方法によって解決された。「あるいは」とは、この考え方を最初に公表したのは 1805 年のルジャンドルであるが、その 10 年前の 1795 年に、当時 20 歳だったガウスによっても発見されていたことが彼の数学日記からわかっているから。天才青年ガウスにとってこの最小二乗法はあまりに自明なことで、誰も

がすでに使っているものと思い込んで特に公表しようと思っていなかったらしい。最小二乗法とは簡単に言えば「絶対値のかわりに二乗を使うといい」というものである。

つまり 2m 小さく見積もった場合、すなわち「－2m のズレ」は二乗すれば「4 のズレ」になる。もちろん 2m 大きく見積もった「＋2m のズレ」も二乗すれば「4 のズレ」である。絶対値の時と同じように、元のズレがプラスであろうとマイナスであろうと必ず「ズレの二乗」は 0 以上の値となる。これを合計した値が最も小さくなるものを「真の値」として推測すればよいのではないか、というのがルジャンドルやガウスの発見である。絶対値と違って二乗の計算の式展開はめちゃくちゃ楽である。場合分けも必要なく、式を整理する方法だけなら中学生にだってできるし、高校生になればその整理された式を微分することもできる。すなわち、この「真の値」を true の頭文字で t とでもしたとき、観測値と真の値のズレの二乗の合計値が最小となる t を求めよ、という計算は、大数学者ならずとも、あるいはコンピューターがなくとも、数式を整理して微分すれば答えられる計算になるのである。ゆえに、「絶対値じゃなく、二乗する」というちょっとした発想の転換は、その後、統計学の進歩を大きく加速させることに繋がった。

なぜ、平均値は真実を捉えることができるのか？

「科学の王者」ガウスの貢献

では、なぜ平均値が中央値より優れているのだろうか？1 つには、因果関係の洞察という観点で、平均値のほうが中央値よりも関心のあることに対する直接的な答えとなっていることが多いという点。因果関係の洞察を行う関心は、多くの場合、何らかの結果を示す値の総量を最大化したい、逆に最小化したいということに向けられるが、「何らかの要因を変えれば結果の値の総量がどうなるか」ということに対して、中央値はその答えを与えない。

例えば、現状把握をするうえで平均値を使うことが不適切な例として（この文は誤解を招くので無視したほうがよい）、毎日 300 円ずつ買い物をする 8 人の顧客と 2100 円ずつ買い物をする 1 人の駄菓子屋でランダムに当たりくじを出す日、2100 円の客が 3 千円買い物をした。この状況でくじを「出している日」と「出していない日」の中央値はともに 300 円。300 円の 8 人のうち 3 人が 400 円でも、中央値は 300 円。これではくじに効果がなかった、ということになる。この状況を平均値で比較すると 1 人当たり 100 円増えて、総量としては 900 円増との結果が得られる。

要するに、それがデータの現状把握として適切だろうとなかろうと、仮にその売り上げ増が一部の極端な人間のみに集中していようと、**全体として売り上げがいくら変わるのか、という総量の増減を示すのに平均値の方が適している**。仮に「中央値が 100 円増えた」という結果が得られても、総量への影響がどうなるのかは計算できない。

また、平均値をバラつきのあるデータの背後になる「真の値」と考えるとよいのか、カール・フリードリヒ・ガウスは 1809 年の「天体運行論」で、現在の統計学でも採用

されている決定的な考え方を示した。肩こり用の磁石の性能表示にもガウスが使われているが、彼は「科学の王者」とも称された偉大な数学者かつ物理学者である。ガウスの発想が他と大きく異なって素晴らしいのは、他の数学者とは全く逆の、ゴールから遡るような問いを立てたところ。つまり、「平均値を使うことが真の値の良い推定方法となる条件とは何か」と考えて、その結果としてガウス分布あるいは正規分布と呼ばれるバラつき方の法則性に辿り着いた。そして、データのバラつき方が正規分布に従っているのであれば、最小二乗法が最も良い推定方法であり、その結果、平均値が最も良い推定値となる、という結論を得た。
正規分布とは「ふつうの広がり」のこと

2）分散と標準偏差（s²、s）

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。

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

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

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

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

＊　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 65kg、80kg、73kg、63kg、70kg 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。

標準偏差が示す「たいていのデータの範囲」

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。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。

現状把握に便利な四分位点

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り」

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、先ほども述べたように平均値±SD（正確には 1.96SD）の範囲に 95%のデータが存在する。

平均値と標準偏差を「洞察」に使ってみる

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？

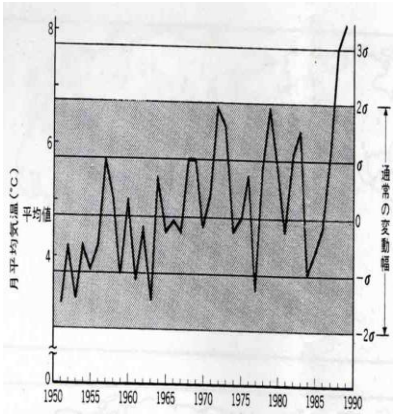
- **異常気象とは、一般に過去に経験した現象から大きく外れた現象。人が一生の**

間にまれにしか経験しない現象。

- 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

（東京の１月の平均気温）



３）その他の数値

（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　1,384 人のうち 98 人

B 地区では、対象者　8,011 人のうち 436 人　であった。

問題４

両者の百分率を式で表してください。

このように A 地区　7.1%、B 地区　5.4%となって、両者の相対的な大きさがわかります。

表 6 は健康法に関する調査結果です。回答数を百分率で示すとわかりやすい。

（2）重みづけ平均；

表 7 はある地域の集団での A、B、C それぞれの職業における 1 日 1 人あたりの食品群別摂取量です。表 7 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、油脂摂取量について求めると、

$$\bar{x}=\frac{7.4+4.8+3.2}{3}=5.1$$

しかし、被調査員がそれぞれの職業について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

$$\bar{x}=\frac{7.4\times300+4.8\times100+3.2\times50}{300+100+50}=6.4$$

A、B、C の職業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、 $n=n_1+n_2+.....+n_m$

空気の平均分子量と重みづけ平均値

空気の平均分子量

空気は色々な気体の混合物

空気の分子はない

混合気体の分子量

混合気体の構成する気体の分子量で重みつき平均で表す。

窒素 78%、酸素 21%、アルゴン 1 %（大気下層）

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、次のようになる。

平均分子量＝28×0.78＋32×0.21＋40×0.01＝28.96

問題 5

表 7 で動物性タンパク、果実について、それぞれ重みづけ平均を式で表してください。

（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 6

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

問題 7

資料が 2,5,8,6,8,10,15,8 のとき、中央値はどれですか、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）；

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

（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。

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

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。

数学Ⅰに掲載している範囲、四分位数、四分位範囲、箱ヒゲ図を掲載

（７）四分位数、四分位偏差 IQR、パーセンタイル値（百分位数）、範囲

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

最大	97
最小	10
平均	55.8
範囲	87

上限	94.0
Q3	75.0
Me	56.0
Q1	38.0
下限	12.0

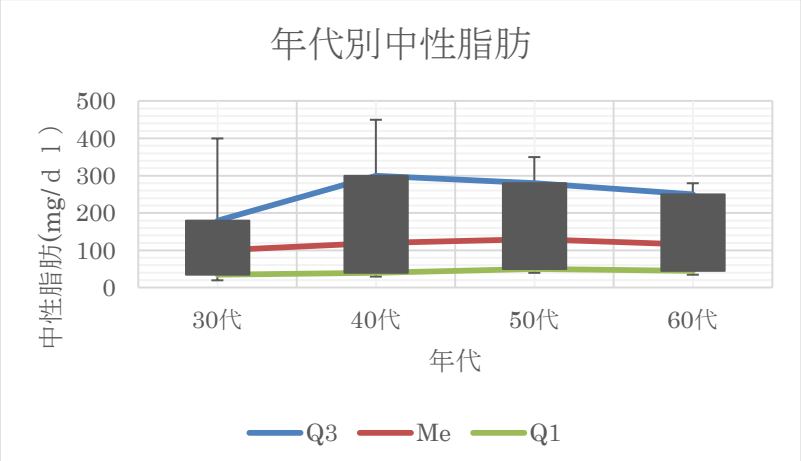
30 代、40 代、50 代、60 代の中性脂肪データのパーセンタイル値を示す。

（単位；mg/dl）

	30 代	40 代	50 代	60 代
上限	400	450	350	280
Q3	180	300	280	250
Me	100	120	130	115
Q1	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
Q3	180	300	280	250
Me	100	120	130	115
Q1	35	40	50	45
Q1－下限	15	10	10	10

	30 代	40 代	50 代	60 代
Q3	180	300	280	250
Me	100	120	130	115
Q1	35	40	50	45



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。

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 四分位数となる。

データの分布する幅が範囲（range）で範囲＝最大値－最小値となる。範囲は、すべてのデータを含むので分布の左右のすそ野、一般にはデータの少ない部分の影響を受けてしまう。そこで、中央値の前後で全体のデータ数の 50%を含む幅を使うことがある。中央値を挟んで 50%とは、第 1 四分位数と第 3 四分位数の間であり、この幅を四分位範囲（interquartile range）と呼ぶ。

四分位範囲＝（3 四分位数）－（第 1 四分位数）

また、中央値からの第 1、第 3 四分位数の偏差を求めて平均した四分位偏差（quartile deviation）も散布度の指標となる。

四分位偏差＝｛(3 四分位数）－（第 1 四分位数)｝÷2

Q－Q プロット、P－P プロット

パーセンタイル値を用いて２つの分布の形を比較する、もしくはある分布が確率分布のどれかにどれほど一致しているかを知る方法として、確率プロットの１つの方法は、２つの分布で同じパーセントに対するパーセンタイル値をそれぞれ求め、XY 平面にプロットしていく Q－Q プロットである。

また、Q－Q プロットは異なる変数どうしで作成することもできる。例えば大学入試の理解や社会科で選択科目があった場合、選択科目 A の８０点が、選択科目 B の７７点に相当するといったように、2 つの異なる科目の得点を等化(equating)する場合に、利用可能である。Q－Q プロットとは逆に同じ変数値に対するパーセンタイルをそれぞれ求めて XY 軸に作図したものを P－P プロットという。

また、データの散らばり方を比較するために平均値と標準偏差からは変動係数を導いたが、これと同様に中央値と四分位偏差から単位や値の異なる分布の散らばり方を比較するための係数を求めることができる。それが四分位偏差係数（coefficient of quartile deviation）で、

四分位偏差係数＝四分位偏差÷中央値×100 で求まる。また、ヒストグラムに対して、箱ヒゲ図がある。図 3－14 は、コレステロールのデータを箱ヒゲ図に作成したもので、箱ヒゲ図の作り方は、まず各四分位数を求め、第1四分位と第3四分位の間に箱を描く。その箱に中央値のところに線を引く。箱の左右に延びている直線がヒゲであるが、これは四分位偏差の 1.5 倍の長さを限度とし、実際にはその範囲の内側の点までを結んでいる。ヒゲより外側にあるデータは、1 個ずつ丸印で示す。

単独の分布にも使うが、箱ヒゲ図が最も有効なのは図 3－15 のように複数の分布を同時に観察したい場合である。この図は、小学 4 年生から中学 3 年生までの男児の身長の図であるが、身長の伸びの様子と、中学 1、2 年でばらつきが最大になることがわかる。

問題 8

1mの規格で作られた鉄棒の長さの平均 100cm、標準偏差 3.2cm、6 個のボールベアリングの長さの平均 2cm、標準偏差 0.029cm のとき、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。

確率分布

例えば、サイコロ 1 個を 1 回投げた時に起こることは、1 から 6 の 6 通りの目のどれからが出ます。この 1～6 の値が確率変数（この場合は特定範囲の自然数となるので、離散確率変数といいます）となります。また、体重や身長の場合には、確率変数は 65.3 kg とか 165.5 ㎝とか自然数とは限らないこともあります（この場合は、連続確率変数といいます)。

今、サイコロの目のように 1～6 の目の出る全体の現象を E で表します。そのときに確率変数は x で表現し、それぞれの目の出る現象は E_1、E_2、E_3、E_4、E_5、E_6 で表します（添え字の 1～6 が確率変数です)。そして、それぞれの目の出る確率は P（E_1)、P（E_2)、P（E_3)、P（E_4)、P（E_5)、P（E_6）となります。もちろん、この場合の確率はそれぞれ 1/6 です。

次に、もう少し複雑な場合を考えてみます。

今、サイコロを 2 個投げた場合の確率変数 x は 2～12 の 11 通りとなります。それぞれの確率は P（E_2）$=\frac{1}{36}$、P（E_3）$=\frac{2}{36}$、P（E_4）$=\frac{3}{36}$、P（E_5）$=\frac{4}{36}$、P（E_6）$=\frac{5}{36}$、P（E_7）$=\frac{6}{36}$、P（E_8）$=\frac{5}{36}$、P（E_9）$=\frac{4}{36}$、P（E_{10}）$=\frac{3}{36}$、P（E_{11}）$=\frac{2}{36}$、P（E_{12}）$=\frac{1}{36}$となります。

確率関数

このように、確率変数 x によってそれぞれの目の出る確率が決まってきます。このような場合、P（E_x）は x に対する関数 f（x）と考えることができます。これを**確率関数**といいます。P（E_x）＝f（x）で x＝2、3、・・・・、12 となります。

この特徴を式で表すと次のようになります。P（2）$=\frac{1}{36}$、P（3）$=\frac{2}{36}$、・・・・・、P（12）$=\frac{1}{36}$、　f（x）≧0、$\sum_{x=2}^{12} f(x) = 1$。

期待値

期待値は、確率変数 x と確率関数 f（x）を使って、すべての確率変数について、確率関数 f（x）を掛け合わせて合計したもの E（x）をいいます。

$$\mathrm{E(x)} = \sum_{x=2}^{12} xf(x)$$

期待値は、確率変数 x の計算上の平均値となります。この 2 つのさいころの場合は

3　正規分布と推定

3－1　正規分布

正規分布とは、ヒストグラムの形が左右対称のキレイな呼び鈴のようになる分布のこと。ただし、このような滑らかなヒストグラムを得るには、膨大な量のデータが必要です。そこで実際には、入手できたデータで正規分布の「きざし」がみえればよいとします。

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。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。

確率分布は、例えば 1 枚のコインを 50 回投げた時（これを何度も繰り返し行った時）に表の出る回数を横軸に、縦軸にその確率を取ってグラフにした場合は、以下のようなグラフになる。このような分布を確率分布と言っている。

ウェルチの方法による検定は、基本的には t 検定であるが、標準偏差（或いは標準誤差）を大きく見積もった上で、検定を行います。つまり、バラつきを大きくするということです。

自分の関心のある「真の値」の代替として、限られたデータからそれに近いものを算出する、という行為を行う背後には必ず、膨大な数の「あり得たはずの値」が存在している。その「あり得たはずの値」の分布における標準偏差が標準誤差である。一方、標準偏差は元のデータそのもののバラつきを示す指標である。複数のデータから求められた平均値のバラつき（標準誤差）は、必ず元のデータのバラつき（標準偏差）よりも小さい。求めるのに用いたデータの件数、サンプルのサイズが増えれば、増えるほど標準誤差は小さくなる。

２つのグループに差がある理由は何なのか？
例えば、がんになる人、ならない人の理由は？
喫煙と副流煙の関係とか？

差の理由を統計学で検定する。
喫煙の有無とがん患者との割合の差を検定する。
リスク認知と予防につながる。
因果関係を探る統計学。
現在のあらゆることがデータ化されている。
売れる商品と売れない商品の違いを探る。売れる商品の開発につながる。
統計学は現状把握と予測のためと見られているが、実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ある。
この情報から「ピンとくる」勘を働かせるのに、役立つ（研究面では、この情報から考察を加え、独自の理論を確立させる。）。

中心極限定理

この定理は、仮に元のデータが正規分布に従っていなくても、そのデータの値をいくつか足し合わせた値は正規分布に収束するというものである。この定理は、現代統計学の重要な礎となっている。「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。

平均値の標準誤差

平均値、元データのバラつき、データの件数と誤差の関係は、標準誤差（SE；Standard Error）で表す。（平均値の標準誤差）＝（元データの標準偏差）÷√（平均値の計算に用いたデータの件数）である。サンプルサイズとは、集団全体から抜き出されたデータの件数。例えば、75 万個の「4 人のデータから求めた平均値」の平均値は 300 万人の高校生全体の金額の平均値と一致する。この 75 万個の平均値について、分散あるいは標準偏差がどうなっているかということだが、この「4 人のデータから求めた平均値」の標準偏差のことを標準誤差と呼ぶ。

なお、標準誤差はデータから算出された平均値のみに対して存在しているわけではない。データから算出した「割合の標準誤差」はもちろん、データから算出された「標準偏差の標準誤差」というややこしいものもある。いずれにせよ、自分の関心のある「真の値」の代替として、限られたデータからそれに近いものを算出する、という行為を行う背景には、「必ずいま例として挙げた 75 万個の平均値のように、膨大な数の「あり得たはずの値」が存在している。その「あり得たはずの値」の分布における標準偏差が標準誤差である。一方、標準偏差はもとのデータそのもののバラつきを示す指標である。そして複数のデータから求められた平均値のバラつき（標準誤差）は、必ず元のデータのバラつき（標準偏差）よりも小さいものになる。また求めるのに用いたデータの件数、

すなわちサンプルサイズが増えれば増えるほど標準誤差は小さくなる。データの件数が多くなればなるほど、元のデータのうち真の平均値より大きいものだけ、あるいは逆に小さいものだけがサンプルに含まれる確率よりも、真の平均値より大きい値のものと小さい値のものが混在してくる確率のほうが大きくなる。そうするとデータの件数が増えれば増えるほど、真の平均値付近に「データの平均値」はどんどん集まってくることになる。ゆえに、データの件数が増えれば増えるほど、データから求めた平均値のバラつき（標準誤差）は、もとのデータのバラつき（標準偏差）よりも小さくなってくる。データの件数が大きくなればなるほど標準誤差が標準偏差よりも小さくなるという関係性を数学的に表現すると、

平均値の標準誤差＝（元データの標準偏差）/√（平均値の計算に用いたデータの件数）

という関係になる。

平均値と標準偏差を使えば「サンプルサイズ設計」ができる

これまでのデータから求めた平均値と標準偏差を用いて「次の調査でどれくらいの標準誤差にするためにどれくらいのデータの件数（すなわちサンプルサイズ）が必要か」という見積もりを行うことができる。このようなデータの件数の見積もりのことは専門用語でサンプルサイズ設計と呼ぶ。

先ほどの関係式に「標準偏差が 1 千円」という情報を加え、横軸が 1 地域ごとのサンプルサイズ、縦軸がそこから求めた平均値の標準誤差のグラフができる。サンプル数が 4 だと、標準誤差は 500 円、サンプルサイズが 100 人だと、SE は 100 円、サンプルサイズが 2500 人ずつのサンプルが得ることができれば、SE は 20 円となる。

平均値が 4 千円、SE が 100 円の結果で、平均値±2SE の範囲「だいたい 3800 円～4200 円」。平均値が約 4 千円という結果がどこかの地域で得られたとして、それが数十円の単位まで正確、という必要はおそらくない、各地域で 2500 人は大げさすぎ、一方、たった 4 人で標準誤差が 500 円という状況では「平均 4 千円」の結果が得られたとしても、それは「平均予算が 3 千円～5 千円という範囲となり、あまり参考にならない。

このように最終的に得られるであろう誤差と、調査にかかる手間や予算を天秤にかけて、必要なデータの数を見積もるのがサンプルサイズ設計である。こうしたサンプルサイズ設計の考え方が理解できていれば、「とりあえず全数調査」とか「とりあえずビックデータ」といった考え方が適切でない状況が分かる。

割合についての標準誤差

割合の標準誤差とデータの件数の関係は

$$\text{割合の標準誤差}=\sqrt{\frac{\text{割合}\times\ (1-\text{割合})}{\text{データの件数}}}$$

である。例えば、100 件のデータから割合が 90%と算出されれば、この標準誤差は 0.9×0.1÷100 のルート、0.03、つまり 3%となる。これも、割合とはある状態を取る（1）

かとらない（０）かを示すデータの平均値である。よって、平均値の標準誤差とまったく同じ。単に「１か０かを示すデータの分散」をシンプルな式で示すと　データの分散＝割合×（１－割合）となる。したがって、

$$\text{標準偏差} \div \sqrt{\text{データの件数}}$$

という平均値の時の関係式と同じ意味となる。
ちなみに、「データから得られた平均値±2SE のことを平均値の 95%信頼区間と呼ぶこともある。正確には平均値±1.96SE である。

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、－∞～＋∞　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%

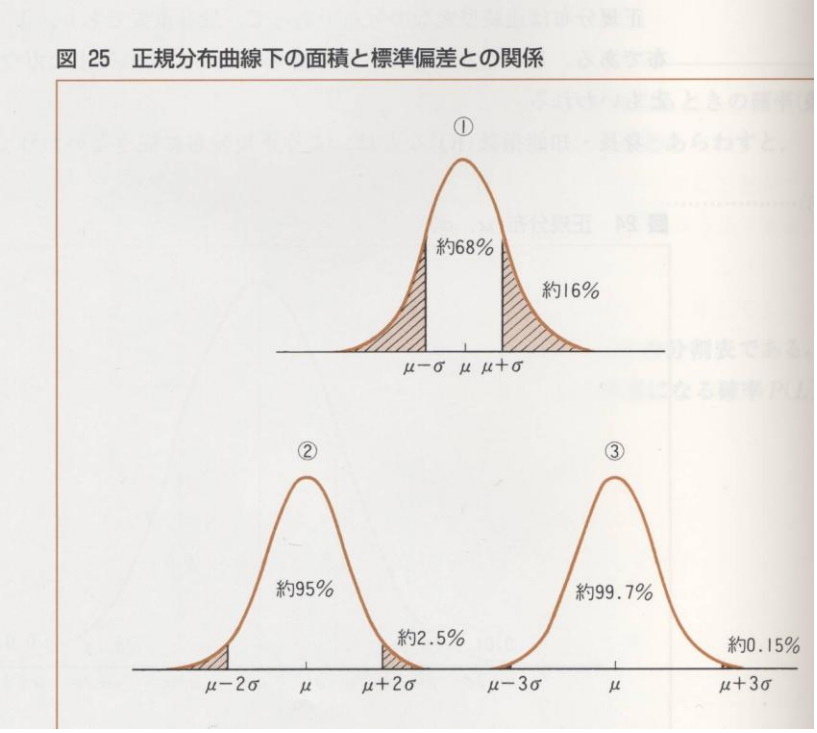
$\mu - 2\sigma$ と $\mu + 2\sigma$ の間の正規分布曲線の面積は全体の約 95%

$\mu - 3\sigma$ 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（μ、σ^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 25 から、171cm 以上は③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。



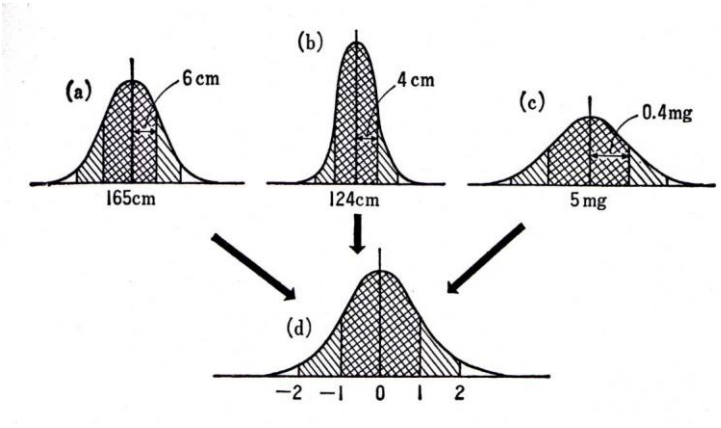
２）正規分布の標準化

先程の例で平均値よりも大きく、175 ㎝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。



（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

（a）の分布で 180 ㎝は（d）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 ＝156ｃｍ、σ ＝5ｃｍの身長分布で 169ｃｍ以上の人の割合を求めたいとします。

この場合まず、169ｃｍに対応する z を求めます。

z ＝（169－156）/5＝2.6

付表 4 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 4 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 と同じことなります。

（この部分に、付表の見方を記した部分を記載してください。）

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。

$$\frac{y_i - 50}{10} = \frac{x_i - \bar{x}}{s}$$

$$y_i = \frac{10}{s}(x_i - \bar{x}) + 50$$

ここで、 s 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、

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

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・モアブルは見つけた「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よい。

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。

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。

正規分布以外の確率分布

二項分布、ポアソン分布、指数関数型分布、カイ 2 乗分布（χ^2分布)、ベイズの定理

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 ）s，σ
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)｝＝√｛(データの平方和）/（N－1）－N×（平均値）^{2}/（N－1)｝＝不偏推定値（unbiased estimate）

または、SD＝√（偏差平方和/N）＝√｛データの平方和/N－（平均値）^{2}｝

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。

「頻度分布、正規分布、正規分布のあてはめ、
平均μに対する最良の推定値は、で与えられる x バーであり、
標準偏差σの最良の推定値としては、　　から s 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。」

<u>平均の分布での標準偏差は標準誤差となる！</u>

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が、これはχ^2分布に関して既に用いた対応する限界P＝0.05とほぼ同等である。**

「・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。

・標本平均の差の分布―　$N(\mu_1-\mu_2、\ \sigma_1^2/n_1+\sigma_2^2/n_2)$

平均の差の推定値「$\mu_1-\mu_2$」が平均の差の標準誤差$\sqrt{(\sigma_1^2/n_1+\sigma_2^2/n_2)}$の何倍になるかを考える。

・比率（割合）の検定；

$z=(p-P_0)/\sqrt{(pq/n)}$、2 群では、$p_1=r_1/n_1$ と $p_2=r_2/n_2$ の平均の割合として

$p=(r_1+r_2)/(n_1+n_2)$ を利用　$z=|p_1-p_2|/\sqrt{(p(1-p)(1/n_1+1/n_2))}$