◎ はじめに

自然科学の多くの分野は実証主義に基づいており、実験結果や観察結果を解釈し、主観を排除して結論が導かれる。しかし、数値１つを得るための測定でさえ、その結果は誤差によりばらついてしまう。そのようなばらつきに対しては、例えば平均をとるとか、最小二乗法を用いるなどにより対処をすることがある。では、そのようにして得られた数値はどこまで信頼してよいのだろうか。そのような問いかけに対する答えは、数値の精度や誤差の伝播の概念を理解することで得られるだろう。しかし、もう少し複雑な、かつあいまいな実験結果があった場合にはどうするべきであろうか。主観を排除しなければいけないのだから、その実験結果が意味することが何々であると「思われる」では望ましくない。誰が判断しても同じ結論を導くことができるための道具が必要であり、その道具を用いて導かれた結論が正しく記述され、その記述に対面した誰もが同じようにその背景にある意味合いを理解できるようでないといけない。ここでまとめた仮説検定の考え方は、100 % の正しさで線引きができないような実験結果に対し、どこに線を引くべきであるかの基準の定め方を示す道具である。この道具の性質と、背景にある統計分布などについてまとめた。はじめは数ページのハンドアウトにまとめるつもりだったが、再学習しながら書き進めるうちに、書き加えるべき内容が次々とでてきて、最終的にまとまるのだろうかと、時に暗澹たる先行きを案じざるを得なかった。なんとかまとめることができたが、自分で分かりにくいと感じたところは、重複も承知の上で何度も強調した。おそらく数学的な立場から見ると緻密ではないとお叱りも受けるだろう。とはいえ、嘘や誤りがないように注意したつもりである。もし誤りがあれば、それは引用先ではなく、著者に責任がある。

◎ 索引

仮説検定と有意差　2
　仮説検定とは　3
　具体的な手順　3
　危険率と過誤　5
　片側検定と両側検定　6
　標本の大きさと p 値、検定力　7
　非有意と差なし　9
　基準率の錯誤　9
　p 値偏重に対する批判　9
　多重比較と第１種の過誤の増加　12
　ではどうすればよいのか　12
確率分布と統計　13
　正規分布を仮定する場合　14
　標本平均の期待値　14
　分散と不偏分散　14
　標準誤差 SEM　15
　母集団、標本、標本平均　16
　信頼区間　17
　エラーバーの使い分け　18
　信頼区間の重なりの有無　19
仮説検定の話、ふたたび　20
　標本の大きさと検定力　21
正規分布以外の分布　22
　二項分布　22
　ポアソン分布　23
　指数分布と一次反応速度式　24
　student の ｔ分布と ｔ検定　25
　t-分布における自由度の決め方　26
　最小二乗法と、a、b 値の検定　26
標本間の比較と種々の検定　28
　二標本検定　28
　χ^2（カイ二乗）分布　29
　パラメトリックな χ^2 検定　30
　F 分布　32
　F 分布に従う１つめの指標値　33
　F 分布に従う２つめの指標値　34
　F 分布に従う３つ目の指標値　35
　分散分析（ANOVA）の考え方　36
ピアソンの χ^2 検定　38
探索的データ解析と箱ひげ図　40
分析に入る前の場合分けのまとめ　41
エクセルによる数値の解析の例　42
　最小二乗法　44
図の作成に用いたプログラムソース　45

◎ 仮説検定と有意差

サイコロを振って、1 の目の出現回数を数える実験を行ったとき、実験結果として次の２通りの結果が得られたとする。

結果１： 12 回の試行の結果、1 が　5 回出現した。

結果２： 60 回の試行の結果、1 が 25 回出現した。

いずれも試行の結果から算出される 1 の出現確率は 5/12 であり、サイコロの出目に偏りがないと考えたときの期待値 2/12 との間に開きがある。この開きが、偶然の誤差、つまり単にばらつきによるものなのか、それともなにか必然的なものであるのかを判断しなくてはならないことがある。必然的とは、そこに偶然以外の何等かの原因があるという意味である。たとえばサイコロであれば重心に偏りがあって、特定の目が出やすい（確率が均等なサイコロではない）などの理由が考えられる。

12 回の試行の結果 1 の目が 2 回出現したのなら、サイコロが正しいということを積極的に疑う理由はない。もちろん、サイコロが正しくないのにそのような結果になったとしてもおかしくない。1 の出目回数が 3 回ならば、偶然そうなったとしても不思議はないと思うだろう。では 12 回の試行の結果 12 回とも 1 の目がでたら。用いたサイコロが正しかったという前提が限りなく怪しい気がするが、かといって、このサイコロの出自を知っていた（真の出目の確率を予め知っていた）のでなければ「このサイコロは公正であったが、この結果は偶然に生じただけである」ということを否定することはできない。しかし、だからといって、なんでも「あり得る、否定できない」と結論する（これは正しい）だけでは、そのサイコロの正しさについて、何も判断できない。なので、3 回ならあり得るのか、4 回なのか、…、10 回なのか、11 回なのか、12 回でもあり得るのか、あるいはあり得ないのか、期待される値からのずれ・偏りに対し、どこまでがあり得ることで、どこからがあり得えないことなのか、多くの人が納得するレベルで線を引いてやる必要がある。これをどこに引くのかが問題である。

この線は、どこに引いても、100 % の正しさにはならない。つまり、上記の実験結果だけから、このサイコロの出目に偏りがあるのか、ないのか（つまり、このあと無限回の試行を行った場合に、それぞれの出目の回数の割合が 1/6 に収束していくのかどうか）を、100 % の正しさで結論することはできない。100 % の完全性を保証するためには全数検査しかないから、現実的には不可能であったり無意味だったりする。しかしながら、あえて 100 % の正しさではないことを承知の上で、この実験で得られた出目の偏りに、意味があるのか無いのか（「有意である」かどうか）を判定することができる。そのために、ある大きさ（size）の標本（sample）を母集団（population）の中から抜き出して検査をし、その結果に基づいて母集団の性質を判断することができる※。この方法を統計検定または、統計的仮説検定、仮説検定などという※2。（「統計検定」は資格の名称でもある。）

※ 母集団とは調査の対象となる数値などについての全体であり、ここでは、当該のサイコロを過去と未来も含めて無限回試行した結果である。標本は、母集団の部分集合である。標本の大きさとは、標本を構成する要素の数であり、ここでは今回試行した回数である。ここで注目している標本や母集団がもつ性質とは、例えば、平均（mean）や分散（variance）である。特に、標本や母集団が正規分布（normal distribution）に従う場合には、この２つの性質を記述するだけでよい。

※2 統計学は、記述統計学と推計統計学（inferential statistics）に大別される。大雑把な言い方になるが、母集団全体を調査対象とし、その分布の仕方などを分析していくのが記述統計学、母集団から標本を無作為に抜き出して、標本の性質に基づいて母集団の性質を推定していくのが推計統計学である。さらに、「原因に対してある確率で結果が生じる」という概念に対し、「この結果があるからには、原因としてある事象があった確率はどの程度であるか」という「逆確率」（原因の確率）（条件付きの確率としても考えられる）や「事後確率」の考え方と、「ベイズの定理」を基にした統計学の体系として「ベイズ統計」もある。ベイズ統計についての詳細は、ここでは扱わない。

○ 仮説検定とは

この仮説検定という手法で判定したいと思っているのは、サイコロの出目の確率が均等であるか、薬の効果の有無、あるいは、ある集団のばらつきのある値が別の集団と異なるかどうか、などである。実験で有限の回数により測定できるのは、サイコロを振った結果の出目の回数、薬の効果があると思われる結果となった回数、一部抽出された標本が示す値などである。つまり、そこに一定の割合や傾向はあっても、他の要因や偶然のばらつきによりいつも同じ結果を与えるとは限らないような事象である。１回の実験を行うだけで真の値がわかってしまう場合には、仮説検定は必要ない。

検証しようとしている作業仮説が直接的に証明できないとき、この仮説に対し、帰無仮説を立て、得られたデータから、帰無仮説が成り立つかどうかを検証することを検定という。証明しようと思っている作業仮説は、（帰無仮説に対する）対立仮説と呼ばれる。

H_1、対立仮説　　積極的に証明したい仮説。実験仮説。作業仮説。検証仮説。
例）差がある、効果がある、相関がある。
H_0、帰無仮説　　主張すべき仮説の逆。
例）差がない、効果がない、相関がない。

帰無仮説が棄却される（間違っていると結論される）なら、立証しようとしていた仮説が正しいことがわかる。つまり「差がないということはない」ならば、「差がある」と言ってよい。ある線引きで帰無仮説を棄却し、100 % の確信をもって言うことはできないが、差がある（ないわけではない）と認めることが妥当であると結論することを「有意差（significant difference）がある」「有意に差が認められる」「差が有意である」などと表現する。

○ 具体的な手順

少し具体的な話にしてみよう。まず２つの母集団を考える。一つは出目確率の均等なサイコロを振ったときの試行結果である。これを無限回繰り返したものを母集団 A としよう。そして、有限回の試行結果は、この母集団 A から抜き出された標本 A である。もう一つの母集団は、いま目の前にあり、実験に供しているサイコロを振ったときの試行結果である。これを無限回繰り返したものを母集団 B としよう。そして、有限回の試行結果は同様に標本 B である。母集団 A、母集団 B はそれぞれの平均値として母平均 A（$\mu_{0A} = 1/6$）、母平均 B（μ_{0B}）を持っているが、無限回の試行を行うわけにいかないから一般的には※母平均を直接知ることはできない。標本 A や標本 B から推定することになる。

※ ビッグデータを扱うことができるようになってきているため、これまで無限に近いとみなされてきたような有限サイズの母集団については、母平均や母分散の値を直接的に算出できる場合もある。

今回の実験に先立って「このサイコロは、出目の確率が均等ではなく本質的に 1 の目が出やすい」という対立仮説 H_1 を立てていたことにしよう。つまり H_1 は、「母集団 B の母平均 μ_{0B} は 1/6 より大きい」、または、「 $\mu_{0B} > \mu_{0A}$ 」などと書くことができる。これに対する帰無仮説 H_0 は $\mu_{0B} = \mu_{0A}$ である。上記の実験結果はこの仮説 H_1 を裏付けるものと言えるのだろうか。

H_1 このサイコロは出目が偏っており、事前に定めた特定の出目が 1/6 の確率より高く出現する。
H_0 このサイコロは出目の偏りがない。事前に定めた特定の出目が 1/6 の確率で出現する。

まずはじめに行うのは、帰無仮説※ H_0（つまり $\mu_{0B} = \mu_{0A} = 1/6$ ）を仮定したとき、何の理由もないの

に、偶然の結果として「上記に示したような偏った実験結果、および、更に極端な実験結果」を生じる確率を計算することである。この値を p 値という（p は probability の頭文字）。

※ 対立仮説 $\mu_{0B} > \mu_{0A}$ に対する対偶は、$\mu_{0B} = \mu_{0A}$ ではないから、$\mu_{0B} < \mu_{0A}$ の場合も考慮する必要があるように思われるだろう。しかし、μ_{0B} が μ_{0A} より小さい値をもつ場合には、「更に極端な実験結果」の極端の方向性が逆になる。つまり、いまある出目の回数が期待値よりも多いこと（のみ）に注目して、偏りのない確率では生じにくいことであると言おうとしているのに対し、$\mu_{0B} < \mu_{0A}$ の場合は、逆に出目の回数が少ない方向に偏ることになる。検定によって明らかにしようとするのが $\mu_{0B} > \mu_{0A}$ だけであるなら、このような偏りには無関心なので、p 値の計算のために $\mu_{0B} = \mu_{0A}$ を帰無仮説として用いて問題ない。ただし、対立仮説 H_1 の選び方（$\mu_{0B} > \mu_{0A}$ とすべきであるのか、あるいは、$\mu_{0B} \neq \mu_{0A}$ とすべきであるのか）については、片側検定と両側検定の項で後述する内容も参照のこと。

ここで、先に p 値の性質をまとめておく。以下の意味は、先を読んでもらえればわかるだろう。
・p 値とは、今回の検定の対象となるデータがどのくらい極端なデータであるのかを示す。本来の性質として「差がない、効果がない」のに、そのような極端なデータを与える確率で表される。
・比較しようとしている対照群との間に「有意に差がある」という結論に至った場合でも、p 値から、本来の性質に「どの程度の確からしさで」「どの程度の量の」差があるのかは決まらない。

帰無仮説の下、偶然により結果１のようになる確率を p_1 としよう。帰無仮説は「出目の偏りがない」のであるから、結果 1「出目が 5 回であった」は「出目が 5 回以上であった」と解釈し、その確率を求める。これは、更に極端な結果である「6 回の出現」や「７回の出現」であっても、同様に帰無仮説に反して「偏りがあった」結果と考えるべきだからである。以下に、その確率を計算式とともに求めておく。また同様に、結果２となる確率（25 回以上となる確率）を p_2 とし、計算結果を示しておく。

$$
\begin{aligned}
p_1 = {} & (1/6)^5 \times (5/6)^7 \times 12!/7!/5! \\
& + (1/6)^6 \times (5/6)^6 \times 12!/6!/6! \\
& + (1/6)^7 \times (5/6)^5 \times 12!/5!/7! \\
& + (1/6)^8 \times (5/6)^4 \times 12!/4!/8! \\
& + (1/6)^9 \times (5/6)^3 \times 12!/3!/9! \\
& + (1/6)^{10} \times (5/6)^2 \times 12!/2!/10! \\
& + (1/6)^{11} \times (5/6)^1 \times 12!/1!/11! \\
& + (1/6)^{12} \times (5/6)^0 \times 12!/0!/12! \\
= {} & 0.03635
\end{aligned}
$$

$$p_2 = 0.0000042$$

このような計算が可能なのは、試行回数が少ないからである。確率を直接計算できない場合には、統計的手法で相当する量を決めることになる。その方法は後述する。

この p 値を事前に定めた有意水準（α を用いることがある）と比較する。ただし、
・p 値（帰無仮説 $\mu_{0B} = \mu_{0A}$ の下で、標本 B の平均値 μ_B が実験で得られた値になるような確率）が有意水準 α よりも小さいならば、帰無仮説は棄却され、$\mu_{0B} = \mu_{0A}$ ではないだろうと結論される。
・有意水準 α とは、p 値と比較するための、いわば閾値である。p 値が α より大きいか小さいかだけを判断するための基準である。p 値が α と近いとか離れているとか考える必要は一切ない。
・有意水準は「事前に」定めるものである。また、どの値を用いたのかを明示しなければならない。
・有意水準には、慣例的に 0.05、0.01、0.001 のいずれか（分野により若干異なる）を用いる。ただし、これらのうち、あるいは他のどの数値を用いるのかは、目的や実験の性質、過誤にあたってのリスクの大きさ等を考慮して決めるべきものである。

・同時に複数の有意水準を用いて表現してもよい。
・帰無仮説を前提とした母集団の確率分布に対し、有意水準以下の確率となるような極端なデータの領域（すなわち、帰無仮説を棄却できるような領域）を、棄却域と言う。
・正規分布に従う母集団があったとき、片側検定における 0.05 の棄却域の閾値は、およそ 1.65 σ に相当する。閾値 2σ 以上では確率 0.023 程度、3σ以上 では 0.0014 程度である。両側検定を有意水準 0.05 で行った場合、棄却域を両極端に設け、確率をそれぞれ 0.025 とするから、およそ「平均値 ±2σ 」より外側となる。有意水準 0.01 の場合は、棄却域は 平均値±2.57σ より外側である。

→ 事前に定めた有意水準が 0.05 の場合。
この結果１は、H_0 を仮定したとき p_1 値 0.036 は有意水準より小さいので H_0（$\mu_{0B} = \mu_{0A}$）は棄却される。従って「H_1 は有意に正しい」（$\mu_{0B} > \mu_{0A}$）、このサイコロは「有意に 1 の目が出やすい」と結論される。偶然のばらつきでは 3.6 % しか生じない事象は滅多に起きるものではない（ = 有意水準より小さい）から、その事象が生じたのであれば、何か別の原因がある可能性を検討する意味があると判断できるという意味である。

→ 事前に定めた有意水準が 0.01 の場合。
この結果１は、H_0 を仮定したとき p 値 0.036 は有意水準より大きいので H_0 は棄却できない。つまりこの実験結果は「有意に 1 の目が出やすいとは言えない」と結論される。ただしこれは、「偏っていない」という結論ではない（つまり、偏っているかもしれないし、偏っていないかもしれない）ことに注意すること。偶然のばらつきで 3.6 % も生じる事象は、十分に想定される範囲内である（ = 有意水準の 1 % より大きい）のだから、ばらつきによりその事象が生じたとしても何の不思議もなく当たり前のこととして受け入れ、別の原因がある可能性を検討することを放棄するという意味である。

○ 危険率と過誤

上記の手続きに従って、対立仮説を検証しようとすると、必ずある確率で誤った結論に至る。これを過誤という。（ここで「過誤」「エラー」とは、手順の間違いなど避けることのできるものではなく、検定という手法に織り込み済みで、一定の割合で必ず生じるものである。）

第１種の過誤：帰無仮説が正しいのに棄却してしまう。
本当は効果がない（真の陰性）のに、陽性であると判断する。これを偽陽性という。

第２種の過誤：帰無仮説が誤っているのに、棄却しない。
本当は効果がある（真の陽性）のに、陽性とはいえない（陰性である）と判断する。これを偽陰性という。

有意水準は危険率とも呼ばれる。同一の実験を繰り返したとき、全実験数に対し、必ず有意水準以下※の割合で、誤った結論（第１種の過誤）を得るからである。つまり、前提として「確率の均等なサイコロ」を用いても、12 回の試行中、事前に定めた特定の出目が 5 回以上出現する確率は、3.6 % である。つまり 28 人のクラス内で全員が 12 回の試行に対し 5 % の有意水準で片側検定を行った場合、期待値として 27 人は「サイコロの出目には偏りがあるとは言えない」と結論し、期待値として 1 人は「サイコロに偏りがあり、有意に 1 の目がでやすい」と結論することになる。しかし、前提（真の陰性）に矛盾する結論（偽陽性）を得た最後の 1 人も、なにか間違ったことをしたわけではない。

※ サイコロ 12 回の試行中の特定の出目の出現確率は、4 回以上 12.5 %、5 回以上 3.6 % なので、ぴったり 5 % を与える回数は存在しない。このため、この検定において第１種の過誤を与えるのが有意水準「以下」の割合（3.6 %）となっている。連続的な値を扱う場合や、連続的であるとみなせるほど標

本が大きい場合には、有意水準（危険率）と等しい割合で第１種の過誤を与えることになる。

このように、帰無仮説 H_0 が実際には正しいにもかかわらず、危険率以下の低い確率で生じた偶然の結果を根拠として、この帰無仮説 H_0 を棄却してしまう場合、このような誤りが第１種の過誤である。第１種の過誤を小さくするために、有意水準（危険率）の数値を下げることができる。しかし、有意水準を下げるだけでは、次に述べる第２種の過誤の可能性が高くなる。

いま、実際には帰無仮説が誤っており、サイコロが正しくなく、その出目に偏りがあった（陽性）としよう。例えば、12 回の試行中に事前に定めた出目が 5 回も出現した（正しいサイコロであった場合の期待値の 2 回より多かった）のが、必然的であったとしよう※。帰無仮説の下でも同じ結果を得る確率が 3.6 % あるため、有意水準を 0.01 としているために、結果１に基づき帰無仮説を棄却できないでいる場合、このような誤りを第２種の過誤という。つまり、慎重になりすぎて、原因があって生じているようなこと（陽性）に対しても偶然のかたよりによる結果かもしれない、陽性とは断定できない（偽陰性）と結論してしまうことである。第２種の過誤の生じる確率を β で表すことがある。

※ 帰無仮説が誤っていることは、今回の標本から算出された確率 $\mu_B = 5/12$ がこのサイコロにおける真の確率であること、すなわち $\mu_{0B} = \mu_B$ を意味しているわけではない。μ_{0B} が「厳密に 1/6 ではない」だけで、ほぼ 1/6 である場合でも、やはり帰無仮説は正しくない。1/6 ではない μ_{0B} の値がいくらであるのかによって変化するが、いずれにしても第２種の過誤が生じる確率は 0 ではない。

結果２は、有意水準として 0.01 を用いても、0.001 を用いても、p_2 値はそれより小さい。すなわち、帰無仮説 H_0 はこれらの有意水準の下に棄却され、H_1「サイコロの出目には偏りがある」が有意に示されたという。とはいえ、4 ppm ほどの確率で、第１種の過誤を犯している可能性が残る。すなわち、正しいサイコロでも 4 ppm ほどの確率で、このような実験結果２を与える（偽陽性）。この確率はごくわずかのように見えて、宝くじの高額当選よりずっと高い（100 組の組番号と 6 桁の数字の下 5 桁で１位当選が決まる場合、0.1 ppm（1000 万分の 1）である）。これを無視してもよいかどうかは、その対象のデータの性質や、過誤にあたっての影響の大小等を踏まえ、別に議論されるべきである。

○ 片側検定と両側検定

対立仮説がなにを主張しようとしているのかの関心事に、方向性を伴う場合（ある数値より大きい、など）には片側検定を、方向性がない場合（単に差があるかどうか、など）には両側検定を用いる。片側検定をする方が、ある極端側における棄却域は広くなるので、有意差が認められやすくなる。だからといって、本来両側検定をするべきものに対し、有意差を見出すための目的で片側検定をするといったことは避けなければならない。片側検定が認められるのは、事前知識などに基づき偏りがある一方向にしか生じないことが確定（推定ではない）している場合、および、実験結果に逆の偏りが生じることを許容しており区別して拾い上げる必要がない場合に限られる。片側検定と両側検定で、検定結果に以下に述べるような違いがあるので、いずれを用いたのかを明記することが必要である。

「サイコロの特定の出目確率が 1/6 より大きい（$\mu_{0B} > \mu_{0A} = 1/6$）」かどうかを検証する場合は上記のように片側検定を行い、棄却域（すなわち帰無仮説を棄却できるようなデータの範囲）を累計で 5 % までの確率を与える n 回以上の出目回数とした。しかし、「出目が偏っていて、特定の出目が 1/6 より出やすいかもしれないし、1/6 より出にくいかもしれない（$\mu_{0B} \neq \mu_{0A}$）」ということを検証する場合には、出目回数が期待値より少ない場合でも、やはり偏りがあったと言わなければならない。この両側検定の場合は、棄却域（第１種の過誤を生じる領域でもある）を両側に二分して置くので、有意水準を 0.05 とする場合、閾値を累計確率が 2.5 % までとなる出目回数としなければならない。

確率の均等なサイコロを用いた場合、60 回試行（期待値は 10 回）における確率※を計算してみたところ、事前に定めた特定の出目が 4 回以下しか出現しない確率は 2.0 %、5 回以下は 5.1 %となった。また、15 回以上で 6.5 %、16 回以上 3.4 %、17 回以上 1.6 % となった。

※ 以上の計算結果は、あくまでも事前に定めた目がそのような回数になる確率である。60 回の試行ののちに、その回数以下、または以上の出目が見つかる（そのような出目が 1 から 6 のどれでもよい）確率ではない。そのような場合は、上記よりももっと数値が大きく、およそ 6 倍となる。ただし、各目の出現は独立ではない（1 の目がでたときに、同時に他の目が出る可能性はない）ので、詳細な計算はここでは行わないが、正確には単純に 6 倍するだけではない。

つまり、H_1「事前に定めた特定の出目が 1/6 より出やすい（$\mu_{0B} > \mu_{0A}$）」ことを検証しようとして片側検定を行った場合、60 回の試行に対し、15 回の出現では帰無仮説が棄却できないので「その目が有意に出やすいとは言えない」と結論し、16 回以上の出現ではじめて「その目が出やすいことが有意に示された」と結論することになる。その場合、その目が 1 回も出現しなかったとしても上記結論に変更はなく「その目が有意に出やすいとは言えない」としか述べることができない。また、H_1「事前に定めた特定の出目に偏りがある（$\mu_{0B} \neq \mu_{0A}$）」ことを検証しようとする場合は、両側検定を行う。その場合、同じ有意水準 0.05 を用いても、片側検定では帰無仮説を棄却できた 16 回の出目回数（16 回以上では $p = 3.4\ \% > 2.5\ \%$）では帰無仮説を棄却できない。17 回以上の出目回数、または 4 回以下の出目回数の際に、帰無仮説が棄却されるため出目に偏りがあることが有意であるということになる。

12 回試行の場合、出現回数が 0 回の確率が、すでに 11.2 % あり H_0 を棄却できない。他方、4 回以上 12.5 %、5 回以上 3.6 %、6 回以上で 0.8 % となる。従って、有意水準 0.05 で両側検定を行って帰無仮説 H_0 を棄却できるのは、その事前に定めた特定の出目の出現回数が 6 回以上の時のみである。

○ 標本の大きさと p 値、検定力

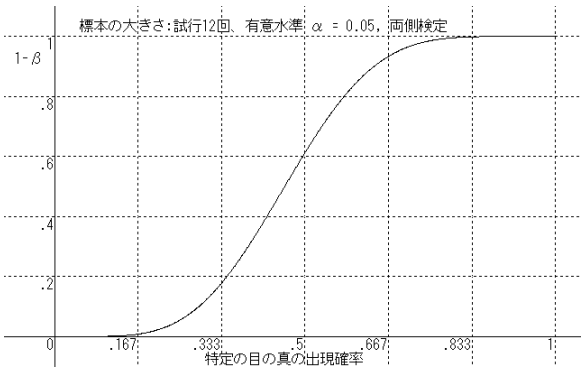
上でも例示したように、標本の大きさが大きい（母集団から抜き出した標本の要素の数が多い）ほど、有意差は検出されやすくなる。つまり、同じ割合の出現回数に対し、標本の大きさが大きいほど p 値は小さくなり、有意水準が同じであっても、帰無仮説を棄却しやすい。

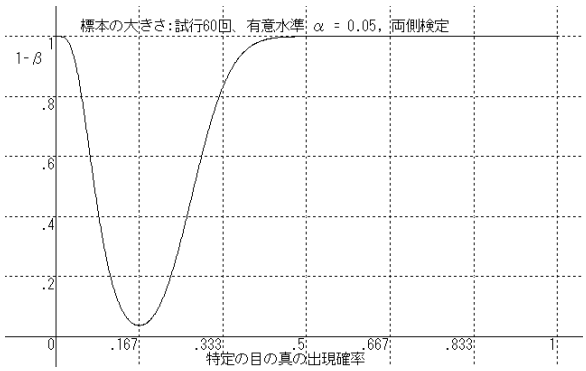
第２種の過誤についてあらためて考えよう。第２種の過誤は、帰無仮説が誤っているのにこれを棄却できないものである。すなわち、サイコロの特定の目の真の出目確率が 1/6 以外のあらゆる数値のうちどれかであったとしたときに、偶然の結果として有意水準内に収まってしまったため、有意差があるとは認められないという場合である。このような過誤を生じる確率 β を 1 から引いたものを検定力という。すなわち、検定力とは、帰無仮説が誤っているときに正しく棄却できる確率である。

検定力は、有意水準、標本の大きさ、および効果量に依存して決まる。標本が大きいほど、効果量が大きいほど、検定力は高くなる。ここで、効果量とは、帰無仮説が成立するような母集団のもつ期待値に対し、検定により差があるかないかを判断しようとしている母集団のもつ期待値 μ_{0B} がどれだけずれているかを表す指標である。つまり出目に偏りがあるかもしれないサイコロについて検定しようとしているのであれば、その真の出目確率 μ_{0B} が 1/6 ($= \mu_{0A}$) であるなら効果量がゼロであるということになる。検定力を算出するためには、効果量 $\mu_{0B} - \mu_{0A}$ について、別の方法でおよその値を推定しなくてはならないが、だからと言って、この効果量が事前に正確にわかるとは限らない。なお、真の効果量がわかっているなら、そもそも統計的な手法に頼り母平均 μ_{0B} を推定する必要はない。

検定力曲線を示す。グラフの横軸には効果量に対応する量として、このサイコロの特定の目の真の出目確率 μ_{0B} をおく。確率なので当然 0 から 1 の範囲である。縦軸には「検定力」$1-\beta$ として、それ

ぞれの確率（μ_{0B} について 0 から 1 まで）をもつ偏ったサイコロで試行を行い、帰無仮説が棄却される確率を取る。左は 12 回の試行回数、右は 60 回の試行回数を用いた両側検定で、有意水準はともに 0.05 としている。つまり縦軸は、12 回試行の場合は 6 回以上の出目の出現確率に等しい。また、60 回試行の場合は 4 回以下および 17 回以上の出現確率の和に等しい。

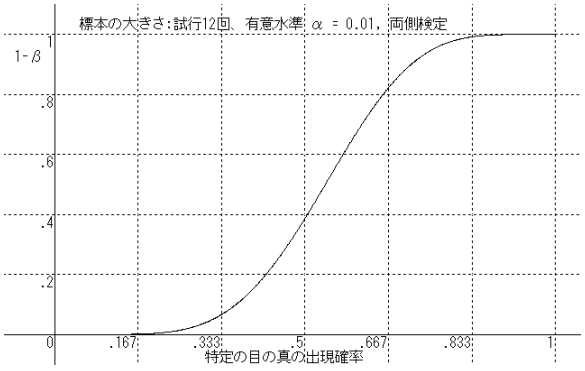


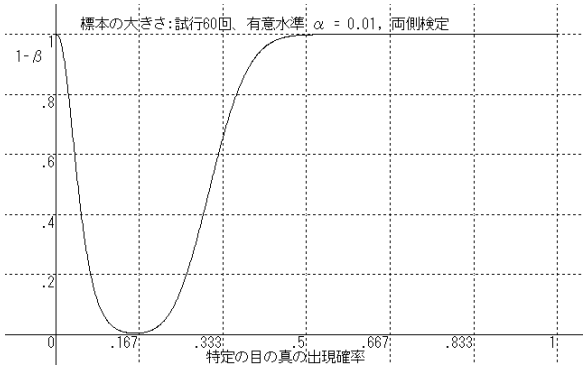


これらのグラフの形状は、帰無仮説 H_0 の下での出現確率が 1/6 であるから、実際には偏ったサイコロを使ったとして、その偏りが小さく、真の出目確率が 1/6 付近である（効果量が小さい）場合には、帰無仮説を棄却しにくいということを示している。真の出目確率と 1/6 との差がゼロに近づく極限では、1-β は α に収束する※。また、12 回試行では両側検定していても 0 回の出現でも H_0 を棄却できないため、結果的に片側検定と同じになり、図は、確率 1/6 の左右で非対称になっている。

※ この収束については、検定したサイコロの真の出目確率が「厳密には 1/6 ではないがほぼ 1/6 である」という収束条件を考えてみればよい。厳密には 1/6 ではないのだから、帰無仮説は正しくない。しかし出目確率はほぼ 1/6 なので、帰無仮説が正しい偏りのないサイコロを振った場合と限りなく近い挙動を示す。検定の結果、帰無仮説を棄却する割合は α に限りなく等しくなる。つまり、帰無仮説が正しくない時に棄却されない割合が β であるから、この収束条件下では $\alpha+\beta \rightarrow 1$ である。

さらに、有意水準 0.01 で両側検定する場合の検定力曲線も示しておこう。棄却域の限界値は 0.005 であるから、12 回試行の場合は、7 回以上の出目の出現確率に等しく、また、60 回試行の場合は、2 回以下および 19 回以上の出現確率に等しい。有意水準を下げることにより、当然のことながら、帰無仮説を棄却しにくくなっており、従って検定力も相対的に小さくなる。





一般に望ましいとされる検定力の目安は、0.8（以上）である。これに従えば、有意水準を 0.05 とした場合でも、効果量「偏ったサイコロだったとして、その偏りがどの程度なのか」が小さい場合、つまり「それ以上偏っていたら検出しようと想定しているサイコロの出目の確率の範囲」が 0.05 から 0.33 までの間（検定力が 0.8 を下回る範囲）である場合には、60 回の試行という標本の大きさでは、標本の大きさが十分ではないために検定力が不足であるとみなさざるを得ない。

もう少しだけ、具体的に書いておこう。サイコロの製造過程において致命的な誤りがあり、正しいサイコロの中に、「６の目の位置に１が印字されたサイコロ（$\mu_{0B} = 2/6$）」がいくつか混ざっているとする。これらのサイコロを投げて１の目の出る回数で検出することを試みよう。有意水準 0.05 で片側検定したとき、間違ったサイコロを「出目確率が有意に 1/6 より大きい」として拾い上げることがどの程度可能なのだろうか。検定力曲線を調べれば、その答えがわかる。試行 12 回（5 回以上で陽性）では、37 % しか拾い上げることができず、残りは有意差なしと判定されるだろう。標本の大きさを大きくし、それぞれのサイコロにつき 60 回の試行を行った場合（16 回以上で陽性）では 89 % まで検出力が上がる。その結果、間違ったサイコロの１割に対しては「有意差なし」として拾いこぼしてしまうことが避けられないが、９割に対しては正しく「有意差あり」と結論することができる。

○ 非有意と差なし　：　検定力の不明な有意差なしは、差がないことを示唆しない

帰無仮説が棄却されない場合でも、第２種の過誤なのかも知れず、帰無仮説が正しいから棄却されなかったとは限らない。「差があるとみなすことが妥当だとは判断できない」だけであり「差がないと判断された」わけではない。つまり、有意差が認められないからと言っても帰無仮説が「採択された」と考えるのは誤りである。仮説検定の手法で、帰無仮説を「採択する」、すなわち「全く差がない・効果がない」ことを積極的に示すことは、原理的にできない。なお、資料や解説において「帰無仮説を採択する」という表現がなされることがあるが、少なくとも消極的に採択するのみであり「帰無仮説が正しい」という意味でつかわれていることはない（はず）。どのような意図であっても、誤解のないように「帰無仮説を採択する」という表現を使用しないように心掛けた方がよい。

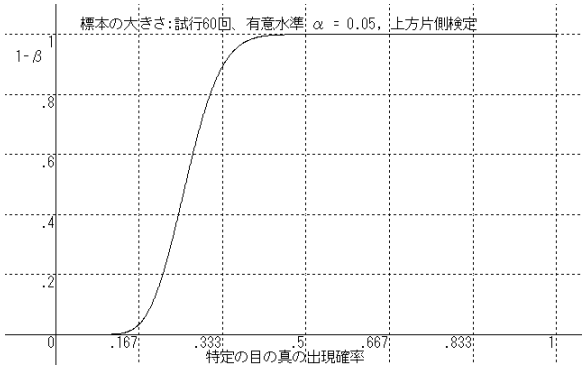
また、たとえば何かの悪影響の有無を調べるために、統計調査を行って、検定力が不足していたためなのに「有意な差は認められなかった」と述べたとしたら、暗に差がないことを示唆しているように受け取られないだろうか。しかし、検定力が小さいならば、第２種の過誤の可能性が高いのだから、この調査結果は判断基準になり得ない。「少なくともある量以上の悪影響はない」という保証をしたいなら、その最低の効果量に対する検定力が 少なくとも 0.8 以上あることを示さないといけない。

想定される効果量（上のサイコロの例では、特定の出目確率の 1/6 からのずれ）に対し、算出された検定力 $1-\beta$ の値が 0.8 より大きいことが示されているならば、「第２種の過誤が生じている確率 β は 0.2 より小さい」「その効果量以上の差をもつのに H_0 を棄却していないということはないだろう」「差があったとしても想定される効果量より小さかっただろう」と言うことができる。ただしこの想定される効果量が小さいほど、検定力を 0.8 以上にするために必要な標本の大きさは大きくなる。

○ 基準率の錯誤　：　有意差ありは、真の陽性であることを強くは示唆しない

ここでは、検定されたデータの解釈において、有意差ありと判定されたものが、p 値がいくら小さくても、実際に効果や差を持っているかどうかを判断するための基準にはならないことを示す。基準率とは、母集団の中で真の陽性の含まれる割合である。今、20000 個のサイコロの中に、1 の出目確率が 2/6 あるようなサイコロ（$\mu_{0B} = 2/6$）が 20 個混ざっており、これを拾い出すために、すべてのサイコロについて、60 回ずつ試行を行い、H_1: $\mu_{0B} > \mu_{0A}$ について有意水準 0.05 で片側検定を行う。ただし残りのサイコロは出目確率が均等である（$\mu_{0A} = 1/6$）。今ここでは、この１の出目の確率が $\mu_{0B} = 2/6$ である偏りのあるサイコロを真の陽性と表現しよう。$\mu_{0A} = 1/6$ のサイコロを真の陰性と表現しよう。

基準率とは、標本抽出の回数全体（$n_A + n_B = 20000$）の中で真の陽性（$n_B = 20$）にあたる割合であるから 0.1 % である。検定力は、有意水準（$\alpha = 0.05$）、標本の大きさ（試行回数 $n = 60$）、効果量（出現確率 $\mu_{0B} = 2/6$）により決まる。次ページに、1 の出目の出現確率（μ_{0B}）を 0 から 1 まで変えて



算出した検定力曲線を示した。この条件での検定力（μ_{B0} = 2/6 のサイコロに対して有意差ありと判定される確率）は 89.2 % と算出されるから、真に陽性である 20 個のサイコロに対して、期待値として 2 個は誤って偽陰性を与えるが、残り 18 個は正しく「有意に偏りあり」とみなされることになる。問題は、実際には真に陰性である（出目確率に偏りのない※）サイコロについてである。有意水準 0.05 で検定しているので、19980 個の真に陰性であるサイコロのうち 5 %、すなわち 999 個程度は偽陽性を与えるだろう※2。すなわち、この検定を行うことにより、20000 個のサイコロに対し 1017 個のサイコロが有意に出目の確率に偏りがあると陽性判定され、そのうち実際に真の偏りを有するものは 18 個にすぎない。すなわち、真の陽性はたったの 1.8 % でしかない。p 値が有意水準 1/20 を下回ったケースのみ有意差ありと結論しているのだから、「この陽性判定は 1/20 しか間違わない」と言っているように思われる（これは誤りであるが、そのような誤解が多々ある）にも関わらず、有意差ありと判定（陽性判定）した中の 98 % 程度以上は偽陽性、すなわち第１種の過誤である。

※ とはいえ、今は 1 の出目回数しか問題にしていないから、1 の出目の確率が 1/6 でさえあれば、他はどうであっても、たとえば他の 5 面すべてが 6 になっていようと、知ったことではない。

※2 試行回数 60 回では、確率分布が連続とみなせるほど標本が十分に大きいわけではないため、実際には、有意水準（危険率）よりやや小さい確率で偽陽性を与える。出目回数 15 回では帰無仮説が棄却されず、16 回以上の出現確率が 3.385 % なので、偽陽性の数は、999 個ではなく 676 個である。しかしこの数に基づいて計算しても、陽性を与えた中で実際に真の陽性であったものは 2.7 % にすぎない。

このように基準率が小さい場合は、偽陽性の割合が増え、陽性と判定されたものが実際に真の陽性である割合は小さくなる。極端な例を考えてみれば、すぐにわかる。基準率が限りなくゼロに近い、またはゼロである場合である。かならず有意水準（すなわち危険率）の割合を最大として第１種の過誤を生じ、偽陽性を与えるのに、真の陽性がゼロであるならば、有意差ありと判定されたものすべてが偽陽性であることは言うまでもない。<u>p 値が保証しているのは、効果がないものでも効果があると判定される確率である。効果があると判定されたもののうち、実際に効果がある確率ではない。</u>

（参考）
http://id.fnshr.info/2014/12/17/stats-done-wrong-toc/
　ダメな統計学
引用： そして、お願いですから、次に誰かが「この結果は p<0.05 で有意だから、これが偶然である確率は 20 分の 1 しかない！」と言うのを聞くことがあったら、私のためにその連中の頭を統計の教科書でぶったたいてください。
引用： 訳注：実際に教科書でぶったたいてどんな結果がもたらされたとしても、訳者は責任を負いかねるので、読者諸氏は注意されたい。

○ p 値偏重に対する批判

「p 値が任意に定めた有意水準を切ったかどうかだけで、商業上や政策上の決定を下したり、科学的な結論を導いたりするべきではない」（アメリカ統計学会の声明、参考にて引用したリスト中の３番の私訳）。

p 値が任意の有意水準より小さければ、帰無仮説が棄却されるという考え方自体は、ひとつのアプローチとして間違っているものではないが、その適用、ならび、解釈に対して、適切に行われていない場合が多いこととを受けての批判もある。基準率の錯誤でも述べたように、ある標本の p 値が 0.05 より小さく有意差ありと判定されたからといって、そのことから、その標本において実際に（少なくとも 95 % 以上の可能性で）差や効果があるという判断を下すことはできない。標本の大きさ n を大きくしさえすれば、意味のないほど小さな差（小さな効果量）であっても有意差ありと判定されやすくなる傾向がある。逆に標本の大きさ n が十分ではないために、検定力が不足していれば、見逃してはいけない差を有意差なしと判定してしまうことになる。また、統計データから p 値を算出する際に、恣意的にその値を操作することが比較的やさしいため、脆弱なデータを支持するために使用されてしまうことがある。そもそも、有意水準 0.05 で検定するということは、本質的には差異がない現象に対しても 20 回に 1 回は有意に差が認められてしまうということである。だから、たとえば有意差を認められる結果になるまで何度でも検定を繰り返し（多重比較）、そのことを伏せておけば（悪意あるトリミング）、あたかも科学的な方法で有意差を見出したかのように捏造することができる。もちろんこれは正しい方法ではないが、第三者からは判別できない。

（参考）
https://www.amstat.org/newsroom/pressreleases/P-ValueStatement.pdf
AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND P-VALUES
Provides Principles to Improve the Conduct and Interpretation of Quantitative Science
（アメリカ統計学会 The American Statistical Association (ASA) が、p 値偏重主義から脱却するという声明を発表。2016.03.07）
引用：The statement's six principles, many of which address misconceptions and misuse of the p-value, are the following:
1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

（参考）
http://www.jspp.gr.jp/doc/jspp2013kikaku1.pdf
効果量と信頼区間： p 値だけでは不充分
（専修大学人間科学部教授 大久保街亜氏による資料。APA アメリカ心理学会の投稿ガイドなどを例にとり、仮説検定の結果に対し、その効果量や信頼区間を合わせて報告すべきことを解説している。）

（参考）
http://www.nature.com/news/psychology-journal-bans-p-values-1.17001
Psychology journal bans P values
（社会心理学系のジャーナル、Basic and Applied Social Psychology (BASP) が、帰無仮説有意性検定およびそれに類する統計学的処理を禁止すると発表。2015.02.16）

○ 多重比較と第１種の過誤の増加

同一の実験系に対し、仮説検定を繰り返すことを多重比較という。多重比較は、第１種の過誤の確率を上げる。つまり、有意水準（すなわち危険率） 0.05 で検定するとき、第１種の過誤の生じる確率は 0.05 である。この実験系で、2 回繰り返して検定を行ったときに、そのいずれかで第１種の過誤が生じる確率は、$1-(1-0.05)^2 = 0.098$ で、約 2 倍となる。14 回では $1-(1-0.05)^{14} = 0.512$ と 5 割を超える。45 回では $1-(1-0.05)^{45} = 0.901$ と 9 割を超える。

同一の検定を有意差が認められるまで繰り返すことが正しい行為ではないことは、比較的わかりやすい。見落としてしまいがちであるが、次に挙げるような例も、多重比較に相当する。

「ある薬品が、ヒトの健康に影響を与える」ことを有意であるかどうか、判定したい。すると帰無仮説は「その薬品はヒトの健康に影響しない」である。そこで、適切な終点（endpoint、評価項目と基準のこと）を定めて、A（血圧が上昇する）、B（体重が増加する）、C（頭髪が抜け落ちる）の３つの項目について実験し、それぞれについて有意水準 0.05 で検定したとしよう。また、前提として帰無仮説が正しかった（その薬品が真の陰性であった）ものとしておく。

はじめからこの薬品が血圧上昇に影響があるかどうかについてのみ着目しており、この薬品によって太ろうが禿になろうが関係ないという立場※であるなら多重比較には相当しないのであるが「人の健康に影響を与えるかどうか」が焦点であるなら、最大で有意水準と同じ 0.05 の危険率で A について第１種の過誤を生じ、B についても同じ危険率で第１種の過誤を生じ、C についても同じ危険率で第１種の過誤を生じる。すなわち、全体としては 0.1426 の危険率となり、つまり 20 回に 3 回程度は、いずれか一つまたは複数の健康チェック項目において、有意に影響ありと判定されることになる。もしこのチェック項目が 14 種類あると、それだけで、本質的には全く影響を持たない薬品に対してでも、たとえば毎日飲み食いしている食事や飲料水に対してさえ、全体として 0.512 の危険率、つまり 2 回に１回以上、チェック項目のいずれかにおいて有意に影響を及ぼすと判定してしまうことになる。

※ ただし、このような立場なら、はじめから項目 B や項目 C についての実験は行われないだろう。なお、デブや禿が不健康であるという主張をしようとしているわけではない。

また、標本 A と標本 B の間に相関や差異があるかどうか※を判定するだけであれば問題ないが、３群以上の標本、たとえば標本 A、標本 B、標本 C の間に相関や差異があるかどうかを調べようとして、A-B 間、B-C 間、A-C 間で検定を行うといった場合も、やはり多重比較になってしまう。これを避けるためには、帰無仮説として A = B = C を用いて、χ^2 検定、F 検定、分散分析など（いずれも後述）を行うことができる。ただし、帰無仮説が棄却された場合、A-B 間、B-C 間、A-C 間のいずれかに（この場合は）差異がある（「A = B = C」ではない）ことがわかるが、そのどの２群間に差があるのかはわからないことになる。（下位検定として、A-B 間、B-C 間、A-C 間で検定を行うなどする必要がある。）

※ 当然のことながら同一の母集団から取り出しても標本が完全に一致することはほとんどない。「標本平均が同じであるとみなせる」という表現は、「それぞれの標本を抜き出した母集団が同じ母平均をもつと判断することが妥当である」ことを示している。つまり、「標本間に差があるとみなせる」は、「標本を抜き出した母集団の間に差があると判断することが妥当である」ことを意味している。

○ ではどうすればよいのか

仮説検定で有意差を見出すことは出発点であり、目的にしてはならない。有意差ありと認められた一群に真の陽性（効果あり、差あり）が含まれている確率は、有意差なしと判定された一群よりも上が

っているのは確かである。しかし、有意差ありと判定されただけで真の陽性であると考えることはできない。真の陽性を探し当てるには、更に実験を重ねての再現性の有無の確認、別法による検討などが必要になるだろう。また有意差がないと判断された場合も、第２種の過誤が必ず起きることを忘れてはならない。また、効果量や信頼区間などを検討し、検定力がどのくらいであるのか、基準率がどのくらいであるのかなどについて考慮することで、これらを定量的に議論できるようになるだろう。

◎ 確率分布と統計

上で行っているように、サイコロの特定の出目の有無、コインの表裏、薬理効果の有無、ある事象の出現回数などのように離散的な事象は、それぞれの確率を想定することにより、有限の試行回数の結果どうなるかを数えあげることができた。しかし、そのような場合でも、試行回数が多いと厳密な計算は大変になるし、測定値のように連続値を扱う場合や、本質的な確率を設定することが難しい場合などは、他の方法を用いることが必要となる。それが確率分布である。

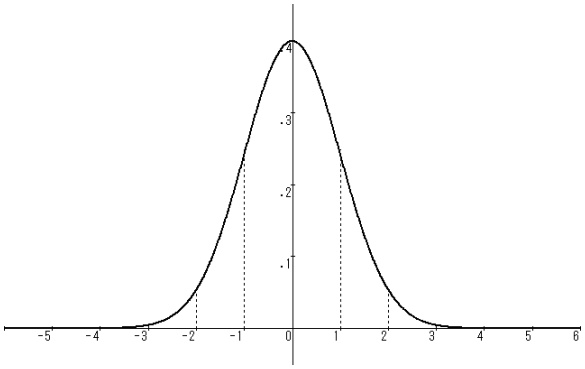
仮説検定とは、ある特性をもつ２つの母集団があるとき、その両方もしくは一方から標本を抜き出して比較することで、その２つの母集団の分布や特性、特に平均値に差があるとみなすことが妥当であるかどうかを判定するための手法※であった。実験に該当させるならば、将来も含めて何度でも繰り返したときに得られるはずの個々の結果の集合が母集団にあたり、抜き出された標本とは、実際に有限回の試行結果ということになる。比較しようとしているもの（２つの母集団の性質に違いがあるかどうかを判断するために、その２つの母集団からそれぞれ抽出して比較しようとしている標本）は、たとえば、一つは均等な確率に従って出目を与える正しいサイコロの試行結果を表す数値の一群※2 であり、これと比較するもう一つは出目の確率が偏っているかもしれない（偏っていることを作業仮説として考えている）サイコロの試行結果の一群である。たとえば、薬効の有無を調べたいとき※3 は、他の条件をそろえる必要があるため、一つは偽薬を投与した場合の結果の一群であり、これと比較するもう一つは実薬を投与した場合の結果の一群である。

※ まずは母集団のもつ平均が同じかどうか（母平均の推定値としての標本平均を、同じであるとみなしてよいかどうか）について調べているが、母分散が同じかどうか（母分散の推定値としての標本不偏分散が同じとみなしてよいかどういか）についても、仮説検定で調べることが可能である（F 検定、後述を参照）。また、繰り返しておくが、有意差が見られたからと言って、２つの母集団の間に確実に差があるということもできないし、また、有意差がないからと言って同じであるという判定もできない。また、有意差があるとわかったとき、どの程度の差があるかどうかということを集団の分布特性から推定することは可能であるが、仮説検定の結果だけから言うことはできない。

※2 抽出された標本として有限回の試行を行った場合には、たとえ公正なサイコロを用いていても、実験から算出される特定の目の出目確率は、いつも 1/6 を与えるわけではないから、標本ごとの値は確率分布に従うことになる。一方、サイコロの試行回数を無限大に近づけることにより、特定の出目の出現確率は（1/6 に等しいかどうかは別にして）１つの数字に収束するが、これは無限大の大きさの標本を抽出した場合に「標本平均の分布」の分散がゼロに収束してデルタ関数を与えるのであって、母集団が非常に狭い分布をもっているからではない。詳細は標準誤差についてのところで述べる。

※3 ここでは検定の方法や考え方が主題なので、詳細は省略するが、薬理効果の試験にあたっては、盲検法（ブラインドテスト；被験者が偽薬か実薬かを知らされずに行われる）や、二重盲検法（ダブルブラインドテスト；投与する者と被験者のいずれもが、偽薬か実薬かを知らされずに行われる）などにより、偽薬効果（思い込みの影響）を分離するなどの配慮も必要となる。

○ 正規分布を仮定する場合



まずは単純な説明のためにある集団（母集団そのものや、そこから抜き出した標本）が示す分布が、正規分布に従うものと仮定しよう。その上で、ここまでに出てきた p 値、有意水準（危険率）、第１種の過誤、第２種の過誤、検定力といった言葉を、図との関係で理解していこう。

右図は、標準正規分布を確率密度表示したものである。正規分布では、期待値である平均値と、最頻値（モード）、中央値（メジアン）は一致しているので、一般に平均値 μ と分散 σ^2（または標準偏差 σ ）の２つがわかれば、その正規分布についてわかったと言ってよい。ここでは標準化されているので、標準偏差 $\sigma = 1$、平均値 $\mu = 0$ である。なお、確率密度なので、グラフが成す面積（$-\infty$ から $+\infty$ の範囲での積分値）は規格化されており、1 となっている。

なお、この後の話の展開を読み進めるにあたって、混乱しないためには、扱おうとしている分布が、次のいずれについてのものであるのかを明確にしておく必要がある。

a） 母集団の分布
b） 母集団から抜き出した１つの標本の示す分布
c） 母集団から独立に抜き出した標本について、標本平均（などの推定値）が示す分布

○ 中心極限定理に従う母集団から抽出した標本平均の期待値は母平均に一致する

中心極限定理に従う場合、母集団がどのような分布をしていようと、その母集団から独立に抜き出された標本が大きければ、標本ごとの平均値（標本平均）μ は、母平均 μ_0 からのずれ $(\mu - \mu_0)/\sigma_m$ が標準正規分布に従う。ここで、知りたいのは母集団の性質であるから、母平均 μ_0 は現時点で未知であり、抽出した標本から推定したい値である。そのため、標本は、母平均などの情報を知らずにランダムに抽出するので、これより大きい値をもつ要素が多く抽出されることもあり、逆もあり、結果として、個々の標本平均はバラつくからである。そのため、可能な組み合わせすべて（または十分な数）について抜き出された標本平均の平均（標本平均の期待値）は、母平均に等しくなる。

○ 分散と不偏分散、不偏分散の期待値は母不偏分散に一致する

分散とは、ある母集団、または１つの標本について、「偏差（平均値と各測定値の差）の平方和」を標本の大きさ n で割ったものである。不偏分散とは「偏差の平方和」を「標本の大きさ n 」の代わりに「自由度 n-1 」で割ったものである。分散、不偏分散 σ^2 の平方根は、標準偏差、標本標準偏差 σ である。ただし、n が大きい極限では $(n-1)/n = 1-(1/n)$ は 1 に収束するので、分散と不偏分散は一致する（したがって、標準偏差と標本標準偏差も一致する）ので、標本の大きさが十分に大きい場合にはこの違いを意識しなくても問題ない。n-1 を自由度と言うのは、標本の大きさ n に対し、その平均値を指定することで測定値を n-1 個までは自由に決めることができるのに対し、最後の一つは平均値と残りの数値から算出されてしまうためである。

標本平均の期待値は母平均に一致するが、一般に、母集団から抽出された１つの標本に対して分散（偏差の平方和/n）を求めると、その期待値は母分散より小さく、標本の大きさが n のとき、(n-1)/n 倍に見積もられてしまうことになる。一方で、不偏分散の期待値※は、ある母集団から標本を抽出

した際に、母集団と標本とで不偏な（変わらない）統計値である。

※ 標本１つでは、標本平均も標本不偏分散も、母平均や母分散と一致するわけではない。可能な限りの標本の取り出し方の組み合わせすべてについて平均をとったもの（すなわち期待値）は一致する。

これらのことを、具体的に数値を用いて確かめてみよう。たとえば大きさ 3 の {-1, 0, 1} を母集団とする。母平均は 0 である。母分散に寄与する偏差の平方和は、$(-1-0)^2 + (0-0)^2 + (1-0)^2 = 2$、これを自由度 2 で割ると 1 となる。これが母不偏分散※である。この中から大きさ 2 の標本として非復元抽出できるのは {-1, 0}、{-1, 1}、{0, 1} の３通りである。この標本平均は、-0.5、0、0.5 であるから、その期待値（平均）は 0 で、母平均と同じである。また、標本不偏分散を求めると、自由度 1 なので偏差の平方和と等しく、$(-1+0.5)^2 + (0-0.5)^2 = 0.5$、$(-1-0)^2 + (1-0)^2 = 2$、$(0-0.5)^2 + (1-0.5)^2 = 0.5$ となる。期待値は、この３つについて平均をとり、母不偏分散と同じく 1 となることがわかる。ここで、それぞれ不偏分散とする代わりに、分散を求めると、すなわち自由度の代わりに集団の大きさで割ると、母分散は 2/3 = 0.67 となり、標本分散は 0.25、1、0.25 の平均から 0.5 となってしまい、互いに一致しないし、不偏分散とも一致しない。なお、これらが不偏分散 1 に対して 2/3 倍、1/2 倍であるのは、(n-1)/n 倍の n=3 および n=2 のケースに相当する。

※ 一般に無限の大きさをもつ母集団を扱う場合は、$(n-1)/n = 1-(1/n)$ が 1 に収束するため、不偏分散と分散が等しいものと扱ってよいが、有限の大きさの母集団の場合には、この近似が成立しない。

（参考）
http://mathtrain.jp/huhenbunsan

○ 標準誤差 SEM

母集団から同じ大きさの標本を複数抜き出して、それぞれの平均値（標本平均）の平均（「標本平均の期待値」の近似※）を算出し、母平均を推定する手順について考える。いま、「一定の大きさの標本を多数、互いに独立に抜き出した※2場合の、標本平均が示す分布について、その分散」を、添え字 m（平均 mean より）をつけて、σ^2_m と表記することにしよう。

※ 十分な数の標本が抜き出されていない場合は、近似的な推定値に過ぎない。

※2 この独立性を保証するためには、復元抽出が必要となる。すなわち、一度取り出した数値を次の抜き出し操作の前に母集団に戻すので、１つの標本中に母集団中の同じ要素を複数回抜き出す可能性もある。母集団の大きさを 5 としたとき、従って大きさ n の標本は 5^n 通りについて検討することになる。（復元抽出においては、有限母集団より大きい標本を抜き出すことが可能である。）この場合でも、標本平均の期待値は母平均に一致する。また、不偏分散の期待値は母集団の分散（偏差の平方和を n-1 ではなく n で割ったもの）に一致する。なお、上の例で行った「重複せずに」標本を抜き出す方法は、非復元抽出にあたる。抜き出した標本の中の１つ目の要素が、「重複しないこと」を要請することで２つ目以降の要素の選択に影響を与えてしまうため、独立な抽出ではない。ただし、母集団の大きさが抜き出す標本よりずっと大きいなら、この影響は無視できる。

（参考）http://www.f-denshi.com/000TokiwaJPN/17kakto/060prob.html
ときわ台学/統計学/くっまんす王国（１）
（F 氏による資料。標本統計用語に数値を入れて計算しながら具体的に確認する内容の解説）

抜き出した標本の大きさが十分に大きいならば、大数の法則により、標本平均のすべては母平均と

（ほとんど）同じ値を示すので、母平均の推定値である「標本平均」の分布は非常に狭くなり、標本平均の不偏分散 σ_m^2 も小さくなる。極限では標本平均の不偏分散はゼロに収束し、標本平均の分布はデルタ関数となる。抜き出す標本の大きさが母集団の大きさに比べてずっと小さい場合であっても、その標本平均の期待値が母平均に一致する点は同じであるが、個々の標本平均はばらつきはじめ、独立に抜き出された標本が十分な数存在すれば、その標本平均の分布が正規分布を示すのは上述した通りである。このように標本平均の分布の不偏分散は、標本の大きさに依存することになる。一般に推定値の標準偏差のことを「標準誤差」とよび、「標本平均の標本標準偏差」 σ_m のことを「平均値の標準誤差」（Standard Error, または Standard Error of Mean）と呼ぶ。この値は、個々の標本の不偏分散の期待値や、母分散（ともに標本の大きさに依存しない）とは一致しないのは明白であろう。

「標本平均の不偏分散」σ_m^2 と母分散 σ_0^2 との関係は、標本の大きさを n として、$\sigma_m^2 = \sigma_0^2/n$ で表される。母分散は、一般には未知であるから、その推定値として標本個々の標本不偏分散（またはその平均値）などを σ として用い、$\sigma_m = \sigma/\sqrt{n}$ と推定して用いることができる。すなわち、母集団がどんなに広い分布をもっていても、その平均値の真の値は１つ存在し、その値を推定するためには、大きい標本をとる方が精確※であることを示している。この式は n が無限大なら分散 σ_m^2 はゼロに収束し、精確な平均値が得られ真の値※2に収束するだろうこと、すなわち大数の法則に対応する。また、標準誤差 σ_m を 1/k 倍にするためには、推定に用いる標本の大きさを k^2 倍にしなくてはならない。

※ ある数値に対し、偶発誤差によるばらつきが小さいか、または標本の大きさが大きくなることにより、標本平均の標準誤差（SEM）が小さいことを、その数値の精度が高いと表現する。系統誤差（バイアス）が小さいことをその数値が正確であると表現する。

※2 偶然誤差によるばらつきは n を大きくすることで打ち消せるが、系統誤差は残される。

○ 母集団、標本、標本平均のそれぞれの分布についてのまとめ

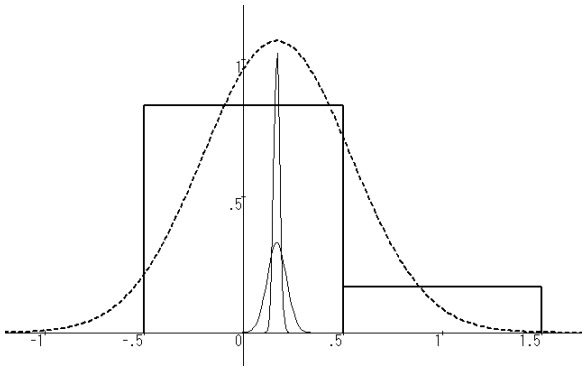
母集団とは、通常は直接その真の値を知り得ないために、標本から推定したい値のばらつきを含む集合と考えてよいが、有限の大きさをもち、真の値を原理的に知ることができるものでも構わない。

a) 母集団の母平均を μ_0 、母分散を σ_0^2 と表しておく。
母集団から独立に抜き出す標本の大きさを n とし、抜き出した個数を L とする。
原理的に、L は大きいほど良い近似を与えるが、推定において L = 1 の場合もある。

b) ここから抜き出した個々の標本は、標本平均 μ と、その標本内での不偏分散 σ^2 をもつ。
大数の法則より、n が大きいほど母集団の性質をよく表す。$n \to \infty$ のとき
$\mu \to \mu_0$ 、$\sigma^2 \to \sigma_0^2$
一般に、n が十分に大きくない限り、個々の標本平均や標本不偏分散は、母集団と一致しない。
しかし、n に無関係に、標本平均の期待値、標本不偏分散の期待値は母集団と一致する。
つまり $L \to \infty$ のとき
$\Sigma\mu/L \to \mu_0$ 、$\Sigma\sigma^2/L \to \sigma_0^2$

c) 標本平均について分布を考える。標本平均の平均は μ_m 、標本平均の不偏分散は、σ_m^2 をもつ。
「標本平均の集合」は大きさＬの１つの標本とみなしてもよい。
L が十分な数であれば、つまり $L \to \infty$ のとき
μ_m は標本平均の期待値となるので、$\mu_m \to \mu_0$
n に無関係に $\sigma_m^2 \to \sigma_0^2/n$ 、（さらに n も大きければ $\sigma^2/n \to \sigma_0^2/n = \sigma_m^2$ ）

従って、μ_0 の推定値として使うことのできる量は、μ または μ_m など。
また、σ_0^2 の推定値として使うことのできる量は、σ^2 または、$\sigma_m^2 \times n$ のいずれか。
ただし、母集団が無限の大きさを持つ場合は、有限の L や n では正しく一致する保証はない。



具体的な例を示しておこう。右図は、確率が均等なサイコロを振った結果、特定の出目の確率を示している。母集団の確率分布は、0 または 1 のヒストグラムで表現される。このヒストグラムは、面積の和が 1 となるように表現するため、横軸（特定の出目の出現確率）は、0 以下および 1 以上にまで延長されている。この分布に正規分布を仮定して n を十分に大きいものとして分散の計算を行ったところ、標準偏差 $\sigma_0 = 0.3727$ であった。この分布曲線を破線で示した。n により母集団の確率分布は変化しないはずだから、この標準偏差もまた（n が十分に大きいなら）一定のままであるとみなしてよい。母集団から大きさ 60 および 600 で抽出した標本について、n が十分に大きいものとみなし、標本平均の分布について正規分布を仮定して併せて表示した。ただし、標本平均の分散は母分散/n（σ_0^2/n）で求めた。また同じ図内に収まるよう、縦方向に 1/25 倍して示している。分散の小さい（半値幅の狭い）方が n = 600 での標本平均の分布で、この場合の平均値の標準誤差は $\sigma_m = 0.0152$ であった。

○ 信頼区間

母平均を知る目的で母集団から取り出した標本の平均が、母平均 μ_0 を中心とし、分散 σ_m^2 で正規分布するものとする。すなわち、標本平均が、$\mu_0 \pm 2\sigma_m$ の外の領域にくる確率は、0.05 である。逆に $\mu_0 \pm 2\sigma_m$ の範囲内に、母集団から独立に取り出された標本平均のうち 95 % が含まれていることになる。実際の手順としては、母平均を知りたいために母集団から標本を抽出しているわけであるから、視点を変えてやって、抽出された１つの標本の平均 μ に対し、$\mu \pm 2\sigma_m$ の範囲※を見てやれば、その範囲内に母平均 μ_0、つまり標本から推定したい「真の値」（未知ではあるが変動しない値）が存在している逆確率が 95 % であることを示す。この領域のことを、95 % の信頼区間という言い方をする。すなわち、95 % 信頼区間とは、実験者には知り得ない母平均に対して設定されるものではなく、抽出された標本ひとつずつに対して設定される領域である。なお、抽出された標本の平均値 $\mu \pm \sigma_m$ の範囲は、68 % の信頼区間に相当する。信頼区間をもう少し精度の高い数値として示しておくならば、95 % では $\mu \pm 1.96\sigma_m$、99 % では $\mu \pm 2.57\sigma_m$、99.9 % では $\mu \pm 3.29\sigma_m$ などである。

※ 式中の σ_m が標本平均の不偏分散に基づく値であり、母分散 σ_0^2 や標本の不偏分散 σ^2 とは異なることを改めて注意しておく。σ_m^2 の値は取り出した標本の大きさに依存する。

ただし、標本平均の不偏分散 σ_m は標本の取り出しを何度も行って分布を調べないと直接求めることができないから、母分散の推定値である標本の不偏分散をもとに推定し※、σ^2/n としてもよい。同じ大きさで抜き出された標本数が十分に多い場合は、その標本平均の分布から上述の標準誤差（SEM）σ_m を求め、標本平均の期待値 μ_m（すなわち標本平均の平均）に対し、$\mu_m \pm 2\sigma_m$ などとして表す。

※ これは、標本の不偏分散の期待値が母分散に一致するからである。そのため、同じ母集団から抜き出した複数の標本があるなら、それらの不偏分散の平均を用いる方がよい近似となるだろう。σ_m^2 の値は、母分散 σ_0^2 の値が既知ならば標本の大きさ以外には依存しないが、標本の不偏分散 σ^2 （期待値ではないため、標本ごとに異なる値をもつ）を基に推定した場合は、標本ごとに異なる。

○ 標準偏差（標本標準偏差）と標準誤差　　エラーバーの使い分け

測定値を図示する際に、値の範囲を図上にエラーバーとして表示することがある。これは、次のように目的に応じて使い分けなければならない。

母集団の分散が関心の対象である場合：標本標準偏差 SD を用いて表示する。n が大きくなると、標本の不偏分散は母分散に近づく。ある測定値のばらつきを表すために、複数の測定値の一群をひとつの標本とみなし、その標本平均 μ に対し、$\mu\pm\sigma$ や $\mu\pm2\sigma$ として範囲を示す。

真の値すなわち母集団の平均が関心の対象である場合：標準誤差 SEM を用いて表示する。n が大きくなると、標準誤差、すなわち標本平均の不偏分散は 0 に近づく。測定値の信頼区間を表したい場合には、$\mu\pm\sigma_m$ や $\mu\pm2\sigma_m$ などとする。

これらが表しているものの違いをあきらかにするために、具体例を見ておこう。たとえば、母集団 S として、ある地域 S における 18 歳男性の集合を考える。この母集団の平均体重 μ_{0s} を、別の地域 T における 18 歳男性の統計値 μ_{0t} = 62.5 kg と比較する目的で、無作為に 100 名を抜き出した標本１つについて調べた。この標本の平均体重は μ_s = 55.0 kg、標本標準偏差は σ_s = 5.0 kg であった。この μ_s = 55.5 kg の背景にある μ_{0s} が μ_{0t} = 62.5 kg との間に差があるかどうかを検定するため、この標本の 95 % 信頼区間を考える。95 % 信頼区間は、$\mu_m\pm2\sigma_m$ で表される。また、この時の σ_m は、$\sigma_m = \sigma_0/\sqrt{n} \fallingdotseq \sigma_s/\sqrt{n}$ で推定される。$\sigma_s/\sqrt{n} = 5/\sqrt{100} = 0.5$ なので、95 % 信頼区間は、55±1 kg となる。従って、母集団 S の真の平均体重は、95 % の逆確率でこの範囲に含まれていたはずである。この範囲に、比較すべき μ_{0t} = 62.5 kg という数値が入っていないため、地域 S における 18 歳男性の平均体重 μ_{0s} は、地域 T における 18 歳男性の平均体重と、95 % 以上の確率で一致しない。すなわち、有意水準（1 − 0.95 = ）0.05 で、有意に差が認められると結論してよい。

この 100 名の標本の内訳は、典型的にはおそらく次のようになっているだろう。40.5 kg から 45.5 kg の体重の者が 2 から 3 人、45.5 kg から 50.5 kg が 14 から 13 人、50.5 kg から 55.5 kg が 34 人、55.5 kg から 60.5 kg が 34 人、60.5 kg から 65.5 kg が 14 から 13 人、65.5 kg から 70.5 kg が 2 から 3 人。n が大きいときには標本の不偏分散は母分散の良い近似になっていたはずであるから、n = 100 が十分に大とみなせるなら、また説明の都合上、この標本平均 μ_s が母平均 μ_{0s} を正しく推定できていたならば、1000 名の標本をとった場合、同じ分散を持ち、分布は各階級について先ほどの人数を 10 倍したような分布になるだろう。ただし、40.5 kg 以下や 70.5 kg 以上の階級に属する者、1000 人中に 2 から 3 人も見えてくるかもしれない。また、さらに、母集団もこれらに比例した分布となっているだろう。このような母集団から１名の標本を抜き出したとする。その標本平均は、その１名の体重そのものである。そうすると、標本平均の分散 σ_m^2 は、標本分散 σ_s^2 に等しい。（標本数が十分に多ければ、母分散にも等しくなる。）このとき n = 1 なので、$\sigma_m = \sigma_s/\sqrt{n}$ が成立していることがわかる。また、大きな同じ母集団から 100 名の標本を複数回、復元抽出で抜き出したとしよう。この n = 100 の標本が複数あったとして、その標本平均がいつも同一の値をとる保証はなにもない。しかし、20 回のうち 19 回は、標本の平均値が母平均 $\mu_{0s}\pm1$ kg の範囲に収まるだろうと推測される。もし n = 10000 の標本を取り出すことが可能であるなら、その平均値は（母集団の標準偏差 σ_{0s} = 5.0 kg が正しいものとして）母平均 $\mu_{0s}\pm0.1$ kg の中に収まることになるだろう。

このように 95 % 信頼区間（$\pm2\sigma_m$ で表される） 55±1 kg は、母平均 μ_0 の推定範囲である。n が大きいほど狭くなり、100 名の標本のデータのばらつき、そしておそらくはかなり正しく近似できているであろう母集団のデータのばらつき（$\pm2\sigma$ で表したとき）55.5±10 kg よりもずっと狭くなる。

○　２つの母集団の比較　　信頼区間の重なりの有無

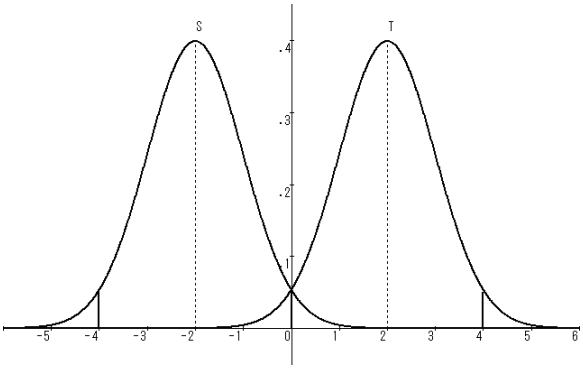
ここで、当然の疑問がでてくる。比較基準である地域 T における 18 歳男性の統計値 μ_{0t} = 62.5 kg という数値に信頼区間が示されていないが、これがあればどうなのであろうか。これによっても、当然結論は変わってくるだろう。データの範囲が、分散や標本標準偏差ではなく、95 % 信頼区間で示されているものとして次のように解釈される（ただし、以下の例において 55.5±10 のような場合、中央値の表し方 55.5 が適切かどうかはここでは議論しないことにする）。

S 地域 55.5±1、 T 地域 62.5±1　の場合：有意な差がある。
S 地域 55.5±1、 T 地域 62.5±10 の場合：有意な差があるとは認められない。
S 地域 55.5±10、T 地域 62.5±1　の場合：有意な差があるとは認められない。
S 地域 55.5±5、 T 地域 62.5±5　の場合：有意な差があるとは認められない。
S 地域 55.5±5、 T 地域 65.5±5　の場合：有意な差がある（と認める閾値上である）。

一方の信頼区間にもう一方の標本平均が入らなくても、信頼区間に重なりがある場合、つまり２つの正規分布の重なりの面積が有意水準を超える場合、帰無仮説（有意に差がない）を棄却できない※。

※ ２つの信頼区間に重なりがあり有意差を認めない場合でも、仮説検定したときに必ず同じ結果を与えることを保証するものではない。（仮説検定では、有意差ありとなる場合がある。）

（参考）http://id.fnshr.info/2014/12/17/stats-done-wrong-06/#toc1
【翻訳】ダメな統計学（6）有意であるかないかの違いが有意差でない場合
有意差が見逃されるとき



一番下の閾値上の例についてみてみよう。右に２つの σ_m = 1 の正規分布 S と T を、95 % 信頼区間が接するように描いた図を示す※。ここで、前提となっていることとして、知りたいこと（検証仮説）が、「２つの母集団の期待値（母平均 μ_{0s} と μ_{0t}、または、この文脈ではいわゆる「真の値」）が異なる」かどうかであり、比較しようとしているものが、ともにある母集団から抽出された標本平均である。それぞれの母集団から抽出されている標本の数は１かもしれないし、複数かもしれないが、この図に描かれているのは、標本平均の分布で、母集団や標本の分布ではない。いずれも、適切な処理によりそれぞれの標本平均の分散を正しく推定できているものとする。そのため、この分布は、それぞれの母集団がもつ真の値がどこにあるかの逆確率であると言ってよい。母平均の真の値がわかっていないので、１本の線として示すことができずに、信頼区間の領域で示している。

※ この２つの正規分布曲線のそれぞれの中央の位置は、それぞれの母集団 S、T から抽出した標本についての「標本平均」μ_s、μ_t、または複数の標本を抽出している場合は「標本平均の平均」であり、母平均の真の値 μ_{0s} や μ_{0t} ではない。

この場合の帰無仮説は、「２つの母集団がもつ真の値は同じである」なので、これを棄却するためには、２つの分布に重なりがないか、またはその重なりが十分に（有意水準より）小さいことが必要な条件となる。２つの独立事象が同時に生じる確率、重なり部分の面積は、S の 95 % 信頼区間より大きい側として 2.5 % あり、T の 95 % 信頼区間より小さい側として 2.5 % あるので、計 5 % である。

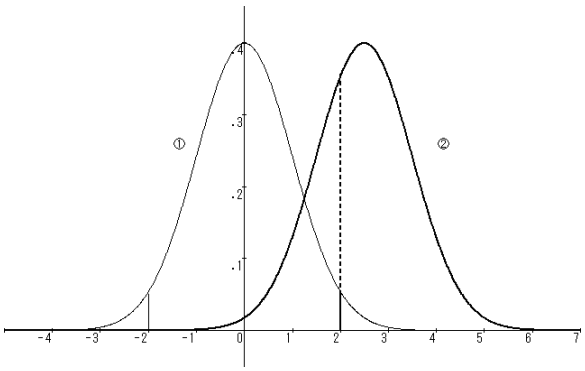
別の説明をしよう。S の真の値が 95 % 信頼区間の限界値である 0 より小さい可能性は 97.5 % ある。T の真の値が 95 % 信頼区間の限界値である 0 より大きい可能性は 97.5 % ある。この２つが同時に成立するなら、S の真の値と T の真の値が同じはずはない。確率は小さいが、逆に、S の真の値が 95 % 信頼区間の限界値である 0 より大きい可能性 2.5 % と T の真の値が 95 % 信頼区間の限界値より小さい可能性 2.5 % が同時に成立しても、S の真の値と T の真の値は同じではあり得ない。これ以外の組み合わせの時には、S の真の値と T の真の値は同じ値である可能性を否定できない。そこで、S の真の値と T の真の値が同じであるはずがないと言ってよい確率を計算すると、$0.975^2 + 0.025^2 = 0.95125$ となる。つまり「（ほぼ）95 % の確実性で S と T の真の値は異なる」と言ってよい※。もし、２つの分布がもっと近づき、従って 95 % 信頼区間が重なってくるような場合には、２つの真の値が異なる確実性は 95 % より下がってしまうことになる。

※ 仮説検定の手順では、有意に差が認められても「ある確実性で差がある」と結論することができなかった。この点で、信頼区間の比較は、仮説検定と異なる。ただし、一般的に（大雑把な言い方をすれば、仮説検定では比較したい数値の間に $2\sigma_m$ の開きがあれば有意差ありとみなしたものが、信頼区間同士の比較をする場合には、推定平均値間に $4\sigma_m$ の開きがないと差があるとみなせないので）、信頼区間同士の比較の方が、集団の間に差があるとみなすためのハードルが高くなる。（これは、ある数値１点を分布に対して比較するのが仮説検定であり、分布と分布を比較するのが信頼区間同士の比較であることに由来し、どの程度の不確実性まで考慮して行われるかの違いでもある。）

◎ 仮説検定の話、ふたたび

もういちど仮説検定の手順について考えることにしよう。ここで、仮説検定による検証仮説は「母集団２が母集団１とでは、平均値が異なる」である。従って、帰無仮説は「母集団２と母集団１は平均値が等しい」である。母集団２から抜き出した標本平均の値（点推定）を用いて、母集団１から抜き出した標本平均の分布（区間推定）と比較するという手順になる。

平均 μ_{01} をもつ母集団１がある場合、この母集団１から独立に抽出された標本の平均 μ_1 の母平均 μ_{01} からのずれ $(\mu_1 - \mu_{01})/\sigma_m$ は、右図の①のような確率分布に従う。



①のグラフは「標本平均の分布」を示す。すなわち、母集団そのものがもつ分布ではない。母集団からある大きさの標本を独立に抜き出した場合の、それぞれの平均値の分布であり、「標本平均の分散」σ_m^2 は、母分散や、１つずつの標本の不偏分散の期待値とは異なり、抜き出した標本の大きさに依存する量である。

ここで、母集団２から抜き出した標本の平均 μ_2 が、$-2\sigma_{m1}$ から $2\sigma_{m1}$ の範囲、すなわち①のグラフの 95 % 信頼区間内にある場合、帰無仮説は棄却されないため有意差なしと判定されることになる。

母集団２の平均 μ_{02} が $(\mu_{02} - \mu_{01})/\sigma_m = 0$ である場合、前提として帰無仮説は正しい。この母集団２から任意に抽出された標本の平均 μ_2 の母平均 μ_{02} からのずれ $(\mu_2 - \mu_{02})/\sigma_m$ は、結果として①と同じ確率分布に従う※。そのため、母集団２から抜き出した標本の平均は、有意水準（危険率）と同じ割合で①のグラフの棄却域に入ることになり、帰無仮説を棄却する。これが第１種の過誤である。

※ いつも成立するわけではないが、ここでは、暗黙のうちに母集団２から抜き出した標本について、

標本平均の分散 σ_{m2}^2 が母集団１からのもの σ_{m1}^2 と同じであることを仮定している。母集団１と母集団２で母分散が同じであり、抜き出した標本の大きさが等しいなら、この仮定は正しい。

次に、母集団２の真の平均が $\mu_{02}/\sigma_m = 2.5$ であった場合（すなわち、帰無仮説は正しくなかった場合）は、任意に抽出された標本平均について μ_2/σ_m の分布は、図の②のようになる。このとき、標本の平均値 μ_2 が棄却域の限界値である $2\sigma_m$ （図中に破線で示した）よりも小さな値をとる場合には、この検定により（誤って）帰無仮説が棄却されないことになる。これが第２種の過誤である。すなわち、第２種の過誤の生じる確率 β は、②のグラフにおいて $2\sigma_m$ より左の部分の面積に等しい。②のグラフも全面積が１なので、$2\sigma_m$ より右の部分の面積は $1-\beta$ であり、（正しく）帰無仮説を棄却する確率にあたる。つまり、この面積が検定力を示している。

母集団２と母集団１の真の平均の差 $(\mu_{02} - \mu_{01})/\sigma_m$ は、いわゆる効果量であるが、これが大きいほど、すなわちグラフ①に比べてグラフ②が右へ移動するほど、検定力（棄却域の限界値より右側の面積）が上がり、第２種の過誤を生じる可能性は小さくなる。グラフがほとんど重ならないほど離れてしまえば、母集団２から抽出した標本の平均は、常にグラフ①の棄却域に入ることになり、第２種の過誤はほとんど生じない。この条件では、検定力は１である※。逆に、効果量が小さい場合には、②は①に近づき、①の信頼区間の中に重なる部分が増え、第２の過誤の割合が上昇する。その極限では、②は①と完全に重なる。その場合には、$1-\beta$ が α に収束する※2。

※ 検定力が１に近づけば、第２種の過誤は減り、偽陰性は生じなくなる。しかし、検定力は、前述のように想定している効果量によって変化する。そして、母集団２の真の平均値は不明で検定が行われる。だから、実際には②のグラフを①のグラフと同じ座標軸に対して書くための情報は不足しており、検定結果（有意差の有無）だけから信頼区間の重なりの程度を推測することはできない。

※2 両側検定では棄却域限界の右側の面積は $\alpha/2$ なのであるが、左側の棄却域を含むので α になる。

○ 標本の大きさと検定力

母集団から取り出ささされた標本平均の分散 σ_m^2 は、その標本の大きさを n とするとき、母分散の 1/n になることを上に示した。また、上の説明で１つの座標系に２つの正規分布曲線を重ねて示しているものは、それぞれ、標本平均の分布である。すなわち、集団１と集団２の真の平均値の間に一定の差つまり絶対的な効果量（$\mu_{02} - \mu_{01}$）があったとしても、検定に用いる標本の大きさが小さいほど、σ_m が大きくなるので、相対的な効果量 $(\mu_{02} - \mu_{01})/\sigma_m$ は小さくなり、検定力が下がる。言い換えれば、実験の結果に事前に想定した効果量の差があっても、標本の大きさが十分ではない場合には、原理的に、その差を有意差として検出できない。逆に、有意差なしと判定した内容も、標本の大きさ不足であったことに由来するのであれば、差がないことを意味しない。すなわち、標本の大きさが本質的に不足しているならば、その実験自体が無意味となる。とはいえ、いつも過剰に大きな標本を用いるべきとも言えない。余分なコストが掛かったりする以外に、検定力が上がりすぎると、本来、検出する必要のない微量の差（実験的には意味のないであろう差）まで有意差として検出されてしまう。そのため、実験に先立って、どのような目的で、どの程度効果量の差を、どんな有意水準で検出したいのかに基づき、標本の大きさや解析法についてあらかじめよく計画を立てておく必要がある。

（参考） http://www.med.akita-u.ac.jp/~doubutu/IACUC/appropriate.html
　動物福祉の観点からみた生物医学研究における適正な動物数
（秋田大学 バイオサイエンス研究サポートセンター 資料）

◎ 正規分布以外の分布

以上、概略を述べるために使用した標本平均が「同じ分散値の」「正規分布に従う」という仮定は、標本の大きさが十分に大きいわけではない場合に、いつも成立するわけではないから、実際の検定にあたっては、自分が検定しようとしている標本の母集団や標本平均の分布の特性に合わせて最も適切な方法をとる必要がある。標本平均が、正規分布ではなくてもなんらかの確率分布に従うことを前提とした検定を、「パラメトリック（Parametric）な検定」とよぶ。正規分布を仮定して説明してきた詳細のいくつかは、正規分布ではない場合、厳密には成り立たない場合もある。しかし、大筋の理解としては、帰無仮説を仮定したときの確率分布を考え、標本の平均値等、着目している統計量が棄却域に入る場合には、有意差ありと判定するという解釈で、問題ないだろう。

代表的な確率分布としては、二項分布、t 分布などがあり、これらを仮定した検定は、二項検定、t 検定などと呼ばれる。パラメトリックな仮説検定を行う場合の帰無仮説の設定の仕方は、正規分布で説明したものと原理的に同じである。また、母集団の確率分布を仮定せずに行うものを「ノンパラメトリック（Non-parametric, または exceedance）な検定」と呼ぶ※。ノンパラメトリックな検定は、標本の大きさが小さく、正しい分布が仮定できないような場合に行われることが多い。

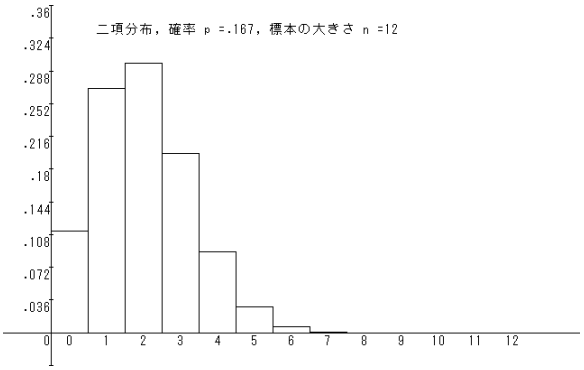
※ ノンパラメトリックな検定も、帰無仮説を設定し、これを棄却できるかどうかを検討する仮説検定であるが、標本の期待値が母集団のもつ確率分布のどの位置にあるのかという観点以外で帰無仮説を設定する点が異なる。たとえばノンパラメトリックな検定の代表的なものに、ウィルコクソンの順位和検定がある。詳細は、別の資料を参照すること。

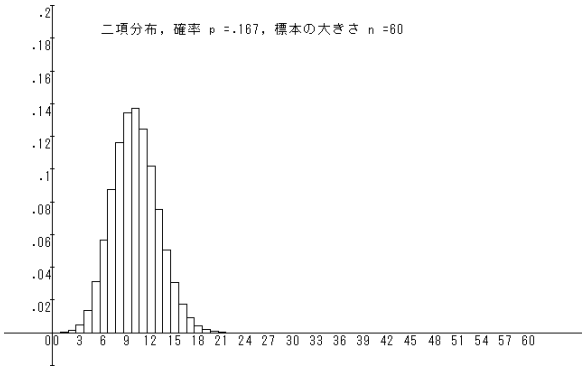
（参考） http://www.weblio.jp/content/パラメトリックな手法とノンパラメトリックな手法

（参考） http://bio-info.biz/statistics/test_wilcoxon_rank_sum.html
　バイオインフォマティックス入門　ウィルコクソンの順位和検定

○ 二項分布

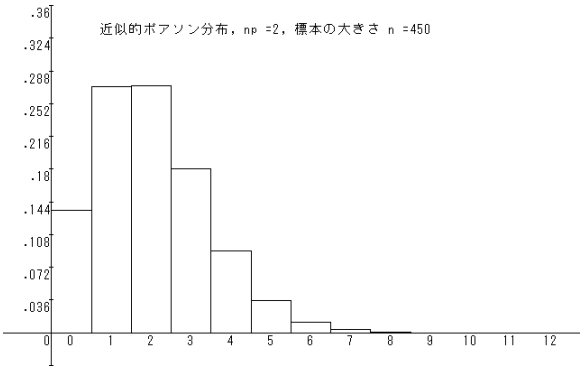
コインの表裏の出現回数、始めに例示したサイコロの特定の出目の出現頻度などは二項分布に従う。確率 1/6 で生じる事象を、試行 12 回、試行 60 回繰り返して、出現頻度ごとの確率（実現数/試行回数）の分布を図示しておく。なお、横軸の出現頻度は、これを試行回数で割れば、標本ごとの標本平均（12 回または 60 回の試行における出現確率）となり、標本の大きさに依存しない値に帰着する。いずれも期待値の頻度（ n = 12 では 2、n = 60 では 10）となる階級を最頻値（モード）としており、n が大きいほど頻度の階級は細分化され連続関数に近づく。なお、標本の大きさ（すなわち試行回数）が大きく（p = 0.5 では、おおむね n > 25 程度以上※）、計算が困難な場合では、正規分布を仮定してもよい（二項分布は、n が大きい極限で正規分布に一致する）。





※ http://www.lbm.go.jp/ohtsuka/envsta/envsta04.html
環境統計学の講義ノート（琵琶湖博物館 大塚泰介氏 による資料）

○ ポアソン分布

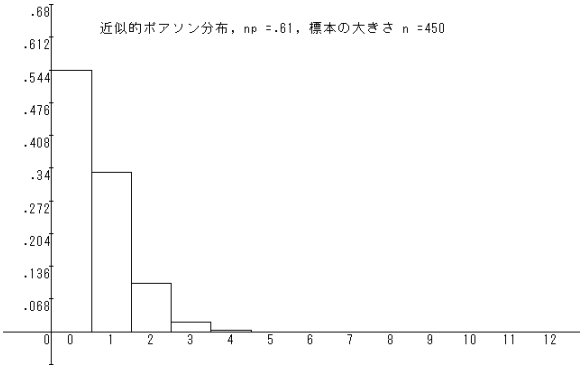


稀にしか起きない現象を長期間計数するなど、低確率の離散的な数値を大量に集計すると、ポアソン分布に従う。右図は λ = np = 2、n = 450 の条件で二項分布により表示した近似的ポアソン分布である。

二項分布は、p を固定して n を増やすと、最後には正規分布に収束する。しかし、n と p の積が一定であれば、二項分布のグラフの形状がほぼ同じになるので、この収束は p に反比例しており、p が小さいほど大きな n が必要となる。ポアソン分布は、二項分布において n と p の積を一定に保つことにより n が小さい場合に見られた形状の非対称性（正規分布からのずれ）を、n が大きくなってもそのまま保っているものと捉えてよい。

ポアソン分布に従う観測事象の歴史的に有名な例として（以下、Wikipedia からの引用※）
統計学者のボルトキーヴィッチによる「プロイセン陸軍で馬に蹴られて死亡した兵士数」の例が知られている。ボルトキーヴィッチは著書 "Das Gesetz der kleinen Zahlen "（The Law of Small Numbers）において、プロイセン陸軍の 14 の騎兵連隊の中で、1875 年から 1894 年にかけての 20 年間で馬に蹴られて死亡する兵士の数について調査しており、1 年間当たりに換算した当該事案の発生件数の分布がパラメータ 0.61 のポアソン分布によく従うことを示している。（引用ここまで）

※ https://ja.wikipedia.org/wiki/ポアソン分布



λ = np = 0.61、n = 450 の条件※で二項分布により表示した近似的ポアソン分布を右図に示す。頻度 0 がおよそ半分強あるから、上の例に当てはめるなら、20 年間のうち 10 から 11 年は該当する事案は発生せず、6 から 7 年は 1 件程度、残りは 2 件乃至まれに 3 件発生することを示す確率分布であることがわかる。なお、右図は、p = 0.61/14、n = 14 の条件で表示した二項分布の図（図示しない）ともほぼ一致する。

※ ここでは二項分布の式から近似して作図したため n を明示しているが、ポアソン分布の形状はひとつのパラメータ λ = np のみに依存し、本質的には n に依存しない。十進 BASIC を用いて計算しているが、その制限で n を 450 より増やすのが難しいのでこの値にしているが、450 という数値に特別な意味はない。ただし、ポアソン分布を二項分布に還元して解釈するならば、n に適当な値を定めて意味を持たせてもよい。プロイセン陸軍の騎兵連隊の所属人数を 10 万人と仮定するなら、$n = 1\times10^5$ として、ある 1 人の騎兵隊員が 1 年以内に馬に蹴られて死ぬ確率が $p = 6.1\times10^{-6}$ とモデル化できる。上のポアソン分布は、この確率での判定を、10 万人分繰り返した結果の二項分布であると解釈される。ある 1 年間に馬に蹴られた死亡者が誰も出ない確率は、$(1-6.1\times10^{-6})^{10^5} = 0.543$ となり、上のグラフとほぼ一致していることがわかる。また、n = 14 とし、p = 0.61/14 = 0.0436 の条件で表示した二項分布として解釈するなら、1 騎兵連隊あたり 1 年以内に平均して 4.36 %の確率で 1 人が死亡

している。この場合に、ある１年間に誰も馬に蹴られた死亡者が出ない確率は、(1-0.0436)^14 = 0.536 である。ただし、後者での解釈は n が十分に大きいとは言えないことによる近似の悪さが誤差の原因となっている。すなわち、後者の二項分布では、１騎兵連隊内から馬に蹴られて死亡する軍人が複数生じることを無視することによる誤差がある。これにより、１年間に 15 件以上の事案が生じる確率が、完全なゼロとなってしまっている（$n = 1\times10^5$ と置いた二項分布ならば、１年間に 1×10^5 件の事案が発生する、すなわち馬に蹴られて騎兵連隊が全滅！する可能性も数学的なゼロではない）。とはいえ、ヒストグラムを見てわかるように、今回の例では１年間に複数の事案が発生すること自体が稀なので、そう悪い近似とは言えないと思われる。

○ ポアソン分布、指数分布と一次反応速度式

ポアソン分布のパラメータ $\lambda = np$ は、ある稀にしか起きない離散事象が単位時間あたりに発生する回数であると表現される。もちろん、この回数は事象が生じる対象の群の大きさ n に依存するのは当然であるが、この対象の群の大きさは経時変化していないとみなしているわけである。

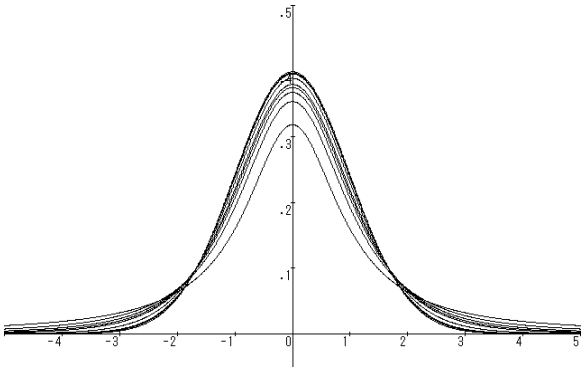
このパラメータ λ は、化学の用語でいうならば、一次反応の反応速度係数 k に相当する量である。いま、励起状態分子 M* があったとしよう※。こいつは、一次反応の速度式に従って減衰する。つまり、輻射、項間交差、内部変換などにより減衰していくのであるが、その単位時間あたりの減少量（単位時間に失活事象の発生する件数）は、（分子の種類、温度や溶媒などの周囲の環境のパラメータを除けば）励起状態分子の濃度のみに依存し、この速度式は、$-d[M^*]/dt = k[M^*]$ と書き表される。もし定常状態にあって $[M^*]$ が一定に保たれるなら、この速度は経時変化せず一定となる。

※ これを二項分布に還元しようとするならば、励起状態分子の失活が「稀に起きる離散事象」に対応する。つまり、励起状態の分子１個が単位時間内に失活する確率を p として（ただしこの単位時間は「離散事象が稀である」「確率 p が十分に小さい」と言えるほど短く設定する）、単位時間ごとに、n = 分子の数（つまり１モルあたり $N_A = 6.02 \times 10^{23}$ 個）の回数の二項判定（M* が失活するかしないか）を繰り返していることを意味する。化学では、分子１個ずつを数える代わりに物質量で数えるのが一般的なので、n を物質量とするなら、p は１モルの分子につき（十分に短い）単位時間内に（1つの分子の）失活が生じる確率（n が $1/N_A$ 倍となる代わりに p は N_A 倍になる。ただしそれでも p が十分に小さいとみなせるように単位時間をとらないと二項分布に還元できない）である。

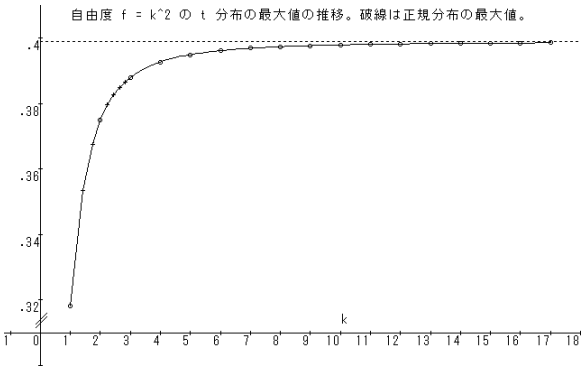
先の式が微分型なので、これを積分し、t=0 における初濃度を $[M^*]_0$ とするなら、ある時刻 t における励起状態分子濃度を表す式は、$[M^*](t) = [M^*]_0 \exp(-kt)$ となり、あらたな供給がない限り、励起状態にある分子は、指数関数的に減少することがわかる。また、ある時間間隔を置いたときに、励起状態分子濃度が以前の半量になる時間を半減期というが、1/2 の代わりに 1/e となる時間を寿命 τ と呼び、$\tau = 1/k$ で与えられる。この指数関数的な減衰は、化学では非常の多数の分子の集合に対しての生存数（というか濃度）で表していることになるが、確率として解釈することも可能である。つまり、１つの励起状態分子が時間 t の経過後に生存している確率 $\exp(-kt)$ に、元の分子の数（濃度）$[M^*]_0$ を乗じたために、その時刻における分子の数（濃度）の期待値になっているわけである。このように指数関数的に減少していく確率分布 $f(t) = \exp(-kt)$ のことを、「指数分布」という。

自然界でこのような減衰をするものには、他にも放射性元素の放射崩壊などが有名であるが、もっと卑近な例として、たとえばある機械の（経年による劣化の蓄積等は無視できて偶発的にある確率でのみ故障すると仮定したときの）購入後に故障するまでの時間のモデルとして、また、周辺環境の経時変化が無視できると考えた場合の店舗等に客が訪れる時間間隔のモデルとして、用いられる。

○ student の t 分布と t 検定



t 分布とは、自由度 f（1 以上の自然数）に依存し、f が大きくなる極限では正規分布に収束するような左右対称型の分布である。右図に f = 1 から f = 8 までの t 分布を示した※。標準正規分布曲線では（標本平均の）標本標準偏差 σ_m に相当する横軸の数字は、t 値と呼ばれる。t = 0 の関数値※2 は、f = 1 で最も小さくなっており、f の増加とともに単調に増加する。代わりにおよそ ±2 より外側で、f が小さいほど関数値が大きくなっているのが特徴である。



※ t-分布の関数定義は、次のページ内の、ガンマ関数に関する性質の解説に基づいて書き下した。
http://www.tamagaki.com/math/Statistics402.html

※2 右にその推移を図示した。横軸は k=√f である。また、点線で標準正規分布における最大値を示した。

t 分布は、信頼区間を算出したり、検定により２つの集団の母平均に差があるかどうかを調べたりする場合に使用する「標本平均の分布」である。ただし、母集団が正規分布に従うことがわかっており、その母分散が未知だが共通の値を持ち、かつ、抜き出した標本の大きさ n が小さい（おおむね 30 未満※）の時に適用する。母分散 σ_0^2 が既知なら、標準誤差 σ_m は、式 $\sigma_m^2 = \sigma_0^2/n$ に基づいて正しく推定できるため、t-分布を考える必要はない。また、n が大きければ、自由度 f も大きくなるので正規分布を仮定してよい。また、例えば n が小さいときに二項分布に従うような例では、母集団が正規分布に従っていないのであるから、 t 分布とそれを仮定した t 検定を用いることは不適切である。

※ http://www.geisya.or.jp/~mwm48961/statistics/sample3.htm
（高校数学に関するウェブ資料集、http://www.geisya.or.jp/~mwm48961/koukou/index_m.htm より。簡潔に要点がまとめられており、見やすい。）

正規分布に従う母集団から「十分に大きく」かつ同じ大きさ n の複数の標本を取り出したとき、標本平均の分布が正規分布に従い、その分散 σ_m^2 は、母分散 σ_0^2 を用いて、$\sigma_m^2 = \sigma_0^2/n$ と表せることをすでに見た。母分散が未知の場合、標本の不偏分散（複数の標本がある場合は、その平均値）σ^2 を母分散の推定値として用いることが可能であった。この推定が可能である理由は、標本ごとの不偏分散は、その期待値が母分散と等しいためである。しかしながら、標本の大きさが小さい場合は、標本平均の分布が正規分布に従うという近似は厳密には正しくない。このような場合に標本平均の分布がどのようになるのかを厳密に調べたものを t 分布という。t 分布を用いた検定における指標である t 値は、正規分布を仮定した場合の標準誤差 SEM に相当する。すなわち、$SEM = \sqrt{\sigma_m^2} = \sqrt{(\sigma_0^2/n)} = \sigma_0/\sqrt{n}$ であった。ただし、ここで σ_0^2 は母分散である。これを標本の不偏分散 σ^2 で置き換えたものが指標としての t 値であり、標本の大きさ n が大きければ、t 値は標準誤差 σ_m に収束する。すなわち、正規分布に従う母集団２から大きさ n の標本を取り出したとき、標本平均が μ_2 であり、標本不偏分散が σ_2^2 であったとき、これを同じ母分散 σ_0^2 を持つと仮定される母集団１の平均 μ_{01} と比較したい場合の指標 t は、$(\mu_2 - \mu_{01})/(\sigma_2/\sqrt{n})$ で与えられる※。このようにして求めた指標である t 値を基準に、その n に対応する自由度 f の分布を調べ、棄却域に含まれるかどうかを判断することになる。

※ σ_0 を σ_2 に置き換えた以外は正規分布の時の指標と同じ形である。例えば、正規分布を仮定した場

合の 95 % 信頼区間の閾値 $\pm 2\sigma_m$ を、σ_m または、($\sigma_0/\sqrt{n}$) で割れば、指標値 2 が得られる。

自由度 f が小さい場合の t 分布は、標準正規分布では棄却域に相当する領域で大きな値を持つから、たとえば 95 % 信頼区間について、正規分布を仮定したときの $\mu_1 \pm 2\sigma_m$ ($\mu_1 \pm 1.96\sigma_m$) は、t 分布においてはもう少し広い範囲を持つことになる。たとえば極端な例を出すと、「自由度 f = 2 の t-分布」を仮定したとき、95 % 信頼区間は、t 分布表※によれば、$\mu_1 \pm 4.3(\sigma_1/\sqrt{n})$ である。また、この信頼区間（または棄却域）の閾値は自由度 f とともに変化する。そのため、２つの標本を比較しようとする際、標本の大きさ（したがって、自由度 f）が大きく異なる場合は、同じ t 値であってもそのデータがどのくらい極端であるのかの基準が異なり、直接の比較ができないので注意しなければならない。自由度ごとの代表的な信頼区間の閾値は、t 分布表を参照するか、t 分布を与える関数を積分※2するか、あるいはエクセルの組み込み関数 TINV などを用いることで知ることができる。

※ たとえば http://www.biwako.shiga-u.ac.jp/sensei/mnaka/ut/tdistinvtab.html
（滋賀大学助手 中川雅央氏による資料。他にも、各種の統計数値表ほか、いろいろな資料が充実している。教材資料のインデックス http://www.biwako.shiga-u.ac.jp/sensei/mnaka/ut/data.html からいろいろ探してみると面白い。）

※2 累積分布関数に数値を代入するか、確率密度関数を数値積分するなどする。数値積分を行う際には、グラフの形状が左右対称なので、数値積分の開始位置を t = 0 としてよい。

○ t-分布における自由度の決め方

t 分布において重要な意味をもつ自由度 f は次のように決める。

１つの集団から抜き出した標本の大きさが n で、この平均値をある値と比較したいときは f = n − 1 である。不偏分散を計算するときの自由度と同じ考え方である。

対応の無い検定を行う場合。たとえば、２つのクラスのテストの平均点に差があったときに、その差が統計的に有意であるかどうかの検定を行う場合のように、それぞれの母集団が無関係である場合。２つの母集団 A と B から、取り出された２つの標本の平均値を比較するとき、標本の大きさをそれぞれ n_A、n_B とするなら、$f = (n_A - 1) + (n_B - 1)$ とする。ただし、t 検定の前提として、２つの標本の大きさがあまり異なっていてはいけない。

対応のある（繰り返しのある）検定を行う場合。たとえば、あるクラスのテストの平均点が１回目と２回目で変化したとき、この差が統計的に有意であるかどうかの検定を行う場合。データの数は 2n 個あるが、n 組の数値の差を調べるのだから、自由度は n − 1 とする。

（参考）http://mizumot.com/handbook/wp-content/uploads/b93393073fa7698302d96237f6f86c63.pdf
（関西大学准教授 水本篤氏による資料；対応の無い検定と、対応のある検定について、エクセルを用いた具体的な解析法が書かれています。）

○ 最小二乗法による回帰直線の決定と、傾き a と切片 b の値に対する t-検定

y のみに誤差を含む n 組の測定値 (x_i, y_i) があるとき、これらの点を通る直線で尤もらしいものを最小二乗法で決めることができる。最小二乗法では、座標上の測定値と回帰直線 y = ax + b との間の y 方向の距離（残差、推定値(ax_i+b)と測定値 y_i の差）の二乗和 $\Sigma[(y_i - (ax_i + b))^2]$ が最小になるよ

うに a と b を定めた。このように求めた a, b の値は、その算出に用いた測定値の組に対して一番尤もらしい推定値となっている。しかし、用いた測定値 y_i のそれぞれには誤差によるばらつきが載っており、そのばらつき方によっては、a, b の値もまた変化したであろう。つまり、a, b の値にも測定値の誤差が伝播しているはずである。従って最小二乗法で求めた a および b の値に対し、どの程度の精度で正しいかを示す必要がある。この目的で傾き a および切片 b それぞれに標準誤差※（SE, σ）を求めることができる。（エクセルの分析ツールを用いれば、簡便に行うことも可能である。）

※ 一般に推定値の標準偏差を標準誤差と呼ぶ。ここでは平均値の標準誤差（SEM）ではないので、添え字として m を用いなかった。

傾きや切片については、ゼロなのかゼロではないのかが関心の対象であることが多い。すなわち、「傾きや切片が小さいがゼロではない」のか、「ゼロであることを否定できない」のかの間には大きな違いがあるだろう。そのため、たとえば最小二乗法で求めた a, b に対しても、標準誤差に基づき信頼区間を算出し、その信頼区間内にゼロという値が含まれるかどうかを調べる必要がある。具体的には、次式に従い、傾きおよび切片の標準誤差を求め、その上で t 検定※を行う。

※ t 値として、a/σ_a および b/σ_b を求め、これを自由度 n−2 の t 分布について、有意水準 5 % での閾値などと比較する。ただし、σ_a は傾き a に対する標準誤差、σ_b は切片 b に対する標準誤差。t 値を比較するべき閾値は、t-分布表の他、エクセル組み込み関数 = TINV(0.05, n−2) からも求めることができる。なお、自由度が n−2 なのは、n 組のデータから、a, b の２つを求めたためである。）

n 個の測定値 y_i を直線 $y = \alpha x + \beta$ の周りに本質的に x によらない母分散 σ_{y0}^2 で正規分布している母集団から抜き出した標本と考える。回帰式 $y = ax + b$ における a, b は α, β の推定値である。残差の分散 σ_y^2 は、残差の二乗和を自由度 n−2 で割ったものであり、y 値の母分散 σ_{y0}^2 の不偏推定量である。σ_y は推定値 y の標準誤差である。また、回帰式 $y = ax + b$ の傾き a、切片 b の値も α, β の周りに正規分布する。x_i, y_i についての平均 $\mu_x = \Sigma[x_i]/n$、$\mu_y = \Sigma[y_i]/n$ は、回帰直線上の点である。また、$S_{xx} = \Sigma[(x_i - \mu_x)^2]$、$S_{yy} = \Sigma[(y_i - \mu_y)^2]$、$S_{xy} = \Sigma[(x_i - \mu_x)(y_i - \mu_y)]$、$\Delta = n\Sigma[x_i^2] - (\Sigma[x_i])^2$ と定義すると、$a = S_{xy}/S_{xx}$、$b = \mu_y - a\mu_x$ であり、各種分散は次のようになる。

残差分散 $\sigma_y^2 = \Sigma[(y_i - (ax_i + b))^2] / (n - 2)$
$= \Sigma[((y_i - \mu_y) - a(x_i - \mu_x))^2] / (n - 2)$
$= (S_{yy} - 2aS_{xy} + a^2S_{xx}) / (n - 2)$
$= (S_{yy} - aS_{xy}) / (n - 2)$

傾き a の分散 $\sigma_a^2 = \sigma_{y0}^2 \times n/\Delta$

切片 b の分散 $\sigma_b^2 = \sigma_{y0}^2 \times \Sigma[x_i^2]/\Delta$

残差分散の１行目から２行目への式変形は、<u>測定値の重心、点(μ_x, μ_y)が回帰直線上にある</u>ことを用いた。ここで、σ_{y0}^2 は一般に未知なので、残差分散 σ_y^2 で置き換えて計算し、代わりに、回帰式の傾き a および 切片 b が（厳密には正規分布ではなく）自由度 n−2 の t-分布に従うものとしてよい。

（参考）http://www.cc.u-ryukyu.ac.jp/~fukami/p0.pdf
（琉球大学教授 深水孝則氏による資料）

（参考）http://www2.mmc.atomi.ac.jp/web13/2014/leastsquares.pdf
（跡見学園女子大学教授 山澤成康氏による資料）

◎ 標本間の比較と種々の検定

ここまでに見てきたいくつかの分布の形に基づいて、点推定された標本平均の値が、帰無仮説の下に形成するはずの確率分布において、どの位置に相当するかを知ることができる。すなわち、この仮説検定は、点推定された値と区間推定との比較に相当する。また、区間推定された標本平均間の差の有無を、信頼区間の重なりの有無で原理的に検定できることを図の上で確認してきた。

次に、区間推定された標本平均間の比較をする方法についてみることにしよう。このような検定を「二標本検定」と呼ぶ。また、χ^2 検定や、F 検定もこの目的で用いることができる。

○ 二標本検定

未知であるが共通の母分散をもつ２つの集団、母集団 A（母分散 σ_{0A}^2、母平均 μ_{0A}）、母集団 B（母分散 σ_{0B}^2、母平均 μ_{0B}）（前提条件より、$\sigma_{0A}^2 = \sigma_{0B}^2 = \sigma_0^2$）からそれぞれ、標本 A（大きさ n_A、標本平均 μ_A）と、標本 B（大きさ n_B、標本平均 μ_B）を抜き出したとき、その標本平均の差（$\mu_B-\mu_A$）がどのような分布に従うかを考える。標本 A、B の標本平均がどのような分布に従うのかはすでに見た通りであり、($\mu_A-\mu_{0A}$) および ($\mu_B-\mu_{0B}$) は正規分布に従い、その標本平均の不偏分散は、$\sigma_{mA}^2 = \sigma_0^2/n_A$、$\sigma_{mB}^2 = \sigma_0^2/n_B$ である。従って、正規分布の性質から、($\mu_B-\mu_{0B}$) − ($\mu_A-\mu_{0A}$) もまた正規分布に従い、その不偏分散は $\sigma_{mB}^2 + \sigma_{mA}^2$ となる。ここで、この標本平均の差の不変分散について整理すると、$\sigma_{mB}^2 + \sigma_{mA}^2 = \sigma_0^2(1/n_B + 1/n_A)$ である。そのため、指標 Z として、

$$Z = ((\mu_B-\mu_{0B}) - (\mu_A-\mu_{0A})) / (\sigma_{mB}^2 + \sigma_{mA}^2)\hat{}0.5$$
$$= ((\mu_B-\mu_A) - (\mu_{0B}-\mu_{0A})) / (\sigma_0(1/n_B + 1/n_A)\hat{}0.5)$$

を計算してやると、これが標準正規分布に従う。分子は、(標本平均の差)と(既知の母平均の差)の差の形になっている。つまり、指標 Z の代わりに $Z' = (\mu_B-\mu_A) / \sigma_0(1/n_B + 1/n_A)\hat{}0.5$ を計算してやると、(既知の母平均の差)を中央値として、（分母によって標準化されているので）分散 1 で正規分布する。得られた２つの標本平均の差が、想定される量と比べて偶然とは言えない程度に大きいときに、このような指標 Z について正規分布を仮定した両側検定を行ってやれば、棄却域に入ることになる。なお、多くの場合母分散 σ_0^2 は未知なので、この推定値として「プールされた不偏分散」※で置き換えることができ、t 検定を行うことになる。その場合の自由度は（$n_A + n_B - 2$）となる。

※ プールされた不偏分散は、標本 A、標本 B の不変分散について、それぞれの自由度で重みづけをした平均したものであり、次式で表される。

$$\sigma^2 = [(n_A-1)\sigma_A^2 + (n_B-1)\sigma_B^2] / [(n_A-1) + (n_B-1)]$$

（参考）http://racco.mikeneko.jp/Kougi/2011a/STAT/2011astat11.pdf
２標本ｔ検定
（関西大学 教授 浅野晃氏による講義資料）

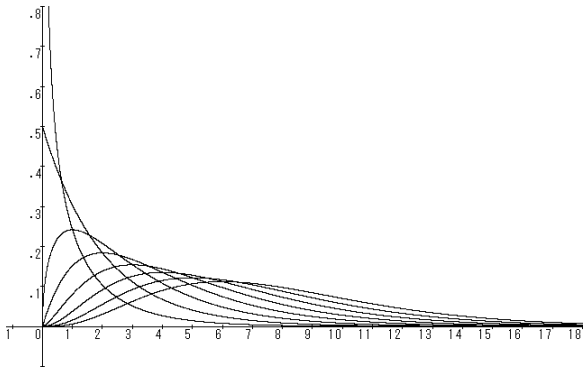
（参考）http://sysplan.nams.kyushu-u.ac.jp/gen/edu/MarineStatistics/2016/index.html
（九州大学 准教授 木村元氏による講義資料。第 11 回、第 12 回など。）

なお、２つの標本平均が等しいとみなせる※かどうかは、この分布の（$\mu_{0B}-\mu_{0A}$）= 0 のケースとして検定できるから、母分散が既知であるならば、指標として $Z' = (\mu_B-\mu_A) / \sigma_0(1/n_B + 1/n_A)\hat{}0.5$ を用いて正規分布による検定を行い、母分散が未知であるならば、$Z' = (\mu_B-\mu_A) / \sigma(1/n_B + 1/n_A)\hat{}0.5$ を用

いて t-検定を行えばよい。（いうまでもなく、後者において σ は、その二乗がプールされた不偏分散であり、自由度は （$n_A + n_B - 2$）となる。

※ 「２つの標本平均が等しいとみなせる」とは、「２つの標本が同じ母平均をもつ母集団から抜き出された標本である」と表現される場合もある。

○ χ^2（カイ二乗）分布



標準正規分布に従う確率変数の２乗がなす確率分布を、自由度１の χ^2 分布という。さらに一般には、標準正規分布に従う k 個の確率変数の２乗の和がなす確率分布は、自由度 k の χ^2 分布に従う。右図に自由度 f = 1 から f = 8 までのカイ二乗分布を示した。

すでに、説明したように、母分散 σ_0^2 をもつ母集団から抜き出した標本平均は、その標本数が十分に多いとき、標本平均の分散 σ_m^2 をもつ正規分布に従う。標本平均の平均は、母平均のよい推計値である。ある標本の平均 μ が、母平均 μ_0 からずれている場合、標準正規分布では、分布の平均から左右のいずれかにずれた位置にあることになり、その偏差 $(\mu-\mu_0)/\sigma_m$（偏差の分母に書かれた σ_m は、標本平均のもつ正規分布を標準化するためである）には、正または負の符号が付く。この偏差を二乗したものは、上の定義に合致するので、自由度１の χ^2 分布に従う。そのためχ^2 分布の最小値は 0 であるが、これはもとの標本平均の偏差 $(\mu-\mu_0)/\sigma_m$ が 0 の場合、すなわち、標本平均が母平均に等しい場合を意味する。もし、標本平均が母平均（や、同じ母分散を持つ別の集団の平均）から極端に離れている領域に棄却域を定めて両側検定する場合、これと対応する検定は、カイ二乗分布においては片側検定となり、右側棄却域を定める。この閾値は、χ^2 分布の関数を積分するか、または χ^2 分布表から求める。エクセルなどでは、組み込み関数 CHIINV を用いることもできる。

ここで「確率変数」とは、偏差 $(\mu-\mu_0)/\sigma_m$ である。この偏差の二乗が自由度１の χ^2 分布に従うのと同様に、同じ母分散をもつ母集団から抜き出した同じ大きさ※ n の（すなわち、標本平均の分散も共通である）標本が k 個あるとき、「k 個の確率変数の２乗の和」すなわち、k 個の標本平均の「偏差の二乗和」 $\Sigma[((\mu-\mu_0)/\sigma_m)^2] = \Sigma[(\mu-\mu_0)^2/\sigma_m^2] = \Sigma[(\mu-\mu_0)^2 n/\sigma_0^2] = \Sigma[n(\mu-\mu_0)^2]/\sigma_0^2$ は、上の定義より自由度 k の χ^2 分布に従う。ただし、母平均 μ_0 が未知であるならその推定値として標本平均の群間平均 μ_{ave} を用いてもよいが、これにより自由度は k ではなく k-1 になる。

※ なお、説明を簡便にするために標本の大きさがすべて等しいという前提を置いたが、χ^2 分布や F 分布を考えるとき、この条件は必ずしも必須ではない。n は、標本平均の分散 σ_m^2 と母分散 σ_0^2 とを関連づけるために必要な量である。具体的に見てみよう。今 k 個の群を、I 群（n_1 の大きさの標本が k_1 個）、II 群（n_2 の大きさの標本が k_2 個）（ただし $k = k_1 + k_2$ ）に再分割することが可能であったとしよう。I 群については、k_1 個の標本平均の「偏差の二乗和」は $\Sigma[((\mu-\mu_0)/\sigma_{m1})^2] = \Sigma[n_1(\mu-\mu_0)^2]/\sigma_0^2$ で、自由度 k_1 の χ^2 分布に従う。また、II 群についても同様に、k_2 個の標本平均の「偏差の二乗和」は $\Sigma[((\mu-\mu_0)/\sigma_{m2})^2] = \Sigma[n_2(\mu-\mu_0)^2]/\sigma_0^2$ で、自由度 k_2 の χ^2 分布に従う。なお、χ^2 分布の性質として、互いに独立な複数の χ^2 分布に従う変数の和は、やはり χ^2 分布に従い、自由度はそれぞれの和になる。つまり、I 群、II 群の標本平均の「偏差の二乗和」の和、$\Sigma[n_1(\mu-\mu_0)^2]/\sigma_0^2 + \Sigma[n_2(\mu-\mu_0)^2]/\sigma_0^2$ は、自由度 $k_1 + k_2 = k$ の χ^2 分布に従う。いま２群に分割したが、これを拡張してやれば、$\Sigma[n(\mu-\mu_0)^2]/\sigma_0^2$（k 個の標本についての偏差の二乗和）が、標本の大きさが同じであることを全く要請せずに自由度 k の χ^2 分布に従うことがわかる。

（参考）　http://www.geisya.or.jp/~mwm48961/statistics/kai2.htm

母平均未知の条件下で２つの標本平均を比較する場合、自由度は f = 2-1 = 1 である。これは、標本 I（大きさ n_1、標本平均 μ_1、標本平均の群内不偏分散 σ_{m1}^2）、標本 II（大きさ n_2、標本平均 μ_2、標本平均の群内不偏分散 σ_{m2}^2）とし、母平均の推定値※として、標本の大きさの平方根で重みをつけた標本平均間の平均 $\mu_{ave} = (n_1^{0.5}\mu_1+n_2^{0.5}\mu_2)/(n_1^{0.5}+n_2^{0.5})$、（図的には μ_1 と μ_2 の間を σ_{m1} 対 σ_{m2} の比で内分した点）を用いれば、$(\mu_1-\mu_{ave})/\sigma_{m1} = (\mu_2-\mu_{ave})/\sigma_{m2}$ となるから、正規分布による仮説検定をする場合に、標本 I と標本 II の検定結果が一致することと対応している。なお、95 % 信頼区間が標本平均に対して $\pm 2\sigma_m$ の範囲であったことを考慮すると、$(\mu_2-\mu_1)$ が $2\sigma_{m1} + 2\sigma_{m2}$ 以上離れている場合、この２つの標本平均は、有意水準 5 % で有意に差があると結論してよい。このことが、指標値が 5 % 水準での χ^2 分布の右側棄却域に入ることと対応する。

※ 母平均の推定値には、一般的には標本の大きさで重みをつけた標本平均の群間平均を用いるようです。この量は、つまり、標本の区切りを取り払った全データについての平均です。

さて、母平均既知のとき、１つの標本平均と母平均※を比較する場合、自由度１であった。これは、上の段落との比較で考えるならば、χ^2 検定では、既知の母平均の信頼区間が、標本平均の信頼区間と同じ程度であることを仮定していることになる。だから、既知の母平均の信頼区間がずっと狭い（高い精度で求められた値である）場合は、少し厳しい検定をしていることになる。しかし実際には、χ^2 検定を、１つの標本平均を（標本平均よりもずっと高い精度で）既知の母平均と比較する目的で使うことは少ない。標本平均の信頼区間がわかった時点で片が付いてしまうからである。おそらく、想定されるのは、複数の標本間の平均値の比較になるだろう。これについては次項で説明する。また、さらにいえば、χ^2 分布を考える上での前提事項である「母分散既知」という条件に当てはまることも現実的には少ないだろうから、複数の標本間の平均の間の比較について、F 検定を行うことになる場合も多いのだが、順を追って見ていくことにしよう。

※ 実験で得た値を、母平均すなわち真の値として、既知の定数と比較する場合もあるだろう。このような定数のうちには、実験的に決められたような量もある。このような値の信頼区間は、数値に SEM が添えられている場合にはその値、添えられていない場合には記載されている最後の桁と同程度と考えるのが妥当であろう。比較の相手が、これら定数の組み合わせなどで算出されるような量である場合には、その信頼区間は、誤差の伝播の公式に基づき求めることになる。

○ χ^2 検定は、３群以上の比較に応用できる。パラメトリックな χ^2 検定

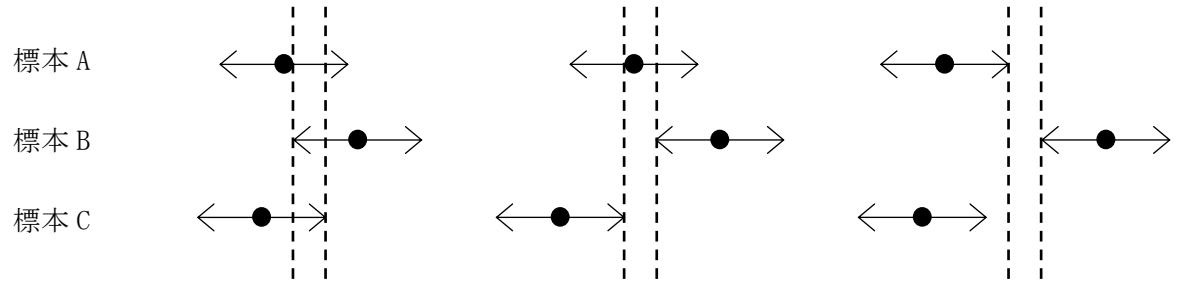
２群の平均値に有意に差があるとみなせることを検定する方法については、上で学んだ。これを A、B、C の３群に当てはめる場合、同じように２群間の比較をすることを考えると、組み合わせの数は、A-B 間、A-C 間、B-C 間の３通りである。もし k 個の群なら k(k-1)/2 通りとなり、群の数の二乗に比例して増える。多重比較による第１種の過誤の増加も見逃せないし、手間の増加も無視できない。

実は、これを１回の検定で判定することができる。つまり、同じ母分散 σ_0^2 もつ母集団から抽出された、同じ※大きさ n の３群以上（A、B、C…）の標本があるとき（標本の数を k とする）「これらの標本平均が同じとみなせる値である」「$\mu_A = \mu_B = \mu_C$ …」を帰無仮説として、自由度 k-1 の χ^2 検定をすることができる。いま、帰無仮説の下では、k 個の標本平均には、母平均から明らかに離れた無関係な値を含んでいないことが保証されるので、これらの標本平均は、標本平均の分散※2 $\sigma_m^2 = \sigma_0^2/n$ で、母平均 μ_0 の周りに正規分布するはずである。そのために $\Sigma[(\mu-\mu_0)^2]/\sigma_m^2 = \Sigma[n(\mu-\mu_0)^2]/\sigma_0^2$ は χ^2 分布に従うのである。

※ 標本の大きさが同じであることは、実は必須ではない。上の項での議論を参照。

※2 考えている意味から、標本平均の群間の不変分散であるが、帰無仮説の下では定義的に群内不偏分散と等しい。

３つ以上ある群（標本）における関係を、次図において整理しておこう。ここでは、いずれも標本 A、B、C の３群で比較をしようとする場合について、標本平均のばらつき（群間のばらつき）の度合いにより、３つの場合にわけて例示した。また、黒丸で標本平均の位置を示した。また矢印の範囲で示されているのが信頼区間（±2SEM など）である（標本そのもの分散の範囲（±2SD など）を示すものではない）。



群間ばらつき：	小	中	大
全群の重なり：	有	無	無
孤立した群　：	無	無	有

話を簡単にするために標本の大きさをすべて同じとすると、χ^2 検定の指標値は、$\Sigma[n(\mu-\mu_0)^2]/\sigma_0^2 = \Sigma[(\mu-\mu_0)^2]/\sigma_m^2$ である。上図における横軸は、標本平均の推定母平均との偏差 $(\mu-\mu_0)/\sigma_m$ に相当する量であり、σ_m で標準化されている（χ^2 分布の項の２段落目参照）。すなわち、χ^2 検定の指標値は、標本平均の推定値からの偏差に対し、二乗和をとったものである。なお、これを標本の数の自由度で割って、標本平均の群間の不変分散として表し、標本平均の群内の不偏分散と比較するのが F 値に相当する（後述）。

上図の左の例では、３標本の信頼区間に重なりがある※。そのため、これら３群が同じ母平均をもつ母集団から抽出された標本である可能性を否定できない。すなわち、帰無仮説（３標本の平均値が同じである）を棄却することができないため、定めた有意水準の下に３標本の平均値の間に有意差がない。また、すべての標本に共通した信頼区間の重なりがあるのだから、A-B 間、A-C 間、B-C 間で下位検定しても、いずれも帰無仮説は棄却できないだろう。これに対し、中央、または右の例では、信頼区間に重なりのない標本の組が存在する。標本 B は、標本 C と（右例では標本 A とも）標本平均が等しいとはみなせないことが、図から判断できるだろう。すなわち、帰無仮説（３標本の平均値が同じである）は棄却される※2 ことになる。

※ 標本の大きさにより、信頼区間の幅は変化するので、「最大の平均値をもつ標本と、最小の平均値をもつ標本の信頼区間の重なりだけで完全に判断できる」と考えるわけにはいかない。

※2 図から判断するのとは異なり、指標値に従って帰無仮説が棄却された場合は、「k 個の標本の平均がすべて同じであるとは見なせない」だけであり、どの標本平均が離れた値であるのかについては、判定できない。その場合には、少なくとも一つ以上あることが判明した、平均値が極端値である標本がどれであるのかを、別の方法、たとえば下位検定としての多重比較（群を分割しての χ^2 検定や、二群間の比較）に頼ることになる。ただし、母平均未知の条件で χ^2 検定を行う場合は、k 個の群から k

より小さい別の群を抜き出すと、母平均の推定制度が悪くなるため検定の精度が落ちることにも注意が必要である。

具体的には、まず $\Sigma[n(\mu-\mu_0)^2]/\sigma_0^2$ を計算する。n は標本の大きさ、μ はそれぞれの標本平均であり、Σ は群（標本）の個数 k に亘っての合計をとるという意味である。なお、σ_0^2 は既知の母分散、μ_0 は母平均である。母平均が未知の場合は、適切な推定値（ただし、母平均の推定値に対しても μ_0 の記号を使い続ける場合がある）を用い、自由度を k ではなく k-1 とする。k 個の標本平均が、数値が母平均と完全に一致しているなら（そもそも、これが確認できた時点で検定が必要とは思えないが）偏差の二乗和はゼロである。標本平均にはばらつきがあるなら、偏差の二乗和は、ゼロよりも大きい値をとるはずである。この値が、χ^2 分布表などから確認した棄却域（偏差の二乗和なので、片側に設ける）に入った場合、（有意水準以下で生じる偶然の結果である場合を除いて）母平均（またはその推定値）から明らかに離れたデータが少なくとも一つはあったとみなされる。すなわち、帰無仮説が棄却されることになる。

次項での F 分布では、このような図的関係を、信頼区間の幅を標本平均の群内不偏分散と読み変え、これが、標本平均がどのように分布しているのか（群間の不偏分散）との比で判定する内容であることを見ていく。

（参考）　http://www012.upp.so-net.ne.jp/doi/biostat/CT39/distribution.pdf
　正規分布・t 分布・χ^2 分布・F 分布とは何か？
　（土居正明氏による資料、χ^2 分布・次項の F 分布ともに、この資料の説明がわかりやすい。）

○ F 分布

上で述べた χ^2 分布に従う指標 $\Sigma[n(\mu-\mu_0)^2]/\sigma_0^2$ を計算するためには、母分散 σ_0^2 が既知であることが必要であるが、近似的には標本不偏分散の群間の※平均値 σ^2_{ave} で推定できる※2。この推定値で置き換えられた新たな指標は、特に標本の数 k が小さい場合、χ^2 分布に従うということがよい近似ではなくなる。この $F = \Sigma[n(\mu-\mu_0)^2]/\sigma^2_{ave}$ （を自由度で割ったもの※3）は、F 分布に従う※4。また、指標値として、F = [群間変動の不偏分散]/[群内変動の不偏分散] を用いる※5場合もある（後述）。

※ ここで、群間とは、扱っている複数の標本間の比較であることを示す。これに対して、群内とは一つの標本内でのという意味である。

※2 ここでは、標本の不偏分散についての k 個の平均であり、期待値ではないため、母分散の推定値でしかない。母分散の推定値として、標本不偏分散を用いようとする場合、$\Sigma[n(\mu-\mu_0)^2/\sigma^2]$ （不偏分散は、Σ で和をとる前の各項に含まれていることに注意）と $\Sigma[n(\mu-\mu_0)^2/]\sigma^2_{ave}$ は一致しないことに注意する。具体的な数字を使って言うならば、2 と 4 の平均は 3 であっても、1/2 + 1/4 は 1/3 の２倍には等しくないと言うことと同じである。しかし後者の方がよりよい近似であることが知られている。

※3 正しい指標値は、χ^2 検定の指標の母分散を推定値に置き換えただけの $\Sigma[n(\mu-\mu_{ave})^2]/\sigma^2_{ave}$ ではなく、これを自由度（k-1）で割ったものである。導出の詳細は、後述。もし、標本の大きさがすべて同じであるならば、χ^2 分布で扱った指標は、$\Sigma[n(\mu-\mu_0)^2]/\sigma_0^2 = \Sigma[(\mu-\mu_0)^2]n/\sigma_0^2 = \Sigma[(\mu-\mu_0)^2]/\sigma_m^2 \div \Sigma[(\mu-\mu_0)/\sigma_m)^2]$ であった（最後を = ではなく ÷ で結んでいるのは、σ_m^2 が、実際には標本平均の不偏分散の期待値（したがって、厳密に $\sigma_m^2 = \sigma_0^2/n$。母分散が既知の場合に算出される）ではなく、標本ごとに異なる値をもつ不偏分散 σ^2 に基づき算出されるためである。これを自由度 k-1 で割るという意味は、定性的には、群間の標本平均の不変分散に相当する量であることを示す。とはいえ、後述の議論からもわかるように、F 分布における指標は、必ず不偏分散の比の形をしている。

※4 正規分布している母集団から大きさ n の小さい標本を抜き出したとき、その標本の分布が厳密には正規分布には従わなかった。この分布を厳密に表現したものが t-分布であった。t-分布は、標本の大きさ n に対し、自由度 n-1 というパラメータが必要となった。このことと、状況は似ている。母分散が未知なので不偏推定量で代替したために、標本の数 k が小さいときに厳密には χ^2 分布に従わなくなった分布について、これを厳密に表現したものが F 分布である。F 分布では、標本の数 k に対し（母平均が未知ならば、母平均の代わりに標本平均の群間平均 μ_{ave} を用いるため、k ではなく）自由度 k-1 というパラメータを新たに導入する。もとより χ^2 分布は１つのパラメータとして自由度 n-1 を要求していた。その結果、F 分布ではパラメータとして２つの自由度（k-1, Σ(n-1)）を指定する必要がある。ここで、Σ は k 個の標本に亘る和であるから、すべての標本で大きさが同じであるならば Σ(n-1) = k(n-1) である。

※5 同一条件下におけるデータのばらつき（意味的に SEM ではなく SD を基準とするもの）は、偶発的な要因による誤差により生じる。これは標本を大きくして平均をとると打ち消すことが可能なばらつきである。これに対し、実験条件を変えた標本間には、誤差によるばらつき以外に、何らかの要因による系統的な変化がみられる。これは、標本平均の差として現れる。これが群間変動である。群内変動は、偶発的な誤差によるばらつきにより生じるが、標本の大きさを考慮するということは、 σ の代わりに σ_m で規格化し、単に「標本内の分散」ではなく「標本平均の群内不偏分散」としていることを意味する（原理的に、後者は信頼区間による議論の意味合いをもつ）。なお、この群内変動（偶発的な要因による標本平均のばらつき）をノイズと表現し、実験条件を変えた標本間に見られる系統的な変化（ばらつきを取り除いたあとの変化）を要因（実験条件の変化）による出力、すなわちシグナルと読むならば、この指標値 F が S/N 比そのものを表していることがわかる。観測されたばらつきを群内変動とそれ以外の要因によるものとに分解していく手法を分散分析と呼ぶ。詳細は後述。

F 分布の定義は、次のようになる。一般に２つの独立なカイ二乗変数 χ^2_1、χ^2_2 があり、それぞれの自由度を f_1、f_2 とするとき、これら二変数の比 $F = (\chi^2_1/f_1) / (\chi^2_2/f_2)$ が従う分布を、自由度(f_1, f_2)の F 分布と呼ぶ※。

※ この定義からも明らかなように、$F = (\chi^2_1/f_1) / (\chi^2_2/f_2)$ が自由度(f_1, f_2)の F 分布に従うならば、分母と分子を入れ替えた $F^{-1} = (\chi^2_2/f_2) / (\chi^2_1/f_1)$ もまた自由度(f_2, f_1)の F 分布に従う。すでにお気づきのように、カイ二乗変数である偏差の平方和を自由度で割るので、分子、分母ともに不偏分散の形になっている。

○ F 分布に従う１つめの指標値

まず、上に示した $F = \Sigma[n(\mu-\mu_0)^2]/\sigma^2_{ave}/(k-1)$ について、定義に沿って F 分布に従うことを順に確認していこう。

1) すでに見たように、$\Sigma[(\mu-\mu_0)^2/\sigma_m^2] = \Sigma[n(\mu-\mu_0)^2]/\sigma_0^2$ は（母平均が未知なので）自由度 k-1 で χ^2 分布に従うことがわかっている。ただし、σ_m^2 は σ_0^2/n により求めた群内の標本平均の不変分散である。これを C_0 とし、（実在するが、値が未知であるために指標値の計算に用いることのできない）母分散を、標本不偏分散の群間平均 σ^2_{ave} で置き換えたものを C と置くことにする。すなわち、$C = \Sigma[n(\mu-\mu_0)^2]/\sigma^2_{ave} = C_0 \times(\sigma^2_{ave}/\sigma_0^2)^{-1}$ である。

2) $\Sigma[n-1]\times\sigma^2_{ave}/\sigma_0^2$ は、自由度 Σ[n-1] で χ^2 分布に従う。ただし、この Σ は標本 k 個に亘っての和であることを示す。従って、もし k 個の標本の大きさが同であるなら、Σ[n-1] = k(n-1) である。また、σ^2_{ave} は標本不偏分散の群間平均である。導出は以下の通り。はじめに k 個の標本の大きさが同

じであるものとして考え、のちに標本の大きさが同じでない場合に一般化することにする。共通の母集団（正規分布に従い、母分散 σ_0^2、母平均 μ_0）から抜き出した、k 個の標本（共通の大きさ n、標本平均それぞれ μ）について、標本の各要素の値を x で表すことにする。そうすると、群内の標本不偏分散は、$\sigma^2 = (1/(n-1))\Sigma[(x-\mu)^2]$ である。これを k 個にわたって平均すると、標本不偏分散の群間平均は $\sigma^2_{ave} = (1/k(n-1))\Sigma\Sigma[(x-\mu)^2]$ となる。ただし、１つ目の Σ は標本の数について 1 から k の範囲の和、２つ目の Σ は標本ごとに要素の数 1 から n の範囲の和である。この式に k(n-1) を乗じたもの $k(n-1)\sigma^2_{ave} = \Sigma\Sigma[(x-\mu)^2]$ について、式の両辺を母分散 σ_0^2 で割ると、$k(n-1)\times\sigma^2_{ave}/\sigma_0^2 = \Sigma\Sigma[(x-\mu)^2]/\sigma_0^2 = \Sigma\Sigma[((x-\mu)/\sigma_0)^2]$ は、各要素の群内標本平均との偏差（x-μ）を母標準偏差 σ_0 で割って標準化したことになり、標準正規分布に従う母集団から抜き出した x についての偏差の二乗の和とみなすことができて、定義から χ^2 分布に従い、自由度は ΣΣ として和をとる偏差平方の数、つまり k(n-1) である。さて、実は k 個の標本の大きさが同じである必要が無いことを示すために、χ^2 変数どうしの和がやはり χ^2 変数になるという性質を用いる。すなわち、χ^2 分布の項で展開したように、k 個の群を、それぞれ同じ大きさの標本からなる２つの群、I 群（n_1 の大きさの標本が k_1 個）、II 群（n_2 の大きさの標本が k_2 個）（ただし $k = k_1 + k_2$ ）に再分割することが可能だったとする。I 群、II 群内ではそれぞれ大きさが揃っているのだから、上の議論に従うと、$k_1(n_1-1)\times\sigma^2_{ave}/\sigma_0^2$ および $k_2(n_2-1)\times\sigma^2_{ave}/\sigma_0^2$ は、ともに χ^2 分布に従い、自由度は、それぞれ $k_1(n_1-1)$、$k_2(n_2-1)$ となる。この２つの和は、$k_1(n_1-1)\times\sigma^2_{ave}/\sigma_0^2 + k_2(n_2-1)\times\sigma^2_{ave}/\sigma_0^2 = [k_1(n_1-1) + k_2(n_2-1)]\times\sigma^2_{ave}/\sigma_0^2$ で、これも χ^2 分布に従い、自由度もそれぞれの和となるので、$[k_1(n_1-1) + k_2(n_2-1)]$ である。さらに一般化を進めれば、$[k_1(n_1-1) + k_2(n_2-1)]$ は $\Sigma[n-1]$ になるので、この段落はじめの結論になる。

3) ２つの（独立である）カイ二乗変数があり、それぞれの自由度がわかったので、F 分布の定義に従うように立式してみる。ただし、分子に 1) の C_0 が、分母に 2) の $\Sigma[n-1]\sigma^2_{ave}/\sigma_0^2$ が来るようにしよう。分子分母ともに、それぞれに自由度で割ったものにしないといけないので、この指標値 F の分子は $C_0/(k-1)$ である。今、1) で見たように、$C_0 = C\times(\sigma^2_{ave}/\sigma_0^2)$ であったから、分子は、$C_0/(k-1) = C\times(\sigma^2_{ave}/\sigma_0^2)/(k-1)$ である。また、分母は $(\sigma^2_{ave}/\sigma_0^2)$ である。分子、分母の共通の $(\sigma^2_{ave}/\sigma_0^2)$ は約分されて、$F = C/(k-1)$ となる。つまり、指標 $F = \Sigma[n(\mu-\mu_0)^2]/\sigma^2_{ave}/(k-1)$ は、F 分布に従い、その時の自由度は $(k-1, \Sigma[n-1])$ である。（ただし、母平均 μ_0 は、未知なので推定値を用いる。標本の大きさで重みをつけた標本平均の群間平均など。この量は、全標本の群を取り払ってしまった全体に対しての平均とも等しい。）

このように分数の形にして計算してやると、共通の母分散（未知だが固定値）を約分して、式の上から消してしまえるので、指標値を計算可能な量とすることができる。なお、指標 F の式について、分母の順を入れ替えてやると、$F = [(\Sigma[n(\mu-\mu_0)^2]/(k-1))]/\sigma^2_{ave}$ の形に書きなおすことができる。

○ F 分布に従う２つめの指標値、 F = [群間変動の不偏分散]/[群内変動の不偏分散]

F 分布に従う指標値は、他にも選ぶことができる。上項の説明を、別の視点でもういちど見直してみよう。上の 1) では「$\Sigma[n(\mu-\mu_0)^2]/\sigma_0^2$ は、自由度 k-1 で χ^2 分布に従う」、2) では「$\Sigma\Sigma[(x-\mu)^2]/\sigma_0^2$ は、自由度 $\Sigma[n-1]$ で χ^2 分布に従う。」という結論を得ていた。ただし、μ は（群内の）標本平均、μ_0 は母平均の推定値である。F 値の定義に従えば、これらをそれぞれの自由度で割った値の比もまた、F 分布に従う。すなわち、 $F = (\Sigma[n(\mu-\mu_0)^2]/(k-1)) / (\Sigma\Sigma[(x-\mu)^2]/\Sigma[n-1])$ もまた、自由度 $(k-1, \Sigma[n-1])$ で F 分布に従う。

はじめに示した指標値 F と別に立式したように見えて、実は、この２つめの指標ははじめに示した指標と同一である。２つめの指標における式の分母 $\Sigma\Sigma[(x-\mu)^2]/\Sigma[n-1]$ は、すべての標本の大きさが n で共通の場合には、$\Sigma\Sigma[(x-\mu)^2]/(k(n-1)) = \Sigma[\Sigma[(x-\mu)^2]/(n-1)]/k$ と書きなおすことができ、標本それぞれの群内の標本不偏分散 $\Sigma[(x-\mu)^2]/(n-1)$ について、k 個の標本に亘って足して、k で割ってい

るから、標本不偏分散の群間平均(σ^2_{ave})となることからも理解できるだろう。$\Sigma\Sigma[(x-\mu)^2]/\Sigma[n-1]$ を「（標本の）群内変動の不偏分散」という。また自由度で割る前の、誤差・ばらつきによる変動の平方和である $\Sigma\Sigma[(x-\mu)^2]$ を「群内平方和」という。

また、分子 $\Sigma[n(\mu-\mu_0)^2]/(k-1)$ は、k 個の標本についての標本平均が示す不偏分散 $\Sigma[(\mu-\mu_0)^2]/(k-1)$ に、標本ごとの大きさ n で重みをつけた形をしている※。この分子の式を「（標本平均の）群間変動の不偏分散」という。群（標本）ごとに変えた実験における要因、独立変数が、結果にどのように差を与えたのかの指標である。また、$\Sigma[n(\mu-\mu_0)^2]$ の部分を標本平均の偏差の「群間平方和」という。

※ この n は、式の導出の途中では、標本平均の信頼区間を表すために必要な標本平均の分散 σ_m^2 を、分母と共通の母分散 σ_0^2 に変換して約分するために必要な係数であった（$\Sigma[(\mu-\mu_0)^2]/\sigma_m^2 = \Sigma[n(\mu-\mu_0)^2]/\sigma_0^2$）。すなわち、この n による重みづけは、分子の標本平均の変動を分母と同じ標本変動に変換して比をとれるようにするために必須である。（逆に、分子に対する重みづけではなく、分母を n で割っているものと捉え、分子分母ともに標本平均の変動としてから比をとると考えてもよい。この方が標本平均のもつ信頼区間として考えたときにわかりやすい。）

このように、F 検定では、不偏分散の比を指標として片側検定を行う。群間でのばらつき（分子）の方が分内でのばらつき（分母）より大きいならば、群間の差が有意となる。すなわち、指標 F 値が大きい値をとるほど（帰無仮説の下では、滅多におきないような、p 値の小さい）標本の分布であることがわかる。従って、上記で求めた指標値 F 値が、F 分布表から確認して、上側棄却域に含まれるならば、帰無仮説（k 個の標本平均は、ばらつきの範囲内で同じであるとみなしてよい）が棄却される。

○ F 分布に従う３つ目の指標値、F 分布による等分散性の検定

F 分布の定義では、分子、分母に、χ^2 分布に従う変数であればどんなものを選んでもよい。χ^2 分布とは、標準正規分布に従う母集団から抜き出した n 個の変数の二乗の和であった。ある母集団から抜き出した n 個の要素をもつ標本 I について、不偏分散は $\sigma_1^2 = \Sigma[(x-\mu_1)^2]/(n_1-1)$ なので、この不偏分散と母分散の比に自由度を掛けたものは、$\sigma_1^2/\sigma_0^2\times(n_1-1) = \Sigma[((x-\mu_1)/\sigma_0)^2]$ のように表すことができ $\mu_1 = \mu_0$ の条件の下で χ^2 分布に従う。共通の母集団から抜き出した別の標本 II についても同様に、$\sigma_2^2/\sigma_0^2\times(n_2-1) = \Sigma[((x-\mu_2)/\sigma_0)^2]$ は、$\mu_2 = \mu_0$ の条件の下で χ^2 分布に従う。ここで、これら２つの χ^2 変数を自由度で割ったものの比は F 分布に従うから、$(\sigma_1^2/\sigma_0^2)/(\sigma_2^2/\sigma_0^2) = \sigma_1^2/\sigma_2^2$ は、$\mu_1 = \mu_2 = \mu_0$ のとき、自由度$(n_1-1,\ n_2-1)$ で F 分布に従う。つまり、共通の母分散をもつ母集団から抜き出された２つの標本において※、「標本不偏分散の比が F 分布に従う」。これにより、この比は、標本不偏分散に差が見られたとき、その偏りが、どの程度の確率で起こり得るのかを示す指標となっている。すなわち、等分散性（２つの標本の分散が等しいとみなせるかどうか）の検定となる。「２つの標本の不偏分散の比」は、不偏分散が一致していれば１であり、いつも１以上になるように、分子に、より大きな不偏分散を充てることにすると、不偏分散に差があるほど、この指標が大きな値をとることになるため、棄却域は右側にのみ定めればよい。従ってこの検定も片側検定となる。

※ 条件「標本平均がそれぞれ母平均に等しいとみなせる場合」が満たされる場合に成り立つ。しかし、「標本 I、標本 II が同じ分散で正規分布しているとみなしてよい」ならば、この条件は不要になる。すなわち、標本 I の不偏分散は $\sigma_1^2 = \Sigma[(x-\mu_1)^2]/(n_1-1)$ であるが、標本平均にかかわらず、これを母分散ではなく２つの標本に共通の分散 σ_s^2 で標準化できるからである。なお、この第２の条件は、共通の母集団から抜き出された「標本 I、標本 II の大きさが十分に大きい」ならば成り立ちそうである。逆に、標本の大きさが十分に大きくなければ、標本は正規分布ではなく、t 分布に従うと考えるべきかもしれない。その場合は厳密に F 分布に従うのではなく、近似的に従うことになる。

○ 分散分析（ANOVA）の考え方（の入り口の手前）

ここでは分散分析のすべてをまとめることはできないので、分散分析とはどのようなことを指すのかを示し、分散分析について調べるときに出てくる用語などについて概観しておくことにしよう。さらに詳細については、成書等を参照してほしい。

分散分析（ぶんさんぶんせき、英: analysis of variance、略称: ANOVA）は、観測データにおける変動を「誤差変動」と「要因による変動」および「各要因の交互作用による変動」に分解することによって、「要因」および「交互作用」の効果を判定する※統計的仮説検定の一手法であると位置づけられる。ここで「変動」とは「（偏差の）平方和」（= 標本不偏分散×自由度）を意味する。

※ もっと簡単な言葉に言い換えると、条件を変えて実験をしたときに見られた平均値の差が、条件を変えたことに由来すると判断するのが妥当なのか、偶発的なばらつきによる結果であると判断すべきかどうかを検定しようということである。また、複数の条件を変えた場合には、どの条件が効果的であったかを知るということでもある。だから単純な話「条件を変えて実験しても、それぞれの群において平均値は同じである（要因による効果がない）」を帰無仮説として仮説検定を行うことになる。ただし、この「要因」（または因子、factor）は、検定結果の解釈において「原因」を意味する言葉ではないことに注意。実験方法によっては、因果関係であると解釈される場合もあるかもしれない。しかし、分散分析やＦ検定は、差や相関関係の有無を判定するだけであり、検定結果そのものが因果関係を示唆することはない。

成書や他のウェブ資料等を参照する際に、混乱しないようにあらかじめ注意しておくと、分散分析を扱う際、「標本」が意味するものがあいまいであることがある。実験条件を変えたときに結果に差があるかどうかを見たいならば、実験条件ごとに標本があり、標本間に差があるかどうかを見るというのがこれまでの扱い方であった。しかし、場合によっては、比較したい実験条件も含めてすべての実験結果をまとめて１つの標本として扱ってしまう場合もある。そこで、なるべく「標本」の用語に依存しない言い換えなどが行われているようである※。たとえば、後述するように、実験の条件は「要因と、その水準」で表されるから、実験条件の同じ実験結果の一群を「同じ水準の群」として扱い、今まで標本平均と読んでいたものは「水準平均」、標本間の差は「水準間の差」と表現するなど。

※ この理由の部分は、単なる推測である。

標本不偏分散の推定値は母分散の不偏推定量であった。すなわち、標本不偏分散の期待値が母分散と一致した。ここで、標本ひとつずつについてみてやれば、当然のことながらその不偏分散はばらついている。そこで、一つずつの標本について、その不偏分散（平方和を自由度で割ったもの）のことを、その標本の「平均平方」という用語で言い換える。

「要因」または「因子」とは、結果になんらかの影響を及ぼすだろうと考えている事柄である。それぞれの要因について、適用される条件の有無、強さなどによる区分、種類の違いなど、要因を構成する条件を「水準」という。比較して分析する以上、ひとつの要因は、少なくとも２水準以上で構成される。たとえば要因（投薬）について、水準（投薬の有、無。または真薬と偽薬）、また、たとえば飲み物について、要因Ａ（温度）の水準として（冷たい、室温、熱い）、要因Ｂ（種類）の水準として（水、日本茶、紅茶、コーヒー）、要因Ｃ（量）、要因Ｄ（頻度）、要因Ｅ（時間帯）…のように、さまざまな要因に対し、いろいろな基準で水準を決めることができる。すべての要因やすべての水準を尽くすことは不可能であるし、あまり意味がないことのように思われる。あらかじめ実験結果に影響を与えるであろう要因について適切な水準を選ぶことが必要となる。そして、実験においては、意図して変化させる以外の要因については、完全にそろえるか、あるいは平均化処理によって打ち消せる

ように選ぶなどが重要となる※。

※ いうまでもないが、「冷たい水」と「熱いコーヒー」とで実験結果に差がでたとき、温度要因が効いたのか種類要因が効いたのか（または交互作用によるものなのか）を区別できない。

同じ実験条件（要因の水準）における実験結果を一つの群（標本）として扱う※と、実験で得られた数値等は、たとえ実験条件（要因の水準）がそろっていてもばらつく。これを「誤差変動」という。これは群内（標本内）の変動である。標本の大きさを考慮すると（同一の母集団から取り出した、つまり同じ実験条件で繰り返したときに得られるだろう）標本平均の不偏分散に換算することができる。これが群内変動の不変分散である。またもし、要因（実験条件）が実験結果に大きく影響を与えるならば、群間（標本間）で数値が系統的に※2 変化するだろう。この群（標本）ごとの違い（要因の水準による違い）が群間の変動である。分散分析では、まず、データごとの平均からのずれ、すなわち変動の大きさ（すなわち平方和の数値）を、この２つの変動に振り分ける※3 ような処理を行う。続いて F = [群間変動の不偏分散]/[群内変動の不偏分散] を指標として片側検定を行うことになる。

※ 群（標本）の大きさにあたるものは、「測定の繰り返し数」などと呼ばれることもある。

※2 ここで「系統的」と言っているのは、平均をとるなどにより打ち消すことができないという意味。

※3 ずれの直接の指標としては、偏差の平方和（= 不偏分散×自由度）を用いる。分散（×自由度）を分解していって実験結果を分析するということから、分散分析と呼ばれる。なお、もし要因数が１ならば、単純に F 検定（特に２群間の比較をするだけであるなら「二標本検定」でもよい）を行うだけで済ませることが可能である。この場合、帰無仮説（要因の水準の異なる標本間で、標本平均が等しいとみなせる）が棄却された場合、要因が実験結果に影響していると結論できる。要因数が多い場合、比較する標本数が複数であるような場合には分散分析の考え方に沿った分析が必須になる。複数の要因がある場合は、要因 A による変動、要因 B による変動、要因 A−B 間の交互作用による変動…、とさらに分割することになる（後述）。交互作用まで含めると、要因が増えるほど複雑は分析が必要になるし、解析の精度も落ちるだろう。計画段階で、一つの実験の中に多くの要因を盛り込まないようにするべきである。

標本のデータに要因がどのように働いているのを「構造」という。１つの要因しか考えない場合は、「一元配置」の構造を考える。すなわち、実験条件の違いは、一次元の表で表され、実験結果は $y = ax + b$ のような線形関係となるだろう。ここで、a は要因による効果の有無や大きさを表し、x が要因の水準である。切片 b は要因影響がない場合の出力に相当する。また、２つの要因を考えているならば、「二元配置」の構造を考える。つまり実験の評価基準は、要因 A が表の行方向ならば、要因 B が表の列方向になる。A という要因を $A_1, A_2, \cdots, A_i, \cdots, A_p$ までの p 個の水準に分け、B という要因を $B_1, B_2, \cdots, B_j, \cdots, B_q$ と q 個の水準に分けたとするならば、実験条件の組み合わせは、p×q 通りの表で表される。この表の一つのセルが一つの群（標本）として扱うことができ、その中に、さらに複数「繰り返し数」の測定値が入ることになるだろう。一つのセル（または群、標本）の実験条件が (A_i, B_j) と表され、その時の測定値（のひとつ）を x_{ij} と書くならば、この数値を「構造モデル」（要因がどのように測定値に影響を与えているのかを表す式）を書き下し、その式に従って「分散（×自由度、すなわち平方和）を分解していく」のが分散分析の手続きの第一歩である。

$$x_{ij} = \mu_0 + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}$$

ただし、μ_0 は全体の平均、α_i は要因 A による効果で j に依存しない（B の水準に依らない）値、β_j は要因 B による効果で i に依存しない（A の水準に依らない）値、$(\alpha\beta)_{ij}$ は、A_i、B_j による交互作用の効果※、ε_{ij} は偶発誤差によるばらつきであるため同じセル（標本）の中に十分に多くの測定値があ

るならば平均化により打ち消すことができるものである。α_i や β_j は、全体平均と水準内の平均とを比較すればわかる。

※ 交互作用とは、要因の組み合わせ方によって効果の出方に差がある場合をいう。飲み物の好みにおいて種類、加糖の有無の２要因で調べたとき、コーヒーは加糖が好きだがウーロン茶や緑茶は無糖がよいなどと結果がでれば、これは交互作用によるものである。

（参考）https://staff.aist.go.jp/t.ihara/anova.html
　分散分析
（産業技術研究所 井原俊英氏、新重光氏による「化学標準物質の不確かさ」に関するコンテンツ）

（参考）http://elsur.jpn.org/resource/anova.pdf
　読めば必ず分かる分散分析の基礎
（小野滋氏による解説資料）

（参考）http://www.shiga-med.ac.jp/~koyama/stat/com-anova.html
［基本解説→多群比較のための ANOVA; analysis of variance　分散分析法］

◎ ピアソンの χ^2 検定（ノンパラメトリックな検定）、G 検定

ピアソンの χ^2 検定は、上の項で説明したパラメトリックな χ^2 検定と形が類似している※が、全く別のものであると考える方がわかりやすい。ここで検定の対象となるのは、ある集団において特定の事象が生じる頻度が、別の集団と異なるかどうかである。サイコロの特定の出目の有無だったり、特定の血液型の出現率だったり、何かの効果の有無の割合、といった名義尺度ごとの頻度（別の言い方をすると、分割表で結果が表されるような統計データ）について、その平均値が集団間で差があるかどうかを調べるものであり、連続分布する値の平均値などを対象とするものではない。とはいえ、この二項分布（や多項分布）で与えられる確率分布が、正規分布とみなして差し支えない程度に大きな標本が対象である※2。χ^2 分布に従う統計量はいくつかあるが、ピアソンは次の指標 Y が χ^2 分布に従うことを示した。また、指標 Y に対応する 指標 G を用いることもできる※3。次の式において、O は観察された量、E は理論値や比較すべき母平均、期待値、または、複数の群間の平均などである。

$Y = \Sigma[(O-E)^2/E]$　　：ピアソンのカイ二乗検定で用いる指標

$G = 2\,\Sigma[O\times\ln(O/E)]$　：G 検定で用いる指標

※ 指標 Y が、パラメトリックな χ^2 検定における $\Sigma[n(\mu-\mu_0)^2]/\sigma_0^2$ と類似の形を持つという意味である。ただし、期待値 E と σ_0^2/n とが 1 対 1 で対応することを保証するという意味ではない。

※2 正規分布以外の一般的な分布に対しても、近似的には指標値 Y が χ^2 分布に従うものとして扱うことができるので、この検定の適用が可能である。しかしながら、（ウィキペディアより引用）期待値 E が小さい（標本数が小さい、または観測数が少ない）場合は、二項分布を正規分布ではうまく近似できないため、この場合には尤度比検定の 1 つである G 検定を用いるのがより適切である。（引用ここまで）（引用先）　https://ja.wikipedia.org/wiki/カイ二乗検定

※3 データ処理にコンピュータや関数電卓を使用するようになると、指数計算自体が比較的簡便に行える。そのため、いままでピアソンのカイ二乗検定を用いていた場面では、G 検定を用いる方がよい。（以下、ウィキペディアより引用）カイ二乗検定は分布関数への適合性や分割表における独立性の検

定に広く用いられてきたが、実は対数尤度の近似に基づくものであり、一方 G 検定は対数尤度を直接用いる方法である。カイ二乗検定はカール・ピアソンによって計算の容易な方法として導入されたのであるが、コンピュータの普及によって G 検定も決して煩雑な方法ではなくなってきた。特に 1994 年に出版されたソーカルとロルフの教科書（「生物統計学」第 3 版：参考文献）で推奨され、広く利用されるようになった。（引用ここまで）（引用先）　https://ja.wikipedia.org/wiki/G 検定

ピアソンの χ^2 検定には、大別して２種類の比較がある。ひとつは「適合度検定」、もう一つは「独立性検定」である。

「適合度検定」では、帰無仮説として「観測された頻度分布が理論と同じである」を立て、これが成り立つかどうかを検定する。次のように表を作成する。ここでは、はじめに例示したサイコロの１の目の出目頻度について、理論（1/6）と同じであることを帰無仮説とする。この場合、1 と 1 以外の目が出ることは独立ではないから、自由度は 1 である。（もし、１から６までの出目頻度をそれぞれカウントしていたら、足し合わせる $(O-E)^2/E$ は 6 個となり、自由度は 5 となる。）

表　実験１および実験２の結果に対するピアソンの χ^2 検定および G 検定の適用結果

	実験１			実験２		
出目	1	1 以外	合計	1	1 以外	合計
O 観察された頻度	5	7	12	25	35	60
E 期待値	2	10	12	10	50	60
$(O-E)^2/E$	4.5	0.9		22.5	4.5	
指標 Y	5.4			27.0		
$O\times\ln(O/E)$	4.581	-2.497		22.907	-12.484	
指標 G	4.169			20.847		
参考値※	CHIINV(0.0364, 1) = 4.378			CHIINV(4.2×10^{-6}, 1) = 21.171		

※ 実験 1 および実験 2 に対して計算した p 値を与えるような、自由度 1 のカイ二乗値

右側棄却域 5 % の閾値は、χ^2 分布表によれば自由度 1 のとき 3.841 であるので、２つの実験結果はともに棄却域に入ることがわかる。すなわち、5 % の有意水準で「有意に 1 の出目確率は 1/6 と比較して偏っている」と結論できる。また、右側棄却域 1 % での棄却域の閾値は、自由度 1 のとき 6.635 である。実験結果１（n=12）では、1 % の有意水準では「有意に 1 の出目確率が 1/6 と比較して偏っているとは言えない」と結論される。なお、これらの結果は、すべて p 値に基づく仮説検定の結果と一致している。参考値として、それぞれの実験結果についての帰無仮説の下での p 値が棄却域限界に対応するような χ^2 指標値を求め、示している。

独立性検定では、２群以上に対して算出されたある事象の出現頻度について、帰無仮説として「すべ等しいとみなせる」ことに基づいて検定する。この場合、理論などから期待値を計算する代わりに、群全体での平均値に基づく期待値を計算する。以下に具体例を示す。

O 観測された頻度	事象 A	事象 B	事象 C	標本内合計
標本１	10	15	25	50
標本２	25	25	50	100
標本３	25	45	80	150
標本４	45	60	95	200
階級別合計	105	145	250	500（総計）

E 期待値（式）	事象 A	事象 B	事象 C	標本内合計
標本１	50 ×105/500	50 ×145/500	50 ×250/500	
標本２	100×105/500	100×145/500	100×250/500	
標本３	150×105/500	150×145/500	150×250/500	
標本４	200×105/500	200×145/500	200×250/500	

E 期待値（値）	事象 A	事象 B	事象 C	標本内合計
標本１	10.5	14.5	25	50
標本２	21	29	50	100
標本３	31.5	33.5	75	150
標本４	42	58	100	200

$(O-E)^2/E$	事象 A	事象 B	事象 C	標本内合計
標本１	0.0238	0.0172	0	0.0411
標本２	0.7619	0.5517	0	1.3136
標本３	1.3413	3.9478	0.3333	5.6224
標本４	0.2143	0.0690	0.25	0.5333

この 12 個のセルの合計 $Y = \Sigma[(O-E)^2/E] = 7.5103$ であった。階級ごとに平均を出しているから、標本数は 4 だが、これに関する自由度は 4-1 = 3、事象 A、事象 B、事象 C が互いに独立ならこれに関する自由度は 3 である。もし、事象 A、事象 B、事象 C が互いに排他的に生じるならば、これに関する自由度は 3-1 = 2 である。今、後者を前提として、自由度 (4-1)×(3-1) = 6 で χ^2 検定すると、右側棄却域の閾値は 12.5916 である。すなわち、Y 値はこれより小さいため、帰無仮説は棄却されない。すなわち、標本のとりかたによって、事象 A、事象 B、事象 C の出現する割合に有意差は認められないと結論される。

◎ 探索的データ解析と箱ひげ図

特に解析の初期フェーズにおいて、数学的な緻密性よりも、データが正規分布に従うという仮定をせずに、試行錯誤的に解析する方法を、探索的データ解析と呼ぶ。統計計算を伴わずにデータのもつ特徴を図的に表そうとする試みでもある。たとえば、一連の数値データを図として表す場合に、箱ひげ図が用いられることがある。箱ひげ図の作成においては、まず中央値、四分位点を求める。一連の数値データの数が n であるなら、この一連のデータを昇順（または降順）に並び変えたときの n/2 番目のデータ（偶数の場合は前後の２個の平均）が中央値になる。また、n/4 番目ならびに 3n/4 番目のデータ（その位置に応じて前後のデータの重み付き平均）が第一および第三四分位点である。中央値を太線で表し、第一および第三四分位点を下限、上限とした箱を書く。この定義より箱の中には全データの半数が入る。これは、正規分布に従う場合には、中央値は平均に相当し、第一および第三四分位点は公算誤差（0.67σ）に相当する。ひげの両端は、最小値と最大値であったり、上下の 10 % 点であったりする。ただし、箱の上下から第一四分位点と第三四分位点の距離（箱の高さ）の 1.5 倍以上離れている（正規分布では 2.7σ より外相当。0.7 % 程度の）データは、はずれ値として扱われる。

（参考）　http://bdastyle.net/tools/boxplot/page2.html
　箱ひげ図の作成
（hawcas 氏による Excel を使ったビジネスデータ分析ツールの作成過程のコンテンツ）

◎ 分析に入る前の場合分けのまとめ

→ 母集団の全データ、または、母平均、母分散等の統計値が入手可能な場合がある。

→ 標本を抽出して、そこから推計しなければいけない場合

1) 知りたいことが何かを確認する。

２群の推定平均値の差の有無　… 信頼区間の比較、SEM によるエラーバー
正規分布を仮定した仮説検定
母分散未知で共通、かつ、標本が小さい： t-検定
分散分析（ANOVA）でもよい。χ^2 検定、F 検定
母分散が異なる場合 … ウェルチの検定（t 値への補正）

２群の推定分散値の差の有無　… 標本不偏分散の比較、SD によるエラーバー
F 検定（等分散性検定）
F 値 = ２群の不偏分散の比（分子＞分母）

３群以上の推定平均値の差の有無　　… ２群間の比較を繰り返すと多重比較になる
信頼区間を並べて比較することとほぼ同値
母分散既知 ： χ^2 検定
母分散未知だが共通、かつ、標本数が少ない ： F 検定
F 値 = [群間変動の不偏分散]/[群内変動の不偏分散]
帰無仮説は $\mu_A=\mu_B=\mu_C\cdots$ など。

平均値の差が、どの要因によるものであるか
（要因ごとに平均値に与える影響に差があるかどうか）　…　分散分析（ANOVA）

2) 標本の分布はどうなっているかを確認する

まず標本の数値をヒストグラムなどにしてみて、およその分布の形を確認する。
いずれも場合も、標本平均の期待値をもって、母平均（真の値）を推定する。
この推定がどの程度正しいのかを、標準誤差、信頼区間の幅をもって表すことができる。

・標本の分布に規則性がある

標本が十分に大きいとき　…　標本平均は正規分布に従うものと扱うことができる
母分散が既知　→　標本平均の分散を、母分散から $\sigma_m^2 = \sigma_0^2/n$ で求める
母分散が未知　→　標本平均の分散を、標本不偏分散から $\sigma_m^2 = \sigma^2/n$ で推定する
標本の大きさが十分に大きいとは言えないとき
母集団が正規分布に従う　…　標本平均は t-分布に従うものと扱うことができる。
母分散が既知　→　標本平均の分散 $\sigma_m^2 = \sigma_0^2/n$ を利用してよい
母分散が未知　→　標本平均の信頼区間を、t-分布を適用して推定する
母集団が 0 と 1 の二値である →　標本平均の分布に、二項分布を適用する。
その他　…　適切な分布を判断する。

・標本分布に規則性が見当たらない

ノンパラメトリックな検定しか行えない。
順位付けが可能なら、順位和で検定を行うなど。
推定値＋エラーバーの代わりに箱ひげ図などを使う。

◎ エクセルによる数値の解析の例

十進 BASIC のプログラムを用い、標準正規分布に従う乱数を 25 個×30 組発生させ、カンマ区切りテキストとしてエクセルのシートに（25 行、30 列で）貼りつけて解析した。
［データ］タブの［分析］グループ内［データ分析］→ 分析ツール「基本統計量」「ヒストグラム」

（結果）

表　基本統計量の比較

	A	B	C	D	E	F	G	H	I	J	K
13		n	m	σ_n	σ	σ^2	信頼区間の			m= 0.000	m= 0.029
14		データの個数	平均	標準誤差	標準偏差	分散	区間幅(95.0%)	上限	下限	信頼区間に上の値が含まれる	
15	全体	750	0.028867	0.036336	0.995109	0.990243	0.071333	0.1002	-0.04247	1	1
16	列1	25	-0.12509	0.206683	1.033415	1.067947	0.426573	0.301482	-0.55166	1	1
17	列2	25	-0.12074	0.185974	0.929868	0.864654	0.383831	0.263095	-0.50457	1	1
18	列3	25	0.259877	0.205749	1.028743	1.058311	0.424644	0.684522	-0.16477	1	1
19	列4	25	-0.13364	0.156307	0.781537	0.6108	0.322603	0.188961	-0.45624	1	1
20	列5	25	-0.36641	0.162619	0.813095	0.661123	0.335629	-0.03078	-0.70204	0	0
21	列6	25	-0.15684	0.182146	0.910732	0.829432	0.375932	0.219095	-0.53277	1	1
22	列7	25	0.148049	0.226486	1.132428	1.282393	0.467443	0.615492	-0.31939	1	1
23	列8	25	-0.01513	0.214184	1.07092	1.14687	0.442054	0.426919	-0.45719	1	1
24	列9	25	0.127371	0.228774	1.14387	1.308438	0.472166	0.599538	-0.3448	1	1
25	列10	25	0.14042	0.221251	1.106253	1.223796	0.456639	0.597059	-0.31622	1	1
26	列11	25	0.080668	0.227802	1.139011	1.297346	0.470161	0.550829	-0.38949	1	1
27	列12	25	0.35101	0.188662	0.943312	0.889837	0.38938	0.74039	-0.03837	1	1
28	列13	25	0.09221	0.224426	1.122131	1.259179	0.463193	0.555403	-0.37098	1	1
29	列14	25	0.095238	0.258276	1.291378	1.667656	0.533055	0.628293	-0.43782	1	1
30	列15	25	0.134643	0.158253	0.791263	0.626097	0.326617	0.461261	-0.19197	1	1
31	列16	25	-0.37223	0.162392	0.811959	0.659278	0.33516	-0.03707	-0.70739	0	0
32	列17	25	0.165374	0.206596	1.032981	1.06705	0.426394	0.591767	-0.26102	1	1
33	列18	25	0.109797	0.171114	0.85557	0.731999	0.353162	0.462958	-0.24337	1	1
34	列19	25	-0.27806	0.221766	1.108831	1.229506	0.457703	0.179641	-0.73577	1	1
35	列20	25	-0.24095	0.16749	0.837448	0.701319	0.345681	0.104732	-0.58663	1	1
36	列21	25	-0.00131	0.221231	1.106155	1.223579	0.456598	0.455286	-0.45791	1	1
37	列22	25	0.281348	0.221185	1.105923	1.223066	0.456503	0.737851	-0.17515	1	1
38	列23	25	-0.02175	0.1772	0.886	0.784996	0.365723	0.343972	-0.38747	1	1
39	列24	25	0.206521	0.183163	0.915815	0.838717	0.37803	0.584551	-0.17151	1	1
40	列25	25	0.264388	0.213764	1.068822	1.142381	0.441188	0.705576	-0.1768	1	1
41	列26	25	-0.0719	0.197597	0.987984	0.976112	0.40782	0.335916	-0.47972	1	1
42	列27	25	0.148671	0.196467	0.982335	0.964982	0.405488	0.554159	-0.25682	1	1
43	列28	25	0.149317	0.174158	0.87079	0.758274	0.359444	0.508761	-0.21013	1	1
44	列29	25	-0.06996	0.196122	0.980612	0.9616	0.404777	0.334814	-0.47474	1	1
45	列30	25	0.085129	0.17476	0.873802	0.76353	0.360688	0.445817	-0.27556	1	1
46	平均		0.028867	0.1994	0.997	0.994009					

データ分析ツールで表示させた基本統計量から一部を抜粋して表にまとめた。
H, I 列は 95 % 信頼区間の上限と下限。J, K 列はその区間内に母平均が含まれるかの判定。
D46, E46 は、根二乗平均。セル J13、K13 は表示形式で "m=" を表示させているが、数値。
K13 = C15、 H15 = C15 + G15、 I15 = C15 - G15、 J15 = IF(AND($H15>J$13,$I15<J$13),1,0)

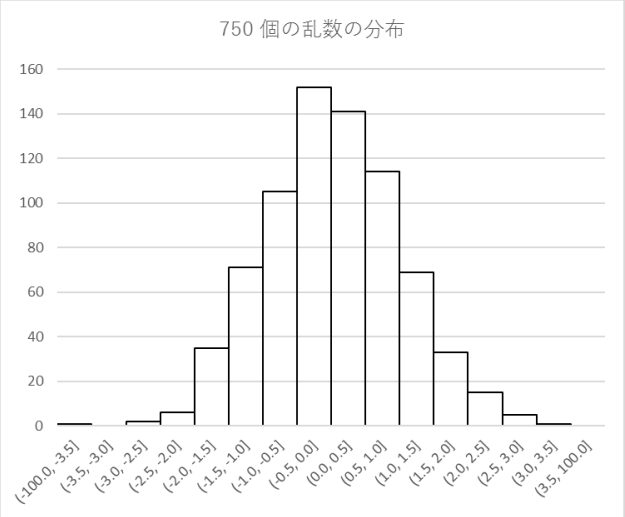


図 750 個のデータの分布
　　標準偏差 = 0.9951
　　母標準偏差 σ_0 = 1.0000

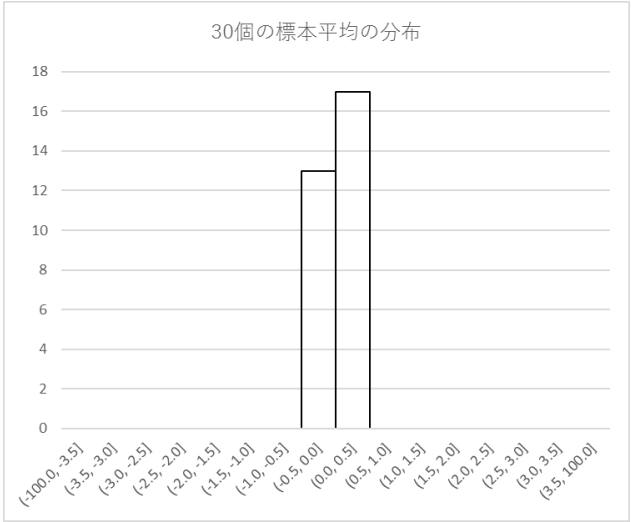


図 標本平均の分布（標本の大きさ = 25）
　　n = 25 の標本の標準誤差の平均 = 0.1994

→ エクセル D 列に表示される標準誤差の値は、D15 = SQRT(F15/B15) で計算される値と一致する。すなわち、$\sigma_m = \sqrt{(\sigma^2/n)}$ で求められている。従って、厳密には 68.3 % 信頼区間の意味は持たない。ただし、σ^2 はその標本における不偏分散値。また、エクセル G 列に表示される 95 % 信頼区間の幅は、G15 = D15*TINV(0.05, B15-1) で計算される値と一致する。すなわち、標本平均が n-1 の自由度で t-分布しているとみなして求められている。

→ n = 25 の 30 個の標本について、標準誤差（平均値の標準偏差）の根二乗平均値は、0.1994 であった。この期待値は、母標準偏差と標本の大きさから $\sigma_0/\sqrt{25}$ で計算され、0.2000 である。

→ n = 25 の 30 個の標本について、平均値は、標準偏差 0.1893 でばらついた（最大 0.3510、最小 -0.3722）が、その平均値は n = 750 の全体に対する平均と一致し（これは当然）、0.02887 であった。平均値 30 個の標準偏差 0.1893 は、30 個の標本ごとに求めた標準誤差の根二乗平均 0.1994 およびその期待値 0.2000 とも比較的よく一致している。

→ n = 25 の 30 個の標本について、不偏分散は、標準偏差 0.2603 でばらついた（最大 1.6677、最小 0.6108）が、平均値（0.9940, SEM = 0.0475）は、標準正規分布（母集団：乱数発生のアルゴリズムより、母標準偏差 σ_0 = 1.0000 ）とも、n = 750 の標本（母集団２、不偏分散 0.9902）とも一致しているとみなして差し支えない。

→ 標準正規分布に従う母集団から抜き出した、n = 750 の標本として、平均は 0.0289 であり、95 % 信頼区間は、-0.0425 ～ 0.1001 であった。この区間は母平均 m_0 = 0 を含んでいた。ただし、95 % 信頼区間なので、同じ試行を繰り返した場合、期待値として 20 回に 1 回は含まないことがある。

→ n = 25 の標本として、１列目 25 個のデータの平均は、-0.1251 であり、95 % 信頼区間は、-0.5517 ～ 0.3015 であった。この区間は母平均 m_0 = 0 を含んでいた。

→ 同様に、2 列目以降も調べたところ、5 列目と 16 列目の n = 25 の標本においては、95 % 信頼区間に母平均が含まれないことがわかった。n = 25 の標本 30 個について評価したので、期待値として 1.5 個の標本において 95 %信頼区間に母平均が含まれないはずであるので、この 2 つという数値は、ほぼ期待値通りである。

○ 最小二乗法

［データ］タブの［分析］グループ内［データ分析］→ 分析ツール「回帰分析」

表 回帰分析の実行結果例

x	y
1	2
2	3
3	4
4	6
5	7
6	8

概要

回帰統計	
重相関 R	0.993859
重決定 R2	0.987755
補正 R2	0.984694
標準誤差	0.29277
観測数	6

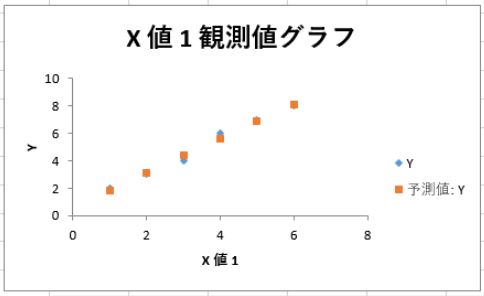
分散分析表

	自由度	変動	分散	観測された分散	有意 F
回帰	1	27.65714	27.65714	322.6667	5.65E-05
残差	4	0.342857	0.085714		
合計	5	28			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	0.6	0.272554	2.201398	0.092508	-0.15673	1.356731
傾き → X 値 1	1.257143	0.069985	17.96292	5.65E-05	1.062832	1.451454

残差出力

観測値	予測値: Y	残差
1	1.857143	0.142857
2	3.114286	-0.11429
3	4.371429	-0.37143
4	5.628571	0.371429



A 列、B 列に入力した数値の組 6 点について、回帰分析を行い、y = 1.26x + 0.6 の式を得た。

（同時にではないが）a=0 または b=0 であるような「本来の」直線 L_0 : y = ax + b の上下に、正規分布に従い、同じ残差分散でばらついて存在している点の集合から、無作為に（今回と同数の）6 個を拾い上げて直線回帰をしたときに、偶然の偏りの結果として今回得られたような傾き a、切片 b の値が得られてしまう確率を p 値として表している。また、直線 L_0 の周囲の点の集合から、無作為に 6 個を拾い上げて直線回帰をしたときに得られる切片や傾きを確率分布（自由度 n-2 = 4 の t-分布）として表したとき今回得られた a や b が中央値からどの程度外れた値に相当するのかを、（その確率分布における標本標準偏差の何倍であるかという値である）t 値として示している。

切片の p 値が 0.093 ということは、本来は切片がゼロであったとしても、サンプリング時のばらつきにより 0.6 という値が算出されてしまう可能性が 9.3 % もあったということである。この値が 5 % 以下になるのは、切片の値が（下限 95 % の）-0.16 よりも小さかった場合か、あるいは（上限 95 % の）1.36 よりも大きかった場合である。この区間内にゼロが入るということは、ゼロである可能性を 5 % の有意水準で否定できないということを示す。（正規分布する推定値に対する ± 2σ に相当する誤差範囲に相当し、有意水準 5 % でこの区間内のいずれであってもおかしくないとみなされる。）

従って、得られた回帰直線 L : y = 1.26(6) x + 0.6(3) （カッコ内は、標準誤差）について傾きは 5 % の有意水準で 0 ではないと言えるが、切片は 5 % の有意水準で 0 ではないとは言えない。（0 であるとも言っていない。データの数が少ない、ばらつきが大きいなどの理由で誤差範囲が広いので、5 % の危険率では 0.6 と 0.0 を区別できる程度の精度がない。）

◎（参考）　図の作成に用いた 十進 BASIC プログラムソース（の一部）

十進 BASIC は、以下の URI より入手してください。
!　http://hp.vector.co.jp/authors/VA008683/

○ 二項分布の確率計算とヒストグラムの作図

```
LET B = 60      ! 試行回数、または「標本の大きさ」。十進 1000 桁モードで 450 まで。
DIM A(0 TO B)   ! 各階級（各出現回数）の確率
LET p = 1/6     ! 対象とする事象の生じる確率
LET S = 0       ! 確率の積算
LET MAX = 0     ! 最頻値（モード）に対する階級の確率を決めるための変数：表示調整用
LET CU = 1      ! ヒストグラムの横方向拡大倍数

FOR k = 0 TO B
   LET A(k) = p^k * (1-p)^(B-k) * FACT(B) / FACT(k) / FACT(B-k)
   LET  S = S + A(K)

   PRINT USING "  ###":k;                ! 出現回数
   PRINT USING "  %.#######   ":A(k);    ! 確率
   PRINT USING "  ### 回以下":k;         ! 出現回数
   PRINT USING "  ###.##":  s*100;       ! 確率積算(%)
   PRINT USING "  ### 回以上":k+1;       ! 出現回数
   PRINT USING "  ###.##": (1-s)*100     ! 確率積算(%)

   IF MAX < a(K) THEN LET MAX =a(k)      ! より大きな値が見つかれば MAX に代入する
NEXT k
PRINT
! PRINT S     ! 全事象の確率の和が１になっていることを確認する。

SET WINDOW (B+1)*(-0.1)/CU ,(B+1)*1.1/CU, -CEIL(MAX*30)/250, CEIL(MAX*30)/25
SET axis COLOR 1
DRAW axes ((B+1)*1.2,CEIL(MAX*30)/250)

LET TITLE$ =  "二項分布, 確率 p =" & STR$(ROUND(p,3)) & ", 標本の大きさ n =" & STR$(B)
PLOT TEXT, AT B/10/CU, CEIL(MAX*30)/25*0.95 : TITLE$

FOR k = 0 TO B
   PLOT LINES : k,0; k,a(k); k+1,a(k); k+1,0
NEXT k

SET TEXT JUSTIFY "center","top"
LET ST = INT(B/20)
IF ST = 0 THEN LET ST = 1
FOR k = 0 TO B STEP ST
   PLOT TEXT, AT k+0.5, 0 : STR$(k)
NEXT k

END
```

○ 二項分布検定の検定力曲線の作図

```
SET WINDOW -0.1, 1.05, -0.1, 1.1
SET axis COLOR 1
DRAW grid(1/6,0.2)
LET B = 60                    ! 試行回数、または「標本の大きさ」
DIM A(0 TO B)                 ! 各階級ごとの確率を収納するための配列変数
LET title$ = "有意水準 α = 0.05, 両側検定"

! 有意水準と比較対象の母平均（p=1/6）は、棄却域の範囲として指定する。
! 範囲指定のとき、0 to 0 では、0 を含んでしまうので注意。
!   棄却域                5 %                    1%
! 12 回 両側検定      : 0 → -1, 6   → B ; 0 → -1, 7   → B
! 12 回 上方片側検定 : 0 → -1, 5   → B ; 0 → -1, 6   → B
! 12 回 下方片側検定 : 0 → -1, B+1 → B ; 0 → -1, B+1 → B
! 60 回 両側検定      : 0 →  4, 17  → B ; 0 →  2, 19  → B
! 60 回 上方片側検定 : 0 → -1, 16  → B ; 0 → -1, 18  → B
! 60 回 下方片側検定 : 0 →  4, B+1 → B ; 0 →  3, B+1 → B

FOR p = 0 TO 1 STEP 0.001     ! 検定対象とする母集団の母平均（真の値）。
   FOR k = 0 TO B
      LET A(k) = p^k * (1-p)^(B-k) * FACT(B) / FACT(k) / FACT(B-k)
   NEXT k
   LET S = 0
   FOR k = 0 TO 4             ! 左側棄却域の範囲を記述。
      LET  S = S + A(K)
   NEXT  K
   FOR k = 17 TO B            ! 右側棄却域の範囲を記述。
      LET  S = S + A(K)
   NEXT k
   PLOT LINES : p, S;
   ! PRINT p; S               ! 関数値を数値データとして欲しい場合有効にする
NEXT p

PLOT TEXT ,AT 0.3, -0.1  :"特定の目の真の出現確率"
PLOT TEXT ,AT -0.08, 0.9 :"1-β"
PLOT TEXT, AT 0.05,1.01  :"標本の大きさ:試行" & STR$(B) & "回、" & title$

END
```

○ 正規分布曲線の作図

```
DEF f(x) =1/(SQR(2*PI)*sigma) * EXP(-(x-mu)^2/2/sigma^2)
LET sigma = 1           ! 標準偏差（標準化のために 1）
LET mu1 = 0             ! 平均 mu に代入するための値。（標準化のため 0）
LET mu2 = 2.5           ! ２本目の正規分布曲線を描く場合の効果量

SET WINDOW -7,7,-0.05,0.45
SET axis COLOR 1
DRAW GRID (1, 0.1)
SET LINE width 2
```

```
LET mu = mu1                ! この位置で、平均以外に標準偏差も指定できる。
CALL normaldistribution     ! サブルーチンの呼び出し
LET mu = mu2
CALL normaldistribution     ! サブルーチンの呼び出し

SUB normaldistribution      ! 正規分布曲線を描く部分の内部サブルーチン
   FOR x = -7 TO 7 STEP 0.01
      PLOT LINES : x, f(x);
   NEXT x
   PLOT LINES
END SUB

END
```

○ t-分布の作図

```
DECLARE EXTERNAL FUNCTION a                          ! 外部関数使用の宣言
DEF t(x,f) = a(f)/SQR(f) * (1+x^2/f)^(-(f+1)/2)      ! 関数定義。f は自由度。
                                                     ! 2 進モードでは、f = 300 まで計算可。
SET WINDOW -5,5,-0.05,0.5
SET axis COLOR 1
DRAW axes (1,0.1)

FOR f = 1 TO 8
   FOR x = -5 TO 5 STEP 0.01
      PLOT LINES : x, t(x,f);
   NEXT x
   PLOT LINES
NEXT f

END

EXTERNAL  FUNCTION a(n)   ! ガンマ関数を含む部分を外部関数として定義。引数 n は自然数。
LET s = 1
IF INT(n/2)*2 = n THEN
   LET k = n/2
   FOR j = 2*k-1 TO 1 STEP -2
      LET s = s * j
   NEXT J
   LET s = s/2^k/fact(k-1)
   LET a = s
ELSE
   LET k = INT(n/2)
   FOR j = 2*k-1 TO 1 STEP -2
      LET s = s / j
   NEXT J
   LET s = fact(k)*2^k * s
   LET a = s/PI
END IF
END FUNCTION
```

○ カイ二乗分布

```
DECLARE EXTERNAL FUNCTION G                          ! 外部関数使用の宣言
DEF f(x,k) = (1/2)^(k/2) * x^(k/2-1)* EXP(-x/2) /G(k/2)

LET R = 18                  ! 表示の右端
LET alpha = 0.05            ! 右側棄却域の限界確率

SET axis COLOR 1
SET WINDOW -(R/15), R, -0.1, 0.8
DRAW axes (1, 0.1)

PRINT "右側棄却域の限界値（上方確率=";alpha;")"
LET dx = 0.001

LET k = 1                   ! k=1 では x → 0 のとき、f → ∞
FOR x = dx TO R STEP dx
   PLOT LINES : x, f(x,k);
NEXT x
PLOT LINES

FOR k = 2 TO 8
   LET t = 0                ! 区間求積のため。ひとつ前の x の値
   LET s = 0                ! ゼロからの累積値
   LET flag = 1             ! 閾値印字のフラグ
   FOR x = 0 TO R STEP dx
      PLOT LINES : x, f(x,k);
      LET s = s + (f(x,k)+f(t,k))*dx/2    ! 台形にて区間求積
      LET t = x
      ! PRINT x; s
      IF s > 1-alpha AND flag = 1 THEN    ! 棄却域に入ったら印字※
         PRINT "自由度";k; " 閾値"; x
         LET flag = 0
      END IF
   NEXT x
   PLOT LINES
NEXT k
! ※ 自由度２のとき、下に凸な関数であるのに、台形で区分求積しているため、
!   数値積分 s が dx = 0.001 で 0.5 % 程度大きくなるため、閾値はやや小さめ表示される。
END

EXTERNAL FUNCTION G(k)      ! ガンマ関数。ただし引数は(1/2)の自然数倍
IF INT(k) = k THEN
   LET A = FACT(k-1)
ELSE
   LET A = SQR(PI)
   FOR n = 1 TO 2*k-2 STEP 2
      LET A = A * n/2
   NEXT n
END IF
LET G = A                   ! 戻り値の指定
END FUNCTION
```

○ 正規分布に従う乱数の生成

```
! テキストの出力結果をエクセルなどに張り付けて使用することを前提としています。
! 他の用途に使用する場合は、※ 行で、適宜配列に格納するなどしてください。

DEF f(x) =1/(SQR(2*PI)*sigma) * EXP(-(x-mu)^2/2/sigma^2)
LET sigma = 1                  ! 標準偏差 σ
LET mu    = 0                  ! 平均 μ
LET NumberOfTheSample  =  30           ! 標本の数：カンマ区切りテキストで出力する際の列数
LET SampleSize         =  25           ! 標本の大きさ：1列の出力における行数
LET NumberOfTheData = NumberOfTheSample * SampleSize   ! 発生させる乱数の全数
LET Fmax = f(mu)

RANDOMIZE
FOR j = 1 TO NumberOfTheData
   LET flag = 0
   DO
      LET k = (RND-0.5)*2*10*sigma + mu   ! μ±10σ の区間内で乱数を平均的に生成
      LET m = RND
      IF f(k)/Fmax > m THEN               ! 発生した乱数 k を採用するかどうかの評価

         PRINT k;                         ! ※

         IF INT(j/NumberOfTheSample)=j/NumberOfTheSample THEN
            print
         ELSE
            PRINT ",";
         END IF
         LET flag = 1
      END IF
   LOOP UNTIL flag = 1
NEXT j

END

! 500 個の乱数を発生させるのに、4000回程度の試行をしているので、Box-Muller 法などと比べて
! 効率はかなり悪いですが、アルゴリズムのわかりやすさを優先させてみました。
! また実質上は、さほどではありませんが、±10σ以上離れたはずれ値が発生する確率を
! 考慮していないなどの問題点もあります。(F(10σ)=7.7E-23程度です。)
! 実用上は、こんな方法もあるという程度に見ていただければ幸いです。
```