Volume 21, Number 1, pages 10-14

特集 feature

論文の書き方4

検定の考え方とt検定

星川佳広　CSCS, NSCAジャパン編集委員, 東海学園大学准教授

本稿は、「論文の書き方1：投稿論文を書こう」（2013年10月号）の続編パート3です。NSCAジャパンウェブサイトに掲載中の「投稿論文（事例報告）の書き方」を参照しながら読み進めてください。

【投稿論文（事例報告）の書き方】
NSCAジャパンウェブサイトTOP　→　[指導者の育成]　→　［事例報告・研究論文］、「投稿要領」部分

前回は、「手持ちの材料」を整理する方法として、基本的な統計量（平均値、標準偏差等）について解説しました。今回は、検定の考え方とt検定（tテスト）について紹介します。検定とは、例えばa群とb群のデータがあってその平均値に差があるときに、その差が本当に差といえるかどうかを確率的に考え、判断しようというものです。

投稿論文を作る際、特に被検者数の少ない事例報告では、検定が絶対に必要か？　と言われれば、そんなことはありません。検定を使わずに説明したほうが適切な場合もあります。しかし、検定を理解し、適切に使いこなせるようになると、「手持ちの材料」を上手に整理することができます。それによって論文も書きやすくなります。

検定には、検定の目的や検定の対象となるデータ構造により、様々な方法（テスト）があります。後半に紹介するt検定は、S&C専門職にとって使用頻度が高い検定方法のひとつです。

1. 標本と母集団

検定の基本として、まずは“標本”と“母集団”の違いを理解する必要があります。例として、日本人と韓国人の身長を比較する場合を考えてみましょう。このとき、日本人と韓国人全員の身長を測定することは現実的には不可能です。そこで、実際には何名かの日本人と韓国人を抽出して調べることになります。このとき日本人、韓国人全員が“母集団”で、抽出して調べた人が“標本”です。標本の身長から、母集団の身長、すなわち日本人と韓国人全体で身長にはどれくらい差があるか（あるいは差がないか）を、確率的に推測しようという考え方が検定です。母集団で考えた場合、日本人と韓国人の身長の平均値は、それぞれ1つしかないはずです。例えば日本人は165.5cmで韓国人は166.7cmというような値です（数値は例で、本当の値ではありません）。しかし、標本の平均については、標本の取り方（どういう人を抽出したか、何名抽出したかなど）によって、その都度、変わります。当然、標本の被検者数が多くなるほど、母集団と同じ平均値が得られる可能性は高くなります。しかし被検者数が少ない場合、母集団と同じ平均値が得られる可能性は下がり、ときには日本人160.0cm、韓国人170.0cmなどと母集団とは大きく異なる平均値が出てくることだってありえます。標本数が少なくなるほど、母集団とは異なる平均値が出てくる確率は高まります。

ここで理解しておくべき重要事項は、私たちが観察や調査で得るデータは、多くの場合“母集団”の値ではなく“標本”として扱うべき値であることで

す。そして標本から母集団を推測するときに使うのが検定です。

逆に考えると、全数調査をするタイプの事例報告――例えば巨人と阪神の選手の身長を比較する際に、各チーム全員の身長を調べたならば、検定は不要です。この場合、もし巨人、阪神それぞれの平均身長が182cm、183cmであったとき、その差が統計的に“有意”かどうかを考えるほうが本質的におかしいわけです。母集団全員を調べたわけですから、論文に堂々と「阪神の選手のほうが巨人の選手よりも身長が1cm高い」と書くことができます。

2. $p<0.05$

論文では“$p<0.05$”という記述をよくみかけます。この$p<0.05$は「有意水準5％未満」を表しています（有意水準は危険率ともいいます）。統計に不慣れな読者の場合、これを“5％未満だと統計的に意味ある差（有意差）で、5％以上だと差がない”と漠然と理解していると思います。では、5％とは何が起こる確率が5％なのでしょうか？

ここで前回、分布を説明する際に使ったコインの例を再び使います。ルールは、コインを振って表が出たら駒を＋1前進し、裏が出たらその場にとどまるというものでした。駒はスタート時には位置0にいます。こういうルールの下でコインを6回振った後、駒がいる位置はどこになるでしょうか。

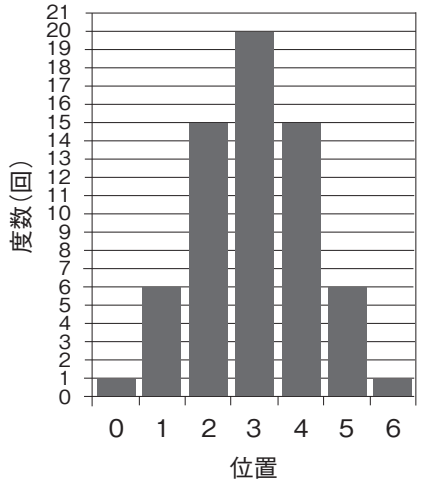

図1　駒の位置の度数分布

そのヒストグラムが**図1**でした。駒の位置で最も可能性が高いのは位置3です。その確率はヒストグラムによれば64回中20回、つまり20÷64＝0.3125（約31％）です。では、位置6にいる確率はどうでしょう？　表→表→表→表→表→表とコインの表が6回続くのは64回中たったの1回、すなわち1÷64＝1.56％です。

今、このルールでゲームを始め、コインを6回振った後、結果的に駒が位置6に来たとします。そして、そこにあなたの友達が現れて、何も知らない友達は駒が位置6にある状況だけを見ます。あなたは友達にルールを説明し、駒のスタート位置はどこであったか？と尋ねます。友達はどう考えるでしょうか。

おそらく友達は「スタート位置が0であった」とは考えないと思います。そう考えるのは危険すぎる。なぜならスタート位置が0であったと仮定すると、駒が位置6にいる確率はたったの1.56％しかない。これは滅多にないことだ。だったらこの仮定は間違っていて、スタートは位置0ではなく、もっと前から（＋3など）スタートしたに違いない。そう考えたほうが確率的に合理的だ…。

検定の考え方はこれに似ています。前述の日本人と韓国人の例を使って説明します。ここに日本人と韓国人の身長の標本データがあるとし、この標本から日本人と韓国人で身長に差があるかどうかを検定します。標本データの平均は日本人160.0cm、韓国人170.0cmであったとします。

①まず、両データが同じ母集団から抽出されたと仮定する。同じ母集団ということはすなわち「日本人と韓国人で身長に差はない」という仮説を立てる（これを帰無仮説といいます）。

②同じ母集団から抽出したデータ群で、平均値に10cmもの差が出る可能性はあるだろうか？　その確率はどれくらいだろう？ その確率を分布から調べてみると1.56％であった（数値は例です）。

③よって、日本人160.0cm、韓国人170.0cmという標本データは、この帰無仮説の下では1.56％の確率でしか起こりえない現象である。もし“5％未満”を滅多にないことの判断基準とすれば、1.56％は滅多にないことに相当する。

④滅多にないことが今起きているわけだから、この標本データは同じ母集団から抽出したものとは考えにくい。つまり「日本人と韓国人で身長に差はない」という仮説のほうが間違っている（これを帰無仮説の棄却といいます）。

⑤すなわち、手元にある日本人と韓国人の標本データはもともと異なる母集団から抽出されている――日本人と韓国人の身長は統計的に有意に差がある、と考えるのが合理的。

つまり、$p<0.05$とは、帰無仮説の下でそのことが起こる確率が5％未満、ということです。この5％は、帰無仮説を棄却する判断基準になるので有意水準と呼ばれます。一方で、1.56％という数字は滅多にないことではあるけれども全くないわけではありません。日本人と韓国人の身長が母集団で同じであっても、標本データでは10cmの

差が出てしまうこともありうるわけです。この場合、帰無仮説を棄却するほうが間違っているわけです。つまり、有意と判断する基準を設定することは、裏を返せば間違いの危険を犯す基準設定でもあります。したがって有意水準のことを危険率ともいいます[1]。

通常、有意水準は5％や1％が使われます。なぜ5％や1％の設定を利用するかは、特別な根拠があるわけではなく、単に慣例です。逆にいうと、特別にこれを否定する根拠もないので、みなさんが「手持ちの材料」を整理するときも、まずは5％、1％を利用すればOKです。

3. t検定

t検定は、2つのデータ群を比較し、その差が有意といえるか検定するものです。レギュラーとサブメンバー、トレーニング前とトレーニング後など、私たちS&C専門職には2つのデータ群を比較したい、ということがよくあります。その際、役に立つ方法がt検定です。

今、a群とb群の2つの標本データがあるとします。a群、b群の母集団をそれぞれA、Bとして、知りたいのはAとBに差があるかないか、です。それを標本であるa群、b群のデータを使って統計的に検定します。

帰無仮説は「AとBは差がない」です。前述したようにまず仮説として母集団A、Bは同じ平均値で同じように正規分布していると仮定します。こう仮定すると、a群とb群の平均値の差は0になる可能性が高いはずです。逆に言えば、a群の平均値とb群の平均値が大きく異なる確率は低いはずです。

正規分布から抽出した標本の標準化した平均値[2]は、標本数ごとのt分布という確率分布に従うことがわかっています。このt分布を利用して確率を求めるのでt検定といいます。

4. Excelを使ったt検定

例として、a群、b群それぞれに10名の身長が入力されたワークシートの例を**表1**に示しました。a群がセルB2:B11に、b群がセルC2:C11にデータが入力されています。Excelでは、“=TTEST(範囲1, 範囲2, 尾部, 検定の種類)”関数によってt検定の確率(*p*値)を得ることができます。カッコ内の範囲1にはa群のB2:B11を、範囲2にはb群のC2:C11などと、データの領域を指定します。尾部は、片側検定(1を指定)か両側検定(2を指定)かですが、通常は2(両側検定)を指定します[3]。検定の種類には、1,2,3のいずれかを指定しますが、Excelでは関数として3種類のt検定が用意されています。

以下にその3種類について説明しますが、いずれの場合もTTEST関数の結果として*p*値が計算されます。この*p*値が設定した有意水準(例えば5％とか1％)より小さければ両群には有意差あり、大きければ有意差なしと判断します。

1)一対の標本による平均の検定(検定の種類に、1を指定)

表1のa群の身長データが就寝前、b群は起床後の身長データとし、同じ行には同一人物の就寝前と起床後のデータが入力されているとします。このように同じ被検者に対して一対のデータがあり、その差を検定する場合にこの

表1　ワークシートの例

	A	B	C
1	被検者	a	b
2	1	175.2	177.6
3	2	170.3	171.5
4	3	166.4	166.6
5	4	178.8	183.0
6	5	182.1	185.1
7	6	162.3	167.8
8	7	169.4	170.7
9	8	187.5	187.1
10	9	160.1	161.3
11	10	157.9	159.3
12	平均	171.0	173.0
13	標準偏差	9.8	9.8

1　母集団には差がないにもかかわらず標本データから差があると判断してしまうことを、統計上の“第1種の誤り”といいます。逆に、母集団に差があるにもかかわらず標本データから差がないと判断してしまうことを、“第2種の誤り”といいます。事例報告で被検者数が少ない場合、問題になるのはどちらかというと“第2種の誤り”のほうです。被検者数が少ないと母集団を精度よく推測できないためです。したがって、被検者数が少ない場合は、検定をしないほうがよい場合もあります。

2　具体的には、正規分布に従う分布から標本をとることを繰り返し、(標本平均－母平均)／(標本標準偏差／$\sqrt{\text{標本数}}$)を求めると、この値は標本数に依存したt分布を示す。

3　片側検定は、a群がb群よりも小さい(大きい)とわかっている場合に使います。

方法を利用します。このようなt検定を、「対のあるt検定」とか「関連2群のt検定」と言ったりします。S&C専門職としては、同じ被検者をトレーニング前後で測定した場合などはこの方法を使って、トレーニングの有効性を検証します。

2)等分散を仮定した2標本による検定(検定の種類に、2を指定)

表1のa群の身長データが陸上競技部10名、b群の身長データが水泳部10名の身長データとします。つまり、a群とb群のデータは異なる被検者から得られています。このように独立した2つの群の差を検定する場合にこの方法を利用します。このようなt検定を、「対のないt検定」とか「独立2群のt検定」と言ったりします。

ただし、検定の種類に2を指定する場合は、a群とb群のデータの分散(=標準偏差の2乗)が同じ程度の場合に利用します。**表1**のデータはa群とb群で標準偏差がほぼ9.8で同じなのでこちらの方法を利用します。

3)分散が等しくないと仮定した2標本による検定(検定の種類に、3を指定)

上記と同じく独立した2つの群の差異を検定する場合に利用します。2)との違いは、a群とb群のデータの分散が著しく異なる場合はこちらを利用します。私たちが取得するデータは必ずしも同じようにバラつくとは限らず、a群はバラつきが大きい一方で、b群はバラつきが小さいということがありえます。また、a群とb群で被検者数が大きく異なる場合も、データのバラつきは両群で異なる可能性が高いです。このような場合は、この方法の利用が適しています。

t検定は標本データが正規分布する母集団から抽出されていることを前提とするので、厳密的には、t検定を実施する前に、①標本データが正規分布から抽出されたといえるかどうか(正規性)の確認と、②標本両群の分散は等しいと仮定できるかどうか(等分散性)の確認が必要です[4]。これらが満たされていない場合は、ノンパラメトリックの検定という方法を使いますが、これらについては別書で理解を深めてください。

5. 注意事項

図2は、**表1**のワークシートをグラフ化したものです。同じデータを元につくったグラフですが、**図2左**は上記1)の関連2群のt検定、**図2右**は2)の独立2群のt検定のイメージです。t検定で得られるp値は、前者が0.0067で後者は0.653で大きく異なり、前者では有意、後者では有意ではないと判断されます。データ構造を理解し、適切な方法を使うことの重要性が理解できると思います。

図2左のような関連2群のt検定の場合、帰無仮説を前提――例えば「就寝前と起床後で身長は差がない」とした場合、$p=0.0067$ということは、**図2左**のようになることが、ほとんどありえないというわけです。

ここで注意してほしいことは、統計的有意性、あるいはp値の大小は、a群とb群の差の大小を表しているわけではないということです。**図2**の左右のグラフではp値が全く異なりますが、a群とb群の差は等しく2.0cmです。p値が"0.0067"などときわめて小さい値であると、何か両群にはとても大きな差があるかのように勘違いしがちですが、そういうことではありません。両群の差の大きさの意味づけは、検定の結果とは切り離して、あなた自身が考えなければいけないものです。仮に**図2左**のデータにおいて、就寝前から起床後に全員の身長が+0.1cmであったとしましょう。そのデータにt検定を施すとp値は、0.0000000…

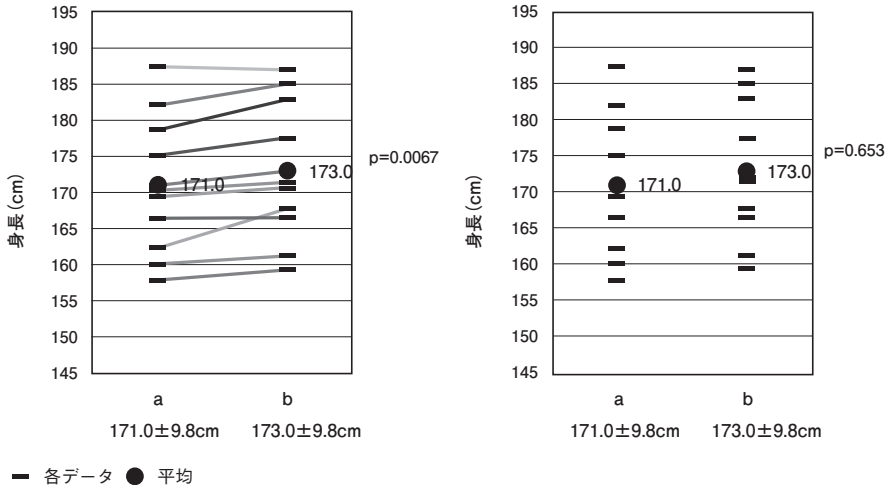

図2　関連2群(左)と独立2群(右)のイメージ
データは同じでもt検定のp値は異なる

4　S&Cの事例報告や研究論文の場合、多くは被検者数が限定的なので、よほど標本データの分布がいびつでないかぎり、正規性や等分散性の検定を実施したとしてもそれが否定されることは少ない。したがって、論文内でわざわざ正規性や等分散性の確認についてまで言及しない場合が多い。

という値になります。帰無仮説を前提とすると、全員が全員同じように0.1cm増えるという確率はきわめて低いからです。この検定の結果に基づけば、“就寝前と比べて起床後では有意に身長が高い”ということは間違っていませんが、しかしその上で、この+0.1cmという身長変化の大きさがどのような意味を持つのかは、p値とは別に考えなければいけないのです。

また、t検定では平均値の差の大きさだけがp値に影響するわけではないことも、注意事項として知っておくとよいと思います。つまり、平均値の大きさが同じであったとしても、標準偏差の大きさによってp値は変わります。その例を**図3**に示しました。**図3左**は**表1**のワークシートデータを元にし、**図3右**はそれと平均値は同じだが、標準偏差がより小さいようにしてあります。独立2群におけるt検定の場合、**図3右**ではp値は0.048と5％水準では有意差と判定されます。なぜなら、標本データの標準偏差が小さい（バラつきが小さい）ということは、母集団の平均値もその小さいバラつきの範囲内に入っている可能性が高いからです。逆にデータのバラつきが大きければ、母集団の平均値が入ると考えられる範囲も広くなります。論文を書こうとする際、データ整理の段階で外れ値が存在して標準偏差が大きい場合など、その外れ値をデータとして含めるかどうかでp値が大きく変わることがありえます。◆

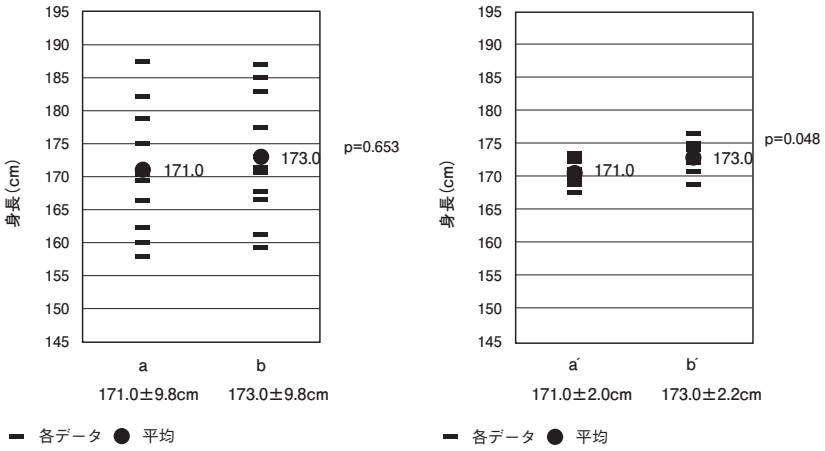

図3　平均値が同じで標準偏差が異なる場合のt検定でのp値の違い
左：標準偏差が大きい場合、右：標準偏差が小さい場合