Volume 21, Number 4, pages 6-11

特集 feature

論文の書き方7

相関と回帰（その1）

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6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

前回までに紹介したt検定や分散分析は、2つもしくは3つ以上のグループ間で観察された差が本当に差といえるかどうか調べるものでした。今回からの「相関」、「回帰」はそれらと異なり、2つの変数間の関係性を調べるものです。私たちS&C専門職には2つの事項がお互いに関係しているのか、全く関係ないのかを知りたいときもよくあります。例えば、あなたが野球選手にスクワットの指導をしたとします。指導後には選手のプレースピードが大幅に改善されたように見えました。しかし、論文を書こうとすれば単にそう見えただけでは説明になりません。スクワット1RMの増加がスピード増加（50m走タイムなど）と関係しているのか？　関係しているとすればそれはどの程度か？　スクワット1RMが1kg増加すると50m走タイムはどれくらい向上すると考えられるのか？「相関」、「回帰」を理解すると、このようなことを数値として示すことができるようになります。

今回（その1）では、「相関」、「回帰」の基礎とExcelの関数やグラフ作成機能を使ったその分析法、結果を解釈する際の注意事項について説明します。次号（その2）では、Excelの「データ分析」機能を使って「相関」、「回帰」をより詳しく理解していきます。

1. 相関係数

今、ここにXとYという2つの事項があるとします。例えばXが身長、Yが体重などをイメージしてください。この2つはお互いにどのような関係にあるのか？　それを調べるためにXのデータ（$x_1, x_2, x_3\cdots$）を横軸に、Yのデータ（$y_1, y_2, y_3\cdots$）を縦軸に配置してグラフを作ります。その様子を示した例が**図1**です。**図1左**では、xが大きくなるとyも大きくなっています。身長と体重も一般的にはこのような関係にあると思います。一方で、**図1中央**では、xとyはあまり関係がないようにみえます。被検者を肥満傾向の人と限定すれば、身長と体重は**図1中央**のようになることもあると思います。**図1右**のデータは、xが大きくなるとyは小さくなっています。例えばXを最大酸素摂取量、Yを3,000m走タイムなどとすると、このような関係になります。

このように2つの事項の関係性を調べるのが相関分析です。そして2つの変数間の関係性の強さを表すのが相関係数で、一般的に小文字のrを使って表現します[1]。rは−1〜1の間の数値をとります。xとyに関係性がない（弱い）場合は、rは0に近い数値になりま

1　Excelはrを自動的に計算してくれるが、その中身は　$r=\dfrac{\text{XとYの共分散}}{(\text{Xの標準偏差}\times\text{Yの標準偏差})}=\dfrac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2}\sqrt{\Sigma(y_i-\bar{y})^2}}$

す。**図1**では中央のグラフがそうです。一方で、xとyの関係性がより強くなるほどrは1もしくは−1に近づきます。xが大きくなるとyも大きくなる場合**(図1左)**は、rがプラスで正の相関、逆にxが大きくなるとyが小さくなる場合**(図1右)**は、rがマイナスで負の相関です。rが1もしくは−1ならば、xとyは完全に一直線上にプロットされます。

どれくらいの相関係数であれば“強い”相関といえるのか、あるいは相関があまりないと考えたほうが良いのか？　これは後述するようにデータ数によって異なり一概にはいえません。しかしおおよその目安[2]としては、0〜0.3相関がないor弱い、0.3〜0.5ふつう、0.5〜0.7強い、0.7〜0.9とても強い、0.9〜1ほぼ完全な相関とみてよいと思います(負の相関の場合も同様)。

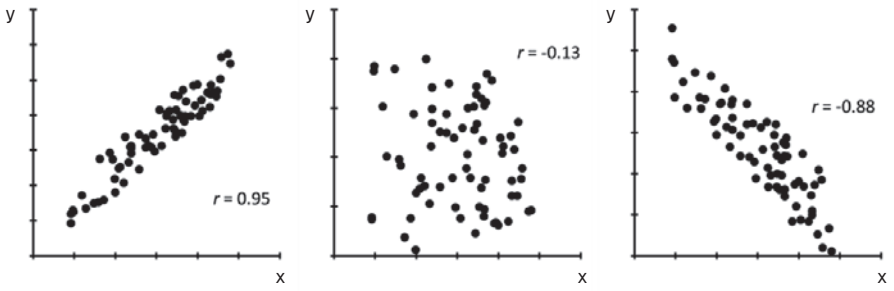

図1　XとYの関係性

表1　ワークシート例

	A	B	C	D	E	F	G
1	選手	50m走(秒)	ベースランニング(秒)				
2	A	6.6	15.1		相関係数	0.84820908	
3	B	6.1	13.6		回帰係数	1.55918031	
4	C	7	15.5		切片	4.32530915	
5	D	6.4	13.9				
6	E	7.2	16.1		被検者数	20	
7	F	6.3	14.2		t	6.79423759	
8	G	7.4	16.2		P	1.155E-06	＊＊
9	H	7.6	16.4				
10	I	6.8	14.7				
11	J	7	14.9				
12	K	7.3	14.6				
13	L	6.5	14.2				
14	M	7.1	14.7				
15	N	6.8	15.2				
16	O	6.7	15.7				
17	P	7.8	16.3				
18	Q	6.6	14.7				
19	R	7.2	15.4				
20	S	7.7	16.5				
21	T	7.2	15.8				

2. グラフを描こう

Excelのグラフ作成機能(散布図)を使うと、「相関」や「回帰」分析の理解が深まります。まずはともかくデータを使ってグラフを描いてみましょう。あなたは大学野球チームでS&Cコーチを務めているとします。あなたは「直線で足の速い選手がベースランニングも速いかどうか？」について知りたいと思いました。選手20名を対象に、直線50m走のタイムとベースランニング(1周約110m)のタイムをストップウォッチで測定し、**表1**ワークシート例のA1：C21の領域のデータを得ました。

図2左は、**表1**のB2：C21の領域を選択しグラフ(散布図)を作成したものです。**図2左**からは、やはり50m走が速い選手はベースランニングも速い傾向にあることが読み取れます。しかしその一方で、50m走が速い割にはベースランニングが遅い選手も見受けられます。両者は完全に一致しているわけではありません(完全に一致しているときが$r = 1$)。では、両者の関係性はどれくらいといえるだろうか。それを数字で示すのが相関係数です。また、ベースランニングタイム(Y)と50m走タイム(X)の関係式を求めたり、それを使って50m走タイムからベースランニングタイムを予測したりすることが回帰分析になります。

相関分析ではxとyの2つの変数を

2　Hopkins,W.G.:A scale of magnitudes for effect statistics https://www.sportsci.org/resource/stats/index.html　参照

対等に扱います。一方、回帰分析ではxを基準にしてyを関係づけたり予測したりします。そこで回帰分析では、yは説明されたり予測されたりする変数なので「目的変数」とか「従属変数」と呼び、xはyを説明する変数なので「説明変数」とか「独立変数」と呼びます。回帰分析によって2つの事項の関係性を調べる際、どちらを説明変数としてどちらを目的変数とするかは、分析の目的や書こうとしている論文の目的、あるいはデータの解釈の仕方(5節参照)によって決める必要があります。

さて、Excelで作った**図2左**のグラフに回帰直線――XとYの関係性を最も適切に表す直線[3]を引いてみましょう。それにはグラフ上の任意のデータ(●)にマウスを当て右クリック→「近似曲線の追加」を選択します。するとExcel 2007以降では**図3**のボックスが出現します(Excelの以前のバージョンでも同様のボックスが出現)。このボックスは**図2左**のグラフの各データ(●)に対する回帰直線(曲線)の描き方を指定するものです。

今、**図2左**のグラフではXとYの両者は直線的な関係にあるように見受けられるので、ここでは「線形近似」を選択しましょう。また**図3**ボックスの下方にある「グラフに数式を表示する」と「グラフにR-2値を表示する」にチェックを入れます。「グラフに数式を表示する」の上にある「切片」にチェックを入れると、下で述べる回帰直線が強制的にその切片を通るように計算されてしまうのでここではチェックを入れません。そして「閉じる」をクリックすると、**図2右**のようにグラフ上に回帰直線、およびその回帰式、決定係数(R^2)が出現します。

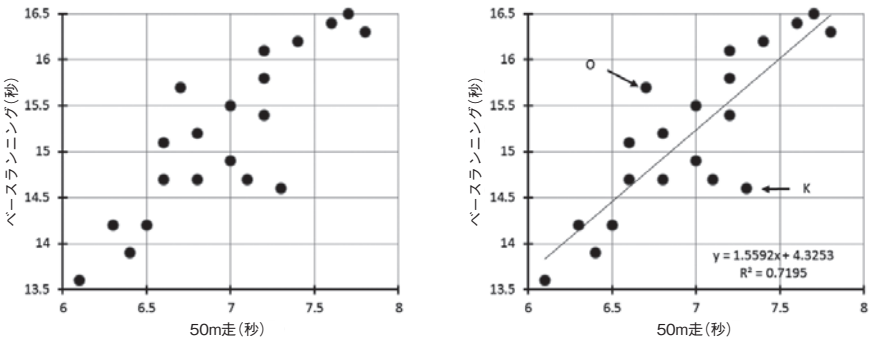

図2　散布図グラフの例(右は、左を元に近似線、回帰式、R^2値を追加したもの)

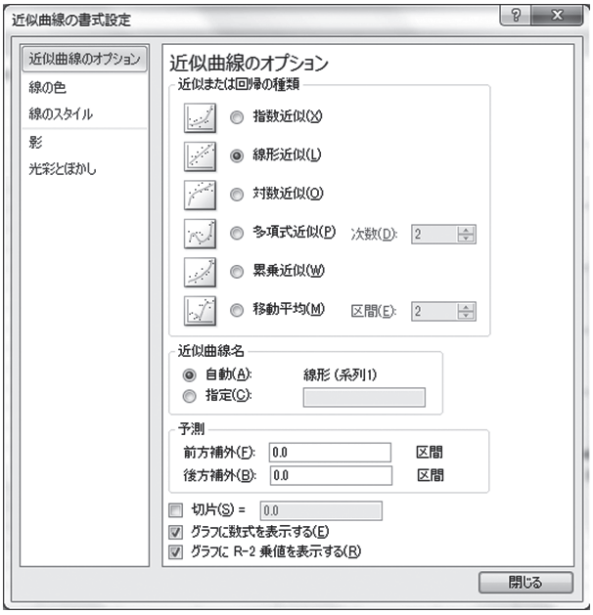

図3　近似直線(曲線)の書式設定(Excel2007以降)

ここで決定係数(R^2)は相関係数(r)とは定義が異なる指標ですが、結論としては決定係数は相関係数の二乗値と等しくなります(次号解説)。すなわち$R^2 = r^2$。**図2右**の例では$R^2 = 0.7195$なので、$r = \pm\sqrt{0.7195} = \pm 0.848\cdots$で、相関係数は0.848か-0.848のどちらかということになります。今、50m走タイムとベーストランニングタイムは、明らかに正の相関関係にあるので0.848、つまり両者は相関係数約0.85の関係にあり、とても強い関係にあると理解できます。

また回帰式$y = 1.5592x + 4.3253$は、回帰直線を一次関数(y=ax+b)の形で表現したものです。つまり、**表1**のデータに基づくと、平均的には「ベーストランニングのタイム(秒) = 1.5592 × 50m走のタイム(秒) + 4.3253」となると考えられるわけです。ここでaのことを「回帰係数(あるいは傾

3　各データ(●)から直線(曲線)までのy軸方向の距離を「予測の誤差」あるいは「残差」という。すべてのデータに対して予測の誤差を計算し、その二乗値の総和(残差平方和)が最も小さくなるような、つまり予測の誤差が最も小さくなる直線(曲線)がxとyの関係性を最も適切に表す直線(曲線)。それを回帰直線(曲線)と呼ぶ。Excelは回帰直線(曲線)を自動的に計算してくれる。

き）」、bを「切片」といいます。回帰係数aが1.5592ということは、50m走タイムが1秒速く（あるいは遅く）なれば、ベースランニングのタイムは平均的には1.5592秒速く（あるいは遅く）なることを表します。

通常、論文で何も断りなく相関・回帰分析と書かれていれば、それはxとyの関係性を直線で結んでいると考えて良いです。このような回帰を「線形回帰」といいます。しかし、データによってはxとyは直線的な関係になく、yがxに対して二次関数や指数関数的な増え方をすることもあると思います。そのような場合、**図3**において指数近似、対数近似、累乗近似などを選択します。これらは線形回帰に対して「非線形回帰」と呼ばれます。はじめて「論文を書こう」とする場合は、基本となる線形回帰を利用することが多いので、本稿では線形回帰のみを扱います。非線形回帰についても知りたい場合はより詳しい資料を参考にしてください。

もう一度**図2**に戻ります。回帰直線を引くことの利点のひとつは、回帰直線から外れた傾向にあるデータを表出できる点です。例えば**図2右**では、↘のついた被検者KとOのデータです。被検者Kの50m走は7.3秒とそれほど速くないのですが、ベースランニングは14.6秒と全体の5位にランクしています。一方で被検者Oは、50m走は6.7秒とかなり速いにもかかわらず、ベースランニングは15.7秒とどちらかと言うと遅いほうに分類されます。50m走が6.7秒ならば、上述の回帰式に基づくと1.5592×6.7＋4.3252≒14.8秒でベースランニングが走れるはずなのです。しかし実際には15.7秒と14.8秒よりも0.9秒も遅い。つまり被検者Oに関していえば、走るスピードそのものというよりはベースランニングのコース取りなどの技術的な部分に課題があるのだろうと考えられます。一方で被検者Kの場合、ベースランニングタイムをさらに短縮しようとすれば、まずは単純な直線走の走力向上に取り組むべきだと考えられます。

このように回帰分析を用いると、トレーニングの方向性を合理的に判断することができます。また論文を書くときにおいても、回帰分析の結果に基づいて被検者をグループ分けしたり抽出したりする――例えば50m走タイムの割にベースランニングが遅い選手のみを選んでトレーニングするなど――と、結果をよりクリアにすることができます。

3. Excel関数による係数の算出

グラフを描くとXとYの関係性が視覚的にわかるので、上述のようにまずはグラフによってデータの全体像を把握することが肝要です。しかし、Excelには相関係数や回帰係数、切片を求めるための関数も用意されていて、関数を使えばグラフを作成しなくてもすぐにそれぞれの数値を得ることができます。相関係数はcorrel関数、回帰係数はslope関数、切片はintercept関数によって求めることができます。**表1**のワークシート例でいうと、セルF2には“=CORREL（C2：C21，B2：B21）”と入力されていて、その関数の結果、50m走とベースランニングの相関係数0.848…が示されています。同様にセルF3に“=SLOPE（C2：C21，B2：B21）”、セルF4に“=INTERCEPT（C2：C21，B2：B21）”と入力されていて、それぞれ回帰係数1.559…と切片4.325…が示されています。それぞれの数値は**図2右**の回帰式と確かに同じになっています。相関分析ではxとyを対等に扱うので“=CORREL（B2：B21，C2：C21）”のようにB2：B21（説明変数）とC2：C21（目的変数）が入れ替わっても同じ数値が得られますが、slope関数、intercept関数では、C2：C21，B2：B21の順を間違えると異なる数値が計算されるので注意してください。

4. 相関係数の有意性

相関分析によって論文を書こうとする場合、相関係数のみならずその有意性も示せれば主張もしやすくなります。論文内に“$r=0.56$（$p<0.05$）”などと書かれているのを見たことがあるかと思います。これは相関係数が0.56でその相関係数が5％水準で有意であることを示していますが、ところで相関係数が有意とはどのようなことを意味するのでしょうか？

多くの場合、私たちが手にできるデータは標本です（1-2月号参照）。標本でXとYに相関があるからといって母集団でもそういえるとは限らない。標本データから確率的に考えたときに母集団にも相関があるといえる――これが相関係数の有意性です。相関分析での帰無仮説は「母集団でXとYは関係ない」つまり「母集団の相関係数＝0」です。標本データからこの帰無仮説が棄却されれば、母集団においてもXとYは関係性がある――有意な相関となります。

この計算には脚注の計算式（t値）[4]が自由度n－2（nは被検者数）のt分

4　$t=\frac{r\times\sqrt{n-2}}{\sqrt{1-r^2}}$　r 相関係数、n 被検者数

布に従うことを利用します。**表1**のワークシートで相関係数の有意性を調べてみましょう。まずセルF6には被検者数を求めるために“=COUNTA(A2：A21)”が入力されています。セルF7にはt値を求めるために、セルF2（相関係数）とセルF6（被検者数）の値を利用しながら脚注の式“=F2＊SQRT(F6-2)/SQRT(1-F2＊F2)”と入力されています。セルF8には、セルF7のt値のt分布（自由度n－2）における上側確率を求めるためにtdist関数を用いて“=TDIST(F7,F6-2,1)”が入力してあります。その結果、セルF8には1.155E-06（0.000001155 …）と非常に小さい値が計算されています。すなわち、**表1**の50m走とベースランニングのデータには、1％水準（$p < 0.01$）で有意な相関があると判定できます。

どれくらいの相関係数ならば有意と判定されるかは被検者数によって変わります。例えば同じ$r = 0.5$が得られているとしても、それが被検者30名のデータから得られている場合と、被検者10名から得られている場合では有意性の判定が違います（前者ならば有意で後者ならば有意ではないと判定される）。なぜならば、被検者30名からのほうが、より確からしく母集団の相関係数も0.5に近いといえる――すなわち、より確からしく「母集団の相関係数＝0」の帰無仮説を棄却できるからです。

逆に言えば、被検者数が十分でないと母集団の相関係数は、標本の相関係数から精度よく推定できないのです。したがって事例報告などで被検者数が少ない場合（6名程度）、$r = 0.7$と一見高く思える相関係数が得られたとしても、「母集団の相関係数＝0」の帰無仮説が棄却できずに相関係数は有意となりません。もちろん、有意かどうかだけが論文の観点ではないので、有意性がなければ論文を書いてはいけないということではありません。しかし分析結果の解釈は、たとえ高い相関係数であってもそれは有意でないことを踏まえた上で行なうことになります。

この逆で被検者数が非常に多い場合（100名程度）は、$r = 0.2$程度しかなかったとしても相関係数は有意と判定されます。これは相関係数の有意性が「母集団の相関係数＝0」すなわち「母集団でXとYは全く関係ない」という帰無仮説を棄却しているにすぎないからです。つまり帰無仮説の棄却は「XとYは全く関係ないことはない」といっているだけなのです。したがって、有意な相関性を発見したからといって、$r = 0.2$がどれほどの意味をもつかは、統計の結果とは別に論文や分析の目的とあわせて考える必要があります。互いに全く関係ないと思われていた事項に実は弱いながらも相関性を認めたことを主張するならば$r = 0.2$でも意味があるでしょう。しかし多くの場合$r = 0.2$程度ならば、“相関がある”と主張するよりも、実質的な相関はないとして解釈したほうがスムーズに論文が書けるように思います。

5. 相関関係と因果関係

さて相関係数や回帰式によって2つの事項の関係性を数値で表すことができたら、次はその結果の解釈です。論文ならば「考察」に記述する内容です。統計は単に数字上の操作でしかないので、その結果をどう解釈するかは分析した人が自分で考えなければなりません。

相関、回帰分析の結果を解釈するときに注意しなければならないことは、XとYが相関関係にあるからといってXとYに因果関係があるとは限らないことです。私たちは相関、回帰分析を行なうと、Xを原因、Yを結果としてXとYを直接的な因果関係として解釈したくなるものです。Yの理由を知りたい！、その答えはXだ――こうすっきりと解釈できるならどれほど論文も書きやすいかわかりません。しかしXとYは本当に因果関係にあるのか？
このことは常に注意しつつデータを読む癖が必要です。

「朝食を摂る子どもほど学校のテストの点が良い」――確かに朝食の摂取率とテストの点は相関関係にあります。しかしだからといって、朝食を食べたらすぐさまテストの点が上がるほど単純でしょうか。両者に直接的な因果関係があるかは両者の相関関係だけからはわかりません 。最近“ビッグデータ”がよく話題になりますが、ITの発達によりかつては分析対象にもならなかったデータ同士で相関関係が調べられるようになってきています。今後ますます、「朝食」と「テスト」のような予想外な事項間の相関関係はいろいろな場面で出てくることでしょう。したがって、なおさらその結果をどう解釈するかが重要になってきます。

おそらく朝食を毎日摂るような家庭環境の子どもは、生活習慣も規則正しく、勉強する習慣が身についているので、その結果学校のテストもよくなるのでしょう。この解釈の場合「朝食」と「テスト」の相関関係の背景には、「家庭環境」や「生活習慣」という共通の別の要因が作用しています。このようなXとYに共通に働く要因のことを交絡要因といいます**（図4）**。交絡は、私たちS&C領域では非常に身近な問題です。例えば、NSCA会員ならばよくわかると思いますが、身長の高さは多くのスポーツパフォーマンスと相関関係

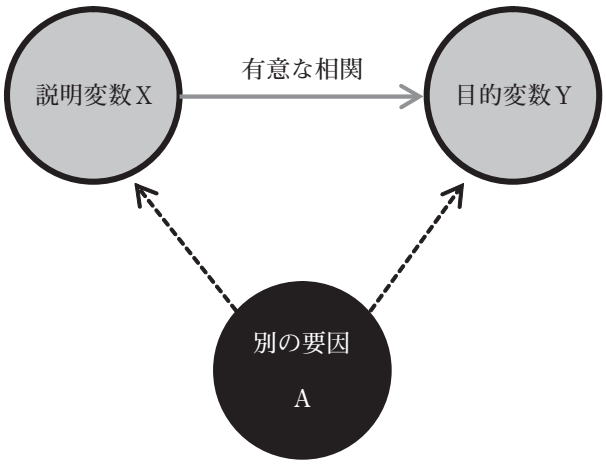

図4　交絡要因のイメージ

にあります。しかし、身長が高いことそのものがスポーツパフォーマンスに直接的にかかわっているよりは、交絡要因として体重や筋力、年齢(競技歴)などが背景に存在していることは少なくありません。したがって、私たちが相関、回帰分析によって論文を書こうとする場合、あるいはそのために実験設定を考える場合、交絡要因の可能性にも目配りしておく必要があります。

相関、因果関係については、交絡以外にも様々に解釈上の注意点があります。ウィキペディア(インターネット上のフリー百科事典)には「相関関係と因果関係」という項目が立てられているくらいです。注意事例がいくつか掲載されているので参考にしてください。

6. エビデンスの信頼性

では、どうすれば因果関係を示すことができるでしょうか？　ここでは深く立ち入りませんが、これは研究デザインの問題になります。本連載1回目に「根拠に基づく医療(エビデンス)」について少し書きました。しかし一口にエビデンスといっても、それが生みだされた研究のデザインによって信頼性のレベルが異なります。例えば、タバコ(原因X)とがん(結果Y)の因果関係を示そうとする際、単に喫煙習慣とがんの罹患率の相関関係を示した場合と、現時点でがんでない人がこれから喫煙する程度を記録し、何年後かにがんが発生する頻度はどれくらいであったかを示した場合では、後者のほうがエビデンスとしての信頼性は当然ながら高いわけです。一般的に相関関係を記述しただけの研究は、エビデンスの信頼性はあまり高くないとされています(しかしだからといって報告の価値がないわけではありません)。

このエビデンスの信頼性は私たちS&C領域においてもあてはまります。例えば、筋力(原因X)と足の速さ(結果Y)の関係性を示そうとする際、単に現時点での「筋力」と「足の速さ」の相関をとるだけよりも、実際にトレーニングによって筋力を高めたときに足の速さがどうなったかを検証したほうが、両者の因果関係は主張しやすい。したがって「論文を書く」ことを想定すれば、まずは比較的容易に調べることができる相関関係を報告し、次にその相関関係からヒントを得た上で実際にトレーニングをやって再度報告する、という2段階で考えるのが良いと思います。筋力を上げれば足が速くなるのか？――実施が長期にわたり、選手にもあなた自身にも負担が大きいトレーニング実験にやみくもに取り組むよりは、まずは相関の程度を把握し、その上で試みたほうが見通しがつきやすい。また仮に、仮説どおりでなかったとしても他の問題点にも気づきやすいでしょう。相関があった→トレーニング実験してみた→仮説どおりに成功した(因果関係あり)。相関があった→トレーニング実験してみた→仮説どおりにはいかなかった(因果関係がなかった)→他の要因が考えられそう。この一連の流れがまさに、論文を書くことを通してあなた自身の専門性が強化されていく過程なのだと思います。◆