Volume 20, Number 10, pages 10-14

特集 feature

論文の書き方3

基本統計量

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2です。本文中のWEB例とは、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」のことです。WEB例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】WEB例
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

事例報告で被検者数が少ない場合、「手持ちの材料」一つひとつをすべて検討し、それを論文に記述することも可能です。しかし、少ないデータを元に強い主張はしにくいものです[1]。やはり、ある程度の被検者数やデータ数を基に、集団としての傾向を示したほうが論文の主張もしやすくなります。ここで、集団の特徴を数値で表すために『統計』の出番となります。今回は「手持ちの材料」を整理するための基本的な統計量、平均値、標準偏差、分布について説明します。

1. 平均値と標準偏差

集団の特徴を表そうとする場合、その集団を代表する数値と、その集団のバラつき具合を表す数値（散布度といいます）が必要になります。前者で最もよく使われるのが「平均値」であり、後者においては「標準偏差」が使われます。

論文を読んだとき、図表や本文中に「被検者の年齢は17.5 ± 0.3歳であった」などの記述を見たことがあるかと思います。これは、被検者全体の年齢は、平均が17.5歳で標準偏差が0.3歳であることを意味します。±の代わりに、17.5（0.3）歳のようにカッコで表現することもあります（WEB例参照）。標準偏差はSD（Standard Deviation）と略すこともあります。

平均値については多くを説明しなくてもよいでしょう。各データの集計値をデータ数で割ったものが平均値です。標準偏差については、以下、例を挙げながら説明します。

表1は、A、B、Cの3チームに各5名の被検者がいて、各被検者のあるテストの得点結果とします。チームの特徴を知るためにまずは平均値を求めてみましょう。すぐにわかると思いますがどのチームも平均は50.0点です。ですから、平均でみればこの3チームには違いがないことになります。この結果から、「どのチームも平均は50点で、チームごとの差異はなかった」と言うことは間違っていません。が、何か物足りないような気がしませんか。

平均値は、確かに集団の特徴を表す最重要な指標です。しかし、それだけで十分かといえばそうではありませ

表1　データ例

チームA	40、45、50、55、60
チームB	30、40、50、60、70
チームC	10、30、50、70、90

1　データが持つ価値が高ければ、少ないデータでも十分に強い主張ができます。オリンピック選手を対象とした報告、特殊な状況の報告では、被検者数1でも十分に報告に値します。

ん。チームAは得点が40～60点、チームBは得点が30～70点、チームCは得点が10～90点と、各チームで得点のバラつき（散布度）が違います。この集団内のバラつきの程度を表現しようとするのが、標準偏差です。平均値と標準偏差がそろうと、その集団の特徴はかなり明確になります。

標準偏差は、

$$\sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n}} \quad \cdots ①$$

ただし、x_iは各データ、$\bar{x}$は平均、nはデータ数

で計算されます。

難解な式に見えるかもしれませんが、①式の$(x_i-\bar{x})$の部分は、各データx_iが平均値$\bar{x}$からどれくらい離れているか、つまりこの値が大きければそのデータは平均値から遠く、この値が0に近ければ平均値に近いというわけです。この（各データ－平均値）のことを「偏差」といいますが、データごとに偏差を出し、それを合算すれば集団内のバラつきが大きいか小さいかわかるということになります。しかし、偏差はプラスの場合もマイナスの場合もあるので、全部の偏差を足し合わせるとプラスとマイナスが相殺されて0になってしまいます。これでは集団内のバラつきの程度がわかりません。そこで偏差を2乗、すなわち$(x_i-\bar{x})^2$とすれば、必ずプラスになるので、集団内の各データが平均値からどれくらい離れているかを確実に足し合わせることができます。したがって①式は、各データの偏差の2乗値を全部合計し（「偏差平方和」といいます）、それをデータ数（n）で割ったものを、ルート（√）により2乗する前の元の状況に戻した、ということを意味しています。

表1の例では、チームA、B、Cそれぞれの標準偏差は、7.1、14.1、28.3になります。ですから、「チームA、B、Cの得点結果はそれぞれ、50.0 ± 7.1、50.0 ± 14.1、50.0 ± 28.3点」であったと記述されていれば、読者は、チームA、B、Cで平均値は同じでもチーム内メンバーの得点は、チームAはバラつきが小さく、チームCには良い得点のものも悪い得点のものも含まれているのだな、ということが読み取れます。

ちなみに、①式の√の中、

$$\frac{\Sigma(x_i-\bar{x})^2}{n}$$

も「分散」という散布度を表す統計量の一つです。つまり標準偏差=√分散です。

Excelでは関数を使うことで、簡単に平均値、標準偏差、分散を知ることができます。それぞれ“=AVERAGE（範囲）”、“=STDEVP（範囲）”、“=VARP（範囲）”で求めることができます[2, 3]。

2. 中央値

「平均値」に似たものに「中央値」があります。「中央値」は「平均値」とともに、教科書には必ず載っている重要な統計量ですが、実際にはS&Cに関する論文で、平均値ではなく中央値が利用されることはほとんどありません。しかし、中央値の概念を知っておくことは重要だと思います。

中央値とは、データを小さい順（または大きい順）に並べたときに、ちょうど中央にくる値のことです（データが複数の場合は、中央の2つのデータの平均値）。例えば5つのデータがあるならば3番目の値が中央値です。

集団内の各データが平均値を中心に対称的にバラついている場合、中央値は平均値に等しくなります。しかし私たちが取得するデータは、必ずしもそのように理想的にバラついてくれるとは限りません。集団の傾向を表す代表値として、平均値よりも中央値のほうが適している場合も多いことが実際です。

例えば、あるサッカークラブの年俸がここ数年どう変わっているか？　ということを調べたいとします。この場合、全選手の年棒の一覧を手に入れて、年度ごとに「平均値」を出してみることが手っ取り早いです。おそらくある年は1,254万円、次の年は1,350万円、などの平均値が出てくると思います。

ところが、ある年だけ非常に有名な外国人選手が移籍してきて、その選手に3億円の年棒を払っていたらどうでしょう。年棒の「平均値」はその年だけ3,000万円などと、かさ上げされることになってしまいます。事情を知っている人ならばそれでも構いませんが、数字だけを見る人は、その年はあたかも選手全員の年棒が上がったかのような印象をもちかねません。

このようなとき、中央値をその集団の代表値とすればよいわけです。中央値は、真ん中の値であるので、3億円の選手が入った年も、例年と異ならず1,200～1,400万円の数字に落ち着くと思います。このように、特に離れ値を含む集団では、平均値よりも中央値のほうがその集団の特性を適切に表す場合があります。

被検者数が少ない場合、平均値を集団の代表値として使うと、離れ値が強く影響することを知っておく必要があ

2　範囲には、例えばセルA2からA8までデータの平均をするならば、「A2:A8」が入る。

3　=STDEVP（範囲）や=VARP（範囲）は、範囲にあるデータを母集団データとした場合の標準偏差、分散です。Excelには似た関数に=STDEV（範囲）、=VAR（範囲）がありますが、これらは範囲にあるデータを標本データとした場合で、①式のnの部分がn-1で計算されます。母集団と標本については次回ご説明します。

ります。Excelを使うと、あまりよく考えないでも簡単に平均値を求めることができてしまうので、より一層この点には注意が必要です。Excelで中央値は“=MEDIAN(範囲)”で計算できます。

3. 度数分布

論文を書き出す段階までくれば、集団を平均値と標準偏差のみで集約してOKですが、「手持ちの材料」を吟味する段階では、「度数分布」を調べて集団の全体像を捉えておくことも役に立ちます。

例として、**表2**に被検者20名のBMI値(A2:B21)と、それを元にした度数分布表(D1:E10)のワークシート例を示しました。度数分布とは、ある基準範囲内に何名が入るか、**表2**で言えばBMIが19～21の人が3名、BMIが21～23の人が5名、といったことを示すものです。Excelでは度数分布表も、データ→データ分析[4]→ヒストグラムによって、簡単に作ることができます。

図1は、**表2**の度数分布表をグラフにしたものです。このように度数分布をグラフ化したものは「ヒストグラム」(histogram)といいます。通常ならば、ヒストグラムは富士山のような形になって、平均値近くで度数が多く、両端に離れるほど度数が少なくなります。**表2**のBMIの平均値は23.3(セルB22)ですが、ヒストグラムを作成すると、この「平均値」が集団を適切に代表するかどうかを確認することができます。また、被検者F(BMI=32.3)が離れ値であることも押さえておくことができます。

ここで注意が必要なことは、ヒストグラムは必ずしも富士山型になるとは限らないことです。例えば、「100点満点のテストで平均は50点であった」と聞けば、通常は**図2左**のようなヒストグラムのイメージを持つと思います。しかし、最近はよく二極化などと言われますが、**図2右**のようなヒストグラムになっている可能性もあるわけです。この場合も、平均値は**図2左**と変わらず50点です。仮に**図2右**のような場合は、平均値を集団の代表値として「手持ちの材料」を整理しても、適切な論文は書けないでしょう。むしろ二極化しているそのことを論文のテーマとするほうが良いと判断すべきです。これは「分布」を検討してはじめて見えてくることです。

表2　被検者A～U20名のBMIと度数分布表

	A	B	C	D	E	F
1	被検者	BMI		基準	度数(人数)	
2	A	23.8		19未満	2	
3	B	22.2		19～21未満	3	
4	C	28.1		21～23未満	5	
5	D	23.9		23～25未満	5	
6	E	20.9		25～27未満	3	
7	F	32.3		27～29未満	1	
8	G	21.6		30～31未満	0	
9	H	21.5		31～33未満	1	
10	I	26.1		以上	0	
11	K	26.2				
12	L	23.2				
13	M	19.8				
14	N	18.7				
15	O	26.4				
16	P	18.0				
17	Q	21.0				
18	R	22.3				
19	S	21.8				
20	T	24.7				
21	U	23.1				
22	平均値	23.3				
23	標準偏差	3.4				

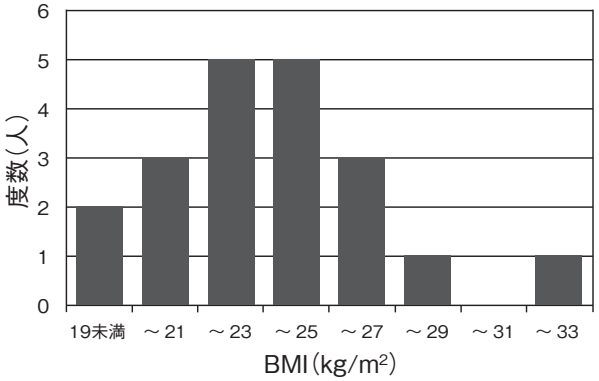

図1　**表2**のデータを元にしたヒストグラム

4　Excelのバージョンによっては、ツール→分析ツール。メニューに分析ツールがない場合、事前にアドインで「分析ツール」を挿入しておく必要がある。

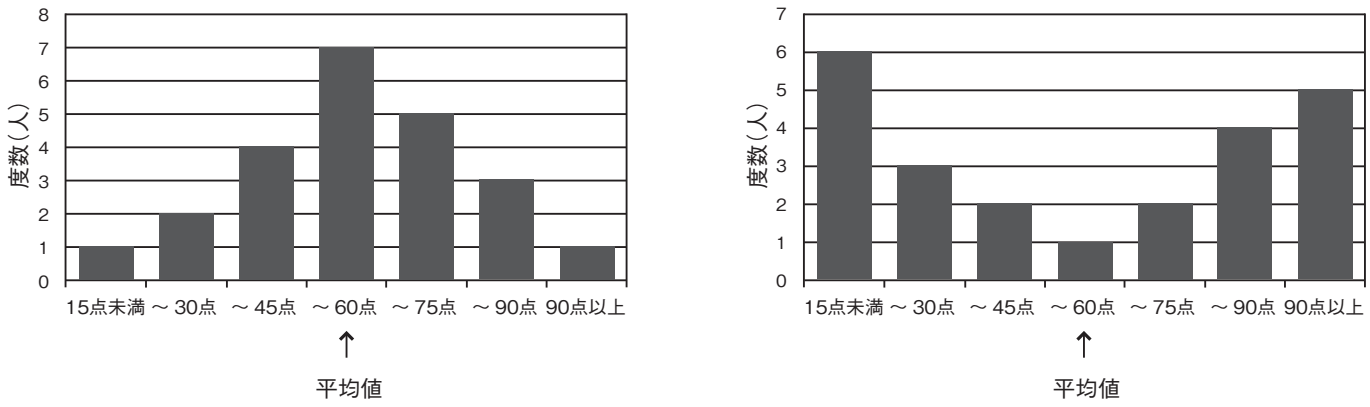

図2　ヒストグラムの例(平均値に度数が多い例[左]と少ない例[右])

4. 正規分布

度数分布を発展させて、正規分布についても簡単に説明しておきます。次回から紹介する統計的検定では正規分布を仮定することが多いので、正規分布とは何かについて事前の知識があると都合が良いです。

図1、**図2**のヒストグラムは、たかだか数十人のデータを元につくられたものでした。これを1万人、10万人と被検者数を増やしたときに描かれるであろうヒストグラムが正規分布です。正規分布は**図3**の太線のような形をします。つまり、被検者数が少ない状況では**図2右**のような偏ったヒストグラムが出現することもありえますが、非常に多くの被検者のデータが得られれば、平均値を中心に対称的なきれいな分布になるであろう、と考えるわけです。実際に私たちのまわりの多くの指標、身長、体重、BMI、腹囲、1RM、50m走タイム…いずれもデータがたくさん集まると、正規分布が現れます[5]。

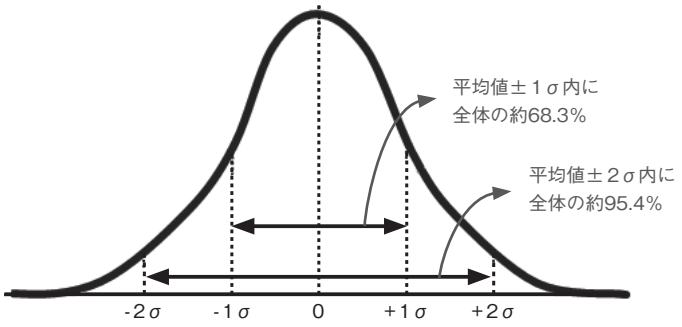

図3　正規分布(σ：標準偏差)

もうひとつ、違う例を使って正規分布を説明します。今、コインを振って、コインの表が出たら、駒を+1前進し、裏が出たらその場に留まることをします。駒ははじめ原点(0)にいます。コインを振ることを2回行なった後、駒の位置はどこになるでしょうか。コインの表と裏が出る確率は等しく1/2とします。

駒の位置は、コインの表裏の出方によって変わるので、何試行かやってみて駒の位置の度数分布を検討してみます(1試行終わったらまた原点に戻って再試行とします)。

1試行目：コインは、表→表と出ました。
　　駒の位置は、2です。
2試行目：コインは、表→裏と出ました。
　　駒の位置は、1です。
3試行目：コインは、裏→表と出ました。
　　駒の位置は、1です。
4試行目：コインは、裏→裏と出ました。
　　駒の位置は、0です。

4試行の度数分布は、位置0が1回、1が2回、2が1回で**図4A**のヒストグラムになります。このあと何試行か繰り返したとしても、確率的に、位置0、1、2に1:2:1のヒストグラムになるはずです。同様のことを、コインを4回、6回振ったあとの駒の位置の度数分布を検討すると、**図4B、C**のヒストグラムになるはずです。そして、コインを振る回数をもっと増やしていくと、正規分布(**図4D**)に近づきます。

このコインの例でわかる重要なことは、正規分布に従うデータは、平均値を中心にある確率でバラつく(分布する)ということです。コインを6回振ったときに、表→表→表→表→表→表と表が連続して6回続く確率は、64回に1回(1.56％)しかありません(**図4C**の位置6)。しかし、珍しいことではあるけれども可能性が全くない(確率が0)ということでもありません。つまり正規分布のカーブは度数の分布である

5　厳密には、「現れると考えられます」のほうが適切。必ずしも正規分布ではないために生じる問題も別にあるのですが、ここでは正規分布すると考えてください。

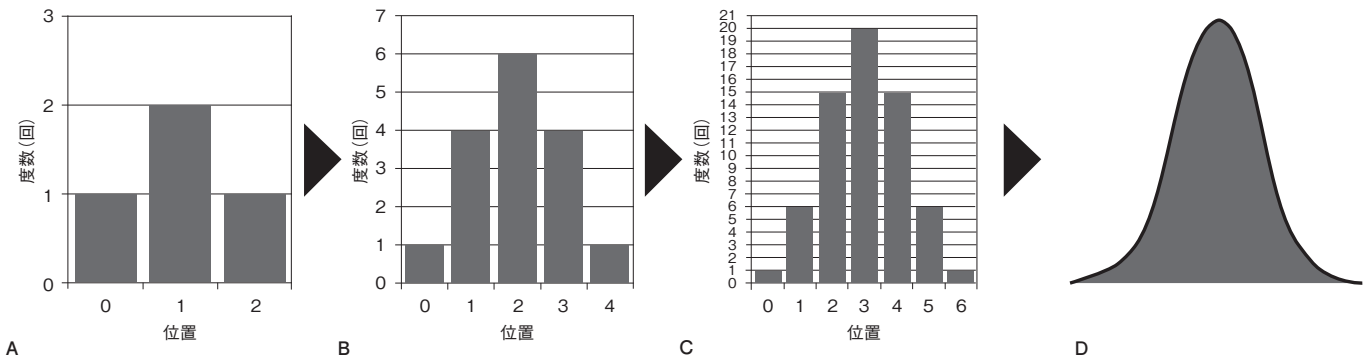

図4　度数分布から正規分布
A：2回、B：4回、C：6回コインを振ったときの駒の位置の度数分布（本文参照）、D：正規分布

と同時に、ある値が出現する確率を表しているともいえます。平均値周りの値は出現する確率が高くて、山の両端の値は出現確率が低い。そしてこの確率が、次回から説明する統計的検定につながってきます（この駒の位置の確率は次回も使いますので覚えておいてください）。

集団データが正規分布に従う場合、確率的に、平均値から±1標準偏差内に集団データの68.3％が含まれることがわかっています（**図3**）。したがって、論文内で「被検者10名の年齢は17.5 ± 0.3歳であった」と書かれている場合、被検者10名のうち6名か7名の年齢は、17.2～17.8歳の間であるということが推測できます。また面白いことに平均±1標準偏差のところで、正規分布のカーブは、上が凸のカーブから下に凸のカーブに切り替わります。

さらに、集団データの95.4％は平均値から±2標準偏差内に含まれることがわかっています。逆に考えれば、平均値より2標準偏差以上大きいか、小さいデータというのは、全体の4.6％しかなく、もしそういうデータが存在していればそれはかなり稀なデータだ、とわかるわけです。◆

Information

NSCAジャパン＆日本健康運動指導士会 第2回合同学術大会

テーマ：「健康的に年を重ねるために」

■日時：2014年3月1日（土）　9：00～17：45
■会場：国立オリンピック記念青少年総合センター 大ホール
■講演者（予定）：
寛仁親王妃信子殿下
下光 輝一（公益財団法人健康・体力づくり事業財団理事長）
森谷 敏夫（NSCAジャパン理事長）
安部 孝（インディアナ大学客員教授）
阿部 良仁（NSCAジャパン事務局長）
斉藤 智子（健康運動指導士、日本健康運動指導士会群馬県支部長）
谷ノ口 昭太郎（CSCS*D、NSCA-CPT*D、認定検定員、有限会社オールフォア）
藤竹 晃平（CSCS、NSCA-CPT、レベルⅠ認定）
■受講料：
NSCAジャパン会員・健康運動指導士会会員　6,000円
非会員　12,000円
■CEU：0.6（カテゴリーA）
■その他資格：　健康運動指導士・健康運動実践指導者、ADI/JAFA AQUA、高齢者体力つくり支援士
※それぞれ単位申請予定

12月上旬、受付開始予定です。詳細は受付開始後、ウェブサイトにてご覧ください。