課題研究ハンドブック Chapter６（試作版）

～統計と統計学～

このシリーズには時折“**統計**”という言葉が出てきます。「**集団における個々の要素の分布を調べ、その集団の傾向・性質などを数量的に統一的に明らかにすること。また、その結果として得られた数値**」ですが（『広辞苑』）、高校生の皆さんにはなじみ薄いかもしれません。そもそも、統計とはどんな理由で生まれたのか？　そのあたりから簡単に紹介しましょう。

統計の歴史

源流１：徴税や徴兵のベース

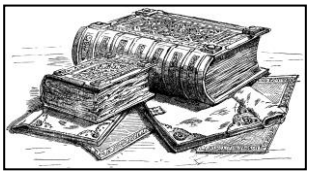

図１. The Domesday Book from “Historic Byways and Highways of Old England” (Andrews, 1900)＠Wikipedia。

統計の源流の一つは、きわめて政治的なもの、すなわち、国家による**徴税**や**徴兵**の基礎データの集積でした。例えば、中世イングランドでは 1086 年にノルマン朝が土地台帳（Domesday Book；図１）を作成、相続や土地争いの際の参考にされました。日本では徴税や徴兵に戸籍が使われましたが、最古の戸籍はおよそ 1200 年前です。

こうして、最初は権力者が“統計”を握っていたわけです。しかし、現代では、巨大企業がデータ集積を把握する状態になりつつあるようです。

源流２：大量なデータを処理、政策・ビジネスに役立てる

近年、Web で膨大なデータをリアルタイムで集積・分析するビッグデータが注目されています。この先駆けは 16 世紀の『**死亡週報（Bills of mortality)**』かもしれません。毎週ロンドンにおける死亡者数と死因を記録したもので、1662 年、社会統計の萌芽である「**政治算術**[1]」の先駆者 J・グラントが分析します。これが刻々集計されるデータを分析、政策に活かそうとする最初の試みです。数年後、ロンドンにペストが流行すると、『死亡週報』を購入している富裕層はいち早く察知、逃げ出します。一方、取り残された貧困層に多くの死者が出ます。

また、データ分析からビジネスが誕生した例に**生命保険**があげられます。従来、様々な人たちが葬式費用等をまかなうため組合を作りました。しかし、掛け金の計算がうまくいかず、挫折します。誰が何時死ぬか、把握しないと年齢ごとの掛け金が計算できないのです。この問題の突破口（**ブレークスルー**）は、天文学者の E・ハレーがドイツの一都市の人口データから死亡表（現在の**生命表**）を作ったことでした。死亡表から年齢層ごとの平均余命を割り出し、掛け金を計算することができました。こうして最初の生命保険会社が 1762 年に設立されます。

源流３：確率を考える

統計の源流の３つめは**確率**です。これも有名な話ですが、ある日、賭博師から相談を受けた哲学者のパスカルが数学者のフェルマーと書簡を交わしながら解いたのがきっかけの一つです。

なお、相談の一つは「**１つのサイコロを４回投げる。１回でも６が出た場合は勝ち」という賭けに挑戦すると、勝つことができた。しかし、「２つのサイコロを 24 回投げる。そこで６のゾロ目（６と６）が１回でも出た場合は勝ち」という賭けでは、「同じ確率から、勝てるだろう」と思ったのに、勝てなくなった。これは何故か？**」です。数学が得意の方は考えてみましょう。

[1] “Political Arithmetic”の略、17 世紀頃に社会現象を統計的に分析、将来を予想しようとします。

統計学とは何か？

統計学は大きく 2 つの分野、（1）対象集団を調べ、規則性・法則性を見出す“**記述統計**”と、（2）対象集団（母集団）のデータの一部をサンプル（標本）として抽出、規則性・法則性を見出す“**推測統計**”に分かれます。推測統計とは製品の不良品の抜き取り検査が典型です（全製品をチェックできないので、一部の製品を調べて不良品率を推測）。ここでは皆さんにとってなじみがありそうな事例で、記述統計を紹介しましょう（関西学院大学総合政策学部、2012）。

成績について“記述統計”してみよう

高等学校の３年生 40 名を対象に、以下の質問を実施した上で、英語や数学、社会の得点を調べたとします。

質問１：英語は得意ですか、不得意ですか？

質問２：数学は得意ですか、不得意ですか？

質問３：一週間の勉強時間（分）を答えて下さい。

図２はデータシートのイメージです。この数値データを統計ソフト等によって集計します。そして、グループごとにデータ数をまとめた結果を「**度数**」または「**頻度**」と呼び、度数分布表を作ります（表１）。

番号	組	氏名	性別	質問1	質問2	質問3	英語	数学	
1	1	上田	男	得意	得意	180	62	31	‥
2	1	鳥飼	女	不得意	不得意	231	50	71	‥
3	1	小沢	女	不得意	得意	151	78	55	‥
4	1	永沢	男	不得意	不得意	145	58	28	‥
5	1	山脇	男	得意	得意	221	51	74	‥
6	1	岡村	男	得意	得意	164	73	55	‥
7	1	小西	女	得意	不得意	108	50	31	‥
8	1	細川	女	不得意	得意	109	46	25	‥
9	1	山本	男	不得意	不得意	131	59	25	‥
10	1	土肥	男	得意	得意	156	51	39	‥
：	：	：	：	：	：	：	：	：	‥

図２．データシートの例

次に、この度数を縦軸にして**ヒストグラム**（**度数分布図**）を描きましょう。英語は左右対称のいわゆる**正規分布型**に分布しますが（図３ａ）、社会は「**右裾が長い分布**」になります。高校生の皆さんも、こうやってデータの全体像をつかむことが統計学の第一歩であることを是非覚えて下さい。

表１．度数分布表の例

度数分布表の一例：１組の英語の得点分布

階級	階級値	度数(頻度)	累積度数	相対度数	累積相対度数
31−40点	35	2	2	5.00%	5.00%
41−50点	45	10	12	25.00%	30.00%
51−60点	55	14	26	35.00%	65.00%
61−70点	65	10	36	25.00%	90.00%
71−80点	75	4	40	10.00%	100.00%
計		40			

階級：	データを分類するグループ
階級値：	階級を代表する値。一般に（階級の上限値＋階級の下限値）/2
度数(頻度)：	その階級に含まれるデータ数
累積度数：	度数を下の階級から積み上げた値
相対度数：	その階級の度数÷データ総数
累積相対度数：	相対度数を下の階級から積み上げた値

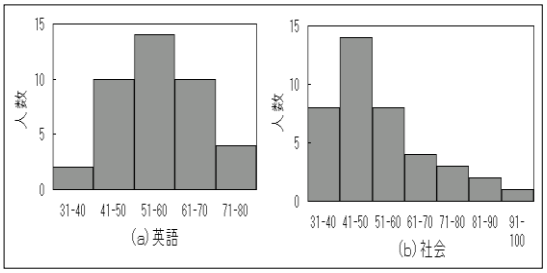

図３．（ａ）英語と（ｂ）社会の点数分布

基本統計量とは？

次に、「**基本統計量**」としてよく使われる数値を説明しましょう。

（１）**代表値**：比較の際、分布を代表する値で、主に次の３種類があります。

①**平均値**（**Mean**）

②**中央値**（**Median**）：中位数、メジアンとも呼び、データを大小順に並べた時の中央の値。

③**最頻値**（**Mode**）：モードとも呼び、最も頻繁に現れるデータの値。

ところが、これらの値が一致しないことがあります。例えば、10 人の学生に１週間の勉強時間を尋ねたところ、表２の分布例１になったとします。平均は 3.3 時間、中位数は 1.5 時間、最頻値は１時間です。

表２．１週間の勉強時間の分布例

学生10名の勉強時間の分布例1

	勉強時間																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
度数（人数）	5	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

学生10名の勉強時間の分布例2

	勉強時間																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
度数（人数）	0	0	0	0	0	0	0	0	1	2	2	2	3	0	0	0	0	0	0	0

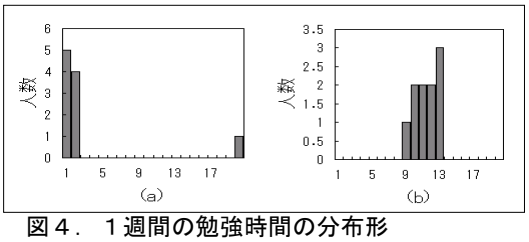

図４．１週間の勉強時間の分布形

このうちどれが代表値としてふさわしいのでしょうか？　これを判断するためには、ヒストグラムによる分布形からの判断が必要です(図４ａ)。その結果、平均 3.3 は、20 という他の値から飛び離れた値(＝「**はずれ値**」)に影響された結果だとわかります。

この場合、中位数（1.5）や、最頻値（１）の方が代表値としてふさわしいのです。では、表２の分布例２では？　この場合、平均 12.4、中位数 11.5、最頻値 13 です（図４ｂ）。12.4 や 13 は分布を代表するには大きすぎ、中位数の 11.5 がふさわしいようです。

表３．３つのデータセット

3つのデータセット（東京大学教養学部統計学教室編、1991、p35）

	0	1	2	3	4	5	6	7	8	9	10
A	1	0	0	2	0	4	0	2	0	0	1
B	1	1	1	1	0	2	0	1	1	1	1
C	0	0	0	1	2	4	2	1	0	0	0

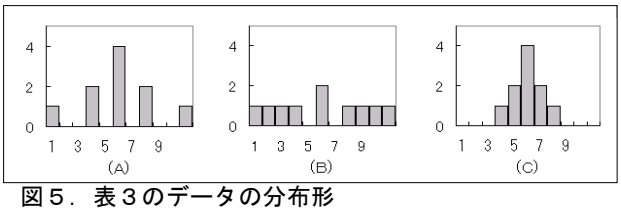

図５．表３のデータの分布形

（２）**データの散らばり具合**：図４で明らかなように、データの分布を判断する場合、平均も大事ですが、むしろ散らばり具合が重要です。表３のデータセットを見て下さい。A、B、Cとも平均値、中央値、最頻値はすべて５ですが、ヒストグラムの形はかなり異なります（図５）このような時、代表値だけではデータの分布を表現しきれないのです。

（３）**散らばりの尺度**：こうして、データの分布を見る際に重要になってくるのが、「データの散らばり具合」を表す指標です。これには、ふつう以下の二つの値が利用されます。

①**分散**（variance）：平均からの乖離の二乗の平均

②**標準偏差**（standard deviation）：分散の平方根Ｓ

「分散」の計算式を式１に示します。この二つは、後述する「データの標準化」で用いる重要な概念です。統計では、こんな風に計算しているのだ、と覚えておいて下さい。

式１．分散の方程式（教科書によって右の式を使うこともあります）

$$S^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\quad,\qquad S^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

簡単な統計の例について

ここで、皆さんにも関心があるかもしれないテーマを２つとりあげましょう。１つはデータの間の相関です。そして、もう一つはデータの標準化、すなわち偏差値です。

散布図と相関～二次元量的データの記述～

さて、２種類のデータの関係として、図２に戻って「勉強時間が長いほど、点数は高い？」という課題を取り上げましょう。興味をおぼえる方もいらっしゃるでしょう。

まず、データの全体像を把握するため、図２にでている 10 名分のデータから、勉強時間と得点を二次元の平面上にプロットし、全体像を把握する**散布図**を作ります。エクセル等のグラフ機能を使えば簡単です。それが図６です。英語の得点と勉強時間は関係が薄そうですが、数学では勉強時間と関係がありそうです。しかし、これはたんなる印象に過ぎません。

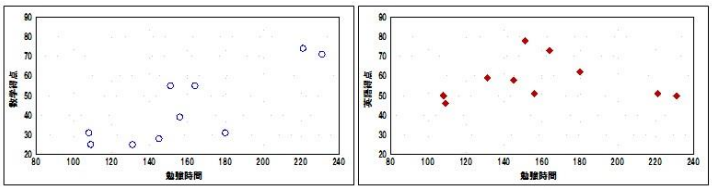

図６．数学（左）と英語（右）の得点と勉強時間

式２．相関係数の式

$$r=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$$

そこで、二つのデータの関係を数値的に表す指標＝「**相関係数**」が登場します。計算式を式２に示しますが、こちらも記憶しておくと後々便利です。なお相関係数はｒで示されます。

図２の 10 人の資料で計算してみましょう。勉強時間は平均 159.6 分(標準偏差 41.7 分)、英語 57.8 点（10.6 点)、数学 43.4 点（18.9点）です。相関係数は、英語がｒ＝-0.0022、数学が 0.8307 です。相関係数は＋１から－１までの値を示しますが、以下の特徴があります。

①相関係数がプラスの値の場合は、勉強時間と得点は**正の相関**（完全に相関すると＋１）
②相関係数が０に近いと、勉強時間と得点は**無関係**
③相関係数はマイナスの値は、勉強時間と得点は**逆相関**（完全に逆相関すると－１）

図７は図６の２つの散布図をあわせて、かつ、回帰直線を付け加えたものです。この図から明らかなように、勉強時間が多いと数学の得点が高い（正の相関がある）のに対して、英語の得点は勉強時間とほとんど無関係という結果になりました。

もっとも、相関は「二つの資料の関係が深そうだ」を意味するだけですから、「因果（原因と結果）関係」を証明しているわけではありません。つまり、「勉強時間が長いと数学の得点が高い」のか、「数学の得点が高い人はたくさん勉強する」のどちらかは不明です。

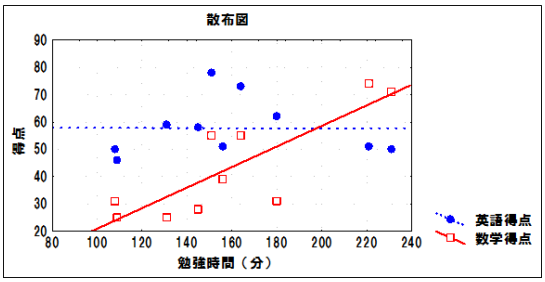

図７　英語と数学の比較

データの標準化～二つの量的データを比較する方法～

最後に、平均や分散が異なる２つのデータを比較する方法も説明しましょう。実は、これが「**偏差値**」の基礎です。上の 10 人の例に戻り、このうちの一人、英語 73 点、数学 55 点の生徒さんが「１科目で大学受験をする場合、どちらで受験するのが有利か？」を考えます。

まず、２つのデータの「標準化」が必要です。そこで、

（１）平均点の違いの影響を取り除くため、まず、この方の得点から平均点を差し引きます。
- **英語：73－57.8＝15.2 点**
- **数学：55－43.4＝11.6 点**

（２）次に、データの散らばり具合を調整します。これが「標準化（正規化・基準化)」です。具体的には上記の値を「標準偏差」で割ります（なお、標準化すると平均は０、分散は１にそろいます)。英語と数学の得点の標準偏差は 10.6 と 18.9 ですから、
- **英語：15.2÷10.6＝1.4337...**
- **数学：11.6÷18.9＝0.6137...**

（３）日本でよく用いられている学力偏差値はこの数値に 10 をかけてから、50 を足すので、
- **素点が 73 だった英語は 64.3**
- **素点が 55 だった数学は 56.1**

したがって、やはり英語で受験した方がよさそうだ、という結果になりました。

引用文献

関西学院大学総合政策学部編（2012）『基礎演習ハンドブック』関西学院大学出版会。
新村出編（1998）『広辞苑第５版』岩波書店。

2016 年 12 月
編集：関西学院大学総合政策学部・関西学院千里国際高等部