情報の授業における統計の扱いについて

名古屋高等学校　中西渉*

2006 年 10 月 21 日

概要

高校の数学に「確率・統計」があった頃も，大学入試でほとんど出題されないため，統計的分野を授業で扱うことはほとんどなかった。その後何度かカリキュラムが変わり，教科「情報」もできたが，統計にまともに向き合っていない状況に変わりはない。

しかし，コンピュータを使うのが当り前，大量のデータを扱うことが容易になったのに，統計をろくに知らないというのは非常にもったいない話だ。そこで本校では情報の授業において統計の基本的な部分に触れることにした。本発表ではその実践内容を紹介する。

1　高校の授業における統計

1.1　数学

高校の数学で統計を扱うのは数学 A「場合の数と確率」と数学 B「統計とコンピュータ」，数学 C「確率分布」「統計処理」である。しかし，多くの学校は数学 B・C のこの分野を履修しない。というのは，入試で数学 B や C を出題する大学のほとんどが「数列」「ベクトル」や「行列とその応用」「式と曲線」のみを出題範囲としているからだ[*1]。

したがって彼らは数学では統計についてほとんど学ばない。期待値だけは習っているが分散や標準偏差を知っているのは稀であり，それもたいていは計算に終始してしまって，その数値の意味するところはわからないままだ。

1.2　情報

では情報はどうか。実は学習指導要領には「統計」という単語は一回しか登場しない。情報 C の「(3) 情報の収集・発信と個人の責任」に関する「内容の取扱い」として

> イ[*2]については，適切な題材を選び，情報の収集から分析・発信までを含めた一連の実習を中心に扱うようにする。情報の分析については，表計算ソフトウェアなどの簡単な 統計 分析機能やグラフ作成機能などを扱うようにする。

と触れているだけだ。もちろん「簡単な統計分析」というのがどの程度のことなのかは示されていない。いくつかの教科書を見ると，表計算ソフトの AVERAGE，MAX，MIN，RANK，STDEVP といった関数を使った例がある。しかし基本的に計算させっぱなしで，その値を読み取ることはしても，値自体を評価することはしていない。ある教科書は演習で標準偏差を計算させるものの，その意味は「値のばらつき云々」というように漠然としか書いていない。だとしたら，紙の上でやっていたことと何が違うというのだろう。むしろ計算式がわからなくなった分だけ後退しているとさえ言えるのではないか。

1.3　統計を知らない大人の代表—教員

そんな風だから，高校では統計についてほとんど学習しない。大学でも統計に関する授業を選択しなければそのままだろう。論文を書くときに教官に指摘されるなどして必要性に気づけばいいが，そういったチャンスに恵まれなければそのまま社会人になっていく。

* watayan@meigaku.ac.jp

*1 「確率分布」を出題する大学（それ自体非常に少ないが）でも「確率の計算」までを範囲とすることが多い。

*2 情報通信ネットワークを活用した情報の収集・発信

私見だが，教員の多くはそんな風に統計を知らないまま現職に至っているのではないだろうか。たしかに職員会議には生徒の成績データから得られた数値がずらっと並ぶ。時には小数点以下の数字まで見つめているが，その値がその桁まで有効かどうかを考えたことがあるかといえばはなはだ疑問だ。しかしそのような意見を耳にしたことは一度もない。そうして，なぜか決まって「小数第 1 位まで」算出したデータがいつまでも作られ続ける。

2　本校での実践

以下，本校の情報の授業における統計の扱いについて説明する。

2.1　値のばらつき

2.1.1　ビュッフォンの針

ビュッフォンの針というのは

> 等間隔に平行線をひき，その間隔の半分の長さの針を無作為に落としたとき，針が平行線と共有点を持つ確率はどれだけか。

という問題である。もちろん積分によってその確率を求めることはできるが，せっかくパソコンがあるのだからモンテカルロ法によって近似値を求めることにする。

実習では，PEN[*3]用のプログラムを生徒に配布し，実行させてその値を記録させる（図 1）。「この値の逆数は君たちがよく知ってる値なんだが...結果には誤差があるから数回やってその真ん中くらいの値を見てごらん」と言って予想させると，何人かが「ひょっとして π？」と口にする。これで答えにたどり着いて一件落着なのだが，1000 本も針を落として計算した割には誤差が大きいことを強調して，次の話につなげる。

2.1.2　期待値って？

正常なサイコロなら各目の出る確率は 1/6 だから，60 回振ればそれぞれ 10 回くらいずつ出る...はずである。しかし表計算のマクロを使って実際に

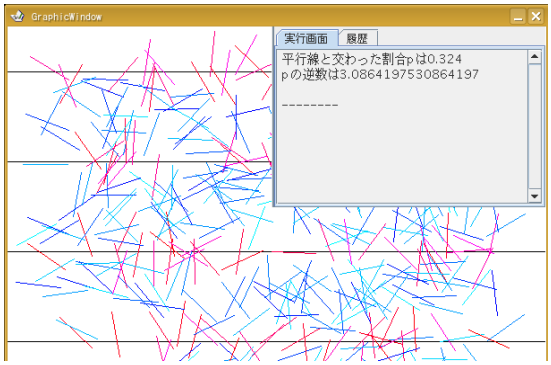
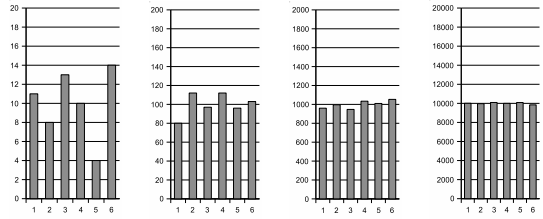


図 1　ビュッフォンの針

やってみると，4 回や 5 回くらいの誤差が出ることも珍しくない。しかし 600 回，6000 回，60000 回やってみると，その相対的誤差はだんだん小さくなっていく（図 2）。これは期待値だけでは説明できないことだ。そこでその値のばらつきを評価するためのシミュレーションを行なう。



(a) 60 回　(b) 600 回　(c) 6000 回　(d) 60000 回

図 2　サイコロを振って各目の出た回数のグラフ

2.1.3　円周率を求める

正方形の的に一辺を半径とする 1/4 円の当たりを作っておき，ランダムに点をばらまく。当たりの確率は $\pi/4$ であるから，当たりの点の数を総数で割ればおよそ $\pi/4 = 0.785...$ になるだろう。この方法によって円周率の近似値が求められることを説明して，配布した PEN 用のプログラムを実行させてみる（図 3）。すると中にはかなり近い値がでる生徒もいるが，けっこう誤差が大きい者もいる。この誤差の大きさが点の個数に影響することはサイコロのシミュレーションから予想されるので，それを評価してみる。

[*3] 初学者向けのプログラミング学習環境で，Java の実行環境があれば利用できる。
`http://www.media.osaka-cu.ac.jp/PEN/`

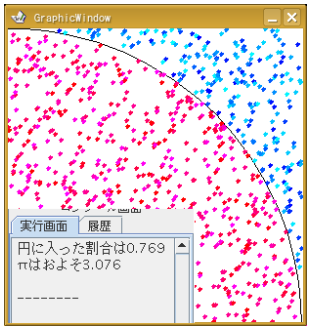


図 3　円周率の近似値を求める

いちいち記録をとるのは面倒なので，表計算のマクロで計算するプログラムを配布して実行させる。点の数を 10，100，1000，10000 と増やしていき，それぞれ 15 回ずつ上記のシミュレーションを行なって結果を表示させる。その 15 回の最大値と最小値の差—これが値のバラツキとみなせるだろう—をとって，回数との関係を考察しようということだ（図 4）。

	A	B	C	D	E
1	回数	10	100	1000	10000
2		実行	実行	実行	実行
3					
4	1回目	0.9	0.72	0.775	0.7834
5	2回目	0.9	0.79	0.774	0.7900
6	3回目	0.7	0.75	0.787	0.7821
7	4回目	0.8	0.81	0.775	0.7889
8	5回目	0.4	0.83	0.766	0.7917
9	6回目	0.7	0.78	0.794	0.7907
10	7回目	0.6	0.75	0.784	0.7835
11	8回目	0.7	0.85	0.790	0.7852
12	9回目	0.6	0.71	0.791	0.7824
13	10回目	0.9	0.76	0.798	0.7793
14	11回目	0.7	0.82	0.773	0.7860
15	12回目	0.8	0.80	0.786	0.7816
16	13回目	0.6	0.74	0.809	0.7805
17	14回目	0.9	0.77	0.786	0.7889
18	15回目	0.9	0.78	0.794	0.7852
19	最大	0.9	0.85	0.809	0.7917
20	最小	0.4	0.71	0.766	0.7793
21	最大-最小	0.5	0.14	0.043	0.0124

図 4　回数とバラツキの関係

たしかに回数が増えればバラツキは減るが，反比例しているわけではない。よく見ると，回数が 100 倍になるとバラツキがおよそ 1/10 倍になる，つまり**バラツキは回数のルートに反比例している**ことをこの表から読み取らせることができる（といってもかなり強引に誘導する必要があるが）。

これでとりあえず一つの重要な結論が得られた。次のテーマに移ろう。

2.2　正規分布

2.2.1　平均，標準偏差

統計的にデータを扱う上で重要な値として，平均と標準偏差があることを紹介する。どのような計算式で得られるかを一応見せてはおくが，それはさほど重要だとは考えていない。数学の授業なら話は別だが，我々はパソコンで即座に計算できるのだし，ここで重視したいのはその値が持つ意味なのだから。

生徒にとって，平均は容易にイメージできるが，標準偏差はそうはいかない。「平均との差の二乗の平均のルート」といってもピンとこないし，大雑把に「平均との差の平均」といってしまうと標準偏差でなく平均偏差になってしまう。標準偏差だとこういうことができる，というメリットについて説明したい。

2.2.2　正規分布

そこでいくつかの度数分布図を見せて，数が多いと度数分布は「正規分布」といわれる決まった形に近付くことが多いということを説明する（根拠は述べない）。そして標準偏差はそのグラフの平べったさを表すことを説明する。

しかしそれだけでは標準偏差の定量的な意味がわからない。そこで

- $m \pm \sigma$ に 68.3% が含まれる。
- $m \pm 2\sigma$ に 95.4% が含まれる。
- $m \pm 3\sigma$ に 99.7% が含まれる。
- $m \pm 1.64\sigma$ に 90% が含まれる。
- $m \pm 1.96\sigma$ に 95% が含まれる。
- $m \pm 2.58\sigma$ に 99% が含まれる。

というように具体的な値がわかっていることを説明し，これを実際のデータ（ある年の身体測定の結果を用いた）で検証する実習を行なう（図 5）。

これによって，正規分布では σ を単位にして割合が決まるという事実が示される。ついでに偏差値の計算式を教えて「偏差値 60 というのは $m+\sigma$，偏差値 40 は $m-\sigma$ だから，偏差値 40〜60 の間に全体の約 68% が含まれる。ということは残り 32% を上下に振り分けると，偏差値 60 というのは上位 16% く

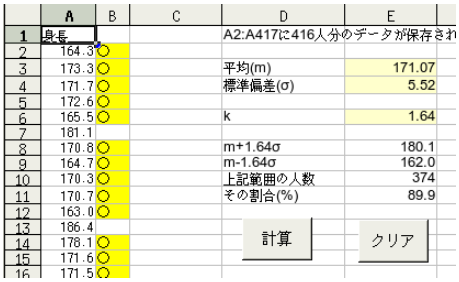


図 5　$m \pm k\sigma$ に含まれる割合

らいの位置にいるということになる」といった説明もできる（分布が正規分布になっていることが前提ではあるが)。

2.3　推定

2.3.1　母平均の推定

本当なら中心極限定理を根拠に公式を作るのだが，これまでに得た結果を用いて次のようにコジツケることができる：

> 1 つのデータが $m \pm 1.96\sigma$ の範囲に含まれる確率は 95%，つまり 95% の確率で m からの値のバラツキは 1.96σ 以下になる。n 個の標本の平均をとれば m からのバラツキは（回数のルートに反比例するのだから）$\frac{1.96\sigma}{\sqrt{n}}$ 以下になる。つまり標本平均と m の差が $\frac{1.96\sigma}{\sqrt{n}}$ 以下になる確率は 95% なのだから，m の 95% 信頼区間は 標本平均 $\pm \frac{1.96\sigma}{\sqrt{n}}$ である。

公式ができたところで，いくつか計算練習をさせてみる。

2.3.2　母比率の推定

母比率の 95% 信頼区間 $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$ まで導くのは無理なので，これについては公式を与えて計算練習をさせることにする。

出来合いの公式を与えてしまうのはつまらないが，母比率の方が計算結果を楽しめるのだ。たとえば「1 の目が 100 回中 20 回出たサイコロの，1 の目の出る確率の 95% 信頼区間」と「...10000 回中 2000 回...」を比較すると，2.1.2 で期待値だけではわからなかったことがわかってくる。

2.3.3　検定

もう一つ踏み込めば検定の話に持ち込むこともできるのだが，実際に授業をしてみると帰無仮説といった考えや，「... H_0 は棄却できないので，学力が等しくないとは断言できない（しかし等しいと断言するわけではない)」のような多段の否定が難しいらしく[*4]，かえって生徒が混乱してしまったようなので 2 年目以降は推定だけを教えることにしている。

正直なことをいえば，検定で結論を断じるおもしろさを味わってほしいという思いもあるのだが，丸め誤差以外にも値がばらつく要因があるということを生徒が実感できれば，ある程度目標は果たしたことになるのだから無理はしないでおこう。

3　雑感

正直いってこの一連の話は強引なこじつけが多く，数学屋としては抵抗がないわけではない。しかし生徒たちがいずれ統計的手法を使うときになればその場で改めて勉強せざるをえないのだから，正確さは現時点ではさほど重要ではない。それよりも推定などといった考え方のイメージや必要性をアピールすることこそが必要だと考える。だからこの時点では公式や値を暗記することは要求せず，実際に計算せよ，その結果を見よ，イメージせよということを強調している。

これが「情報」の授業でやるべき内容なのかと言われれば疑問はなくもないが，しかし他に彼らに統計を教えるチャンスはないのが現実だ。我々の判断が正しいかどうかを云々するのはまだ先のことだと考えている。

[*4] しばらくコンピュータの前を離れて計算ばかりやっているので飽きてきたというところもあったような気はする。