| | |
|---|---|
| 高校生のための現代数学講座 | 東京大学 |
| 「確率と統計」 | 玉原国際セミナーハウス |
| 講義 (5)　宮本　安人 | 2016 年 7 月 23 日 |

# 「正規分布と偏差値」

**シンプソンのパラドックス．** 100 人の生徒を持つＡ高校とＢ高校が，同じ数学の試験を行った結果，左の表のようになりました（表の数字は [5] より引用）．左の表から,「Ａ高校の方が数学に強い」となります.

| | 全生徒 |
|---|---|
| Ａ高校 | 平均 78 点 (100 人) |
| Ｂ高校 | 平均 72 点 (100 人) |

| | 理系 | 文系 |
|---|---|---|
| Ａ高校 | 平均 80 点 (90 人) | 平均 60 点 (10 人) |
| Ｂ高校 | 平均 90 点 (10 人) | 平均 70 点 (90 人) |

次に，理系と文系とに分けて内訳をみると，右の表になりました．右の表から，理系でも文系でも「Ｂ高校の方が数学に強い」となりま…，あれ !?，どこがおかしかったのだろう….

同じようなことは，次のＡ大学とＢ大学の就職率に関する表でも起きています．大学全体（左の表）ではＡ大学の方が就職率が高いのに，理工学系と人文学系とに分けて内訳（右の表）をみると，共にＢ大学の方が就職率が高くなりました．両大学の進学説明会では，きっと，Ａ大学は左の表を使ってアピールし，Ｂ大学は右の表を使ってアピールすることでしょう．一体，どちらが適切なのでしょうか？

| | 全学生 |
|---|---|
| Ａ大学 | 78%(100 人) |
| Ｂ大学 | 72%(100 人) |

| | 理工学系 | 人文学系 |
|---|---|---|
| Ａ大学 | 80% (72 人/90 人) | 60% ( 6 人/10 人) |
| Ｂ大学 | 90% ( 9 人/10 人) | 70% (63 人/90 人) |

これは，集団全体の結果と部分集団の結果が必ずしも一致しないことを示唆しています．この現象は，それ以前から知られていましたが，アメリカの統計学者シンプソンによる指摘（[6]）以降，シンプソンのパラドックスと呼ばれています.

この現象が起こる理由を説明するためには，多変量解析の知識が必要となりますのでここでは触れません．しかし，一番目の例のカラクリを簡単に説明すると,「数学の試験なので，全生徒の平均は，理系学生が多いＡ高校の方が初めから有利」，ということです．この例から，教訓を得るとすると,「前提条件（理系が多いか？文系が多いか?）が異なると，全体と部分では結論が異なることがあるので注意しよう」ということになるでしょう.

| | 英数の総正答率 |
|---|---|
| A君 | 55.4%(61 問/110 問) |
| B君 | 35.4%(39 問/110 問) |

| | 数学 | 英語 |
|---|---|---|
| A君 | 100 問中 60 問正解 | 10 問中 1 問正解 |
| B君 | 10 問中 9 問正解 | 100 問中 30 問正解 |

最後に，A君とB君は英語と数学の試験を受け，表のようになりました．B君にとってみれば，数学も英語もA君より正答率が高いのに，総正答率では負けているので，総正答率で成績が決まった場合，文句も言いたくなるでしょう．前提条件が異なる場合，どのような指標を用いて比較すればよいのでしょうか？

**全ての道は正規分布へ．** コインを投げて表の出た回数を競うゲームを考えます．ルールは簡単で，$N$ 回コインを投げ表の出た回数が得点となります．但し表と裏は各 1/2 の確率で起こるものとします．$N = 2$ の場合を考えます．表表，表裏，裏表，裏裏の全 4 通りなので，0 点となる確率は 1/4，1 点となる確率は 1/2，2 点となる確率は 1/4 となります．次に，$N = 3$ の場合を考えます．表表表，表表裏，表裏表，表裏裏，裏表表，裏表裏，裏裏表，裏裏裏の全 8 通りなので，0 点，1 点，2 点, 3 点となる確率は，それぞれ，1/8, 3/8, 3/8, 1/8 となります．同様に，$N = 4$ の場合，全 16 通りとなり，0 点，1 点，2 点, 3 点，4 点となる確率は，それぞれ，1/16, 1/4. 3/8, 1/4, 1/16 となります．$N$ をだんだん大きくしたときに，規則性はないのでしょうか？

規則性を探る前段階として，$N$ 回投げるゲームのときに，$k$ 点 $(0 \leq k \leq N)$ となる確率を求めることを考えます．下の図は，○を表，●を裏とするとき，$N = 14$ で $k = 8$ となる例です．

○○●●○○○●●●○○●○

この図を見ると，$N$ 回投げるゲームでは，全ての場合の数は，コインを一回投げるごとに○か●の 2 通りあるので全部で $2^N$ 通り，そのうち $k$ 点となるのは，$N$ 個のなかから○を $k$ 個選ぶ場合の数と同じなので，${}_NC_k$ 通りとなります．ここで，

$$ {}_NC_k = \frac{N!}{k!(N-k)!} = \frac{N(N-1)(N-2)\cdots(N-k+1)}{k(k-1)\cdots 3\cdot 2\cdot 1}. $$

よって確率は，${}_NC_k/2^N$ となります．

次に，$g(k)$ を次で定義します：

$$ g(k) := \frac{{}_NC_k}{2^N}. $$

$N$ が大きいときのグラフを描くとき，問題点が３つ出てきます : (1) グラフがだんだん右に移動する，(2) 高さがだんだん低くなる，(3) 幅がだんだん広くなる．これら

の問題点を解決するのが，少し技巧的ですが次の変数変換です：

$$f(x) = a_N g\left(\frac{N}{2} + x\sqrt{N}\right).$$

$g$ の中の $N/2$ で (1) の右に行くことを抑え，$a_N$ を適当に大きくなる増大する数列とすることで (2) の低くなることを抑え，$\sqrt{N}$ によって (3) の幅が広くなるのを抑えています．この変数変換によって，$f(x)$ の概形を正しくとらえることができます．この考え方（＝変数変換）は，単純ですが数学や物理学でよく用いられる考え方で，スケール変換（または尺度変換）と呼ばれています．ところで，$x\sqrt{N}/4$ は必ずしも自然数ではないですが，ここでは $x\sqrt{N}/4$ が自然数となるような場合を考えるということにします．若干，高度な式変形が必要ですが，下記の計算を乗り越えると結論にたどり着きます：

$$\begin{aligned} f\left(x + \frac{1}{\sqrt{N}}\right) - f(x) &= \frac{a_N}{2^N}\left({}_N C_{N/2+x\sqrt{N}+1} - {}_N C_{N/2+x\sqrt{N}}\right) \\ &= \frac{a_N}{2^N}{}_N C_{N/2+x\sqrt{N}} \frac{-2x\sqrt{N} - 1}{N/2 + x\sqrt{N} + 1} \\ &= f(x)\frac{-2x\sqrt{N} - 1}{N/2 + x\sqrt{N} + 1}. \end{aligned}$$

両辺，$1/\sqrt{N}$ で割り，極限 $N \to \infty$ を考えると，

$$f'(x) = \lim_{N\to\infty} \frac{f\left(x + \frac{1}{\sqrt{N}}\right) - f(x)}{\frac{1}{\sqrt{N}}} = \lim_{N\to\infty} \frac{f(x)(-2Nx - \sqrt{N})}{\frac{N}{2} + x\sqrt{N} + 1} = -4xf(x).$$

大学で行う常微分方程式の理論を用いると，$f(x) = Ce^{-2x^2}$ のみが，この関係式を満たす関数であることが分かります．従って，$f(x)$ はこの関数に近づくことが予想されます．もう一度変数変換 $x \mapsto x/2$ し，特に $C := 1/\sqrt{2\pi}$ とした関数

$$f_0(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

は定積分が 1 となる特別な関数で（$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = 1$），ガウス分布または標準正規分布と呼ばれています．例えば，さまざまな測定における「誤差」の分布は，測定の回数を大きくするとこの関数に従うことが数学的に証明されており，中心極限定理と呼ばれています．この定理は，適当なスケール変換によって $f_0(x)$ に収束することを主張しているので，$f_0(x)$ のグラフを見れば，データの散らばり具合や平均からのズレが分かることになります．この関数のグラフを描いてみましょう．

**偏差値とは何か．** データの分布が標準正規分布 $f_0(x)$ の場合，$T(x) = 10x + 50$ を偏差値と定義します．この分布の場合，平均は明らかに $x = 0$ なので $T(0) = 50$．これは平均では偏差値 50 となることが分かります．$x = 1$ の場合は，偏差値は $T(1) = 60$，$x = -0.5$ の場合は，偏差値は $T(-0.5) = 45$ となります．さまざまな平均や散らばり具合を持ったデータに対しても，中心極限定理によって適当なスケール変換で $f_0(x)$ となるので，データの種類に関わらず平均値からどの程度離れているか (どの程度稀なことか) を測ることができます．日本人の男性の身長の統計によると，例えば 18 歳のある男子の身長が 168.3cm で体重が 71.6kg の場合，身長と体重の偏差値はそれぞれ 45.0 と 60.0 となります (平均は身長 171.12cm, 体重 62.64kg)．ちなみに，体重の偏差値が 60 ですが，この場合，18 歳男子全体の中で重い方から 15.9%のところに位置しています．同様に偏差値 45 は，身長の高い方から 69.2%の位置です．偏差値を用いると，平均からどの程度離れているか，または全体の中でのどこに位置しているかが，身長や体重などデータの種類に関わりなく分かります．

さて，偏差値というと馴染みがあるのが，学力偏差値だと思います．ここでは，学力試験の結果を表すのに用いる偏差値を，学力偏差値と呼ぶことにします．その話の前に，ケトレー―を紹介したいと思います．アドルフ・ケトレー―(1796–1874) はベルギーの天文学者，数学者，統計学者，社会学者で，初期の頃は天文学を研究していました．当時，新しい学問であった統計学は，主に天文学で応用されておりました．元々数学や天文学を研究していたケトレー―は，自然科学のみならず，犯罪率，結婚率，自殺率など社会学的なデータに対して統計学の手法を持ち込み,「あらゆるデータは正規分布に従う」と考えました ([3])．この背景には，1810 年にフランスの数学者ピエール＝シモン・ラプラスによって，中心極限定理が証明されたことがあります ([7])．また，ケトレー―は社会学的データのみならず身体的データについても研究しました．スコットランドの数千人の兵士の「胸囲」を測定し，その数値の分布が天体観測の結果と同様に，正規分布に従うと主張しました ([2])．

時代は下り，1950 年代，東京都港区の中学校の理科教諭であった桑田昭三 (1928–2016) は，どの高校を受験するかを指導するときに，中学浪人を避けるためになるべくブレの少ない客観的評価の必要性を感じていました ([1])．というのは,『当時の受験校選びはベテラン教師の経験と勘が頼り．教師の言葉は生徒，父母にとっては“ご託宣”のようなもの ([4]) で，本人や父母が納得する指標が存在しなかった』からです．進学指導の科学への挑戦を決意し 3 年の模索の末に，ケトレー―の法則に出会いました ([1]).『入学試験における受験生の学力分布は，正規分布であるとみなすことができる』と 仮定し，可能な限り大勢の受験生の学力データを集め，この仮説を実証的に検証しました ([1])．また，確率論や統計学に行きつき，当時は心理学で用いられていた偏差値を応用し，学力偏差値を考案しました．便利な指標は同僚教師から区内，都内へとまたたく間に広がりました．評判を聞きつけたテスト業者の進

学研究会から誘われ，1963 年に転職し偏差値や学力評価の研究を行いました．

# 参考文献

[1] 桑田 昭三,『偏差値の生みの親・桑田昭三氏へのインタビュー』, SHIKEN: JALT Testing & Evaluation SIG Newsletter. **14** (2), October 2010, pp. 6–10.

[2] 小林 祐児, 『『平均値』をどう読むか-社会学的統計史』, http://www.jma-net.com/1839 (2016 年 7 月 6 日)

[3] 「アドルフ・ケトレー」『フリー百科事典　ウィキペディア日本語版』, http://ja.wikipedia.org/wiki/アドルフ・ケトレー (2016 年 7 月 6 日)

[4] 梁瀬 誠一,『[新教育の森] キーワードの軌跡　今週のテーマは ··· 偏差値　受験生減少で崩壊か』, 毎日新聞朝刊, 1999 年 2 月 6 日.

[5] マスオ,『高校数学の美しい物語』, http://mathtrain.jp/simparadox (2016 年 7 月 5 日)

[6] E. H. Simpson, *The Interpretation of Interaction in Contingency Tables*, Journal of the Royal Statistical Society, Series B **13** (1951), 238–241.

[7] 山本 誠志,『図解 統計がわかる本』, 学研教育出版, 2013 年.