論文数分布へのポアソン近似

片　岡　佑　作

目　　次

要　　旨

この論文は紀要に提出される年間論文数をポアソン分布で近似することを考える。x をポアソン変数、j をとりうる数値として以下のような表がある。

表

j	0	1	2	3	4
n_j/n	0.250	0.441	0.162	0.103	0.044
p_j	0.287	0.358	0.224	0.093	0.029

注）n：標本数、n_j : j に関する度数

この表の2行目を見て標本平均 $\bar{x}$ を計算すると $\bar{x} = 1.250$ である。さらに $\lambda = 1.250$ とおいてポアソン確率 $Pr(x = j)$ をポアソン分布表から読むと以下のようになる。つまり、$Pr(x = 0) = 0.287$、$Pr(x = 1) = 0.358$、$Pr(x = 2) = 0.224$、$Pr(x = 3) = 0.093$、$Pr(x = 4) = 0.029$ である。これらの数値は表の3行目にある。すぐわかるように対応するこれらの数値のくいちがいの大きさはわずかである。

次に提出された論文数に関する T- スコア（偏差値）の問題に移る。T- スコアの考え方は本質的に正規変数を平均50、分散 10^2 の正規分布に移しかえるものである。すでに見ているように論文数 j はポアソン変数によってよく近似することができる。こうして紀要に提出された論文数について T- スコアを考えることはあきらかに不適当である。

キーワード：正規分布、ポアソン近似、T スコア、z スコア、汚染されたデータ

1　序

この論文の主要な目的の1つは、教員が報告する（年間）論文数データがポアソン分布にしたがうことを1つの例から示す点にある。これは直観的にあきらか、つまり論文本数のほとんどは0か1に近いが、例外的に報告論文数の多い教員も存在するというものであるが、この点を実際に調べた最近の研究はあまりない。ポアソン分布にしたがう変数のその他の例としては交通事故件数、1ページあたり誤植数などがある（ポアソン分布については井上［1］、Fisz［6］、Johnson-Kotz-Kemp［7］を見るとよい）。

以下、具体的に述べると調査対象はA大学における教員別の紀要論文数についての資料から特定のページをとり出し、そこにリストされたものである。1つ注意をしたいのはリストにある教員群は1965年から1995年までにある期間のみ在籍しているものであって例えば異なる教員A_1、A_2について在籍期間、在籍時点も異なる。したがって教員A_1、A_2の論文数としては対応する期間をT_1、T_2とし、A_1/T_1、A_2/T_2とする必要がある。T_1、T_2については同一資料の前半のページに紀要の発行時点がリストされているのでこれを使った。

教員A_1が最初と最後に報告した紀要の発行時点をt_{10}、t_{11}として$T_1 = t_{11} - t_{10} + 1$とするのである。したがって、$T_1 \leq$（教員$A_1$の在籍期間）で$T_1$は在籍期間のほぼ下限である点に注意する必要がある。

1）こうしてA_j/T_j　$j = 1, \cdots$について例えば

i	A_j/T_j
0	0～0.5
1	0.5～1.0
⋮	

のようなクラス分けをするとiがpoisson変数になっている点を以下に示す。また、ここでの注意点としては対象とする標本群をいく分せばめる必要があるということである。これは制限をかけることによってcontaminated dataの混入を防ぐことができるからである。つづいて

2）論文数などのpoisson変数はその分布のピークが左によっているが、こうした集団群（$j = 1, \cdots\cdots$）の1つに偏差値（D_{ji}）の計算をすることがいかにナンセンスであるかを以下具体的に示す（Dは欧文献でTスコアと書かれる。また、もとの正規化したものはzスコアとされる）。

偏差値の分布は平均50、分散10^2の正規であり、もちろん左右対称、$Pr(50 > D_{ji}) = 0.5$など

であるが元の変数が poisson であれば、そこから作られた偏差値 D'_{ji} については $Pr(50 > D'_{ji}) > 0.5$ である。原因は問題の poisson の分布のピークが極端に左によっているからである。

3）つづいて偏差値計算の注意点をあげたうえで 1 つの簡単な例を示す。集団の分布形のピークが左にあるようなベルヌーイ分布を考えると（ピークが左にあるということは能力が平均的にいく分おちる構成員が多数で、かつすぐれた力をもつ構成員もわずかなから存在する、という仮定である)、このケースの偏差値を決めるものは構成員 1 人 1 人の score ではなく、集団に対する高能力保持者の割合のみである。こうしてこの割合が高まればこの集団の偏差値は当然すべて下がる。また、異なる集団間で高能力保持者数が一定であれば、構成員数が少ない集団の偏差値は大集団のそれよりもつねに下方に位置し、もし、すべての集団の構成員について並べかえをすれば小集団がきまって下方に位置する。

以下、1）を 2、2)、3）を 3 において証明をつけて示す。

2　ポアソン近似

表 2－1－1、表 2－1－2 は A 大学が刊行する紀要から集計された教員別の報告論文数のリストである。詳しく言うと、表の第 3 列目が論文数、第 2 列目はある教員が最初の論文を報告した年から最後の論文刊行年までの期間を示している。例えば 1）の教員であれば、期間は 4（年間)、論文数は 3 である。4 列目の 0.75 は 1 年あたりの報告論文数（論文数/期間）である。0.75 をこのように解釈するのは適当ではないかもしれないが、いちおう近似的な値として採用する。また期間において、ブランクの部分は論文数 1 に対応している。

ところで、とり上げる問題は次のようなものである。

⑴　論文本数の分布はどのようなものか。
⑵　より正確には教員が報告する 1 年間あたりの論文数の分布はどうなるか。
⑶　以上の経験分布を理論から説明する場合、あてはめる分布は何か。

まず、表 2－1－1、表 2－1－2 から必要とされる頻度分布を作成すると表 2－2、表 2－3 のようになる。

表 2－1－1

	期間	本数			期間	本数	
1）	4	3	0.75	38）	17	8	0.47
2）	2	2	1.0	39）		1	
3）	3	2	0.67	40）	1	2	2.0
4）		1		41）	7	3	0.43
5）		1		42）	10	7	0.7
6）		1		43）	5	2	0.4
7）		1		44）	6	3	0.5
8）	2	3	1.5	45）	5	6	1.2
9）	16	37	2.31	46）	6	4	0.67
10）		1		47）		1	
11）	5	4	0.8	48）	3	3	1.0
12）	3	2	0.67	49）	4	2	0.5
13）	6	3	0.5	50）	4	3	0.75
14）	4	5	1.25	51）	2	2	1.0
15）		1		52）	8	4	0.5
16）	10	8	0.8	53）	16	10	0.625
17）	4	6	1.5	54）	4	2	0.5
18）		1		55）	10	16	1.6
19）	4	5	1.25	56）		1	
20）		1		57）	18	3	0.166
21）	10	3	0.3	58）	3	2	0.67
22）		1		59）	17	59	3.47
23）	2	4	2.0	60）		1	
24）	4	6	1.5	61）	14	4	0.29
25）	6	2	0.33	62）	6	5	0.83
26）		1		63）	17	9	0.53
27）	2	2	1.0	64）		1	
28）	3	4	1.33	65）	12	10	0.83
29）	10	5	0.5	66）		1	
30）	10	6	0.6	67）	14	12	0.86
31）		1		68）		1	
32）	1	2	2.0	69）		1	
33）	3	3	1.0	70）		1	
34）	6	5	0.83	71）	9	9	1.0
35）	2	2	1.0	72）	13	22	1.69
36）	2	2	1.0	73）	11	10	0.91
37）		1		74）		1	

表 2－1－2

	期間	本数			期間	本数	
1）		1		32）	10	18	1.80
2）	3	2	0.67	33）	11	12	1.09
3）	8	6	0.75	34）	7	6	0.86
4）		1		35）	13	26	2.0
5）		1		36）	6	3	0.5
6）	10	4	0.4	37）	15	5	0.33
7）	5	3	0.6	38）	17	3	0.18
8）		1		39）	16	12	0.75
9）	3	3	1.0	40）		1	
10）		1		41）	3	6	2.0
11）		1		42）		1	
12）	14	5	0.36	43）		1	
13）	3	2	0.67	44）	4	2	0.5
14）		1		45）	15	4	0.27
15）		1		46）		1	
16）		1		47）		1	
17）	17	17	1.0	48）		1	
18）		1		49）	7	5	0.71
19）	3	2	0.67	50）	10	6	0.6
20）	13	6	0.46	51）	19	20	1.05
21）		1		52）	16	13	0.81
22）	2	2	1.0	53）	3	2	0.67
23）	6	6	1.0	54）	18	11	0.61
24）	10	8	0.8	55）	7	2	0.29
25）	16	25	1.56	56）		1	
26）		1		57）		1	
27）	7	3	0.43	58）	1	2	2.0
28）	2	3	1.5	59）	1	2	2.0
29）	20	65	3.25	60）	25	27	1.08
30）		1		61）	9	9	1.0
31）		1		62）	3	2	0.67

表 2－1 の説明：

1965 年から 1995 年までについての教員別の論文本数を第 3 列目に記してある。第 2 列目はある教員が報告した最初の論文刊行年と最後の刊行年との間かくである（単位は年)、例えば番号 1）について言えば 4：年数、3：本数となっている。この 4 は年間あたりの論文数を算出するさいにもちいる。第 4 列目の 0.75 は第 3 列/第 2 列である。(1 年間あたりの本数の近似値として 3/4 を引用している)。本数については資料の pp. 212–213 の数値を用いたが、間かくについては pp. 3–209 までを参照にした。標本サイズはそれぞれ 74、62、である。

表 2－2

論文数	p. 212	p. 213	
1	22	22	44
2	14	10	24
3	10	6	16
4	6	2	8
5	5	3	8
6	4	6	10
7	1	0	1
8	2	1	3
9	2	1	3
10	2	0	2
⋮			

注）7 本までの標本数：111、p. 212 とあるのは引用資料のページ番号である。

ところですぐ気づくように以上は観測値に欠落があるケースである。つまり、1 本も論文を書かない教員のリストはない。とりあえず論文数データが poisson 分布にしたがうとすると、論文数 $\leqslant 7$ までにおいて表 2－2 より

$$\text{標本平均} = \frac{279}{111} \div 2.51 = \tilde{\lambda}$$

であるが、これは論文数が 0 に対応する標本を含んでいない。もしそのケースの n_0 が存在すれば $\tilde{\lambda}$ は当然小さくなるはずである。つまり

$$\tilde{\lambda}(n_0) = \frac{279}{n_0 + 111}$$

とした場合の $\tilde{\lambda}(n_0)$ が poisson 分布の母数になる。ここで n_0 に適当な値を選ぶにはどうするか？ n_0 を指定する前に、$\tilde{\lambda} = 2.51$ の大きさについて考えると $\tilde{\lambda}(n_0)$ は $\tilde{\lambda}(n_0) < 2$ であるはずである。なぜなら poisson 変数が 1、2 で n_j は表 2－2 より 44、24 であるから $\tilde{\lambda}(\ ) > 2$ とすると変数 2 に対応する理論値が 44 を上まわってしまう。こうして $\tilde{\lambda}(n_0)$ は

$$\tilde{\lambda}(n_0) = \frac{279}{n_0 + 111} < 2$$

でなければならない。これを解くと

$$279 < 2n_0 + 222$$
$$28.5 < n_0$$

そこで $n_0 = 29$ とすると

$$\tilde{\lambda}(\ ) = \frac{279}{29+111} = 1.9928$$

となるが直観的には 1.9928 は大き過ぎる。n_0 を大きく指定すれば $\tilde{\lambda}$ は小さくなるのでとりあえず $n_0 = 40$ で計算値と poisson からの理論値を作成すると表 2−3 のようになる。

表 2−3

		データからの計算値	p_j
j	n_j	$n_j/(n_0+111)$	$e^{-\lambda}\lambda^j(j)^{-1}$
0	40(＝ n_0)	0.2649	0.1576
1	44	0.2914	0.2912
2	24	0.1589	0.2690
3	16	0.1060	0.1657
4	8	0.0530	0.0765
5	8	0.0530	0.0283
6	10	0.0662	0.0087
7	1	0.0066	0.0023

注）$\lambda = 279/(40+111) = 1.8477$、$n_0 = 40$、ポアソン変数の分布は $\{\exp(-\lambda)\}\lambda^j(j!)^{-1}$、$j \geqq 0$ である。

表 2−3 からすぐ気づくようにくいちがいの程度は少し大きい。しかし $j = 1$，2 で計算値、p_j（理論値）とも大きさについての順序は保たれている。さらに $j = 0$，1 で n_j が同一と仮定した場合の計算値、p_j（理論値）は表 2−4 のようになる。

表 2－4

		データからの計算値	p_j
j	n_j	$n_j/(n_0+111)$	$e^{-\lambda}\lambda^j(j!)^{-1}$
0	44（＝ n_0）	0.2839	0.1653
1	44	0.2839	0.2975
2	24	0.1548	0.2678
3	16	0.1032	0.1607
4	8	0.0516	0.0723
5	8	0.0516	0.0260
6	10	0.0645	0.0078
7	1	0.0065	0.0008

注）$\lambda = 279/(44+111) = 1.80$，$n_0 = n_1 = 44$。

この表 2－4 も表 2－3 とあまりちがいはない。要約すると表 2－2 には論文数が 0 本の標本は除かれているが観測データ全体に poisson 分布をあてはめてみると n_0（0 本に対応する標本数）は 40～50 程度と推定され、λ（ポアソンの母数）はおよそ 1.8 程度になっている。つまり表 2－2 の標本数は 111 であるが、論文を 1 本も報告しない教員を含めると、1965 年～1995 年で執筆者に入れかわりはあっても平均の執筆本数は 1.8 程度ということである。注意をしておくが、とり上げている議論には論文本数、教員の入れかわり、欠落している教員数、これらすべてが含まれている（poisson の λ についてその推定値は標本平均と標本分散がある。後者については λ とのくいちがいの程度は大きい。理由はデータのうち、0 から遠く離れた値を除いたからである）。

すでに述べているように以上の解析をそのまま進めるのはあまり意味がない。そこで 1 年間あたりの論文数を見ると表 2－5 のようになる。注意したいのは 1965～1995 年で 1 本のみの場合は考慮することはできないという点である。一応、報告回数が 2 回以上の標本をとり出した。例えば 1965 年～1970 年で報告が 2 回の場合、問題の数値は 2/6 ÷ 0.33 である。0.33 は年間あたり論文数の近似値である。つまりかりに当該の教員が 1965 年～1975 年に在籍したとしてもこの場合 0.33 とカウントするということである。そうするとあつかう対象は 1965 年～1995 年に在籍し、かつ論文を 2 回以上報告した標本群ということになる。こうした工夫で論文数データへの poisson 分布のあてはまりの良さはかなり改善されるだろう。

くり返すと表 2－5 は 2 回以上の報告があった教員について資料 pp. 212–213 の論文数データを整理したものである。n（標本サイズ）は 88 にまでおちるが、n_j の分布の散らばりの形を見るとこのケースももちろん poisson 分布に近いのがわかる。

表 2－5

j	1 年あたり論文数	n_j
0	～0.5	23
1	～1.0	44
2	～1.5	11
3	～2.0	7
4	2.0 をこえる	3
		88

注）1965 年から 1995 年までに論文数が 2 本以上あった教員の度数分布である。ただし特定の 1 年間に 2 本を報告したが、これ以外に記録のない標本は除いた。
資料の pp. 212–213 までを取り上げ、論文がのった年の最初と最後を資料の前半のページから調査した。

このケースの $\hat{\lambda}$ を計算すると $\hat{\lambda} = 99/88 = 1.125$ である。そうしてデータからの計算値と p_j（理論値）を示せば以下のようになる。

表 2－6

j	(1)　n_j/n	(2)　p_j	(1)－(2)
0	0.2614（＝ 23/88）	0.3247	－0.0633
1	0.5000	0.3652	0.1348
2	0.1250	0.2054	－0.0804
3	0.0795	0.0770	0.0025
4	0.0341	0.0217	0.0124

注）$n = 88$

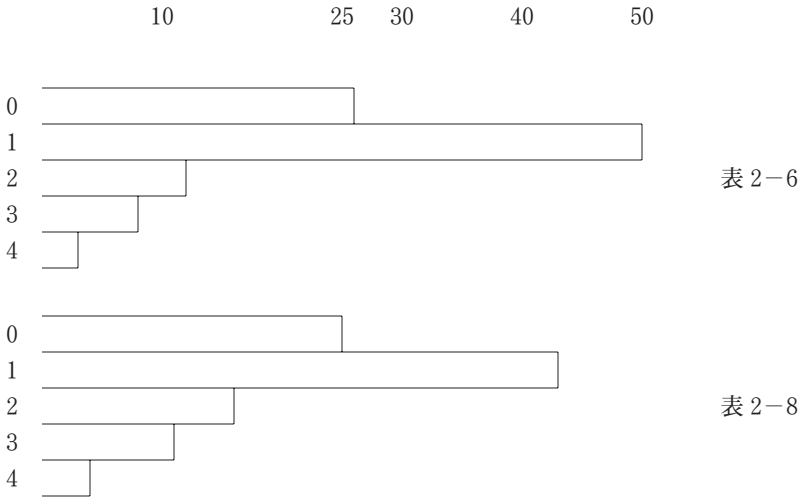
ここであてはまりの程度を見るために $v = \sum_{j=0}^{4}(n_j - np_j)^2/(np_j)$ の値を計算すると $v = 8.8644$、他方自由度 3 のカイ 2 乗分布の上例（右側）5%点は 7.81 だから「H_0：データは poisson 分布にしたがう」という仮説はこの時点では一応すてられる。つまりこれはあてはまりがさほどよくないことを言っている。理由の 1 つは 1965 年～1995 年のあいだで論文本数を 2 としているため contaminated data が含まれるからである。ただし 8.8644 と 7.81 では、あまりちがいはない。以下で示すが 1965 年～1995 年での論文数を 3 以上とするとあてはまりの程度は大きく改良される（表 2－7）。

表 2－7

j	1 年あたり論文数	n_j
0	～0.5	17
1	～1.0	30
2	～1.5	11
3	～2.0	7
4	2.0 をこえる	3
		68

注）n（標本サイズ）は 68 である。ここでは 1965 年～1995 年のあいだに論文を 3 回以上執筆した教員を対象とする。

図 2－1



注）図 2－1 は表 2－6、表 2－8 の相対度数分布をヒストグラムにしたものである。

表 2－8 から poisson 変数の平均（$\hat{\lambda}$）は $\hat{\lambda} = 1.250$ である。ここで p_j（理論値）、データからの計算値を以下に示す。

表 2－8

j	(1)　n_j/n	(2)　p_j	(1)－(2)
0	0.2500	0.2865	－0.0365
1	0.4412	0.3581	0.0831
2	0.1618	0.2238	－0.0620
3	0.1029	0.0933	0.0096
4	0.0441	0.0291	0.0150

かい離の程度を表 2－6 の結果と比較すると、いく分小さくなっている。論文本数を 3 以上に制限したことで contaminated data が少なくなり、データ計算値への理論値のあてはまりがよくなったと考えられる。以上から H_0：poisson 分布である、として検定統計量を見ると

$$v = \sum_{j=0}^{4}(n_j - np_j)^2/(np_j)$$
$$= 3.3877$$

となる。ここで $n_0 = 17$、$np_0 = 19.482$、$n = 68$ などである。v は H_0 のもとで自由度 3 のカイ 2 乗分布にしたがい、この分布の上例 5%点は 7.81 だから、$v = 3.3877$ では H_0 をすてることはできない。したがって以上の内容は表 2－8 のデータが poisson 分布にしたがっていることをつよく示唆している。

くり返えすがあてはまりの程度がよくなった理由の 1 つは報告回数を制限したことによって標本内の同質性が高まった点にある。つまり 1965 年～1995 年間の回数が 1～2 回では論文執筆能力を判定するための情報が不足しており、短期間にまれに報告をする教員も含まれてしまう。そこでこうしたケースを除いて普通の意味で執筆をしている教員（標本）について調査をすると、論文執筆数の分布には poisson 分布がよくあてはまり、年間平均 1 本までの累積確率がデータではほぼ 0.70 であり、ポアソンの理論値も累積で 0.64 までに届くということである。

3　論文数の T スコア

言う必要はあまりないが、元のデータに余分の情報がない表 2－5、表 2－6 についていわゆる偏差値との関連を示す。標本平均は $\bar{x} = 0.8125$ であるがこの場合偏差値 50 を下回る標本数は 67 であり、全体に占める割合は 67/88 ≒ 0.7614 ＞ 0.5 となる。元の分布が正規であればこの割合はほぼ 0.5 であるから、これはここで示した論文数データが偏差値計算についていかに場ちがいであることを言っている。理由は元のデータの分布の歪みにあり、そうした場合偏差値計算はその意味を完全に失うのである（歪んだ分布では中央部分（平均値、あるいは中央値）を定義することができない）、表 3－1、

表3－2を見るとよい。ついでに言っておくとこの偏差値計算が可能であるには

表3－1

j	n_j	v_j	z_j	$vol(v < v_j)$	$Pr(z < z_j),\ z \sim N(0,\ 1)$
0	23	0.25	－1.1275		
		0.5	－0.6264	0.2614	0.264
1	44	0.75	－0.1253		
		0.8125	0.0	0.580	0.50
		1.0	0.3758	0.7614	0.649
2	11	1.25	0.8769		
		1.5	1.3780	0.8864	0.916
3	7	1.75	1.8791		
		2.0	2.3802	0.9659	0.991
4	3	2.25	2.8813		

注）表2－6の代表値v_jを正規化したz_jを第4列目に示す。ここで$z_j = (v_j - \bar{x})$/s.d.、s.d. ＝ 標準偏差 ＝ 0.4989である。$\bar{x} \fallingdotseq 0.8125$を下回る標本数は67となる。$j = 1$の区間は0.5～1.0だからいく分詳しく計算すると（0.8125－0.5）: 0.5 ＝ n': 44から$n' = 27.5$、このn'が0.5～0.8125のあいだにおちる標本数になる。そうするとこの場合51が$\bar{x} \fallingdotseq 0.8125$を下回る標本数になる。その割合は$51/n = 0.580 > 0.5$である。0.580も0.5をはるかに上回る。

第5、6列目にv_j、z_jに対応する累積確率を示した。例えば$\bar{x} = 0.8125 = v_j$のとき、z_jは0であり、$vol(v < v_j) = 0.580$、ここでもしzが$N(0,\ 1)$にしたがえば、$Pr(z < 0) = 0.5$だから問題にしている分布がいかに歪んでいるかがわかる。

表3－2

j	n_j	v_j	z_j	$vol(v < v_j)$	$Pr(z < z_j),\ z \sim N(0,\ 1)$
0	17	0.25	－1.1615		
		0.5	－0.6969	0.2500	0.242
1	30	0.75	－0.2323		
		0.875	0.0	0.588	0.50
		1.0	0.2323	0.6912	0.591
2	11	1.25	0.6969		
		1.5	1.1615	0.8529	0.877
3	7	1.75	1.6261		
		2.0	2.0907	0.9559	0.982
4	3	2.25	2.5553		

注）$\bar{x} = 0.8175$，s.d. ＝ 0.5381である。$\bar{x}$を下回る標本数は比例配分による計算では40となる。その割合は$40/68 \fallingdotseq 0.588 > 0.5$でありこの0.588も0.5をはるかにこえる。

1．とり上げる複数の集団の標本数がそれぞれかなり大きい。つまり $n \geqslant 200$ 程度。
2．元のデータの分布に極端な歪みがない。
3．異なる集団間においても潜在的能力のちがいにあまり差はない。

などの要請がある。3 から順にコメントをすると偏差値計算の適用範囲はほとんどが受験生間の学力テストに関するものであり、異なる集団でテスト結果にちがいが生じたのは問題の難易度のみによるという考え方である。芝－南風原［4，pp. 37-38］を見るとよい。

2 については最近の新司法試験に関する法務省文書が参考になる。この試験の論述問題については受験者数の大標本性により、複数の採点委員が必要であるが、その採点委員に対して以下のような要請がある。つまり 100 点満点の問題に対して素点の分布を

答案枚数の 5%：100～75 点
25%：74～58 点
40%：57～42 点
30%：41～ 0 点

とするのである。これは素点の分布が 50 のまわりで左右対称をしており、かつ 50 の附近でボリュームが大きくなっている。この要請は素点が正規分布になるようにせよ、というものではないが、上の割合を正規分布にあてはめると、平均 50、s.d. が 15.197～15.238 になる。こうした要請があるのは素点の分布が近似的にも正規（正規分布は左右対称）でないと、偏差値に変換した得点が意味をもたないからである。素点の分布の歪みについて注意をうながす文献には以下がある。

- 偏差値はデータの集まりが大量であり、正規分布することを仮定したものである（佐々木［2，p. 28］）。
- 実際には「理論的に正規分布」の仮定が成り立たないときに誤差が大（住田［5，p. 70］）。
- 分布が正規分布からかけ離れ、大きく歪んでいるときには、上に述べた線形変換による標準得点から得点分布上の位置を正しく知ることはできない（芝－渡部［3，p. 29］）

次に 1 の集団の標本サイズについても法務省文書が参考になる。つまり 2009 年の新司法試験の合格者数目安は 2,500～2,900 であり、例として労働法（選択）の科目は 3,077 名、全体の 31%が受験し、他方労働法の考査委員数は 5 名、したがって合格者について委員 1 人あたり答案枚数は

合格者（労働法専攻）の枚数/5

＝（2500～2900）×0.31×0.2

≒（155～180）

となる。そうすると倍率を2倍としても考査委員に割りあてられる答案枚数は310～360枚はあるということになる（ただし考査委員は短答式で下限のスコアに到達しない答案については採点をしない）。つまりこれは偏差値計算では各集団の標本数が200程度ないと話にならない点を意味している。

また、もとの理論は $x_i \sim N(\mu, \sigma^2)$ のとき、$(x_i-\mu)/\sigma \sim N(0, 1)$ だから、これに対応して $(x_i-\bar{x})/s$ を作るとすると標本数が大きいときにのみ $(x_i-\mu)/\sigma \fallingdotseq (x_i-\bar{x})/s$ である（s^2, σ^2 はそれぞれ標本分散、母分散を表す）。

サイズに関する文献には以下のようなものがある。

- たとえば1つの学級の得点分布を正規化したりすることの意味はほとんどない（芝－南風原［4, p. 38］）。
- 偏差値はデータの集まりが大量であり、正規分布することを仮定したものである（佐々木［2, p. 28］）。

最後に集団の標本サイズ、左にピークのある歪みのある分布、相対的な評価（とくに相対的に下位の被評価者にペナルティを課すケース）に関して1つの例を上げよう。

いま集団 $t = 1, \cdots\cdots, T$ について標本数を一般に n_t と書く。また素点を集団 t の個人についてすべて x_{0t} としよう。つづいて

1．集団 t について m_t 名の素点が $x_{1t} = x_{0t}+\varepsilon_{0t}$, $\varepsilon_{0t} > 0$ に移り、他は x_{0t} を維持とする。

2．そうするとこの集団の平均は $x_{0t}+\varepsilon_{0t}(m_t/n_t)$、他の n_t-m_t 名の偏差値はすべて50をきる。

3．つづいて基準化をするために s_t^2（標本分散）を計算すると

$$s_t^2 = (1/n_t)\left\{\left(\frac{m_t}{n_t}\varepsilon_{0t}\right)^2(n_t-m_t)+\varepsilon_{0t}^2(1-m_t/n_t)^2 m_t\right\}$$
$$= \frac{\varepsilon_{0t}^2}{n_t^2}(n_t-m_t)m_t$$

となる。

4．そうすると下位の n_t-m_t 名について基準化された値、$z(n_t-m_t)$ は

$$
\begin{aligned}
*)\quad z(n_t-m_t) &= -m_t(n_t-m_t)^{-1/2}m_t^{-1/2} \\
&= -m_t^{-1/2}(n_t-m_t)^{-1/2} \\
&= -\left(\frac{\hat{p}_t}{1-\hat{p}_t}\right)^{1/2} \\
&< 0
\end{aligned}
$$

となる。ただし $\hat{p}_t = m_t/n_t$ である。他方、上位の $z(m_t)$ については

$$
\begin{aligned}
z(m_t) &= \varepsilon_{0t}(1-n_t/m_t)\cdot(\varepsilon_{0t}^2 n_t^{-2}(n_t-m_t)m_t)^{-1/2} \\
&= (1-\hat{p}_t)^{1/2}\,(\hat{p}_t)^{-1/2} \\
&> 0
\end{aligned}
$$

となる。確認のため $z(m_t)$、$z(n_t-m_t)$ の和を作ると

$$
\begin{aligned}
&m_t z(m_t)+(n_t-m_t)\,z(n_t-m_t) \\
&= m_t\cdot\frac{\sqrt{1-\hat{p}_t}}{\sqrt{\hat{p}_t}}-(n_t-m_t)\cdot\frac{\sqrt{\hat{p}_t}}{\sqrt{1-\hat{p}_t}} \\
&= 0
\end{aligned}
$$

となる。

以上から次の点が読みとれる。

1）$z(m_t)$、$z(n_t-m_t)$ ともこの基準化した値（偏差値 $=10\cdot z(m_t)+50$、あるいは $10\cdot z(n_t-m_t)+50$）は x_{0t}、$\varepsilon_{0t}(>0)$ にはまったく依存しない。

2）くり返すがこの場合の偏差値は集団 t において、素点が大きい m_t 名の n_t に関する割合 m_t/n_t に依存するだけである。

3）$m_t/n_t = \hat{p}_t(0<\hat{p}_t<1)$ が大きくなるだけで上位の m_t 名の基準化した値、$z(m_t)$ は下がる。またとくに強調したいのは $\hat{p}_t\uparrow 1$ で下位の $z(n_t-m_t)$ はマイナスの方向へ大きくなる。

4）ここで $\hat{p}_t = m_t/n_t$ だからもし $t=1,\ 2$ で $m_1=m_2=m=$ 一定とすると $n_t\downarrow$ は $\hat{p}_t\uparrow$ を意味す

るから集団 $t = 1,\ 2$ で $n_1 < n_2$ であれば上の＊）によって集団内の素点 $x_{0t}+\varepsilon_{0t}$、x_{0t} に関係なく集団 1 の下位 m_1 名の偏差値は集団 2 の下位 m_2 名の偏差値よりもつねに小さくなる（$z(n_1-m_1) < z(n_2-m_2)$ である）。つまりペナルティは小集団のものがまっ先に受ける。

5）以上の内容は集団の性質を

$$Pr(x_{1t} = x_{0t}+\varepsilon_{0t}) = p_t,\ \ 0 < p_t < 1,\ \ \varepsilon_{0t} > 0$$
$$Pr(x_{1t} = x_{0t}) = 1-p_t = q_t$$

とし、分布のピークが左によっているとすればさらに $p_t < 0.5$ を仮定することに等しい。そうして m_t/n_t が p_t の推定値となっている。

注

1）本論文は経済学部内研究会報告（2009 年 5 月 13 日）を整理したものである。研究会参加者の方々からは数多くの有益なコメントをいただいた。また、レフェリーの 2 人からもお教えを受けた。ここにお礼申し上げる。

2）以下のように記号を定め contaminated data についてコメントを加える。

$A^{(0)}$：観察期間に在籍したが論文本数が 0 の教員群

$A^{(1)}$：期間内に在籍し 1 本以上の論文がある教員群

$A^{(2)}$：期間内に 1 本以上の論文がある学外の研究者、あるいは在籍した大学院生。

ここで簡単化のために $A^{(2)}$ から $A^{(0)}$、$A^{(1)}$ への移動はないものとしよう。さらに $a_i^{(j)} \in A^{(j)} (j = 1,\ 2)$ となる個人 $a_i^{(j)}$ の時点 t の論文数を $a_{it}^{(j)}$ とすると $a_{it}^{(j)} = 0,\ 1,\ 2,\ \cdots\cdots$ である。他方 $a_i^{(0)} \in A^{(0)}$ についてはすべての t について $a_{it}^{(0)} = 0$ であり、$a_{it}^{(0)}$ は本論文に引用する資料には表示されない。$a_{it}^{(j)},\ j = 1,\ 2$ の期間（近似したもの）を $T_i^{(j)}$ として $\sum a_{it}^{(j)}/T_i^{(j)}$ をプールし変数変換したものが poisson と予想したが（本稿後半）、$T_i^{(j)}$ が小さい $a_{it}^{(j)}$ は observe された回数が少ないことを意味し、これらを考察対象から除くのがよい。より詳しく言うと次のようになる。

1．本来の考察対象は $A^{(0)}$、$A^{(1)}$ の群であるが、$A^{(0)}$ のデータは入手不能（missing）、$A^{(2)}$ に属するデータが引用の資料に混入している。

2．資料の全体はいわゆるパネルデータからなるが、observe された期間が短いデータは信ぴょう性に乏しい。例えばこれは $t = t^*$ で $a_{it^*}^{(j)} \geqslant 1 (j = 1,\ 2)$ のみがあり、$a_{it}^{(j)} = 0,\ t \neq t^*,$ となるような $a_{it}^{(j)}$ を言う。

3．上記 2 で t の 2 カ所のみで $a_{it}^{(j)} = 1$、それ以外では $a_{it}^{(j)} = 0$ のデータも本文後半では除いてモデルのあてはめをした。もちろん $T_i^{(j)}$ の大きさで除く考え方もある。

4．上記、2、3 のような操作をすると $A^{(2)}$ に属する contaminated data を部分的に排除することができる。他方、$A^{(1)}$ に入る正常な $a_{it}^{(1)}$ を誤って除いているかも知れない。

5．在籍期間を示すデータがないので（本稿では近似値を使う）、$a_{it}^{(j)} = 0,\ j = 1,\ 2$ とあったとしても確

かに observe されて $a_{it}^{(j)} = 0$ か、もともと対応するデータがなくて $a_{it}^{(j)} = 0$ かは判別がつかない場合もある（上記 2 に関連、また Cohen, A.C. Jr［8］）。

参考文献

［1］ 井上勝雄『新・よくわかる統計学の考え方』ミネルヴァ書房、2008 年。

［2］ 佐々木正文『社会科学系学生のための統計学』共立出版、1999 年。

［3］ 芝祐順 - 渡部洋『入試データの解析』新曜社、1988 年。

［4］ 芝祐順 - 南風原朝和『行動科学における統計解析法』東京大学出版会、1992 年。

［5］ 住田幸次郎『初歩の心理教育統計法』ナカニシヤ出版、1988 年。

［6］ Fisz, M.: *Probability Theory and Mathematical Statistics,* John Wiley & Sons, New York, 1962.

［7］ Johnson, N.L., S. Kotz, and A.W. Kemp: *Univariate Discrete Distributions,* Second Edition, John Wiley & Sons, New York, 1992.

［8］ Cohen, A. Clifford Jr: Estimation in the Truncated Poisson Distribution When Zeros and Some Ones Are Missing, Journal of the American Statistical Association 55, 1960, pp. 342–348.

資料

1．京都産業大学、京都産業大学学術誌論文名一覧、1965–1995、1995 年。

2．法務省、平成 21 年新司法試験の実施について（新司法試験における採点及び成績評価等の実施方法・基準について）、www.moj.go.jp/SHIKEN/index.html

Poisson Approximations for the Distribution of the Number of Journal Articles

Yusaku KATAOKA

Abstract

This paper investigates Poisson approximations for the number of journal articles submitted during period of one year. Let x be Poisson random variables. Then we observe the following frequencies of appearance of values of j.

Table

j	0	1	2	3	4
Frequency	0.250	0.441	0.162	0.103	0.044
Probability	0.287	0.358	0.224	0.093	0.029

In the central row of this table we compute the sample mean $\bar{x} = 1.250$. Moreover let us compute the corresponding probabilities $Pr(x = j)$ for the Poisson distribution with Poisson parameter $= 1.250$. Observing these probabilities from Poisson distribution table, we have $Pr(x = 0) = 0.287$, $Pr(x = 1) = 0.358$, $Pr(x = 2) = 0.224$, $Pr(x = 3) = 0.093$, $Pr(x = 4) = 0.029$. These values are presented in the lower row of Table. As we see, these probabilities differ but little from the corresponding frequencies.

We second turn to the T-score related to the number $j = 0, 1, \ldots$ of journal articles submitted. The method of T-score is essentially one of transforming the normal variables into deviates of $N(50, 100)$. As has already been stated, the number j are well approximated by Poisson random variables. Thus it is apparently irrelevant to consider the T-score to the number of journal articles submitted.

Keywords : normal distribution, poisson approximation, T-score, z-score, contaminated data