Bulletin of Faculty of Liberal Arts, Nagasaki University
Vol. 3

病気欠勤の統計資料の分析

藤　沢　秀　雄

1. 序　　論

統計資料が離散的な値をとる場合，これらの資料の頻度分布にポアソン分布

$$e^{-\lambda}\lambda^{x}/x! \qquad (x=0,1,2,\cdots;\ \lambda>0)$$

がうまくあてはまることが多い。

もしポアソン分布がうまくあてはまらないようなときには，ポアソン分布の特性値 λ それ自身がピアソンの第III型の分布

$$\alpha^{\rho}\lambda^{\rho-1}e^{-\alpha x}\Gamma(\rho) \qquad (\lambda>0;\ \alpha,\rho>0)$$

に従うという仮定を導入することによって，うまくあてはめることができることがある。

この方法によって得られた分布は負の二項分布と呼ばれ，事故や病気等に関する統計資料にうまく適合することがすでに知られている。事故件数の資料の場合においては，このような確率変数 λ の導入は，その集団の団員の各自がもっている事故発生の危険性が一様でないということを表現する方法として解釈される。しかしながらこのような統計資料の分析については，十分な研究がなされていない。そこでまずその実際の資料に基づいてその分析方法を考えてみた。

2. 実例とその分析の数学的形式化

某事業所において，或る3つの集団について或る一定期間内における各人の病気欠勤の回数を調べ，その資料を基にしてこれら3つの集団の健康状態が同じであるか否かを判定しようとする試みがなされた。

もしこの分布がポアソン分布

$$P\{X=x\}=e^{-\lambda}\lambda^{x}/x! \qquad (x=0,1,\cdots) \tag{1}$$

に従うものであるならば，欠勤回数の分布は λ によって決定されるから，各集団の病気欠勤に関する健康状態は λ の値によって表わされるとみなされる。この場合，平均欠勤回数 $\bar{x}=\sum x_i/n$ が λ の不偏有効推定量であるから，各集団の平均欠勤回数を計

表 1. 病気欠勤の回数の分布

回 数	I 群	II 群	III 群	計
0	328	324	321	973
1	47	41	54	142
2	20	15	11	46
3	12	8	11	31
4	6	4	3	13
5	0	3	1	4
6	1	3	3	7
7	0	1	0	1
8	1	0	0	1
9	0	1	1	2
10	0	0	1	1
計	415	400	406	1221

算することによって，各集団の病気欠勤に関する健康状態を知ることができる。そこでまず各集団について，欠勤回数の分布が実際にポアソン分布に従うか否かの適合度の検定を行なってみた。その結果各群ともポアソン分布からの有意なへだたりがみられた。

表 2. ポアソン分布への適合度検定

回 数 x	I 群		II 群		III 群	
	観測数	期 待 値	観測数	期 待 値	観測数	期 待 値
0	328	281.56	324	268.13	321	271.75
1	47	109.23	41	107.25	54	109.10
2	20	21.20	15	21.45	11	21.90
≧3	20	3.01	20	3.17	20	3.25
計	415		400		406	

$\bar{x}$	.3880	.4000	.4015
$\chi^2_{(2)}$	139.08	148.52	128.51
P	$<.001$	$<.001$	$<.001$

このように各群とも欠勤回数の分布がポアソン分布に従わないことが分ったので，この分布について次のような仮定を設けた。

各人の欠勤回数 x はポアソン分布

$$p(x;\lambda)=e^{-\lambda}\lambda^{x}/x! \qquad (\lambda>0;\ x=0,1,\cdots) \tag{2}$$

に従うものとする。しかしながらλは各人に固有な値であって（勿論期間の長さには関係する），すべての人について決して同一ではない。すなわち各集団とも λ に関してそれぞれ独自の分布を持っていると考えるわけであるが，それらはすべてピアソンの第III型

$$dF(\lambda)=\frac{\alpha^{-\rho}}{\Gamma(\rho)}\lambda^{\rho-1}e^{-\lambda/\alpha}d\lambda \qquad (\alpha,\rho>0) \tag{3}$$

の分布で表わされるものと仮定する。

この型の分布は ρ と α によって定まるから，このような仮定は各集団の病気欠勤に関する健康状態が ρ と α によって特徴づけられると解釈することを意味する。

ρ と α は調査の期間によって変るけれども，もし各人の λ が期間の長さに一様に比例するならば，ρ の値は期間の長さに無関係で，α のみが期間の長さに比例する。

もし λ の分布について上の仮定がなりたつならば，一集団の中から x 回の欠勤をした人が見出される確率は

$$\begin{aligned}P\{X=x\}=f(x;\rho,\alpha)&=\int_0^\infty p(x;\lambda)dF(\lambda)\\&=\left(\frac{1}{1+\alpha}\right)^{\rho}\frac{\Gamma(\rho+x)}{\Gamma(\rho)\Gamma(x+1)}\left(\frac{\alpha}{1+\alpha}\right)^{x}\end{aligned} \tag{4}$$

によって与えられる。これは，

$$f(x;\rho,\alpha)=\left(\frac{1}{1+\alpha}\right)^{\rho}\binom{-\rho}{x}\left(-\frac{\alpha}{1+\alpha}\right)^{x} \tag{5}$$

とも変形できるので，負の二項分布と呼ばれ，特に $\rho=1$ のときには幾何分布と呼ばれる。この分布の平均値 $E(X)$ および平均値の周りの k 次の積率 μ_k は

$$E(X)=\rho\alpha \tag{6}$$

$$\mu_2=V(X)=\rho\alpha(1+\alpha) \tag{7}$$

$$\mu_3=\rho\alpha(1+\alpha)(1+2\alpha) \tag{8}$$

$$\mu_4=\rho\alpha(1+\alpha)\{1+3\alpha(1+\alpha)(2+\rho)\} \tag{9}$$

であり，また歪度 β_1 および尖度 β_2 は

$$\beta_1=\mu_3^2/\mu_2^3=\left(4+\frac{1}{\alpha(1+\alpha)}\right)\frac{1}{\rho} \tag{10}$$

$$\beta_2=\mu_4/\mu_2^2=3+\frac{1}{\rho}\left(6+\frac{1}{\alpha(1+\alpha)}\right) \tag{11}$$

である。このことから ρ の値が十分大きければ，この分布は正規分布に近くなることが分る。一方 α が十分小さければ分散 $V(X)$ は平均値 $E(X)$ に近くなるから，このときにはポアソン分布に近いことが分る。

3. 負の二項分布の特性値 ρ, α の推定量の求め方

（i） ρ が既知のとき，$\alpha^*=\bar{x}/\rho$ は α の不偏かつ有効推定量である。

（証） X の特性函数は $(1+\alpha-\alpha e^{it})^{-\rho}$ なる故，$\alpha^*=\bar{x}/\rho$ の確率分布は

$$g(\alpha^*;\alpha)=\left(\frac{1}{1+\alpha}\right)^{n\rho}\frac{\Gamma(n\rho+n\rho\alpha^*)}{\Gamma(n\rho)\Gamma(n\rho\alpha^*+1)}\left(\frac{\alpha}{1+\alpha}\right)^{n\rho\alpha^*} \tag{12}$$

であることが容易に見出される。これから

$$E(\alpha^*)=\alpha\,,\qquad V(\alpha^*)=\frac{\alpha(1+\alpha)}{n\rho} \tag{13}$$

が得られる。また

$$E\left[\left(\frac{\partial\log f}{\partial\alpha}\right)^2\right]=E\left[\left(\frac{X-\rho\alpha}{\alpha(1+\alpha)}\right)^2\right]=\left(\frac{1}{\alpha(1+\alpha)}\right)^2V(X)=\frac{\rho}{\alpha(1+\alpha)}$$

であるから

$$V(\alpha^*)=1/nE\left[\left(\frac{\partial\log f}{\partial\alpha}\right)^2\right]. \tag{14}$$

故に α^* は α の不偏有効推定量である。

（ii） ρ, α が共に未知である場合，まず最尤法によってこれらの推定量を求めてみる。

ρ，α の尤度函数 $L(x_1,\cdots,x_n;\rho,\alpha)$ は

$$L=\left(\frac{1}{1+\alpha}\right)^{n\rho}\left(\frac{\alpha}{1+\alpha}\right)^{\Sigma x_i}[(\Gamma(\rho)]^{-n}\prod_{i=1}^{n}\frac{\Gamma(\rho+x_i)}{x_i!} \tag{15}$$

であるから

$$\log L=-n\rho\log(1+\alpha)+n\bar{x}\log\left(\frac{\alpha}{1+\alpha}\right)-n\log\Gamma(\rho)+\sum\log\Gamma(\rho+x_i)$$

$$-\sum\log x_i!$$

$$\frac{\partial\log L}{\partial\alpha}=\frac{n}{\alpha(1+\alpha)}(\bar{x}-\rho\alpha)$$

$$\frac{\partial \log L}{\partial \alpha}=-n\log(1+\alpha)+\sum_{x_i\neq 0}\left\{\frac{1}{\rho+x_i-1}+\cdots+\frac{1}{\rho+1}+\frac{1}{\rho}\right\}$$

故に最尤推定量 $\hat{\rho}$, $\hat{\alpha}$ は

$$\bar{x}=\rho\alpha \tag{16}$$

$$n\log(1+\alpha)=\sum_{x_i\neq 0}\left\{\frac{1}{\rho+x_i-1}+\cdots+\frac{1}{\rho+1}+\frac{1}{\rho}\right\} \tag{17}$$

を解くことによって得られるが，この方程式の解を求めるのは困難である。

一方 $\bar{x}=\sum x_i/n$, および $u^2=\sum(x_i-\bar{x})^2/(n-1)$ の平均値及び分散は,

$$E(\bar{x})=\rho\alpha\,,\qquad V(\bar{x})=\frac{1}{n}\rho\alpha(1+\alpha)\,, \tag{18}$$

$$\left.\begin{aligned}&E(u^2)=\rho\alpha(1+\alpha)\,,\\&V(u^2)=\frac{1}{n}\rho\alpha(1+\alpha)+\frac{2}{n-1}\rho^2\alpha^2(1+\alpha)^2+\frac{6}{n}\rho\alpha^2(1+\alpha)^2\end{aligned}\right\} \tag{19}$$

であることから，$\bar{x}$, u^2 はそれぞれ $\rho\alpha$, $\rho\alpha(1+\alpha)$ の不偏一致推定量である。

従って ρ, α の最尤解 $\hat{\rho}$, $\hat{\alpha}$ の第１一近似として

$$\alpha^*=u^2/\bar{x}-1\,,\qquad \rho^*=\bar{x}/\alpha^* \tag{20}$$

を求め，以下逐次近似で $\hat{\rho}$, $\hat{\alpha}$ を求めればよい。

4.　負の二項分布のあてはめ

前節の方法に基づいて，表１に対する ρ, α の最尤解を求め，その結果を表３に掲げた。

表 3.　表 1 に対する ρ, α の最尤解

		I	II	III
標本平均	$\bar{x}$	.38795	.40000	.40148
不偏分散	u^2	.89019	1.21805	1.21866
第１近似	ρ^*	.29967	.19559	.19725
	α^*	1.29460	2.04512	2.03542
最 尤 解	$\hat{\rho}$	.25962	.18032	.22960
	$\hat{\alpha}$	1.49430	2.21828	1.74861

この表から第１近似値 ρ^*, α^* は最尤解にかなり近い値を示していることが分る。

表４は上の結果を用いて，負の二項分布への適合度の検定を行なったものである。

期待値 np_x $(x=0,1,\cdots)$ は

$$p_0=\left(\frac{1}{1+\alpha}\right)^{\rho},\qquad p_x=p_{x-1}A_x \qquad (x=1,2,\cdots) \tag{21}$$

但し
$$A_x=\left(\frac{\alpha}{1+\alpha}\right)(\rho+x-1)/x \qquad (x=1,2,\cdots) \tag{22}$$

によって計算すれば容易に求められる。

表 4. 負の二項分布への適合度検定

回数 x	I 群		II 群		III 群	
	観測数	期待値	観測数	期待値	観測数	期待値
0	328	327.33	324	323.98	321	321.89
1	47	50.91	41	40.26	54	47.02
2	20	19.21	15	16.38	11	18.39
3	12	8.67	8	8.21	11	8.69
4	6	4.23	4	4.50	3	4.47
≧5	2	4.65	8	6.66	6	5.54
計	415		400		406	
$\chi^2_{(3)}$	3.864		.460		5.144	
P	.280		.925		.170	

表 4 は欠勤回数の分布が負の二項分布にうまく適合していることを示している。

このことは病気欠勤に関して各人のもっている特性値（欠勤率）λ について，どの集団もピアソンの第 III 型の分布

$$dF(\lambda)=f(\lambda)d\lambda=\frac{\alpha^{-\rho}}{\Gamma(\rho)}\lambda^{\rho-1}e^{-\lambda/\alpha}d\lambda$$

を持っているという仮定が当を得たものであることを意味する。

$$f'(\lambda)=\frac{\alpha^{-(\rho+1)}}{\Gamma(\rho)}\lambda^{\rho-1}e^{-\lambda/\alpha}\left[\frac{\alpha(\rho-1)}{\lambda}-1\right] \tag{23}$$

であるから，$f(\lambda)$ は $\rho<1$ のときには $(0,\infty)$ において単調減少であり，$\rho\geqq1$ のときには $\alpha(\rho-1)$ で最大値を取る。そして ρ の値が大きくなるにつれて，最大値を取る点の位置は右へ移行する。このことから ρ の値が小さいということは，その集団には λ の小さい人（すなわち滅多に病気欠勤をしないような健康体の持主）が非常に多いことを意味していると解釈され，大抵の場合どの集団もこれに属していると考えられる。実際表 3 は表 1 における各群の ρ の推定値が共に 1 より小であることを示して

いる。また

$$\frac{\partial}{\partial\alpha}f(\lambda)=\frac{1}{\alpha^2}(\lambda-\rho\alpha)f(\lambda) \tag{24}$$

であるから，十分大きな λ においては，α が大きくなるにつれて $f(\lambda)$ が大きくなる。このことから α が大きいということは，λ が他の人に比べて極端に大きい人（ひんぱんに病気欠勤をするような人）が可成りいることを意味していると解釈される。

5.　負の二項分布の特性値 ρ，α の性質

λ の確率密度函数 $f(\lambda)$ における ρ，α の性質は勿論そのまま負の二項分布

$$p_x=f(x;\rho,\alpha)=\left(\frac{1}{1+\alpha}\right)^{\rho}\frac{\Gamma(\rho+x)}{\Gamma(\rho)\Gamma(x+1)}\left(\frac{\alpha}{1+\alpha}\right)^{x}\qquad(x=0,1,\cdots)$$

にもあらわれている。

任意の自然数 x について

$$p_{x+1}-p_x=\left(\frac{1}{1+\alpha}\right)\left[1-\frac{\alpha(\rho-1)}{1+x}\right]p_x \tag{25}$$

がなりたつから，$\rho<1$ のときには勿論，$\rho<1+\dfrac{1}{\alpha}$ のときには数列 $\{p_x\}_{x=0}^{\infty}$ は減少数列をなす。特に $\rho_0=\left(\dfrac{1}{1+\alpha}\right)^{\rho}$ は ρ が 0 に近づくにつれて 1 に近づく。従って ρ の値が小さい集団では無欠勤者の見出される確率 p_0 は $\sum_{x=1}^{\infty}p_x$ に比べて非常に大きくなる。もし逆に ρ が 1 より大きく α が十分大きくて $\dfrac{\rho\alpha}{1+\alpha}$ が 1 に近いとき，p_0 は小さくなり数列 $p_0, p_1, p_2, \cdots$ は減少度の小さいものになる。そして $\rho>1+\dfrac{1}{\alpha}$ のときには，数列 $\{p_x\}_{x=0}^{\infty}$ は $[\alpha(\rho-1)]-1$ または $[\alpha(\rho-1)]$ で最大値を取るようになる。ρ の値が十分大きいときには，負の二項分布は正規分布に近くなることは前にも述べた通りである。

表 5 は表 3 に与えられている ρ，α の各組について確率分布を求めてみたものである。

この表において分布 I と分布 II とを比較してみるとき，p_0 の値および p_4 以下のものの値は分布 II の方が大きいことに気が付く。p_0 の値は α の値が大きくなるにつれて小さくなるべきであるのに，この場合において逆に大きくなっているのは，ρ の値が少さくなっているということが大きく影響しているからである。ρ と α の積（即ち平均値）は両者とも殆んど差はないのであるから，結局 p_0 の値は α よりも ρ によって大きく変動することがこのことから察せられる。今 p_0 の ρ，α についての

表 5. 負の二項分布

	I	II	III
0	.788759	.809962	.792830
1	.122680	.100671	.115806
2	.046289	.040951	.045295
3	.020887	.020515	.021416
4	.010197	.011243	.011000
5	.005204	.006479	.005920
6	.002733	.003856	.003283
7	.001464	.002346	.001858
8	.000796	.001452	.001068
9	.000438	.000909	.000622
10	.000243	.000575	.000365
≧11	.000310	.001041	.000537
$\rho\alpha$	.38795	.40000	.40148
ρ	.25962	.18032	.22960
α	1.49430	2.21823	1.74861

偏微係数をそれぞれ求めてみると，

$$\frac{\partial p_0}{\partial \rho}=-p_0\log(1+\alpha)=-p_0\left[\sum_{k=1}^{\infty}\frac{1}{k}\left(\frac{\alpha}{1+\alpha}\right)^k\right] \tag{26}$$

$$\frac{\partial p_0}{\partial \alpha}=-p_0\left(\frac{\rho}{1+\alpha}\right) \tag{27}$$

であるから，$\rho<\alpha$ に対して

$$\frac{\partial p_0}{\partial \rho}<\frac{\partial p_0}{\partial \alpha}<0 \tag{28}$$

がなりたつ。

従って $\rho<1$, $\alpha>1$ であるような分布について考えるとき（通常の集団はこれに属すると考えられる），p_0 の値は ρ の微小変化にも大きく変動し，その際の α の多少の変動は問題にならない。

一方十分大きな x においては，α の値が大きくなるにつれて p_x が大きくなることは，p_x の α についての偏微係数

$$\frac{\partial}{\partial \alpha}p_x=\frac{1}{\alpha(1+\alpha)}[x-\rho\alpha]p_x \tag{29}$$

からも容易に分る。これはαが大きくなるにつれて欠勤回数の多い人が見出される確

率が大きくなることを意味している。

しかしながら ρ の値が小さくなっていてもやはり p_4 以下のものが大きいということは，x の値が十分大きいような x の領域においては，p_x の値は α の値によって大きく左右されることを示唆している。実際 p_x の ρ についての偏微係数

$$\frac{\partial}{\partial\rho}p_x=\left[-\log(1+\alpha)+\frac{1}{\rho}+\cdots+\frac{1}{\rho+x_{i-1}}\right]p_x \tag{30}$$

において，数列 $\left\{\dfrac{1}{\rho+n-1}\right\}_{n=1}^{\infty}$ は単調減少数列であり，$\lim_{n\to\infty}\left(\sum_{k=1}^{n}\dfrac{1}{k}-\log n\right)=C$ (Euler の常数）であることに注意すれば，このことは容易に理解できる。

これらのことから α は数列 $\{p_x\}_{x=0}^{\infty}$ の各値を均す役割をし，ρ はこの数列で最高の値を示す項の位置を移行させるのに大きな役割を果たしている。

ρ, α のこのような性質は，これらのものについての検定の問題を考えるときに，よく注意しておかなければいけない。

6.　負の二項分布の特性値 ρ, α についての検定法

負の二項分布においては，$\rho<1$ の場合と $\rho>1$ の場合とでは，数列 $\{p_x\}_{x=0}^{\infty}$ の形態に大きな相違がみられる。従って ρ の値が正確に分らなくても，$\rho>1$ であるか $\rho<1$ であるかということは分っている場合が多い。

ρ の値が大きければ負の二項分布は正規分布に近くなる。したがってこのような場合には ρ, α の値について考えるよりも，むしろ平均値 m $(=\rho\alpha)$，分散 σ^2 $(=\rho\alpha+\rho\alpha^2)$ の値そのものについて考える方が好ましくなる。また ρ の値が余り大きくなくても，α の値が小さいとポアソン分布と殆んど変らなくなるから，この場合にはむしろ平均値 m の値についてのみ考える方がよい。このように ρ の値が大きいか，または α の値が小さいときには，それぞれ正規分布やポアソン分布の場合と同様に取り扱えばよいから，ここでは α の値が或る程度大きく，ρ の値がさほど大きくない場合についてのみ考える。

元来 ρ の値が小さいということは，λ の値の小さいもの（即ちこの場合には滅多に病欠しないような人）が非常に多いということを意味しているから，このような場合には，p_0 の値そのものについて考えることは大変意義のあるものとなる。実際 p_0 の値は ρ が十分小さくなれば 1 に近い値を示し，ρ の微小変化に対して大きく変動する。しかも α の値が或る程度大きいときには，α の値の多小の変動には余り影響さ

れない。それ故このような場合には，p_0 についての検定の問題は ρ についての検定問題に十分代用できるものとなる。

一方任意の正整数 N に対して $P\{X \geqq N\}=\sum_{x=N}^{\infty} p_x$ は N 回以上欠勤した人の見出される確率であるから，ρ の値が余り大きくない集団においては $\sum_{x=N}^{\infty} p_x$ についての検定の問題はそれ自身重要な意義をもっている。しかも N が十分大きければ $\sum_{x=N}^{\infty} p_x$ は ρ の多少の変動には余り影響されず，α の値によって大きく変動する性質があるので α についての検定にも十分代用できる。

以上のような考察に基づいて，表1に対して次のような検定を行なってみた。

III 群は I 群と II 群の中間の分布をしているので（表 3)，I 群と II 群の 2 つの集団について分布が同一であるか否かの検定を試みた。

（i） p_0 の差の検定（非欠勤者率の差の検定）

これは 2×2 分割表を作成し χ^2 検定を行えばよい。

回　数	I	II	計
0	328	324	652
≧1	87	76	163
計	415	400	815

$\chi^2=.490$

$P=.49$

（ii） $\sum_{x=N}^{\infty} p_x$ の差の検定（多数回欠勤者の率の差の検定）

この場合 N は出来るだけ大きい方が望ましいが，N を余り大きく取ると頻度（観測数）が少なくなるので，N として 5 を採用した。

回　数	I	II	計
0—4	413	392	805
5 以上	2	8	10
計	415	400	815

$\chi^2=3.873$

$P=.049$

以上二つの検定から I 群と II 群とは分布に有意な相違がみられ，それは両者の α に違いがあることによると解釈される。

なお次に掲げてあるのは，表1と同一調査対象者を職種別に分類しなおしたものである。A 職種は主として事務的仕事，B 職種は肉体労働を主とするものである。

表 6.　職種別病気欠勤回数の分布

回数	A 職	B 職	計
0	352	621	973
1	33	109	142
2	10	36	46
3	4	27	31
4	1	12	13
5	0	4	4
6	2	5	7
7	0	1	1
8	0	1	1
9	0	2	2
10	0	1	1
計	402	819	1221
$\bar{X}$	.20149	.49206	.39640
u^2	.45057	1.39939	1.10504
ρ	.16879	.26077	.22010
α	1.19373	1.88695	1.80100

この表から B 職種の群は A 職種の群よりも ρ, α が共に大きくなっている。

この表について p_0 (無欠勤者の率) の差の検定を行なってみたところ，有意な差がみられた。

回数	A	B	計
0	352	621	973
≧1	50	198	248
計	402	819	1221

$\chi^2=30.92$
$P<.001$

B 群は ρ だけでなく α も大きくなっているので，このように有意な差が生じたものと思われる。勿論この場合両群の α に有意な差があることは表 6 から明らかである。

7. 結　　論

一般に二つの集団が同一母集団からとりだされたものであるか否かの有意性の検定を行なう際に，母集団の分布の型が不明の場合には，$2\times n$ 分割表を作成し χ^2 を用いて斉次性の検定を行なうのが通常である。しかしながらこのように負の二項分布に従

うことが分っているときには，前節に述べたような観点に基づいて 2×2 分割表を作成し，χ^2 を用いて有意性の検定を行なうことにより，効果的に分布の相違点を見出すことができる。なお負の二項分布の平均値は ρ と α の積であるから，ポアソン分布の場合のように平均値だけを論じることはよくないということに注意しなければいけない。

参 考 文 献

[1] GREENWOOD, M. and YULE, G. U.: *An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurence of multiple attacks of disease or of repeated accidents.* J.R.S.S., 83 (1920), p. 255.

[2] NEWBOLD, E.: *Practical application of the statistics of repeated events, particularly to industrial accidents.* J.R.S.S., 90 (1927), p. 487.