2014年6月13日　統計数理研究所　オープンハウス

Conway-Maxwell-Poisson分布の一般化

井本　智明　　リスク解析戦略研究センター　特任研究員

序論

生物の死亡数や事故発生数等の個数データがどのような分布に従うのか、という問題は古くから関心がもたれている。このような問題に対し、ポアソン分布はポアソン過程や小数法則等の導出背景や唯一つのパラメータで簡潔に表現されるという利点から重宝されてきた。また、ポアソン分布のパラメータが他の分布に従うと仮定して導出される混合ポアソン分布は発生傾向の異なる事象が混合している場合も考慮に入れることができる分布として注目を集めてきた。

混合ポアソン分布はその導出背景から、大きな観測値が見られる割合が元のポアソン分布よりも大きくなる。このような性質を持つ分布を裾の長い分布と呼び、逆に、大きな観測値が見られる割合がポアソン分布よりも小さくなる分布を裾の短い分布と呼ぶ。

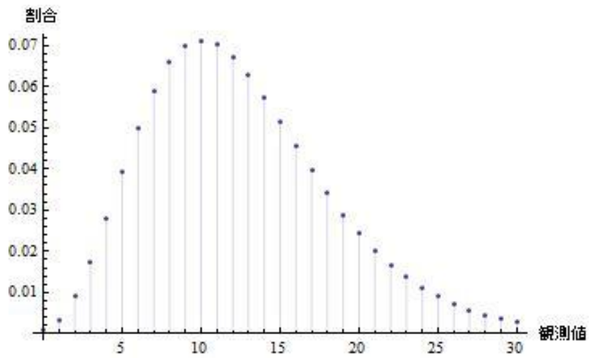

図1 裾の長い分布

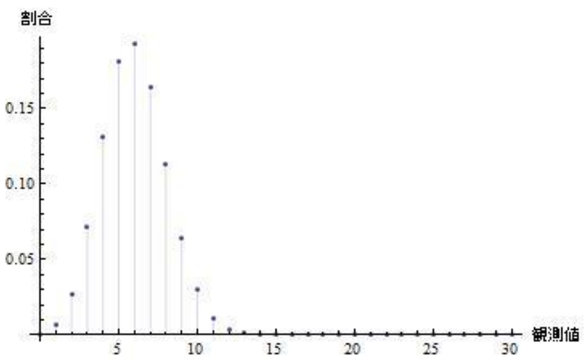

図2 裾の短い分布

離散型分布族ではこの裾の長短に関する尺度として、分散の平均に対する比で定義される分散指数が用いられる。ポアソン分布はこの値が1として特徴づけられることから、分散指数が1よりも大きくなる分布は（ポアソン分布に対して）過大分散型分布、1より小さくなる分布は(ポアソン分布に対して)過小分散型分布と呼ばれる。ポアソン分布と同様に古くから知られている二項分布と負の二項分布はそれぞれ過小分散型分布と過大分散型分布に分類される。

この分散指数の値が取り得る範囲が広い分布を考えることで、裾の長短について柔軟に対応できる分布を考えることができる。

Conway and Maxwell (1962)によって導出されたConway-Maxwell-Poisson分布は

$$\mathrm{P}(X=x)=\frac{\theta^x}{Z(\theta,r)(x!)^r},\ x=0,1,\ldots$$

という確率関数を持ち、ただし、$r,\theta>0$, $Z(\theta,r)=\sum_{k=0}^{\infty}\frac{\theta^k}{(k!)^r}$,

- $r\to 0$ のとき、負の二項分布の特別な形である幾何分布
- $r=1$のとき、ポアソン分布
- $r\to\infty$ のとき、二項分布の特別な形であるベルヌーイ分布

となる分布である。
この分布は、$r<1$ のとき過大分散型、$r>1$ のとき過小分散型となる柔軟な分布としてShmueli (2005)が指摘し、注目を集めた。

問題点と解決案

Conway-Maxwell-Poisson分布に含まれる二つのパラメータはその平均と分散を同時にコントロールでき、そのためこの分布の平均と裾の長短を同時にコントロールできる。しかし、二つしかパラメータを持たないために、それ以外の事柄については柔軟性を有さない。

例えば、観測値0の頻度を考えたとき、平均の位置と裾の長短にしか対応できないConway-Maxwell-Poisson分布では実際の頻度と大きなずれを生じさせてしまうことがある。

観測値0の頻度に対する柔軟性を持つ分布を考えるには負の二項分布を含むConway-Maxwell-Poisson分布を考えればよい。

これは元のConway-Maxwell-Poisson分布に含まれている幾何分布は観測値0が必ず最頻値をとる場所となってしまう一方、その一般化形である負の二項分布は正の値上ならばどこにでも最頻値をとれることから、幾何分布よりも観測値0の頻度について柔軟に値をとれる、という考えからくるものである。

一般化Conway-Maxwell-Poisson分布

確率関数

$$\mathrm{P}(X=x)=\frac{\Gamma(\nu+x)^r\theta^x}{C(r,\nu,\theta)x!},\ x=0,1,\ldots$$

を持つ確率分布を考える。

ただし、$r<1,\nu>0,\theta>0$, $C(r,\nu,\theta)=\sum_{k=0}^{\infty}\frac{\Gamma(\nu+k)^r\theta^k}{k!}$.

この分布は

- $\nu=1$のとき、Conway-Maxwell-Poisson分布
  ⇒*分布の裾の長短に柔軟*
- $r\to 1$のとき、負の二項分布
  ⇒*観測値0の頻度がConway-Maxwell-Poisson分布より柔軟*

となる。

最頻値について調べると

- $r<0$, もしくは $r<1,\nu>1$ のとき、単峰型分布となる。
- $0<r<1,0<\nu<1,\theta\nu^r<1$ のとき、一つの最頻値を0でとる双峰型分布となることがある。

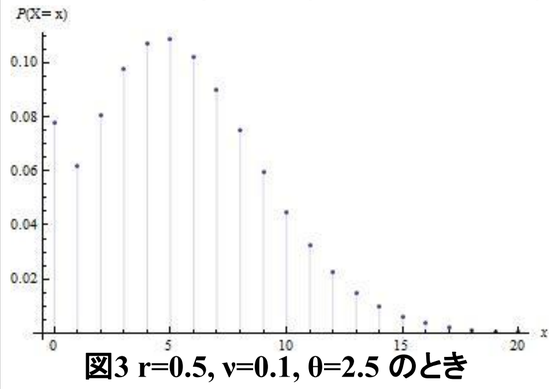

図3 r=0.5, v=0.1, θ=2.5 のとき

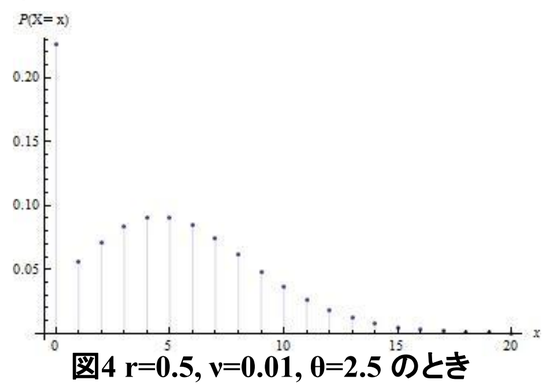

図4 r=0.5, v=0.01, θ=2.5 のとき

・データへの当てはめ例とその比較

Table 1: The number of spots in southern pine beetle, *Dendroctonus frontalis Zimmermann* (Coleopetra: Scolytidae), in Southeast Texas (Lin, 1985)

Count	Observed	COMP	GCOMP	Count	Observed	COMP	GCOMP
0	1169	927.60	1168.66	11	2	0.04	1.99
1	144	372.48	152.55	12	0	0.02	1.32
2	92	149.57	80.48	13	0	0.01	0.87
3	54	60.06	49.82	14	1	0.00	0.57
4	29	24.12	32.52	15	0	0.00	0.38
5	18	9.68	21.68	16	0	0.00	0.25
6	10	3.89	14.58	17	0	0.00	0.16
7	12	1.56	9.83	18	0	0.00	0.10
8	6	0.63	6.63	19+	1	0.00	0.1
9	9	0.25	4.45				
10	3	0.10	2.98	Total	1550	1550.00	1550.00
COMP:	$\hat{r}=0.0000$	GCOMP:	$\hat{r}=0.8750$	$\log L$		−1755.49	−1552.49
	$\hat{\theta}=0.4015$		$\hat{\nu}=0.1011$	χ^2		939.53	10.48
			$\hat{\theta}=0.9699$	d.f.		8	7
				p-value		0.000	0.163

Table 2: The number of roots produced by 270 shoots of the apple cultivar Trajan (Ridout et al., 1998)

Count	Observed	ZICOMP	GCOMP	Count	Observed	ZICOMP	GCOMP
0	64	64.00	64.00	10	17	12.45	12.65
1	10	6.48	7.25	11	12	9.03	9.18
2	13	11.93	11.99	12	5	6.25	6.33
3	15	17.61	17.23	13	2	4.14	4.16
4	21	22.21	21.68	14	3	2.63	2.62
5	18	24.79	24.37	15	0	1.61	1.59
6	24	25.06	24.85	16	0	0.95	0.92
7	21	23.28	23.30	17	1	1.18	1.10
8	23	20.10	20.28				
9	21	16.28	16.50	Total	270	270.00	270.00
ZICOMP:	$\hat{w}=0.2281$	GCOMP:	$\hat{r}=0.3826$	$\log L$		−673.19	−672.40
	$\hat{r}=0.5463$		$\hat{\nu}=0.0001$	χ^2		11.21	9.45
	$\hat{\theta}=2.6896$		$\hat{\theta}=3.3061$	d.f.		10	10
				p-value		0.341	0.490

参考文献

1. **Conway, R. W., and Maxwell, W. L. (1962). A queueing model with state dependent service rates. Journal of Industrial Engineering, 12, 132--136.**
2. **Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54, 127--142.**
3. **Lin, S-K. (1985). Characterization of lightning as a disturbance to the forest ecosystem in East Texas. M.Sc. thesis, Texas A&M University, College Station.**
4. **Ridout, M., Demetrio, C. G. B., and Hinde, J. (1998). Models for count data with many zeros. International Biometric Conference, Cape Town, December.**