Proceedings of Joint Conference of Korean Data And Information Science Society and The Korean Data Analysis Society, April 30 - May 1, 2004, pp. 111-119

유사이항분포와 유사다항분포의 통계적 성질

안성진[1] · 정연선[2]

요 약

유사이항분포와 유사다항분포를 소개하고 베타분포와 Dirichlet 분포와의 관계를 밝힘으로써 심플렉스상에서 정의되는 성분데이터의 분석을 위한 새로운 방법을 제시하는 토대를 마련하고자 한다.

1. 서론

우리는 확률변수 Y가 베타분포 즉, $Y\sim BETA(\alpha,\beta)$ 이고, 확률변수 X가 이항분포 즉, $X\sim B(\alpha+\beta-1,p)$ 이면, $P(Y<p)=P(X\geq\alpha)$이라는 사실을 알고 있다. 그리고 차원을 높여서 확률벡터 $\boldsymbol{Y}$가 Dirichlet 즉, $\boldsymbol{Y}\sim DIRI(b_0,b_1,\cdots,b_r)$ 이고, 확률벡터 $\boldsymbol{X}$가 다항분포 즉, $\boldsymbol{X}\sim MULT(n,p_0,p_1,\cdots,p_r)$ 이면, 역시 $P(\boldsymbol{Y}<\boldsymbol{p})=P(\boldsymbol{X}\geq\boldsymbol{b})$ 임을 알고 있다. 그렇지만 이는 베타분포와 Dirichlet 분포의 모수가 모두 양의 정수인 경우에만 적용할 수 있는 성질이다. 본 논문에서는 베타분포와 이항분포 그리고, 다항분포와 Dirichlet 분포의 이런 성질을 모수가 양의 실수인 경우에도 적용할 수 있도록 확장하는 과정에서 만나게 된 '유사이항분포'와 '유사다항분포'에 대해 연구하려한다. 안&정(2002)은 유사이항분포와 유사다항분포를 확률적 분류규칙에 적용하고 있다.

안&정(2002)은 다음과 같은 가정을 이용해서 성분벡터 $\mathbf{x}$와 $\mathbf{y}$의 관련성을 측정하였다.

[가정 1] $\mathbf{x},\mathbf{y}$는 각각 모수 $\boldsymbol{a}=(a_0,\ldots,a_f)$, $\boldsymbol{b}=(b_0,\ldots,b_r)$를 가지는 Dirichlet 분포를 따르는 $f+1$-차원과 $r+1$-차원의 확률벡터이다.

[가정 2] $a^*=\sum_{i=0}^{f}a_i$, $b^*=\sum_{j=0}^{r}b_j$ 라 하고, $\alpha_i=\log\frac{(a_i/a^*)}{(a_0/a^*)}=\log\frac{a_i}{a_0}$ 그리고 $\beta_j=\log\frac{b_j}{b_0}$ 라 할 때 $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_f)$와 $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_r)$의 결합밀도함수는 다음과 같은 $r+f$-차원의 다변량 정규분포를 따른다:

$$(\boldsymbol{\alpha},\boldsymbol{\beta})\sim MN_{f+r}(\boldsymbol{\mu},\boldsymbol{\Sigma}),\quad \boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}_\alpha\\ \boldsymbol{\mu}_\beta\end{bmatrix},\quad \boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_\alpha & \boldsymbol{\Sigma}_{\alpha\beta}\\ \boldsymbol{\Sigma}_{\beta\alpha} & \boldsymbol{\Sigma}_\beta\end{bmatrix}.$$

[1]660-701 경남 진주시 가좌동 900번지, 경상대학교 컴퓨터정보통신연구원, 통계정보학과 교수.
[2]660-701 경남 진주시 가좌동 900번지, 경상대학교 통계정보학과 박사과정

위의 가정과 다변량 정규분포의 조건부 분포를 이용하면 $\boldsymbol{\alpha}, \boldsymbol{x}$가 주어졌을 때 $\mathbf{y}$의 확률밀도함수는 식 (1)으로부터 구할 수 있다:

$$f(\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{x}) = \int f_y(\boldsymbol{y}|\boldsymbol{b}) f_b(\boldsymbol{b}|\boldsymbol{\alpha}, \boldsymbol{x}) d\boldsymbol{b} \tag{1}$$

여기서, $f_y(\boldsymbol{y}|\boldsymbol{b}) = \dfrac{\Gamma(b^*)}{\prod_{j=0}^{r}\Gamma(b_j)} \prod_{j=0}^{r} y_j^{b_j-1}$.

정규분포와 로지스틱 정규분포는 로지스틱 변환에 의한 일대일 대응관계에 있기 때문에 $f(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{x})$로 부터 $f_b(\boldsymbol{b}|\boldsymbol{\alpha}, \boldsymbol{x})$를 구할 수 있다. 즉, $\delta_i = \dfrac{b_i}{b^*}$, $i=0, \ldots, r$ 로 정의하고 $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_r]$라 하면, $\boldsymbol{\delta}|\boldsymbol{\alpha}, \boldsymbol{x}$는 로지스틱 정규분포를 따르게 되고 이로부터 $f_b(\boldsymbol{b}|\boldsymbol{\alpha}, \boldsymbol{x})$ 는 다음과 같이 얻어질 수 있다:

$$f_b(\boldsymbol{b}|\boldsymbol{\alpha}, \boldsymbol{x}) = \frac{b^*}{|2\pi\Sigma_L|^{1/2} \prod_{j=0}^{r} b_j} \exp\left[-\frac{1}{2}\left(\log\frac{\boldsymbol{b}}{b_0} - \mu_L\right)' \Sigma_L^{-1}\left(\log\frac{\boldsymbol{b}}{b_0} - \mu_L\right)\right], \tag{2}$$

여기서 $\log\dfrac{\boldsymbol{b}}{b_0} = \left(\log\dfrac{b_1}{b_0}, \ldots, \log\dfrac{b_r}{b_0}\right)'$.

Dirichlet 분포의 차원이 일차원인 경우는 베타분포가 된다. 베타분포의 확률밀도함수와 $b_1|\alpha, x$의 밀도함수는 각각 식(3), (4)와 같이 주어진다.

$$f_y(y|b_1) = \frac{\Gamma(b^*)}{\Gamma(b_1)\Gamma(b^*-b_1)} y^{b_1-1}(1-y)^{b^*-b_1-1}, \tag{3}$$

$$f_b(b_1|\boldsymbol{\alpha}, \boldsymbol{x}) = \frac{b^*}{\sqrt{2\pi}\sigma_L b_1(b^*-b_1)} \exp\left[-\left(\log\frac{b_1}{(b^*-b_1)} - \mu_L\right)^2 / 2\sigma_L^2\right], \tag{4}$$

여기서 $\mu_L = \mu_{\beta_1} + \sigma_{\beta_1} \sum_{i=1}^{f} \rho_{1i}\left(\dfrac{\alpha_i - \mu_{\alpha_i}}{\sigma_{\alpha_i}}\right)$, $\sigma_L^2 = \sigma_{\beta_1}^2\left(1 - \sum_{i=1}^{f}\rho_{1i}^2\right)$.

베타분포의 모수가 양의 정수인 경우는 이항분포와의 관계를 이용해서 $f(y|\alpha, x)$의 근사식을 구할 수 있다. 즉, W를 이항분포 $B(n, y)$의 확률변수로 두고 L을 로지스틱 정규분포 $LN(\mu_L, \sigma_L^2)$의 확률변수로 두었을 때, $F(y|\boldsymbol{\alpha}, \boldsymbol{x})$는 다음과 같이 얻어진다:

$$F(y|\alpha, x) = \int f(y|\alpha, x) dy = \int_y \int_{b_1} f_y(y|b_1) f_L(b_1|\alpha, x) db_1 dy$$

$$= \int_{b_1} F_y(y|b_1) f_L(b_1|\alpha, x) db_1 = \int_{b_1} [1 - F_W(b_1; n, y)] f_L(b_1|\alpha, x) db_1$$

$$= \int_{b_1} P\{W \geq b_1\} f_L(b_1|a,x)db_1 \quad = E^L[P\{W \geq b^*L|L\}]$$

$$= P\{W \geq b^*L\}.$$

그렇지만 베타분포의 모수가 양의 정수가 아닌 실수인 경우는 위의 식을 이용할 수 없다. 본 논문에서는 베타분포의 모수가 실수인 경우에도 위의 성질을 이용할 수 있는 일반화된 분포를 정의할 것이다.

2. 유사이항분포와 유사다항분포의 성질

가. 유사이항분포

유사이항분포는 이산형에서의 이항분포를 연속형으로 확장한 형태의 밀도함수로 이항분포가 가지고 있는 특성을 그대로 가지고 있다.

[정의 1] 다음과 같은 밀도함수를 가지는 확률변수 z는 모수 t, p를 가지는 유사이항분포(Quasi Binomial Distribution)로 정의되며 $QB(t, p)$ 라고 표기될 것이다.

$$f(z|t,p) = \frac{1}{B(t,p)} \frac{\Gamma(t-1)}{\Gamma(z)\Gamma(t-z)} p^{z-1}(1-p)^{t-z-1}\ ;\ 0<p<1\ ,\ 1 \leq z \leq t$$

여기서 함수 $B(t,p)$는 다음과 같이 주어진다:

$$B(t,p) = \int_1^t \frac{\Gamma(t-1)}{\Gamma(z)\Gamma(t-z)} p^{z-1}(1-p)^{t-z-1} dz$$

[성질 1] $B(t, p)$의 성질

1. $B(t,p) \leq 1$
2. $t \to \infty$ 또는 $p \to \frac{1}{2}$ 이면, $B(t,p) \to 1$
3. $tp > 5$ 이면 $B(t,p) \doteq 1$ ($p > 0.5$ 이면, $t(1-p) > 5$)
4. $t_1 > t_2$이면 $B(t_1,p) \geq B(t_2,p)$
5. $B(t,p) = B(t,1-p)$
6. $B(t,p) \doteq B(t-1,p)$, $t > 10$

t, p에 따른 $B(t,p)$값들의 일부가 [부표 1]에 제시되어있다.

[성질 2] 유사이항분포의 점근적 성질

1. $z \sim QB(t,p)$ 이고 $t \to \infty$ or $p \to \frac{1}{2}$ 이면,

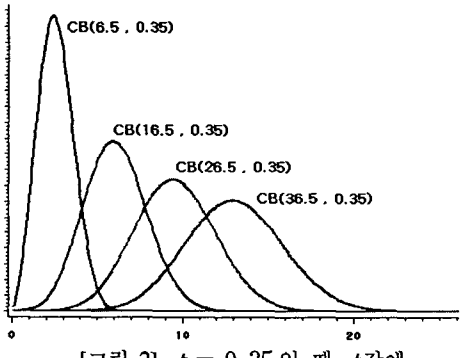
$E[z] \simeq [(t-2)p+1]$

$Var(z) \simeq (t-2)p(1-p)$

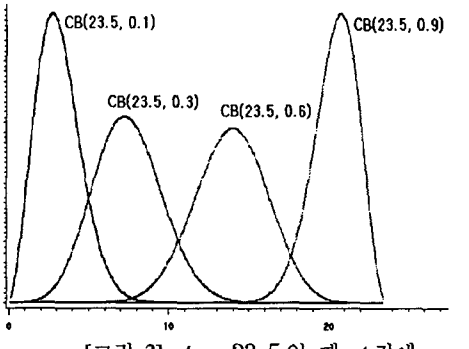
2. $t \to \infty$ 이면 $QB(t,p) \dot{\sim} N((t-2)p+1, (t-2)p(1-p))$
3. $Y \sim Beta(\alpha, \beta)$, $Z \sim QB(t+1, p)$ 이면 (단, $t = \alpha+\beta$),
   $P(Y \leq p) \simeq P(Z \geq \alpha + \frac{1}{2})$ 이다.

유사이항분포의 세 번째 성질은 베타분포와 이항분포의 관계를 확장한 것으로써 Dirichlet 분포를 따르는 성분데이터분석을 위한 중요한 실마리가 된다. 유사이항분포의 성질에 대한 증명은 [부록 1]에 제시되어 있다.

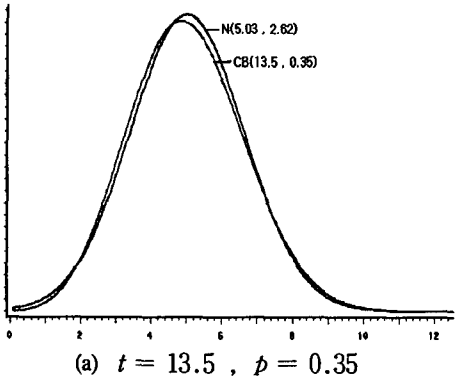
t값과 p값에 따른 유사이항분포의 형태가 [그림 2]와 [그림 3]에 제시되어있다. [그림 4]는 t값이 클 때 정규분포에 근사한다는 것을 보여준다. 베타분포와 유사이항분포의 누적분포의 관계는 [그림 5]에 제시되어 있다.



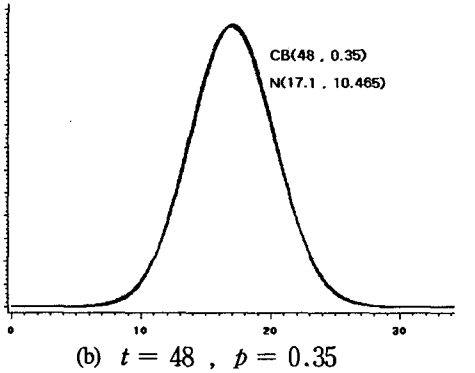
[그림 2] $p = 0.35$일 때 t값에 따른 유사이항분포



[그림 3] $t = 23.5$일 때 p값에 따른 유사이항분포

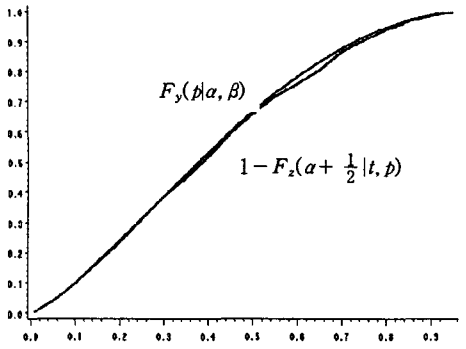


(a) $t = 13.5$, $p = 0.35$

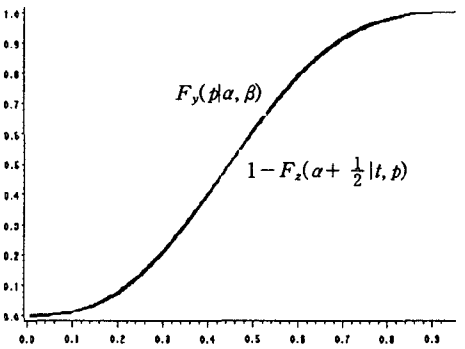


(b) $t = 48$, $p = 0.35$

[그림 4] t값에 따른 유사이항분포와 정규분포의 비교



(a) $t = 4.5$, $b = 3.4$



(b) $t = 8.5$, $b = 6.5$

[그림 5] 베타분포와 유사이항분포의 누적분포 비교

나. 유사다항분포

유사다항분포는 일변량에서의 유사이항분포를 r차원으로의 확장한 분포로 이산형의 다항분포에 대응되는 분포다.

[정의 2] 다음과 같은 밀도함수를 가지는 확률벡터 $\boldsymbol{z} = (z_1, \ldots, z_r)$는 모수 t, $\boldsymbol{p}$를 가지는 유사다항분포(Quasi Multinomial Distribution)라고 정의되며 $QM_r(t, p_1, \cdots, p_r)$로 표기될 것이다.:

$$f_r(\boldsymbol{z}|t,\boldsymbol{p}) = \frac{1}{M_r(t,\boldsymbol{p})} \frac{\Gamma(t-r)}{\Gamma\left(t-\sum_{i=1}^{r} z_i\right)\prod_{i=1}^{r}\Gamma(z_i)} \left(1-\sum_{i=1}^{r} p_i\right)^{t-\sum_{i=1}^{r} z_i - 1} \prod_{i=1}^{r} p_i^{z_i-1}$$

$$; \ 0<p_i<1 \ , \ 1\le z_i < t, \ \sum_{i=0}^{r} z_i = t$$

여기서 함수 $M_r(t, \boldsymbol{p})$는 다음과 같이 주어진다:

$$M_r(t,\boldsymbol{p}) = \int \frac{\Gamma(t-r)}{\Gamma\left(t-\sum_{i=1}^{r} z_i\right)\prod_{i=1}^{r}\Gamma(z_i)} \left(1-\sum_{i=1}^{r} p_i\right)^{t-\sum_{i=1}^{r} z_i - 1} \prod_{i=1}^{r} p_i^{z_i-1} dz$$

[성질 3] $M_r(t, p)$의 성질

1. $M_r(t, p_1, \ldots, p_r) \le 1$
2. $t \to \infty$ 또는 $p_i \to \frac{1}{r+1}$ 이면 $M_r(t, p_1, \ldots, p_r) \to 1$
3. $t_1 > t_2$ 이면 $M_r(t_1, p_1, \ldots, p_r) > M_r(t_2, p_1, \ldots, p_r)$
4. $M_r(t, a_k, a_1, \ldots, a_r) = M_r(t, a_1, a_k, \ldots, a_r)$
5. $M_r(t, p_1, \ldots, p_r) \doteq M_r(t-1, p_1, \ldots, p_r)$

[성질 4] 유사다항분포의 점근적 성질

1. $z \sim QM_r(t, \boldsymbol{p})$ 이고 $t \to \infty$이면,

   $E(z_i) \simeq (t-r-1)p_i + 1,$

   $Var(z_i) \simeq (t-r-1)p_i(1-p_i)$

   $Cov(z_i, z_j) \simeq -(t-r-1)p_i p_j$

   $Corr(z_i, z_j) \simeq -\sqrt{\dfrac{p_i p_j}{(1-p_i)(1-p_j)}}$

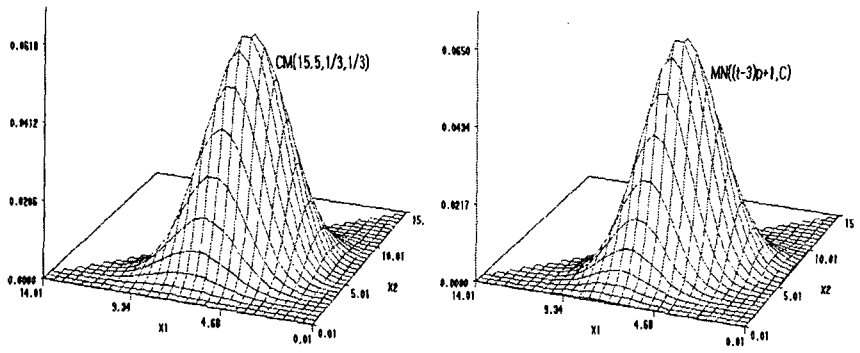
2. $t \to \infty$ 이면 $QM_r(t, \boldsymbol{p}) \doteq MN((t-r-1)\boldsymbol{p}+\mathbf{1}, \Sigma)$

   (단, 다변량 정규분포의 공분산 Σ는 유사다항분포의 공분산 행렬과 동일)

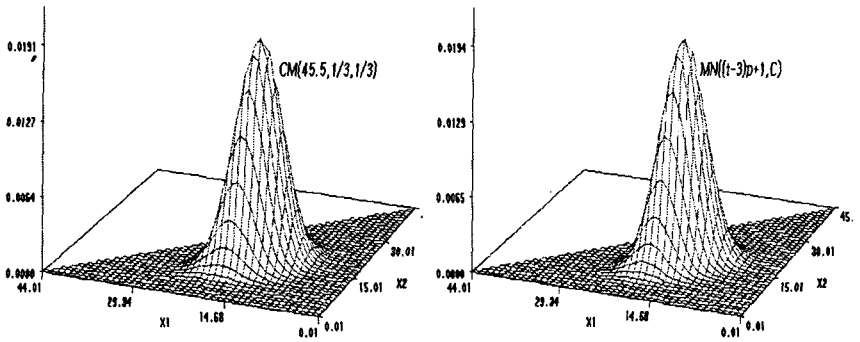
3. $\boldsymbol{Y} \sim Dirichlet(\boldsymbol{\beta})$, $\boldsymbol{Z} \sim QM(t+r, \boldsymbol{p})$ 이면, $\left(\text{단}, t = \sum_{j=0}^{k} \beta_j\right)$,

   $P(\boldsymbol{Y} \le \boldsymbol{p}) \simeq P(\boldsymbol{Z} \ge \boldsymbol{\beta} + \frac{1}{2}\mathbf{1})$ 이다. 단, $\mathbf{1} = [1, 1, \ldots, 1]'$

[그림 6]에서 t값이 커지면 유사다항분포는 다변량정규분포로 근사한다는 것을 보여주고 있다.



(a) $r = 2$; $t = 15.5$, p_1 1/3 , $p_2 = 1/3$ 일때



(b) $r = 2$; $t = 45.5$, $p_1 = 1/3$, $p_2 = 1/3$ 일때

[그림 6] 유사다항분포와 다변량정규분포의 밀도함수 비교

3. 결론

본 논문에서 소개한 두 확률밀도함수인 유사이항분포와 유사다항분포는 심플렉스 공간상의 확률밀도함수인 Dirichlet 분포와의 관련성을 통해 성분 데이터의 분석에 이용될 수 있다. 심플렉스 공간상의 벡터에 대한 분석을 위해서는 지금까지 성분벡터를 로지스틱 변환시켜 정규분포로 근사시키는 방법이 많이 사용되었다. Dirichlet 분포는 심플렉스 상의 거의 모든 형태의 데이터를 아주 잘 설명할 수 있다는 것이 몽테카를로 시뮬레이션을 통해 알려졌다.

[부표 1] $B(t,p)=\int_1^t \frac{\Gamma(t-1)}{\Gamma(z)\Gamma(t-z)} p^{z-1}(1-p)^{t-z-1}dz$

t \ p	0.01	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
2	5.487	2.348	1.752	1.515	1.386	1.306	1.253	1.218	1.196	1.183	1.179
3	3.575	1.707	1.357	1.222	1.150	1.107	1.079	1.061	1.049	1.043	1.041
4	2.863	1.473	1.220	1.126	1.079	1.051	1.034	1.024	1.017	1.014	1.013
5	2.476	1.348	1.150	1.080	1.046	1.028	1.017	1.011	1.007	1.005	1.005
6	2.228	1.270	1.108	1.054	1.029	1.016	1.009	1.005	1.003	1.002	1.002
7	2.054	1.217	1.081	1.038	1.019	1.010	1.005	1.003	1.002	1.001	1.001
8	1.924	1.179	1.062	1.027	1.013	1.006	1.003	1.002	1.001	1.000	1.000
9	1.822	1.150	1.049	1.020	1.009	1.004	1.002	1.001	1.000	1.000	1.000
10	1.740	1.127	1.039	1.015	1.006	1.003	1.001	1.001	1.000	1.000	1.000
11	1.740	1.127	1.039	1.015	1.006	1.003	1.001	1.001	1.000	1.000	1.000
12	1.673	1.109	1.031	1.011	1.004	1.002	1.001	1.000	1.000	1.000	1.000
13	1.617	1.094	1.025	1.009	1.003	1.001	1.000	1.000	1.000	1.000	1.000
14	1.568	1.082	1.021	1.007	1.002	1.001	1.000	1.000	1.000	1.000	1.000
15	1.527	1.072	1.017	1.005	1.002	1.001	1.000	1.000	1.000	1.000	1.000
16	1.490	1.063	1.014	1.004	1.001	1.000	1.000	1.000	1.000	1.000	1.000
17	1.458	1.056	1.012	1.003	1.001	1.000	1.000	1.000	1.000	1.000	1.000
18	1.429	1.050	1.010	1.002	1.001	1.000	1.000	1.000	1.000	1.000	1.000
19	1.404	1.044	1.008	1.002	1.001	1.000	1.000	1.000	1.000	1.000	1.000
20	1.360	1.036	1.006	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000
25	1.279	1.022	1.003	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
30	1.224	1.014	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

[부록 1] 유사이항분포의 점근적 성질 증명

1. 기댓값과 분산의 증명

$$E(z-1)=\int_1^t \frac{1}{B(t,p)}\frac{\Gamma(t-1)}{\Gamma(z)\Gamma(t-z)}(z-1)p^{z-1}(1-p)^{t-z-1}dz$$

$$=(t-2)p\frac{1}{B(t,p)}\int_1^t \frac{\Gamma(t-2)}{\Gamma(z-1)\Gamma(t-z)}p^{z-2}(1-p)^{t-z-1}dz$$

$z-1=y$로 두면,

$$\int_1^t \frac{\Gamma(t-2)}{\Gamma(z-1)\Gamma(t-z)} p^{z-2}(1-p)^{t-z-1}dz = \int_0^{t-1} \frac{\Gamma(t-2)}{\Gamma(y)\Gamma(t-1-y)} p^{y-1}(1-p)^{t-y-1}dy$$

$$= B(t-1,p) + \int_0^1 \frac{\Gamma(t-2)}{\Gamma(y)\Gamma(t-1-y)} p^{y-1}(1-p)^{t-y-1}dy$$

$\therefore\ E[z] \simeq [(t-2)p+1]$

$$E[(z-1)(z-2)] = \int_1^t \frac{1}{B(t,p)} \frac{\Gamma(t-1)}{\Gamma(z)\Gamma(t-z)}(z-1)(z-2)p^{z-1}(1-p)^{t-z-1}dz$$

$$= (t-2)(t-3)p^2 \frac{1}{B(t,p)} \int_1^t \frac{\Gamma(t-3)}{\Gamma(z-2)\Gamma(t-z)} p^{z-3}(1-p)^{t-z-1}dz$$

$$\simeq (t-2)(t-3)p^2$$

$\therefore\ Var(z) = E(z^2) - E(z)^2 \simeq (t-2)p(1-p)$

2. $Y \sim Beta(\alpha,\beta)$, $Z \sim QB(t+1,p)$ 이면, $P(Y \le p) \simeq P(Z \ge \alpha + \frac{1}{2})$ 증명

$$P(Z > \alpha+\tfrac{1}{2} \mid t+1,p) = \frac{1}{B(t+1,p)} \int_{\alpha+\frac{1}{2}}^{t+1} \frac{\Gamma(t)}{\Gamma(z)\Gamma(t-z+1)} p^{z-1}(1-p)^{t-z}dz$$

$$\frac{d}{dp}\left(\frac{1}{B(t+1,p)}\right) = \frac{1}{B(t+1,p)^2} \int_0^{t+1} \frac{\Gamma(t)}{\Gamma(z)\Gamma(t-z+1)} \times$$

$$[(z-1)p^{z-2}(1-p)^{t-z} - (t-z)p^{z-1}(1-p)^{t-z-1}]dz$$

$$= \frac{1}{B(t+1,p)^2}\left(\int_0^{t+1} \frac{z\Gamma(t)}{\Gamma(z)\Gamma(t-z+1)} p^{z-2}(1-p)^{t-z}dz \right.$$

$$+ \int_0^{t+1} \frac{z\Gamma(t)}{\Gamma(z)\Gamma(t-z+1)} p^{z-1}(1-p)^{t-z-1}dz$$

$$- \int_0^{t+1} \frac{\Gamma(t)}{\Gamma(z)\Gamma(t-z+1)} p^{z-2}(1-p)^{t-z}dz$$

$$\left. - \int_0^{t+1} \frac{t\Gamma(t)}{\Gamma(z)\Gamma(t-z+1)} p^{z-1}(1-p)^{t-z-1}dz \right)$$

$$= \frac{1}{B(t+1,p)}\left(\frac{E[z]}{p} + \frac{E[z]}{1-p} - \frac{1}{p} - \frac{t}{1-p} \right)$$

$$= \frac{1}{B(t+1,p)} \frac{t-1}{1-p}\left(\frac{B(t,p)}{B(t+1,p)} - 1 + o_1(t) \right) \simeq 0$$

$$\frac{d}{dp} P(z > \alpha+\tfrac{1}{2} \mid t+1,p) = \frac{1}{B(t+1,p)} \int_{\alpha+\frac{1}{2}}^{t+1} \frac{\Gamma(t)}{\Gamma(z)\Gamma(t-z+1)} \times$$

$$[(z-1)p^{z-2}(1-p)^{t-z} - (t-z)p^{z-1}(1-p)^{t-z-1}]dz$$

$$-\left(\frac{d}{dp}\frac{1}{B(t+1,p)}\right) \int_{\alpha+\frac{1}{2}}^{t+1} \frac{\Gamma(t)}{\Gamma(z)\Gamma(t-z+1)} p^{z-1}(1-p)^{t-z}dz$$

$$= \frac{1}{B(t+1,p)}\left(\int_{\alpha+\frac{1}{2}}^{t+1} \frac{\Gamma(t)}{\Gamma(z-1)\Gamma(t-z+1)} p^{z-2}(1-p)^{t-z}dz \right.$$

$$
\begin{aligned}
&- \int_{\alpha-\frac{1}{2}}^{t+1} \frac{\Gamma(t)}{\Gamma(z)\Gamma(t-z)} p^{z-1}(1-p)^{t-z-1} dz \\
&+ \int_{\alpha-\frac{1}{2}}^{\alpha+\frac{1}{2}} \frac{\Gamma(t)}{\Gamma(z)\Gamma(t-z)} p^{z-1}(1-p)^{t-z-1} dz \Big) + o_2(t) \\
&= \frac{1}{B(t+1,p)} \int_{\alpha-\frac{1}{2}}^{\alpha+\frac{1}{2}} \frac{\Gamma(t)}{\Gamma(z)\Gamma(t-z)} p^{z-1}(1-p)^{t-z-1} dz + o_2(t) \\
&= \frac{(t-1)B(t,p)}{B(t+1,p)} \Big(F(\alpha+\tfrac{1}{2}|t,p) - F(\alpha-\tfrac{1}{2}|t,p) \Big) + o_2(t) \\
&= \frac{(t-1)B(t,p)}{B(t+1,p)} [f(\alpha|t,p) + o_3(t)] + o_2(t)
\end{aligned}
$$

$$
\begin{aligned}
P\big(z > \alpha+\tfrac{1}{2}|p\big) &= \int_0^p \frac{d}{dy} P\big(z > \alpha+\tfrac{1}{2}|y\big) dy \\
&= \int_0^p \frac{(t-1)B(t,y)}{B(t+1,y)} [f(\alpha|t,y) + o_3(t)] + o_2(t)\, dy \\
&= \int_0^p \frac{1}{B(t+1,y)} \frac{\Gamma(t)}{\Gamma(\alpha)\Gamma(t-\alpha)} y^{\alpha-1}(1-y)^{t-\alpha-1} dy + o(t) \\
&(t \to \infty \text{ 이면, } o(t) \to 0,\ B(t+1,p) \to 1) \\
&\simeq \int_0^p \frac{\Gamma(t)}{\Gamma(\alpha)\Gamma(t-\alpha)} y^{\alpha-1}(1-y)^{t-\alpha-1} dy = P(Y < p)
\end{aligned}
$$

$\therefore$ $Y \sim Beta(\alpha, \beta)$, $Z \sim QB(t+1, p)$이면, $P(Y \le p) \simeq P(Z \ge \alpha + \frac{1}{2})$ ∎

참고문헌

1. Aitchison, J. & Begg, C.B. (1976). Statistical diagnosis when cases are not classified with certainty. *Biometrika,* Vol. 63, 1-12.

2. Aitchison, J. and Shen, S.M. (1980). Logistic-normal distributions : some properties and uses. *Biometrika,* Vol. 67, 261-272

3. Aitchison, J. (1982). The Statistical analysis of Compositional Data. *Journal of Royal Statistical Society,* Vol. 2, pp 139-177.

4. Aitchison, J. (1985). A general Class of Distributions on the Simplex. *Journal of Royal Statistical Society,* Vol 47, pp 136-146

5. Brehm, J., Gates, S., and Gomez, B. (1998). A Monte Carlo comparison of Methods for Compositional Data Analysis. *Prepared for presentation at the 1998 annual meeting of the Society for Political Methodology*

6. 안성진, 정연선(2002), 성분데이터에 대한 확률적 분류규칙, *Journal of the Korean Data Analysis Society,* Vol4, pp 437-450.