연세경제연구 제 Ⅷ 권 제 2 호
Yonsei Economic Studies, Vol. Ⅷ, No. 2 (September 2001)

방사선 역학연구에서 사용되는 통계적 방법의 소개

김 병 수

일본 히로시마와 나가사키에 각각 소재하고 있는 방사선영향연구소(RERF)는 지난 50여 년 동안 원폭 생존자들을 추적하여 역학조사 자료를 집계, 분석하고 있으며 電離 放射線이 인간에게 미치는 여러 가지 영향, 특히 발암에 대한 장기 효과를 평가하고 있다. RERF의 연구결과는 방사선 방어에 대한 국제적 기준을 마련하는 중요한 근거가 되어 왔다. 본 논문에서는 우선 RERF에서 추적조사하고 있는 네 가지 동집단 자료를 소개하고, 그 중 가장 중요한 Life Span Study(LSS)자료에 기초하여 발암의 위해도를 추정하는 통계적 모형의 근거와 이 과정에서 포아송 회귀분석이 어떻게 적용되고 있는지를 예와 함께 소개하고자 한다.

Ⅰ. 서 론

일본의 방사선영향연구소(Radiation Effects Research Foundation : RERF)는 1947년 미국의 국립과학원(National Academy of Sciences)에 의하여 세워진 원폭희생자위원회(Atomic Bomb Casualty Commission : ABCC)를 발전 승계한 연

연세대학교 상경대학 응용통계학과, 서울특별시 서대문구 신촌동 134, 120-749.

본 논문은 아시아 연구 기금의 2000년도 연구지원사업의 지원을 받아 수행되었다. 본 논문 수행을 위하여 1999년부터 2000년 겨울기간 동안 일본 히로시마 소재 방사선영향연구소(RERF)를 방문하였으며, 방문기간중에 편의를 제공하여 준 RERF 당국과 특히 많은 학문적 조언을 하여 준 통계학과의 Dr. Dale Preston과 Dr. Donald Pierce에게 감사를 표한다.

구기관으로서 1975년에 설립되었고, 일본 후생성을 통한 일본정부와 미국 에너지성과의 계약관계에 의하여 국립과학원을 통한 미국 정부로부터 균등하게 지원을 받는 일본의 비영리 사단법인(a private nonprofit Japanese foundation)이다. RERF(당시는 ABCC)는 1950년 히로시마와 나가사키의 특별국세조사에 기초하여 약 12만 명에 이르는 원폭피해자 동집단(cohort)을 구축하고 이를 壽命調査(Life Span Study : LSS)라고 부르며, 지금까지 약 반 세기에 걸쳐 LSS를 추적조사하고 있다. 이 LSS자료는 여러 측면에서 電離방사선(ionizing radiation)이 인간에 미치는 長期효과를 평가하는데 가장 중요한 자료가 되고 있으며, LSS자료에 기초한 분석결과는 방사선 방어에 대한 국제적 기준을 마련하는 근거가 되어 왔다.[1)]

본 논문에서는 우선 LSS자료를 포함한 RERF의 중요한 세 가지 동집단자료를 소개하고,[2)] 지난 1980년부터 개발되기 시작한, 群을 이룬 자료(grouped data)에 포아송 회귀모형을 적용시키는 과정에 대하여 간단히 소개하기로 한다.[3)] 포아송 회귀분석은 실제로 1980년 중반 이후 방사선 위해평가에서 활발하게 적용되었고, 이 목적을 위하여 개발된 통계분석 소프트웨어인 EPICURE[4)]는 상대위해도, 초과 위해도(excess risk) 등에 대한 數理的 모형을 추정하는데 매우 유용하게 사용된다. 최근에 발표된 두 종류의 방사선 위해평가 보고[5)]에서 초과위해도가 포아송 회귀모형에서 구체적으로 어떠한 함수형태로 구성되고 추정되었는지를 살펴본다.

본론을 시작하기 전에 방사선의 단위에 대하여 언급하기로 한다. 방사선에는 알파, 베타, 감마, 엑스선, 중성자선과 같이 여러 종류가 있으며 일종의 에너지 흐름으로 정의할 수 있다. 방사선의 단위는 다음 네 가지가 있다.[6)] 조사선량, 흡수선량, 등가선량, 그리고 유효선량이다. 조사선량은 공간상으로 방출되는 방사선양의 강도이고 단위는 렌트겐(R)이다. 흡수선량은 방사선의 종류와 관계없

1) Shigematsu and Mendelsohn [16], Preston [12].
2) RERF [14].
3) Laird and Olivier [11], Frome [7].
4) Preston *et al.* [13].
5) Thompson *et al.* [17], Cardis *et al.* [4].
6) 편웅범 [1].

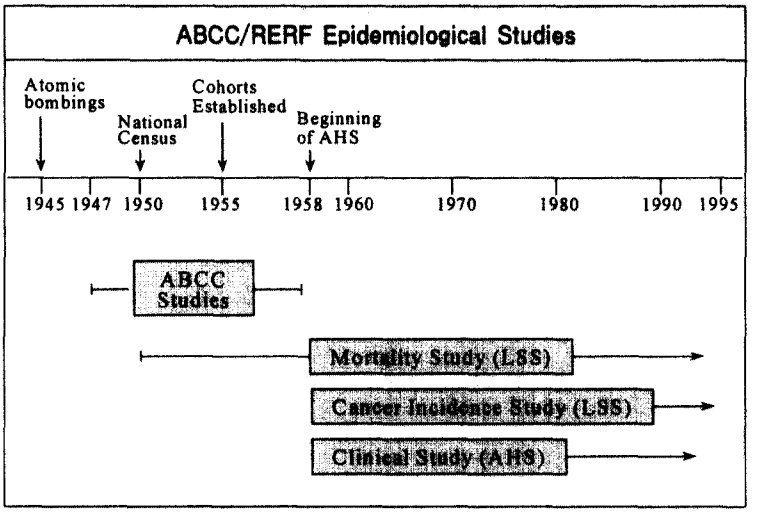
이 모든 방사선이 어떤 물질에 미치는 영향을 파악하기 위하여 만들어진 개념이고 단위는 그레이(gray, Gy)이다. 1gray = 100rad이고 rad는 구측도이다. 방사선의 성질 중 가장 중요한 것은 인체에 미치는 영향이다. 방사선의 종류와 관계없이 방사선의 영향을 파악할 수 있는 단위가 필요하다. 등가선량은 방사선의 종류에 따른 생물학적 효과의 차이를 방사선 가중치로 보정한 양으로 흡수선량에 방사선 가중치를 곱하여 계산하고 단위는 시버트(sievert)이다. 1Sv = 100rem이고 rem은 구측도이다. 방사선에 대한 감수성은 인체 부위에 따라 다르다. 즉, 같은 등가선량에서도 조직의 종류에 따라 발암이나 치사 위험도에 차이가 있다. 따라서 등가선량으로는 피폭자의 위험을 적절하게 나타낼 수 없다. 유효선량은 전신에 대한 방사선 영향을 종합적으로 나타내기 위하여 조직가중치를 등가선량에 곱하여 계산하며, 단위는 등가선량의 단위와 같은 시버트이다. 조직 가중치의 경우 생식선이나 적색골수가 다른 조직에 비해 매우 높은 값을 가진다. 그리고 일반인들에게 가장 필요하다고 생각되는 단위가 유효선량이다. 이외에도 커르마(Kerma)는 물체내에서 방사선에 의하여 방출되는 에너지를 말한다. 그러나, 일반적으로는 피부 표면상의 방사선량의 측도로 간주되며, 그레이(Gy)로 표시된다.

Ⅱ. RERF의 역학연구 자료

RERF는 네 가지 동집단 자료를 추적조사하고 있는데 LSS자료, AHS자료, F1 자료, 그리고 In Utero Study가 그것이다. 이 자료를 순서대로 소개하기로 한다.

1950년 일본 정부는 원폭 생존자들을 대상으로 국세조사를 실시하였는데 이 때 응답한 히로시마, 나가사키의 주민들은 약 19만 5,000명이었다. LSS자료는 이 19만 5,000명으로 구성된 Master Sample에서 일부가 선택되었고, 또 1950년부터 1953년 동안 실시된 특별국세조사에서 원폭투하 당시 市에 살지 않았던 (Not in the Cities : NIC) 것으로 확인된 3만 2,000명을 포함하고 있다. 초기 LSS표본은 9만 9,393명으로 구성되었고, ① 폭탄투하 당시 폭심에서 2,000m 내

〈그림 1〉 RERF에서 LSS 및 관련 조사를 시작한 시점

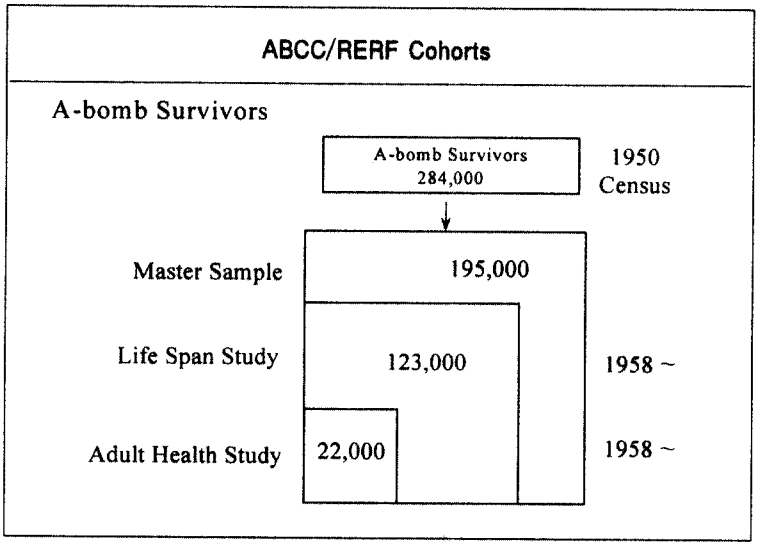


주: Life Span Study (LSS), Adult Health Study (AHS).
자료: Preston, D., Personal Communication.

에 있었고, 히로시마나 나가사키에 주민등록(koseki)을 둔 생존자의 대부분, ② 폭탄투하 당시 폭심에서 2,500~9,999m에 있었고, 히로시마와 나가사키에 주민등록을 둔 생존자들로서 ①의 집단과 연령-성별로 대응된 표본, ③ 폭탄투하 당시 시내에 있지 않았고(NIC) 1950년에 히로시마와 나가사키에 살던 주민들이거나, 혹은 도시에 있었으나 폭심에서 1만 m 밖에 있던 주민들의 연령-성별로 대응된 표본으로 구성되었다.[7] 그러다가 근접 폭로군(proximally exposed group)의 크기를 늘리기 위하여 1970년에 Master Sample에서 2,500m 이내에 있는 사람들을 모두 포함시켰다. 또 1980년대 초 나가사키의 대조군을 확장하기 위하여 2,500~9,999m에 있는 'Proper but Not Selected' 범주에 속하는 8,900명, 'Reserved Part in the Master Sample' 범주에 속하는 2,510명을 추가하여 모두 1만 1,393명이 추가되었다. 1985년에 LSS는 좀더 확장되어 동집단 크기는 12만 321명이 되었고 이 동집단을 LSS-E85라고 부른다. 지난 20여 년 동안 RERF는 LSS자료 분석에 NIC群을 포함시키지 않고 있다. 이는 사회-경제적 지표와 배

7) Thompson *et al.* [17].

〈그림 2〉 원폭 희생자의 Master Sample, LSS 자료와 AHS 자료와의 관계



자료: Preston, D., Personal Communication.

경 발병률면에서 NIC군이 폭로군과 다르기 때문이다.[8]

LSS동집단 개체의 사망 여부는 거주지와 관계없이 주민등록(koseki)을 정기적으로 살펴봄으로써 거의 100% 가깝게 확인할 수 있다. 종양등록(tumor registery)은 히로시마에는 1957년, 나가사키에는 1958년에 각각 설립되었으며, RERF의 전신인 ABCC의 기술적 지원을 받아 각 市의 종양등록제도는 癌例관련 정보를 수집하고 보관하는데 매우 효과적인 방법이 되고 있으며, 비치명적 암의 病因에서 전리방사선의 역할을 평가하는데 매우 중요한 자료가 되고 있다.[9]

AHS는 Adult Health Study를 나타내며, LSS동집단의 부분집합을 취하여 매 2년마다 의료검진을 실시하여 얻은 자료이며, 약 2만 명을 대상으로 1958년에 시작하여 지금에 이르고 있다. LSS자료와 관련 부분 집합, 발생 시점 등은 〈그림 1〉과 〈그림 2〉에서 찾아볼 수 있다.[10]

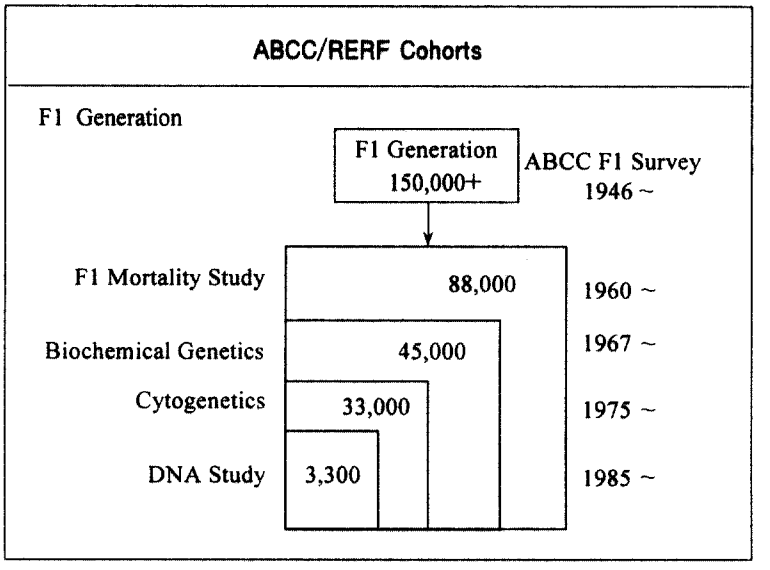
F1자료는 1946년에 시작되었고, 원폭생존자의 첫 번째 후손의 유전학 연구가

8) Thompson *et al.* [17].

9) Kato [10].

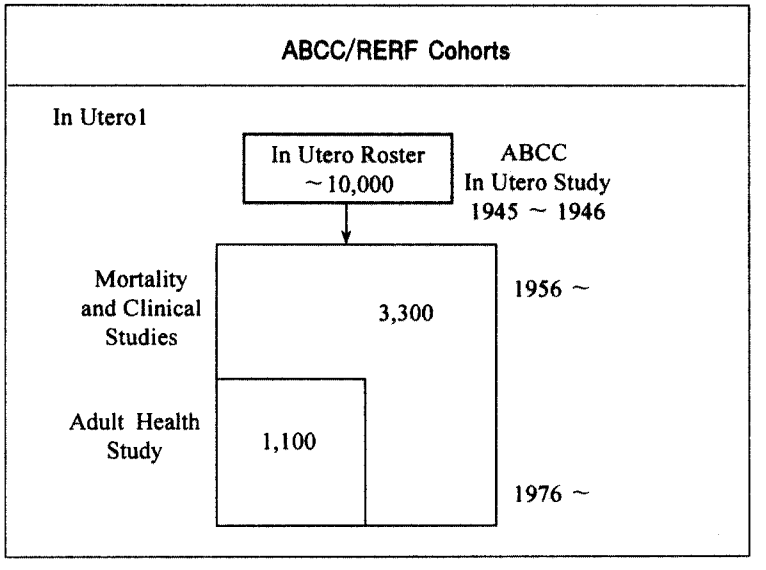
10) Hamilton [8].

〈그림 3〉 RERF의 F1 연구자료와 연구방법의 세분화 과정



자료: Preston, D., Personal Communication.

〈그림 4〉 RERF의 In Utero자료의 수집 시점과 AHS와의 관계



자료: Preston, D., Personal Communication.

주목적이었다. 시간 흐름에 따라 F1자료의 연구방법론이 세분화되는 과정은 〈그림 3〉과 같다. 끝으로 폭탄투하 당시 모태에 있었던 약 1만 명을 대상으로 실시하는 In Utero Study도 계속 추적 조사되고 있으며, 이 중 3,300명에 대하여서는 1956년부터 사망 및 임상 조사가 실시되었고, 이 중 1,100명은 1976년부터 AHS에 편입되었다(〈그림 4〉 참조).

LSS자료는 개개인을 사망할 때까지 추적 조사하는 역학자료이므로 중도 절단된 실패 시간을 관찰하는 것이고, EPICURE User's Guide[11]에 수록되어 있는 TBFCHRT.DAT과 유사하다. LSS자료는 1985년 증보판(LSS-E85)을 기준으로 12만 321명으로 구성되어 있으나, 이 중에서 NIC群을 제외하고, DS86[12]선량이 알려지지 않은 사람도 제외하고, DS86커르마>4Gy도 제외하며, 1958년 1월 1일 이전에 죽거나 암 발병한 사람을 제외하면, 실제 자료분석 대상은 약 8만 명 정도가 된다. 그리고 연구대상을 피폭 당시의 연령(13개 범주), DS86가중 조직선량(16범주), 달력시간기간(7범주), 성별, 도시(2범주)로 나누면 모두 3,640개 칸(cell)을 가진 분할표를 구성하게 된다. Thompson *et al.* [17]은 3,640개 칸 중에 약 2,900개에 대하여 개체가 하나 이상 있음을 보고하고 있다. 각 칸 內에서 年人員(person-year)의 계산과 암 발병 事例는 EPICURE[13]의 DATAB 프로그램을 이용하여 계산할 수 있다.

Ⅲ. 군을 이룬 자료와 포아송 회귀모형

군을 이룬 자료(grouped data)에 포아송 회귀모형을 사용하는 이론적 배경은

11) Preston *et al.* [13] p. 16.

12) DS86는 Dosimetry System 86의 약자로 개개 원폭피해자의 피폭선량을 추정하는 체계이고, 1986년 이전까지 사용한 T65D(Tentative 65 Dose)를 대체한 개념이다. RERF에서는 1987년부터 DS86체계를 사용하고 있다. T65D는 실험자료에 기초한 반면, DS86는 적정한 모형을 설립하고 몬테칼로 방법에 의한 계산에 기초되어 있으며, 당시까지의 과학적 기술이 만들어 낼 수 있는 최선의 방법으로 간주된다(Roesch [15]).

13) Preston *et al.* [13].

1980년 초에 발표된 Holford [9], Aitkin and Clayton [2], Laird and Olivier [11], Frome [7]에 의하여 정립되었다. 이에 앞서 Frome *et al.* [6]은 포아송 회귀모형에서 최대우도추정량(MLE)을 구하는 계산절차가 (포아송분포 가정하에서) 수정된 최소카이제곱 방법과 일치하고, 이 두 가지 계산절차는 선형회귀모형의 일반적 추정절차인 반복 재가중 최소제곱법과 같음을 보고하였다. Aitkin and Clayton [2]은 위험(hazard)모형의 대수선형 모형이 포아송평균의 대수선형 모형과 연관되어 있으므로 중도 절단된 생존분석에 포아송 회귀분석을 적용할 수 있음을 보고하였다.

Holford [9]는 비율(rate : $\frac{\text{사건의 숫자}}{\text{폭로된 전체 기간}}$로 정의됨)에 대한 자료 분석과 생존자료 분석은 결국 같은 우도함수에 기초하는 것을 보이고 있다. 우선 加算자료(count data)와 같은 비율 자료의 경우 많은 수의 사람들이 여러 가지 共變數(나이, 성, 기간, 종양의 단계 등)에 의하여 분류된 分割表를 생각하여 보자. i번째 칸의 비율 λ_i를 다음과 같이 정의할 수 있다.

$$\lambda_i = \frac{x_i}{T_i}$$

단, $x_i = i$번째 칸의 事例數

$T_i = i$번째 칸의 모집단이 관찰된 전체시간

그리고 λ_i에 대하여 다음과 같은 대수선형식을 가정한다.

$$\lambda_i = \exp(\alpha_i + \boldsymbol{z}_i'\boldsymbol{\beta}) \tag{1}$$

단, $\boldsymbol{z}_i$ = 공변수 벡터

$\boldsymbol{\beta}$ = 母數 벡터

식 (1)의 모형은 다른 방법으로도 얻을 수 있다. 가령 i번째 칸에 속하는 h번째 개인($h = 1, \cdots, H_i$)이 실패할 때까지 걸리는 시간 t_{ih}를 관찰한다고 하자. 그리고 중도절단을 나타내는 δ_{ih}를 식 (2)와 같이 정의한다.

$$\delta_{ih} = \begin{cases} 1, & i\text{번째 칸에서 } h\text{번째 개인이 실패하면} \\ 0, & \text{그렇지 않으면} \end{cases} \tag{2}$$

통상적으로 t_{ih}는 모수 λ_i를 갖는 지수분포, $\mathcal{E}(\lambda_i)$[14]를 가정한다. 이 때 $\{t_{ih}, \delta_{ih}\}$가 관찰되면 λ_i의 우도함수 l_e는 식 (3)과 같이 얻어진다.

$$l_e(\lambda_i;\ t_{ih}) = \prod_h \lambda_i^{\delta_{ih}} e^{-\lambda_i t_{ih}} = \lambda_i^{x_i} e^{-\lambda_i T_i}, \tag{3}$$
$$\text{단, } T_i = \sum_h t_{ih}$$

식 (3)에서 λ_i에 대한 충분 통계량은 $(T_i,\ x_i)$임을 알 수 있다. Holford [9]는 다음의 정리를 증명함으로써 포아송분포, 지수분포의 모수 β의 최대우도추정량이 다항분포와 관련되어 있음을 보이고 있다.

Holford [9]

$\lambda_i = \exp(\alpha + \boldsymbol{z}_i' \boldsymbol{\beta})$, 단, $\boldsymbol{z}_i = (z_{i1}, \cdots, z_{ir})$, $\boldsymbol{\beta}' = (\beta_1, \cdots, \beta_r)$이라고 하자. β에 대한 최대우도 추정량은 다음 세 가지 경우 모두 동일하다.

① $x_i \sim \mathcal{P}(\lambda_i T_i)$[15]

② $t_{ih} \sim \mathcal{E}(\lambda_i)$, $i = 1, \cdots, I$, $h = 1, \cdots, H_i$

③ $\boldsymbol{x}_{I\times 1} \sim M(x_+,\ \boldsymbol{P}_{I\times 1})$[16]

단, $\boldsymbol{x}_{I\times 1} = (x_1, \cdots, x_I)$, $\boldsymbol{P}_{I\times 1} = (P_1, \cdots, P_I)$, $p_i = \dfrac{T_i \exp(\alpha + \boldsymbol{z}_i' \boldsymbol{\beta})}{\sum_i T_i \exp(\alpha + \boldsymbol{z}_i' \boldsymbol{\beta})}$,

---

14) $X \sim \mathcal{E}(\lambda)$라 함은 X의 확률밀도함수가 $f_x(x) = \lambda e^{-\lambda x}$, $\lambda > 0$, $x > 0$임을 의미한다.

15) $X \sim \mathcal{P}(\lambda)$라 함은 포아송분포를 말하며 X의 확률질량함수가 $f_x(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$, $\lambda > 0$, $x = 0, 1, \cdots$이 됨을 의미한다.

16) $\boldsymbol{x}_{r\times 1} \sim M(n,\ \boldsymbol{\pi}_{r\times 1})$은 다항분포를 의미하며, 확률질량함수가 다음과 같음을 의미한다. $f_x(\boldsymbol{x}) = \binom{n}{x_1, \cdots, x_r} \prod_{i=1}^{r} \pi_i^{x_i}$. 단, $\boldsymbol{\pi}_{r\times 1} = (\pi_1, \cdots, \pi_r)$, $\pi_i \geq 0$, $\sum \pi_i = 1$, $\boldsymbol{x}_{r\times 1} = (x_1, \cdots, x_r)$, $\sum x_i = n$.

$$i=1,\cdots,I,\ \sum p_i=1,\ x_+=\sum x_i.$$

증명 : ②의 경우 대수우도함수 l^*의 커널 부분은 식 (4)와 같다.

$$l^* = x_+\alpha + \sum_i x_i z_i'\beta - \sum_i \exp(\alpha + z_i'\beta)T_i \qquad (4)$$

단, $T_i=\sum_h t_{ih}$. 식 (4)를 α와 β 각각에 대하여 도함수를 구하고 0으로 놓으면 식 (5)를 얻게 되는데, 결국 ①, ②의 경우 β에 대한 최대우도 추정법도 식 (5)의 解로 얻어지게 된다.

$$\sum_i x_i z_{ij} - \frac{x_+\sum_i z_{ij}T_i\exp(z_i'\beta)}{\sum_i T_i\exp(z_i'\beta)} = 0, \qquad j=1,\cdots,r \qquad (5)$$

이 정리는 포아송모형이나 지수분포모형의 모수추정에 다항분포빈도 x_i에 기초된 추정 알고리즘을 이용할 수 있음을 말하고 있다. 구체적으로는 분할표 분석에 추정 알고리즘으로 많이 사용되는 반복비례적합(Iterative Proportional Fitting : IPF) 알고리즘이나 Newton-Raphson 방법을 사용할 수 있다.

Holford [9]는 이어서 각 구간별로 常數 위험(constant hazard)을 가정하는 경우 Cox의 비례위험모형도 기본적으로는 식 (5)와 같은 우도함수를 취하고, 결국 생존자료 분석에도 분할표 분석에 사용되는 대수선형 모형을 적용할 수 있음을 보이고 있다. 즉, Cox모형은 $\lambda(t;\ z)=\exp(\alpha^*(t)+z'\beta)$와 같은 대수-선형 위험함수를 가정하고 있으며, 이 위험함수는 추적 조사기간을 K개의 구간$(\tau_1,\ \tau_2]$, $(\tau_2,\ \tau_3]$, $\cdots$, (τ_K,∞)으로 나누고, 각 구간마다 常數위험을 가정함으로써 $\alpha^*(t)\approx\alpha_k$, $\tau_k<t\le\tau_{k+1}$로 근사할 수 있다. 이렇게 근사된 모형의 대수대수우도함수의 커널 부분은 l_s는 식 (6)과 같다.

$$l_s = \sum_k \alpha_k x_{+k} + \sum_{i,k} x_{ik} z_i'\beta - \sum_{i,k}\exp(\alpha_k + z_i'\beta)T_{ik} \qquad (6)$$

단, x_{ik}와 T_{ik}는 각각 k번째 구간, i번째 칸에서 관찰된 실패 횟수와 추적 조사된 전체길이를 나타낸다. 식 (6)과 식 (4)는 기본적으로 같은 형태이므로 생존분석에 대수선형 모형을 적용할 수 있음을 알 수 있다.

Holford [9]의 결과에 기초하여 Laird and Olivier [11]는 동집단 연구자료를 분석함에 있어 群을 이룬 자료를 사용하는 이론적 근거를 명확하게 제시하고 있다. Laird and Olivier [11]는 다음 두 가지 조건 (a), (b)하에서 포아송분포에서 생성된 분할표의 각 칸의 평균에 대한 대수선형 모형이 생존자료의 대수선형 위험모형과 같음을 보이고 있다.

(a) 생존분포가 조각별(piec6ye wise) 지수분포를 이룬다.
(b) 공변수가 모두 범주형이다.

Cox의 비례위험모형은 식 (7)과 같이 나타낼 수 있다.

$$\lambda(t, z) = \lambda_0(t)\exp(z'\beta) \tag{7}$$

단, $\lambda_0(t)$는 기준 위험함수로서 모수 형태를 취할 수도 있으며, 식별되지 않은 상태로 남겨 놓을 수도 있다. 그리고 $z' = (z_1, \cdots, z_r)$은 모두 범주형 공변수이다. 실패시간 t가 조각별 지수분포를 따른다는 것은, 시간축이 K개의 서로 배반이고 전체를 이루는(exhaustive) 구간, $\Omega_1, \cdots, \Omega_K$로 구성되어 있다고 하고, 각 구간마다 常數위험함수를 가정하면 된다. 그러면, 식 (7)은 $t \in \Omega_k$에 대하여 식 (8)과 같다.

$$\lambda(t, z) = \lambda_k\exp(z'\beta), \qquad t \in \Omega_k,\ k = 1, \cdots, K \tag{8}$$

단, λ_k는 Ω_k 구간 내에서의 常數위험을 나타낸다.

$z = (z_1, \cdots, z_r)$은 모두 범주형 공변수로 가정하였다. 공변수 z_j가 취할 수 있는 가능한 수준의 개수를 l_j라고 할 때, 시간축을 포함하여 (t, z)가 취할 수

있는 가능한 수준의 개수는 $K\times l_1\times\cdots\times l_r$개가 된다. (t, z)가 취할 수 있는 한 수준을 指數$(k_0, k_1, \cdots, k_r)$로 나타내기로 한다. 이 때 (t, z)로써 분할표를 구성하면서 Ω_{k_0}와 $(k_1, \cdots, k_r)$로 분류된 칸에 해당되는 위험률을 $\theta_{k_0 k_1\cdots k_r}$이라 하자. 그러면 식 (8)은 식 (9)와 같이 표기된다.

$$\lambda(t, z)=\theta_{k_0 k_1\cdots k_r},\qquad t\in\Omega_{k_0},\quad z=(k_1, \cdots, k_r) \tag{9}$$

식 (8)의 우변에 대수선형 모형을 취하고, Bishop, Fienberg and Holland [3]의 u항으로 표기하며 식 (10)과 같은 포화모형을 구성한다.

$$\begin{aligned}\theta_{k_0 k_1\cdots k_r} &= u + u_{0(k_0)} + u_{1(k_1)} + \cdots + u_{r(k_r)}\\ &\quad + u_{12(k_1 k_2)} + \cdots + u_{12\cdots r(k_1\cdots k_r)}\end{aligned} \tag{10}$$

한편, 정리 1은 식 (10)의 좌변에 전체폭로시간 $T_{k_0 k_1\cdots k_r}$을 곱해 주면 결과되는 양이 바로 $(k_0, k_1, \cdots, k_r)$칸의 포아송 평균이 됨을 말해 주고 있다.

Laird and Olivier [11]가 두 번째로 밝힌 내용은 Holford [9]가 이미 언급한 내용으로 식 (4)와 식 (6)이 비례하므로 생존분석에 대수선형 모형을 사용할 수 있다는 내용과, 정리 1의 결과를 종합하여 조각별 지수분포 생존자료에 기초한 우도함수와 포아송분포로부터 생성된 분할표의 우도함수는 같음을 보이고 있다. 우선 지수분포를 이루고 공변수가 없는 생존자료의 경우를 살펴보기로 하자. 확률밀도함수(pdf) $f(t)=\theta e^{-\theta t}$, 위험함수 $\lambda(t)=\theta$, 생존함수 $S(t)=1-F(t)=e^{-\theta t}$(단, $F(t)=\int_0^t f(x)dx$)가 된다. $F(t)$에서 추출된 H개의 독립적 관찰치가 우측 중도 절단된 경우 우도함수는 식 (11)과 같다.

$$l(\theta)=\prod_{h=1}^{H}\theta^{\delta_h}e^{-\theta t_h}=\theta^x e^{-\theta T} \tag{11}$$

단, $x=\sum_{h=1}^{H}\delta_h$, $T=\sum_{h=1}^{H}t_h$, δ_h는 중도절단을 나타내는 표시자로 식 (2)와 유사하게 정의된다. 식 (11)에서 x는 관찰된 사망건수나 실패횟수이고, T는 전체

〈그림 5〉 LSS자료의 분할표 구성

k번째 부분 구간	
칸 \ 선량	$1 \cdots j \cdots J$
1 ⋮ i ⋮ k	(i, j, k)

폭로를 나타낸다. 즉, T는 그 표본 전체가 실패 위험에 놓여 있는 시간을 나타낸다. 식 (11)에서 θ의 최대우도 추정량은 x/T임을 쉽게 알 수 있다.

이제 T가 주어졌고 $E(x|T)=T\theta$일 경우 x가 포아송 분포를 이룬다고 하면 포아송 분포하에서 우도함수는 식 (12)와 같다.

$$l_p(\theta) \propto \frac{(T\theta)^x e^{-T\theta}}{x!} \propto l(\theta) \tag{12}$$

식 (12)를 통하여 포아송 분포의 평균발생을 θ의 최대우도추정량을 구하여도 x/T로 얻어진다. 따라서 위의 두 가지 다른 표본 분포하에서 얻은 우도함수는 서로 비례하므로 θ의 최대우도 추정량을 구하고, 그 추정량의 점근적 성질을 밝히는 데에는 두 우도함수를 서로 바꾸어서 사용하여도 무방하다. 동결과는 조각별 지수분포의 형태로 쉽게 확장되고, 범주형 설명변수가 있는 경우로도 확장이 되나, 표기가 복잡해지므로 본 논문에서는 구체적 기술을 생략하기로 한다. Laird and Olivier [11]에서 자세한 확장과정을 볼 수 있다.

위와 같은 논거에 기초하여 LSS생존자료를 분석하는 접근으로 나이, 시간, 선량 등과 같은 연속형 변수를 범주화하여 범주형 공변수로 간주하면서, 동시에 군을 이룬 자료를 구성할 수 있다. LSS자료를 〈그림 5〉의 분할표처럼 구성하면 기본적 자료는 우선 (i,j,k)칸의 **事例數** x_{ijk}, 연인원 PY_{ijk}, 그리고 공변수 z_{ijk}가 된다. z_{ijk}는 도시, 성별, 피폭시의 나이에 따라 분류된 칸 중 i번째 칸,

j번째 선량군, 그리고 추적조사 기간의 k번째 부분구간과 연관된 공변수 벡터를 나타낸다($i=1, \cdots, I,\ j=1, \cdots, J,\ k=1, \cdots, K$). 그리고 λ_{ijk}는 (i,j,k) 칸의 단위 시간당 위험을 나타낸다. λ_{ijk}는 경우에 따라 $\lambda(t;\ \boldsymbol{z},\ d)$, 즉 선량 d에 폭로된 개인이 $\boldsymbol{z}$의 공변수를 가질 때 시점 t에서의 위험함수로 일반화될 수도 있다.

앞에서도 언급하였듯이 Thompson *et al.* [17]은 피폭 당시의 연령을 13개 범주, 선량을 10개 범주, 달력시간 기간을 7개 범주로 나누어서 성별과 도시별 범주를 포함하면 모두 3,640개의 칸을 가진 분할표를 구성하였고, 이 중 2,900여 개 칸에 대하여 개체가 1개 이상 있음을 보고하였다. 연인원 PY_{ijk}와 단위시간당 위험률 λ_{ijk}를 常數로 간주할 수 있다면 Holford [9]와 Laird and Olivier [11]의 결과를 적용할 수 있다. 즉, 조각별 지수분포에 기초한 생존자료의 우도함수와 $x_{ijk} \sim \mathcal{P}(PY_{ijk}\lambda_{ijk}),\ i=1, \cdots, I,\ j=1, \cdots, J,\ k=1, \cdots, K$에 기초한 우도함수는 같게 되어 결국 (군을 이룬) LSS자료의 표본분포로서 포아송분포를 사용할 수 있게 된다.

방사선 역학 연구를 포함한 역학연구에서 일반적으로 위험함수에 대한 다음 두 가지 단순한 모형을 고려해 볼 수 있다.

$$\lambda(t;\ \boldsymbol{z}, d) = \lambda_0(t;\ \boldsymbol{z})\rho_R(d) \tag{13}$$

단, $\lambda_0(t;\ \boldsymbol{z})$는 선량에 노출되지 않았을 때의 율(rate), 즉 배경률을 나타내며 $\rho_R(d)$는 상대 위해도(relative risk)함수이다. 한편, 선량 d에 대한 추가 위해도가 시간이나 다른 공변수에 의존하지 않는 순수한 加法性 초과 위해도(additive excess risk)모형을 식 (14)와 같이 고려할 수도 있다.

$$\lambda(t;\ \boldsymbol{z}, d) = \lambda(t;\ \boldsymbol{z}) + \rho_A(d) \tag{14}$$

단, $\rho_A(d)$는 절대 위해도(absolute risk)함수이다. 식 (13)~식 (14)의 모형에 공변수에 의한 선량효과 수정 요인을 감안한다면 식 (15)~식 (16)과 같이 더 현

실적인 모형을 세울 수도 있다.

$$\lambda(t;\ \boldsymbol{z}, d) = \lambda_0(t;\ \boldsymbol{z})\rho_R(d;\ t,\ \boldsymbol{z}) \tag{15}$$

$$\lambda(t;\ \boldsymbol{z}, d) = \lambda_0(t;\ \boldsymbol{z}) + \rho_A(d;\ t,\ \boldsymbol{z}) \tag{16}$$

식 (15)~식 (16)의 경우 상대 위해도 함수와 절대 위해도 함수는 구별이 되지 않는다. 실제로 방사선 역학연구에서 식 (13)~식 (14)는 너무 단순한 형태이고, 식 (15)~식 (16)은 너무 복잡한 형태가 되므로 그 중간 정도에서 유용한 모형을 찾는 것이 일반적이다.

LSS자료는 중도 절단된 실패시간의 자료이므로 1972년에 발표된 Cox의 비례위험모형에 따라 식 (17)과 같이 기술할 수 있다.

$$\lambda_0(t)\exp(\boldsymbol{z}'\beta) \tag{17}$$

그러나, Cox의 모형은 지수함수로써 선량-반응을 나타내고 있는데, 사실 선량-반응함수로서 지수함수에 대한 이론적·경험적 논거가 충분하지 않다. 따라서 모형 해석의 편의상 LSS자료 분석에는 식 (18)과 같이 초과 상대 위해도(ERR) 모형이 유용하다.[17]

$$\lambda_0(t;\ \boldsymbol{z})[1+\rho(d,\ \beta_1)\varepsilon(\boldsymbol{z},\ \beta_2)] \tag{18}$$

단, $\rho(d,\ \beta_1)$은 선량-반응곡선을 기술하는 함수이고, $\varepsilon(\boldsymbol{z},\ \beta_2)$는 초과 상대 위해도가 시간이나 그 외 다른 요인에 따라 어떻게 변하는가를 나타내는 효과 수정요인이다.

17) Preston [12].

Ⅳ. 포아송 회귀모형의 적용 예

Frome [6]은 Doll and Hill [5]이 영국 의사들을 대상으로 추적 조사한 흡연과 암 발병자료에 포아송 회귀모형을 적용하였다. 추적 조사된 자료를 흡연 기간에 따라 9개 범주로 나누었고, 매일 흡연량을 7개 범주로 나누어서 각 칸마다 연인원(PY_i)과 폐암 발병건수(X_i)를 조사하였다.[18] 조각별 지수 생존함수를 가정하면 X_i는 식 (19)와 같은 포아송분포를 갖는다.

$$X_i \sim \mathcal{P}(PY_i \cdot \lambda(l_i, d_i)) \tag{19}$$

단, $\lambda(l_i, d_i)$는 흡연 기간 l_i와 흡연량 d_i에 대한 위험률을 나타낸다. $\lambda(l_i, d_i)$에 대한 모형으로는 식 (20)과 같은 가법 초과 위해도 모형을 설정하였는데 이는 동물실험과 역학 연구로부터 도출된 이론적 모형이다.

$$\lambda(l_i, d_i) = (\gamma + \alpha d_i^{\theta}) l_i^{\beta} = \lambda_0(l_i) + \rho_A(d_i) \tag{20}$$

단, γ, α, θ, β는 모수로서 γ는 배경(비흡연자) 발생률을 나타내고, αd^{θ}는 흡연이 폐암 사망률에 미치는 효과를 나타낸다. Frome *et al.* [6]은 반복 재가중 최소제곱법을 이용하여 α, β, γ, θ를 각각 추정하였고, $\hat{\beta} \cong 4.5 (SE = 0.34)$, $\hat{\theta} \cong 1.20 (SE = 0.40)$을 보고하고 있다.

Cardis *et al.* [4]는 미국, 영국, 캐나다 3개국의 원자력산업 종사자 9만 5,673명의 역학조사 자료를 종합하여 사망과 암 발병에 대한 방사선 위해도 평가 결과를 보고하고 있다. 즉, 상기 3개국 중 어느 한 나라의 핵 산업체에 6개월 이상 근무하였고 전리 방사선에 대한 외부 폭로를 감시 받아온 9만 5,673명의 근로자 중 사망자의 사망 원인을 분석하여, 감마 방사선이 낮은 농도로 오랫동안

18) Frome [7] 〈표 1〉을 참조.

폭로됨으로써 얼마만큼의 발암 효과를 유발하는지를 평가하고자 하였다. 그 결과 모두 7개 핵산업체에서 자료가 수집되었고, 연인원(PY) 총 212만 4,526명이 위해도에 노출되어 있었고(at risk), 1만 5,825명이 사망하였으며, 그 중 3,976명은 암으로 사망한 것을 알았다. 조사 자료는 〈그림 5〉와 같은 분할표로 구성되었는데, 시버트 단위의 누적 선량은 11개의 범주, 그 외 교락 효과로서 성별, 실제 연령(5년 구간), 달력기간(5년 범주), 그리고 사회경제적 지표와 연구 모집단(7개 산업체)을 고려하였다. 11개 선량군에 따른 사망의 추세를 검정하기 위하여 식 (21)과 같은 선형 상대위험모형을 구성하였다.

$$\lambda_{ijk} = \lambda_{(0)ik}[1 + \beta d_{ij}] \tag{21}$$

단, $\lambda_{(0)ik}$는 (i,k)칸에서 방사선에 노출되지 않았을 때의 율, 즉 배경률을 나타내고, d_{ij}는 i번째 칸, j번째 선량群의 가중 평균된 선량(단위 : Sievert)을 나타낸다. 식 (19)에서 β는 단위 시버트에 대한 초과 상대 위해도(ERR)를 나타낸다. 자료의 분석은 EPICURE[19]를 사용하였다. 모든 암으로 인한 혹은 모든 원인으로 인한 사망과 방사선량과의 유의적 관계는 찾을 수 없었다. 그러나 원폭 생존자료와 高線量에 노출된 다른 모집단자료의 분석에서 일치적으로, 그리고 강한 증거를 가지고, 線量과 연관되었던 chronic lymphocytic leukemia를 제외한 leukemia로 의한 사망은 누적 선량과 유의적인 연관을 나타내었다(p값 = 0.046, 119사망). 즉, 이 경우 $\hat{\beta}=2.18$로 계산되었으며 90% 신뢰구간은 0.1～5.7이었다.

Thompson *et al.* [17]은 1958년부터 1987년 동안 일본 원폭 생존자 추적 조사 자료인 LSS-E85에 기초하여 고형암에 대한 전리 방사선의 위해도 평가 결과를 보고하고 있다. 연구 동집단의 크기는 7만 9,972명이었고, 고형암 사례는 8,613건을 기록하고 있다. 군을 이룬 형태로 만들기 위하여 〈표 1〉과 같이 연속형 변수들도 범주화하였다.

이렇게 하여 분류된 각 칸마다 연인원(PY_{ijk})과 사례수(x_{ijk})를 계산하는 것 이외에 공변수로서 피폭 당시 연령의 평균, 평균 실제 연령(mean attained age),

19) Preston *et al.* [13].

〈표 1〉 분할표를 구성하기 위한 변수의 범주화

피폭 당시의 나이(age at exposure)	e	13
조직선량(organ dose)	d	10
달력시간기간(carlendar time period)	t	7
성 별(sex)	s	9
도 시(city)	c	2
전체범주 개수		3,640

분석 대상 조직의 감마선과 중성자 선량의 평균, 그리고 가중 조직 선량의 평균이 계산되었다. 이 평균들은 연인원으로 가중 평균하였고, EPICURE의 DATAB 프로그램을 이용하여 계산하였다. 고려된 각각의 조직과 부위에 대하여 표준적인 분석이 실시되었다. 즉, 식 (22)와 같은 일반적인 초과 상대 위해도모형을 사용하여 추정과 가설 검정을 실시하였다.

$$\lambda_0(c, s, a, b)[1 + \rho(d)\varepsilon(c, s, e, t, a)] \tag{22}$$

단, $\lambda_0(\cdot)$는 배경률, 즉 선량이 0인 사람들의 발병률을 나타내는 모수적 모형이고, $\rho(\cdot)$는 선량-반응함수, 그리고 $\varepsilon(\cdot)$는 선량-효과 수정을 기술하고 있다. 식 (22)에서 배경률에 대한 표준 모형으로는 식 (23)이 사용되었다.

$$\lambda_0(c, s, a, b) = \exp[\beta_{1s} + \beta_{2c} + \beta_{3b} + \beta_{4s}\log(a) + \beta_{5s}\log^2(a)] \tag{23}$$

그리고 갑상선암과 피부암의 경우는 식 (21)을 확장하여 AHS구성원 여부를 모형에 포함시켜 β_6AHS가 식 (23)에 추가되며, AHS는 AHS 구성원 여부를 나타내는 二進型 변수를 나타낸다. 그리고 선량-반응함수로서는 $\rho(d) = \gamma_1 d$가 표준모형으로 사용되었고, d는 시버트로 표시된 RBE_{10}[20]가중 선량을 나타낸

20) RBE_{10}은 상대생물학적효과(Relative Biological Effectiveness)의 약자이고 하첨자 10은 감마선보다 중성자가 10배의 효과를 미친다는 의미이다.

다. 그리고 선량-반응곡선의 선형성에 대한 검정을 식 (24)와 같은 선형-2차 선량-반응모형에 기초하였다.

$$\rho(d) = \gamma_1 d + \gamma_2 d^2 \tag{24}$$

그리고 피부암의 경우는 식 (25)와 같은 선형-스플라인 모형을 사용하였다.

$$\rho(d) = \gamma_1 d + \gamma_2 (d-1)_+ \tag{25}$$

단, $(x)_+ = \max(x, 0)$를 나타낸다.

표준적인 효과수정에 대한 검정은 식 (26)과 같은 선량-효과 수정 함수에 기초하였다.

$$\varepsilon(\boldsymbol{z}) = \exp(\boldsymbol{z}'\beta) \tag{26}$$

단, $\boldsymbol{z}$는 共變數 벡터로서 $\boldsymbol{z} = (s, e, c, \log(\text{time since exposure}))$이다.

결국 식 (22)~식 (26)의 모형을 사용하고, 여러 단계의 가설검정 절차를 거쳐 효과 수정요인으로 성별(s), 피폭시 연령(e)을 찾았다. 따라서, 각각의 부위나 조직에 대한 잉여 상대 위험도를 식 (27)과 같이 성별과 피폭시 연령별로 추정하였다.

$$1 + \beta_{es} d \tag{27}$$

각 조직과 부위에 따라 식 (22), 식 (23), 식 (27)에 기초하여 β_{es}를 성별과 연령별로 추정한 결과는 Thompson *et al.* [17]이 자세히 보고하고 있으며, 갑상선암의 경우 남자의 경우 피폭시 연령이 0~9년이면 $\widehat{\beta_{es}} = 9.39$로 얻어지고, 10~19년이면 $\widehat{\beta_{es}} = 2.60$, 그리고 그 이후에는 모두 음의 값으로 얻어진다. 이러한 추세에는 여성의 경우도 유사하다. 따라서, 갑상선암은 폭로시 연령이 선량-반응에 중요한 요인이 됨을 알 수 있다.

V. 맺는 말

본 논문에서는 지난 50년 가까이 원폭생존자들을 대상으로 추적 조사해 온 일본 RERF의 LSS자료를 중심으로 방사선 역학연구에 포아송 회귀모형이 어떻게 적용되고 있으며, LSS자료를 群을 이룬 자료를 구성하여 포아송 회귀분석을 적용하는 이론적 근거를 살펴보았다. 이 포아송회귀분석은 EPICURE프로그램과 함께 LSS자료를 분석하는 중요한 도구가 되었다. LSS자료에 근거한 방사선 위해도 추정치는 세계적으로 방사선 보호 기준을 마련하는데 중요한 자료원이 되어 왔다.

초기 LSS자료 분석은 고형암과 백혈병에 대한 피폭효과를 검색하고, 방사선 폭로와 연관된 위해도를 추정하는데 주안점을 두었다. 그러나, 이와 관련된 정보가 증가하면서 선량-반응에 대한 기본적 질문 이외에 다른 문제에 관심을 가지게 되었다. 관심대상은 구체적인 선량-반응의 형태, 성별효과, 피폭시 연령효과, 그리고 초과 위해도에 대한 연령-시간 유형 등을 분석하는 일들이다. 이와 같은 새로운 문제들을 해결하는 데에는 표준적인 생존분석방법에 대한 확장과 대안이 요구된다. 이러한 맥락에서 최근에 Preston [12]은 Cox의 비례위험모형의 제약점과 해석상의 어려움을 들어 대안으로 초과 상대 위해도모형의 사용을 제안하고 있다. 또 개개 관찰치에 기초한 준 우도방법과 군을 이룬 자료에 기초한 포아송 회귀분석방법을 비교하면서 후자가 기준 위험률을 모형화하는데 더 유연한 방법임을 지적하고 있다.

▣ 참 고 문 헌 ▣

1. 편응범, "알면 유용한 방사선의 단위", 『식품의약품정보』, 1998, 12, 겨울호(식품의약품안전청 소식지).

2. Aitkin, M. and M. Clayton, "The Fitting of Exponential, Weibull, and Extreme

Value Distributions to Complex Censored Survival Data Using GLIM," *Applied Statistics*, 29(2), 1980, pp. 156～163.

3. Bishop, Y. M. M., Fienberg, S. E. and P. W. Holland, *Discrete Multivariate Analysis*, Cambridge: MIT Press, 1975.
4. Cardis, E., Gilbert, E. S., Carpenter, L., Howe, G., Kato, I., Armstrong, B. K., Beral, V., Cowper G., Douglas A., Fix J., Kaldor J., Lave C., Slamon L., Smith P. G., Voelz G. L. and L. D. Wiggs, "Effects of Low Doses and Low Dose Rates of External Ionizing Radiation: Cancer Mortality among Nuclear Industry Workers in Three Countries," *Radiation Research*, 142, 1995, pp. 117～132.
5. Doll, R. and A. B. Hill, "Morality of British Doctors in Relation to Smoking : Observations on Coronary Thrombosis," in *Epideomological Study of Cancer and Other Disease*, W. Haenszel (ed.), National Cancer Institute Monograph 19, Washington D.C. : US Government Printing Office, pp. 205～268.
6. Frome, E. L., Kutner, M. H. and J. J. Beauchamp, "Regresson Analysis of Poisson-distributed Data," *Journal of the American Statistical Association*, 68(344), 1973, pp. 935～940.
7. _____, "The Analysis of Rates using Poisson Regression Models," *Biometrics* 39, 1983, pp. 665～674.
8. Hamilton, H. B., "Data Resources for the Major Cohort Studies: The Adult Health Study," in R. L., Prentice and D. J. Thompson (ed.), *Atomic Bomb Survivor Data: Utilization and Analysis*, Philadelphia: SIAM, pp. 18～32.
9. Holford, T. R., "The Analysis of Rates and Survivorship using Log-linear Models," *Biometrics*, 36, 1980, pp. 299～305.
10. Kato, H., "Data Resources for Life Span Study," in R. L., Prentice and D. J. Thompson (ed.), *Atomic Bomb Survivor Data: Utilization and Analysis*, Philadelphia: SIAM, 1983, pp. 3～17.
11. Laird, N. and D. Olivier, "Covariance Analysis of Censored Survival Data Using Log-linear Analysis Technique," *Journal of the American Statistical Association*, 76, 1981, pp. 231～240.
12. Preston, D. L., "Beyond Dose-response: Describing Long-term Health Effects of Radiation Exposure," *Bulletin of the International Statistical Institute*, 53rd

Session Proceedings, Book 1, 2001, pp. 301～304.

13. _____, Lubin, J. H., Pierce, D. A. and M. E. McConney, *Epicure, User's Guide*, 1998, Seattle: Hirosoft International Corporation.
14. Radiation Effects Research Foundation, Radiation Effects Research Foundation, A Brief Description, Hiroshima: RERF, 1993.
15. Roesch, W. C. (ed.), *US-Japan Joint Reassessment of Atombic Bomb Radiation Dosimetry in Hiroshima and Nagasaki*, Vol. I, Vol. II, 1987, Hiroshima: RERF.
16. Shigematsu, I. and M. L. Mendelsohn, "The Readiation Effects Research Foundation of Hiroshima and Nagasaki: Past, Present and Future," *Journal of the American Medical Association*, 274(5), 1995, pp. 425～426.
17. Thompson, D. E., Mabuchi, K., Ron, E., Soda, M., Tokunaga, M., Ochikubo, S., Sugimoto, S., Ikeda, T., Terasaki, M., Izumi, S. and D. Preston, "Cancer Incidence in Atomic Bomb Survivors, Part II: Solid Tumors, 1958～1987," *Radiation Research*, 137, 1994, pp. S17～S67.