

# 산업재해 사례인자의 범주형 분석

지경택 · 송영호\*<sup>†</sup> · 정국삼\*\*

(주)한라 · \*충북대학교 대학원 안전공학과 · \*\*충북대학교 안전공학과  
(2001. 9. 12. 접수 / 2001. 10. 30. 채택)

## Categorical Analysis for the Factors of Industrial Accident Cases

Kyung-Tek Jhee · Young-Ho Song\*<sup>†</sup> · Kook-Sam Chung\*\*

HALLA · \*Graduate School, Department of Safety Engineering, Chungbuk National University

\*\*Department of Safety Engineering, Chungbuk National University

(Received September 12, 2001 / Accepted October 30, 2001)

**Abstract :** This study aimed to search for the fundamental accident causes using a categorical analysis, a kind of statistical methods. As the analysis methods, correlation analysis, independence test and logistic regression analysis were used. And the SPSS package, a general - purpose mathematical library, was used to obtain statistical characteristics. As the result of this study, the accident causes associated with factor of 'lost working days' were factors such as 'employed periods', 'sex', 'types of accident', 'month'. In case of applying independence test method, the most important cause was the factor of 'month'. In case that logistic regression analysis method was applied, the cause contributed to the increase of occurrence probability of major injury were factors such as 'suffocation & intoxication', 'temporary construction structure', 'less than 6 month'. On the basis of these results, the plan for accident prevention and the proper investment for accident prevention expenditure could be carried out in each workshop.

**Key Words :** categorical analysis, accident cause, correlation analysis, logistic regression analysis

### 1. 서 론

국내의 현행 산업재해사고의 통계는 산업재해보험 가입 사업장의 업무상 재해로 인정된 산재자만을 대상으로 하여, 산업재해조사 조사표에 단순 구분된 재해 형식과 원인 요소별로 분류하여 백분율로 산출하고 있어 다분히 피상적이다. 따라서, 산업재해 요인의 과학적인 분석을 기하지 못하고 있어 근본적인 재해 발생의 원인 규명에 한계가 있는 실정이다.<sup>1)</sup>

그러므로, 본 연구에서는 국내의 충북지역에서 발생한 산업재해를 대상으로 통계적 기법인 범주형 분석을 이용하여 재해발생에 있어서의 요인별 상관성, 요인간의 영향성 및 중요요인을 파악하여 보다 효과적인 산업재해 예방의 대책과 안전관리 활동 방안을 수립하는데 필요한 기초자료로 활용하고자

하였다.

본 연구에 사용된 통계학적 기법을 설명하면 다음과 같다.

첫째, 산업재해 사례의 data를 분석하여 근로손실 일수 및 재해형태와 기타 변수들간의 상관분석을 통하여 상관도를 도출하였다.

둘째, 명목 및 서열척도의 범주형 변수(categorical variables)들을 분석하여 변수간의 교차분석을 통하여 독립성과 관련성을 분석하였다.

셋째, 종속변수와 독립변수들간의 인과관계와 독립변수들의 변동에 따라 종속변수의 변화를 예측하기 위하여, logistic regression analysis을 적용하였다.

위와 같은 분석을 위하여 사용된 산업재해 사례 data는 모두 2,133건으로, 충북지역의 사업장에서 1997년 한 해 동안 수집한 data로 하였고, 분석에 사용된 통계 package는 SPSS(Statistical Package for the Social Science ver. 8.0)을 사용하였다. 그리고, 통계분석을 용이하게 하기 위하여 data의 coding을 변경하거나 변환을 시도하였다.

<sup>†</sup>To whom correspondence should be addressed.  
sshae5@trut.chungbuk.ac.kr

## 2. 통계적 분석

### 2.1. 상관도 해석

산업 재해 사례에서 요인들 간의 상관도 계산 방법의 기본적인 개념은 다음과 같다. 인자  $l$  또는 인자  $m$ 이 동시에 일어나는 재해 사례는 인자  $l$  또는 인자  $m$ 이 각각 일어난 사례에 대하여 어느 정도의 비율을 갖는가?하는 문제로 인자  $l$ 이 일어난 재해사례 집합을 A, 인자  $m$ 이 발생한 재해 사례 집합을 B 라 하고 각각의 요소 수를  $n(\cdot)$ 로 표시하면 인자  $l$ , 인자  $m$  사이의 상관도  $\rho_{lm}$ 은

$$\rho_{lm} = \frac{n(A \cap B)}{n(A \cup B)} \quad (1)$$

로 표시된다.

이때, 상관도의 실제 계산을 위해 재해사례 Data 를  $X_i$ 로 하면,

$X_i = [0, 1, 0, 1, \dots, 0, 1]$ 로 재해 사례에 1 : 1로 대응하고 있다.

재해 사례의 분류 인자를 순서대로 번호를 부여 하면,  $i$ 번째 사례의  $j$ 번째 인자에 있어서의 data는 다음과 같은 것이 된다.

$X_{ij} = 1$  ( $i$ 번째 사고의  $j$ 번째 인자가 일어난 경우)  
 $0$  ( $i$ 번째 사고의  $j$ 번째 인자가 일어나지 않은 경우)

여기서,  $i = 1, 2, 3, \dots, N$  (재해 사례의 총 건수)이고,  $j = 1, 2, 3, \dots, K$  (인자의 총 수)이다.

변수  $X_{ij}$ 를 위와 같이 정의하기로 하면, 인자  $l$ 과  $m$ 이 재해 사례 총 건수 가운데 함께 일어나고 있는 경우를 A, 인자  $l$ 이 일어나고, 인자  $m$ 이 일어나지 않는 경우의 수를 B, 그리고, 인자  $l$ 이 일어나지 않고 인자  $m$ 이 일어나는 경우의 수를 C라고 하면,

$$A = \sum_{i=1}^N X_{il}X_{im} \quad (2)$$

$$B = \sum_{i=1}^N X_{il}(1 - X_{im}) \quad (3)$$

$$C = \sum_{i=1}^N (1 - X_{il})X_{im} \quad (4)$$

로 되어, 앞서 정의되었던 상관도  $\rho_{lm}$ 은 다음 식과 같이 표시할 수 있다.

$$\rho_{lm} = \frac{A}{(A+B+C)} \quad (5)$$

여기에서 정의된  $\rho_{lm}$ 은 인자  $l$ 과 인자  $m$ 이 어느 정도의 관련성을 갖고 있는가를 나타내는 양이 된다.<sup>2)</sup>

### 2.2. 독립성 검정

두 변수간의 관련성의 여부와 검정 결과 만일 상호 관련성이 있다면, 그 정도는 얼마나 되는지를 알아보기 위하여 다음과 같이 독립성 검정 방법을 채택하였다.

#### 2.2.1. $\chi^2$ 통계량

만일 두 변수가 각각  $m, n$ 개의 범주를 갖는다면 교차표는  $(m \times n)$ 개의 cell을 이룬다. 이들 두 변수가 갖는 빈도분포에 의하여  $(m \times n)$ 개의 각 cell에 두 변수가 결합된 기대빈도를 구하고 실제 측정된 빈도와 비교함으로써  $\chi^2$ 값을 산출해내게 되는데, 독립성 검정에서 가장 많이 이용되는 것이 바로 Pearson의 카이제곱 통계량(Pearson's Chi-square Statistics)이다.<sup>3)</sup> Pearson의 카이제곱 통계량은

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{(E_i)} \quad (6)$$

여기서,  $O_i$ 는  $i$ 번째 범주에 대한 관측빈도를 말하고,  $E_i$ 는  $i$ 번째 범주에 대한 기대빈도,  $k$ 는 범주수로 정의되는데, 귀무가설이 맞을 때에는 자유도(degree of freedom ; df)가  $(m-1)(n-1)$ 인  $\chi^2$  분포를 따르게 된다.

결정원칙은 검정 통계량보다 더 큰 값이 나올 확률인 P값 (P - value)을 구해서 유의수준 ( $\alpha = 0.05$ )보다 작으면 귀무가설을 기각하게 되어 두 변수는 서로 독립이 아니라는 결론을 얻게 된다.

#### 2.2.2. 관련성 척도 ( $\lambda$ )

$\chi^2$  검정 결과는 두 변수간에 통계적으로 의미있는 관련성이 있는지는 알려주지만, 이를 근거로 두 변수간에 관련성이 어느 정도인지는 판단할 수 없다.  $\chi^2$ 값이 크다고 반드시 관련성이 강하다고는 볼 수 없으며, 관련성은 약하더라도 표본이 크면  $\chi^2$  값은 크게 나타날 수 있다. 그러므로, 관련성의 정도를 알기 위해서는 별도의 측정방법이 필요하다.

본 연구에서는 이 중에서 가장 무난하고 해석이 명확한  $\lambda$ 를 채택하여 관련성을 측정하기로 한다.  $\lambda$ 는 식(7)과 같이 정의되는데, 취할 수 있는 값의 범

위는  $0 \leq \lambda \leq 1$ 으로,  $\lambda = 0$ 일 때 두 변수는 완전히 독립적이고,  $\lambda = 1$ 일 때 두 변수는 완전한 관계임을 의미한다.

$$\lambda = \frac{\sum (f_i - F_d)}{(N - F_d)} \quad (7)$$

여기서,  $N$ 은 사례수이고,  $f_i$ 는 독립변수의 각 카테고리 내의 최빈도수를 말하며,  $F_d$ 는 종속 변수의 최빈도수를 말한다.<sup>4)</sup>

### 2.3. Logistic regression analysis

Logistic regression analysis는 자료의 분포형태가 비선형적이고, 종속변수가 이원화, 즉 단지 두 개의 값만을 취할 때, 종속변수와 독립변수들간의 인과관계를 추정하는 기법이다.<sup>5)</sup>

Logistic regression analysis로 어떤 사건이 일어날 확률을 직접적으로 추정할 수 있다. 하나 이상의 독립변수에 대하여, 이 모델은 다음과 같이 나타낼 수 있다.<sup>3)</sup>

$$P(x) = \frac{e^z}{(1 + e^z)} \quad (8)$$

여기서,  $z$ 는 선형결합을 나타내는데, 다음과 같이 나타낼 수 있다.

$$Z = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p \quad (9)$$

## 3. 분석결과 및 고찰

분석을 위해 수집된 산업재해사례 data들은 모두 25개의 항목들로 구성되어 있는데, 신뢰성 있는 분석을 위하여 10개 항목으로 재분류하였고, 또 각각의 항목들에 대해서 통계분석기법을 처리하는데 있어서 용이하도록 code화하였다.

또, 재해사례 조사 중 미기재 항목이나 기록이 불충분한 자료는 가용항목들만 code화하고, 나머지 항목들은 분류불능으로 처리하였다.

### 3.1 상관도 해석

재해사고의 과정상 중요한 변수라고 할 수 있는 사고의 발생형태와 근로손실일수에 영향을 줄 수 있는 기타 변수들간의 상관관계를 상관도 분석을 통하여 그 결과를 도출해 내었는데, 이 결과를 Table

1 및 Table 2에 나타내었다. 이들 Table에서 보는 바와 같이 가장 높은 상관도를 보이는 것은 Table 1의 경우 '61일~90일 - 6개월 미만'에서는 0.161이었고, Table 2에서는 '추락 - 6개월 미만'의 경우가 0.189로서 항목간 높은 교차빈도를 보일 경우 상관도도 높

Table 1. Correlation degrees between 'lost working days' and 'employed periods'

employed periods [yr] \ lost working days	-0.5	05-1	1-2	2-3	3-4	5-10	10-20	20-
1-14	.015	.011	.029	.010	.016	.020	.038	.067
15-28	.064	.046	.051	.055	.038	.036	.017	.011
29-60	.160	.106	.094	.054	.072	.071	.047	.009
61-90	.161	.052	.068	.073	.039	.064	.048	.010
91-120	.115	.066	.060	.048	.052	.047	.014	.010
121-150	.081	.037	.039	.039	.027	.039	.029	
151-180	.048	.048	.027	.039	.040	.015	.030	.014
181-270	.139	.056	.062	.025	.031	.027	.020	.003
271-365	.039	.020	.010	.023	.033	.009	.019	.029
366-730	.044	.039	.033	.015	.037	.043	.021	.044
death	.022	.024	.032	.029	.005	.038	.007	.022

Table 2. Correlation degrees between 'types of accident' and 'employed periods'

employed periods [yr] \ types of accident	-0.5	05-1	1-2	2-3	3-4	5-10	10-20	20-
electric shock	.012	.012	.004	.011	.012	.022		
personal disease	.005	.008	.008		.006	.017	.009	.043
traffic accident	.037	.103	.158	.092	.058	.067	.011	.031
dropping	.122	.070	.055	.055	.044	.050	.030	.010
upset&collapse	.175	.067	.028	.034	.040	.041	.027	.006
pneumoc oniosis						.011	.058	.244
suffocation& intoxication	.002		.004	.006				
falling	.189	.062	.049	.040	.043	.042	.018	.003
collision	.041	.034	.031	.033	.021	.037	.028	.008
narrowness	.187	.085	.115	.069	.059	.069	.033	.017
fire&explosion	.013	.012	.008	.011			.018	

게 나타났다.

**3.2. 독립성 검정**

Table 3은 '근로손실일수'와 '발생월' 항목간의  $\chi^2$  검정을 행한 결과이다. 양측검정의 접근 유의 수준이 설정 유의 수준( $\alpha=0.05$ )보다 작으므로 "두 변수가 상호 독립이다"라는 귀무가설을 기각하게 되어 대립가설, 즉 "두 변수가 상호 독립이 아니다."를 채택하였다. 즉, 근로손실일수는 재해가 발생하는 월에 따라 다르다는 결론을 얻었다. 또 우도비 검정(Likelihood Ratio Chi-Square) 결과도 같은 결과를 나타내고 있었다. 연관성의 정량적 척도인  $\lambda$ 값은 0.075로서 두 변수간에는 다소 미약한 연관성을 나타내고 있었다.

Table 3. Result of  $\chi^2$ -test on 'month' by 'lost working days'

Chi-Square Tests			
	Value	of	Approx. Sig.
Pearson Chi-Square	662.221 <sup>a</sup>	110	.000
Likelihood Ratio	659.289	110	.000
Linear-by-Linear Association	87.086	1	.000
N of Valid-Cases	2133		

a 22 cells(16.7%) have expected count less than 5. The minimum expected count is 1.39

**Directional Measures**

			Value	Asymp. Std. Error	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	.075	.010	.000
		month Dependent	.103	.011	.000
		lost working days Dependent	.043	.012	.000
	Goodman and Kruskal tau	month Dependent	.030	.002	.000b
		lost working days Dependent	.033	.003	.000b

a Not assuming the null hypothesis  
b Based on chi-square approximation

**3.3. Logistic Regression Analysis**

Logistic 회귀분석에서는 상관도 분석에서 나온 결과를 바탕으로 반응변수, 즉 '근로손실일수' 항목에 큰 영향을 미친다고 판단되는 독립변수 네 가지(재해형태, 기인물, 근속기간, 업종)를 채택하여 반응변수의 발생확률을 예측하였다. 이 분석기법에서는 반응변수가 2원화 되어야 하므로 반응변수 '근로손실일수' 항목을 경상(근로손실일수 56일 이하) 및 중상(근로손실일수 56일 초과)으로 나누어 독립변수들에 따라 각각의 발생확률을 계산하였다. Table 4에 logistic 회귀분석의 결과를 나타내었는데, 유의 수준으로 볼 때 독립변수 중 반응변수에 영향을 주는 것은 '연령', '근속기간', '기인물' 항목이었으며, 회귀식은 식(10)과 같다.

$$Z = 0.9717 + 0.0018(\text{accident types}) - 0.0385(\text{originated objects}) - 0.0552(\text{employed periods}) + 0.0745(\text{ages}) \quad (10)$$

그리고, 독립변수에 따라서 종속변수의 발생확률을 구하여 고찰하였는 바, 이 확률값은 회귀계수와 독립변수들의 곱의 합으로서 나타나는 선형회귀식

Table 4. Result of logistic regression analysis

· Number of selected cases : 2133 · Number rejected because of missing data : 0 · Number of cases included in the analysis : 2133  Dependent Variable : lost working days -2 Log Likelihood : 2635.8918 -2 Log Likelihood : 2589.731 Goodness of Fit : 2127.392 Cox R <sup>2</sup> : .021 Nagelkerke R <sup>2</sup> : .021						
	Chi-Square	df	Significance			
Model	46.161	4	.0000			
Block	46.161	4	.0000			
Step	46.161	4	.0000			
Variable	B	S.E	df	Sig.	R	Exp(B)
accident types	.0018	.0154	1	.9060	-.0836	1.0018
originated objects	-.0385	.0085	1	.0000	-.0836	.9622
employed periods	-.0552	.0209	1	.0081	-.0436	.9463
ages	.0745	.0202	1	.0002	.0663	1.0774
Const.	.9717	.1979	1	.0000		

에서 선형결합 값을 구한 후 식(8)과 (9)에 대입하여 구하였다. 즉, 증상재해의 발생확률이 높은 경우는 재해발생 형태 중에서는 '질식중독', '추락', '개인질병' 등이고, 기인물 중에서는 '건조설비', '화학설비' 등이었으며, 재해자의 연령은 '45-49세 미만', '55-59세 미만'과 같이 연령이 높고 근속기간은 '6개월 미만'과 같이 짧은 경우에 발생확률이 높게 나타났는데, 이는 안전교육을 충분히 실시하지 않은 채 위험작업에 종사하게 되어 재해가 발생한 것으로 사료된다.

#### 4. 결 론

산업재해 사례에 있어서의 재해발생 요인에 대한 제반 분석을 범주형 자료 분석기법을 이용하여 분석한 결과, 다음과 같은 결론을 얻을 수 있었다.

1) 상관도의 분석 결과, 변수 '근로손실일수'와 다른 변수간의 가장 높은 값을 나타내는 경우는 '(29-60일) - 업종(제조업)'인 경우 0.198, '(29-60일) - 성별(남자)'인 경우 0.230, '(181-270일) - 발생 월(6월)'인 경우 0.154, '(61-90일) - 근속기간(6개월 미만)'인 경우 0.161의 값을 나타내었다.

2)  $\chi^2$ -Test로 독립성 검정을 한 결과, 독립변수 중

근속기간, 기인물, 재해발생월, 성별, 업종, 재해형태 등이 종속변수인 근로손실일수와 상호관련이 있는 것으로 나타났고,  $\lambda$ 값을 적용시킨 결과 근로손실일수와 가장 관련성이 높은 요인은 재해발생월로 나타났다.

3) 근로손실일수를 정상 및 증상으로 나누어 반응변수와 독립변수의 인과관계를 추정하여, 증상재해의 발생확률을 높게 하는 변수요인들을 파악할 수 있었다.

#### 참고문헌

- 1) 홍광수, 정국삼, "상관분석을 응용한 산업재해사례 요인의 고찰," 산업안전학회지, Vol. 14, No. 1, pp. 141~149, 1999.
- 2) 정국삼, 재해사고분석론, 충북대학교 안전공학과, 1993.
- 3) 정충영, 최이규, SPSSWIN을 이용한 통계분석, 무역경영사, 1998.
- 4) 홍종선 외 2명, 조사방법과 통계자료분석, 전영사, 1996.
- 5) 배일섭, 정영숙, SPSS 한글프로그램과 통계분석기법, 대구대학교 출판부, 1998.