Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.26, NO.3, Jun. 2016 ISSN 2288-2715(Online)
http://dx.doi.org/10.13089/JKIISC.2016.26.3.799

개인정보유출 사고의 분포 추정에 관한 연구*

황 윤 희,† 유 진 호‡
상명대학교

A Study on the Distribution Estimation of Personal Data Leak Incidents*

Yoon-hee Hwang,† Jinho Yoo‡
Sangmyung University

요 약

본 논문은 국내 개인정보유출사고 발생의 패턴을 찾고 어떤 분포를 따르는지 확인한 연구이다. 이를 위해 2011년도부터 2014년도까지 언론에 보도된 개인정보유출사고를 사용하였다. 조사결과를 바탕으로 'K-S통계량' 방법론을 사용하여 개인정보유출사고의 통계적 분포를 추정하였고, 적합도 검정을 실시하였다. 그 결과 '유의수준 95%에서 포아송분포와 지수분포 모두 높은 적합도를 지닌다.'는 사실을 정략적으로 입증하였고, 이를 통해 1년에 평균 12번씩 대형 개인정보유출사고가 발생되어 언론에 보도되었다는 것을 확인할 수 있었다. 본 연구는 향후 기업 및 조직의 개인정보 유출사고의 발생예측 및 정보보호 투자금액선정 등 보안경제성 분석에 유용하게 활용될 것으로 전망된다.

ABSTRACT

To find the pattern of personal data leak incidents and confirm which distribution is suitable for, this paper searched the personal data leak incidents reported by the media from 2011 to 2014. Based on result, this research estimated the statistical distribution using the 'K-S Statistics' and tested the 'Goodness - of - Fit'. As a result, the fact that in 95% significance level, the Poisson & Exponential distribution have high 'Goodness-of-Fit' has been proven quantitatively and, this could find it for major personal data leak incidents to occur 12 times in a year on average. This study can be useful for organizations to predict a loss of personal data leak incidents and information security investments and furthermore, this study can be a data for requirements of the cyber-insurance.

Keywords: The Date of Personal Data Leak Incidents, Cumulative Distribution Function(CDF), Optimal Distribution Estimation

I. 서 론

최근 ICBM(IoT, Cloud, BigData, Mobile)이 유망사업으로 대두됨에 따라, IT 보안기술에 대한 관심이 증가하고 있다. 특히 2014년도에는 카드 3사의 대형 개인정보유출 사건이 발생한 후 개인정보보호에 대한 경각심도 크게 증가하였다.

이러한 흐름에 맞춰 다양한 논문지를 통해 개인정보유출 사고예방을 주제로 많은 논문이 게재되었다. 하지만 실제적으로 개인정보유출사고패턴의 분포에 대한 연구사례는 많지 않은 것으로 파악된다.

본 연구는 개인정보유출 사고발생의 패턴과 이에 대한 통계적 분포를 추정하려 한다. 이를 위해 관측데이터와 기존 이론상의 7가지 누적분포를 비교한 후, 가장 작은 유사성 차이(Similarity Gap)값을

Received(08. 24. 2015), Modified(1st: 10. 12. 2015 2nd: 10. 19. 2015), Accepted(10. 20. 2015)
* 본 연구는 2015년도 상명대학교 교내연구비를 지원받아 수행하였음.
† 주저자, denper1227@naver.com
‡ 교신저자, jhyoo@smu.ac.kr(Corresponding author)

갖는 누적분포의 종류를 식별하고자 한다. 그리고 관측 데이터와 누적분포함수와의 유사성을 측정할 시 'K-S 통계량'을 사용해 계산한 후, 가장 적합한 분포를 찾고자 'χ^2-통계량'의 적합도 검정 방법론을 이용하고자 한다.

II. 이론적 배경 및 문제제기

일반적으로 관측 데이터는 가장 높은 확률을 나타내는 평균과 분산 등의 가치에 대한 분포를 찾아내는 것이 분포 추정의 원리이다[3]. 이 원리에 따라 관측 데이터를 정형화된 누적분포함수(Cumulative Distribution Functions, CDF)와 비교해 어떠한 분포를 따르는지 추정하고자 한다.

누적분포함수(Cumulative Distribution Function, 이하 CDF)는 다양한 용도로 여러 분야에서 응용되어지고 있다. 그 중 관측 데이터에 최적의 근사 값을 지닌 분포추정을 구하는 다양한 선행연구가 존재하였다. Jun and Yoo[5]는 강우 시간분포를 예측하기 위해 베타 분포를 이용하였다. Kim and Cho[6]는 Time of Arrival(TOA)를 통한 목표물과의 거리 추정의 정확성을 높이기 위해 감마분포의 특성을 이용하고 그에 따른 알고리즘을 제안하였다. Riddhi[13]는 불완전 베타 함수의 불특정 요소를 베타 함수의 표준화로 간주한다고 주장하였다. Yim and Yang[4]은 해상 부유체 모델의 표본 데이터에 가장 높은 적합도를 지닌 누적분포함수를 탐색한 후 이에 대한 오차 평가를 실시하는 방안을 제시하였다.

관측 데이터의 분포 적합도를 검증하기 위한 방법론은 오랫동안 연구되어왔다. Kim 등[7]은 열차 소음 주파수의 청감반응평가실험을 수행하여 응답분포모형에 가장 높은 적합도를 지닌 곡선을 Hosmer-Lemeshow 검정과 적합도 검정을 통해 분석하였다. Song and Jung[14]은 다항 로짓 회귀모형의 적합성을 평가하기 위한 다양한 검정 방법들을 비교, 평가하였다.

위의 연구들은 관측 표본데이터에 대해 가장 높은 적합도를 갖는 분포함수를 예측한다는 점에서 본 연구에 적용될 수 있다. 본 연구에서는 개인정보유출사고에 초점을 맞춰 'K-S 통계량' 방법론과 '카이제곱 적합도 검정'을 통해 가장 적합한 누적분포함수를 찾고자 한다. 이 때, 본 연구에서는 분포추정에 가장 많이 사용되는 7가지의 함수(포아송, 지수, 정규, 감마, 베타, 와이블, 카이제곱)를 비교평가 대상으로 정하였다. 일곱 가지 분포함수 중 가장 밀접한 연관성은 지닌 포아송과 지수분포의 누적확률 계산식을 나타내면 다음과 같다[11].

x에 대한 포아송누적분포함수(Poisson CDF) F_{poiss}의 누적확률 p_{poiss}는 다음과 같이 나타낸다.

$$p_{poiss} = F_{poiss}(x \parallel \lambda) = e^{-\lambda} \sum_{i=0}^{floor(x)} \frac{\lambda^i}{i!} \quad (1)$$

여기서, λ는 형상 변수, $floor(x)$은 x값과 같거나 큰 정수에 대한 바닥 값, $i!$는 i의 팩토리얼(factorial)을 의미한다.

포아송분포는 이산 확률 분포로서, 단위 시간동안 특정 사건이 몇 번 발생하는지 표현하는 분포이다. 주로 이상 현상이 일어나는 사건에 이용되는 분포이다[15].

표본 데이터 x에 대한 지수누적분포함수(Exponential CDF) $F_{\exp}$의 누적확률 $p_{\exp}$의 식은 다음과 같다.

$$p_{\exp} = F_{\exp}(x \parallel \mu) = \int_0^x \frac{1}{\mu} e^{\frac{t}{\mu}} dt = 1 - e^{\frac{x}{\mu}} \quad (2)$$

여기서, μ는 CDF의 형상(shape)을 결정짓는 변수로써 데이터 평균을 의미하고, e는 지수(exponential), t는 적분을 위한 변수, $F(x \parallel \mu)$는 μ로 결정되는 x에 대한 CDF를 의미하고, 심벌 '$\parallel$'는 주어진 형상 변수에 대응한다.

지수분포는 인접 사건 간의 시간간격을 표현할 때 주로 사용되는 분포이다[1]. 사건이 서로 독립적이라는 조건 하에 평균발생횟수가 포아송분포를 따른다면, 다음 사건이 일어날 때까지의 대기시간은 지수분포를 따른다는 특징을 가지고 있다[16].

Lee[10]는 기준시간당 평균적으로 수신되는 바이러스 메일의 수가 포아송분포를 따르고, 바이러스 메일이 전송되는 시간간격의 분포가 지수분포에 적합한지 연구하였다. 하지만 해당 연구는 분석방법론에 있어 'χ^2 적합도 검정'에 한정되어 있다.

본 논문은 선행연구를 더욱 발전시켜 개인정보유출사고에 적용하고자 'χ^2 적합도 검정'을 통해 최적

의 분포를 정량적으로 입증하고자 한다. 또한 시뮬레이션을 통한 'K-S 통계량' 검증을 통해 모델의 적합성을 수치뿐만 아니라 시각적으로 확인하고자 한다.

III. 연구방법론

3.1 가설설정

본 연구에서는 개인정보유출사고 발생시기의 통계적 분포를 추정하고자 다음의 가설을 설정하였다.

H_0: 개인정보유출 사고의 평균 발생 수는 포아송 분포를 따를 것이다.
H_1: 개인정보유출 사고의 평균 발생 수는 포아송 분포를 따르지 않을 것이다.

가설설정에 앞서 이론적 배경을 통해 지수분포는 포아송분포와 높은 연관성을 지닌다는 사실을 확인할 수 있다. 포아송분포가 일정한 시간 내에 발생하는 사건의 수에 대한 분포라면 지수분포는 특정 사건이 일어나는 시간간격에 대한 분포를 알고자 할 때 적용된다[10]. 따라서 2011년도에서 2014년도까지의 "개인정보유출 사고 발생의 수가 포아성분포를 따른다."는 가설과 "개인정보유출의 시간간격은 지수분포를 따른다."는 가설을 동시에 입증하고자 한다. 따라서 본 연구에서는 다음의 가설도 수립하였다.

H'_0: 개인정보유출 사고발생의 시간간격은 지수분포를 따를 것이다.
H'_1: 개인정보유출 사고발생의 시간간격은 지수분포를 따르지 않을 것이다.

3.2 표본데이터의 확보

Table 1. Date of Personal Data Leak Incidents

Date		Exposure Organization	Date Gap
2011	01-01	**capital	0
	04-10	**capital	99
	04-11	**electronic corporation	1
	05-07	****industrial complex	26
	05-18	**capital	11
	05-19	***secondhand	1
		market	
	07-26	**card	68
	07-28	***portal	2
	08-13	***video service	16
	08-20	****producer	7
	09-01	****card	12
	09-19	****government department	18
	11-26	**game company	68
2012	01-03	*****, **wire service	38
	01-05	**card	2
	03-12	**resort	67
	05-17	***broadcaster	66
	07-20	**wire service	64
	09-29	maternity clinic	71
2013	02-06	***water purifier company	130
	02-07	**town office	1
	05-11	****lab	93
	05-28	***insurance company	17
	06-25	***goverment agency	28
	07-02	**wire service	7
	07-04	***game company	2
	07-21	***forum	17
	08-23	******game company	33
	10-03	***software company	41
	10-30	**national police agency	27
	12-11	**bank	42
2014	01-08	****, **, **card	28
	02-26	****, ****, ***association	49
	03-06	**wire service	8
	03-07	**social commerce	1
	03-16	**veteran association	9
	04-05	***chicken company	20
	04-13	**education	8
	04-16	****cosmetics	3
	05-09	****cosmetics	23
	07-14	**education	66
	09-13	*****video service	61
	09-18	****retailers	5
	10-12	***retailers	24
	10-19	****retailers	7
	12-02	*****wire service	44
	12-09	******power plant	7

Table 1은 최근 4년 동안의 유출상황을 시일별로 정리한 표이다. 특히 '시간간격(Date Gap)'은 표본 데이터의 유효한 가설설정을 위해 개인정보유출 시일 간의 차이를 연산한 값이다. 다음은 'Date Gap'에 대한 누적분포 값을 산출(Table 2)하고, 이를 그래프(Figure 1)로 제시한 결과이다.

Table 2. Date Gap of Personal Data Leak Incidents

Date Gap(x)	Frequency	Cumulative Percent(y)
0	2	4.2
1	4	12.5
2	3	18.8
3	1	20.8
5	1	22.9
7	4	31.3
8	2	35.4
9	1	37.5
11	1	39.6
12	1	41.7
16	1	43.8
17	2	47.9
18	1	50.0
20	1	52.1
23	1	54.2
24	1	56.3
26	1	58.3
27	1	60.4
28	2	64.6
33	1	66.7
38	1	68.8
41	1	70.8
42	1	72.9
44	1	75.0
49	1	77.1
61	1	79.2
64	1	81.3
66	2	85.4
67	1	87.5
68	2	91.7
71	1	93.8
93	1	95.8
99	1	97.9
130	1	100

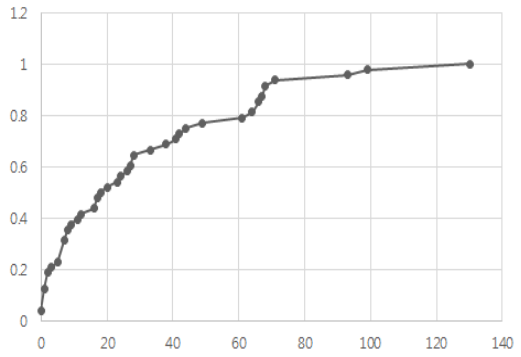

Fig. 1. CDF of Personal Data Leak Incidents

IV. 적합도 검정

앞서 설명한 두 가설을 증명하기 위해 카이제곱 적합도 검정을 적용하였다. 통계방법론을 통해 관측 데이터 분포와 포아성분포 및 지수분포와의 차이를 모두 검증하였다.

일반적으로 관찰결과로 얻은 데이터가 이론과 잘 일치하는지를 확인하는 것을 적합도 검정(goodness of fit test)이라고 한다. 이 때, 관측도수와 이론적인 기대도수가 부합하는지를 검정하기 위해 χ^2-통계량을 이용하였다. 기대도수를 E, 관측도수를 O라고 가정하면 χ^2-통계량 식은 다음과 같다.

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{3}$$

4.1 포아송분포의 적합도 검정

포아성분포의 적합도 검정을 위해 2011년부터 2014년까지의 개인정보유출 발생빈도(Occurrence Frequency)를 산출하였다.

발생한 사건빈도(Occurrence Frequency)를 분석한 결과 4년 동안 평균 1개월에 1번씩 유출사고가 있었다는 것을 알 수 있다. 다음은 도출된 평균값을 포아송분포 식에 대입하여 기대도수를 구하였다. 대입 식은 다음과 같다.

$$p_{poiss} = F_{poiss}(x \parallel \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{4}$$

Table 3. Occurence Frequency of Personal Data Leak Incidents

Time of Occurrence		Occurrence Frequency
2011	01	0
	02	0
	03	0
	04	2
	05	3
	06	0
	07	2
	08	2
	09	2
	10	0
	11	1
	12	0
2012	01	2
	02	0
	03	1
	04	0
	05	1
	06	0
	07	1
	08	0
	09	1
	10	0
	11	0
	12	0
2013	01	0
	02	2
	03	0
	04	0
	05	2
	06	1
	07	3
	08	1
	09	0
	10	2
	11	0
	12	1
2014	01	3
	02	1
	03	3
	04	3
	05	1
	06	0
	07	1
	08	0
	09	2
	10	2
	11	0
	12	2

여기서 λ는 '평균발생횟수'를 나타내고, x는 30일을 기준으로 하는 'Occurence Frequency(사건의 발생횟수)'를 의미한다.

그리고 x값에 해당하는 빈도수를 조사하여 'Actual Frequency(실제빈도)'와 'Actual Percent(실제확률)'의 값을 구하였다. 'Expected Percent(이론적확률)'는 포아송분포의 식에 평균발생횟수를 λ에 대입하여 나온 결과 값이다.

χ^2-통계량은 적합도 검정 통계량(Value of 'Goodness-of-Fit')을 모두 합한 '0.106551772'이다. 이 값은 χ^2분포표에 의해 유의수준 95%의 기각값인 '$\chi^2_{0.05}(3\,d.f.)=0.99104$'보다 훨씬 작으므로 포아송분포를 따른다고 할 수 있다. 즉, 일정기간 동안의 개인정보유출 사고발생 수는 유의수준 5%에서 포아송분포를 따르므로, 귀무가설 H_0를 채택한다.

Table 4. The Value of 'Goodness-of-Fit' on the Poisson distribution

x	Actual Frequency	Actual Percent	Expected Percent	Value of 'Goodness-of-Fit'
0	21	0.438	0.367879441	0.0133655
1	11	0.229	0.367879441	0.052428859
2	11	0.229	0.183939721	0.011038555
3	5	0.104	0.06131324	0.029718858
Chi-Square Test Statistics				**0.106551772**
P-Value[$\chi^2_{0.05}(3\,d.f.)$]				**0.99104**

4.2 지수분포의 적합도 검정

개인정보유출의 발생빈도가 포아송분포를 따르는 것으로 검증되었기 때문에 사고발생간격은 지수분포를 따른다는 것을 실증적으로 검증하고자 한다.

이를 위해 관측데이터의 누적분포함수와 이론적 통계분포를 비교하여 최적의 누적분포함수를 찾는 'K-S 통계량' 검증과 'χ^2 적합도 검정' 방법론을 사용하고자 한다.

4.2.1 최적 누적분포함수 추정

Table 5는 7가지 분포함수의 결과 값 산출을 위

Table 5. Excel Codes to Calculate The Cumulative Probability of 7 Distribution Types

Distribution Types (a)	Excel codes to calculate the cumulative probability (b)	Variable Types (c)
Poisson	POISSON.DIST	(x, μ, cumulative)
Exponential	EXPON.DIST	(x, λ, cumulative)
Normal	NORM.DIST	(x, μ, σ, cumulative)
Gamma	GAMMA.DIST	(x, α, β, cumulative)
Beta	BETA.DIST	(x, α, β, cumulative)
Weibull	WEIBULL.DIST	(x, α, β, cumulative)
Chi-Square	CHISQ.DIST	(x, k, cumulative)

한 함수(b)와 추정에 필요한 모수값(c)을 정리한 표이다.

가장 높은 적합도의 누적분포함수를 추정하기 위한 평가방법과 절차는 다음과 같다. 우선, (1)변수 x값(Date Gap)을 누적분포함수식에 대입한다. 그리고 (2)각 함수의 조건에 만족하는 임의의 형상변수 값을 대입하여 $F(x)$값을 도출한다. (3)도출된 $F(x)$값에 대하여 'K-S 통계량(Kolmogorov-Smirnov test)'을 적용한다.

'K-S 통계량'은 Kolmogorov-Smirnov test, 실증적 데이터와 확률분포의 비교분석에 사용될 수 있는 방법론이다. 'K-S 통계량'을 통해 표본 데이터의 실증적 분포와 누적분포그래프 사이의 거리 값을 정량화할 수 있다[2]. 이 때, 주어진 누적분포함수 $F(x)$와 실증적 분포함수인 $F_n(x)$의 차이비교를 위한 식은 다음과 같다.

$$D_n = \max[|F_n(x) - F(x)|] \quad (5)$$

여기서 D_n은 'K-S 거리'를, n은 '표본데이터 지점'을 나타낸다. 이를 통해 두 누적분포그래프 사이의 거리는 관측데이터 지점과 이론적 분포 값과의 차이에 대한 절대 값임을 알 수 있다.

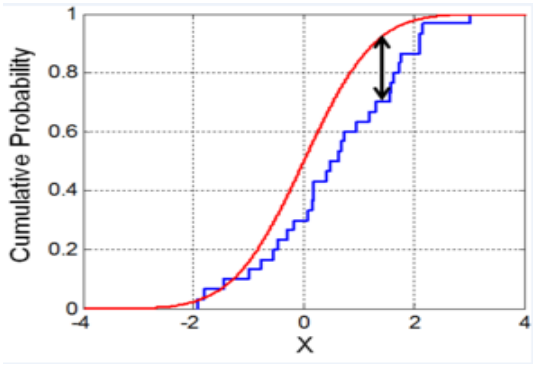

Fig. 2. Illustration of the 'K-S statistic'

'K-S 통계량' 방법론을 본 연구에 적용하여 (2)에서 도출된 결과($F(x)$)값과 관측데이터의 누적퍼센트 값($F_n(x)$)의 차이에 대한 절대 값을 구하였다. (4)각각의 x값에 대한 '$|F_n(x) - F(x)|^2$'의 값들을 모두 합산하여 'Similarity Gap'을 측정한다. 이를 식으로 정리하면 다음과 같다.

$$\text{Similarity Gap} = \sum (F_n(x) - F(x))^2 \quad (6)$$

이 때, 'Similarity Gap'은 해당 통계분포와 관측데이터 분포의 차이 값을 의미하며, 'Similarity Gap'의 크기가 작으면 작을수록 관측데이터와 가장 높은 유사도를 가지게 된다. (5)합계가 최솟값이 나오도록 누적분포함수의 형상변수 값을 조정한다. (1)~(5)의 과정을 반복적으로 시뮬레이션하여 나온 값을 Table 4에 정리하였다.

표에서 나타나듯이, 모수값 "$\lambda = 0.0354$"의 지수

Table 6. The Result of 'K-S Statistics' on 7 Distribution Types

Distribution Types and Shape Parameters			Values of Shape Parameters		Similarity Gap
Poisson	μ		21	-	1.95207733
Exponential	λ		**0.0354**	**-**	**0.11652498**
Normal	μ	σ	20	31	0.14070352
Gamma	α	β	10	2	1.60548
Beta	α	β	2	2	0.15190157
Weibull	α	β	1	28	0.11666471
Chi-Square	k		3	-	0.11910662

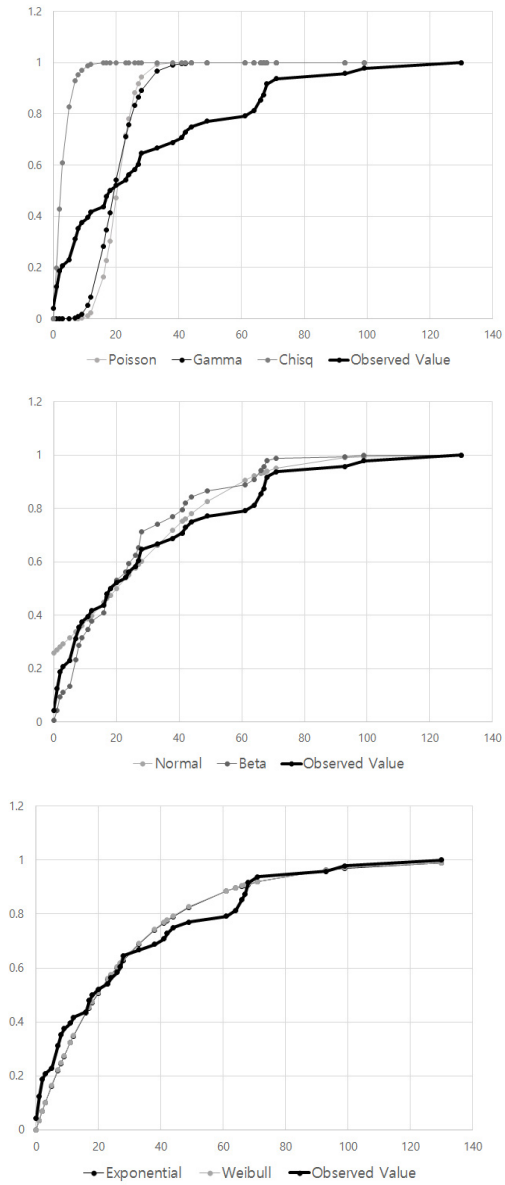

Fig. 3. Comprehensive Cumulative Distribution Function

분포가 "0.11652498"로 가장 작은 유사도 차이 값(Similarity Gap)을 지니고 있는 것으로 나타났다. 이 때, λ는 반복 시뮬레이션을 통해 산출된 값을 의미한다. 와이블분포도 매우 작은 유사도의 차이 값을 가지는데, 이를 통해 "와이블분포는 지수분포 같은 통계적인 분포를 흉내 낼 수 있는 가능성이 있다[16]."는 이론이 실제 관측치로 입증되었다. 또한 카이제곱 분포도 유사한 것으로 측정되었다.

Figure 3은 실제 값들을 도식화 한 후 정리한 그래프이다. 검은색 굵은 그래프(Observed Value)는 표본1데이터이자 기준그래프를 나타낸다. 그림을 살펴보면 다른 그래프와 비교하여 '지수분포'와 '와이블분포'가 기준그래프와의 유사도가 가장 높은 것을 확인할 수 있다.

4.2.2 χ^2 적합도 검정

4.2.1의 과정에서 지수분포, 와이블분포, 카이제곱 분포 모두 높은 적합도를 가지므로 추가적으로 χ^2적합도 검정을 시행하였다. Table 7,8,9는 관측데이터의 실제 빈도수와 누적분포함수의 액셀 가상시뮬레이션 값을 비교한 χ^2적합도 검정의 결과이다.

x값인 'Date Gap(시간간격)'은 위의 누적분포 비교 시 언급한 변수와 동일한 값으로, 범위를 지정하여 빈도수를 측정하였다. 실제 데이터 상에서도 x값에 대응하는 빈도수가 대부분 '1'이므로 이를 조정하기 위해 괄호 안의 수를 기준으로 구분하였다. 그리고 내림차순으로 분류된 7개의 x값에 따라 결과값을 측정하였다.

Actual 값과 Expected 값을 비교한 결과, 지수분포, 와이블 분포, 카이제곱 분포의 χ^2-통계량은 '0.132326', '0.137753', '3.518911'로 나타났다. 지수분포와 와이블분포의 카이제곱 검정값은 각각의

Table 7. The Value of 'Goodness-of-Fit' on the Exponential distribution

Date Gap (x)	Actual Frequency	Actual Percent	Expected Percent	Value of 'Goodness-of-Fit'
1(0~9)	18	0.375	0.471229	0.019651
2(10~19)	6	0.125	0.191359	0.023012
3(20~29)	7	0.1458333	0.219484	0.024714
4(30~49)	6	0.125	0.191359	0.023012
5(50~79)	8	0.1666667	0.246631	0.025927
6(80~99)	2	0.0416667	0.068352	0.010418
7(100~)	1	0.0208333	0.034781	0.005593
Chi-Square Test Statistics				**0.132326**
P-Value[$\chi^2_{0.05}(6\,d.f.)$]				**0.999954**

Table 8. The Value of 'Goodness-of-Fit' on the Weibull distribution

Date Gap (x)	Actual Frequency	Actual Percent	Expected Percent	Value of 'Goodness-of-Fit'
1(0~9)	18	0.375	0.474212	0.020757
2(10~19)	6	0.125	0.192882	0.02389
3(20~29)	7	0.1458333	0.221199	0.025678
4(30~49)	6	0.125	0.192882	0.02389
5(50~79)	8	0.1666667	0.248523	0.026961
6(80~99)	2	0.0416667	0.068937	0.010788
7(100~)	1	0.0208333	0.035084	0.005788
Chi-Square Test Statistics				**0.137753**
P-Value[$\chi^2_{0.05}(6\ d.f.)$]				**0.999948**

Table 9. The Value of 'Goodness-of-Fit' on the Chi-Square distribution

Date Gap (x)	Actual Frequency	Actual Percent	Expected Percent	Value of 'Goodness-of-Fit'
1(0~9)	18	0.375	0.99956	0.390247
2(10~19)	6	0.125	0.88839	0.655978
3(20~29)	7	0.1458333	0.928102	0.65935
4(30~49)	6	0.125	0.88839	0.655978
5(50~79)	8	0.1666667	0.953988	0.649772
6(80~99)	2	0.0416667	0.427593	0.34832
7(100~)	1	0.0208333	0.198748	0.159265
Chi-Square Test Statistics				**3.518911**
P-Value[$\chi^2_{0.05}(6\ d.f.)$]				**0.741452**

기각값인 '0.999954', '0.999948'보다 작기 때문에 관측데이터가 2가지 분포를 따른다 할 수 있다. 반면, 카이제곱분포는 기각값인 '0.741452'보다 큰 값을 가지므로 적합하다고 볼 수 없다. 지수분포와 와이블 분포 모두 적합성을 갖지만, 이 중에서 지수분포가 가장 최소값을 갖는 것으로 나타났다. 즉, 개인정보유출사고 발생의 시간간격은 '지수분포'와 가장 유사하다고 할 수 있고, 귀무가설인 H'_0를 채택한다.

V. 결 론

본 논문은 개인정보유출사고 발생관련 실증적 데이터에 대한 정량적 확률분포 추정을 시도하였고 개인정보유출 사고 발생을 설명해주는 가장 유사한 통계적 분포를 찾고자 하였다.

카이제곱 적합도 검정의 결과, 유의수준 95%에서 개인정보유출사고의 평균발생횟수는 포아송 분포를 따르는 것으로 입증되었다. 또한, 개인정보유출사고 발생의 시간간격은 관측데이터의 누적분포와 7가지 통계적 분포를 비교한 결과 지수분포와 가장 가깝다는 사실을 입증할 수 있었다.

해당 결과는 대형 개인정보유출사고 발생빈도를 예측하고 확률을 측정하여 앞으로 발생할 수 있는 사고에 대비해 투자해야하는 금액 등을 산정하는 다양한 정량적 보안경제성 분석에 기초자료로 유용하게 활용될 수 있을 것으로 판단된다. 특히, 사고발생예측을 통해 손실을 예상하고 그에 따른 정보보안 예산의 투자금액산정연구 뿐만 아니라 개인정보 유출사고에 대한 통계적 분포추정을 수행함으로써 보안경제성 연구 활성화에 활용될 것으로 기대된다.

본 연구는 관측데이터가 2011년도부터 2014년도까지의 4개년의 데이터에 한정되어 있다는 한계점을 지닌다. 이러한 데이터 부족의 한계점을 극복하기 위해 관련데이터의 양을 증가시키고 더욱 정교화 하고자 한다. 향후 본 연구는 2015년의 개인정보유출사고 데이터를 확보하여 연구모델의 예측 검증력을 평가하고자 한다. 또한 데이터를 수집함에 따라 개인정보유출사고 발생률을 기관·업종별로 분류하고, 각각의 분포를 추정하여 연구결과의 신뢰도를 향상시킬 예정이다.

References

[1] Cha Jae Bok, "Gamma distribution," 2013. 10.10 http://ktword.co.kr/abbr_view.php?m_temp1=4413

[2] David Vose, Quantitative Risk Analysis : A Guide to Monte Carlo Simulation Modelling, John Wiley & Sons, 605 Third Avenue, New York, NY 10158-0012, USA, pp.126-132, 1996.

[3] David Vose, "Fitting Distributions to Dat

a and why you are probably doing it wrong," 2010.02.15, http://www.vosesoftware.com/whitepapers/Fitting%20distributions%20to%20data.pdf.

〔4〕 Jeong-Bin Yim and Won-Jae Yang, "Estimating Cumulative Distribution Funtions with Maximum Likelihood to Sample Data Sets of a Sea Floater Model," Journal of Navigation and Port Research, 37(5), pp.453-461, Oct. 2013.

〔5〕 Jun Chang Hyun and Yoo Chul Sang, "Application of the Beta Distribution for the Temporal Quantification of Storm Events," Journal of Korean Water Resources Association, 45(6), pp.531-544. Mar. 2015.

〔6〕 Kim Jin Ho, Kim Hyeong Seok and Cho Sung Ho(2013), "A Ranging Algorithm for IR-UWB in Multi-Path Environment Using Gamma Distribution," The Journal of Korea Information and Communications Society, 38B(2), pp. 146-153, Feb. 2013.

〔7〕 Kim Phillip, Ahn Soyeon, Jeon Hyesung, Lee Jae Kwan, Park Sunghyun, Chang Seoil, Park Ilgun, Jung Changu and Kwon Segon, "Classification Accuracy Test of Hearing Laboratory Test Models for Railway Noise at Station Platform," Trans. Korean Soc. Noise Vib. Eng., 25(4), pp.299-305, Mar. 2015.

〔8〕 KISA, Information Security Survey, "Insurance and Reporting against Security Incidents", pp.26, Dec, 2010.

〔9〕 Korea Insurance Development Institute, CEO Report, "Activation Methods of Private Information Liability Insurance", CR 2012-04, pp.11, Dec. 2012.

〔10〕 Lee Jung Hoon, "Goodness of Fit for Probability Model," Master's Degree Thesis, Semyung University, Aug. 2004.

〔11〕 MATLAB(2008a), Programming, MATLAB Version 7.6 (R2008a).

〔12〕 Ministry of Science, ICT, and Future Planning, The Study of Broadcasting and Communications Policy, "A Study on Estimating Economic Damages from Internet Incidents for Cybersecurity Insurance", 13-Jinheong-098, pp.66, Nov. 2013.

〔13〕 Riddhi D., "Beta Function and its Applications," Department of Physics and Astronomy, The University of Tennessee, 2008.10.27, http://sces.phys.utk.edu/~moreo/mm08/Riddi.pdf.

〔14〕 Song Mi Kyung and Jung In Kyung, "Comparison of Goodness-of-Fit Tests using Grouping Strategies for Multinational Logit Regression Model," The Korean Journal of Applied Statistics, 26(6), pp.889-902, Oct. 2013.

〔15〕 Wikipedia, Tutorial for Poisson distribution, 2016.04.24, https://en.wikipedia.org/wiki/Poisson_distribution.

〔16〕 Wikipedia(2015), Tutorial for Exponential distribution, 2016.03.17, https://en.wikipedia.org/wiki/Exponential_distribution.

〈저자소개〉

황 윤 희 (Yoon-hee Hwang) 정회원
2015년 2월: 상명대학교 경영학과 졸업
2015년 3월~현재: 상명대학교 지식보안경영학과 석사과정
〈관심분야〉 기업보안, 정보보호, MIS

유 진 호 (Jinho Yoo) 종신회원
1992년 2월: 고려대학교 수학과 졸업
1994년 2월: 고려대학교 통계학과 석사
2010년 2월: 고려대학교 정보보호 박사
1993년 11월~1999년 12월: 한국전자통신연구원 연구원
2000년 1월~2004년 9월: IBM KOREA 전문차장
2004년 10월~2012년: KISA 인터넷문화진흥단장
2013년~현재: 상명대학교 경영학과 교수
〈관심분야〉 정보보호, 개인정보보호, MIS, 인터넷윤리