

Full Length Research Paper

Risk analysis and assessment on construction operation based on human factors and empirical Bayesian theory

Kai-Chang Sun¹ and Jian-Lan Zhou^{2*}

¹College of Hydraulic and Environmental Engineering, China Three Gorges University, Yichang, Hubei, 443002 China.

²Automation College, Huazhong University of Science and Technology, Wuhan, Hubei, 430074 China.

Received 3 January, 2013; Accepted 2 March, 2013

This study was focused on the roles of human factors in construction accident and also dealt with the probabilities of four levels of injury. We used an empirical Bayesian method and the human factors analysis and classification system framework to analyze the probability distributions of the severity of accidents of high risk operations in hydropower construction. Accident severity in four levels of injury was modeled: severe injury, one death, two deaths, and three deaths. The results show the behavior characteristics of workers and factors influencing their operation violations. The calculation of posterior distributions of the levels of injury enables us to rank the factors with respect to their risk of injury. The study revealed that lack of the ability to determine hazards is the direct reason in many accidents; resource management, inadequate supervision and supervisory violations also play important roles in the occurrence of accidents.

Key words: Empirical Bayesian analysis, human factors, work accidents, risk analysis and assessment.

INTRODUCTION

In the southwest China, there are numerous large-scale hydropower projects that have been built and put into operation. The construction projects are large-scale, with long construction period, and are extremely complex with high safety risks. At present, there are no systematic and thorough study on the hazards identification, risk evaluation and control management on the hydropower construction project at home and abroad. Though our country has built laws, technical specifications, and procedures, there are no related concrete contents of the construction safety hazards. On the other hand,

strengthening the construction accident statistics work is an important part of the safety production. Construction accident statistics is the basic management work in the construction safety production. Statistical data are used to analyze the key factors of the accidents to obtain the cause and regulation of the accidents.

Hazard risk assessment involves many factors, which is a dynamic system with interaction factors. Therefore, the establishment of the evaluation indicator set must be scientific, rational and comprehensive. This study begins from the view of common hazard identification and

*Corresponding author. E-mail: zhoujl1999@163.com

Author(s) agree that this article remain permanently open access under the terms of the [Creative Commons Attribution License 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

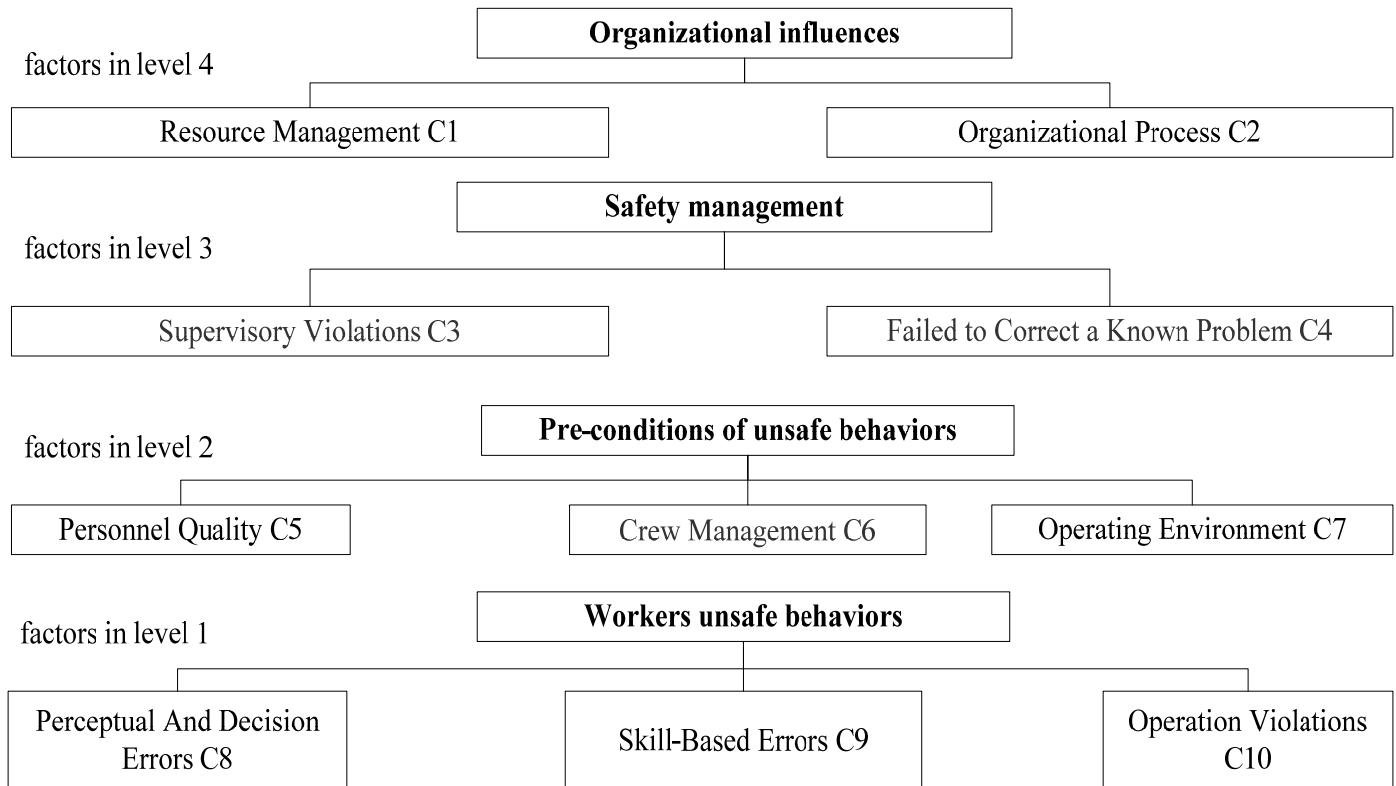


Figure 1. Four levels of the HFACS framework.

evaluation, takes into account the characteristics of construction safety hazard, and deeply analyzes the main factors influencing hazard risk assessment. On the basis of the human factors analysis and classification system framework (HFACS), we use classical Bayesian theory to calculate and sort the importance degree of the factors.

METHODOLOGY

Human factors analysis and classification system framework (HFACS)

The safety assessment structure model of high-risk operation and risk factors causing accidents are shown in Figure 1, which is also a HFACS framework (Wiegmann and Shappell, 1997, 2001). This study uses the HFACS framework to identify the kind of factors that are of most importance in the accidents.

The HFACS framework classified the human fault factors involved into four levels. L1 level factors were the focus of the past accident investigation, which are the personnel behavior mistakes causing the accident (i.e. construction workers unsafe behaviors). This level can be further divided into four categories: decision-making mistake, skill-based errors, perception errors and violation operations. L2 level factors (precondition for unsafe behaviors-potential/obvious mistakes) refer to the obvious psychological or environmental factors' influence on L1 factors, such as operation environment that does not meet the required conditions and are more likely to cause accidents. L3 level factors (unsafe supervision-the potential mistakes) are the onsite supervision factors that lead to unsafe behaviors. L4 level factors (organizational influences-the

potential mistakes) are described as the wrong decisions made by the high-level managers' influence on the low-level workers' behavior.

In HFACS framework, the higher level can influence the adjacent lower level. Adjusting and replacing some factors according to the characteristics of construction can strengthen the framework's independence and generality; and makes it possible to quickly identify the key factors in the whole system. The amendment HFACS framework is more in line with the actual hydropower construction project.

Statistical framework

There are two basic concepts in Bayesian statistics: the prior distribution and posterior distribution.

Prior distribution is a probability distribution of the overall distribution parameter θ . The fundamental idea of Bayesian School is that any statistical inference on the overall distribution of the parameter θ , in addition to samples provided by the information must also provide for a prior distribution. It is an indispensable element for making statistical inferences. They believe that the prior distribution does not need to have an objective basis, and can be partially or fully based on subjective belief (Shappell and Wiegmann, 2001); posterior distribution, according to the samples distribution and the prior distribution of the unknown parameters, with the method of conditional probability distribution to find out the conditional distribution of unknown parameters under the known samples. Because of this, distribution obtained after the sampling is called posterior distribution. The key of Bayesian inference method is that any inference must only rely on the posterior distribution, and no longer involve in sample distribution (de Lapparent, 2006).

The severity of the injury accident cases is divided into four levels: severe injury, one death, two deaths and three deaths. K indicates the mutually exclusive type of injury accident (that is, event space). Y_{ij} is the result of the accident type i leading to the injury type j , which is represented by an array of discrete random variable, $Y_i = (Y_{i,1}, \dots, Y_{i,k})'$.

For all $j=1, \dots, K$, Y_{ij} takes the value 1 if the severity of accident is the same with level j and the value 0 otherwise. A probability measure P_i on BY , the σ -algebra of the K elementary events, is associated with each other.

$P_{i,j} = P_i(Y_{i,j} = 1)$ is the probability that the accident i has an injury of type j . Therefore, we state that $Y_i | P_i \xrightarrow{id} M(1, P_i), P_i = (P_{i,1}, \dots, P_{i,k})'$, where id stands for 'independently distributed': Y_i is distribution conditionally on P_i with a Multinomial probability distribution with parameter 1 and P_i (Leonard, 1977). The corresponding conditional probability density functions is then

$$f_{Y_i | P_i}(y_i) = \prod_{j=1}^k p_{i,j}^{y_{i,j}}, \quad p_{i,j} \in [0,1], \quad y_{i,j} \in \{0,1\}. \tag{1}$$

Where $\sum_{j=1}^k p_{i,j} = 1$ and $\sum_{j=1}^k y_{i,j} = 1$

Because there are different accident configurations, understood as different circumstances and consequences, there are variations of the probabilities of the types of injury according to many factors. Due to the unpredictable nature of accident, for each accident i , P_i is random, and uses known accident model in line with the Bayesian analysis principle.

In a Bayesian context, the beliefs one can have about the family of distributions of probabilities of the types of injury are summarized through prior distributions. They represent the state of knowledge about the individual distributions of the types of injury before observing the sample $y = (y_1, \dots, y_n)'$. For convenience, it is assumed that they all belong to the same family of probability distributions. A Dirichlet distribution is used because it can be mixed conveniently with the Multinomial distribution and it results in a Dirichlet posterior distribution. For all $i = 1, \dots, n$, $p_i \xrightarrow{id} D(\theta_{i,1}, \dots, \theta_{i,k})$, where $\theta_{i,k} > 0, j = 1, \dots, K$, are shape parameters. The probability density function for the prior distribution is

$$\pi_{p_i | \theta_i}(p_i) = \frac{\Gamma(\sum_{i=1}^K \theta_{i,j})}{\prod_{j=1}^K \Gamma(\theta_{i,j})} \prod_{j=1}^K p_{i,j}^{\theta_{i,j}-1}, \quad p_{i,j} \in [0,1], \theta_{i,j} > 0 \tag{2}$$

where Γ is the gamma function:

$$\Gamma(s) = \int_0^{\infty} e^{-x} x^{s-1} dx, \Gamma(s) = (s-1)\Gamma(s-1). \tag{3}$$

Using the conditional distribution (1) and the prior distribution (2), the joint distribution of (Y_i, P_i) is defined:

$$\psi_{Y_i, p_i | \theta_i}(y_i, p_i) = \frac{\Gamma(\sum_{i=1}^K \theta_{i,j})}{\prod_{j=1}^K \Gamma(\theta_{i,j})} \prod_{j=1}^K p_{i,j}^{y_{i,j} + \theta_{i,j} - 1}. \tag{4}$$

The marginal distribution of Y_i is derived using (4) by integrating over P_i :

$$g_{Y_i | \theta_i}(y_i) = \frac{\Gamma(\sum_{j=1}^K \theta_{i,j})}{\prod_{j=1}^K \Gamma(\theta_{i,j})} \int_{[0,1]^K} \prod_{j=1}^K p_{i,j}^{y_{i,j} + \theta_{i,j} - 1} dp_{i,1} \dots dp_{i,K} \\ = \prod_{k=1}^K \left(\frac{\theta_{i,k}}{\sum_{j=1}^K \theta_{i,j}} \right)^{y_{i,k}} \tag{5}$$

The posterior distribution of $p_i | y_i$ is obtained using the transition formula of Bayes:

$$h_{p_i | y_i, \theta_i} = \frac{\psi_{Y_i, p_i | \theta_i}(y_i, p_i)}{g_{Y_i | \theta_i}(y_i)} \\ = \frac{\Gamma(\sum_{j=1}^K \theta_{i,j} + y_{i,j})}{\prod_{j=1}^K \Gamma(\theta_{i,j} + y_{i,j})} \prod_{j=1}^K p_{i,j}^{y_{i,j} + \theta_{i,j} - 1} \tag{6}$$

Where is a Dirichlet distribution:

$$p_i | y_i, \theta_i \xrightarrow{id} D(\theta_{i,1} + y_{i,1}, \dots, \theta_{i,K} + y_{i,K}). \tag{7}$$

Empirical Bayesian analysis is performed in two steps: first, estimate the unknown hyper parameters; secondly, compute the posterior distributions using the estimated parameters and analyze the results.

Estimation of hyper parameters

As stated above, before accident occurs, construction organizations attempt to improve the workers' safety consciousness, use safety equipment, strengthen safety supervision and management, and control the probability distribution of P_i of the severity of accident. Also, the construction environment and accident attributes will play different roles when the corresponding accident happens. From the statistical standpoint, it means that one can explain the values of the shape parameters of the distributions according to some exogenous factors.

Let $X_i \in X \subset R^p$ be a $(p, 1)$ array of explanatory variables about observation i , and let $X = (X_1, \dots, X_n)'$ be the full rank (n, p) matrix of explanatory variables about the observed sample. Let $\forall k = 1, \dots, K, \beta_k \in \Theta_k \subset R^p$ be a $(p, 1)$ array of (unknown) weights measuring the causal effects of explanatory variables on the shape of the probability distribution through $\theta_{i,k} \cdot \beta_j$. Measure the effects of the explanatory variables X_i on the importance of the outcome j in the probability distribution of accident severity. $X, \Theta_1 \times \dots \times \Theta_k$ are assumed to be compact subsets of R^p . In order to maintain consistency with the strict positivity of the shape parameters, it is assumed for the rest of the paper that:

$$\theta_{i,j} = \exp(x_i' \beta_j) \tag{8}$$

The only available observation we have for an individual i is that severe injury, one death, two deaths and three deaths. In order to infer on the values of the unknown hyper parameters of the model, the marginal probability distributions of $Y_i, i = 1 \dots n$, are used to build the sample log-likelihood function of observed $y_i, i = 1 \dots n$:

$$\ell(\beta | y, x) = \prod_{i=1}^n \prod_{k=1}^K \left(\frac{\exp(x_i' \beta_k)}{\sum_{j=1}^K \exp(x_i' \beta_j)} \right)^{y_{i,k}} \tag{9}$$

Looking at (9), we see there is no objective relationship between the value of the log-likelihood function and the values of the parameters as long as we do not set further identification restrictions. In fact, we must choose a benchmark outcome k^* which for $\beta_{k^*} = 0$, and all the remaining estimable parameters are expressed as differences with respect to β_{k^*} . The maximization of (9) with respect to the unknown parameters gives us asymptotically unbiased and efficient estimates we use to compute the posterior probability distribution of accident severity.

Analysis of result

Hyper parameters are interpreted in the sense of their causal influences on the shapes and moments of probability distributions of the types of accident. The posterior distribution represents the state of knowledge concerning π_i after the observations have been combined with the prior information. Posterior distributions represent our updated beliefs about probability distribution of the types of accident after accidents happened.

The Bayesian estimator of the expected rate of an accident with injury of level j for individual i is given by the posterior mean of $p_{i,j}$ (Gwet, 2002b):

$$E(p_{i,j} | y_i, \hat{\alpha}_i, \hat{\beta}_i) = \frac{\theta_{i,j} + y_{i,j}}{\sum_{k=1}^K (\theta_{i,k} + y_{i,k})} \tag{10}$$

And the posterior variance of the distribution is

$$V(p_{i,j} | y_i, \hat{\alpha}_i, \hat{\beta}_i) = \frac{(\theta_{i,j} + y_{i,j})(\sum_{k=1}^K (\theta_{i,k} + y_{i,k}) - (\theta_{i,j} + y_{i,j}))}{(\sum_{k=1}^K (\theta_{i,k} + y_{i,k}))^2 (1 + \sum_{k=1}^K (\theta_{i,k} + y_{i,k}))} \tag{11}$$

MODEL CALCULATION AND ANALYSIS

Factors analysis

Select 59 accident cases occurring in the construction peak period of the Xiluodu project and the Xiangjiaba project. The accidents are not described as name but serial number from 1 to 59. The statistical process is to determine in turn whether the HFACS factors are the reasons causing these accidents. In the 59 accidents,

there are 28 serious injuries, 27 cases with one person death, 3 cases with two persons death and 1 case with three persons death. According to the HFACS factors, we get descriptive statistics of the accident frequency, as shown in Table 1.

- i) Resource management,
- ii) Organizational process,
- iii) Supervisory violations,
- iv) Failed to correct a known problem,
- v) Personal quality,
- vi) Crew management,
- vii) Operating environment,
- viii) Perceptual and decision error,
- ix) Skill-based errors,
- x) Operation violations.

In Table 1, of the 59 cases, 47.46% severe injury, 45.76% one death, 5.08% two deaths, and 1.70% three deaths.

Super parameter calculation: Super parameter can be calculated by formula (12):

$$\theta_{i,j} = \exp(x_i' \beta_{i,j}) \tag{12}$$

Let $x_i \in X \subset R^p$ be an array in the $[p, 1]$ interval, x_i is the full rank matrix (n, p) of the explanatory variables of the observed samples and $\beta_{i,j}$ measures the influence of probability distribution of the explanatory variable x_i to the accident severity.

The value of x_i is: x_i = the sum of edge frequency of the factors in the accidents/1000. Where, the number of 1000 means that there are ten factors and the edge frequency is expressed as a percentage.

The edge frequency of the influence factors are given in Table 1. According to the inclusion relationship in Table 1, we can obtain the values of $x_i, \beta_{i,j}$ and $\theta_{i,j}$. ($i=1,2,\dots,59$; $j=1,2,3,4$). For example, x_i corresponding to accident 1 is shown as follows:

$$x_1 = (55.93+67.8+62.71+52.54+74.58+62.71+50.85+40.68+83.05+66.10)/1000=0.62$$

The value of $\beta_{i,j}$ is:

$\beta_{i,j}$ = The sum of edge frequency of the factors corresponding to the injury type j in the accidents/1000.

For example, $\beta_{1,j}$ represents the sum of edge frequency of the factors corresponding to the injury type j in the accident 1. The values of $\beta_{1,j}$ in accident 1 are shown as follows:

Table 1. All HFACS factors' edge frequency.

Factors	Severe injury	One death	Two deaths	Three deaths	Edge frequency
(1)	50	42.31	7.69	0	44.07
Not (1)	45.45	48.49	3.03	3.03	55.93
(2)	52.5	40	5	2.5	67.8
Not (2)	36.84	57.9	5.26	0	32.2
(3)	54.55	40.91	4.54	0	37.29
Not (3)	43.24	48.65	5.41	2.7	62.71
(4)	45.16	51.61	3.23	0	52.54
Not (4)	50	39.29	7.14	3.57	47.46
(5)	60	33.33	6.67	0	25.42
Not (5)	43.18	50	4.55	2.27	74.58
(6)	54.54	40.91	4.55	0	37.29
Not (6)	43.24	48.65	5.41	2.7	62.71
(7)	40	53.33	6.67	0	50.85
Not (7)	55.17	37.93	3.45	3.45	49.15
(8)	54.17	45.83	0	0	40.68
Not (8)	42.86	45.71	8.57	2.86	59.32
9()	30	60	10	0	16.95
Not (9)	51.02	42.86	4.08	2.04	83.05
(10)	50	45	5	0	33.9
Not (10)	46.15	46.15	5.13	2.57	66.1
edge frequency	47.46	45.76	5.08	1.7	

$$\beta_{1,1} = (45.45+52.50+43.24+45.16+43.18+43.24+40.00+54.17+51.02+46.15) / 1000 = 0.46$$

$$\beta_{1,2} = (48.49+40.00+48.65+51.61+50.00+48.65+53.33+45.83+42.86+46.15) / 1000 = 0.48$$

$$\beta_{1,3} = (3.03+5.26+5.41+3.23+4.55+5.41+6.67+0+4.08+5.13) / 1000 = 0.04$$

$$\beta_{1,4} = (3.03+2.50+2.70+0+2.27+2.70+0+0+2.04+2.57) / 1000 = 0.02$$

The calculation of prior probability and posterior probability

Based on the results of the super parameter calculation, the formulas 2 and 6, we calculated prior probability and posterior probability of the 59 accidents.

Calculation results show the prior and posterior distribution of the accident samples. The difference among the prior distributions is not obvious, which indicates the researchers could not easily find out the occurring law of serious accidents based on the empirical prior distribution. But the posterior distribution makes up

this shortage; it dramatically reflects the difference between the serious accidents and other accidents. For example, the number 32, 34, 39 and 46, their posterior distribution value is small, and the difference value between the posterior distribution and the prior distribution is negative, this feature shows that using Bayesian method to analyze the accident cause is feasible. According to the descending order of standard deviation, we get Table 2.

In accordance with the value of standard deviation, we sort the factors in descending order, and get the order of importance degree of the various factors in Table 2. It is the order of factors of the accident severity degree based on Bayesian theory under the new HFACS framework. To sum up the above 59 accidents, we may know the sample size is small, the accident is different from the experiment and is unrepeated, and the accident analysis is strong subjective. Such feature is suitable for the Bayesian statistical method.

The calculation and analysis of expectation

The expectation of each factor is shown in Table 3. We use the data in Table 3 to draw the broken line chart shown as Figure 2 for further analysis. We can draw the following conclusions from Figure 2: the expectations of severe injury accidents are almost over 0.3, "(5) personal quality", "(3) supervisory violations", "(6) crew

Table 2. The priori probability, posteriori probability and standard deviation of each factor by the descending order of standard deviation.

Factors	Priori probability	Posteriori probability	Standard deviations	
(8)	9.9676	16.8422	4.861076	1
(4)	9.9966	16.3475	4.490764	2
(6)	9.7212	15.9816	4.426771	3
(3)	9.6343	15.8303	4.381234	4
(10)	9.6773	15.7449	4.290441	5
Not (2)	9.6692	15.6051	4.197315	
(5)	9.5569	15.4921	4.19682	6
Not (9)	10.0507	15.9222	4.151777	
Not (7)	9.8722	15.7315	4.143151	
(1)	9.6939	15.5409	4.134453	7
Not (1)	10.1376	15.9145	4.084885	
Not (5)	10.0734	15.8378	4.076046	
(7)	10.0096	15.7677	4.071592	8
(2)	10.0717	15.8187	4.063743	9
Not (10)	10.0779	15.7525	4.012548	
Not (3)	10.1251	15.7021	3.943535	
Not (6)	10.0734	15.6121	3.916452	
(9)	9.4096	14.9054	3.886117	10
Not (4)	9.8817	15.0883	3.681622	
Not (8)	9.9245	15.0009	3.589557	

Table 3. The expectations of the influencing factors.

Factors	Severe injury	One death	Two deaths	Three deaths
(1)	0.3193	0.2982	0.1896	0.1793
(2)	0.3251	0.3089	0.1919	0.1839
(3)	0.3283	0.3026	0.1993	0.1875
(4)	0.3106	0.2924	0.1946	0.1852
(5)	0.3383	0.313	0.1951	0.1913
(6)	0.3279	0.2849	0.1914	0.1796
(7)	0.3009	0.3067	0.1944	0.1791
(8)	0.3283	0.2947	0.1908	0.1871
(9)	0.3194	0.3181	0.1833	0.1797
(10)	0.319	0.2561	0.1827	0.179

management”, and “(8) perceptual and decision error”, these four factors are prone to cause serious injured; the expectations of one death accidents evenly distribute on both sides of 0.3, the value of “(10) operation violations” is significantly lower than other factors, but the values of “(9) skill-based errors”, “(5) personal quality” and “(2) organizational process” are relatively higher; in the cases of two deaths accidents, “(3) supervisory violations”, “(4) failed to correct a known problem” and “(5) personal quality” most easily induce the accidents; and in the cases of three deaths accidents, “(5) personal quality”, “(3) supervisory violations” and “(8) perceptual and decision

error” most easily result in the accidents.

The calculation and analysis of variance

The posterior variance of each factor is shown in Table 4. We use the data in Table 4 to draw the posterior variance broke line chart of the accident severity of the factors, shown as Figure 3. We can draw the following conclusions from Figure 3: the posteriori variance of each factor to the four types of severity smoothly distributes at a mean value. The severe injury accidents and the one

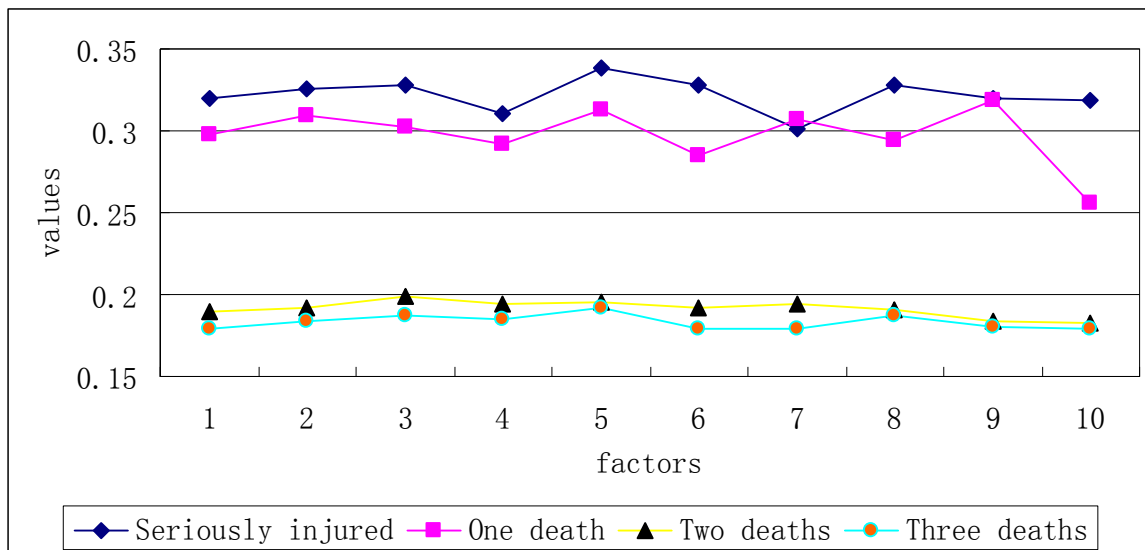


Figure 2. The expectation broke line chart of the ten factors causing the four types of accidents.

Table 4. The posterior variances of the influencing factors.

Factors	Severe injury	One death	Two deaths	Three deaths
(1)	0.0317	0.0308	0.0237	0.0224
(2)	0.0318	0.0305	0.023	0.0224
(3)	0.0323	0.0307	0.0234	0.0225
(4)	0.0311	0.0317	0.0229	0.0222
(5)	0.0329	0.0299	0.0238	0.0226
(6)	0.0322	0.0307	0.0233	0.0225
(7)	0.0305	0.0319	0.0233	0.0221
(8)	0.0321	0.0311	0.0226	0.0223
(9)	0.0298	0.0328	0.0243	0.0227
(10)	0.0317	0.0311	0.0234	0.0225

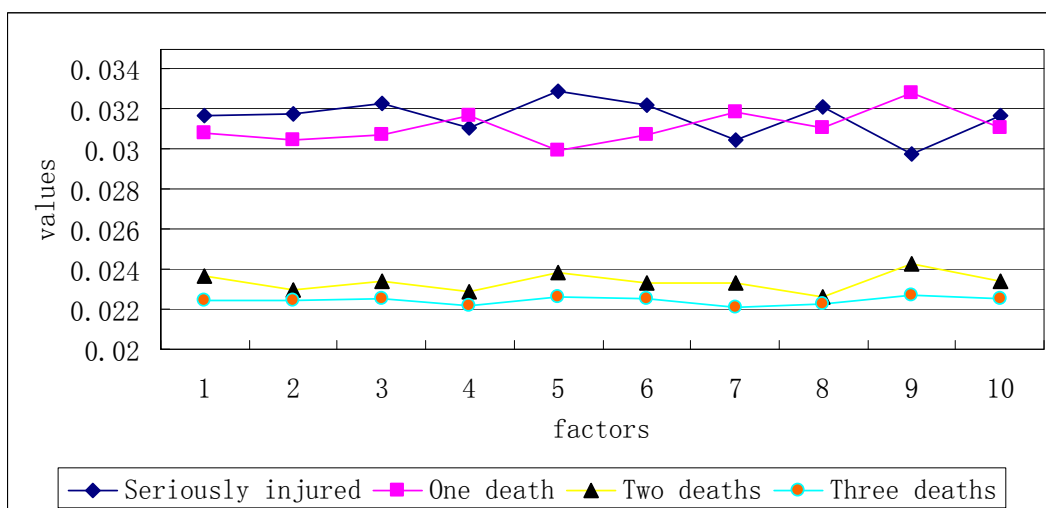


Figure 3. The variance broke line chart of the ten factors causing the four types of accidents.

death accidents are significantly different from zero in level 4%, and the two deaths accidents and the three deaths accidents are significantly different from zero in level 3%, which indicate the distribution consistency of accident severity of the samples is good.

Conclusion

Through the comparative analysis of the results calculated by the traditional statistical method and the Bayesian statistical method, we can find that the results calculated by the traditional statistical methods laid emphasis on the organizational and supervisory levels, and ignored the issue of workers. The results of Bayesian statistical method, however, appear more "humanization". The proportion of the judgment and the physical state of the workers increases. We have referred in the above analysis that lack of the capacity of determining hazards is the direct reason in many accidents. It also conforms to China's actual situation: lack of professional skills training of the workers, lower average educational level. Meantime, in the results of Bayesian statistical analysis, the proportions of resource management, inadequate supervision and supervisory violations are relatively higher, which indicates the defects of management rules and malfeasances of managers and supervisors play an important role in causing of accident. Such result is the same with it obtained by the classical statistical analysis.

So, Bayesian statistics method can learn from experience, combine historical information and sample likelihood function together, make a set of statistical method more flexible, more visual and easily understood than classical statistical method. Bayesian statistics method is widely used in the application of measurement model. Especially in the fewer samples, point estimate and interval estimate can get more accurate results than classical statistical method; secondly, we use Bayesian posterior distribution to take account of the losses caused by type one and type two errors, therefore it is more of practical than classical statistical method; in addition, in dealing with the problems of redundant parameters, Bayesian statistics method can directly integral off superfluous parameters in posterior density, which is far more convenient than classical statistical method.

Conflict of Interests

The author(s) have not declared any conflict of interests.

ACKNOWLEDGEMENT

This paper is supported by the national natural science fund project of China (50909045, 51079078).

REFERENCES

- Gwet K (2002b). Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. In: *Statistical Methods for Inter-rater Reliability Assessment*.
- Leonard T (1977). A Bayesian approach to some multinomial estimation and pretesting problems, *J. Am. Stat. Association* 72:869-874.
- de Lapparent M (2006). Empirical Bayesian analysis of accident severity for motorcyclists in large French urban areas. *Accid. Anal. Prev.* 38:260-268. <http://dx.doi.org/10.1016/j.aap.2005.09.001>; PMID:16280119
- Shappell SA, Wiegmann DA (2001). Applying reason: the human factors analysis and classification system (HFACS). *Hum. Factors Aerospace.* 1:59-86.
- Wiegmann DA, Shappell SA (1997). Human error analysis of post accident data: applying theoretical taxonomies of human error. *Int. J. Aviat. Psychol.* 7:67-81. http://dx.doi.org/10.1207/s15327108ijap0701_4