Reappraisal on the estimation of parameters of the hybrid lognormal distribution

Shigeru KUMAZAWA
Former JAERI (Predecessor of JAEA)

ABSTRACT

The hybrid log-normal (HLN) distribution is the probability distribution of positive variates X that the transformation hyb(ρX) = ρX + ln(ρX) are normally distributed with the mean E[hyb(ρX)] and the variance V[hyb(ρX)]. The HLN distribution model has been applied to various data of biology and social statistics as well as radiological statistics with the multiple linear regression: normal rank z_i vs variables (x_i, ln x_i) of descending ordered data. The paper presents a single regression method of the HLN distribution model using the EXCEL functions LINEST and SOLVER, to maximize the R-squared with respect to ρ (> 0), including the calculation of standard error. A similar approach is feasible to expand the single regression method of data plotted on the hybrid-hybrid section paper as a comprehensive section paper, which hybridizes four popular section papers (linear-linear, linear-log, log-log and log-linear) with five additional new section papers (linear-hybrid, hybrid-log, log-hybrid, hybrid-linear and hybrid-hybrid) that serve to connect the conventional section papers smoothly.

Keywords: hybrid lognormal distribution, simple regression, hybrid-hybrid section paper.

INTRODUCTION

The Log-normal distribution was introduced by Galton and McAlister in 1879 to provide the replacement of normal distributions that appeared the skewed distribution in vital and social statistics. Since 1965 the log-normal distribution has been applied to interpret the characteristics of the distribution of annual doses incurred by workers under the regulatory control (Gale, 1965). In 1977 the United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) reported the comprehensive results of the lognormal analysis of occupational annual dose statistics over world-wide nations, which was essential for establishing the system of dose limitation in the 1977 Recommendations of the International Commission on Radiological Protection (ICRP). The statistics of occupational annual doses under the regulatory control, however, has been pointed out to deviate from the often-observed lognormal distribution, because of the effect of dose limits or levels of radiation control (Gale, 1965; UNSCEAR, 1977).

In 1980 we proposed the hybrid lognormal distribution to replace the skewed lognormal distribution in radiation protection statistics by introducing an exposure control parameter ρ with the inverse unit of dose that should explain the degree of active feedback dose control depending on the magnitude of exposure (Kumazawa, Shimazaki and Numakunai, 1980; Kumazawa and Numakunai, 1981). Developing the computer package of the HLN analysis, the HLN distribution model was applied to complete the report (EPA- 520/1-84-005, 1984) on the U.S. occupational exposure to review comprehensively for the year 1980 including the trends for the years 1960-1985, as one of revision tasks on the Federal radiation protection guidance for occupational exposure, approved by the President Reagan in January 27, 1987.

In statistics the hybrid lognormal distribution was discussed to clarify the characteristics of the distribution as well as to verify the genesis of HLN distributions with the martingale central limit theorem, including examples to show the feasibility of wide application (Kumazawa and Ohashi, 1986). Especially in the case of the lognormal suffered by a constraint of reducing the more the occurrence of data against the larger-value, the HLN distribution is adequate to interpret the data in terms of risk control or some constraint of underlying phenomena incurred by data.

It is natural to use the multiple linear regression for the HLN analysis in practice: the regression model is defined as $z_i = \alpha + \beta x_i + \gamma \ln x_i + \varepsilon_i$ for the ascending ordered data $\{x_i | i = 1, n\}$ where the normal cumulative distribution function $\Phi(z_i) \approx \text{prob}\{X \leq x_i\}$, which is $(i - 0.375)/(\text{n} + 0.25)$ (Blom, 1958). In 1945 this regression model was first used to analyze the particle size distribution of products ground in tube mill (Fagerholt, 1945; A. Hald, 1948). We use the similar model for the HLN analysis with the awareness of the positive-value for the active control parameter $\rho = \beta / \gamma$ per unit dose or unit quantity of data.

Based on the HLN genesis, it is adequate to introduce the hybrid function $\text{hyb}(x) = x + \ln(x)$ and to define the HLN distribution as $\text{hyb}(\rho X) \sim N(\mu, \sigma^2)$ where μ is $E[\text{hyb}(\rho X)]$ and σ^2 is $V[\text{hyb}(\rho X)]$. The function $\text{hyb}(\rho X)$ is almost the same to the logarithmic function $\ln(\rho X)$ in the region of $\rho X < 0.1$ and almost proportional to the linear function of ρX in the region of $\rho X > 5$ but it is neither logarithmic nor linear for ρX between 0.1 and 5 or so. The region of ρX from 0.1 to 5 has the special significance to concentrate our efforts on risk control of occupational exposure to ionizing radiation. To find the effective region of risk control is to estimate the parameter ρ with the standard error reasonably.

The paper discusses the simple regression model on the HLN distribution using the hybrid function: putting $h_i = \text{hyb}(\rho x_i)$ of ascending ordered data $\{x_i | i = 1, n\}$ and their ranks $\{z_i | i = 1, n\}$of the normal probability, the regression model is discussed as $h_i = \mu + \sigma z_i + \varepsilon_{hi}$ or $z_i = \alpha' + \gamma h_i + \varepsilon_{zi}$. The former contains the parameter ρ in the dependent variable but ρ is assumed that it might be estimated separately in advance. This is accomplished by maximizing the R-squared for ρ using the EXCEL functions LINEST and SOLVER, including the adjustment to replace degrees of freedom from n - 2 to n - 3. The calculation of the standard error of estimate ρ is derived from the F test method of the multiple regression relating to R^2 (Okuno, Kume, Haga and Yoshizawa, 1977).

In radiological statistics a pair of positive-value data $\{u_i, v_i | \, i = 1, n\}$ both often varies ranging from the logarithmic to the linear region continuously connected by the hybrid interface region because of the underlying phenomena due to stochastically multiplicative and additive interactions: e.g., chromosome aberrations vs radiation dose, etc. Therefore this paper presents, in discussion section, an introductory method of the simple regression model $\text{hyb}(\nu \cdot v_i) = \alpha + \beta \, \text{hyb}(\tau \cdot u_i) + \varepsilon_{2Di}$ assumed as the known parameters τ and ν but simultaneously solving by the EXCEL functions LINEST and SOLVER to maximize $R^2(\tau, \nu)$ by τ and ν with the replacement of degrees of freedom from n - 2 to n - 4. Some examples are given to clarify the simple regression of the HLN distribution and the hybrid-hybrid section paper that contains basically nine types of linear relationships.

METHOD

The density function of the hybrid lognormal distribution is given as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\left(\rho + \frac{1}{x}\right)\exp\left[-\frac{(\mathrm{hyb}(\rho x) \ -\mu)^2}{2\sigma^2}\right] \qquad (x, \rho, \sigma > 0). \tag{1}$$

Putting $t = \rho x$, we have $f(t)dt = f(x)dx = \phi(z)dz$ where $\phi(z)$ is the density function of the standard normal cumulative function $\Phi(z)$ and $\mathrm{hyb}(\rho x) = \mathrm{hyb}(t) = t + \ln(t) = \mu + \sigma z$ or $z = (\mathrm{hyb}(t) - \mu)/\sigma$. Then the simple regression model is

$$h_i = \mathrm{hyb}(t_i) = \mu + \sigma z_i + \varepsilon_{hi}, \qquad i = 1, 2, \cdots, n \tag{2}$$

where $\mathrm{E}[\varepsilon_{hi}] = 0, \mathrm{V}[\varepsilon_{hi}] = \sigma_h^2$, $\mathrm{Cov}\left[\varepsilon_{hi}, \varepsilon_{hj}\right]_{i \neq j} = 0$ and $\varepsilon_{hi} \sim N(0, \sigma_h^2)$. For the known parameter ρ the simple regression model can be solved in the usual way. The unknown parameter ρ needs to discuss on the simple regression in terms of the estimate of ρ with the standard error because of $h_i = \mathrm{hyb}(\rho x_i)$.

For the comparison the multiple linear regression model and another simple regression model are given:

$$z_i = \alpha + \beta x_i + \gamma \ln x_i + + \varepsilon_i \quad (\beta, \gamma > 0, \ \rho = \beta/\gamma\ ), \tag{3}$$

$$z_i = \alpha' + \gamma h_i + \varepsilon_{zi} \quad (\ \rho, \gamma > 0, \ \ h_i = \mathrm{hyb}(\rho x_i)\ ). \tag{4}$$

To estimate the parameter ρ there are several approaches, e.g. a Bayesian procedure of the estimation of the hybridization parameter ρ for a specific prior distribution for ρ, μ and σ (Groer and Uppuluri, 1991) and the maximum likelihood estimation of the parameters of the hybrid lognormal distribution (Sont, 1991), etc. As a practical way this paper is to reasonably solve the simple regression model in Equation (2) with the unknown parameter ρ. This attains to maximize the coefficient of determination R^2 of Equation (2) for ρ.

The R-squared of Equation (2) is derived as follows:

$$R^2(\rho) = \frac{S_{zh}(\rho)^2}{S_{hh}(\rho)\, S_{zz}} = \frac{\left(\rho S_{xz} + S_{yz}\right)^2}{\left(\rho^2 S_{xx} + \rho S_{xy} + S_{yy}\right) S_{zz}}, \tag{5}$$

where $S_{hh}(\rho) = \sum_n {H_i}^2$, $S_{zz} = \sum_n {Z_i}^2$, $S_{zh} = \sum_n Z_i H_i$, $S_{xx} = \sum_n {X_i}^2$ $S_{yy} = \sum_n {Y_i}^2$, $S_{xy} = \sum_n X_i Y_i$, $S_{xz} = \sum_n X_i Z_i$ and $S_{yz} = \sum_n Y_i Z_i$, and deviations $H_i = h_i - E[h_i]$, $Z_i = z_i - E[z_i]$, $X_i = x_i - E[x_i]$ and $Y_i = \ln x_i - E[\ln x_i]$.

The estimation of the parameter ρ is the maximum of $R^2(\rho)$ with respect to ρ:

$$\frac{\partial R^2(\rho)}{\partial \rho} = 0 \qquad \therefore \hat{\rho}_{R^2} = \frac{S_{xy}S_{yz} - S_{xz}S_{yy}}{S_{xy}S_{xz} - S_{xx}S_{yz}} = \frac{A_{zx}}{A_{zy}}, \tag{6}$$

where $A_{zx}\ and\ A_{zy}$ are the cofactors of $S_{zx}\ and\ S_{zy}$, respectively, of the following matrix S:

$$S = \begin{pmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{yx} & S_{yy} & S_{yz} \\ S_{zx} & S_{zy} & S_{zz} \end{pmatrix} \tag{7}$$

The estimate $\hat{\rho}_{R^2}$ in Equation (6) is equal to the estimate of ρ to maximize the R-squared as well as to minimize the sum of squared errors (residuals) in Equations (3) and (4) but it is different to the estimate $\hat{\rho}_h$ to minimize the sum of squared errors $S_{he}(\rho)$ in Equation (2). However, the sum of squared errors $S_{ze}(\rho)$ in Equation (4) and $S_{he}(\rho)$ in Equation (2) satisfy the following equation:

$$1 - R^2(\rho) = \frac{S_{he}(\rho)}{S_{hh}(\rho)} = \frac{S_{ze}(\rho)}{S_{zz}}. \tag{8}$$

For the estimate $\hat{\rho}_{R^2}$ as the maximum of $R^2(\rho)$ in Equation (2), Equation (8) results in:

$$1 - R^2(\hat{\rho}_{R^2}) = \frac{S_{he}(\hat{\rho}_{R^2})}{S_{hh}(\hat{\rho}_{R^2})} = \frac{S_{ze}(\hat{\rho}_{R^2})}{S_{zz}} = \frac{|S|}{A_{zz}S_{zz}}, \tag{9}$$

where $|S|$ is the determinant of S and $A_{zz} = S_{xx}S_{yy} - {S_{xy}}^2$, the cofactor of the element S_{zz} of matrix S.

The residual variance $V_{ze}(\hat{\rho}_{R^2}) = S_{ze}(\hat{\rho}_{R^2})/(n-3)$ provides the standard error $se(\hat{\rho}_{R^2})$ as $S_{ze}(\rho) = S_{ze}\big(\hat{\rho}_{R^2} + se(\hat{\rho}_{R^2})\big) = S_{ze}(\hat{\rho}_{R^2}) + V_{ze}(\hat{\rho}_{R^2})$ and $R^2(\rho) = R^2\left(\hat{\rho}_{R^2} + se(\hat{\rho}_{R^2})\right)$ is obtained as follows:

$$\frac{S_{ze}(\rho)}{S_{zz}} = \frac{S_{ze}(\hat{\rho}_{R^2}) + V_{ze}(\hat{\rho}_{R^2})}{S_{zz}} = \frac{S_{ze}(\hat{\rho}_{R^2})}{S_{zz}}\frac{n-2}{n-3} = \frac{|S|}{A_{zz}S_{zz}}\frac{n-2}{n-3} = 1 - R^2(\rho)$$

$$R^2(\rho) = \frac{\big(\rho S_{xz} + S_{yz}\big)^2}{\big(\rho^2 S_{xx} + 2\rho S_{xy} + S_{yy}\big)S_{zz}} = 1 - \frac{|S|}{A_{zz}S_{zz}}\frac{n-2}{n-3} = \frac{c_z}{S_{zz}}. \tag{10}$$

Equation (10) is solved as the quadratic equation with respect to ρ:

$$\rho^2\big(c_z S_{xx} - {S_{xz}}^2\big) + 2\,\rho\big(c_z S_{xy} - S_{xz}S_{yz}\big) + c_z S_{yy} - {S_{yz}}^2 = 0,$$

$$\rho = \frac{-\big(c_z S_{xy} - S_{xz}S_{yz}\big) \pm \sqrt{c_z A_{zz}\{S_{zz}R^2(\hat{\rho}_{R^2}) - c_z\}}}{\big(c_z S_{xx} - {S_{xz}}^2\big)},$$

$$\therefore\ se(\hat{\rho}_{R^2}) = \frac{\rho_2 - \rho_1}{2} = \frac{\sqrt{c_z A_{zz}\{S_{zz}R^2(\hat{\rho}_{R^2}) - c_z\}}}{\big(c_z S_{xx} - {S_{xz}}^2\big)}. \tag{11}$$

The range between ρ_1 and ρ_2 of the simple regression in Equation (2) is given in terms of $R^2(\rho)$ for maximization or $1 - R^2(\rho)$ for minimization. Thus, $\rho = \hat{\rho}_{R^2} + se(\hat{\rho}_{R^2})$ satisfies the following equation:

$$R^2(\rho) = R^2(\hat{\rho}_{R^2}) - \frac{1 - R^2(\hat{\rho}_{R^2})}{n-3} \quad or \quad 1 - R^2(\rho) = 1 - R^2(\hat{\rho}_{R^2}) + \frac{1 - R^2(\hat{\rho}_{R^2})}{n-3} \tag{12}$$

The standard error of $\hat{\rho}_{R^2}$ is to calculate as $(\rho_2 - \rho_1)/2$ because $\hat{\rho}_{R^2} - \rho_1$ is not always equal to $\rho_2 - \hat{\rho}_{R^2}$. In the case of distribution characteristics to be purely lognormal-dominant or normal-dominant, the HLN analysis can be performed to select the estimate $\rho = 0.1/x_n$ for the negative or too much small value of the estimate ρ (lognormal-dominant) or to select the estimate $\rho = 5/x_1$ for the infinitive or too much large value of the estimate ρ (normal-dominant). Thus, the HLN model provides the analysis approach of data ranging from the lognormal-dominant to the normal-dominant systematically.

RESULTS OF ANALYSIS

Using the well-known data on the number of words per sentence in 60 sentences taken from a certain of Toynbee's "A Study of History" (Wilks, 1948), the left panel in Figure 1 shows the graph of $R^2(\rho)$ so that it should change from the log-normal ($\rho \to 0$) to the normal ($\rho \to +\infty$) via the hybrid lognormal model by increasing the value of ρ. $R^2(\rho \to 0)$ is larger than $R^2(\rho \to +\infty)$ but the largest is $R^2(\rho = \hat{\rho}_{R^2})$. The right panel is the enlarged view of the bold curve on the left panel to show how the standard error $\mathrm{se}(\hat{\rho}_{R^2})$ is decided based on Equation (12).

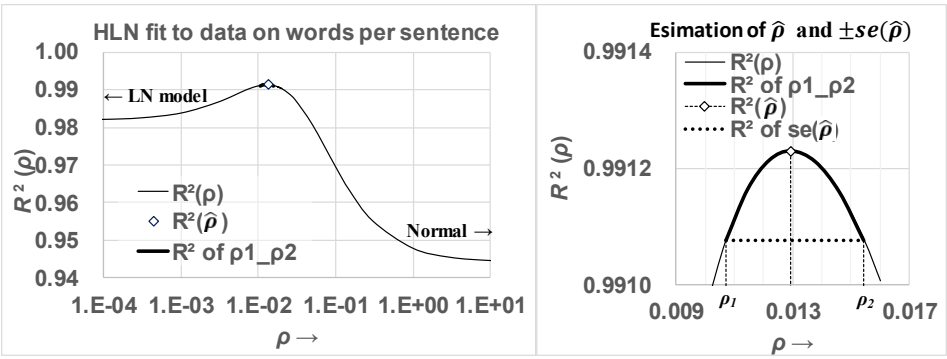


Figure 1. The graph of $R^2(\rho)$ in Equation (2) by ρ and the interval of ρ for $\hat{\rho}_{R^2} \pm$ se $(\hat{\rho}_{R^2})$.
Left panel is $R^2(\rho)$ ranging from the LN to the Normal model, and right panel shows how to decide se $(\hat{\rho}_{R^2})$: $\hat{\rho} = \hat{\rho}_{R^2}$.

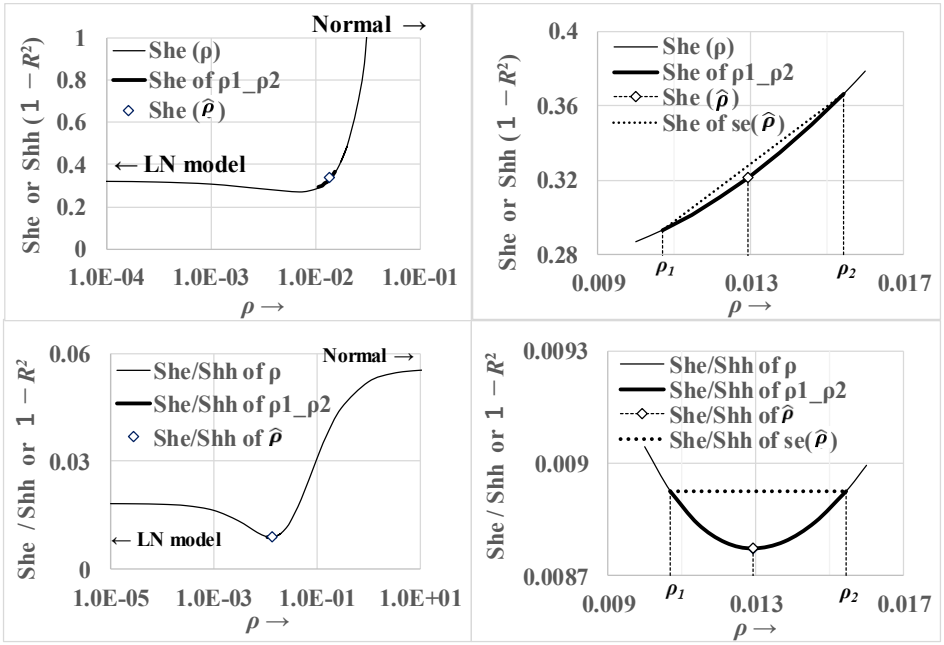


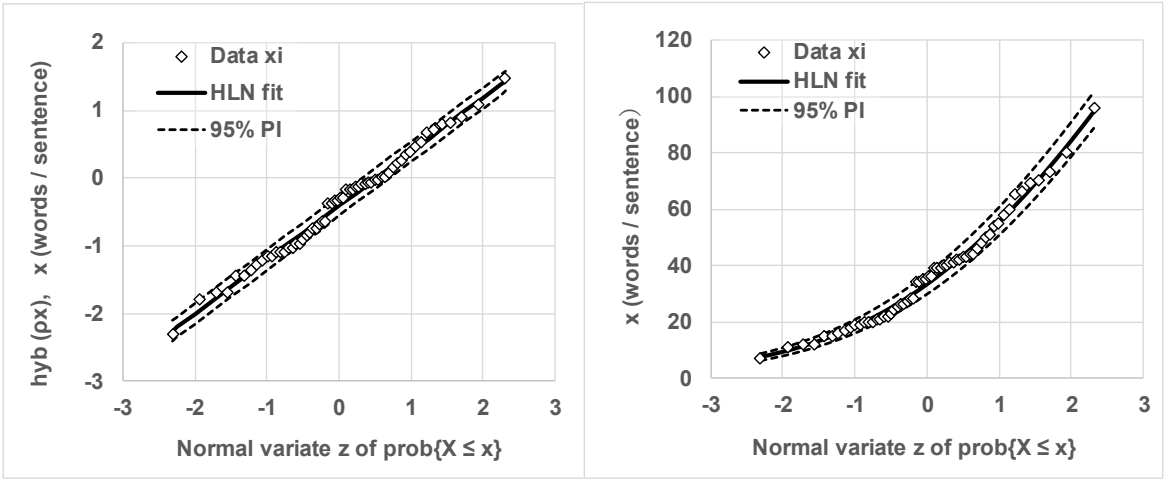
Figure 2. Other options to calculate the standard error of the estimate $\hat{\rho}_{R^2}$.
Upper panel: based on $S_{he}(\rho)$, not practicable; lower panel: based on 1 - $R^2(\rho)$, another good option.

Figure 2 shows other options to calculate the standard error $\mathrm{se}(\hat{\rho}_{R^2})$. The upper panel shows a method based on minimizing $S_{he}(\rho)$ of Equation (2), which is somewhat complicated to use. The lower panel shows another good method based on minimizing $1 - R^2(\rho)$, namely $S_{he}(\rho)/S_{hh}(\rho)$.

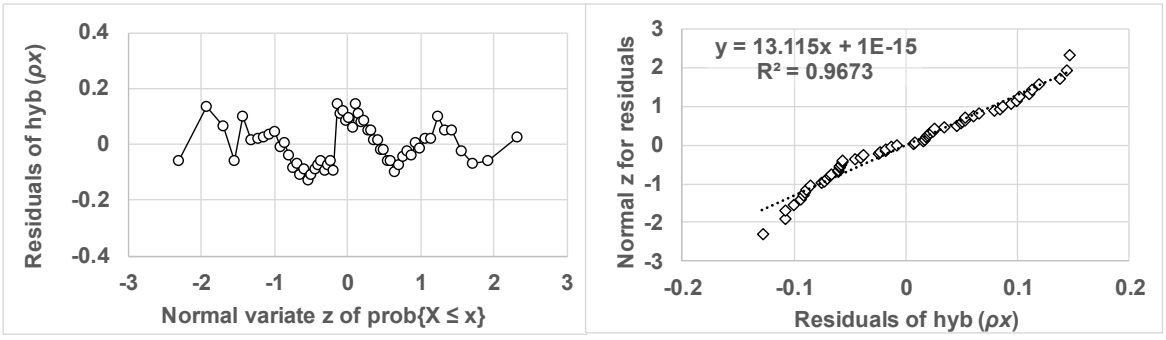
These examples confirm that the above-mentioned method is valid and reasonable to estimate ρ with the

standard error as the simple regression model. According to the method, the resultant statistics on the data of words per sentence are $\hat{\mu} = 0.4084 \pm 0.0096$, $\hat{\sigma} = 0.7972 \pm 0.00985$, $\hat{\rho} = 0.01294 \pm 0.00236$, and $R^2 = 0.9912$. Figure 3 (a) shows the good fitting to the data all of which are within the 95% prediction interval. The number of words per sentence might be the result of making sentences to clarify the thought considering the readability. Interpreting the graph, Toynbee would write the sentences lognormally for words / sentence less than 8 but hybrid lognormally for words / sentence above 44. Thus, there is a possibility to avoid a long sentence in Toynbee. Figure 3 (b) shows approaching the normality in the longer sentence. The normality of the residuals is pretty good shown in Figure 2 (d).

The HLN distribution tells us how to cope with increasing risks and the effective range of risk control in terms of the control parameter ρ. For example, the statistical fluctuation in running speed is important for football players to maintain flexible mobility as well as to avoid exhaustion. The distribution of running speeds during the game shows the HLN distribution characteristics. Eyes speed data during viewing still / moving pictures are also hybrid lognormally distributed. Thus, somewhat similarities exit like the radiation dose in terms of risk management to gain the benefit during a small risk but to avoid the significant risk.



(a) HLN probability plots of data of words/sentence (b) Normal probability plots of data



(c) Residuals: $\varepsilon_{hi} = \text{hyb}(\hat{\rho}_{R^2} x_i) - \hat{\mu} - \hat{\sigma} z_i$ vs z_i (d) Probability plots of ascending ordered $\{\varepsilon_{hi}\}$.

Figure 3. The HLN analysis of 60 ascending ordered data on the number of words per sentence.

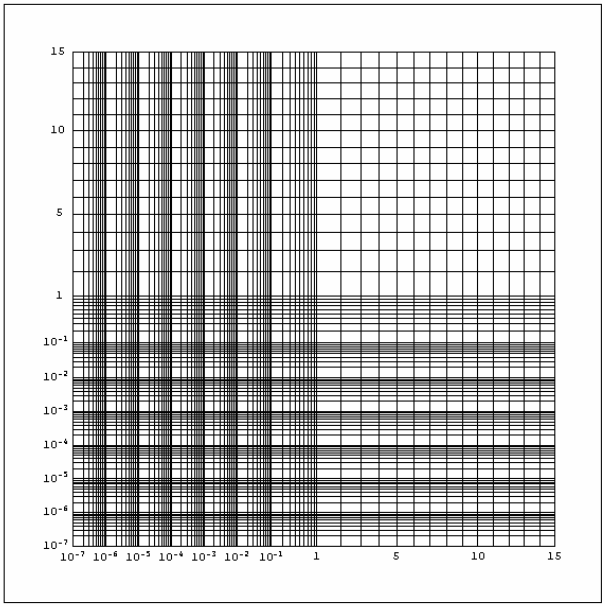
DISCUSSION

The hybrid lognormal distribution is relating to the multiple linear regression method by Fagerholt (1945) and Hald (1948). We, as stuff of former JAERI, developed it so that it should reasonably interpret how to cope with radiation exposure risk with some uncertainty. Such risk control issues are common in environmental exposure statistics, health risk statistics, economic/social statistics, etc. Therefore, the HLN distribution should be discussed widely by statisticians because of the issues still unsolved.

The variation of the HLN distribution needs for X > a > 0 as $\mathrm{hyb}\big(\rho(X-a)\big) \sim N(\mu, \sigma^2)$ and for $0 < a < X < b$ as $\mathrm{hyb}(\rho\,(X-a)/(b-X)) \sim N(\mu, \sigma^2)$, the latter of which calls "the hybrid S_B distribution" like the Johnson's S_B distribution. These have been developed to analyze data in radiation protection and nuclear environmental safety assessment. Radiological data relating to the Fukushima DAIICHI accident shows examples of the variation to be the HLN distribution. The systematic random number generation for the radiological uncertainty analysis has been developed a system of normal family distributions that contains the lognormal, hybrid lognormal and positive-value normal distributions in the region defined $X > 0$, $X > a > 0$ and $0 < a < X < b$. All of these use the simple regression analysis.

The data relating to radiation emission processes, radiation interactions with matters, and radiation dose to man in various exposure situations are resultant effects due to stochastically multiplicative and additive interactions. The data other than radiation protection shows the similar effects with simultaneously multiplicative and additive uncertainty. Depending on the dominancy between multiplicative and additive components, the variation characteristics changes to be purely logarithmic, to be logarithmic in the lower side but linearly in the upper side and to be purely linear. The hybrid function found from the development the HLN distribution covers the wide range of variations from the logarithmic region to the linear region continuously connected both via the hybrid interface region.

If the hybrid function expresses as a scale like the logarithmic scale, the new scale is defined, called "the hybrid scale." It consists of three major regions, the logarithmic for $\rho x < 0.1$ of $\mathrm{hyb}(\rho x)$, the truly hybrid for $0.1 \le \rho x \le 5$ of $\mathrm{hyb}(\rho x)$ and the linear region for $\rho x > 5$ of $\mathrm{hyb}(\rho x)$. Suppose the ascending ordered data of positive value $\{x_i | i = 1, n\}$, all $\mathrm{hyb}(\rho x_i)$ are in the logarithmic region for $\rho < 0.1/x_n$ and in the linear region for $\rho > 5/x_1$. Thus, the hybrid scale can express the log-, truly-hybrid- and linear-region of data with $\mathrm{hyb}(\rho x)$ by selecting a proper value of ρ. Applying the hybrid scale to the vertical and horizontal axes graduated with a hybrid scale, a new section paper is provided (Figure 4). This is called "the hybrid-hybrid section paper" like the term of the log-log one. In radiation protection, quantities often change by orders of magnitude that should be modeled by the logarithmic scale but the magnitude of exposure in control is managed in the same order of magnitude that should be modeled by the hybrid scale gradually approaching the linear scale. The hybrid-hybrid section paper contains nine section paper regions of four conventional ones (linear-linear, linear-log, log-log and log-linear) and five interface ones (linear-hybrid, hybrid-log, log-hybrid, hybrid-linear and hybrid-hybrid). Each of five interface regions is corresponding to five types of new independent section papers (Figure 4 (b), (c)).



(a) A hybrid-hybrid section paper

Nine types of section papers

Semi-Log **linear-log**	**Semi-Hybrid** **linear-hybrid**	**Normal** **linear-linear**
Log-Hybrid **hybrid-log**	**Hybrid-Hybrid** **hybrid-hybrid**	**Semi-Hybrid** **hybrid-linear**
Log-Log **log-log**	**Log-Hybrid** **log-hybrid**	**Semi-Log** **log-linear**

Conventional section papers
linear-linear, linear-log, log-log & log-linear

New section papers: connecting above papers
linear-hybrid, hybrid-log, log-hybrid, hybrid-linear, hybrid-hybrid

(b) Nine types of section papers included in it.

Semi-Log section paper $y = \alpha + \beta \ln(x) + \varepsilon$ $x = \exp[(y - \alpha)/\beta] + \varepsilon'$	Semi-Hybrid section paper $y = \alpha + \beta\, \mathrm{hyb}(\tau x) + \varepsilon$ $x = \tau^{-1}\, \mathrm{cyb}[(y - \alpha)/\beta] + \varepsilon'$	Normal section paper $y = \alpha + \beta x + \varepsilon$ $x = (y - \alpha)/\beta + \varepsilon'$
Log-Hybrid section paper $\mathrm{hyb}(\nu y) = \alpha + \beta \ln(x) + \varepsilon$ $x = \exp[(\mathrm{hyb}(\nu y) - \alpha)/\beta] + \varepsilon'$	Hybrid-Hybrid section paper $\mathrm{hyb}(\nu y) = \alpha + \beta\, \mathrm{hyb}(\tau x) + \varepsilon$ $x = \tau^{-1}\, \mathrm{cyb}[(\mathrm{hyb}(\nu y) - \alpha)/\beta] + \varepsilon'$	Semi-Hybrid section paper $\mathrm{hyb}(\nu y) = \alpha + \beta x + \varepsilon$ $x = (\mathrm{hyb}(\nu y) - \alpha)/\beta + \varepsilon'$
Log-Log section paper $\ln(y) = \alpha + \beta \ln(x) + \varepsilon$ $x = \exp[(\ln(y) - \alpha)/\beta] + \varepsilon'$	Log-Hybrid section paper $\ln(y) = \alpha + \beta\, \mathrm{hyb}(\tau x) + \varepsilon$ $x = \tau^{-1}\, \mathrm{cyb}[(\ln(y) - \alpha)/\beta] + \varepsilon'$	Semi-Log section paper $\ln(y) = \alpha + \beta x + \varepsilon$ $x = (\ln(y) - \alpha)/\beta + \varepsilon'$

(c) Nine types of linear equations for nine types section papers organized by the hybrid function

Figure 4. An example of the hybrid-hybrid section paper that contains nine types of section papers.
This section paper allows any linear relationships among linear, logarithmic, power and exponential functions.
The inverse of hyb(x) is denoted as cyb (x), called "the cyb function" that is the hybrid between exp(x) and x.

Because of the tri-regionality of the hybrid scale as log, hybrid and linear regions, the linear relationship on the hybrid-hybrid section paper sufficiently approximates nine types of linear relationships corresponding to nine types of section papers shown in Figure 4 (c). For data $\{x_i, y_i | i = 1, n\}$ sorted by x_i in ascending order, where y_i does not mean $y_i = \ln(x_i)$ in Equations after (5), a hybrid-hybrid simple regression model is given by:

$$v_i = \alpha + \beta u_i + \varepsilon_{2Di} \quad \leftarrow \quad \mathrm{hyb}(\nu y_i) = \alpha + \beta\, \mathrm{hyb}(\tau x_i) + \varepsilon_{2Di} \quad (i = 1, \cdots, n). \quad (13)$$

For the known parameters τ and ν the simple regression model can be solved in the usual way. For the unknown τ and ν, scaling parameters with the inverse unit of x and y, respectively, the similar approach of the HLN model is applicable to attain the global maximum of $R^2(\tau, \nu)$ using LINEST and SOLVER, with the nuisance parameters α, β and the residual ε_{2Di}. This approach is to find the best linearity of plotting data on the hybrid-hybrid section paper as a comprehensive section paper, despite the linearity of the data ranging across any several types of nine section papers.

The R-squared of Equation (13) is $R^2 = {S_{uv}}^2/S_{uu}S_{vv}$ where S_{uu}, S_{vv} and S_{uv} are the sum of squared deviations and the sum of deviation products for $u_i \; and \; v_i$, respectively, and using $\mathrm{hyb}(\tau x_i)$ and $\mathrm{hyb}(\nu y_i)$,

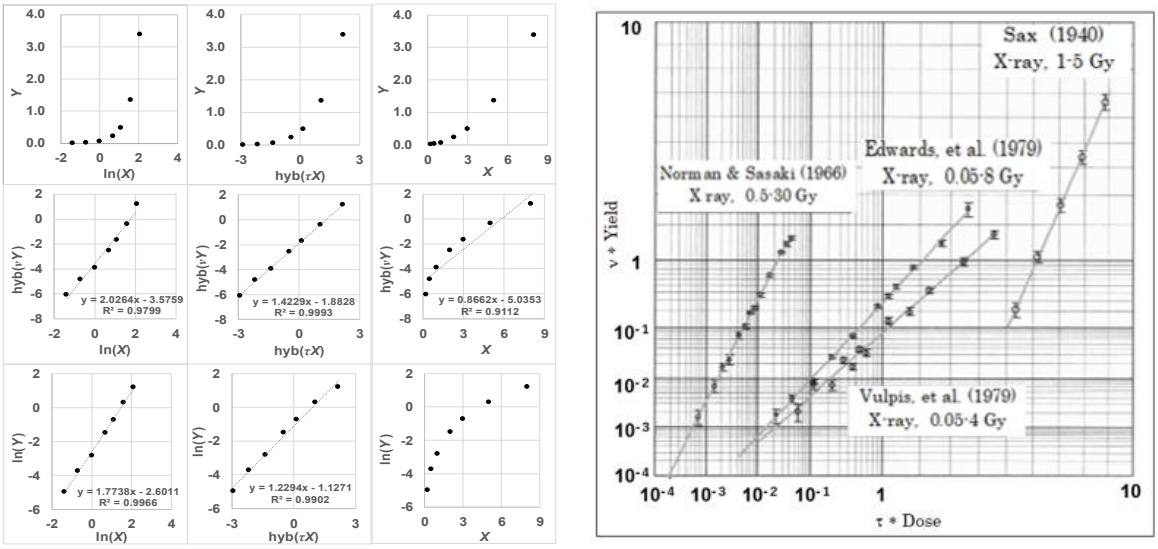
$$R^2 = \frac{{S_{uv}}^2}{S_{uu}S_{vv}} = \frac{\left(\tau\nu S_{xy} + \tau S_{xy_L} + \nu S_{yx_L} + S_{x_Ly_L}\right)^2}{\left(\tau^2 S_{xx} + 2\tau S_{xx_L} + S_{x_Lx_L}\right)\left(\nu^2 S_{yy} + 2\nu S_{yy_L} + S_{y_Ly_L}\right)}, \qquad (14)$$

$$S_{uu} = \tau^2 S_{xx} + 2\tau S_{xx_L} + S_{x_Lx_L} \leftarrow \sum(\tau(x_i - E[x_i]) + \ln(x_i) - E[\ln(x_i)])^2 = \sum(\tau X_i + X_{Li})^2,$$
$$S_{vv} = \nu^2 S_{yy} + 2\tau S_{yy_L} + S_{y_Ly_L} \leftarrow \sum(\nu(y_i - E[y_i]) + \ln(y_i) - E[\ln(y_i)])^2 = \sum(\nu Y_i + Y_{Li})^2,$$
$$S_{uv} = \tau\nu S_{xy} + \tau S_{xy_L} + \nu S_{x_Ly} + S_{x_Ly_L} \leftarrow \sum(\tau X_i + X_{Li})(\nu Y_i + Y_{Li}).$$

Based on equations satisfied $\partial R^2/\partial\tau = 0, \partial R^2/\partial\nu = 0$ and $\partial^2 R^2/\partial\tau\partial\nu = 0$, the estimation of parameters τ and ν are given as follows:

$$\hat{\tau} = \frac{S_{xx_L}S_{yx_L} - S_{xy}S_{x_Lx_L}}{S_{xy}S_{xx_L} - S_{xx}S_{x_Ly}}, \qquad \hat{\nu} = \frac{S_{xy_L}S_{yy_L} - S_{xy}S_{y_Ly_L}}{S_{xy}S_{yy_L} - S_{yy}S_{xy_L}}. \qquad (15)$$

The sum of squared errors (residuals) $S_{2De}(\hat{\tau},\hat{\nu})$ and the residual variance $V_{2De}(\hat{\tau},\hat{\nu}) = S_{2De}(\hat{\tau},\hat{\nu})/f_e$, $f_e = n - 4$, provide the level of reducing R^2 from the maximum $R^2(\hat{\tau},\hat{\nu})$ to $R^2(\hat{\tau},\hat{\nu}) - \left(1 - R^2(\hat{\tau},\hat{\nu})\right)/f_e$. Then their standard errors $\mathrm{se}(\hat{\tau})$ and $\mathrm{se}(\hat{\nu})$ are obtained from the value of τ and ν to satisfy the level.



(a) Data (Edwards et al.) plotted on nine papers (b) Several data plots on hybrid-hybrid section paper

Figure 5. An example of chromosome aberration data plotted on the hybrid-hybrid section paper.

Figure 5 shows examples of the hybrid-hybrid analysis on several experimental chromosomal aberration data (Indrawati and Kumazawa, 2000). Figure 5 (a) is the nine types of plots on frequencies of dicentrics observed in human lymphocytes irradiated by x-rays at 1 Gy/min (Edwards et el. 1979). The hybrid-hybrid plot is the best maximum of $R^2(\hat{\tau},\hat{\nu})$. Figure 5 (b) shows several plots of various types of chromosome aberration data, the linearity of which is attained across different regions of the hybrid-hybrid section paper.

The hybrid-hybrid analysis has been applied to various data in vital and biological data as well as radiation protection: basic trend of dose reduction due to the Fukushima DAIICHI accident (Kumazawa, 2013), resuspension factor vs time after deposition (Kumazawa, 2014), sky-shine dose vs photon source distance (Kumazawa, 2015 in Japanese), various types of dose vs response (Kumazawa, 2001, 2015), and others, e.g., $\mathrm{hyb}[\rho P/(b-P)]\ vs\ age$ by income category (low to high) cited from the WHO GHO 2015 statistics: P is population ($< b$).

CONCLUSION

This paper presented a practical method for a single regression model of hybrid log normal distribution. It also showed that this method is applicable to the calculation of the linear relationship plotted on the hybrid-hybrid section paper, whose axes are graded in the hybrid scale of log and linear scales.

References

1. F. Galton (1879). "The geometric mean, in vital and social statistics," Proc Roy Soc 29,365-367.
2. D. McAlister (1879). "The law of the geometric mean," Proc Roy Soc 29,367-376.
3. H.J. Gale (1965). "The lognormal distribution and some examples of its application in the field of radiation protection," United Kingdom Atomic Energy Authority, AERE-R 4736.
4. UNSCEAR (1977). "Annex E Occupational Exposure, Sources and Effects on Ionizing Radiations." United Nations Scientific Committee on the Effects of Atomic Radiation, Sales No. E.77.IX.1, New York.
5. ICRP (1977). "ICRP Publication 26 Recommendations of the ICRP, Annals of the ICRP 1(3).
6. S. Kumazawa, J. Shimazaki and T. Numakunai (1980). "Numerical algorithm of statistics on the hybrid lognormal distribution," the 2nd symposium on the applied statistics, D3-1~10 (Tokyo, Oct 25, 1980).
7. S. Kumazawa and T. Numakunai (1981). "A new theoretical analysis of occupational dose distributions indicating the effects of dose limits," Health Phys 41, 465-449.
8. S. Kumazawa, D.R. Nelson and A.C.B. Richardson (1984). "Occupational exposure to ionizing radiation in the United States – A comprehensive review for the year 1980 and a summary of trends for the years 1960-1985," US Environmental Protection Agency, EPA 520/1-84-005.
9. Federal register (1987). "Radiation protection Guidance to Federal Agencies for occupational exposure – Recommendations approved by the President," January 27, 1987.
10. S. Kumazawa and Y. Ohashi (1986). "The hybrid lognormal distribution and Its application (in Japanese)," Jpn J. Appl. Stat. 1581), 1-14.
11. G. Blom (1958). "Statistical estimates and transformed beta variables," New York: John Wiley & Sons.
12. G. Fagerholt (1945). "Particle size distribution of products ground in tube mill," dissertation.
13. A. Hald (1948). "Statistical theory with engineering applications." Cited a slightly expanded version of the Danish textbook, New York: John Wiley & Sons (1952).
14. T. Okuno, H. Kurume, T. Haga and T. Yoshizawa (1971). "Multiple regression analysis in Japanese," monograph, Tokyo, JUSE, Ltd.
15. S.S. Wilks (1948). "Elementary statistical analysis," Princeton, New Jersey, Princeton University Press.
16. I. Indrawati and S. Kumazawa (2000). "Analysis of chromosome aberration data by hybrid-scale model," Japan Atomic Energy Agency, JAERI-Research 2005-005.
17. A.A. Edwards, D.C. Lloyd and R.J. Purrott (1979). "Radiation induced chromosome aberrations and the Poisson distribution," Rad Environ Biophys 16, 89-100.
18. S. Kumazawa (2013). "Distribution of doses to residents evacuated after the Fukushima nuclear power station accident," Proc of Int. Sym. on Environmental monitoring and dose estimation of residents after accident of TEPCO's Fukushima Daiichi Nuclear Power Station, Topics_5-04.pdf.
19. S. Kumazawa (2014). "Reappraisal of predictive models for resuspension," Progress in JNST, 4, 875-878.