*Environmental Health Perspectives*
*Vol. 87, pp. 131–134, 1990*

# Models for Association in Bivariate Survival Data

## by Yoshinori Fujii*

**This paper reviews dependence models for bivariate survival data, classifying them into the four groups: the shock model, the Freund model, the Clayton model, and the mixture model. The paper then concentrates on the mixture model, discussing the testing problem for the equality of marginal distributions under the Weibull type baseline hazard assumption. The new test proposed recently by Fujii is introduced, and its characteristic is studied with respect to the test proposed by Nayak and the sign test by simulation study.**

## Introduction

We often face multivariate survival data such as familial data, matched pairs, and different components of a system. When we study such data, it is important to construct the model that represents dependence. Several models for multivariate survival data have been proposed. To begin with, we briefly review the four models: the shock model, the Freund model, the Clayton model, and the mixture model; these models have been proposed for different situations. We clarify the type of data to which each model can be applied.

We next consider a testing problem in the mixture model that may be used in more situations than the other models. Under the constant baseline hazard assumption, Salvia and Bollinger (*1*), Cantor and Knapp (*2*), and Nayak (*3*) proposed the test for the equality of marginal distributions. To relax the constant hazard assumption, which is too strong in practice, Fujii (*4*) developed a test based on Weibull type baseline hazard functions. We review the test and compare the powers of the test with the test proposed by Nayak (*3*) and also with the sign test by simulation study.

## Review of Bivariate Survival Models

In this section we review four bivariate survival models: the shock model, the Freund model, the Clayton model, and the mixture model. The shock model and the Freund model were proposed for the bivariate extensions of exponential distribution. These models are often used in industrial life testing and can be also extended to various bivariate survival models. On the other hand, the Clayton model and the mixture model are semiparametric models such that the marginal distributions may be unknown. We briefly review these four models below.

### The Shock Model

Consider three events called shock, $S1$, $S2$, and $S3$, which are independent Poisson process $Z_1(t,\lambda_1)$, $Z_2(t,\lambda_2)$ and $Z_{12}(t,\lambda_{12})$, respectively. If $S1(S2)$ occurs, individual 1 (individual 2) fails. If $S3$ occurs, both individuals 1 and 2 fail. Then it follows that a paired failure times $(X,Y)$ of individuals 1 and 2 has the following survival function:

$$F(s,t) \equiv P\{X > s, Y > t\} = \exp[-\lambda_1 s - \lambda_2 t - \lambda_{12}\max(s,t)]$$

Marshall and Olkin (*5*) introduced this idea for a version of bivariate exponential family. We denote this distribution by $\mathrm{SBE}(\lambda_1,\lambda_2,\lambda_{12})$. This distribution has the following properties:

*a*) The marginal distributions of both $X$ and $Y$ follow exponential distribution.
*b*) Lack of memory property
$$P\{X > s_1 + t, Y > s_2 + t \mid X > t, Y > t\} = P\{X > s_1, Y > s_2\}.$$
*c*) $\mathrm{Min}(X,Y)$ follows exponential distribution.
*d*) This distribution is not absolutely continuous with respect to Lebesgue measure because the set $\{X = Y\}$, which is Lebesgue measure 0, has positive probability.
*e*) If $(X,Y)$ follows $\mathrm{SBE}(\lambda_1,\lambda_2,\lambda_{12})$, then $(aX,aY)$ follows $\mathrm{SBE}(\lambda_1/a,\lambda_2/a,\lambda_{12}/a)$ for $a > 0$.

This model may be extended easily to various bivariate survival models using the transformations of $X$ and $Y$. Marshall and Olkin (*5*) also consider the Weibull type extension.

### The Freund Model

Consider the process with four states, O, A, B, AB: O, both individuals 1 and 2 survive; A, only individual

*Department of Mathematics, Miyazaki University, Miyazaki 889-21, Japan.

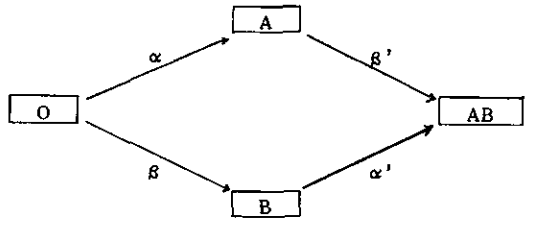


FIGURE 1. The transition probabilities between each two states.

1 fails; B, only individual 2 fails; AB, both individuals 1 and 2 fail. Assume the Markov property such that the transition probability does not depend on time. The transition probabilities between each two states are expressed in Figure 1.

The Markov property in the diagram leads to the following joint density of $X$ and $Y$,

$$f(s,t) = \begin{cases} \alpha\,\beta' \exp\{-\beta' t - (\alpha + \beta - \beta')s\} & \text{for } 0 < s < t \\ \alpha'\beta \exp\{-\alpha' s - (\alpha + \beta - \alpha')t\} & \text{for } 0 < t < s \end{cases}$$

Freund (*6*) proposed this model for bivariate exponential distribution and showed this model has the following properties: The marginal densities are not exponential distribution; $X$ and $Y$ are independent if and only if $\alpha = \alpha'$ and $\beta = \beta'$; this distribution belongs to the exponential family. The dependence between $X$ and $Y$ is essentially such that the failure of either individual 1 or 2 changes the hazard function of the other individual. This model is extended by relaxing the Markov property and the estimation of parameters involved in the model are considered (*7*).

## The Clayton Model

Clayton (*8*) considers the following function:

$$\theta(F(s,0),F(0,t)) = \frac{\dfrac{\partial^2 F(s,t)}{\partial s\,\partial t} F(s,t)}{\dfrac{\partial F(s,t)}{\partial s}\dfrac{\partial F(s,t)}{\partial t}}$$

This function is invariant under the monotone transformation of $X$ or $Y$. Supposing that $\theta(s,t)$ does not depend on $s$ and $t$, Clayton (*8*) proposed $\theta$ as a measure for association for bivariate survival data. From this formulation it follows that $\theta$ is not less than 1; this model does not depend on the marginal distributions; the survival function is given by

$$F(s,t) = \begin{cases} \{F(s,0)^{1-\theta} + F(0,t)^{1-\theta} - 1\}^{1/(1-\theta)} & \text{if } \theta > 1 \\ F(s,0)\,F(0,t) & \text{if } \theta = 1 \end{cases}$$

The conditional hazard function of $X$ given $\cdot$ is defined by

$$H(S \mid \cdot) = \frac{f(s \mid \cdot)}{F(s \mid \cdot)}$$

where $f(s \mid \cdot)$ and $F(s \mid \cdot)$ are density function and survival function given $\cdot$, respectively. Then in terms of the hazard function, this model may be represented by

$$h(s \mid Y = t) = \theta\, h(s \mid Y > t)$$

We call this model the Clayton model. The parameter $\theta$ is larger as dependence is stronger, so we can easily understand the meaning of the parameter. The estimations of the parameter $\theta$ were investigated by Oakes (*9*), Clayton and Cuzick (*10*), and others. This model is attractive, but the assumption that $\theta(s,t)$ is constant seems to be too strong in practice. The extension of this model is considered in the next subsection.

## The Mixture Model

Suppose that a pair of random variables $(X,Y)$ has a common unobservable random variable $Z$, called frailty. Also suppose that given $Z = z$, $X$ and $Y$ are independently distributed with hazard function $z\lambda_1(s)$, $z\lambda_2(t)$, respectively. We call $\lambda_i(t)$ $(i = 1,2)$ the baseline hazard function. What is the meaning of the random variable $Z$? In univariate case, the similar random variable $Z$ was introduced to represent a heterogeneous factor in Swedish mortality data (*11*). When we deal with survival data, there are many situations where we cannot suppose the homogeneity of the data, so we need to introduce the heterogeneous factor in the probability model. In the case of multivariate survival data, one possible reason of dependence is due to the individuals share of some heterogeneous factor. If this is the case, $Z$ indicates the common heterogeneous factor such as genetic factors or environmental factors.

Gamma distribution family is used for the distribution of $Z$ (*11*). Hugaard (*12*) also introduced three parameter distribution family $P(\alpha,\delta,\theta)$ and investigated its properties. These properties are as follows.

*a*) The parameter space of this family is $(0,1] \times (0,\infty) \times [0,\infty) \cup \{0\} \times (0,\infty) \times (0,\infty)$. It is neither open nor closed and is not a Cartesian product.

*b*) This family $P(\alpha,\delta,\theta)$ is wide. Because it includes gamma distribution, stable distribution, and inverse Gaussian distribution.

*c*) The Laplace transform is

$$L(s) = \exp\left[-\frac{\delta}{\alpha}\{(\theta + s)^\alpha - \theta^\alpha\}\right].$$

*d*) Let $Z_1, \ldots, Z_n$ be independent $Z_i \sim P(\alpha\delta_i,\theta)$. The distribution of $\Sigma Z_i$ is then $P(\alpha,\Sigma\delta_i,\theta)$.

Supposing the distribution family $P(\alpha,\delta,\theta)$ for the common factor $Z$, it follows that the survival function of $(X,Y)$ is given by

$$F(s,t) = \exp[-\delta\{(\Lambda_1(s) + \Lambda_2(t) + \theta)^\alpha - \theta^\alpha\}/\alpha]$$

where $\Lambda_i(t) = \int_0^t \lambda_i(u)\,du$ $(i = 1,2)$.

Especially if we assume the gamma distribution of $Z$, which is equivalent to suppose $P(0,\delta,\theta)$, then the survival function is

$$F(s,t) = (F(s,0)^{-1/\delta} + F(0,t)^{-1/\delta} - 1)^{-\delta}\ \delta > 0,$$

which is equivalent to Clayton model. This shows that the mixture model is a wider family of distribution than Clayton model, but we cannot easily understand the meaning of the parameter $(\alpha,\delta,\theta)$.

The four models above have been generated in different situations, and it is important to choose the model according to the type of data. Since the shock model and Freund model use the idea of stochastic process, we must suppose the pair of failure times are recorded by a single clock from a common origin. These models are available for life testing of two component system or a person's paired organs. However, double clocks are natural when we study the paired survival data such as familial data and matched pair data. The Clayton model and the mixture model can be applicable to the data on the survival times that are recorded by separate two clocks.

## Test for Equality of Marginal Distributions

Among the four models reviewed in the previous section, the mixture model is frequently used in statistical inference for the baseline hazard function. In this section we consider the testing problem for a pair of survival data in the mixture model. Under the constant baseline hazard assumption such that $\lambda_i(t) = \lambda_i$ $(i = 1,2)$, Salvia and Bollinger (*1*), Cantor and Knapp (*2*), and Nayak (*3*) proposed the test for the equality of marginal distributions by using the statistic $T = X/Y$. The statistic $T$ is not dependent on the distribution of $Z$. Here we discuss the optimal $C(\alpha)$-test proposed by Fujii (*4*) under the Weibull type hazard assumption. $T$ is also useful in this case.

### The Optimal $C(\alpha)$-Test

If we suppose the following baseline hazard functions are given by

$$\lambda_i(t) = \beta_i t^{\gamma+1}\ (i = 1,2)$$

then the conditional density function of $(X,Y)$ given $Z = z$ is given by

$$f(x,y) = z^2(\gamma + 1)^2\beta_1\beta_2 x^\gamma y^\gamma \times \exp\{-z(\beta_1 x^{\gamma+1} + \beta_2 y^{\gamma+1})\}.$$

We consider testing the hypothesis $H_0{:}\beta_1 = \beta_2$ versus $H_1{:}\beta_1 \neq \beta_2$, treating $\gamma$ as a nuisance parameter. Like the previous authors, we consider the statistic $T = X/Y$. Under the above model, $T$ has marginal density

$$f_T(t) = \frac{(\gamma + 1)\delta t^\gamma}{(1 + \delta t^{\gamma+1})^2},\ \delta = \beta_1/\beta_2$$

which does not depend on $z$. Furthermore, $T$ is the maximum invariant statistic with respect to $Z$. This property of $T$ holds if and only if the baseline hazard is in the Weibull type.

We construct the optimal $C(\alpha)$-test for the hypothesis. When $T_1, T_2, \ldots, T_n$ are observed, we have the log likelihood as follows

$$L(\gamma,\delta,t_i) = n\log(\gamma + 1)\delta + \gamma\sum_{i=1}^{n}\log t_i - 2\log(1 + \delta t^{\gamma+1}).$$

The efficient score with respect to $\delta$ is

$$\frac{d\,L(\gamma,\delta,t_i)}{d\,\delta} = \sum_{i=1}^{n}\frac{1 - t_i^{\gamma+1}}{1 + t_i^{\gamma+1}}$$

and the Fisher information matrix is

$$i(\theta) = \frac{n}{3}\begin{pmatrix} 1 & 0 \\ 0 & \dfrac{3 + \pi^2}{3(1 + \gamma)^2}\end{pmatrix}.$$

Then the optimal $C(\alpha)$-test statistic is

$$W = \frac{3}{n}\left\{\sum_{i=1}^{n}\frac{1 - t_i^{\hat\gamma+1}}{1 + t_i^{\hat\gamma+1}}\right\},$$

where $\hat\gamma$ is maximum likelihood estimator under the null hypothesis. Since the likelihood function is concave, we can easily calculate $\hat\gamma$ by the Newton iterative method.

### The Power Comparison

We investigate the power property of the Nayak test, sign test, and the optimal $C(\alpha)$-test under Weibull type hazard assumption. The Nayak test rejects the null hypothesis if the statistic

$$U = \frac{1}{n}\sum_{i=1}^{n} T_i/(1 + T_i)$$

is smaller than a constant. In the case of the Weibull type hazard, it follows that statistic $U$ asymptotically follows normal distribution. The mean and variance of $U$ is obtained as

$$\mu_0 = E(U) = A(\gamma,\delta) - A(\gamma,1/\delta) + 1/2$$
$$n\,\sigma_0^2 = n\,\mathrm{Var}(U) = 2A(\gamma,\delta) - 2B(\gamma,1/\delta) - 2B(\gamma,\delta) + 3/4 - \mu_0^2$$

where

$$A(\gamma,\delta) = \int_0^1 \frac{1}{(1 + t)^2(1 + \delta t^{\gamma+1})}\,dt$$

$$B(\gamma,\delta) = \int_0^1 \frac{1}{(1 + t)^3(1 + \delta t^{\gamma+1})}\,dt.$$

**Table 1. Simulation result of power for Nayak test, sign test, and optimal C (α)–test.**

| | $\gamma = 0.0$ | | | $\gamma = 0.2$ | | | | $\gamma = 0.4$ | | | $\gamma = 1.0$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| δ | Nayak | Sign | C (α) | Nayak | Sign | C (α) | δ | Nayak | Sign | C (α) | Nayak | Sign | C (α) |
| 1.0 | 4.99 | 5.08 | 5.00 | 3.50 | 5.13 | 5.14 | 1.0 | 2.46 | 5.04 | 5.10 | 0.57 | 5.04 | 5.14 |
| 1.2 | 18.88 | 15.11 | 18.87 | 14.12 | 15.38 | 18.25 | 1.2 | 10.32 | 16.06 | 18.29 | 3.88 | 15.79 | 18.85 |
| 1.4 | 38.84 | 31.84 | 38.75 | 32.32 | 31.67 | 38.17 | 1.4 | 26.74 | 31.78 | 38.84 | 12.13 | 31.92 | 39.87 |
| 1.6 | 61.10 | 50.03 | 60.57 | 54.48 | 48.56 | 60.69 | 1.6 | 47.03 | 48.94 | 60.25 | 26.28 | 58.95 | 60.17 |
| 1.8 | 77.45 | 64.56 | 76.86 | 72.07 | 65.31 | 76.34 | 1.8 | 65.88 | 64.54 | 76.39 | 43.42 | 64.05 | 76.87 |
| 2.0 | 87.68 | 77.28 | 87.26 | 83.47 | 77.35 | 86.78 | 2.0 | 79.36 | 77.68 | 87.07 | 60.53 | 76.84 | 87.22 |

This shows that under the null hypothesis, the mean is 1/2 for any $\gamma$, but the variance increases as $\gamma$ becomes larger. Note that this test is one-sided.

We next consider the sign test. Let $R$ = number of $i$ such that $T_i \leqslant 1$, then the signed test for $H_0 : \delta = 1$ against $H_1 : \delta \geqslant 1$ is given by

$$\phi(T_1, T_2, \ldots, T_n) = \begin{cases} 1 \text{ if } R > C \\ \theta \text{ if } R = C \\ 0 \text{ if } R < C \end{cases}$$

where $C$ and $\theta$ are determined by the level of significance $\alpha$.

Table 1 shows the powers of the three tests at sample size $n = 50$ obtained by a simulation. The simulation is carried out by generating 10,000 uniform random digits for each set of $(\gamma, \delta)$. Note that because the Nayak test is one-sided, we must correct the optimal $C(\alpha)$-test to one-sided test for comparison.

The table shows that the Nayak test is good at neighborhood of $\gamma = 0$, but it tends to be more inferior to the other tests as $\gamma$ becomes larger. Especially, the Nayak test is not better than Sign test at $\gamma = 1$. On the other hand the optimal $C(\alpha)$-test is as good as the Nayak test at $\gamma = 0$ but superior to the other tests for $\gamma > 0$. Note that its power is almost constant even if $\gamma$ becomes large.

The author appreciates the suggestions and encouragement given by T. Yanagawa.

## REFERENCES


1. Salvia, A. A., and Bollinger, R. C. Testing equality of correlated exponential variables. IEEE Trans. Reliability R-33: 374–376 (1984).
2. Cantor, A. B., and Knapp, R. G. A test of the equality of survival distributions based on paired observations from conditionally independent exponential distributions. IEEE Trans. Reliability R-34: 342–246 (1985).
3. Nayak, T. K. Testing equality of conditionally independent exponential distributions. Commun. Stat. A17: 807–820 (1988).
4. Fujii, Y. Testing for the equality of marginal distributions on positively dependent bivariate survival data. Commun. Stat. B18: 633–642 (1988).
5. Marshall, A. W., and Olkin, I. A multivariate exponential distribution. J. Am. Stat. Assoc., 62: 30–44 (1967).
6. Freund, J. E. A bivariate extension of the exponential distribution. J. Am. Stat. Assoc., 56: 971–977 (1961).
7. Aalen, O., Borgan, O., Keiding, N., and Thormann, J. Interaction between life history events. Nonparametric analysis for prospective and retrospective data in the presence of censoring. Scand. J. Stat. 7: 161–171 (1980).
8. Clayton, D. G. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, Biometrika 65: 141–151 (1978).
9. Oakes, D. Semiparametric inference in a model for association in bivariate survival data. Biometrika 73: 353–361 (1986).
10. Clayton, D. G., and Cuzick, J. Multivariate generalizations of the proportional hazards model. J. Roy. Soc. Stat. A147: 82–117 (1985).
11. Vaupel, J. W., Manton, K. G. and Stallard, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16: 439–454 (1978).
12. Hougaard, P. Survival models for heterogenous populations derived from stable distributions. Biometrika 73: 387–396 (1986).