—Review—

Applications of Statistics to Medical Science, IV
Survival Analysis

Hiroshi Watanabe

Department of Mathematics, Nippon Medical School

Abstract

The fundamental principles of survival analysis are reviewed. In particular, the Kaplan-Meier method and a proportional hazard model are discussed. This work is the last part of a series in which medical statistics are surveyed.
(J Nippon Med Sch 2012; 79: 176–181)

Key words: statistical test, statistical inference, survival analysis, Kaplan-Meier method, proportional hazard model

1. Introduction

In the previous works[1–3], we studied the conceptual framework of statistical analysis and actual procedures. The present work is devoted to survival analysis. A typical problem is to estimate statistically how long subjects may survive or remain free from a relapse after treatment.

Textbooks[4,5] are rich sources of information on medical statistics. For details of survival analysis, see[6].

2. Survival Probability

Let $S(t)$ be survival probability, that is, $S(t)$ gives (conditional) probability with which a subject may survive at time $t(>0)$ who is assumed to be alive at $t=0$. Obviously, $S(0)=1$ holds because a subject was alive at $t=0$, and $S(t)$ decreases with respect to t because a subject is at risk to "die" at each instant. An example of $S(t)$ is shown in **Figure 1**. In this example, the risk of death seems to be high for $0<t<20$ and $t>60$. In general, the magnitude of risk varies with time. In Section 4, this statement will be made more precise.

Kaplan-Meier Method

Let us make statistical inference of survival probability $S(t)$ by a sample shown in **Table 1 (A)**. This sample consists of data for 10 subjects: each subject was followed up for some period after a certain treatment. A final state, i.e., a state at the end of a follow-up was "endpoint(1)" or "censored(0)". The former means that a subject "died" by the reason under consideration, whereas the latter means that a subject was "alive" at the end of a follow-up, that a subject "died" by another reason we were not interested in, or that the final state was unknown. Note that a follow-up may be terminated by several reasons before a subject meets an endpoint.

We read the data given in **Table 1 (A)** as follows: at $t=0$, there were 10 subjects to be followed up;

Correspondence to Hiroshi Watanabe, Department of Mathematics, Nippon Medical School, 2–297–2 Kosugi-cho, Nakahara-ku, Kawasaki, Kanagawa 211–0063, Japan
E-mail: watmath@nms.ac.jp
Journal Website (http://www.nms.ac.jp/jnms/)

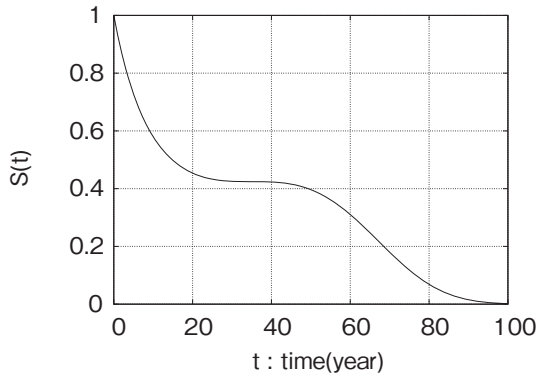

Fig. 1 An example of survival probability $S(t)$, i.e., probability that a subject may be alive at time t under the assumption that the subject was alive at $t=0$.

at t=2, no subject in 10 met an endpoint, and 1 was censored (9 remain);

at t=4, 1 subject in 9 met an endpoint, and no one was censored (8 remain);

at t=5, no subject in 8 met an endpoint, and 1 was censored (7 remain);

at t=7, 1 subject in 7 met an endpoint, and no one was censored (6 remain);

at t=10, 1 subject in 6 met an endpoint, and (then) 1 was censored (4 remain),

and so on. We have assumed that events happened at t=2, 4, 5, 7, 10, 12, 14, 15, although the precise time of each event is not known. Then we can make inference on $S(t)$ as follows:

$$S(t)=1, \qquad 0<t<2, \quad (1)$$

$$S(t)=1\cdot\frac{10}{10}, \qquad 2<t<4, \quad (2)$$

$$S(t)=1\cdot\frac{10}{10}\cdot\frac{8}{9}, \qquad 4<t<5, \quad (3)$$

$$S(t)=1\cdot\frac{10}{10}\cdot\frac{8}{9}\cdot\frac{8}{8}, \qquad 5<t<7, \quad (4)$$

$$S(t)=1\cdot\frac{10}{10}\cdot\frac{8}{9}\cdot\frac{8}{8}\cdot\frac{6}{7}, \qquad 7<t<10, \quad (5)$$

$$S(t)=1\cdot\frac{10}{10}\cdot\frac{8}{9}\cdot\frac{8}{8}\cdot\frac{6}{7}\cdot\frac{5}{6}, \qquad 10<t<12, \quad (6)$$

$$S(t)=1\cdot\frac{10}{10}\cdot\frac{8}{9}\cdot\frac{8}{8}\cdot\frac{6}{7}\cdot\frac{5}{6}\cdot\frac{2}{4}, \qquad 12<t<14, \quad (7)$$

$$S(t)=1\cdot\frac{10}{10}\cdot\frac{8}{9}\cdot\frac{8}{8}\cdot\frac{6}{7}\cdot\frac{5}{6}\cdot\frac{2}{4}\cdot\frac{2}{2}, \qquad 14<t<15. \quad (8)$$

The survival probability $S(t)$ calculated as above is drawn as curve A in **Figure 2**. Curve A is discontinuous at t=4, 7, 10, 12, i.e., when subjects met endpoints. These times are called *failure times*. Note that the thus obtained function $S(t)$ is not the true (population) survival probability but its statistical inference. This method is referred to as the **Kaplan-Meier method**.

Table 1 (A)

Subject	Period (months)	Final state
1	2	0
2	4	1
3	5	0
4	7	1
5	10	1
6	10	0
7	12	1
8	12	1
9	14	0
10	15	0

Table 1 (B)

Subject	Period (months)	Final state
11	1	1
12	3	1
13	3	0
14	5	1
15	7	1
16	9	1
17	10	0
18	11	1
19	12	0

Examples of data sets of the Kaplan-Meier method. Group A (or B) consists of 10 (or 9, respectively) subjects. The subjects were followed up for some period of time after certain treatments, and their final states were recorded. (0 and 1 mean "censored" and "endpoint," respectively.)

Log-Rank Test

We next discuss how to compare (true) survival probabilities for two groups. Let $S_A(t)$ and $S_B(t)$ be Kaplan-Meier survival probabilities for the data given in **Table 1 (A) and (B)**, respectively. Curve B in **Figure 2** corresponds to $S_B(t)$.

The problem that arises is whether the survival probabilities for A and B are significantly different or not. In what follows, we introduce a method referred to as the **log-rank test**.

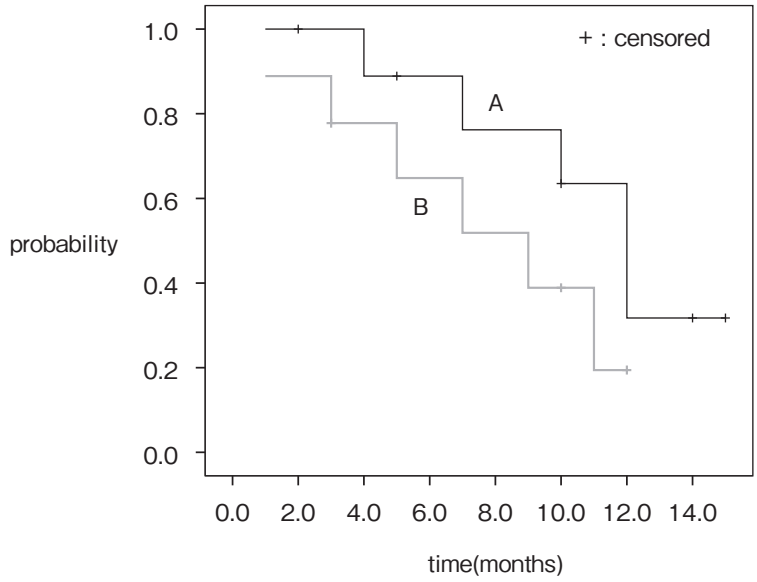

Fig. 2 Examples of Kaplan-Meier survival probabilities. The curves A and B show survival probabilities that are obtained by means of the Kaplan-Meier method from the data given in Table 1 (A) and (B), respectively. The mark "+" indicates a censored case.

Table 2

	group A	group B	total
at an endpoint	0	1	1
remaining	10	8	18
total	10	9	19

A contingency table for a log-rank test. This table is produced from Table 1 (A) and (B) for $t=1$. No (or 1) subject meets an endpoint, and 10 (or 8) subjects remain to be followed up in group A (or B, respectively).

Table 3

	group A	group B	total
at an endpoint	a_j	b_j	a_j+b_j
remaining	c_j	d_j	c_j+d_j
total	a_j+c_j	b_j+d_j	n_j

A general form of contingency table for a log-rank test: a_j (or b_j) subjects meet endpoints, and c_j (or d_j) subjects remain to be followed up in group A (or B, respectively) at, $t=t_j$, the jth failure time.

We make a contingency table at each failure time. We make no table at a time when no subject meets an endpoint, and we have only censored subject(s). For example, a contingency table for t=1 is as **Table 2**, and we do nothing for t=2. A general form of a contingency table for the jth failure time t_j is shown in **Table 3**.

We define a log-rank statistic X^2_{LR} by

$$X^2_{\mathrm{LR}}=\frac{(O_{\mathrm{A}}-E_{\mathrm{A}})^2}{E_{\mathrm{A}}}+\frac{(O_{\mathrm{B}}-E_{\mathrm{B}})^2}{E_{\mathrm{B}}}, \tag{9}$$

where

$$O_{\mathrm{A}}=\sum_j a_j, \tag{10}$$

$$O_{\mathrm{B}}=\sum_j b_j, \tag{11}$$

$$E_{\mathrm{A}}=\sum_j \frac{(a_j+b_j)(a_j+c_j)}{n_j}, \tag{12}$$

$$E_{\mathrm{B}}=\sum_j \frac{(a_j+b_j)(b_j+d_j)}{n_j}. \tag{13}$$

The statistic X^2_{LR} approximately obeys a chi-squared distribution with 1 degree of freedom under the null-hypothesis H_0: true (population) survival probabilities for groups A and B are the same. Note that the summands in the right sides of (12) and (13) are the expected numbers of events under H_0. Then, a test for H_0 with significance level α is done according to the rule: reject H_0 if and only if $X^2_{\mathrm{LR}} > \chi^2_1(1-\alpha)$, where $\chi^2_1(1-\alpha)$ denotes the $(1-\alpha)\times 100$th percentile of a chi-squared distribution with 1 degree of freedom.

Another test called a **Mantel-Haenszel test** for the above null-hypothesis H_0 uses the following statistic

Table 4

subject	W (period)	X (exposure)	Y (exposure)	Z (final state)
1	w_1	x_1	y_1	z_1
2	w_2	x_2	y_2	z_2
3	w_3	x_3	y_3	z_3
⋮	⋮	⋮	⋮	⋮
n	w_n	x_n	x_n	z_n

A general form of data set to which a proportional hazard model with 2 exposure variables can be fitted. The jth subject, who has exposure variables $X=x_j$ and $Y=y_j$, was followed up for a time period of length w_j and the final state was z_j. The exposure measured by variable X may be, e.g., "treatment" ($X=1$) and "placebo" ($X=0$), whereas Y may be age of each subject. The meaning of final states is the same as in Table 1.

$$X^2_{\mathrm{MH}}=\frac{(O_{\mathrm{A}}-E_{\mathrm{A}})^2}{V}=\frac{(O_{\mathrm{B}}-E_{\mathrm{B}})^2}{V}, \tag{14}$$

where

$$V=\sum_j \frac{(a_j+b_j)(c_j+d_j)(a_j+c_j)(b_j+d_j)}{n_j^2(n_j-1)}. \tag{15}$$

The statistic X^2_{MH} approximately obeys a chi-squared distribution with 1 degree of freedom under H_0.

Remarks. 1) The above two tests using X^2_{LR} and X^2_{MH}, respectively, give, in most cases, the same result. In fact, the test using X^2_{MH} is a special case of a more general procedure called the *Mantel-Haenszel test* that is applicable to a *log-rank situation*.

2) Mantel-Haenszel tests for more than 2 groups are also possible. The statistic that is used for this purpose is mathematically more complicated and is given in terms of a matrix.

3) If the samples have no censored cases, we can use the Wilcoxon rank sum test with respect to the 2 sets of failure times. Consider the data given in **Table 1 (A) and (B)** with the final state 1 (an endpoint) for all the subjects. Then, we can apply the Wilcoxon rank test to the 2 sets {2,4,5,7,10,10,12,12,14,15} and {1,3,3,5,7,9,10,11,12}. Note that a nonparametric method should be used because we cannot assume the normality of probability distributions of failure times.

3. Proportional Hazard Model

In this section, we discuss how to estimate effects of exposure on survival probability.

Data Set

Consider the data for n subjects, as in **Table 4**. The quantities X and Y measure some exposure, and values of X and of Y for the ith subject are denoted by x_i and y_i, respectively. These variables may be quantitative or categorical. For example, x_i=1 (or =0) may mean that the ith subject received (or did not receive, respectively) a certain treatment, and y_i may denote the age of the ith subject. The variable Z denotes final states, i.e., "endpoint(1)" or "censored (0)".

If we ignore the columns for X and Y, **Table 4** has the same structure as **Table 1 (A) (and as (B))**. If we ignore the column for Y and break the table into a group with the final state 1 and a group with the final state 0, we obtain a set of tables as **Table 1 (A) and (B)** that are combined.

Our aim is to make an inference on the (true) survival probability $S(x, y, t)$ for a subject who has exposure given by $X=x$ and $Y=y$. The number of exposure variables (2 in **Table 4**) can be generalized to an arbitrary positive integer.

Statistical Model

To obtain the survival probability $S(x, y, t)$ as a function of x, y, and t, we adopt a statistical model called the proportional hazard model. Our task is to determine the optimal values of parameters included in this model so that the model may describe our experimental data. In spite of an apparently special form of the proportional hazard model, it is, in fact, based on the general concept of *hazard* as explained in Section 4 and has wide applications.

Let us introduce the proportional hazard model. We make the following assumptions about the survival probability $S(x, y, t)$:

1. The function $S(x, y, t)$ depends on x, y, and t through the form:

$$S(x, y, t) = S_0(t)^{c(x, y)}, \tag{16}$$

where $c(x, y)$ is a function of x and y independent of t and $S_0(t)$ is a function of t independent of x and y.

2. The function $c(x, y)$ has the following form:

$$c(x, y) = e^{a+b_1x+b_2y}, \tag{17}$$

where e=2.71828⋯ is a mathematical constant.

The function $S_0(t)$ on the right side of (16) is a (virtual) prototype survival probability, and $S(x, y, t)$ is assumed to be a modification of $S_0(t)$ by $c(x, y)$. The meaning of the assumption (16) will be explained in Section 4 by using the concept of hazard. On the other hand, (17) implies that 2 kinds of exposure measured by X and Y, respectively, have mutually independent effects on subjects as risk. The model that is characterized by the assumptions (16) and (17) is called the **proportional hazard model** or **Cox's model**.

If $b_1>0$, $c(x, y)$ given by (17) is increasing with respect to x, and $S(x, y, t)$ given by (16) is decreasing with respect to x because $0 \leqq S_0(t) \leq 1$. Therefore, the larger X is, the less likely a subject will survive. Similarly, if $b_1<0$, the smaller X is, the less likely a subject will survive. In both cases (i.e., if $b_1 \neq 0$), the exposure measured by X is a risk factor.

Statistical Inference

Optimal coefficients b_1 and b_2 are determined by the method of *partial likelihood*. (The value of a is of no use because we are interested in the effects of exposure.) We, however, cannot write explicit formulae for the optimal values. Instead, we numerically obtain them by using software packages, such as SAS and SPSS. Such programs output the optimal values of coefficients b_1, b_2, their standard errors SE(b_1), SE(b_2), and a log likelihood statistic $-2 \log L$ (or $\log L$). Then, in the same way as logistic regression[3], we can make use of the outputs for the following aims: 1) to produce interval estimation for population values b_{1*} and b_{2*} corresponding to b_1 and b_2, respectively; 2) to perform a Wald test for statistical significance of the coefficients b_1 and b_2; and 3) to perform a likelihood ratio test by means of a log likelihood statistic $-2 \log L$. In fact, the description in the subsection "Interval estimation and statistical tests" in Section 4 of[3] can be read in the context of a proportional hazard model by interpreting the logits (31) and (32) in[3] as exponents of $c(x, y)$ in (17).

Interpretation of Significance Tests

Suppose that X is a "treatment/placebo" variable, and Y denotes age for each subject. If the population parameter b_{1*} is significantly nonzero, it implies that the treatment has a real effect. Furthermore, we have solved a possible confounding by age on the effect of the treatment because the effect of age Y was estimated and separated in the proportional hazard model.

Let us define a reduced model with only the "treatment/placebo" variable X but without the age variable Y. This model can be fitted to the data given in **Table 4** by ignoring the Y column. Suppose that the log likelihood statistics of the full model (with X and Y) and of the reduced model (without Y) are significantly different. Then, we conclude that the variable Y is meaningful; hence, age is a real confounding factor. For this aim, we can also perform a Wald test for the full model to test the significance of b_2.

Remark. The likelihood ratio test and the Wald test may not give the same conclusion. In general, the log likelihood test is preferable.

4. Hazard

In this section, we discuss the concept of hazard.

Hazard Function

Let $S(t)$ be a survival probability as shown in **Figure 1**. The values $S(t)$ and $S(t+\Delta t)$ are the probabilities with which a subject may be alive at t and $t+\Delta t$, respectively, under the assumption that the subject is alive initially (at t=0). The difference $S(t)-S(t+\Delta t)$, therefore, gives the probability that the subject may meet an endpoint in a time interval [t,

$t+\Delta t$]. If Δt is a small positive number, the difference $S(t+\Delta t)-S(t)$ is proportional to Δt, and the proportional constant is called a *derivative* of $S(t)$, which is written as $\frac{dS(t)}{dt}$, hence

$$S(t)-S(t+\Delta t)\approx-\frac{dS(t)}{dt}\Delta t. \tag{18}$$

We thus see that the probability that a subject who is alive at t may meet an endpoint in a time interval $[t,t+\Delta t]$ is given by

$$\frac{S(t)-S(t+\Delta t)}{S(t)}\approx-\frac{1}{S(t)}\frac{dS(t)}{dt}\Delta t, \tag{19}$$

because the subject may be alive at t with probability $S(t)$. We write the right side of (19) as $h(t)\Delta t$. Namely, we put

$$\frac{dS(t)}{dt}=-h(t)S(t). \tag{20}$$

The function $h(t)$ measures instantaneous magnitude of risk at time t, i.e., it gives probability that a subject who is alive at t may meet an endpoint *per unit time* just after the time t. We refer to $h(t)$ as a **hazard function**. The larger $h(t)$ is, the more likely a subject may "die" at t. In particular, if the function $h(t)$ is a constant $\lambda>0$, the corresponding survival probability $S_0(t)$ is given by

$$S_0(t)=e^{-\lambda t},\quad t>0. \tag{21}$$

For a time-dependent function $h(t)$, we can mathematically determine survival probability $S(t)$ with $h(t)$ as its hazard function.

Proportional Hazard Model

Assume again that a hazard function is a constant λ. Then, (21) implies that

$$S_0(t)^c=e^{-c\lambda t}, \tag{22}$$

where c is an arbitrary positive constant. Namely, the hazard function for $S_0(t)^c$ is $c\lambda$, i.e., c times the hazard function for $S_0(t)$.

The above fact is generalized for a time-dependent hazard. Let $h_0(t)$ and $h(t)$ be hazard functions for survival probabilities $S_0(t)$ and $S_0(t)^c$, respectively. Then, it holds that

$$h(t)=ch_0(t), \tag{23}$$

where c is independent of time t. Namely, the hazard function for $S_0(t)^c$ is c times that for $S_0(t)$.

In view of (23), we can interpret the assumption (16) of a proportional hazard model. Let $h(x,\ y,\ t)$ and $h_0(t)$ be hazard functions for survival probabilities $S(x,\ y,\ t)$ and $S_0(t)$, respectively. Then, the assumption (16) is equivalent with

$$h(x,\ y,\ t)=c(x,\ y)h_0(t). \tag{24}$$

Note that $c(x,\ y)$ is independent of t. The hazard function $h_0(t)$ is called a *baseline hazard function*. Thus, in a proportional hazard model, exposure acts on each subject as risk so that the subject is $c(x,\ y)$ times more likely to "die" at each instant.

Hazard Ratio

If X is a categorical variable with a value 0 or 1, the hazard functions for $x=1$ and for $x=0$ are given by $h(1,\ y,\ t)$ and $h(0,\ y,\ t)$, respectively, hence (24) implies

$$\frac{h(1,\ y,\ t)}{h(0,\ y,\ t)}=\frac{c(1,\ y)}{c(0,\ y)}. \tag{25}$$

Furthermore, using the assumption (17), we have the following formula for a *hazard ratio*:

$$\frac{h(1,\ y,\ t)}{h(0,\ y,\ t)}=\frac{e^{a+b_1\cdot 1+b_2y}}{e^{a+b_1\cdot 0+b_2y}}=e^{b_1}. \tag{26}$$

By means of (26), we can produce a confidence interval of a hazard ratio from that of b_1 (precisely, of b_{1*}).

References

1. Watanabe H: Applications of Statistics to Medical Science, I. J Nippon Med Sch 2011; 78: 274–279.
2. Watanabe H: Applications of Statistics to Medical Science, II. J Nippon Med Sch 2012; 79: 31–36.
3. Watanabe H: Applications of Statistics to Medical Science, III. J Nippon Med Sch 2012; 79: 115–120.
4. Rosner B: Fundamentals of Biostatistics, 7th ed., 2011; Brooks/Cole, Boston.
5. Armitage P, Berry G, Matthews JNR: Statistical Methods in Medical Research, 4th ed., 2002; Blackwell Science, Massachusetts.
6. Kleinbaum DG, Klein M: Survival analysis, 2nd ed., 2005; Springer, New York.

(Received, February 13, 2012)
(Accepted, April 26, 2012)