

Occurrence/Exposure Rate

GUTTI JOGESH BABU & M. BHASKARA RAO

Volume 3, pp. 1199–1201

In

Encyclopedia Of Actuarial Science
(ISBN 0-470-84676-3)

Edited by

Jozef L. Teugels and Bjørn Sundt

© John Wiley & Sons, Ltd, Chichester, 2004

Occurrence/Exposure Rate

Introduction

Suppose someone who is healthy enters a hazardous employment. Let π denote the probability that he/she will die of cancer after the employment is taken. Let π' denote the probability of a randomly selected person, dying because of cancer, in the population at large. A comparison of the probabilities π and π' would enable the researcher to assess quantitatively how risky the hazardous employment is. These probabilities are unknown to begin with. In order to estimate the probabilities, we need to collect data. Let us focus exclusively on the estimation of the probability π . We now consider the population of individuals who enter into the hazardous employment. Let X be the life span of an individual randomly selected from the population. Let F be the cumulative distribution function of X . The quantity that is more informative than the probability π is the so-called *risk ratio* ρ , which is defined as (Probability of Death due to Cancer)/(Average Life Span), that is,

$$\rho = \frac{\pi}{\int_0^{\infty} t dF(t)}. \quad (1)$$

The entity ρ is also called *specific occurrence/exposure rate*. It is the rate of deaths due to cancer during an average life span. In order to estimate the risk ratio ρ , we need to estimate the distribution function F , in addition. The distribution function F has two components. Two subpopulations can be identified within the population: those who die of cancer with probability π and those who die of other causes with probability $1 - \pi$. Let F_1 be the cumulative distribution function of lifetimes of individuals in the subpopulation who die of cancer and F_2 that of the subpopulation who die of other causes. The distribution function F is indeed a mixture (*see Mixture of Distributions*) of F_1 and F_2 , that is,

$$F = \pi F_1 + (1 - \pi)F_2. \quad (2)$$

Thus, now we need to estimate π , F_1 , and F_2 in order to estimate the risk ratio. For this purpose, we follow a cohort (*see Cohort*) of n individuals randomly selected from the population. It is impossible to

follow the individuals until all of them die. On the basis of practical and financial considerations, we could follow the individuals only for a certain length M of time. If we monitor the individuals in the sample during the time interval $[0, M]$, the risk ratio cannot be estimated nonparametrically. We need to modify the definition of risk ratio. The *modified risk ratio* is given by

$$\rho_M = \frac{\pi F_1(M)}{M\bar{F}(M) + \int_{[0,M]} t dF(t)} = \frac{\pi_M}{\int_{[0,M]} \bar{F}(t) dt}, \quad (3)$$

where \bar{F} is the survival function associated with the distribution function F , that is, $\bar{F} = 1 - F$ and $\pi_M = \pi F_1(M)$. The denominator of the ratio is essentially the average life span of individuals in the population in the presence of deterministic right-censoring at M . The numerator is the probability of dying due to cancer in the interval $[0, M]$. For a general introduction to Survival Models in the context of Actuarial Studies, see [7]. For additional discussion on risk ratio, see [3].

The model we are entertaining here is more general than the competing risks model (*see Competing Risks*). It is easier to explain the mechanics of a competing risk model in the language of machines and parts. Suppose a machine runs on two components. Let X_1 be the life span of Component 1 and X_2 be that of Component 2 with distribution functions F_1 and F_2 , respectively. Assume that X_1 and X_2 are independently distributed. Suppose that the machine breaks down if and only if at least one of the components fails. Let X be the life span of the machine with distribution function F . Then

$$F = \pi F_1 + (1 - \pi)F_2, \quad (4)$$

where $\pi = P(X_1 \leq X_2)$. Thus, in the competing risks model, the entity π is determined by the distribution functions F_1 and F_2 uniquely. In our model, π has no structure. The main purpose of the note is to outline nonparametric estimation of the risk ratio.

Nonparametric Estimation

All individuals in the sample are followed up to the time M . In such a study, four possibilities arise with reference to any member of the cohort.

2 Occurrence/Exposure Rate

1. The individual dies of cancer.
2. The individual dies of other causes.
3. The individual leaves the study.
4. The individual is alive at Time M .

Let C denote the time at which an individual leaves the cohort study. Traditionally, C is called the *censoring random variable*. Let G be its distribution function. Let us introduce a binary random variable for any randomly selected individual from the population by

$$\begin{aligned} \bar{\Delta} &= 1 && \text{if the individual dies of cancer} \\ &= 0 && \text{if the individual dies of other causes.} \end{aligned} \quad (5)$$

We assume that C and $(X, \bar{\Delta})$ are independently distributed. Associated with each individual in the population we have a triplet $(X, \bar{\Delta}, C)$, which is not observable in its entirety. What can be recorded for each individual is (Y, Δ) , where $Y = \min\{X, C\}$ and

$$\begin{aligned} \Delta &= 1 && \text{if } Y = X \text{ and } \bar{\Delta} = 1, \\ &= 0 && \text{if } Y = X \text{ and } \bar{\Delta} = 0, \\ &= -1 && \text{if } Y = C, \\ &= -2 && \text{if } Y = M. \end{aligned} \quad (6)$$

The four possible values of Δ represent the four possibilities 1 to 4 outlined at the beginning of this section. Let $(Y_1, \Delta_1), \dots, (Y_n, \Delta_n)$ be the observations on the n individuals in the study. One may view the observations as independent realizations of (Y, Δ) . We use the Kiefer–Wolfowitz theory of **nonparametric statistics** to estimate π_M and $\gamma_M = \int_{[0, M]} \bar{F}(t) dt$. For Kiefer–Wolfowitz theory, see [4, 6]. For technical derivation of the estimates, see [1, 2].

Let \hat{G} and \hat{F} be the Kaplan–Meier estimators of G and F , respectively, based on the data $(Y_1, \Delta_1), \dots, (Y_n, \Delta_n)$ see [4, 5, 8]. The **generalized maximum likelihood estimators** à la Kiefer–Wolfowitz of π_M and γ_M are given by

$$\hat{\pi}_M = \frac{1}{n} \sum_{i=1}^n I(\Delta_i = 1) (\hat{G}(Y_i -))^{-1} \quad (7)$$

and

$$\hat{\gamma}_M = \int_{[0, M]} \hat{F}(t) dt, \quad (8)$$

respectively. Consequently, the generalized maximum likelihood estimate of ρ_M is given by

$$\hat{\rho}_M = \frac{\hat{\pi}_M}{\hat{\gamma}_M}. \quad (9)$$

The asymptotic distribution of $\hat{\rho}_M$ is normal and details of the asymptotic distribution can be found in [2].

A Numerical Example

Suppose a sample of size $n = 7$ with $M = 7$ yields the following data: $(Y_1, \Delta_1) = (1.7, 1)$; $(Y_2, \Delta_2) = (2.3, 0)$; $(Y_3, \Delta_3) = (4.4, -1)$; $(Y_4, \Delta_4) = (4.5, 1)$; $(Y_5, \Delta_5) = (4.9, -1)$; $(Y_6, \Delta_6) = (6.0, 0)$; $(Y_7, \Delta_7) = (6.1, 0)$. The data are arranged according to increasing values of Y_i s. The first individual died of cancer at time 1.7, the second died of other causes at time 2.3, the third left the study at time 4.4, the fourth died of cancer at time 4.5, the fifth left the study at time 4.9, the sixth died of other causes at time 6.0, and the seventh died of other causes at time 6.1. The Kaplan–Meier estimate of the distribution of C gives probability mass 3/15 at time 4.4, probability mass 4/15 at time 4.9, and probability mass 8/15 to the interval $(4.9, \infty)$. Consequently, $\hat{\pi}_M = \frac{9}{28}$. The Kaplan–Meier estimate of the distribution function F gives the following distribution (for this, data on both causes of death need to be combined):

X :	1.7	2.3	4.5	6.0	6.1
Pr :	8/56	8/56	10/56	15/56	15/56

Consequently,

$$\begin{aligned} \hat{\gamma}_M &= 1 \times 1.7 + (48/56) \times 0.6 + (40/56) \times 2.2 \\ &\quad + (30/56) \times 1.5 + (15/56) \times 0.1 = 4.08. \end{aligned}$$

The generalized maximum likelihood estimate of ρ_M is given by $(9/28)/4.08 = 0.08$.

References

- [1] Babu, G.J., Rao, C.R. & Rao, M.B. (1991). Nonparametric estimation of survival functions under dependent competing risks, in *Nonparametric Functional Estimation and Related Topics*, G. Roussas, ed., Kluwer Academic Publishers, New York, pp. 431–441.
- [2] Babu, G.J., Rao, C.R. & Rao, M.B. (1992). Nonparametric estimation of specific occurrence/exposure rate in risk

-
- and survival analysis, *Journal of the American Statistical Association* **87**, 84–89.
- [3] Howe, G.R. (1983). Confidence interval estimation for the ratio of simple and standardized rates in cohort studies, *Biometrics* **39**, 325–331.
- [4] Johansen, S. (1978). The product limit estimator as a maximum likelihood estimator, *Scandinavian Journal of Statistics* **5**, 195–199.
- [5] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.
- [6] Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *The Annals of Mathematical Statistics* **27**, 887–906.
- [7] London, D. (1988). *Survival Models and Their Estimation*, ACTEX Publications, Winsted, CT.
- [8] Miller, R.G. Jr (1981). *Survival Analysis*, Series in Probability and Mathematical Statistics, John Wiley, New York.

GUTTI JOGESH BABU & M. BHASKARA RAO