Censored Data

In classical statistics, the observations are frequently assumed to include *independent* random variables $X_1, \ldots, X_n$, with X_i having the density function

$$f_i^\theta(x) = \alpha_i^\theta(x) S_i^\theta(x),$$

where $\alpha_i^\theta(x)$ is the hazard function, $S_i^\theta(x)$ is the survival function, and θ is a vector of unknown parameters (*see* **Survival Distributions and Their Characteristics**). Then inference on θ may be based on the **likelihood** function,

$$L(\theta) = \prod_i f_i^\theta(X_i),$$

in the usual way. In survival analysis, however, one can rarely avoid various kinds of incomplete observation. The most common form of this is *right-censoring* where the observations are

$$(\tilde{X}_i, D_i), \quad i = 1, \ldots, n, \tag{1}$$

where D_i is the indicator $I\{\tilde{X}_i = X_i\}$, and $\tilde{X}_i = X_i$, the true survival time, if the observation of the lifetime of i is uncensored and $\tilde{X}_i = U_i$, the time of right-censoring, otherwise. Thus, $D_i = 1$ indicates an uncensored observation, $D_i = 0$ corresponds to a right-censored observation. Other kinds of incomplete observation will be discussed below.

Survival analysis, then, deals with ways in which inference on θ may be performed based on the censored sample (1). We would like to use the function

$$\begin{aligned} L^c(\theta) &= \prod_i \alpha_i^\theta(\tilde{X}_i)^{D_i} S_i^\theta(\tilde{X}_i) \\ &= \prod_i f_i^\theta(\tilde{X}_i)^{D_i} S_i^\theta(\tilde{X}_i)^{1-D_i} \end{aligned} \tag{2}$$

for inference, but there are two basic problems:

1. The presence of censoring may alter the hazard function of the lifetime X_i, i.e. the conditional distribution of X_i, given that i is alive at $t(X_i \geq t)$ *and uncensored at* t $(U_i \geq t)$, may be different from what it was in the uncensored case, i.e. just given $X_i \geq t$ (*dependent censoring*).
2. The observed right-censoring times, U_i, may contain information on θ (*informative censoring*).

An example of a dependent censoring scheme would be if, in a clinical trial with survival times as the outcome variables, one removed patients from the study while still alive and when they appeared to be particularly ill (or particularly well), so that patients remaining at risk are not representative of the group that would have been observed in the absence of censoring. In other words, dependent censoring represents a dynamic version of what in an epidemiologic context would be termed a **selection bias**. An example is provided below (Example 1). Mathematical formulations of independent censoring (conditions on the joint distribution of X_i and U_i) may be given, and it may be shown that several frequently used models for the generation of the times of right-censoring satisfy these conditions. The difficulty in a given practical context lies in the fact that the conditions may be impossible to verify, since they refer to quite hypothetical situations.

The second concept mentioned, noninformative censoring, is simpler and relates to the fact that if censoring is informative, then a more efficient inference on θ may be obtained than the one based on (2); see below.

Independent Censoring

The general definition of independent censoring given by Andersen et al. [2], Section III.2.2 for multivariate counting processes has the following interpretation for the special case of survival analysis with time-fixed **covariates**. The basic (uncensored) model is that conditional on covariates $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ the lifetimes $X_1, \ldots, X_n$ are independent, X_i having the hazard function

$$\alpha_i^\theta(t|\mathbf{Z}_i) \approx P^{\theta\phi}(X_i \in I_{\mathrm{d}t}|X_i \geq t, \mathbf{Z})/\,\mathrm{d}t. \tag{3}$$

Here, $I_{\mathrm{d}t}$ is the interval $[t, t + \mathrm{d}t)$ and $P^{\theta\phi}$ is the joint distribution of $X_1, \ldots, X_n, \mathbf{Z}$ *and* the censoring times. Note that the hazard function only depends on θ, i.e. ϕ is a nuisance parameter. Because of the conditional independence of X_i it follows that

$$P^{\theta\phi}(X_i \in I_{\mathrm{d}t}|\mathcal{F}_{t-}) \approx \alpha_i^\theta(t|\mathbf{Z}_i) I\{X_i \geq t\}\,\mathrm{d}t,$$

where the *history* $\mathcal{F}_{t-}$ contains $\mathbf{Z}$ and all information on $X_1, \ldots, X_n$ from the interval $[0, t)$, i.e. values of X_i for i with $X_i < t$ and the information that $X_j \geq t$

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a11008

for j with $X_j \geq t$. Let there now be given right-censoring times $U_1, \ldots, U_n$ and define the enlarged history $\mathcal{G}_t$ as the one containing $\mathcal{F}_t$ *and* all information on $U_1, \ldots, U_n$ from the interval $[0, t]$, i.e. values of $U_i \leq t$ and the information that $U_j \geq t$ for those j where $U_j \geq t$. The condition for independent censoring is then that

$$P^{\theta\phi}(X_i \in I_{\mathrm{d}t} | \mathcal{F}_{t-}) = P^{\theta\phi}(X_i \in I_{\mathrm{d}t} | \mathcal{G}_{t-}). \qquad (4)$$

It follows that *simple type I* censoring, where all U_i are equal to a *fixed* time, u_0, and *simple type II* censoring, where all U_i are equal to the kth smallest lifetime $X_{(k)}$ for some k between 1 and n, are both independent, since the right-censoring times in these cases give rise to no extra randomness in the model; that is, $\mathcal{F}_t = \mathcal{G}_t$.

In some models, $U_1, \ldots, U_n$ are assumed to be independent given $\mathbf{Z}$ and $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are independent identically distributed (iid). Then the assumption (4) reduces to

$$\alpha_i^{\theta}(t|\mathbf{Z}_i) \approx P^{\theta\phi}(X_i \in I_{\mathrm{d}t} | X_i \geq t, U_i \geq t, \mathbf{Z}) / \mathrm{d}t \qquad (5)$$

and it is fulfilled, e.g. if U_i and X_i are independent given Z_i. This is, for instance, the case in the *simple random* censorship model where $U_1, \ldots, U_n$ are *iid* and independent of $X_1, \ldots, X_n$.

Some authors take the condition (5) (which is less restrictive than (4)) as the definition of independent censoring; see, for example, [6], p. 128. However, (4) may be generalized to other models based on counting processes and both (4) and (5) cover the most frequently used mathematical models for the right-censoring mechanisms. These include both the models already mentioned, i.e. simple type I, type II and random censorship and various generalizations of these (e.g. progressive type I censorship (cf. Example 2, below), general random censorship, and randomized progressive type II censorship; see, [2, Section III.2.2]). Earlier contributions to the definition and discussion of independent censoring are the monographs by Kalbfleisch & Prentice [13], p. 120 and Gill [7], Theorem 3.1.1 and the papers by Cox [5], Williams & Lagakos [16], Kalbfleisch & MacKay [12] and Arjas & Haara [3], all of whom give definitions that are close or equivalent to (5). Another condition for independent censoring, stronger than (5) but different from (4), is discussed by Jacobsen [11].

From (4) and (5) it is seen that censoring is allowed to depend on covariates as long as these are included in the model for the hazard function of the lifetime distribution in (3). Thus, an example of a *dependent* censoring scheme is one where the distribution of U_i depends on some covariates that are not included there. This is illustrated in the following example.

Example 1: Censoring Depending on Covariates

Suppose that iid binary covariates, $Z_1, \ldots, Z_n$, have

$$P^{\theta\phi}(Z_i = 1) = 1 - P^{\theta\phi}(Z_i = 0) = \phi,$$

and that $X_1, \ldots, X_n$ are iid with survival function $S(t)$. The **Kaplan–Meier estimator** $\widehat{S(t)}$ based on the X_i then provides a consistent estimate of $\theta = S(\cdot)$, the marginal distribution of X_i. This may be written as

$$S(t) = \phi S_1(t) + (1 - \phi) S_0(t),$$

where $S_j(t)$, for $j = 0, 1$, is the conditional distribution given $Z_i = j$. Note that these may be different, e.g. $S_1(t) < S_0(t)$ if individuals with $Z_i = 1$ are at higher risk than those with $Z_i = 0$. Define now the right-censoring times U_i by

$$U_i = u_0, \text{ if } Z_i = 1, \qquad U_i = +\infty, \text{ if } Z_i = 0.$$

Then, for $t < u_0$ the Kaplan–Meier estimator will still consistently estimate $S(t)$, while for $t > u_0$, $\widehat{S(t)}/\widehat{S(u_0)}$ will estimate $S_0(t)/S_0(u_0)$. If, however, the covariate is included in the model for the distribution of X_i, i.e. $\theta = [S_0(\cdot), S_1(\cdot)]$, then $\widehat{S_j(t)}$, the Kaplan–Meier estimator based on individuals with $Z_i = j$, $j = 0, 1$, will consistently estimate the corresponding $S_j(t)$, also based on the right-censored sample (though, of course, no information will be provided about $S_1(t)$ for $t > u_0$).

It is seen that censoring is allowed to depend on the *past* and on external (in the sense of conditionally independent) random variation. This means that if, in a lifetime study, sex and age are included as covariates, then a right-censoring scheme, where, say, every year, one out of the two oldest women still alive and uncensored is randomly (e.g. by flipping a coin) chosen to be censored, is independent. However, a right-censoring scheme depending on the *future* is dependent. This is illustrated in the following example.

DOI: 10.1002/0470011815.b2a11008

Example 2: Censoring Depending on the Future

Suppose that, in a clinical trial, patients are accrued at calendar times $T_1, \ldots, T_n$ and that they have iid lifetimes $X_1, \ldots, X_n$ (since entry) independent of the entry times. The study is terminated at calendar time t_0 and the entry times are included in the observed history, i.e. $Z_i = T_i$ in the above notation. If, at t_0, all patients are traced and those still alive are censored (at times $U_i = t_0 - T_i$) and, for those who have died, their respective lifetimes, X_i, are recorded, then this right-censoring is independent (being *deterministic*, given the entry times, so-called progressive type I censoring).

Consider now, instead, the situation where patients are only seen, for instance, every year, i.e. at times $T_i + 1, \ldots, T_i + k_i \leq t_0$ and suppose that if a patient does not show up at a scheduled follow-up time, then this is because he or she has died since last follow-up and the survival time is obtained. Suppose, further, that for the patients who are alive at the time, $T_i + k_i$, of their last scheduled follow-up, and who die before time t_0, there is a certain probability, ϕ, of obtaining information on the failure, whereas for those who survive past t_0 nothing new is learnt. If these extra survival times are included in the analysis and if everyone else is censored at k_i, then the right-censoring scheme is dependent. This is because the fact that patient i is censored at k_i tells the investigator that this patient is likely not to die before t_0 and the right-censoring, therefore, depends on the future. To be precise, if the average probability of surviving past t_0, given survival until the last scheduled follow-up time is $1 - \pi$, then the probability of surviving past t_0, given censoring at the time of the last scheduled follow-up, is $(1 - \pi)/[\pi(1 - \phi) + 1 - \pi]$, which is 1 if $\phi = 1$, $1 - \pi$ if $\phi = 0$, and between $1 - \pi$ and 1, otherwise.

If, alternatively, everyone still alive at time $T_i + k_i$ were censored at k_i, then the censoring would be independent (again being deterministic given the entry times).

Another censoring scheme that may depend on the future relative to "time on study", but not relative to calendar time, occurs in connection with testing with replacement, see, for example, [8].

Let us finally in this section discuss the relation between independent right-censoring and **competing risks**. A competing risks model with two causes of failure, d and c, is an inhomogeneous **Markov process** $W(\cdot)$ with a transient state 0 ("alive"), two absorbing states d and c and two cause-specific hazard functions $\alpha_{0d}(t)$ and $\alpha_{0c}(t)$, e.g. Andersen et al. [1]. This generates two random variables:

$$X = \inf[t : W(t) = d]$$

and

$$U = \inf[t : W(t) = c],$$

which are incompletely observed since the observations consist of the transition time $\widetilde{X} = X \wedge U$ and the state $W(\widetilde{X}) = d$ or c reached at that time. The elusive concept of "independent competing risks" (e.g. [13, Section 7.2]) now states that in a population where the risk c is not operating, the hazard function for d is still given by $\alpha_{0d}(t)$. This condition is seen to be equivalent to censoring by U being independent. However, since the population where a given cause of failure is eliminated is usually completely hypothetical in a biological context, this formal equivalence between the two concepts is of little help in a practical situation and, as is well known from the competing risks literature (e.g. [4, 15], and [13, Chapter 7]), statistical independence of the random variables X and U cannot be tested from the incomplete observations $[\widetilde{X}, W(\widetilde{X})]$. What can be said about the inference on the parameter $\theta = \alpha_{0d}(\cdot)$ based on these data is that consistent estimation of θ may be obtained by formally treating failures from cause c as right-censorings, but that this parameter has no interpretation as the d failure rate one would have had in the hypothetical situation where the cause c did not operate.

For the concept of independent censoring to make sense, the "uncensored experiment" described in the beginning of this section should, therefore, be meaningful.

Likelihoods: Noninformative Censoring

The right-censored data will usually consist of

$$(\widetilde{X}_i, D_i, \mathbf{Z}_i; i = 1, \ldots, n)$$

and, under independent censoring, the likelihood can then be written using **product-integral** notation

$$L(\theta,\phi) = P^{\theta\phi}(\mathbf{Z}) \prod_i \prod_{t>0} \alpha_i^{\theta}(t)^{D_i(\mathrm{d}t)} [1 - \alpha_i^{\theta}(t)\mathrm{d}t]^{1-D_i(\mathrm{d}t)}$$
$$\times \gamma_i^{\theta\phi}(t)^{C_i(\mathrm{d}t)} [1 - \gamma_i^{\theta\phi}(t)\,\mathrm{d}t]^{1-C_i(\mathrm{d}t)}. \quad (6)$$

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.
This article is © 2005 John Wiley & Sons, Ltd.
This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a11008

Here, $D_i(\mathrm{d}t) = I\{X_i \in I_{\mathrm{d}t}\}$, $C_i(\mathrm{d}t) = I\{U_i \in I_{\mathrm{d}t}\}$, and $\alpha_i^{\theta}(t)$ and $\gamma_i^{\theta\phi}(t)$ are the conditional hazards of failure and censoring, respectively, given the past up until t- (including covariates). The likelihood (6) may be written as

$$L(\theta, \phi) = L^c(\theta)L^*(\theta, \phi),$$

with $L^c(\theta)$ given by (2) and where the contributions from censoring and covariates are collected in $L^*(\theta, \phi)$. Thus, the function (2), which is usually taken as the standard censored data likelihood, is, under independent censoring, a **partial likelihood** on which a valid inference on θ may be based. It is only the full likelihood for θ if $L^*(\theta, \phi)$ does not depend on θ, which is the case if censoring (and covariates) are *noninformative*. Thus, noninformative censoring is a statistical concept (while the concept of independent censoring is *probabilistic*) and means that the conditional hazard of censoring $\gamma_i^{\theta\phi}(t)$ does, in fact, not depend on θ, the parameter of interest.

An example of an informative right-censoring scheme could be in a study with two competing causes of failure and where only one of the two cause-specific failure rates is of interest; if the two cause-specific failure rates are *proportional* (as in the so-called Koziol–Green model for random censoring, [14]), then the failures from the second cause (the censorings) will carry information on the shape of the hazard function for the failure type of interest. It is, however, important to notice that even if the censoring is informative, then inference based on (2) will still be valid (though not fully efficient) and as it is usually preferable to make as few assumptions as possible about the distribution of the right-censoring times, the (partial) likelihood (2) is often the proper function to use for inference.

Other Kinds of Incomplete Observation

When observation of a survival time, X, is right-censored, then the value of X is only known to belong to an interval of the form $[U, +\infty)$. This is by far the most important kind of censoring for survival data, but not the only one. Thus, the observation of X is **interval-censored** if the value of X is only known to belong to an interval $[U, V)$ and it is said to be *left-censored* if $U = 0$.

It was seen above that under independent right-censoring a right-censored observation, U_i, contributed to the partial likelihood function with a factor $S^{\theta}(U_i)$, which was also the contribution to the *full* likelihood under noninformative censoring. Similarly, concepts of independent and noninformative interval-censoring may be defined as leading to a contribution of $S^{\theta}(U_i) - S^{\theta}(V_i)$ to, respectively, the partial and the full likelihood. These concepts have received relatively little attention in the literature; however, this way of viewing censoring is closely related to the concept of **coarsening at random**.

Formally, **grouped data**, where for each individual the lifetime is known only to belong to one of a fixed set of intervals $[u_{k-1}, u_k)$ with $0 = u_0 < u_1 < \cdots < u_m = +\infty$, are also interval-censored. However, the fact that the intervals are the same for everyone simplifies the likelihood to a binomial-type likelihood with parameters $p_k^{\theta} = S^{\theta}(u_{k-1}) - S^{\theta}(u_k), k = 1, \ldots, m$.

Let us finally remark that while, following Hald [9; 10, p. 144], *censoring* occurs when we are able to sample a complete population but individual values of observations above (or below) a given value are not specified, truncation corresponds to sampling from an incomplete population, i.e. from a conditional distribution (*see* **Truncated Survival Times**). Left-truncated samples, where an individual is included only if his or her lifetime exceeds some given lower limit, also occur frequently in the analysis of survival data, especially in epidemiologic studies where hazard rates are often modeled as a function of age and where individuals are followed only from age at diagnosis of a given disease or from age at employment in a given factory.

References

[1] Andersen, P.K., Abildstrom, S. & Rosthøj, S. (2002). Competing risks as a multistate model. *Statistical Methods in Medical Research.* **11**, 203–215.

[2] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

[3] Arjas, E. & Haara, P. (1984). A marked point process approach to censored failure data with complicated covariates, *Scandinavian Journal of Statistics* **11**, 193–209.

[4] Cox, D.R. (1959). The analysis of exponentially distributed life-times with two types of failure, *Journal of the Royal Statistical Society, Series B* **21**, 411–421.

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.
This article is © 2005 John Wiley & Sons, Ltd.
This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a11008

[5] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[6] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[7] Gill, R.D. (1980). Censoring and stochastic integrals, *Mathematical Centre Tracts* **124**, Mathematisch Centrum, Amsterdam.

[8] Gill, R.D. (1981). Testing with replacement and the product limit estimator, *Annals of Statistics* **9**, 853–860.

[9] Hald, A. (1949). Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point, *Skandinavisk Aktuarietidsskrift* **32**, 119–134.

[10] Hald, A. (1952). *Statistical Theory with Engineering Applications*. Wiley, New York.

[11] Jacobsen, M. (1989). Right censoring and martingale methods for failure time data, *Annals of Statistics* **17**, 1133–1156.

[12] Kalbfleisch, J.D. & MacKay, R.J. (1979). On constant-sum models for censored survival data, *Biometrika* **66**, 87–90.

[13] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[14] Koziol, J.A. & Green, S.B. (1976). A Cramér-von Mises statistic for randomly censored data, *Biometrika* **63**, 465–474.

[15] Tsiatis, A.A. (1975). A nonidentifiability aspect of the problem of competing risks, *Proceedings of the National Academy of Sciences* **72**, 20–22.

[16] Williams, J.A. & Lagakos, S.W. (1977). Models for censored survival analysis: constant sum and variable sum models, *Biometrika* **64**, 215–224.

(*See also* **Survival Analysis, Overview**)

PER KRAGH ANDERSEN

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.
This article is © 2005 John Wiley & Sons, Ltd.
This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a11008