Basic Concepts

Md. Moyazzem Hossain
Professor
Department of Statistics and Data Science
Jahangirnagar University
Dhaka, Bangladesh
Email: hossainmm@juniv.edu

Biostatistics

Biostatistics (a combination of biology and statistics; sometimes referred to as **biometry** or **biometrics**) is the application of statistics to a wide range of topics in biology. The science of biostatistics encompasses the design of biological experiments, especially in medicine, pharmacy, agriculture and fishery; the collection, summarization, and analysis of data from those experiments; and the interpretation of, and inference from, the results.

Survival time

- Time from a starting point until an event of interest occurs. For example,
  - Time from birth to death
  - Duration of marriage
  - Time from treatment to relapse of a specific disease
- The starting point is sometimes harder to define than the endpoint
  - When is the actual beginning of a disease? Using time of diagnosis instead?
- Endpoint not necessarily something negative, such as illness or death
  - Can be, e.g., recovery from a disease

Survival time, cont.

- Also known as time-to-event data
- Rarely useful to calculate mean survival time
  - Requires that endpoint actually occurs and is observed for all subjects
  - Survival data rarely normally distributed
- Survival data analyzed by using special methods

Notation

- We denote by a capital **T** the random variable for a person's survival time.
- Since T denotes time, its possible values include all nonnegative numbers; that is, **T** can be any number equal to or greater than zero.

$$T = \text{survival time } (T \geq 0)$$

random variable

Notation

- We denote by a small letter **t** any specific value of interest for the random variable capital **T**.
- For example, if we are interested in evaluating whether a person survives for more than 5 years after undergoing cancer therapy, small **t** equals 5; we then ask whether capital **T** exceeds 5.

$$\text{Survives} > 5 \text{ years?}$$
$$T > t = 5$$

Notation

- We denote the small letter **d** to define a (0,1) random variable indicating either failure or censorship.
- That is, **d=1** for **failure** if the event occurs during the study period, or **d=0** if the survival time is **censored** by the end of the study period.

$$d = (0,\ 1) \text{ random variable}$$

$$= \begin{cases} 1 & \text{if failure} \\ 0 & \text{censored} \end{cases}$$

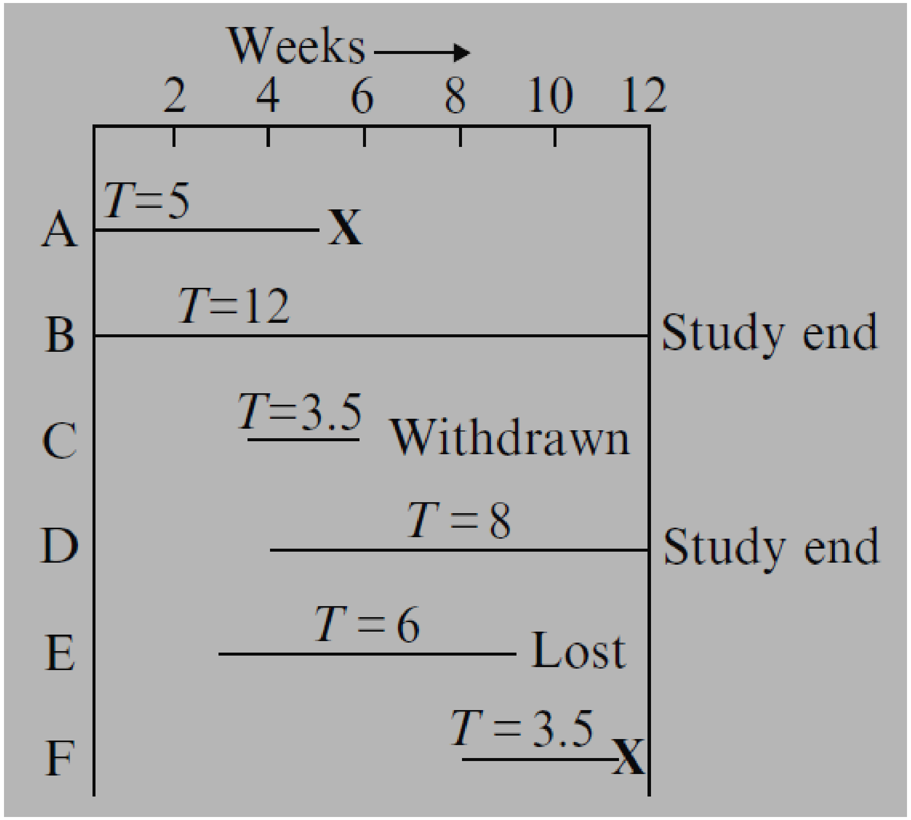
Censoring

- Exact time of the endpoint is not known
  - Event of interest not observed for all subjects at the end of the study
- Survival time is only partly known (e.g., “at least as large as”)
- Often due to data collected during a limited period of time
- Three main types
  - Right-censoring
  - Left-censoring
  - Interval-censoring

Reasons for Censoring

- A person does not experience the event before the study ends;
- A person is lost to follow-up during the study period;
- A person withdraws from the study because of death (if death is not the event of interest) or some other reason (e.g., adverse drug reaction or other competing risk)

Example of Censoring

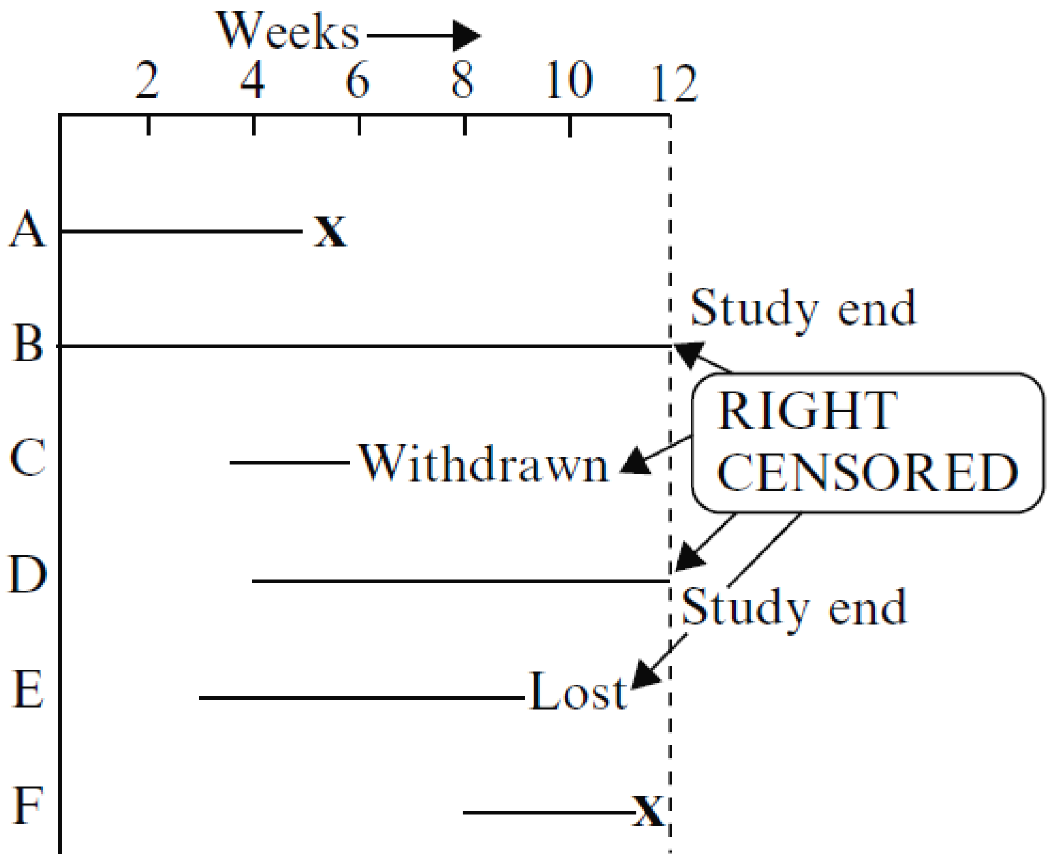


In **summary**, of the six persons observed, two get the event (persons A and F) and four are censored (B, C, D, and E).

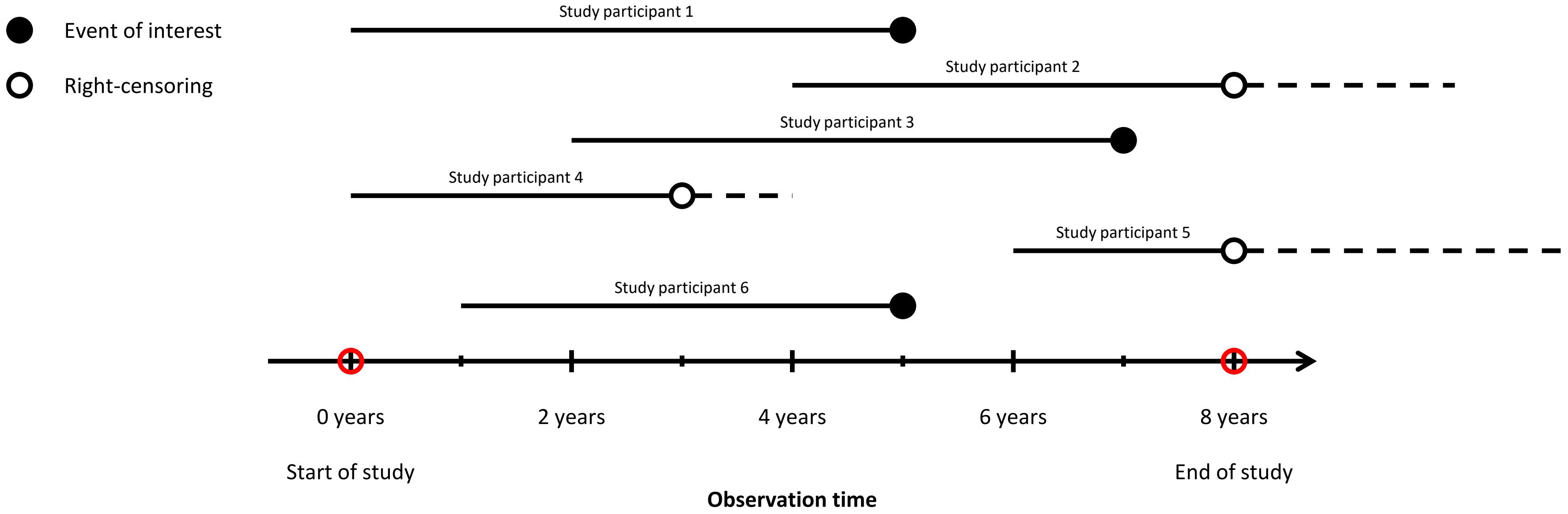
X $\Longrightarrow$ Event occurs

Right Censoring

Right-censored: true survival time is equal to or greater than observed survival time



Right Censoring, cont.



Right Censoring, cont.

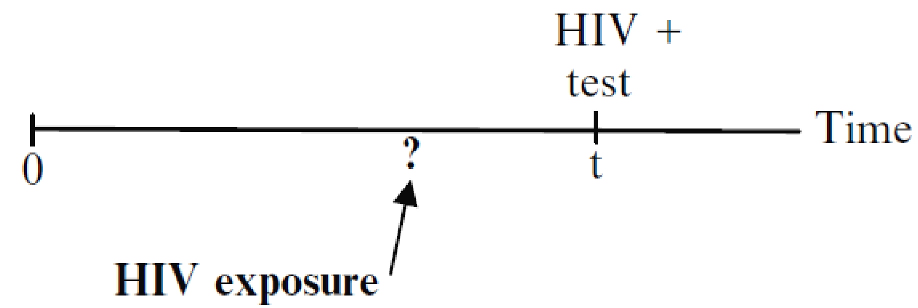
- Right-censoring due to study termination or loss to follow-up
- Several possible reasons for being lost to follow-up
  - Not responding to questionnaires or attending scheduled hospital visits
  - Study withdrawal
  - Moving or emigration
  - Death (by a cause other than that being studied)
- A common phenomenon
  - Routinely handled in survival analysis

Right Censoring, cont.

- Right-censoring due to study termination or loss to follow-up
- Several possible reasons for being lost to follow-up
  - Not responding to questionnaires or attending scheduled hospital visits
  - Study withdrawal
  - Moving or emigration
  - Death (by a cause other than that being studied)
- A common phenomenon
  - Routinely handled in survival analysis

Left Censoring

Left-censored: true survival time is less than or equal to the observed survival time



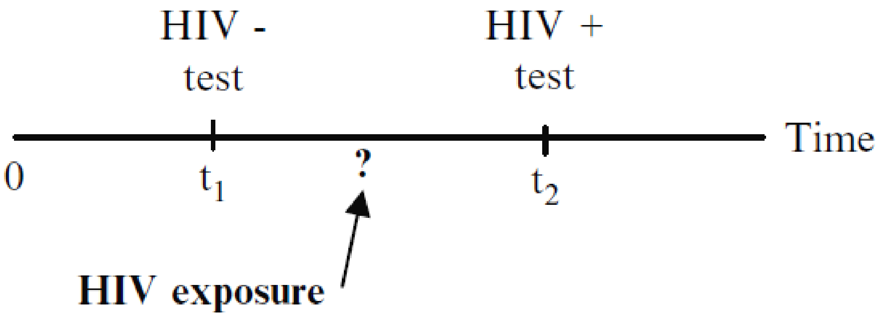
- If a person is left-censored at time t, we know they had an event between time 0 and t, but we do not know the exact time of event.

Event occurs between 0 and t
but
do not know the exact time.

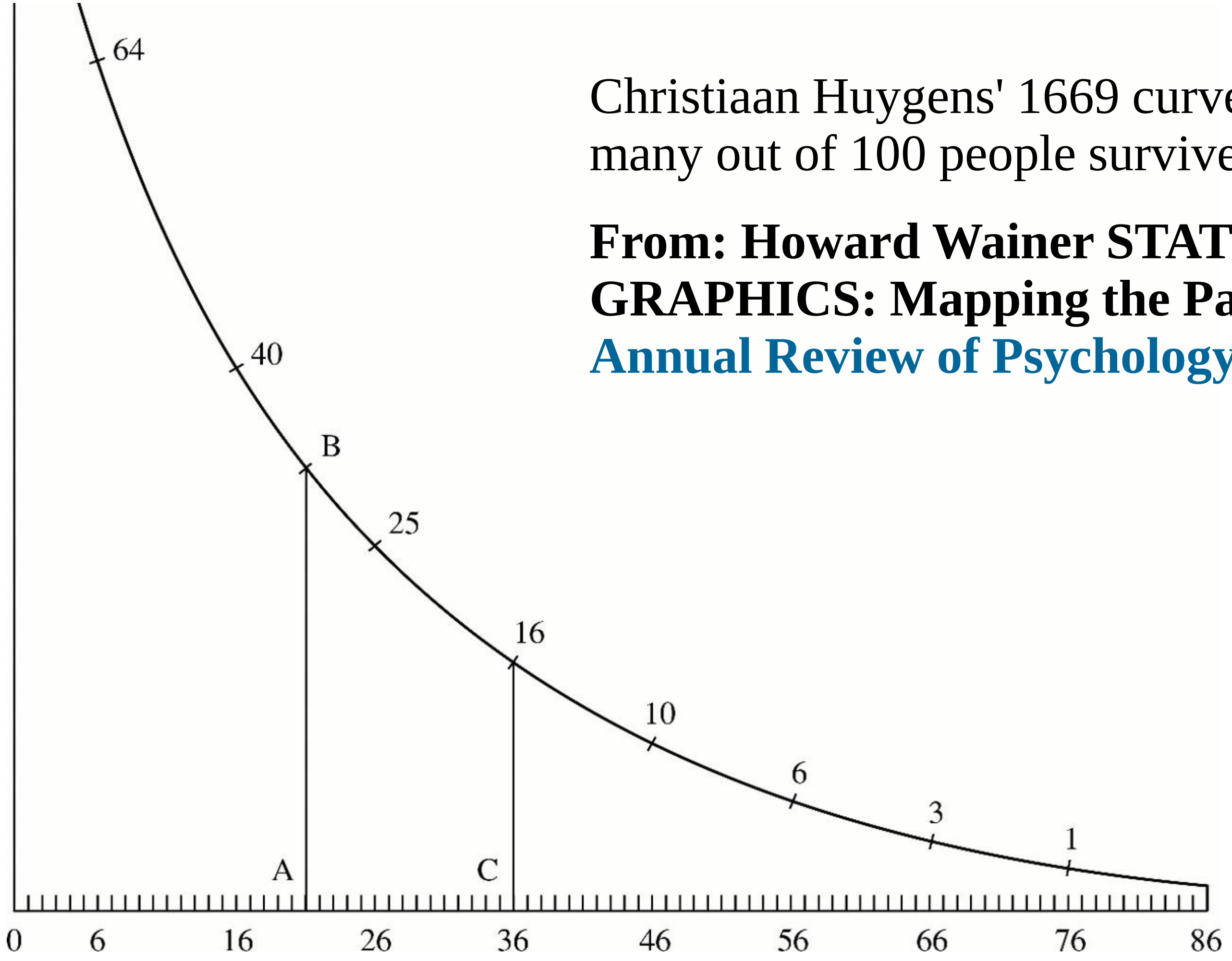
Interval Censoring

Interval-censored: true survival time is within a known time interval

A subject may have had two HIV tests, where he/she was HIV negative at the time (say, t1) of the first test and HIV positive at the time (t2) of the second test. In such a case, the subject's true survival time occurred after time t1 and before time t2, i.e., the subject is interval censored in the time interval (t1, t2).



Early example of survival analysis, 1669



Christiaan Huygens' 1669 curve showing how many out of 100 people survive until 86 years.

From: Howard Wainer STATISTICAL GRAPHICS: Mapping the Pathways of Science. **Annual Review of Psychology**. Vol. 52: 305-335

What is survival analysis?

- Statistical methods for analyzing longitudinal data on the occurrence of events.
- Events may include death, injury, beginning of illness, recovery from illness (binary variables) or transition above or below the clinical threshold of a meaningful continuous variable.
- Accommodates data from randomized clinical trial or cohort study design.

Objectives of survival analysis

- **Estimate time-to-event for a group of individuals**, such as time until second heart-attack for a group of MI patients.
- **To compare time-to-event between two or more groups**, such as treated vs. placebo MI patients in a randomized controlled trial.
- **To assess the relationship of co-variables to time-to-event**, such as: does weight, insulin resistance, or cholesterol influence survival time of MI patients?

Note: expected time-to-event = 1/incidence rate

Why use survival analysis?

- Logistic regression can predict the presence or absence of events but not time until events and it can not handle time dependent covariates.

- Linear regression can not handle censoring well or time dependent covariates or the fact that time can only be positive.

Survival Analysis Steps

- Get some data and make sure it is valid.
- Estimate the survival/hazard functions.
- Compare the functions between groups.
- Assess the impact of predictors on survival rates.

Survival Function or Curve

- **Let *T* denote the survival time**
- **S(t) = P(surviving longer than time t)= *P(T > t)***
- ***The function S(t) is also known as the cumulative survival function. 0$\leq$ S(t) $\leq$ 1***

$$\hat{S}(t) = \frac{\text{number of patients surviving longer than } t}{\text{total number of patients}}$$

$$\therefore S(t) = \int_t^{\infty} f(x)\,dx\ ; \quad \text{where} \quad f(x) = \text{pdf}$$

Survival function

- The goal of survival analysis is to estimate and compare survival experiences of different groups.

- Survival experience is described by the cumulative survival function:

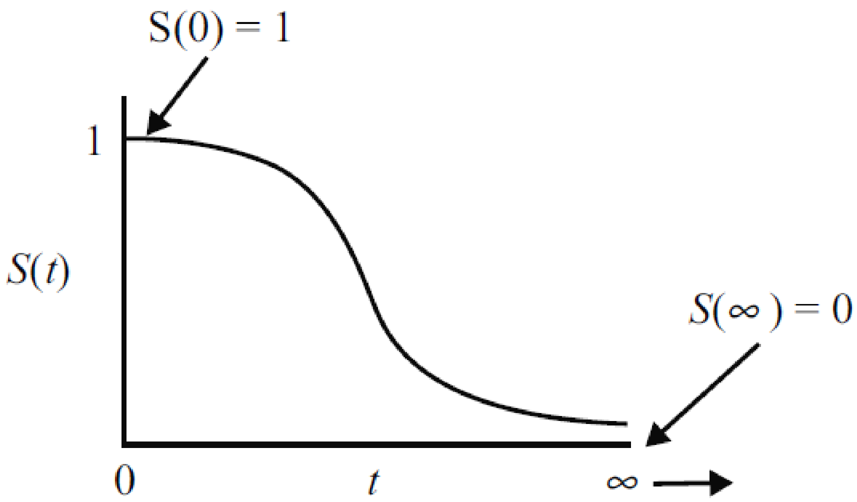
$$S(t) = 1 - P(T \le t) = 1 - F(t)$$

F(t) is the CDF of f(t), and is "more interesting" than f(t).

- Example: If t=100 years, S(t=100) = probability of surviving beyond 100 years.

Characteristics survivor functions

- They are nonincreasing; that is, they head downward as t increases;
- At time t=0, S(0)=1; that is, at the start of the study, since no one has gotten the event yet, the probability of surviving past time 0 is one;
- At time t=1, S(1)=0; that is, theoretically, if the study period increased without limit, eventually nobody would survive, so the survivor curve must eventually fall to zero.



Hazard Function

- **The hazard function *h(t)* of survival time *T* gives the *conditional failure rate***
- **The hazard function is also known as the *instantaneous failure rate*, *force of mortality*, and *age-specific failure rate***
- ***The hazard function gives the risk of failure per unit of time during the aging process***

Hazard Function

Given

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}$$

Conditional probabilities: $P(A|B)$

$P(t \le T < t + \Delta t \mid T \ge t)$
$=$ P(individual fails in the interval
$[t, t + \Delta t] \mid$ survival up to time t)

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

Cumulative Hazard Function

The cumulative hazard function is defined as

$$H(t) = \int_0^t h(x)dx$$

Relationship of S(t) and h(t)

$$f(t) = S(t) \times h(t)$$

Relations

$$\text{Hazard from density and survival}: h(\text{t}) = \frac{f(t)}{S(t)}$$

$$\text{Survival from density}: \text{S(t)} = \int_t^{\infty} f(u)du$$

$$\text{Density from survival}: \text{f}(t) = -\frac{dS(t)}{dt}$$

$$\text{Density from hazard}: \text{f}(t) = h(t)e^{(-\int_0^t h(u)du)}$$

$$\text{Survival from hazard}: \text{S(t)} = e^{(-\int_0^t h(u)du)}$$

$$\text{Hazard from survival}: h(\text{t}) = -\frac{\text{d}}{\text{dt}}\ln S(t)$$

Example

Suppose, $f(t) = e^{-t} \quad ; \quad t \geq 0$

Find S(t), and h(t).

We know that, $S(t) = 1 - F(t)$

$$F(t) = \int_0^t f(x)dx = \int_0^t e^{-x}dx = 1 - e^{-t}$$

$$S(t) = 1 - \left(1 - e^{-t}\right) = e^{-t}$$

$$f(t) = h(t) \times S(t)$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{e^{-t}}{e^{-t}} = 1$$

Practice

Suppose, $f\left(t\right)=\lambda e^{-\lambda t} \quad ; \quad t\geq 0$

Find S(t), and h(t).

Thanks for Your Attention