Weeks 11-12

Analysis of Survival Data

Dr. Esam Mahdi

Department of Mathematics, Statistics and Physics

Qatar University

Learning Objectives

- Distinguish between retrospective studies (case-control studies) and Prospective studies (cohort studies).
- Distinguish between categorical and survival data (why we can’t use logistic regression in survival data?).
- Deal with survival data and basic survival analysis.
- Learn about Survival Data; Kaplan-Meier curve and comparison of survival distributions.
- Simple regression and correlation.
- Pair-matched case-control studies: Model and analysis.

Retrospective Versus Prospective Studies

- **Biomedical research data may come from two fundamental designs:**
  1) ***Retrospective studies*** (also called case-control studies): gather past data from selected cases and controls to determine differences, if any, in exposure to a suspected risk factor.
     - **The advantages** of a retrospective study are that it is economical and provides answers to research questions relatively quickly because the cases are already available.
     - **Disadvantages** are due to the inaccuracy of the exposure histories and uncertainty about the appropriateness of the control sample.
  2) ***Prospective studies*** (also called cohort studies) are epidemiological designs in which one enrolls a group of persons and follows them over certain periods of time; examples include occupational mortality studies and clinical trials.
- The case-control study is focused on a particular disease.
- The cohort study design focuses on a particular exposure rather than a particular disease as in case-control studies.

Why not Logistic Regression?

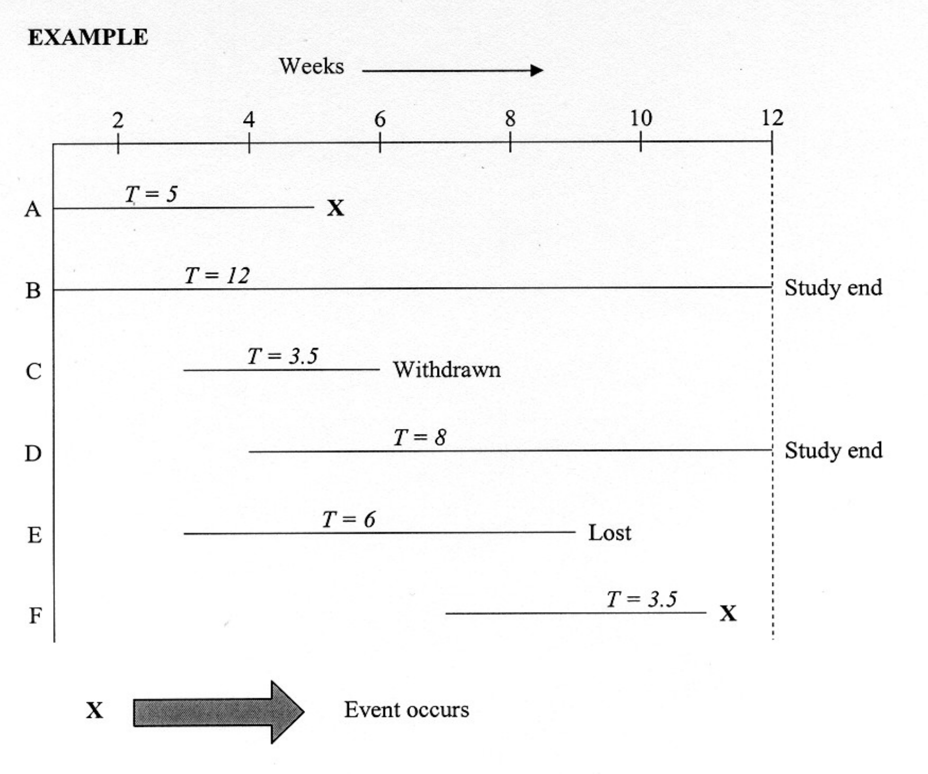
- A cohort study consists of randomized clinical trials where follow-up starts from the date of enrollment and randomization of each subject.
- **Survival analysis**, which was developed to deal with data resulting from cohort studies, is also focused on the occurrence of an *event*, such as death or relapse of a disease, after some initial treatment—a binary outcome.
- The logistic regression can't be used to model a survival data. The reasons are:
  a) Survival data studies have staggered entry and subjects are followed for varying lengths of time; they do not have the same probability for the event to occur even if they have identical characteristics, a basic assumption of the logistic regression model.

  b) Each member of the cohort belongs to one of three types of termination:
    1) Subjects still alive on the analysis date.
    2) Subjects who died on a known date within the study period.
    3) Subjects who are lost to follow-up after a certain date.
  - That is, for many study subjects, the observation may be terminated before the occurrence of the main event under investigation: for example, subjects in types 1 and 3.

Objectives of Survival Analysis

- Estimate time-to-event for a group of individuals, such as time until second heart-attack for a group of myocardial infarction (MI) patients.
- To compare time-to-event between two or more groups, such as treated vs. **placebo** MI patients in a randomized controlled trial.
  - **Placebo** is known as a fake treatment (e.g. smoking an electronic cigarette with no nicotine or drinking fake bill to grow head hair) that is known to have no medical effect.
- 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**$= \dfrac{\mathbf{1}}{\textbf{incidence rate}}$

Survival Time



- In prospective studies, the important feature is not only the outcome event, such as death, but the time to that event, the *survival time*. To determine the survival time T, three basic elements are needed:
  - A time origin or starting point.
  - An ending event of interest.
  - A measurement scale for the passage of time.
- These may be, for example, the life span T from birth (starting point) to death (ending event) in years (measurement scale)
- Time-to-event: The time from entry into a study until a subject has a particular outcome.
- It is important to note that for some subjects in the study a complete survival time may not be available due to censoring.
- Censoring: Subjects are said to be censored if they are lost to follow up or drop out of the study, or if the study ends before they die or have an outcome of interest. They are counted as alive or disease-free for the time they were enrolled in the study.

Censoring

- Some subjects may not be observed for the full time to failure or event.
- Random censoring arises in medical applications with animal studies, epidemiological applications with human studies, or in clinical trials.
- Observation is terminated before the occurrence of the event.

Censoring may occur in one of the following forms:

- Loss to follow-up (the patient may decide to move elsewhere).
- Dropout (a therapy may have such bad effects that it is necessary to discontinue treatment).
- Termination of the study.
- Death due to a cause not under investigation (e.g., suicide).

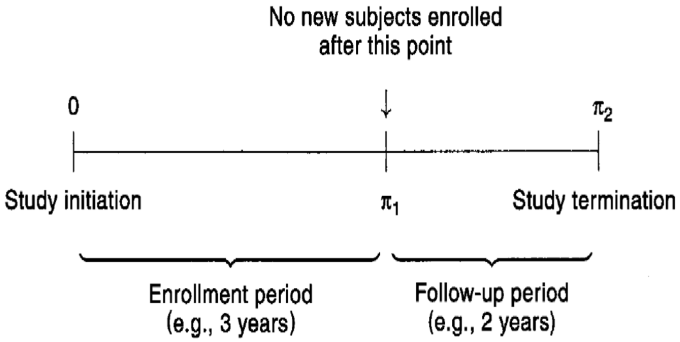

No new subjects enrolled after this point
0
π_2
Study initiation
π_1
Study termination
Enrollment period (e.g., 3 years)
Follow-up period (e.g., 2 years)

Figure 11.3 Clinical trial.

Right, Left, and Interval Censored Data

Right-censored data

- Suppose you're conducting a study on tumor duration.
- You're ready to complete the study and run your analysis, but some patients in the study are still have the tumor, so you don't know exactly how long their tumor will last.
- These observations would be right-censored.
- The "failure," or recover from tumor in this case, will occur after the recorded time.

Left-censored data

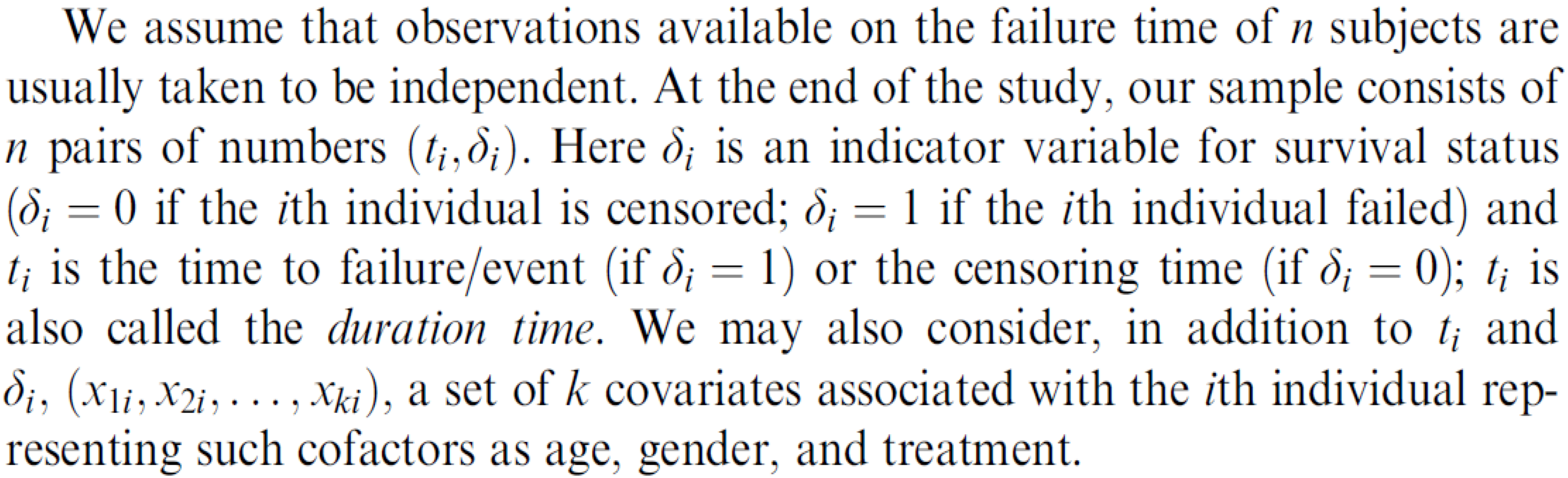
- Suppose you survey some patients in your study at the 300-day, but they already recover from tumor.
- You know they had recover before 200 days, but don't know exactly when.
- These are therefore left-censored observations.
- The "failure" occurred before a particular time.

Interval-censored data

- If we don't know exactly when some tumors were disappeared but we know it was within some interval of time.
- These observations would be interval-censored.
- We know the "failure" occurred within some given time period.

Assumptions with Censoring Data

- **The assumptions with censoring data**
- A person who is censored at t should be representative of all those subjects with the same values of explanatory variables who survive to t.
- Survival condition and reason of loss are independent.



We assume that observations available on the failure time of n subjects are usually taken to be independent. At the end of the study, our sample consists of n pairs of numbers (t_i, δ_i). Here δ_i is an indicator variable for survival status ($\delta_i = 0$ if the ith individual is censored; $\delta_i = 1$ if the ith individual failed) and t_i is the time to failure/event (if $\delta_i = 1$) or the censoring time (if $\delta_i = 0$); t_i is also called the *duration time*. We may also consider, in addition to t_i and δ_i, $(x_{1i}, x_{2i}, \ldots, x_{ki})$, a set of k covariates associated with the ith individual representing such cofactors as age, gender, and treatment.

Survival Function and Hazard (Risk) Function

- **Distribution of the survival time T from enrollment or starting point to the event of interest is characterized by either one of two equivalent functions:**

1. **The survival function**, denoted $S(t)$, also known as the survival rate, is defined as the probability that a person survives longer than t units of time:

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

for example, if times are in years, $S(2)$ is the two-year survival rate, $S(5)$ is the five-year survival rate, and so on.

2. **The hazard or risk function** $\lambda(t)$ gives the instantaneous failure rate which approximates the proportion of subjects dying or having events per unit time around time t.

A graph of $S(t)$ versus t is called a survival curve (Figure 11.2).

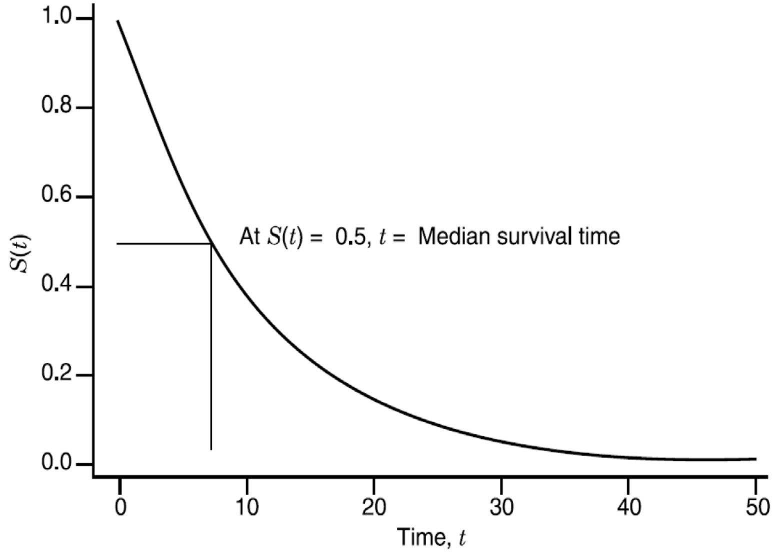


Figure 11.2 General form of a survival curve.

Hazards Model and Relative Risk

- $\lambda(t)$ is also known as the force of mortality and is a measure of the proneness to failure as a function of the person's age.
- When a population is subdivided into two subpopulations, E(exposed) and E'(nonexposed), by the *presence or absence of a certain characteristic* (an exposure such as smoking), each subpopulation corresponds to a hazard or risk function, and the ratio of two such functions,

$$\mathrm{RR}(t) = \frac{\lambda(t; E)}{\lambda(t; E')}$$

is called the *relative risk* associated with exposure to factor E; that is risk of the espoused subjects *relative* to the risk of nonexposed subjects.

Proportional Hazards Model

Proportional Hazards Model (PHM):

- When the magnitude of an effect is constant $RR(t) = \rho$, then PHM is

$$\lambda(t; E) = \rho\lambda(t; E')$$

- Another way to express this model is

$$\lambda(t) = \lambda_0(t)e^{\beta x}$$

where $\lambda_0(t) = \lambda(t; E')$ and the indicator (or *covariate*) x is defined as

$$x = \begin{cases} 0 & \text{if unexposed} \\ 1 & \text{if exposed} \end{cases}$$

$\lambda(t; E')$ is the hazard function of the unexposed subpopulation
The regression (*Cox's regression*) coefficient $\beta_0$represents the relative risk on the log scale.

Kaplan-Meier Curve

- Let $t_1 < t_2 < \cdots < t_k$ be the distinct observed death times in a sample of size n from a homogeneous population with survival function $S(t)$ to be estimated
- $k \leq n$; k could be less than n because some subjects may be censored and some subjects may have events at the same time.
- Let n_i be the number of subjects at risk at a time just prior to t_i $(1 \leq i \leq k$; these are cases whose duration time is at least $t_i)$.
- Let d_i the number of deaths at t_i.
- The survival function $S(t)$ is estimated by

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

which is called the *product-limit estimator* or *Kaplan–Meier estimator.*

Kaplan-Meier Curve

- **The explanation as follows:**
  - $d_i\ /n_i$ is the proportion (or estimated probability of having an event in the interval from t_{i-1} to t_i,
  - $1 - d_i\ /n_i$ represents the proportion (or estimated probability of surviving that same interval), and the product in the formula for $\hat{S}(t)$ follows from the product rule for probabilities.

- The 95% confidence of $S(t)$ is given by

$$\hat{S}(t) \exp[\pm 1.96 \hat{s}(t)]$$

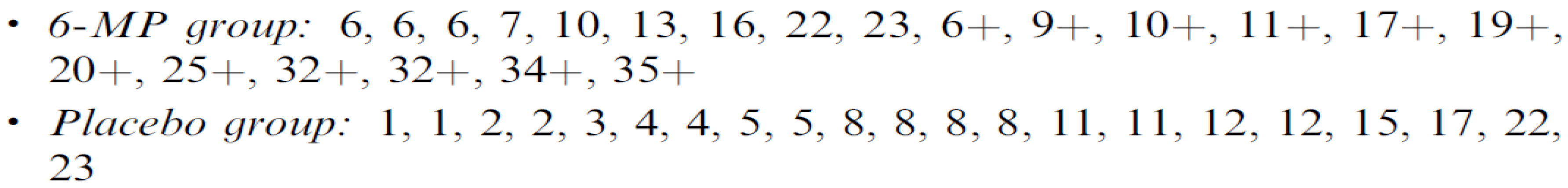
where

$$\hat{s}^2(t) = \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}$$

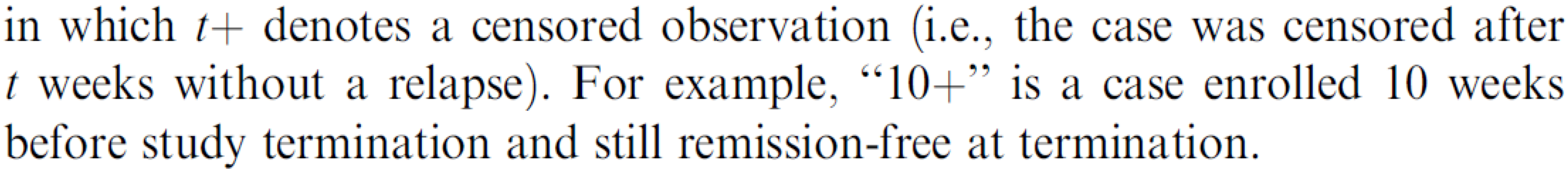
- **Limitations of Kaplan-Meier:**
  - Mainly descriptive.
  - Doesn’t control for covariates.
  - Requires categorical predictors.
  - Can’t accommodate time-dependent variables.

Example 11.1

- The remission times of 42 patients with acute leukemia were reported from a clinical trial undertaken to assess the ability of the drug 6-mercaptopurine (6-MP) to maintain remission (See Figure 11.4).
- Each patient was randomized to receive either 6-MP or placebo.
- The study was terminated after one year; patients have different follow-up times because they were enrolled sequentially at different times.
- Times in weeks were:



  - *6-MP group:* 6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+
  - *Placebo group:* 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23



in which *t*+ denotes a censored observation (i.e., the case was censored after *t* weeks without a relapse). For example, "10+" is a case enrolled 10 weeks before study termination and still remission-free at termination.

Example 11.1 (Cont.)

- According to the product-limit method, survival rates for the 6-MP group are calculated by constructing a table such as Table 11.1 with five columns;

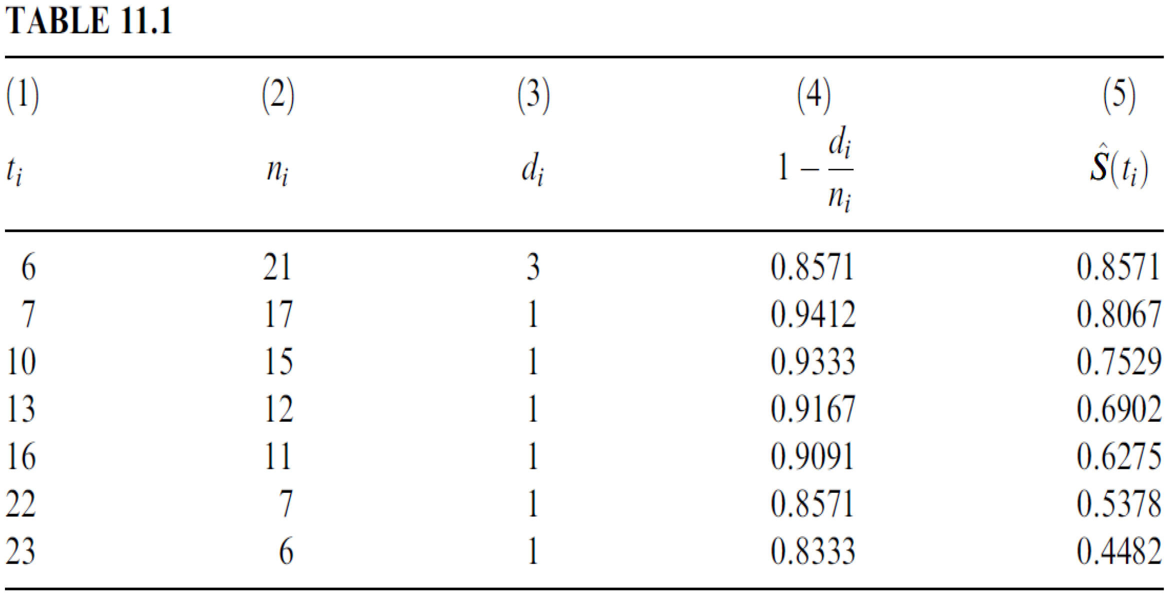


TABLE 11.1

(1) t_i	(2) n_i	(3) d_i	(4) $1 - \frac{d_i}{n_i}$	(5) $\hat{S}(t_i)$
6	21	3	0.8571	0.8571
7	17	1	0.9412	0.8067
10	15	1	0.9333	0.7529
13	12	1	0.9167	0.6902
16	11	1	0.9091	0.6275
22	7	1	0.8571	0.5378
23	6	1	0.8333	0.4482

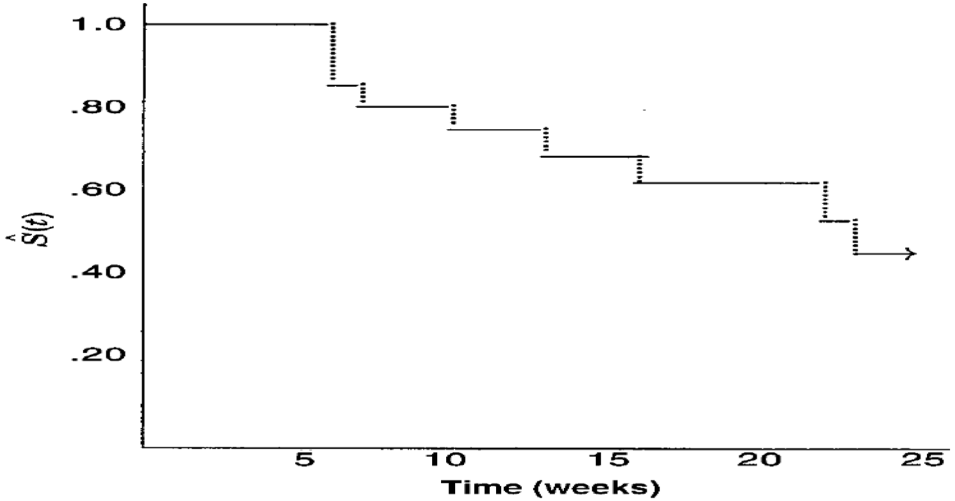


Figure 11.4 Survival curve: drug 6-MP group.

From Table 11.1, we have, for example:

7-week survival rate is 80:67%

22-week survival rate is 53:78%

and a 95% confidence interval for $S(7)$ is (0.6531, 0.9964).

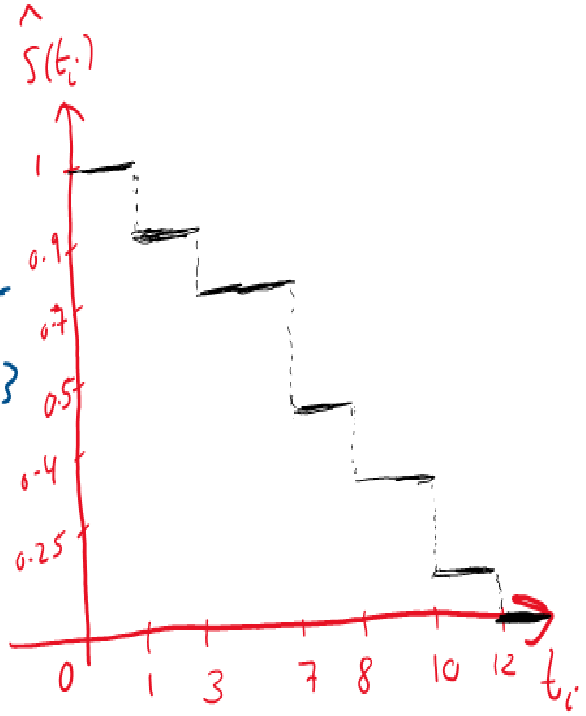
Example

- The survival time (in years) from first heart attack till death is measured for 10 patients and are reported below

Survival time: 1, 3, 3, 6+, 7, 7, 8, 10, 11+, 12

- Calculate the Kaplan-Meier estimator (product-limit estimator) and find C.I. for $S(3)$.

t_i	n_i	d_i	$1-d_i/n_i$	$\hat{S}(t_i)$
1	10	1	0.9	0.9
3	9	2	0.778	$0.9\times0.778=0.70$
7	6	2	0.667	$0.70\times0.667=0.467$
8	4	1	0.75	$0.467\times0.75=0.35$
10	3	1	0.667	$0.35\times0.667=0.233$
12	1	1	0	0

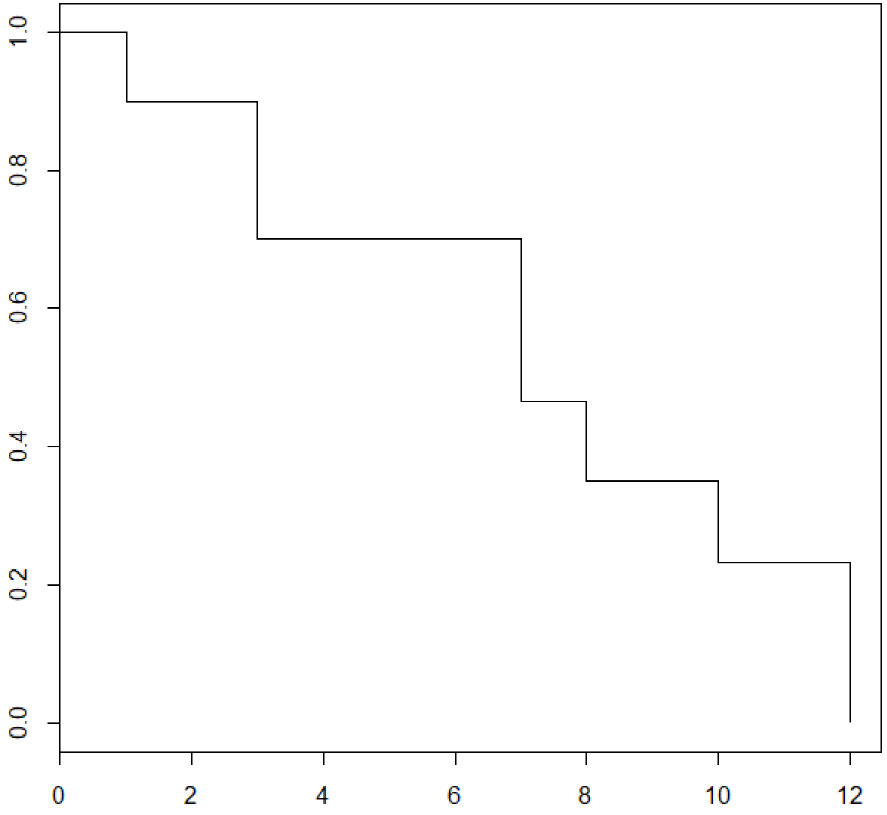


R-code of Previous Example

```
library(survival)
time <- c(1,3,3,6,7,7,8,10,11,12)
status <- c(1,1,1,0,1,1,1,1,0,1)
fit <- survfit(Surv(time, status)~1)
plot(fit,conf.int="none")
summary(fit)
```

```
Call: survfit(formula = Surv(time, cens) ~ 1)

time n.risk n.event survival std.err lower 95% CI upper 95% CI
   1     10       1    0.900  0.0949       0.7320        1.000
   3      9       2    0.700  0.1449       0.4665        1.000
   7      6       2    0.467  0.1658       0.2326        0.936
   8      4       1    0.350  0.1602       0.1427        0.858
  10      3       1    0.233  0.1431       0.0701        0.776
  12      1       1    0.000     NaN           NA           NA
```



Comparison of Survival Distribution

Null hypothesis: There are no differences between survival curves

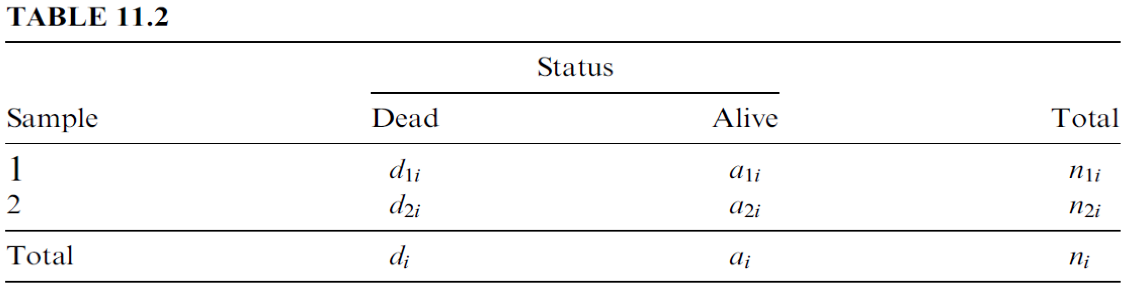


TABLE 11.2

	Status		
Sample	Dead	Alive	Total
1	d_{1i}	a_{1i}	n_{1i}
2	d_{2i}	a_{2i}	n_{2i}
Total	d_i	a_i	n_i

After constructing a 2 × 2 table for each uncensored observation, the evidence against the null hypothesis The evidence against the null hypothesis is summarized in the standardized statistic

$$z = \frac{\theta}{[\mathrm{Var}_0(\theta)]^{1/2}}$$

which is referred to the standard normal percentile $z_{1-\alpha}$ for a specified size α of the test. We may also refer z^2 to a chi-square distribution at 1 degree of freedom.

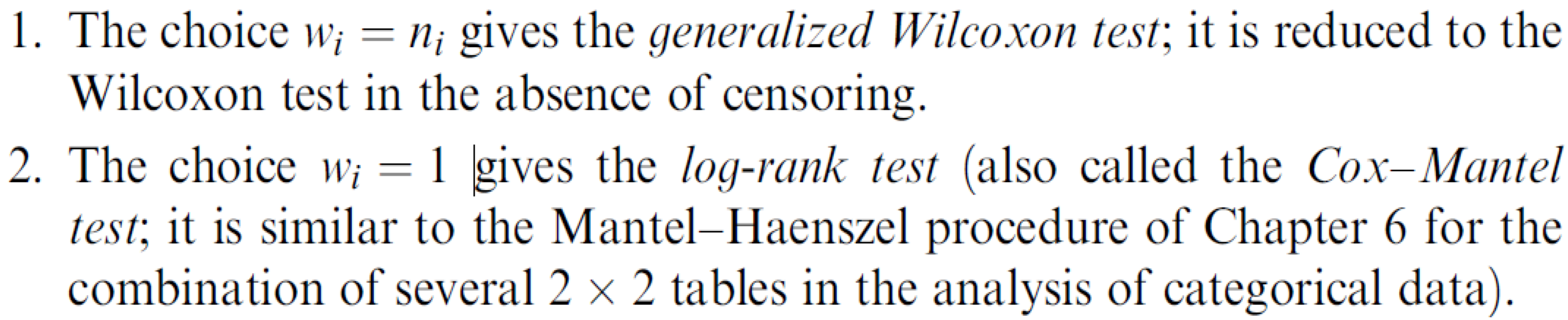
$$\theta = \sum_{i=1}^{m} w_i [d_{1i} - E_0(d_{1i})]$$

$$\mathrm{Var}_0(\theta) = \sum_{i=1}^{m} w_i^2 \ \mathrm{Var}_0(d_{1i}) = \sum_{i=1}^{m} \frac{w_i^2 n_{1i} n_{2i} a_i d_i}{n_i^2 (n_i - 1)}$$

where w_i is the weight associated with the 2 × 2 table at t_i and $E_0(d_{1i}) = \frac{n_{1i} d_i}{n_i}$

Comparison of Survival Distribution (Cont.)

- **There are two important special cases:**



1. The choice $w_i = n_i$ gives the *generalized Wilcoxon test*; it is reduced to the Wilcoxon test in the absence of censoring.
2. The choice $w_i = 1$ gives the *log-rank test* (also called the *Cox–Mantel test*; it is similar to the Mantel–Haenszel procedure of Chapter 6 for the combination of several 2 × 2 tables in the analysis of categorical data).

- **Which test should we use?**
  - The generalized Wilcoxon statistic : puts more weight on the beginning observations, and because of that its use is more powerful in detecting the effects of short-term risks.
  - The log-rank statistic : puts equal weight on each observation. Hence, by default, is more sensitive to exposures with a constant relative risk (the proportional hazards effect; in fact, we have derived the log-rank test as a score test using the proportional hazards model).

Examples 11.2

- Refer back to the clinical trial (Example 11.1) to evaluate the effect of 6-mercaptopurine (6-MP) to maintain remission from acute leukemia.
- The results of the tests indicate a highly significant difference between survival patterns of the two groups (Figure 11.5).

$$\text{Generalized Wilcoxon}: \quad \chi^2 = 13.46(1\text{ df}); p < 0.0002$$

$$\text{Log-rank}: \quad \chi^2 = 16.79(1\text{ df}); p = 0.0001$$

- The generalized Wilcoxon test shows a slightly larger statistic, indicating that the difference is slightly larger at earlier times;
- However, the log-rank test is almost equally significant, indicating that the use of 6-MP has a long-term effect (i.e., the effect does not wear off).

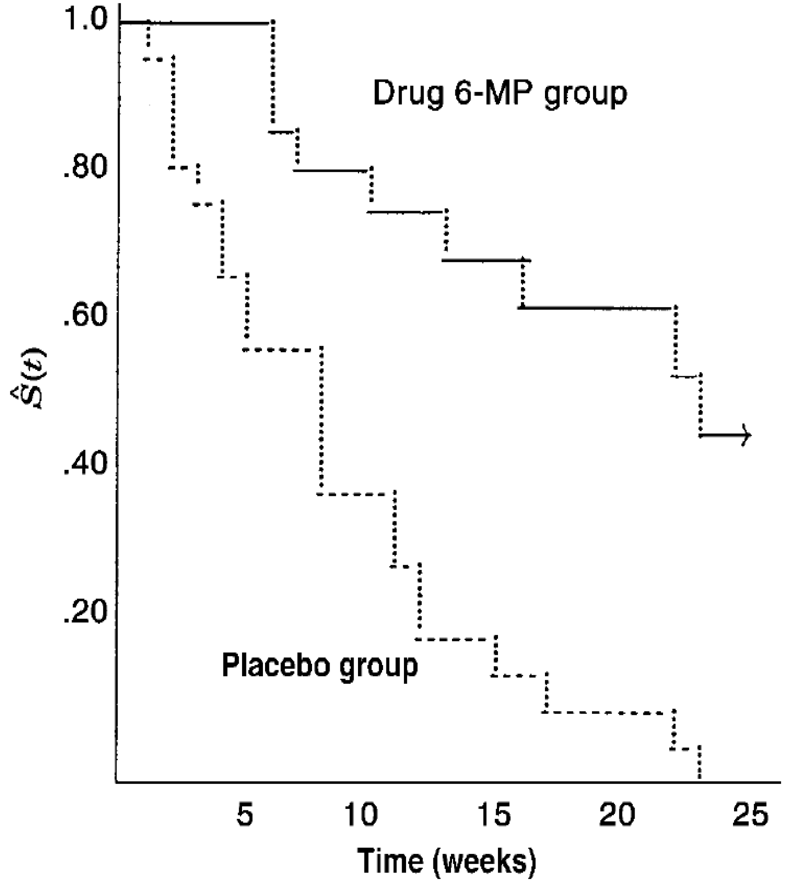


Figure 11.5 Two survival curves: drug 6-MP group and placebo group.

Simple Regression Analysis: Model and Approach

- **Simple regression model** is used when only one predictor or independent variable is available for predicting the survival of interest.
- **Cox's regression model** or proportional hazards model (PHM) expresses a relationship between X and the hazard function of T as follows:

$$\lambda(t \mid X = x) = \lambda_0(t)e^{\beta x}$$

In this model, $\lambda_0(t)$ is an unspecified baseline hazard (i.e., hazard at $X = 0$) and β is an unknown regression coefficient.

Measures of Association

- For the case of a binary covariate with the conventional coding

$$X_i = \begin{cases} 0 & \text{if the patient is not exposed} \\ 1 & \text{if the patient is exposed} \end{cases}$$

- It can be seen that from the proportional hazards model, the **relative risk (RR)** ratio

$$RR = e^{\beta} = \frac{\lambda(t; \text{exposed})}{\lambda(t; \text{unexposed})}$$

- The term exposed may refer to a risk factor such as smoking, or a patient’s characteristic such as race (white/non-white) or gender (male/female).

Measures of Association (Cont.)

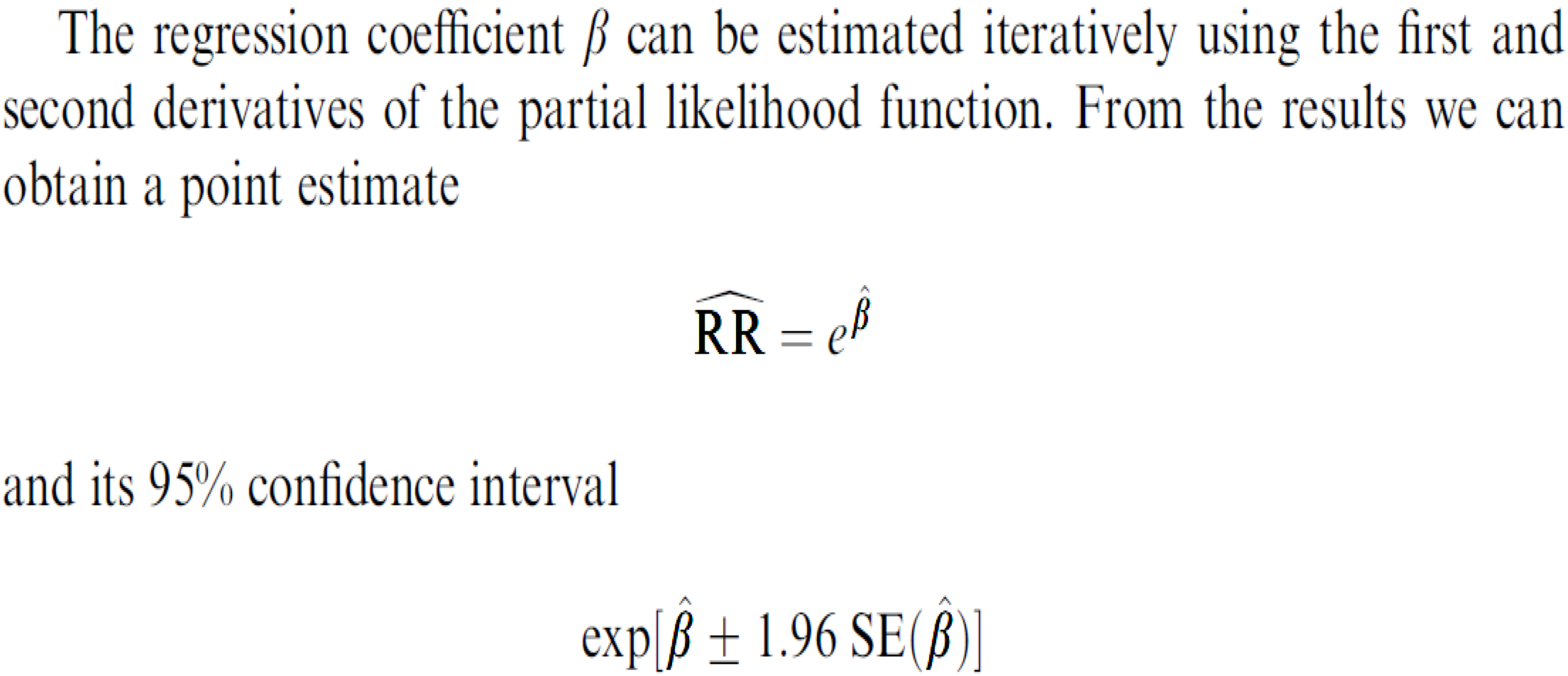
- For the case of a continuous covariate and any value x of X, the **relative risk (RR)** ratio due to a 1-unit increase in the value of $X = x + 1$ versus $X = x$.

$$RR = e^{\beta} = \frac{\lambda(t; X = x + 1)}{\lambda(t; X = x)}$$

- For an m- unit increase in the value of X, say $X = x + m$ versus $X = x$, the corresponding **relative risk** is

$$e^{m\beta}$$

Measures of Association (Cont.)



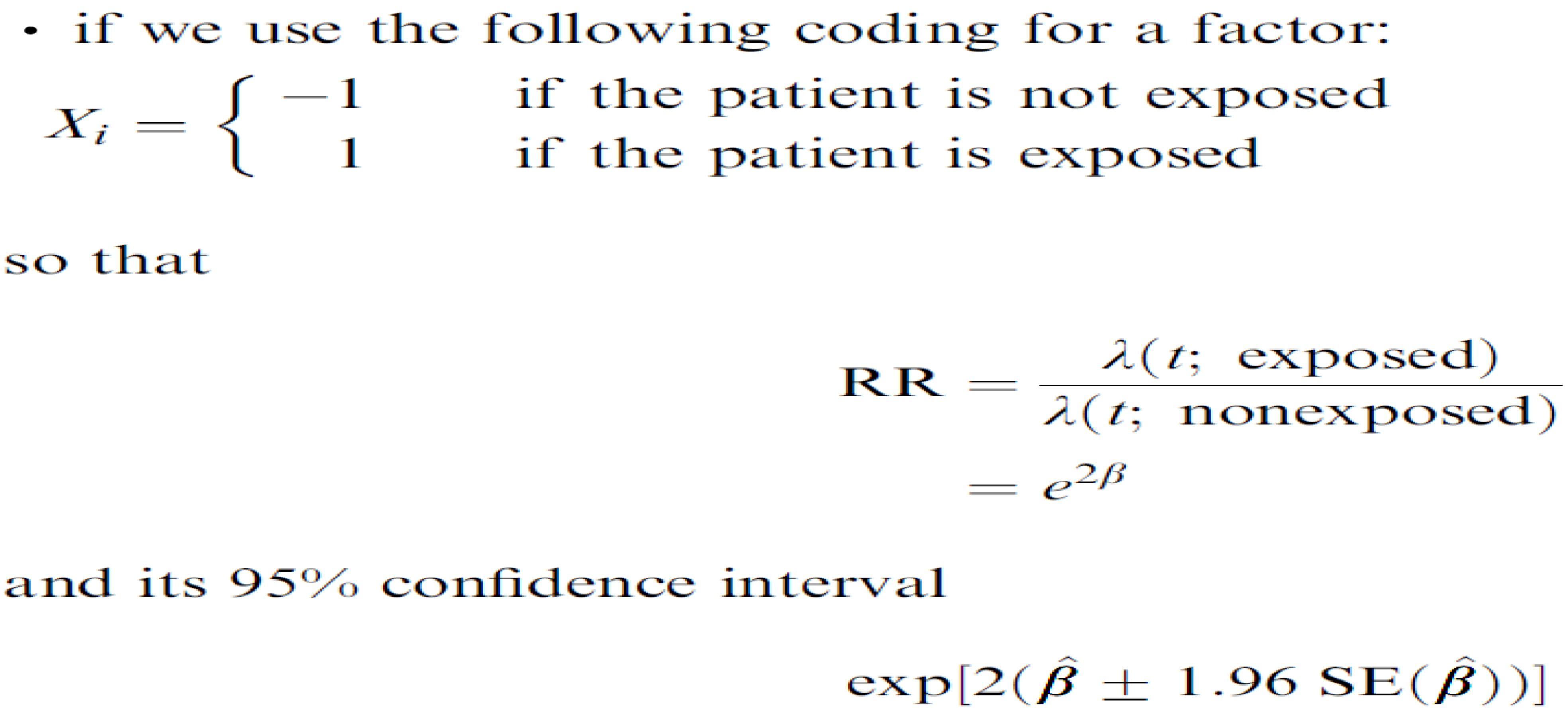
The regression coefficient β can be estimated iteratively using the first and second derivatives of the partial likelihood function. From the results we can obtain a point estimate

$$\widehat{\mathrm{RR}} = e^{\hat{\beta}}$$

and its 95% confidence interval

$$\exp[\hat{\beta} \pm 1.96\,\mathrm{SE}(\hat{\beta})]$$

Measures of Association (Cont.)



- if we use the following coding for a factor:

$$X_i = \begin{cases} -1 & \text{if the patient is not exposed} \\ 1 & \text{if the patient is exposed} \end{cases}$$

so that

$$\begin{aligned} \text{RR} &= \frac{\lambda(t;\ \text{exposed})}{\lambda(t;\ \text{nonexposed})} \\ &= e^{2\beta} \end{aligned}$$

and its 95% confidence interval

$$\exp[2(\hat{\beta} \pm 1.96\ \text{SE}(\hat{\beta}))]$$

Examples 11.3 & 11.4

- A group of patients who died of acute myelogenous leukemia were classified into two subgroups according to the presence or absence of a morphologic characteristic of white cells (Table 11.3).
- Patients termed AG positive were identified by the presence of Auer rods and/or significant granulature of leukemic cells in the bone marrow at diagnosis.
- These factors were absent for AG-negative patients.
- Investigate the relationship between survival time of AG-positive patients and white blood count (WBC) in two different ways using either
  a) X=WBC or
  b) X =log(WBC).

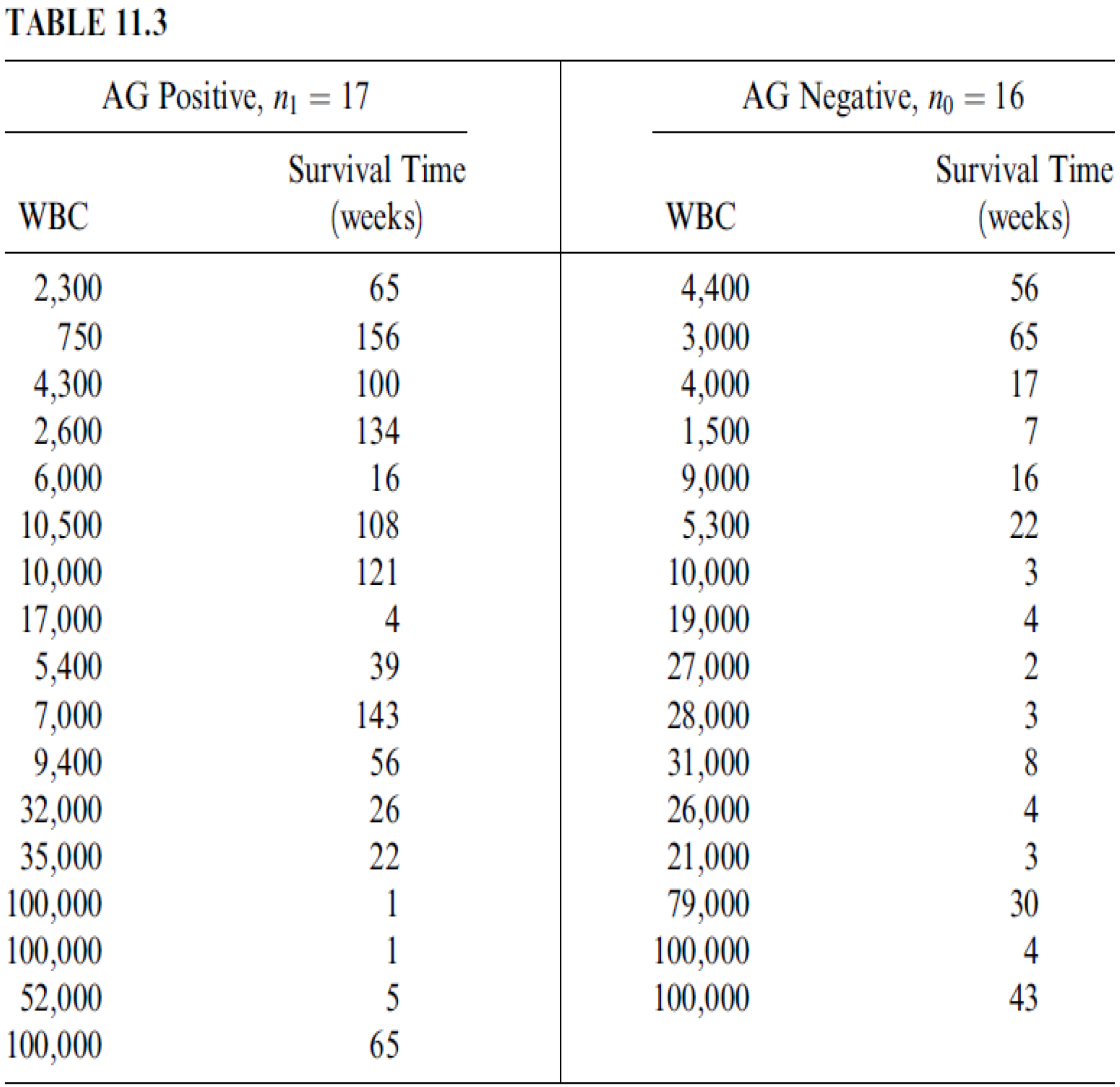
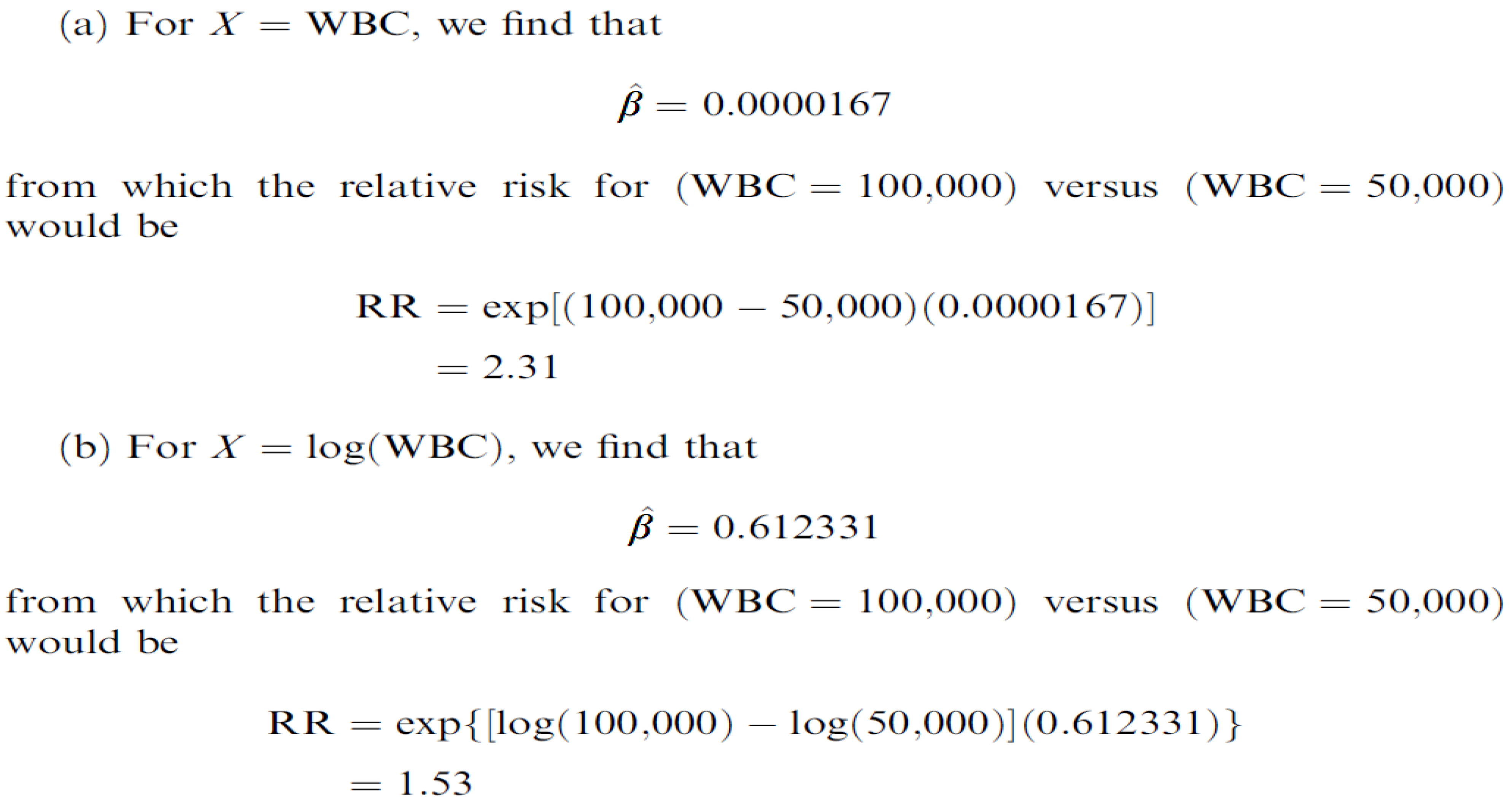


TABLE 11.3

AG Positive, $n_1 = 17$		AG Negative, $n_0 = 16$	
WBC	Survival Time (weeks)	WBC	Survival Time (weeks)
2,300	65	4,400	56
750	156	3,000	65
4,300	100	4,000	17
2,600	134	1,500	7
6,000	16	9,000	16
10,500	108	5,300	22
10,000	121	10,000	3
17,000	4	19,000	4
5,400	39	27,000	2
7,000	143	28,000	3
9,400	56	31,000	8
32,000	26	26,000	4
35,000	22	21,000	3
100,000	1	79,000	30
100,000	1	100,000	4
52,000	5	100,000	43
100,000	65		

Examples 11.3 & 11.4 (Cont.)



(a) For X = WBC, we find that

$$\hat{\beta} = 0.0000167$$

from which the relative risk for (WBC = 100,000) versus (WBC = 50,000) would be

$$\begin{aligned} \text{RR} &= \exp[(100{,}000 - 50{,}000)(0.0000167)] \\ &= 2.31 \end{aligned}$$

(b) For $X = \log(\text{WBC})$, we find that

$$\hat{\beta} = 0.612331$$

from which the relative risk for (WBC = 100,000) versus (WBC = 50,000) would be

$$\begin{aligned} \text{RR} &= \exp\{[\log(100{,}000) - \log(50{,}000)](0.612331)\} \\ &= 1.53 \end{aligned}$$

Pair-Matched Case-Control Studies

- The simplest example of pair-matched data occurs with a single binary exposure (e.g., smoking versus nonsmoking).
- The data for outcomes can be represented by a 2 × 2 table (Table 11.7) where $(+, -)$ denotes (exposed, unexposed).
- The most suitable statistical model for making inferences about the odds ratio θ is to use the conditional probability of the number of exposed cases among the discordant pairs.

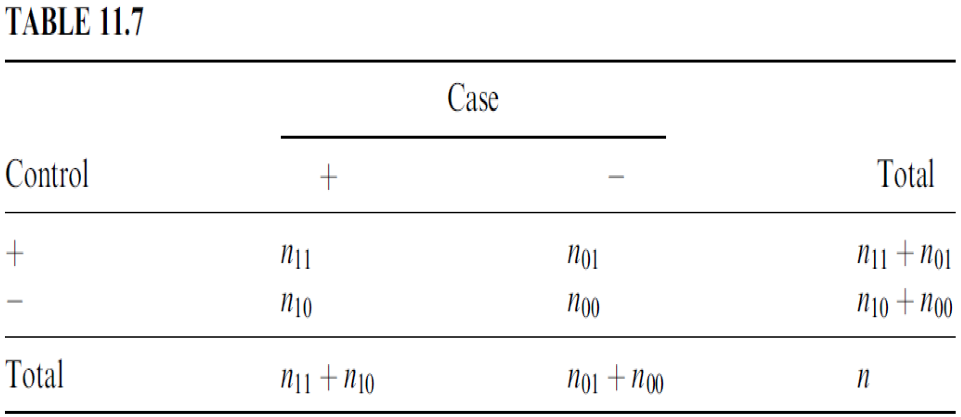


TABLE 11.7

	Case		
Control	+	−	Total
+	n_{11}	n_{01}	$n_{11}+n_{01}$
−	n_{10}	n_{00}	$n_{10}+n_{00}$
Total	$n_{11}+n_{10}$	$n_{01}+n_{00}$	n

- Given $(n_{10}+n_{01})$as fixed, n_{10} has the binomial distribution $\mathrm{B}(n_{10}+n_{01},\pi)$, that is, the binomial distribution with n_{10} $n = n_{10}+n_{01}$trials, each with probability of success

$$\pi = \frac{\mathrm{OR}}{1+\mathrm{OR}}$$

- Using the binomial model above, one can estimate the odds ratio, the results are

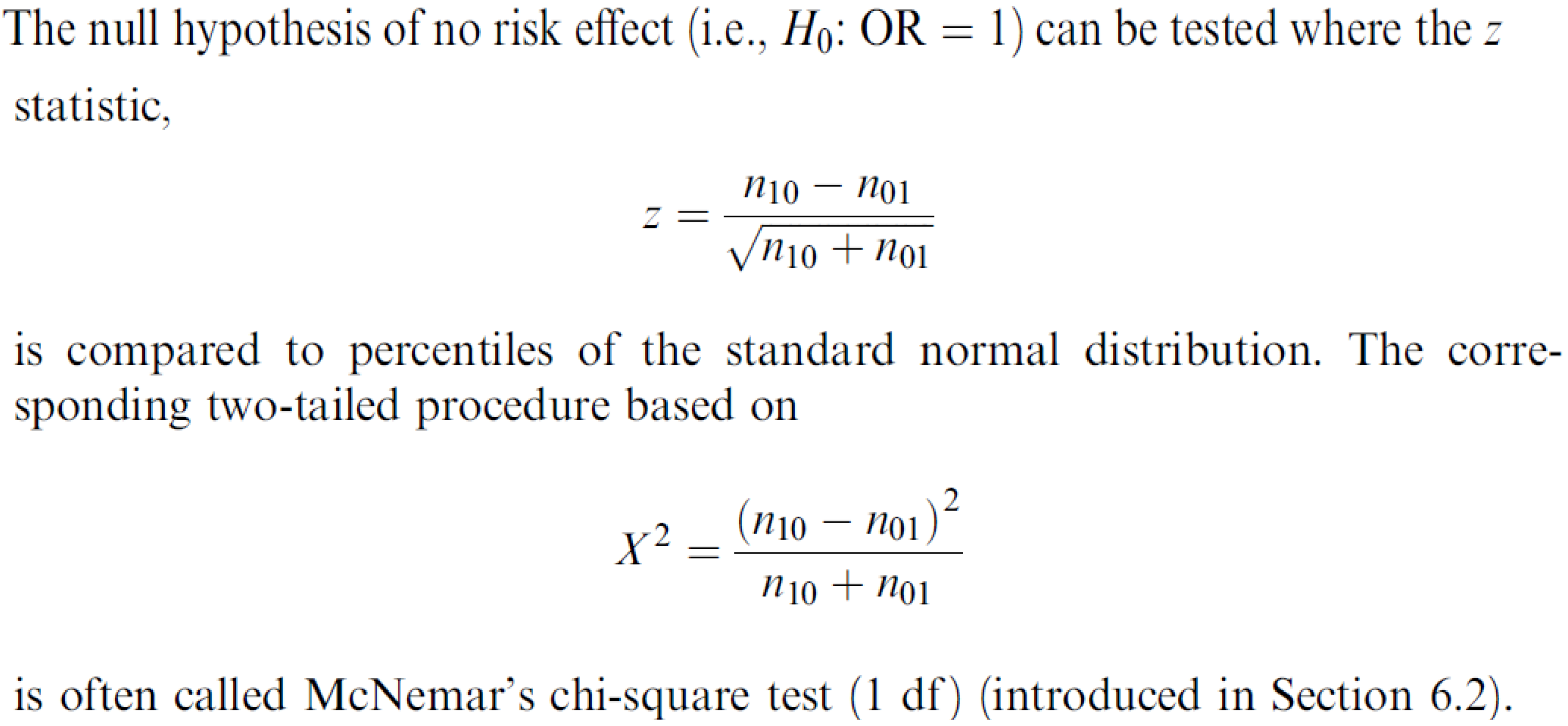
$$\widehat{\mathrm{OR}} = \frac{n_{10}}{n_{01}}$$

$$\widehat{\mathrm{Var}}(\widehat{\mathrm{OR}}) = \frac{n_{10}(n_{10}+n_{01})}{n_{01}^3}$$

For example, with large samples, a 95% confidence interval for the odds ratio is given by

$$\widehat{\mathrm{OR}} \pm (1.96)[\widehat{\mathrm{Var}}(\widehat{\mathrm{OR}})]^{1/2}$$

Pair-Matched Case-Control Studies (Cont.)



The null hypothesis of no risk effect (i.e., H_0: OR $= 1$) can be tested where the z statistic,

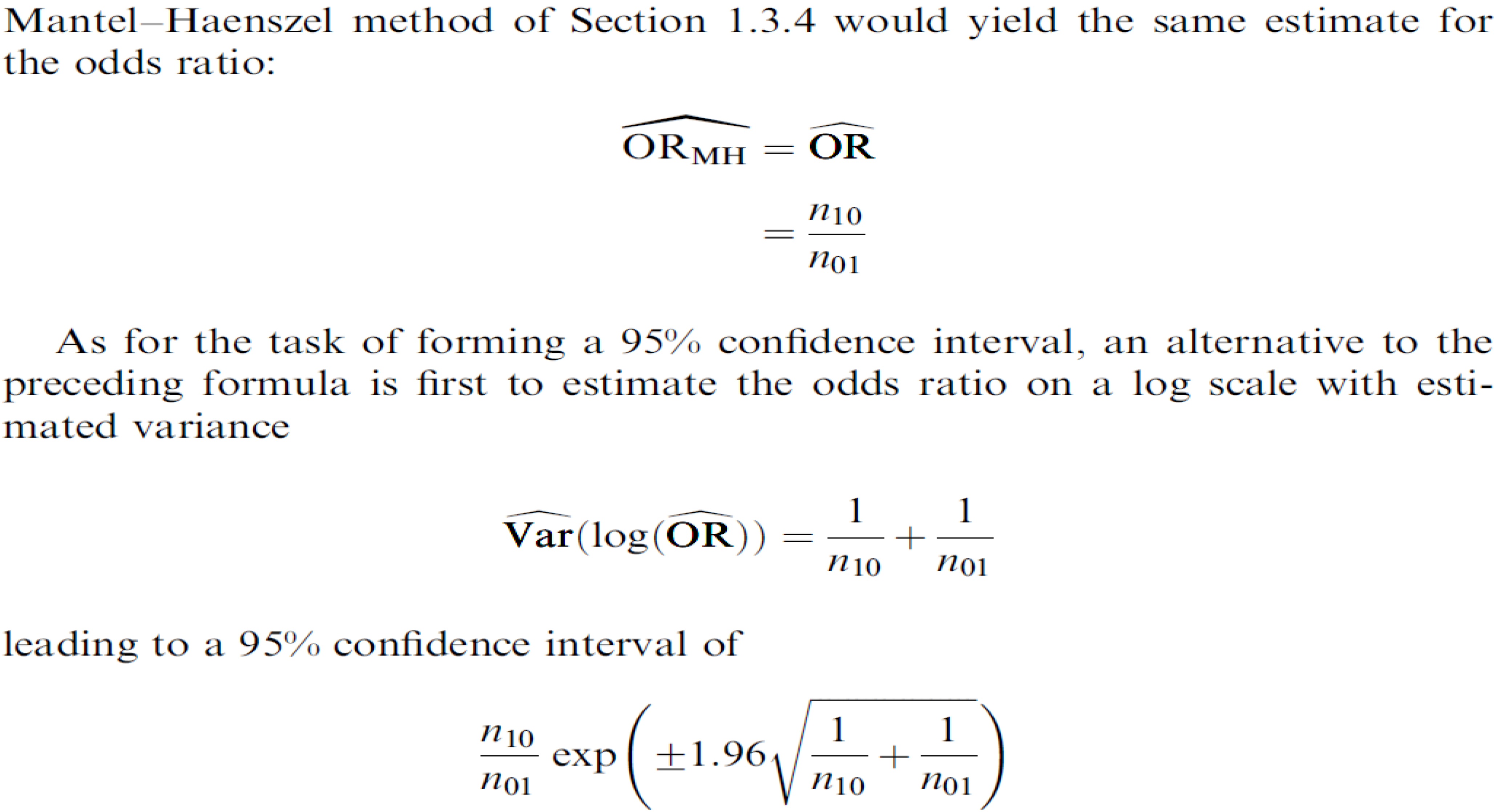
$$z = \frac{n_{10} - n_{01}}{\sqrt{n_{10} + n_{01}}}$$

is compared to percentiles of the standard normal distribution. The corresponding two-tailed procedure based on

$$X^2 = \frac{(n_{10} - n_{01})^2}{n_{10} + n_{01}}$$

is often called McNemar's chi-square test (1 df) (introduced in Section 6.2).

Pair-Matched Case-Control Studies (Cont.)



Mantel–Haenszel method of Section 1.3.4 would yield the same estimate for the odds ratio:

$$\widehat{\mathrm{OR}_{\mathrm{MH}}} = \widehat{\mathbf{OR}}$$

$$= \frac{n_{10}}{n_{01}}$$

As for the task of forming a 95% confidence interval, an alternative to the preceding formula is first to estimate the odds ratio on a log scale with estimated variance

$$\widehat{\mathbf{Var}}(\log(\widehat{\mathbf{OR}})) = \frac{1}{n_{10}} + \frac{1}{n_{01}}$$

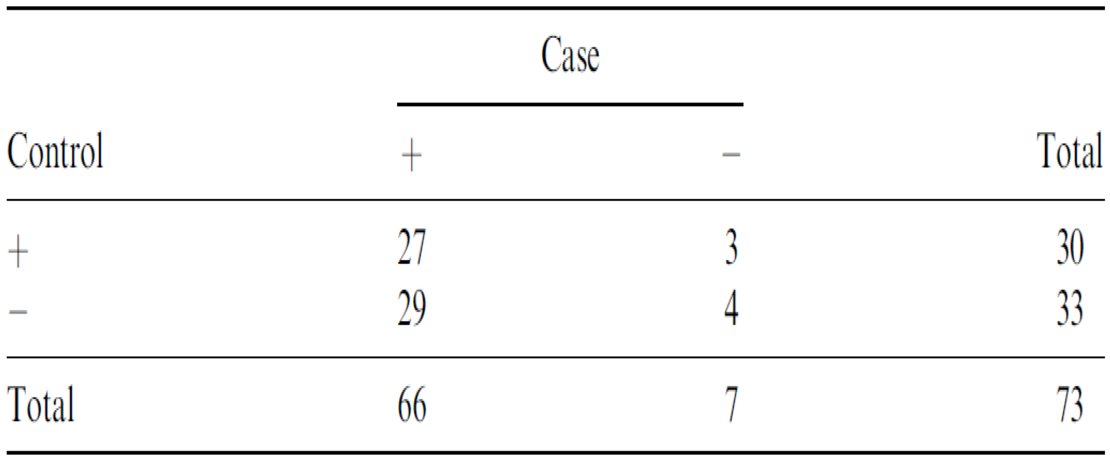
leading to a 95% confidence interval of

$$\frac{n_{10}}{n_{01}} \exp\left(\pm 1.96\sqrt{\frac{1}{n_{10}} + \frac{1}{n_{01}}}\right)$$

Example 11.7

- In a study of endometrial cancer in which the investigators identified 63 cases occurring in a retirement community near Los Angeles, California from 1971 to 1975, each disease person was matched with $R = 4$ controls who were alive and living in the community at the time the case was diagnosed, who were born within one year of the case, who had the same marital status, and who had entered the community at approximately the same time. The risk factor was previous use of estrogen (yes/no) and the data in Table 11.9 were obtained from the first-found matched control (the complete data set with four matched controls will be given later).
- **Compute the Odds Ratio and the 95% CI.**

TABLE 11.9



	Case		
Control	+	−	Total
+	27	3	30
−	29	4	33
Total	66	7	73

An application of the methods above yields

$$\widehat{\mathbf{OR}} = \frac{n_{10}}{n_{01}} = \frac{29}{3} = 9.67$$

and a 95% confidence interval for OR is $(2.95, 31.74)$.