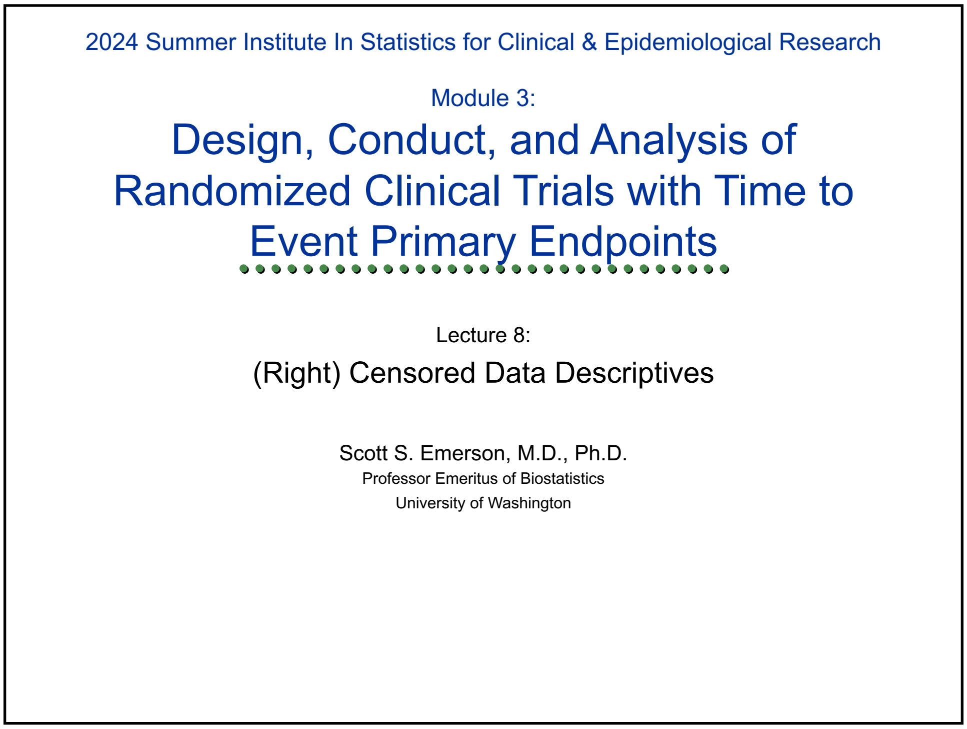


2024 Summer Institute In Statistics for Clinical & Epidemiological Research

Module 3:

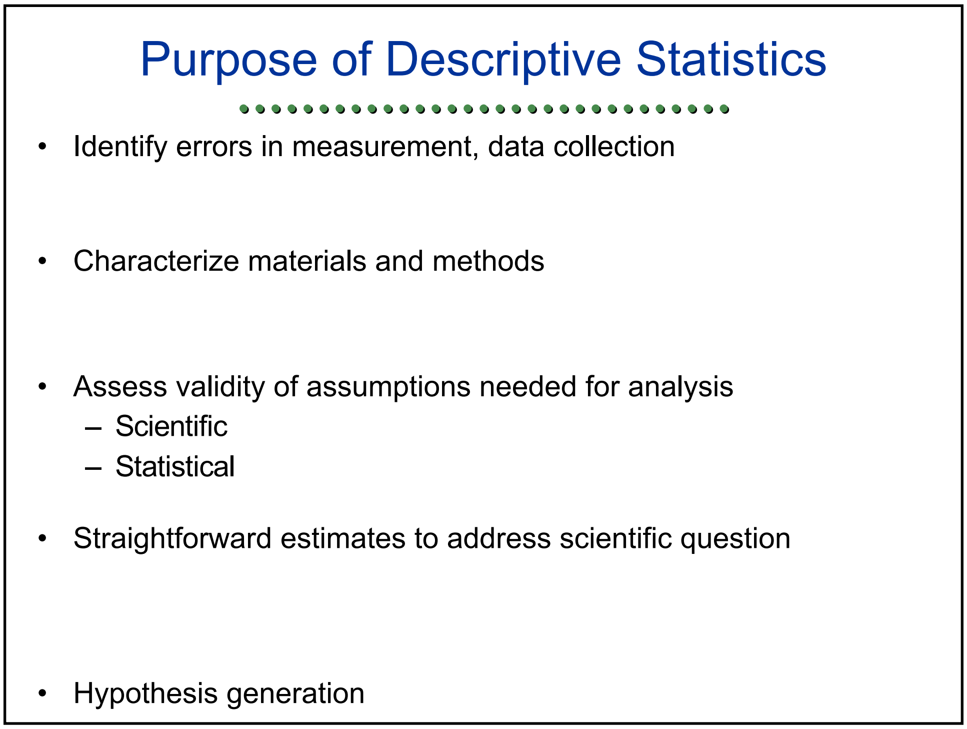
Design, Conduct, and Analysis of Randomized Clinical Trials with Time to Event Primary Endpoints

Lecture 8:

(Right) Censored Data Descriptives

Scott S. Emerson, M.D., Ph.D.
Professor Emeritus of Biostatistics
University of Washington

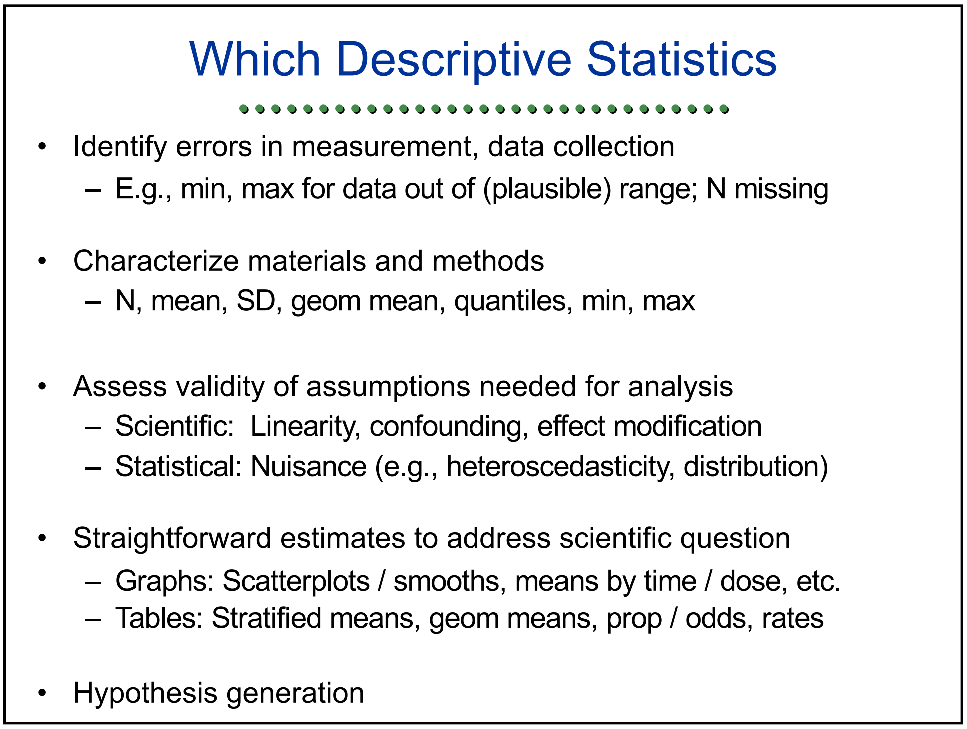
1



Purpose of Descriptive Statistics

- Identify errors in measurement, data collection
- Characterize materials and methods
- Assess validity of assumptions needed for analysis
  - Scientific
  - Statistical
- Straightforward estimates to address scientific question
- Hypothesis generation

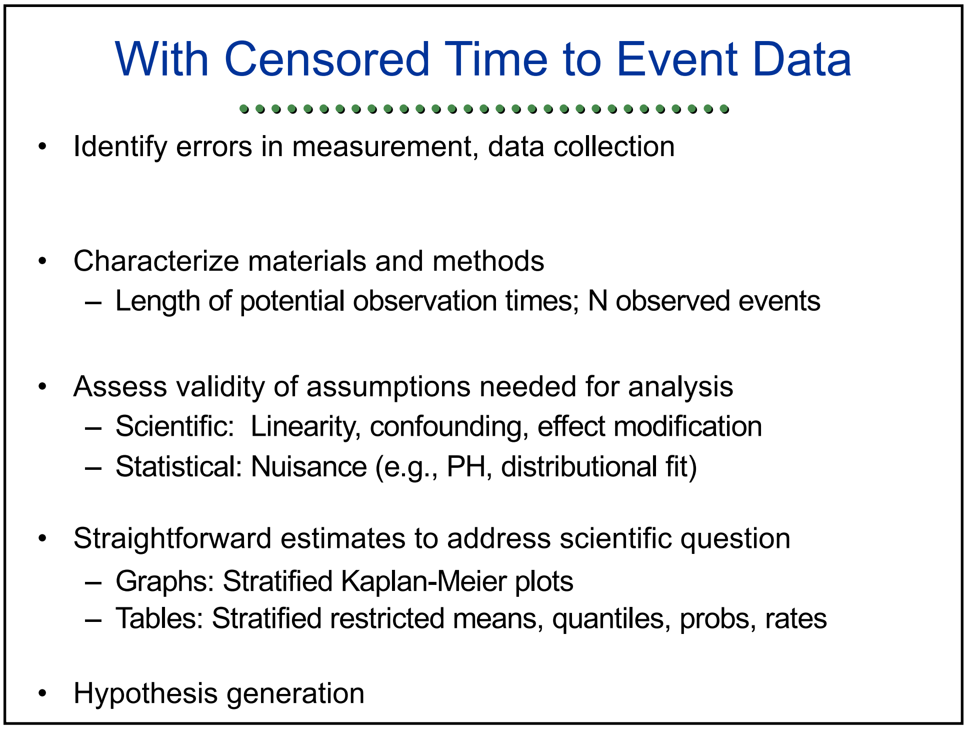
2



Which Descriptive Statistics

- Identify errors in measurement, data collection
  - E.g., min, max for data out of (plausible) range; N missing
- Characterize materials and methods
  - N, mean, SD, geom mean, quantiles, min, max
- Assess validity of assumptions needed for analysis
  - Scientific: Linearity, confounding, effect modification
  - Statistical: Nuisance (e.g., heteroscedasticity, distribution)
- Straightforward estimates to address scientific question
  - Graphs: Scatterplots / smooths, means by time / dose, etc.
  - Tables: Stratified means, geom means, prop / odds, rates
- Hypothesis generation

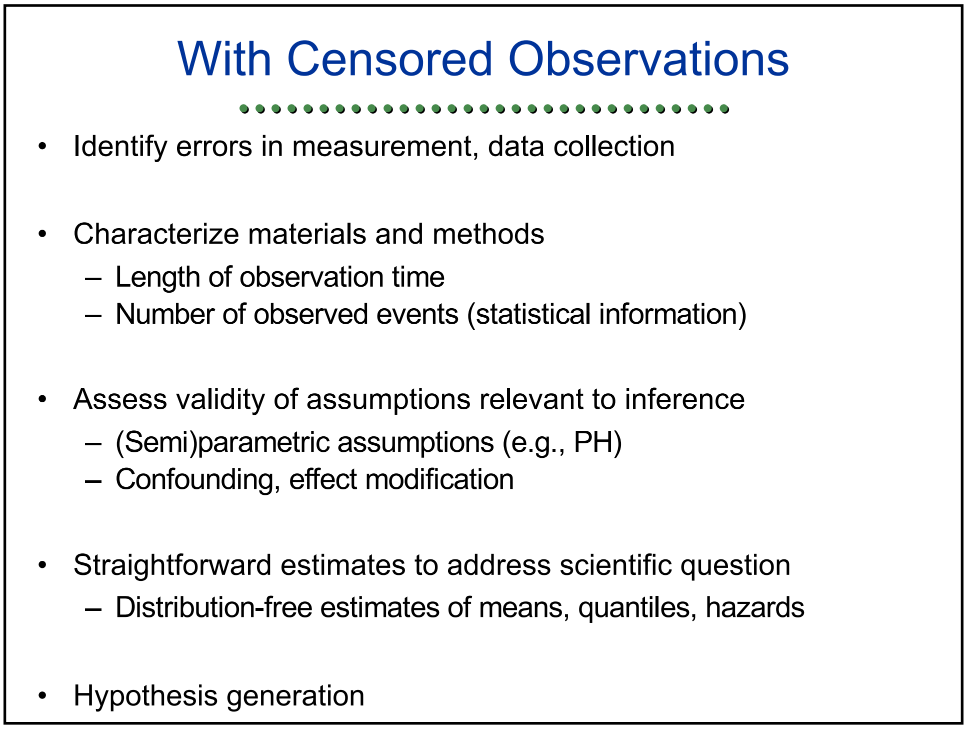
3



With Censored Time to Event Data

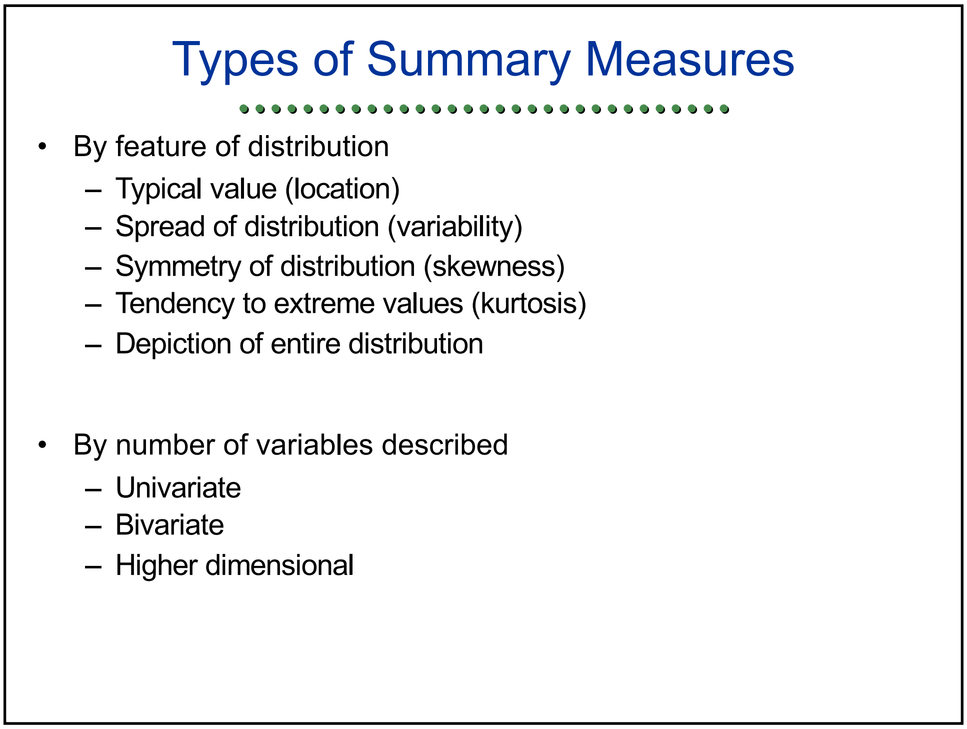
- Identify errors in measurement, data collection
- Characterize materials and methods
  - Length of potential observation times; N observed events
- Assess validity of assumptions needed for analysis
  - Scientific: Linearity, confounding, effect modification
  - Statistical: Nuisance (e.g., PH, distributional fit)
- Straightforward estimates to address scientific question
  - Graphs: Stratified Kaplan-Meier plots
  - Tables: Stratified restricted means, quantiles, probs, rates
- Hypothesis generation

4



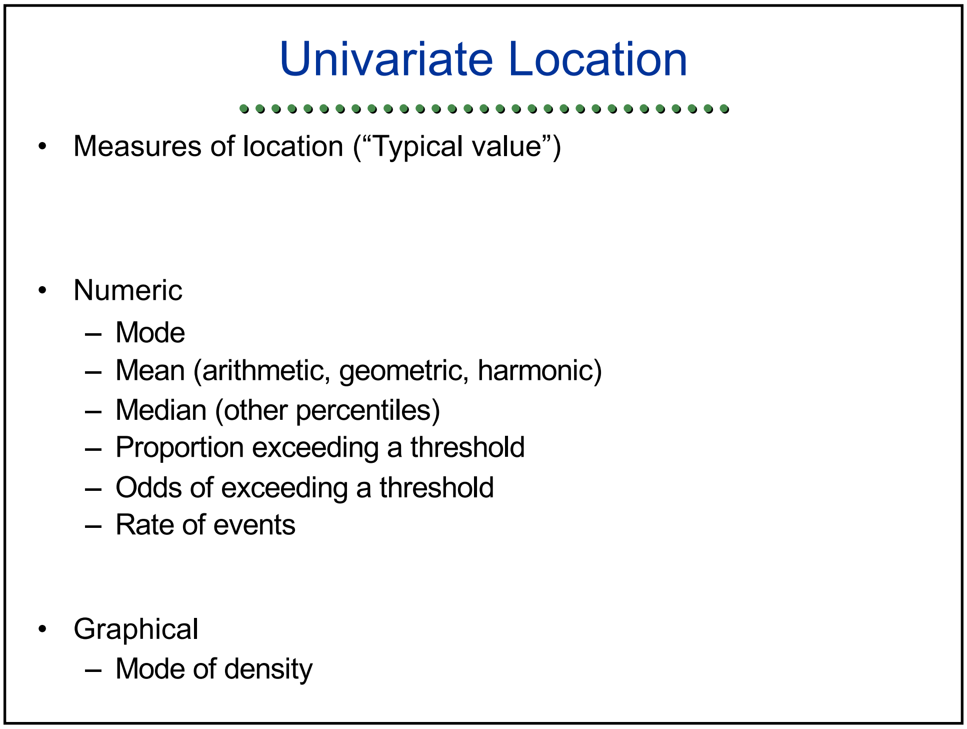
With Censored Observations

- Identify errors in measurement, data collection
- Characterize materials and methods
  - Length of observation time
  - Number of observed events (statistical information)
- Assess validity of assumptions relevant to inference
  - (Semi)parametric assumptions (e.g., PH)
  - Confounding, effect modification
- Straightforward estimates to address scientific question
  - Distribution-free estimates of means, quantiles, hazards
- Hypothesis generation



Types of Summary Measures

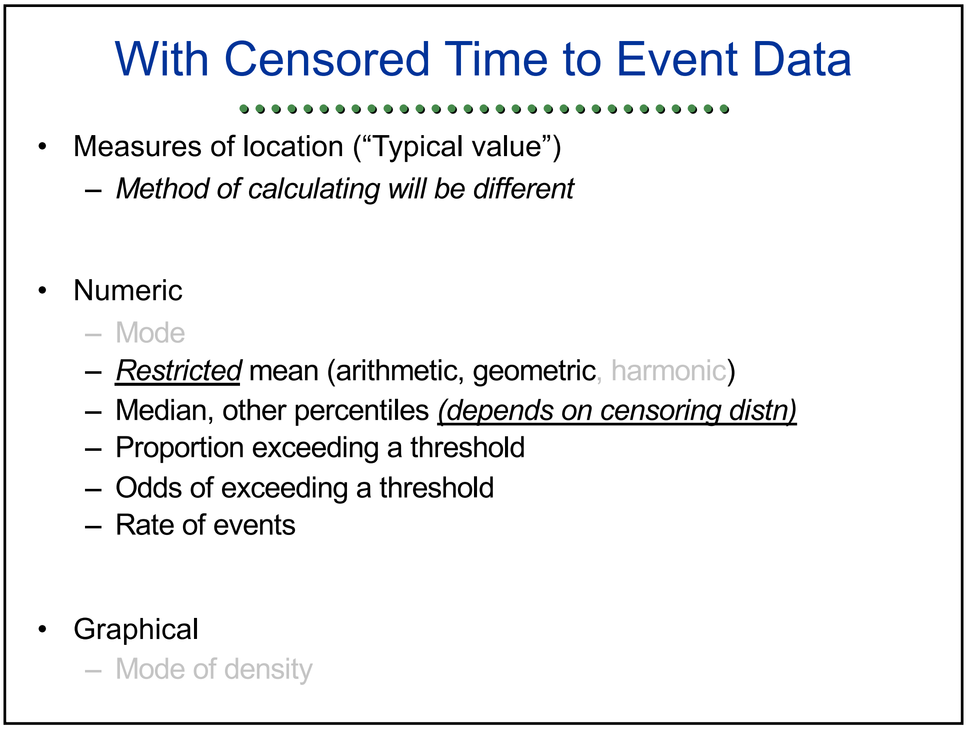
- By feature of distribution
  - Typical value (location)
  - Spread of distribution (variability)
  - Symmetry of distribution (skewness)
  - Tendency to extreme values (kurtosis)
  - Depiction of entire distribution
- By number of variables described
  - Univariate
  - Bivariate
  - Higher dimensional



Univariate Location

- Measures of location (“Typical value”)
- Numeric
  - Mode
  - Mean (arithmetic, geometric, harmonic)
  - Median (other percentiles)
  - Proportion exceeding a threshold
  - Odds of exceeding a threshold
  - Rate of events
- Graphical
  - Mode of density

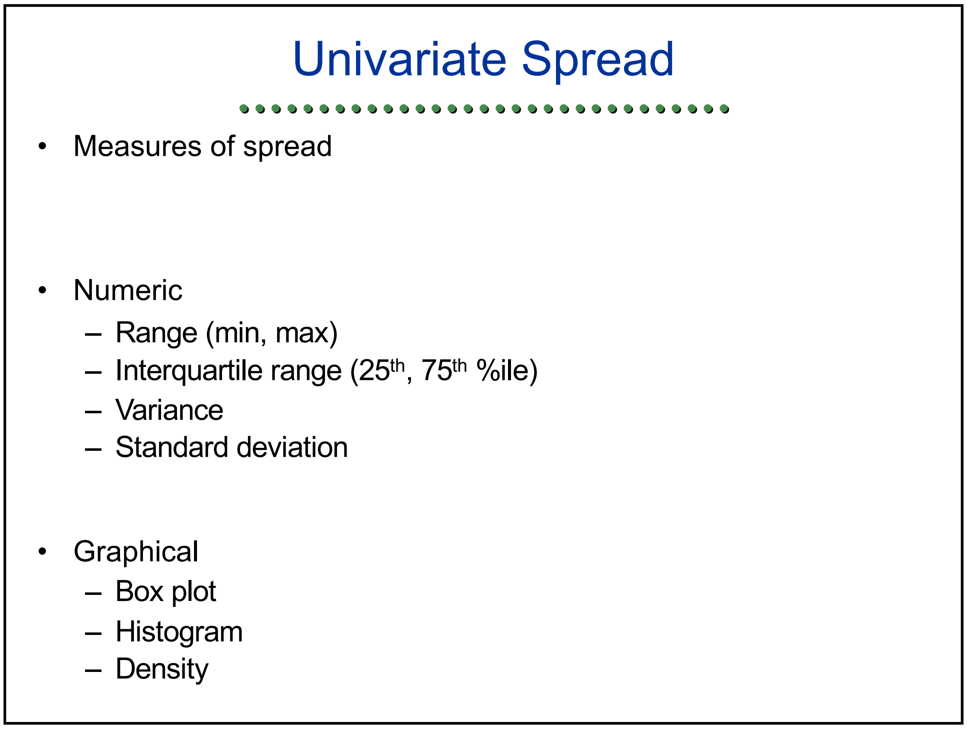
7



With Censored Time to Event Data

- Measures of location (“Typical value”)
  - *Method of calculating will be different*
- Numeric
  - Mode
  - *Restricted* mean (arithmetic, geometric, harmonic)
  - Median, other percentiles *(depends on censoring distn)*
  - Proportion exceeding a threshold
  - Odds of exceeding a threshold
  - Rate of events
- Graphical
  - Mode of density

8



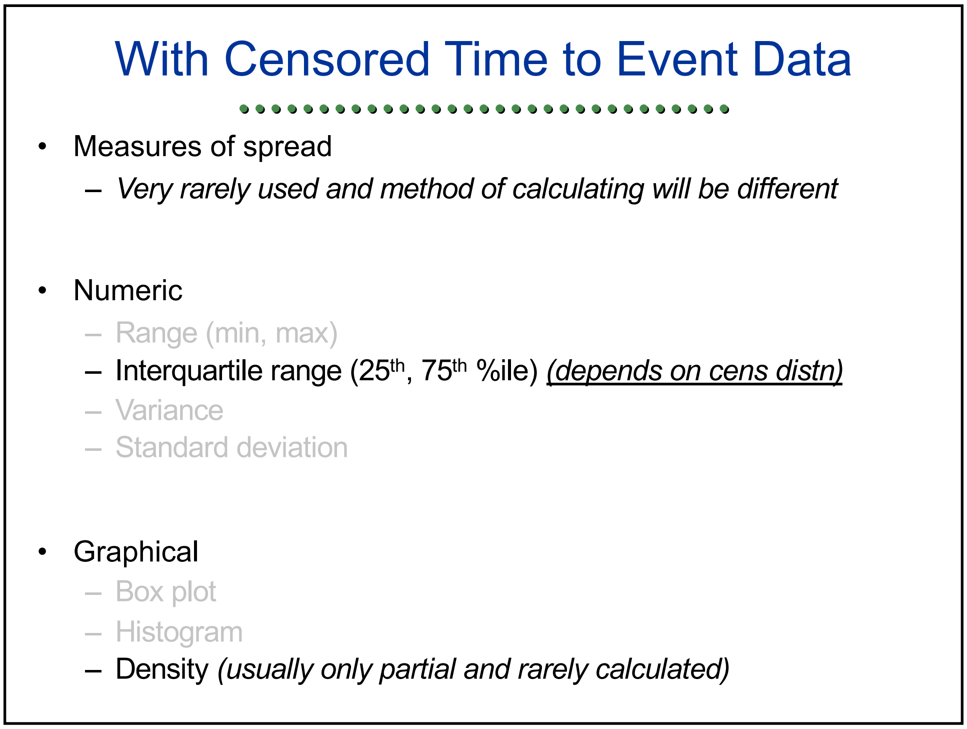
Univariate Spread

- Measures of spread

- Numeric
  - Range (min, max)
  - Interquartile range (25th, 75th %ile)
  - Variance
  - Standard deviation

- Graphical
  - Box plot
  - Histogram
  - Density

9



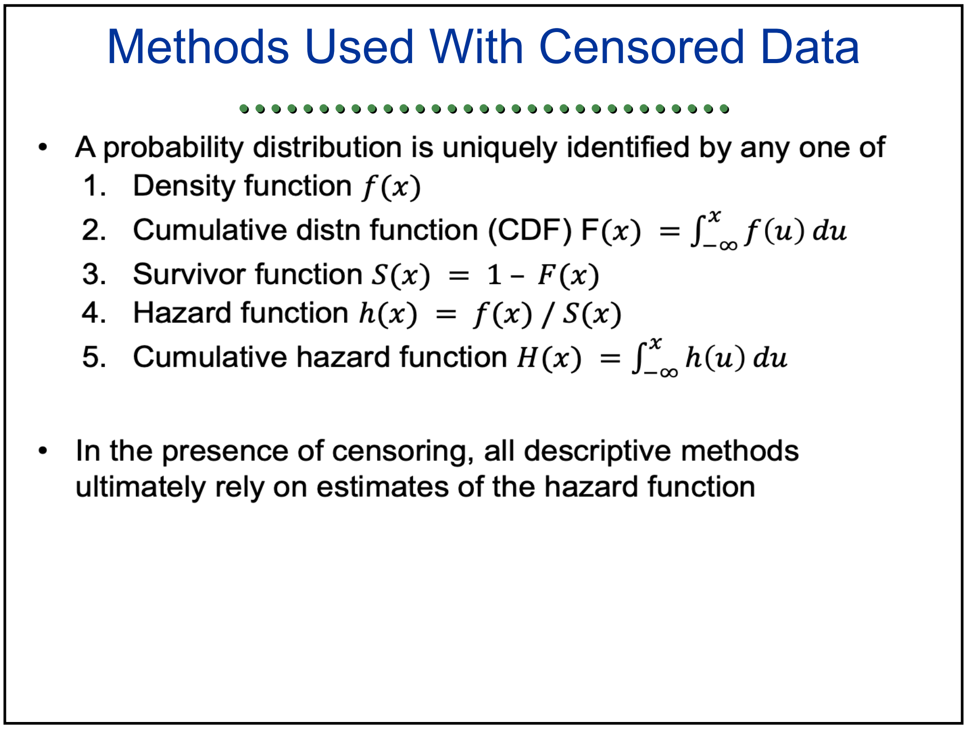
With Censored Time to Event Data

- Measures of spread
  - *Very rarely used and method of calculating will be different*

- Numeric
  - Range (min, max)
  - Interquartile range (25th, 75th %ile) *(depends on cens distn)*
  - Variance
  - Standard deviation

- Graphical
  - Box plot
  - Histogram
  - Density *(usually only partial and rarely calculated)*

10

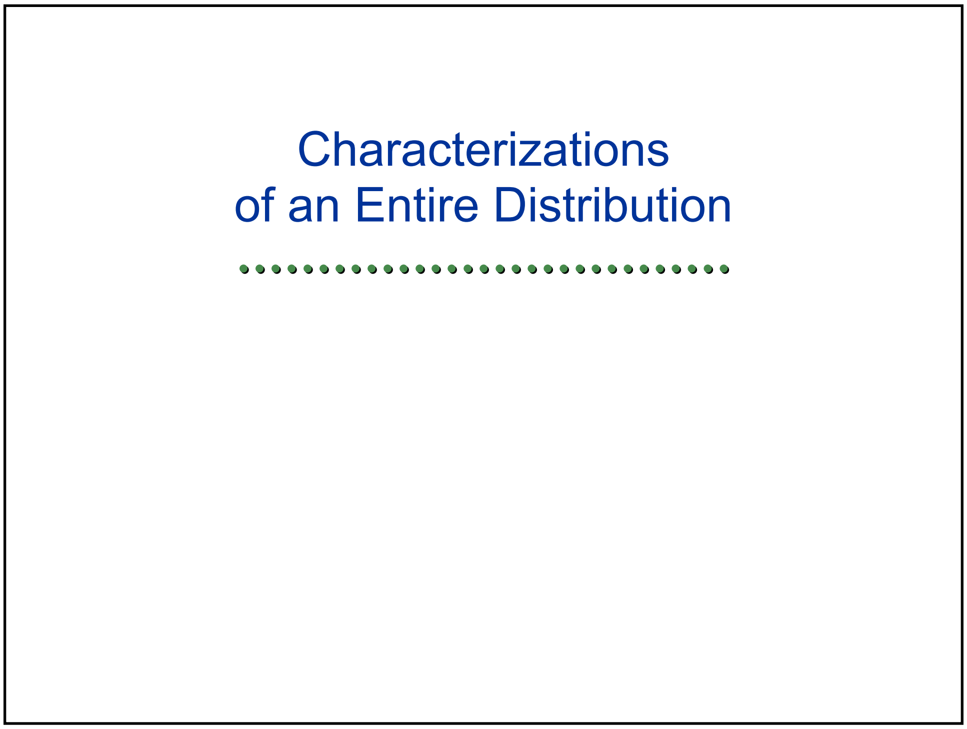


Methods Used With Censored Data

- A probability distribution is uniquely identified by any one of
  1. Density function $f(x)$
  2. Cumulative distn function (CDF) $\mathrm{F}(x) = \int_{-\infty}^{x} f(u)\, du$
  3. Survivor function $S(x) = 1 - F(x)$
  4. Hazard function $h(x) = f(x) \,/\, S(x)$
  5. Cumulative hazard function $H(x) = \int_{-\infty}^{x} h(u)\, du$

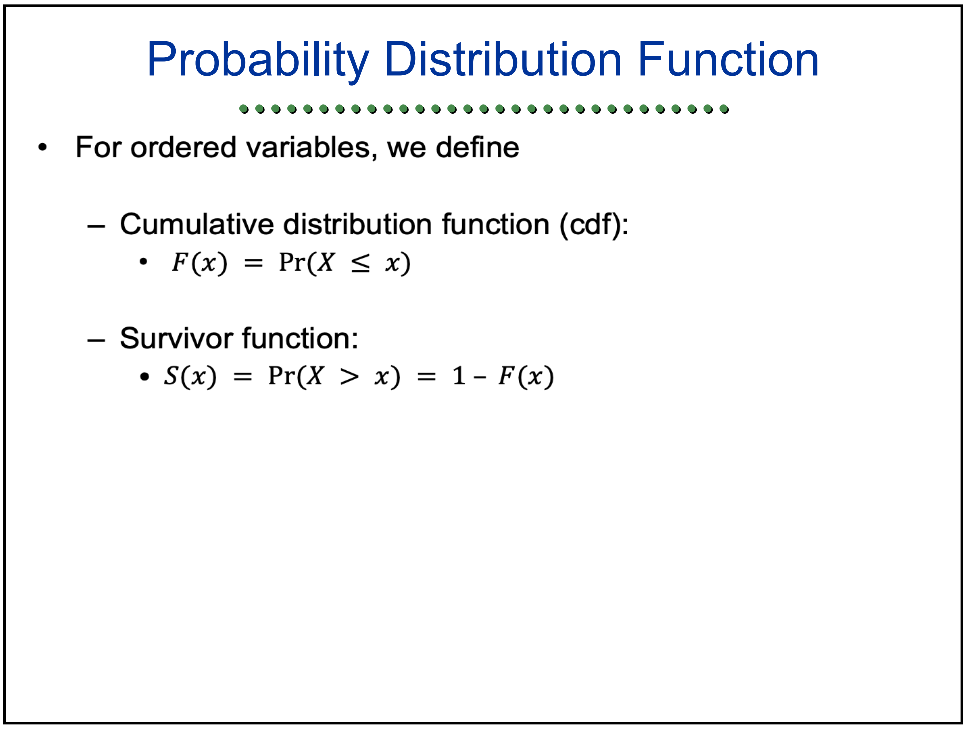
- In the presence of censoring, all descriptive methods ultimately rely on estimates of the hazard function

11



Characterizations of an Entire Distribution

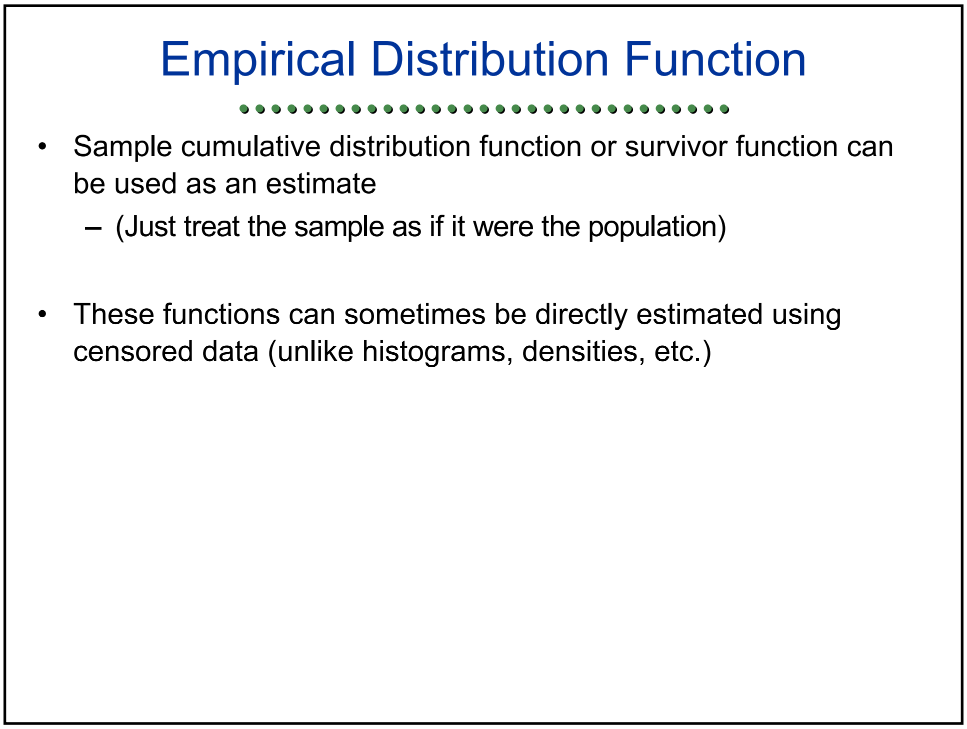
12



Probability Distribution Function

- For ordered variables, we define
  - Cumulative distribution function (cdf):
    - $F(x) = \Pr(X \leq x)$
  - Survivor function:
    - $S(x) = \Pr(X > x) = 1 - F(x)$

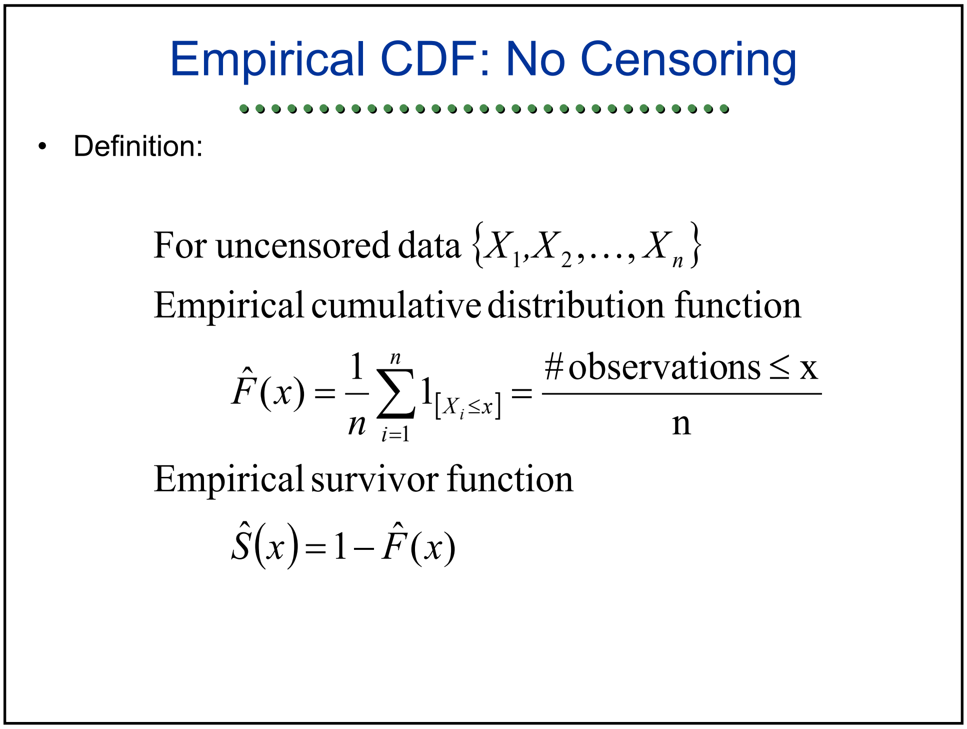
13



Empirical Distribution Function

- Sample cumulative distribution function or survivor function can be used as an estimate
  - (Just treat the sample as if it were the population)
- These functions can sometimes be directly estimated using censored data (unlike histograms, densities, etc.)

14



Empirical CDF: No Censoring

- Definition:

For uncensored data $\{X_1, X_2, \ldots, X_n\}$

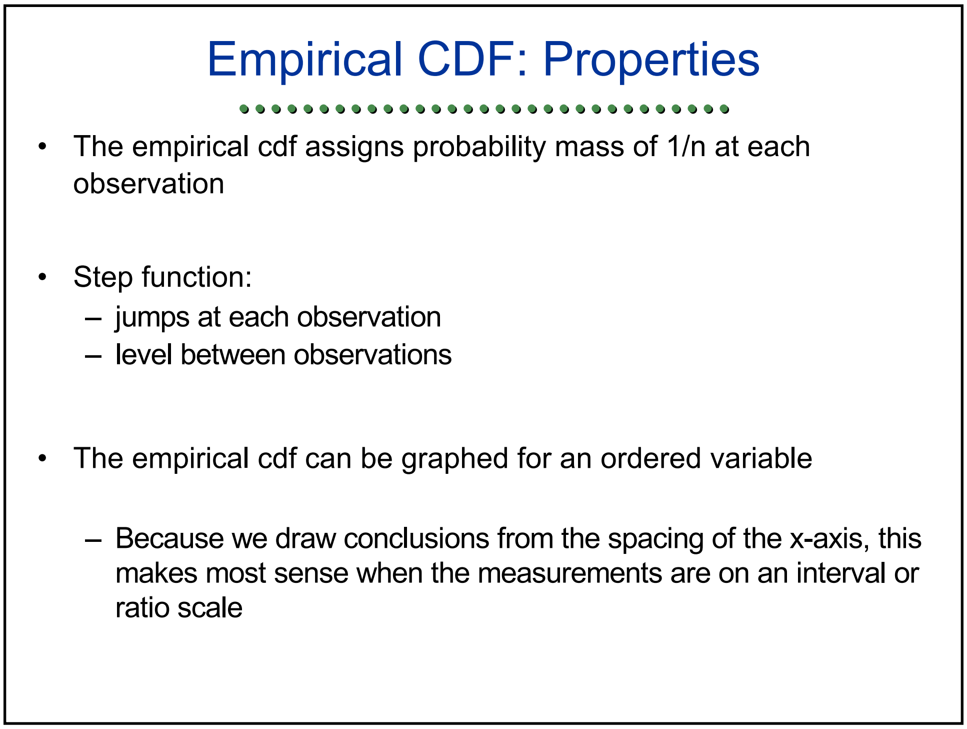
Empirical cumulative distribution function

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{[X_i \le x]} = \frac{\#\text{observations} \le \text{x}}{\text{n}}$$

Empirical survivor function

$$\hat{S}(x) = 1 - \hat{F}(x)$$

15



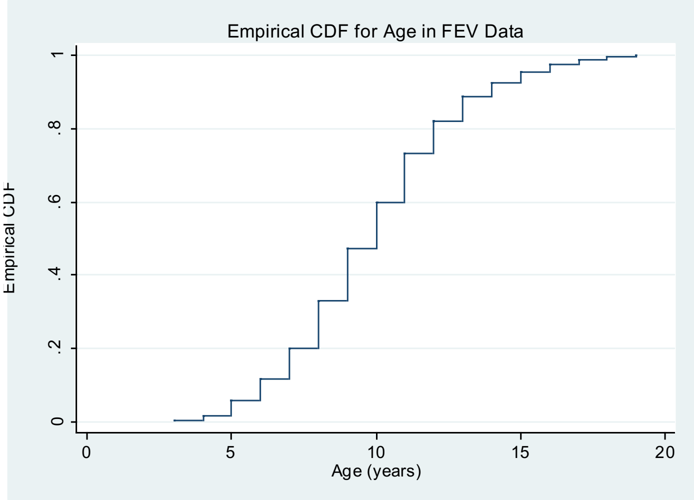
Empirical CDF: Properties

- The empirical cdf assigns probability mass of 1/n at each observation
- Step function:
  - jumps at each observation
  - level between observations
- The empirical cdf can be graphed for an ordered variable
  - Because we draw conclusions from the spacing of the x-axis, this makes most sense when the measurements are on an interval or ratio scale

16

Ex: Age CDF (FEV data)

- From an observational dataset exploring associations between smoking and lung function in children

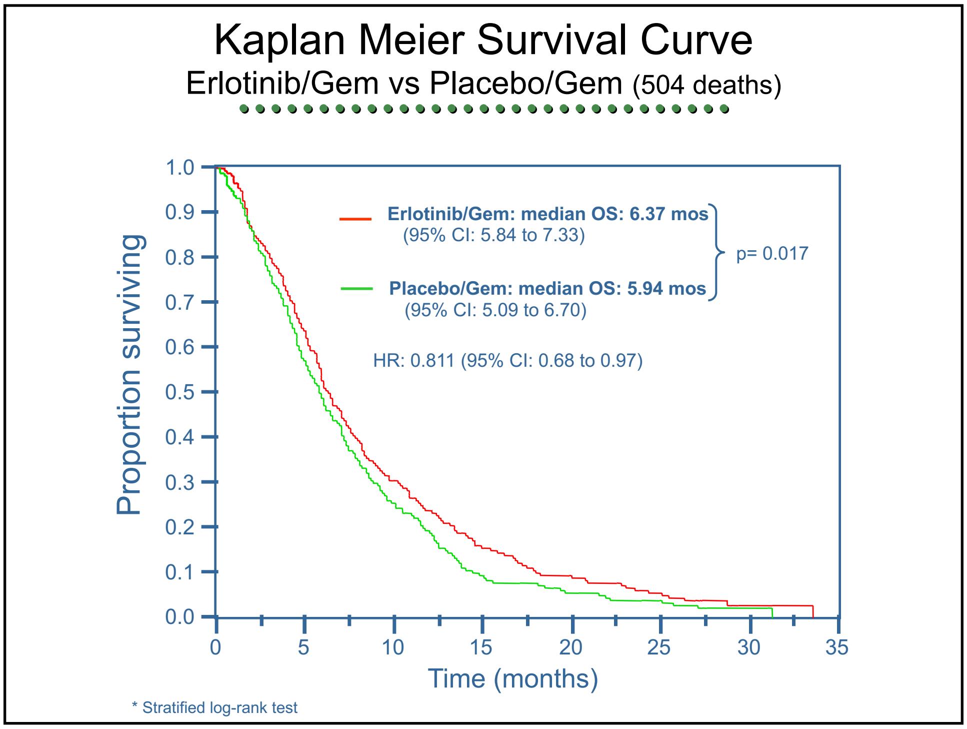


17

Survivor Curves

- In biomedicine, we typically look at the “survivor” curves for times to an event, rather than the CDF
- Note that we can “see” many common sample statistics from a plot of any survival curve
- (With censored data, we will use the Kaplan-Meier estimate, rather than the empirical CDF, to obtain the survival curve)

18



Kaplan Meier Survival Curve

Erlotinib/Gem vs Placebo/Gem (504 deaths)

Erlotinib/Gem: median OS: 6.37 mos (95% CI: 5.84 to 7.33)

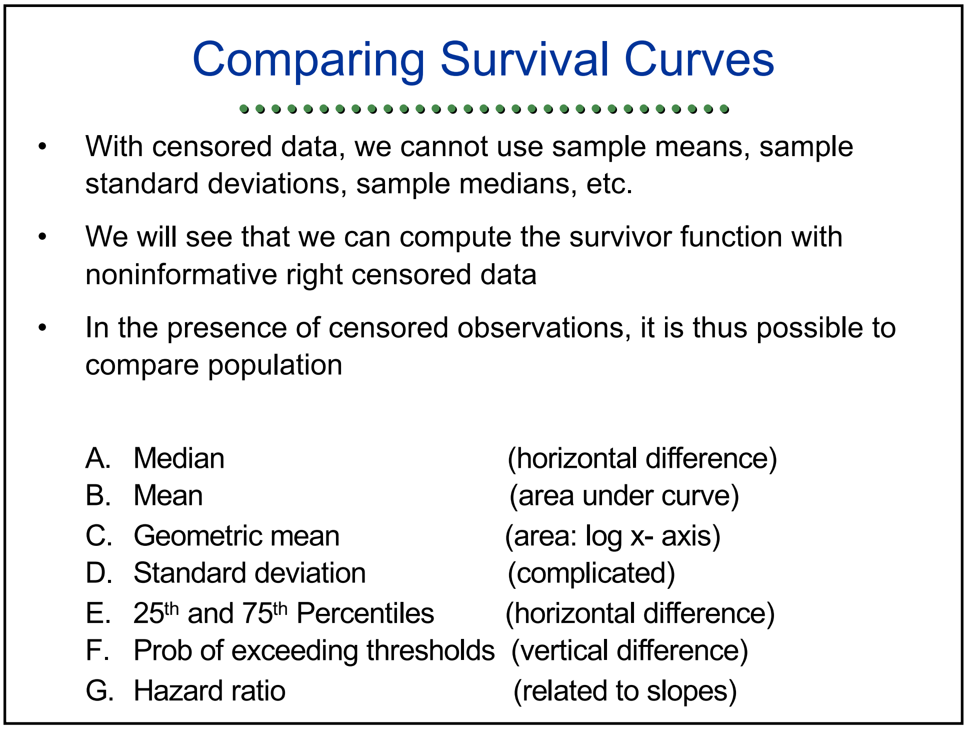
Placebo/Gem: median OS: 5.94 mos (95% CI: 5.09 to 6.70)

p= 0.017

HR: 0.811 (95% CI: 0.68 to 0.97)

* Stratified log-rank test

19

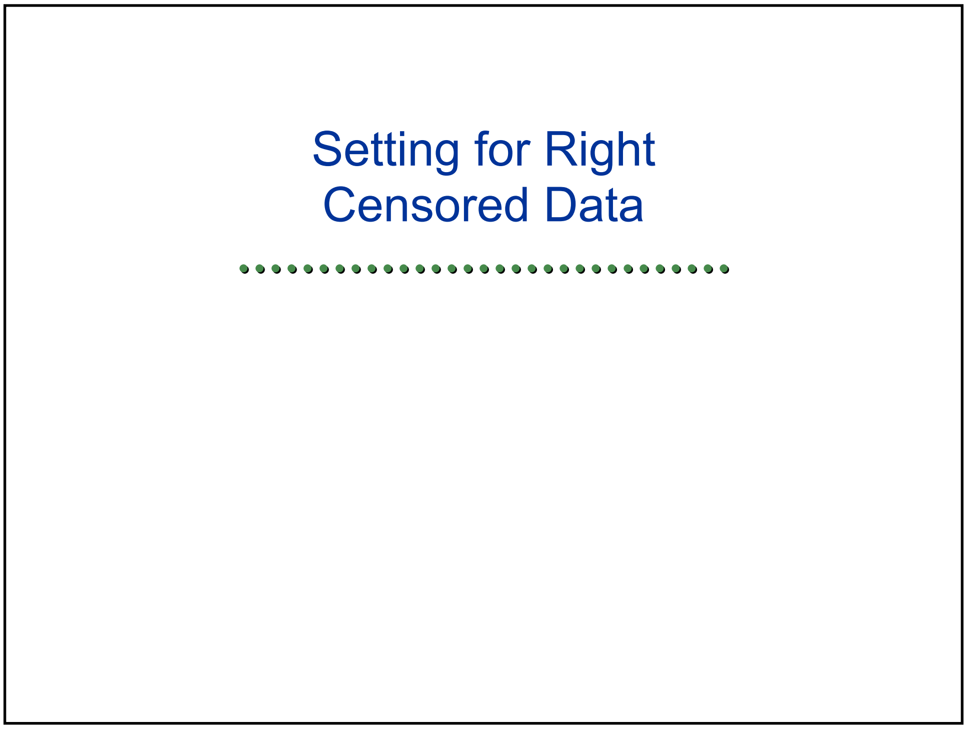


Comparing Survival Curves

- With censored data, we cannot use sample means, sample standard deviations, sample medians, etc.
- We will see that we can compute the survivor function with noninformative right censored data
- In the presence of censored observations, it is thus possible to compare population

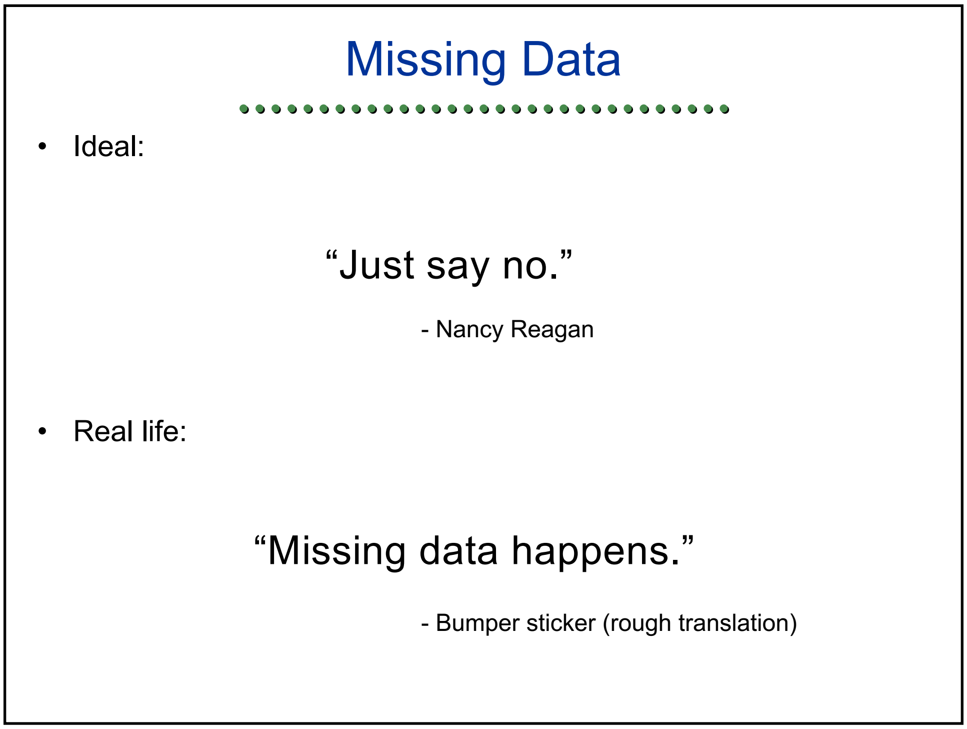
A.	Median	(horizontal difference)
B.	Mean	(area under curve)
C.	Geometric mean	(area: log x- axis)
D.	Standard deviation	(complicated)
E.	25th and 75th Percentiles	(horizontal difference)
F.	Prob of exceeding thresholds	(vertical difference)
G.	Hazard ratio	(related to slopes)

20



Setting for Right Censored Data

21



Missing Data

- Ideal:

“Just say no.”

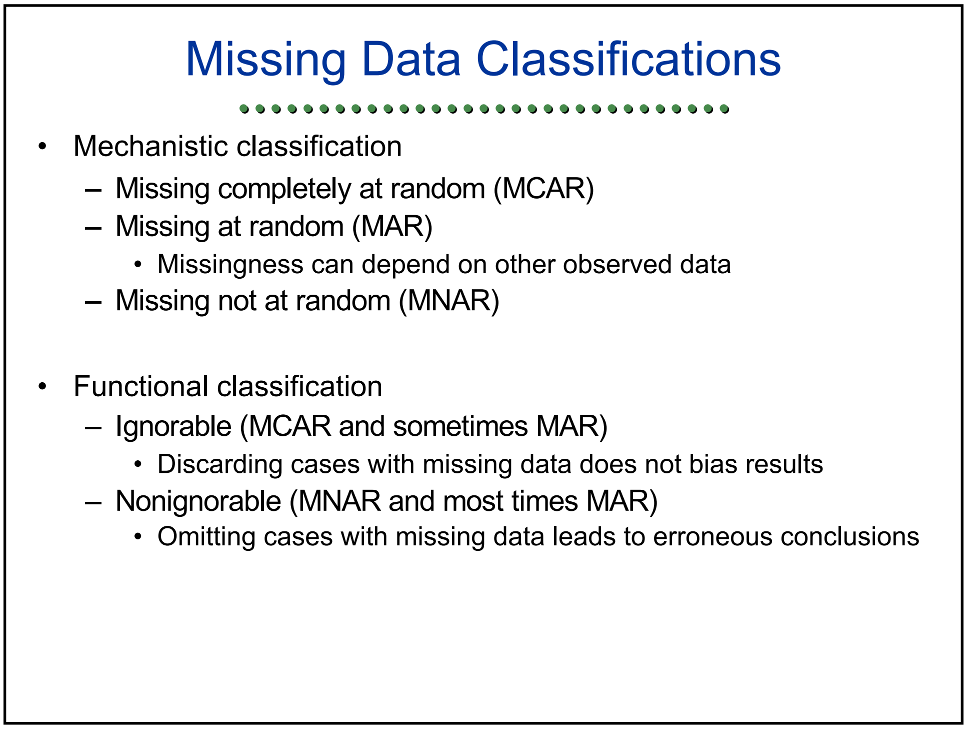
- Nancy Reagan

- Real life:

“Missing data happens.”

- Bumper sticker (rough translation)

22

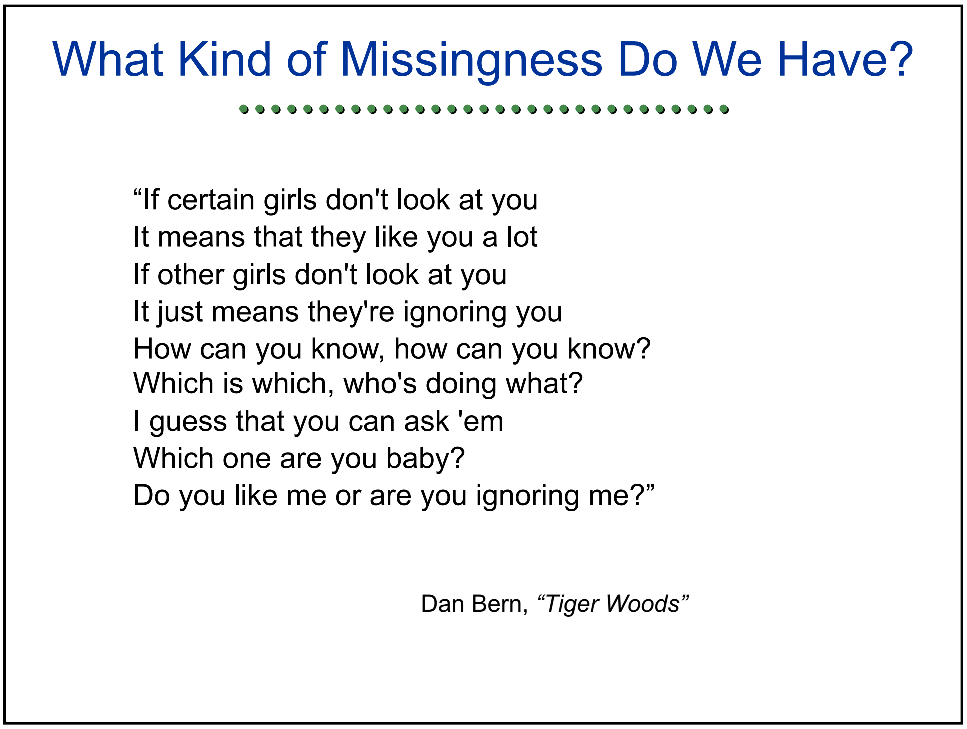


Missing Data Classifications

- Mechanistic classification
  - Missing completely at random (MCAR)
  - Missing at random (MAR)
    - Missingness can depend on other observed data
  - Missing not at random (MNAR)

- Functional classification
  - Ignorable (MCAR and sometimes MAR)
    - Discarding cases with missing data does not bias results
  - Nonignorable (MNAR and most times MAR)
    - Omitting cases with missing data leads to erroneous conclusions

23

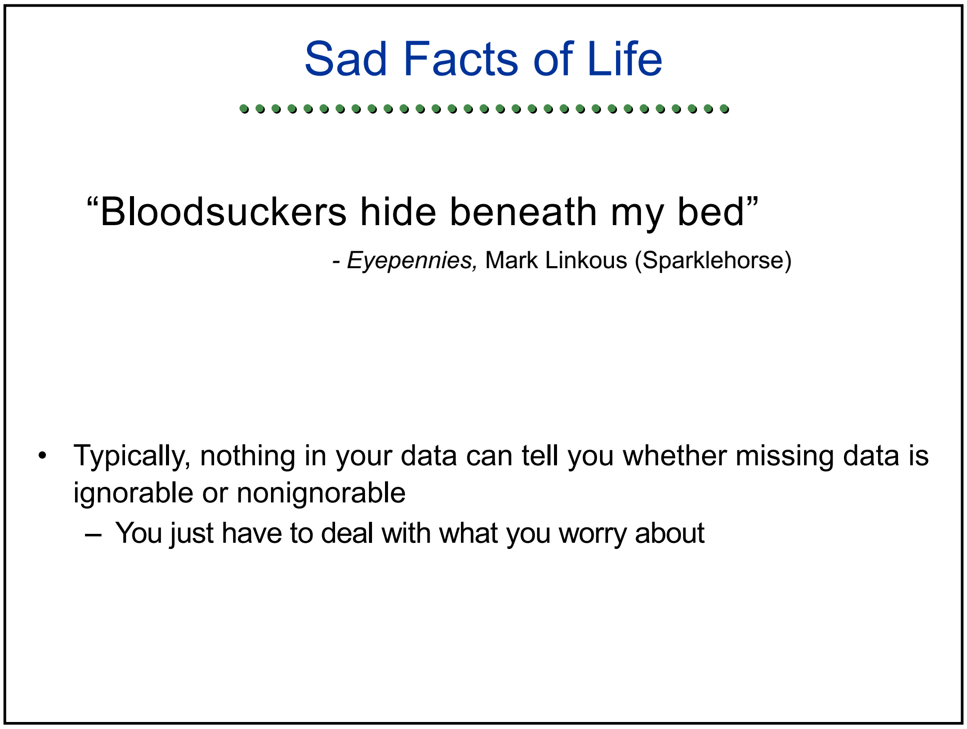


What Kind of Missingness Do We Have?

“If certain girls don't look at you
It means that they like you a lot
If other girls don't look at you
It just means they're ignoring you
How can you know, how can you know?
Which is which, who's doing what?
I guess that you can ask 'em
Which one are you baby?
Do you like me or are you ignoring me?”

Dan Bern, *“Tiger Woods”*

24

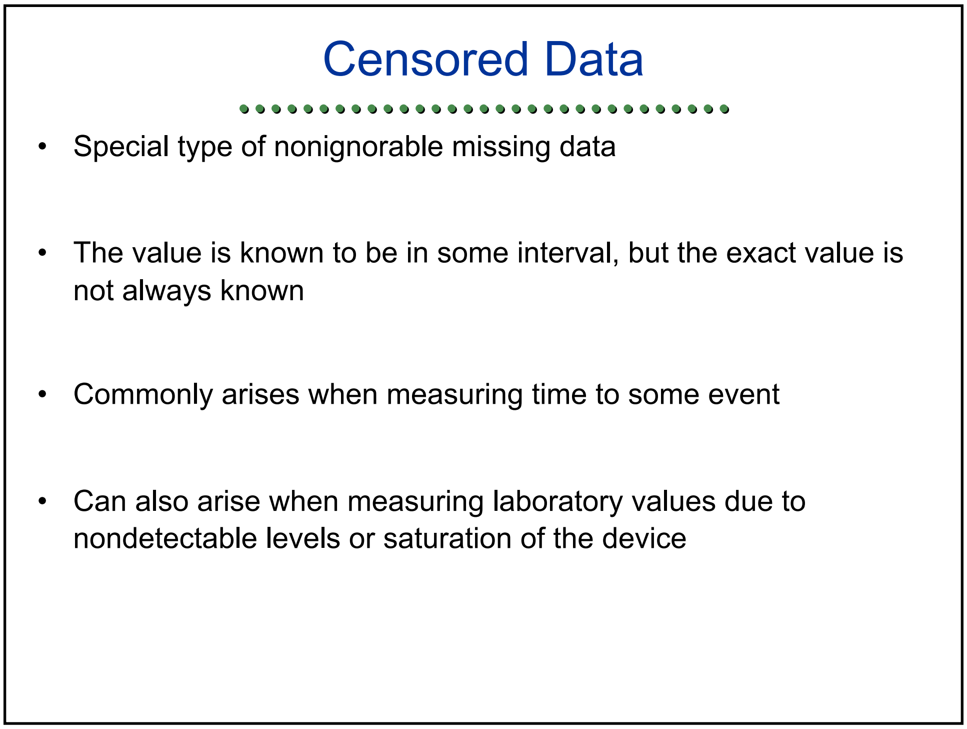


Sad Facts of Life

> "Bloodsuckers hide beneath my bed"
>
> \- *Eyepennies,* Mark Linkous (Sparklehorse)

- Typically, nothing in your data can tell you whether missing data is ignorable or nonignorable
  - You just have to deal with what you worry about

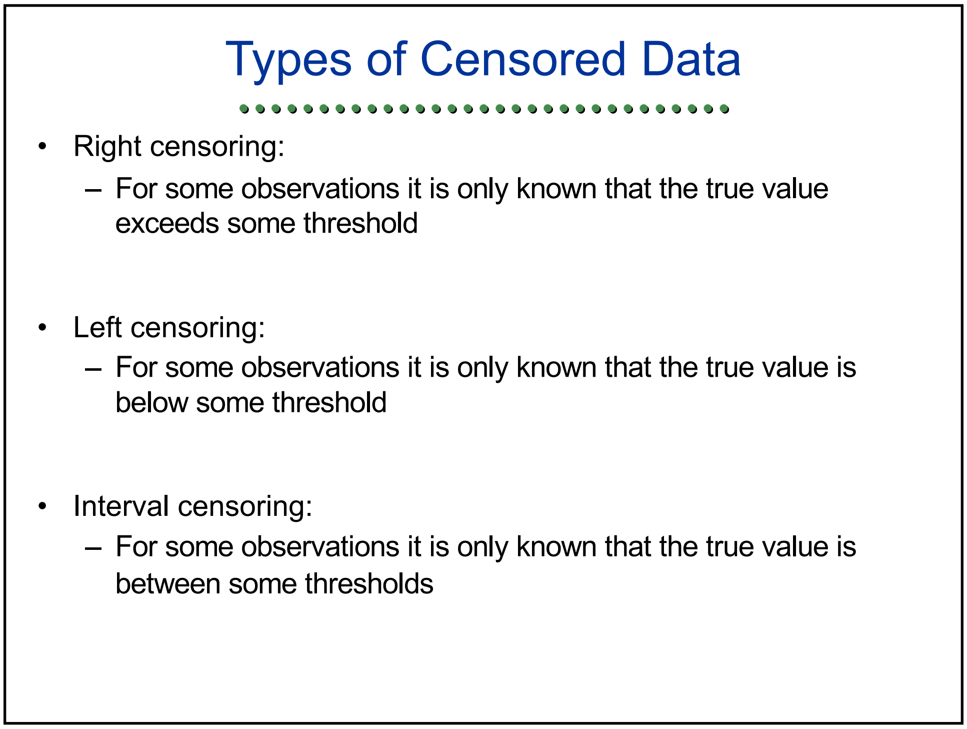
25



Censored Data

- Special type of nonignorable missing data
- The value is known to be in some interval, but the exact value is not always known
- Commonly arises when measuring time to some event
- Can also arise when measuring laboratory values due to nondetectable levels or saturation of the device

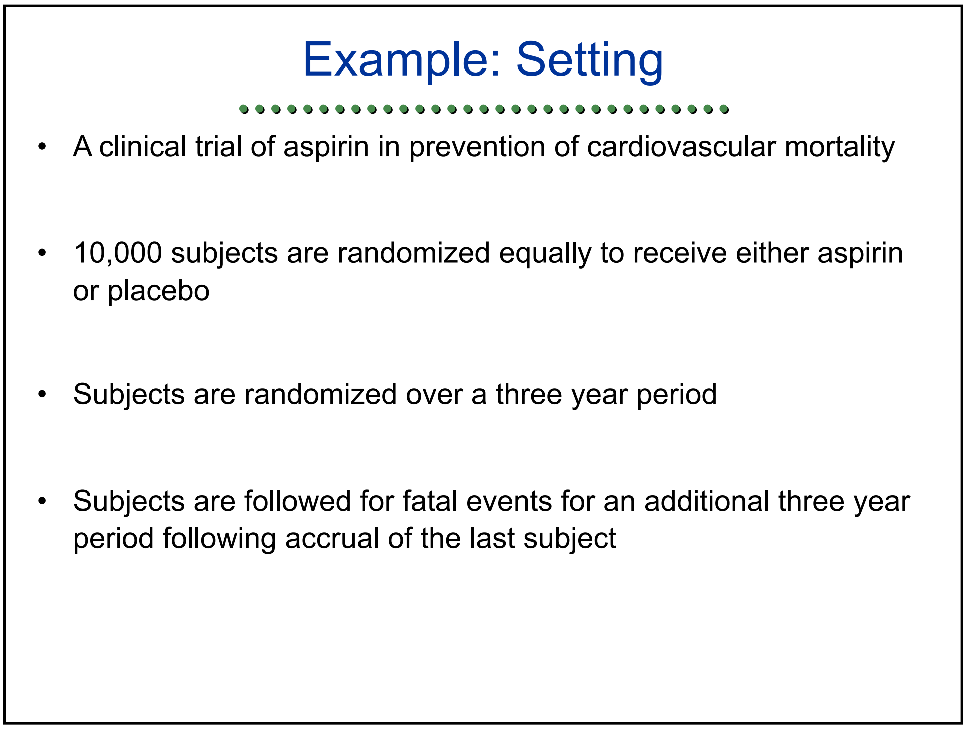
26



Types of Censored Data

- Right censoring:
  - For some observations it is only known that the true value exceeds some threshold
- Left censoring:
  - For some observations it is only known that the true value is below some threshold
- Interval censoring:
  - For some observations it is only known that the true value is between some thresholds

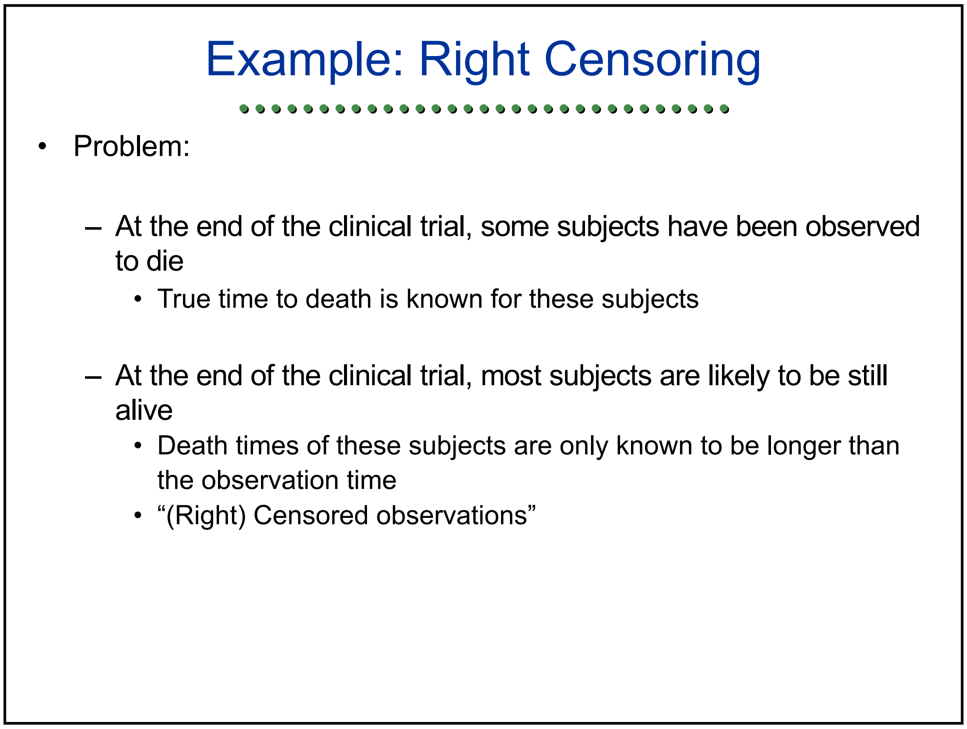
27



Example: Setting

- A clinical trial of aspirin in prevention of cardiovascular mortality
- 10,000 subjects are randomized equally to receive either aspirin or placebo
- Subjects are randomized over a three year period
- Subjects are followed for fatal events for an additional three year period following accrual of the last subject

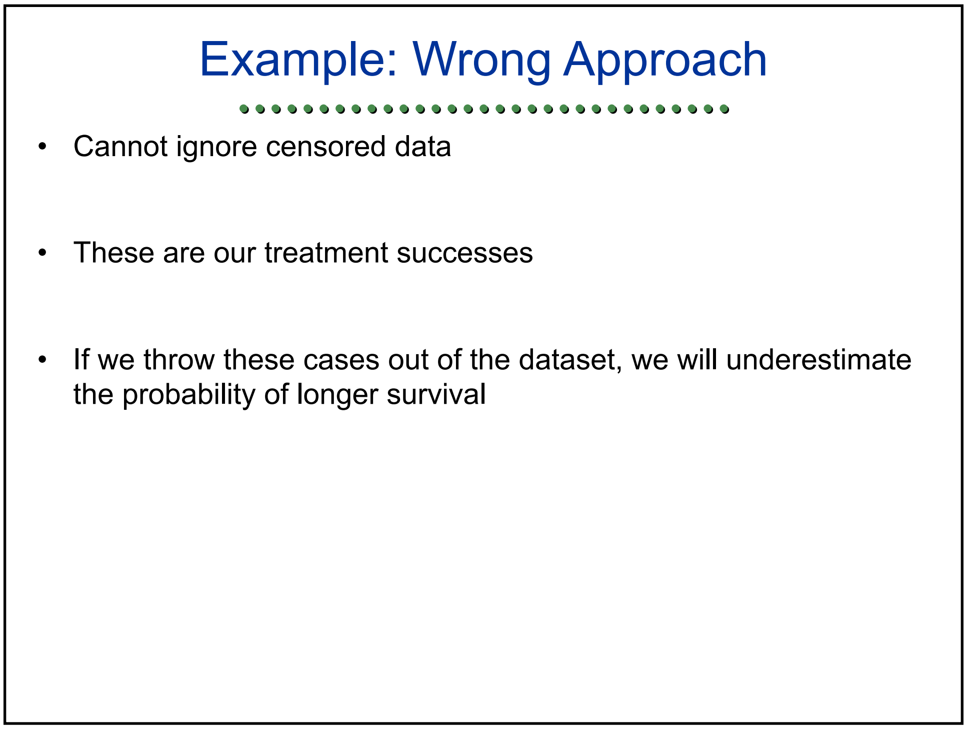
28



Example: Right Censoring

- Problem:
  - At the end of the clinical trial, some subjects have been observed to die
    - True time to death is known for these subjects
  - At the end of the clinical trial, most subjects are likely to be still alive
    - Death times of these subjects are only known to be longer than the observation time
    - “(Right) Censored observations”

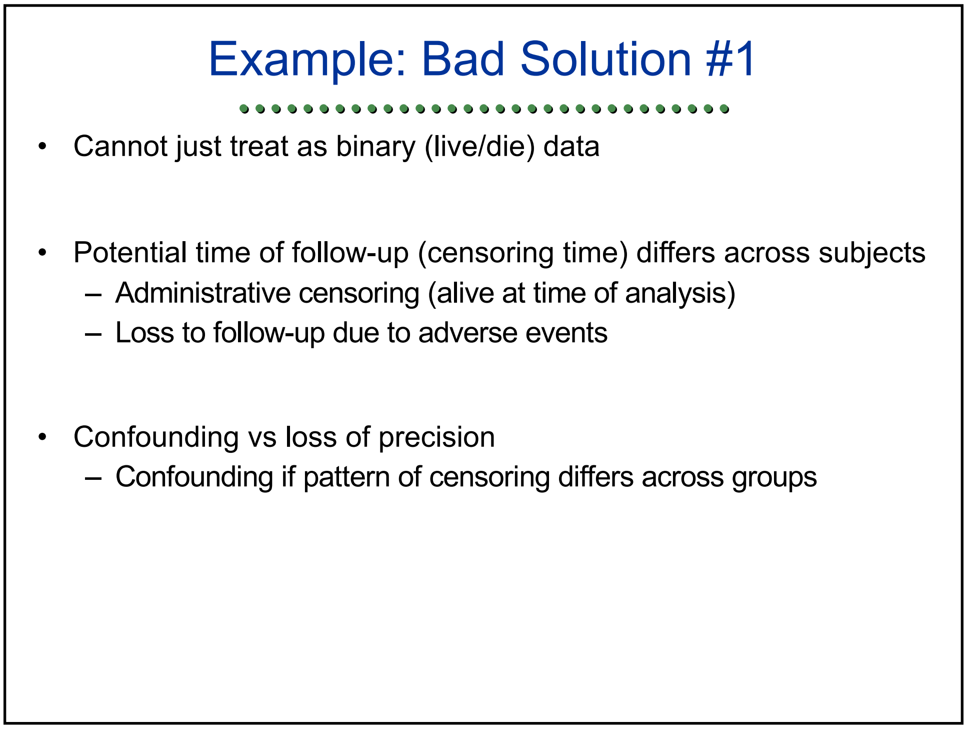
29



Example: Wrong Approach

- Cannot ignore censored data
- These are our treatment successes
- If we throw these cases out of the dataset, we will underestimate the probability of longer survival

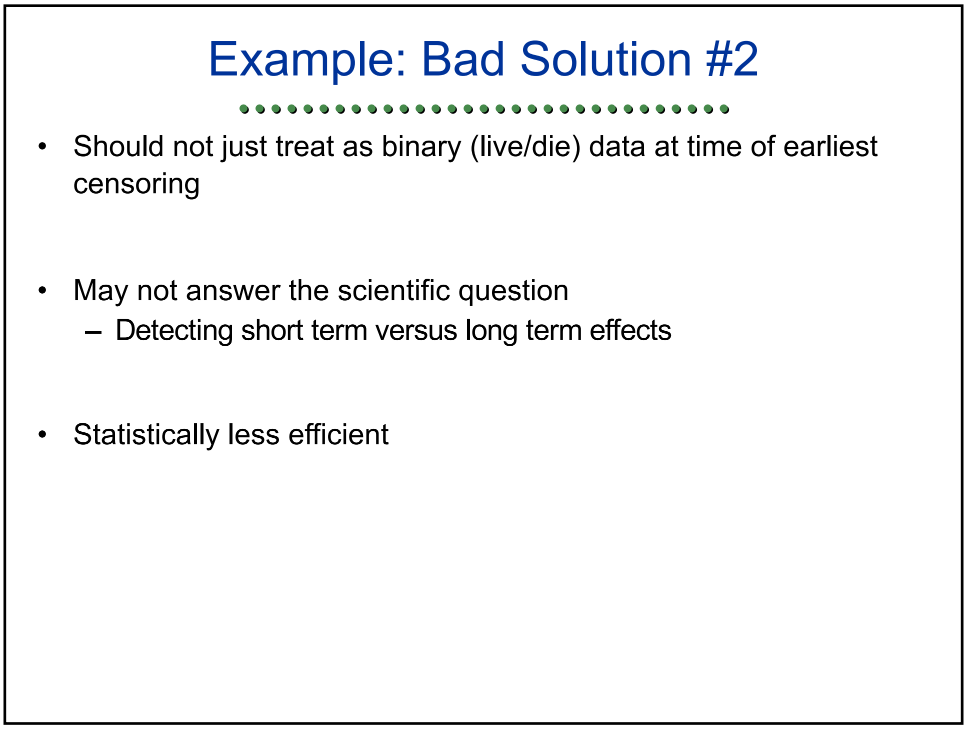
30



Example: Bad Solution #1

- Cannot just treat as binary (live/die) data
- Potential time of follow-up (censoring time) differs across subjects
  - Administrative censoring (alive at time of analysis)
  - Loss to follow-up due to adverse events
- Confounding vs loss of precision
  - Confounding if pattern of censoring differs across groups

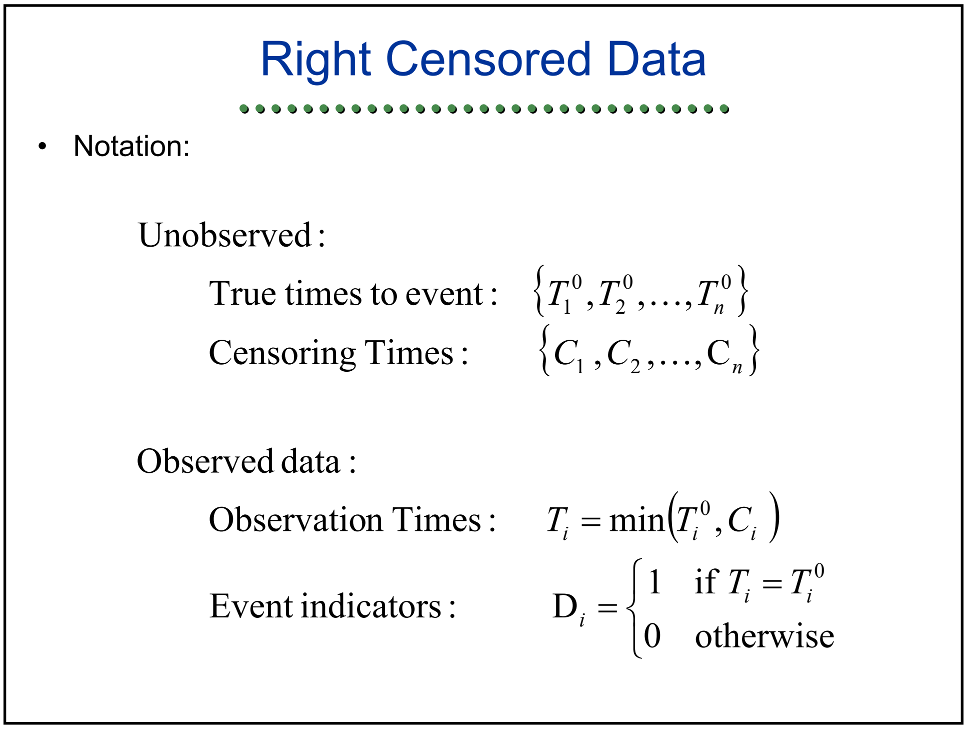
31



Example: Bad Solution #2

- Should not just treat as binary (live/die) data at time of earliest censoring
- May not answer the scientific question
  - Detecting short term versus long term effects
- Statistically less efficient

32



Right Censored Data

- Notation:

Unobserved:

True times to event: $\left\{T_1^0, T_2^0, \ldots, T_n^0\right\}$

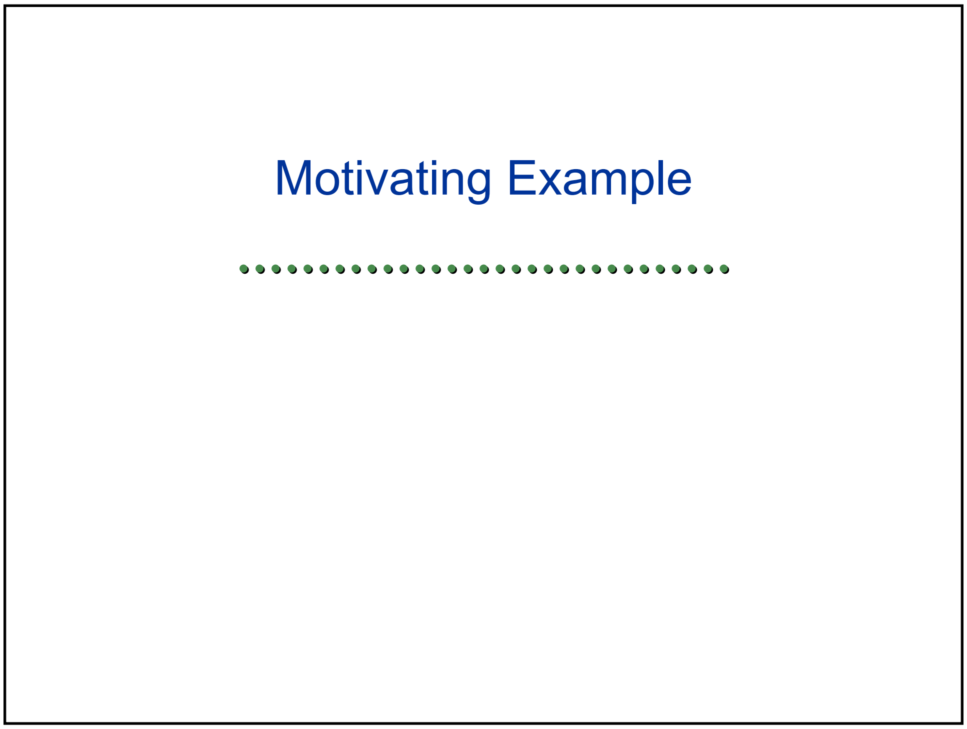
Censoring Times: $\left\{C_1, C_2, \ldots, \mathrm{C}_n\right\}$

Observed data:

Observation Times: $T_i = \min\left(T_i^0, C_i\right)$

Event indicators: $\mathrm{D}_i = \begin{cases} 1 & \text{if } T_i = T_i^0 \\ 0 & \text{otherwise} \end{cases}$

33



Motivating Example

34

Motivating Example

- Hypothetical study of subject survival
- Subjects accrued to study and followed until time of analysis
- Study done at three centers, which started the studies in three successive years
- Censoring time thus differs across centers

35

Data by Date (Real Time)

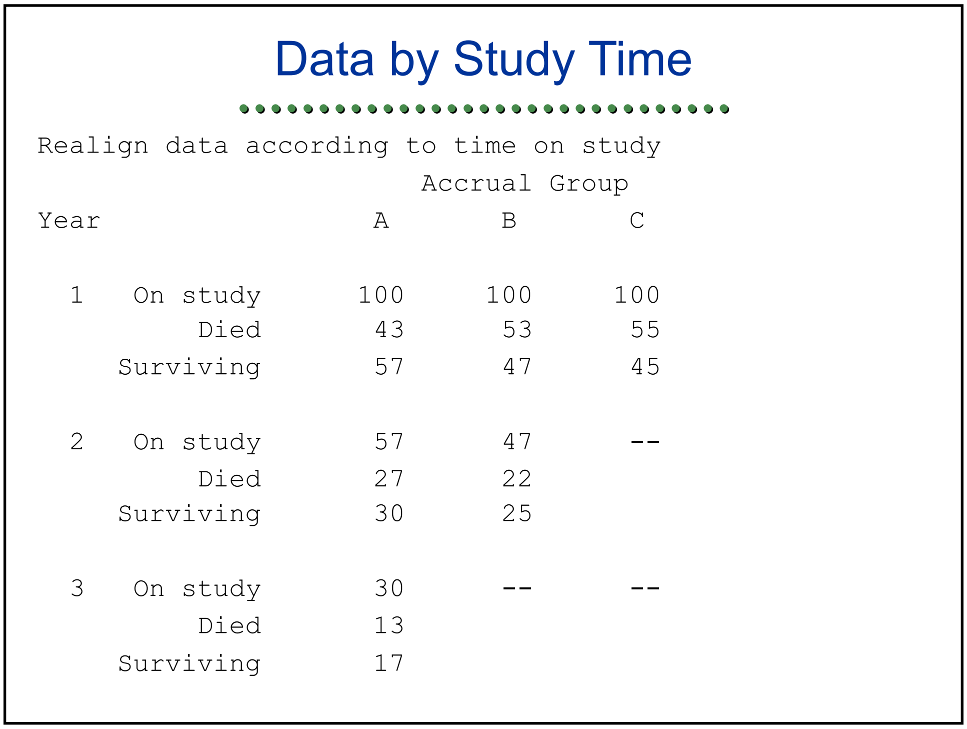
```
Staggered study entry by site
                          Accrual Group
Year                    A       B       C

2010   On study       100      --      --
           Died        43
      Surviving        57

2011   On study        57     100      --
           Died        27      53
      Surviving        30      47

2012   On study        30      47     100
           Died        13      22      55
      Surviving        17      25      45
```

36



Data by Study Time

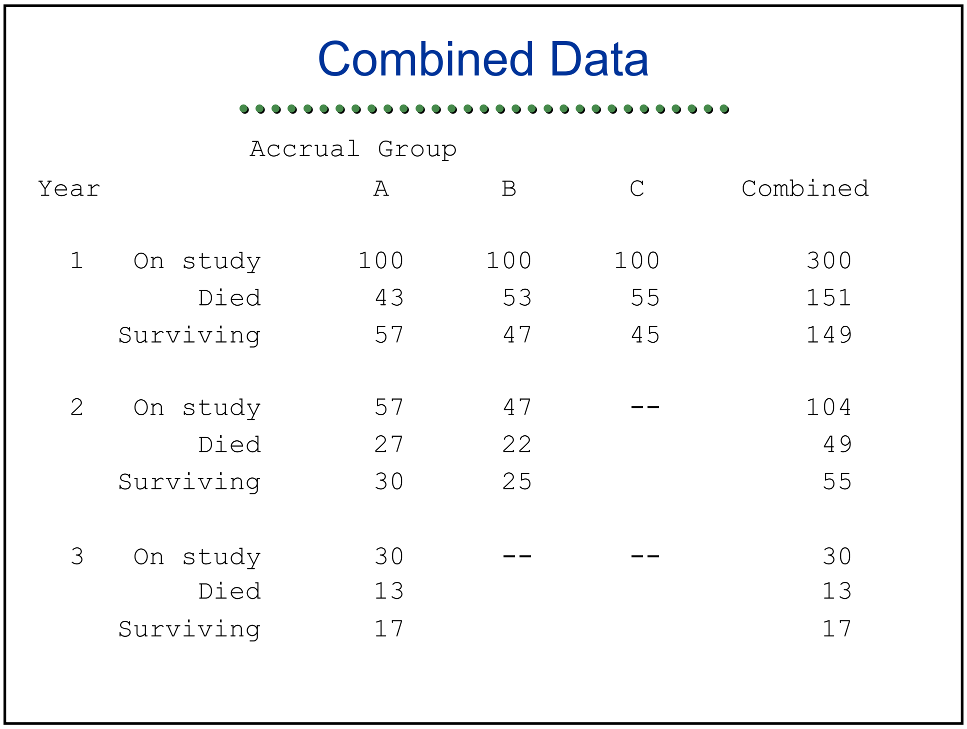
```
Realign data according to time on study
                                Accrual Group
Year                     A         B         C


  1    On study        100       100       100
            Died        43        53        55
       Surviving        57        47        45

  2    On study         57        47        --
            Died        27        22
       Surviving        30        25

  3    On study         30        --        --
            Died        13
       Surviving        17
```

37



Combined Data

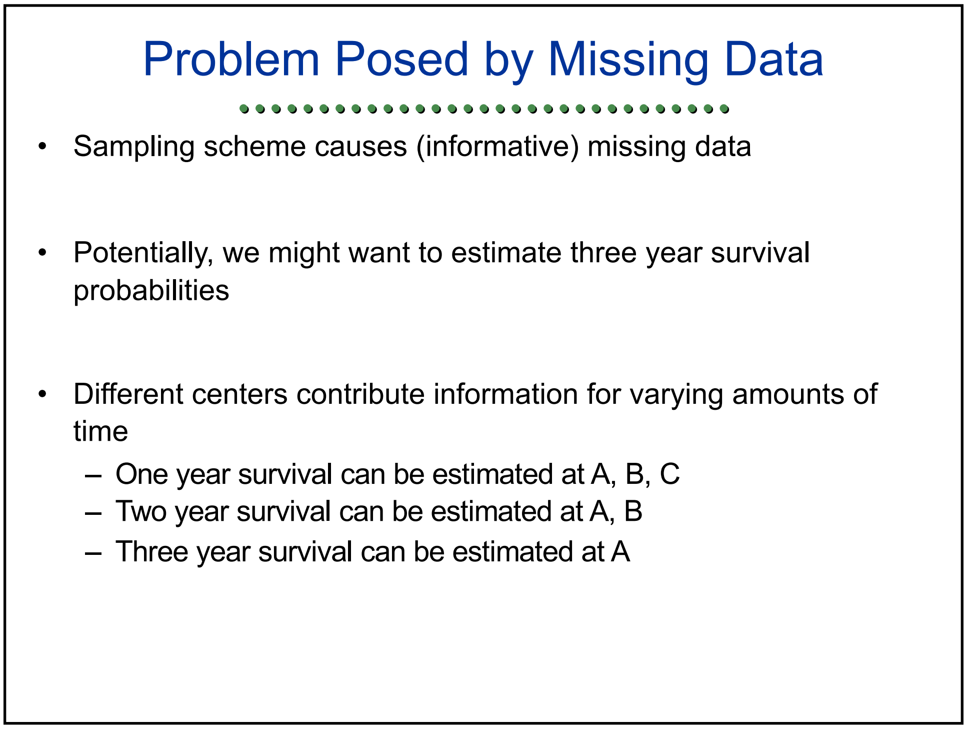
```
                 Accrual Group
Year                     A         B         C       Combined

  1    On study        100       100       100           300
            Died        43        53        55           151
       Surviving        57        47        45           149

  2    On study         57        47        --           104
            Died        27        22                      49
       Surviving        30        25                      55

  3    On study         30        --        --            30
            Died        13                                  13
       Surviving        17                                  17
```

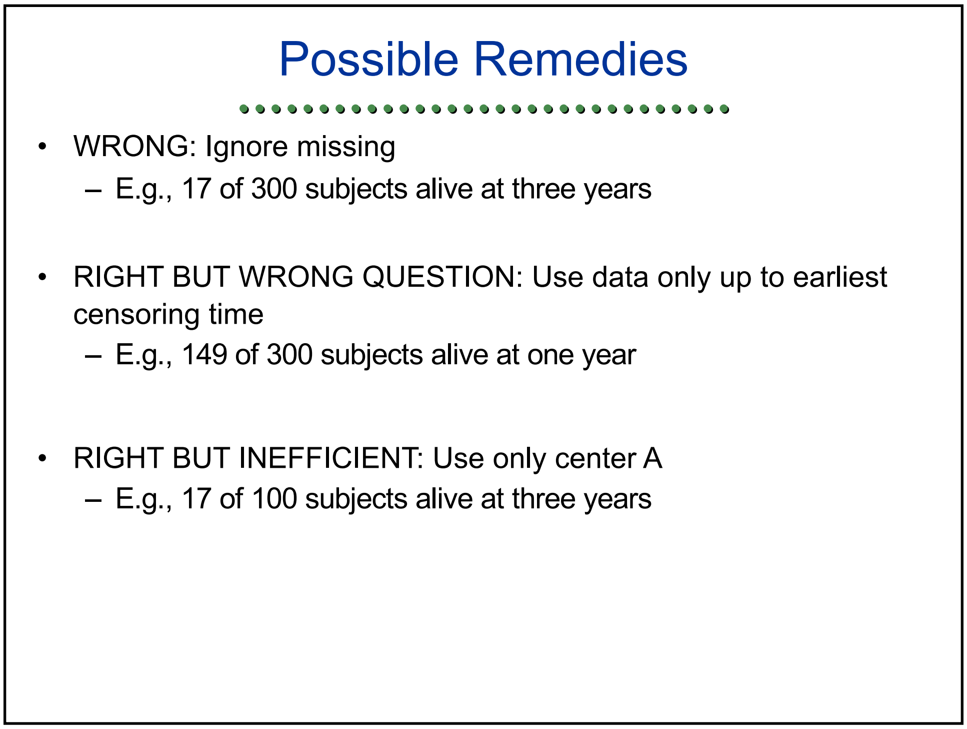
38



Problem Posed by Missing Data

- Sampling scheme causes (informative) missing data
- Potentially, we might want to estimate three year survival probabilities
- Different centers contribute information for varying amounts of time
  - One year survival can be estimated at A, B, C
  - Two year survival can be estimated at A, B
  - Three year survival can be estimated at A

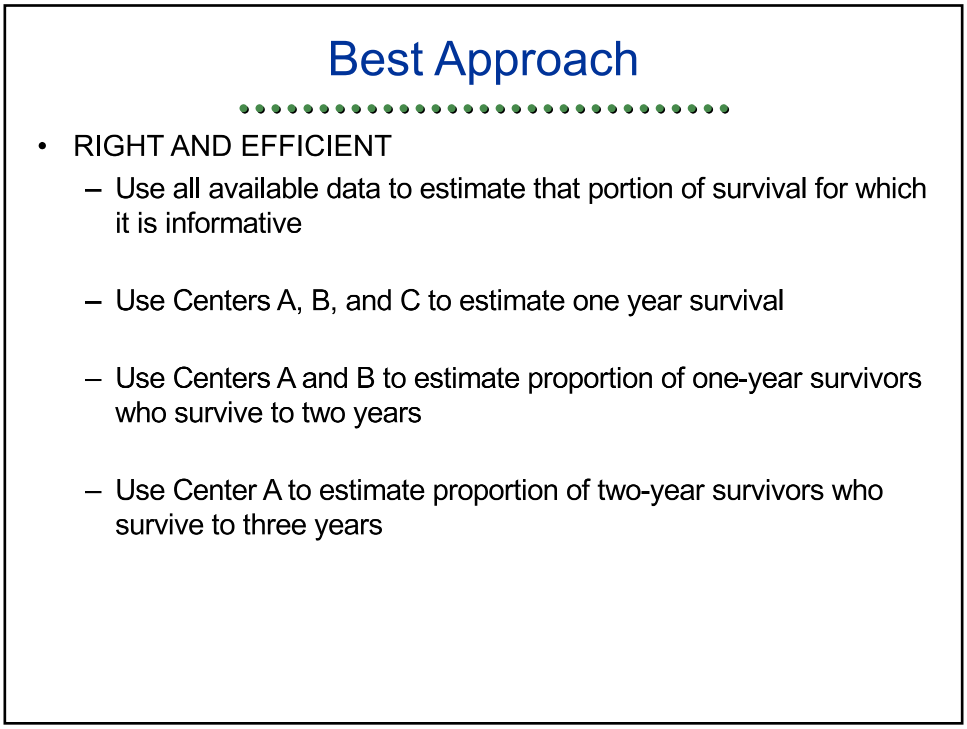
39



Possible Remedies

- WRONG: Ignore missing
  - E.g., 17 of 300 subjects alive at three years
- RIGHT BUT WRONG QUESTION: Use data only up to earliest censoring time
  - E.g., 149 of 300 subjects alive at one year
- RIGHT BUT INEFFICIENT: Use only center A
  - E.g., 17 of 100 subjects alive at three years

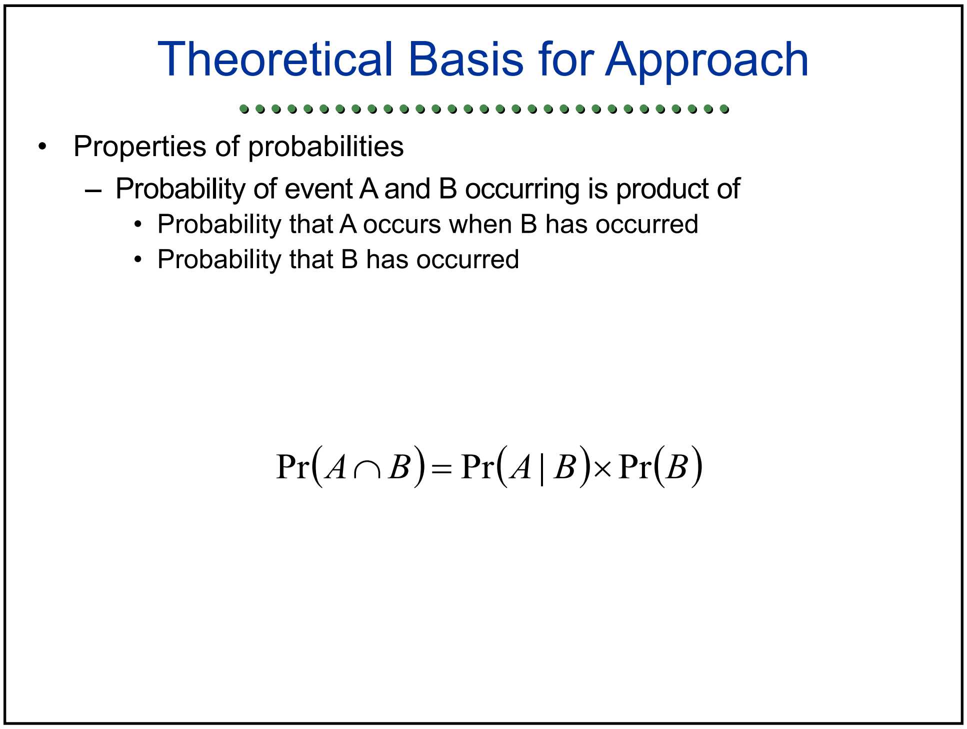
40



Best Approach

- RIGHT AND EFFICIENT
  - Use all available data to estimate that portion of survival for which it is informative
  - Use Centers A, B, and C to estimate one year survival
  - Use Centers A and B to estimate proportion of one-year survivors who survive to two years
  - Use Center A to estimate proportion of two-year survivors who survive to three years

41

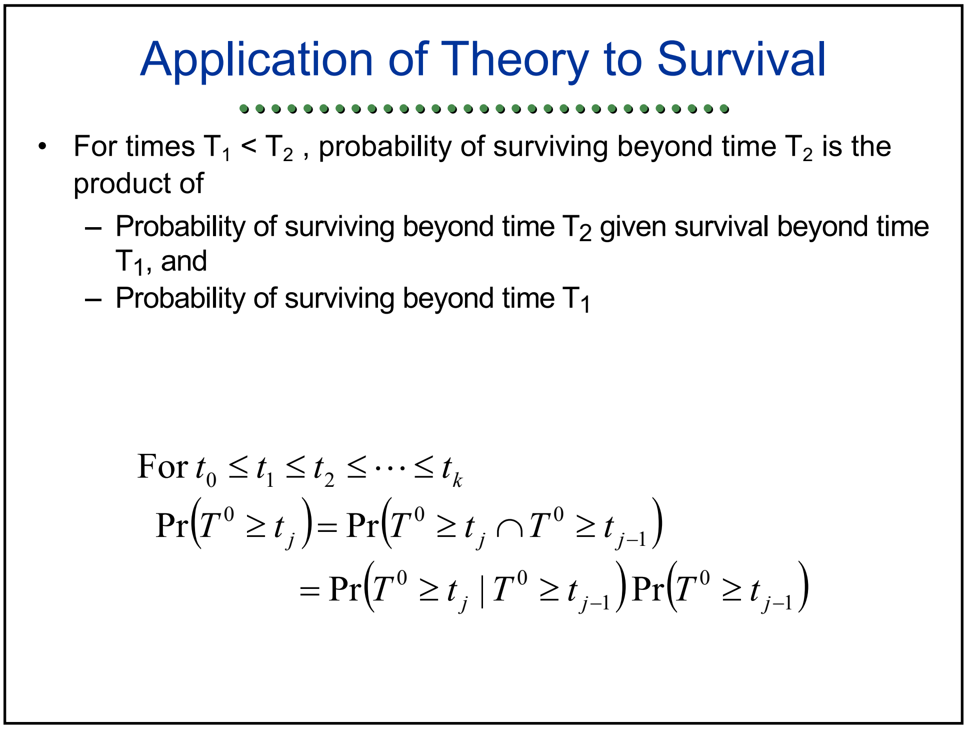


Theoretical Basis for Approach

- Properties of probabilities
  - Probability of event A and B occurring is product of
    - Probability that A occurs when B has occurred
    - Probability that B has occurred

$$\Pr\left(A \cap B\right) = \Pr\left(A \mid B\right) \times \Pr\left(B\right)$$

42



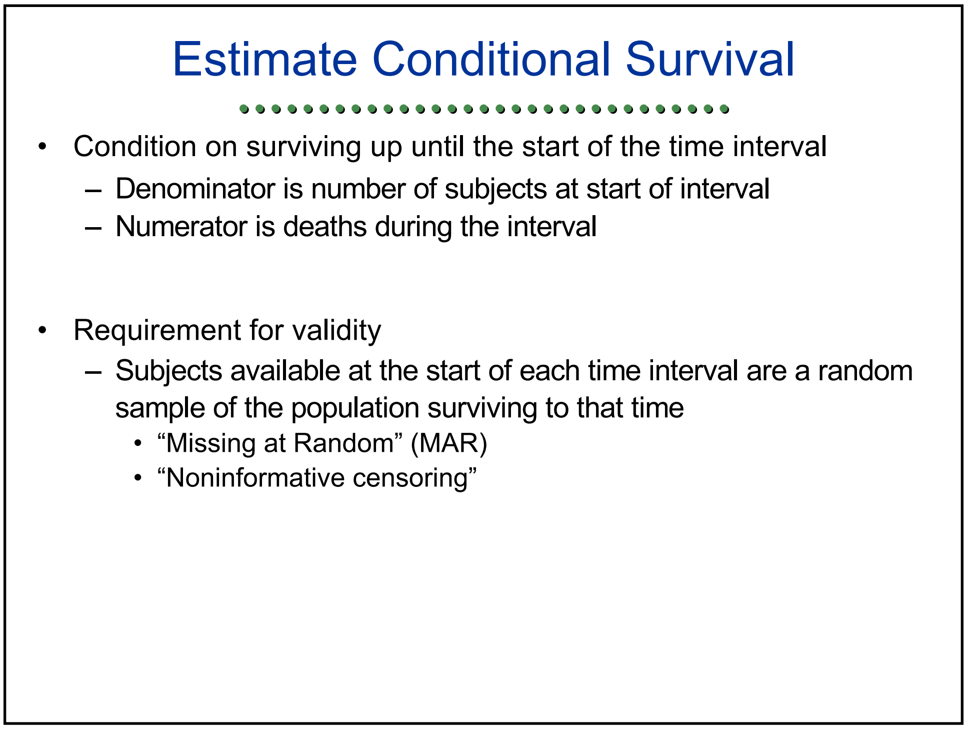
Application of Theory to Survival

- For times $T_1 < T_2$, probability of surviving beyond time T_2 is the product of
  - Probability of surviving beyond time T_2 given survival beyond time T_1, and
  - Probability of surviving beyond time T_1

$$\text{For } t_0 \le t_1 \le t_2 \le \cdots \le t_k$$

$$\begin{aligned}\Pr\left(T^0 \ge t_j\right) &= \Pr\left(T^0 \ge t_j \cap T^0 \ge t_{j-1}\right)\\ &= \Pr\left(T^0 \ge t_j \mid T^0 \ge t_{j-1}\right)\Pr\left(T^0 \ge t_{j-1}\right)\end{aligned}$$

43



Estimate Conditional Survival

- Condition on surviving up until the start of the time interval
  - Denominator is number of subjects at start of interval
  - Numerator is deaths during the interval

- Requirement for validity
  - Subjects available at the start of each time interval are a random sample of the population surviving to that time
    - “Missing at Random” (MAR)
    - “Noninformative censoring”

44

Estimate Survival Probability

- Estimate probability of survival at the endpoint of each time interval

- Multiply the conditional probabilities for all intervals prior to the time point of interest

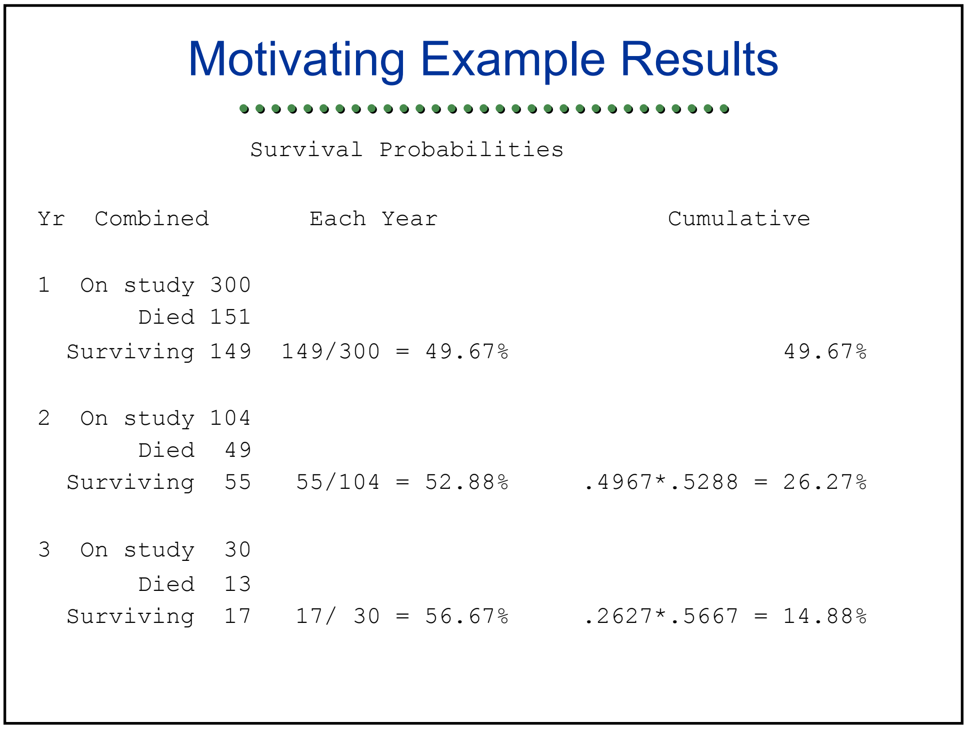
45

Application to Example

- Within interval conditional probabilities
  - Use A, B, C to estimate $Pr\ (T^0 \geq 1)$
  - Use A, B to estimate $Pr\ (T^0 \geq 2 \mid T^0 \geq 1)$
  - Use A to estimate $Pr\ (T^0 \geq 3 \mid T^0 \geq 2)$

- Multiply to obtain unconditional cumulative survival
  - $Pr\ (T^0 \geq 1)$
  - $Pr\ (T^0 \geq 2) = Pr\ (T^0 \geq 2 \mid T^0 \geq 1)\ \ Pr\ (T^0 \geq 1)$
  - $Pr\ (T^0 \geq 3) = Pr\ (T^0 \geq 3 \mid T^0 \geq 2)\ \ Pr(T^0 \geq 2)$

46



Motivating Example Results

```
                Survival Probabilities

Yr  Combined       Each Year                Cumulative

1  On study 300
        Died 151
  Surviving 149  149/300 = 49.67%                      49.67%

2  On study 104
        Died  49
  Surviving  55   55/104 = 52.88%     .4967*.5288 = 26.27%

3  On study  30
        Died  13
  Surviving  17   17/ 30 = 56.67%     .2627*.5667 = 14.88%
```

47