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 7:

Quiz and Discussion

Densities, Cumulative Distribution Functions, Hazard Functions

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

1

Quiz Format

- Occasionally I record a quiz on the material we have covered
  - **Sometimes I include material to motivate upcoming topics**
- I present a few multiple choice questions
  - Successive questions are somewhat related
    - All questions should be answered before discussion
  - During an in-person class, I would allow only about 20-30 seconds for students to record their answers
    - Recording their answers was more for me to see where we stood, rather than to really assign grades
  - We then discuss the answers I would give to the questions
- I strongly urge the same process for this online course
  - Pause the recording for 30 seconds to choose an answer

2

Question 1

We are interested in determining the most common age at death for males and females.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for each sex?

3

3

Question 2

For 60 year olds celebrating their birthday, we are interested in determining the probability that they will die before turning 61.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for each sex?

4

4

Question 3

We are interested in determining the age at which males and females have 50% probability of dying within the next year.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for each sex?

5

5

Question 4

We want to know the first age at which males and females have higher risk of dying than they did the prior year.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for each sex?

6

6

Question 5

We are interested in determining the probability of males and females surviving to receive social security payments at age 65.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for each sex?

7

7

Question 6

We are interested in determining the age range during which males have at least twice the immediate risk of death of females.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for the age range?

8

8

Question 7

We are interested in determining the age at which there are equal numbers of males and females in the US.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for the age?

9

9

Question 8

- Which continent has the lowest highest point, and what is the name of that mountain?

10

10

Answers and Discussion

11

11

Terminology

- **Epidemiology** is particularly interested in describing patterns of disease states and events over space and time
  - Random variables measure the disease state
  - Probabilities for the random variables are further characterized by the time of measurement and/or the time the disease state develops
- **Probability and Statistics** define quantities that will hold for any random variable, not just time or event related

12

12

Epidemiology: Prevalence

There are multiple ways to define probabilities and rates. Here I will use the "frequency" definition by trying to relate the epidemiologic terms to a denominator and a numerator

- **Prevalence** describes for some specified space / time the proportion of individuals having a characteristic of interest
  - (Here I describe point rather than period prevalence)
- Denominator: Number of individuals in specified space / time
  - Time should be a single point, space may be an area
  - Time can be measured by the calendar (January 1, 2024)
  - Time can be measured by an event (date of birth, randomization)
- Numerator: Among individuals counted in the denominator, how many have the particular characteristic
  - E.g., prevalence of rheumatoid arthritis in the U.S. population on January 1, 2024

13

13

Epidemiology: Cumulative Incidence

There are multiple ways to define probabilities and rates. Here I will use the "frequency" definition by trying to relate the epidemiologic terms to a denominator and a numerator

- **Cumulative Incidence** describes for some specified space and interval of time the proportion of individuals newly developing a characteristic of interest
- Denominator: Number of individuals in specified space who are followed for the specific interval of time
  - Time is an interval, space may be an area
  - Time can be measured by the calendar (Jan 1 – Dec 31, 2024)
  - Time can be measured from some event (date of birth, randomization)
- Numerator: Among individuals counted in the denominator, how many develop the particular characteristic over time
  - E.g., incidence of newly diagnosed rheumatoid arthritis in the U.S. population during 2024

14

14

Epidemiology: Incidence Rate

There are multiple ways to define probabilities and rates. Here I will use the "frequency" definition by trying to relate the epidemiologic terms to a denominator and a numerator

- **Incidence rate** describes for some specified space and infinitesimal interval of time the proportion of individuals developing a characteristic of interest
- Denominator: Number of individuals in specified space who are at risk for developing the characteristic (i.e., do not yet have the characteristic) at a specified time
  - Time is a single point (or nearly so), space may be an area
  - Time can be measured by the calendar (January 1, 2024)
  - Time can be measured from an event (birth, randomization)
- Numerator: Among individuals at risk in the denominator, how many develop the particular characteristic at specified time
  - E.g., incidence rate of developing RA on 50th birthday

15

15

Probability and Statistics: Density

There are multiple ways to define probabilities and rates. Here I will use the "frequency" definition by trying to relate the statistics terms to a denominator and a numerator

- **Density** describes for some population of measurements the proportion that are equal to some particular value
- Denominator: Number of individuals in specified population
- Numerator: Among individuals counted in the denominator, how many have specific value
  - E.g., proportion of U.S. population who weigh 200 pounds
  - (Technically a density needs to be integrated over an interval, say 199.5 to 200.5 pounds)

16

16

Probability and Statistics: CDF

There are multiple ways to define probabilities and rates. Here I will use the "frequency" definition by trying to relate the statistics terms to a denominator and a numerator

- **Cumulative distribution function (CDF)** describes for some population of measurements the proportion that are less than or equal to some particular value
- Denominator: Number of individuals in specified population
- Numerator: Among individuals counted in the denominator, how many have a measurement less than or equal to a specific value
  - E.g., proportion of U.S. population who weigh up to 200 pounds
  - (This is obtained by integrating a density from 0 to 200 pounds)
- In biostatistics, we often talk about the **Survival function,** which for continuous variables is just **1 minus the CDF**

17

17

Probability and Statistics: Hazard

There are multiple ways to define probabilities and rates. Here I will use the "frequency" definition by trying to relate the statistics terms to a denominator and a numerator

- **Hazard function** describes for some population of measurements greater than or equal to some particular value the proportion that exactly equal to the particular value
- Denominator: Number of individuals in specified population with measurements greater than or equal to the particular value
- Numerator: Among individuals counted in the denominator, how many have a measurement exactly equal to a specific value
  - E.g., Among the U.S. population who survives to age 65, how many die at age 65 exactly.
  - (We shall find that this is the density divided by 1 minus the CDF)

18

18

Question 1

We are interested in determining the most common age at death for males and females.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for each sex?

19

19

Question 1

We are interested in determining the most common age at death for males and females.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for each sex? **M: 85yrs F: 88 yrs**

20

20

Question 2

For 60 year olds celebrating their birthday, we are interested in determining the probability that they will die before turning 61.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for each sex?

21

21

Question 2

For 60 year olds celebrating their birthday, we are interested in determining the probability that they will die before turning 61.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function (discretized to 1 year)
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for each sex? **M: 1.1% F: 0.7%**

22

22

Question 3

We are interested in determining the age at which males and females have 50% probability of dying within the next year.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for each sex?

23

23

Question 3

We are interested in determining the age at which males and females have 50% probability of dying within the next year.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function. (discretized to 1 year)
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for each sex? **M: 107 yrs F: 109 yrs**

24

24

Question 4

We want to know the first age at which males and females have higher risk of dying than they did the prior year.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for each sex?

25

25

Question 4

We want to know the first age at which males and females have higher risk of dying than they did the prior year.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for each sex? **10 years**

26

26

Question 5

We are interested in determining the probability of males and females surviving to receive social security payments at age 65.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for each sex?

27

27

Question 5

We are interested in determining the probability of males and females surviving to receive social security payments at age 65.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for each sex? **M: 80.3% F: 87.8%**

28

28

Question 6

We are interested in determining the age range during which males have at least twice the immediate risk of death of females.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for the age range?

29

29

Question 6

We are interested in determining the age range during which males have at least twice the immediate risk of death of females.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for the age range? **15 – 32 years**

30

30

Question 7

- We are interested in determining the age at which there are equal numbers of males and females in the US.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for the age?

31

31

Question 7

- We are interested in determining the age at which there are equal numbers of males and females in the US.

a) In epidemiologic terms, this quantity is best related to
   A. Prevalence.
   B. Cumulative incidence.
   C. Incidence rate.
   D. None of the above.

b) In statistical terms, this quantity can best be related to
   A. Hazard function.
   B. Cumulative distribution function.
   C. Density function.
   D. None of the above

c) What is your best guess for the age? **56 years**

32

32

Looking Ahead: Inference

- Most statistical inference is based on summarizing distributions (differences in distributions) by a single number
  - Means, geometric means, proportions, odds, hazards, medians
- It is useful to consider how those summary measures can be visualized when comparing graphs of the entire distribution
  - Density (histogram)
  - Survival function (or cumulative distribution function)
  - Hazard function
- And because we often use ratios (multiplicative scales) also consider graphs with log transformed axes

33

Looking Ahead: Regression Models

- The most commonly used statistical methods for comparing two samples can be viewed as special cases of a regression model
- Relatively distribution-free regression models
  - Linear (robust SE): Diff of means (proportions)
  - Linear on logs (robust SE): Ratio of geometric means
  - Poisson (robust SE): Ratio of means (proportions, rates)
  - Logistic: Odds ratios
  - Proportional hazards: Ratios of (weighted avg) hazards
- Regression models with greater dependence on the distribution
  - Exponential: Ratios of means, quantiles, hzds
  - Weibull: Ratios of quantiles, hazards
  - Accel failure time: Ratios of quantiles

34

Example: 2009 SSA Data

- I will use U.S. Social Security Administration data from 2009
  - Compare mortality / survival of males and females
- We will consider how we can visualize summary measures
  - Modes
  - Means
  - Medians and other quantiles
  - Geometric means
  - Proportions or odds
  - Hazards
- We will consider how we can visualize
  - Multimodality
  - Associations: differences or ratios

35

Detecting Associations with Full Data

- No association
  - Graphs of densities, survival funtions, or hazards will show coincident curves
- Association
  - Curves are different somewhere
- Association in some specific summary measure
  - Curves differ in a particular aspect

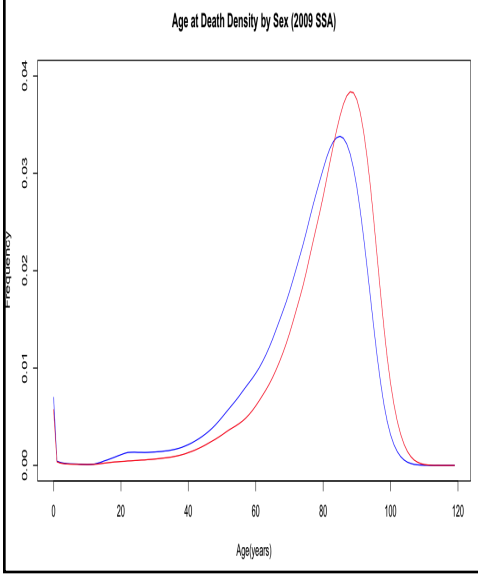
36

Detecting Effect Modification with Full Data

- Effect modification depends on the summary measure used to quantify "effect"
- No effect modification using some specific summary measure
  - Graphs of densities, survival functions, or hazards will have same appearance with respect to the corresponding aspect
- I am just displaying two groups: males (blue), females (red)
  - To talk about effect modification, we would need to further divide those groups according to some other variable
    - E.g., race or ethnicity
  - But we can talk about what similarity of effect would look like as we consider different strata

37

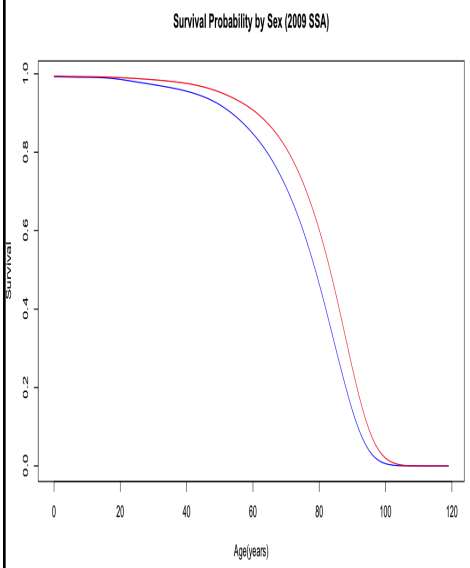
Comparing Densities: Difficult



- **Mode / Multimodality**
  - **Age at highest peak**
- Survival probability at t_0
  - Area to the right of t_0
- *Median (other quantiles)*
  - *Area under curve to the left of median age will equal ½*
- *Means*
  - *Not easily visualized*
- *Hazard*
  - *Density divided by area under curve to the right*

38

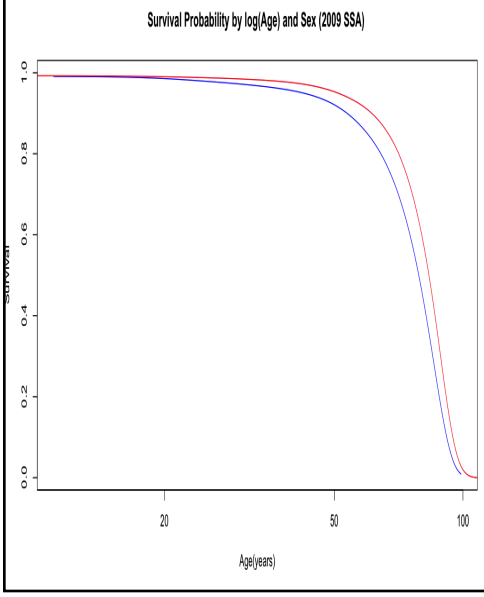
Comparing Survival Curves: S(t) vs t



- Multimodality
  - Change in convexity
- **Difference in survival at t_0**
  - **Vertical separation at t_0**
- **Difference in quantiles**
  - **Horizontal separation at *p***
- **Difference in means**
  - **Area between curves**
- Hazard
  - Slope divided by height of the curve

39

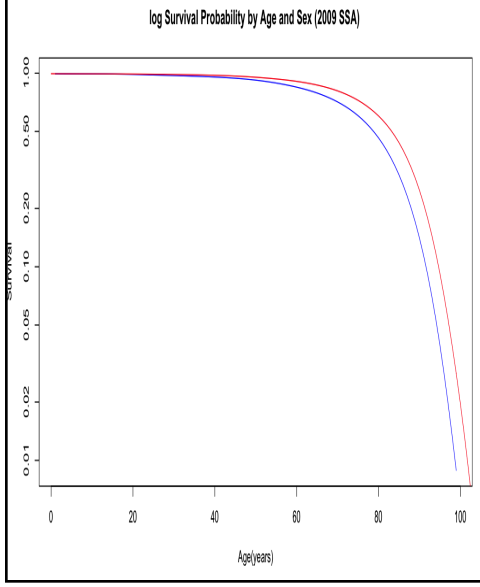
Comparing Survival Curves: S(t) vs log t



- **Difference in survival at t_0**
  - **Vertical separation at t_0**
- **Log ratio of quantiles**
  - **Horizontal separation at *p***
- **Log ratio of geometric means**
  - **Area between curves**
- Hazard
  - Slope divided by height of the curve

40

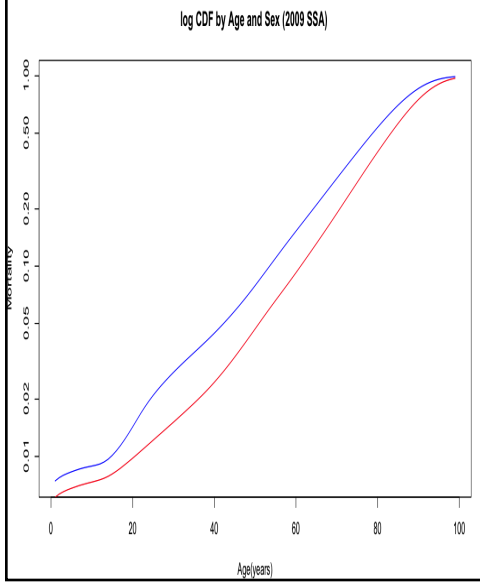
Comparing Survival Curves: log S(t) vs t



- **Log ratio of survival at t_0**
  - **Vertical separation at t_0**

41

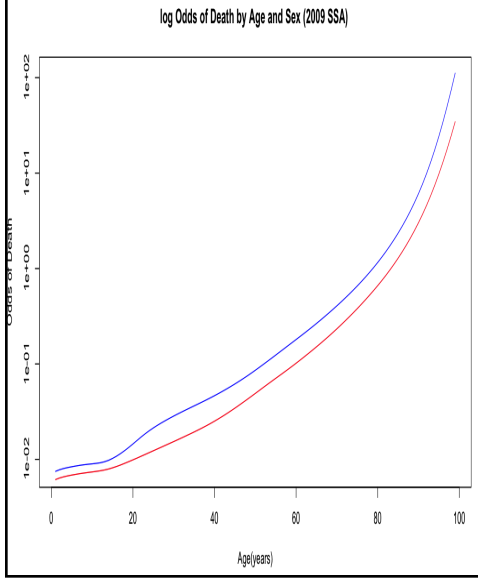
Comparing CDF : log (1 - S(t)) vs t



- **Log ratio of event cumulative probability at t_0**
  - **Vertical separation at t_0**

42

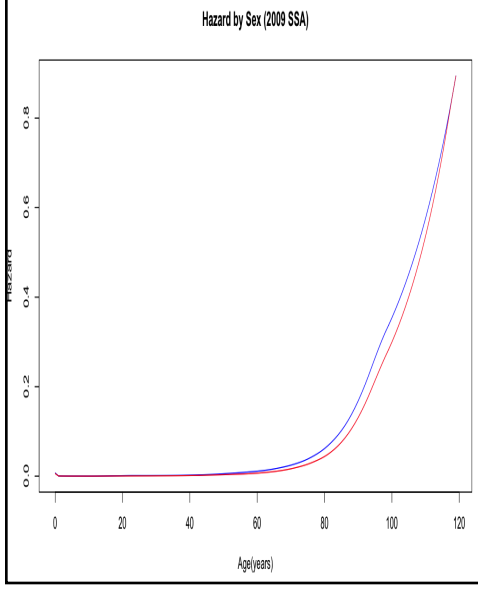
Odds of Death : log [(1-S(t))/S(t)] vs t



- **Log odds ratio of event at t_0**
  - **Vertical separation at t_0**

43

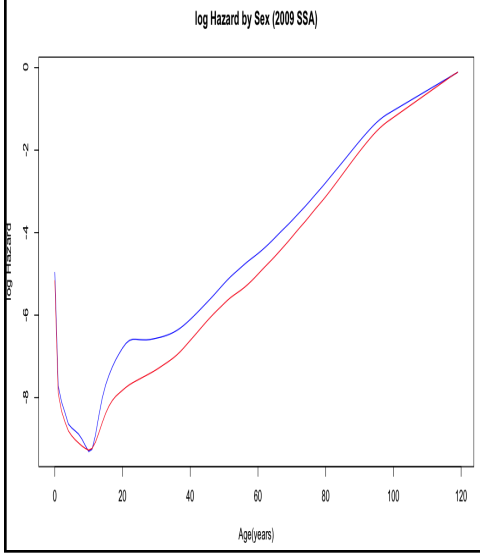
Comparing Hazards: λ(t) vs t



- Difference in survival at t_0
  - Difference in exponentiated area under curve to left of t_o
- **Difference in hazards at t_0**
  - **Vertical separation at t_0**

44

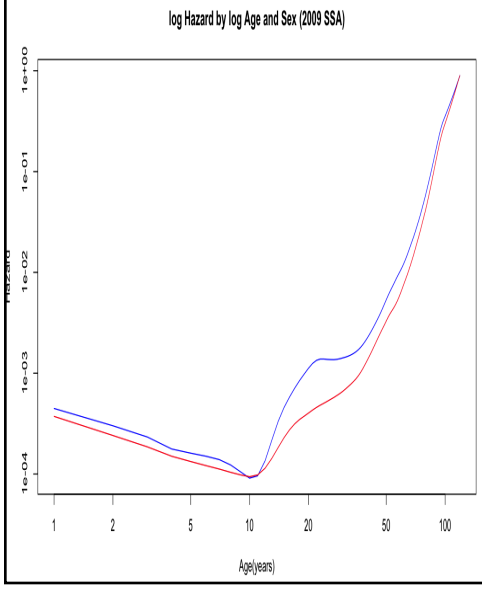
Comparing Hazards: log λ(t) vs t



- **Log hazard ratio at t_0**
  - **Vertical separation at t_0**
    - **PH if parallel**

45

Comparing Hazards: log λ(t) vs log t



- **Shape of distribution**
  - **Weibull is straight line**
    - **PH if parallel**
  - **Exponential is flat line**
- **Log hazard ratio at t_0**
  - **Vertical separation at t_0**
    - **PH if parallel**

46

Question 8

Which continent has the lowest highest point, and what is the name of that mountain?

47

47

Question 8

Which continent has the lowest highest point, and what is the name of that mountain?

Australia:	**Mt. Kosciuszku**	**7,310 ft**
Antarctica:	**Vinson Massif**	**16,050 ft**
Europe:	**Mt. Elbrus**	**18,510 ft**
Africa:	**Mt. Kilimanjaro**	**19,341 ft**
North America:	**Denali (fmr Mt. McKinley)**	**20,310 ft**
South America:	**Aconcagua**	**22,838 ft**
Asia:	**Mt. Everest / Sagarmatha / Chomolungma**	**29,032 ft**

48

48