Biost 517
Applied Biostatistics I

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

First Quiz and Discussion

October 14, 2011

1

Question 1

- For the purposes of **detecting errors** in the data, the most useful descriptive statistic is

A. Mean
B. Median
C. Geometric mean
D. Standard deviation
E. Minimum, maximum
F. 25th and 75th percentiles
G. Histogram

2

Question 2

- For the purposes of describing **materials and methods** for the study, the most useful descriptive statistic to describe the **central tendency (location)** of the data is

A. Mean
B. Median
C. Geometric mean
D. Standard deviation
E. Minimum, maximum
F. 25th and 75th percentiles
G. Histogram

3

Question 3

- For the purposes of describing **materials and methods** for the study, the most useful descriptive statistic to describe the **spread (variability)** of the data is

A. Mean
B. Median
C. Geometric mean
D. Standard deviation
E. Minimum, maximum
F. 25th and 75th percentiles
G. Histogram

4

Question 4

- For the purposes of assessing **the possibility of confounding** in the study, the most useful descriptive statistic to use in **stratified analysis** of the data is

A. Mean
B. Median
C. Geometric mean
D. Standard deviation
E. Minimum, maximum
F. 25th and 75th percentiles
G. Histogram

5

Question 5

- For the purposes of assessing **validity of technical assumptions** for the study, the most useful descriptive statistic to use in **stratified analysis** of the data is

A. Mean
B. Median
C. Geometric mean
D. Standard deviation
E. Minimum, maximum
F. 25th and 75th percentiles
G. Histogram

6

Question 6

- For the purposes of obtaining **preliminary estimates of association** for the study, the most useful descriptive statistic to use in **stratified analysis** of the data is

A. Mean
B. Median
C. Geometric mean
D. Standard deviation
E. Minimum, maximum
F. 25th and 75th percentiles
G. Histogram

7

Question 7

- For the purposes of **exploring effects within subgroups** for the study, the most useful descriptive statistic to use in **stratified analysis** of the data is

A. Mean
B. Median
C. Geometric mean
D. Standard deviation
E. Minimum, maximum
F. 25th and 75th percentiles
G. Histogram

8

Question 8

- Summary statistics for age and blood levels in PBC

	N	Mean	SD	Min	Max
age	418	50.7	10.4	26.3	78.4
bili	418	3.22	4.41	0.3	28.0
albumin	418	3.50	0.42	1.96	4.64
cholest	284	370	232	120	1775

- Which variables might have **substantial outliers**?
  - A. Age
  - B. Bilirubin
  - C. Albumin
  - D. Cholesterol
  - E. Cannot tell

9

Question 9

- A cohort study is conducted in cystic fibrosis patients in order to assess the age at which the patients become colonized with *Pseudomonas.*
- Some patients are noted to already be colonized at their very first clinic visit
- In the resulting data, we would characterize the data on these subjects as
  - A. Missing
  - B. Left censored
  - C. Right censored
  - D. Interval censored
  - E. Any of the above

10

Question 10

- A cohort study is conducted in cystic fibrosis patients in order to assess the age at which the patients become colonized with *Pseudomonas.*
- Some patients are noted to be still not colonized at the time of data analysis
- In the resulting data, we would characterize the data on these subjects as
  - A. Missing
  - B. Left censored
  - C. Right censored
  - D. Interval censored
  - E. Any of the above

11

Question 11

- A cohort study is conducted in cystic fibrosis patients in order to assess the age at which the patients become colonized with *Pseudomonas*.
- Some patients were noted to be free of colonization at one visit, but colonized at their next visit one year later
- In the resulting data, we would characterize the data on these subjects as
  - A. Missing
  - B. Left censored
  - C. Right censored
  - D. Interval censored
  - E. Any of the above

12

Question 12

- A cohort study is conducted in cystic fibrosis patients in order to assess the age at which the patients become colonized with *Pseudomonas.*
- At some visits, the clinical staff forgot to obtain samples to see whether the patients were colonized
- In the resulting data, we would characterize the data for these visits as

A. Missing
B. Left censored
C. Right censored
D. Interval censored
E. Any of the above

13

Question 13

- In the presence of **censored observations**, the **descriptive statistics of interest include**

A. Sample median
B. Sample mean
C. Sample geometric mean
D. Sample standard deviation
E. Sample 25th and 75th Percentiles
F. Sample proportion with observed events
G. All of the above
H. None of the above

14

Question 14

- In the presence of **censored observations**, it is never **possible to estimate population**

A. Median
B. Mean
C. Geometric mean
D. Standard deviation
E. 25th and 75th Percentiles
F. Probability of exceeding thresholds
G. All of the above
H. None of the above

15

Question 15

- A survival study was conducted at three centers. Subjects known to be alive at the end of each year:

	Year 0	Year 1	Year 2	Year 3
Site A	100	60	30	20
Site B	100	40	30	
Site C	100	80		

- The best estimate of the **one year survival** is

A. 40%
B. 60%
C. 80%
D. Impossible to obtain

16

Question 16

- A survival study was conducted at three centers.
  Subjects known to be alive at the end of each year:

	Year 0	Year 1	Year 2	Year 3
Site A	100	60	30	20
Site B	100	40	30	
Site C	100	80		

- The best estimate of the **two year survival** is
  - A. (30 + 30 + 80) / 300 = 46.7%
  - B. (30 + 30) / 200 = 30%
  - C. (180 / 300) x (30 + 30) / 100 = 36%
  - D. Impossible to obtain

17

Question 17

- A survival study was conducted at three centers.
  Subjects known to be alive at the end of each year:

	Year 0	Year 1	Year 2	Year 3
Site A	100	60	30	20
Site B	100	40	30	
Site C	100	80		

- The best estimate of the **three year survival** is
  - A. (20 + 30 + 80) / 300 = 43.3%
  - B. 20 / 100 = 20%
  - C. (180 / 300) x (60 / 100) x (20 / 30) = 24%
  - D. Impossible to obtain

18

Question 18

- In the presence of **censored observations**, it is **usually easiest to estimate population**

  - A. Mean
  - B. Median and other percentiles
  - C. Geometric mean
  - D. Standard deviation
  - E. Probability of exceeding thresholds
  - F. All of the above
  - G. None of the above

19

Survey 19

- What should be the order of the missions of the School of Public Health among (in alphabetical order)
  - Education
  - Public Health Practice
  - Research

  - A. In my opinion it is (list the order)
  - B. I have an opinion, but I don't want to tell you
  - C. I have no opinion

20

Questions 20 and 21

- Write in your answer at the bottom of the answer sheet

20. Which four famous bands from the "British Invasion" of the 1960s were the first to be inducted into the Rock and Roll Hall of Fame?

21. What is the best movie that has ever been made?

21

Answers and Discussion

22

Question 1

- For the purposes of **detecting errors** in the data, the most useful descriptive statistic is

A. Mean
B. Median
C. Geometric mean
D. Standard deviation
E. Minimum, maximum
F. 25th and 75th percentiles
G. Histogram

23

Role of Minima, Maxima

- In monitoring clinical trials, need to distinguish between
  - Adverse experiences that affect nearly everyone
  - Adverse experiences that affect relatively few
- While not quite an error, minima and maxima are also the most important for detecting individual level toxicities
  - Descriptive statistics stratified by treatment group

24

Question 2

- For the purposes of describing **materials and methods** for the study, the most useful descriptive statistic to describe the **central tendency (location)** of the data is

A. Mean
B. Median
C. Geometric mean
D. Standard deviation
E. Minimum, maximum
F. 25th and 75th percentiles
G. Histogram

25

Question 3

- For the purposes of describing **materials and methods** for the study, the most useful descriptive statistic to describe the **spread (variability)** of the data is

A. Mean **(proportion above important thresholds)**
B. Median
C. Geometric mean
D. Standard deviation
E. Minimum, maximum
F. 25th and 75th percentiles
G. Histogram

26

Question 4

- For the purposes of assessing **possibility of confounding** for the study, the most useful descriptive statistic to use in **stratified analysis** of the data is

A. Mean *(if adjust on additive scale)*
B. Median
C. Geometric mean *(if adjust multiplicatively)*
D. Standard deviation
E. Minimum, maximum
F. 25th and 75th percentiles
G. Histogram

27

Question 5

- For the purposes of assessing **validity of technical assumptions** for the study, the most useful descriptive statistic to use in **stratified analysis** of the data is

A. Mean
B. Median
C. Geometric mean
D. Standard deviation
E. Minimum, maximum
F. 25th and 75th percentiles
G. Histogram

28

Question 6

- For the purposes of obtaining **preliminary estimates of association** for the study, the most useful descriptive statistic to use in **stratified analysis** of the data is

A. Mean
B. Median
C. Geometric mean
D. Standard deviation
E. Minimum, maximum
F. 25th and 75th percentiles
G. Histogram

29

Inference for Means

- Most common parameter used as a basis for statistical inference is the mean
  - Proportions = mean of binary variable
  - log Geometric mean = mean of log transformed data
- Tends to reflect a wide variety of differences between distributions
  - E.g., extremely sensitive to changes in the tail of distributions
- Statistical theory allow us to know the sampling distribution, and thus allows us to do inference

30

Question 7

- For the purposes of **exploring effects within subgroups** for the study, the most useful descriptive statistic to use in **stratified analysis** of the data is

A. Mean
B. Median
C. Geometric mean
D. Standard deviation
E. Minimum, maximum
F. 25th and 75th percentiles
G. Histogram

31

Question 8

- Summary statistics for age and blood levels in PBC

	N	Mean	SD	Min	Max
age	418	50.7	10.4	26.3	78.4
bili	418	3.22	4.41	0.3	28.0
albumin	418	3.50	0.42	1.96	4.64
cholest	284	370	232	120	1775

- Which variables might have **substantial outliers**?

A. Age
B. Bilirubin
C. Albumin
D. Cholesterol
E. Cannot tell

32

Question 9

- A cohort study is conducted in cystic fibrosis patients in order to assess the age at which the patients become colonized with *Pseudomonas.*
- Some patients are noted to already be colonized at their very first clinic visit
- In the resulting data, we would characterize the data on these subjects as

A. Missing

B. Left censored

C. Right censored

D. Interval censored

E. Any of the above

33

Question 10

- A cohort study is conducted in cystic fibrosis patients in order to assess the age at which the patients become colonized with *Pseudomonas.*
- Some patients are noted to be still not colonized at the time of data analysis
- In the resulting data, we would characterize the data on these subjects as

A. Missing

B. Left censored

C. Right censored

D. Interval censored

E. Any of the above

34

Question 11

- A cohort study is conducted in cystic fibrosis patients in order to assess the age at which the patients become colonized with *Pseudomonas.*
- Some patients were noted to be free of colonization at one visit, but colonized at their next visit one year later
- In the resulting data, we would characterize the data on these subjects as

A. Missing

B. Left censored

C. Right censored

D. Interval censored

E. Any of the above

35

Question 12

- A cohort study is conducted in cystic fibrosis patients in order to assess the age at which the patients become colonized with *Pseudomonas.*
- At some visits, the clinical staff forgot to obtain samples to see whether the patients were colonized
- In the resulting data, we would characterize the data for these visits as

A. Missing

B. Left censored

C. Right censored

D. Interval censored

E. Any of the above

36

Question 13

- In the presence of **censored observations**, the **descriptive statistics of interest include**

A. Sample median
B. Sample mean
C. Sample geometric mean
D. Sample standard deviation
E. Sample 25th and 75th Percentiles
F. Sample proportion with observed events
G. All of the above
H. None of the above

37

Question 14

- In the presence of **censored observations**, it is never **possible to estimate population**

A. Median
B. Mean
C. Geometric mean
D. Standard deviation
E. 25th and 75th Percentiles
F. Probability of exceeding thresholds
G. All of the above
H. None of the above

38

Question 15

- A survival study was conducted at three centers. Subjects known to be alive at the end of each year:

	Year 0	Year 1	Year 2	Year 3
Site A	100	60	30	20
Site B	100	40	30	
Site C	100	80		

- The best estimate of the **one year survival** is

A. 40%
B. 60%
C. 80%
D. Impossible to obtain

39

Question 16

- A survival study was conducted at three centers. Subjects known to be alive at the end of each year:

	Year 0	Year 1	Year 2	Year 3
Site A	100	60	30	20
Site B	100	40	30	
Site C	100	80		

- The best estimate of the **two year survival** is

A. (30 + 30 + 80) / 300 = 46.7%
B. (30 + 30) / 200 = 30%
C. (180 / 300) x (30 + 30) / 100 = 36%
D. Impossible to obtain

40

Question 17

- A survival study was conducted at three centers. Subjects known to be alive at the end of each year:

	Year 0	Year 1	Year 2	Year 3
Site A	100	60	30	20
Site B	100	40	30	
Site C	100	80		

- The best estimate of the **three year survival** is
  - A. (20 + 30 + 80) / 300 = 43.3%
  - B. 20 / 100 = 20%
  - **C. (180 / 300) x (60 / 100) x (20 / 30) = 24%**
  - D. Impossible to obtain

41

Question 18

- In the presence of **censored observations**, it is **usually easiest to estimate population**
  - A. Mean
  - **B. Median and other percentiles**
  - C. Geometric mean
  - D. Standard deviation
  - **E. Probability of exceeding thresholds**
  - F. All of the above
  - G. None of the above

42

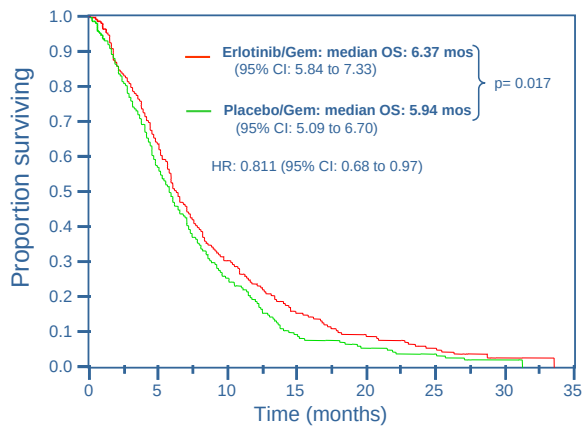
Question 14

- In the presence of **censored observations**, it is never **possible to estimate population**
  - A. Median
  - B. Mean
  - C. Geometric mean
  - D. Standard deviation
  - E. 25th and 75th Percentiles
  - F. Probability of exceeding thresholds
  - G. All of the above
  - **H. None of the above**

43

Kaplan Meier Survival Curve

Erlotinib/Gem vs Placebo/Gem (504 deaths)



* Stratified log-rank test

44

Comparing Survival Curves

- In the presence of **censored observations**, it is never **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)

45

Questions 20 and 21

- Write in your answer at the bottom of the answer sheet

20. Which four famous bands from the "British Invasion" of the 1960s were the first to be inducted into the Rock and Roll Hall of Fame?

    The Beatles, The Rolling Stones, The Kinks, The Who

21. What is the best movie that has ever been made?

    Nashville

46