Biost 536 / Epi 536 Categorical Data Analysis in Epidemiology

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

Lecture 1:
Course Structure;
Setting and One-Sample Studies

September 25, 2014

1

The Use of Statistics to Answer Scientific Questions

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

2

General Philosophy

"Everything should be as simple as possible, but no simpler."

- A. Einstein (paraphrased)

3

Lecture Outline

- Course Structure
- Course Overview
- Categorical Data Setting
- One-Sample Studies
  - Binomial proportions / odds
  - Multinomial proportions
  - Incidence rates
- Two-Sample Studies
  - Comparing proportions
  - Comparing odds
  - Comparing rates

4

Course Structure

5

Course Structure

- Instructor: Scott S. Emerson, M.D., Ph.D.
  - » Fair warning
- TAs:
  - Scott Coggeshall
  - Jean Morrison
- Time and Place:
  - Lectures: 1:30 - 3:20 am TuTh HST 639
  - Data Analysis:
    - 12:30 - 1:20 pm Tu HST 739
    - Alternative TBD

6

Assumed Prior Knowledge

- This course covers advanced categorical analysis including
  - Directly and indirectly standardized rates,
  - logistic, conditional logistic, and Poisson regression
- Material may be covered “out of order” compared to an introductory course
- Equivalent of Biost 515/518 or Biost 513
  - Mantel-Haenszel statistic
  - Logistic regression
  - (Poisson regression)
  - Proportional hazards regression
- Equivalent of Epi 513
- Or permission of instructor

7

Recording of Lectures

- Camtasia
  - Audio and computer video on web
  - Posted approximately 24 hours after class
  - Lecture will typically be posted as “Part A” and “Part B”
- No guarantees: “Mistakes happen”
  - Lecture may be re-recorded for online component of course
  - Recording may differ slightly from what was presented in class

8

Textbooks

- Recommended
  - Hosmer, Lemeshow, and Sturdivant
  - Kleinbaum and Klein
- Other references
  - Agresti (1 introductory text, 2 higher level texts)
  - Fleiss, Levin, and Paik
  - Bishop, Fienberg, and Holland
  - Lachin

9

Computer Software

- Extensively used for data analysis
- Students may use any program that will do what is required, however
  - Stata is used heavily in Biostat 537, 540
    - However, an effort is underway to migrate to R in future years
    - Stata commands will be provided in lecture
  - R is free, but has a harder learning curve
    - Some R functions will be provided on the website
  - Help will presume the use of Stata or R

10

Stata

- Extremely flexible statistical package
  - Interactive
  - Excellent complement of biostatistical methods
- Graphical, report capabilities suboptimal
- Available in microcomputer lab
- Supplementary info on web page
- Syntax introduced in lectures as needed

11

R

- The ultimate in flexible statistical languages
  - Interactive
  - Many user-supplied functions
- Graphical functions generally very good
- Open source and free
- Supplementary info on web page
- Syntax provided (sometimes) as supplemental materials on webpages
- A collection of functions (**`uwIntroStats`**) based on functions that I have written will be made available on the webpages
  - www.emersonstatistics.com/R
  - Caveat emptor

12

Computer Software: Comments

- Designed for people who know statistics, but do not want to write basic functions
- Tries to be all things to all people
  - Much output that you will not want
  - Much output that I will recommend against
- In developing , **`uwIntroStats`** we have tried to provide output that I tend to think best
  - Defaults will correspond to the analyses I tend to look at

13

Guiding Principles

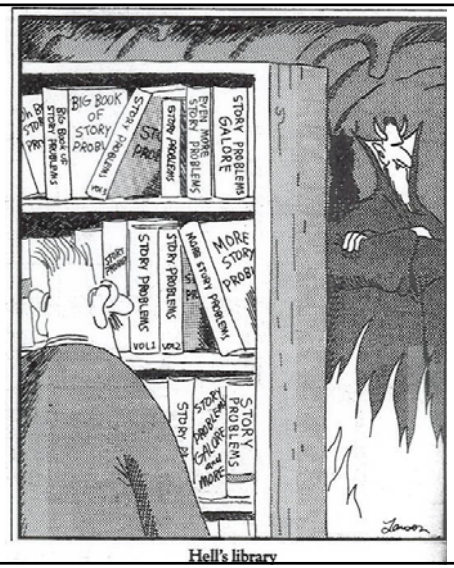
- This is a course in biostatistics, not statistical software
- I will tell you how you can get the statistics I teach you to use
  - There are often multiple ways
  - I tend to teach one of them
- I will not explain every number that appears on the printout

14

Written Homeworks

- Approximately 7 homeworks: analysis of real data
  - Questions directed toward specific analyses
    - But questions will still be stated in as scientific terms (as opposed to statistical) terms as possible
  - Work handed in is expected to be organized scientifically
    - I expect nicely formatted tables, figures
    - Unedited Stata or R output is totally unacceptable
- Homeworks must be submitted:
  - Electronically via Catalyst dropbox
  - On-time (exceptions only in the most dire circumstances)
  - Anonymized MS Word compatible document
    - You must use ID codes supplied for each homework
    - It is your responsibility to ensure no indication of your identity as document author, etc.

15



Hell's library

16

Errors to Avoid

- Unedited Stata / R output is TOTALLY unacceptable
- Any assignments that are handed in should be only your work
- Electronically submitted homeworks should be anonymized
  - Use ID code provided to you
  - Remove identifying information from your files
  - Name file appropriately
    - **Pay VERY CLOSE attention to the formats of the file names**
- Submission of homeworks and grades must be on time.

17

Peer Grading

- Keys to the homeworks will be available on the web pages after the deadline for submission
  - My answers typically go beyond what I expected you to do
  - Extra information will be identified in special fonts
  - You are responsible for any new information that I provide in the homework keys, even if that information is not otherwise presented in class
  - Annotated Stata / R output will often be included
- Each student is expected to use the key to grade another student's paper
  - Double blind: both submitted homework and comments should be anonymous
  - Appeals of grades are decided by TAs and instructor

18

Lectures

- Mix of didactic lectures, class discussion
- Preparatory materials will sometimes be posted on the webpages
  - Videos of didactic lectures posted at least 48 hours in advance
  - Reading materials
- Occasionally the first part of the lecture will be used to
  - Take a brief quiz on the subject matter
    - Basic knowledge / judgment
    - Answers handed in will be graded on good faith effort
  - We will then discuss the reasoning that should be used to answer the quiz questions and current homework assignments
    - Participation in the discussion is required

19

Lab: Discussion Section

- Two types of activities: Simulations and Data analysis
- Simulations
  - Each student will perform a very limited number of simulations
  - In discussion section, we will expand on these results
- Data analysis to answer scientific questions
  - More realistic than that which is given on written homeworks
  - We will discuss the approach to the whole problem
- Requirements:
  - Written solution handed in
    - Graded on "good faith effort"
  - Participation in discussion is required
    - I will often call on students at random
    - It is okay to be wrong, but not okay to be inattentive

20

Class Absences

- We recognize that students are involved in research that often requires attendance at scientific meetings
  - We always try to accommodate such travel
    - (Even when held in Honolulu in November)
  - Please make arrangements in advance so we can handle
    - Peer grading
    - Quizzes
    - Midterms
- We also presume that everyone has a life outside of work
  - Please let us know when illness, family emergencies, etc. requires modification of course schedules for you.
- I am not a big believer in "correspondence courses", but it has worked for students in my classes in the past
  - You definitely need to get my permission if you are planning on rarely attending class

21

Grading

- 20% Homeworks and peer grading (approx 7)
  - Unless modified by TAs or the instructor, the grade assigned by your peer grader is the official grade for your homework
- 10% Quizzes and discussion
- 30% Two Midterms (in class, closed book)
- 20% Data analysis project, report, and oral defense
- 20% Final Exam (in class, closed book)

22

Course Web Pages

- Address: www.emersonstatistics.com/b536/
- Content
  - Syllabus
  - Lecture handouts
  - Recordings of lectures (and discussions if I think it worthwhile)
  - Homework assignments and keys
  - Datasets
  - Supplemental materials not discussed in class
    - Handouts, noteworthy emails

23

Course Content

- Categorical response variable
  - Binary, counts, ordered categorical, unordered categorical
- Sampling schemes
  - Cross-sectional, cohort, case-control
- Summarization of distribution
  - Proportion (mean), odds
  - Difference (proportion), Ratio (proportion, odds)
- One and two sample inference
  - Large sample, small sample (unconditional exact tests)
- Adjusted analyses
  - Stratified analyses
  - Logistic, Poisson, conditional logistic regression
- Prediction

24

Course Overview

25

Categorical Data Analysis…

- Categorical response variables
- Binary indicators of an event
  - Incidence of cancer, prevalence of diabetes, death…
- Counts of events
  - Asthma exacerbations, actinic keratoses, …
- Ordered categorical
  - Stage of cancer, severity of edema, …
- Unordered categorical
  - Findings on screening CT, allelic variants, …

26

…in Epidemiology: Study Design

- Observational studies
  - Surveillance
  - Cross-sectional
  - Cohort
  - Case-control
- Interventional
  - Randomized clinical trials (cohort)

27

…in Epidemiology: Estimand

- Univariate
  - Incidence rate; hazard; “force of mortality”
  - Cumulative incidence; risk
  - Prevalence
- Bivariate
  - Risk difference
    - Attributable risk
  - Relative risk (rate ratio; odds ratio)
    - Relative risk difference
    - Attributable risk percent
  - Population attributable risk
- Adjusted
  - Confounders, precision
- Effect modification

28

Modeling Decisions

- Associations (variable importance) vs prediction
- Proportion vs odds vs (average) hazard
- Difference vs ratio
- Modeling predictor of interest (exposure)
- Modeling effect modification
- Covariates for adjustment: confounders, precision
- Method of adjustment:
  - Stratification (weighting?)
  - Dummy variables
  - Linear
  - Transformed linear
  - Splines / smooths

29

Relevance to Biost 536

- We will need to make many decisions
  - We cannot make an informed decision without understanding the distinctions among our many choices
    - Scientific interpretation
    - Statistical behavior
- The validity of our scientific generalizations will be greatly affected by when we make the decisions
  - Exploratory data-driven analyses
    - Attempts to model the data generation process
    - Prone to high false discovery rate
  - Confirmatory hypothesis driven analyses
    - Attempts to model our scientific question

30

Breslow (*Int. Stat. Rev*, **67**, pp. 252-255)

"As a medical statistician, I am appalled by the large number of irreproducible results published in the medical literature. There is a general, and likely correct, perception that this problem is associated more with statistical, as opposed to laboratory, research. I am convinced, however, that results of clinical and epidemiological investigations could become more reproducible if only the investigators would apply more rigorous statistical thinking and adhere more closely to well established principles of the scientific method. While I agree that the investigative cycle is an iterative process, I believe that it works best when it is hypothesis driven."

31

Breslow (*Int. Stat. Rev*, **67**, pp. 252-255)

"The epidemiology literature is replete with irreproducible results stemming from the failure to clearly distinguish between analyses that were specified in the protocol and that test the a priori hypotheses whose specification was needed to secure funding, and those that were performed post-hoc as part of a serendipitous process of data exploration."

32

What is the Estimand?

- In every data analysis, we will first need to make sure we know
  - What we are trying to estimate
  - Why we are trying to estimate it
    - (What will we do with the answer?)
  - What is the "burden of proof"?
- "Estimand": What we are trying to estimate?
  - Scientifically
    - Population, time frame, clinical measure
    - Prediction vs association vs effect modification
  - Statistically
    - Summary measure
    - Measure used for comparison
    - Covariate adjustment
    - (Handling of missing data)

33

Why a Special Biostatistics Course?

- "Everything is regression."
  - Scott Emerson (as quoted by a Biost 570 in 1997)
- In Biost 518 I stress the "general regression model"
  - Some summary measure (mean, geometric mean, odds, …)
  - A link function (additive vs multiplicative models)
  - A "linear predictor" of "predictof of interest" (POI), effect modifiers, confounders, precision variables
- What are the special features of categorical data?
  - Mean-variance relationship: e.g., Mean= p Variance= $p(1-p)$
  - Discreteness: e.g., Y= 0 or 1
  - Restricted range on parameter: e.g., $0 \leq p \leq 1$
  - Really: the combination of the three
    - Often estimate zero variance in "sparse" data
- Other issues related to regression are not really different

34

Overview of Setting

Scientific Method

35

Purpose of Statistics

- Statistics is about science
  - (Science in the broadest sense of the word)
- Science is about proving things to people
  - (The validity of any proof rests solely on the willingness of the audience to believe it)
- It is rare that a scientific question is answered in one study
- Instead, we always have a progression of studies
  - Hypothesis generating studies
  - Pilot / screening studies
  - Confirmatory studies

36

First Stage of Scientific Investigation

- Hypothesis generation
- Observation
- Measurement of existing populations
- Disadvantages:
  - Confounding
  - Limited ability to establish cause and effect

37

Further Stages of Scientific Investigation

- Refinement and confirmation of hypotheses
- Experiment
  - Ideally: an interventional study
  - At least: a designed observational study in an independent setting
- Elements of experiment
  - Overall goal
  - Specific aims (hypotheses)
  - Materials and methods
  - Collection of data
  - Analysis
  - Interpretation; Refinement of hypotheses

38

Do You Need Statistics?

- Two question test (Both must be YES)
- In a deterministic world, do YOU know how to answer your question?
  - Is the question answerable in the real world?
  - How do you use a number to answer the scientific question?
- In a world subject to variation, do YOU know how you would answer your question if you had the entire population?

39

Statistical Tasks

- Understand overall goal
- Refine specific aims (stat hypotheses)
- Materials and methods: Study design
- Collection of data: Advise on QC
- Analysis
  - Describe sample (materials and methods)
  - Analyses to address specific aims
- Interpretation

40

Overview of Setting

Statistical Hypotheses

41

Statistical Questions

- Clustering of observations
  - E.g, identifying subcategories of disease
- Clustering of variables
  - E.g, identifying genetic pathways
- Quantification of distributions
  - Estimating mean, median, …
- Comparing distributions
  - Existence of an association; direction of "first order trend"
  - Linearity / nonlinearity of "effect" vs exact "dose-response"
  - Effect modification
- Prediction of individual observations
  - Binary (classification): diagnosis, prognosis
  - Continuous: normal ranges, surrogate measurements

42

Refining Scientific Hypotheses

- Statistical hypotheses precisely define
  - the intervention (or risk factor)
  - the outcome
    - advise on precision of measurement
  - the target population(s)
    - covariates
  - "tend to" (the standards for comparison)
    - summary measures
    - relevance of absolute or relative standards

43

Statistical Role of Variables

- "Response" or "Outcome"
  - Can be either the "effect" or the "cause"
- "Grouping Variable(s)"
  - Primary scientific question
    - Predictor of interest (POI)
    - Effect Modifiers
      - Target of inference vs reproducible summarization
  - Adjustment for covariates
    - Confounders
    - Precision variables

44

Classification of Methods

- One sample studies
  - Grouping variable is constant
  - No covariates
- Two sample studies
  - Grouping variable is binary
  - No covariates
- *K* sample studies
  - Grouping variable is (treated as) unordered categorical
  - No covariates (though dummy variable for predictor of interest)
- Regression studies
  - Grouping variable can be continuous
  - Possible modeling of effect modifiers
  - Possible adjustment for confounders and precision variables

45

Quiz
(Pre-test and Survey)

U.S. Mortality Statistics by Sex

46

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?.

47

Question 2

- 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?.

48

Question 3

- We are interested in determining the 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?.

49

Question 4

- 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?.

50

Question 5

- 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?.

51

Question 6

- 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?.

52

Question 7

- Who was the University of Washington President who traveled to New England in order to entice young unmarried women to move to Seattle?

53

Categorical Data Analysis in Epidemiology

Where are we going?

Much of epidemiologic research revolves around disease surveillance and identification of risk factors for incidence of disease or mortality.

Statistically, “disease” and “mortality are almost always binary variables, and we often dichotomize exposure, as well.

54

Categorical Data Analysis…

- Categorical response variables
- **Binary indicators of an event**
  - **Incidence of cancer, prevalence of diabetes, death…**
- Counts of events
  - Asthma exacerbations, actinic keratoses, …
- Ordered categorical
  - Stage of cancer, severity of edema, …
- Unordered categorical
  - Findings on screening CT, allelic variants, …

55

…in Epidemiology: Study Design

- **Observational studies**
  - **Surveillance**
  - **Cross-sectional**
  - **Cohort**
  - **Case-control**
- Interventional
  - Randomized clinical trials (cohort)

56

…in Epidemiology: Estimand

- **Univariate**
  - **Incidence rate; hazard; "force of mortality"**
  - **Cumulative incidence; risk**
  - **Prevalence**
- **Bivariate**
  - **Risk difference**
    - **Attributable risk**
  - **Relative risk (rate ratio; odds ratio)**
    - **Relative risk difference**
    - **Attributable risk percent**
  - **Population attributable risk**
- Adjusted
  - Confounders, precision
- Effect modification

57

Epidemiologic Estimands

Incidence Rates, Incidence, Prevalence

58

Binary Data in Epidemiology

- Focus of epidemiologic research (in broad categories)
  - Disease surveillance: morbidity, mortality
  - Identifying risk factors or protective factors for disease / death
  - Evaluating preventive strategies and treatments
- Scientific classification of binary data
  - "Clinical outcomes"
    - Diagnosis of disease, cure, disease progression, death
  - "Exposures" sometimes dichotomized
    - Genotypes, environmental exposures, behavioral exposures
  - Indicators of risk status
    - Time, place, population

59

Risk Sets

- Most often, we recognize that the probability of an event depends in some way upon time
- In many cases, that time dependence is something we merely want to adjust for as we compare different groups
  - It is not as important to contrast the event probability over time
- Examples
  - Comparing rates of cancer between smokers and nonsmokers, we know cancer occurs more often late in life
  - Comparing exposure groups for premature delivery, we know that probability of delivery increases as pregnancy continues
- We thus find it convenient to couch many of our analyses of binary data in terms that also consider "time to event"

60

Incidence and Mortality Rates (Hazards)

- We are often interested in the rate (over time) at which individuals convert from being "event-free" to having had the event
  - Time can be calendar time, age, study time …
  - (They differ in what we call time zero)
- At each point in time, we essentially compute a proportion
  - Denominator: Individuals who are currently "event-free"
  - Numerator: Among those in the denominator, who converts in the next instant
- Referred to as
  - Epidemiology: incidence and mortality rates, force of mortality
  - Statistics and probability: hazard function

61

Hazard Function Notation

- For each individual in some group of interest, T measures the time the event will occur
  - $Y(t)$ is thus an indicator that the event has occurred prior to t
  - T might be infinity

Hazard function (continuous T): for very small Δt

$$\lambda(t) = \Pr(t \le T < t + \Delta t \mid t \le T)$$
$$= \frac{\Pr(t \le T < t + \Delta t)}{\Pr(t \le T)} = \frac{f(t)}{1 - F(t)}$$

$F(t)$ is cumulative distribution function

$f(t)$ is density

62

Hazard Rate Based Inference

- When the changing conversion rate is just a nuisance to our primary question, we still have to worry that time might be
  - An effect modifier and/or
  - A confounder and/or
  - A precision variable.
- Most often we choose some way to adjust for those roles by
  - Using weighted averages of the hazard (e.g., standardized rates)
  - Adjusting in a regression model
    - Poisson models adjusting for person-time at risk
    - Proportional hazards regression models
    - Parametric regression models

63

(Cumulative) Incidence and Mortality

- Sometimes we choose a specific interval of time of greatest interest
  - E.g., incidence of cancer within one year, teenage mortality
- Usually estimated with a simple proportion
  - Denominator: Individuals who are "event-free" at time *a*
  - Numerator: Individuals experiencing event between *a* and *b*
- It does relate to the hazard

(Cumulative) incidence between times a and b

$$\Pr(a \le T < b \mid a \le T) = 1 - e^{-\int_a^b \lambda(u)\,du}$$

64

(Cumulative) Incidence Based Inference

- Note that if the hazard function is (nearly) constant over some small period of time then

(Cumulative) incidence between times a and b

$$\Pr(a \le T < b \mid a \le T) = 1 - e^{-\int_a^b \lambda(u)\,du} = 1 - e^{-\int_a^b \lambda\,du} = 1 - e^{-\lambda(b-a)}$$

- This "piecewise exponential" model is often used as a basis for inference
  - The "exponential distribution" has a constant hazard

65

Prevalence

- Sometimes we are interested in how many affected individuals there are in a population at a given point in time
  - Sometime "point prevalence" relative to some event (e.g., birth)
- Most often estimated by a proportion
  - Denominator: Individuals still alive in the population at time *t*
  - Numerator: Individuals still alive and have had the event prior to time *t*
- Note the dependence on
  - The hazard rate, and
  - The case-fatality rate (for this disease and competing risks)

66

Answers and Discussion

67

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?.

68

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**

69

Question 2

- 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?.

70

Question 2

- 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?. **M: 107 yrs F: 109 yrs**

71

Question 3

- We are interested in determining the 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?.

72

Question 3

- We are interested in determining the 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**

73

Question 4

- 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?.

74

Question 4

- 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%**

75

Question 5

- 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?.

76

Question 5

- 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**

77

Question 6

- 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?.

78

Question 6

- 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**

79

Question 7

- Who was the University of Washington President who traveled to New England in order to entice young unmarried women to move to Seattle?

80

Question 7

- Who was the University of Washington President who traveled to New England in order to entice young unmarried women to move to Seattle?

Asa Mercer and his brothers cleared land for UW in 1861.

Asa Mercer was the only college graduate in town, so became the university president.

In 1864, he recruited the "Mercer girls" to Seattle.

What was the 1960s TV show based on this story?

81

One Sample Inference for Binomial Proportions

Large Samples

(Uncensored)

(reF: Lecture notes and from Biost 517, lecture 11 and recordings from Nov 9, 2012)

82

Binary Random Variables

- Many variables can take on two values
  - For convenience code as 0 or 1
    - Vital status: "Dead" 0 is (alive) or 1 (dead)
    - Sex: "Female" is 0 (male) or 1 (female)
    - Intervention: "Tx" is 0 (control) or 1 (new therapy)
- Sometimes dichotomize variables
  - For scientific reasons (statistically less precise)
    - Blood pressure less than 160 mm Hg
    - PSA less than 4 ng/ml
    - Serum glucose less than 120 mg/dl

83

Statistical Hypotheses

- Binary variable has Bernoulli (binomial) distn
  - p is the proportion of the population with the random variable equal to 1
  - p is also the population mean for the random variable
- Scientific questions often translated into a statistical question about parameter p
- Equivalently, we can phrase statistical question about the odds parameter $o = p / (1 - p)$
  - We can easily transform estimates (point and interval) about proportion p to odds o

84

Exact Distribution

- Here, we do not have to rely on asymptotic (large sample) theory
- A binary variable must follow the Bernoulli distribution: $Y_i \sim B(1,p)$
  - *"I could have been a famous singer if I had someone else's voice" – Road to Joy, Conor Oberst (Bright Eyes)*
- Sums of independent Bernoulli random variables must be binomial: $S = Y_1 + \ldots + Y_n \sim B(n,p)$
- We can use the exact binomial distribution to compute our probabilities
  - (Well, computers can)

85

Binomial Distribution

- Probability theory provides a formula for the distribution of binomial random variables

$$\text{Data } Y_1,\ldots,Y_n \overset{\text{iid}}{\sim} B(1,p)$$

$$\Downarrow$$

$$S=\sum_{i=1}^{n} Y_i = Y_1+\cdots+Y_n \sim B(n,p)$$

$$\text{For } k=0,1,\ldots,n: \quad \Pr(S=k)=\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

86

Exact Point Estimate Using MLE

- The sample mean is the "maximum likelihood estimate" (MLE)

$$\text{Data } Y_1,\ldots,Y_n \overset{\text{iid}}{\sim} B(1,p) \quad E(Y_i)=p \quad Var(Y_i)=p(1-p)$$

$$\text{Point estimate:} \qquad \hat{p}=\bar{Y}=\frac{1}{n}\sum_{i=1}^{n} Y_i=\frac{Y_1+\cdots+Y_n}{n}$$

- In a wide variety of settings, including binomial, normal, Poisson, exponential distributions, we can show
  - MLE of mean is unbiased: tend to the true value on average
  - MLEs are consistent: tend to the true value for large *n*
  - MLEs are (asymptotically) efficient: smallest variance possible

87

Hypothesis Testing: General Approach

- Is data too unusual for some hypothesized value?
  - Too high? Too low? Too extreme?

For independent random variables $Y_1, Y_2, \ldots Y_n \sim F(\theta)$

Observe data $Y_1 = y_1, \ldots, Y_n = y_n$

Compute statistic $T = T(y_1,\ldots,y_n) = t$

(often use estimate $T=\hat{\theta}$)

Test $H_0: \theta = \theta_0$, calculate P values by

$$\text{Upper one-sided:} \quad P_{upper} = \Pr[T(Y_1,\ldots,Y_n) \geq t;\ \theta=\theta_0]$$

$$\text{Lower one-sided:} \quad P_{lower} = \Pr[T(Y_1,\ldots,Y_n) \leq t;\ \theta=\theta_0]$$

$$\text{Two-sided (easy):} \quad 2\times\min(P_{lower}, P_{upper}, 0.5)$$

88

Exact Tests for a Proportion

- Use binomial distribution under the null
  - (But let a computer do it for you)

For $S \sim B(n, p)$ and observation $S = k$:

Test $H_0 : p = p_0$, calculate P values by

$$\text{Upper one-sided:} \quad P_{upper} = \Pr[S \geq k; p_0] = \sum_{i=k}^{n} \frac{n!}{i!(n-i)!} p_0^{\,i} (1-p_0)^{n-i}$$

$$\text{Lower one-sided:} \quad P_{lower} = \Pr[S \leq k; p_0] = \sum_{i=0}^{k} \frac{n!}{i!(n-i)!} p_0^{\,i} (1-p_0)^{n-i}$$

$$\text{Two-sided (easy):} \quad 2 \times \min(P_{lower}, P_{upper}, 0.5)$$

89

Stata: Tests for Proportion

- Syntax
  - `"bitest var = #p"`
- Provides exact test that proportion = #p
- Gives upper and lower one-sided, two-sided P values
  - Two-sided P value is computed under a slightly more complicated rule, but is valid (and better than "easy" approach)
- Note: The p value for a test of H_0: $p = p_0$ is also the p value for a test of H_0: $o = o_0 = p_0 / (1 - p_0)$

90

Confidence Interval: General Approach

- For what parameter values would we regard the data "typical"?
  - (CI will "cover" true value of parameter with desired probability)

For independent random variables $Y_1, Y_2, \ldots Y_n \sim F(\theta)$

Observe data $Y_1 = y_1, \ldots, Y_n = y_n$

Compute statistic $T = T(y_1, \ldots, y_n) = t$

(often use estimate $T = \hat{\theta}$)

Compute $100(1-\alpha)\%$ confidence intervals by

$$\text{Upper bound:} \quad CI_u(t) = \left\{ \theta^* : \alpha \leq \Pr[T(Y_1, \ldots, Y_n) \leq t;\ \theta = \theta^*] \right\}$$

$$\text{Lower bound:} \quad CI_l(t) = \left\{ \theta^* : \Pr[T(Y_1, \ldots, Y_n) \leq t;\ \theta = \theta^*] \leq 1-\alpha \right\}$$

$$\text{Two-sided CI:} \quad CI_2(t) = \left\{ \theta^* : \frac{\alpha}{2} \leq \Pr[T(Y_1, \ldots, Y_n) \leq t;\ \theta = \theta^*] \leq 1-\frac{\alpha}{2} \right\}$$

91

Exact Confidence Interval for Proportions

- Use the binomial distribution
  - (But let a computer do it for you)

An exact $100(1-\alpha)\%$ confidence interval for p based on observation $S = k$ is $(\hat{p}_L, \hat{p}_U)$ where an iterative search is used to find

$$\Pr[Y \leq k;\ p = \hat{p}_U] = \sum_{i=0}^{k} \frac{n!}{i!(n-i)!} \hat{p}_U^{\,i} (1-\hat{p}_U)^{n-i} = \alpha/2$$

$$\Pr[Y \geq k;\ p = \hat{p}_L] = \sum_{i=k}^{n} \frac{n!}{i!(n-i)!} \hat{p}_L^{\,i} (1-\hat{p}_L)^{n-i} = \alpha/2$$

92

Stata: Exact CI for Proportion

- Syntax
  - "`ci varlist, binomial`"
    - Provides exact confidence intervals
    - (Standard errors are based on asymptotics)
- Note: The CI for a proportion *p* can be transformed to a CI for odds *o:* CI is $(o_l, o_u) = (p_l / (1-p_l), p_u / (1-p_u))$

93

Role of Approximate Distributions

- In one sample problems, statistical inference based on the exact binomial distribution is the best thing to use
  - It is extremely nice of Stata to (usually) do that for you
- When we get to two sample problems and adjusted analyses, we most often use asymptotic (large sample) approximations
  - When *n* is large, *S* is approximately normal with mean *np* and variance *np(1-p)*
  - When *n* is large and *p* is small, *S* is approximately Poisson with mean *np* and variance *np*
- It is useful to consider the accuracy of these approximations in one sample problems in order to explore issues with
  - the discrete nature of the data, and
  - the mean-variance relationship

94

Point Estimate

- Use the sample mean

$$\text{Data } Y_1,\ldots,Y_n \overset{\text{iid}}{\sim} B(1,p) \quad E(Y_i)=p \quad Var(Y_i)=p(1-p)$$

$$\text{Point estimate:} \qquad \hat{p}=\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i=\frac{Y_1+\cdots+Y_n}{n}$$

95

Approximate Distribution

- Use the central limit theorem

$$\text{Data } Y_1,\ldots,Y_n \overset{\text{iid}}{\sim} B(1,p) \quad E(Y_i)=p \quad Var(Y_i)=p(1-p)$$

$$\hat{p}=\bar{Y} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

- NOTE: Unlike in the normal distribution, with Bernoulli data
  - the sample mean can only be one of *n+1* different values
  - there is a mean – variance relationship

96

Discreteness: Use Continuity Correction

- In one sample problem we often make a continuity correction in order to handle the discreteness

$$\Pr\left(\hat{p} \le \frac{k}{n}\right) = \Pr\left(\hat{p} \le \frac{k+0.5}{n}\right)$$

$$\Pr\left(\hat{p} \ge \frac{k}{n}\right) = \Pr\left(\hat{p} \ge \frac{k-0.5}{n}\right)$$

97

Mean-Variance: Alternative Approaches

- When using MLEs and likelihood methods, there are three statistics we commonly use
- Wald statistic uses asymptotic distribution of the MLE, plugging the MLE into the formula for the variance
- Score statistic uses asymptotic distribution of the MLE, plugging a hypothesized value of the parameter into the formula for the variance
- Likelihood ratio (LR) statistic compares the asymptotic distribution of the MLE with the plug-in estimator to the asymptotic distribution using the hypothesized value of the parameter

98

CI Through Inverted Statistics

- With Wald intervals, the CI will often look like

  (estimate) ± (crit value) × (std error)

- With score or likelihood ratio intervals, we search for the hypothesized values that would be rejected with the corresponding test statistic
  - A computer can do this easily, though most programs do not
- In the absence of a mean-variance relationship, all three intervals may be the same
- In a one-sample problem with a mean-variance relationship, the score and LR intervals often are most often the same.

99

Asymptotic CI: Wald statistic

- Often we can just use best estimate of *p* in standard error for confidence intervals and ignore the continuity correction
  - *np* and *n(1-p)* must be large

$$100(1-\alpha)\%\ \text{CI for } p: \qquad \hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

100

Asymp CI: Wald w/ Continuity Correction

- We can add a continuity correction to the Wald interval

$100(1-\alpha)\%$ CI for p: $(\hat{p}_L, \hat{p}_U)$

$$\hat{p}_L = \hat{p} - \frac{1}{2n} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\hat{p}_U = \hat{p} + \frac{1}{2n} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

101

Asymptotic CI: Score Statistic

- We consider mean-variance relationship and continuity correction
  - Requires quadratic formula or iterative search
  - (Quadratic formula can be easily implemented in Excel, etc.)

$100(1-\alpha)\%$ CI for p: $(\hat{p}_L, \hat{p}_U)$

$$\hat{p}_L = \hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}_L(1-\hat{p}_L)}{n}}$$

$$\hat{p}_U = \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}_U(1-\hat{p}_U)}{n}}$$

102

Asymptotic CI: Best Approach

- We do best by considering mean-variance relationship and continuity correction
  - Requires quadratic formula or iterative search
  - (Quadratic formula can be easily implemented in Excel, etc.)

$100(1-\alpha)\%$ CI for p: $(\hat{p}_L, \hat{p}_U)$

$$\hat{p}_L = \hat{p} - \frac{1}{2n} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}_L(1-\hat{p}_L)}{n}}$$

$$\hat{p}_U = \hat{p} + \frac{1}{2n} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}_U(1-\hat{p}_U)}{n}}$$

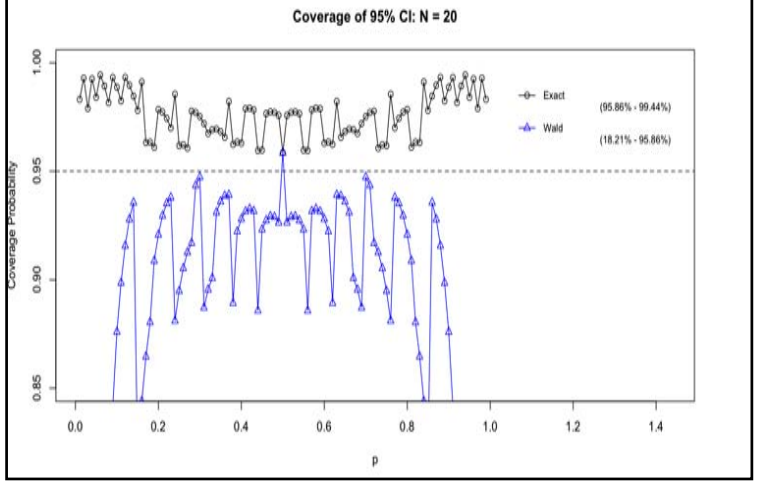
103

Evaluating Methods

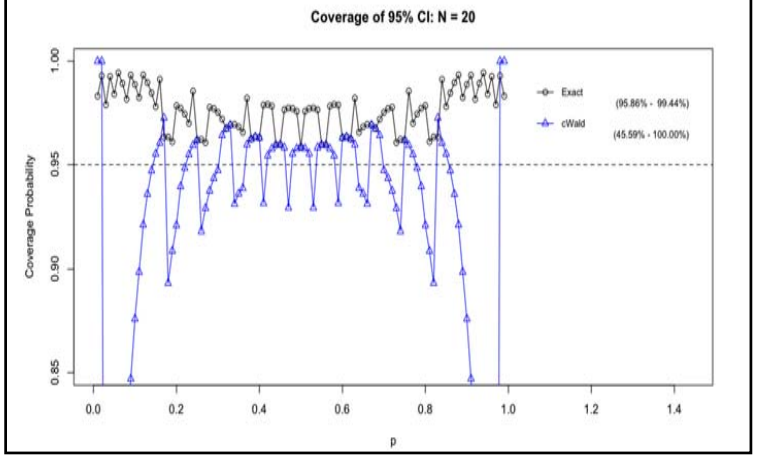
- We can compare how the various CI behave in small samples
  - E.g., 95% of CI should include the true value of the parameter *p*
- We consider coverage probabilities of CI as a function of
  - Nominal confidence level
  - True value of *p*
  - Sample size

104

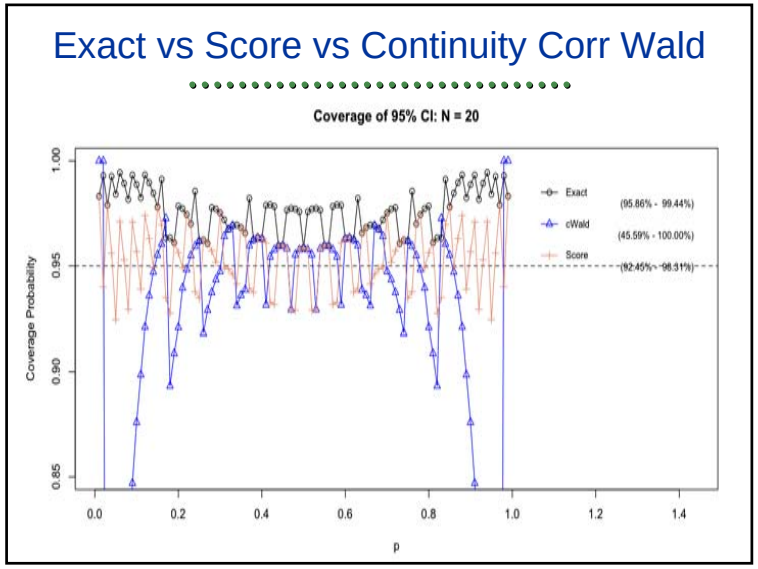
Exact vs Wald



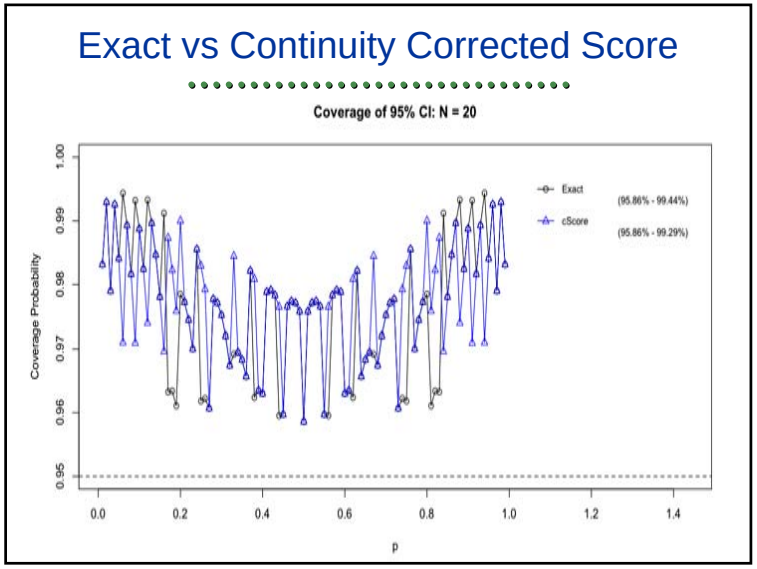
Exact vs Continuity Corrected Wald



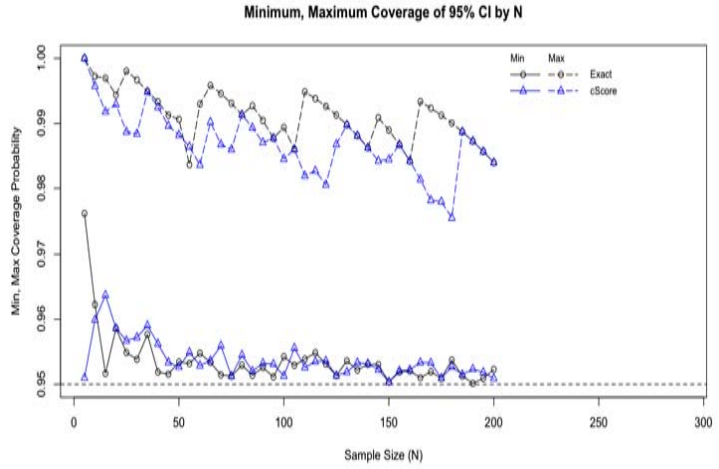
Exact vs Score vs Continuity Corr Wald



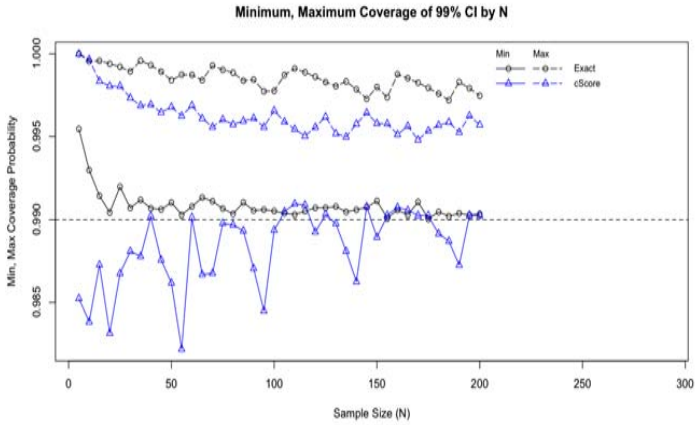
Exact vs Continuity Corrected Score



As Sample Size Varies 5 – 200 by 5



Coverage of 99% CI by n



Special Case: 0 (or n) Events

- The Wald based CI fail badly when you have observed 0 (or n) events
  - Estimated proportion is 0 (or 1), so variance estimate is 0
- Exact CI are always an option
- But we can use a score based interval
  - One-sided 95% upper bound ➔ "Rule of 3"
  - One-sided 97.5% upper bound ➔ "Rule of 3.69"
    - Corresponds better to a two-sided 95% CI

111

Upper Conf Bnd for 0 Events

- Exact upper confidence bound when all observations are 0

Suppose $Y \sim B(n, p)$ and $Y = 0$ is observed

Exact $100(1-\alpha)\%$ upper confidence bound for p is $\hat{p}_U$

$$\Pr[Y = 0; \hat{p}_U] = (1 - \hat{p}_U)^n = \alpha$$

$$\Downarrow$$

$$\hat{p}_U = 1 - \alpha^{1/n}$$

112

Large Sample Approximation

$$(1-\hat{p}_U)^n = \alpha \quad \Rightarrow \quad n\log(1-\hat{p}_U) = \log(\alpha)$$

For small $\hat{p}_U$ $\qquad \log(1-\hat{p}_U) \approx -\hat{p}_U$

so for large n $\quad \Rightarrow \quad \hat{p}_U \approx -\dfrac{\log(\alpha)}{n}$

113

Elevator Stats: 0 Events in n trials

- "Three over n rule"
  - log (.05) = -2.9957
  - In large samples, when 0 events observed, the 95% upper confidence bound for p is approximately 3 / n
- "Three over n rule"
  - log (.025) = -3.688879
  - In large samples, when 0 events observed, the 97.5% upper confidence bound for p is approximately 3.69 / n
- 99% upper confidence bound
  - log (.01) = -4.605
  - Use 4.6 / n as 99% upper confidence bound

114

Take Home Message

- Exact methods are conservative due to discreteness
- At 95% confidence level, asymptotic methods behave well even at very small sample size *(normal distribution of estimates is not the issue)* PROVIDING you
  - Address mean-variance relationship (most important)
  - Address discreteness (also pretty important in one sample case)
- If you want good behavior in extreme tails (e.g., 99% confidence), a larger sample size is needed
  - This is highly relevant in genomics *p* >> *n* problems
- LOOKING AHEAD: Two sample problems, covariate adjustment
  - We will continue to stress the mean-variance relationship
  - However, continuity corrections are generally not advised
    - The discreteness in two groups tends to cancel each other out

115

One Sample Inference for Binomial Odds

Large Samples

(Uncensored)

(reF: Lecture notes and from Biost 517, lecture 11 and recordings from Nov 9, 2012)

116

Odds

- Odds of being in group of interest:
  - Definition
    - Dichotomize observations as with proportion
    - Ratio of proportion to 1 minus proportion

$$\hat{p} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$\hat{o} = \frac{\hat{p}}{1-\hat{p}}$$

117

Odds: Uses

- As with proportions, except:
- Odds is less easily understood by a lay person
  - Odds of rolling 1 on a die: 1/5
- However, odds (actually log odds) is often more convenient in modeling due to greater range of possible values
  - proportion is between 0 and 1
  - odds is between 0 and infinity
  - log odds is any real number

118

Inference with Odds

- Odds are a "monotonic transformation" of p
  - $p_1 > p_2$ ➔ $o_1 = p_1 / (1 - p_1) > o_2 = p_2 / (1 - p_2)$
- Thus the most straightforward way to get CI for odds is to
  - Get CI for p: (p_L, p_U)
  - Transform to the CI for odds: $(p_L / (1 - p_L),\ p_U / (1 - p_U))$
- In more complex modeling, however (e.g., logistic regression) we often use other formulas based on using the "delta method" with
  - Identity link (less often)
  - Log link (most often)
    - We exponentiate a CI for the log odds
    - (a log transformation is also "monotonic"

119

Approximate Distribution

- Use the central limit theorem

$$\text{Data } Y_1,\ldots,Y_n \overset{\text{iid}}{\sim} B(1,p) \quad E(Y_i) = p \quad Var(Y_i) = p(1-p)$$

$$\hat{p} = \overline{Y} \overset{\cdot}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

$$\hat{o} = \frac{\hat{p}}{(1-\hat{p})} \overset{\cdot}{\sim} N\left(o = \frac{p}{(1-p)}, \frac{p}{n(1-p)^3}\right)$$

$$\log(\hat{o}) \overset{\cdot}{\sim} N\left(\log(o), \frac{1}{np(1-p)}\right)$$

120

One Sample Inference for Rates

Large Samples

(reF: Lecture notes and from Biost 517, lecture 11 and recordings from Nov 9, 2012)

121

Incidence Rates

- In some studies, we make inference about rates of some event over space and / or time
  - E.g., Estimation of cancer incidence rates
    - Number of new cases of cancer diagnosed per person – year of observation
  - E.g., Number of colon polyps that grow in a person during a 3 year period
  - E.g., Number of respiratory tract infections in cystic fibrosis patients

122

Incidence Rates

- A mean, normalized to a standard period of time and a standard area of space (population)
- Most often, inference is based on a probability model involving the Poisson distribution
- Assumptions that lead to Poisson
  - In a small interval of space and time, only one event can occur
  - The number of events occurring in nonoverlapping intervals are independent
- Alternatively, Poisson approximation to binomial

123

Incidence Rates: Data

- Typically, the data for incidence rate data consist of
  - Length of time-space interval a subject is under observation
    - E.g., “Person – years” of observation
  - Number of events observed in that subject
  - Quite often, aggregate data is all that is presented
    - Total person – years of observation
    - Total number of events across subjects

124

Point Estimate

- Use the "sample mean"

Data $X_1,\ldots,X_n$ independent with $X_i \sim P(\lambda t_i)$ (t_i known)

$E(X_i) = \lambda t_i \quad Var(X_i) = \lambda t_i$

$$Y = \sum_{i=1}^{n} X_i \sim P(\lambda_0 t) \text{ with } t = \sum_{i=1}^{n} t_i$$

Point estimate : $\hat{\lambda} = \frac{Y}{t}$

125

Approximate Distribution

- From central limit theorem

Data $X_1,\ldots,X_n$ independent with $X_i \sim P(\lambda t_i)$ (t_i known)

$E(X_i) = \lambda t_i \quad Var(X_i) = \lambda t_i$

$$Y = \sum_{i=1}^{n} X_i \sim P(\lambda_0 t) \text{ with } t = \sum_{i=1}^{n} t_i$$

$$\hat{\lambda} = \frac{Y}{t} \sim N\left(\lambda, \frac{\lambda}{t}\right)$$

126

Continuity Correction

- As with the binomial distribution, the number of events is discrete
  - We do not usually bother with the continuity correction, but it would make sense

$$\Pr\left(\hat{\lambda} \le \frac{k}{t}\right) = \Pr\left(\hat{\lambda} \le \frac{k+0.5}{t}\right)$$

$$\Pr\left(\hat{\lambda} \ge \frac{k}{t}\right) = \Pr\left(\hat{\lambda} \ge \frac{k-0.5}{t}\right)$$

127

Asymptotic CI: Best Approach

- We do best by considering mean-variance relationship and continuity correction
  - Requires quadratic formula or iterative search

100(1-α)% CI for λ : $\left(\hat{\lambda}_L, \hat{\lambda}_U\right)$

$$\hat{\lambda}_L = \hat{\lambda} - \frac{1}{2t} - z_{1-\alpha/2}\sqrt{\frac{\hat{\lambda}_L}{t}}$$

$$\hat{\lambda}_U = \hat{\lambda} + \frac{1}{2t} + z_{1-\alpha/2}\sqrt{\frac{\hat{\lambda}_U}{t}}$$

128

Asymptotic CI: Elevator Stats

- Often we can just use best estimate of λ in standard error for confidence intervals and ignore the continuity correction
  - number of events and t must be large

$$100(1-\alpha)\%\ \text{CI for } \lambda: \quad \hat{\lambda} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\lambda}}{t}}$$

129

Asymptotic P values: Best

- We do best by considering mean-variance relationship and continuity correction

P values for $H_0: \lambda = \lambda_0$:

$$\text{Lower one-sided P}: \quad P_{lower} = \Pr\left(Z \le \frac{\hat{\lambda} + \frac{1}{2t} - \lambda_0}{\sqrt{\lambda_0 / t}}\right)$$

$$\text{Upper one-sided P}: \quad P_{upper} = \Pr\left(Z \ge \frac{\hat{\lambda} - \frac{1}{2t} - \lambda_0}{\sqrt{\lambda_0 / t}}\right)$$

$$\text{Two-sided P}: \quad 2 \times \min\left(P_{lower}, P_{upper}, 0.5\right)$$

130

Asymptotic P values: Elevator

- We still consider mean-variance relationship but ignore continuity correction

P values for $H_0: \lambda = \lambda_0$:

$$\text{Lower one-sided P}: \quad P_{lower} = \Pr\left(Z \le \frac{\hat{\lambda} - \lambda_0}{\sqrt{\lambda_0 / t}}\right)$$

$$\text{Upper one-sided P}: \quad P_{upper} = \Pr\left(Z \ge \frac{\hat{\lambda} - \lambda_0}{\sqrt{\lambda_0 / t}}\right)$$

$$\text{Two-sided P}: \quad 2 \times \min\left(P_{lower}, P_{upper}, 0.5\right)$$

131

Stata Commands

- "ir *countvar timevar*"
  - ir = "incidence rates"
  - timevar = person – years (or area)

132

Exact Inference

- In the one sample problem, exact inference is possible
- It is not as common to use exact inference for Poisson rates, however
  - Usually considering Poisson approximation to the binomial
  - Most often we are in a two (or more) sample setting
- We often use a log link for the rates

Data $X_1,\ldots,X_n$ independent with $X_i \sim P(\lambda t_i)$ (t_i known)

$$Y=\sum_{i=1}^{n} X_i \sim P(\lambda_0 t) \text{ with } t=\sum_{i=1}^{n} t_i$$

$$\hat{\lambda}=\frac{Y}{t} \sim N\left(\lambda,\frac{\lambda}{t}\right) \quad \Rightarrow \quad \log(\hat{\lambda}) \sim N\left(\log(\lambda),\frac{1}{t\lambda}\right)$$

133

Incidence Rates: Comments

- The assumption that incidence rate data might follow the Poisson distribution is a very strong one
- Usually the rate is changing over time, which causes the data to be more variable than the Poisson analysis might allow fo
- But many times, the real reason we are using a Poisson analysis is just as an approximation to the binomial distribution in the presence of a very low probability of event

134

One Sample Inference for Multinomial Probabilities

Large Samples

(Uncensored)

135

Generalization from Binomial

- We sometimes have categorical data that can take on more than two values
  - Unordered (nominal): e.g., disease diagnosis
  - Ordered qualitatitive: e.g., stage of cancer
  - (Ordered quantitative: e.g., number of asthma exacerbations)
- One approach to inference is to generalize the results for binomial proportions and odds
  - Consider each group against all the others
    - The only choice for unordered categories
    - Multiple comparisons due to multiple parameters
  - For each possible threshold, consider probabilities of being above vs below that threshold
    - Proportional odds model assumes a constant odds as the threshold varies

136

One Sample Inference for Categorical Data

Summary

137

Final Comments

- Our observations about inference with categorical data from one sample studies will guide us as we progress to
  - Two sample studies
  - Stratified analyses
  - Regression analyses
- As we go beyond one sample studies, we will have to consider
  - Mean-variance relationships,
  - Discreteness of the outcome space, and
  - Adequacy of normal approximations
    - (We will generally not have exact methods in regression with categorical data)

138

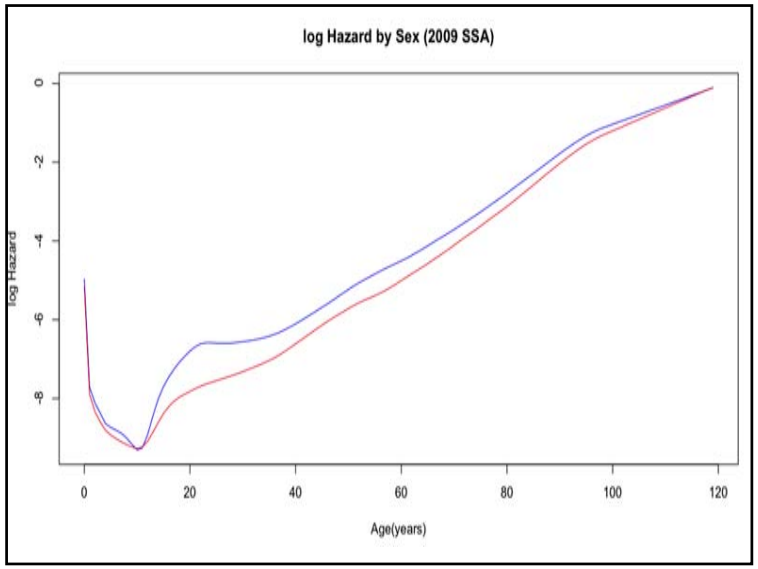
Epidemiologic Measures of Association (Intro)

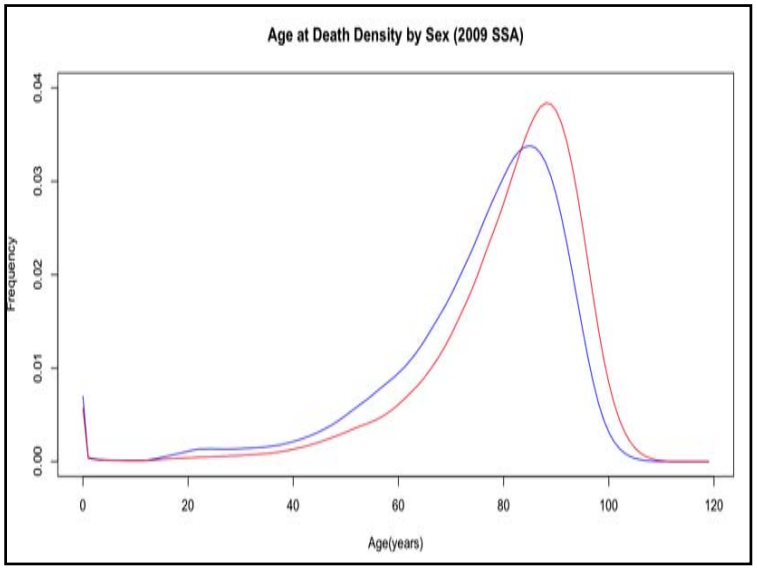
139

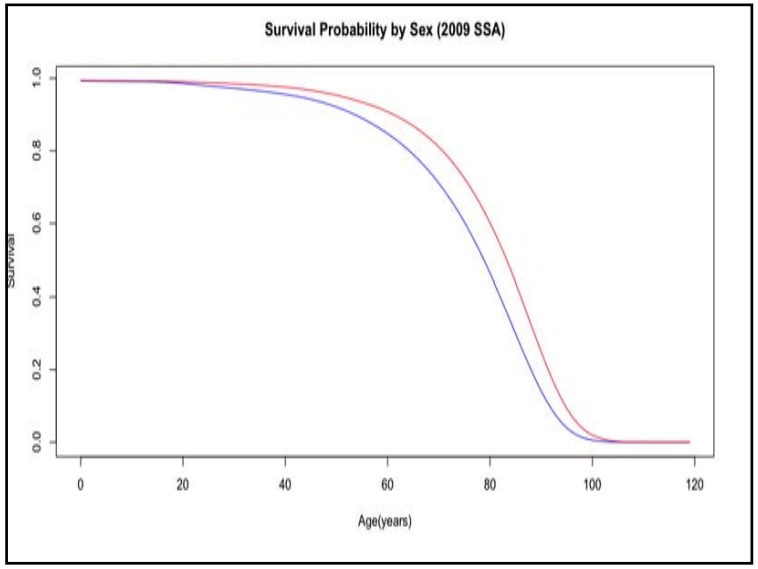
Comparing Distributions

- Consider the following three graphs depicting the mortality experience predicted using US Social Security Data from 2009
- What measures would you look at descriptively to decide that there is an association between sex and mortality?
- In particular, how can you read from the survival curves
  - Comparisons of survival at a particular time
  - Comparisons of medians (quantiles) of the distributions
  - Comparisons of means
  - Comparisons of geometric means
  - Comparisons of hazards
- Is there effect modification?

140


log Hazard by Sex (2009 SSA)
log Hazard
0
-2
-4
-6
-8
0
20
40
60
80
100
120
Age(years)


Age at Death Density by Sex (2009 SSA)
Frequency
0.00
0.01
0.02
0.03
0.04
0
20
40
60
80
100
120
Age(years)


Survival Probability by Sex (2009 SSA)
Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
Age(years)