Biost 536 / Epi 536
Categorical Data Analysis in Epidemiology

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

Lecture 5:
Standardized Rates

October 9, 2014

1

Lecture Outline

- Review
  - Stratified Analyses
  - Prevalence, incidence, hazard (incidence rate)
- Example: Colorectal Cancer Incidence in US Whites

2

Review

Stratified Analyses

3

Two Sample Studies

- Summary comparison between two populations

General approach:

$Y_{ij} \sim F_i(y)$ (distribution of data in group i)

$\theta = d(F_1, F_0)$ (some comparison of distributions)

Most common approach:

$\omega_i = \omega(F_i)$ (summarize distribution: mean, median, ...)

$\theta = \begin{cases} \omega_1 - \omega_0 \\ \omega_1 / \omega_0 \end{cases}$ (contrast across distributions: difference, ratio)

4

Two Sample Studies: Binary Outcome

- Summary comparison between two populations

$Y_{ij} \sim B(1, p_i)$ (Bernoullidistribution of data in group i)

$$\omega_i = \begin{cases} p_i \\ \dfrac{p_i}{(1-p_i)} \end{cases}$$ (summarizedistribution: proportion, odds)

$$\theta = \begin{cases} \omega_1 - \omega_0 \\ \omega_1 / \omega_0 \end{cases}$$ (contrastacrossdistributions: difference, ratio)

5

What Changes with Stratifying Variables?

- Suppose we know have *K* independent strata
  - Strata perhaps defined by combinations of other covariates

$Y_{ikj} \sim B(1, p_{ik})$ (Bernoullidata in group i of stratum k)

$$\omega_{ik} = \begin{cases} p_{ik} \\ \dfrac{p_{ik}}{(1-p_{ik})} \end{cases}$$ (summarizedistribution: proportion, odds)

$$\theta_k = \begin{cases} \omega_{1k} - \omega_{0k} \\ \omega_{1k} / \omega_{0k} \end{cases}$$ (contrastacrossdistributions: difference, ratio)

$$\theta = \frac{\sum_1^K w_k \theta_k}{\sum_1^K w_k}$$ (overalleffect: weightedaverage)

6

How Should We Choose Weights?

- It depends
- Do we think there is no effect modification (for measure used)?
  - Any weighted average will be consistent for the same estimand
  - Choose "efficiency weights" to maximize precision (minimize SE)

If $\theta_k = \theta$ in each independent stratum:

$$\frac{\sum_1^K w_k \theta_k}{\sum_1^K w_k} = \frac{\sum_1^K w_k \theta}{\sum_1^K w_k} = \theta \frac{\sum_1^K w_k}{\sum_1^K w_k} = \theta$$

If $\hat{\theta}_k \sim N(\theta, V_k)$ in k-th stratum:

$$\frac{\sum_1^K w_k \hat{\theta}_k}{\sum_1^K w_k} \sim N\left(\theta, \frac{\sum_1^K w_k^2 V_k}{\left(\sum_1^K w_k\right)^2}\right)$$ (G-M theorem: weights $w_k = \frac{1}{V_k}$ minimizevariance)

7

How Should We Choose Weights?

- It depends
- Do we think there might be some effect modification?
  - Do we want to prove effect modification exists?
    - Perform a test for differences among the stratum specific θ_k
      - ANOVA like test across unordered strata
      - First order trend across ordered strata
  - Do we want to quantify the effect modification?
    - Estimate a trend across ordered strata?
      - Use "contrast" instead of "weighted average"
    - Estimate each stratum specific θ_k separately?
  - Do we just want to describe an average effect?
    - Choose "importance weights" relevant to some population?
    - Choose "efficiency weights" if believe only slight effect modification or if unsure of best "importance weights"?

8

How Should We Choose Weights?

- It depends
- Do we just want to describe an average effect?
  - Choose "importance weights" relevant to some population?

If θ_k might differ in each independent stratum:

$$\frac{\sum_1^K w_k \theta_k}{\sum_1^K w_k} = \theta_{\vec{w}} \qquad \text{(estimand depends on choice of weights } \vec{w})$$

If $\hat{\theta}_k \sim N(\theta, V_k)$ in k-th stratum:

$$\frac{\sum_1^K w_k \hat{\theta}_k}{\sum_1^K w_k} \sim N\left(\theta_{\vec{w}}, \frac{\sum_1^K w_k^2 V_k}{\left(\sum_1^K w_k\right)^2}\right) \qquad \text{(more precision for some estimands than others)}$$

9

Importance Weights: RD

- For RD, common choices try to estimate public health impact of policy
  - What happens if exposed became unexposed?
    - Standardize to "exposed group":
  - What happens if unexposed became exposed?
    - Standardize to "unexposed group":
- In Stata: `cs …, rd`
  - `istandard` option standardizes to "exposed" distribution
  - `estandard` option standardizes to "unexposed" distribution
  - `standard(`*`varname`*`)` option for arbitrary distribution

10

Importance Weights: RR

- For RR, we truly have three approaches that could be considered
  - Average the stratum specific proportions for each group, then take ratio
    - Standardized Rate Ratio: ratio of "directly standardized rates"
    - Stata: cs, cc, ir
  - Take the ratio in each stratum, then use weighted average
    - (Not really a very interpretable quantity)
  - Take the ratio in each stratum, then use weighted geometric mean
    - Equivalent to using log link on the stratum specific proportions
    - Hence, equivalent to Poisson regression

11

Weighted Analyses with RR

$$Y_{ikj} \sim B(1, p_{ik}) \qquad \text{(Bernoulli data in group } i \text{ of stratum } k)$$

$$\theta = \frac{\sum_1^K w_k p_{1k}}{\sum_1^K w_k p_{0k}} \qquad \text{(SRR: ratio of directly standardized rates)}$$

$$\theta = \frac{\sum_1^K w_k \frac{p_{1k}}{p_{0k}}}{\sum_1^K w_k} \qquad \text{(weighted average of stratum specific RR)}$$

$$\theta = \exp\left[\frac{\sum_1^K w_k \log\left(\frac{p_{1k}}{p_{0k}}\right)}{\sum_1^K w_k}\right] \qquad \text{(weighted geometric mean of stratum specific RR)}$$

12

Review

Prevalence, (Cumulative) Incidence, Hazard

13

Epidemiologic Measures of Disease

- Prevalence
  - Probability of being diseased in population
  - A combination of incidence and case-fatality / cure
    - (We usually prefer to estimate these separately)
- (Cumulative) incidence
  - Defined over some finite interval of time
  - Probability of new diagnoses among previously non diseased
- Incidence rate
  - Instantaneous rate of new disease
  - Probability of new diagnoses among previously non diseased ***over some infinitesimal period of time***

14

Alternative Perspectives: Binary Data

- Suppose we observe colorectal cancer incidence over 14 years

  US born: 62,668

  non US: 11,026
- We can use census data to estimate denominators
  - E.g., 16,082,630 people x 14 years = 225,156,822 person years
- We could have estimated the proportion of people who would get colorectal cancer within 14 years:

  US born: 62,668 / 16,082,630 = 0.390% get cancer in 14 years

  non US: 11,026 / 1,031,721 = 1.069% get cancer in 14 years
- Can we just divide by 14 to get the incidence rate (hazard)?

15

Example: Constant Hazard of 1%

- Suppose we have exponential data

Year	At risk	Events
1	100,000	1,000
2	99,000	990
3	98,010	980
4	97,030	970
5	96,060	961
Cumulative	100,000	4,901 = 4.90%

(dividing by 5 gives less than 1%)

16

Example: Constant Hazard of .01%

- Suppose we have exponential data

Year	At risk	Events
1	100,000	10.000
2	99,990	9.999
3	99,980	9.998
4	99,970	9.997
5	99,960	9.996
Cumulative	100,000	49.99 = .04999%

(dividing by 5 = close approximation to .01%)

17

(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

Survivor function

$$S(a) = \Pr(T > a) = e^{-\int_0^a \lambda(u)\,du}$$

18

(Cumulative) Incidence and Mortality

- Cumulative incidence: relationship to hazard

(Cumulative) incidence between times a and b

$$\Pr(T > a) = e^{-\int_0^a \lambda(u)\,du}$$

$$\Pr(a \le T < b \mid a \le T) = \frac{\Pr(a \le T < b)}{\Pr(a \le T)} = \frac{\Pr(a \le T) - \Pr(b < T)}{\Pr(a \le T)}$$

$$= 1 - \frac{\Pr(b < T)}{\Pr(a \le T)} = 1 - \frac{e^{-\int_0^b \lambda(u)\,du}}{e^{-\int_0^a \lambda(u)\,du}} = 1 - e^{-\int_a^b \lambda(u)\,du}$$

19

Constant Hazard

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

(Cumulative) incidence between times a and b

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

$$= 1 - e^{-\int_a^b \lambda\,du}$$

$$= 1 - e^{-\lambda(b-a)}$$

- When hazard functions are constant, the distribution of *T* is the exponential distribution

20

Piecewise Exponential

- We divide time into small intervals
  - Small age intervals will have common risk function (hazard)
  - (Hopefully similar birth cohorts, as well, unless modeled)
  - Small follow-up time intervals will have (nearly) constant hazard
- The incidence rate function (hazard) is thus approximated as a "step-function"
  - Constant over small intervals with jumps between intervals
  - "Piecewise exponential"
- Properties of exponential thus allow us to combine data according to person-years of observation
  - Independent intervals are independent, both within and between individuals

21

Reformulation of the Problem

- Divide all data into strata based on
  - Time (age)
  - Other baseline covariates (sex, race, ethnicity, …)
- Assume constant hazard in stratum *s*: λ_s
- For the *i*-th observation and each stratum *s* define
  - Let t_{is} be the time that individual was observed to be at risk (without yet having event) while belonging to stratum *s*
  - Let Y_{is} be an indicator that the individual had an event during the time he/she belonged to stratum *s*

$$\Pr(Y_{is}=1)=1-e^{-\lambda_s t_{is}}$$

22

Approximation for Rare Events

- We expect $\lambda_s t_{is}$ to be small:
  - Usually applied in rare disease, and
  - We have divided time and covariates into many strata
- We can use the Poisson approximation to the binomial

$$\Pr(Y_{is}=1)=1-e^{-\lambda_s t_{is}} \approx \lambda_s t_{is}$$
$$Y_{is} \sim B\left(1, 1-e^{-\lambda_s t_{is}}\right) \Rightarrow Y_{is} \sim P(\lambda_s t_{is})$$

- Sums of independent Poisson random variables having constant rate allow us to just sum the observation time

$$Y_{is} \sim P(\lambda_s t_{is}) \Rightarrow \sum_{i=1}^{N} Y_{is} \sim P\left(\lambda_s \sum_{i=1}^{N} t_{is}\right)$$

23

Large Sample Inference

- Assuming total time of observation while at risk is sufficiently large, a central limit theorem applies:
  - Normal approximation to the Poisson

$$Y_{is} \sim P(\lambda_s t_{is}) \Rightarrow \sum_{i=1}^{N} Y_{is} \sim P\left(\lambda_s \sum_{i=1}^{N} t_{is}\right) \Rightarrow \sum_{i=1}^{N} Y_{is} \sim N\left(\lambda_s \sum_{i=1}^{N} t_{is}, \lambda_s \sum_{i=1}^{N} t_{is}\right)$$

$$\Rightarrow \hat{\lambda}_s = \frac{\sum_{i=1}^{N} Y_{is}}{\sum_{i=1}^{N} t_{is}} \sim N\left(\lambda_s, \frac{\lambda_s}{\sum_{i=1}^{N} t_{is}}\right) \Rightarrow se\left(\frac{\sum_{i=1}^{N} Y_{is}}{\sum_{i=1}^{N} t_{is}}\right) = \sqrt{\frac{\lambda_s}{\sum_{i=1}^{N} t_{is}}}$$

$$\hat{se}(\hat{\lambda}_s) = \sqrt{\frac{\hat{\lambda}_s}{\sum_{i=1}^{N} t_{is}}}$$

24

If Hazard Rate Not Constant

- Sums of independent Poissons are still Poisson distributed
  - But now we are estimating a weighted average of incidence rates
  - Weighting by time of observation
- Assuming total time of observation while at risk is sufficiently large, and the distribution of different rates is "well-behaved":

$$Y_i \sim P(\lambda_i t_{is}) \Rightarrow \sum_{i=1}^{N} Y_i \sim P\left(\sum_{i=1}^{N} \lambda_i t_i\right) \Rightarrow \sum_{i=1}^{N} Y_{is} \sim N\left(\sum_{i=1}^{N} \lambda_i t_i, \sum_{i=1}^{N} \lambda_i t_i\right)$$

$$\Rightarrow \hat{\theta} = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} t_i} \sim N\left(\frac{\sum_{i=1}^{N} \lambda_i t_i}{\sum_{i=1}^{N} t_i}, \frac{\sum_{i=1}^{N} \lambda_i t_i}{\left(\sum_{i=1}^{N} t_i\right)^2}\right) \Rightarrow \hat{se}\left(\frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} t_i}\right) = \frac{\sqrt{\sum_{i=1}^{N} \lambda_i t_i}}{\sqrt{\left(\sum_{i=1}^{N} t_i\right)^2}}$$

$$\hat{se}(\hat{\theta}) = \frac{\sqrt{\hat{\theta}}}{\sqrt{\sum_{i=1}^{N} t_{is}}}$$

25

Convoluted Reasoning That Works Here

- Under the assumption of rare events that have a piecewise constant hazard
- "Memorylessness" of exponential says we would get the same distribution if we had observed 225,156,822 independent people for 1 year
  - Binomial data for which we are estimating *p*
  - Our person-years of observation can be treated like *n*
  - (Our measuring person-years *while at risk* guarantees the relevance of this proportion as an estimate of a hazard)
- Rare events means that (1-*p)* is approximately 1

$$\hat{\theta} = \frac{N_{Events}}{PY} \sim N\left(\frac{\sum_{i=1}^{N} \lambda_i t_i}{\sum_{i=1}^{N} t_i}, \frac{\sum_{i=1}^{N} \lambda_i t_i}{\left(\sum_{i=1}^{N} t_i\right)^2}\right) \Rightarrow \hat{se}(\hat{\theta}) = \sqrt{\frac{\hat{\theta}}{PY}}$$

26

Piecewise Exponential

- 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
  - The exponential distribution is "memorylessness"
    - Independent intervals are independent
      - Within or between individuals
  - Also be thought of as Poisson approximation to binomial and/or times between events in Poisson process

27

Example: Incidence of Colorectal Cancer by Birthplace

28

Example

- We are interested in exploring the incidence of colorectal cancer by birthplace among whites in the US
- Cases identified through the SEER registry 1973-1987
- Available data
  - US, 25 non-US, unknown
  - Age in 5 year groups
  - Sex
- Denominator data from US census data

29

Specific Aim

- For the purposes of this example:
  - Compare incidence rates
    - Persons known to be born in US
    - Persons known to be born outside the US
  - Pretend subjects with missing data are missing completely at random
    - Consider methods for missing data in discussion section
- Compare rates separately for males and females
  - "Subgroup analyses"
- Adjust for known association between cancer incidence and age

30

Effect Modification by Age

- Strongly suspect that there would be effect modification by age
  - Analysis models could consider RD, RR, OR
  - At most one of those measures could have no effect modification by age
  - Most likely they all do
- However,…
  - Want to avoid information overload of having to provide association between cancer incidence and birthplace for each age
  - Instead provide some sort of summary across all ages
    - If effect modification is only magnitude (not direction) of effect, this will only affect quantification of association, not qualification

31

Choosing Measure of Association

- Choices: RD, RR, OR
- For credibility of statistical inference, we need to avoid issues with multiple comparisons
  - Choose analysis methods prior to looking at the data
- For teaching purposes we will
  - First discuss the general advantages and disadvantages of each measure when used in adjusted analyses
  - Then examine the data to illustrate the issues we would try to anticipate
    - I am not advocating that this would be done in order to choose the analysis model

32

A priori: Relative Advantages of RD

- Absolute difference is easily understood, but "disambiguate"
  - Often need to make clear that a difference of, say, 10% means 20% vs 30% or 60% vs 70%, not 20% vs 22% or 60% vs 66%
- Magnitude is similar whether consider RD of A-B or B-A
  - 20% vs 30% has RD of 10%; 30% vs 20% has RD of -10%
- Magnitude is similar whether discussing probability of event or probability of being event-free
  - 10% difference in mortality is -10% difference in survival
- Collapsible across strata defined by precision variables
  - Population effect is same as within stratum (individual) effect
- Highly relevant when considering impact of public health policy
  - What proportion of the population will benefit?

33

A priori: Relative Disadvantages of RD

- May not be as useful to highlight risk factors for rare diseases
  - Difference between 0.005% and 0.0005% may seem obscure
- May be more prone to effect modification
  - If baseline risk in unexposed varies markedly across adjustment strata, a strong association must be modified
  - Impossible to have a RD of 20% across all strata if baseline risk varies from 10% to 90%

34

Example: Anticipating Issues with RD

- Incidence of colorectal cancer by birthplace separately by sex
  - Adjusting for age
- Effect modification in RD by age?
- Do we care about effect modification by age?
  - If so, are we interested in
    - Establishing existence of effect modification?, OR
    - Quantifying association within each age group?
  - If not, how will we summarize association across age groups?
    - Is there confounding of the birthplace - colorectal cancer incidence association by age?
      - If so, we will want to adjust in some way to make the birthplace groups more comparable
    - Would we gain precision by adjusting for age?
      - How would we adjust in order to gain such precision?

35

Example: Effect Modification using RD

- Effect modification in RD by age?
- Young ages have extremely low incidence rate (hazard) expected in either group
- Older ages have higher hazard
- If there is no effect modification by age, there is unlikely to be a difference in risk at any age that we would ever care about
  - We would be limited by the incidence at the lowest ages

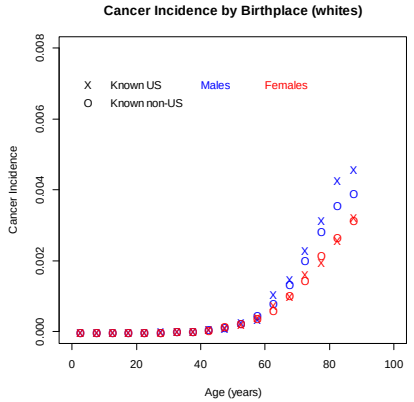
36

Example: Descriptive Statistics for EM

- Plot incidence rate as a function of age
  - Parallel lines would suggest no effect modification on RD
- We will actually plot cumulative incidence per year within 5 year age groups
  - We assume incidence rate (hazard) is nearly constant for observations of similar age, similar birth-year cohort, and small period of follow-up
  - Then we can collapse person-years of observation within that stratum

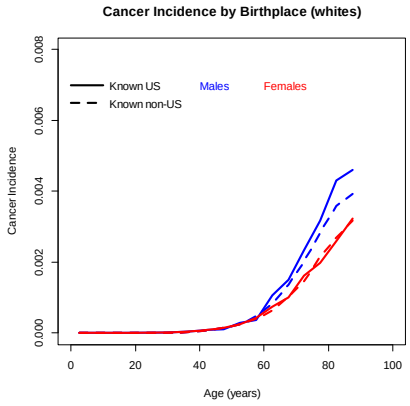
37

Example: Descriptive Statistics for EM



38

Example: Continuous Approximation



39

Example: Effect Modification for RD?

- Age stratum specific of effect measured by RD:
  - Vertical separation between solid (US born) and dashed (non-US born) curves of same color (blue= males, red= females)
- Among males:
  - Clearly non-parallel lines over age
  - Greater separation (RD) at higher ages
- Among females:
  - Lines appear more parallel (coincident) over age
    - (If there is no effect of birthplace at any age, there cannot be effect modification)

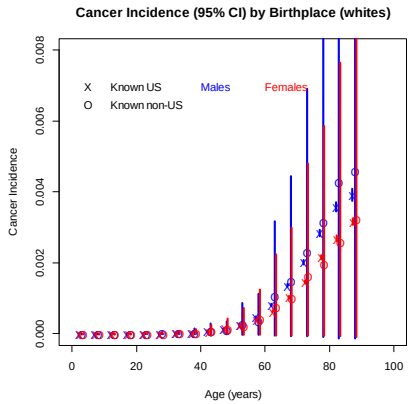
40

Do We Care About Effect Modification?

- Is our question primarily one of effect modification?
  - Either establishing existence of EM or quantifying effect in each age subgroup?
- Should we try to answer questions about effect modification before we know there is an effect?
- Is there an advantage of finding some overall tendency?
  - What would we have to report in the presence of effect modification?
    - A p value testing for effect modification?
    - The estimated RD in every stratum?
  - Do we have sufficient precision to handle multiple comparisons?

41

Precision? Including Stratum Specific CI



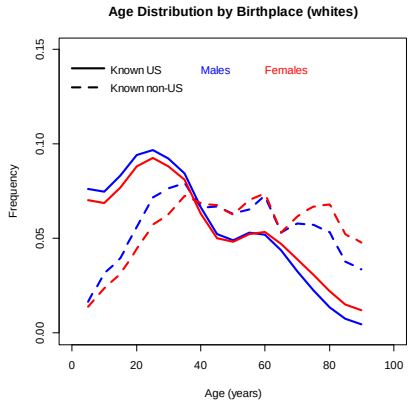
42

Suppose We Do Not Care About EM?

- We need to find some way to summarize over the different ages
- Possibilities
  - Look at incidence across all ages (unadjusted)
    - Should we worry about confounding?
  - Directly standardized rates
    - Standardize to some population
  - Regression
    - Poisson
    - Logistic?
    - RR regression with Gaussian family?

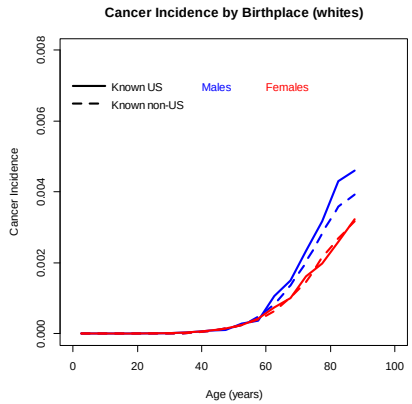
43

Confounding? Age – Birthplace Assoc



44

Confounding? Age – Cancer Assoc



45

Confounding by Age

- We definitely see (and know *a priori*) that age is associated with cancer incidence
  - We believe that to be causal
  - We believe that to be independent of birthplace
- We definitely see a different distribution of age across groups defined by birthplace
  - This is not surprising given demographics of immigration over the years
- Hence, if we decide to analyze by RD, we would want to adjust for age
  - But a straight line fit would not be very good

46

A priori: Relative Advantages of RR

- Risk ratio is relatively easily understood, but "disambiguate"
  - We sometimes report a RR of 1.1 as being 10% higher risk
  - We then need to make clear that 10% higher means 20% vs 22% or 60% vs 66%, not 20% vs 30% or 60% vs 70%
- Useful to highlight risk factors for rare diseases
  - Difference between 0.005% and 0.0005% is RR of 10
- Stays constant when each group is "contaminated" with subjects who are not susceptible
  - Suppose RR is 1.25 in true target population: 20% vs 25%
  - In a "contaminated" population in which 50% are not susceptible, the rates would be 10% vs 12.5%, still RR of 1.25
- Collapsible across strata defined by precision variables
  - Population effect is same as within stratum (individual) effect

47

A priori: Relative Disadvantages of RR

- People sometimes have difficulty recognizing same RR when expressed as B/A rather than A/B
  - 20% vs 25% is RR of 1.25, sometimes stated as 25% increase
  - 25% vs 20% is RR of 0.80, sometimes stated as 20% decrease
- Magnitude is different when discussing probability of event or probability of being event-free
  - 20% vs 25% mortality is mortality RR of 1.25 (inverse: 0.80)
  - 80% vs 75% survival is survival RR of 0.938 (inverse: 1.067)
- With common diseases would be prone to effect modification
  - E.g., RR of 2.0 is only possible when baseline risk below 50%
- When considering impact of public health policy also need to know "baseline" risk in relevant stratum
  - RR of 2.0 more important when risk is 5% than when risk is 0.05%

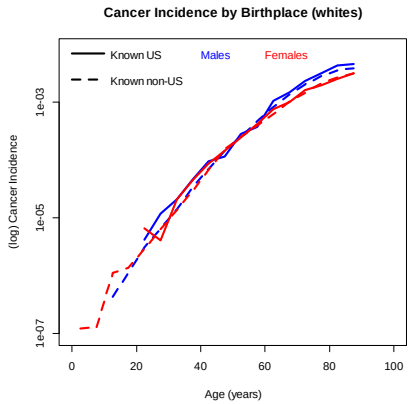
48

Example: Descriptive for EM with RR

- Plot ***log*** incidence rate as a function of age
  - (Use incidence rate labeling with log spacing)
  - Parallel lines would suggest no effect modification on RR
- We will actually plot cumulative incidence per year within 5 year age groups
  - We assume incidence rate (hazard) is nearly constant for observations of similar age, similar birth-year cohort, and small period of follow-up
  - Then we can collapse person-years of observation within that stratum

49

Incidence (on log scale) vs Age



50

Findings

- Some suggestion of effect modification still evident on RR
- Still suggestion of strong age-cancer association
  - But it is a bit more linear so may allow easier modeling
- (Age – birthplace association is still present in data as shown before)

51

Large Sample Inference for log Rate

- Assuming total time of observation while at risk is sufficiently large, a central limit theorem applies:
  - Normal approximation to the Poisson along with “delta method”

$$\hat{\lambda}_s = \frac{\sum_{i=1}^{N} Y_{is}}{\sum_{i=1}^{N} t_{is}} \sim N\left(\lambda_s, \frac{\lambda_s}{\sum_{i=1}^{N} t_{is}}\right) \quad \Rightarrow \quad \log\left(\hat{\lambda}_s\right) \sim N\left(\log(\lambda_s), \frac{1}{\lambda_s \sum_{i=1}^{N} t_{is}}\right)$$

52

A priori: Relative Advantages of OR

- Approximates RR with rare diseases
  - Can be estimated from case-control study
- Magnitude is the same when discussing probability of event or probability of being event-free
  - 20% vs 25% mortality is mortality OR of 1.33 (inverse: 0.75)
  - 80% vs 75% survival is survival OR of 0.75 (inverse: 1.33)
- Perhaps less prone to effect modification
  - Every OR contour is defined for every baseline risk in unexposed
  - (Gains in parsimony may make up for "less understood" odds)

53

A priori: Relative Disadvantages of OR

- People do not intuitively understand odds as well as they do probability
- People sometimes have difficulty recognizing same OR when expressed as B/A rather than A/B
- When considering impact of public health policy also need to know "baseline" risk in relevant stratum
  - OR of 2.0 more important when risk is 50% than when risk is 1%

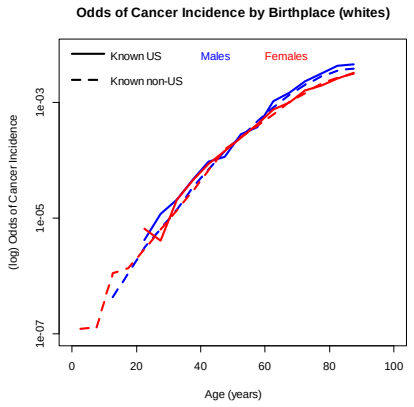
54

Example: Descriptive for EM with OR

- Plot ***logit*** incidence rate as a function of age
  - Parallel lines would suggest no effect modification on RR
- We will actually plot cumulative incidence per year within 5 year age groups
  - We assume incidence rate (hazard) is nearly constant for observations of similar age, similar birth-year cohort, and small period of follow-up
  - Then we can collapse person-years of observation within that stratum

55

(log) Odds of Cancer Incidence by Age



56

Findings

- What did we expect?
- With a rare disease, OR is approximately RR

57

Large Sample Inference for log Odds

- Assuming total time of observation while at risk is sufficiently large, a central limit theorem applies:
  - Normal approximation to the Poisson along with "delta method"

$$\hat{\lambda}_s = \frac{\sum_{i=1}^{N} Y_{is}}{\sum_{i=1}^{N} t_{is}} \approx N\left(\lambda_s, \frac{\lambda_s}{\sum_{i=1}^{N} t_{is}}\right) \quad \Rightarrow \quad \log\left(\frac{\hat{\lambda}_s}{1-\hat{\lambda}_s}\right) \approx N\left(\log\left(\frac{\lambda_s}{1-\lambda_s}\right), \frac{1}{\lambda_s(1-\lambda_s)\sum_{i=1}^{N} t_{is}}\right)$$

58

Stata: Standardized Incidence Rates

- `ir casevar expvar timevar, by() standard()`
  - Stata only allows one variable in by()
  - May need to create a new variable
    - `E.g., .g strata= 10*agectg + male`
  - Default standardization is Mantel-Haenszel
- Will provide stratum specific RR as well as standardized rates

59