6. What is the rate is estimating in simplest case:
   a. Model: \( P[T \geq t] = \exp(-\lambda t)(= S(t)) \)
      i. Hence rate is \( -\frac{d}{dt}\log(S(t)) = \lambda \)
   b. Suppose every individual is observed between \( C \) years and \( K \) years
      i. For ex., \( C = 20 \) and \( K = 60 \)
   c. Then
   \[
   E[\text{time}] = \int_C^K \exp(-t\lambda) dt + \int_K^\infty \exp(-t\lambda) dt - C
   = \exp(-C\lambda)(1 + C\lambda) - \exp(-K\lambda)(1 + K\lambda)/\lambda + C \exp(-K\lambda) - C
   = \lambda + \frac{C^2\lambda^2}{K} + O(\lambda^3)
   \]

7. Summarize overall age groups
   a. Notation: Strata width \( \Delta \)
   b. Crude rate for standard population is
   \( R = \sum_i R_i P_i / \sum_i P_i \)
   c. Is study group doing better than standard population?
      i. Differences in rates might arise from differences in ages.

   - Get \( \sum \frac{d_i(P_i\Delta)/(P_i\Delta)}{\sum R_iP_i\Delta} = \sum \frac{\lambda_i(P_i\Delta)}{\sum R_iP_i\Delta} \)
   - Multiplier is called Comparative Mortality Figure (CMF)
   Se: 1 pp. 34f.
   - Could also standardize rate
   - Remove standard rates from denominator
   - Get \( \sum \frac{d_i(P_i\Delta)/(P_i\Delta)}{\sum R_iP_i\Delta} = \sum \frac{\lambda_i(P_i\Delta)}{\sum R_iP_i\Delta} \)
   - Calculate crude rate for study population as if
   - Age-specific rates are same as for study population
   - Age group sizes were proportional to standard population
   - Called Directly standardized rate
   B&D2: 2.2b

8. What if people in age group are studied for different times?
   iv. Estimate cumulative hazard \( \Lambda(t) \) at some time point
      - Estimate as \( \sum_i \Delta_i \hat{\lambda}_i \)
      - Sum is over all intervals ending at or before \( t \)
      - thing you estimate is equivalent to \( P[ \text{event time} < t] \)
   B&D1: 2.4

9. Measuring differences and variation in rates
   a. Raw scale:
      i. \( \beta_i = \lambda_i - \lambda_{0i} \) is called excess risk.
      ii. Interpretation in absolute terms
   b. Log scale:
   i. \( \beta_i = \log(\lambda_{1i}) - \log(\lambda_{0i}) \)
   ii. \( q_i = \exp(\beta_i) \) is called relative risk.
   iii. Then \( q_i = \lambda_{1i}/\lambda_{0i} \)
   iv. Interpretation in relative terms
   v. Simple case of log linear model
      \[
      \log(\lambda_{ki}) = \log(\lambda_{0i}) + \sum_{k=1}^{\infty} \beta_{ki} \text{ if } k = 1
      \]
      \[
      = \log(\lambda_{0i}) + \beta_k \text{ if } k = 0
      \]
      This looks like a two-way analysis of variance model.
   Se: 7 p. 200
   - Perhaps can use \( \log(\lambda_k(t)) = \alpha + \beta t + \beta_k = 1_{k=1}^\infty \beta_i \)
   - Advantages when trying to estimate difference without estimating either \( \lambda_{1i} \) or \( \lambda_{0i} \), to be seen later
   Se: 2 pp. 61–65
   vi. If additivity fails, we say there is an interaction.
      - If additivity can be restored by useful transformation we say it is removable interaction
      - We only consider increasing transformations.
      - Won’t happen if order changes with strata
      - Won’t happen if one rate drops and other increases as we move across strata.
      - If interaction is not removable, it is called essential interaction.
   vii. Interactions are sometimes said to modify effects
      - Intuitively, presence of interaction indicates effect of one variable depends on level of the other
• Mathematically this interpretation is problematic, since effects interacting are formally symmetric.
• Effects are often heuristically symmetric as well.
c. Decision of raw vs. log scale might be based on making same value of β hold for a variety of i
B&D1: 2.9
d. Attributable Risk
i. Assume homogeneous rates in each risk group
( perhaps after standardizing)
i. How much of a disease rate arises because of an elevated risk factor?
ii. Naive answer λ1 − λ0
iv. Too high, because the increased rate applies only to subgroup
v. \[
\left( \frac{p_0 - p}{p_0 + (1 - p_0)} \right) \times \left( \frac{\lambda_1 \lambda_0}{\lambda_1 + \lambda_0} \right)
\]
vi. Called Attributable risk (AR)
vii. Estimate using \[ p_1 = \frac{P_1}{P_1 + P_0}, \lambda_1 = \hat{\lambda}_1 \]
\[ (\hat{P}_1 / (\hat{P}_1 + \hat{P}_0)) \approx (P_1 / (P_1 + P_0)) \]
Se: 2 pp. 51–56
D. Measuring agreement between diagnosis and presence of disease
1. Requires knowledge of “Gold Standard”
a. Diagnostic procedure close to truth that is not routinely used because of
i. invasiveness
ii. cost
2. Sensitivity and Specificity:

II. Designing studies
A. Aim
1. Estimate rates for various groups
a. Often rates vary by time
b. Often want long time course
2. Usually more importantly, to compare rates for 2 groups
a. Groups are often defined by exposure to risk factor or medication
b. Ex.,
   i. cancer cases after benzene exposure
   ii. communicable disease cases after vaccination
   iii. seizure history after medication
c. Generally want to describe association between to variables
   i. Ex, risk factor and disease
   ii. Often want to say one variable causes another
   Se: 2 p. 77
B. Three study designs
1. Best case: designed experiment
   a. Assemble group with qualities you would like to explore
   b. Randomly divide among 2 (or more) treatments to be compared
      i. Avoids having groups systematically different from each other based on some quality linked to response to treatment
      ii. Chance of getting imbalance rolled into error probability
         Se: 9 pp. 271–277
c. Often stratify on things you think may be important
   i. Definition:
      ● Split data into groups or strata
      ● Perform analysis separately on each group
      ● Reassemble results
   ii. Ex., trial of vaccine for children’s ear infection
      ● Stratify based on whether child has had previous infection
      ● If you think age is important, might even divide into groups of 2 with very similar ages, and give new medication to 1 of every 2
      ● Stratification with strata of size 1 or 2 is called matching.
      ● If strata represent continuous variable like age, generally split into bins and attach estimate to center of bin
d. Makes association imply causality B&D2: 1
2. Cohort Study
   a. Group of subjects is identified whose status will be followed
      i. Group often identified by
         ● Birth records of an area with centralized record keeping
         ● health plan membership
         ▶ VA, Kaiser, etc.
         ● military records
         ● employment records
      ii. Often this trait is exposure to some toxic chemical
data is collected
   i. generally after subjects are identified
      ● Studies can take a long time if followup is long
         ▶ Ex., cohorts exposed to radiation in 30’s and 40’s are still being followed up on
      ▶ Accurate follow up is essential
         Certain kinds of losses to followup can be accounted for if we know about them
         Losses to followup related to the disease are a problem.
   ii. occasionally before subjects are identified
      ● called a historical cohort study, or, more ambiguously, a retrospective study.
      ● Advantage: process is typically much faster.
      ● Disadvantages:
         ▶ Determining who to include is harder.
         ▶ Data available is limited to what was collected in past,
         Especially important if you want to measure
things that will be influenced by disease process
Data sufficient to test innovative hypotheses is likely not available from existing records what can be remembered (correctly)
• sometimes done via chart review
• The phenomenon of systematic errors arising from incorrect recollection is called recall bias.
  ▷ People who have disease are more likely to recall (perhaps mistakenly) exposure to putative cause

iii. sometimes a group with some rare trait is collected separately from rest of population.

c. What you can estimate:

i. rates of various diseases in each group
  • Can expect precision to be low for rare diseases
ii. proportion of exposure ONLY IF sampling not stratified (or oversampled) based on exposure
  • hence can estimate attributable risk only in this situation
  Se: 2 pp. 50–51

iii. important to avoid selection bias
  • Arises when
    ▷ being included for study is linked to outcome
    ▷ mechanism might vary among subgroups of interest
  • Possible when data are collected prospectively:
    ▷ One group might be more likely to volunteer
      Ex., subjects with family history might be more likely to volunteer for periodic examinations
  ▷ Also happens when one group is less likely to be followed up
    Ex., subjects with family history may be more diligent about attending follow up
  ▷ Unusual in these studies
  • More common when data are collected historically:
    ▷ Ex., people who died might be overrepresented because you have everyone’s death certificate
    ▷ Ex., people undeserved by medical care might be less likely to have chart to review
    ▷ Ex., people with identifiable risk factor might have been followed better.

iv. Conclusions apply only to group from which cohort was drawn
  • if you draw your cohort from a group healthier than general population, rates you see will be lower
    ▷ Ex., employed people
    ▷ Phenomenon is known as healthy worker effect.