BIOSTAT 651
Notes #8: Analysis of Binary Data

- Lecture Topics:
  - Measures of association
  - Sampling mechanisms
  - Potential biases
  - Examples

Data Structure

- Example: Consider a study of liver cancer patients (n=120) who have refused conventional therapy. Such patients were randomized to receive either an experimental treatment ($X_i = 1$) or placebo ($X_i = 0$). Patients were then followed for one year, with the response defined as alive ($Y_i = 0$) or dead ($Y_i = 1$). A total of 20 patients refused to be randomized, insisting on receiving the placebo.

  The observed data are provided in the following table:

	Y=0	Y=1	total
X=0	27	43	70
X=1	10	40	50
total	37	83	120

Measures of Frequency

- For now, ignore treatment ...
- *Risk* of death: $P(Y_i = 1)$,

$$\widehat{P}(Y_i = 1) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{83}{120} = 0.692$$

- *Odds* of death,

$$\begin{aligned} \text{Odds}_i &= \frac{P(Y_i = 1)}{P(Y_i = 0)} \\ \widehat{\text{Odds}}_i &= \frac{83/120}{37/120} = 2.24 \end{aligned}$$

Measures of Frequency (continued)

- *Odds* is sometimes used to estimate *risk*

π	$\pi/(1-\pi)$
0.02	0.020
0.04	0.042
0.06	0.064
0.08	0.090
0.1	0.111
0.2	0.250
0.3	0.429
0.4	0.667
0.5	1

Measures of Association

- Returning to the liver cancer example, we now focus on comparing the treatment and placebo groups

X	$\widehat{\pi}_j$
0	43/70=0.61
1	40/50=0.80

where $\pi_j \equiv P(Y_i = 1|X_i = j)$

- Risk ratio

$$
\begin{aligned}
RR &= \frac{\pi_1}{\pi_0} \\
\widehat{RR} &= \frac{\widehat{\pi}_1}{\widehat{\pi}_0} = 1.31
\end{aligned}
$$

  - excess relative risk: $(RR - 1) \times 100\%$
    in our example: 31%

- Risk difference:

$$
\begin{aligned}
RD &= \pi_1 - \pi_0 \\
\widehat{RD} &= 0.8 - 0.61 = 0.19
\end{aligned}
$$

Difference versus Ratio

- Risk difference and ratio may yield very different interpretations of exactly the same data set
- e.g., Suppose a flu vaccine is being evaluated, with the risk of the UM650 virus as given in the following table

X	$\widehat{\pi}_j$
0	0.01
1	0.003

  - $\widehat{RR} =$
  - $\widehat{RD} =$

Difference versus Ratio (continued)

- e.g., Suppose a second flu vaccine is being evaluated, this time with the risk of the UM651 virus given by:

X	$\widehat{\pi}_j$
0	0.9
1	0.7

  - $\widehat{RR} =$
  - $\widehat{RD} =$

Measures of Association (continued)

- Odds ratio:

$$
\begin{aligned}
OR &= \frac{\text{odds}_1}{\text{odds}_0} \\
&= \frac{P(Y_i = 1|X_i = 1)/P(Y_i = 0|X_i = 1)}{P(Y_i = 1|X_i = 0)/P(Y_i = 0|X_i = 0)} \\
&= \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}
\end{aligned}
$$

- e.g., in the liver cancer example,

$$\widehat{OR} \quad =$$

  compare to relative risk: $\widehat{RR} = 1.31$

- The OR is often used to approximate the RR

OR as an Estimator of RR

- How accurately the OR approximates the RR depends on baseline risk
  - consider the table below, where $RR = 1.5$

π_0	OR
0.02	1.51
0.04	1.53
0.06	1.55
0.08	1.57
0.1	1.59
0.2	1.71
0.3	1.91
0.4	2.25
0.5	3.00

Odds Ratio: Further Considerations

- In addition to its relationship with the RR, the OR is often viewed as an interesting measure in its own right
  - OR can be estimated consistently for biased samples (ex. case-control design)
  - OR is easily computed using logistic regression

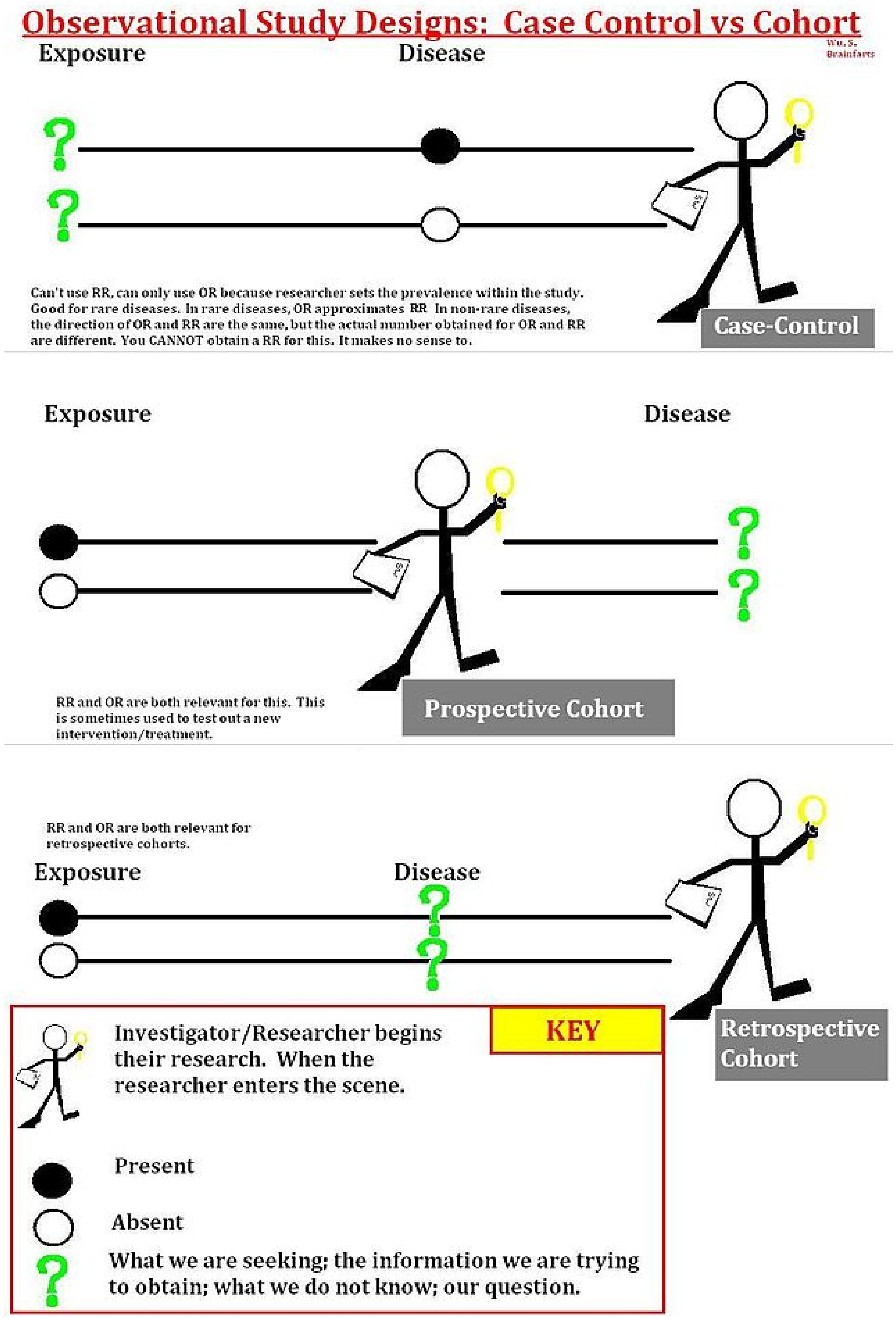
- At this point, it is useful to consider the commonly used study designs . . .

Observational Study: Study Designs

- Cohort Study:
  - subjects sampled independently of outcome status followed (prospectively or retrospectively) to ascertain outcome
  - RR and OR are both relevant
- Case Control Study:
  - subjects sampled based on outcome status
  - e.g., select 100 *cases* ($Y_i = 1$) and 300 *controls* ($Y_i = 0$) then, obtain treatment/exposure information
  - often used when studying rare diseases
  - Can't use RR
- Cross-sectional Study:
  - both covariate and outcome status are obtained at the same time point often, a common calendar date
  - RR and OR are both relevant

Study Designs: Cohort Studies

- A cohort study may be either *prospective* or *retrospective*
  - Prospective cohort: response variate has *not* been observed at the start of the study
  - Retrospective cohort: response variate has already been observed by the time the study began
- Prospective designs are considered to be less prone to bias
- Retrospective studies are often more cost- and time-efficient
  - e.g, using large pre-collected databases


Observational Study Designs: Case Control vs Cohort
Exposure
Disease
Wu, S.
Brainfarts
Can't use RR, can only use OR because researcher sets the prevalence within the study. Good for rare diseases. In rare diseases, OR approximates RR In non-rare diseases, the direction of OR and RR are the same, but the actual number obtained for OR and RR are different. You CANNOT obtain a RR for this. It makes no sense to.
Case-Control
Exposure
Disease
RR and OR are both relevant for this. This is sometimes used to test out a new intervention/treatment.
Prospective Cohort
RR and OR are both relevant for retrospective cohorts.
Exposure
Disease
Retrospective Cohort
KEY
Investigator/Researcher begins their research. When the researcher enters the scene.
Present
Absent
What we are seeking; the information we are trying to obtain; what we do not know; our question.

[wikipedia]

Study Designs: Comparisons

- Simulation: comparison between cohort vs case-control designs
  - Smoking is a risk factor for the colorectal cancer. It can increase the risk twice.
  - Assumptions:
    - Risk for the cancer among non-smoker: 0.05
    - Prevalence of smoking: 20%
  - Studies
    - Cohort study with 10,000 samples
    - Cohort study with 5,000 non-smoker vs 5,000 smokers
    - Case-control study with 5,000 cases vs 5,000 controls.

Study Designs: Comparisons

- Settings:
  - Y=1 (cancer) vs 0 (no-cancer)
  - X=1 (smoker) vs 0 (non-smoker)
  - RR=2
  - $P(X = 1) = 0.2$
  - $P(Y = 1|X = 0) = \pi_0 = 0.05$
  - $P(Y = 1|X = 1) = 0.1$

Cohort Study

- Sample 10,000 healthy individual without considering smoking status, and follow them several years.
- The observed data (after several years of follow up)

	$Y=0$	$Y=1$	total
$X=0$	7613	385	7998
$X=1$	1808	194	2002
total	9421	579	10000

∘ $\widehat{\pi}_0 =$ $\widehat{\pi}_1 =$

∘ $\widehat{RR} =$

∘ $\widehat{OR} =$

Cohort Study: Use exposures

- Sample 5000 healthy smokers and 5000 healthy non-smokers.
- The observed data (after several years of follow up)

	$Y=0$	$Y=1$	total
$X=0$	4748	252	5000
$X=1$	4465	535	5000
total	9213	787	10000

- ○ $\widehat{\pi}_0 =$ $\widehat{\pi}_1 =$

- ○ $\widehat{RR} =$

- ○ $\widehat{OR} =$

Case-Control

- Sample 5000 cancer patients and 5000 heathy controls.
- Investigate their smoking history.

	$Y=0$	$Y=1$	total
$X=0$	4022	3358	7380
$X=1$	978	1642	2620
total	5000	5000	10000

- ◦ $\widehat{\pi}_0 =$ $\widehat{\pi}_1 =$
- ◦ $\widehat{RR} =$
- ◦ $\widehat{OR} =$

Case-Control

- $P(Y = 1|X)$ cannot be estimated, so RR.
- The OR can be accurately estimated
  - use the *Exposure odds ratio*

$$
\begin{aligned}
EOR &= \frac{\text{odds}(X = 1|Y = 1)}{\text{odds}(X = 1|Y = 0)} \\
&= \frac{P(X = 1|Y = 1)}{P(X = 0|Y = 1)} \cdot \frac{P(X = 0|Y = 0)}{P(X = 1|Y = 0)} \\
&= \ldots \\
&= OR
\end{aligned}
$$

Misclassification Bias

- Misclassification:
  - e.g., some subjects with $Y = 1$ are mistakenly classified as $Y = 0$
  - if random, OR is generally biased towards the null

    if non-random, bias can be in either direction
- Examples:
  - recall bias (e.g., case-control study)

Recall bias

- Colorectal cancer example (Case-Control)
- 20 % of previous-smokers without cancer misidentify them as non-smokers.

	Y=0	Y=1	total
X=0	4231	3271	7502
X=1	769	1729	2498
total	5000	5000	10000

◦ $\widehat{OR} =$

Selection Bias

- Selection:
  - Sample obtained is not representative of the population intended to be analyzed
- Key: Does the selected sample accurately represent the target population?
  - if not (resulting from the selection mechanism): *selection bias*

Confounding

- Even in the absence of selection or misclassification, bias can still occur
- e.g., Suppose that there is an *unmeasured* covariate, C
  - *confounding* occurs when:

  (i) C is associated with X

  (ii) C is associated with Y (i.e., adjusting for X)
- Confounding can lead to substantial bias

Example: Confounding

- Example: A study was carried out to investigate the association between alcohol consumption (X_i) and lung cancer Y_i. A random sample of $n = 220$ Ann Arbor residents was classified based on whether they drank alcohol ($X_i = 1$) or not ($X_i = 0$). The cohort was then followed for 30 years and classified based on whether they had been diagnosed with lung cancer ($Y_i = 1$) or not ($Y_i = 0$).

  Observed data are summarized by the following table:

	Y_i=0	Y_i=1	total
X_i=0	91	19	110
X_i=1	19	91	110
total	110	110	220

- Odds ratio: $\widehat{OR} =$

Example: Confounding (continued)

- However, *if* information on smoking status S_i *had been recorded*, the following data would have been observed

  for non-smokers, $S_i = 0$

	Y_i=0	Y_i=1	total
X_i=0	90	9	99
X_i=1	10	1	11
total	100	10	110

  and for smokers, $S_i = 1$

	Y_i=0	Y_i=1	total
X_i=0	1	10	11
X_i=1	9	90	99
total	10	100	110

- The apparent association between alcohol consumption and lung cancer was completely due to *confounding* by smoking