Advanced Methods in Biostatistics

STAT 438

Winter 2022 (1221)[1]

Cameron Roopnarine[2] Yeying Zhu[3]

11th January 2022

[1]Online Course until January 27th, 2022

[2]LaTeXer

[3]Instructor

Contents

Chapter 1

Introduction

WEEK 1
5th to 7th January

About this Course

Three topics covered in this course:

- Causal Inference.
- Missing Data.
- Measurement Error.

Basics in Biostatistics

Review:

- Experimental Studies vs. Observational Studies.
- Statistics of Interest.
- Using Regression Models.
- Association vs. Causation.

Research Questions

Questions to ask when studying a disease:

- Which factors are associated with a given disease? These so-called risk factors are sometimes referred to as predictors, explanatory variables, covariates, independent variables, or exposure variables, etc.
- Which factors are associated with the duration of a given disease?
- Correlation (Association) does not imply causation.
- Ultimately, we want to ask: which factors cause the disease, or which factors determine the duration of the disease?

Types of Studies

- Experimental studies.
- Observational studies.

1.1 Experimental Studies

- In an experimental study, the investigator can manipulate the main (risk) factor of interest, while controlling for other factors.
- In a randomized experimental study, such as a clinical trial, eligible people are randomly assigned to one of two or more groups. One group receives the treatment (such as a new drug) while the control group receives nothing or an inactive placebo.
- Due to randomization, the investigator can control for both known and unknown factors, while investigating, typically, a treatment comparison.

Randomization and Causal Inference:

- Randomization is the perfect/golden design for causal inference.
- Random assignment of treatment (exposure) ensures balance across study arms with respect to observed and unobserved risk factors.
- Direct comparisons between treatment groups can be made.
- Any difference can be attributed to the causal effect of treatment.
- Randomization is not always feasible due to ethical/economic reasons.
- Even the treatment is randomized, the participant may not comply with the assigned treatment: compliance issue.

1.2 Observational Studies

- These studies are typically based on sampling populations with subsequent measurement of various factors of interest. In this setting, we cannot even take advantage of a naturally occurring experiment that changed risk factor status conveniently.
- It is sometimes useful to use these studies to look at the natural history of a disease, but any attempt to identify causality between a risk factor and outcome must be done with great caution.
- There is no experimental setting, as study participants typically self-reflect their exposure categories. Nevertheless, in large part due to ethics, such studies are most often to what we have access in Biostatistics.

Examples of Observational Studies

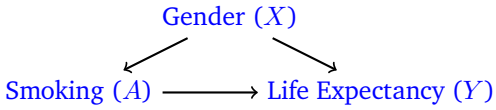
1. – **Risk factor**: cigarette smoking.
   – **Outcome**: bladder cancer.
2. – **Risk factor**: distance of home from hazardous waste site.
   – **Outcome**: respiratory disease.

- Three most popular observational studies:

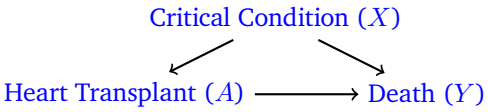
1. Cross-sectional studies.
2. Cohort studies.
3. Case-control studies.

- No control over which subjects have the exposure and which do not.
- Exposed and Unexposed groups may be quite different with respect to other subject characteristics.
- Differences in the outcome are not only due to the (risk) factor of interest, but also because of the masking effect of other covariates (confounders).

Confounding Issue



Another Example of Confounding



1.2.1 Cross-sectional Studies

- Individuals are selected from the target population and their status with respect to the risk factor and the disease status is ascertained at the same time.
- The data represents a snapshot view of the relation between the risk factor and the event occurrence.
- Surveys are often cross-section in nature where associations are of interest and less priority is given to establishing causation.
- Advantage: cross-sectional studies are typically short.
- Disadvantage: a serious problem with such cross-sectional studies is the inability to determine whether the disease outcome or the risk factor occurred first, again this makes causal inferences more problematic or almost impossible.

1.2.2 Cohort Studies

- Cohort studies typically include obtaining two groups from a pre-determined # of individuals, one possessing and the other not possessing a risk factor of interest. Subsequent counts of cases (and non-cases) of a disease of interest are then recorded.
- Much more often than not, cohort studies are prospective, but there are retrospective (or historical) cohort studies as well.

Table representing simple cohort study with sampling based on risk-factor status:

	Disease		
Risk Factor	Present (D)	Absent (D^c)	Total
Present (E)	a	b	n_1
Absent (E^c)	c	d	n_2

- $a \sim \text{BIN}\big(n_1, \mathbb{P}(D \mid E)\big)$.
- $c \sim \text{BIN}\big(n_2, \mathbb{P}(D \mid E^c)\big)$.

1.2.3 Case-control Studies

- In case-control studies, the direction of sampling differs from that of cohort studies. Specifically, the investigator selects a pre-determined # of disease cases and non-cases (i.e., controls), then looks retrospectively to see the # of individuals with and without the risk factor in each group.
- Case-control studies are retrospective studies.

Table representing simple case-control study with sampling based on disease status:

	Disease	
Risk Factor	Present	Absent
Present	a	b
Absent	c	d
Total	n_1	n_2

- $a \sim \text{BIN}\big(n_1, \mathbb{P}(E \mid D)\big)$.
- $b \sim \text{BIN}\big(n_2, \mathbb{P}(E \mid D^c)\big)$.

WEEK 2
10th to 14th January

Statistics of Interest

- Relative Risk.
- Excess Risk.
- Odds Ratio.
- Others: such as attributable risk, hazard ratio.

1.3 Relative Risk

The **relative risk** (RR) of an outcome (e.g., disease) D associated with a binary risk factor E is:

$$\text{RR} = \frac{\mathbb{P}(D \mid E)}{\mathbb{P}(D \mid E^c)},$$

where $0 \leq \text{RR} < \infty$.

Remarks:

(1) The upper limit in practice typically will have a finite constraint. Noting that $\mathbb{P}(D \mid E) \leq 1$, we have

$$\text{RR} \leq \frac{1}{\mathbb{P}(D \mid E^c)} < \infty,$$

assuming $\mathbb{P}(D \mid E^c) \neq 0$.

(2) If there exists absolutely no association between D and E, this results in $\text{RR} = 1$, that is, this will

happen when $\mathbb{P}(D \mid E) = \mathbb{P}(D \mid E^c)$.

(3) If RR > 1, there is greater risk or probability of D when E is present versus absent.

(4) If RR < 1, there is lower risk or probability of D when E is present versus absent.

RR Calculation

- Recall the table for a cohort study.

Risk Factor	*Disease* Present (D)	Absent (D^c)	Total
Present (E)	a	b	n_1
Absent (E^c)	c	d	n_2

  Then,
$$\widehat{\text{RR}} = \frac{a/(a+b)}{c/(c+d)} = \frac{a/n_1}{c/n_2}.$$

- To make inference, we have, approximately,
$$\log(\widehat{\text{RR}}) \sim \mathcal{N}\Big(\log(\text{RR}), \text{Var}\big(\log(\text{RR})\big)\Big),$$
  where
$$\text{Var}\big(\log(\text{RR})\big) \approx \frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}.$$

- The (approximate) 95% confidence interval for $\log(\text{RR})$ is
$$\log(\widehat{\text{RR}}) \pm 1.96\sqrt{\widehat{\text{Var}}\big(\log(\widehat{\text{RR}})\big)}.$$

- The (approximate) 95% confidence interval for RR is:
$$\exp\Big\{\log(\widehat{\text{RR}}) \pm 1.96\sqrt{\widehat{\text{Var}}\big(\log(\widehat{\text{RR}})\big)}\Big\}$$

For RR, we have
$$\text{Var}\big(\log(\widehat{\text{RR}})\big) \approx \frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}.$$

Proof: Define $\hat{p}_a = \frac{a}{n_1}$ and $\hat{p}_c = \frac{c}{n_2}$. Assuming the exposed and unexposed groups are independent, we have
$$\begin{aligned}\text{Var}\big(\log(\widehat{\text{RR}})\big) &= \text{Var}\big(\log(\hat{p}_a) - \log(\hat{p}_c)\big) \\ &= \text{Var}\big(\log(\hat{p}_a)\big) - \text{Var}\big(\log(\hat{p}_c)\big).\end{aligned}$$
Using, Taylor's approximation, we have
$$\begin{aligned}\log(\hat{p}_a) &\approx \log(p_a) + \frac{\mathrm{d}\log(p_a)}{\mathrm{d}p_a}(\hat{p}_a - p_a) \\ &= \log(p_a) + \frac{(\hat{p}_a - p_a)}{p_a}.\end{aligned}$$

Since $a \sim \text{BIN}(n_1, p_a)$,

$$\begin{aligned}\mathsf{Var}\big(\log(\hat{p}_a)\big) &\approx \frac{\mathsf{Var}(\hat{p}_a)}{p_a^2} \\ &= \frac{\mathsf{Var}\left(\frac{a}{n_1}\right)}{p_a^2} \\ &= \frac{n_1 p_a(1-p_a)}{n_1^2 p_a^2} \\ &= \frac{1-p_a}{n_1 p_a}.\end{aligned}$$

Therefore,

$$\widehat{\mathsf{Var}}\big(\log(\hat{p}_a)\big) = \frac{1-\hat{p}_a}{n_1\hat{p}_a} = \frac{b}{a(a+b)}.$$

Similarly,

$$\widehat{\mathsf{Var}}\big(\log(\hat{p}_c)\big) = \frac{d}{c(c+d)}.$$

Therefore,

$$\begin{aligned}\widehat{\mathsf{Var}}\big(\log(\widehat{\text{RR}})\big) &= \frac{b}{a(a+b)} + \frac{d}{c(c+d)} \\ &= \frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}.\end{aligned}$$

Remarks:

(1) Relative risk (or sometimes called **risk ratio**) is a common measure of the disease-exposure association from cohort studies.

(2) In general, the relative risk is *not* symmetric in the role of D and E, that is,

$$\frac{\mathbb{P}(D \mid E)}{\mathbb{P}(D \mid E^c)} \neq \frac{\mathbb{P}(E \mid D)}{\mathbb{P}(E \mid D^c)}.$$

1.4 Excess Risk

While RR is a relative measure of risk, it is sometimes of interest to look at absolute measures of risk. One such measure is *excess risk*.

The **excess risk** (ER) is:

$$\text{ER} = \mathbb{P}(D \mid E) - \mathbb{P}(D \mid E^c),$$

where $-1 \leq \text{ER} \leq 1$.

Remark:

(1) ER $= 0$ means no excess risk (null value).

(2) ER > 0 means greater risk of D for E versus E^c.

(3) ER < 0 means lower risk of D for E versus E^c.

ER Calculation

- Recall the table for a cohort study.

Risk Factor	*Disease* Present (D)	Absent (D^c)	Total
Present (E)	a	b	n_1
Absent (E^c)	c	d	n_2

  Then,
$$\widehat{\text{ER}} = \frac{a}{a+b} - \frac{c}{c+d} = \hat{p}_a - \hat{p}_c.$$

- To make inference, we have, approximately,
$$\widehat{\text{ER}} \sim \mathcal{N}\big(\text{ER}, \text{Var}\big(\widehat{\text{ER}}\big)\big),$$
  where
$$\text{Var}(\widehat{\text{ER}}) \approx \frac{ab}{(a+b)^3} + \frac{cd}{(c+d)^3}.$$

- The (approximate) 95% confidence interval for ER is:
$$\widehat{\text{ER}} \pm 1.96\sqrt{\widehat{\text{Var}}\big(\widehat{\text{ER}}\big)}.$$

For ER, we have
$$\text{Var}(\widehat{\text{ER}}) \approx \frac{ab}{(a+b)^3} + \frac{cd}{(c+d)^3}.$$

Proof: Define $\hat{p}_a = \frac{a}{n_1}$ and $\hat{p}_c = \frac{c}{n_2}$. Note that $a \sim \text{BIN}(n_1, p_a)$ and $c \sim \text{BIN}(n_2, p_c)$. Hence,
$$\begin{aligned}
\text{Var}(\widehat{\text{ER}}) &= \text{Var}(\hat{p}_a - \hat{p}_c) \\
&= \text{Var}(\hat{p}_a) + \text{Var}(\hat{p}_c) \\
&= \text{Var}\left(\frac{a}{n_1}\right) + \text{Var}\left(\frac{c}{n_2}\right) \\
&= \frac{p_a(1-p_a)}{n_1} + \frac{p_c(1-p_c)}{n_2}.
\end{aligned}$$

Therefore,
$$\begin{aligned}
\widehat{\text{Var}}(\widehat{\text{ER}}) &= \frac{\hat{p}_a(1-\hat{p}_a)}{n_1} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_2} \\
&= \frac{ab}{(a+b)^3} + \frac{cd}{(c+d)^3}.
\end{aligned}$$

1.5 Odds Ratio

The **odds** of disease for the *exposed group* is

$$\frac{\mathbb{P}(D \mid E)}{\mathbb{P}(D^c \mid E)} = \frac{\mathbb{P}(D \mid E)}{1 - \mathbb{P}(D \mid E)}.$$

The **odds** of disease for the *unexposed group* is

$$\frac{\mathbb{P}(D \mid E^c)}{\mathbb{P}(D^c \mid E^c)} = \frac{\mathbb{P}(D \mid E^c)}{1 - \mathbb{P}(D \mid E^c)}.$$

The **odds ratio** for measuring the association of disease with the exposed versus unexposed groups is

$$\text{OR} = \frac{\mathbb{P}(D \mid E)/\,\mathbb{P}(D^c \mid E)}{\mathbb{P}(D \mid E^c)/\,\mathbb{P}(D^c \mid E^c)} = \frac{\mathbb{P}(D \mid E)/[1 - \mathbb{P}(D \mid E)]}{\mathbb{P}(D \mid E^c)/[1 - \mathbb{P}(D \mid E^c)]}.$$

Remarks:

- OR $= 1$ means no association between D and E.
- OR > 1 means greater odds of disease when E is present.
- OR < 1 means lower odds of disease when E is present.

OR Calculation

- For general study with binary disease and exposure (risk factor):

Risk Factor	*Disease* Present (D)	Absent (D^c)
Present (E)	a	b
Absent (E^c)	c	d

  Here,

$$\widehat{\text{OR}} = \frac{\mathbb{P}(D \mid E)/\,\mathbb{P}(D^c \mid E)}{\mathbb{P}(D \mid E^c)/\,\mathbb{P}(D^c \mid E^c)} = \frac{\left(\frac{a}{a+b}\right)/\left(\frac{b}{a+b}\right)}{\left(\frac{c}{c+d}\right)/\left(\frac{d}{c+d}\right)} = \frac{ad}{bc}.$$

- To make inference, we have approximately,

$$\log(\widehat{\text{OR}}) \sim \mathcal{N}\big(\log(\text{OR}), \text{Var}\big(\log(\widehat{\text{OR}})\big)\big),$$

  where

$$\text{Var}\big(\log(\widehat{\text{OR}})\big) \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}.$$

Remark: OR is symmetric in roles of D and E:

$$\frac{\mathbb{P}(E \mid D)/\,\mathbb{P}(E^c \mid D)}{\mathbb{P}(E \mid D^c)/\,\mathbb{P}(E^c \mid D^c)} = \frac{\left(\frac{a}{a+c}\right)/\left(\frac{c}{a+c}\right)}{\left(\frac{b}{b+d}\right)/\left(\frac{d}{b+d}\right)} = \frac{ad}{bc}.$$

Therefore, the OR for D associated with E is equal to the OR for E associated with D. It is this symmetry that makes OR a popular "risk" measure for case-control studies, where sampling is done on disease status, not risk factor status.

1.6 Comments

The various types of probabilities that may be of interest:

- Joint probabilities: $\mathbb{P}(D, E)$, $\mathbb{P}(D, E^c)$, $\mathbb{P}(D^c, E)$, and $\mathbb{P}(D^c, E^c)$.
- Marginal probabilities: $\mathbb{P}(D)$, $\mathbb{P}(E)$, $\mathbb{P}(D^c)$, and $\mathbb{P}(E^c)$.
- Conditional probabilities: $\mathbb{P}(D \mid E)$, $\mathbb{P}(D \mid E^c)$, $\mathbb{P}(E \mid D)$, and $\mathbb{P}(E \mid D^c)$.

Cross-sectional Study:

- All the above probabilities can be estimated by the observed proportions if the sampling is simple random sampling.

Cohort Study:

- $\mathbb{P}(D \mid E)$, $\mathbb{P}(D^c \mid E)$, $\mathbb{P}(D \mid E^c)$, and $\mathbb{P}(D^c \mid E^c)$ can be estimated.
- Marginal probabilities $\mathbb{P}(D)$, $\mathbb{P}(E)$, and joint probabilities such as $\mathbb{P}(D, E)$ cannot be estimated.
- RR, ER, and OR can be estimated.

Case-control Study:

- Only $\mathbb{P}(E \mid D)$, $\mathbb{P}(E^c \mid D)$, $\mathbb{P}(E^c \mid D^c)$, and $\mathbb{P}(E \mid D^c)$ can be estimated.
- RR and ER cannot be estimated.
- OR can be estimated. Furthermore, RR $\approx$ OR when the disease is rare.

If the disease is rare in a case-control study (i.e., $\mathbb{P}(D) \approx 0$), we have RR $\approx$ OR.

Proof:

$$\begin{aligned}
\text{OR} &= \frac{\mathbb{P}(D \mid E)/\mathbb{P}(D^c \mid E)}{\mathbb{P}(D \mid E^c)/\mathbb{P}(D^c \mid E^c)} \\
&= \frac{\mathbb{P}(D \mid E)}{\mathbb{P}(D \mid E^c)} \frac{\overbrace{\mathbb{P}(D^c \mid E^c)}^{\approx 1}}{\underbrace{\mathbb{P}(D^c \mid E)}_{\approx 1}} \\
&\approx \frac{\mathbb{P}(D \mid E)}{\mathbb{P}(D \mid E^c)} \\
&= \text{RR}.
\end{aligned}$$

1.7 Regression Models

- Linear model.
- Log-linear model.
- Probit model.
- Logistic regression model.

Notation:

- X: exposure variable of interest.
- D: disease status.
- P_x: $\mathbb{P}(D = 1 \mid X = x)$, that is, how the risk of disease changes according to the exposure variable.

1.7.1 Linear Model

$$P_x = \mathbb{P}(D = 1 \mid X = x) = \alpha + \beta x.$$

- $\alpha = P_{x=0}$: the baseline risk.
- $\beta = P_{x+1} - P_x$: excess risk with 1 unit increase in exposure.

Drawbacks:

(1) Possible to produce $\hat{P}_x < 0$ or $\hat{P}_x > 1$.

(2) Can't be directly applied to case-control data.

1.7.2 Log-Linear Model

$$\log(P_x) = \log\big(\mathbb{P}(D = 1 \mid X = x)\big) = \alpha + \beta x.$$

- $\alpha = \log(P_{x=0})$: the log baseline risk.
- β: log relative risk associated with 1 unit increase in exposure.

Drawbacks:

(1) Possible to produce $\hat{P}_x > 1$.

(2) Can't be directly applied to case-control data.

1.7.3 Probit Model

$$P_x = \mathbb{P}(D = 1 \mid X = x) = \Phi(\alpha + \beta x),$$

where $\Phi(u)$ is the cdf of a standard normal distribution.

- $\alpha = \Phi^{-1}(P_{x=0})$.
- $\beta > 0$: the risk increases as X increases.
  $\beta < 0$: the risk increases as X decreases.

Drawbacks:

(1) There is no natural interpretation of α and α in terms of association.

(2) Can't be directly applied to case-control data.

1.7.4 Logistic Regression Model

$$P_x = \mathbb{P}(D = 1 \mid X = x) = \frac{1}{1 + \exp\big\{-(\alpha + \beta x)\big\}}.$$

- $\alpha = \log\left(\frac{P_{x=0}}{1-P_{x=0}}\right)$: the log odds of disease at baseline.
- β: log odds ratio associated with 1 unit increase in exposure.

Advantages:

(1) $0 < \hat{P}_x < 1$.

(2) $\exp\{\beta\}$: the odds ratio, which is symmetric with respect to D and E if both are binary.

(3) Can be applied to case-control data.

Remarks:

(1) "Correlation does not imply causation."

(2) Regression models tell us correlational/associational relationship between the exposure and the disease outcome

(3) *Conclusion*: We need better tools to define causality

(4) *Solution*: Potential outcomes framework (Chapter 2).