Study Design: Overview

Outline

- Introduction to study design
- Measures of effect commonly used for categorical data
- Cohort Studies
- Case-control studies
- Intervention studies (clinical trials)
- Alternative study designs
- Confounding, interaction, and standardization
  - Mantel-Haenszel method

Exposure-Disease Relationship

Hypothetical exposure–disease relationship

		Disease: Yes	Disease: No	
Exposure	Yes	a	b	$a + b = n_1$
	No	c	d	$c + d = n_2$
		$a + c = m_1$	$b + d = m_2$	

Study Design

- Prospective study
  - A group of disease-free individuals are identified at one point in time and are followed over a period of time until some of them develop the disease. The development of disease over time is then related to other variables measured at baseline, generally called exposure variables. The study population in a prospective study is often called a cohort.

Study Design cont.

- Retrospective study (case-control study)
  - Two groups of individuals are initially identified
    - A group that has the disease under study (the case)
    - A group that does not have the disease (the control)
    - An attempt is then made to relate their prior health habits to their current disease status.
- Cross-sectional study
  - A study population is ascertained at one point in time. All participants in the study population are asked about their current disease status and their current or past exposure status.
  - Also called prevalence study
    - Prevalence of disease at one point in time is compared between exposed and unexposed individuals.
    - In prospective study, one is interested in the incidence rather than the prevalence.

Intervention Studies

- Intervention study, or clinical trial, is an experiment applied to
  - existing patients, in order to decide upon an appropriate therapy,
  - those presently free of symptoms, in order to decide upon an appropriate preventive strategy.
- Giving treatments to the subjects in the study.
  - Drugs
  - Hospital procedures
  - Field trials of vaccines
- Control groups are often used
  - Placebo group.

Study Design cont.

- A study looked at the effects of oral contraceptive (OC) use on heart disease in women 40 to 44 years of age. It found that among 5000 current OC users at baseline, 13 women develop a myocardial infarction (MI) over a 3-year period, whereas among 10,000 non-OC users, 7 develop an MI over a 3-year period.
- This is a prospective design.
  - All patients are disease free at baseline and had their exposure (OC use) measured at that time.
  - They were followed for 3 years, during which some developed disease, and others remained disease free.

- A hypothesis: an important factor for breast cancer is age at first birth.
- An international study was set up to test the hypothesis.
  - Breast cancer cases were identified among women in selected hospitals in the United States, Greece, Yugoslavia, Brazil, and Japan.
  - Controls were chosen from women of comparable age who were in the hospital at the same time as the cases, but who did not have breast cancer.
  - All women were asked about their age at first birth.
  - The set of women with at least one birth was arbitrarily divided into two categories:
    - Women whose age at first birth ≤ 29
    - Women whose age at first birth ≥ 30
- This is a retrospective study.
  - Breast-cancer cases were identified together with controls who were in the hospital at the same time as the cases but who did not have breast cancer and were of comparable age to the controls.
  - Pregnancy history (age at first birth) of cases and controls was compared.

- Suppose a study is performed concerning infant blood pressure. All infants born in a specific hospital are ascertained within the first week of life while in the hospital and have their blood pressure measured in the newborn nursery. The infants are divided into two groups: a high-blood-pressure group, if their blood pressure is in the top 10% of infant blood pressure based on national norms, and a normal-blood-pressure group, otherwise. The infants blood-pressure group is then related to their birthweight (low if <88 oz and normal otherwise).
- This is an example of a cross-sectional study.
  - The blood pressures and birthweights are measured at approximately the same point in time.
  - Prevalence is known.

Study Design

- Prospective study is more definitive
  - Gold standard
- Retrospective study has a greater chance of bias
  - Selection bias
    - A milder series of case participants who are still alive is used.
    - Control selection is related, often unexpected, to the exposure.
  - Recall bias
- Retrospective study is much less expensive.
  - Often used as preliminary steps
- Cross-sectional studies have the same problems as case-control studies.

Risk Difference

Let

p_1 = probability of developing disease for exposed individuals

p_2 = probability of developing disease for unexposed individuals

The **risk difference** is defined as $p_1 - p_2$. The **risk ratio** or **relative risk** is defined as p_1/p_2.

$$\hat{p}_1 \sim N(p_1, p_1q_1/n_1), \hat{p}_2 \sim N(p_2, p_2q_2/n_2)$$

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}\right)$$

Therefore, if $p_1q_1/n_1 + p_2q_2/n_2$ is approximated by $\hat{p}_1\hat{q}_1/n_1 + \hat{p}_2\hat{q}_2/n_2$, then

$$Pr\left(p_1 - p_2 - z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \le \hat{p}_1 - \hat{p}_2 \le p_1 - p_2 + z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\right) = 1 - \alpha$$

$$p_1 - p_2 \le \hat{p}_1 - \hat{p}_2 + z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$\text{and } \hat{p}_1 - \hat{p}_2 - z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \le p_1 - p_2$$

Risk Ratio

$$\hat{RR} = \hat{p}_1 / \hat{p}_2$$

Assuming the normal approximation to the binomial distribution is valid.

Sampling distribution of $\ln(\hat{RR})$ more closely follows a normal distribution than $\hat{RR}$.

$$\begin{aligned} Var[\ln(\hat{RR})] &= Var[\ln(\hat{p}_1) - \ln(\hat{p}_2)] \\ &= Var[\ln(\hat{p}_1)] + Var[\ln(\hat{p}_2)] \end{aligned}$$

Delta Method The variance of a function of a random variable $f(X)$ is approximated by

$$Var[f(X)] \cong [f'(X)]^2 Var(X)$$

Use the delta method to find the variance of $\ln(\hat{p}_1)$, $\ln(\hat{p}_2)$, and $\ln(\widehat{RR})$.

In this case $f(X) = \ln(X)$. Because $f'(X) = \dfrac{1}{X}$, we obtain

$$Var[\ln(\hat{p}_1)] \cong \frac{1}{\hat{p}_1^2} Var(\hat{p}_1) = \frac{1}{\hat{p}_1^2}\left(\frac{\hat{p}_1\hat{q}_1}{n_1}\right) = \frac{\hat{q}_1}{\hat{p}_1 n_1}$$

$\hat{p}_1 = a/n_1$, $\hat{q}_1 = b/n_1$. Therefore,

$$Var[\ln(\hat{p}_1)] = \frac{b}{an_1}$$

Also, using similar methods,

$$Var[\ln(\hat{p}_2)] = \frac{\hat{q}_2}{\hat{p}_2 n_2} = \frac{d}{cn_2}$$

It follows that

$$Var[\ln(\widehat{RR})] = \frac{b}{an_1} + \frac{d}{cn_2}$$

or $$se[\ln(\widehat{RR})] = \sqrt{\frac{b}{an_1} + \frac{d}{cn_2}}$$

$$\left[\ln(\widehat{RR}) - z_{1-\alpha/2}\sqrt{\frac{b}{an_1} + \frac{d}{cn_2}},\quad \ln(\widehat{RR}) + z_{1-\alpha/2}\sqrt{\frac{b}{an_1} + \frac{d}{cn_2}}\right]$$

Odds Ratio

If the probability of a success = p, then the **odds in favor of success** = $p/(1-p)$.

Let p_1, p_2 be the underlying probability of success for two groups. The **odds ratio** (***OR***) is defined as

$$OR = \frac{p_1/q_1}{p_2/q_2} = \frac{p_1 q_2}{p_2 q_1} \qquad \text{and is estimated by} \qquad \hat{OR} = \frac{\hat{p}_1 \hat{q}_2}{\hat{p}_2 \hat{q}_1}$$

Equivalently, if the four cells of the 2×2 contingency table are labeled by a, b, c, d, as they are in Table 13.1, then

$$\hat{OR} = \frac{[a/(a+b)] \times [d/(c+d)]}{[c/(c+d)] \times [b/(a+b)]} = \frac{ad}{bc}$$

The **disease-odds ratio** is the odds in favor of disease for the exposed group divided by the odds in favor of disease for the unexposed group.

The **exposure-odds ratio** is the odds in favor of being exposed for diseased subjects divided by the odds in favor of being exposed for nondiseased subjects.

Odds Ratio

- Odds ratio greater than 1
  - a greater likelihood of disease among the exposed than among the unexposed
- Odd ratio less than 1
  - A greater likelihood of disease among the unexposed than among the exposed
- If the disease is rare odds ratio will be approximately the same as the relative risk.

- Odds ratio is particularly useful for case-control studies since we cannot directly estimate either the risk difference or the risk ratio.
- Let *A*, *B*, *C*, *D* represent the true number of subjects in the reference population, corresponding to cells *a*, *b*, *c*, and *d* in our sample.

Hypothetical exposure–disease relationships in a sample and a reference population

		Sample			Population		
		Disease			Disease		
		Yes	No		Yes	No
Exposed	Yes	*a*	*b*	Exposed	Yes	*A*	*B*
	No	*c*	*d*		No	*C*	*D*

- In a case-control study, we assume a random fraction f_1 of subjects with disease and a random faction f_2 of subjects without disease in the reference population are included in our study sample.

$$\hat{RR} = \frac{a/(a+b)}{c/(c+d)}$$

$$= \frac{f_1 A/(f_1 A + f_2 B)}{f_1 C/(f_1 C + f_2 D)}$$

$$= \frac{A/(f_1 A + f_2 B)}{C/(f_1 C + f_2 D)}$$

The true relative risk in the reference population is

$$RR = \frac{A/(A+B)}{C/(C+D)}$$

$$\hat{OR} = \frac{ad}{bc}$$

$$= \frac{f_1 A(f_2 D)}{f_2 B(f_1 C)}$$

$$= \frac{AD}{BC} = OR$$

Point and Interval Estimation for the Odds Ratio (Woolf Procedure) Suppose we have a 2×2 contingency table relating exposure to disease, with cell counts a, b, c, d as given in Table 13.1.

(1) A point estimate of the true odds ratio (OR) is given by $\hat{OR} = ad/bc$.

(2) An approximate two-sided $100\% \times (1 - \alpha)$ CI for OR is given by (e^{c_1}, e^{c_2}), where

$$c_1 = \ln(\hat{OR}) - z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

$$c_2 = \ln(\hat{OR}) + z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

(3) In a prospective or a cross-sectional study, the CI in (2) should only be used if $n_1\hat{p}_1\hat{q}_1 \geq 5$ and $n_2\hat{p}_2\hat{q}_2 \geq 5$ where

n_1 = the number of exposed individuals

$\hat{p}_1$ = sample proportion with disease among exposed individuals and $\hat{q}_1 = 1 - \hat{p}_1$

n_2 = the number of unexposed individuals

$\hat{p}_2$ = the sample proportion with disease among unexposed individuals, and $\hat{q}_2 = 1 - \hat{p}_2$

(4) In a case–control study, the CI should only be used if $m_1\hat{p}_1^*\hat{q}_1^* \geq 5$ and $m_2\hat{p}_2^*\hat{q}_2^* \geq 5$ where

m_1 = the number of cases

$\hat{p}_1^*$ = the proportion of cases that are exposed, and $\hat{q}_1^* = 1 - \hat{p}_1^*$

m_2 = the number of controls

$\hat{p}_2^*$ = the proportion of controls that are exposed, and $\hat{q}_2^* = 1 - \hat{p}_2^*$

(5) If the disease under study is rare, then $\widehat{OR}$ and its associated $100\% \times (1 - \alpha)$ CI can be interpreted as approximate point and interval estimates of the risk ratio. This is particularly important in case–control studies where no direct estimate of the risk ratio is available.