

[illegible][illegible][illegible]

# Overview

## 1. Common study designs (epidemiology)

- Prospective, retrospective , and cross-sectional study

## 2. Commonly used measures of effect for categorical data

- Risk Difference, Risk Ratio, and Odds Ratio
- Parameters estimation with confidence intervals

## 3. Multiple Logistic regression

# Study Design

**Table 13.1** Hypothetical exposure–disease relationship

		Disease		
		Yes	No	
Exposure	Yes	<i>a</i>	<i>b</i>	$a + b = n_1$
	No	<i>c</i>	<i>d</i>	$c + d = n_2$
		$a + c = m_1$	$b + d = m_2$	

- $n_1 = a + b$  exposed subjects (a have disease)
- $n_2 = c + d$  unexposed subjects (c have disease)
- Three main study designs: prospective, retrospective, and cross-sectional study design
- **Prospective (Cohort) study design:** a group of disease-free individuals is identified at one point in time and are followed over a period of time until some of them develop the disease
  - development of disease over time is then related to other variables (exposure variables) measured at baseline
  - **Cohort:** study population in a prospective study

- **Retrospective (case-control) study:** two groups of individuals are initially identified:

- (1) Cases: a group that has the disease under study

- (2) Controls: a group that does not have the disease under study
    - relate their prior health habits to their current disease status

- **Cross-sectional (Prevalence) study:** study population is ascertained at a single point in time

- Collect current disease status and current or past exposure status from study participants

- Prevalence of disease at one point in time is compared between exposed and unexposed individuals (v.s. prospective study: incidence rather than the prevalence of the disease)



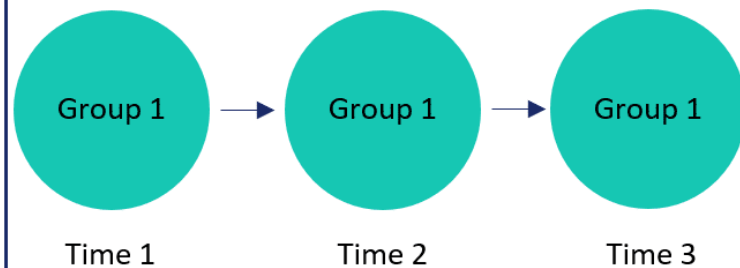
- A **prospective** study is usually more **definitive**
  - patient's knowledge of their **current** health habits is more precise than recall of their past health habits
- A **retrospective** study has a **greater chance of bias** for two reasons
  - **Selection bias**: more **difficult to obtain** a **representative** sample of people who already have the disease in question
  - **Recall bias**: individuals with the disease or their surrogates may tend to give **biased answers** about prior health habits if they believe there is a relationship between these prior health habits and the disease
  - **much less expensive** to perform and can be completed in **much less time** than a prospective study

# Two main study designs

## Cohort Study

### Longitudinal Study

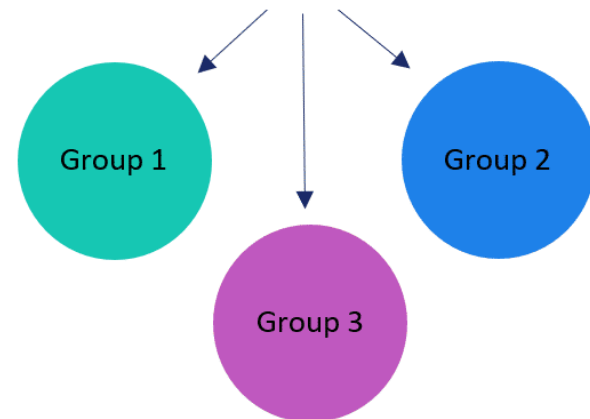
*Same Group*  
Compared over time



## Cross-sectional Study

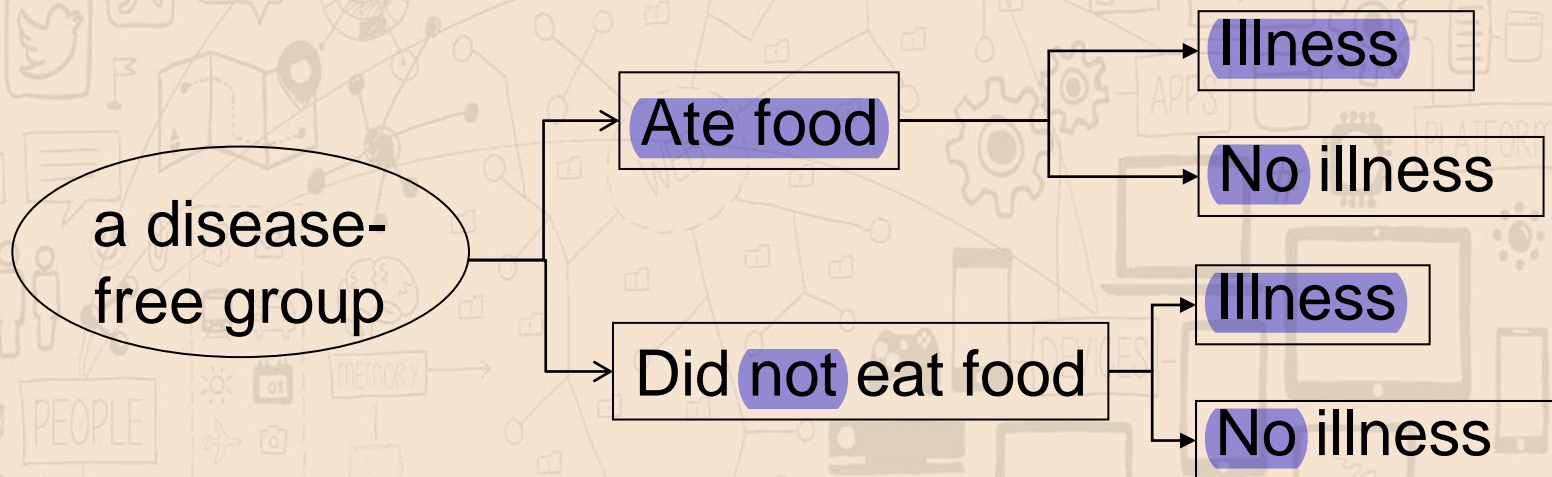
### Cross-Sectional Study

*Different Groups*  
Compared at the same time



# An Example of Cohort Study

- A disease-free group in which outbreak (expected) to occur later
- Compare attack rates among people who ate and did not eat certain food(s)
- Higher attack rates among people eating a food (compared to those not eating it) suggest the food might be associated with illness



# Relative Risk / Risk Ratio (RR)

- Measure of association for a cohort study
- Compares proportion of people who ate the food who became ill with the proportion of people did not eat the food who became ill

$$\text{Relative risk} = \frac{\text{attack rate among exposed}}{\text{attack rate among unexposed}}$$

- Q: How much more likely is it for people who ate the food to become ill than people not eating the food?



# Example: Outbreak of Salmonella at a Hospital

Returning to the outbreak of salmonella:

- 212 (37%) of 571 attending lunch became ill \*exposed\*
- 12 (7%) of 165 not attending lunch became ill \*unexposed\*

$$\text{Relative risk} = \frac{\text{attack rate (attended)}}{\text{attack rate (did not attend)}} = \frac{37\%}{7\%} = 5.3$$

- A relative risk of 5.3 : people who attended the luncheon were about 5 times more likely to become ill than those who did not attend

→ Attending the luncheon might be a risk factor for salmonellosis in this outbreak

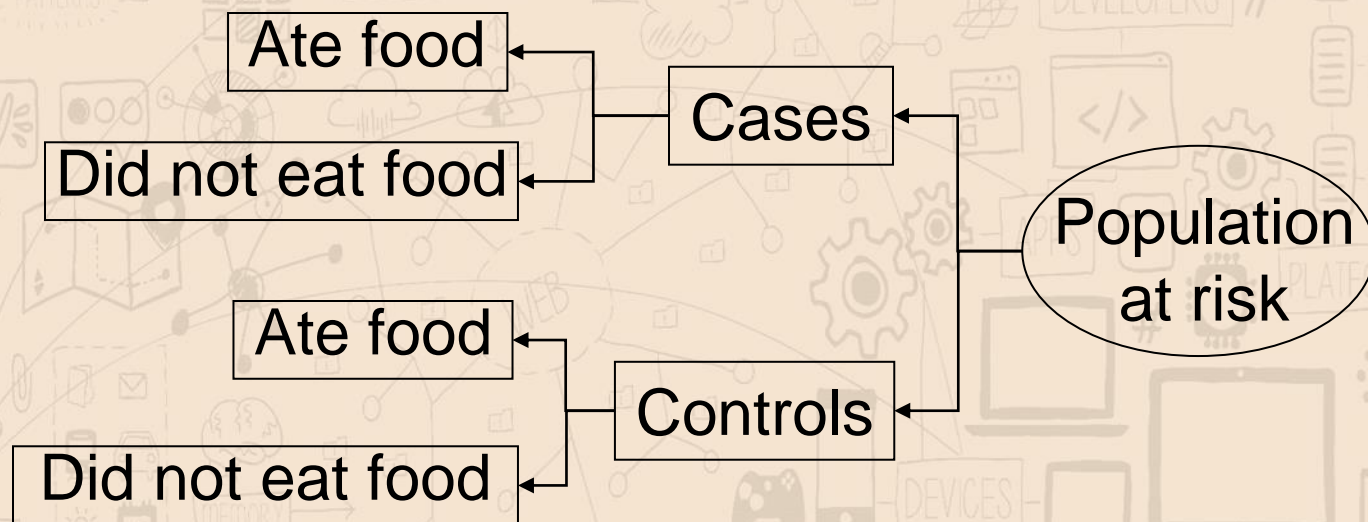


# Magnitudes of Relative Risk

- **Close to 1.0** = risk of disease is **similar among** people eating and not eating the food → food **not associated with illness**
- **Greater than 1.0** = risk of disease is **higher among** people eating the food than people not eating the food → food could be **risk factor**
- **Less than 1.0** = risk of disease is **lower among** people eating the food than people not eating the food → food could be **“protective factor”**
- **Magnitude** reflects strength of association between eating food and illness

# Case-Control Study

- Also called **retrospective study**: 2 groups are identified
- **Cases** (people with **illness**) and **controls** (people with **no illness**)
- **Compare** foods eaten by **cases** and **controls**
- Foods more commonly **eaten by cases than controls** might be **associated with illness**





# Example: Outbreak of Botulism in Vancouver, B.C.

- Two sisters and their mother from Vancouver developed signs and symptoms suggestive of botulism
- 36 cases of botulism among customers of Restaurant X
- Case-control study undertaken
  - 20 (91%) of 22 cases ate beef dip sandwich
  - 3 (14%) of 22 controls ate beef dip sandwich





# Odds Ratio

- Measure of association for a case-control study
- Compares odds of cases having eaten a certain food to odds of controls having eaten the food

$$\text{odds ratio} = \frac{\text{odds of eating food among cases}}{\text{odds of eating food among controls}}$$

- Answers for “How much higher is the odds of eating the food among cases than controls?”

# Odds Ratio

- **Close to 1.0** = odds of eating food is similar among cases and controls → no association between food and illness
- **Greater than 1.0** = odds of eating food among cases is higher than among controls → food could be risk factor
- **Less than 1.0** = odds of eating food among cases is lower than among controls → food could be “protective factor”
- **Magnitude** reflects strength of association between illness and eating the food

> Case-control studies

# Odds Ratio Calculation

- Let  $p$  = probability of a success  $\rightarrow$  the **odds** in favor of success =  $p/(1-p)$
- $p_1, p_2$  : the odds in favor of success are computed for each of two proportions  $\rightarrow$  ratio of odds (**OR**): useful measure for **relating the two proportions**

Table 13.2

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

		Sample				Population	
		Disease				Disease	
		Yes	No			Yes	No
Exposed	Yes	<i>a</i>	<i>b</i>	Exposed	Yes	<i>A</i>	<i>B</i>
	No	<i>c</i>	<i>d</i>		No	<i>C</i>	<i>D</i>

Let  $p_1, p_2$  : underlying probability of success for two groups

OR :

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

Let  $a, b, c, d$ : label of the four cells of the  $2 \times 2$  contingency table

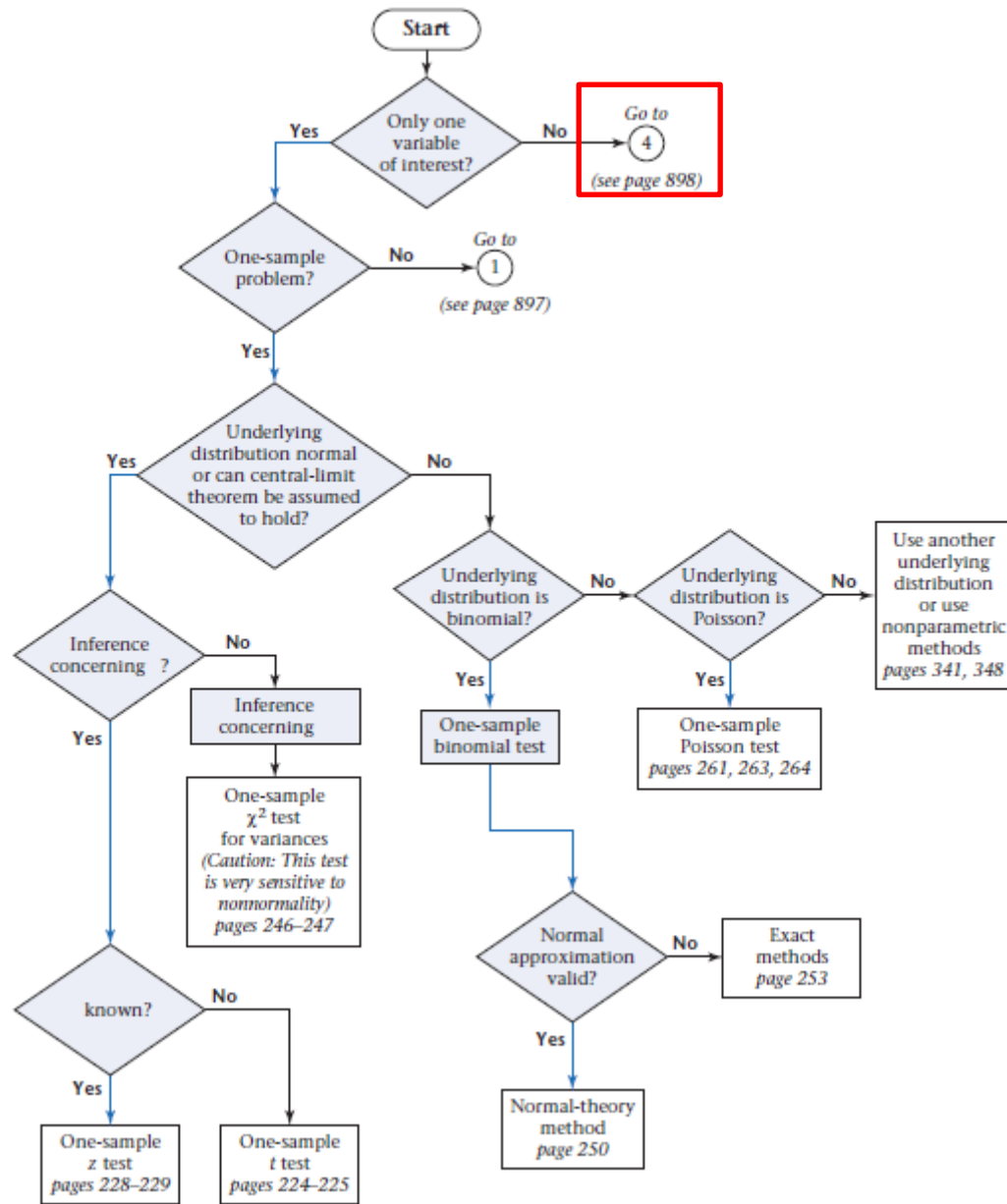
Interpretation of OR:

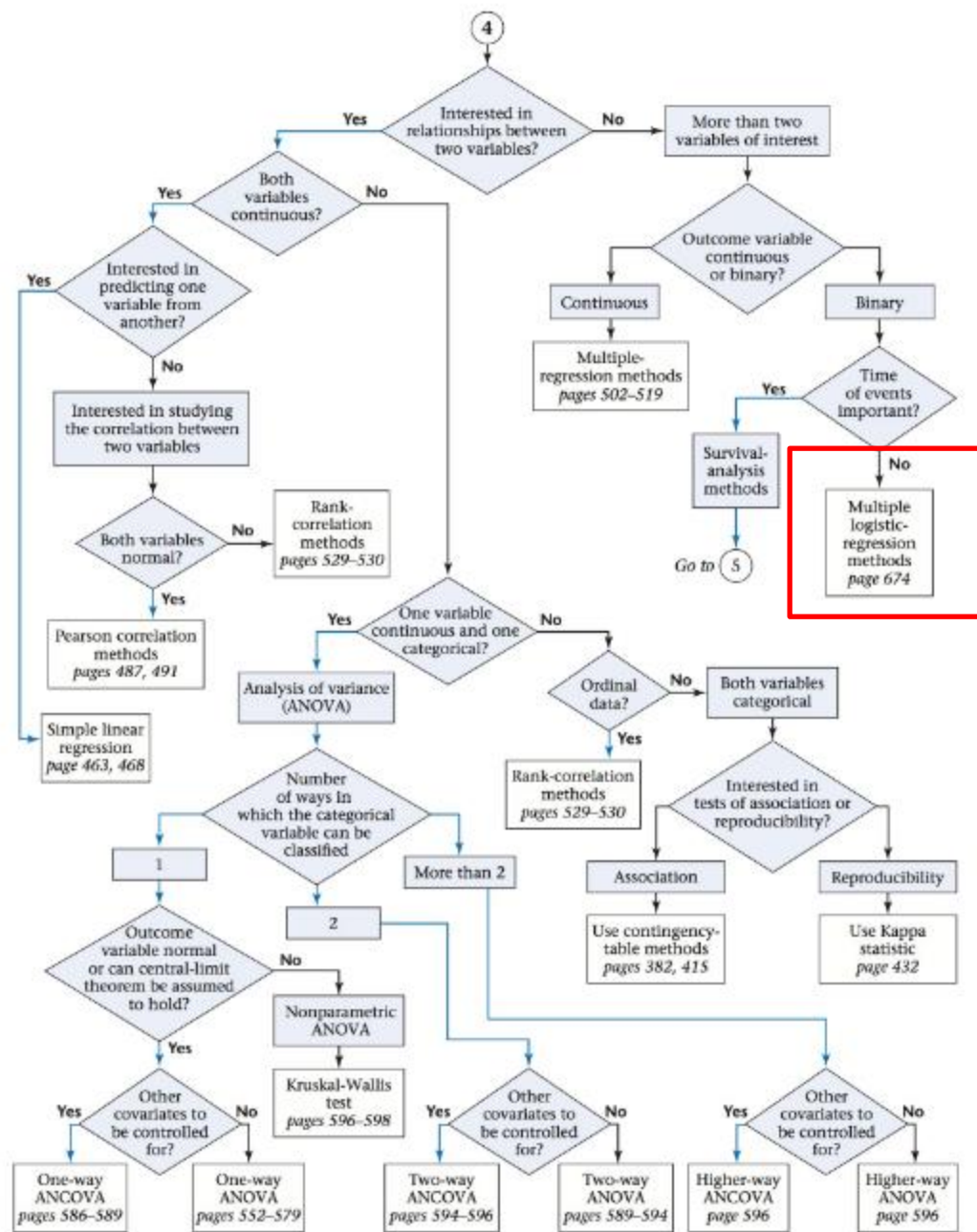
$$\hat{OR} = \frac{[a / (a + b)] \times [d / (c + d)]}{[c / (c + d)] \times [b / (a + b)]} = \frac{ad}{bc}$$

**disease-odds ratio**: odds in favor of disease for the exposed group divided by the odds in favor of disease for the unexposed group



FIGURE 7.18 Flowchart for appropriate methods of statistical inference





# Example on Relative Risk / Risk Ratio – Cancer

**Question: Estimate the RR for breast cancer for women with a late age at first birth ( $\geq 30$ ) compared with women with a nearly age at first birth ( $\leq 29$ ) based on the data below.**

**TABLE 10.1** Data for the international study in Example 10.4 comparing age at first birth in breast-cancer cases with comparable controls

Status	Age at first birth		Total
	$\geq 30$	$\leq 29$	
Case	683	2537	3220
Control	1498	8747	10,245
Total	2181	11,284	13,465

Source: Based on *WHO Bulletin*, 43, 209–221, 1970.

# Example on Relative Risk / Risk Ratio – Cancer

Solution :

The estimated OR is given by

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

$$\widehat{OR} = \frac{683(8747)}{2537(1498)} = 1.57$$

- An estimate of the RR



# Multiple Logistic Regression

$$p = \alpha + \beta_1 x_1 + \dots + \beta_k x_k$$

- $p$  is the probability of disease
- Right hand side:  $< 0$  or  $> 1$  for certain values of  $x_1, \dots, x_k$ 
  - predicted probabilities  $< 0$  or  $> 1 \rightarrow$  impossible
- logit (logistic) transformation of  $p$  is often used as the dependent variable
- The **logit transformation**  $\text{logit}(p)$ :  $\text{logit}(p) = \ln[p/(1-p)]$
- logit transformation:  $-\infty$  to  $+\infty$

## Multiple Logistic-Regression Model

- $x_1, \dots, x_k$ : independent variables
- $y$ : binomial-outcome variable with probability of success =  $p$

multiple logistic-regression model:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \dots + \beta_k x_k$$

Or

$$p = \frac{e^{\alpha + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\alpha + \beta_1 x_1 + \dots + \beta_k x_k}}$$

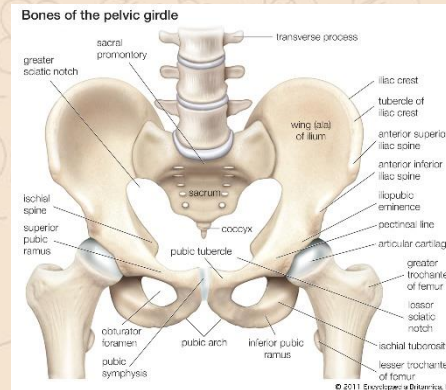
# Interpretation of Logistic-Regression Parameters

- Multiple logistic regression: analog to multiple linear regression
- Dichotomous exposure variable ( $x_j$ ): 1 (if present) and 0 (if absent)
  - OR associates this exposure variable to dependent variable:

$$\hat{OR} = e^{\hat{\beta}_j}$$

# Example on Infectious Disease

- *Chlamydia trachomatis* : microorganism that has been established as an important cause of nongonococcal urethritis, pelvic inflammatory disease



- A study of risk factors for *C. trachomatis* in 431 female college students
- Because multiple risk factors may be involved, several risk factors must be controlled for simultaneously in analyzing variables associated with *C. trachomatis*



$\text{logit}(p_A)$ ,  $\text{logit}(p_B)$ : logit of the probability of success for individuals A and B:

$$\text{logit}(p_A) = \alpha + \beta_1 x_1 + \dots + \beta_{j-1} x_{j-1} + \beta_j(1) + \beta_{j+1} x_{j+1} + \dots + \beta_k x_k$$

$$\text{logit}(p_B) = \alpha + \beta_1 x_1 + \dots + \beta_{j-1} x_{j-1} + \beta_j(0) + \beta_{j+1} x_{j+1} + \dots + \beta_k x_k$$

**TABLE 13.18** Multiple logistic-regression model relating prevalence of *C. trachomatis* to race and number of lifetime sexual partners

Risk factor	Regression coefficient $(\hat{\beta}_j)$	Standard error $se(\hat{\beta}_j)$	$z$ $[\hat{\beta}_j / se(\hat{\beta}_j)]$
Constant	-1.637		
Black race	+2.242	0.529	+4.24
Lifetime number of sexual partners among users of nonbarrier <sup>a</sup> methods of contraception <sup>b</sup>	+0.102	0.040	+2.55

<sup>a</sup>Barrier methods of contraception include diaphragm, diaphragm and foam, and condom; nonbarrier methods include all other forms of contraception or no contraception.

<sup>b</sup>This variable is defined as 0 for users of barrier methods of contraception.

Source: From McCormack, et al., "Infection with Chlamydia Trachomatis in Female College Students," *American Journal of Epidemiology*, 1985 121: 107-115.



# Estimation of ORs in Multiple Logistic regression for Dichotomous Independent Variables

- $x_j$ : dichotomous exposure variable (1 [present] / 0 [absent])

- OR in multiple logistic-regression model:  $\hat{OR} = e^{\hat{\beta}_j}$

*odds in favor of success if  $x_j = 1$*

*odds in favor of success if  $x_j = 0$*

\*after controlling for all other variables

Two-sided  $100\% \times (1-\alpha)$  CI for the true OR:

$$\left[ e^{\hat{\beta}_j - z_{1-\alpha/2} \text{se}(\hat{\beta}_j)}, e^{\hat{\beta}_j + z_{1-\alpha/2} \text{se}(\hat{\beta}_j)} \right]$$

- Only one risk factor in the model:  $E$  (1 [exposed] / 0 [unexposed]) and a dichotomous disease variable  $D$

$$\log[p/(1-p)] = \alpha + \beta E$$

# Example on Multiple Logistic Regression – Infectious Disease

**Question:** Estimate the odds in favor of infection with *C. trachomatis* for African American women compared with Caucasian women after controlling for previous sexual experience and provide a **95% CI** about this estimate.

**TABLE 13.18** Multiple logistic-regression model relating prevalence of *C. trachomatis* to race and number of lifetime sexual partners

Risk factor	Regression coefficient ( $\hat{\beta}_i$ )	Standard error $se(\hat{\beta}_i)$	$\left[ \hat{\beta}_i / se(\hat{\beta}_i) \right]^z$
Constant	-1.637		
Black race	+2.242	0.529	+4.24
Lifetime number of sexual partners among users of nonbarrier <sup>a</sup> methods of contraception <sup>b</sup>	+0.102	0.040	+2.55

<sup>a</sup>Barrier methods of contraception include diaphragm, diaphragm and foam, and condom; nonbarrier methods include all other forms of contraception or no contraception.

<sup>b</sup>This variable is defined as 0 for users of barrier methods of contraception.

Source: From McCormack, et al., "Infection with Chlamydia Trachomatis in Female College Students," *American Journal of Epidemiology*, 1985 121: 107–115.

# Example on Multiple Logistic Regression – Infectious Disease

**Solution :**

$$\widehat{OR} = e^{2.242} = 9.4$$

**Conclusion:** odds in favor of infection for African American women are nine times as great as those for Caucasian women after controlling for previous sexual experience.

- because  $Z_{1-\alpha/2} = Z_{.975} = 1.96$  and  $se(\hat{\beta}_j) = 0.529$   
95% CI for OR is given by:

$$\left[ e^{2.242 - 1.96(0.529)}, e^{2.242 + 1.96(0.529)} \right] = (e^{1.205}, e^{3.279}) = (3.3, 26.5)$$



TABLE 3 The normal distribution

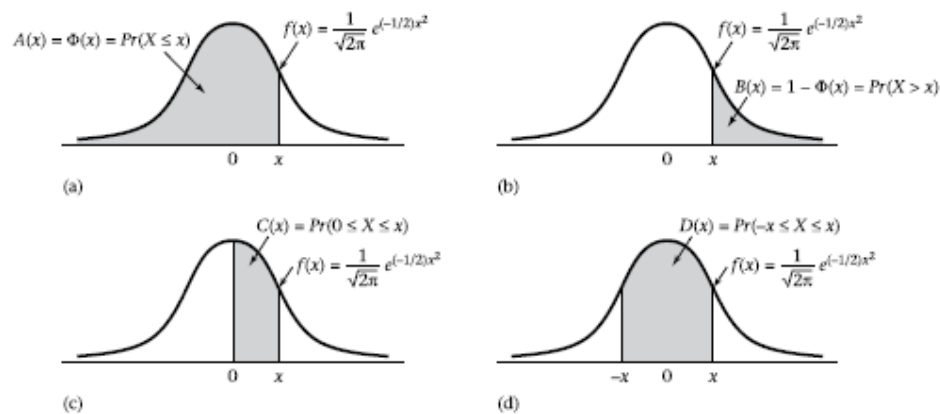


TABLE 3 The normal distribution (continued)

$x$	$A^a$	$B^a$	$C^c$	$D^d$
1.82	.9656	.0344	.4656	.9312
1.83	.9664	.0336	.4664	.9327
1.84	.9671	.0329	.4671	.9342
1.85	.9678	.0322	.4678	.9357
1.86	.9686	.0314	.4686	.9371
1.87	.9693	.0307	.4693	.9385
1.88	.9699	.0301	.4699	.9399
1.89	.9706	.0294	.4706	.9412
1.90	.9713	.0287	.4713	.9426
1.91	.9719	.0281	.4719	.9439
1.92	.9726	.0274	.4726	.9451
1.93	.9732	.0268	.4732	.9464
1.94	.9738	.0262	.4738	.9476
1.95	.9744	.0256	.4744	.9488
1.96	.9750	.0250	.4750	.9500
1.97	.9756	.0244	.4756	.9512
1.98	.9761	.0239	.4761	.9523
1.99	.9767	.0233	.4767	.9534
2.00	.9772	.0228	.4772	.9545
2.01	.9778	.0222	.4778	.9556
2.02	.9783	.0217	.4783	.9566
2.03	.9788	.0212	.4788	.9576
2.04	.9793	.0207	.4793	.9586
2.05	.9798	.0202	.4798	.9596
2.06	.9803	.0197	.4803	.9606
2.07	.9808	.0192	.4808	.9615
2.08	.9812	.0188	.4812	.9625
2.09	.9817	.0183	.4817	.9634
2.10	.9821	.0179	.4821	.9643
2.11	.9826	.0174	.4826	.9651
2.12	.9830	.0170	.4830	.9660
2.13	.9834	.0166	.4834	.9668
2.14	.9838	.0162	.4838	.9676
2.15	.9842	.0158	.4842	.9684
2.16	.9846	.0154	.4846	.9692
2.17	.9850	.0150	.4850	.9700
2.18	.9854	.0146	.4854	.9707
2.19	.9857	.0143	.4857	.9715
2.20	.9861	.0139	.4861	.9722
2.21	.9864	.0136	.4864	.9729
2.22	.9868	.0132	.4868	.9736
2.23	.9871	.0129	.4871	.9743
2.24	.9875	.0125	.4875	.9749
2.25	.9878	.0122	.4878	.9756
2.26	.9881	.0119	.4881	.9762
2.27	.9884	.0116	.4884	.9768
2.28	.9887	.0113	.4887	.9774
2.29	.9890	.0110	.4890	.9780
2.30	.9893	.0107	.4893	.9786
2.31	.9896	.0104	.4896	.9791
2.32	.9898	.0102	.4898	.9797
2.33	.9901	.0099	.4901	.9802
2.34	.9904	.0096	.4904	.9807
2.35	.9906	.0094	.4906	.9812
2.36	.9909	.0091	.4909	.9817
2.37	.9911	.0089	.4911	.9822
2.38	.9913	.0087	.4913	.9827

# Summary

- Main epidemiological study designs and relevant effect estimates:

Study Design	Prospective cohort study	Case and Control
Effect Estimate	Relative Risk/ Risk Ratio	Odds ratio

- **Multiple logistic regression:**  $p = \alpha + \beta_1 x_1 + \dots + \beta_k x_k$

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \dots + \beta_k x_k$$

- OR: links exposure variable to the dependent variable:  
95% CI for OR:

$$\hat{OR} = e^{\hat{\beta}_j}$$

$$\left[ e^{\hat{\beta}_j - z_{1-\alpha/2} \text{se}(\hat{\beta}_j)}, e^{\hat{\beta}_j + z_{1-\alpha/2} \text{se}(\hat{\beta}_j)} \right]$$