Epidemiologic Measure of Association

Basic Concepts

Measures of Disease Occurrence:

$$\begin{aligned} \text{Prevalence} &= \text{The amount of disease already present in the population} \\ P &= \frac{\text{Number of cases}}{\text{Number of People}} \end{aligned}$$

$$\begin{aligned} \text{Risk} &= \text{The likelihood that an individual will contract the disease} \\ R &= \frac{\text{\# New cases of disease during given period}}{\text{Population at risk}} \end{aligned}$$

$$\begin{aligned} \text{Incidence Rate} &= \text{How fast new occurrences of disease arise} \\ IR &= \frac{\text{\# New cases of disease during given period}}{\text{Total person-time of observation}} \end{aligned}$$

Confidence Interval for population characteristic:

$$\text{Best guess of population characteristic} \pm 1.96 \times \text{Standard Error}$$

95% CI for Prevalence P

$$P \pm 1.96 \times \sqrt{\frac{P \times Q}{\text{Sample Size}}}, \quad \text{where} \quad Q = 1 - P$$

95% CI for Population Mean

$$\overline{X} \pm 1.96 \times \frac{\text{Standard Deviation}}{\sqrt{\text{Sample Size}}}$$

		Disease		
Exposure		**Present**	**Absent**	**Total**
	Yes	A	B	N_1 = A+B
	No	C	D	N_2=C+D
	Total	M_1=A+C	M_2=B+D	$N_1+N_2=M_1+M_2$

Epidemiologic Measures of Disease Association

$$\text{Risk Difference} = RD = R_1 - R_0 = \frac{A}{A+B} - \frac{C}{C+D}$$

$$\text{Risk Ratio} = RR = \frac{R_1}{R_0} = \frac{A \times (C+D)}{C \times A + B}$$

$$\text{Odds of disease} = \frac{\text{Probability of disease}}{1 - \text{Probability of disease}}$$

$$\text{Odds Ratio} = OR = \frac{\text{Odds of disease among the exposed}}{\text{Odds of disease among the non-exposed}}$$

$$\text{Odds Ratio} = OR = \frac{A \times D}{B \times C}$$

95% Confidence Interval for the Risk Difference (RD):

(1) Estimate RD from data

(2) Estimate Standard Error of RD as follows:

$$\text{se(RD)} = \sqrt{\frac{R_1 \times (1-R_1)}{N_1} + \frac{R_0 \times (1-R_0)}{N_2}}$$

(3) 95% CI for RD is given by:

$$RD \pm 1.96 \times \text{se(RD)}$$

95% Confidence Interval for Risk Ratio (RR):

(1) Compute $\log(RR)$ from data

(2) Estaimate Standard Error of $\log(RR)$ as follows:

$$\text{se}(\log(\text{RR})) = \sqrt{\frac{B}{AN_1} + \frac{D}{CN_2}}$$

(3) Construct a 95% CI for $\log(RR)$

$$\log(RR) \pm 1.96 \times \text{se}(\log(\text{RR}))$$

(4) Take the anti-log (natural log)

$$exp(\log(RR) \pm 1.96 \times \text{se}(\log(\text{RR})))$$

95% Confidence Interval for the Odds Ratio (OR):

(1) Compute $\log(OR)$ from data

(2) Estaimate Standard Error of $\log(OR)$ as follows:

$$\text{se}(\log(\text{OR})) = \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}}$$

(3) Construct a 95% CI for $\log(OR)$

$$\log(OR) \pm 1.96 \times \text{se}(\log(\text{OR}))$$

(4) Take the anti-log (natural log)

$$exp(\log(OR) \pm 1.96 \times \text{se}(\log(\text{OR})))$$

The Framingham Heart Study

Data for this example was obtained from the textbook "Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data," by William D. Dupont. The description of the study and the variable names are taken from the textbook verbatim.

The Framingham Heart Study has collected long-term follow-up and cardiovascular risk factor data on almost 5,000 residents of the town of Framingham, MA. Recruitment of patients started in 1948. At the time of the baseline exams there were no effective treatments for hypertension. Permission to use a subset of the 40-year data from this study was obtained by the author. The data consist of 4,699 individuals who were free of coronary heart disease at their baseline exam. (W. Dupont, Statistical Modeling for biomedical Researchers: A Simple introduction to the analysis of complex data. Cambridge University Press, 2009. Second Edition).

Baseline variables

sbp = Systolic blood pressure in mmHg
dbp = Diastolic blood pressure in mmHg
age = Age in years
scl = Serum cholesterol in mg/100ml,
bmi = Body mass index (kg/m2)
sex = Gender coded as 1 if male 2 if female

Follow-up variables

followup = The subject's follow-up in days
chdfate = 1 if patient develops CHD at the end of follow-up; 0 otherwise

Calculated Variables

obesity = 1 if subject's BMI greater than 30; 0 otherwise
htn = Hypertension; 1 if SBP greater than 140 or DBP greater than 90

Table I. Summary measures of main variables

Variable	Sample Size	Mean	Standard Deviation
sbp	4699	132.8	22.8
dbp	4699	82.5	12.7
scl	4666	228.3	44.5
age	4699	46.0	8.5
bmi	4690	25.6	4.1

Table II. Hypertension by obesity

		Hypertension		
		Yes	**No**	**Total**
Obesity	**Yes**	346	235	581
	No	1205	2904	4109
	Total	1551	3139	4690

Table III. CHD morbidity by obesity				
		CHD		
		Yes	**No**	**Total**
Obesity	**Yes**	241	340	581
	No	1231	2878	4109
	Total	1472	3218	4690

Table IV. CHD morbidity by obesity			
		CHD	
		Number who Developed CHD	**Total person-years contributed**
Obesity	**Yes**	241	11378.0
	No	1231	92197.7
	Total	1472	103575.6

Measures of Disease Occurrence

1. **Estimate the prevalence of hypertension in the Framingham population**

$$\mathrm{P} = \mathrm{Prevalence} = \frac{1551}{4690} \times 100 = 33.1\%$$

2. **Construct a 95% confidence interval for the true prevalence of hypertension**

$$(1) \quad P = 0.33$$

$$(2) \quad Q = 1 - P = 0.67$$

$$(3) \quad \begin{aligned} \mathrm{Lower} &= 0.33 - 1.96 \times \sqrt{\tfrac{P \times Q}{4690}} = 0.33 - 1.96 \times 0.0069 = 0.317 \\ \mathrm{Upper} &= 0.33 + 1.96 \times \sqrt{\tfrac{P \times Q}{4690}} = 0.33 - 1.96 \times 0.0069 = 0.34.3 \end{aligned}$$

3. **Estimate the prevalence of obesity in the Framingham population**

$$\mathrm{Prevalence\ of\ Obesity} = \frac{581}{4690} \times 100 = 12.4\%$$

4. **Construct 95% confidence intervals for the true mean systolic and diastolic blood pressure**

$$\begin{aligned} \text{SBP} \quad &: \quad \begin{array}{l} \text{Lower:} 132.8 - 1.96 \times \frac{22.8}{\sqrt{4699}} = 132.1 \\ \text{Upper:} 132.8 + 1.96 \times \frac{22.8}{\sqrt{4699}} = 133.5 \end{array} \\ \text{DBP} \quad &: \quad \begin{array}{l} \text{Lower:} 82.5 - 1.96 \times \frac{12.7}{\sqrt{4699}} = 82.1 \\ \text{Upper:} 82.5 + 1.96 \times \frac{12.7}{\sqrt{4699}} = 82.9 \end{array} \end{aligned}$$

5. **Estimate the incidence rate of developing CHD in the Framingham population**

$$\text{IR} = \text{Incidence Rate} = \frac{1472}{103575.6} = 142 \text{ per } 10{,}000 \text{ person-years}$$

6. **Estimate the incidence rate of developing CHD among the obese in the Framingham population**

$$\text{IR}_1 = \text{Incidence Rate} = \frac{241}{11378.0} = 212 \text{ per } 10{,}000 \text{ person-years}$$

7. **Estimate the incidence rate of developing CHD among the non-obese in the Framingham population**

$$\text{IR}_0 = \text{Incidence Rate} = \frac{1231}{92197.7} = 134 \text{ per } 10{,}000 \text{ person-years}$$

8. **What is the risk of developing CHD among obese subjects?**

$$\text{R}_1 = \text{Risk} = \frac{241}{581} \times 100 = 41.5\%$$

9. **What is the risk of developing CHD among non-obese subjects?**

$$\text{R}_0 = \text{Risk} = \frac{1231}{4109} \times 100 = 30.0\%$$

Epidemiologic Measures of Disease Association

10. What is the risk difference of developing CHD among the obese and the non-obese?

$$\text{RD} = \text{Risk Difference} = 41.5\% - 30.0\% = 11.5\%$$

11. What is the risk ratio of developing CHD of the obese compared to the non-obese?

$$\text{RR} = \text{Risk Ratio} = \frac{41.5\%}{30.0\%} = 1.38$$

12. Is obesity associated with hypertension? Hint: Use the odds ratio to estimate the degree of association

$$\text{OR} = \text{Odds Ratio} = \frac{A \times D}{C \times B} = \frac{241 \times 2878}{1231 \times 340} = 1.66$$

Yes, there is an association between obesity and hypertension, since the OR =1.66 is greater than one. Also, the risk ratio is greater than one. Furthermore, the 95% confidence interval for the true OR does not contain 1.0, i.e., 95% CI = 1.38-1.99. The 95% confidence interval for the true RR is 1.24-1.54, which excludes 1.0 as well.

13. Of the 2049 males in the study, 823 developed CHD, whereas 650 of the 2650 women developed CHD. Is there an association between gender and CHD morbidity in the Framingham population?

From the data provided, the following 2 by 2 table is constructed:

CHD morbidity by sex				
		CHD		
		Yes	**No**	**Total**
Sex	**Male**	823	1226	2049
	Female	650	2000	2650
	Total	1473	3226	4699

$$\begin{aligned} \text{Risk}_{males} &= 0.402 \\ \text{Risk}_{females} &= 0.245 \\ \text{RR} &= 1.67 \\ \text{Lower 95\% CI} &= 1.50 \\ \text{Upper 95\% CI} &= 1.78 \end{aligned}$$

The estimated risk ratio, 1.67, exceeds 1.0, and the corresponding 95% confidence interval for the true risk ratio does not contain 1.0. Thus, there is an association between gender and the development of CHD, with males having about 70% higher risk than females.

The Honolulu Heart Study

The Honolulu Heart Study is a prospective examination of coronary heart disease and stroke among a cohort of men of Japanese ancestry born between 1900 and 1919 and residing on the island of Oahu in 1965. The target population consisted of 11,148 men aged 45 to 68, and 8,006 from this group participated in the baseline examination between 1965 and 1968. Researchers collected data on height, weight, demographic material, medical history, and socio-cultural factors. Repeat examinations were performed near the 2nd and 6th anniversaries of baselines, with response rates of 95% and 90%, respectively. Mortality data were determined through ongoing review of death certificates, as part of the routine surveillance of the cohort through 25 years of followup. Remarkably, the survival status is unknown for only four men. For each death, a panel of study physicians determined the underlying cause and classified it according to ICD-8.

Although the investigators are interested in studying the association between coronary heart disease (CHD) and risk factors, the data could be used to study the association between smoking and cancer mortality.

Table I. Cancer mortality by smoking(ever)

		Cancer Death		
		Yes	**No**	**Total**
Smoking	**Yes**	619	4308	4927
	No	161	2000	2161
	Total	780	6308	7088

Table II. Cancer mortality by smoking (ever)

		Cancer Death	
		Number who died of cancer	**Total person-years contributed**
Smoking	**Yes**	619	94390.2
	No	161	43156.5
	Total	780	137546.7

Epidemiologic Measures of Disease Association

1. **What is the risk difference of death from cancer among the smokers and the non-smokers?**

$$\text{RD} = \text{Risk Difference} = \frac{619}{4927} - \frac{161}{2161} = 0.1256 - 0.0745 = 0.0511$$

2. **What is the risk ratio of cancer death of smokers compared to non-smokers? Interpret the estimated risk ratio.**

$$\text{RR} = \frac{\frac{619}{4927}}{\frac{161}{2161}} = \frac{0.1256}{0.0745} = 1.686$$

3. **The 95% confidence interval for the true risk ratio of cancer death is: 1.43-1.99. What do you conclude about the association between cancer death and smoking? Explain**

   First, the estimated risk ratio is 1.686 >1. Second, the 95% confidence interval does not contain 1.0. Therefore, it appears there is a positive association between smoking and cancer death.

4. **What are the odds of cancer death among smoker? How about among non-smokers?**

$$\begin{aligned} \text{Odds of cancer for smokers} &= \frac{0.1256}{1-0.1256} = 0.1436 \\ \text{Odds of cancer for non-smokers} &= \frac{0.0745}{1-0.0745} = 0.0805 \end{aligned}$$

5. **What is the odds ratio of cancer death of smokers compared to non-smokers?**

$$\text{OR} = \frac{0.1436}{0.0805} = 1.785$$

Or

$$\text{OR} = \frac{619 \times 200}{161 \times 4308} = 1.785$$

6. **Construct and interpret a 95% confidence interval for the odds ratio.**

$$
\begin{aligned}
&(1) \quad \text{OR} = 1.785 \\
&(2) \quad \log(\text{OR}) = \log(1.785) = 0.5794 \\
&(3) \quad \text{se}(\log(\text{OR})) = \sqrt{\frac{1}{619} + \frac{1}{4308} + \frac{1}{161} + \frac{1}{2000}} = 0.0925 \\
&(4) \quad \begin{array}{rcl} \text{Lower 95\% CI for } \log(\text{OR}) & = & 0.5794 - 1.96 \times 0.0925 = 0.3981 \\ \text{Upper 95\% CI for } \log(\text{OR}) & = & 0.5794 + 1.96 \times 0.0925 = 0.7607 \end{array} \\
&(5) \quad \begin{array}{rcl} \text{Lower 95\% CI for OR} & = & \exp(0.3981) = 1.49 \\ \text{Upper 95\% CI for OR} & = & \exp(0.7607) = 2.14 \end{array}
\end{aligned}
$$

7. **What is the mortality rate (incidence rate) of cancer death among smokers?**

$$\text{IR} = \text{Incidence Rate for smokers} = \frac{619}{94390.2} = 656 \text{ per } 100{,}000 \text{person-years}$$

8. **What is the mortality rate (incidence rate) of cancer death among non-smokers?**

$$\text{IR} = \text{Incidence Rate for non-smokers} = \frac{161}{43156.5} = 373 \text{ per } 100{,}000 \text{person-years}$$

9. **Estimate the mortality rate ratio (incidence rate ratio) and compare to the risk ratio and odds ratio.**

$$\text{IRR} = \text{Incidence Rate Ratio} = \frac{656}{373} = 1.76$$

The three estimates of disease association are very similar since the risk of disease is small (less than 10%). Precisely, the risk of cancer death in the population is 5.1%.

Food-borne Example

The table below summarizes the data corresponding to the food-borne example

		Disease (Gastroenteritis)		
		Present	**Absent**	**Total**
Exposure (Eating meat dish at picnic)	**Yes**	**63**	**25**	**88**
	No	**1**	**6**	**7**
	Total	**64**	**31**	**95**

introduced in class.

Use the information above to:

1. **Estimate the risk of gastroenteritis among the exposed and non-exposed**

$$
\begin{aligned}
R_1 &= \frac{63}{88} = 0.716 \\
R_0 &= \frac{1}{7} = 0.143
\end{aligned}
$$

2. Estimate the risk difference and corresponding standard error

$$
\begin{aligned}
RD &= \frac{63}{88} - \frac{1}{7} = 0.573 \\
se(RD) &= \sqrt{\frac{R_1 \times (1 - R_1)}{n_1} + \frac{R_0 \times (1 - R_0)}{n_2}} = \sqrt{\frac{0.716 \times 0.284}{88} + \frac{0.143 \times 0.857}{7}} = 0.141
\end{aligned}
$$

3. Compute a 95% confidence interval for the true risk difference

$$
\begin{aligned}
\text{Lower} &= 0.573 - 1.96 \times 0.141 = 0.297 \\
\text{Upper} &= 0.573 + 1.96 \times 0.141 = 0.849
\end{aligned}
$$

4. Construct a 95% confidence interval for the risk ratio of gastroenteritis of the exposed compared to the non-exposed

(1) $RR = 5.01$

(2) $\log(RR) = 1.612$

(3) $se(\log(RR)) = \sqrt{\frac{B}{A \times n_1} + \frac{D}{C \times n_2}} = \sqrt{\frac{25}{63 \times 88} + \frac{6}{1 \times 7}} = 0.928$

(4) $\text{Lower:} 1.612 - 1.96 \times 0.928 = -0.208; \quad \text{Upper:} 1.612 + 1.96 \times 0.928 = 3.431$

(5)
$$
\begin{aligned}
\text{Lower CI for RR} &= \exp(-0.208) = 0.812 \\
\text{Upper CI for RR} &= \exp(3.431) = 30.910
\end{aligned}
$$

5. Construct a 95% confidence interval for the odds ratio of gastroenteritis of the exposed compared to the non-exposed.

(1) $OR = \frac{63 \times 6}{1 \times 25} = 15.12$

(2) $\log(OR) = 2.716$

(3) $se(\log(OR)) = \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}} = 1.106$

(4) $\text{Lower:} 2.716 - 1.96 \times 1.106 = 0.549; \quad \text{Upper:} 2.716 + 1.96 \times 1.106 = 4.883$

(5)
$$
\begin{aligned}
\text{Lower CI for OR} &= \exp(0.549) = 1.73 \\
\text{Upper CI for OR} &= \exp(4.883) = 132.04
\end{aligned}
$$