Measures of Association

Causal Inference

The **Causal Effect** of an exposure for an individual is the difference in the outcomes (Y) for that individual if given the exposure (Y^1) versus not given the exposure Y^0). These two outcomes are referred to as the **counterfactual outcomes** for that individual. For example, the causal effect of lifetime smoking starting at age 20 on the development of CHD for a person is the difference in the CHD counterfactual outcome if that person started smoking at age 20 (smoker), compared to the CHD counterfactual outcome if that person did not start smoking at age 20 (non-smoker). Since only one of these outcomes can be observed in reality (depending on whether the individual actually smoked or did not smoke), the causal effect on the individual level cannot be measured.

The **average causal effect of a risk factor for a population** is the difference in average counterfactual outcomes, $E(Y^1)$, when all members of the population receive the exposure and the average counterfactual outcomes ,$E(Y^0)$, when none of the members of the population receive it. **Average Causal Measures of Effect** can be defined as

Casual Risk Difference = $E(Y^1) - (E(Y^0)$.

Causal Risk Ratio = $E(Y^1) / E(Y^0)$.

Causal Odds Ratio = $[E(Y^1) /(1-E(Y^1))] /[E(Y^0) /(1-E(Y^0))]$

(N.B. E(Y) refers to the statistical term of "expected value" of Y and represents the average (mean) of y)

If the population is comprised of some members who receive the exposure and others who do not, then it may be possible to estimate each average counterfactual outcomes and the average causal effect of the exposure in the absence of **confounding** and **bias**. For example, in the absence of confounding and bias, the incidence of coronary heart disease from a group of non-smokers (**a factual outcome**) is a valid estimate for the counterfactual outcome for a group of smokers had they not smoked. This assumes an ability to "exchange" factual outcomes of the non-smokers with the counterfactual outcomes for the smokers and vice versa (Greenland S, Robins JM. Identifiability, Exchangeability, and Epidemiological Confounding. *Intl J Epidemiol* 1986;15:412-419.). This implies that the observed difference (ratio) in incidence of coronary heart disease in the two comparison groups (**a measure of association**) is an estimate of the average causal effect of smoking.

Hernan and Robin (Estimating causal effects from epidemiological data. J *Epidemiol Community Health* 2006;60:578-586.) demonstrate how these causal effect measures can be estimated by measures of association from epidemiologic studies that are free of bias and confounding (e.g. randomized experiments) or from observation

studies by conditioning the values of the confounding. They proposed using the method of inverse probability weighting (IPW) to estimate causal effect to control for known confounders. This method is similar to standardization and will be discussed in future lectures.

Measures of Association

In epidemiology a **measures of association** compares the outcome measurement (Prevalence, Incidence Rate, …) in groups of subjects that are defined by categories of a risk factor of interest (exposure**).** Mathematically, measures of association are either ratios or differences of an outcome measure in these groups. In the absence of bias and confounding, these measures of association provide estimates for the causal effect of the risk factor on the outcome (measure of effect). The following table presents some commonly used measures of association in epidemiology. As mentioned previously, the outcome measurement of choice is typically determined by the scale of the outcome.

Table Some Common Measures of Association.

Outcome	Outcome Measure	Measure of Association
Nominal	Proportions (Cumulative Incidence, Estimated Risk) Incidence Rates Odds	Risk Ratio/Risk Difference Rate Ratio/Rate Difference Odds Ratio
Ordinal	Above Measures Medians (Percentiles)	Above Measures Ratio/Difference of Medians (Percentiles)
Continuous	Above Measures Averages	Above Measures Ratio/Difference of Averages

Binary Outcome

A binary outcome may reflect a natural dichotomy (e.g. dead versus alive) or can be created from nominal, ordinal, or continuous outcome by collapsing categories (e.g. New York Heart Association Classes III and IV versus Classes I and II (Goldman et al.,1981)) or specifying a threshold value for a continuous variable to separate high from low outcomes (e.g. hypertension (SBP $\geq$ 140) versus no hypertension). If the exposure of interest is also binary (e.g. current smoker vs. non-smoker), then data relating the exposure to the outcome can be displayed in a 2x2 table as follows"

2x2 Table Displaying the Relationship between a Binary Exposure (E) and Disease (D).

	D		
	+	-	Total
E +	a	b	N_1
E -	c	d	N_0
Total	M_1	M_0	T

Data displayed in 2x2 tables arise from a variety of study designs: cohort studies, case control studies, cross-sectional studies, and experimental studies. Often the choice of the appropriate measure of effect may depend on the type of study design. For example, the odds ratio or some function of the odds ratio is typically the only measure of effect that is calculated from most case control studies.

If there are no losses-to-follow-up or losses due to competing risks, then the Cumulative Incidence of disease for the exposure groups provides estimates for the risk of disease for each group.

Estimated Risk of Disease

Exposed Subjects (E+) $R_1 = a/N_1$

Non-exposed Subjects (E-) $R_0 = c/N_0$

Common measures of association based on ratio or differences of estimated risks are shown in following table.

Measure of Effect	**Formula**
Risk Ratio (Relative Risk)	$RR = R_1 / R_0$
Risk Difference (Attributive Risk)	$RD = R_1 - R_0$
Disease Odds Ratio	$OR = [R_1/(1-R_1)]/[R_0/(1-R_0)]$

Similar measures can be calculated with data from cross-sectional studies using prevalence rather than incidence as an outcome measures.

Incidence Rate data for a binary exposure can be displayed in a slightly different table

	D+	Person-time	Incidence Rate
E +	a	K_1	$R_1 = a/K_1$
E -	c	K_0	$R_0 = c/K_0$
Total	M_1	T	

Measures of Association can be calculated by taking the ratio or difference of the Incidence Rates

$$\text{Rate Ratio} = R_1 / R_0$$

$$\text{Rate Difference} = R_1 - R_0$$

The value for a Rate Ratio is a pure number. However, Rate Difference, like Incidence Rates, is measured in case/person-time.

Example: The following tables display the relationship between current smoking and the incidence death during 24 years of follow-up from the FHS teaching data set.

	Death			
	+	-	Total	Estimated Risk
Smoker	788	1393	2181	0.3613
Non-Smoker	762	1491	2253	0.3382

Risk Ratio = RR = 0.3613 / 0.3382 = 1.0683

Risk Difference = RD = 0.3613 – 0.3382 = 0.0231

	Deaths	Person-years	Incidence Rate (deaths/10,000py)
Smoker	788	44,440.38	177.3162
Non-Smoker	762	46,675.20	163.2559

Rate Ratio = RR = 177.3162 / 163.2559 = 1.0861

Rate Difference = RD = (177.3162 – 163.2559)cases/(10,000py)
= 14.0603 (cases/10,000py)

It should be noted that the label RR may refer to either a Risk Ratio calculation or a Rate Ratio calculation. Often the term **Relative Risk** (also labeled RR) is used to describe either calculation. However, as shown in the previous example, the value for the Risk Ratio and Rate Ratio will tend to differ, and the use of a single tem (Relative Risk) to describe two different results may be confusing.

Attributable Proportions

If R_1 is the risk of developing an outcome among an exposed subject, then a question of interest might be how much of the magnitude of R_1 is actually caused by the exposure. If R_0 is the risk that an exposed subject would have had in the absence of the exposure then $(R_1 - R_0)$ is the extra risk that is caused by the exposure and the proportion of R_1 that is attributed to the exposure is

$$(R_1 - R_0)/R_1 = (R_1/R_0 - R_0/R_0) / R_1/R_0 = (RR - 1)/RR$$

This quantity can be estimated using the estimated risks (cumulative incidence) from population data and is referred to as the **Attributable Proportion among the Exposed** (Ashcengrau and Seage), **Attributable Fraction** (in Stata and in Rothman KJ. Epidemiology: An Introduction 2nd Edition. Oxford University Press 2012), and **Attributable Risk Percent (**Hennekens and Buring**).**

A similar question for consideration is what proportion of the **average risk** in a population is attributed to some members of that population having the exposure of interest. This is a more of a public health consideration as the answer is linked to a specified population with a specific prevalence of exposure (p). The average risk is a population is

$$R_T = pR_1 + (1-p)R_0$$

The portion of the average risk that is attributable to the exposure is

$$\begin{aligned}
(R_T - R_0)/R_T &= [pR_1 + (1-p)R_0 - R_0]/ [pR_1 + (1-p)R_0] \\
&= [pR_1 - pR_0]/ [pR_1 - pR_0 + R_0] \\
&= [p(R_1 - R_0)] / [p(R_1 - R_0) + R_0] \\
&= [p(R_1/R_0 - R_0/R_0] / [p(R_1/R_0 - R_0/R_0) + R_0/R_0] \\
&= [p(RR - 1)] / [p(RR - 1) + 1]
\end{aligned}$$

This quantity also can be estimated using the estimated risks (cumulative incidence) from population data and is referred to as the **Attributable Proportion in the Total Population** (Ashcengrau and Seage), **Attributable Fraction for the Population** (Stata, Rothman), **Population Attributable Risk Percent (**Hennekens and Buring**).**

Example: The following table displays the relationship between current smoking and the incidence death during 24 years of follow-up from the FHS teaching data set.

	Death			
	+	-	Total	Estimated Risk
Smoker	788	1393	2181	0.3613
Non-Smoker	762	1491	2253	0.3382
Total	1550	2884	4434	0.3496

Risk Ratio = RR = 0.3613 / 0.3382 = 1.0683

p = Prevalence of Smoking = 2181/4434 = 0.4919

Attributable Fraction – Exposed = (1.0683 – 1)/1.0683 = 0.0639

Attributable Fraction – Population = [0.4919(1.0683-1)]/ [0.4919(1.0683-1) +1]
= 0.0325

Number Needed to Harm and Number Needed to Treat

The **Number Needed to Harm** is the number of subjects, if given a harmful exposure ($R_1 > R_0$), would cause the one case of disease. For example, the following table displays the relationship between current smoking and death during 24 years of follow-up from the FHS teaching data set.

	Death			
	+	-	Total	Estimated Risk
Smoker	788	1393	2181	0.3613
Non-Smoker	762	1491	2253	0.3382

Risk Difference = RD = 0.3613 – 0.3382 = 0.0231
= 3613/10,000 –3382/10,000 = 231/10,000

This result implies that if 10,000 subjects smoked rather than not smoking then 231 extra deaths would occur. Therefore, if 43.29 subjects smoked (rather than not smoking) then 1 extra death would occur. This number of based on the following formula:

231/10000 = (231/231 / 10,000/231) = 1/43.29 = 1/RD

The general formula for the **Number Needed to Harm (NNH)** is

$$NNH = 1/RD$$

Similarly is an exposure (treatment) lowers the risk of developing an outcome ($R_1 < R_0$) then the **Number Needed to Treat (NNT)** to prevent one case of disease is

$$NNT = 1/(R_0 - R_1) = 1/|RD|$$

Regression Coefficients

Estimates for most of the effect measures that were presented in these notes can also be obtained from **regression models** that describe an outcome measure as a function of the exposure of interest.

The general formula for a straight line to describe linear relationship between two variables X and Y is

$$Y = mX + b$$

where

b = Y-intercept = value for Y when X=0

m = slope = change in Y when X changes by one unit

In epidemiology research, Y (dependent variable) represents a function of outcome measure and X (independent variable) represents an exposure (E) and the model is usually written as

$$f(\text{outcome measure}) = B_0 + B_1E$$

where

B_1 = slope = change in f(outcome measure)/ unit change in E

When the outcome is binary scale and the outcome measure is a proportion (P: prevalence, cumulative incidence) then the **logistic regression model** to describe the relationship between and exposure (E) and the outcome. The logistic regression model uses a function of P (logit: natural logarithm of the P/(1-P)) for Y, yielding the following formula

$$\log(P/(1-P)) = B_0 + B_1E$$

where

$$B_1 = \Delta\log(P/(1-P)) \text{ / unit } \Delta \text{ in E}$$

If the exposure is binary, labeled 1 for exposed subjects and 0 for non-exposed subjects, then

$$\begin{aligned} B_1 &= [\log(P_1/(1-P_1)) - \log(P_0/(1-P_0))] / (1-0) \\ &= \log[P_1/(1-P_1)) / (P_0/(1-P_0))] \\ &= \log(OR) \end{aligned}$$

where P_1 = outcome measure for exposed subjects and P_0 = outcome measure for non-exposed subjects.

Example: The following table, taken from the FHS teaching data set, shows the relationship between smoking at the 1956 exam and the incidence of death during 24 years of follow-up

	Death			
	+	-	Total	Odds
Smoker	788	1393	2181	788/1393 = 0.5657
Non-Smoker	762	1491	2253	762/1491 = 0.5111

Odds Ratio = OR = 0.5657 / 0.5111 = 1.1068

If a logistic regression model were fit to these data the resulting fitted model is

$$\log(P/(1-P)) = -0.6713 + 0.1015(\text{Smoker})$$

$$B_1 = 0.1015 = \log(OR)$$

$$OR = e^{0.1015} = 1.1068$$