

BASIC CONCEPTS IN EPIDEMIOLOGY AND BIOSTATISTICS

Epidemiology is the study of the distribution and determinants of health-related states or events in specified populations. **Study** is interpreted to mean observation, recording or testing; **distribution** can be spatial, temporal or demographic; the **determinants** can be physical, biological, behavioural, environmental etc; the **health-related states** are incidence of disease, death, treatment reaction, health-care provision etc.; epidemiological studies may be

- **ANALYTIC:** designed to examine and test associations and/or causal relationships
- **DESCRIPTIVE:** containing non-quantitative assessment, data summary and display
- **EXPERIMENTAL:** study conditions at the control of the experimenter

Epidemiological studies are concerned with

- disease definition
- aetiology (ie mechanistic cause)
- population incidence
- identification of risk factors
- disease control/prevention elimination

A **risk factor** is an aspect of behaviour, environmental exposure, inborn biological characteristic etc. that is associated with the rate/frequency of occurrence of a health-related condition; typically risk factors will be associated with **increased** risk of a particular outcome (e.g. smoking status, age).

MEASURES OF DISEASE FREQUENCY

- **Disease Rate:** The **disease rate** is the relative frequency with which a disease event is observed in a defined population in a specified time period. Generically, the disease rate is defined by means of a “numerator/denominator” calculation.

$$\text{RATE} = \frac{\# \text{ EVENTS IN TIME PERIOD}}{\text{AVERAGE POP}^N \text{ SIZE DURING THE PERIOD}} \quad (1)$$

usually reported $\times 10^k$ for some k (that is, for example, per 1000 or per 100000 etc. population, per unit time). The denominator in the calculation can also be a “person-time” figure, that is, the total amount of time on study, aggregated for all individuals in the population; this allows calculations for studies in which each individual is tracked during the study period, rather than merely taking a population average.

- **Mortality Rate:** The **mortality rate** is the death rate in a defined population in a specified time period.

$$\frac{\# \text{ DEATHS}}{\# \text{ AT RISK OF DEATH}} \quad (2)$$

- **Morbidity:** Morbidity is the departure from health - change from disease-free to diseased - of an individual
- **Disease Incidence and Incidence Rate:** A disease **incident** is the observation of a **new** case of the disease. The **incidence rate** is the number of new cases in a defined population in a specified time period, divided by the at-risk population size, or person-time measure.

$$\frac{\# \text{ NEW CASES}}{\# \text{ AT RISK OF DEATH IN UNIT TIME}}$$

- **Disease Prevalence and Prevalence Rate:** **Disease Prevalence** is the number of individuals in a defined population at a specified time (**point prevalence**) or in a specified time period (**period prevalence**). The prevalence rate is the prevalence per number of individuals, in the specified period.

COMPARISON OF RATES AND RATE STANDARDIZATION

The objective of many epidemiological studies is to **compare** occurrence of the health-related condition in two or more populations, or for two or more levels of a risk factor. This is often achieved by comparison of rates.

Definition 1 *Crude, Specific and Standardized rates.*

- A **crude rate** is as defined in (1); the number of cases divided by the number of individuals at risk (in unit time)
- A **specific rate** is the rate defined for a particular identified sub population, or **stratum**
- A **standardized rate** is a population-level measure of incidence adjusted for population substructure, usually a weighted average of specific rates.

Suppose that two populations A and B say, have subpopulations defined by $K + 1$ partitioning points. Let N_A, N_B be the population sizes and

$$N_A = \sum_{k=1}^{K+1} N_{Ak} \quad N_B = \sum_{k=1}^{K+1} N_{Bk}$$

where (N_{Ak}, N_{Bk}) are the number of individuals in group $k = 1, \dots, K + 1$. Similarly, let

$$D_A = \sum_{k=1}^{K+1} D_{Ak} \quad D_B = \sum_{k=1}^{K+1} D_{Bk}$$

be the corresponding numbers of deaths (or incidences). Then we have

	A	B
CRUDE RATE	$R = \frac{D_A}{N_A}$	$\frac{D_B}{N_B}$
GROUP SPECIFIC RATE	$R_{.k} = \frac{D_{Ak}}{N_{Ak}}$	$\frac{D_{Bk}}{N_{Bk}}$

Note that

$$R_A = \sum_{k=1}^{K+1} \left(\frac{N_{Ak}}{N_A} \right) R_{Ak} \quad R_B = \sum_{k=1}^{K+1} \left(\frac{N_{Bk}}{N_B} \right) R_{Bk}$$

Definition 2 *Crude Rate Ratio.*

The **crude rate ratio**, **CRR**, is defined by

$$CRR(A, B) = \frac{R_A}{R_B}$$

Definition 3 *Group-specific Rate Ratio.*

The **group-specific rate ratio**, **SRR**, for group k is defined by

$$GSRR(A, B, k) = \frac{R_{Ak}}{R_{Bk}}$$

We seek a means of comparing populations A and B in a global sense whilst still recognizing the population substructure. We achieve this through **standardization**, which is carried out using either **direct** or **indirect** standardization. Consider some standardizing population S , of size N_S which has the same population substructure with N_{Sk} individuals and D_{Sk} deaths in group k , $k = 1, \dots, K + 1$.

Definition 4 Direct Standardized Rate.

The *directly standardized rate, DSR*, for population A standardized with respect to population S is

$$DSR(A; S) = R_A^{(S)} = \sum_{k=1}^{K+1} \left(\frac{N_{Sk}}{N_S} \right) R_{Ak} \quad (3)$$

and in general for any two populations P_1 and P_2 , the directly standardized rate for population P_1 standardized with respect to population P_2 is

$$DSR(P_1; P_2) = R_{P_1}^{(P_2)} = \sum_{k=1}^{K+1} \left(\frac{N_{P_2k}}{N_{P_2}} \right) R_{P_1k}$$

Definition 5 Directly Standardized Rate Ratio.

The *directly standardized rate ratio, SRR*, for population A against population B , standardized with respect to population S is

$$SRR(A, B; S) = \frac{R_A^{(S)}}{R_B^{(S)}} = \frac{\sum_{k=1}^{K+1} N_{Sk} R_{Ak}}{\sum_{k=1}^{K+1} N_{Sk} R_{Bk}} = \frac{\sum_{k=1}^{K+1} \left(\frac{N_{Sk}}{N_S} \right) R_{Ak}}{\sum_{k=1}^{K+1} \left(\frac{N_{Sk}}{N_S} \right) R_{Bk}} = \frac{\sum_{k=1}^{K+1} \left(\frac{N_{Sk}}{N_S} \right) \left(\frac{D_{Ak}}{N_{Ak}} \right)}{\sum_{k=1}^{K+1} \left(\frac{N_{Sk}}{N_S} \right) \left(\frac{D_{Bk}}{N_{Bk}} \right)}.$$

If S is chosen to be identical to B , then

$$SRR(A; S) \equiv SRR(A, S; S) = \frac{R_A^{(S)}}{R_S^{(S)}} = \frac{\sum_{k=1}^{K+1} \left(\frac{N_{Sk}}{N_S} \right) \left(\frac{D_{Ak}}{N_{Ak}} \right)}{\sum_{k=1}^{K+1} \left(\frac{N_{Sk}}{N_S} \right) \left(\frac{D_{Sk}}{N_{Sk}} \right)} = \frac{\sum_{k=1}^{K+1} N_{Sk} R_{Ak}}{\sum_{k=1}^{K+1} D_{Sk}} = \frac{\sum_{k=1}^{K+1} N_{Sk} R_{Ak}}{D_S}$$

and therefore

$$SRR(A; S) = \frac{R_A^{(S)}}{R_S^{(S)}} = \frac{\sum_{k=1}^{K+1} N_{Sk} R_{Ak}}{D_S} = \frac{\sum_{k=1}^{K+1} \left(\frac{N_{Sk}}{N_S} \right) R_{Ak}}{D_S / N_S} = \frac{\sum_{k=1}^{K+1} \left(\frac{N_{Sk}}{N_S} \right) R_{Ak}}{R_S} \quad (4)$$

Definition 6 Indirect Standardization.

The *indirectly standardized rate, ISR*, for population A standardized with respect to population S is defined in terms of the *expected* number of deaths in population A

$$E_A = \sum_{k=1}^{K+1} N_{Ak} R_{Sk}$$

and hence

$$ISR(A; S) = \frac{E_A}{N_A} = \frac{1}{N_A} \sum_{k=1}^{K+1} N_{Ak} R_{Sk} = \sum_{k=1}^{K+1} \left(\frac{N_{Ak}}{N_A} \right) R_{Sk}$$

Definition 7 Indirectly Standardized Rate Ratio.

The *indirectly standardized rate ratio, ISRR*, for population A against population B , standardized with respect to population S is

$$ISRR(A, B; S) = \frac{ISR(A; S)}{ISR(B; S)} = \frac{\sum_{k=1}^{K+1} \left(\frac{N_{Ak}}{N_A} \right) R_{Sk}}{\sum_{k=1}^{K+1} \left(\frac{N_{Bk}}{N_B} \right) R_{Sk}}$$

Definition 8 Standardized Mortality Ratio.

The *standardized mortality ratio, SMR*, for population A , standardized with respect to population S is

$$SMR(A; S) = \frac{D_A}{E_A} = \frac{\sum_{k=1}^{K+1} N_{Ak} R_{Ak}}{\sum_{k=1}^{K+1} N_{Ak} R_{Sk}} = \frac{\sum_{k=1}^{K+1} \left(\frac{N_{Ak}}{N_A} \right) R_{Ak}}{\sum_{k=1}^{K+1} \left(\frac{N_{Ak}}{N_A} \right) R_{Sk}} = \frac{R_A^{(A)}}{R_S^{(A)}} \quad (5)$$

The difference between direct and indirect standardization is now evident; direct standardization applies a weighting to the test population rates, with weights determined by the subpopulation structure in the **standardizing** population - from (3) and (4)

$$SRR(A; S) = \frac{\sum_{k=1}^{K+1} w_k^{(S)} R_{Ak}}{\sum_{k=1}^{K+1} w_k^{(S)} R_{Sk}} \quad \text{where } w_k^{(S)} = \frac{N_{Sk}}{N_S}.$$

Indirect standardization applies a weighting to the test and standardizing population rates, with weights determined by the subpopulation structure in the **test** population - from (5)

$$SMR(A; S) = \frac{\sum_{k=1}^{K+1} w_k^{(A)} R_{Ak}}{\sum_{k=1}^{K+1} w_k^{(A)} R_{Sk}} \quad \text{where } w_k^{(A)} = \frac{N_{Ak}}{N_A}.$$

2 × 2 TABLE REPRESENTATIONS FOR EPIDEMIOLOGICAL STUDIES

In the notation of the (prospective/retrospective) tree: let

- E be the event that an individual is exposed
- F be the event that an individual suffers failure during the study period
- S be the event that an individual is selected for inclusion in the study.

Cohort and case-control studies can be represented using the following combinations of events.

COHORT STUDY

	E	E'
F	$E \cap F$	$E' \cap F$
F'	$E \cap F'$	$E' \cap F'$

CASE-CONTROL STUDY

	$E \cap S$	$E' \cap S$
$F \cap S$	$E \cap F \cap S$	$E' \cap F \cap S$
$F' \cap S$	$E \cap F' \cap S$	$E' \cap F' \cap S$

The individuals in a cohort study are a random sample from the general population, whereas those in a case-control study are specifically selected according to their outcome (mortality/morbidity) status. The data from such a study can be represented by a 2×2 table

		EXPOSURE		
		E	E'	TOTAL
OUTCOME	F	n_{11}	n_{12}	$n_{1.}$
	F'	n_{21}	n_{22}	$n_{2.}$
	TOTAL	$n_{.1}$	$n_{.2}$	$n_{..}$

The probabilities of interest to be investigated are

$$\begin{aligned} \pi_1 &= P(F|E) && \text{the probability of failure in the EXPOSED group} \\ \pi_0 &= P(F|E') && \text{the probability of failure in the UNEXPOSED group} \end{aligned}$$

but also let $\theta = P(E)$ be the probability of exposure, and $\phi = P(F)$ be the probability of failure, in the population. In the case-control, it is assumed that S and E are conditionally independent given F , that is

$$P(S|E \cap F) = P(S|E' \cap F) = P(S|F) \qquad P(S|E \cap F') = P(S|E' \cap F') = P(S|F').$$

or equivalently

$$P(E|S \cap F) = P(E|S' \cap F) = P(E|F) \qquad P(E|S \cap F') = P(E|S' \cap F') = P(E|F').$$

In a case-control study, probabilities and conditional probabilities involving S are not available.

ESTIMATION OF KEY QUANTITIES

- COHORT STUDY:

$$P(E) \quad \text{by } \hat{\theta} = \frac{n_{.1}}{n_{..}} \qquad P(F|E) \quad \text{by } \hat{\pi}_1 = \frac{n_{11}}{n_{.1}}$$

$$P(F) \quad \text{by } \hat{\phi} = \frac{n_{1.}}{n_{..}} \qquad P(F|E') \quad \text{by } \hat{\pi}_0 = \frac{n_{12}}{n_{.2}}$$

- CASE-CONTROL STUDY: Consider first the ratio

$$\begin{aligned} \frac{P(E \cap F \cap S)}{P(E' \cap F \cap S)} &= \frac{P(E) P(F|E) P(S|E \cap F)}{P(E') P(F|E') P(S|E' \cap F)} = \frac{P(E)}{P(E')} \times \frac{P(F|E)}{P(F|E')} \times 1 \\ &= \frac{P(F)}{P(F')} \times \frac{P(E|F)}{P(E'|F)} \\ &= \frac{P(E|F)}{P(E'|F)} \end{aligned}$$

This is the *odds on exposure* in the cases. Hence, let

$$\Omega_1 = \frac{P(E|F)}{P(E'|F)} \qquad \Omega_0 = \frac{P(E|F')}{P(E'|F')}$$

be the odds on exposure in the cases and controls respectively, and finally let

$$\psi = \frac{\Omega_1}{\Omega_0}$$

be the *odds ratio*. A similar calculation involving the ratios

$$\frac{P(E \cap F \cap S)}{P(E \cap F' \cap S)} \qquad \frac{P(E' \cap F \cap S)}{P(E' \cap F' \cap S)}$$

can be used to define the *odds on being a case* in the exposed and unexposed groups respectively

$$\begin{aligned} \frac{P(E \cap F \cap S)}{P(E \cap F' \cap S)} &= \frac{P(E)}{P(E')} \times \frac{P(F|E)}{P(F'|E)} \times \frac{P(S|E \cap F)}{P(S|E \cap F')} = 1 \times \frac{P(S|E \cap F)}{P(S|E \cap F')} \times \frac{\pi_1}{1 - \pi_1} \quad (6) \\ \frac{P(E' \cap F \cap S)}{P(E' \cap F' \cap S)} &= \frac{P(E')}{P(E')} \times \frac{P(F|E')}{P(F'|E')} \times \frac{P(S|E' \cap F)}{P(S|E' \cap F')} = 1 \times \frac{P(S|E' \cap F)}{P(S|E' \cap F')} \times \frac{\pi_0}{1 - \pi_0}. \end{aligned}$$

But, for ψ , from above

$$\Omega_1 = \frac{P(E|F)}{P(E'|F)} = \frac{P(F|E)}{P(F'|E)} \frac{P(E)}{P(E')} = \frac{\pi_1}{\pi_0} \times \frac{\theta}{1 - \theta}$$

and

$$\Omega_0 = \frac{P(E|F')}{P(E'|F')} = \frac{P(F'|E)}{P(F|E')} \frac{P(E)}{P(E')} = \frac{1 - \pi_1}{1 - \pi_0} \times \frac{\theta}{1 - \theta}$$

and hence

$$\psi = \frac{\Omega_1}{\Omega_0} = \frac{\pi_1/\pi_0}{(1 - \pi_1)/(1 - \pi_0)} = \frac{\pi_1(1 - \pi_0)}{\pi_0(1 - \pi_1)}$$

Finally, as by assumption

$$\frac{P(S|E \cap F)}{P(S|E \cap F')} = \frac{P(S|E' \cap F)}{P(S|E' \cap F')}$$

we have that, from (6)

$$\frac{\omega_1}{\omega_0} = \frac{P(F|E)/P(F'|E)}{P(F|E')/P(F'|E')} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} \times \frac{P(S|E \cap F)/P(S|E \cap F')}{P(S|E' \cap F)/P(S|E' \cap F')} = \frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)} = \psi.$$

Therefore, by replacing the probabilities by observed frequencies

$$\begin{array}{ll} P(E \cap F \cap S) \text{ by } n_{11} & P(E' \cap F \cap S) \text{ by } n_{12} \\ P(E \cap F' \cap S) \text{ by } n_{21} & P(E' \cap F' \cap S) \text{ by } n_{22} \end{array}$$

we can estimate ψ by

$$\hat{\psi} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

but cannot estimate any other quantities.

MEASURES OF EFFECT

- *Rate of Incidence or Risk in Exposed and Unexposed Groups*

$$\pi_1 = P(F|E)$$

$$\pi_0 = P(F|E')$$

- *Odds on Incidence in Exposed and Unexposed Groups*

$$\omega_1 = \frac{\pi_1}{1-\pi_1}$$

$$\omega_0 = \frac{\pi_0}{1-\pi_0}$$

- *Risk Difference*

$$RD = \pi_1 - \pi_0$$

- *Risk Ratio or Relative Risk*

$$RR = \frac{\pi_1}{\pi_0}$$

- *Odds ratio*

$$OR = \frac{\pi_1/\pi_0}{(1-\pi_1)/(1-\pi_0)} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)}$$

- *Excess Relative Risk*

$$ERR = \frac{\pi_1 - \pi_0}{\pi_0}$$

- *Attributable Risk*

$$AR = \frac{\pi_1 - \pi_0}{\pi_1}$$