CHAPTER 1

INTRODUCTION TO BIOSTATISTICS AND EPIDEMIOLOGY

1.1 BIOSTATISTICS AND EPIDEMIOLOGY: TERMINOLOGY

Definition 1.1.1 ***Epidemiology***

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.;*

Definition 1.1.2 *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 (cause)
- population incidence
- identification of risk factors
- disease control/prevention elimination

Definition 1.1.3 ***Risk Factor***

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).*

1.1.1 MEASURES OF DISEASE FREQUENCY

Definition 1.1.4 ***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.

$$\text{RATE} = \frac{\text{\# EVENTS IN TIME PERIOD}}{\text{AVERAGE POP}^{\text{N}}\text{ SIZE DURING THE PERIOD}} \tag{1.1.1}$$

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.

Definition 1.1.5 ***Mortality Rate***

The ***mortality rate*** *is the death rate in a defined population in a specified time period.*

$$\frac{\#\ DEATHS}{\#\ AT\ RISK\ OF\ DEATHS}. \tag{1.1.2}$$

Definition 1.1.6 ***Morbidity***

Morbidity *is the departure from health - change from disease-free to diseased - of an individual*

Definition 1.1.7 ***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.*

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

Definition 1.1.8 ***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.*

1.1.2 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 1.1.9 ***Crude, Specific and Standardized rates.***

- *A* ***crude rate*** *is as defined in (1.1.1); 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

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

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

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

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

$$\begin{array}{lccc} & & \text{A} & \text{B} \\ \text{CRUDE RATE} & R = & \dfrac{D_A}{N_A} & \dfrac{D_B}{N_B} \\ \text{GROUP SPECIFIC RATE} & R_{.k} = & \dfrac{D_{Ak}}{N_{Ak}} & \dfrac{D_{Bk}}{N_{Bk}} \end{array}$$

Note that

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

Definition 1.1.10 ***Crude Rate Ratio.***

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

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

Definition 1.1.11 ***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, $k = 1, ..., K+1$.

Definition 1.1.12 ***Direct Standardized Rate.***

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

$$DSR\,(A; S) = R_A^{(S)} = \sum_{k=1}^{K+1} \left(\frac{N_{Sk}}{N_S} \right) R_{Ak} \tag{1.1.3}$$

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 1.1.13 ***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

$$SRR\,(A;S) = \frac{R_A^{(S)}}{R_S^{(S)}} = \frac{\sum_{k=1}^{K+1} N_{Sk}R_{Ak}}{D_S} = \frac{\sum_{k=1}^{K+1}\left(\frac{N_{Sk}}{N_S}\right)R_{Ak}}{D_S/N_S} = \frac{\sum_{k=1}^{K+1}\left(\frac{N_{Sk}}{N_S}\right)R_{Ak}}{R_S} \tag{1.1.4}$$

Definition 1.1.14 ***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

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

Definition 1.1.15 ***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 1.1.16 ***Standardized Mortality Ratio.***

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

$$SMR(A;S) = \frac{D_A}{E_A} = \frac{\sum_{k=1}^{K+1} N_{Ak} R_{Ak}}{\sum_{k=1}^{K+1} N_{Ak} R_{Sk}} = \frac{\sum_{k=1}^{K+1} \left(\frac{N_{Ak}}{N_A}\right) R_{Ak}}{\sum_{k=1}^{K+1} \left(\frac{N_{Ak}}{N_A}\right) R_{Sk}} = \frac{R_A^{(A)}}{R_S^{(A)}} \tag{1.1.5}$$

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 (1.1.3) and (1.1.4)

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

Indirect standardization applies a weighting to the test and standardizing population rates, with weights determined by the subpopulation structure in the **test** population - from (1.1.5)

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

1.2 TYPES OF EPIDEMIOLOGICAL STUDY AND THEIR ANALYSIS

1.2.1 ELEMENTARY STUDY DESIGN

Definition 1.2.1 ***Observational Study***

An ***observational study*** *is a study that does not involve any intervention by the experimenter, and where* ***exposure*** *to a risk factor is determined by nature.*

Definition 1.2.2 ***Experimental Study***

An ***experimental study*** *is a study in which the conditions, in particular exposure to the risk factor, are at the control of the experimenter.*

The two main types of observational study are the **cohort** and **case-control study.** Experimental studies include **clinical trials,** or **controlled** or **randomized controlled trials**.

Definition 1.2.3 ***Cohort Study***

A ***cohort study*** *is a study carried out on a defined population, where inclusion in the study* ***is not influenced*** *by exposure or outcome, and where either exposure is determined at the start of the study, and outcome determined subsequently (a* ***prospective*** *study), or both exposure and outcome are determined throughout the course of the study or* ***follow-up period*** *(a* ***retrospective*** *study).*

Definition 1.2.4 ***Case-Control Study***

*A **case-control study** is a study carried out on a defined population, where inclusion in the study is **completely determined** by outcome (whether the individual is a **case** or a **control**), and where exposure status is determined during the course of the study. A case-control study can be **retrospective** (with cases gathered from health records) or **prospective** (with cases included as the are recorded during the study period). The controls should be drawn from the same population as the cases, and should preferably be individually matched to the cases in terms of other risk or confouding factors.*

1.2.2 TREE REPRESENTATIONS FOR STUDY DESIGN

Let

- E be the event that an individual is exposed to the risk factor
- 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.

A **prospective** study design tree regards these events to occur in the sequence

$$E \rightarrow F \rightarrow S$$

whereas a **retrospective** study design tree regards these events to occur in the sequence

$$S \rightarrow F \rightarrow E$$

but in either case we will be principally interested in the comparison of incidence rates in the exposed and unexposed groups, that is, the conditional probabilities

$$P\left(F|E\right) \qquad P\left(F|E'\right). \tag{1.2.6}$$

Note that the prospective study is formulated in terms of the structured probability

$$P\left(E \cap F \cap S\right) = P(E)P(F|E)P(S|E \cap F)$$

that is, it includes the probabilities in (1.2.6) explicitly, whereas the retrospective study is formulated as

$$P\left(E \cap F \cap S\right) = P(S)P(F|S)P(E|F \cap S)$$

which does not.

It follows that we will attempt to estimate the probabilities

$$P\left(E \cap F \cap S\right)$$

and the other conditional probabilities from the study data.

1.2.3 2×2 TABLE REPRESENTATIONS FOR EPIDEMIOLOGICAL STUDIES

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		
OUTCOME	E	E'	TOTAL
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

$\pi_1 = P(F|E)$ the probability of failure in the EXPOSED group

$\pi_0 = P(F|E')$ the probability of failure in the UNEXPOSED group

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

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

or equivalently

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

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

1.2.4 COHORT AND CASE-CONTROL STUDIES

Consider a general tree-based version of an observational study; let S denote the inclusion of a subject in the study, E denote exposure and F denote incidence; then

$$P(E \cap F \cap S) = P(E)P(F|E)P(S|E \cap F). \tag{1.2.7}$$

We will use this factorization to deduce estimable quantities from different observational studies that comprise the S "margin" of a $2 \times 2 \times 2$ events table.

1.2.5 COHORT STUDY

In a cohort study, the defining feature is that E and F are independent of S so that (1.2.7) and the events table becomes

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

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

as the S and S' margins are **identical.** In particular,

$$P(E \cap F|S) = P\left(E \cap F|S'\right) = P(E \cap F)$$

and the S margin can be used to estimate the required probabilities. It follows that **all of the following quantities are estimable:**

- **MARGINAL PROBABILITIES OF EXPOSURE AND INCIDENCE**

$$\begin{aligned} \theta &= P(E) = P(E \cap F \cap S) + P\left(E \cap F' \cap S\right) + P\left(E \cap F \cap S'\right) + P\left(E \cap F' \cap S'\right) \\ &= P(E \cap F) + P\left(E \cap F'\right) \end{aligned}$$

and

$$\phi = P(F) = P(E \cap F) + P\left(E' \cap F\right)$$

with estimates

$$\widehat{\theta} = \frac{n_{.1}}{n_{..}} \qquad \widehat{\phi} = \frac{n_{1.}}{n_{..}}$$

- **INCIDENCE PROBABILITIES IN THE EXPOSED/UNEXPOSED GROUPS**

$$\begin{aligned} \pi_1 &= P(F|E) = \frac{P(E \cap F)}{P(E)} \\ \pi_0 &= P(F|E') = \frac{P(E' \cap F)}{P(E')} \end{aligned}$$

with estimates

$$\widehat{\pi}_1 = \frac{n_{11}}{n_{.1}} \qquad \widehat{\pi}_0 = \frac{n_{12}}{n_{.2}}$$

- **EXPOSURE PROBABILITIES IN THE CASE AND CONTROL GROUPS**

$$\begin{aligned} \gamma_1 &= P(E|F) = \frac{P(E \cap F)}{P(F)} \\ \gamma_0 &= P(E|F') = \frac{P(E \cap F')}{P(F')} \end{aligned}$$

with estimates

$$\widehat{\gamma}_1 = \frac{n_{11}}{n_{1.}} \qquad \widehat{\gamma}_0 = \frac{n_{21}}{n_{2.}}$$

- **ODDS ON INCIDENCE IN THE EXPOSED AND UNEXPOSED GROUPS**

$$\begin{aligned} \omega_1 &= \frac{\pi_1}{1-\pi_1} = \frac{P(E\cap F)}{P(E\cap F')} \\ \omega_0 &= \frac{\pi_0}{1-\pi_0} = \frac{P(E'\cap F)}{P(E'\cap F')} \end{aligned}$$

with estimates

$$\widehat{\omega}_1 = \frac{\widehat{\pi}_1}{1-\widehat{\pi}_1} = \frac{n_{11}}{n_{21}} \qquad \widehat{\omega}_0 = \frac{\widehat{\pi}_0}{1-\widehat{\pi}_0} = \frac{n_{12}}{n_{22}}$$

- **ODDS ON EXPOSURE IN THE CASE AND CONTROL GROUPS**

$$\begin{aligned} \Omega_1 &= \frac{\gamma_1}{1-\gamma_1} = \frac{P(E\cap F)}{P(E'\cap F)} \\ \Omega_0 &= \frac{\gamma_0}{1-\gamma_0} = \frac{P(E\cap F')}{P(E'\cap F')} \end{aligned}$$

with estimates

$$\widehat{\Omega}_1 = \frac{\widehat{\gamma}_1}{1-\widehat{\gamma}_1} = \frac{n_{11}}{n_{12}} \qquad \widehat{\Omega}_0 = \frac{\widehat{\gamma}_0}{1-\widehat{\gamma}_0} = \frac{n_{21}}{n_{22}}$$

- **ODDS RATIO**

$$\psi = \begin{cases} \dfrac{P(F|E)/P(F'|E)}{P(F|E')/P(F'|E')} = \dfrac{P(E\cap F)/P(E\cap F')}{P(E'\cap F)/P(E'\cap F')} = \dfrac{\omega_1}{\omega_0} = \dfrac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} \\[2ex] \dfrac{P(E|F)/P(E'|F)}{P(E|F')/P(E'|F')} = \dfrac{P(E\cap F)/P(E'\cap F)}{P(E\cap F')/P(E'\cap F')} = \dfrac{\Omega_1}{\Omega_0} = \dfrac{\gamma_1/(1-\gamma_1)}{\gamma_0/(1-\gamma_0)} \end{cases}$$

with estimate

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

1.2.6 CASE-CONTROL STUDY

In a case-control study, the defining feature is that E is independent of S **given** F and **given** F' so that (1.2.7) is unchanged, but we have the following relationships

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

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

Now, consider the estimation of

$$\Omega_1 = \frac{P(E|F)}{P(E'|F)} = \frac{\gamma_1}{1-\gamma_1}$$

We have from (1.2.7) that

$$\frac{P(E\cap F\cap S)}{P(E'\cap F\cap S)} = \frac{P(E)P(F|E)P(S|E\cap F)}{P(E')P(F|E')P(S|E'\cap F)} = \frac{P(F|E)P(E)}{P(F|E')P(E')} = \frac{P(E|F)}{P(E'|F)} \qquad (1.2.8)$$

as, by assumption, $P(S|E \cap F) = P(S|E' \cap F)$. A similar factorization is possible for

$$\Omega_0 = \frac{P(E|F')}{P(E'|F')} = \frac{\gamma_0}{1-\gamma_0}$$

However, if we try to proceed in the same way for the **incidence ratio** in cases and controls, then

$$\frac{\pi_1}{\pi_0} = \frac{P(F|E)}{P(F|E')} = \frac{P(E|F)}{P(E'|F)}\frac{P(E')}{P(E)} = \Omega_1 \left(\frac{1-\theta}{\theta}\right) \tag{1.2.9}$$

and the simplification cannot proceed further; we have no way of estimating θ, the exposure rate in the population. Furthermore, for the **odds on incidence** in the exposed group, we might try a similar approach to above and examine

$$\frac{P\left(E \cap F \cap S\right)}{P\left(E \cap F' \cap S\right)} = \frac{P(E)P(F|E)P(S|E \cap F)}{P(E)P(F'|E)P(S|E \cap F')} = \frac{P(F|E)P(S|E \cap F)}{P(F|E')P(S|E \cap F')} = \omega_1 \frac{P(S|E \cap F)}{P(S|E \cap F')} \tag{1.2.10}$$

and again the simplification cannot proceed further as

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

is indeterminate. However

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

and hence

$$\frac{P\left(E \cap F \cap S\right)/P\left(E \cap F' \cap S\right)}{P\left(E' \cap F \cap S\right)/P\left(E' \cap F' \cap S\right)} = \frac{\gamma_1/\left(1-\gamma_1\right)}{\gamma_0/\left(1-\gamma_0\right)} = \frac{\Omega_1}{\Omega_0} = \psi$$

and also, from (1.2.8)

$$\frac{P(E|F)/P(E'|F)}{P(E|F')/P(E'|F')} = \frac{\dfrac{P(F|E)}{P(F|E')}\dfrac{P(E)}{P(E')}}{\dfrac{P(F'|E)}{P(F'|E')}\dfrac{P(E)}{P(E')}} = \frac{P(F|E)/P(F|E')}{P(F'|E)/P(F'|E')} = \frac{\pi_1/\pi_0}{\left(1-\pi_1\right)/\left(1-\pi_0\right)} = \frac{\pi_1/\left(1-\pi_1\right)}{\pi_0/\left(1-\pi_0\right)} = \psi \tag{1.2.11}$$

and finally, by (1.2.11)

$$\frac{P(E|F)/P(E'|F)}{P(E|F')/P(E'|F')} = \frac{P(F|E)/P(F'|E)}{P(F|E')/P(F'|E')} = \frac{\pi_1/\left(1-\pi_1\right)}{\pi_0/\left(1-\pi_0\right)} = \frac{\omega_1}{\omega_0} = \psi$$

It follows that **only the following quantities are estimable in the absence of other knowledge**

- **EXPOSURE PROBABILITIES IN THE CASE AND CONTROL GROUPS**

$$\begin{aligned}\gamma_1 &= P(E|F) = \frac{P(E \cap F)}{P(F)} \\ \gamma_0 &= P(E|F') = \frac{P(E \cap F)}{P(F')}\end{aligned}$$

with estimates

$$\widehat{\gamma}_1 = \frac{n_{11}}{n_{1.}} \qquad \widehat{\gamma}_0 = \frac{n_{21}}{n_{2.}}$$

- **ODDS ON EXPOSURE IN THE CASE AND CONTROL GROUPS**

$$\begin{aligned}\Omega_1 &= \frac{\gamma_1}{1-\gamma_1} = \frac{P(E\cap F)}{P(E'\cap F)}\\ \Omega_0 &= \frac{\gamma_0}{1-\gamma_0} = \frac{P(E\cap F')}{P(E'\cap F')}\end{aligned}$$

with estimates

$$\widehat{\Omega}_1 = \frac{\widehat{\gamma}_1}{1-\widehat{\gamma}_1} = \frac{n_{11}}{n_{12}} \qquad \widehat{\Omega}_0 = \frac{\widehat{\gamma}_0}{1-\widehat{\gamma}_0} = \frac{n_{21}}{n_{22}}$$

- **ODDS RATIO**

$$\psi = \begin{cases} \dfrac{P(F|E)/P(F'|E)}{P(F|E')/P(F'|E')} = \dfrac{P(E\cap F)/P(E\cap F')}{P(E'\cap F)/P(E'\cap F')} = \dfrac{\omega_1}{\omega_0} = \dfrac{\pi_1/\left(1-\pi_1\right)}{\pi_0/\left(1-\pi_0\right)} \\ \\ \dfrac{P(E|F)/P(E'|F)}{P(E|F')/P(E'|F')} = \dfrac{P(E\cap F)/P(E'\cap F)}{P(E\cap F')/P(E'\cap F')} = \dfrac{\Omega_1}{\Omega_0} = \dfrac{\gamma_1/\left(1-\gamma_1\right)}{\gamma_0/\left(1-\gamma_0\right)} \end{cases}$$

with estimate

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

EXAMPLE: THE LIMITATIONS OF CASE CONTROL STUDIES
An illustration of why case-control studies are limited in their usefuleness is presented below; fixing $\gamma_1 = 0.2$ and $\gamma_0 = 0.1$ and changing the size of the CONTROLS group

TABLE 1

	$E\cap S$	$E'\cap S$	TOTAL
CASES	20	80	100
CONTROLS	100	900	1000
TOTAL	120	980	1100

TABLE 2

	$E\cap S$	$E'\cap S$	TOTAL
CASES	20	80	100
CONTROLS	500	4500	5000
TOTAL	520	4580	5100

Then clearly if we estimate γ_1 and γ_0, we recover the true values 0.2 and 0.1, and in each case

$$\text{TABLE 1: } \widehat{\psi} = \frac{20\times 900}{80\times 100} = \frac{9}{4} \qquad \text{TABLE 2: } \widehat{\psi} = \frac{20\times 4500}{80\times 500} = \frac{9}{4}$$

but if we try to estimate, for example π_1 and π_0 in the same way that we would for a cohort study, we get different results from the two tables

$$\text{TABLE 1:} \quad \widehat{\pi}_1 = \frac{20}{120} = \frac{1}{6} \qquad \widehat{\pi}_0 = \frac{80}{980} = \frac{4}{49}$$

$$\text{TABLE 2:} \quad \widehat{\pi}_1 = \frac{20}{520} = \frac{1}{26} \qquad \widehat{\pi}_0 = \frac{80}{4580} = \frac{4}{229}$$

This result follows from the fact that the row totals, corresponding to the total numbers of cases and controls, $n_{1.}$ and $n_{2.}$, are fixed by the experimenter, and we do **not have a random sample of exposed and unexposed individuals** from the population. In a cohort study, only the total cohort size, $n_{..}$, is fixed.

1.3 MEASURES OF EFFECT

Below is presented a comprehensive list of possible measure of effect that may be of interest in different contexts:

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

$$\begin{aligned}\pi_1 &= P(F|E) \\ \\ \pi_0 &= P(F|E')\end{aligned}$$

- *Odds on Incidence* in Exposed and Unexposed Groups

$$\begin{aligned}\omega_1 &= \frac{\pi_1}{1-\pi_1} \\ \omega_0 &= \frac{\pi_0}{1-\pi_0}\end{aligned}$$

- *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*

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

Recall from previous discussions that all of these quantities are estimable from cohort studies, as the parameters π_0, π_1 are estimable, along with the other parameters identified in sections 1.2.5 and 1.2.6 θ, ϕ and γ_0, γ_1 and ψ. Case-control studies, however, are only able to give information about γ_0, γ_1 and ψ, and functions of these parameters. Under the *rare disease hypothesis*, however, where ϕ, and hence π_0, π_1 are small, case-control studies can be used to learn (approximately) about the risk ratio/relative risk π_1/π_0; this follows as, in that case, for $i = 0, 1$,

$$\frac{\pi_i}{1-\pi_i} = \pi_i(1-\pi_i)^{-1} = \pi_i + \pi_i^2 + \pi_i^3 + ... \approx \pi_i$$

so that

$$\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} \approx \frac{\pi_1}{\pi_0}.$$

1.4 PROBABILITY MODELS FOR EPIDEMIOLOGICAL STUDIES

1.4.1 BINOMIAL MODEL

Discrete random variable X counts the number of failures in a sequence of independent and identical binary (0/1, success/failure) trials.

$$f_X(x) = \binom{n}{x}\theta^x (1-\theta)^{n-x} \qquad x = 0, 1, ..., n$$

for $0 \leq \theta \leq 1$. Then

$$E_{f_X}[X] = n\theta \qquad Var_{f_X}[X] = n\theta(1-\theta)$$

1.4.2 POISSON MODEL

Discrete random variable X is suitable for count data

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x = 0, 1, 2, ...$$

for $\lambda > 0$. Then

$$E_{f_X}[X] = \lambda \qquad Var_{f_X}[X] = \lambda.$$

The Poisson distribution is the limiting case of the binomial as $n \to \infty$, with $\lambda = n\theta$ fixed.

1.4.3 GAMMA MODEL

Continuous random variable X

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad x \in \mathbb{R}^+$$

for $\lambda > 0$. Then

$$E_{f_X}[X] = \frac{\alpha}{\beta} \qquad Var_{f_X}[X] = \frac{\alpha}{\beta^2}$$

Special case: If $\alpha = \nu/2$ for some integer ν, and $\beta = 1/2$, then

$$Gamma\left(\frac{\nu}{2}, \frac{1}{2}\right) \equiv \chi^2_\nu$$ the **Chi-squared distribution with νdegrees of freedom.**

1.4.4 NORMAL MODEL

Continuous random variable X

$$f_X(x) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \qquad x \in \mathbb{R}$$

for $\lambda > 0$. Then

$$E_{f_X}[X] = \mu \qquad Var_{f_X}[X] = \sigma^2$$

The Normal distribution can be used as an approximation to binomial and Poisson models using the Central Limit Theorem,

$$X \sim Binomial(n, \theta) \Longrightarrow X \dot{\sim} Normal(n\theta, n\theta(1-\theta))$$

$$X \sim Poisson(\lambda) \Longrightarrow X \dot{\sim} Normal(\lambda, \lambda)$$

1.4.5 BETA MODEL

Continuous random variable X on $[0,1]$

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad x \in [0,1]$$

for $\alpha, \beta > 0$. Then

$$E_{f_X}[X] = \frac{\alpha}{\alpha+\beta} \qquad Var_{f_X}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

1.4.6 MULTINOMIAL MODEL

Generalization of the binomial; instead of two possible outcomes on each experiment, suppose that there are K; suppose that n experiments are carried out. Let $X = (X_1, ..., X_K)$ be the counts of numbers of each type of result, that is, X_i is the discrete random variable counting the number of times outcome i is observed. The multinomial model the vector random variable X has joint mass function for vector $x = (x_1, ..., x_K)$

$$f_X(x) = \binom{n}{x_1, x_2, ..., x_K} \theta_1^{x_1}\theta_2^{x_2}...\theta_K^{x_K}$$

where $0 \leq \theta_i \leq 1$ and $\theta_1 + ... + \theta_K = 1$ and

$$\binom{n}{x_1, x_2, ..., x_K} = \frac{n!}{x_1!x_2!...x_K!}.$$

Can show that marginally, for $i = 1, ..., K$, $X_i \sim Binomial(n, \theta_i)$ and conditionally

$$X_i | X_j, j \neq i \sim Binomial\left(n - \sum_{j\neq i} x_j, \frac{\theta_i}{1 - \sum\limits_{j \neq i}\theta_j}\right)$$

1.4.7 DIRICHLET MODEL

Generalization of the beta; let $X = (X_1, ..., X_K)$. The Dirichlet model has joint pdf for the vector vector $x = (x_1, ..., x_K)$ such that

$$0 \leq x_i \leq 1 \qquad \sum_{i=1}^{K} x_i = 1$$

given by

$$f_X(x) = \frac{\Gamma(\alpha_1 + \alpha_2 + ... + \alpha_K)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_K)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} ... x_K^{\alpha_K - 1}$$

where $\alpha_i > 0$ for all i. Can show that marginally, for $i = 1, ..., K$

$$X_i \sim Beta\left(\alpha_i, \sum_{j \neq i}\alpha_j\right)$$

and conditionally

$$X_i | X_j, j \neq i \sim \left(1 - \sum_{j\neq i}\alpha_j\right) \times Beta(\alpha_i, \alpha_K)$$

1.4.8 MULTIVARIATE NORMAL

Generalization of the normal, constructed as follows; let $Z_1, ..., Z_K$ be independent and identically distributed standard $(\mu = 0, \sigma = 1)$Normal random variables; define the transformed vector variable

$$X = VZ + \mu$$

where $Z = (Z_1, ..., Z_K)^T$ and $\mu = (\mu_1, ..., \mu_K)^T$ are $(K \times 1)$ column vectors, and V is a $K \times K$ matrix. Then it can be shown that

$$f_X(x) = \frac{1}{(2\pi)^{K/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}$$

where $\Sigma = V^T V$ is a $K \times K$ symmetric matrix. This is the joint pdf of the multivariate normal, and we write

$$X \sim N_K(\mu, \Sigma)$$

It can be shown that

$$E_{f_X}[X] = \mu \qquad Var_{f_X}[X] = \Sigma$$

that is, Σ is the variance-covariance matrix, so that

$$Cov_{f_{X_i,X_j}}[X_i, X_j] = [\Sigma]_{ij}.$$

Some other results also follow from standard theory.

- If A is a $D \times K$ matrix, and $Y = AX$, then

$$Y \sim N_D\left(A\mu, A\Sigma A^T\right)$$

- If X is partitioned $\left(X_{(1)}, X_{(2)}\right)$ where $X_{(1)}$ is a K_1-dimensional vector, $X_{(2)}$ is a K_2-dimensional vector, with $K = K_1 + K_2$; then writing

$$\mu = \left(\mu_{(1)}, \mu_{(2)}\right)^T \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where Σ_{11} is $(K_1 \times K_1)$, $\Sigma_{12} = \Sigma_{21}^T$ is $(K_1 \times K_2)$ and Σ_{22} is $(K_2 \times K_2)$, we have that marginally

$$X_{(1)} \sim N_{K_1}\left(\mu_{(1)}, \Sigma_{11}\right)$$

in particular

$$X_i \sim N\left(\mu_i, \sigma_i^2\right)$$

and conditionally

$$X_{(1)}|X_{(2)} = x_{(2)} \sim N_{K_1}\left(\mu_{(1)} + \Sigma_{12}\Sigma_{22}^{-1}\left(x_{(2)} - \mu_{(2)}\right), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$