

## Analyzing simple epidemiologic data

### 4.1 Overview and concepts

In this chapter we focus on analyzing one- and two-sample measurements for risk or prevalence data, incidence rate data, and case-control data. At the point of analyzing data that has already been collected we are assuming that the study subjects are representative of the target population (minimal selection bias), and that the data was measured accurately (minimal measurement error). Now, we are interested in these basic questions:

1. What is the point estimate? (estimation)
2. What is the variability of this estimate? (confidence interval)
3. Is this estimate consistent with a reference value? ( $p$  value and Type I error)
4. What is the chance of detecting a meaningful difference from the reference, if one exists? (power and Type II error)
5. How many subjects are required to be able to detect a meaningful difference, if one exists? (sample size)

The point estimate enables epidemiologic inference (What is the risk of disease occurrence or the prevalence of a condition?). The precision or variability of the estimate enables statistical inference (Does random error limit valid epidemiologic inferences?). Next, we might be interested in knowing whether this point estimate is consistent with a reference value. Finally, if it is “consistent,” was there a sufficient sample size to detect a meaningful difference if one existed? We don’t want to mistakenly infer no difference when there was insufficient data to draw this conclusion. In summary, we need to address estimation, precision, comparisons, and sample size.

#### 4.1.1 Estimation

In epidemiology, we are generally interested in the following measures of occurrence:

- Prevalence (number of existing cases)
- Incidence (occurrence of new cases)
- Time until an event occurs
- Rate
- Risk

#### 4.1.2 Confidence intervals

The width of confidence intervals provide information on the precision and/or variability of our estimation process. For starters, we are generally interested in calculating a 95% confidence interval (CI). But what is a 95% CI? A 95% confidence interval represents a statistical procedure, that if it could be repeated many times, we would construct intervals that would cover the “true” value 95% of the time, on average. The procedure we use depends on what we know or what we assume about the distribution of a specific measure.

First, does this measure follow a known distribution? Second, if we are not sure, can we safely assume it follows a known distribution? Third, can we directly construct a confidence interval using this known distribution? Using today’s computers, this is usually straightforward. Fourth, can we construct a reasonable confidence interval using a normal distribution approximation? Fifth, if we have no idea of the underlying distribution of a specific measure, can we use simulation methods to construct the confidence interval?

In general, we use the following methods to construct confidence intervals for epidemiologic measures:

- Normal distribution approximation methods
- Exact methods using known distributions
- Exact method approximations using derived formulas
- Resampling (simulation) methods for unknown distributions

#### Confidence interval using normal approximation

Epidemiologists are familiar with the normal distribution to construct approximate confidence interval. For some estimate, say  $R$ , the confidence interval is constructed using this formula:

$$R_L, R_U = R \pm Z \times SE(R), \quad (4.1)$$

where  $\pm Z$  are the quantile values of the standard normal distribution density curve such that the area between  $-Z$  and  $+Z$  is equal to the confidence level.  $SE(R)$  is the standard error of the measure  $R$ . For a 95% confidence interval,  $P(-Z \leq z \leq +Z) = 0.95$ , where  $Z = 1.96$ . This means that  $P(Z \leq -1.96) = P(z \geq 1.96) = 0.025$ . In truth, 1.96 is only an approximation. To get the precise  $Z$  quantile values that correspond to a specific confidence level, we use the `qnorm` function. For a given probability  $p$  (area under the normal distribution density curve), where  $p = P(Z \leq q)$ , the `qnorm` function

calculates the corresponding quantile value  $q$ . The following code does the trick:

```
> conf.level <- 0.90
> qnorm((1 + conf.level)/2)
[1] 1.644854
> conf.level <- 0.95
> qnorm((1 + conf.level)/2)
[1] 1.959964
> conf.level <- 0.99
> qnorm((1 + conf.level)/2)
[1] 2.575829
```

Now, the  $SE(R)$  is often calculated from some formula that has been derived. Recall that the standard error is the standard deviation of the means from theoretical repeated sampling from the same distribution. We'll see examples, in the sections that follow.

### Confidence interval using exact and exact approximations

If we know (or assume to know) the underlying distribution, then it is possible to calculate an exact confidence interval. This is achieved letting a modern computer do all the work. We will see three exact confidence interval examples:

- Poisson count using the tail method
- Binomial proportion using R's (`binom.test`) function
- Risk ratio measure using a resampling method (bootstrapping)

Before the wide availability of computing power, formulas were derived to approximate exact confidence intervals for selected epidemiologic measures. We'll see a few examples in the sections that follow.

#### 4.1.3 $p$ values and Type I error

A natural question that arises from evaluating an epidemiologic measure is: how consistent is it compared to some reference value? For example, suppose that  $\hat{R}$  proportion of hospitalized patients suffered an adverse event last year, and our goal was to have not more than  $R_0$  proportion experiencing an adverse event. How consistent is our experience ( $\hat{R}$ ) with a reference value of  $R_0$ . To answer this, we test the null hypothesis that  $R_0 = R$ . Under this null hypothesis we calculate the probability  $p$  of observing a value of  $\hat{R}$  or greater:  $p = P(R \geq \hat{R})$ . If this  $p$  value is high, then our experience ( $\hat{R}$ ) is more consistent with the null hypothesis. If this  $p$  value is low, then our experience is less consistent with the null hypothesis. Similar to confidence intervals, to calculate the  $p$  value we must either know or make assumptions about the underlying distribution.

How low should the  $p$  value be for us to conclude that  $\hat{R}$  is not consistent with  $R_0$ ? There is no hard fast rule. However, sometimes a decision rule is used to reject the null hypothesis if  $p \leq \alpha$ , where  $\alpha$  is arbitrarily set to be small. Therefore, if the null hypothesis is true, we are willing to incorrectly reject the null hypothesis  $\alpha\%$  of the time. This is called the Type I error and is often set at  $\alpha = 0.05$ .

Unfortunately, a  $p$  value is driven by both the sample size and the effect size (difference between  $R_0$  and  $\hat{R}$ ). With very large sample sizes even small, nonimportant effect sizes can result in very small  $p$  values. Likewise, small sample sizes can result in large  $p$  values and suggest that an observed measure is consistent with a reference value. We address these issues in the next sections.

This has been an example of a one-sample measure ( $\hat{R}$ ). These concepts equally apply to two-sample measures of association such as rate ratios, risk ratios, and odds ratios. The reference value for these measures of association will be the null value 1 for no association.

## The $p$ value function

### 4.1.4 Power and Type II error

### 4.1.5 Sample size calculations

Sample size calculation integrate all the previous concepts:

- Select an effect size?
- Select  $\alpha$  (Type I error)
- Select  $(1 - \beta)$  (Type II error)
- Calculate the sample size

## 4.2 Evaluating a single measure of occurrence

In terms of implementing analytic methods using R, we use the following steps:

1. Assign input values
2. Do calculations
3. Collect results

Using this approach prepares one for writing functions later.

### 4.2.1 Poisson count (incidence) and rate data

In this section, we address the count of new cases (incidence). For convenient, we assume the occurrence of new cases follows a Poisson distribution. The Poisson distribution is a discrete probability distribution with the following density function:

$$P(X = x) = \frac{x^{-\lambda} \lambda^x}{x!}, \quad (4.2)$$

where  $X$  is the random variable,  $x$  is the observed count, and  $\lambda$  is the expected count. And, here is the Poisson distribution function:

$$P(X \leq x) = \sum_{k=0}^x \frac{k^{-\lambda} \lambda^k}{k!}. \quad (4.3)$$

In R, we use the `dpois` and `ppois` functions for the density and distribution functions, respectively. For example, in the United States, the rate of meningococcal disease is about 1 case per 100,000 population per year [4]. In San Francisco, with a population of about 800,000, we expected about 8 cases per year ( $\lambda$ ). In a given year, what is the probability of observing exactly 6 cases? Using the `dpois` function,

```
> dpois(x = 6, lambda = 8)
[1] 0.1221382
```

there is about a 12% chance of observing 6 cases exactly. In a given year, what is the probability of observing more than 10 cases? The chance of observing more than 10 cases is the  $1 - P(X \leq 10)$ :

```
> 1 - ppois(q = 10, lambda = 8)
[1] 0.1841142
```

Therefore, there is about an 18% chance of observing more than 10 cases in a given year.

### Estimation

In epidemiology, we calculate the average or per-capita rate ( $r$ ) as the count of new cases ( $x$ ) divided by the person-time at risk ( $PT$ ).

$$r = \frac{x}{PT} \quad (4.4)$$

Again, for convenience, we assume the count  $x$  has a Poisson distribution, and we consider  $PT$  a fixed quantity.

## Confidence intervals

### *Normal approximation*

Here is the standard error of an incidence rate, and the formula to construct a confidence interval using a normal approximation:

$$\begin{aligned} \text{SE}(r) &= \sqrt{x/PT^2} \\ r_L; r_U &= r \pm Z \times \text{SE}(r) \end{aligned}$$

Consider 8 cases of cancer from 85,000 person-years at risk for a rate of 9.4 cases per 100,000 person-years [3]. Here we calculate a 90% confidence interval using a normal approximation method,

```
> #assign input value
> conf.level <- 0.90
> Z <- qnorm(0.5*(1 + conf.level))
> x <- 8
> PT <- 85000
> mult <- 100000
> #do calculations
> r <- x/PT
> SE.r <- sqrt(x/PT^2)
> LL <- r - Z*SE.r
> UL <- r + Z*SE.r
> #collect results
> cbind(x, PT, rate=mult*r, lower=mult*LL, upper=mult*UL)
      x   PT  rate lower upper
[1,] 8 85000 9.4118 3.9384 14.885
```

If we plan on using this method often, it is convenient to convert your steps into a function:

```
cipois.norm <- function(x, PT = 1, conf.level = 0.95, mult = 1){
  Z <- qnorm(0.5*(1 + conf.level))
  r <- x/PT
  SE.r <- sqrt(x/PT^2)
  LL <- r - Z*SE.r
  UL <- r + Z*SE.r
  cbind(x, PT, rate=mult*r, lower=mult*LL, upper=mult*UL,
        conf.level=conf.level, multiplier=mult)
}
```

Notice, we followed the same steps: assign input values (default values were assigned to function arguments), do calculations, and collect results (into a matrix object). We set the default PT=1, which calculate a confidence interval for a Poisson count. We also added a multiplier in order to change the rate to a more interpretable number (e.g., rate per 100,00 person-years). Here we test our function:

```
> cipois.norm(8, 85000, 0.90, 100000)
      x    PT  rate  lower  upper conf.level multiplier
[1,] 8 85000 9.4118 3.9384 14.885      0.9      1e+05
```

### Exact approximation

For low counts, an exact confidence interval is more accurate. A normal distribution is symmetric; however, a low count will have an asymmetric distribution. We will use the Poisson distribution to calculate an exact confidence, but first we start with an exact approximation to the Poisson distribution using Byar's method [3].

*Byar's confidence limits:*

$$r_L, r_U = (x + 0.5) \left( 1 - \frac{1}{9(x + 0.5)} \pm \frac{Z}{3} \sqrt{\frac{1}{(x + 0.5)}} \right)^3 / PT \quad (4.5)$$

Here is Byar's method converted into a function:

```
cipois.byar <- function(x, PT = 1, conf.level = 0.95, mult = 1) {
  Z <- qnorm(0.5*(1+conf.level))
  Zinsert <- (Z/3)*sqrt(1/(x+0.5))
  r <- x/PT
  LL <- ((x+0.5)*(1-1/(9*(x+0.5))-Zinsert)^3)/PT
  UL <- ((x+0.5)*(1-1/(9*(x+0.5))+Zinsert)^3)/PT
  cbind(x, PT, rate=mult*r, lower=mult*LL, upper=mult*UL,
        conf.level=conf.level, multiplier=mult)
}
```

Suppose we observe a rate of 3 cases per 2500 person-years (12 cases per 10,000 person-years). What is the 90% confidence interval using Byar's method:

```
> cipois.byar(3, 2500, 0.90, 10000)
      x    PT rate  lower  upper conf.level multiplier
[1,] 3 2500  12 3.3446 25.437      0.9      10000
```

### Exact methods

*The tail method:*

```
cipois.exact <- function(x, PT = 1, conf.level = 0.95, mult = 1) {
  f1 <- function(x, ans, alpha = alp) {
    ppois(x, ans) - alpha/2
  }
  f2 <- function(x, ans, alpha = alp) {
    1 - ppois(x, ans) + dpois(x, ans) - alpha/2
  }
}
```

```

alp <- 1 - conf.level
interval <- c(0, x * 9)
r <- x/PT
UL <- uniroot(f1, interval = interval, x = x)$root/PT
if(x == 0) {
  LL <- 0
} else {
  LL <- uniroot(f2, interval = interval, x = x)$root/PT
}
cbind(x, PT, rate=mult*r, lower=mult*LL, upper=mult*UL,
      conf.level=conf.level, multiplier=mult)
}

```

Suppose we observe a rate of 3 cases per 2500 person-years (12 cases per 10,000 person-years). What is the 90% confidence interval using exact tail method for the Poisson distribution?

```

> cipois.exact(3, 2500, 0.90, 10000)
      x  PT rate lower upper conf.level multiplier
[1,] 3 2500  12 3.2708 31.015      0.9      10000

```

*The Daly method:* In 1992, Dr. Leslie Daly published a SAS macro that uses the Gamma distribution to calculate exact confidence intervals for a Poisson count. The SAS macro was translated into the R language.

```

cipois.daly <- function(x, PT = 1, conf.level = 0.95, mult = 1) {
  r <- x/PT
  if(x != 0) {
    LL <- qgamma((1 - conf.level)/2, x)/PT
    UL <- qgamma((1 + conf.level)/2, x + 1)/PT
  } else {
    if(x == 0) {
      LL <- 0
      UL <- -log(1 - conf.level)/PT
    }
  }
  cbind(x, PT, rate=mult*r, lower=mult*LL, upper=mult*UL,
        conf.level=conf.level, multiplier=mult)
}

```

Suppose we observe a rate of 3 cases per 2500 person-years (12 cases per 10,000 person-years). What is the 90% confidence interval using exact Daly method for the Poisson distribution?

```

> cipois.daly(3, 2500, 0.90, 10000)
      x  PT  rate lower upper conf.level multiplier
[1,] 3 2500 0.94118 3.2708 31.015      0.9      10000

```

**Comparison**

Comparison to a “fixed” reference value.

**Power and sample size****4.2.2 Binomial risk and prevalence data****Estimation**

$$R = \frac{x}{N} \quad (4.6)$$

**Confidence intervals**

*Normal approximation*

$$SE(R) = \sqrt{\frac{x(N-x)}{N^3}} \quad (4.7)$$

R does have a function for testing a one sample proportions. However, it was more informative to learn how to create your own function using known statistical formulas. Why? Because R or another software package may not have the specific function you need, and by learning how to create you own functions you will be able to solve many more problems effectively and efficiently. Here is the same analysis using R’s prop.test function (which additionally provides a confidence interval).

```
> prop.test(x=39, n=215, p=.15)

1-sample proportions test with continuity correction

data: 39 out of 215, null probability 0.15
X-squared = 1.425, df = 1, p-value = 0.2326
alternative hypothesis: true p is not equal to 0.15
95 percent confidence interval:
 0.1335937 0.2408799
sample estimates:
      p
0.1813953
```

*Exact approximation*

*Wilson’s confidence limits:* Baby Rothman, p. 132.

$$R_L, R_U = \frac{N}{N + Z^2} \left[ \frac{x}{N} + \frac{Z^2}{2N} \pm Z \sqrt{\frac{x(N-x)}{N^3} + \frac{Z^2}{4N^2}} \right] \quad (4.8)$$

**Table 4.1.** Cohort study, person-time data

	Exposure	
	Exposed	Unexposed
Number of new cases	$x_1$	$x_0$
Person-time at risk	$PT_1$	$PT_0$

*Exact methods*

To assess whether an observed one sample proportion ( $R = x/N$ ) differs from an alternative value one can calculate a p value based on the binomial distribution (`binom.test`) or a normal distribution approximation to the binomial distribution (covered in previous section). The approximation using the normal distribution is satisfactory when the expected number of “successes” ( $x$ ) and the “failures” ( $N-x$ ) are both larger than 5. Now, here is the same one sample analysis but using an exact method with R’s `binom.test` function:

```
> binom.test(x=39, n=215, p=.15)
```

```
Exact binomial test
```

```
data: 39 and 215
number of successes = 39, number of trials = 215, p-value = 0.2135
alternative hypothesis: true probability of success is not equal to 0.15
95 percent confidence interval:
 0.1322842 0.2395223
sample estimates:
probability of success
 0.1813953
```

**Comparison****Power and sample size****4.3 Evaluating two measures of occurrence****4.3.1 Comparing two rate estimates: Rate ratio****Estimation**

$$rr = \frac{r_1}{r_0} = \frac{x_1/PT_1}{x_0/PT_0} \quad (4.9)$$

**Table 4.2.** Cohort study, binomial data

	Exposure	
	Exposed	Unexposed
Number of new cases	$x_1$	$x_0$
Persons at risk	$N_1$	$N_0$

**Confidence intervals**

*Normal approximation*

$$SE[\log(rr)] = \sqrt{\frac{1}{x_1} + \frac{1}{x_0}} \tag{4.10}$$

*Exact approximation*

*Exact methods*

**Comparison**

**Power and sample size**

**4.3.2 Comparing two risk estimates: Risk ratio and disease odds ratio**

```
##Table set up
##      Disease
##Exposure Yes No Total
##   Yes  x1  .  n1
##   No  x0  .  n0
```

**Estimation**

$$RR = \frac{R_1}{R_0} = \frac{x_1/N_1}{x_0/N_0} \tag{4.11}$$

$$DOR = \frac{R_1/(1 - R_1)}{R_0/(1 - R_0)} = \frac{x_1(N_0 - x_0)}{x_0(N_1 - x_1)} \tag{4.12}$$

**Confidence intervals**

$$SE[\log(RR)] = \sqrt{\frac{1}{x_1} - \frac{1}{N_1} + \frac{1}{x_0} - \frac{1}{N_0}} \tag{4.13}$$

*Normal approximation*

```

###Risk Ratio CI from baby Rothman, p. 135
rr.wald <- function(x, conf.level = 0.95){
  ##prepare input
  x1 <- x[1,1]; n1 <- sum(x[1,])
  x0 <- x[2,1]; n0 <- sum(x[2,])
  ##calculate
  p1 <- x1/n1 ##risk among exposed
  p0 <- x0/n0 ##risk among unexposed
  RR <- p1/p0
  logRR <- log(RR)
  SElogRR <- sqrt(1/x1 - 1/n1 + 1/x0 - 1/n0)
  Z <- qnorm(0.5*(1 + conf.level))
  LCL <- exp(logRR - Z*SElogRR)
  UCL <- exp(logRR + Z*SElogRR)
  ##collect
  list(x = x,
       risks = c(p1 = p1, p0 = p0),
       risk.ratio = RR,
       conf.int = c(LCL, UCL),
       conf.level = conf.level
      )
}

```

Here we test the code:

```

> ##From Jewell, p. 83
> tab7.4 <- matrix(c(178, 79, 1411, 1486), 2, 2)
> dimnames(tab7.4) <- list("Behavior type" = c("Type A", "Type B"),
+                          "CHD Event" = c("Yes", "No"))
> tab7.4
          CHD Event
Behavior type Yes  No
Type A  178 1411
Type B   79 1486
> rr.wald(tab7.4)
$x
          CHD Event
Behavior type Yes  No
Type A  178 1411
Type B   79 1486

$risks
      p1      p0
0.112020 0.050479

```

```

$risk.ratio
[1] 2.2191

$conf.int
[1] 1.7186 2.8654

$conf.level
[1] 0.95

```

*Exact approximation*

*Exact methods*

```

rr.boot <- function(x, conf.level = 0.95, replicates = 5000){
  ##prepare input
  x1 <- x[1,1]; n1 <- sum(x[1,])
  x0 <- x[2,1]; n0 <- sum(x[2,])
  ##calculate
  p1 <- x1/n1 ##risk among exposed
  p0 <- x0/n0 ##risk among unexposed
  RR <- p1/p0
  r1 <- rbinom(replicates, n1, p1)/n1
  x0.boot <- x0.boot2 <- rbinom(replicates, n0, p0)
  x0.boot[x0.boot2==0] <- x0.boot2 + 1
  n0.denom <- rep(n0, replicates)
  n0.denom[x0.boot2==0] <- n0.denom + 1
  r0 <- x0.boot/n0.denom
  rrboot <- r1/r0
  rrbar <- mean(rrboot)
  alpha <- 1 - conf.level
  ci <- quantile(rrboot, c(alpha/2, 1-alpha/2))
  ##collect
  list(x = x,
       risks = c(p1 = p1, p0 = p0),
       risk.ratio = RR,
       rrboot.mean = rrbar,
       conf.int = unname(ci),
       conf.level = conf.level,
       replicates = replicates)
}

```

Here we test the code:

```

> rr.boot(tab7.4)
$x

```

**Table 4.3.** Case-control study, binomial data

	Cases Controls	
Exposed	<i>a</i>	<i>b</i>
Unexposed	<i>c</i>	<i>d</i>

```

                CHD Event
Behavior type Yes  No
Type A 178 1411
Type B  79 1486

```

```

$risks
      p1      p0
0.112020 0.050479

```

```

$risk.ratio
[1] 2.2191

```

```

$rrboot.mean
[1] 2.2443

```

```

$conf.int
[1] 1.7235 2.9085

```

```

$conf.level
[1] 0.95

```

```

$replicates
[1] 5000

```

**Comparison**

**Power and sample size**

**4.3.3 Comparing two odds estimates: Odds ratio**

**Estimation**

**Confidence intervals**

*Normal approximation*

*Exact approximation*

*Exact methods*

Fisher method

mid-p method

**Comparison**

**Power and sample size**

**Table 4.4.** Deaths among subjects who received tolbutamide and placebo in the University Group Diabetes Program (1970), stratifying by age

	Disease	No disease	Total
Exposed	$a$	$b$	$a + b$
Nonexposed	$c$	$d$	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

#### 4.4 Problems

- Using the  $2 \times 2$  table format displayed in Table 4.4, create and test a function that takes four integers and calculate the risk ratio ( $RR$ ) and the odds ratio ( $OR$ ).

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

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

- Using the  $2 \times 2$  table format displayed in Table 4.4, create a function that takes a  $2 \times 2$  matrix and calculates the risk ratio ( $RR$ ) and odds ratio ( $OR$ ).

## Control confounding with stratification methods

In this chapter, we review two stratification methods to control for confounding:

- Pooling methods
- Standardization methods

### 5.1 Pooling methods

#### 5.1.1 Cohort studies with risk (binomial) data

Risk ratio

$$RR_{MH} = \frac{\sum_i \frac{a_i N_{0i}}{T_i}}{\sum_i \frac{b_i N_{1i}}{T_i}} \quad (5.1)$$

$$\text{Var}[\log(RR_{MH})] = \frac{\sum_i \left( \frac{M_{1i} N_{1i} N_{0i}}{T_i^2} - \frac{a_i b_i}{T_i} \right)}{\left( \sum_i \frac{a_i N_{0i}}{T_i} \right) \left( \sum_i \frac{b_i N_{1i}}{T_i} \right)} \quad (5.2)$$

```
rr.mh <- function(x, conf.level=0.95){
  ai <- x[1,1,]; bi <- x[1,2,]
  ci <- x[2,1,]; di <- x[2,2,]
```

**Table 5.1.** Stratified tables for a cohort study with risk data

Outcome	Exposure		Total
	Exposed	Nonexposed	
Cases	$a_i$	$b_i$	$M_{1i}$
Noncases	$c_i$	$d_i$	$M_{0i}$
Total	$N_{1i}$	$N_{0i}$	$T_i$

**Table 5.2.** Stratified tables for a cohort study with incidence rate data

Outcome	Exposure		Total
	Exposed	Nonexposed	
Cases	$a_i$	$b_i$	$M_i$
Person-time at risk	$PT_{1i}$	$PT_{0i}$	$T_i$

```

NOi <- bi+di; N1i <- ai+ci
MOi <- ci+di; M1i <- ai+bi
Ti <- ai+bi+ci+di
RRmh <- sum(ai*NOi/Ti)/sum(bi*N1i/Ti)
num <- sum(M1i*N1i*NOi/Ti^2-ai*bi/Ti)
dem <- sum(ai*NOi/Ti)*sum(bi*N1i/Ti)
Z <- qnorm((1+conf.level)/2)
SElogRRmh <- sqrt(num/dem)
LL <- exp(log(RRmh)-Z*SElogRRmh)
UL <- exp(log(RRmh)+Z*SElogRRmh)
list(data=x, risk.ratio=RRmh, conf.int=c(LL,UL))
}

##test function
##read UGDP data
ud <- read.table("http://www.medepi.net/data/ugdp.txt",
                 header=T, sep=",")

str(ud)
ud[1:6,]
tab <- table(ud$Status,ud$Treatment,ud$Agegrp)[,2:1,2:1]
tab

rr.mh(tab)

```

### Risk difference

$$RD_{MH} = \frac{\sum_i \frac{a_i N_{0i} - b_i N_{1i}}{T_i}}{\sum_i \frac{N_{1i} N_{0i}}{T_i}} \quad (5.3)$$

$$\text{Var}(RD_{MH}) = \frac{\sum_i \left( \frac{N_{1i} N_{0i}}{T_i} \right)^2 \left[ \frac{a_i d_i}{N_{1i}^2 (N_{1i} - 1)} + \frac{b_i c_i}{N_{0i}^2 (N_{0i} - 1)} \right]}{\left( \sum_i \frac{N_{1i} N_{0i}}{T_i} \right)^2} \quad (5.4)$$

### 5.1.2 Cohort studies with incidence rate data

#### Rate ratio

**Table 5.3.** Stratified tables for a case-control data

Outcome	Exposure		Total
	Exposed	Nonexposed	
Cases	$a_i$	$b_i$	$M_{1i}$
Controls	$c_i$	$d_i$	$M_{0i}$
Total	$N_{1i}$	$N_{0i}$	$T_i$

$$IR_{MH} = \frac{\sum_i \frac{a_i PT_{0i}}{T_i}}{\sum_i \frac{b_i PT_{1i}}{T_i}} \quad (5.5)$$

$$\text{Var}[\log(IR_{MH})] = \frac{\sum_i \left( \frac{M_i PT_{1i} PT_{0i}}{T_i} \right)^2}{\left( \sum_i \frac{a_i PT_{0i}}{T_i} \right) \left( \sum_i \frac{b_i PT_{1i}}{T_i} \right)} \quad (5.6)$$

**Rate difference**

$$ID_{MH} = \frac{\sum_i \frac{a_i PT_{0i} - b_i PT_{1i}}{T_i}}{\sum_i \frac{PT_{1i} PT_{0i}}{T_i}} \quad (5.7)$$

$$\text{Var}(ID_{MH}) = \frac{\sum_i \left( \frac{PT_{1i} PT_{0i}}{T_i} \right)^2 \left( \frac{a_i}{PT_{1i}^2} + \frac{b_i}{PT_{0i}^2} \right)}{\left( \sum_i \frac{PT_{1i} PT_{0i}}{T_i} \right)^2} \quad (5.8)$$

**5.1.3 Case control studies****Odds ratio ratio**

$$OR_{MH} = \frac{\sum_i \frac{a_i d_i}{T_i}}{\sum_i \frac{b_i c_i}{T_i}} \quad (5.9)$$

$$\text{Var}[\log(OR_{MH})] = \frac{\sum_i G_i P_i}{2(\sum_i G_i)^2} + \frac{\sum_i (G_i Q_i + H_i P_i)}{2(\sum_i G_i \sum_i H_i)} + \frac{\sum_i H_i Q_i}{2(\sum_i H_i)^2} \quad (5.10)$$

where

$$G_i = \frac{a_i d_i}{T_i}, \quad H_i = \frac{b_i c_i}{T_i}, \quad P_i = \frac{(a_i + d_i)}{T_i}, \quad Q_i = \frac{(b_i + c_i)}{T_i}$$

**5.2 Standardization methods****5.2.1 Direct standardization****5.2.2 Indirect standardization**