4.3 Confidence Intervals for Relative Risk, Odds Ratio, and Incidence Rate Ratio

Definition. Suppose in population 1 of size N_1, there are X_1 cases of interest, whereas in population 2 of size N_2 there are X_2 such cases. The population prevalence proportions are $p_1 = X_1/N_1$ and $p_2 = X_2/N_2$. The **relative risk** of the cases of interest in the two populations is the ratio $RR = \dfrac{p_1}{p_2} = \dfrac{X_1/N_1}{X_2/N_2}$. Suppose x_1 cases were observed among n_1 individuals in sample 1 drawn from population 1, and x_2 cases were observed among n_2 individuals in an independent sample 2 drawn from population 2. The **estimated relative risk** is the ratio of the two estimated prevalence proportions $\widehat{RR} = \dfrac{\hat{p}_1}{\hat{p}_2} = \dfrac{x_1/n_1}{x_2/n_2}$.

It is customary to construct first a confidence interval for the natural logarithm of the relative risk $\ln(RR)$, say, $[lcl, ucl]$, and then exponentiate both end-points of this interval to obtain a confidence interval for the relative risk RR, $[e^{lcl}, e^{ucl}]$.

The formula for a $100(1-\alpha)\%$ CI for $\ln(RR)$ is based on a normal distribution: $\ln(\widehat{RR}) \pm z_{1-\alpha/2} \cdot \widehat{SE}(\ln(\widehat{RR}))$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-percentile of a standard normal distribution, and $\widehat{SE}(\ln(\widehat{RR})) = \sqrt{\dfrac{1}{x_1} - \dfrac{1}{n_1} + \dfrac{1}{x_2} - \dfrac{1}{n_2}}$.

From here, a $100(1-\alpha)\%$ CI for RR can be computed as

$$\Big[\exp\big\{\ln(\widehat{RR}) - z_{1-\alpha/2} \cdot \widehat{SE}(\ln(\widehat{RR}))\big\}, \exp\big\{\ln(\widehat{RR}) + z_{1-\alpha/2} \cdot \widehat{SE}(\ln \widehat{RR})\big\}\Big]$$

$$= \Big[\frac{\widehat{RR}}{\exp\big\{z_{1-\alpha/2} \cdot \widehat{SE}(\ln(\widehat{RR}))\big\}}, \widehat{RR} \cdot \exp\big\{z_{1-\alpha/2} \cdot \widehat{SE}(\ln(\widehat{RR}))\big\}\Big].$$

Example. Two hundred subjects were recruited for a study on ischemic /uh-**skee**-muhk/ heart disease (narrow heart arteries) treatment. The subjects were categorized into two groups: 103 individuals with high cholesterol level, and 97 with a normal level. In the first group, 39 subjects responded to the treatment, whereas in the second group 53 responded.

We are given $x_1 = 39$, $n_1 = 103$, $x_2 = 53$, and $n_2 = 97$. First, we estimate the prevalence proportions of responding to the treatment. In the first group, $\hat{p}_1 = x_1/n_1 = 39/103 = 0.378641$. In the second group, $\hat{p}_2 = x_2/n_2 =$

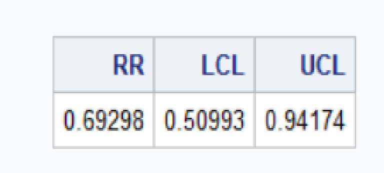
$53/97 = 0.546392$. The estimate of the relative risk in this case is $\widehat{RR} = \dfrac{\hat{p}_1}{\hat{p}_2} = \dfrac{39/103}{53/97} = 0.692984$. Next, we compute the estimated standard error of log-$\widehat{RR}$, $\widehat{SE}(\ln(\widehat{RR})) = \sqrt{\dfrac{1}{x_1} - \dfrac{1}{n_1} + \dfrac{1}{x_2} - \dfrac{1}{n_2}} = \sqrt{\dfrac{1}{39} - \dfrac{1}{103} + \dfrac{1}{53} - \dfrac{1}{97}} = 0.156496$. Finally, we calculate a 95% CI for the population relative risk.

$$
\begin{aligned}
&= \Big[\frac{\widehat{RR}}{\exp\big\{z_{1-\alpha/2} \cdot \widehat{SE}(\ln(\widehat{RR}))\big\}}, \widehat{RR} \cdot \exp\big\{z_{1-\alpha/2} \cdot \widehat{SE}(\ln(\widehat{RR}))\big\}\Big] \\
&= \Big[\frac{0.692984}{\exp\big\{(1.96)(0.156496)\big\}}, (0.692984) \cdot \exp\big\{(1.96)(0.156496)\big\}\Big]
\end{aligned}
$$

$= [0.509931, 0.941749]$. Note that since the confidence interval doesn't cover 1, we can conclude that the population prevalence proportions are different. □.

In SAS:

```
data cases;
input x1 n1 x2 n2;
cards;
39 103 53 97
;

%let conf_level=95; *choices 90, 95, 99, etc.;
data RR_CI;
 set cases;
  RR=(x1/n1)/(x2/n2);
   z=-probit((1-0.01*&conf_level)/2);
    SE=sqrt(1/x1-1/n1+1/x2-1/n2);
     LCL=RR/exp(z*SE);
      UCL=RR*exp(z*SE);
  keep RR LCL UCL;
 run;

 proc print data=RR_CI noobs;
 run;
```



RR	LCL	UCL
0.69298	0.50993	0.94174

In R:

```
x1<- 39
n1<- 103
x2<- 53
n2<- 97

conf.level<- 95
z<- -qnorm((1-0.01*conf.level)/2)
SE<- sqrt(1/x1-1/n1+1/x2-1/n2)
print(RR<- (x1/n1)/(x2/n2))
```

```
0.6929841
```

```
print(LCL<- RR/exp(z*SE))
```

```
0.5099338
```

```
print(UCL<- RR*exp(z*SE))
```

```
0.9417437
```

Definition. The **odds** in favor of an event A are defined as the ratio of the probabilities of A and the complement of A (denoted by $\bar{A}$). We write

$$\text{odds of } A = \frac{\mathbb{P}(A)}{\mathbb{P}(\bar{A})} = \frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)}.$$

For example, if the odds are 5 to 4, then $\frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)} = \frac{5}{4}$. Solving, we get $\mathbb{P}(A) = 5/9$, that is, the entire probability space is divided into 5+4=9 parts, of which A takes 5, and $\bar{A}$ takes 4.

Definition. Suppose X individuals in a population of size N constitute some cases of interest. The **odds in favor of the cases** are $X/(N - X)$. If we consider two independent populations with respective parameters X_1 and N_1, and X_2 and N_2, then the **odds ratio** is defined as $OR = \dfrac{X_1/(N_1 - X_1)}{X_2/(N_2 - X_2)}$. The **empirical estimator of the odds ratio** is $\widehat{OR} = \dfrac{x_1/(n_1 - x_1)}{x_2/(n_2 - x_2)}$, based on two independent samples with parameters x_1, x_2, n_1, and n_2.

The standard error of log-$\widehat{OR}$ has the expression

$$\widehat{SE}(\ln(\widehat{OR})) = \sqrt{\frac{1}{x_1} + \frac{1}{n_1 - x_1} + \frac{1}{x_2} + \frac{1}{n_2 - x_2}},$$

and a $100(1-\alpha)\%$ **confidence interval for the population odds ratio** can be found according to the formula

$$\Big[\frac{\widehat{OR}}{\exp\big\{z_{1-\alpha/2} \cdot \widehat{SE}(\ln(\widehat{OR}))\big\}},\ \widehat{OR} \cdot \exp\big\{z_{1-\alpha/2} \cdot \widehat{SE}(\ln(\widehat{OR}))\big\}\Big].$$

Example. In our example above, $x_1 = 39$, $n_1 = 103$, $x_2 = 53$, and $n_2 = 97$. The estimated odds ratio is $\widehat{OR} = \dfrac{x_1/(n_1 - x_1)}{x_2/(n_2 - x_2)} = \dfrac{39/(103-39)}{53/(97-53)} = 0.505896$, and the estimated standard error of log-$\widehat{OR}$ is $\widehat{SE}(\ln(\widehat{OR})) = \sqrt{\dfrac{1}{39} + \dfrac{1}{103-39} + \dfrac{1}{53} + \dfrac{1}{97-53}} = 0.287856$. Thus, a, say, 99% CI for OR is

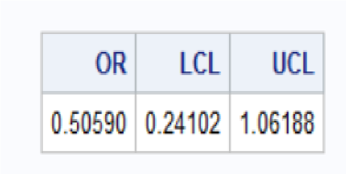
$$\Big[\frac{0.505896}{\exp\big\{(2.576)(0.287856)\big\}},\ (0.505896)\ \exp\big\{(2.576)(0.287856)\big\}\Big]$$

$=$ [0.241004, 1.061936]. Since the interval covers 1, it can be concluded that the odds are the same in both samples.

In SAS:

```
%let conf_level=99; *choices 90, 95, 99, etc.;

data OR_CI;
set cases;
 OR=(x1/(n1-x1))/(x2/(n2-x2));
  z=-probit((1-0.01*&conf_level)/2);
   SE=sqrt(1/x1+1/(n1-x1)+1/x2+1/(n2-x2));
    LCL=OR/exp(z*SE);
     UCL=OR*exp(z*SE);
  keep OR LCL UCL;
```

```
run;

proc print data=OR_CI noobs;
run;
```



OR	LCL	UCL
0.50590	0.24102	1.06188

In R:

```
conf.level<- 99
z<- -qnorm((1-0.01*conf.level)/2)
SE<- sqrt(1/x1+1/(n1-x1)+1/x2+1/(n2-x2))
print(OR<- (x1/(n1-x1))/(x2/(n2-x2)))
```

```
0.5058962
```

```
print(LCL<- OR/exp(z*SE))
```

```
0.2410159
```

```
print(UCL<- OR*exp(z*SE))
```

```
1.061884
```

□

Definition. Suppose in a sample from population 1, n_1 cases were observed during T_1 person-years. In a sample from an independent population 2, n_2 cases were observed during T_2 person-years. Denote by λ_1 and $\lambda 2$ the respective population incidence rates. The **incidence rate ratio** is defined as $IRR = \lambda_1/\lambda_2$. We estimate this quantity by $\widehat{IRR} = (n_1/T_1)/(n_2/T_2)$.

A $100(1-\alpha)\%$ CI for IRR has the form

$$\Big[\frac{\widehat{IRR}}{\exp\big\{z_{1-\alpha/2}\cdot\widehat{SE}(\ln(\widehat{IRR}))\big\}},\ \widehat{IRR}\cdot\exp\big\{z_{1-\alpha/2}\cdot\widehat{SE}(\ln(\widehat{IRR}))\big\}\Big]$$

where $\widehat{SE}(\ln(\widehat{IRR}))=\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}$.

Example. In a controlled experiment, patients with a cirrhosis (a chronic liver disease) were randomized to receive peniccilamine /peh-nuh-**si**-luh-meen/ (treatment group) or a placebo (control group). In the treatment group, there were 52 deaths during the course of 873.5 patient-years, whereas in the control group, there were 69 death during 834.5 patient-years.

The estimated incidence rate in the treatment group is $\hat{\lambda}_1 = 52/873.5 = 0.059531$, or roughly, 6 per 100 patient-years. In the control group, the incidence rate is estimated as $\hat{\lambda}_2 = 69/834.5 = 0.082684$, or about 8.3 per 100 patient-years. The incidence rate ratio has an estimate $\widehat{IRR} = 0.059531/0.082684 = 0.719975$. A 99% CI for IRR is

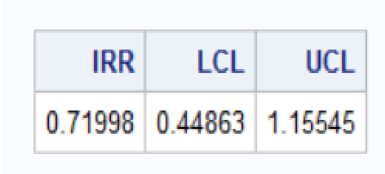
$$\Big[\frac{0.719975}{\exp\big\{(2.576)\sqrt{\frac{1}{52}+\frac{1}{69}}\big\}},\ (0.719975)\cdot\exp\big\{(2.576)\sqrt{\frac{1}{52}+\frac{1}{69}}\big\}\Big]$$

$=[0.448613,\ 1.155482]$. Since the interval includes the unity, we conclude that the true population incidence rates are equal, and the treatment is not effective.

In SAS:

```
data cases;
input n1 T1 n2 T2;
cards;
52 873.5 69 834.5
;

%let conf_level=99; *choices 90, 95, 99, etc.;

data IRR_CI;
set cases;
 IRR=(n1/T1)/(n2/T2);
  z=-probit((1-0.01*&conf_level)/2);
   SE=sqrt(1/n1+1/n2);
```

```
    LCL=IRR/exp(z*SE);
     UCL=IRR*exp(z*SE);
 keep IRR LCL UCL;
run;

proc print data=IRR_CI noobs;
run;
```



IRR	LCL	UCL
0.71998	0.44863	1.15545

In R:

```
n1<- 52
T1<- 873.5
n2<- 69
T2<- 834.5
conf.level<- 99
z<- -qnorm((1-0.01*conf.level)/2)
SE<- sqrt(1/n1+1/n2)
print(IRR<- (n1/T1)/(n2/T2))
```

```
0.7199754
```

```
print(LCL<- IRR/exp(z*SE))
```

```
0.4486274
```

```
print(UCL<- IRR*exp(z*SE))
```

```
1.155446
```

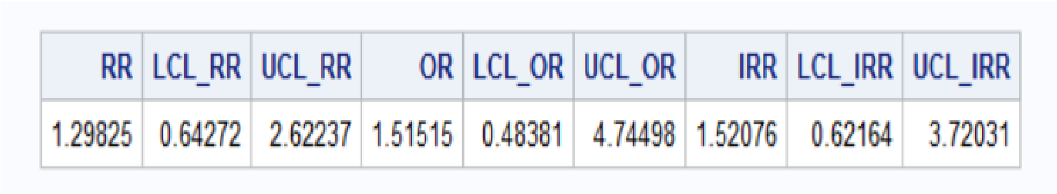
□

Example. On April 2, 2013, there were 56 workers who were present in the building at the time of a "lab leak" when a significant amount of Mycobacterium tuberculosis virus escaped. Everyone was placed into quarantine and symptoms of tuberculosis were monitored. The file "TB_Symptoms_Data.csv" contains gender, if symptoms developed, and the date symptoms developed, or June 1, 2013, the date when the monitoring ended.

Using the raw data, we estimate relative risk, odds ratio, and incidence rate ratio for women and men, and compute 95% CI for the three population parameters.

In SAS:

```
proc import out=TBdata datafile="./TB_Symptoms_Data.csv"
dbms=csv replace;

data TBstats (drop=SubjectN Gender TB_Symptoms Date_Symptoms);
 set TBdata;
  if gender='F' and TB_Symptoms=1 then Fcases+1;
   if gender='M' and TB_Symptoms=1 then Mcases+1;
     if gender='F' then Ftotal+1;
     if gender='M' then Mtotal+1;
      if gender='F' then Fduration+Date_Symptoms-'2Apr2013'd;
       if gender='M' then Mduration+Date_Symptoms-'2Apr2013'd;
  if _N_=56;
 run;

%let conf_level=95; *choices 90, 95, 99, etc.;

data TBstats;
 set TBstats;
   RR=(Fcases/Ftotal)/(Mcases/Mtotal);
    OR=(Fcases/(Ftotal-Fcases))/(Mcases/(Mtotal-Mcases));
  IRR=(Fcases/Fduration)/(Mcases/Mduration);
        z=-probit((1-0.01*&conf_level)/2);
   SE_lnRR=sqrt(1/Fcases-1/Ftotal+1/Mcases-1/Mtotal);
 SE_lnOR=sqrt(1/Fcases+1/(Ftotal-Fcases)+1/Mcases+1/(Mtotal-Mcases));
 SE_lnIRR=sqrt(1/Fcases+1/Mcases);

    LCL_RR=RR/exp(z*SE_lnRR);
     UCL_RR=RR*exp(z*SE_lnRR);
       LCL_OR=OR/exp(z*SE_lnOR);
         UCL_OR=OR*exp(z*SE_lnOR);
       LCL_IRR=IRR/exp(z*SE_lnIRR);
             UCL_IRR=IRR*exp(z*SE_lnIRR);
 keep RR OR IRR LCL_RR UCL_RR LCL_OR UCL_OR LCL_IRR UCL_IRR;
run;

 proc print data=TBstats noobs;
 var RR LCL_RR UCL_RR OR LCL_OR UCL_OR IRR LCL_IRR UCL_IRR;
 run;
```



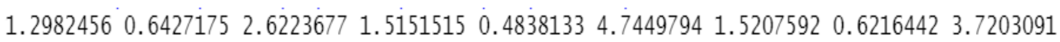
RR	LCL_RR	UCL_RR	OR	LCL_OR	UCL_OR	IRR	LCL_IRR	UCL_IRR
1.29825	0.64272	2.62237	1.51515	0.48381	4.74498	1.52076	0.62164	3.72031

Note that even though all the point estimators are above 1 (meaning women were hit worse than men), the confidence intervals all cover 1, and so the differential effect of gender was statistically insignificant.

In R:

```
TBdata<- read.csv(file="./TB_Symptoms_Data.csv", header=TRUE, sep=",")

female.workers <- subset(TBdata, Gender=="F")
male.workers<- subset(TBdata, Gender=="M")

Fcases<- sum(female.workers$TB_Symptoms)
Mcases<- sum(male.workers$TB_Symptoms)
Ftotal<- nrow(female.workers)
Mtotal<- nrow(male.workers)

Fduration<- as.numeric(sum(as.Date(female.workers$Date_Symptoms) -as.Date("2013-
04-02")))
Mduration<- as.numeric(sum(as.Date(male.workers$Date_Symptoms) -as.Date("2013-
04-02")))

conf.level<- 95

RR<- (Fcases/Ftotal)/(Mcases/Mtotal)
    OR<- (Fcases/(Ftotal-Fcases))/(Mcases/(Mtotal-Mcases))
       IRR<- (Fcases/Fduration)/(Mcases/Mduration)
```

```
z<- -qnorm((1-0.01*conf.level)/2)

SE.lnRR<- sqrt(1/Fcases-1/Ftotal+1/Mcases-1/Mtotal)
     SE.lnOR<- sqrt(1/Fcases+1/(Ftotal-Fcases)+1/Mcases+1/(Mtotal-Mcases))
        SE.lnIRR<- sqrt(1/Fcases+1/Mcases)

LCL.RR<- RR/exp(z*SE.lnRR)
     UCL.RR<- RR*exp(z*SE.lnRR)
        LCL.OR<- OR/exp(z*SE.lnOR)
          UCL.OR<- OR*exp(z*SE.lnOR)
            LCL.IRR<- IRR/exp(z*SE.lnIRR)
              UCL.IRR<- IRR*exp(z*SE.lnIRR)

print(c(RR,LCL.RR,UCL.RR,OR,LCL.OR,UCL.OR,IRR,LCL.IRR,UCL.IRR))
```



```
1.2982456 0.6427175 2.6223677 1.5151515 0.4838133 4.7449794 1.5207592 0.6216442 3.7203091
```

□