Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

Demo for Chapter 15, part 4: Inference for Odds Ratios

Jinko Graham

Testing whether $OR = 1$

- The chi-square test assesses the null hypothesis that $OR = 1$ (no association between exposure and disease) against the alternative hypothesis that $OR \neq 1$ (an association).

```
mydf <- data.frame(case=c(1350,7),control=c(1296,61)) # Doll and Hill's data
rownames(mydf) <- c("smoker","non-smoker")
mydf
```

```
##            case control
## smoker     1350    1296
## non-smoker    7      61
```

```
chisq.test(mydf)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  mydf
## X-squared = 42.37, df = 1, p-value = 7.552e-11
```

Using R to get point and interval estimates of the OR.

- For a 2×2 table, the calculations are easy enough to do with a hand calculator or spreadsheet.
- Below we present some R code that illustrates data frame manipulation in R.
  - This will give us code that will generalize to the case of multiple exposures, as in the Doll and Hill dataset where smoking status has four levels.
  - If you are not interested in learning R, don't struggle to understand each line of code.
  - Instead focus on understanding the *purpose* of each line of code.
    - There will be similar code in assignment 2 and you will need to understand the steps to make simple modifications as necessary.

Add Odds of Case to Data Frame

```
mydf <- data.frame(group=c("smoker","non-smoker"),
                   case=c(1350,7),
                   control=c(1296,61)) # Doll and Hill's data
library(dplyr) # for the mutate() function
mydf <- mutate(mydf,total = case+control,prcase  = case/total,
               odds = prcase/(1-prcase))
mydf
```

```
##         group case control total    prcase      odds
## 1      smoker 1350    1296  2646 0.5102041 1.0416667
## 2 non-smoker    7      61    68 0.1029412 0.1147541
```

- `mutate()` is used to create new variables from existing ones and add them to our data frame.
- In this example, the variables `total`, `prcase` and `odds` are created and added to `mydf`.
  - Notice that the calculation of `odds` can use the newly-created variable `prcase`.

Extract the Baseline Odds

- As a baseline for comparison, we will use the group of non-smokers. Let's extract their estimated odds of lung cancer.

```
odds0 <- mydf[mydf$group=="non-smoker","odds"]
odds0
```

```
## [1] 0.1147541
```

- We can grab elements of a dataframe by referencing the desired rows and columns inside square brackets.
- E.G, `mydf[1,2]` will grab the element in the 1st row and 2nd column of `mydf`.
- Above, we indicate the desired row with the logical condition `mydf$group=="non-smoker"`,
  - `mydf$group` refers to the variable `group` in the data frame `mydf`.
  - Grab only the row for the non-smoking group.
- The desired column has the variable name `odds`.

Extracting Other Baseline Data

- Extract the number of non-smoker cases and controls analogously.

```
c <- mydf[mydf$group=="non-smoker","case"]
c
```

```
## [1] 7
```

```
d <- mydf[mydf$group=="non-smoker","control"]
d
```

```
## [1] 61
```

Add the SE for the log-OR

```
mydf <- mutate(mydf, OR=odds/odds0,
               se.logOR = sqrt(1/case + 1/control + 1/c + 1/d))
mydf
```

```
##         group case control total    prcase      odds       OR  se.logOR
## 1      smoker 1350    1296  2646 0.5102041 1.0416667 9.077381 0.4009525
## 2 non-smoker    7      61    68 0.1029412 0.1147541 1.000000 0.5643591
```

Add the CI

```
critval <- qnorm( 0.025, lower.tail=FALSE)
mydf <- mutate(mydf,
               lowerCI=round(exp(log(OR) - critval*se.logOR),3),
               upperCI=round(exp(log(OR) + critval*se.logOR),3))
mydf <- mutate(mydf, prcase=round(prcase,3),
               odds=round(odds,3), OR=round(OR,3),
               se.logOR=round(se.logOR,3))
mydf
```

```
##        group case control total prcase  odds    OR se.logOR lowerCI
## 1     smoker 1350    1296  2646  0.510 1.042 9.077    0.401   4.137
## 2 non-smoker    7      61    68  0.103 0.115 1.000    0.564   0.331
##   upperCI
## 1  19.918
## 2   3.023
```

- The `se.logOR` and CI for the non-smokers are not defined because we are using the non-smokers as the baseline group in our calculations.
  - Technically, the `se.logOR` is 0 and the CI is exactly 1, by definition.
- As a result, we set these to be the missing data code `NA` in the baseline group:

```
mydf[mydf$group=="non-smoker",
     c("se.logOR", "lowerCI", "upperCI")] <-
  c(NA, NA, NA)
mydf
```

```
##         group case control total prcase  odds    OR se.logOR lowerCI
## 1      smoker 1350    1296  2646  0.510 1.042 9.077    0.401   4.137
## 2 non-smoker    7      61    68  0.103 0.115 1.000       NA      NA
##   upperCI
## 1  19.918
## 2      NA
```

- Notice how we referenced multiple columns at once with `c("se.logOR","lowerCI","upperCI")`.

More Than Two Exposure Levels

- Doll and Hill's data with smokers classified by the average number of cigarettes per day:

		case	control
Number of	25+	340	182
cigarettes	15-24	445	408
per day	1-14	565	706
	0	7	61

- Can use the last row with 0 cigs per day (unexposed) as a baseline group, and calculate our ORs for each level of exposure.
- Here is where the R code we wrote can pay off. We essentially repeat the code, but now refer to the baseline group as `"0"` instead of `"non-smoker"`.

```r
mydf <- data.frame(group=c("25+", "15-24", "1-14", "0"),
                   case=c(340,445,565,7),
                   control=c(182,408,706,61))

library(dplyr)
mydf <- mutate(mydf,total = case+control,
               prcase  = case/total,odds = prcase/(1-prcase))
odds0 <- mydf[mydf$group=="0","odds"]
c <- mydf[mydf$group=="0","case"]
d <- mydf[mydf$group=="0","control"]
mydf <- mutate(mydf, OR=odds/odds0,
               se.logOR = sqrt(1/case + 1/control + 1/c + 1/d))
critval <- qnorm( 0.025, lower.tail=FALSE)
mydf <- mutate(mydf,
             lowerCI = round(exp(log(OR) - critval*se.logOR),3),
             upperCI = round(exp(log(OR) + critval*se.logOR),3))
mydf <- mutate(mydf, prcase=round(prcase,3), odds=round(odds,3),
               OR=round(OR,3), se.logOR=round(se.logOR,3))
mydf[mydf$group=="0",c("se.logOR", "lowerCI", "upperCI")] <-
  c(NA, NA, NA)
```

```
mydf
```

```
##   group case control total prcase  odds     OR se.logOR lowerCI upperCI
## 1   25+  340     182   522  0.651 1.868 16.279    0.409   7.296  36.325
## 2 15-24  445     408   853  0.522 1.091  9.505    0.405   4.298  21.018
## 3  1-14  565     706  1271  0.445 0.800  6.974    0.403   3.165  15.365
## 4     0    7      61    68  0.103 0.115  1.000       NA      NA      NA
```