Stats 111 HW1

Viraj Vijaywargiya

2023-01-19

```
library(epitools)
library(rmeta)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(nnet)
```

```
prop.comp <- function( x, estimate="all", conf.level=.95, transpose=FALSE ){
    if( transpose ) x <- t(x)
    rslt <- vector( "list", length=3 )
    names( rslt ) <- c( "riskdiff", "riskratio", "oddsratio" )
    diff.rslt <- suppressWarnings(prop.test( x, conf.level=conf.level ))
    rslt[[1]] <- rslt[[2]] <- rslt[[3]] <- epitab( x, method="riskratio", pvalue="chi2", conf.level=con
    colnames( rslt[[1]] )[5] <- "riskdiff"
    rslt[[1]][,5] <- c(0,diff(rev(diff.rslt$estimate)))
    rslt[[1]][2,6:7] <- diff.rslt$conf.int
    colnames( rslt[[3]] )[5] <- "oddsratio"
    rslt[[3]][,5:8] <- suppressWarnings(epitab( x, method="oddsratio", pvalue="chi2", conf.level=conf.l
    if(is.null(names(dimnames(x)))){
        for(i in 1:3){
            colnames(rslt[[i]])[c(1,3)] <- c("Outcome=0", "Outcome=1")
            rownames(rslt[[i]]) <- c("Group=1", "Group=2")
            }
    }
    if( is.element( estimate, c("all", "oddsratio") ) ){
        if(is.null(names(dimnames(x)))){
            warning( "Estimated probabilities represent Pr[ Outcome | Group ]. For estimates of
            Pr[ Group | Outcome ], change the value of 'transpose'.")
        }
        else
            warning( paste("Estimated probabilities represent Pr[", names(dimnames(x))[2],
            "|",names(dimnames(x))[1], "]. For estimates of
```

```
            Pr[", names(dimnames(x))[1], "|",names(dimnames(x))[2], "], change the value of 'transpose'
        }
    if( estimate == "riskdiff" ) return(rslt[[1]])
    else if( estimate == "riskratio" ) return(rslt[[2]])
    else if( estimate == "oddsratio" ) return(rslt[[3]])
    else return(rslt)
}
```

1) **1a)** Binomial. Each trial has a binary outcome of success or failure (yes or no).

   1b) Poisson. It is used for when the response is the number of times something occurs.

   1c) Normal.

   1d) Multinomial.

   1e) Multinomial.

   1f) Poisson.

2) **2a**) Null Hypothesis HO: p = 0.5. Alternate Hypothesis HA: p > 0.5.

   2b) pˆ = n1/n. Therefore, pˆ = 300/500 = 0.6.

   2c)

```
prop.test(300, 500, 0.5, alternative="greater")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  300 out of 500, null probability 0.5
## X-squared = 19.602, df = 1, p-value = 4.768e-06
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.5625085 1.0000000
## sample estimates:
##   p
## 0.6
```

p-value = 4.768e-06. Therefore, we reject the null hypothesis (p = 0.5) and conclude evidence for the alternate hypothesis (p > 0.5). Thus, they have statistical evidence that they deliver early a majority of the time.

2d) p1 = early, p2 = ontime, p3 = late. Null hypothesis H0: p1 = 0.75, p2 = 0.24, p3 = 0.01. Alternate hypothesis HA: at least one of the equalities does not hold.

2e)

```
chisq.test(x=c(300,192,8), p=c(0.75,0.24,0.01))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(300, 192, 8)
## X-squared = 60, df = 2, p-value = 9.358e-14
```

p-value = 9.358e-14. Since p-value < 0.05, we reject the null hypothesis (p1 = 0.75, p2 = 0.24, p3 = 0.01) and conclude evidence for the alternate hypothesis that atleast one of the equalities does not hold. Therefore, the company cannot state that they deliver early 75% of the time, 24% of the time they deliver on time and deliver late only 1% of the time.

3) **3a**) It is a randomized experimental study.

3b)

```
fiber = matrix(c(140,200,60,50),2,2)
rownames(fiber) = c("Low fiber","High fiber")
colnames(fiber) = c("Disease no","Disease yes")
prop.comp(fiber)
```

```
## Warning in prop.comp(fiber): Estimated probabilities represent Pr[ Outcome | Group ]. For estimates
##          Pr[ Group | Outcome ], change the value of 'transpose'.

## $riskdiff
##         Outcome=0  p0 Outcome=1  p1 riskdiff      lower       upper   p.value
## Group=1       140 0.7        60 0.3      0.0         NA          NA        NA
## Group=2       200 0.8        50 0.2     -0.1 -0.1850734 -0.01492665 0.0141764
##
## $riskratio
##         Outcome=0  p0 Outcome=1  p1 riskratio     lower     upper   p.value
## Group=1       140 0.7        60 0.3 1.0000000        NA        NA        NA
## Group=2       200 0.8        50 0.2 0.6666667 0.4812002 0.9236164 0.0141764
##
## $oddsratio
##         Outcome=0  p0 Outcome=1  p1 oddsratio     lower    upper   p.value
## Group=1       140 0.7        60 0.3 1.0000000        NA       NA        NA
## Group=2       200 0.8        50 0.2 0.5833333 0.3783222 0.899439 0.0141764
```

The estimated probability of Disease for a seniors with Low fiber diet is 0.3 and for seniors with High fiber diet is 0.2. As a result, the risk difference is -0.1. Going from Low Fiber to High, results in an estimated difference in risk of -0.1, risk will be 0.1 lower. Lower and upper designate the 95% CI, (-0.1850734, -0.01492665).

3c) The odds ratio (OR) estimate is 0.583. This to say that the odds of having Colonic disease for seniors with high fiber diet is 0.583 of that of seniors with low fiber diet (so seniors with low fiber diet have almost twice the odds).

3d) Null hypothesis H0: OR = 1. Alternate hypothesis HA: OR != 1. p-value = 0.0141764. Since p-value < 0.05, we reject the null hypothesis (OR = 1) and conclude the alternate hypothesis (OR != 1). Therefore, it is expected that the odds for the high fiber group is not the same as the odds for the low fiber group.

4) **4a**) Null hypothesis H0: p(ij) = p(i)p(j) for all i=1,2,3 and j=1,2. That is X and Y are independent. Alternate hypothesis HA: at least one combination of i and j has p(ij) != p(i)p(j). That is X and Y are not independent.

4b)

```
smoke.school = matrix(c(1168,1823,1380,188,416,400),3,2)
rownames(smoke.school) = c("0 parents smoke","1 parent smokes","2 parents smoke")
colnames(smoke.school) = c("Smoke no","Smoke yes")
chisq.test(smoke.school)$expected
```

```
##                  Smoke no Smoke yes
## 0 parents smoke 1102.712  253.2882
## 1 parent smokes 1820.776  418.2244
## 2 parents smoke 1447.513  332.4874
```

The expected number of children smokers whose parents do not smoke is 253.2882.

4c)

```
chisq.test(smoke.school)
```

```
##
##  Pearson's Chi-squared test
##
## data:  smoke.school
## X-squared = 37.566, df = 2, p-value = 6.959e-09
```

p-value = 6.959e-09. Since the p-value < 0.05, we reject the null hypothesis (p(ij) = p(i)p(j) for all i=1,2,3 and j=1,2) and conclude the alternate hypothesis (at least one combination of i and j has p(ij) != p(i)p(j)). Therefore, X and Y are not independent, that is the smoking status of the parents is expected to have an effect on the smoking status of the child.

4d) We cannot conclude that the smoking status of the parent causes the smoking status of the child because the study is observational, and there can be other cofounding variables.

4e) The explanatory variable, smoking status of the parents, can be viewed as an ordinal variable because there is a reason to believe that the outcomes can be ordered with respect to the nature of the study or data. That is, the order of the parents smoking status is 0, 1, 2 (0 = No parent smokes, 1 = 1 parent smokes, 2 = both parents smoke).

5) **5a**) Type of sampling used: Binomial. Type of study: Observational study. Prospective.

5b) Type of sampling used: Multinomial. Type of study: Randomized Experiment. Prospective.

5c)

i) Estimated probability of choosing the Low-Fat diet and dropping out of the study = 10/312 = 0.032.
ii) Estimated probability of dropping out of the study if or given the individual chose the Low-Fat diet = 10/104 = 0.096.
iii) Estimated probability of dropping out of the study = 50/312 = 0.16. 5d) Yes. Sampling method in part b is multinomial and it is a randomized experiment, which means cause and effect conclusions generally can be made. Therefore, we can estimate the probability that an individual with a desire to lose weight will choose the Low-Fat diet.

6)

```
trial = matrix(c(92,87,8,23),2,2)
rownames(trial) = c("Placebo","Ursodiol")
colnames(trial) = c("Negative","Positive")
prop.comp(trial)
```

```
## Warning in prop.comp(trial): Estimated probabilities represent Pr[ Outcome | Group ]. For estimates
##          Pr[ Group | Outcome ], change the value of 'transpose'.
```

```
## $riskdiff
##         Outcome=0        p0 Outcome=1        p1  riskdiff      lower    upper
## Group=1        92 0.9200000         8 0.0800000 0.0000000         NA       NA
## Group=2        87 0.7909091        23 0.2090909 0.1290909 0.02679585 0.231386
##             p.value
## Group=1          NA
## Group=2 0.008441788
##
## $riskratio
##         Outcome=0        p0 Outcome=1        p1 riskratio    lower   upper
## Group=1        92 0.9200000         8 0.0800000  1.000000       NA      NA
## Group=2        87 0.7909091        23 0.2090909  2.613636 1.225322 5.57494
##             p.value
## Group=1          NA
## Group=2 0.008441788
##
## $oddsratio
##         Outcome=0        p0 Outcome=1        p1 oddsratio    lower    upper
## Group=1        92 0.9200000         8 0.0800000   1.00000       NA       NA
## Group=2        87 0.7909091        23 0.2090909   3.04023 1.291383 7.157442
##             p.value
## Group=1          NA
## Group=2 0.008441788
```

6a) p1 = 0.2090909, p0 = 0.08. RD = 0.1290909, RR = 2.613636, OR = 3.04023. log(RR) = 0.9607424, log(OR) = 1.111933.

6b) Estimated variance = (p1 * (1 - p1)) / n1 + (p2 * (1 - p2)) / n2 = 0.002699. Standard error = sqrt(variance) = 0.0520. CI = point estimate +- (critical value * standard error). For a 95% CI, the critical value is typically 1.96. 95% CI for the risk difference: 0.1290909 +- (1.96 * 0.0520) = (0.026, 0.232). 95% CI for the log relative risk: 0.9279 +- (1.96 * 0.0520) = (0.824, 1.032). 95% CI for the log odds ratio: 0.9279 +- (1.96 * 0.0520) = (0.824, 1.032).

6c) Lower limit for the odds ratio = exp(0.824) = 2.28 and Upper limit for the odds ratio = exp(1.032) = 2.82. Lower limit for the relative risk = exp(0.824) = 2.28 and Upper limit for the relative risk = exp(1.032) = 2.82.

The estimated odds ratio of 2.613636 means that the odds of a positive response in the Ursodiol group are 2.6 times higher than the odds of a positive response in the placebo group. The 95% CI for the odds ratio (2.28, 2.82) indicates that, given the sample size and assuming the normal approximation holds, there is a 95% chance that the true odds ratio in the population falls between 2.28 and 2.82.

The estimated relative risk of 2.613636 means that the probability of a positive response in the Ursodiol group is 2.6 times higher than the probability of a positive response in the placebo group. The 95% CI for the relative risk (2.28, 2.82) indicates that, given the sample size and assuming the normal approximation holds, there is a 95% chance that the true relative risk in the population falls between 2.28 and 2.82.

In both cases, the lower limit of the 95% CI is above 1, which indicates that the Ursodiol group is more likely to have a positive response than the placebo group. Therefore, Ursodiol has a positive effect on the disappearance of gallstones with a 95% probability.

7) The sex of the subject (female yes or no) can be considered a cofounder when trying to study the association between high protein diet (explanatory) and having high blood pressure (response). This is because the body weight of males are usually more than females which means they have to consume a higher protein diet, potentially resulting in a higher blood pressure. Also, workout intensity for males is expected to be more than that of females, which means they might consume more protein in their diet.