Inference on Two Sample Proportions in R

BIOST 514/517

Discussion - Week 6

Proportions

Interest in examining a binary variable in a population

- Death
- Cancer relapse

```r
psa <- read.table("../psa.txt",header=TRUE)

dat <- psa[!is.na(psa$grade),]

relapse24 <- dat$inrem=="no" & dat$obstime < 24
hiGrade <- dat$grade > 2
```

Inference on Two Proportions

May want to compare proportions between two groups by their difference $p_1 - p_2$

- Treatment versus control
- Low grade versus high grade tumor

Difference in Sample Proportions

Use $\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$ to estimate $p_1 - p_2$

$$\hat{p}_1 - \hat{p}_2 \dot{\sim} N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

Testing the Difference in Proportions

Test the null hypothesis $p_1 - p_2 = 0$ against $p_1 - p_2 \neq 0$

- Under the null $p_1 = p_2 = p$, so the standard deviation is

$$\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- We estimate p with $\hat{p}$ the pooled sample proportion

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

- Then we can use a Z-test with

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Confidence Intervals for the Difference in Proportions

As in the one-sample case, we do not assume the null hypothesis for the standard deviation, and we use our best guess for p_1 and p_2

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

prop.test for Difference in Proportions

Can still use `prop.test(x, n)`, with defaults

- Testing $p_1 - p_2 = 0$ against $p_1 - p_2 \neq 0$
- Continuity correction
- 95% confidence with argument `conf.level=0.95`

prop.test for Difference in Proportions

```
diffGrade <- (xHigrade/nHigrade)-
  (xLograde/nLograde)
```

We estimate the proportion experiencing relapse within 24 months is 0.04 higher in a group with high grade tumors than in a group with low grade tumors based on our data.

prop.test for Difference in Proportions

```
diffTestCorrect <- prop.test(x=c(xLograde,xHigrade),
                             n=c(nLograde,nHigrade))
diffTestCorrect
```

```
##
##  2-sample test for equality of proportions with continui
##  correction
##
## data:  c(xLograde, xHigrade) out of c(nLograde, nHigrade
## X-squared = 5.8362e-31, df = 1, p-value = 1
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.3847789  0.3097789
## sample estimates:
## prop 1 prop 2
## 0.4000 0.4375
```

prop.test for Difference in Proportions

```
diffTestCorrect$conf.int
```

```
## [1] -0.3847789  0.3097789
## attr(,"conf.level")
## [1] 0.95
```

Our data are consistent with the proportion relapsing within 24 months in a high grade tumor group being 0.38 lower to 0.31 higher than in a low grade tumor group. Because 0 is in our confidence interval, we would not be surprised if the true proportions were similar between groups.

Odds

A different way of looking at probabilities

$$o = \frac{p}{1-p} \implies \hat{o} = \frac{\hat{p}}{1-\hat{p}}$$

```
oLograde <- (xLograde/nLograde)/
  ((nLograde-xLograde)/nLograde)
oHigrade <- (xHigrade/nHigrade)/
  ((nHigrade-xHigrade)/nHigrade)
oLograde
```

```
## [1] 0.6666667
```

```
oHigrade
```

```
## [1] 0.7777778
```

Odds Ratio

Compare the relative difference in odds instead of absolute difference in proportions

$$\text{OR} = \frac{o_1}{o_2}$$

```
orGrade <- oHigrade/oLograde
orGrade
```

```
## [1] 1.166667
```

We estimate the odds of relapse within 24 months for a group with high grade tumors are 1.17 times the odds for a group with low grade tumors based on our data.

Alterantively, we can say that the odds of relapse within 24 months for a group with high grade tumors are $(\widehat{\text{OR}} - 1) \cdot 100\% \approx 16.67\%$ higher than for a group with low grade tumors based on our data.

Odds Ratio from a 2x2 Table

	$E=0$	$E=1$
$D=0$	a	b
$D=1$	c	d

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

```r
tabGrade <- table(relapse24,hiGrade)
tabGrade
```

```
##            hiGrade
## relapse24 FALSE TRUE
##     FALSE    15    9
##     TRUE     10    7
```

```r
(15*7)/(9*10)
```

```
## [1] 1.166667
```

Inference on the Odds Ratio

$$\log \widehat{\mathrm{OR}} \dot{\sim} N\left(\log \mathrm{OR}, \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)$$

95% Confidence Interval for the Odds Ratio

$$\exp\left(\log\widehat{\mathrm{OR}} \pm \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)$$

```
orSE <- sqrt((1/15)+(1/9)+(1/10)+(1/7))
exp(log(orGrade) + c(-1,1)*qnorm(0.975)*orSE)
```

```
## [1] 0.3272565 4.1591563
```

fisher.test for Inference on the Odds Ratio in R

`fisher.test` function includes CI for the odds ratio

- Default to 95% confidence interval with argument `conf.level=0.95`
- Estimates slightly different than "by hand"

Confidence Interval for the Odds Ratio in R

```
infGrade <- fisher.test(tabGrade)
infGrade$conf.int
```

```
## [1] 0.2695186 4.9506212
## attr(,"conf.level")
## [1] 0.95
```

These data are consistent with the odds of relapse in 24 months in a high grade tumor group being 0.27 to 4.95 times those of a low grade tumor group. Because 1 is in our confidence interval, it would not be surprising if the true odds of relapse were similar between groups.