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

R Demo for Chapter 14 : Inference for Proportions

Jinko Graham

Confidence Intervals

- Recall: WHI has 16,608 women aged 50 to 79 years randomized to receive either estrogen plus progestin (EP; $n_1 = 8506$), or a placebo ($n_2 = 8102$). After five years, 166 or those in the EP group had developed invasive breast cancer, compared to 122 in the placebo group.

```
numCancer1 <- 166; n1 <- 8506 #EP group
p1hat <- numCancer1/n1
numCancer2 <- 122; n2 <- 8102 #Placebo group
p2hat <- numCancer2/n2

phatDiff <- p1hat - p2hat
phatDiff
```

```
## [1] 0.004457626
```

- The difference in proportions is about 0.0044

- 95% CI is estimate $\pm$ margin of error, where
  - estimate of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2 = 0.0044$
  - margin of error is a critical value times standard error of difference.

```
zstar <- qnorm((1-.95)/2,lower.tail=FALSE) #get critical value
se <- sqrt(p1hat*(1-p1hat)/n1 + p2hat*(1-p2hat)/n2) #get se
me <- zstar*se #get margin of error

CI <- c(phatDiff - me,phatDiff + me)
CI
```

```
## [1] 0.000498629 0.008416622
```

The 95% CI is about $(0.0005, 0.008)$

R Lookup Table

CI for Difference of Proportions

- $n_1 = 8506$ women receive EP and $n_2 = 8102$ receive placebo. 166 in EP group and 122 in placebo group get breast cancer. Table to help understand R code:

R variable	Notation	Value
`numCancer1`	–	166
`n1`	n_1	8506
`p1hat`	$\hat{p}_1$	166/8506
`numCancer2`	–	122
`n2`	n_2	8102
`p2hat`	$\hat{p}_2$	122/8102
`phatDiff`	$\hat{p}_1 - \hat{p}_2$	0.0044
`zstar`	z^*	1.96
`se`	$\sqrt{\hat{p}_1 * (1 - \hat{p}_1)/n_1 + \hat{p}_2 * (1 - \hat{p}_2)/n_2}$	0.00202
`CI`	$(\hat{p}_1 - \hat{p}_2) \pm z^* * \sqrt{\hat{p}_1 * (1 - \hat{p}_1)/n_1 + \hat{p}_2 * (1 - \hat{p}_2)/n_2}$	(0.0005, 0.008)

Test statistic

- Under $H_0 : p_1 - p_2 = 0$ or $p_1 = p_2 = p$. Both the EP and placebo populations have the same proportion p of cancer cases.
- Calculating the test statistic requires the pooled-sample estimate of p.

```
phat <- (numCancer1+numCancer2)/(n1+n2) #pooled-sample estimate of p
se <- sqrt(phat*(1-phat)*(1/n1+1/n2)) #based on phat
z <- phatDiff/se
z
```

```
## [1] 2.199707
```

- The test statistic is about 2.2

Hypothesis test

- For the hypothesis test of $H_0 : p_1 - p_2 = 0$ vs. $H_a : p_1 - p_2 \neq 0$, we have:

```
pval<-2*pnorm(abs(z),lower.tail=FALSE)
pval
```

```
## [1] 0.02782772
```

* The pvalue is about 0.03.

- We therefore reject H_0 at the 5% level.
  - There is statistical evidence that women taking EP have a higher risk of invasive breast cancer than those taking the placebo.

R Lookup Table

Testing $H_0 : p_1 - p_2 = 0$ vs. $H_a : p_1 - p_2 \neq 0$.

- $n_1 = 8506$ women receive EP and $n_2 = 8102$ receive placebo. 166 in EP group and 122 in placebo group get breast cancer.

R variable	Notation	Value
`numCancer1`	–	166
`n1`	n_1	8506
`p1hat`	$\hat{p}_1$	166/8506
`numCancer2`	–	122
`n2`	n_2	8102
`p2hat`	$\hat{p}_2$	122/8102
`phatDiff`	$\hat{p}_1 - \hat{p}_2$	0.0044
`phat`	$\hat{p}$	$\frac{166+122}{8506+8102}$
`se`	$\sqrt{\hat{p} * (1 - \hat{p}) * (1/n_1 + 1/n_2)}$	0.00203
`z`	$\frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p} * (1-\hat{p}) * (1/n_1 + 1/n_2)}}$	2.2
`pval`	$2 * P(Z \geq \lvert z \rvert)$	0.03

Test and CI using prop.test()

- Can also use `prop.test()` function to get p-value and CI.

```
numCancer <- c(numCancer1,numCancer2)
n <- c(n1,n2)
prop.test(numCancer,n,conf.level=0.95,correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  numCancer out of n
## X-squared = 4.8387, df = 1, p-value = 0.02783
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.000498629 0.008416622
## sample estimates:
##     prop 1     prop 2
## 0.01951564 0.01505801
```

Software Notes

- Arguments to `prop.test()`:
  - The first two arguments are the numbers of successes (cancers) and number of trials (women), respectively.
  - `conf.level` is the level or coverage probability C of the interval (default $= 0.95$).
  - `correct` specifies whether to apply a "continuity correction" that improves the statistical inference when the total size of the sample is small. The default is `correct=TRUE`, but I set `correct=FALSE` to re-create the results from using the formulas in the text.
- Output:
  - Mostly like the output of `t.test()`
  - `X-squared` is the square of the test statistic Z that we have discussed. This has a chi-squared (χ^2) distribution with one df.
  - When the alternative hypothesis is two-sided, the p-value from the χ^2 test is equivalent to the p-value from the Z-test.