STAT 8010 R Session 5

Whitney Huang

6/14/2023

Contents

Session Objectives

- To gain experience with R, a programming language and free software environment for statistical computing and graphics.
- To perform categorical data analysis using `R`

Confidence interval for proportions

Cancer treatment survival rate

Researchers in the development of new treatments for cancer patients often evaluate the effectiveness of new therapies by reporting the *proportion* of patients who survive for a specified period of time after completion of the treatment. A new genetic treatment of 870 patients with a particular type of cancer resulted in 330 patients surviving at least 5 years after treatment. *Estimate* the proportion of all patients with the specified type of cancer who would survive at least 5 years after being administered this treatment.

```
n = 870; x = 330
# point estimate
phat <- x / n
phat
```

```
## [1] 0.3793103
```

```
# 95% CI for p
alpha = 0.05
ME = qnorm(1 - alpha / 2) * sqrt(phat * (1 - phat) / n)
phat + c(-1, 1) * ME
```

```
## [1] 0.3470683 0.4115524
```

```
# function
ci_prop <- function(x, n, alpha){
  phat <- x / n
  ME = qnorm(1 - alpha / 2) * sqrt(phat * (1 - phat) / n)
  return(phat + c(-1, 1) * ME)
}
ci_prop(330, 870, 0.05)
```

```
## [1] 0.3470683 0.4115524
```

Bird flu example

Among 900 randomly selected registered voters nationwide, 63% of them are somewhat or very concerned about the spread of bird flu in the United States.

- What is the point estimate for p, the proportion of U.S. voters who are concerned about the spread of bird flu?
- Construct a 95% CI for p

```
n = 900; x = 900 * .63
# point estimate
phat <- x / n
phat
```

```
## [1] 0.63
```

```r
# 95% CI for p
ci_prop(900 * .63, 900, 0.05)
```

```
## [1] 0.5984574 0.6615426
```

Sample size calculation example

A researcher wants to estimate the proportion of voters who will vote for candidate A. She wants to estimate to within 0.05 with 90% confidence.

```r
# True proportion is .9
p = 0.9; alpha = 1 - 0.9
n = p * (1 - p) * (qnorm(1 - alpha / 2) / 0.05)^2
n
```

```
## [1] 97.39956
```

```r
# True proportion is .6
p = 0.6
n = p * (1 - p) * (qnorm(1 - alpha / 2) / 0.05)^2
n
```

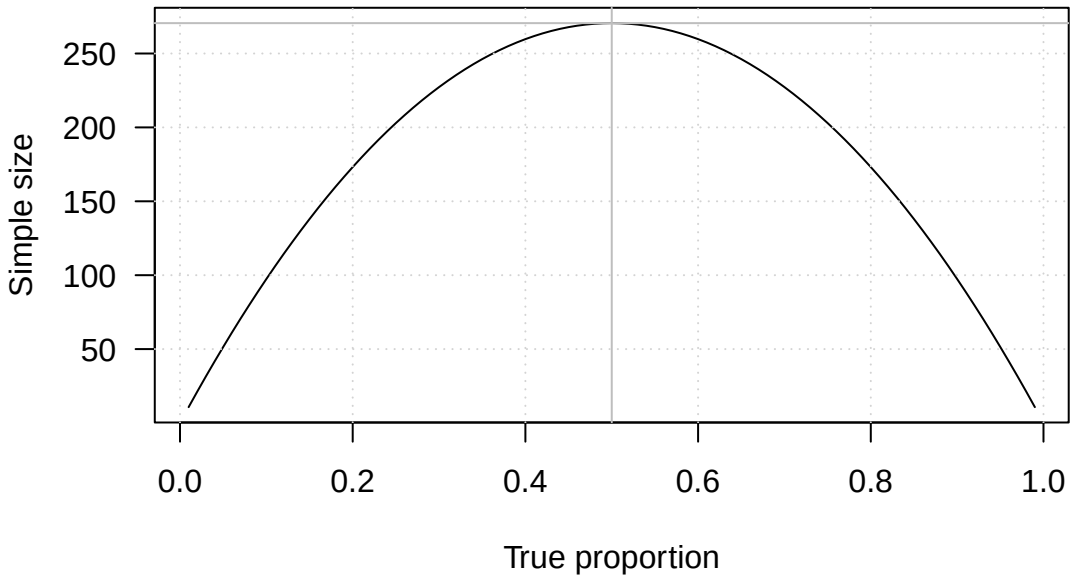
```
## [1] 259.7322
```

```r
# True proportion is .5
p = 0.5
n = p * (1 - p) * (qnorm(1 - alpha / 2) / 0.05)^2
n
```

```
## [1] 270.5543
```

```r
# Just for fun
p <- seq(0.01, 0.99, 0.01)
n = p * (1 - p) * (qnorm(1 - alpha / 2) / 0.05)^2

plot(p, n, type = "l", xlab = "True proportion", ylab = "Simple size", las = 1)
abline(v = 0.5, col = "gray")
abline(h = max(n), col = "gray")
grid()
```



Hypothesis testing

Bird flu example

Among 900 randomly selected registered voters nationwide, 63% of them are somewhat or very concerned about the spread of bird flu in the United States. Conduct a hypothesis test at .01 level to assess the research hypothesis: $p > .6$.

```
phat = 0.63
n = 900; x = phat * 900;
p_null = .6; alpha = 0.01
# Test statistic
zobs <- (phat - p_null) / sqrt(p_null * (1 - p_null) / n)
# P-vale of the right-tailed test
pnorm(zobs, lower.tail = F)
```

```
## [1] 0.03309629
```

```
# The Z-test here is in fact equlievent to chi-square test

prop.test(x, n, p = p_null, conf.level = 1 - alpha,
          alternative = "greater", correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  x out of n, null probability p_null
## X-squared = 3.375, df = 1, p-value = 0.0331
## alternative hypothesis: true p is greater than 0.6
## 99 percent confidence interval:
##  0.5918879 1.0000000
## sample estimates:
##    p
## 0.63
```

```r
# With Yates' continuity correction
prop.test(x, n, p = p_null, conf.level = 1 - alpha,
          alternative = "greater", correct = TRUE)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  x out of n, null probability p_null
## X-squared = 3.2512, df = 1, p-value = 0.03569
## alternative hypothesis: true p is greater than 0.6
## 99 percent confidence interval:
##  0.5913242 1.0000000
## sample estimates:
##    p
## 0.63
```

Wilson score CI

Proportion of CU vegetarian

```r
n = 25; x = 0
ci_prop(0, 25, 0.05)
```

```
## [1] 0 0
```

```r
## Wilson score CI
library(PropCIs)
scoreci(x, n, conf.level = .95)
```

```
##
##
##
## data:
##
## 95 percent confidence interval:
##  0.0000 0.1332
```

```r
## Check
prop.test(x, n, conf.level = 0.95, correct = F)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  x out of n, null probability 0.5
## X-squared = 25, df = 1, p-value = 5.733e-07
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.0000000 0.1331923
## sample estimates:
## p
## 0
```

Here the rule of three provides a quick and reasonable approximation

Inference for $p_1 - p_2$

A Simple Random Simple of 100 CU graduate students is taken and it is found that 79 strongly agree that they would recommend their current graduate program. A Simple Random Simple of 85 USC graduate students is taken and it is found that 52 strongly agree that they would recommend their current graduate program. At 5 % level, can we conclude that the proportion of strongly agree is higher at CU?

```
x <- c(79, 52); n <- c(100, 85)
phat <- x / n
pbar <- sum(x) / sum(n)
# Test statistic
zobs <- -diff(phat) / sqrt(pbar * (1 - pbar) / n[1] + pbar * (1 - pbar) / n[2])
# P-value
pnorm(zobs, lower.tail = F)
```

```
## [1] 0.003937328
```

```
prop.test(x, n, alternative = "greater", correct = F)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  x out of n
## X-squared = 7.0618, df = 1, p-value = 0.003937
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.06847019 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.7900000 0.6117647
```

```
# With Yates' continuity correction
prop.test(x, n, alternative = "greater", correct = F)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  x out of n
## X-squared = 7.0618, df = 1, p-value = 0.003937
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.06847019 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.7900000 0.6117647
```

Chi-Squared Tests

Example: Testing Mendel's theories

Among its many applications, Pearson's χ^2 test was used in genetics to test Mendel's theories of natural inheritance. Mendel crossed pea plants of pure yellow strain (dominant strain) plants of pure green strain. He predicted that second generation hybrid seeds would be 75% yellow and 25% green. One experiment produced $n = 8023$ seeds, of which $X_1 = 6022$ were yellow and $X_2 = 2001$ were green.

- Use Pearson's χ^2 test to assess Mendel's hypothesis.

```
x1 = 6022; x2 = 2001; n = 8023; p1 = .75; p2 = .25
# chi square test for p1 = .75
prop.test(x1, n, p = p1, correct = F)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  x1 out of n, null probability p1
## X-squared = 0.014999, df = 1, p-value = 0.9025
## alternative hypothesis: true p is not equal to 0.75
## 95 percent confidence interval:
##  0.7410061 0.7599381
## sample estimates:
##        p
## 0.750592
```

```
# Chi square test for p1 = .25
prop.test(x2, n, p = p2, correct = F)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  x2 out of n, null probability p2
## X-squared = 0.014999, df = 1, p-value = 0.9025
## alternative hypothesis: true p is not equal to 0.25
## 95 percent confidence interval:
##  0.2400619 0.2589939
## sample estimates:
##        p
## 0.249408
```

```
# Z test for p1 = .75
## test statistic
zobs <- (x1 / n - p1) / sqrt((p1 * p2) / n)
zobs^2
```

```
## [1] 0.01499855
```

```
## P-value
2 * (1 - pnorm(zobs))
```

```
## [1] 0.902528
```

```
# Yates' continuity correction
prop.test(x1, n, p = p1)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  x1 out of n, null probability p1
## X-squared = 0.012007, df = 1, p-value = 0.9127
## alternative hypothesis: true p is not equal to 0.75
## 95 percent confidence interval:
##  0.7409430 0.7599996
## sample estimates:
##        p
## 0.750592
```

Color preference example

In Child Psychology, color preference by young children is used as an indicator of emotional state. In a study of 112 children, each was asked to choose favorite color from the 7 colors indicated below. Test if there is evidence of a preference at the $\alpha = .05$ level.

Color	Blue	Red	Green	White	Purple	Black	Other
Frequency	13	14	8	17	25	15	20

```
x <- c(13, 14, 8, 17, 25, 15, 20)
chisq.test(x, correct = F)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  x
## X-squared = 11, df = 6, p-value = 0.08838
```

```
#Check
ek <- (sum(x) * 1 / length(x))
chisq <- sum((x - (sum(x) * 1 / length(x)))^2 / ek)
chisq
```

```
## [1] 11
```

```
## P-value
1 - pchisq(chisq, 6)
```

```
## [1] 0.08837643
```

Gender vs. Handness example

```r
x <- c(43, 9, 44, 4)
data <- matrix(x, nrow = 2, ncol = 2, byrow = TRUE)
dimnames(data) = list(Gender = c("M", "F"), Right = c("R", "L"))
data
```

```
##       Right
## Gender  R L
##      M 43 9
##      F 44 4
```

```r
chisq.test(data, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 1.7774, df = 1, p-value = 0.1825
```

```r
#Check
ek <- outer(rowSums(data), colSums(data)) / 100
chisq <- sum((data - ek)^2 / ek)
chisq
```

```
## [1] 1.777415
```

Marital status example

```r
x <- c(581, 487, 455, 477)
data <- matrix(x, nrow = 2, ncol = 2, byrow = TRUE)
dimnames(data) = list(Child = c("M", "D"), Parent = c("M", "D"))
data
```

```
##      Parent
## Child   M   D
##     M 581 487
##     D 455 477
```

```r
chisq.test(data, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 6.2088, df = 1, p-value = 0.01271
```

Purdue enrollment data example

The following contingency table contains enrollment data for a random sample of students from several colleges at Purdue University during the 2006-2007 academic year. The table lists the number of male and female students enrolled in each college. Use the two-way table to conduct a χ^2 test from beginning to end. Use $\alpha = .01$.

```
table <- matrix(c(378, 99, 104, 262, 175, 510), 3, 2)
colnames(table) <- c("Female", "Male")
rownames(table) <- c("Liberal Art", "Science", "Engineering")
table
```

```
##             Female Male
## Liberal Art    378  262
## Science         99  175
## Engineering    104  510
```

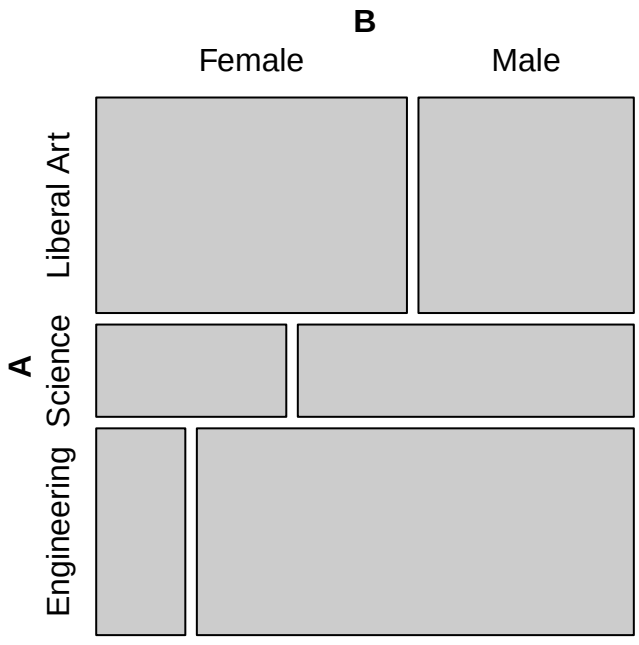
```
chisq.test(table, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 236.47, df = 2, p-value < 2.2e-16
```

```
library(vcd)
```

```
## Loading required package: grid
```

```
mosaic(table)
```



The Lady Testing Tea Example

```
TeaTasting <- matrix(c(3, 1, 1 ,3), nrow = 2,
                     dimnames = list(Guess = c("Milk", "Tea"),
                                     Truth = c("Milk", "Tea")))
TeaTasting
```

```
##       Truth
## Guess  Milk Tea
##   Milk    3   1
##   Tea     1   3
```

```
fisher.test(TeaTasting, alternative = "greater")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  TeaTasting
## p-value = 0.2429
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.3135693       Inf
## sample estimates:
## odds ratio
##   6.408309
```