CUNY MSDA 606 Spring 2016 Ch. 6 Graded Problems

James Topor

Chapter 6 Problems 6.6, 6.12, 6.20, 6.28, 6.44, 6.48

6.6
n = 1012; p = .46; 95% conf level; ME = .03

a) *We are 95% confident that between 43% and 49% of Americans in this sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.*

FALSE - the confidence interval applies to the population, not the sample.

b) *We are 95% confident that between 43% and 49% of Americans support the decision of the U.S. Supreme Court on the 2010 healthcare law.*

True

c) *If we considered many random samples of 1,012 Americans, and we calculated the sample proportions of those who support the decision of the U.S. Supreme Court, 95% of those sample proportions will be between 43% and 49%.*

True

d) *The margin of error at a 90% confidence level would be higher than 3%.*

FALSE - the margin of error will be lower than 3% due to a smaller critical value (90% = 1.64) being applied when compared to the critical value used when computing a 95% confidence interval (1.96). We can see this by calculating the margin or error for the critical values at 95% and 90%:

```
p = .46
n = 1012

# calculate margin of error at 90% confidence level
1.64 * sqrt( (p * (1 - p)) / n)
```

```
## [1] 0.02569386
```

```
# calculate margin of error at 95% confidence level
1.96 * sqrt( (p * (1 - p)) / n)
```

```
## [1] 0.0307073
```

6.12
n = 1259; p = .48

a) *Is 48% a sample statistic or a population parameter? Explain.*

48% is a sample statistic representing the proportion of sampled Americans who believe the use of marijuana should be made legal.

b) *Construct a 95% confidence interval for the proportion of US residents who think marijuana should be made legal, and interpret it in the context of the data.*

We are given n = 1259, p = .48

First, find the standard error SE = sqrt((p * (1-p)) / n) =>

```
n <- 1259
p <- .48

SE <- sqrt( (p * (1-p)) / n)
SE
```

```
## [1] 0.01408022
```

Then we calculate our usual confidence interval metrics: p +- critical value * SE where the critical value will be 1.96 since were are asked to find a 95% confidence interval:

```
lowerB <- p - 1.96 * SE
upperB <- p + 1.96 * SE
c(lowerB, upperB)
```

```
## [1] 0.4524028 0.5075972
```

So our 95% confidence interval is (0.45, 0.50)

c) *A critic points out that this 95% confidence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Is this true for these data? Explain.*

We have no information on how the data was collected, nor do we have any information regarding the sample distribution. As such, we have no way of determining whether the normal model is a good approximation in this instance.

d) *A news piece on this survey's findings states, "Majority of Americans think marijuana should be legalized." Based on your confidence interval, is this news piece's statement justified?*

No this news piece's statement is not justified since the 95% confidence interval spans only (0.45, .50) and both of these values are below the amount required for a majority.

6.20

If we wanted to limit the margin of error of a 95% confidence interval to 2%, about how many Americans would we need to survey ?

p = .48

.02 >= 1.96 * sqrt((p * (1 - p)) / n) =>

.02^2 >= 1.96^2 * (p * (1 - p)) / n =>

n = (1.96^2 * (p * (1 - p))) / .02^2

Solve for 'n':

```
p = .48
( 1.96^2 * (p * (1 - p)) ) / .02^2
```

```
## [1] 2397.158
```

So the sample size n would need to be at least 2398 (rounding up 2397.158) if we wanted to limit the margin of error to 2%.

6.28
DIFFERENCE OF TWO PROPORTIONS

Calculate a 95% confidence interval for the difference in proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data.

First we must calculate the standard error for the difference between two sample proportions:

SE = sqrt((pCal * (1 - pCal) / nCal) + (pOr * (1 - pOr) / nOr)) =>

```
pCal <- .08
nCal <- 11545
pOr <- .088
nOr <- 4691
SE = sqrt( (pCal* (1 - pCal) / nCal) + (pOr * (1 - pOr) / nOr) )
SE
```

```
## [1] 0.004845984
```

We then use our calculated SE to calculate the bounds of the 95% confidence interval: (pCal - pOr) +- 1.96 * SE

```
lowerB <- (pCal - pOr) - 1.96 * SE
upperB <- (pCal - pOr) + 1.96 * SE
c(lowerB, upperB)
```

```
## [1] -0.017498128  0.001498128
```

So our 95% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived is: (-0.017, 0.0014)

As such, we are 95% confident that the proportion of Californians who are sleep deprived differs by between -1.7% and 0.14% from the proportion of Oregonians who are sleep deprived.

6.44

TESTING FOR GOODNESS OF FIT USING CHI-SQUARE

a) *Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.*

H0: There is no evidence that barking deer prefer to forage in certain habitats over others

HA: There is evidence that barking deer prefer to forage in certain habitats over others

b) *What type of test can we use to answer this research question?*

To answer this research question we can make use of a chi-square goodness of fit test.

c) *Verify whether the conditions for a chi-square test have been met:*

1) *Is each case indepenent?* We have no information at all regarding whether the observations were either random or independent.
2) *Each particular scenario must have at least 5 expected cases*; Each case has at least 426/4 expected cases so this condition is satisfied
3) $df > 1$; df is 4-1 = 3 in this study

Unless we assume that the observations were random and independent, we cannot conclude that all of the preconditions for a chi-square test have been met. As such, we can only proceed with our hypothesis test if we assume the observations were random and independent.

d) *Do these data provide convincing evidence that barking deer prefer to forage in certain habitats over others? Conduct an appropriate hypothesis test to answer this research question.*

```
W = 4
C = 16
D = 61
O = 345

n <- W + C + D + O
n
```

```
## [1] 426
```

```
expP <- n/4
expP
```

```
## [1] 106.5
```

```
# Calculate the chi-square statistics (O-E)^2/E
chisW <- (W - expP)^2/expP
chisC <- (C - expP)^2/expP
chisD <- (D - expP)^2/expP
chisO <- (D - expP)^2/expP

chisTot <- chisW + chisC + chisD + chisO
chisTot
```

```
## [1] 214.4319
```

```
dfChi <- 4 - 1

# find the p value
pchisq ( q = chisTot, df = dfChi, lower.tail = FALSE)
```

```
## [1] 3.20845e-46
```

The p value is approximately ZERO, as such we reject the null hypothesis and conclude that barking deer do, in fact, prefer to forage in some habitats over others.

6.48

TESTING FOR INDEPENDENCE IN TWO WAY TABLES

n = 50739

a) *What type of test is appropriate for evaluating if there is an association between coffee intake and depression?*

We can use a chi-square test of independence to evaluate whether there is an association between coffee intake and depression.

b) *Write the hypotheses for the test you identified in part (a).*

H0: Coffee intake and depression are independent. Depression levels do not vary by the amount of coffee a woman consumes.

H1: Coffee intake and depression are not independent. Depression levels do vary relative to the amount of coffee a woman consumes.

c) *Calculate the overall proportion of women who do and do not suffer from depression.*

The overall proportion of women who suffer from depression is: 2607 / 50739 = .0513

```
2607 / 50739
```

```
## [1] 0.05138059
```

The overall proportion of women who do not suffer from depression is: 48132 / 50739 = .9486

```
48132 / 50739
```

```
## [1] 0.9486194
```

d) *Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic:*

The expected count for the highlighted cell is:

```
eCount <- 2607 * 6617 / 50739
eCount
```

```
## [1] 339.9854
```

The contribution of this cell to the test statistic is:

```
(373 - eCount)^2 / eCount
```

```
## [1] 3.205914
```

e) *The test statistic is chi^2 = 20.93. What is the p-value?*

We have 5-1 = 4 degrees of freedom, so the p-value is:

```
# find p value
pchisq ( q = 20.93, df = 4, lower.tail = FALSE)
```

```
## [1] 0.0003269507
```

f) *What is the conclusion of the hypothesis test?*

Since the p-value is approximately ZERO, we reject the null hypothesis and conclude that coffee intake and depression are not independent. Depression levels do vary relative to the amount of coffee a woman consumes.

g) *One of the authors of this study was quoted on the NYTimes as saying it was "too early to recommend that women load up on extra coffee" based on just this study. Do you agree with this statement? Explain your reasoning.*

Absolutely. First, we have no way of knowing whether the samples were either random or independent, nor can we be certain that some other confounding variable isn't responsible for the results of the study.