Chapter 6 - Inference for Categorical Data

Joshua Registe

2010 Healthcare Law. (6.48, p. 248) On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 46% of 1,012 Americans agree with this decision. At a 95% confidence level, this sample has a 3% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.

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

a) False, we are 95% confident that between 43$ and 49% of americans in the population support the decision of the US supreme court.

b) True, this correctly states the population estimate with our sample statistics and margin of error.

c) False, unless we know the true population statistic is, we can not say the what the random sample proportions should be.

d) False The margin of error at a 90% confidence level would be smaller at a 90% confidence interval because less certainty is needed.

Legalization of marijuana, Part I. (6.10, p. 216) The 2010 General Social Survey asked 1,259 US residents: "Do you think the use of marijuana should be made legal, or not" 48% of the respondents said it should be made legal.

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

a) 48% is a sample statistic since this is a survey of US residents that is likely used to estimate the entire population statistic.

b)

```
n <- 1259
p <- 0.48

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


CV<- qnorm(.975)

paste0("95% Confidence interval is [", round(p - SE*CV,2),", ", round(p+SE*CV,2),"]")
```

```
## [1] "95% Confidence interval is [0.45, 0.51]"
```

c) the conditions that need to be met for the sampling to be able to estimate the population statistic are for the samples to be independent, which is acheived through random sampling. and second, for the proportions * number of samples has to be greater than 10 which is satisfied.

d) based on the confidence interval, it seems to be more likely that majority of Americans do not think that, but there is a probability that this statement is true since the upper limit of our confidence interval is 0.51.

Legalize Marijuana, Part II. (6.16, p. 216) As discussed in Exercise above, the 2010 General Social Survey reported a sample where about 48% of US residents thought marijuana should be made legal. 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?

if we want to limit our margin of error for the confidence interval to 2, 2%, we would need the amount of samples shown below.

```
z<-qnorm(.975)

SE<- 0.02/z

paste0("We would need to sample ",ceiling((p*(1-p))/SE^2)," participants to redunce margin of error to 2%.")
```

```
## [1] "We would need to sample 2398 participants to redunce margin of error to 2%."
```

Sleep deprivation, CA vs. OR, Part I. (6.22, p. 226) According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insuffient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents. Calculate a 95% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data.

```
p_cal <-0.08
p_org <-0.088

n_cal <-11545
n_org <- 4691

CV<-qnorm(.975)

#checking for independence
0.08*11545
```

```
## [1] 923.6
```

```
(1-0.08)*11545
```

```
## [1] 10621.4
```

```
0.088*4691
```

```
## [1] 412.808
```

```
(1-0.088)*4691
```

```
## [1] 4278.192
```

```
paste("values all satisfy independence check")
```

```
## [1] "values all satisfy independence check"
```

```
SE_cal<- sqrt(p_cal*(1-p_cal)/n_cal)

SE_org<- sqrt(p_org*(1-p_org)/n_org)

SE_org_cal <- sqrt(SE_org+SE_cal)

lowerlim<- (p_cal-p_org)- CV*SE_org_cal
upperlim<- (p_cal-p_org)+ CV*SE_org_cal
paste0("95% Confidence interval for the California survey is [", round(p_cal - SE_cal*CV,2),", ", round
```

```
## [1] "95% Confidence interval for the California survey is [0.08, 0.08]"
```

```r
paste0("95% Confidence interval for the Oregon survey is [", round(p_org - SE_org*CV,2),", ", round(p_o
```

```
## [1] "95% Confidence interval for the Oregon survey is [0.08, 0.1]"
```

```r
paste0("95% Confidence interval for the difference in proportions between Oregon and California is [",
```

```
## [1] "95% Confidence interval for the difference in proportions between Oregon and California is [-0.
```

Barking deer. (6.34, p. 239) Microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China were examined from 2001 to 2002. In this region woods make up 4.8% of the land, cultivated grass plot makes up 14.7% and deciduous forests makes up 39.6%. Of the 426 sites where the deer forage, 4 were categorized as woods, 16 as cultivated grassplot, and 61 as deciduous forests. The table below summarizes these data.

Woods	Cultivated grassplot	Deciduous forests	Other	Total
4	16	67	345	426

(a) Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.
(b) What type of test can we use to answer this research question?
(c) Check if the assumptions and conditions required for this test are satisfied.
(d) Do these data provide convincing evidence that barking deer pre- fer to forage in certain habitats over others? Conduct an appro- priate hypothesis test to answer this research question.

a) null hypothesis - there is no preference for deer of where they prefer to forage. alternate hypothesis - deer prefer to forage in certain areas over others

b) we can use goodness of fit or chi squared test to answer this research question

c) we assume indepence and we assume that all categorical values have at least 5 expected cases. we can check this by:

```
n <-426
expect_woods<-.048*n
expect_grass<-.147*n
expect_forest<-.396*n
expect_other<-.409*n
c(expect_forest,expect_grass,expect_other,expect_woods)
```

```
## [1] 168.696  62.622 174.234  20.448
```

These are all greater than 5 so we are OK.

d) to approximate deer preference, we compute the z scores of the observed and expected values as such

```
observed_woods<-4
observed_grass<-16
observed_forest<-67
observed_other<-345

z_w<-(observed_woods- expect_woods)^2/expect_woods
z_g<-(observed_grass- expect_grass)^2/expect_grass
z_f<-(observed_forest- expect_forest)^2/expect_forest
z_o<-(observed_other- expect_other)^2/expect_other

z<-abs(sum(c(z_w,z_g,z_f,z_o)))
?chisq.test()
```

```
## starting httpd help server ... done
```

```r
x<-c(observed_forest,observed_grass,observed_other,observed_woods)
y<-c(expect_forest,expect_grass,expect_other,expect_woods)
pchisq(z,3, lower.tail = F)
```

```
## [1] 1.144396e-59
```

```r
chisq.test(x,y)
```

```
## Warning in chisq.test(x, y): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  x and y
## X-squared = 12, df = 9, p-value = 0.2133
```

Because of our low P value, we reject the null hypothesis that there is no preference for where deers choose to forage.

Coffee and Depression. (6.50, p. 248) Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician-diagnosed depression, and also asked about the use of antidepressants. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption.

		Caffeinated coffee consumption					
		≤ 1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	≥ 4 cups/day	Total
Clinical depression	Yes	670	373	905	564	95	2,607
	No	11,545	6,244	16,329	11,726	2,288	48,132
	Total	12,215	6,617	17,234	12,290	2,383	50,739

(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?
(b) Write the hypotheses for the test you identified in part (a).
(c) Calculate the overall proportion of women who do and do not suffer from depression.
(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. $(Observed - Expected)^2/Expected$).
(e) The test statistic is $\chi^2 = 20.93$. What is the p-value?
(f) What is the conclusion of the hypothesis test?
(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.

a) Chi squared test is appropriate for this evaluation.
b) null hypothesis - there is no relationship between the amount of coffee consumed and clinical depression alternate hypothesis - coffee has an effect on clinical depression
c)

```
(P_dep<-2607/50739)
```

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

```
(p_nodep<-1-P_dep)
```

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

d) the expected count for the highlighted cell is

```
(exp<-6617*P_dep)
```

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

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

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

e) If the test statistic is 20.93, degrees of freedom is 4 our p-value is

```
pchisq(20.93,4, lower.tail = F)
```

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

f) The conclusion of the hypothesis test is that we can reject our null hypothesis and there is sufficient evidence that coffee consumption and clinical depression are related.
g) Based on this study, women should not drink as much coffee as they want without potentially increasing the risk of depression unless more data is gathered and provides evidence otherwise.