ds606_hw6_jagdishchhabria

Jagdish Chhabria

March 29, 2019

Chapter 6 - Inference for Categorical Data

Practice: 6.5, 6.11, 6.27, 6.43, 6.47 Graded: 6.6, 6.12, 6.20, 6.28, 6.44, 6.48

6.6 2010 Healthcare Law. 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. False. The confidence interval is for the population inference, 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. A lower confidence level would translate to a lower margin of error because of a lower z-statistic (test statistic).

6.12 Legalization of marijuana, Part I. 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. It is a sample statistic which can be used as an unbiased estimator of the population parameter.

(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.

```
n=1259
p1=0.48
(std.err.gss=sqrt((p1*(1-p1))/n))
```

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

```
(me.gss=1.96*std.err.gss)
```

```
## [1] 0.02759723
```

```
(ci.gss.lo=p1+me.gss)
```

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

```
(ci.gss.hi=p1-me.gss)
```

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

The 95% confidence interval is (0.4524, 0.5075)

(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.

Assuming the sample has been collected randomly, the given sample size of 1259 is lower than 10% of the US population, and the sample proportion of 48% means that there were more than 10 successes and 10 failures. So the sampling distribution is expected to be normal, and the confidence interval is accurate.

(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, the statement is not justified because 48% does not constitute a majority.

6.20 Legalize Marijuana, Part II. As discussed in Exercise 6.12, 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 ?

```
(n1=(0.48*0.52)/(0.02/1.96)^2)
```

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

In order to limit the margin of error to 2%, we'd need to survey 2,398 Americans.

6.28 Sleep deprivation, CA vs. OR, Part I. According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient 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.ore=0.088
n.cal=11545
n.ore=4691
(me.sleep.hi=(p.cal-p.ore)+1.96*sqrt(((p.cal)*(1-p.cal)/n.cal) + ((p.ore)*(1-p.ore)/n.ore)))
```

```
## [1] 0.001498128
```

```
(me.sleep.lo=(p.cal-p.ore)-1.96*sqrt(((p.cal)*(1-p.cal)/n.cal) + ((p.ore)*(1-p.ore)/n.ore)))
```

```
## [1] -0.01749813
```

The 95% confidence interval for the difference in proportion of sleep-deprived Californians and Oregonites is (-0.01749, 0.00148).

6.44 Barking deer. 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 61 345 426

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

H0: The observed proportion of the different microhabitats where the barking deer forage is in line with the expected proportion of these sites

Ha: The observed proportion of the different microhabitats where barking deer forage is different from the expected proportion of these sites

(b) What type of test can we use to answer this research question? A chi-square test for a one-way table (difference between observed and expected counts) can be used to answer this research question.

(c) Check if the assumptions and conditions required for this test are satisfied. We assume that the sampled data is independent, and it can be seen that the expected count for each category is greater than 5, so the conditions for this test are satisfied.

(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.

```
microhabitats.obs<-c(4,16,61,345)
(microhabitats.exp<-c(0.048*426, 0.147*426, 0.396*426, 0.409*426))
```

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

```
n<-length(microhabitats.obs)
df<-n-1
chi<-(microhabitats.obs-microhabitats.exp)^2/microhabitats.exp
(chi.stat<-sum(chi))
```

```
## [1] 284.0609
```

```
(p.val<-1-pchisq(chi.stat, df))
```

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

Since the chi-square statistic is so large that the p-value is 0, the null hypothesis can be rejected i.e. it can be inferred that the barking deer have a preference where they forage which is different from the relative proportion of these microhabitat sites in nature.

6.48 Coffee and Depression. 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.

(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression? A chi-square test for a two-way table (contingency table) 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).

H0: There is no association between the proportion of women with clinical depression and their consumption of caffeinated coffee.

Ha: There is an association between the proportion of women with clinical depression and their consumption of caffeinated coffee.

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

```
dep<-2607
no_dep<-48132
total<-dep+no_dep

(dep.prop<-dep/total)
```

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

```
(no_dep.prop<-no_dep/total)
```

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

The overall proportion of women who do suffer from depression is 5.14%. The overall proportion of women who do not suffer from depression is 94.86%.

(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.

```
(two_six.exp<-(2607*6617)/50739)
```

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

```
(chi<-(373-two_six.exp)^2/two_six.exp)
```

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

The contribution of this cell to the test statistic is 3.205914.

(e) The test statistic is $\#2 = 20.93$. What is the p-value?

```
chi.test<-20.93
n=5
k=2
df=(n-1)*(k-1)
(p_val<-1-pchisq(chi.test, df))
```

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

(f) What is the conclusion of the hypothesis test? Since the p-value is so small, the null hypothesis can be rejected i.e. there does seem to be an association between caffeinated coffee consumption and depression.

(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.

Yes, I agree with this statement because while it can be inferred that there is an association between the caffeinated coffee consumption and the clinical depression, it has not been proven that there is causality. This is because based on the data provided, this seems to be an observational study and not an experimental study. There might be confounding variables involved.