Homework\_6

Priya Shaji

3/28/2019

## Inference for Categorical Data

### Ques\_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.

1. 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.

Answer a)

FALSE. The confidence interval applies to the entire population and not just to the sample taken.

1. 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.

Answer b)

FALSE. The 43% to 49% confidence interval range is not a probability taken for the sample mean.

1. 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%.

Answer c)

TRUE.

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

Answer d)

FALSE. For the same standard error (SE), 90% confidence interval will have a lower margin of error since we are throwing a narrower net. The z-score for 90% confidence interval is only 1.645 against 1.96 for a 95% confidence interval.

### Ques\_6.12

Legalization of marijuana, Part I. The 2010 General Social Survey asked 1,259 US res- idents: “Do you think the use of marijuana should be made legal, or not?” 48% of the respondents said it should be made legal.44

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

Answer a)

48% is a sample statistic. 48% of 1,259 US residents surveyed (sample data)

1. 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.

Answer b)

SE = sqrt((0.48 \* (1 - 0.48))/1259); SE # standard error

## [1] 0.01408022

LI = 0.48 - (qnorm(0.975)\*SE); LI # lower limit

## [1] 0.4524033

HI = 0.48 + (qnorm(0.975)\*SE); HI # upper limit

## [1] 0.5075967

1. 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.

Answer c)

There are 2 conditions for the sampling distribution of p̂ being nearly normal -

1. Observations are independent. The sample taken must have been and should not be more than 10% of the population. For this survey, we can reasonably say that this is true.
2. Success-failure condition. The sample size must also be sufficiently large - that is the success and failure rates must be greater than 10. In this case, this condition is true - 48% and 52% of 1,259 are greater than 10.
3. 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?

Answer d)

Yes, it is justified based on our confidence interval since our upper limit is around 51%. Since this is barely a majority though, it is also probably true that based on our findings, the new piece of survey may not be justified.

### Ques\_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 ?

Answer)

# ME = (qnorm(0.975) \* sqrt((0.48 \* (1 - 0.48))/n) margin of error  
  
# sample size at 2% margin of error and 95% control interval   
n = (0.48 \* (1 - 0.48)) / (0.02 / (qnorm(0.975)))^2; ceiling(n)

## [1] 2398

### Ques\_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 insu

Answer)

We assume that the sampling is near normal and thus we can apply the formula for the difference of 2 proportions.

SE\_cal = sqrt((0.08 \* (1 - 0.08))/11545); SE\_cal #standard error for California

## [1] 0.002524887

SE\_org = sqrt((0.088 \* (1 - 0.088))/4691); SE\_org #standard error for Oregon

## [1] 0.004136243

#standard error for the difference in proportion between Oregon and California   
SE\_org\_cal = sqrt(SE\_cal + SE\_org); SE\_org\_cal

## [1] 0.08161575

#confridence interval for the difference between the proportions of Oregonians and Californians  
#who are sleep deprived  
  
LL = (0.088 - 0.08) - (qnorm(0.975)\*SE\_org\_cal); LL # lower limit

## [1] -0.1519639

UL = (0.088 - 0.08) + (qnorm(0.975)\*SE\_org\_cal); UL # upper limit

## [1] 0.1679639

The 95% confidence interval between the difference in the proportion of sleep depreviation between Oregonians and Californians is between ~-15% to ~17%. This means that the proportion of Oregonians who are sleep deprived can be as much as 15% less than Californians or the proportions of Oregonians that are sleep deprived can be as much as 17% more than Californians for 95% of the samples taken with the given sample sizes. This means that there may also be no difference in the proportion of sleep deprived Californians and Oregonians since the interval includes 0.

### Ques\_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.62

Woods Cultivatedgrassplot Deciduousforests Other Total 4 16 67 345 426

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

Answer a)

H0: The barking deer does not prefer a certain habitat to forage. HA: The barking deer has certain habitats that it prefer to forage

1. What type of test can we use to answer this research question?

Answer b)

A chi-square test. Since we have cases that can be classified into several groups (deer habitats where they forage), we can determine if the forage habitats proportion is representative of the land make up.

1. Check if the assumptions and conditions required for this test are satisfied.

Answer c)

There are two conditions that must be checked before performing a chi-square test:

Independence. Each case that contributes a count to the table must be independent of all the other cases in the table. We assume that all the barking deer habitat variables (4) are independent of each other

Sample size / distribution. Each particular scenario (i.e. cell count) must have at least 5 expected cases. The woods habitat has only 4.8 cases - so this is in the lower boundary of what is accepted. We assume though that this is an acceptable count.

1. 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.

Answer d)

H0: The barking deer does not prefer a certain habitat to forage. HA: The barking deer has certain habitats that it prefer to forage.

# Z\_woods = (Obs\_cnt\_woods - Null\_cnt\_woods)/SE; SE = sqrt(Null\_cnt\_woods)  
# = (Obs\_cnt\_woods - Null\_cnt\_woods)^2/Null\_cnt\_woods  
 Z\_woods\_chisq = (0.9 - 4.8)^2/4.8  
  
# Z\_grass = (Obs\_cnt\_grass - Null\_cnt\_grass)/SE; SE = sqrt(Null\_cnt\_grass)  
# = (Obs\_cnt\_grass - Null\_cnt\_grass)^2/Null\_cnt\_grass  
 Z\_grass\_chisq = (3.8 - 14.7)^2/14.7  
   
# Z\_forest = (Obs\_cnt\_forest - Null\_cnt\_forest)/SE; SE = sqrt(Null\_cnt\_forest)  
# = (Obs\_cnt\_forest - Null\_cnt\_forest)^2/Null\_cnt\_forest  
 Z\_forest\_chisq = (14.3 - 39.6)^2/39.6   
  
# Z\_oth = (Obs\_cnt\_oth - Null\_cnt\_oth)/SE; SE = sqrt(Null\_cnt\_oth)  
# = (Obs\_cnt\_oth - Null\_cnt\_oth)^2/Null\_cnt\_oth  
 Z\_oth\_chisq = (81.0 - 40.9)^2/40.9   
   
Z\_all = Z\_woods\_chisq + Z\_grass\_chisq + Z\_forest\_chisq + Z\_oth\_chisq; Z\_all

## [1] 66.7306

1 - pchisq(Z\_all, 3)

## [1] 2.14273e-14

Since the p-value for the chi square distribution is very small, we can strongly reject the null hypothesis that deers have no preference in the habitats were they forage.

### Ques\_6.48

Coffee and Depression. Researchers conducted a study investigating the relationship between ca↵einated co↵ee 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 ca↵einated co↵ee 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 ca↵einated co↵ee consumption.63

1. What type of test is appropriate for evaluating if there is an association between co↵ee intake and depression?

Answer a)

chi square test to test for independence in 2 way table

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

Answer b)

H0: There is no relationship between coffee consumption and clinical depression.

HA: There is a relationship between coffee consumption and clinical depression.

1. Calculate the overall proportion of women who do and do not su↵er from depression.

Answer c)

p\_depressed = (2607/50739) \* 100; round(p\_depressed, digits=2) #women who are depressed

## [1] 5.14

p\_normal = (48132/50739) \* 100; round(p\_normal, digits=2) #women who are normal

## [1] 94.86

1. Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. (Observed

Answer d)

exp\_cnt = 6617 \* 0.0514; round(exp\_cnt,digit=0)

## [1] 340

obs\_cnt = 373  
  
(obs\_cnt - exp\_cnt)^2/exp\_cnt

## [1] 3.179824

1. The test statistic is

Answer e)

df = (5-1)\*(2-1); df

## [1] 4

1 - pchisq(20.93,df)

## [1] 0.0003269507

1. What is the conclusion of the hypothesis test?

Answer f)

Since the p-value is very small (0.0003), for similar samples there is only a 0.0003 chance that we reject the null hypothesis that there is no relationship between depression and coffee consumption among women.

1. 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.64 Do you agree with this statement? Explain your reasoning.

Answer g)

Yes, I agree with his statement. Based on this study, there is a very weak relationship between coffee consumption and depression among women.