HW6_StephenJones

Stephen Jones

March 30, 2019

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 doesn't refer to the sample, but any random sample from the population.

(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; this use of the confidence interval is appropriate with margin of error for random samples.

(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 US Supreme Court, 95% of those sample proportions will be between 43% to 49%.

True; this outlines the concept of what a confidence interval is.

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

False; with a lower confidence level, the margin of error would decrease with the range of the confidence interval.

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.

48% is a sample statistic. If the statistic passes specific criteria, it may be applied to the population.

(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.

```r
z <- 1.96
n <- 1259
p <- 0.48
SE <- ((p*(1 - p))/n)^0.5
low <- p - (z * SE)
high <- p + (z * SE)
cat("The confidence interval is (",low,",",high,")")
```

```
## The confidence interval is ( 0.4524028 , 0.5075972 )
```

(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.

Sampling distribution appears to be nearly normal and composed of independent observations. Proportions of "successes"" and "failures" are checked below.

```r
#p & n from prior code chunk
r1 <- p*n
r2 <- (1-p)*n
cat("Since",r1,"&",r2,"are sufficiently large, the success-failure criterion is met.")
```

```
## Since 604.32 & 654.68 are sufficiently large, the success-failure criterion is met.
```

(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 entirely justified. Not all–not even most–randomly composed samples of the sample will feature a majority that approve of legalization.

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?

$$ME = z \times SE \implies ME = z \times \sqrt{\left(\frac{p(1-p)}{n}\right)} \implies n = \frac{p \times (1-p) \times z^2}{ME^2}$$

```r
#for 95% confidence interval, z-score is 1.96
z <- 1.96

#value of p does not change from before
p <- 0.48

#from above
ME<- 0.02

n <- (p*(1-p)*(z^2))/(ME^2)
cat("We would have to sample",ceiling(n),"people to limit the margin of error to 2%.")
```

```
## We would have to sample 2398 people to limit the margin of error to 2%.
```

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.

```
z<-1.96
pCA<-.08
nCA<-11545
pOR<-.088
nOR<-4691

seCAsq<-(pCA*(1-pCA))/nCA
seORsq<-(pOR*(1-pOR))/nOR

#calculate the difference in standard error
se28<-sqrt(seCAsq + seORsq)

low <- pCA-pOR - (z * se28)
high <- pCA-pOR + (z * se28)
cat("The confidence interval is (",low,",",high,")")
```

```
## The confidence interval is ( -0.01749813 , 0.001498128 )
```

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.

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

H_0: there is no significant difference; deer have no preference over others
H_A: there is a significant difference; deer have a preference

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

Chi-square.

(c) Check if the assumptions and conditions required for this test are satisfied.

Independence is observed, as all habitats are independent. Sample sizes are marginally acceptable, with the lower limit of 5 challenged by inlcusion of 4 units of land classified as “woods”.

(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.

```r
land<-c(4,16,61,345)
finpct<-1-.048-.147-.396
pct<-c(.048,.147,.396,finpct)

deer<-as.data.frame(cbind(land,pct))
deer$ratio<-deer$pct*426

deer$pct<-NULL

chisq.test(deer)
```

```
##
##  Pearson's Chi-squared test
##
## data:  deer
## X-squared = 145.37, df = 3, p-value < 2.2e-16
```

```r
chi<-((deer$land-deer$ratio)^2)/deer$ratio

chisum<-sum(chi)

p_val <- 1 - pchisq(chisum,df=3)

cat("The secondary p-value agrees with the output above with p-value of",p_val)
```

```
## The secondary p-value agrees with the output above with p-value of 0
```

We reject the hull hypothesis and conclude that there is a significant difference in the foraging habitats, and the deer prefer some habitats over others.

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 is appropriate.

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

H_0: there is no association between coffee intake and depression
H_A: there is an association between coffee intake and depression

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

```
fTOT <- 50739
fDEP <- 2607
fFINE <- 48132

fDEPpct<-fDEP/fTOT
fFINEpct<-fFINE/fTOT

cat("There are",fDEP,"(",round(fDEPpct*100,2),"%) depressed women and",fFINE,"(",round(fFINEpct*100,2),
```

```
## There are 2607 ( 5.14 %) depressed women and 48132 ( 94.86 %) women who are not depressed in the samp
```

(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. $(\text{Observed} - \text{Expected})^2/\text{Expected}$.

```
exp<-fDEPpct*6617
cat("The expected value for the highlighted cell is",exp,"\n")
```

```
## The expected value for the highlighted cell is 339.9854
```

```
obs<-373

st<-((obs-exp)^2)/exp

cat("The resulting contribution is",st,"as calculated above.")
```

```
## The resulting contribution is 3.205914 as calculated above.
```

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

```
p_val<-pchisq(20.93,df=4,lower.tail=FALSE)

cat("The p-value is",p_val)
```

```
## The p-value is 0.0003269507
```

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

We reject the null hypothesis and accept the alternative and conclude that there is an association between coffee intake 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, as an avid coffee drinker I recommend it to everyone, however, more data need to be collected and more trials conducted. This assumes that depression is equally likely among all women in the sample.