charleyferrari_week6hw

Charley Ferrari

October 26, 2015

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. In this sample we have perfect information, so we are 100% certain 46% of this sample supports the decision.

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, the sample allows us to make an inference about the source population.

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%.

False. We don't know what the true population proportion is. Our confidence interval gives us a range of possible values for the true population proportion.

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

False. It would be lower, since we're allowed to be less confident.

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 shouldbe made legal.

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

48% is a sample statistic, since it's calculated based on a 1,259 sample of the total US 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
n <- 1259
p <- 0.48
se <- sqrt((p*(1-p))/n)
t <- qt(0.975,n-1)
moe <- t*se
CI <- c(p-moe, p+moe)
CI
```

```
## [1] 0.4523767 0.5076233
```

This confidence interval means that, with 95% certainty, we can say that the proportion of Americans that think marijuana should be legalized is between 45.2% and 50.8%.

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.

This is true for the data, because we're looking at a proportion $\hat{p}$. As long as the sample observations are independent and the sample size is large enough such that the propotion pro or against (p or 1-p) is greater than 10, $\hat{p}$ will be normally distributed. Since our p's are hovering around 50%, we can be save with a sample size of 1,259.

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?

We cannot reject the hypothesis that the proportion of Americans who think marijuana should be legalized is above 50, however we also cannot reject the hypothesis that the proportion is below 50.

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 ?

$$MOE = t \times \sqrt{\frac{p(1-p)}{n}}$$

We want to solve for n:

$$n = \frac{t^2 p(1-p)}{MOE^2}$$

```r
p <- 0.48
z <- qnorm(0.975)
moe <- 0.02
n <- (t^2 * p * (1-p))/moe^2
n
```

```
## [1] 2401.689
```

(Using z instead of t to more quickly solve this problem)

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 insuficient 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.

H_0: $\mu_{California} - \mu_{Oregon} = 0$ H_a: $\mu_{California} - \mu_{Oregon} \neq 0$

```
pc <- 0.08
po <- 0.088
nc <- 11545
no <- 4691
se <- sqrt((pc*(1-pc)/nc)+(po*(1-po)/nc))
tstat <- (pc-po)/se
pvalue <- pt(tstat,no-1)
pvalue
```

```
## [1] 0.01423446
```

This p-value is greater than 0.05, so we cannot reject the null hypothesis that the two means are equal.

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 make 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: Barking deer don't prefer one habitat over another, and therefore their foraging proportions equal the proportion of habitats on the island

H_a: At least one (by necessity two) of the proportion of habitats are not equal to the proportions on the island, meaning the barking deer have preferences in foraging habitats.

b. What type of test can we use to answer the research question?

We can use a chi-square test to answer this question

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

The assumptions and conditions are not satisfied. There are only 4 sites that are 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.

The assumptions aren't met, but lets try this based on the percentages calculated on what we h ave.

We're given percentages of habitat in Hainan, so I'll convert the number of sites given to percentages:

```r
n <- 426

pw <- 4/n
pg <- 16/n
pf <- 61/n
po <- 345/n

pw0 <- 0.048
pg0 <- 0.147
pf0 <- 0.396
po0 <- 1-pw0-pg0-pf0

chisquare <- (pw-pw0)^2/pw0 + (pg-pg0)^2/pg0 + (pf-pf0)^2/pf0 + (po-po0)^2/po0

chisquare
```

```
## [1] 0.6668097
```

```r
# number of groups = 4

ng <- 4
df <- ng-1

1-pchisq(chisquare,df)
```

```
## [1] 0.8809814
```

Our p-value is 0.88, which is well above 0.05. We can safely reject the null hypothesis that barking deer have no preference in where they forage.

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 for evaluating if there is an association between coffee intake and depression.

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

H_0: The amount of coffee one drinks is not related to depression H_a: The amount of coffee one drinks is related to depression

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

```
ntotdep <- 2607
ntot <- 50739
ptot <- ntotdep/ntot

ptot
```

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

```
1-ptot
```

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

d. Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic: $\frac{(O_k - E_k)^2}{E_k}$

```
ntot2cup <- 6617
Ek <- (ntotdep*ntot2cup)/ntot
Ek
```

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

```
chipart <- (373-Ek)^2/Ek
chipart
```

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

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

```
df <- (5-1)*(2-1)
chi2 <- 20.93

1-pchisq(chi2,df)
```

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

f. What is the conclusion of the hypothesis test?

The p-value is below 0.05, so we cannot reject the null hypothesis that coffee doesn't cause depression.

g. I agree with this statement, since according to the chi-square test, we found no link between coffee consumption and depression.

6.48: Choose a test

We would like to test the following hypotheses:

H_0: $p = 0.1$ H_a: $p \neq 0.1$

The sample size is 120 and the sample proportion is 8.5%. Determine which of the tests below are appropriate, and explain your reasoning.

This is mostly straight forward, and we should use the z-test for a proportion. However, np is dangerously close to 10. We might want to exercize caution when applying this test.