DATA 606 - Homework 6

Joshua Sturm

11/12/2017

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.

Solution

False. The confidence interval is for the entire population, not just 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.

Solution

True. The confidence interval tells us about the entire 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%.

Solution

False. The confidence interval only tells us about the population mean, not the proportions.

(d)

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

Solution

False. Since we don't need to be as certain with a lower confidence level, the interval will be more narrow, and so the margin of error will be lower.

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.

Solution

It comes from sample data, so it is a sample statistic.

(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.

Solution

$SE = \sqrt{\frac{p \cdot (1-p)}{n}}$ $ME = z^* SE$

```
n <- 1259
p <- 0.48
z <- qnorm(1-.05/2)
SE <- sqrt((p*(1-p)) / n)
ME <- z*SE
ci.lower <- p - ME
ci.upper <- p + ME
```

We are 95% sure that between 45.24% and 50.76% of Americans are in favour of marijuana being made legal. $ ### (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.

Solution

We can use the normal approximation because all the conditions are satisfied. Samples are assumed to be independent, make up less than 10% of the population, and the success-failure condition is valid.

(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?

Solution

No. The interval is mostly below the 50% mark, so it's unfair to call it 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?

Solution

$ME = z^* SE$
$SE = \sqrt{\frac{p \cdot (1-p)}{n}}$
$\frac{ME}{z} = \sqrt{\frac{p \cdot (1-p)}{n}} \rightarrow \left(\frac{ME}{z}\right)^2 = \frac{p \cdot (1-p)}{n}$
Plugging in our values, we get:
$n \geq (1.96)^2 \times \frac{0.48(1-0.48)}{(0.02)^2}$
Solving for n, we need at least 2398 people to ensure a margin of error $\leq 2\%$ with 95% confidence.

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 insuffcient 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.

Solution

The sample was randomly selected, and made up of less than 10% of the population. The success-failure condition is also satisfied, so we can use the normal approximation. Formula 6.9: $SE_{\hat{p_1}-\hat{p_2}} = \sqrt{SE_{\hat{p_1}}^2 + SE_{\hat{p_2}}^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

```
n.ca <- 11545
p.ca <- 0.08

n.or <- 4691
p.or <- 0.088

p.diff <- p.ca - p.or

se.ca <- (p.ca*(1-p.ca)) / n.ca
se.or <- (p.or*(1-p.or)) / n.or
SE <- sqrt(se.ca + se.or)

z <- qnorm(1-.05/2)
ME <- z*SE

ci.l <- p.diff - ME
ci.u <- p.diff + ME
```

We are 95% confident that the difference in California's and Oregon's populations with sleep deprivation is between -0.017498 and 0.001498.

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.

\begin{center}

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.

Solution

H_0: Deer have no foraging location preference. H_A: Deer do have a foraging location preference.

(b)

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

Solution

We can use a chi-square test.

(c)

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

Solution

There are two conditions needed to perform a chi-square test. Each case must be independent of the others, and Each scenario must have at least 5 expected cases. We'll assume that the cases are independent.

```
n <- 426
pct <- c(0.048, 0.147, 0.396, 1-0.048-0.147-0.396)
expct <- pct * n
expct
```

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

Since each scenario has at least 5 expected cases, the second condition is 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.

Solution

$\chi^2 = \frac{(O_1 - E_1)^2}{E_1}$

```
observed <- c(4, 16, 61, 345)

chi.sq <- sum(((observed - expct)^2) / expct)

k <- NROW(observed)
df <- k - 1

pchisq(chi.sq, df = df, lower.tail = F)
```

```
## [1] 2.799724e-61
```

Since the p-value is ≈ 0, we reject the null hypothesis, and conclude that deer, indeed, have a preference as to 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.

{

		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?

Solution

We can use the chi-square test for two-way tables.

(b)

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

Solution

H_0: There is no association between caffeinated coffee consumption and depression in women. H_A: There is an association between caffeinated coffee consumption and depression in women.

(c)

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

Solution

Women who suffer from depression: $\frac{2607}{50739} = 0.05138059 \approx 5.14\%$. Women who do not suffer from depression: $\frac{48132}{50739} = 0.9486194 \approx 94.86\%$.

(d)

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

Solution

Expected $\text{Count}_{row\ i,\ col\ j} = \frac{(\text{row } i \text{ total}) \times (\text{column } k \text{ total})}{\text{table total}}$. Expected count: $\frac{2607}{50739} \times 6617 = 339.9854 \approx 340$. Cell's contribution to test statistic: $\frac{(373-339.9854)^2}{339.9854} = 3.205914$.

(e)

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

Solution

df = (number of rows minus 1) $\times$ (number of columns minus 1)

```
chi.sq <- 20.93
df <- (2 - 1)*(5 - 1)
p <- pchisq(chi.sq, df = df, lower.tail = F)
p
```

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

(f)

What is the conclusion of the hypothesis test?

Solution

Since $p = 0.0003269507 < 0.05$, we reject the null hypothesis, and conclude that there is an association between women drinking caffeinated coffee, and experiencing 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.

Solution

I agree with the statement, because this study was observational, not experimental, so we can't draw conclusions from it; there may be other factors involved.