homework_ch6_JimmyNg

Jimmy Ng

October 25, 2018

Question 6.6

(a) False. The confidence intervals (CI) is not describing this sample; the CI should be used to address the population, not the sample. In this case, the proportion is already found/calculated for this particular sample, which is 46%.
(b) True. That is the 95% confidence intervals. We are 95% confident that the true proportion (from the population) would fall between these intervals, given the sampling methodology met the criteria, e.g. random sampling, independence, etc.
(c) True. If we created a sampling distribution using 1012 in each sample, we should expect to see that the point estimate fall between these intervals 95% of the time.
(d) False. Since the critical value of 90% is less than 95% (1.64 < 1.96), and the margin of error is equal to the critical value multiplied by the standard error (SE), we should expect to see that the margin of error decreases with 90% confidence when the SE remains constant.

Question 6.12

(a) It's a sample statistics. It is derived from the 1259 samples, but we can use this sample statistics to generate a meaningful parameter for (estimating) the population.
(b) The SE is equal to sqrt((0.48)*(1 - 0.48)) / sqrt(1259) = 0.01408022. The margin of error (ME) for 95% CI is equal to 1.96 multiplied by SE, i.e. 0.02759723. Putting them together, we have the 95% CI [45.2%, 50.8%]. The result indicated that we are 95% confident that roughly 45% to 51% of the population is in support of legalizing marijuana in the US.
(c) As long as the observations are independent and less than 10% of the population, e.g. each data point collected is independent, meaning the individuals' opinion in this survey are not influenced by other survey-takers, and we can meet the "success-failure" condition (in this case, np >= 10 and n(1 -p) >= 10), we can approximate the data using normal distribution.
(d) It's hard to justify this claim when the intervals is crossing 50, i.e. between 45.2% and 50.8%. We cannot say for sure that majority of Americans are supporting it.

Question 6.20

We can set this up as following,

0.02 = 1.96 * (sqrt((p)(1-p)) / sqrt(n)) 0.02 = 1.96 * (sqrt((0.48)*(0.52)) / sqrt(n)) 0.02 = 1.96 (0.4995998 / sqrt(n)) 0.01020408 = 0.4995998 / sqrt(n) sqrt(n) = 48.96079 n = 2397.159

We need at least 2398 participants for the survey in order to achieve the result.

Question 6.28

```
ca <- 0.08
or <- 0.088
ca_n <- 11545
or_n <- 4691
```

```r
critical_value <- 1.96

# margin of error (me)
ca_me <- critical_value * ( sqrt((ca)*(1-ca)) / sqrt(ca_n) )
or_me <- critical_value * ( sqrt((or)*(1-or)) / sqrt(or_n) )

print(paste("the 95% confidence intervals of CA sleep deprivation is between ",
      round(100 * (ca - ca_me), 1),
      "% and ",
      round(100 * (ca + ca_me), 1),
      "%",
      sep = ""))
```

```
## [1] "the 95% confidence intervals of CA sleep deprivation is between 7.5% and 8.5%"
```

```r
print(paste("the 95% confidence intervals of OR sleep deprivation is between ",
      round(100 * (or - or_me), 1),
      "% and ",
      round(100 * (or + or_me), 1),
      "%",
      sep = ""))
```

```
## [1] "the 95% confidence intervals of OR sleep deprivation is between 8% and 9.6%"
```

```r
# calculate the 95% confidence interval for the difference between CA and OR
diff <- ca - or
se <- sqrt( ((ca * (1 - ca)) / ca_n) + ((or * (1 - or)) / or_n) )
me <- critical_value * se

print(paste("the 95% confidence intervals for the difference between CA and OR is between ",
      round(100 * (diff - me), 1),
      "% and ",
      round(100 * (diff + me), 1),
      "%",
      sep = ""))
```

```
## [1] "the 95% confidence intervals for the difference between CA and OR is between -1.7% and 0.1%"
```

The 95% confidence intervals of CA sleep deprivation is between 7.5% and 8.5%, whereas the 95% confidence intervals of OR sleep deprivation is between 8% and 9.6%. The 95% confidence interval of the difference between the proportions of CA and OR is between -1.7% and 0.1%. Judging from the result, there's no significant difference of sleep deprivation between CA and OR residents. They have similar proportions of residents who self-identify as sleep deprived.

Question 6.44

(a) Ho: deer has no preference in choosing where to forage in certain habitats, whereas the Ha: deer has a preference in choosing where to forage.
(b) We can choose a chi-square distribution test for this categorical variable.
(c) Independence is met; however, the "minimium-bin-count" condition is not, i.e. one of the value (Wood) has less than 5 cases.
(d) Although one of the conditions is not met, we can still apply the chisq.test() to test for our hypothesis. However, we should add the argument "correct = T" in order to apply Yate's continuity correction for fixing the problem. Below is the result and we can see that the p-value is close to 0. In other words, we

can reject our null hypothesis. Indeed, deer has a preference in choosing where to forage in certain habitats.

```
deer <- matrix(c(4, 16, 67, 345),
               nrow = 1,
               dimnames = list(c("deer"),
                               c("W", "Cg", "Df", "O")))
chisq.test(deer, correct = T)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  deer
## X-squared = 714.17, df = 3, p-value < 2.2e-16
```

Question 6.48

(a) A chi-square distribution test would be appropriate to test the relationship between these two categorical variables.

(b) Ho: there is no relationship between level of coffee in-take and clinical depression; whereas the Ha: there is a relationship between coffee in-take and clinical depression.

(c) Proportion of women who suffered from depression: 2607 / 50739 = 0.05138059, whereas the proportion of women who did not suffer from depression: 48132 / 50739 = 0.9486194.

(d) (2607 / 50739) * (6617) = 339.9854 is the expected count. (Observed - Expected)^2 / Expected: (373 - 339.9854)^2 / 339.9854 = 3.205914.

(e) df = (R - 1) * (C - 1), i.e. (2 - 1) * (5 - 1) = 4. The Chi-squared value is 20.93 with df = 4, the p-value is equal to pchisq(q = 20.93, df = 4, lower.tail = F) = 0.0003269507.

(f) The result indicated that there's a significant relationship between coffee in-take and clinical depression. The p-value is much lower than 0.05 or even 0.01. It seems that women who drink more coffee are less likely to get depression. We should reject the null hypothesis.

(g) Yes it is too early to call for that. First of all, this is a quasi-experimental design. We cannot draw direct causation from here. For instances, we cannot assign women to be depressed or not, neither can we assign (or control) women to drink more or less of coffee. This is just an observational study and we need better control for this type of study in order to rule out any confounding variable.