MSDS 5043 - Assignment 4

Olivia Samples, Lipscomb University

10/13/2019

Question 1: Problem 6.6

a. False, this confidence interval applies to the total population, not the sample population.
b. True, this is the total population for which the confidence interval applies.
c. False, this is not necessarily true. While 95% of the total population lies within that range, the sample populations might not exactly represent that total population.
d. False, the margin of error would be less because we would be less confident resulting in a narrower margin. #### Question 2: Problem 6.33 We will use the 2-proportion z-test for this hypothesis test. Here, proportions associated with c are controlled and with t are truck drivers.

We have that the standard error estimate is reasonable because these are independent samples and the point estimate is nearly normal and unbiased.

```
pc = 35/292
nc = 292

pt = 35/203
nt = 203

pooled = 70/(nc + nt)
nc*pooled >= 10
```

```
## [1] TRUE
```

```
nt*pooled >= 10
```

```
## [1] TRUE
```

```
nc*(1-pooled) >= 10
```

```
## [1] TRUE
```

```
nt*(1-pooled) >= 10
```

```
## [1] TRUE
```

Our null hypothesis is $H_0 : p_c = p_t$. We are testing to see if the alternative is significantly different, or if $H_A : p_c \neq p_t$. We will test this at a 95% confidence level, or with a significance level of $\alpha = 0.05$.

Now we will calculate the test statistic.

```
#point estimate
pe = pc - pt

#standard error
se = sqrt(pooled*(1-pooled)*((1/nc) + (1/nt)))

#z-score
z = pe/se
z
```

```
## [1] -1.650356
```

```
p.value = 2*pnorm(z)
p.value
```

```
## [1] 0.09887008
```

And so, our p-value is greater than our significance value which means that we fail to reject our null hypothesis. Furthermore, the data does not provide sufficient evidence of a difference between the proportions of truck drivers and non-transportation workers who get less than 6 hours of sleep per day. #### Question 3: Problem 6.42 (a) Respondents for each category 1, 2, 3, 4 in 2010 are as follows:

```
n = 1019
r1 = .38 * n
r2 = .16 * n
r3 = .40 * n
r4 = .06 * n
t.o = c(r1, r2, r3, r4)
t.o
```

```
## [1] 387.22 163.04 407.60  61.14
```

b. Our null hypothesis is $H_0 : p_c = p_t$, or the distribution of responses in 2001 matches the distribution of responses in 2010. We are testing to see if $H_A : p_c \neq p_t$, or if the distribution of responses does not match.
c. The expected number of respondents for each category 1, 2, 3, 4 in 2001 are as follows (such that the null hypothesis is true):

```
n = 1019
r1 = .37 * n
r2 = .12 * n
r3 = .45 * n
r4 = .06 * n
t.t = c(r1, r2, r3, r4)
t.t
```

```
## [1] 377.03 122.28 458.55  61.14
```

d. We will be using the X^2 goodness of fit test to test this hypothesis. We have a simple random sample and we have verified that all expected counts are ≥ 5. Our hypothesis was stated above. We will test this at a 95% confidence level, or with a significance level of $\alpha = 0.05$.

```
#chi squared statistic
#z1 = (t.t[1]-t.o[1])^2/t.t[1]
#z2 = (t.t[2]-t.o[2])^2/t.t[2]
#z3 = (t.t[3]-t.o[3])^2/t.t[3]
#z4 = (t.t[4]-t.o[4])^2/t.t[4]
#x2 = z1 + z2 + z3 + z4
#x2


#chi squared test in r
x = t.o
p = c(.37,.12,.45,.06)
chisq.test(x, p=p)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  x
## X-squared = 19.523, df = 3, p-value = 0.0002131
```

We have that the p-value is 0.002131 which is less than 0.05, so we reject the null hypothesis. There is sufficient evidence to suggest that peoples' beliefs on the origin of human life have changed since 2001.

Question 4: Problem 6.50

a. We will be using the X^2 goodness of fit test to test this hypothesis. We have a simple random sample and we have verified that all expected counts are ≥ 5. Our hypothesis is as follows. H_0 : the proportion of the ANES sample = the proportion of population. H_A : at least one proportion of the ANES sample is different than the proportion of population. We will test this at a 95% confidence level, or with a significance level of $\alpha = 0.05$.

```
pp = c(.18, .22, .37, .23)
x = c(83, 121, 193, 103)
chisq.test(x, p=p)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  x
## X-squared = 300.44, df = 3, p-value < 2.2e-16
```

We have that the p-value is significantly lower than 5% such that there is sufficient evidence to reject the null Hypothesis. And so, the ANES sample is not representative of the population distribution of US residents. (b) (i) Here we would assume that region affects the feeling about the country's direction, rather than the other way around. Thus, region is the explanatory variable and feeling about the country's direction is the response variable. (ii) We would like to determine if there are actualy differences in the peoples' beliefs based on regions. Our hypothesis test is as follows.

H_0 : There is no difference in beliefs between the four regions. H_A : There is some difference in beliefs between the four regions.

iii. We have multiple random samples with all expected counts ≥ 5. We will test this at a 95% confidence level, or with a significance level of $\alpha = 0.05$.

```
R <- as.table(rbind(c(29, 44, 62, 36), c(54, 77, 131, 67)))
Xsq <- chisq.test(R)
Xsq
```

```
##
##  Pearson's Chi-squared test
##
## data:  R
## X-squared = 0.66724, df = 3, p-value = 0.8809
```

And so, we have a p-value of 0.8809 which is greater than 0.05. Thus, there is not sufficient evidence to reject the null Hypothesis. In other words, we cannot reject that there is no difference in beliefs between the four regions.

Question 5

In August of 2012, news outlets ranging from the Washington Post to the Huffington Post ran a story about the rise of atheism in America. The source for the story was a poll that asked people, “Irrespective of whether you attend a place of worship or not, would you say you are a religious person, not a religious person or a convinced atheist?” This type of question, which asks people to classify themselves in one way or another, is common in polling and generates categorical data. In this lab we take a look at the atheism survey and explore what’s at play when making inference about population proportions using categorical data.The report is availabe on Cavas.

The data

The data is available on Canvas. Load the csv data first. Then using the command below, create a new data frame called us12 that contains only the rows in atheism associated with respondents to the 2012 survey from the United States.

```
atheism = read.csv("/Users/osamples/Desktop/MSDS5043/atheism.csv", head = TRUE, sep =
",")

us12 <- subset(atheism, atheism$nationality == "United States" & atheism$year == "2012")
head(us12)
```

```
##         nationality    response year
## 49926 United States non-atheist 2012
## 49927 United States non-atheist 2012
## 49928 United States non-atheist 2012
## 49929 United States non-atheist 2012
## 49930 United States non-atheist 2012
## 49931 United States non-atheist 2012
```

5.1:

Calculate the proportion of atheist responses. Does it agree with the percentage in Table 6? If not, why?

```
w = table(us12$response)
w
```

```
##
##     atheist non-atheist
##          50         952
```

```
prop = 50/length(us12$response)
prop
```

```
## [1] 0.0499002
```

Yes, Table 6 gives 5% of the United States population in 2012 were atheists. This is the same as what our proportion gives us here.

5.2

Although formal confidence intervals and hypothesis tests don't show up in the report, suggestions of inference appear at the bottom of page 6: "In general, the error margin for surveys of this kind is ± 3% – 5% at 95% confidence."

What is the margin of error for the estimate of the proportion of atheists in US in 2012?

```
se = sqrt(prop*(1-prop)/length(us12$response))
me = 1.96 * se
me
```

```
## [1] 0.01348211
```

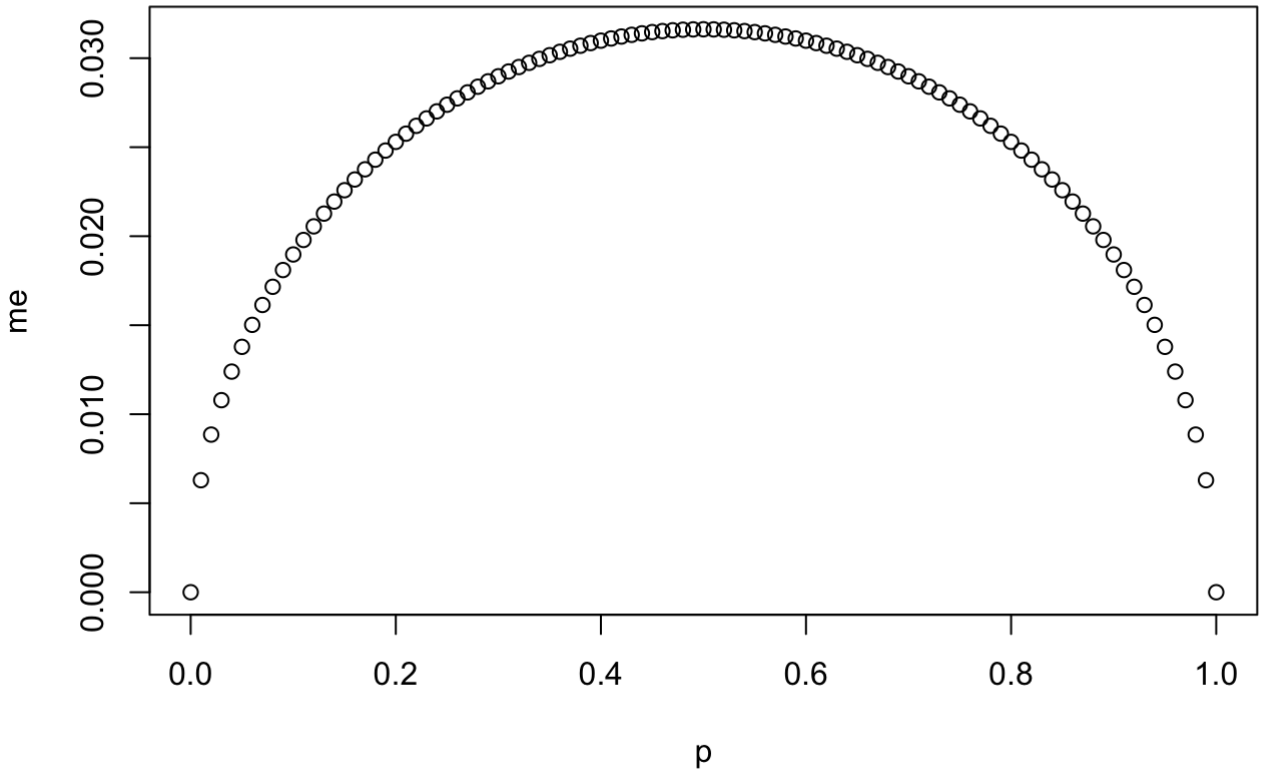
And so, our margin of error at a 95% confidence interval is 0.0135. ###### 5.3 How does the proportion affect the margin of error The formula for the standard error: $SE = \sqrt{p\,(1 - p)/n}$. This is then used in the formula for the margin of error for a 95% confidence interval: $ME = 1.96 \times SE = 1.96 \times \sqrt{p\,(1 - p)/n}$. Since the population proportion p is in this ME formula, it should make sense that the margin of error is in some way dependent on the population proportion. We can visualize this relationship by creating a plot of ME vs. p.

The first step is to make a vector p that is a sequence from 0 to 1 with each number separated by 0.01.

We can then create a vector of the margin of error (me) associated with each of these values of p using the familiar approximate formula (ME = 2 SE).

Lastly, we plot the two vectors against each other to reveal their relationship.

```
n <- 1000
p <- seq (0, 1, 0.01)
me <- 2 * sqrt(p*(1-p)/n)
plot(me ~ p)
```



Describe the relationship between p and me . As the proportion increases to 50%, the margin of error increases. As the proportion increases from 50 to 100%, the margin of error decreases. In other words, the margin of error reaches a maximum at p = 0.5.

5.4

If you refer to Table 6, you'll find that Australia has a sample proportion of 0.1 on a sample size of 1040, and that Ecuador has a sample proportion of 0.02 on 400 subjects. Let's suppose for this exercise that these point estimates are actually the truth. Then given the shape of their respective sampling distributions, do you think it is sensible to proceed with inference and report margin of errors, as the reports does?

```
aus12 <- subset(atheism, atheism$nationality == "Australia" & atheism$year == "2012")
a = table(aus12$response)
a
```

```
##
##     atheist non-atheist
##         104         935
```

```
aprop = 104/length(aus12$response)
aprop
```

```
## [1] 0.1000962
```

```
#Australia margin of error
se = sqrt(aprop*(1-aprop)/length(aus12$response))
me = 1.96 * se
me
```

```
## [1] 0.01824968
```

```
ec12 <- subset(atheism, atheism$nationality == "Ecuador" & atheism$year == "2012")
e = table(ec12$response)
e
```

```
##
##     atheist non-atheist
##           8         396
```

```
eprop = 8/length(ec12$response)
eprop
```

```
## [1] 0.01980198
```

```
#Ecuador Margin of Error
se = sqrt(eprop*(1-eprop)/length(ec12$response))
me = 1.96 * se
me
```

```
## [1] 0.01358553
```

These margin of errors (United States, Australia, and Ecuador) are less than 3-5%, and so the report should not proceed with their original margin of errors if they are testing at 95% confidence.