Week6-Lab-VI

Aamol Gote

10/11/2019

**Excercise 1**

In the first paragraph, several key findings are reported. Do these percentages appear to be sample statistics (derived from the data sample) or population parameters?

These percentages are the sample statistics derived from the sample data. Sample data is collected by the worldwide poll conducted by WIN-Gallup International.

**Excercise 2**

The title of the report is “Global Index of Religiosity and Atheism”. To generalize the report’s findings to the global human population, what must we assume about the sampling method? Does that seem like a reasonable assumption?

Assumptions

1. Sampling is random
2. Sample groups are independent.
3. Sample size = 51,927, we can assume sample size is less than 10% of overall world population.

Assumptions are reasonable

download.file("http://www.openintro.org/stat/data/atheism.RData", destfile = "atheism.RData")  
load("atheism.RData")

**Excercise 3**

What does each row of Table 6 correspond to? What does each row of atheism correspond to?

nrow(atheism)

## [1] 88032

Each row in table 6 corresponds to survey results for each country, for each country what are the percentages of person been Religios Not Religios Convinced Atheist Don’t Know/No Response

Each row also has the sample size for each country.

Each row in atheism data corresponds to indivisdual surveyed items, to indicate whether person is athiest or non athiest belonging to certain country along with year.

**Excercise 4**

Using the command below, create a new dataframe called us12 that contains only the rows in atheism associated with respondents to the 2012 survey from the United States. Next, calculate the proportion of atheist responses. Does it agree with the percentage in Table 6? If not, why?

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

atheistSubset <- subset(us12, response == 'atheist')  
atheistCount <- nrow(atheistSubset)  
totalCount <- nrow(us12)  
nonatheistSubset <- subset(us12, response == 'non-atheist')  
nonatheistCount <- nrow(nonatheistSubset)  
proportion <- atheistCount/totalCount  
proportion

## [1] 0.0499002

Proportion of atheist response = 0.0499002, rounding it would be 0.05, percentage wise 5%, which matches percentage of 5% in the table 6.

**Excercise 5**

Write out the conditions for inference to construct a 95% confidence interval for the proportion of atheists in the United States in 2012. Are you confident all conditions are met?

nrow(us12)

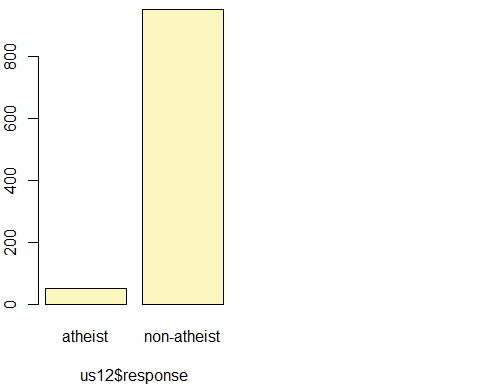
## [1] 1002

1. Samples are randomly selected
2. Observations are independent
3. Sample size of US is 1022, it is definitely less than 10% of the US population.
4. Number of Success failure conditions needs to greater than 10. np > 10 and n(1-p) > 10 p =0.05 np = 1002*0.05 = 50 which is greater than 10 n(p-1) = 1002*(1-0.05) = 952 which is greater than 10.

Based on the point we are confident that conditions are met.

inference(us12$response, est = "proportion", type = "ci", method = "theoretical", success = "atheist")

## Single proportion -- success: atheist   
## Summary statistics:



## p\_hat = 0.0499 ; n = 1002   
## Check conditions: number of successes = 50 ; number of failures = 952   
## Standard error = 0.0069   
## 95 % Confidence interval = ( 0.0364 , 0.0634 )

**Excercise 6**

Based on the R output, what is the margin of error for the estimate of the proportion of the proportion of atheists in US in 2012?

SE = 0.0069 (Based on above output) Z = 1.96 (95% Confidence interval) ME = Z \* SE

ME <- 1.96 \* 0.0069  
ME

## [1] 0.013524

ME = 0.013524

**Excercise 7**

Using the inference function, calculate confidence intervals for the proportion of atheists in 2012 in two other countries of your choice, and report the associated margins of error. Be sure to note whether the conditions for inference are met. It may be helpful to create new data sets for each of the two countries first, and then use these data sets in the inference function to construct the confidence intervals.

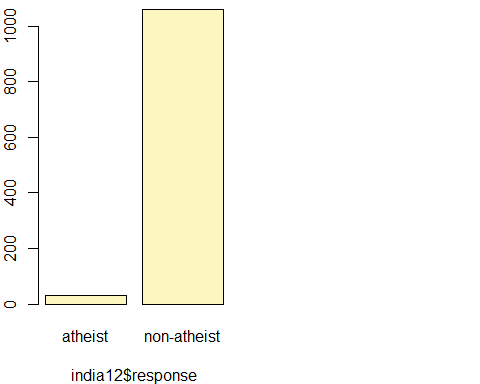
2 Countries of choice India China

india12 <- subset(atheism, nationality == "India" & year == "2012")  
china12 <- subset(atheism, nationality == "China" & year == "2012")

ME For India

inference(india12$response, est = "proportion", type = "ci", method = "theoretical", success = "atheist")

## Single proportion -- success: atheist   
## Summary statistics:



## p\_hat = 0.0302 ; n = 1092   
## Check conditions: number of successes = 33 ; number of failures = 1059   
## Standard error = 0.0052   
## 95 % Confidence interval = ( 0.0201 , 0.0404 )

India Standard error = 0.0052

meIndia <- 1.96 \* 0.0052  
meIndia

## [1] 0.010192

India Margin of Error (ME) = 0.010192

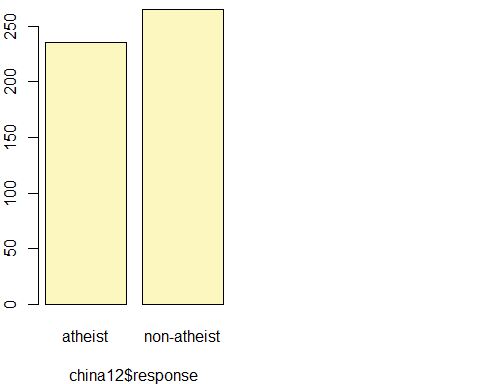
Conditions

1. Observations are independent.
2. Assumption is samples are independent.
3. Success/Failure number of observations have to greater than 10. In this case number of successes(Athiest) = 33; number of failures = 1059 (non athiest). Both values are greater than 10.
4. Sample size 1092 is definitely less than 10% of India population

ME For China

inference(china12$response, est = "proportion", type = "ci", method = "theoretical", success = "atheist")

## Single proportion -- success: atheist   
## Summary statistics:



## p\_hat = 0.47 ; n = 500   
## Check conditions: number of successes = 235 ; number of failures = 265   
## Standard error = 0.0223   
## 95 % Confidence interval = ( 0.4263 , 0.5137 )

China Standard error = 0.0223

meIndia <- 1.96 \* 0.0223   
meIndia

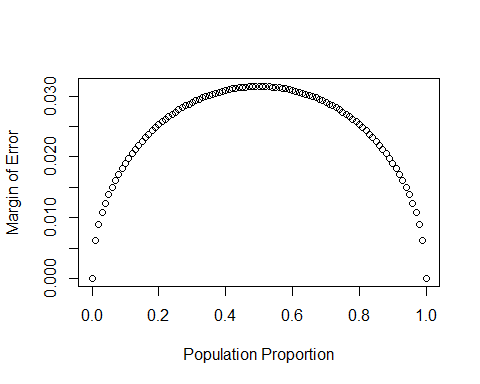
## [1] 0.043708

Conditions

1. Observations are independent.
2. Assumption is samples are independent.
3. Success/Failure number of observations have to greater than 10. In this case number of successes(Athiest) = 235; number of failures = 265 (non athiest). Both values are greater than 10.
4. Sample size 500 is definitely less than 10% of India population

China Margin of Error (ME) = 0.043708

n <- 1000  
p <- seq(0, 1, 0.01)  
me <- 2 \* sqrt(p \* (1 - p)/n)  
plot(me ~ p, ylab = "Margin of Error", xlab = "Population Proportion")

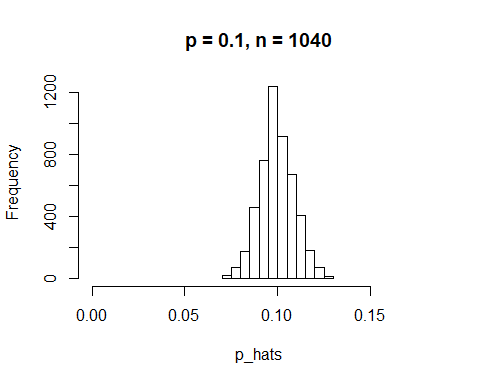


**Excercise 8**

Describe the relationship between p and me.

1. The plot follows bell curve.
2. When proportion moves away from 0.5 to 0.0 to 1.0, then margin of error decreases.
3. p and me move in opposite direction when p is between 0.5 to 1.0, which is inverse correlation.

p <- 0.1  
n <- 1040  
p\_hats <- rep(0, 5000)  
  
for(i in 1:5000){  
 samp <- sample(c("atheist", "non\_atheist"), n, replace = TRUE, prob = c(p, 1-p))  
 p\_hats[i] <- sum(samp == "atheist")/n  
}  
  
hist(p\_hats, main = "p = 0.1, n = 1040", xlim = c(0, 0.18))



mean(p\_hats)

## [1] 0.09969

sd(p\_hats)

## [1] 0.009287382

**Excercise 9**

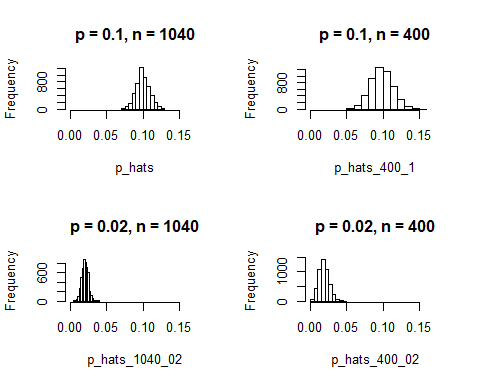
Describe the sampling distribution of sample proportions at n=1040 and p=0.1. Be sure to note the center, spread, and shape. Hint: Remember that R has functions such as mean to calculate summary statistics.

1. The sampling distribution is unimodal, symmetrical and follows normal distribution.
2. Mean is 0.09969 and Standard deviation is 0.009287382.
3. Plot is centered around 0.09969 (near to 0.01). Plot is nearly normal.

**Excercise 10**

Repeat the above simulation three more times but with modified sample sizes and proportions: for n=400 and p=0.1, n=1040 and p=0.02, and n=400 and p=0.02. Plot all four histograms together by running the par(mfrow = c(2, 2)) command before creating the histograms. You may need to expand the plot window to accommodate the larger two-by-two plot. Describe the three new sampling distributions. Based on these limited plots, how does n appear to affect the distribution of p^? How does p affect the sampling distribution?

p <- 0.1  
n <- 400  
p\_hats\_400\_1 <- rep(0, 5000)  
  
for(i in 1:5000){  
 samp <- sample(c("atheist", "non\_atheist"), n, replace = TRUE, prob = c(p, 1-p))  
 p\_hats\_400\_1[i] <- sum(samp == "atheist")/n  
}  
  
p <- 0.02  
n <- 1040  
p\_hats\_1040\_02 <- rep(0, 5000)  
  
for(i in 1:5000){  
 samp <- sample(c("atheist", "non\_atheist"), n, replace = TRUE, prob = c(p, 1-p))  
 p\_hats\_1040\_02[i] <- sum(samp == "atheist")/n  
}  
  
p <- 0.02  
n <- 400  
p\_hats\_400\_02 <- rep(0, 5000)  
  
for(i in 1:5000){  
 samp <- sample(c("atheist", "non\_atheist"), n, replace = TRUE, prob = c(p, 1-p))  
 p\_hats\_400\_02[i] <- sum(samp == "atheist")/n  
}  
  
par(mfrow = c(2, 2))  
hist(p\_hats, main = "p = 0.1, n = 1040", xlim = c(0, 0.18))  
hist(p\_hats\_400\_1, main = "p = 0.1, n = 400", xlim = c(0, 0.18))  
hist(p\_hats\_1040\_02, main = "p = 0.02, n = 1040", xlim = c(0, 0.18))  
hist(p\_hats\_400\_02, main = "p = 0.02, n = 400", xlim = c(0, 0.18))



1. Larger the n lesser the spread length
2. Distribution is mormals in all 4 cases
3. P-value changes the centre of distributions, when p-value increases, the centre of distribution towards lower side.

**Excercise 11**

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?

australia12 <- subset(atheism, nationality == "Australia" & year == "2012")  
ecuador12 <- subset(atheism, nationality == "Ecuador" & year == "2012")  
  
australia12\_aethist <- subset(australia12, response == "atheist")  
ecuador12\_aethist <- subset(ecuador12, response == "atheist")  
nrow(australia12\_aethist)

## [1] 104

nrow(ecuador12\_aethist)

## [1] 8

We can assume that sampling is random and observation results are independent, but in this case success and failure conditions of >= 10 fails as Ecuador success number of items is 8 which is less than 10. So it is not sensible to proceed with inference and report margin of errors

**On Your Own**

1. Answer the following two questions using the inference function. As always, write out the hypotheses for any tests you conduct and outline the status of the conditions for inference.
2. Is there convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012? Hint: Create a new data set for respondents from Spain. Form confidence intervals for the true proportion of athiests in both years, and determine whether they overlap.

NULL Hypothesis : There is no difference in the the aethism index between year 2005 and 2012

Ho : p2005 - p2012 = 0

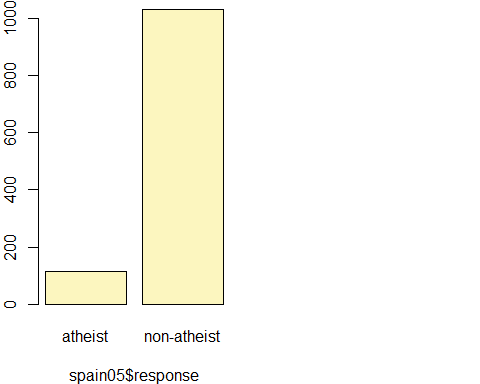
Alternative Hypothesis : There is difference in the the aethism index between year 2005 and 2012

Ha : p2005 - p2012 = 0

spain05 <- subset(atheism, nationality == "Spain" & year == "2005")  
spain12 <- subset(atheism, nationality == "Spain" & year == "2012")

inference(spain05$response, est = "proportion", type = "ci", method = "theoretical",   
 success = "atheist")

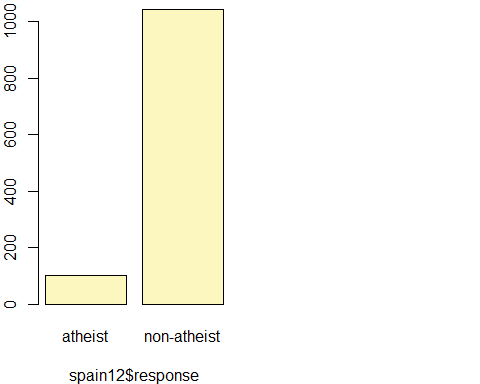
## Single proportion -- success: atheist   
## Summary statistics:



## p\_hat = 0.1003 ; n = 1146   
## Check conditions: number of successes = 115 ; number of failures = 1031   
## Standard error = 0.0089   
## 95 % Confidence interval = ( 0.083 , 0.1177 )

inference(spain12$response, est = "proportion", type = "ci", method = "theoretical",   
 success = "atheist")

## Single proportion -- success: atheist   
## Summary statistics:



## p\_hat = 0.09 ; n = 1145   
## Check conditions: number of successes = 103 ; number of failures = 1042   
## Standard error = 0.0085   
## 95 % Confidence interval = ( 0.0734 , 0.1065 )

95% Confidence Interval for Spain 2005 = (0.083, 0.1177) 95% Confidence Interval for Spain 2012 = (0.0734, 0.1065)

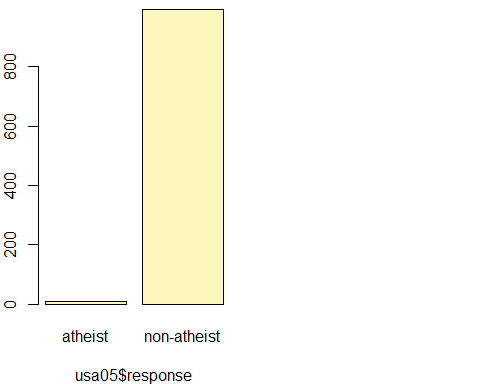
95% confidence interval overlap in for 2005 and 2012. There is good amount of overalal with very slight decrease in year 2012, which provides strong evidence that there is no difference in aethism index for year 2005 and 2012.

1. Is there convincing evidence that the United States has seen a change in its atheism index between 2005 and 2012?

usa05 <- subset(atheism, nationality == "United States" & year == "2005")  
usa12 <- subset(atheism, nationality == "United States" & year == "2012")

inference(usa05$response, est = "proportion", type = "ci", method = "theoretical", success = "atheist")

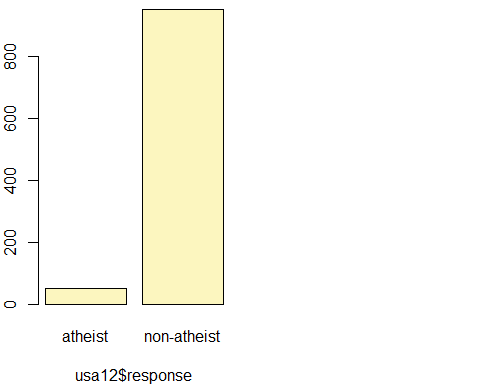
## Single proportion -- success: atheist   
## Summary statistics:



## p\_hat = 0.01 ; n = 1002   
## Check conditions: number of successes = 10 ; number of failures = 992   
## Standard error = 0.0031   
## 95 % Confidence interval = ( 0.0038 , 0.0161 )

inference(usa12$response, est = "proportion", type = "ci", method = "theoretical", success = "atheist")

## Single proportion -- success: atheist   
## Summary statistics:



## p\_hat = 0.0499 ; n = 1002   
## Check conditions: number of successes = 50 ; number of failures = 952   
## Standard error = 0.0069   
## 95 % Confidence interval = ( 0.0364 , 0.0634 )

95% Confidence Interval for USA 2005 = (0.0038, 0.0161) 95% Confidence Interval for USA 2012 = (0.0364, 0.0634)

There is no overlap so we can reject NULL hypothesis. There is convincing evidence to prove that there is difference in the aethism index between year 2005 ans 2012.

1. If in fact there has been no change in the atheism index in the countries listed in Table 4, in how many of those countries would you expect to detect a change (at a significance level of 0.05) simply by chance? Hint: Look in the textbook index under Type 1 error.

There has been no change in the atheism index in the countries listed in Table 4 which indicates NULL hypoetesis to be true. So to expect countries to detect a change (at a significance level of 0.05) simply by chance could be dues to Type Error 1. Since significance level is 0.05 which indicates 95% has no change and remaining 5% there was change by chance. Table 4 has 39 countries So countries with change by chance = 39 \* 0.05 = 1.95 = 2 Countries.

1. Suppose you’re hired by the local government to estimate the proportion of residents that attend a religious service on a weekly basis. According to the guidelines, the estimate must have a margin of error no greater than 1% with 95% confidence. You have no idea what to expect for p. How many people would you have to sample to ensure that you are within the guidelines? Hint: Refer to your plot of the relationship between p and margin of error. Do not use the data set to answer this question.

ME = z \* SE z (for 95% confidence) = 1.96 SE = SQRT(0.5\*(1-0.5)/n) ME = 0.01

The largest margin of error occurs when p = 0.5, therefore, to make a conservative estimate, we should use this proportion to ensure we will sample the necessary amount of people.

p=0.5

0.01 = 1.96 \* (SQRT(0.5\*(1-0.5)/n))

(0.01/1.96)^2 = 0.5\*0.5/n

n = 9604

Sample size has to be: **9604**