lab5

jason grahn

10/24/2018

# Inference for numerical data

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

## Observations: 1,000  
## Variables: 13  
## $ fage <int> NA, NA, 19, 21, NA, NA, 18, 17, NA, 20, 30, NA,...  
## $ mage <int> 13, 14, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16,...  
## $ mature <fct> younger mom, younger mom, younger mom, younger ...  
## $ weeks <int> 39, 42, 37, 41, 39, 38, 37, 35, 38, 37, 45, 42,...  
## $ premie <fct> full term, full term, full term, full term, ful...  
## $ visits <int> 10, 15, 11, 6, 9, 19, 12, 5, 9, 13, 9, 8, 4, 12...  
## $ marital <fct> married, married, married, married, married, ma...  
## $ gained <int> 38, 20, 38, 34, 27, 22, 76, 15, NA, 52, 28, 34,...  
## $ weight <dbl> 7.63, 7.88, 6.63, 8.00, 6.38, 5.38, 8.44, 4.69,...  
## $ lowbirthweight <fct> not low, not low, not low, not low, not low, lo...  
## $ gender <fct> male, male, female, male, female, male, male, m...  
## $ habit <fct> nonsmoker, nonsmoker, nonsmoker, nonsmoker, non...  
## $ whitemom <fct> not white, not white, white, white, not white, ...

## 1

### What are the cases in this data set? How many cases are there in our sample?

The cases in this dataset, otherwise known as observations, are births recorded in North Carolina. glimpse() shows us there are 1000 observations. We can also run summary() and review any of the categorical variables like mature. The summation of which comes to 1000.

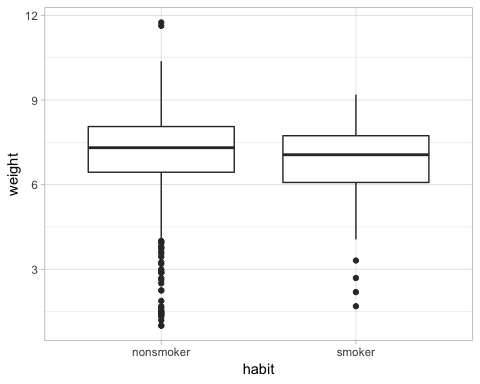
summary(nc)

## fage mage mature weeks   
## Min. :14.00 Min. :13 mature mom :133 Min. :20.00   
## 1st Qu.:25.00 1st Qu.:22 younger mom:867 1st Qu.:37.00   
## Median :30.00 Median :27 Median :39.00   
## Mean :30.26 Mean :27 Mean :38.33   
## 3rd Qu.:35.00 3rd Qu.:32 3rd Qu.:40.00   
## Max. :55.00 Max. :50 Max. :45.00   
## NA's :171 NA's :2   
## premie visits marital gained   
## full term:846 Min. : 0.0 married :386 Min. : 0.00   
## premie :152 1st Qu.:10.0 not married:613 1st Qu.:20.00   
## NA's : 2 Median :12.0 NA's : 1 Median :30.00   
## Mean :12.1 Mean :30.33   
## 3rd Qu.:15.0 3rd Qu.:38.00   
## Max. :30.0 Max. :85.00   
## NA's :9 NA's :27   
## weight lowbirthweight gender habit   
## Min. : 1.000 low :111 female:503 nonsmoker:873   
## 1st Qu.: 6.380 not low:889 male :497 smoker :126   
## Median : 7.310 NA's : 1   
## Mean : 7.101   
## 3rd Qu.: 8.060   
## Max. :11.750   
##   
## whitemom   
## not white:284   
## white :714   
## NA's : 2   
##   
##   
##   
##

## 2

### Make a side-by-side boxplot of habit and weight. What does the plot highlight about the relationship between these two variables?

nc %>%   
 #There are NAs in the data that need to be removed.  
 filter(!is.na(habit)) %>%   
 #now plot the results.  
 ggplot(aes(x = habit, y = weight)) +  
 geom\_boxplot() +  
 theme\_light()



by(nc$weight, nc$habit, mean)

## nc$habit: nonsmoker  
## [1] 7.144273  
## --------------------------------------------------------   
## nc$habit: smoker  
## [1] 6.82873

We definitely have an observed difference. While the medians are similar, they *are* different. The boxplot shows smokers have a tighter tolerance for weight, and that non-smokers have much more variation with weights.

## 3

### Check if the conditions necessary for inference are satisfied. Note that you will need to obtain sample sizes to check the conditions. You can compute the group size using the same by command above but replacing mean with length.

by(nc$weight, nc$habit, length)

## nc$habit: nonsmoker  
## [1] 873  
## --------------------------------------------------------   
## nc$habit: smoker  
## [1] 126

The conditions seem to be met: The 1,000 observations are likely less than 10% of the population to ensure the sampling is random, which fulfills independence. We have more than 30 samples, which satifies large number central limit theorem. Though distributions shown on box and whisker plots have outliers, these seem reasonable for the sample size.

## 4

### Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different.

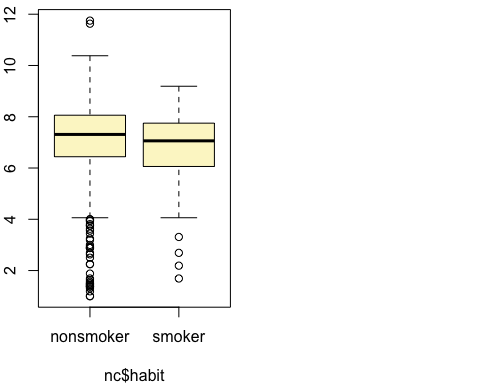
\* H0 = The average weights of babies born to smoking and non-smoking mothers are not different  
\* HA = The average weights of babies born to smoking and non-smoking mothers are different

## 5

### Change the type argument to “ci” to construct and record a confidence interval for the difference between the weights of babies born to smoking and non-smoking mothers.

#inference(y = nc$weight,   
# x = nc$habit,   
# est = "mean",   
# type = "ht",   
# null = 0,   
# alternative = "twosided",   
# method = "theoretical")  
  
inference(y = nc$weight,   
 x = nc$habit,   
 est = "mean",   
 type = "ci",   
 null = 0,   
 alternative = "twosided",   
 method = "theoretical")

## Response variable: numerical, Explanatory variable: categorical  
## Difference between two means  
## Summary statistics:  
## n\_nonsmoker = 873, mean\_nonsmoker = 7.1443, sd\_nonsmoker = 1.5187  
## n\_smoker = 126, mean\_smoker = 6.8287, sd\_smoker = 1.3862



## Observed difference between means (nonsmoker-smoker) = 0.3155  
##   
## Standard error = 0.1338   
## 95 % Confidence interval = ( 0.0534 , 0.5777 )

The 95% Confidence interval is (0.0534 , 0.5777 ) and this command also ends up making the same boxplot we made before.

# Inference for categorical data

## 1

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

The *second* paragraph states:

\* 59% of the world said that they think of themselves as religious person, 23% think of themselves as not religious whereas 13% think of themselves as convinced atheists. However, when we compare this to the Irish population, only 47% consider themselves religious, placing Ireland low on the index of being religious coming in at position 43 out of 57 countries.

It is not feasible to conduct this level of analysis with *every* individual in the world. These appear to be sample statistics.

## 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?

We must assume the this is a random sample that has a high enough *n* to account for a reasonable percentage of the global population and that each interviewee / respondant was part of a simple random selection process.

## 3

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

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

## Observations: 88,032  
## Variables: 3  
## $ nationality <fct> Afghanistan, Afghanistan, Afghanistan, Afghanistan...  
## $ response <fct> non-atheist, non-atheist, non-atheist, non-atheist...  
## $ year <int> 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 20...

#View(atheism)

The rows in Table 6 of [the article](https://sidmennt.is/wp-content/uploads/Gallup-International-um-tr%C3%BA-og-tr%C3%BAleysi-2012.pdf) is titled GLOBAL RELIGIOSITY AND ATHEISM INDEX FOR 2012. The table provides a country, the sample size of those surveyed, and the percent of respondants as they self-selected their level of religiousity.

The observations in the atheism data are a given country, the year of the survey, and the respondants selection of atheist or non-atheist. This appears to be individual line-item responses.

nat.summary <- atheism %>%   
 group\_by(nationality, response) %>%   
 summarize(count = n())  
  
head(nat.summary, 10)

## # A tibble: 10 x 3  
## # Groups: nationality [6]  
## nationality response count  
## <fct> <fct> <int>  
## 1 Afghanistan non-atheist 1031  
## 2 Argentina atheist 90  
## 3 Argentina non-atheist 1903  
## 4 Armenia atheist 10  
## 5 Armenia non-atheist 485  
## 6 Australia atheist 104  
## 7 Australia non-atheist 935  
## 8 Austria atheist 200  
## 9 Austria non-atheist 1805  
## 10 Azerbaijan non-atheist 509

## 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.

us12 <- subset(atheism, nationality == "United States" & year == "2012")  
  
#or the tidy version of same  
us12.tidy <- atheism %>%   
 filter(nationality == "United States",  
 year == "2012")

### Next, calculate the proportion of atheist responses.

us12atheists <- us12.tidy %>%   
 group\_by(response) %>%   
 summarise(response\_count = n()) %>%  
 mutate(proportion = response\_count / nrow(us12.tidy)) %>%   
 filter(response == "atheist")  
  
us12atheists

## # A tibble: 1 x 3  
## response response\_count proportion  
## <fct> <int> <dbl>  
## 1 atheist 50 0.0499

### Does it agree with the percentage in Table 6? If not, why?

Yes, this agrees with the percentage in Table 6, which shows a US atheist population of 5%. The proportion of atheists by calculation of the openintro data is 4.99%, which is close enough to 5% to make this acceptable as agreeing.

## 5

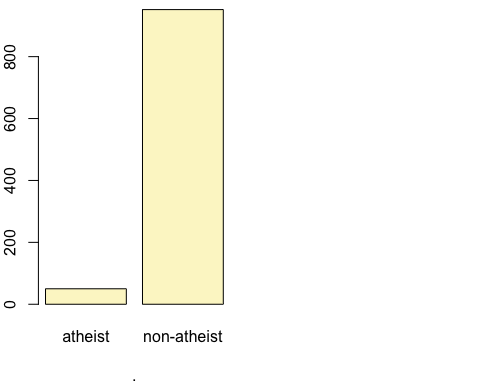
### Write out the conditions for inference to construct a 95% confidence interval for the proportion of atheists in the United States in 2012. Are you confident all conditions are met?

The conditions are:

1. The observations are independent. This condition is met because assume that the samples are randomly selected.   
2. This also requires a sample size of np >= 10 and n(1-p)>=10. This condition is met.  
 \* np = 1002\*0.05 = 50, which is >= 10.   
 \* n(1-p) = 1002\*0.95 = 952, which is also >= 10.

us12$response %>%   
inference(.,  
 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 )

#describe(us12$response)

## 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?

Assuming a 95% confidence interval:

#we'll need to make the confidence interval value for .95  
a <- 1 - 0.05/2  
ci.95 <- qnorm(a)  
  
#one way is to use the SE from above and manufacture the Margin of error  
margin.error <- ci.95\* 0.0069  
margin.error

## [1] 0.01352375

#but that's not the best for reproducibility. Let's make the code do the thinking  
p <- us12atheists$proportion  
q <- 1-p   
  
#the use P\*Q to make the standard error  
stderr <- sqrt( (p\*q) / nrow(us12.tidy) )  
#stderr  
  
#then use the confidence interval to make the margin of error:  
margerr <- ci.95 \* stderr   
margerr

## [1] 0.01348187

The margin of error is either 0.0135238 or 0.0134819 depending on how it’s calculated. Either way we can round this to 0.0135.

## 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.

fr12 <- subset(atheism, nationality == "France" & year == "2012")  
  
#how many values available from n to determine if   
fr.rows <- nrow(fr12)  
  
#test conditions  
#This also requires a sample size of np >= 10 and n(1-p)>=10 to determine if n is appropriate size  
fr\_condition\_1 <- if\_else((fr.rows\*0.05) >= 10, "Sample size is large and passes condition 1", "Sample size is not large and does not pass condition 1")   
   
fr\_condition\_2 <- if\_else((fr.rows\*0.95) >= 10, "Sample size is large and passes condition 2", "Sample size is not large and does not pass condition 2")  
  
fr\_condition\_1

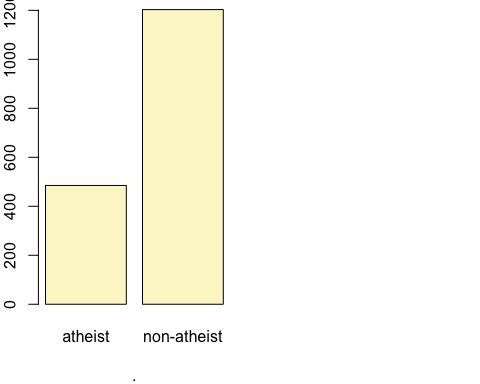
## [1] "Sample size is large and passes condition 1"

fr\_condition\_2

## [1] "Sample size is large and passes condition 2"

fr12$response %>%   
inference(.,  
 est = "proportion",   
 type = "ci",   
 method = "theoretical",   
 success = "atheist")

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



## p\_hat = 0.2873 ; n = 1688   
## Check conditions: number of successes = 485 ; number of failures = 1203   
## Standard error = 0.011   
## 95 % Confidence interval = ( 0.2657 , 0.3089 )

ch12 <- subset(atheism, nationality == "China" & year == "2012")  
  
#how many values available from n to determine if   
ch.rows <- nrow(ch12)  
  
#test conditions  
#This also requires a sample size of np >= 10 and n(1-p)>=10 to determine if n is appropriate size  
ch\_condition\_1 <- if\_else((ch.rows\*0.05) >= 10, "Sample size is large and passes condition 1", "Sample size is not large and does not pass condition 1")   
ch\_condition\_1

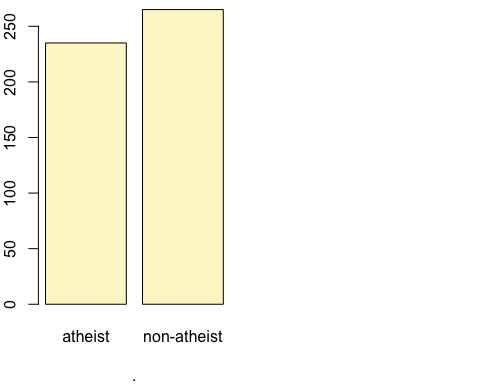
## [1] "Sample size is large and passes condition 1"

ch\_condition\_2 <- if\_else((ch.rows\*0.95) >= 10, "Sample size is large and passes condition 2", "Sample size is not large and does not pass condition 2")   
ch\_condition\_2

## [1] "Sample size is large and passes condition 2"

ch12$response %>%   
inference(.,  
 est = "proportion",   
 type = "ci",   
 method = "theoretical",   
 success = "atheist")

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

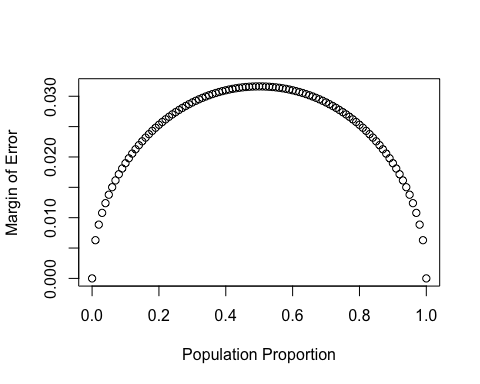


## p\_hat = 0.47 ; n = 500   
## Check conditions: number of successes = 235 ; number of failures = 265   
## Standard error = 0.0223   
## 95 % Confidence interval = ( 0.4263 , 0.5137 )

## 8

### Describe the relationship between p and me.

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")



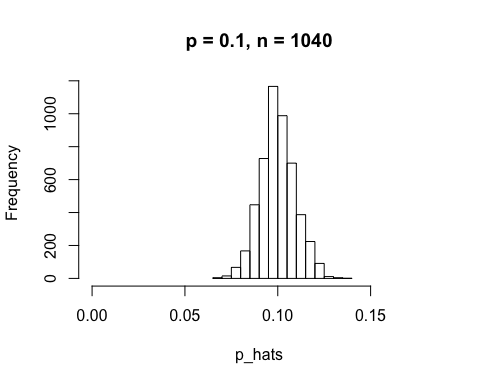
The relationship between p and me is certainly non-linear! The graph shows a proportion of 0.5 is the proportion with the peak margin of error, following a strong arc of 0 ME at 0 proportion to 0 ME and 1.0 proportion (100%). We could say there’s an inverse correlation between p and me as they move away from 0.5; but this is curvature, not straightline.

## 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.

p <- 0.1  
n <- 1040  
p\_hats <- rep(0, 5000)  
  
set.seed(30)  
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))



summary(p\_hats)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.06538 0.09423 0.10000 0.10013 0.10577 0.13654

describe(p\_hats)

## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 5000 0.1 0.01 0.1 0.1 0.01 0.07 0.14 0.07 0.08 0.05  
## se  
## X1 0

The sampling distribution is quite normal, with a mean and median of 0.1. The range is 0.07, and with a skew of 0.09 and a standard error of 0 we can be assured there is no tailing.

## 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.

#n=400 and p=0.1  
p <- 0.1  
n <- 400  
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  
 }  
describe(p\_hats)

## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 5000 0.1 0.02 0.1 0.1 0.01 0.05 0.16 0.1 0.11 -0.18  
## se  
## X1 0

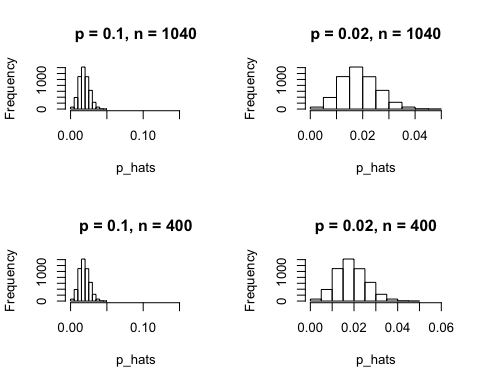
#n=1040 and p=0.02  
p <- 0.02  
n <- 1040  
p\_hats <- rep(0, 5000)  
for (i in 1:5000) {  
 samp <- sample(c("atheist", "non\_atheist"), n, replace = TRUE, prob = c(p,1 - p))  
 p\_hats[i] <- sum(samp == "atheist")/n  
 }  
describe(p\_hats)

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 5000 0.02 0 0.02 0.02 0 0.01 0.04 0.03 0.22 -0.03 0

#n=400 and p=0.02  
p <- 0.02  
n <- 400  
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  
 }  
describe(p\_hats)

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 5000 0.02 0.01 0.02 0.02 0.01 0 0.05 0.04 0.32 -0.05 0

#plotting  
par(mfrow = c(2,2))  
hist(p\_hats, main = "p = 0.1, n = 1040", xlim = c(0, 0.18))  
hist(p\_hats, main = "p = 0.02, n = 1040", xlim = c(0.0001, 0.05))  
hist(p\_hats, main = "p = 0.1, n = 400", xlim = c(0, 0.18))  
hist(p\_hats, main = "p = 0.02, n = 400", xlim = c(0.0001, 0.06))



### Describe the three new sampling distributions. Based on these limited plots, how does n appear to affect the distribution of p̂ ? How does p affect the sampling distribution?

\* The first has a matching `mean` and `median` of 0.1, with a range of 0.11 and skew of 0.12. Skewness indicates there might be a slight right tail, which the histogram confirms.  
\* The second ALSO has a matching `mean` and `median`, but this time of 0.02. The range is much smaller at 0.03, but skewness increases to 0.23. with a range of 0.11 and skew of 0.12. Skewness indicates there might be a slight right tail, providing support of a longer right tail. The histogram confirms this as well.  
\* The third have matching mean and median as well, both of 0.02. Range increases to 0.05 and skewness also increases to 0.34.

The size of n affects the shape and spread of the distribution, with larger samples providing closer estimates of the population proportion of p. The spread, also known as variability will decrease as sample size n increases. The larger n also changes the distribution shape to become more normal.

## 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?

It seems ok to move forward with the inference. Australia meets the conditions for the sampling distribution of p. Ecuador is a different story though. The data does not pass the success-failure condition as np is only 8. Eight is less than the prerequisite 10 necessary to continue.