Homework4

Aeriel Belk

10/24/2018

List data in data subdirectory

list.files('../data')

## [1] "homicide-data.csv"

## Homework Assignment

## Estimating the proportion of unsolved homicides in each city

library(readr)  
library(dplyr)  
library(ggplot2)  
library(lubridate)  
library(tidyr)  
library(broom)  
library(purrr)  
library(scales)

1. Read in the data as an R object named homicides.

homicides <- read\_csv('../data/homicide-data.csv')

1. Create a new column called city\_name that combines the city and state like this “Baltimore, MD”.

homicides <- homicides %>%   
 unite(city\_name, city, state, sep = ', ', remove = FALSE)

1. Create a dataframe called unsolved with one row per city that gives the total number of homicides for the city and the number of unsolved homicides (those for which the disposition is “Closed without arrest” or “Open/No arrest”).

onlyunsolved <- homicides %>%   
 mutate(unsolved = disposition == 'Closed without arrest' |  
 disposition == 'Open/No arrest') %>%   
 group\_by(city\_name, unsolved) %>%   
 count() %>%   
 filter(unsolved == TRUE) %>%   
 ungroup() %>%   
 select(-'unsolved') %>%   
 rename(n\_unsolved = n)  
  
unsolved <- homicides %>%   
 group\_by(city\_name) %>%   
 count() %>%   
 ungroup() %>%   
 left\_join(onlyunsolved, by = 'city\_name') %>%   
 mutate(n\_unsolved = replace(n\_unsolved, city\_name == 'Tulsa, AL', 0))  
  
head(unsolved)

## # A tibble: 6 x 3  
## city\_name n n\_unsolved  
## <chr> <int> <dbl>  
## 1 Albuquerque, NM 378 146  
## 2 Atlanta, GA 973 373  
## 3 Baltimore, MD 2827 1825  
## 4 Baton Rouge, LA 424 196  
## 5 Birmingham, AL 800 347  
## 6 Boston, MA 614 310

1. For the city of Baltimore, MD, use the prop.test function to estimate the proportion of homicides that are unsolved, as well as the 95% confidence interval for this proportion. Print the output of the prop.test directly in your RMarkdown, and then save the output of prop.test as an R object and apply the tidy function from the broom package to this object and pull the estimated proportion and confidence intervals from the resulting tidy dataframe.

baltimore <- filter(unsolved, city\_name == 'Baltimore, MD')  
  
baltimore\_prop <- prop.test(baltimore$n\_unsolved, baltimore$n)  
baltimore\_prop

##   
## 1-sample proportions test with continuity correction  
##   
## data: baltimore$n\_unsolved out of baltimore$n, null probability 0.5  
## X-squared = 239.01, df = 1, p-value < 2.2e-16  
## alternative hypothesis: true p is not equal to 0.5  
## 95 percent confidence interval:  
## 0.6275625 0.6631599  
## sample estimates:  
## p   
## 0.6455607

baltimore\_prop <- baltimore\_prop %>%   
 tidy()  
  
baltimore\_prop

## # A tibble: 1 x 8  
## estimate statistic p.value parameter conf.low conf.high method  
## <dbl> <dbl> <dbl> <int> <dbl> <dbl> <chr>   
## 1 0.646 239. 6.46e-54 1 0.628 0.663 1-sam…  
## # ... with 1 more variable: alternative <chr>

The estimated proportion is 0.646 and the estimated confidence interval is 0.628, 0.663.

1. Now use what you learned from running prop.test for one city to run prop.test for all the cities. Your goal is to create the figure shown below, where the points show the estimated proportions of unsolved homicides in each city and the horizontal lines show the estimated 95% confidence intervals.

unsolved <- unsolved %>%   
 mutate(prop\_results = map2(.x = n\_unsolved, .y = n, .f = prop.test),  
 prop\_results = map(.x = prop\_results, .f = tidy)) %>%   
 unnest(prop\_results, .drop = TRUE) %>%   
 select(city\_name, estimate, conf.low, conf.high)

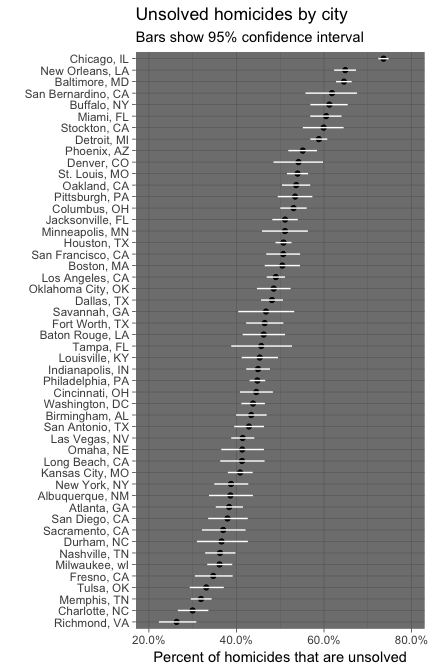
## Warning in .f(.x[[i]], .y[[i]], ...): Chi-squared approximation may be  
## incorrect

unsolved

## # A tibble: 51 x 4  
## city\_name estimate conf.low conf.high  
## <chr> <dbl> <dbl> <dbl>  
## 1 Albuquerque, NM 0.386 0.337 0.438  
## 2 Atlanta, GA 0.383 0.353 0.415  
## 3 Baltimore, MD 0.646 0.628 0.663  
## 4 Baton Rouge, LA 0.462 0.414 0.511  
## 5 Birmingham, AL 0.434 0.399 0.469  
## 6 Boston, MA 0.505 0.465 0.545  
## 7 Buffalo, NY 0.612 0.569 0.654  
## 8 Charlotte, NC 0.300 0.266 0.336  
## 9 Chicago, IL 0.736 0.724 0.747  
## 10 Cincinnati, OH 0.445 0.408 0.483  
## # ... with 41 more rows

1. Create the plot shown below.

unsolved %>%   
 filter(city\_name != 'Tulsa, AL') %>%   
 ggplot(aes(x = estimate, y = reorder(city\_name, estimate))) +  
 geom\_point() +  
 geom\_errorbarh(aes(xmin = conf.low, xmax = conf.high),   
 height = 0,  
 color = 'white') +  
 theme\_dark() +  
 labs(x = 'Percent of homicides that are unsolved',  
 y = '') +  
 scale\_x\_continuous(labels = percent,  
 limits = c(0.2, 0.8)) +  
 ggtitle('Unsolved homicides by city',  
 subtitle = 'Bars show 95% confidence interval')



1. All of the code for this should be in an RMarkdown document. Render this to a pdf and then push to your GitHub repository. Go on GitHub and make sure that everything made in online.

**Note** I can’t knit to pdf. It doesn’t work on my computer for some reason I don’t understand. So, I just knit to word, I hope you understand! Thanks!