midterm

jason grahn

10/26/2018

# Problem 1:

## N. 1-1

Download smoking.csv (or smoking.txt) and read corresponding data into R. Example command in R: MyData <- read.csv(file="path/TheDataIWantToReadIn.csv",header=TRUE, sep=",")Make sure to include the code/command.

library(readr)  
smoking <- read\_csv("smoking.csv")

## Parsed with column specification:  
## cols(  
## gender = col\_character(),  
## age = col\_integer(),  
## maritalStatus = col\_character(),  
## highestQualification = col\_character(),  
## nationality = col\_character(),  
## ethnicity = col\_character(),  
## grossIncome = col\_character(),  
## region = col\_character(),  
## smoke = col\_character(),  
## amtWeekends = col\_integer(),  
## amtWeekdays = col\_integer(),  
## type = col\_character()  
## )

#View(smoking)

## N. 1-2

How many observations are there in this data set? How many variables, and what are they?

#how many observations and variables are shown with glimpse()  
glimpse(smoking)

## Observations: 1,691  
## Variables: 12  
## $ gender <chr> "Male", "Female", "Male", "Female", "Fema...  
## $ age <int> 38, 42, 40, 40, 39, 37, 53, 44, 40, 41, 7...  
## $ maritalStatus <chr> "Divorced", "Single", "Married", "Married...  
## $ highestQualification <chr> "No Qualification", "No Qualification", "...  
## $ nationality <chr> "British", "British", "English", "English...  
## $ ethnicity <chr> "White", "White", "White", "White", "Whit...  
## $ grossIncome <chr> "2,600 to 5,200", "Under 2,600", "28,600 ...  
## $ region <chr> "The North", "The North", "The North", "T...  
## $ smoke <chr> "No", "Yes", "No", "No", "No", "No", "Yes...  
## $ amtWeekends <int> NA, 12, NA, NA, NA, NA, 6, NA, 8, 15, NA,...  
## $ amtWeekdays <int> NA, 12, NA, NA, NA, NA, 6, NA, 8, 12, NA,...  
## $ type <chr> NA, "Packets", NA, NA, NA, NA, "Packets",...

The glimpse() function shows us both the number of variables and rows. The number of rows in the data is 1691, and the number of variables is 12. They are:

\* gender  
\* age  
\* maritalStatus  
\* highestQualification  
\* nationality  
\* ethnicity  
\* grossIncome  
\* region  
\* smoke  
\* amtWeekends  
\* amtWeekdays  
\* type

#What is the 300th observation of nationality?  
smoking$nationality[[300]]

## [1] "British"

## N. 1-3

Create a numerical summary for age and compute the interquartile range.

summary(smoking$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16.00 34.00 48.00 49.84 65.50 97.00

age.iqr <- describe(smoking$age, IQR = TRUE) %>%   
 select(IQR)

The numerical summary for age is shown above. The IQR is 31.5.

Compute the relative frequency distribution for gender.

table(smoking$gender)/nrow(smoking)

##   
## Female Male   
## 0.5706682 0.4293318

How many males are in the sample? Include both the code/command and the output/graph.

male.count <- nrow(  
 smoking %>%   
 filter(gender == "Male")  
)

There are 726 males in the sample.

## N. 1-4

Using numerical summaries and a side-by-side box plot, determine if male smokers are as old as female smokers. Include both the code/command and the output/graph.

s.g <- smoking %>%   
 filter(smoke == "Yes") %>%   
 select(age, gender)  
  
#group the male data together  
s.g.m <- s.g %>% filter(gender == "Male")  
  
#group the female data together  
s.g.f <- s.g %>% filter(gender == "Female")

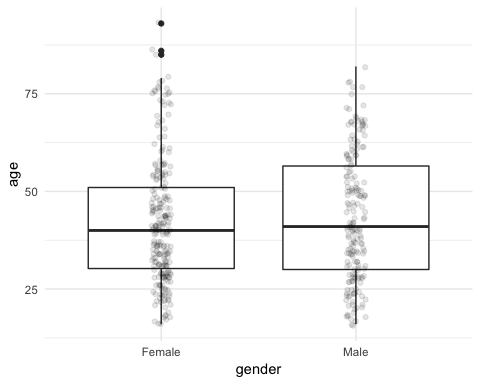
#age summary for male smokers  
summary(s.g.m$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16.00 30.00 41.00 43.35 56.50 82.00

#age summary for female smokers  
summary(s.g.f$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16.00 30.25 40.00 42.21 51.00 93.00

s.g %>%   
 ggplot(aes(x = gender, y = age)) +   
 geom\_boxplot() +  
 geom\_jitter(alpha = 0.1, width = 0.05) +  
 theme\_minimal()



## N. 1-5

Create a bar chart or frequency table for maritalStatus, what is the proportion for Divorced, Single, Married, and Widowed, respectively? What can you interpret from these numbers? Include both the code/command and the output/graph.

table(smoking$maritalStatus)/nrow(smoking)

##   
## Divorced Married Separated Single Widowed   
## 0.09520993 0.48018924 0.04021289 0.25251331 0.13187463

We can interpret from these numbers that most people in the smokers data frame are married. Single is next, then widowed.

# Problem 2

# N. 2-1

Suppose we’re flipping an unfair coin that we know only lands heads 30% of the time. Please simulate this flip 10 times, what is the proportion of heads?

outcomes.coin <- c("heads", "tails")  
  
#build a dataset of an UNFAIR coin toss  
sim\_unfair\_coin <- sample(outcomes.coin, size = 10, replace = TRUE, prob = c(0.3, 0.7))  
  
unfair.tidy <- tibble(sim\_unfair\_coin)  
  
prop.heads <- unfair.tidy %>%   
 group\_by(sim\_unfair\_coin) %>%   
 summarise(coin\_count = n()) %>%  
 mutate(freq = coin\_count / nrow(unfair.tidy)) %>%   
 filter(sim\_unfair\_coin == "heads") %>%   
 select(freq)

The proportion of heads is 0.4.

If you simulate this flip 100 times, what is the proportion of heads now? Include both the code/command and the output/graph.

sim\_unfair\_coin.large <- sample(outcomes.coin, size = 100, replace = TRUE, prob = c(0.3, 0.7))  
  
unfair.tidy.large <- tibble(sim\_unfair\_coin)  
  
prop.heads.large <- unfair.tidy.large %>%   
 group\_by(sim\_unfair\_coin) %>%   
 summarise(coin\_count = n()) %>%  
 mutate(freq = coin\_count / nrow(unfair.tidy)) %>%   
 filter(sim\_unfair\_coin == "heads") %>%   
 select(freq)

The proportion of heads is 0.4.

## N.2-2

Suppose we’re flipping an unfair dice and the corresponding probability of landing 1, 2, 3, 4, 5, and 6 is 0.05, 0.1, 0.15, 0.2, 0.3, and 0.2, respectively. If you simulate this flip 10 times, what is the proportion of land on side 5?

outcomes.dice <- c(1, 2, 3, 4, 5, 6)  
  
#build a dataset of an UNFAIR coin toss  
sim.unfair.dice <- sample(outcomes.dice, size = 10, replace = TRUE, prob = c(0.05, 0.1, 0.15, 0.2, 0.3, 0.2))  
  
unfair.dice.tidy <- tibble(sim.unfair.dice)  
  
prop.5 <- unfair.dice.tidy %>%   
 group\_by(sim.unfair.dice) %>%   
 summarise(dice.count = n()) %>%  
 mutate(freq = dice.count / nrow(unfair.dice.tidy)) %>%   
 filter(sim.unfair.dice == 5) %>%   
 select(freq)

The proportion of landing on side 5 is 0.1.

Simulate this flip 100 times, what is the proportion of side 5 now? Include both the code/command and the output/graph.

#build a dataset of an UNFAIR coin toss  
sim.unfair.dice.large <- sample(outcomes.dice, size = 100, replace = TRUE, prob = c(0.05, 0.1, 0.15, 0.2, 0.3, 0.2))  
  
unfair.dice.tidy.large <- tibble(sim.unfair.dice.large)  
  
prop.5.large <- unfair.dice.tidy.large %>%   
 group\_by(sim.unfair.dice.large) %>%   
 summarise(dice.count = n()) %>%  
 mutate(freq = dice.count / nrow(unfair.dice.tidy.large)) %>%   
 filter(sim.unfair.dice.large == 5) %>%   
 select(freq)

The proportion of landing on side 5 is 0.35.

## N. 2-3

Compare the proportions in each questions above, what conclusion can you draw? Does the number of simulations affect the proportions? If so, how? Please explain in details.

Running the simulation only 10 times, we have a situation where we might not land on a give value *at all*. However, with running more simulations, we get ever closer to the probabilities provided for the cheater dice. Even with a cheater dice, this is due to central limit theorem. With the cheater dice, we’ve set the conditions, from there, the dice rolls are independent in that rolling once has no conditional bearing on what happens afterward, AND we have a sufficiently high number of dice rolls… though there is at least some diminished independece as we HAVE set the conditional parameters.

# Problem 3:

## N. 3-1

Download countyComplete.csv (or countyComplete.txt) and read corresponding data into R. Example command in R: MyData <- read.csv(file="path/TheDataIWantToReadIn.csv", header=TRUE, sep=","). Include the code/command.

countyComplete <- read\_csv("countyComplete.csv")

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## state = col\_character(),  
## name = col\_character(),  
## FIPS = col\_integer(),  
## pop2010 = col\_integer(),  
## pop2000 = col\_integer(),  
## veterans = col\_integer(),  
## housing\_units = col\_integer(),  
## households = col\_integer(),  
## per\_capita\_income = col\_integer(),  
## median\_household\_income = col\_integer(),  
## private\_nonfarm\_establishments = col\_integer(),  
## private\_nonfarm\_employment = col\_integer(),  
## nonemployment\_establishments = col\_integer(),  
## firms = col\_integer(),  
## manufacturer\_shipments\_2007 = col\_integer(),  
## mercent\_whole\_sales\_2007 = col\_integer(),  
## sales = col\_integer(),  
## sales\_per\_capita = col\_integer(),  
## accommodation\_food\_service = col\_integer(),  
## building\_permits = col\_integer()  
## # ... with 1 more columns  
## )

## See spec(...) for full column specifications.

#glimpse(countyComplete)

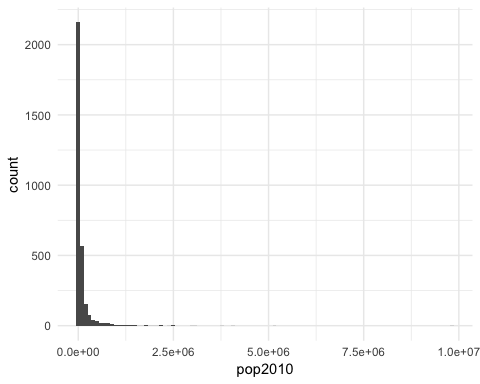
## N. 3-2

Make a histogram of pop2010, how can you descript its distribution, bellshaped or normal? Is it right skewed or left sewed? Include both the code/command and the output/graph.

countyComplete %>%   
 select(pop2000, pop2010) %>%   
 psych::describe() %>%   
 select(skew)

## skew  
## pop2000 15.51  
## pop2010 14.34

countyComplete %>%   
 select(pop2010) %>%   
 ggplot() +  
 geom\_histogram(aes(x = pop2010), bins = 100) +   
 theme\_minimal()



The distribution of pop2010 definitely has a strong right skew. Based on this, I would NOT treat the population data as “normal”. When we evaluate the skew of pop2010, we also see it has a value of 14.34 (down from the skewness of pop2000). This is a clear indicator toward the right skewness of the data.

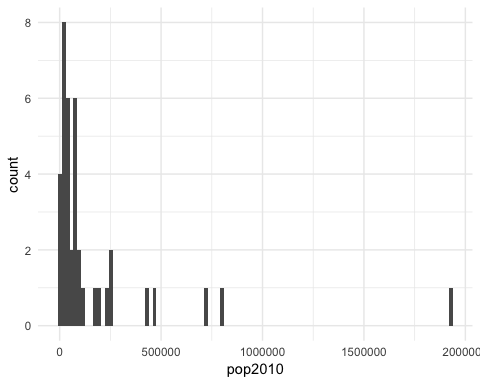
## N. 3-3

Create a new subset named Washington which contains only the observations of Washington, and then make a histogram of pop2010, how can you describe its distribution, bell-shaped or normal? Is it right skewed or left sewed? Compare this with question 1. Include both the code/command and the output/graph.

wa.subset <- countyComplete %>%   
 filter(state == "Washington")  
  
wa.subset %>%   
 select(pop2000, pop2010) %>%   
 psych::describe() %>%   
 select(skew)

## skew  
## pop2000 3.81  
## pop2010 3.71

wa.subset %>%   
 ggplot() +  
 geom\_histogram(aes(x = pop2010), bins = 100) +   
 theme\_minimal()

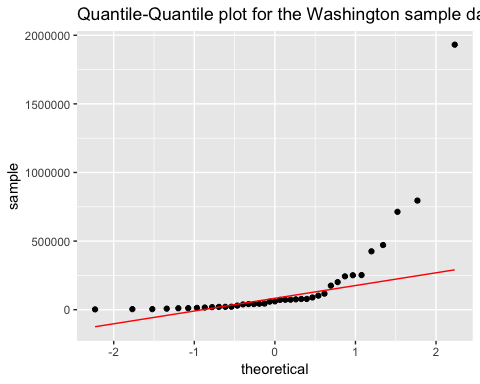


The 2010 data from Washington is much less skewed than the entire countyComplete data, with a skewness of 3.71. We still see a right tail, but it is much less pronounced.

## N. 3-4

Based on the subset Washington, make a normal probability plot of pop2010.

wa.subset %>%   
 ggplot(aes(sample = pop2010)) +   
 stat\_qq() +  
 stat\_qq\_line(color = "red") +  
 labs(title = "Quantile-Quantile plot for the Washington sample data")

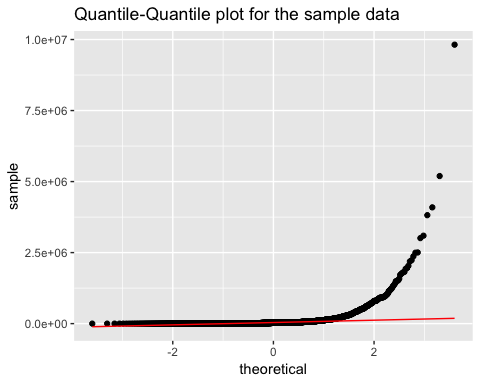


Do all of the points fall on the line?

Most defininitely do many points *not* fall on the line.

How does this plot compare to the probability plot of the original data? Include both the code/command and the output/graph.

countyComplete %>%   
 #filter(state != "Washington") %>%   
 ggplot(aes(sample = pop2010)) +  
 stat\_qq() +  
 #scale\_y\_log10() + if I were going to model this, it'd have to be in log scale.   
 stat\_qq\_line(color = "red") +  
 labs(title = "Quantile-Quantile plot for the sample data")



When we evaluate the qqplot for Washington against the entire population, the charts show similar behaviors. Each provides indications of long right tails with their rapid appearance of growth on the far right. The numbers are too large to see, but I’m willing to bet the left side of the entire population is slightly curved upward, similar to Washington.

## N. 3-5

Suppose the variable pop2010 has a normal distribution, what is the probability that pop2010 is greater than 102,410?

county.pop <- countyComplete$pop2010  
  
pnorm(q = 102410, mean = mean(county.pop), sd = sd(county.pop))

## [1] 0.5053257

Assuming normalicy, the probability pop2010 is greater than 102,410 is 0.5053257.

What is the probability that pop2010 is between 190,000 and 1,000,000? Include both the code/command and the output/graph.

(oneninety.thou <- 1- pnorm(q = 190000, mean = mean(county.pop), sd = sd(county.pop)))

## [1] 0.3846546

(one.mil <- pnorm(q = 1000000, mean = mean(county.pop), sd = sd(county.pop)))

## [1] 0.9980239

one.mil - oneninety.thou

## [1] 0.6133693

The probability that pop2010 is between 190,000 and 1,000,000 is 0.6133693.

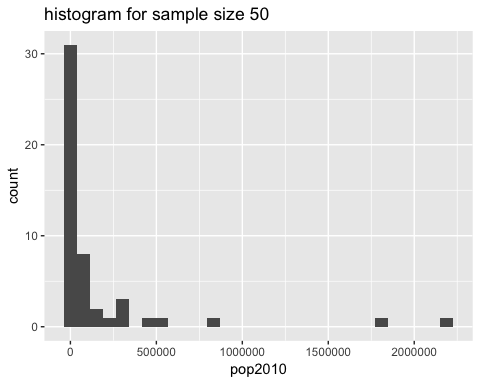
# Problem 4

## N. 4-1

Collect a simple random sample of size 50 from pop2010 in data set countyComplete.csv (or countyComplete.txt). Describe the distribution of this sample. How does it compare to the distribution of the population? Using this sample, what is your best point estimate of the population mean? Include both the code/command and the output/graph.

sample.county <- countyComplete %>%   
 select(pop2010) %>%   
 rep\_sample\_n(size = 50, reps = 1)   
  
sample.county.histogram <- sample.county %>%   
 ggplot() +  
 geom\_histogram(aes(pop2010)) +  
 labs(title = "histogram for sample size 50")  
  
sample.stats <- sample.county %>%   
 describe()  
  
sample.county.histogram

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



sample.stats

## vars n mean sd median trimmed mad min max  
## replicate 1 50 1.0 0.0 1 1 0.00 1 1  
## pop2010 2 50 163024.7 408525.4 23787 59528 28609.73 539 2189641  
## range skew kurtosis se  
## replicate 0 NaN NaN 0.00  
## pop2010 2189102 3.75 14.19 57774.21

population.stats <- countyComplete %>%   
 select(state, pop2010) %>%   
 describe()

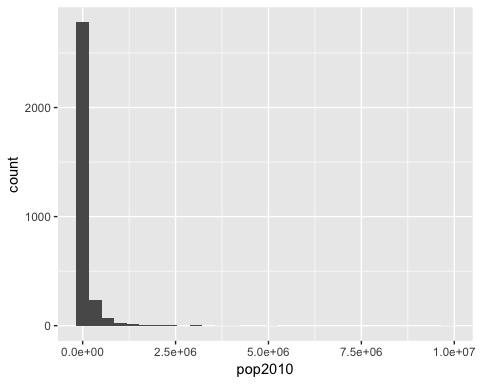
## Warning in describe(.): NAs introduced by coercion

## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning  
## Inf

## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf

population.histogram <- countyComplete %>%   
 ggplot() +   
 geom\_histogram(aes(x = pop2010))  
  
population.histogram

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



population.stats

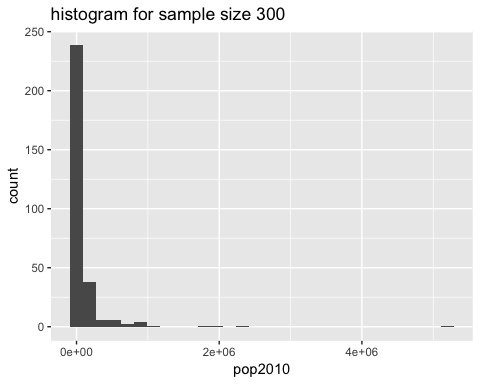
## vars n mean sd median trimmed mad min max  
## state\* 1 3143 NaN NA NA NaN NA Inf -Inf  
## pop2010 2 3143 98232.75 312901.2 25857 41699.34 27404.38 82 9818605  
## range skew kurtosis se  
## state\* -Inf NA NA NA  
## pop2010 9818523 14.34 344.64 5581.3

The distribution of the sample has a mean of 125,291 with a median 22,695 so already we can tell this is a right skewed sample. When we look at the variable skew, we see it has a value 6.04 which reinforces the right bias of the variable. The spread is over 1 million. The sample would indicate the population mean is 108741.

1. Now, collect a simple random sample of size 300 from pop2010 in data set countyComplete.csv (or countyComplete.txt). Describe the distribution of this sample. How does it compare to the distribution of the population? Using this sample, what is your best point estimate of the population mean? Compare your point estimates from the question above (question 1), what conclusion can you draw? Include both the code/command and the output/graph.

sample.county.large <- countyComplete %>%   
 select(pop2010) %>%   
 rep\_sample\_n(size = 300, reps = 1)   
  
sample.county.large.histogram <- sample.county.large %>%   
 ggplot() +  
 geom\_histogram(aes(pop2010)) +  
 labs(title = "histogram for sample size 300")   
 #this histogram looks much more normal with a log-scale X.   
 #scale\_x\_log10()   
  
sample.county.large.histogram

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



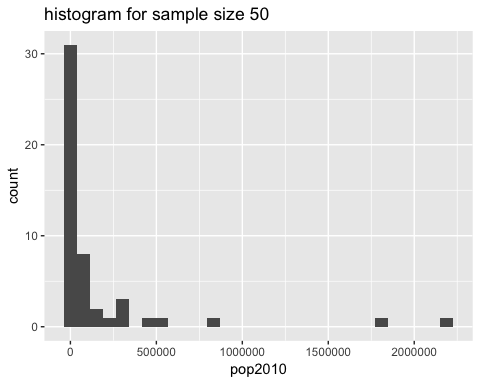
sample.stats.large <- sample.county.large %>%   
 describe()

The question of distribution of this dataset is starting to feel quite redundant as the data isn’t normal to begin with.. The distribution of the sample has a mean of 125,291 with a median 22,695 so already we can tell this is a right skewed sample. When we look at the variable skew, we see it has a value 6.04 which reinforces the right bias of the variable. The spread is over 1 million. The sample would indicate the population mean is 99847, where the true mean is 98233. The larger sample better approximates the true mean.

1. Create and compare histograms of sample size 50 and 300, which one is closer to symmetry? Why or why not? Include both the code/command and the output/graph.

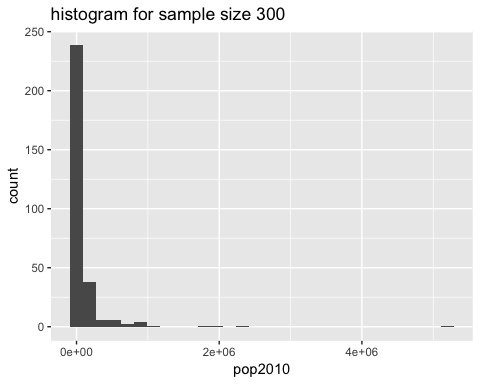
#We drew the histograms already, so we only need to call them back again.  
sample.county.histogram

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



sample.county.large.histogram

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



What we remember from the population.histogram is that the bin for the lowest value was somewhere in the thousands. Both sample sizes have relatively similar shaped distributions to each other and to the population with extremely long right tails and zero symmetry. Neither of these are symmetrical samples and neither will *ever* be symmetrical samples unless we utilize a log-scale for the X axis.