PH142 Data Project - The Effect of Population Density on Rates of Chlamydia in California Counties (2003)

Madeleine Wang, Tiffany Chung, Ahmed Amorsi, Afroze Khan, Rayni Wells

October-December 2021

```
# importing necessary libraries
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(readxl)
```

1. [2 marks] The first part of our PPDAC framework is to identify the problem you are addressing with these data. State the question you are trying to answer and let us know what type of question this is in terms of the PPDAC framework. A question statement should be as specific as possible. For example: Do students who regularly get 8 hours of sleep have fewer visits to the health center? This question is an example of an etiologic or causal question.

How does the population density of each county affect the rate of Chlamydia in 2003 per sex?

2. [2 marks] Why is this question interesting or important? You could talk here about how existing data/studies suggest this might be important, how the findings might make an impact, how the findings might be used, or why you are personally interested in this question.

This question is interesting because Chlamydia is one of the most common STIs in the United States and it would be interesting to investigate the different determinants that affect prevalence of Chlamydia. This can inform allocation of funding to Chlamydia prevention to certain areas/counties.

3. [2 marks] What is the target population for your project? Why was this target chosen? (i.e., what was your rationale for wanting to answer this question in this specific population?)

The target population is males and females in California that may be at risk of Chlamydia. We are looking at this population because we are looking to inform this population and counties that may have high Chlamydia rates on how to combat it and potential causes of higher Chlamydia rates.

4. [2 marks] What is the sampling frame used to collect the data you are using? It may be helpful here to read any protocol papers, trial registration records, '.Readme' files or documentation that are associated with your dataset. If you have trouble identifying how the records/individuals were sampled, confirm with your supporting GSI that your dataset will be usable for the purposes of the class. Describe why you think this sampling strategy is appropriate for your question. To what group(s) would you feel comfortable generalizing the findings of your study and why?

This data comes from cases with an estimated diagnosis date from 2001 to the year we selected (2003) from California Confidential Morbidity Reports and/or Laboratory reports that were submitted to CDPH. This sampling strategy is apporpriate for our question because we needed the exact number of cases of Chlamydia in each county and this data set receives that data from hospital records. Male and female Californians and perhaps generalize further to people in the United States because population density is a general variable that can determine how compact certain areas are and how Chlamydia can be spread in crowded environments.

5. [2 marks] Write a brief description (1-4 sentences) of the source and contents of your dataset. Provide a URL to the original data source if applicable. If not (e.g., the data came from your internship), provide 1-2 sentences saying where the data came from. If you completed a web form to access the data and selected a subset, describe these steps (including any options you selected) and the date you accessed the data.

URL: https://data.chhs.ca.gov/dataset/stds-in-california-by-disease-county-year-and-sex/resource/563ba92b-8ac5-48ec-9afd-2f515bbbad66?filters=Disease%3AChlamydia%7CSex%3AMale

We received the data from the California Health and Human Services Open Data Portal. The data contains information such as the disease, year, county, number of cases, etc. We selected a subset of this data by filtering for a specific disease - Chlamydia.

6. [1 mark] Write code below to import your data into R. Assign your dataset to an object. Make sure to include and annotate this code in your submission (you can use a # to comment out regular text within code chunks to annotate).

```
# reading in STD dataset
std_data <- read_csv('stds-by-disease-county-year-sex.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Disease = col_character(),
##   County = col_character(),
##   Year = col_double(),
##   Sex = col_character(),
##   Cases = col_double(),
##   Population = col_double(),
##   Rate = col_double(),
##   `Lower 95% CI` = col_double(),
##   `Upper 95% CI` = col_double(),
##   `Annotation Code` = col_character()
## )
```

7. [3 marks] Write code in R (included in your submission with annotation) to answer the following questions:

```
# What are the dimensions of the dataset?
dim(std_data)
```

```
## [1] 9558   10
```

```
head(std_data)
```

```
## # A tibble: 6 x 10
##   Disease   County      Year Sex     Cases Population  Rate `Lower 95% CI`
##   <chr>     <chr>      <dbl> <chr>   <dbl>      <dbl> <dbl>          <dbl>
## 1 Chlamydia California  2001 Female  75941   17339700  438            435.
## 2 Chlamydia California  2001 Male    24885   17173042  145.           143.
## 3 Chlamydia California  2001 Total  101590   34512742  294.           293.
## 4 Chlamydia California  2002 Female  81583   17554666  465.           462.
## 5 Chlamydia California  2002 Male    28521   17383624  164.           162.
## 6 Chlamydia California  2002 Total  110759   34938290  317            315.
## # ... with 2 more variables: Upper 95% CI <dbl>, Annotation Code <chr>
```

```
# What are the variable names of the variables in your dataset?
names(std_data)
```

```
##  [1] "Disease"         "County"          "Year"            "Sex"
##  [5] "Cases"           "Population"      "Rate"            "Lower 95% CI"
##  [9] "Upper 95% CI"    "Annotation Code"
```

```
# Print the first 6 rows of the dataset
head(std_data)
```

```
## # A tibble: 6 x 10
##   Disease   County      Year Sex     Cases Population  Rate `Lower 95% CI`
##   <chr>     <chr>      <dbl> <chr>   <dbl>      <dbl> <dbl>          <dbl>
## 1 Chlamydia California  2001 Female  75941   17339700  438            435.
## 2 Chlamydia California  2001 Male    24885   17173042  145.           143.
## 3 Chlamydia California  2001 Total  101590   34512742  294.           293.
## 4 Chlamydia California  2002 Female  81583   17554666  465.           462.
## 5 Chlamydia California  2002 Male    28521   17383624  164.           162.
## 6 Chlamydia California  2002 Total  110759   34938290  317            315.
## # ... with 2 more variables: Upper 95% CI <dbl>, Annotation Code <chr>
```

8. [2 marks] Use the data to demonstrate a data visualization skill we have covered during Part I of the course. Choose a visualization relevant to your stated problem. Include your code in your submission. For example, you could visualize the distribution of our outcome with a histogram, or use a bar graph to represent the distribution of your exposure variable.

```
# Reading in square mileage per county dataset (specifically in 2010 but square mileage of counties has
# source: https://www.indexmundi.com/facts/united-states/quick-facts/california/land-area#table
county_data <- read_xlsx('county-sq-mileage-official.xlsx', col_types = c("text", "numeric", "skip"))
```

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting numeric in B46 / R46C2: got '3775.40 '
```

```
county_data <- county_data %>% rename('County' = 'county', 'sqMileage' = 'sq-mileage')
county_data <- county_data %>% select('County', 'sqMileage')
str(county_data)
```

```
## tibble [58 x 2] (S3: tbl_df/tbl/data.frame)
##  $ County   : chr [1:58] "Alameda" "Alpine" "Amador" "Butte" ...
##  $ sqMileage: num [1:58] 739 738 595 1636 1020 ...
```

Chlamydia Cases in 2003 per County w/ Square Mileage of County

```
# cleaning data set to get desired fields
std_data_cleaned <- std_data %>%
  filter(Year == 2003, Disease == 'Chlamydia') %>%
  select(-'Lower 95% CI', -'Upper 95% CI', -'Annotation Code')
std_data_cleaned
```

```
## # A tibble: 177 x 7
##    Disease   County      Year Sex     Cases Population  Rate
##    <chr>     <chr>      <dbl> <chr>   <dbl>      <dbl> <dbl>
##  1 Chlamydia California  2003 Female  85153   17782868  479.
##  2 Chlamydia California  2003 Male    31007   17606060  176.
##  3 Chlamydia California  2003 Total  116385   35388928  329.
##  4 Chlamydia Alameda     2003 Female   3780     747441  506.
##  5 Chlamydia Alameda     2003 Male     1143     719746  159.
##  6 Chlamydia Alameda     2003 Total    4928    1467187  336.
##  7 Chlamydia Alpine      2003 Female     NA        596   NA
##  8 Chlamydia Alpine      2003 Male       NA        653   NA
##  9 Chlamydia Alpine      2003 Total       3       1249  240.
## 10 Chlamydia Amador      2003 Female     NA      16572   NA
## # ... with 167 more rows
```

```
std_county_num <- std_data_cleaned %>% filter(Sex == "Total")

# merge data sets
std_data_final <- merge(std_data_cleaned, county_data,by = 'County')
str(std_data_final)
```

```
## 'data.frame':    174 obs. of  8 variables:
##  $ County    : chr  "Alameda" "Alameda" "Alameda" "Alpine" ...
##  $ Disease   : chr  "Chlamydia" "Chlamydia" "Chlamydia" "Chlamydia" ...
##  $ Year      : num  2003 2003 2003 2003 2003 ...
##  $ Sex       : chr  "Female" "Male" "Total" "Female" ...
##  $ Cases     : num  3780 1143 4928 NA NA ...
##  $ Population: num  747441 719746 1467187 596 653 ...
##  $ Rate      : num  506 159 336 NA NA ...
##  $ sqMileage : num  739 739 739 738 738 ...
```

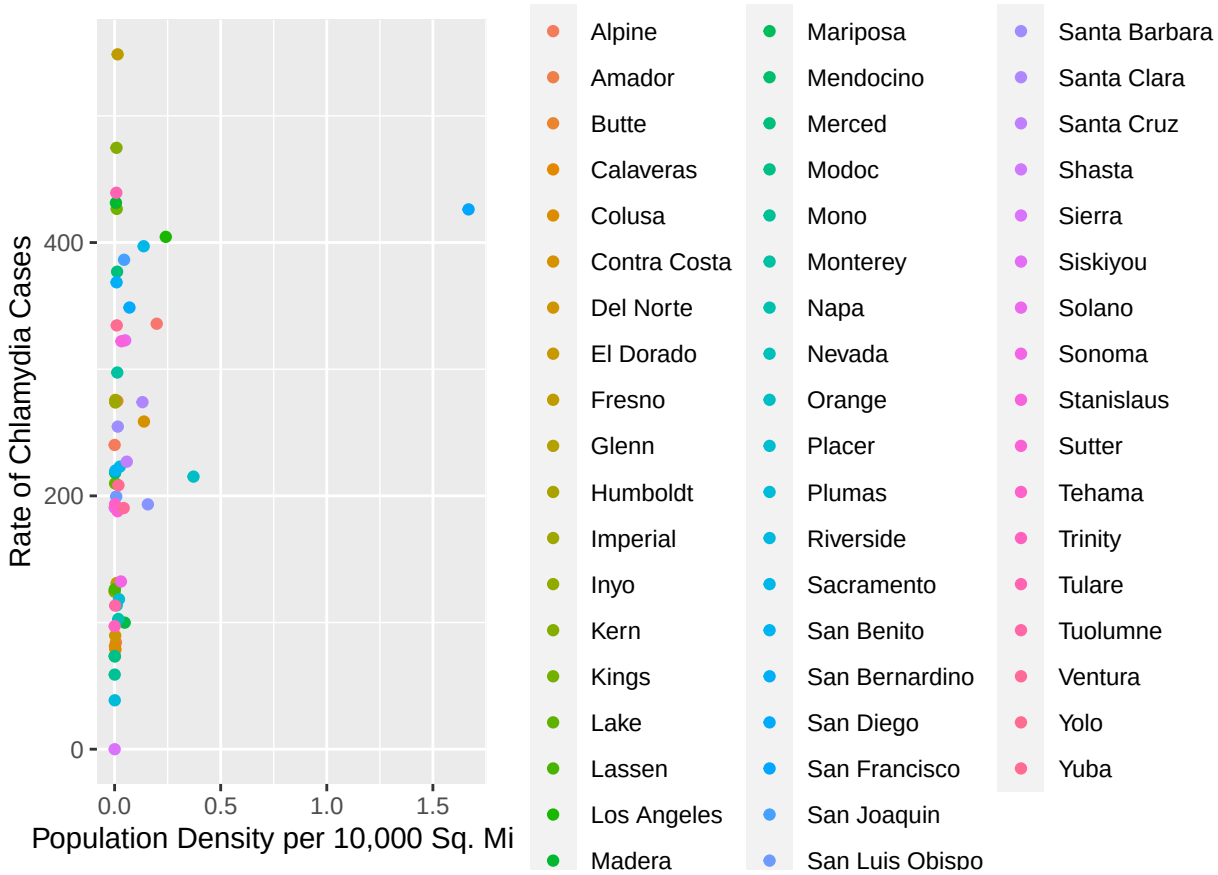
```
# population density mutation
std_data_final <- std_data_final %>%
  mutate(populationDensity = Population/sqMileage)

# creating variable of population density per 100 square miles to make visualization
# easier to read
std_data_final <- std_data_final %>%
  mutate(populationDensityper10000 = populationDensity/10000) %>%
  filter(Sex == "Total")
head(std_data_final)
```

```
##      County   Disease Year   Sex Cases Population  Rate sqMileage
## 1   Alameda Chlamydia 2003 Total  4928    1467187 335.9    739.02
## 2    Alpine Chlamydia 2003 Total     3       1249 240.2    738.33
## 3    Amador Chlamydia 2003 Total    31      36776  84.3    594.58
## 4     Butte Chlamydia 2003 Total   579     210623 274.9   1636.46
## 5 Calaveras Chlamydia 2003 Total    34      43186  78.7   1020.01
## 6    Colusa Chlamydia 2003 Total    16      19685  81.3   1150.73
##   populationDensity populationDensityper10000
## 1       1985.314335              0.1985314335
## 2          1.691655              0.0001691655
## 3         61.852064              0.0061852064
## 4        128.706476              0.0128706476
## 5         42.338801              0.0042338801
## 6         17.106532              0.0017106532
```

Population Density per County vs. Rate of Chlamydia Cases in California in 2003

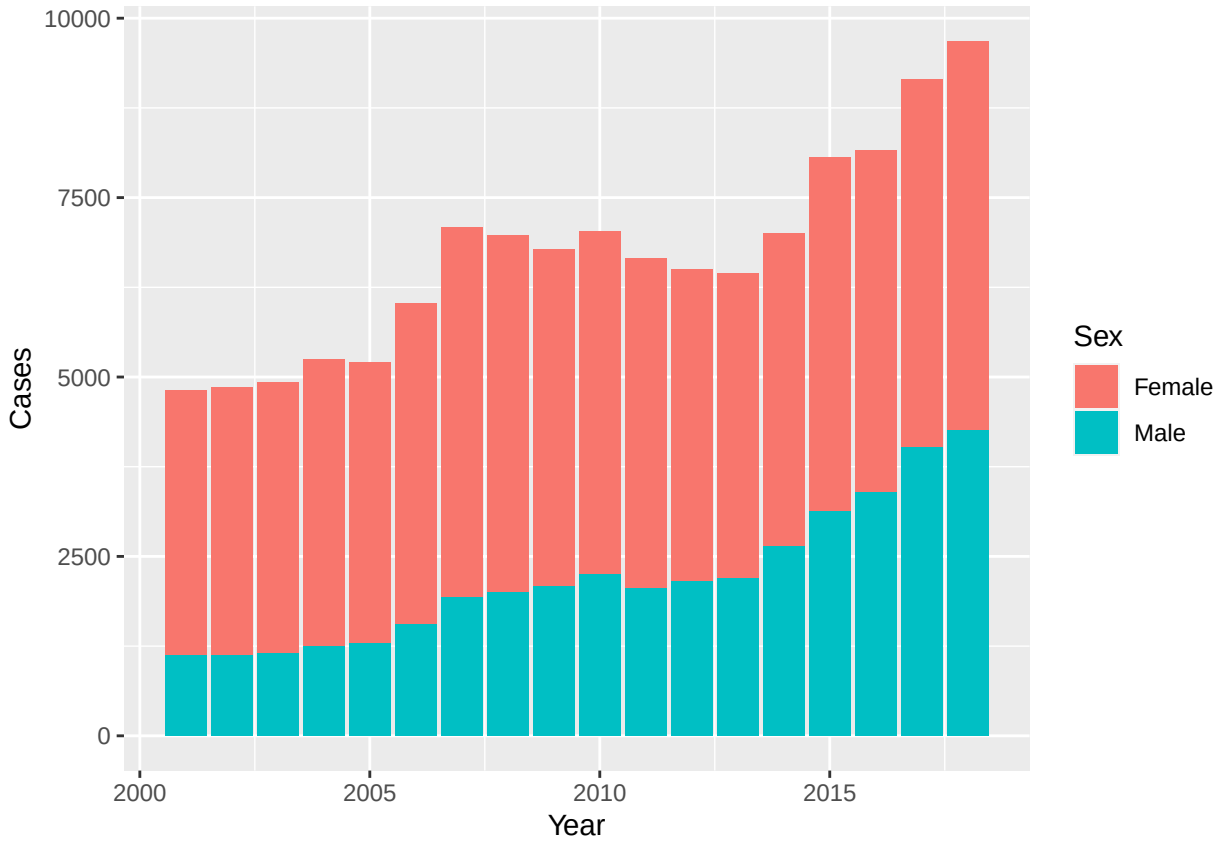
```
std_data_scatter <-ggplot(std_data_final, aes(x = populationDensityper10000, y = Rate)) +
  geom_point(aes(col = County)) +
  labs(y = "Rate of Chlamydia Cases",
       x = "Population Density per 10,000 Sq. Miles")
std_data_scatter
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```




```
# removed values come from not having any cases of Chlamydia in that specific
# county
```

```
# breakdown of Alameda County's rate of chlamydia per sex
chlamydiadf <- filter(std_data, Disease == "Chlamydia")
alamedadf <- filter(chlamydiadf, County == "Alameda")
finaldf <- filter(alamedadf, Sex == "Male" | Sex == "Female")
bar1 <- ggplot(data=finaldf, aes(x = Year, y= Cases)) +geom_bar(aes(fill = Sex), stat="identity")
bar1
```


10000
7500
5000
2500
0
Cases
2000
2005
2010
2015
Year
Sex
Female
Male

9. [2 marks] Describe the skill that you are demonstrating and interpret your findings. For example, if you have created a histogram, describe the central tendency, shape of the distribution, etc.

We created a scatter plot to demonstrate the population density versus the rate of Chlamydia cases in each county. This distribution shows a moderate, positive, linear association between the population density and rate of chlamydia cases. There are a few outliers in the data that affect the strength, such as the one with a population density $>$ 1.5 people per 1000 square mile. We also created a bar plot to see the number of cases in a single county, in this case Alameda county, over time and also separating the cases by sex using fill. It seems like the bar plot there was an increase in cases 2005-2007 before there was a plateau and slight decrease in cases from 2007-2015. After 2015, there was a steep increase in cases again. The proportion of female to male cases seems to roughly stay consistent over time.

10. [1 mark] Include your work for Part I.

11. [2 marks] Calculate a marginal probability based on your outcome variable. Provide an equation (using probability notation) that describes this probability. For example, if my outcome variable is height in inches, I might calculate the probability that an individual in the dataset has a height of greater than 60 inches. P(height $\geq$ 60) =?. This would be a marginal probability. You may need to first add a new variable to your dataset to calculate your probability of interest, such as a binary variable indicating whether height is greater than 60 inches. There is a resource video about how to code such variables that could be helpful!

P(rate of Chlamydia > average rate of chlamydia) = 0.4310345 = 43.10% The marginal probability of the rate of chlamydia being greater than the average rate of chlamydia among all states is approximately 43.10%.

```
# Added column to indicate whether the rate is greater than the average
average_rate <- mean(std_data_final$Rate)
average_rate
```

```
## [1] 232.7621
```

```
std_data_final <- std_data_final %>% mutate(chlamydia_greater_than_avg = (Rate > average_rate))
head(std_data_final)
```

```
##      County   Disease Year   Sex Cases Population  Rate sqMileage
## 1   Alameda Chlamydia 2003 Total  4928    1467187 335.9    739.02
## 2    Alpine Chlamydia 2003 Total     3       1249 240.2    738.33
## 3    Amador Chlamydia 2003 Total    31      36776  84.3    594.58
## 4     Butte Chlamydia 2003 Total   579     210623 274.9   1636.46
## 5 Calaveras Chlamydia 2003 Total    34      43186  78.7   1020.01
## 6    Colusa Chlamydia 2003 Total    16      19685  81.3   1150.73
##   populationDensity populationDensityper10000 chlamydia_greater_than_avg
## 1       1985.314335                0.1985314335                      TRUE
## 2          1.691655                0.0001691655                      TRUE
## 3         61.852064                0.0061852064                     FALSE
## 4        128.706476                0.0128706476                      TRUE
## 5         42.338801                0.0042338801                     FALSE
## 6         17.106532                0.0017106532                     FALSE
```

```
# Marginal Probability Calculation
# probability = Sum of trues in greater_than_avg/# of individuals in the data frame
# (# of individuals = 58)
m_prob_chlamydia <- sum(std_data_final$chlamydia_greater_than_avg)/58
m_prob_chlamydia
```

```
## [1] 0.4310345
```

12. [2 marks] Using any two variables in your dataset (or derived variables), calculate a conditional probability. Provide an equation (using probability notation) that describes this probability and then use R to calculate it.

Probability of the chlamydia rate being greater than the average rate given the population density = P(Chlamydia Rate > Average Rate | Population Density > Population Density Average) = P((Chlamydia Rate > Avg Rate) & P(Population Density > Population Density Average))/(P(Population Density > Population Density Average)) = 0.7643678 = 76.43%

The conditional probability of the chlamydia rate being greater than the average rate of chlamydia among all counties given the population density being greater than the population density average among all California counties is approximately 76.43%

```
# Conditional Probability Calculation
std_data_with_avg <- std_data_final %>% filter_at(vars(populationDensity),all_vars(!is.na(.)))
population_avg <- mean(std_data_with_avg$populationDensity)
std_data_with_avg <- std_data_with_avg %>% mutate(popDensity_greater_than_avg = (populationDensity > pop
head(std_data_with_avg)
```

```
##      County   Disease Year   Sex Cases Population  Rate sqMileage
## 1   Alameda Chlamydia 2003 Total  4928    1467187 335.9    739.02
## 2    Alpine Chlamydia 2003 Total     3       1249 240.2    738.33
## 3    Amador Chlamydia 2003 Total    31      36776  84.3    594.58
## 4     Butte Chlamydia 2003 Total   579     210623 274.9   1636.46
## 5 Calaveras Chlamydia 2003 Total    34      43186  78.7   1020.01
## 6    Colusa Chlamydia 2003 Total    16      19685  81.3   1150.73
##   populationDensity populationDensityper10000 chlamydia_greater_than_avg
## 1       1985.314335                0.1985314335                      TRUE
## 2          1.691655                0.0001691655                      TRUE
## 3         61.852064                0.0061852064                     FALSE
## 4        128.706476                0.0128706476                      TRUE
## 5         42.338801                0.0042338801                     FALSE
## 6         17.106532                0.0017106532                     FALSE
##   popDensity_greater_than_avg
## 1                        TRUE
## 2                       FALSE
## 3                       FALSE
## 4                       FALSE
## 5                       FALSE
## 6                       FALSE
```

```
head(population_avg)
```

```
## [1] 649.7702
```

```
# Probability of both chlamydia rate and population density being greater than their averages
and_probability <- std_data_with_avg %>% select(chlamydia_greater_than_avg, popDensity_greater_than_avg
  filter(popDensity_greater_than_avg == TRUE & chlamydia_greater_than_avg == TRUE)
and_probability
```

```
##   chlamydia_greater_than_avg popDensity_greater_than_avg
## 1                       TRUE                        TRUE
```

```
## 2                       TRUE                       TRUE
## 3                       TRUE                       TRUE
## 4                       TRUE                       TRUE
## 5                       TRUE                       TRUE
## 6                       TRUE                       TRUE
## 7                       TRUE                       TRUE
```

```r
# 7/58 rows for and probability

# Conditional Probability Calculation
m_prob_population <- sum(std_data_with_avg$popDensity_greater_than_avg)/57
c_pop_rate_prob <- (7/58)/(m_prob_population)
c_pop_rate_prob
```

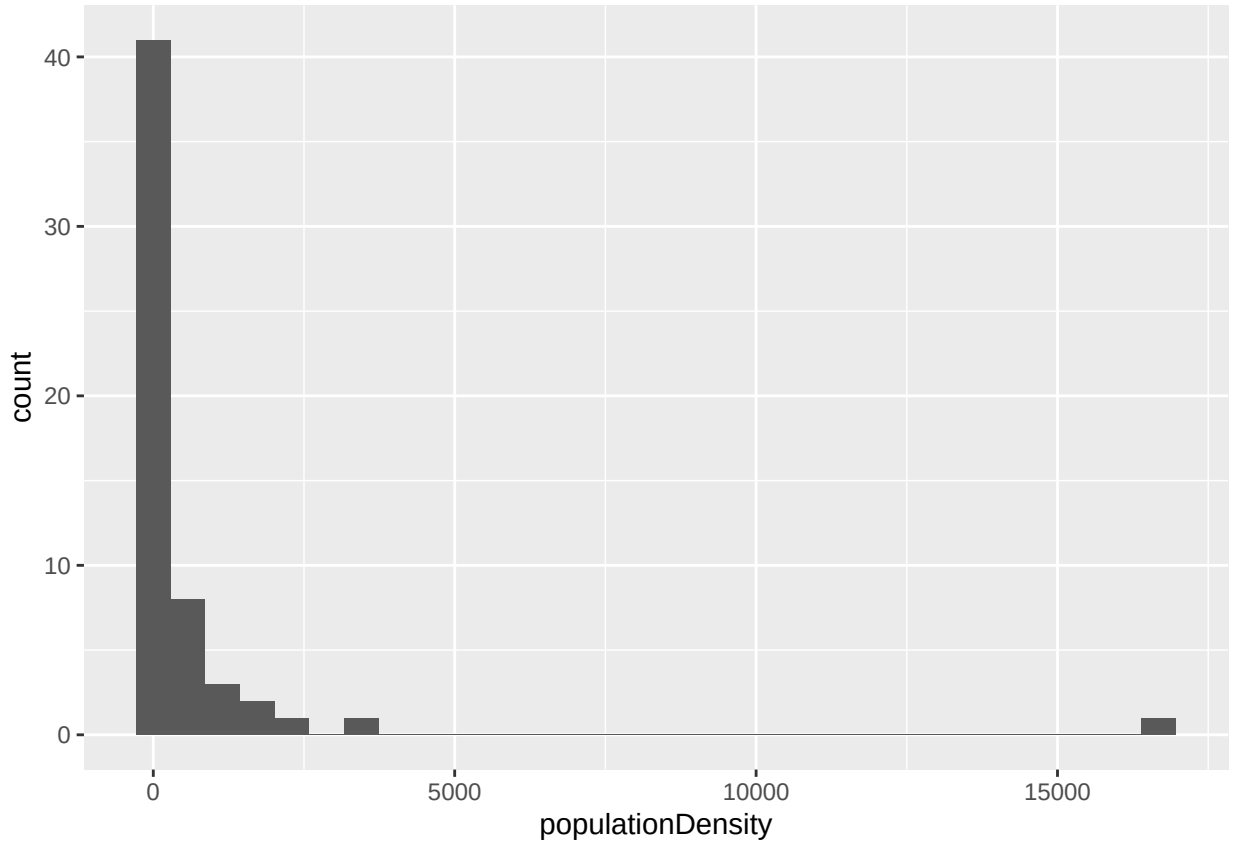
```
## [1] 0.7643678
```

13. [2 marks] Does your dataset contain a continuous variable? If it does, does the distribution of that variable appear to be normal? Justify your answer using a plot. If your data does not contain a continuous variable, give an example related to your dataset of a hypothetical variable that is continuous. That is, imagine what a continuous variable could be in relation to your dataset and topic of interest. For this hypothetical variable, describe what you imagine its shape might be, and how you would check whether or not it is normally distributed.

Population density is a continuous variable in our dataset. The distribution is not normal; it is right skewed with an outlier.

```
pop_histogram <- ggplot(std_data_final, aes(x=populationDensity)) +
  geom_histogram(bins = 30)
pop_histogram
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



14. [4 marks] Does your dataset contain a binary variable? If so, does this variable meet the criteria to be considered binomially distributed? If so, describe this variable in terms of n and p. Calculate a probability based on this variable, first write the formula for the probability and then using R to calculate the probability (you do not need to calculate the probability by hand). If your data does not contain a binary variable, you can create one based on an underlying continuous variable or a categorical variable with > 2 levels to answer this question.

Chlamydia_greater_than_avg is a binary variable because each observation either falls into TRUE or FALSE categories, there are a fixed number of observations, each observation is independent, and the probability of each observation is 50% (either true or false).

p = probability of true if chlamydia rate is greater than average n = number of counties x = number of successes

P(x) = nCx * pˆx * (1 - p)ˆ(n-x) = 0.1022 = 10.22%

The probability of the Chlamydia rate being greater than the average using a binomial distribution calculation is 10.22%.

```
# Converts T/F to 0s and 1s
head(std_data_final)
```

```
##      County   Disease Year   Sex Cases Population  Rate sqMileage
## 1   Alameda Chlamydia 2003 Total  4928    1467187 335.9    739.02
## 2    Alpine Chlamydia 2003 Total     3       1249 240.2    738.33
## 3    Amador Chlamydia 2003 Total    31      36776  84.3    594.58
## 4     Butte Chlamydia 2003 Total   579     210623 274.9   1636.46
## 5 Calaveras Chlamydia 2003 Total    34      43186  78.7   1020.01
## 6    Colusa Chlamydia 2003 Total    16      19685  81.3   1150.73
##   populationDensity populationDensityper10000 chlamydia_greater_than_avg
## 1       1985.314335                0.1985314335                      TRUE
## 2          1.691655                0.0001691655                      TRUE
## 3         61.852064                0.0061852064                     FALSE
## 4        128.706476                0.0128706476                      TRUE
## 5         42.338801                0.0042338801                     FALSE
## 6         17.106532                0.0017106532                     FALSE
```

```
std_data_final$chlamydia_greater_than_avg <- as.integer(std_data_final$chlamydia_greater_than_avg == TRU
head(std_data_final)
```

```
##      County   Disease Year   Sex Cases Population  Rate sqMileage
## 1   Alameda Chlamydia 2003 Total  4928    1467187 335.9    739.02
## 2    Alpine Chlamydia 2003 Total     3       1249 240.2    738.33
## 3    Amador Chlamydia 2003 Total    31      36776  84.3    594.58
## 4     Butte Chlamydia 2003 Total   579     210623 274.9   1636.46
## 5 Calaveras Chlamydia 2003 Total    34      43186  78.7   1020.01
## 6    Colusa Chlamydia 2003 Total    16      19685  81.3   1150.73
##   populationDensity populationDensityper10000 chlamydia_greater_than_avg
## 1       1985.314335                0.1985314335                         1
## 2          1.691655                0.0001691655                         1
## 3         61.852064                0.0061852064                         0
## 4        128.706476                0.0128706476                         1
## 5         42.338801                0.0042338801                         0
## 6         17.106532                0.0017106532                         0
```

```r
samples <- c()
for (i in seq_len(nrow(std_data_final)))
{sample_index <- std_data_final %>% slice_sample(n=58, replace = TRUE)
num_greater_than_avg <- sample_index %>% summarize(sum(chlamydia_greater_than_avg))
samples <- append(samples,num_greater_than_avg)
}

avged_samples <- mean(samples)
```

```
## Warning in mean.default(samples): argument is not numeric or logical: returning
## NA
```

```r
binom_prob <- dbinom(avged_samples, 58, 25/58)
binom_prob
```

```
## [1] NA
```

```r
set.seed(10)
sample_index <- std_data_final %>% slice_sample(n=58, replace = TRUE)
num_greater_than_avg <- sample_index %>% summarize(sum(chlamydia_greater_than_avg))
num_greater_than_avg
```

```
##   sum(chlamydia_greater_than_avg)
## 1                              34
```

```r
# Binomial Probability Calculation- How to calculate? pbinom or dbinom?
# When we ran the random sample the number of successes = 24, so we set x = 24 in our
# dbinom() function
binom_prob <- dbinom(24, 58, 25/58)
binom_prob
```

```
## [1] 0.1022143
```

16. [1 mark] Include parts I and II of your project.

17. [2 marks] Identify a statistical test to apply to your data. This must be a statistical test that we cover in part III of the course. Name the statistical test you have chosen and explain why this is the appropriate test for these data. For example, if I have pre- and post-intervention measurements of morning sleepiness recorded as a quantitative variable, I might choose a paired t test, because the paired t-test is appropriate for continuous outcome data in 2 groups that are inherently related.

Chi-Squared Test of Independence because we can draw two categorical variables from our data set of whether a person is male or female and whether they have chlamydia or not. With this, we can create a two-way table with sex as one category and whether a person has the disease or not as the other category.

18. [2 marks] What assumptions are required by the testing method you chose? Are these assumptions met by your data? How did you assess this? For example, one of the assumptions of the t-test is that the data are normally distributed, so you might choose to assess this with a histogram, or a q-q plot.

Sample Conditions: • Independent SRSs from >= 2 population, with each individual classified according to one category (i.e., each individual can only belong to one cell in the table so the categories need to be mutually exclusive) - This condition is met because each individual is classified according to one category (i.e. male and has chlamydia). Each category is mutually exclusive

```
# Actual Counts Table
# (take data from std_data_cleaned and take the row that has the total
# counts in California in 2003 for this table)
std_count_by_sex <- std_data_cleaned %>%
  filter(County == "California")
std_count_by_sex
```

```
## # A tibble: 3 x 7
##   Disease   County      Year Sex     Cases Population  Rate
##   <chr>     <chr>      <dbl> <chr>   <dbl>      <dbl> <dbl>
## 1 Chlamydia California  2003 Female  85153   17782868  479.
## 2 Chlamydia California  2003 Male    31007   17606060  176.
## 3 Chlamydia California  2003 Total  116385   35388928  329.
```

```
female_count <- 85153
male_count <- 31007
total_female <- 17782868
total_male <- 17606060
chlamydia_by_sex <-
  data.frame(chlamydia   = c("Chlamydia", "No Chlamydia",    "Chlamydia", "No Chlamydia"),
             sex = c("Male", "Male", "Female", "Female"),
             actual_count  = c(31007,  17697715, 85153, 17575053),
             stringsAsFactors = FALSE)

# sample, row and column total calculations
chlamydia_by_sex <-
  chlamydia_by_sex %>%
  group_by(chlamydia) %>%
  mutate(row_total = sum(actual_count)) %>%
  ungroup()

chlamydia_by_sex <-
  chlamydia_by_sex %>%
  group_by(sex) %>%
  mutate(column_total = sum(actual_count)) %>%
  ungroup()

chlamydia_by_sex <- chlamydia_by_sex %>%
  mutate(total_sample = sum(actual_count))

chlamydia_by_sex
```

```
## # A tibble: 4 x 6
##   chlamydia    sex    actual_count row_total column_total total_sample
```

```
##   <chr>        <chr>         <dbl>     <dbl>        <dbl>        <dbl>
## 1 Chlamydia    Male          31007    116160     17728722     35388928
## 2 No Chlamydia Male       17697715  35272768     17728722     35388928
## 3 Chlamydia    Female        85153    116160     17660206     35388928
## 4 No Chlamydia Female     17575053  35272768     17660206     35388928
```

```
# Expected Counts Calculation
chlamydia_by_sex_w_expected <- chlamydia_by_sex %>%
   mutate(expected_count = (row_total*column_total)/total_sample)
chlamydia_by_sex_w_expected
```

```
## # A tibble: 4 x 7
##   chlamydia    sex    actual_count row_total column_total total_sample
##   <chr>        <chr>         <dbl>     <dbl>        <dbl>        <dbl>
## 1 Chlamydia    Male          31007    116160     17728722     35388928
## 2 No Chlamydia Male       17697715  35272768     17728722     35388928
## 3 Chlamydia    Female        85153    116160     17660206     35388928
## 4 No Chlamydia Female     17575053  35272768     17660206     35388928
## # ... with 1 more variable: expected_count <dbl>
```

Expected counts conditions for chi-squared test for independence • Ei >= 5 for at least 80% of the cells (Expected is well over 5 for all of the cells. - see above chlamydia_by_sex_w_expected) • All Ei > 1 (Expected is well over 1 for all of the cells. - see above chlamydia_by_sex_w_expected) • If table is 2X2, all four cells need Ei >= 5 (Expected is well over 5 for all of the cells. - see above chlamydia_by_sex_w_expected)

19. [2 marks] Clearly state the null and alternative hypotheses for your test. Null hypothesis: The probability distribution for chlamydia in males is equal to the probability distribution of chlamydia in females, so sex and chlamydia are independent. Alternative Hypothesis: The probability distribution for chlamydia in males is different from the probability distribution, so sex and chlamydia are dependent.

20. [2 marks] Conduct the statistical test. Include the R code you used to generate your results. Annotate your code to help us follow your reasoning.

```
# test statistic calculation
chlamydia_by_sex_w_expected <- chlamydia_by_sex_w_expected  %>%
  mutate(diff_sq = (actual_count - expected_count)^2,
         ratio = diff_sq/expected_count, test_statistic = sum(ratio))
chlamydia_by_sex_w_expected
```

```
## # A tibble: 4 x 10
##   chlamydia    sex    actual_count row_total column_total total_sample
##   <chr>        <chr>         <dbl>     <dbl>        <dbl>        <dbl>
## 1 Chlamydia    Male          31007    116160     17728722     35388928
## 2 No Chlamydia Male       17697715  35272768     17728722     35388928
## 3 Chlamydia    Female        85153    116160     17660206     35388928
## 4 No Chlamydia Female     17575053  35272768     17660206     35388928
## # ... with 4 more variables: expected_count <dbl>, diff_sq <dbl>, ratio <dbl>,
## #   test_statistic <dbl>
```

```
test_stat <- 27234.46
# calculate p-value with test statistic
# df = (r-1)(c-1) = 1 * 1 = 1
p_val <- pchisq(q = test_stat, df = 1, lower.tail = F)
p_val
```

```
## [1] 0
```

21. [4 marks] Present your results in a clear summary. This should include both a text summary and a table or figure with appropriate labeling. For example, if your outcome and predictor/exposure variables are both binary, this might be a 2x2 table. If your method was regression, you might present your regression line graphically. Include your code and annotations.

The p-value from the chi-sq test was very small, nearly 0, so we reject the null hypothesis that the rates for females with the disease and males with the disease are equal and independent and we fail to reject the alternative. In the 2x2 table below, the proportion of females with chlamydia is much higher than the proportion of males with chlamydia so we can conclude that there may be a correlation between sex and whether or not an individual has chlamydia. In the previous question, we calculated the test statistic and then calculated the p-value from that test statistic.

```
two_way <-
  data.frame(chlamydia   = c("Chlamydia", "No Chlamydia",    "Chlamydia", "No Chlamydia"),
             sex = c("Male", "Male", "Female", "Female"),
             actual_count  = c(31007,  17697715, 85153, 17575053),
             stringsAsFactors = FALSE)
two_way
```

```
##      chlamydia    sex actual_count
## 1    Chlamydia   Male        31007
## 2 No Chlamydia   Male     17697715
## 3    Chlamydia Female        85153
## 4 No Chlamydia Female     17575053
```

```
# see above for p-value calculation
p_val
```

```
## [1] 0
```

22. [4 marks] Interpret your findings. Include a statement about the evidence, your conclusions, and the generalizability of your findings. Our analyses and conclusions depend on the quality of our study design and the methods of data collection. Any missteps or oversights during the data collection process could potentially change the outcome of what we are trying to find. Consider the methods used to collect the data you analyzed. Was there any potential issue in how the participants were selected/recruited, retained, or assessed that may have impacted the outcome of your analysis/visualization? Were there any potential biases that you might be concerned about? Were there factors that were not measured or considered that you think could be important to the interpretation of these data?

From part 3, we found that our findings are in favor of our alternative that gender and chlamydia are dependent. This finding could be true, however there should be more investigation on the influence of sex on chlamydia.

The data was collected from California Confidential Morbidity Reports and/or Laboratory Reports that were submitted to the California Department of Public Health through the California Reportable Disease Information Exchange (CalREDIE) as well as unique county developed systems. The date for the data was also estimated. Some factors that may not have been measured are: individuals with chlamydia did not submit a report and the date that reported may be inaccurate which could lead to changes in the outcome of our analysis.

It's also interesting how there are almost 3x more total chlamydia cases for women than men. A potential bias for the greater chlamydia cases for women might be because women tend to visit the doctor or OBGYN more, and therefore are regularly tested, than men would visit their doctors. We can further explore this in the future.

23. [1 mark] Create a statement of contribution. This is now common in journal articles. For example, the American Journal of Epidemiology provides the following instructions to authors: "Authorship credit should be based on criteria developed by the International Committee for Medical Journal Editors (ICMJE): 1) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; 2) drafting the article or reviewing it and, if appropriate, revising it critically for important intellectual content; 3) final approval of the version to be published. Authors should meet all conditions. In addition, each author must certify that he or she has participated sufficiently in the work to believe in its overall validity and to take public responsibility for appropriate portions of its content. Author names should be listed in ScholarOne and author contributions should be detailed in the cover letter (e.g., "Author A designed the study and directed its implementation, including quality assurance and control. Author B helped supervise the field activities and designed the study's analytic strategy. Author C helped conduct the literature review and prepare the Methods and the Discussion sections of the text.").

Author Tiffany Chung assisted in coding, attended meetings and submitted projects. Author Rayni Wells assisted in coding, attended meetings and office hours, and directed communication between the GSI. Author Afroze Khan suggested use of the STI dataset and attended group meetings. Author Ahmed Morsi attended meetings and contributed to the discussion on the direction of our data project. Author Maddy Wang directed group discussions, organized meetings, and cleaned the dataset in R. During meetings, Maddy took group input and implemented the code in R studio for parts 1 and 2.