PBH 142 Data Project

1. What is the problem you are addressing with this data? Express this in terms of the PPDAC framework.

   In terms of the PPDAC framework, our problem is “What is the relationship between gender and gonorrhea rates in Alameda county in 2001?” Our plan is to use the data of the total case numbers of female and male gonorrhea cases in Alameda county in 2001. The study design is an observational annual count of the number of STD cases each year in California. To analyze our data, we will use a two-sample t-test, which will yield a conclusion to whether there is a difference in number of cases for each sex. With our results, we will be able to generate new ideas for future analysis, such as investigating how different circumstances in each sex affects the rate of gonorrhea.

2. What is the target population for your project? Why was this target chosen?

The target population is men and women in Alameda county with STDs. We chose this target because we were interested in understanding whether gender plays a factor in the likelihood of having STDs.

3. What is the sampling frame used to collect the data you are using. 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?

The sampling frame of this dataset is generally California residents. Subjects were found from the California Confidential Morbidity Reports and Laboratory Reports sent to the California Department of Public Health and filtered through a specific surveillance definition. Data was collected from diagnoses estimated around 2001 up to the present day.

We would feel comfortable generalizing our findings on gonorrhea prevalence over the course of 2001 to 2017 to a general population of people in Alameda county. We would not be able to generalize gonorrhea prevalence for other counties in California because we do not know if other factors such as race, environment, and socioeconomic status may have contributed to Gonorrhea prevalence in Alameda county that may not be applicable to other counties.

4. 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.

Our datatset is a csv file that contains information on the type of STD, county, year, sex, cases, population, rate, lower 95% CI, and upper 95% CI from 2001 to 2017. This dataset was retrieved from the California Health and Human Services Agency (CHHSA) Open Data website. https://data.chhs.ca.gov/dataset/stds-in-california-by-disease-county-year-and-sex/resource/563ba92b-8ac5-48ec-9afd-2f515bbbad66

5. Import your data into R. Assign your dataset to an object. Completed.

6. What are the dimensions of the dataset? ii) Provide a list of variable names. iii) Print the first six rows of the dataset.
    a. Dimensions of the Dataset are 11 columns and
    b. Variable Names are _id, Disease, County, Year, Sex, Cases, Population, Rate, Lower 95% CI, and Upper 95% CI.
    c. Print first six rows

7. Use the data to demonstrate a statistical concept from Part I of the course. Describe the concept that you are demonstrating and interpret the findings. You should use a combination of code and written explanation.

Part II_____

1) See Part 1
2) Describe a quantity you will estimate in your problem using probability notation. Are you planning to calculate marginal probabilities/conditional probabilities?

P(having gonorrhea | female resident of Alameda County in 2001). We will calculate the conditional probability that any randomly selected female resident of Alameda County in 2001 has gonorrhea.

3) Describe type of theoretical distribution relevant to to your data

    A Binomial Probability Distribution would be relevant to our data because the outcome (a female having gonorrhea or a female not having gonorrhea) is a binary outcome that can be better visualized and evaluated through a probability distribution.

4) Use the data you have to **demonstrate** a statistical concept from Part II of the course. Describe the concept that you are demonstrating and interpret the findings.

Using the filtered dataset (only including residents of Alameda county in 2001 with gonorrhea), we can calculate the probability of having gonorrhea given that you are a female.

P(having gonorrhea | female in Alameda in 2001): 1106 cases/ 746,596 total female population = 0.00148139 * 100 = 0.1481%

```{r overall sex-specific rate across all counties}
gon_rate_overall <- std_data %>% filter(Disease == "Gonorrhea",
                                County == "California",
                                Year == "2001",
                                Sex =="Female")
gon_rate_overall

overall_rate <- 10442/17339700
```

The overall sex-specific rate in 2001 across all counties was 60.2 per 100,000 females, or 0.0602%.

After finding this probability through the filtered dataset, we can create a **binomial probability distribution** in order to compare the rate in Alameda county to the overall sex rate in 2001. Therefore, we can explore the probability of seeing numbers as high as those in Alameda compared to the total population of California.

```{r}
1 - pbinom(q = 1106, size = 746596, p = overall_rate)
```

This function calculated the probability of seeing more than 1106 cases in Alameda's female population in 2001, given the total sex-specific rate of gonorrhea in 2001.
Therefore the probability of seeing a rate greater than 0.1481% (Alameda) is around 0 given that the overall female rate in California was 0.0602%. This means that it is highly improbable, nearly impossible, to see a rate as high as Alameda County in 2001.

Part III _____

5)

a) **Identify the statistical test that you applied to your data (must be a concept from part III)**

   A two sample t-test will be applied to our dataset.

b) **What assumptions are required by the method you chose in 5a? Describe how you assessed whether these assumptions are met by your dataset.**

The conditions for a two sample t-test are that the populations are normally distributed, independent groups that were selected through a simple random sample. The means and standard deviations of the population are unknown. In our dataset, we filtered our dataset to years 2001 to 2017, people with cases of gonorrhea, and gender. The dataset we identified also stated that the individuals in the database were selected randomly and not a product of solely self reporting. Finally, the standard deviation and average of women infected with gonorrhea is not given in the dataset.

c) **Explain why this test is appropriate for the data you have and the questions you are trying to answer. Use at least one visualization technique and include both the output and the r code that generated it.**

The original dataset contained cases of gonorrhea grouped by year. We realized that by default, this dataset would have an inherent trend since gonorrhea cases from one year most likely were affected by the cases of gonorrhea from the previous year due to the fact that the incidence and prevalence of gonorrhea affect the transmission rate. If one year has many cases, it is likely that those cases go on to produce more cases.

To account for this, in the following analysis we make the assumption that there is no trend for gonorrhea cases, meaning that in this analysis we are assuming that there is no year to year variation of Gonorrhea transmission from 2001 to 2017. In the new column, new_cases, we assume that each year's total gonorrhea case number varies randomly around the overall mean of the original dataset's case population, separated into two groups, male and female population.

A two sample t-test compares samples from independent populations. We chose to use a two sample t-test because we are interested in comparing the means between two groups. The groups are females with gonorrhea and males with gonorrhea. These groups come from two different underlying means. The following code shows that these theoretical cases drawn from a simple random sample do in fact vary randomly around the mean and sd from the original data set. The histograms that are generated show that these data are not necessarily normally distributed, but given the other factors, Normality isn't required for the t-test.

_____code starts here_____

```{r}
library(dplyr)
library(ggplot2)
library(readr)
library(broom)

# Reading in the original dataset csv file
```

```
STDS_data <- read_csv("STDS_data.csv")

#Creating a new dataset that filters Gonorrhea cases in Alameda county
STD_data_gon <- STDS_data %>% filter(County == "Alameda") %>% filter(Sex %in%
c("Male", "Female")) %>% filter(Disease == "Gonorrhea")
STD_data_gon

#calculating the mean and sd by sex of the original gonorrhea dataset
STD_statistics <- STD_data_gon %>% group_by(Sex) %>%
  summarise(mean_gon = mean(Cases),
            sd_gon = sd(Cases))
STD_statistics

#randomly sampling cases from a population with the mean and sd from the original
population of gonorrhea cases per year in Alameda county
set.seed(12345)
samples_female <- round(rnorm(17, 920.000, 171.3210), 0)
samples_male <- round(rnorm(17, 1223.118 , 507.2217), 0)

#adding the new randomly sampled cases into the original dataset
STD_gon <- STD_data_gon %>% arrange(Sex) %>% mutate(new_cases =
c(samples_female, samples_male))
STD_gon

#histogram visualization of the random sample (not very Normally distributed)
hist(STD_gon$new_cases[STD_gon$Sex=="Female"])
hist(STD_gon$new_cases[STD_gon$Sex=="Male"])
```

_____end of code_____

d) State the null and alternative hypothesis for this test.

Null Hypothesis: The mean number of gonorrhea cases in men in Alameda County between 2001 and 2017 is the same as the mean number of gonorrhea cases in women in Alameda County between 2001 and 2017.

Alternative Hypothesis: There is a difference between the average rate of men with gonorrhea and the average rate of women with gonorrhea in Alameda County between 2001 and 2017.

6) **Include the R code you used to generate your results -- annotate your code to help us follow your reasoning** (Jeanne)

Please post the following code as is:
_____start of the code_____

```{r}
library(dplyr)
library(ggplot2)
library(readr)
library(broom)

# Reading in the original dataset csv file
STDS_data <- read_csv("STDS_data.csv")

#Creating a new dataset that filters Gonorrhea cases in Alameda county
STD_data_gon <- STDS_data %>% filter(County == "Alameda") %>% filter(Sex %in% c("Male", "Female")) %>% filter(Disease == "Gonorrhea")

#calculating the mean and sd by sex of the original gonorrhea dataset
STD_statistics <- STD_data_gon %>% group_by(Sex) %>%
  summarise(mean_gon = mean(Cases),
            sd_gon = sd(Cases))
STD_statistics

#randomly sampling cases from a population with the mean and sd from the original population of gonorrhea cases per year in Alameda county
set.seed(12345)
samples_female <- round(rnorm(17, 920.000, 171.3210), 0)
samples_male <- round(rnorm(17, 1223.118 , 507.2217), 0)
#adding the new randomly sampled cases into the original dataset
STD_gon <- STD_data_gon %>% arrange(Sex) %>% mutate(new_cases = c(samples_female, samples_male))

#calculating a two-sided, two-sample t-test for the cases sampled
t.test(STD_gon$new_cases[STD_gon$Sex=="Female"],
STD_gon$new_cases[STD_gon$Sex=="Male"], alternative = "two.sided", paired = F)
```

_____ This is the end of the code_____

7) **Present your results in a clear summary. This should include both a text summary and table or figure with appropriate labelling. (Doris)**

From the t-test performed above, the t-statistic was calculated to be -3.7536 with a p-value of 0.001457. The t-statistic indicates a negative difference between our sample data and our null hypothesis, i.e. there is a lower chance of gonorrhea if one is female versus male. Since the p-value is smaller than the significance level of 0.05, we reject the null hypothesis. The 95% confidence interval calculated was -866.9545 to -244.6927.

Table of T-Test Values

t	-3.7536
df	17.974
p	0.001457
95% CI	-866.9543 to -244.6927
Mean of x	924.4118
Mean of y	1480.2353

Summary of Statistics

Sex	Mean Cases of Gonorrhea	SD of Cases of Gonorrhea
Female	920.000	171.3210
Male	1223.118	507.2217

8) **Interpret your findings. Include a statement about the strength of this testing, your conclusions and the generalizability of your findings. (Doris + everyone else too!)**

Using a method that created confidence intervals to capture the true parameter 95% of the time, the confidence interval computed is (-866.95, -244.69). This indicates that 95% of such intervals will contain the true value. Additionally, since the p-value is $0.001 < P < 0.01$, there is strong evidence against the null hypothesis that the mean number of gonorrhea cases in men is the same as the mean number of gonorrhea cases in women.

Although samples used in a t-test are assumed to be normally distributed and have similar variance between groups, a t-test is robust even when there are violations to these assumptions. By definition of two sample t-tests, regardless of if the populations tested are normal and have differing variances, these samples are comparable. We can thus conclude that our data and its results are robust.

Our original dataset is not normally distributed because there is a clear year to year trend of gonorrhea cases, but for the purposes of this analysis we assumed that each year's total population for those with gonorrhea varied randomly around the overall mean of the whole dataset's case numbers that were grouped by sex. We randomly sampled 17 total case populations (one total number of cases of gonorrhea per year), and because of this we cannot generalize these findings to the general public.

Since our dataset consisted of individuals in Alameda County over the course of several years, the findings of our data cannot be extrapolated to the general public. Furthermore, we recognize that STDs have trends and are contingent on cultural stigma, access to resources, education, and socioeconomic status, so our data alone cannot take into account these social and environmental factors that influence gonorrhea rates. Our findings extend to the data we collected on gonorrhea rates in Alameda county in the years we sampled.

9) Peer Evaluation form!