Project Part 3

Hyun Suk Ryoo (Max Ryoo)

Question

With recent medical and technological advances being made today, the life expectancy of Humans have significantly increased as shown by the previous project (Project #1). It was evident that life expectancy for both males and females increased by a great factor. The new question that the researcher thought of was whether there was a difference in the life expectancy between Males and Females. Which population had a higher mean life expectancy rate? Is this difference statistically significant?

Data

Description

The data set that will be covered is titled 'NCHS - Death rates and life expectancy at birth' . This data set was collected by the researcher from data.gov under the health section. The metadata was updated in August 20, 2018. This dataset from the U.S Department of Health & Human Services highlights the difference in age-adjusted death rates and life expectancy at birth by race and sex from the 1900 until 2015.

The data was collected from the U.S Department of Health & Human services, which collected their samples from 1900 to 2015. The columns are separated into 5 columns of year, race, sex, Average Life Expectancy (Years), and Age-adjusted Death Rate. The years are represented as a numeric. The race is a string value that takes either the values of White, Black, or All Races. The Sex column is a string value that takes either the values of Male, Female, Both Sexes. The Average Life Expectancy is shown as a numeric number that represents how many years a person born in a certain year will have on average. The Age-adjusted represents the deaths per 100,000 which is calculated based on the 2000 U.S. standard population.

Relevance

This data that was collected had a lot of entries as well as variables. Not all of this data is completely necessary. The researcher focused on the 'Sex' and 'Average Life Expectancy' columns. From the 'Sex' column the researcher subset to only include "Male" and "Female" for the remaining entires were "Both Sexes", which was not relevant to the study. The researcher subset the data to only include entries from the years 1900-2014. This was because the data for 2015 was null.

Test

The test that was used to detect whether there is a statistically significant difference between the two Sexes for life expectancy was "Two-sample t-test". This test was chosen due to the researcher having quantitative data for two populations. We didn't know the standard deviation of the population, which lead the researcher to use a two-sample t-test. The null hypothesis will be that there is no difference between the two Sexes for Life Expectancy

Assumptions

There are assumptions that must be looked at before conducting a two sample t-test. The two-sample t-test requires either two independent normal populations or two large enough independent samples such that the Central Limit Theorem holds. In this dataset there are 114 (2014-1900) entries of Life Expectancy of both women and men, which equate to having 228 entries to conduct this test, which is seemed to be an adequate sample for this study. Also, the data was collected originally by census to represent the population. Therefore, we can use this data to represent the population. This method also follows the law of large numbers and the central limit theorm holds.

Testing

The two sample t-test was chosen to see if there is a significant difference in life expectancy between the two Sexes. This means that their difference of life expectancy is not zero and their probability of the difference not being zero is statistically significant (alpha = 0.05). The below code was run to conduct this test.

```
## R code
```

```
setwd("/Users/maxryoo/Documents/Fall 2018/STAT3080/Project1")
dataset <- read.csv("deathrateslifeexpect.csv")
names(dataset) <- c("Year", "Race", "Sex", "Average_Life_Expectancy",
                    "Age_Adjusted_Death_Rate")
female <- dataset$Average_Life_Expectancy[dataset$Sex == "Female"]
male <- dataset$Average_Life_Expectancy[dataset$Sex == "Male"]
female <- female[-1]
male <- male[-1]
t.test(female, male, mu=0, alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  female and male
## t = 5.8835, df = 679.17, p-value = 6.307e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.464248 6.934592
## sample estimates:
## mean of x mean of y
##  66.73536  61.53594
```

We can see from the above output that the mean life expectancy between 1900 and 2014 was 66.74 and 61.54 for females and males respectively. We can also see that the probability of the true difference of means is zero is 6.307e-09, which is below our alpha level.

Conclusion & Generalization

Based on the results of this that showed a low p value, we can infer that there is a significant difference in means of life expectancy between males and females. The mean of females was 66.74 and the mean of males was 61.54. Through the two sample t-test we can see that females had a significantly higher life expectancy. We thus reject the null hypothesis. Because the data was collected by U.S. Department of Health & Centers for Disease control to represent the entire population, we are able to relay this same finding to the population. The conclusion thus becomes that in America there is a statistically significant difference between males and females in the aspect of life expectancy where the female population has a higher life expectancy.

References

1. <https://catalog.data.gov/dataset/age-adjusted-death-rates-and -life-expectancy-at-birth-all-races-both-sexes-united-sta-1900>
2. http://www.nccp.org/media/releases/release_34.html