Project

Will Curkan, Alina Ila, Rachel Uc, Jocelyn Serrot, Henry Zerep

2022-11-18

Introduction

Abortion is a sensitive topic in the United States with the legendary court ruling of *Roe vs. Wade* being overturned in June 2022. While sensitive, it is still an interesting topic because of the different viewpoints: are we killing a baby, or are we killing a fetus? The biology is irrelevant, though, in statistical analyses. Because of the amount of data, it is a good topic of analyses and we shall compare statistics among the categorized demographics of recorded abortions, to see if there are any disparities*****

We are using the "Pregnancies, Births and Abortions in the United States: National and State Trends by Age" dataset called `NationalAndStatePregnancy_PublicUse` from the Guttmacher Institute website which sources its data from numerous different organizations like the World Health Origanization (WHO) and UNICEF.

We will use the columns: state, year, abortionslt15, abortions1517, abortions1819, abortions1519, abortionslt20, abortions2024, abortions2529, abortions3034, abortions3539, abortions40plus, which are the number of abortions in the age range. For example `abortionslt15` is the number of abortions of the given U.S. state and year of girls less than 15 years old, and `abortions2024` is the number of abortions of ladies of ages 20-24.

Some questions to ask:

- Is there a statistically significant difference in mean abortion rates among age groups of [young] women.
- Is there a statistically significant difference in mean abortion rates among states
- Is there a statistically significant difference in mean abortion rates among the years.

```
## [1] 912  10
```

```
## round(cor(abortions[,-c(1,2,102,103)]),2)
```

```
abortions[is.na(abortions)] <- 0
```

Lets test the normality of the columns in the dataset. There are 103 columns (features) in the dataset, so it isn't practical to plot them all, but we can loop a shapiro test using significance level $\alpha = .05$ and have it tell us which columns are normal.

```
# Create a vector to keep track of which columns contain normally distributed data
normal_columns <- c()
i <- 0

for (column in colnames(abortions))
{
```

```
  # If the class if numeric, tests its normality
  if (class(abortions[,column]) == 'numeric' |
      class(abortions[,column]) == 'integer' )
  {
    # If any of the columns are normal, print them
    if ((shapiro.test(abortions[,column])$p.value > .05)) {
      normal_columns[i] <- column
      i <- i + 1
      }
  }
}
normal_columns # No columns have normally distributed data
```

```
## NULL
```

We can see that no columns contain normally distributed data. Because there is no normally distributed data, we will use resampling methods to estimate population attributes.

First, let's look at the mean difference in age groups