

STATS 205: Final Project Write-Up

Brian Liu

6/14/2019

1. Background of the data and why it is interesting or important

The data we are using is the data from WHO suicide statistics from Kaggle. This gives population-based statistics on suicide rate...

2. Explanation of the method studied and its properties

3. Data analysis or simulation study

We will use the crude rate of suicide per 100,000 people.

This analysis provides information on age-standardized rates...

```
who_suicide_statistics_df <- read.csv("who_suicide_statistics.csv")
head(who_suicide_statistics_df)
```

```
##   country year    sex      age suicides_no population
## 1 Albania 1985 female 15-24 years         NA      277900
## 2 Albania 1985 female 25-34 years         NA      246800
## 3 Albania 1985 female 35-54 years         NA      267500
## 4 Albania 1985 female  5-14 years         NA      298300
## 5 Albania 1985 female 55-74 years         NA      138700
## 6 Albania 1985 female  75+ years         NA       34200
```

```
colnames(who_suicide_statistics_df)
```

```
## [1] "country"    "year"       "sex"        "age"        "suicides_no"
## [6] "population"
```

Filter and save countries with missing suicide rate.

```
library(tidyverse)
```

```
## Registered S3 methods overwritten by 'ggplot2':
```

```
##   method      from
## [.quosures   rlang
## c.quosures   rlang
## print.quosures rlang
```

```
## -- Attaching packages ----- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.1    v purrr   0.3.2
## v tibble  2.1.1    v dplyr  0.8.1
## v tidyr   0.8.3    v stringr 1.4.0
## v readr   1.3.1    v forcats 0.4.0
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
filtered_suicide_df <- drop_na(who_suicide_statistics_df, "suicides_no")
head(filtered_suicide_df)
```

```
##      country year    sex      age suicides_no population
## 25 Albania 1987 female 15-24 years          14    289700
## 26 Albania 1987 female 25-34 years           4    257200
## 27 Albania 1987 female 35-54 years           6    278800
## 28 Albania 1987 female  5-14 years           0    311000
## 29 Albania 1987 female 55-74 years           0    144600
## 30 Albania 1987 female  75+ years            1     35600
```

After filtering countries with missing suicide rate, take a random sample of 100 countries and make sure each continent has approximately equal countries.

$$\frac{100 \text{ countries}}{7 \text{ continents}} \approx 14 \text{ to } 15 \text{ countries per continent}$$

Filter countries by continent:

```
library(countrycode)
filtered_suicide_df$continent <- countrycode(sourcevar = filtered_suicide_df[, "country"],
                                             origin = "country.name",
                                             destination = "continent")
```

```
## Warning in countrycode(sourcevar = filtered_suicide_df[, "country"], origin = "country.name", : Some
```

```
## Warning in countrycode(sourcevar = filtered_suicide_df[, "country"], origin = "country.name", : Some
```

```
head(filtered_suicide_df)
```

```
##      country year    sex      age suicides_no population continent
## 25 Albania 1987 female 15-24 years          14    289700    Europe
## 26 Albania 1987 female 25-34 years           4    257200    Europe
## 27 Albania 1987 female 35-54 years           6    278800    Europe
## 28 Albania 1987 female  5-14 years           0    311000    Europe
## 29 Albania 1987 female 55-74 years           0    144600    Europe
## 30 Albania 1987 female  75+ years            1     35600    Europe
```

```
write.csv(filtered_suicide_df, 'filtered_suicide.csv')
```

```
# Get list of continents
```

```
unique(filtered_suicide_df$continent)
```

```
## [1] "Europe" "Americas" "Asia" "Oceania" "Africa" NA
```

Notably, there are countries that are not on any of the listed continents. Let's see which ones those are:

```
not_in_a_continent = filtered_suicide_df[is.na(filtered_suicide_df$continent),]
write.csv(not_in_a_continent, 'not_in_a_continent.csv')
head(not_in_a_continent)
```

```
##      country year    sex      age suicides_no population continent
## 32317 Rodrigues 2001 female 15-24 years            0      NA      <NA>
## 32318 Rodrigues 2001 female 25-34 years            0      NA      <NA>
## 32319 Rodrigues 2001 female 35-54 years            0      NA      <NA>
## 32320 Rodrigues 2001 female  5-14 years            0      NA      <NA>
## 32321 Rodrigues 2001 female 55-74 years            0      NA      <NA>
## 32322 Rodrigues 2001 female  75+ years             0      NA      <NA>
```

```

unique(not_in_a_continent$country)

## [1] Rodrigues                Virgin Islands (USA)
## 141 Levels: Albania Anguilla Antigua and Barbuda Argentina ... Zimbabwe
# Get seven dataframes, filtered by list of countries for each continent.
europe_suicide = filtered_suicide_df[filtered_suicide_df$continent == 'Europe',]
head(europe_suicide)

##   country year    sex      age suicides_no population continent
## 25 Albania 1987 female 15-24 years         14      289700      Europe
## 26 Albania 1987 female 25-34 years          4      257200      Europe
## 27 Albania 1987 female 35-54 years          6      278800      Europe
## 28 Albania 1987 female  5-14 years          0      311000      Europe
## 29 Albania 1987 female 55-74 years          0      144600      Europe
## 30 Albania 1987 female  75+ years          1       35600      Europe
americas_suicide = filtered_suicide_df[filtered_suicide_df$continent == 'Europe',]

```

4. Interpretation of the results or discussion