An Analysis of Suicide Rates in 1970

MATH 4650 Project 1
3/26/2021
Alexander Kahanek and Jonah Turner

Introduction

Suicides are a tragic event, but it would be beneficial to know their patterns. These patterns could help to identify months, or seasons, of higher or lower suicide rates, allowing for a more strategic allocation of resources. Our dataset includes the number of suicides per month in the year 1970. By conducting a visual analysis of the data and two goodness-of-fit tests, we can identify whether there is a relationship between the time of year and the suicide rate. The two main questions we have are:

1. Are there variances between the probability of a suicide being in any given month?
2. Are there variances between the probability of a suicide being in any given season?

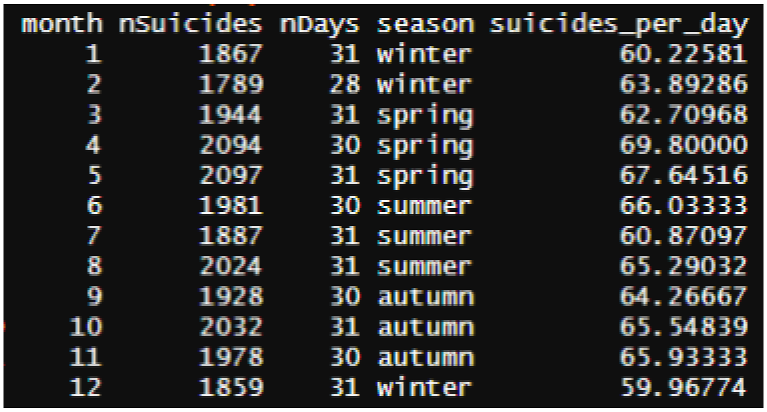
Our seasons are identified to be: Winter = {December, January, February}, Spring = {March, April, May}, Summer = {June, July, August}, Autumn = {September, October, November}.

What does our data look like?

The original dataset has three main columns: *(month)* an integer indicator for the month, *(nDays)* the number of days included in the month, and *(nSuicides)* the total number of suicides for that month. No changes were made to the actual data; however, the columns were renamed and two extra columns were added. Namely, a column *(season)* was added for the season in which a month falls (Winter, Spring, Summer, and Autumn) and a column *(suicides_per_day)* was calculated for the number of suicides per day. We assume our data to be independent and identically distributed. For the data we observed the mean number of suicides per month to be 1956.667 and the variance to be 9212.424. When normalizing to the suicide rate based on a daily granularity, we found the average rate to be 64.34869, with a variance of 9.040442.

The following code was used to load our packages and data, then perform all data manipulation:

```r
library(dplyr)
library(ggplot2)
library(cowplot)

raw <- read.csv('suicides1970.csv') %>%
  rename(
    "month" = Month
    ,"nSuicides" = Number
    ,"nDays" = Days
  ) %>%
  mutate(
    season = ifelse(month < 3 | month == 12, "winter",
                    ifelse(month < 6, "spring",
                           ifelse(month < 9, "summer", "autumn")))
    ,suicides_per_day = nSuicides / nDays)

raw %>% head(20)
```



```
month nSuicides nDays season suicides_per_day
    1      1867    31 winter         60.22581
    2      1789    28 winter         63.89286
    3      1944    31 spring         62.70968
    4      2094    30 spring         69.80000
    5      2097    31 spring         67.64516
    6      1981    30 summer         66.03333
    7      1887    31 summer         60.87097
    8      2024    31 summer         65.29032
    9      1928    30 autumn         64.26667
   10      2032    31 autumn         65.54839
   11      1978    30 autumn         65.93333
   12      1859    31 winter         59.96774
```

Figure 1

Basic Statistics of Suicides in 1970

To get an idea of the distribution of the number of suicides per month, a box plot was graphed as an exploratory visualization for the data.

```
ggplot(raw) +
  geom_boxplot(aes(x=nSuicides)) +
  labs(title = "The distribution of the number of suicides, per month in 1970"
       ,x = "number of suicides, per month") +
  theme(axis.title.y=element_blank()
        ,axis.text.y=element_blank()
        ,axis.ticks.y=element_blank())
```

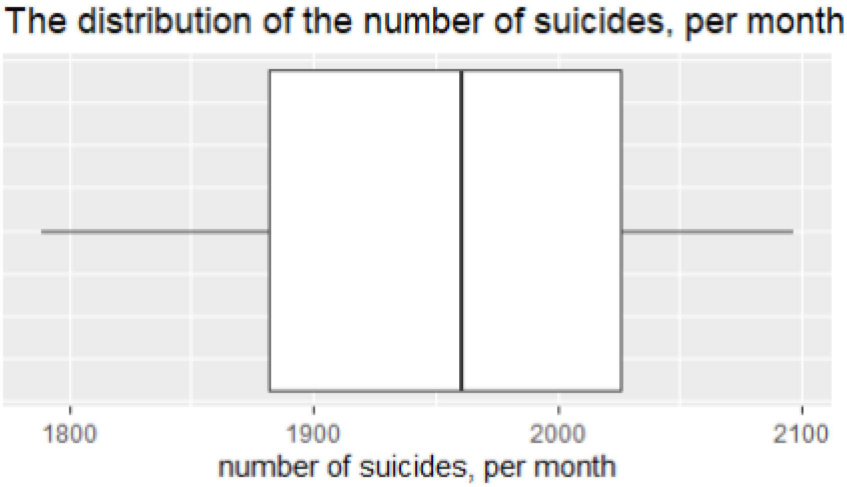


Figure 2

From Figure~2 we can see that the median number of suicides per month is slightly higher than 1,950; specifically, it is 1,961, with the most extreme months having as few as 1,789 suicides and as many as 2,097. 50% of the data - the second and third quartiles - lies between 1,882 and

2,026. February is the leftmost data point, as the box plot does not account for the different lengths of the months.

To get an idea of the Empirical Cumulative Density of the number of suicides per month, we also visualized this using the following code:

```
ecdf(raw$nSuicides) %>%
  plot(main = "Empirical Cumulative Distribution of the number of suicides in 1970"
       ,xlab = "Number of suicides, per month"
       ,ylab = "Cumulative probability")
```

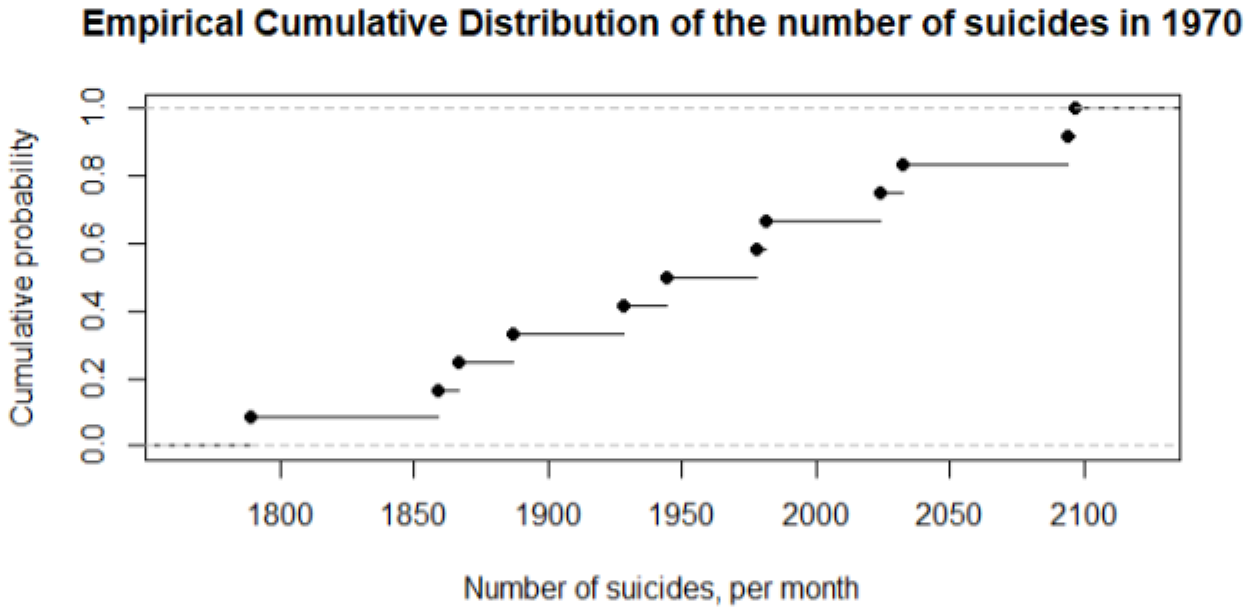


Figure 3

From Figure~3 we can see that we have too few data points to get a smooth curve; however, our curve does slightly resemble the curve of a typical standard distribution. Effectively, this graph has the same information as the box plot, except at a higher granularity. Figure~3 tells us that in the first 25% of the data, the number of suicides for any given month will be less than 1,887. Likewise, 50% of the data includes months where the number of suicides is less than 1,978. The ECDF shows the distribution within the quantiles more clearly than the box plot, with the steps showing data more narrowly distributed near the median.

To compare the differences of the number of suicides per month, with a normalization of the number of days, we compared the two distributions across our twelve months.

```
p1 <- ggplot(raw) +
  geom_bar(aes(x=as.factor(month), y=nSuicides), stat = "identity") +
  labs(title = "The number of suicides, per month"
       ,x = "Index of Month"
```

```r
       ,y = "Number of suicides, per month")

p2 <- ggplot(raw) +
  geom_bar(aes(x=as.factor(month), y=suicides_per_day), stat = "identity") +
  labs(title = "The number of suicides, per day for each month"
       ,x = "Index of Month"
       ,y = "Number of suicides, per day")

plot_grid(p1, p2)
```

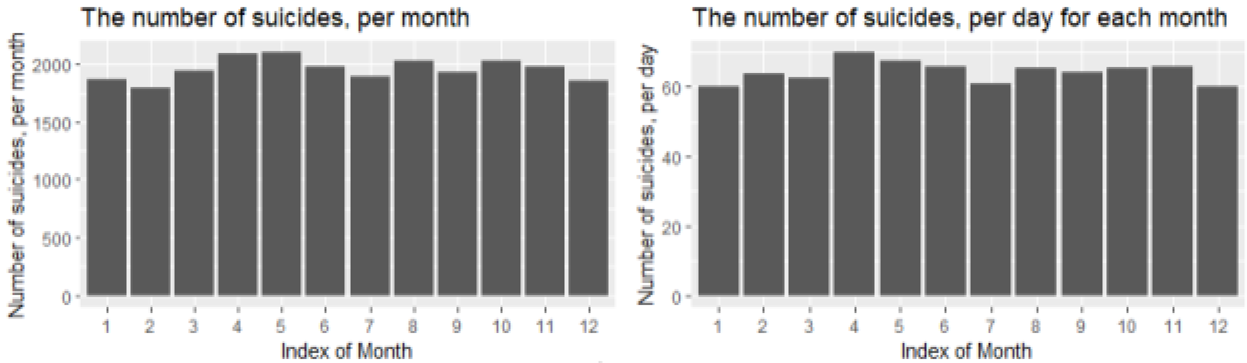


Figure 4

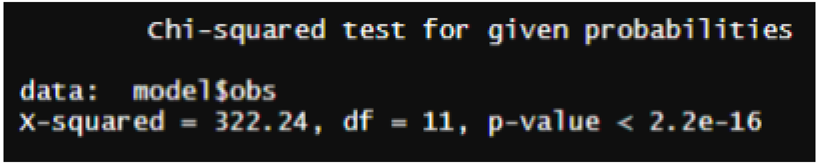
From Figure~4 we can see that there are drastic changes in how the months are ranked when normalizing the number of suicides to a per day granularity. We specifically see a large change in February where initially, on a per month granularity, we notice that February is lower than January and March. However, when normalizing to the number of suicides per day, we see that in actuality February has a higher rate than January and March. Due to this, we will be using the normalized suicide rate, based on daily granularity, for our multinomial modeling as well as the Pearson's Chi-Square tests for the months and seasons.

Modeling suicides to test for variances between months

To test the null hypothesis of a constant suicide rate among months, i.e., $H_0 = p(suicide \mid month_i) = \#\{month_i\}/365$, with the alternate hypothesis being there is a variation in the suicide rate among months, i.e., $H_A = p(suicide \mid month_i) \neq \#\{month_i\}/365$,where $\#\{month_i\}$ refers to the number of days in the respective month, a Pearson's Chi-Squared Test was conducted on a created multinomial model. The probabilities used in the creation of the multinomial model - the probability that a suicide falls in a given month based on the daily suicide rate - were extrapolated from the original dataset. From there we ran 1 experiment, where 100,000 samples were randomly assigned months based on the normalized probabilities. In essence, this simulates 100,000 suicides and their distribution throughout the possible 12 months.

```r
total_spd <- 772.1843 # the total sum of the number of suicides per day

probs <- raw %>% # the probability of a suicide being in a month
```

```
  group_by(month) %>%
  summarise (
    p = sum(suicides_per_day) / total_spd
  ) %>%
  select(p) %>%
  unlist() %>% as.numeric()


number_of_experiments <- 1 # how many times to draw K samples
number_of_samples <- 100000 # the amount of samples per experiment

multinomial_model_obs <- rmultinom(n = number_of_experiments, size =
number_of_samples, prob = probs)

model <- multinomial_model_obs %>% # find observations and probabilities of our
model
  as.data.frame() %>%
  rename("obs" = V1) %>%
  mutate(
    month = 1:n()
    ,p = obs / number_of_samples)

chisq.test(model$obs, p = raw$nDays/365)
```



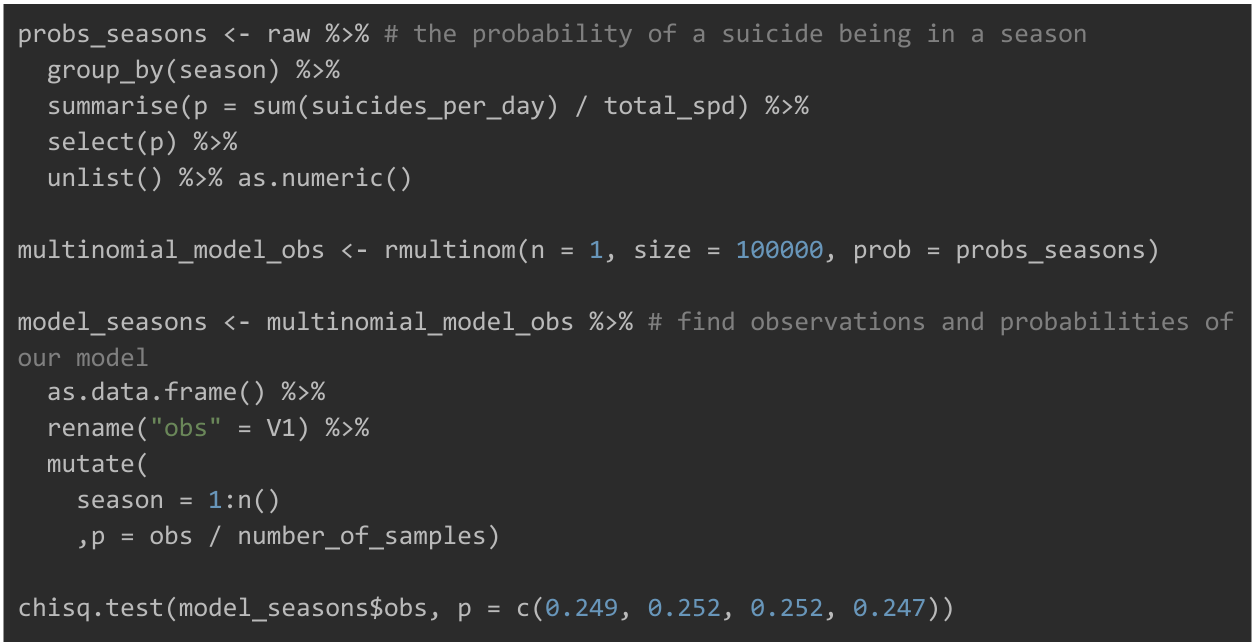
```
        Chi-squared test for given probabilities

data:  model$obs
X-squared = 322.24, df = 11, p-value < 2.2e-16
```

Figure 5

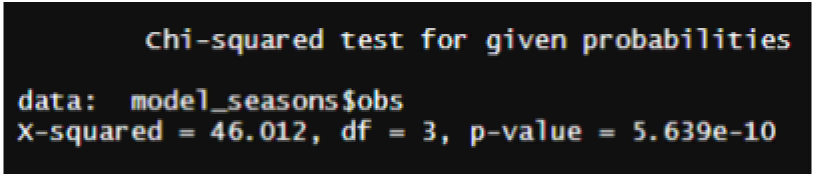
From Figure~5 we see our resulting p-value is $< 2.2e^{-16}$ from Pearson's Chi-Squared test, which is considerably lower than the given significance level $\alpha = 0.05$. The null hypothesis that the suicide rate remains constant among the months is therefore rejected and the alternative hypothesis is accepted. Thus, we conclude there is a statistically significant difference between the suicide rates among months.

Modeling suicides to test for variances between seasons

The same procedure was used for modeling the suicides with a random multinomial distribution and conducting the Pearson's Chi-Squared test, with the only change being that the months were grouped into the four seasons. The null hypothesis for this test is that the suicide rate is constant over the four seasons, i.e. $H_0 = p(suicide \mid season_i) = \#\{season_i\}/365$, with the alternate hypothesis being that the suicide rate varies among the seasons, i.e. $H_A = p(suicide \mid season_i) \neq \#\{season_i\}/365$, where $\#\{season_i\}$ refers to the number of days in the respective season.



```
probs_seasons <- raw %>% # the probability of a suicide being in a season
  group_by(season) %>%
  summarise(p = sum(suicides_per_day) / total_spd) %>%
  select(p) %>%
  unlist() %>% as.numeric()

multinomial_model_obs <- rmultinom(n = 1, size = 100000, prob = probs_seasons)

model_seasons <- multinomial_model_obs %>% # find observations and probabilities of
our model
  as.data.frame() %>%
  rename("obs" = V1) %>%
  mutate(
    season = 1:n()
    ,p = obs / number_of_samples)

chisq.test(model_seasons$obs, p = c(0.249, 0.252, 0.252, 0.247))
```



```
        Chi-squared test for given probabilities

data:  model_seasons$obs
X-squared = 46.012, df = 3, p-value = 5.639e-10
```

Figure 6

From Figure~6 we see our resulting p-value is $5.639e^{-10}$ from Pearson's Chi-Squared test, which is considerably lower than the given significance level $\alpha = 0.05$. The null hypothesis that the suicide rate does not vary among the seasons is therefore rejected, and the alternate hypothesis of a varying suicide rate is accepted. Thus, we conclude that there is a statistically significant variation in the suicide rates of the four seasons.

Conclusions

In summary, we found that there are statistically significant differences between the suicide rates of months and seasons; however, the reasons for these variances are unknown. We also notice from the box plot that the median number of suicides per month is 1,961, where 50% of our data lies *(Q1 to Q3)* between 1,882 and 2,026 suicides per month. The graph of the ECDF makes more clear the distribution of the number of suicides, with the steps growing more narrow around the median and larger towards the edges of the distribution. It can also be found that normalizing the number of suicides into a per day granularity, as opposed to a per month granularity, shows better results for comparing the months directly. Due to this, the goodness of fit tests were normalized to test for a constant suicide rate based on a per day granularity, which resoundingly rejected both null hypotheses, with sufficiently low p-values to reduce the possibility of Type 1 errors almost completely. We therefore conclude that there are statistically significant differences in the daily suicide rate among the months and seasons of 1970.