Analysis of the Vietnam Draft in 1970

MATH 4650 Final Project
4/27/2021
Alexander Kahanek and Jonah Turner

Introduction

The 1970 Vietnam Draft selected men born between January 1st, 1944 and December 31, 1950 by purportedly randomly ranking birthdays to create the order of the draft. They accomplished this by writing each day of the year, from 1 to 366, on a slip of paper. These paper slips were then randomly drawn from a jar one at a time. The purpose of this analysis is to judge the fairness of this draft system in randomly selecting the birthdates of the draftees.

What does our data look like?

The original dataset has four columns: *(Month)*, which is the plaintext identification of the month, *(Month_Number)*, which is the integer representation of the month, *(Day_of_year)*, which is the given day of the year out of a total 366, and *(Draft_No)*, which is the ordered integer value of when that day was chosen. We assume the draft number is not independent, as they are of ordinal value, meaning that they were chosen in consecutive order as opposed to randomly. However, we do assume the day of the years to be chosen independently and randomly. The following code was used to load our packages and data:

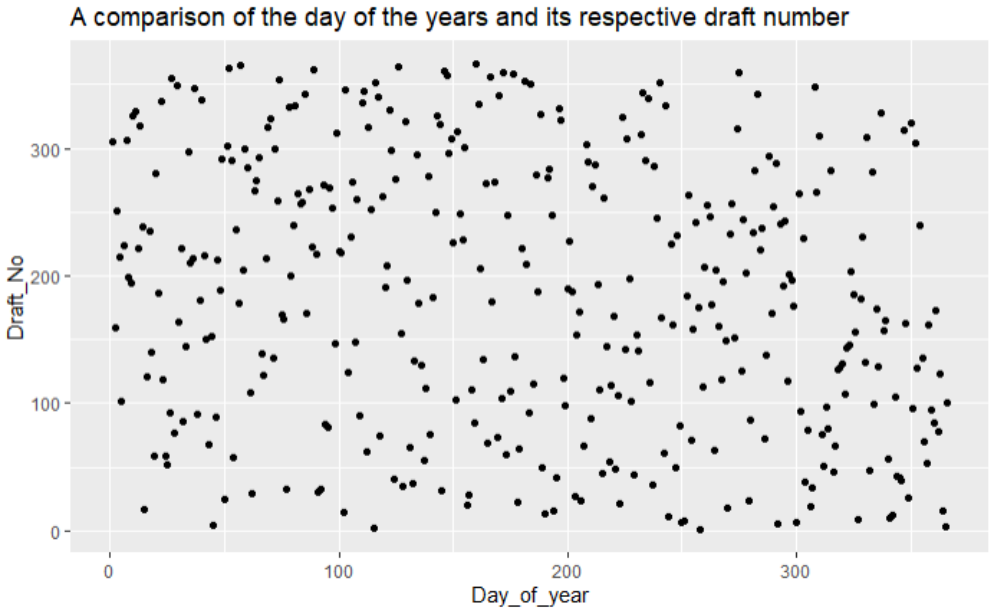
```
library(dplyr) # for piping
library(ggplot2) # for plotting
library(FSA) # for dunn's test

raw <- read.csv('1970lottery.csv')
raw %>% head()
```

Month	Month_Number	Day_of_year	Draft_No
Jan	1	1	305
Jan	1	2	159
Jan	1	3	251
Jan	1	4	215
Jan	1	5	101
Jan	1	6	224

An Analysis of the Draft Number and Day of the Year

To analyze the relationship between the draft number and the day of the year, we plotted a scatterplot of the two variables, found the Pearson's coefficient between them, and finally performed a simple linear regression. To get an idea of the relationship between the day of the year and the chosen draft number, we present the following:

```
raw %>%
  ggplot(aes(x = Day_of_year, y = Draft_No)) +
  geom_point() +
  labs(title = "A comparison of the day of the years and its respective draft
number")
```



From our scatterplot we can see that there appears to be no pattern among the comparisons of the chosen day of the year and their draft numbers. This tells us that the data was chosen with a high amount of randomness. Although, if you look close enough, you can see a slight trend of lower draft numbers for later days of the year, and slightly more higher draft numbers for the beginning of the days. Along with this scatterplot, we found the Pearson's Correlation Coefficient to be *(R = -0.226)* and *(R^2 = 0.05109)* showing a weak correlation between the two variables. This effectively means that we can weakly account for ~5% of the variance in the draft number to be affected by the day of the year, or vice versa.

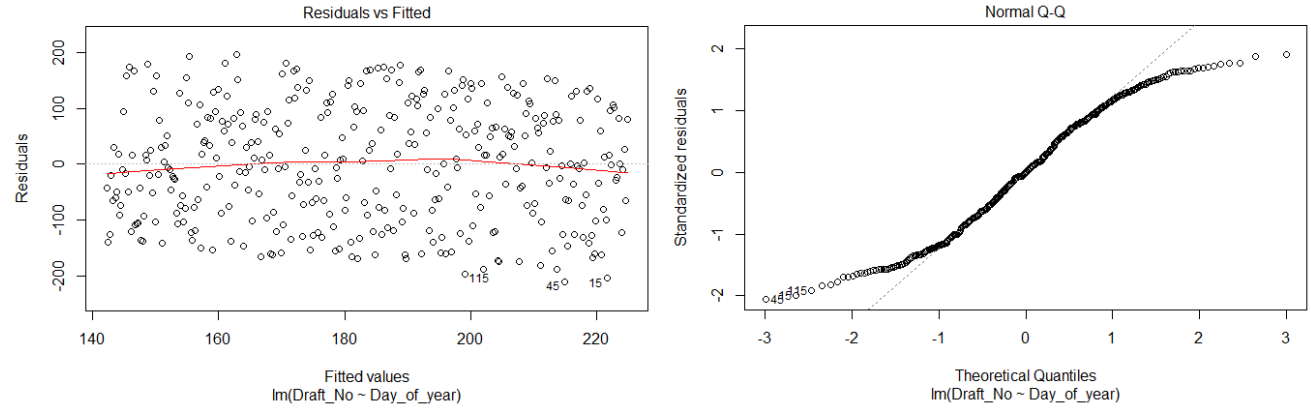
```
model <- lm(Draft_No ~ Day_of_year, data = raw)
model %>% summary()
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-210.837  -85.629   -0.519   84.612  196.157

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 225.00922   10.81197  20.811  < 2e-16 ***
Day_of_year  -0.22606    0.05106  -4.427 1.26e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 103.2 on 364 degrees of freedom
Multiple R-squared:  0.05109,   Adjusted R-squared:  0.04849
F-statistic:  19.6 on 1 and 364 DF,  p-value: 1.264e-05
```

The linear regression model was not strong, with an adjusted R-squared of .04849 and an F-statistic of 19.6. The p-value of $1.264e^{-5}$ is sufficiently low to verify the significance of the test. The linear regression did show a negative relationship between the day of the year and draft number. Meaning the intercept of ~225 is subtracted from for each day of the year times our slope. Our linear regression line would hence be $\hat{Y} = (-0.22606)X + 225.00922$, where $\hat{Y}$ is

the draft number and X is the day of the year. To further analyze the effectiveness of the linear regression, a graph of the residuals and a Q-Q plot of the residuals were produced.

```
model %>% plot()
```



From our Residuals graph, we can see a consistent spread around 0, showing that the variance appears to be stable throughout all the fitted values; however, we do notice a slight pattern of the residuals getting lower as the fitted values increase. The Q-Q graph shows us that the data is fairly close to a normal distribution between the quantiles -1 to 1. When we get past these quantiles, we notice the outliers deviate from a normal distribution heavily. This could be another indicator of the issues of our linear model, as we did notice a trend in the beginning days and lower days to have slight variations in the "true randomness".

While this linear model is relatively weak, it remains the strongest among the several transformations that were applied. Square root and log transformations were applied to each variable separately and to both variables together, and a reciprocal transformation was applied to the dependent variable, but none of the transformations resulted in models that were stronger than the original linear regression with no transformations performed. The results of each transformation are below:

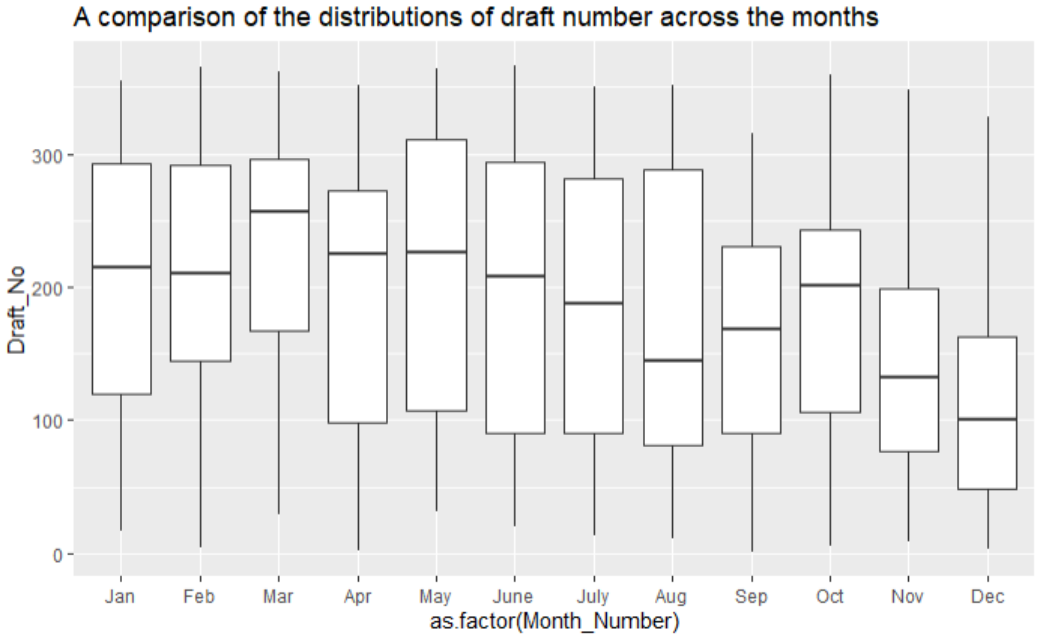
Transformations	Residual Standard Error	Adjusted R-squared	F-statistic	p-value
Draft_No ~ Day_of_year	103.2	**0.04849**	19.6	**1.264e-05**
log(Draft_No) ~ Day_of_year	0.9519	0.03698	15.02	0.0001264
sqrt(Draft_No) ~ Day_of_year	4.395	0.04575	18.5	2.186e-05
1/Draft_No ~ Day_of_year	0.06456	0.004044	2.482	0.116
Draft_No ~ log(Day_of_year)	104.4	.02702	11.14	9.334e-04

Draft_No ~ sqrt(Day_of_year)	103.7	0.03993	16.18	6.998e-05
log(Draft_No) ~ log(Day_of_year)	.9583	.02418	10.04	1.657e-03
sqrt(Draft_No) ~ sqrt(Day_of_year)	4.411	0.03881	15.74	8.769e-05

An Analysis of the Draft Number difference across months.

To determine if there was bias between the variance of ranks chosen across the months, we plotted a boxplot of the monthly distribution of draft numbers and performed an Analysis of Variance test.

```
raw %>%
  ggplot() +
  geom_boxplot(aes(x = as.factor(Month_Number), y = Draft_No)) +
  scale_x_discrete(labels = c("Jan", "Feb", "Mar", "Apr", "May", "June", "July",
"Aug", "Sep", "Oct", "Nov", "Dec")) +
  labs(title = "A comparison of the distributions of draft number across the
months")
```



The side-by-side boxplots show a bias of the draft selection in picking dates in the later months of the year earlier. This could be attributed to the procedure of the draft and the mixing of the days in a shoebox. This can be observed by the medians of each month as well as the 25th to 75th percentile being much lower and smaller than most months in November and December.

To further analyze the differences in selection order between the months, an analysis of variance test (ANOVA) was conducted on the draft numbers between the months of 1970.

```
aov(Draft_No ~ Month, data = raw) %>% summary()
```

```
             Df  Sum Sq Mean Sq F value  Pr(>F)
Month        11  290863   26442   2.466 0.00552 **
Residuals   354 3795414   10722
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA test shows that there is significantly more variance between months as opposed to within months. The F-value of 2.466 suggests that for every Mean Squared Error point within months, there are 2.466 Mean Square Error points between the months. This is statistically significant at the 0.01 level *(p = 0.00552)*, showing that it is likely there was bias in the randomization of the written slips for the days of the years.

To get an idea of how well our ANOVA test performs, we need to look at the length and variance of each month. This is because the ANOVA test is parametric and assumes homogeneity of variance, meaning the data should be normally distributed with approximately equal variances between groups. To get a grasp of our data distributions we used the following:

```
raw %>%
  group_by(Month) %>%
  summarise(
    month_number = max(Month_Number)
    ,length = n()
    ,variance = var(Draft_No)) %>%
  arrange(month_number)
```

Month <fctr>	month_number <int>	length <int>	variance <dbl>
Jan	1	31	9951.258
Feb	2	29	10806.677
Mar	3	31	9182.495
Apr	4	30	11961.954
May	5	31	13218.966
Jun	6	30	13892.478
Jul	7	31	12015.323
Aug	8	31	12708.189
Sep	9	30	7595.321
Oct	10	31	9366.523
Nov	11	30	8910.961
Dec	12	31	9036.123

From the above we see that our data has non-parametric attributes and does not exhibit homogeneity of variance. Specifically, our lengths in months differ slightly by ±1 from 30 and the variances differ heavily, especially in the later quarter of the year. Due to these reasons, we believe a Kruskal-Wallis test would be more statistically appropriate than a typical ANOVA test. To perform the test we used the following:

```
kruskal.test(Draft_No ~ Month_Number, data = raw)
```

```
        Kruskal-Wallis rank sum test

data:  raw$Draft_No by raw$Month_Number
Kruskal-Wallis chi-squared = 25.946, df = 11, p-value = 0.006611
```

With a p-value of .00661, the Kruskal-Wallis test is statistically significant at the .01 alpha level, leading us to reject the null hypothesis of the same median of Draft Number among months. The sufficiently high chi-squared value of 25.946 supports this conclusion. To further pinpoint which months statistically differ, we conducted a Dunn's test across the 12 months.

```
dunnTest(Draft_No ~ Month, data = raw, method = 'bh')
```

The Dunn's test, using the Benjamin-Hochberg p-value adjustment, shows that only certain differences among the months are statistically significant at $\alpha = .05$: December and January *(p = 0.039034328)*, December and February *(p = 0.048345322)*, December and March *(p = 0.006972504)*, December and May *(p = 0.042921439)* , and November and March *(p = 0.049305736)*. The Benjamin-Hochberg adjustment method was selected to control the False Discovery Rate, as controlling the False Discovery Rate is more reasonable in this analysis than controlling the Family Wise Error Rate. Even when using the Bonferroni adjustment method, some statistically significant differences between months remain. However, the months found by the Benjamin-Hochberg adjustment all fall within the first and last quarter of the year, further strengthening the claim that the drawing process favored selecting dates later in the year before dates earlier in the year.

Conclusion

In general, we found evidence that there are variations in the randomness between the chosen draft number for the days of the year, despite the initial scatterplot results. From the linear model we found that the best performing model was one without any transformations, yet it appeared to be a weak predictor still. Despite this, we found that most of its residuals are of a normal distribution, from the -1 to 1 quantiles, with the outliers being not normal. This is also suggested in our boxplot, which shows that there appears to be bias in the distribution of draft numbers in the beginning and later parts of the year. To confirm this, we performed an ANOVA and Kruskal-Wallis test, both of which had statistically significant results, confirming that there is a statistically significant difference of the variances and medians between months. To further pinpoint the months that varied, we used a Dunn's test, which found statistically significant differences between *{December}* and *{January, February, March, May}* as well as significant differences between *{November}* and *{March}*.

From all of these results, we can conclude with high confidence that the drafting methods used by the US government in 1970 were not statistically random. There is a bias for choosing citizens that were born in December and November to be drafted before those that were born in the beginning of the year.