Average Time Required to Start a Business

Erin Courtemanche, Megan Johnson, Gabriel Taveira | MA 214-3 | 29 April, 2024

I. Executive Summary

Entrepreneurship is possible no matter where a person is from or what their gender may be. But do these various factors still affect the amount of time it takes on average for a person to start a business?

The data set used is from the World Bank website and outlines the average time individuals in 188 countries across six continents. It includes data starting in 2003 and going up through 2019.

We have selected two models. One can be used to test if there is a significant difference between the time it takes to start a business in African countries and European countries. The other allows us to observe if there is a general difference in startup times across continents, as well as how the findings are impacted when gender is controlled.

Based on our results, we found that there is convincing evidence that Africa has a longer startup time, on average, than Europe. We also found that there is an overall difference in startup times across all continents that is amplified when controlling gender.

II. Introduction

One of the largest problems faced by entrepreneurs today is inefficient and unnecessary regulations that can make it a difficult and time-consuming process to even start a business.

The dataset used was collected by the World Bank as part of a desire to see how various regulations surrounding business startups may affect the time it takes to start businesses. The author also acknowledges that in countries where regulations are strict, informality when starting a business may increase. This is because entrepreneurs would rather subvert regulations than go through long and often costly processes. Firms that fall into this informal category tend to grow slower because they have less credit and fewer employees.

The methodology of picking the average time it takes to start a business is based on the number of calendar days it takes to complete the necessary procedures to legally operate a business. Some of these procedures can be completed faster for a cost, so the fastest procedure is selected independent of cost.

Some limitations presented by this data set are that data points are only taken from each country's largest city and may not be representative of more rural or suburban areas. Additionally, the data only focuses on limited liability companies of a prespecified size. The time measures used are also subject to judgment as when different sources report different responses, the median of the values is taken. Lastly, the methodology used operates under the assumption that businesses have full knowledge of the information required to start a business and do not waste any time on these procedures.

The survey used to collect this data is standardized by the World Bank and assumes a business' standard size (small or medium), legal form (LLC), location (a country's largest city), and nature of operation. Surveys are carried out by over 9,000 local experts such as lawyers, accountants, business consultants, etc.

III. Data Characteristics

We have cleaned the dataset to only include the years 2018 and 2019. Additionally, we have combined the male and female datasets, which were previously separate, for the sake of simplicity, understandability, and model fitting.

All variables taken together include Country (qualitative), Continent (qualitative), Gender (qualitative), and Startup Time (quantitative). We separated 2018 and 2019 data to view the two years separately, specifically since we are most interested in the most recent data from 2019.

IV. 2-Population Z-Test for Difference of Means

First, we wanted to see if there was a significant difference in the mean average startup time for a business between the continents of Africa and Europe. To this end, we decided to run a 2-population z-test for difference of means. Our null hypothesis is that the true mean time to start a business in 2019 in Europe is the same as it is in Africa. Our alternative hypothesis is that the true mean time to start a business in 2019 in Europe is different than it is in Africa.
Assumptions:

- Independent Random Sampling: The World Bank states that it collects data from each country's largest city on a random basis. This satisfies the conditions. Additionally, data obtained from one country should not affect the data obtained from another country.
- Normality: We have 54 variables for average time (in days) to start a business in 2019 in Africa, and 42 variables for average time (in days) to start a business in 2019 in Europe. Since 54 and 42 are both greater than 30, the Normality condition is met through the Central Limit Theorem.

Running the Test:

Because all assumptions are satisfied, we can run the test (see appendix for details). Since our obtained p-value, 0.0217, is less than alpha 0.05, we reject the null hypothesis; there is convincing evidence that the true mean time to start a business in 2019 in Europe is different from Africa.

Upon further research, we are 95% confident that the interval (1.0845,14.0266) contains the true mean difference in the number of days it takes to start a business in Africa than in Europe. Because this interval only contains values greater than 0, we have reason to believe that the number of days it takes to start a business in Africa is greater than in Europe.

V. One-Way ANOVA

Next, a one-way ANOVA test was conducted to test the differences between each continent in average startup times. Our null hypothesis is that the true mean startup time for business across six continents is the same for each continent. Our alternative hypothesis is that at least one of the true mean startup times for businesses across these six continents is different.

Assumptions:

- Independent Random Sampling: In this case, the World Bank states that it collects data randomly from each country's largest city on a random basis. Satisfying this condition.

Additionally, data obtained from one country should not affect the data obtained from another country.

- Normality: While the data shown in the qqplot is not perfectly normal, it is still fair to say that this data meets the normality assumption. Because of the imperfections mentioned earlier regarding the World Bank's data collection methodology (see introduction for details), it is reasonable for this plot to not be perfectly linear.
- Homoscedasticity: This is met through the distribution of various boxplots across all tests, from groups that have met Normality conditions and influenced by the same factors. The box plots for the "Startup Times by Continent" are shown in the appendix. We then create boxplots of the residuals for this data to check that the standard deviations of all residuals are constant. From the data, we see that the standard deviations of residuals are similar, so this data passes this assumption.

Running the Test:

Since all assumptions for running a one-way ANOVA test are met, we can now run the test. Since the test returns a p-value of 0.0285, which is less than a 5% significance level, we reject the null hypothesis; there is convincing evidence that at least one of the true mean startup times across the six continents is different. Although we have convincing evidence for different mean startup times, we can use our observed sample means from our data summaries for which continents take shorter and longer. From these, we notice average startup times from least to greatest are as follows: Europe (13.383), North America (18.36), Asia (18.56), Oceania (18.96), Africa (20.94), and South America (42.63). These findings can help us to prioritize assisted entrepreneurship efforts across the less efficient continents.

VI. **Two-Way ANOVA**

Additionally, we wanted to see if gender had any influence on whether there is a difference in the true mean startup times across continents when controlling gender, to see if gender is influential as a variable. To this end, we ran a two-way ANOVA test controlling gender. Our null hypothesis is that there is not a difference in the true mean startup times across continents when controlling gender. Our alternative hypothesis is that there is a difference in the true mean startup times across continents when controlling gender.

Assumptions:

- Independence: it is reasonable to assume that the data obtained from one country would not affect the data obtained from another country.
- Normality: While the data shown in the qqplot is not perfectly normal, it is still fair to say that this data meets the normality assumption. Because of the imperfections mentioned earlier regarding the World Bank's data collection methodology (see introduction for details) it is reasonable for this plot to not be perfectly linear.
- Homoscedasticity: residuals vs predicted values plot

Running the Test:

Because all our assumptions are satisfied, we can run a Two-way ANOVA test in R Studio (see appendix for details). Since our test returns the p-value, 0.000000377, is less than alpha 0.05, we reject the null hypothesis. There is convincing evidence that there is a difference in the true mean startup times across continents when controlling gender. We also see that gender can significantly influence startup time since the p-value is close to 0. This tells us that the effect of continents becomes more significant after considering the effect of gender.

VII. Conclusion

The results of this investigation suggest that both continental geography and gender are influential in the average time required to start a business. It can be reasonably inferred by our two-population z-test that entrepreneurs in Africa may face longer delays in starting a business than those in Europe. According to our one and two-way ANOVA tests, there also appears to be a difference in startup times between genders on different continents. We believe this disparity is strong enough to consider further attention through policymaking.

The findings regarding the differences in startup times for Africa and Europe suggest that more developed countries may have more streamlined procedures when it comes to business startup, which could make the process of starting a business faster on average. Some policy recommendations may support global equity overall. Our findings point to the need for focused changes in particular areas, like in Africa in this case, to further streamline the startup process. If the procedure of starting a business is made more accessible with less stringent regulations, then more entrepreneurs would have the resources to start a business and to do so in a shorter period.

This research also draws attention to gender inequality in startups, highlighting the need for more research to fully understand the underlying causes of these variations and to create gender equality-driven initiatives in entrepreneurship.

VIII. Appendix

Code for 2-population z-test on differences in Africa & Europe

```
x <- Startups.2019.Combined[Startups.2019.Combined$Continent == 'Africa' , 'Time']
y <- Startups.2019.Combined[Startups.2019.Combined$Continent == 'Europe' , 'Time']
u1 <- mean(x)
u2 <- mean(y)
var1 <- var(x)
var2 <- var(y)
n1 <- length(x)
n2 <- length(y)
SE <- sqrt((var1/n1)+(var2/n2))
TS <- (u1-u2)/SE
dfNumerator <- ((var1/n1) + (var2/n2))^2
dfDenominator <- (var1^2/(n1-1)) + (var2^2/(n2-1))
df <- dfNumerator/dfDenominator
conf <- 0.95
a <- 1 - conf
z <- qnorm(1-a/2, df)
lower <- u1-u2-z*SE
```

```
upper <- u1-u2+z*SE
### for u1 != u2 ###
2*pnorm(abs(TS), df, lower.tail = F)
```

Code for one-way ANOVA test on differences between continents

```
### conditions: independent (World Bank collection methodology), SD's are the same (Figure 2), normal distribution (Figure 3) ###
boxplot(Startups.2019.Combined$Time, main = 'Startup Times')
```

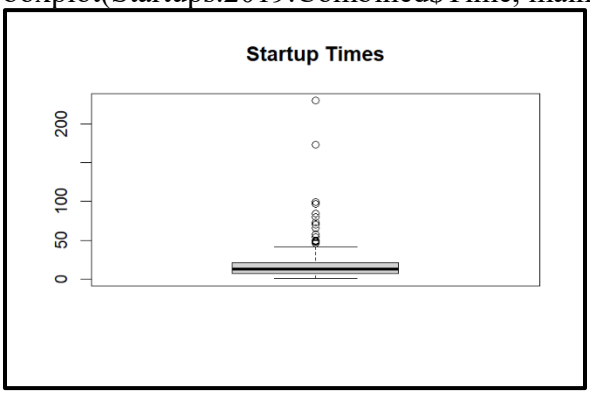

Figure 1

```
boxplot(Time~Continent, data = Startups.2019.Combined, main = 'Startup Times by Continent')
```

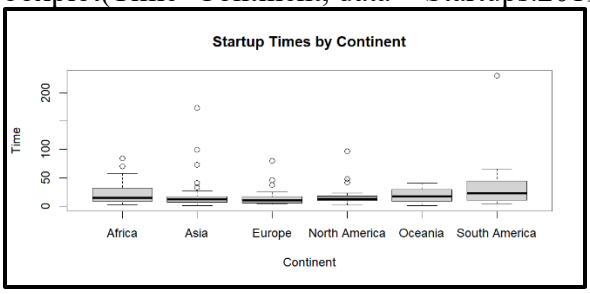

Figure 2

```
fit <- aov(Time~Continent, data = Startups.2019.Combined)
Startups.2019.Combined$Residuals <- residuals(fit)
qqnorm(Startups.2019.Combined$Residuals, main = 'Normal Q-Q Plot of Residuals');
qqline(Startups.2019.Combined$Residuals)
```

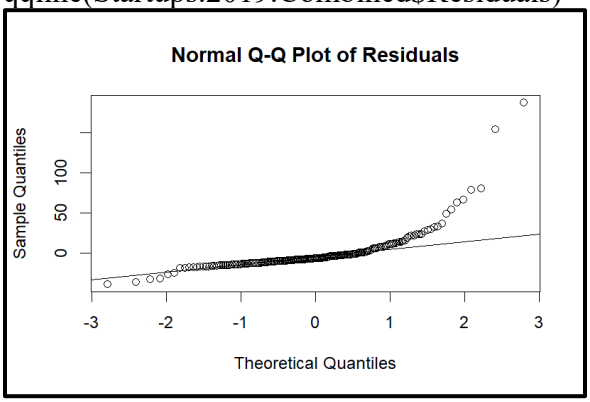

Figure 3

```
boxplot(Residuals~Continent, data = Startups.2019.Combined, main = 'Boxplot of Residuals')
```

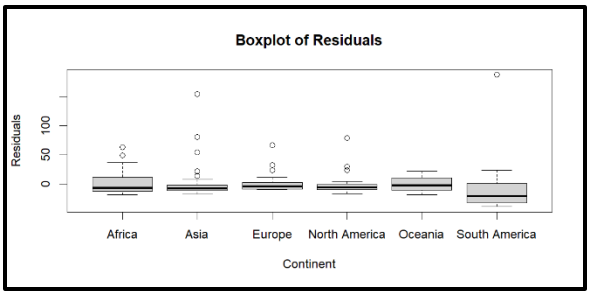

Figure 4

summary(fit)

```
             Df Sum Sq Mean Sq F value Pr(>F)
Continent     5   8178    1636   2.568 0.0285 *
Residuals   182 115931     637
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5

p-value will be in upper right corner under Pr(>F)

africa <- Startups.2019.Combined[Startups.2019.Combined$Continent == 'Africa', 'Time']
summary(africa)

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.50    8.50   13.75   20.94   31.38   84.00
```

asia <- Startups.2019.Combined[Startups.2019.Combined$Continent == 'Asia', 'Time']
summary(asia)

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00    7.00   11.00   18.56   16.50  173.00
```

europe <- Startups.2019.Combined[Startups.2019.Combined$Continent == 'Europe', 'Time']
summary(europe)

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
3.500   5.125   9.250  13.383  16.125  80.000
```

north.america <- Startups.2019.Combined[Startups.2019.Combined$Continent == 'North America', 'Time']
summary(north.america)

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.50    9.20   12.00   18.31   17.50   97.00
```

oceania <- Startups.2019.Combined[Startups.2019.Combined$Continent == 'Oceania', 'Time']
summary(oceania)

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.50    9.00   16.50   18.96   28.75   41.00
```

south.america <- Startups.2019.Combined[Startups.2019.Combined$Continent == 'South America', 'Time']
summary(south.america)

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
4.00   11.12   22.00   42.63   41.75  230.00
```

Code for two-way ANOVA test on difference in continents controlling gender

for two-way ANOVA###

difference in continents controlling gender

conditions: independent (World Bank collection methodology), SD of residuals are the same (Figure 7), residuals are normally distributed (Figure 6)

```
fit <- aov(Time~Continent + Gender, data = Startups.2019)
### order of independent.column does not matter ###
qqnorm(residuals(fit), main = 'Normal Q-Q Plot of Residuals'); qqline(residuals(fit))
```

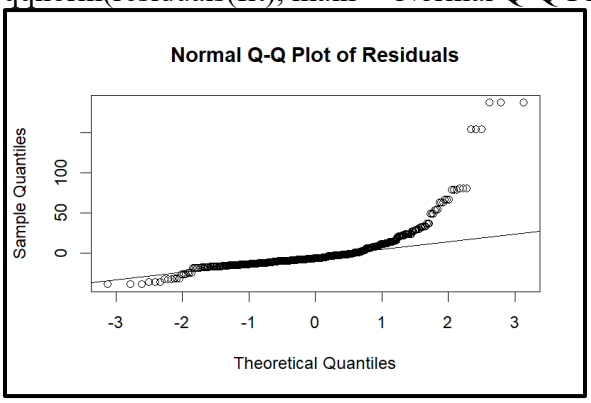

Figure 6

```
plot(predict(fit), residuals(fit), main = 'Residual vs Predicted Values Plot')
```

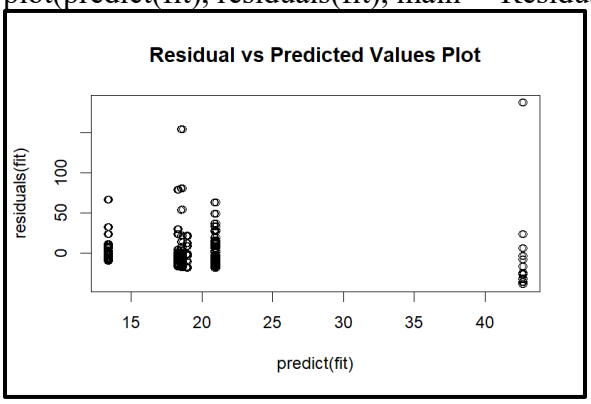

Figure 7

```
summary(fit)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Continent     5  24532    4906   7.844 3.77e-07 ***
Gender        2      1       1   0.001    0.999
Residuals   556 347753     625
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 8

###look at p-value for factor NOT being controlled ###