Exploring Global Wealth Trends Beyond Billions: An In-depth Analysis of 2023

Jaimin Shah
Goergen Institute for Data Science
University of Rochester
Rochester, USA
jshah15@ur.rochester.edu

Manasvi Patwa
Goergen Institute for Data Science
University of Rochester
Rochester, USA
mpatwa@ur.rochester.edu

Sai Mourya Buchi
Department of Computer Science
University of Rochester
Rochester, USA
sbuchi@ur.rochester.edu

Shyam Shah
Goergen Institute for Data Science
University of Rochester
Rochester, USA
sshah77@ur.rochester.edu

***Abstract*—This paper presents an in-depth analysis of global wealth trends in 2023, focusing on a comprehensive dataset of billionaires worldwide. The study employs various statistical techniques to explore patterns in wealth distribution, demographics, and industry involvement among billionaires.**

***Index Terms*—Billionaires, Wealth Trends, Statistical Analysis, Global Economy, Industry, Demographics**

I. Introduction

In this project, we investigate a comprehensive dataset encompassing 2640 billionaires worldwide for the specified year. This dataset, featuring 35 attributes, provides intricate statistical details about the billionaires' businesses, industries, and personal information. Our analysis employs rigorous random sampling techniques to extract meaningful inferences, which will be meticulously compared against the overall statistics of the dataset. This thorough examination aims to enhance our comprehension of global billionaire demographics and wealth distribution patterns.

II. Data Collection

Our study makes use of an extensive database that includes details on the billionaires around the world, providing a nuanced picture of their wealth distribution, industry, and demographics worldwide. Table 1 describes some of the key features of our dataset. In this project, we perform our tests on the sample data obtained by random sampling from the population data. We have also compared our inferences on sample data with actual population data to check the accuracy of our inferences.

TABLE I
Key Features of the Dataset

Key Features	Description
Rank	Ranking of the billionaire in terms of wealth
FinalWorth	Final net worth of the billionaire in U.S. dollars
Category	Category or industry in which the billionaire's business operates
Age	Age of the billionaire
Country	Country in which the billionaire resides
Industries	Industries associated with the billionaire's business interests
CountryOfCitizenship	Country of citizenship of the billionaire
Organization	Name of the organization or company associated with the billionaire
SelfMade	Indicates whether the billionaire is self-made (True/False)
BirthMonth	Birth month of the billionaire
BirthDay	Birth day of the billionaire
CPI_Country	Consumer Price Index (CPI) for the billionaire's country
CPI_Change_Country	CPI change for the billionaire's country
GDP_Country	Gross Domestic Product (GDP) for the billionaire's country
Gross_Tertiary_Education Enrollment	Enrollment in tertiary education in the billionaire's country
Gross_Primary_Education Enrollment_Country	Enrollment in primary education in the billionaire's country
Life_Expectancy_Country	Life expectancy in the billionaire's country
Tax_Revenue_Country_Country	Tax revenue in the billionaire's country
Total_Tax_Rate_Country	Total tax rate in the billionaire's country
Population_Country	Population of the billionaire's country
Latitude_Country	Latitude coordinate of the billionaire's country
Longitude_Country	Longitude coordinate of the billionaire's country

III. Descriptive Statistics

1) Absolute frequency table for gender is as follows,

TABLE II
GENDER COUNTS

Gender	Count
F	337
M	2303

2) Absolute frequency table for Countries is as follows,

TABLE III
COUNTRY COUNTS

Country	Count
Algeria	1
Argentina	6
Armenia	1
Australia	43
Austria	12
Bahamas	2
Belgium	4
Bermuda	2
Brazil	44
British Virgin Islands	1
Cambodia	1
Canada	42
Cayman Islands	3
Chile	6
China	523
Colombia	1
Cyprus	5
Czech Republic	9
Denmark	7
Egypt	4
Eswatini	1
Finland	7
France	36
Georgia	1
Germany	112
Greece	3
Guernsey	1
Hong Kong	68
Hungary	3
India	160
Indonesia	25
Ireland	4
Israel	26
Italy	55
Japan	39
Kazakhstan	7
Latvia	1
Liechtenstein	1
Luxembourg	1
Malaysia	15
Mexico	14
Monaco	17
Morocco	2
Nepal	1
Netherlands	10
New Zealand	2
Nigeria	3
Norway	9
Panama	1
Peru	4
Philippines	14
Poland	5
Portugal	1
Qatar	2
Romania	3
Russia	79
Singapore	46
South Africa	5
South Korea	32
Spain	25
Sweden	26
Switzerland	78
Taiwan	46
Tanzania	1
Thailand	28
Turkey	25
Turks and Caicos Islands	1
Ukraine	6
United Arab Emirates	17
United Kingdom	84
United States	755
Uruguay	1
Uzbekistan	1
Vietnam	6

3) Absolute frequency barplot of industries is as follows,

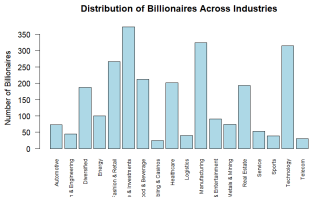


Fig. 1. Number of Billionaires in different industry

4) Absolute frequency barplot of gender is as follows,

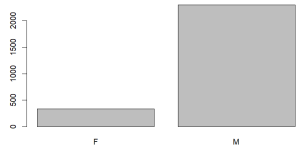


Fig. 2. Number of billionaires of each gender

5) Summaries of Center for Age is as follows,

TABLE IV
SUMMARY STATISTICS

Statistic	Value
Min.	18.00
1st Quartile	56.00
Median	65.00
Mean	65.14
3rd Quartile	75.00
Max.	101.00

6) Side by Side boxplot for gender is as follows,

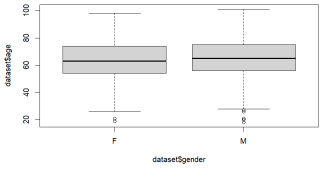


Fig. 3. Side-by-Side Boxplot

7) Scatterplot for age is as follows,

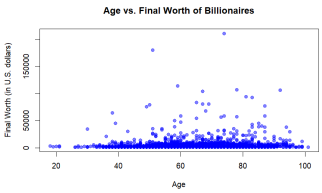


Fig. 4. Scatterplot of age vs. final worth

8) Histogram for age is as follows,

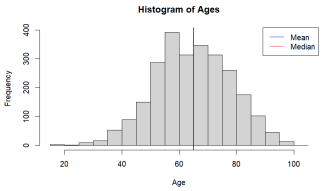


Fig. 5. Histogram of Age

9) Histogram for final worth before transformation is as follows,

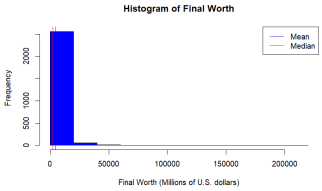


Fig. 6. Histogram of Final Worth

10) QQ plot of final worth is as follows,

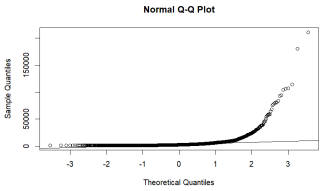
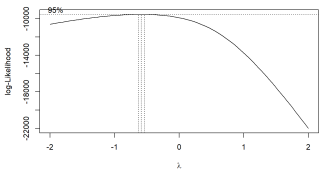


Fig. 7. Q-Q plot of final worth

11) Box Cox Transformation



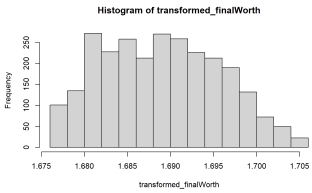


Fig. 8. After Box Cox Transformation

IV. Inferential Statistics

In the inferential statistics section of our analysis, we aim to unravel significant insights from our dataset by posing a series of questions and employing diverse statistical techniques:

A. Is average final worth of inherited billionaires equal to the average final worth of self-made billionaires?

n_1 = 1029, n_2 = 471, x_1 = 4530.321, x_2 = 5467.941

Since $n_1 \geq 30$ and $n_2 \geq 30$, we can apply central limit theorem.

$\mu_1 =$ the average final worth of inherited billionaires

$\mu_2 =$ the average final worth of self-made billionaires.

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$\alpha = 0.05$

Since we don't have population variance, this is a hypotheses test of two independent samples with unknown variances. Therefore, we will use Welch Two Sample t-test.

$$\bar{X} \sim t_{n-1}\big(\mu_1 - \mu_2, \sqrt{s_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\big)$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{1}$$

t = -1.4057

We compare this t-value with a t distribution with df = 733.82

$$\begin{aligned} \text{p-value} : \text{p} &= 2 \cdot P(T \leq t) \\ &= 2 \cdot \text{pt}(-1.4057, \text{df} = 733.82) \\ &= 0.1602 \end{aligned}$$

Since $p > \alpha$, we fail to reject the null hypothesis.

Conclusion: We don't have sufficient evidence to conclude that the average final net worth of self made billionaires and inherited billionaires is significantly different.

$$\begin{aligned} 95\%\ \text{CI} &= (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ &= (-2247.064, 371.824) \end{aligned}$$

With 95% confidence interval we can say that our true difference in mean will lie between (-2247.064, 371.824).

μ_1= 4465.397, μ_2= 4970.411

$\mu_1 - \mu_2$ = -505.0133

True difference in population means is -505.0133, which certainly lies in our confidence interval found above.

B. Is there a significant difference in the distribution of billionaires across four longitudinal regions of the Earth, as determined by a Goodness of Fit test?

C. Is there a significant association between the continent of origin of billionaires and the industries in which they are involved?

D. Is average age of billionaires in USA equal to the average age of billionaires in China?

n_1 = 433, n_2 = 288, x_1 = 66.78522, x_2 = 59.12500

Since $n_1 \geq 30$ and $n_2 \geq 30$, we can apply central limit theorem.

$\mu_1 =$ the average age of billionaires in USA

$\mu_2 =$ the average age of billionaires in China

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$\alpha = 0.05$

Since we don't have population variance, this is a hypotheses test of two independent samples with unknown variances. Therefore, we will use Welch Two Sample t-test.

$$\bar{X} \sim t_{n-1}\big(\mu_1 - \mu_2, \sqrt{s_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\big)$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{2}$$

t = 9.0803

We compare this t-value with a t distribution with df = 717.11

$$\begin{aligned}\text{p-value : p} &= 2 \cdot P(T \geq t) \\ &= 2 \cdot (1 - \text{pt}(9.0803, \text{df} = 717.11)) \\ &< 2.2e - 16\end{aligned}$$

Since $p << \alpha$, we reject the null hypothesis.

Conclusion: We have sufficient evidence to conclude that the average age of billionaires in USA and billionaires in China is significantly different.

$$\begin{aligned}95\%\ \text{CI} &= (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ &= (6.001081, 9.318364)\end{aligned}$$

With 95% confidence interval we can say that our true difference in mean will lie between (6.001081, 9.318364).

μ_1= 67.30596, μ_2= 58.69643

$\mu_1 - \mu_2$ = 8.609532

True difference in population means is 8.609937, which certainly lies in our confidence interval found above.

E. Is the variance in the final net worth of self-made billionaires less than the variance in the final net worth of inherited billionaires?

n_1 = 1029, n_2 = 471, $\frac{s_1^2}{s_2^2}$ = 0.5903325

Since $n_1 \geq 30$ and $n_2 \geq 30$, we can apply central limit theorem.

σ_1 = the variance in the final net worth of self-made billionaires

σ_2 = the variance in the final net worth of inherited billionaires

$H_0 : \sigma_1^2 \geq \sigma_2^2$

$H_1 : \sigma_1^2 < \sigma_2^2$

$\alpha = 0.05$

To compare these variances, we use one sided upper hypothesis test for two population variances which is F test.

$$F = \frac{s_1^2}{s_2^2} \sim F_{1028,\ 470}$$

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} \tag{3}$$

F = 0.59033

$n_1 - 1$ = 1028

$n_2 - 1$ = 470

$$\begin{aligned}\text{p-value : p} &= P(F < F_{\text{obs}}) \\ &= 1 \text{ - pf}(0.59033, 1028, 470) \\ &= 2.795e - 12\end{aligned}$$

Since $p << \alpha$, we reject the null hypothesis.

Conclusion: We have sufficient evidence to conclude that the variance in the final worth of self-made billionaires is less than the variance in final worth of inherited billionaires.

$$\begin{aligned}95\%\ \text{CI} &= \left(0, \frac{1}{F_\alpha} \cdot \frac{s_1^2}{s_2^2}\right) \\ &= (0, 0.6708215)\end{aligned}$$

With 95% confidence interval we can say that our true ratio of variances will lie between (0, 0.6708215).

σ_1^2 = 89402684, σ_2^2 = 112660877

$\frac{\sigma_1^2}{\sigma_2^2}$ = 0.7935557

True ratio of variances in population is 0.7935557, which lies outside our 95% confidence interval found above, which means that we need a more precise confidence interval(i.e, even lower α) to contain the true ratio of variances.

F. Is variance in the age of billionaires in US equal to the age of billionaires in China?

n_1 = 433, n_2 = 288, $\frac{s_1^2}{s_2^2}$ = 2.515894

Since $n_1 \geq 30$ and $n_2 \geq 30$, we can apply central limit theorem.

σ_1 = the variance in the age of billionaires in US

σ_2 = the variance in the age of billionaires in China

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

$\alpha = 0.05$

To compare these variances, we use hypothesis test for two population variances which is F test.

$$F = \frac{s_1^2}{s_2^2} \sim F_{432,\ 287}$$

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} \tag{4}$$

F = 2.5159

$n_1 - 1$ = 432

$n_2 - 1$ = 287

$$\begin{aligned} \text{p-value : p} &= 2 \cdot P(F \geq F_{\text{obs}}) \\ &= 2 \cdot (1 \text{ - pf}(2.5159, 432, 287)) \\ &= 4.441e-16 \end{aligned}$$

Since $p << \alpha$, we reject the null hypothesis.

Conclusion: We have sufficient evidence to conclude that the variance in in the age of billionaires in US equal to the age of billionaires in China.

$$\begin{aligned} \text{95\% CI} &= \left(\frac{1}{F_{1-\frac{\alpha}{2}}} \cdot \frac{s_1^2}{s_2^2}, \frac{1}{F_{\frac{\alpha}{2}}} \cdot \frac{s_1^2}{s_2^2} \right) \\ &= (2.031767, 3.101180) \end{aligned}$$

With 95% confidence interval we can say that our true ratio of variances will lie between (2.031767, 3.101180).

σ_1^2 = 192.1888, σ_2^2 = 84.45836

$\frac{\sigma_1^2}{\sigma_2^2}$ = 2.275545

True ratio of variances in population is 2.275545, which certainly lies in our confidence interval found above.

G. Is there a significant difference in the mean final worth of billionaires across different continents of origin, as determined by a one-way analysis of variance (ANOVA)?

Normality assumptions:

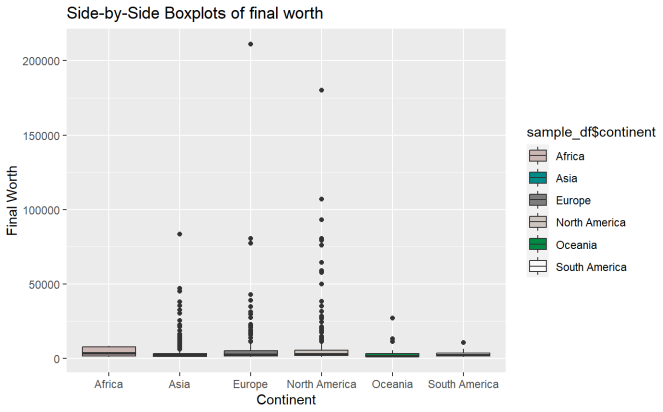


Fig. 9. Side-by-Side Boxplots of final worth

From side-by-side boxplots of final worth, we can see that the variances of final worth for different continents are drastically different. So, we can't use parametric ANOVA. We will use Non-parametric ANOVA for this question.

$H_0 : \eta_1 = \eta_2 = \eta_3 = ... = \eta_6$

$H_1 : H_0^c$

$\alpha = 0.05$

For this hypothesis, we use Kruskal-Wallis rank sum test as number of observations in each continent, $n_i > 5$:

$H \sim \chi_{k-1}^2$

$$\begin{aligned} H &= \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1) \\ &= 57.486 \end{aligned}$$

$$\begin{aligned} p &= Pr(\chi_{k-1}^2 > H) \\ &= 4.016e^{-11} \end{aligned}$$

Since $p << \alpha$, we reject the null hypothesis.

Now we use Dunn's test for post-hoc analysis.

According to bonferroni correction,

$\alpha^* = \frac{\alpha}{\binom{6}{2}} = 0.00333$

TABLE V
DUNN'S TEST (BONFERRONI)

Col Mean - Row Mean	Africa	Asia	Europe	North Am	Oceania
Asia	1.039736 1.0000				
Europe	0.346446 1.0000	-3.648319 0.0020*			
North Am	-0.229786 1.0000	-7.403688 0.0000*	-2.888914 0.0290		
Oceania	0.995316 1.0000	0.168448 1.0000	1.352879 1.0000	2.368960 0.1338	
South Am	0.677871 1.0000	-0.562746 1.0000	0.778816 1.0000	1.915113 0.4161	-0.51080 1.0000

The pairwise comparisons of Asia-Europe and Asia-North America are the ones where $p < \alpha$, therefore we reject the null hypothesis for those pairs. We have sufficient evidence to say that,

i) The median final worth of billionaires in Asia and the median final worth of the billionaires in Europe are significantly different.

ii) The median final worth of billionaires in Asia and the median final worth of the billionaires in North America are significantly different.

H. Is there a significant difference in the average age among individuals from different continents (ANOVA)?

Normality Assumptions:

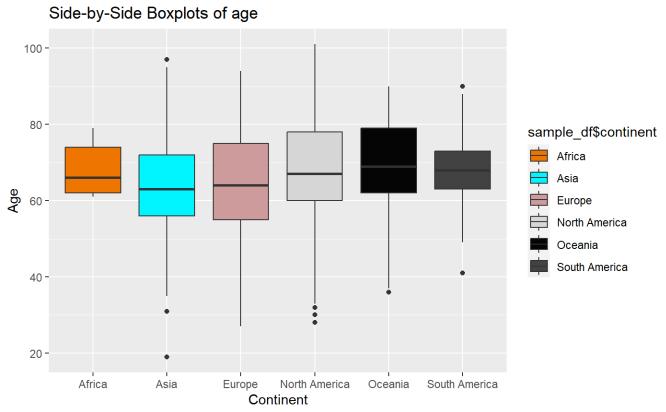
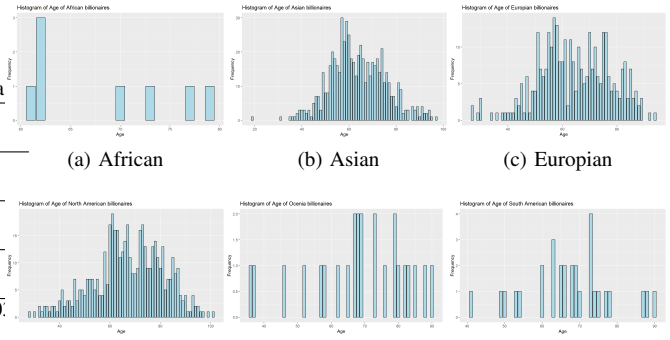


Fig. 10. Side-by-Side Boxplots of age

The equal variance assumptions looks reasonable here.



(a) African (b) Asian (c) Europian

(d) North American (e) Ocenia (f) South American

Fig. 11. Histograms of Age of billionaires

Each group follows a reasonably normal distribution as well. The observations in each group are independent as well.

$H_0 : \mu_1 = \mu_2 = ... = \mu_6$

$H_1 : H_0^c$

$\alpha = 0.05$

Applying One Way ANOVA, we get,

$s_b^2 = 795.7$

$s_w^2 = 163.4$

$F = \frac{s_b^2}{s_w^2}$

F = 4.87

p-value = Pr(>F) = 0.000201

Since $p < \alpha$, we reject the null hypothesis. We have sufficient evidence to conclude that at least one of the equalities in the null hypothesis doesn't hold. Now we use post-hoc analysis for further investigation with bonferroni correction.

$$\alpha^* = \frac{\alpha}{\binom{6}{2}} = 0.00333$$

I. Can we infer a significant correlation between the Gross Domestic Product (GDP) of a billionaire's country and their final worth?

n = 1500, r = 0.02207552

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

$\alpha = 0.05$

For inference on correlation:

$$T \sim t_{n-2}\left(0, \sqrt{\frac{1-r^2}{n-2}}\right)$$

$$t = r.\sqrt{\frac{n-2}{1-r^2}}$$

t = 1.0997

$$\begin{aligned} p &= 2 \cdot Pr(T \geq 1.0997) \\ &= 0.2716 \end{aligned}$$

Since $p > \alpha$, we fail to reject the null hypothesis.

Conclusion: We have sufficient evidence to conclude that the GDP of Billionaire's country and their final worth are linearly correlated.

$$\begin{aligned} 95\%\ \text{CI} &= r \pm t_{\frac{\alpha}{2}, n-2}\sqrt{\frac{1-r^2}{n-2}} \\ &= (-0.02224410, 0.07890135) \end{aligned}$$

With 95% confidence interval we can say that our true ratio of variances will lie between (-0.02224410, 0.07890135).

ρ = 0.04377624

True correlation in population is 0.03758884, which certainly lies in our confidence interval found above.

J. Does the distribution of billionaires' birth months and days significantly deviate from a uniform pattern, particularly with a majority born in January and the first four days of each month?

Expected count for each month = 125

Since all the expected counts are > 5, we can apply central limit theorem.

$H_0 : p_1 = 0.083, p_2 = 0.083, p_3 = 0.083, p_4 = 0.083, p_5 = 0.083, p_6 = 0.083, p_7 = 0.083, p_8 = 0.083, p_9 = 0.083, p_{10} = 0.083, p_{11} = 0.083, p_{12} = 0.083$

H_1 : at least one of these proportions does not hold

$\alpha = 0.05$

Let O be the observed frequencies

Let E be the expected frequencies

TABLE VI
OBSERVED AND EXPECTED VALUES OF ALL MONTHS

C	01	02	...	12
O	298	100	...	99
E	125	125	...	125

We will test this hypothesis using the Goodness-of-fit test,

$X^2 \sim \chi^2_{k-1}$

$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

$X^2 = 292.98$

df = 11

$$\begin{aligned} p\text{-value} &= \Pr(\chi^2 > X^2) \\ &< 2.2 \times 10^{-16} \end{aligned}$$

Since $p << \alpha$, we reject the null hypothesis.

Conclusion: We have sufficient evidence to conclude that the months are not in equal proportions.

Expected count for each day = 0.032

H_0: $p_1 = 0.032, p_2 = 0.032, \ldots, p_{31} = 0.032$

H_1: at least one of these probabilities does not hold

$\alpha = 0.05$

Let O be the observed frequencies

Let E be the expected frequencies

TABLE VII
OBSERVED AND EXPECTED VALUES OF ALL DAYS

C	01	02	...	31
O	388	35	...	13
E	0.032	0.032	...	0.032

We will test this hypothesis using the Goodness-of-fit test,

$X^2 \sim \chi^2_{k-1}$

$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

df = 30

$$\begin{aligned} p\text{-value} &= \Pr(\chi^2 > X^2) \\ &< 2.2e-16 \end{aligned}$$

Since $p << \alpha$, we reject the null hypothesis.

Conclusion: We have sufficient evidence to conclude that the probabilities for each day are not equal to 0.032.

K. Is there a significant difference in the proportion of self-made billionaires between the United States and France?

n_1 = 433, n_2 = 25, x_1 = 307, x_2 = 10

p_1 = 0.709, p_2 = 0.4

Normality Assumptions:

$n_1.p_1$ = 306.99 > 5

$n_1(1-p_1)$ = 126 > 5

$n_2.p_2$ = 10 > 5

$n_2(1-p_2)$ = 15 > 5

Thus, we can apply central limit theorem.

$H_0 : p_1 - p_2 = 0$

$H_1 : p_1 - p_2 \neq 0$

$\alpha = 0.05$

Under $H_0 : p_1 - p_2 = p$, so we can estimate their common value p

$$p = \frac{x_1 + x_2}{n_1 + n_2} = 0.6921$$

To compare these proportions, we use 2-sample test for equality of proportions without continuity correction.

$$\text{z} \sim N\left(p_1 - p_2, \sqrt{p(1-p)\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)}\right)$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)}}$$

z = 3.2546

$$\text{p-value} : \text{p} = 2 * P(Z < 3.2546) = 0.001136$$

Since $p << \alpha$, we reject the null hypothesis.

Conclusion: We have sufficient evidence to conclude that there is a significant difference in the proportion of self-made billionaires between the United States and France.

$$95\%\ \text{CI} = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} = (0.1122625, 0.5057514)$$

With 95% confidence interval we can say that our true ratio of variances will lie between (0.1122625, 0.5057514).

p_1 = 0.7152318, p_2 = 0.4166667

$p_1 - p_2$ = 0.2985651

True ratio of variances in population is 0.2985651, which certainly lies in our confidence interval found above.

In the context of comparing two proportions, it is noteworthy that the normal distribution and the chi-squared distribution are interconnected.

If the null hypothesis is not rejected in the chi-squared test, it suggests that the two variables are independent, indicating that the two proportions associated with these variables are likely the same. In other words, there is no significant difference between the two proportions.

Conversely, if the null hypothesis is rejected in the chi-squared test of independence, it signifies an association between the two variables. This implies that the two proportions are not the same, indicating a significant difference between them.

Interestingly, these conclusions obtained from chi-squared test of independence align with the conclusions drawn from a two-sample proportion test.

H_0 : Countries and self made billionaires are independent

H_1 : Countries and self made billionaires are associated

$\alpha = 0.05$

Observed Contingency Table:

TABLE VIII
OBSERVED FREQUENCIES

	Self-Made Status	
Country	True	False
United States	307	126
France	10	15

Expected Contingency Table:

TABLE IX
EXPECTED FREQUENCIES

	Self-Made Status	
Country	True	False
United States	299.69651	133.303493
France	17.30349	7.696507

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$
$$= 10.59135$$

$$p = 1 - pchisq(10.59135, 1)$$
$$= 0.00113618$$

Since $p < \alpha$, we reject the null hypothesis.

Conclusion: We have sufficient evidence to conclude that there is an association between country and self made billionaires.

L. Is the proportion of self-made billionaires in eastern countries equal to the proportion of self-made billionaires in western countries?

n_1 = 634, n_2 = 866, x_1 = 487, x_2 = 542

p_1 = 0.768, p_2 = 0.626

Normality Assumptions:

$n_1.p_1$ = 486.912 > 5

$n_1(1-p_1)$ = 147.088 > 5

$n_2.p_2$ = 542.116 > 5

$n_2(1-p_2)$ = 323.884 > 5

Thus, we can apply central limit theorem.

$H_0 : p_1 - p_2 = 0$

$H_1 : p_1 - p_2 \neq 0$

$\alpha = 0.05$

Under $H_0 : p_1 - p_2 = p$, so we can estimate their common value p

$$p = \frac{x_1 + x_2}{n_1 + n_2}$$
$$= 0.686$$

To compare these proportions, we use 2-sample test for equality of proportions without continuity correction.

$$\text{z} \sim N\left(p_1 - p_2, \sqrt{p(1-p)\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)}\right)$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)}}$$

z = 5.8539

$$\text{p-value : p} = 2 * P(Z < 5.8539)$$
$$= 4.497e^{-09}$$

Since $p << \alpha$, we reject the null hypothesis.

Conclusion: We have sufficient evidence to conclude that there is a significant difference in the proportion of self-made billionaires in eastern countries and western countries.

$$95\%\ \text{CI} = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$
$$= (0.09625292, 0.18829258)$$

With 95% confidence interval we can say that our true ratio of variances will lie between (0.09625292, 0.18829258).

p_1 = 0.754955, p_2 = 0.6366013

$p_1 - p_2$ = 0.1183536

True ratio of variances in population is 0.1183536, which certainly lies in our confidence interval found above.

In the context of comparing two proportions, it is noteworthy that the normal distribution and the chi-squared distribution are interconnected.

If the null hypothesis is not rejected in the chi-squared test, it suggests that the two variables are independent, indicating that the two proportions associated with these variables are likely the same. In other words, there is no significant difference between the two proportions.

Conversely, if the null hypothesis is rejected in the chi-squared test of independence, it signifies an association between the two variables. This implies that the two proportions are not the same, indicating a significant difference between them.

Interestingly, these conclusions obtained from chi-squared test of independence align with the conclusions drawn from a two-sample proportion test.

H_0 : Countries and self made billionaires are independent

H_1 : Countries and self made billionaires are associated

$\alpha = 0.05$

Observed Contingency Table:

TABLE X
OBSERVED FREQUENCIES

Comparison	True	False
Eastern	487	147
Western	542	324

Expected Contingency Table:

TABLE XI
EXPECTED FREQUENCIES

Comparison	True	False
Eastern	434.924	199.076
Western	594.076	271.924

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$
$$= 34.39581$$

$$p = 1 - pchisq(34.39581, 1)$$
$$= 4.497e^{-09}$$

Since $p < \alpha$, we reject the null hypothesis.

Conclusion: We have sufficient evidence to conclude that there is an association of eastern and western countries with self made billionaires.

M. Can we predict if a billionaire is self-made or not based on their final worth, age, and their country of residence using logistic regression?

Since the data of the final worth of the billionaires was too skewed, we restricted our population data to only the billionaires who had their final worth less than or equal to 5 billion $. We took a random sample of this population for logistic regression.

H_0 : Explanatory variables do not help explain log-odds response

H_1 : At least one explanatory variable helps explain log-odds response

$$\text{selfMade} \sim \text{finalWorth} + \text{age} + \text{country} \quad (5)$$

```
Call:
glm(formula = selfMade ~ finalWorth + age + country, family = "binomial",
    data = shuffled_data)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  1.742e+01  3.956e+03   0.004  0.99649
finalWorth                  -1.472e-04  5.346e-05  -2.753  0.00591 **
age                          1.046e-02  4.267e-03   2.450  0.01427 *
countryAndorra              -3.559e+01  5.595e+03  -0.006  0.99492
countryArgentina            -1.750e+01  3.956e+03  -0.004  0.99647
countryArmenia              -2.390e-01  5.595e+03   0.000  0.99997
countryAustralia            -1.706e+01  3.956e+03  -0.004  0.99656
countryAustria              -1.625e+01  3.956e+03  -0.004  0.99672
countryBahamas              -1.859e-01  5.595e+03   0.000  0.99997
countryBahrain              -3.558e+01  5.595e+03  -0.006  0.99493
countryBelgium              -1.854e+01  3.956e+03  -0.005  0.99626
countryBermuda              -3.548e+01  4.842e+03  -0.007  0.99415
countryBrazil               -1.828e+01  3.956e+03  -0.005  0.99631
countryBritish Virgin Islands -1.034e-01  5.595e+03   0.000  0.99999
countryCambodia             -2.231e-01  5.595e+03   0.000  0.99997
countryCanada               -1.660e+01  3.956e+03  -0.004  0.99665
countryCayman Islands       -1.782e+01  3.956e+03  -0.005  0.99641
countryChile                -1.741e+01  3.956e+03  -0.004  0.99649
countryChina                -1.418e+01  3.956e+03  -0.004  0.99714
countryCyprus               -1.020e-01  4.333e+03   0.000  0.99998
countryCzech Republic       -1.605e+01  3.956e+03  -0.004  0.99676
countryDenmark              -3.555e+01  5.595e+03  -0.006  0.99493
countryEgypt                -1.714e+01  3.956e+03  -0.004  0.99654
countryFinland              -1.870e+01  3.956e+03  -0.005  0.99623
countryFrance               -1.726e+01  3.956e+03  -0.004  0.99652
countryGeorgia               1.696e-01  5.595e+03   0.000  0.99998
countryGermany              -1.839e+01  3.956e+03  -0.005  0.99629
countryGreece               -1.800e+01  3.956e+03  -0.005  0.99637
countryGuernsey             -2.444e-01  5.595e+03   0.000  0.99997
countryHong Kong            -1.717e+01  3.956e+03  -0.004  0.99654
countryHungary              -2.791e-01  4.568e+03   0.000  0.99995
countryIndia                -1.808e+01  3.956e+03  -0.005  0.99635
countryIndonesia            -1.716e+01  3.956e+03  -0.004  0.99654
countryIreland              -1.673e+01  3.956e+03  -0.004  0.99663
countryIsrael               -1.707e+01  3.956e+03  -0.004  0.99656
countryItaly                -1.810e+01  3.956e+03  -0.005  0.99635
countryJapan                -1.655e+01  3.956e+03  -0.004  0.99666
countryKazakhstan           -1.562e+01  3.956e+03  -0.004  0.99685
countryLatvia                5.614e-02  5.595e+03   0.000  0.99999
countryLebanon              -1.962e-01  4.845e+03   0.000  0.99997
countryLiechtenstein        -3.532e+01  5.595e+03  -0.006  0.99496
countryLuxembourg           -3.511e+01  5.595e+03  -0.006  0.99499
countryMalaysia             -1.760e+01  3.956e+03  -0.004  0.99645
countryMexico               -1.750e+01  3.956e+03  -0.004  0.99647
countryMonaco               -1.853e+01  3.956e+03  -0.005  0.99626
countryMorocco              -3.558e+01  4.840e+03  -0.007  0.99413
countryNepal                -3.542e+01  5.595e+03  -0.006  0.99495
countryNetherlands          -1.710e+01  3.956e+03  -0.004  0.99655
countryNew Zealand          -2.680e-01  5.595e+03   0.000  0.99996
countryNorway               -1.854e+01  3.956e+03  -0.005  0.99626
countryOman                 -3.319e-01  5.595e+03   0.000  0.99995
countryPanama               -3.564e+01  5.595e+03  -0.006  0.99492
countryPeru                 -3.565e+01  4.835e+03  -0.007  0.99412
countryPhilippines          -1.852e+01  3.956e+03  -0.005  0.99627
countryPoland               -4.546e-02  4.416e+03   0.000  0.99999
countryPortugal             -3.521e+01  5.595e+03  -0.006  0.99498
countryQatar                -1.792e+01  3.956e+03  -0.005  0.99639
countryRomania              -2.860e-01  4.567e+03   0.000  0.99995
countryRussia               -1.402e-01  3.988e+03   0.000  0.99997
countrySingapore            -1.722e+01  3.956e+03  -0.004  0.99653
countrySlovakia             -1.875e-01  4.845e+03   0.000  0.99997
countrySouth Africa         -2.569e-01  4.565e+03   0.000  0.99996
countrySouth Korea          -1.808e+01  3.956e+03  -0.005  0.99635
countrySpain                -1.820e+01  3.956e+03  -0.005  0.99633
countrySweden               -1.770e+01  3.956e+03  -0.004  0.99643
countrySwitzerland          -1.731e+01  3.956e+03  -0.004  0.99651
countryTaiwan               -1.743e+01  3.956e+03  -0.004  0.99648
countryTanzania             -3.525e+01  5.595e+03  -0.006  0.99497
countryThailand             -1.747e+01  3.956e+03  -0.004  0.99648
countryTurkey               -1.767e+01  3.956e+03  -0.004  0.99644
countryTurks and Caicos Islands -3.617e-01  5.595e+03   0.000  0.99995
countryUkraine              -2.553e-01  4.333e+03   0.000  0.99995
countryUnited Arab Emirates -1.799e+01  3.956e+03  -0.005  0.99637
countryUnited Kingdom       -1.665e+01  3.956e+03  -0.004  0.99664
countryUnited States        -1.676e+01  3.956e+03  -0.004  0.99662
countryUruguay              -3.553e+01  5.595e+03  -0.006  0.99493
countryVietnam              -1.294e-01  4.271e+03   0.000  0.99998
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2485.6  on 2044  degrees of freedom
Residual deviance: 1955.9  on 1968  degrees of freedom
  (62 observations deleted due to missingness)
AIC: 2109.9

Number of Fisher Scoring iterations: 16
```

From the above table we can see that only significant values are for final worth and age. From this we can say that final worth and age are the only significant explanatory variables.

From the results of the previous regression model we can make another model.

H_0 : Explanatory variables do not help explain log-odds response

H_1 : At least one explanatory variable helps explain log-odds response

$$\text{selfMade} \sim \text{finalWorth} + \text{age} \tag{6}$$

TABLE XII
COEFFICIENTS

Variable	**Estimate**	**Std. Error**	**Pr($>\|z\|$)**
(Intercept)	1.513×10^{0}	2.587×10^{-1}	5.848×10^{-9} ***
finalWorth	-1.324×10^{-4}	4.607×10^{-5}	0.00406 **
age	-5.391×10^{-3}	3.730×10^{-3}	0.14831
Signif. codes	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1		

TABLE XIII
MODEL SUMMARY

Metric	**Value**
(Dispersion parameter for binomial family taken to be 1)	
Null deviance	2485.6 on 2044 degrees of freedom
Residual deviance	2474.3 on 2042 degrees of freedom
(62 observations deleted due to missingness)	
AIC	2480.3
Number of Fisher Scoring iterations	4

From the result found above, we can say that final worth is the only significant explanatory variable here.

Upon comparing two models, based on the AIC values of the two models we can see that the first model was better, even though country was not useful.