STAT 6730 Final Project: Data Exploration with Gapminder

Shuang Liu, Qiwei Men

`Gapminder` contains population, GDP per Capita and life expectancy for 142 countries in 5 continents from 1952 to 2007 in increments of 5 years. The first part of the project describes how development history unfold differently on different continents and in different countries over the 5 decades. Of particular interest is which countries are the leaders in this racing of development, and which countries strive in chasing the leaders. The second part of the project reveals the unusual dropping patterns and the relevant history. Although many countries experienced growth in the 5 decades, some countries encountered significant drops due to political events or diseases. The third part, which has a deviant interest, explores whether the leading digit in "naturally occurring numbers" (i.e., population and GDP per Capita) follows the famous Benford's law. Techniques used in part I and II are pipes, data manipulation and visualization. Part III employs Monte Carlo simulation.

1.1 Population

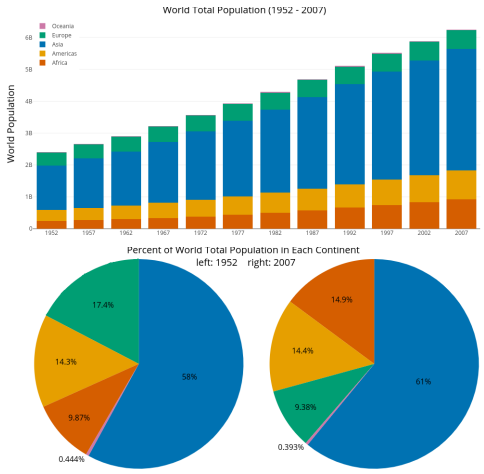

Figure 1: Change of World Population

From 1952 to 2007, world population has experienced the fastest growth in human history. The total population has grown from 2.4 billion to 6.3 billion (Fig.1 upper panel). Population in each continent grew gradually, while Africa grew faster than Europe. In 1952, Africa occupies 9.87% percent of the world population. In 2007, this percentage has increased to 15%. In 1952, Europe occupies 17.4%, while in 2007 its percentage has decreased to 9.38%. The share of Asia, Americas and Oceania has been relatively stable. (Fig.1 lower panel)

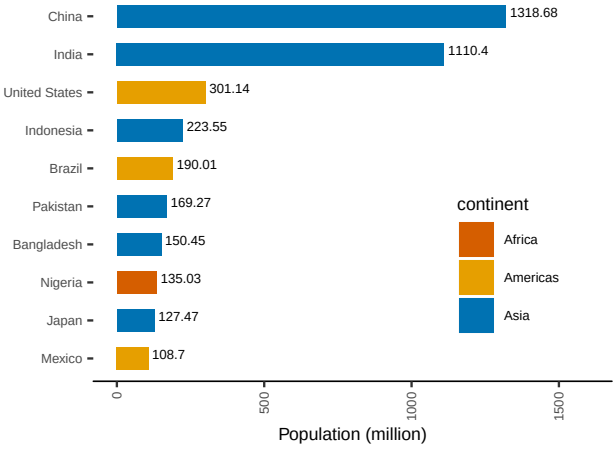

Figure 2: 10 countries with the largest population (2007)

As of 2007, Asian countries accounted for more than half of the 10 most populous countries in the world (Fig.2). East and South Asia are still the most densely populated regions. The gap between China and India, which are the two most populous countries, is narrowing (supplemental animations, animation 2).

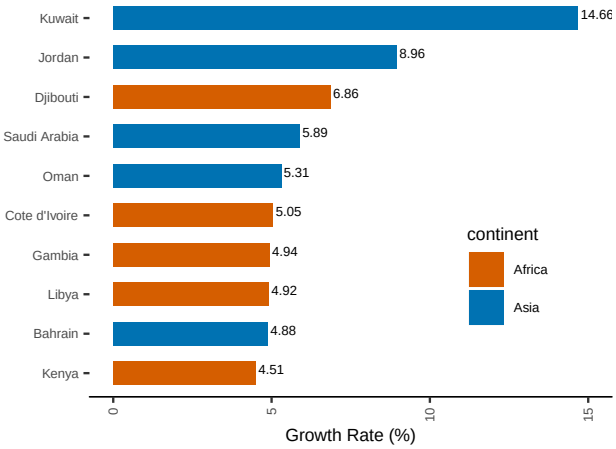

Figure 3: 10 countries with the fastest population growth (1952 - 2007)

In terms of growth rates, countries with the fastest population growth are all from Asia and Africa (Fig.3), which is consistent with the trend we see in Fig.1. An interesting phenomenon is that the Middle East countries, especially the oil-producers such as Kuwait, Jordan, Saudi Arabia and Oman have experienced the most rapid growth, mainly due to the rapid develop-

ment of the oil industry and the influx of immigrants after World War II.

1.2 GDP per Capita

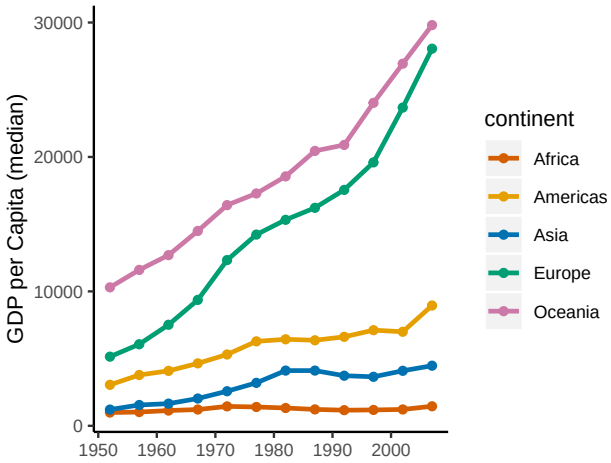

Figure 4: Change in median GDP per Capita in each continent (1952 - 2007)

By observing the changes in GDP per Capita over the 5 decades, we can see that world wealth was growing continuously, especially in European countries and Oceania countries (Fig.4). This is mainly because of the US assistance and the process of European integration after the war. On the other hand, the economic in Africa had almost stagnated. The gap between different regions was expanding, which also indicates the problem of development imbalance.

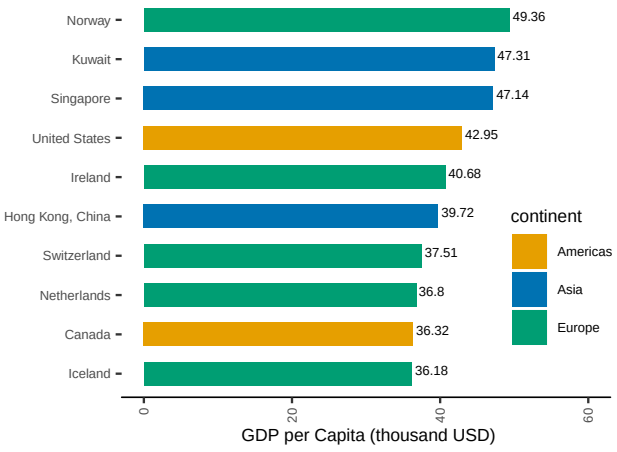

Figure 5: 10 countries with the largest GDP per Capita (2007)

Half of the top 10 countries with the highest GDP per Captia are from Europe, followed by Asia and North America, which are the most developed areas in the world (Fig.5).

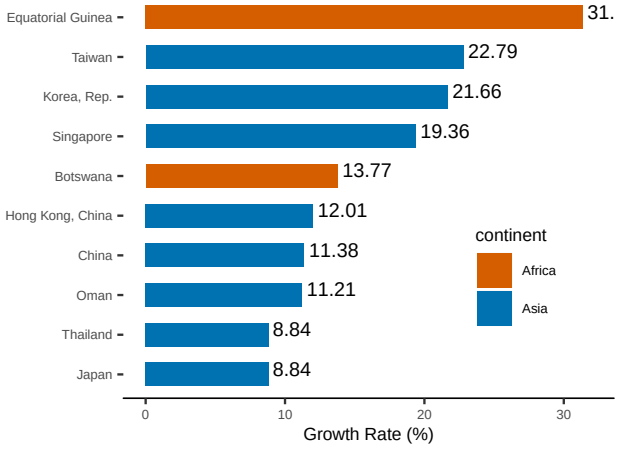

Figure 6: 10 countries with the fastest growth rate of GDP per Capita (1952 - 2007)

On the other hand, the fastest growing economies are mainly from Asia and Africa (Fig.6). Emerging economies in East and Southeast Asia like South Korea, Taiwan, Singapore, Hong Kong experienced a rapid growth. They were called The Four Asian Dragons because of this achivement.

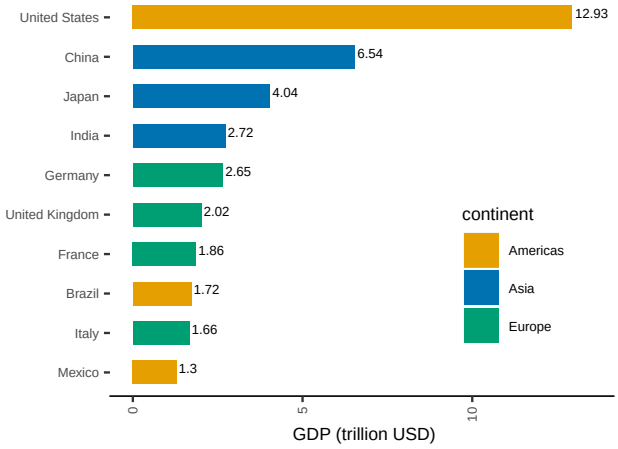

Figure 7: 10 countries with the largest GDP (2007)

In terms of total GDP, the United States is still the largest economy in the world as of 2007 (Fig.7). Total GDP shows strength of countries. The relationship between them profoundly affects the political and economic structure of the world.

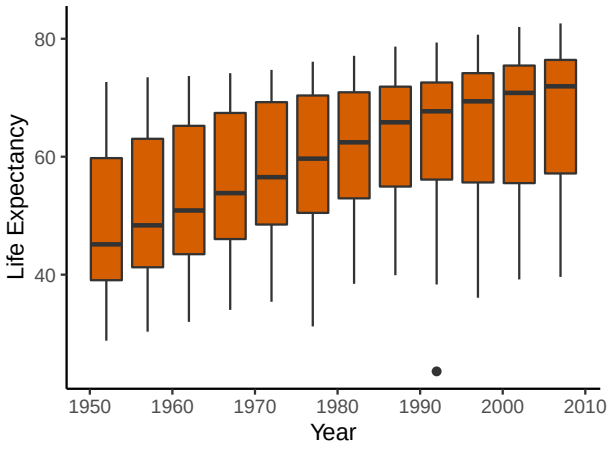

Figure 8: Life Expectancy Growth (1952-2007)

1.3 Life Expectancy

Life expectancy increased smoothly in the 5 decades. The median is rising in the box, which implys that the distribution of life expectancies has changed from a right-skewed distribution to a left-skewed distribution (Fig.8). More people live a longer life because of the development of medical and health services .

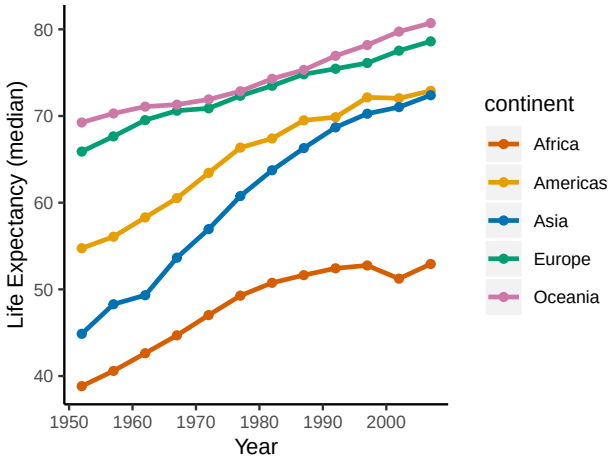

Figure 9: Change of median life expectancy in each continent (1952 - 2007)

Not like what we saw in GDP per capita, the discrepancy in life expectancies between continents is getting smaller, except Africa (Fig.9). All the other four continents have improved their median life expectancies to over 70 years, while that of Africa is still below 60. From Fig.10 and Fig.11, we can see that the leaders of life expectancy are very similar to that of GDP per Capita. Developed European countries dominate the richest and longest-lived ranking list, while Asian and African countries experienced the greatest improvement.

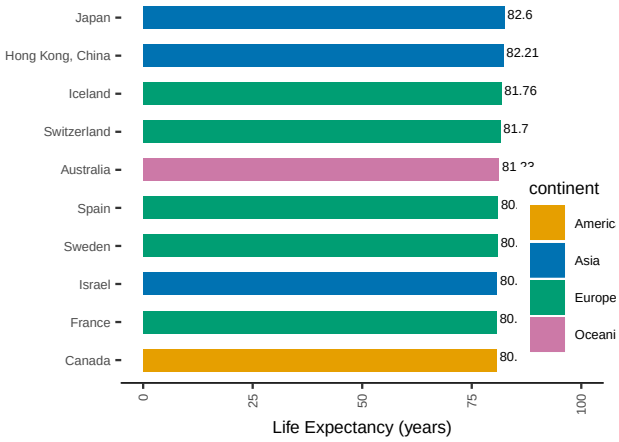

Figure 10: 10 countries with the highest life expectancy (2007)

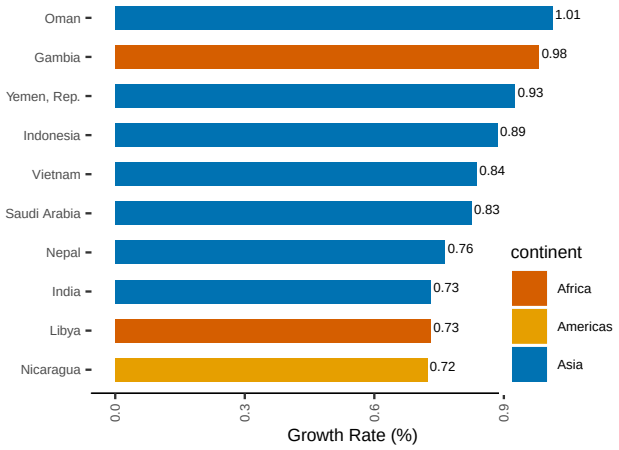

Figure 11: 10 countries with the greatest life expectancy growth (1952 - 2007)

2 Unusual Patterns

Almost half of European countries experienced a dramatic decrease of GDP per capita at the begining of 1990s (Fig.12, Fig.13). This is because of the break-up of Eastern bloc countries in 1989. Former socialist states struggled to adapt to free-market systems after they changed their political systems. Germany is the only expection in this recession. The unification masks this process happened in East German.

HIV / AIDS epidemic is taking a devastating toll on the population of many sub-Saharan countries. In the nine countries with an adult HIV prevalence of 10

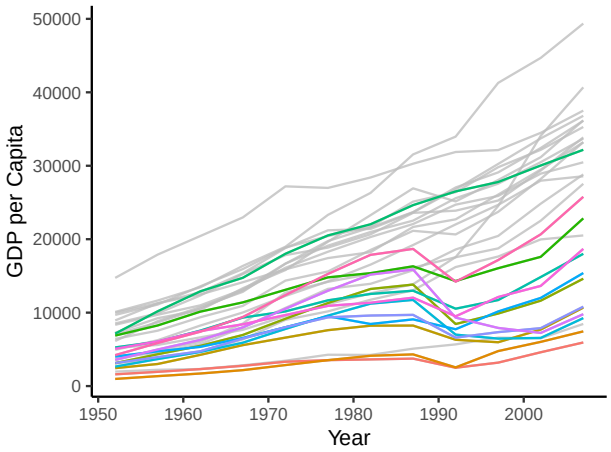

Figure 12: GDP per Capita change in European countries (colored: former socialist states; gray: non-socialist states)

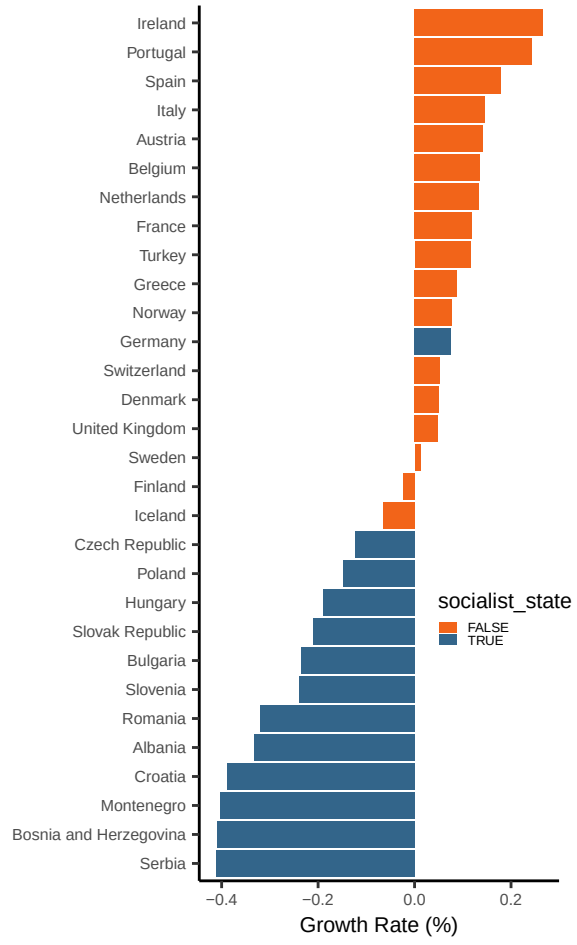

Figure 13: GDP per Capita growth of European countries (1987-1992)

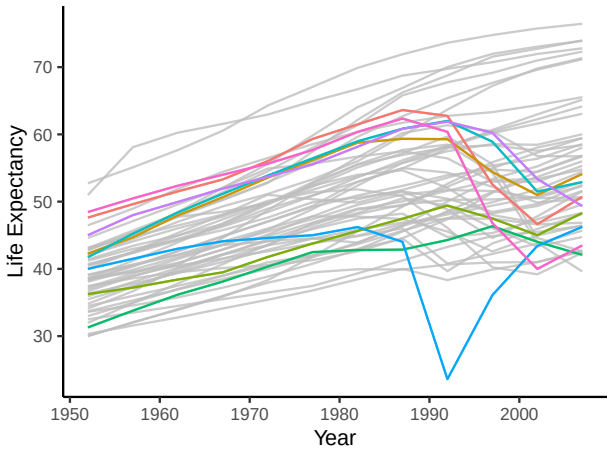

Figure 14: Life expectancy change in Africa countries (colored: HIV prevelance > 10%; gray: HIV prevelance <= 10%)

percent or more (Botswana, Kenya, Malawi, Mozambique, Namibia, Rwanda, South Africa, Zambia and Zimbabwe), the impact of AIDS is even more dramatic: more than 10 years of life expectancy have already been lost to AIDS (Fig.14).

3 Benford's Law

The law states that in many naturally occurring collections of numbers, the leading significant digit is likely to be small. For example, in sets that obey the law, 1 appears as the leading significant digit about 30% of the time, while 9 appears as the leading significant digit less than 5% of the time. More formally, a set of decimal numbers is said to satisfy Benford's law if the leading digit d ($d \in \{1, ..., 9\}$) occurs with probability $P(d) = log_{10}(\frac{d+1}{d}) = log_{10}(1 + \frac{1}{d})$. The leading digits in such a set thus have the distribution shown in Fig.15.

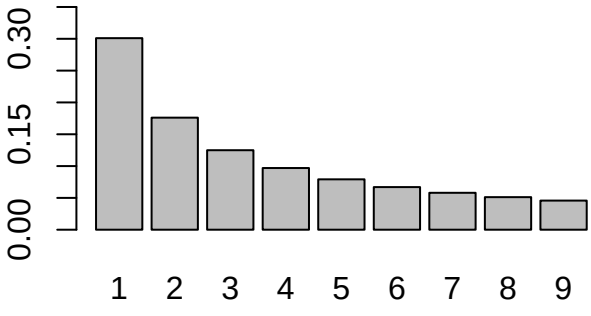

Figure 15: Benford's Law

Fig.16 compares the observed distribution of the leading digits in population in 2007 to the distribution

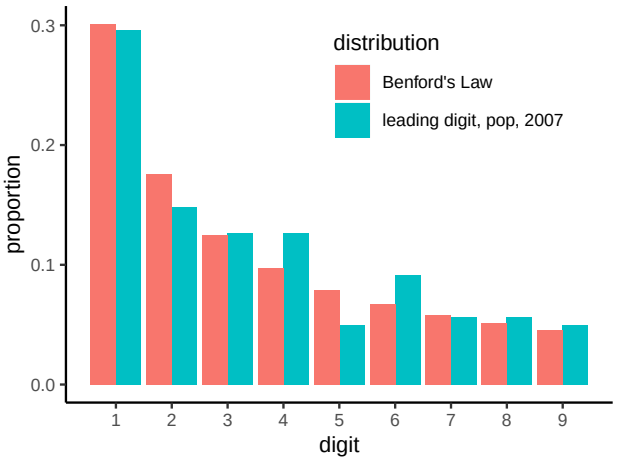

Figure 16: Compare the distribution of leading digit of population in 2007 to Benford's law

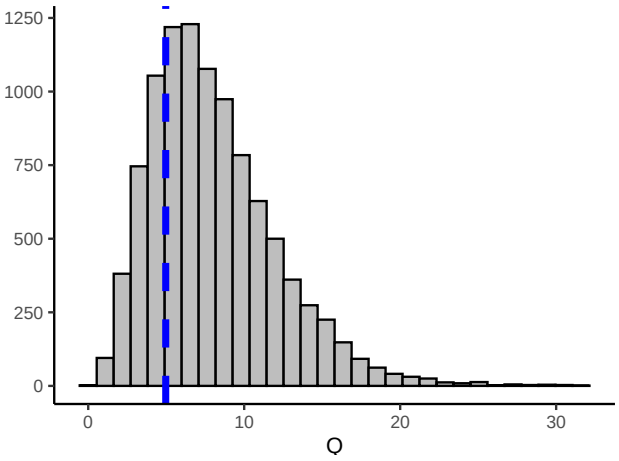

Figure 17: Sampling distribution of Q and the critical value

predicted by Benford's law. At first glance, the two distributions are quite similar. What statistical test may be used to test whether the leading digit of population in 2007 follows Benford's law or not? Chi-square Goodness-of-fit test might be a good choice. It tests if a sample of data came from a population with a specific distribution. It can be applied to discrete distributions. It can be used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.

- H_0: The leading digits of population in 2007 follows Benford's law.
- H_a: The leading digits of population in 2007 does not follow Benford's law.
- Chi-square Statistic: $Q_{k-1} = \sum_{i=1}^{k} \frac{(O_i - np_i)^2}{np_i}$ The larger the value of this statistic, the larger the discrepancy between the two distributions being tested.
  - k: number of digits (=9)
  - $k-1$: degree of freedom (=8)
  - O_i: observed count of digit i
  - n: total number of observations (= 142)
  - p_i: probability of digit i predicted by Benford's law.

Let's estimate the p-value by Monte Carlo simulation. 10,000 samples of digits of size 142 were drawn from Benford's law. For each sample, the value of Q was calculated. Fig.17 shows the sampling distribution of Q. From the observed sample, we calculated the value of Q, which is the critical value. It is shown as a blue vertical line in Fig.17. The estimated p-value is the "area" under the the sampling distribution to the right of the critical value. The critical value is **4.97**. The estimated p-value is **0.76**.

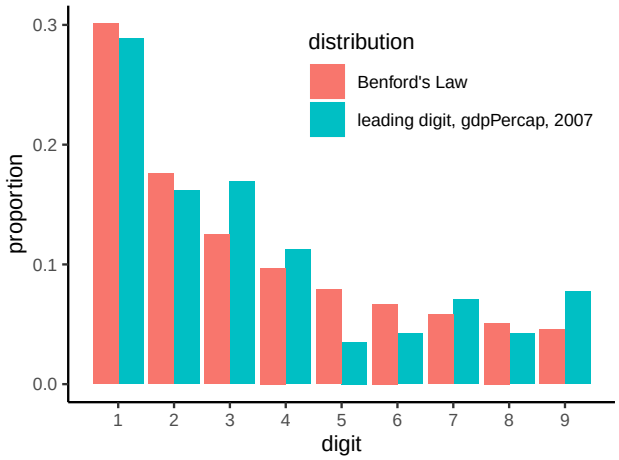

Figure 18: Compare the distribution of leading digit of gdpPercap in 2007 to Benford's law

The same comparison was made between the leading digit of GDP per Capita in 2007 to Benford's law (Fig.18), and the same Chi-square test was carried out (Fig.19). The critical value is **11.28**. The estimated p-value is **0.18**.

Both p-values can be considered large, indicating that the leading digits in both empirical distributions follow Benford's law well. P-value is an indicator of similarity here. It seems the leading digits of population in 2007 follows Benford's law more closely than that of GDP per Capita 2007. One can more or less confirm this by visually checking Fig. 16 and Fig. 18.

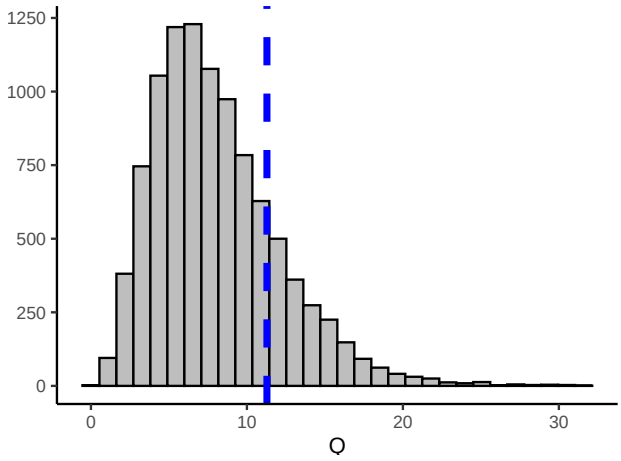

Figure 19: Sampling distribution of Q and the critical value

References

- WORLD DEMOGRAPHIC TRENDS : REPORT OF THE SECRETARY-GENERAL
  - Symbol:E/CN.9/1999/5
  - Session / Year:32
  - Agenda Item(s):6Distribution:GEN
  - Area:UNDOC
- https://en.wikipedia.org/wiki/Benford%27s_law
- https://en.wikipedia.org/wiki/Eastern_Bloc