Applying Advanced Statistics to Evaluate the Fortune 1000 Dataset

Ahmed Malik

Stockton University

CIST-3327 Probability and Applied Statistics

Byron Hoy

14 December 2023

Abstract

The statistical relevance of a dataset including the Top 1000 revenue-generating corporations from Fortune is examined in this research. The study explores the financial indicators of these top companies using various analytical methods, statistical models, and methodologies to uncover underlying trends and linkages. It delves into conditional probabilities, joint and marginal probability distributions, and applies important statistical concepts like Tchebysheff's theorem and the characteristics of various probability density functions. By doing this, the research aims to clarify the relationship between financial indicators like profit and revenue, as well as the ways in which they affect businesses. This thorough statistical analysis offers insights into the larger economic trends and business dynamics in addition to highlighting the Fortune 1000 businesses' present financial situation.

Applying Advanced Statistics to Evaluate the Fortune 1000 Dataset

Living in a capitalist society, like the United States of America, gives the chance to develop large corporations. These companies earn massive amounts of money and hold large amounts of equity. These numbers are reported and are public, so looking at them can give us an understanding of which companies are the top earners in the country. Fortune magazine publishes the renowned Fortune 500 yearly, which ranks the top 500 companies in the US based on revenue. The superset of the Fortune 500 is the Fortune 1000, which provides a larger data set and more insight into the top companies in the country. From this dataset, we can derive many interesting notes, facts, and statistics about the companies listed.

Every company has either stayed in its rank (denoted S), gone down in rank (denoted D), or has gone up in rank (denoted U). In addition, each company either has a positive (denoted +) or negative profit (denoted -). We can list the sample space as, (S, +): Stayed the same in rank, positive profit, (S, -): Stayed the same in rank, negative profit, (D, +): Gone down in rank, positive profit, (D, -): Gone down in rank, negative profit, (U, +): Gone up in rank, positive profit, (U, -): Gone up in rank, negative profit. With these 6 events defined, we can find the probability of each by dividing the number of companies that fit each event by the total number of companies in the dataset. The probabilities are:

Stayed the Same in Rank, Positive Profit **(S, +)**: Probability is 5.8%

Stayed the Same in Rank, Negative Profit **(S, -)**: Probability is 2.2%

Gone Down in Rank, Positive Profit **(D, +)**: Probability is 33.1%

Gone Down in Rank, Negative Profit **(D, -)**: Probability is 4.1%

Gone Up in Rank, Positive Profit **(U, +)**: Probability is 49.7%

Gone Up in Rank, Negative Profit **(U, -)**: Probability is 5.1%

With this information, we can find obscure probabilities like, what would the probability of a random company chosen in the data set either staying in the same rank or gone down in rank, and having positive profit? The answer to that would be 38.9%.

Using these probabilities, we can also create a table to help visualize the events in totality

	Positive Profit	Negative Profit	Total
Stayed in the Same Rank	5.8	2.2	8
Down in Rank	33.1	4.1	37.2
Up in Rank	49.7	5.1	54.8
Total	88.6	11.4	100

Let's clearly define event A to be any change in rank, so both going up and down but not staying in the same rank. Let event B represent the company made positive profits, and event C represent the company made negative profits. We can determine if event A is independent from events B and C by using the independence formula.

$P(A \text{ and } B) = P(A) \times P(B) \rightarrow 82.8\% \neq 81.512\%$

$P(A \text{ and } C) = P(A) \times P(C) \rightarrow 9.2\% \neq 10.488\%$

So, we can conclude that the events are not independent.

Let's conduct a sample, from the 1000 companies ranked, we can sample any 100 companies. To find the total number of samples of 100 companies that could be selected,

$\frac{1000!}{100!(1000-100)!} = 6.39 \times 10^{139}$, this is just the base amount of samples that could be created out of the 1000 companies, to get a more in-depth sample, we can further increase our specifications. In the data set, we see that 114 companies are unprofitable. If 100 companies are selected, we can find the probability that the sample consists of exactly 70 profitable companies and 30 that are unprofitable. Probability = $\frac{C^{114}_{30} \times C^{886}_{70}}{C^{1000}_{100}} = 4.81 \times 10^{-8}$. This indicates that this is a very rare event under the given conditions.

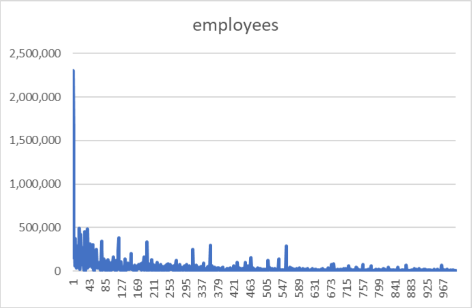
In the top 10 companies listed, 3 companies have a market value in the trillions (Amazon, Apple, and Alphabet), while the other 7 have a market value of less than a trillion dollars. Large corporations are usually subjugated to various laws and regulations. Suppose it is reported that 40% of Trillion-Value Companies and 20% of Sub-Trillion-Value Companies are in favor of adopting a new environmental sustainability policy. A company chosen at random from these top 10 is found to favor the sustainability policy. To determine the chance that a randomly picked company from the top ten, which supports a new environmental policy, is a 'Trillion-Value Company', we use a straightforward statistical method called Bayes' Theorem. This approach combines what we know about the overall makeup of the top ten companies (3 out of 10 are Trillion-Value Companies) with the information on which companies support the policy (40% of Trillion-Value Companies and 20% of the others). With Bayes' Theorem, we calculate the probability that a company that supports the policy is a Trillion-Value Company. The calculation tells us there's about a 46.15% chance that a company from the top ten that supports the policy has a market value in the trillions.

Examining different parts of the data set can provide interesting information even if it is simple in nature. When examining the “profits_percent_change” column, which represents the profits changed in 2021 over the previous year, we can get a simple understanding of the probability distribution for any negative value in this column. Just for simplicity's sake, we can take a look at the first 20 companies on the list. Let Y denote the position in the list at which each company with negative profits is found. The probability distribution for Y would be $P(Y = i) = 0$ for all positions i except for 9 and 12. $P(Y = 9) = 0.5$ and $P(Y = 12) = 0.5$. Since these are the only positions in the first 20 companies with a negative profits percent change, the probability is only distributed over two positions, Y = 9 and Y = 12. This number would be smaller when examining all 1000 companies because the probabilities would be distributed over 142 (the total number of negative values in the negative profits percent change) positions instead of only 2.

Using our findings above, we can go into more detail by setting up a binomial probability distribution. Say, for example, consider a random sample of 30 companies selected from the Fortune 1000 dataset. It is known that 142 out of the 1000 companies in the dataset have a negative profits percent change. Using a binomial distribution, we can calculate the probability of finding exactly 5 companies with a negative profits percent change in this sample of 30 companies. Setting up the equation, P(X = 5) = $(c_5^{30})0.142^5(1 - 0.142)^{30-5} = 0.1788$. So, the probability of finding exactly 5 companies with a negative profit percent change in any random sample of 30 companies from the dataset is 17.88%.

Analyzing of the top 100 companies from the Fortune 1000 list, a geometric probability distribution can be applied to understand the likelihood of encountering companies with revenues exceeding $50 billion. With a calculated probability of 'success' being 73%, the probability of encountering the first company exceeding this revenue threshold is highest at the very top of the list, at 73% for the first rank. However, as we move down the ranks, this probability diminishes sharply, dropping to 19.71% by the second rank, and further decreasing to a mere 5.32% by the third. This trend continues, showcasing a rapid decline in likelihood, whereby the tenth rank, the probability is just about 0.0006%. This analysis underscores a significant characteristic of the top Fortune 1000 companies - a higher concentration of extremely high revenue companies at the top, with a swift decrease as we progress through the ranks. This pattern is a classic demonstration of the geometric distribution, where the probability of encountering a 'success' diminishes as one moves further down the list of trials, in this case, the company ranks

Initially, the number of employees in each company might seem like a mundane detail. However, a deeper analysis shows fascinating insights in examining the top 100 companies. We observe a trend where companies tend to have fewer employees as their rank increases.



9% of these companies (the top 100) have fewer than 10,000 employees, with this likelihood increasing as we move towards higher ranks (or lower in the list). In this context, if we begin our examination from rank 100 and move upwards, the negative binomial distribution can be applied to predict the probability of encountering a specific number of companies with fewer than 10,000 employees. For instance, we might ask: What is the probability that starting from rank 100 and moving upwards, the first company with fewer than 10,000 employees will be found on the second trial? This translates to $P(X = 2) = (1 - 0.09)^{2-1} \times 0.09 = 8.19\%$.

A different but also interesting aspect of the data set is the assets column in which rank (especially towards the middle of the data set) doesn't seem to indicate how much a company can have in assets. All different types of companies have different amounts of assets. Focusing on the bottom 100, we discover an interesting dynamic when applying the hypergeometric distribution to their asset values. Within this subset, 66 of 100 companies have assets valued at less than $5 billion, we can designate this as 'defective' for the sake of this analysis. Given this scenario, we explore the probability of randomly selecting five companies from this group and finding that all of them are 'defective' (i.e., each having assets less than $5 billion). Applying hypergeometric we get $P(X = 0) = \frac{c_{0}^{66} \times c_{5-0}^{100-66}}{c_{5}^{100}} = 11.33\%$. This probability shows that many companies in the lower ranks of the Fortune 1000 list have fewer assets. This helps us understand how assets are distributed among these companies. This kind of information is can be useful for making investment decisions, analyzing the market, and conducting economic research

In applying Tchebysheff's theorem to the assets of Fortune 1000 companies, we find insightful bounds on asset distribution. With an average asset value of 60,632 and a standard deviation of 264,239.34, we use the theorem to determine the spread of assets. Tchebysheff's theorem states that at least $1 - \frac{1}{k^2}$ of data lies within k standard deviations of the mean. For $k = 2$, this is $1 - \frac{1}{4} = 0.75$ or 75%. Therefore, at least 75% of the companies are expected to have assets within 2 standard deviations of the mean. With 1000 companies in total, this translates to at least 750 companies having assets in the range of 60,632 ± 2×264,239.34. This analysis helps us understand the asset distribution among the top companies.

We can model company revenue distribution among the Fortune 1000 companies. Firstly, we have to construct a hypothetical probability density function (PDF), f(y), for their revenue, denoted as Y in hundreds of billions of dollars. This PDF will be defined as:

$f(y) = y,\ 0 \leq y \leq 0.5$ for revenues up to \$50 billion, and

$f(y) = 1,\ 1 \leq y \leq 1.5$ for revenues between \$50 billion and \$150 billion.

To analyze this distribution, we first calculated the cumulative distribution function (CDF), F(y).

For the range 0 to 0.5, the CDF is $F(y) = \int_{0}^{y^2} ydy = \frac{y^2}{2}$, evaluating this at 0.5 gives us 0.125.

For the range of 0.5 to y, the CDF is $F(y) = \int_{0}^{y^2} 1dy = y - 0.5$. We have to combine both integrals to get the CDF for y in the range $0.5 \leq y \leq 1.5$, which results in $F(y) = 0.125 + (y - 0.5)$. Using this CDF, we calculated the probability of companies having revenues in specific ranges. For example, the probability of a company's revenue being between \$0 and \$50 billion (0 ≤ Y ≤ 0.5) was found to be 12.5% (F(0.5) = 0.125). Similarly, the probability of a company's revenue falling between \$50 billion and \$120 billion (0.5 ≤ Y ≤ 1.2) can be calculated, $P(0.5 \leq y \leq 1.2) = F(1.2) - F(0.5) = 0.825 - 0.125 = 0.7$, which means the probability of this scenario is 70%. We can go further with this model and calculate the expected value. $E(Y)_1 = \int_{0}^{0.5} y^2 dy = \frac{1}{24} and\ E(Y)_2 = \int_{0.5}^{1.5} ydy = 1$, so the expected is $E(Y) = E(Y)_1 + E(Y)_2 = \frac{25}{24}$. This gives us the average revenue (in hundreds of billions of dollars) expected based on our hypothetical distribution.

Continuing the hypothetical scenarios, we can represent the proportions of a company's total annual revenue and profit, respectively, earned on a particular workday. These proportions are bounded between 0 and 1. The joint behavior of Y1 and Y2 can be described by a density function $f(y_1, y_2) = y_1 + y_2$ for $0 \le y_1 \le 1, 0 \le y_2 \le 1$ and $f(y_1, y_2) = 0$ elsewhere. With this model, two probabilities can be explored. The probability $P(Y_1 < \frac{1}{2}, Y_2 > \frac{1}{4})$, represents the likelihood of a company earning less than half its daily average revenue and more than a quarter of its daily average profit on a given day. And the probability $P(Y_1 + Y_2 \le 1)$, signifying the chance of the combined daily revenue and profit proportions not exceeding the company's daily average. Solving the first probability we get $\int_{0}^{1/2} \int_{1/4}^{1} (y_1 + y_2) dy_2 dy_1 = \frac{21}{64}$ or 37.5%. Solving the second probability we get $\int_{0}^{1} \int_{0}^{1-y_1} (y_1 + y_2) dy_2 dy_1 = \frac{1}{3}$ or 33%. This result indicates the probability that the combined proportions of daily revenue and profit for a company will not exceed its daily average.

Continuing with this analysis, we are going to need to find our marginal density functions for Y1 and Y2 by integrating the joint density function over the range of the other variable. This process yielded the marginal densities as $fY_1(y_1) = y_1 + 1/2$ and $fY_2(y_2) = y_2 + 1/2$. To delve deeper, we can calculate the conditional probability P(Y1 ≥ 1/2 | Y2 ≥ 1/2). This probability represents the likelihood of a company earning at least half of its average daily revenue, given that it has already made at least half of its average daily profit. It's a measure of how the performance in one financial area (profit) might influence or relate to performance in another (revenue). $\frac{\int_{1/2}^{1}\int_{1/2}^{1}(y_1+y_2)dy_2dy_1}{\int_{1/2}^{1}(\frac{1}{2}+y_2)dy_2} = \frac{15/32}{5/8} = \frac{3}{4}$. This indicates that there is a 75% chance that a company is earning at least half of its average daily revenue, given that it has already made at least half of its average daily profit. This reflects a degree of dependency between the two metrics, illustrating how strong performance in profit is often associated with a strong performance in revenue on any given day.

Using statistics we can learn many interesting facts about data sets. In this case, we took a look at the dataset containing numerous numerical information about Fortune's top 1000 companies ranked by revenue from greatest to least. The data here can be used to see which companies are strong in certain industries and aspects and which companies are lacking in certain areas. The information and statistics discovered can be useful for investors and others alike. Overall, the study provided meaningful insight into the biggest companies in the country.

References

https://www.kaggle.com/datasets/surajjha101/fortune-top-1000-companies-by-revenue-2022/data

Wackerly, Dennis D., et al. Mathematical Statistics with Applications. Brooks/Cole, Cengage Learning, 2008.