The Gold Standard: The NBD Fit to Gold Medals Won at the 2020 Summer Olympics

Anonymous

February 25th 2022

1 Executive Summary

The contents of this paper majorly include an analysis of how the Negative Binomial Distribution (NBD) can be utilized to estimate the count distribution of the number of gold medals won by a given country's team in the 2020 Summer Olympics held in Tokyo, Japan. Moreover, the final model is used to analyze the number of bronze and silver medals won by a given country's team in the 2020 Summer Olympics.

The results display that the regular Unit Time NBD most optimally fits the count distribution of the number of gold medals won by a given country's team, in comparison to its spiked and Non-Unit Time variants. Additionally, the parameters derived from fitting the distribution of gold medals won fit more optimally to the distributions of the silver and bronze medals won than they fit the original gold medal data.

2 Introduction

The dataset was extracted directly from Kaggle, with the notes detailing that the user manually scraped the information from the Olympics homepage. It became quickly evident upon comparing summary statistics with publicly available information on participation that the dataset did not include all of the countries participating in the Olympics, only the countries that had participated and won at least 1 medal in competition. In order to account for the difference, the rest of the countries' medal data was manually inputted from a list found online.

The raw information included lists the country's name, the number of gold medals won, the number of silver medals won, the number of bronze medals won, the total number of medals won and the ranking based on the total number of medals won. The number of gold medals was chosen for the main unit of analysis for both objective and subjective reasons. For one, given the wide array of values for total number of medals won, the data is much more effectively visualized as a right-truncated model by minimizing the range of x values in the data. Additionally, a gold medal being the most prestigious award given at the Olympics, the investigation into the countries' heterogeneity and its relation to the silver and bronze medal distributions are of particular interest. Given each country having its own data entry, it is understandably of disaggregate format.

Table 1: Tokyo 2020 Medals Example Entry

Country	**Gold Medal**	**Silver Medal**	**Bronze Medal**	**Total**	**Rank By Total**
United States	39	41	33	113	1

As such, individuals in this dataset are not buyers or groups, but countries. Although the Winter Olympics are currently occurring, the Summer Olympics was chosen for analysis due to its popularity qualitatively and the number of countries quantitatively. This was performed to maximize the number of individuals in the model, with 77 more countries participating in the Summer Olympics than the Winter Olympics. 206 countries total participated in the Tokyo games, which is a relatively small magnitude for N_x in comparison to many of the other count datasets analyzed in this course. This has potential to affect the model's fit if the distribution of gold medals won does not conform to the trends depicted by the NBD due to lack of data.

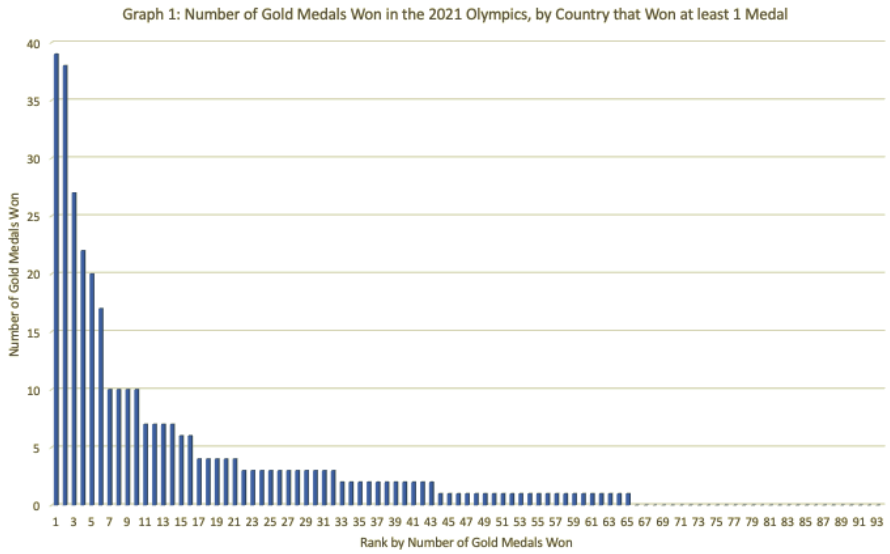
Given that there is a finite number of events in the Olympics, there are 340 total gold medals to be competed for by the countries in the games. This creates a noticeable inconsistency when referring to gold medals won as count data, with the Beta-Binomial being the best choice for optimal fit. However, given the high level of competition amongst each country, the mutual desire to attain the highest achievement at each event, and the fact that every medal won is for a unique event in which a given country may not have the same level of aptitude, the act of winning a medal cannot be described as a simple "choice." As the largest number of medals won by a country being 39, approximately 10% of the upper limit, the limit can be taken as negligible for the purpose of analysis.

Now assuming that the Unit Time NBD is suitable, each country has a unique propensity to win a certain number of gold medals in the Olympics, λ. In summary, the main analysis tests whether the number of gold medals in a given Olympic games can be described by each country spinning their respective Poisson wheel, with other parameters r and α characterizing size and scale of the underlying distribution.

3 Optimal Model Choice Analysis

3.1 Initial Data Evaluation

A bar graph depicting the total number of gold medals won by each country's rank in total number of gold medals won was initially conceived to visualize the raw data.



As shown in Graph 1 above, it is evident that only 10 countries won at least or more than 10 gold medals. This fact implies that each of these countries are statistical outliers when included in a comparison with the vast majority of the other countries participating in The Olympics. Additionally, all of these countries have teams among the top 13 largest in number of athletes participating and consistently rank amongst the highest-winning teams in every Summer Olympics hosted in the past 30 years. Thus, 10 was chosen as the value for right-censoring for every model created.

Before any models are fit to the data, it is necessary to hypothesize the potential fit of the model given the dataset's intrinsic qualities. It is likely that there will be high levels of heterogeneity, corresponding to a low r value, as the highest-winning countries majorly dominate events in comparison to their under-performing competitors, many of which did not even win one gold model. There has been much academic research on the subject of why the top countries perform considerably better than their lower ranked peers.

For example, in a paper published in Economic and Political Weekly, the generalized term Effectively Participating Population (EPP) is used to explain why different countries' populations around the world do not have the same access to competitive sports, understandably influenced by a multitude of factors indicating general quality of life. Overall, levels of financial support for sports, national wealth, social mobility and national culture for competitive sport were found to be correlated with Olympic performance. Given a lower supply of athletes in one country and an equal demand as the rest, the participation qualifications must necessarily decrease in order for a country to have a complete team at all. For example, enrollment rates of sports participation measured by memberships

in sports organizations is 0.01% to 1% in underdeveloped countries compared to 20% to 25% in European countries. Understandably, this would create high levels of heterogeneity found in the data, leading to a low r value. This additionally informs how each country has a unique λ informed by their intrinsic qualities such as EPP.

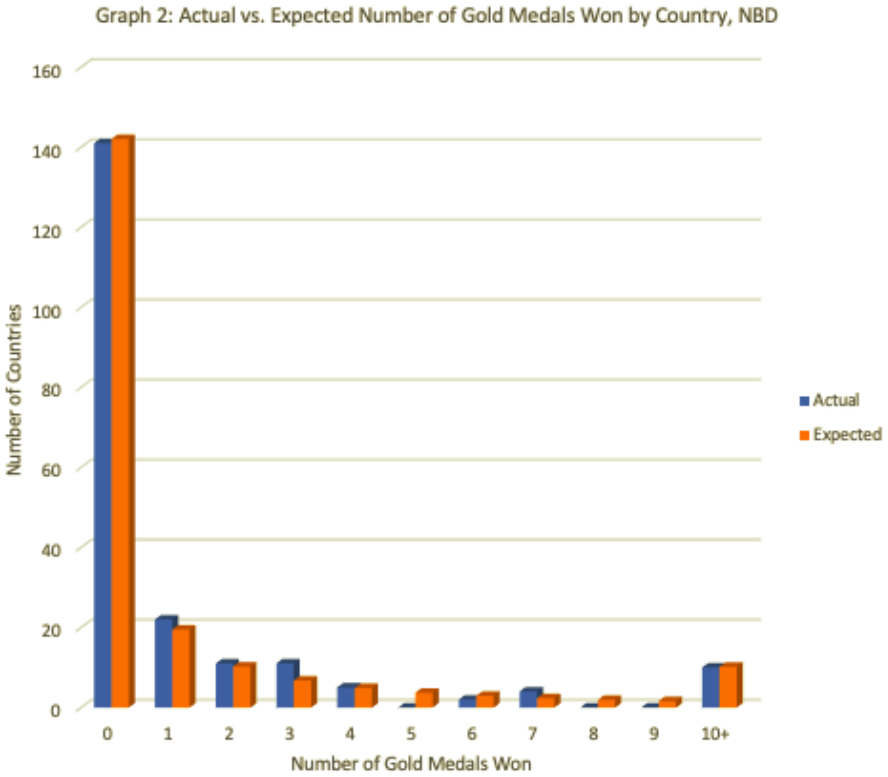
3.2 Unit Time NBD

The Maximum Likelihood Estimation (MLE) method was utilized in order to fit the Unit Time NBD to the disagreggated data. From this, the r and α estimates were obtained, being 0.149 and 0.095 respectively. After, the data was aggregated into the right-truncated format previously acknowledged.

Table 2: Aggregated Gold Medal Win Data

$\boldsymbol{x}$	0	1	2	3	4	5	6	7	8	9	10+
N_x	141	22	11	11	5	0	2	4	0	0	10

Using this aggregated data, a histogram depicting the actual and expected values for the number of countries that won a given number of gold medals was constructed.



A Chi-Square goodness of fit test was utilized to assess the fit of the model on the data, with a p-value of 0.162 obtained. This is below the standard of $p = 0.2$ for a great fit, possible for many reasons. Given the fact that there are less than 5 data entries in 5 of the data cells, there is simply not enough data to determine whether or not noise or outliers are influencing the fit outcomes. Additionally, the highly concentrated number of countries with 0 gold medal wins points to the disparity in EPP previously mentioned. It is evident that intrinsic qualities of a given country inform their number of gold medal wins. As such, further investigation into the fit of this model and the creation of new models are necessary to complete the analysis.

3.3 Non-Unit Time NBD

Given the subpar fit of the Unit Time NBD to the number of gold medal wins by country data, it is evident that new measures must be conceived to better represent the underlying distribution, in light of a convincing explanation. It appeared plausible that certain countries may never win a gold medal, only competing in the Olympics to be apart of a world tradition as opposed to attempting to win on a high level. However, the model predicts a higher number of countries to win 0 gold medals than the actual data displays. This indicates

that a spike at 0 would be ineffective, potentially due to the relatively small number of countries participating altogether. A spike at 1 quantitatively works, but the notion that there are a surplus of countries that only seriously compete in 1 event is far-fetched, as all countries send more than 1 adequate athlete to compete in multiple events. This points to a fundamental problem in the dataset at large: each country participates in a different number of events, an issue that can be addressed with a Non-Unit Time NBD.

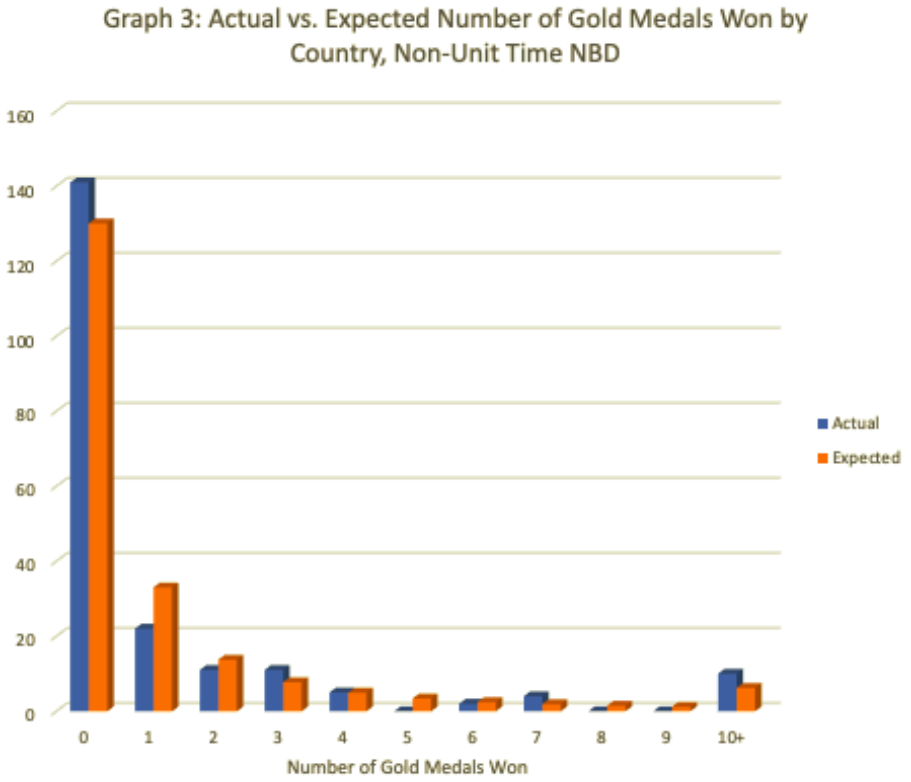
Due to the difference in number of events competed in by country, this dataset is at odds with the individual Poisson distribution essential to the NBD. The Poisson necessarily states that each individual has an equal number of opportunities to spin for a number of event occurrences. As the individuals in this model may not participate in the same number of events as the rest, they do not have equal opportunities to spin the Poisson wheel. This means that it is likely that a Non-Unit Time NBD may work with a given t value representing the number of events a given country participates in.

Table 3: Updated Tokyo 2020 Medals Example Entry

Country	**Gold Medal**	**Silver Medal**	**Bronze Medal**	**Total**	**Team Size**
United States	39	41	33	113	613

However, the regular NBD still may perform adequately due to a country not necessarily having an equal opportunity to win a specific event, with athletes always varying in skill level, and the possibility that smaller teams will only send their most capable athletes to compete, which may give them a higher chance of winning in those selected events. For example, Jamaica has a team consisting of just 50 athletes, but achieved more gold medals than Spain, which has a team of 321 athletes, due to Jamaica's specific domination in sprinting events. These probabilities could cancel out and allow for a subpar fit on a regular NBD, at the very least, and potentially hamper the fit of a Non-Unit Time NBD.

MLE was again used to fit the Non-Unit Time NBD to the disaggregate data, returning parameter values for r and α of 2.198 and 93.20 respectively. This indicates that the model predicts each country to have very similar likelihoods of winning a given amount of gold medals in comparison to the others, taking the number of Poisson spins as number of players on each team in this depiction. After, the data was similarly aggregated into the right-truncated format as before. A histogram was constructed depicting the actual and expected values calculated from the Non-Unit Time NBD.



A Chi-Square goodness of fit test yielded a p-value of 0.0266, a noticeable downgrade in quality from the Unit Time NBD that is visually evident when examining the histogram. It is likely that the specialization of certain small teams, the existence of a large number of athletes competing for 1 gold medal in certain events, and the relatively small size of the dataset are contributing to the poor results. Thus, the regular NBD was chosen as a basis of analysis to assess the distributions for the number of silver and bronze medals.

3.4 Means and Zeroes Estimation

As an extra measure to ensure the relative robustness of the Unit-Time NBD, the Means and Zeroes Estimation will be calculated. This allows for an extra opportunity to examine whether or not the model truly captures the underlying behavior of the given process.

The mean of the data is first estimated by summing the total number of gold medals earned by each country by the total number of countries, $\mu = 1.65$. Next, a random value is given to α, which is multiplied by μ to obtain r. From there, an experimental *P(0)* is calculated from the formula $(\alpha/\alpha + 1)^r$ and the actual *P(0)* is calculated by dividing the total number of countries with 0 gold

medal wins by the total number of countries participating. By minimizing the squared error between these two *P(0)* values, the parameters can be reestimated.

r and α are determined to be 0.154 and 0.0933 respectively. Although slightly different from the regular NBD fit, the error is likely a result from within sample estimates rather than a lack of conformity to the data. The difference in r is less than 0.005, while the difference in α is slightly larger, being approximately 0.30. However, given that α is simply a scale factor, this is unlikely to represent underlying model behavior, with r being a true measure of the similar heterogeneity. Despite the Means and Zeroes estimation leveraging less of the actual data, its parameters have a slightly better fit to the gold medal data with a p-value of 0.177, although this is likely due to the lack of data entries.

4 Model Application

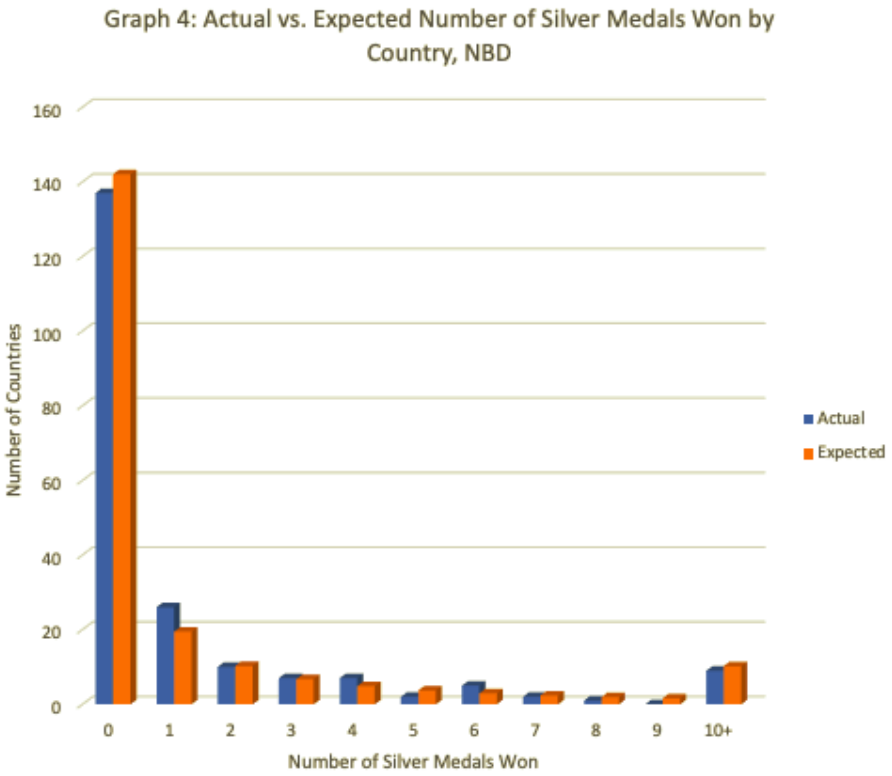
To further the investigation of whether or not the NBD accurately fits the medal wins data, the parameters derived from the gold medal wins distribution will be applied to the silver and bronze medal distributions and assessed for fit. Realistically, the aptitude for winning for each country should remain constant no matter the level of award to be won. Therefore, the application of such parameters acts as another measure of model robustness and may potentially lead to interesting conclusions being drawn from the data.

4.1 Silver Medal Data

The same method of MLE was performed on the disaggregated silver medal win data, then aggregated accordingly after to perform the Chi-Square goodness of fit test and create a histogram depicting the actual and expected values for the number of countries that won a given number of silver medals.

Table 4: Aggregated Silver Medal Win Data

$\boldsymbol{x}$	0	1	2	3	4	5	6	7	8	9	10+
N_x	137	26	10	7	7	2	5	2	1	0	9



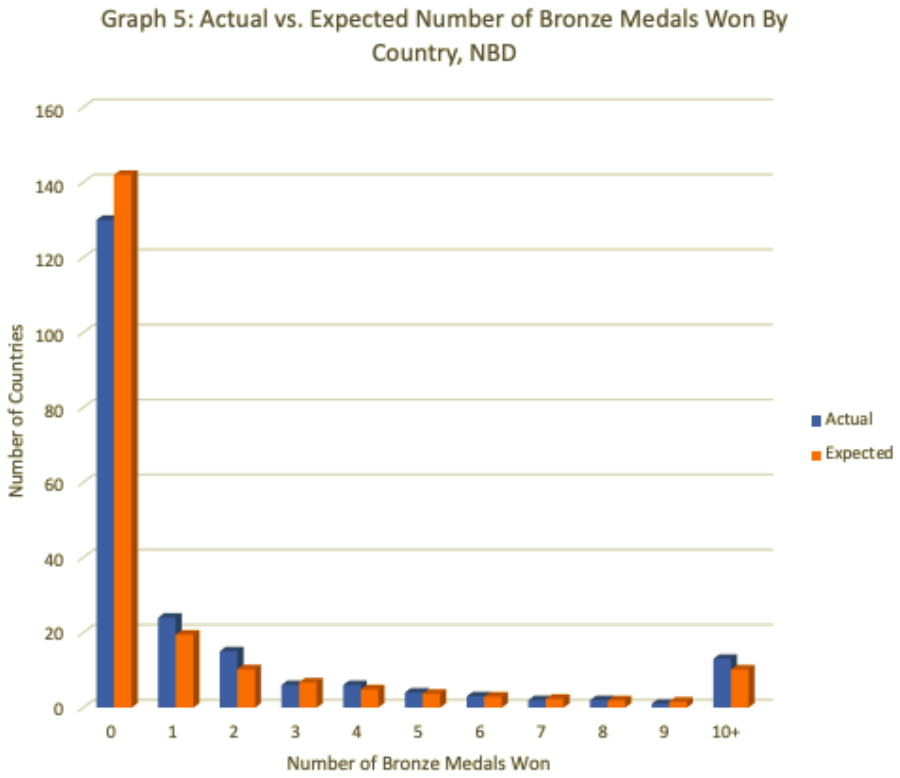
The p-value obtained for the silver medal wins distribution was 0.450.

4.2 Bronze Medal Data

The same method of MLE was performed on the disaggregated bronze medal win data, then aggregated accordingly after to perform the Chi-Square goodness of fit test and create a histogram depicting the actual and expected values for the number of countries that won a given number of bronze medals.

Table 5: Aggregated Bronze Medal Win Data

$\boldsymbol{x}$	0	1	2	3	4	5	6	7	8	9	10+
N_x	130	24	15	6	6	4	3	2	2	1	13



The p-value obtained for the bronze medal wins distribution was 0.681.

4.3 Implications

The increasingly improved fit from higher to lower levels of achievement by Olympic awards displays some interesting implications on the dataset and NBD as a whole. From a purely quantitative perspective, it is noticeable visually from the histograms and mathematically from the p-values that the silver and bronze medal wins data conform much more to the expected values than the gold medal wins data does. Specifically, the gold medal wins data contains 0's for countries having 5, 8 and 9 wins, while the silver and bronze variants display the downwards trend typical of an NBD. Given the small number of data entries, it is possible that this is pure coincidence; however, some attention must be given to the story that addresses increasing fit improvement.

It is believable that countries are able to achieve 2nd and 3rd place more frequently than 1st, as the top competitors will consistently take the 1st position, making it statistically more probable for any given competitor to spin a silver or bronze given 1 spot and 1 competitor are already taken out of the race entirely. As not all of the top teams compete against each other in every event, there

is more opportunity for lower ranked countries to take the lesser spots on the podium when it can be considered a given that 1st will be won by a top competitor. Additionally, there may be some validity to the notion that the best athletes from lower ranked teams may solely vie for 3rd or middle-of-the-pack positions, given the high-achieving nature of athletes from countries with high EPP. Naturally, this would create a trend typical of a standard NBD as reflected in the increasing p-values.

5 Further Research

Without a doubt, a larger number of countries should be included in this dataset in order to assess the full scope of the NBD on Olympic medal count data. As this is unlikely to occur given the limited number of nations in the world that participate in the Olympics any given year, the model should be used on data from past Olympics to make a direct assessment on its robustness.

Additionally, the Non-Unit Time NBD must be improved to accommodate for the difference in number of events for each country. Given the time constraints of this assessment, it was not possible to manually research the number of events each company participated in, while acknowledging more than 1 athlete from 1 country participating in an event, and more. It is likely that this investigation would lead to a more generalize model that displays the underlying propensity of a given country to win medals, all else held equal.

6 Acknowledgements

Thanks so much for getting through my paper! I hope you enjoyed (at least some aspect of) it.

7 Bibliography

https://www.npr.org/sections/tokyo-olympics-live-updates/2021/08/07/1025452727/how-home-field-advantage-gives-olympic-host-countries-an-edge-and-more-gold-meda

https://www.jstor.org/stable/pdf/40277720.pdf?refreqid=excelsior

https://www.businessinsider.in/sports/news/ranked-every-nation-at-the-tokyo-olympics-by-the-number-of-athletes-they-have-taken-to-japan/slidelist/84690956.cmsslideid=84691079

https://www.businessinsider.in/sports/news/ranked-every-nation-at-the-tokyo-olympics-by-the-number-of-athletes-they-have-taken-to-japan/slidelist/84690956.cmsslideid=84691079

Olympics.com