Is age associated with success at the Olympics?

Steven Leung, Brandon Lam, Sam Quist, Ruben De la Garza

2021/11/26 (updated: 2022-05-23)

Contents

Summary

Here we attempt to perform a hypothesis test on the question, "Is the proportion of athletes younger than 25 that win a medal greater than the proportion of athletes of age 25 or older that win a medal?". Our final result was very conclusive, since we got a p-value of 1 in our hypothesis test. We used simulation methods to generate our null hypothesis and placed our observed test statistic on it to find a visually consistent answer with the p-value. We did not have enough statistical evidence to say that the proportions mentioned are not equal.

Introduction

It is known that Olympic athletes require to train year after year for their shot at winning a medal for their country. In addition to physical strength, the right mental state is very important for success in these events. It is a common conception to think that the younger the athlete, the stronger and the more likely it is for him/her to win a spot on the podium. But is that enough to win a medal? Does experience play a more important role? For this project we will attempt to make a hypothesis test to answer the question - is the proportion of athletes younger than 25 that win a medal greater than the proportion of athletes of age 25 or older that win a medal?

Limitations and assumptions:

- The age threshold of 25 years old was chosen as this is the median age of the athletes in the data set. If time was provided, a way to make this analysis more robust would be to make some research and talk to some domain experts to find out if 25 years old is a good threshold to set for this hypothesis test.

- The data set contains information from the years 1896 - 2016, therefore, the analysis is taking into consideration all of these records, and the result should be interpreted as the comparison of the proportions mentioned within that time span. The analysis could be improved and give more specific insight if athletes were grouped by years (for example, before 1950 and after 1950) or by season (winter/summer games).
- The same athlete could appear in the same event for several years. If this is the case, each appearance will be taken as a different record, since we are taking into account each combination of athlete-event-games.

Methods

R programming language (R Core Team 2021) and the following R packages were used to perform the analysis:

- tidyverse (Wickham et al. 2019)
- knitr (Xie 2021)
- infer (Bray et al. 2021)
- broom (Robinson, Hayes, and Couch 2021)
- docopt (de Jonge 2020)
- kableExtra (Zhu 2021)

Also, python language (Python Core Team 2019a) and the following packages were used for the EDA:

- os (Python Core Team 2019b)
- altair (VanderPlas et al. 2018)
- pandas (McKinney 2010)

Exploratory Data Analysis

Here is the URL of our data source:

https://github.com/rfordatascience/tidytuesday/tree/master/data/2021/2021-07-27

Direct download links to individual CSV files:

The only file we need for our purpose:

https://github.com/rfordatascience/tidytuesday/raw/master/data/2021/2021-07-27/olympics.csv

Other files:

https://github.com/rfordatascience/tidytuesday/raw/master/data/2021/2021-07-27/athlete_events.csv
https://github.com/rfordatascience/tidytuesday/raw/master/data/2021/2021-07-27/noc_regions.csv
https://github.com/rfordatascience/tidytuesday/raw/master/data/2021/2021-07-27/regions.csv

Based on the source page, we understand that we really need the `olympics.csv` file which is the cleaned version of the file `athlete.csv`. The other 2 files only contains redundant information as far as our analytic objective is concerned. So we are going to do EDA on the `olympics.csv` file here.

The data dictionary is available here:

https://github.com/rfordatascience/tidytuesday/tree/master/data/2021/2021-07-27#olympicscsv

Table 2: Table 1: Summary information of data

	id	name	sex	age	height	weight	team
	Min. : 1	Length:261642	Length:261642	Min. :10.0	Min. :127	Min. : 25	Length:261642
	1st Qu.: 34755	Class :character	Class :character	1st Qu.:21.0	1st Qu.:168	1st Qu.: 60	Class :character
	Median : 68198	Mode :character	Mode :character	Median :24.0	Median :175	Median : 70	Mode :character
	Mean : 68291	NA	NA	Mean :25.6	Mean :175	Mean : 71	NA
	3rd Qu.:102109	NA	NA	3rd Qu.:28.0	3rd Qu.:183	3rd Qu.: 79	NA
	Max. :135571	NA	NA	Max. :97.0	Max. :226	Max. :214	NA
	NA	NA	NA	NA	NA's :51574	NA's :54263	NA

Table 3: Table 2: Sample rows from data

id	name	sex	age	height	weight	team	noc	games	year	sea
37280	Anna-Lena Katarina Fritzon	F	22	169	60	Sweden	SWE	1988 Winter	1988	Wi
2848	James Kanati Allen	M	21	173	64	United States	USA	1968 Summer	1968	Su
29215	Ivanka Peneva Dolzheva	F	16	162	60	Bulgaria	BUL	1952 Summer	1952	Su
116641	Bruny Surin	M	29	180	81	Canada	CAN	1996 Summer	1996	Su
44211	August Gttinger	M	35	NA	NA	Switzerland	SUI	1928 Summer	1928	Su

variable	class	description
id	double	Athlete ID
name	character	Athlete Name
sex	character	Athlete Sex
age	double	Athlete Age
height	double	Athlete Height in cm
weight	double	Athlete weight in kg
team	character	Country/Team competing for
noc	character	noc region
games	character	Olympic games name
year	double	Year of olympics
season	character	Season either winter or summer
city	character	City of Olympic host
sport	character	Sport
event	character	Specific event
medal	character	Medal (Gold, Silver, Bronze or NA)

Let's load the data and find out more.

The data, after removing rows with `age` missing because we focus on `age` in our analysis, has 261642 rows and 15 columns. Here are more information as we examine the dataframe.

Let's also sample a few rows from the data:

Points to note:

1. Each observation is an athlete-event-games key. In other words, an athlete could participate in more than 1 event in the same Olympic game, and the same athlete can participate in multiple Olympic games;
2. The columns do not have missing values after removing the rows with age missing. Even though `medal` appears to have a lot of missing values ('NA'), they are really not missing because the meaning of not

having Gold, Silver or Bronze would only mean that the athlete concerned did not win any medal. This is normal because there is a very small number of medals per event. For the purpose of this EDA, we may just treat 'NA' as a category together with Gold, Silver and Bronze.

3. id is the unique identifier for an athlete and noc is the unique identifier for an NOC which most often represents a country except IOA (standing for "Individual Olympic Athlete") which is the representation for athletes without an NOC and similary ROT (standing for "Refugee Olympic Team").

Age Distribution

Let's see what the age distribution looks like for all athletes:

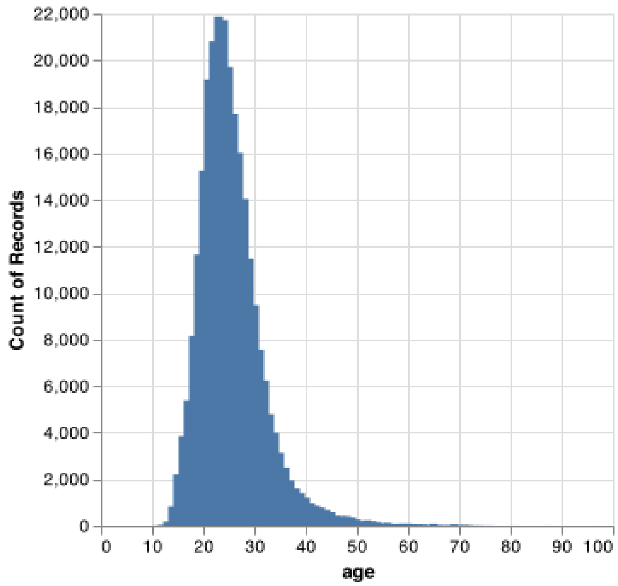

Figure 1: Figure 1: Age distribution

As can be seen above, the age peaks at 23 years old and the distribution is bell-shaped and right-skewed, which means that there are a few older athletes that compete in the olympics.

Age vs Numeric features

Let's explore how age correlates with other numeric features.

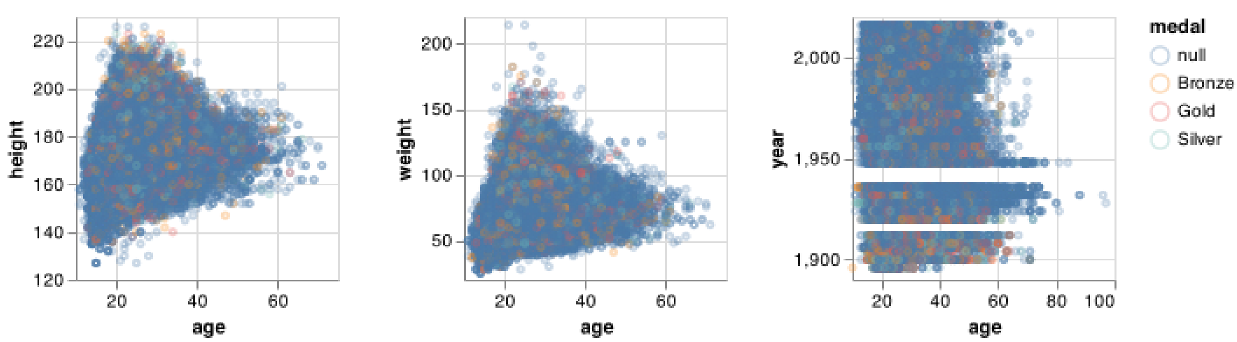

Figure 2: Figure 2: Age vs Numeric features

It seems difficult to visualize when the class imbalance is with `medal`, when not having a medal is the majority. Now we try again with the data only with medals and look at the data again.

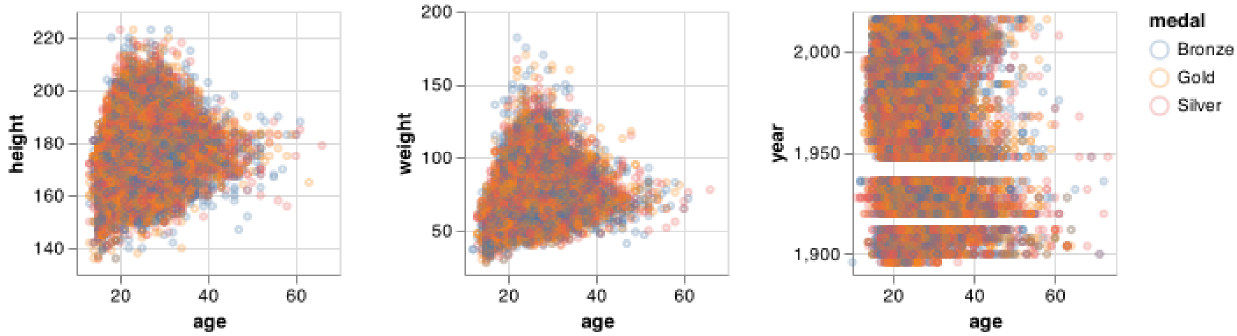

Figure 3: Figure 3: Age vs Numeric features (only with medals)

Some high-level insights:

1. There is some apparent correlation of between height and age and between weight and age for those who got medals; and
2. The maximum age of athletes getting medals seemed to shrink between 1960s and 1980s, and it seemed to increase again till now.

Perhaps we should simply just look at the relationship between age and medals...

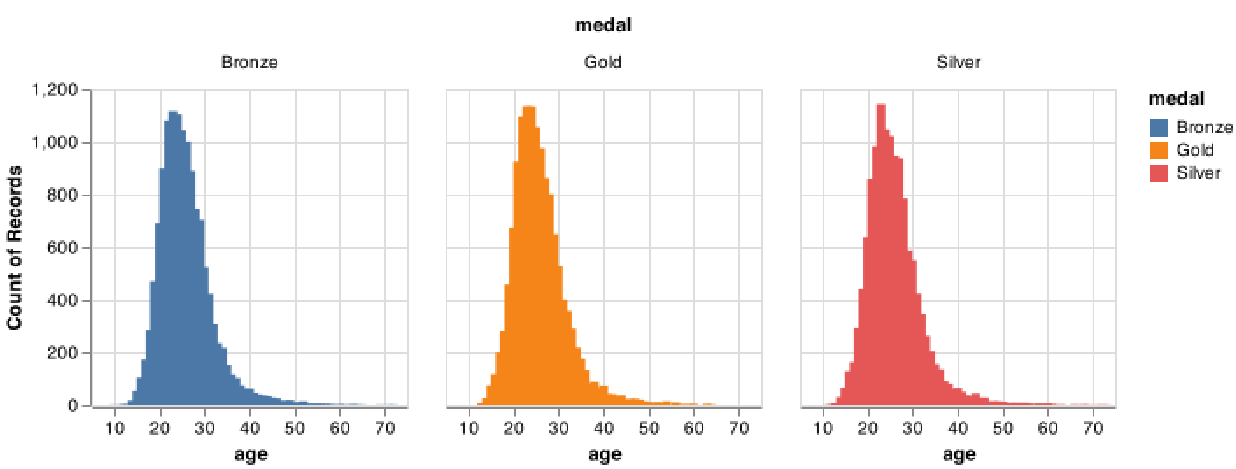

Figure 4: Figure 4: Age vs Medals

All the modes for `Gold`, `Silver` and `Bronze` appear to be the same as the overall age distribution as seen in Figure 1 above.

After this initial analysis, we can see that the age threshold of 25 years old lies in a very good spot, a little higher than the mean. This makes it harder to intuitively predict what the result of the test will be. Let's follow this with the analysis!

Hypothesis Test

Analysis

To answer our question, we will perform a hypothesis testing. First, we'll define H_0 and H_A as below:

Table 4: Table 3. Data summary

age	medal	n	prop
Under	17939	131134	0.137
Above	21112	130508	0.162

H_0 : the proportion of athletes under 25 that win a medal is equal to the proportion of athletes 25 and older that win a medal.

H_A : the proportion of athletes under 25 that win a medal is greater to the proportion of athletes 25 and older that win a medal.

We will then

1. Compute the observed test statistic from original sample,
2. Use the null model to generate 100 random permuted samples from the original sample and calculate their corresponding r test statistics,
3. Generate the null distribution using these r test statistics,
4. Check if the observed test statistic computed in (1) falls on the distribution,
5. Calculate the p-value to verify the result

The code used to perform the analysis and create this report can be found here: https://github.com/UBC-MDS/olympic_medal_htest

Results & Discussion

We can see from the table above that there are 131,134 athletes under age of 25 and 13.68% of them got a medal in the event, while there are 130,508 athletes equal to or above the age of 25 and 16.18% of them got a medal in the event.

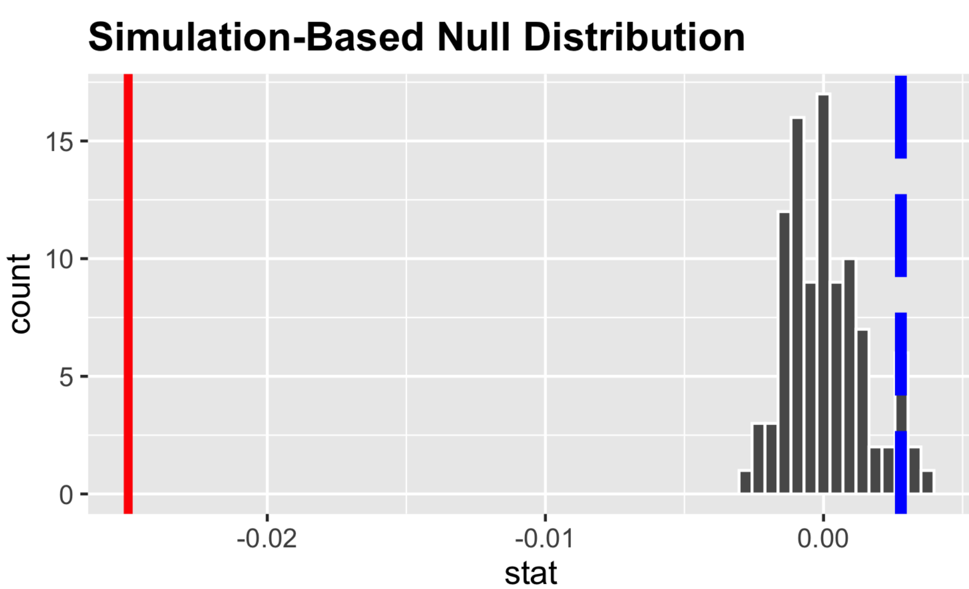

Figure 5: Figure 5. Hypothesis testing result

After generating the null distribution, and placing our observed test statistic on the plot in figure 5, we can

Table 5: Table 4. p-value for the test.

p-value
1

see that the observed test statistics (red line) falls within the significance threshold (blue line), therefore we fail to reject H_0.

The test statistic is -0.025, which is the portion of medal athletes under 25 minus the portion of medal athletes equal to or above 25. It is far outside the null distribution in the graph as our alternative hypothesis is "the proportion of medal athletes under 25 is greater than the proportion of medal athletes equal to or above 25", but the test statistic suggests that the portion of medal athletes under 25 is less than the portion of medal athletes equal to or above 25. This is a complete reverse of the alternative hypothesis.

The p-value calculated is 1 and it is higher than the α of 0.05. It leads us to the same conclusion:

We fail to reject the null hypothesis and conclude that it is not statistically significant that the proportion of athletes younger than 25 that win a medal is greater than the proportion of athletes of age 25 or older that win a medal.

The results show that athletes under 25 have not been more successful in the olympics in comparison to athletes who are 25 and older. We can attribute this result to two different factors:

1. The olympics have many different types of events, and these events have been changing through the years. Many of these sports are dominated by older athletes, since they require more experience and hours put into the sport, rather than physical dexterity. Examples for these events could be art competitions (sculpturing, music, among others in the 1940's), archery (1900's), shooting (1900's).

2. For the majority of events, experience still plays a very important role in winning a medal.

We also found a couple of papers (Singh 2021) and (Elmenshawy, Machin, and Tanaka 2015) that suggest that it is more likely for an athlete to win a medal when he has more experience.

References

Bray, Andrew, Chester Ismay, Evgeni Chasnovski, Simon Couch, Ben Baumer, and Mine Cetinkaya-Rundel. 2021. *Infer: Tidy Statistical Inference*. https://CRAN.R-project.org/package=infer.

de Jonge, Edwin. 2020. *Docopt: Command-Line Interface Specification Language*. https://CRAN.R-project.org/package=docopt.

Elmenshawy, Ahmed R, Daniel R Machin, and Hirofumi Tanaka. 2015. "A Rise in Peak Performance Age in Female Athletes." *Age* 37 (3): 1–8.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, 51–56.

Python Core Team. 2019a. *Python: A dynamic, open source programming language*. Python Software Foundation. https://www.python.org/.

———. 2019b. *Python: A dynamic, open source programming language*. Python Software Foundation. https://www.python.org/.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2021. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Singh, Dr Preet Deep. 2021. "Olympic Medals: Matter of Nerves." *Available at SSRN 3901321*.

VanderPlas, Jacob, Brian E Granger, Jeffrey Heer, Dominik Moritz, Kanit Wongsuphasawat, Arvind Satyanarayan, Eitan Lees, Ilia Timofeev, Ben Welsh, and Scott Sievert. 2018. "Altair: Interactive Statistical Visualizations for Python." *Journal of Open Source Software* 3 (32): 1057.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* https://CRAN.R-project.org/package=kableExtra.