ALY_6010 Module 1 Report

Junchen Yi

College Professional Studies in Analytics, Northeastern University

ALY_6010: Probability Theory and Introductory Statistics

Na Yu

Submission Date: 05/11/2023

Introduction:

The Olympic Games have always been a showcase of athletic and international friendship, captivating audiences around the world. The dataset I selected summarizes the history of these global competitions, documenting the performance and background of Olympic athletes over more than a century. The data in this dataset includes the first modern Olympic games from 1896 to 2016.

Analysis:

Variables of Interest:

The key variables that were the focus of my analysis included:

Athlete's Age (age): To explore the distribution of athletes' ages, particularly among gold medalists.

Gender (sex): To determine the medal distribution between male and female athletes.

Medal Type (medal): To ascertain the count of different types of medals (Gold, Silver, and Bronze).

Year of the Olympics (year): To observe trends in the number of participating athletes over the years.

Data Cleaning Process:

1. Exclusion of Non-Medalists: Records of athletes who did not win any medals were filtered out to focus the analysis on medal winners.
2. Removal of NA Values: Rows containing NA values across any column were removed to ensure the quality and completeness of the analysis.
3. Normalization of Column Names: All column names were converted to lowercase to maintain consistency and ease of reference.
4. Column Selection: The 'id' and 'noc' columns were deemed unnecessary for the analysis and so we removed it.
5. Column Renaming: The 'team' column was renamed to 'country' to more accurately reflect the variable it represents.

Analysis Step:

1. Counting the age distribution

In this step, I used the table() function to count the number of athletes of different age groups in the data set. table(olympics_data$age) generates a frequency table showing the counts of athletes in each age group, which helps to understand the age structure of the participating athletes. Due to the big range of age (from 13 to 58), I do not show the table in the report.

2. Cross-tabulation of gender and medal type

A cross-table is created with table(olympics_data$sex, olympics_data$medal) which shows the relationship between gender and the type of medals won. The crosstab was then formatted with ftable() to make it more concise and readable.

	Var1	Var2	Freq
1	F	Bronze	772
2	M	Bronze	1574
3	F	Gold	834
4	M	Gold	1595
5	F	Silver	785
6	M	Silver	1541

3. Selecting data for the 2016 Summer Olympics

Using the subset() function, I filtered the data for the 2016 Summer Olympics. The conditions are that the year is equal to 2016 and the season is summer. A simplified table was then generated by cross-tabbing gender and medal type with the xtabs() function.

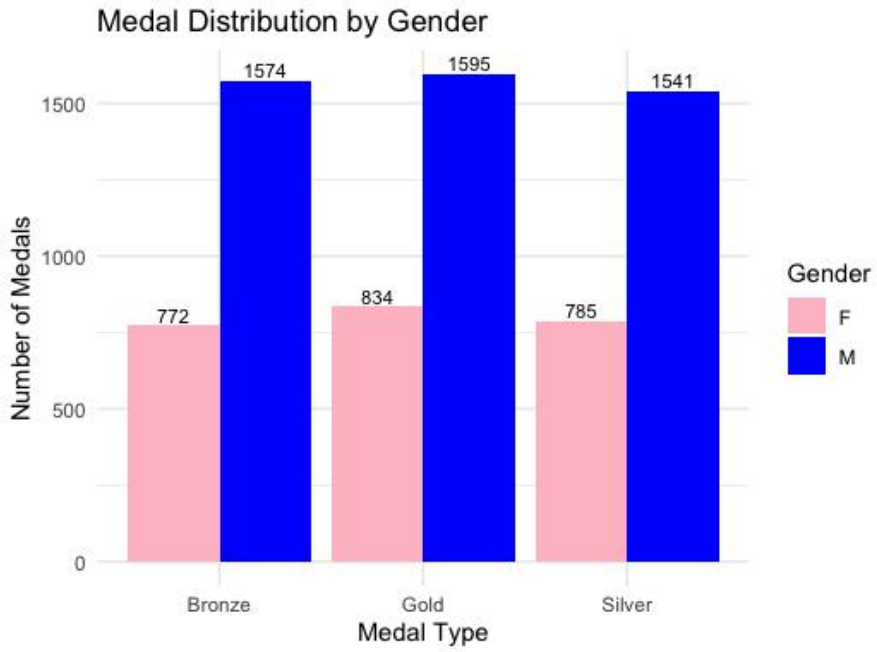
	sex	medal	Freq
1	F	Bronze	70
2	M	Bronze	104
3	F	Gold	82
4	M	Gold	92
5	F	Silver	94
6	M	Silver	75

4. Histograms of gender and medal distributions

First, the crosstab was converted to a dataframe as.data.frame(flat_medal_table), in order to be used in the ggplot2 package. A histogram was then created to show the number of medals won by different genders. geom_bar(stat = "identity") indicates that I want a bar graph, with the height of the bar determined by the value of the Freq column. position_dodge() is used to separate the bars for males and females so that they are displayed side-by-side. geom_text() is used to display specific values above each bar.

It's clear that men win significantly more Olympic medals than women do (almost twice as many overall), suggesting men's dominance in competitive events.

Flaw: The specific number of male and female athletes was not analyzed, and the above assumptions should be based on the same number of male and female athletes.

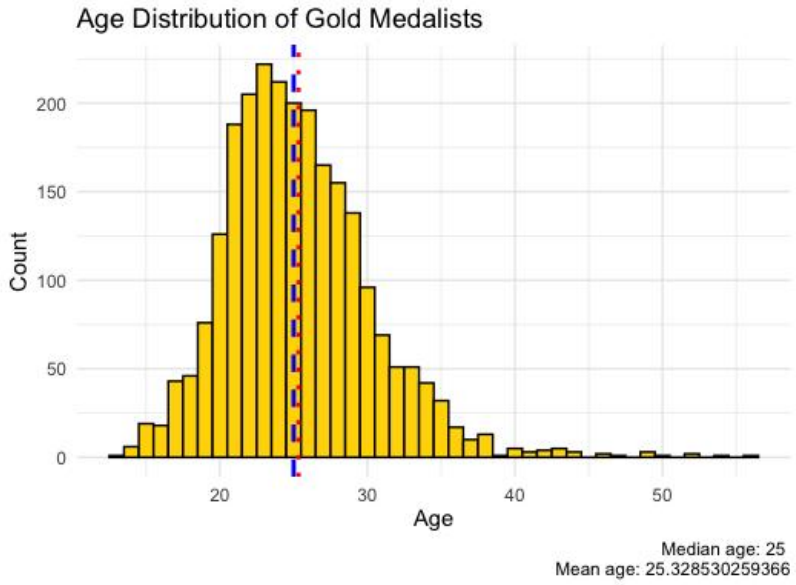


5. Age distribution of gold medalists

In this step, all the athletes who won gold medals were filtered by filter() function. I then calculated the median and mean age of these gold medalists, and finally used geom_histogram() to create a histogram to show the age distribution of these gold medalists, and geom_vline() to add reference lines indicating the median and mean age.

The average age of Olympic gold medalists is 25.3 years old, and the median age of 25 years old suggests that most gold medalists are around 25 years old. Therefore, we can conclude that athletes reach the height of their competitiveness in their young adulthood. During this period, athletes attain peak physical condition, technical maturity, and competition experience.

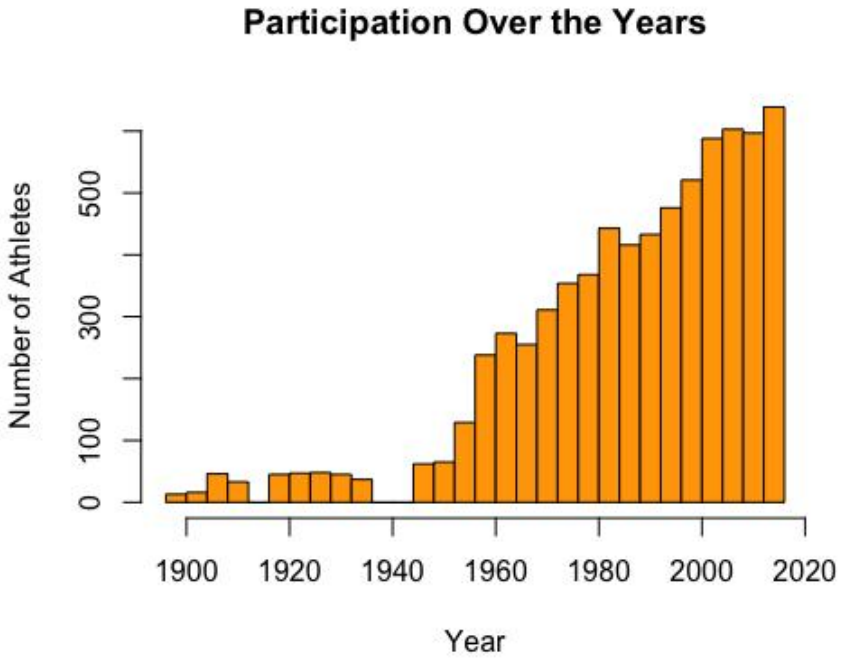
Flaw: Different sports have a big impact on the age distribution of athletes; programs like gymnastics may tend to favor younger athletes, while programs like shooting and soccer may have a wider age range.



6. Histogram of the number of participants by year

The final step in the analysis is to use the hist() function of base R to draw a histogram showing the number of athletes participating in the Olympics by year. the breaks parameter is used to set the width of the grouping of the histograms, which in this case is 4 years to match the Olympic cycle. This histogram can be used to analyze changes in size and trends in participation in the Olympics.

It is clear that the number of Olympic athletes has been increasing throughout the century, indicating that the Olympic Games are becoming more and more important to all countries and are becoming a major international event.



Reference:

Biswas, B. P. (2023, July 20). Olympic data . Kaggle. https://www.kaggle.com/datasets/bhanupratapbiswas/olympic-data