Analysis of competitors at the London 2012 Olympic games

Summary

This report aims to provide details on a statistical analysis of the physical traits of the athletes who competed in the London 2012 Olympics. This includes an analysis on the competitors' height, weight and age and an inspection on these parameters on swimmers and comparing them between medal winners and non-medal winners.

Introduction

In this Investigation I analyse the height, weight and age of 2012 Olympians and then analyse their spread and distribution with histograms and normal Q-Q plots. I then look at the three attributes in regards to swimmers, comparing them between medallists and non-medallists using t-tests to compare the averages of both groups.

Methods

Summary statistics are used to give a brief overview of what the spread of data looks like. Histograms then are used to portray which points of the data are of the highest density. The Normal Q-Q plot are used to determine whether a statistic fits a normal distribution model or not. Boxplots are used to portray the skewness of the data. Hypothesis tests are used to see if the mean of two data sets are equal.

All the calculations were performed in R, and the code to produce all the results is given in the Appendix.

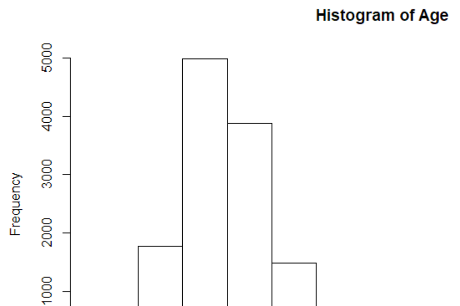
Results

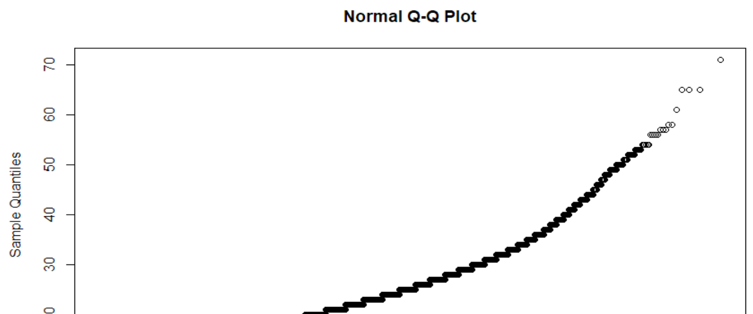
Analysis of competitors' height, weight and age

Summary Statistics:

Age(years)

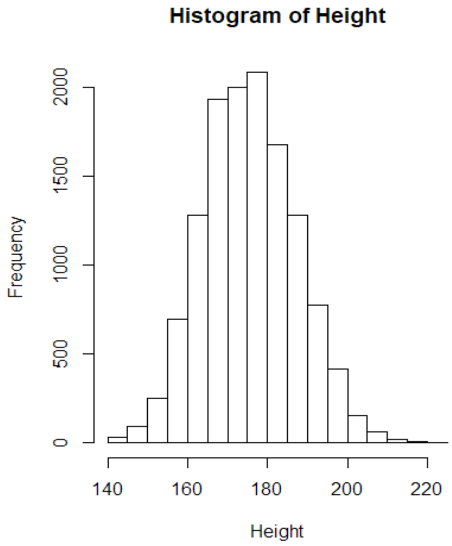
Min: 13; Max: 71; Lower Quartile: 22; Upper Quartile: 29; Median: 25; Mean: 25.96; Standard Deviation: 5.68

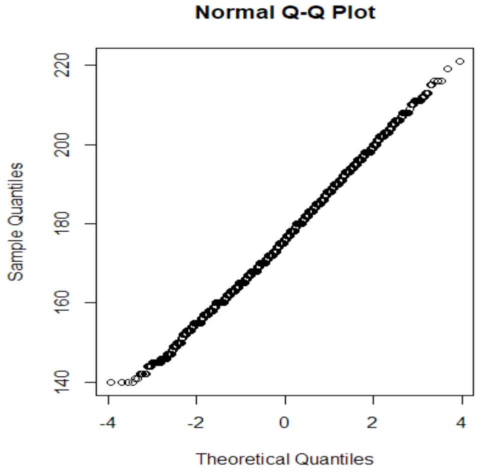




Height(cm)

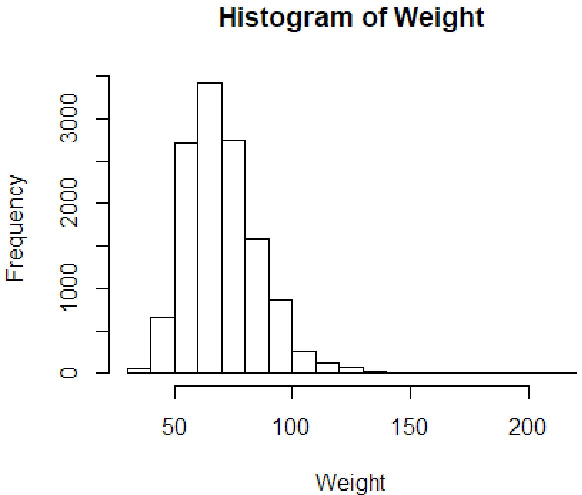
Min: 140; Max: 221; Lower Quartile: 168; Upper Quartile: 184; Median: 176; Mean: 176.30; Standard Deviation: 11.45

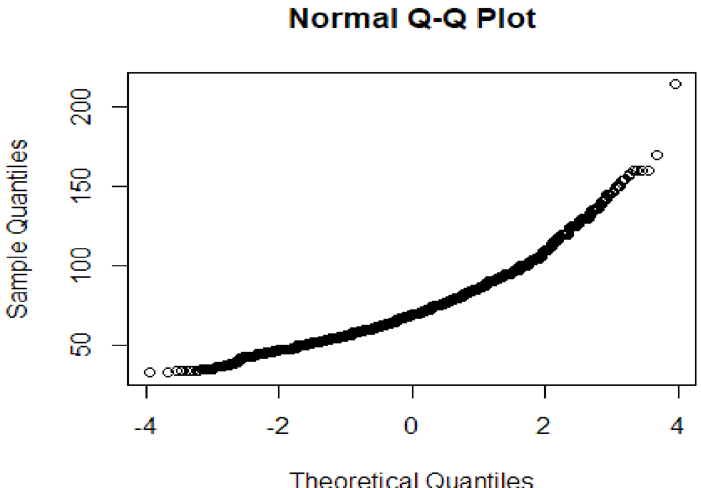

Histogram of Height


Normal Q-Q Plot

Weight(kg)

Min:33; Max: 214; Lower Quartile: 60; Upper Quartile: 80; Median: 69; Mean: 71.32; Standard Deviation: 15.86

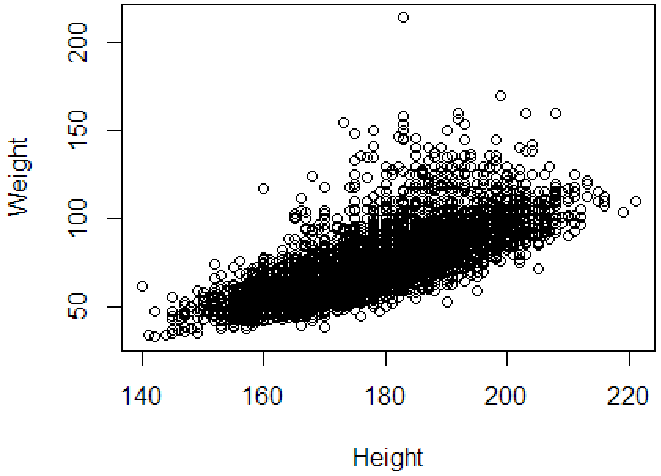

Histogram of Weight


Normal Q-Q Plot

Comments on base statistics: Upon first glance, it is very obvious that the data has a lot of spread in all three categories. The spread in age is explainable by the role of the Olympics to each sport. For example, in men's boxing the Olympics serves as a platform for the best amateurs to display their skills before starting a professional career, therefore skewing age lower whilst archery, where the Olympics serves as the highest level of competition, has older, more experienced competitors thus skewing the age higher. The spread in weight and height can be explained by sports with weight classes. For example, men's boxing's lowest weight division, light flyweight, has a catchweight of 49kg, however the super heavyweight division has no maximum weight and starts at 91kg.

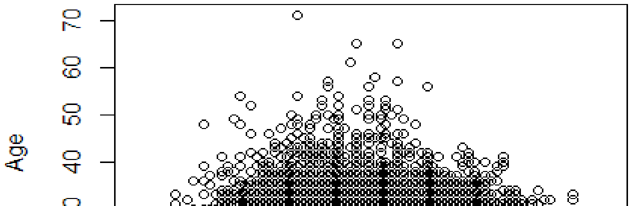
Comments on normal Q-Q plots: The plots for age and weight show that they are incompatible with a normal model. However, height's plot forms a straight line meaning the height could be normally distributed.

Scatterplot: Height-Weight



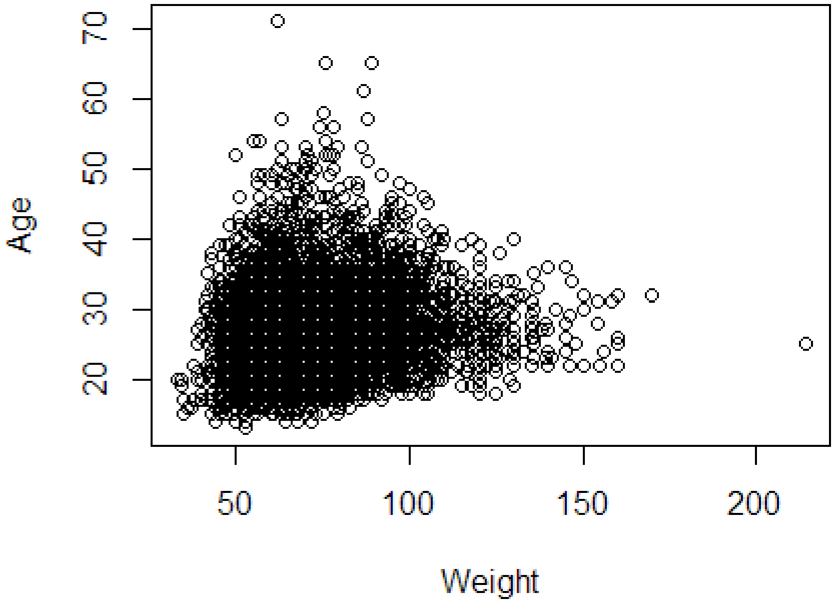
Notes: The dense area seems to portray a slightly positive correlation between these two statistics. There are outliers though. Most notably, the highest weighing competitor, Ricardo Blas Jr., weighing in at 214. Ricardo Blas Jr. is a statistical anomaly, even for a Judo heavyweight he shouldn't be at that sort of mass at his height.

Scatterplot: Height-Age



Notes: There seem to be no correlation between Height and Age.

Scatterplot: Weight-Age



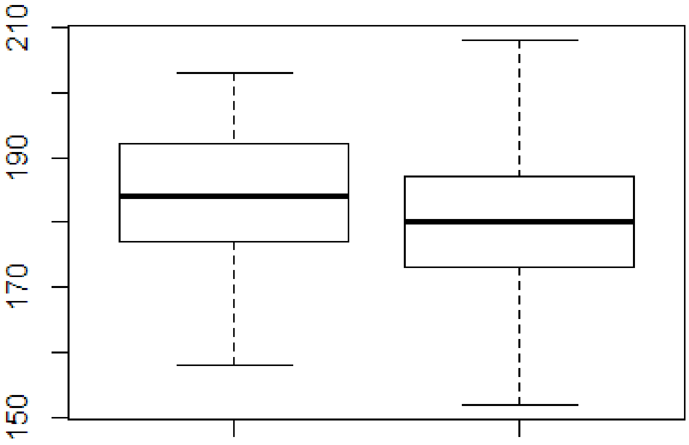
Notes: There seem to be no correlation between Weight and Age.

The association between medal-winning and height, weight and age amongst swimmers

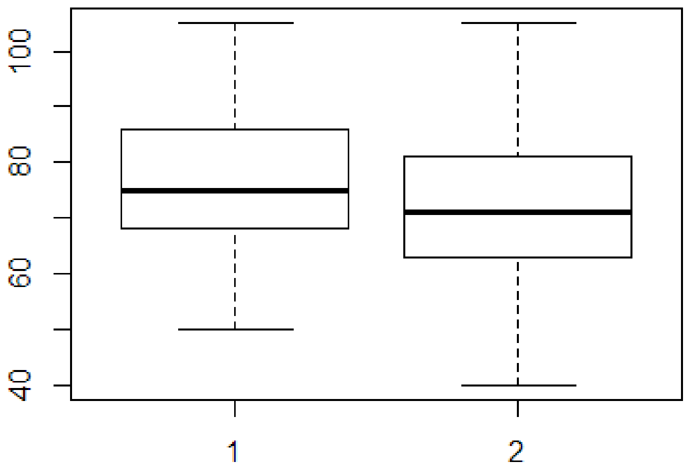
Boxplots:

Medal-winning Swimmers (Left) Vs. Swimmers who didn't win a medal (Right)

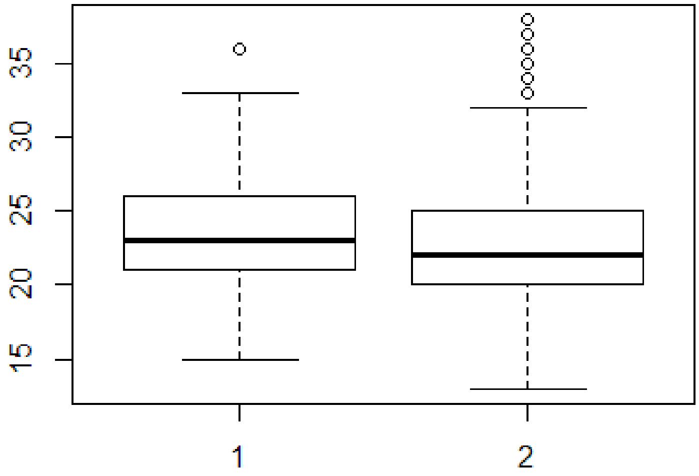
Height:



Weight:



Age:



Comments: The range of heights and weights are a lot narrower for the winners of medals. There are 5 more outliers in age for non-medallist than medallists. All of these boxplots are positively skewed except from height in non-medallists, which is normally distributed.

We investigate this further using hypothesis tests at a 95% significance level. Suppose that μ_{medal} is the mean weight of medal-winning swimmers and that $\mu_{\text{non-medal}}$ is the mean weight of non-medal-winning swimmers. A two-sample t-test of $H_0 : \mu_{\text{non-medal}=}\mu_{\text{medal}}$ versus the two-sided alternative $H_1 : \mu_{\text{non-medal}=/=}\mu_{\text{medal}}$ has a statistic t=5.3802 with corresponding *p*-value of 1.722 x 10^{-7} indicating that the alternate hypothesis is true.

I then repeated this for age and height the following were what I got as results.

Height:

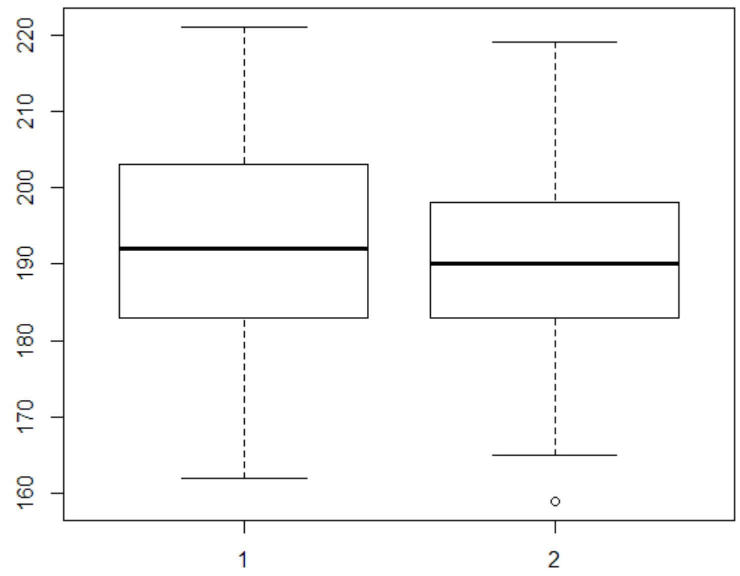
t=5.7744, p-value= 2.233 x 10^{-8}, null hypothesis rejected.

Age:

t=3.4554, p-value= 0.0006435, null hypothesis rejected.

Height of Basketball Players Vs. Volleyball Players

Box Plot: Basketball Player Height (Left) Vs. Volleyball Player Height (Right)



Upon first glance the heights seem somewhat similar with volleyball having a narrower spread. To further this investigation I took a t-test comparing $\mu_{Basketball}$ and $\mu_{Volleyball}$ with H_0: $\mu_{Basketball} = \mu_{Volleyball}$ and H_1: $\mu_{Volleyball}$ =/= $\mu_{Basketball}$. The test statistic was equal to 2.5341 with p-value 0.01154. at a 95% confidence level I had to reject H_0 and accept that the two means are not equal.

Conclusions

Most of my t-tests resulted in my null hypothesis being rejected. The average weight, height and age of an Olympic medallist in swimming was not the same as that of an Olympic swimmer with no medals. I also found out that the average height of an Olympic basketball player was not the same as an Olympic volleyball player. A limitation with the swimmers' statistics is that since there were a lot of very young swimmers that were no where near their physical peak to be true contenders for medals it had an impact on data as they would be a lot more underdeveloped, to bypass this I could have set a minimum age limit on the data.

Word count = 888

Appendix: The R code for producing the results and plots in this report is as follows.

```
athlete.data <- read.csv("athlete_data_2012.csv", header = TRUE) #Load the data
attach(athlete.data)#Attach the column names in the spreadsheet to the statistics
summary(Age)
summary(Height)
summary(Weight)
sd(Age)
sd(Height,na.rm=TRUE)
sd(Weight,na.rm=TRUE)#Calculate Summary Data and the Standard Deviations
hist(Age)
qqnorm(Age)
hist(Height)
qqnorm(Height)
hist(Weight)
qqnorm(Weight)
plot(Height,Weight)
plot(Height,Age)
plot(Weight,Age) #Plots the histograms, normal Q-Q plots and scattergraphs.

#Now to start coding for the boxplots and t-tests


  Medallist=athlete.data[athlete.data["Sport"] == "Swimming" &
!is.na(athlete.data["Medal"]), "Height"]
  NonMedallist=athlete.data[athlete.data["Sport"] == "Swimming" &
is.na(athlete.data["Medal"]), "Height"]# Seperates Medalists and Non-Medalists

  boxplot(Medallist,NonMedallist)
  t.test(Medallist,NonMedallist)

  Medallist=athlete.data[athlete.data["Sport"] == "Swimming" &
!is.na(athlete.data["Medal"]), "Weight"]
  NonMedallist=athlete.data[athlete.data["Sport"] == "Swimming" &
is.na(athlete.data["Medal"]), "Weight"]# Seperates Medalists and Non-Medalists

  boxplot(Medallist,NonMedallist)
  t.test(Medallist,NonMedallist)


  Medallist=athlete.data[athlete.data["Sport"] == "Swimming" &
!is.na(athlete.data["Medal"]), "Age"]
  NonMedallist=athlete.data[athlete.data["Sport"] == "Swimming" &
is.na(athlete.data["Medal"]), "Age"]# Seperates Medalists and Non-Medalists

  boxplot(Medallist,NonMedallist)
  t.test(Medallist,NonMedallist)

  #Coding For Basketball and Volleyball players

  Basketball=athlete.data[athlete.data["Sport"] == "Basketball", "Height"]
  Volleyball=athlete.data[athlete.data["Sport"] == "Volleyball", "Height"]

  boxplot(Basketball,Volleyball)
  t.test(Basketball,Volleyball)
```