

Análises Olímpíadas

Consultor Responsável:

Davi Esmeraldo

Requerente:

Rogério Santana



Conteúdo

1	Introdução	2
2	Metodologia	3
2.1	Frequência Relativa	3
2.2	Média	3
2.3	Mediana	4
2.4	Quartis	4
2.5	Variância	5
2.5.1	Variância Populacional	5
2.5.2	Variância Amostral	5
2.6	Desvio Padrão	5
2.6.1	Desvio Padrão Populacional	6
2.6.2	Desvio Padrão Amostral	6
2.7	Boxplot	7
2.8	Gráfico de Dispersão	7
2.9	Coeficiente de Correlação de Pearson	8
3	Análises	9
3.1	Série Histórica nas Olimpíadas de verão	9
3.2	Correlação entre número de medalhas e idade	10
3.3	Quantidade de pódios por continente	13
3.4	Distribuição dos IMCs por jogos de verão e inverno	13
3.5	Top 5 participantes	16
3.6	Países com menor número de participações	17
4	Conclusão	18

1 Introdução

O presente relatório tem como principal objetivo promover o entendimento de como as características dos participantes podem influenciar no ganho de medalhas. Ademais, para que fosse possível tal finalidade, foi realizado um estudo do o banco dados disponibilizado, esse, constituído de dados referentes a jogos Olímpicos.

Correlato a isso, o mesmo contém 271116 linhas e 12 colunas, onde cada linha corresponde a um atleta individual e, nesse sentido, é preenchida com informações pessoais como peso, altura, gênero, entre outros.

Para com que fosse possível gerar tais conclusões, foi utilizado o software 'R' versão 4.0.3 .

2 Metodologia

2.1 Frequência Relativa

A frequência relativa é utilizada para a comparação entre classes de uma variável categórica com c categorias, ou para comparar uma mesma categoria em diferentes estudos.

A frequência relativa da categoria j é dada por:

$$f_j = \frac{n_j}{n}$$

Com:

- $j = 1, \dots, c$
- n_j = número de observações da categoria j
- n = número total de observações

Geralmente, a frequência relativa é utilizada em porcentagem, dada por:

$$100 \times f_j$$

2.2 Média

A média é a soma das observações dividida pelo número total delas, dada pela fórmula:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

Com:

- $i = 1, 2, \dots, n$
- n = número total de observações

2.3 Mediana

Sejam as n observações de um conjunto de dados $X = X_{(1)}, X_{(2)}, \dots, X_{(n)}$ de determinada variável ordenadas de forma crescente. A mediana do conjunto de dados X é o valor que deixa metade das observações abaixo dela e metade dos dados acima. Com isso, pode-se calcular a mediana da seguinte forma:

$$med(X) = \begin{cases} X_{\frac{n+1}{2}}, & \text{para } n \text{ ímpar;} \\ \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}, & \text{para } n \text{ par;} \end{cases}$$

2.4 Quartis

Os quartis são separatrizes que dividem o conjunto de dados em quatro partes iguais. O primeiro quartil (ou inferior) é o conjunto que delimita os 25% menores valores, o segundo representa a mediana e é o valor que ocupa a posição central (ou seja, metade dos dados estão abaixo dela e a outra metade está acima) e o terceiro delimita os 25% maiores valores. Inicialmente deve-se calcular a posição do quartil:

- Posição do primeiro quartil P_1 :

$$P_1 = \frac{n+1}{4}$$

- Posição da mediana (segundo quartil) P_2 :

$$P_2 = \frac{n+1}{2}$$

- Posição do terceiro quartil P_3 :

$$P_3 = \frac{3 \times (n+1)}{4}$$

Com n sendo o tamanho da amostra. Dessa forma, $X_{(P_i)}$ é o valor do i -ésimo quartil, onde $X_{(j)}$ representa a j -ésima observação dos dados ordenados.

Se o cálculo da posição resultar em uma fração deve-se fazer a média entre o valor que está na posição do inteiro anterior e do seguinte ao da posição.

2.5 Variância

A variância é uma medida que avalia o quanto que os dados estão dispersos em relação à média, em uma escala ao quadrado da escala dos dados.

2.5.1 Variância Populacional

Para uma população, a variância é dada por:

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

Com:

- X_i = i -ésima observação da população
- μ = média populacional
- N = tamanho da população

2.5.2 Variância Amostral

Para uma amostra, a variância é dada por:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

Com:

- X_i = i -ésima observação da amostra
- \bar{X} = média amostral
- n = tamanho da amostra

2.6 Desvio Padrão

O desvio padrão é a raiz quadrada da variância. Avalia o quanto os dados estão dispersos em relação à média.

2.6.1 Desvio Padrão Populacional

Para uma população, o desvio padrão é dado por:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}$$

Com:

- X_i = i-ésima observação da população
- μ = média populacional
- N = tamanho da população

2.6.2 Desvio Padrão Amostral

Para uma amostra, o desvio padrão é dado por:

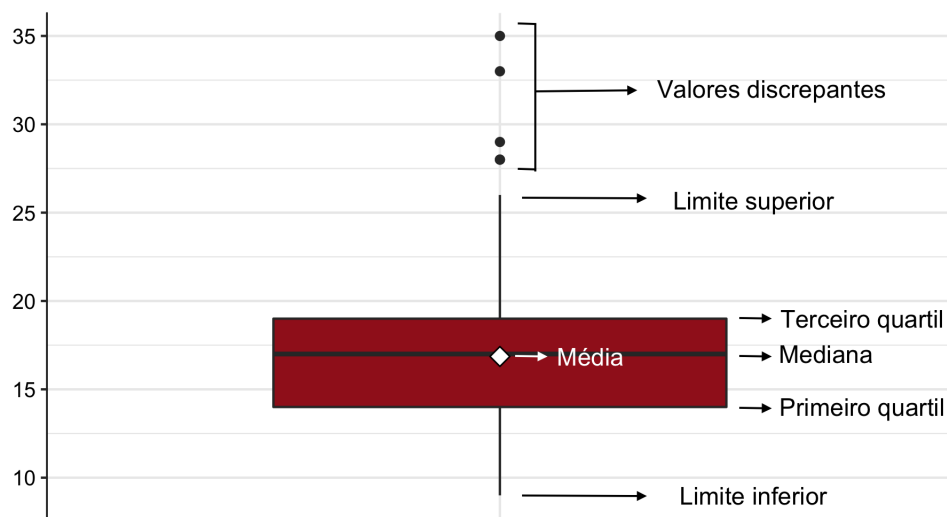
$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

Com:

- X_i = i-ésima observação da amostra
- \bar{X} = média amostral
- n = tamanho da amostra

2.7 Boxplot

Figura 1: Exemplo de boxplot



O boxplot é uma representação gráfica na qual se pode perceber de forma mais clara como os dados estão distribuídos. A figura abaixo ilustra um exemplo de boxplot.

A porção inferior do retângulo diz respeito ao primeiro quartil, enquanto a superior indica o terceiro quartil. Já o traço no interior do retângulo representa a mediana do conjunto de dados, ou seja, o valor em que o conjunto de dados é dividido em dois subconjuntos de mesmo tamanho. A média é representada pelo losango branco e os pontos são *outliers*. Os *outliers* são valores discrepantes da série de dados, ou seja, valores que não demonstram a realidade de um conjunto de dados.

2.8 Gráfico de Dispersão

O gráfico de dispersão é uma representação gráfica utilizada para ilustrar o comportamento conjunto de duas variáveis quantitativas. A figura abaixo ilustra um exemplo de gráfico de dispersão, onde cada ponto representa uma observação do banco de dados.

2.9 Coeficiente de Correlação de Pearson

O coeficiente de correlação de Pearson é uma medida que verifica o grau de relação linear entre duas variáveis quantitativas. Este coeficiente varia entre os valores -1 e 1. O valor zero significa que não há relação linear entre as variáveis. Quando o valor do coeficiente r é negativo, diz-se existir uma relação de grandeza inversamente proporcional entre as variáveis. Analogamente, quando r é positivo, diz-se que as duas variáveis são diretamente proporcionais.

O coeficiente de correlação de Pearson é normalmente representado pela letra r e a sua fórmula de cálculo é:

$$r_{Pearson} = \frac{\sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \times \sqrt{\sum_{i=1}^n y_i^2 - n\bar{y}^2}}$$

Onde:

- x_i = i-ésimo valor da variável X
- y_i = i-ésimo valor da variável Y
- \bar{x} = média dos valores da variável X
- \bar{y} = média dos valores da variável Y

Vale ressaltar que o coeficiente de Pearson é paramétrico e, portanto, sensível quanto à normalidade (simetria) dos dados.

3 Análises

3.1 Série Histórica nas Olimpíadas de verão

Figura 2: Gráfico de linhas dos gêneros masculino e feminino com o decorrer dos anos.

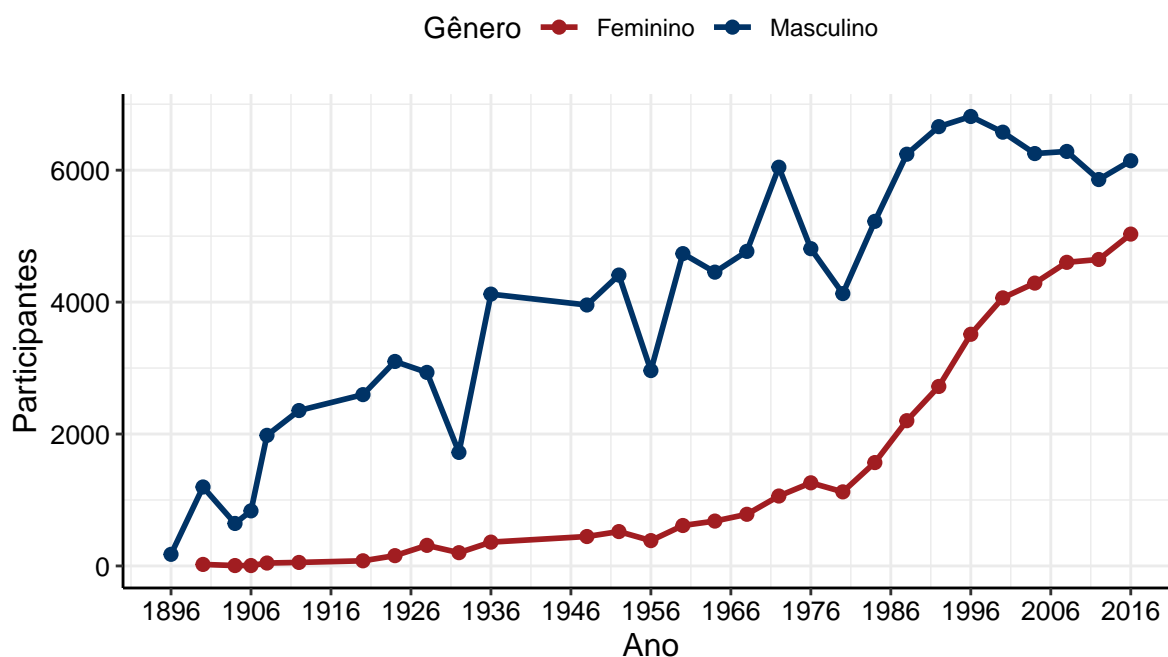


Tabela 1: Frequência de classes da variável gênero feminino em anos citados

¹ Ano	Frequência	Porcentagem
1900	23	4,08%
1980	1123	21,4%
2016	5031	45%

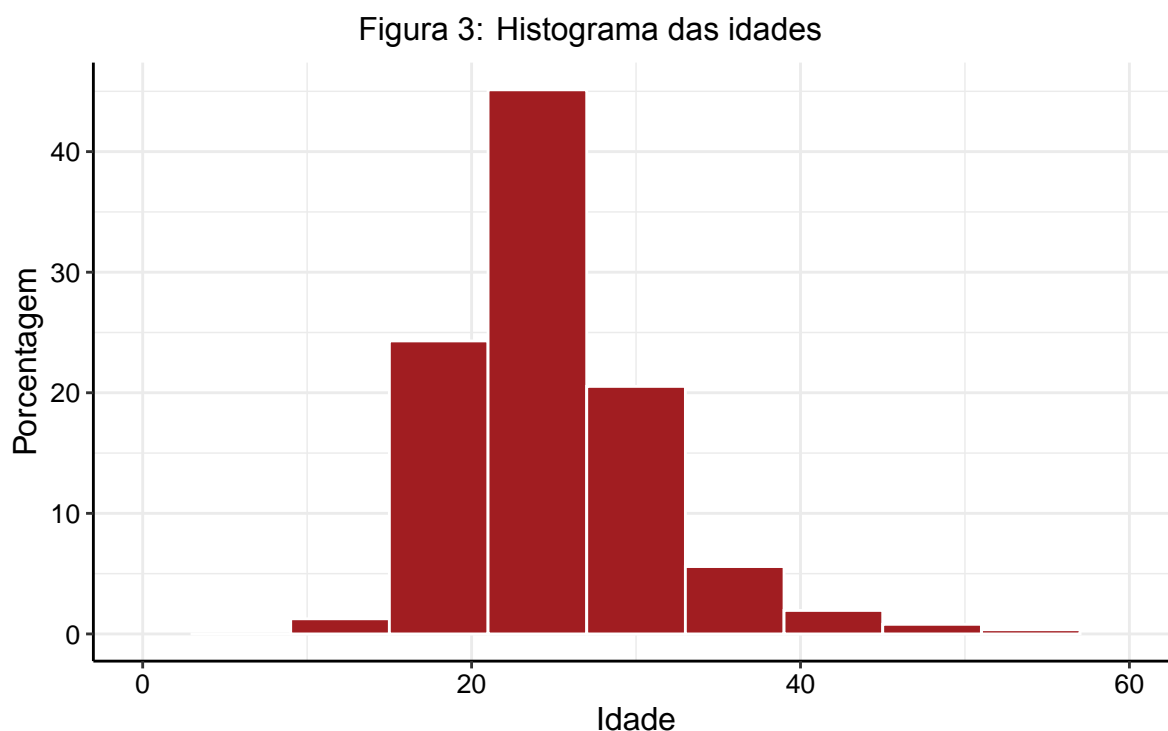
¹Fonte. Banco de dados disponibilizado

Percebe-se, com auxílio do gráfico acima, a participação destoante entre gêneros em eventos olímpicos, sendo notória a prevalência numérica de participações masculinas.

Análogo a isso, a participação do gênero feminina nos Jogos Olímpicos de fato veio a ocorrer pela primeira vez apenas no ano de 1900. Desde então, observa-se o aumento dessa, com destaque aos anos posteriores a 1980 e, especificamente, ao ano de 2016, no qual ocorreram os Jogos Olímpicos no Rio de Janeiro em que cerca de 45% dos participantes eram mulheres.

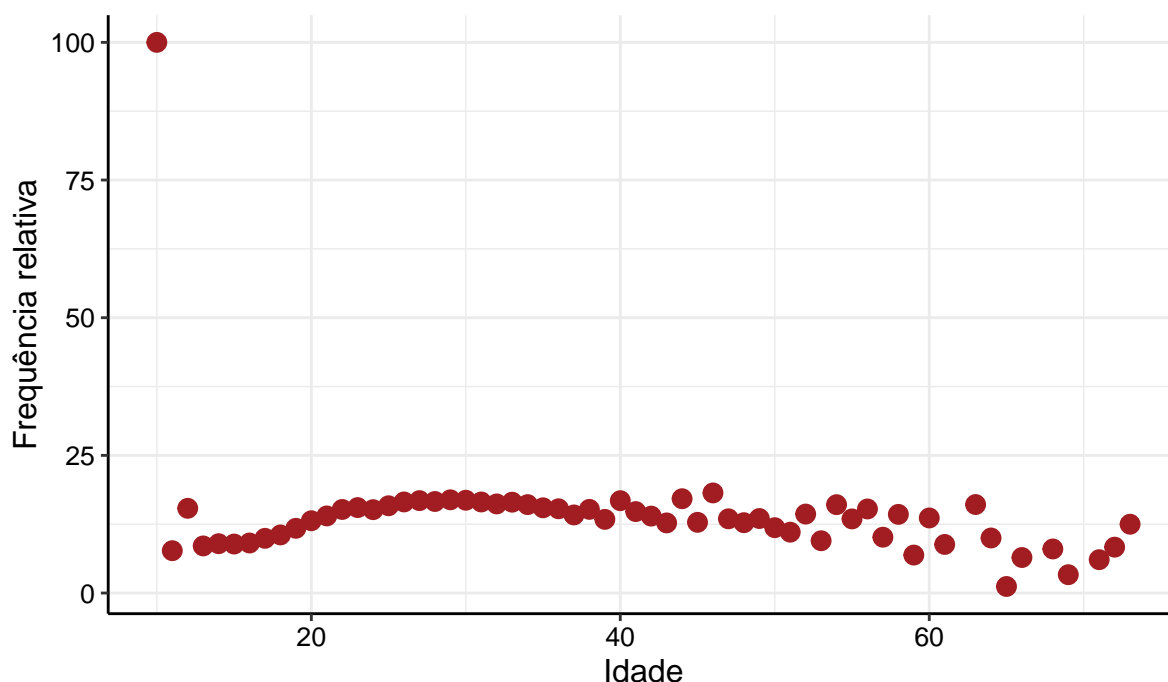
3.2 Correlação entre número de medalhas e idade

Com intuito de se entender a distribuição de idades dos atletas, obteve-se o seguinte gráfico.



Mediante a esse, percebe-se que atletas jovens com idades próximas de 25 anos são maioria em contexto olímpico. Além disso pode ser observado a participação de atletas com diversas idades, sendo perceptível a presença de observações que se distanciam das demais.

Figura 4: Gráfico de dispersão da frequência relativa dos atletas ganhadores de medalhas pela idade dos mesmos



Quadro 1: Coeficiente de correlação de Pearson entre as variáveis Idade e Medalhas.

Variáveis	Coeficiente de Pearson
Medalhas Idade	-0,3

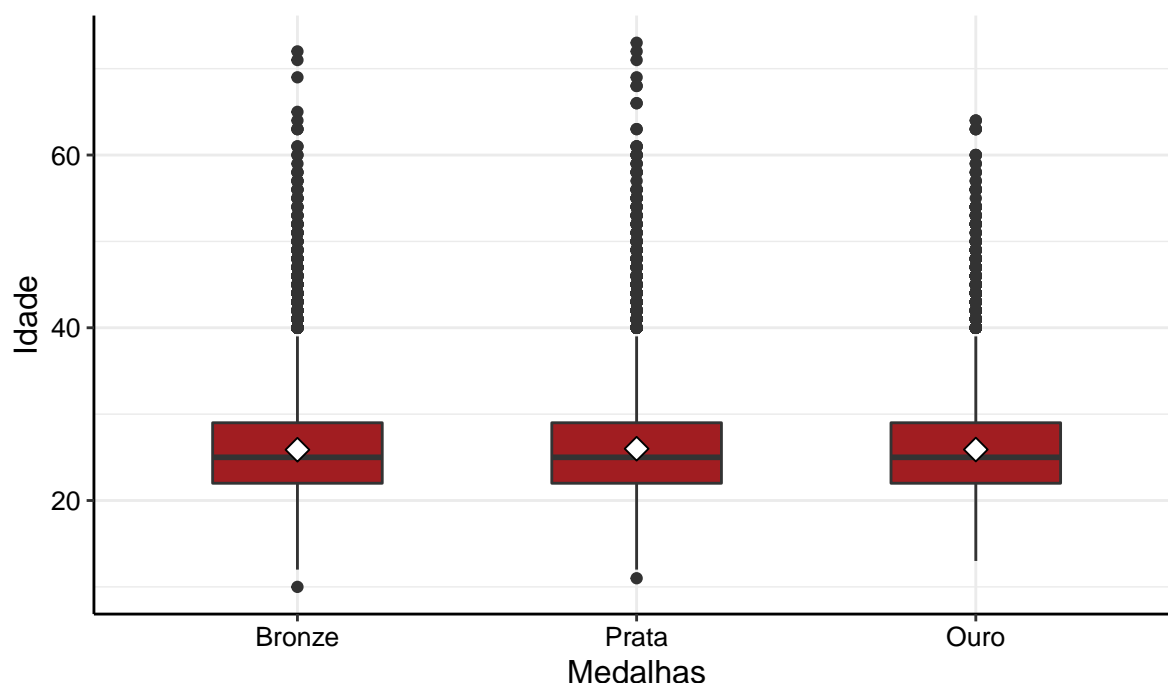
A fim de se entender uma possível correlação entre número de medalhas e idade, tem-se uma análise das frequências relativas viabilizada por meio da divisão do número de medalhas obtidas por faixa etária pela quantidade de atletas pertencentes a essa.

No banco presente, havia um apenas um atleta com a idade mínima de 10 anos, sendo que o mesmo foi premiado com uma medalha. Desse modo, acredita-se que com uma amostra mais robusta a frequência esperada seria diferente de 100%, ou seja, não necessariamente há uma vantagem absoluta de atletas com essa idade específica. Em contraposição, ao comparar o coeficiente de correlação apresentado ao considerar e desconsiderar tal singularidade, observou-se uma pequena variação decimal desse.

Nesse sentido, o coeficiente de correlação de Pearson pode auxiliar o entendimento de que há uma relação fraca entre idade e número de medalhas, essa caracterizada como de grandeza inversamente proporcional, ou seja, há uma tendência de

que quanto maior a idade do atleta, maior a dificuldade de que ele seja premiado com uma medalha.

Figura 5: Boxplot da idade por medalhas



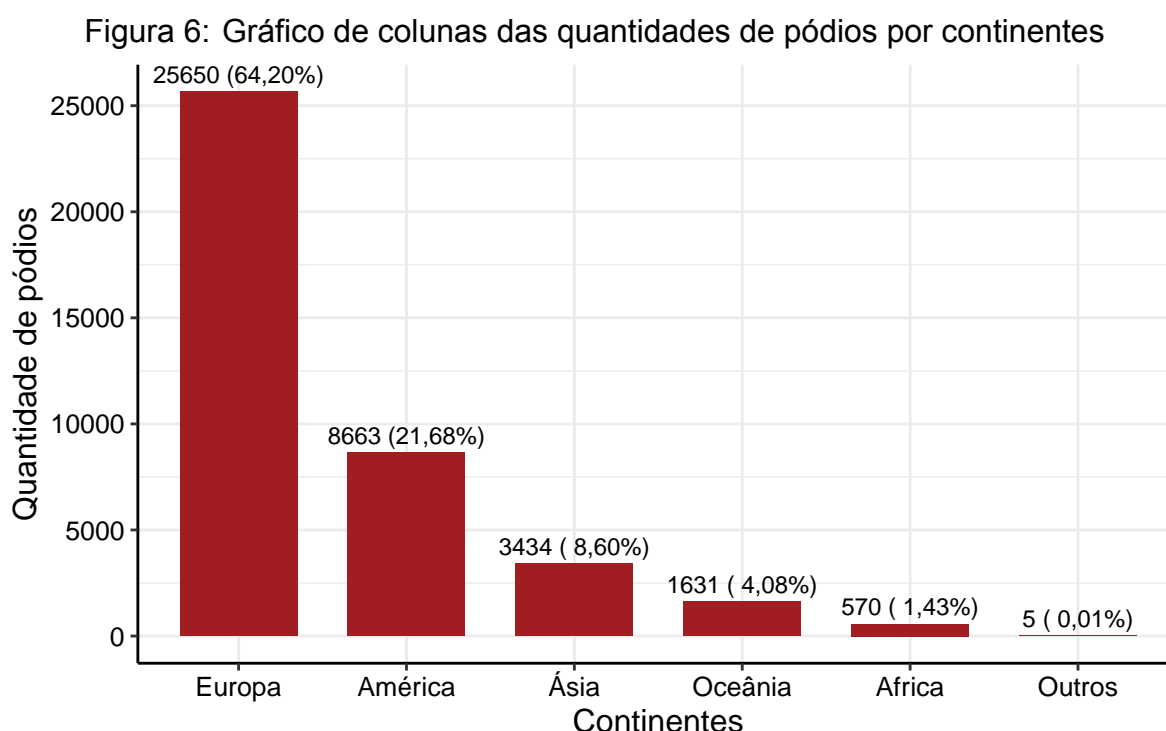
Quadro 2: Medidas resumo da idade dos participantes por tipo de medalhas

Estatística	Bronze	Prata	Ouro
Média	25,9	26,0	25,9
Desvio Padrão	5,8	6,0	5,9
Mínimo	10	11	13
1º Quartil	22	22	22
Mediana	25	25	25
3º Quartil	29	29	29
Máximo	72	73	64

Ao considerar o tipo de medalhas, segundo Figura 5 e Quadro 2, pode-se observar que as médias das idades de atletas olímpicos são muito semelhantes entre a classificação de qual medalha foi conquistada. No que tange ao desvio padrão, as idades dos atletas que conquistaram medalha de bronze apresentaram uma menor dispersão quando comparada aos demais, apesar da proximidade dos valores em questão.

3.3 Quantidade de pódios por continente

Nessa seção, serão expostas as quantidades de pódios por continentes. Desse modo, tem-se a seguir um gráfico de colunas, no qual estão representadas as quantidades e suas respectivas porcentagens.



Como pode ser visto, o continente em que mais atletas fizeram parte de um pódio Olímpico foi o Europeu, com 25650 atletas, o que representa cerca de 64% desses. Em contrapartida, o continente Africano, com 570 atletas, apresentou 45 vezes menos do que o continente Europeu, sendo assim, o que menos apresentou competidores donos de pódios. Já a Oceânia, Ásia e América, encontram-se em posições intermediárias.

Outrossim, a coluna "Outros" remete aos Participantes Olímpicos Independentes (IOP) ou Atletas Olímpicos Individuais/Independentes (IOA), que participam sob a bandeira Olímpica.

3.4 Distribuição dos IMCs por jogos de verão e inverno

Faz-se válido ressaltar que o Índice de massa corporal, conhecido pela sigla 'IMC', corresponde a um índice para calcular o peso ideal de um indivíduo. Além disso, ele é calculado por meio da divisão da massa em quilogramas de uma pessoa por sua altura

em metros elevada ao quadrado. Por conseguinte, a interpretação desse índice consiste na verificação de em qual faixa esse se encontra e o diagnóstico correspondente.


Tabela 2: Interpretação do 'IMC'

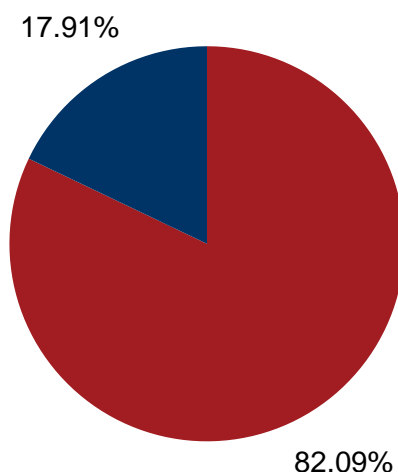
¹ Intervalo	Classificação	Grau de obesidade
Menor que 18,5	Magreza	0
Entre 18,5 e 24,9	Normal	0
Entre 25,0 e 29,9	Sobrepeso	1
Entre 30,0 e 39,9	Obesidade	2
Maior que 40	Obesidade grave	3

¹Fonte.

Calculadora de IMC- Programa Saúde Fácil

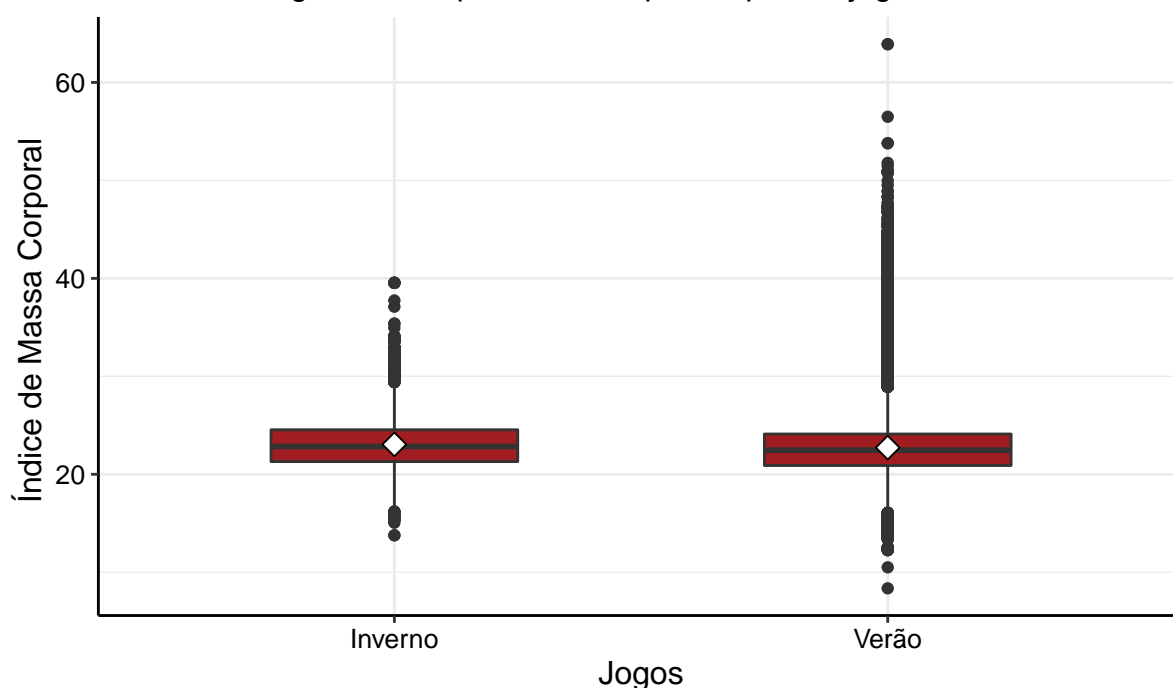
Figura 7: Gráfico de setores dos tipos de jogos

Jogos  Inverno  Verão



Os jogos ocorridos no verão são maioria no banco de dados analisado, correspondendo a aproximadamente 82%. Tal estatística segue dessa forma, em razão, da transição da coexistência dos Jogos de Inverno e de Verão no mesmo ano até 1992. Posteriormente, foram escalonados de forma que os Jogos de Inverno ocorressem em um ciclo de quatro anos começando em 1994, depois no verão em 1996, no inverno em 1998 e assim por diante.

Figura 8: Boxplot dos IMC pelos tipos de jogos



Quadro 3: Medidas resumo dos IMC

Estatística	Inverno	Verão
Média	23,0	22,7
Desvio Padrão	2,5	3,0
Mínimo	13,8	8,4
1º Quartil	21,3	20,9
Mediana	22,8	22,5
3º Quartil	24,5	24,1
Máximo	39,5	63,9

Com auxílio do quadro e gráfico acima, percebe-se que o IMC dos atletas foi, em média, levemente superior durante os jogos de inverno do que nos de verão. Por outro lado, os índices durante os jogos de verão variaram mais, apresentando um valores de "IMC" mínimo de 8,4 e um máximo de 63,9 .

A título de curiosidade, buscou-se entender, se dentre as observações mais destoantes da média há padrões como, por exemplo, se essas correspondem a atletas de um mesmo esporte. Nesse viés, observou-se que dentre os 10 atletas com menor IMC em jogos de verão, a maioria competia pela modalidade Ginástica e, por outro lado, os atletas que apresentaram os maiores índices nas mesmas condições atuavam em Levantamento de peso e Judô. Em contrapartida, durante os jogos de inverno, os atletas com menores IMC praticavam Cross Country Skiing e os com mais Bobsleigh.

3.5 Top 5 participantes

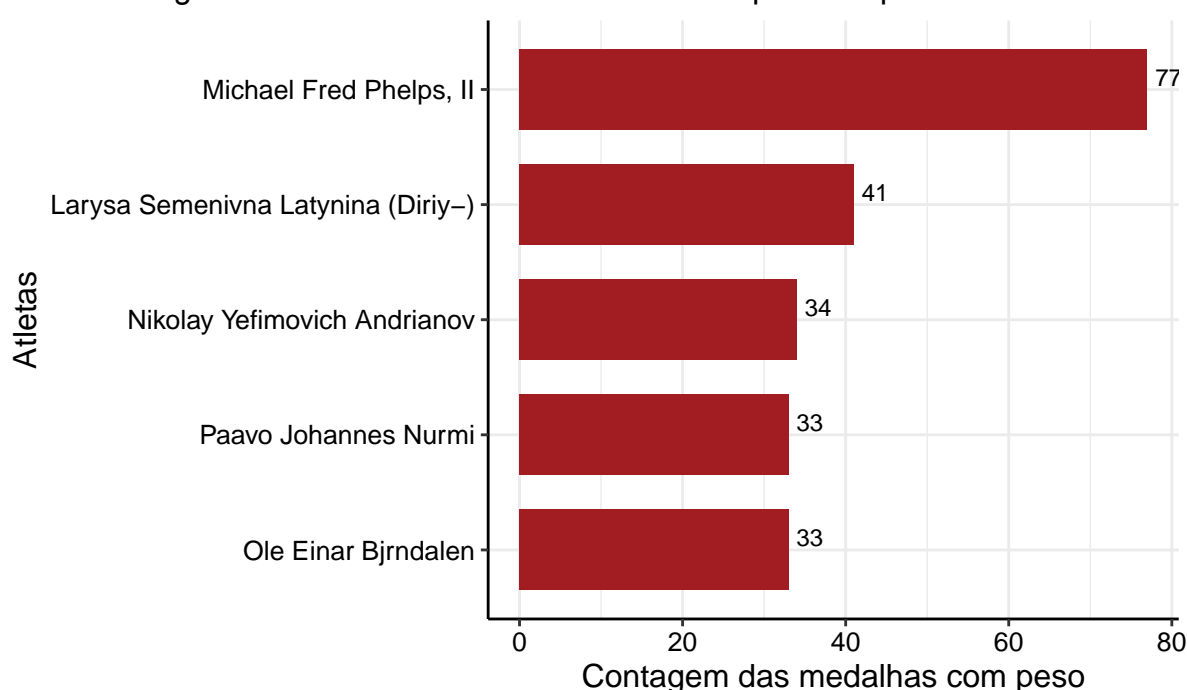
Para a confecção do gráfico a seguir foram usados os seguintes pesos:

Quadro 4: Pesos atribuídos para as medalhas

Medalha	Pontos
Ouro	3
Prata	2
Bronze	1

Por meio desses, obteve-se :

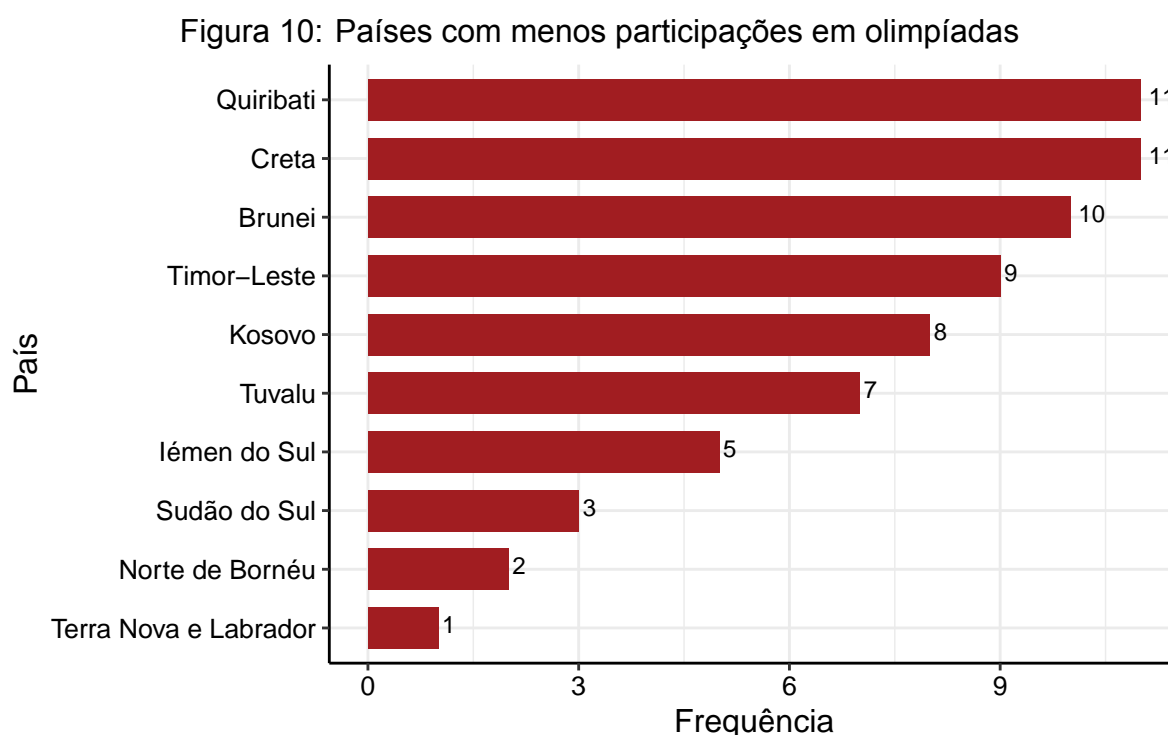
Figura 9: Gráfico de barras dos 5 atletas que mais pontuaram



Como pode ser visto, o nadador Michael Phelps teve um total de 77 pontos, o que o garante uma primeira posição de efetividade notória, com mais que o dobro de pontos que qualquer atleta abaixo da segunda colocação, a qual pertence a ex ginasta Larysa Latynina. Além disso, o também ginasta Nikolay Andrianov obteve a terceira posição do ranking acima, com 34 pontos. Por fim, os atletas em quarta e quinta posição, Paavo NurmiNikolay e Ole Einar Bjrndalen, respectivamente, obtiveram pontuação empatadas em 33.

3.6 Países com menor número de participações

Já referente às participações por países, observou-se a participação de no total 230 países diferentes, dentre os quais, listam-se a seguir os que menos participaram.



Quadro 5: Medidas resumo do número de países participantes

Estatística	Participações
Média	1179
Mínimo	1
Mediana	182
Máximo	18853

Dentre esses, destacam-se os 10 que menos participaram durante as edições do Jogos Olímpicos. Desse modo, nota-se Terra Nova e Labrador, província canadense, com apenas 1 participação, Norte de Bornéu com 2, Sudão do Sul, 3, e assim por diante, segundo Figura 10. Em síntese, percebe-se que os países citados apresentam número de participações olímpicas bem inferiores a quaisquer medidas de tendência central apresentadas, média e mediana .

4 Conclusão

A partir dos resultados obtidos, foi possível observar em primeiro plano a forma com que as participações Olímpicas por gêneros evoluíram ao longo do tempo, sendo as masculinas majoritárias durante esse contexto. De forma similar, a correlação negativa da idade dos atletas e ganho de medalhas também foi vislumbrada.

Análogo a isso, pôde ser visualizado o conjunto de países que, no decorrer dos eventos, menos participaram, sendo Terra Nova e Labrador o principal destaque com menos participações, assim como também vislumbrada a quantidade de pódios agrupadas por continente percebendo-se a hegemonia numérica das participações europeias.

Por fim, acredita-se ter sido proporcionada a compreensão de como foram distribuídos os índices de massa corporal dos atletas em diferentes tipos de eventos, percebendo-se a sutil superioridade dos IMC's durante os jogos de inverno em comparativo aos de verão. De forma correlata, acredita-se, também, ter sido proporcionada a compreensão do desempenho dos atletas que mais se destacaram no quesito ganho de medalhas em um ranking. Sendo esses, em ordem decrescente de pontuação: Michael Fred Phelps com liderança absoluta, Larysa Semenivna, Nikolay Yefimovich, Paavo Johannes e Ole Einar.