

Relatório sobre os atletas nas Olimpíadas 2000-2016

Consultores Responsáveis:

Tales Vaz

Requerente:

João Neves

Brasília, 16 de janeiro de 2025.



Sumário

	Página
1 Introdução	3
2 Referencial Teórico	4
2.1 Frequência Relativa	4
2.2 Média	4
2.3 Mediana	4
2.4 Quartis	5
2.5 Variância	5
2.5.1 Variância Populacional	5
2.5.2 Variância Amostral	6
2.6 Desvio Padrão	6
2.6.1 Desvio Padrão Populacional	6
2.6.2 Desvio Padrão Amostral	7
2.7 Coeficiente de Variação	7
2.8 Boxplot	7
2.9 Gráfico de Dispersão	8
2.10 Tipos de Variáveis	9
2.10.1 Qualitativas	9
2.10.2 Quantitativas	9
2.11 Coeficiente de Correlação de Pearson	10
3 Análises	11
3.1 Top 5 países com o maior número de mulheres medalhistas	11
3.2 Valor do IMC por esporte	12
3.3 Top 3 medalhistas gerais por quantidade de cada tipo de medalha	13
3.4 Variação Peso por altura	15
4 Conclusão	17

top3<-c("Nome","Medalha","Quantidade","Frequência","Quantidade2","Acumulado","Acumulad

1 Introdução

Este relatório tem como objetivo realizar análises estatísticas que visam otimizar o desempenho de todos os atletas que participaram das Olimpíadas de Verão entre os anos de 2000 até 2016.

Nesse projeto é apresentado quatro análises sobre diversos aspectos, características e feitos esportivos, com o propósito de auxiliar no melhor entendimento sobre os atletas, dentre elas, foi questionado quais foram os 5 países que mais tiveram mulheres medalhistas, qual o IMC dos atletas em cada esporte, este sendo, a ginástica, futebol, judô, atletismo e badminton, também foi foco de estudo quais eram os 3 atletas com mais medalhas no período de 2000 até 2016 e, por fim, buscou-se correlacionar o peso com a altura de cada atleta e entender sua variação. Para essas análises, a filtragem de apenas atletas premiados pelo seu desempenho com pelo menos uma medalha foi aplicada.

A coleta dos dados foi feita pela House of Excellence, contendo 67474 observações em 9 variáveis diferentes com informações importantes para a compreensão dos diversos aspectos que dão as características de cada atleta, entre as variáveis coletadas estão o nome do indivíduo, o seu gênero, idade, altura e peso, o país que representou nessas edições dos Jogos Olímpicos, o esporte em que compete, a modalidade ou evento ou categoria que, dentro do esporte, o atleta em questão disputou e, por último, se conquistou medalha ou não e, caso tenha conquistado, o tipo dela.

Toda a análise estatística, gráficos, tabelas e quadros feitos neste relatório foram produzidos a partir do software RStudio na versão “R-4.4.3”.

2 Referencial Teórico

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:

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

2.4 Quartis

Os quartis são separatrizes que dividem o conjunto de dados em quatro partes iguais. O primeiro quartil (ou inferior) delimita os 25% menores valores, o segundo representa a mediana, 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 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. Ele 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 Coeficiente de Variação

O coeficiente de variação fornece a dispersão dos dados em relação à média. Quanto menor for o seu valor, mais homogêneos serão os dados. O coeficiente de variação é considerado baixo (apontando um conjunto de dados homogêneo) quando for menor ou igual a 25%. Ele é dado pela fórmula:

$$C_V = \frac{S}{\bar{X}} \times 100$$

Com:

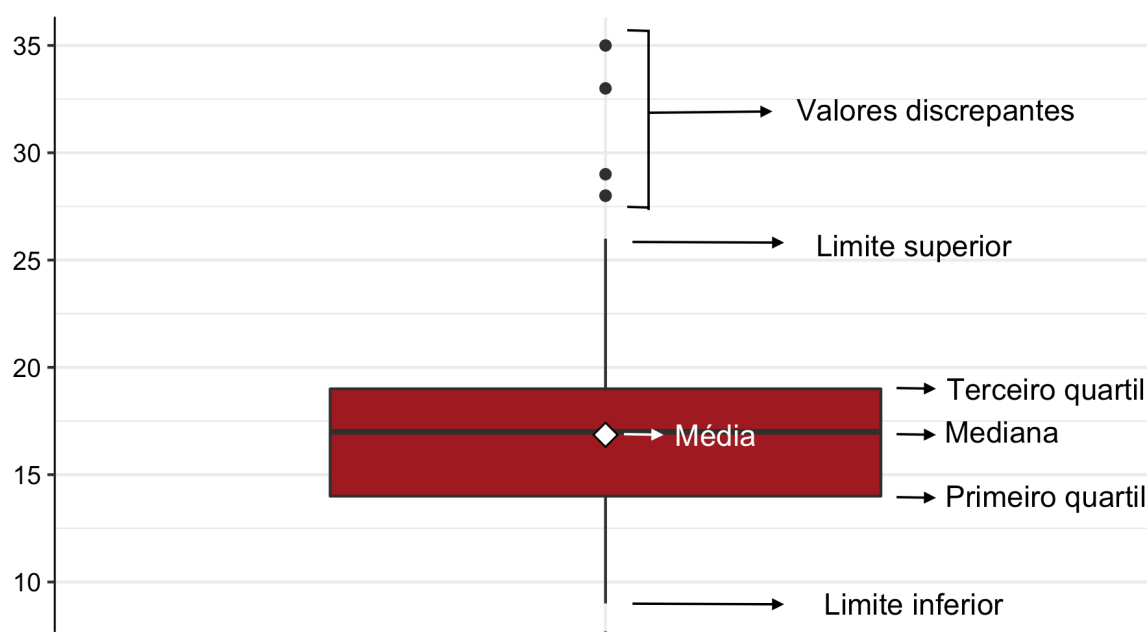
- S = desvio padrão amostral
- \bar{X} = média amostral

2.8 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.

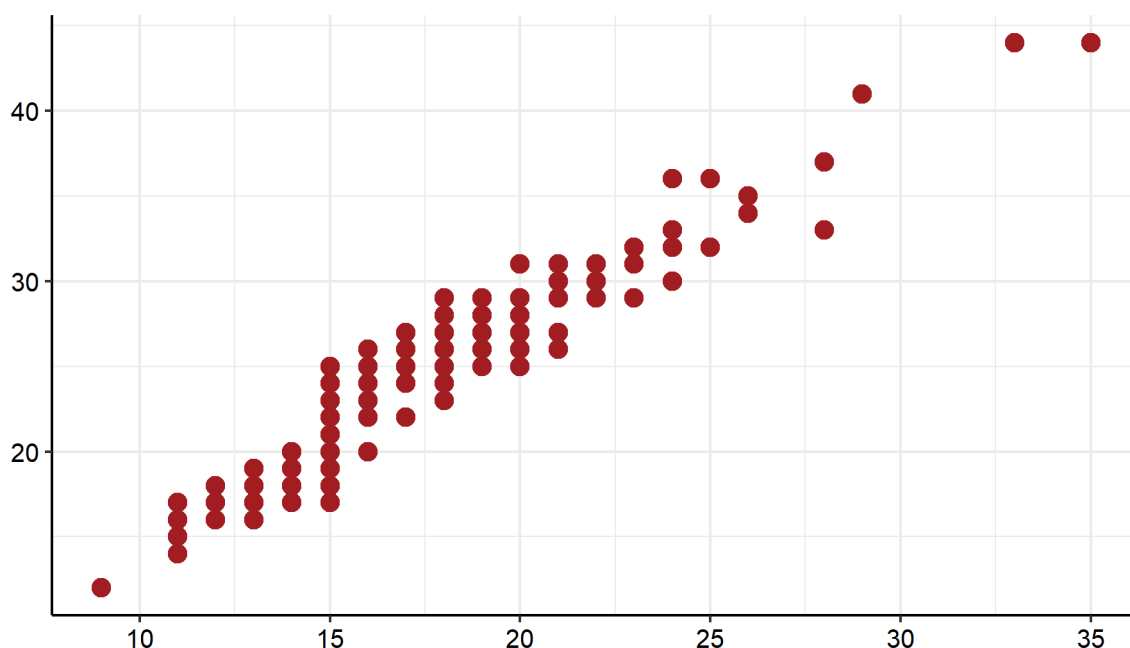
Figura 1: Exemplo de boxplot



2.9 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.

Figura 2: Exemplo de Gráfico de Dispersão



2.10 Tipos de Variáveis

2.10.1 Qualitativas

As variáveis qualitativas são as variáveis não numéricas, que representam categorias ou características da população. Estas subdividem-se em:

- **Nominais:** quando não existe uma ordem entre as categorias da variável (exemplos: sexo, cor dos olhos, fumante ou não, etc)
- **Ordinais:** quando existe uma ordem entre as categorias da variável (exemplos: nível de escolaridade, mês, estágio de doença, etc)

2.10.2 Quantitativas

As variáveis quantitativas são as variáveis numéricas, que representam características numéricas da população, ou seja, quantidades. Estas subdividem-se em:

- **Discretas:** quando os possíveis valores são enumeráveis (exemplos: número de filhos, número de cigarros fumados, etc)
- **Contínuas:** quando os possíveis valores são resultado de medições (exemplos: massa, altura, tempo, etc)

2.11 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

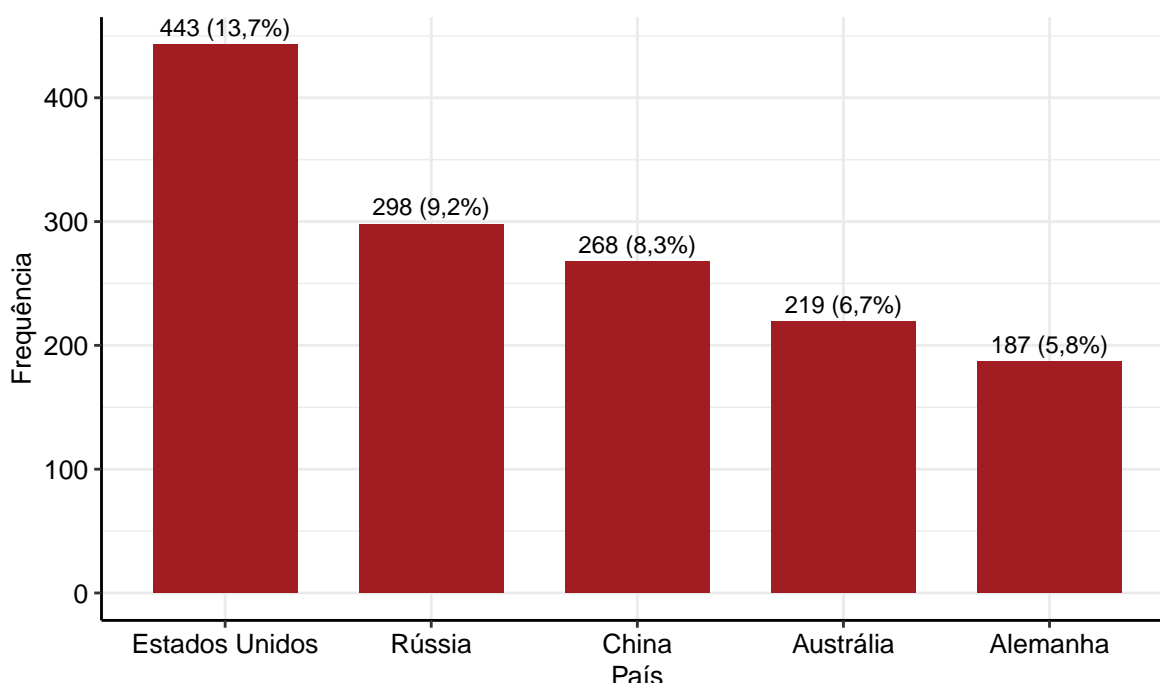
Vale ressaltar que o coeficiente de Pearson é paramétrico e, portanto, sensível quanto à normalidade (simetria) dos dados.

3 Análises

3.1 Top 5 países com o maior número de mulheres medalhistas

Nesta análise, foi feito o estudo da variável quantitativa discreta número de medalhistas femininas, pela variável qualitativa nominal país. Foram coletados dados de 94 países em que tiveram alguma mulher medalhista e, as 5 nações que tiveram mais representantes femininas no pódio durante todo o período entre as Olimpíadas de 2000 em Sydney até a de 2016 no Rio de Janeiro foram observadas e analisadas.

Figura 3: Gráfico de colunas do número de medalhas por país



Analisando cada nação, percebe-se por meio da **Figura 3** que o número de medalhistas é desigual. O gráfico indica que os 5 países que mais tiveram medalhistas na categoria feminina nos Jogos Olímpicos entre as edições de 2000 até 2016 foram, em primeiro lugar de maneira isolada, os Estados Unidos com 443 medalhistas, em seguida, a Rússia com 298, na terceira posição se encontra a China com 268, em quarto colocado a Austrália com 219 e mais afastada dos demais, a Alemanha com 187 medalhistas em quinto colocado fechando o ranking das nações analisadas. Observa-se uma notória liderança dos Estados Unidos com 4.5 pontos percentuais de diferença e 145 medalhistas em comparação ao segundo colocado na análise, Rússia. Dessa forma, indicando que existe, nos Jogos Olímpicos, a soberania de atletas estadunidenses na categoria feminina no período analisado.

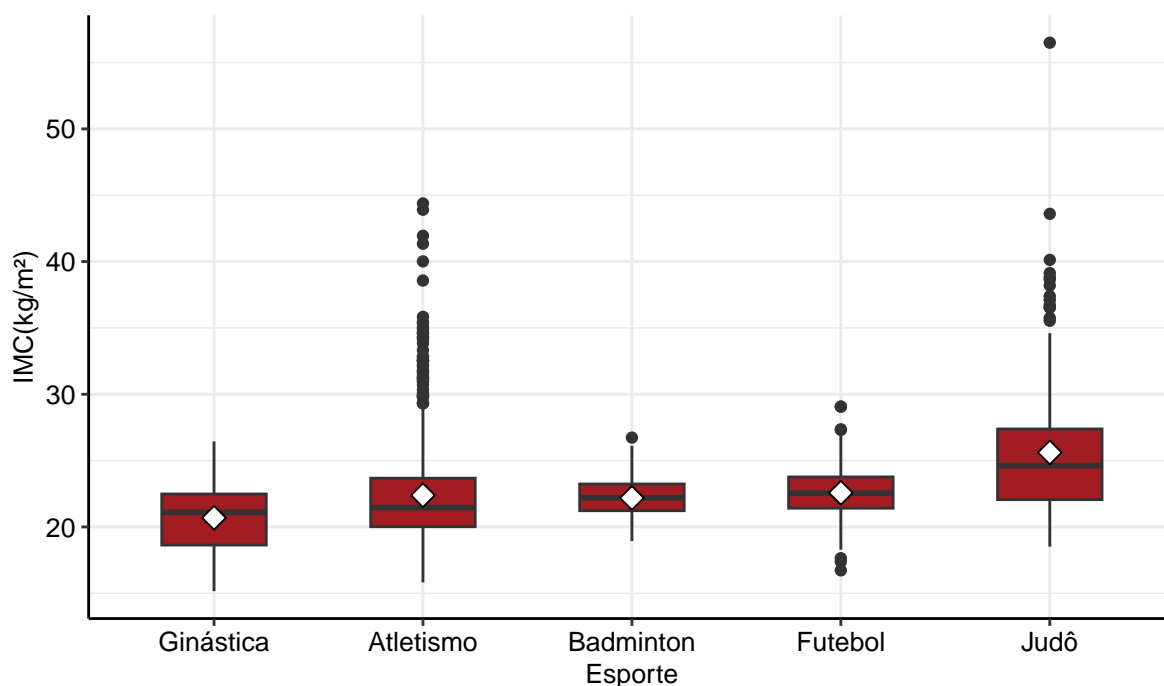
3.2 Valor do IMC por esporte

O IMC, Índice de Massa Corporal, é uma medida simples e internacional utilizada para avaliar o estado nutricional de uma pessoa, considerando sua altura e peso, podendo classificar os indivíduos em diferentes faixas de peso. O cálculo desse índice se dá pela fórmula:

$$IMC = (\text{peso}/\text{altura}^2)$$

Tendo isso em vista, foi feita uma análise descritiva da variável quantitativa contínua IMC de atletas olímpicos em 5 diferentes modalidades, uma variável qualitativa nominal, sendo essas: Atletismo, Ginástica, Badminton, Futebol e Judô.

Figura 4: Gráfico Boxplot do IMC do atleta por esporte



Quadro 1: Medidas resumo do IMC por esporte

Estatística	Atletismo	Badminton	Futebol	Ginástica	Judô
Média	22,39	22,18	22,57	20,69	25,61
Desvio Padrão	4,01	1,59	1,77	2,42	5,05
Variância	16,10	2,52	3,13	5,86	25,51
Mínimo	15,82	18,94	16,73	15,16	18,52
1º Quartil	20,02	21,22	21,41	18,63	22,06
Mediana	21,46	22,20	22,55	21,09	24,62
3º Quartil	23,67	23,23	23,77	22,48	27,38
Máximo	44,38	26,73	29,07	26,45	56,50

Avaliando o IMC dos esportes pela **Figura 4**, é perceptível que cada esporte possui um IMC específico em que a maioria dos participantes tendem, sendo, por exemplo, necessário um Índice de Massa Corporal menor para atletas da ginástica, enquanto os atletas de judô tendem a ser mais pesados.

Pelo **Quadro 1** é possível observar uma grande dispersão nos valores das modalidades do atletismo e no judô, contendo o desvio padrão de 4,01 e 5,05 respectivamente, isso se deve pelo fato de existir divisões de peso no caso do judô, indo de extra-peso-leve até o peso-pesado, e no caso do atletismo possui desde atletas de lançamento de peso e até maratonistas, nos quais possuem uma composição corporal diferente. Além disso, percebe-se a existência de diversos outliers, valores atípicos, nesses esportes, contendo até um valor de IMC maior que 56 no judô e maior que 44 no caso do atletismo.

Outro caso interessante de se analisar é a simetria dos valores do IMC dos atletas do Badminton e do futebol, isto é, bem concentrados e de baixo desvio-padrão, sendo respectivamente de 1,59 e 1,77, indicando que exista, uma relação entre peso e altura a ser buscada entre eles.

Apesar da facilidade do uso desse índice, é importante ressaltar que ele não leva em conta fatores como a composição corporal, a distribuição de gordura ou a saúde metabólica individual.

3.3 Top 3 medalhistas gerais por quantidade de cada tipo de medalha

Nesta análise, o objetivo é descobrir quem são os três atletas que mais conquistaram alguma medalha no período de observação, a variável qualitativa nominal nome, e quais foram os tipos de medalhas, uma variável qualitativa ordinal, conquistas por eles, com isso, entender se existe relação entre o medalhista e cada tipo de medalha conquistada. Para isso, foi feita uma análise descritiva desses 3 atletas em questão.

Figura 5: Gráfico de colunas bivariado da frequência e do tipo de medalha por atleta

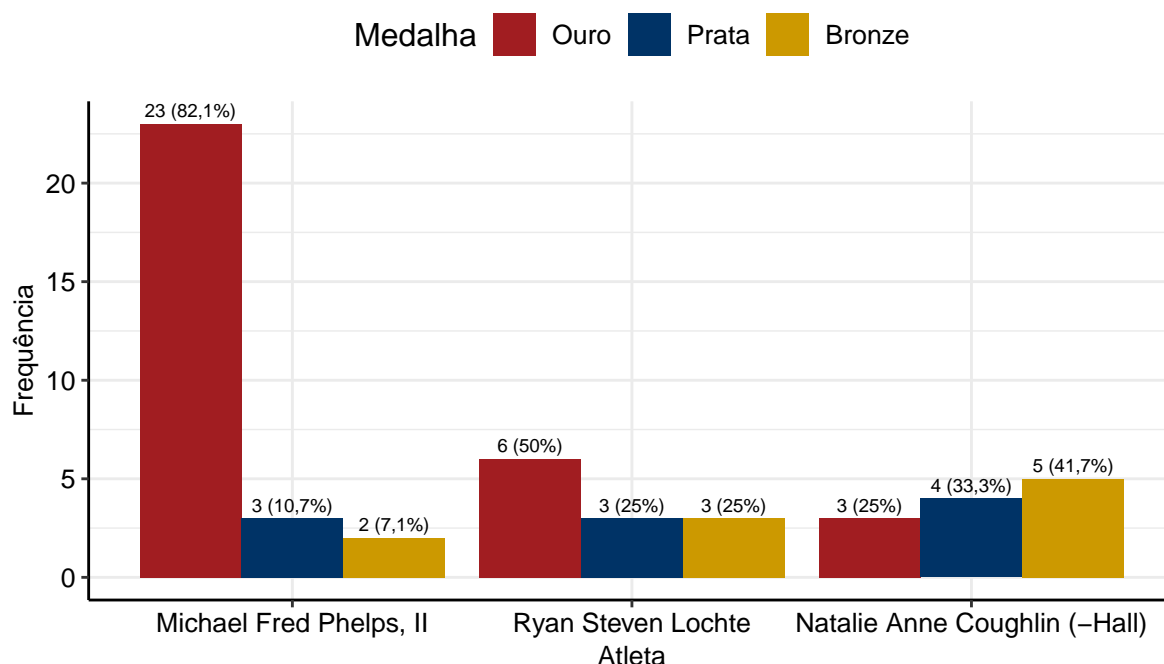


Tabela 1: Frequência e tipo das medalhas

Medalha	Frequência	Porcentagem
Ouro	32	61,54%
Prata	10	19,23%
Bronze	10	19,23%
Total	52	100,00%

Para essa análise podemos observar os 3 atletas que mais conquistaram medalhas nas Olimpíadas 2000-2016, pela **Figura 5**, é possível observar que o estadunidense Michael Phelps foi o atleta mais vezes campeão, sendo em 23 oportunidades, mais do que qualquer outro atleta, sendo 82,1% de todas as suas medalhas e também foi o que mais vezes ganhou alguma medalha, com um total de 28 no período de quatro Olimpíada. Já a atleta Natalie Coughlin que, entre os 3 atletas, foi a que menos teve medalhas de ouro com apenas 3 entre as 12 totais conquistadas, a mesma quantidade de medalhas de Ryan Lochte porém o mesmo conquistou em sua maioria ouro, sendo 6 delas.

Pelo **Tabela 1** é possível observar que entre os atletas analisados há uma tendência ao tipo da medalha conquistada, o ouro, uma vez que essa, representa 61,54% das 52 medalhas conquistadas por eles, 32 no total.

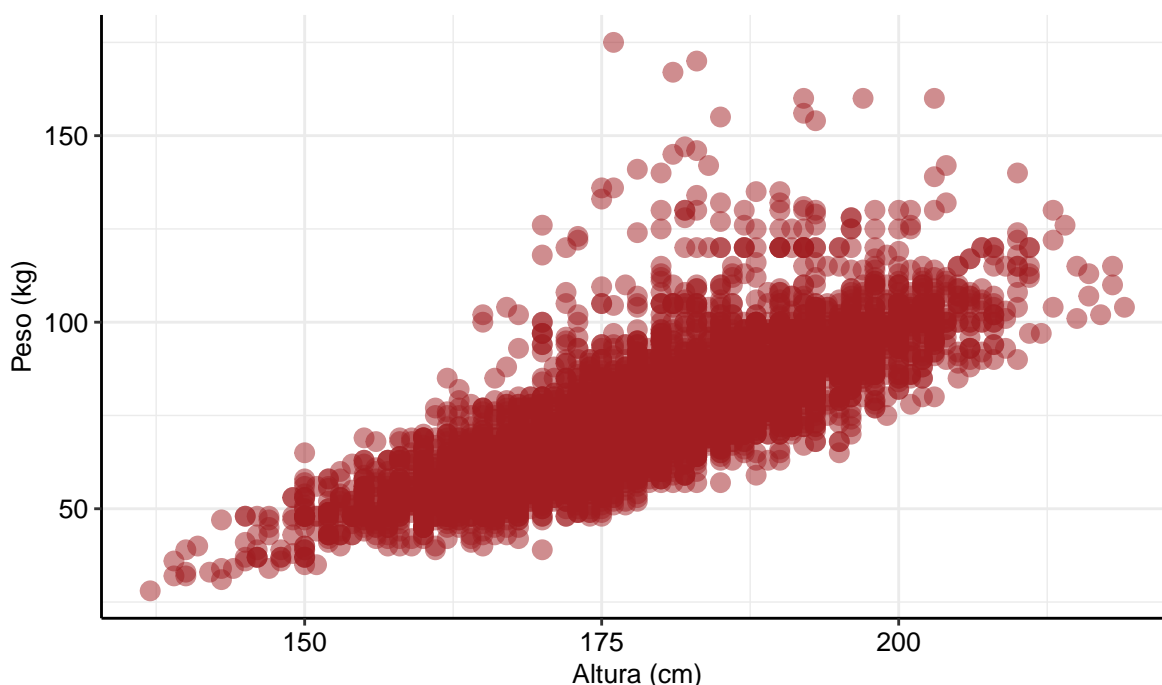
Todos os 3 indivíduos analisados com mais medalhas são atletas de natação, isso se deve pelo fato de existirem diversas divisões nesse esporte o que dá a chance de

disputarem por mais medalhas.

3.4 Variação Peso por altura

A relação entre a altura e o peso do atleta é um importante dado ao estudar sobre o desempenho esportivo. Com o objetivo de compreender melhor essas duas variáveis quantitativas contínuas, foi feita uma análise descritiva sobre a variação entre essas duas medidas, além disso, buscou-se compreender a intensidade da relação entre elas.

Figura 6: Gráfico de dispersão do peso e da altura do atleta



Quadro 2: Medidas resumo da altura e peso dos atletas

Estatística	Altura (cm)	Peso (kg)
Média	178,21	74,14
Desvio Padrão	11,71	16,25
Variância	137,05	264,11
Mínimo	137,00	28,00
1º Quartil	170,00	63,00
Mediana	178,00	72,00
3º Quartil	186,00	84,00
Máximo	219,00	175,00

Analisando o gráfico da **Figura 6** juntamente com o Coeficiente de Correlação de Pearson de $r = 0,795$. É possível determinar que a relação entre o peso e a altura do

atleta é forte e positiva, ou seja, é diretamente proporcional, à medida que o peso do atleta aumenta, a altura tende a aumentar proporcionalmente e vice-versa.

O **Quadro 2** também serve de parâmetro para observar que, entre os atletas, existe uma grande amplitude entre essas características, na altura por exemplo, o mais alto mede 2,19 metros enquanto o menor mede 1,37 metros, já no peso, o mais pesado possui 175 quilos e o mais leve tem 28 quilos apenas. Além disso, é possível observar que os valores do peso variam mais em comparação aos valores da altura, evidenciado pelo coeficiente de variação de 6,56 para a altura e de 21,91 para o peso, isso se deve pelo fato de não se possível existir muita diferença na altura de um atleta médio, porém, o peso pode variar bastante entre eles, dependendo do esporte praticado.

4 Conclusão

O projeto realizado, teve como objetivo fornecer uma análise estatística detalhada sobre atletas que participaram dos Jogos Olímpicos de Sidney em 2000 até a edição de 2016 sediada no Rio de Janeiro. foram objeto de estudo para essas análises 4 tópicos que de maneira sucinta auxiliaram para um entendimento amplo sobre tendências nos esportes e em características dos atletas.

Nas análises pôde ser compreendido a representatividade de mulheres e quais países desempenham melhor na categoria feminina através do questionamento dos 5 países em que mulheres mais vezes conquistaram medalhas. Além disso, o estudo sobre o IMC demonstrou que certas proporções corporais têm mais tendência para um melhor desempenho no esporte que outras. Saber quais são os atletas que mais vezes chegaram na posição de destaque é importante para questionar seus treinamentos para que seus feitos possam ser repetidos por outros. Entender que o existe uma proporção entre a variação do peso e a altura de atletas de altíssimo nível é importante para buscar essa tal relação nos futuros atletas para otimizar seus desempenhos.