Relatório Olimpíadas 2000 - 2016

Consultores Responsáveis:
Felipe Bretas

Requerente:
João Neves

Brasília, 7 de dezembro de 2024.

ESTAT
consultoria estatística

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 Boxplot 7
2.8 Gráfico de Dispersão 8
2.9 Tipos de Variáveis 8
2.9.1 Qualitativas 8
2.9.2 Quantitativas 9
2.10 Coeficiente de Correlação de Pearson 9
3 Análises 10
3.1 Relação de mulheres medalhistas por país 10
3.2 Relação do IMC por esporte 11
3.3 Relação entre medalhistas e o tipo de medalha 13
3.4 Relação entre peso e altura 15
4 Conclusões 17

1 Introdução

Este relatório tem como objetivo um estudo mais aprofundado, por meio de análises descritivas, sendo utilizado diversos gráficos, tabelas, quadros e também analisadas suas correlações, sobre a performance de atletas nas Olimpíadas dos anos 2000 até 2016, a fim de entender melhor o desempenho e possíveis relações desses atletas medalhistas entre essas edições desse grandioso torneio. Neste relatório foram feitas 4 análises, sendo estas: o número de mulheres medalhistas por país, sendo analisado os 5 maiores ganhadores; se existe uma possível relação entre atletas de determinados esportes e sua taxa de IMC; se existe algum tipo de relação entre o atleta medalhista e o tipo de medalha que ele conquistou e por fim, se existe alguma relação entre o peso e a altura desses atletas.

O banco de dados foi fornecido pelo próprio cliente, possuindo uma boa amostra para fins de estudo contendo um total de 9 variáveis e 10.017 observações presentes para a confecção da análise, possuindo variáveis tanto qualitativas, discretas e contínuas, como país de origem de cada atleta, nomes, quanto quantitativas, nominais e ordinais, como altura e peso dos atletas medalhistas.

Os softwares utilizados para a realização destas análises são o R na versão 4.3.3. e o Rstudio.

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:

- $j = 1, ..., c$
- $n_j =$ 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:

- $i = 1, 2, ..., n$
- $n =$ 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:

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

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
- $\mu =$ média populacional
- $N =$ tamanho da população

2.5.2 Variância Amostral

Para uma amostra, a variância é dada por:

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

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:

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

Com:

- $X_i =$ i-ésima observação da população
- $\mu =$ média populacional
- $N =$ tamanho da população

2.6.2 Desvio Padrão Amostral

Para uma amostra, o desvio padrão é dado por:

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

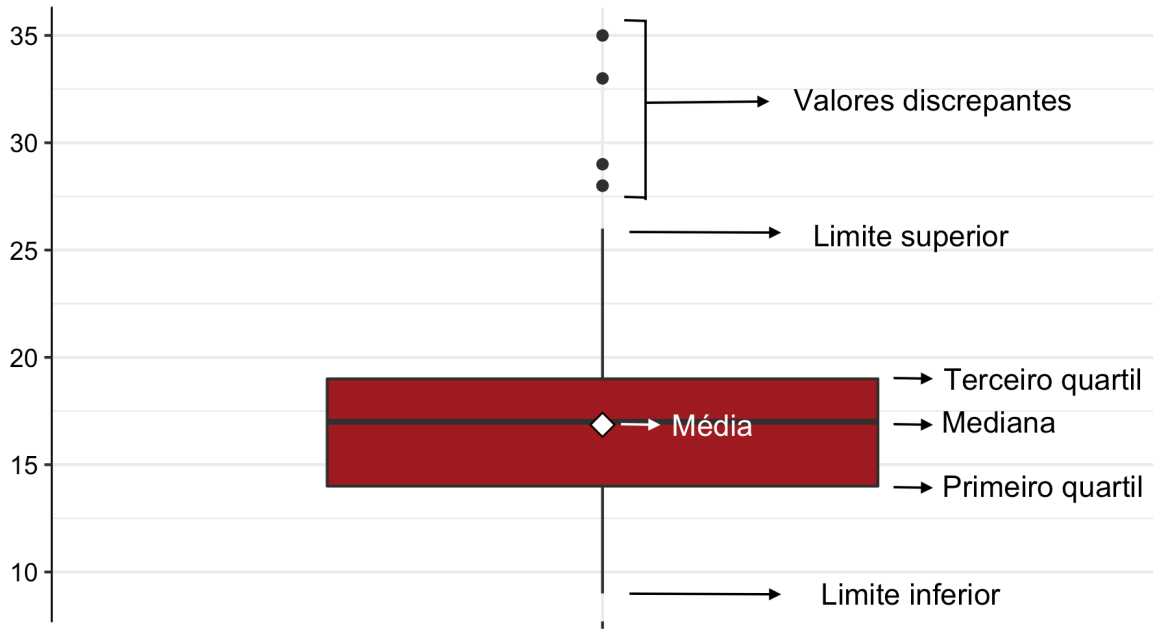
Com:

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

2.7 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.

Figura 1: Exemplo de boxplot



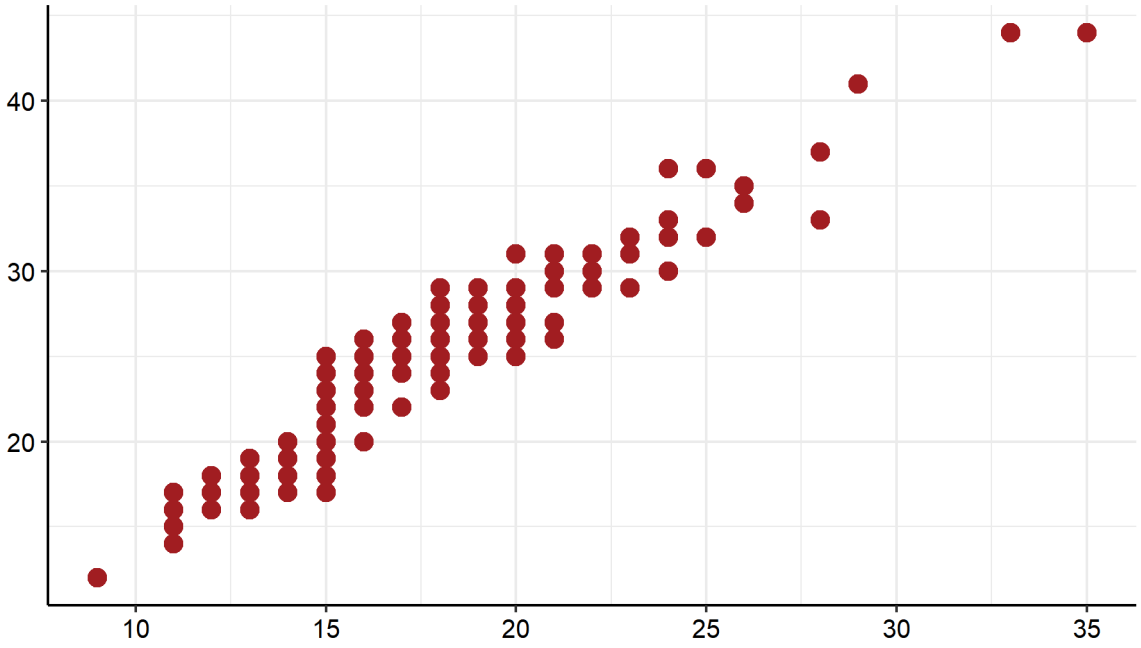
A porção inferior do retângulo diz respeito ao primeiro quartil, enquanto a superior indica o terceiro quartil. Já o traço no interior do retângulo representa a mediana do conjunto de dados, ou seja, o valor em que o conjunto de dados é dividido em dois subconjuntos de mesmo tamanho. A média é representada 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.

Figura 2: Exemplo de Gráfico de Dispersão



2.9 Tipos de Variáveis

2.9.1 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)

2.9.2 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)

2.10 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:

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

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

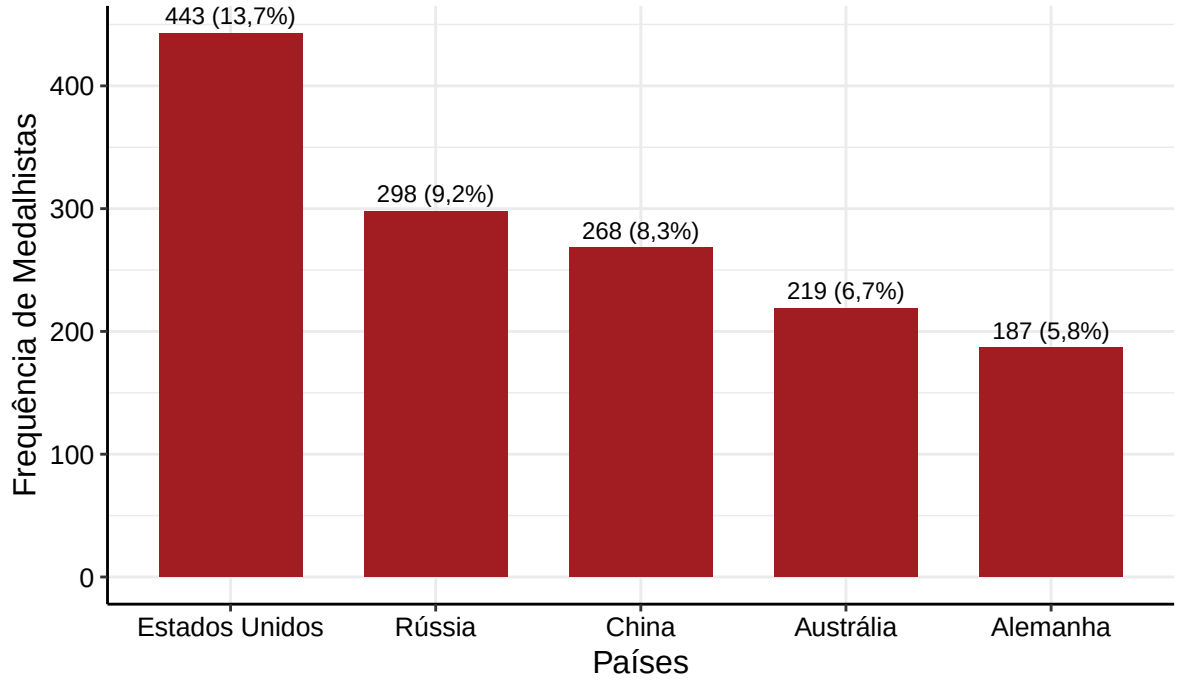
3.1 Relação de mulheres medalhistas por país

Desde o século XX, as mulheres vêm ganhando espaço em diversas áreas de atuação que antes eram consideradas exclusivamente dos homens. Uma conquista muito grande do público feminino foi o ingresso nas Olimpíadas e, com a virada do século, as participações se tornaram ainda maiores.

Para a confecção desta análise, foram utilizadas as variáveis: “Sexo” (qualitativa nominal), “Medalha” (qualitativa ordinal) e “País” (qualitativa nominal) que significam respectivamente: o sexo dos atletas participantes; se o atleta ganhou medalha e seu tipo, sendo estes: ouro, prata ou bronze; o país que o atleta representou.

Utilizando dessas variáveis, este estudo tem como finalidade compreender quais foram os países que tiveram o maior número de medalhistas femininas nas Olimpíadas de 2000 até 2016, analisando os 5 que se saíram mais vencedores nessas edições, como é mostrado na figura a seguir:

Figura 3: Gráfico de colunas do total de medalhas femininas por país



Como pode ser observado pela **Figura 3**, os Estados Unidos possuem o maior número de atletas femininas que conquistaram medalhas olímpicas, com um total de 443, representando 13,7% no total de medalhistas mulheres das Olimpíadas de 2000 até 2016, havendo uma diferença de 4,5% para o segundo colocado Rússia. Essa grande diferença percentual entre o primeiro e segundo colocado pode ser explicada por fatores externos que podem estar afetando essa grande diferença entre os dois.

Vale a pena ser ressaltado a grande variedade de continentes presentes dentro do top 5 medalhistas femininas, sendo estes: América, Europa, Ásia e Oceania, mostrando que existe uma variedade muito grande entre as atletas ganhadoras.

Algo que também deve ser observado é a frequência relativa acumulada entre esses 5 países mais vencedores os quais equivalem a 46,7% do total de medalhas ganhas por atletas femininas em todas as Olimpíadas em análise.

3.2 Relação do IMC por esporte

O Índice de Massa Corporal, mais conhecido como IMC, que possui como fórmula:

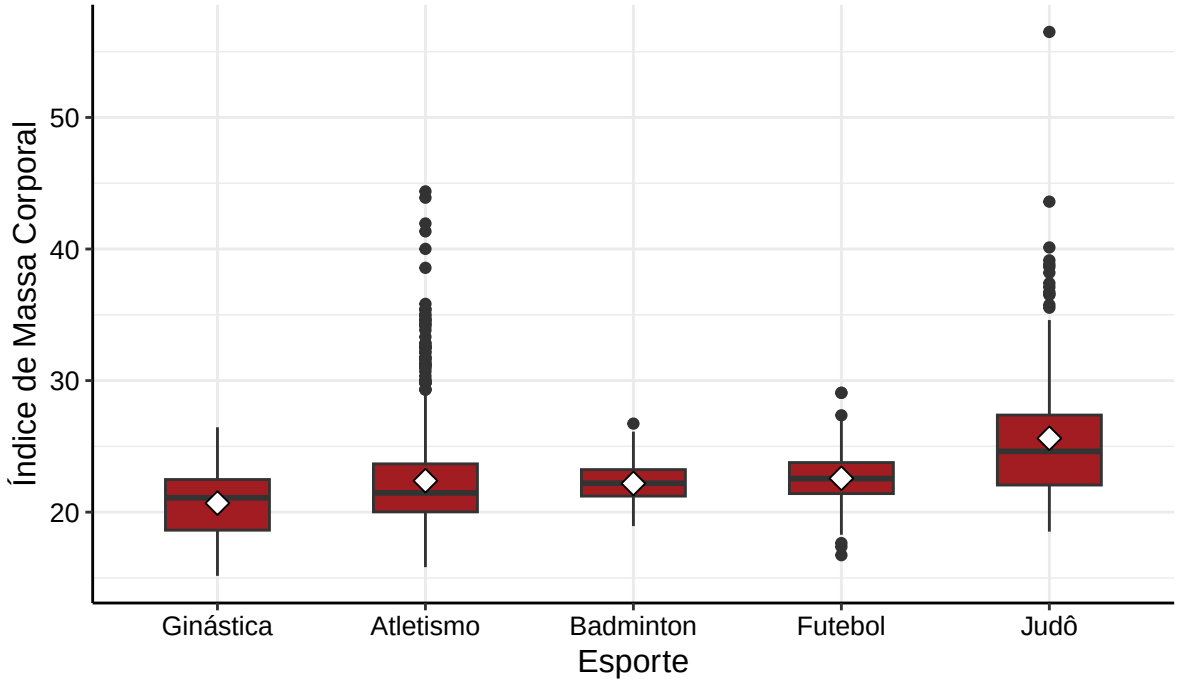
$$IMC = \frac{peso(kg)}{(altura(m))^2}$$

É uma medida internacional importante que tem como objetivo identificar se uma pessoa está com peso ideal, abaixo ou acima deste, podendo reconhecer possíveis condições de baixo peso, sobrepeso ou obesidade, ajudando na prevenção de doenças graves, como diabetes e hipertensão.

Para esta análise, foram utilizadas as variáveis: “Altura (m)” (quantitativa contínua), “Peso (kg)” (quantitativa contínua) e “Esporte” (qualitativa nominal), significando respectivamente: altura de cada atleta participante em metros; peso de cada atleta em quilogramas; esporte que o competidor participou.

A fim de tentar compreender se existe uma possível relação entre o IMC de alguns atletas e seus respectivos esportes, foi feita uma análise descritiva demonstrado pela figura e quadro abaixo:

Figura 4: Boxplot do IMC por esporte



Quadro 1: Medidas resumo do Índice de Massa Corporal

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	3,12	5,86	25,52
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	23,24	23,77	22,48	27,38
Máximo	44,38	26,73	29,07	26,45	56,50

Como pode ser observado pela **Figura 4** em conjunto com o quadro de medidas resumo, os atletas de atletismo são os que possuem a segunda maior variabilidade entre os atletas analisados, existindo uma grande diferença entre os valores de mínimo e de máximo. Há também uma diferença de somente 0,93 pontos entre o valor de sua média e sua mediana.

O badminton por sua vez é o que apresenta a menor variabilidade entre os esportes em questão, com uma diferença somente de 7,79 entre seus valores de máximo e mínimo. Algo que também chama a atenção nesse esporte são seus valores de média e de mediana, os quais apresentam valores muito semelhantes entre si, com uma diferença entre os dois de somente 0,02.

O futebol apresenta valores muito semelhantes quando comparado com o badminton, apresentando também uma diferença de somente 0,02 entre seus valores de média e mediana. Esse esporte também possui o segundo menor valor quando se trata de desvio padrão.

A ginástica é a que apresenta o menor valor de média que os demais esportes analisados, com um valor de 20,69. Também pode ser observado um aumento de variabilidade quando analisado com badminton e futebol, em que esse esporte apresenta o menor valor de mínimo entre todos em questão. Esse esporte também possui a menor mediana que os demais, com o valor igual a 21,09. Esse resultado menor que os demais, pode ser explicado por valores externos que vão além do banco de dados fornecido, podendo estar interferindo nesse valor.

O judô é o esporte que possui o maior valor de média, desvio padrão e mediana entre todos os esportes analisados. Ele também é o que possui a maior diferença entre os valores de mínimo e máximo, sendo este igual a 37,98.

Algo que também vale a pena ser observado é o grande número de outliers nos competidores de atletismo, judô e futebol, evidenciando a diversidade possível para os atletas entrarem em suas competições. Porém, mesmo com essa grande presença desse outliers, o atletismo ainda apresenta a segunda menor mediana (21,46) de IMC por atletas por esporte.

Como pode ser observado pelo quadro acima, a média entre os atletas de atletismo, badminton e futebol são bastante parecidas, mesmo com a grande variabilidade presente entre os profissionais de atletismo, todas um pouco acima de 22.

Como citado anteriormente, existe uma grande variabilidade presente entre os atletas de atletismo e principalmente os de judô, os quais possuem um desvio padrão respectivamente de 4,01 e 5,05. Algo que evidencia ainda mais essa grande variabilidade, são os valores mínimos e máximos para esses esportes, sendo seus valores de 15,82 e 18,52 como mínimo e 44,38 e 56,50 como máximo para seus respectivos esportes. Em contrapartida, os atletas de badminton e futebol apresentaram os desvios padrões mais baixo, 1,59 e 1,77, respectivamente, evidenciando uma baixa variabilidade em relação aos demais.

3.3 Relação entre medalhistas e o tipo de medalha

O sonho de cada atleta que adentra uma Olimpíadas é conquistar alguma medalha para seu país. Para isso é necessário anos de dedicação, treinos intensos e muito talento para ter uma chance de conquistar a tão sonhada medalha. Graças à combinação de todos esses fatores, existem competidores que conseguem atingir um maior sucesso que os demais, adquirindo um grande total de medalhas ao longo de vários anos.

A fim de tentar entender melhor quais são os atletas que possuem mais medalhas entre as Olimpíadas de 2000 até 2016, foi feita uma análise descritiva utilizando as variáveis: “Nomes” (qualitativa nominal), “Medalha” (qualitativa ordinal), significando respectivamente: o nome dos atletas participantes; medalha conquistada e seu tipo, sendo estes: ouro, prata e bronze. Além dessas variáveis, foi utilizado a figura e a tabela a seguir:

Figura 5: Gráfico de colunas bivariado de medalhas por atleta

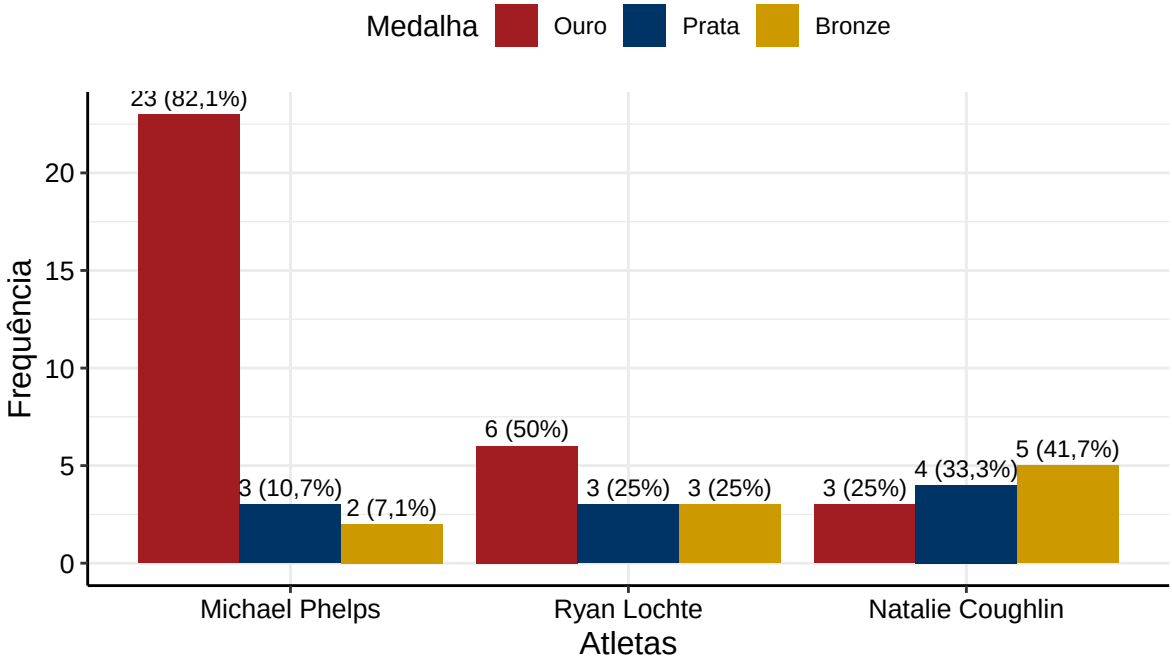


Tabela 1: Frequências da variável Medalha

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

Como pode ser observado pela **Figura 5**, o atleta Michael Phelps é quem detém o maior número de medalhas olímpicas entre 2000 até 2016, totalizando 28 medalhas, dentre elas 23 sendo de ouro, representando 82,10%. Para fechar o top 3 tem a atleta Natalie Coughlin e Ryan Lochte, ambos empatados no total de medalhas com 12 cada.

Como é evidenciado pelo quadro acima em conjunto com a **Figura 5**, existe uma certa relação quanto ao tipo da medalha conquistada e o atleta que a conquistou. Essa relação fica ainda mais evidente quando analisado o número de medalhas de

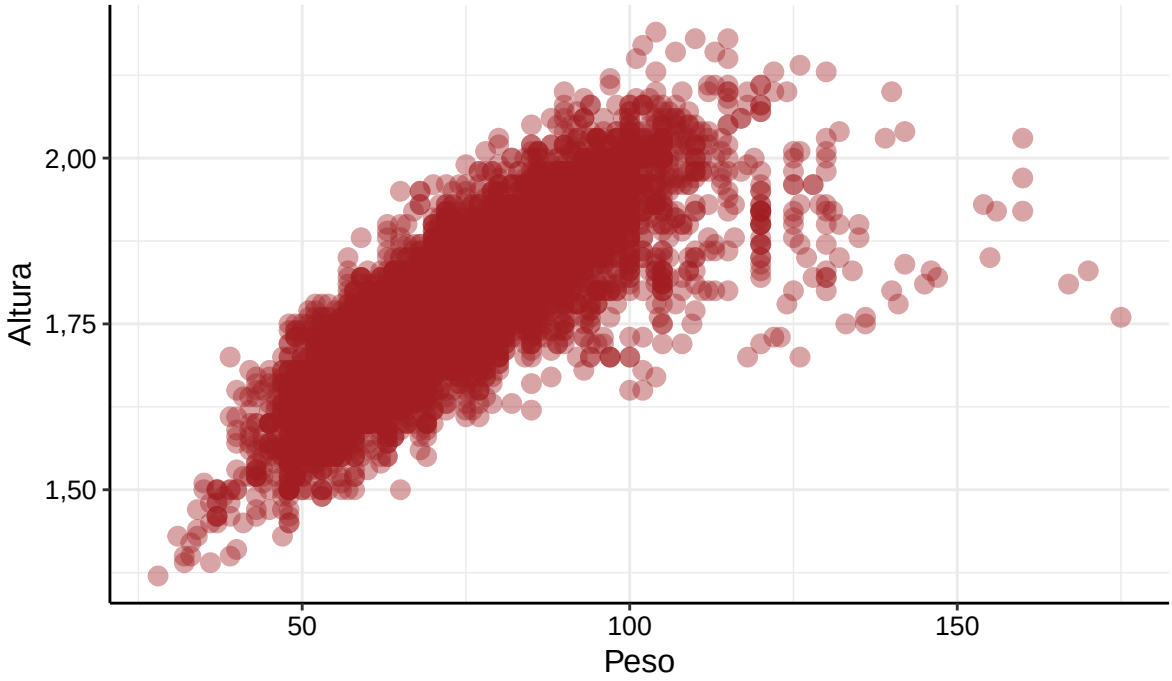
ouro conquistadas pelos 3 maiores ganhadores das Olimpíadas de 2000 até 2016, representando 61,54% do total de medalhas ganhas por esses atletas. Porém, dessas 32 medalhas de ouro, 28 são do Michael Phelps, o equivalente a 87,50% desse total.

3.4 Relação entre peso e altura

Algo bastante comum na fase de crescimento de todos os seres vivos é o aumento tanto de peso quanto de altura, podendo variar bastante os dois dependendo de cada indivíduo. Porém, isso nem sempre acontece, havendo uma grande variedade possível de alturas e pesos diferentes para todos. Nas Olimpíadas não é diferente, em muitos esportes, possuir uma certa faixa de peso ou uma altura maior, acaba se tornando uma vantagem.

A fim de tentar compreender melhor essa possível relação, foi feita uma análise descritiva, estudando sua correlação, utilizando das variáveis: “Peso (kg)” (quantitativa contínua) e “Altura (m)” (quantitativa contínua), possuindo seus respectivos significados: peso em quilogramas de cada atleta; altura em metros de cada competidor. Para compreender melhor esta análise foi feita a seguinte figura e o seguinte quadro:

Figura 6: Gráfico de dispersão do peso pela altura



Quadro 2: Medidas resumo do Peso e da Altura

Estatística	Peso	Altura
Média	74,14	1,78
Desvio Padrão	16,25	0,12
Variância	264,11	0,01
Mínimo	28,00	1,37
1º Quartil	63,00	1,70
Mediana	72,00	1,78
3º Quartil	84,00	1,86
Máximo	175,00	2,19

Como pode ser observado pela **Figura 6**, uma grande parte da amostra analisada possui entre 1,50 e 2,00 metros, e de forma semelhante, os atletas se encontram na faixa de peso entre 50 a 100 quilos, sendo destacada pela área mais densa do gráfico.

Para analisar sua correlação, foi utilizado o coeficiente de correlação de Pearson, o qual obteve um valor de r = 0,795, demonstrando que a relação é forte e positiva entre o peso e a altura das pessoas, o qual mostra que as duas variáveis são diretamente proporcionais, ou seja, quanto maior a altura de um indivíduo, maior tende a ser seu peso.

Como pode ser observado pelo quadro acima, a altura não apresenta uma grande variação com somente 0,12 de desvio padrão. Contudo, o peso existe uma grande variabilidade o qual possui 16,25 de desvio padrão mostrando que apesar de os competidores não haverem muita diferença quando se trata da altura, o peso muda consideravelmente, também sendo evidenciado pela sua diferença de 147 quilogramas entre o indivíduo mais pesado e o mais leve.

4 Conclusões

A análise sobre a relação entre o número de mulheres medalhistas, é possível observar a grande quantidade de medalhas conquistadas pelo Estados Unidos. Essa quantia maior que os demais, pode ser explicado por diversos fatores externos que afetam o desempenho dos atletas que não é mostrado somente pela análise do banco de dados.

Sobre a análise da possível relação entre o IMC dos atletas e seus respectivos esportes, não é possível dizer se ela realmente existe. Contudo, quando analisado a existência de uma relação entre o peso e altura, sendo essas as duas variáveis utilizadas no cálculo do Índice de Massa Corporal, é factível uma correlação entre essas duas.

Como visto na análise entre o medalhista e o tipo de medalha conquistada, é possível concluir que existe algum tipo de relação entre essas duas variáveis, havendo uma predominância de certos nomes em determinadas medalhas.

A partir das análises feitas neste relatório, pode-se concluir que existem várias variáveis e fatores externos que possam influenciar o desempenho desses atletas em seus respectivos esportes, aonde diversas dessas variáveis podem acabar se relacionando entre si, afetando ainda mais a conquista de uma ou mais medalhas.