Projeto House of excellence

Consultores Responsáveis:
Maria Beatriz Cunha Barros

Requerente:
João Vitor Neves

Brasília, 12 de novembro 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	Histograma	5
	2.5	Tipos de Variáveis	5
		2.5.1 Qualitativas	5
		2.5.2 Quantitativas	6
	2.6	Teste de Normalidade de Shapiro-Wilk	6
	2.7	Teste de Normalidade de Anderson-Darling	7
	2.8	Coeficiente de Correlação de Kendall	7
	2.9	Teste de Correlação de Postos de Kendall	7
3	Análises		10
	3.1	Top 5 países com maior número de mulheres medalistas	10
	3.2	Valor IMC por esporte	11
	3.3	Top 3 medalhistas gerais por quantidade de cada tipo de medalha	13
	3.4	Variação peso por altura	14
	3.5	Conclusões	15

1 Introdução

Este relatório tem como objetivo principal a realização de análises estatísticas, a fim de ponderar e avaliar o desempenho dos atletas de elite da companhia House of Excellence ao longo de 5 olimpíadas. Através da utilização de alguns critérios estatísticos,as avaliações tendem esclarecer 4 tópicos pertinentes à empresa, os quais são descritas como a compreensão dos países com o maior número de mulheres medalistas, a verificação e a possível correlação entre os IMCs de determinados esportes, a relação entre medalistas gerais e as medalhas conquistadas, além de entender a relação entre o peso e a altura dos atletas da equipe.

O banco de dados foi coletado e disponibilizado pela própria empresa. Trata-se de uma amostra composta por 5 variáveis qualitativas nominais, sendo essas o nome, o gênero, a idade, o time, o esporte e a modalidade. Além disso, também dispõe de 3 variáveis quantitativas contínuas, sendo essas a idade, a altura e o peso. Também conta com uma variável do tipo qualitativa ordinal descrita como o tipo da medalha conquistada pelo atleta.

Por fim, o software utilizado para o desenvolvimento da pesquisa foi a versão 4.4.1 do programa R, o qual se trata de uma linguagem de programação estatística e gráfica, além de ser uma ferramenta gratuita e específica para análise de dados.

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 Histograma

O histograma é uma representação gráfica utilizada para a visualização da distribuição dos dados e pode ser construído por valores absolutos, frequência relativa ou densidade. A figura abaixo ilustra um exemplo de histograma.

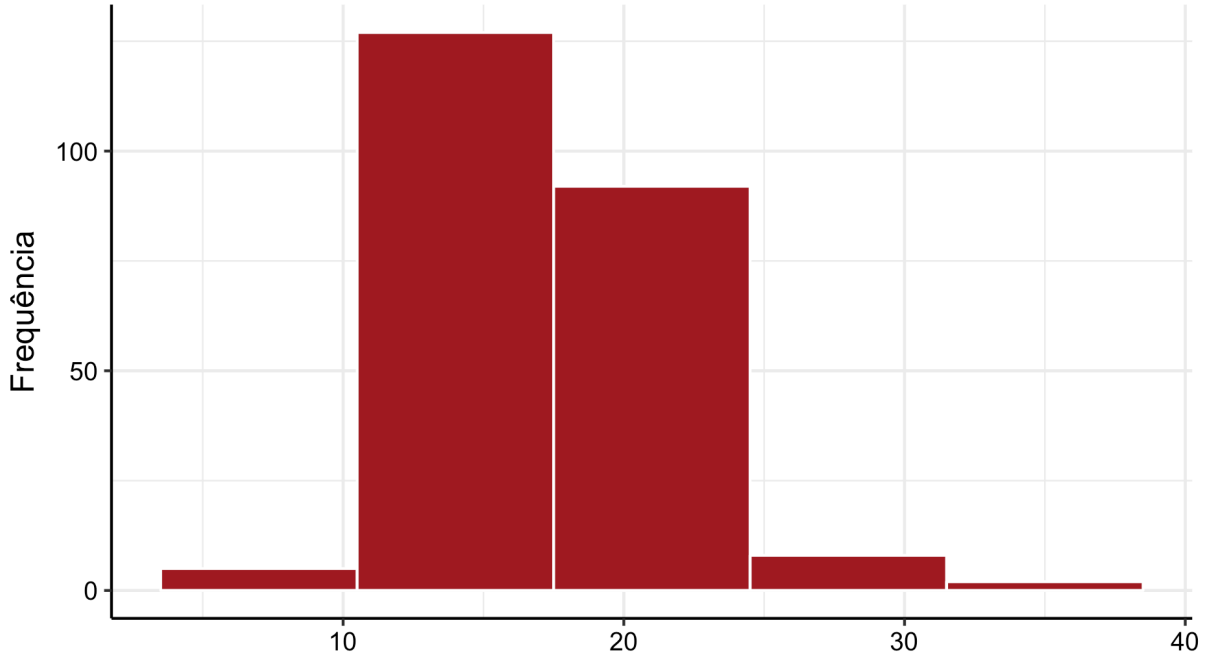


Figura 1: Exemplo de histograma

2.5 Tipos de Variáveis

2.5.1 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)

2.5.2 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)

2.6 Teste de Normalidade de Shapiro-Wilk

O **Teste de Shapiro-Wilk** é utilizado para verificar a aderência de uma variável quantitativa ao modelo da Distribuição Normal, sendo mais recomendado para amostras pequenas. A suposição de normalidade é importante para a determinação do teste a ser utilizado. As hipóteses a serem testadas são:

$$\begin{cases} H_0 : \text{A variável segue uma distribuição Normal} \\ H_1 : \text{A variável segue outro modelo} \end{cases}$$

A amostra deve ser ordenada de forma crescente para que seja possível obter as estatísticas de ordem. A estatística do teste é dada por:

$$W = \frac{1}{D}\left[\sum_{i=1}^{k} a_i \left(X_{(n-i+1)} - X_{(i)}\right)\right]$$

Com:

- K aproximadamente $\frac{n}{2}$
- $X_{(i)} =$ estatística de ordem *i*
- $D = \sum_{i=1}^{n}(X_i - \bar{X})^2$, em que $\bar{X}$ é a média amostral
- $a_i = $ constantes que apresentam valores tabelados

2.7 Teste de Normalidade de Anderson-Darling

O **teste de Normalidade de Anderson-Darling** é utilizado para verificar se uma amostra aleatória $X_1, X_2, ..., X_n$ de uma variável quantitativa segue uma distribuição Normal de probabilidade ou não. O teste possui as seguintes hipóteses:

$$\begin{cases} H_0 : \text{A variável segue uma distribuição Normal} \\ H_1 : \text{A variável segue outro modelo} \end{cases}$$

Se a hipótese nula for verdadeira, espera-se que o p-valor esteja acima do nível de significância α.

2.8 Coeficiente de Correlação de Kendall

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

O coeficiente de correlação de Kendall é normalmente representado pela letra τ, e sua fórmula de cálculo é:

$$\tau = \frac{C - D}{\frac{n(n-1)}{2}}$$

Onde:

- $C =$ número de pares concordantes
- $D =$ número de pares discordantes
- $n =$ tamanho da amostra

Os pares (x_i, y_i) e (x_j, y_j) são considerados concordantes se ambas as partes concordam, ou seja, se $x_i > x_j$ e $y_i > y_j$ ou se $x_i < x_j$ e $y_i < y_j$.

Já os pares (x_i, y_i) e (x_j, y_j) são discordantes se as partes discordam, ou seja, se $x_i > x_j$ e $y_i < y_j$ ou se $x_i < x_j$ e $y_i > y_j$.

2.9 Teste de Correlação de Postos de Kendall

Esse teste tem como objetivo verificar, por meio da comparação de postos, se existe independência entre as variáveis, avaliando a concordância e discordância dos pares.

As variáveis em estudo podem ser qualitativas ordinais ou quantitativas. Assim, o total de pares é $\binom{n}{2}$, em que n é o tamanho da amostra e $\binom{n}{2}$ representa a combinação das n observações da amostra tomadas de duas a duas. Considere, então, que N_c representa o número de pares concordantes e N_d é o número de pares discordantes. Os pares são concordantes se ambos os valores de X e Y de uma observação (um par) são maiores que os valores de X e Y de outra observação; os pares são discordantes se os valores das variáveis de uma observação diferem os valores de outra observação em direções opostas (por exemplo, $X_1 > X_2$ e $Y_1 < Y_2$).

As hipóteses para esse teste podem ser escritas como:

$$\begin{cases} H_0 : X \text{ e } Y \text{ são independentes (não há correlação entre elas)} \\ H_1 : \text{Há correlação de Kendall entre } X \text{ e } Y \end{cases}$$

A estatística do teste pode ter duas formas que variam conforme a presença de empates entre os pares:

a) Sem empates: $\tau = N_c - N_d$

Considerando H_0 verdadeira, essa estatística tem:

i) Distribuição exata apresentada em um tabela se o tamanho da amostra n for menor que 60.

ii) Aproximada pela Normal Padrão em caso de n grande:

$$w_p = z_p \frac{\sqrt{n(n-1)(2n+5)}}{18}$$

b) Com empates: $\tau = \dfrac{N_c - N_d}{N_c + N_d}$

Considerando H_0 verdadeira, essa estatística tem:

i) Distribuição exata apresentada em um tabela se o tamanho da amostra n for menor que 60.

ii) Aproximada pela Normal Padrão em caso de n grande:

$$w_p = z_p \frac{\sqrt{n(n-1)(2n+5)}}{18}$$

Para realizar a comparação dos pares e concluir se serão concordantes ou discordantes, pode-se utilizar as seguintes regras de decisão:

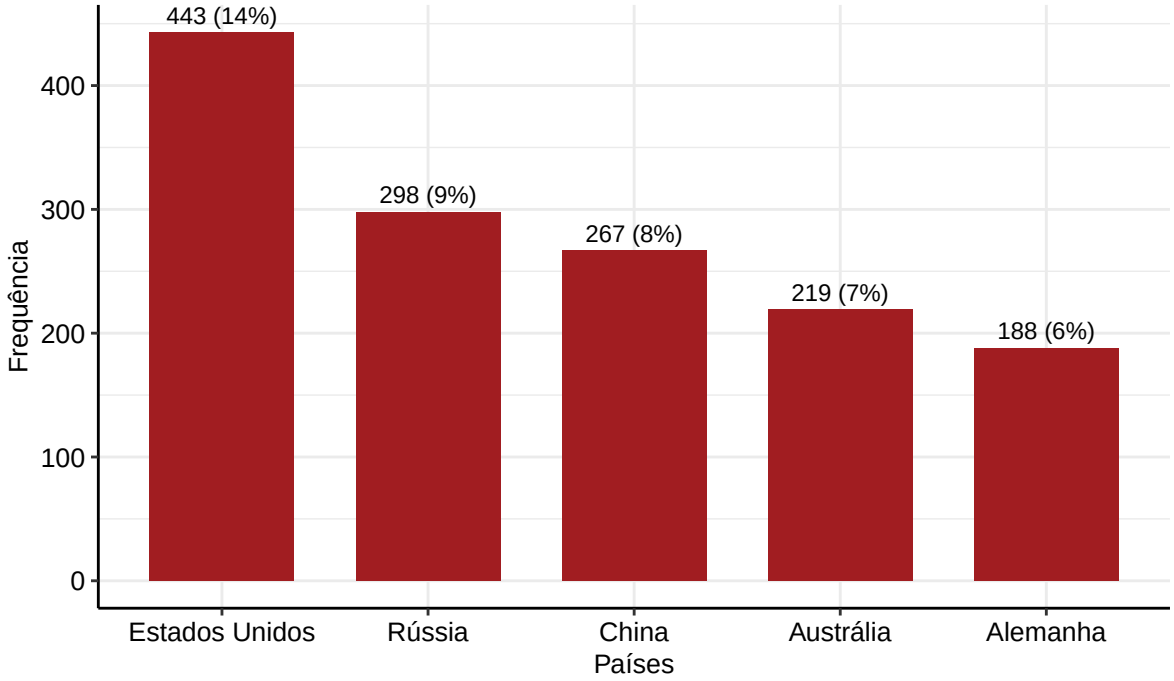
- Se $\frac{Y_j - Y_i}{X_j - X_i} > 0$, os pares são **concordantes** (adicione 1 a N_c)
- Se $\frac{Y_j - Y_i}{X_j - X_i} < 0$, os pares são **discordantes** (adicione 1 a N_d)
- Se $\frac{Y_j - Y_i}{X_j - X_i} = 0$, ocorreu **empate** (adicione 0,5 a N_c e a N_d)
- Se $X_j = X_i$, não há comparação

3 Análises

3.1 Top 5 países com maior número de mulheres medalistas

Esta análise tem o objetivo de verificar os países que possuem o maior número de mulheres medalistas e, dessa forma, montar um ranking com os 5 países mais pontuados. Para tal, foram utilizadas as variáveis Gênero e Time, as quais se tratam de variáveis qualitativas nominais, e a variável Medalha, se tratando de uma variável qualitativa ordinal. Será utilizado um gráfico de colunas com o intuito de ilustrar as análises.

Figura 1: Gráfico de Colunas do Top 5 países com maior número de mulheres medalistas



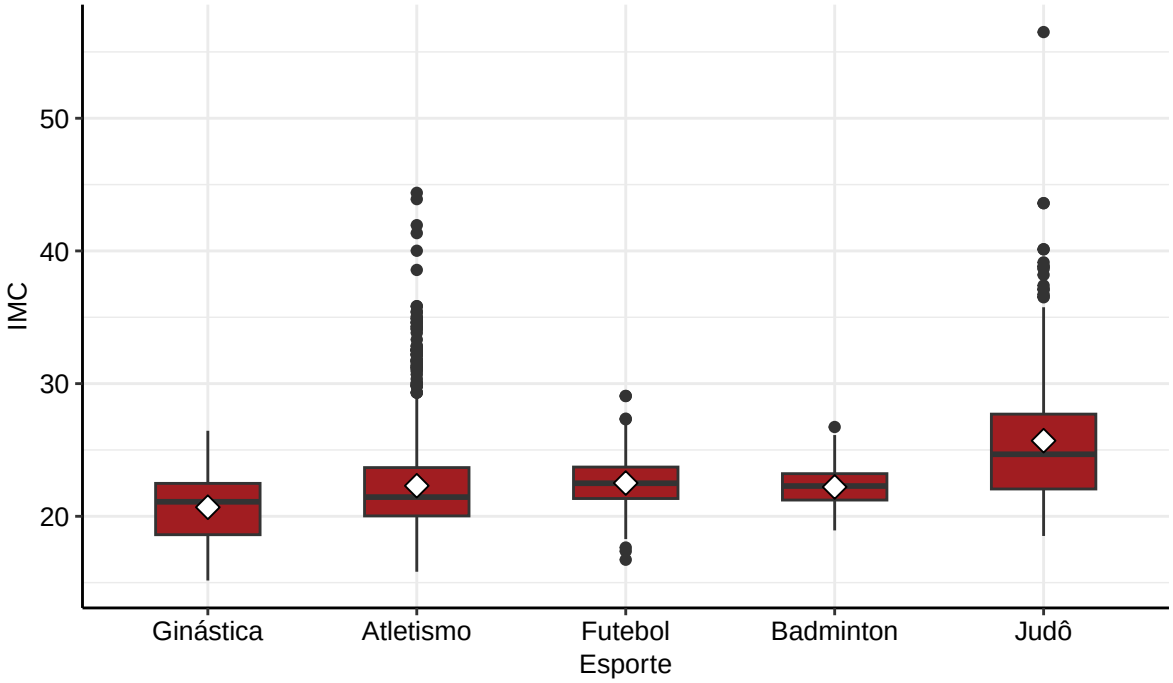
Ao selecionarmos apenas as ganhadoras mulheres dos dados disponibilizados, podemos perceber que representam 3245 ganhadores, ou seja, cerca de 44,48% do total de medalistas.

Outrossim, percebe-se por meio da Figura 1 que o país com maior número de ganhadoras mulheres são os Estados Unidos, representando cerca de 13,7% dentro desse secção denominada Top 5, seguido pela Rússia (9%), pela China (8,22%), pela Austrália (6,74%) e pela Alemanha (5,79%). A figura 1 também permite observar que continentes como a América, Europa, Ásia e Oceania, possuem uma forte representação feminina nas olimpiadas, com foco na Europa, a qual desponta duas vezes no top 5.

3.2 Valor IMC por esporte

Essa análise tem como objetivo verificar as tendências dos indices de massa corporal dos atletas e compara-los por meio do esporte que praticam, além de apurar se existem ligações entre esses dois fatores. Será utliziado a variável categórica qualitativa nominais Time, além da variável categórica qualitativa contínua denominada IMC.

Figura 2: Boxplot do IMC pelo tipo de esporte



Quadro 1: Medidas resumo do IMC

Estatística	Atletismo	Badminton	Futebol	Ginástica	Judô
Média	22,30	22,21	22,51	20,68	25,70
Desvio Padrão	3,86	1,50	1,73	2,38	5,12
Variância	14,92	2,26	2,99	5,67	26,23
Mínimo	15,82	18,94	16,73	15,16	18,52
1º Quartil	20,03	21,22	21,34	18,61	22,06
Mediana	21,45	22,28	22,49	21,09	24,68
3º Quartil	23,67	23,21	23,71	22,48	27,70
Máximo	44,38	26,73	29,07	26,45	56,50

Ao analisarmos as médias de cada esporte, podemos observar que, em comparação aos outros esportes, os atletas da ginástica possuem um valor de IMC médio mais baixo (20,68), enquanto o judô apresenta IMC médio mais alto (25,7). Segundo a figura 2, é visivel um grande número de outliners, ou seja, valores discrepantes, que estão for do intervalo esperado.

$$\begin{cases} H_0 : \text{A variável segue uma distribuição normal} \\ H_1 : \text{A variável não segue uma distribuição normal} \end{cases}$$

Quadro 2: P-valor do teste de distribuição Shapiro-Wilk para a variável IMC

Grupo	P-valor	Decisão do teste
IMC	<0,001	Rejeita H_0

A fim de verificarmos se há diferença entre os IMC entre diferentes tipos de esporte, foi realizado primeiro o teste de normalidade Shapiro-Wilk, o qual indicou que a váriavel analisada não segue uma distribuição Normal de probabilidade(P<0,001), ou seja, é necessário a realização de um teste não paramétrico.

$$\begin{cases} H_0 : \text{As médias do IMC entre os esportes são iguais} \\ H_1 : \text{A média de pelo menos um esporte é diferente} \end{cases}$$

Quadro 3: P-valor do teste não paramétrico Kruskal-Wallis para os diferentes esportes

Grupo	P-valor	Decisão do teste
Esportes IMC	<0,001	Rejeita H_0

Posteriormente foi realizado o teste Kruskal-Wallis(P<0,001), o qual apontou que existe ao menos uma diferença significativa entre os grupos de esportes quando comparados entre si.

$$\begin{cases} H_0 : \text{Não há diferença significativa entre os pares de grupos comparados} \\ H_1 : \text{Há diferença significativa entre os pares de grupos comparados} \end{cases}$$

Quadro 4: P-valor para o teste post-hoc Dunn para comparação de IMC entre diferentes esportes

Grupo	P-valor	Decisão do teste
Esportes IMC	0,519	Rejeita H_0

Ademais, vale ressaltar que, ao comparar os esportes a partir do teste de Dunn, o judô se destacou com grandes diferenças entre sua média de IMC em comparação com outros esportes, enquanto entre atletismo e badminton(p = 0,519) não foram encontradas diferenças significativas entre as médias.

Tabela 1: Esporte por categoria de IMC

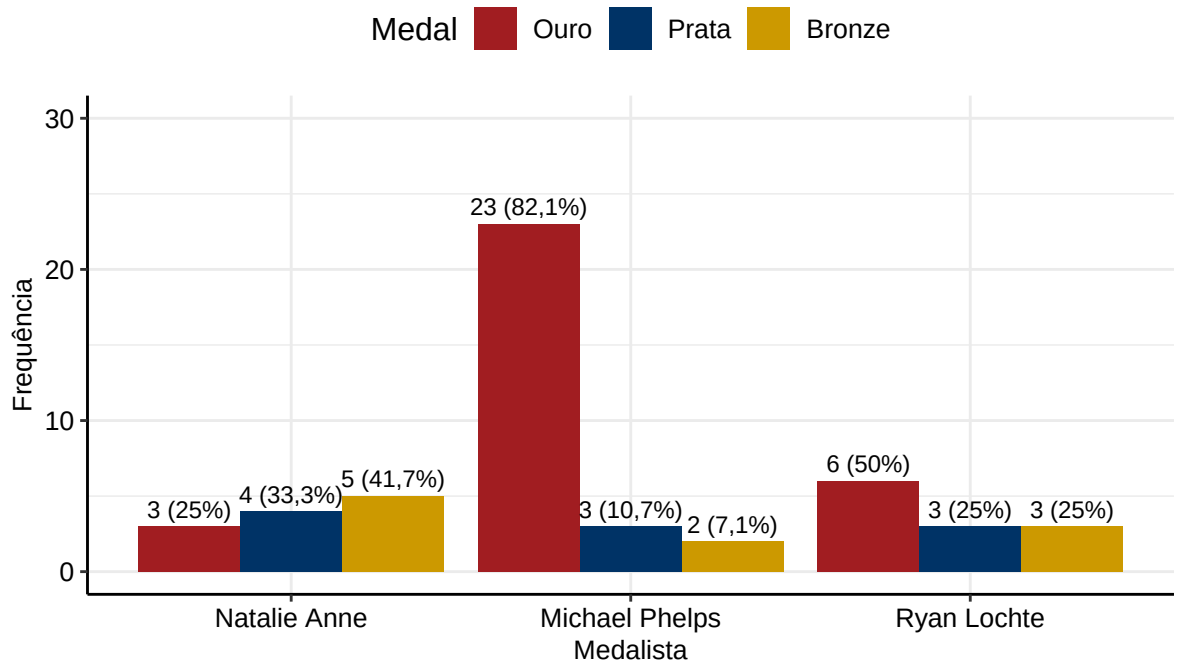
	Atletismo	Badminton	Futebol	Ginástica	Judô
Abaixo	96	0	6	77	0
Normal	690	116	468	257	161
Obesidade	52	0	0	0	45
Sobrepeso	95	3	31	7	74

Percebe-se por meio da tabela a frequência de cada esporte dentro das categorias de IMC, as quais são classificadas como abaixo do ideal (IMC < 18.5), normal (18.5 < IMC < 25), sobrepeso (25 < IMC < 30) e obesidade (IMC > 30). Foi possível afirmar que o grupo de atletismo apresentam maior número de atletas com IMC abaixo do ideal (53,63%), enquanto os atletas do judo e badminton ficaram de fora dessa categoria.

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

Esta análise tem o objetivo de apontar os 3 maiores medalistas gerais, além de observar a quantidade de cada tipo de medalha que cada um destes atletas conquistou. Para isso, foram utilizadas as variáveis Nome, que se trata de uma variável qualitativa nominal, e Medalha, que se categoriza como uma variável qualitativa ordinal.Outrossim, será utilizado um gráfico de colunas e um quadro indicativo do p-valor do teste que será utilizado, a fim de ilustrar as análises.

Figura 3: Gráfico de Colunas dos três medalistas mais pontuados e suas respectivas quantidades de medalhas.



Segundo a figura 3, foi observado uma diferença de 16 medalhas entre o primeiro colocado, o nadador Michael Phelps, e o segundo e terceiro colocados, os quais empataram em quantidade de medalhas. O valor de medalhas de ouro de Phelps chega a ser discrepante entre os outros competidores, o que indica uma maior vantagem do nadador sobre os outros atletas.

$$\begin{cases} H_0 : \text{As variáveis medalhas e medalistas são independentes} \\ H_1 : \text{As variáveis medalhas e medalistas não são independentes} \end{cases}$$

Quadro 5: P-valor do teste de independência (Teste Qui-Quadrado) entre as variáveis medalista e medalha

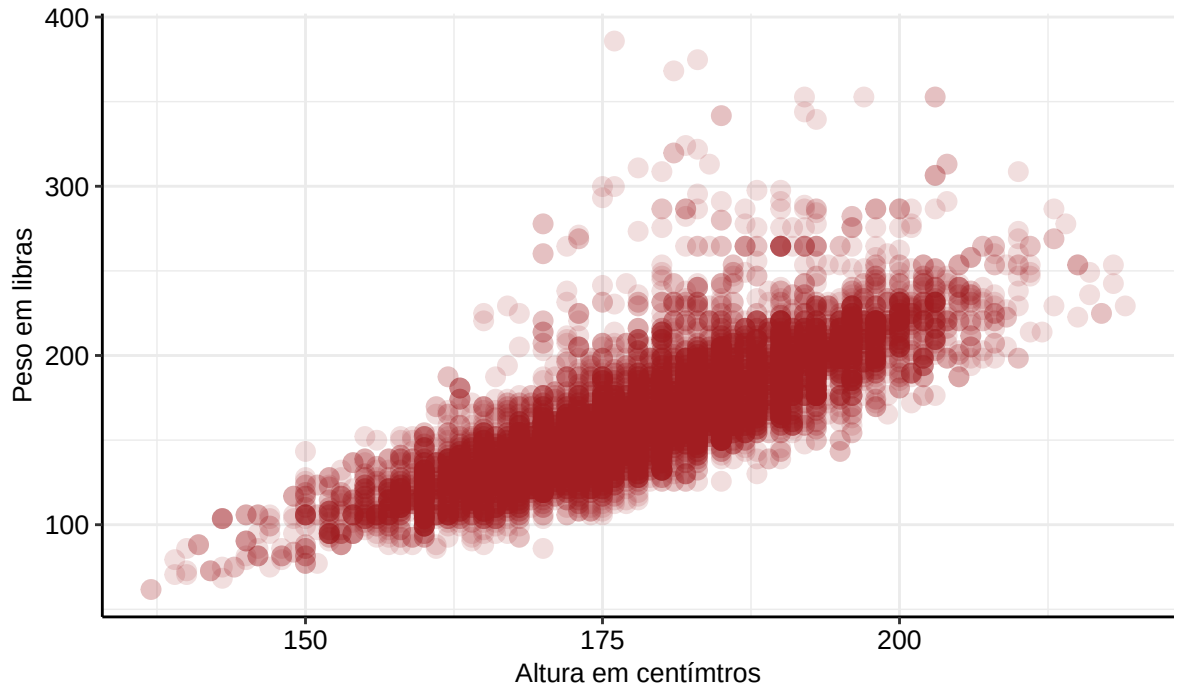
Grupo	P-valor	Decisão do teste
Nome Medalha	0.012	Rejeita H_0

Analisando o Quadro 5, que representa o teste de associação entre as variáveis medalista e medalha, é visível que há uma associação estatisticamente significativa entre as variáveis analisadas (P = 0,012), ou seja, é possível afirmar que a frequência de medalha parece estar associada ao medalista.

3.4 Variação peso por altura

Esta análise tem o objetivo de compreender se há uma relação entre o peso e altura dos atletas. Para tal, foram utilizadas as variáveis Peso e Altura, as quais se tratam de variáveis quantitativas contínuas, e será utilizado um quadro de medidas resumo,um quadro indicativo do p-valor do teste que será utilizado e um gráfico de dispersão a fim de ilustrar as análises.

Fugura 4: Gráfico de dispersão da relação entre Altura e Peso dos atletas



A figura 4 sugere que, à medida que a altura aumenta, o peso também tende a aumentar, indicando uma correlação positiva entre essas duas variáveis. Além disso, é possível afirmar que há uma certa dispersão dos pontos, ou seja, indivíduos de mesma altura podem ter pesos distintos.

$$\begin{cases} H_0 : \text{Não existe uma associação significativa entre o Peso e a Altura} \\ H_1 : \text{Existe uma associação significativa entre o Peso e a Altura} \end{cases}$$

Quadro 6: P-valor do teste de coorelação entre a Altura e o Peso

Variáveis	**P-valor**	**Decisão do teste**
Peso Altura	<0,001	Rejeita H_0

Analisando o Quadro 6, que representa o teste de correlação entre as variáveis peso e altura, é visível que existe dependência entre as variáveis(P<0,001), ou seja, há uma forte evidência que o peso e a altura do atleta estão associados de forma diretamente proporcional, quando um aumenta, o outro também tende a aumentar.

3.5 Conclusões

Em princípio, a análise acerca das mulheres medalistas mostrou que elas representam cerca de 44,48% do total de medalistas, com os Estados Unidos se

destacando com o maior numero de ganhadoras, seguidos pela Rússia, China, Austrália e Alemanha. Essesn países dominam o ranking das medalhas femininas, sendo os Estados Unidos responsáveis por quase 14% das medalhas no Top 5.

Já em relação às características físicas, observamos que os atletas dde ginástica apresentam IMC médio mais baixo, enquanto os judocas têm o IMC mais alto. A análise do IMC também revelou a diferença entre os esportes, com o judô destacando-se devido a um número elevado de outliers e uma diferença substancial de IMC, comparado aos outros esportes. Além disso, análise das frequências dos IMCs por esporte revelou que o atletismo possui a maior proporção de atletas com IMC abaixo do idea, enquanto judô e badminton não apresentaram atletas nessa categoria. Esses achados evidenciam a variabilidade nas características físicas dos atletas em diferentes esportes, o que pode influenciar diretamente no seu desempenho.

Por fim, ao verificar o quadro de medalhas de atletas individuais, como o nadador e maior medalista Michael Phelps, mostrou uma discrepância significativa em seu número de medalhas de ouro, indicando uma vantagem sobre os demais medalistas, sugerindo que a frequência de medalhas está de fato relacionado ao perfil do medalista. Outrossim, quando se trata da relação entre o peso e a altura dos competidores premiados, há uma forte dependência entre as duas variáveis, com o aumento de um fator tendendo a se associar ao aumento do outro.

Em resumo, esses resultados conectam as características físicas dos atletas com seu desempenho nas Olimpíadas, proporcionando uma compreensão mais aperfeiçoada acerca de características como o valor do IMC, do peso e da altura, e suas devidas relações com o número de medalhas e os esportes praticados. Ou seja, o estudo revela padrões interessantes sobre como os aspectos físicos de cada competidor podem influenciar o desempenho pessoal nos jogos olímpicos.