Capítulo 2

Estatística Descritiva

A estatística descritiva é a parte da Estatística que abrange métodos destinados a resumir a informação contida nos dados, destacando os aspectos mais marcantes. Nessa etapa da análise algumas medidas são calculadas para descrever as informações disponíveis, possibilitando o estudo do comportamento de uma variável. O interesse principal é caracterizar o conjunto de dados a partir de medidas que resumam a informação, por exemplo, representando a tendência central dos dados ou a forma pela qual estes dados estão dispersos.

2.1 Medidas resumo

Para descrever uma variável quantitativa de um conjunto de dados pode-se calcular as medidas resumo dessa variável. Essas medidas buscam sumarizar as informações disponíveis sobre o comportamento de uma variável quantitativa, em que, o interesse é caracterizar o conjunto de dados através de medidas que resumam a informação nele contida. As medidas resumo mais conhecidas são:

Medidas de posição ou de tendência central: são medidas ao redor das quais as observações das variáveis quantitativas tendem a se agrupar (exemplo: *média e mediana*).

Medidas de dispersão: são medidas que quantificam a dispersão dos dados, ou seja, quantifica o quanto as observações quantitativas estão distantes entre si (exemplo: *variância e desvio padrão*).

2.1.1 Medidas de Tendência Central

Há varias formas de medir o centro de uma distribuição. As duas formas mais utilizadas, e que serão abordadas com detalhe a seguir, são a média e a mediana.

Média: A chamada média aritmética, é a mais comum das medidas de tendência central. Considerando uma variável Y com observações representadas por $y_1, y_2, \ldots, y_n$; a média desse conjunto é a soma dos valores dividida pelo número total de observações (n), isto é,

$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{2.1}$$

em que, $\sum$ (sigma maiúsculo) é a notação matemática usual para somatório.

Mediana: A mediana é o 50^0 percentil, isto é, o valor abaixo do qual recaem 50% dos valores da amostra, ou seja, é o valor que ocupa a posição central dos dados ordenados. A mediana será denotada por md.

Para calcular essas medidas será considerado o conjunto de dados "Dieta.xlsx" (disponível na plataforma Moodle). Esse conjunto de dados é composto por 4 variáveis, são elas:

- **Dieta**: representa o tipo de dieta que o indivíduo foi submetido; 50 indivíduos foram submetidos a um tipo de dieta, denominada de Dieta A, e 50 indivíduos foram submetidos a outro tipo de dieta, denominada de Dieta B.

- **Hipertensao**: representa se o indivíduo estava hipertenso após ser submetido pela dieta (Sim) ou se o indivíduo se curou da hipertensão após o término da dieta (Nao); todos os indivíduos observados apresentavam hipertensão no começo do estudo, ou seja, antes dele ser submetido a dieta.

- **Perda**: representa a proporção de perda de peso, em %, após 5 meses de dieta (fim do estudo). Essa medida é calculada utilizando a seguinte expressão:

$$\frac{Peso\ Inicial - Peso\ Final}{Peso\ Inicial}, \tag{2.2}$$

  em que, $Peso\ Inicial$: peso do indivíduo em Kg antes do início da dieta; $Peso\ Final$: peso do indivíduo em Kg após 5 meses de dieta (fim do estudo).

- **Idade**: representa a idade do indivíduo em anos completos.

Se o interesse for calcular a proporção média de perda de peso de todos os indivíduos utilizando o `Software R`, após a importação do conjunto de dados (ver Seção 1.3.3), basta utilizar o seguinte comando:

```
mean(dados$Perda)
```

Observar que no argumento da função `mean` (função que retorna o valor médio de um vetor de dados) foi utilizado o comando `dados$Perda`, esse comando retorna o vetor com todas as observações da matriz "dados" que correspondem a coluna "Perda". Ou seja, o argumento da função `mean` é composto pelo vetor com os valores que se tem interesse em calcular a média.

Se essa linha de comando for executada de forma correta, o `Software R` irá retornar o valor `15.7804`. Ou seja, a proporção média de perda de peso de todos os indivíduos foi de $15,78\%$.

Supor, agora, que o pesquisador esteja interessado em verificar qual dieta é mais eficiente para a perda de peso. Ou seja, para responder essa pergunta é necessário calcular a proporção média de perda de peso para cada uma das duas dietas. Para calcular esses valores utilizando o `Software R` basta utilizar o seguinte comando:

```
tapply(dados$Perda,dados$Dieta,mean)
```

O comando `tapply` está sendo utilizado aqui para calcular as médias para cada um dos dois tipos de dieta. Nesse comando o primeiro argumento (`dados$Perda`) é composto pelo vetor com os valores que se tem interesse em calcular a medida descritiva (nesse caso a média); o segundo argumento (`dados$Dieta`) é composto pelo vetor que representa os grupos em que se deseja calcular a medida descritiva (nesse caso são dois grupos: Dieta A e Dieta B); o terceiro argumento é composto pela função padrão do `Software R` utilizada para calcular a medida

descritiva de interesse (nesse caso `mean`, pois o interesse é calcular a **média** para cada um dos grupos de dieta).

Se essa linha de comando for executada de forma correta, o `Software R` irá retornar as seguintes informações:

A	B
11.0398	20.5210

Ou seja, a proporção média de perda de peso dos indivíduos submetidos a Dieta A foi de $11,04\%$, enquanto que a proporção média de perda de peso dos indivíduos submetidos a Dieta B foi de $20,52\%$. Conclui-se, então, que os indivíduos submetidos a Dieta B tendem a perder mais peso se comparado com os indivíduos submetidos a Dieta A. Portanto, aparentemente, a Dieta B é mais eficiente na perda de peso dos indivíduos se comparada com a Dieta A.

De modo análogo é possível calcular, utilizando o `Software R`, a proporção **mediana** de perda de peso de todos os indivíduos, para isso basta utilizar o seguinte comando:

```
median(dados$Perda)
```

Observar que o argumento da função `median` (função que retorna a mediana de um vetor de dados) é o mesmo utilizado na função `mean`, ou seja, aqui também é utilizado como argumento da função o comando `dados$Perda`, comando que retorna o vetor com todas as observações da matriz "dados" que correspondem a coluna "Perda". Ou seja, o argumento da função `median` é composto pelo vetor com os valores que se tem interesse em calcular a mediana.

Se essa linha de comando for executada de forma correta, o `Software R` irá retornar o valor `15.39`. Ou seja, a proporção mediana de perda de peso de todos os indivíduos foi de $15,39\%$. Pode-se dizer, ainda, que 50% dos indivíduos observados apresentam uma proporção de perda de peso menor ou igual a $15,39\%$; assim como, 50% dos indivíduos observados apresentam uma proporção de perda de peso maior ou igual a $15,39\%$.

Supor, novamente, o interesse em verificar qual dieta é mais eficiente para a perda de peso. Ou seja, para responder essa pergunta pode-se, também, calcular a proporção mediana de perda de peso para cada uma das duas dietas. Para calcular esses valores utilizando o `Software R` basta utilizar o comando `tapply` da seguinte maneira:

```
tapply(dados$Perda,dados$Dieta,median)
```

Veja que o comando `tapply` foi utilizado de forma similar ao cálculo da média separada pelo grupo de dieta, a única diferença está no último argumento, que agora é `median`, pois o interesse agora é retornar a proporção mediana de perda de peso para cada uma das duas dietas. Se essa linha de comando for executada de forma correta, o `Software R` irá retornar as seguintes informações:

A	B
10.445	20.615

Ou seja, a proporção mediana de perda de peso dos indivíduos submetidos a Dieta A foi de $10,44\%$, enquanto que a proporção mediana de perda de peso dos indivíduos submetidos a Dieta B foi de $20,61\%$. Conclui-se, então, que os indivíduos submetidos a Dieta B tendem a perder mais peso se comparado com os indivíduos submetidos a Dieta A. Portanto, aparentemente, a Dieta B é mais eficiente na perda de peso dos indivíduos se comparada com a Dieta A. Conclusão idêntica a observada quando calculado as proporções médias de perda de peso dos indivíduos para os dois grupos de tipo de dieta.

As medidas de tendência central (média e mediana) podem ser utilizadas em conjunto para auxiliar a análise dos dados onde, em determinadas situações, uma dessas medidas pode ser mais conveniente do que a outra. Por exemplo, se um ou mais valores são muito discrepantes do que o geral das observações, a média será muito influenciada por este valor, tornando-a, assim, inadequada para representar aquele conjunto de dados. Neste caso, como a mediana não é afetada por valores discrepantes, seu uso seria mais adequado para representar o centro dos dados. Como regra geral, é preciso usar essas medidas com o cuidado de não distorcer as informações e características dos dados que se está analisando.

Como exemplo, considerar o número de falhas de sistema por ano em 8 fábricas de disjuntores dadas por:

$$16 \quad 18 \quad 15 \quad 22 \quad 24 \quad 23 \quad 15 \quad 62$$

Nesse caso o número médio de falhas no sistema por ano é de $24,38\ falhas$, enquanto que o número mediano de falhas no sistema por ano é de $20\ falhas$. Veja que, nesse caso, a medida que melhor descreve a variável número de falhas de sistema por ano é a **mediana**, pois as observações tendem a estar mais próximas do valor 20 do que do valor $24,38$, sendo 24 o segundo maior valor observado, ou seja, o valor da média não está descrevendo de forma coerente o comportamento da variável número de falhas no sistema por ano. Isso ocorre pois o valor 62 é muito discrepante se comparado com os demais valores, como a média é influenciada por valores discrepantes (valores muito baixos ou muito altos se comparado com os demais) a mesma não é recomendada para descrever o comportamento da tendência central de uma variável quantitativa na presença desses valores discrepantes.

2.1.2 Medidas de Dispersão

Apesar das medidas de tendência central fornecerem uma ideia do comportamento das variáveis, elas podem esconder valiosas informações. Essas medidas podem não ser suficientes para descrever e discriminar o comportamento por completo das variáveis quantitativas de um conjuntos de dados. Enquanto a média possa ser uma das medidas estatísticas mais importante, é também igualmente importante conhecer a dispersão das observações. Tal como no caso das medidas de tendência central, aqui também será apresentado algumas medidas de dispersão.

Amplitude: A amplitude é simplesmente a distância entre o maior e o menor valor do conjunto de dados, e será denotada por Δ, logo,

$$\Delta = \max(x) - \min(x), \tag{2.3}$$

em que, $\max(x)$ é o maior valor observado do conjunto de dados e $\min(x)$ o menor valor observado. A amplitude pode ser criticada, como medida de dispersão, pelo fato de não dizer nada sobre a distribuição do conjunto de dados, a não ser onde ela começa e termina. A utilização de apenas essas duas observações extremas não é muito digna de confiança. No entanto ela retorna uma informação importante, a amplitude nada mais é do que a variabilidade máxima observada para uma variável quantitativa.

Desvio Médio Absoluto (DMA): O Desvio Médio (DM), como seu nome indica, é obtido pelo cálculo do desvio de cada observação a contar da média, ou seja, calcula-se as distâncias de cada observação em relação a média $(x_i - \bar{x})$, $i = 1, \ldots, n$; para obter uma medida que mensure a variabilidade basta calcular a média desses desvios, ou seja, soma-se esses desvios e divide-se o resultado pelo número total de observações (n). Enquanto o DM

possa parecer uma medida útil, ele não tem utilidade, pois na soma, os desvios positivos sempre se cancelam com os negativos, retornando média zero. Este problema pode ser contornado se for ignorado todos os sinais negativos e for tomado a média dos valores absolutos dos desvios, logo, obtêm-se o DMA, dado por,

$$DMA = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| \,. \tag{2.4}$$

Observe que o DMA nada mais é do que o calculo da distância média das observações em relação a média. Ou seja, quanto maior for essa medida, se comparada com a média observada, maior será a variabilidade das observações; consequentemente, quanto menor for essa medida, se comparada com a média observada, menor será a variabilidade das observações.

Desvio Quadrático Médio (DQM): Apesar do DMA ser intuitivamente uma boa medida de dispersão, ele não oferece facilidade de manejo matemático; uma dificuldade é o problema relacionado a determinação da função valor absoluto. Uma alternativa a função valor absoluto é elevar ao quadrado cada desvio, isto é,

$$DQM = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \,. \tag{2.5}$$

Variância e Desvio-padrão Populacionais: O DQM é uma boa medida de variabilidade, desde que, seja possível selecionar todos os indivíduos de uma população. Ou seja, o DQM pode ser considerado como sendo a Variância Populacional, denotada usualmente por σ^2. O problema em se utilizar a Variância Populacional ($\sigma^2 = DQM$) é que, ao elevar os desvios $(x_i - \bar{x})$ ao quadrado, a unidade da medida de variabilidade fica elevada ao quadrado. Supor que um pesquisador tenha calculado a Variância Populacional da idade dos indivíduos de uma dada população, e esse valor foi de $\sigma^2 = 33\ anos^2$; veja que essa medida não tem interpretação, pois $anos^2$ não é interpretável. Para resolver esse problema basta aplicar a raiz quadrada no valor obtido, ou seja, $\sigma = \sqrt{\sigma^2} = 5,74\ anos$. Essa medida ($\sigma$) é conhecida como Desvio-padrão Populacional e tem interpretação, pois está na mesma unidade dos dados observados (no caso $anos$).

Variância e Desvio-padrão: Como na prática dificilmente um pesquisador tem acesso a todos os indivíduos de uma população, se o interesse é avançar um pouco e for utilizado a amostra para fazer inferências sobre a população, é melhor utilizar o divisor $n-1$ em (2.5), em lugar de n. Tecnicamente, isto faz com que a variância da amostra seja um estimador não-viciado da variância da população (esse assunto será abordado com mais detalhes ao longo do texto). Portanto, a variância, referente a variável X de um conjunto de dados, é definida por,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 \,, \tag{2.6}$$

Para manter a mesma unidade dos dados originais, é conveniente definir o desvio-padrão, definido por,

$$s = \sqrt{s^2}. \tag{2.7}$$

Na prática, para calcular a distância média das observações em relação a média, com mais eficiência, utiliza-se o desvio padrão (2.7). Portanto, quanto maior for essa medida,

se comparada com a média observada, maior será a variabilidade das observações; consequentemente, quanto menor for essa medida, se comparada com a média observada, menor será a variabilidade das observações. Veja que a interpretação do desvio-padrão e do DMA é a mesma, no entanto o desvio-padrão é matematicamente uma medida mais eficiente. Por esse motivo que, na prática, opta-se por utilizar o desvio-padrão como alternativa ao DMA.

Intervalo Interquartil (IIQ): O intervalo interquartil é calculado pela diferença entre o terceiro e primeiro quartis. Os quartis são medidas que permitem dividir a distribuição dos dados em quatro partes iguais quanto ao número de elementos de cada uma, ou seja:

(a) **Primeiro Quartil (Q_1)**: Dado um conjunto ordenado de valores, o primeiro quartil, Q_1, é o valor que divide o conjunto em duas partes tal que um quarto (ou 25%) dos valores seja menor do que Q_1 e três quartos (ou 75%) sejam maiores que Q_1.

(b) **Segundo Quartil (Q_2)**: É igual à mediana, ou seja, o valor abaixo do qual recaem 50% dos valores da amostra, é o valor que ocupa a posição central dos dados ordenados.

(c) **Terceiro Quartil (Q_3)**: Dado um conjunto ordenado de valores, o terceiro quartil, Q_3, é o valor que divide o conjunto em duas partes, tal que, três quartos (ou 75%) dos valores sejam menores do que Q_3 e um quarto (ou 25%) dos valores sejam maiores do que Q_3.

Portanto, o Intervalo Interquartil é definido por,

$$IIQ = Q_3 - Q_1. \tag{2.8}$$

O valor do IIQ não é influenciado por valores discrepantes, diferentemente do desvio-padrão, que é influenciado. No entanto, assim como a amplitude, a utilização de apenas essas duas observações não é muito digna de confiança. Por esse motivo, deve-se tomar muito cuidado ao utilizar o IIQ como medida de variabilidade. No entanto, assim como o desvio-padrão, quanto maior for essa medida, se comparada com a média observada, maior será a variabilidade das observações; consequentemente, quanto menor for essa medida, se comparada com a média observada, menor será a variabilidade das observações.

Agora, essas medidas serão calculadas considerando o conjunto de dados "Dieta.xlsx" (disponível na plataforma Moodle). Se o interesse for calcular a amplitude da proporção de perda de peso de todos os indivíduos utilizando o `Software R`, após a importação do conjunto de dados (ver Seção 1.3.3), basta utilizar o seguinte comando:

```
diff(range(dados$Perda))
```

Observar que aqui foram utilizadas duas funções, uma como argumento da outra, as funções `diff` e `range`. A função `range` retorna um vetor com duas observações, o valor mínimo e máximo de um vetor de dados. Ou seja, o argumento da função `range` é composto pelo vetor de observações que se tem interesse em obter o menor e maior valores observados em um vetor de dados, que nesse caso é o vetor obtido pelo comando `dados$Perda`. Aqui a função `diff` está sendo utilizada para calcular a diferença entre o segundo e o primeiro elemento do vetor obtido pelo comando `range(dados$Perda)`, ou seja, está retornando a diferença entre o maior e menor valor do vetor com as observações da perda de peso dos indivíduos compostos pelo conjunto de dados "Dieta.xlsx".

Se essa linha de comando for executada de forma correta, o `Software R` irá retornar o valor `31.32`. Ou seja, a amplitude da proporção de perda de peso de todos os indivíduos foi de $31,32\%$. Pode-se dizer, também, que a variabilidade máxima da proporção de perda de peso de todos os indivíduos foi de $31,32\%$

Supor, agora, que o pesquisador esteja interessado em comparar a variabilidade máxima da proporção de perda de peso entre os dois tipos de dietas. Para realizar essa comparação é necessário calcular a amplitude da proporção média de perda de peso para cada uma das duas dietas. Para realizar esses cálculos utilizando o `Software R` basta executar as seguintes linhas de comando:

```
maxmin <- tapply(dados$Perda,dados$Dieta,range)

diff(maxmin$A)

diff(maxmin$B)
```

O comando `tapply` está sendo utilizado aqui para retornar dois vetores (vetores $A e $B, ver a saída do software apresentada logo abaixo) com duas observações cada, esses vetores são compostos pelo valor mínimo e máximo da proporção de perda de peso para cada um dos dois tipos de dieta. Nesse comando o primeiro argumento (`dados$Perda`) é composto pelo vetor com os valores da proporção de perda de peso de todos os indivíduos observados; o segundo argumento (`dados$Dieta`) é composto pelo vetor que representa o tipo de dieta em que os indivíduos foram submetidos (Dieta A ou Dieta B); o terceiro argumento é composto pela função padrão do `Software R` utilizada para calcular a medida de interesse (nesse caso `range`). Veja que aqui foi criado o objeto `maxmin` a partir da linha de comando `tapply(dados$Perda,dados$Dieta,range)`, esse objeto é uma lista, que se for executado no software R retornará o seguinte resultado:

```
> maxmin
$A
[1]  0.45 26.28

$B
[1] 11.14 31.77
```

Com o objeto `maxmin` criado basta executar as linhas de comando `diff(maxmin$A)` para obter o valor `25.83` e `diff(maxmin$B)` para obter o valor `20.63`. Ou seja, a variabilidade máxima da proporção de perda de peso dos indivíduos submetidos a Dieta A foi de $25,83\%$, enquanto que a variabilidade máxima da proporção de perda de peso dos indivíduos submetidos a Dieta B foi de $20,63\%$. Conclui-se, então, que os indivíduos submetidos a Dieta B tendem a ter uma variabilidade máxima da proporção de perda de peso menor se comparado com os indivíduos submetidos a Dieta A.

Para calcular o desvio-padrão da proporção de perda de peso de todos os indivíduos utilizando o `Software R`, basta utilizar o seguinte comando:

```
sd(dados$Perda)
```

Veja que, no argumento da função `sd` (função que retorna o valor do desvio-padrão de um vetor de dados) foi utilizado o argumento `dados$Perda` (vetor com os valores que se tem interesse em calcular o desvio-padrão).

Se essa linha de comando for executada de forma correta, o `Software R` irá retornar o valor `6.78741`. Ou seja, a variabilidade da proporção de perda de peso de todos os indivíduos é de $6,79\%$. Para verificar se essa variabilidade é alta ou baixa, é necessário verificar se o valor

do desvio padrão é alto ou baixo se comparado com o valor da média, pois o desvio-padrão retorna o valor da distância média em relação a média observada. Como visto anteriormente a proporção média de perda de peso de todos os indivíduos foi de $15,78\%$, ou seja, se comparado com o valor do desvio-padrão $(5,79\%)$, conclui-se que a variabilidade da proporção de perda de peso de todos os indivíduos é baixa.

Supor, agora, que o pesquisador esteja interessado em comparar a variabilidade da proporção de perda de peso entre os dois tipos de dietas. Para realizar essa comparação é necessário calcular o desvio-padrão da proporção média de perda de peso para cada uma das duas dietas. Para realizar esses cálculos utilizando o `Software R` basta executar a linha de comando abaixo e o `Software R` irá retornar as seguintes informações:

```
> tapply(dados$Perda,dados$Dieta,sd)
        A         B
4.739729  4.974854
```

Ou seja, a variabilidade da proporção de perda de peso dos indivíduos submetidos a Dieta A é de $4,74\%$, enquanto que a variabilidade da proporção de perda de peso dos indivíduos submetidos a Dieta B é de $4,97\%$. Como esses valores são muito próximos, praticamente iguais, pode-se concluir que os indivíduos submetidos a Dieta A e a Dieta B tendem a apresentar a mesma variabilidade na proporção de perda de peso.

De modo análogo é possível calcular, utilizando o `Software R`, o IIQ da proporção de perda de peso de todos os indivíduos, para isso basta utilizar o seguinte comando:

```
IQR(dados$Perda)
```

Observar que o argumento da função `IQR` (função que retorna o IIQ de um vetor de dados) é composto pelo vetor com os valores que se tem interesse em calcular o IIQ.

Se essa linha de comando for executada de forma correta, o `Software R` irá retornar o valor `10.295`. Ou seja, o IIQ da proporção de perda de peso de todos os indivíduos foi de $10,29\%$.

Supor, novamente, o interesse em comparar a variabilidade da proporção de perda de peso entre os dois tipos de dietas. Ou seja, para responder essa pergunta pode-se, também, calcular o IIQ da proporção de perda de peso para cada uma das duas dietas. Para realizar esses cálculos utilizando o `Software R` basta executar a linha de comando abaixo e o `Software R` irá retornar as seguintes informações:

```
> tapply(dados$Perda,dados$Dieta,IQR)
     A      B
6.4425 7.9750
```

Ou seja, a variabilidade da proporção de perda de peso dos indivíduos submetidos a Dieta A é de $6,44\%$, enquanto que a variabilidade da proporção de perda de peso dos indivíduos submetidos a Dieta B é de $7,97\%$. A interpretação aqui é a mesma observada quando foram comparados os desvios-padrões, ou seja, como esses valores são muito próximos, pode-se concluir que os indivíduos submetidos a Dieta A e a Dieta B tendem a apresentar a mesma variabilidade na proporção de perda de peso.