UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
Programa de Pós-Graduação em Matemática
Mestrado Profissional - PROFMAT/CCT/UFCG

CORRELAÇÃO LINEAR E REGRESSÃO LINEAR SIMPLES NO CONTEÚDO DE MATEMÁTICA DO ENSINO MÉDIO

Matheus Vinícius Francelino Queiroz

Trabalho de Conclusão de Curso

Orientador: Prof. Dr. Alexsandro Bezerra Cavalcanti

Campina Grande - PB
Junho/2020

Q3c Queiroz, Matheus Vinícius Francelino.
Correlação linear e regressão linear simples no conteúdo de matemática do ensino médio / Matheus Vinícius Francelino Queiroz. – Campina Grande, 2020.
72 f. : il. color.

Dissertação (Mestrado em Matemática) – Universidade Federal de Campina Grande, Centro de Ciências e Tecnologia, 2020.
"Orientação: Prof. Dr. Alexsandro Bezerra Cavalcanti".
Referências.

1. Estatística Descritiva. 2. Correlação Linear. 3. Regressão Linear Simples. 4. Matemática – Estudo e Ensino. I. Cavalcanti, Alexsandro Bezerra. II. Título.

CDU 519.2(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA SEVERINA SUELI DA SILVA OLIVEIRA CRB-15/225

UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
Programa de Pós-Graduação em Matemática
Mestrado Profissional - PROFMAT/CCT/UFCG

CORRELAÇÃO LINEAR E REGRESSÃO LINEAR SIMPLES NO CONTEÚDO DE MATEMÁTICA DO ENSINO MÉDIO

por

Matheus Vinícius Francelino Queiroz

Trabalho de Conclusão de Curso apresentado ao Corpo Docente do Programa de Pós-Graduação em Matemática - CCT - UFCG, na modalidade Mestrado Profissional, como requisito parcial para obtenção do título de Mestre.

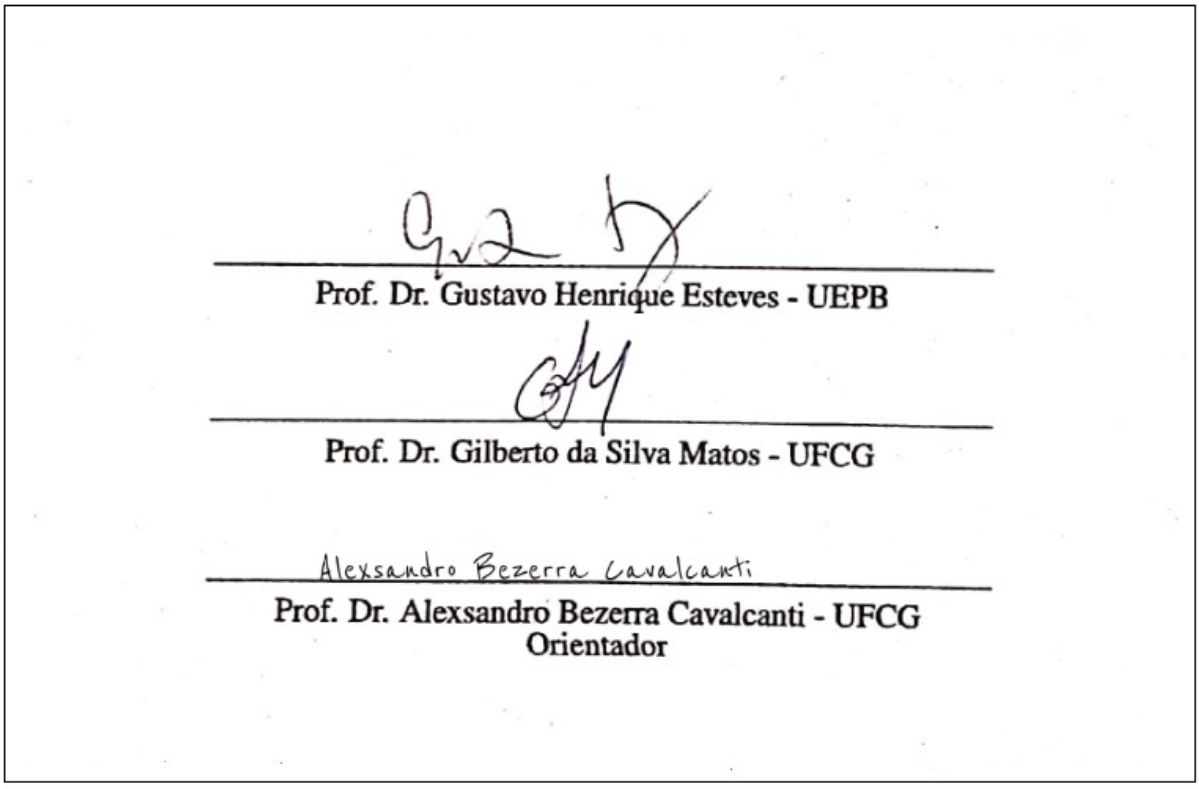
CORRELAÇÃO LINEAR E REGRESSÃO LINEAR SIMPLES NO CONTEÚDO DE MATEMÁTICA DO ENSINO MÉDIO

por

Matheus Vinícius Francelino Queiroz

Trabalho de Conclusão de Curso apresentado ao Corpo Docente do Programa de Pós-Graduação em Matemática - CCT - UFCG, modalidade Mestrado Profissional, como requisito parcial para obtenção do título de Mestre.

Aprovado por:



Universidade Federal de Campina Grande
Centro de Ciências e Tecnologia
Unidade Acadêmica de Matemática
Curso de Mestrado Profissional em Matemática em Rede Nacional

Junho/2020

Dedicatória

Aos meus pais Elídia e Zezinho e às minhas avós Mariinha e Lídia.

Agradecimentos

Primeiramente, agradeço a Deus por me proporcionar todas as vitórias que vem acontecendo em minha vida e por colocar nela todas as pessoas maravilhosas que citarei nestes agradecimentos.

Agradeço,

Ao meu Pai Zezinho por toda a força que me deu e pelas longas viagens que realizamos até Campina Grande, eram 600 km percorridos toda sexta-feira, mas ele sempre estava lá ao meu lado. Sem esquecer que meus irmãos Lucas, Marcos, e Suzy e minha namorada Héryca também partilharam de algumas dessas viagens para me fazer companhia, sou extremamente grato.

À minha Mãe Elídia por ser essa mulher incrível que mesmo em meio as dificuldades, sempre se preocupa com seus filhos e não mede esforços para ajudá-los. Muito obrigado, a senhora é a MELHOR Mãe do mundo, tenho orgulho de ser seu filho.

Às minhas avós Mariinha e Lídia por sempre me colocarem em suas orações e pelos incentivos que elas me dão em todos os aspectos da minha vida.

À minha boadrasta Suzana e meus irmãos Lucas, Suzy e João Arthur por sempre aguardarem a minha chegada e a de meu pai das viagens de Campina Grande, além da alegria e entusiasmos que só eles têm.

Ao meu padrasto Nal e meu irmão Marcos pela força que eles têm me dados indiretamente, só nós aqui de casa sabemos a provação que estamos passando e sei que Deus vai nos dar essa vitória, eu creio.

À minha namorada Héryca pelo companheirismo, calma e dedicação que ela tem tido ao longo desta jornada. Tenho muita sorte de tê-la ao meu lado.

Ao meu melhor amigo Miguel pelos momentos de descontração, brincadeiras, incentivos e conselhos. Tudo me ajudou muito e continua me ajudando como pessoa e profissional.

Aos meus antigos professores/colegas de trabalho e, ainda, amigos para a vida toda: o pessoal da EREM AIRES GAMA. Desde minha época de estudante sempre fui muito incentivado por eles e isso não foi diferente quando me tornei colega de trabalho. Não é mesmo? Gilberto, Ledjane, Cláudia, Dona Zezé, Luciana, Liliane, Cida Melo, Evandro, Jaqueline, Dona Artemes, Richard, Geane, Adelmo, Ana Paula, Gorethe, Lane, Marla, Nazaré, Rose e Sérgio. Sem esquecer do pessoal da segurança Élio e Bob, assim como a meninas da limpeza Mirian, Cielma e Cirlane e da cozinha Gilda e Edna.

Aos meus colegas de mestrado que estão concluindo e aqueles que ficaram ao longo da caminhada, só nós sabemos cada barreira que enfrentamos. Quero aqui dar destaque a algum destes: Lucielma, minha companheira de viagens e longas conversas (isto quando ela não estava dormindo) obrigado por cada momento; Airtonelton, meu ex-professor de graduação e amigo, sempre disposto a ajudar tanto em relação aos assuntos do mestrado quando fora dele; por falar em ajudar no mestrado não poderia deixar de citar Marília e Bruno, pense em duas pessoas que sempre tem o que você precisa; sem esquecer de Márcio, Sandra pela grande afinidade que tínhamos; e por fim Wagner, Hydayane, Eduardo, Rejane, Teófilo, Renato Geraldo e Daniel vocês são pessoas incríveis e os levarei para sempre em minha memória.

Aos meus amigos de Serra Talhada: Tiago Melo (e também compadre), pelo incentivo que me deu para realizar o ENA e pelas espetaculares e divertidas aulas que tive com você na FAFOPST; e Isaías Lima, sempre disposto a tirar minhas dúvidas, dar dicas, corrigir erros e muito mais, você me ajudou MUITO a passar no concurso do IF-Sertão, muito obrigado.

Ao meu amigo Matheus Dantas, por todas as conversas, discussões, brincadeiras, desabafos e compartilhamentos das mesmas emoções que passamos no mestrado. Pode ter certeza que aprendi muito com você meu xará.

Aos meus professores do programa: Deise Mara Barbosa de Almeida, Denilson da Silva Pereira, Jaime Alves Barbosa Sobrinho, Luiz Antônio da Silva Medeiros, Marcelo Carvalho Ferreira e Romildo Nascimento de Lima, aprendi muito com cada um de vocês.

Ao meu orientador Alexsandro Bezerra Cavalcanti, o senhor foi excelente em suas aulas e orientações, sempre muito calmo e disposto a ajudar, mesmo em fins de semana, muito obrigado professor.

À UFCG por ter me concedido a oportunidade de realizar este mestrado incrível.

Por fim, agradeço à Sociedade Brasileira da Matemática - SBM pelo oferecimento deste Curso em Rede Nacional.

Resumo

Neste trabalho, propomos a inclusão dos conteúdos de Correlação Linear e o cálculo dos coeficientes da reta de Regressão Linear Simples no conjunto de conteúdos de Matemática do Ensino Médio, uma vez que tais assuntos são uma aplicação direta dos conceitos da Estatística Descritiva. Para tanto, realizamos, num primeiro momento, uma fundamentação teórica a respeito dos principais resultados relacionados à Estatística Descritiva. Em seguida, desenvolvemos os conteúdos de Correlação Linear e Regressão Linear Simples a partir da fundamentação teórica realizada anteriormente. Por fim, apresentamos uma proposta didática aliada a resolução de problemas para o ensino da Correlação e Regressão Lineares utilizando o Software Matemático GeoGebra, que é uma das Tecnologias de Informação e Comunicação (TIC) mais utilizadas nos últimos anos para o Ensino da Matemática.

Palavras Chaves: Estatística Descritiva. Correlação Linear. Regressão Linear Simples.

Abstract

In this work, we propose the inclusion of the Linear Correlation and the Simple Linear Regression calculation of coefficients on the High School Mathematics curriculum, since such subjects are a direct application of the concepts of Descriptive Statistics. For this purpose, we present, at first, a theoretical foundation regarding the main results related to Descriptive Statistics. Then, we proceed to build upon said theoretical foundation the Linear Correlation and Simple Linear Regression curricular contents. Finally, we present a didactic proposal combined with a problem-solving methodology for teaching Linear Correlation and Regression using the GeoGebra Mathematical Software, which is one of the most used Information and Communication Technologies (ICT) in recent years for Mathematic Teaching.

Keywords: Descriptive Statistics. Linear Correlation. Simple Linear Regression.

Lista de Figuras

Lista de Tabelas

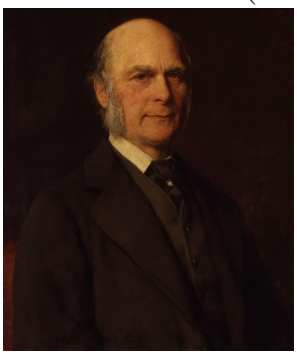
Sumário

Capítulo 1

Introdução

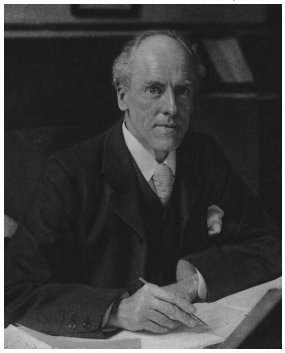
Correlação e Regressão são métodos estatísticos que compreendem a análise de dados amostrais para saber se e como duas ou mais variáveis quantitativas estão relacionadas entre si. Os principais nomes que associamos a este estudo são os dos britânicos: Francis Galton (Figura 1.1) e Karl Pearson (Figura 1.2). Este trabalho se dedica ao caso da correlação Linear e Regressão Linear Simples, que é a análise envolvendo apenas duas variáveis quantitativas.

Figura 1.1: Francis Galton (1822-1911)



Fonte: [11] http://galton.org/

Figura 1.2: Karl Pearson (1857-1936)



Fonte: [12] https://karlpearson.org/

De modo geral, dizemos que existe uma correlação entre duas variáveis quando uma delas está, de alguma forma, relacionada com a outra. Esta relação pode ser verificada visualmente por meio do Diagrama de Dispersão, que consiste em um sistema de eixos ortogonais: o eixo horizontal representa os valores da variável independente, ou explicativa, que denotamos por X e o eixo vertical representa os valores da variável dependente, ou resposta, que denotamos por Y. Em particular, a Correlação Linear mede a força ou grau de relacionamento linear entre duas variáveis através do chamado Coeficiente de Correlação Linear (ou, Coeficiente de Correlação de Pearson, em homenagem ao matemático Karl Pearson).

Por outro lado, a Regressão Linear Simples explicita a forma como ocorre essa relação, isto é, ela nos dá uma equação que descreve o comportamento de uma das variáveis em função do comportamento da outra. De acordo com Bussab [3] (2017, p.505)

> O uso do termo regressão deve-se a Galton, por volta de 1885, quando investigava relações entre características antropométricas de sucessivas gerações. Uma de suas constatações era de que "cada peculiaridade de um homem é transmitida aos seus descendentes, mas, em média, numa intensidade menor". Por exemplo: embora pais com baixa estatura tendam a ter filhos também com baixa estatura, estes têm altura média maior do que a altura média de seus pais. O mesmo ocorre, mas em direção contrária, com pais com estatura alta.

Os conteúdos de Correlação Linear e o cálculo dos coeficientes da reta de Regressão Linear Simples possuem bases na Estatística Descritiva; em outras palavras, aqueles podem ser vistos como aplicações desta. Assim, uma vez previsto, na Base Nacional Comum Curricular - BNCC, o estudo da Estatística Descritiva no Ensino Médio, os alunos deste nível de ensino teriam totais condições de compreender o desenvolvimento de tais conteúdos.

Quando de posse dos conhecimentos relacionados à Correlação Linear e à Regressão Linear Simples, os alunos adquirem um ganho substancial em sua formação, visto que esses conteúdos são ferramentas fortes e eficazes para estudos relacionados a muitas áreas do conhecimento, como por exemplo: na química, no estudo das transformações dos gases e solubilidade de algumas substâncias; na biologia, quando se pretende relacionar a altura e o diâmetro de árvores ao longo do tempo ou a disponibilidade de alimento com crescimento de peixes; na engenharia, quando se analisa o crescimento populacional de certa região para estimar uma vazão para o abastecimento de água ou para a determinação de uma estação de tratamento de esgoto; etc.

É sabido que a BNCC do Ensino Médio garante que a área de Matemática e suas Tecnologias deve assegurar ao estudante o desenvolvimento de cinco competências[1] específicas. Relacionadas a cada uma delas, é indicado um rol de habilidades a serem alcançadas com o propósito do pleno desenvolvimento da respectiva competência.

A habilidade EM13MAT510[2] da Competência Específica 5, possui a seguinte redação "Investigar conjuntos de dados relativos ao comportamento de duas variáveis numéricas, usando tecnologias da informação, e, se apropriado, levar em conta a variação e utilizar uma reta para descrever a relação observada." (BRASIL [1] 2017, p.533). Analisando cui-

[1] Estas competências estão descritas no apêndice B.

[2] O primeiro par de letras (EM) indica a etapa de ensino, neste caso a etapa é o Ensino Médio. O primeiro par de números (13) indica que as habilidades descritas podem ser desenvolvidas em qualquer série do Ensino Médio, conforme definição dos currículos. A segunda sequência de letras indica a área, ou seja, MAT = Matemática e suas Tecnologias. Finalmente, os números finais indicam a competência específica à qual se relaciona a habilidade (1º número) e a sua numeração no conjunto de habilidades relativas a cada competência (dois últimos números).

dadosamente esta habilidade, percebe-se que ela propõe um estudo superficial a respeito de Correlação Linear e Regressão Linear Simples, sem que seja necessário apresentar tais assuntos de forma direta.

Todavia, a área de Matemática e suas Tecnologias não propõe, em outra oportunidade, o tratamento formal de tais conteúdos, o que seria extremamente viável, tendo em vista a complementaridade da habilidade EM13MAT510 e o ganho na formação do aluno, como já destacado anteriormente.

Portanto, visando o pleno desenvolvimento da formação do aluno e a complementaridade da habilidade EM13MAT510, este trabalho tem como finalidade alcançar os objetivos (geral e específicos) elencados na seção a seguir.

1.1 Objetivos

1.1.1 Objetivo Geral

Propor a inserção dos conteúdos de Correlação Linear e o cálculo dos coeficientes da reta de Regressão Linear Simples no Ensino Médio como uma aplicação dos conceitos da Estatística Descritiva.

1.1.2 Objetivos Específicos

- Apresentar uma fundamentação teórica a respeito dos principais resultados da Estatística Descritiva;
- Expor os conteúdos relacionados à Correlação Linear e à Regressão Linear Simples a partir do estudo da Estatística Descritiva;
- Desenvolver uma proposta didática para o ensino dos conteúdos de Correlação Linear e o cálculo dos coeficientes da reta de Regressão Linear Simples utilizando o Software Matemático GeoGebra [13].

1.2 Organização

Este TCC está organizado da seguinte maneira. Além deste, temos os seguintes capítulos:

- Capítulo 2: Apresentamos os aportes teóricos relacionados à Estatística Descritiva;
- Capítulos 3 e 4: Desenvolvemos os conteúdos de Correlação Linear e Regressão Linear Simples, respectivamente, a partir do estudo da Estatística Descritiva;

- Capítulo 5: Apresentamos uma proposta didática para o ensino dos conteúdos de Correlação Linear e Regressão Linear Simples utilizando o Software Matemático GeoGebra;
- Capítulo 6: Apresentamos as considerações finais do trabalho;
- Por fim, as Referências Bibliográficas e os Apêndices, nesta ordem.

Capítulo 2

Fundamentação Teórica

2.1 Conceitos Fundamentais e Definições

A estatística é a ciência que realiza pesquisas com o intuito de coletar, organizar, analisar e interpretar dados de uma pequena parte de um grupo maior, de modo que possamos conhecer algo sobre esse grupo maior e, a partir daí, sermos capazes de tomar uma série de decisões. Esse é um objetivo comum e importante da estatística: aprender sobre um grande grupo pela análise de dados de alguns de seus membros.

Nesse contexto, há dois tipos de conjuntos de dados que têm significados especiais: *população* e *amostra*. A seguir serão apresentadas definições formais desses conceitos e de outros termos básicos.

Dados são coleções de observações (por exemplo: medidas, gêneros, resposta de pesquisas).

População é o conjunto de todos os elementos ou resultados sob investigação.

Amostra é qualquer *subconjunto* da população.

Se faz necessário que os dados amostrais devem ser selecionados de modo apropriado, ou seja, que tais dados sejam representativos da população do qual são extraídos, caso contrário as conclusões a respeito da população podem sair distorcidas. "Se os dados não forem coletados de modo apropriado, podem ser de tal maneira inúteis que nenhuma manipulação estatística poderá salvá-los" (TRIOLA [9], 2013, p.4).

A cada elemento da população, ou da amostra, associamos uma característica de interesse que será chamada de **variável**. As variáveis, por sua vez, classificam-se em:

Qualitativas (ou categóricas): é uma variável onde seus possíveis valores são expressos por atributos ou qualidades. Tais variáveis ainda podem ser reclassificadas em dois tipos:

- Nominal: não existe uma ordenação no conjunto dos possíveis resultados. Exemplos: sexo, cor dos olhos, estado civil.

- Ordinal: existe uma ordenação no conjunto dos possíveis resultados. Exemplos: escolaridade (ensino fundamental, ensino médio, superior), estágio de uma doença (inicial, intermediário, terminal), dia de observação (domingo, segunda, ..., sábado).

Quantitativas (ou numéricas): é uma variável onde seus possíveis valores são expressos por números. De forma análoga as variáveis qualitativas, as quantitativas também podem ser reclassificadas em dois tipos:

- Discreta: os seus possíveis valores variam em um conjunto finito ou enumerável, em geral, resultantes de contagens. Exemplos: número de filhos, idade (em anos), número de cigarros fumados por dia.

- Contínua: os seus possíveis valores variam em um subconjunto dos números reais, em geral, resultantes de mensurações. Exemplos: salário, altura, peso.

2.2 Distribuição de Frequências

De acordo com Triola [9] (2013, p.39) "Uma distribuição de frequências (ou tabela de frequência) mostra como o conjunto de dados é divido entre todas as várias categorias (ou classes), listando todas as categorias juntamente com o número de valores de dados em cada uma delas". Em particular, uma distribuição de frequências nos ajuda a entender a natureza da distribuição de uma variável.

Apresentaremos a seguir um conjunto de conceitos fundamentais para a elaboração e análise das distribuições de frequências.

1. **Dados Brutos** - São os dados originais obtidos após a coleta e que não se encontram organizados numericamente.

2. **Rol** - São os dados brutos organizados em uma determinada ordem (crescente ou decrescente).

3. **Amplitude Total** (AT) - É a diferença obtida entre o maior e o menor valor observado da variável sob estudo.

4. **Frequência Absoluta** (n_i) - É o número de vezes em que cada elemento da variável se repete na amostra ou o número de elementos pertencentes a i-ésima classe, $i = 1, ..., k$.

   A soma das frequências absolutas é igual ao número total de observações

$$\sum_{i=1}^{k} n_i = n. \tag{2.1}$$

5. **Frequência Relativa** (f_i) - É a razão entre a i-ésima frequência absoluta e o número total de observações (n)

$$f_i = \frac{n_i}{n}, \text{ para } i = 1, ..., k. \tag{2.2}$$

Pode-se expressar esse resultado em termos percentuais multiplicando a frequência relativa por 100. A soma das frequências relativas deve ser igual a 1. De fato,

$$\sum_{i=1}^{k} f_i = \frac{\sum n_i}{n} = \frac{n}{n} = 1. \tag{2.3}$$

6. **Frequência Acumulada** (F_i) - É a soma da frequência da i-ésima classe mais as frequências de todas as classes anteriores.

Para que uma análise da distribuição de frequências seja feita de forma eficiente, em geral, é interessante organizar os dados em tabelas. Tais tabelas podem representar dois tipos de valores: pontuais ou agrupados em classes. Vejamos quando utilizar cada uma delas.

2.2.1 Tabela de Frequências Pontuais

É uma tabela onde os valores da variável aparecem individualmente. Esse tipo de distribuição é utilizado geralmente para representar uma variável discreta, com pouca variedade de valores ou variáveis qualitativas.

Os Exemplos 2.1 e 2.2 a seguir tratam dos casos de uma variável discreta com pouca variedade de valores e uma variável qualitativa, respectivamente.

Exemplo 2.1: A Tabela 2.1 nos mostra a distribuição das frequências (absoluta, relativa e acumulada) da variável *frequência semanal*, que foi obtida a partir de uma entrevista com 20 moradores de um bairro que fazem uso de um parque ali situado.

Tabela 2.1: Frequências da variável: *frequência semanal*

Frequência semanal	Frequência absoluta (n_i)	Frequência relativa (f_i)	Porcentagem ($100 f_i\%$)	Frequência acumulada (F_i)
1	3	0,15	15%	3
2	4	0,20	20%	7
3	6	0,30	30%	13
4	3	0,15	15%	16
5	2	0,10	10%	18
6	1	0,05	5%	19
7	1	0,05	5%	20
Total	20	1,00	100%	-

Fonte: Iezzi [6] (2013, p.74).

Exemplo 2.2: A Tabela 2.2 nos mostra a distribuição das frequências da variável *meio de transporte*, que os moradores do bairro do Exemplo 2.1 utilizam para chegar ao parque que fazem uso.

Tabela 2.2: Frequências da variável: *meio de transporte*

Meio de transporte	Frequência absoluta (n_i)	Frequência relativa (f_i)	Porcentagem ($100f_i\%$)	Frequência acumulada (F_i)
Carro	5	0,25	25%	5
Ônibus	5	0,25	25%	10
A pé	10	0,50	50%	20
Total	20	1,00	100%	-

Fonte: Iezzi [6] (2013, p.74).

2.2.2 Tabela de Frequências Agrupadas em Classes

É uma tabela onde os valores da variável aparecem agrupados em classes, que são intervalos de variação da variável. Esse tipo de distribuição é indicado para representar uma variável contínua ou discreta com uma grande variedade de valores.

A seguir, definiremos alguns termos-padrão utilizados na discussão e construção de tabelas de frequências agrupadas em classes.

1. **Números de classes** (k): De acordo com Bussab [3] (2017, p.16) "a escolha dos intervalos é arbitrária e a familiaridade do pesquisador com os dados é que lhe indicará quantas e quais classes (intervalos) devem ser usadas". Contudo, deve-se levar em consideração que um número pequeno de classes faz com que se perca informações e, por outro lado, com um número grande de classes, pode haver alguma classe com uma frequência pequena ou até mesmo nula, apresentando uma distribuição irregular e prejudicando a análise.

   Em relação à escolha do número de classes para a construção de uma tabela de frequências, Bussab (2017) menciona que normalmente faz-se o uso de 5 a 15 classes ; por outro lado, Triola (2013) afirma que este número deve estar entre 5 e 20.

   Abaixo são apresentados dois critérios que nos dão uma ideia de como escolher o número de classes (k):

   (a) **A regra da Raiz Quadrada:** $k = 5$, para $n \leqslant 25$ e $k \cong \sqrt{n}$, para $n > 25$;

   (b) **A regra de Sturges:** $k = 1 + 3,3 \log n$.
   Onde n é o número de observações.

2. **Limites inferiores de classe (l_i):** são os menores números que podem pertencer às diferentes classes.

3. **Limites superiores de classe (L_i):** são os maiores números que podem pertencer às diferentes classes.

4. **Amplitude da Classe (Δ_i):** É a diferença entre os limites superior e inferior da classe.

$$\Delta_i = L_i - l_i, \text{ para } i = 1, ..., k. \tag{2.4}$$

5. **Pontos médios das classes (s_i):** são os valores no meio do intervalo. Cada ponto médio de classe pode ser encontrado somando-se o limite inferior de classe ao limite superior de classe e dividindo-se a soma por 2. Ou seja,

$$s_i = \frac{L_i + l_i}{2}, \text{ para } i = 1, ..., k. \tag{2.5}$$

Com esses conceitos definidos temos condições de construir tabelas de frequências agrupadas em classes. Os Exemplos 2.3 e 2.4 a seguir tratam dos casos de uma variável contínua e uma variável discreta com grande variedade de valores, respectivamente.

Exemplo 2.3: Na Tabela 2.3 temos a distribuição das frequências da variável *renda familiar mensal* (em salários mínimos) dos moradores do bairro do Exemplo 2.1.

Tabela 2.3: Frequências da variável: *renda familiar mensal*

Renda familiar mensal	Frequência absoluta (n_i)	Frequência relativa (f_i)	Porcentagem ($100 f_i\%$)	Frequência acumulada (F_i)
$5 \vdash 8$	2	0,10	10%	2
$8 \vdash 11$	5	0,25	25%	7
$11 \vdash 14$	7	0,35	35%	14
$14 \vdash 17$	4	0,20	20%	18
$17 \vdash 20$	2	0,10	10%	20
Total	20	1,00	100%	-

Fonte: Iezzi [6] (2013, p.74).

Perceba que estamos utilizando a notação $l_i \vdash L_i$ para indicar que o intervalo contém o limite inferior, mas não contém o limite superior, outras possibilidades são $l_i \dashv L_i$, $l_i \nvdash L_i$ e $l_i - L_i$. Uma notação equivalente é $[l_i, L_i), (l_i, L_i], [l_i, L_i]$ e (l_i, L_i).

Exemplo 2.4: A Tabela 2.4 nos mostra a distribuição das frequências da variável *idade* dos moradores do bairro do Exemplo 2.1.

Tabela 2.4: Frequências da variável: *idade*

Idade	Frequência absoluta (n_i)	Frequência relativa (f_i)	Porcentagem ($100f_i\%$)	Frequência acumulada (F_i)
$18 \vdash 26$	3	0,15	15%	3
$26 \vdash 34$	7	0,35	35%	10
$34 \vdash 42$	4	0,20	20%	14
$42 \vdash 50$	3	0,15	15%	17
$50 \vdash 58$	3	0,15	15%	20
Total	20	1,00	100%	-

Fonte: Iezzi [6] (2013, p.74).

2.3 Gráficos

Os gráficos constituem um importante instrumento de análise e interpretação de um conjunto de dados. Segundo Guedes [5] (2005, p.17)

> Gráfico é um recurso visual da Estatística utilizado para representar um fenômeno. Sua utilização em larga escala nos meios de comunicação social, técnica e científica, devem-se tanto à sua capacidade de refletir padrões gerais e particulares do conjunto de dados em observação, como à facilidade de interpretação e a eficiência com que resume informações dos mesmos.

Em relação as tabelas, os gráficos apresentam um grau menor de detalhamento, porém, estes têm a vantagem de, rápida e concisamente, informar sobre a variabilidade de um conjunto de dados. "Uma representação gráfica pode colocar em evidência as tendências, as ocorrências ocasionais, os valores mínimos e máximos e as ordens de grandezas dos fenômenos que estão sendo observados" (GUEDES [5], 2005, p.17).

Na elaboração de um gráfico é essencial que ele apresente: um título (que norteará o leitor quanto a natureza do conteúdo), onde, quando e por quanto tempo o fato em estudo se destaca e uma escala adequada. Todo gráfico deve zelar pela simplicidade, clareza e veracidade nas informações.

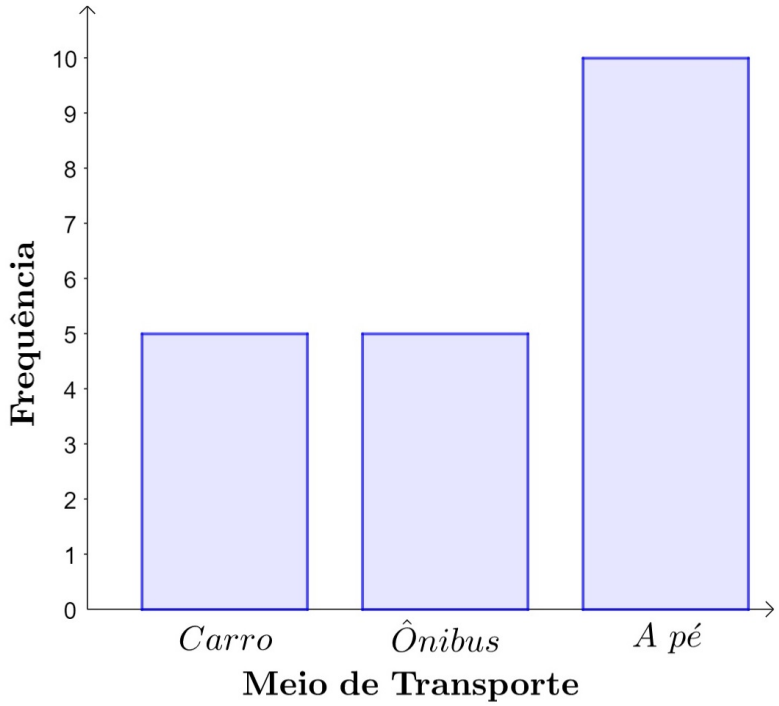
2.3.1 Gráficos para Variáveis Qualitativas

Existem uma grande diversidade de gráficos que são utilizados para representar variáveis qualitativas; contudo, iremos tratar aqui apenas dos gráficos de *barras* e de *setores* que são os mais frequentes em pesquisas e textos científicos.

Gráfico de Barras - É um gráfico formado por retângulos paralelos (horizontais ou verticais), onde uma das dimensões é proporcional a frequência (n_i ou f_i) da variável em estudo, e a outra arbitrária, porém igual para todos os retângulos. Este tipo de gráfico é recomendado quando se deseja comparar grandezas.

Exemplo 2.5: A Figura 2.1 mostra o gráfico em barras para a variável *meio de transporte* do Exemplo 2.2.

Figura 2.1: Gráfico de barras para a variável *meio de transporte*

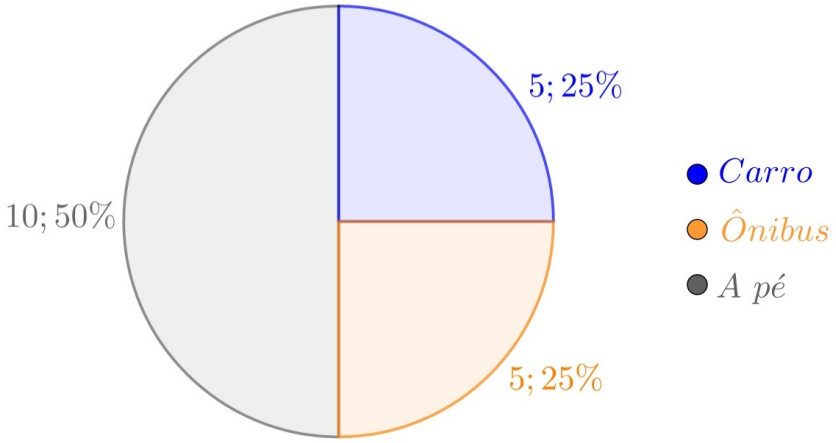


Fonte: Produção do autor.

Gráficos de Setores - É um gráfico onde a variável em estudo é projetada em círculo de raio arbitrário, representando o todo, dividido em setores com áreas proporcionais às frequências das partes. Este tipo de gráfico é recomendado quando se deseja comparar as frequências das partes com o todo.

Exemplo 2.6: A Figura 2.2 mostra o gráfico de setores para a variável *meio de transporte*.

Figura 2.2: Gráfico de setores para a variável *meio de transporte*



Fonte: Produção do autor.

2.3.2 Gráficos para Variáveis Quantitativas

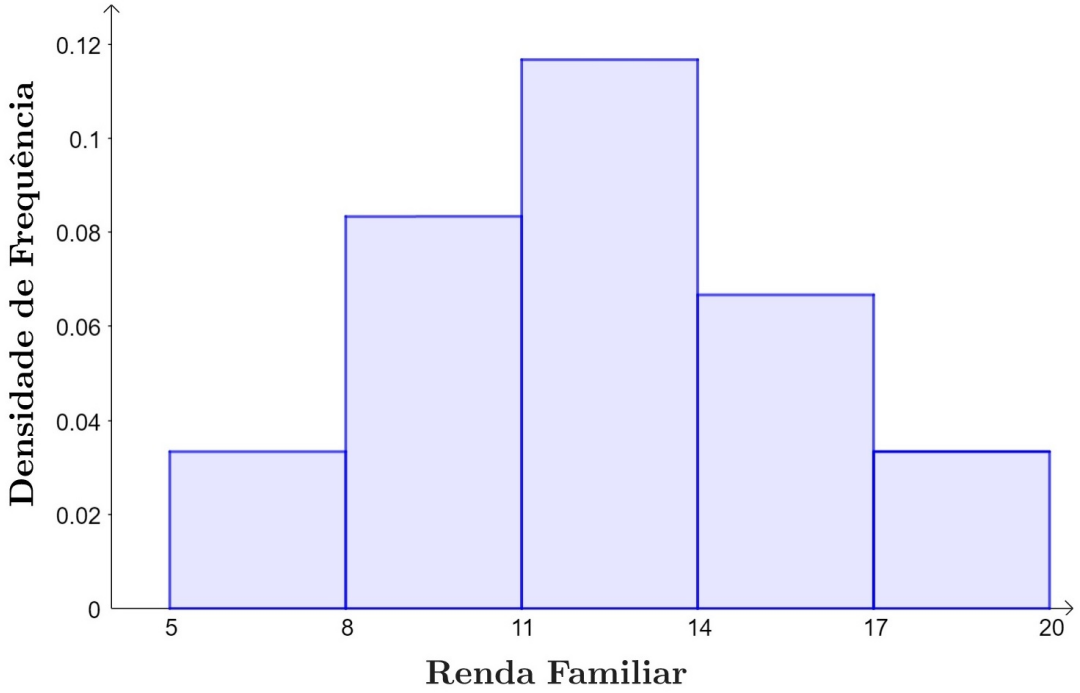
Podemos considerar uma maior variedade de gráficos que podem representar uma variável quantitativa, como por exemplo: Histograma, *Box Plots*, Ramo-e-Folha etc. Aqui vamos tratar, em particular, do Histograma. Para informações a respeito do *Box Plots* e do Ramo-e-Folha o leitor pode consultar [3] (BUSSAB, 2017).

Histograma - Este tipo de gráfico é adequado para representar uma distribuição de frequência para variáveis quantitativas contínuas ou para variáveis quantitativas discretas com uma grande variedade de valores. Ele é formado por retângulos justapostos, onde uma das dimensões é proporcional a amplitude da classe (Δ_i) representada e a área de cada retângulo é proporcional à respectiva frequência, n_i ou f_i. As classes são localizadas no eixo horizontal e as frequências no eixo vertical.

A fim de que a área de cada retângulo seja proporcional a f_i (n_i), a sua altura deve ser igual a f_i/Δ_i (n_i/Δ_i), que é chamada densidade de frequência da i-ésima classe. Quando a amplitude das classes forem todas iguais a Δ, a densidade de frequência da i-ésima classe passa a ser f_i/Δ (n_i/Δ). Estabelecida esta convenção, a área total do histograma será igual a 1 (n).

Exemplo 2.7: A Figura 2.3 abaixo mostra o histograma para a variável *renda familiar*, do Exemplo 2.3.

Figura 2.3: Histograma da variável *renda familiar*



Fonte: Produção do autor.

2.4 Medidas de Tendência Central

Nas seções 2.2 e 2.3, vimos, respectivamente, como resumir um conjunto de dados em tabelas de frequências e como representá-los graficamente. Frequentemente, se faz necessário resumir ainda mais estes dados, apresentando um ou mais valores que sejam *representativos* da variável em estudo. Habitualmente, fazemos uso das seguintes medidas de tendência central: Média Aritmética, Mediana ou Moda.

2.4.1 Média Aritmética

A média aritmética (ou simplesmente média) é, em geral, a mais importante e, provavelmente, a mais utilizada de todas as medidas de tendência central.

Considere um conjunto de n observações $x_1, ..., x_n$, a média aritmética é definida como a medida encontrada pela soma de todas as observações, dividida pelo número de observações. Expressando essa definição em termos matemáticos:

$$\text{média} = \frac{x_1 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{2.6}$$

Se no conjunto das n observações tivermos: n_1 iguais a x_1, n_2 iguais a $x_2, ..., n_k$ iguais a x_k, então (2.6) pode ser reescrita como

$$\text{média} = \frac{n_1x_1 + n_2x_2 + ... + n_kx_k}{n} = \frac{\sum_{i=1}^{k} n_ix_i}{n}. \tag{2.7}$$

Se os dados são uma amostra de uma população, a média é representada por $\overline{x}$ (lê-se "x barra"); entretanto, se os dados são a população inteira, então representamos a média por μ.

Uma forma equivalente para a média aritmética é dada por

$$\text{média} = \sum_{i=1}^{n} x_i f_i, \tag{2.8}$$

onde recordamos que $f_i = n_i/n$ é a frequência relativa da observação x_i.

Exemplo 2.8: Considere a variável *frequência semanal*, do Exemplo 2.1. Veja que

$$\overline{x} = \frac{3\times 1 + 4\times 2 + 6\times 3 + 3\times 4 + 2\times 5 + 6 + 7}{20} = \frac{64}{20} = 3,2.$$

Esse resultado nos diz que a frequência semanal média dos 20 moradores que fazem uso do parque em seu bairro é de 3,2 dias. Note que, neste caso, representamos a média por $\overline{x}$, pois os dados em questão representam uma amostra da população (todos os moradores do bairro).

Cálculo Aproximadado da Média Aritmética

Ao lidarmos com um conjunto de dados agrupados em classes, podemos encontrar uma medida aproximada para a média aritmética convencionando que na sua fórmula o valor da observação (x_i) seja substituído pelo ponto médio (s_i) da i-ésima classe.

Exemplo 2.9: Consideremos a variável *renda familiar mensal*, do Exemplo 2.3. A Tabela 2.5 mostra a distribuição de frequência desta variável, bem como o ponto médio de cada classe.

Tabela 2.5: Pontos médios das classes da Tabela 2.3.

Renda familiar mensal	Frequência absoluta (n_i)	Frequência relativa (f_i)	Ponto médio (s_i)
$5 \vdash 8$	2	0,10	6,5
$8 \vdash 11$	5	0,25	9,5
$11 \vdash 14$	7	0,35	12,5
$14 \vdash 17$	4	0,20	15,5
$17 \vdash 20$	2	0,10	18,5
Total	20	1,00	–

Fonte: Tabela 2.3.

Portanto, a renda familiar mensal média aproximada é

$$\begin{aligned}\overline{x} &\approx \frac{2\times 6,5+5\times 9,5+7\times 12,5+4\times 15,5+2\times 18,5}{20}\\ &= \frac{247}{20}=12,35 \text{ salários mínimos.}\end{aligned}$$

2.4.2 Mediana

A mediana de um conjunto de dados é a medida que deixa 50% das observações abaixo dela e 50% das observações acima dela e, em geral, é representada por Md.

Sejam $x_{(1)},x_{(2)},...,x_{(n)}$ os valores de um conjunto de dados de tal modo que:

$$x_{(1)} \leqslant x_{(2)} \leqslant ... \leqslant x_{(n)}.$$

Com essa notação, a mediana pode ser definida como

$$Md = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, \text{ se } n \text{ é ímpar;} \\ \\ \frac{x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}}{2}, \text{ se } n \text{ é par.} \end{cases} \tag{2.9}$$

Exemplo 2.10: Consideremos a distribuição da variável *frequência semanal*, do Exemplo 2.1. Como $n=20$, a mediana será a média aritmética dos números que ocupam a 10ª e a 11ª posição após a ordenação dos valores:

$$1,1,1,2,2,2,2,3,3,3,3,3,3,4,4,4,5,5,6,7.$$

Como podemos ver, tais números são $x_{(10)}=3$ e $x_{(11)}=3$, assim

$$Md=\frac{x_{(10)}+x_{(11)}}{2}=\frac{3+3}{2}=3 \text{ dias.}$$

Cálculo Aproximado da Mediana

Por intermédio do histograma utilizamos o fato de as áreas dos retângulos serem proporcionais às frequências das classes e, através de uma regra de três simples, podemos encontrar um valor aproximado para a mediana.

Retomemos o Exemplo 2.3, que trata da variável *renda familiar mensal*. Vamos, inicialmente, calcular o valor exato da mediana desta variável: como $n=20$, a mediana será a média aritmética dos números que ocupam a 10ª e 11ª posição após a ordenação dos valores; tais números são $12,1$ e $12,4$, respectivamente. Logo,

$$Md = \frac{12,1 + 12,4}{2} = 12,25 \text{ salários mínimos.}$$

Por meio do histograma desta variável devemos encontrar o ponto das abscissas que acumula 50% das observações abaixo (ou acima) dele . As áreas dos dois primeiros retângulos acumulam 35% do total, os três primeiros acumulam 70%, portanto a mediana é um número situado entre 11 e 14. Em outras palavras, a mediana é o ponto das abscissas do terceiro retângulo, de modo que a área do retângulo de base $Md - 11$ e mesma altura que o retângulo de base $14 - 11$, seja 15% (35% dos dois primeiros retângulos mais 15% do terceiro, resultam 50%). Agora, utilizando uma regra de três simples, obtemos o seguinte resultado:

$$\frac{Md - 11}{15} = \frac{14 - 11}{35} \Rightarrow Md = 11 + 15 \times \frac{3}{35} = 12,28 \text{ salários mínimos.}$$

Em muitos casos, é preferível utilizar a mediana como medida de tendência central em vez da média, isto ocorre pelo fato dela ser uma medida resistente a valores atípicos, ao passo que a média não é. Afirmar que uma medida não é resistente significa dizer que um único valor atípico no conjunto de dados pode afetar radicalmente o valor desta medida. Vejamos um exemplo nesse sentido.

Exemplo 2.11: Considere o consumo mensal de água, em metros cúbicos, de uma residência nos nove primeiros meses de um ano: 33, 31, 34, 32, 34, 32, 102, 34 e 30. Calculando a média mensal de consumo, obtemos:

$$\bar{x} = \frac{33 + 31 + 34 + 32 + 34 + 32 + 102 + 34 + 30}{9} = \frac{362}{9} \approx 40,2\ m^3.$$

O valor de $40,2\ m^3$ encontrado para a média não representa, com fidelidade, uma medida de tendência central: o consumo mensal dessa residência aponta para um valor entre 30 e 35 metros cúbicos; além disso, dos 9 valores registrados, 8 são menores que a média e "distantes", ao menos, 6 unidades dela e apenas 1 valor é maior que a média, estando muito distante dela.

Nessa situação, a média foi afetada por um valor atípico do consumo, que destoa dos demais: o valor de $102\ m^3$, que pode ser explicado por algum fator não corriqueiro dentro do mês em questão.

Por outro lado, como $n = 9$ a mediana será o valor que ocupa a 5^a posição após a ordenação dos valores:

$$30, 31, 32, 32, 33, 34, 34, 34, 102.$$

Como podemos perceber, tal valor é $33\ m^3$ e representa uma medida de centralidade mais fiel ao conjunto de dados.

2.4.3 Moda

A moda de um conjunto de dados é o valor que ocorre com maior frequência. A moda é representada por Mo.

Quando lidamos com valores não-agrupados, a moda é facilmente encontrada: basta fazer uso da definição, isto é, procurar o valor que mais se repete.

Exemplo 2.12: A moda da variável *frequência semanal* do Exemplo 2.1 é $Mo = 3$ dias.

É interessante observar que um conjunto de dados pode ter mais de uma moda, em tal caso dizemos que este conjunto de dado é *multimodal*. Em outros casos, um conjunto de dados pode apresentar uma distribuição que não possua nenhum valor predominante, assim dizemos que este conjunto de dados é *amodal*.

Cálculo Aproximado da Moda

Quando os dados estão agrupados a moda é calculada de forma aproximada, fazendo uso da classe com maior frequência. A classe que apresenta a maior frequência é denominada *classe modal*.

Existem algumas fórmulas que nos dão um padrão de como determinar a moda de um conjunto de dados, eis duas delas:

(a) **Moda bruta:**

$$Mo = \frac{l_i + L_i}{2} = s_i, \tag{2.10}$$

isto é, a moda é o ponto médio da classe modal;

(b) **Regra de Czuber:**

$$Mo = l_i + \frac{h_i(n_i - n_{i-1})}{(n_i - n_{i-1}) + (n_i - n_{i+1})} \tag{2.11}$$

onde

i é a classe modal;

l_i é o limite inferior da classe modal;

h_i é a amplitude da classe modal;

n_i é a frequência absoluta da classe modal;

n_{i-1} é a frequência absoluta da classe anterior à classe modal;

n_{i+1} é a frequência absoluta da classe posterior à classe modal.

Exemplo 2.13: Vamos calcular a moda da variável *Renda familiar mensal* (em salários mínimos) do Exemplo 2.3.

Utilizando a fórmula da moda bruta juntamente com a Tabela 2.5, temos que a moda da variável em questão é $Mo = 12,5$ salários mínimos. Por outro lado, utilizando a regra de Czuber e, mais uma vez, a Tabela 2.5, a moda será

$$Mo = 11 + \frac{3(7-5)}{(7-5)+(7-4)} = 11 + \frac{6}{5} = 12,2 \text{ salários mínimos.}$$

2.5 Medidas de Dispersão

Quando se fala sobre a dispersão de um conjunto de dados, estamos nos referindo a variabilidade destes dados. Em particular, as medidas de dispersão quantificam se os dados de uma determinada amostra estão ou não próximos uns dos outros. Se os dados estão próximos, há uma pequena variabilidade; se estão afastados, há uma grande variabilidade e se os dados forem todos iguais, a variabilidade é zero.

Nesta seção, apresentaremos as principais medidas de dispersão: *amplitude total, desvio médio, variância* e *desvio-padrão*. Com exceção à primeira, todas tem a média aritmética como ponto de partida.

2.5.1 Amplitude Total

A *amplitude total* (AT) de um conjunto de dados é a diferença entre o maior e o menor valor observado.

Pelo fato de a amplitude total não utilizar os valores intermediários ela perde informações de como os dados estão distribuídos e/ou concentrados. No entanto, esta medida é bem interessante quando o conjunto de dados é pequeno e como ela é de fácil cálculo e compreensão, é muito utilizada em algumas situações específicas como por exemplo no controle estatístico de processo (para maiores informações ver seção 14.2 de [9]).

2.5.2 Desvio Médio

A diferença entre cada valor observado (x_i), para $i = 1, ..., n$, e a média $(\overline{x})$ é chamada de *desvio* (d_i), isto é,

$$d_i = x_i - \overline{x}. \tag{2.12}$$

O desvio é utilizado quando queremos analisar a dispersão ou o grau de concentração dos valores em torno da média. Contudo, é fácil ver que, para qualquer conjunto de dados, a soma dos desvios é igual a zero, ou seja,

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (x_i - \overline{x}) = 0. \tag{2.13}$$

Como (2.13) ocorre, a soma dos desvios não é uma boa medida de dispersão. Uma saída para este impasse é tomar a soma do módulo de cada desvio, obtendo assim

$$\sum_{i=1}^{n} |d_i| = \sum_{i=1}^{n} |x_i - \overline{x}|. \tag{2.14}$$

Note, agora, que (2.14) zera se e somente se todos os dados forem iguais.

Em muitos casos o uso de (2.14) pode trazer dificuldades quando analisamos conjuntos de dados com números diferentes de observações. Dessa maneira, é comum exprimir (2.14) como uma média, ficando assim definida a medida de dispersão chamada *desvio médio* (DM)

$$DM = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}. \tag{2.15}$$

É fácil perceber que se no conjunto das n observações tivermos: n_1 iguais a x_1, n_2 iguais a $x_2, ..., n_k$ iguais a x_k, então (2.15) pode ser reescrita como

$$DM = \frac{\sum_{i=1}^{k} n_i |x_i - \overline{x}|}{n} = \sum_{i=1}^{k} f_i |x_i - \overline{x}|. \tag{2.16}$$

2.5.3 Variância

Uma outra medida de dispersão pode ser construída quando, ao invés de trabalhar com o módulo dos desvios, considerarmos o quadrado de cada desvio,

$$\sum_{i=1}^{n} (d_i)^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2. \tag{2.17}$$

De forma análoga ao que mencionamos no caso de (2.14), o uso de (2.17) pode trazer dificuldades quando analisamos conjuntos de dados com números diferentes de observações. Desse modo, mais uma vez, exprimiremos (2.17) em forma de média, ficando assim definida a medida de dispersão denominada *variância*:

$$\text{Variância} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}. \tag{2.18}$$

Como ocorreu no desvio médio, se no conjunto das n observações tivermos: n_1 iguais a x_1, n_2 iguais a $x_2, ..., n_k$ iguais a x_k, então podemos reescrever (2.18) da seguinte forma

$$\text{Variância} = \frac{\sum_{i=1}^{k} n_i (x_i - \overline{x})^2}{n} = \sum_{i=1}^{k} f_i (x_i - \overline{x})^2. \tag{2.19}$$

Para o cálculo da variância, podemos utilizar uma fórmula um pouco mais simples e usual. Para tanto, realizando as seguintes manipulações algébricas em (2.18), obtemos

$$\begin{aligned}
\frac{\sum (x_i - \overline{x})^2}{n} &= \frac{\sum (x_i^2 - 2x_i\overline{x} + \overline{x}^2)}{n} \\
&= \frac{\sum x_i^2}{n} - 2\overline{x}\frac{\sum x_i}{n} + \overline{x}^2\frac{\sum 1}{n} \\
&= \frac{\sum x_i^2}{n} - 2\overline{x}\overline{x} + \overline{x}^2\frac{n}{n} \\
&= \frac{\sum x_i^2}{n} - 2\overline{x}^2 + \overline{x}^2 = \frac{\sum x_i^2}{n} - \overline{x}^2
\end{aligned}$$

ou seja,

$$\text{Variância} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2. \tag{2.20}$$

Assim, (2.20) é uma expressão mais usual para o cálculo da variância, uma vez que precisamos somente das seguintes quantidades para determiná-la: n, $\sum x_i^2$ e $\overline{x}$. E no caso de observações repetidas,

$$\text{Variância} = \frac{\sum_{i=1}^{k} n_i x_i^2}{n} - \overline{x}^2. \tag{2.21}$$

Se os dados são uma *amostra* de uma população, a variância é representada por *Var*; por outro lado, se os dados são a *população* inteira, então representamos a variância por σ^2.

Na maioria dos casos é preferencial utilizarmos a variância em vez do desvio médio, isto porque além de eliminarmos o módulo, estamos potencializando os afastamentos, dando ênfase aos desvios em relação à média.

2.5.4 Desvio Padrão

Quando calculamos a variância, ela nos fornece um resultado que é dado em unidades quadráticas (por exemplo, se os dados são expressões em *cm*, a variância será expressa em cm^2), este fato pode conduzir a erros de interpretação. Para sanar esta "dificuldade" é comum utilizar o desvio padrão, que é definido como a raiz quadrada da variância.

$$\text{Desvio Padrão} = \sqrt{\text{Variância}}. \tag{2.22}$$

De modo análogo à variância, se os dados são uma *amostra*, o desvio padrão é representada por dp; por outro lado, se os dados são a *população* inteira, então representamos o desvio padrão por σ.

É interessante mencionar que tanto o desvio médio quanto o desvio padrão indicam, em média, qual será o erro cometido ao tentar substituir cada observação pelo valor da média.

Exemplo 2.14: Considere, novamente, a variável *frequência semanal* do Exemplo 2.1. Vamos calcular as medidas de dispersão para esta variável.

Como o valor máximo e mínimo dos dados são 7 e 1, respectivamente, a amplitude total será igual a

$$AT = 7 - 1 = 6 \text{ dias.}$$

Já foi calculado anteriormente que a frequência semanal média é de $3,2$ dias; assim, os desvios $x_i - \bar{x}$ são $-2,2;\ -1,2;\ -0,2;\ 0,8;\ 1,8;\ 2,8$ e $3,8$.

Com isso, utilizando a expressão (2.16), o desvio médio será igual a

$$\begin{aligned} DM &= \frac{3|-2,2| + 4|-1,2| + 6|-0,2| + 3|0,8| + 2|1,8| + 1|2,8| + 1|3,8|}{20} \\ &= 1,26 \text{ dias.} \end{aligned}$$

Por outro lado, utilizando (2.21), a variância será igual a

$$\begin{aligned} Var &= \frac{3(1)^2 + 4(2)^2 + 6(3)^2 + 3(4)^2 + 2(5)^2 + 1(6)^2 + 1(7)^2}{20} - (3,2)^2 \\ &= 2,56 \text{ dias.} \end{aligned}$$

Em consequência, o desvio padrão será igual a

$$dp = \sqrt{2,56} = 1,6 \text{ dias.}$$

Cálculo Aproximado da Variância

O cálculo aproximado da variância quando os dados estão agrupados em classes pode ser feito de modo análogo ao caso da média aritmética, ou seja, basta trocar em (2.21) cada observação (x_i) pelo respectivo ponto médio (s_i) da i-ésima classe.

Exemplo 2.15: Considere o caso da variável renda familiar mensal, do Exemplo 2.3. Já vimos que a renda familiar mensal média é de $12,35$ salários mínimos; assim, o valor aproximado da variância é igual a

$$\begin{aligned} Var &\approx \frac{2(6,5)^2 + 5(9,5)^2 + 7(12,5)^2 + 4(15,5)^2 + 2(18,5)^2}{20} - (12,35)^2 \\ &= 11,2275 \text{ (salários mínimos)}^2. \end{aligned}$$

Portanto, o desvio padrão aproximado é igual a

$$dp \approx \sqrt{11,2275} = 3,35 \text{ salários mínimos.}$$

Padronização de Variáveis

A padronização de uma variável é um recurso bastante utilizado e consiste em subtrair de todos os valores de uma variável a sua média e dividir o resultado pelo desvio padrão da respectiva variável. Desse modo, se X representa uma variável com média $\overline{x}$ e desvio padrão $dp(X)$, e se sobre cada x_i, para $i = 1, ..., n$, fizermos a mudança de variável

$$Z_x = \frac{x_i - \overline{x}}{dp(X)},$$

teremos a variável padronizada Z_x.

A nova variável Z_x terá as seguintes propriedades:

Propriedade 1: *Possui média zero, isto é,* $\overline{Z}_x = 0$.

Assim a padronização corresponde a deslocar o centro (dado pela média) de um conjunto de dados para a origem do sistema cartesiano.

Propriedade 2: *Desvio padrão passa a ser igual a 1, isto é,* $dp(Z_x) = 1$.

Dessa forma fica simplificado a comparação de conjuntos padronizados.

Capítulo 3

Correlação Linear

No Capítulo anterior nossa preocupação foi organizar e resumir informações relacionadas a uma única variável. Tendo isto em vista, aprendemos a construir tabelas de frequências, gráficos, calcular medidas de tendência central e dispersão. Contudo, é frequente estarmos interessados em analisar relações que podem existir entre duas ou mais variáveis. Neste caso, as técnicas que foram desenvolvidas são insuficientes para realizar este tipo de análise.

Considerando o cenário descrito acima, faz-se necessário o desenvolvimento de novas medidas que são capazes de realizar um estudo entre o relacionamento de duas ou mais variáveis. Neste trabalho, iremos nos ater às relações entre duas variáveis.

3.1 Tipos de Relação

Ao lidarmos com duas variáveis, quando queremos estudar o relacionamento entre elas, podemos ter três situações:

(a) as duas variáveis são qualitativas;

(b) as duas variáveis são quantitativas; e

(c) uma variável é qualitativa e a outra é quantitativa.

Nesses três casos, as técnicas de análise de dados a serem desenvolvidas são diferentes e cada uma tem suas próprias peculiaridades. Ficaremos restritos à análise de dados com duas variáveis quantitativas, pois seu desenvolvimento se alinha com o objeto de estudo deste trabalho.

3.1.1 Relações Determinísticas

É comum encontrarmos, em livros de matemática do Ensino Médio, relações entre duas variáveis que estão perfeitamente ligadas através de uma definição, e que podem ser

expressas por meio de uma sentença matemática. Alguns exemplos deste tipo de relação são:

- a área e o lado de um quadrado: $A = l^2$, onde A é a área e l é o lado;
- o comprimento e o raio de uma circunferência: $C = 2\pi r$, onde C é o comprimento e r é o raio;
- a soma dos ângulos internos de um polígono e o número de lados deste polígono: $S = 180 \cdot (n-2)$, aonde S é a soma dos ângulos internos e n é o número de lados.

Esses exemplos caracterizam as relações conhecidas como *relações determinísticas*, onde dado o valor para uma variável sabe-se dizer, exatamente, qual será o valor da outra.

3.1.2 Relações Aleatórias

Ao contrário das relações determinísticas, as relações aleatórias não estão ligadas através de uma definição bem estabelecida, em outras palavras, estas são bem menos precisas que aquelas, ou seja, dado um certo valor para uma variável não sabemos dizer qual será o valor da outra com exatidão. Exemplos deste tipo de relação são:

- o peso e a estatura de um grupo de pessoas;
- a altitude e a temperatura de uma região;
- a nota em matemática e em estatística de uma turma.

Veja, por exemplo, que no caso peso-estatura pode ocorrer variações que não somos capazes de predizer: estaturas diferentes podem corresponder a pesos iguais ou estaturas iguais podem corresponder a pesos diferentes. Todavia, em média, quanto maior for a estatura do indivíduo, maior será o seu peso.

Quando duas variáveis estão ligadas por uma relação aleatória, dizemos que existe **correlação** entre elas. De acordo com Triola [9] (2013, p.416) "existe uma correlação entre duas variáveis quando os valores de uma variável estão relacionados, de alguma forma, com os valores da outra variável".

Em linhas gerais, a correlação é o estudo que compreende a análise de *dados amostrais* para saber se e como duas variáveis quantitativas estão relacionadas entre si. Neste trabalho iremos estudar o caso da correlação *linear*, que é aquela que mede a força ou grau de relacionamento linear entre duas variáveis.

3.2 Diagrama de Dispersão

Como vimos anteriormente, se duas variáveis não estiverem relacionadas deterministicamente então, para um valor fixo da primeira variável, o valor da segunda variável será aleatório (como foi o caso do exemplo peso-estatura). Comumente, a variável cujo valor x for fixado será representada por X e denominada variável independente ou explicativa. Para um x fixo, a segunda variável será aleatória; representamos essa variável aleatória e seu valor observado por Y e y, respectivamente, e a chamamos de variável dependente ou resposta.

A distribuição conjunta destas variáveis pode ser organizada em tabelas de dupla entrada, formando um conjunto de n pares ordenados $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$; dessa forma é possível realizar um estudo sobre a associação entre elas. Além deste tipo de análise, também temos procedimentos analíticos e gráficos mais refinados.

Um destes gráficos mencionados acima, e o mais utilizado, chama-se *diagrama de dispersão*, ele nos permite verificar, visualmente, qual é o tipo de associação que existe entre duas variáveis. Segundo Triola [9] (2013, p.416)

> Antes de fazermos quaisquer análises estatísticas formais, devemos usar um diagrama de dispersão para explorar os dados visualmente. Podemos examinar o diagrama de dispersão em relação a quaisquer padrões distintos e em relação a valores atípicos, que são os pontos distantes dos demais pontos.

Este diagrama consiste em um sistema de eixos ortogonais: o eixo horizontal representa os valores da variável X (ou independente) e o eixo vertical representa os valores da variável Y (ou dependente). Cada (x_i, y_i) dos dados observados é representado graficamente como um ponto neste sistema.

Exemplo 3.1: Na Figura 3.1, temos o diagrama de dispersão das variáveis:

X : número de anos de serviço;

Y : número de clientes de agentes de uma companhia de seguros.

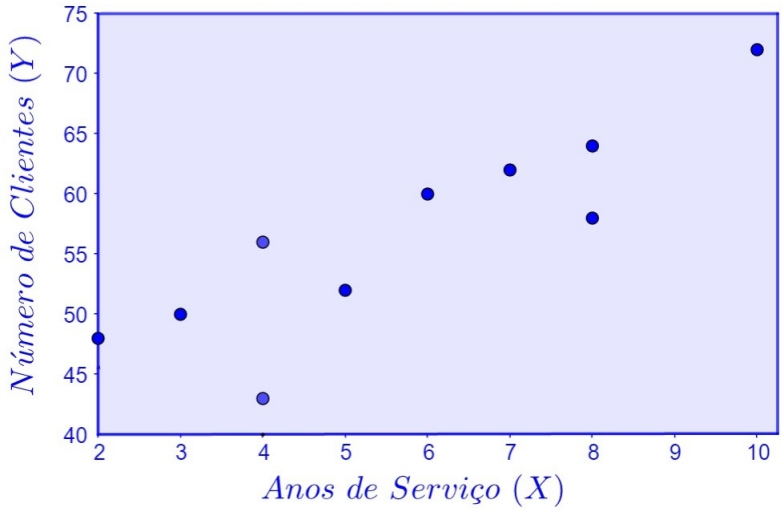
Os dados estão na Tabela 3.1. Veja que parece existir uma associação entre as variáveis, o diagrama mostra um padrão claro de reta, ou linear. Neste caso dizemos que há uma correlação *positiva* (ou *direta*) entre X e Y, pois à medida que os valores de X acrescem, os respectivos valores de Y também tendem a crescer.

Tabela 3.1: Número de anos de serviço (X) e número de clientes (Y) de agentes de uma companhia de seguros.

Agentes	Anos de serviços (X)	Número de clientes (Y)
A	2	48
B	3	50
C	4	56
D	5	52
E	4	43
F	6	60
G	7	62
H	8	58
I	8	64
J	10	72

Fonte: Bussab [3] (2017, p.86).

Figura 3.1: Diagrama de dispersão para as variáveis X: anos de serviço e Y: número de clientes.



Fonte: Produção do autor.

Exemplo 3.2: Na Figura 3.2, temos o diagrama de dispersão das variáveis:

X : renda bruta mensal;

Y : porcentagem da renda gasta em saúde,

obtidas a partir de uma pesquisa feita com dez famílias com renda bruta mensal entre 10 e 60 salários mínimos; os dados das variáveis se encontram na Tabela 3.2.

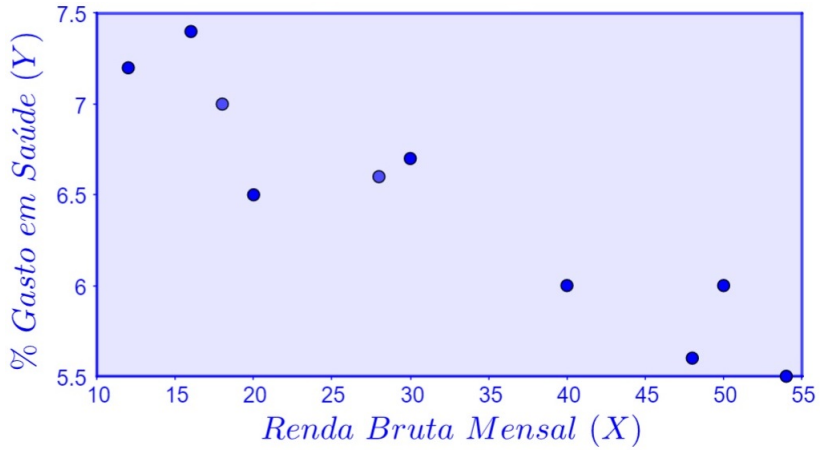
Mais uma vez o diagrama mostra um padrão linear claro. Contudo, este caso apresenta uma correlação *negativa* (ou *inversa*) entre X e Y, pois à medida que os valores de X acrescem, os respectivos valores de Y tendem a decrescer.

Tabela 3.2: Renda bruta mensal (X) e porcentagem da renda gasta em saúde (Y) para um conjunto de famílias.

Família	Renda bruta mensal (X)	% gasta em saúde (Y)
A	12	7,2
B	16	7,4
C	18	7,0
D	20	6,5
E	28	6,6
F	30	6,7
G	40	6,0
H	48	5,6
I	50	6,0
J	54	5,5

Fonte: Bussab [3] (2017, p.87).

Figura 3.2: Diagrama de dispersão para as variáveis X: renda bruta mensal e Y: % renda gasta com saúde.



Fonte: Produção do autor.

Exemplo 3.3: Oito indivíduos foram submetidos a um teste (máximo = 100 pontos) sobre conhecimento de língua estrangeira e, em seguida, mediu-se o tempo (em minutos) gasto para cada um aprender a operar uma determinada máquina. Com isso foram medidas as variáveis:

X : resultado obtido no teste;

Y : tempo necessário para aprender a operar a máquina.

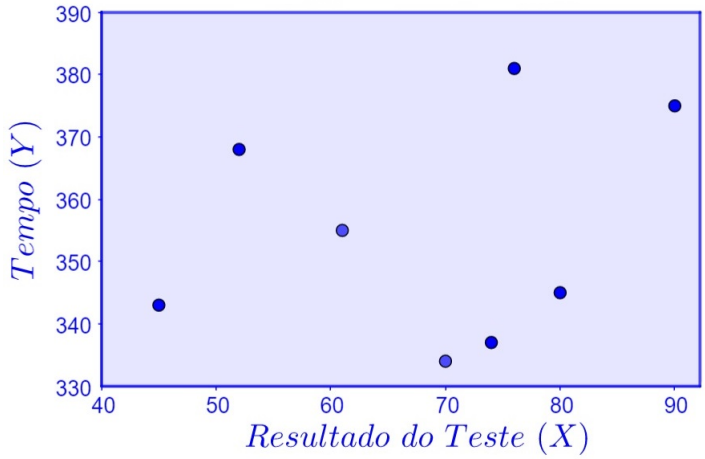
Os dados coletados encontram-se na Tabela 3.3. A Figura 3.3 mostra o diagrama de dispersão das variáveis X e Y. Diferente dos Exemplos 3.1 e 3.2, neste, não parece existir uma associação entre as variáveis, perceba que quando aumentamos os valores da variável X, não sabemos dizer o que pode ocorrer com os respectivos valores de Y. Neste caso, dizemos que *não existe* correlação entre as duas variáveis ou que a correlação é *nula*.

Tabela 3.3: Resultado do teste (X) e tempo de operação de máquina (Y) para oito indivíduos.

Indivíduo	Resultado do teste (X)	Tempo (Y)
A	45	343
B	52	368
C	61	355
D	70	334
E	74	337
F	76	381
G	80	345
H	90	375

Fonte: Bussab [3] (2017, p.88).

Figura 3.3: Diagrama de dispersão para as variáveis X: resultado do teste e Y: tempo de operação.



Fonte: Produção do autor.

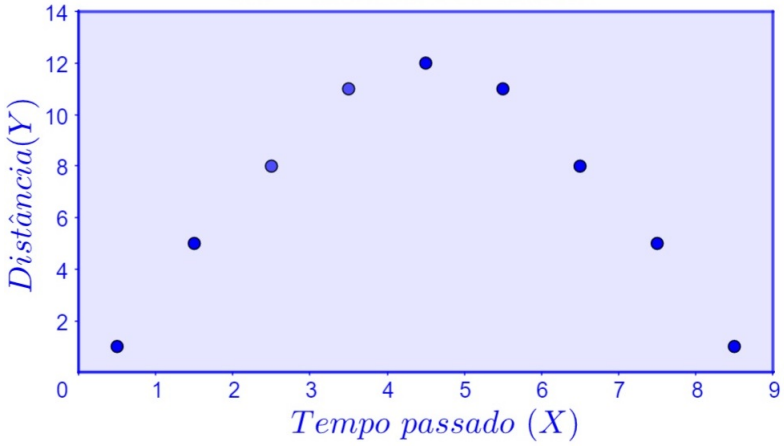
Exemplo 3.4: Um objeto foi lançado para cima e mediu-se as seguintes variáveis:

X : tempo passado após o lançamento do objeto;

Y : distância do objeto em relação ao chão.

A Figura 3.4 mostrar o diagrama de dispersão para as variáveis X e Y. Note que os pontos apresentados no diagrama têm como "imagem" uma curva, sugerindo uma correlação *não-linear* entre X e Y.

Figura 3.4: Diagrama de dispersão para as variáveis X: tempo passado e Y: distância em relação ao solo.



Fonte: Produção do autor.

3.3 Coeficiente de Correlação Linear

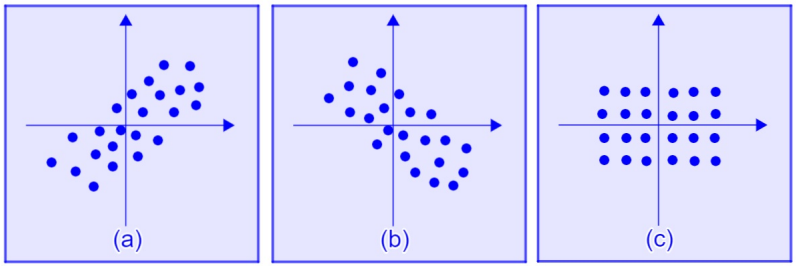
Com o auxílio do diagrama de dispersão fomos capazes de analisar e compreender o comportamento conjunto de duas variáveis, verificando a presença ou ausência de associação entre elas, e, caso havendo associação, definimos o seu tipo: correlação linear (positiva/negativa) ou correlação não-linear.

No entanto, é necessário o desenvolvimento de uma medida que possa quantificar esta associação. Neste trabalho, iremos nos concentrar na associação envolvendo o caso da correlação linear, isto é, vamos definir uma medida de associação chamada *coeficiente de correlação linear*, que é útil para medir a força da correlação linear entre valores de duas variáveis X e Y em uma *amostra*. De acordo com Bussab [3] (2017), essa medida de associação fornece a proximidade dos dados a uma reta.

Para tanto, consideremos um diagrama de dispersão como o da Figura 3.5 (a) no qual a origem do sistema de eixos ortogonais foi colocada no centro da nuvem de pontos, através de uma transformação conveniente. Perceba que os dados possuem uma correlação linear

positiva (ou direta) e que a maioria dos pontos se encontram situados no primeiro e terceiro quadrantes. No primeiro quadrante as coordenadas dos pontos são sempre positivas e no terceiro as coordenadas são sempre negativas; assim, o produto das coordenadas de qualquer desses dois quadrantes sempre será positivo. Ao somar o produto das coordenadas dos pontos obteremos um número positivo, tendo em vista a existência de mais produtos positivos do que negativos.

Figura 3.5: Tipos de correlação entre duas variáveis



Fonte: Bussab [3] (2017, p.89) (Adaptada).

Por outro lado, um diagrama de dispersão como o da Figura 3.5 (b) apresenta um tipo de correlação linear negativa (ou inversa) e, realizando de forma análoga os passos feito acima, a soma dos produtos das coordenadas será negativa.

Por fim, um diagrama como o da Figura 3.5 (c) não apresenta nenhum tipo de correlação linear entre as variáveis. Note que os pontos estão igualmente dispersos entre os quatros quadrantes; portanto, para cada produto das coordenadas que resultar em um número positivo, teremos um resultado negativo simétrico. Somando-se os produtos das coordenadas dos pontos, o resultado será igual a zero. Outra variante para esta situação é quando os pontos no diagrama de dispersão estão próximos de um formato circular, neste caso, a soma dos produtos será aproximadamente zero.

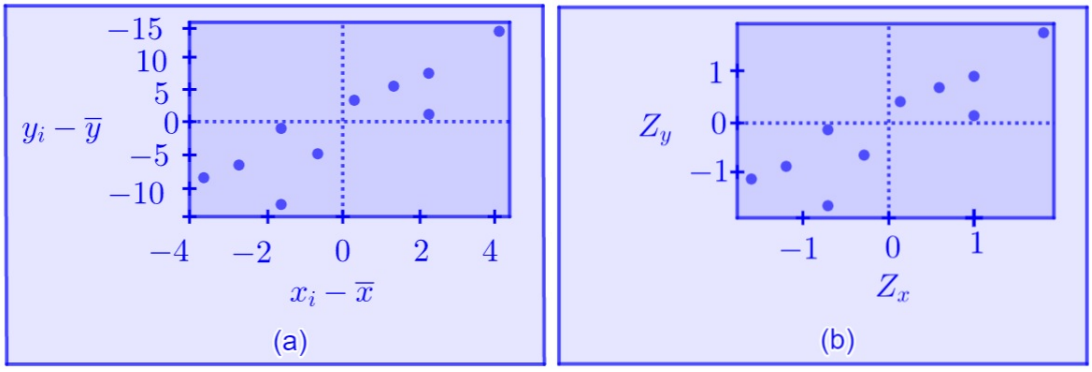
Agora, com o auxílio das três situações descritas acima, temos condições de desenvolver o coeficiente de correlação linear para, finalmente, defini-lo formalmente.

3.3.1 Cálculo do Coeficiente de Correlação Linear

Consideremos os dados da Tabela 3.1. Pelo que vimos anteriormente, nosso primeiro objetivo é colocar a origem do sistema de eixos ortogonais no centro da nuvem de pontos. Parece ser bem razoável deslocarmos a origem para o ponto de coordenadas $\overline{x} = 5,7$ e $\overline{y} = 56,5$, nesta ordem, que são as médias dos valores das variáveis X e Y, respectivamente. A partir desta transformação, uma observação x_i da variável X (resp. y_i da variável Y) passa a ter abscissa $x_i - \overline{x}$ (resp. ordenada $y_i - \overline{y}$) no novo diagrama de dispersão, veja a Figura 3.6 (a). Assim, os valores das novas coordenadas $(x_i - \overline{x}, y_i - \overline{y})$ estão mostradas na quarta e quinta coluna da Tabela 3.4.

Analisando o diagrama de dispersão da Figura 3.6 (a) percebemos que existe uma diferença de escalas entre os eixos, isto decorre do fato de que a variabilidade da variável Y é muito maior se comparada com a da variável X. Podemos constatar isso através do desvio padrão de ambas as variáveis, onde $dp(Y) = 8,11$ e $dp(X) = 2,41$. Para que possamos reduzir os eixos a uma mesma escala, dividiremos cada um dos desvios que estão presentes na quarta e quinta coluna da Tabela 3.4 pelo seu respectivo desvio padrão. Os novos valores (Z_x e Z_y) obtidos encontram-se na sexta e sétima coluna da Tabela 3.4, eles são, respectivamente, os *valores padronizados* das variáveis X e Y, e o novo diagrama de dispersão com a escala dos eixos ajustada está presente na Figura 3.6 (b).

Figura 3.6: Mudança de origem e escala



Fonte: Bussab [3] (2017, p.91) (Adaptada).

Finalmente, na oitava coluna, está presente os produtos das coordenadas padronizadas e sua soma, $8,769$, que, como esperado, é positiva. Para concluirmos a definição dessa medida de associação mencionemos a seguinte observação feita por Bussab [3] (2017, p. 90)

> A soma dos produtos das coordenadas depende, e muito, do número de pontos. Considere o caso de associação positiva: a soma acima tende a aumentar com o número de pares (x,y) e ficaria difícil comparar essa medida para dois conjuntos com números diferentes de pontos. Por isso, costuma-se usar a média da soma dos produtos das coordenadas.

Portanto, calculando a média dos produtos das coordenadas padronizadas, obtemos $r = 8,769/10 = 0,877$; é comum representarmos a correlação entre duas variáveis X e Y pela letra r. Assim, o grau de associação linear entre as variáveis em questão está quantificado em $0,877$.

Tabela 3.4: Cálculo do coeficiente de correlação linear.

Agentes	Anos (X)	Clientes (Y)	$x_i - \overline{x}$	$y_i - \overline{y}$	$\frac{x_i-\overline{x}}{dp(X)} = Z_x$	$\frac{y_i-\overline{y}}{dp(Y)} = Z_y$	$Z_x \cdot Z_y$
A	2	48	$-3,7$	$-8,5$	$-1,54$	$-1,05$	$1,617$
B	3	50	$-2,7$	$-6,5$	$-1,12$	$-0,80$	$0,846$
C	4	56	$-1,7$	$-0,5$	$-0,71$	$-0,06$	$0,043$
D	5	52	$-0,7$	$-4,5$	$-0,29$	$-0,55$	$0,160$
E	4	43	$-1,7$	$-13,5$	$-0,71$	$-1,66$	$1,179$
F	6	60	$0,3$	$3,5$	$0,12$	$0,43$	$0,052$
G	7	62	$1,3$	$5,5$	$0,54$	$0,68$	$0,367$
H	8	58	$2,3$	$1,5$	$0,95$	$0,19$	$0,181$
I	8	64	$2,3$	$7,5$	$0,95$	$0,92$	$0,874$
J	10	72	$4,3$	$15,5$	$1,78$	$1,91$	$3,400$
Total	57	565	0	0	–	–	$8,769$

Fonte: Bussab [3] (2017, p.90).

Com base no que foi exposto até o momento, temos condições de definir formalmente o coeficiente de correlação linear.

Definição. *Sendo (x_i, y_i) as observações individuais de cada elemento de uma amostra de tamanho n das variáveis X e Y, chamaremos de coeficiente de correlação linear amostral entre essas variáveis o valor*

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{dp(X)}\right)\left(\frac{y_i-\overline{y}}{dp(Y)}\right). \tag{3.1}$$

Em outras palavras, esta medida de associação é dada pela média dos produtos dos valores padronizados das variáveis.

Para efeitos de cálculos, (3.1) não é uma expressão muito usual, pois o seu desenvolvimento se torna trabalhoso para um número elevado de observações. Pensando nisso, podemos utilizar as expressões (2.6), (2.20) e (2.22) e realizar algumas manipulações algébricas em (3.1), com efeito

$$\begin{aligned}
\frac{1}{n}\sum\left(\frac{x_i-\overline{x}}{dp(X)}\right)\left(\frac{y_i-\overline{y}}{dp(Y)}\right) &= \frac{1}{n}\sum\left(\frac{x_i-\overline{x}}{\sqrt{\frac{\sum x_i^2}{n}-\overline{x}^2}}\right)\left(\frac{y_i-\overline{y}}{\sqrt{\frac{\sum y_i^2}{n}-\overline{y}^2}}\right) \\
&= \frac{1}{n}\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\frac{\sum x_i^2-n\overline{x}^2}{n}}\sqrt{\frac{\sum y_i^2-n\overline{y}^2}{n}}} \\
&= \frac{1}{n}\frac{\sum(x_iy_i-x_i\overline{y}-\overline{x}y_i+\overline{x}\overline{y})}{\sqrt{\left(\frac{\sum x_i^2-n\overline{x}^2}{n}\right)\left(\frac{\sum y_i^2-n\overline{y}^2}{n}\right)}}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\cancel{n}} \frac{\sum x_i y_i - \bar{y}\sum x_i - \bar{x}\sum y_i + \bar{x}\bar{y}\sum 1}{\frac{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}}{\cancel{n}}} \\
&= \frac{\sum x_i y_i - \bar{y}n\bar{x} - \bar{x}n\bar{y} + \bar{x}\bar{y}n}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}} \\
&= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}}
\end{aligned}
$$

isto é,

$$
r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}. \tag{3.2}
$$

Perceba que (3.2) é uma expressão que pode ser operacionalizada de modo mais prático, tendo em vista que precisamos apenas das seguintes quantidades para determinar o coeficiente de correlação linear: n, $\bar{x}$, $\bar{y}$, $\sum x_i y_i$, $\sum x_i^2$ e $\sum y_i^2$.

Exemplo 3.5: Vamos calcular o coeficiente de correlação linear para as variáveis: *renda bruta mensal* (X) e *porcentagem da renda gasta em saúde* (Y) do Exemplo 3.2. Inicialmente, tomemos como base os dados da Tabela 3.2 para a construção da Tabela 3.5 abaixo.

Tabela 3.5: Cálculo do coeficiente de correlação linear para as variáveis renda bruta mensal (X) e porcentagem da renda gasta em saúde (Y).

Família	Renda (X)	% (Y)	$x_i y_i$	x_i^2	y_i^2
A	12	7,2	86,4	144	51,84
B	16	7,4	118,4	256	54,76
C	18	7,0	126,0	324	49,00
D	20	6,5	130,0	400	42,25
E	28	6,6	184,8	784	43,56
F	30	6,7	201,0	900	44,89
G	40	6,0	240,0	1.600	36,00
H	48	5,6	268,8	2.304	31,36
I	50	6,0	300,0	2.500	36,00
J	54	5,5	297,0	2.916	30,25
Total	316	64,5	1952,4	12.128	419,91

Fonte: Tabela 3.2.

Da tabela acima temos:

- $n = 10$;
- $\overline{x} = \frac{316}{10} = 31,6$;
- $\overline{y} = \frac{64,5}{10} = 6,45$;
- $\sum x_i y_i = 1952,4$;
- $\sum x_i^2 = 12.128$;
- $\sum y_i^2 = 419,91$.

Por fim, utilizando (3.2), obtemos

$$r = \frac{19,52,4 - 10(31,6)(6,45)}{\sqrt{(12.128 - 10(31,6)^2)(419,91 - 10(6,45)^2)}} = -0,94.$$

Assim, o grau de associação linear entre as variáveis está quantificado em $-0,94$, que é, como esperado, uma correlação negativa.

3.3.2 Propriedades do Coeficiente de Correlação Linear

Anunciamos a seguir algumas propriedades que o coeficiente de correlação linear possui. Essas propriedades nos ajudam a compreender e interpretar melhor tal coeficiente.

Propriedade 1: *O valor de r está sempre entre* -1 *e* 1*, inclusive. Isto é,*

$$-1 \leqslant r \leqslant 1.$$

Prova: Inicialmente, veja que utilizando as expressões (2.18) e (2.22) para o cálculo do desvio padrão, podemos reescrever (3.1) da seguinte forma

$$\begin{aligned} r &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{dp(X)}\right)\left(\frac{y_i - \overline{y}}{dp(Y)}\right) \\ &= \frac{\frac{1}{n}\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{\sum(x_i-\overline{x})^2}{n}}\sqrt{\frac{\sum(y_i-\overline{y})^2}{n}}}, \end{aligned}$$

isto nos dá

$$r^2 = \frac{\left(\frac{1}{n}\sum(x_i - \overline{x})(y_i - \overline{y})\right)^2}{\left(\frac{\sum(x_i-\overline{x})^2}{n}\right)\left(\frac{\sum(y_i-\overline{y})^2}{n}\right)}. \tag{3.3}$$

Agora, considere a seguinte função na variável real t

$$f(t) = \frac{1}{n}\sum_{i=1}^{n}(A+tB)^2,$$

onde $A = x_i - \overline{x}$ e $B = y_i - \overline{y}$.

Visto que $(A+tB)^2 \geqslant 0$, temos que $f(t) \geqslant 0$ para todo t. Desenvolvendo, obtemos

$$\begin{aligned} f(t) &= \frac{1}{n}\sum(A^2 + 2tAB + t^2B^2) \\ &= \frac{1}{n}\sum A^2 + 2t\frac{1}{n}\sum AB + t^2\frac{1}{n}\sum B^2. \end{aligned}$$

Desse modo, $f(t)$ é uma expressão quadrática em t. Em geral, se uma expressão quadrática $g(t) = at^2 + bt + c$ tem a propriedade de que $g(t) \geqslant 0$ para todo t, isto significa que seu discriminante $b^2 - 4ac$ deve ser $\leqslant 0$. Aplicando essa conclusão à função $f(t)$, com $a = \frac{1}{n}\sum B^2$, $b = 2\frac{1}{n}\sum AB$ e $c = \frac{1}{n}\sum A^2$, obtemos

$$4\left(\frac{1}{n}\sum AB\right)^2 - 4\left(\frac{1}{n}\sum B^2\right)\left(\frac{1}{n}\sum A^2\right) \leqslant 0.$$

Isto nos fornece,

$$\frac{\left(\frac{1}{n}\sum AB\right)^2}{\left(\frac{1}{n}\sum A^2\right)\left(\frac{1}{n}\sum B^2\right)} \leqslant 1 \Longleftrightarrow \frac{\left(\frac{1}{n}\sum(x_i-\overline{x})(y_i-\overline{y})\right)^2}{\left(\frac{\sum(x_i-\overline{x})^2}{n}\right)\left(\frac{\sum(y_i-\overline{y})^2}{n}\right)} \leqslant 1. \tag{3.4}$$

A partir de (3.3) e (3.4) chegamos a conclusão de que $r^2 \leqslant 1$, ou ainda, $-1 \leqslant r \leqslant 1$, que é o resultado desejado.

Propriedade 2: *O valor de r não muda se todos os valores de qualquer das variáveis forem convertidos para uma escala diferente.*

Prova: Mudar a escala de todos os valores de alguma das variáveis significa multiplicar cada um deles por uma certa constante $k > 0$. Suponha que alteramos a escala da variável X; assim, cada observação x_i passa a ser kx_i e a média passa a ser $k\overline{x}$. Com isso,

$$\begin{aligned} r &= \frac{\sum(kx_i)y_i - n(k\overline{x})\overline{y}}{\sqrt{(\sum(kx_i)^2 - n(k\overline{x})^2)\left(\sum y_i^2 - n\overline{y}^2\right)}} \\ &= \frac{\sum kx_iy_i - nk\overline{x}\overline{y}}{\sqrt{\left(\sum k^2x_i^2 - nk^2\overline{x}^2\right)\left(\sum y_i^2 - n\overline{y}^2\right)}} \\ &= \frac{\cancel{k}(\sum x_iy_i - n\overline{x}\overline{y})}{\cancel{k}\sqrt{\left(\sum x_i^2 - n\overline{x}^2\right)\left(\sum y_i^2 - n\overline{y}^2\right)}} \\ &= \frac{\sum x_iy_i - n\overline{x}\overline{y}}{\sqrt{\left(\sum x_i^2 - n\overline{x}^2\right)\left(\sum y_i^2 - n\overline{y}^2\right)}}, \end{aligned}$$

que é exatamente o que afirma a propriedade.

Propriedade 3: *O valor de r não é afetado pela escolha de x_i e y_i.*

Em outras palavras, se todos os valores de x_i forem trocados pelos respectivos valores de y_i e vice-versa, então o valor de r não se altera. Esta propriedade é justificada pela comutatividade da multiplicação de números reais.

Propriedade 4: *r mede a intensidade de uma relação linear.*

Essa medida de associação não é eficiente para medir intensidade de uma relação que não seja linear; por exemplo, a relação presente na Figura 3.4.

Propriedade 5: *r é muito sensível a valores atípicos, no sentido de que um único valor atípico pode afetar drasticamente seu valor.*

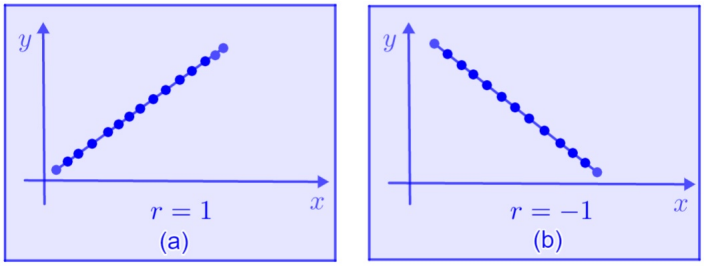
Essa propriedade é justificada pelo fato de o coeficiente de correlação linear ser uma medida de associação que se baseia em médias, e como vimos na subseção 2.4.1 a média é uma medida sensível a valores atípicos.

3.3.3 Interpretação do Coeficiente de Correlação Linear

Nesta subseção vamos fazer um cruzamento entre a discursão feita, na seção 3.2, a respeito do diagrama de dispersão, e os possíveis valores que o coeficiente de correlação linear pode assumir.

Quando o valor de r é exatamente igual a 1, dizemos que existe uma correlação *positiva perfeita* entre as variáveis X e Y. Graficamente, isto significa que todos pontos no diagrama de dispersão estão alinhados sobre uma mesma reta ascendente, veja a Figura 3.7 (a). Por outro lado, quando o valor de r é exatamente igual a -1, dizemos que há uma correlação *negativa perfeita* entre as variáveis e os pontos no diagrama de dispersão estão alinhados sobre uma mesma reta descendente, como na Figura 3.7 (b).

Figura 3.7: Coeficiente de correlação linear: $r = 1$ e $r = -1$

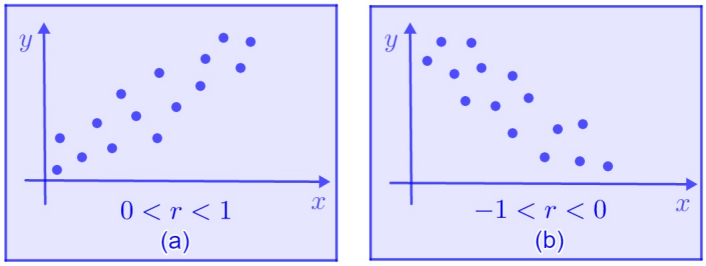


Fonte: Produção do autor.

Ao passo que $0 < r < 1$, a correlação é dita positiva e pode variar entre forte, moderada ou fraca, de acordo com o valor de r estar mais próximo de 1 ou não, isto é, quanto mais próximo de 1 estiver o valor de r, mais forte será a correlação, e quanto mais afastado de 1 estiver o valor de r, mais fraca será a correlação. De modo análogo, se $-1 < r < 0$, a correlação é dita negativa e pode variar entre forte, moderada ou fraca, de acordo com o valor de r estar próximo de -1 ou não.

Em relação a interpretação gráfica, tanto para $0 < r < 1$ quanto para $-1 < r < 0$, existe um certo distanciamento entre os pontos no diagrama de dispersão, a Figura 3.8 (a) e (b) ilustra esses dois casos, respectivamente.

Figura 3.8: Coeficiente de correlação linear: $0 < r < 1$ e $-1 < r < 0$



Fonte: Produção do autor.

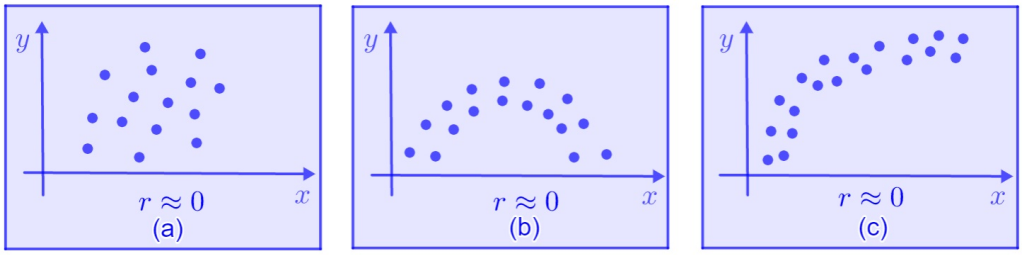
Para as variáveis *anos de serviço* (X) e *número de clientes* (Y), do Exemplo 3.1, o valor do coeficiente de correlação linear foi de $0,877$. Podemos interpretar essa informação dizendo que os anos de serviço de agentes de uma companhia de seguros tem uma forte correlação positiva (note que $0,877$ está próximo de 1) com o número de clientes que ele possui, em outros termos, quanto mais anos de serviço o agente tiver, mais clientes ele tenderá a ter. Perceba também que o diagrama de dispersão da Figura 3.1 é semelhante ao da Figura 3.8 (a).

Consideremos, agora, as variáveis *renda bruta mensal* (X) e *porcentagem da renda gasta em saúde* (Y) do Exemplo 3.2. O valor do coeficiente de correlação linear para essas variáveis foi calculado no Exemplo 3.5 e é igual a $-0,94$. Isto nos diz que a renda bruta mensal tem uma forte correlação negativa (veja que $-0,94$ está próximo de -1) com a porcentagem da renda gasta em saúde, ou seja, quanto maior for a renda bruta mensal, menor, tenderá a ser, a porcentagem desta renda gasta em saúde. Note a semelhança entre os diagramas de dispersão das Figuras 3.2 e 3.8 (b).

Quando o valor de $r \approx 0$, dizemos que *não existe* correlação entre as variáveis X e Y. Temos que ter cuidado ao interpretar essa situação, pois quando afirmamos que não existe correlação quando $r \approx 0$, estamos nos referindo à linear. Em suma, quando $r \approx 0$ pode ocorrer dois casos: não existe nenhum tipo de correlação entre as variáveis ou existe uma correlação entre as variáveis, contudo, está correlação é não-linear.

Graficamente, o primeiro caso, significa que os pontos no diagrama de dispersão ficam dispersos entre si, sem ter um padrão claro a ser seguido; no segundo caso, os pontos seguem uma curva plana (esta curva pode ser modelada através de uma função polinomial, exponencial, logarítmica, etc.). A Figura 3.9 mostra alguns casos particulares.

Figura 3.9: Coeficiente de correlação linear: $r \approx 0$



Fonte: Produção do autor.

Utilizando (3.2) e calculando o coeficiente de correlação linear para as variáveis *resultado do teste* (X) e *tempo de operação de máquina* (Y) do Exemplo 3.3, obtemos $r = 0,238$. Este valor nos diz que o resultado obtido no teste de língua estrangeira não possui correlação (note que $0,238$ está próximo de 0) com o tempo necessário para aprender a operar a máquina. Esta última afirmação fica evidente quando analisamos o diagrama de dispersão na Figura 3.3. Veja que à medida que o resultado obtido no teste aumenta, não se sabe, nem mesmo em média, qual seria o tempo necessário para aprender a operar a máquina.

Erro de Interpretação Envolvendo Correlação

O *New York Pizza Connection*, ou Princípio da pizza, é uma "lei econômica" bem-humorada, mas geralmente historicamente precisa, proposta pelo nova-iorquino Eric M. Bram, que observou em 1980 que, desde o início dos anos 60, o preço de uma fatia de pizza correspondia, com uma precisão incrível, o custo de uma viagem de metrô na cidade de Nova York.

O termo "*Pizza Connection*" referente a esse fenômeno foi cunhado no início de 2002 pelo colunista do "New York Times" Clyde Haberman, quando em seu artigo "*Will Subway Fares Rise? Check at Your Pizza Place*" (A tarifa do metrô aumentará? Verifique na sua Pizzaria), o mesmo escreveu que, na cidade de New York, a tarifa do metrô e o custo da fatia de pizza "tinham andado paralelamente por décadas".

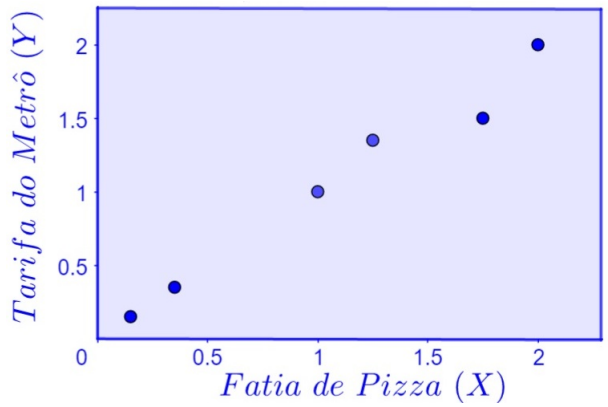
Uma amostra aleatória das variáveis *custo (em dólares) de uma fatia de pizza* (X) e a *Tarifa do Metrô* (Y), coletada na segunda metade do século 20 e início do século 21, está listada na Tabela 3.6. A Figura 3.10 mostra o diagrama de dispersão para essas variáveis.

Tabela 3.6: Custo de uma Fatia de Pizza (X) e Tarifa do Metrô (Y).

Ano	Custo da Pizza (X)	Tarifa do Metrô (Y)
1960	0,15	0,15
1973	0,35	0,35
1986	1,00	1,00
1995	1,25	1,35
2002	1,75	1,50
2003	2,00	2,00

Fonte: Triola [9] (2013, p. 415).

Figura 3.10: Diagrama de dispersão para as variáveis X: fatia de pizza e Y: tarifa do metrô.



Fonte: Produção do autor.

Vamos realizar uma análise para esse conjunto de dados para verificar se existe ou não correlação entre as variáveis em questão.

Inicialmente, podemos perceber, pela Tabela 3.6, que os pares de valores Pizza/Tarifa são praticamente os mesmos. Veja que o diagrama de dispersão sugere que há uma correlação positiva entre o custo da Fatia de Pizza e o custo da Tarifa do Metrô. Agora, calculando o coeficiente de correlação linear para esse conjunto de dados, obtemos $r = 0,987$, o que caracteriza uma forte correlação positiva entre as variáveis, como esperávamos.

A partir da análise feita acima, podemos concluir que há uma correlação entre o custo da fatia de pizza e a tarifa de metrô. Entretanto, não podemos concluir que um aumento no custo da fatia de pizza cause um aumento na tarifa do metrô, isto porque a correlação não implica causalidade. O que pode ter ocorrido no nosso exemplo é ambos os custos serem afetados por variáveis ocultas, "uma **variável oculta** é uma variável que afeta as variáveis em estudo, mas que não está incluída nele" (TRIOLA [9], 2013. p.423).

Um erro bastante comum na interpretação de resultados que envolvem correlação é pensar que ela implica causalidade, mas como vimos, isso não é verdade.

Capítulo 4

Regressão Linear Simples

4.1 Conceitos Básicos

No Capítulo anterior, apresentamos métodos que nos auxiliaram a explorar a presença ou ausência de relação linear entre duas variáveis emparelhadas, e a quantificar a força dessa relação através do Coeficiente de Correlação Linear amostral r. Agora, vamos *explicitar algebricamente* a forma dessa relação através do **Modelo de Regressão Linear Simples**. Não diferente da Correlação Linear, a Regressão Linear Simples também trabalha com um conjunto de dados amostrais.

O Modelo de Regressão Linear Simples (ou simplesmente o Modelo Linear) nos fornece uma equação, chamada de *equação de regressão*, que descreve o comportamento de uma variável em função do comportamento da outra. O gráfico da equação de regressão é uma reta, chamada de *reta de regressão*, que melhor se ajusta ao conjunto de dados amostrais emparelhados das variáveis em estudo.

Quando lidamos com um modelo, sob o ponto de vista da estatística, devemos ter em mente que as relações entre as variáveis quase nunca são exatas, determinísticas (como foi visto na seção 3.1). Elas, em geral, incluem flutuações aleatórias. Logo, qualquer modelo estatístico é constituído por duas componentes:

Modelo = Componente Sistemática + Componente Aleatória

Com este fato em mente, temos condições de definir o Modelo Linear.

Definição. *Sendo (x_i, y_i), onde $i = 1, ..., n$, as observações individuais de cada elemento de uma amostra de tamanho n, a equação de regressão é dada por*

$$y_i = ax_i + b + e_i, \quad i = 1, ..., n \tag{4.1}$$

onde a e b são os parâmetros do modelo e e_i representa a componente aleatória do modelo.

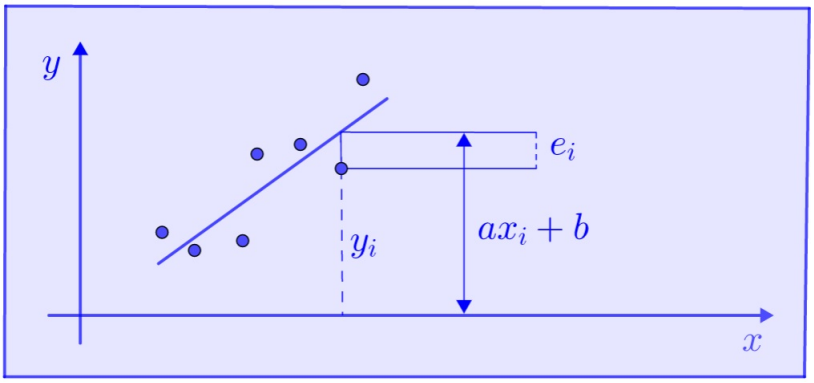
A componente sistemática do modelo é a média $ax_i + b$, referente à variável Y, que será representada por

$$\mu(x_i) = ax_i + b. \tag{4.2}$$

É importante destacar que a componente sistemática pode ser entendida como uma média devido as suposições que são estabelecidas a cerca das variáveis X e Y na subseção 4.2.1.

A componente aleatória e_i, é conhecida como o erro que se comete ao tentar modelar a relação entre as variáveis X e Y. A Figura 4.1 ilustra a definição acima.

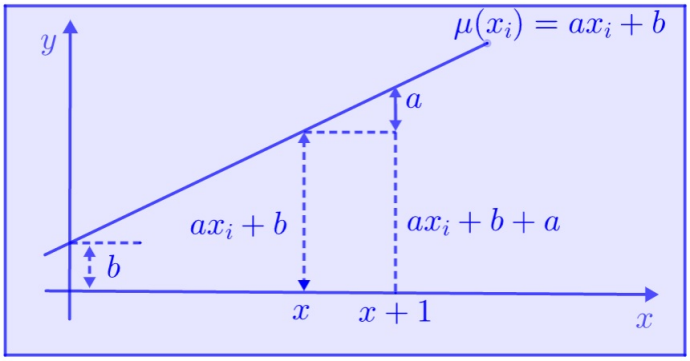
Figura 4.1: Modelo de Regressão Linear Simples



Fonte: Produção do autor.

No modelo (4.2), o parâmetro b é o intercepto, isto é, representa o ponto onde a reta de regressão corta o eixo das ordenadas; o parâmetro a, é o coeficiente angular da reta, ele representa o quanto Y varia em média para um aumento de uma unidade da variável X. Esses parâmetros estão representados na Figura 4.2.

Figura 4.2: Representação dos Parâmetros a e b.



Fonte: Bussab [3] (2017, p.464) (Adaptada).

O modelo (4.2) é chamado linear, pois ele representa uma reta. Todavia, Bussab [3] (2017, p.451) observa que em casos mais gerais, o termo linear refere-se ao modo como os parâmetros entram no modelo, ou seja, de forma linear. Um exemplo é o modelo $\mu(x_i) = ax_i^2 + bx_i + c$, que embora represente uma parábola graficamente, é linear em a, b e c. Em contra partida, $\mu(x_i) = ae^{bx_i}$ não é um modelo linear em a e b.

4.2 Estimação dos Parâmetros

O objetivo desta seção é utilizar dados amostrais emparelhados para estimar a equação de regressão. É notório que dispondo apenas dos dados amostrais, não podemos achar os valores exatos dos parâmetros a e b, mas com esses mesmo dados, podemos estimá-los.

4.2.1 Suposições para as Variáveis X e Y

Antes de estimar, de fato, os parâmetros a e b, se faz necessário estabelecer algumas suposições acerca das variáveis X e Y envolvidas. Vejamos quais são essas suposições:

1ª Suposição: A variável X é por hipótese controlada e não está sujeita a variações aleatórias.

2ª Suposição: Para dado valor x_i de X, os erros e_i distribuem-se ao redor da média $ax_i + b$ com média zero, isto é, se e representa a variável erro, então

$$\overline{e} = 0. \tag{4.3}$$

3ª Suposição: Os erros tenham a mesma variabilidade em torno dos níveis de X, ou seja,

$$Var(e) = \sigma^2. \tag{4.4}$$

Em outras palavras, a 2ª e 3ª suposição nos dizem que a média e a variância de e não dependem do valor de x_i.

4ª Suposição: Os erros e_i sejam não-correlacionados, ou seja, $cor(e_i, e_j) = 0, \forall\, i \neq j$.

Essas suposições são feitas com base em justificativas formais. Como tais justificativas não implicam em ganho substancial para o propósito destas notas, não as apresentaremos aqui.

4.2.2 Método dos Mínimos Quadrados (MMQ)

Um dos métodos mais utilizados para a estimação de parâmetros é o Método dos Mínimos Quadrados (MMQ). Isto se dá pelo fato do MMQ exigir uma série de suposições mínimas (as apresentadas na subseção anterior) para ser aplicado, enquanto outros métodos exigem, além das citadas, outras suposições; um exemplo disto é o Método da Máxima Verossimilhança que exige que os erros possuam distribuição normal (o leitor interessado no assunto pode consultar [7] (MEYER, 1983).

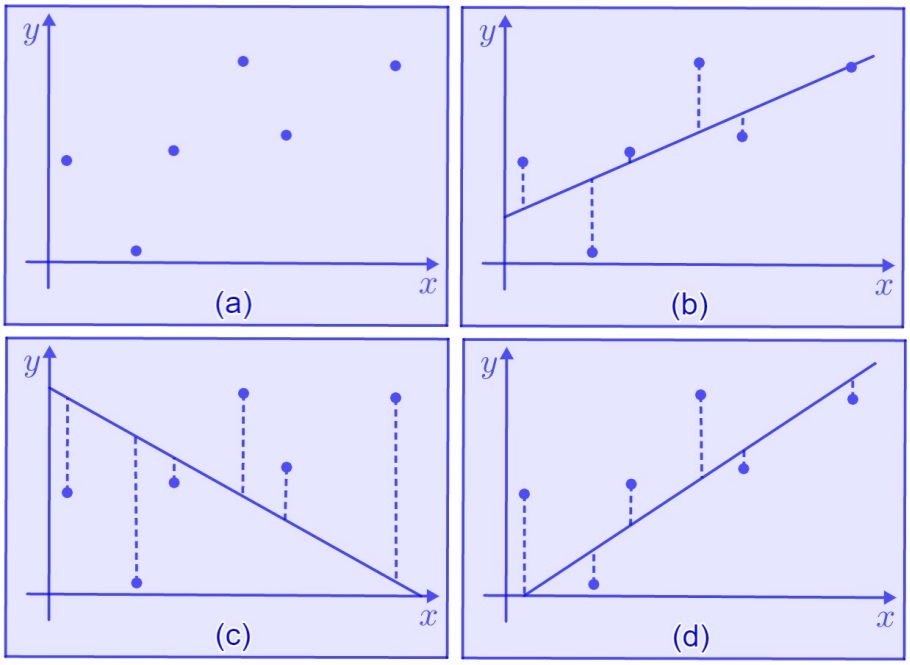
Segundo Bussab [3] (2017, p.317), o MMQ foi

> introduzido por Gauss em 1794, mas que primeiro apareceu com esse nome no apêndice do tratado de Legendre, *Nouvelles Méthodespour la Determination des Orbites des Comètes*, publicado em Paris em 1806. Gauss somente viria a publicar seus resultados em 1809, em Hamburgo. Ambos utilizaram o princípio em conexão com problemas de Astronomia e Física.

Para o desenvolvimento do MMQ consideremos o problema de escolher uma reta para representar um conjunto de n pontos, $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, não necessariamente colineares. Para cada reta candidata, o MMQ analisa as n diferenças entre cada valor y_i e o valor na reta, correspondente ao respectivo valor x_i. A reta escolhida é aquela que apresenta a menor soma de quadrados de tais diferenças.

A Figura 4.3 mostra a ideia do ajuste por mínimos quadrados a um conjunto de seis pontos, apresentados na Figura 4.3 (a).

Figura 4.3: Ajuste da reta por mínimos quadrados.



Fonte: Charnet [4] (2008, p.29) (Adaptada).

Para cada ponto (x_i, y_i), estamos traçando um segmento vertical cujo comprimento é o módulo da diferença entre y_i e o valor da reta em x_i (perceba que essa diferença é exatamente e_i quando o isolamos em 4.1).

A reta na Figura 4.3 (b) é a reta escolhida pelo MMQ. Perceba que na Figura 4.3 (c) todos os segmentos são maiores, se comparados com os da Figura 4.3 (b). Por outro lado, a Figura 4.3 (d) apresenta segmentos de comprimentos bem próximos dos da Figura 4.3 (b); todavia, no cálculo da soma dos mínimos quadrados das diferenças, o resultado ainda é menor para a reta da Figura 4.3 (b).

Se equacionarmos o problema da escolha da reta que melhor se ajusta ao conjunto dos n pontos, o MMQ consiste em encontrar os valores dos parâmetros a e b da expressão (4.1) de forma que minimizem a soma dos quadrados dos erros (ou diferenças), que como vimos são dados por

$$e_i = y_i - (ax_i + b),\ i = 1, 2, ..., n. \tag{4.5}$$

Perceba que ao considerarmos a Soma dos Quadrados dos Erros (SQE) em (4.6), teremos, para cada valor de a e b, um resultado diferente para essa soma de quadrados. Em outros termos, obtemos a quantidade de informação perdida pelo Modelo Linear.

$$SQE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2. \tag{4.6}$$

Portanto, minimizando a SQE estamos também minimizando a quantidade de informação perdida. Assim, nosso objetivo é encontrar o mínimo da função SQE nas variáveis reais a e b.

Para encontar esse mínimo, devemos obter as seguintes derivadas parciais:

$$\frac{\partial}{\partial a} \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

e

$$\frac{\partial}{\partial b} \sum_{i=1}^{n} [y_i - (ax_i + b)]^2.$$

Para um estudo relacionado a funções reais de duas variáveis e derivadas parciais, indicamos [8] (STEWART, 2016) ao leitor. Dando continuidade, se denominarmos por $\hat{a}$ e $\hat{b}$ os valores que minimizam a função, teremos o seguinte sistema:

$$-2 \sum_{i=1}^{n} [y_i - (\hat{a}x_i + \hat{b})] x_i = 0$$

$$-2 \sum_{i=1}^{n} [y_i - (\hat{a}x_i + \hat{b})] = 0,$$

ou ainda,

$$\sum_{i=1}^{n} x_i y_i - \hat{a}\sum_{i=1}^{n} x_i^2 - \hat{b}\sum_{i=1}^{n} x_i = 0 \tag{4.7}$$

$$\sum_{i=1}^{n} y_i - \hat{a}\sum_{i=1}^{n} x_i - n\hat{b} = 0 \tag{4.8}$$

que é denominado *sistema de equações normais.*

Pela equação (4.8), obtemos

$$\begin{aligned} & n\hat{b} = \sum y_i \; - \; \hat{a}\sum x_i \\ \Rightarrow \quad & \hat{b} = \frac{1}{n}\sum y_i \; - \; \hat{a}\frac{1}{n}\sum x_i \\ \Rightarrow \quad & \hat{b} = \overline{y} \; - \; \hat{a}\overline{x}, \end{aligned}$$

e, substituindo na equação (4.7), temos

$$\begin{aligned} & \sum x_i y_i \; - \; \hat{a}\sum x_i^2 \; - \; (\overline{y} \; - \; \hat{a}\overline{x})\sum x_i \; = \; 0 \\ \Rightarrow \quad & \sum x_i y_i \; - \; \hat{a}\sum x_i^2 \; - \; (\overline{y} \; - \; \hat{a}\overline{x})\, n\overline{x} \; = \; 0 \\ \Rightarrow \quad & \sum x_i y_i \; - \; \hat{a}\sum x_i^2 \; - \; n\overline{xy} \; + \; n\hat{a}\overline{x}^2 \; = \; 0 \\ \Rightarrow \quad & \hat{a}\left(\sum x_i^2 \; - \; n\overline{x}^2\right) \; = \; \sum x_i y_i \; - \; n\overline{x}\overline{y} \\ \Rightarrow \quad & \hat{a} = \frac{\sum x_i y_i \; - \; n\overline{x}\overline{y}}{\sum x_i^2 \; - \; n\overline{x}^2}. \end{aligned}$$

Em resumo, obtemos

$$\hat{a} = \frac{\displaystyle\sum_{i=1}^{n} x_i y_i \; - \; n\overline{x}\overline{y}}{\displaystyle\sum_{i=1}^{n} x_i^2 \; - \; n\overline{x}^2} \tag{4.9}$$

e

$$\hat{b} = \overline{y} \; - \; \hat{a}\overline{x}. \tag{4.10}$$

Observe que a solução encontrada poderia representar tanto um mínimo quanto um máximo da função SQE. Todavia, tal função não possui um máximo, bastar notar que, para qualquer reta que passe acima (ou abaixo) dos pontos, sempre podemos apontar uma outra reta cuja soma dos quadrados dos erros seja ainda maior.

Finalmente, substituindo as expressões (4.9) e (4.10) em (4.2), teremos um estimador para $\mu(x_i)$, dado por

$$\hat{\mu}(x_i) = \hat{a}x_i + \hat{b}, \tag{4.11}$$

e que iremos representá-lo da seguinte forma

$$\hat{y} = \hat{a}x + \hat{b}. \tag{4.12}$$

A expressão (4.12) é conhecida como *equação de regressão* e o gráfico de tal equação é denominado *reta de regressão* (ou reta dos mínimos quadrados ou, ainda, reta ajustada).

4.3 Determinação da Reta de Regressão

As expressões para $\hat{a}$ e $\hat{b}$ são muito convenientes para efeito de cálculos, uma vez que precisamos apenas das seguintes quantidades para determiná-los: n, $\overline{x}$, $\overline{y}$, $\sum x_i y_i$ e $\sum x_i^2$.

Exemplo 4.1: Vamos determinar a equação de regressão para as variáveis: *número de anos de serviço* (X) e *número de clientes de agentes de uma companhia de seguros* (Y) do Exemplo 3.1. Tomemos os dados da Tabela 3.1 como base para a construção da Tabela 4.1 a seguir.

Tabela 4.1: Determinação da reta de regressão para as variáveis número de anos de serviço (X) e número de clientes (Y).

Agentes	Anos (X)	Clientes (Y)	$x_i y_i$	x_i^2
A	2	48	96	4
B	3	50	150	9
C	4	56	224	16
D	5	52	260	25
E	4	43	172	16
F	6	60	360	36
G	7	62	434	49
H	8	58	464	64
I	8	64	512	64
J	10	72	720	100
Total	57	565	3.392	383

Fonte: Tabela 3.1.

Da tabela acima obtemos:

- $n = 10$;
- $\overline{x} = 5,7$;
- $\overline{y} = 56,5$;
- $\sum x_i y_i = 3.392$;
- $\sum x_i^2 = 383$.

Agora, utilizando (4.9) e (4.10), obtemos

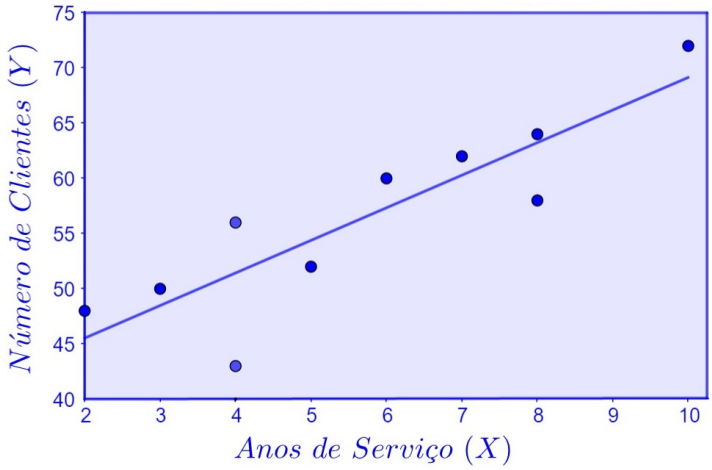
$$\hat{a} = \frac{3.392 - 10(5,7)(56,5)}{383 - 10(5,7)^2} = 2,95$$

e

$$\hat{b} = 56,5 - 2,95(5,7) = 39,68.$$

Portanto, a equação de regressão para as variáveis anos de serviço e número de clientes é dada por $\hat{y} = 2,95x + 39,68$. Esta equação nos informa que um determinado agente terá, em média, um acrescimo de $2,95$ unidades no número de clientes no decorrer de um ano de serviço. A Figura 4.4 mostra o diagrama de dispersão com a reta de regressão ajustada ao conjunto de pontos.

Figura 4.4: Exemplo 4.1: Reta de regressão ajustada



Fonte: Produção do autor.

Exemplo 4.2: Neste exemplo vamos determinar a equação de regressão para as variáveis: *renda bruta mensal* (X) e *porcentagem da renda gasta em saúde* (Y) do Exemplo 3.2. Notemos que as quantidades que necessitamos para calcular as estimativas $\hat{a}$ e $\hat{b}$ já foram determinadas no Exemplo 3.5,

- $n = 10$;
- $\overline{x} = 31,6$;
- $\overline{y} = 6,45$;
- $\sum x_i y_i = 1952,4$;
- $\sum x_i^2 = 12.128$.

Desse modo, utilizando (4.9) e (4.10), obtemos

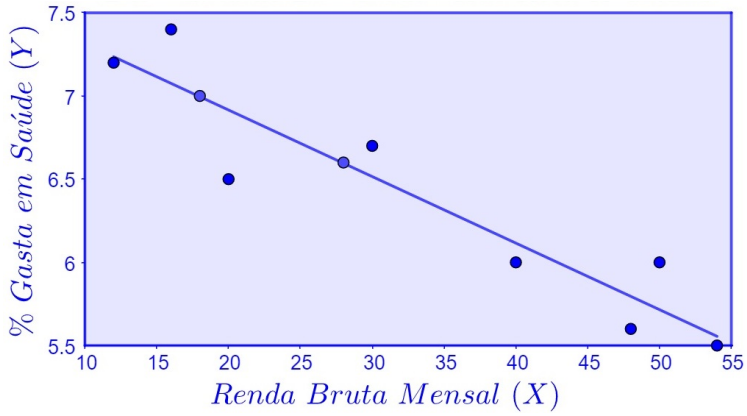
$$\hat{a} = \frac{1.952,4 - 10(31,6)(6,45)}{12.128 - 10(31,6)^2} = -0,04$$

e

$$\hat{b} = 6,45 - (-0,04)(31,6) = 7,72.$$

Finalmemte, a equação de regressão para as variáveis renda bruta mensal e porcentagem da renda gasta em saúde é dada por $\hat{y} = -0,04x + 7,72$. Esta equação nos informa que a porcentagem da renda gasta em saúde terá, em média, um decréscimo de $-0,04$ unidade quando a renda bruta mensal sofrer um acréscimo de um salário mínimo. A Figura 4.5 mostra o diagrama de dispersão com a reta de regressão ajustada.

Figura 4.5: Exemplo 4.2: Reta de regressão ajustada



Fonte: Produção do autor.

A partir destes dois exemplos pontuamos algumas observações importantes, que facilitam o estudo da Regressão Linear Simples.

Observação 4.1 *O sinal do estimador â coincide com o sinal do coeficiente de correlação linear. Para ter uma verificação parcial dessa observação, compare os sinais de â nos Exemplos 4.1 e 4.2 com os respectivos sinais de r das variáveis estudadas.*

Observação 4.2 *O estudo da Regressão Linear Simples é melhor aproveitado após ser apresentado o conteúdo de Correlação Linear, tendo em vista que, uma vez confirmada a correlação linear como sendo o tipo de associação estabelecida entre as variáveis, podemos utilizar o Modelo Linear de forma segura. Além disso, as quantidades necessárias para determinar os estimadores â e $\hat{b}$ são praticamente as mesmas para determinar o coeficiente de correlação linear, sem que seja necessário realizar cálculos adicionais.*

Observação 4.3 *Não se pode extrapolar o conjunto de dados na Regressão Linear Simples. Isto significa que, dado um certo valor x que não pertença ao intervalo limitado pelo menor*

e maior valor observado da variável X, não é possível utilizar a equação de regressão para predizer qual seria o valor esperado para a variável Y. Uma ilustração disso seria no Exemplo 4.1, onde não podemos tomar um x menor que 2 ou maior que 10 para dizer, em média, qual seria o número de clientes de um determinado agente.

Capítulo 5

Proposta Didática com Auxílio do Software GeoGebra

Como já foi mencionado, a análise de Correlação Linear e Regressão Linear Simples é um instrumento de medição extremamente poderoso que estuda o relacionamento entre duas variáveis quantitativas emparelhadas. Todavia, para colocar esta análise em prática é necessário muito cuidado e perspicácia da parte de quem a executa, isto no sentido de inserir dados e realizar os cálculos corretamente, haja vista que um erro (por menor que seja) pode gerar resultados falsos, conduzindo a erros de interpretação.

Percebe-se, ainda, que para um conjunto de dados que possua uma grande quantidade de observações, realizar os cálculos do Coeficiente de Correlação Linear, dos estimadores $\hat{a}$ e $\hat{b}$ e construir o diagrama de dispersão com a reta de regressão ajustada é uma tarefa penosa e que consome uma grande quantidade de tempo.

Todos estes fatos podem ser um dificultador ou até mesmo um desestímulo para o aluno que está tendo o primeiro contato com a Correlação e a Regressão. Nesse sentido, faz-se necessário a busca de recursos que facilitem o ensino desses conteúdos. Felizmente, nas duas últimas décadas, as tecnologias de informação e comunicação (TIC) evoluíram de forma significativa, afetando o processo de aprendizagem nos diversos campos e níveis escolares.

Tendo isto em vista, utilizar essas ferramentas tecnológicas, em nossa prática pedagógica, contornariam as possíveis dificuldades que os alunos venham a apresentar. Além do mais, os documentos oficiais que norteiam a educação no Brasil recomendam o uso dessas ferramentas, por exemplo, a BNCC "propõe que os estudantes utilizem tecnologias, como calculadoras e planilhas eletrônicas, desde os anos iniciais do Ensino Fundamental" (BRASIL [1], 2017, p.518).

Segundo este documento, a inserção de tais tecnologias possibilita que os estudantes possam ser estimulados a desenvolver o pensamento computacional, por meio da interpretação e da elaboração de diagramas, gráficos e algoritmos. Em contrapartida, as Orientações Curriculares para o Ensino Médio, Brasil [2] (2006, p.87), destacam que

> Não se pode negar o impacto provocado pela tecnologia de informação e comunicação na configuração da sociedade atual. Por um lado, tem-se a inserção dessa tecnologia no dia-a-dia da sociedade, a exigir indivíduos com capacitação para bem usá-la; por outro lado, tem-se nessa mesma tecnologia um recurso que pode subsidiar o processo de aprendizagem da Matemática. É importante contemplar uma formação escolar nesses dois sentidos, ou seja, a Matemática como ferramenta para entender a tecnologia, e a tecnologia como ferramenta para entender a Matemática.

Portanto, seguindo as recomendações da BNCC e as Orientações Curriculares para o Ensino Médio, o objetivo deste capítulo é desenvolver uma proposta de atividade para os conteúdos de Correlação Linear e Regressão Linear Simples utilizando o *Software* Matemático GeoGebra. *Software* este que é mencionado como ferramenta auxiliar no processo de ensino aprendizagem em diversos tralbalhos de ensino e pesquisa em todo o mundo.

5.1 O *Software* GeoGebra

5.1.1 Contexto Histórico

O GeoGebra (aglutinação das palavras **Geo**metria e Al**Gebra**) é um *software* de matemática dinâmica para todos os níveis de ensino, foi objeto da tese de doutorado de Markus Hohenwanter na Universidade de Salzburgo, Áustria (2001). Atualmente, Markus lidera uma grande equipe de programadores e pesquisadores entusiasmados no desenvolvimento do GeoGebra para aprendizagem e o ensino da matemática nas escolas.

Podendo ser utilizado em ambientes *online* e *offline*, o GeoGebra reúne, em um único pacote fácil de se usar, elementos da Geometria, Álgebra, Planilha de Cálculo, Gráficos, Probabilidade, Estatística e Cálculos Simbólicos. Este *software* possui uma comunidade que é formada por milhões de usuário em praticamente todos os países. Este feito o tornou líder na área de *softwares* de matemática dinâmica, auxiliando o ensino e a aprendizagem em Ciência, Tecnologia, Engenharia e Matemática.

Com uma ampla possibilidade de uso, o Geogebra foi construído em java, o que possibilita ser executado virtualmente em qualquer sistema operacional como: *Android*, *Windows*, *Windows Phone*, *Unix/Linux* e *iOS*. É um programa multiplataforma, podendo ser executado em Computadores, Tablets e Smartphones. Para Computadores, pode ser encontrado para *donwload* de forma livre e gratuita via *internet* no site http://www.geogebra.org, e para Tablets e Smartphones, está disponível nas lojas de aplicativos, também de forma livre e gratuita.

Podemos também citar, além das mencionadas, outras vantagens que trás o GeoGebra:

- Interface amigável, com vários recursos sofisticados;
- Ferramenta de produção de aplicativos interativos em páginas *WEB*;

- Disponível em vários idiomas para milhões de usuários em torno do mundo.

Atualmente, estão espalhados por todo o mundo 62 Institutos Internacionais de GeoGebra, 6 destes estão localizados no Brasil. Tais institutos são organizações sem fins lucrativos e foram criados devido à ampla divulgação e uso do *software*, onde professores e pesquisadores trabalham juntos para promover o ensino e a aprendizagem da Matemática apoiando e desenvolvendo as seguintes atividades:

- Desenvolver materiais gratuitos para oficinas;
- Oferecer oficinas para professores e para futuros formadores;
- Desenvolver e implementar novas funcionalidades do *software* GeoGebra;
- Desenvolver um sistema de apoio *online* para professores;
- Avaliar e melhorar as atividades de desenvolvimento profissional e materiais;
- Projetar e implementar tópicos de pesquisa;
- Comunicações em conferências nacionais e internacionais.

5.1.2 Interface Gráfica

Nesta subseção trataremos dos seguintes assuntos: instalação do *software* GeoGebra no sistema operacional *windows* e apresentação de sua interface gráfica. Estamos utilizando como referência o GeoGebra Clássico 6; logo, tudo que aqui for mencionado refere-se a esta versão. A Figura 5.1 apresenta a logo do GeoGebra.

Figura 5.1: Logo do GeoGebra

Fonte: Produção do autor.

Para realizar a instalação no *Windows* basta seguir os passos abaixo. A instalação em outros sistemas operacionais é relizada de modo análogo,

1º) Acesse o site [14] https://www.geogebra.org/download;

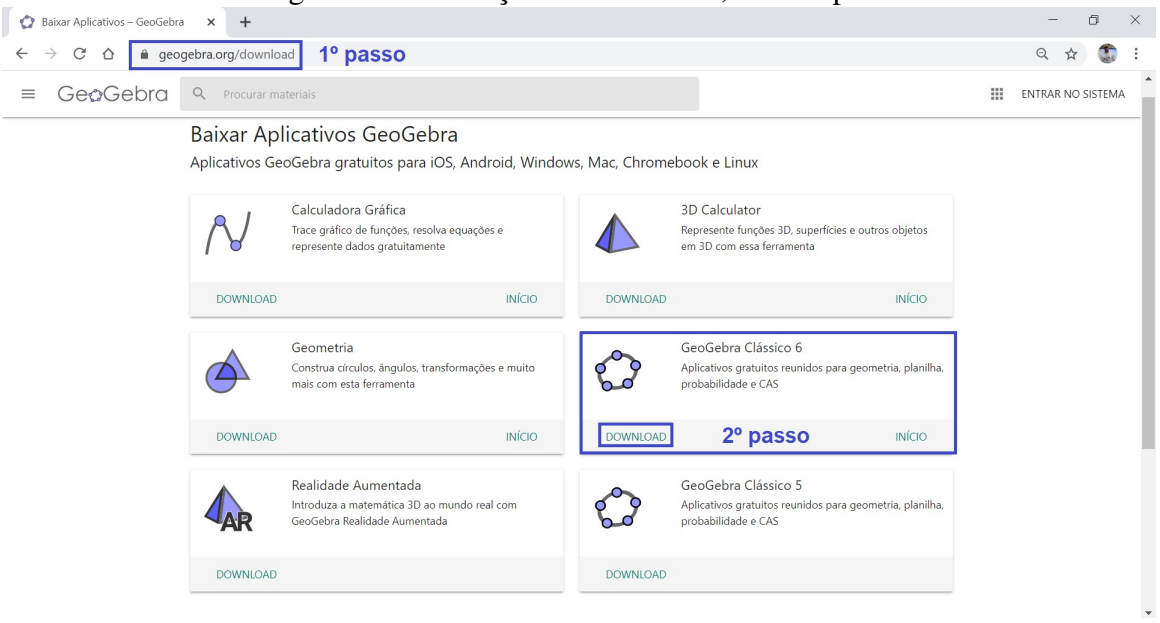
2º) Selecione o GeoGebra Clássico 6 e clique em Download;

3º) Na pasta "Downloads" execute o aplicativo baixado.

Obs: A Figura 5.2 ilustra o 1º e 2º passo e a 5.3 ilustra o 3º.

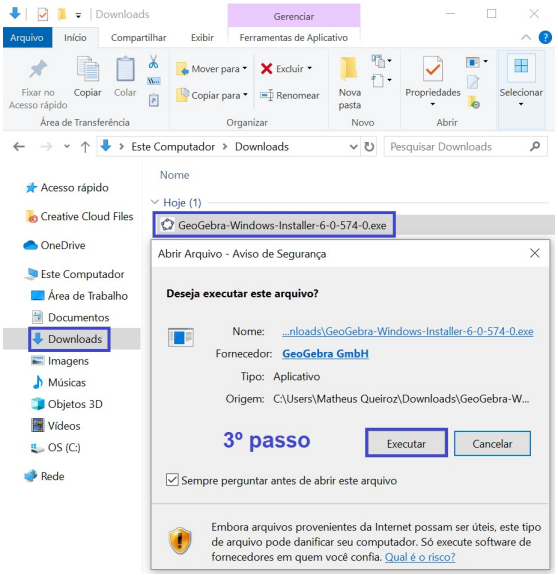
Dessa forma o GeoGebra está pronto para uso.

Figura 5.2: Instalação do GeoGebra, 1º e 2º passo



Fonte: Produção do autor.

Figura 5.3: Instalação do GeoGebra, 3º passo



Fonte: Produção do autor.

A interface gráfica do GeoGebra está organizada em seis regiões:

1. **Barra de Menus**

   A Barra de Menus disponibiliza opções como para salvar o projeto em arquivo (.ggb) e para controlar configurações gerais.

2. **Barra de Ferramentas**

   A Barra de Ferramentas concentra todas as ferramentas úteis para construir pontos, retas, figuras geométricas, obter medidas de objetos construídos, entre outros. Cada ícone dessa barra esconde outros ícones que podem ser acessados clicando com o mouse em seu canto inferior direito.

3. **Janela de Álgebra**

   Região em que é exibida as coordenadas, equações, medidas e outros atributos dos objetos construídos.

4. **Entrada**

   Campo de entrada para digitação de comandos.

5. **Janela de Vizualização**

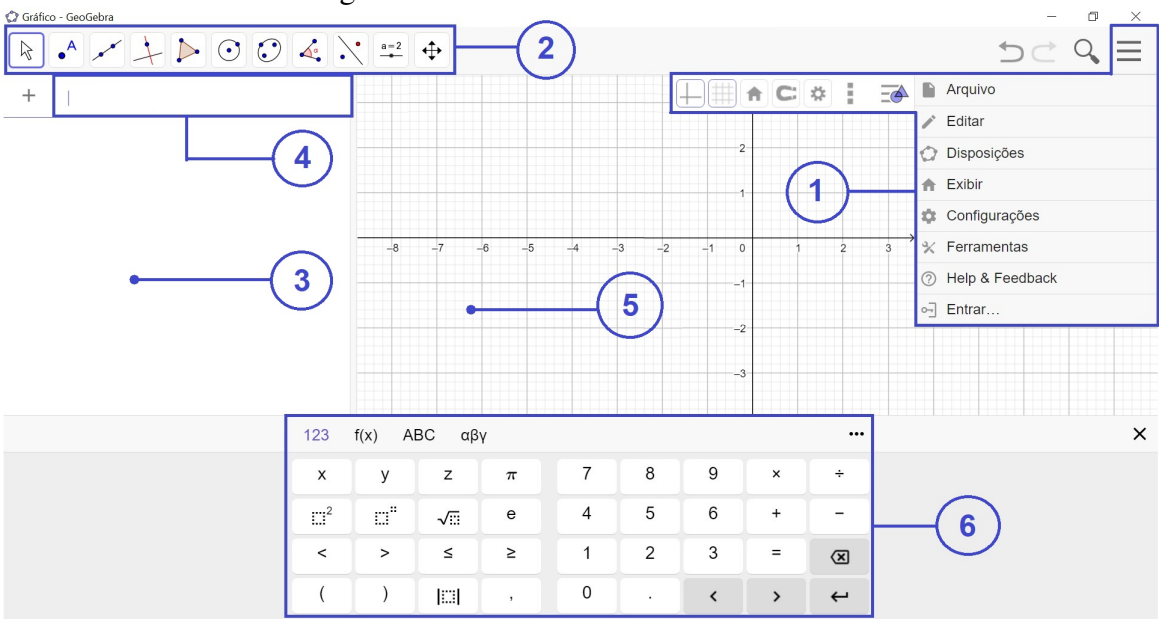
   Região de visualização gráfica de objetos que possuam representação geométrica e que podem ser desenhados com o mouse usando ícones da Barra de Ícones ou comandos digitados na Entrada

6. **Teclado Virtual**

   O Teclado Virtual possui uma listagem de comandos predefinidos que auxiliam na digitação.

   A Figura 5.4 a seguir apresenta a interface do *software* GeoGebra.

Figura 5.4: Interface Gráfica do GeoGebra



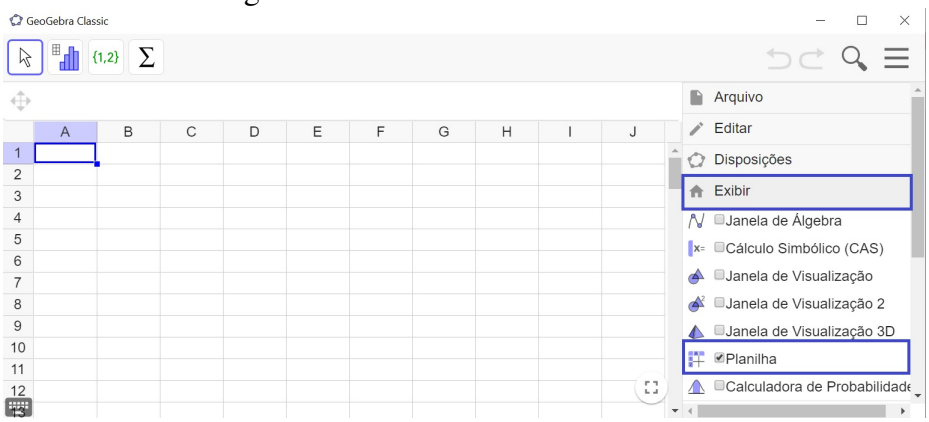
Fonte: Produção do Autor.

5.1.3 GeoGebra: Correlação Linear e Regressão Linear Simples

Vamos, agora, ensinar a configurar o GeoGebra para que ele fique no formato de Planilha, possibilitando a inseção de dados emparelhados e, consequentemente, a análise de Correlação Linear e Regressão Linear Simples.

Inicialmente, na *Barra de Menus* clique na guia *Exibir*. Nesta guia estarão selecionadas a *Janela de Algebra* e a *Janela de Visualização*, deixe apenas selecionado o campo *Planilha*. A interface do GeoGebra ficará igual a que se encontra na Figura 5.5.

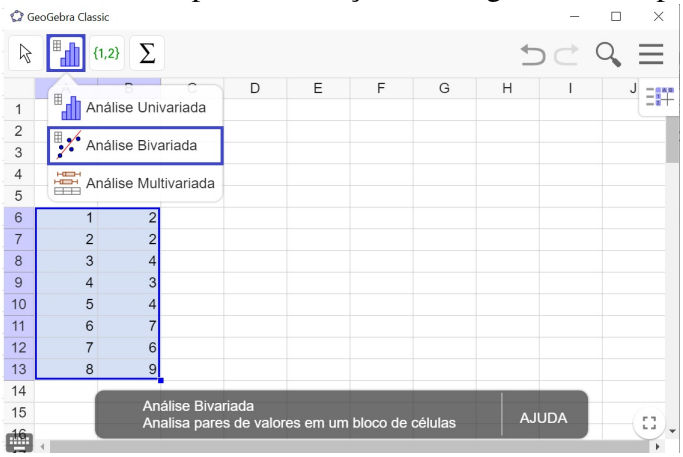
Figura 5.5: GeoGebra formato Planilha



Fonte: Produção do autor.

Para construir o Diagrama de Dispersão, insira e selecione os dados que serão plotados no diagrama. Em seguida, clique no ícone  e selecione a opção *Análise Bivariada*, de acordo com a Figura 5.6.

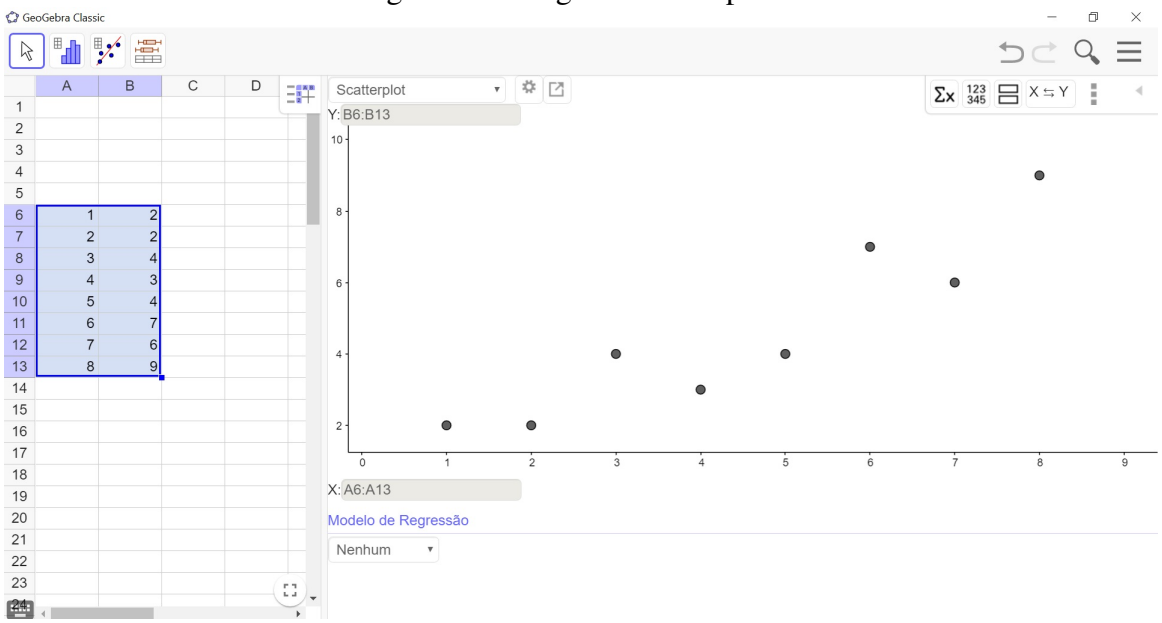
Figura 5.6: Passos para construção do Diagrama de Dispersão



Fonte: Produção do autor.

Realizando esses passos o GeoGebra gera automaticamente o Diagrama de Dispersão, ficando com o aspecto da Figura 5.7.

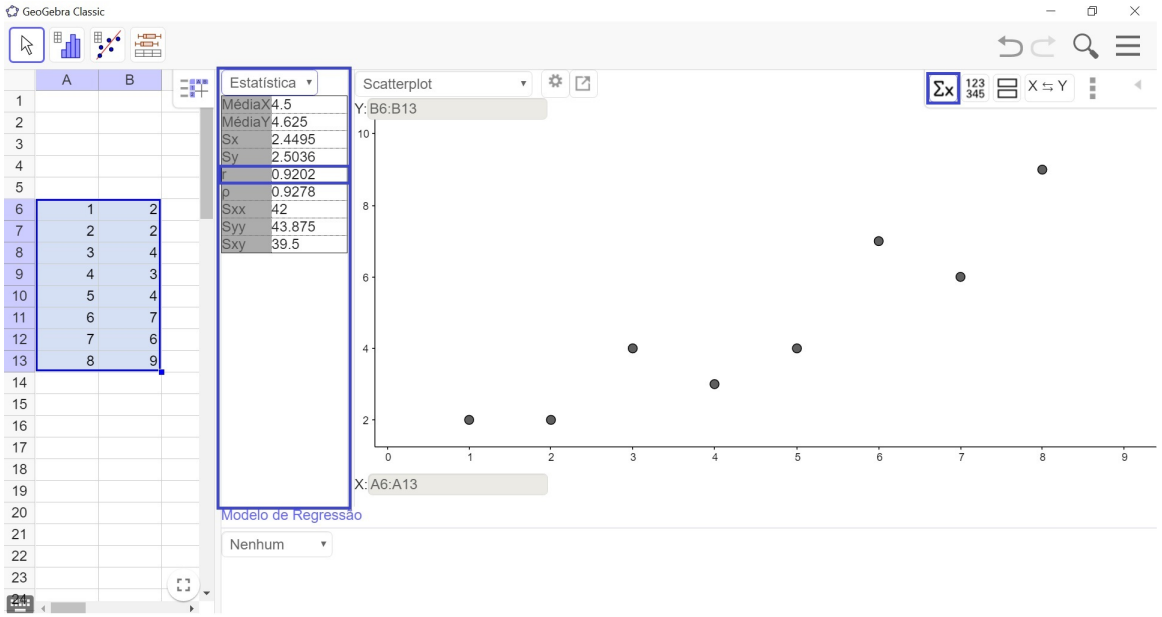
Figura 5.7: Diagrama de Dispersão



Fonte: Produção do autor.

Para calcular o Coeficiente de Correlação Linear (r), basta clicar no ícone Σx (Exibir Estatística). Com isso, aparecerá uma janela contendo uma tabela com várias estatísticas referentes ao conjunto de dados inserido, dente elas, o r, conforme Figura 5.8.

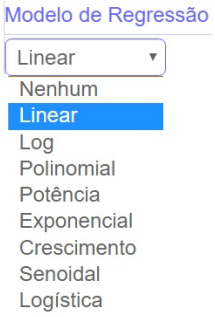
Figura 5.8: Coeficiente de Correlação Linear (r)



Fonte: Produção do autor.

Finalmente, para determinar a Equação de Regressão, bem como sua reta ajustada, clique na guia que se encontra logo abaixo de *Modelo de Regressão*, Figura 5.9. Aparecerá varios modelos de regressão (Linear, Logarítimica, Polinomial, etc.), escolha o modelo Linear.

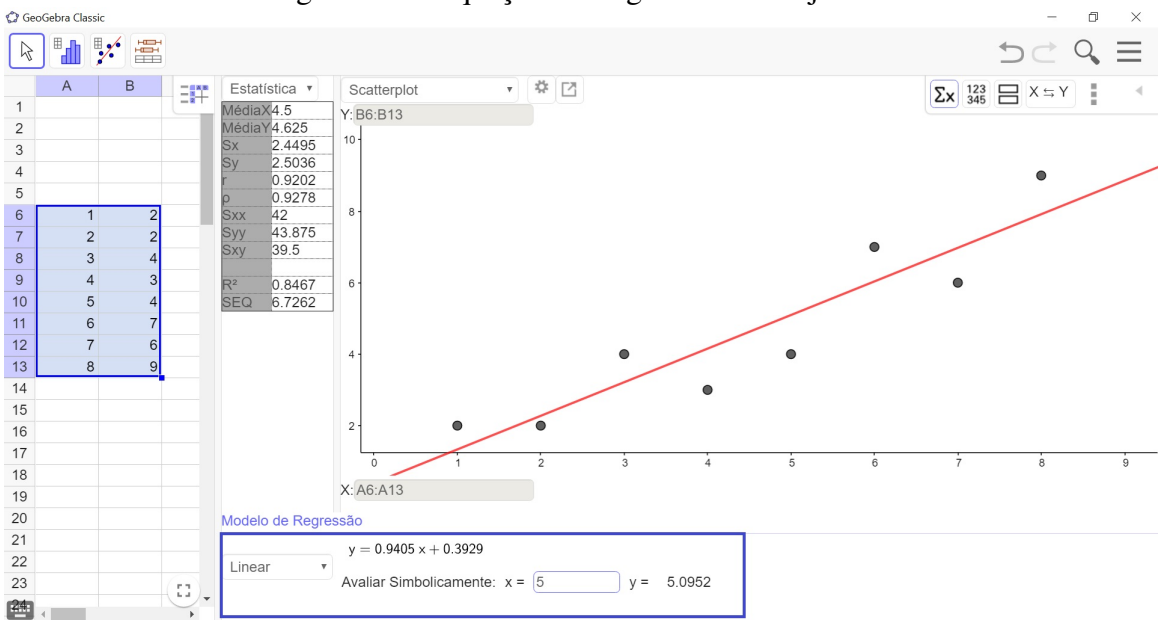
Figura 5.9: Selecionando o modelo Linear



Fonte: Produção do autor.

Assim, o GeoGebra gerará a Equação de Regressão e apresentará a reta ajustada no Diagrama de Dispersão. Também é possível predizer valores da variável Y para um certo valor x da variável X. A Figura 5.10 esclarece os pormenores.

Figura 5.10: Equação de Regresão e reta ajustada



Fonte: Produção do autor.

Observação 5.1 *A ordem em que as variáveis são inseridas altera o Diagrama de Dispersão e a Equação de Regressão, ou seja, os pares de dados (x,y) e (y,x) geram resultados distintos, exceto para o Coeficiente de Correlação Linear.*

Observação 5.2 *Os dados utilizados nas ilustrações acima são fictícios.*

5.2 Sequência Didática

Sequência didática é definida por Zabala ([10], 1998, p.18 - grifos do autor) como "*um conjunto de atividades ordenadas, estruturadas e articuladas para a realização de certos objetivos educacionais, que têm um princípio e um fim conhecidos tanto pelos professores como pelos alunos*". Assim, o objetivo desta seção é apresentar uma sequência didática para o ensino da análise de Correlação Linear e Regressão Linear Simples, ou seja, apresentar um conjunto de atividades encadeadas para tornar mais eficiente o processo de apredizagem destes conteúdos.

5.2.1 Público-Alvo e Apresentação do Conteúdo

Propõe-se que a análise de Correlação Linear e Regressão Linear Simples seja apresentada para alunos do 3º ano do Ensino Médio. A escolha deste público se deve ao fato de que os conteúdos da Estatística Descritiva, tais como: medidas de tendência central e dispersão, construção e interpretação de tabelas de frequência e representações gráfica, já terem sido apresentados por completo aos mesmos, tendo em vista a necessidade de tais conteúdos para o estudo que se segue.

Uma vez selecionado o público-alvo, partimos para a forma de exposição do conteúdo.

Com relação ao material teórico e exemplos voltados para os conteúdos de Correlação Linear e Regressão Linear Simples, o professor pode fazer uso do que foi desenvolvido na introdução deste trabalho e nos capítulos 3 e 4.

Abaixo estão elencados cinco momentos que o professor pode seguir para realizar a apresentação do conteúdo:

Primeiro momento - Apresentação do contexto histórico dos temas que serão abordados em sala, mencionando os precursores e suas contribuições para a estatística e as outras áreas do conhecimento.

Segundo momento - Introdução dos conteúdos e conceitos por meio de um exemplo ou de uma situação-problema ou, ainda, de uma situação "motivadora", fazendo com que os alunos pensem nas possíveis estratégias e soluções para o exemplo/situação.

Terceiro momento - Retomada e resolução do exemplo/situação proposto(a) no segundo momento.

Quarto momento - Formalização e a sistematização teórica utilizando a linguagem e o rigor matemático e estatístico necessário, além da clareza e precisão nas definições, bem como as justificativas lógicas nas demonstrações (quando cabíveis).

Quinto momento - Aplicação de atividades, classificadas em Básicas, Complementares e Avaliativas.

As atividades citadas no quinto momento serão explanadas de forma detalhada a frente.

5.2.2 Atividades

Atividade Básica

A atividade básica (ver Apêndice A.1) tem como finalidade colocar o aluno para praticar a construção de Diagramas de Dispersão (com e sem reta ajustada) e os cálculos que determinam o Coeficiente de Correlação Linear e os estimadores $\hat{a}$ e $\hat{b}$ da Reta de Regressão. Propõe-se que ela seja realizada em sala e de forma individual, as questões marcadas com (*calculadora*) faculta ao aluno o uso da calculadora. Nesta atividade, o professor entra com o papel de auxiliador, tirando dúvidas e corrigindo possíveis erros.

Atividade Complementar

A atividade complementar (ver Apêndice A.2) visa por em ação a capacidade de análise, compreensão e interpretação do educando ao que se refere à correlação Linear e Regressão Linear Simples. A proposta é que os cinco primeiros problemas desta atividade sejam realizados no laboratório de informática da unidade escolar com a ajuda do *software* GeoGebra, o professor pode optar por aplicá-los individualmente ou em duplas.

O sexto problema é um trabalho em equipe. Este item tem como objetivo trabalhar o planejamento, a organização e as comunicações oral e escrita. Ele irá contribuir para o desenvolvimento de atitudes, tais como: dividir tarefas e se comprometer com elas, ajudar os colegas, lidar com diferentes opiniões, fazer uma exposição oral com desenvoltura e etc.

Neste estágio, o professor deve apresentar o GeoGebra para os alunos, tal como fizemos na seção 5.1, além de entrar com o papel de auxiliador, tirando dúvidas, dando dicas e corrigindo erros.

Atividade Avaliativa

A atividade avaliativa (ver Apêncide A.3) tem como objetivo checar o conhecimento que o aluno adquiriu acerca dos conteúdos apresentados. Esta atividade possui cinco problemas que estão divididos em dois grupos: aqueles que podem ser solucionados com o auxílio do GeoGebra, tais problemas apresentam a marca (*GeoGebra*), e os que não podem ter este auxílio. Como o Geogebra será utilizado, a atividade deve ser aplicada no laboratório de informática da unidade escolar e de forma individual. Assim o professor conseguirá detectar quais alunos obtiveram uma apredizagem significativa e quais não a obtiveram.

É claro que a atividade avaliativa não é o único instrumento de avaliação que o professor dispõe para verificar o conhecimento adquirido pelo aluno. No decorrer das aulas, o docente pode criar outras oportunidades de avaliação, como por exemplo:

- solicitar aos alunos que expliquem, na lousa, oralmente ou, ainda, no GeoGebra, exercícios e resolução de problemas contidos nas atividades básica e complementar;
- observar as interações (aluno/professor e aluno/aluno) dos estudantes durante a exposição dos conteúdos e resolução das atividades;
- propor que elaborem, individualmente ou em grupo, uma atividade ou situação-problema para um colega resolver individualmente ou em grupo.

Capítulo 6

Considerações Finais

Como foi destacado na introdução deste trabalho, a Correlação Linear e a Regressão Linear Simples são tópicos da Estatística Descritiva que não são tratados formalmente na BNCC do Ensino Médio dentro do conjunto de conteúdos da área de Matemática e suas Tecnologias. Porém, nesta pesquisa foi mostrado que é possível e viável trabalhar esses conteúdos com alunos que estejam neste nível de ensino, além de termos desenvolvido uma proposta didática para tanto.

Nesse sentido, visando a imediata necessidade da disseminação da estatística e o pleno desenvolvimento da formação do aluno nesta área, esperamos que, ao findo deste trabalho, professores de Matemática de todo o Brasil, que lecionam em turmas do 3º ano do Ensino Médio, passem a trabalhar os conteúdos de Correlação Linear e Regressão Linear Simples com seus alunos, utilizando para tanto a proposta didática que aqui foi desenvolvida. Esperamos também que esses conteúdos sejam incluídos, em futuras alterações que venham a ocorrer na BNCC, no conjunto de conteúdos da área de Matemática e suas tecnologias.

É fato que estimular novos pensamentos e proporcionar uma aprendizagem significativa é o papel de todo professor. Nós professores, como agentes da educação, não podemos fugir destas obrigações, muito pelo contrário, temos sempre que inovar e quando necessário reinovar nossas práticas pedagógicas, sempre pensando no futuro, mas nunca esquecendo das necessidades do presente.

Em particular, e agora parafraseando meu orientador Alexsandro Bezerra em uma de suas espetaculares aulas, atualmente a estatística é uma área da matemática pouca difundida na educação básica, logo cabe a nós (professores de matemática) refletirmos a respeito da prática estatística na sala de aula e realizar esta difusão.

Referências Bibliográficas

[1] BRASIL. *Base Nacional Comum Curricular (BNCC).* Brasília: MEC, 2017.

[2] BRASIL. Secretaria de Educação Básica, *Orientações Curriculares para o Ensino Médio*: Ciências da Natureza, Matemática e suas Tecnologias; Volume 2; Brasília: Ministério da Educação, 2006.

[3] BUSSAB, Wilton O.; MORETTIN, Pedro A.; *Estatística Básica.* 9ª ed. - São Paulo: Saraiva, 2017.

[4] CHARNET, Reinaldo et al.; *Análise de Modelos de Regressão Linear: com Aplicações.* 2ª Ed. - São Paulo: Editora da UNICAMPI, 2008.

[5] GUEDES, Terezinha A.; *Projeto de Ensino Aprender Fazendo Estatística.* Universidade Estadual de Maringá, 2005.

[6] IEZZI, Gelson; HAZZAN, Samuel; DEGENSZAJN, David M.; *Fundamentos de Matemática Elementar: Volume 11.* 9ª Ed. - São Paulo: Atual, 2013.

[7] MEYER, Paul L.; *Probabilidade: Aplicação à Estatística.* 2ª ed. - Rio de Janeiro: LTC, 1983.

[8] STEWART, James. *Cálculo: Volume 2.* Tradução: Helena Maria Ávila de Castro. 8ª ed. - São Paulo, SP: Cengage Leaening, 2016.

[9] TRIOLA, Mario F.; *Introdução à Estatística: Atualização da Tecnologia.* 11ª Ed. - Rio de Janeiro: LTC, 2013.

[10] ZABALA, Antoni. *A Prática Educativa: Como Ensinar.* Porto Alegre: Artmed, 1998.

[11] `http://galton.org/`. Último acesso em 17 Mar. 2020.

[12] `https://karlpearson.org/`. Último acesso em 17 Mar. 2020.

[13] `https://www.geogebra.org/`. Último acesso em 19 Mar. 2020.

[14] `https://www.geogebra.org/download`. Último acesso em 19 Mar. 2020.

Apêndice A

Atividades Aplicadas

A.1 Atividade Básica

1. Construa o Diagrama de Dispersão referente as variáveis X e Y dadas abaixo.

X	5	8	7	10	6	7	9	3	8	2
Y	6	9	8	10	5	7	8	4	6	2

2. Complete o esquema de cálculo do Coeficiente de Correlação Linear para os valores das Variáveis X e Y:

X	6	8	10	12	14
Y	14	12	10	14	16

Temos:

Observação	X	Y	$x_i y_i$	x_i^2	y_i^2
1	6	14	84	36	196
____	____	____	____	____	____
____	____	____	____	____	____
____	____	____	____	____	____
5	12	14	168	144	196
Total	____	____	____	____	____

Logo:

$$r = \frac{\_\_\_\_ - \_\_\_\_(\_\_\_\_)(\_\_\_\_)}{\sqrt{(\_\_\_\_ - \_\_\_\_(\_\_\_\_)^2)(\_\_\_\_ - \_\_\_\_(\_\_\_\_)^2)}} =$$

3. Complete o esquema para o cálculo dos estimadores $\hat{a}$ e $\hat{b}$ da Equação de Regressão para os valores das variáveis X e Y:

X	2	4	6	8	10	12	14
Y	30	25	22	18	15	11	10

Temos:

Observação	X	Y	$x_i y_i$	x_i^2
1	2	30	60	4
____	____	____	____	____
____	____	____	____	____
____	____	____	____	____
____	____	____	____	____
____	____	____	____	____
7	14	10	140	196
Total	____	____	____	____

Logo:

$$\hat{a} = \frac{\_\_\_\_ - \_\_\_\_(\_\_\_\_)(\_\_\_\_)}{\_\_\_\_ - \_\_\_\_(\_\_\_\_)^2} =$$

e

$$\hat{b} = \_\_\_\_ - \_\_\_\_(\_\_\_\_) =$$

4. Construa o Diagrama de Dispersão, com reta ajustada, referente aos dados da questão anterior.

5. (*Calculadora*) Pretendendo-se estudar a relação entre as variáveis *Consumo de Energia Elétrica* (X) e *Volume de Produção nas Empressas Industriais* (Y), faz-se uma amostragem que inclui vinte empressas, computando-se os seguintes valores:

$$\sum x_i = 11,34,\ \sum y_i = 20,72,\ \sum x_i^2 = 12,16,\ \sum y_i^2 = 84,96\ e\ \sum x_i y_i = 22,13$$

Determine:

(a) O Coeficiente de Correlação Linear.

(b) A Equação de Regressão onde Y é a variável independente e X é a variável dependente.

(c) A Equação de Regressão onde X é a variável independente e Y é a variável dependente.

A.2 Atividade Complementar

1. Numa amostra de cinco operários de uma dada empresa foram observadas as variáveis *Anos de Experiência num dado Cargo* (X) e *Tempo* (Y), em minutos, gasto na execução de uma certa tarefa relacionada com esse cargo.

   As observações são apresentadas na tabela abaixo:

X	1	2	4	4	5
Y	7	8	3	2	2

   Você diria que a variável X pode ser usada para explicar a variação de Y? Justifique.

2. Considere os resultados de dois testes, X e Y, obtidos por um grupo de alunos da escola **A**:

X	11	14	19	19	22	28	30	31	34	37
Y	13	14	18	15	22	17	24	22	24	25

   (a) Verifique, pelo Diagrama de Dispersão, se existe correlação linear.

   (b) Em caso afirmativo, calcule o Coeficiente de Correlação Linear.

   (c) Escreva, em poucas linhas, as conclusões a que chegou sobre a relação entre essas variáveis.

3. A tabela abaixo apresenta a produção de uma indústria de 2010 a 2018:

Ano	2010	2011	2012	2013	2014	2015	2016	2017	2018
Quantidade	34	36	36	38	41	42	43	44	46

   (a) Determine o Coeficiente de Correlação Linear.

   (b) Use para o tempo uma variável auxiliar, por exemplo: $x'_i = x_i - 1980$, e refaça o cálculo do Coeficiente de Correlação Linear. Compare o resultado obtido com o do item (a).

   (c) Qual é o cuidado que devemos ter ao fazer uso de uma variável auxiliar?

   (d) Determine a reta ajustada para este conjunto de dados e interprete-a.

   (e) Qual produção estimada para 2019.

4. A tabela abaixo indica o valor Y do *aluguel* e a *idade* X de cinco casas.

X	10	13	5	7	20
Y	4	3	6	5	2

   (a) Obtenha a equação de regressão ajustada, $\hat{y} = \hat{a}x + \hat{b}$. Em seguida, construa o diagrama de dispersão e a reta ajustada.

   (b) Você acha que o modelo adotado é razoável?

   (c) Qual o significado do estimador $\hat{a}$ nesse caso?

   (d) E do $\hat{b}$?

5. Os dados abaixo referem-se a *meses de experiência* de dez digitadores e o *número de erros cometidos* na digitação de determinado texto.

Meses X	1	2	3	4	5	6	7	8	9	10
Erros Y	30	28	24	20	18	14	13	10	7	6

   (a) Represente graficamente esse conjunto de dados.

   (b) Assumindo que um modelo de regressão linear é adequado, determine os estimadores da Equação de Regressão, represente a reta de regressão no gráfico feito anteriormente e interprete seu resultado.

   (c) Qual o número esperado de erros para um digitador com 5 meses de experiência?

6. (*Atividade de Pesquisa*) Junte-se com mais quatro colegas de sala que moram no seu bairro, ou próximo dele. Na sua vizinhança, selecione 20 pares de pai e filho e extraia as seguintes variáveis

   X : *altura dos pais*;

   Y : *altura dos filhos*.

   Utilizando a análise de Correlação Linear e Regressão Linear Simples, estudem o relacionamento entre as variáveis X e Y e elaborem um relatório dos resultados obtidos.

A.3 Atividade Avaliativa

1. Muitas vezes a determinação da capacidade de produção instalada para certo tipo de indústria em certas regiões é um processo difícil e custoso. Como alternativa, pode-se estimar a capacidade de produção através da escolha de uma outra variável de medida mais fácil e que esteja linearmente relacionada com ela.

   Suponha que foram observados os valores para as variáveis: capacidade de produção instalada, potência instalada e área construída. Com base num critério estatístico, qual das variáveis você escolheria para estimar a capacidadede produção instalada?

X: cap. prod. inst. (ton.)	4	5	4	5	8	9	10	11	12	12
Y: potência inst. (1.000 kW)	1	1	2	3	3	5	5	6	6	6
Z: área construída (100 m)	6	7	10	10	11	9	12	10	11	12

$$\begin{array}{lll} \sum x = 80, & \sum y = 38, & \sum z = 100, \\ \sum x^2 = 736, & \sum y^2 = 182, & \sum z^2 = 1.048, \\ \sum xy = 361, & \sum xz = 848, & \sum yz = 411, \end{array}$$

2. (*GeoGebra*) Abaixo estão os dados referentes à porcentagem da população economicamente ativa empregada no setor primário e o respectivo índice de analfabetismo para algumas regiões metropolitanas brasileiras.

Regiões metropolitanas	Setor primário	Índice de anafalbetismo
São Paulo	2,0	17,5
Rio de Janeiro	2,5	18,5
Belém	2,9	19,5
Belo Horizonte	3,3	22,2
Salvador	4,1	26,5
Porto Alegre	4,3	16,6
Recife	7,0	36,6
Fortaleza	13,0	38,4

   (a) Faça o Diagrama de Dispersão.

   (b) Você acha que existe uma dependência linear entre as duas variáveis?

   (c) Determine o Coeficiente de Correlação Linear.

   (d) Existe alguma região com comportamento diferente das demais? Se existe, elimine o valor correspondente e recalcule o coeficiente de correlação.

3. A tabela abaixo apresenta valores que mostram como o comprimento de uma barra de aço varia conforme a temperatura:

Temperatura (C°)	10	15	20	25	30
Comprimento (mm)	1,003	1,005	1,010	1,011	1,014

Com a ajuda de uma variável auxiliar, determine:

(a) O Coeficiente de Correlação Linear.

(b) A Equação de Regressão.

(c) O valor estimado do comprimento da barra para a temperatura de 18°C.

(d) O valor estimado do comprimento da barra para a temperatura de 35°C.

4. (*GeoGebra*) Um laboratório está interessado em medir o efeito da temperatura sobre a potência de um antibiótico. Dez amostras de 50 gramas cada foram guardadas a diferentes temperaturas, e após 15 dias mediu-se a potência. Os resultados estão no quadro abaixo.

Temperatura	30°		50°			70°			90°	
Potência	38	43	32	26	33	17	27	23	14	21

(a) Faça a representação gráfica dos dados.

(b) Ajuste a reta de regressão, da potência como função da temperatura.

(c) O que você acha desse modelo?

(d) A que temperatura a potência média seria nula?

5. (*GeoGebra*) Os dados abaixo correspondem às variáveis *renda familiar* (X) e *gasto com alimentação* (Y) numa amostra de dez famílias, representadas em reais.

X	300	500	1.000	2.000	3.000	5.000	7.000	10.000	15.000	20.000
Y	150	200	600	1.000	1.500	2.000	2.500	4.000	6.000	8.000

Obtenha a Equação de Regressão ajustada, $\hat{y} = \hat{a}x + \hat{b}$.

(a) Qual a previsão do gasto com alimentação para uma família com renda de 17.000 reais?

(b) Qual a previsão do gasto para famílias com excepcional renda, por exemplo 100.000 reais? Você acha esse valor razoável? Por quê?

(c) Se você respondeu que o valor obtido em (b) não é razoável, encontre uma explicação para o ocorrido.

Sugestão: interprete a natureza das variáveis X e Y e o comportamento de Y para grandes valores de X.

Apêndice B

Competências Específicas da Área de Matemática e suas Tecnologias do Ensino Médio

Competência Específica 1

Utilizar estratégias, conceitos e procedimentos matemáticos para interpretar situações em diversos contextos, sejam atividades cotidianas, sejam fatos das Ciências da Natureza e Humanas, ou ainda questões econômicas ou tecnológicas, divulgados por diferentes meios, de modo a consolidar uma formação científica geral.

Competência Específica 2

Articular conhecimentos matemáticos ao propor e/ou participar de ações para investigar desafios do mundo contemporâneo e tomar decisões éticas e socialmente responsáveis, com base na análise de problemas de urgência social, como os voltados a situações de saúde, sustentabilidade, das implicações da tecnologia no mundo do trabalho, entre outros, recorrendo a conceitos, procedimentos e linguagens próprios da Matemática.

Competência Específica 3

Utilizar estratégias, conceitos e procedimentos matemáticos, em seus campos – Aritmética, Álgebra, Grandezas e Medidas, Geometria, Probabilidade e Estatística –, para interpretar, construir modelos e resolver problemas em diversos contextos, analisando a plausibilidade dos resultados e a adequação das soluções propostas, de modo a construir argumentação consistente.

Competência Específica 4

Compreender e utilizar, com flexibilidade e fluidez, diferentes registros de representação matemáticos (algébrico, geométrico, estatístico, computacional etc.), na busca de solução e comunicação de resultados de problemas, de modo a favorecer a construção e o desenvolvimento do raciocínio matemático.

Competência Específica 5

Investigar e estabelecer conjecturas a respeito de diferentes conceitos e propriedades matemáticas, empregando recursos e estratégias como observação de padrões, experimentações e tecnologias digitais, identificando a necessidade, ou não, de uma demonstração cada vez mais formal na validação das referidas conjecturas.