

## **ESTATÍSTICA APLICADA NA ANÁLISE DA RENDA FAMILIAR ANUAL E O PREÇO DA CASA DE CIDADES DOS ESTADOS UNIDOS**

**Geraldo Magela Barbosa** é mestre em Engenharia pela Escola Politécnica da Universidade de São Paulo – USP, Engenheiro Mecânico com ênfase em Produção pela Universidade Federal de Itajubá – UNIFEI. É licenciado em Matemática pelo Centro Universitário de Itajubá – UNIVERSITAS e professor da FATEC-SEBRAE.

### **Resumo**

A estatística está inserida em todas as áreas do conhecimento. Deve-se, portanto, fazer uso da sua aplicação, o seu entendimento e sua interpretação como ferramenta de pesquisa. Neste artigo a estatística será aplicada na interpretação, análise e tomada de decisão sobre o conjunto de dados de uma amostra da renda familiar anual e o preço da casa de cidades dos Estados Unidos. Na análise dos dados serão utilizadas medidas estatísticas, dispersão dos valores e construção de gráficos. O modelo estatístico obtido pelo método de regressão linear será testado e comparado com o grau de relação linear da renda familiar anual e o preço da casa.

**Palavras-chave:** Análise de Dados; Medidas Estatísticas; Regressão; Correlação.

**Editor Geral**

Prof. Dr. Mário Pereira Roque Filho

**Organização e Gestão**

Prof. Ms. Clayton Pedro Capellari

**Correspondência**

Alameda Nothmann, nº 598 Campos Elíseos, CEP 01216-000 São Paulo – SP, Brasil.

+55 (11) 3224.0889 ramal: 218

E-mail: [f272dir@cps.sp.gov.br](mailto:f272dir@cps.sp.gov.br)

## Abstract

Statistics is inserted in all areas of knowledge. Therefore, one must make use of its application, its understanding and its interpretation as a research tool. In this article the statistics will be applied in the interpretation, analysis and decision making on the data set of a sample of the annual family income and the price of the house of cities in the United States. In the analysis of the data will be used statistical measures, dispersion of the values and construction of graphs. The statistical model obtained by the linear regression method will be tested and compared with the degree of linear relationship of the annual family income and the price of the house.

**Key words:** Data Analysis, Statistical Measures, Regression, Correlation.

## Introdução

A Estatística pode ser definida como um conjunto de métodos e processos quantitativos que servem para estudar e medir os fenômenos coletivos. Ao se estudar os fenômenos coletivos, o mais importante são os fatos que envolvem os elementos, como eles se relacionam e qual o seu comportamento. Os dados devem ser levantados através de uma pesquisa científica. Sendo esta pesquisa elaborada com metodologia e exatidão dos seus objetivos.

Os dados estatísticos estão no dia a dia das pessoas, e o conjunto dos métodos e processos da Estatística faz a interpretação desses dados, que podem passar pelo domínio da Estatística Descritiva para o campo da Estatística Inferencial com a contribuição da Teoria das Probabilidades. A Estatística é fundamental na análise de dados provenientes de quaisquer processos onde exista variabilidade, estando assim, interessada nos métodos e processos quantitativos que servem para a coleta, organização, resumo, apresentação e análise desses dados, bem como na obtenção de conclusões válidas e na tomada de decisões a partir de tais análises.

A Estatística Descritiva trata da organização e sintetização os dados estatísticos. Isto envolve o cálculo e a interpretação de medidas numéricas como a média aritmética, a mediana e o desvio padrão, assim como a constatação e uso de dispositivos gráficos.

Na vida, sempre que for necessário tomar uma decisão, deve-se levar em conta um grande número de fatores. Quando se analisa o mundo que nos cerca, identifica-se que todos os acontecimentos, sejam eles culturais ou naturais, envolvem um grande número de variáveis. As diversas ciências têm a pretensão de conhecer a realidade, e de interpretar os acontecimentos e os fenômenos, baseadas no conhecimento das variáveis intervenientes, consideradas importantes nesses eventos. Estabelecer relações, encontrar, ou propor, leis explicativas, é papel próprio da ciência. Para isso, é necessário controlar, manipular e medir as variáveis que são consideradas relevantes ao entendimento do fenômeno analisado.

Conforme Bussab e Morettin (2017) um pesquisador, em alguma fase do seu trabalho, depara com o problema de analisar e entender um conjunto de dados relevante ao seu particular objeto de estudo. Precisar-se-á então, trabalhar os dados para transformá-los em informações, para compará-los com outros resultados, ou ainda para julgar sua adequação a alguma teoria ou hipótese. Pode-se dizer que a essência da *ciência* é a observação e que seu objetivo básico é a *inferência*, que pode ser *dedutiva* (na qual se argumenta das premissas às conclusões) ou *indutiva* (por meio da qual se vai do específico ao geral).

Segundo Karmel e Polasek (1972) a palavra Estatística muitas vezes é utilizada como significado de dados, como ocorre quando se fala de *estatística da renda nacional* ou *estatísticas vitais*, mas também é usada como referência a todo campo do conhecimento. A estatística tem uma aplicação muito ampla, mas os seus princípios e a sua prática independem do campo onde é aplicada.

Para Silva, Silva, Gonçalves e Murolo (2010) o pesquisador que utiliza no seu estudo um conjunto de dados se vê obrigado a trabalhar com grande quantidade de valores numéricos. Esses valores numéricos observados são chamados dados estatísticos. A Estatística ensina métodos racionais para trabalhar os dados estatísticos observados e a obtenção de informações a respeito de um fenômeno e também permitir a tomada de decisões.

De acordo com Ramos (2000) a Estatística se preocupa com a coleta, a organização, a descrição, a análise e a interpretação de dados. Mas, o mais importante, é que possibilita a tomada de decisões com base em fatos, em vez de simples opiniões. Métodos estatísticos vêm sendo aplicados em todo o mundo, com o objetivo de fazer com que dados, que aparentemente nada significam, sejam corretamente interpretados e, portanto, melhores decisões sejam obtidas e tomadas.

A aplicação da estatística ocorre em todas áreas do conhecimento, pois contribui para a obtenção das informações necessárias ao planejamento, através de estimativas, previsões e simulações que sustentaram as decisões a serem tomadas. O atual universo globalizado sofre influências de algumas variáveis, como as situações socioeconômicas, políticas ou mercadológicas. Nesse contexto, as projeções estatísticas são indicadores que nos orientam sobre o comportamento dos mercados das diversas áreas e de sua evolução (Tiboni, 2010).

Conforme Martins e Domingues (2019) a *estatística descritiva* é a organização, sumarização e descrição de um conjunto de dados. Com a construção de gráficos, como o *histograma* (*representação gráfica através de retângulos adjacentes em que a base corresponde aos intervalos da classe, e a altura é dada pela frequência absoluta*), tabelas, e com o cálculo de *medidas* (por exemplo: a *média aritmética*, a *mediana* e a *moda*) tomando como referência uma coleção de dados numéricos, pode-se melhor compreender o comportamento da variável expressa no conjunto de dados sob análise.

Na Estatística o conceito de *População* considera um conjunto finito ou infinito de valores, que possui ao menos uma característica em comum entre todos os seus elementos componentes. A *amostra* é um subconjunto ou uma parte da totalidade de observações abrangidas pela população, e que deve ser um *subconjunto representativo da população*. A coleta dessa amostra deve ser feita com base em métodos adequados, chamados de *amostragem* (Ramos, Almeida e Araújo, 2013).

*População* (ou *universo*) é a totalidade de itens, objetos ou pessoas, que deve apresentar pelo menos uma característica comum. *Amostra* é uma parte representativa da população que é selecionada para análise, sendo então a amostra um subconjunto finito da população.

Segundo Spiegel (2006) a *estatística inferencial* são métodos que tornam possível a estimação de características sobre parâmetros de uma população baseadas nos resultados amostrais. Os métodos da *estatística inferencial* possibilitam a extrapolação das informações a partir do subconjunto de dados (amostra), ou seja, procura inferir, induzir e estimar as leis de comportamento da população e assim estabelecer conclusões sobre a população em estudo.

Para Webster (2006) a característica de uma amostra ou de uma população que está sendo analisada é definida na estatística como *variável*. A variável pode ser *discreta* onde o conjunto de dados pode ser enumerados, por exemplo número de filhos. E ainda pode ser *variável contínua* onde seus dados podem assumir qualquer valor real, por exemplo medidas de altura entre: um metro sessenta e três centímetros e um metro e sessenta e quatro centímetros, são infinitas as possibilidades, de forma que não se pode enumerá-las.

A evolução possibilitou que a estatística seja aplicada em diversas atividades econômicas ou empresariais, e em todas as áreas de conhecimento. Assim como também na utilização para a obtenção de conclusões válidas, por meio de análises juntamente com interpretações de dados. Pois através da coleta de dados é admissível obter o conhecimento da realidade geográfica e social, recursos nativos, humanos e financeiros, as perspectivas da sociedade em relação à empresa, assim podendo determinar metas melhores. Devido a sua importância, a estatística está em todas as áreas de conhecimento, sendo uma ferramenta utilizada na interpretação, análise de dados, tomada de decisões e situação com componentes de imprevisibilidade ou multiplicidade de variáveis pertinentes, fornecendo elementos para controle, gestão e melhoria constante de processos.

Na Estatística as informações numéricas são obtidas com a finalidade de acumular informação para a tomada de decisão. A estatística pode ser vista como um conjunto de técnicas para planejar experimentos, obter dados e organizá-los, resumi-los, analisá-los, interpretá-los e deles extrair conclusões. As técnicas clássicas da estatística foram delineadas para serem as melhores possíveis sob rigorosas suposições.

É importante lembrar que geralmente nas grandes empresas ou na economia, enfim em áreas que englobam muitos dados, usa-se a amostra como forma de execução da pesquisa. Como se torna muito trabalhoso e com um custo muito alto pesquisar a população toda, faz-se um estudo preliminar da mesma a fim de encontrar os parâmetros e então buscar a amostra que tem confiabilidade no seu resultado, sendo menos trabalhoso e com menor custo.

A área financeira possui uma grande aplicabilidade das pesquisas estatísticas, pois se dedica aos lucros, aos custos, aos gastos, às avaliações, às análises dos processos, entre outras considerações. É fundamental para um gestor ter uma ampla visão da empresa para que ele possa tomar atitudes. Os números lhe fornecem interpretações com mais exatidão e permite maior confiabilidade na ação

Mesmo seguindo essa linha de raciocínio, a amostra é uma ferramenta fundamental e se bem aplicada permite que o administrador tome as decisões com confiança. As escolhas somente apresentarão resultados eficientes se tiverem como base dados de exatidão ou de alguma forma comprovada pelo uso adequado dessas amostras.

A situação real de uma empresa, de como ela se situa em relação ao mercado, é garantida pela média aritmética. Por exemplo, ao analisar a quantidade de peças vendidas semestralmente em uma autopeça, há como se ter um controle maior do estoque. Assim também a mediana, a moda, a média ponderada e a média geométrica são responsáveis por análises das observações, procurando um número central estratégico. A variância e o desvio padrão visam medir o distanciamento de seus dados ou observações em relação à média aritmética.

Na indústria, o *controle estatístico de processos* é uma ferramenta que utiliza a estatística com o objetivo de fornecer informações para um diagnóstico mais eficaz na prevenção e detecção de falhas (defeitos), identificando suas causas em tempo real, o que, conseqüentemente, auxilia no aumento da produtividade (resultados) da empresa, evitando desperdícios de matéria-prima, insumos e produtos.

No mercado financeiro e nas instituições bancárias os métodos estatísticos são empregados para modelagem financeira e econômica, visando modelar o

comportamento do crédito, da inadimplência, a movimentação de ações e previsões de taxas de juros, possibilitando estabelecer estratégias para a concessão de empréstimos que maximizem os lucros.

Na engenharia de produção de uma indústria, os métodos estatísticos são empregados para o *planejamento e controle da produção*, visando à implantação de técnicas administrativas eficientes que garantam menores custos e maiores lucros, na estimação de receitas, previsão de estoques e demandas e, principalmente, conhecimento do mercado e do seu cliente.

Nos sistemas implantados na empresa, a estatística colabora para obter materiais, controle de estoque e compra de matéria-prima. A administração em uma empresa é facilitada com a estatística, tomando como base os movimentos futuros dos cenários econômicos. A estatística incentiva e influencia os administradores a terem reações de maneira inteligente e eficaz em relação às informações e dados disponíveis, para que desta forma tenha uma reflexão antes da tomada de decisão.

Os empresários, ao usarem a estatística, têm mais condições de decidir com mais precisão o controle da organização, isto é, além de conhecerem a importância da estatística é necessário decidir que providências tomar com base nos seus resultados. Visto que a estatística é uma ferramenta fundamental tanto no início, no desenvolvimento e na manutenção das empresas, pois monitoram o progresso e tem uma contribuição nas melhorias dos resultados.

## **1. Metodologia**

Em termos metodológicos o presente estudo se classifica, quanto aos objetivos como descritivo, quanto aos procedimentos como estudo de caso e quanto à abordagem do problema como quantitativo. O estudo de caso foi adotado como método nesta pesquisa, pois permite a construção do raciocínio usando-se conceitos teóricos aplicáveis em apenas um caso analisado.

A investigação de estudo de caso, tem como base várias fontes de evidências e beneficia-se do desenvolvimento prévio de proposições teóricas para concluir a coleta e análise de dados. O estudo de caso é uma estratégia de pesquisa para compreender um método que abrange tudo em abordagens específicas de coletas e análise de dados.

Trata-se de uma pesquisa de caráter quantitativo com o uso de procedimentos estatísticos para analisar um conjunto de dados de uma amostra. Quanto aos meios, a pesquisa desenvolveu-se em estudos bibliográficos a fim de apresentar conceitos, métodos e ferramentas relacionadas a pesquisa.

A pesquisa bibliográfica pode ser classificada como o primeiro passo de toda pesquisa científica, concedendo a compreensão de que a solução de um problema e a relação de estudos de campo podem ser alcançado por meio dela.



A estatística aplicada nessa pesquisa relaciona os dados de uma amostra da Renda Familiar Anual e o Preço da Casa de cidades dos Estados Unidos. Em seguida uma explanação das definições e descrição das medidas que serão utilizadas na análise dos dados. Serão utilizadas as medidas estatísticas: média aritmética, mediana, moda e a medida de dispersão dos valores (variância e desvio padrão). O modelo estatístico (matemático) da relação da Renda Familiar Anual e o Preço da Casa será feito pela equação de regressão linear. Essa modelagem será testada com o conjunto de dados amostrais e comparada com o resultado da correlação linear.

O procedimento técnico será realizado, com base na apresentação dos métodos explanados de forma teórica, que consiste no estudo profundo e exaustivo da análise estatística dos dados amostrais que permitirá seu amplo e detalhado conhecimento da variação da Renda Familiar Anual e o Preço da Casa de cidades dos Estados Unidos.

## **2. A Análise Exploratória de Dados e as Medidas Estatísticas**

A *Análise Exploratória de Dados* foi impulsionada pelo estatístico norte-americano John Wilder Tukey, que incentivava os estatísticos a explorarem os dados e possivelmente formular hipóteses que poderiam levar a novas coletas de dados e experimentos.

Tukey definiu a *Análise Exploratória de Dados* em 1961 como: "Procedimentos para analisar dados, técnicas para interpretar os resultados de tais procedimentos, formas de planejar a reunião dos dados para tornar sua análise mais fácil, mais precisa ou mais exata e todos os cálculos e resultados da estatística (matemática) que se aplicam a análise de dados". A análise exploratória de dados emprega grande variedade de técnicas gráficas e quantitativas, visando maximizar a obtenção de informações, descobrir a tendência dos dados, detectar o comportamento do fenômeno e criar modelos matemáticos que representem os dados.

O livro *Exploratory Data Analysis* (Análise Exploratória de Dados) publicado em 1977 por John Wilder Tukey foi desenvolvido para mostrar que se dava ênfase demais em estatística aos testes de hipóteses (análise confirmatória de dados) e que se precisava dar mais ênfase ao uso de dados para sugerir hipóteses a se testarem.

- 
- A *Análise Exploratória de Dados* nos fornece um extenso repertório de métodos para um estudo detalhado dos dados, antes de adaptá-los. Nessa abordagem, a

finalidade é obter dos dados a maior quantidade possível de informação, que indique modelos plausíveis a serem utilizados numa fase posterior, a *análise confirmatória de dados ou inferência estatística*.

- A *Análise Exploratória de Dados*, antigamente chamada apenas de *Estatística Descritiva*, constitui o que a maioria das pessoas entende como Estatística, e inconscientemente usa no dia a dia. Consiste em resumir e organizar os dados coletados através de tabelas, gráficos e medidas estatísticas, e a partir dos dados resumidos procurar alguma regularidade ou padrão nas observações dando uma visão global dos dados.

Durante as últimas décadas tem-se desenvolvido uma série de tipos de análise de dados que se situaram entre a estatística descritiva e a estatística inferencial. A Análise Exploratória, por exemplo, considera que os dados são constituídos de regularidades ou tendências e desvios ou variabilidades. Dessa forma, a proposta é que se estude grande parte das perspectivas com as ferramentas possíveis, com o propósito de extrair o máximo de informação possível e gerar novas hipóteses. As descrições dos dados podem ser feitas com a Estatística Descritiva, analisando-se as tabelas, os gráficos, as medidas de tendência central e a variabilidade.

A Análise Exploratória de Dados tem como objetivo observar o que os dados podem nos dizer além da modelagem formal. Explorar os dados e formular hipóteses que podem levar a novas coletas de dados e experimentos.

O estudo e análise da variabilidade dos dados para tirar o máximo de informação possível pode ser feito com algumas *Medidas Estatísticas*. Além do gráfico e da tabulação que representam os dados, pode-se encontrar um valor típico, ou seja, um número que melhor represente a distribuição ou um conjunto deles. Esses valores recebem o nome de *Medidas de Tendência Central*.

Dentre as *Medidas Estatísticas* as mais importantes são as *Medidas de Tendência Central* que são: a *média aritmética*, *mediana* e a *moda*. A *medida de dispersão* representa a variabilidade existente dentro de um conjunto de dados, e a medida mais utilizada é o *desvio padrão*.

Conforme Vieira (2014) para entender as características gerais de um conjunto de

dados, é preciso a sua organização, apresentação, e em seguida a análise e interpretação desses dados e deles extrair conclusões. Isso pode ser feito por meio de *Medidas de Tendência Central* que são a *média aritmética (média)*, *mediana* e a *moda* e as de *Variabilidade* tais como a *variância* e *desvio padrão*.

### 3. Medidas de Tendência Central e as Medidas de Dispersão

Segundo Martins e Donaire (1990) as medidas estatísticas possibilitam representar um conjunto de dados relativos à observação de determinado fenômeno de forma resumida. Fornecendo uma ideia geral a respeito do fato ou fenômeno analisado. Essas medidas são as *Medidas de Posição ou Medidas de Tendência Central*, pois representam os fenômenos pelos seus valores médios, em torno dos quais tendem a concentrar-se os dados.

#### 3.1. Medidas de Tendência Central

As Medidas de Tendência Central: a *Média Aritmética (Média)*, a *Mediana* e a *Moda* são as mais importantes para a análise na estatística.

**1) Média Aritmética:** é a medida mais utilizada por ser robusta do ponto de vista matemático pois considera no seu cálculo todos os valores da distribuição e para a interpretação das discrepâncias entre eles. Podemos interpretar a média aritmética como o ponto de equilíbrio dos desvios da distribuição. A média aritmética também é utilizada nos cálculos estatísticos avançados.

De acordo com Lapponi (2005, p.88) na análise do procedimento de cálculo da média aritmética pode-se concluir que:

- todos os valores da variável são incluídos no cálculo da média aritmética,

- a média aritmética está posicionada de forma equilibrada entre os valores ordenados da amostra,
- a média aritmética não é uma medida resistente, pois ela é sensível à presença de dados extremos (dados com valores bem diferentes da maioria dos dados coletados),
- nas amostras ou variáveis com histograma simétrico, os valores da mediana, a moda e a média aritmética coincidem, seus valores são iguais,
- a média aritmética é fácil de incluir em funções matemáticas.

Para Costa Neto (2002), a Média Aritmética para elementos não agrupados é definida a seguir, seja  $X_i$  ( $i = 1, 2, \dots, n$ ) o conjunto de dados, define-se sua média aritmética ou, simplesmente, média como:

$$\text{Média Aritmética (Média): } \bar{X} = \frac{\sum X_i}{n} \quad (1)$$

**2) Mediana:** é uma medida posicional, seu valor se encontra na posição central. Supondo que a distribuição, de dados não agrupados, está em ordem crescente a *mediana* é o valor que divide os elementos em duas partes iguais.

Conforme McClave, Benson e Sincich (2009, p.56-57) a mediana de um conjunto de dados quantitativos é o número do meio quando as medidas são organizadas em ordem crescente ou decrescente para dados não agrupados. Na análise do procedimento de obtenção da mediana pode-se concluir que:

- a mediana é de muito valor quando se está descrevendo grandes conjuntos de dados,
- se o conjunto é caracterizado por um histograma a estimativa da mediana é o ponto sobre o eixo da abscissa que divide o histograma em duas regiões de áreas iguais,
- em algumas situações, com valores extremos, a mediana pode ser uma medida de tendência central melhor que a média aritmética,
- a mediana é menos sensível (não é afetada por valores extremos da amostra),
- nos dados de uma amostra, a mediana possui um valor único.

De acordo com Martins e Domingues (2019) a Mediana para dados não agrupados, que estejam em ordem crescente é assim definida:

- Se  $n$  for ímpar, a *Mediana* será o elemento central (de ordem  $\frac{n+1}{2}$ )
- Se  $n$  for par, a *Mediana* será a Média entre os elementos centrais (de ordem  $\frac{n}{2}$  e  $\frac{n}{2} + 1$ ).

**3) Moda:** é o valor que é mais frequente (que mais se repete) em um conjunto de elementos de uma distribuição.

Segundo Bruni (2017, p.53) na análise do procedimento de cálculo da moda pode-se concluir que:

- é fácil o cálculo da moda,
- a moda não é afetada por valores extremos,
- o resultado da moda pode estar afastado do centro das observações,
- a distribuição pode ter mais de uma moda,
- no cálculo da moda, não se usa todos os dados disponíveis.

A Moda para dados não agrupados é de fácil identificação pela simples observação do elemento que apresenta maior frequência em uma série de valores.

### 3.2. Medidas de Dispersão

As Medidas de Dispersão são medidas estatísticas utilizadas para avaliar o grau de variabilidade ou dispersão de um conjunto de valores em torno da média aritmética. As medidas de variação medem o grau com que os dados tendem a se distribuir em torno de um valor central que, geralmente, é a média aritmética, e o *desvio padrão* é a medida mais usada para esse estudo.

A representatividade da média aritmética é medida pelo grau de variabilidade ou de dispersão. Utiliza-se na estatística as medidas de variância e desvio padrão para

analisar as características de dispersão dos elementos de um conjunto em torno da média aritmética.

### 1) Variância:

Dado o conjunto de valores:  $x_1, x_2, \dots, x_n$  e a sua *Média Aritmética*  $\bar{x}$ .

Obtém-se os *Quadrados das Diferenças* entre os valores e a *Média Aritmética*:

$$\left(x_1 - \bar{x}\right)^2, \left(x_2 - \bar{x}\right)^2, \dots, \left(x_n - \bar{x}\right)^2$$

A Variância da População é a *Média Aritmética dos Quadrados das Diferenças*:

$$\text{Variância Populacional: } \sigma^2 = \frac{\left(x_1 - \bar{x}\right)^2 + \left(x_2 - \bar{x}\right)^2 + \dots + \left(x_n - \bar{x}\right)^2}{n}$$

ou simplesmente:

$$\text{Variância Populacional: } \sigma^2 = \frac{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}{n} \quad (2)$$

A Variância Amostral tem a redução de um grau de liberdade no denominador  $n-1$ .

$$\text{Variância Amostral: } s^2 = \frac{\left(x_1 - \bar{x}\right)^2 + \left(x_2 - \bar{x}\right)^2 + \dots + \left(x_n - \bar{x}\right)^2}{n-1}$$

ou simplesmente:

$$\text{Variância Amostral: } s^2 = \frac{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}{n-1} \quad (3)$$

## 2) Desvio Padrão:

Conforme Costa, Epprecht e Carpinetti (2005) a variabilidade dos valores medidos em um produto são corrigidos através do estudo da dispersão dos valores em torno de um valor central (média aritmética) que é o *Desvio Padrão*. E o desvio padrão é utilizado na melhoria do controle estatístico da qualidade do produto.

O Desvio Padrão Populacional (utiliza toda a quantidade  $n$  do conjunto de dados) é definido como a *Raiz Quadrada da Variância Populacional*:

$$\text{Desvio Padrão Populacional: } \sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \quad (4)$$

O Desvio Padrão Amostral tem a redução de um grau de liberdade no denominador  $n-1$ . O valor que isso resulta representa uma estimativa melhor do desvio padrão da população da qual a amostra foi extraída.

$$\text{Desvio Padrão Amostral: } \sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \quad (5)$$

Interpretação do desvio padrão:

Segundo Martins e Domingues (2019, p.45) para qualquer distribuição amostral com média  $\bar{x}$  e desvio padrão  $S$ , há as *Regras Empíricas*, com a seguinte interpretação:

- o intervalo  $\bar{x} \pm S$  contém entre 60% e 80% de todas as observações amostrais.

- o intervalo  $\bar{x} \pm 2S$  contém aproximadamente 95% das observações amostrais.
- o intervalo  $\bar{x} \pm 3S$  contém aproximadamente 100% das observações amostrais

## 4. Regressão e Correlação Linear

A análise de regressão linear descreve a relação existente entre duas variáveis através de uma equação de reta ajustada aos pontos. E o grau de relação entre duas variáveis aleatórias é estudado pela correlação linear conforme o coeficiente de correlação de Pearson.

### 4.1. Regressão Linear Simples

Para Webster (2006) a análise de Regressão Linear é um conjunto de métodos e técnicas para o estabelecimento de uma reta que interprete a relação funcional entre duas variáveis com boa aproximação. A análise de regressão foi desenvolvida pelo cientista inglês Sir Francis Galton (1822-1911). Galton estudou o padrão de hereditariedade das alturas dos adultos humanos. Ele descobriu que crianças com pais muito altos ou muito baixos tendem a “regredir” para a altura média da população adulta. E esse fenômeno Galton denominou “regressão” ou retorno a média.

A análise de regressão linear simples compreende a análise de dados amostrais para saber como duas variáveis estão relacionadas, uma com a outra, em uma população. O estudo da regressão linear simples tem como resultado a tentativa de estabelecer uma *equação matemática linear (linha reta)* que *descreve o relacionamento* entre duas variáveis. A finalidade da equação de regressão linear simples é de *estimar* ou *predizer* valores de uma variável com base em valores conhecidos da outra.



As equações lineares (linha reta) são importantes porque servem para aproximar muitas relações da vida real, e que são relativamente fáceis de calcular e de interpretar.

De acordo com Stevenson (1981, p.348), é importante reconhecermos que na equação obtida de regressão linear simples:

- a equação obtida se trata de uma relação média entre os pontos do conjunto de valores da amostra e os pontos da reta ajustada,
- o resultado final de um cálculo na equação não obterá necessariamente um valor exato se comparado com a base de dados utilizada no equacionamento,
- não se deve extrapolar valores, que não estejam dentro do intervalo do conjunto de dados usados no equacionamento. Porque para fora do intervalo dos elementos do conjunto a relação pode não se verificar.

A análise de regressão é usada com o propósito de desenvolver um *modelo estatístico (matemático)* para uma amostra de  $n$  unidades que possa ser usado para estimar uma variável dependente ( $Y$ ) em função dos valores das variáveis explicativas ou independentes:  $X_i(x_1, x_2, \dots, x_n)$ . A estatística oferece meios de encontrar a equação ajustada de uma reta que passa o mais próximo possível dos pares de pontos de uma amostra:  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  pelo método do mínimo da soma ao quadrado dos desvios verticais entre os pontos teóricos (pontos da reta estimada) e os pontos experimentais (pares ordenados da amostra). Essa reta linear simples que melhor se ajusta aos dados amostrais é obtida pelo conhecido Método dos Mínimos Quadrados.

Equação da Reta de Mínimos Quadrados (*Regressão Linear Simples*), segundo Martins e Domingues (2019) é assim definida:

$$\text{Equação da Reta Ajustada: } \hat{Y} = a + bX_i \quad (6)$$

$$\text{Sendo: } b = \frac{\sum X_i Y_i - \frac{(\sum X_i) \cdot (\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}} \quad (7)$$

$$a = \frac{\sum Y_i}{n} - b \cdot \frac{\sum X_i}{n} \quad (8)$$

Onde:

$\hat{Y}$  = o valor da previsão (valor ajustado) de  $Y$  para uma observação de  $X_i$

$X_i$  = o valor de  $X$  para a observação  $i$

$a$  = ponto de intersecção entre a reta de regressão estimada e o eixo  $Y$

$b$  = inclinação da reta ajustada (coeficiente angular)

Na equação de regressão linear simples os valores  $\hat{Y}$  são preditos com base em valores dados ou conhecidos de  $X_i$ .

O coeficiente  $b$  da equação de regressão linear simples mede a variação que ocorre

em  $\hat{Y}$ , em *média*, por unidade de variação em  $X_i$ .

De acordo com Anderson, Sweeney e Williams (2011, p.447) as hipóteses sobre o modelo de regressão linear são:

- para cada observação da variável independente, o erro no resultado da variável dependente para uma grande série de experimentos é uma variável aleatória com uma média igual a zero,
- a variância do erro é a mesma (constante), para todo valor da variável independente,
- os valores de erros associados a duas observações quaisquer são independentes. Isto é, o erro associado no resultado de uma variável dependente não afeta o erro associado com outro valor da variável dependente,

- o erro no resultado da variável dependente é uma variável aleatória normalmente distribuída.

A reta de regressão linear simples satisfaz as seguintes propriedades: a) *a soma dos desvios verticais* entre os valores dos pontos observados e os pontos da reta ajustada é zero b) *a soma dos quadrados dos desvios verticais* entre os valores dos pontos observados e os pontos da reta ajustada é *mínima*. Isto é, nenhuma outra reta daria menor soma de quadrados de tais desvios.

Então a *Equação da Reta Ajustada obtida pela Regressão Linear Simples* através do *Método dos Mínimos Quadrados* é a melhor medida, *em média*, entre os valores dos pontos observados e os pontos da reta ajustada, ou seja, a *melhor reta de ajustamento*.

## 4.2. Correlação Linear

*Correlação Linear* é o grau de relação linear existente entre duas variáveis contínuas e normalmente distribuídas. Indica o grau de aderência ou qualidade do ajuste dos pares ordenados dos valores da amostra a uma equação linear (uma reta), de acordo com Costa (2015).

O *coeficiente de correlação* é uma medida da intensidade de associação existente entre duas variáveis quantitativas, e sua fórmula de cálculo foi proposta por *Karl Pearson* em 1896. Por essa razão, é também *denominado coeficiente de correlação linear de Pearson*, conforme Callegari-Jacques (2003).

O *Coefficiente de Correlação* independe das unidades de medida das variáveis. Trata-se de um número adimensional. Somando ou subtraindo e multiplicando ou dividindo um valor constante e arbitrário em ambas as variáveis, o Coeficiente de Correlação não se altera, segundo Fonseca, Martins e Toledo (1985).

Para Spiegel (2006) pode-se determinar, de modo *qualitativo*, quão bem uma reta representa a relação entre as variáveis, mediante a observação direta do próprio diagrama de dispersão. Ao tratar de modo *quantitativo* o problema da dispersão dos dados amostrais, em relação a reta, será necessário estabelecer *medidas de correlação*. O *coeficiente de correlação* mede o grau de relação entre as variáveis. O

coeficiente de correlação mede a excelência do ajustamento aos dados pela equação de reta realmente considerada.

O grau de associação linear entre duas variáveis é medido pelo coeficiente de correlação linear. Quando o coeficiente de correlação é igual a  $+1$ , diz-se que as duas variáveis possuem uma correlação linear positiva perfeita. Quando o coeficiente de correlação é igual a  $-1$ , diz-se que as variáveis possuem uma correlação linear negativa perfeita (Sicsú e Dana, 2012).

De acordo com Kazmier (1982, p.304) as hipóteses para a correlação linear são:

- a relação entre as duas variáveis é linear,
- ambas as variáveis são aleatórias,
- para cada variável, as variâncias condicionais, são iguais (*homoscedasticidade*),
- para cada variável, as distribuições condicionais são todas distribuições normais.

Conforme Karmel e Polasek (1972) um alto grau de correlação linear entre as variáveis não pode ser usado como evidência de causa e efeito, a menos que tenha sido cuidadosamente investigada a possibilidade desse coeficiente de correlação ser devido a ação de outras variáveis conhecidas. E o melhor procedimento é usar como prova uma relação de causa e efeito que pode ser testada com base no intervalo dos dados observados.

A análise de *correlação linear* dá um número que *resume o grau de relacionamento* entre duas variáveis. A análise de *correlação* é útil para trabalho exploratório, quando um pesquisador procura determinar quais variáveis são potencialmente importantes e o interesse está basicamente no *grau ou força do relacionamento*.

O *Coefficiente de Correlação Linear de Pearson* ( $r$ ), que mede a força de uma relação linear entre duas variáveis é dado a seguir, segundo Toledo e Ovalle (1985):

$$r = \frac{\sum X_i Y_i - \frac{(\sum X_i) \cdot (\sum Y_i)}{n}}{\sqrt{\left[ \sum X_i^2 - \frac{(\sum X_i)^2}{n} \right] \cdot \left[ \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} \right]}} \quad (9)$$

*Interpretação do grau de intensidade do Coeficiente de Correlação Linear de Pearson*

Conforme Crespo (2009) o grau de intensidade do Coeficiente de Correlação Linear de Pearson pode ser interpretado segundo a seguinte escala:

Se  $0,6 \leq |r| \leq 1 \Rightarrow$  Correlação Linear Forte entre as variáveis.

Se  $0,3 \leq |r| < 0,6 \Rightarrow$  Correlação Linear Moderada entre as variáveis.

Se  $0 < |r| < 0,3 \Rightarrow$  Correlação Linear Fraca entre as variáveis.

A magnitude do coeficiente de correlação indica quão mais próximo da reta estão os pontos individuais. A correlação está relacionada com a dispersão do conjunto de pontos em torno da reta. Quanto menor a dispersão dos pontos em torno da reta mais alto é o valor do coeficiente de correlação.

Para Stevenson (1981) um *coeficiente de correlação linear positivo* entre duas variáveis indica que valores altos (baixos) de uma variável corresponde a valores altos (baixos) da outra variável. Um *coeficiente de correlação linear negativo* entre duas variáveis indica que valores altos (baixos) de uma variável corresponde a valores baixos (altos) da outra variável. Um *coeficiente de correlação linear que se aproxima de zero ou é nulo* significa pouca ou nenhuma relação entre as variáveis.

## 5. Aplicação das Medidas Estatísticas e da Regressão e Correlação Linear

As *Medidas Estatísticas e a Regressão e Correlação Linear* serão aplicadas para analisar os dados de cidades dos Estados Unidos avaliadas em relação a Renda Familiar Anual e o Preço da Casa onde moram que estão no *Places Rated Almanac, 2000 (Almanaque das Localidades (Cidades) que foram Seleccionadas (Classificadas) no ano de 2000 nos Estados Unidos)*. As informações do *Places Rated Almanac, 2000* estão no livro *Estatística aplicada à administração e economia* do autor ANDERSON, D.R.; SWEENEY, D.J.; WILLIAMS, T.A. 2. ed. São Paulo: Cengage Learning, 2011, na página 116. A apresentação desses dados está a seguir na Tabela 1.

Tabela 1. Renda Familiar Anual (US\$) e o Preço da Casa (US\$) de cidades dos Estados Unidos

Cidade (Estados Unidos)	Renda Familiar Anual (US\$) (x1000)	Preço da Casa (US\$) (x1000)
----------------------------	---	------------------------------------

Washington, DC	97,4	191,9
Philadelphia, PA	87,0	151,5
Detroit, MI	85,3	145,0
Denver, CO	82,6	161,9
Atlanta, GA	82,4	126,9
Hartford, CT	89,1	162,1
Cleveland, OH	79,2	135,8
Madison, WI	78,8	145,2
Fort Lauderdale, FL	75,8	145,3
Nashville, TN	77,3	125,9

Fonte: *Places Rated Almanac, 2000. Estados Unidos*. Esses dados estão no livro *Estatística aplicada à administração e economia* do autor: Anderson, Sweeney e Williams (2011) na 2ª edição. São Paulo: Cengage Learning, na página 116.

## 5.1. Medidas Estatísticas na Análise da Renda Familiar Anual e o Preço da Casa

### a) Média da Renda Familiar Anual e do Preço da Casa

#### . Média Aritmética da Renda Familiar Anual.

Com a Equação (1) e os valores da Tabela 1. tem-se que:

$$\bar{X}_R = \frac{97,4+87,0+85,3+ \dots +78,8.+75,8+77,3}{10} = \frac{834,9}{10} = 83,49 \quad (\times 1000)$$

Logo a média aritmética da renda familiar anual é:  $\bar{X}_R = \text{US\$ } 83.490,00$ :

A Renda Familiar Anual que mais se aproxima do valor Média Aritmética de  $\bar{X}_R = \text{US\$ } 83.490,00$  são as cidades: Detroit (US\$ 85.300,00), Denver (US\$ 82.600,00) e Atlanta (US\$ 82.400,00). Essas são, portanto, as localidades com valores da Renda Familiar Anual mais próximos da média aritmética.

. *Média Aritmética do Preço da Casa*

Com os valores da Tabela 1. e a Equação (1) determina-se que:

$$\bar{x}_P = \frac{191,9 + 151,5 + 145,0 + \dots + 145,2 + 145,3 + 125,9}{10} = \frac{1.491,5}{10} = 149,15 \quad (\times 1000)$$

Portanto a média aritmética do preço da casa é:  $\bar{x}_P = \text{US\$}149.150,00$

O Preço da Casa que mais se aproxima do valor da Média Aritmética de  $\bar{x}_P = \text{US\$}149.150,00$  são as cidades: Philadelphia (US\$ 151.500,00), Madison (US\$ 145.200,00) e Fort Lauderdale (US\$ 145,300,00). Essas são, portanto, as localidades com valores do Preço da Casa mais próximos da média aritmética.

Significa que o Preço Médio da Casa nos Estados Unidos, nessa amostra de 10 cidades, nem sempre são dos proprietários com Renda Familiar Anual média.

b) *Mediana da Renda Familiar Anual e do Preço da Casa*

. *Mediana da Renda Familiar Anual*

Dada a definição da Mediana para número par de dados:

- Se  $n$  for par, a Mediana será a Média entre os elementos centrais (de ordem  $\frac{n}{2}$  e  $\frac{n}{2} + 1$ ).

Com os valores da *Renda Familiar Anual* da Tabela 1. e colocando-os em ordem crescente fica:

Dados em Ordem Crescente: 75,8; 77,3; 78,8; 79,2; 82,4; 82,6; 85,3; 87,0; 89,1; 97,4



Sendo  $n = 10$  (par), tem-se que:  $\frac{n}{2} = \frac{10}{2} = 5$  e que  $\frac{n}{2} + 1 = \frac{10}{2} + 1 = 6$ , ou seja,

os elementos: 82,4 e 82,6 . Logo a Mediana da *Renda Familiar Anual* é:

$$\text{Mediana: } Md_R = \frac{82,4 + 82,6}{2} = \frac{165}{2} = 82,50 \quad (\times 1000) \quad Md_R = \text{US\$ } 82.500,00$$

A *Renda Familiar Anual* que mais se aproxima da Mediana  $Md_R = \text{US\$ } 82.500,00$

são as cidades: Denver (US\$ 82.600,00) e Atlanta (US\$ 82.400,00) com valor muito próximo da mediana.

### . *Mediana do Preço da Casa*

Com a definição da Mediana para número par de dados tem-se que:

Dados em Ordem Crescente: 125,9; 126,9; 135,8; 145,0; 145,2; 145,3; 151,5; 161,9; 162,1; 191,9

Sendo  $n = 10$  (par), tem-se que:  $\frac{n}{2} = \frac{10}{2} = 5$  e que  $\frac{n}{2} + 1 = \frac{10}{2} + 1 = 6$ , ou seja,

os elementos: 145,2 e 145,3. Logo a Mediana do *Preço da Casa* é:

$$\text{Mediana: } Md_P = \frac{145,2 + 145,3}{2} = \frac{290,5}{2} = 145,25 \quad (\times 1000) \quad Md_P = \text{US\$ } 145.250,00$$

O Preço da Casa que mais se aproxima da Mediana  $Md_P = \text{US\$ } 145.250,00$  são as cidades: Madison (US\$ 145.200,00) e Fort Lauderdale (US\$ 145.300,00).

E que a cidade de Madison corresponde também a um dos valores próximos da Média Aritmética que foi demonstrado no item anterior. Confirmando que tanto a Média Aritmética como a Mediana são medidas de tendência central.

### c) *Desvio Padrão Amostral da Renda Familiar Anual e do Preço da Casa*

#### . *Desvio Padrão Amostral da Renda Familiar Anual*

Com a Equação (5) e os valores da Tabela 1. calcula-se:

$$S_R = \sqrt{\frac{(97,4-83,49)^2 + (97,4-83,49)^2 + \dots + (75,8-83,49)^2 + (77,3-83,49)^2}{10-1}}$$

$$S_R = \sqrt{\frac{380,389}{9}} = \sqrt{42,26544444} = 6,501187926 \quad (\times 1000) \quad S_R = \text{US\$ } 6.501,19$$

Logo o desvio padrão da renda familiar anual é:  $S_R = \text{US\$ } 6.501,19$

*Interpretação do desvio padrão pelas Regras Empíricas:*

Para a Renda Familiar Anual com Média Aritmética  $\bar{x}_R = \text{US\$ } 83.490,00$

monta-se o intervalo  $\bar{x}_R \pm S_R$ :

$$\bar{x}_R \pm S_R = 83.490,00 \pm 6.501,19 = (76.988,81 ; 89.991,19)$$

Com os dados da Tabela 1. conclui-se que no intervalo entre US\$ 76.988,81 e US\$ 89.991,19 estão 8 cidades com *Renda Familiar Anual* nesta faixa, ou seja,  $(8/10) \times 100 = 80\%$  das observações. Isto é: o intervalo compreendido entre a média *menos um desvio padrão* e a média *mais um desvio padrão* contém 80% de todos os valores da Renda Familiar Anual, que corresponde a 8 cidades que estão na faixa entre US\$ 76.988,81 e US\$ 89.991,19. O resultado confirma a Regra Empírica que o referido intervalo deveria conter de 60% a 80% das observações.

Para a Renda Familiar Anual com Média Aritmética  $\bar{x}_R = \text{US\$ } 83.490,00$  e

Desvio Padrão Amostral  $S_R = \text{US\$ } 6.501,19$  constrói-se o intervalo  $\bar{x}_R \pm 2S_R$ :

$$\bar{x}_R \pm 2S_R = 83.490,00 \pm 2(6.501,19) = (70.487,62 ; 96.492,38)$$

Com os dados da Tabela 1. conclui-se que no intervalo entre US\$ 70.487,62 e US\$ 96.492,38 estão 9 cidades com *Renda Familiar Anual* nesta faixa, ou seja,  $(9/10) \times 100 = 90\%$  das observações. Isto é: o intervalo compreendido entre a média *menos duas vezes o desvio padrão* e a média *mais duas vezes o desvio padrão* contém 90% de todos os valores da Renda Familiar Anual, que corresponde a 9 cidades que estão na faixa entre US\$ 70.487,62 e US\$ 96.492,38. O resultado confirma a Regra Empírica que o referido intervalo deveria conter aproximadamente 95% das observações.

. *Desvio Padrão Amostral do Preço da Casa*

Com os valores da Tabela 1. e a Equação (5) determina-se:

$$S_P = \sqrt{\frac{(191,9-149,15)^2 + (151,5-149,15)^2 + \dots + (145,3-149,15)^2 + (125,9-149,15)^2}{10-1}}$$

$$S_P = \sqrt{\frac{3.424,845}{9}} = \sqrt{380,5383333} = 19,50739176 \quad (\times 1000) \quad S_P = \text{US\$ } 19.507,39$$

Portanto o desvio padrão do preço da casa é:  $S_P = \text{US\$ } 19.507,39$

*Interpretação do desvio padrão pelas Regras Empíricas:*

Para o Preço da Casa com Média Aritmética  $\bar{X}_P = \text{US\$ } 149.150,00$

tem-se o intervalo  $\bar{X}_P \pm S_P$ :

$$\bar{X}_P \pm S_P = 149.150,00 \pm 19.507,39 = (129.642,61 ; 168.657,39)$$

Com os dados da Tabela 1. conclui-se que no intervalo entre US\$ 126.642,61 e US\$ 168.657,39 estão 7 cidades com *Preço da Casa* nesta faixa, ou seja,  $(7/10) \times 100 = 70\%$  das observações. Isto é: o intervalo compreendido entre a média *menos um desvio padrão* e a média *mais um desvio padrão* contém 70% de todos os valores do Preço da Casa, que corresponde a 7 cidades que estão na faixa entre US\$ 126.642,61 e US\$

168.657,39. O resultado confirma a Regra Empírica que o referido intervalo deveria conter de 60% a 80% das observações.

Para o Preço da Casa com Média Aritmética  $\bar{x}_R = \text{US\$ } 149.150,00$  e

Desvio Padrão Amostral  $s_R = \text{US\$ } 19.507,39$  tem-se o intervalo  $\bar{x}_P \pm 2s_P$ :

$$\bar{x}_P \pm 2s_P = 149.150,00 \pm 2(19.507,39) = (110.135,22 ; 188.164,78)$$

Com os dados da Tabela 1. conclui-se que no intervalo entre US\$ 110.135,22 e US\$ 188.164,78 estão 9 cidades com Preço da Casa nesta faixa, ou seja,  $(9/10) \times 100 = 90\%$  das observações. Isto é: o intervalo compreendido entre a média *menos duas vezes o desvio padrão* e a média *mais duas vezes o desvio padrão* contém 90% de todos os valores do Preço da Casa, que corresponde a 9 cidades que estão na faixa entre US\$ 110.135,22 e US\$ 188.164,78. O resultado confirma a Regra Empírica que o referido intervalo deveria conter aproximadamente 95% das observações.

## 5.2. Análise da Regressão Linear da Renda Familiar Anual e o Preço da Casa

A Equação da Reta de Mínimos Quadrados (Regressão Linear) pode ser obtida utilizando-se da construção dos valores da Tabela 2. que facilita os cálculos dos coeficientes  $a$  e  $b$  para montagem da equação estimada da Reta:

Tabela 2. Valores para obter a Equação da Reta de Mínimos Quadrados

Cidade (Estados Unidos)	Renda Familiar Anual (US\$)	Preço da Casa (US\$)			
	$x_i$	$y_i$	$x_i^2$	$x_i \cdot y_i$	$y_i^2$
	(x1.000)	(x1.000)	(x1.000.000)	(x1.000.000)	(x1.000.000)

					000)
Washington, DC	97,4	191,9	9.486,76	18.691,06	36.825,61
Philadelphia, PA	87,0	151,5	7.569,00	13.180,50	22.952,25
Detroit, MI	85,3	145,0	7.276,09	12.368,50	21.025,00
Denver, CO	82,6	161,9	6.822,76	13.372,94	26.211,61
Atlanta, GA	82,4	126,9	6.789,76	10.456,56	16.103,61
Hartford, CT	89,1	162,1	7.938,81	14.443,11	26.276,41
Cleveland, OH	79,2	135,8	6.272,64	10.755,36	18.441,64
Madison, WI	78,8	145,2	6.209,44	11.441,76	21.083,04
Fort Lauderdale, FL	75,8	145,3	5.745,64	11.013,74	21.112,09
Nashville, TN	77,3	125,9	5.975,29	9.732,07	15.850,81
Σ (Total)	834,9	1.491,5	70.086,19	125.455,60	225.882,07

Fonte: Elaborado pelo autor

A Equação (6) que é a Equação da Reta de Mínimos Quadrados (*Regressão Linear Simples*) dada por  $\hat{y} = a + bX_i$  precisa do cálculo dos coeficientes  $a$  e  $b$  para se estabelecer o seu equacionamento.

Determina-se o coeficiente  $b$  com os dados da Tabela 2. e a Equação (7). Os cálculos estão a seguir:

$$b = \frac{125.455,60 - \frac{(834,9) \cdot (1.491,5)}{10}}{70.086,19 - \frac{(834,9)^2}{10}} = \frac{930,265}{380,389} = 2,445562306 \quad \text{Logo: } b = 2,4456$$

Com o coeficiente  $b = 2,4456$ , os dados da Tabela 2. e a Equação (8) o cálculo do coeficiente  $a$  é assim determinado:

$$a = \frac{1.491,50}{10} - 2,4456 \cdot \frac{834,9}{10} = -55,033144 \quad \text{Então: } a = -55,03$$

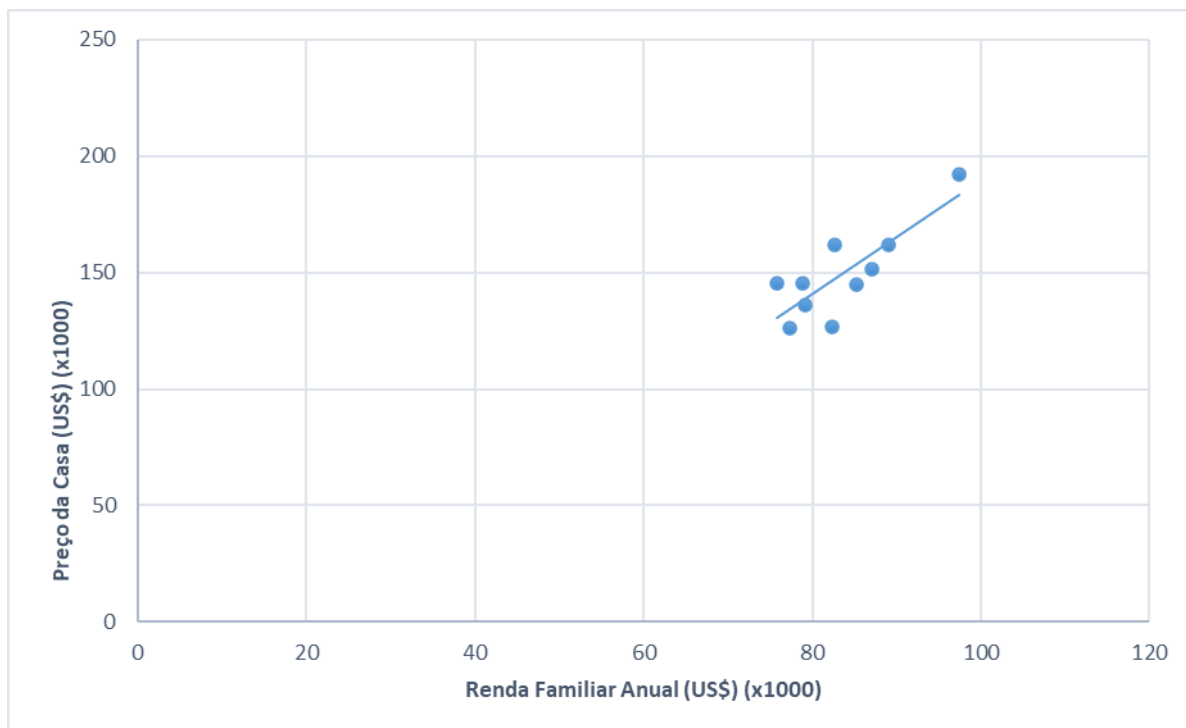
Portanto, com valores em dólares (US\$), a Equação da Reta estimada è:

$$\hat{Y} = -55,03 + 2,4456 X_i \quad (\times 1000).$$

O significado da Equação da Reta de Mínimos Quadrados  $\hat{Y} = -55,03 + 2,4456 X_i$  ( $\times 1000$ ) com valores em dólares (US\$), indica que o coeficiente 2,4456 ( $\times 1000$ ) representa, que para Cada Unidade de Variação Positiva da Renda Familiar Anual (X), o Preço da Casa (Y) cresce US\$ 2.445,60.

A representação da *Reta de Mínimos Quadrados (Regressão Linear)* e a Dispersão (“nuvem” de pontos) dos valores da Renda Familiar Anual e o Preço da Casa estão no Gráfico 1.

Gráfico 1. Dispersão (“nuvem” de pontos) da Renda Familiar Anual e o Preço da Casa e a Reta Ajustada de Mínimos Quadrados (Regressão Linear)



Fonte: Elaborado pelo autor

Ao testar o modelo matemático (estatístico) obtido pela Equação da Reta de Mínimos

Quadrados, com valores em dólares (US\$):  $\hat{Y} = -55,03 + 2,4456 X$  (x1000) e ao utilizar por exemplo um valor de US\$ 90 (x1000) de Renda Familiar Anual obtém-se:

Para  $x = 90$  substituindo na equação:  $\hat{Y} = -55,03 + 2,4456 X_i$  (x1000)

logo  $\hat{Y} = -55,03 + 2,4456 \cdot (90) = 165,074$  (x1000) = 165.074,00

Conclui-se que o valor estimado do Preço da Casa será de US\$ 165.074,00 para a Renda Familiar Anual de US\$ 90 (x1000). Esse resultado está compatível com os valores da Tabela 2., pois a tabela mostra que para o valor da Renda Familiar Anual igual a US\$ 89,1 (x1000) apresenta o valor do Preço da Casa de US\$ 162.100,00, ou seja, o modelo matemático (estatístico) da Equação da Reta se apresenta bem ajustado aos pontos.

### 5.3. Correlação Linear na Análise da Renda Familiar e o Preço da Casa

Para se obter o *Coefficiente de Correlação Linear de Pearson* ( $r$ ), utilizam-se os dados da Tabela 2. e a Equação (9):

O valor do coeficiente de correlação linear de Pearson, é calculado como segue:

$$r = \frac{1225.455,60 - \frac{(834,9)(1.491,5)}{10}}{\sqrt{\left[70.086,19 - \frac{(834,9)^2}{10}\right] \left[225.882,07 - \frac{(1.491,5)^2}{10}\right]}} = \frac{930,265}{\sqrt{[380,389][3.424,845]}}$$

$$r = \frac{930,265}{\sqrt{1.302.773,365}} = \frac{930,265}{1.141,390978} = +0,815027469$$

Portanto o Coeficiente de Correlação Linear de Pearson é  $r = +0,815$ .

Pelo *Coefficiente de Correlação Linear de Pearson* obtido, conclui-se que o valor  $r = +0,815$  pertence ao intervalo  $0,6 \leq |r| \leq 1$ , que significa uma *Correlação Linear Forte* entre as variáveis e que indica uma boa semelhança dos comportamentos das variáveis.

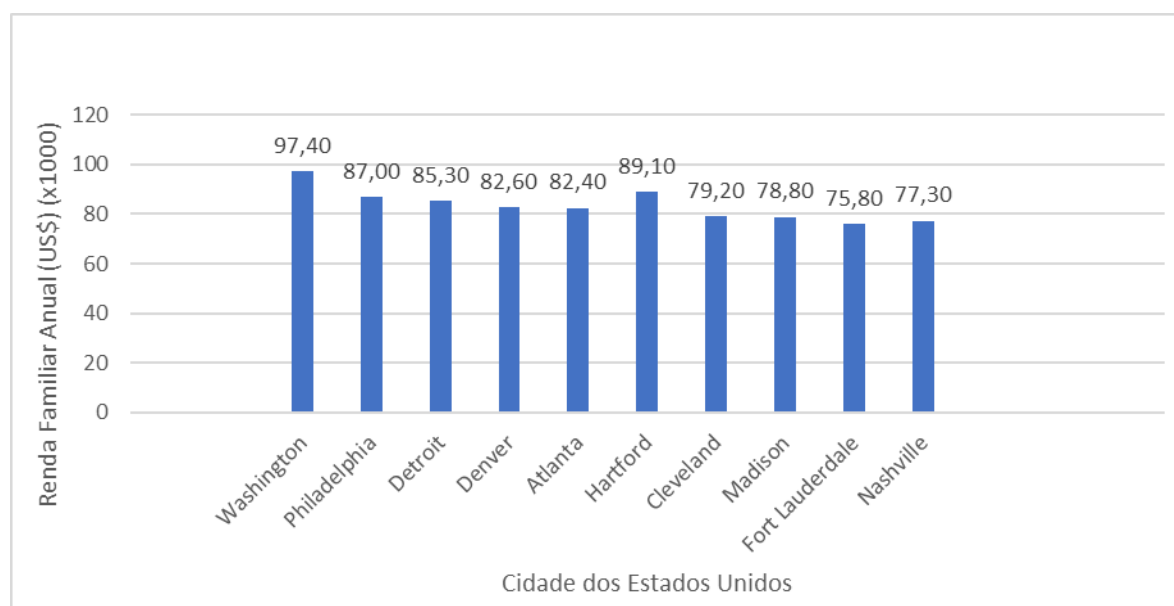
*Interpretação do grau de intensidade do Coeficiente de Correlação Linear de Pearson:*

O grau de intensidade de 81,50% do *Coefficiente de Correlação Linear de Pearson* mostra um resultado muito significativo. Comprova que há uma *Correlação Linear Forte* entre as variáveis.

O sinal positivo (+) do *Coefficiente de Correlação Linear* mostra que as duas variáveis são ascendentes (crescentes) no mesmo sentido de forma diretamente proporcional. E indica que valores altos (baixos) de uma variável corresponde a valores altos (baixos) da outra variável. Confirma-se que há uma *Relação Linear Positiva Forte entre a Renda Familiar Anual e o Preço da Casa*. Mais especificamente um aumento da *Renda Familiar Anual* está associada a um aumento do *Preço da Casa*.

**A representação de como se comporta o valor da Renda Familiar Anual (com os valores da Tabela 1.) na amostra das cidades dos Estados Unidos está mostrado no Gráfico 2.**

Gráfico 2. Evolução da Renda Familiar Anual nas cidades dos Estados Unidos



Fonte: Elaborado pelo autor

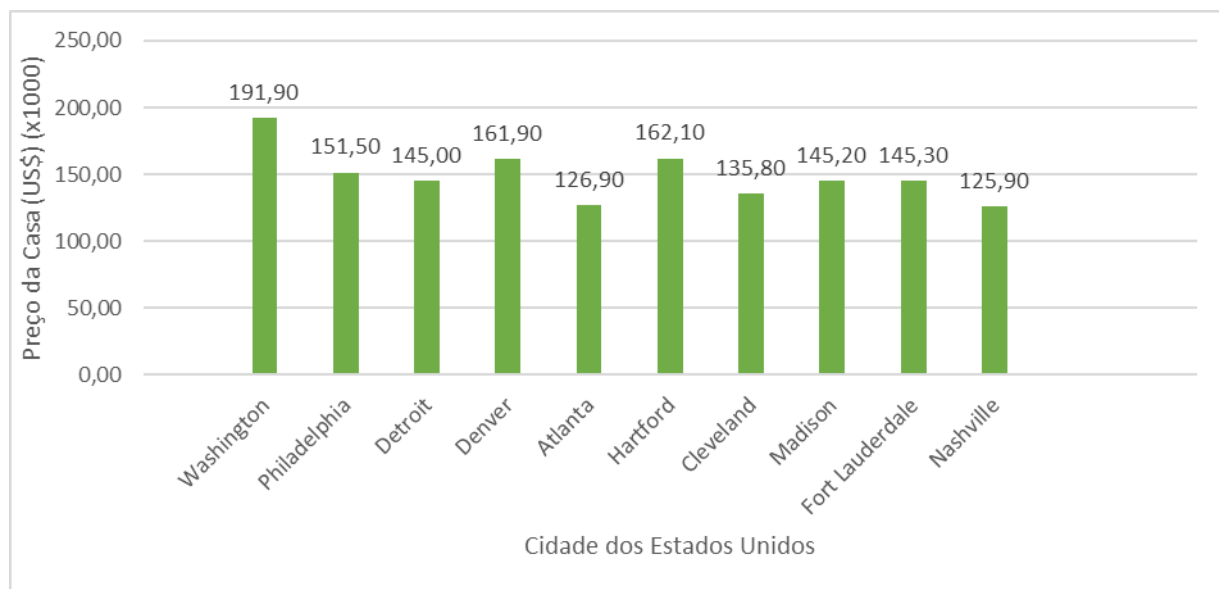


A Renda Familiar Anual no Gráfico 2. mostra que os maiores valores são em Washington (US\$ 97.400,00), Hartford (US\$ 89.100,00) e Philadelphia (US\$ 87.000,00). E que os menores valores de Renda Familiar Anual são visualizados no gráfico e que correspondem a Madison (US\$ 78.800,00), Nashville (US\$ 77.300,00 e Fort Lauderdale (US\$ 75.800,00).

Verifica-se também que há uma certa estabilidade, com variação pouco significativa da Renda Familiar Anual entre a Philadelphia (US\$ 87.000,00), Detroit (US\$ 85.300,00), Denver (US\$ 82.600,00) e Atlanta (US\$ 82.400,00) com variação absoluta máxima na ordem de 5,6% na Renda Familiar Anual. E que essa estabilidade ocorre ainda entre Cleveland (US\$ 79.200,00), Madison (US\$ 78.800,00), Fort Lauderdale (US\$ 75.800,00) e Nashville (US\$ 77.300,00) variando com valor absoluto máximo de 4,5% em relação a Renda Familiar Anual, ou seja, esses valores se distribuem de forma uniforme praticamente.

**A representação de como se comporta o valor do Preço da Casa (com os valores da Tabela 1.) na amostra das cidades dos Estados Unidos está mostrado no Gráfico 3.**

Gráfico 3. Evolução do Preço da Casa nas cidades dos Estados Unidos



A Evolução do Preço da Casa mostrado no Gráfico 3. é nítido a verificação dos maiores valores sendo na capital Washington (US\$ 191.900,00), Denver (US\$ 161.900,00) e Hartford (US\$ 162.100,00). E que os menores valores do Preço da Casa são Atlanta (US\$ 126.900,00), Cleveland (US\$ 135.800,00) e Nashville (US\$ 125.900,00).

Conclui-se também que há uma certa estabilidade, com variação pouco significativa do Preço da Casa entre a Philadelphia (US\$ 151.500,00), Detroit (US\$ 145.000,00) com variação absoluta da ordem de 4,5% no Preço da Casa. E que essa estabilidade ocorre ainda entre Madison (US\$ 145.200,00), Fort Lauderdale (US\$ 145.300,00), com valores praticamente idênticos, variando com valor absoluto de 0,07% em relação ao Preço da Casa, ou seja, esses valores se distribuem de forma uniforme.

Comprova-se que as cidades de Washington e Hartford possuem os maiores valores em Renda Familiar Anual e em Preço das Casas.

E que o menor valor tanto como em Renda Familiar Anual e no Preço das Casas é a cidade de Nashville.

A estabilidade da Renda Familiar Anual e do Preço das Casas ficaram por conta das cidades de Philadelphia, Detroit, Madison e Fort Lauderdale.

## 6. Resultados e Conclusões

As Medidas Estatísticas calculadas, da Renda Familiar Anual e do Preço da Casa, obtidas no item 5.1 estão resumidas na Tabela 3.

Tabela 3. Resumo das Medidas Estatísticas

Medidas Estatísticas	Renda Familiar Anual	Preço da Casa
Média Aritmética	$\bar{X}_R = \text{US\$ } 83.490,00$	$\bar{X}_P = \text{US\$ } 149.150,00$
Mediana	$Md_R = \text{US\$ } 82.500,00$	$Md_P = \text{US\$ } 145.250,00$
Desvio Padrão	$S_R = \text{US\$ } 6.501,19$	$S_P = \text{US\$ } 19.507,39$

*Fonte: Elaborado pelo autor*

### *Conclusões da Média Aritmética e da Mediana na variação da Renda Familiar Anual*

A Média Aritmética da Renda Familiar Anual  $\bar{x}_R = \text{US\$ } 83.490,00$  e a Mediana  $Md_R = \text{US\$ } 82.500,00$  oscilaram em valor absoluto 1,20%, significando que esses valores de Medidas de Tendência Central estão muito próximos validando o estudo e que confirma a tendência ao centro das medidas. E que o valor de Renda Familiar Anual de  $\text{US\$ } 82.600,00$  é o que mais se aproxima da Média Aritmética com margem de variação absoluta 1,08%, e da Mediana com valor de variação absoluta de 0,12%. Sendo assim a cidade de Denver com  $\text{US\$ } 82.600,00$  de Renda Familiar Anual é o valor com melhor tendência central das 10 cidades estudadas dos Estados Unidos. Significa que as famílias que vivem em Denver têm o valor com melhor tendência central da Renda Familiar Anual na amostra das 10 cidades estudadas dos Estados Unidos.

### *Conclusões da Média Aritmética e da Mediana na variação do Preço da Casa.*

O valor da Aritmética do Preço da Casa  $\bar{x}_P = \text{US\$ } 149.150,00$  e o da Mediana  $Md_P = \text{US\$ } 145.250,00$  variam com valor absoluto em torno de 2,69%, o que significa esses valores de Medidas de Tendência Central estão muito próximos validando o estudo, e que confirma a tendência ao centro das medidas. E que o valor do Preço da Casa  $\text{US\$ } 145.300,00$  é o que mais se aproxima da Média Aritmética com margem de valor absoluto 2,65%, e da Mediana com valor de variação absoluta de 0,0344%. Sendo assim a cidade de Fort Lauderdale com o Preço da Casa de  $\text{US\$ } 145.300,00$  é o valor com melhor tendência central da amostra das 10 cidades dos Estados Unidos. Significando que as famílias que vivem em Fort Lauderdale possuem o valor com melhor tendência central do Preço da Casa nos Estados Unidos em relação a amostra das 10 cidades estudadas.

### *Conclusões da variação do Desvio Padrão em torno da Média Aritmética da Renda Familiar Anual:*

O Intervalo da *Renda Familiar Anual* compreendido entre a média *menos* um desvio padrão e a média *mais* um desvio padrão é de:

$$\bar{x}_R \pm S_R = 83.490,00 \pm 6.501,19 = (76.988,81 ; 89.991,19)$$

A Tabela 4. com os valores em Ordem Crescente da Renda Familiar Anual mostra que as localidades que estão no Intervalo de US\$ 76.988,81 até US\$ 89.991,19 são as cidades: Nashville (US\$ 77.300,00), Madison (US\$ 78.800,00), Cleveland (US\$ 79.200,00), Atlanta (US\$ 82.400,00), Denver (US\$ 82.600,00), Detroit (US\$ 85.300,00), Philadelphia (US\$ 87.000,00) e Hartford (US\$ 89.100,00). Isto é: o intervalo compreendido entre a média *menos um desvio padrão* e a média *mais um desvio padrão* contém 80% de todos os valores da Renda Familiar Anual, que corresponde a 8 cidades que estão na faixa entre US\$ 76.988,81 e US\$ 89.991,19.

Tabela 4. Valores em Ordem Crescente da Renda Familiar Anual e o Preço da Casa

Cidade (Estados Unidos)	Renda Familiar Anual (US\$)	<p><i>Fonte:</i> <i>Elaborado</i> <i>pelo autor</i></p> <p>O intervalo da Renda Familiar Anual compreendido entre a média <i>menos</i> duas vezes o desvio padrão e a</p>	Cidade (Estados Unidos)	Preço da Casa (US\$)	
Fort Lauderdale, FL	75,8			Nashville, TN	125,9
Nashville, TN	77,3			Atlanta, GA	126,9
Madison, WI	78,8			Cleveland, OH	135,8
Cleveland, OH	79,2			Detroit, MI	145,0
Atlanta, GA	82,4			Madison, WI	145,2
Denver, CO	82,6			Fort Lauderdale, FL	145,3
Detroit, MI	85,3			Philadelphia, PA	151,5
Philadelphia, PA	87,0			Denver, CO	161,9
Hartford, CT	89,1			Hartford, CT	162,1
Washington, DC	97,4		Washington, DC	191,9	

média *mais* duas vezes o desvio padrão é de:

$$\bar{x}_R \pm 2S_R = 83.490,00 \pm 2(6.501,19) = (70.487,62 ; 96.492,38)$$

Nesse Intervalo da Renda Familiar Anual de US\$ 70.487,62 até US\$ 96.492,38 com exceção da capital Washington (US\$ 97.400,00), sendo que este valor de US\$ 97.400,00 é um valor discrepante (*outlier*), ou seja, foge a dimensão esperada, todas as outras cidades estão dentro desse intervalo.

Conforme a Tabela 4 essas cidades são: Fort Lauderdale (US\$ 75.800,00), Nashville (US\$ 77.300,00), Madison (US\$ 78.800,00), Cleveland (US\$ 79.200,00), Atlanta (US\$ 82.400,00), Denver (US\$ 82.600,00), Detroit (US\$ 85.300,00), Philadelphia (US\$ 87.000,00) e Hartford (US\$ 89.100,00). Isto é: o intervalo compreendido entre a média *menos duas vezes o desvio padrão* e a média *mais duas vezes o desvio padrão* contém 90% de todos os valores da Renda Familiar Anual, que corresponde a 9 cidades que estão na faixa entre US\$ 70.487,62 e US\$ 96.492,38.

*Conclusões da variação do Desvio Padrão em torno da Média Aritmética do Preço da Casa:*

O Intervalo da *Preço da Casa* compreendido entre a média *menos* um desvio padrão e a média *mais* um desvio padrão é de:

$$\bar{x}_p \pm S_p = 149.150,00 \pm 19.507,39 = (129.642,61 ; 168.657,39)$$

Na Tabela 4. com os valores em Ordem Crescente do Preço da Casa as localidades que estão no Intervalo de US\$ 129.642,61 até US\$ 168.657,39 são as cidades: Cleveland (US\$ 135.800,00), Detroit (US\$ 145.000,00), Madison (US\$ 145.200,00), Fort Lauderdale (US\$ 145.300,00), Philadelphia (US\$ 151.500,00), Denver (US\$ 161.900,00) e Hartford (US\$ 162.100,00). Isto é: o intervalo compreendido entre a média *menos um desvio padrão* e a média *mais um desvio padrão* contém 70% de todos os valores do Preço da Casa, que corresponde a 7 cidades que estão na faixa entre US\$ 126.642,61 e US\$ 168.657,39.

O intervalo compreendido entre a média *menos* duas vezes o desvio padrão e a média *mais* duas vezes o desvio padrão é de:

$$\bar{x}_p \pm 2S_p = 149.150,00 \pm 2(19.507,39) = (110.135,22 ; 188.164,78)$$

Nesse Intervalo o Preço da Casa de US\$ 110.135,22 até US\$ 188.164,78 com exceção da capital Washington, DC (US\$ 191.900,00), sendo que este valor de US\$ 191.900,00 é um valor discrepante (*outlier*), ou seja, foge a dimensão esperada, todas as outras cidades estão dentro desse intervalo.

Conforme a Tabela 4. estas cidades são: Nashville (US\$ 125.900,00), Atlanta (US\$ 126.900,00), Cleveland (US\$ 135.800,00), Detroit (US\$ 145.000,00), Madison (US\$ 145.200,00), Fort Lauderdale (US\$ 145.300,00), Philadelphia (US\$ 151.500,00), Denver (US\$ 161.900,00) e Hartford (US\$ 162.100,00). Isto é: o intervalo compreendido entre a média *menos duas vezes o desvio padrão* e a média *mais duas vezes o desvio padrão* contém 90% de todos os valores do Preço da Casa, que corresponde a 9 cidades que estão na faixa entre US\$ 110.135,22 e US\$ 188.164,78

*Conclusões sobre a Equação da Reta estimada com os dados da Renda Familiar Anual e o Preço da Casa obtida no item 5.2.*

A Equação da Reta de Mínimos Quadrados obtida no item 5.2. é  $\hat{Y} = -55,03 + 2,4456X_i$  (x1000), com valores em dólares (US\$).

A avaliação e análise dessa Equação mostra que, por exemplo, se o Preço da Casa (Y) for nulo, ou seja,  $0 = -55,03 + 2,4456X_i$  (x1000 US\$), o valor da Renda Familiar Anual (X) é de aproximadamente  $X = \text{US\$ } 22,50$  (x1000) significa que a Renda Familiar Anual tem um valor estimado de US\$ 22.500,00, ou seja, as famílias, dentre as 10 cidades estudadas dos Estados Unidos, têm uma Renda Anual mínima (patamar mínimo de sobrevivência).de US\$ 22.500,00.

O acréscimo estimado do Preço da Casa (Y) é de US\$ 2.445,60 para cada unidade de variação positiva da Renda Familiar Anual (X), por exemplo, se a Renda Familiar Anual for  $X=78$  (US\$ 78.000) e passar para  $X=79$  (US\$ 79.000), tem-se que

Para  $x= 78$  substituindo na equação:  $\hat{Y} = -55,03 + 2,4456X_i$  (x1000)

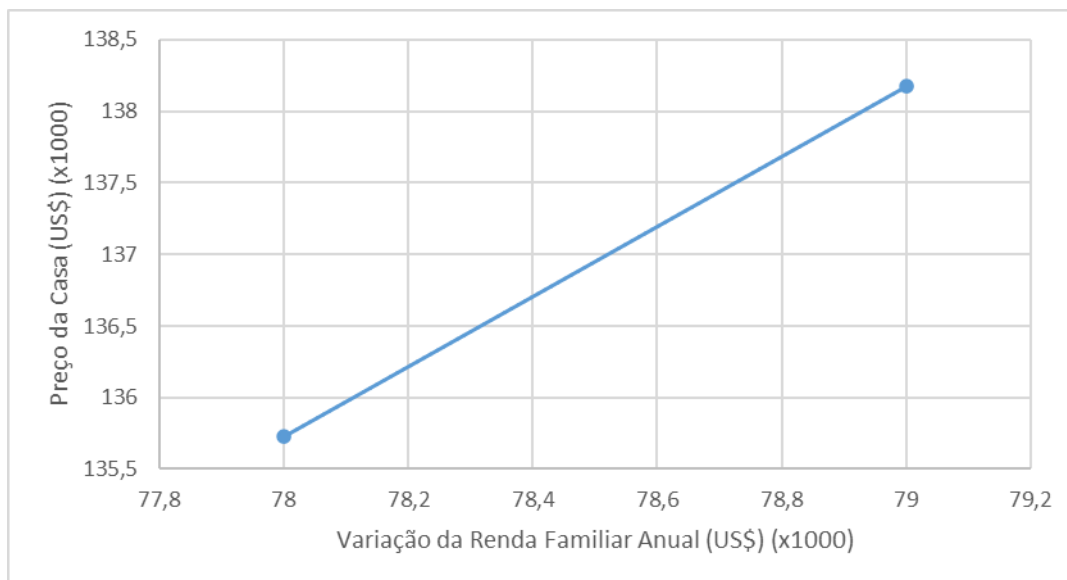
logo  $\hat{Y} = -55,03 + 2,4456 \cdot (78) = 135,7268$  (x1000) = 135.726,80

Para  $x= 79$  substituindo na equação:  $\hat{Y} = -55,03 + 2,4456X_i$  (x1000)

logo  $\hat{Y} = -55,03 + 2,4456 \cdot (79) = 138,1724$  (x1000) = 138.172,40.

O Gráfico 4. mostra a Reta de Regressão Linear estimada na variação da Renda Familiar de US\$ 78.000,00 para US\$ 79.000,00 e conseqüentemente a variação do Preço da Casa de US\$ 135.726,80 para US\$ 138.172,40.

Gráfico 4. Reta de Regressão estimada para unidade de variação positiva da Renda Familiar Anual



*Fonte: Elaborado pelo autor*

O valor da Renda Familiar Anual que passou de US\$ 78.000,00 para US\$ 79.000,00 gerou respectivamente o valor do Preço da Casa de US\$ 135.726,80 para US\$ 138.172,40. Esse resultado de variação do Preço da Casa confirma o valor acrescido de US\$ 2.445,60, ou seja,  $US\$ 138.172,40 - US\$ 135.726,80 = US\$ 2.445,60$ . Confirma-se que para cada unidade de variação positiva da Renda Familiar Anual o acréscimo estimado do Preço da Casa é de US\$ 2.445,60.

*Conclusões sobre o Coeficiente de Correlação Linear de Pearson obtido com os dados da Renda Familiar Anual e do Preço da Casa obtidas no item 5.3.*

O Coeficiente de Correlação Linear de Pearson que foi calculado no item 5.3.entre os valores da Renda Familiar Anual e o Preço da Casa para uma amostra de 10 cidades dos Estados Unidos é  $r = +0,815$  (81,50%). Conclui-se que o valor  $r = +0,815$  pertence ao intervalo  $0,6 \leq |r| \leq 1$  e a Interpretação do grau de intensidade do Coeficiente de Correlação Linear de Pearson significa dizer que há um a *Correlação Linear é Forte entre as variáveis*, indicando uma boa semelhança dos comportamentos das variáveis e no mesmo sentido de forma proporcional.

O Coeficiente de Correlação Linear mede a excelência do ajustamento aos dados pela Equação de Reta. A Equação da Reta de Mínimos Quadrados obtida no item 5.2. que



é  $\hat{Y} = -55,03 + 2,4456 X_i$  (x1000), com valores em dólares (US\$), mostra o valor do coeficiente angular  $b = +2,4456$  (inclinação da reta ajustada), sendo um valor positivo (+). Verifica-se que o Coeficiente de Correlação Linear de Pearson  $r = +0,815$  também é positivo (+), Correlação Positiva ou Direta. Conclui-se então que, sendo o Coeficiente de Correlação Linear com sinal (+) indica uma Equação de Reta Crescente (Ascendente) e que é confirmado com o valor positivo (+) da inclinação da reta ajustada  $b = +2,4456$ .

Conclui-se que o comportamento da *Renda Familiar Anual (US\$)* e do *Preço da Casa (US\$)* é no mesmo sentido com grau de intensidade de 81,50% sendo muito significativo, e que existe uma forte associação linear positiva entre essas variáveis. E que as duas variáveis são crescentes no mesmo sentido de forma diretamente proporcional. Significa que um aumento da *Renda Familiar Anual (US\$)* está associado a um aumento do *Preço da Casa (US\$)*, ou de outra forma, o *Preço da Casa (US\$)* aumenta quando a *Renda Familiar Anual (US\$)* cresce.

Neste trabalho, analisou-se a aplicação das Medidas Estatísticas da média aritmética e da mediana, onde se obteve resultados satisfatórios sobre a evolução da Renda Familiar Anual e do Preço da Casa de uma amostra de 10 cidades dos Estados Unidos. Após a obtenção do conjunto de dados da Renda Familiar Anual e o Preço da Casas, o objetivo foi descobrir alguma propriedade entre esses dados. O conhecimento geral das características dos valores amostrais, mais o conhecimento específico, proporcionou analisar a relação entre a Renda Familiar Anual e o Preço da Casa. Localidades com valores médios mais próximos da média aritmética, da mediana foram definidas. Significando que o Preço Médio da Casa nos Estados Unidos nem sempre são dos proprietários com Renda Familiar Anual Média.

As Medidas de Dispersão ou Variabilidade apresentaram o intervalo compreendido entre a média *menos* um desvio padrão e a média *mais* um desvio padrão que obtiveram valores com 80% da Renda Familiar Anual e 80% do Preço da Casa. O resultado confirmou a Regra Empírica que o referido intervalo deveria conter de 60% a 80% das observações.

A correlação e a regressão linear se referem à estimação de relações entre pares de dados. A finalidade última da correlação linear é sintetizar o grau de relacionamento,

enquanto que o propósito da regressão é estabelecer uma equação matemática que descreva o relacionamento.

Pode-se analisar que as informações coletadas da amostra de dados da Renda Familiar Anual e o Preço da Casa de cidades dos Estados Unidos proporcionaram a estimação da equação da reta pela análise de Regressão Linear. Essa equação gerou um ajustamento quase perfeito aos dados amostrais que reproduziram de forma altamente confiável as características importantes dos dados. Na análise de Regressão Linear o modelo estatístico foi testado e, mesmo com a existência de *outliers* (valores discrepantes), os resultados gerados foram muito próximos do conjunto de valores coletados e que reproduziu de forma confiável a relação entre a Renda Familiar Anual e o Preço da Casa para a mostra das 10 cidades dos Estados Unidos.

Pode-se afirmar ainda, que na Correlação Linear o grau de relação linear existente entre as duas variáveis foi de relação forte entre as variáveis, que confirma o grau de aderência entre a Renda Familiar Anual e o Preço da Casa. Verifica-se uma boa semelhança dos comportamentos das variáveis e no mesmo sentido de forma proporcional o que valida de forma confiável este estudo com boa qualidade nos resultados.

Um ponto que não pode deixar de ser destacado refere-se ao gráfico de dispersão dos pontos com a estimativa da reta de regressão linear que mostraram de forma rápida pouca dispersão entre os valores, o que comprova uma boa relação linear entre as variáveis. Comprova-se assim, a variação proporcional da relação entre a Renda Familiar Anual e o Preço da Casa das cidades dos Estados Unidos que foram estudadas.

## **Considerações Finais**

O uso de modelos matemáticos constitui um aspecto importante na análise estatística.

Os fundamentos lógicos da utilização de modelos residem no fato de que eles simplificam problemas, eliminando detalhes não importantes. A representatividade da amostra obtida diretamente da fonte de dados reproduziu de forma altamente confiável o estudo desta pesquisa. Tomando como base o modelo estatístico da equação de regressão linear, que sendo testado produziu resultados satisfatórios dentro do intervalo do conjunto de dados. A equação de reta ajustada aos pontos pela regressão linear correspondeu de forma satisfatória aos dados da Renda Familiar Anual e do Preço da Casa e que foi confirmado com o valor do coeficiente de correlação que gerou uma relação forte entre as variáveis.

Os resultados apresentados revelaram que o objetivo geral deste trabalho foi alcançado. O objetivo de descrever, interpretar e obter conclusões que consistiu em identificar os métodos de análise estatística foram alcançados. No entanto, esses métodos foram tratados no presente trabalho como métodos de características matemáticas. Fica claro a relação entre os dados coletados no estudo com a bibliografia apresentada, que vem ao encontro dos objetivos gerais e específicos.

Com base nos resultados desta pesquisa, abrem-se possibilidades para novos estudos. Para o estudo futuro de análise estatística em um conjunto de dados, podem ser incorporadas técnicas de estatística multivariada que constituem uma extensão das técnicas de correlação e regressão. Trata-se de situações que envolvem três ou mais variáveis e incidem relações tanto lineares como curvilíneas. Essas técnicas, em particular a análise de regressão, constituem instrumentos poderosos, largamente usados em administração e economia.

Deve-se incluir nos próximos estudos, como sugestão, a estatística não-paramétrica que envolve testes de significância e estimação, usando métodos que em geral são rápidos e que exigem hipóteses. Esses métodos abrangem dados sob a forma de postos ou sinais e que costumam perder informação.

A realização desta pesquisa demonstrou que o ajustamento aos dados da Renda Familiar Anual e o Preço da Casa, evidencia resultados bem satisfatórios confirmados com o coeficiente correlação com forte relação entre as variáveis. Os resultados a que

se chegaram deixam campo aberto para outras pesquisas na mesma área, como construção de intervalos de confiança e testes de significância.

A estatística mostra-se, cada vez mais, como uma poderosa ferramenta para a análise e avaliação de dados, em todas as áreas do conhecimento. A estatística contribui, nos estudos realizados envolvendo dados numéricos, proporcionando resultados significativos que são usados para uma tomada de decisão com a maior precisão possível.

### **Referências**

ANDERSON, D.R.; SWEENEY, D.J.; WILLIAMS, T.A. Estatística aplicada à administração e economia. 2. ed. São Paulo: Cengage Learning, 2011.

BRUNI, A.L. Estatística aplicada à gestão empresarial. 4. ed. São Paulo: Atlas, 2017.

BUSSAB, W.O.; MORETTIN, P.A. Estatística básica. 9. ed. São Paulo: Saraiva, 2017.

CALLEGARI-JACQUES, S. M. Bioestatística: princípios e aplicações. Porto Alegre: Artmed, 2003.

COSTA, A.F.B.; EPPRECHT, E.K.; CARPINETTI, L.C.R. Controle estatístico de qualidade. 2. ed. São Paulo: Atlas, 2005.

COSTA, G.G.O. Estatística aplicada à informática e às suas novas tecnologias. v.2. Rio de Janeiro: Ciência e Moderna, 2015.

COSTA NETO, P.L. Estatística. 2. ed. São Paulo: Edgar Blücher, 2002.

CRESPO, A.A. Estatística fácil. 19. ed. São Paulo: Saraiva, 2009.

FONSECA, J.S.; MARTINS, G.A.; TOLEDO, G.L. Estatística aplicada. 2. ed. São Paulo: Atlas, 1985.

KARMEL, P.H., POLASEK, M. B. Estatística geral e aplicada para economistas. São Paulo: Atlas, 1972.

KAZMIER, L.J. Estatística aplicada a economia e administração. São Paulo: McGraw-Hill, 1982.

LAPPONI, J.C. Estatística usando Excel. 4. ed. Rio de Janeiro: Elsevier, 2005.

MARTINS, G.A.; DONAIRE, D. Princípios de estatística. 4. ed. São Paulo: Atlas, 1990.

MARTINS, G.A.; DOMINGUES, O. Estatística geral e aplicada. 6. ed. São Paulo: Atlas, 2019.

McCLAVE, J.T.; BENSON, P.G.; SINCICH, T. Estatística para administração e economia. 10. ed. São Paulo: Pearson Prentice Hall, 2009.

RAMOS, E.M.L.S.; ALMEIDA, S.S.; ARAÚJO, A.R. Controle estatístico da qualidade. Porto Alegre: Brookman, 2013.

RAMOS, A.W. CEP para processos contínuos e em bateladas. São Paulo: Edgar Blücher, 2000.

SILVA, E.M.; SILVA, E.M.; GONÇALVES, V.; MUROLO, A.C. Estatística. 4 ed. v.1. São Paulo: Atlas, 2010.

STEVENSON, W.J. Estatística aplicada à administração, São Paulo: Harper e Row do Brasil, 1981.

SPIEGEL, M.R. Estatística. 3. ed. São Paulo: Pearson Makron Books, 2006.

SICSÚ, A. L.; DANA, S. Estatística aplicada: análise exploratória de dados. São Paulo: Saraiva, 2012.

TIBONI, C.G.R. Estatística básica: para os cursos de administração, ciências contábeis, tecnólogos e de gestão. São Paulo: Atlas, 2010.

TOLEDO, G.L.; OVALLE, I.I. Estatística básica. 2. ed. São Paulo: Atlas, 1985

VIEIRA, S. Estatística para a qualidade. 3. ed. Rio de Janeiro: Elsevier, 2014.

WEBSTER, A.L. Estatística aplicada à administração e economia. 3. ed. São Paulo: McGraw-Hill, 2006.