Faculdade XPe

Capítulo 01

ESTATÍSTICA PARA CIÊNCIA DE DADOS

Arnaldo Satoru Gunzi

2024

Introdução à Estatística

A Estatística está em todos os lugares: nos testes farmacêuticos para avaliar se um remédio funciona ou não, nas pesquisas de opinião pública, no controle de qualidade dos produtos, na análise de risco de crédito de bancos, nos forecasts de preços e demanda, nos jogos de cassino... a lista é infindável.
A estatística estuda, entre outros assuntos, coletar, organizar, analisar, interpretar e apresentar dados. Em essência, ela nos ajuda a tomar decisões informadas, a identificar tendências, a fazer previsões e a entender a variabilidade inerente aos fenômenos naturais e sociais. Outra aplicação prática é a de não ser enganado pelos números: ao saber interpretar dados e gráficos, conseguimos discernir entre dados concretos e narrativas sobre eles.

Figura 1 - Ilustração sobre Estatística

Objetivo do curso

O objetivo deste curso é apresentar conceitos de probabilidade, estatística descritiva e estatística inferencial, utilizando uma abordagem prática e computacional.
O curso compreende análises descritivas de dados, noções de probabilidade, análises através de gráficos, análise de distribuições de probabilidades discretas e contínuas, testes de hipóteses, regressão linear e regressão logística. Esta apostila tem o foco mais teórico. No final de cada capítulo, há um link para atividades computacionais. As aulas gravadas complementarão o material apresentado.

Qual a diferença entre Probabilidade e Estatística?

A probabilidade e a estatística são campos correlatos. Em termos bem gerais, a probabilidade calcula as chances de um evento ocorrer dada uma população conhecida, ao passo que a estatística trata de descrever, analisar e tirar conclusões de uma amostra da população.

Probabilidade: É o ramo da matemática que estuda a chance ou a possibilidade de que um determinado evento ocorra, situações incertas. Por exemplo, se você joga um dado, a probabilidade de sair o número 6 é de 1 em 6, porque existem 6 lados no dado, mas só um deles é o número 6.

Estatística: É como a gente conta e analisa amostras, além de inferir conclusões. Exemplo: aplico um novo remédio a um grupo teste, e um placebo em um grupo de controle. Como avaliar se o remédio funcionou ou não, a partir dos dados desta amostra das pessoas do mundo? Como mostrar que o efeito não foi por acaso?

De modo simplificado, com probabilidades, conheço a população inteira e faço perguntas sobre ela; já com estatística, trabalho com amostras da população para descrever suas características e fazer inferências sobre ela.

Aplicações possíveis

Há uma infinidade de aplicações de probabilidade e estatística. Citando algumas:

a. **Detecção de fraudes:** para estimar as chances de uma nova transação de cartão de crédito ser ou não uma fraude, o modelo preditivo armazenado em nuvem pode receber os dados dessa transação via api, em formato json, processar esses dados e retornar a probabilidade de ser uma fraude. Caso a probabilidade seja alta, a transação pode ser bloqueada automaticamente.
b. **Análise de Risco:** Como as empresas usam a probabilidade para avaliar riscos financeiros.
c. **Previsão de Demanda:** Utilização de estatísticas para prever a demanda por produtos e serviços.
d. **Ensaios Clínicos:** Uso de testes estatísticos para avaliar a eficácia de novos medicamentos.
e. **Testes A / B para sites:** Qual dos designs funciona melhor?
f. **Controle de Qualidade:** Utilização de amostragem e gráficos de controle para monitorar processos de produção.
g. **Simulações Monte Carlo:** Método para resolver problemas complexos através de simulações probabilísticas e para dimensionamento de projetos.
h. **Probabilidade em Jogos de Azar:** Como calcular as chances de ganhar em jogos como pôquer, roleta etc.

Conforme se vê, há uma gama enorme de aplicações possíveis. Embarque nesta jornada!

Estatística Descritiva e Inferencial

A estatística pode ser dividida em duas áreas principais: estatística descritiva e estatística inferencial.

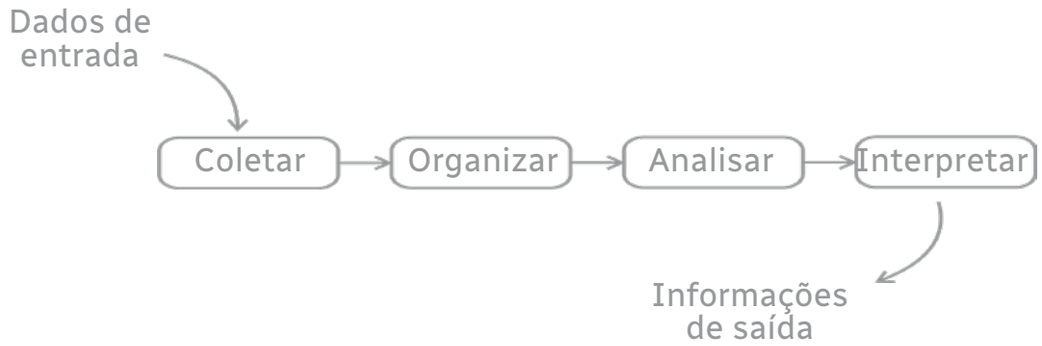
Estatística Descritiva

A estatística descritiva envolve a descrição de uma base de informações: organização e a apresentação dos dados de forma resumida e informativa, por meio do uso de tabelas, gráficos e medidas resumo.
Em estatística descritiva, serão discutidos tópicos de medidas de centralidade (média, mediana) e medidas de dispersão (desvio padrão, quartis, amplitude), além de gráficos diversos, como o boxplot.

Estatística Inferencial

A estatística inferencial, como o próprio nome implica, faz inferências a partir das informações que temos: generalizações de conclusões a partir de uma amostra para uma população maior.
Tópicos como intervalo de confiança, teste de hipótese e teorema do limite central serão discutidos em estatística inferencial.
Esta lógica segue o processo de análise de dados descrito na figura a seguir: primeiro organizamos e analisamos os dados existentes, para então inferir, interpretar e fornecer informações de saída.

Figura 2 - Ilustração sobre Estatística Inferencial



Fonte: Adaptado de Smailes & McGrane, 2012

Em resumo:

- **Estatística Descritiva:** Descreve e organiza os dados que você já tem.
- **Estatística Inferencial:** Usa os dados que você tem para fazer inferências ou previsões sobre uma população maior.

Na prática, essa diferença é mais para organizar o pensamento, porque utilizamos ferramentas de ambas para chegar a conclusões úteis para a tomada de ação, seja numa empresa, seja em nossas vidas.

O que são Medidas de Centralidade e Dispersão

As medidas de centralidade referem-se a valores típicos de uma variável, isto é, um valor em torno do qual uma grande proporção de outros valores está centralizada. As principais são: **Média Aritmética** e **Mediana**.

Também é importante saber como os dados se espalham ou o quão variadas são as observações em torno dessa medida central e, para isso, utilizamos as medidas de dispersão: **Variância, Desvio Padrão, Coeficiente de Variação, Amplitude** e **Quartis**.

Imagine que você trabalha em um grande hipermercado varejista. Cada linha do conjunto de dados é um dia, e cada valor é o preço praticado para o produto café.

Figura 3 Variável "Preco_Cafe", contendo 30 observações

Preco_Cafe
4,77
4,67
4,75
4,74
4,63
4,56
4,59
4,75
4,75
4,49
4,41
4,32
4,68
4,66
4,42
4,71
4,66
4,46
4,36
4,47
4,43
4,4
4,61
4,09
3,73
3,89
4,35
3,84
3,81
3,79

Para começar a explorar a variável, iremos tomar a **Média Aritmética**. A Média Aritmética é uma medida estatística que é calculada somando os valores da variável e dividindo pela quantidade de valores. Pode ser representada pela equação:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Onde x são os valores individuais de cada observação e n é a quantidade de observações.

Em Estatística, às vezes, trabalhamos com amostras, e, às vezes, trabalhamos com toda a população. A notação $\bar{x}$ indica que a média se originou de uma amostra, e a notação μ (letra grega mi) indica que a média se originou de uma população. População, nesse contexto, seria se estivéssemos trabalhando com todo histórico existente de preços do café. Não é o nosso caso, pois estamos trabalhando com uma amostra de trinta dias de variações de preço.

Aplicando a fórmula na nossa variável de estudo, que são as variações do preço do café, temos:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{132{,}79}{30} = 4{,}42$$

Agora que já sabemos o preço médio, precisamos de uma medida de dispersão para saber o quão os demais preços destoam desse preço médio. Para esse objetivo, utilizaremos o **Desvio Padrão**, que pode ser calculado pela fórmula:

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

Onde: x_i são os valores individuais de cada observação, $\bar{x}$ é a média aritmética e n é a quantidade de observações.
A moda é uma medida de tendência central que representa o valor mais frequente em um conjunto de dados. Em outras palavras, é o valor que aparece com maior frequência em uma distribuição de dados.
Há outras medidas como a média geométrica, média harmônica, mas devido ao menor uso delas, fica a cargo do aluno interessado pesquisar a respeito.
Assim como na Média Aritmética, temos notações para desvio padrão amostral e populacional. A literatura nos sugere utilizar a notação s para Desvio Padrão amostral e a letra grega sigma σ para desvio padrão populacional. Há uma pequena diferença entre o cálculo do desvio padrão da amostra s e no desvio padrão da população σ. No s utilizamos n-1 em seu denominador, já no σ utilizamos apenas o n.
Outra medida de dispersão muito utilizada na estatística é a variância, que nada mais é do que o desvio padrão elevado ao quadrado.

Calculando o desvio padrão dos preços, temos:

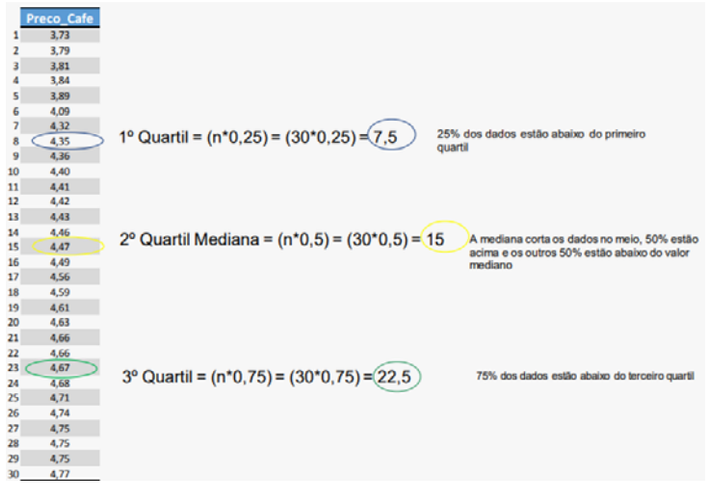
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{3}{30-1}} = \sqrt{0,1} = 0,32$$

A interpretação fica: Os preços praticados para o café variam em média 0,32 centavos em torno do seu preço médio. O Desvio Padrão nos dá a noção de variabilidade utilizando a própria unidade de medida da variável.
O Coeficiente de Variação é uma razão entre o Desvio Padrão e a Média Aritmética.

$$\frac{s}{\bar{x}} * 100 = \frac{0,31}{4,42} * 100 = 7,01\%$$

A interpretação fica: Os preços praticados para o café variam em média 7,01% em torno do preço médio.
Outro conjunto de medidas de dispersão bastante usado são os quartis. Eles são valores que dividem a variável em quatro partes iguais, e assim, cada parte representa 25% da variável.
Os quartis se dividem em **Primeiro Quartil (Q1), Segundo Quartil (Q2)** e **Terceiro Quartil (Q3)**. Para calcular os quartis, os valores da variável estudada devem estar ordenados do menor para o maior.
O Primeiro Quartil, ou Q1, é um valor que deixará 25% dos dados abaixo. O Segundo Quartil, ou Q2, é um valor que deixará 50% dos dados abaixo e 50% dos dados acima dele, ou seja, é um valor que corta os dados ao meio. Já o Terceiro Quartil, ou Q3, é um valor que deixará 75% dos dados abaixo dele.

Figura 4 - Quartis

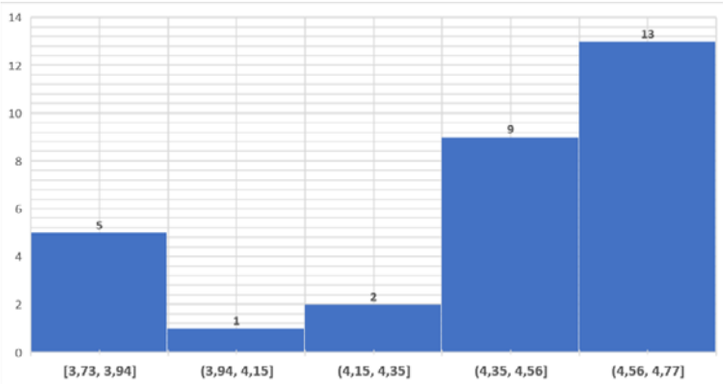


De forma análoga aos quartis, tempos os percentis. A diferença é que, ao invés de dividir em quatro, a amostra é dividida em 100 partes.
A **Amplitude** nada mais é do que o intervalo entre o maior valor e o menor valor. Como o menor preço foi de R$3,73 e o maior preço foi R$4,77, podemos dizer que na amostra estudada os preços variam entre R$3,73 e R$4,77, e a amplitude de R$ 1,04.

Análise de Dados Através de Gráficos

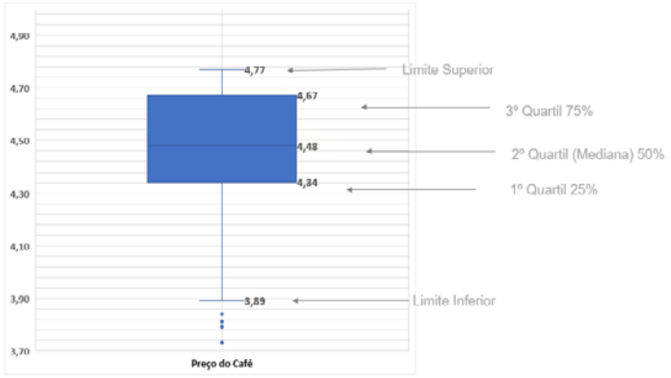
Um gráfico é a maneira visual de exibir variáveis. Normalmente, é mais fácil para qualquer pessoa entender a mensagem de um gráfico do que aquela embutida em tabelas ou sumários numéricos.
O **Histograma** é uma representação gráfica em barras de uma variável, dividida em classes. A altura de cada barra representa a frequência com que o valor da classe ocorre. Vejamos um histograma para apresentar a variação do preço do café.

Figura 5 - Histograma dos Preços Praticados para o Café



Repare que cada barra corresponde à frequência de um intervalo de preços. Interpretando, temos: a primeira barra nos informa que tivemos cinco registros em que o preço praticado foi entre R$3,73 e R$3,94. A segunda barra nos informa que tivemos um registro no qual o preço estava entre R$3,94 e R$4,15 etc.
Outro gráfico bastante utilizado na Estatística é o **Boxplot**, que é baseado nos quartis e é um modo rápido de visualizar a distribuição dos dados. Vejamos um boxplot para os preços do café.

Figura 6 - Boxplot dos Preços Praticados para o Café

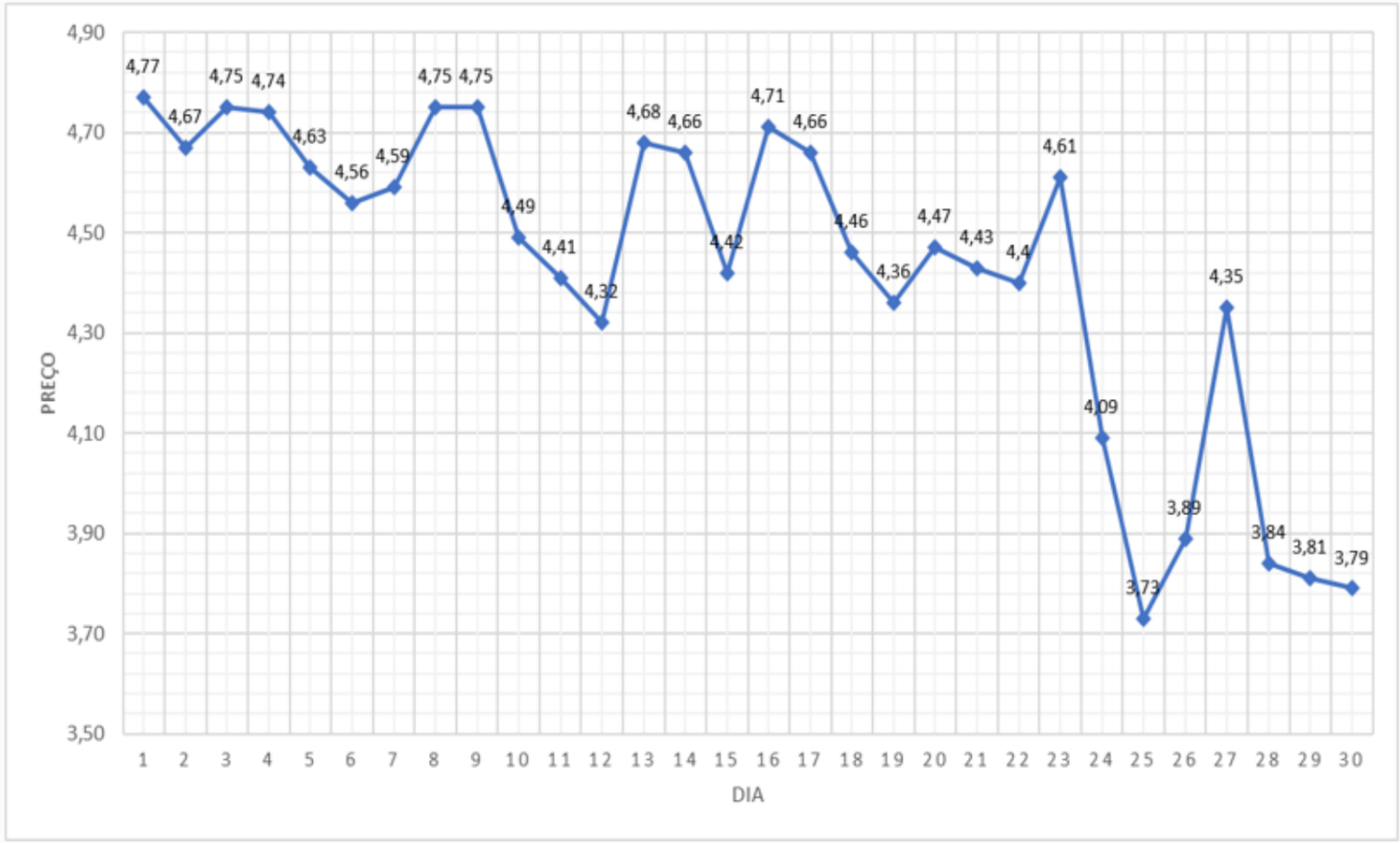


Além dos quartis, o boxplot também nos dá o limite inferior e o limite superior. No boxplot, vemos que o limite superior R$4,77, ou seja, de acordo com a distribuição dos preços, um valor acima do de R$4,77 é um outlier. Já o limite inferior é R$3,89, ou seja, preços abaixo desse valor são considerados outliers.
Valores outliers, sejam superiores ou inferiores, devem ser investigados para compreender o que houve naquela observação, pode ter sido de fato um evento raro ou apenas um erro na entrada de dados.
O intervalo interquartil (IQR) nada mais é que subtrair o terceiro quartil pelo primeiro quartil. Uma vez calculado o IQR, para chegar nos valores limites, a fórmula fica:

$$IQR = 3ºQuartil - 1ºQuartil$$
$$Limite\ Inferior = 1ºQuartil - (1.5 * IQR)$$
$$Limite\ Superior = 3ºQuartil + (1.5 * IQR)$$

Se desejarmos visualizar a evolução dos preços ao longo do tempo, é recomendado utilizar um **gráfico de linhas** (também chamado de gráfico de séries temporais). Ele é bastante simples. Basta plotar a variável no eixo vertical y e o tempo no eixo horizontal x. Cada ponto é representado por um marcador e ligado ao ponto seguinte por uma reta.

Figura 7 - Gráfico de linha da evolução dos preços do café durante os dias do mês

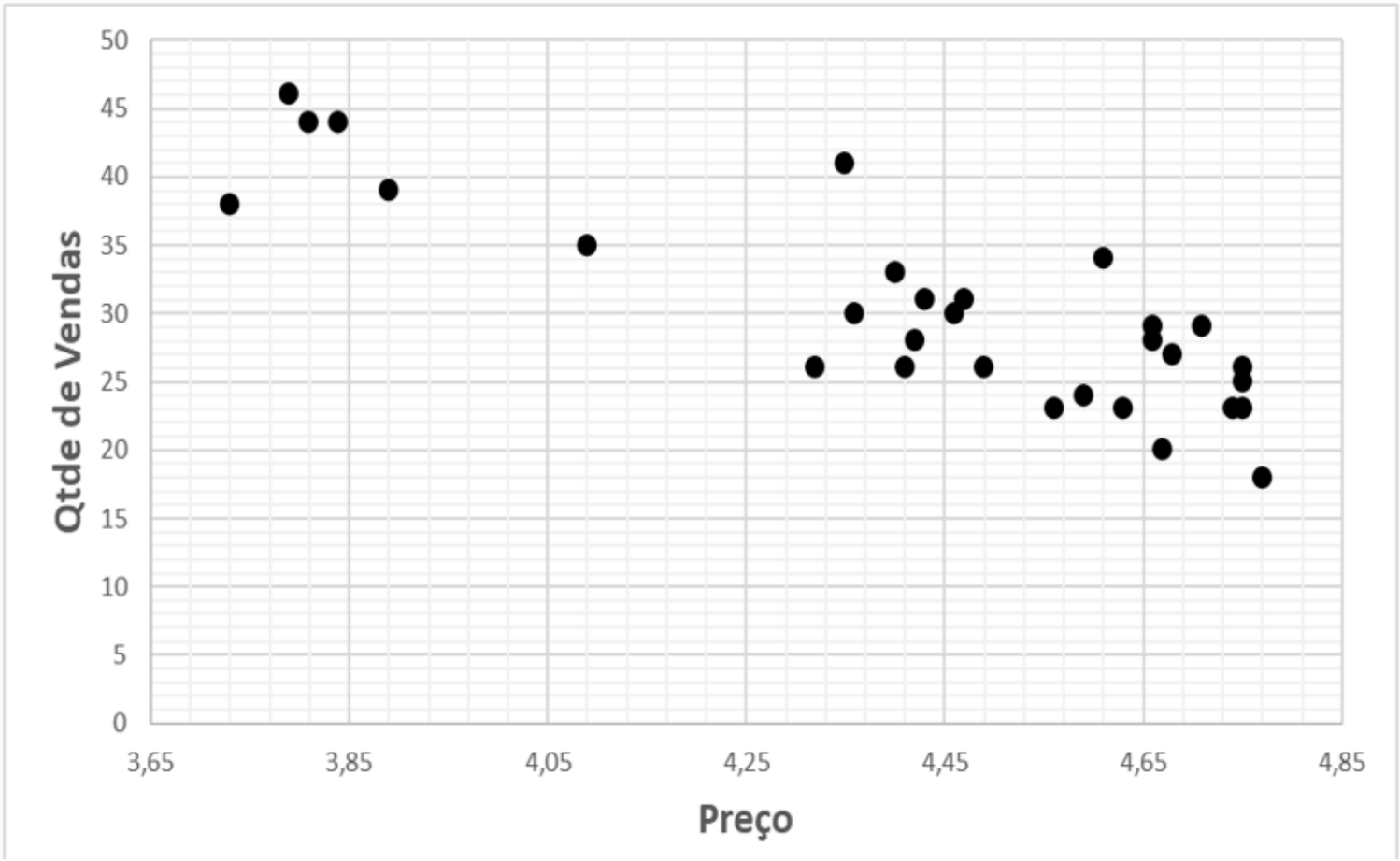


Supondo que você precise analisar se existe relação entre o preço do café com as vendas do café. Vamos adicionar mais uma variável.

Figura 8 - Variável Preço do Café e a variável Vendas do Café

Preco_Cafe	Vendas_Cafe
4,77	18
4,67	20
4,75	23
4,74	23
4,63	23
4,56	23
4,59	24
4,75	25
4,75	26
4,49	26
4,41	26
4,32	26
4,68	27
4,66	28
4,42	28
4,71	29
4,66	29
4,46	30
4,36	30
4,47	31
4,43	31
4,4	33
4,61	34
4,09	35
3,73	38
3,89	39
4,35	41
3,84	44
3,81	44
3,79	46

Figura 9 - Relação entre o preço do café e as vendas do café



Há inúmeras outras maneiras gráficas de exibir dados. Outras muito utilizadas são os gráficos de barras e o gráfico de setores (ou gráfico de pizza). Para direcionar nossos esforços focaremos nos que foram apresentados.

Estatística Computacional – Análise Exploratória de Dados com Python

Vide link a seguir para acompanhar o código de apoio para análise exploratória de dados com Python.

https://github.com/asgunzi/Estatistica_Analise_Dados

Faculdade XPe

xpeducacao.com.br