



Fundação

**CECIERJ**

Consórcio **cederj**

Centro de Educação Superior a Distância do Estado do Rio de Janeiro

## **Métodos Estatísticos**

**Volume Único**

Rafael Canellas Ferrara Garrasino



**GOVERNO DO  
Rio de Janeiro**

**SECRETARIA DE CIÊNCIA,  
TECNOLOGIA, INOVAÇÃO E  
DESENVOLVIMENTO SOCIAL**

**UNIVERSIDADE  
ABERTA DO BRASIL**

**MINISTÉRIO DA  
EDUCAÇÃO**



Apoio:



**FAPERJ**

Fundação Carlos Chagas Filho de Amparo  
à Pesquisa do Estado do Rio de Janeiro

# Fundação Cecierj / Consórcio Cederj

www.cederj.edu.br

## Presidente

Carlos Eduardo Bielschowsky

## Vice-presidente

Marilvia Dansa de Alencar

## Coordenação do Curso de Tecnólogo em Turismo

Claudia Fragelli

## Material Didático

### Elaboração de Conteúdo

Rafael Canellas Ferrara Garrasino

### Direção de Design Instrucional

Cristine Costa Barreto

### Coordenação de Design Instrucional

Bruno José Peixoto

Flávia Busnardo da Cunha

Paulo Vasques de Miranda

### Design Instrucional

Anna Maria Osborne

Jacks Williams Peixoto Bezerra

José Meyohas

Paulo Alves

### Coordenação de Produção

Fábio Rapello Alencar

### Revisão Linguística e Tipográfica

Carolina Godoi

### Ilustração

Clara Gomes

### Capa

Clara Gomes

### Programação Visual

Alexandre d'Oliveira

Filipe Dutra

Larissa Averbug

### Produção Gráfica

Patrícia Esteves

Ulisses Schnaider

Copyright © 2016, Fundação Cecierj / Consórcio Cederj

Nenhuma parte deste material poderá ser reproduzida, transmitida e gravada, por qualquer meio eletrônico, mecânico, por fotocópia e outros, sem a prévia autorização, por escrito, da Fundação.

G242m

Garrasino, Rafael Canellas Ferrara.

Métodos Estatísticos. Volume único. / Rafael Canellas Ferrara Garrasino. – Rio de Janeiro : Fundação Cecierj, 2016.

320p.; 19 x 26,5 cm.

ISBN: 978-85-458-0012-5

1. Estatística. I. Título.

CDD: 519.5

Referências Bibliográficas e catalogação na fonte, de acordo com as normas da ABNT.  
Texto revisado segundo o novo Acordo Ortográfico da Língua Portuguesa.

# Governo do Estado do Rio de Janeiro

## Governador

Luiz Fernando de Souza Pezão

## Secretário de Estado de Ciência, Tecnologia, Inovação e Desenvolvimento Social

Gabriell Carvalho Neves Franco dos Santos

## Instituições Consorciadas

### CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca

Diretor-geral: Carlos Henrique Figueiredo Alves

### FAETEC - Fundação de Apoio à Escola Técnica

Presidente: Alexandre Sérgio Alves Vieira

### IFF - Instituto Federal de Educação, Ciência e Tecnologia Fluminense

Reitor: Jefferson Manhães de Azevedo

### UENF - Universidade Estadual do Norte Fluminense Darcy Ribeiro

Reitor: Luis César Passoni

### UERJ - Universidade do Estado do Rio de Janeiro

Reitor: Ruy Garcia Marques

### UFF - Universidade Federal Fluminense

Reitor: Sidney Luiz de Matos Mello

### UFRJ - Universidade Federal do Rio de Janeiro

Reitor: Roberto Leher

### UFRRJ - Universidade Federal Rural do Rio de Janeiro

Reitor: Ricardo Luiz Louro Berbara

### UNIRIO - Universidade Federal do Estado do Rio de Janeiro

Reitor: Luiz Pedro San Gil Jutuca



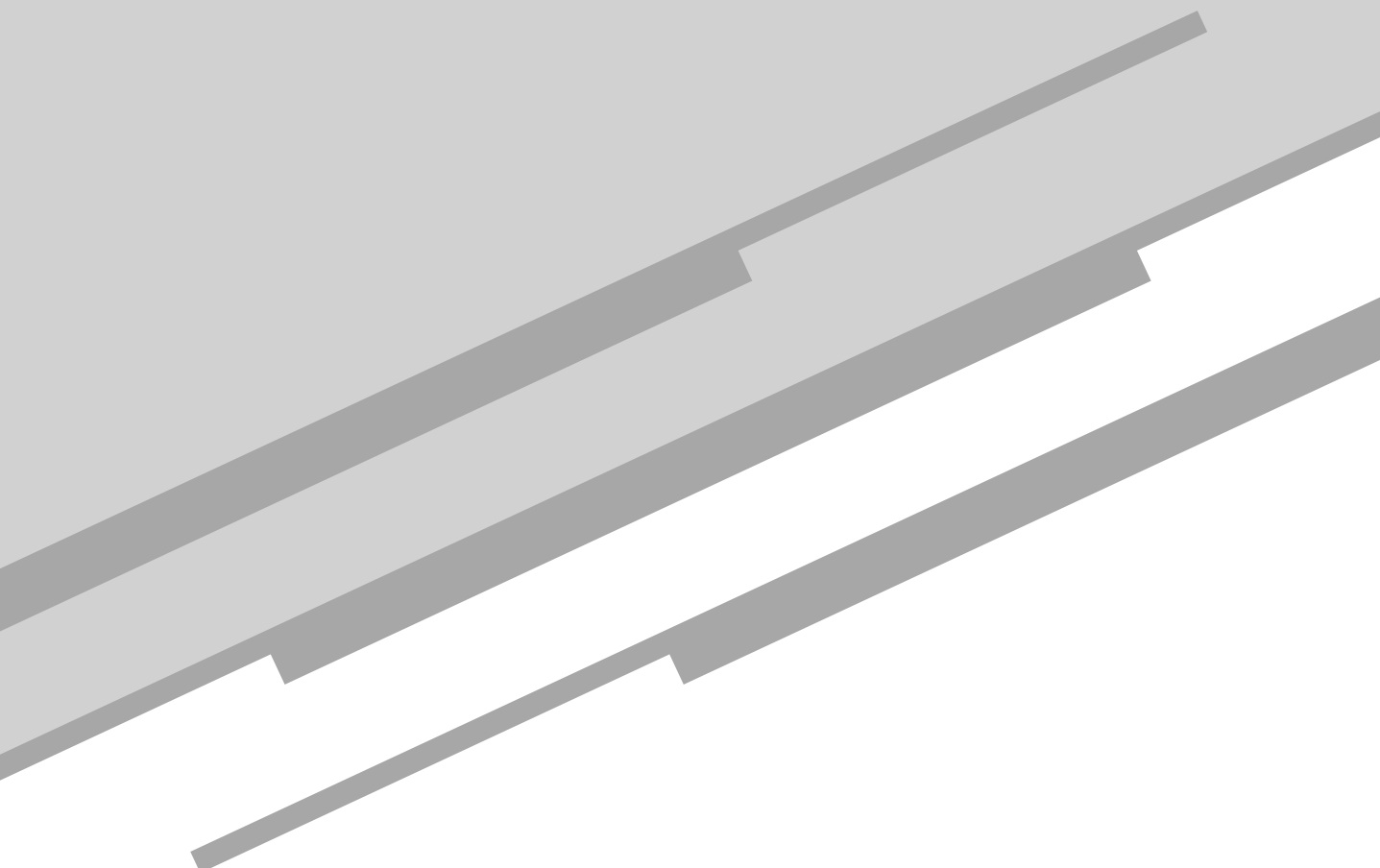
# Sumário

<b>Aula 1 • Estatística para o quê mesmo? .....</b>	<b>7</b>
<i>Rafael Canellas Ferrara Garrasino</i>	
<b>Aula 2 • Aquele negócio que “troça” essa coisa.....</b>	<b>29</b>
<i>Rafael Canellas Ferrara Garrasino</i>	
<b>Aula 3 • Senhor, um minuto da sua atenção, por favor!.....</b>	<b>49</b>
<i>Rafael Canellas Ferrara Garrasino</i>	
<b>Aula 4 • Coloca em um <i>Tapeware</i> com etiqueta identificando .....</b>	<b>69</b>
<i>Rafael Canellas Ferrara Garrasino</i>	
<b>Aula 5 • Vai uma pizza, aí?.....</b>	<b>89</b>
<i>Rafael Canellas Ferrara Garrasino</i>	
<b>Aula 6 • Estou ao lado do baixinho careca! .....</b>	<b>115</b>
<i>Rafael Canellas Ferrara Garrasino</i>	
<b>Aula 7 • Três para cá, três para lá!.....</b>	<b>141</b>
<i>Rafael Canellas Ferrara Garrasino</i>	
<b>Aula 8 • A onda, o bigode e o russo .....</b>	<b>165</b>
<i>Rafael Canellas Ferrara Garrasino</i>	
<b>Aula 9 • Z de Zorro .....</b>	<b>191</b>
<i>Rafael Canellas Ferrara Garrasino</i>	
<b>Aula 10 • Vai subir comigo ou vai descer?.....</b>	<b>209</b>
<i>Rafael Canellas Ferrara Garrasino</i>	
<b>Aula 11 • O próximo sorteado é... ..</b>	<b>235</b>
<i>Rafael Canellas Ferrara Garrasino</i>	
<b>Aula 12 • A sorte para todos .....</b>	<b>263</b>
<i>Rafael Canellas Ferrara Garrasino</i>	
<b>Aula 13 • Cada macaco no seu galho .....</b>	<b>285</b>
<i>Rafael Canellas Ferrara Garrasino</i>	
<b>Referências.....</b>	<b>315</b>



# Aula 1

Estatística para o quê mesmo?



*Rafael Canellas Ferrara Garrasino*

## **Meta**

Estabelecer a importância da estatística para a estruturação de planejamentos de uma empresa, de uma organização ou até mesmo de uma entidade pública.

## **Objetivos**

Esperamos que, após o estudo desta aula, você seja capaz de:

1. reconhecer a importância da estatística em diversas áreas;
2. identificar, dentro de um problema inicial, as informações que serão necessárias para buscar uma solução;
3. diferenciar todo o processo estatístico – do surgimento da demanda à análise do relatório final.

## Introdução

Senhores, imaginem que estão responsáveis por uma empresa de turismo. Como prever quantos clientes deverá procurar para oferecer os seus serviços ou como estimar quantos pacotes deverá deixar previamente reservados para uma temporada? Qual tipo de mídia utilizar para captar mais clientes potenciais? Como detectar os pontos que, necessariamente, precisam ser revistos para melhorar o seu atendimento?

São perguntas como essas que iremos responder no decorrer desta disciplina. Antes, faz-se necessário compreender melhor o real propósito da Estatística, ou seja, como utilizá-la para obter sua maior eficiência e como chegar às informações que serão capazes de gerar dados suficientes para que a leitura do cenário em questão seja feita.

Deste modo, com exemplos de casos factíveis, juntos entenderemos alguns problemas e necessidades que recorrem à estatística como ferramenta para fazer efetiva leitura da situação ou para prever melhor perspectiva do negócio. Assim, em uma das várias definições nos dicionários, estatística se trata de informações e/ou fatos numéricos coletados, organizados sistematicamente e estudados.

## Estatística e tomada de decisão

De uma maneira mais completa, podemos dizer que estatística é uma parte da matemática que transforma dados em informações úteis para a tomada de decisão. Estas informações podem ser de todo o tipo, ou seja, desde a idade das pessoas que frequentam o teatro, do seu bairro, até os valores gastos por famílias nas compras de Natal. Estas informações não precisam ser necessariamente numéricas, pois elas podem ser de tipos de operadoras de celular, de marca de carros, de destinos de viagens ou o clássico par “sim” e “não”. Obviamente, se o estudo envolver apenas valores, ganharemos agilidade na análise da informação. Caso contrário, teremos de transformar as respostas em números para enfim processá-las em dados úteis às decisões.

Nesse momento, imagino que já esteja um pouco cansado de ouvir falar em tomada de decisão. Eu, pelo menos, não aguento mais escrever isto, mas dessa forma podemos deixar bem clara a meta principal da estatística que é auxiliar na tomada de decisão.

Assim, antes, mesmo com a existência secular da estatística, muitas empresas recorriam à experiência e sensibilidade de seus gerentes para a

tomada de decisão. Posteriormente, notaram que com o uso da estatística as informações chegariam de forma mais precisa e claras ao gerente. Deste modo, essas informações, somadas às habilidades do gerente, permitiriam uma decisão mais bem embasada.

Ainda assim, deixando clara a real finalidade da estatística, ela ainda é usada para outros meios como, por exemplo, ilustrar várias ações em uma única informação. Este recurso é muito utilizado em transmissões esportivas. Contudo, assim como pode ser uma excelente ferramenta para agregar valor à transmissão, pode ser confusa ou inútil dependendo da maneira que for empregada. Como exemplos, temos:



**Figura 1.1:** Show do Intervalo!

A primeira estatística gerada pela emissora é clara e de grande utilidade. Ao fazer a sua leitura, qualquer telespectador, seja com base estatística ou não, será capaz de interpretar sua real finalidade. O jogador Joãozinho é o que mais chuta a gol daquele time. Logo, ou é o que recebe mais bolas em condições de finalização ou é o chamado “fominha”. Portanto, uma boa maneira do time adversário evitar gols é marcando melhor o Joãozinho.

A segunda é uma que comumente presencio e que gera mais discussão. Pensem comigo. Você acaba de ligar a televisão e está no intervalo. Você não sabe o placar da partida. Surge a informação de que o goleiro Muralha fez 3 de 4 defesas difíceis. A sua primeira conclusão será que o time dele levou um gol, correto? Pois é. Está errado. Não! A sua conclusão está perfeita! A televisão é que está errada.

Por várias vezes, presenciei dados como este e o placar estava zerado. Ora, se o goleiro foi exigido em 4 defesas difíceis e apenas conseguiu concluir 3 delas, significa que uma das quatro entrou no gol – ou será que outro jogador fez esta defesa? Seria então uma defesa difícil que iria para fora? A bola parou sobre a linha? Este é um dos casos que, na necessidade de gerar dados, sem entender ao certo a sua real utilidade, foi gerada uma estatística confusa e contraditória.

A terceira estatística exibida é uma das que se encaixa com perfeição naqueles casos em que o gerente, com sua experiência, complementa a leitura dos dados. Neste caso, o gerente será o comentarista esportivo. Ao receber a informação da distância total percorrida pelo atleta durante a partida, em primeiro lugar, irá dizer se aquela é uma distância acima do esperado para a posição dele. Este primeiro comentário irá determinar o real esforço do atleta. Contudo, nesse momento, o comentarista deverá acrescentar mais uma nota que vai **ratificar** ou não o real esforço do atleta ou se a longa distância por ele percorrida foi apenas uma consequência das suas tarefas. Vejamos: imaginem que na partida a equipe teve várias cobranças de falta e escanteios e que o atleta em questão seja o cobrador oficial. Somente com este detalhe, o comentarista vai deixar claro que, apesar de ter percorrido um longo percurso durante a partida, uma boa parte dele foi se deslocando em velocidade baixa por vários pontos distantes para cumprir tarefas corriqueiras. Portanto, talvez a sua longa distância percorrida na partida não seja tão impressionante assim.

### Retificar

Significa tornar autêntica a aprovação de; validar; comprovar; confirmar. Não confundir ratificar com retificar. Retificar é consertar. Por isto as oficinas de automóveis também são chamadas de *Retíficas de Motores*.

## Atividade 1

### Atende aos objetivos 1 e 2

Suponhamos que seja responsável por uma empresa que oferece serviços de roteiros por pontos turísticos de uma cidade qualquer. Você pretende reestruturar o seu serviço fazendo uma nova seleção dos pontos que serão o destino do passeio. Como obter uma seleção que potencialmente será interessante aos seus clientes?

---



---



---



---



---

---

---

---

---

---

---

---

---

---

---

### **Resposta comentada**

Existem algumas opções de montar a seleção, contudo, para todas será necessária uma pesquisa, por mais simples que seja. Sendo o objetivo da empresa montar um passeio que atenda de forma genérica seus clientes, deveremos pesquisar os pontos mais visitados pela maioria dos turistas que visitam a sua cidade. Desejando algo mais bem direcionado ao seu público, podemos pesquisar com os clientes atuais (antes da alteração) os pontos que possuem mais interesse ou simplesmente fazer uma pesquisa de idade, por exemplo, e montar um roteiro baseado especificadamente àquela faixa etária.

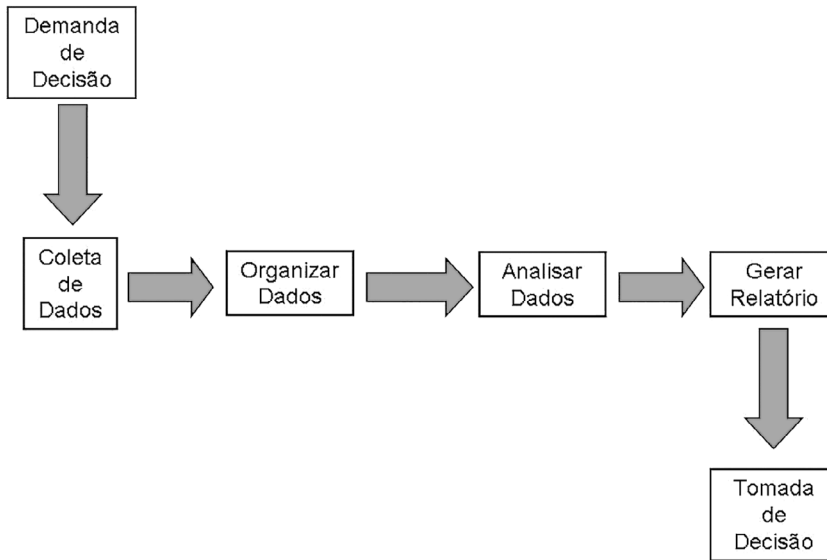
---

---

---

## **O processo estatístico**

Para que possam tomar decisões, empresas precisam de várias informações e uma delas será de origem “estatística”. Neste processo decisório, a estatística será primordial em duas etapas: fornecendo métodos para a coleta dos dados e gerando ferramentas para a análise destes – transformando-os em informações úteis. A **Figura 1.2** ilustra esse processo de forma resumida:



**Figura 1.2:** Processo estatístico de demanda e tomada de decisão.

O processo se inicia com a necessidade de se tomar uma decisão. Isto é, aumentar o seu negócio, investir mais em algum meio de comunicação para divulgar sua marca, alavancar sua produção na expectativa de mais vendas em um futuro próximo etc. Inspirado nesta necessidade de prever ou fazer uma leitura de um cenário específico surge a demanda de decisão.

Com o surgimento da demanda é possível estabelecer quais tipos de dados serão necessários para que possamos chegar à conclusão ou à decisão a ser tomada. Este ponto é bastante importante, pois a coleta de mais dados que o necessário poderá prolongar as próximas etapas com o manuseio de informações que não serão úteis no processo decisório. Isso também acontece como o seu inverso, isto é, coletar menos informações que o necessário acabará implicando em refazer a pesquisa para complementar os dados faltantes.

Com os dados enfim coletados, em quantidade satisfatória, eles serão submetidos a processos de organização para que sua análise possa ser feita de maneira mais ágil. Este processo varia de acordo com o tipo de dados os quais serão analisados, como de acordo com o método que será utilizado para a análise.

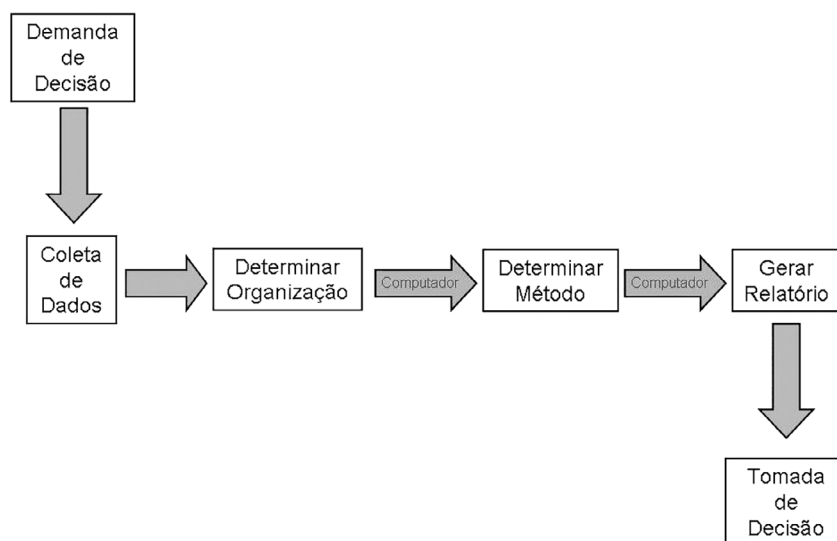
A análise se dá imediatamente após a organização dos dados. Para tal, a Estatística fornece inúmeras opções de ferramentas de análise que devem ser utilizadas de acordo com os dados fornecidos, o tipo

de situação, dentre outras condições iniciais que serão explicadas mais apropriadamente.

Depois de feita a análise, gera-se um relatório, no qual a quantidade inicial de dados coletados transforma-se, em sua maior parte das vezes, em alguns resultados. Este relatório é o grande encerramento do processo estatístico, ou seja, a transformação de todo um processo e alguns números que terão grande relevância para quem irá estudá-los.

Por fim, de posse do relatório, o responsável pela decisão determinará, baseado no resultado, qual medida deverá ser tomada como resposta à demanda originalmente gerada pela empresa.

Ainda assim, é importante ressaltar que tanto o processo de organização dos dados, quanto a sua análise, são processos longos e exaustivos. Daí, na sua maioria das vezes, mesmo quando coordenado por um estatístico, eles são feitos em computadores com o auxílio de programas que podem ser desde o Excel até sistemas sofisticados voltados, exclusivamente, para a Estatística. Então, podemos dizer que na etapa da organização e na etapa da análise o indivíduo interfere, apenas, determinando qual a melhor organização dos dados e qual o melhor método estatístico de análise a ser utilizados, pois o restante ficará por conta dos computadores. Sendo assim, a **Figura 1.3** ilustra de maneira mais completa o processo.



**Figura 1.3:** Processo estatístico, amplo, de demanda e tomada de decisão.

Portanto, podemos concluir que o processo estatístico é algo muito mais intelectual do que braçal. Não requer cálculos longos e intermináveis operações algébricas. Cabe apenas o indivíduo orientar o processo e no final fazer a leitura do relatório. Contudo, para que chegue a este ponto, ele precisa ter uma boa bagagem de conhecimento das ferramentas estatísticas, como também uma maturidade de leitura de resultados.

## Atividade 2

### Atende ao objetivo 3

Ao fim de um projeto, toda a equipe estratégica foi reunida para finalizar o processo. Naquele momento o diretor da empresa perguntou a cada um dos seus funcionários qual a atividade executada por eles e obteve as seguintes respostas:

**Mario:** “Tive acesso às planilhas e a partir delas pude determinar a média de custo das novas matérias-primas, o tempo de entrega e as recomendações dos outros compradores dessas.”

**Suzana:** “Entrei em contato com todos os fabricantes de matéria-prima que se enquadram na necessidade da nossa fábrica. Perguntei preço e tempo de entrega. Posteriormente, entrei em contato com alguns dos compradores dessas mesmas matérias-primas e perguntei qual a impressão que tinha delas.”

**Valter:** “Determinei qual será o novo fornecedor de matéria-prima que iremos adotar a partir de hoje.”

**Adriana:** “Montei uma planilha na qual cada fornecedor tinha registrado o preço da matéria-prima que vendia, o tempo de entrega e as qualificações dadas pelos clientes atuais.”

**Fernando:** “Após várias reclamações do setor de produto, identifiquei que o nosso atual fornecedor além de ter um custo alto, tem atrasado algumas entregas e muitas delas com não conformidades.”

**Patrícia:** “Fiz um *ranking* dos fornecedores pelo custo, tempo de entrega e qualificação dos compradores.”

Em vista dos *feedbacks* apresentados, determine por qual etapa do processo estatístico cada funcionário ficou responsável.

---

---

---

---

---

---

---

---

---

---

---

---

### **Resposta comentada**

Fernando, ao detectar um problema, *gerou a demanda* de substituir o fornecedor. Logo, ficou com a primeira parte.

Suzana entrou em contato com todos os potenciais novos fornecedores e fez a *coleta dos dados*; a segunda parte.

Adriana pegou os dados coletados pela Suzana e *organizou-os*, executando a terceira parte.

Mario, de posse dos dados organizados, *fez a análise* deles, que é a quarta parte.

Patrícia transformou a análise do Mario em dados mais resumidos, *elaborando um relatório*, a quinta parte.

Valter, com o relatório de Mario, *tomou a decisão* de qual será o novo fornecedor, executando a sexta e última etapa.

---

---

---

---

### **Casos de estudo**

Todos os casos a seguir são meramente fictícios, mas ao mesmo tempo factíveis. A ideia inicial será apenas desenvolver o processo estatístico de maneira intuitiva em casos simples. Posteriormente, com o decorrer do curso, deixaremos tudo mais sofisticado.

## Caso da conta de luz



**Figura 1.4:** Visão noturna de Hong Kong (China/2011).

Fonte: [http://pt.wikipedia.org/wiki/Ficheiro:1\\_hong\\_kong\\_panorama\\_2011\\_dusk\\_victoria\\_peak.jpg](http://pt.wikipedia.org/wiki/Ficheiro:1_hong_kong_panorama_2011_dusk_victoria_peak.jpg) – chensiyuan

Um condomínio empresarial se engajou na campanha de redução de gastos com energia, dos escritórios, como medida preventiva de preservação dos bens em escassez. Para tal, precisará gerar um relatório que seja reduzido, claro e de fácil entendimento para todos os 1.500 escritórios divididos em 10 blocos. Para que tenhamos um caso de nível baixo de complexidade, partiremos da premissa de que todos os ambientes empresariais têm o mesmo tamanho e igual quantidade de cômodos e de usuários.

A demanda necessária, neste momento, será a meta que cada escritório deverá bater na conta de luz. Este valor pode ser arbitrário ou inspirado em algum resultado estipulado por qualquer entidade voltada para estes assuntos. Sendo assim, suponhamos que a meta será de R\$ 150,00 mensais.

Neste momento temos o problema e a meta. Agora precisamos iniciar a coleta dos dados. Se o foco é exclusivamente o valor da conta de luz, a coleta fica facilmente determinável: será o valor da conta de cada apartamento. Portanto, fica decidido que uma ou mais pessoas será responsável por coletar o valor da fatura de luz, do último mês, de cada escritório.

Aqui cabe mais uma ressalva, pois a simples informação da última fatura não necessariamente representará a realidade. Suponhamos que no último mês tivemos uma anormal elevação da temperatura. Este fato pode ser providencial para subir o valor com o consumo de ar-condicionado e ventiladores. O mesmo, no sentido inverso, aconteceria em um mês de temperaturas extremamente baixas. De qualquer forma, para manter a simplicidade do caso, permaneceremos com um mundo perfeito de faturas idênticas mensalmente.

Seguindo, temos os valores das últimas faturas de todos os escritórios. Agora se faz necessário organizá-las, pois lidar com 1.500 valores diferentes ficará complicado demais. Uma boa opção de organizar estes dados é reuni-los em faixas, isto é, determinar algumas faixas de valores e contabilizar quantos escritórios estão compreendidos entre eles. Por exemplo: se a meta é de R\$ 150,00, podemos criar a Faixa 1 para os usuários que possuem baixo consumo, entre R\$ 0,00 (que seria um escritório vazio) e R\$ 80,00. A Faixa 2 ficará entre R\$ 80,01 e R\$ 150,00. A Faixa 3 abrange os valores entre R\$ 150,01 e R\$ 200,00. A última faixa, Faixa 4, valores acima de R\$ 200,00.

Com esta organização, não importa o valor individual de cada fatura, afinal uma dentro de um universo de 1.500 faturas pode ter uma **significância** baixa. Assim, agrupando em faixas, podemos lidar com grupos e identificar melhor o comportamento do condomínio em um todo. Portanto, fazendo a contagem das faturas de acordo com o seu valor e as faixas previamente determinadas, tivemos o seguinte resultado:

## Significância

Designação de valores numéricos que permitam avaliar a probabilidade de aceitar-se como verdadeira uma hipótese falsa. Na estatística, o termo significância está relacionado à ideia de ser válido, possuir importância. Ao afirmar que um resultado é significativo para um estudo, estamos dizendo que ele possui uma importante relevância ao mesmo. Assim como, inversamente, dizendo que não há significância, estamos afirmando que ele pode ser desprezado, pois não possui importância para o estudo.

**Tabela 1.1:** Faixa de valores x Quantidade de escritórios

Faixa	Quantidade
R\$ 0,00 até R\$ 80,00	475
R\$ 80,01 até R\$ 150,00	388
R\$ 150,01 até R\$ 200,00	423
Acima de R\$ 200,00	214

Com os dados devidamente organizados, faz-se necessário analisá-los. Como estamos falando de dados simples e uma única informação (o valor da conta), uma simples análise percentual será suficiente. Deste modo, veremos qual o percentual do total de moradores que está compreendido em cada faixa.

**Tabela 1.2:** Faixa de valores x Quantidade de escritórios x Percentual

Faixa	Quantidade	Percentual
R\$ 0,00 até R\$ 80,00	475	31,67%
R\$ 80,01 até R\$ 150,00	388	25,87%
R\$ 150,01 até R\$ 200,00	423	28,20%
Acima de R\$ 200,00	214	14,27%

Já é possível agora obtermos um resultado para este estudo. Baseado na análise dos dados, temos que 57,53% dos escritórios conseguiriam bater a meta estipulada, enquanto 42,47% não conseguiriam. A tomada de decisão será feita baseada no resultado. Isto é: será que a meta foi exagerada? Para esta pergunta podemos dizer que não, pois não somente a maioria cumpriu a meta, como a faixa com maior quantidade de escritórios é a mais distante da meta (abaixo de R\$ 80,00). Segunda tomada de decisão: é possível melhorar este resultado? Novamente inspirado nos dados, vemos que uma quantidade considerável está na primeira faixa acima da meta. Logo, é possível fazer uma campanha com estes escritórios para que consigam diminuir seus gastos a ponto de alcançar o valor estipulado. Outra pergunta que pode ser feita: é possível melhorar a análise destes dados? Sim, se pegarmos a Faixa 2 e a Faixa 3 e dividirmos cada uma em mais duas novas faixas, poderemos identificar quantos apartamentos estão de fato próximo da meta para batê-la ou para ultrapassá-la.

## Caso do desperdício no restaurante



**Figura 1.5:** Desperdício de alimentos no preparo de refeições.

Você é o gerente de um restaurante e notou que, no manuseio dos alimentos, os cozinheiros têm desperdiçado muitos ingredientes. Este desperdício pode até ser explicado, mas, com o intuito de tentar reduzi-lo, optou por fazer um levantamento geral das informações.

Sua primeira ação foi escolher quais alimentos serão objeto do seu estudo. Deste modo, após conversar com os cozinheiros, concluiu que os alimentos que mais geravam desperdício eram: ovos, azeitonas, cebolas, tomates e batatas.

É importante notar que somente nesta primeira etapa já foi feito, mesmo que sutilmente, um processo estatístico. Você pesquisou com os cozinheiros os alimentos, reuniu as informações, gerou um relatório (mesmo que mentalmente) e concluiu quais serão estudados. Aqui, deixamos bem claro que, mesmo para fazer um processo estatístico, podemos redigir para que seja possível ter outros processos estatísticos dentro de um mesmo processo de elaboração ou de otimização.

Em seguida, orientou a todos que acumulassem, em vasilhas diferentes, os alimentos da lista desperdiçados para que pudessem contabilizar a perda. Após duas semanas, o resultado foi transformado em uma planilha:

**Tabela 1.3:** Alimentos x Desperdício

Alimentos	Desperdício
Ovos	68 unid
Azeitonas	143 unid
Cebolas	35 unid
Tomates	28 unid
Batatas	45 unid

Repare que aqui tivemos algumas etapas do processo estatístico. Você obteve a demanda, coletou dados (reunindo o desperdício em vasilhas) e organizou as informações (lançamento em uma planilha). O próximo passo será analisar as informações para gerar um relatório. Contudo, note que as informações geradas não possuem muita relevância. Vamos refletir sobre as azeitonas. Aquela quantidade desperdiçada é muito grande? Insignificante? Pode ser comparada com a quantidade de batatas desperdiçadas? Enfim, precisamos de mais dados para fazer a análise. Aqui, mais uma vez **corroboramos** com a necessidade de sempre cogitar otimizar o processo estatístico a partir de uma reflexão sobre o mesmo em todas as suas etapas.

Como resultado da reflexão que fizemos no parágrafo anterior, optamos por coletar também o total de cada alimento manuseado. Com isto, podemos comparar o total manuseado com a quantidade desperdiçada e, assim, medir o real desperdício de cada alimento. Perceba que aqui não foi necessariamente feito todo o processo estatístico. Apenas fizemos um pequeno ajuste na parte da coleta dos dados. Sendo assim, com as novas informações coletadas e organizadas com as anteriores chegamos a uma nova tabela:

**Tabela 1.4:** Alimentos x Desperdício x Total

Alimentos	Desperdício	Total
Ovos	68	320
Azeitonas	143	1900
Cebolas	35	110
Tomates	28	100
Batatas	45	500

### Corroborar

Dar força a; aduzir provas da verdade de; confirmar, comprovar. Comumente, utiliza-se corroborar no mesmo sentido de *ratificar*.

Agora, com o desperdício podendo ser comparado com o total manuseado é possível concluir quais alimentos tiveram o maior desperdício percentual. Isto é: dos alimentos estudados, quais devem, supostamente, receber uma maior atenção por possuírem o menor índice de aproveitamento. Compreenda que, para esta ocasião, usaremos *aproveitamento* no sentido de produto comprado e efetivamente utilizado.

Com os dados complementares coletados, já é possível fazer uma análise prévia do real desperdício dos alimentos. A próxima tabela trouxe as informações fazendo um percentual de desperdício, comparando a quantidade desperdiçada com a quantidade total comprada.

**Tabela 1.5:** Alimentos x Desperdício x Total x Percentual de desperdício

<b>Alimentos</b>	<b>Desperdício</b>	<b>Total</b>	<b>% Desp.</b>
Ovos	68	320	21,25%
Azeitonas	143	1900	7,53%
Cebolas	35	110	31,82%
Tomates	28	100	28,00%
Batatas	45	500	9,00%

Feita a análise, podemos notar que mesmo tratando-se de um exemplo simplório, podemos ilustrar o quanto um estudo minucioso pode deflagrar resultados diferentes. Antes, sem fazer a análise, as azeitonas estariam no topo da lista de alimentos com mais desperdício. Sua quantidade de 143 unidades desperdiçadas era mais do que o dobro da segunda quantidade desperdiçada (ovos com 68). Com esta leitura prematura, teríamos as atenções voltadas para, de fato, o alimento com maior quantidade desperdiçada, mas ao mesmo tempo o alimento com o menor índice de desperdício. Isto é: mesmo desperdiçando muitas quantidades, as azeitonas são usadas em larga escala. Logo, a alta quantidade reduz o impacto das desperdiçadas.

Agora vejamos o tomate, que na leitura apenas das quantidades desperdiçadas ficaria em último lugar. Seu índice de desperdício é bem alto: 28%. Estamos dizendo que, praticamente, a cada quatro tomates comprados, um é desperdiçado. Isto é: se fizermos um movimento para reduzir o desperdício do tomate, por consequência o restaurante passará a comprar menos, pois uma boa parte da compra é para suavizar as perdas.

Ainda assim, com informações complementares e uma análise do obtido podemos melhorar o estudo. Note que até o momento estamos apenas nos atentando para as quantidades desperdiçadas. Mas, e se fôssemos direcionar uma parte do foco do estudo para o valor (dinheiro) desperdiçado? Para isto, iríamos coletar novas informações que pudessem fornecer o custo destes produtos e complementar a análise conforme a próxima tabela:

**Tabela 1.6:** Alimentos x Desperdício x Total x Percentual de desperdício x Valor do desperdiçado

Alimentos	Desperdício	Total	% Desp.	R\$ Desp.
Ovos	68	320	21,25%	R\$ 20,40
Azeitonas	143	1900	7,53%	R\$ 17,16
Cebolas	35	110	31,82%	R\$ 14,00
Tomates	28	100	28,00%	R\$ 9,80
Batatas	45	500	9,00%	R\$ 20,25

Veja como o cenário mudou mais uma vez. Agora, as atenções estão voltadas para ovos e batatas, itens que sequer foram comentados anteriormente. Daí, reforçamos a conclusão que estamos repetindo exaustivamente: o processo estatístico, apenas de ser bem desenhado, pede que seja revisto e melhorado constantemente. Conforme aparecem novas informações, novos questionamentos ou perspectivas, não previstos anteriormente, sobressaem. Contudo, isto não signifique que será sempre assim. A experiência do responsável pelo processo já permitirá que ele se antecipe a algumas destas necessidades e muitas outras poderão ser previstas com uma demanda bem desenhada.

Voltando ao caso em questão, uma demanda bem desenhada seria suficiente para evitar essas pequenas adequações no processo. Pois vejamos: se o objetivo fosse apenas mensurar o alimento com maior quantidade desperdiçada, o primeiro formato seria suficiente e elegeria a azeitona como foco de um ajuste no manuseio de tal forma que reduzisse o desperdício.

Contudo, se o foco fosse sobre o alimento com maior desperdício dentro da quantidade comprada, o segundo modelo seria o ideal. E, com isto bem amarrado, o processo já seria iniciado com as duas informações coletadas, pois, sabendo que o balizador será o percentual de desperdício, o gerente terá ciência da necessidade também da quanti-

dade total de cada alimento. Logo, a conclusão será que as cebolas possuem maior índice de desperdício, exigindo uma melhora no manuseio para que o índice seja reduzido e, por consequência direta, a quantidade total comprada também diminua.

Por fim, se a preocupação fosse a perda financeira em alimentos desperdiçados, desde o início o gerente saberia da necessidade de obter quantidades e valores financeiros envolvidos. Portanto, o último Modelo seria o mais adequado. Nele, a conclusão imediata seria que tanto ovos quanto batatas urgem por uma atenção no que tange à perda financeira.

## Conclusão

Após a leitura desta aula, já é possível ter uma breve noção de como, por alto, funciona o processo estatístico e o quanto ele pode ser importante para uma determinada atividade. Atente-se de que até o momento não usamos termos técnicos, não foram feitos cálculos e tampouco geramos gráficos. O propósito real é apenas amadurecer a ideia da construção de um pensar estatístico e a partir daí absorver toda a sua metodologia.

Uma das partes que precisa ser reforçada é a necessidade de bagagem de conhecimento e maturidade. Obviamente, isto é algo que se adquire com o tempo. Contudo, com as próximas aulas e seus inúmeros exemplos, será possível construir uma pequena gama de casos que lhe proporcionará tomar decisões ou melhor elaborar um processo com antecedência.

Obviamente, para que isso surta efeito, será necessário um esforço mínimo por sua parte: fazer exercícios. E serão muitos! Não adianta achar que apenas lendo conseguirá absorver o necessário, pois isto não funciona. O professor, normalmente, sabe bastante, porque está habitualmente resolvendo exercícios – mesmo que sejam os que ele próprio sugere em sala de aula.

Portanto, mãos à obra que a jornada será longa!!!

## ***Atividade final***

*Atende aos objetivos 2 e 3*

Suponhamos que os mesmos funcionários da Atividade 2 estejam trabalhando na empresa da Atividade 1 e que, coincidentemente, eles ficarão responsáveis pelas mesmas etapas do processo estatístico. O problema da empresa da Atividade 1 já foi levantado. Sendo assim, simule tarefas para os funcionários com o propósito de termos um processo estatístico completo nesta empresa.

This image shows a single sheet of white paper with horizontal ruling lines. The lines are evenly spaced and run across the width of the page. There are no margins, text, or other markings on the paper.

### Resposta comentada

Fernando precisa detectar um problema, isto é, *gerar a demanda*. Neste caso, podemos sugerir que ele detectou que o pacote atual do roteiro não tem vendido tão bem quanto o de seus concorrentes.

Suzana fará a *coleta dos dados*. Portanto, ela pode, por exemplo, levantar os pontos turísticos mais visitados em quantidade de pessoas. De igual modo, pode pesquisar a idade dos clientes da sua empresa, bem como

questionar aos clientes atuais quais os pontos do presente pacote mais lhe agradam e quais que deveriam ser incluídos.

Adriana *organizará* os dados. Montará uma planilha com os clientes entrevistados agrupando-os por idade. Poderá também gerar uma relação com os pontos turísticos da cidade e quantidade de visitantes no último mês.

Mario vai *fazer a análise* dos dados. Vai determinar faixas etárias e contabilizar quantos clientes estão compreendidos em cada uma delas. Estipular a média de visitantes por dia em cada ponto turístico.

Patrícia *elaborará um relatório*. Ela indicará quais dos pontos, que fazem parte do pacote, possuem maior média de visitação e que devem permanecer no mesmo. Indicará, também, quais que não fazem parte do pacote, mas que possuem grande quantidade de visitação, e que, supostamente, deveriam ser incluídos no mesmo. Também pode excluir as faixas etárias que possuem menor incidência de clientes, dando preferência às faixas com maior quantidade de clientes.

Valter *tomará a decisão*. Vai afirmar quais destinos devem permanecer e quais devem ser retirados do pacote, como também quais que não fazem parte e devem ser incluídos. Tudo isto baseado nas informações recebidas.



## Resumo

Nesta aula, pudemos constatar o quanto um processo estatístico pode ser útil para uma empresa escolher um novo serviço a prestar, um novo fornecedor ou até mesmo estabelecer metas a serem alcançadas. Notamos que, após um estudo estatístico, o indivíduo ou organização que o faz está mais bem preparado para decisões ou novas ações que supostamente estará por vir.

Reconhecida a importância da estatística em um aspecto generalizado, fomos apresentados ao processo estatístico. Quais as etapas que o compõe, qual a ordem de executá-lo, em quais momentos o homem interfere e quais momentos os computadores assumem o papel deles, entre outras.

Com o processo estatístico desenhado, fizemos uma prévia de quais informações são importantes para dar prosseguimento ao mesmo, bem como constatamos como uma escolha ruim pode implicar em refazer todo o processo ou induzir à respostas sem significância.

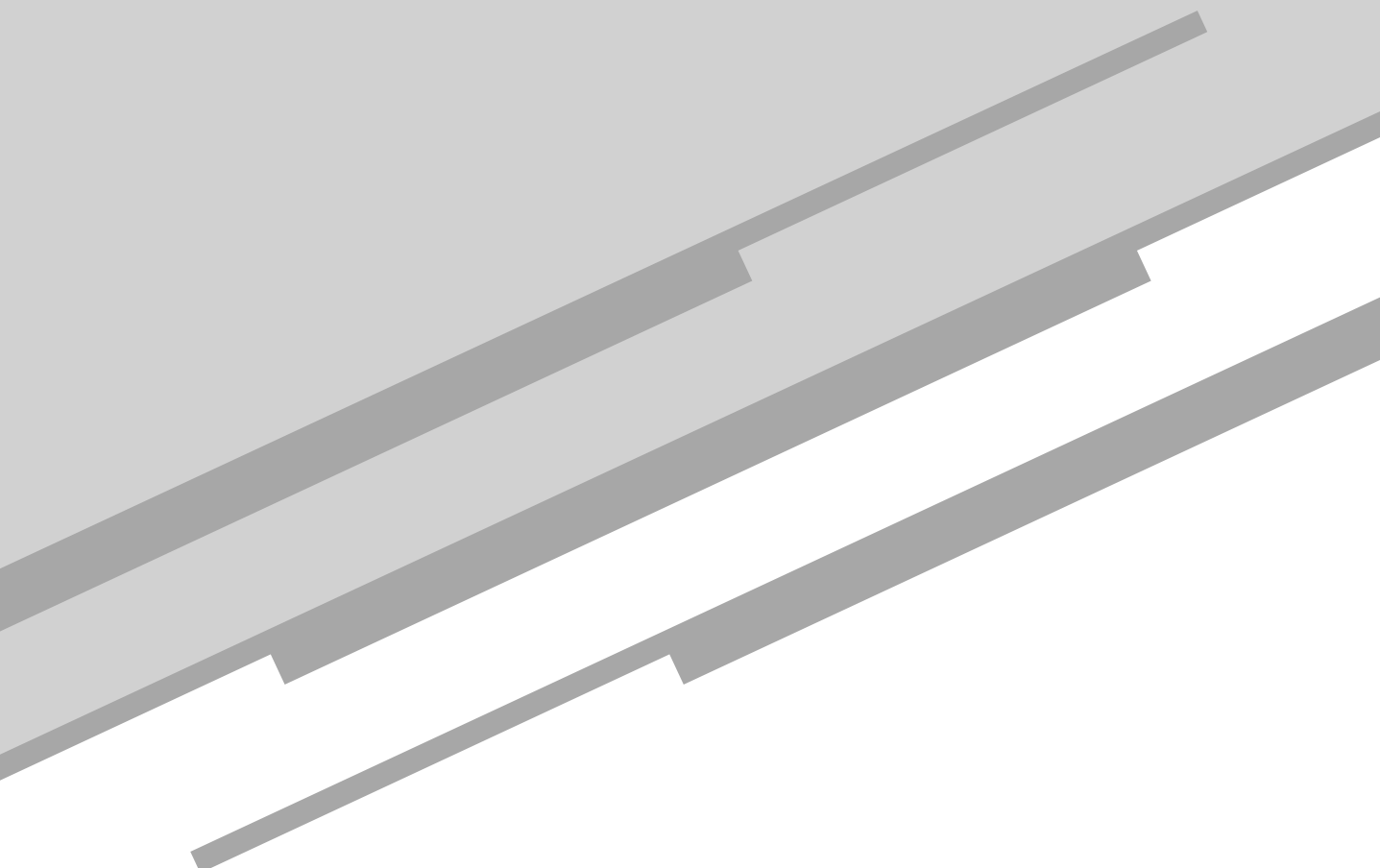
## **Informação sobre a próxima aula**

Na próxima aula, iremos nos aprofundar sobre o vocabulário específico da estatística e seus elementos básicos. Como os dados podem se apresentar e serem classificados também serão objetos do nosso estudo.



# Aula 2

Aquele negócio que “troça” essa coisa



*Rafael Canellas Ferrara Garrasino*

## **Meta**

Apresentar o vocabulário básico – palavra técnica e/ou chave – a ser utilizado no decorrer do curso.

## **Objetivos**

Esperamos que, após o estudo desta aula, você seja capaz de:

1. distinguir os termos população e amostra utilizados na Estatística;
2. estabelecer a classificação de variáveis;
3. desenvolver a leitura de um texto estatístico sem dificuldades.

## Introdução

Senhores, vocês acabaram de ligar a televisão e está passando um documentário médico. Por possuir interesse neste assunto, resolve assistir. Mas, assim que aparece o médico, a primeira frase lhes causa espanto: “Teremos de usar uma tesoura Ultracision para esta histerectomia...”. E agora? Do que ele está falando? Será que conseguirão compreender o restante do programa?

Vamos desligar a televisão! Vamos para um bar! Gostou, né? Na mesa, está você e mais alguns amigos. Um deles pede uma caipirinha e, ao ser servido pelo garçom, ergue o copo e pergunta a todos da mesa: “Uma solução saturada com corpo de chão é heterogênea ou homogênea?”. E aí, vai voltar para casa?

Não adianta! Quando o assunto envolve conhecimento específico, normalmente existe um vocabulário técnico para se referir aos elementos e ações. Isto não significa, necessariamente, que ao saber o significado das palavras consiga entender claramente o que está sendo dissertado naquele assunto, mas pelo menos sabe o que está envolvido. E, convenhamos, nada mais elegante do que alguém falando sobre um assunto usando seu vocabulário técnico do que um leigo com as clássicas frases: “Qual o nome daquele trocinho que faz assim?”; “Isso me lembra aquela coisa que deixa o negócio estranho!”



**Figura 2.1:** Termos técnicos e ruídos na comunicação.

Sabemos bem que estagiário comumente “sofre”, mas no caso da tirinha houve pelo menos um erro. Assim, ou o chefe superestimou a capacidade do estagiário ou o estagiário deveria dominar o assunto quando foi contratado e, por algum motivo, foi aprovado sem que se confirmassem isto.

Aqui no curso de Estatística, mesmo voltando para o público de Turismo, que é bem distinto do público de um curso de Matemática ou Engenharia, o uso de termos técnicos será comumente necessário. Logo, nada mais justo que, no início deste curso, eles sejam devidamente apresentados.

## **Tipos de estatística**

Na realidade, Estatística só existe uma. Contudo, dentro dela, de acordo com os seus objetivos e o que se pretende fazer, existem ramos específicos.

### **Estatística descritiva**

Este ramo coleta, resume e apresenta os dados. As tarefas que fizemos na Aula 1 se encaixam na definição, isto é, tarefas como fazer o levantamento de dados, organizá-los em uma planilha e depois ilustrá-los em um gráfico são chamadas de Estatística Descritiva.

Um dos objetivos mais comuns da Estatística Descritiva é “desenhar” como os elementos estudados se distribuem, o seu comportamento em relação a uma tendência central (assunto o qual veremos mais a frente), suas relações entre si e a variabilidade deles.

### **Estatística inferencial**

Também conhecida como Estatística Indutiva, ela trabalha exclusivamente com a análise e interpretação dos dados, isto é, usa-se generalizações, comparações que podem ser através de princípios ou simplesmente lógicas.

Alguns autores gostam de dizer que a Estatística é uma arte e que a Estatística Inferencial tem como objetivo chegar às respostas corretas, mesmo com um pré-determinado grau de acerto para questões pontuais. Outros preferem definir que a Estatística Inferencial tem como principal potencial tirar conclusões sobre populações (neste caso, população se refere ao termo estatístico que veremos à frente).

De uma maneira simplista e remetendo aos casos que fizemos na Aula 1, podemos induzir a ideia de que na primeira parte do processo estatístico estávamos lidando com Estatística Descritiva e em seguida, na segunda parte, lidávamos com Estatística Inferencial.

## Elementos básicos

Como em toda área técnica, a Estatística também possui um vocabulário básico. Nele estão contidos os termos que mais vezes serão empregados no decorrer do curso.

### População

O termo *população*, que já foi usado anteriormente, refere-se a todos os itens ou indivíduos que serão estudados para se tirar uma conclusão. Portanto, suponhamos que vamos fazer uma pesquisa sobre a intenção de votos para prefeito do Rio de Janeiro. Nossa população a estudar será, necessariamente, todos os eleitores do município do Rio de Janeiro.

Contudo, o termo população induz à ideia especificamente de pessoas. Entretanto, não se aplica apenas a estes casos. Vejamos outros exemplos: se fossemos fazer um estudo sobre resultados na distribuição da primeira carta em mesas de pôquer em um cassino. Pois bem! O jogo envolve cartas e o resultado será sempre um tipo de carta. Logo, nossa população envolverá as cartas do baralho. Note que não mais o termo população será referente a pessoas.

Vejamos outro exemplo: resolvemos estudar o faturamento de uma rede de hotéis em um país nos últimos três anos. Para este estudo, a nossa população será a receita obtida por cada hotel, mês a mês, nos últimos três anos.

Em suma, uma população não necessariamente envolve pessoas e também pode se apresentar de vários tamanhos, desde ligeiramente pequena (exemplo das cartas), grande (exemplo dos hotéis), muito grande ou até mesmo infinita.

### Amostra

A palavra *amostra* será tão utilizada quanto *população*. Ela nada mais é do que uma parcela, um pedaço ou uma parte específica de uma população selecionada para ser estudada.

A finalidade da amostra é reduzir um suposto estudo com uma população muito grande, tomando como objeto apenas uma parte dele. Isto é: quando temos uma população muito grande, pode ser muito

exaustivo ou custoso analisar cada elemento dela. Logo, toma-se uma amostra que irá representar o todo, chegando-se a resultados próximos.

Um dos exemplos clássicos de recorrer a uma amostra é o estudo de intenções de voto como citamos anteriormente. Para se ter uma ideia de como estão as chances de vitória de um candidato, em uma específica eleição, não precisa necessariamente questionar todos os eleitores daquela região. Imagine o quanto demorado seria perguntar para todas as pessoas do município do Rio de Janeiro em quem pretende votar para prefeito.

Exatamente por isso, opta-se por um grupo de pessoas escolhidas aleatoriamente para representar o todo, isto é, uma amostra será determinada para dar um resultado que represente toda a população. Obviamente, sabemos que o resultado não será precisamente idêntico, mas com a “margem de erro” (veremos futuramente do que se trata) chegaremos a um resultado bem próximo.

Ao lidarmos ainda com pessoas, podemos citar estudos que simulam a quantidade de reais (R\$) que cada turista pretende gastar no próximo verão na cidade do Rio de Janeiro. Imagine o trabalho que seria colocar pessoas em todas as saídas dos aeroportos e da rodoviária para fazer a cada turista a mesma pergunta. Daí, mais uma vez estipula-se uma amostra que, com certa margem de confiança, irá representar com segurança toda a população que interessa ao estudo.

Deste modo, se preferirmos, podemos exemplificar sem usar pessoas. Suponhamos uma fábrica de *chips* para microcomputadores. O gerente técnico decide levantar quantos *chips*, ao término do processo de produção, estão com algum tipo de falha. Vamos tentar **mensurar** um número que represente a quantidade de *chips* que uma fábrica produz. Agora, vamos tentar imaginar testar *chip* por *chip*. Pois é. Seria um trabalho tão árduo e desnecessário que, provavelmente, comprometeria o prazo de entrega para os clientes, bem como o custo final do produto. Daí, qual a solução? Adotar uma parte de todos os *chips* produzidos e testá-los. A partir de estudos estatísticos poderemos comprovar que, usando apenas uma amostra desta população, é possível determinar qual o percentual de *chips* que termina o processo de fabricação em condições de uso.

Por fim, se dentro da população cada pessoa, objeto, dado ou item estudado tiver a igual chance ou oportunidade de ser escolhido para formar a amostra, denominamos que é uma amostra aleatória. Em todos os casos citados, estávamos lidando com amostras aleatórias.

## Mensurar

Determinar a medida de; medir; “... tão altos que uma tibia deles mensurava um homem” (Vinícius de Moraes).

## Atividade 1

### Atende ao objetivo 1

Dados os casos abaixo, determine em cada um qual se refere à *população* e qual se refere à *amostra*. Justifique.

a) Uma pesquisa feita em um colégio particular que irá determinar quantos meninos existem na oitava série.

---



---



---

b) Uma pesquisa feita com os carros populares da cidade do Rio de Janeiro para determinar, dentre os modelos importados, quantos são roubados.

---



---



---

c) Uma pesquisa em uma banca de jornal para determinar quantas revistas sobre moda encalham ou não são vendidas até ao final do mês.

---



---



---

### **Resposta comentada**

a) População: todos os alunos do colégio particular em questão. Amostra: composta pelos alunos da oitava série.

a) População: veículos populares na cidade do Rio de Janeiro. Amostra: Veículos importados populares na cidade do Rio de Janeiro.

a) População: todas as revistas da banca de jornal em questão. Amostra: todas as revistas de moda desta mesma banca de jornal.

---



---



---

## Variável

O tempo todo estudaremos pessoas, objetos, dados, sejam eles de uma população ou de uma amostra. Para estudarmos, precisaremos das suas características, valores, informações etc. Isto que precisaremos é o que chamamos de *variável*.

Cada objeto de um estudo pode ter mais de uma variável e, conforme a própria definição da palavra, com o passar do tempo, estas variáveis podem mudar. Suponhamos um estudo com os alunos deste curso. Logo, você é um dos objetos dele. Vejamos quantas variáveis você possui: idade, sexo, nacionalidade, altura, peso, cor dos cabelos, estado civil e número de filhos. Note que apesar de ter citado apenas oito variáveis, existem uma infinidade delas que você mesmo pode levantar. Outro fator importante é perceber que algumas são representadas por número (medidas) e outras por palavras. Sendo que quase todas, exceto pela nacionalidade, se forem questionadas daqui a um ano, podem denotar em respostas diferentes.

As variáveis nem sempre precisam estar associadas diretamente às características físicas ou sociais de um indivíduo. Vejamos na perspectiva de um gerente de hotel. Para ele, a cor do cabelo de um hóspede é tão importante quanto a frequência com que ele corta as unhas dos pés. Ainda assim, um hóspede para ele pode possuir diversas variáveis como: quantidade de diárias, consumo de frigobar, gastos com serviços de lavanderia, despesas no café da manhã etc.

O mesmo valerá se estivermos lidando com elementos mais abstratos como investimentos. Suas características podem ser a rentabilidade anual, a variação da rentabilidade nos últimos meses, suas tarifas e descontos embutidos etc. Como foi visto, um objeto de estudo pode ter desde uma variável até diversas. Estas variáveis possuem comportamentos distintos, logo podendo ser caracterizadas de formas diferentes: as Qualitativas e as Quantitativas.

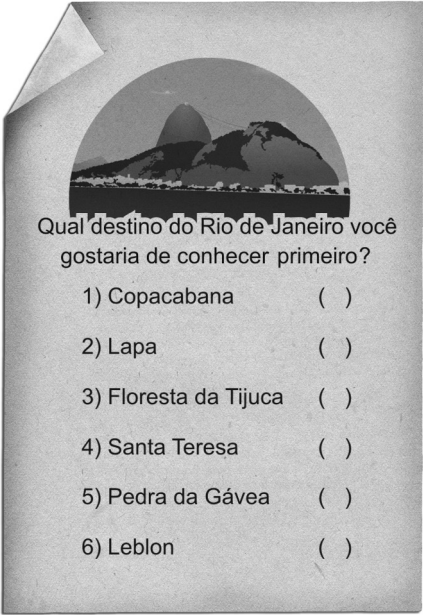
### Variáveis qualitativas

São variáveis que, conforme o próprio nome diz, expressam uma qualidade ou atributo. A maneira mais prática de identificar uma variável qualitativa é quando usamos uma palavra para expressá-la. Comumente, variáveis qualitativas também são chamadas de *variáveis categóricas*.

No exemplo no qual usamos você como objeto para levantar algumas de suas variáveis, podemos afirmar que, das citadas, temos como variáveis qualitativas a cor dos seus cabelos, seu sexo, nacionalidade e estado civil.

Como falamos antes, uma parte da Estatística se destina na organização destes dados. Esta organização se faz baseada nas variáveis que estão sendo consideradas para o estudo. Em um estudo que trabalhe com variáveis qualitativas, formas de mensuração podem ser feitas na escala nominal e na escala ordinal.

Na *escala nominal* não existe necessariamente uma hierarquia entre as variáveis. Isto é: em um questionário sobre já ter visitado o Tibet, as opções de respostas são “sim” e “não”. Essas variáveis não possuem entre si uma relação que determine qual vale mais que a outra, quem vem primeiro etc. – ainda que optemos por enumerar as respostas. Assim, digamos que o questionário é sobre qual destino você primeiro gostaria de conhecer no Rio de Janeiro conforme a **Figura 2.2**.

A questionnaire form with a header image of Rio de Janeiro's skyline, including Sugarloaf Mountain. Below the image is the question: "Qual destino do Rio de Janeiro você gostaria de conhecer primeiro?". Below the question is a numbered list of six destinations, each followed by a pair of parentheses for a response.

Qual destino do Rio de Janeiro você gostaria de conhecer primeiro?	
1) Copacabana	( )
2) Lapa	( )
3) Floresta da Tijuca	( )
4) Santa Teresa	( )
5) Pedra da Gávea	( )
6) Leblon	( )

**Figura 2.2:** Locais de visita  o na cidade do Rio de Janeiro.

Note, na **Figura 2.2**, que a numera  o provavelmente foi criada para facilitar aos pesquisadores inserir as informa  es em uma tabela. Ao inv  s de digitar Santa Tereza sempre que um entrevistado der essa resposta, ele coloca apenas 4. Contudo, nessa escala nominal n  o necessariamente signifique que 6 (Leblon) seja maior que 2 (Lapa).

No entanto, o pesquisador pode até optar por agrupar as respostas. Isto é: formar o grupo Praia, composto por Copacabana e Leblon; o grupo Ecoturismo, com Floresta da Tijuca e Pedra da Gávea; o grupo Boêmia, formado por Lapa e Santa Tereza. Mas, ainda assim, não teremos uma relação de ordem ou superioridade entre os grupos. Teremos os dados mais bem organizados, mas a prioridade continuará tão subjetiva quanto antes.

Na *escala ordinal* já possível estabelecer um critério de ordem nos objetos de estudo de acordo com a sua variável em questão. Vejamos, por exemplo, uma pesquisa que trate os cargos dos funcionários de uma empresa. Teríamos como opções: diretor, gerente, analista, assistente e estagiário. Dentro destas variáveis é possível estabelecer uma ordem de acordo com o grau de importância dentro da empresa.

Se estivéssemos fazendo uma pesquisa sobre a satisfação dos seus clientes com a comida do seu novo *chef*. Neste caso, dentro das respostas do formulário teríamos: muito insatisfeito, insatisfeito, indiferente, satisfeito e muito satisfeito. Com estas características é possível estabelecer um critério para ordenar as variáveis de tal forma que os clientes que aprovassem o seu novo *chef* ficassem no topo do resultado.

Ainda nessa última pesquisa é possível fazer um agrupamento semelhante ao que foi feito com as respostas da Figura 2.2. A sugestão inicial seria de muito insatisfeito e insatisfeito no grupo de clientes a rever o atendimento, indiferente no grupo de clientes a conquistar mais rápido e muito satisfeito e satisfeito no grupo de clientes a manter a fidelização. Sendo assim, percebe-se que continua possível manter uma hierarquia entre os grupos. O grupo para manter a fidelização permanece no topo, seguido pelos demais que, quando desmembrados, também ficavam abaixo dele.

Esta ordenação dos dados não se trata apenas de algo estético ou pragmático. Posteriormente, quando tivermos estudando uma amostra ou população em relação às suas tendências centrais ou *percentis*, notará que somente com as mesmas devidamente ordenadas isto será possível.

## Variáveis quantitativas

De maneira simétrica às variáveis qualitativas, as variáveis quantitativas são as expressas por números. Isto é: trata-se de um resultado de medição, contagem ou após cálculos matemáticos. Elas também são conhecidas como *variáveis numéricas*, por motivos óbvios.

Retomando ao que levantamos no início do Tópico 2.3, temos que, das suas características, as variáveis quantitativas são idade, peso, altura e quantidade de filhos. Note que para estas variáveis os valores têm comportamentos distintos. Enquanto para peso e altura o resultado é obtido após um processo de medição. Para quantidade de filhos e idade o resultado é fruto de um processo de contagem. Esta diferenciação vai determinar uma subdivisão nas variáveis quantitativas: discretas e contínuas.

As *variáveis quantitativas discretas* são frutos de processo de contagem. Logo, só admitem valores inteiros. Sendo objeto de estudo a quantidade de lâmpadas queimadas devolvidas em uma loja específica, a variável será quantitativa discreta. Caso o objeto seja a quantidade de crianças aprovadas no terceiro bimestre em uma determinada escola, a variável será também uma quantitativa discreta.

Outros casos geram mais polêmica, como, por exemplo, idade. Ao perguntar quantos anos uma pessoa tem, supostamente a resposta será um número inteiro. Mesmo não sendo o dia exato de seu aniversário, raramente uma pessoa responde que possui 18 anos e meio ou 29 anos e três quartos. Contudo, sempre existe um implicante ou alguém dizendo que a idade pode ser um número quebrado. Logo, não seria uma variável quantitativa discreta. Enfim, o propósito não é se aprofundar em temas polêmicos, mas, sim, conceituar bem a ideia de variável quantitativa discreta. Portanto, deixemos estas discussões para o pós-aula.

Por sua vez, *variáveis quantitativas contínuas* são resultados de processos de medição ou operações matemáticas. Portanto, admitem valores não inteiros, isto é, qualquer valor do conjunto dos reais. Exemplos comuns são: altura de uma pessoa, peso de um produto, variação do câmbio de uma moeda estrangeira, faturamento de uma empresa etc. Todos estes permitem a possibilidade de assumirem valores decimais. Por fins práticos e como forma de padronizar os resultados no decorrer do curso, iremos adotar a metodologia de casas decimais conforme o quadro a seguir.



Arredondamento: o padrão será de duas casas decimais. Contudo, durante a execução dos cálculos, cabe ao aluno optar em trabalhar com todas as casas e só arredondar ao final ou efetuar a cada operação um novo arredondamento. Fica salientado que o aluno que optar pelo segundo método, potencialmente encontrará resultados ligeiramente diferentes devido às perdas nos arredondamentos durante os cálculos. O padrão para arredondamento será: se a última casa a arredondar for menor do que 5, os números restantes não sofrerão alteração: 2,351 vira 2,35; 15,4728 vira 15,47; 23,990678 vira 23,99. Caso a última casa a arredondar seja igual ou maior do que cinco, aumentaremos em uma unidade a segunda casa decimal: 2,438 vira 2,44; 18,3682 vira 18,37; 42,995032 vira 43.

Assim como nas variáveis qualitativas, as variáveis quantitativas precisam ser organizadas pelos mesmos motivos. Contudo, como as formas de obtenção dessas variáveis podem ser diferentes, faz-se necessário um parâmetro para que sejam classificáveis. O melhor parâmetro para valores reais é o zero, que neste caso será um zero absoluto.

As variáveis quantitativas podem ser mensuradas em uma *escala de razão*, também conhecida como *escala proporcional*. Nesta, o valor zero é absoluto. Por exemplo, o peso de um produto. Podemos afirmar que se o peso for zero, o produto inexistente. Da mesma forma que se um produto pesa 60 quilos, pode-se dizer que ele tem o dobro de peso de um produto de 30 quilos.

Assim, outro exemplo é a renda de um indivíduo. Ao comparar uma pessoa que possui renda de R\$ 15.000 com uma pessoa de renda de R\$ 3.000, está correto afirmar que um ganha cinco vezes mais do que o outro. Neste caso, o parâmetro zero é existente. Isto é: zero reais significa a ausência de dinheiro.

Por sua vez, quando não temos um zero absoluto, isto é, o zero é resultado de uma convenção, dizemos que as variáveis quantitativas serão mensuradas por uma *escala intervalar*. O exemplo mais comum é a medição da temperatura em Celsius. Pode-se dizer que um dia com a tempera-

tura de 5° foi 10° mais frio do que o dia anterior que teve temperatura de 15°. Mas não é correto dizer que a temperatura caiu em um terço de um dia para outro. Isto porque o zero na Escala Celsius não é absoluto; ele é uma convenção adotada para quando a água congela. Esta afirmação só poderia ser feita se a temperatura tivesse sido medida na Escala Kelvin, na qual o zero significa a total ausência de movimentação molecular.

## Atividade 2

### Atende aos objetivos 2 e 3

Em uma empresa de serviços hospitalares existem aproximadamente 5.000 funcionários de diversas idades e formação. Foi feita uma pesquisa com seus funcionários homens, questionando a idade, a renda mensal e a formação. Dito isto, responda o que se pede.

a) Qual é a população deste estudo?

---



---

b) Qual é a amostra deste estudo?

---



---

c) Qual ou quais são as variáveis deste estudo?

---



---

d) Classifique as variáveis identificadas em qualitativas ou quantitativas.

---



---

### **Resposta comentada**

a) Funcionários da empresa de serviços hospitalares.

b) Funcionários homens da mesma empresa.

c) Idade, renda mensal e formação.

d) Idade e renda mensal são quantitativas; formação é qualitativa.

---



---



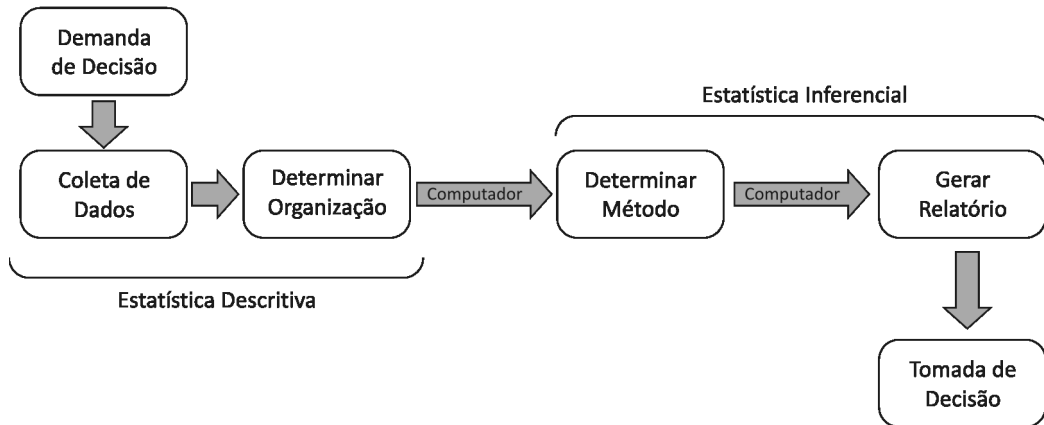
---

## Conclusão



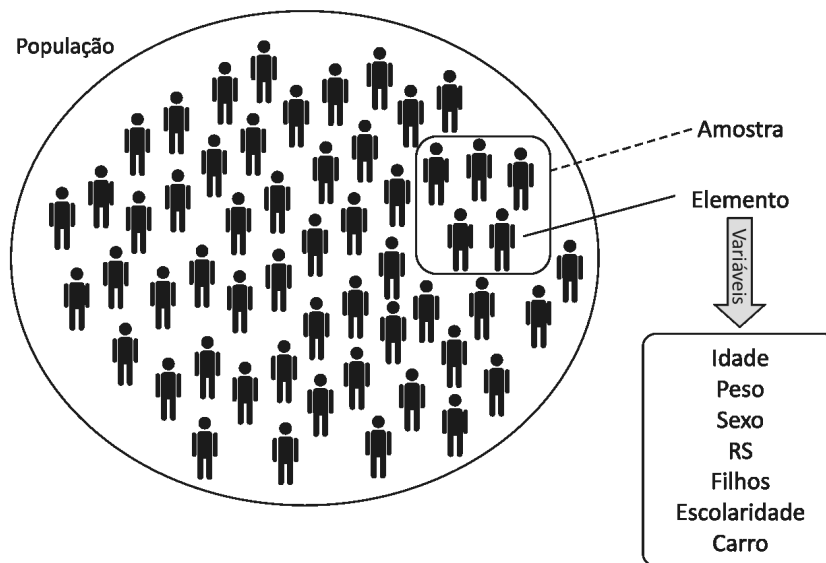
**Figura 2.3:** Termos técnicos da estatística.

Neste Aula, pudemos ser apresentados ao vocabulário básico da Estatística. Obviamente, com o aprofundar do conteúdo, novos termos surgirão, mas também serão apresentados mais oportunamente. Inicialmente, descobrimos que existem dois ramos da Estatística que andam separadamente, mas ao fim se complementam, formando o estudo final. A **Figura 2.4** é uma melhoria da **Figura 1.2** da Aula anterior com o detalhamento dos ramos apresentados.



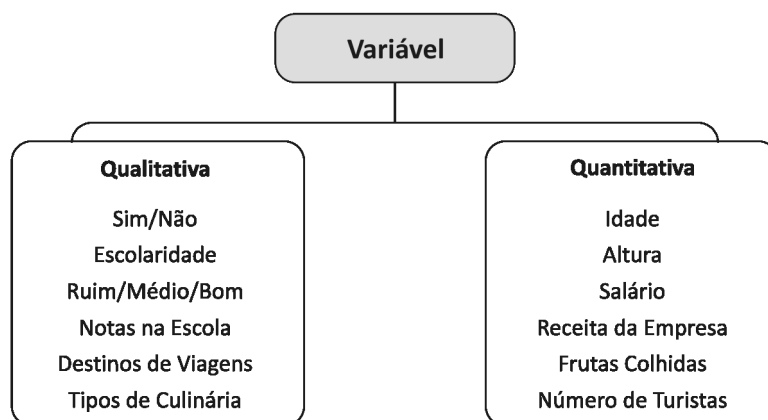
**Figura 2.4:** Processos estatística descritiva e inferencial para tomada de decisão.

Posteriormente, tivemos a oportunidade de entender o que é *população* no sentido estatístico e o que é uma *amostra*. Notamos que, na existência de uma população excessivamente grande, podemos recorrer a uma parte dela (amostra) para obter uma leitura por igual de todos. Esta leitura é feita através de uma das várias características que cada elemento da amostra/população pode ter, que chamamos de *variável*. A **Figura 2.5** ilustra a ideia de amostra fazendo parte de uma população, ambos sendo compostos por um grupo de elementos, nos quais existem diversas informações que são as variáveis.



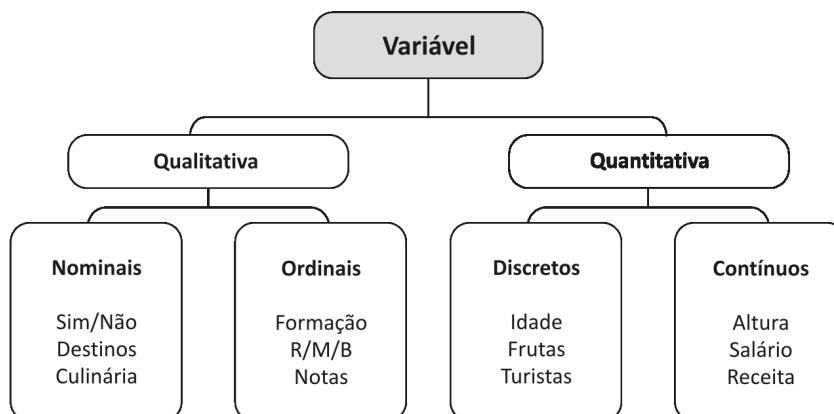
**Figura 2.5:** Amostra de população e variáveis.

Daí entendemos que o grande foco do nosso estudo será a variável de cada elemento. Ela que será, após coletada e organizada, a informação propriamente dita que estudaremos. Para tal, foi categorizado que elas se dividem em dois tipos, conforme a **Figura 2.6**:



**Figura 2.6:** Variáveis qualitativa e quantitativa.

Observamos, também, que dependendo do tipo de variável, ela possui uma forma diferente de ordenação, quando possível, além de maneiras diferentes de serem geradas. A possibilidade de uma variável ser fruto de uma contagem ou uma medição, ou até mesmo a inviabilidade de ser ordenada, nos obriga a subdividi-las em categorias distintas, conforme a **Figura 2.7**:



**Figura 2.7:** Variáveis qualitativa e quantitativa e as suas categorias.

Assim, podemos praticamente encerrar esta Conclusão, pois, se houve uma clara concepção do que foi explicado nesta aula, para recordar, a simples consulta às últimas figuras será suficiente. Cabe, apenas, a execução dos exercícios propostos para fixação.

---

---

---

---

---

---

### **Atividade final**

---

---

---

---

---

---

*Atende aos objetivos 2 e 3*

1. Identifique no formulário a seguir, quais perguntas terão respostas que serão classificadas como variáveis quantitativas e quais serão classificadas como qualitativas. Justifique:

a) Qual é o seu nome?

---

---

b) Qual é a sua idade?

---

---

c) Você é filho único?

---

---

d) Considerando notas inteiras entre 0 (pior nota) e 10 (melhor nota), como avaliaria nosso refeitório ou o restaurante onde almoças ou jantas?

---

---

e) Quanto gasta em média no nosso refeitório ou no restaurante onde almoças ou jantas?

---

---

2. Identifique as variáveis a seguir como ordinárias, discretas, nominais ou contínuas. Justifique.

a) Peso de um pedaço de carne.

---

---

b) Capitais dos Estados do Brasil.

---

---

c) Patentes da Polícia Militar.

---

---

d) Quantidade de veículos que passaram pela Ponte Rio-Niterói.

---

---

e) Limite do amigo-oculto da família.

---

---

3. Faça o arredondamento dos valores abaixo conforme foi acordado nesta aula:

a) 18,4683

---

b) 23,120008

---

c) 68,369

---

d) 71,95

---

e) 48,991

---

**Resposta comentada**

1.

a) Qualitativa.

b) Quantitativa.

c) Qualitativa.

d) Quantitativa.

e) Quantitativa.

2.

- a) Contínua.
- b) Nominal.
- c) Ordinária.
- d) Discreta.
- e) Contínua.

3.

- a) 18,47.
- b) 23,12.
- c) 68,37.
- d) 71,95.
- e) 48,99.



## Resumo

Nesta aula, pudemos conhecer os termos mais utilizados na Estatística e suas funcionalidades. Assim, a compreensão de um texto estatístico ficou menos complexa. O mesmo valerá quanto a elaborar um texto também estatístico. Outro objetivo alcançado foi a classificação das *variáveis*, objeto principal da Estatística. Com elas classificadas, a sua ordenação, agrupamento e método de análise ficam mais dinâmicos. Isto é: foi dado um grande passo na parte de elaboração e interpretação de um estudo estatístico.

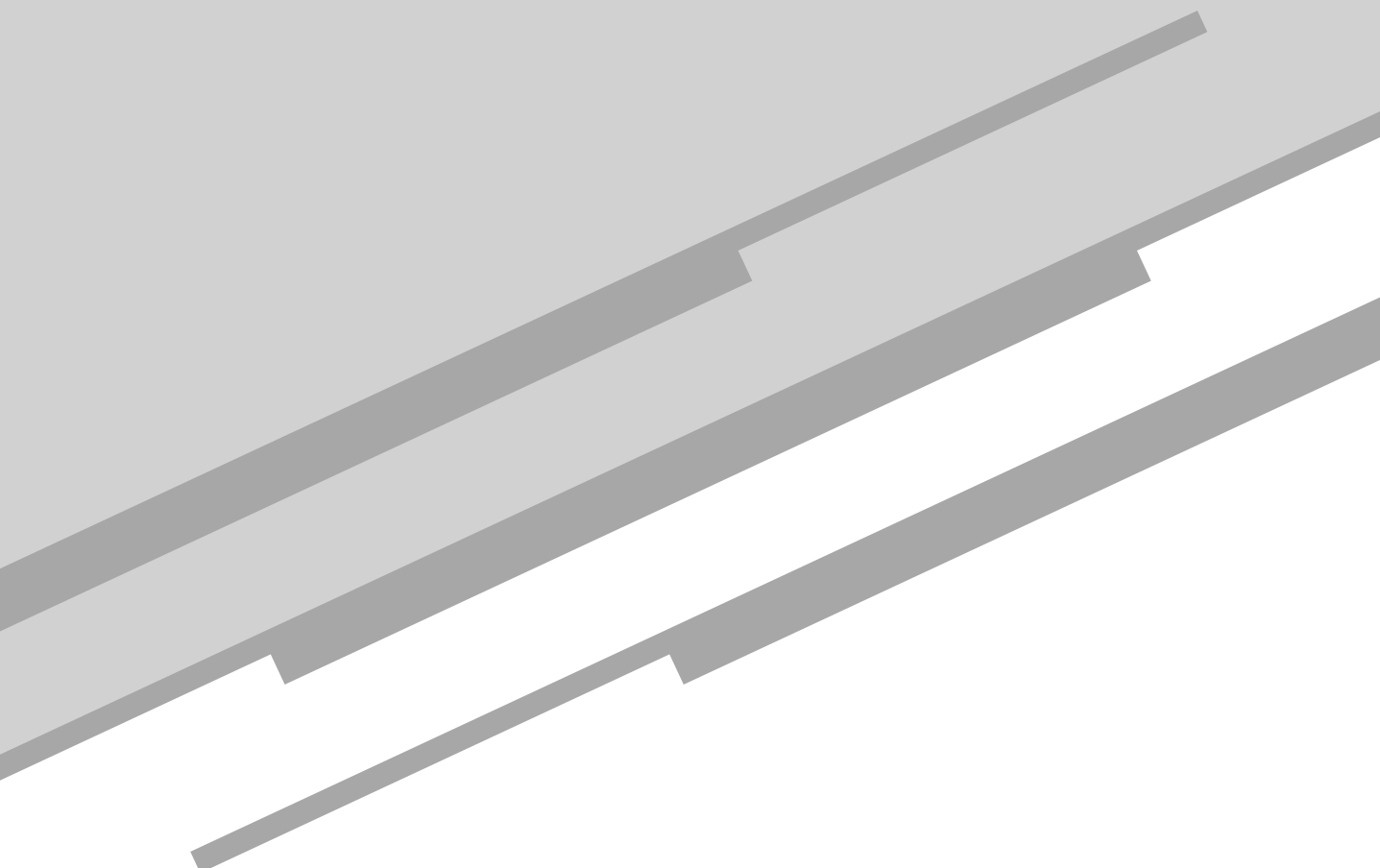
## Informação sobre a próxima aula

Na próxima aula, iremos entender melhor como montar uma pesquisa. Quais os tipos de formulários, como usá-los e como prepará-los. Veremos alguns erros comuns e situações que potencialmente poderão induzir o resultado de uma pesquisa; deste modo invalidando-a.



# Aula 3

Senhor, um minuto da sua atenção, por favor!



*Rafael Canellas Ferrara Garrasino*

## **Meta**

Apresentar os elementos básicos de uma pesquisa – com ênfase nos tipos de perguntas que existem e no modo de elaborá-las para o eficaz estudo estatístico.

## **Objetivos**

Esperamos que, após o estudo desta aula, você seja capaz de:

1. estabelecer as fontes de dados, primárias ou secundárias, para um estudo estatístico;
2. identificar os tipos de questionários que podem ser feitos em cenários específicos: perguntas de múltipla escolha, dicotômicas ou abertas;
3. estabelecer pesquisas que sejam diretas ao propósito do estudo em questão.

## Introdução

Senhores, cada dia que passa somos submetidos a pesquisas. Após ligar para uma central de *telemarketing*, sempre existe a tal pesquisa de opinião. Uma famosa rede de farmácias, aqui no Rio de Janeiro, possui terminais nos próprios caixas para que se possa avaliar o atendimento. Cursos de pós-graduação fazem comumente pesquisas com seus alunos e para alguém que passe mais de cinco horas na rua é impossível não ter a oportunidade de preencher um formulário ou de responder algumas perguntas. Imagine, então, alguém que passe horas navegando pela *internet*.

Pois bem. Na Estatística, como já sabemos, tudo começa com um processo de coleta de dados. Para tal, existem algumas maneiras de acessá-los – desde questionários até experimentos sofisticados. Como fazê-los? A quais recorrer? Como funcionam?

Certa vez, na época do colégio, a professora pediu para que coletássemos preços de produtos para um trabalho sobre porcentagem. A ordem foi clara. Deveríamos pegar encartes de supermercados diferentes e recortar alguns produtos com seus respectivos preços.

Como morava na parte mais central da Tijuca, que dispõe de muitos supermercados, em uma tarde tinha algo em torno de dez encartes diferentes. Já em casa, recortei uma média de 15 produtos de cada encarte. Pronto! Já possuía uma amostra de 150 elementos. Bastava aguardar a aula para exibir a minha dedicação.

No dia da aula, após as orientações da professora para o trabalho, descobri o problema que tinha em mãos. O objetivo era comparar o preço de um mesmo produto em supermercados diferentes e calcular a diferença percentual entre eles. Por ironia, percebi que não tinha um produto repetido sequer. Acredito que na vontade de mostrar eficiência, optei instintivamente pela maior diversidade de itens possível. Qual seria a causa deste meu problema?

Obviamente o problema estava em uma amostra incompatível com o estudo em questão. Naquele momento, não sabia mais afirmar se a ordem da professora foi incompleta, confusa ou se, simplesmente na ansiedade de começar a pensar no que iria fazer, só ouvi uma parte.

Percebam que em uma situação simplista do cotidiano podemos notar o quanto um estudo estatístico pode ser comprometido se a coleta dos dados não for compatível com as necessidades. Naquele momento, tínhamos 150 elementos na amostra e nenhum era útil, enquanto que, possivelmente, um colega ao lado com apenas 8 elementos seria capaz de desenvolver o trabalho com total eficiência.

## As fontes de dados

Como sabemos, todo estudo estatístico é voltado para os dados. Eles são o objeto maior desta Ciência. Para tal, o processo de acesso a eles é tão importante como qualquer etapa do processo. Contudo, por ser a primeira etapa, pode comprometer todo o restante se não for bem executado.

Os dados podem ser classificados quanto ao tipo de fonte de duas maneiras: as Fontes Primárias e as Fontes Secundárias.

### Fontes Primárias

Em uma analogia com alimentos, é possível afirmar que as Fontes Primárias são como o leite recém tirado da vaca. São as informações coletadas diretamente para um determinado estudo. Isto é: uma mesma entidade faz a coleta dos dados e os analisa. O processo é todo feito por uma única empresa, organização ou pessoa. Um bom exemplo de fontes primárias é o censo realizado pelo IBGE. A própria instituição faz a pesquisa, organiza os dados, faz a análise deles e chega a um resultado.

### Fontes Secundárias

Na mesma analogia com alimentos, a Fonte Secundária seria o leite transformado em queijo. Isto é: a entidade (que pode ser uma pessoa, empresa ou organização) acessa as informações originais (Fontes Primárias) de outro estudo e com elas, a partir daí, elabora seu processo estatístico.

Voltemos ao estudo do IBGE. Naquele momento tínhamos uma fonte primária de dados. Contudo, quando um pesquisador qualquer desenvolve um estudo sobre homens acima de 40 anos, por exemplo, e opta por aproveitar os dados do IBGE, para o pesquisador estes dados serão de uma fonte secundária.

Para desenvolver um estudo estatístico, como já falamos, precisamos de dados. De acordo com as circunstâncias e necessidades, você pode optar por fazer a própria coleta ou recorrer aos dados de uma pesquisa já feita por outra pessoa. Ao realizar a pesquisa, estes dados serão de uma fonte primária. Quando se usa de outra pesquisa, eles serão uma fonte secundária.

### A coleta de dados

Ainda sobre a coleta de dados é importante conhecer as quatro maneiras mais comuns de se fazê-lo.

## Dados de uma organização ou indivíduo

Esta forma de coleta se enquadra no perfil de fontes secundárias. O processo estatístico será feito por uma entidade, contudo os dados que serão utilizados foram cedidos por outra entidade ou outro indivíduo.

Isto é muito comum em trabalhos acadêmicos e publicações. Neles, o autor acessa as informações de outros estudos e toma conclusões ou dá prosseguimento ao estudo original. Encontramos também em notícias de jornais ou colunas especializadas. Normalmente, nestes casos, não se trata de um estudo estatístico, mas pode ser usado, como exemplo, a forma como recorrem aos dados de fontes primárias para usá-los como fonte secundária.

## Experimento e testes

Este método é muito comum na indústria e na engenharia. Para tirar conclusões sobre a eficácia de um determinado produto são feitos exaustivos testes. Ali são medidos os resultados, o funcionamento ou a durabilidade do produto em questão para, posteriormente, serem avaliados dentro de um processo estatístico.

Suponhamos que uma fábrica de artefatos voltados para esportes radicais deseje lançar um novo modelo de corda de segurança. É imprescindível que na embalagem contenha informações básicas como limite de peso suportado ou quantas vezes pode ser utilizado até ser descartado. Para tal, podemos especular que várias desta corda serão utilizadas. Todas serão testadas colocando pesos de vários tamanhos de forma crescente até arrebentar. Quando isto acontecer, será anotado o peso que provocou o arrebentar da corda. Ao final, terá o peso que arrebentou cada corda. A partir daí iniciará a segunda parte do processo estatístico até se concluir qual é o peso limite para o uso seguro da nova corda.

## Observações

É muito comum que confundam o método de experimentos com observações. Contudo, a diferença crucial está no fato de que em experimentos as ações foram projetadas e induzidas, enquanto em observações o pesquisador não interfere em momento algum.

Geralmente, utiliza-se o *método de observações* em estudos que envolvam comportamento, principalmente em empresas. Escolhe-se um grupo de um determinado setor e é observado como se comporta em equipe, em dinâmicas etc.

---

---

---

## Atividade 1

---

---

---

### Atende ao objetivo 1

Determine nas pesquisas a seguir se as fontes dos dados são primárias ou secundárias. Justifique.

a) Um engenheiro arremessou 50 pratos, do topo de uma mesa, para testar a sua durabilidade.

---

---

---

b) De posse das marcações de tempo feitas pela equipe Ferrari, um repórter da Rede Tombo fez um estudo para apresentar no programa noturno da emissora.

---

---

---

c) Alunos de um curso técnico de meio ambiente fizeram questionamentos aos visitantes do Jardim Botânico para apresentar, posteriormente, um relatório ao professor.

---

---

---

d) Um médico coletou as publicações dos Ministérios da Saúde de vários países e, em seguida, iniciou um experimento para o tratamento de queda de cabelos.

---

---

---

### Resposta comentada

a) Fonte primária: o próprio engenheiro está gerando dados para seu estudo.

b) Fonte secundária: o repórter está utilizando dados coletados por outra entidade.

c) Fonte primária: os alunos coletaram e analisarão os dados.

d) Fonte secundária: apesar de o estudo ser do médico, os dados foram gerados por outras organizações.

---

---

---

## Pesquisa

Talvez seja o mais conhecido de uma maneira geral. Normalmente utilizado para medir satisfação de um grupo de pessoas com um produto, programa ou serviço. Também pode ser utilizado na simulação de mudanças ou lançamento de um novo produto.

Um estudo sobre cremes de pele, por exemplo, pode ser feito entrevistando os seus respectivos consumidores. Contudo, se for feito como experimento projetado, talvez tenha um resultado diferente. A diferença estará que no primeiro estamos medindo a sensibilidade dos seus consumidores, enquanto que no segundo determinamos a eficiência do produto no ponto de vista técnico. Para este curso, usaremos na maior parte das vezes a *pesquisa* como fonte de dados. Logo, iremos nos aprofundar um pouco mais nesta.

## Métodos de pesquisa

Hoje, com a globalização e as demais ferramentas tecnológicas de comunicação, é possível ter acesso às pessoas nas mais diferentes formas. Ainda assim, os métodos mais utilizados para se fazer uma pesquisa são: o *Questionário* e a *Entrevista*. Ambos, contudo, são compostos por perguntas que, após respondidas, fornecerão informações, que reunidas possibilitarão uma conclusão sobre um determinado produto, serviço, pessoa, hábito etc.

Em um *questionário*, basicamente temos um formulário no qual a pessoa irá preencher suas respostas conforme solicitadas. Em uma perspectiva ampla, o questionário não deixa de ser uma entrevista, contudo, a ausência de uma pessoa fazendo as perguntas permite que seja classificado diferentemente.

A sua principal característica é a possibilidade do público-alvo fornecer os dados reservadamente. Isto é: não necessariamente precisamos ter uma pessoa envolvida com a pesquisa por perto tomando nota das respostas. Esta característica pode ser considerada boa e ruim ao mesmo tempo.

No ponto de vista de honestidade nas respostas, a possibilidade de o entrevistado poder opinar sem um terceiro ao lado é muito boa. Assim, não fica coagido ou influenciado a dar apenas respostas positivas. Sozinho, o questionado pode ser mais honesto e fornecer respostas que mudem a forma de se interpretar o resultado. Contudo, a mesma possibilidade de responder sozinho pode ser ruim caso as perguntas não

sejam claras ou suficientemente autoexplicativas. Um questionado, que tiver dificuldades, pode, por não ter a quem recorrer, fornecer respostas “de qualquer jeito” comprometendo o resultado.

Antes, os questionários eram encontrados em panfletos, correspondência ou encartes de produtos. Hoje já recebemos estes questionários por e-mail, mensagem (SMS), ao encerrar uma ligação ou nos próprios *sites* onde fazemos compras ou consultas. Este fácil acesso, a quem será supostamente questionado, pode ser considerado ruim se avaliarmos que nem sempre a própria pessoa estará respondendo ou, quando realmente for, pode estar fazendo sem atenção apenas para encerrar. Todavia, permite que possamos obter informações de uma infinidade de pessoas, independentemente da distância, inclusive.

Já na *entrevista* temos, necessariamente, uma pessoa envolvida com o processo de coleta de dados que irá interagir com o questionado. Esta pessoa, que fará as perguntas, anotará as respostas e, quando necessário, interagirá com o entrevistado. Muitas das vezes, nas quais a pessoa acaba interagindo com o questionado, trata-se de um momento de dúvida ou de melhor esclarecimento do que está sendo perguntado. Contudo, não muito comumente, existem casos nos quais o entrevistador acaba influenciando a resposta. Este evento não pode jamais ocorrer. Qualquer entrevista que houver a suspeita de influência deve ser sempre descartada.

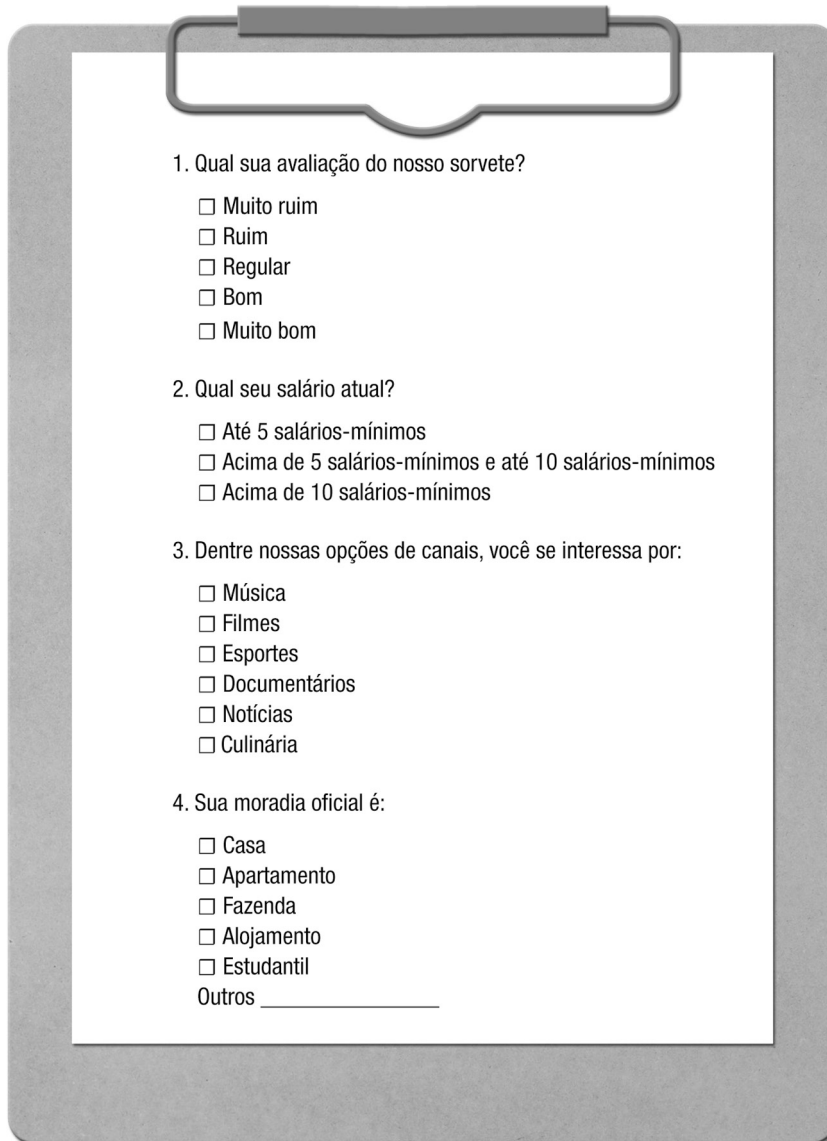
Um entrevistador habilidoso é capaz de obter respostas com mais naturalidade das pessoas, detalhes que poucos notariam e, obviamente, um resultado mais eficaz que um questionário. Todavia, este é um recurso de alto custo e, talvez por isto, algumas empresas optem pelo questionário.

Ao elaborar um questionário ou uma entrevista, é necessário sempre fazer um “piloto”. Isto é, um teste no qual, com um grupo misto de pessoas, as perguntas serão testadas. Nele, verificamos se consideraram as perguntas claras, as opções de respostas suficientes, a ordem dos questionamentos – entre outros detalhes que precisam ser revistos e testados antes de “ir para a rua”.

Basicamente, um questionário ou entrevista é composto por perguntas. Contudo, existem tipos específicos de perguntas. Cada uma com uma característica, finalidade e estruturação. Vejamos elas: as mais comumente utilizadas são as de *múltipla escolha*, pois além de fornecer aos entrevistados as opções que te interessa, este método é bastante popular por ser utilizado em avaliações escolares, por exemplo.

Deste modo, ao estruturar uma pergunta de múltipla escolha é necessário primeiro prever quais os resultados que serão do seu interesse,

pois todas deverão constar no questionário para o entrevistado responder. Ainda assim, imprevistos podem acontecer conforme comentaremos ao final da aula. Vejamos antes alguns exemplos de perguntas na **Figura 3.1**.



1. Qual sua avaliação do nosso sorvete?

- ☐ Muito ruim
- ☐ Ruim
- ☐ Regular
- ☐ Bom
- ☐ Muito bom

2. Qual seu salário atual?

- ☐ Até 5 salários-mínimos
- ☐ Acima de 5 salários-mínimos e até 10 salários-mínimos
- ☐ Acima de 10 salários-mínimos

3. Dentre nossas opções de canais, você se interessa por:

- ☐ Música
- ☐ Filmes
- ☐ Esportes
- ☐ Documentários
- ☐ Notícias
- ☐ Culinária

4. Sua moradia oficial é:

- ☐ Casa
- ☐ Apartamento
- ☐ Fazenda
- ☐ Alojamento
- ☐ Estudantil
- Outros \_\_\_\_\_

**Figura 3.1:** Pesquisa “Conhecendo um pouco mais de você”.

Note que a estrutura de uma pergunta de múltipla escolha é amigável e já conhecida. Ainda assim, apesar de óbvia, podemos ter dúvidas. Vejamos os casos exemplificados na **Figura 3.1**

A pergunta 1 é bastante clara e direta. Após marcar a sua opção, o leitor terá fornecido sua avaliação do sorvete apreciado. Feito isto, a aná-

lise dos dados será simplificada pelo indivíduo responsável pelo estudo, pois este já estruturou as respostas em opções que lhe satisfaçam.

A segunda pergunta, baseado no que interessa ao pesquisador, também estratifica as faixas salariais de tal forma que ele possa dividir os entrevistados em classes. Contudo, existem detalhes nesta questão que podem ser melhorados. O primeiro é mudar a referência de salário-mínimo para valores. Nem todo entrevistado consegue estimar o valor do salário-mínimo para responder, fato que talvez obrigue a pessoa a estipular a quantidade de salários-mínimos que recebe e a coloque em uma faixa salarial. O outro detalhe seria aumentar as opções para melhor distribuir os entrevistados em camadas mais distintas.

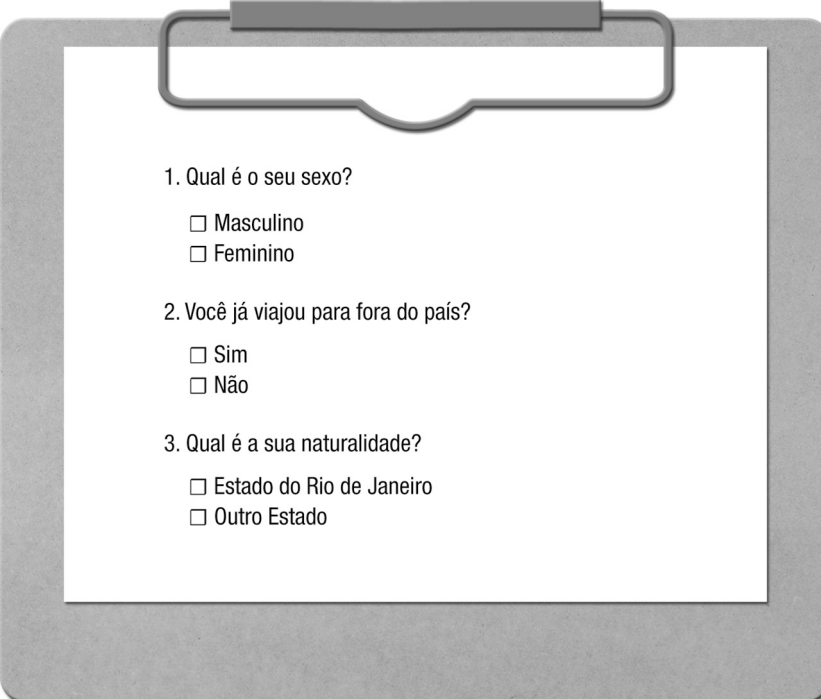
A terceira pergunta é objetiva, mas ao mesmo tempo confusa. Nela, não está especificado se o entrevistado deve marcar apenas a opção que mais lhe atrai ou se pode marcar mais de uma. Esta pergunta será traumática para quem analisa os dados, pois, potencialmente, muitos entrevistados marcarão diversas opções. E, de acordo com o objetivo, eles deverão ser desconsiderados. Para este caso faz-se necessário que esteja claro “selecione apenas a opção que mais lhe interessa” ou a chance de escolher as que de fato agradam ao entrevistado. Uma terceira possibilidade é pedir que o entrevistado enumere de 1 a 6 a preferência, deixando claro que 1 será a opção que mais gosta e 6 a que menos tem interesse ou o contrário – de acordo com a escolha do entrevistador.

Repare que nas perguntas anteriores não se faz necessário avisar que é para marcar apenas uma opção, pois dificilmente uma pessoa recebe um salário que esteja entre duas faixas distintas ou, tampouco, alguém considere um sorvete bom e muito ruim ao mesmo tempo.

A quarta questão é a mais interessante. Primeiro porque é bastante direta. Por mais que uma pessoa possua mais de uma residência (tenha outras para viagens, por exemplo), ela deixa bem claro que se trata da residência oficial. Outro fator interessante foi desmembrar fazenda e alojamento estudantil. Isto se deve pelo interesse neste tipo de resposta. E, por fim, a opção *outros* com espaço para preenchimento. Nem sempre temos todas as opções possíveis, o que pode limitar a resposta do entrevistado. Eu, particularmente, tenho um amigo que mora nos Estados Unidos em um barco – por motivos financeiros e pela praticidade de burlar a fiscalização de imigrantes. Ainda assim, esta opção também ajudaria, por exemplo, uma pessoa que mora em um barraco e ficaria em dúvida em dizer se seria uma casa.

Essa opção *outros* ajuda a estruturar melhor a pergunta em outras oportunidades. A pessoa responsável, pela entrevista, deverá reunir todas as respostas *outros* e verificar se alguma se encaixa nas opções já oferecidas. Caso negativo, irá estudá-las para verificar se devem permanecer como *outros* apenas ou se, pelo menos uma parte delas, deve formar uma nova opção.

As perguntas de múltipla escolha que possuem apenas duas opções de respostas são conhecidas como perguntas **dicotômicas**. Uma vantagem nestas perguntas é limitar ao máximo as opções dos entrevistados direcionando os dados do estudo. A **Figura 3.2** possui alguns exemplos:



1. Qual é o seu sexo?

☐ Masculino

☐ Feminino

2. Você já viajou para fora do país?

☐ Sim

☐ Não

3. Qual é a sua naturalidade?

☐ Estado do Rio de Janeiro

☐ Outro Estado

**Figura 3.2:** Pesquisa com perguntas dicotômicas.

Note que um dos exemplos mais comuns de perguntas dicotômicas é o clássico binômio sim/não. Contudo, outras opções são bastante comuns. Ainda assim é importante se atentar que na tentativa de restringir as opções de respostas, com uma pergunta dicotômica, você pode obter um resultado muito superficial ou insuficientemente claro.

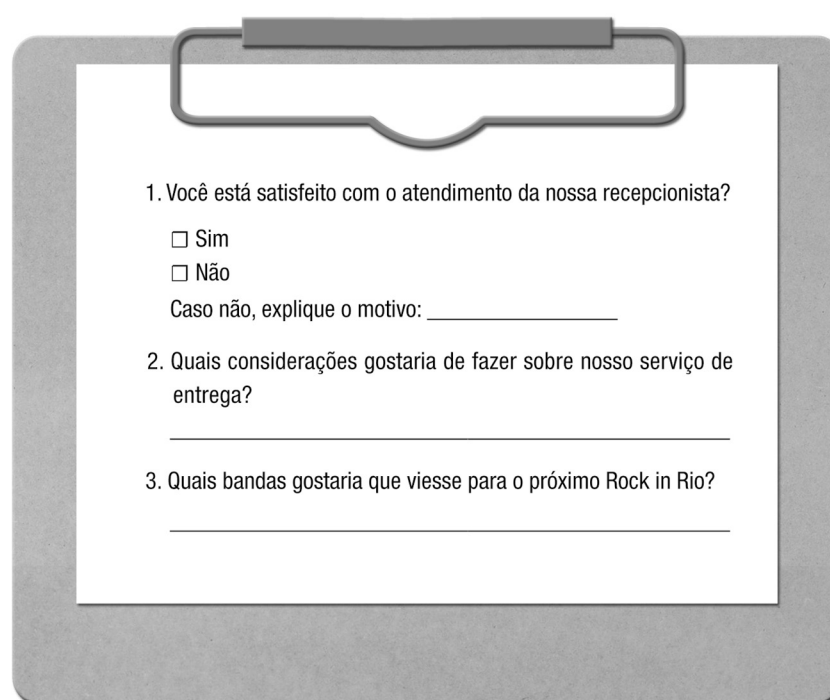
Vejamos o exemplo da pergunta 3 da **Figura 3.2**. Nela, só será possível medir quantos entrevistados nasceram no Estado do Rio de Janeiro

## Dicotomia

Classificação em que se divide cada coisa ou cada proposição em duas, subdividindo-se cada uma destas em outras duas; e, assim, sucessivamente. Divisão em dois ramos. Divisão de um gênero em duas espécies que absorvem o total.

e quantos nasceram fora dele. Logo, se, por algum acaso, for necessário medir a quantidade de paulistas entrevistados, isto não será possível. Neste sentido, indo mais além, sequer será possível determinar se algum estrangeiro foi entrevistado.

Por sua vez, outro formato é a *pergunta aberta*. Nesta o entrevistado tem total liberdade de escrever sobre o que lhe foi questionado. Para ele, isto será confortável e prático. Contudo, para quem irá lidar com os dados, pode-se tornar um processo exaustivo e inconclusivo. A **Figura 3.3** ilustra alguns exemplos!



1. Você está satisfeito com o atendimento da nossa recepcionista?

☐ Sim

☐ Não

Caso não, explique o motivo: \_\_\_\_\_

2. Quais considerações gostaria de fazer sobre nosso serviço de entrega?

\_\_\_\_\_

3. Quais bandas gostaria que viesse para o próximo Rock in Rio?

\_\_\_\_\_

**Figura 3.3:** Pesquisa com perguntas abertas.

Note que a primeira pergunta da **Figura 3.3** era originalmente dicotômica, mas houve a necessidade de “abrir” a opção para o entrevistado relatar sobre o motivo que o desagradou. Aqui, possivelmente o responsável pelo questionário está esperando algo relacionado com a dicção, simpatia ou prestatividade da recepcionista. Contudo é possível que tenhamos respostas envolvendo os trajés, a beleza da pessoa ou coisas inesperadas como achar que o nome era feio.

Essa abertura na pergunta permite que o responsável tenha um cenário em mãos sobre o assunto, organizando as respostas, e, caso seja

necessário no futuro, poderá dividi-lo em várias perguntas sobre a satisfação do cliente em relação à dicção, à simpatia, à prestatividade, entre outras características da recepcionista.

A segunda pergunta e a terceira são abertas ao extremo. Nelas as opções de respostas são infinitas, dando ao organizador da pesquisa um trabalho árduo de reunir as informações e transformá-las em um cenário. Ainda assim, após perguntas como estas é possível fazer uma nova pesquisa com questionamentos em múltipla escolha com opções mais bem direcionadas. Isto é: no caso da terceira pergunta, após coletadas e agrupadas as respostas por bandas citadas, é possível realizar uma nova pesquisa com as 5 ou 10 mais escolhidas, reduzindo as opções de aceitação.

## ===== **Atividade 2** =====

### *Atende aos objetivos 2 e 3*

Com base nos objetivos a seguir, determine se as perguntas a serem feitas devem ser de múltipla escolha, dicotômicas ou abertas. Considere a melhor possibilidade que atinja a necessidade. Justifique.

a) Medir qual dos quatro principais times de futebol do Rio de Janeiro possui a maior torcida.

---

---

---

b) Estipular uma nota que avalie o desfile de uma determinada escola de samba.

---

---

---

c) Constatar a quantidade de pessoas que possui nível superior em uma empresa.

---

---

---

d) Montar uma pré-seleção para determinar os finalistas em uma premiação de melhor filme.

---

---

---

### **Resposta comentada**

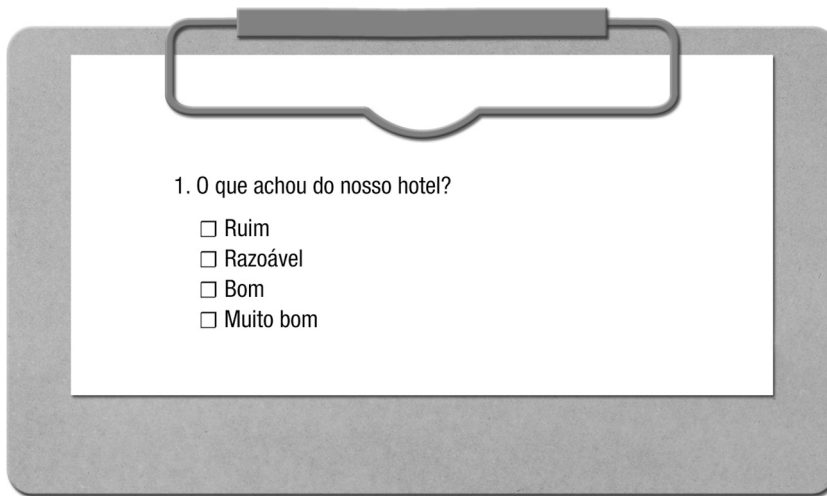
- a) Múltipla escolha é a melhor opção. Uma alternativa para cada um dos quatro times. Após encerrada a pesquisa, a contabilização ficará mais ágil.
  - b) Permitir uma pergunta aberta vai dar margem aos entrevistados estipularem qualquer tipo de nota – desde números que podem variar até números com casas decimais. Para facilitar, o ideal é múltipla escolha com as notas pré-determinadas. Isto facilitará a contabilização final e induzirá os entrevistados a optarem entre as faixas que interessem ao pesquisador.
  - c) Essa definitivamente precisa ser dicotômica. O entrevistado vai responder apenas “sim” se possuir nível superior e “não” caso contrário. Qualquer margem diferente desta não terá validade para o pesquisador.
  - d) Para caso como esse faz-se necessária uma pergunta aberta. Cada entrevistado escreverá o filme que gostaria que estivesse na final. Depois o pesquisador reunirá todas as respostas, arrumará as que foram escritas erradas e contabilizará as mais votadas para compor, enfim, a pesquisa de quem deverá ganhar. Esta, sim, na final, deverá ser e múltipla escolha apenas com os finalistas.
- 
- 
- 

### **Possíveis problemas**

Mesmo conhecendo as formas de se obter dados e as opções de perguntas para elaborar um formulário a ser utilizado em uma pesquisa, alguns problemas podem surgir eventualmente. A maior parte deles surge no momento em que o questionado irá responder. Em alguns, o próprio questionado nota que a pergunta está confusa e acaba respondendo como “acha que eles queriam saber”. Outros sequer são notados por quem responde, nem por quem elaborou, produzindo respostas que não deveriam ser aceitas. Isto é, a pergunta fica dúbia, mas quem vai responder entende apenas uma das interpretações dela (a errada) e a responde com tanta convicção que sequer cogita uma possível segunda interpretação. Por exemplo, um restaurante que optou em colocar música ao vivo questionou aos seus clientes em um formulário: “A música lhe agradou?” O cliente entendeu apenas que a ideia da pergunta era questionar se a música ao vivo incomodou ou não o jantar dele e, como não afetou o seu programa, optou por responder *sim*. Contudo, o dono

do restaurante queria saber se o estilo musical (samba, por exemplo) agradou à clientela. Neste caso, se o mesmo cliente tivesse entendido ou cogitado essa segunda abordagem da pergunta, teria respondido *não*, pois não gosta nem um pouco de samba.

Ressalto que perguntas demasiadamente subjetivas devem ser evitadas. O foco da pergunta precisa ser claro e objetivo para que quem for responder saiba de fato do que se trata. Suponhamos que seja feita uma pesquisa de satisfação com clientes de um hotel.



1. O que achou do nosso hotel?

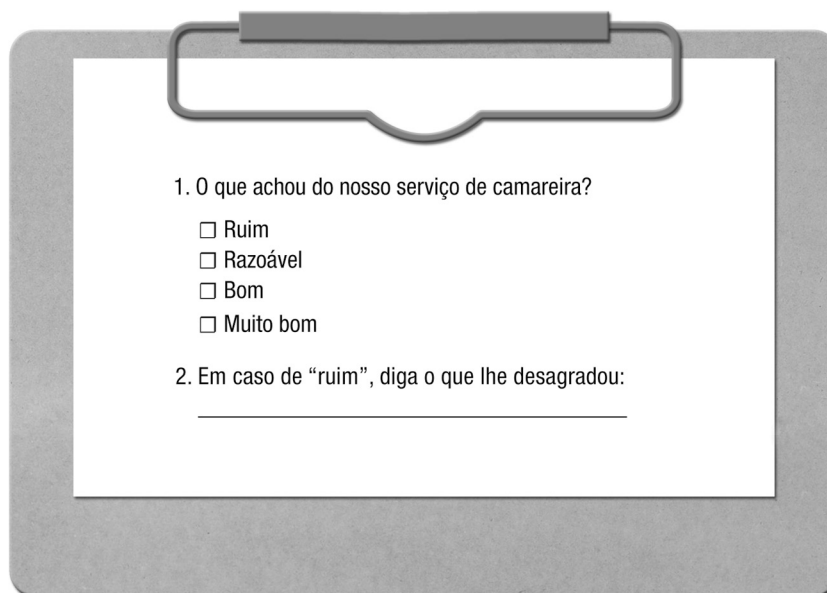
- ☐ Ruim
- ☐ Razoável
- ☐ Bom
- ☐ Muito bom

**Figura 3.4:** Pesquisa com problema de significância para o pesquisador.

Poucos clientes terão dificuldades em responder a essa pergunta. Contudo, poucas respostas terão uma significância para o pesquisador, uma vez que a pergunta é tão subjetiva que um cliente pode avaliar apenas a recepção, outro o quarto e um terceiro apenas a limpeza do banheiro. Logo, sendo o objetivo do pesquisador levantar dados pontuais como recepção, quarto ou banheiro, o ideal é que a pergunta “amarre” estes quesitos para o cliente avaliar. Ainda assim é possível que alguns clientes entendam que a pergunta deseja avaliar o hotel em um todo, incluindo atendimento, comida e outros elementos. Mesmo assim, como ter uma noção precisa de como foi avaliada a resposta?

Deste modo, imagine um cliente que gostou das instalações, achou a comida agradável, mas em um dia específico se aborreceu com o porteiro. Agora considere que apenas por este incidente ele optou por marcar a opção “Ruim” apenas para avisar ao gerente que algo de errado aconteceu. Ora, como o pesquisador vai saber o que provocou aquela ava-

liação “ruim”? Como diferenciar este cliente de outro que de fato achou tudo muito ruim e marcou a mesma opção? Para isto enfatizamos que as perguntas precisam ser bem direcionadas. O ideal é perguntar sobre os quartos, a recepção, o atendimento, a comida etc. e fornecer uma opção de preencher o que lhe desagradou em caso de avaliações negativas.

A imagem mostra um formulário de pesquisa desenhado para parecer uma folha de papel presa a um bloco de notas cinza. O formulário contém duas perguntas numeradas. A primeira pergunta tem quatro opções de resposta com caixas de seleção. A segunda pergunta é aberta e inclui uma linha para o cliente escrever.

1. O que achou do nosso serviço de camareira?

- ☐ Ruim
- ☐ Razoável
- ☐ Bom
- ☐ Muito bom

2. Em caso de “ruim”, diga o que lhe desagradou:

\_\_\_\_\_

**Figura 3.5:** Pesquisa com significância positiva para o pesquisador.

Neste momento temos uma pergunta totalmente direcionada a um ponto específico do seu negócio. Como elemento agregador, inserimos a opção do cliente dizer o que motivou a avaliar como “ruim” o serviço em questão. Contudo, algo pode ser melhorado. Note que temos duas opções para avaliações satisfatórias (bom e muito bom) e apenas uma para o oposto (ruim). O ideal é que sempre possamos dividir a avaliação em partes iguais. Colocar uma opção mediana (Razoável) para separar as boas das ruins é um bom meio de dar uma opção para o cliente indiferente. Outra sugestão é que forneça quantidades iguais de opções boas e ruins. Se possível, uma de cada, pois diferenciar algo ruim de muito ruim não é uma tarefa que qualquer um consiga fazer com habilidade. E, convenhamos, na maior parte das vezes, tudo o que queremos é saber se o cliente está satisfeito, indiferente ou insatisfeito. Quanto mais detalhes, mais dados e menos foco.

A maneira como a pergunta é elaborada também pode causar confusão. Para tal, é sempre recomendável que evite termos técnicos, linguagem muito sofisticada ou qualquer artifício que possa complicar a compreensão do que está sendo indagado. Em certos momentos existem algumas pessoas que implicam com coisas bobas, mas que não deixam de fazer algum sentido. Ao perguntar “Qual sua avaliação da nossa comida?”, sempre vai aparecer um metido a interpretação que vai dizer “Minha avaliação é muito boa, mas a comida é ruim!” Sim! A pergunta pode ser entendida em como é a maneira de avaliar a comida do entrevistado, mas todos entenderam que, na realidade, o que estava sendo perguntando era sobre a própria comida.

A ordem das perguntas pode comprometer também o resultado. Vejamos que ao perguntar primeiro sobre a satisfação acerca de algo, um cliente insatisfeito pode responder as demais perguntas com a ideia de insatisfação em mente. Assim, perguntas sobre outros fatores podem receber uma avaliação mais baixa por influência desta recordação.

1. Você está satisfeito com o seu salário?

☐ Sim  
☐ Não

2. Você considera que a empresa lhe paga um salário dentro da média do mercado para o seu cargo?

☐ Sim  
☐ Não

**Figura 3.6:** Influência da ordem das perguntas no resultado final da pesquisa.

Ao responder a primeira pergunta da **Figura 3.6**, o entrevistado pode fornecer uma resposta tendenciosa na segunda pergunta – imagine que o indivíduo não está satisfeito com o salário dele por motivos de egoísmo próprio! Isto é: na cabeça dele, a insatisfação se dá porque gos-

taria de ganhar o triplo ou um salário astronômico. Aqui pouco importa se o salário dele está adequado com o mercado ou não. Ele responderá conforme sua sensibilidade pessoal e vontade de estar ganhando muito mais, sendo isto justo ou não. Ao passar para a segunda pergunta, com a ideia em mente de que gostaria de ganhar mais, o entrevistado responderá que a empresa não paga um salário compatível com o mercado.

Deste modo, para que as respostas sejam autênticas, o ideal é que primeiro se questione a percepção do entrevistado sobre o salário e o mercado. Mesmo considerando que poucos, de fato, são capazes de avaliar isto. Assim ele tende a ser mais justo na sua resposta. Depois, com a perspectiva de mercado respondida, ele poderá avaliar sua satisfação com o salário em si.

## Conclusão

Já entendemos como funciona na teoria o processo estatístico, ratificamos a importância dos dados e, nesta aula, também pudemos notar que a fonte de coleta dos dados é tão importante quanto complexa de se fazer. Elaborar perguntas é fácil. A dificuldade é elaborar perguntas as quais as respostas te atendam de forma efetiva. Isto é: não elaboramos perguntas, pois estamos nos preparando para receber as respostas. Para isto, como costume dizer informalmente, é necessário elaborar um formulário de perguntas à prova de “idiotas”.

Em vista disso, vimos nos exemplos anteriores que perguntas simples podem gerar dúvidas ou até mesmo respostas caóticas. Ainda assim, existe uma gama de opções que induziriam a erros na pesquisa e é praticamente impossível prever todos eles. Contudo, com a experiência e o bom senso é possível prever potenciais erros. Somado isto ao questionário piloto, reduzimos bastante as chances de algo dar errado.

## Atividade final

### Atende aos objetivos 2 e 3

Elabore um questionário de pesquisa com os três tópicos abaixo. Justifique.

a) Determinar quais as principais fontes que as pessoas recorrem para obter informações sobre a cidade ou país que irão visitar.

---

---

---

---

---

---

---

---

b) Agrupar os possíveis tipos de destinos e determinar a preferência de cada entrevistado.

---

---

---

---

---

---

---

---

c) Criar algumas categorias nas quais o turista gasta e descobrir em qual ele prefere gastar mais.

---

---

---

---

---

---

---

---

### **Resposta comentada**

a) Para o primeiro tópico a sugestão de uma pergunta de múltipla escolha é a mais adequada. Nela, podemos colocar opções como revistas, internet, dicas de amigos, entre outras. O principal é não criar redundância que pode confundir o entrevistado. Por exemplo: ter a opção revistas, *internet* e revistas eletrônicas. Revistas eletrônicas estão categorizadas como *internet*.

- b) No segundo tópico, o principal é ter bem determinado quais tipos de destinos você pretende estudar. Podemos criar tipos como: a Praia, a Serra, o Ecológico, o Histórico/Cultural etc. Mas é importante que na pergunta esteja bem claro o questionamento sobre qual é a característica principal do destino que o fez escolher viajar. Vejamos: no Rio de Janeiro é possível ser praia, ecológico e histórico/cultural. Assim como em São Paulo pode ser histórico/cultural, gastronômico e balada noturna.
- c) O terceiro tópico é o mais interessante! Todo turista possui uma opção na qual vai concentrar os seus gastos. Alguns optam em não economizar no hotel, pois preferem conforto acima de tudo. Outros economizam em tudo para gastar em compras. Alguns direcionam os gastos para passeios. Para este Tópico, uma pergunta de múltipla escolha com as opções mais comuns irá determinar o perfil dos entrevistados.
- 
- 

## Resumo

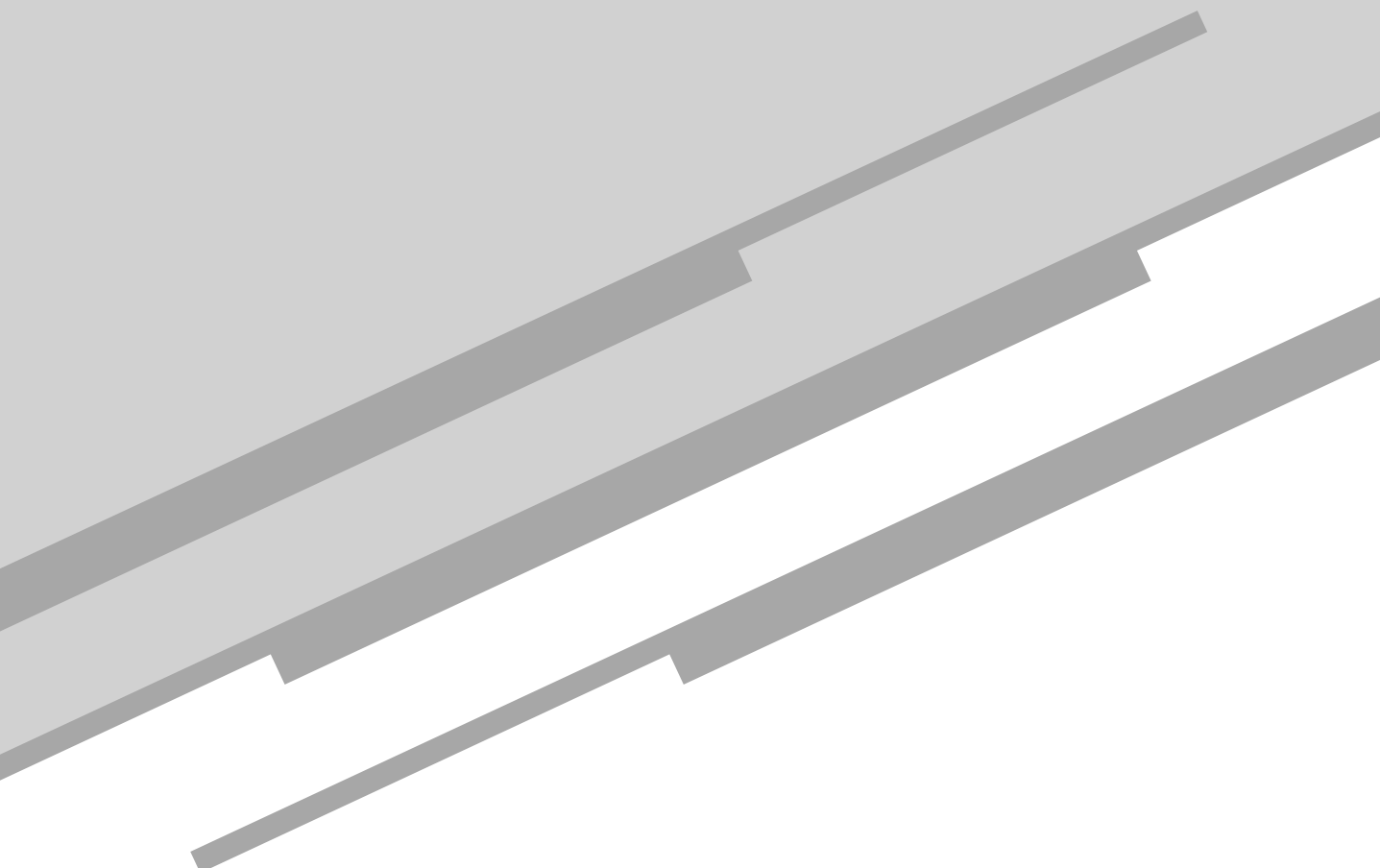
Nesta aula podemos perceber que elaborar *perguntas* não é tão simples quanto imaginávamos. Com o passar do tempo e adquirindo experiência, costuma-se dizer que elaborar *questionários* é uma arte. Existem formatos determinados e tudo precisa ser repensado com antecedência, fora os empecilhos que ainda assim podem acontecer. Foi possível conhecer as principais *fontes de dados* para um pesquisador, como se classificam e como chegar até elas. Conhecemos os principais formatos de questionários e como se aplicam de acordo com a necessidade do entrevistador. Por fim, tentamos levantar algumas eventualidades que podem ocorrer para que, ao fazer a sua primeira pesquisa, já tenha alguma experiência adquirida.

## Informação sobre a próxima aula

Na próxima aula iremos ver as diversas maneiras de organizar os dados. Perceberemos como a necessidade do estudo pode alterar a ordenação dos dados. Por fim, seremos apresentados à Tabela de Frequência – um excelente instrumento para lidar com amostras numéricas grandes.

# Aula 4

Coloca em um *Tapeware* com etiqueta identificando



## **Meta**

Organizar a estruturação de dados de uma pesquisa, para que esta atenda às necessidades da mesma e ao método que será utilizado para sua respectiva análise.

## **Objetivos**

Esperamos que, após o estudo desta aula, você seja capaz de:

1. organizar os dados de uma amostra categórica em uma Tabela de Contingência;
2. classificar os dados de uma amostra numérica;
3. criar tabela de frequência para grandes amostras.

## Introdução

Ao apresentarmos as formas de se organizar os dados de um estudo, para que facilite a análise do mesmo, iremos abordar desde uma simples ordenação até métodos mais completos de organização de amostras em grandes quantidades. Todo este propósito será para que, nas próximas aulas, ao iniciarmos a análise dos dados, estejamos preparados para estruturá-los, de tal forma que o foco fique diretamente apontado para os métodos estatísticos propriamente ditos.

Desse modo, imaginem uma pesquisa com mais de quinhentas pessoas entrevistadas. Agora, visualize a quantidade de dados que uma pesquisa dessas é capaz de gerar. Pois bem, com esta imagem em mente, pense em tudo isso desorganizado sobre uma mesa e você precisando tirar uma conclusão sobre o estudo. Complicado, não acha?

Enfim, é humanamente impossível fazer uma análise correta com dados desorganizados. Assim, por mais que conte com a ajuda de um computador, ele também vai exigir que os dados estejam em ordem para a maioria das funções, pois, caso contrário, vai acusar algum erro ou apontará um resultado distorcido.

A organização dos dados pode seguir critérios dos mais variados. Simplesmente colocar em ordem crescente já é uma maneira de organizar os dados. Desse modo, os maiores e menores valores ficam evidentes. Em ordem alfabética é um recurso interessante quando o foco do estudo são dados categóricos – daí encontrar o nome fica mais fácil. De igual modo, pode-se organizar os dados em ordem cronológica, de acordo com o acontecimento dos eventos, pois, respeitar a sequência de uma série de acontecimentos é importante, dependendo da análise que será feita.

## Dados categóricos

Exatamente por não serem números, os dados categóricos limitam a sua forma de organização. Ainda assim, as poucas opções de organização desses tipos de dados são de grande utilidade para uma melhor análise ou até mesmo para uma consulta a eles.

### Ordem alfabética

Talvez a maneira mais simples e conhecida de se organizar os dados, que é a ordem alfabética, dependendo do objetivo do estudo, será bastante eficiente. Imaginemos uma festa com a lista de convidados na

porta. Por mais que os nomes sejam inseridos nela em ordem desorganizada, pois depende da venda dos ingressos ou do responsável lembrar dos nomes, ao término, com a lista propriamente dita finalizada, é necessário organizar os nomes em ordem alfabética – pois imaginem o martírio que será para a pessoa que recebe os convidados ficar procurando os nomes, de ponta a ponta, cada vez que alguém chega!

Ainda assim, alguém pode alegar que, na lista, existirão convidados comuns e convidados VIPs e misturar os nomes. Neste caso, mesmo em ordem alfabética, algum problema pode ser gerado. Sendo assim, teríamos duas listas: convidados comuns e convidados VIPs. Entretanto, ambas estariam com os nomes organizados em ordem alfabética.

Essa estratégia de dividir a relação de dados em menores listas e depois organizar cada uma, individualmente, em ordem alfabética é uma estratégia bastante válida, pois nem sempre podemos manter todos os elementos em uma única organização. Vejamos o caso de uma pesquisadora de preços de mercado, conforme a Tabela a seguir.

**Tabela 4.1:** Preços de produtos de limpeza e de alimentos

Limpeza		Frios		Frutas		Biscoitos	
Produto	\$	Produto	\$	Produto	\$	Produto	\$
Água sanitária		Mortadela		Banana		Água e sal	
Amaciante		Presunto		Goiaba		Maizena	
Detergente		Queijo bola		Melancia		Recheado	
Esponja		Queijo prato		Morango		Torrada	
Sabão em pó		Salame		Uva		Waffle	

A missão da pesquisadora é ir a cada supermercado de uma região e anotar o preço de uma lista específica de produtos para compará-los depois. Essa lista, se organizada em ordem alfabética em um todo, permitiria encontrar cada produto pelo nome. Contudo, em um mercado, os produtos ficam separados por tipo. Logo, separando por tipo de produto, agilizaria a localização de cada um na lista, ao passar pelo seu corredor específico. Isto é: ao passar pelo corredor de produtos de limpeza, por exemplo, bastaria a pesquisadora atentar apenas aos produtos separados na primeira coluna, ganhando, com isto, mais dinâmica. Obviamente, o fato de a lista exemplificada na **Tabela 4.1** ser pequena, dispensa qualquer tipo de organização. Entretanto, considere que estamos, o tempo todo, falando de amostras grandes.

## Tabela de contingência

O recurso da tabela de contingência é muito útil quando desejamos contabilizar a quantidade de respostas para cada opção do questionário. Sua maior vantagem é a ágil consulta aos resultados sem grandes dificuldades, isto é, toda pesquisa com várias respostas pode ficar resumida em uma única tabela de poucas linhas e colunas.

Desse modo, suponhamos que vamos levantar o consumo de opções “pratos para o jantar” em um restaurante sofisticado. Neste, o serviço é fechado por um preço único, ou seja, o cliente recebe três opções de entradas e três opções de pratos principais, escolhe um de cada e, independentemente da escolha, o preço será o mesmo pela refeição.

Restaurantes que funcionam dessa forma procuram se estruturar antes de abrir, deixando, praticamente, tudo pronto. Para tal, é sempre importante ter uma previsão de quantos pratos de cada tipo, possivelmente, irá servir naquela noite. Portanto, com o intuito de estimar a demanda da clientela, cada noite são anotadas as combinações de cada cliente, tanto de entrada, quanto de prato principal. Entretanto, organizar estas informações exige o recurso da tabela de contingência. Vejamos, na Tabela a seguir, como ficou o resultado da pesquisa.

**Tabela 4.2:** Estimativa de demanda de clientes/Combinação de pedidos

<b>Opções</b>	Filet do bira	Risoto de frango	Fritada de bacalhau	<b>Total</b>
Sopa de siri	97	75	98	270
Salada colorida	50	42	68	160
Carpaccio	41	58	71	170
<b>Total</b>	188	175	237	600

Note que, com uma rápida consulta à **Tabela 4.2**, podemos afirmar que ali foram contabilizados os pedidos de 600 clientes. Ainda em uma consulta rápida, identificaremos a última *coluna* como a que determina o total de pedidos de cada entrada e a última *linha*, como a que totaliza o número de pratos principais pedidos.

Desse modo, pela consulta da última *coluna* é possível concluir que a sopa de siri é a entrada mais pedida com 270 escolhas. De igual modo, pela última *linha*, fritada de bacalhau é o prato principal mais escolhido com 237 pedidos. Em vista disto, ao cruzarmos as *linhas* com *colunas* co-

meçaremos a identificar a combinação dos pedidos. Quando a terceira coluna se encontra com a terceira linha, identificamos que 42 pessoas pediram salada colorida e risoto de frango, assim como concluímos que 41 pessoas pediram carpaccio e filet do Bira, ao cruzarmos a segunda coluna com a quarta linha. Esta tabela de contingência, como pudemos ver, é prática na consulta de variáveis simples, como, por exemplo, tipo de entrada e tipo de prato principal, bem como na consulta de variáveis combinadas na pesquisa, isto é, a escolha combinada de cada cliente.

## ===== **Atividade 1** =====

### *Atende ao objetivo 1*

Uma gerente arquivou as vendas de camisetas regatas para o carnaval, catalogando por tamanho e cor. Para tal, criou um código, no qual a primeira parte é a cor (C de cinza, B de branca e A de azul) e a segunda, o tamanho (P de pequeno, M de médio e G de grande). A organização ficou conforme a Tabela a seguir:

A/G	B/M	C/P	A/P	B/M	B/P	A/M	A/P	C/P	A/G
C/P	B/G	A/M	B/P	C/P	A/P	B/M	B/G	B/P	A/P
B/M	A/P	C/M	B/M	C/M	B/P	B/P	C/M	A/M	B/G
B/P	C/G	C/P	B/M	A/P	A/G	A/P	C/P	A/P	B/M

Organize estas informações em uma tabela de contingência, para que a gerente possa contabilizar as vendas, e explique quais critérios você adotou para a organização.

### Resposta comentada

Devemos, basicamente, contar quantas camisetas regatas foram vendidas de acordo com o tamanho e cor. Ao final, teremos a seguinte Tabela de Contingência:

Camisetas	Peq.	Médio	Grande	Total
Azul	8	3	3	14
Branca	6	7	3	16
Cinza	6	3	1	10
<b>Total</b>	20	13	7	40

### Dados Numéricos

Enquanto, nos dados categóricos, a preocupação estava remetida à informação na forma de palavra, nos dados numéricos, o que importa é o próprio número, o valor e/ou o resultado. Por ser muito plural, a sua organização depende do objetivo de cada pesquisa.

### Ordem crescente ou decrescente

Tanto em uma ordem, como na outra, o resultado imediato desse tipo de estruturação é a rápida visualização dos maiores e menores resultados nos extremos da listagem. Por exemplo: se fôssemos fazer uma relação com todas as contas que pagamos em um determinado mês e organizássemos pela ordem decrescente, no topo, estariam as contas de maior valor, isto é: potencialmente estarão aluguel, prestação do carro, empréstimo do banco etc. Ao estruturar nesta sequência, e você precisando fazer cortes no orçamento, bastará iniciar a busca pelas possíveis contas no topo. Logo, quanto mais no topo estiver o compromisso que puder evitar nos próximos meses, maior será o corte.

Todavia, nem sempre a escolha desse ordenamento é tão simples, pois afinal, em diversos casos, existem mais de uma informação a ser considerada. Daí cabe ao pesquisador, conhecendo o objetivo do estudo, apontar qual das informações é que irá determinar o critério do ordenamento. Suponhamos, então, que seja o gerente de um *buffet* para

festas e eventos diversos. Com o tempo, andou notando “a quebra” de muitos itens do serviço. Para ter certeza disto, resolveu tomar nota de quantos itens do *buffet* eram quebrados em cada evento, formando, ao final de seis meses, a Tabela que segue:

**Tabela 4.3:** Quantidade de itens quebrados x Perda financeira

Itens	Quebrados	Perda
Cinzeiros	77	R\$ 45,30
Copos	49	R\$ 53,50
Jarros	8	R\$ 118,20
Pratos	19	R\$ 66,70
Taças	35	R\$ 58,30

Repare que as informações estão em ordem alfabética, mas precisamos determinar um critério de ordenamento de acordo com as variáveis numéricas. Nesse momento, são duas variáveis em questão: quantidade de itens quebrados e o valor estimado de perda com as quebras. A escolha do critério será preponderante para a leitura inicial da Tabela, pois, de acordo com esta organização, teremos um novo item em destaque no topo.

Suponhamos, então, que você tenha optado por dar atenção às quantidades quebradas. Ao ordenar por esta característica, teremos no topo o cinzeiro, seguido pelos copos, taças, pratos e jarros, necessariamente nesta ordem. Assim, quando iniciarmos a leitura, a primeira coisa que ficará em evidência é que cinzeiro é o item que mais se quebra nos eventos. Contudo, mudando o foco do estudo para a perda financeira proporcionada pela quebra dos itens, nossa planilha mudará totalmente de formato. Cinzeiro, que antes estava no topo, por ser o item que mais se quebra, agora está no final por possuir um menor valor agregado. Da mesma forma com os jarros, cujo índice de quebra é baixo, aparecerá agora no topo por ser um item muito caro.

## Princípio de Pareto ou 80/20

O *Princípio de Pareto* pressupõe que, na maioria dos fenômenos, 80% das consequências são decorrentes de 20% das causas. Isto é: a maior parte dos eventos está concentrada em uma pequena parcela de ações.

Esse Princípio foi inspirado em um dos estudos de **Vilfredo Pareto**, que notou que 80% da renda estava concentrada apenas nas mãos de 20% da população. Daí, esta mesma observação pode ser extrapolada para diversas áreas. Por exemplo: em restaurantes, 80% da receita é originária de 20% das opções do cardápio. Em uma estrada, 80% dos acidentes acontecem em 20% dos pontos de risco desta. Ou simplesmente que você passa 80% da sua vida em apenas 20% da área da sua casa.

Para o nosso caso, esse Princípio será útil no que tange à leitura da planilha. Isto é: ao ordenar os dados de acordo com a necessidade do estudo, potencialmente teremos um 80/20 ou algo bem próximo disto. Vejamos o exemplo na Tabela que segue:

**Tabela 4.4:** Materiais x Quantidade x Custos unitário e Total

Item	Qtde.	Custo unit.	Custo total
Hastes	30	R\$ 28,30	R\$ 849,00
Placas	15	R\$ 45,90	R\$ 688,50
Parafusos	150	R\$ 0,77	R\$ 115,50
Molas	80	R\$ 1,30	R\$ 104,00
Porcas	150	R\$ 0,55	R\$ 82,50

Na **Tabela 4.4**, temos a relação de materiais que serão necessários para a produção de um item em particular. Para cada item, temos a quantidade necessária, o custo de cada unidade e o custo total.

Ao totalizarmos as informações, temos que, para produzir um item, o custo total será de aproximadamente R\$ 1.840,00, dos quais 46% são representados pelo custo das hastes e 37% pelo custo das placas. Nesse momento, temos mais de 80% dos custos concentrados em apenas dois itens. Esses representam, respectivamente, 7% e 3,5% do total de quantidades unitárias para a montagem.



### Vilfredo Pareto

Nascido em 1848, faleceu em 1923. Matemático, economista e sociólogo.

Italiano, seu maior trabalho foi o *Tratado Geral de Sociologia*, mas também fez diversas publicações na área da economia e matemática.

No início do século XX desenvolveu uma relação intelectual com Mussolini sendo nomeado Senador.

Fonte da imagem: [http://pt.wikipedia.org/wiki/Ficheiro:Vilfredo\\_Pareto.jpg](http://pt.wikipedia.org/wiki/Ficheiro:Vilfredo_Pareto.jpg)

Podemos dizer então que, mesmo não sendo precisamente um 80/20, tivemos um Pareto (forma mais informal de se chamar o Princípio) nos custos do produto, pois nem sempre teremos o acerto de 80% e 20%, mas sempre teremos, na maioria dos eventos, a maior parte dos valores associados às menores quantidades de itens.

## Tabela de Frequência

Comumente, trabalharemos com amostras grandes e tentar resumir, pelo menos, estes dados em grupos é sempre uma boa iniciativa para compreendermos melhor como estão se comportando. Para tal, o recurso da Tabela de Frequência é a opção mais comum, pois, como o próprio nome diz, ela determina a frequência dos eventos.

A montagem de uma Tabela de Frequência requer alguns passos. O primeiro é determinar a amplitude da *amostra*. Esta amplitude irá determinar o quão distante o menor dos valores da amostra está do maior desta. Isto é: trata-se apenas da subtração do menor valor, do maior valor.

Posteriormente, precisamos determinar em quantas classes ou faixas iremos dividir nosso estudo. Neste quesito existem algumas opções. A mais comumente utilizada é o próprio critério do pesquisador, isto é, ele irá arbitrar uma quantidade que achar conveniente.

Outra opção é uma regra que determina que, caso tenhamos uma amostra com quantidade inferior a 25 dados, usaremos 5 classes. Caso seja maior que 25 dados, o número de classes será igual à raiz quadrada do número de dados. Por exemplo, tendo 91 dados, a raiz quadrada de 91 é aproximadamente 9,54, nos indicando usar 9 classes.

Já o Método de Kelley estipula a quantidade de classes baseado em uma tabela previamente montada, de acordo com quantidade de dados, conforme a Tabela que segue:

**Tabela 4.5:** Quantidade de dados x Quantidade de classes

Dados	5	10	25	50	100	200	500
Classes	2	4	6	8	10	12	15

Desse modo, se estamos trabalhando com uma amostra de 100 dados, teremos uma Tabela de Frequência com 10 classes ou faixas.

Nessa sequência, outro método conhecido é a Regra de Sturges. Por este, a quantidade  $K$  de classes é determinada pela fórmula a seguir. Segundo esta, o cálculo é resultado do logaritmo Neperiano do número de dados ( $N$ ), vezes 3,3 mais 1.

$$K = 1 + 3,3 \ln N$$

Desse modo, com a quantidade de dados que pode variar nesta fórmula é possível prever os números de classes – conforme a próxima tabela:

**Tabela 4.6:** Quantidade de dados x Previsão de número de classes

Dados	Classes	Dados	Classes
1	1	De 47 até 90	7
2	2	De 91 até 181	8
De 3 até 5	3	De 182 até 362	9
De 6 até 11	4	De 363 até 724	10
De 12 até 22	5	De 725 até 1448	11
De 23 até 46	6	De 1449 até 2896	12

Agora que já temos a quantidade de classes, podemos dar prosseguimento à construção da tabela de frequência. Contudo, convém apresentarmos um exemplo para que fique mais fácil o entendimento das próximas etapas. Suponhamos, então, que foi feito um estudo com o tempo em minutos das ligações feitas de uma residência, durante o mês de abril. Na Tabela que segue temos todas as marcações:

**Tabela 4.7:** Tempo, em minutos, de ligações telefônicas realizadas em abril – dados ordenados de forma aleatória

1,35	2,78	5,16	4,77	2,34	1,28	3,88	4,12	3,44	5,26
4,28	3,61	2,73	3,15	1,18	4,7	3,91	2,22	1,68	4,88
2,28	1,08	5,12	4,18	3,48	2,77	3,12	4,81	5,61	4,12
5,39	3,48	2,18	2,67	3,79	5,22	4,18	1,91	2,88	3,08
4,91	1,99	2,18	3,64	5,18	4,81	5,68	4,28	3,94	4,56

Inicialmente, para qualquer processo de construção de tabela de frequência, precisamos organizar os dados. Como já falamos antes, uma

das melhores maneiras é colocá-los em ordem crescente para podermos, mais facilmente, identificar o menor e o maior e, depois, selecioná-los de acordo com as classes. Então, organizando os dados temos:

**Tabela 4.8:** Tempo, em minutos, de ligações telefônicas realizadas em abri – dados ordenados de forma crescente

1,08	1,18	1,28	1,35	1,68	1,91	1,99	2,18	2,18	2,22
2,28	2,34	2,67	2,73	2,77	2,78	2,88	3,08	3,12	3,15
3,44	3,48	3,48	3,61	3,64	3,79	3,88	3,91	3,94	4,12
4,12	4,18	4,18	4,28	4,28	4,56	4,7	4,77	4,81	4,81
4,88	4,91	5,12	5,16	5,18	5,22	5,26	5,39	5,61	5,68

Com as informações ordenadas, já podemos afirmar que a ligação mais rápida durou 1,08 minutos, enquanto a mais demorada foi de 5,68 minutos. Logo, a **amplitude** dessa amostra é de 4,60 minutos.

## Amplitude

Amplitude é o “tamanho” real da amostra, isto é, a “distância”, calculada dentro da unidade de medida da amostra, entre o menor e o maior valor desta – ou simplesmente a diferença entre os extremos da amostra

Estamos falando de uma amostra com 50 dados e podemos determinar a quantidade de classes, baseados em qualquer um dos métodos apresentados. Pelo método da raiz quadrada, como são mais de 25 dados, a quantidade sugerida será a raiz quadrada de 50, que arredondando para baixo fica em 7 classes. Já no Método de Kelley, consultando a **Tabela 4.5**, a quantidade sugerida é de 8. Pela Regra de Sturges, consultando a **Tabela 4.6**, a quantidade de classes será de 7. Por fim, pelo critério próprio do pesquisador podemos induzir 7 classes.

Com a quantidade de classes definidas, iremos determinar a *amplitude do intervalo*, isto é, qual será o critério para a contagem de dados em cada classe. O cálculo da amplitude do intervalo é resultado da divisão da amplitude pela quantidade de classes. Portanto, na realidade, tudo que faremos agora é dividir a amplitude da amostra em pedaços iguais. Para este caso específico: 7 pedaços iguais. Portanto, com uma amplitude de 4,60 minutos e 7 classes, temos uma amplitude de intervalo de 0,6571, que iremos arredondar para nossa conveniência para 0,66 (o arredondamento da amplitude do intervalo, quando necessário, sempre se faz “para cima”). Sendo assim, as faixas serão compostas da seguinte maneira:

**Tabela 4.9:** Divisão da amplitude x Quantidade de classes

Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6	Classe 7
De 1,08 até 1,73	De 1,74 até 2,39	De 2,4 até 3,05	De 3,06 até 3,71	De 3,72 até 4,37	De 4,38 até 5,03	De 5,04 até 5,7

Basicamente, o que fizemos foi pegar o menor valor (ligação mais rápida) e somar uma amplitude do intervalo chegando em 1,74. Logo, a primeira classe começa com a ligação mais rápida e vai até 1,73, que é o tempo imediatamente menor ao 1,74. Assim, a segunda classe vai começar em 1,74 e terminar antes de 2,4 (2,4 é resultado de 1,74 mais um intervalo de classe). Assim sucessivamente.

Agora que temos as classes determinadas com seus respectivos intervalos, podemos determinar quantos dados das amostras estão compreendidos em cada uma. A Tabela que segue ilustra essa contagem:

**Tabela 4.10:** Classes x Frequência

Classes	Frequência
De 1,08 até 1,73	5
De 1,74 até 2,39	7
De 2,4 até 3,05	5
De 3,06 até 3,71	8
De 3,72 até 4,37	10
De 4,38 até 5,03	7
De 5,04 até 5,7	8
Total	50

A leitura que esta Tabela faz, basicamente, é que existem, na amostra, 5 ligações que duraram entre 1,08 e 1,73 minutos, 7 ligações que duraram entre 1,74 e 2,39 minutos e assim por diante. Esta contagem de quantos

dados estão compreendidos em cada classe se chama *frequência*. Com a contagem individual por classe, podemos iniciar o processo de acumular a contagem, conforme vamos passando de classes. Vejamos a próxima tabela:

**Tabela 4.11:** Classes x Frequência x Frequência acumulada

Classes	Frequência	Freq. acumulada
De 1,08 até 1,73	5	5
De 1,74 até 2,39	7	12
De 2,4 até 3,05	5	17
De 3,06 até 3,71	8	25
De 3,72 até 4,37	10	35
De 4,38 até 5,03	7	42
De 5,04 até 5,7	8	50
Total	50	50

O que vemos agora é que, conforme passamos a contagem da primeira para a segunda classe, passamos a falar de 12 ligações de um total de 50. Posteriormente, adentrando à terceira classe, já estamos falando de 17 ligações do total de 50 e, assim, sucessivamente. Esta contagem acumulativa de dados, conforme vamos passando pelas classes chama-se *frequência relativa*.

Desse modo, após ser feita a contagem de ligações em cada classe, o próximo passo é determinar o percentual que aquela quantidade representa dentro da amostra. Vejamos a próxima Tabela:

**Tabela 4.12:** Classes x Frequência x Freq. acumulada x Freq. relativa

Classes	Frequência	Freq. acumulada	Freq. Relativa
De 1,08 até 1,73	5	5	10,00%
De 1,74 até 2,39	7	12	14,00%
De 2,4 até 3,05	5	17	10,00%
De 3,06 até 3,71	8	25	16,00%
De 3,72 até 4,37	10	35	20,00%
De 4,38 até 5,03	7	42	14,00%
De 5,04 até 5,7	8	50	16,00%
Total	50	50	100,00%

Calcularemos agora que as 8 ligações compreendidas nas 4 Classes representam 16% do total de 50 ligações da amostra. Também podemos citar que a quinta classe possui a maior frequência, isto é, maior quantidade de dados compreendidos nela (10 ligações), logo, por consequência óbvia, maior frequência relativa: 20% do total de ligações.

Por fim, assim como fizemos com a frequência, iremos acumular as Frequências Relativas também – conforme a próxima Tabela:

**Tabela 13.4:** Classes x Frequência x Frequência acumulada x Frequência relativa x Frequência Relativa acumulada

Classes	Frequência	Freq. acumulada	Freq. relativa	Freq. rel. acum.
De 1,08 até 1,73	5	5	10,00%	10,00%
De 1,74 até 2,39	7	12	14,00%	24,00%
De 2,4 até 3,05	5	17	10,00%	34,00%
De 3,06 até 3,71	8	25	16,00%	50,00%
De 3,72 até 4,37	10	35	20,00%	70,00%
De 4,38 até 5,03	7	42	14,00%	84,00%
De 5,04 até 5,7	8	50	16,00%	100,00%
Total	50	50	100,00%	100,00%

A leitura da contagem acumulativa das frequências relativas é idêntica à da frequência acumulada, com o diferencial de estarmos falando de percentuais. Esta levará o nome de *frequência relativa acumulada*.

Enfim, temos a tabela de frequência das ligações completas. Ante, o que era um grupo de 50 tempos de ligações espalhados, agora são dados organizados, prontos para serem lidos e interpretados. As leituras podem ser feitas das mais diversas maneiras, de acordo com a conveniência do estudo. Uma das leituras, por exemplo, é identificar que é muito parecida a frequência em cada classe, logo, não existe um tempo predominante. Melhor dizendo: não podemos dizer que há maioria absoluta de ligações de tempo curto, médio ou longo.

## Atividade 2

### Atende aos objetivos 2 e 3

A tabela a seguir é resultado de uma pesquisa, contabilizando quantos amigos cada usuário possui em sua conta no *Facebook*. Organize estes dados, monte uma tabela de frequência e explique quais critérios você adotou para a organização.

722	735	1.183	797	984	3.769	1.515	1.367	692
886	1.146	369	825	791	792	1.368	930	831
849	901	977	1.559	808	612	785	870	766

### Resposta comentada

Ordenando os dados:

369	612	692	722	735	766	785	791	792
797	808	825	831	849	870	886	901	930
977	984	1.146	1.183	1.367	1.368	1.515	1.559	3.769

Determinando o número de classes: pelo Método da Raiz Quadrada temos aproximadamente 5,2; por Kelley, são 7; a Regra de Sturges, sugere 6. Por opção própria, utilizarei 6. A amplitude da amostra é 3.400, logo a amplitude do intervalo será 567 (já arredondando o resultado). Com isto, a tabela de frequência ficará assim:

Intervalos		Frequência		Freq. Rel.	
De	Até		Acum.		Acum.
369	936	18	18	66,67%	66,67%
936	1.503	6	24	22,22%	88,89%
1.503	2.070	2	26	7,41%	96,30%
2.070	2.637	0	26	0,00%	96,30%
2.637	3.204	0	26	0,00%	96,30%
3.204	3.771	1	27	3,70%	100,00%

## Conclusão

O processo estatístico é composto por várias etapas, todas elas com sua importância. Nesta aula, vimos que a organização dos dados, dentro do processo, é algo imprescindível para que possamos fazer uma melhor leitura da amostra em questão. Contudo, ficou claro que é importante, desde o início, saber quais os objetivos dos estudos, pois, de acordo com eles, a melhor maneira de organizar os dados pode ser alterada consideravelmente. Ter em mãos os dados organizados implica, diretamente, em conseguir enxergar com mais facilidade padrões, características predominantes da amostra etc. Assim, valores extremos ficam em destaque, repetições também. Dessa forma, fica mais fácil tirar conclusões.

## Atividade final

### Atende aos objetivos 2 e 3

Foi feita uma pesquisa com o valor de gorjeta que alguns hóspedes dão quando recebem novas toalhas nos quartos em que estão hospedados. A Tabela que segue resume esta pesquisa:

R\$ 3,14	R\$ 0,61	R\$ 0,79	R\$ 1,22	R\$ 2,42	R\$ 0,30	R\$ 0,18	R\$ 2,99	R\$ 1,93	R\$ 2,48
R\$ 2,48	R\$ 2,48	R\$ 2,48	R\$ 2,48	R\$ 1,20	R\$ 0,35	R\$ 1,00	R\$ 1,20	R\$ 3,45	R\$ 3,45
R\$ 1,80	R\$ 2,78	R\$ 2,78	R\$ 2,14	R\$ 0,19	R\$ 0,19	R\$ 0,19	R\$ 0,19	R\$ 0,30	R\$ 1,30
R\$ 0,45	R\$ 1,46	R\$ 1,46	R\$ 1,46	R\$ 1,46	R\$ 3,40	R\$ 1,13	R\$ 2,06	R\$ 2,71	R\$ 1,61

Desse modo, crie a tabela de frequência para organizar estes dados e explique quais critérios você adotou para a organização.

### **Resposta Comentada**

Ordenando os dados temos:

R\$ 0,18	R\$ 0,19	R\$ 0,19	R\$ 0,19	R\$ 0,19	R\$ 0,30	R\$ 0,30	R\$ 0,35	R\$ 0,45	R\$ 0,61
R\$ 0,79	R\$ 1,00	R\$ 1,13	R\$ 1,20	R\$ 1,20	R\$ 1,22	R\$ 1,30	R\$ 1,46	R\$ 1,46	R\$ 1,46
R\$ 1,46	R\$ 1,61	R\$ 1,80	R\$ 1,93	R\$ 2,06	R\$ 2,14	R\$ 2,42	R\$ 2,48	R\$ 2,48	R\$ 2,48
R\$ 2,48	R\$ 2,48	R\$ 2,71	R\$ 2,78	R\$ 2,78	R\$ 2,99	R\$ 3,14	R\$ 3,40	R\$ 3,45	R\$ 3,45

Determinando a quantidade de classes pelo Método da Raiz Quadrada teríamos 6 (já arredondando para baixo); por Kelley, ficam sugeridas 7 classes; por Sturges, seriam 6 classes. Optaremos por seis classes.

A amplitude total da amostra é de R\$ 3,27. Com isto, escolhendo 6 classes, teremos uma amplitude de intervalo de R\$ 0,55 (já arredondado). Logo, a Tabela de Frequência ficará assim:

Classes		Frequência		Freq. Relativa	
De	Até		Acum.		Acum.
R\$ 0,18	R\$ 0,72	10	10	25,00%	25,00%
R\$ 0,73	R\$ 1,27	6	16	15,00%	40,00%
R\$ 1,28	R\$ 1,82	7	23	17,50%	57,50%
R\$ 1,83	R\$ 2,37	3	26	7,50%	65,00%
R\$ 2,38	R\$ 2,92	9	35	22,50%	87,50%
R\$ 2,93	R\$ 3,47	5	40	12,50%	100,00%
Classes		40	40	–	100,00%

## Resumo

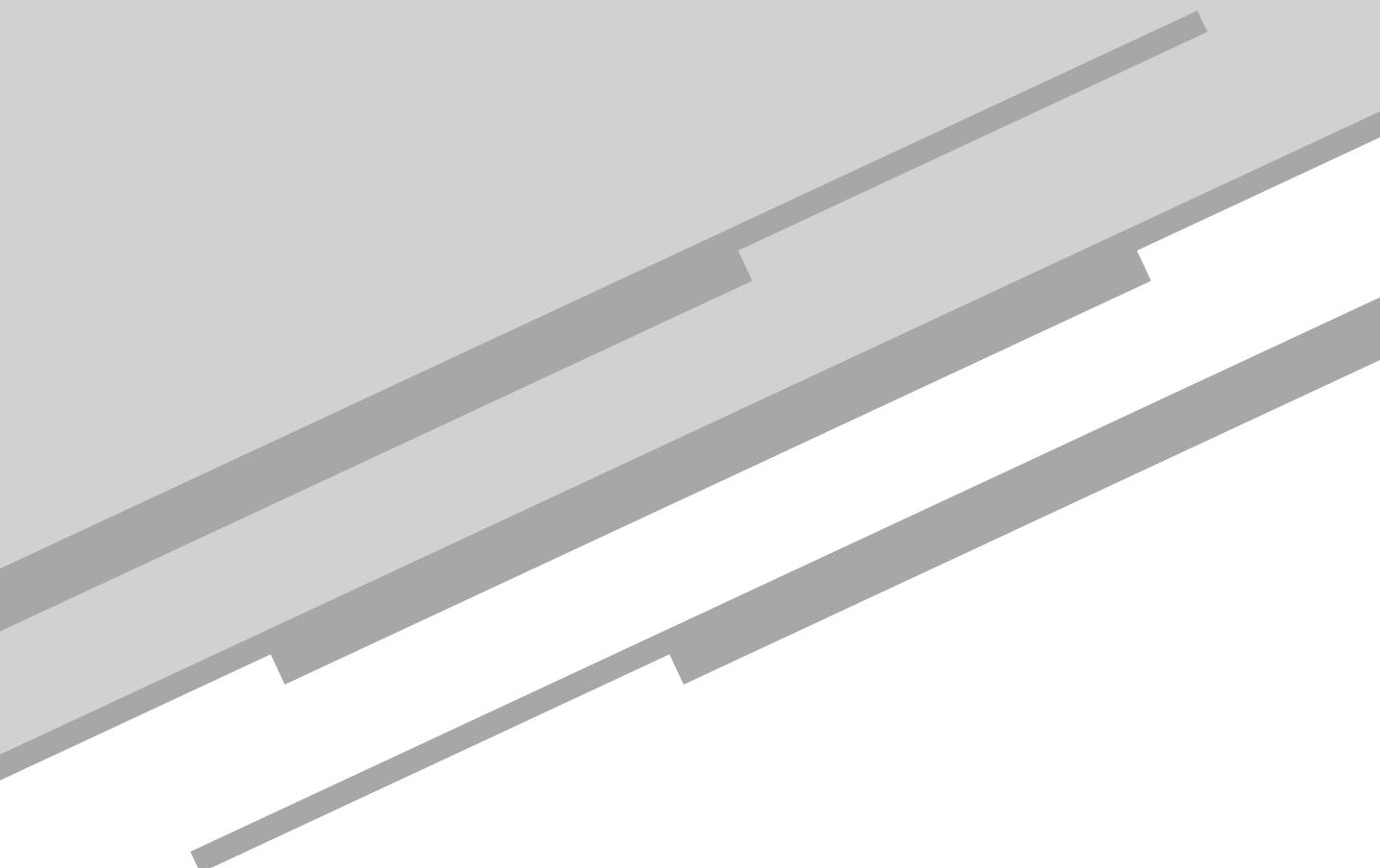
Nesta aula, fomos apresentados a métodos de organização de dados que mais facilitam o processo de leitura da amostra. Para os casos de dados categóricos, vimos que uma simples organização, em ordem alfabética ou por setor, já ajuda bastante. Contudo, para amostras com dados de mais de uma variável, o recurso da tabela de contingência é um excelente instrumento para contabilizar as informações e fazer uma interpretação prévia. Já com dados numéricos, percebemos que a ordenação crescente ou decrescente também ajuda – ainda mais para amostras que possuem características que se encaixam no Princípio de Pareto. Desse modo, ao ordenarmos na forma crescente, este padrão ficará em evidência. Por fim, ainda sobre dados numéricos, aprendemos como montar uma tabela de frequência – Recurso este que auxilia resumindo uma grande amostra em classes, sem camuflar os padrões e comportamento dos dados que compõem a amostra.

## **Informação sobre a próxima aula**

Na próxima aula, iremos ver como exibir alguns resultados e conhecer os principais tipos de gráficos que temos à nossa disposição. Ainda: como escolher estes gráficos e qual o propósito de cada um deles. Falaremos, também, de alguns erros comuns que podem comprometer a apresentação de resultados.

# Aula 5

Vai uma pizza, aí?



*Rafael Canellas Ferrara Garrasino*

## **Meta**

Apresentar recursos gráficos, mais conhecidos, que ilustram dados e resultados de processos estatísticos.

## **Objetivos**

Esperamos que, após o estudo desta aula, você seja capaz de:

1. identificar tipos de recurso gráfico;
2. fazer uso, de forma correta, de recursos gráficos em situações para os quais são aplicáveis;
3. listar erros comuns que podem comprometer o efetivo uso de recursos gráficos.

## Introdução

Um das artes implícitas na Estatística é a capacidade de transformar resultados ou uma gama de informações em imagens, neste caso, gráficos. Diversas vezes nos deparamos com uma quantidade absurda de informações que precisam ser exibidas em um relatório ou para um determinado público e que, se não optarmos por uma solução gráfica, com certeza, ficará algo confuso, cansativo e desinteressante.

O recurso gráfico possui, como maior vantagem, a capacidade de resumir informações em uma única imagem e de torná-las potencialmente compreensíveis para quem as lê. Contudo, para que isto seja possível, é imprescindível que a pessoa escolha corretamente o gráfico de acordo com as informações a exibir, além de não cometer pequenos erros comuns que acabam comprometendo o resultado final.

Existem diversas opções de gráficos para utilizar como recurso em uma apresentação. Contudo, como dito, alguns possuem uma característica que condicionam a ser utilizados em alguns casos mais específicos. Por sua vez, outros são mais abrangentes no que se refere à aceitação em resultados distintos. De qualquer forma, é importante dominar qual tipo de gráfico fica mais bem adaptado a um tipo de informação. E, quando se utiliza corretamente o gráfico para o seu propósito original, costume, particularmente, dizer que ficou muito mais elegante.

Entretanto, isto não é comumente analisado por quem elabora o gráfico de uma apresentação. Em diversos momentos, o critério para escolha do gráfico envolve decisões como “este é mais bonito”. Obviamente que o desejo de toda pessoa que vai ilustrar o resultado da sua pesquisa é que ele seja bonito. Porém, se não for claro, direto e se não atingir os seus objetivos iniciais, vai ser apenas uma figura bela sem sentido. Portanto, nesta aula, iremos conhecer os modelos mais comuns e para quais situações eles mais bem se adequam, bem como ampliar as opções de recursos gráficos, conhecendo os propósitos originais de cada um. De igual modo, discernir a funcionalidade dos recursos, podendo, então, escolher corretamente o tipo de gráfico para cada situação.

## Gráfico pizza

Indiscutivelmente, o mais famoso de todos. Toda pessoa ao elaborar o resultado de algum projeto já cogitou em primeiro lugar usar um gráfico pizza. Entretanto, ele é o mais seletivo quanto a sua aplicação. Isto é: trata-se do tipo de gráfico que mais chances possui de ser usado em situações não aplicadas a ele.

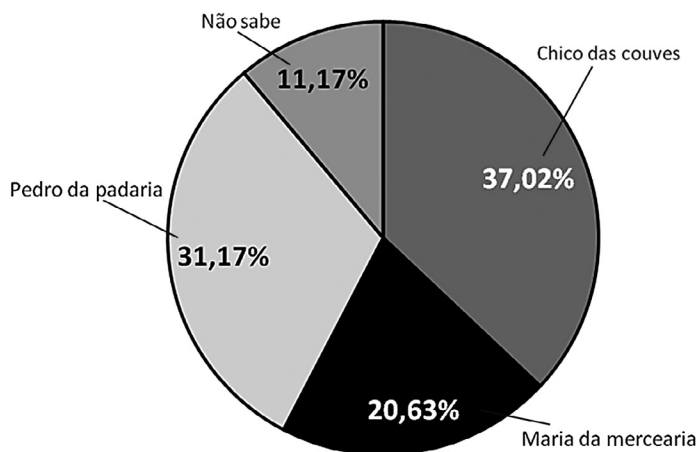
Assim, o seu formato remete à ideia de totalidade. Logo, obrigatoriamente, a soma das informações exibidas em um gráfico pizza sempre tem de totalizar 100%. Uma boa maneira de associar esta ideia é recordando de quando aprendemos fração. Normalmente, a professora utiliza a representação pizza para dar a ideia de totalidade. Ali, cada fatia da pizza representa uma parte dela, mas juntando todos os pedaços voltamos a ter uma pizza inteira.

Suponhamos, então, uma campanha para prefeito de uma cidade do interior. Temos três candidatos disputando a vaga. Foi feita uma pesquisa com a intenção de voto de cada eleitor, fornecendo a opção de dizer que não sabe ainda em quem votar. O resultado foi consolidado na **Tabela 5.1**.

**Tabela 5.1:** Candidatos x Eleitores x Percentual de intenção de votos

Candidatos	Eleitores	Percentual
Chico das couves	5.830	37,02%
Maria da mercearia	3.250	20,63%
Pedro da padaria	4.910	31,17%
Não sabe	1.760	11,17%
Total	15.750	100,00%

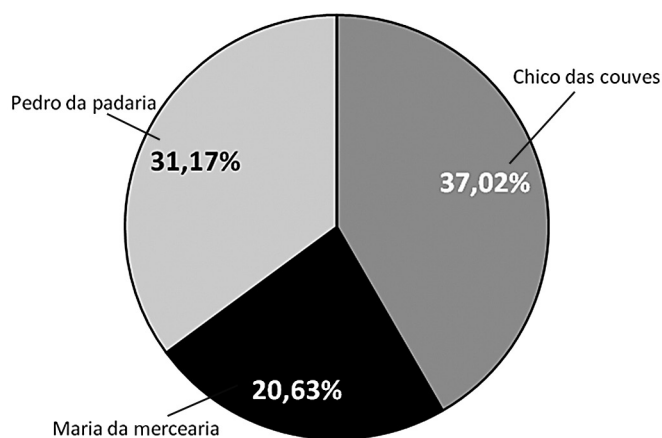
Com os dados coletados e organizados, calculamos o percentual de cada candidato e das pessoas que não sabem ainda em quem votar. Agora, com estas informações já é possível gerar um gráfico pizza conforme a **Figura 5.1**.



**Figura 5.1:** Gráfico pizza: percentuais de intenção de votos e inclusão da opção “Não sabe”.

Notem que as informações ficaram claras ao utilizar o recurso do gráfico pizza. Cada opção assumiu um tamanho diretamente relacionado ao seu percentual de respostas. Observem, também, que a soma desses percentuais totalizaram 100%, isto é, todas as pessoas entrevistadas.

É importante salientar que caso o pesquisador tenha esquecido ou optado por não selecionar, por exemplo, as pessoas que responderam que não sabiam ainda em quem votar, o gráfico adaptaria o tamanho das fatias para gerar uma figura inteira. Contudo, a soma dos resultados não totalizaria 100%. Isto deixaria a informação ilustrada sem sentido algum, causando confusão ao leitor. Vejamos esta hipótese na **Figura 5.2**.



**Figura 5.2:** Gráfico pizza – percentuais de intenção de votos sem a inclusão da opção “Não sabe”.

Como foi dito, a forma inteira da pizza foi gerada. Contudo, a soma dos percentuais ficou aquém, totalizando pouco menos de 90%. Isto se deu pela exclusão de uma parte da pesquisa que completaria os 100% que a “pizza” representa.



É sempre importante que um gráfico tenha um título. Mesmo contendo o nome do gráfico na legenda da figura ou que se insira o título em destaque, você facilita a compreensão por parte do leitor ao nomear o gráfico.

É importante ressaltar que nem toda informação relacionada à percentual deva ser ilustrada em gráfico pizza. Mais à frente veremos um caso para ilustrar melhor.

## Gráfico de barras

Este talvez seja o segundo gráfico mais conhecido. Seu maior recurso é a comparação entre valores. Pelo seu formato de barras proporcionais, fica mais fácil identificar que a maior barra é o parâmetro com maior valor e, em contrapartida, a menor barra é o parâmetro de menor valor.

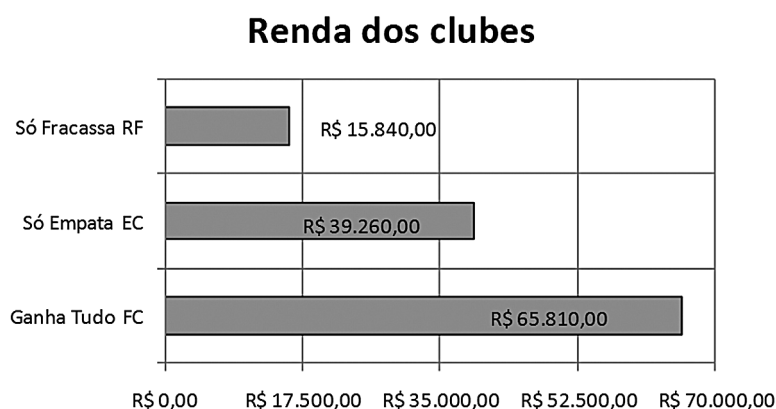
Por diversas vezes, ao optar por este recurso, o pesquisador decide por abrir mão do próprio valor e dedica-se apenas ao formato geométrico gerado. Isto é: o que importa, em alguns casos, é quem tem o maior valor, o segundo maior e assim por diante. Optando-se por este tipo de gráfico, esta informação fica facilmente observável.

Assim, suponhamos um estudo com a receita obtida com a venda de ingressos de cada time futebol de um determinado estado. Colhida as informações chegamos à **Tabela 5.2**.

**Tabela 5.2:** Time x Renda

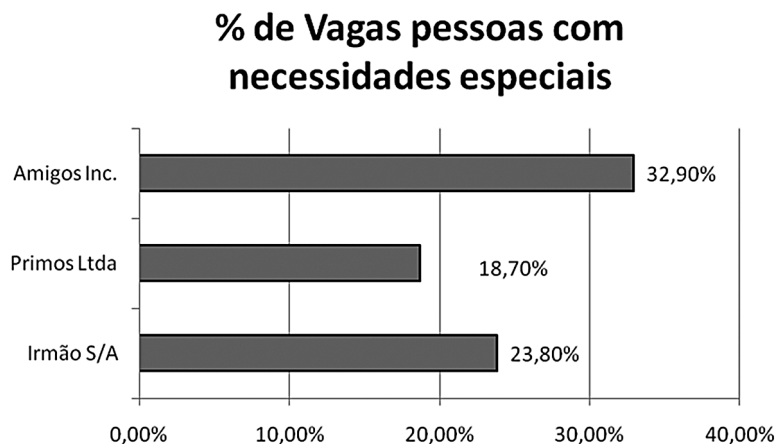
Time	Renda
Ganha tudo FC	R\$ 65.810,00
Só empata EC	R\$ 39.260,00
Só fracassa	R\$ 15.840,00

Repare que a Tabela por conta própria já explicita a maior renda, bem como a segunda e a terceira. Contudo, ao optar pelo recurso gráfico, a opção do gráfico de barras deixará tão evidente quanto. Vejamos na **Figura 5.3**.

**Figura 5.3:** Gráfico de barras – time x renda de clubes.

Note que mesmo com a exibição dos valores das rendas, fica visível pelo tamanho da barra, que a renda do Ganha Tudo FC é a maior dentre os três Times. Logo, ao inserir o valor de cada parâmetro, o pesquisador optou por não somente destacar o maior pelo recurso gráfico, mas também em detalhar o tamanho. De qualquer forma, o gráfico de barras cumpriu exatamente o seu propósito original: destacar os valores pelo tamanho.

Como foi dito anteriormente, nem todo caso que envolva percentual deve ser ilustrado pelo gráfico pizza. Casos nos quais os valores não possuem relação entre si e, por consequência, não totalizam 100% devem ser ilustrados através de outros gráficos. Vejamos na **Figura 5.4** o exemplo de uma pesquisa feita com três empresas, as quais foram questionadas acerca de quantas vagas reservam para pessoas portadoras de necessidades especiais.



**Figura 5.4:** Gráfico de Barras – % de vagas para pessoas portadoras de necessidades especiais.

Repare que o percentual de vagas reservadas para pessoas com necessidades especiais da empresa Amigos Inc. não tem relação com o percentual de vagas da empresa Primos Ltda, tampouco com o percentual de vagas da empresa Irmão S/A. Logo, associar essas informações em um gráfico pizza não faria sentido. Além disto, o propósito deste gráfico de barras é ilustrar o quanto cada empresa reserva e qual reserva mais, portanto, a opção do gráfico de barras foi acertado.



Apesar do lema “Mais é menos”, utilizar efeitos e detalhes ao construir um gráfico é valorizar o seu trabalho. Contudo, exageros podem comprometer a sua boa intenção.

## Histograma

Seu intuito é ilustrar uma tabela de frequência. Portanto, trata-se de um recurso muito útil quando se refere à redução de muitas informações em poucas. Uma tabela de frequência, como vimos anteriormente, é um recurso de resumir uma amostra grande de dados em classes. Logo, é fato o ganho de otimização de resultado utilizando uma tabela de fre-

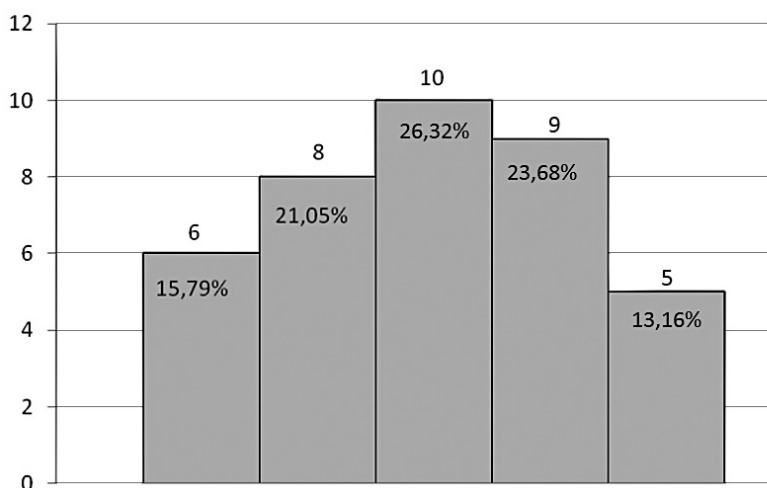
quência. Contudo, esse ganho pode ser melhorado ao ilustrar esta mesma tabela usando apenas um gráfico.

O histograma é um gráfico de barras verticais sem espaço entre elas. Cada barra representa uma classe (faixa) e a respectiva quantidade de dados que ela contém. Vejamos a **Tabela 5.3** como exemplo de uma tabela de frequência qualquer.

**Tabela 5.3:** Faixas e Frequências

Faixa	Freq.	Freq. acum.	% Freq.	% Freq. acum.
De 1,15 até 2,15	6	6	15,79%	15,79%
De 2,16 até 3,16	8	14	21,05%	36,84%
De 3,17 até 4,17	10	24	26,32%	63,16%
De 4,18 até 5,18	9	33	23,68%	86,84%
De 5,19 até 6,19	5	38	13,16%	100%
Total	38	38	100%	100%

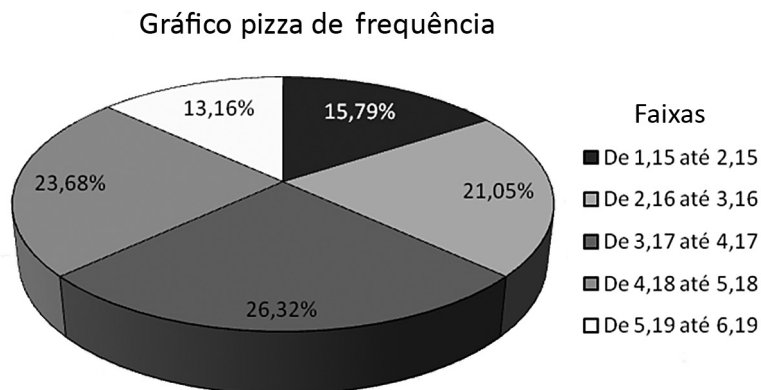
Com a tabela de frequência gerada, basta gerar o histograma utilizando apenas as frequências de cada faixa. Sendo assim, temos a **Figura 5.5**:



**Figura 5.5:** Histograma gerado a partir da Tabela 5.3 (Faixas e frequências).

Note que visualizando a Figura 5.5 fica claro que a quantidade de dados nas faixas vai aumentando até a terceira e depois volta a decair. Isto, graficamente falando, é o intuito principal do histograma.

Mais uma vez podemos ratificar que informações percentuais podem e devem ser utilizadas em gráficos diferentes do gráfico pizza. Os percentuais de frequência de cada classe são suficientes para gerar o histograma conforme vemos na própria Figura 5.5. Contudo, como existe uma relação de totalidade entre eles, é possível utilizar um gráfico de pizza conforme a **Figura 5.6**.



**Figura 5.6:** Gráfico Pizza gerado a partir dos dados da Tabela 5.3 (Classes e Frequências).



Legenda é um recurso imprescindível para um gráfico, seja qual for o seu estilo. Com a legenda, as chances do leitor não ficar “perdido” na leitura do gráfico são bem maiores – além de demonstrar um cuidado seu com quem irá acessar o seu material ou pesquisa.

Veja que, tecnicamente falando, não existe erro algum em gerar esse gráfico pizza. Os valores fazem sentido e completam 100%. Contudo, como pretendemos representar uma tabela de frequência, o histograma é mais completo, pois, além de indicar o percentual, acusa o movimento (crescimento/decrescimento) dos dados e a relação (tamanho) entre eles.

## Atividade 1

### Atende ao objetivo 2

Um estagiário está responsável por ilustrar três estudos distintos, podendo, inclusive, utilizar o mesmo tipo de gráfico em estudos diferentes. Determine e justifique quais as opções que ele possui para cada estudo.

- a) Uma pesquisa sobre o faturamento total dos cinco maiores hotéis de uma determinada cidade no último verão.

---

---

---

---

---

---

- b) Uma pesquisa na qual perguntaram a 50 turistas quanto cada um gastou em compras na atual viagem.

---

---

---

---

---

---

- c) Uma pesquisa na qual perguntaram aos hóspedes de um hotel qual o motivo da viagem (lazer, negócios ou estudos).

---

---

---

---

---

---

### **Resposta comentada**

- a) Como são informações individuais, sem relação entre si, só podemos ilustrar com um gráfico de barras. Cada barra representará o faturamento de um hotel. Se os cinco hotéis pesquisados fossem os únicos da cidade, poderíamos utilizar um gráfico de pizza. A “pizza” inteira simbolizaria toda a receita com hospedagem naquela cidade, dentro daquele período, e cada fatia representaria o faturamento individual de cada hotel.

b) Como se trata de uma pergunta aberta, teremos todos os tipos de respostas. Valores aleatórios diversos surgirão para serem estudados. Logo, a primeira etapa desta ilustração será organizar os dados mediante tabela de frequência. Tendo a amostra resumida em uma tabela de frequência, podemos ilustrá-la com um histograma, gráfico de barras ou de pizza conforme dito anteriormente. Contudo, uma tabela de frequência sempre fica mais bem ilustrada em um histograma.

c) O estudo possui como espaço amostral os hóspedes de um determinado hotel. Cada um deu uma resposta dentro das fornecidas. Portanto, uma ilustração com gráfico pizza será o mais adequado. Contudo, não podemos negar que ele também poderia ser ilustrado com um gráfico de barras, o que deixaria a representação gráfica menos elegante.



## Diagrama de Pareto

Na aula anterior vimos o Princípio de Pareto. Notamos o quanto ele pode ser útil para determinar a parte principal que requer mais atenção em uma amostra. Agora, veremos como ele pode ser ilustrado.

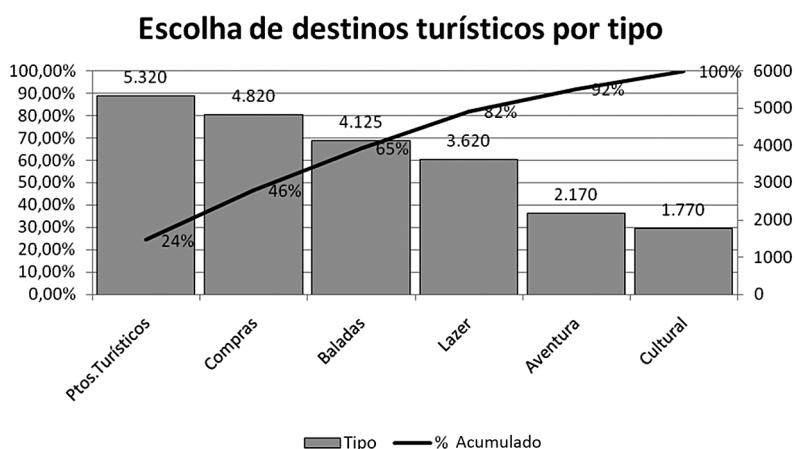
O intuito será o mesmo que vimos anteriormente. Iremos posicionar inicialmente o maior resultado, depois o segundo maior e assim sucessivamente até chegarmos ao último. Assim, ficarão em destaque os maiores valores para que o leitor possa identificá-los mais rapidamente, além de identificar visualmente o decrescimento da amostra.

Contudo, se apenas deixarmos assim a ilustração, teremos um simples gráfico de colunas. O diferencial será na inclusão de uma linha de tendência a qual irá determinar a evolução acumulada do percentual de participação dos elementos da amostra. Vejamos um exemplo: suponhamos que uma companhia de turismo receptivo tenha feito um levantamento com a quantidade de clientes que optam por cada tipo de destino. Dentre as opções foram selecionadas as seguintes: pontos turísticos (englobavam os pontos clássicos da cidade), aventura (trilhas e escaladas), compras (shopping e mercados), cultural (museus e casas antigas), lazer (passeio de barco, praia e estádios) e baladas (bares, restaurantes e casa noturnas). Após a contagem feita chegaram aos seguintes resultados representados na **Tabela 5.4**.

**Tabela 5.4:** Tipo de evento turístico x Quantidade de usuários x % Acumulado

Tipo	Qtde.	% Acuml.
Ptos.Turísticos	5320	24,38%
Compras	4820	46,46%
Baladas	4125	65,36%
Lazer	3620	81,95%
Aventura	2170	91,89%
Cultural	1770	100,00%

Repare que, como vimos na aula anterior, estudos como este podem ser ilustrados em uma tabela mais bem detalhada. Contudo, como nosso objetivo é apenas ilustrá-los em um gráfico, este pequeno punhado de dados será suficiente. Vejamos como fica a **Figura 5.7**.

**Figura 5.7:** Diagrama de Pareto: evento turístico x Quantidade de usuários x % Acumulado.

Como foi dito, os maiores valores ficam em destaque logo no início da figura. Isto já é suficiente para o leitor que desejar ter acesso rapidamente às informações mais importantes. Além disto, nota-se que graficamente podemos perceber o quanto são maiores os primeiros valores dos demais.

Contudo, o objetivo aqui é na parte que se refere ao Diagrama de Pareto. Ele é visível na linha auxiliar que corta o gráfico. Note que ela faz o movimento oposto das colunas. Enquanto estas estão em movimento

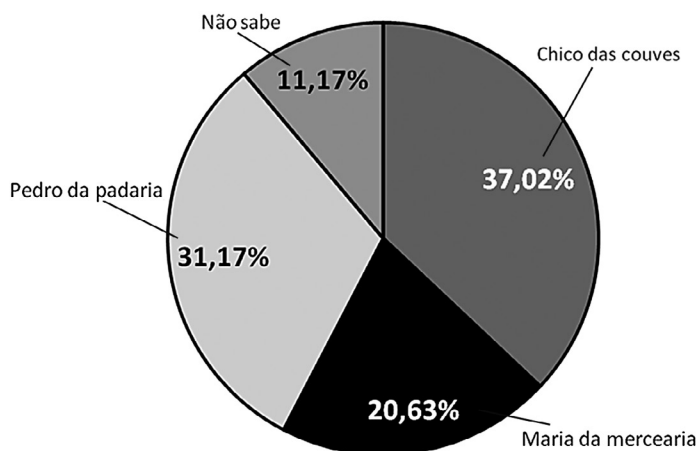
decrecente, a linha segue um movimento crescente. Isto se dá porque ela acumula o percentual de participação dos elementos da amostra. Isto é: quando estamos em Pontos Turísticos, falamos de aproximadamente 24% das respostas da pesquisa. Ao passarmos para Compras, temos aproximadamente 46% das respostas, das quais em torno de 24% delas já sabemos que se referem a Pontos Turísticos. Assim sucessivamente. Note que, com este recurso, podemos graficamente identificar os maiores valores e visualizar, também, que com os dois primeiros valores estamos falando de quase metade das respostas. Portanto, temos praticamente uma excelente ferramenta para resumir informações e de alta praticidade para fazer a leitura dos dados.

## Atividade 2

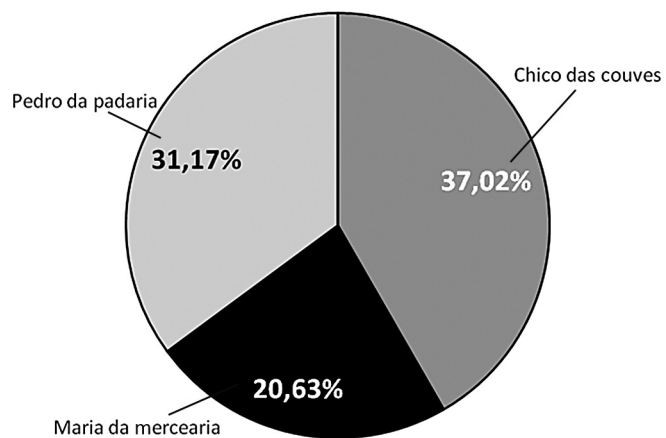
### Atende ao objetivo 3

Inspirado nas dicas para uma boa apresentação, considere, dentre as figuras que já vimos nesta aula, o que pode ser revisto nas citadas. Justifique.

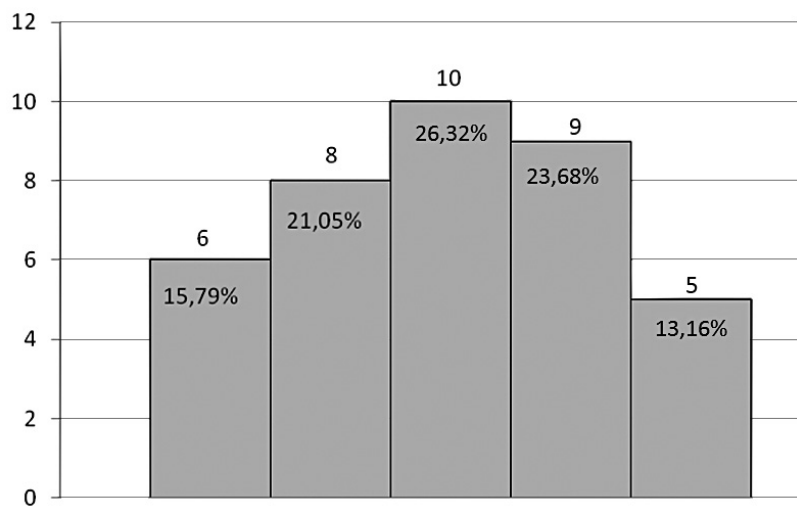
a) Relativa à Figura 5.1.



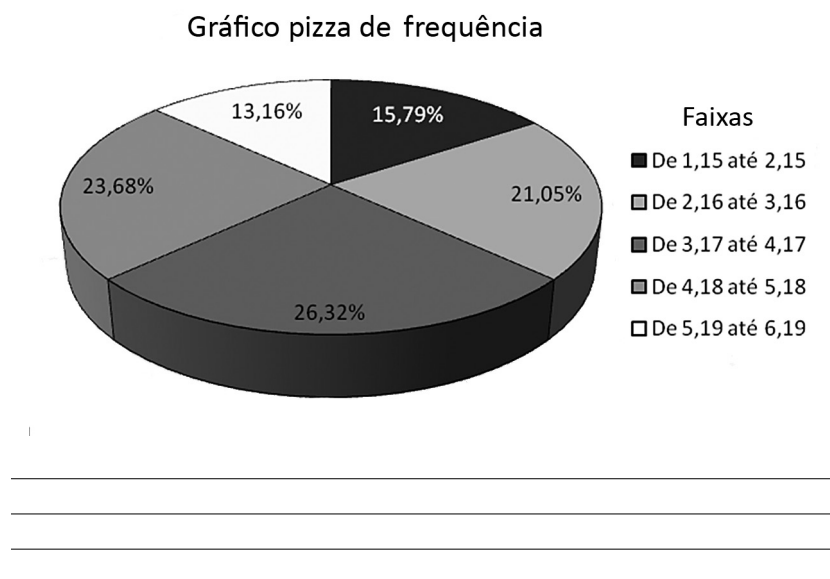
b) Relativa à Figura 5.2.



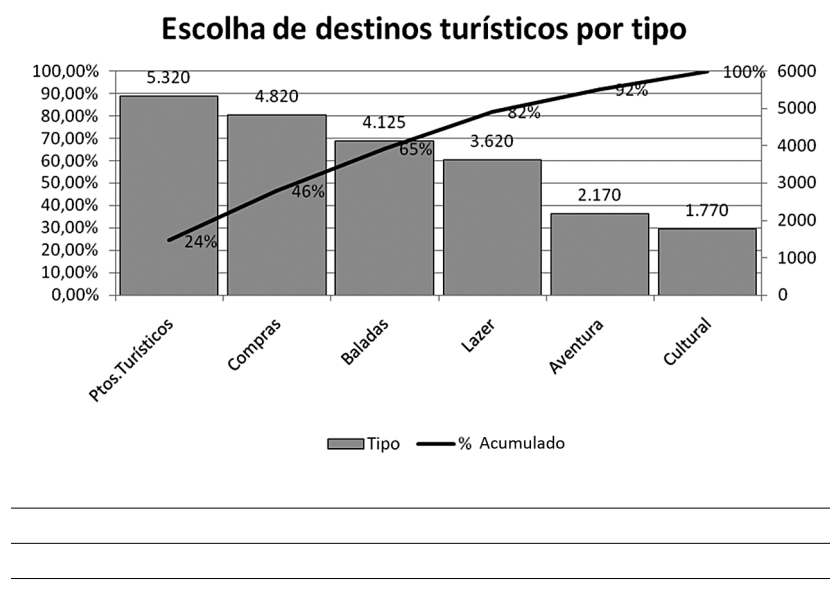
c) Relativa à Figura 5.5.



d) Relativa à Figura 5.6.



e) Relativa à Figura 5.7.



**Resposta comentada**

- a) Relativa à Figura 5.1 – O principal é a ausência de título. Nota-se que a imagem fica até perdida no meio do box de figura sem um título. Um fundo diferenciado também ajudaria bastante, mas sem exageros.
- b) Relativa à Figura 5.2 – Basicamente a mesma reflexão assinalada quanto à Figura 5.1.-
- c) Relativa à Figura 5.5 – Mais uma vez percebemos de início a gritante falta de um título. Contudo, a legenda também se faz necessária. Percebam que não se tem a menor noção do quê cada coluna representa.
- d) Relativa à Figura 5.6 – Nota-se um excesso de informação neste gráfico. Alguns valores ficam próximos uns dos outros sem entender de qual informação se trata. Esse é o típico caso que por já possuir, naturalmente, muita informação (colunas e linha de Pareto), optar por fundos e muitos números vai tornar a figura exagerada e confusa.
- e) Relativa à Figura 5.7 – Existe uma evidente poluição na tentativa de fazer um trabalho gráfico. O fundo ficou destoante com a figura em um todo e, como se não fosse suficiente, ele atrapalha no contraste das cores do gráfico – que é o principal.

**Série temporal**

Em alguns estudos nos deparamos com amostras das quais cada valor está associado a um determinado momento de um espaço de tempo. Isto é: existe uma relação direta entre o valor e o momento no qual ele ocorreu. Quando sua amostra se comporta de tal maneira, temos uma série temporal.

Nas séries temporais, a relação valor/tempo é muito importante, tanto que ao, por descuido, inverter a posição de um valor com outro, de um momento diferente, teremos um novo significado para esta amostra. Vejamos um exemplo: suponhamos que um hotel tenha feito o levantamento da quantidade de diárias vendidas nos últimos dois anos em cada mês. O resultado foi organizado na **Tabela 5.5**:

**Tabela 5.5:** Quantidade de diárias de hotel vendidas nos anos 2011 e 2012

2011												
Mês	jan	fev	mar	abr	mai	jun	jul	ago	set	out	nov	dez
Diárias	812	872	630	590	510	685	799	513	481	459	566	863

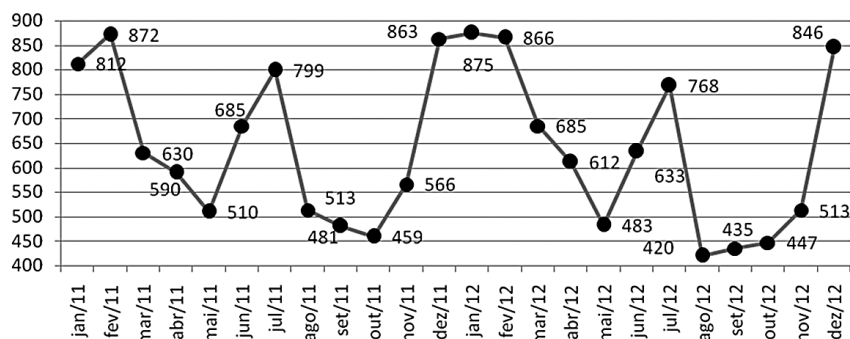
2012												
Mês	jan	fev	mar	abr	mai	jun	jul	ago	set	out	nov	dez
Diárias	875	866	685	612	483	633	768	420	435	447	513	846

## Sazonal

Relativo à estação do ano. Diz-se que algo é sazonal quando ocorre com maior frequência em um período do ano que nos demais. Por exemplo: venda de ovos de páscoa, aparelhos de ar-condicionado, pacotes de viagens, brinquedos etc.

Repare que os números indicam que, conforme esperado, a quantidade de diárias obedece a um padrão **sazonal**. Temos mais diárias nos períodos de férias e festividades que nos demais meses. Isto nos remete ao que foi dito anteriormente. Se, por algum acaso, trocássemos as quantidades de um mês *dezembro* qualquer com um mês *agosto* qualquer, teríamos uma informação confusa. Não saberíamos explicar porque poucas diárias em um período de muitas e, de igual modo, muitas em um período de poucas. Daí, a forte relação valor/tempo. Contudo, isso não significa que seja impossível acontecer. Basta supor que naquele mês de dezembro o hotel, por motivos aleatórios, precisou fazer uma obra inviabilizando muitos quartos. Assim, no citado mês de agosto, podemos considerar que tivemos um grande evento na cidade ou promoção de diárias.

Voltando à amostra, sabemos que se tivéssemos apenas a Tabela, conseguiríamos tirar algumas conclusões sobre o comportamento das diárias no passar do tempo. Entretanto, o recurso gráfico para Séries Temporais torna ainda mais fácil essa leitura. Vejamos na **Figura 5.8**.

**Figura 5.8:** Gráfico para séries temporais: quantidade de diárias de hotel vendidas nos anos 2011 e 2012.

Conforme foi dito, ao gerar o gráfico ficou muito mais evidente o comportamento dos valores na linha do tempo. Ficam gritantes os períodos de alta e de baixa temporada. Isto se dá independentemente da existência ou não dos valores no gráfico. O seu formato já é suficiente para uma leitura completa da variação do movimento no hotel.

É importante notar que essa mesma amostra poderia ser ilustrada por um gráfico de colunas, respeitando-se a ordem dos dados. Contudo, teria uma poluição de imagens e muitas colunas. Deste modo, quando se sabe que tudo que queremos é apenas esboçar o comportamento dos dados, esta opção que tivemos fica mais “limpa” e elegante.



Apesar da vontade de ilustrar todos os valores ser grande, é importante considerar que muitos valores podem poluir a imagem – como tivemos na Figura 5.8. Quando possível, limite-se a colocar apenas os valores que serão comentados ou destacados. Os demais, por proximidade, poderão ser entendidos. Assim a imagem fica mais “limpa” e o leitor menos confuso.

## Gráfico de dispersão

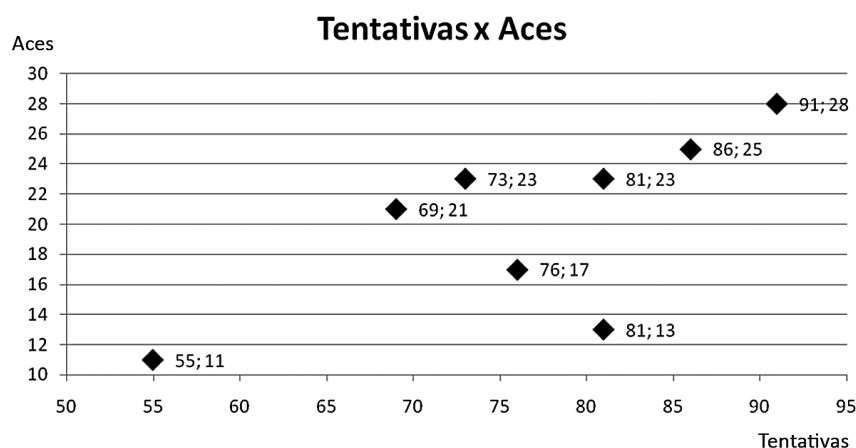
Em diversas situações nos deparamos com estudos nos quais precisamos analisar a suposta relação entre duas amostras. Para tal, existe um artifício estatístico que será comentado em um momento mais oportuno. Contudo, de antemão, podemos conhecer o gráfico de dispersão.

Este tipo de gráfico é exclusivo para relação entre duas amostras. Ele, basicamente, posiciona os valores de uma das amostras no eixo X e os valores da outra no eixo Y. Com isto, seu resultado é bem próximo de um gráfico cartesiano. Pontos serão posicionados de acordo com a combinação X e Y. Vejamos um exemplo: um time de voleibol resolveu utilizar do artifício estatístico para aprimorar um dos recursos mais eficiente deste esporte – o ponto de saque (*ace*). Seu técnico contabilizou quantas tentativas cada jogador praticou e quantos *aces* conseguiu. Com isto, chegou nos dados apresentados na **Tabela 5.6**.

**Tabela 5.6:** Jogador x Tentativas x Acertos

Jogador	Tentativas	Acertos
Rogério	86	25
Fernando	73	23
Alíton	91	28
Marcelo	55	11
Jefferson	81	13
Formiga	69	21
Thomaz	76	17
Dudu	81	23

Nesse caso, em específico, a leitura das informações em uma simples tabela não se mostra muito prático e tampouco fácil. Existem diversas possibilidades. É possível que quanto mais um jogador pratique, mais acertos tenha. Contudo, nessa ótica estamos desconsiderando o talento natural de cada um. Outra perspectiva é que a prática leva aos acertos, mas que depois de certas quantidades o cansaço ou a repetição exaustiva pode comprometer o resultado. Também podemos supor que a sorte tem um grande fator associado ao resultado. Entretanto, como interpretar isso através dos números? Como disse, este tipo de estudo requer uma análise algébrica que será vista à frente, mas um gráfico de dispersão poderá ajudar no entendimento da relação entre as amostras em questão. Vejamos na **Figura 5.9**.

**Figura 5.9:** Gráfico de dispersão: Tentativas x Acertos (pontos de saque).

No gráfico, temos que no eixo X estão as *tentativas* e no eixo Y a quantidade de *aces* (pontos de saque). Repare que o primeiro ponto a aparecer corrobora com a ideia de poucas tentativas, poucos acertos. Inversamente podemos falar o mesmo dos três pontos no canto superior direito. Eles reforçam a ideia de muitas tentativas e muitos acertos. Contudo, se nos atentarmos para o segundo ponto mais baixo, temos a impressão de que muitas tentativas não significam necessariamente muitos acertos ou até mesmo desenvolver a ideia de cansaço após certa quantidade de tentativas. Mas também teremos impressões opostas se nos atentarmos para o ponto que representa o jogador Fernando (73;23). Ele praticamente foi um dos melhores e pouco praticou.



Gráficos como o de dispersão não exigem legenda. Contudo, vários pontos soltos podem ser difíceis para o leitor compreender o que se passa. Aconselha-se colocar, pelo menos, o nome de alguns deles para que possam identificar mais facilmente os que pretende comentar. Para o caso exemplificado, Figura 5.9, se tivesse colocado o nome de uns três jogadores, a visualização dos pontos que citei seria mais rápida.

Deste modo, a análise gráfica por si apenas nem sempre será suficiente. Recursos algébricos para casos como este são imprescindíveis. Todavia, não se faz dos gráficos um recurso dispensável. Eles são complementares e, quando utilizados juntos, tornam mais fácil a interpretação dos dados.

### Atividade 3

#### Atende ao objetivo 1

Identifique o nome dos gráficos ilustrados a seguir:

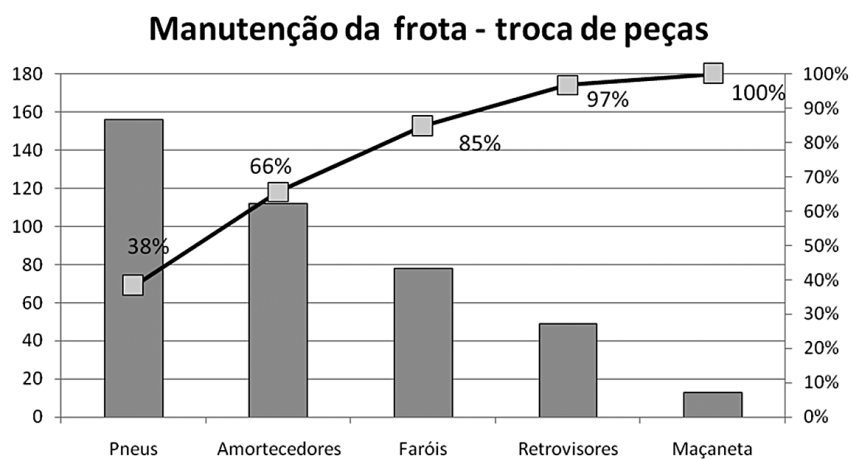


Figura A: \_\_\_\_\_

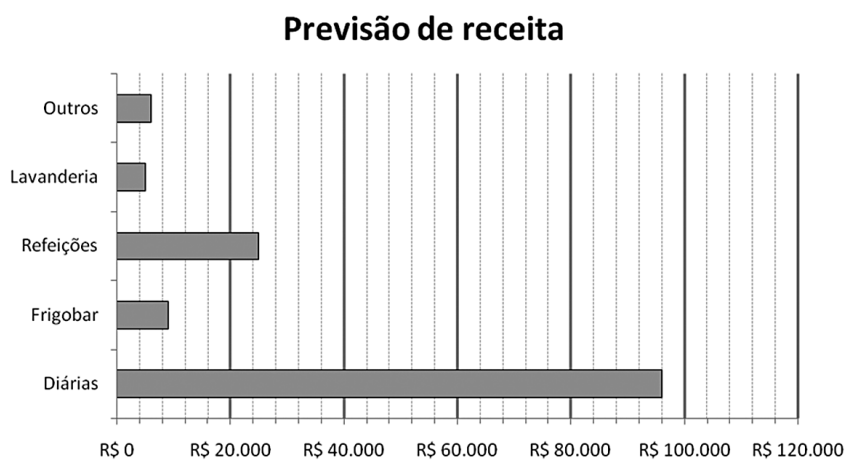


Figura B: \_\_\_\_\_

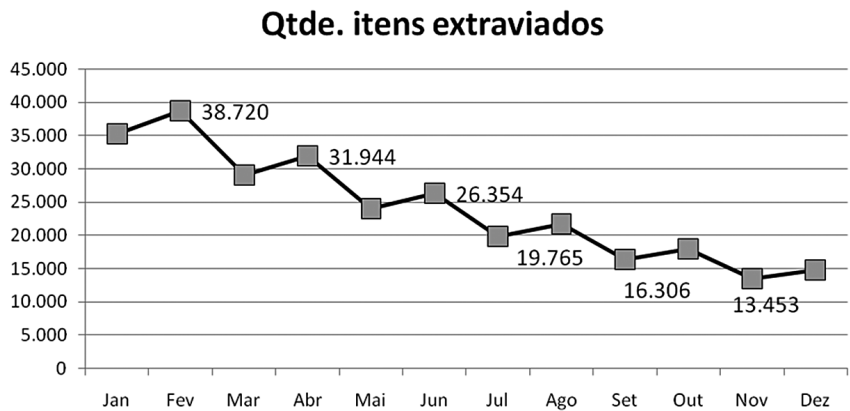


Figura C: \_\_\_\_\_

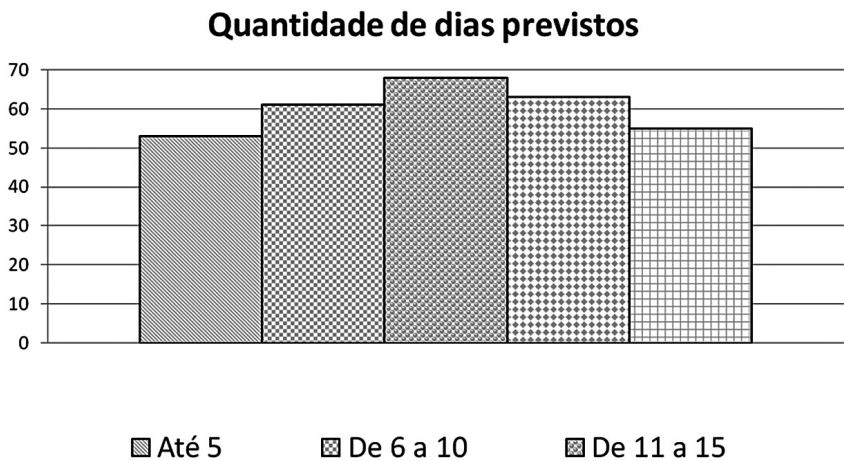


Figura D: \_\_\_\_\_

**Resposta comentada**

Figura A: Diagrama de pareto

Figura B: Gráfico de barras

Figura C: Séries temporais

Figura D: Histograma

=====

## Conclusão

Como vimos até agora, o processo estatístico envolve várias etapas e todas com um elevado grau de importância no todo. Contudo, não basta efetuar um excelente processo estatístico e comprometer todo o trabalho com uma apresentação ineficaz ou nada compatível com ele.

Gráficos, em uma visão genérica, foram feitos para ilustrar resultados e informações de muitos detalhes em uma maneira mais prática para a fácil compreensão do leitor. Contudo, a aplicação dos gráficos é rigorosamente restrita aos tipos de situações para os quais foram desenvolvidos.

Uma exibição com um gráfico não adequado ao caso em questão pode gerar diversas críticas ao final da mesma e, mesmo sendo um excelente trabalho, pode acabar ficando como o ponto mais evidente da sua apresentação. Isto se deve pelo momento da exibição, obviamente, ser o que mais marca os espectadores. O mesmo acontece com gráficos excessivamente poluídos ou confusos. Normalmente em casos como estes, os espectadores, por não conseguirem compreender as informações, fazem muitas perguntas, demoram a entender o que estava sendo ilustrado e, talvez, por isso, uma apresentação simples, fica longa, cansativa e, na maior parte das vezes, termina com muitas pessoas sem entender.

De fato, ilustrar dados com gráficos é uma arte. Entretanto, uma arte nada complexa, com poucas regras que, ainda assim, são flexíveis. O princípio número um desta arte é o bom senso. O segundo é experiência. Veja outros gráficos com um ponto de vista crítico. Tente absorver o que queria ser ilustrado e o que foi feito. Assimile as boas iniciativas feitas e, acima de tudo, assimile as iniciativas que não deram certo. O objetivo final é sempre ser claro. Para tal, basta ser direto, mas sem economia. Esqueça que você fez o estudo e domina todas aquelas informações. Para você, elas serão sempre óbvias. Desenvolva um gráfico sempre pensando que quem vai “lê-lo” é um completo leigo no assunto. Isto fará com que elabore gráficos que atingirão diretamente o seu público-alvo.

## Atividade final

### Atende aos objetivos 1 e 2

Foi encomendada uma pesquisa que seria dividida em três etapas, da qual você está responsável pela ilustração dos resultados de cada uma delas. A primeira parte consiste basicamente em questionar aos hóspedes de um hotel qual parte mais agradou: atendimento; decoração; alimentação e bebidas; conforto ou segurança. Feito isto, irão selecionar a opção menos escolhida e pedir para os clientes darem uma nota de 0 a 10 que necessariamente deverá ser um valor inteiro. Estas notas serão organizadas da seguinte forma: menores que 4; de 4 até 7, acima de 7. Depois disto, compararão o total de notas acima de 7 que este quesito recebeu com o total de notas também acima de 7 que ele mesmo recebeu nos últimos onze meses.

Determine os gráficos que deverão ser utilizados para ilustrar cada etapa do estudo, justificando suas escolhas:

---

---

---

---

---

---

---

---

---

---

### **Resposta comentada**

O primeiro estudo pede o gráfico pizza. Existem algumas opções e cada cliente escolherá uma delas. Logo, a relação de totalidade entre os resultados será suficiente para ser representado por uma “pizza”. O segundo, como foi organizado em classes, por indução vai gerar uma tabela de frequência (mesmo com apenas 3 classes). Portanto, o uso de um histograma é o mais adequado. Por fim, o terceiro irá recorrer a uma série histórica dos dados em uma linha de tempo ininterrupta. Então, temos como melhor das opções uma série temporal.

---

---

---

---

## Resumo

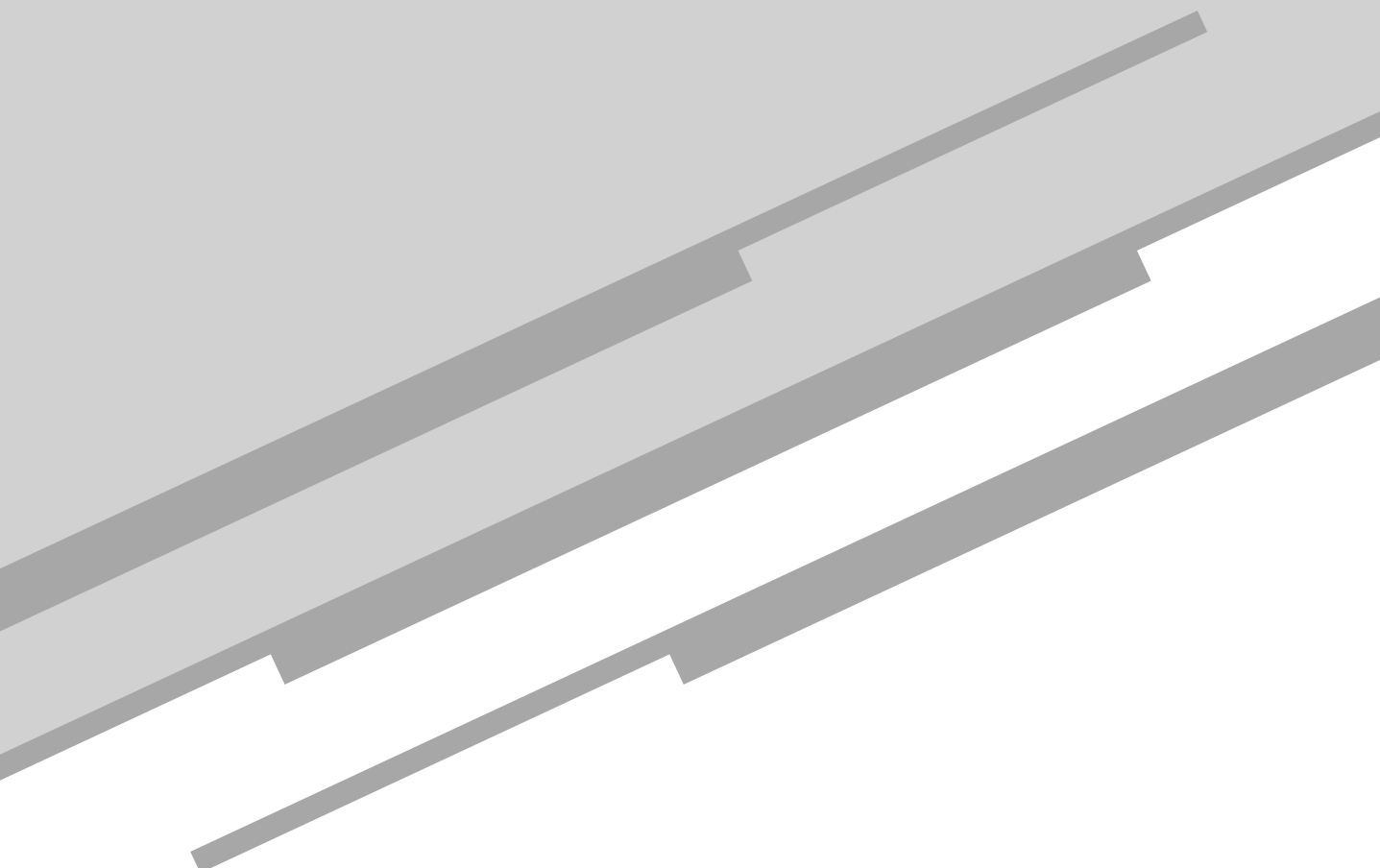
Nesta Aula fomos apresentados as mais famosas formas de ilustrar os dados e/ou resultados de um estudo estatístico. Conhecemos o gráfico pizza que, mesmo sendo o mais famoso de todos, é o mais rígido na sua aplicação. Vimos que o gráfico de barras é bastante abrangente nas suas utilidades e, potencialmente, pode ser usado como um recurso de segurança. Em seguida tivemos o histograma que basicamente se utiliza para ilustrar tabelas de frequência que, em contrapartida, pode ser ilustrada de outras formas. Posteriormente, vimos o diagrama de Pareto, que é ótimo para exibir os elementos que compõem a maior parte do seu estudo ou os que causarão maior impacto. A série temporal foi apresentada pela primeira vez. Esta possui uma aplicação bem específica, mas, talvez por isso, seja a mais clara de se determinar quando deva ser utilizada. Assim como o gráfico de dispersão, que também é bastante pontual e focado em um tipo de estudo em especial.

## Informação sobre a próxima aula

Na próxima Aula, finalmente, iniciaremos os trabalhos diretamente com os dados. Iremos fazer cálculos, tirar breves conclusões sobre amostras e determinaremos a postura dos dados que a compõem. Basicamente, após aprender a coletar, organizar e ilustrar dados, começaremos a dar tratamento a eles. Portanto, preparem as calculadoras que agora vai começar a ficar animado!

# Aula 6

Estou ao lado do baixinho careca!



*Rafael Canellas Ferrara Garrasino*

## Meta

Apresentar recursos estatísticos que estabeleçam como dados de uma amostra se comportam quando se determina um valor central.

## Objetivos

Esperamos que, após o estudo desta aula, você seja capaz de:

1. determinar a Média Aritmética de uma amostra;
2. indicar a Mediana de uma amostra;
3. apontar a Moda de uma amostra;
4. especificar os 4 Quartis de uma amostra.

## Introdução

Comumente, usamos referências para especificar algo, sua localização ou até mesmo um grupo de pessoas. Quando estamos esperando alguém, jamais falamos que estamos ao lado de uma pessoa caucasiana de altura mediana e com cabelos bem cortados. Se você quer ser encontrado, precisa determinar um ponto de referência único para que a pessoa que irá ao seu encontro o faça com precisão. Em momentos como esses, geralmente, usamos baixinhos carecas, gordinhos de camisa vermelha, idoso em cadeira de rodas ou a menina rebelde toda de preto e cabelos roxo. Isso se dá porque, provavelmente, estas referências irão se destacar das demais ao seu redor.

Deste modo, o mesmo pode ser dito quando analisamos a estratégia defensiva de um time de futebol. Como podemos afirmar que um determinado time está, claramente, usando a famosa retranca? Usamos a bola como referência! Daí fica claro que todos os seus jogadores estão posicionados “atrás da linha da bola”. Talvez sem uma referência deste tipo ficasse mais complexo afirmar o posicionamento excessivamente defensivo deste time.

Na Estatística, o mesmo se faz necessário. Temos uma amostra com certa quantidade de dados e precisamos determinar como estes dados estão se comportando e como estão se agrupando. Contudo, sem uma referência, fica tudo muito vago e sem precisão. A utilização de um valor central típico, como balizador para determinar o agrupamento dos dados de uma amostra ao seu redor, é chamado de *Tendência Central*.

O que veremos a partir daqui são os métodos mais conhecidos de se determinar um valor central típico para ser utilizado quando precisarmos fazer uma leitura do comportamento da amostra. De igual modo, também serão destacados: fornecer opções de ferramentas estatísticas capazes de auxiliar o aluno ao analisar o comportamento dos dados quando definido um ponto central como referência; entender o processo de determinação destes pontos de acordo com a quantidade de dados de uma amostra; conhecer as limitações de cada uma destas ferramentas e quando não são unicamente aplicáveis.

## Média aritmética

Talvez a mais conhecida de todas e comumente chamada apenas de Média. Pode-se dizer que é a mais justa de todas no que se refere a consi-

derar os dados de uma amostra. Para ela, todos os dados de uma amostra têm o mesmo peso. Isto é, ao determinar a média de uma amostra, levamos em consideração, apenas, o valor do próprio dado.

A maioria das pessoas já construiu, uma vez na vida, uma frase usando termos como “... na média ...” e/ou já calculou a média de algo. Assim, o que quase todos já calcularam foi a média escolar. Pegávamos a nota da primeira prova, somávamos a nota da segunda prova e dividíamos por dois. O cálculo da média é feito exatamente assim e independente da quantidade de dados da sua amostra. Vejamos a fórmula no formato matemático:

$$\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$$

**Figura 6.1:** Média aritmética – fórmula no formato matemático.

Apesar de assustadora, a fórmula não diz nada muito diferente do que foi feito com as notas das provas. Ela afirma: some todos os dados que possui e divida por esta mesma quantidade de dados. Isto é: somou a nota das duas provas e dividiu por dois, que é a quantidade de provas que estamos estudando. Vejamos um exemplo (**Tabela 6.1**) com um pouco mais de dados para confirmar que entendemos: suponhamos uma turma com vinte alunos e suas respectivas idades.

**Tabela 6.1:** Alunos x Idades

João	Pedro	Fernanda	Maria	Jorge	Ana	Diogo	Claudia	Thiago	Carol
19	21	23	22	23	24	19	24	20	21
Felipe	Bruno	Thais	Julia	Telma	Fabio	Nina	Carlos	Cris	Igor
21	18	20	19	20	21	22	23	19	22

Como dito, para calcularmos a média, precisamos somar todos os dados da amostra e dividir pela quantidade de dados. Vejamos:

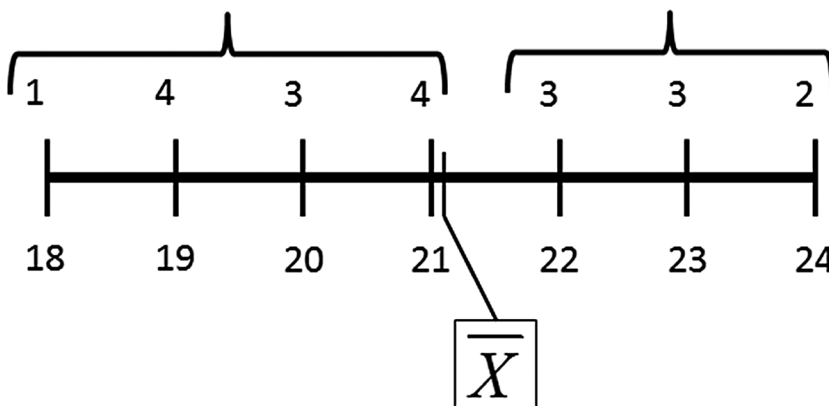
$$\overline{X} = \frac{19+21+23+22+23+24+19+24+20+21+21+18+20+19+20+21+22+23+19+22}{20}$$

$$\overline{X} = \frac{421}{20} = 21,05$$

**Figura 6.2:** Média aritmética: soma x divisão por quantidade de dados.

O que concluímos ali é que a média de idade da turma é de 21,05 anos. Obviamente, sabemos que, em raros casos (1 em 365 e veremos mais a frente porque deste número), as pessoas possuem idades redondas. O esperado é que tenhamos idades quebradas, como 20 anos, 5 meses e 12 dias. Contudo, por fins práticos, optamos em determinar uma idade redonda para cada aluno. Todavia, o resultado não foi redondo, como era de se esperar, mas ainda assim representa bem próximo da realidade a idade média da turma. Se tivéssemos feito o cálculo com as idades precisas de cada um, mais preciso seria nosso resultado. Mas o que representa este 21,05? Em que pode ser útil?

Inicialmente, ao falar que a média da turma é de 21,05, podemos ter uma grande segurança ao preparar, por exemplo, uma aula voltada para alunos com esta idade. Citar referências muito antigas ou voltadas para pessoas muito maduras, mesmo novas demais, talvez não tenha tanto impacto quanto as específicas para esta faixa etária. Outra grande utilidade é notar como os dados estão se comportando em relação a esta média. Isto é: quantos dados eu tenho para cada idade e, deste total de dados, quantos estão acima, abaixo, próximo, e distante da média.



**Figura 6.3:** Média aritmética da turma: 21,05

Podemos ver que temos 12 alunos com idade abaixo da média e 8 acima da média. Percebemos, também, que quase um terço (7 alunos) do total das idades está bem próxima da média (21 e 22). Essas leituras já são bastante úteis para uma melhor compreensão da amostra em questão. Contudo, ao mesmo tempo que a média considerar todos os dados de forma igual é ponto positivo, também pode ser um ponto negativo, pois desta forma valores discrepantes da amostra irão compro-

meter diretamente o resultado. Vejamos um exemplo (**Tabela 6.2**) no qual, usando a mesma turma, tenhamos a entrada de uma aluna que decidiu voltar mesmo depois de muito tempo. A amostra ficará da seguinte forma:

**Tabela 6.2:** alunos x idade discrepante

João	Pedro	Fernanda	Maria	Jorge	Ana	Diogo	Claudia	Thiago	Carol	Alzira
19	21	23	22	23	24	19	24	20	21	78
Felipe	Bruno	Thais	Julia	Telma	Fabio	Nina	Carlos	Cris	Igor	
21	18	20	19	20	21	22	23	19	22	

Calculando a nova média, temos 23,76 anos. Isto é: a entrada da aluna Alzira alterou a média da turma, consideravelmente, em 2,75 anos. Melhor dizendo, aumentou a média da turma em praticamente 13%. Tal aumento é muito elevado! Isto só ocorreu porque entrou um dado excessivamente elevado para a amostra em questão. O mesmo poderia acontecer se, em vez da Alzira, tivesse entrado o Leozinho com 1 ano. Ele diminuiria a média em quase um ano.

Deste modo, devemos deixar claro que apenas a análise da média não é algo confiável para uma leitura de amostra. Em alguns casos, o resultado não fará sentido algum, pois como falamos é um resultado matemático. Ele apenas opera os dados fornecidos, não sendo capaz de interpretar a situação que envolve estes dados. Vejamos o mesmo caso da turma original, mas em uma situação estapafúrdia (**Tabela 6.3**). Em vez de perguntar a idade de cada aluno, iremos perguntar quantos seios cada aluno possui. Teremos o seguinte resultado:

**Tabela 6.3:** alunos com seio x alunos sem seio

João	Pedro	Fernanda	Maria	Jorge	Ana	Diogo	Claudia	Thiago	Carol
0	0	2	2	0	2	0	2	0	2
Felipe	Bruno	Thais	Julia	Telma	Fabio	Nina	Carlos	Cris	Igor
0	0	2	2	2	0	2	0	2	0

Ao calcular a quantidade média de seios da turma, temos como resultado 1. Ora, é possível dizer que na média cada aluno tem um seio? Isto é, na média, os homens possuem um seio e as mulheres também? Claro que não! O que aconteceu aqui foi o uso indevido de um recurso matemático para calcular algo inapropriado. Não faz sentido calcular a quantidade de seios de uma turma. Sabemos que as mulheres (exceto por uma fatalidade) possuem sempre 2 seios e os homens nenhum.

**Atividade 1**

*Atende ao objetivo 1*

Foi feita uma pesquisa com a quantidade de dias em que cada visitante, a uma determinada cidade, pretendia ficar. Determine, caso fosse o gerente de um hotel desta cidade, quantas diárias está prevendo para cada hóspede, baseado nesses dados e apenas na média:

2	5	3	4	6	2	3	2	3	6
4	5	5	2	3	6	4	7	2	5
3	4	5	3	4	2	3	6	5	6
4	6	7	5	7	6	2	3	4	6

**Resposta comentada**

$$\overline{X} = \frac{170}{40} = 4,25$$

Em vista disto, como um hóspede não pode ficar com diárias não inteiras, estipulamos que é esperado em torno de 4 diárias para cada hóspede na média. Isto é: em um mês com 31 dias, deveremos ter, caso esteja lotado, por volta de 8 hóspedes passando por cada quarto.

**Mediana**

Podemos dizer que a *Mediana* é uma medida estritamente geográfica. Isto é: ela apenas determina em qual ponto iremos dividir a amostra em duas partes com quantidades iguais de dados. Neste quesito, pouco

importa para a mediana se os dados que ficaram antes dela totalizam um valor muito menor que os que ficaram após ela. Como disse, ela apenas divide a amostra ao meio e nada mais.

Ressalto, também, que o termo *mediana* é muito utilizado na Geometria e no Desenho Geométrico para determinar o segmento de reta que divide uma reta, semirreta ou outro segmento de reta ao meio.

Assim, como a mediana faz uma divisão na quantidade de dados, pouco importa os valores desses. A única informação necessária a ela é a própria quantidade de dados de uma amostra. A fórmula da mediana é representada da seguinte forma:

$$\text{Mediana} = \frac{n+1}{2}$$

**Figura 6.4:** Fórmula para o cálculo da mediana.

Deste modo, tudo o que devemos fazer para determinar a mediana de uma amostra é somar uma unidade ao total da quantidade de dados dessa e dividir por dois. Contudo, dois pontos precisam de uma atenção especial. O primeiro é que a mediana determina a posição do termo do meio. Logo, cabe à pessoa retornar à amostra e determinar qual é o dado naquela posição. A segunda, uma consequência da primeira, é que a amostra esteja sempre na ordem crescente. Vejamos um exemplo (**Tabela 6.4**) com a amostra de voos saindo de São Paulo para Paris nos últimos sete meses.

**Tabela 6.4:** Mediana: amostra com 7 dados

5	7	4	8	3	9	6
---	---	---	---	---	---	---

Como estamos falando de uma amostra com 7 dados, iremos somar uma unidade a 7 e posteriormente dividir por dois. Com isto, nossa mediana será 4. Entretanto, todo cuidado é pouco, pois não estamos falando que o valor 4 é a mediana. Estamos afirmando que a mediana está posicionada na 4ª posição. Contudo, antes precisamos ordenar de forma crescente nossa amostra (**Tabela 6.5**), para depois marcar a mediana:

**Tabela 6.5:** Mediana – ordenação de dados na forma crescente

3	4	5	<b>6</b>	7	8	9
---	---	---	----------	---	---	---

Note que a 4ª posição, depois de ordenada a amostra na forma crescente, pertence ao dado de valor 6. Isto é: a mediana dessa amostra é 6 – relativa aos voos de São Paulo para Paris. Assim, como dizemos que a mediana divide a amostra ao meio, faz todo o sentido a mediana ser 6, pois, antes desta temos 3 dados e após mais 3 dados.

Todavia, o sucesso desta operação, se notarmos com frieza, só foi possível porque a amostra é composta por um número ímpar de dados, que somado a um dá um valor par, que dividido por dois dará um resultado inteiro. Mas, e se a amostra for composta por um número par de dados? Ao somar uma unidade teremos um valor ímpar, o qual não permite uma divisão exata por dois. Como faremos? Suponhamos, então, uma pesquisa (**Tabela 6.6**) com a quantidade de turistas vindo do México nos últimos doze meses para a cidade de Pindamonhangaba. Os resultados foram:

**Tabela 6.6:** Dados de quantidade de turistas, nos últimos 12 meses, vindos do México para Pindamonhangaba – ordenação aleatória

345	822	633	278	458	694	578	844	531	768	483	419
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Temos nesse momento uma amostra de 12 dados. O seu cálculo da mediana permanece inalterado. Vejamos:

$$Mediana = \frac{12+1}{2} = \frac{13}{2} = 6,5$$

**Figura 6.5:** Cálculo da mediana.

Notamos, então, que não temos na nossa amostra a posição 6,5. Todavia, já sabemos que a posição 6,5 é antecedida pela posição 6 e sucedida pela posição 7. Logo, quando tivermos uma mediana de valor não inteiro, iremos recorrer à média das posições que, respectivamente, a

antecede e a sucede. Isto é: faremos a média da 6ª e da 7ª posição. Vamos então ordenar os dados na forma crescente (**Tabela 6.7**) para melhor enxergar quem é o 6º e o 7º dados desta amostra:

**Tabela 6.7:** Dados de quantidade de turistas, nos últimos 12 meses, vindos do México para Pindamonhangaba – ordenação na forma crescente

278	345	419	458	483	<b>531</b>	<b>578</b>	633	694	768	822	844
-----	-----	-----	-----	-----	------------	------------	-----	-----	-----	-----	-----

Temos, então, que o dado da 6ª posição é o 531 e o da 7ª posição é o 578. Agora faremos a média deles: cálculo o qual espero que já esteja dominado por vocês. Iremos somar os dois ( $531 + 578$ ) e dividir o resultado por dois (soma dos dados, dividida pela quantidade de dados). O resultado será 554,5, ou, se preferir, podemos arredondar para 555. Até porque ter 554,5 turistas é pouco provável. Este novo resultado, de 555, será a mediana. Ele irá ocupar a tal posição que não existia antes – a posição de número 6,5. Com isto, a amostra (**Tabela 6.8**) fica da seguinte forma:

**Tabela 6.8:** Dados de quantidade de turistas, nos últimos 12 meses, vindos do México para Pindamonhangaba – inclusão do dado “mediana”

278	345	419	458	483	531	<b>555</b>	578	633	694	768	822	844
-----	-----	-----	-----	-----	-----	------------	-----	-----	-----	-----	-----	-----

Note que agora tudo faz sentido. Conseguimos dividir a amostra em duas partes iguais. Do total da amostra, seis dados ficam antes da mediana e seis ficam após a mesma. Isto é: podemos dizer que na metade dos meses tivemos menos de 555 turistas na cidade de Pindamonhangaba, enquanto na outra metade tivemos mais de 555.

A mediana, que fique claro, por ser uma medida estritamente de posicionamento, não é afetada por valores extremos. Podemos trocar qualquer valor dessa amostra, mesmo por um valor cem vezes maior, que a sua mediana continuará entre a 6ª e a 7ª posição. Isto é uma vantagem que ela possui sobre a *média*. Contudo, sua desvantagem é que não faz uma leitura mais abrangente da amostra. Mas, quem sabe, as duas juntas são de grande valia.

## Atividade 2

*Atende aos objetivos 1 e 2*

Foi feita uma pesquisa em uma empresa na qual, por hábito, seus funcionários viajam muito a trabalho. O intuito é saber até que ponto a viagem de trabalho está afetando o lazer de seus funcionários. Perguntaram, então, quantas vezes ao ano eles viajam (desde fora da cidade, até para fora do país) por lazer. As respostas foram:

2	5	3	6	8	1	7	9	12	5
4	8	3	8	1	10	9	7	6	4
2	7	3	8	4	6	2	7	4	3

Em vista disto, determine sua média, sua mediana e trace uma linha com as quantidades respondidas na pesquisa, posicionando a média e a mediana para que melhor possamos entender como as respostas estão se comportando.

[illegible]

### Resposta comentada

Ordenando os dados temos:

1	1	2	2	2	3	3	3	3	4
4	4	4	5	5	6	6	6	7	7
7	7	8	8	8	8	9	9	10	12

Determinando a *média*:

$$\bar{X} = \frac{164}{30} = 5,47$$

O resultado da média deu uma dízima periódica. Contudo, para maior praticidade, adotaremos apenas duas casas decimais – até porque estamos fazendo um estudo o qual as respostas são números inteiros. Logo, se optássemos por uma casa decimal, para este caso em particular, não estaríamos cometendo um exagero. Portanto, adotaremos, então, 5,5 como média.

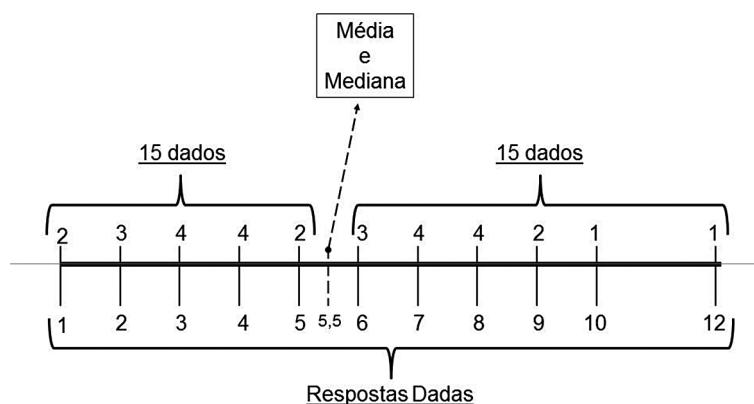
Determinando a *mediana*:

$$Mediana = \frac{30+1}{2} = 15,5$$

Como sabemos, o cálculo da mediana determina a posição na qual ela está posicionada. Como não temos a posição 15,5, pelas regras já citadas, iremos determinar a média entre os valores respectivamente nas 15ª e 16ª posições, que são as posições que “cercam” o resultado 15,5.

Na 15ª posição, temos o valor 5 e na 16ª posição temos 6. Logo, a *média* e ao mesmo tempo a *mediana*, deste estudo, será 5,5.

Deste modo, com estas informações em mãos, podemos traçar uma reta que ilustre o posicionamento dos dados para melhor entendermos como se comportam.



Notemos, por fim, que a *média* ficou igual à *mediana* da amostra, isto é, temos uma distribuição igualmente dividida por ambas as medidas calculadas. Note que, na parte de baixo da figura, temos os valores respondidos. Por sua vez, há na parte de cima a quantidade que cada valor foi respondido. Com isto, temos a certeza de que antes da *média* e da *mediana*, temos 15 dados ou respostas e o mesmo após essas. Oportunamente, iremos classificar esta amostra de acordo com a distribuição dos dados “ao redor” da *média* e da *mediana*.

---

---

## Moda

De uma maneira simplista, podemos dizer que, apesar de ser um recurso estatístico, a *moda* muito se assemelha ao conceito que comumente usamos para a palavra que dá seu nome. Assim, quando falamos que algo está na moda, estamos de uma maneira geral dizendo que aquilo é o mais usado, a tendência do momento, isto é, o mais repetido. Para o recurso estatístico, com este mesmo nome, a sua aplicação não é diferente. Basicamente, a moda determina o que mais foi respondido, o valor mais frequente, a medida mais usada e assim sucessivamente. Por isso, assim como dizemos que a *mediana* apenas divide a amostra em duas partes, podemos dizer que a *moda* apenas determina a resposta mais repetida e nada além. Vejamos um exemplo (**Tabela 6.9**) de uma pesquisa sobre países que cada 1 de 15 turistas pretende visitar, em uma futura oportunidade, mediante os seguintes indicativos em uma lista pré-elaborada: 6 = França; 9 = Chile; 15 = México; 18 = Egito; 24 = EUA.

**Tabela 6.9:** Turistas x cinco países que podem ser visitados

6	9	9	9	9
18	18	18	18	15
24	24	24	24	24

Contabilizando os dados, temos que, com 5 respostas, a *moda* é EUA, por ser o país mais citado. Note que, logo após, tivemos Chile e Egito com 4 respostas. Contudo, isto não importa, pois para a *moda* só consideramos o mais repetido e nada mais.

Deste modo, surge a dúvida de como seria se não tivéssemos o México nas células e outro resultado final da pesquisa (**Tabela 6.10**). Vejamos como ficaria:

**Tabela 6.10:** Turistas x quatro países que podem ser visitados

EUA	Chile	França	Chile	Egito
Chile	EUA	Chile	França	EUA
Egito	EUA	Chile	EUA	Egito

Agora temos dois países com a maior quantidade de respostas: EUA e Chile com 5 repetições. Portanto, mesmo havendo um empate, iremos afirmar que a *moda* desta nova amostra é EUA e Chile. Com isto, surge uma nova possibilidade ao substituírmos todas as respostas França por Egito (**Tabela 6.11**). A Amostra ficará da seguinte maneira:

**Tabela 6.11:** Turistas x três países que podem ser visitados

EUA	Chile	Egito	Chile	Egito
Chile	EUA	Chile	Egito	EUA
Egito	EUA	Chile	EUA	Egito

Note que mais uma vez temos um empate: EUA, Chile e Egito com 5 respostas cada. Todavia, pode haver a dúvida de como eleger a *moda*, tendo em vista que todos os dados da amostra estão empatados com mais votos. Ora, a regra da *moda* é sinalizar o mais votado e temos três enquadrados nesse requisito. Portanto, por mais que a *moda* seja todos os elementos da amostra, ela deverá ser anunciada. Então, a moda agora é EUA, Chile e Egito. Mas, e se fosse ao contrário? Isto é: nenhum dado respondido se repetisse. Vejamos:

**Tabela 6.12:** Turistas x três países que podem ser visitados

EUA	Chile	França	China	Egito
Cuba	Japão	Irã	Angola	Sudão
Irlanda	Suíça	Peru	Itália	Austria

Veja que temos mais uma vez um empate geral. Contudo, um empate coletivo de apenas uma resposta para cada um. Logo, não podemos dizer que houve *moda*, pois, diferente do empate anterior, neste não temos três tendências (EUA, Chile e Egito), mas, sim, respostas avulsas. Logo, para casos como este, afirmamos que não temos *moda* ou classificamos como *amodal*.

### Atividade 3

Atende aos objetivos 1, 2 e 3



Fonte: <http://www.flickr.com/photos/victorcamillo/5202318942/sizes/m/in/photostream/>  
- victorcamillo

Em uma cidade praiana, foi feito um levantamento com os “bugueiros” sobre quantos serviços de passeio, nas dunas, eles fazem por semana, chegando-se, assim, aos resultados apresentados na próxima tabela. Com estas as informações, determine sua média, mediana e moda:

21	18	15	24	18	12	21	27	30	15	9	12	15	21	24
9	30	21	15	12	27	27	24	18	6	9	21	24	27	30
9	18	15	18	24	30	27	21	9	12	15	18	24	21	15

---

---

---

---

---

---

---

---

---

---

---

---

**Resposta comentada**

Logo de início, devemos organizar os dados na ordem crescente:

6	9	9	9	9	9	12	12	12	12	15	15	15	15	15
15	15	18	18	18	18	18	18	21	21	21	21	21	21	21
24	24	24	24	24	24	27	27	27	27	27	30	30	30	30

Calculando a média, teremos:

$$\overline{X} = \frac{858}{45} = 19,1$$

Calculando em seguida a mediana, temos que ela é a 23ª posição que está ocupada pelo dado 18.

$$Mediana = \frac{45+1}{2} = 23$$

Por fim, pela tabela, na qual contamos cada dado, concluímos que a moda é 15 e 21 passeios.

<b>Dado</b>	6	9	12	15	18	21	24	27	30
<b>Qtde.</b>	1	5	4	7	6	7	6	5	4

---

---

---

## Quartil

Como o próprio nome já diz, seu objetivo é dividir a amostra em quartos. Logo, dividiremos em quatro partes iguais. Para tal, como consequência, teremos de determinar os quatro quartis. Vejamos a seguir.

### 1º Quartil

O primeiro quartil, necessariamente, dividirá a amostra em duas partes: antes dele estarão 25% da amostra e após os 75% restante, isto é, o 1º Quartil irá determinar o primeiro quarto da amostra e os três quartos restantes dela. Seu cálculo é feito da seguinte forma:

$$Q_1 = \frac{n+1}{4}$$

**Figura 6.6:** Fórmula para o cálculo do 1º Quartil.

De igual modo feito no cálculo da mediana, estamos lidando com a quantidade de dados da amostra. Logo, o processo apontará a posição que está o dado que determina o 1º Quartil. Portanto, seguindo os mesmos padrões, além da amostra estar em ordem crescente, permanece necessário (após usar a fórmula) retornar à amostra para que, com a posição calculada, apontar qual o dado que de fato será o 1º Quartil. Vejamos isto (**Tabela 6.13**) na amostra que segue – com dados aleatórios e já em ordem para a nossa comodidade.

**Tabela 6.13:** 1º Quartil X dados aleatórios organizados em ordem crescente

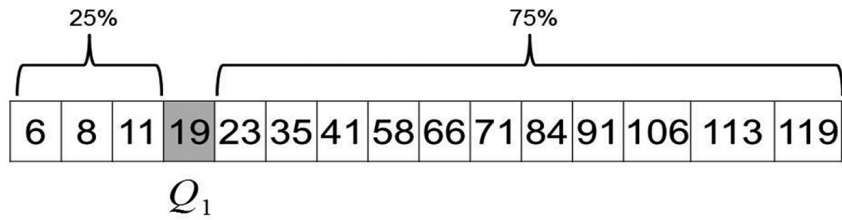
6	8	11	19	23	35	41	58	66	71	84	91	106	113	119
---	---	----	----	----	----	----	----	----	----	----	----	-----	-----	-----

Utilizando a fórmula, temos o seguinte resultado:

$$Q_1 = \frac{15+1}{4} = 4$$

**Figura 6.7:** Cálculo do 1º Quartil: dados da **Tabela 6.13**

Logo, o 1º Quartil está posicionado na 4ª posição, sendo; portanto, o dado 19. Com isto, nossa amostra fica com a seguinte divisão:



**Figura 6.8:** Indicação do 1º Quartil: 19.

Deste modo, podemos afirmar agora que 25% dos dados da amostra são de valores inferiores a 19 e o restante superiores ao mesmo.

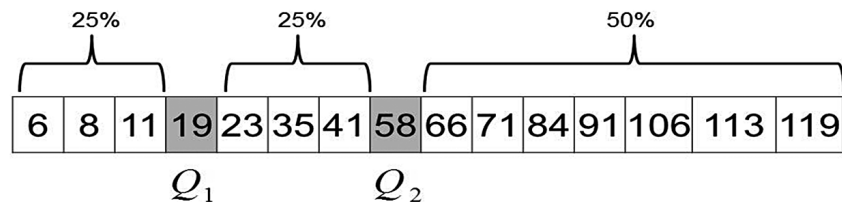
## 2º Quartil

Se o 1º Quartil separa o primeiro quarto da amostra dos demais, por intuição afirmamos que o 2º Quartil separa os dois primeiros quartos dos demais. Logo, estamos falando, necessariamente, de dividir a amostra ao meio. Para esta função já temos um recurso estudado: a Mediana. Portanto, seguramente podemos afirmar que a mediana e o 2º Quartil são necessariamente a mesma coisa. Com isto, torna-se desnecessário repetir o que já foi dito anteriormente. Vamos apenas reforçar, utilizando o mesmo exemplo da **Tabela 6.13**:

$$\text{Mediana} = Q_2 = \frac{15+1}{2} = 8$$

**Figura 6.9:** Mediana x 2º Quartil: cálculo.

Então, na 8ª posição estará o 2º Quartil que, remetendo à **Tabela 6.13**, indica que ele será o dado 58. Vejamos, então, como ficará a amostra com esta nova divisão:



**Figura 6.10:** Indicação do 2º Quartil: 58.

Agora temos a amostra dividida em três partes. A primeira, que antecede o 1º Quartil, compreende os primeiros 25% dos dados da amostra. A segunda, entre o 1º e 2º Quartil, é composta pelos segundos 25% dos dados da amostra. Por fim, após o 2º Quartil, temos os dois últimos 25% dos dados da amostra ou a segunda metade desta.

### 3º Quartil

De forma simétrica ao 1º Quartil, o 3º Quartil determina os primeiros 75% dos dados da amostra e os últimos 25% dos dados dessa. Isto é: ele separa a amostra em duas partes, ou seja, a primeira composta por três quartos dela e a segunda por um quarto da mesma. O seu cálculo deve proceder como os demais anteriores. Contudo, utilizando a fórmula a seguir:

$$Q_3 = 3 \cdot \frac{n+1}{4}$$

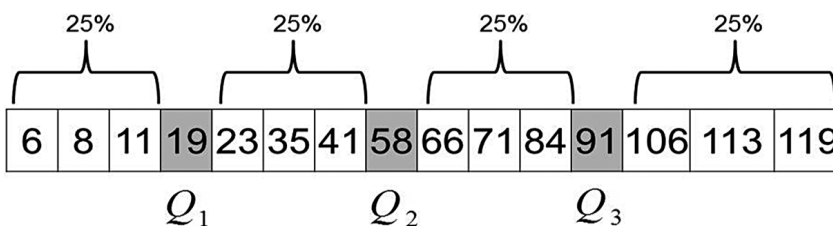
**Figura 6.11:** Fórmula para o cálculo do 3º Quartil.

Mais uma vez, vale reforçar que estas fórmulas apontam apenas a posição do dado que está no lugar do quartil em questão. Vejamos como ficará o exemplo já usado para os demais Quartis.

$$Q_3 = 3 \cdot \frac{15+1}{4} = 12$$

**Figura 6.12:** Fórmula para o cálculo do 3º Quartil.

Portanto, o 3º Quartil está localizado na 12ª posição que está ocupada pelo dado 91. Com isto, a divisão final dos dados ficará da seguinte forma:



**Figura 6.13:** indicação do 3º Quartil: 91

Agora temos nossa amostra com todos os quartos determinados. Portanto, podemos afirmar que a dividimos em quatro pedaços, contendo, igualmente, 25% dos dados da amostra.



Como estamos lidando com uma divisão, por quatro, é possível que obtenhamos 4 resultados diferentes:

1. se for uma divisão exata, o valor obtido será a própria posição do quartil, não havendo mais o que fazer;
2. se a divisão der um resultado do tipo 3,5, por exemplo, procederemos como no caso da mediana. Será feita a média entre os dados que o “cercam”. Neste caso, o da 3ª e 4ª posição;
3. se a divisão der um resultado do tipo 7,25, arredondaremos para baixo, adotando a 7ª posição como resultado;
4. se a divisão der um resultado do tipo 9,75, arredondaremos para cima, adotando a 10ª posição como resultado.

Com o recurso dos quartis passamos a ter um novo recurso para a leitura da amostra. Chama-se *amplitude interquartil* a diferença entre o 3º e o 1º Quartil. Vimos, anteriormente, que a amplitude é a diferença entre o maior elemento da amostra e o menor elemento desta mesma amostra. Contudo, sabemos também que em uma amostra podemos ter valores extremos que comprometem essa amplitude, tornando-a um valor elevado. Recorrendo à amplitude interquartil, temos um valor que não é influenciado por extremos, pois tanto o 1º quanto o 3º quartil não são influenciados por estes extremos. Logo, o resultado da subtração entre eles também não será. Valores que não são influenciados por extremos são chamados de *medidas resistentes*. Dentre as que estudamos temos a moda, a amplitude interquartil e os três quartis.

## Atividade 4

**Atende aos objetivos 1, 2, 3 e 4**

Os dados abaixo foram obtidos em uma pesquisa sobre a idade com que cada pessoa fez a sua primeira viagem sem a companhia de um responsável. Com eles, determine: média, mediana, moda, quartis, amplitudes e trace uma reta para ilustrar melhor o comportamento dos dados dessa amostra.

18	23	24	19	20	19	21	22	21	19	23	18	21	24	19	21	18
24	18	21	20	19	21	22	19	21	20	22	23	24	21	20	24	22

This image shows a single sheet of white paper with horizontal blue or grey ruling lines. The lines are evenly spaced and run across the width of the page. There are no margins, text, or other markings on the paper.

### Resposta comentada

1. Ordenando os dados, temos:

18	18	18	18	19	19	19	19	19	19	20	20	20	20	21	21	21
21	21	21	21	21	22	22	22	22	23	23	23	24	24	24	24	24

2. Calculando as medidas, temos os seguintes resultados:

$$\bar{X} = \frac{711}{34} = 20,9$$

$$Q_1 = \frac{34+1}{4} = 8,75$$

$$\text{Mediana} = Q_2 = \frac{34+1}{2} = 17,5$$

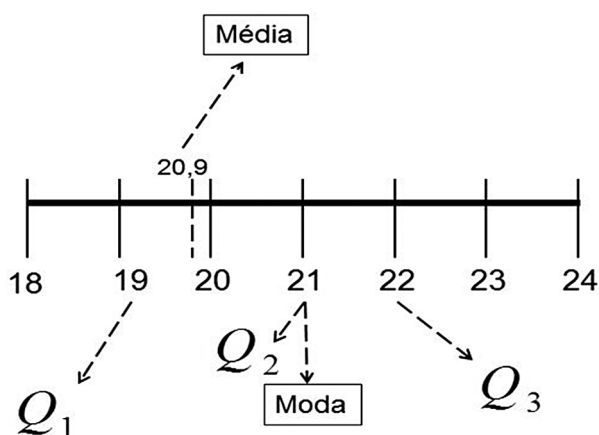
$$Q_3 = 3 \cdot \frac{34+1}{4} = 26,25$$

Pelos cálculos, temos que a média de idade é 20,9 (a opção de uma casa decimal é por conta de trabalharmos com valores inteiros na pesquisa, mas usar até três seria um recurso cauteloso indicado). O 1º Quartil indicou 8,75, o qual pela regra arredondamos para a 9ª posição que é o dado 19. O 2º Quartil indicou a posição 17,5, que pela regra pede a média entre a 17ª e 18ª posição, as quais são 21. Logo, a média será o próprio 21. Por fim, o 3º Quartil acusou a posição 26,25 que, seguindo a regra, será arredondada para 26ª posição ocupada pelo dado 22.

3. Usaremos uma tabela para calcular a repetição de cada dado e, assim, a moda da amostra.

Dado	18	19	20	21	22	23	24
Qtde.	4	6	4	8	4	3	5

Agora que temos que a moda dessa amostra é 21 anos, organizaremos as informações em uma reta e, por fim, determinaremos os demais valores pedidos.



4. Como o maior dado da amostra é 24 e o menor 18, temos que amplitude é 6 mediante a subtração entre 3º e 1º Quartil, temos 3 como amplitude interquartil.

## Conclusão

É fato que, para analisarmos uma amostra, precisamos de algumas referências. Simplesmente dizer que “é muito grande” nem sempre é suficiente, pois o que é muito grande para uma formiga, pode não ser muito grande para um elefante. Contudo, se tivermos dentro da própria amostra um parâmetro para balizar e dar condições de afirmarmos que é grande, pequeno, reto, torto, feio ou bonito, tudo fica mais fácil e incontestável.

Com a *média* de uma amostra, por mais que ela esteja suscetível a influências de valores discrepantes dos demais, podemos usá-la para identificar se um resultado está dentro do que era esperado. Com ela, podemos prever que tipo de informação potencialmente ocorrerá no próximo evento.

Com os *quartis* (nesse assunto incluímos a *mediana* que é um dos quartis), podemos dividir melhor os dados da amostra em partes, isolando os valores centrais que, possivelmente, influenciarão a média por estarem muito distantes dos demais. Logo, com essa parceria já alcançamos uma melhor leitura da amostra.

Por fim, com a *moda* poderemos notar padrões de repetição. Padrões que poderão indicar a possibilidade de um evento influenciar na média e, assim, ter mais probabilidade de ocorrer. Portanto, a reunião dessas informações é bastante necessária para uma melhor compreensão de uma amostra que estamos prestes a estudar.

## Atividade final

*Atende aos objetivos 1, 2, 3 e 4*

Foi feita uma pesquisa com alguns viajantes, perguntando quantos dias eles acreditam que são suficientes para esgotar todas as atividades que gostariam de fazer na cidade de Paraty/RJ. As respostas foram acumuladas na Tabela que segue:

5	6	4	8	9	7	4	6	5	3	4
4	4	6	7	3	6	9	5	7	3	8
5	7	9	3	6	8	8	5	4	7	9
6	4	3	5	9	6	4	7	8	3	4

Com estes dados, determine as medidas de tendência central apresentadas e, baseados nesses resultados, responda aos questionamentos a seguir:

- a) se fôssemos entrevistar mais uma pessoa e ela dissesse 8 dias, seria esta uma resposta esperada de acordo com o nosso estudo?
- b) considerando todos os dados disponíveis, estaria correto esperar uma variação entre as estadias com até 11 dias de variação?
- c) desconsiderando a parte que potencialmente compromete o resultado, podemos esperar uma variação na estadia de até 3 diárias?
- d) é possível que tenhamos nas próximas respostas uma grande quantidade de diárias a partir de 7?

This image shows a single sheet of white paper with horizontal blue or grey ruling lines. The lines are evenly spaced and run across the width of the page. There are no margins, text, or other markings on the paper.

### Resposta comentada

Vamos, inicialmente, como de praxe, ordenar os dados e, posteriormente, calcular as medidas de tendência central conhecidas:

3	3	3	3	3	3	4	4	4	4	4
4	4	4	4	5	5	5	5	5	5	6
6	6	6	6	6	6	7	7	7	7	7
7	8	8	8	8	8	9	9	9	9	9

$$\bar{X} = \frac{253}{44} = 5,75$$

$$Q_1 = \frac{44+1}{4} = 11,25$$

$$\text{Mediana} = Q_2 = \frac{44+1}{2} = 22,5$$

$$Q_3 = 3 \cdot \frac{44+1}{4} = 33,75$$

Com a *média* calculada de 5,75, podemos responder à letra A, na qual, segundo o nosso resultado, é esperado um resultado de aproximadamente 6. Logo, 8 está um pouco além do esperado. Não significa que não possa acontecer, mas a tendência é que seja por volta de 5,75.

A letra B versa, especificamente, da *amplitude total da variação*, pois ela pede para considerar todos os dados. Portanto, em uma amostra na qual o menor dado é 3 e o maior é 9, temos uma amplitude 6 diárias. Com isso, esperar uma variação de 11 diárias entre estadias é um exagero que não convém com o estudo feito.

Com os resultados dos *quartis*, temos que suas respectivas posições serão 11<sup>a</sup> no primeiro Quartil, média entre 22<sup>a</sup> e 23<sup>a</sup> para o segundo e 34<sup>a</sup> posição no terceiro. Logo, os quartis serão respectivamente 4, 6 e 7. Assim, podemos calcular a amplitude interquartil, que é a amplitude que descarta os valores extremos. Então, a amplitude interquartil é de 3 diárias, o que faz com que tenhamos que concordar com a pergunta da letra C.

Por fim, podemos notar que a *moda* é 4 diárias com oito repetições, seguido de 6 diárias com 7 repetições. Portanto, se tivéssemos que dizer qual quantidade de diárias potencialmente será respondida mais vezes, nas próximas perguntas, seria 4 e/ou 6. Logo, a partir de 7 diárias continua sendo possível, mas bem menos provável.

---

---

---

## Resumo

Nesta aula, vimos os recursos principais para determinar medidas que serão utilizadas como referências para auxiliar na leitura de uma amostra. A primeira medida central apresentada foi a *média* que, por considerar o valor de cada dado individualmente e de forma igual para todos, denota em um resultado referente às próprias informações passadas. Contudo, pode ser influenciado por valores exageradamente altos ou baixos em relação aos demais.

Em seguida, vimos medidas que não são influenciadas por estes valores extremos. Contudo, elas apenas interpretam o posicionamento dos dados. Isto é: se mudarmos os valores de todos, mas mantivermos a mesma quantidade de dados, tanto a *mediana* quanto os *quartis* estarão posicionados nos mesmos lugares.

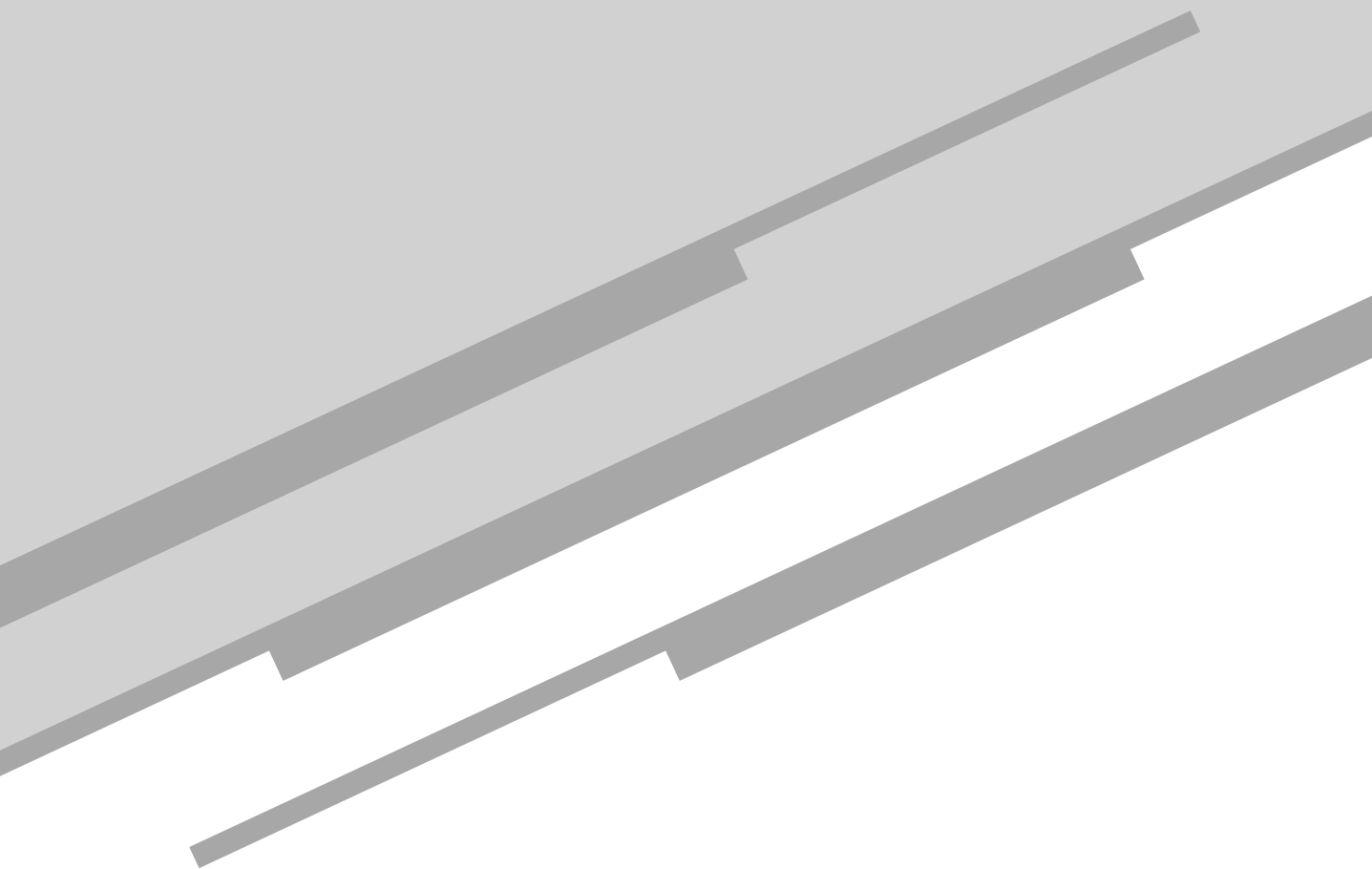
Vimos também a *moda*, que é uma medida preocupada apenas com a repetição mais frequente de uma informação. Assim, também não é vulnerável a valores extremos, levando o nome de *medida resistente*.

## Informação sobre a próxima aula

Na próxima aula, daremos prosseguimento às medidas numéricas descritivas, sendo que falaremos desta vez sobre Variação. Isto é: como os dados estão variando em relação à medida central da amostra estudada. Até lá!

# Aula 7

Três para cá, três para lá!



*Rafael Canellas Ferrara Garrasino*

## Meta

Estabelecer as formas de medir a dispersão dos valores de um conjunto de dados, conhecida como variação ou *spread*.

## Objetivos

Esperamos que, após o estudo desta aula, você seja capaz de:

1. determinar a Variância de uma amostra;
2. fazer uso do desvio-padrão de uma amostra;
3. definir a variação de uma amostra pelo Coeficiente de Variação.

## Introdução

Na aula anterior, vimos alguns instrumentos da Estatística para que possamos determinar o comportamento de um grupo de dados em relação a um valor central. Contudo, é importante também verificar como cada um desses dados comporta-se em relação a este valor central, isto é, estão todos “distantes” dele? Estão distribuídos de forma uniforme?

Deste modo, o que veremos a seguir é uma maneira complementar ao uso de valores centrais como instrumento de análise de uma amostra. Vimos, assim, que eles sozinhos não são tão precisos e fornecem uma leitura incompleta. Com os instrumentos que serão apresentados a partir de agora teremos mais recursos que, juntos com os já apresentados anteriormente, vão nos possibilitar, enfim, fazer uma leitura mais apropriada da amostra em questão.

Neste processo, você será capacitado a analisar um grupo de dados em relação a como se comporta de uma maneira ampla entre si. De igual modo, também poderão conceber a ideia de distribuição de uma amostra por partes para delimitar a possibilidade de cada evento ocorrer.

## Variância

Vamos supor um estudo com as quantias gastas por um grupo de dez pessoas em uma loja de suvenires (lembranças) de um determinado museu. As informações foram organizadas e lançadas na Tabela 7.1.

**Tabela 7.1:** Consumo individual x valores de suvenires

Consumo individual									
Jorge	Ana	Pedro	Maria	Flávia	Rita	Bruno	Edu	Carla	Marcio
R\$ 15	R\$ 19	R\$ 33	R\$ 28	R\$ 22	R\$ 29	R\$ 18	R\$ 20	R\$ 30	R\$ 26

Baseado no que já vimos, podemos afirmar que a média de gastos foi de R\$ 24,00. Entretanto, essa informação avulsa pode não representar muito, pois, dentre os indivíduos estudados, temos, por exemplo, o Pedro que gastou bem acima da média (R\$ 33,00), isto é, R\$ 9,00 a mais que a média. Na verdade, praticamente 38% a mais que a média. Também, como exemplo, temos o Jorge que gastou menos do que todos (R\$ 15,00). Ele, coincidentemente, gastou R\$ 9,00 a menos que a média, ou os mesmo 38% – só que para baixo.

Deste modo, notem que, em uma amostra com uma quantidade pequena como esta, é até possível analisar individualmente como cada um se comportou em relação à média. Mas vamos imaginar uma amostra com centenas ou milhares de dados. Isso ficaria longo, extenuante e sem sentido! Para tal, poderíamos sugerir calcular, por exemplo, a média das diferenças entre cada consumidor em relação à média dos consumos deles próprios. Organizando as diferenças, teremos os valores contidos na **Tabela 7.2**.

**Tabela 7.2:** Consumo individual x Diferenças individuais, quanto à média da amostra

Consumo individual									
Jorge	Ana	Pedro	Maria	Flávia	Rita	Bruno	Edu	Carla	Marcio
R\$ 15	R\$ 19	R\$ 33	R\$ 28	R\$ 22	R\$ 29	R\$ 18	R\$ 20	R\$ 30	R\$ 26

Diferenças individuais para a média de consumo de R\$ 24,00									
Jorge	Ana	Pedro	Maria	Flávia	Rita	Bruno	Edu	Carla	Marcio
–R\$ 9	–R\$ 5	R\$ 9	R\$ 4	–R\$ 2	R\$ 5	–R\$ 6	–R\$ 4	R\$ 6	R\$ 2

Em vista disto, com as diferenças calculadas, para obtermos a média das diferenças, basta fazer o somatório delas e dividir pelo total de dados. Todavia, iremos nos deparar com um pequeno problema. Qual? É que a soma das diferenças foi zero.

$$\bar{X} = \frac{-9 - 5 + 9 + 4 - 2 + 5 - 6 - 4 + 6 + 2}{10}$$

$$\bar{X} = \frac{0}{10} = 0$$

**Figura 7.1:** Média das diferenças.

Isso se deu porque como a média é uma posição central, dentre os dados estudados, ao somar a diferença desses mesmos dados em relação à média, tenderemos a um resultado nulo. Ora, os valores menores que a média terão diferenças negativas e os valores maiores que a média terão diferenças positivas. Como dito, a média é uma posição central entre eles. Logo, é esperado que eles se anulassem, caso somados. Entretanto, isto não significa que não seja possível calcular uma espécie de média

entre as diferenças dos dados de uma amostra em relação à sua própria média. Para tal, teremos de optar por um recurso matemático que faça os resultados negativos ficarem positivos de forma que, ao somar todas as diferenças, eles não se anulem mais. Existem duas opções: a primeira opção seria multiplicar todas as diferenças por menos um. De fato, os valores negativos ficarão positivos. Contudo, os positivos ficarão negativos. Logo, o problema permanecerá. A segunda opção é elevar todas as diferenças ao quadrado. Assim, os valores negativos ficarão positivos e os positivos assim permanecerão. Vejamos como ficarão os dados após utilizar este recurso na **Tabela 7.3**:

**Tabela 7.3:** Diferenças individuais elevadas ao quadrado

Consumo individual									
Jorge	Ana	Pedro	Maria	Flávia	Rita	Bruno	Edu	Carla	Marcio
R\$ 15	R\$ 19	R\$ 33	R\$ 28	R\$ 22	R\$ 29	R\$ 18	R\$ 20	R\$ 30	R\$ 26

Diferenças individuais para a média de consumo de R\$ 24,00									
Jorge	Ana	Pedro	Maria	Flávia	Rita	Bruno	Edu	Carla	Marcio
-R\$ 9	-R\$ 5	R\$ 9	R\$ 4	-R\$ 2	R\$ 5	-R\$ 6	-R\$ 4	R\$ 6	R\$ 2

Diferenças individuais elevadas ao quadrado									
Jorge	Ana	Pedro	Maria	Flávia	Rita	Bruno	Edu	Carla	Marcio
R\$ 81	R\$ 25	R\$ 81	R\$ 16	R\$ 4	R\$ 25	R\$ 36	R\$ 16	R\$ 36	R\$ 4

Agora, com a eliminação da possibilidade da soma das diferenças ser zero, após utilizado o recurso de elevar os resultados ao quadrado, podemos dar prosseguimento ao suposto cálculo da média das diferenças propriamente dita.

$$\overline{X} = \frac{81 + 25 + 81 + 16 + 4 + 25 + 36 + 16 + 36 + 4}{10}$$
$$\overline{X} = \frac{324}{10} = 32,4$$

**Figura 7.2:** Suposto cálculo da média das diferenças

Este processo que acabamos de fazer, chamando de suposta média das diferenças, leva o nome de Variância e deve ser feito rigorosamente

nesta ordem: primeiro calculamos a diferença de cada dado da amostra em relação à sua própria média. Posteriormente, elevamos todas as diferenças ao quadrado. Em seguida, fazemos a soma desses novos resultados. Por fim, dividimos pela quantidade de dados menos um. Esta etapa final é a única parte que será ligeiramente diferente do que fizemos e, em algumas aulas à frente, veremos porque dividir pela quantidade de dados menos e um, e não apenas pela quantidade de dados.



A parte do processo no qual somamos o quadrado das diferenças calculadas é também chamado de Soma dos Quadrados. No exemplo utilizado (**Figura 7.2**), podemos afirmar que a Soma dos Quadrados é 324.

A fórmula da variância de uma amostra é representada na **Figura 7.3**. Contudo, note que no primeiro contato iria causar enorme estranheza. Mas agora que sabemos como funciona o processo, talvez não seja tão necessário utilizá-la, basta seguir as etapas citadas.

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

**Figura 7.3:** Fórmula da variância.

Deste modo, se quiséssemos calcular corretamente a variância do exemplo da loja de brindes, bastaria apenas “ajustar” o denominador e corrigir a sigla, pois o restante foi feito corretamente. Vejamos na **Figura 7.4**:

$$\begin{aligned}\bar{X} &= \frac{81+25+81+16+4+25+36+16+36+4}{9} \\ \bar{X} &= \frac{324}{9} = 36\end{aligned}$$

**Figura 7.4:** Cálculo da variância.



**Resposta comentada**

O primeiro passo a ser tomado é calcular a média dessa amostra. A soma de dias é 138 e uma vez dividida pela quantidade de processos (30), teremos a média de 4,6. Agora podemos calcular a diferença individual de cada processo em relação à média da amostra, conforme a próxima Tabela.

Processo	Dif.	Processo	Dif.	Processo	Dif.	Processo	Dif.	Processo	Dif.
F193165	-0,6	F654552	-3,6	F678583	3,4	F633606	-0,6	F698539	-1,6
F567283	4,4	F657114	-1,6	F684168	2,4	F647098	3,4	F719103	-1,6
F594338	5,4	F658227	-3,6	F688348	2,4	F653719	-1,6	F721890	0,4
F608530	5,4	F660190	0,4	F688441	-1,6	F676810	-2,6	F722799	-0,6
F623716	-3,6	F668504	-0,6	F692244	-2,6	F673885	0,4	F672149	4,4
F629219	-1,6	F671953	4,4	F696574	-3,6	F729738	-3,6	F673520	-1,6

Com as diferenças individuais devidamente calculadas, já podemos calcular o quadrado de cada uma e, em seguida, a soma dos quadrados, conforme a tabela apresentada em seguida.

Processo	Dif <sup>2</sup>	Processo	Dif <sup>2</sup>	Processo	Dif <sup>2</sup>	Processo	Dif <sup>2</sup>	Processo	Dif <sup>2</sup>
F193165	0,36	F654552	13	F678583	11,6	F633606	0,36	F698539	2,56
F567283	19,4	F657114	2,56	F684168	5,76	F647098	11,6	F719103	2,56
F594338	29,2	F658227	13	F688348	5,76	F653719	2,56	F721890	0,16
F608530	29,2	F660190	0,16	F688441	2,56	F676810	6,76	F722799	0,36
F623716	13	F668504	0,36	F692244	6,76	F673885	0,16	F672149	19,4
F629219	2,56	F671953	19,4	F696574	13	F729738	13	F673520	2,56
Soma do Quadrados = 249,68									

Por fim, dividindo pela quantidade de dados menos um, finalmente, temos a variância desta amostra que é 8,61. Note que mais uma vez, a Fórmula “tão assustadora” nem foi necessária. Bastou apenas uma sequência de operações matemáticas triviais que, nada mais é do que a fórmula.

## Desvio-Padrão

Anteriormente, vimos como se calcula a variância de uma amostra. Contudo, seu resultado talvez não tenha muito significado, pois ele é resultado de uma soma de quadrados. Isto é: como precisamos utilizar o recurso de elevar as diferenças ao quadrado para eliminar a possibilidade de se anularem, acabamos obtendo um resultado exageradamente alto. Mas nem por isso podemos dizer que a variância não possui utilidade, pois podemos considerá-la como parte de um processo. Deste modo, se optarmos por não pararmos o nosso cálculo ao achar a variância, podemos dar um tratamento melhor a ela e, assim, chegar a um resultado com significado para o estudo estatístico.

Assim, como recorremos a um recurso matemático para evitar a anulação das diferenças, podemos recorrer a outro recurso para minimizar o elevado resultado obtido por conta do primeiro processo. Para tal, iremos obter a raiz quadrada da variância. Com isso, o resultado será bem menor e terá um real significado como veremos a seguir. Mas, antes, é necessário nomear esta nova operação: chama-se Desvio-Padrão a raiz quadrada da variância. Vejamos como ficará a fórmula na **Figura 7.5**:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

**Figura 7.5:** Cálculo da raiz quadrada da variância.

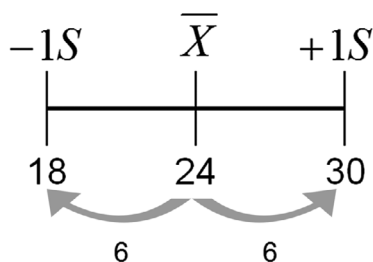
Note que apesar de termos ligeiramente aumentado a Fórmula, na realidade apenas acrescentamos mais uma etapa elementar no processo que desmembramos antes. Logo, se já tínhamos a variância do exemplo da loja de brindes, podemos agora calcular o seu desvio-padrão. Será necessário apenas calcular a raiz quadrada do resultado já conhecido. Logo:

$$S = \sqrt{S^2} = \sqrt{36} = 6$$

**Figura 7.6:** Cálculo do desvio-padrão.

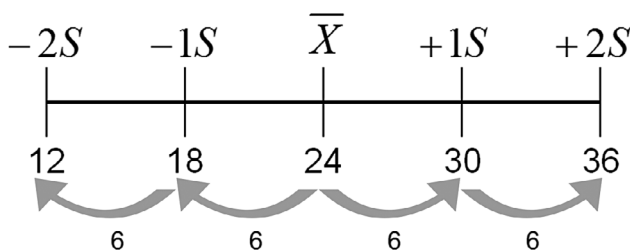
Agora que sabemos que o desvio-padrão do exemplo em questão é 6 e podemos falar sobre seu significado e utilidade. O desvio-padrão será útil para identificar como os dados comportam-se em relação à média dessa mesma amostra. Contudo, antes, é necessário entender o que chamamos de distribuir o desvio-padrão para os dois lados.

Assim, como sabemos, a média é uma medida de posição central. Quando falamos que vamos distribuir um desvio-padrão para cada lado, estamos necessariamente falando de somar um desvio-padrão à média e subtrair um desvio-padrão à média, obtendo agora dois novos pontos à nossa distribuição. Vejamos a **Figura 7.7** que usa os valores do exemplo em questão:



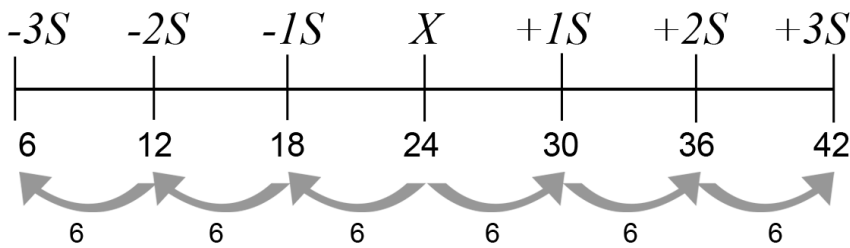
**Figura 7.7:** Distribuição do desvio-padrão a partir da média.

A média de 24 está posicionada ao meio. Após somar um desvio-padrão de 6 à média, obteremos um novo ponto no valor: 30. Isso é chamado intimamente de “jogar um desvio-padrão para a direita”. De forma análoga, subtraindo o mesmo desvio-padrão de 6 da média, temos um novo ponto no valor: 18. É o que chamamos intimamente de “jogar um desvio-padrão para a esquerda”. Dando prosseguimento, faremos o que é chamado de jogar dois desvios-padrão para os dois lados. A **Figura 7.8** irá representar o resultado desta ação:



**Figura 7.8:** Distribuição de dois desvios-padrão a partir da média.

Note que o processo permanece simplório. Basta somar um desvio-padrão ao ponto da extrema direita (+1S) para chegar ao ponto que representa dois desvios-padrão à direita (+2S) e, posteriormente, subtrair um desvio-padrão ao ponto da extrema esquerda (-1S) para chegar ao ponto que representa dois desvios-padrão à esquerda (-2S). Iremos determinar agora os últimos pontos que nos interessam nesta etapa do processo. Vejamos a **Figura 7.9**:



**Figura 7.9:** Distribuição de três desvios-padrão a partir da média



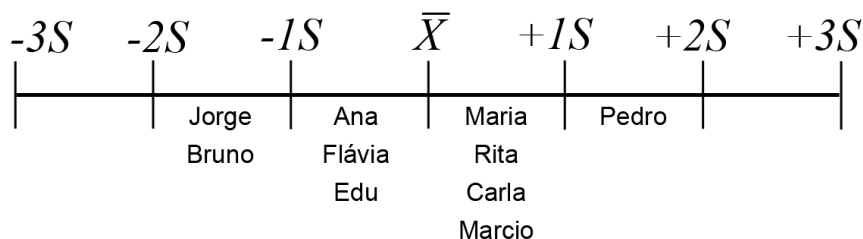
Não existe uma regra fixa para determinar o plural de desvio-padrão. Alguns estudiosos da Língua Portuguesa defendem que é facultativo o plural de palavras que são compostas por dois substantivos. Já outros defendem que o ideal é que o segundo assuma o papel de adjetivo e, assim, somente o primeiro iria para o plural. De qualquer forma, está correto dizer desvios-padrão ou desvios-padrão.

Temos agora os três desvios-padrão “jogados para a direita e para a esquerda”. Com este cenário, já podemos analisar nossa amostra de maneira mais completa, conforme prometemos anteriormente.

Deste modo, ao distribuir os desvios-padrão, temos um cenário que mais bem representa nossa amostra. Isso se dá porque cada “pedaço” desse novo cenário vai representar uma posição da qual a probabilidade de ocorrer varia. Isto é: quando mais próximo este “pedaço” estiver da média, maior a probabilidade deste evento acontecer. Conforme ele se vai se afastando, menor será a probabilidade. Melhor dizendo:

no “pedaço” entre a média e um desvio-padrão (+1S), os resultados ali contidos possuem mais chances de ocorrer do que os contidos entre um desvio-padrão (+1S) e dois desvios-padrão (+2S). De forma simétrica, podemos dizer o mesmo dos pedaços contidos “do outro lado da média”.

É importante destacar que, comumente, trabalhamos com até “três desvios-padrão para cada lado”. O motivo disso é que, como sabemos, quanto mais distante da média, menor a chance de ocorrer. Portanto, além de três desvios-padrão, as chances são tão remotas que sequer vale a pena esmiuçar. Mas veremos isso com mais detalhes na próxima aula. Voltemos, então, ao exemplo da loja de brinde e vamos contabilizar quantos resultados estão contidos em cada pedaço da análise da amostra que acabamos de montar. Vejamos a **Figura 7.10**:



**Figura 7.10:** Resultados contidos em cada pedaço da amostra.

Como dito, quanto mais próximo estivermos da média, maior a chance do evento ocorrer. Logo, mais pessoas terão consumido valores próximos a ela. Conforme nos afastamos e a probabilidade diminuindo, menos pessoas aparecerão. Note que nenhuma pessoa consumiu os dois “pedaços” das extremidades, pois, como dito, as chances de ocorrer são baixas.

Além disto, temos outras leituras para fazer com este novo cenário montado. Podemos notar que a distribuição está dividida igualmente “ao redor” da média. Isto é: cinco pessoas para “cada lado”. Contudo, notará que do lado esquerdo temos três pessoas no primeiro “pedaço” e duas no segundo. Enquanto do lado direito temos quatro no primeiro “pedaço” e apenas uma no segundo. Isso se deu porque os valores maiores que a média estão mais próximos dela, do que os valores menores que a média.

## Atividade 2

*Atende aos objetivos 1 e 2.*

Baseado nos dados da Atividade 1, determine seu desvio-padrão, monte o seu cenário com três desvios-padrão e posicione a quantidade de dados existentes, bem como cada uma das partes desse cenário montado.

---

---

---

---

---

---

---

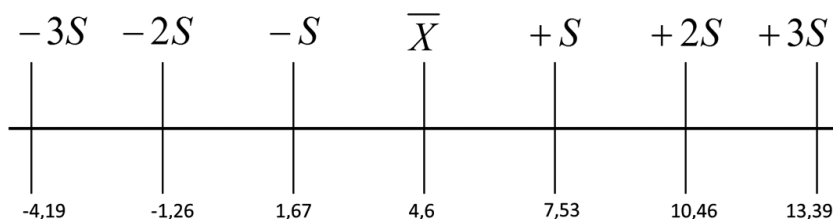
---

---

---

### **Resposta comentada**

Já de posse da variância da Atividade 1 (8,61), basta apenas calcular sua raiz quadrada para chegarmos ao desvio-padrão da amostra em questão. Logo, o resultado do desvio-padrão será 2,93. Em seguida, de posse da média (4,6), montaremos o cenário com as quantidades de dados em cada parte dele:



-4,19 ... -1,26 ... 1,67 ... **4,6** ... 7,53 ... 10,46 ... 13,39

A solução do exercício encerra-se na construção da figura anterior. Contudo, alguns comentários precisam ser feitos. O primeiro é sobre a montagem que, conforme vamos “andando” com os desvios-padrão para a esquerda, notamos que entramos na faixa de números negati-

vos. Ora, como estamos falando de quantidade de dias e sabemos que é inviável um processo demorar quantidade de dias negativos para ser encerrado, isto não faz muito sentido. Entretanto, devemos respeitar a montagem e ir até ao final. O outro comentário é sobre como, à direita da média, temos mais dados distantes dela do que próximo. Isso é uma consequência de uma distribuição irregular, assunto que falaremos em uma próxima aula.



## Coeficiente de variação

De posse da média e do desvio-padrão de uma amostra é possível recorrer a um método de medir a dispersão dessa amostra, que é o coeficiente de variação. Enquanto a média e o desvio-padrão “devolvem” resultados na mesma medida dos dados, isto é, em um estudo sobre idade, a média e o desvio-padrão será em anos, o coeficiente de variação será sempre um resultado percentual independente do tipo de dado estudado. Isto se dá porque ele, basicamente, calcula a dispersão da amostra em relação à média dela. Sua fórmula é:

$$CV = \left( \frac{S}{\bar{X}} \right) \cdot 100\%$$

**Figura 7.11:** Fórmula do coeficiente de dispersão.

Voltemos ao exemplo da loja de brindes. Sua média foi de R\$ 24,00 e o desvio-padrão de R\$ 6,00. Logo, pela fórmula, o coeficiente de variação é de 25%. Este percentual é aceitável, se considerarmos que quanto maior o percentual, maior é a dispersão dos dados. Isto é: quanto maior o percentual, mais espalhados eles estão. Obviamente, mesmo sabendo que o nível de dispersão aumenta conforme aumenta o coeficiente de

variação, percentuais aleatórios talvez não sejam suficientes para uma boa interpretação, ou seja, os 25% calculados anteriormente podem ser considerados baixos, mas quão baixos? Para isso, é sempre comum comparar o coeficiente de variação de duas ou mais amostras relativas para obter um parâmetro mais real.

Vejamos, por exemplo, que em paralelo ao estudo da loja de brindes do museu tenha sido feito um estudo na loja de brindes de outros dois museus na mesma cidade. Suponhamos, então, que no museu concorrente A a média foi de R\$ 29,00 com desvio-padrão de R\$ 8,50. Já no museu concorrente B, a média foi R\$ 23,00 com desvio-padrão de R\$ 1,50. Com isso, os coeficientes de variação desses museus foram respectivamente 29% e 6,5%.

Note, também, que por ter a maior média, a loja de brindes do museu concorrente A seria a mais atrativa de todas. Contudo, com um coeficiente de variação também maior de todos, indica que seus valores estão muito dispersos. Logo, a média alta pode ser um resultado de duas compras individuais altas e o restante abaixo do movimento das outras lojas. Isto é: sem essas duas compras que foram eventuais (com baixa probabilidade de acontecer), talvez a média fosse mais baixa do que as demais e, assim, a loja não fosse a mais atrativa. Portanto, comparando os resultados dos coeficientes de variação, temos que o da loja do museu concorrente B, por ser o menor, é o menos disperso. Isto significa que os valores estão menos “espalhados”. Ora, isso é uma boa característica. Os valores estando próximos à média denotam uma maior chance de um evento acontecer igual a ela. Logo, a previsão fica mais factível.

### Atividade 3

*Atende aos objetivos 1, 2 e 3*

Na Tabela que segue, temos o resultado de duas pesquisas, envolvendo encomendas feitas com dois fornecedores por um mesmo restaurante. Nela constam os números dos pedidos e seus respectivos valores. Determine, pelo coeficiente de variação, qual das amostras é a mais dispersa.

Fornecedor D				Fornecedor E			
Pedido	Valor	Pedido	Valor	Pedido	Valor	Pedido	Valor
D648114	R\$ 2.459	D785370	R\$ 4.071	E071469	R\$ 10.707	E768904	R\$ 4.232
D687757	R\$ 1.432	D874943	R\$ 8.484	E079574	R\$ 4.477	E775132	R\$ 6.495
D716185	R\$ 4.019	D911773	R\$ 5.892	E091421	R\$ 2.564	E022940	R\$ 1.729
D718196	R\$ 6.935	D948328	R\$ 2.280	E115546	R\$ 3.983	E033738	R\$ 3.337
D720723	R\$ 5.300	D952492	R\$ 4.049	E219414	R\$ 3.231	E256888	R\$ 4.016
D734982	R\$ 6.558	D996977	R\$ 7.699	E236030	R\$ 10.787	E315570	R\$ 12.341

This image shows a single sheet of white paper with horizontal blue ruling lines. The lines are evenly spaced and run across the width of the page. There are no margins, text, or other markings on the paper.

**Resposta comentada**

Primeiro teremos de calcular a média de cada amostra. Portanto, temos:

$$\overline{X}_D = \frac{59.178}{12} = 4.931$$

$$\overline{X}_E = \frac{67.899}{12} = 5.658$$

Agora, podemos calcular a diferença de cada pedido, para a sua respectiva média, como primeira etapa do cálculo da variância. A próxima tabela traz esses valores:

Fornecedor D				Fornecedor E			
Pedido	Diferença	Pedido	Diferença	Pedido	Diferença	Pedido	Diferença
D648114	-R\$ 2.473	D785370	-R\$ 860	E071469	R\$ 5.049	E768904	-R\$ 1.426
D687757	-R\$ 3.499	D874943	R\$ 3.552	E079574	-R\$ 1.181	E775132	R\$ 837
D716185	-R\$ 912	D911773	R\$ 961	E091421	-R\$ 3.094	E022940	-R\$ 3.930
D718196	R\$ 2.003	D948328	-R\$ 2.651	E115546	-R\$ 1.675	E033738	-R\$ 2.321
D720723	R\$ 368	D952492	-R\$ 883	E219414	-R\$ 2.427	E256888	-R\$ 1.642
D734982	R\$ 1.626	D996977	R\$ 2.768	E236030	R\$ 5.128	E315570	R\$ 6.682

A próxima etapa, como sabemos, será calcular o quadrado dessas diferenças. Vejamos isto a seguir:

Fornecedor D				Fornecedor E			
Pedido	Diferença²	Pedido	Diferença²	Pedido	Diferença²	Pedido	Diferença²
D648114	R\$ 6.115.280	D785370	R\$ 740.338	E071469	R\$ 25.488.168	E768904	R\$ 2.032.932
D687757	R\$ 12.242.925	D874943	R\$ 12.617.349	E079574	R\$ 1.395.940	E775132	R\$ 699.734
D716185	R\$ 832.162	D911773	R\$ 923.311	E091421	R\$ 9.573.388	E022940	R\$ 15.442.300
D718196	R\$ 4.013.415	D948328	R\$ 7.027.956	E115546	R\$ 2.804.986	E033738	R\$ 5.386.248
D720723	R\$ 135.439	D952492	R\$ 779.405	E219414	R\$ 5.889.645	E256888	R\$ 2.697.015
D734982	R\$ 2.644.497	D996977	R\$ 7.661.164	E236030	R\$ 26.299.367	E315570	R\$ 44.655.149

Basta apenas fazer a soma dos quadrados e dividir o resultado pela quantidade de dados menos um de cada amostra para obtermos as suas respectivas variâncias:

$$S_D^2 = \frac{55.733.241}{11} = 5.066.658$$

$$S_E^2 = \frac{142.364.872}{11} = 12.942.261$$

De posse das variâncias, basta calcular a raiz quadrada de cada uma para obtermos o desvio-padrão de cada amostra:

$$S_D = \sqrt{S_D^2} = \sqrt{5.066.658} = 2.251$$

$$S_E = \sqrt{S_E^2} = \sqrt{12.942.261} = 3.598$$

Por fim, com as médias e desvios-padrão de cada amostra, poderemos, finalmente, calcular os coeficientes de variação:

$$CV_D = \left( \frac{2.251}{4.931} \right) \cdot 100\% = 45,6\%$$

$$CV_E = \left( \frac{3.598}{5.658} \right) \cdot 100\% = 63,6\%$$

Podemos concluir que ambos possuem alto coeficiente de variação. Isto é: ambas as amostras possuem valores muito dispersos e bastante “espalhados”. Ainda assim, podemos dizer que a amostra do fornecedor D é a menos dispersa das duas.

## Conclusão

Quanto maior a necessidade de se ter um estudo estatístico com resultados fidedignos ou, pelo menos, factível ao cenário analisado, mais complexo ele terá de ser. Para tal, o uso de diversos recursos estatísticos faz-se necessário. Vimos na aula anterior que a média, a mediana e o quartil são importantes para tirarmos conclusões sobre uma amostra. Contudo, sozinhos não teriam o mesmo impacto de um resultado obtido também com outros instrumentos estatísticos – como as apresentadas hoje.

O desvio-padrão, de fato, é uma ferramenta poderosa para analisar uma amostra. Contudo, também sozinho perderá toda sua importância. Todavia, quando acompanhado da média forma uma dupla muito eficiente na Estatística. Engana-se quem acha que a média é apenas uma parte do processo de se obter o desvio-padrão. Apesar de, sim, fazer

parte do processo, ao final, ela complementa o desvio-padrão para que obtenhamos melhores conclusões sobre o estudo. Em contrapartida, a variância é apenas uma parte do processo para obtenção do desvio-padrão. O seu resultado não terá tanta importância para estudo simplesmente porque ele pode ser melhorado, chegando-se ao desvio-padrão.

Deste modo, com o binômio média/desvio-padrão podemos melhor estruturar o posicionamento dos dados de uma amostra. Com esta estratégia de montar a amostra em partes, poderemos definir com mais clareza a possibilidade dos eventos acontecerem, tirar prognósticos com alguma precisão e, por consequência, melhor estruturar nossos negócios com expectativa de movimento, gastos, receita etc.

### Atividade final

*Atende aos objetivos 1, 2 e 3.*

Dois aplicadores do mercado de ações resolveram comprar as suas carteiras de títulos, mostrando apenas em qual empresa investe, a sigla desta empresa na Bolsa de Valores e a cotação atual de cada uma, conforme a tabela que segue:

Investidor 1			Investidor 2		
Nome	Código	Cotação	Nome	Código	Cotação
Providência	PRVI3	R\$ 6,00	Usiminas	USIM6	R\$ 13,00
Grendene	GRND3	R\$ 9,44	Localiza	RENT3	R\$ 34,14
Odontoprev	ODPV3	R\$ 30,46	Eternit	ETER3	R\$ 9,16
Usiminas	USIM3	R\$ 15,76	Lojas Americanas	LAME3	R\$ 16,55
Usiminas	USIM5	R\$ 12,28	Lojas Americanas	LAME4	R\$ 18,09
Cetip	CTIP	R\$ 31,98	Amil	AMIL3	R\$ 19,52
Vale	VALE3	R\$ 42,61	Eletropaulo	ELPL3	R\$ 38,77
Vale	VALE5	R\$ 41,63	Eletropaulo	ELPL4	R\$ 50,00
Klabin	KLBN3	R\$ 8,90	Brasil	BBAS3	R\$ 29,15
Klabin	KLBN4	R\$ 8,81	Ecorodovias	ECOR3	R\$ 15,95

De posse destas informações, faça o que se pede:

a) Monte o cenário do Investidor 1, usando três desvios-padrão.

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

b) Determine quais as faixas de cotação que possuem mais possibilidade de ocorrer na carteira do Investidor 1.

---

---

---

---

---

---

---

---

---

---

---

---

c) Refaça as tarefas a e b, mas agora para o Investidor 2.

---

---

---

---

---

---

---

---

---

---

---

---

This image shows a single sheet of white paper with horizontal blue or grey ruling lines. The lines are evenly spaced and run across the width of the page. There are approximately 20 lines visible. The paper has a slight shadow on the right side, suggesting it's resting on a surface.

d) Compare o coeficiente de variação da carteira dos dois investidores e diga qual possui a amostra mais dispersa:

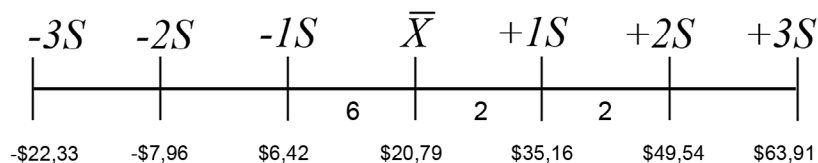
[illegible]

### Resposta comentada

a) Precisamos, inicialmente, calcular a média, a variância e, por fim, o desvio-padrão. Partindo da premissa que o processo do cálculo já está consolidado, vamos aos valores finais:

$$\begin{aligned}\overline{X}_1 &= \frac{207,87}{10} = 20,79 \\ S_1^2 &= \frac{1.859,32}{9} = 206,59 \\ S_1 &= \sqrt{S_1^2} = \sqrt{206,59} = 14,37\end{aligned}$$

Com os dados, podemos montar o cenário da carteira do Investidor 1 com o posicionamento dos dados de acordo com a faixa.



b) Como as faixas próximas à média são as com maiores probabilidades de ocorrerem, podemos dizer que, para a carteira deste investidor, os valores mais prováveis de acontecer são entre \$ 6,42 e \$ 35,16.

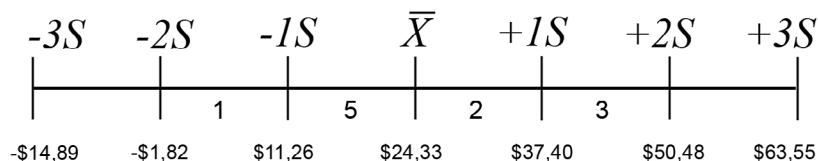
c) Iremos refazer o mesmo processo, mas agora com o Investidor 2. Vejamos:

$$\bar{X}_2 = \frac{244,33}{10} = 24,43$$

$$S_2^2 = \frac{1.538,26}{9} = 170,92$$

$$S_2 = \sqrt{S_2^2} = \sqrt{170,92} = 13,1$$

Novamente, de posse dos dados, iremos montar o cenário.



Com o cenário montado, podemos concluir que, na carteira do Investidor 2, os valores com maiores possibilidades de acontecer estão contidos entre \$ 11,26 e \$ 37,40.

d) Primeiro precisamos calcular os coeficientes de variação de cada carteira:

$$CV_1 = \left( \frac{14,37}{20,79} \right) \cdot 100\% = 69,14\%$$

$$CV_2 = \left( \frac{13,1}{24,33} \right) \cdot 100\% = 53,73\%$$

Deste modo, podemos concluir que estamos falando de duas carteiras que formam amostras bem dispersas, sendo a do Investidor 1 mais dispersa que a do Investidor 2.

---

---

## Resumo

Nesta aula, podemos conhecer mais um excelente instrumento de análise estatística: o desvio-padrão. Junto com ele, podemos notar que apesar de bons instrumentos, nenhum possui eficácia, quando trabalhado sozinho. Deste modo, ficou claro que o uso de vários instrumentos ao mesmo tempo sempre irá valorizar o seu estudo e possibilitar, por conseguinte, que alcance um resultado mais fidedigno e precioso. Em vista disto, foi possível notar que, mesmo sendo calculado através de fórmulas que, supostamente, assustariam no primeiro contato, o desvio-padrão é obtido após uma série de operações elementares em sequência. Isso irá cada vez mais ratificar que a Estatística não é uma ciência de grande complexidade matemática. O seu verdadeiro teor estará sempre presente na parte de interpretação dos dados. O caminho para chegar até esses dados não será tão complicado, quanto se parece.

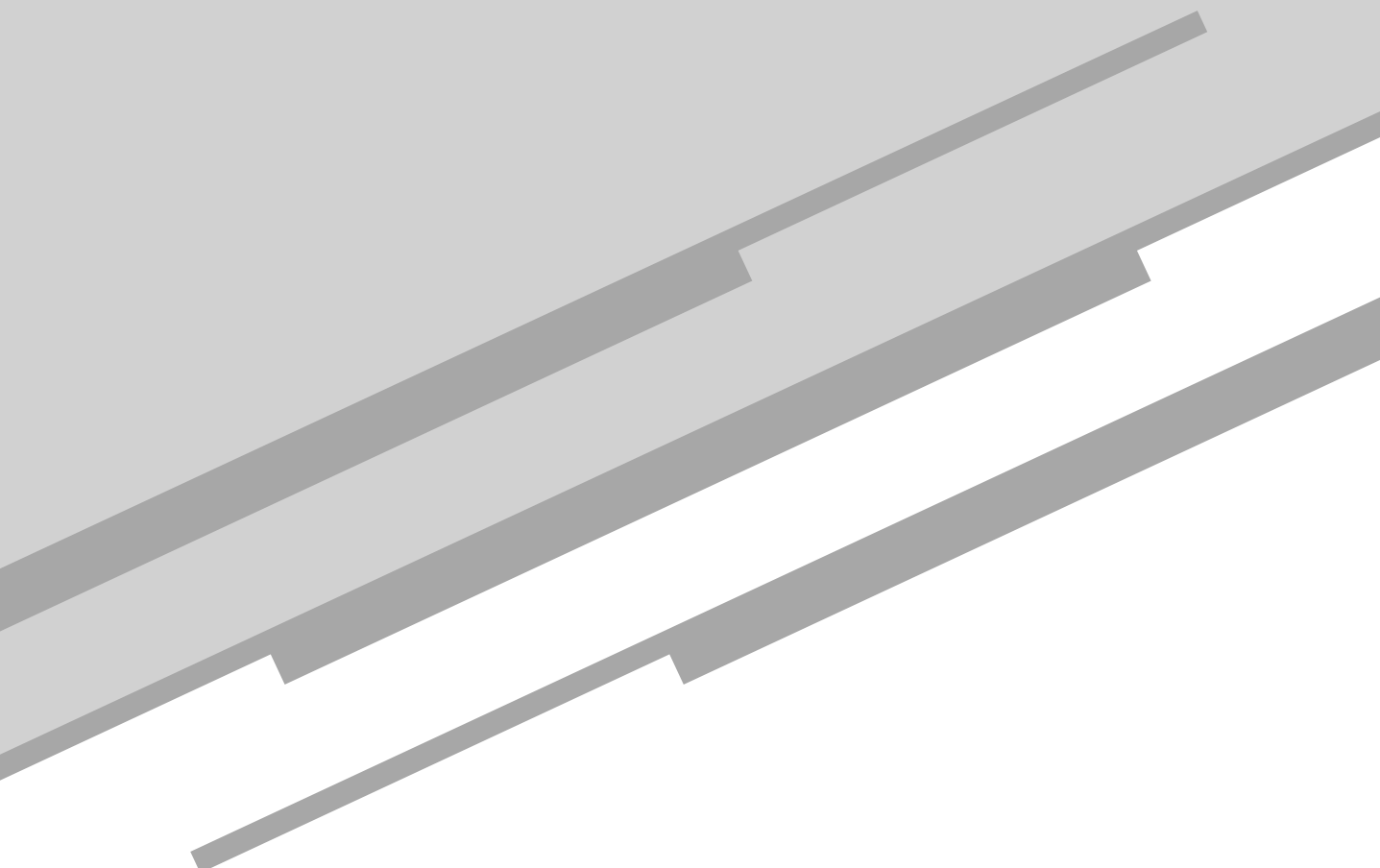
## Informação sobre a próxima aula

Na próxima aula, agora que conseguimos fazer uma boa leitura de uma amostra e determinar como os seus dados comportam-se, iremos, finalmente, dar nome aos padrões que esses dados podem adotar. Iremos focar diretamente no formato da amostra que, junto com os instrumentos apresentados, irão consolidar uma leitura completa do seu estudo. Até lá!



# Aula 8

A onda, o bigode e o russo



*Rafael Canellas Ferrara Garrasino*

## Meta

Distinguir as formas de distribuição de valores.

## Objetivos

Esperamos que, após o estudo desta aula, você seja capaz de:

1. identificar o tipo de formato de uma distribuição de valores;
2. recorrer ao Resumo de Cinco Números para determinar o formato de uma distribuição;
3. fazer uso do *Box-Plot* para ilustrar uma distribuição e interpretar as informações ali contidas;
4. utilizar a Regra Empírica para determinar a quantidade de dados compreendidos em um intervalo de dados;
5. utilizar a Regra de Chebyshev para mensurar a quantidade de dados compreendidos em um intervalo de dados.

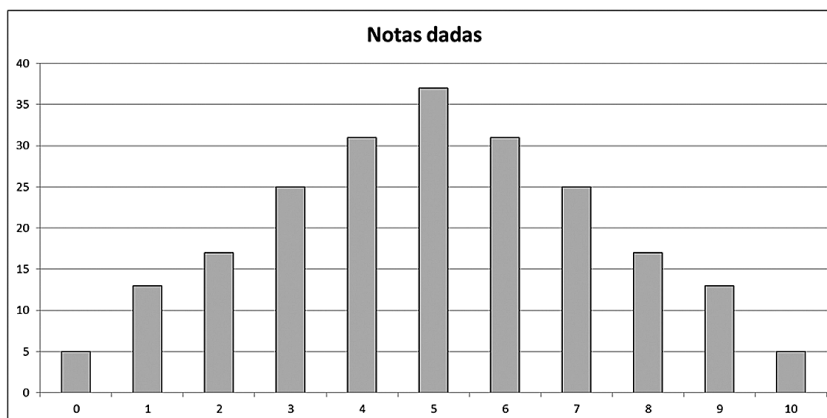
## Introdução

Suponhamos uma pesquisa na qual cada pessoa deveria dar uma nota variando de 0 (zero) até 10 (dez), apenas usando números inteiros, para avaliar o serviço de determinado restaurante. Após a coleta das informações, organizadas de forma crescente e estruturadas em uma tabela de frequência, chegaram ao seguinte resultado, conforme a **Tabela 8.1**.

**Tabela 8.1:** Avaliação de serviço de restaurante / Estrutura em frequência

Intervalo	Qtde.	Qtde. acumulada	Freq.	Freq. acumulada
De 0 até 2,5	35	35	15,98%	15,98%
De 2,5 até 5	93	128	42,47%	58,45%
De 5 até 7,5	56	184	25,57%	84,02%
De 7,5 até 10	35	219	15,98%	100%
<b>Total</b>	<b>219</b>	<b>-</b>	<b>100%</b>	<b>-</b>

É possível notar que, com as informações agora disponíveis, não será viável tirar alguma conclusão dessa distribuição. Podemos, sim, dizer que mais da metade das notas são menores que 5; que uma pequena quantidade é superior a 7, e assim por diante. Mas como identificar como esses valores estão se comportando? Para tal, foi feito um gráfico com cada nota e quantidade de vezes citada. Vejamos na **Figura 8.1**.



**Figura 8.1:** Notas dadas x Quantidade de vezes de cada nota.

Com a ilustração em questão, já podemos notar que a quantidade de *notas* dadas vai aumentando conforme se aproxima da nota 5 e, logo em seguida, começa a decrescer. A quantidade de notas dadas se repete em cada faixa. Perceba que a quantidade de notas 0 dadas é exatamente a mesma de notas 10, assim 1 e 9, 2 e 8, e assim por diante. Logo, sendo assim, repare que apenas com os dados até a nota 5 seria capaz de gerar o mesmo gráfico, bastando apenas “espelhar” o que acontece antes da nota 5 para depois da nota 5, pois basicamente o que temos é uma repetição inversa de cada lado.

Assim, quando o assunto é estatística, a forma como os dados se apresentam é bastante importante e, para tal, comumente iremos nos referir como a **simetria** da distribuição. Na maioria dos casos, inicialmente apenas precisamos saber se estamos falando de uma distribuição simétrica ou assimétrica e, se esse for o nosso foco, estaremos então falando sobre o *formato da distribuição*.

## Simetria

Qualidade de simétrico. Correspondência em tamanho, forma ou arranjo de partes em lados opostos de um plano, seta ou ponto, tendo cada parte de um lado a sua contraparte, em ordem reversa, no outro lado. Proporção correta das partes de um corpo ou de um todo entre si, quanto a tamanho e forma.

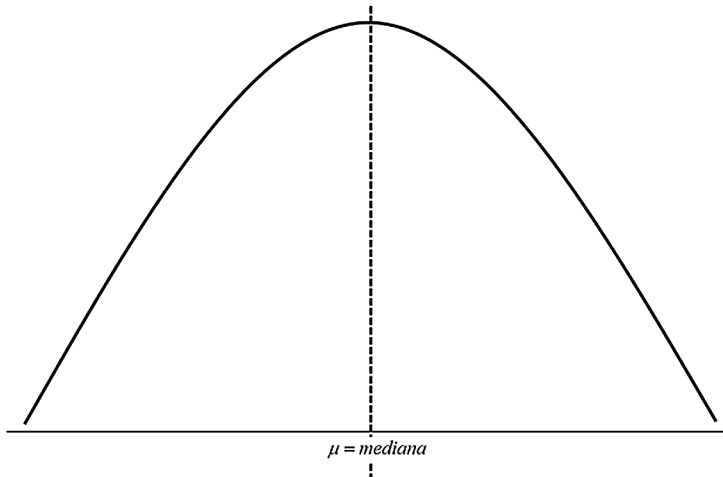
## Formato

Como foi dito, a preocupação com o formato de uma distribuição faz parte do processo estatístico. Esta preocupação, de início, está voltada para um breve resumo: é simétrico ou assimétrico. Contudo, é necessário determinar meios mais práticos e precisos para se chegar a uma conclusão. Imagine fazer o mesmo que foi desenvolvido no exemplo das notas, mas com uma distribuição de mais de 1.000 dados. Seria um trabalho extremamente longo e cansativo. Para isto, optou-se por determinar o formato de uma distribuição utilizando, apenas, suas respectivas mediana e média.

Basta recordar que o principal balizador que utilizamos no exemplo anterior foi a quantidade de dados antes e depois do meio da *distribuição*. Logo, se tivermos a média que determina o ponto médio dentre os valores fornecidos e a mediana que divide todos os dados, independentemente de valor, em duas partes com iguais quantidades, poderemos identificar se estamos falando de uma distribuição simétrica ou assimétrica.

Em vista disso, quando a média e a mediana coincidem sobre um mesmo ponto, podemos afirmar que estamos falando que seu formato é

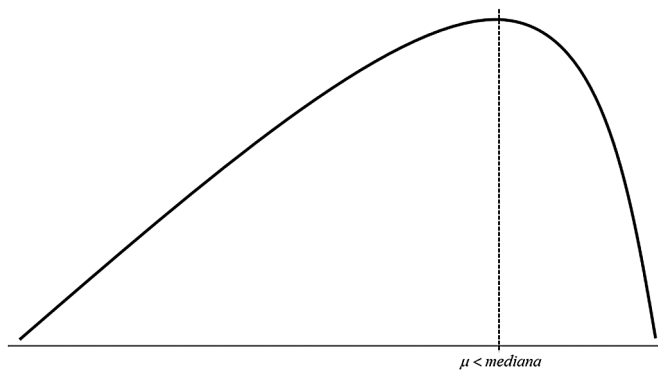
simétrico, isto é, ela estava dividida ao meio não somente em relação à quantidade de dados, mas quanto à incidência de valores. A **Figura 8.2** ilustra um tipo de distribuição simétrica.



**Figura 8.2:** Média e mediana sobre um mesmo ponto: formato simétrico.

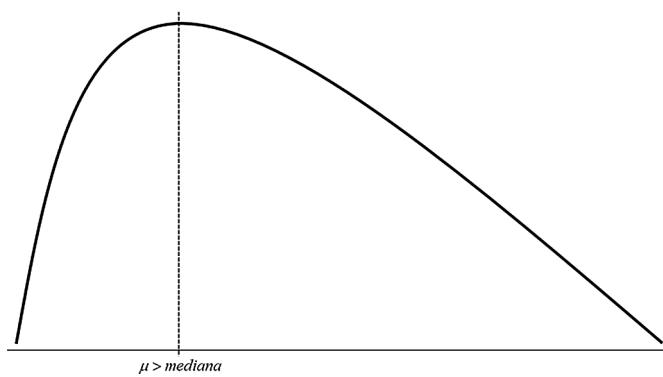
Por padrão, os formatos são ilustrados por curvas, nas quais os pontos mais altos são os de maior incidência de dados. À esquerda, estão os primeiros dados (ou de valores menores) e, à direita, os últimos (ou de maiores valores). Note que, se traçássemos uma linha não retilínea, ligando as extremidades do gráfico – do exemplo das notas –, teríamos um resultado igual ao da **Figura 8.2**. Contudo, nem sempre a média e a mediana coincidem no mesmo ponto. Deste modo, quando isto não ocorrer, teremos um formato assimétrico, que poderá ser à esquerda ou à direita. Este detalhe – de para qual lado o formato terá assimetria – variará de acordo com a relação entre média e mediana.

Nos casos em que a média é inferior à mediana, isto é, está à esquerda dela, trata-se de formatos assimétricos à esquerda. Sendo assim, podemos concluir que os valores menores dessa amostra são muito pequenos em relação aos demais, puxando assim a média para baixo. Dizemos que distribuições com este formato têm uma cauda. Por analogia, fala-se também que é uma onda em deslocamento na direção da sua arrebenção. Vejamos um exemplo na **Figura 8.3**.



**Figura 8.3:** Média inferior à mediana: formato assimétrico à esquerda

De forma análoga, quando a média é maior do que a mediana, os maiores valores da amostra são desproporcionalmente maiores que os demais, puxando, assim, a média para cima. Temos, então, uma distribuição assimétrica à direita. Não se fala de cauda, pois, como sabemos, cauda fica atrás, mas podemos manter a analogia com uma onda que, neste caso, quebrou e seguiu caminho. Vejamos na **Figura 8.4**.



**Figura 8.4:** Média maior que a mediana: formato assimétrico à direita

É importante notar que nem sempre as distribuições terão formas (desenho) exatamente como aqui ilustradas. Essas figuras são apenas exemplos para se ter uma noção de como a *distribuição* está, supostamente, se comportando. Contudo, é possível que possua uma concavidade mais acentuada ou inversa. Ainda assim, o que importa é o *conceito do formato* e como seria um esboço deste.

---

---

**Atividade 1**

---

---

**Atende ao objetivo 1**

A seguir, temos os resultados obtidos em três pesquisas diferentes, nas quais houve a prioridade de se calcular as suas respectivas médias e medianas.

Estudo 1: média 6,5 horas e mediana 5,3 horas.

Estudo 2: média 7,8 horas e mediana 7,8 horas.

Estudo 3: média 7,1 horas e mediana 8,1 horas.

Baseado nessas informações, indique qual das ordens a seguir remete à correta classificação do formato de cada estudo.

- a) Simétrica, assimétrica à esquerda e assimétrica à direita.
- b) Assimétrica à esquerda, assimétrica à direita e simétrica.
- c) Assimétrica à direita, simétrica e assimétrica à esquerda.
- d) Assimétrica à esquerda, simétrica e assimétrica à direita.
- e) Assimétrica à direita, assimétrica à esquerda e simétrica.

**Resposta comentada**

O Estudo 1, por possuir uma média superior à mediana, indica uma assimetria à direita. O Estudo 2 é simétrico, por ter uma igualdade entre média e mediana. Por fim, o Estudo 3, que possui a mediana superior à média, indica um formato oposto ao do Estudo 1; logo, uma assimetria à esquerda.

Portanto, a resposta correta é a opção da letra C.

---

---

---

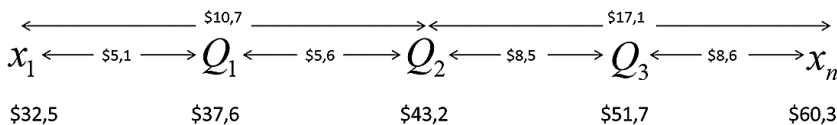
**O Resumo de cinco números**

Como vimos, através da média e da mediana, podemos prever o formato de uma distribuição de dados. Contudo, este tipo de informação pode ser necessário apenas em um momento preliminar. Comumente, conforme vamos esmiuçando os dados de um estudo, mais detalhes passam a ser necessários. Com isso, apenas o formato se torna incompleto.

Já quando a necessidade extrapola o simples formato, o Resumo dos Cinco Números passa a ser uma técnica muito útil por descrever de forma mais completa a mesma *distribuição*. Deste modo, enquanto com o formato temos apenas uma espécie de rascunho da distribuição, com o resumo dos cinco números teremos não somente o formato propriamente dito, mas também a distância entre pontos- chave da amostra.

Assim, como diz o próprio nome da técnica, são necessários cinco pontos. São eles: o menor dado da amostra ( $x_1$ ); o primeiro quartil ( $Q_1$ ); a mediana ( $Q_2$ ), o terceiro quartil ( $Q_3$ ), o maior dos dados ( $x_n$ ). O processo envolve, basicamente, medir a distância entre eles, para determinar a dispersão entre eles e, assim, “traçar” mais detalhadamente o formato da *distribuição*.

Nesse contexto, suponhamos um estudo sobre quanto cada pessoa costuma gastar em deslocamentos, com táxi, durante uma viagem. Após coleta e organização dos dados, os seguintes pontos foram determinados, arrumados e tiveram suas respectivas distâncias calculadas conforme a **Figura 8.5**.



**Figura 8.5:** Estudo de gastos nos deslocamentos por táxi.

Note que, ao compararmos a distância do menor dado até à mediana, com a distância do maior dado também até a mediana, concluímos que a *distribuição* é assimétrica à direita. Contudo, existem outros dados e mais valores para serem utilizados no comparativo, que determina de forma mais completa o formato desta amostra. Para tal, quando se trata do Resumo dos cinco números, a **Tabela 8.2** passa as combinações possíveis de resultados e as conclusões a partir deles.

**Tabela 8.2:** Resumo dos cinco números

Comparativos	Assimétrico à esquerda	Simétrico	Assimétrico à direita
$(Q_2 - x_1) \times (x_n - Q_2)$	$(Q_2 - x_1) > (x_n - Q_2)$	$(Q_2 - x_1) = (x_n - Q_2)$	$(Q_2 - x_1) < (x_n - Q_2)$
$(Q_2 - Q_1) \times (Q_3 - Q_2)$	$(Q_2 - Q_1) > (Q_3 - Q_2)$	$(Q_2 - Q_1) = (Q_3 - Q_2)$	$(Q_2 - Q_1) < (Q_3 - Q_2)$
$(Q_1 - x_1) \times (x_n - Q_3)$	$(Q_1 - x_1) > (x_n - Q_3)$	$(Q_1 - x_1) = (x_n - Q_3)$	$(Q_1 - x_1) < (x_n - Q_3)$

A conclusão imediata que tiramos é a de que, quando as diferenças calculadas possuem o mesmo valor, estamos lidando com uma distribuição simétrica. Contudo, quando as diferenças obtidas à esquerda da mediana são maiores que as obtidas à direita, trata-se de uma distribuição assimétrica à esquerda. Logo, de forma análoga, quando as maiores diferenças são obtidas à direita da mediana, trata-se, então, de uma distribuição assimétrica à direita.

## Atividade 2

### Atende ao objetivo 2

Uma pesquisa levantou a idade com que cada entrevistado conseguiu o primeiro emprego. Após coletadas as informações, elas foram lançadas na tabela que segue:

16	22	18	21	20	17	24	25	17	18
22	19	17	21	17	22	25	26	16	19
19	20	23	21	18	24	27	17	22	16
20	24	16	18	21	19	21	24	18	26

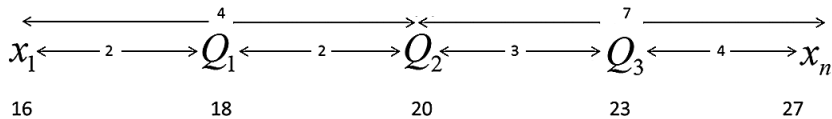
De posse dessas informações e recorrendo ao resumo dos cinco números, determine o formato dessa distribuição.

### Resposta Comentada

O primeiro passo neste exercício deve ser organizar, em ordem crescente, os dados, para que possamos enumerar os cinco números em questão. Ordenando, então, temos a seguinte tabela:

16	16	16	16	17	17	17	17	17	18
18	18	18	18	19	19	19	19	20	20
20	21	21	21	21	21	22	22	22	22
23	24	24	24	24	25	25	26	26	27

Feito isto, já é possível determinar os cinco números e calcular suas respectivas distâncias, conforme podemos ver na próxima figura.



Vamos montar, por fim, uma tabela com o comparativo entre as diferenças e, assim, chegar a uma conclusão final. Vejamos na tabela que segue.

Comparativos	Resultado	Conclusão
$(Q_2 - x_1) \times (x_n - Q_2)$	$(4) < (7)$	Assimétrico à direita
$(Q_2 - Q_1) \times (Q_3 - Q_2)$	$(2) < (3)$	Assimétrico à direita
$(Q_1 - x_1) \times (x_n - Q_3)$	$(2) < (4)$	Assimétrico à direita

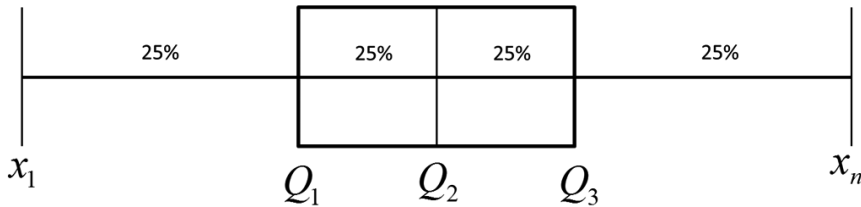
Logo, estamos lidando com uma *distribuição assimétrica à direita*.

### Box-Plot

Assim, como na técnica que envolvia considerar a média e a mediana para determinar o formato, existe uma forma específica de ilustração para o resumo dos cinco números também. O chamado gráfico *Box-Plot* é a ilustração do resultado obtido com o resumo dos cinco números.

Aquelas ilustrações que utilizamos até agora (a **Figura 8.5** e a figura da Resposta Comentada da Atividade 2) tiveram e terão uma utilidade, pois poderão ser utilizadas como rascunho para chegar ao novo resultado final: *Box-Plot*.

A **Figura 8.6** ilustra um gráfico *box-plot* sem valores. O intuito será apenas apresentá-la e compreender como se faz a sua leitura.



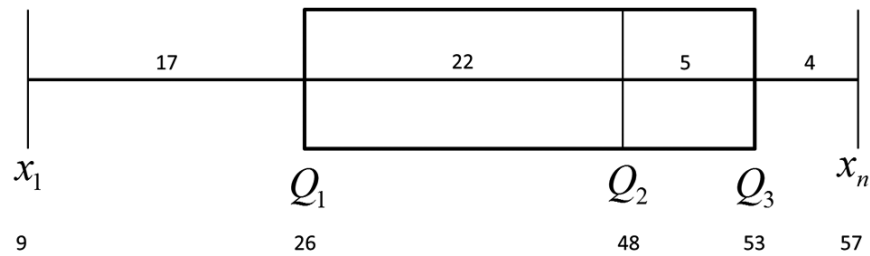
**Figura 8.6:** Ilustração de gráfico *box-plot* sem valores.

À primeira vista, é possível reconhecer alguma semelhança com as outras ilustrações que fizemos anteriormente. Isto foi proposital, para nos acostumarmos com esta estruturação com os cinco pontos e a distância entre eles. De fato, poucas são as diferenças conceituais entre as figuras.

As duas linhas verticais na extremidade da figura indicam, respectivamente, o menor e o maior dado da amostra. O retângulo (*box*), que é algo novo, representa os dados compreendidos entre o primeiro e o terceiro quartil. Ali estão 50% dos dados da distribuição, dados estes que, se analisados separadamente, obterão uma nova média, tendo em vista que os dados extremos seriam descartados, por estarem além do retângulo. As duas laterais do retângulo indicam, respectivamente, o primeiro e o terceiro quartil, assim como a linha vertical que o corta representa a mediana.

Igualmente como estávamos fazendo antes, cada pedaço de linha horizontal entre os pontos irá representar a distância entre eles. A novidade, diferentemente de antes, é que conforme esta distância aumenta ou diminui, estes pedaços também sofrerão mudanças para acompanhar, proporcionalmente, os valores que eles representam.

Na **Figura 8.6**, tivemos um caso de distribuição simétrica. Vejamos, agora, na Figura 8.7, um caso de distribuição assimétrica à esquerda, para melhor identificarmos como o gráfico *box-plot* se comportará com a mudança de valores.



**Figura 8.7:** Gráfico box-plot de distribuição assimétrica à esquerda.

Note que, conforme aumenta a distância entre os pontos, mais longe um estará do outro e vice-versa. Isto facilitará na parte de identificar o formato. Com a figura nitidamente “esticada” para a esquerda, facilita a interpretação de ser uma distribuição assimétrica à esquerda. Isto fica bem claro com o formato do próprio retângulo, que tem um lado muito maior em relação ao outro, quando tomamos a linha da mediana como referência. As extremidades também possuem tamanhos discrepantes entre si, fato que também facilita na identificação do formato. Estas extremidades também são chamadas por bigode (*whisker*).

Deste modo, a criação do gráfico *Box-Plot* se torna importante, pois, além de indicar os pontos e distâncias, ele deixa claro o formato da sua distribuição. Ele pode ser simplesmente desenhado, bastando que se tenha o mínimo cuidado de tentar manter o tamanho das distâncias dentro de uma proporcionalidade com os valores que elas representam, como pode ser gerado por programas (o Excel, por exemplo).

===== **Atividade 3** =====

*Atende ao objetivo 3*

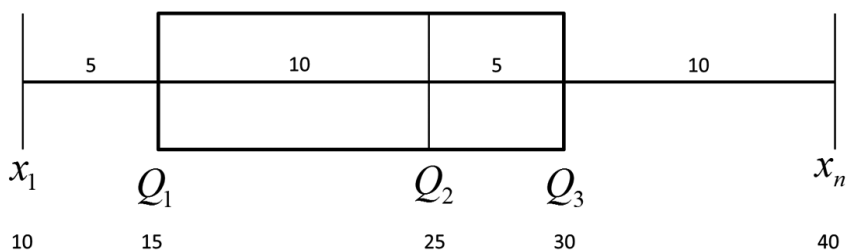
Os dados da tabela a seguir foram coletados em uma pesquisa que questionava até quantos minutos é aceitável um voo atrasar em situações que não envolvam catástrofes da natureza.

30	15	25	20	15	30	10	25	35	40
15	25	30	35	15	25	40	20	25	15
10	35	20	15	40	30	25	20	15	20

Organize os dados e, rascunhando um gráfico *box-plot*, determine o formato dessa distribuição.

### Resposta comentada

O primeiro passo será colocar os dados em ordem crescente. Feito isso, você deverá identificar a posição e o respectivo dado que ocupa essa posição para o menor dado, o primeiro quartil, segundo quartil (mediana), terceiro quartil e maior dado. De posse dessas informações, será possível fazer um rascunho do gráfico *box-plot* como na figura a seguir.



Note que, se nos balizarmos apenas por um dos parâmetros do resumo dos cinco números, podemos tirar conclusões precipitadas. A tabela que segue indica qual a conclusão que seria tomada, caso analisássemos de forma singular o formato da distribuição.

Comparativos	Resultado	Conclusão
$(Q_2 - x_1) \times (x_n - Q_2)$	$(15) = (15)$	Simétrico
$(Q_2 - Q_1) \times (Q_3 - Q_2)$	$(10) > (5)$	Assimétrico à esquerda
$(Q_1 - x_1) \times (x_n - Q_3)$	$(5) < (10)$	Assimétrico à direita

Analisando por cada parâmetro, fica a dúvida, pois cada um indica um formato diferente e, daí, surgiria a dúvida. No caso, o que temos, na realidade, é um caso de simetria, pois, se atentarmos ao detalhe dos valores, a diferença em que um parâmetro indica uma assimetria à direita é a mesma que, sob a ótica de outro parâmetro, indica uma assimetria à esquerda. Com isso, pode-se dizer que existe quase um “empate” entre as assimetrias, o qual deveria ser definido pelo terceiro parâmetro que, obviamente, indicou a simetria como resultado final.

## Posicionamento dos dados

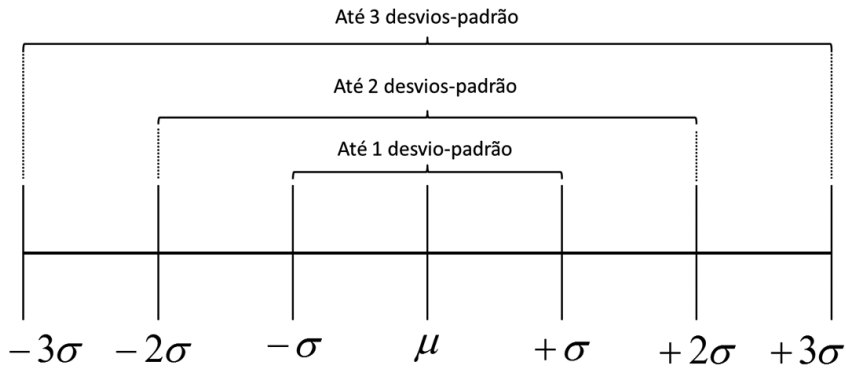
Comumente, em uma distribuição, os dados tendem a ficar agrupados em determinadas partes, isto é, uma grande quantidade fica concentrada em um pedaço da distribuição, outra quantidade fica em outro pedaço, e assim por diante. Normalmente, isso varia de acordo com o formato da distribuição em questão.

## Regra empírica

Conforme já foi visto nesta aula, em distribuições nas quais a mediana e a média estão coincidentes, temos uma distribuição simétrica. Logo, quando se comporta dessa forma, os dados tendem a se concentrar ao redor delas de forma igual, isto é, a quantidade de dados que se encontra a uma distância à esquerda da média é a mesma que se encontra à direita dela.

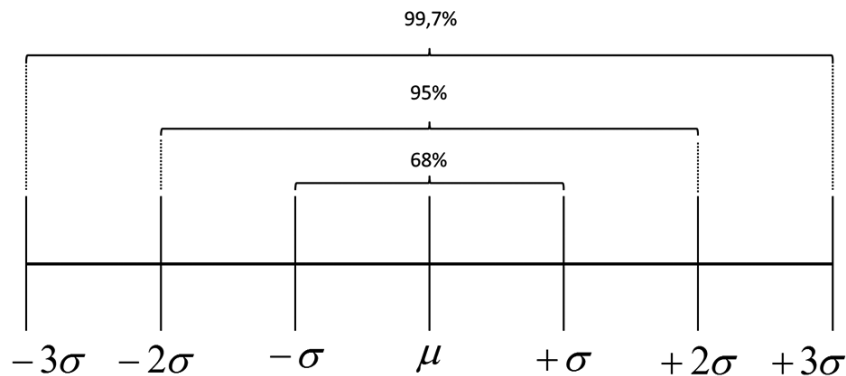
Com base nessa ideia, mas agora utilizando o desvio-padrão como unidade de medida para a distância que determinará os agrupamentos, tem-se a *Regra Empírica*. Ela basicamente distribuirá três desvios-

-padrão para cada lado da média, criando os agrupamentos que serão citados. A **Figura 8.8** ilustra melhor esses agrupamentos.



**Figura 8.8:** Distribuição de três desvios-padrão pela regra empírica.

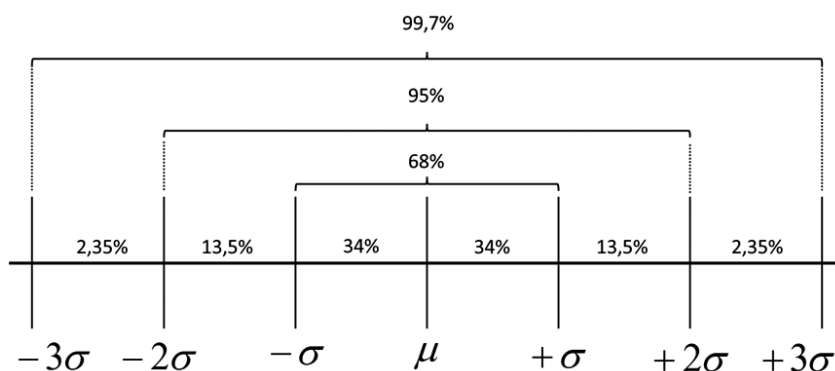
Os agrupamentos ficam, então, divididos em dados compreendidos em até 1 desvio-padrão, até 2 desvios-padrão e em até 3 desvios-padrão. A regra empírica determinará que um percentual do total de dados de uma distribuição tende a ficar compreendido em cada um desses agrupamentos. Vejamos os percentuais na **Figura 8.9**.



**Figura 8.9:** Percentual do total de dados de uma distribuição a partir da regra empírica.

Pela **Figura 8.9**, entende-se que, segundo a regra empírica, 68% do total dos dados possuem um valor que fica entre  $-1$  desvio-padrão da média até  $+1$  desvio-padrão da média. Se aumentarmos mais um desvio-padrão para cada lado deste agrupamento, lidaremos com 95% do total dos dados dessa distribuição. Seguindo com mais um desvio-padrão, têm-se um

agrupamento que consolida 99,7% do total dos dados estudados. Por fim, conclui-se que, pela regra empírica, apenas 0,3% dos dados estão abaixo de  $-3$  desvios-padrão ou acima de  $+3$  desvios-padrão. Vejamos a regra empírica mais detalhada na **Figura 8.10**.



**Figura 8.10:** Percentual do total de dados de uma distribuição a partir da Regra Empírica: estrutura detalhada.

Com essa estruturação detalhada, é possível determinarmos o percentual de dados em uma parte específica de uma distribuição. Além disso, fica claro que conforme mais distante da média, menos dados encontramos. Por conclusão, podemos dizer que é cada vez menor a possibilidade de um dado possuir um valor muito maior ou menor do que a média. Lembrando sempre que estas interpretações consideram o *desvio-padrão* como referência para algo muito distante ou próximo da média.

## Atividade 4

### Atende ao objetivo 4

Uma fábrica controla a qualidade dos seus produtos baseando-se no tempo gasto na fabricação, isto é, quanto mais tempo demorar, maiores são as chances de que o operário responsável o tenha feito com cuidado e atenção. Logo, menores serão as chances de esse produto apresentar algum tipo de defeito. No último estudo feito, o tempo médio de fabricação de cada produto foi de 3,8 minutos com um desvio-padrão de 24 segundos. De acordo com o tempo de fabricação, cada produto recebe uma classificação e, daí, segue para um setor específico.

Os produtos que demoram entre 3,8 e 4,2 minutos são classificados como aprovados, assim como os que demoram entre 3,4 e 3,8 minutos. Produtos cujos tempos sejam menores que 2,6 minutos são classificados como reprovados. Se o tempo de fabricação for entre 4,6 e 5 minutos, o produto é considerado pronto. Demorando menos que 3,4 e mais que 3 minutos, são orientados a serem revistos. Ficando o tempo entre acima de 2,6 e abaixo de 3 minutos, o produto vai para reciclagem. Caso o tempo de fabricação seja superior a 5 minutos, o produto deverá ser analisado. Nos casos em que o tempo de cada produto esteja entre 4,2 e 4,6 minutos, eles serão desmontados.

Sabe-se que o setor gerencial analisa os produtos classificados como *prontos*, *aprovados* e *reprovados*, enquanto o setor de produção se encarrega dos produtos que devem ser vistos, e o setor técnico fica com os demais. Atualmente, a linha de produção trabalha com 20.000 por mês. Desse modo, com quantos produtos o setor técnico, baseado nesses dados, espera lidar por mês?

### Resposta comentada

Estruturando os tempos com os desvios-padrão como balizadores, temos a seguinte organização, conforme a figura que segue.

$$\mu = 3,8$$

$$\mu + \sigma = 4,2; \mu + 2\sigma = 4,6; \mu + 3\sigma = 5,0$$

$$\mu - \sigma = 3,4; \mu - 2\sigma = 3,0; \mu - 3\sigma = 2,6$$

Utilizando o nome dado por cada faixa pela fábrica, organizamos os tempos, como pode ser visto a seguir.

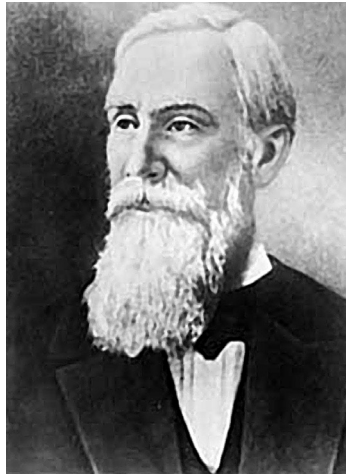
*Rep.*  $\leftarrow -3\sigma \rightarrow$  *Rcl.*  $\leftarrow -2\sigma \rightarrow$  *Rever*  $\leftarrow -\sigma \rightarrow$  *Apr.*  $\leftarrow \mu \rightarrow$  *Apr.*  $\leftarrow \sigma \rightarrow$  *Desm*  $\leftarrow 2\sigma \rightarrow$   
*Pronto*  $\leftarrow 3\sigma \rightarrow$  *Anal*

Comparando cada faixa estipulada pela fábrica, para ficar sob a responsabilidade do setor técnico, com a Regra Empírica, temos que os produtos classificados como Análise, Desmontar e Reciclar somarão 16% do total da distribuição: 3.200 produtos.



### Regra de Chebyshev

De forma similar à regra empírica, a *Regra de Chebyshev* separa a distribuição em agrupamentos balizados pelo desvio-padrão. Contudo, como não trabalha exclusivamente com distribuições simétricas, a regra de Chebyshev não possui valores iguais aos da regra empírica; inclusive, ela não se dá por uma enumeração de percentuais – ela é obtida através de uma fórmula, conforme a **Figura 8.11**.



Pafnuti Lvovitch Tchebychev (em russo: Пафнутий Львович Чебышёв), nascido em Okatowo, circunscrição de Borovsk, perto de Moscou, em 4 de maio de 1821, faleceu em São Petersburgo em 26 de novembro de 1894. Foi matemático. É conhecido por seu trabalho no domínio da probabilidade e estatística. A desigualdade de Tchebychev é utilizada para provar a lei fraca dos grandes números e o Teorema de Bertrand-Tchebychev (1845–1850). Os polinômios de Tchebychev são assim chamados em sua homenagem. Em eletrônica analógica, existe uma família de filtros chamada *Filtros de Tchebychev*.

Fonte: [http://pt.wikipedia.org/wiki/Pafnuti\\_Tchebychev](http://pt.wikipedia.org/wiki/Pafnuti_Tchebychev)

---

$$\left(1 - \frac{1}{k^2}\right) \cdot 100\%$$

**Figura 8.11:** Fórmula de cálculo da regra de Chebyshev.

Pela fórmula, usamos a variável  $k$  para representar entre quantos desvios-padrão estarão contidos os dados que iremos mensurar. Estes valores para  $k$  precisam ser necessariamente positivos e maiores do que 1, pois, sendo iguais a 1, teremos resultado nulo, conforme a **Figura 8.12**.

*Sendo  $k = 1$*

$$\left(1 - \frac{1}{1^2}\right) \cdot 100\% = (1 - 1) \cdot 100\% = 0 \cdot 100\% = 0\%$$

**Figura 8.12:** regra de Chebyshev.

A variável  $K$  precisa ser valor positivo e maior do que 1, para não se ter resultado nulo.

Em vista disto, partindo-se da mesma demonstração, é possível concluir por qual motivo  $k$  não pode ser um valor inferior a 1, pois resultará em um percentual negativo.

## Atividade 5

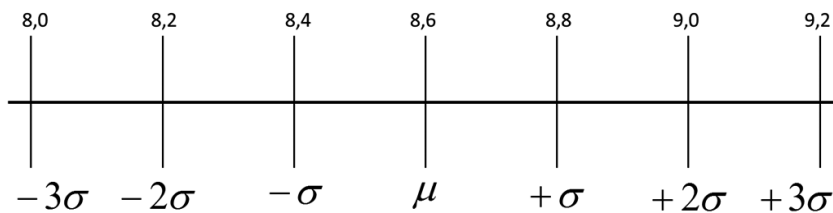
### Atende ao objetivo 5

Certo museu resolveu estipular o tempo que cada visitante demorava percorrendo os corredores das suas principais obras, para que pudesse determinar quantas pessoas poderiam entrar em cada faixa de horário. Após algumas semanas, notou-se que o tempo médio era de 8,6 minutos com 0,2 minutos de desvio-padrão.

Determine qual é o percentual de pessoas que deva demorar menos do que 8 minutos, considerando-se que não foi possível estabelecer se tratava-se de uma distribuição simétrica ou assimétrica.

### Resposta Comentada

Como é desconhecido o formato da distribuição, é recomendado o uso da regra de Chebyshev – já que a Regra Empírica só é utilizada em distribuições simétricas. Para tal, agora precisamos determinar com qual faixa de dados estamos lidando. Vejamos que, na figura que segue, determinam-se os tempos em função dos desvios-padrão:



Note que a relação de distância padronizada entre os desvios-padrão foi utilizada apenas para facilitar o posicionamento das informações. Contudo, como sabemos que não se trata necessariamente de uma distribuição simétrica, talvez não tenhamos quantidades iguais entre certas partes. Ainda sobre a figura, podemos notar que a quantidade de visitantes que interessa, neste momento, ao museu está posicionada na área

imediatamente à esquerda de três desvios-padrão menos a média. Logo, recorrendo à fórmula, concluímos que aproximadamente 88,89% dos dados estão compreendidos entre três desvios-padrão. Portanto, aproximadamente 11,11% estão compreendidos abaixo de 8 minutos ou acima de 9,2 minutos.

Essa parte é a mais crucial da Regra de Chebyshev, quando aplicada em distribuições assimétricas. Concluímos que 11,11% das pessoas demoram menos de 8 minutos ou mais de 9,2 minutos. Se estivéssemos lidando com uma distribuição simétrica, bastava obter a metade desse resultado para garantir que, aproximadamente, 5,56% das pessoas demorariam menos de 8 minutos. Todavia, a distribuição não é simétrica, logo, não podemos garantir que exista igual divisão entre os dados abaixo de 8 minutos e acima de 9,2 minutos. Portanto, o correto é dizer que a quantidade de pessoas que demora menos do que 8 minutos é, aproximadamente, entre 0% e 11,11%.



## Conclusão

Assim como um engenheiro precisa de detalhes técnicos sobre o terreno antes de sobre ele levantar um edifício ou como um médico precisa de informações sobre o atual estado de saúde de um paciente antes de iniciar uma cirurgia, para que um estudo estatístico consiga gerar informações fidedignas, faz-se necessário saber com que tipo de *distribuição* ele está lidando. Isto porque, conforme vimos, de acordo com o tipo de distribuição, teremos uma abordagem diferente.

Essas leituras que podem ser feitas de uma distribuição, conforme seu formato apresentado, irão colaborar consideravelmente para outras técnicas que, mais à frente, serão vistas, assim como da mesma forma que algumas técnicas que vimos antes desta aula foram úteis, concedendo dados para determinar o formato e/ou para montar os cenários a estudar.

A leitura de uma distribuição de dados, de acordo com uma das regras dissertadas, é de extrema importância para que seja possível detectar onde está a maior parte dos dados. Conforme é possível detectar a maior incidência de dados, também é possível detectar qual resultado

dentre aquele estudo tem maior possibilidade de ocorrer. Este tipo de informação é de máxima importância, se quisermos montar prognósticos baseados em uma distribuição de dados.

Das regras citadas, fica a recordação de que a *Regra de Chebyshev* não exige um formato específico. Logo, pode ser aplicada em qualquer situação, mas, em contrapartida, como não possui uma igual divisão entre as partes, não nos retorna muita precisão. Do outro lado, temos a *Regra Empírica*, que é exigente no que se refere onde será aplicada (somente formatos simétricos). Contudo, compensa por ser mais precisa e nos dar a flexibilidade de determinar o percentual de dados entre intervalos divididos ao mesmo.

### ===== **Atividade final** =====

*Atende aos objetivos 1, 2, 3, 4 e 5*

Uma das etapas no momento de inscrição em um curso preparatório consiste em fazer uma prova de nivelamento para que, de acordo com a nota, o aluno seja realocado em uma turma específica. Com a finalidade de não criar grandes desigualdades entre alunos de uma mesma turma, a direção estabeleceu que os alunos que obtivessem nota inferior a 2,0 ficariam na turma de intensidade máxima. Os alunos com nota entre 2,0 e 8,0 ficariam na turma mesclada. Já os alunos com nota superior a 8,0 iriam para a turma de dificuldade máxima.

No último exame de nivelamento, coincidentemente, a média das notas foi 5,0 e o desvio-padrão foi 1,5. Organizando os dados de maneira global, puderam notar que a nota mais baixa foi 0,2 e a mais alta foi 9,8. O primeiro quartil foi representado pela nota 2,8; o terceiro quartil, pela nota 7,2; a mediana, pela nota 5,0.

Em vista do exposto, responda ao que se pede:

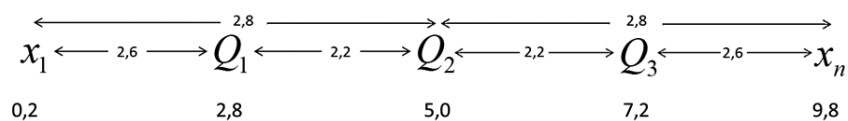
- a) Baseado apenas pela média e pela mediana, trata-se de que tipo de distribuição?
- b) Baseado no resumo dos cinco números, trata-se de que tipo de distribuição?
- c) Faça um esboço do gráfico *box-plot*.

- d) Utilizando a regra empírica, quantos alunos terão na turma mesclada, sabendo-se que 1.000 se inscreveram?
- e) Refaça a letra D, considerando agora a regra de Chebyshev.

[illegible]

### Resposta comentada

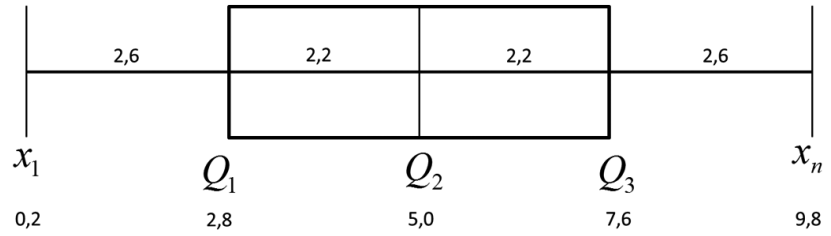
- Como a mediana e a média são iguais, por este critério, podemos dizer que é uma distribuição simétrica.
- Por cautela, é importante organizar os dados para que os cinco números sejam mais bem detectados. Vejamos na figura que segue.



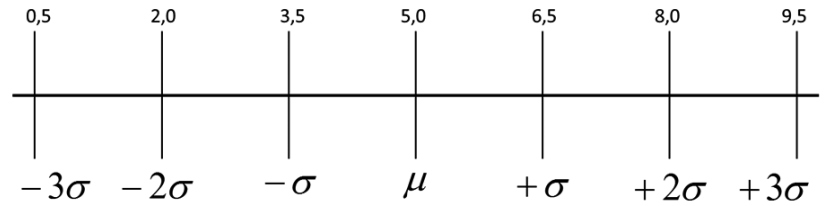
Agora, com os dados visíveis e as distâncias calculadas, podemos, enfim, determinar o tipo de distribuição, conforme na tabela a seguir:

Comparativos	Resultado	Conclusão
$(Q_2 - x_1) \times (x_n - Q_2)$	$(2,8) = (2,8)$	Simétrico
$(Q_2 - Q_1) \times (Q_3 - Q_2)$	$(2,2) = (2,2)$	Simétrico
$(Q_1 - x_1) \times (x_n - Q_3)$	$(2,6) = (2,6)$	Simétrico

c)



d) Primeiro, devemos organizar as notas e as turmas de acordo com a média e o desvio-padrão, conforme a próxima figura.



Agora é possível afirmar que a turma mesclada é composta por alunos que tiveram notas compreendidas entre a média menos dois desvios-padrão e a média mais dois desvios-padrão. Logo, pela Regra Empírica, podemos afirmar que, do total de alunos, 95% ficarão nesta turma ou que dos 1.000, 950 estarão nesta turma.

e) Para a Regra de Chebyshev, vale a mesma arrumação feita na figura anterior; contudo, teremos de recorrer à fórmula para obter o percentual de dados compreendidos entre dois desvios-padrão. Vejamos, então, a próxima figura.

Sendo  $k = 2$

$$\left(1 - \frac{1}{2^2}\right) \cdot 100\% = (1 - 0,25) \cdot 100\% = 75\%$$

Logo, 75% do total de alunos, segundo a Regra de Chebyshev, estarão presentes na turma mesclada, ou, em números, 750 dos 1.000 alunos inscritos estarão na tal turma questionada.

## Resumo

Nesta aula, descobrimos que uma distribuição de dados pode se apresentar de três maneiras distintas quanto ao formato. A chamada *simétrica*, na qual os dados estão distribuídos de igual forma ao redor da média/mediana. As distribuições *assimétricas* são as que possuem uma pequena quantidade de dados de valores extremos em uma das suas pontas, comprometendo a média, deixando-os distante desta.

Pudemos ver, também, que existem duas formas de se determinar o formato de uma distribuição. Embora mesmo sendo diferentes, uma pode complementar a outra, conforme comentado na Atividade 3.

Por fim, fomos apresentados a dois tipos de técnicas para determinar a quantidade de dados presentes em uma partição da distribuição em questão. Uma é mais precisa; contudo, bastante restrita no que se refere à sua aplicação. A outra, de ampla aplicação, mas como visto na Atividade Final, de baixa precisão, quando aplicada em distribuições simétricas.

## Informação sobre a próxima aula

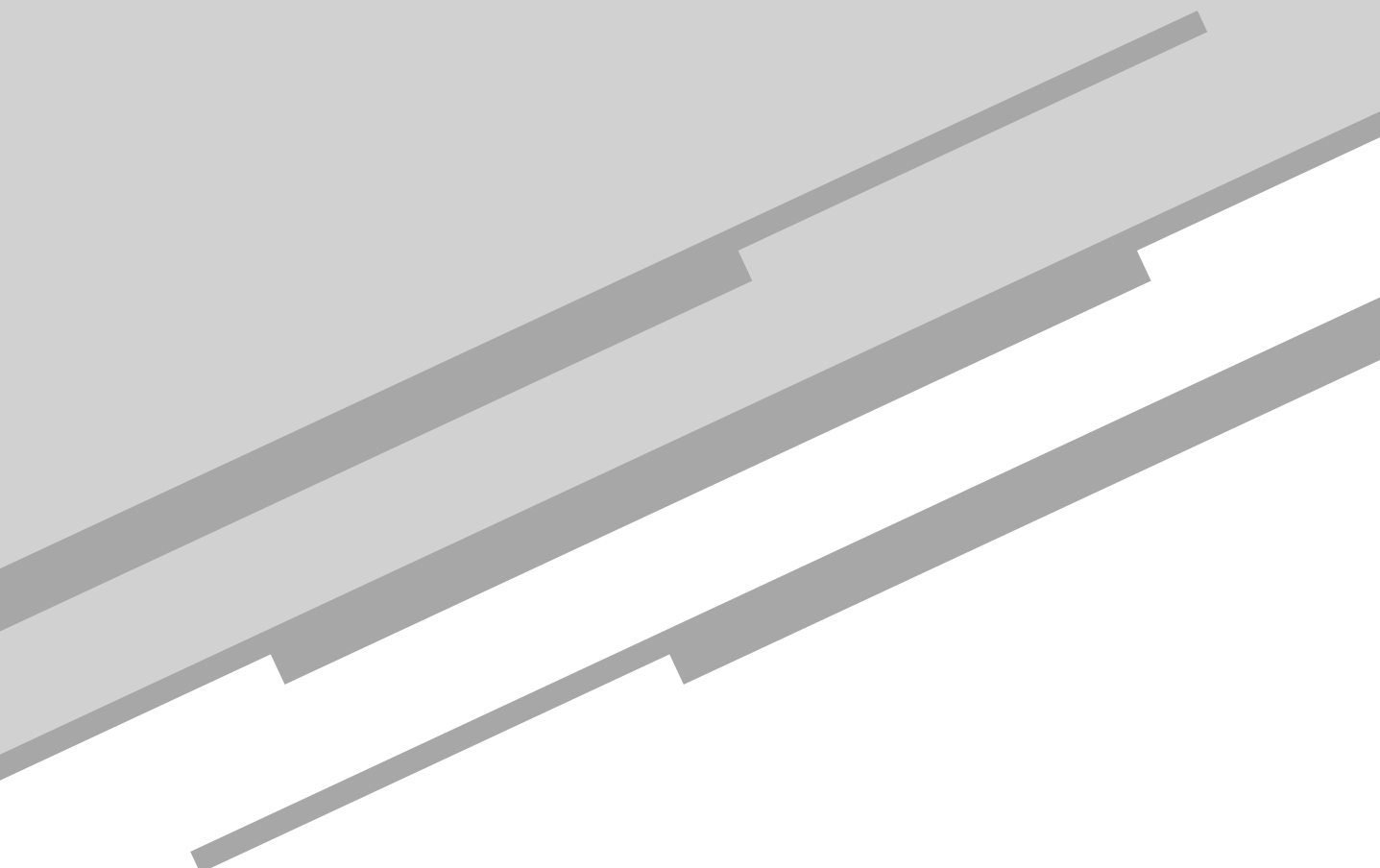
Na próxima aula, veremos que, com a quantidade de informações que temos à nossa disposição, é necessário resumir em algo único e preciso. Passaremos a lidar com um resultado que, na forma de uma espécie de índice, indicará várias informações referentes a um dado e sua relação com a amostra que estamos estudando.

Então, não sai daí!

Mas, se sair, volta logo!

# Aula 9

Z de Zorro



*Rafael Canellas Ferrara Garrasino*

## Meta

Apresentar o Escore Z como ferramenta de avaliação de um dado de uma amostra e em relação a ela mesma.

## Objetivos

Esperamos que, após o estudo desta aula, você seja capaz de:

1. reconhecer o conceito de Escore Z;
2. calcular o Escore Z;
3. analisar se um dado é um *outlier*.

## Introdução

Imagine uma equipe de corrida. Nela, existem alguns corredores que treinam todos os dias com um técnico marcando os tempos. Agora, vamos supor que um dos corredores, em um dia qualquer, diminuiu seu tempo de corrida. De fato, não temos dúvidas de que houve uma melhoria no seu rendimento. Mas como mensurar esta melhoria?



Fonte: [http://pt.wikipedia.org/wiki/Ficheiro:Zorro\\_pencil\\_sketch\\_CVN.jpg](http://pt.wikipedia.org/wiki/Ficheiro:Zorro_pencil_sketch_CVN.jpg) - CVN405

Zorro é um personagem fictício, criado em 1919 pelo escritor Johnston McCulley. Mascarado, com uma capa negra, espada e seu cavalo negro chamado Tornado, ele defendia os fracos e oprimidos na Califórnia durante o Velho Oeste. Sua identidade secreta era Don Diego de La Veja – um membro da aristocracia local. Sua marca registrada, após entrar em ação, era sempre deixar um Z na parede, chão ou qualquer lugar, usando sua espada. O personagem ficou famoso por vários anos após estrelar também em revistas em quadrinhos, desenhos animados e diversos filmes.

Para responder à pergunta anterior, creio que alguém tenha pensado em comparar os tempos e medir a variação percentual entre eles. Este método pode ser um bom ponto de partida, mas ele pode, ao mesmo tempo, significar nada. Afinal, se imaginarmos um aluno que tirou zero na primeira prova, ao tirar zero na segunda, ele pode dizer que dobrou a nota de uma prova para a outra. Assim, sendo menos radical, um aluno que tira 1 (um) em uma prova e 2 (dois) em outra, também pode-se dizer que seu rendimento aumentou em 100%, mas ao mesmo tempo sabemos que a evolução dele foi pífia. Contudo, a mesma variação de 100% que nada significou para estes dois casos, pode ser de grande valor para um aluno que tirou 5 (cinco) na primeira prova e 10 (dez) na segunda. Voltando para os corredores, uma redução de 5% no tempo, mesmo sendo um baixo percentual, pode ter grande importância para o atleta em questão, pois – em corridas – milésimos de segundos são valiosos. Então, como fazer?

Há uma solução mais sofisticada, que é comparar a melhoria deste atleta com os demais, pois assim seria possível determinar se esta melhoria foi de fato significativa ou insignificante. De igual modo, também podemos comparar o novo tempo desse atleta com o tempo dos demais. Daí seria possível dizer se a melhoria dele o fez um dos mais rápidos, se o colocou entre os que estão na média ou sequer fez diferença, pois ainda está longe do rendimento do resto da equipe. Note que esta avaliação se aproxima bastante do que estivemos fazendo nas últimas aulas! Logo, se encaixa melhor no propósito do curso.

Ressalte-se, contudo, que já temos instrumentos suficientes que nos permitem ir além do simples “olhar um tempo”, “olhar o outro”, e dizer se acha que foi bom ou não. Podemos, então, montar uma distribuição com todas as marcações de tempo, obter a média, o desvio-padrão e posicionar o novo tempo dentro da amostra criada. De posse destas informações, podemos afirmar se o novo tempo está acima ou abaixo da média, o quanto acima ou abaixo ele está, e assim por diante. Mas será que podemos tornar mais sofisticada esta análise?

## Escore Z

Sem dúvidas, o uso da média e do desvio-padrão como instrumentos para balizar um dado em relação a uma amostra é algo importante e necessário! Contudo, isto ficaria mais fácil se tivéssemos algo como uma unidade de medida. A verdade é que uma unidade de medida não existe, mas existe um índice de medição que facilita este processo. Chama-se *Escore Z*.

Assim, de uma forma bem sintética, o *Escore Z* resume-se em um valor que indica a “distância” que um dado se encontra em relação à média. Deste modo, como ele considera o desvio-padrão, pode-se dizer que ele indica a “distância” em desvios-padrão que o dado em questão está da média.

Na aula anterior, por alguns momentos, falamos, por exemplo, que tal dado estava a “menos dois desvios-padrão da média” e, assim, concluímos que era um valor menor que a média – logo, estava abaixo dela e a uma distância de dois desvios-padrão. Também falamos que um dado estava a “mais um desvio-padrão da média”. Então, era possível concluir que se tratava de um valor maior do que a média, estando assim acima dela e a uma distância de um desvio-padrão. Agora, toda esta informação estará resumida em um único valor, um número, que será suficiente para tirar estas mesmas conclusões e outras novas. Isto é: temos agora um indicador no qual, em um único valor, estarão resumidas diversas informações.

### **Atividade 1**

#### ***Atende ao objetivo 1***

Descreva qual é a importância do *Escore Z* em uma análise estatística.

---

---

---

---

---

---

---

---

#### ***Resposta comentada***

A importância do *Escore Z*, em uma análise estatística, está em resumir em um único resultado diversas informações importantes, como a posição de um dado em relação à média, a “distância” dele em desvios-padrão da média, dentre outras.

---

---

---

---

## Cálculo do Escore Z

Vamos retornar ao caso dos corredores e supor alguns valores. Imagine que a média do tempo dos corredores, para uma determinada distância, está em 48 segundos e que um deles reduziu seu tempo para 53 segundos. De posse destas duas únicas informações, podemos apenas afirmar que ele está com um tempo acima da média em 5 segundos.

Agora vamos sofisticar este estudo. Suponhamos que o desvio-padrão da distribuição de tempo desses corredores é de 2,5 segundos. Com esta nova informação, podemos elaborar melhor nossa análise. Já é possível afirmar que o novo tempo deste corredor está a dois desvios-padrão acima da média. Portanto, temos neste momento uma conclusão mais completa. O atleta diminuiu seu tempo, mas ainda assim está acima da média e a uma consideravelmente distância de alcançá-la (dois desvios-padrão é razoavelmente distante). Logo, ele precisará melhorar muito, se quiser destacar-se dentro da equipe.

Note, então, que fizemos apenas duas operações matemáticas básicas para obter esta conclusão. Calculamos, por meio de uma subtração, a “distância” do novo tempo até a média e, depois, através de uma divisão, determinamos a quantos desvios-padrão esta “distância” estava da média. É exatamente com estes cálculos simples que determinamos o Escore Z de um dado. Vejamos na **Figura 9.1**:

$$Z = \frac{x - \bar{X}}{\sigma}$$

$\left\{ \begin{array}{l} Z : \text{Escore Z} \\ x : \text{Dado a analisar} \\ \bar{X} : \text{Média da amostra} \\ \sigma : \text{Desvio-padrão da amostra} \end{array} \right.$

**Figura 9.1:** Fórmula de cálculo do Escore Z.

Vamos calcular o Escore Z do atleta citado anteriormente. Temos que o dado a analisar é o novo tempo dele de 53 segundos. A média é 48 segundos e o desvio-padrão de 2,5 segundos. Aplicando estes valores na fórmula, obtemos o resultado 2 (dois). Este resultado significa a “distância” em desvios-padrão da média da amostra e, como é um resultado positivo, conclui-se que está acima da mesma. Logo, pode-se afirmar que o novo tempo está 2 (dois) desvios-padrão acima da média da amostra.

Neste contexto, suponhamos outro atleta que diminuiu seu tempo para 43 segundos. Mantendo os mesmo dados da amostra, ao calcular o Escore Z, o resultado é -2. O resultado ser negativo é uma possibilidade para o Escore Z. Em vista disto, sempre que temos um resultado negativo, sabemos que este dado está abaixo da média. Logo, para este atleta, seu novo tempo está a menos dois desvios-padrão da média. Podemos concluir que ele conseguiu um tempo abaixo da média dos seus colegas de equipe e, por estar em uma distância considerável da mesma, os demais terão de se esforçar muito para alcançar este mesmo feito.

Assim, como calculamos pontualmente o Escore Z para um atleta em questão, o mesmo pode ser feito para uma seleção de dados que for conveniente à sua escolha. Por exemplo, foi feita uma pesquisa com vários grupos de adolescentes que optaram por passar férias na Disney. Dentre várias informações coletadas, a quantidade de dólares destinados para compras foi uma das selecionadas para um estudo. Após análise e cálculos, que já vimos como são feitos nas aulas anteriores, concluíram que a média é de US\$ 3.200,00 com desvio-padrão de US\$ 300,00. De posse destes dados, um grupo de amigos resolveu comparar suas respectivas quantias destinadas para compras. Vejamos os valores deles na **Tabela 9.1**:

**Tabela 9.1:** Turistas x Dólares destinados, individualmente, às compras

Nome	Quantia (US\$)
Jorge Fernando	3.900,00
Maria Fernanda	2.800,00
Pedro Henrique	2.500,00
Ana Paula	4.100,00
João Paulo	3.600,00
Rita de Cássia	3.300,00
Carlos Pedro	3.100,00

O uso do recurso do Escore Z, nesta situação, é muito útil para se ter uma noção se a quantia reservada por cada um deste grupo de amigos foge muito ou pouco da média estudada para esta situação. O cálculo segue o mesmo feito, anteriormente, para o caso dos corredores. Vejamos os resultados na próxima **Tabela 9.2**:

**Tabela 9.2:** Turistas x Dólares destinados, individualmente, às compras x Escore Z

Nome	Quantia (US\$)	Escore Z
Jorge Fernando	3.900,00	2,33
Maria Fernanda	2.800,00	-1,33
Pedro Henrique	2.500,00	-2,33
Ana Paula	4.100,00	3,00
João Paulo	3.600,00	1,33
Rita de Cássia	3.300,00	0,33
Carlos Pedro	3.100,00	-0,33

A primeira conclusão que tiramos é que existe a possibilidade do Escore Z resultar em um valor não inteiro. Isto se dá porque dados de uma amostra não necessariamente estão posicionados à mesma distância que os desvios-padrão estão da média. Deste modo, eles podem estar localizados entre os desvios-padrão. As demais conclusões são relacionadas diretamente aos resultados obtidos.



*Casas decimais:* para o Escore Z de valores não inteiros é necessário apenas duas casas decimais. Mais do que isso é desprezível, conforme poderão ver nas aulas mais à frente. É importante lembrar que, ainda assim, é importante usar o chamado *arredondamento científico* quando o resultado tiver mais de duas casas decimais. Em vista disto, quando a terceira casa decimal for igual ou maior do que 5, arredondaremos para cima a segunda casa decimal. O valor 3,138 arredondando fica 3,14. Quando a terceira casa decimal for menor do que cinco, mantemos a segunda casa decimal como está. O valor 4,781 fica como 4,78.

De igual modo, é possível notar que duas pessoas levaram quantidades de dólares que estão bem próximas à média: Rita de Cássia e Carlos Pedro. A primeira, por possuir um Escore Z positivo, conclui-se que foi um valor acima da média, enquanto o segundo, pelo resultado negativo,

está abaixo da média. Mas ambos estão à mesma “distância” da média, 0,33 desvios-padrão. Isto é: estão a menos de meio desvio-padrão da média; estão muito próximos dela.

O mesmo pode-se dizer de João Paulo e Maria Fernanda. Estão à mesma “distância” da média, sendo que o primeiro acima (Escore Z positivo) e a segunda abaixo (Escore Z negativo). Estão a um pouco mais de um desvio-padrão da média. Seguindo a mesma linha, Jorge Fernando e Pedro Henrique se posicionaram a iguais 2,33 desvios-padrão da média, sendo o primeiro acima (Escore Z positivo) e o segundo abaixo (Escore Z negativo).

Por fim, temos Ana Paula, que está a 3 desvios-padrão acima da média. É a mais distante de todos da média. Recordando a aula em que falamos de média pela primeira vez, podemos afirmar que, por estar muito distante, ela compromete a média, “puxando” a mesma para cima.

## Atividade 2

### Atende ao objetivo 2

Um estudo levantou a quantidade de quartos disponíveis em cada hotel classificado com, no mínimo, três estrelas para receber dirigentes e pessoas importantes durante a Copa do Mundo em uma determinada cidade-sede. Conclui-se que a média é de 58 quartos e desvio-padrão de 9 quartos. Por sua vez, uma rede de hotéis dessa cidade possui cinco unidades que se enquadram no perfil necessário. A unidade Praia Bela possui 71 quartos disponíveis, a unidade Centro Histórico possui 53 quartos, a unidade Zona Leste tem 38 quartos, a unidade Lago das Garças tem 65 quartos e a unidade Mercado Popular tem 46 quartos. Determine o Escore Z de cada unidade e interprete os resultados.

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

### **Resposta comentada**

**Praia Bela:** o Escore Z é igual a 1,44. Está a quase 1,5 desvios-padrão acima da média. É um bom resultado, pois está se destacando dos concorrentes.

**Centro Histórico:** o Escore Z é igual a  $-0,56$  (atenção ao arredondamento). Está abaixo da média, mas bastante próximo dela – quase meio desvio-padrão. Pode ser considerado um resultado razoável.

**Zona Leste:** o Escore Z é igual  $-2,22$ . Está muito abaixo da média – mais de dois desvios-padrão. É um resultado ruim, pois está se destacando negativamente em relação aos concorrentes.

**Lago das Garças:** o Escore Z é igual a 0,78 (atenção ao arredondamento). Relativamente próximo à média, quase um desvio-padrão, mas, por estar acima, pode ser considerado um resultado bom.

**Mercado Popular:** o Escore Z é igual a  $-1,33$ . Resultado abaixo da média e a uma “distância” considerável – quase 1,5 desvios-padrão. É um resultado preocupante, mas menos drástico do que o da unidade Zona Leste.

---

---

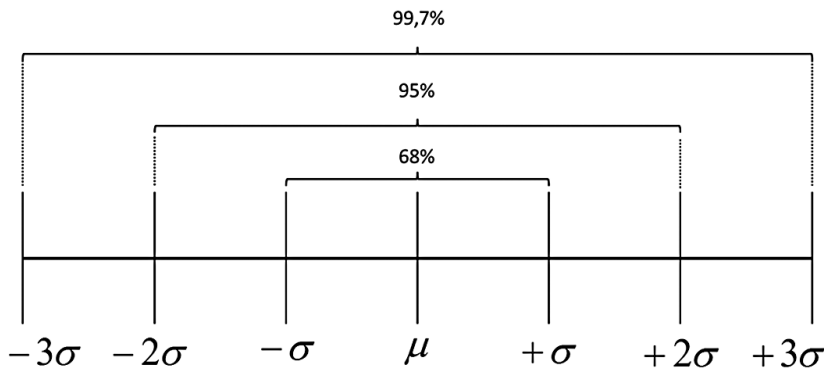
---

---

### **Outlier**

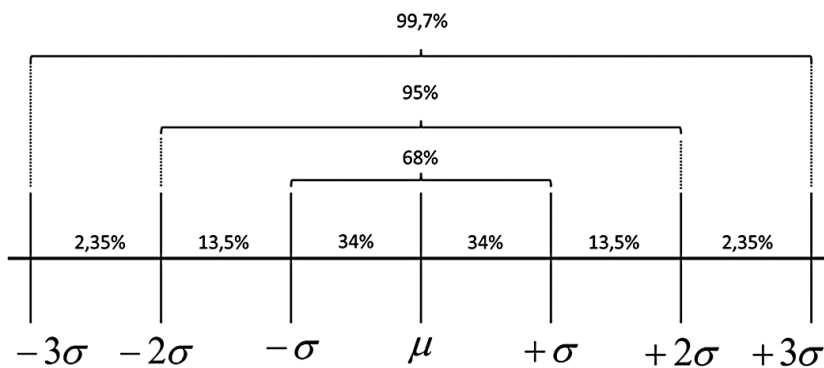
Como vimos, o *Escore Z* resume em um único resultado uma gama de informações importantes. Ainda assim, ele permite que outras conclusões mais bem elaboradas possam ser afirmadas, se recordamos de alguns conceitos já apresentados nas nossas aulas. Vamos recordar da *Regra Empírica*.

Na Aula 8, foi apresentada a Regra Empírica, a qual afirma que, em uma distribuição simétrica, conforme um dado se “afasta” da média em desvios-padrão, é possível prever a probabilidade de este evento ocorrer. A probabilidade de um dado estar contido em um intervalo de um desvio-padrão ao redor da média é de 68%, de até dois desvios-padrão é de 95% e até três desvios-padrão é de 99,7%. A **Figura 9.2** ajudará a recordar deste conceito:



**Figura 9.2:** Distribuição simétrica e a regra empírica.

Foi visto, também, que esta *probabilidade* pode ser desmembrada em cada intervalo delimitado pelos desvios-padrão, conforme a **Figura 9.3**:



**Figura 9.3:** Distribuição simétrica, regra empírica e desmembramento da probabilidade.

De posse desta regra, é possível afirmar que dados que se encontrem acima de mais de três desvios-padrão da média, bem como dados abaixo de menos três desvios-padrão da média possuem baixa possibilidade

de ocorrência. Deste modo, como pode ser visto, a probabilidade de um dado estar abaixo de menos três desvios-padrão é de 0,15%; igualmente para estar acima de mais três desvios-padrão.

Neste sentido, sabemos que um dado posicionado exatamente a três desvios-padrão abaixo da média possui Escore Z igual a -3. Sabemos, também, que um dado posicionado a exatos três desvios-padrão acima da média possui Escore Z igual a 3. Logo, concluímos que a probabilidade de um dado possuir um Escore Z menor ou igual a -3 é de 0,15% e o mesmo para possuir um Escore Z maior ou igual a 3. Portanto, são eventos de baixa probabilidade. Estes eventos serão chamados de *Outlier*.

Em vista disto, agora sabemos que, conforme o Escore Z cresce na direção de 3, menor será a sua probabilidade de ocorrer. Por sua vez, passando de 3, torna-se um *outlier*. De forma análoga, conforme o Escore Z de um dado diminui em direção a -3, menor será a sua probabilidade de ocorrer e, ao ultrapassar -3, também será chamado de *outlier*. Assim, de maneira simétrica, podemos afirmar que quanto mais próximo da média, maior é chance de aquele evento ocorrer. Logo, concluímos que quanto mais próximo de 0 for o Escore Z, maiores são as chances do evento.

Vamos, então, retomar o exemplo da equipe de corredores para reforçar o conceito de *outlier*. Suponhamos que quatro atletas se candidatarão para as vagas disponíveis e cada um fez a sua marcação de tempo para comparar com a média e o desvio padrão já medidos (respectivamente 48 segundos e 2,5 segundos). O primeiro candidato, chamado Juca Tartaruga, marcou 56 segundos. Calculando seu Escore Z, obteremos 3,2. Isto é: ele está a mais de 3 desvios-padrão acima da média. Logo, não somente podemos dizer que ele não está apto para entrar na equipe, como também ele é um evento raro, melhor dizendo, dificilmente teremos candidatos com tempo alto como este. Eventos com Escore Z maiores do que 3, além de raros, comprometem “para cima” a média de forma considerável.

Fernando Mediano foi o segundo candidato e marcou o tempo de exatos 50 segundos. Sendo assim, seu Escore Z é de 0,8. Isto significa que está quase 1 desvio-padrão acima da média. Portanto, seu tempo é um evento de grande probabilidade, isto é, muitos da equipe devem fazer este mesmo tempo. Talvez sua entrada na equipe não traga melhores resultados, além de correr o risco de elevar um pouco a média.

O terceiro candidato chamado Eduardo Acelerado fez o tempo de 42 segundos. Com este tempo, seu Escore Z é de -2,4. Ele está a quase 2,5 desvios-padrão abaixo da média. Logo, um tempo de relativa baixa probabilidade. Poucos fazem este tempo. Ele, com certeza, agregará valor à equipe, pois diminuirá a média, exigindo cada vez mais dos seus colegas.

Por fim, o candidato Usain Bolt marcou o tempo de 38 segundos. Seu Escore Z é de -4. Trata-se de um evento raríssimo estar quatro desvios-padrão abaixo da média. As chances são raras de encontrar outro atleta com este tempo. Sendo assim, se as equipes que competem com esta possuírem uma média e desvio-padrão próximos, a vitória deste atleta será quase sempre certa.

### Atividade 3

#### Atende aos objetivos 2 e 3

A administração de um aeroporto de determinada cidade resolveu recolher o tempo de atraso de todos os voos que de lá decolaram no último mês. Seu objetivo era identificar algum padrão nos atrasos, seja por um motivo em comum, seja pela falta de pontualidade de alguma companhia aérea. Ainda assim, para que este estudo trouxesse à tona apenas atrasos que atendessem ao objetivo inicial, foram desconsiderados os atrasos provocados por tempo ruim ou outras situações adversas das quais nem as companhias nem o aeroporto fossem capazes de interferir. Recolhidas as informações, chegaram à conclusão de que a média de atraso dos voos era de 22,1 minutos; com 3,7 minutos de desvio-padrão. Neste contexto, uma pequena companhia, que opera alguns trajetos de lá, aproveitou estes dados da pesquisa para avaliar seus voos que tiveram atraso na decolagem. O voo 7413 atrasou 18,6 minutos; o voo 6689 atrasou 24,2 minutos; o voo 8463 atrasou 34,3 minutos. Em vista disto, determine o Escore Z destes voos, identifique os *outliers* e tire conclusões a respeito.

---

---

---

---

---

---

---

[illegible]**Resposta comentada**

Antes de calcularmos o Escore Z, é importante destacar que qualquer atraso da companhia deve ser avaliado como algo de elevada importância, para que não ocorra novamente. O que será avaliado aqui é o atraso de cada voo em comparação com os demais voos que decolaram do mesmo aeroporto no período em questão.

**Voo 7413:** o Escore Z é igual a -0,95 (atenção ao arredondamento). O atraso deste voo está a quase 1 desvio-padrão abaixo da média. A probabilidade de ocorrer este evento é considerada alta, mas o fato de estar abaixo da média denota em uma menor gravidade em relação aos voos dos seus concorrentes, por exemplo.

**Voos 6689:** o Escore Z é igual a 0,57 (atenção ao arredondamento). Está pouco mais de meio desvio-padrão acima da média. Há uma grande probabilidade de que outros voos tenham um atraso com tempo próximo a este.

**Voo 8463:** o Escore Z é igual a 3,30 (atenção ao arredondamento). Estamos falando de um atraso que se encontra a mais de 3 desvios-padrão acima da média. É um evento raríssimo, que exige extrema atenção. Este evento é um *outlier*, e a ocorrência dele, mesmo com remotas chances, fez com que a média fosse mais alta.

## Conclusão

Nas análises estatísticas, lidamos com diversos dados, muitos resultados e uma gama de informações que juntas devem culminar em uma decisão. A possibilidade de lidarmos com um índice que represente algumas das diversas informações ao mesmo tempo e nos conduza a uma interpretação precisa é algo indispensável. Para tal, temos agora o Escore Z. Só o fato de o Escore Z permitir tirar tantas conclusões já é uma vantagem. Acrescente a isto o fato de que a obtenção dele é resultado de apenas duas operações básicas, e temos a certeza de que estamos sofisticando o estudo estatístico sem complicar o processo. De igual modo, também ressaltamos que inserimos dentre as opções de análise, já apresentadas, um índice capaz de mensurar o posicionamento de um dado dentro de uma amostra e validar a sua importância para a mesma.

A importância do Escore Z, nos estudos estatísticos, vai muito além do que vimos nesta aula. Mais à frente, nos depararemos com ele novamente, tornando cada vez mais regular a sua utilização. Tantas são as vezes que ele é utilizado, que é comum presenciarmos pessoas que falam apenas Z, quando se referem ao *Escore Z*. Isto se dá porque, na oratória de explicar um estudo ou um cálculo, repetir várias vezes seu nome por completo se torna cansativo. Daí será possível entender por que muitos falam apenas “calcule o Z deste dado”, “consulte o Z dele na tabela tal” e assim por diante.

### Atividade final

#### Atende aos objetivos 2 e 3

Um abatedouro de aves resolveu estudar o peso que cada frango possui no momento do envio para o seu principal cliente. Após coletar todos os pesos, chegou-se a conclusão de que, na média, os frangos eram enviados com 7,1 kg e desvio-padrão de 830 gramas. De posse dessas informações, responda ao que se pede:

- a) Um frango que pese 8,4 kg pode ser considerado um *outlier*?

---

---

---

---

---

---

b) Qual evento tem mais chances de ocorrer: um frango pesar 6,4 kg ou 7,5 kg?

---

---

---

---

---

c) Quais são as faixas de peso que um frango pode ter para ser considerado *outlier*?

---

---

---

---

---

---

---

---

---

---

d) Quanto pesa um frango que tem seu Escore Z igual a -1,62?

---

---

---

---

---

---

### **Resposta comentada**

a) O Escore Z do frango em questão é 1,57, e não pode ser considerado um *outlier*.

b) O Escore Z do frango que pesa 6,4 kg é -0,84 e o Escore Z do frango que pesa 7,5 kg é 0,48. Para esta análise não importa se o Escore Z é positivo ou negativo. O que tem mais chance de ocorrer é o que possui Escore Z mais próximo de 0. Logo, é o frango que pesa 7,5 kg.

c) Para ser considerado *outlier* ele precisa ter Escore Z inferior a -3 ou superior a 3. Utilizando a fórmula para Escore Z igual a -3, conforme a **Figura 9.3**, concluímos que qualquer peso igual ou inferior a 4,61 kg será considerado *Outlier*.

$$Z = \frac{x - \bar{X}}{\sigma} \therefore -3 = \frac{x - 7,1}{0,83} \therefore -2,49 = x - 7,1 \therefore x = 4,61$$

Utilizando agora a fórmula para Escore Z igual a 3, conforme a **Figura 9.4**, concluímos que qualquer peso igual ou maior que 9,59 kg será considerado *outlier*.

$$Z = \frac{x - \bar{X}}{\sigma} \therefore 3 = \frac{x - 7,1}{0,83} \therefore 2,49 = x - 7,1 \therefore x = 9,59$$

Logo, para ser considerado *Outlier*, um frango precisa pesar igual ou menor do que 4,61 kg ou a partir de 9,59 kg.

d) Segundo os cálculos apresentados na Figura 9.5, podemos afirmar que um frango com Escore Z igual a -1,62 pesa aproximadamente 5,755 kg.

$$Z = \frac{x - \bar{X}}{\sigma} \therefore -1,62 = \frac{x - 7,1}{0,83} \therefore -1,3446 = x - 7,1 \therefore x = 5,755$$

---

---

---

---

## Resumo

Nesta aula, notamos que é necessário ter um parâmetro para avaliar um dado em relação a uma amostra e que, com tantas informações que coletamos e calculamos, a necessidade de ter isso de maneira mais ágil e clara é real. A partir de duas operações matemáticas básicas, chegamos a um índice que atenderá a esta necessidade: o Escore Z. Desse índice, podemos avaliar melhor o posicionamento de um dado em relação a uma amostra, a possibilidade de ocorrência de eventos similares a eles, o quanto ele possivelmente influencia a média, dentre outras informações. Esse índice passou a resumir uma quantidade considerável de informações referente ao dado em questão e a relação que ele mantém com sua respectiva amostra.

## Informação sobre a próxima aula

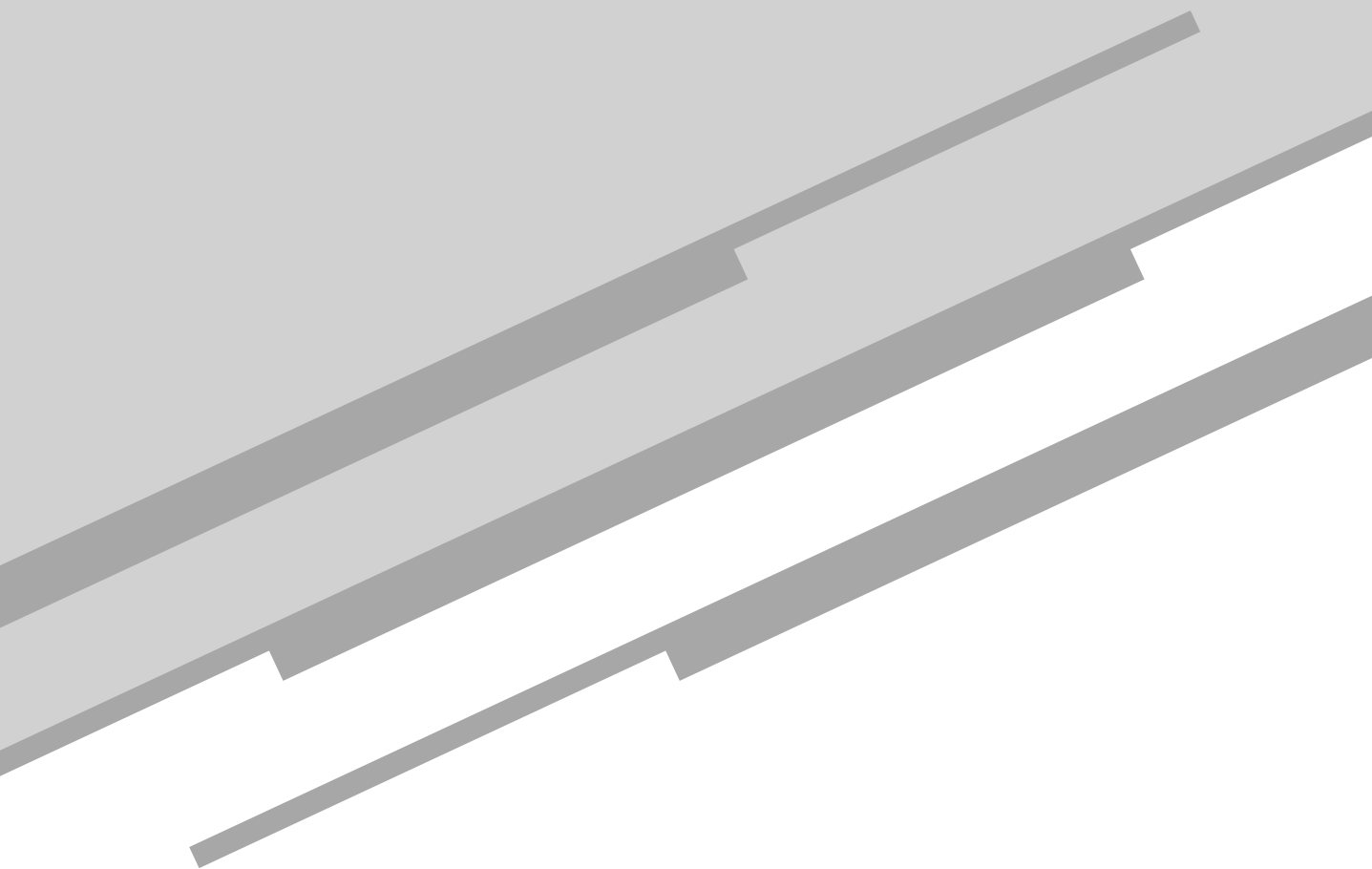
Na próxima aula, analisaremos duas amostras entre si. Verificaremos como uma se comporta em relação à outra. Determinaremos se uma tem influência sobre a outra e que tipo de influência está ocorrendo.

Até lá!



# Aula 10

Vai subir comigo ou vai descer?



*Rafael Canellas Ferrara Garrasino*

## Meta

Apresentar a covariância e o coeficiente de correlação como instrumentos de avaliação do comportamento de duas amostras entre si.

## Objetivos

Esperamos que, após o estudo desta aula, você seja capaz de:

calcular a covariância;

1. interpretar o resultado da covariância;
2. interpretar um gráfico de dispersão;
3. calcular o coeficiente de correlação;
4. interpretar o resultado do coeficiente de correlação.

## Introdução

Vamos supor um estudo que questione quanto cada pessoa gasta em uma viagem. Provavelmente, teremos diversos valores que podem variar, por exemplo, de R\$10,00 a R\$1.000,00 ao dia. Obviamente, um estudo deste tipo está consideravelmente vago. Nestes gastos, podem estar incluídos valores necessários, como a diária do hotel, alimentação, transporte local etc. Logo, passa a ser imprescindível um maior detalhamento, isto é, um foco em uma parte deste gasto.

Estamos, neste momento, com um estudo mais detalhado e coletamos apenas o valor que cada pessoa gasta em diária de hotel em uma determinada cidade. Com esta especificidade, já é possível tirar uma conclusão mais próxima da realidade, pois sabemos que estamos lidando apenas com aquele tipo de custo. É possível, agora, determinar quanto, em média, cada visitante gasta em diárias, o quanto está variando esta amostra, quais respostas dadas podem ser consideradas *outliers* e se estão comprometendo o resultado. Suponhamos que os resultados obtidos sejam os da **Tabela 10.1**.

**Tabela 10.1:** Turistas x Valores gastos em diária de hotel

R\$ 250	R\$ 390	R\$ 320	R\$ 410	R\$ 270
R\$ 360	R\$ 350	R\$ 430	R\$ 280	R\$ 340
R\$ 290	R\$ 330	R\$ 390	R\$ 220	R\$ 310
R\$ 400	R\$ 280	R\$ 360	R\$ 370	R\$ 300



Prezado aluno, fica nesse momento a sugestão para que você calcule a média e o desvio-padrão da amostra da **Tabela 10.1** como exercício de fixação. Ao término, é esperado que encontre uma média de R\$ 332,50 e um desvio-padrão de R\$ 57,297.

Ao analisar apenas os resultados obtidos dessa amostra, podemos tirar diversas conclusões, conforme já estipulamos no parágrafo anterior. Contudo, outras conclusões podem ser tiradas, se considerarmos mais informações além do valor da diária informado. Isto é: será que todas as pessoas que participaram dessa pesquisa estavam hospedadas em um mesmo bairro? Afinal, sabemos que, de acordo com o bairro, o valor da diária de um hotel pode variar consideravelmente. Será que todas estão hospedadas em um mesmo tipo de hotel? Já sabemos que um hotel de cinco estrelas será mais caro que um de duas estrelas, que será mais caro que um simples albergue.

Deste modo, na tentativa, então, de melhorar o que estamos falando e evitar essas dúvidas levantadas, vamos supor que, na realidade, a pesquisa apenas perguntou até quanto a pessoa estava disposta a pagar por uma diária de hotel. Aqui, com raras exceções, eliminamos a possibilidade de diferença de localidade ou tipo de hotel. Assim, estamos falando basicamente de quanto a pessoa está disposta a pagar: uma previsão particular. Se existir hotel cinco estrelas no melhor bairro com uma diária dentro do previsto, com certeza, a pessoa pedirá um quarto por lá. Caso contrário, procurará em outro hotel dentro das suas expectativas.

Em vista disto, dando continuidade ao estudo, suponhamos que, além de quanto a pessoa está disposta a pagar por uma diária, foi perguntada a idade também. Agora é possível traçar um perfil dos entrevistados. Podemos saber, baseados na média, na variação da amostra e nos valores extremos, que tipo de público foi entrevistado. Os supostos valores foram apresentados na **Tabela 10.2**

**Tabela 10.2:** Turistas x Idade

19,3	30	24,7	31,6	20,8
27,7	27	33,1	21,6	26,2
22,4	25,4	30	17	23,9
30,8	21,6	27,7	28,5	23,1



Prezado aluno, fica a sugestão de praticar com os valores da **Tabela 10.2** os mesmos cálculos da **Tabela 10.1**, como indicado anteriormente. Ao final, será encontrada uma média de 25,62 anos e um desvio-padrão de 4,392 anos.

Repare, então, que neste momento temos dois estudos e duas conclusões, ambos obtidos separadamente. Um falando unicamente de valores de diárias e outro sobre a idade dos entrevistados. Contudo, sabemos que ambos os dados foram obtidos em uma mesma pesquisa. Isto é: são duas amostras que, na realidade, foram obtidas juntas. Logo, estamos falando de uma única amostra com duas variáveis (variável valor diária e variável idade). Será que é possível associar estas informações e montar um estudo usando as duas ao mesmo tempo? Vamos, deste modo, ampliar o campo da Estatística, em que você irá exercer seus estudos com análise que envolve duas amostras ao mesmo tempo. Apresentaremos, com isto, os conceitos de covariância e coeficiente de correlação e como podem ser interpretados – seja como um número, seja como uma tomada de decisão.

## Covariância

Estamos agora com uma pesquisa que envolve duas informações. Assim, no vocabulário da Estatística, estamos com uma amostra com duas variáveis. Cada indivíduo entrevistado respondeu sua idade e o quanto está disposto a pagar por uma diária. É importante ressaltar que, quando estamos falando de amostras com mais de uma variável, a ordem das respostas não interfere no resultado. Entretanto, precisamos ser rigorosos com a organização das respostas no que se refere ao entrevistado. Isto é: se Pedro falou que tem 29 anos e pretende gastar até R\$ 400,00, este par de informações deverá sempre estar associado entre si, caso contrário não terá mais validade para o estudo. Mantendo as respostas associadas entre si, isto é, “deixando juntas” as respostas, montamos o perfil de cada entrevistado. Sendo assim, a **Tabela 10.3** apresenta os dados organizados conforme foram colhidos, entrevistado a entrevistado.

**Tabela 10.3:** Turistas x Diária + Idade

Entrv.	1	2	3	4	5	6	7	8	9	10
<b>Diária</b>	R\$ 250	R\$ 390	R\$ 320	R\$ 410	R\$ 270	R\$ 360	R\$ 350	R\$ 430	R\$ 280	R\$ 340
<b>Idade</b>	19,3	30	24,7	31,6	20,8	27,7	27	33,1	21,6	26,2
Entrv.	11	12	13	14	15	16	17	18	19	20
<b>Diária</b>	R\$ 290	R\$ 330	R\$ 390	R\$ 220	R\$ 310	R\$ 400	R\$ 280	R\$ 360	R\$ 370	R\$ 300
<b>Idade</b>	22,4	25,4	30	17	23,9	30,8	21,6	27,7	28,5	23,1

Repare que, agora, se perguntado quais foram as respostas do entrevistado 8, podemos afirmar que foi R\$ 430 de limite para diárias e ele tem 33,1 anos. Assim, é possível, pontualmente, montar o perfil de cada entrevistado. Logo, algumas conclusões podem ser tiradas. Deste modo, apenas observando a **Tabela 10.3**, podemos afirmar que o entrevistado 8 é o mais velho e o disposto a pagar mais caro por uma diária de hotel. Em contrapartida, o entrevistado 14 é o mais novo e o que menos deseja gastar em diária. Podemos, a partir destas duas informações, concluir que conforme a pessoa fica mais velha, mais ela se preocupa com o próprio conforto. Logo, estará mais disposta a pagar por um valor alto de diária? Obviamente, na Estatística, nada se responde apenas “olhando” ou com conclusões tiradas com um par de informações. Portanto, não podemos responder ao questionamento recém-feito. Contudo, é possível estipular um valor que tenha representatividade para interpretarmos a relação entre diária de hotel e idade e, assim, tentar responder à pergunta feita.

Ao associar as duas respostas de cada entrevistado, passamos a ter duas sequências de valores que podem ter uma relação linear entre si. Esta relação pode ser positiva, pode ser negativa, como pode ser inexistente. Chamamos de *Covariância* a medida numérica que mede a força entre esta relação, ou seja, entre duas variáveis de uma mesma amostra. Ela é obtida através da fórmula na **Figura 10.1**:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X - \bar{X})(Y - \bar{Y})}{n - 1}$$

**Figura 10.1:** Fórmula de cálculo da covariância.

Apesar de assustar no primeiro contato, a fórmula da covariância requer apenas as quatro operações básicas e um pouco de atenção. Mas, antes, precisamos saber quais dados serão necessários para este cálculo. Usando na pesquisa que estamos estudamos, precisaremos dos dados sobre diária (será o  $X$ ), a média das diárias (será o  $\bar{X}$ ), os dados sobre idade (será o  $Y$ ), a média das idades (será o  $\bar{Y}$ ) e a quantidade de pessoas entrevistadas (será o  $n$ ). Devemos destacar que não existe uma obrigatoriedade para quem será  $X$  e quem será  $Y$ . É obrigatório, apenas, que mantenham dados e amostras com a mesma letra. Mantendo a organização das respostas entre si e, agora, com as letras determinadas, temos a seguinte organização na **Tabela 10.4**:

**Tabela 10.4:** Turistas x Dados sobre diária e idade

Entrv	X	Y	Entrv	X	Y
1	R\$ 250	19,3	11	R\$ 290	22,4
2	R\$ 390	30	12	R\$ 330	25,4
3	R\$ 320	24,7	13	R\$ 390	30
4	R\$ 410	31,6	14	R\$ 220	17
5	R\$ 270	20,8	15	R\$ 310	23,9
6	R\$ 360	27,7	16	R\$ 400	30,8
7	R\$ 350	27	17	R\$ 280	21,6
8	R\$ 430	33,1	18	R\$ 360	27,7
9	R\$ 280	21,6	19	R\$ 370	28,5
10	R\$ 340	26,2	20	R\$ 300	23,1

Com os dados organizados de acordo com a letra que adotamos, já é possível iniciar o processo de cálculo da *covariância*. A primeira etapa será subtrair todos os dados de diária de hotel ( $X$ ) pela sua respectiva média ( $\bar{X}$ ) e subtrair todos os dados de idade ( $Y$ ) pela sua respectiva média ( $\bar{Y}$ ). Os resultados estão na **Tabela 10.5**:

**Tabela 10.5:** Turistas x Organização de dados para cálculo da covariância

Média diárias			R\$ 332,50		Média idades			25,62	
Entrv	X	$X - \bar{X}$	Y	$Y - \bar{Y}$	Entrv	X	$X - \bar{X}$	Y	$Y - \bar{Y}$
1	R\$ 250	-R\$ 83	19,3	-6,32	11	R\$ 290	-R\$ 43	22,4	-3,22
2	R\$ 390	R\$ 58	30	4,38	12	R\$ 330	-R\$ 3	25,4	-0,22
3	R\$ 320	-R\$ 13	24,7	-0,92	13	R\$ 390	R\$ 58	30	4,38
4	R\$ 410	R\$ 78	31,6	5,98	14	R\$ 220	-R\$ 113	17	-8,62
5	R\$ 270	-R\$ 63	20,8	-4,82	15	R\$ 310	-R\$ 23	23,9	-1,72
6	R\$ 360	R\$ 28	27,7	2,08	16	R\$ 400	R\$ 68	30,8	5,18
7	R\$ 350	R\$ 18	27	1,38	17	R\$ 280	-R\$ 53	21,6	-4,02
8	R\$ 430	R\$ 98	33,1	7,48	18	R\$ 360	R\$ 28	27,7	2,08
9	R\$ 280	-R\$ 53	21,6	-4,02	19	R\$ 370	R\$ 38	28,5	2,88
10	R\$ 340	R\$ 8	26,2	0,58	20	R\$ 300	-R\$ 33	23,1	-2,52

Cada entrevistado possui agora dois resultados obtidos. O entrevistado 1 possui o resultado -R\$ 83 (fruto da subtração da sua resposta R\$ 250 pela média das diárias R\$ 332,5) e o resultado -6,32 (fruto da subtração da sua idade 19,3 pela média das idades 25,62). Devemos, neste momento, multiplicar os dois resultados de cada entrevistado  $((X - \bar{X})(Y - \bar{Y}))$ , conforme a **Tabela 10.6**:

**Tabela 10.6:** Turistas x  $((X - \bar{X})(Y - \bar{Y}))$ 

Média diárias			R\$ 332,50			Média idades			25,62		
Entrv	X	$X - \bar{X}$	Y	$Y - \bar{Y}$	Multp.	Entrv	X	$X - \bar{X}$	Y	$Y - \bar{Y}$	Multp.
1	R\$ 250	-R\$ 83	19,3	-6,32	521,40	11	R\$ 290	-R\$ 43	22,4	-3,22	136,85
2	R\$ 390	R\$ 58	30	4,38	251,85	12	R\$ 330	-R\$ 3	25,4	-0,22	0,55
3	R\$ 320	-R\$ 13	24,7	-0,92	11,50	13	R\$ 390	R\$ 58	30	4,38	251,85
4	R\$ 410	R\$ 78	31,6	5,98	463,45	14	R\$ 220	-R\$ 113	17	-8,62	969,75
5	R\$ 270	-R\$ 63	20,8	-4,82	301,25	15	R\$ 310	-R\$ 23	23,9	-1,72	38,70
6	R\$ 360	R\$ 28	27,7	2,08	57,20	16	R\$ 400	R\$ 68	30,8	5,18	349,65
7	R\$ 350	R\$ 18	27	1,38	24,15	17	R\$ 280	-R\$ 53	21,6	-4,02	211,05
8	R\$ 430	R\$ 98	33,1	7,48	729,30	18	R\$ 360	R\$ 28	27,7	2,08	57,20
9	R\$ 280	-R\$ 53	21,6	-4,02	211,05	19	R\$ 370	R\$ 38	28,5	2,88	108,00
10	R\$ 340	R\$ 8	26,2	0,58	4,35	20	R\$ 300	-R\$ 33	23,1	-2,52	81,90
Somatório da coluna					2.575,50	Somatório da coluna					2.205,50

Por fim, é feito o somatório dessas multiplicações (em uma coluna, obtivemos 2.575,5 e, na outra, 2.205,5), chegando a um resultado único de 4.781 (soma das duas colunas). Este resultado será dividido pela quantidade de pessoas entrevistadas menos um (são 20 entrevistados; logo, dividiremos por 19), obtendo 251,632. Este resultado é a *covariância* da amostra – comparando diárias de hotel com idades. Mas qual é o significado dele?

Como dito anteriormente, a covariância está relacionada com a relação linear entre duas variáveis de uma mesma amostra. Foi dito também que esta relação pode se apresentar de três formas. Uma relação positiva ocorre quando, conforme os valores de uma variável ( $X$ ) da amostra crescem, os valores da outra variável ( $Y$ ) também crescem. Isto é: eles seguem uma tendência a se comportarem de maneira igual. Logo, quando os valores de uma variável diminuem ( $X$ ), os da outra variável ( $Y$ ) também diminuirão. Para casos de relação positiva entre duas variáveis, a Covariância será sempre um valor positivo.

Neste contexto, duas variáveis de uma mesma amostra também podem ter uma relação negativa. Isto é: quando os valores de uma variável ( $X$ ) crescem, os valores da outra variável ( $Y$ ) decrescem. Eles passam a ter uma relação inversa entre si. Nestes casos, a Covariância terá um valor negativo.

Por fim, quando não existe relação alguma entre as variáveis, elas se comportam aleatoriamente e sem um padrão sequer. Isto é: quando alguns valores de uma variável ( $X$ ) crescem, os valores da outra variável ( $Y$ ) decrescem, mas, ainda assim, quando outros valores de uma variável ( $X$ ) crescem, os valores da outra variável ( $Y$ ) também crescem.

Na pesquisa que estamos estudando, o valor da Covariância foi de 251,632: um valor positivo. Logo, uma relação positiva entre as variáveis. A conclusão que podemos tirar é que, para este estudo, conforme a idade do entrevistado aumenta, mais ele está disposto a pagar por uma diária de hotel. Obviamente, não iremos padronizar isso como uma verdade absoluta para todas as pessoas, mas para este estudo, em questão, isto não deixa dúvidas.

## Atividade 1

*Atende aos objetivos 1 e 2*

Foi feita uma pesquisa com algumas pessoas na qual perguntou-se a quantidade de dias em que cada uma ficou na cidade e uma estimativa de horas que gastou, por dia, com atividades culturais. As respostas, após coletadas, estão organizadas a seguir:

Pessoa	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Dias	16	13	15	14	10	8	14	9	10	8	13	13	13	6	11
Horas	6,5	5,5	5,5	5,5	4,5	2,5	3,5	2,5	3	2,5	4,5	4,5	5,5	2,5	3,5

De posse dessas informações, determine a covariância dessa amostra e interprete o resultado.

[illegible]

**Resposta comentada**

A pergunta da questão foi clara: qual é a covariância da amostra? Contudo, ela poderia ser: determine a relação entre as variáveis desta amostra. Implicitamente, estaria perguntando a mesma coisa. Entretanto, não estaria sendo questionada de maneira mais óbvia. Deste modo, o primeiro passo será calcular a média de cada variável. A variável “Dias na cidade” possui média 11,53, enquanto a variável “Horas culturais” possui média 4,13. Agora, iremos subtrair cada dado da variável “Dias” pela sua respectiva média e cada dado da variável “Horas” pela sua respectiva média. O resultado está na tabela a seguir:

Pessoa	Dias	$X - \bar{X}$	Horas	$Y - \bar{Y}$
1	16	4,47	6,5	2,37
2	13	1,47	5,5	1,37
3	15	3,47	5,5	1,37
4	14	2,47	5,5	1,37
5	10	-1,53	4,5	0,37
6	8	-3,53	2,5	-1,63
7	14	2,47	3,5	-0,63
8	9	-2,53	2,5	-1,63
9	10	-1,53	3	-1,13
10	8	-3,53	2,5	-1,63
11	13	1,47	4,5	0,37
12	13	1,47	4,5	0,37
13	13	1,47	5,5	1,37
14	6	-5,53	2,5	-1,63
15	11	-0,53	3,5	-0,63

Os resultados das subtrações de cada pessoa entrevistada deverão ser multiplicados entre si para, posteriormente, serem todos somados. Vejamos a próxima tabela:

Pessoa	Dias	$X - \bar{X}$	Horas	$Y - \bar{Y}$	Multp.
1	16	4,47	6,5	2,37	10,57
2	13	1,47	5,5	1,37	2,00
3	15	3,47	5,5	1,37	4,74
4	14	2,47	5,5	1,37	3,37
5	10	-1,53	4,5	0,37	-0,56
6	8	-3,53	2,5	-1,63	5,77
7	14	2,47	3,5	-0,63	-1,56
8	9	-2,53	2,5	-1,63	4,14
9	10	-1,53	3	-1,13	1,74
10	8	-3,53	2,5	-1,63	5,77
11	13	1,47	4,5	0,37	0,54
12	13	1,47	4,5	0,37	0,54
13	13	1,47	5,5	1,37	2,00
14	6	-5,53	2,5	-1,63	9,04
15	11	-0,53	3,5	-0,63	0,34
<b>Somatório da coluna</b>					48,43

Por fim, dividimos o resultado do somatório pela quantidade de pessoas entrevistadas menos um. Logo, dividindo 48,43 por 14, obteremos a covariância 3,46. Através deste resultado, concluímos que existe uma relação positiva entre dias de viagem e horas de passeios culturais. Isto é: conforme mais tempo a pessoa fica na cidade, mais tempo gasta com passeios culturais e vice-versa.

---

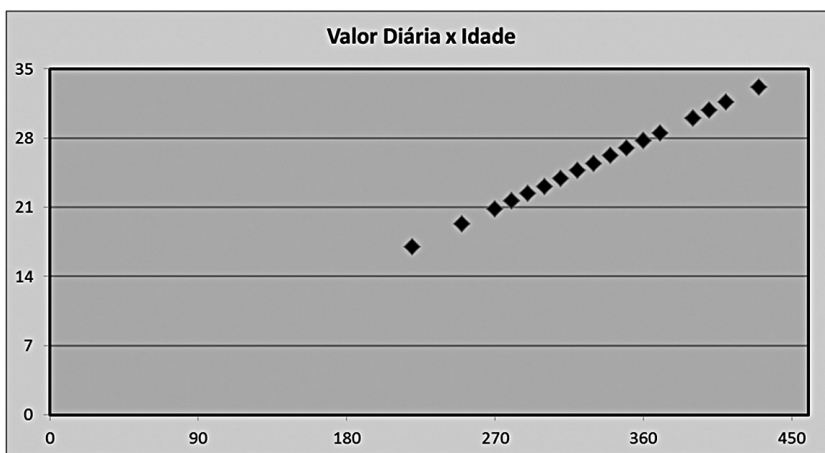
---

---

---

## Gráfico de dispersão

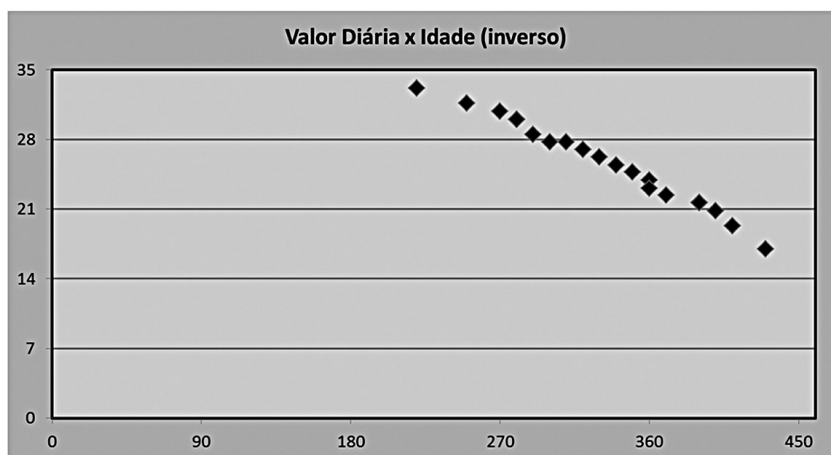
Ao falar de relação entre duas variáveis de uma mesma amostra, falamos também de *Gráfico de Dispersão*. Este tipo de gráfico ilustra exatamente esta relação, fazendo com que seja possível representar o comportamento das respostas e, graficamente, interpretar o seu comportamento. Para a sua construção, usamos o mesmo critério feito para o cálculo da covariância. Assim, uma variável da amostra que chamamos de  $X$  será representada no eixo  $x$ , e a variável da amostra que chamamos de  $Y$  será representada no eixo  $y$ . Com isto, vários pontos surgirão no gráfico – “desenhando” a possível relação entre as variáveis. Assim, usando os dados da pesquisa sobre valor da diária e idade, foi gerado o gráfico da **Figura 10.2**. Vejamos o comportamento dos dados, notando que no eixo  $x$ , temos os valores de diárias e, no eixo  $y$ , as idades:



**Figura 10.2:** Gráfico de dispersão – Diária x Idade.

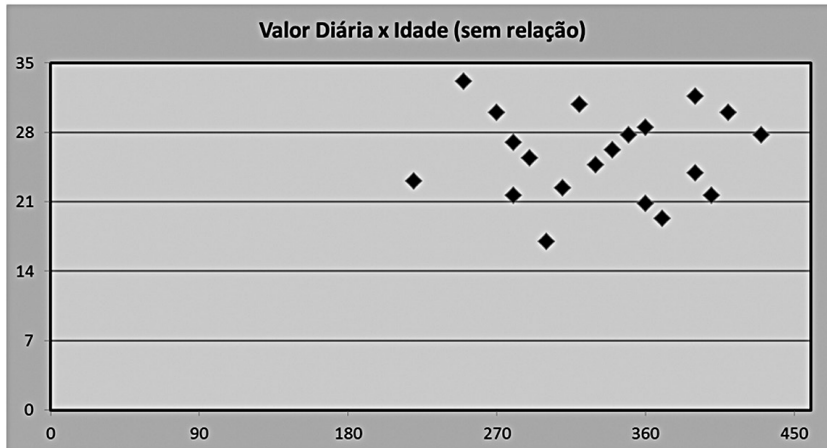
Desse modo, cada ponto é um entrevistado. Esses pontos são obtidos pela intercessão da diária respondida (eixo  $x$ ) com a idade (eixo  $y$ ). O primeiro ponto, que aparece na extrema esquerda, é o entrevistado 14, pois é o que informou o menor valor de diária. Ele também é o mais baixo dos pontos porque é o mais novo dos entrevistados. De maneira análoga, o ponto mais alto e mais à direita é o entrevistado 8 – por ser o mais velho e o que respondeu a maior diária.

Em vista disto, graficamente, é fácil identificar a relação positiva das duas variáveis. Note que, conforme o valor da diária aumenta (anda para a direita no eixo  $x$ ), o valor da idade também aumenta (sobe no eixo  $y$ ). Vejamos agora, na **Figura 10.3**, um exemplo de relação inversa – ao supormos um estudo similar a este que usamos:



**Figura 10.3:** Gráfico de dispersão – Diária x Idade (inverso).

A relação das variáveis nesse gráfico (**Figura 10.3**) está tão óbvia quanto no anterior. Contudo, estamos agora lidando com uma relação negativa. Isto é: quando uma variável cresce, a outra decresce. Isto fica claro se olharmos os dois pontos extremos do gráfico. O primeiro à esquerda possui a maior idade de todas. Entretanto, possui o menor valor de diária. Já o último ponto, na extrema direita, possui o maior valor de diária, mas a menor idade. Como foi dito, além das relações positivas e relações negativas, existem casos nos quais não existe relação alguma. Vejamos, então, a **Figura 10.4**:



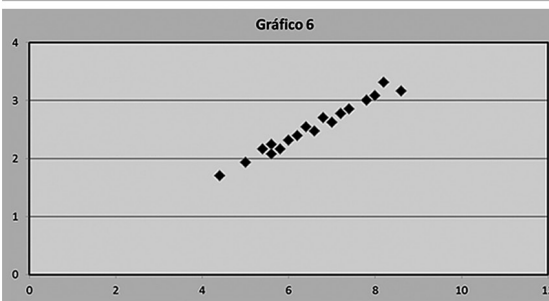
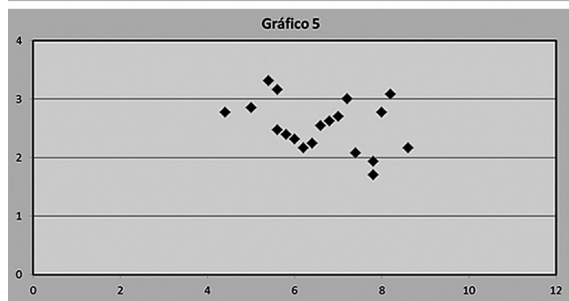
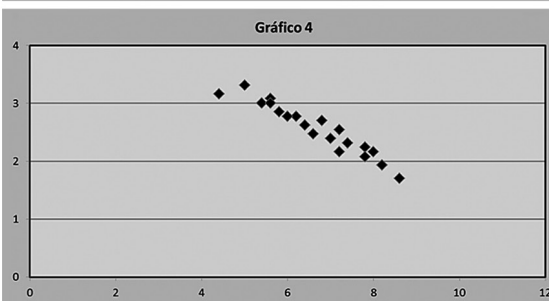
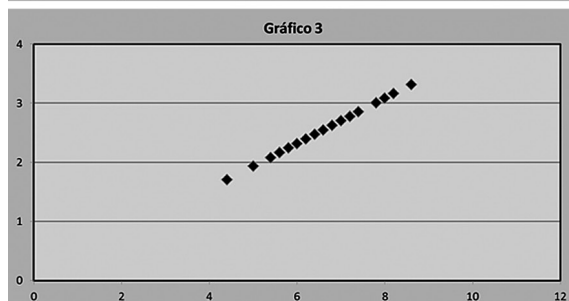
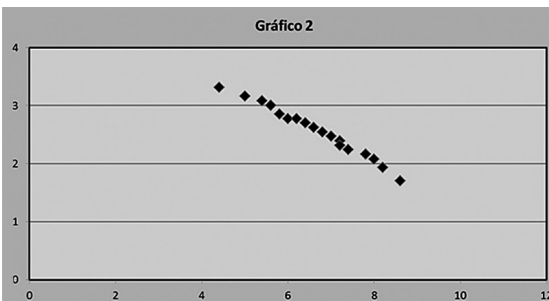
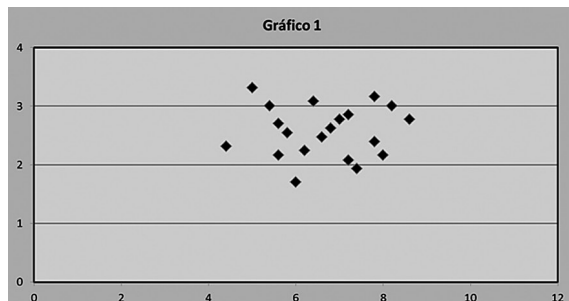
**Figura 10.4:** Gráfico de dispersão – Diária x Idade (sem relação).

Repare no ponto localizado à extrema esquerda. Ele parece até estar propositalmente deslocado dos demais para chamar mais atenção. Indiscutivelmente, ele possui o menor valor de diária, mas sua idade é intermediária em relação às demais idades. Da mesma forma, destaca-se o ponto mais alto de todos, quase posicionado abaixo da letra S da palavra SEM do título do gráfico. Ele representa a maior idade de todas e, ao mesmo tempo, a segunda menor diária. Apenas com estes dois pontos destacados, já é possível notar que não existe uma relação implícita entre todos os pontos. Ainda assim, analisando o gráfico de forma geral, nota-se que é apenas um monte de pontos espalhados, sem um padrão sequer. Daí, podemos concluir que não existe uma relação entre as duas variáveis estudadas.

## Atividade 2

### Atende ao objetivo 2

Alguns estudos foram feitos e, após os dados terem sido coletados e organizados, obtiveram os seis gráficos apresentados a seguir. Identifique que tipo de relação existe em cada um deles. Justifique.




---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

**Resposta comentada**

Gráfico 1 – Nenhuma relação associada

Gráfico 2 – Relação negativa

Gráfico 3 – Relação positiva

Gráfico 4 – Relação negativa

Gráfico 5 – Nenhuma relação associada

Gráfico 6 – Relação positiva

---

---

---

---

## Coeficiente de correlação

Como visto, tanto a covariância quanto o gráfico de dispersão podem ajudar a determinar a relação entre duas variáveis de uma mesma amostra. Contudo, cada um possui uma limitação. A covariância permite que se obtenha qualquer resultado. Logo, nunca saberemos se um resultado de, por exemplo, -18,68 representa uma relação negativa forte ou fraca. De igual modo, como no caso do valor das diárias e das idades, no qual foi obtido 251,632, trata-se de um valor alto, mas necessariamente estamos falando de uma relação positiva forte?

Já o gráfico de dispersão, por motivos óbvios, possui a limitação da visualização. Isto é: se a relação, tanto positiva quanto negativa, não for muito clara, será difícil identificar. Em alguns casos, existem sutis relações que sequer são passíveis de serem notadas graficamente. Com isso, se não fizermos uma abordagem alternativa, corremos o risco de interpretar equivocadamente.

Deste modo, para que possamos chegar a um resultado mais fidedigno, que analise a relação entre duas variáveis de uma mesma amostra, temos a medida numérica chamada *coeficiente de correlação*. Ela será interpretada da mesma forma que a covariância, isto é, valores positivos implicam relações positivas, e valores negativos implicam relações ne-

gativas. Porém, por só admitir resultados entre -1 e +1, ela é mais precisa na sua conclusão. A **Figura 10.5** ilustra a fórmula do seu cálculo:

$$r = \frac{\text{cov}(X, Y)}{S_X S_Y}$$

**Figura 10.5:** Fórmula de cálculo do Coeficiente de Correlação

De maneira resumida, o cálculo do coeficiente de correlação trata de dividir a covariância pelo produto dos desvios-padrão das variáveis que estão sendo estudadas. Logo, por mais que o coeficiente de correlação seja mais preciso que a covariância, deveremos calculá-la também. Isto remete basicamente à variância, que passa a ser uma das etapas para o cálculo do desvio-padrão.

Deste modo, como dito, o resultado do coeficiente de correlação varia apenas entre -1 e +1. Esta limitação permite prever uma maior precisão para estes resultados. O conceito de que um resultado negativo denota uma relação negativa e um resultado positivo denota uma relação positiva se mantém. Contudo, a precisão agora fica por conta do valor obtido – quanto mais próximo do zero for o resultado, mais fraca será a relação. Logo, analogamente, quanto mais distante do zero, mais forte será a relação, sendo que, quando o resultado for precisamente -1 ou +1, dizemos que temos, respectivamente, uma relação negativa perfeita ou uma relação positiva perfeita. Vamos, então, retornar ao caso das diárias e da idade que estamos usando desde o início desta aula. Já calculamos sua covariância e fizemos a leitura pelo gráfico de dispersão. Agora, iremos determinar seu Coeficiente de Correlação. Para tal, precisamos da covariância (já calculada) e dos desvios-padrão (que sugerimos que calculassem no início da aula). Aplicando estes valores na fórmula, temos o seguinte resultado, conforme a **Figura 10.6**:

$$r = \frac{\text{cov}(X, Y)}{S_X S_Y} = \frac{251,632}{57,297 \cdot 4,392} = \frac{251,632}{251,648} = 0,99994$$

**Figura 10.6:** Cálculo do coeficiente de correlação – Diárias x Idade

O resultado é, indiscutivelmente, uma relação positiva perfeita. Deixar o resultado com tantas casas decimais foi apenas uma solução exagerada para ilustrar o valor final. Mas, de fato, fazendo o arredondamento, conforme já comentamos em outras aulas, o resultado é +1. Com isto, podemos confirmar que 251,632 era um valor difícil de prever o significado. Contudo, com o coeficiente de correlação ficou mais claro e óbvio o entendimento da relação entre as variáveis em questão.

### Atividade 3

*Atende aos objetivos 4 e 5*

Utilizando os dados da Atividade 1, determine o coeficiente de correlação entre as variáveis “Dias de viagem” e “Horas de eventos culturais” para, posteriormente, interpretar o resultado.

This image shows a single sheet of white paper with horizontal ruling lines. The lines are evenly spaced and run across the width of the page. There are no margins, text, or other markings on the paper.

### Resposta comentada

Para determinarmos o coeficiente de correlação, precisamos da covariância entre as variáveis em questão e os respectivos desvios-padrão de cada variável. A covariância foi calculada na própria Atividade 1, e o resultado foi 3,46. Os desvios-padrão precisam ser calculados. Através dos cálculos já ensinados em aulas anteriores, chegaremos ao resultado de 2,92 para o desvio-padrão de “Dias de viagem” e de 1,37 para o desvio-padrão de “Horas em eventos culturais”. De posse destes dados, basta aplicá-los à fórmula, conforme a próxima figura:

$$r = \frac{\text{cov}(X,Y)}{S_X S_Y} = \frac{3,46}{2,92 \cdot 1,37} = \frac{3,46}{4,0004} = 0,865$$

O coeficiente de correlação é 0,865. Logo, por ser positivo, trata-se de uma relação positiva. Por se tratar de um valor próximo de +1, trata-se de uma relação forte.



### Conclusão

Vimos nesta aula que, igualmente nas anteriores, a necessidade de aprimorar um resultado nos leva a outras formas de analisar uma amostra. Especificamente, nesta aula, foi dissertado sobre a análise de duas variáveis de uma mesma amostra entre si. Isto porque, em algumas pesquisas, são feitas mais de uma pergunta. Com isto, relacionar as diferentes respostas de cada indivíduo (variáveis) entre as dos outros indivíduos pode nos trazer mais conclusões que ajudarão na leitura do cenário que está para ser criado.

Assim, o que precisa ficar claro nessa análise que será feita é que o resultado remete a uma interpretação unicamente associada ao estudo em questão. Portanto, quando concluímos que, na pesquisa sobre valor de diária e idade, quanto mais velha for a pessoa, mais ela está disposta a pagar mais caro pelo seu conforto, foi uma interpretação estritamente para aquele cenário em questão. Isto é: não pode ser generalizado para qualquer situação, mesmo que pareça fazer sentido.

Outro fator que precisa ser lembrado é que, mesmo existindo uma relação (seja positiva ou negativa) entre duas variáveis, esta relação nem sempre é necessariamente explicada apenas por elas duas. Suponhamos que uma pesquisa levantou alguns dados gerais de uma cidade durante um mês. Ao final, resolveram analisar a relação entre o tempo médio de deslocamento dos veículos em uma longa avenida principal e a quantidade de ligações para a Defesa Civil daquela cidade. Foi notada uma forte relação positiva entre estas duas variáveis, mas a explicação não está nelas, e sim em uma terceira, que é o índice de chuvas neste dia. Isto é: a quantidade de ligações para a Defesa Civil não tem relação direta no tempo de deslocamento em um ponto da cidade, tão pouco a sua leitura inversa. Entretanto, a quantidade de chuvas interfere no deslocamento pela cidade (seja por alagamentos ou simplesmente por excesso de cautela dos motoristas), como influencia nas ligações para a Defesa Civil – pedindo ajuda sobre alagamentos, deslizamentos, entre outras. Logo, quanto mais chove, mais lento fica o trânsito e mais ligações temos. Mas note que as duas últimas variáveis só fazem sentido com a variável da chuva.

Por fim, o último cuidado que precisa ser tomado é que não podemos interpretar tudo com extrema frieza e rigidez. O *coeficiente de correlação*, assim como a *covariância* ou qualquer outra medida que citamos nas demais aulas são resultados obtidos após diversas operações matemáticas. Logo, são incapazes de interpretar a parte da realidade envolvida. Podemos, para melhor ilustrar, imaginar que ao calcular a relação entre duas variáveis de uma amostra, obtiveram o resultado 0,95. Pelo conceito já dito nesta aula, estamos falando de uma relação positiva muito forte e quase perfeita. Mas a pergunta que fica é: o que estamos analisando? E se uma variável for a quantidade de copos de café vendidos no aeroporto de Belo Horizonte e a outra variável a quantidade de cangurus albinos que nasceram na Austrália? Mesmo com este resultado indiscutível, vocês seriam capazes de afirmar que quanto mais cafezinhos são vendidos no aeroporto de Belo Horizonte, mais cangurus albinos nascem na Austrália? Obviamente, a intervenção humana continua sendo necessária nos estudos estatísticos, para que os números, dentro da sua frieza e rigidez, não sejam os únicos a ditar o que se passa em cada cenário.

**Atividade final**

Atende aos objetivos 1, 2, 4 e 5

Foi feito um levantamento no qual eram coletadas as temperaturas máximas diárias numa certa cidade durante um determinado mês. Nesse mesmo estudo, para cada dia, tomaram nota da quantidade de cocos vendidos pelas principais praias da cidade. Os resultados foram agrupados na tabela que segue:

Dia	Temp.	Cocos	Dia	Temp.	Cocos	Dia	Temp.	Cocos
1	35,2	1.300	11	24,4	2.200	21	39,1	1.200
2	36,3	1.900	12	30,9	3.900	22	36,6	1.300
3	36,7	2.300	13	31,2	4.200	23	35,9	1.800
4	36,8	3.500	14	31,5	1.200	24	36,2	2.900
5	31,9	4.100	15	33,9	1.300	25	38,7	4.600
6	30,4	4.300	16	35,6	1.800	26	38,9	5.800
7	28,4	1.100	17	36,1	2.800	27	39,5	5.600
8	28,2	1.100	18	38,9	4.700	28	33,2	1.100
9	25,7	1.600	19	38,6	5.200	29	32,6	1.300
10	25,1	2.000	20	38,5	5.500	30	31,5	1.700

De posse desses valores, calcule a covariância e analise o seu resultado. Depois, calcule o coeficiente de correlação e analise o seu resultado.

[illegible]

### **Resposta comentada**

Para iniciarmos o cálculo da covariância, faz-se necessário primeiro calcular as médias. A média das temperaturas é 33,88 e a média da quantidade de cocos vendidos é 2.776,67. Agora é possível calcular a covariância, conforme a próxima tabela:

Dia	Temp.	$X - \bar{X}$	Cocos	$Y - \bar{Y}$	Dia	Temp.	$X - \bar{X}$	Cocos	$Y - \bar{Y}$	Dia	Temp.	$X - \bar{X}$	Cocos	$Y - \bar{Y}$
1	35,2	1,32	1.300	-1.477	11	24,4	-9,48	2.200	-577	21	39,1	5,22	1.200	-1.577
2	36,3	2,42	1.900	-877	12	30,9	-2,98	3.900	1.123	22	36,6	2,72	1.300	-1.477
3	36,7	2,82	2.300	-477	13	31,2	-2,68	4.200	1.423	23	35,9	2,02	1.800	-977
4	36,8	2,92	3.500	723	14	31,5	-2,38	1.200	-1.577	24	36,2	2,32	2.900	123
5	31,9	-1,98	4.100	1.323	15	33,9	0,02	1.300	-1.477	25	38,7	4,82	4.600	1.823
6	30,4	-3,48	4.300	1.523	16	35,6	1,72	1.800	-977	26	38,9	5,02	5.800	3.023
7	28,4	-5,48	1.100	-1.677	17	36,1	2,22	2.800	23	27	39,5	5,62	5.600	2.823
8	28,2	-5,68	1.100	-1.677	18	38,9	5,02	4.700	1.923	28	33,2	-0,68	1.100	-1.677
9	25,7	-8,18	1.600	-1.177	19	38,6	4,72	5.200	2.423	29	32,6	-1,28	1.300	-1.477
10	25,1	-8,78	2.000	-777	20	38,5	4,62	5.500	2.723	30	31,5	-2,38	1.700	-1.077

Com esses resultados, faremos a multiplicação dos valores de  $[X - \bar{X}]$  e  $[Y - \bar{Y}]$ , dia a dia, conforme a tabela que segue.

Dia	$X - \bar{X}$	$Y - \bar{Y}$	Multp.	Dia	$X - \bar{X}$	$Y - \bar{Y}$	Multp.	Dia	$X - \bar{X}$	$Y - \bar{Y}$	Multp.
1	1,32	-1.477	-1.944,3	11	-9,48	-577	5.468,7	21	5,22	-1.577	-8.224,9
2	2,42	-877	-2.118,6	12	-2,98	1.123	-3.351,3	22	2,72	-1.477	-4.011,6
3	2,82	-477	-1.342,6	13	-2,68	1.423	-3.819,3	23	2,02	-977	-1.969,6
4	2,92	723	2.109,7	14	-2,38	-1.577	3.757,7	24	2,32	123	285,7
5	-1,98	1.323	-2.624,6	15	0,02	-1.477	-24,6	25	4,82	1.823	8.782,4
6	-3,48	1.523	-5.306,3	16	1,72	-977	-1.676,6	26	5,02	3.023	15.167,1
7	-5,48	-1.677	9.193,7	17	2,22	23	51,7	27	5,62	2.823	15.857,7
8	-5,68	-1.677	9.529,1	18	5,02	1.923	9.648,7	28	-0,68	-1.677	1.145,7
9	-8,18	-1.177	9.629,1	19	4,72	2.423	11.430,1	29	-1,28	-1.477	1.895,1
10	-8,78	-777	6.821,7	20	4,62	2.723	12.572,7	30	-2,38	-1.077	2.566,1

Em seguida, faremos o somatório dos produtos calculados  $\left([X - \bar{X}] \cdot [Y - \bar{Y}]\right)$ . O resultado será 89.498,33. Por fim, dividiremos pelo total de dias menos um (29), obtendo a covariância de valor 3.086,15. Podemos concluir apenas que se trata de uma relação positiva entre as variáveis dessa amostra. Com a covariância calculada, precisamos apenas do desvio-padrão de cada variável para iniciar o cálculo do coeficiente de correlação. Recorrendo ao cálculo do desvio-padrão, já ensinado em aulas anteriores, chegaremos ao valor de 4,42 para as temperaturas e de 1.595,62 para a quantidade de cocos vendidos. Agora, basta aplicar à fórmula, conforme a próxima figura:

$$r = \frac{\text{cov}(X, Y)}{S_X S_Y} = \frac{3.086,15}{4,42 \cdot 1.595,62} = \frac{3.086,15}{7.052,64} = 0,436$$

Com resultado de 0,436 para o coeficiente de correlação, confirmamos a conclusão tirada, com a covariância, de que a relação entre as variáveis é positiva. Contudo, agora é possível determinar o quão forte é esta relação. Como o valor está mais próximo de zero do que de +1, devemos afirmar que é uma relação positiva não muito forte. Isto é: para este estudo, nem sempre quando tivemos altas temperaturas, tivemos grandes vendas de coco. Talvez isso se dê porque em dias de semana, mesmo estando muito quente, o movimento nas praias não é tão grande.

## Resumo

Nesta aula, vimos como é possível relacionar duas variáveis de uma mesma amostra e tirar conclusões sobre como uma supostamente influencia a outra e vice-versa. De início, a primeira medida que nos foi apresentada, chamada de *covariância*, aparentava ser bastante útil para nossos estudos. Em seguida, aprendemos a ler *gráficos de dispersão*, que são gráficos usados para ilustrar o comportamento de duas variáveis em uma mesma amostra. Notamos que existem três tipos de relações possíveis entre essas variáveis e que, dependendo do comportamento delas, estas relações podem ser facilmente identificadas. Por fim, com a constatação de que a covariância pode retornar um resultado com pouca precisão e que o gráfico de dispersão assume o risco de apresentar formatos incapazes de serem interpretados, surge o *coeficiente de correlação*. Tal medida nova é capaz de não somente determinar que tipo de relação as variáveis possuem entre si, como também o quanto intensa esta relação se apresenta.

## Informação sobre a próxima aula

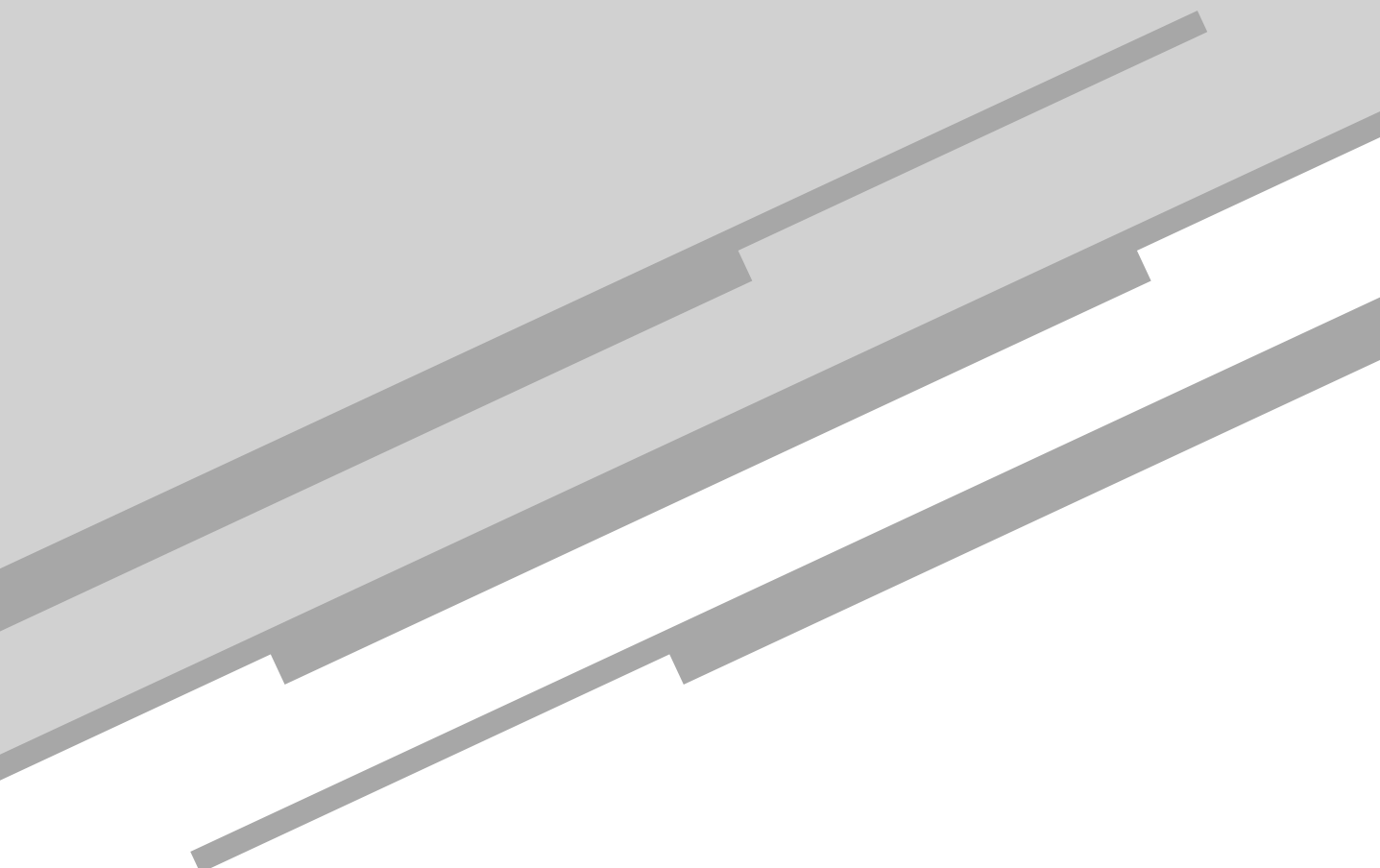
Na próxima aula, começaremos a falar de *probabilidade*. Aprenderemos o seu vocabulário, como determinar as chances de ocorrer um evento e algumas particularidades. Infelizmente, não iremos descobrir um método de adivinhar os números da Mega-Sena, mas ainda assim, seu eu fosse você, continuaria por aqui.

Um abraço!



# Aula 11

O próximo sorteado é...



*Rafael Canellas Ferrara Garrasino*

## **Meta**

Apresentar a análise combinatória como instrumento para determinar o espaço amostral de eventos possíveis.

## **Objetivos**

Esperamos que, após o estudo desta aula, você seja capaz de:

1. calcular arranjos com e sem repetição;
2. calcular permutações;
3. calcular combinações;
4. calcular permutações com repetições.

## Introdução

Nas aulas anteriores, falamos muito a palavra “probabilidade”. Comentamos que a probabilidade de um evento *outlier* é muito baixa. Afirmamos que a probabilidade de ocorrer um evento com valor muito próximo da média é alta. Mas, até então, não conceituamos probabilidade, tampouco nos aprofundamos em como determinar a probabilidade de algo.

Em vista disso, nesta aula, iniciaremos o estudo da probabilidade de um evento. Não falaremos diretamente de probabilidade, mas de uma parte que a antecede que é de extrema importância: a análise combinatória. Vale ressaltar, então, que pretendemos iniciar o universo da probabilidade com a primeira etapa do processo, que é quantificar as possibilidades de ocorrência de um evento, além de apresentar as diversas formas de contagem que existem e em quais casos elas se enquadram.

Desse modo, se perguntarmos para quaisquer pessoas, mesmo que não sejam entendidas do assunto, qual a probabilidade de dar cara em um jogo de “cara ou coroa”, quase todas dirão 50%. Isto se dá porque, intuitivamente, elas sabem que existem duas opções de resultado (cara ou coroa). Logo, metade das chances vai para cada um deles. Este raciocínio está correto, ou seja, estimar quantas opções de resultado existem, para, depois concluir a probabilidade, é de fato o primeiro passo de um cálculo probabilístico.

Agora, suponhamos que a estimativa não seja mais de um jogo de cara ou coroa. Vamos imaginar um homem acordando bem cedo, ainda escuro, com sono e preguiça de acender a luz. Ele vai até a gaveta das meias, pega duas sem enxergar direito, veste o resto da roupa e vai trabalhar. Quais as chances de ele ter escolhido duas meias da mesma cor e não passar vergonha? Obviamente, se fosse conosco, com certeza chegaríamos ao trabalho com uma meia vermelha e uma verde. Afinal, quando algo pode dar errado, normalmente dá. Mas, deixando de lado a face debochada da vida e voltando para o lado matemático, precisaríamos contar quantas meias de cada cor ele possui e estimar quantos pares de meias ele poderia montar com elas. Notem que não é mais algo tão intuitivo assim como o jogo de “cara ou coroa”. Imagino que algum de vocês sugeriu, como solução para esse impasse, acender a luz, pois dá bem menos trabalho do que ficar fazendo contas. Podemos concordar com isso, mas, no momento, teremos de fazer contas também. Vamos lá?

## Arranjos

Ressalte-se que sempre que precisarmos estimar a quantidade de possibilidades de resultados para um evento estamos lidando com um processo de contagem. O processo de contagem, por possuir diversos formatos, possui métodos diferentes de resolução. Cada método se encaixa em uma situação apropriada, de acordo com as condições do caso em questão. Geralmente, o primeiro a ser apresentado é o *arranjo*, principalmente por sua simplicidade.

Assim, suponhamos uma criança chamada Rita. Ela ganhou um jogo de blocos de madeira. Cada bloco possui uma letra do alfabeto. Rita estava brincando com os blocos que possuem a letra do seu nome: R, I, T e A. Em um determinado momento, ela escondeu os quatro blocos atrás de si. Pegou um deles sem olhar e era a letra T. Depois o escondeu, misturou os quatro sem olhar e pegou novamente: era a letra R. Depois disso, ela voltou a brincar com todos os blocos do alfabeto.

Vamos, então, retornar à brincadeira de Rita. Ela tinha quatro blocos diferentes ocultos atrás dela. Ela escolheu um sem olhar, checou qual era a letra e o devolveu. Depois, voltou a pegar um, dentre esses quatro, sem olhar. Naquele momento ela acabara de formar o resultado TR. Mas a pergunta que fica é: quantos resultados diferentes ela conseguiria fazer com aquela mesma brincadeira? Para solucionar esse caso, podemos tentar escrever todos os resultados, conforme a **Figura 11.1**:

RI	RT	RA	IR
IT	IA	TR	TI
TA	AR	AI	AT
RR	II	TT	AA

**Figura 11.1:** Máximo de resultados diferentes para a brincadeira de se escolher, de forma aleatória, 2 de 4 blocos de madeira com uma letra em cada, mas com restituição de bloco antes escolhido.

De fato, encontrar todas as soluções de maneira manual não é algo tão complexo ou exaustivo assim. Mas, suponhamos que Rita tenha resolvido

brincar com as letras do estado onde nasceu e ela seja de Pernambuco. Já temos uma atividade mais longa e cansativa! Ficaria mais interessante desenvolver um método de calcular os resultados dessa brincadeira de maneira fácil – independentemente da quantidade de letras.

Nesse sentido, vejamos passo a passo da **Figura 11.2**. Primeiro, estabelecemos duas células, dois campos ou espaços, ou seja, o que considerar conveniente. O importante é que, como no jogo inicial, Rita vai retirar dois blocos. Precisamos ter registrados esses dois movimentos (passo 1). Em seguida, faremos o primeiro movimento de Rita, que é retirar um bloco ao acaso. Vamos lembrar que são quatro blocos ocultos. Logo, ela possui quatro opções disponíveis (passo 2). Em seguida, ela fará a segunda escolha aleatória de blocos. Mas como ela devolve aos restantes o bloco que acabou de retirar, para esta segunda escolha ela volta a ter quatro opções (passo 3). Por fim, sacramentamos a solução transformando em uma multiplicação (passo 4), obtendo o resultado de 16 resultados diferentes para essa brincadeira, igualmente como estimamos na **Figura 11.1**.

Passo 1: \_\_\_\_\_

Passo 2: 4 \_\_\_\_\_

Passo 3: 4      4

Passo 4: 4   x   4

**Figura 11.2:** Método de se calcular o máximo de resultados diferentes para a brincadeira de se escolher, de forma aleatória, 2 de 4 blocos de madeira com uma letra em cada, mas com restituição de bloco antes escolhido.

Notem que, agora, com essa metodologia, nem será tão cansativo fazer a mesma brincadeira com as letras do estado de Pernambuco. Acompanhem pela **Figura 11.3**:

Passo 1: \_\_\_\_\_

Passo 2: 10 \_\_\_\_\_

Passo 3: 10 10

Passo 4: 10 x 10

**Figura 11.3:** Método de se calcular o máximo de resultados diferentes para a brincadeira de se escolher, de forma aleatória, 2 de 10 blocos de madeira com uma letra em cada, mas com restituição de bloco antes escolhido.

Desse modo, continuamos com duas rodadas de escolha de blocos. Logo, dois campos (passo 1). Em seguida, estimamos a quantidade de opções de blocos disponíveis para Rita retirar. Assim, como a palavra Pernambuco tem 10 letras, usaremos o número 10 (passo 2). Depois, lembrando que Rita devolve o bloco escolhido, voltamos a ter 10 opções para a segunda retirada (passo 3). Por fim, sacramentando a multiplicação chegamos ao resultado de impressionantes 100 resultados diferentes para esta segunda brincadeira.

É importante destacar que, com essa metodologia consolidada, qualquer variação de exercício será facilmente solucionada. Vamos supor uma urna com quatro bolas: uma vermelha, uma verde, uma azul e uma amarela. Uma pessoa coloca a mão dentro da urna e, sem olhar, escolhe uma das bolas. Alguém toma nota da cor da bola e a devolve para a urna. Daí, outra pessoa coloca a mão dentro da urna e escolhe uma bola também sem olhar. O mesmo alguém anota a cor da bola. Se quiséssemos determinar quantos resultados diferentes poderíamos obter com esse exercício, bastaria apenas repetir a solução da **Figura 11.2**. Mudamos a ambientação da situação, mas a solução, que é genérica, permanece a mesma.

Por outro lado, caso tenhamos mais movimentos de retirada de bolas ou escolha de blocos, caberá a quem está calculando apenas inserir mais um campo. No recém citado caso das bolas coloridas, caso tivéssemos três retiradas, sendo todas com reposição de bolas após anotada a cor, teríamos três campos respectivamente preenchidos com o número quatro. Portanto, passaríamos a ter 64 resultados distintos. Contudo, que fique bem claro: sempre após anotada no bloco a cor da bola ou o exem-

plô que for, o elemento é devolvido, voltando a fazer parte do grupo original. Entretanto, e se não houver devolução?

Podemos intuitivamente, inspirados nos casos que exemplificamos até agora, montar uma fórmula para casos desse tipo. Isto é: quando tivermos  $n$  opções para sortear ao acaso  $r$  vezes, com reposição do item escolhido, o cálculo da quantidade de arranjos distintos será feito pela fórmula:

$$AR_{n,r} = n \cdot n \cdot \dots \cdot n = n^r$$

Note que esse caso só fará sentido se a ordem do sorteio interferir no resultado final. No sorteio dos blocos, quando quero formar uma palavra, a ordem faz diferença, isto é, sorteando as letras R e I, dependendo da ordem, formam-se duas palavras distintas (Ir e Ri). Contudo, se quiséssemos apenas montar um grupo de letras, o grupo de letras I e R será igual ao grupo R e I, logo a ordem do sorteio não interfere. Para situações nas quais a ordem do sorteio interfere diretamente no resultado, teremos outro tipo de abordagem, que será vista posteriormente.

Retornemos então à urna com as quatro bolas coloridas (uma vermelha, uma verde, uma azul e uma amarela). Suponhamos que uma pessoa retire uma bola aleatoriamente dessa urna e a coloque sobre um copo com o número 1. Vamos fingir que foi a bola de cor verde. Em seguida, mantendo a bola verde sobre o copo, a pessoa retira outra bola sem olhar para dentro da urna, colocando-a em seguida sobre um copo com o número 2. A bola retirada é de cor azul. Visualizando como foi feita a atividade com a urna e as bolas coloridas, vamos calcular a quantidade de resultados diferentes possíveis conforme a **Figura 11.4**:

Passo 1: \_\_\_\_\_

Passo 2: 4 \_\_\_\_\_

Passo 3: 4      3

Passo 4: 4 x 3

**Figura 11.4:** Método de se calcular o máximo de resultados diferentes para a brincadeira de se escolher, de forma aleatória, 2 de 4 bolas coloridas, mas sem substituição de bola antes escolhida.

Note que o procedimento é muito parecido como o feito na atividade em que as bolas, após retiradas, eram repostas. A diferença básica está agora na segunda rodada (passo 3) que, por não haver devolução da bola que foi antes sorteada, a quantidade de bolas disponíveis para sorteio diminui de 4 para 3. Com isso, para esse tipo de atividade teremos 12 tipos de resultados diferentes. Vejamos agora um caso maior.

Assim, suponhamos uma turma com 25 alunos. A professora pediu que eles organizassem uma equipe de tal forma que elessem um presidente da turma para representá-los, um comunicador interno para divulgar entre os alunos as novidades, notícias e informações novas da escola, bem como um mensageiro que irá recolher todos os trabalhos que devem ser entregues, além de pegá-los corrigidos para devolver aos donos. Uma informação bastante importante, que deve ser destacada, foi que a professora exigiu que nenhum aluno acumule mais de uma função. Vejamos na **Figura 11.5** quantas combinações diferentes de equipes poderão ser feitas nessa turma:

Passo 1: \_\_\_\_\_

Passo 2: 25 \_\_\_\_\_

Passo 3: 25    24 \_\_\_\_\_

Passo 4: 25    24    23

Passo 5: 25 x 24 x 23

**Figura 11.5:** Método de se calcular o máximo de resultados diferentes para três vagas, em um conjunto de 25 alunos.

A solução é idêntica ao exemplo anterior, apenas mudamos a quantidade de elementos disponíveis para sorteio e a quantidade de “vagas” a preencher. Note que pouco importa, para fins matemáticos, o nome de cada “vaga”. Estamos somente preocupados com três “vagas” disponíveis, além do fato de que o arranjo de alunos em vagas diferentes forma resultados diferentes.



Caso seja necessária uma fórmula para os novos casos que estamos estudando agora, precisamos diferenciá-los dos anteriores. Nos primeiros, os itens escolhidos eram devolvidos ao grupo original e a ordem do sorteio fazia diferença. Agora, os itens sorteados não são devolvidos, mas continuamos com uma situação na qual a ordem do sorteio faz diferença.

Por sua vez, para esse novo caso, em que temos  $n$  opções para sorteio sem repor os elementos sorteados e com  $r$  sorteios, a fórmula para o cálculo do número de resultados diferentes é:

$$A_{n,r} = (n)(n-1)(n-2)\dots(n(r-1))$$

É importante lembrar que o número de sorteios não pode ser maior que o número de elementos disponíveis, caso contrário faltarão elementos para serem sorteados.

No primeiro passo, temos as três “vagas” disponíveis a serem preenchidas por uma turma de 25 alunos (candidatos). Após o primeiro aluno ser selecionado para uma delas, restam 24 para a segunda. Feita a segunda escolha, restam 23 alunos para a terceira e última vaga. Por fim, calculando o produto dos valores temos 13.800 maneiras diferentes de arrumar essa equipe na turma em questão.

## Atividade 1

### Atende ao objetivo 1

Uma empresa abriu processo seletivo para quatro funções: digitador, recepcionista, controle de correspondência e controle de documentação. Como os anúncios de vagas foram disponibilizados para todos os can-

didatos, a empresa recebeu 40 pessoas dispostas a preencher uma das quatro vagas em questão. Com estas informações, calcule o que se pede:

a) De quantas maneiras diferentes podemos preencher as quatro funções, considerando que um mesmo funcionário pode exercer até as quatro ao mesmo tempo.

---

---

---

---

---

b) De quantas maneiras diferentes podemos preencher as quatro funções, considerando que um mesmo candidato não pode acumular tarefas.

---

---

---

---

---

### **Resposta comentada**

a) Nesta hipótese, temos a possibilidade de repetição. Isto é: poderemos alocar um mesmo funcionário em mais de uma tarefa. Logo, após selecionado para a primeira vaga, ele volta a disputar a seguinte e assim sucessivamente. Portanto, o cálculo ficará da seguinte forma:

$$\_ \cdot \_ \cdot \_ = 40 \cdot 40 \cdot 40 = 40^3 = 64.000$$

A solução, tanto pela fórmula, quanto pela montagem dos cálculos manualmente, necessariamente precisa resultar no mesmo valor. Contudo, vale destacar que o segundo método é considerado matematicamente mais elegante.

b) Nesta hipótese, um mesmo candidato não pode assumir vagas diferentes. Logo, após selecionado para uma delas, ele automaticamente para de disputar as demais. Isso quer dizer, em suma, que não haverá reposição de elementos disponíveis após uma seleção feita. O cálculo ficará da seguinte forma:

$$\_ \cdot \_ \cdot \_ = 40 \cdot 39 \cdot 38 = 59.280$$

Assim como na resposta anterior, para esta é independente a opção de resolução, continuando ainda assim a opção manual ser a mais elegante dentre as duas.

## Permutações

Na seção Arranjos, vimos o que é um *arranjo*, que pode ser com ou sem repetições. Quando o arranjo que iremos estudar não tiver repetições e a quantidade de rodadas for igual ao número de elementos disponíveis, temos rodadas suficientes para que todos os elementos sejam sorteados. Chamamos este caso específico de arranjo de *permutação*. Contudo, precisamos enfatizar que, para estes casos, continua prevalecendo a condição de que a ordem dos elementos sorteados difere do resultado final.

Desse modo, suponhamos que o técnico de basquete vai montar o time. Entretanto, como ele gostaria de que todos os jogadores passassem por todas as posições para ganhar experiência, ele sorteia quem vai ocupar cada posição. No momento, ele tem dez jogadores e dez vagas: pivô; pivô reserva; ala esquerdo; ala esquerdo reserva; ala direito; ala direito reserva; armador esquerdo; armador esquerdo reserva; armador direito; armador direito reserva. Na **Figura 11.6**, podemos verificar com quantas combinações diferentes ele pode montar o time.

Vagas: \_\_\_\_\_

Opções: \_10\_ \_9\_ \_8\_ \_7\_ \_6\_ \_5\_ \_4\_ \_3\_ \_2\_ \_1\_

Cálculo:  $10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$

**Figura 11.6:** Cálculo do número máximo de permutações acerca de 10 jogadores para 10 vagas.

Resolvendo de maneira manual, o primeiro passo será determinar a quantidade de vagas a serem preenchidas. Note que pouco importa os nomes e a ordem das vagas. O necessário é saber que são dez vagas diferentes que implicam em um resultado diferente para cada combinação feita com os jogadores.



O cálculo de uma permutação é como o cálculo de um arranjo sem reposições e com etapas em quantidades suficientes que todos sejam escolhidos. Logo, a fórmula matemática, para um caso com  $m$  opções disponíveis, pode ser representada como:

$$m \cdot (m-1) \cdot (m-2) \cdot \dots \cdot 2 \cdot 1$$

Para casos específicos como esses, usa-se o que chamamos *fatorial*. Isto nada mais é do que a representação matemática de um cálculo como feito para permutações. O fatorial é expresso com o símbolo da exclamação.

$$m! = m \cdot (m-1) \cdot (m-2) \cdot \dots \cdot 2 \cdot 1$$

Nesse sentido, com as vagas previamente estruturadas, resta determinar quantos candidatos estão disponíveis. Pela **Figura 11.6** é possível notar que, por não ter reposição, a cada jogador sorteado, temos menos um candidato para a próxima vaga até sobrar apenas um para a última. Daí, por fim, consolidamos o cálculo com a multiplicação dos valores chegando a um total de 3.628.800 maneiras diferentes de montar o time.

## Atividade 2

### Atende ao objetivo 2

Anagrama é um jogo de letras no qual escolhemos uma frase ou uma palavra e embaralhamos as letras que as compõem para formar novas frases ou palavras. Na maior parte das vezes, em um anagrama, não

existe um objetivo de formar outra palavra que de fato exista. A ideia é apenas tentar combinar, ao máximo possível, as letras que formam uma palavra em outras palavras, respeitando a quantidade de cada letra da palavra original e usando todas elas. Com essa definição, determine quantos anagramas diferentes podemos formar com as letras da palavra *Pernambuco*.

---

---

---

---

---

---

### **Resposta comentada**

Como o objetivo do anagrama é usar todas as letras de uma palavra, sem que sobre uma sequer ou tampouco o uso além do original na palavra principal, ressaltando-se que *Pernambuco* não tem letras repetidas, estamos falando de um caso de permutação. Na próxima figura temos a solução manual da questão.

Letras: \_\_\_\_\_

Opções: \_10\_ \_9\_ \_8\_ \_7\_ \_6\_ \_5\_ \_4\_ \_3\_ \_2\_ \_1\_

Cálculo:  $10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$

Caso prefiram, o cálculo da solução pode ser feito através do recurso do fatorial. Por esta opção basta apenas considerar que são 10 letras diferentes a serem combinadas sem repetição. Sendo assim, o cálculo fica como na próxima figura.

$$10! = 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 3.628.800$$

Logo, para este caso específico de anagrama, teremos 3.628.800 permutações diferentes.

---

---

---

---

## Combinações

Falamos bastante até aqui de casos nos quais a ordem dos elementos sorteados, necessariamente, compromete o resultado final. Mas como seria em um caso no qual a ordem não interfira no resultado final? Para esses casos, usaremos o termo *combinação*.

Vamos, então, imaginar cinco irmãos: Marcio, Maria, Marcelo, Mariana e Mauricio. Eles pretendem fazer uma festa surpresa no aniversário da mãe. Dentre as várias atividades programadas para o evento, eles planejaram recitar o poema favorito dela – só que em dupla. Para que não haja favorecimento, eles optaram por colocar o nome de cada um deles em um pedaço de papel e depois pedir para que o pai, que organiza o evento, sorteie os dois nomes dos filhos que recitarão o poema.

A situação em questão é um típico caso em que a ordem dos sorteios não interferirá no resultado final. Suponhamos que o primeiro sorteado seja o filho Marcio, e depois o filho Marcelo. Esse resultado é exatamente o mesmo se ambos fossem sorteados em ordem diferente, ou seja, primeiro Marcelo e depois Marcio, pois, no fim, serão os dois que irão recitar o poema. Vejamos na próxima figura como ficariam os resultados possíveis, sem considerar o fato de resultados iguais:

[Marcio; Maria] [Marcio; Marcelo] [Marcio; Mariana] [Marcio; Mauricio]  
 [Maria; Marcio] [Maria; Marcelo] [Maria; Mariana] [Maria; Mauricio]  
 [Marcelo; Marcio] [Marcelo; Maria] [Marcelo; Mariana] [Marcelo; Mauricio]  
 [Mariana; Marcio] [Mariana; Maria] [Mariana; Marcelo] [Mariana; Mauricio]  
 [Mauricio; Marcio] [Mauricio; Maria] [Mauricio; Marcelo] [Mauricio; Mariana]

**Figura 11.7:** Combinação: resultados possíveis, sem considerar resultados iguais.

Note que temos 20 duplas possíveis, mas, como foi dito anteriormente, a ordem não interfere no resultado. Logo, temos duplas repetidas que devem ser retiradas. Vejamos, na figura que segue, quantas combinações diferentes teremos após retirarmos as repetições:

[Marcio; Maria] [Marcio; Marcelo] [Marcio; Mariana] [Marcio; Mauricio]  
~~[Maria; Marcio]~~ [Maria; Marcelo] [Maria; Mariana] [Maria; Mauricio]  
~~[Marcelo; Marcio]~~ ~~[Marcelo; Maria]~~ [Marcelo; Mariana] [Marcelo; Mauricio]  
~~[Mariana; Marcio]~~ ~~[Mariana; Maria]~~ ~~[Mariana; Marcelo]~~ [Mariana; Mauricio]  
~~[Mauricio; Marcio]~~ ~~[Mauricio; Maria]~~ ~~[Mauricio; Marcelo]~~ ~~[Mauricio; Mariana]~~

**Figura 11.8:** Combinação: resultados possíveis, considerando a exclusão de duplas repetidas.

Com a nova contagem de possíveis duplas diferentes, nosso resultado cai para 10 opções. Obviamente, para este caso, tivemos um exemplo simples por estarmos lidando com valores baixos. Mas e se fossem 20 pessoas disputando 5 vagas em uma equipe de recitar poemas? Com certeza lidaremos com um processo longo e penoso. Portanto, tentar compreender de fato o que foi feito torna-se indispensável.

O primeiro passo foi ilustrado na **Figura 11.7**. Ali, cogitamos todas as possibilidades possíveis se a ordem não interferisse no resultado final. Esse tipo de cálculo já foi trabalhado na parte de arranjos desta mesma aula. Em seguida, na **Figura 11.8**, retiramos as repetições. Nessa parte é que está a diferença. Para determinar a quantidade de resultados que deverão ser considerados repetidos, devemos imaginar um novo caso à parte, ou seja, uma situação que conte quantas maneiras diferentes podemos sortear a mesma dupla. Necessariamente, para este novo caso, estamos falando de calcular uma permutação com duas opções e dois sorteios. Por fim, basta dividir o resultado do cálculo feito pela **Figura 11.7** pelo cálculo feito pela **Figura 11.8**. Vejamos como fica, matematicamente falando, na **Figura 11.9**:

$$\text{Figura 7: } A_{5,2} = 5 \cdot 4 = 20$$

$$\text{Figura 7: } AR_{2,2} = 2! = 2 \cdot 1 = 2$$

$$\text{Final: } \frac{\text{Figura 7}}{\text{Figura 8}} = \frac{20}{2} = 10$$

**Figura 11.9:** Cálculo de combinação: 5 candidatos para 2 vagas.



Aproveitando o conceito de fatorial, é possível escrever uma fórmula que calcule as combinações em um caso com  $n$  candidatos e  $r$  vagas:

$$C_{n,r} = \frac{n!}{r!(n-r)!}$$

Vale lembrar que esta fórmula, assim como o método manual que estamos desenvolvendo é apenas uma de várias maneiras de se chegar ao resultado final. Fica a escolha de cada um por qual optar. Contudo, quanto mais opções disponíveis melhor.

Optando, agora, pelo caso longo e penoso dos 20 candidatos para 5 vagas em uma equipe para recitar poemas, o cálculo de quantas equipes diferentes podem ser formadas será ágil e prático. Basta calcular a quantidade de arranjos para 20 opções em 5 sorteios para ter o total de equipes, mesmo com formações iguais. Depois, a permutação de 5 vagas com 5 opções para determinar de quantas ordens diferentes podemos sortear as mesmas 5 pessoas para uma equipe de 5 vagas. Por fim, faremos a divisão para obter o resultado das equipes que, de fato, são diferentes, isto é, após eliminadas as repetições. Vejamos na próxima figura:

$$\text{Passo 1: } A_{20,5} = 20 \cdot 19 \cdot 18 \cdot 17 \cdot 16 = 1.860.480$$

$$\text{Passo 2: } AR_{5,5} = 5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$$

$$\text{Final: } \frac{\text{Passo 1}}{\text{Passo 2}} = \frac{1.860.480}{120} = 15.504$$

**Figura 11.10:** Cálculo de combinação: 20 candidatos para 5 vagas.

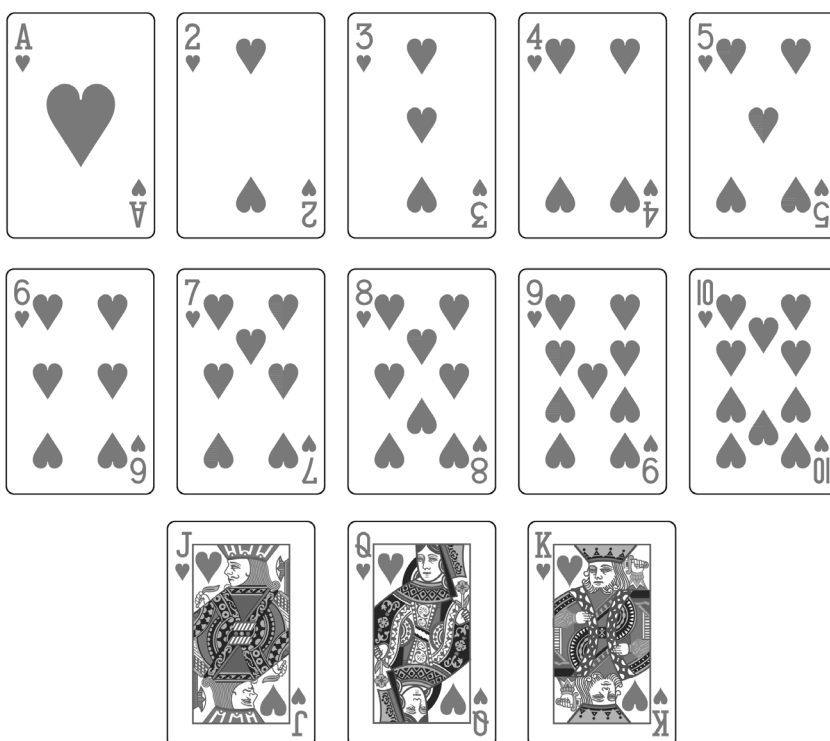
Chegamos ao resultado de 15.504 equipes diferentes de 5 integrantes com 20 candidatos disponíveis. O que mais importa para este caso é que não foi necessário montar todas as combinações e retirar as repetidas como foi feito no caso dos irmãos. Intuitivamente, montamos um modelo matemático que resolve esse tipo de problema de maneira mais dinâmica.

### Atividade 3

#### Atende ao objetivo 3

Um grupo de amigos, para ocupar o tempo, inventou um jogo de cartas. Nesse jogo, eles usam apenas um dos quatro naipes disponíveis com um total de 13 cartas (Ás, três figuras e nove números). O jogo é sempre entre dois jogadores.

O primeiro jogador pega uma carta do bolo embaralhado sem saber qual é e, depois, pega mais uma. Em seguida, o segundo jogador pega uma carta do mesmo bolo e a seguir pega mais uma. Agora, cada jogador tem duas cartas. Vence quem tiver maior pontuação com a soma das duas cartas, sendo que o Ás vale um ponto, as de figuras valem onze pontos e as de números valem seus respectivos valores impressos. Em caso de empate na soma, ganha o jogador que tiver a carta de valor mais alto – exemplo: 1 jogador tirou as cartas 1 e 4 e o outro jogador tirou as cartas 2 e 3. A carta com o valor 4 ganha.



Determine quantos pares de resultados diferentes o primeiro jogador terá à sua disposição nesse duelo de cartas.

[illegible]**Resposta comentada**

É fato que, nesse tipo de situação, a ordem da retirada das cartas não interfere no resultado final. Isto é: se o jogador retirar um Ás e a carta de número 10, isso fará o mesmo sentido que se tivesse tirado a carta de número 10 primeiro e o Ás em seguida. Contudo, não é a mesma coisa que tirar a carta 6 e a carta 5, apesar de também somar 11, pois o par que foi pedido no enunciado é diferente (Ás e 10 é um par, 6 e 5 é outro par). Logo, concluímos que estamos falando de um caso de combinação. Nesse caso, uma combinação de 13 opções com dois sorteios. A solução fica da seguinte forma:

*Passo 1:*  $A_{13,2} = 13 \cdot 12 = 156$

*Passo 2:*  $AR_{2,2} = 2! = 2 \cdot 1 = 2$

$$\text{Final: } \frac{\text{Passo 1}}{\text{Passo 2}} = \frac{156}{2} = 78$$

Desejando, o cálculo também pode ser feito pela fórmula que recorre ao fatorial:

$$C_{13,2} = \frac{13!}{2!(13-2)!} = \frac{13!}{2! \cdot 11!} = \frac{13 \cdot 12 \cdot 11!}{2! \cdot 11!} = \frac{13 \cdot 12}{2!} = 78$$

Logo, o primeiro jogador terá à sua disposição 78 pares diferentes de cartas. Agora, fica a sugestão para você praticar e determinar quantos pares o segundo jogador terá à disposição, logo após o primeiro retirar seu par.

## Permutações com repetições

É possível que tenham notado que alguns casos usados até agora admitem cenários perfeitos demais para serem resolvidos. Vamos então sofisticar nossa capacidade de calcular permutações, mas desta vez com elementos repetidos. Suponhamos novamente a brincadeira de anagrama, mas agora com a palavra PATA. Para que possamos identificar melhor como chegaremos ao resultado, matematicamente, iremos primeiro montá-lo manualmente. Vejamos na próxima figura, mas com o detalhe de que uma das letras A da palavra será grifada, de tal forma que possamos diferenciar uma da outra.

PATa PAaT PTaA PaAT PaTA

APTa APaT ATaP ATPa AaPT AaTP

---

TPAa TPaA TAPa TAaP TaPA TaAP

aATP aAPT aPTA aPAT aTPA aTAP

**Figura 11.11:** Permutação com repetição: palavra “pata”.

Note que obtivemos 24 combinações supostamente diferentes. Isso porque, na realidade, a diferença está no posicionamento da letra *a* maiúscula e da letra *a* minúscula, que foram diferenciadas apenas por fins especiais. Se retornarmos a letra *a* minúscula para maiúscula igual às demais, notaremos que, na verdade, temos apenas 12 resultados diferentes, pois, por exemplo, as combinações PATa e PaTA ficarão exatamente iguais: PATA. O problema é que feito à mão, um exemplo simples como este é possível. Contudo, podemos nos deparar com palavras como *guanabara* ou *presidente*.

Voltemos, então, ao caso da PATA para tentarmos modelar uma maneira de calcular a quantidade de resultados diferentes. Inicialmente, fica indiscutível a necessidade de calcular a quantidade de resultados considerando que todas as letras sejam diferentes. Essa parte já foi mostrada na seção Permutações desta mesma aula. Podemos fazer manualmente ou recorrendo à fórmula conforme a próxima figura, pois o que importa é o resultado correto ao final:

$$\begin{array}{ccccccc} \_4\_ & \times & \_3\_ & \times & \_2\_ & \times & \_1\_ \\ m! = 4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24 \end{array}$$

**Figura 11.12:** Cálculo de permutação com repetição: palavra “pata”.

Agora foi calculada a quantidade total de resultados, mesmo sabendo que teremos repetições por conta de letras iguais. Iremos, enfim, eliminar as duplicidades. Para isso, podemos pensar pontualmente em um resultado apenas. Suponhamos a ordem TPaA (mantivemos uma letra minúscula apenas para enfatizar que mudaremos a ordem da vogal *a*, mas sem mudar o resultado). Note que existe um segundo resultado mantendo as consoantes T e P na ordem TP que é TPAa. Ao retornar a letra *a* minúscula para maiúscula, fica evidente que ambas formavam a mesma palavra. Logo, podemos induzir o raciocínio de que, a cada arrumação feita, teremos uma idêntica, apenas trocando a posição da letra *a*. Resta determinar como calcular esta duplicidade. Vejamos na figura que segue:

$$\begin{array}{ccccccc} \_T\_ & \times & \_P\_ & \times & \_2\_ & \times & \_1\_ \\ m! = 2! = 2 \cdot 1 = 2 \end{array}$$

**Figura 11.13:** Permutação: cálculo de duplicidade de letras iguais.

Veja que calculamos as duplicidades como se fosse um exemplo separado. Isto é: mantendo as vogais da palavra TAPA em uma posição fixa, consigo arrumar as duas vogais *a* em duas posições diferentes (mesmo que o resultado seja igual). Logo, metade dos resultados que obtivermos será repetido. Portanto, devemos dividir o total de combinações que fizemos, supondo que todas as letras são diferentes, por dois. Vamos para um exemplo mais complexo e daí consolidar o raciocínio.



Para casos de permutações nos quais temos elementos repetidos, o cálculo imediato pode também ser feito por meio de fórmula. Suponhamos uma palavra com  $n$  letras, da qual uma delas se repete  $m$  vezes. O cálculo se dá:

$$P_n^m = \frac{n!}{m!}$$

Caso outra letra se repita, por exemplo  $k$  vezes, basta considerar esta informação também no denominador.

$$P_n^{m;k} = \frac{n!}{m! \cdot k!}$$

Esta fórmula é usada para quantas letras repetidas existirem, contanto que a quantidade de repetições de cada uma delas seja considerada individualmente no denominador.

Suponhamos a palavra BANANA. Se considerarmos que todas as letras são diferentes, a quantidade de anagramas que poderemos fazer será 720. Como sabemos que as letras  $n$  e  $a$  se repetem, devemos eliminar os resultados iguais. A letra  $n$  aparece duas vezes. Logo, podemos mudá-la duas vezes de posição (como vimos em PATA) sem alterar o resultado. Com isso, dividiremos 720 por 2, chegando a 360 resultados. Contudo, ainda assim, temos as repetições da letra  $a$ . Precisamos calcular de quantas formas diferentes podemos mudar a letra  $a$  em um resultado sem alterar o resultado. Vejamos na **Figura 11.14**:

\_B\_ x \_N\_ x \_N\_ x \_3\_ x \_2\_ x \_1\_

$$m! = 3! = 3 \cdot 2 \cdot 1 = 6$$

**Figura 11.14:** Permutação com repetição: cálculo de duplicidade da letra a na palavra banana.

Notem que mantendo as letras *b* e *n* fixadas em uma posição, podemos mudar as letras *a* seis vezes de posição e, ainda assim, obter o mesmo resultado. Com isso, devemos dividir o resultado parcial que temos de 360 resultados por 6 para eliminar esses resultados iguais, chegando, enfim, ao valor de 60 maneiras diferentes de arrumar as letras da palavra BANANA.

#### ==== **Atividade 4** =====

##### *Atende ao objetivo 4*

Um gestor possui à sua disposição seis funcionários, sendo um gerente, um coordenador, dois analistas e dois estagiários. A sala onde trabalharão é estreita e comprida ao mesmo tempo. Por isso, as mesas ficarão enfileiradas uma ao lado da outra. As seis mesas são iguais. Determine de quantas formas diferentes, considerando apenas cargos, esse gestor pode arrumar seus funcionários na sala em questão.

---

---

---

---

---

---

---

---

---

---

---

---

**Resposta comentada**

Ao considerarmos que o gerente será representado pela letra G, o coordenador pela letra C, cada analista por A1 e A2 e os estagiários por E1 e E2, simularemos duas combinações:

G C A1 A2 E1 E2

G C A2 A1 E1 E2

Considerando os indivíduos um a um, é correto afirmar que temos duas arrumações diferentes na sala. Contudo, o gestor deseja analisar a arrumação, utilizando como parâmetros apenas os cargos. Logo, as duas arrumações são iguais. Baseados nesta conclusão, podemos afirmar que o exercício proposto é sobre permutações com repetições.

O caso em questão é composto por seis elementos, dos quais, se considerarmos que são todos distintos, teremos 720 combinações. Como não são todos diferentes à vista do gestor, precisamos retirar as repetições calculando de quantas formas diferentes podemos arrumar os analistas sem alterar o resultado e depois o mesmo com os estagiários.

G C \_2\_ x \_1\_ E1 E2

G C A2 A1 \_2\_ x \_1\_

Como ambos os cargos possuem a mesma quantidade de repetições, ambos terão o mesmo resultado. Logo, dividiremos o resultado parcial de 720 por 2 (referente aos analistas) e depois novamente por 2 (referente aos estagiários), obtendo como resposta 180 arrumações diferentes na sala, considerando apenas os cargos como diferencial.

Caso opte por resolver pela fórmula, basta considerar que são 6 elementos, sendo que um deles (A) se repete duas vezes e outro (E) se repete também 2 vezes. A resolução fica da seguinte forma:

$$P_n^{A;E} = \frac{n!}{A! \cdot E!} \therefore P_6^{2;2} = \frac{6!}{2! \cdot 2!} = 180$$

## Conclusão

É indiscutível que, para se determinar a probabilidade de um evento ocorrer, precisamos antes estipular quantas maneiras diferentes de ocorrências são possíveis para disputar com o evento que temos interesse. Por isso, o cálculo de todas as possibilidades é algo imprescindível para qualquer estudo probabilístico.

O cálculo de todas as possibilidades, como vimos, depende do formato e objetivo do caso em questão. Precisamos considerar se iremos escolher todos os elementos disponíveis, se elementos sobrarão e/ou serão repostos para serem escolhidos novamente e se existem elementos repetidos. Toda essa variedade de combinações permite que o cálculo em questão seja complexo, caso apenas não tenha compreendido a essência do método.

Desse modo, exatamente para não correr o risco de tornar algo que não precisa ser complexo em algo complicado, foi mostrado como se resolve cada situação nos dois moldes possíveis: manualmente e usando fórmulas. Como enfatizado algumas vezes, o método manual é considerado mais elegante, pois permite uma diversidade de montagens na elaboração, além de exercitar diretamente a compreensão e a interpretação do problema. Utilizando apenas fórmulas, o indivíduo corre o risco de ser apenas um repetidor de cálculos, fator que, com o tempo, pode comprometer a compreensão dos casos e induzir à utilização errada da fórmula.

Por fim, é importante deixar claro que o uso de fórmulas não é e nem deve ser recriminado. Trata-se de uma solução tão válida quanto o método manual. Contudo, o primeiro exige que o indivíduo se esforce mais, além de praticar mais raciocínio matemático. De qualquer forma, fica a sugestão de duas soluções, as quais podem ser usadas juntas, inclusive para confirmar os resultados obtidos.

## ===== **Atividade final** =====

*Atende aos objetivos 1, 2, 3 e 4*

Suponhamos uma urna na qual foram colocadas setes bolas: uma bola vermelha com o número 1, uma bola verde com o número 2, uma bola

azul com o número 3, uma bola vermelha com o número 4, uma bola amarela com o número 5, uma bola verde com o número 6 e uma bola vermelha com o número 7. Agora, diversas brincadeiras serão feitas!

a) Uma pessoa vai retirar uma bola com a mão direita e depois uma bola com a mão esquerda. Quantos números diferentes, de dois dígitos, podemos formar considerando o da mão direita a dezena e o da mão esquerda a unidade?

---

---

---

---

---

---

---

b) Uma pessoa vai retirando uma bola de cada vez e colocando sobre a mesa, uma ao lado da outra, sempre posicionando a última sorteada ao lado direito da anteriormente sorteada – até que acabem as bolas. Quantos números diferentes de sete dígitos podemos montar?

---

---

---

---

---

---

---

c) Três bolas são sorteadas seguidamente. Ao final, quem as sortear somar o valor dos números nelas gravados. Quantos números diferentes podemos somar nessa brincadeira?

---

---

---

---

---

---

---

d) Uma pessoa retira uma bola de cada vez, colocando sobre uma mesa e posicionando-as da esquerda para a direita, conforme são retiradas. Ao final, nota-se a sequência de cores estabelecida. Determine quantas sequências diferentes de cores podemos obter com essa brincadeira.

---

---



---



---



---



---



---

### Resposta Comentada

a) Para esse experimento em questão, temos um caso de *arranjo*, pois não sorteamos todas as bolas e a ordem com que elas são sorteadas faz diferença. Isto é: se sorteamos primeiro a bola de número 2 e depois a bola de número 5, teremos o número 25, mas se a ordem do sorteio for inversa, teremos o número 52. Além disso, temos de nos atentar de que não existe reposição. Logo, uma bola sorteada não é devolvida para as demais. Com isso, o cálculo pelos dois métodos apresentados será:

$$\underline{\quad}7\underline{\quad} \times \underline{\quad}6\underline{\quad}$$

$$A_{7,2} = 7 \cdot 6 = 42$$

b) Para este caso, todas as bolas são retiradas até acabarem. Elas não são repostas durante o sorteio e, o mais importante, a ordem do sorteio faz diferença, pois conforme vão sendo sorteadas, vão formando um número. Logo, uma sequência diferente. Para casos com essas características, usamos o que aprendemos para *permutações sem repetições*. O cálculo, fazendo pelos dois métodos será da seguinte forma:

$$\underline{\quad}7\underline{\quad} \times \underline{\quad}6\underline{\quad} \times \underline{\quad}5\underline{\quad} \times \underline{\quad}4\underline{\quad} \times \underline{\quad}3\underline{\quad} \times \underline{\quad}2\underline{\quad} \times \underline{\quad}1\underline{\quad}$$

$$7! = 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5.040$$

c) Este caso é mais sofisticado. Temos uma situação na qual apenas três bolas serão sorteadas e, ao final, seus valores serão somados. Aqui já fica claro que a ordem não vai interferir no resultado, pois se sorteamos nesta ordem as bolas 1, 2 e 3, a soma será 6. Igualmente, se tivéssemos sorteado nesta ordem as bolas 3, 2 e 1. Estamos falando necessariamente

de um caso de *combinação*. Para resolver no método manual, primeiro determinaremos o total de combinações possíveis, considerando que a ordem não irá interferir. Em seguida, calcularemos quantas maneiras diferentes as três bolas sorteadas podem ser escolhidas. Por fim, retiraremos estas repetições do total inicialmente calculado. Na próxima figura, temos estes dois passos, além do cálculo direto com fórmulas:

$$\text{Passo 1: } \underline{\quad 7 \quad} \times \underline{\quad 6 \quad} \times \underline{\quad 5 \quad}$$

$$\text{Passo 2: } \underline{\quad 3 \quad} \times \underline{\quad 2 \quad} \times \underline{\quad 1 \quad}$$

$$C_{7,3} = \frac{7!}{3!(7-3)!} = 35$$

d) Para esta última brincadeira, ao identificarmos que todas as bolas serão retiradas sem reposição, sabemos que é um caso de permutação. Cabe identificar se a ordem será preponderante para o resultado ou não. Note que, para esta brincadeira, o que está sendo considerado são as cores, não os valores. Logo, se sortearmos primeiro a bola de número 1 e depois a de número 4, o resultado será o mesmo que sortear primeiro a de número 4 e depois a de número 1, pois, na realidade, o que fizemos foi sortear primeiro uma bola vermelha e depois outra bola vermelha. Com isso, temos uma situação de *permutação com repetições*. O cálculo seguirá exatamente como feito no caso da palavra BANANA conforme a figura que segue:

$$\text{Passo 1: } \underline{\quad 7 \quad} \times \underline{\quad 6 \quad} \times \underline{\quad 5 \quad} \times \underline{\quad 4 \quad} \times \underline{\quad 3 \quad} \times \underline{\quad 2 \quad} \times \underline{\quad 1 \quad}$$

$$\text{Repetições Vermelhas: } \underline{\quad 0 \quad} \underline{\quad 0 \quad} \underline{\quad 0 \quad} \underline{\quad 0 \quad} \underline{\quad 3 \quad} \times \underline{\quad 2 \quad} \times \underline{\quad 1 \quad}$$

$$\text{Repetições Verdes: } \underline{\quad 0 \quad} \underline{\quad 0 \quad} \underline{\quad 0 \quad} \underline{\quad 0 \quad} \underline{\quad 0 \quad} \underline{\quad 2 \quad} \times \underline{\quad 1 \quad}$$

$$P_7^{3;2} = \frac{7!}{3! \cdot 2!} = 420$$

## Resumo

Nesta aula, aprendemos a calcular a quantidade de possibilidades que um cenário específico é capaz de gerar. Vimos e ratificamos que não existe um método único de fazê-lo, pois podemos trabalhar com a montagem manual ou diretamente com aplicações de fórmulas matemáticas.

Ainda no que diz respeito a métodos diferentes, podemos constatar que não somente o método possui diversificação. A maneira de lidar com as informações, de acordo com o caso que está sendo estudado, faz toda a diferença. Precisamos considerar sempre se todos os elementos disponíveis estão sendo aproveitados, se existe reposição ou se depois de selecionado este elemento fica indisponível para demais escolhas. Não bastante, existe uma sutileza que compromete toda a análise que é a parte da ordem das escolhas interferir no resultado.

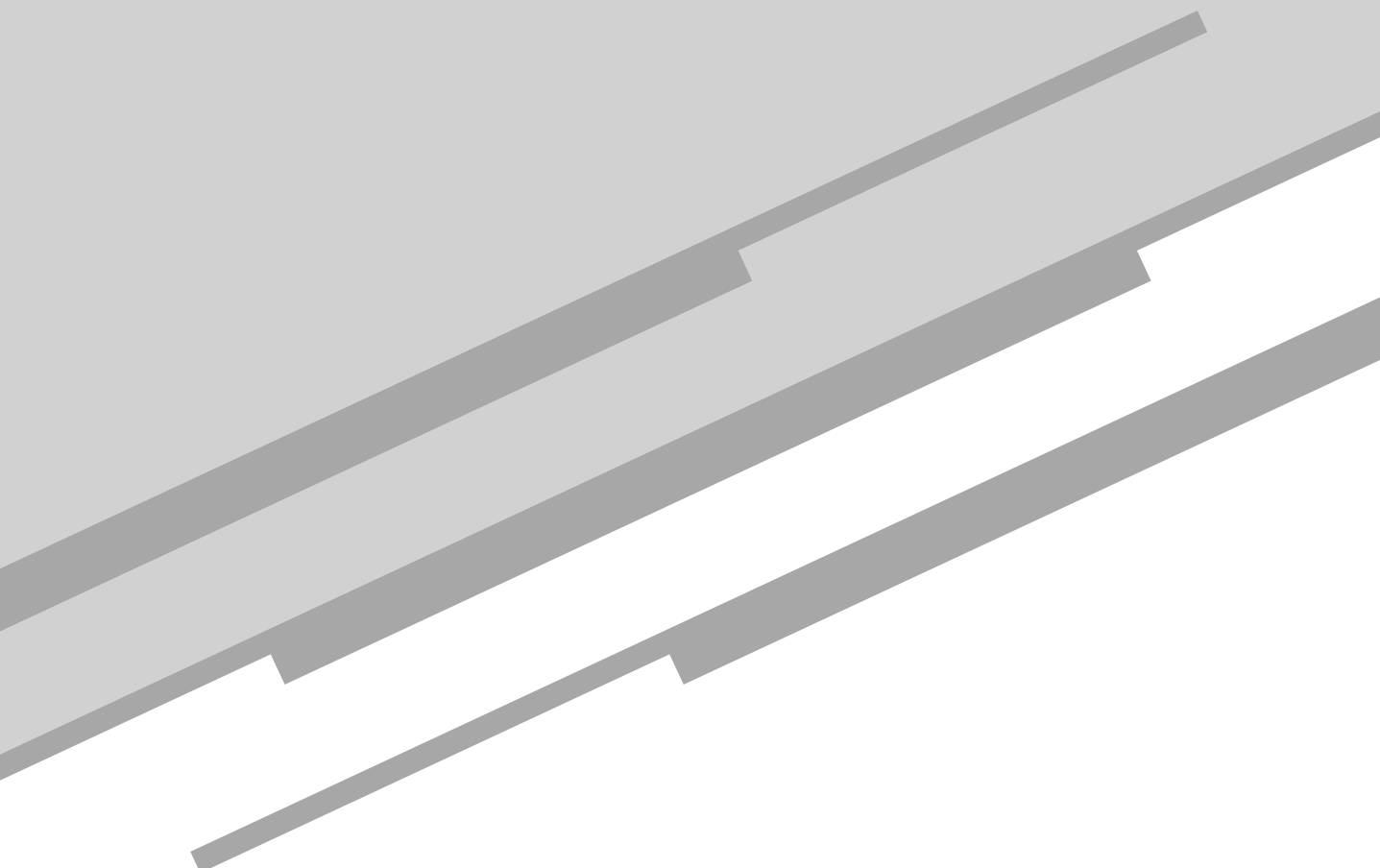
O que não podemos negar é que, fazendo uma leitura imparcial, a parte de cálculo de combinações, permutações e arranjos se trata de algo simplório, envolvendo operações básicas da matemática e, se assim optar, simples aplicações de fórmulas. A grande dificuldade, e esta precisa ser reconhecida, é a interpretação. Mais uma vez estivemos de frente com um estudo matemático no qual o seu maior obstáculo é o domínio da língua portuguesa e não a tão temida matemática.

## Informação sobre a próxima aula

Na próxima aula, enfim, iniciaremos os cálculos probabilísticos. Finalmente, nos aventuraremos no mundo de determinar a probabilidade de acontecer um evento. Isto se aplicará ao que mais interessa a você: que é aprender a calcular quais as reais chances de ganhar na loteria. Talvez não fiquem muito satisfeitos com os resultados, mas daí já não é culpa nossa. Um abraço!

# Aula 12

A sorte para todos



*Rafael Canellas Ferrara Garrasino*

## Meta

Apresentar a Probabilidade Básica como instrumento para se determinar as chances de ocorrência de fato ou resultado dentre diversas opções disponíveis.

## Objetivos

Esperamos que, ao final desta aula, você seja capaz de:

1. dominar o vocabulário específico e suas definições;
2. calcular a probabilidade marginal de um evento;
3. calcular a probabilidade combinada de um evento;
4. calcular a probabilidade condicional de um evento.

## Introdução

Como vimos na aula anterior, para que possamos determinar a *probabilidade* de qualquer tipo de evento, faz-se necessário primeiro determinar todas as possibilidades disponíveis para aquela ocasião em questão. Para esta aula, iremos montar previamente um caso, que será usado exaustivamente, para ilustrar cada situação que será apresentada.

Nesse caso, usaremos um artifício muito comum chamado *Tabela de Contingência*. Com esta, podemos explicitar todos os dados individualmente, além de ficar mais fácil consultar outras combinações de informações que forem convenientes. Suponhamos, então, que uma empresa de viagens resolveu contabilizar todos os clientes que atenderam no ano passado – detalhando a informação pelo sexo e faixa etária. Por motivos óbvios, cada cliente só poderia ser classificado como homem ou mulher. Já a faixa etária foi separada em quatro categorias: criança (até 12 anos), jovem (de 13 até 25 anos), adulto (de 26 até 55 anos), idoso (a partir de 55 anos). Assim, chegaram ao resultado que consta na **Tabela 12.1**.

**Tabela 12.1:** Classificação de clientes de empresa de viagens, por gênero e faixa etária

	Criança	Jovem	Adulto	Idoso	Total
Homem	85	113	230	217	645
Mulher	96	158	268	175	697

Note que, conforme dito anteriormente, a consulta é quase instantânea à *Tabela de Contingência*. Se quisermos saber o total de mulheres que usou o serviço da empresa de viagens, bastará olhar para o final da linha com o nome *Mulher* e identificar que foram 697. O mesmo, se quisermos saber o total de idosos, ou seja, ao final da coluna com o nome *Idoso*, indica-se um total de 392. De igual modo, há informações combinadas como: a quantidade de jovens do sexo masculino é encontrada ao cruzarmos a coluna *Jovem* com a linha *Homem*. Encontramos um total de 113 jovens do sexo masculino.

Em vista do exposto, ressaltamos o propósito de construirmos o vocabulário específico, bem como de desenvolvermos a capacidade de discernimento quanto a cada cenário, para que seja possível prevermos com qual qualidade de probabilidade estaremos lidando. Apresentare-

mos e mostraremos, então, o processo de cálculo de cada tipo de probabilidade, de acordo com a situação exigida.

Agora, com este caso apresentado, podemos, enfim, iniciar a teoria prevista para esta aula.

## Vocabulário

O vocabulário específico é sempre necessário porque iremos, por diversas vezes, recorrer a termos deste mesmo vocabulário para podermos deixar claro do que estamos falando. Conforme poderão ver mais à frente, a troca de uma única palavra em uma frase mudará completamente a forma de se enxergar o caso em questão. Logo, de maneira análoga, o mau entendimento de uma única palavra será capaz de comprometer todo o estudo em questão. Vamos aos termos!

Cada resultado possível dentre os que temos interesse em um estudo é chamado de *evento*. No caso criado, no início desta aula, temos vários eventos: ser homem, ser mulher, ser idoso, ser criança, e assim por diante. Note que o evento, necessariamente, possui apenas uma característica para se distinguir dos demais. Contudo, é possível que um evento tenha mais de uma característica. Neste caso, ele passa a se chamar *evento combinado*. Podemos exemplificar com idosos do sexo feminino. Esse tipo de evento possui duas características, pois o indivíduo em questão é um idoso e é uma mulher ao mesmo tempo. Existem outros diversos nesse mesmo caso: criança do sexo masculino, jovem do sexo feminino, e assim por diante. Para o tal caso em questão, por motivos óbvios, não existe a possibilidade de um evento combinado possuir mais de duas características, pois não é possível que um indivíduo seja homem e mulher ao mesmo tempo ou faça parte de duas faixas etárias simultaneamente. Entretanto, em outras situações, é possível que tenhamos eventos combinados com mais de duas características.

As características de um evento se comportam de tal maneira que, ao enumerá-las, podemos montar todas as possibilidades disponíveis em relação a elas. No caso construído, ao falarmos de homem e mulher, temos todas as possibilidades referentes à característica sexo. Sendo assim, diz-se que “homem” é o *complemento* de “mulher” nesse quesito, assim como “idoso” e “jovem” são os complementos para “criança” e “adulto” no quesito faixa etária.

A coletânea de todos os eventos possíveis de um estudo leva o nome de *espaço amostral*. Logo, no caso desta aula, temos um espaço amostral composto por crianças, jovens, adultos, idosos, mulheres e homens.

Neste contexto, ainda dentro do que se refere aos complementos, surgem dois termos imprescindíveis para resumir, de maneira rápida, o comportamento dos eventos de um espaço amostral: coletivamente exaustivos; mutuamente excludentes. Deste modo, quando temos todos os eventos que sozinhos descrevem por inteiro o espaço amostral, dizemos que eles são *coletivamente exaustivos*. Ao falar que no meu espaço amostral temos homens e mulheres, no que se refere ao sexo, eu já illustrei todas as possibilidades existentes. Logo, estas duas características são coletivamente exaustivas. Já quando falamos de duas características que não podem ocorrer ao mesmo tempo, dizemos que são *mutuamente excludentes*. Retornando ao caso desta aula, pode-se afirmar que “idoso” e “criança” são mutuamente excludentes, pois um mesmo indivíduo não pode ser uma criança e um idoso ao mesmo tempo.

## Atividade 1

### Atende ao objetivo 1

Suponhamos que você pretenda fazer um estudo com um baralho tradicional de quatro naipes (ouros, paus, espadas e copas) com treze cartas cada naipe. Neste caso, temos números (2 até 10), figuras (valetes, dama, rei) e o ás. Usando esse baralho como objeto para exemplificar, apresente o que se pede:

- a) três exemplos de evento;
- b) quatro exemplos de eventos combinados;
- c) exemplo de complemento;
- d) exemplo de mutuamente excludente;
- e) exemplo de coletivamente exaustivos;
- f) exemplo de espaço amostral.

---



---



---



---

---

---

---

---

---

---

---

---

---

---

### **Resposta comentada**

- a) Aqui, precisamos exemplificar características únicas em uma carta do baralho, dentre elas: ser de copas; ser de um naipe preto; ser um número; ser uma figura; ser ás etc.
- b) Para esta, precisamos combinar, pelo menos, mais de uma característica: ser um número de cor vermelha; ser uma dama de espadas; ser um número par e maior que cinco; ser um número preto do naipe de paus menor que sete e múltiplo de três etc.
- c) Precisamos responder características que se complementam: ser preta ou vermelha; ser número; figura ou ás etc.
- d) Aqui, precisamos de características que não podem acontecer ao mesmo tempo, como: ser preta e vermelha; ser número e figura; ser de copas e de ouros etc.
- e) Esta opção se assemelha bastante ao complemento. Portanto, as respostas também se encaixam em ambas as perguntas.
- f) Obrigatoriamente, neste exercício, só temos um espaço amostral, que é o próprio baralho em questão.

---

---

---

---

### **Probabilidade marginal**

A *probabilidade marginal* de um evento nada mais é do que um comparativo direto entre a quantidade de vezes que o evento do seu interesse ocorre contra a quantidade total de eventos disponíveis no espaço amostral. Para tal, é considerado que cada evento que compõe o espaço amostral possua iguais chances de ocorrência. Logo, se considerarmos

$X$  como a quantidade de ocorrências do evento do seu interesse e  $T$  o total de ocorrências do espaço amostral, a fórmula que calcula a probabilidade marginal é a ilustrada na **Figura 12.1**:

$$P = \frac{X}{T}$$

**Figura 12.1:** Fórmula de cálculo da probabilidade marginal.

Assim, obrigatoriamente, a quantidade de eventos que são do seu interesse será um número menor ou igual ao espaço amostral – por conta, inclusive, da própria definição de espaço amostral. Logo, podemos afirmar que o numerador dessa fórmula será sempre menor ou igual ao denominador, obrigando o resultado a ser, necessariamente, um valor que pode variar de 0 a 1. Esse resultado é a *probabilidade marginal* do evento, representada no formato decimal, que depois deve ser passada para percentual, multiplicando por 100%.



Apesar do intuito de dar ênfase ao que está sendo dito, muitos jogadores de futebol cometem um erro básico ao afirmar que as chances serão de 120% ou algo do tipo. O resultado da probabilidade de um evento, necessariamente, varia entre 0 e 1; logo, entre 0% e 100%. Qualquer resultado diferente disso é bobagem, pois sabemos que um evento com 0% de ocorrência é um evento impossível e, com 100% de ocorrência, é um evento certo.

Ao retornarmos ao caso desta aula, vejamos qual será a probabilidade de, ao pegar uma ficha aleatória de todos os clientes estudados, selecionar uma pessoa do sexo feminino. Precisaremos, para este cálculo, da quantidade de eventos *mulher* e da quantidade total de eventos do espaço amostral. Consultada a tabela de contingência, vemos que são 697 mulheres e 1.342 clientes no total. O cálculo ficará conforme a **Figura 12.2**:

$$P = \frac{X}{T} = \frac{697}{1.342} = 0,5194 = 51,94\%$$

**Figura 12.2:** Cálculo de probabilidade marginal: ficha de cliente mulher.

Portanto, existem quase 52% de chances de se selecionar uma ficha de uma cliente mulher ao acaso. É importante que, conforme a pessoa domine os conceitos envolvidos, ela seja capaz de prever se cometeu algum deslize durante os cálculos. Encontrar mais do que 100% de chances já foi citado como um erro grave, mas o caso que acabamos de calcular, apenas olhando para a tabela de contingência, já era visível, uma vez que tinha mais mulheres do que homens. Portanto, obter um resultado inferior a 50% indicaria um possível erro durante os cálculos.

Neste cenário, ainda falando do mesmo caso, vejamos agora as chances de se selecionar ao acaso a ficha de um cliente que seja idoso. Pela tabela de contingência, temos um total de 392 idosos e os mesmos 1.342 clientes no espaço amostral. Vejamos a **Figura 12.3** com o cálculo desse exemplo:

$$P = \frac{X}{T} = \frac{392}{1.342} = 0,2921 = 29,21\%$$

**Figura 12.3:** Cálculo de probabilidade marginal: ficha de cliente idoso.

Logo, existem quase 30% de chances de se selecionar uma pessoa idosa dentre os clientes cadastrados. Note que, se tivéssemos calculado também a probabilidade de escolher um cliente criança, depois um cliente jovem e, por fim, um cliente adulto, teríamos a probabilidade do cliente adulto como a maior das quatro – por ser o evento com maior incidência, depois seguido pelo idoso, que já calculamos. Outro fator que devemos considerar é que essas quatro faixas etárias são mutuamente excludentes e coletivamente exaustivas. Portanto, pela definição destes dois termos, elas juntas formam todo o espaço amostral – o que, por consequência, significa que a soma das probabilidades das quatro, necessariamente, precisa totalizar 100%. O mesmo aconteceria se somássemos a probabilidade de escolher um homem com a probabilidade de escolher uma mulher.

## Atividade 2

*Atende ao objetivo 2*

Com base no baralho padrão citado na Atividade 1, determine a probabilidade de escolher ao acaso:

- a) carta da cor vermelha;
- b) carta do naipe de espadas;
- c) carta que seja um ás;
- d) carta que seja um número par.

This is a blank sheet of white paper with horizontal blue or grey ruling lines, typical of notebook paper. The lines are evenly spaced and run across the width of the page. There is no handwriting or other markings on the paper.**Resposta comentada**

- a) O baralho tem 52 cartas, sendo 26 de cor vermelha e 26 de cor preta. Logo, metade é vermelha e metade é preta. Portanto, são 50% de chance. Outra maneira de calcular é considerar que existem duas cores no total: uma é preta e outra é vermelha. Ambas, utilizando a fórmula, resultarão nos mesmos 50%.
- b) Existem 13 cartas do naipe de espadas contra 52 no total do baralho ou, se preferir, existe o naipe de espada contra 4 naipes do baralho. Ambos os casos, quando calculados na fórmula, indicarão 25%.
- c) Existem 4 ases no baralho, um de cada naipe, contra um total de 52 cartas no baralho. Utilizando a fórmula, chegaremos ao resultado de 7,69%.

d) No baralho, em cada naipe, existem 5 cartas de número par (2, 4, 6, 8 e 10). Logo são, no total, 20 cartas de número par contra um total de 52 cartas no baralho. Pela fórmula, o resultado será 38,46%.

---

## Probabilidade combinada

Como vimos no início desta aula, um evento pode ter mais de uma característica, formando, assim, um evento combinado. Então, intuitivamente, quando lidarmos com uma situação deste tipo, não calcularemos mais uma probabilidade marginal, mas, sim, uma *probabilidade combinada*. O procedimento é basicamente o mesmo da probabilidade marginal. Contudo, iremos considerar a combinação dos eventos que nos interessam para determinar quantas vezes eles ocorrem. A fórmula será a mesma: iremos comparar a quantidade de eventos que nos interessam contra a quantidade total de eventos disponíveis.

Dentro do caso desta aula, existem inúmeras formas de ilustrar esta situação. Suponhamos que desejamos calcular a probabilidade de escolher, ao acaso, a ficha de um cliente que seja um homem jovem. Agora, não adianta apenas pegar a quantidade de homens ou apenas a quantidade de jovens, pois precisamos necessariamente considerar a quantidade de homens que são jovens; ou jovens que são homens (a recíproca é válida na probabilidade). Olhando para a tabela de contingência, temos 158 ocorrências e os mesmos 1.342 no total. A **Figura 12.4** ilustra o cálculo:

$$P = \frac{X}{T} = \frac{158}{1.342} = 0,1177 = 11,77\%$$

**Figura 12.4:** Cálculo de probabilidade combinada: homem jovem.



Existe uma propriedade especial para a probabilidade combinada. Assim, em casos nos quais os eventos são mutuamente excludentes e coletivamente exaustivos, a soma das probabilidades condicionais com uma mesma característica em comum resulta na probabilidade marginal desta característica repetida. No caso desta aula, se somarmos a probabilidade condicional de ser homem jovem com a probabilidade condicional de ser mulher jovem, teremos a probabilidade marginal de ser apenas jovem. O mesmo, se somarmos as probabilidades condicionais de ser mulher criança, mulher jovem, mulher adulta e mulher idosa, ou seja, teremos a probabilidade marginal de ser apenas mulher.

Em vista disso, suponhamos que desejássemos calcular a probabilidade de escolher ao acaso uma mulher idosa. Temos aqui um evento com duas características: ser mulher e ser idosa. Consultando a tabela de contingência, veremos que são, no total, 175 ocorrências deste evento. O resultado será de acordo com a **Figura 12.5**:

$$P = \frac{X}{T} = \frac{175}{1.342} = 0,1304 = 13,04\%$$

**Figura 12.5:** Cálculo de probabilidade combinada: mulher idosa.

Note que, deste modo, são 13,04% as chances de se escolher uma mulher idosa e que, conforme calculamos anteriormente, as chances de escolher apenas um idoso é de 29,21%. Portanto, pela propriedade da probabilidade combinada, a soma das probabilidades combinadas relacionadas a ser idoso (mulher idosa e homem idoso), necessariamente soma à probabilidade marginal de ser apenas idoso. Logo, subtraindo 13,04% de 29,21%, obteremos o resultado de 16,17%, que é a probabilidade combinada de ser homem e idoso. Desejando, pode fazer pela fórmula, que obterá o mesmo resultado.



com características que nos atendem contra 52 do baralho completo. Logo, pela fórmula, temos a probabilidade combinada de 19,23%.

e) Em cada naipe vermelho só existe um ás, totalizando 2 cartas que nos interessam contra as 52 do baralho completo. Logo, pela fórmula, a probabilidade é de 3,85%.

## Probabilidade condicional

Em algumas situações, já sabemos que um dos eventos que nos é interessante, de fato, ocorreu. Logo, por consequência, restringimos o nosso espaço amostral. Para casos como esse, trabalhamos com o que é chamado de *probabilidade condicional*. Nela, estamos interessados em um evento combinado, mas, como um deles é certo que ocorreu, restringimos o espaço amostral de tal forma que outros eventos que são mutuamente excludentes a eles não comprometam o resultado.

Vamos exemplificar com o caso que estamos usando nesta aula. Assim, suponhamos que você pretenda retirar uma ficha aleatória dos clientes. Quais as chances de essa ficha ser de uma criança mulher, sendo que já sabe que é de uma mulher? Note que, ao afirmar que já sabemos que se trata de uma ficha de uma mulher, não faz sentido considerar os homens. Iremos, agora, considerar apenas todas as faixas etárias possíveis, mas apenas as mulheres como espaço amostral.

De forma intuitiva, pela explicação de como faríamos no exemplo recém-citado, é possível montar a fórmula para o cálculo da probabilidade combinada. Temos dois eventos marginais, que chamaremos de  $A$  e  $B$ . Queremos saber a probabilidade de acontecer um evento com as características  $A$  e  $B$ , sendo que  $B$  é certo. A fórmula fica conforme a **Figura 12.6**:

$$P(A \mid B) = \frac{P(A \text{ e } B)}{P(B)}$$

**Figura 12.6:** Fórmula de cálculo da probabilidade condicional.

A fórmula se lê: “*probabilidade de acontecer A e B, sabendo que B aconteceu*”. Ela pede que seja dividida a probabilidade combinada de  $A$  e  $B$  pela probabilidade marginal de  $B$ .

Assim, retornando ao exemplo criado, o que acabamos de chamar de  $A$  e  $B$  serão, respectivamente, os eventos criança e mulher. Então, fazendo a leitura da fórmula com essas informações, temos “a probabilidade de ser criança e mulher, sabendo que é mulher”. Agora vamos calcular a probabilidade combinada criança e mulher e a probabilidade marginal mulher. A probabilidade marginal mulher já foi calculada anteriormente, inclusive na **Figura 12.2**, e é de 51,94%. A probabilidade combinada criança e mulher não foi calculada ainda. Então, deveremos calcular, conforme a **Figura 12.7**, ressaltando-se que se usa o mesmo procedimento aqui ensinado.

$$P = \frac{96}{1.342} = 0,0715 = 7,15\%$$

**Figura 12.7:** Cálculo de probabilidade combinada: “criança e mulher”.

Agora, com as duas probabilidades necessárias para calcular a probabilidade condicional, basta aplicar os resultados à fórmula, conforme a **Figura 12.8**, na qual  $C$  é *ser criança* e  $M$  é *ser mulher*.



Com a possibilidade de se manipular algebricamente a fórmula da probabilidade condicional, podemos definir novas fórmulas que também calculem a probabilidade marginal de um evento  $B$  e que calcule a probabilidade combinada de  $A$  e  $B$ :

$$P(B) = \frac{P(A \text{ e } B)}{P(A \text{ I } B)}$$

$$P(A \text{ e } B) = P(A \text{ I } B) \cdot P(B)$$

Fica agora por sua conta qual método usar. O que importa é ter muitos à disposição.

$$P(C \mid M) = \frac{P(C \cap M)}{P(M)} = \frac{0,0715}{0,5194} = 0,1377 = 13,77 \%$$

**Figura 12.8:** Cálculo de probabilidade condicional: criança e mulher.

É importante destacar que quando já temos as probabilidades combinadas e marginais necessárias em mãos, o uso desta fórmula, dessa maneira, é prático. Contudo, quando estamos partindo do zero, podemos apenas inserir os valores diretamente da tabela de contingência para obter o mesmo resultado. Neste caso, para substituir a probabilidade combinada criança e mulher, usaremos o real valor para crianças e mulheres contido na tabela de contingência – que é 96. Para a probabilidade marginal de mulher, iremos usar a quantidade real de mulheres na tabela de contingência – que é 697.

### Atividade 4

*Atende ao objetivo 4*

Continuamos com o mesmo baralho padrão das atividades anteriores. Vamos responder qual é a probabilidade de retirar ao acaso:

- valeta de copas, sabendo-se que é uma carta de copas;
- carta de um número preto, sabendo-se que a carta é um número;
- figura de espadas, sabendo-se que é uma carta preta.

[illegible]

### **Resposta comentada**

Para as três questões desta atividade, iremos resolver pelo método mais completo, pois obriga a prática do cálculo de todas as probabilidades envolvidas. Contudo, optando pelo método mais ágil, basta ter atenção para se chegar ao mesmo resultado – com tolerância para uma casa decimal nos casos de arredondamento.

a) A probabilidade combinada valete e copas é calculada com 1 contra 52, que dá 1,92%. A probabilidade marginal copas é 1 contra 4 (ou 13 contra 52), que dá 25%. Fazendo o cálculo da probabilidade condicional de 1,92% contra 25%, temos 7,68%.

b) A probabilidade combinada de carta de número e preta é calculada com 18 contra 52, que dá 34,62%. A probabilidade marginal de ser número é calculada com 36 contra 52, que dá 69,23%. Logo, a probabilidade condicional de 34,62% contra 69,23% será de 50%.

c) Aqui, se interpretarmos de maneira literal, a condicional (carta preta) não faz parte da combinação dos eventos que nos interessam (figura e espada). Todavia, ela está implícita na informação ser espada – o que vai apenas fazer o cálculo um pouco sofisticado, mas nada diferente do original.

A probabilidade combinada figura e espada é calculada com 3 contra 52, que dá 5,77%. A probabilidade marginal de ser carta preta é a já conhecida 50%. A probabilidade condicional de 5,77% contra 50% é de 11,54%.



### **Conclusão**

Como dito na aula anterior, para se calcular a probabilidade de um evento é necessário conhecer seu espaço amostral e o quanto o evento do seu interesse se repete. Entretanto, vimos também nesta aula que o cálculo vai além disso. É necessário entender o que está sendo calculado, para saber como recolher os dados e proceder com eles. Para cada tipo de situação, teremos um tipo diferente de abordagem.

Deste modo, uma maneira de melhor perceber isso é calculando as tão prometidas chances de se ganhar na Mega-Sena. Primeiro, temos de entender quanto representa nosso interesse. Apesar de o jogo pa-

drão, que possui seis números, poder ser organizado em 720 ordens diferentes (ver **Figura 12.9**), todos eles representam o mesmo jogo, isto é, quem apostou 5, 10, 15, 20, 25 e 30 vai ganhar o mesmo prêmio que quem apostou 30, 25, 20, 15, 10 e 5, pois, no final, o que importa é o conjunto das dezenas. Logo, temos a favor do nosso interesse apenas uma possibilidade.

$$\text{—————} \quad \text{—————} \quad \text{—————} \quad \text{—————} \quad \text{—————} \quad \text{—————}$$

$$6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 6! = 720$$

**Figura 12.9:** Total de ordens diferentes do jogo Mega-Sena: aposta padrão de 6 números.

Faz-se necessário, agora, calcular o tamanho do espaço amostral contra o qual a nossa única chance de interesse irá disputar. Esse espaço amostral será composto por todas as combinações possíveis feitas com seis números dentre as 60 unidades disponíveis para a aposta. Lembro que, assim como para a aposta, para as combinações que irão compor o espaço amostral, a ordem não interfere no resultado final, isto é, o que importa é o agrupamento de seis combinações. Recordando, então, a aula anterior, temos aqui um caso de combinação de 60 números arrumados seis a seis. Vejamos na **Figura 12.10** como fica:

$$C_{60,6} = \frac{60!}{6!(60! - 6!)} = \frac{60!}{6! \cdot 54!} = 50.063.860$$

**Figura 12.10:** Cálculo do espaço amostral total do jogo Mega-Sena.

Como podemos ver, temos impressionantes mais de 50 milhões de combinações contra a única que nos interessa. Calculando a probabilidade, chegamos à constrangedora, ou obscena, chance de vitória de menos de 0,000002%. É algo motivador, não?

Já que estamos falando de notícias nada agradáveis, vale aproveitar o momento para acabar com um mito que se repete, mas não tem fundamento algum. No sorteio dos números da Mega-Sena, as bolinhas que representam cada número estão misturadas e são sorteadas aleatoriamente. Com isso, todas possuem a mesma chance de serem escolhidas. Portanto, é uma besteira sem tamanho falar “quase que acertei”, quando sai a dezena 15 e você apostou na dezena 14, por exemplo. Isso só faria

sentido se as bolinhas viessem correndo em ordem crescente em fileira e uma pessoa retirasse aleatoriamente. Como não é o que acontece, falar “quase” para quando sair o número 15 no lugar de 14, que você apostou, faz tanto sentido quanto falar “quase” para quando sair o número 15 no lugar de 58 que você apostou.

Ok ... prometo não estragar mais o seu dia!

## ===== **Atividade final** =====

### *Atende aos objetivos 1, 2, 3 e 4*

Foi feita uma pesquisa com 5.000 pessoas sobre o destino de sua próxima viagem. Essa pesquisa era composta por duas perguntas. A primeira questionava que tipo de principal motivação leva a pessoa a escolher o destino que tinha em mente. As opções eram praia, serra, floresta, cultural ou profissional.

A segunda pergunta consistia em saber se a pessoa estava indo ou não pela primeira vez ao tal destino. Em ambas as perguntas, cada entrevistado só podia escolher uma opção. Depois de feitas as perguntas, coletadas e organizadas as respostas, chegaram às seguintes conclusões:

- 664 pessoas disseram que iriam pela primeira vez a um destino de praia;
- 538 pessoas disseram que não iriam pela primeira vez a um destino de serra;
- 403 pessoas disseram que iriam pela primeira vez a um destino de floresta;
- 711 pessoas disseram que não iriam pela primeira vez a um destino cultural;
- 369 pessoas disseram que iriam pela primeira vez a um destino profissional.

Ainda:

- um total de 3.745 pessoas disse que o destino escolhido seria a primeira vez por lá;
- um total de 947 pessoas marcou o destino serra;
- um total de 754 pessoas escolheu o destino profissional;
- um total de 1.312 pessoas escolheu o destino praia.



---

---

---

---

---

---

---

---

### Resposta comentada

a) Arrumando as informações dadas e completando a tabela de contingência com as demais, temos o seguinte resultado ilustrado na tabela:

	Praia	Serra	Floresta	Cultural	Profissional	Total
1ª vez	664	409	403	401	369	3.745
2ª ou mais	648	538	472	711	385	1.255
Total	1.312	947	875	1.112	754	5.000

b) A probabilidade marginal floresta será determinada com 875 (total de pessoas que responderam floresta) contra 5.000 (total de pessoas entrevistadas). O resultado será 17,5%.

c) a probabilidade condicional de serra e primeira vez, sabendo-se que é a primeira vez que será calculada com a probabilidade combinada *serra e primeira vez* contra a probabilidade marginal *primeira vez*. Lembrando que pode ser feito pelo processo mais rápido, mas para praticar mais os cálculos, faremos da maneira mais completa.

A probabilidade combinada *serra e primeira vez* será através de 409 (quantidade de pessoas que irão pela primeira a vez à serra) contra 5.000 (total de pessoas entrevistadas), obtendo como resultado 8,18%. Já a probabilidade marginal de primeira vez será 3.745 (total de pessoas que irão para o destino escolhido pela primeira vez) contra 5.000 (total de pessoas entrevistadas), resultando em 74,9%. Logo, a probabilidade condicional será o resultado de 8,18% contra 74,9%, que é 10,92%.

d) A probabilidade combinada de praia e não ser a primeira vez será calculada com 648 (total de pessoas que irão para a praia, mas não será a primeira vez) contra 5.000 (nosso espaço amostral). Temos então como resultado 12,96%.

e) Outra situação de probabilidade condicional, sendo agora de destino cultural e não ser a primeira vez, sabendo que estamos falando de

alguém que não fará a viagem pela primeira vez. Precisaremos da probabilidade combinada de destino cultural e não ser a primeira vez e da probabilidade marginal não ser a primeira vez.

A probabilidade combinada de destino cultural e não ser a primeira vez será obtida com 711 (quantidade de pessoas que não irão pela primeira vez a um destino cultural) contra 5.000 (espaço amostral). A probabilidade será de 14,22%. Já a probabilidade marginal não ser a primeira vez é resultado de 1.255 (total de pessoas que escolheram um destino que não será a primeira vez) contra os 5.000 habituais, chegando a 25,1%. Assim, a probabilidade condicional questionada será encontrada após 14,22% contra 25,1%, que dá 56,65%.

---

---

## Resumo

Nesta aula, completamos, com o que vimos na aula anterior, como se calcula a probabilidade de ocorrer um evento. Ficou claro o que previmos anteriormente: a necessidade de primeiro arrumar todo o espaço amostral, para que possamos facilmente identificar a incidência dos eventos que nos interessam e os demais que estão competindo com ele.

Posteriormente, foi passado que existem três formas de se calcular a probabilidade de um evento, mas cada uma só se aplica a um tipo de situação específica. Quando estamos lidando com um evento que possui apenas uma característica, deveremos calcular a *probabilidade marginal* dele. Contudo, se este evento possuir mais de uma característica, estamos então lidando com um evento combinado. Logo, iremos calcular a *probabilidade combinada*. Por fim, vimos que podemos restringir o espaço amostral quando sabemos que uma das características de um evento combinado de fato ocorreu. Para tal, usaremos a *probabilidade condicional*.

Obviamente, não nos restringimos apenas a isso. Percebemos que conforme aumenta nossa intimidade com os conceitos envolvidos, mais fácil fica prever se cometemos algum equívoco durante o processo. Além disso, vimos também que podemos interagir resultados entre si, para otimizar nossos cálculos e pular etapas.

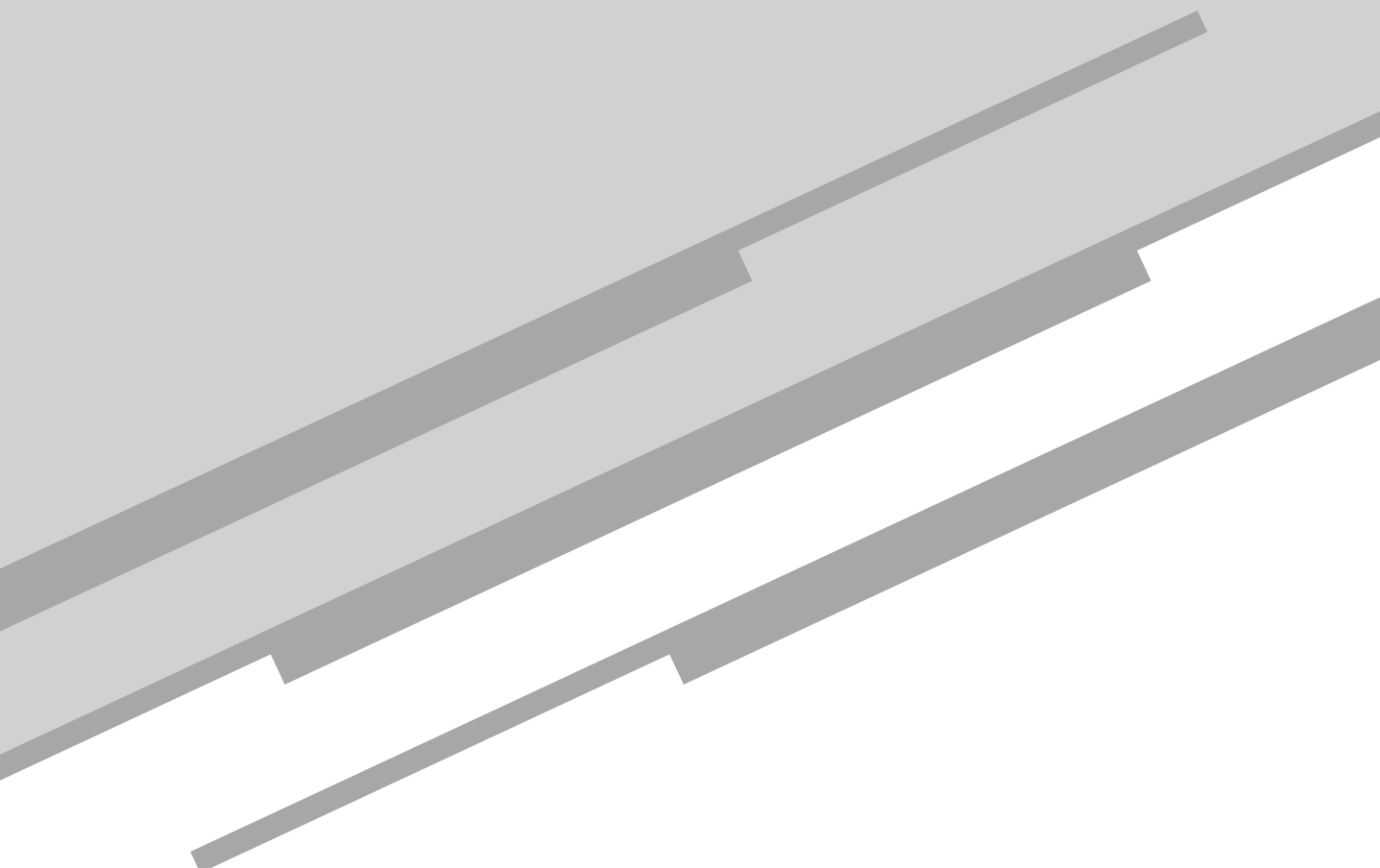
## **Informação sobre a próxima aula**

Na próxima aula, iremos lidar diretamente com um instrumento para destrinchar um estudo probabilístico. Com algumas informações e o uso correto desse instrumento, seremos capazes de obter todas as informações faltantes do estudo em questão e, assim, tirar conclusões mais completas sobre ele.

Então, nos vemos lá, que provavelmente será na próxima página.

# Aula 13

Cada macaco no seu galho



*Rafael Canellas Ferrara Garrasino*

## **Meta**

Apresentar o instrumento de probabilidade chamado Árvore de Decisão.

## **Objetivos**

Esperamos que, ao final desta aula, você seja capaz de:

1. desenvolver a Árvore de Decisão combinada;
2. elaborar a Árvore de Decisão condicional;
3. fazer uso, ao mesmo tempo, e de forma integrada, das Árvores Condicional e Combinada.

## Introdução

Na aula anterior, vimos como calcular a *probabilidade* de um evento ocorrer. Para tal, montamos um suposto cenário de pesquisa. Contudo, os cálculos, conforme eram feitos, ficavam relegados a respostas de exercícios. Não existia uma maneira de organizar essas informações para facilitar a leitura e rápida compreensão por parte de quem vai aproveitá-las para tomar alguma decisão. Daí surge a necessidade de algum tipo de instrumento ou técnica que nos ajude na organização desses novos dados, não somente no sentido de armazená-los, mas também no que se refere ao entendimento de todos eles.

Assim, obviamente, ter acesso a algum artifício que seja capaz de atender a todas essas necessidades citadas já é um grande benefício. Entretanto, se ele puder ser mais abrangente, oferecendo algum recurso adicional, será melhor ainda. De fato, este artifício existe. Trata-se de um instrumento chamado *Árvore de Decisão*. Ele se divide em dois modelos: combinada e condicional.

Desse modo, como foi mostrado na aula passada, existem alguns tipos de probabilidades, dentre elas: a probabilidade combinada e a probabilidade condicional. Exatamente por existirem formas diferentes de lidar com cada tipo de probabilidade (as duas citadas em questão), existem dois tipos de *Árvore de Decisão* – cada uma voltada para um tipo específico de probabilidade.

Temos então o seguinte cenário: foi feita uma pesquisa com todos os hóspedes de um hotel. Essa pesquisa era composta por duas perguntas. A primeira, o próprio entrevistador respondia, pois se tratava de identificar o sexo do entrevistado. A segunda era saber se o entrevistado estava visitando aquele destino pela primeira vez. Após feita a pesquisa, os dados foram resumidos de forma precária da seguinte forma:

- 332 homens estavam naquele destino pela primeira vez;
- foram entrevistadas 867 pessoas;
- um total de 331 pessoas estava visitando aquele destino pelo menos pela segunda vez;
- foram entrevistadas 378 mulheres.

De posse dos dados em questão, vamos agora trabalhar com relevantes conceitos para determinarmos:

- a) A Tabela de Contingência dessa pesquisa com todos os campos preenchidos;

- b) A probabilidade de um entrevistado ser mulher;
- c) A probabilidade de um entrevistado ser homem e já ter ido àquele destino;
- d) A probabilidade de um entrevistado ser mulher e estar indo pela primeira vez àquele destino;
- e) A probabilidade de escolher uma pessoa que está indo pela primeira vez àquele destino, sabendo que é um homem;
- f) A probabilidade de escolher uma pessoa que já foi antes àquele destino, sabendo-se que é mulher.

A partir de agora, mediante os desafios propostos, começaremos a explicitar o nosso raciocínio.

- a) A Tabela de Contingência completa deve ser como a tabela que vem a seguir:

**Tabela 13.1:** Modelo da Tabela de Contingência

	1ª vez	2ª ou mais	Total
Mulher	204	174	378
Homem	332	157	489
Total	536	331	867

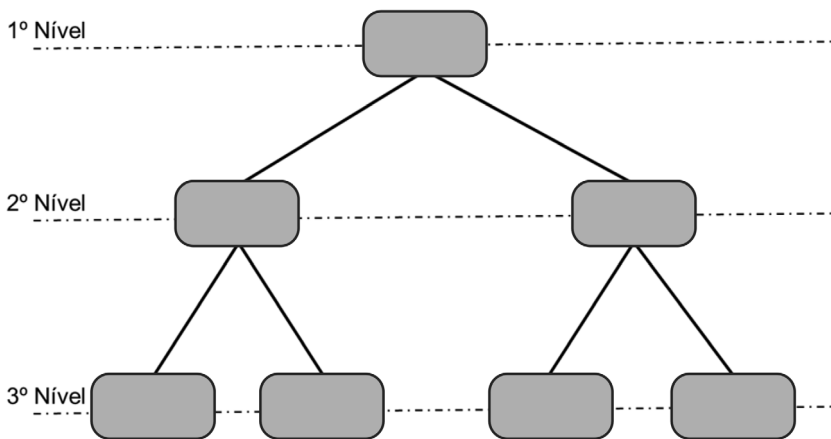
- b) No que diz respeito à probabilidade de um entrevistado ser mulher. Para tal, temos 378 mulheres contra um total de 867 entrevistados. Isto indica uma probabilidade de 46,6%.
- c) No que concerne à probabilidade de um entrevistado ser homem e já ter ido àquele destino, temos 157 homens que já foram lá antes contra 867 entrevistados. Isto dá uma probabilidade de 18,11%.
- d) Com relação à probabilidade de um entrevistado ser mulher e estar indo pela primeira vez àquele destino, temos 204 mulheres que foram pela primeira vez contra 867 entrevistados. Temos uma probabilidade de 23,53%.
- e) Por sua vez, quanto à probabilidade de se escolher uma pessoa que está indo pela primeira vez àquele destino, sabendo-se que é um homem, temos 332 homens que foram pela primeira contra 489 homens. Assim, a probabilidade é de 67,89%.

f) Por fim, quanto à probabilidade de se escolher uma pessoa que já foi antes àquele destino, sabendo-se que é mulher, temos 174 mulheres que já foram lá antes contra 378 mulheres. Há, então, a probabilidade de 46,03%.

Deste modo, após relembrarmos e fixarmos o conteúdo agora trabalhado, relativo à Aula 12, vamos, neste momento, nos deter à Árvore de Decisão, certo?

## Árvore de decisão

A Árvore de Decisão, conforme já falamos, possui diversos atributos e iremos discorrer sobre eles aos poucos. O primeiro atributo será o da *organização dos dados*. Inicialmente, iremos mostrar um modelo qualquer para que já tenham um primeiro contato e, assim, possam acompanhar os comentários feitos. Vejamos então a **Figura 13.1**:



**Figura 13.1:** Dados organizados que se desdobram em níveis.

A organização dos dados se dá de cima para baixo. No primeiro nível, o mais alto, estão todos os entrevistados (ou todos os dados) da pesquisa. Descendo um nível, temos um desdobramento das informações. Isto é: os dados que estavam todos consolidados no primeiro nível serão separados conforme uma característica qualquer estipulada pela pesquisa. Abaixo do segundo nível, no terceiro nível, temos um segundo desdobramento. Esse novo desdobramento respeita a primeira separação de acordo com a característica determinada e daí faz uma nova separação – baseada em uma nova característica. Habitualmente, não se trabalha com as linhas denominando os níveis. Contudo, neste primeiro contato, optamos por isto para facilitar a compreensão.

Obrigatoriamente, o primeiro nível não tem separação (círculo, retângulo ou bolinha, que depende da arte do pesquisador), pois ali estão todas as informações consolidadas. Os níveis seguintes não possuem um regimento sobre a quantidade de separações. Isto se dá porque depende, necessariamente, do tipo de característica que estaremos lidando.

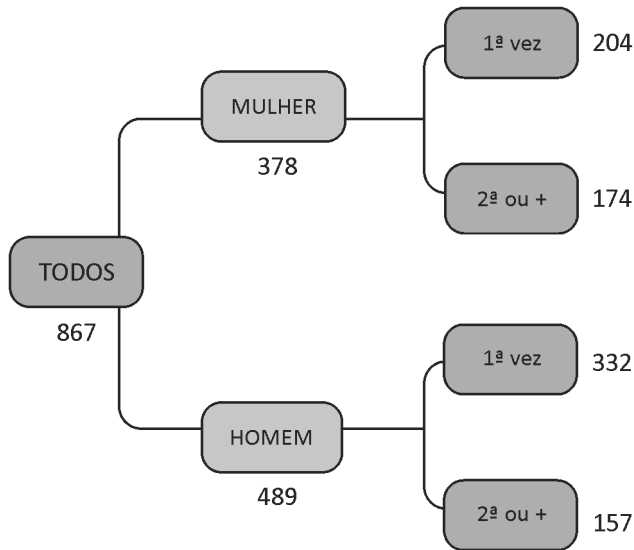
É importante frisar que, comumente, lidamos com árvores de decisão com até três níveis. Contudo, apesar de tornar mais complexo, é possível se deparar com modelos com mais níveis. Isso vai depender da quantidade de características associadas que estamos estudando. Por praticidade e padrão, iremos adotar o máximo de três níveis nesta disciplina.



**Atenção!** Para o estudo da Árvore de Decisão é estritamente importante que estejamos trabalhando com dados coletivamente exaustivos (porque iremos ilustrar todas as informações) e mutuamente excludentes (porque uma mesma informação contida em dois lugares ao mesmo tempo tornaria impossível a compreensão dos dados).

Por sua vez, outra particularidade que é muito comum das árvores de decisão é sua habitual *apresentação no formato vertical* (exatamente como na **Figura 13.1**). Em alguns momentos, é possível notar que estejam trabalhando com uma apresentação na horizontal. Na maioria das vezes isso se dá por limitação do espaço e, com isso, o uso horizontal torna mais prático o posicionamento das informações sem ocupar uma página inteira, por exemplo. Aqui, pelo motivo já dito, usaremos a apresentação horizontal.

Vejamos, então, o caso da pesquisa que trabalhamos no exercício apresentado na Introdução desta aula. Vamos, deste modo, inserir os dados da Tabela de Contingência na Árvore de Decisão, para que possamos praticar a organização dos dados. A **Figura 13.2** que ilustra o resultado final.



**Figura 13.2:** Inserção de dados da Tabela de Contingência na Árvore de Decisão.

Repare que cada informação foi lançada de acordo com o detalhamento do seu respectivo nível. No primeiro, temos todos os entrevistados. Logo, um total de 867 pessoas. No segundo, temos um desdobramento pela característica sexo: mulher e homem. Portanto, devemos lançar o total de mulheres (378) na sua respectiva informação e o total de homens (489) na sua também respectiva informação. Por fim, no último nível, temos o desdobramento de quem está viajando pela primeira vez ou indo pela segunda ou mais vezes. Aqui é que existe o detalhe de se respeitar o último desdobramento. No primeiro retângulo, ao topo do terceiro nível, temos a característica “1ª VEZ”. Mas ela está diretamente associada à característica ser mulher do nível anterior. Logo, lançaremos a quantidade de mulheres que está indo pela primeira vez àquele destino (204). Nos demais retângulos, abaixo desse, faremos pelo mesmo raciocínio.

Neste contexto, é esperado que uma ou outra pessoa comente que é algo muito parecido com a Tabela de Contingência. De fato é no que se refere ao lançamento dos dados. Contudo, a leitura das informações é mais dinâmica. Não precisa ficar cruzando linhas e colunas para achar a célula em comum que lhe interesse. Basta seguir a linha do desmembramento do seu interesse. Ainda assim, podemos afirmar que sequer

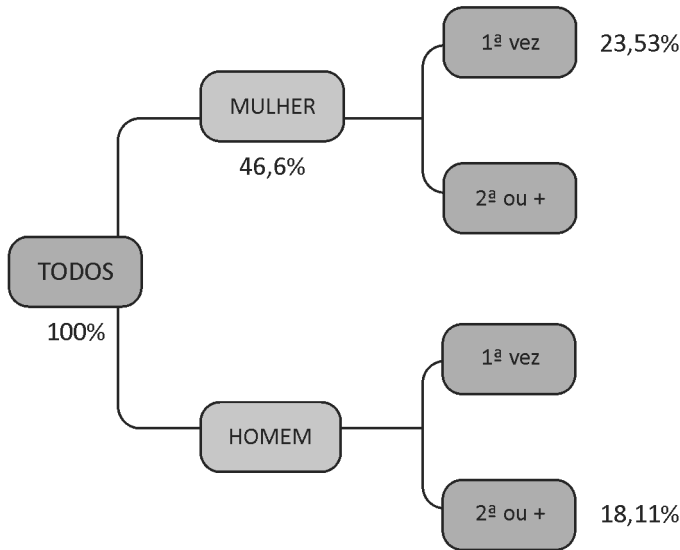
começamos a manusear a Árvore de Decisão, pois tudo que fizemos até agora foi um exemplo para ilustrar a organização e a leitura dos dados. A grande vantagem deste instrumento é que ele não vai trabalhar com os dados absolutos, mas, sim, com os dados percentuais (probabilísticos). Daí, surgirá a incontestável vantagem sobre a Tabela de Contingência, que continuará sendo útil como passo inicial para todas as pesquisas.

## Árvore combinada

Conforme falamos nesta e na aula anterior, dependendo da forma que a informação é apresentada, teremos um tipo diferente de *probabilidade*. Logo, nada mais justo que trabalharmos com uma específica Árvore de Decisão para cada tipo. Aqui, apresentaremos a Árvore de Decisão que lida exclusivamente com a *probabilidade combinada*. Ficando desde já a informação de que não podemos misturar as probabilidades em uma única Árvore de Decisão.

Voltaremos, então, às informações do exercício indicado na Introdução para usarmos como exemplo. Nele, calculamos diversas probabilidades após montarmos a Tabela de Contingência. Algumas dessas probabilidades eram combinadas, outras eram condicionais. As *probabilidades combinadas*, como sabemos bem, são as que lidam com duas ou mais características do nosso interesse. Já as *probabilidades condicionais* são as que restringimos à amostra com alguma característica certa previamente informada.

A primeira informação que aproveitaremos é a obviedade de que todos os entrevistados compõem 100% e, como já visto no exemplo sobre a montagem da Árvore de Decisão, esta informação ficará associada ao primeiro nível. Em seguida, no segundo nível, temos as duas informações: ser homem e ser mulher. Pelo que foi feito na letra b do exercício introdutório, temos que 46,6% dos entrevistados são mulheres. Seguindo para o terceiro nível, usaremos as informações da letra c e da letra d. Na letra c, temos a combinação de características ser homem e já ter visitado anteriormente aquele destino. Logo, o percentual de 18,11% será colocado no campo que remete a esta combinação. Depois, usando as informações da letra d, nas quais estão combinadas as informações ser mulher e estar visitando aquele destino pela primeira vez, colocaremos o percentual calculado de 23,53% no campo referente. Vejamos na **Figura 13.3** como ficará.



**Figura 13.3:** Probabilidade Combinada integrada à Árvore de Decisão – Árvore Combinada.

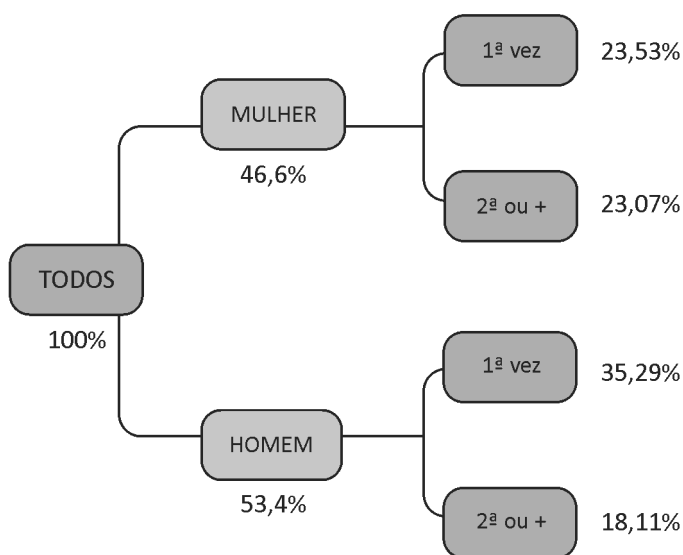
Como podem notar, colocamos apenas as informações que tínhamos em mãos. As demais poderão ser calculadas usando os dados da própria Árvore de Decisão ou os valores da Tabela de Contingência. Como o intuito desta aula é aprimorar o uso da Árvore de Decisão, iremos focar diretamente nela, mas, quando disponíveis os valores da tabela de contingência, ela pode ser utilizada de acordo com a conveniência do aluno.

O primeiro nível, que possui apenas uma informação, está completo. No segundo nível, temos uma informação faltando: a probabilidade de um entrevistado ser homem. Para determinar essa informação, usaremos a nosso favor a probabilidade já conhecida de ser mulher (46,6%). Se todos os entrevistados compõem 100% da amostra, logo, a partir do percentual de mulheres, concluímos que o percentual de homens é a diferença de 53,4%.

Deste modo, seguindo para o terceiro nível, temos dois campos a preencher. O localizado mais acima é o que se refere à combinação de ser mulher e já ter visitado aquele destino antes. Pois bem! Sabemos que 23,53% é o percentual de mulheres que nunca foram àquele destino e que 46,6% é o percentual de mulheres. Recordando o que falamos na aula anterior, vimos que as probabilidades combinadas totalizam a *probabilidade marginal* da característica que possuem em comum. Ora, a probabilidade que temos em comum é ser mulher, pois estamos comparando mulheres que já visitaram o destino antes com mulheres que

nunca visitaram o destino. Portanto, precisaremos apenas fazer a diferença do total de mulheres (46,6%) pela quantidade de mulheres que nunca visitaram o destino (23,53%). Tal operação nos dará o percentual de 23,07%.

Em vista disso, ainda no terceiro nível temos a “informação faltante” correspondente aos homens que nunca visitaram aquele destino antes. Elaboraremos para ele o mesmo raciocínio realizado no parágrafo anterior. Retiraremos o percentual já conhecido de 18,11% do percentual total de homens que é de 53,4% para obter o percentual que nos resta de 35,29%. Sendo assim, completaremos nossa Árvore de Decisão Combinada conforma a **Figura 13.4**:



**Figura 13.4:** Árvore de Decisão Combinada.

Caso esteja inseguro em relação aos seus resultados; ou apenas seja devoto de São Tomé (só acredita vendo), a *Árvore de Decisão Combinada* possui uma particularidade que permite confirmá-los. Todos os níveis, quando somadas suas respectivas probabilidades, totalizarão 100%. Basta testar no exemplo que acabamos de fazer e ficará confirmado.

## Atividade 1

*Atende ao objetivo 1*

Um hotel fez uma breve pesquisa sobre a satisfação do cliente. Nesta pesquisa, constavam duas perguntas: a primeira era para saber se o cliente estava satisfeito com o atendimento prestado. Para esta pergunta existiam duas opções: *Sim* e *Não*. Caso dissesse que sim, lhe era perguntado qual o item que o fez primeiro pensar nesta resposta. Eram dadas duas opções: *Equipe* e *Instalações*. Caso dissesse que não, lhe era perguntado qual o item que o fez primeiro pensar nesta resposta. Novamente lhe eram dadas as mesmas opções: *Equipe* e *Instalações*. Cada entrevistado não podia responder sim e não ao mesmo tempo. Assim como não era possível falar que, tanto Equipe quanto Instalações foram preponderantes na escolha.

Após feita a pesquisa e organizadas as informações, todos os dados foram inseridos em uma tabela de contingência. Contudo, o responsável por essa tabela acabou perdendo o arquivo, restando apenas os dados que seguem. De posse deles, monte a Árvore de Decisão Combinada com as informações disponíveis e complete com as faltantes:

- do total de entrevistados, 19,13% são pessoas insatisfeitas por conta da equipe do hotel;
- do total de entrevistados, 68,5% estão satisfeitos com o hotel;
- do total de entrevistados, 39,72% estão satisfeitos com as instalações do hotel.

[illegible]

---

---

---

---

---

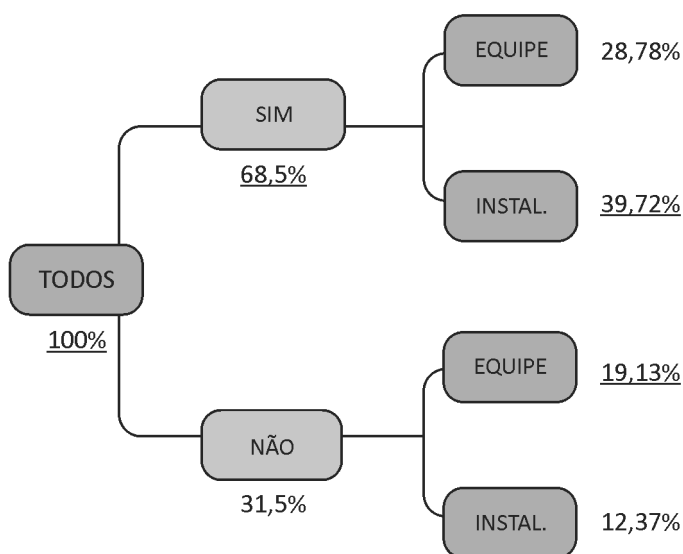
---

---

---

### Resposta comentada

A estrutura da Árvore de Decisão será conforme a figura apresentada a seguir. Pode, é claro, mudar a posição das características (*Sim* pode ficar embaixo, no segundo nível, ou *Equipe* embaixo no terceiro nível) de acordo com a sua conveniência. Os valores sublinhados são os que foram retirados diretamente do enunciado do texto, enquanto os demais foram calculados.



O percentual de *Não* foi obtido com a diferença de *Sim* (68,5%) com o total (100%). O percentual de pessoas satisfeitas com a equipe foi obtido com a diferença de pessoas satisfeitas com as instalações (39,72%) e o total de pessoas satisfeitas (68,5%). O percentual de pessoas insatisfeitas com as instalações foi obtido com a diferença de pessoas insatisfeitas com a equipe (19,13%) com o total de pessoas insatisfeitas (31,5%).

---

---

---

---

## Árvore condicional

Agora, igualmente como feito com a Árvore de Decisão Combinada, trabalharemos com a *Árvore de Decisão Condicional*. Contudo, é necessário falar que existe uma grande diferença entre as duas e esta diferença está no *terceiro nível*. Os dois níveis iniciais serão rigorosamente idênticos.

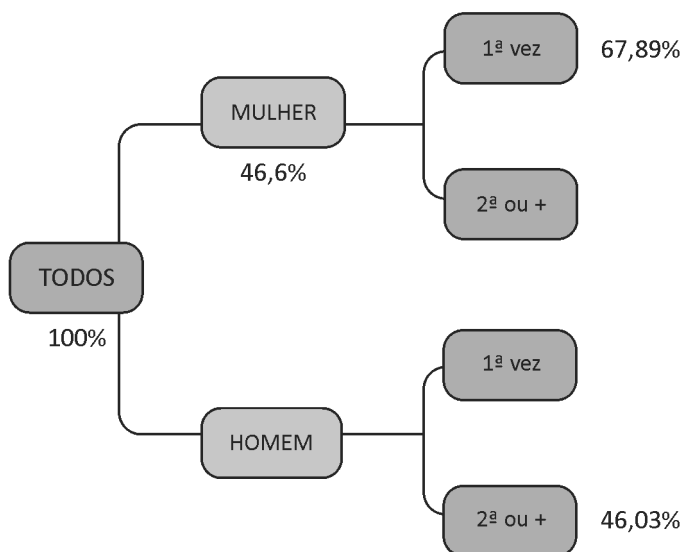
A diferença está no conceito da *probabilidade condicional*. Nela, restringimos o espaço amostral a uma característica a qual sabemos que de fato aconteceu. Isto é: é como se fizéssemos um corte no espaço amostral separando apenas os eventos que possuem a característica que afirmamos que aconteceu. Depois disso, faremos uma nova análise probabilística com as subdivisões que as demais características analisadas formarão. Mas, vamos primeiro preencher um exemplo de Árvore de Decisão Condicional para que, aos poucos, a diferença entre os dois tipos de árvores fique mais perceptível.

Iremos novamente adotar os dados do exercício apresentado na Introdução. Os primeiros passos seguirão como feito na árvore anterior. Adotaremos 100% para o espaço amostral completo no primeiro nível e 46,6% para a característica mulher do segundo nível. Em seguida, usaremos as duas últimas respostas desta atividade.

A penúltima informação que foi pedida no exercício introdutório foi a probabilidade de escolher uma pessoa que estava visitando aquele destino pela primeira vez, mas sabendo-se que é um homem. Isto é: pegaremos no terceiro nível a característica primeira vez associada à característica homem do segundo nível e atribuiremos 67,89% (porcentagem calculada no exercício).

Depois, de posse da última informação pedida, que era a probabilidade de escolher uma pessoa que já tenha visitado aquele destino antes sabendo-se que era mulher, atribuiremos o percentual calculado (46,03%) à característica segunda ou mais vezes do terceiro nível que está associada à característica mulher.

A próxima **Figura 13.5** ilustra como ficará este preenchimento inicial dos campos.

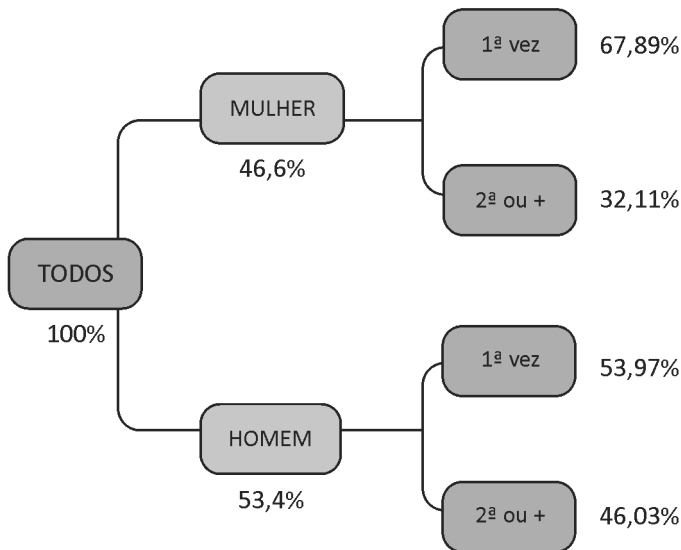


**Figura 13.5:** Árvore Condicional.

Agora iremos preencher os demais campos. O espaço destinado à característica homem será calculado da mesma maneira feita para a Árvore de Decisão Combinada (resultado foi 53,4%), pois já falamos que o primeiro e segundo níveis de ambas as árvores funcionam identicamente. Os dois campos restantes dependerão do entendimento do funcionamento, que é a diferença citada há pouco. Para tal, precisamos entender o que significa aqueles dois percentuais no terceiro nível da Árvore de Decisão que estamos estudando.

Deste modo, foi dito que 67,89% é a probabilidade de escolher uma pessoa que esteja indo pela primeira vez àquele destino sabendo-se que esta pessoa é um homem. Mas o que isto significa?

Isto significa necessariamente que, do total de homens entrevistados, 67,89% foram pela primeira vez àquele destino. Ora, se o terceiro nível significa uma divisão da informação do segundo nível (neste caso ser homem que acabou de ser separado para virar um novo espaço amostral) e sabemos que 67,89% representa uma parte dele, logo a outra parte restante é o que resta para completar o novo espaço amostral homem (lembrando que espaço amostral sempre equivale a 100%). Portanto, o valor faltante é 32,11%. Sendo assim, de maneira análoga, os 46,03% informados representam uma parte do novo espaço amostral (ser mulher). Com isto, o espaço a ser preenchido será a diferença de 53,97%. Vejamos na **Figura 13.6:**



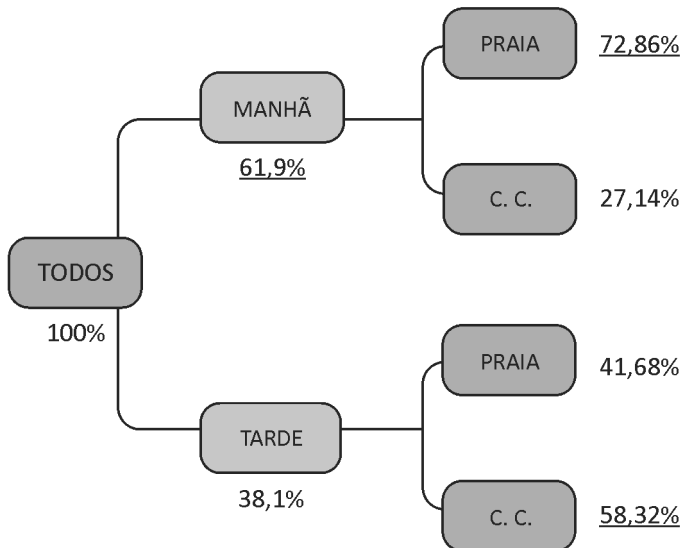
**Figura 13.6:** Árvore Condicional completa.

Mais uma vez temos uma opção para confirmar se os dados estão corretos. Como a Árvore de Decisão Condicional se comporta de maneira igual à Árvore de Decisão Combinada nos dois primeiros níveis, em cada um deles o somatório das probabilidades precisa ser 100%.

Por sua vez, já no terceiro nível, para a Árvore de Decisão Condicional deveremos somar as probabilidades separadamente por grupos, considerando a repartição feita a partir do segundo nível. Isto é: todos os percentuais do terceiro nível (caso existissem mais de dois) associados à característica homem precisam somar 100%. De maneira análoga, todos os percentuais do terceiro nível associados à característica mulher também precisam somar 100% e assim por diante, quantas características tivermos no segundo nível.



segue temos a árvore completa, sendo os valores sublinhados originais do enunciado do exercício desta atividade e os demais calculados.



O percentual de passeios feito de tarde foi calculado igualmente como calculamos a quantidade de homens ou de não homens nos exemplos anteriores, exatamente pela já repetida, exaustivamente, igualdade nos dois primeiros níveis em ambas as árvores. Os demais percentuais foram calculados utilizando a propriedade exclusiva da Árvore de Decisão Condicional. Ao saber que 72,86% dos passeios feitos pela manhã são para praia, lógico que para os centros culturais será 27,14% (restante para 100%). De maneira igual, sabendo que 58,32% dos passeios da tarde são para centros culturais, concluímos que 41,68% são para a praia (restante para 100%).

## Interagindo com as árvores

Tanto quando estávamos falando da Árvore Combinada, quanto estávamos falando da Árvore Condicional, usamos o mesmo cenário inspirado nas informações do exercício da Introdução desta aula. Esse recurso foi intencional para que pudessem notar a possibilidade de trabalhar com ambas as árvores em um mesmo estudo. Contudo, não ire-

mos mais falar da possibilidade de trabalhar com ambas, mas, sim, afirmar a obrigatoriedade de lidar com as duas árvores ao mesmo tempo.

Deste modo, em algumas situações não temos dados suficientes para completar a Árvore Combinada e/ou a Árvore Condicional. Para tal, precisamos recorrer a outra árvore para que possamos aproveitar os dados nela contidos. Isto só é possível porque para calcular uma probabilidade condicional, necessariamente, precisamos da probabilidade combinada das informações em questão. Logo, por consequência, a inversão dessa utilidade se torna possível.

Vamos supor um estudo no qual você deseje calcular a probabilidade do cliente ser homem, sabendo-se que as compras são feitas aos domingos. Recordando a forma de calcular, sabemos que para obtermos o resultado desejado precisaremos da probabilidade de ser um homem que compra aos domingos (chamaremos de  $H$  e  $D$ ) e da probabilidade de ser um comprador de domingo (chamaremos de  $D$ ). A próxima **Figura 13.7** ilustra como será o cálculo utilizando as letras que atribuímos:

$$P(H|D) = \frac{P(H \text{ e } D)}{P(D)}$$

**Figura 13.7:** Fórmula de Cálculo da Probabilidade.

Partindo-se daí, caso tivéssemos apenas a probabilidade condicional ser homem sabendo que faz compras aos domingos e a probabilidade de ser um comprador de domingo, seria possível calcular a probabilidade combinada de ser homem e fazer compras aos domingos. Basta apenas fazer uma manipulação algébrica e obteremos o resultado indicado pela próxima figura.

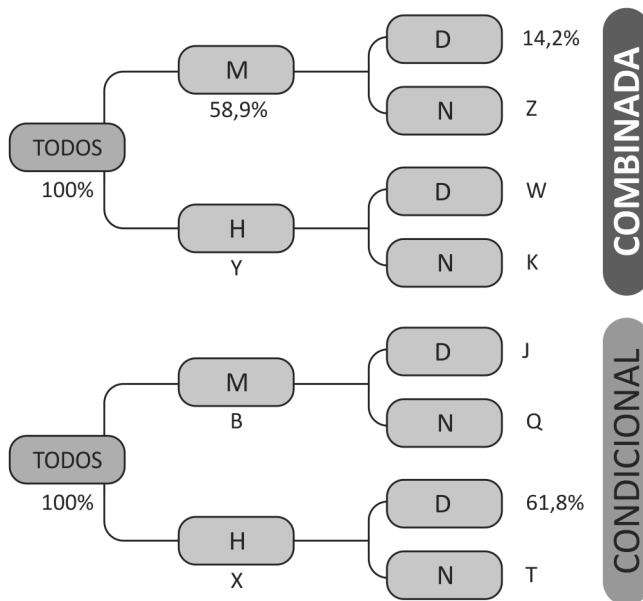
$$P(H|D) \cdot P(D) = P(H \text{ e } D)$$

**Figura 13.8:** Cálculo da Probabilidade Combinada.

Será esse princípio que nos permitirá aproveitar as informações de uma árvore e transferi-las para a outra árvore. Vamos criar um exemplo para melhor ilustrar como isto será feito.

Suponhamos, então, que a lavanderia de um hotel deseja melhor estruturar o seu funcionamento. Para tal, ela deverá considerar a demanda

de toalhas masculinas e de toalhas femininas. Além disso, deverá considerar, também, o estado que essas toalhas lhe são entregues. As toalhas que o cliente espera para devolver dão mais trabalho para serem lavadas, diferentemente das que são rapidamente devolvidas. De posse das informações coletadas, eles conseguiram montar as duas árvores contidas na próxima figura:



**Figura 13.9:** Etapa inicial de criação de Árvores Condicional e Combinada.

Temos duas árvores totalizando 14 campos. Contudo, apenas 5 estão preenchidos. Deveremos determinar os valores dos demais. Os campos que precisam ser preenchidos foram identificados com letras para que possamos identificar de maneira mais ágil no decorrer da solução. É importante não confundir com as letras que simbolizam as informações da pesquisa: M (mulher); H (homem); D (demora a devolver a toalha); N (não demora a devolver a toalha).

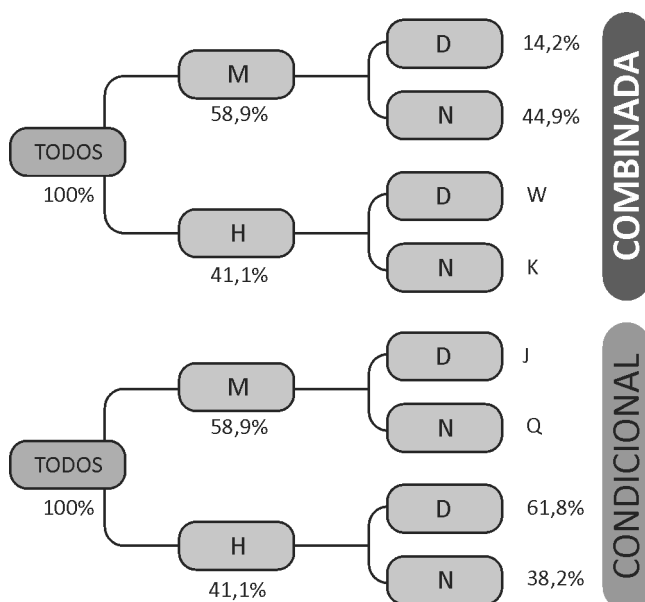
Assinalo que não existe um padrão a ser seguido para a resolução de situações como estas. Temos a nossa disposição, neste momento, uma diversidade de opções de movimentos a serem feitos até chegar ao completo preenchimento das árvores. O conselho padrão é que comecem sempre pelas obviedades. A partir daí, com mais informações chegando, as demais serão detalhes tão óbvios quanto.

Assim, alguns campos são bem óbvios para o rápido preenchimento. O campo Y é a diferença do total de mulheres para o espaço amostral.

Logo, 41,1%. Já os campos B e X são exatamente os mesmo campos de mulheres e homens da outra árvore. Portanto, basta apenas copiar o percentual de uma para a outra. Note que até aqui não nos preocupamos em saber, mesmo que devidamente identificadas, qual árvore é combinada ou condicional – isto porque elas se comportam de maneira igual nos dois primeiros níveis.

Já para preencher o campo Z precisaremos notar que estamos lidando com uma Árvore Combinada. Nela, os percentuais associados a uma mesma característica de um nível acima, necessariamente somam o percentual desta característica comum. Isto é, ali o percentual de ser mulher e demorar (14,2%) somado ao percentual de ser mulher e não demorar (Z) precisa resultar no percentual de ser mulher (58,9%). Logo, o resultado de Z será 44,9%.

Temos, também, o campo T como uma opção de rápida solução. Contudo, devemos nos atentar que agora estamos lidando com uma Árvore Condicional. Neste caso, os percentuais associados a uma mesma característica do nível acima precisam somar 100%. Isto é: a probabilidade de demorar sabendo-se que é homem (61,8%) somada à probabilidade de não demorar sabendo que é homem (T) precisa ser 100%. Fazendo as contas obteremos 38,2%. Neste momento, temos uma nova parcial conforme a figura que segue:



**Figura 13.10:** Etapa intermediária de criação de Árvores Condicional e Combinada.

Até aqui acredito que todos sejam capazes de chegar por conta própria. A novidade está em transferir as informações. Perceba que, para calcular W e K, precisaremos usar informações referentes às mesmas características da outra árvore. De forma simétrica, teremos de fazer o mesmo para calcular J e Q. Vamos pela ordem e calcularemos W e K. Remetendo ao que foi mostrado na Figura 13.8, vamos substituir apenas pelas informações deste exemplo em particular conforme a próxima figura:

$$P(DIH) \cdot P(H) = P(H \text{ e } D) \therefore 61,8\% \cdot 41,1\% = W \therefore W = 25,4\%$$

$$P(NIH) \cdot P(H) = P(H \text{ e } N) \therefore 38,2\% \cdot 41,1\% = K \therefore K = 15,5\%$$

**Figura 13.11:** Cálculo de W e K.

Para quem for mais detalhista, podemos destacar que não necessariamente precisaríamos calcular W e K de maneira igual. Poderíamos calcular apenas W, por exemplo, como acabamos de fazer, e usar para calcular K o mesmo raciocínio que usamos para calcular Z.

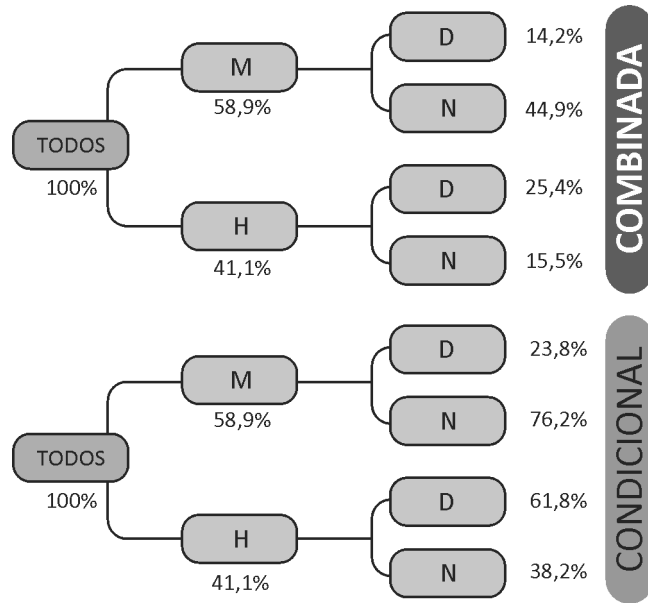
Neste processo, seguindo para os campos J e Q, iremos utilizar o que foi mostrado na **Figura 13.7**, mas substituindo pelas informações que estamos utilizando neste momento. Vejamos na figura que segue.

$$P(DIM) = \frac{P(M \text{ e } D)}{P(M)} \therefore J = \frac{14,2\%}{44,9\%} \therefore J = 23,8\%$$

$$P(NIM) = \frac{P(M \text{ e } N)}{P(M)} \therefore Q = \frac{44,9\%}{44,95} \therefore Q = 76,2\%$$

**Figura 13.12:** Cálculo de J e Q.

Aqui também vale a mesma ressalva feita após a **Figura 13.11**. Não era necessário calcular J e Q usando o mesmo recurso. Poderíamos calcular apenas J, por exemplo, da forma que fizemos e Q utilizando o mesmo mecanismo que foi utilizado para calcular T. De qualquer forma, na próxima figura temos as duas árvores completas:



**Figura 13.13:** Etapa final de criação de Árvore Condiciona e Combinada.

Deste modo, o importante é que tenhamos em mente que não existe um caminho único para se chegar à solução. Uma boa maneira de praticar é refazendo o exemplo, mas utilizando outras combinações de soluções até chegar ao final. Isto não somente vai reforçar o conceito, como também vai ajudar a desenvolver o raciocínio lógico.

### Atividade 3

#### Atende ao objetivo 3

Foi feita uma pesquisa com diversos visitantes de uma determinada cidade. A pesquisa coletava a informação sobre o sexo de cada entrevistado e se eles achavam que não ter almoço incluído nas refeições influenciava na decisão deles em descobrir a culinária local. Após coletadas, as informações foram resumidas, contudo de forma muito superficial, deixando o resultado final incompleto. São elas:

- a) O percentual de homens que responderam *não* foi de 23,9%;
- b) o percentual de pessoas que responderam *não* sabendo que são mulheres é de 27,1%;
- c) 45,7% das pessoas entrevistadas eram homens.

De posse desses dados, desenvolva as Árvore Condicional e Combinada com todas as informações preenchidas:

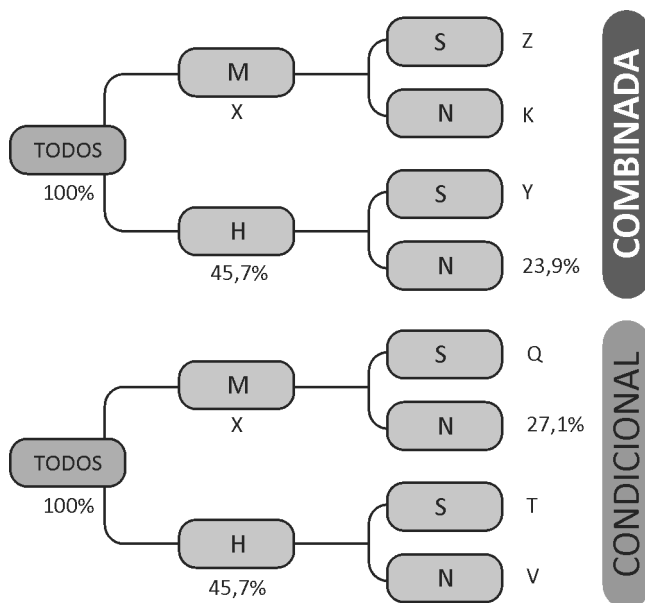
This image shows a single sheet of white paper with horizontal ruling lines. The lines are evenly spaced and run across the width of the page. There are no margins, text, or other markings on the paper.

### Resposta comentada

Antes de montar as árvores é necessário identificar que tipo de informação temos para que não cometamos o erro de inserir o valor em uma árvore errada e, com isto, comprometer do início ao restante da atividade. A informação da *letra a* é referente a uma probabilidade combinada, isto é, o percentual de pessoas que possuem a mesma característica ser homem e ter respondido não.

Na informação da *letra b*, temos uma probabilidade condicional, pois restringimos o espaço amostral ao das mulheres quando garantimos que sabemos que é uma mulher, ficando apenas a característica ter respondido não.

Já a última informação é restrita ao grupo de homens apenas. Com isso, atribuindo aleatoriamente letras para os campos vazios, temos, na figura que segue, a visão inicial das árvores.



Mesmo tendo diversas opções de caminhos para seguir, a sugestão mais fácil é calcular necessariamente nesta ordem: X, Y e Q.

O campo X será calculado considerando a diferença de homens (45,7%) com a pesquisa por completo (100%). O campo Y, através do conceito da probabilidade combinada, será calculado através da diferença faltante entre homens (45,7%) e homens que responderam não (23,9%). O campo Q será calculado pela propriedade da probabilidade condicional, no qual o percentual de pessoas que disseram não sabendo que é mulher (27,1%) e o percentual de pessoas que disseram sim sabendo que é mulher necessariamente somam 100%.

Com essas novas informações, os demais campos podem ser calculados. Seguindo o mesmo raciocínio feito no último exemplo, temos, na próxima figura, como chegar aos valores faltantes:

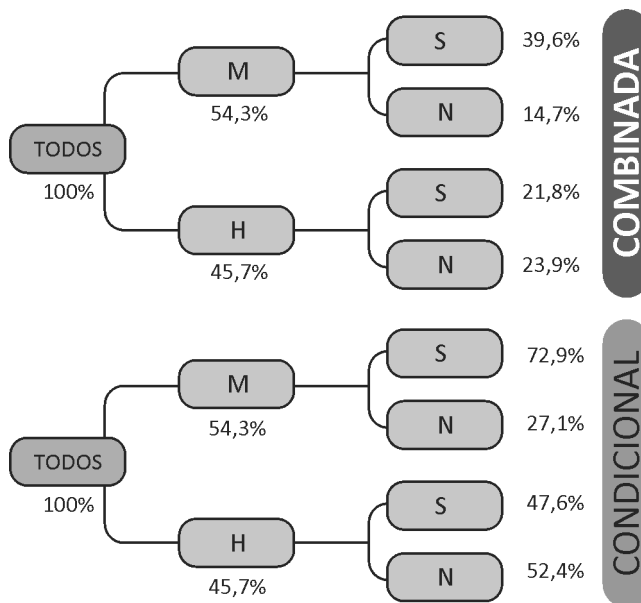
$$P(S|M) \cdot P(M) = P(M \text{ e } S) \therefore 72,9\% \cdot 54,3\% = Z \therefore Z = 39,6\%$$

$$P(N|M) \cdot P(M) = P(M \text{ e } N) \therefore 27,1\% \cdot 54,3\% = K \therefore K = 14,7\%$$

$$P(S|H) = \frac{P(H \text{ e } S)}{P(H)} \therefore T = \frac{21,8\%}{45,7\%} \therefore T = 47,6\%$$

$$P(N|H) = \frac{P(H \text{ e } N)}{P(H)} \therefore V = \frac{23,9\%}{45,7\%} \therefore V = 52,4\%$$

Sendo assim, o resultado final será como a figura que segue:



## Conclusão

É indiscutível o potencial da Árvore de Decisão em um estudo probabilístico. Vimos que, no que se refere à organização das informações e à leitura dos dados, ela é completa e superior à Tabela de Contingência. Não suficiente, ela expande suas qualidades como instrumento para obter dados faltantes e dar a chance de se ter um estudo completo.

Em vista disso, outro fator interessante sobre a Árvore de Decisão está na flexibilidade que ela possui quando estamos lidando com as

duas ao mesmo tempo. Existe uma diversidade de opções a serem seguidas até chegarmos ao preenchimento completo delas. Para uns, isso pode ser considerado um *hobby*, para outros uma grande vantagem de superar obstáculos no decorrer da resolução. O fato é que com tantas opções não existe desculpa para ficar com o estudo pela metade.

Contudo, com tantas vantagens, a Árvore de Decisão ainda é um mecanismo dependente dos conceitos que envolvem a probabilidade condicional e a probabilidade combinada. Isto é: sem o domínio delas e suas respectivas interações, nada poderá ser feito. Logo, as vantagens serão nulas. Portanto, a conclusão imediata é que, independentemente do grau de sofisticação em que estamos, sempre precisaremos dos conceitos básicos para nos dar suporte.

### ===== **Atividade final** =====

#### *Atende aos objetivos 1, 2 e 3*

Um restaurante resolveu acumular todos os pedidos feitos no último mês para tentar encontrar um padrão entre os pedidos dos seus clientes. As informações coletadas eram relacionadas ao tipo de prato e ao tipo de bebida. Primeiro eles contabilizavam os pratos de carne, de frango e de peixe. Depois, verificavam se a pessoa pediu bebida alcoólica ou comum. Sabe-se que uma mesma pessoa não pode pedir mais de um prato e caso tenha pedido dois tipos de bebida (uma alcoólica e outra comum), apenas a primeira que for pedida é que será considerada. Coletadas e organizadas as informações, eles chegaram às seguintes conclusões:

- a) do total de pessoas que pediram carne, 38% não pediram bebida alcoólica;
- b) 18% das pessoas pediram frango e bebida alcoólica;
- c) o percentual de pessoas que pediu peixe é de 34%;
- d) do total de pessoas, 29% pediu carne;
- e) 73% das pessoas que pediram peixe, também pediram bebida alcoólica.

De posse dessas informações complementares, monte as árvores Condicional e Combinada com todas as informações preenchidas.

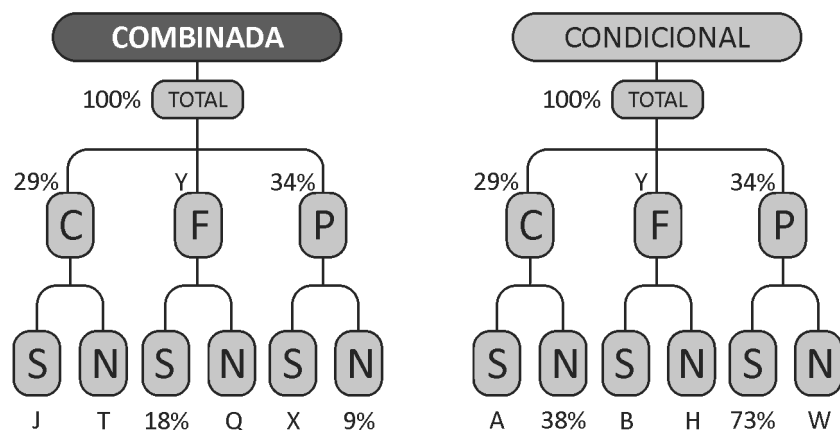
[illegible]

### Resposta comentada

Para montarmos as primeiras árvores com as informações parciais, de início, precisamos identificar cada tipo de dado fornecido para que seja corretamente associado à sua respectiva árvore. Separando os dados, temos que os fornecidos na *letra c* e na *letra d* são marginais, podendo ser associados em ambas as árvores – necessariamente no segundo nível.

Na *letra a* e na *letra e* restringimos o espaço amostral a um tipo de prato. Logo, estamos falando de dados condicionais que, por consequência, irão para a Árvore Condicional. Por fim, a *letra b* restante ficará na Árvore Combinada.

Deste modo, atribuindo letras aleatórias para os campos não preenchidos temos nossas árvores parciais como a próxima figura representa. Para ajudar na leitura, consideramos as seguintes letras para representar as informações da pesquisa: carne (C); frango (F); peixe (P); pediu bebida alcoólica (S); não pediu bebida alcoólica (N).



O campo Y talvez seja o de resposta mais rápida, uma vez que os pedidos de carne (29%), frango (Y) e peixe (34%) precisam somar 100%. Em seguida, escolhendo o campo X, temos que, por ser Árvore Combinada, seu resultado e o campo que representa peixe e bebida não alcoólica (9%) precisam somar 34% (percentual de pessoas que pediram peixe). O mesmo pode ser feito com o campo Q e o que representa frango e bebida alcoólica (18%). Contudo, isto só será possível se já tivermos calculado o valor de Y (37%).

Partindo para a Árvore Combinada, temos o campo A para ser calculado. Sabemos que neste tipo de árvore, necessariamente os campos do

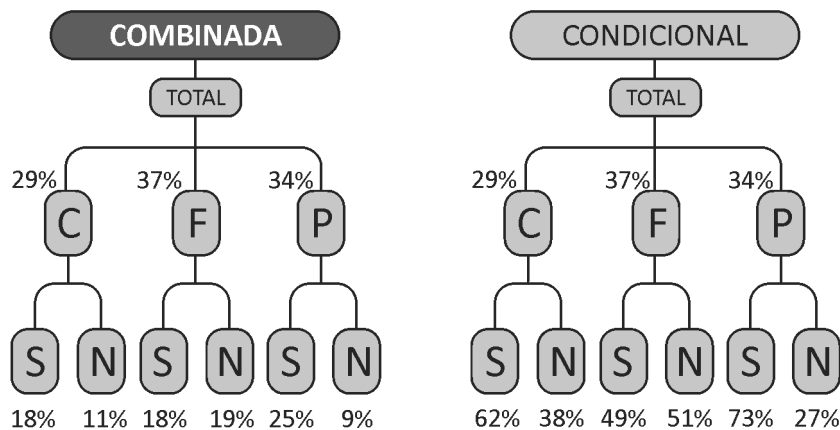
terceiro nível associados à mesma característica do segundo nível somam 100%. Portanto, o campo A e a informação pessoas que não pediram bebida alcoólica sabendo que pediram carne (38%) precisam somar também 100%. O mesmo valerá para o campo W e a informação pessoas que pediram bebida alcoólica sabendo que pediram peixe (73%).

Por fim, faltam os campos B e H que, necessariamente, serão preenchidos com dados importados da Árvore Combinada. Os seus respectivos cálculos, se optarem por fazer de maneira igual, será conforme a próxima figura:

$$P(S|F) = \frac{P(F \text{ e } S)}{P(F)} \therefore B = \frac{18\%}{37\%} \therefore B = 49\%$$

$$P(N|F) = \frac{P(F \text{ e } N)}{P(F)} \therefore H = \frac{19\%}{37\%} \therefore H = 51\%$$

Agora, com todas as informações, podemos finalmente preencher em sua totalidade cada Árvore de Decisão. O resultado consta na figura a seguir:



Mais uma vez, vale ressaltar que a sequência para se obter todos os resultados é totalmente flexível. É importante lembrar da opção de confirmar os resultados, utilizando os métodos de revisão que já foram passados nesta aula. Por fim, fica registrado que, por motivos práticos, exclusivamente nesta atividade não usamos casas decimais. Contudo, é sempre importante utilizar no mínimo duas.

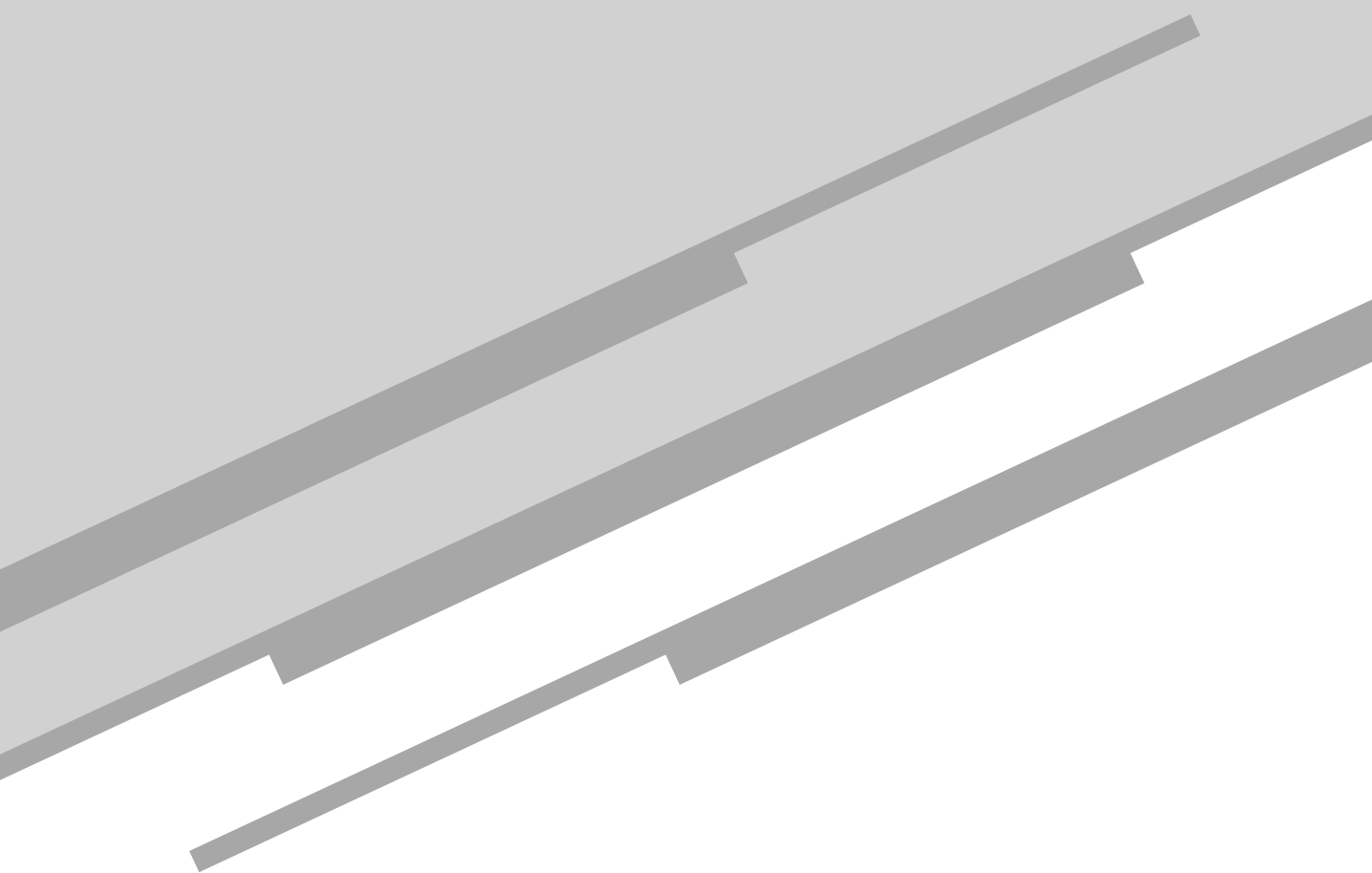
## Resumo

Nesta aula, fomos apresentados à Árvore de Decisão. Primeiro vimos sua estruturação física e como se comportam os campos que a compõem. Campos, níveis e a relação entre eles foi nosso primeiro passo. Isso feito, pudemos aprender como inserir os dados em cada uma delas e como interpretá-los. Quando somos capazes de ler corretamente as árvores, passamos a ter em mãos um poderoso instrumento que, graficamente, resume uma série de informações importantes.

Em seguida, após recordar as propriedades das probabilidades envolvidas foi possível entender como completar os campos que não possuem dados informados. Cada probabilidade tem uma propriedade. Logo, nada mais justo que cada árvore associada a uma probabilidade também tenha uma maneira específica de se comportar. Essa mesma especificidade de cada uma nos permite, ao final, confirmar se os valores estão corretos através da relação que possuem entre eles.

Por fim, com toda essa bagagem, conseguimos consolidar o conhecimento interagindo os dados das árvores entre si. Neste ponto chegamos ao nível máximo de sofisticação do assunto, criando uma independência que nos permite dar prosseguimento a estudos incompletos, mesmo que tenhamos poucas informações.

# Referências



## **Aula 1**

SMAILES, J.; MCGRANE, A. *Estatística aplicada à administração com Excel*. São Paulo: Atlas, 2010.

LEVINE, D.; STEPHAN, D.; KREHBIEL, T.; BERENSON, M. *Estatísticas: Teoria e aplicações*. Rio de Janeiro: LTC, 2008.

TIBONI, C. *Estatística Básica*. São Paulo: Atlas, 2010.

OLIVEIRA, F. *Estatística e Probabilidade*. São Paulo: Atlas, 2011.

LAPPONI, J. *Estatística usando Excel*. São Paulo: Lapponi, 1995.

## **Aula 2**

SMAILES, J.; MCGRANE, A. *Estatística aplicada à administração com Excel*. São Paulo: Atlas, 2010.

LEVINE, D.; STEPHAN, D.; KREHBIEL, T.; BERENSON, M. *Estatísticas: Teoria e aplicações*. Rio de Janeiro: LTC, 2008.

TIBONI, C. *Estatística Básica*. São Paulo: Atlas, 2010.

OLIVEIRA, F. *Estatística e Probabilidade*. São Paulo: Atlas, 2011.

LAPPONI, J. *Estatística usando Excel*. São Paulo: Lapponi, 1995.

## **Aula 3**

SMAILES, J.; MCGRANE, A. *Estatística aplicada à administração com Excel*. São Paulo: Atlas, 2010.

LEVINE, D.; STEPHAN, D.; KREHBIEL, T.; BERENSON, M. *Estatísticas: Teoria e aplicações*. Rio de Janeiro: LTC, 2008.

TIBONI, C. *Estatística Básica*. São Paulo: Atlas, 2010.

OLIVEIRA, F. *Estatística e Probabilidade*. São Paulo: Atlas, 2011.

LAPPONI, J. *Estatística usando Excel*. São Paulo: Lapponi, 1995.

## Aula 4

SMAILES, J.; MCGRANE, A. *Estatística aplicada à administração com Excel*. São Paulo: Atlas, 2010.

LEVINE, D.; STEPHAN, D.; KREHBIEL, T.; BERENSON, M. *Estatísticas: Teoria e aplicações*. Rio de Janeiro: LTC, 2008.

TIBONI, C. *Estatística Básica*. São Paulo: Atlas, 2010.

OLIVEIRA, F. *Estatística e Probabilidade*. São Paulo: Atlas, 2011.

LAPPONI, J. *Estatística usando Excel*. São Paulo: Lapponi, 1995.

## Aula 5

SMAILES, J.; MCGRANE, A. *Estatística aplicada à administração com Excel*. São Paulo: Atlas, 2010.

LEVINE, D.; STEPHAN, D.; KREHBIEL, T.; BERENSON, M. *Estatísticas: Teoria e aplicações*. Rio de Janeiro: LTC, 2008.

TIBONI, C. *Estatística Básica*. São Paulo: Atlas, 2010.

OLIVEIRA, F. *Estatística e Probabilidade*. São Paulo: Atlas, 2011.

LAPPONI, J. *Estatística usando Excel*. São Paulo: Lapponi, 1995.

## Aula 6

SMAILES, J.; MCGRANE, A. *Estatística aplicada à administração com Excel*. São Paulo: Atlas, 2010.

LEVINE, D.; STEPHAN, D.; KREHBIEL, T.; BERENSON, M. *Estatísticas: Teoria e aplicações*. Rio de Janeiro: LTC, 2008.

TIBONI, C. *Estatística Básica*. São Paulo: Atlas, 2010.

OLIVEIRA, F. *Estatística e Probabilidade*. São Paulo: Atlas, 2011.

LAPPONI, J. *Estatística usando Excel*. São Paulo: Lapponi, 1995.

## Aula 7

SMAILES, J.; MCGRANE, A. *Estatística aplicada à administração com Excel*. São Paulo: Atlas, 2010.

LEVINE, D.; STEPHAN, D.; KREHBIEL, T.; BERENSON, M. *Estatísticas: Teoria e aplicações*. Rio de Janeiro: LTC, 2008.

TIBONI, C. *Estatística Básica*. São Paulo: Atlas, 2010.

OLIVEIRA, F. *Estatística e Probabilidade*. São Paulo: Atlas, 2011.

LAPPONI, J. *Estatística usando Excel*. São Paulo: Lapponi, 1995.

## Aula 8

SMAILES, J.; MCGRANE, A. *Estatística aplicada à administração com Excel*. São Paulo: Atlas, 2010.

LEVINE, D.; STEPHAN, D.; KREHBIEL, T.; BERENSON, M. *Estatísticas: Teoria e aplicações*. Rio de Janeiro: LTC, 2008.

TIBONI, C. *Estatística Básica*. São Paulo: Atlas, 2010.

OLIVEIRA, F. *Estatística e Probabilidade*. São Paulo: Atlas, 2011.

LAPPONI, J. *Estatística usando Excel*. São Paulo: Lapponi, 1995.

## Aula 9

SSMAILES, J.; MCGRANE, A. *Estatística aplicada à administração com Excel*. São Paulo: Atlas, 2010.

LEVINE, D.; STEPHAN, D.; KREHBIEL, T.; BERENSON, M. *Estatísticas: Teoria e aplicações*. Rio de Janeiro: LTC, 2008.

TIBONI, C. *Estatística Básica*. São Paulo: Atlas, 2010.

OLIVEIRA, F. *Estatística e Probabilidade*. São Paulo: Atlas, 2011.

LAPPONI, J. *Estatística usando Excel*. São Paulo: Lapponi, 1995.

## Aula 10

SMAILES, J.; MCGRANE, A. *Estatística aplicada à administração com Excel*. São Paulo: Atlas, 2010.

LEVINE, D.; STEPHAN, D.; KREHBIEL, T.; BERENSON, M. *Estatísticas: Teoria e aplicações*. Rio de Janeiro: LTC, 2008.

TIBONI, C. *Estatística Básica*. São Paulo: Atlas, 2010.

OLIVEIRA, F. *Estatística e Probabilidade*. São Paulo: Atlas, 2011.

LAPPONI, J. *Estatística usando Excel*. São Paulo: Lapponi, 1995.

## Aula 11

SMAILES, J.; MCGRANE, A. *Estatística aplicada à administração com Excel*. São Paulo: Atlas, 2010.

LEVINE, D.; STEPHAN, D.; KREHBIEL, T.; BERENSON, M. *Estatísticas: Teoria e aplicações*. Rio de Janeiro: LTC, 2008.

TIBONI, C. *Estatística Básica*. São Paulo: Atlas, 2010.

OLIVEIRA, F. *Estatística e Probabilidade*. São Paulo: Atlas, 2011.

LAPPONI, J. *Estatística usando Excel*. São Paulo: Lapponi, 1995.

## Aula 12

SMAILES, J.; MCGRANE, A. *Estatística aplicada à administração com Excel*. São Paulo: Atlas, 2010.

LEVINE, D.; STEPHAN, D.; KREHBIEL, T.; BERENSON, M. *Estatísticas: Teoria e aplicações*. Rio de Janeiro: LTC, 2008.

TIBONI, C. *Estatística Básica*. São Paulo: Atlas, 2010.

OLIVEIRA, F. *Estatística e Probabilidade*. São Paulo: Atlas, 2011.

LAPPONI, J. *Estatística usando Excel*. São Paulo: Lapponi, 1995.

## **Aula 13**

SMAILES, J.; MCGRANE, A. *Estatística aplicada à administração com Excel*. São Paulo: Atlas, 2010.

LEVINE, D.; STEPHAN, D.; KREHBIEL, T.; BERENSON, M. *Estatísticas: Teoria e aplicações*. Rio de Janeiro: LTC, 2008.

TIBONI, C. *Estatística Básica*. São Paulo: Atlas, 2010.

OLIVEIRA, F. *Estatística e Probabilidade*. São Paulo: Atlas, 2011.

LAPPONI, J. *Estatística usando Excel*. São Paulo: Lapponi, 1995.