

# ESTATÍSTICA

Roberta Mendiondo

Sumário	
<b>Unidade 1: ESTATÍSTICA DESCRITIVA</b>	5
<b>Objetivos</b>	5
<b>Introdução</b>	5
<b>Palavras-chave da Unidade</b>	5
<b>Seção 1: Introdução à estatística</b>	5
Seção 2: <b>Construção e análise de gráficos e tabelas</b>	11
Seção 3: <b>Medidas de tendência central:</b>	20
Seção 4: <b>Medidas de dispersão</b>	22
Desafio	29
Saiba mais	29
Dica de Leitura	29
Finalizando a Unidade	30
Material de apoio	30
Referência Bibliográfica	31
<b>Unidade 2:</b>	33
Objetivos	33
Introdução	33
Palavras-chave da Unidade	33
Seção 1:	33
Seção 2:	35
Seção 3:	40
Seção 4:	41
Desafio	50
Saiba mais	50
Dica de Leitura	50
Finalizando a Unidade	50
Material de apoio	51
Referência Bibliográfica	51
<b>Unidade 3:</b>	53
Objetivos	53
Introdução	53
Palavras-chave da Unidade	53
Seção 1:	53
Seção 2:	56
Seção 3:	60
Seção 4:	66

Desafio.....	69
Saiba mais.....	69
Dica de Leitura .....	69
Finalizando a Unidade .....	69
Material de apoio .....	<b>Erro! Indicador não definido.</b>
Referência Bibliográfica.....	<b>Erro! Indicador não definido.</b>
<b>Unidade 4:</b> .....	69
Objetivos.....	72
Introdução.....	72
Palavras-chave da Unidade.....	72
Seção 1: .....	72
Seção 2: .....	75
Seção 3: .....	87
Seção 4: .....	94
Desafio.....	99
Saiba mais.....	99
Dica de Leitura .....	99
Finalizando a Unidade .....	99
Material de apoio .....	100
Referência Bibliográfica.....	100
<b>Unidade 5:</b> .....	<b>Erro! Indicador não definido.</b>
Objetivos.....	<b>Erro! Indicador não definido.</b>
Introdução.....	<b>Erro! Indicador não definido.</b>
Palavras-chave da Unidade.....	<b>Erro! Indicador não definido.</b>
Seção 1: .....	<b>Erro! Indicador não definido.</b>
Seção 2: .....	<b>Erro! Indicador não definido.</b>
Seção 3: .....	<b>Erro! Indicador não definido.</b>
Seção 4: .....	<b>Erro! Indicador não definido.</b>
Desafio.....	<b>Erro! Indicador não definido.</b>
Saiba mais.....	<b>Erro! Indicador não definido.</b>
Dica de Leitura .....	<b>Erro! Indicador não definido.</b>
Finalizando a Unidade .....	<b>Erro! Indicador não definido.</b>
Material de apoio .....	<b>Erro! Indicador não definido.</b>
Referência Bibliográfica.....	<b>Erro! Indicador não definido.</b>
<b>Unidade 6:</b> .....	<b>Erro! Indicador não definido.</b>
Objetivos.....	<b>Erro! Indicador não definido.</b>
Introdução.....	<b>Erro! Indicador não definido.</b>

Palavras-chave da Unidade.....	<b>Erro! Indicador não definido.</b>
Seção 1: .....	<b>Erro! Indicador não definido.</b>
Seção 2: .....	<b>Erro! Indicador não definido.</b>
Seção 3: .....	<b>Erro! Indicador não definido.</b>
Seção 4: .....	<b>Erro! Indicador não definido.</b>
Desafio.....	<b>Erro! Indicador não definido.</b>
Saiba mais.....	<b>Erro! Indicador não definido.</b>
Dica de Leitura .....	<b>Erro! Indicador não definido.</b>
Finalizando a Unidade .....	<b>Erro! Indicador não definido.</b>
Material de apoio .....	<b>Erro! Indicador não definido.</b>
Referência Bibliográfica.....	<b>Erro! Indicador não definido.</b>

## **Unidade 1: ESTATÍSTICA DESCRITIVA**

### **Objetivos**

Realizar a coleta, descrição e análise dos dados referentes à uma pesquisa

### **Introdução**

A Estatística muitas vezes é entendida apenas como tabelas, índices e gráficos associados à economia, esportes, desemprego, natalidade, dentre outros; no máximo responsável por previsões de resultados eleitorais, mas o trabalho da Estatística contempla o planejamento da pesquisa, a coleta de dados, a organização dos dados para gerar informação, a interpretação e a análise destes dados e apresentação de resultados de forma que possam subsidiar processos de tomada de decisão mais razoáveis.

A Estatística se ocupa do tratamento de dados de maneira que se possa extrair informações deles.

A estatística pode ser dividida em duas partes: a **Estatística Descritiva** e a **Estatística Inferencial**. Nesta unidade trataremos da Estatística Descritiva.

### **Palavras-chave da Unidade**

Tabela, gráfico, frequência, média, desvio-padrão.

### **Seção 1: Introdução à estatística**

Nesta seção, você estudará algumas instituições, índices, eventos e contextos nos quais a Estatística está presente como forma de coletar, analisar e informar a população.

## No Brasil

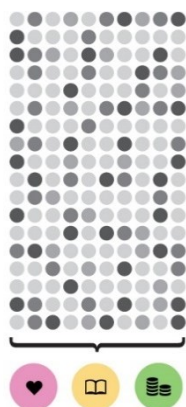
**IBGE** A instituição mais importante que trata de estatísticas no Brasil é o Instituto Brasileiro de Geografia e Estatística – IBGE, o qual realiza, a cada década, o censo populacional, quando pesquisa sobre a **população** do país. Por outro lado, trimestralmente, o IBGE desenvolve a Pesquisa Nacional por Amostras de Domicílios Contínua – PNAD Contínua, que coleta dados de uma **amostra** da população brasileira. Segundo o IBGE, a PNAD Contínua

Visa acompanhar as flutuações trimestrais e a evolução, no curto, médio e longo prazos, da força de trabalho, e outras informações necessárias para o estudo do desenvolvimento socioeconômico do País. Para atender a tais objetivos, a pesquisa foi planejada para produzir indicadores trimestrais sobre a força de trabalho e indicadores anuais sobre temas suplementares permanentes (como trabalho e outras formas de trabalho, cuidados de pessoas e afazeres domésticos, tecnologia da informação e da comunicação etc.), investigados em um trimestre específico ou aplicados em uma parte da amostra a cada trimestre e acumulados para gerar resultados anuais, sendo produzidos, também, com periodicidade variável, indicadores sobre outros temas suplementares. Tem como unidade de investigação o domicílio.

Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao/9173-pesquisa-nacional-por-amostra-de-domicilios-continua-trimestral.html?t=o-que-e>

**IDH** A partir de pesquisas censitárias realizadas pelo IBGE são definidos o Índices de Desenvolvimento Humano – IDH - no Brasil e correspondentes IDH municipais – IDHM.

O IDH faz parte do Programa das nações Unidas para o Desenvolvimento, e segundo este



“ é uma medida resumida do progresso a longo prazo em três dimensões básicas do desenvolvimento humano: renda, educação e saúde. O objetivo da criação do IDH foi o de oferecer um contraponto a outro indicador muito utilizado, o Produto Interno Bruto (PIB) *per capita*, que considera apenas a dimensão econômica do desenvolvimento. Criado por Mahbub ul Haq com a colaboração do economista indiano Amartya Sen, ganhador do Prêmio Nobel de Economia de 1998, o IDH pretende ser uma medida geral e sintética que, apesar de ampliar a perspectiva sobre o desenvolvimento humano, não abrange nem esgota todos os aspectos de desenvolvimento.” (Disponível em : <http://www.br.undp.org/content/brazil/pt/home/idh0.html>)

**IBOVESPA** O Ibovespa é o resultado de uma carteira teórica de ativos (ações). O objetivo do Ibovespa é ser o indicador do desempenho médio das cotações dos ativos de maior negociabilidade e representatividade do mercado de ações brasileiro, representando um índice de retorno total das ações negociadas na bolsa de valores.

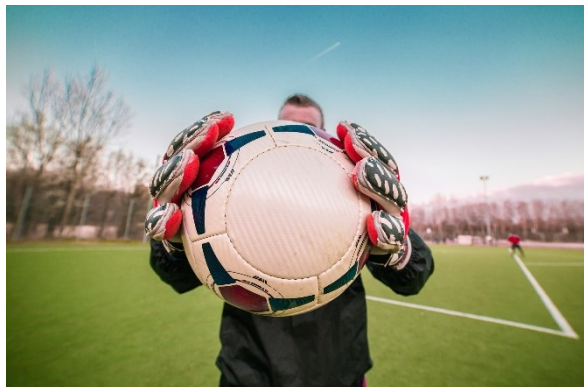


Na página da [BM&fBOVESPA](#) ainda é possível acompanhar as cotações das ações de empresas listadas na Bolsa por meio de gráficos.

**OURO NO FUTEBOL OLÍMPICO O ouro no futebol masculino das Olimpíadas Rio 2016 foi conquistado com Estatística.**

O goleiro Weverton conhecia estatísticas sobre o último batedor, Petersen, este havia batido metade das cobranças para cada canto da rede, o que não ajudaria muito na análise, mas ao analisar melhor os dados, foi observado que em momentos mais tensos, como seria o caso da final olímpica, ele batia para o lado esquerdo; munido desse conhecimento proporcionado pelas análises estatísticas, o goleiro Weverton escolheu o canto certo e garantiu a Neymar a oportunidade de consolidar a medalha de ouro para a seleção olímpica brasileira. Veja uma reportagem completa sobre o impacto da Estatística nessa conquista, a partir de entrevista com o goleiro campeão, [aqui](#) .

**PROCESSOS ELEITORAIS** O Superior Tribunal Eleitoral mantém uma página com estatísticas sobre eleitores, candidatos e resultados das eleições, com informações como escolaridade, faixa etária, sexo, idade, profissão, partido das candidaturas e relação



candidato – vaga. Você pode explorar estas informações diretamente do site do [Superior Tribunal Eleitoral](#) .

### Saiba mais

Para saber mais sobre como são planejadas e realizadas as pesquisas eleitorais assista ao vídeo [Dando IBOPE](#) no qual um personagem, funcionário do IBOPE, expõe os métodos que o instituto utiliza e esclarece algumas dúvidas comuns dos eleitores.

**Afinal, porque muitos de nós nunca fomos entrevistados e mesmo assim, a maioria das pesquisas eleitorais acaba apontando corretamente os resultados das eleições?**

### NO MUNDO

#### **A paralisia infantil erradicada e o Zé Gotinha, o que têm a ver com Estatística?**

Assista o vídeo intitulado [Lembranças de Sofia](#) para entender o papel do método estatístico na erradicação da Poliomielite, doença que atingia milhares de crianças, pelo mundo.

A questão central era como testar a efetividade de um novo medicamento. Por meio do planejamento realizado ao testar a vacina Salk, a bióloga Sofia apresentou alguns aspectos importantes que devem ser levados em consideração.



Fonte: Por USAID -  
USAID Bangladesh,  
Domínio público,

<https://commons.wikimedia.org/w/index.php?curid=1021>

**Terremoto na Itália** Já no primeiro mês do ano de 2017, os italianos sofreram com uma avalanche causada por tremores de terra, que soterrou um hotel de luxo na cordilheira dos Montes Apeninos. Ainda em agosto de 2016, ocorreram 298 mortes causadas por terremoto nas províncias de Marca e Rieti e os tremores não pararam.





A elaboração de estatísticas sobre os tremores se torna fundamental para tomada de decisões pelas autoridades, que proporcionem maior segurança à população. Nesse sentido, o Instituto Nacional de Geofísica e Vulcanologia (Itália) monitora os dados sísmicos e elabora estatísticas que subsidiam possíveis intervenções de segurança.

L'Aquila, Abruzzo, Itália. O escritório de um governo interrompido pelo terremoto de 2009  
[https://it.wikipedia.org/wiki/Terremoti\\_in\\_Italia#/media/File:L'Aquila\\_earthquake\\_prefettura.jpg](https://it.wikipedia.org/wiki/Terremoti_in_Italia#/media/File:L'Aquila_earthquake_prefettura.jpg) (Licença Creative Commons)

**Nessa página**, do INGV é possível observar uma tabela com dados dos últimos terremotos, por magnitude, província de ocorrência, profundidade, e posição geográfica, a qual pode ser exportada em alguns formatos.

**ONU e Objetivos de Desenvolvimento Sustentável** A Comissão de Estatística das Nações Unidas aprovou, em março de 2016, um conjunto de 230 indicadores globais que serão utilizados para monitorar e revisar o cumprimento dos Objetivos de Desenvolvimento Sustentável (ODS) no mundo. Esta comissão estabelece padrões estatísticos globais, conceitos e métodos e a implementação de ações a nível nacional e internacional para o atingimento dos ODS.



Fonte: <https://nacoesunidas.org/pos2015/agenda2030/>

Os indicadores são globais e não necessariamente adequados a contextos nacionais, os quais podem precisar do desenvolvimento de mecanismos específicos para avaliar seus

progressos. Dos países será requerido que gerem um grande volume de dados para que o atingimento dos ODS, por cada país, possa ser monitorado.

Veja [neste vídeo](#) as referências a dados como:

- 50% da população brasileira não tem acesso a esgotamento sanitário.
- 27 milhões de mulheres brasileiras não têm acesso a saneamento básico.
- 1,5 milhões de brasileiras não têm banheiro em suas casas.
- Renda das mulheres que não tem acesso a saneamento básico é 73% menor do que daquelas que tem acesso a esse serviço.

Estatísticas de um país de desigualdades...

**Melhores salários são dos cientistas de dados** Segundo a [Harvard Business Review](#), em 2018, o cientista de dados será o profissional mais cobiçado do século 21, e dentre outras habilidades ele precisa conhecer Estatística para realizar análises sobre [Big Data](#), de forma que entregue às empresas informações relevantes e com potencial para alavancar seus lucros e posição no mercado.

Veja trechos da matéria intitulada Cientista de dados: o profissional mais cobiçado do século 21, da Harvard Business Review Brasil



"A firma de capital de risco Greylock Partners, que investe na fase embrionária de projetos e já bancou nomes como Facebook, LinkedIn, Palo Alto Networks e Workday, está tão preocupada com essa escassez que montou uma equipe especial de recrutamento só para suprir empresas em sua carteira com cientistas de dados. "Quando [uma empresa] tem dados", diz Dan Portillo, chefe da equipe, "é indispensável ter gente para processar isso tudo e tirar insights dali".

*"A profissão da hora na próxima década será a de estatístico." A frase é atribuída a Hal Varian, economista-chefe do Google. "Todo mundo acha que estou brincando, mas quem teria imaginado que a engenharia da computação teria sido a profissão da hora na década de 1990?"*

Você pode observar que contextos distintos, do IBGE aos salários dos Ciência de Dados tanto em seus significados, quanto em sua geografia são tratados pela Estatística e muitos outros além dos que você estudou, nesta seção, fazem uso do método estatístico. O cotidiano de cada pessoa é permeado por estatísticas que fornecem subsídios para várias questões importantes na vida, como a relação de candidatos para uma vaga de concurso, o histórico ou média de temperatura de uma cidade para a qual se deseja viajar, o número

de vagas de emprego para alunos do curso que se pretende fazer, a variação do valor de ações de empresa em que se deseja investir, a eficiência de um medicamento que fará parte de um tratamento de câncer, enfim, muitas são as questões cujas repostas podem ser fundamentadas por estatísticas, por isso é importante que todo profissional adquira competência em elaboração e análises de dados – estatísticas.

## Seção 2: **Construção e análise de gráficos e tabelas**

Qualquer conjunto de **dados**, tal como os tempos de ligações telefônicas, as magnitudes de terremotos, as velocidades de processamento de uma determinada máquina, os percentuais de participação no mercado das empresas de tecnologia, os níveis de suscetibilidade de empresas a uma determinada mudança no mercado, opinião dos alunos quanto à qualidade da universidade na qual estudam, dentre outros, possui informação sobre algum grupo de indivíduos, e característica específica desse grupo na qual se está interessado é que configura a **variável de interesse**.

### **VARIÁVEIS**

Uma característica de interesse e que pode assumir diferentes valores de indivíduo para indivíduo é denominada **variável**, pois de outra forma seria denominada constante.

As variáveis são aquilo que queremos observar e podem ser classificadas em **qualitativas** e **quantitativas**, da seguinte forma:

**Variáveis Quantitativas:** assumem valores numéricos, podendo ser contínuas ou discretas.

**Variáveis discretas:** características que podem ser medidas, as quais assumem somente um número finito ou infinito contável de valores e, dessa forma, assumindo somente valores inteiros. Em geral, resultam de contagens.

**Variáveis contínuas:** características que podem ser medidas e assumem valores contínuos, como na reta real; inteiros ou fracionários (decimais). Em geral, costumam resultar de medições com instrumentos.

**Variáveis Qualitativas:** não assumem valores quantitativos, são definidas por categorias, representando uma classificação dos indivíduos, podendo ser nominais ou ordinais.

**Variáveis nominais:** não existe ordenação dentre as categorias.

**Variáveis ordinais:** existe uma ordenação entre as categorias.

Tipo de Variáveis		Exemplos
Quantitativa	Discreta	Número de páginas de um e-book. Número de eleitores dos municípios brasileiros. Quantidade de projetos sociais de empresas brasileiras.
	Contínua	Peso dos funcionários do setor de marketing de uma empresa. Energia liberada por terremotos no mundo. SalDOS bancários de trabalhadores da construção civil
Qualitativa	Nominal	Tipos de empresas que adotam programas de qualidade. Cidade de nascimento de alunos de Ciências Econômicas da Unileste Tipos de jogos de futebol quanto à dificuldade
	Ordinal	Tamanho de empresas atuantes no setor primário (pequena, média e grande) Classificação de candidatos em concursos públicos. Estágio da doença (inicial, intermediário, terminal)

**Como é possível organizar os dados de uma forma mais eficiente, na qual se possa apresentar uma quantidade maior de informações?**

## TABELAS

### Elementos da Tabela

A disposição de uma tabela pode ser generalizada como mostra a Figura 01 a seguir.

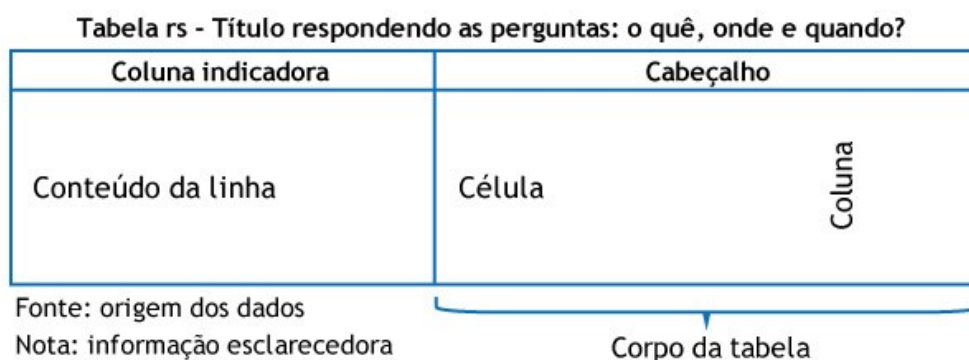


Figura Representação tabular de dados.

### Saiba mais

Veja [neste livro](#), como apresentar tabelas e figuras conforme as regras da Associação Brasileira de Normas Técnicas – ABNT.

### Tabela de Distribuição de Frequências

Se organizarmos dados brutos em ordem crescente teremos um **rol**, porém muitas vezes essa organização não é suficiente, pois existem dados repetidos e a quantidade de vezes que cada dado aparece é importante. Uma forma de organizar um conjunto de dados, sem repetir valores iguais, é por meio de uma tabela onde são apresentadas as frequências de cada uma das categorias.

A disposição tabular dos dados agrupados em classes, juntamente com as frequências correspondentes, se denomina **tabela de distribuição de frequências**.

Você pode contar a frequência de cada dado ou definir certos intervalos de dados, chamados de classes, dos quais se deseja observar a frequência. Na figura abaixo, as classes são as faixas etárias.

Tabela 1

Distribuição dos pacientes hospitalizados pelo SUS segundo sexo, por faixa etária de 0 a 19 anos. Pernambuco, Brasil, 1999.

Faixa etária (anos)	Gênero				Total n
	Masculino n	%	Feminino n	%	
0-4	1.185	59,0	823	41,0	2.008
4-10	1.621	68,4	747	31,6	2.368
10-14	1.892	77,6	544	22,4	2.436
14-19	1.891	78,5	517	21,5	2.408
<b>Total</b>	<b>6.589</b>	<b>71,4</b>	<b>2.631</b>	<b>28,6</b>	<b>9.220</b>

Fonte: MENDONCA, Roberto Natanael da Silva; ALVES, João Guilherme Bezerra; CABRAL FILHO, José Eulálio. Gastos hospitalares com crianças e adolescentes vítimas de violência, no Estado de Pernambuco, Brasil, em 1999. [Cad. Saúde Pública](#), Rio de Janeiro, v. 18, n. 6, p. 1577-1581, Dec. 2002.

### Como decidir o número de classes e o tamanho de cada classe para elaborar uma tabela de distribuição de frequências?

Não há certo, ou errado nesta questão, porém algumas orientações otimizam o processo de construção da tabela de distribuição de frequências e sua apresentação, como:

- Utilizar de 5 a 20 classes, dependendo do volume de dados a ser organizado.
- Manter a mesma largura em cada uma das classes (mesma **amplitude** de classe).
- Usar como limite inferior da primeira classe o menor valor da amostra (porém algumas vezes pode ser conveniente utilizar um valor um pouco menor que esse limite)

Podem ser utilizados alguns critérios matemáticos – fórmulas – para determinar o número de classes (k) e a amplitude de classe.

### Tipos de frequências

Algumas vezes pretendemos comparar conjuntos de dados, e nesse caso a frequência em termos percentuais é fundamental, porque permite a comparação entre amostras com diferentes números de elementos.

**Frequência Absoluta:** gerada na contagem dos dados, é a quantidade de vezes que um elemento da amostra aparece (sua frequência – absoluta)

**Frequência Relativa:** é calculada dividindo-se a frequência absoluta de cada classe da variável pelo número total de observações (número de elementos da amostra ou da população), e multiplicando-se este resultado por 100, obtém-se a frequência relativa em termos percentuais

---

## Saiba mais

### Explorando o conceito de frequência relativa

Leia a reportagem da Folha de São Paulo, baseada em pesquisa do DataFolha com ingressantes na USP, em 2016, sob o título : **Elite está nos cursos mais concorridos da USP; a classe C, nos menos** .

Todos os gráficos do tipo pizza apresentados na matéria, mostram frequências relativas – percentuais - e poderiam ser representados em uma tabela de distribuição de frequências.

Agora considere que o Datafolha tenha entrevistado 2000 estudantes ingressantes na USP, em 2016 (amostra), neste caso, a tabela de distribuição de frequências poderia ser apresentada da seguinte forma:

	Frequência	Frequência	
CLASSE SOCIAL	absoluta	relativa	Frequência relativa (%)
AB	1740	0,87	87
C	260	0,13	13

Fonte: Datafolha



Ou seja, a pesquisa aponta que a maior e mais conceituada universidade pública brasileira, sustentada pelos impostos de todos os cidadãos, admitiu em 2016, 87% de estudantes das classes mais privilegiadas, enquanto somente 13% deles vêm da classe C, e nenhum estudante originário das classes D e E entrou na USP, no mesmo ano.

Fonte: <https://br.freepik.com/fotos-vetores-gratis/escola>>Escola foto criado por freepik - br.freepik.com</a>

**\*\*Não deixe de verificar as frequências relativas na página da reportagem.**

Se considerarmos a classificação proposta pelo IBGE, as classes sociais estão distribuídas pelo critério exclusivo de renda mensal, conforme tabela abaixo:

<b>CLASSE SOCIAL</b>	<b>FAIXA SALARIAL (em salários mínimos)</b>	<b>FAIXA SALARIAL (em reais - 2016)</b>
<b>A</b>	> 20	> 17.600,01
<b>B</b>	10 --  20	8.800,01 --  17.600,00
<b>C</b>	4 --  10	3.720,01 --  8.800,00
<b>D</b>	2 --  4	1.760,01 --  3.720,00
<b>E</b>	0 --  2	0,00 --  1.760,00

## GRÁFICOS

Tabelas são quadros que resumem um conjunto de observações, enquanto os gráficos são formas de apresentação dos dados, que implicam em interpretação mais rápida – visual – do fenômeno estudado.

A escolha do tipo de gráfico mais adequado para representar um conjunto de dados deve ser realizada a partir das respostas de questões como:

- Gráfico realmente é a melhor opção?
- Qual é o público-alvo?
- Qual é o objetivo do gráfico?
- Que tipo de gráfico deve ser usado?
- Como o gráfico deve ser apresentado?



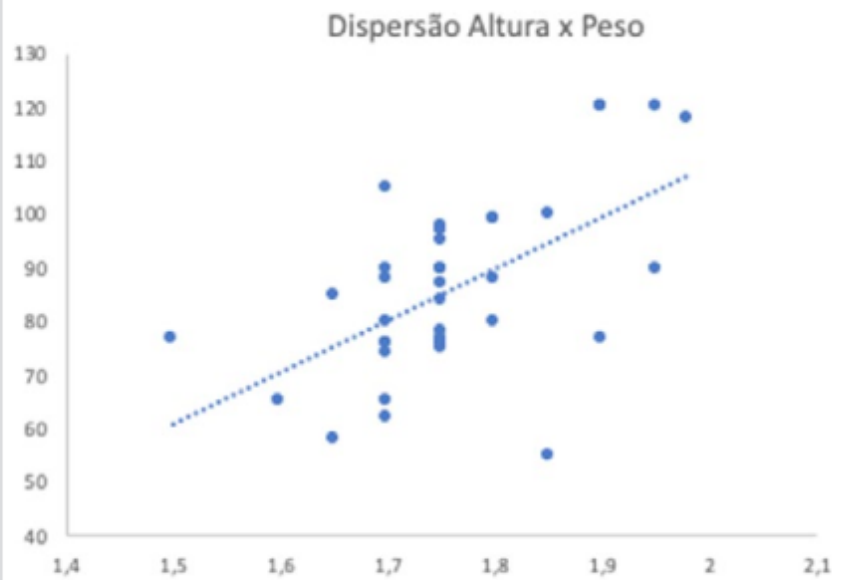
Ao responder estas questões para o seu problema real, você estará mais próximo de uma escolha correta e que trará resultados mais eficazes na transmissão da informação para o público que deseja atingir.

**Tabela resumo de tipos de gráficos mais usuais e sua adequação.**

Tipo/Características	Representação
<p><b>1. Gráfico de colunas</b></p> <p>Representação de uma série por meio de retângulos de mesma base e alturas proporcionais aos respectivos dados.</p> <p>Utilizado quando apresentamos as categorias com palavras curtas.</p>	<p><b>NÚMERO DE MATRÍCULAS NA EDUCAÇÃO SUPERIOR (GRADUAÇÃO E SEQUENCIAL) – BRASIL – 2008-2018</b></p> <p>Fonte: Elaboração própria com base em dados do Censo da Educação Superior 2018.</p>
<p><b>2. Gráfico de barras</b></p> <p>Representação de uma série por meio de retângulos de mesma altura e bases proporcionais aos respectivos dados.</p> <p>Utilizado quando apresentamos as categorias com palavras extensas.</p>	<p><b>Ranking de Índice de Desenvolvimento Humano - IDH - 2015</b> Por país, por ordem de desenvolvimento</p> <p>Fonte: Pnud</p>

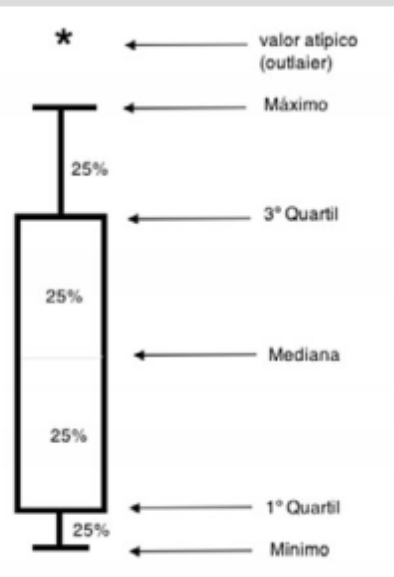
### 6. Diagrama de dispersão (scatterplot) e linha de tendência

O diagrama de dispersão, também conhecido como scatterplot, é a representação do relacionamento entre duas variáveis quantitativas. A linha de tendência é a linha que faz uma aproximação dessa relação.



### 7. Box Plot

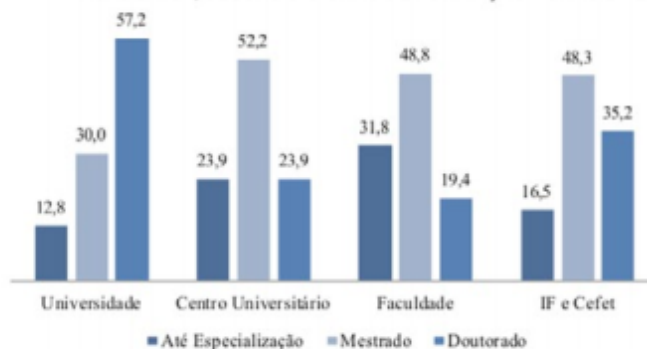
É um gráfico que possibilita interpretar rapidamente o comportamento da distribuição em cada  $\frac{1}{4}$  ou 25% (quartis). É muito útil para observarmos a existência de valores discrepantes (*outliers*) e a simetria da distribuição.



### 3. Colunas justapostas

Descreve simultaneamente duas ou mais categorias para uma variável.

PERCENTUAL DO NÚMERO DE FUNÇÕES DOCENTES EM EXERCÍCIO, POR ORGANIZAÇÃO ACADÊMICA, SEGUNDO O GRAU DE FORMAÇÃO – BRASIL – 2017



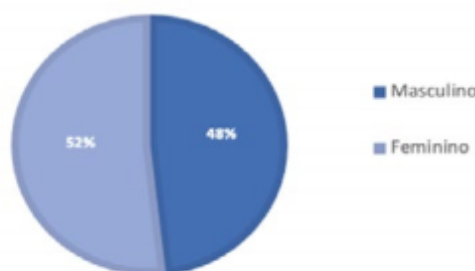
Fonte: Elaborada por Deed/Inep com base nos dados do Censo da Educação Superior.

### 4. Setores

Também conhecido como gráfico de pizza. É utilizado sempre que desejamos ressaltar a participação do dado no total.

Cada setor é obtido por meio de uma regra de três simples e direta, lembrando que 100% corresponde a 360°.

POPULAÇÃO BRASILEIRA, SEGUNDO O SEXO, EM 2019.

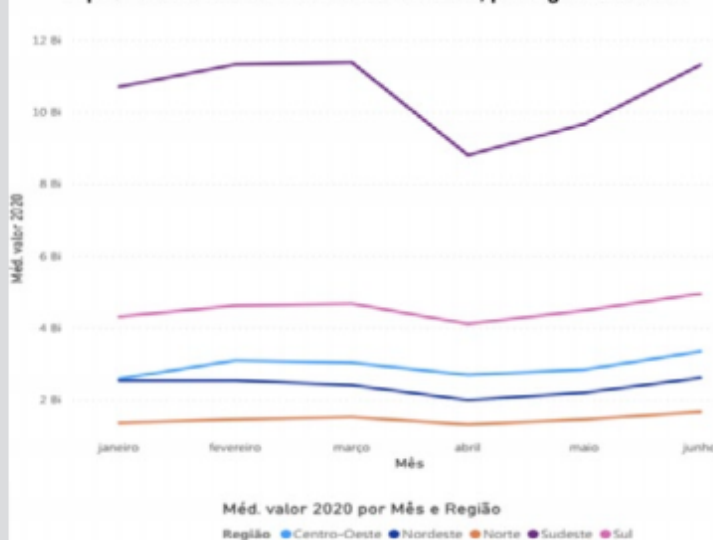


Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Trabalho e Rendimento, Pesquisa Nacional por Amostra de Domicílios Contínua 2012-2019.

### 5. Gráfico de linha ou gráfico de segmentos

Marcamos todos os pares ordenados correspondentes à série estatística e os unimos a partir de uma linha poligonal tracejada ou contínua.

Impactos da Covid-19 no volume de vendas, por região brasileira

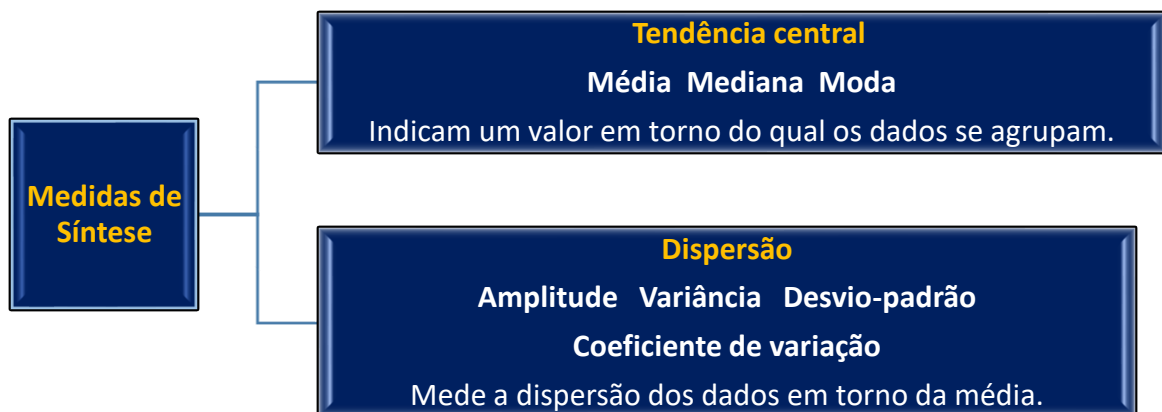


### Seção 3: **Medidas de tendência central**

Medidas de síntese, também chamadas de medidas descritivas, são utilizadas para resumir os dados de uma variável quantitativa, representando o conjunto por meio de valores numéricos. Você estudará dois tipos de medidas de síntese: de **tendência central**, como média, moda e mediana; e de **dispersão** como amplitude, variância, desvio-padrão e coeficiente de variação.

Se as medidas forem calculadas a partir de **dados populacionais**, são denominadas **parâmetros** e se calculadas a partir de **dados amostrais** são denominadas **estatísticas** ou **estimadores**.

A figura abaixo resume a classificação e significados da medidas de síntese.



#### **Medidas Tendência Central**

As medidas de tendência central ajudarão você no resumo, na caracterização de um conjunto de dados, e para tanto você precisa escolher as medidas mais adequadas aos objetivos da sua análise. Por outro lado, se você já receber um conjunto de dados sintetizado, por meio de medidas de tendência central, precisará saber como interpretá-las corretamente.

## O que significam e para que servem medidas tendência central?

### Média Mediana Moda

A **Média** é uma medida que indica o valor típico de um conjunto de dados, e é calculada pela razão entre o somatório dos dados e a quantidade de dados, ou seja, soma-se todos os dados e divide-se pela quantidade de dados do conjunto considerado.

$$Média = \frac{\sum x_i}{n}$$

Onde  $x_i$  = cada um dos dados do conjunto e  $n$  = número de dados que compõe o conjunto

A **Mediana** é o valor que ocupa a posição central do conjunto de dados "ordenados", ou seja, a mediana é um valor que divide o conjunto de dados ao meio, assim em cada parte há uma mesma quantidade de dados.

>> Se o conjunto for formado por um número ímpar de dados, a mediana será o dado do conjunto que ocupa a posição central.

>> Se o conjunto for formado por um número par de dados, a mediana será a média entre os dois dados centrais, e nesse caso, pode não ser um dado do conjunto.

Para calcular a posição da mediana usamos  $P_{md} = \frac{n+1}{2}$ , onde  $n$  = número de dados do conjunto.

A **Moda** é o valor mais frequente do conjunto de dados, indicando o valor que mais vezes aparece no conjunto. Um conjunto de dados pode não ter moda, quando não há valor que ocorra mais de uma vez, ou possuir mais de uma moda quando há valores que se repetem o mesmo número de vezes (sendo a maior frequência).

A moda não é calculada, mas sim "contada", ou seja, conta-se o número de ocorrências de cada valor do conjunto, e a moda será o valor cujo número de ocorrências é o maior.

Na [tabela resumo](#), leia com bastante atenção a última coluna, ela aponta algumas características importantes de cada medida tanto para sua interpretação, quanto para a

escolha da medida mais adequada, se você for o responsável pela elaboração das estatísticas.

Segue um quadro resumo com as medidas de posição.

	Amostras	População	Em que:
Para dados não agrupados.	$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$	$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$	$s^2$ ou $\sigma^2$ = variância $n$ = tamanho da amostra $N$ = tamanho da população $\bar{x}$ = média amostral $\mu$ = média populacional $x_i$ = valor observado $X_i$ = ponto médio da classe
Para dados agrupados em tabelas sem intervalos de classes.	$s^2 = \frac{\sum(x_i - \bar{x})^2 \cdot f_i}{n - 1}$	$\sigma^2 = \frac{\sum(x_i - \mu)^2 \cdot f_i}{N}$	
Para dados agrupados em tabelas com intervalos de classes.	$s^2 = \frac{\sum(X_i - \bar{x})^2 \cdot f_i}{n - 1}$	$\sigma^2 = \frac{\sum(X_i - \mu)^2 \cdot f_i}{N}$	

#### Vídeoaula - Exemplo

Para explorar medidas de posição como média, moda e mediana usando o Excel de, assista a esse [vídeo](#).

Na maioria das vezes, uma medida como a média, por exemplo, não é suficiente para que se tenha clareza sobre o comportamento dos dados de uma amostra, sendo ela uma medida somente de tendência central, e nesse sentido é que surge a necessidade de conhecer medidas que indiquem o quanto os elementos do conjunto se dispersam da média, que é o valor característico da amostra. As medidas que cumprem esse papel são as medidas de dispersão, as quais você estudará na próxima seção.

#### Seção 4: Medidas de dispersão

Para resumir um conjunto de dados é bastante adequado que se conheça uma medida de tendência central (média, moda ou mediana) e uma medida de dispersão (amplitude, variância, desvio-padrão, coeficiente de variação). Você poderá observar que a média será quase sempre necessária, mesmo que não seja a melhor medida para caracterizar a tendência central dos dados, e que o desvio-padrão, em geral, é a medida mais utilizada para medir a dispersão dos dados em torno da média.

### O que significam e para que servem medidas de dispersão?

Amplitude   Variância   Desvio-padrão   Coeficiente de variação

A **amplitude** de um conjunto de dados é calculada pela diferença entre o maior valor e o menor valor de um conjunto de dados quantitativos.

$$\text{Amplitude} = (\text{Valor máximo} - \text{Valor mínimo})$$

#### Exemplo



A **amplitude térmica** de um dia em que a temperatura mínima foi de 16°C e que a temperatura máxima foi de 30°C é de **14°C** (30°C - 16°C).

Uma ideia importante é a de **desvio** em um conjunto de dados. O desvio é a diferença entre o dado considerado e a média do conjunto de dados, ou seja,

$$\text{Desvio} = \text{Dado considerado} - \text{média do conjunto}$$

Essa noção é importante porque a construção das medidas de dispersão variância e o desvio-padrão partem dessa ideia de "desvio" de cada dado em relação à média.

A **Variância** é uma medida de variação dos dados em relação à média, calculada pela média do somatório dos quadrados dos desvios. Uma desvantagem da variância é que sua unidade de medida é diferente da unidade de medida do conjunto de dados, pois como elevamos ao quadrado cada desvio, a unidade de medida da variância fica igual ao

quadrado da unidade de medida dos dados originais. Para superar esse problema é que se calcula e aplica o desvio-padrão (que é a raiz quadrada da variância).

A fórmula para o cálculo da variância de uma população é

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

O **Desvio-padrão** é uma medida que indica o quanto, em média, os valores se dispersam da média do conjunto. Quanto mais dispersos em relação à média os valores estiverem, maior será o desvio-padrão. Essa é a mesma ideia da variância, mas como o cálculo do desvio-padrão inclui a extração da raiz quadrada da variância, a unidade de medida do desvio-padrão é a mesma dos dados estudados.

A fórmula para o cálculo do desvio-padrão de uma população é

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

### Exemplo



Quando uma loja de tecidos vende seus produtos, em geral, o faz por comprimento; considerando que cada tipo de tecido tem uma largura padrão. Se uma loja analisar seus dados de venda mensal do linho, em metros, poderá fazê-lo por meio do cálculo da média mensal de venda e desvio-padrão; e assim terá a média de vendas mensal e ainda uma medida da variabilidade em torno dessa média. Por exemplo,

[ média mensal de vendas de linho = 200 metros (medida de comprimento)  
desvio-padrão = 2,5 metros (medida de comprimento)

Por outro lado, se a loja resolver calcular somente a variância como medida de dispersão, obterá uma medida de variabilidade cuja unidade será o metro quadrado, e não o metro de tecido, como é a realidade de suas vendas e controle. Para o mesmo exemplo, teria

[ média mensal de vendas de linho = 200 metros (medida de comprimento)  
variância = 6,25 metros quadrados (medida de área)



Mesmo que não seja um absurdo, porque podemos medir área dos tecidos, essa mensuração não reflete o problema concreto, porque os tecidos são vendidos por metros de comprimento (sabendo que a largura será sempre a mesma), e portanto a medida de dispersão em metros quadrados dificulta a avaliação da dispersão da quantidade de tecido vendido em torno da média. Nesse sentido é que o desvio-padrão se torna mais adequado que a variância, em muitos contextos, pois apresenta a mesma unidade de medida da variável de estudo (nesse caso, metros de linho puro)

Nesta tabela resumo, leia com bastante atenção a última coluna, ela aponta algumas características importantes de cada medida tanto para sua interpretação, quanto para a escolha da medida mais adequada para representar a variabilidade de dados.



resumo\_medidas\_di  
spersao.pdf

Segue uma tabela resumo com as medidas de dispersão.

### Videoaulas - Exemplo

- 1) Para explorar alguns cálculos de medidas de dispersão, para uma população, assista a esse [vídeo](https://www.youtube.com/watch?v=UdrtnBGSeSw&index=10&list=PLTtZUJqLYbCI9oBV_22ycFJsVeddG4ixd) e observe as falas do professor quanto a comparação dos dois conjuntos de dados.
- 2) Para entender a diferença do **cálculo da variância populacional** e da **variância amostral (em que a divisão é por n-1)**, assista aos vídeos a seguir (procure assistir em ordem)

1º. **Variância Populacional**

[https://www.youtube.com/watch?v=ngNrB\\_7FLGo&list=PLTtZUJqLYbCI9oBV\\_22ycFJsVeddG4ixd&index=11](https://www.youtube.com/watch?v=ngNrB_7FLGo&list=PLTtZUJqLYbCI9oBV_22ycFJsVeddG4ixd&index=11)

2º. **Variância amostral**

[https://www.youtube.com/watch?v=VYKhA0iiC\\_0&list=PLTtZUJqLYbCI9oBV\\_22ycFJsVeddG4ixd&index=12](https://www.youtube.com/watch?v=VYKhA0iiC_0&list=PLTtZUJqLYbCI9oBV_22ycFJsVeddG4ixd&index=12)

3º. **Revisão e justificativa intuitiva para a divisão por n – 1**

[https://www.youtube.com/watch?v=MEvOeS84890&index=13&list=PLTtZUJqLYbCI9oBV\\_22ycFJsVeddG4ixd](https://www.youtube.com/watch?v=MEvOeS84890&index=13&list=PLTtZUJqLYbCI9oBV_22ycFJsVeddG4ixd)

**Algumas considerações sobre Desvio-padrão (DP) e Coeficiente de variação (CV)**

- A partir do desvio padrão e do coeficiente de variação é possível avaliar a homogeneidade do conjunto de dados e, conseqüentemente, se a média é uma medida representativa deste conjunto. Quanto maior forem os valores de desvio padrão e coeficiente de variação, menos representativa será a média. Se o coeficiente de variação for superior a 50% há alta dispersão o que aponta heterogeneidade dos dados. Por outro lado, quanto mais próximo de zero, mais homogêneo é o conjunto de dados e mais representativa será sua média.
- Para comparar a variabilidade entre grupos por meio do desvio-padrão ou variância, os conjuntos de dados tem que possuir mesmo número de observações; mesma unidade de medida; e mesma média.
- O coeficiente de variação, por ser uma medida dada em termos percentuais, é bastante adequado para comparar conjuntos com volume de dados ou unidades de medidas distintas.
- Uma desvantagem do coeficiente de variação é que ele deixa de ser útil quando a média se aproxima de zero, porque uma divisão por quase zero pode inflacionar o CV.
- Quando se verifica que a média não é uma boa medida para representar o conjunto de dados, opta-se pela mediana ou moda, não existindo uma regra para realizar esta escolha. O pesquisador, que entende o contexto de seus dados é quem define a melhor medida a ser considerada, tanto de posição, quanto de dispersão.

**Questão resolvida** (FUNDATEC - Auditor-Fiscal da Receita Estadual (SEFAZ RS)/2009)  
Análise as seguintes assertivas

- I. Média, moda e mediana são medidas de tendência central.
  - II. A amplitude de classe de um conjunto de dados é dada pela diferença entre o maior e o menor valor observado.
  - III. A mediana de um conjunto de dados é dado pelo valor que separa exatamente ao meio o conjunto de dados – 50% abaixo e 50% acima.
  - IV. O desvio-padrão é a raiz quadrada da variância.
- Quais estão corretas?
- A) Apenas I.
  - B) Apenas II.
  - C) Apenas II e III.
  - D) Apenas II e IV.
  - E) Apenas I, III e IV.

### **Resolução**

- I. Média, moda e mediana são medidas de tendência central. **Correta.**
- II. A amplitude de classe é dada pela diferença entre o maior e o menor valor observado no conjunto de dados. **Errada, porque a amplitude de classe diferença entre o maior e o menor valor da classe.**
- III. A mediana de um conjunto de dados é dada pelo valor que separa exatamente ao meio o conjunto de dados – 50% abaixo e 50% acima. **Correta.**
- IV. O desvio-padrão é a raiz quadrada da variância. **Correta.**



Est71.m4a

**Gabarito: E**

**Comentário em áudio:**

### **EXEMPLO ORIENTADO**

A Indústria Ecopneus pretende investir no aumento de produção de uma de suas linhas de pneus. A Linha Free é destinada a automóveis de passeio e a Linha Strong para veículos comerciais leves. A Ecopneus atua em todo Brasil, distribuindo cada uma de suas linhas de produtos para todos os estados da federação, nesse sentido, precisa avaliar informações sobre o mercado automotivo brasileiro, dentre outras, para que possa tomar a decisão de investir, com mais segurança.



Ana é executiva na Ecopneus, e a ela foi atribuída a tarefa de gerar informações que subsidiem a tomada de decisão da empresa, por meio de um relatório que deverá apresentar à direção.

Ana procurou por dados dos quais pudesse extrair informações, encontrando nos indicadores econômicos consolidados do Banco Central do Brasil, os dados que precisava para construir um relatório que indicasse em qual linha de pneus a empresa deveria investir; isso equivale a identificar, inicialmente, qual dos tipos de veículos (de passeio ou comerciais leves) tem tido maior volume de vendas no país e a regularidade dessa venda, já que a empresa não suportaria impactos sazonais negativos.

A direção da empresa é composta por pessoas com pouca ou nenhuma formação estatística, por esse motivo a apresentação do relatório deve ser simples, mas consistente.

#### **Elabore um relatório que contenha informações importantes para a tomada de decisão da Ecopneus, para isso:**

- Defina a pergunta a ser respondida e os dados que precisa para respondê-la.
- Procure por dados em fontes confiáveis.
- Organize os dados relevantes que encontrou, para que esta organização otimize a análise necessária para responder a questão inicial.
- Escolha formas gráficas adequadas para esta apresentação
- Resuma os dados em forma de medidas, focando nos seus significados aplicados a este contexto da Ecopneus.
- Ao final, escreva um texto que aponte para a solução, mas não determine a escolha da linha de pneus, pois provavelmente seriam necessárias outras análises, para além da estatística, nesse processo decisório.

Onde estão os dados? Busque-os! Esse estudo está muito mais próximo do que você enfrentará no mercado de trabalho do que a resolução de uma lista de exercícios – que obviamente não faz parte do mercado profissional.



U1\_Caso\_EcoPneus.  
pdf

**Guia para desenvolvimento do relatório**

## **Desafio**

Pense em um problema do seu interesse, que possa ser solucionado por meio de análise de dados quantitativos.

Busques esses dados, organize-os e calcule as medidas de síntese. A partir do tratamento desses dados e das medidas, elabore um texto que aborde o problema inicial, apontando para a solução, que deve estar fundamentada na análise descritiva que você fará.

Por exemplo, se você desejasse investir em ações da Petrobrás; poderia entrar no site da Bovespa e pesquisar uma série de valores dessa ação, ao longo do tempo (sendo que você definiria de quanto tempo seria sua análise); e a partir da análise gráfica, da média de valor da ação para o período que você escolheu e desvio-padrão poderia avaliar a viabilidade do investimento para você, considerando valor médio da ação, e o quanto os valores se dispersaram dessa média.

### **Ao realizar essa tarefa, você deverá:**

Postar o material selecionado no AVA, no espaço indicado, identificando a fonte de onde foi extraído, segundo as normas da ABNT.

## **Saiba mais**

Faça uma visita ao site do **IBGE** e no menu "Estatística" você terá acesso a estatísticas divididas em temas como: sociais, econômicas e multidomínio. (<https://www.ibge.gov.br/>)

## **Dica de Leitura**

Leia e avalie o artigo sob o título **Os Efeitos da Incerteza sobre a Atividade Econômica no Brasil**, de Ricardo de Menezes Barboza Eduardo Zilberman. Este estudo trata dos impactos da incerteza na economia brasileira, buscando capturar o nível de incerteza vigente na economia brasileira (incerteza doméstica) e em seus principais parceiros comerciais (incerteza externa). No desenvolvimento do estudo são utilizados alguns métodos, mas que em sua essência tomam os conceitos de média e desvio-padrão como conteúdo. Observe que, mesmo em um estudo que se utiliza de

estratégias mais sofisticadas do que uma análise descritiva simples, as medidas de síntese se fazem fundamentais, especialmente média e desvio-padrão.

### **Finalizando a Unidade**

Nesta unidade você estudou algumas aplicações reais da Estatística em contextos distintos espalhados pelo mundo, especialmente em nosso país.

A relevância das estatísticas descritivas tanto em situações cotidianas quanto profissionais fica evidenciada pela necessidade que temos delas para tomada de decisão em várias esferas da vida, afinal quem marca uma festa ao ar livre sem verificar as “previsões” da meteorologia, ou pelo menos as precipitações dos últimos dias, ou seja as estatísticas meteorológicas?

É importante salientar que a Estatística serve aos objetivos do pesquisador, é ele quem define os objetivos da pesquisa, as perguntas que pretende responder, quais dados são necessários, como os organizará, analisará e apresentará, escolhendo os gráficos, tabelas, medidas e aplicativos mais adequados.

O uso de planilhas eletrônicas ou pacotes estatísticos é essencial para o trabalho atual, o qual requer rapidez no tratamento de grandes volumes de dados, por outro lado a habilidade humana de realizar análises tem sido cada vez mais valorizada, num mercado permeado pela ciência de dados, e sempre ávido por extrair informações e gerar conhecimento.

### **Material de apoio**

Vídeos com resolução de problemas envolvendo estatística descritiva.

\* Todos os vídeos indicados foram produzidos pela autora desse texto e estão disponíveis em seu canal no Youtube.

- Gráfico de barras – interpretação <https://youtu.be/JdIb4TYKpAs>
- Medidas de síntese (cálculo e interpretação): média e desvio-padrão. Rol. <https://youtu.be/kx3jh16uzbg>
- Interpretação de média e desvio-padrão <https://youtu.be/ao0BLtriV2Y>
- Medidas de síntese (cálculo e interpretação): média, moda, desvio-padrão, coeficiente de variação. [https://youtu.be/ZM\\_FzJZIXhY](https://youtu.be/ZM_FzJZIXhY)

## **Referências**

### **Básicas**

LARSON, Ron; FARBER, Betsy. Estatística aplicada. 4 ed. São Paulo: Pearson Education do Brasil, 2015.

SARTORIS, Alexandre. Estatística e introdução à econometria. 2. ed. – São Paulo: Saraiva, 2013.

TRIOLA, Mario F. Introdução à estatística. 12. ed. – Rio de Janeiro : LTC, 2017.

### **Complementares**

BUSSAB, W. O.; MORETTIN, P. Estatística Básica. 8. Ed. São Paulo: Atual. 2013.

FREUND, John E. Estatística aplicada: economia, administração e contabilidade. 11. ed. – Porto Alegre : Bookman, 2007.

SHARPE, Norean R.; DE VEAUX, Richard d.; VELLEMAN, Paul F. Estatística aplicada administração, economia e negócios. Porto Alegre: Bookman, 2011.

ROSSI, José W. Econometria e séries temporais com aplicações a dados da economia brasileira / José W. Rossi, Cesar das Neves. - Rio de Janeiro : LTC, 2014.

TAVARES, Marcelo. Estatística aplicada. Universidade aberta do Brasil: Brasília, 2007.



## Unidade 2: **Probabilidade**

### Objetivos

Identificar o espaço amostral de um experimento probabilístico e eventos simples.  
Aplicar probabilidade clássica, probabilidade empírica e probabilidade subjetiva.  
Determinar a probabilidade do complemento de um evento.

### Introdução

Olá, seja muito bem-vindo(a) à Unidade 2!

Nesta unidade você iniciará o estudo da probabilidade. A Unidade está organizada em quatro seções que abordam, na sequência, os seguintes conteúdos:

1. Conceitos básicos: Espaço amostral. Experimento aleatório. Evento. Probabilidade.
2. Tipos de probabilidades: clássica, empírica e subjetiva.
3. Tipos de eventos: mutuamente exclusivos, complementares, independentes.
4. Probabilidade condicional.

O estudo destes conteúdos proporcionará a você o desenvolvimento de pensamento probabilístico que poderá ser mobilizado na resolução de problemas que envolvam chances de que algo ocorra, considerando determinadas características.

### Palavras-chave da Unidade

Espaço amostral, experimento, evento, probabilidade, eventos mutuamente exclusivos, eventos independentes, probabilidade condicional.

## Seção 1: **Conceitos básicos de probabilidade**

### **Experimentos probabilísticos (= aleatórios)**

Se o Centro de Previsão de Tempo e Estudos Climáticos (CPTEC/INPE) prevê que há chance de 80% de chuva em determinada cidade, ou um médico diz que há 10% de chance de cura para certo tipo de câncer tratado com quimioterapia, nesses casos, eles estão afirmando a possibilidade, ou probabilidade, de que um evento específico ocorra (chuva e cura do câncer por quimioterapia). Alguns processos de tomada de decisão podem se basear nessas probabilidades, como o de responder às perguntas

“Devo realizar a comemoração de aniversário na praia?”

“Devo tentar o tratamento por quimioterapia?”

A Estatística possui dois ramos, a estatística descritiva e a inferencial. Na Unidade 1, você estudou sobre o papel da Estatística descritiva, e o segundo ramo da Estatística, a Estatística inferencial, fundamenta-se na probabilidade, por isso é essencial o estudo desta unidade de forma consistente, tanto para cálculo de probabilidades quanto para realizar inferências, previsões baseadas em dados. Esse é o procedimento de todo mercado, atualmente. Não há mais espaço para decisões baseadas em experiência de mercado somente, mas sim em ciência de dados, a qual gera *insights* a partir do estudo dos dados, e pode inferir sobre variáveis de interesse, subsidiando decisões como a de investir ou não em determinada empresa, ou país, a partir dos dados sobre suas ações, e/ou índice de sua bolsa de valores.

### Definições

- **Experimento aleatório (ou probabilístico)** é uma ação, ou tentativa sujeita ao acaso, por meio da qual resultados específicos (contagens, medições ou respostas) são obtidos.
- **Resultado** é o produto de uma única tentativa em um experimento probabilístico.
- **Espaço amostral** é o conjunto de todos os resultados possíveis de um experimento probabilístico.
- **Evento** é um subconjunto do espaço amostral. Ele pode consistir em um ou mais resultados.

#### Notações

$P$  denota probabilidade.

$A, B, C, E$  denotam eventos.

$P(E)$  denota a probabilidade de que o evento  $E$  ocorra

### Exemplo

Em uma *startup* trabalham 6 pessoas Ana, Cléa, João, Pedro, Tony, Carlos. Considere que um colaborador tenha que participar de uma feira cujo tema é empoderamento feminino e tecnologia. O ideal é que se escolha uma das mulheres que trabalham nessa startup. Caso se faça uma escolha aleatória de um dos colaboradores, qual a probabilidade de que a pessoa escolhida seja do sexo feminino?

<a href="https://br.freepik.com/fotos-vetores-gratis/poster">Poster foto criado por creativeart - br.freepik.com</a>



*Nesse contexto, como identificar o que é experimento, espaço amostral, resultado, evento e ainda calcular a probabilidade desejada?*

Então vamos lá... primeiro identificando cada conceito, neste problema.

- Experimento: Escolher um colaborador ao acaso.
- Espaço amostral (todas as possibilidades):  $S = \{\text{Ana, Cléa, João, Pedro, Tony, Carlos}\}$ , ou seja, 6 possibilidades de resultado, quando se escolhe um colaborador ao acaso, nessa startup.
- Resultado: é o produto de uma única tentativa, ou seja, o sexo do colaborador escolhido em uma tentativa.
- Evento (conjunto de possibilidades favoráveis = colaborador escolhido ser do sexo feminino):  $\{\text{Ana, Cléa}\}$ , ou seja, 2 possibilidades de resultados favoráveis.
- Cálculo da Probabilidade de que a pessoa escolhida, ao acaso, seja do sexo feminino:

Chamando de  $A$  o evento: pessoa escolhida ser do sexo feminino, e de  $P(A)$  a probabilidade de  $A$  ocorrer.

$$P(A) = \frac{\text{conjunto de possibilidades favoráveis (quantidade de colaboradores do sexo feminino)}}{\text{todas as possibilidades (número total de colaboradores)}} = \frac{2}{6}$$

$$P(A) = \frac{2}{6} \quad \text{simplificando } (\div 2) \rightarrow P(A) = \frac{1}{3} = 0,333 \dots$$

Portanto, nessa startup, a probabilidade de se escolher, aleatoriamente um colaborador para participar da feira, do sexo feminino é de  $0,333\dots = 33,33\%$ .

## Seção 2: Tipos de probabilidades

*A estratégia que você utilizará para calcular a probabilidade de algo ocorrer depende do tipo de probabilidade que o problema está tratando.*

Existem três tipos:

- **Probabilidade clássica**
- **Probabilidade empírica**
- **Probabilidade subjetiva**

Neste texto, considere que a probabilidade de ocorrência de um evento  $E$  é escrita como  $P(E)$  e lê-se "probabilidade do evento  $E$ ".

### Definição

**Probabilidade clássica** (ou **teórica**) é adequada para tratar problemas em que cada resultado em um espaço amostral  $A$  é igualmente possível de ocorrer (eventos equiprováveis). A probabilidade clássica para um evento  $E$  é dada por:

$$P(E) = \frac{\text{número de elementos no evento } E}{\text{número de elementos no espaço amostral } A}$$

Usando as notações  $n(E)$  e  $n(A)$ , temos

$\left. \begin{array}{l} \text{número de elementos em } E = n(E) \\ \text{número elementos em } A = n(A) \end{array} \right\} \rightarrow$

$$P(E) = \frac{n(E)}{n(A)}$$

### Exemplos

- a) Qual a chance de você ganhar uma aposta com um amigo, que depende de jogar uma moeda e observar se deu cara ou coroa, sendo que você escolheu coroa?



*Resolução*

Como cada moeda tem duas faces, se ela for honesta, há duas possibilidades de resultado, então, em cada jogada:

Espaço amostral = {cara, coroa}  $\rightarrow n(A) = 2$

Evento (face voltada para cima ser coroa) = {coroa}  $\rightarrow n(E) = 1$

$$P(E) = \frac{n(E)}{n(A)} = \frac{1}{2} = 0,5 = 50\%$$

- b) Qual a probabilidade de, em uma gravidez não gemelar, o sexo biológico do bebê ser masculino?

*Resolução*

Em uma gravidez não gemelar, o bebê terá duas possibilidades de sexo, ou seja, o

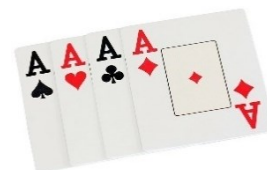
Espaço amostral = {masculino, feminino}  $\rightarrow n(A) = 2$

Evento (bebê do sexo masculino) = {masculino}  $\rightarrow n(E) = 1$

$$P(E) = \frac{n(E)}{n(A)} = \frac{1}{2} = 0,5 = 50\%$$

<https://pixabay.com/pt/photos/%C3%A1s-aces-quatro-diamantes-cora%C3%A7%C3%B5es-164029/>

- c) Qual a probabilidade de você retirar uma carta de um baralho padrão e o naipe da carta ser o de copas?



*Resolução*

Um baralho padrão é formado por 52 cartas, distribuídas em 4 naipes, cada naipe com 13 cartas, então temos 13 possibilidades de tirar cartas de copas

Espaço amostral: formado por 52 cartas  $\rightarrow n(A) = 52$

Evento (carta de copas): 13 cartas de copas  $\rightarrow n(E) = 13$

$$P(E) = \frac{n(E)}{n(A)} = \frac{13}{52} = 0,25 = 25\%$$

Todos esses três eventos foram calculados pela estratégia da probabilidade clássica, pois tratam de eventos equiprováveis, isto é, de eventos cujas chances de ocorrência são as mesmas.

Se um experimento for repetido muitas vezes, formam-se padrões regulares. Esses padrões permitem determinar a probabilidade, chamada de **probabilidade empírica**, que pode ser utilizada mesmo quando os eventos não são equiprováveis.

**Probabilidade empírica** (ou **estatística**) é baseada em observações obtidas de experimentos aleatórios. A probabilidade empírica de um evento  $E$  é a frequência relativa do evento  $E$  dividida pela frequência total.

$$P(E) = \frac{\text{frequência do evento } E}{\text{frequência total}}$$

### Exemplos

1) Um supermercado está conduzindo uma pesquisa pela internet com indivíduos selecionados ao acaso para determinar com que frequência eles compram produtos de encontrados em mercados, por meio da internet. Até o momento, 2.500 pessoas foram

pesquisadas. Os resultados foram registrados na tabela de distribuição de frequências abaixo.

Resposta	Ocorrências
Nunca	1100
Sempre	223
Raramente	625
Regularmente	552

Considerando os dados atuais qual é a probabilidade de que a próxima pessoa pesquisada compre regularmente produtos vendidos em mercados, pela internet?

*Resolução*

O evento é uma resposta "Regularmente".

A frequência desse evento é igual a 552 e a frequência total é 2.500 (total de pessoas pesquisadas)

A probabilidade empírica de a próxima pessoa responder "regularmente" é calculada por:

$$P(\text{Regularmente}) = \frac{552}{2500} = 0,2208 \rightarrow P(\text{Regularmente}) = 0,2208 \times 100 = 22,08\%.$$

Portanto, a probabilidade de que a próxima pessoa pesquisada responda que compra "regularmente" produtos de mercado, pela internet, é de 22,08% .

2) Uma pesquisa perguntou a 500 universitários o que fariam após a formatura: se procurariam emprego ou fariam pós-graduação, ou ambos. As respostas constam no quadro, abaixo.

Resposta	Ocorrências
Emprego	422
Pós-graduação	50
Emprego e pós-graduação	28

A partir dos dados da pesquisa, determine quais as probabilidades de que o próximo estudante pesquisado responda que:

- Procurará emprego
- Cursará pós-graduação
- Procurará emprego e cursará pós-graduação

*Resolução*

a) *Emprego*

$$P(\text{Emprego}) = \frac{28}{500} = 0,844$$

84,4% é a probabilidade de que um estudante pesquisado pretenda procurar emprego após sua formatura.

b) *Pós-graduação*

$$P(\text{Pós – graduação}) = \frac{50}{500} = 0,1$$

10% é a probabilidade de que um estudante pesquisado pretenda cursar pós-graduação após sua formatura.

c) *Emprego e pós-graduação*

$$P(\text{Emprego e Pós – graduação}) = \frac{28}{500} = 0,056$$

5,6% é a probabilidade de que um estudante pesquisado pretenda procurar emprego e cursar pós-graduação, simultaneamente, após sua formatura.

### Lei dos Grandes Números

À medida que um experimento é repetido mais e mais vezes, a probabilidade dada pela frequência relativa de um evento tende a se aproximar da probabilidade real.

**Probabilidade subjetiva** é a probabilidade que resultante de conjeturas e de estimativas intuitivas. Por exemplo, observando as condições de saúde de um paciente e a gravidade da doença, um médico pode sentir que o paciente tem 70% de chance de recuperação; ou o reitor de uma universidade pode prever que a chance de os funcionários entrarem em greve é de 0,15.

Outro exemplo seria a probabilidade de que o governo altere a política econômica, o que pode ser mais provável em períodos de crise. Enfim, a probabilidade subjetiva tem origem não em dados, mas na experiência, na percepção sobre cada evento, sobre o qual muitas vezes nem seria possível aplicar outro tipo de probabilidade.

### Amplitude das probabilidades

- A probabilidade de um evento  $E$  ocorrer sempre assume uma valor de 0 a 1, ou seja,  $0 \leq P(E) \leq 1$ .
- Se for certo que um evento  $E$  ocorrerá, então  $P(E) = 1$ .
- Se for impossível um evento  $E$  ocorrer, então  $P(E) = 0$ .
- Se as chances de um evento  $E$  ocorrer, e de não ocorrer, forem iguais, então  $P(E) = 0,5$ .

### Seção 3: Tipos de Eventos

#### Eventos complementares

A soma das probabilidades de todos os resultados em um espaço amostral é igual a 1 ou 100%, e nesse sentido, quando você sabe a probabilidade de um evento  $E$  ocorrer, pode também encontrar a probabilidade do evento  $E$  não ocorrer, e essa probabilidade de “não  $E$ ” é chamada de probabilidade do **complemento de  $E$** , ou do **evento complementar de  $E$** .

#### Definição

O **evento complementar de  $E$**  é o conjunto de todos os resultados em um espaço amostral que não estão incluídos no evento  $E$ . O complemento do evento  $E$  é denotado por  $\bar{E}$  e é lido como “não  $E$ ”.

#### Exemplo

Considere uma carteira de investimentos composta por 12 ações, 20 debêntures e 18 títulos públicos. Um ativo é selecionado aleatoriamente dessa carteira. Qual a probabilidade de o ativo selecionado não ser uma debênture?

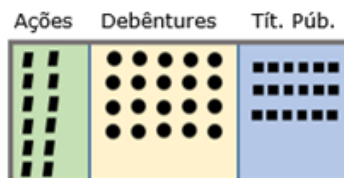
#### Resolução

Chamando  $A$  = ativo do tipo ação;  $D$  = ativo do tipo debênture e  $T$  = ativo do tipo título público, temos:

Ativo	Ocorrências
Ação	12
Debênture	20
Título público	18

$$\rightarrow P(A) = \frac{12}{50} = 0,24 \quad ; \quad P(D) = \frac{20}{50} = 0,4 \quad ; \quad P(T) = \frac{18}{50} = 0,36$$

E a representação gráfica pode ser



Fonte: Elaborada pela autora

Nesse caso, a probabilidade de **ocorrer  $D$**  é:  $P(D) = \frac{20}{50} = 0,4 = 40\%$

E a probabilidade solicitada, que é a de **não ocorrer  $D$** , ou seja, ( $\bar{D}$ ) é:

$$P(\bar{D}) = 1 - 0,4 = 0,6 = 60\%$$



Aplicando a definição de eventos complementares:

O **evento complementar do evento E** é o conjunto de todos os resultados em um espaço amostral que não estão incluídos no evento E.

O **evento complementar do evento D** ( $D = \text{ser debênture}$ ) é o conjunto de todos os resultados no espaço amostral que **NÃO estão no evento D** (12 ações + 18 títulos públicos).

**Notações: União ( $\cup$ ), Intersecção ( $\cap$ )**

**União de eventos**  $P(A \cup B) = P(\text{A ou B})$

**Intersecção de eventos**  $P(A \cap B) = P(\text{A e B})$

### Eventos mutuamente exclusivos

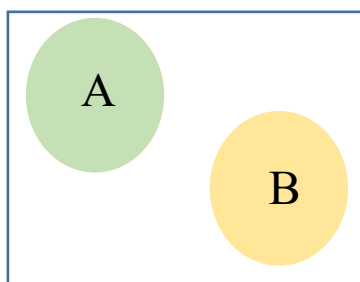
#### Definição

Dois eventos A e B são **mutuamente exclusivos** quando A e B não puderem ocorrer ao mesmo tempo, ou seja, a ocorrência de A exclui a ocorrência de B, e vice-versa.

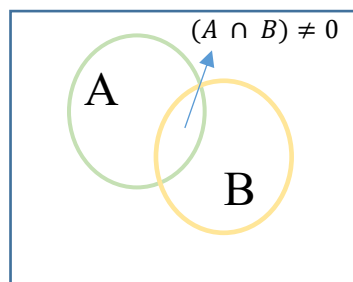
Quando eventos **A e B são mutuamente exclusivos**, eles não possuem resultados em comum, então  $P(A \cap B) = 0$ .

Se os eventos **A e B não são mutuamente exclusivos** eles possuem resultados em comum e  $P(A \cap B) \neq 0$ .

Eventos mutuamente exclusivos



Eventos **não** mutuamente exclusivos



Fonte: Elaborada pela autora

### Exemplos

Considere os eventos:

- Evento A: selecionar aleatoriamente um estudante do sexo masculino.
- Evento B: selecionar aleatoriamente um graduando Ciências econômicas.
- Evento C: selecionar aleatoriamente um estudante do sexo feminino.

Identifique se os eventos são, ou não, mutuamente exclusivos:

- a) A e B
- b) A e C
- c) B e C

*Resolução*

- a) *A e B: não são mutuamente exclusivos, porque um estudante pode cursar Ciências econômicas e ser do sexo masculino (é homem e cursa Economia)*
- b) *A e C: são mutuamente exclusivos, porque se ocorrer A (ser do sexo masculino), não ocorrerá B (ser do sexo feminino) ao mesmo tempo.*
- c) *B e C: não são mutuamente exclusivos, porque um estudante pode cursar Ciências econômicas e ser do sexo feminino (é mulher e cursa Economia).*

### A regra da adição

#### A regra da soma para a probabilidade de A ou B = $P(A \cup B)$

A probabilidade de que os eventos A ou B ocorram, é igual a probabilidade da união dos eventos, ou seja,  $P(A \cup B)$ , e é dada por:

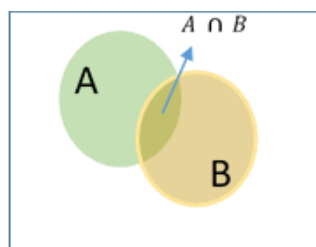
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

**Se os eventos A e B forem mutuamente exclusivos**, então a regra pode ser simplificada, porque como não há intersecção entre os eventos A e B, não é necessário diminuí-la, e a regra fica

$$P(A \cup B) = P(A) + P(B)$$

Esta regra simplificada pode ser estendida para qualquer número de eventos mutuamente exclusivos.

Então, para encontrar a probabilidade de um evento **ou** outro ocorrer, você deve somar as probabilidades individuais de cada evento e subtrair a probabilidade de ambos ocorrerem. Conforme mostrado no diagrama abaixo, chamado de diagrama de Venn, subtrair  $P(A \cap B)$  compensa a dupla contagem da probabilidade dos resultados que ocorrem em A e em B, ao mesmo tempo.



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

### Exemplo

Os volumes de vendas atingidos, em reais, pela Loja EconomiTech e o número de meses em cada nível de vendas, nos últimos três anos, estão registrados no quadro, abaixo.

Volume de vendas (dólares)	Meses
0 - 24.999	3
25.000 - 49.999	5
50.000 - 74.999	4
75.000 - 99.999	8
100.000 - 124.999	10
125.000 - 149.999	2
150.000 - 174.999	3
175.000 - 199.999	1

Considerando esse padrão de vendas, qual seria a probabilidade de que a EconomiTech vendesse entre US\$ 75.000,00 e US\$ 124.999,00?

#### Resolução

Vender entre US\$ 75.000,00 e US\$ 124.999,00 envolve dois níveis de vendas, e como esses eventos (vender entre 75.000 - 99.999) e (vender entre 100.000 - 124.999) não podem ocorrer ao mesmo tempo, esses eventos são mutuamente exclusivos, e então você pode utilizar a regra da soma das probabilidades.

Chamando de:

Evento  $A$  = vender entre 75.000 - 99.999

Evento  $B$  = vender entre 100.000 - 124.999

Temos que  $P(A \text{ ou } B) = P(A) + P(B) = \frac{8}{36} + \frac{10}{36} = \frac{18}{36} = 0,5$

Portanto, a probabilidade de que a EconomiTech vendesse entre US\$ 75.000,00 e US\$ 124.999,00 seria de 50%.

### Eventos Independentes

**Dois eventos são independentes** quando a ocorrência de um deles não afeta a probabilidade de ocorrência do outro.

Eventos que não são independentes são **dependentes**.

### Regra do produto para eventos independentes

Se os **eventos A e B forem independentes**, então a probabilidade de A e B ocorrerem é dada pelo produto

$$P(A \cap B) = P(A) \cdot P(B)$$

Essa regra pode ser estendida para qualquer número de eventos independentes.

### Exemplos de eventos dependentes e independentes

Determine se os eventos são independentes ou dependentes.

- 1) Selecionar uma rainha (evento A) de um baralho normal com 52 cartas, sem reposição, e na sequência selecionar um rei (evento B) do baralho.
- 2) Dirigir a mais de 200 Km/h (evento A) e então sofrer um acidente de carro (evento B).
- 3) Servidor público investir em produtos de uma financeira pequena que oferece altas taxas de remuneração do capital (evento A), e servidor público receber a remuneração referente ao exercício de suas funções, mensalmente (evento B).
- 4) Probabilidade de que as ações da PETROBRÁS aumentem 15%, se o preço do barril de petróleo subir 15% previamente.

#### Resolução

1) *Eventos dependentes. Em um baralho há 52 cartas, distribuídas em 4 naipes; sendo uma rainha e um rei de cada naipe, portanto 4 rainhas e 4 reis. Dessa forma, na primeira retirada (selecionar uma carta e ser rainha = evento A), temos que  $P(A) = \frac{4}{52}$ , e como não houve reposição da carta retirada do baralho, o espaço amostral diminuiu, o baralho passou a ter 51 cartas; e assim a probabilidade da ocorrência de B (selecionar um rei) é  $P(B) = \frac{4}{51}$ , ou seja, a ocorrência de A (selecionar uma carta e ser rainha) interferiu na probabilidade da ocorrência de B (selecionar uma carta e ser rei) porque foi retirado um elemento do espaço amostral, sem reposição, ou seja, o espaço amostral diminuiu.*

*Se houvesse reposição da primeira carta selecionada, os eventos seriam independentes, porque o espaço amostral seria o mesmo, 52 cartas, e então  $P(A) = \frac{4}{52} = P(B)$ .*

- 2) *Eventos dependentes. A dirigir a 200km/h interfere na probabilidade de sofrer um acidente.*
- 3) *Eventos independentes. Um servidor público investir de forma arriscada, não interfere na remuneração mensal pelo exercício de suas funções.*
- 4) *Eventos dependentes. O preço do barril de petróleo subir 15% previamente, interfere na probabilidade de que as ações da PETROBRÁS aumentem 15%.*

### **Eventos mutuamente exclusivos x Eventos independentes**

Não confunda eventos mutuamente exclusivos com eventos independentes!

O fato de dois eventos serem independentes não quer dizer que eles sejam mutuamente exclusivos.

**Dois eventos mutuamente exclusivos são dependentes**, obrigatoriamente, pois a ocorrência de um implica a não ocorrência do outro.

Na próxima seção, você estudará a probabilidade condicional, e então terá uma compreensão completa da regra do produto para eventos que ocorrem em sequência, sejam eles independentes, ou não.

## **Seção 4: Probabilidade Condicional**

Nesta seção, você estudará sobre a probabilidade de dois eventos ocorrerem em sequência, ou seja, como determinar a ocorrência de um evento sabendo que um outro evento ocorreu. Para alguns contextos a ocorrência do primeiro evento não interfere na probabilidade de ocorrência do segundo evento (eventos independentes), em outros contextos há essa interferência (eventos dependentes).

Por exemplo: Qual a probabilidade de ocorrer deslizamentos em comunidades localizadas em morros hoje, sabendo que ocorreu chuva de alta intensidade ontem?

Os eventos em sequência são:

1º. Ocorreu chuva de alta intensidade ontem

2º. Ocorrer deslizamento em comunidades localizadas em morros

Nesse contexto, é fácil perceber que a ocorrência do primeiro evento (chuva muito forte ontem), interfere na probabilidade de ocorrer o segundo evento (deslizamentos hoje).

Note que é dado que o primeiro evento ocorreu, ou seja, é certo, já sabemos que o primeiro evento ocorreu (ou que ocorrerá certamente)! E estamos interessados em calcular a probabilidade de ocorrência do segundo evento. É disso que se ocupa a probabilidade condicional: Sob a condição de que algo já ocorreu (ou ocorrerá), determinar a probabilidade de ocorrência de um segundo evento.

### Importante!

Probabilidade condicional não é igual a probabilidade de ocorrer uma intersecção de eventos, ou seja, probabilidade de ocorrer o evento A e o evento B (ao mesmo tempo), é diferente de probabilidade de ocorrer A sabendo que B já ocorreu (eventos em sequência).

Probabilidade da Intersecção de eventos:  $P(A \cap B) \rightarrow$  Probabilidade de eventos ocorrem ao mesmo tempo.

Probabilidade condicional:  $P(A|B) \rightarrow$  evento B já ocorreu e evento A terá ou não a probabilidade impactada pela ocorrência de B?

### Exemplos

Evento A: chover  
Evento B: aparecer neblina  
Evento C: cair a temperatura

Probabilidade de chover e aparecer neblina:

$$P(\text{ocorrer } A \text{ e ocorrer } B) = P(A \cap B)$$

Probabilidade de aparecer neblina, dado que caiu a temperatura:

$$P(\text{ocorrer } B \text{ sabendo que já ocorreu } C) = P(B|C)$$

### Definição

Uma **probabilidade condicional** é a probabilidade de um evento ocorrer, dado que outro evento já tenha ocorrido (ou seja certo que ocorrerá).

A probabilidade condicional de o evento A ocorrer, dado que o evento B tenha ocorrido, é denotada por  $P(A|B)$  e lê-se "probabilidade de A, dado B".

O cálculo da probabilidade condicional é dado por

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

E pelo Teorema de Bayes 
$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Logo, você pode calcular a **probabilidade condicional** de um evento ocorrer pelas duas expressões, regra do produto e Teorema de Bayes

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$$

### Exemplo

O quadro, abaixo, apresenta os resultados de um estudo no qual os pesquisadores estudaram o QI de um jovem e a presença de um gene específico nele.

Encontre a probabilidade de que o jovem possua um QI alto, dado que ele tem o gene.

QI	Gene presente	Gene ausente	Total
Alto	33	19	52
Normal	39	11	50
Total	72	30	102

(adaptado de LARSON, 2015)

### Resolução

A probabilidade solicitada é a de que um jovem possua um QI alto, dado que ele tem o gene (condição). Então, chamando QI alto =  $QIa$  e gene presente = *tem gene*, queremos:

$$P(QIa|tem\ gene) = ?$$

Usando a regra do produto  $P(A|B) = \frac{P(A \cap B)}{P(B)}$ , temos

$$P(B) = P(tem\ gene) = \frac{72}{102} = 0,7058$$

$$P(A \cap B) = P(QIa \cap tem\ gene) = \frac{33}{102} = 0,3235$$

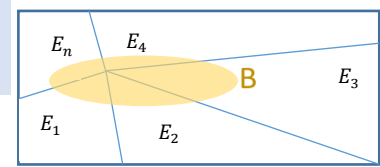
$$P(QIa|tem\ gene) = \frac{0,3235}{0,7058} = 0,458$$

45,8% é a probabilidade de que um jovem possua um QI alto, dado que ele tem o gene.

**Teorema da Probabilidade Total**

Considere os eventos  $E_1, E_2, E_3, \dots, E_n$  que constituem uma partição do espaço amostral  $S$ , ou seja:

- $E_1 \cup E_2 \cup E_3 \cup \dots \cup E_n = S$
- $P(E_i) > 0$ , para todo  $i=1, 2, 3, \dots, n$ .
- $E_i \cap E_j = \emptyset$  para  $i \neq j$ . (são mutuamente exclusivos)



Dessa forma, se  $B$  representa um evento, temos o seguinte teorema, conhecido como Teorema da probabilidade total

$$\sum_{i=1}^n P(E_i \cap B) = \sum_{i=1}^n P(E_i) \cdot P(B|E_i)$$

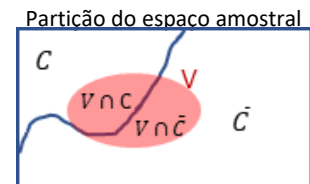
**Exemplo**

Um atleta africano tem 60% de probabilidade de vencer uma maratona, se ela for realizada em um dia de calor. Caso não faça calor durante a prova, sua probabilidade de vencer cai para 25%. Se a previsão meteorológica indicar que a chance de estar calor, durante a prova, é de 30%, qual será a probabilidade deste atleta ganhar a maratona?

*Resolução*

Aplicando o teorema da probabilidade total, pois estar calor e não estar calor são eventos mutuamente exclusivos temos a representação

- $E_1 = C = \text{estar Calor}$
- $E_2 = \bar{C} = \text{não estar Calor}$
- $B = V$



É um caso de probabilidade condicional (estar calor), podemos usar

$P(A|B) = \frac{P(A \cap B)}{P(B)}$  e isolando a intersecção temos  $P(A \cap B) = P(A|B) \cdot P(B)$

Chamando os eventos de

$V = \text{vencer a corrida}$  ;  $C = \text{estar calor}$  ;  $\bar{C} = \text{não calor}$  ;

Sabemos que

$P(V|C) = 0,60$  e  $P(V|\bar{C}) = 0,25$  previsão  $P(C) = 0,30 \rightarrow P(\bar{C}) = 0,70$

Queremos saber a chance de vencer (com calor ou sem calor), então aplicando o teorema da probabilidade total:

$P(V) = P(V \cap C) + P(V \cap \bar{C})$ , como  $P(V \cap C) = P(V|C) \cdot P(C)$  e  $P(V \cap \bar{C}) = P(V|\bar{C}) \cdot P(\bar{C})$

$P(V) = P(V|C) \cdot P(C) + P(V|\bar{C}) \cdot P(\bar{C})$

$P(V) = (0,60 \cdot 0,30) + (0,25 \cdot 0,70) = 0,18 + 0,175 = 0,355 = 35,5\%$  é a probabilidade de o atleta africano vencer a maratona (com ou sem calor).



**Exemplo - Resolução em vídeo** <https://www.youtube.com/watch?v=4fCoH5aIx50>

A inscrição para seleção para um cargo público X, de um órgão do Poder Executivo, fica aberta por poucos dias e impõe três condições a serem comprovadas para que os candidatos possam efetuar a inscrição, que são:

- Possuir experiência profissional;
- Apresentar negativa de antecedentes criminais, atualizada;
- Possuir curso de graduação.

Basta que uma delas não seja comprovada, para que o candidato tenha sua inscrição negada. Admita que as probabilidades dessas condições não serem comprovadas, por um candidato sejam:

- 10% (Experiência),
- 25% (Negativa antecedentes criminais)
- 20% (Graduação).

Considerando somente estas três condições, supondo que os demais elementos da inscrição estejam em conformidade com o exigido, determine:

- a) Qual a probabilidade de que candidato consiga efetivar sua inscrição?
- b) Qual a probabilidade de que apenas uma das condições anteriores não tenha sido comprovada, dado que o candidato não pode efetivar sua inscrição?

*Resolução*

*Temos que*

$$P(\bar{E}) = 0,10 \rightarrow P(E) = 0,90$$

$$P(\bar{A}) = 0,25 \rightarrow P(A) = 0,75$$

$$P(\bar{G}) = 0,20 \rightarrow P(G) = 0,80$$

a)  $I =$  efetivar a inscrição

$$P(I) = P(E \text{ e } A \text{ e } G) = P(E) \times P(A) \times P(G) \text{ porque os eventos são independentes}$$

$$P(I) = 0,54 = 54\%$$

b)  $\bar{I} =$  não conseguir se inscrever  $1C =$  uma condição não comprovada

$$P(1C|\bar{I}) = 1 - P(I) = 1 - 0,54 \rightarrow P(\bar{I}) = 0,46$$

$$P(1C) = P(\bar{E} \text{ e } A \text{ e } G) + P(E \text{ e } \bar{A} \text{ e } G) + P(E \text{ e } A \text{ e } \bar{G})$$

$$P(1C) = (0,10 \times 0,75 \times 0,80) + (0,90 \times 0,25 \times 0,80) + (0,90 \times 0,75 \times 0,20)$$

$$P(1C) = 0,375$$

$$\text{Como } P(1C|\bar{I}) = \frac{P(1C \cap \bar{I})}{P(\bar{I})} = \frac{0,375}{0,46} = 0,815$$

81,5% é a probabilidade de que um candidato não conseguir se inscrever por não atender **apenas uma** das condições para a inscrição.

## Desafio

Formule um problema que envolva probabilidade condicional e probabilidade total. Você pode se inspirar em problemas de livros de probabilidade e estatística, ou elaborar o contexto a partir de probabilidades que você observa no mundo real. Resolva o problema de forma detalhada, indicando quais ideias/regras/fórmulas aplicou para desenvolver a resolução.

### **Ao realizar essa tarefa, você deverá:**

- Postar o material selecionado no AVA, no espaço indicado, identificando a fonte de onde foi extraído, segundo as normas da ABNT.

## Saiba mais

Acesse a matéria sob título "**Processo de recuperação gradual da economia foi interrompida, diz BC**", da Agência Brasil, e avalie o tipo de probabilidade que se aplica, especialmente quando esta trata dos indicadores disponíveis sugerirem **probabilidade** relevante de que o Produto Interno Bruto (PIB) tenha recuado ligeiramente no primeiro trimestre do ano. <http://agenciabrasil.abc.com.br/economia/noticia/2019-05/processo-de-recuperacao-gradual-da-economia-foi-interrompido-diz-bc>

## Dica de Leitura

Leia a Unidade 4 do livro **Estatística aplicada à administração, de Marcelo TAVARES**. Acompanhe os exemplos e problemas resolvidos para contribuir com a consolidação dos conteúdos envolvendo probabilidade. (<https://educapes.capes.gov.br/bitstream/capes/401408/1/PNAP%20-%20Bacharelado%20-%20Modulo%204%20-%20Estatistica%20Aplicada%20a%20Administracao%20-%203ed%202014%20-%20WEB%20-%20atualizado.pdf>)

## Finalizando a Unidade

Nesta unidade, você estudou conceitos e aplicações importantes sobre probabilidade, tanto conceitos básicos como as ideias de experimento, evento, espaço amostral, quanto pensamento probabilístico mais sofisticado quando se utilizou a probabilidade condicional. Conhecendo as ideias principais da probabilidade, você já tem condições de estudar as distribuições de probabilidades, que são especialmente aplicadas a contextos concretos de diversos campos do saber, como à economia, por exemplo; e um outro ramo da estatística, a Estatística inferencial.

Bons estudos!

## Material de apoio

Estas páginas apontam para conjuntos de videoaulas e outros objetos que contribuem para aprendizagem da probabilidade.

### Videoaulas da Khan Academy sobre Probabilidade

<https://sway.office.com/Z2n6e59UYs0ougjY?ref=Link>

Código de incorporação

```
<iframe width="760px" height="500px" src="https://sway.office.com/s/Z2n6e59UYs0ougjY/embed" frameborder="0" marginheight="0" marginwidth="0" max-width="100%" sandbox="allow-forms allow-modals allow-orientation-lock allow-popups allow-same-origin allow-scripts" scrolling="no" style="border: none; max-width: 100%; max-height: 100vh" allowfullscreen mozallowfullscreen msallowfullscreen webkitallowfullscreen></iframe>
```

### \*Videoaulas e outros objetos sobre Probabilidade

<https://sway.office.com/caUm4N64RIXmuNHV?ref=Link>

Código de incorporação

```
<iframe width="760px" height="500px" src="https://sway.office.com/s/caUm4N64RIXmuNHV/embed" frameborder="0" marginheight="0" marginwidth="0" max-width="100%" sandbox="allow-forms allow-modals allow-orientation-lock allow-popups allow-same-origin allow-scripts" scrolling="no" style="border: none; max-width: 100%; max-height: 100vh" allowfullscreen mozallowfullscreen msallowfullscreen webkitallowfullscreen></iframe>
```

\* essa página está sendo atualizada continuamente

Filme que apresenta um problema resolvido por probabilidade condicional.

## Referência Bibliográfica

### Básicas

LARSON, Ron; FARBER, Betsy. Estatística aplicada. 4 ed. São Paulo: Pearson Education do Brasil, 2015.

SARTORIS, Alexandre. Estatística e introdução à econometria. 2. ed. – São Paulo: Saraiva, 2013.

TRIOLA, Mario F. Introdução à estatística. 12. ed. – Rio de Janeiro : LTC, 2017.

### Complementares

BUSSAB, W. O.; MORETTIN, P. Estatística Básica. 8. Ed. São Paulo: Atual. 2013.

FREUND, John E. Estatística aplicada: economia, administração e contabilidade. 11. ed. – Porto Alegre : Bookman, 2007.

SHARPE, Norean R.; DE VEAUX, Richard d.; VELLEMAN, Paul F. Estatística aplicada administração, economia e negócios. Porto Alegre: Bookman, 2011.

ROSSI, José W. Econometria e séries temporais com aplicações a dados da economia brasileira / José W. Rossi, Cesar das Neves. - Rio de Janeiro : LTC, 2014.

TAVARES, Marcelo. Estatística aplicada. Universidade aberta do Brasil: Brasília, 2007.

## Unidade 3: Distribuições de Probabilidades

### Objetivos

- Aplicar conceitos de variável aleatória e distribuição de probabilidades na representação de problemas.
- Resolver problemas envolvendo distribuições de probabilidade discretas.
- Resolver problemas envolvendo distribuições de probabilidade contínuas.

### Introdução

Olá, seja muito bem-vindo(a) à Unidade 3!

Nesta unidade você estudará as distribuições de probabilidade. A Unidade está organizada em quatro seções que abordam, na sequência, os seguintes conteúdos:

1. Variáveis aleatórias.
2. Distribuições de probabilidades discretas
3. Distribuições de probabilidades contínuas.
4. Distribuições de probabilidades em planilha

A aplicação das distribuições de probabilidades é muito ampla, da biologia, passando pela engenharia e até fenômenos da economia podem ser estudados a partir de conhecimentos sobre distribuições de probabilidades, especialmente, da distribuição normal.

### Palavras-chave da Unidade

Distribuição de probabilidade, distribuição binomial, distribuição de Poisson, distribuição normal, variável aleatória.

### Seção 1: Variáveis aleatórias

O resultado de um experimento aleatório, em geral, é produto de uma contagem ou de uma medição. Quando isso ocorre, esse resultado é um possível valor de uma **variável aleatória**.

#### Variável aleatória

Uma **variável aleatória  $X$**  representa um valor numérico associado a cada resultado de um experimento aleatório.

A palavra *aleatória* indica que  $X$  é determinada em função de um objeto escolhido ao acaso. Há dois tipos de variáveis aleatórias: **discretas** e **contínuas**.

### Variável aleatória discreta

Uma variável aleatória é **discreta** quando tem um número finito ou contável de resultados possíveis que podem ser enumerados.

### Variável aleatória contínua

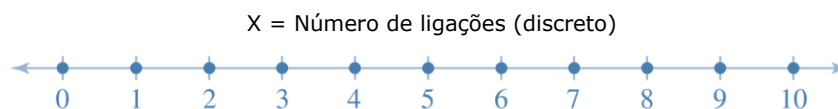
Uma variável aleatória é **contínua** quando tem um número incontável de resultados possíveis, representados por um intervalo na reta numérica (reta que representa o conjunto dos números reais).

#### Dica

Na maioria das aplicações práticas, as variáveis aleatórias discretas representam dados que podem ser contados, enquanto as variáveis aleatórias contínuas representam dados que são resultados de medições.

### Exemplos

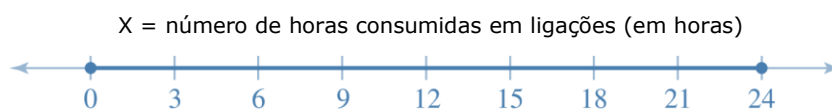
- 1) Considere uma pesquisa sobre "o número de ligações" que uma atendente de *call center* faz em um único dia. Os valores possíveis da variável aleatória  $X$  são 0, 1, 2, 3, 4 e assim por diante. Como o conjunto de resultados possíveis  $\{0, 1, 2, 3, \dots\}$  pode ser listado,  $X$  é uma **variável aleatória discreta**, e pode ser representada por pontos na reta numérica.



$X$  só pode assumir valores inteiros: 0, 1, 2, 3, ...

- 2) Uma forma diferente de conduzir a pesquisa seria a de medir "o tempo diário (em horas)" que uma atendente de *call center* consome fazendo ligações. O tempo gasto fazendo ligações pode ser qualquer número real de 0 a 24 (incluindo frações e decimais), então  $X$  é uma **variável aleatória contínua**.

Quando uma variável aleatória é discreta, você pode listar ou enumerar os valores possíveis que ela pode assumir. Por outro lado, é impossível listar todos os valores para uma **variável aleatória contínua**.



$X$  pode assumir qualquer valor de 0 a 24 (horas)

Você poderá representar os valores que  $X$  pode assumir em um intervalo na reta, mas não poderá enumerar todos os valores possíveis.

### **Média = Esperança matemática $E(X)$**

A média (ou valor esperado, ou ainda a “esperança” matemática) de uma variável aleatória  $X$ , denotada por  $E(X)$ , é uma medida que dá ideia de qual valor de  $X$  seria o esperado, caso o experimento ao qual a variável está associada fosse repetido inúmeras vezes.

**Para uma variável aleatória discreta**, o valor esperado  $E(X)$  é a média ponderada de todos os possíveis valores de  $X$  com pesos iguais às respectivas probabilidades desses valores, matematicamente,

$$E(X) = \sum x \cdot P(X = x)$$

**Para variáveis aleatórias contínuas**, o valor esperado é calculado pela seguinte fórmula:

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x) dx$$

### **Variância**

A variância de uma variável aleatória  $X$  é uma medida de sua dispersão estatística, e corresponde ao valor esperado do quadrado de quanto ela se afasta de seu valor esperado.

O valor dado por  $X - E(X)$  corresponde ao desvio de  $X$  em relação a sua média. Logo, para calcular a variância, quando  $X$  é uma variável aleatória discreta, vale que

$$Var(X) = E[(X - E(X))^2] = E(X^2) - [E(X)]^2$$

Quando  $X$  é uma **variável aleatória contínua**, recorremos ao cálculo integral:

$$Var(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 \cdot f(x) dx$$

É muito importante que você saiba diferenciar variáveis aleatórias discretas e contínuas, porque precisará dessa informação para avaliar a estratégia adequada para tratar cada

variável, mais especificamente, distribuições discretas e contínuas, como estudará a seguir.

## Seção 2: **Distribuições Discretas de probabilidade**

Para cada valor de uma variável aleatória discreta pode ser atribuída uma probabilidade. Ao associar cada valor da variável aleatória com a probabilidade correspondente, você construirá uma **distribuição discreta de probabilidade**.

### **Distribuições discretas**

Uma **distribuição discreta de probabilidade** lista cada valor possível que a variável aleatória pode assumir, com a probabilidade correspondente. Uma distribuição de probabilidade discreta deve satisfazer às seguintes condições:

1. A probabilidade de cada valor da variável aleatória discreta está de 0 e 1.

$$0 \leq P(X) \leq 1$$

2. A soma de todas as probabilidades é 1.

$$\Sigma P(X) = 1$$

As probabilidades podem ser indicadas por frequências relativas, assim uma distribuição de probabilidades discreta pode ser representada por meio de uma tabela, e graficamente por um histograma de frequências relativas.

### **Como construir uma distribuição discreta de probabilidade?**

Considere que  $X$  seja uma variável aleatória discreta com resultados possíveis  $X_1, X_2, \dots, X_n$ .

1. Construa uma distribuição de frequências para os resultados possíveis (tabela)
2. Calcule a soma das frequências.
3. Calcule a estimativa da probabilidade para cada resultado possível, dividindo sua frequência pelo total das frequências.
4. Verifique que cada probabilidade seja um número de 0 a 1, e que a soma seja igual a 1.



**Exemplo**

A Universidade A realizou uma pesquisa com alunos formandos, pela qual buscou saber a quantidade de horas que os formandos trabalhavam por dia.

Os resultados foram representados na tabela de distribuição de frequências, abaixo.

X	Frequência
4	45
6	50
8	60
10	20
12	25
Total	200

Observe que os valores possíveis da variável  $X$  são  $\{4, 6, 8, 10, 12\}$ , conforme observamos à esquerda da tabela. Na coluna à direita da tabela, está registrada a contagem da quantidade de vezes que cada valor ocorreu na pesquisa.

A distribuição de probabilidade de uma variável aleatória discreta é chamada de **função massa de probabilidade**. Nesse caso, pode-se especificar a probabilidade de a variável  $X$  ser igual a um determinado valor  $x$ , o que representa-se por  $P(X = x)$ .

Dividindo cada frequência absoluta pelo total das frequências, para cada um dos valores possíveis da variável  $X$ , você encontrará a tabela de distribuição de probabilidades, conforme, abaixo:

Para encontrar  $P(X = x)$

$$P(X = 4) = \frac{45}{200} = 0,225$$

$$P(X = 6) = \frac{50}{200} = 0,25$$

$$P(X = 8) = \frac{60}{200} = 0,3$$

$$P(X = 10) = \frac{20}{200} = 0,1$$

$$P(X = 12) = \frac{25}{200} = 0,125$$

X	Frequência	$P(X = x)$
4	45	0,225
6	50	0,25
8	60	0,3
10	20	0,1
12	25	0,125
Total	200	1

Nessa tabela, a terceira coluna, também chamada de frequência relativa, é o resultado da divisão da frequência de cada valor pelo total de casos contados (200). Sendo assim, representa a probabilidade de que, escolhendo-se um formando qualquer que tenha respondido à pesquisa, sua resposta tenha sido cada um dos valores da variável  $X$ .

Além da probabilidade associada a cada um dos valores da variável aleatória, também é possível calcular probabilidades para intervalos ou expressões lógicas que representem combinações dos eventos associados a cada valor de  $X$ . Por exemplo, se o interesse for pela probabilidade de se escolher ao acaso um formando e ele trabalhe menos de 8 horas por dia, isso corresponderia a:

$$P(X < 8) = P(X = 4) + P(X = 6) = 0,225 + 0,25 = 0,475$$

Portanto, a probabilidade de se escolher, ao acaso, um formando e ele trabalhe menos de 8 horas por dia é de 47,5%.

Ainda, se o interesse fosse na probabilidade de se escolher um formando, aleatoriamente, que trabalhe a partir de 8 horas por dia, ou seja:

$$P(X \geq 8) = P(X = 8) + P(X = 10) + P(X = 12) = 0,3 + 0,1 + 0,125 = 0,525$$

Por outro lado, você poderia fazer o cálculo da seguinte forma:

$$P(X \geq 8) = 1 - P(X < 8) = 1 - [P(X = 4) + P(X = 6)] = 1 - 0,475 = 0,525$$

Portanto, a probabilidade de se escolher, ao acaso, um formando que trabalhe a partir de 8 horas por dia é de 52,5%.

### Experimentos binomiais

São experimentos aleatórios para os quais os resultados de cada tentativa podem ser reduzidos a dois resultados: sucesso e fracasso. Por exemplo, quando um investidor compra uma ação, ele pode ter lucro ou não, com essa ação. Experimentos probabilísticos em que os resultados podem ser de dois tipos, sucesso ou fracasso, são denominados **experimentos binomiais**.

Um **experimento binomial** deve satisfazer as seguintes condições:

1. Há apenas dois resultados possíveis para cada tentativa, que podem ser classificados como sucesso (S) ou fracasso (F).
2. O experimento tem um número fixo de tentativas, em que cada tentativa é independente das outras.
3. A probabilidade de um sucesso é a mesma para cada tentativa.
4. A variável aleatória  $X$  conta o número de tentativas com sucesso.

Notações para experimentos binomiais

$n$  = Número de tentativas

$p$  = Probabilidade de sucesso em uma única tentativa

$q$  = Probabilidade de fracasso em uma única tentativa ( $q = 1 - p$ )

$X$  = A variável aleatória representa a contagem do número de sucessos em  $n$  tentativas:

$X = 0, 1, 2, 3, \dots, n$

**Exemplo**

Numa fábrica, a probabilidade de o processo produtivo gerar uma peça defeituosa é de 3%. Qual a probabilidade de que, num lote de 100 peças, haja 5 peças defeituosas?

*Resolução*

É dado no enunciado que  $p=0,03$ , que corresponde à probabilidade de ocorrer "sucesso", definido, neste caso, por mais estranho que possa parecer, como "fabricar uma peça defeituosa" – pois a pergunta que queremos responder é sobre "peças defeituosas". Logo,  $q = 1 - p = 1 - 0,03 = 0,97$ .

Como são lotes de 100 peças, temos  $n = 100$ . Por fim, como estamos interessados na probabilidade correspondente a 5 peças defeituosas, queremos saber  $P(X = 5)$ .

Substituindo os valores na fórmula da binomial, temos:

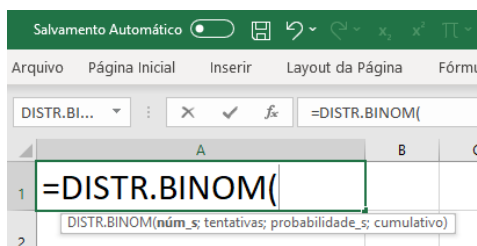
$$P(X = 5) = \frac{100}{5! \cdot 95!} \cdot 0,03^5 \cdot 0,97^{(100-5)}$$

$$= \frac{100 \cdot 99 \cdot 98 \cdot 97 \cdot 96 \cdot 95!}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 95!} \cdot 0,03^5 \cdot 0,97^{95} = 75.287.520 \cdot 0,0000000243 \cdot 0,05538$$

$$= 0,1013$$

Portanto, há uma probabilidade de, aproximadamente, 10% de que existam 5 peças defeituosas em um lote de 100 peças.

A forma usual de calcular essa probabilidade, atualmente, no mundo do trabalho é por meio de pacotes estatísticos, ou mesmo, de uma planilha Excel. Usando o Excel, você utilizaria a função estatística DISTR.BINOM. Veja na imagem abaixo, os argumentos que você deve colocar na função, eles aparecem na tela.



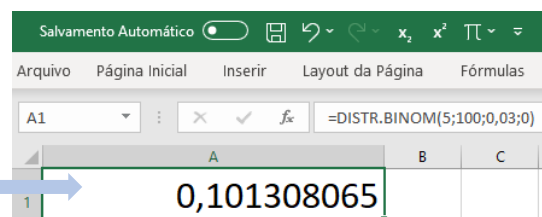
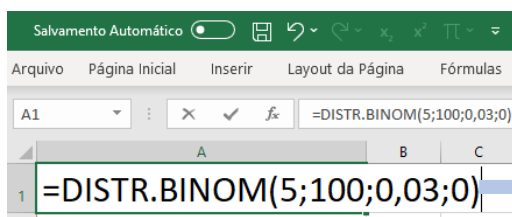
**núm\_s** = número de sucessos =  $X$

**tentativas** = número de tentativas =  $n$

**probabilidade\_s** = probabilidade de sucesso em uma única tentativa =  $p$

**cumulativo** = para a função probabilidade de massa usar 0 ou falso (calcula a probabilidade de ocorrerem exatamente  $X$  sucessos)

Abaixo, você vê a resolução do exemplo anterior, usando a função DISTR.BINOM do Excel.



Veja que, como não poderia deixar de ser, os resultados são os mesmos 10% de probabilidade de se escolher uma peça, ao acaso, e ela ser defeituosa, no lote de 100 peças.

Em um experimento binomial, você está interessado em determinar a probabilidade de um número específico de sucessos em um dado número de tentativas. Se você estiver interessado em saber a probabilidade de que um número específico de ocorrências aconteça dentro de um período contínuo (unidade de tempo, área ou volume), poderá utilizar a distribuição de Poisson. Por exemplo, para determinar a probabilidade de que um colaborador fique doente por 3 dias dentro de um mês, você pode usar a **distribuição de Poisson**.

### **Distribuição de Poisson**

A distribuição de Poisson é uma distribuição discreta de probabilidade de uma variável aleatória  $X$  que satisfaz as seguintes condições:

1. O experimento consiste em contar o número de vezes,  $X$ , que um evento ocorre em um dado intervalo contínuo. O intervalo pode ser de tempo, área, volume ou outro intervalo contínuo.
2. A probabilidade de um evento acontecer é a mesma para intervalos de mesmo tamanho.
3. O número de ocorrências em um intervalo é independente do número de ocorrências em outros intervalos não sobrepostos. A probabilidade de haver exatamente  $X$  ocorrências em um intervalo é:

$$P(X = x) = \frac{\lambda^x \cdot e^{-\lambda}}{X!}$$

Onde:

$e$  = um número irracional aproximadamente igual a 2,7

$\lambda$  = é o número médio de ocorrências por intervalo unitário de referência (taxa média de ocorrência)

### **Exemplo**

Um uma avenida, ocorrem em média 3 acidentes por mês. Qual será a probabilidade de que quatro acidentes ocorram em algum mês, nessa avenida?

*Resolução*

Veja que a variável número de acidentes é apresentada ao longo de um intervalo contínuo, um mês. O problema nos dá uma taxa de variação média desse número de acidentes por mês, que é  $\lambda = 3$  acidentes/mês

Aplicando a fórmula para o cálculo da probabilidade  $P(X = 4)$ , temos

$$P(X = x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!} \quad \rightarrow \quad P(X = 4) = \frac{3^4 \cdot e^{-3}}{4!} \approx \frac{81 \cdot 0,05081}{24} \approx \frac{4,1152}{24} \approx 0,17$$

Portanto, a probabilidade de que quatro acidentes ocorram em algum mês, nessa avenida, é de 17%, aproximadamente.

**Saiba mais**

Para a Distribuição de Poisson, temos:  $E(X) = Var(X) = \lambda$

Por exemplo, considere que em um banco a fila seja composta por 4 pessoas, em média, a cada hora. Isso significa que, considerando que esse fenômeno seja descrito pela Distribuição de Poisson,  $\lambda = 4$ . Sendo a variável aleatória  $X =$  quantidade de pessoas na fila do banco por hora, pode-se representar da seguinte maneira  $X \sim Poi(4)$  e além disso,  $E(X) = Var(X) = 4$ .

**Exemplo (resolução em vídeo)** [https://youtu.be/uYMeb6ga\\_X4](https://youtu.be/uYMeb6ga_X4)

O gestor de um edifício-garagem próximo ao centro da cidade estimou que a média da quantidade de carros que chegam em um período de 1 hora é de 49 carros.

A partir da análise desse contexto, obedecendo à distribuição de probabilidades de Poisson, calcule o desvio-padrão da distribuição.

*Resolução*

A situação proposta pode ser modelada como uma distribuição de Poisson. Para esse tipo de distribuição, é necessário apenas o parâmetro  $\lambda$ , que representa a taxa de ocorrência de "sucessos" em um determinado intervalo. Nesse caso,  $\lambda = 49$  carros em um intervalo 1 hora (intervalo contínuo), e como em uma distribuição de Poisson, a média e a variância são iguais a  $\lambda$ , temos  $\lambda = 49 =$  variância.

O desvio-padrão é a raiz quadrada da variância, e como a variância é igual a média, nessa distribuição, temos *desvio – padrão*  $= \sqrt{49} = 7$  carros/hora

*Desvio-padrão* = 7 carros/hora.

### Seção 3: Distribuições Contínuas de Probabilidade

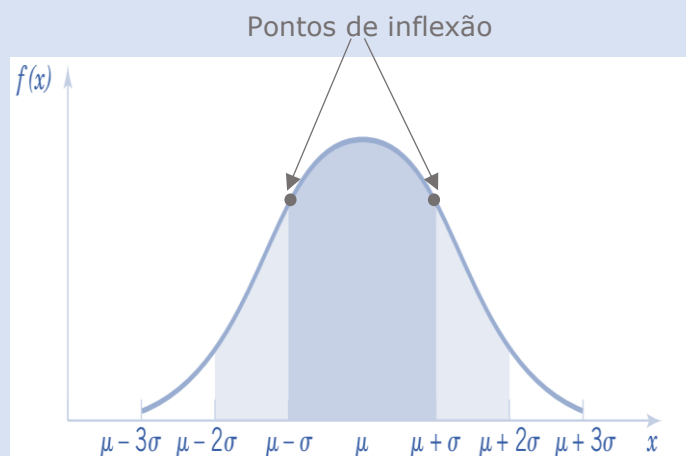
Na Seção 1, você pode estudar a diferença entre variáveis aleatórias discretas e contínuas, e aprendeu que uma variável aleatória contínua tem um número infinito de valores possíveis que podem ser representados por um intervalo em uma reta numérica.

Nesta seção, você estudará a mais importante das distribuições contínuas da estatística — a **distribuição normal**. A relevância das distribuições normais reside na ampla gama de aplicações, em diferentes contextos. As distribuições normais podem ser usadas para modelar muitos conjuntos de medidas na natureza, na indústria e nos negócios. Por exemplo, a pressão sanguínea sistólica dos humanos.

#### Distribuição Normal de probabilidade

Uma **distribuição normal** é uma distribuição de probabilidade contínua para uma variável aleatória  $X$ , cujo gráfico é chamado de **curva normal** (forma de sino), satisfazendo as propriedades listadas a seguir.

1. A média, a mediana e a moda são iguais.
2. Uma curva normal tem forma de sino e é simétrica em torno da média.
3. A área total sob a curva normal é igual a 1.
4. À medida que a curva normal se distancia da média, ela se aproxima do eixo *horizontal*, mas sem tocá-lo.
5. Entre  $\mu - \sigma$  e  $\mu + \sigma$  (no centro da curva), o gráfico se curva (tem concavidade para baixo). O gráfico tem concavidade para cima à esquerda de  $\mu - \sigma$  e à direita de  $\mu + \sigma$ . Os pontos nos quais o gráfico muda a orientação da concavidade são chamados de pontos de inflexão.



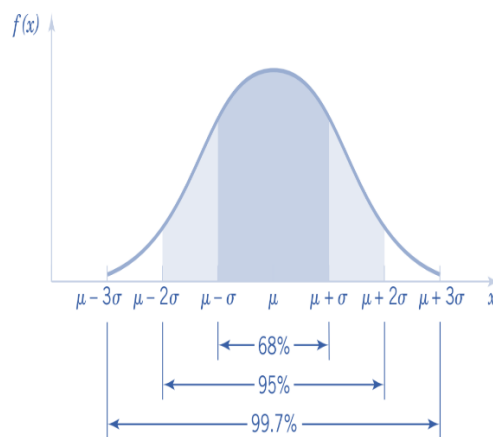
A função densidade de probabilidade da distribuição normal é dada por:

$$f(x) = \frac{1}{\delta\sqrt{2\pi}} \cdot e^{-\left(\frac{(x-\mu)^2}{(2\delta^2)}\right)}$$

Para definir completamente uma distribuição normal, são necessários dois parâmetros: a média e o desvio-padrão. É uma distribuição contínua, infinita para os dois lados, cujo gráfico tem formato de sino, simétrico ao redor da média e com uma largura que depende do desvio-padrão.

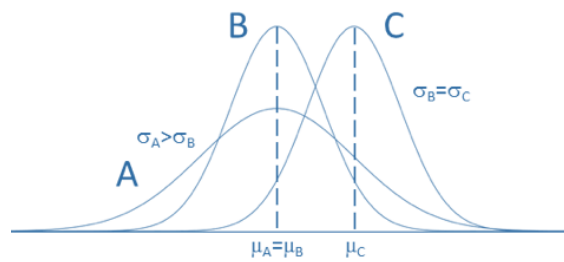
Uma característica importante observada empiricamente na Distribuição Normal são as porcentagens da probabilidade total coberta em determinados intervalos medidos em quantidades de desvios-padrão em relação à média.

- Cerca de 68,26% da probabilidade total está entre  $\mu - \sigma$  e  $\mu + \sigma$ .
- Cerca de 95,44% da probabilidade total está entre  $\mu - 2\sigma$  e  $\mu + 2\sigma$ .
- Cerca de 99,74% da probabilidade total está entre  $\mu - 3\sigma$  e  $\mu + 3\sigma$ .



- Quando o valor da média aumenta, gráfico se desloca para direita. Se o valor da média diminui o gráfico desloca-se para a esquerda.
- Quando o desvio-padrão é grande, a curva é larga, dispersa em relação à média, e o gráfico possui um formato achatado.
- Quando o desvio-padrão é pequeno, a largura da curva é mais estreita, e o gráfico tem um formato mais alto e magro.

Observe, na figura abaixo. As curvas A e B têm mesma média, porém o desvio-padrão de A é maior que o desvio-padrão de B, e isso faz com que a curva A seja mais achatada que a curva B. As curvas B e C têm o mesmo desvio-padrão, então têm o mesmo formato, mas como a média de C é superior à média de B, então a curva C está mais à direita.



Como você pode observar no gráfico, os valores de  $X$  se estendem pelo intervalo  $(-\infty, +\infty)$ , e a probabilidade associada a eles vai diminuindo conforme se afastam da média, de tal forma que a probabilidade é muito pequena (mas nunca igual a zero) quando os valores de  $X$  são extremos. Isso significa que é muito improvável encontrar valores muito distantes da média, tanto para direita quanto para esquerda. Na prática, nem sempre vamos encontrar situações em que o intervalo de valores da variável aleatória seja estritamente infinito em ambos os sentidos. Porém, pela característica explicada anteriormente, em muitas situações podemos considerar a Distribuição Normal uma boa aproximação para a faixa de valores observada.

**Para a Distribuição Normal, temos:**

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

Sabemos que, para variáveis aleatórias contínuas, não podemos calcular a probabilidade associada a um valor específico, apenas probabilidades acumuladas. No caso da Normal, isso envolve cálculos pouco amigáveis. Uma das possibilidades para se lidar com isso é usar ferramentas computacionais, como o Excel.

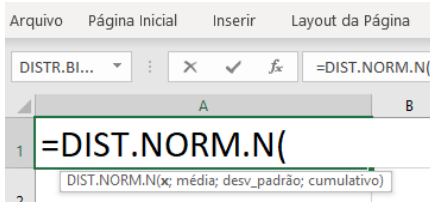
### Exemplo

Uma pesquisa apontou que europeus mantêm seus telefones celulares por 1,5 ano antes de trocar por um novo dispositivo. O desvio padrão é 0,25 ano. Um usuário de telefone celular é selecionado aleatoriamente. Calcule a probabilidade de que um usuário selecionado, ao acaso, mantenha seu telefone atual por menos de 1 ano antes de comprar um novo. Considere que os períodos que as pessoas mantêm seus celulares são normalmente distribuídos e representados pela variável  $X$ .

*Resolução*

*Podemos utilizar a função do Excel, DIST.NORM.N para calcular essa probabilidade*





$x$  = número do qual se quer a probabilidade de ocorrência

**média** = média da distribuição normal considerada

**desv\_padrão** = desvio padrão da distribuição normal considerada

**cumulativo** = para a função probabilidade cumulativa (até algum valor) use VERDADEIRO (ou 1), para função densidade de probabilidade (valor da função no ponto) use FALSO (ou 0).

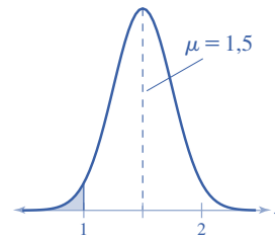
Nesse caso,

$x = 1$

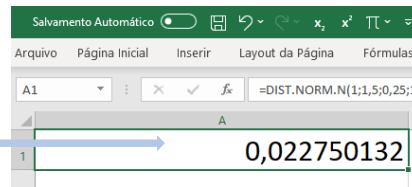
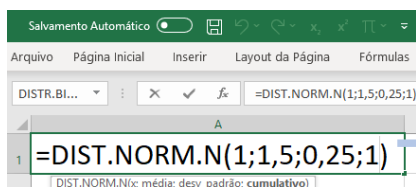
**média** = 1,5

**desv\_padrão** = 0,25

**cumulativo** = 1 (o interesse é a probabilidade para  $x < 1$  anos, então é uma probabilidade acumulada)



Então, na função do Excel, teremos



Portanto, a probabilidade de que um europeu, selecionado aleatoriamente, mantenha seu celular atual por menos de 1 ano, antes de comprar um novo, é de 2,28% (arredondando).

**Dica**

Observe que, no Excel, a probabilidade calculada, quando se quer uma probabilidade acumulada é de **até o valor de X, ou seja,  $P(X \leq x)$** , e isso significa que se o interesse for em probabilidades para valores maiores do que o X que colocamos como argumento, precisaremos diminuir de 1, já que 1 é a probabilidade total de toda distribuição normal. Dessa forma, se você estiver interessado na probabilidade de um europeu manter um celular por **1 ano ou mais**, o cálculo será dado por

$$P(X \geq 1) = 1 - P(X < 1) \longrightarrow P(X \geq 1) = 1 - 0,02275 = 0,97725$$

Probabilidade acumulada encontrada no Excel, de que X seja menor que 1 ano.

Portanto, a probabilidade de que um europeu, selecionado aleatoriamente, mantenha seu celular atual por 1 ano ou mais, antes de comprar um novo, é de 97,73% (arredondando).

Outra forma de tratar distribuições normais é por meio do uso de tabelas com os valores de probabilidades acumuladas da distribuição. O problema é que não se consegue elaborar tabelas com todos os valores para todas as combinações de média e desvio-padrão possíveis. A saída, neste caso, é o uso de uma distribuição normal padrão, e a tabulação das probabilidades acumuladas dessa distribuição, conforme você estudará na próxima seção.

#### Seção 4: **Distribuição Normal Padrão**

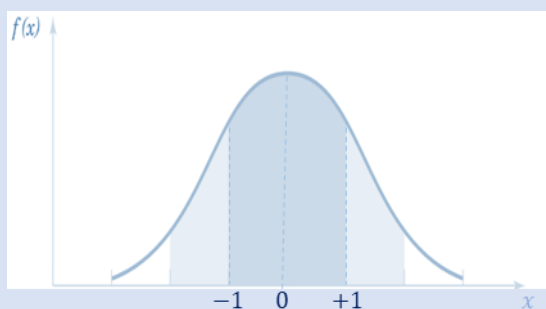
Existem infinitas distribuições normais, cada uma com sua própria média e desvio padrão. A distribuição normal com média 0 e desvio padrão 1 é chamada de **distribuição normal padrão**. A escala horizontal do gráfico da distribuição normal padrão corresponde ao escore-z.

Um escore-z é uma medida de posição que indica o número de desvios padrão em que um valor se encontra a partir da média, e você pode transformar um valor  $x$  em escore-z usando a fórmula:

$$z = \frac{\text{valor} - \text{média}}{\text{desvio padrão}} = \frac{x - \mu}{\sigma}$$

#### **Distribuição normal padrão**

A distribuição normal padrão é uma distribuição normal com média igual a 0 e desvio padrão 1. A área total sob a curva normal é 1



#### **Exemplos**

1) Considere que as notas obtidas na disciplina de Estatística possam ser aproximadas por uma Distribuição Normal, com média  $\mu = 6,5$  e desvio-padrão  $\sigma = 1,5$ . Uma professora pretende recomendar para monitoria os alunos que tirarem uma nota igual ou superior a oito. Qual a probabilidade de um aluno ser recomendado para monitoria?

Resolução

Começamos calculando o z-score para o ponto de interesse ( $x=8$ ):

$$z = \frac{x - \mu}{\sigma} = \frac{8 - 6,5}{1,5} = 1$$

Traduzindo o problema para linguagem estatística, foi solicitado:

$$P(X \geq 8) = P(Z \geq 1)$$

Vale lembrar que a tabela normal padronizada nos dá a probabilidade acumulada entre  $Z = 0$  e o valor  $Z$  procurado. Ou seja, quando você consultar a **Tabela da Distribuição Normal padronizada**,

**Normal padronizada**, 

Como encontrar a probabilidade na tabela normal padrão? Procure os centésimos na coluna  $Z$  e os décimos faltantes na linha superior e cruze linha dos centésimos com a colunas dos décimos – a probabilidade procurada está nesse cruzamento.

Nesse exemplo,  $z = 1,00$  então procure na coluna  $z$  a linha dos centésimos 1,0 e na linha superior o décimo 0; conforme a figura

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015

A probabilidade encontrada é a de a nota ser até 8,5; ou seja,  $P(Z \leq 8,5) = 0,8413$ . Porém o que foi solicitado é  $P(Z \geq 8,5)$ . Então como sabemos que a distribuição normal é simétrica, e que a probabilidade total é 1, temos:

$$P(Z \geq 8,5) = 1 - P(Z \leq 8,5)$$

$$P(Z \geq 8,5) = 1 - 0,8413$$

$$P(Z \geq 8,5) = 0,1587$$

Portanto, a probabilidade de a professora recomendar alunos para monitoria é de 15,87% aproximadamente.

2) Considere agora que os alunos que tirarem menos de cinco serão chamados para fazer aulas de recuperação. Qual a porcentagem da turma que deve ir para recuperação?

De novo, escrevendo em linguagem estatística, temos:

$$z = \frac{x - \mu}{\sigma} = \frac{5 - 6,5}{1,5} = -1$$

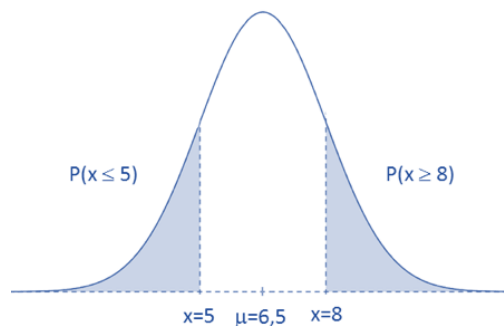
$P(X \leq 5) = P(Z \leq -1)$  consultando a tabela você encontrará 0,1587.

z	0,09	0,08	0,07	0,06	0,05	0,04	0,03	0,02	0,01	0,00
-3,4	0,0002	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003
-3,3	0,0003	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0005	0,0005	0,0005
-3,2	0,0005	0,0005	0,0005	0,0006	0,0006	0,0006	0,0006	0,0006	0,0007	0,0007
-3,1	0,0007	0,0007	0,0008	0,0008	0,0008	0,0008	0,0009	0,0009	0,0009	0,0010
-3,0	0,0010	0,0010	0,0011	0,0011	0,0011	0,0012	0,0012	0,0013	0,0013	0,0013
-2,9	0,0014	0,0014	0,0015	0,0015	0,0016	0,0016	0,0017	0,0018	0,0018	0,0019
-2,8	0,0019	0,0020	0,0021	0,0021	0,0022	0,0023	0,0023	0,0024	0,0025	0,0026
-2,7	0,0026	0,0027	0,0028	0,0029	0,0030	0,0031	0,0032	0,0033	0,0034	0,0035
-2,6	0,0036	0,0037	0,0038	0,0039	0,0040	0,0041	0,0043	0,0044	0,0045	0,0047
-2,5	0,0048	0,0049	0,0051	0,0052	0,0054	0,0055	0,0057	0,0059	0,0060	0,0062
-2,4	0,0064	0,0066	0,0068	0,0069	0,0071	0,0073	0,0075	0,0078	0,0080	0,0082
-2,3	0,0084	0,0087	0,0089	0,0091	0,0094	0,0096	0,0099	0,0102	0,0104	0,0107
-2,2	0,0110	0,0113	0,0116	0,0119	0,0122	0,0125	0,0129	0,0132	0,0136	0,0139
-2,1	0,0143	0,0146	0,0150	0,0154	0,0158	0,0162	0,0166	0,0170	0,0174	0,0179
-2,0	0,0183	0,0188	0,0192	0,0197	0,0202	0,0207	0,0212	0,0217	0,0222	0,0228
-1,9	0,0233	0,0239	0,0244	0,0250	0,0256	0,0262	0,0268	0,0274	0,0281	0,0287
-1,8	0,0294	0,0301	0,0307	0,0314	0,0322	0,0329	0,0336	0,0344	0,0351	0,0359
-1,7	0,0367	0,0375	0,0384	0,0392	0,0401	0,0409	0,0418	0,0427	0,0436	0,0446
-1,6	0,0455	0,0465	0,0475	0,0485	0,0495	0,0505	0,0516	0,0526	0,0537	0,0548
-1,5	0,0559	0,0571	0,0582	0,0594	0,0606	0,0618	0,0630	0,0643	0,0655	0,0668
-1,4	0,0681	0,0694	0,0708	0,0721	0,0735	0,0749	0,0764	0,0778	0,0793	0,0808
-1,3	0,0823	0,0838	0,0853	0,0869	0,0885	0,0901	0,0918	0,0934	0,0951	0,0968
-1,2	0,0985	0,1003	0,1020	0,1038	0,1056	0,1075	0,1093	0,1112	0,1131	0,1151
-1,1	0,1170	0,1190	0,1210	0,1230	0,1251	0,1271	0,1292	0,1314	0,1335	0,1357
-1,0	0,1379	0,1401	0,1423	0,1446	0,1469	0,1492	0,1515	0,1539	0,1562	0,1587

Note que pelos cálculos,  $P(X \leq 5) = P(Z \leq -1) = P(Z \geq 1) = P(X \geq 8) = 0,1587$

Ou seja, espera-se que 15,87% da turma seja encaminhada para recuperação.

A figura abaixo ilustra este raciocínio. A área azul à esquerda de  $X = 5$  (que corresponde a  $Z = -1$ ) é a probabilidade que queremos calcular, e, por simetria, ela é igual à área azul à direita de  $X = 8$  (que corresponde a  $Z = 1$ ).



## Desafio

Pesquise uma situação que possa ser representada por uma distribuição de normal. Você pode identificá-la em seu próprio contexto profissional, ou cotidiano; ou ainda pesquisar em artigos científicos.

Ao realizar essa tarefa, você deverá:

- Postar o material selecionado no AVA, no espaço indicado, identificando a fonte de onde foi extraído, segundo as normas da ABNT.

## Saiba mais

Leia o artigo sob título **Ensinando a distribuição de probabilidade normal utilizando os recursos do Microsoft Excel** e observe as funções do Excel que podem ser utilizadas para resolver problemas que envolvam distribuição normal, inclusive, sem a necessidade de utilizar a tabela de distribuição normal padrão.

<http://www.eumed.net/cursecon/ecolat/br/14/distribuzao-probabilidade-normal.html>

## Dica de Leitura

Leia a Unidade 5 do livro **Estatística aplicada à administração, de Marcelo TAVARES**. Acompanhe os exemplos e problemas resolvidos para contribuir com a consolidação dos conteúdos envolvendo probabilidade, além de estudar outras distribuições de probabilidades.

(<https://educapes.capes.gov.br/bitstream/capes/401408/1/PNAP%20-%20Bacharelado%20-%20Modulo%204%20-%20Estatistica%20Aplicada%20a%20Administracao%20-%203ed%202014%20-%20WEB%20-%20atualizado.pdf>)

## Finalizando a Unidade

Nesta unidade, você estudou conceitos e aplicações importantes sobre variáveis aleatórias e distribuições de probabilidade discretas e contínuas. Conhecendo as ideias associadas às distribuições de probabilidades discretas e contínuas, você terá condições de aplicá-las a diferentes problemas. Nesse sentido, a distribuição normal é a mais importante das distribuições de probabilidades, justamente, por sua ampla gama de aplicações. Na próxima, unidade você estudará importantes ideias da estatística inferencial.

Bons estudos!

## Material de apoio

Estas páginas apontam para conjuntos de videoaulas e outros objetos que contribuem para aprendizagem da probabilidade.

**\*Videoaulas e outros sobre Distribuições de Probabilidade**

<https://sway.office.com/bh89ZhrT6Hq8aQo2?ref=Link>

Código de incorporação

```
<iframe width="760px" height="500px" src="https://sway.office.com/s/bh89ZhrT6Hq8aQo2/embed" frameborder="0" marginheight="0" marginwidth="0" max-width="100%" sandbox="allow-forms allow-modals allow-orientation-lock allow-popups allow-same-origin allow-scripts" scrolling="no" style="border: none; max-width: 100%; max-height: 100vh" allowfullscreen mozallowfullscreen msallowfullscreen webkitallowfullscreen></iframe>
```

\*Essa página está sendo atualizada continuamente pela autora

**Videoaulas da Khan Academy sobre Distribuições de Probabilidade**

<https://sway.office.com/394H4Zsi4ZIqZZGu?ref=Link>

Código de incorporação

```
<iframe width="760px" height="500px" src="https://sway.office.com/s/394H4Zsi4ZIqZZGu/embed" frameborder="0" marginheight="0" marginwidth="0" max-width="100%" sandbox="allow-forms allow-modals allow-orientation-lock allow-popups allow-same-origin allow-scripts" scrolling="no" style="border: none; max-width: 100%; max-height: 100vh" allowfullscreen mozallowfullscreen msallowfullscreen webkitallowfullscreen></iframe>
```

## **Referência Bibliográfica**

### **Básicas**

LARSON, Ron; FARBER, Betsy. Estatística aplicada. 4 ed. São Paulo: Pearson Education do Brasil, 2015.

SARTORIS, Alexandre. Estatística e introdução à econometria. 2. ed. – São Paulo: Saraiva, 2013.

TRIOLA, Mario F. Introdução à estatística. 12. ed. – Rio de Janeiro : LTC, 2017.

### **Complementares**

BUSSAB, W. O.; MORETTIN, P. Estatística Básica. 8. Ed. São Paulo: Atual. 2013.

FREUND, John E. Estatística aplicada: economia, administração e contabilidade. 11. ed. – Porto Alegre : Bookman, 2007.

SHARPE, Norean R.; DE VEAUX, Richard d.; VELLEMAN, Paul F. Estatística aplicada administração, economia e negócios. Porto Alegre: Bookman, 2011.

ROSSI, José W. Econometria e séries temporais com aplicações a dados da economia brasileira / José W. Rossi, Cesar das Neves. - Rio de Janeiro : LTC, 2014.

TAVARES, Marcelo. Estatística aplicada. Universidade aberta do Brasil: Brasília, 2007.

## Unidade 4: Introdução à Estatística Inferencial

### Objetivos

- Aplicar técnicas de amostragem
- Determinar estimadores de uma população.
- Avaliar hipóteses por meio de testes.
- Analisar variâncias

### Introdução

Olá, seja muito bem-vindo(a) à Unidade 4!

Nesta unidade, você estudará estatística inferencial. A Unidade está organizada em quatro seções que abordam, na sequência, os seguintes conteúdos:

1. Amostragem
2. Estimação
3. Teste de hipóteses
4. Análise de variância

O estudo da estatística inferencial permitirá que você realize inferências sobre populações a partir de estatísticas sobre amostras representativas dessas populações. Esse ramo da estatística participa de importantes decisões no mundo contemporâneo, onde não existe espaço para processo decisório que não seja baseado em análise de dados, em inferências.

### Palavras-chave da Unidade

Amostragem, estimadores, intervalo de confiança, teste de hipóteses, análise de variância.

### Seção 1: **Amostragem**

#### ***Técnicas de amostragem***

*O censo realizado a cada 10 anos pelo IBGE, por exemplo, é uma contagem ou medição de toda a população. A realização de um censo gera informações completas, porém é caro e difícil de realizar.*

*Uma **amostragem** é uma contagem ou medição de parte de uma população (uma amostra) e é mais frequentemente aplicada nos estudos estatísticos.*

*Técnicas de amostragem adequadas devem ser implementadas para garantir que as inferências sobre a população (a partir da amostra) sejam válidas, pois quando um estudo for realizado com dados falhos, os resultados serão questionáveis. Mesmo com os melhores*



métodos de amostragem, erros de amostragem podem ocorrer, e esse erro de amostragem se reflete na diferença entre os resultados da amostra e os da população. Nesse sentido, a estatística inferencial possui estratégias para controlar erros de amostragem.

As amostras devem ser representativas da população. Para que as conclusões da teoria de amostragem sejam válidas, as amostras devem ser escolhidas de modo que a amostra possua as mesmas características básicas da população, referentes à variável de interesse.

Existem dois tipos de amostragem, as probabilísticas e as não probabilísticas, as quais serão definidas a seguir.

- **Amostragem probabilística:** todos os elementos da população têm uma probabilidade conhecida e diferente de zero de pertencer à amostra. Por exemplo, 100 colaboradores em capacitação, e você deve selecionar 5 colaboradores. A realização deste tipo de amostragem só é possível se a população estiver acessível e for finita.
- **Amostragem não probabilística:** quando não se conhece a probabilidade de um elemento da população pertencer à amostra. Por exemplo, quando somos obrigados a colher a amostra na parte da população a que temos acesso.

Veja que a utilização de uma amostra probabilística é mais adequada para garantir a representatividade dessa amostra, porque o acaso será o único responsável por eventuais discrepâncias entre população e amostra, e tais discrepâncias impactam no processo de inferência estatística

Os principais esquemas amostrais são apresentados, abaixo:



Quando se decide realizar um estudo a partir de amostra, é preciso elaborar um plano de amostragem. Em um plano amostragem se define quem serão as unidades amostrais, maneira pela qual a amostra será retirada (o tipo de amostragem), e o próprio tamanho da amostra.

### **Amostragens probabilísticas**

- **Aleatória simples** é aquela na qual todos os elementos de uma população têm a mesma chance de serem selecionados. Uma maneira de coletar uma amostra

*aleatória simples é atribuir um número diferente para cada elemento da população e sortear os que vão compor a amostra.*

*Exemplo: Considere o conjunto de todos os candidatos a um cargo, em concurso público. Uma forma de selecionar uma amostra aleatória é atribuir um número a cada candidato e sortear  $n$  deles para compor a amostra.*

- **Amostragem sistemática** é aquela na qual se atribui um número a cada elemento da população ordenada. Essa ordenação é dividida conforme o número de elementos definidos para a amostra, gerando grupos. Um número é selecionado aleatoriamente no primeiro grupo, e os demais elementos da amostra são selecionados em intervalos regulares a partir do número inicial. Uma vantagem da amostragem sistemática é que ela é fácil de ser usada. Contudo, caso ocorra qualquer padrão de regularidade nos dados, esse tipo de amostragem deve ser evitado.

*Exemplo: Para coletar uma amostra sistemática do número de pessoas que moram em um Bairro A, você poderia atribuir um número diferente para cada residência, e escolher aleatoriamente um número no primeiro grupo (por exemplo, residências de 1 a 100, sorteando o número 60) e, a partir dele, selecionar a cada 100 residências (60, 160, 260 e assim por diante) e contar o número celulares em cada residência.*

- **Amostragem estratificada** é adequada para variáveis de interesse que apresentam uma heterogeneidade quando consideradas na população, e estas características heterogêneas permitem a identificação de grupos homogêneos (com alguma característica comum). Nesse caso, você pode dividir a população em subconjuntos que possuam características comuns, os estratos, e realizar uma amostragem dentro de cada estrato, garantindo a representatividade de cada estrato na amostra. Dependendo do foco do estudo, elementos de uma população são divididos em dois ou mais estratos, que compartilham uma característica similar como idade, sexo, grupo étnico ou até mesmo preferência política e uma amostra é então selecionada aleatoriamente de cada um dos estratos. A escolha por uma amostragem estratificada pode assegurar que cada segmento da população estará representado.

*Exemplo: Para coletar uma amostra estratificada do número de pessoas que moram no bairro de Copacabana, na cidade do Rio de Janeiro, você poderia dividir os domicílios em níveis socioeconômicos, pois existem diversos níveis no bairro, já que é composto tanto por classes média e alta como por comunidades, e então, selecionar aleatoriamente residências de cada nível. Ao utilizar uma amostragem*

*estratificada, é necessário que se tome o cuidado de assegurar que todos os estratos forneçam amostras proporcionais às suas reais porcentagens de ocorrência na população. Por exemplo, se 50% das pessoas em Copacabana pertencem ao grupo de renda mais baixa, então a amostra (obtida por amostragem estratificada proporcional) deve ter a mesma proporção de 50% desse grupo.*

- **Amostragem por conglomerado** é adequada para os casos em que a população está distribuída em subgrupos que ocorrem naturalmente, cada um tendo características similares, e seus elementos heterogêneos entre si, representando a população de forma adequada. Para amostrar uma população por meio desse procedimento, divide a população em conglomerados, que são subconjuntos heterogêneos da população, e selecione todos os elementos em um ou mais (mas não em todos) conglomerados sorteados.

*Exemplo: Considere que o interesse seja o número de pessoas que moram nas residências do bairro de Copacabana, na cidade do Rio de Janeiro; para coletar uma amostra por conglomerado, divida as residências em subconjuntos de acordo com as regiões do bairro, selecionando todas as residências em um ou mais, mas não todas as regiões e realize a contagem das pessoas que vivem em cada residência. Para bem utilizar amostragem por conglomerado, deve-se assegurar que todos tenham características semelhantes. Por exemplo, se uma das regiões do bairro tem um percentual maior de pessoas de alta ou de baixa renda, os dados podem não ser representativos da população. No caso de Copacabana, por exemplo, as residências inseridas nas regiões envolvendo comunidades que ficam no bairro, muito provavelmente, vão apresentar rendas mais baixas.*

### **Saiba mais**

Sobre métodos e resultados da [Pesquisa Nacional por Amostra de Domicílios \(PNAD\)](#), consulte o site [www.ibge.com.br](http://www.ibge.com.br).

## **Seção 2: Estimação**

*Quando não se conhece os parâmetros populacionais (média e desvio-padrão, por exemplo) você pode estimar estes parâmetros desconhecidos utilizando dados amostrais; por exemplo, estimar a média de renda de uma população de interesse a partir de uma*

média calculada sobre uma amostra dessa população. Este processo de estimação de parâmetros populacionais é um dos principais resultados da estatística inferencial.

Qualquer que seja a característica que se queira observar em uma população, esta pode ser estimada a partir de dados de uma amostra aleatória, considerando que a amostra seja representativa, dessa população.

A análise de dados amostrais está associada à maioria das decisões seja no âmbito da iniciativa privada, ou do Estado, e em todos os campos da medicina à economia, nesse sentido, a estatística inferencial possui alta relevância, já que estimativas subsidiam decisões que podem impactar a vida da sociedade como um todo.

**Estimativas** são valores descritivos obtidos através dos estudos com as amostras; **Parâmetros** são valores descritivos correspondentes à população.

Dessa forma, o valor da **média amostral** ( $\bar{X}$ ) é uma estimativa pontual da **média populacional** ( $\mu$ ). De modo análogo, o valor do **desvio-padrão amostral** ( $S$ ) constitui uma estimativa do parâmetro **desvio-padrão populacional** ( $\sigma$ ).

*Tipos de estimativas dos parâmetros: a pontual e a intervalar.*

### **Estimativa Pontual**

Uma estimativa é pontual quando temos uma única e melhor estimativa para o parâmetro populacional. Neste caso, com base em dados amostrais, calcula-se um valor da estimativa do parâmetro populacional, obtendo uma estimativa por ponto (ou pontual) do parâmetro considerado.

*Exemplo: Uma amostra aleatória de 200 alunos de uma universidade com 20.000 estudantes apresenta coeficiente de rendimento médio amostral de 5,2. Logo,  $\bar{X} = 5,2$  é uma estimativa pontual do verdadeiro coeficiente de rendimento médio dos 20.000 alunos. A média amostral  $\bar{X}$  é a estimativa mais honesta para o parâmetro  $\mu$  (média da população).*

*Esse tipo de estimador é pontual, pois especifica um único valor para o estimador. Esse procedimento de estimação não permite que você avalie qual o possível tamanho do erro que está cometendo, pois não fornece qualquer informação sobre o quão perto a média da amostra ( $\bar{X}$ ) está da média da população ( $\mu$ ). A estimativa poderia, por exemplo, estar muito perto ou consideravelmente longe. Daí, surge a ideia de construir os intervalos de confiança — estimador intervalar —, que são baseados na distribuição amostral do estimador pontual. Portanto, a estimativa pontual de um parâmetro não apresenta uma medida do possível erro cometido na estimação.*

### Estimativa Intervalar

Uma estimativa é intervalar quando se conhece um intervalo de valores, dentro do qual se acredita que esteja o valor do parâmetro populacional.

Uma estimativa por intervalo para um parâmetro populacional é um intervalo delimitado por dois números, obtidos a partir dos elementos amostrais, onde se espera que esteja o valor do parâmetro populacional, com um dado nível de confiança, ou probabilidade.

Os intervalos de confiança são usados para indicar a confiabilidade de uma estimativa.

#### Exemplos

Problema	Intervalo de confiança
1. Uma amostra de 60 pessoas apresenta média de estaturas igual a 165 cm. Sabendo-se que o desvio-padrão das estaturas é de 10 cm, podemos construir um intervalo de confiança, com nível de confiança de 95%, para a média da população.	Intervalo de confiança para a média populacional, com <b>variância conhecida</b> .
2. Uma amostra das quantidades (em gramas) de um poluente encontrado nas águas da Lagoa da Pampulha apresentou os seguintes valores: (98,4 - 104,6 - 108,3 - 9,8 - 91,7 - 110,5 - 89,0 - 105,2 - 115,9 - 86,4). Não se conhece o desvio-padrão da população. Pode-se construir um intervalo de confiança, com nível de confiança de 90%, para a média da população.	Intervalo de confiança para a média populacional, com <b>variância desconhecida</b> .

Antes de estudar os processos de construção dos intervalos de confiança, é preciso que você entenda o que é **nível de confiança**!

**Nível de confiança** é a probabilidade de que o parâmetro estimado esteja dentro dos limites do intervalo de confiança, ou seja, quando você delimitar um intervalo de confiança, poderá afirmar, com uma probabilidade igual à do nível de confiança, que esse intervalo contém o parâmetro populacional que você quer encontrar.

Exemplo: Se o intervalo de confiança para a média de renda de uma população for igual a [980,00 ; 1.550,00] com nível de confiança de 95%, significa que há uma probabilidade de 95% de que a renda média da população esteja dentro deste intervalo, ou seja, com 95% de certeza a renda média da população está no intervalo de R\$ 980,00 a R\$ 1.550,00.

De forma geral, são aplicados níveis de confiança de: 90%, 95% e 99%, na construção de intervalos de confiança. Os z-escores correspondentes a esses níveis de confiança constam da tabela, abaixo, e você pode utilizá-los diretamente em seus cálculos.

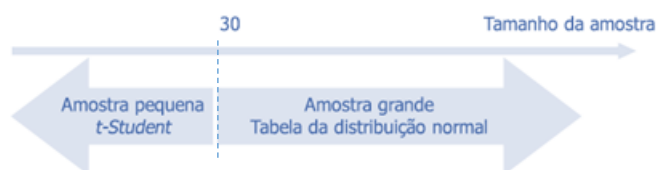
Nível de confiança	$Z_c$
90%	1,645
95%	1,96
99%	2,575

O nível de confiança é uma medida de confiabilidade do procedimento de definição do intervalo de confiança.

**Cuidado!**

Um nível de confiança de 95% **não significa** que 95% dos dados da amostra estejam dentro do intervalo.

Nos exemplos, foram colocados intervalos de confiança para a média populacional, mas no primeiro exemplo, o desvio-padrão populacional é conhecido e a amostra é grande, e no segundo exemplo, desvio-padrão populacional é desconhecido e a amostra é pequena. Portanto, para estimar a média de uma população, utilizando intervalos de confiança, você tem que considerar dois casos: desvio-padrão conhecido e desvio-padrão desconhecido (quando foi estimado de dados amostrais). Por outro lado, também precisa avaliar o tamanho da amostra, ou seja, se trata-se de uma amostra grande ( $n \geq 30$ ), ou uma de amostra pequena ( $n < 30$ ), pois para amostras grandes utiliza-se a Tabela da distribuição Normal e para amostras pequenas a Tabela da distribuição t-Student (desde que o desvio-padrão populacional seja desconhecido, mas que provenha de uma distribuição normal).



**1º. caso**

**Intervalo de confiança da média populacional**

**$\sigma$  conhecido  
amostras grandes.**

Nos casos em que você tem informação acerca do desvio-padrão populacional, o intervalo de confiança encontrado para a média populacional  $\mu$ , com nível de confiança igual a  $(1 - \alpha)$ , é representado por:

$$IC(\mu, 1 - \alpha) = (\bar{X} - Z_{\frac{\alpha}{2}} \cdot \sigma_{\bar{X}})$$

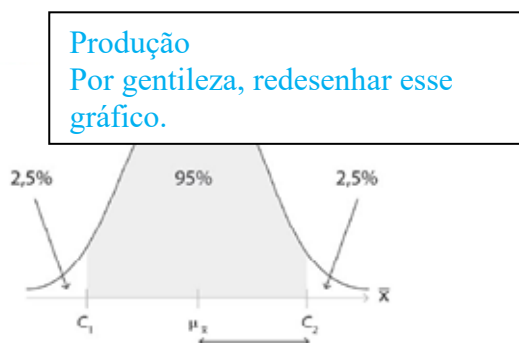
Duas novas medidas surgem na determinação do intervalo de confiança:

- O desvio-padrão da média que é calculado pela divisão do desvio-padrão pela raiz do tamanho da amostra.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- O valor  $Z_{\frac{\alpha}{2}}$ , que determinará a amplitude do intervalo, é o valor tabelado referente à distribuição normal padrão, que você estudou na Unidade 3. Dessa forma, como na construção do intervalo de confiança, vamos considerar valores maiores e menores que a estimativa pontual, devemos fracionar a nossa confiança por dois e buscar seu respectivo valor tabelado.

Por exemplo, na definição de um intervalo de confiança de 95%, queremos que a probabilidade de o valor da média estar compreendido dentro do intervalo de confiança calculado seja de 0,95 ( $1 - \alpha$ ). Restam 0,025 ( $\frac{\alpha}{2}$ ) de cada um dos lados de fora do intervalo, correspondente à probabilidade de que o valor da média não esteja dentro do intervalo (lembrando que, pelo fato da distribuição ser bicaudal, dividimos os 0,05 restantes para as duas caudas). Veja a figura abaixo.



Usando a tabela da distribuição normal padrão, conseguimos identificar que o z-score correspondente à área de 0,025 é de 1,96.

### Exemplo 1

Uma amostra de 61 pessoas apresentou média de estaturas igual a 171 cm. Sabendo-se que o desvio-padrão das estaturas é de 12 cm. Construa um intervalo de confiança para a média da população, com nível de confiança de 95%.

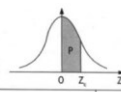
#### Resolução

Primeiro, você precisa buscar o valor tabelado para a confiança.

O método prático consiste em dividir o nível de confiança por dois e em seguida buscar no corpo da tabela normal o seu resultado. O nível de confiança 0,95 dividido por 2 será igual a 0,475. Agora, olhando no centro da tabela Normal esse valor, percebemos que estará na vigésima linha da tabela e em sua sétima coluna. Por fim, para saber qual o valor de, que

utilizaremos, você deverá cruzar o título da linha (1,9) com o título da coluna (0,06) somando ambos para chegar até 1,96.

**Tabela – Distribuição Normal Padrão  $Z \sim N(0, 1)$**   
 Corpo da tabela dá a probabilidade  $p$ , tal que  $p = P(0 < Z < z)$



parte inteira e primeira decimal de $z_1$	Segunda decimal de $z_1$										parte inteira e primeira decimal de $z_1$
	0	1	2	3	4	5	6	7	8	9	
p = 0											
0,0	00000	00399	00798	01197	01595	01994	02392	02790	03188	03586	0,0
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535	0,1
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409	0,2
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173	0,3
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793	0,4
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240	0,5
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490	0,6
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524	0,7
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327	0,8
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891	0,9
1,0	34134	34375	34614	34850	35083	35314	35542	35769	35993	36214	1,0
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298	1,1
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147	1,2
1,3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774	1,3
1,4	41924	42073	42220	42364	42507	42647	42784	42922	43056	43189	1,4
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408	1,5
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449	1,6
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327	1,7
1,8	46407	46485	46562	46638	46712	46784	46855	46926	46995	47062	1,8
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670	1,9
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169	2,0
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574	2,1

A leitura do exemplo fornece os seguintes dados:

- Estimativa Pontual:  $\bar{X} = 1,71$
- Tamanho da amostra ( $n$ ): 61 pessoas.
- Desvio-padrão populacional: 12 centímetros.
- Nível de confiança: 95% ou 0,95.
- $Z_{\frac{\alpha}{2}} = 1,96$
- Calculando o desvio-padrão amostral:  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{61}} = 1,5364$

Substituindo essas informações na definição do intervalo de confiança, temos:

$$IC(\mu, 1 - 5\%) = (171 - 1,96 \cdot 1,5364 ; 171 + 1,96 \cdot 1,5364) = (171 - 3 ; 171 + 3)$$

$$IC(\mu, 95\%) = (168 ; 174)$$

Portanto, você pode dizer, com 95% de confiança, que a média populacional  $\mu$  das estaturas está entre 168 cm e 174 cm. Isso não significa que a probabilidade de o parâmetro  $\mu$  cair nesse intervalo seja de 95%, mas que, se forem extraídas várias amostras independentes e de mesmo tamanho dessa população, espera-se que, em 95% delas, o verdadeiro valor do parâmetro estimado (nesse caso, a média populacional) esteja dentro desse intervalo.

### Fatores que influem na amplitude de um intervalo de confiança

Dada uma estimativa pontual de um parâmetro populacional, pode-se escrever de forma geral a expressão, abaixo, para o intervalo de confiança:



$IC = \text{Estimativa pontual} \pm z \cdot \frac{\sigma}{\sqrt{n}}$	<p><b>Onde</b></p> <p>Estimativa pontual = média amostral.</p> <p><math>z</math> = coeficiente de confiança desejado para um determinado nível de confiança (Distribuição Normal Padrão).</p> <p><math>\sigma</math> = desvio-padrão</p> <p><math>N</math> = número de dados na amostra, isto é, tamanho amostral.</p>
--	--

Observando a fórmula geral para determinação de um intervalo de confiança, podemos ver que os fatores que influenciam na amplitude do intervalo são:

- **Coeficiente de confiança ( $z$ ):** se aumentar o valor de  $z$  (isto é, aumentar o nível de confiança), o intervalo de confiança também aumenta (o  $z$  está no numerador).
- **Tamanho amostral:** se aumentar o tamanho da amostra, o intervalo de confiança diminui ( $n$  está no denominador).
- **Desvio-padrão:** se aumentar o valor de  $\sigma$  (desvio-padrão populacional), o intervalo de confiança também aumenta ( $\sigma$  está no numerador).

### Exemplo 2

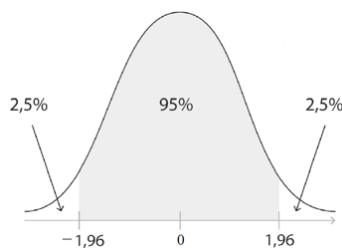
Um engenheiro verificou que a vida útil de um componente eletrônico apresenta um desvio-padrão de 5 horas. Com o objetivo de definir um intervalo de confiança para a média da vida útil desse componente, o engenheiro selecionou uma amostra de 100 unidades do equipamento, obtendo uma média amostral de 500 horas de vida útil. Encontre o intervalo de confiança para a média populacional com um nível de confiança de 95%.

Resolução

Dados no problema:

- Estimativa Pontual (média amostral):  $\bar{X} = 500$
- Tamanho da amostra ( $n$ ):  $n = 100$  equipamentos.
- Desvio-padrão populacional: 5 horas.
- Nível de Confiança: 95% ou 0,95.  $\rightarrow$  sabe-se que  $Z_{\frac{\alpha}{2}} = 1,96$  (da Tabela Normal Padrão)

O gráfico da distribuição Normal padrão será



Calculando o desvio-padrão amostral:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{100}} = \frac{1}{2} = 0,5$$

Substituindo essas informações na definição do intervalo de confiança, teremos:

$$IC(\mu, 95\%) = (500 - 1,96 \cdot 0,5 ; 500 + 1,96 \cdot 0,5) = (500 - 0,98 ; 500 + 0,98)$$

$$IC(\mu, 95\%) = (499,02 ; 500,98)$$

Portanto, o intervalo [499,02 ; 500,98] contém a duração média da vida útil do componente com 95% de confiança, o que significa que, se forem construídos intervalos dessa mesma maneira, para um grande número de amostras, em 95% dos casos os intervalos incluiriam o valor da média populacional  $\mu$ .

Note que o tamanho amostral tem influência na determinação do intervalo de confiança, pois como  $n = 100$  (amostra grande), o intervalo de confiança diminui.

### Exemplo 3

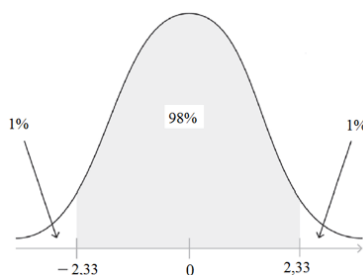
Se para o problema descrito no exemplo anterior, fosse aplicado nível de confiança de 90%, qual seria o intervalo de confiança para a média?

Resolução

Dados no problema

- Estimativa Pontual (média amostral):  $\bar{X} = 500$
- Tamanho da amostra ( $n$ ):  $n = 100$  equipamentos.
- Desvio-padrão populacional: 5 horas.
- Nível de Confiança: 98% ou 0,98.  $\rightarrow$  sabe-se que  $Z_{\frac{\alpha}{2}} = 2,33$  (da Tabela Normal Padrão)

O gráfico da distribuição Normal padrão será



Calculando o desvio-padrão amostral:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{100}} = \frac{1}{2} = 0,5$$

Substituindo essas informações na definição do intervalo de confiança, teremos:

$$IC(\mu, 1 - 2\%) = (500 - 2,33 \cdot 0,5 ; 500 + 2,33 \cdot 0,5) = (500 - 1,165 ; 500 + 1,165)$$

$$IC(\mu, 98\%) = (498,835 ; 501,165)$$

Portanto, o intervalo [498,835 ; 501,165] contém a duração média da vida útil do componente com 95% de confiança, o que significa que, se forem construídos intervalos dessa mesma maneira, para um grande número de amostras, em 98% dos casos os intervalos incluiriam o valor da média populacional  $\mu$ .

2º. caso

**Intervalo de confiança da média populacional**

**$\sigma$  desconhecido  
amostras pequenas**

Quando você não conhece o desvio-padrão populacional, que é a situação real mais comum, deve observar o tamanho da amostra para definir o intervalo de confiança para a média. Neste caso, precisa substituir o desvio-padrão populacional ( $\sigma$ ) pelo desvio-padrão amostral ( $S$ ), o qual é uma boa aproximação do verdadeiro valor.

Pelo **Teorema do Limite Central** (<https://sway.office.com/bh89ZhrT6Hq8aQo2?ref=Link>) sabe-se que, quando o número de elementos da amostra for  $n \geq 30$  (amostra grande), a distribuição das médias é aproximadamente Normal (o valor do coeficiente  $z$  é dado pela Tabela da Distribuição Normal Padrão). Porém, se  $n < 30$  (amostra pequena) devemos utilizar a Distribuição  $t$  (de Student), que é o correto para o desvio-padrão amostral ( $S$ ) e o valor do coeficiente  $t$  é dado pela tabela da distribuição  $t$  – Student.

A forma da distribuição  $t$  – Student é muito parecida com a distribuição normal. A principal diferença entre as duas é que a distribuição  $t$  – Student possui área maior nas caudas.

**Como usar a tabela  $t$  – student?**

Para encontrar os valores de  $t$  na tabela  $t$  – Student, precisamos saber duas coisas: o nível de confiança desejado e o número de graus de liberdade ( $gl = n - 1$ ).

Se sua amostra for composta por 10 elementos e desejar um nível de confiança para a média de 95%, terá:

$n = 10$  (amostra pequena, pois  $n < 30$ ) e 95% de confiança.

Verifique na tabela  $t - Student$ , da seguinte forma:

- Na Linha da tabela:  $n - 1 = 10 - 1 = 9$  graus de liberdade (ou seja, linha 9 da tabela  $t - Student$ )
- Na Coluna da tabela: metade da diferença (100% - 95%), ou seja,  $5\% \div 2 = 2,5\%$

Teremos, então, o valor:  $t = 2,2622$

$t_{1-\frac{\alpha}{2}}$	25%	10%	5%	2,5%	1%	0,5%
Graus de liberdade						
1	1,0000	3,0777	6,3138	12,7062	31,8207	63,6574
2	0,8165	1,8856	2,9200	4,3027	6,9646	9,9248
3	0,7649	1,6377	2,3534	3,1824	4,5407	5,8409
4	0,7407	1,5332	2,1318	2,7764	3,7469	4,6041
5	0,7267	1,4759	2,0150	2,5706	3,3649	4,0322
6	0,7176	1,4398	1,9432	2,4469	3,1427	3,7074
7	0,7111	1,4149	1,8946	2,3646	2,9980	3,4995
8	0,7064	1,3968	1,8595	2,3060	2,8965	3,3554
9	0,7027	1,3830	1,8331	2,2622	2,8214	3,2498
10	0,6998	1,3722	1,8125	2,2281	2,7638	3,1693
11	0,6974	1,3634	1,7959	2,2010	2,7181	3,1058
12	0,6955	1,3562	1,7823	2,1788	2,6810	3,0545
13	0,6938	1,3502	1,7709	2,1604	2,6503	3,0123
14	0,6924	1,3450	1,7613	2,1448	2,6245	2,9768
15	0,6912	1,3406	1,7531	2,1315	2,6025	2,9467
16	0,6901	1,3368	1,7459	2,1199	2,5835	2,9208
17	0,6892	1,3334	1,7396	2,1098	2,5669	2,8982
18	0,6884	1,3304	1,7341	2,1009	2,5524	2,8784
19	0,6876	1,3277	1,7291	2,0930	2,5395	2,8609
20	0,6870	1,3253	1,7247	2,0860	2,5280	2,8453
21	0,6864	1,3232	1,7207	2,0796	2,5177	2,8314
22	0,6858	1,3212	1,7171	2,0739	2,5083	2,8188
23	0,6853	1,3195	1,7139	2,0687	2,4999	2,8073
24	0,6848	1,3178	1,7109	2,0639	2,4922	2,7969
25	0,6844	1,3163	1,7081	2,0595	2,4851	2,7874
26	0,6840	1,3163	1,7056	2,0555	2,4786	2,7787
27	0,6837	1,3137	1,7033	2,0518	2,4727	2,7707
28	0,6834	1,3125	1,7011	2,0484	2,4671	2,7633
29	0,6830	1,3114	1,6991	2,0452	2,4620	2,7564
30	0,6828	1,3104	1,6973	2,0423	2,4573	2,7500

Student é o pseudônimo do químico e matemático inglês William Sealy Gosset (1876-1937), funcionário da cervejaria irlandesa Guinness Brewing Company, em Dublin, no início do século XX, criador da Distribuição  $t$ .

### O intervalo de confiança

Sabendo como encontrar o valor de  $t$  na tabela da distribuição  $t - Student$ , você poderá definir o intervalo de confiança da média populacional, com  $\sigma$  desconhecido e amostras pequenas.

O intervalo de confiança encontrado para a média populacional  $\mu$ , com nível de confiança igual a  $(1 - \alpha)$  é representado por:

$$IC(\mu, 1 - \alpha) = \left( \bar{X} - t_{(n-1);(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}} ; \bar{X} + t_{(n-1);(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}} \right)$$

Onde:

$S$  = desvio-padrão amostral.

$t_{(n-1);(1-\frac{\alpha}{2})}$  = valor tabelado da distribuição  $t - Student$ .

**Exemplo 1**

Uma amostra das quantidades (em gramas) de um poluente encontrado nas águas da Lagoa da Pampulha apresentou os seguintes valores: (98,4 – 104,6 – 108,3 – 9,8 – 91,7 – 110,5 – 89,0 – 105,2 – 115,9 – 86,4). Não se conhece o desvio-padrão da população. Construa um intervalo de confiança, com nível de confiança de 90%, para a média da população.

**Resolução**

Dessa amostra, você deve calcular, por fórmula ou em um aplicativo:

- Média amostral:  $\bar{X} = 91,98$
- Desvio-padrão amostral:  $S = 30,46$
- Tamanho da amostra:  $n = 10$  medidas. Logo, teremos  $n - 1 = 9$  graus de liberdade
- Nível de confiança: 90%
- $t_{9;5\%} = 1,8331$

$t_{1-\frac{\alpha}{2}}$	25%	10%	5%	2,5%	1%	0,5%
<b>Graus de liberdade</b>						
1	1,0000	3,0777	6,3138	12,7062	31,8207	63,6574
2	0,8165	1,8856	2,9200	4,3027	6,9646	9,9248
3	0,7649	1,6377	2,3534	3,1824	4,5407	5,8409
4	0,7407	1,5332	2,1318	2,7764	3,7469	4,6041
5	0,7267	1,4759	2,0150	2,5706	3,3649	4,0322
6	0,7176	1,4398	1,9432	2,4469	3,1427	3,7074
7	0,7111	1,4149	1,8946	2,3646	2,9980	3,4995
8	0,7064	1,3968	1,8595	2,3060	2,8965	3,3554
9	0,7027	1,3830	1,8331	2,2622	2,8214	3,2498

Substituindo essas informações na definição do intervalo de confiança, teremos:

$$IC(\mu, 1 - \alpha) = \left( 91,98 - t_{9;5\%} \cdot \frac{30,46}{\sqrt{10}} ; 91,98 + t_{9;5\%} \cdot \frac{30,46}{\sqrt{10}} \right)$$

$$IC(\mu, 1 - \alpha) = \left( 91,98 - 1,8331 \cdot \frac{30,46}{\sqrt{10}} ; 91,98 + 1,8331 \cdot \frac{30,46}{\sqrt{10}} \right)$$

$$IC(\mu, 1 - \alpha) = \left( 91,98 - 1,8331 \cdot \frac{30,46}{\sqrt{10}} ; 91,98 + 1,8331 \cdot \frac{30,46}{\sqrt{10}} \right)$$

$$IC(\mu, 1 - \alpha) = (91,98 - 17,6569 ; 91,98 + 17,6569)$$

$$IC(\mu, 1 - \alpha) = (74,32 ; 109,64)$$

Então, podemos dizer que, com 90% de confiança, a média populacional  $\mu$  está entre 74,32 gramas e 109,64 gramas.

Isso não significa que a probabilidade de o parâmetro  $\mu$  cair nesse intervalo seja de 90%, mas que, se extrairmos diversas amostras independentes de mesmo tamanho dessa população, espera-se que, em 90% delas, o verdadeiro valor do parâmetro estimado (nesse caso, a média populacional) esteja dentro desse intervalo.

### Exemplo 2

Um gestor de produção selecionou uma amostra aleatória de 10 peças do total produzido em um dia. Seus testes apresentaram vida útil média de 1100 horas, com desvio-padrão de 120 horas. Determine a verdadeira vida útil média dessas peças para um intervalo de confiança de 98%.

#### Resolução

Dessa amostra, você deve calcular, por fórmula ou em um aplicativo:

- Média amostral:  $\bar{X} = 1100$  horas
- Desvio-padrão amostral:  $S = 120$  horas
- Tamanho da amostra:  $n = 10$  peças. Logo, teremos  $n - 1 = 9$  graus de liberdade
- Nível de confiança: 98%
- $t_{9;2\%} = 2,8214$
- Linha da tabela:  $n - 1 = 9$  graus de liberdade (linha 9 da tabela  $t$  – Student)
- Coluna da tabela: metade da diferença (100% - 98%), ou seja,  $2\% \div 2 = 1\%$

$t_{1-\frac{\alpha}{2}}$	25%	10%	5%	2,5%	1%	0,5%
Graus de liberdade						
1	1,0000	3,0777	6,3138	12,7062	31,8207	63,6574
2	0,8165	1,8856	2,9200	4,3027	6,9646	9,9248
3	0,7649	1,6377	2,3534	3,1824	4,5407	5,8409
4	0,7407	1,5332	2,1318	2,7764	3,7469	4,6041
5	0,7267	1,4759	2,0150	2,5706	3,3649	4,0322
6	0,7176	1,4398	1,9432	2,4469	3,1427	3,7074
7	0,7111	1,4149	1,8946	2,3646	2,9980	3,4995
8	0,7064	1,3968	1,8595	2,3060	2,8965	3,3554
9	0,7027	1,3830	1,8331	2,2622	2,8214	3,2498
10	0,6998	1,3722	1,8125	2,2281	2,7638	3,1693

$$IC(\mu, 1 - \alpha) = \left( 1100 - t_{9;1\%} \cdot \frac{120}{\sqrt{10}} ; 1100 + t_{9;1\%} \cdot \frac{120}{\sqrt{10}} \right)$$

$$IC(\mu, 1 - \alpha) = \left( 1100 - 2,8214 \cdot \frac{120}{\sqrt{10}} ; 1100 + 2,8214 \cdot \frac{120}{\sqrt{10}} \right)$$

$$IC(\mu, 1 - \alpha) = (1100 - 107,06 ; 1100 + 107,06)$$

$$IC(\mu, 1 - \alpha) = (992,94 ; 1.207,06)$$

*Portanto, podemos dizer que, com 98% de confiança, a média populacional  $\mu$  (vida útil média das peças produzidas) está entre os valores mínimo de 992,94 horas e máximo de 1.207,06 horas.*

### Seção 3: Testes de hipóteses

Nesta seção, você estudará uma técnica da estatística inferencial, o teste de hipótese. O **teste de hipóteses** é um processo que utiliza estatísticas calculadas a partir de amostras para testar uma afirmação sobre o valor de um parâmetro populacional. Pesquisadores em campos como biologia, psicologia e negócios, dentre outros, contam com os testes de hipóteses para subsidiar processos de tomada de decisões sobre novos medicamentos, tratamentos e estratégias de mercado, investimentos.

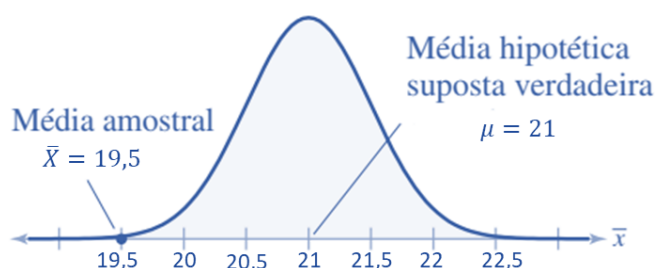
Por exemplo, considere que uma montadora de automóveis informa que seu lançamento híbrido tem média de consumo de combustível de 21 quilômetros por litro. Se você suspeitar que o consumo médio não é de 21 quilômetros por litro, como você poderia mostrar que o anúncio é falso? Obviamente, você não pode testar todos os veículos, mas você ainda pode tomar uma decisão razoável sobre o consumo médio retirando uma amostra aleatória da população de veículos, e medindo o consumo de cada um. Se a média da amostra diferir o suficiente da média que a montadora informa, você pode decidir que a informação está errada.

Por exemplo, para testar que o consumo médio de combustível de todos os veículos híbridos desse tipo é  $\mu = 21$  quilômetros por litro, você retira uma amostra aleatória de  $n = 30$  veículos e mede o consumo de cada um. Você obtém uma média amostral de  $\bar{X} = 19,5$  quilômetros por litro com desvio-padrão amostral  $S = 2,75$  quilômetros por litro. Isso indica que a anúncio do fabricante é falso?

Para decidir,  *você supõe que o anúncio está correto!* Ou seja,  *você supõe que  $\mu = 21$ .* Então, avalia a distribuição amostral das médias (com  $n = 30$ ) obtida de uma população na qual  $\mu = 21$  e  $\sigma = 2,75$ .

Pelo teorema do limite central,  *você sabe que essa distribuição amostral é uma distribuição normal, com média 21 e um erro padrão de  $\frac{2,75}{\sqrt{30}} \approx 0,5$ .*

Na figura, abaixo, observe que sua média amostral ( $\bar{x}$ ) igual a 19,5  *quilômetros por litro* é altamente improvável, pois ela está a, aproximadamente, 3 erros padrão da média anunciada pela montadora! Pensando no que  *você aprendeu sobre distribuições de probabilidade, você pode determinar que, se o anúncio é verdadeiro, então a probabilidade de se obter uma média amostral de 19,5 ou menos é de aproximadamente 0,0013 (1-0,997), pois na distribuição normal cerca de 99,7% da probabilidade total está entre  $\mu - 3\sigma$  e  $\mu + 3\sigma$ .* Este é um evento incomum! Sua suposição de que o anúncio da empresa estava correto, levou a um resultado improvável. Então, ou  *você teve uma amostra muito incomum ou o anúncio é provavelmente falso.* A conclusão é a de que a informação da montadora sobre a média de consumo de seu veículo híbrido, provavelmente, é falsa.



Durante processos decisórios, é conveniente a formular hipóteses sobre as populações de interesse. Essas hipóteses, verdadeiras ou não, são chamadas de  *hipóteses estatísticas.* Em alguns casos, uma hipótese estatística é formulada para ser rejeitada. Por exemplo, desejando decidir se em uma parte de uma praia são encontrados maior número de poluentes do que em outra,  *você formula a hipótese de que não há diferença entre elas, ou seja, que a diferença é devida somente à flutuação das amostras retiradas da mesma população.* Essa hipótese é chamada de  *hipótese nula,* sendo representada por  $H_0$ . Qualquer hipótese diferente de  $H_0$  é chamada de  *hipótese alternativa.*

A inclinação é pela rejeição da hipótese nula ( $H_0 =$  não existe diferença entre as amostras), no caso dessa diferença ser muito grande para ser devida ao acaso, unicamente. E então, conclui-se, que a diferença é  *significativa.* A decisão é baseada na probabilidade de a hipótese nula ( $H_0$ ) estar errada, ou seja, em sua rejeição, e afirmar que duas amostras são



significativamente diferentes. Nesse sentido, os chamados *testes de hipóteses* se constituem em estratégias para subsidiar processos de tomada de decisão.

Em função do caráter aleatório da amostragem, o processo de tomada de decisão é sempre acompanhado de alguma probabilidade de erro.

Dois tipos de erro podem ocorrer:

- 1) Não aceitar uma hipótese quando ela está correta.
- 2) Aceitar (não rejeitar) uma hipótese quando ela é incorreta.

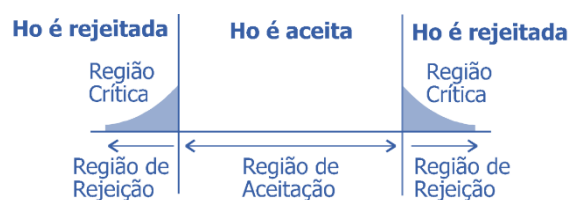
**Nível de significância** é a maior probabilidade, ou risco que se está disposto a correr na rejeição de uma hipótese quando ela é verdadeira. O nível de significância, geralmente, é representado por  $\alpha$ .

No caso concreto, o usual é definir um nível de significância de 5% ou 1%, ainda que possam ser utilizados outros valores. Assim, se você afirmar que uma população A é diferente significativamente de uma população B, ao nível de significância de  $\alpha = 0,01$ , você terá 1% de chance de errar, ou seja, terá uma confiança de 99% de que sua decisão esteja certa. No processo de tomada de decisão, o pesquisador busca sempre diminuir o erro, ao máximo. Uma forma de fazer isso é aumentar a quantidade  $n$  de amostras, mas isso não é sempre possível.

Os passos para você decidir se uma hipótese é verdadeira ou falsa, ou seja, se ela deve ser aceita ou rejeitada, considerando uma determinada amostra, são:

- 1) Determinar a hipótese nula ( $H_0$ ) e a hipótese alternativa ( $H_1$  para tentar rejeitar  $H_0$ ).
- 2) Escolher o nível de significância ( $\alpha$ ) desejado.
- 3) Adequar uma distribuição amostral ao problema de interesse.
- 4) Delimitar as regiões de rejeição e de aceitação (valores tabelados).
- 5) Calcular a estatística da distribuição escolhida, a partir dos valores amostrais obtidos e tomar a decisão.

Para decidir, você deve seguir o seguinte raciocínio: se o valor da estatística calculada (da distribuição adequada) se localizar na região de rejeição, rejeite a hipótese nula, senão a decisão deve ser a de que, ao nível de significância determinada, a hipótese nula não pode ser rejeitada.



### Teste de hipótese para média populacional

Ao selecionar uma amostra de uma população e calcular a média dessa amostra, você pode verificar se uma afirmação sobre a média dessa população está correta. Para isso, você deve verificar se a estatística calculada do teste se encontra na região de rejeição, ou de aceitação da hipótese nula,  $H_0$ .

Nesse contexto, podem ocorrer duas situações distintas:

1) **Desvio-padrão da população é conhecido ou a amostra é considerada grande ( $n \geq 30$ )**, neste caso, a distribuição amostral adequada é a Normal, sendo a estatística-teste dada por:

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Onde

$\bar{X}$  = média amostral

$\mu$ : média populacional

$\sigma$ : desvio-padrão populacional

$n$ : tamanho da amostra.

2) **Desvio-padrão populacional desconhecido e amostra pequena,  $n < 30$** , a distribuição amostral deve ser a *t de Student*, sendo a estatística-teste dada por:

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Onde

$\bar{X}$  = média amostral

$\mu$ : média populacional

$S$ : desvio-padrão amostral

$n$ : tamanho da amostra.

Para amostras grandes,  $n \geq 30$ , a distribuição de  $Z$  (normal padrão) e a distribuição  $t$  de Student, e valores das estatísticas correspondentes apresentam comportamentos muito próximos.

## Exemplos

1) Uma faculdade registrou, nos últimos anos, os testes de QI de seus colaboradores, com média de 115, e desvio-padrão de 20. Com o objetivo de verificar se uma nova equipe de colaboradores é típica dessa faculdade, foi retirada uma amostra aleatória de 50 colaboradores que compõe a nova equipe, cuja média de QI foi de 118. Com uma significância de 5%, teste a hipótese de que esta nova equipe apresente a mesma característica dos colaboradores da faculdade, com relação ao QI.

*Resolução*

*Dados*

$$\mu = 115$$

$$\sigma = 20 \text{ desvio-padrão conhecido}$$

$$n = 50 \text{ amostra grande}$$

$$\bar{X} = 118$$

$$\alpha = 0,05$$

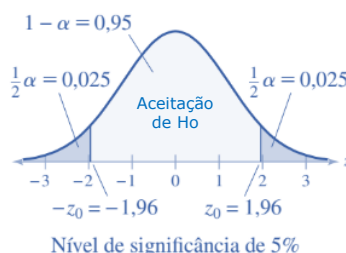


Estadística Z, distribuição Normal

$$H_0: \mu = 115$$

$$H_1: \mu \neq 115$$

$$Z = \frac{118 - 115}{20/\sqrt{50}} = 1,06$$



Como a distribuição normal padrão tem média 0 e desvio-padrão 1,  $Z = 1,06$  está na região de aceitação de  $H_0$ , ou seja,  $\mu = 115$ .

2) O tempo médio consumido por economistas para realizarem certas avaliações de investimentos tem sido de 50 minutos. Um novo procedimento para realização de avaliações de investimentos está sendo implementado, utilizando inteligência artificial. Aplicando esse novo procedimento, foi selecionada uma amostra de 12 pessoas, com um tempo médio de 42 minutos e um desvio-padrão de 11,9 minutos, de realização do mesmo tipo de avaliação de investimentos. Com significância de 5%, teste a hipótese de que a média populacional no novo procedimento de avaliação de investimentos seja menor do que 50.

*Resolução*

**Dados**

$\mu = 50$  minutos

$n = 12$  (amostra pequena)

$S = 11,9$  minutos

$\bar{X} = 42$

$\alpha = 0,05$



Estadística  $t$ , distribuição  $t$  – *student*

$H_0: \mu = 50$

$H_1: \mu < 50$

$$t = \frac{42 - 50}{11,9/\sqrt{12}} = -2,53 \quad , \quad t_{\alpha} = t_{0,05} = -1,796$$

Como a estatística  $t = -2,53$ , que foi encontrada está na região de rejeição de  $H_0$ , a hipótese de que  $\mu = 50$  deve ser rejeitada.



**Comparação entre duas médias**

Para verificar se as duas médias são significativamente diferentes, você pode comparar as médias de duas populações  $\mu_1$  e  $\mu_2$ , obtidas a partir de duas amostras com número de elementos  $n_1$  e  $n_2$ , e variâncias  $S_1^2$  e  $S_2^2$ .

A estratégia consiste em calcular a diferença  $d$  entre as médias,  $d = \mu_1 - \mu_2$ , e pensar: *Se as duas amostras tiverem sido selecionadas da mesma população, existirá pouca diferença entre elas, e assim, o valor de  $d$  será pequeno.*

E nesse sentido, será que o valor absoluto de  $d$  seria tão grande que você não poderia concluir que as amostras foram selecionadas da mesma população, ou de duas populações iguais?

Para responder a essa pergunta, você deverá aplicar um teste ao valor de  $d$ , e para este caso, o teste adequado é o chamado teste *t de Student*.

Considere que você repetirá a amostragem (seleção das amostras), nas duas populações, várias vezes e que calculará as diferenças entre essas médias. Você obterá uma série de valores de  $d$  *normalmente distribuídos*, com média chamada de  $d_m$  e desvio-padrão  $S_d$ , chamado de erro padrão das diferenças.

Assim, como os valores de  $d$  são normalmente distribuídos, e considerando a hipótese nula ( $d = 0$ ), se as duas populações fossem iguais, 95% dos valores de  $d$  estariam dentro do intervalo:  $dm = \pm 2 \cdot S_d$

Portanto, existe probabilidade de 0,95 de que um valor de  $d$ , selecionado aleatoriamente, esteja no intervalo  $\pm 2 \cdot S_d$  e uma probabilidade 0,05, que esteja localizado além dos limites desse intervalo, ou seja, fora dele. Você pode afirmar que 95% dos valores de  $d$  devem ser menores ou iguais ao limite  $2 \cdot S_d$  do intervalo. Dessa forma,

$$d \leq 2 \cdot S_d \rightarrow \frac{d}{S_d} \leq 2$$

Veja que, se  $d$  assumir um valor muito alto, o valor  $\frac{d}{S_d}$  pode ser maior do que 2, e assim, ficaria fora do intervalo de 95%, o que levaria à rejeição da hipótese nula, de igualdade das duas médias. Nesse caso, você poderia afirmar, com probabilidade de errar de 5%, que as duas médias são, significativamente, diferentes.

O modo de calcular o erro padrão das diferenças é dado pela fórmula:

$$S_d = \sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}$$

Onde,

$S_1^2$  e  $S_2^2$  : variâncias das populações

$n_1$  e  $n_2$ : número de dados das amostras.

### Exemplo

1) Considere duas amostras, A e B, cujos parâmetros constam no quadro abaixo.

	amostra A	amostra B
número de dados (n)	200	100
média (m)	2,5	2,8
variância (v)	0,8	0,6

Verifique se a diferença observada entre as duas amostras permite dizer que elas foram coletadas em duas populações distintas, isto é, que as duas médias são significativamente diferentes, ao nível de probabilidade de errar de 0,05.

*Resolução*

$$d = 2,8 - 2,5 = 0,3$$

$$S_d = \sqrt{\left(\frac{0,8}{200} + \frac{0,6}{100}\right)} = 0,1$$

Assim temos que  $\frac{d}{S_d} = \frac{0,3}{0,1} = 3$  que é maior que  $t = 2$ .

Com esse resultado, você pode rejeitar a hipótese nula e afirmar, com probabilidade de erro menor do que 0,05; que as duas médias são significativamente diferentes, ou seja, que a diferença entre as médias é altamente significativa.

#### Seção 4: **Análise de variância**

A análise de variância é um teste de hipótese adequado para comparar mais de duas populações. Por exemplo, considere que você precise comparar nível de endividamento de empresas de três áreas, chamados de tratamentos (metalurgia, moveleira e calçadista). Para realizar a comparação é preciso selecionar uma amostra de empresas de cada área (repetições). No processo de análise de variância, divide-se a variação total de um conjunto de tratamentos a serem comparados por suas correspondentes repetições.

No exemplo, as áreas (metalurgia, moveleira e calçadista) correspondem aos tratamentos.

Existem dois componentes de variação: variação *entre* e variação *dentro*.

- Variação ENTRE é a variação existente entre as médias dos tratamentos, com relação a uma média geral. Essa variação mede a diferença existente entre os tratamentos.
- Variação DENTRO do tratamento é a variação existente entre as repetições de cada tratamento.

As avaliações das repetições *dentro* de cada tratamento correspondem à variação aleatória.

Variação total = variação *entre* + variação *dentro*

ou

Variação total = variação não aleatória + variação aleatória

A variação *entre* tratamentos é atribuída especificamente à variação das médias dos tratamentos com relação à média geral.

A variação *dentro* de cada tratamento corresponde à variação de cada observação com relação à média do tratamento, originada por todas as fontes causadoras de variações nos experimentos (aleatórias), excetuando os tratamentos.

A variação *total* corresponde à variação de cada observação com relação à média geral.

Portanto, as expressões para cálculo são:

$$\text{Variação total} = \text{SQTotal} = \sum_{i=1}^t \sum_{j=1}^r (Y_{ij} - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{rt}$$

Onde,

$\bar{Y}$ : média geral de todos os tratamentos

$r$ : número de repetições

$t$ : número de tratamentos

$$\text{Variação entre} = \text{SQTratamento} = \sum_{i=1}^t \frac{T_i^2}{r_j} - C$$

Onde,

$T_i$ : total do tratamento  $i$

$C$ : fator de correção =  $\frac{(\sum Y_{ij})^2}{tr}$

Como a variação *total* é igual à variação *entre* somada à variação *dentro*; o cálculo da variação *dentro* (aleatória) é dada pela fórmula:  **$SQDENTRO = SQTOTAL - SQENTRE$**

O propósito da análise de variância são os de encontrar estimativas mais precisas para as médias dos tratamentos, diferenças entre médias e testar hipóteses acerca da igualdade de médias dos tratamentos.

As hipóteses na análise de variância são:

$H_0$ :  $t_1 = t_2$  (não há diferença entre as médias dos tratamentos)

$H_1$ : *pelo menos, um dos tratamentos, é diferente dos demais*

Por meio da análise de variância, são obtidos os quadrados médios (*QM*), que são estimativas não viesadas das variâncias contempladas na análise. Essa é a origem do nome

análise de variância. Esses quadrados médios são encontrados dividindo-se a soma de quadrado pelo correspondente grau de liberdade. Dessa forma,

$$QM_{Trat} = \frac{SQ_{Trat}}{G.L.Trat} \qquad QM_{Resíduo} = \frac{SQ_{Resíduo}}{G.L.Resíduo}$$

Para encontrar os graus de liberdade proceda conforme o quadro, abaixo.

Fonte de variação	Graus de liberdade G. L.
Tratamento (entre)	t - 1
Resíduo (dentro)	t (r - t)
Total	tr - 1

Em seguida, você precisará testar se a variância do fator (*entre*) é diferente da variância aleatória (*dentro*). A distribuição amostral que compara duas variâncias, é a **distribuição F**. Nesse sentido, você pode utilizar o teste de F para verificar a validade da hipótese  $H_0$  descrita anteriormente. O teste é apresentado a seguir:

$$F_{calc} = \frac{QM_{Trat}}{QM_{Resíduo}}$$



Distribuição F\_LARSON.pdf

Material sobre a distribuição F

Compare o valor calculado ( $F_{calc}$ ) na análise de variância, com o valor da tabela  $F_{\alpha}(v_1, v_2)$ , sendo  $v_1$  e  $v_2$  os graus de liberdade de tratamentos e resíduos, respectivamente.

Se  $F_{calc} > F_{tab}$ , o experimento foi significativo, o que indica que há uma probabilidade maior do que  $1 - \alpha$  de que pelo menos um dos tratamentos se diferencie dos demais.

O cenário da análise de variância pode ser sintetizado, conforme quadro, abaixo:

Fonte de variação	G. L.	S. Q	Q.M.	F <sub>calc</sub>	$F_{\alpha}(v_1, v_2)$
Tratamento	t - 1	SQ <sub>Trat</sub>	QM <sub>Trat</sub>	F calculado	F tabelado
Resíduo	t (r - 1)	SQ <sub>Resíduo</sub>	QM <sub>Resíduo</sub>		
Total	tr - 1	SQ <sub>Total</sub>			

**Exemplo**

O quadro abaixo, registra os valores assumidos por determinado índice inflacionário em três Estados para um período de cinco meses. Verifique, por meio de uma análise de



variância, se as médias são estatisticamente iguais ou não, com nível de significância de 5%.

Meses	Estados		
	E1	E2	E3
1	1,60	1,20	2,00
2	2,00	1,10	1,80
3	2,20	1,20	1,40
4	1,70	1,30	1,60
5	1,80	1,00	1,90
Total	9,30	5,80	8,70

### Resolução

As hipóteses desta análise de variância são:

$H_0: E1 = E2 = E3$  (as médias dos estados não diferem)

$H_1: ao menos um dos estados difere dos demais, em média.$

As repetições, ou seja, os meses, são independentes, porque são somente repetições.

### Importante

O teste  $F$  para análise de variância é um teste unilateral à direita, sempre, por causa do tipo de hipótese alternativa.

Cálculos das somas de quadrados:

Dados

$r$ : número de repetições (5 meses)

$T_i$ : total de cada tratamento

$C$ : fator de correção =  $\frac{(\sum \sum Y_{ij})^2}{tr}$

$$SQ_{Tratamento} = SQ_{ENTRE} = \sum_{i=1}^t \frac{T_i^2}{r_j} - C$$

$$SQ_{ENTRE} = \frac{9,30^2 + 5,80^2 + 8,70^2}{5} - \frac{(9,30 + 5,80 + 8,70)^2}{3 \cdot 5} = 39,164 + 37,7627 = 1,4013$$

$$SQ_{Total} = \sum Y^2 - \frac{(\sum Y)^2}{rt}$$

$$SQ_{Total} = (1,6^2 + 2,0^2 + 2,20^2 + \dots + 1,90^2) - \frac{(9,30 + 5,80 + 8,70)^2}{3 \cdot 5} = 1,9173$$

$$SQ_{DENTRO} = SQ_{TOTAL} - SQ_{ENTRE} = 1,9173 - 1,4013 = 0,516$$

$$F_{calc} = \frac{QM_{Trat}}{QM_{Resíduo}} = \frac{0,7007}{0,0430} = 16,2953$$

Em síntese:

Fonte de variação	G. L.	S. Q	Q.M.	F <sub>calc</sub>	Significância
Tratamento	3 - 1 = 2	1,4013	0,7007	16,2942	0,0006
Resíduo	3(5-1) = 12	0,5160	0,0430		
Total	3 . 5 - 1 = 14	1,9173			

*Resposta*

*Há diferença significativa entre os estados, porque o F tabelado (tabela de 5% e  $v_1 = 2$  graus de liberdade e  $v_2 = 12$  graus de liberdade) é menor do que o valor F calculado (16,29), e dessa forma, o F calculado fica localizado na região de rejeição da hipótese  $H_0$ .*

## Desafio

Realize uma pesquisa, com uma variável quantitativa do seu interesse. Registre os dados em uma tabela e:

- a) Calcule as medidas de síntese (de tendência central e de dispersão), utilizadas em testes de hipóteses;
- b) Aplique um dos testes de hipóteses que você estudou nesta unidade – observando qual o objetivo do teste, que deve estar de acordo com o seu objetivo de pesquisa.
- c) Registre sua avaliação a partir dos resultados do teste escolhido.

Ao realizar essa tarefa, você deverá:

Postar o material selecionado no AVA, no espaço indicado, identificando a fonte de onde foi extraído, segundo as normas da ABNT.

## Saiba mais

Sobre as distribuições amostrais, utilizadas nos testes de hipóteses, leia as páginas de número 140 a 149, do **livro Estatística aplicada à administração**. Essa leitura pode aprofundar seu entendimento sobre estas distribuições e facilitar a realização de testes de hipóteses. (<https://educapes.capes.gov.br/bitstream/capes/401408/1/PNAP%20-%20Bacharelado%20-%20Modulo%204%20-%20Estatistica%20Aplicada%20a%20Administracao%20-%203ed%202014%20-%20WEB%20-%20atualizado.pdf>)

## Dica de Leitura

Leia a Unidade 6, do livro Estatística aplicada à administração, de Marcelo Tavares. Acompanhe os exemplos e problemas resolvidos para contribuir com a consolidação dos conteúdos envolvendo probabilidade, além de estudar outras distribuições de probabilidades. (<https://educapes.capes.gov.br/bitstream/capes/401408/1/PNAP%20-%20Bacharelado%20-%20Modulo%204%20-%20Estatistica%20Aplicada%20a%20Administracao%20-%203ed%202014%20-%20WEB%20-%20atualizado.pdf>)

## Finalizando a Unidade

Nesta unidade, você estudou como aplicar técnicas de amostragem, determinar estimadores de uma população, avaliar hipóteses por meio de testes e analisar variâncias. O conteúdo que compõe a Estatística inferencial não se esgota nesta unidade, mas a partir dos conhecimentos adquiridos aqui você terá condições de realizar inferências e estudar mais situações e técnicas de aplicação, deste ramo da Estatística.

Muito bons estudos!

### **Material de apoio**

Página com vídeos e problemas resolvidos sobre estatística inferencial  
<https://sway.office.com/52RJA4zfQ1GdfP46?ref=Link>

### **Referência Bibliográfica**

#### **Básicas**

LARSON, Ron; FARBER, Betsy. Estatística aplicada. 4 ed. São Paulo: Pearson Education do Brasil, 2015.

SARTORIS, Alexandre. Estatística e introdução à econometria. 2. ed. – São Paulo: Saraiva, 2013.

TRIOLA, Mario F. Introdução à estatística. 12. ed. – Rio de Janeiro : LTC, 2017.

#### **Complementares**

BUSSAB, W. O.; MORETTIN, P. Estatística Básica. 8. Ed. São Paulo: Atual. 2013.

FREUND, John E. Estatística aplicada: economia, administração e contabilidade. 11. ed. – Porto Alegre : Bookman, 2007.

SHARPE, Norean R.; DE VEAUX, Richard d.; VELLEMAN, Paul F. Estatística aplicada administração, economia e negócios. Porto Alegre: Bookman, 2011.

ROSSI, José W. Econometria e séries temporais com aplicações a dados da economia brasileira / José W. Rossi, Cesar das Neves. - Rio de Janeiro : LTC, 2014.

TAVARES, Marcelo. Estatística aplicada. Universidade aberta do Brasil: Brasília, 2007.

