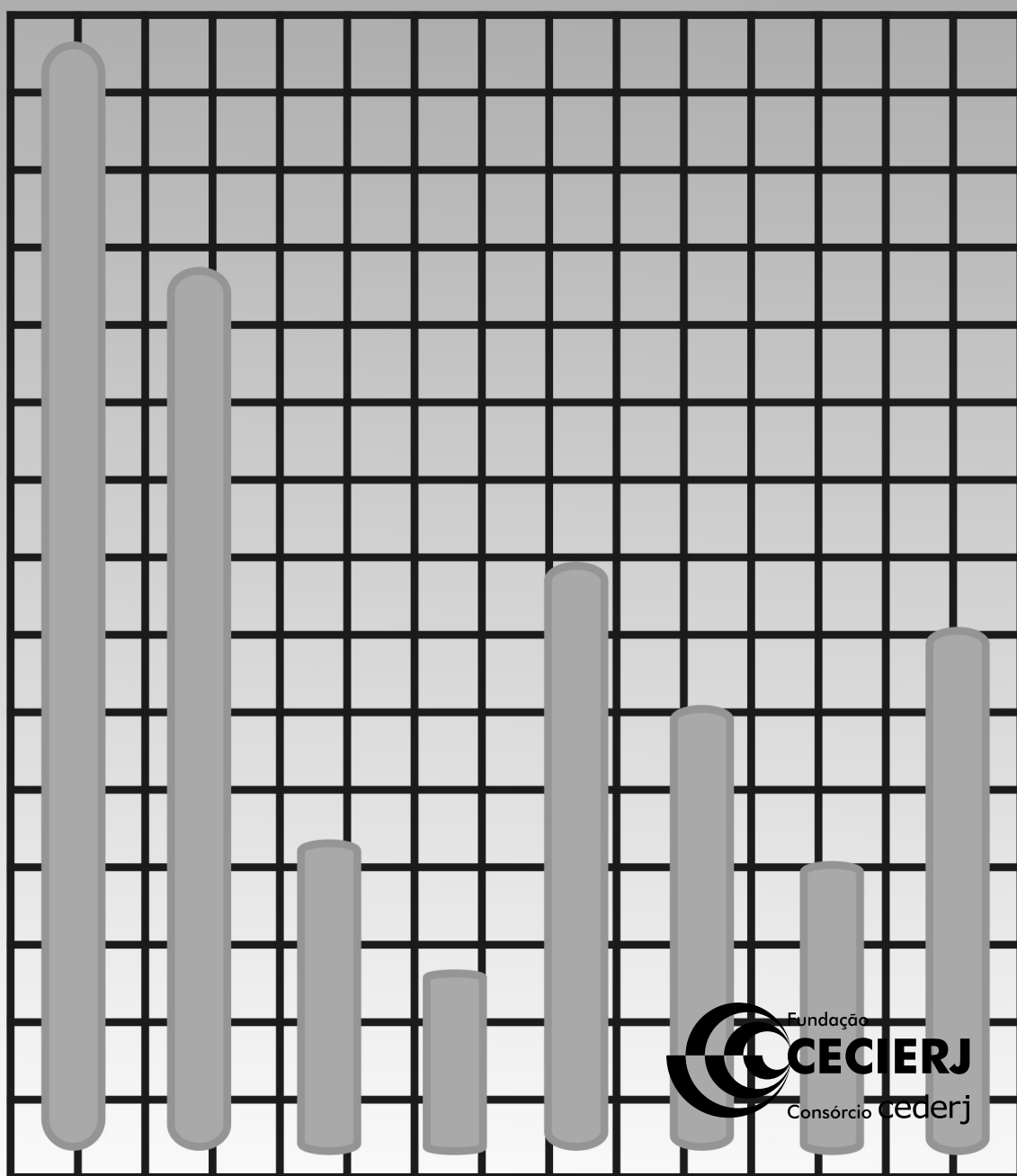


# Estatística Aplicada à Administração





Ministério da Educação – MEC  
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES  
Diretoria de Educação a Distância – DED  
Universidade Aberta do Brasil – UAB  
Programa Nacional de Formação em Administração Pública – PNAP  
Bacharelado em Administração Pública

# ESTATÍSTICA APLICADA À ADMINISTRAÇÃO

Marcelo Tavares



2011

© 2011. Universidade Federal de Santa Catarina – UFSC. Todos os direitos reservados.

A responsabilidade pelo conteúdo e imagens desta obra é do(s) respectivo(s) autor(es). O conteúdo desta obra foi licenciado temporária e gratuitamente para utilização no âmbito do Sistema Universidade Aberta do Brasil, através da UFSC. O leitor se compromete a utilizar o conteúdo desta obra para aprendizado pessoal, sendo que a reprodução e distribuição ficarão limitadas ao âmbito interno dos cursos. A citação desta obra em trabalhos acadêmicos e/ou profissionais poderá ser feita com indicação da fonte. A cópia desta obra sem autorização expressa ou com intuito de lucro constitui crime contra a propriedade intelectual, com sanções previstas no Código Penal, artigo 184, Parágrafos 1º ao 3º, sem prejuízo das sanções cíveis cabíveis à espécie.

T231e	<p>Tavares, Marcelo</p> <p>Estatística aplicada à administração / Marcelo Tavares. – Florianópolis : Departamento de Ciências da Administração / UFSC; [Brasília] : CAPES : UAB, 2011.</p> <p>222p. : il.</p> <p>Inclui bibliografia</p> <p>Bacharelado em Administração Pública</p> <p>ISBN: 978-85-7988-099-5</p> <p>1. Administração – Métodos estatísticos. 2. Estatística. 3. Probabilidades. 4. Educação a distância. I. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Brasil). II. Universidade Aberta do Brasil. III. Título.</p> <p>CDU: 519.2:65</p>
-------	---

*Catálogo na publicação por: Onélia Silva Guimarães CRB-14/071*

**PRESIDENTA DA REPÚBLICA**

*Dilma Vana Rousseff*

**MINISTRO DA EDUCAÇÃO**

*Fernando Haddad*

**PRESIDENTE DA CAPES**

*Jorge Almeida Guimarães*

**UNIVERSIDADE FEDERAL DE SANTA CATARINA**

REITOR

*Alvaro Toubes Prata*

VICE-REITOR

*Carlos Alberto Justo da Silva*

**CENTRO SÓCIO-ECONÔMICO**

DIRETOR

*Ricardo José de Araújo Oliveira*

VICE-DIRETOR

*Alexandre Marino Costa*

**DEPARTAMENTO DE CIÊNCIAS DA ADMINISTRAÇÃO**

CHEFE DO DEPARTAMENTO

*Gilberto de Oliveira Moritz*

SUBCHEFE DO DEPARTAMENTO

*Marcos Baptista Lopez Dalmau*

**DIRETORIA DE EDUCAÇÃO A DISTÂNCIA**

DIRETORIA DE EDUCAÇÃO A DISTÂNCIA

COORDENAÇÃO GERAL DE ARTICULAÇÃO ACADÊMICA

*Liliane Carneiro dos Santos Ferreira*

COORDENAÇÃO GERAL DE SUPERVISÃO E FOMENTO

*Grace Tavares Vieira*

COORDENAÇÃO GERAL DE INFRAESTRUTURA DE POLOS

*Joselino Goulart Junior*

COORDENAÇÃO GERAL DE POLÍTICAS DE INFORMAÇÃO

*Adi Balbinot Junior*

**COMISSÃO DE AVALIAÇÃO E ACOMPANHAMENTO – PNAP**

*Alexandre Marino Costa*

*Claudinê Jordão de Carvalho*

Eliane Moreira Sá de Souza  
Marcos Tanure Sanabio  
Maria Aparecida da Silva  
Marina Isabel de Almeida  
Oreste Preti  
Tatiane Michelin  
Teresa Cristina Janes Carneiro

## **METODOLOGIA PARA EDUCAÇÃO A DISTÂNCIA**

Universidade Federal de Mato Grosso

### **COORDENAÇÃO TÉCNICA – DED**

Tatiane Michelin  
Tatiane Pacanaro Trinca  
Soraya Matos de Vasconcelos

### **AUTOR DO CONTEÚDO**

Marcelo Tavares

### **EQUIPE DE DESENVOLVIMENTO DE RECURSOS DIDÁTICOS CAD/UFSC**

Coordenador do Projeto  
*Alexandre Marino Costa*

Coordenação de Produção de Recursos Didáticos  
*Denise Aparecida Bunn*

Supervisão de Produção de Recursos Didáticos  
*Érika Alessandra Salmeron Silva*

Designer Instrucional  
*Adriana Novelli*  
*Andreza Regina Lopes da Silva*  
*Denise Aparecida Bunn*

Auxiliar Administrativo  
*Stephany Kaori Yoshida*

Capa  
*Alexandre Noronha*

Projeto Gráfico e Finalização  
*Annye Cristiny Tessaro*

Diagramação  
*Rita Castelan Minatto*

Ilustração  
*Livia Remor Pereira*

Revisão Textual  
*Claudia Leal Estevão Brites Ramos*  
*Mara Aparecida Andrade R. Siqueira*

## PREFÁCIO

Os dois principais desafios da atualidade na área educacional do País são a qualificação dos professores que atuam nas escolas de educação básica e a qualificação do quadro funcional atuante na gestão do Estado brasileiro, nas várias instâncias administrativas. O Ministério da Educação (MEC) está enfrentando o primeiro desafio com o Plano Nacional de Formação de Professores, que tem como objetivo qualificar mais de 300.000 professores em exercício nas escolas de Ensino Fundamental e Médio, sendo metade desse esforço realizado pelo Sistema Universidade Aberta do Brasil (UAB). Em relação ao segundo desafio, o MEC, por meio da UAB/CAPES, lança o Programa Nacional de Formação em Administração Pública (PNAP). Esse Programa engloba um curso de bacharelado e três especializações (Gestão Pública, Gestão Pública Municipal e Gestão em Saúde) e visa a colaborar com o esforço de qualificação dos gestores públicos brasileiros, com especial atenção no atendimento ao interior do País, por meio dos Polos da UAB.

O PNAP é um Programa com características especiais. Em primeiro lugar, surgiu do esforço e da reflexão de uma rede composta pela Escola Nacional de Administração Pública (ENAP), pelo Ministério do Planejamento, pelo Ministério da Saúde, pelo Conselho Federal de Administração, pela Secretaria de Educação a Distância (SEED) e por mais de 20 instituições públicas de Ensino Superior (IPES), vinculadas à UAB, que colaboraram na elaboração do Projeto Político Pedagógico dos cursos. Em segundo lugar, esse Projeto será aplicado por todas as instituições e pretende manter um padrão de qualidade em todo o País, mas abrindo margem para

que cada IPES, que ofertará os cursos, possa incluir assuntos em atendimento às diversidades econômicas e culturais de sua região.

Outro elemento importante é a construção coletiva do material didático. A UAB colocará à disposição das IPES um material didático mínimo de referência para todas as disciplinas obrigatórias e para algumas optativas. Esse material está sendo elaborado por profissionais experientes da área da Administração Pública de mais de 30 diferentes instituições, com o apoio de equipe multidisciplinar. Por último, a produção coletiva antecipada dos materiais didáticos libera o corpo docente IPES para uma dedicação maior ao processo de gestão acadêmica dos cursos; uniformiza um elevado patamar de qualidade para o material didático e garante o desenvolvimento ininterrupto dos cursos, sem paralisações que sempre comprometem o entusiasmo dos estudantes.

Por tudo isso, estamos seguros de que mais um importante passo em direção à democratização do Ensino Superior público e de qualidade está sendo dado, desta vez contribuindo também para a melhoria da gestão pública brasileira.

*Celso José da Costa*  
*Diretor de Educação a Distância*  
*Coordenador Nacional da UAB*  
*CAPES-MEC*



# SUMÁRIO

Apresentação.....	11
-------------------	----

## **Unidade 1** – Fases do Método Estatístico, População e Amostra

Fases do Método Estatístico.....	17
Definição do Problema.....	18
Planejamento.....	19
Coleta de Dados.....	19
Organização de Dados.....	20
Apresentação de Dados.....	21
Análise e Interpretação de Dados.....	22
População e Amostras.....	25
Amostragens Probabilísticas.....	31
Amostragens não Probabilísticas.....	40

## **Unidade 2** – Distribuições de Frequências e Representação Gráfica

Distribuições de Frequências.....	49
Distribuições de Frequências de uma Variável Quantitativa Contínua.....	51
Distribuições de Frequências de uma Variável Qualitativa.....	58
Distribuições de Frequências de uma Variável Quantitativa Discreta.....	58
Representação Gráfica.....	61

**Unidade 3 – Medidas de Posição e Dispersão**

Medidas de Posição..... 73

    Média..... 74

    Moda..... 78

    Mediana..... 80

    Separatrizes..... 82

Medidas de Dispersão..... 85

    Amplitude Total..... 86

    Variância..... 87

    Desvio Padrão..... 89

    Coeficiente de Variação..... 90

**Unidade 4 – Probabilidade**

Introdução..... 101

    Experimento Aleatório..... 103

    Espaço Amostral..... 105

    Evento..... 106

Definições de Probabilidades..... 107

Probabilidade Condicional..... 113

    Regra do Produto e Eventos Independentes..... 116

Algumas Regras Básicas de Probabilidades..... 122

**Unidade 5 – Distribuição de Probabilidades Discretas e Contínuas**

Introdução..... 129

Distribuições Discretas..... 131

    Distribuição Binomial..... 132

    Distribuição de Poisson..... 136

Distribuições Contínuas.....	140
Distribuição Normal.....	140
Distribuições Amostrais.....	148
Distribuição <i>t</i> de <i>Student</i> .....	149
Distribuição de Qui-Quadrado.....	153
Distribuição de F.....	156
Noções de Estimação.....	159
Estimação por Intervalos.....	161
Dimensionamento de Amostras.....	164

## **Unidade 6 – Testes de Hipóteses**

Introdução.....	171
Estrutura dos Testes de Hipóteses.....	173
Teste de Hipótese para uma Média.....	178
Teste de Hipótese para a Razão de duas Variâncias.....	183
Teste de Hipótese para a Diferença entre Médias.....	185
Teste de Hipótese para a Diferença entre Proporções.....	198
Teste de Qui-Quadrado de Independência.....	201
Associação entre Variáveis.....	205
Considerações finais.....	219
Referências.....	220
Minicurriculo.....	222



# APRESENTAÇÃO

Seja bem-vindo ao estudo da Estatística!

Esperamos que esta disciplina seja uma experiência interessante e enriquecedora. Pensando nisso, elaboramos o material com cuidado para que você aprenda os principais conceitos associados à Estatística, que vem se tornando cada vez mais importante no competitivo ambiente de negócios e de gestão. Juntos, iremos viajar pelo mundo dos números associados à estatística e suas relações no dia a dia do gestor público.

O principal objetivo é que você tenha a oportunidade de ampliar seu conhecimento sobre o universo da estatística. Dessa forma, não serão feitas neste material deduções e demonstrações matemáticas de expressões. Será tentado um entendimento mais abstrato das expressões a serem utilizadas.

Você já deve estar acostumado a utilizar a estatística, ou ferramentas estatísticas, no seu dia a dia, sem saber que a está utilizando. Se você acha que a estatística se resume apenas a números e a gráficos, está redondamente enganado. Dessa forma, estaremos, a partir de agora, entrando em um mundo no qual os números irão sempre lhe falar ou lhe contar alguma coisa. O seu trabalho usando a estatística passará a ser o de ajudar a planejar a obtenção de dados, a interpretar e a analisar os dados obtidos e a apresentar os resultados de maneira a facilitar a sua tomada de decisões como gestor na área pública.

Para gerar tabelas, gráficos e utilizar técnicas estatísticas, temos uma infinidade de *softwares* que fazem isso automaticamente. Entretanto, para podermos descobrir quais as respostas que os dados podem nos dar para determinados questionamentos, é necessário

que você saiba a teoria estatística e treine suas aplicações por meio de estudos de casos, ou situações.

Sempre surgem, então, perguntas do tipo: quais variáveis devem ser medidas? Como retirar amostras da população que se deseja estudar? Que tipo de análise realizar? Como interpretar os resultados? Espero que no final da leitura deste material você tenha condições de responder de forma clara a essas perguntas e a outras que podem ser feitas.

É necessário termos em mente que a estatística é uma ferramenta para o gestor ou para o executivo, nas respostas aos “porquês” de seus problemas. Contudo, para que ela seja bem utilizada, é necessário conhecer os seus fundamentos e os seus princípios e, acima de tudo, que o gestor desenvolva um espírito crítico e de análise; pois é fácil mentir usando a estatística, difícil é falar a verdade sem usar a estatística.

Atualmente, as empresas têm procurado admitir como gestores os profissionais que tenham um nível de conhecimento de estatística alto, pois esse conhecimento tem resultado em diferença significativa nos processos decisórios.

Para estudar na modalidade a distância o conteúdo da disciplina *Estatística Aplicada à Administração* é preciso que você tenha disciplina intelectual que, para desenvolver, somente praticando e, ainda, uma postura crítica, sistemática. Ou seja, ao invés de você atuar como passivo e concordar com tudo o que diz o texto, você deve **duvidar, contestar, criticar, comentar e descobrir o que o autor quer dizer**. O ato de estudar exige que você **faça exercícios e entenda o que está fazendo**, não sendo apenas um mero executor de fórmulas. Isso implica o entendimento dos conceitos apresentados neste material.

Uma vez que a leitura é uma atividade, você deve ser ativo. Tenha certeza de que um estudante consegue aprender mais do que o outro à medida que se aplica mais e é capaz de uma atividade maior de leitura. E aprende melhor se exigir mais de si mesmo e do texto que tem diante de si.

Para tanto, dividimos o livro em seis Unidades. Na Unidade 1, iremos ver as fases do método estatístico e os conceitos de populações, de amostras e de métodos de amostragem. Nas Unidades 2 e 3, vamos aprender a descrever um conjunto de dados por meio de distribuições de frequências, de medidas de posição e de dispersão. Já nas Unidades 4 e 5, estaremos aprendendo conceitos relacionados a probabilidades, a distribuições discretas e contínuas, além de noções de estimação. E, por fim, na última Unidade, você irá aprender como tomar decisões baseadas nos chamados testes de hipóteses.

Desejo a você bons estudos!

*Professor Marcelo Tavares*





# UNIDADE 1

## FASES DO MÉTODO ESTATÍSTICO, POPULAÇÃO E AMOSTRA

### OBJETIVOS ESPECÍFICOS DE APRENDIZAGEM

Ao finalizar esta Unidade, você deverá ser capaz de:

- ▶ Entender as relações entre as fases do método estatístico e aplicá-las no desenvolvimento de seus projetos;
- ▶ Compreender conceitos básicos relacionados à estatística, como variáveis, estimadores, estimativas, parâmetros, população, amostras; e
- ▶ Entender os diversos tipos de amostragem e saber como aplicá-los quando for desenvolver qualquer tipo de projeto em que se utilize planos amostrais.



# FASES DO MÉTODO ESTATÍSTICO

Caro estudante,

Vamos iniciar nossos estudos de estatística para que você tenha condições de identificar a forma pela qual podemos utilizá-la, seja dentro da pesquisa científica ou na tomada de decisões e na estruturação de projetos.

Além disso, trabalharemos a definição de população e de amostra, bem como a forma de retirar as amostras de uma população; temas de fundamental importância para que você consiga desenvolver trabalhos com resultados de campo de alto nível.

Vamos então aprender esses assuntos? Boa leitura e, qualquer dúvida, não hesite em consultar o seu tutor.

Para realizarmos um estudo estatístico, normalmente, existem várias etapas a serem realizadas. Essas etapas são chamadas de fases do método estatístico. Quando você tiver bem definido essas fases, e tiver condições de realizá-las de forma adequada, a chance de sucesso em um trabalho estatístico ou que envolva estatística será muito maior. Para isso, então, você irá conhecer essas fases ou etapas de forma mais detalhada.

As fases do método estatístico são:

- ▶ definição do problema;
- ▶ planejamento do processo de resolução;
- ▶ coleta dos dados;
- ▶ organização de dados;
- ▶ apresentação de dados; e
- ▶ análise e interpretação dos resultados.

Agora, você verá de forma detalhada cada uma dessas fases. Ao longo da apresentação, iremos detalhando cada uma delas, inserindo-as passo a passo, para que no final você tenha uma ideia das relações entre essas fases.

## DEFINIÇÃO DO PROBLEMA

A primeira fase consiste na definição e na formulação correta do problema a ser estudado. Para isso, você deve procurar outros estudos realizados sobre o tema escolhido, pois, desse modo, você evitará cometer erros que já tenham sido cometidos por outros.

Sendo assim, essa primeira fase pode responder à definição de um problema ou, simplesmente, dar resposta a um interesse de profissionais. Em alguns casos, podem estar envolvidas **variáveis qualitativas e quantitativas**, por exemplo:

Veremos esses conceitos mais adiante nesta Unidade, ao longo da disciplina.

- ▶ a receita do Imposto Territorial e Predial Urbano (IPTU) de cada um dos bairros de uma cidade em vários anos;
- ▶ o tratamento de dados relativos aos desempenhos dos funcionários de um setor de uma prefeitura ao longo de alguns meses;
- ▶ a quantidade de residências em uma cidade que atrasam o pagamento do IPTU em 1, 2, 3, 4, 5 ou mais meses; e
- ▶ o tempo necessário entre o pedido de reparo de uma via pública e a realização do serviço.

Mas não para por aí! Existem outros problemas relacionados à gestão pública que merecem ser resolvidos.

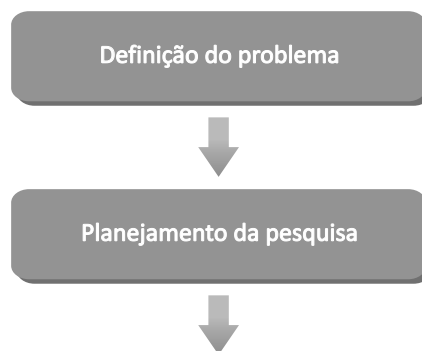
Definição do problema



## PLANEJAMENTO

Após você definir o problema, é preciso determinar um processo para resolvê-lo e, em especial, a forma de como obter informações sobre a variável ou as variáveis em estudo. É nessa fase que devemos decidir pela observação da população ou de uma amostra. Portanto, você precisa:

- ▶ determinar os procedimentos necessários para resolver o problema, em especial, como levantar informações sobre o assunto objeto do estudo;
- ▶ planejar o trabalho tendo em vista o objetivo a ser atingido;
- ▶ escolher e formular corretamente as perguntas;
- ▶ definir o tipo de levantamento – censitário ou por amostragem; e
- ▶ definir cronograma de atividades, custos envolvidos, delineamento da amostra etc.



## COLETA DOS DADOS

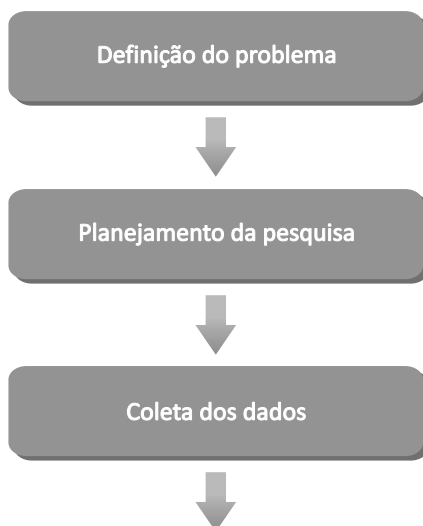
Agora que você já planejou o seu trabalho, vamos para a terceira etapa, que consiste na coleta de dados. Essa fase que deve ser seguida com cuidado, pois dados mal coletados resultam em

estatísticas inadequadas ou que não refletem a situação que você deseja estudar.

Os dados podem ser coletados, por exemplo, por meio de:

- ▶ questionário;
- ▶ observação;
- ▶ experimentação; e
- ▶ pesquisa bibliográfica.

A coleta de dados que você vai fazer pode ser realizada de forma direta com base nos elementos de registros ou pelo próprio pesquisador através de questionários.

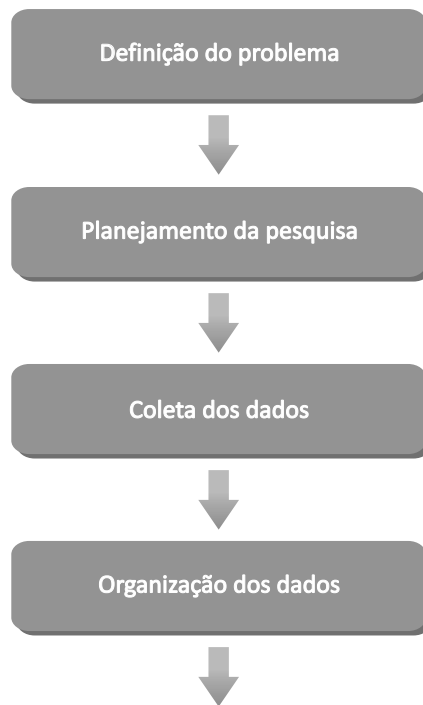


## ORGANIZAÇÃO DE DADOS

\*Atributo estatístico – é toda medida estatística, por exemplo, média.  
Fonte: Elaborado pelo autor.

\*Tratamento estatístico – implica analisar os dados utilizando técnicas estatísticas. Fonte: Elaborado pelo autor.

Agora que temos os dados precisamos organizá-los, pois somente coletar os dados não é suficiente. Essa organização consiste em “resumir” os dados através da sua contagem e agrupamento. Desse modo, obtemos um conjunto de informações que irá conduzir ao estudo do **atributo estatístico\***. Geralmente, essa organização é feita em planilhas eletrônicas (tipo Excel) para posterior **tratamento estatístico\***.

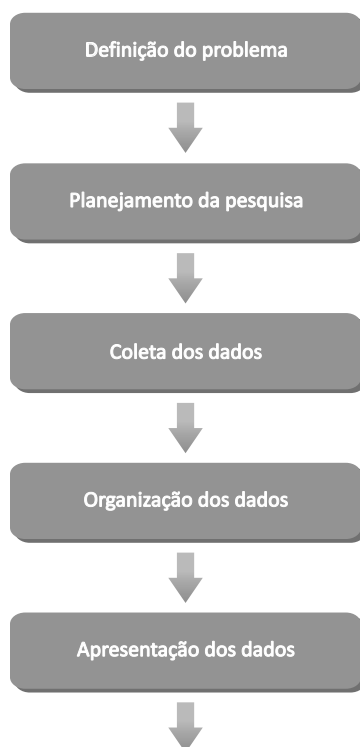


## APRESENTAÇÃO DE DADOS

Agora que temos os dados organizados, precisamos apresentá-los e, para tanto, existem duas formas de apresentação que você poderá utilizar, que não se excluem mutuamente, a saber:

- ▶ apresentação por tabelas; e
- ▶ apresentação por gráficos.

Essas formas de expor as informações coletadas permitem sintetizar uma grande quantidade de dados (números), tornando mais fácil a compreensão do atributo em estudo e permitindo uma futura análise.



Na Unidade 2, ampliaremos nossa discussão quanto à forma de apresentação dos dados, ou seja, detalharemos como montar essas tabelas de distribuição de frequências, quais os tipos de gráficos mais adequados para cada situação que você venha a ter.

\*Médias – são os resultados obtidos por meio da soma de todos os valores, divididos pela quantidade de números que você somou. Fonte: Elaborado pelo autor.

## ANÁLISE E INTERPRETAÇÃO DE DADOS

Nessa fase, você irá calcular novos números com **médias\*** embasadas nos dados coletados. Esses novos números permitem fazer uma descrição do fenômeno em estudo, evidenciando algumas das suas características particulares. Nessa fase, ainda é possível, por vezes, “arriscar” alguma generalização a qual envolverá sempre algum grau de incerteza.

Na fase de análise e de interpretação dos dados, precisamos, ainda, estar muito atentos ao significado das medidas estudadas,



Como quantificar esse grau de incerteza, você irá aprender na Unidade 5.



por exemplo, média e **moda\*** e ao porquê de as utilizarmos. Para verificar as relações entre essas medidas, você deve estar de mente aberta. E, para tanto, é necessário que você conheça a estrutura e o cálculo dessas medidas.

\*Moda – valor que mais se repete em um conjunto de observações. Fonte: Elaborado pelo autor.

Imagine que você esteja envolvido em um estabelecimento de conjecturas e na comunicação da informação de uma forma convincente através da elaboração de relatórios, de textos e de artigos que incluam, por exemplo, gráficos e tabelas. As pessoas que se utilizam da estatística como ferramenta devem ser sensibilizadas para a influência que poderá ter o modo de apresentação da informação na comunicação de resultados, a utilização de diferentes gráficos e/ou de diferentes escalas.

Para compreender essa nossa conversa, analise a Figura 1, que apresenta um resumo de todas essas fases:

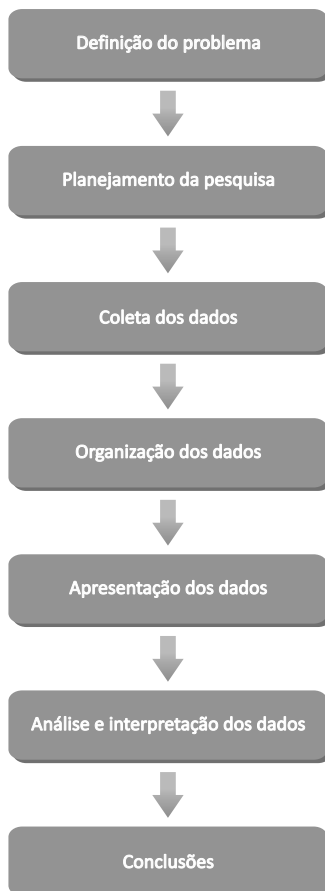


Figura 1: Fases do método estatístico  
Fonte: Elaborada pelo autor

Por fim, é importante destacarmos que para a realização dessa fase de análise faz-se necessário que você tenha o domínio da utilização de planilhas tipo Excel e de *softwares* estatísticos.

## POPULAÇÃO E AMOSTRAS

Antes, precisamos entender o que é uma população e o que é uma amostra. Se considerarmos somente os habitantes de uma cidade que contribuem com o pagamento do IPTU, podemos considerar essas pessoas como sendo a população, pois apresentam características em comum; sendo, nesse caso, o fato de que elas estão na mesma cidade e contribuem, todas, com o imposto do IPTU.

Suponha, todavia, que você queira trabalhar com apenas uma parte dessa população, a qual apresente as mesmas características da população da qual você terá uma amostra, ou seja, uma porção ou fração da população que preserva todas as características importantes dos elementos que a integram.

Nessa população, geralmente, você poderá medir uma variável, por exemplo, a renda dessas pessoas. Assim, você poderá querer calcular a renda média dessa população de pessoas que contribuem com o IPTU (média populacional ( $\mu$ ) que corresponde, geralmente, a um valor desconhecido chamado de parâmetro). Como você não vai medir toda a população, podemos obter uma amostra que represente essa população e, estudando a amostra, você terá condições de calcular a média amostral ( $\bar{x}$ ) que corresponde ao estimador, e o resultado obtido (valor numérico) corresponderá à estimativa. Para entender melhor essa relação, observe a Figura 2, a seguir.

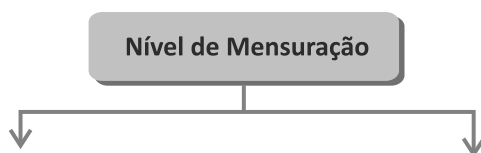


Figura 2: Relações entre estimadores, parâmetros e estimativa  
Fonte: Elaborada pelo autor

Para você entender melhor a figura anterior, verifique que  $\mu$  (média populacional) e  $\sigma$  (desvio padrão populacional) correspondem aos parâmetros (população), e  $\bar{x}$  corresponde ao estimador (amostra) e R\$ 587,00 corresponde à estimativa da renda média populacional (aproximação numérica do valor da população).

Portanto, quando você está estudando uma população inteira (censo) ou realizando uma amostragem, a classificação da variável que está trabalhando será muito importante. Então, vamos ver a classificação das variáveis em relação a sua natureza; as quais podem ser: qualitativas (ordinais ou nominais) e quantitativas (discretas ou contínuas). Essa classificação permitirá, por exemplo, que você defina, posteriormente, o tipo de teste de hipótese a ser utilizado ou o tipo de distribuição de probabilidade que necessitará aplicar para a variável em questão.

Sendo assim, precisamos entender a classificação das variáveis. Então, mãos à obra! A classificação:



- **Variável qualitativa:** faz referência a observações relacionadas a atributos que não apresentam estrutura numérica, como cor dos olhos, classe social, estado civil, nome da empresa etc. Essa variável qualitativa pode ser classificada em:
  - **Nominal:** quando as observações não apresentam nenhuma hierarquia ou ordenamento, como o sexo

dos funcionários de uma prefeitura; estado civil; naturalidade, etc.

- **Ordinal:** quando as observações apresentam uma hierarquia ou um ordenamento, por exemplo, cargo do funcionário de uma empresa (diretor, gerente, supervisor etc.); posição das empresas em relação a pagamento de impostos em um município (primeira, segunda, terceira etc.).



- **Variável quantitativa:** essa variável está relacionada às observações que apresentam uma estrutura numérica associada a contagens ou a mensurações, como quantidade de energia elétrica consumida por uma prefeitura em um mês; número de pessoas atendidas por hora em um determinado setor público etc. Essa variável quantitativa pode ser classificada em:
  - **Discreta:** observações de estrutura numérica estão associadas a valores fixos, ou seja, na maioria dos casos, números inteiros e positivos associados a contagens, como o número de pessoas que pagam seus impostos em dia etc.

- **Contínua:** são todas as observações que representam valores numéricos que podem assumir qualquer valor dentro de um intervalo, ou seja, os números reais, por exemplo, o tempo que pessoas ficam na fila aguardando para serem atendidas; peso dos funcionários de uma prefeitura etc.

Para melhor visualizar essa classificação das variáveis, observe a Figura 3, apresentada a seguir.

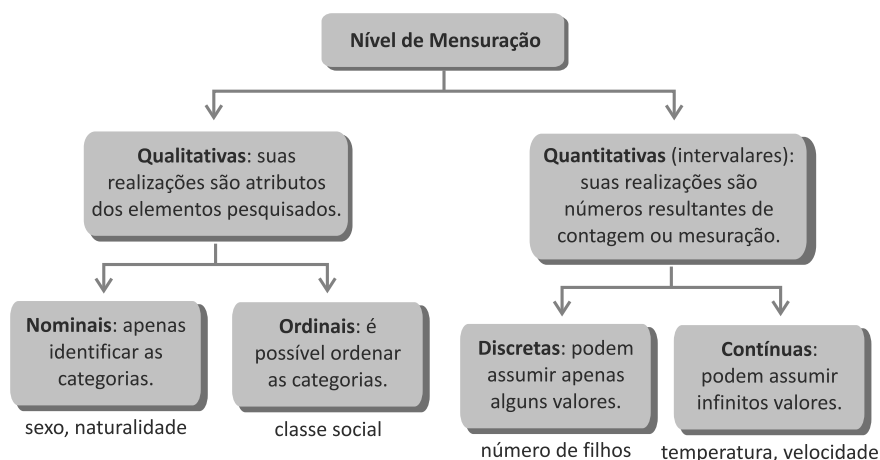


Figura 3: Classificação das variáveis

Fonte: Elaborada pelo autor

Agora que você já conhece e compreendeu a classificação das variáveis, vamos voltar à relação entre amostragens e populações. A amostragem é o estudo das relações existentes entre a amostra e a população de onde ela – a amostra – foi extraída e a forma como ocorreu essa extração.

As principais vantagens da utilização do estudo por **amostras representativas\*** em relação ao **censo\*** são:

- Ocorre uma redução no custo, pois sendo os dados obtidos apenas de uma fração da população, as despesas são menores do que as oriundas de um censo. Tratando-se de grandes populações, podemos obter resultados suficientemente precisos, para serem úteis, de amostras que representam apenas uma pequena fração da população.

\*Amostras representativas – são as amostras que mantêm as características da população de onde ela foi retirada. Fonte: Elaborado pelo autor.

\*Censo – avaliação de toda a população. Fonte: Elaborado pelo autor.

- ▶ Na prática ou no dia a dia das organizações, é necessário que os resultados sejam obtidos com a maior rapidez possível. Portanto, com a amostragem, você pode apurar os dados e sintetizá-los mais rapidamente do que em uma contagem completa. Esse é um fator primordial quando se necessita urgentemente das informações. Se o resultado de uma pesquisa for conhecido muito tempo depois, é bem possível que a situação que você pretendia resolver seja, no momento presente, completamente diferente da que existia no momento da coleta dos dados.
- ▶ Outra vantagem corresponde a maior amplitude e flexibilidade. Em certos tipos de investigação, como ocorre em pesquisas de mercado, temos que utilizar pessoal bem treinado e equipamento de alta tecnologia, cuja disponibilidade é limitada para a obtenção de dados. O censo completo torna-se impraticável e resta a escolha de obter as informações por meio de uma amostra. Portanto, com um número reduzido de entrevistadores, por exemplo, o treinamento a ser aplicado com eles é de qualidade muito maior do que em um grupo maior de entrevistadores.
- ▶ A última vantagem a ser citada aqui é a maior exatidão dos resultados. Em virtude de se poder empregar pessoal de melhor qualidade e mais treinado, e por se tornar exequível a supervisão mais cuidadosa do campo de trabalho e do processamento de dados, favorecendo à redução no volume de trabalho, uma amostragem “pode”, na realidade, proporcionar melhores resultados do que o censo.

Dessa forma, podemos dizer que as amostras a serem trabalhadas devem apresentar uma característica importante que corresponde à **representatividade**. Para que as conclusões da teoria de amostragem sejam válidas, as amostras devem ser escolhidas de modo a serem representativas da população.

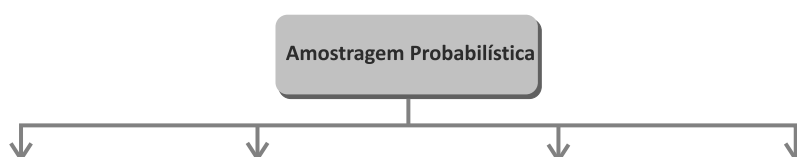
*Antes de darmos continuidade, reflita: como você faria para retirar uma amostra de 300 pessoas que estão em um cadastro de prefeitura que tem 60.000 pessoas? Essa amostra seria representativa da população?*

Uma vez que você tenha decidido realizar a pesquisa selecionando uma amostra da população, é preciso elaborar o **plano de amostragem\***. O plano de amostragem consiste em definir as **unidades amostrais\***, maneira pela qual a amostra será retirada (o tipo de amostragem), e o próprio tamanho da amostra.

Essas unidades amostrais podem corresponder aos próprios elementos da população, quando há acesso direto a eles, ou qualquer outra unidade que possibilite chegar até eles. Você pode considerar como população os domicílios de uma cidade da qual se deseja avaliar o perfil sócioeconômico. A unidade amostral será cada um dos domicílios, que corresponderá aos elementos da população. Caso a unidade amostral seja definida como os bairros, a unidade amostral não corresponderá aos elementos populacionais.

Temos dois tipos principais de amostragem as probabilísticas e as não probabilísticas. Observe a descrição a seguir:

- **Amostragem probabilística:** ocorre quando todos os elementos da população tiveram uma probabilidade ou a chance conhecida e diferente de zero de pertencer à amostra. Por exemplo, imagine que temos 50 funcionários de uma prefeitura em uma atividade de treinamento e você deve selecionar 10 funcionários. A realização desse tipo de amostragem somente é possível se a população for finita e totalmente acessível.

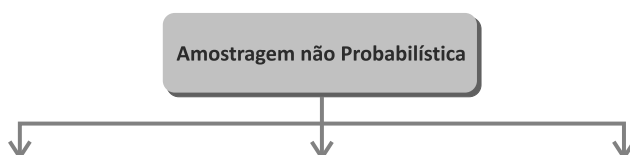


\*Plano de amostragem – plano de como será feita a retirada da amostra da população. Fonte: Elaborado pelo autor.

\*Unidades amostrais – correspondem às unidades selecionadas. Fonte: Elaborado pelo autor.



- **Amostragem não probabilística:** é assim denominada sempre que não conhecemos a probabilidade ou a chance 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, ou seja, os elementos da população a que não temos acesso não têm chance de serem sorteados para compor a amostra.



Você pode notar que a utilização de uma amostra probabilística é melhor para garantir a representatividade da amostra, pois o acaso será o único responsável por eventuais discrepâncias entre população e amostra. Essas discrepâncias são levadas em consideração nas inferências estatísticas.

Vamos, então, detalhar os tipos de amostragens probabilísticas.

## AMOSTRAGENS PROBABILÍSTICAS

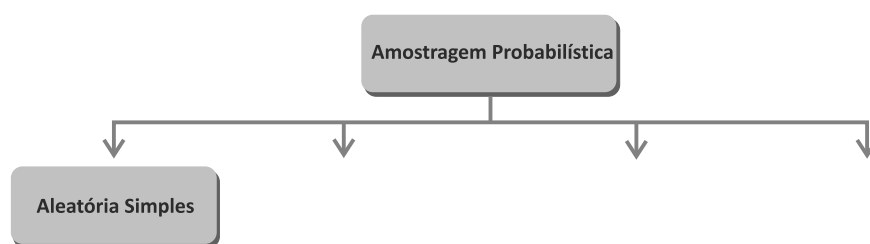
Como já mencionamos, essa amostragem é caracterizada pela chance conhecida de mensurarmos uma amostra. Os principais métodos de amostragem são: aleatória (casual) simples, sistemática, estratificada e conglomerado. Veja a seguir a descrição de cada uma delas.

### Amostragem Aleatória (Casual) Simples

Devemos utilizar a Amostragem Aleatória Simples (AAS) somente quando a população for homogênea em relação à variável que se deseja estudar. Geralmente, atribuímos uma numeração a cada indivíduo da população e através de um **sorteio aleatório** os

elementos que irão compor a amostra são selecionados. Todos os elementos da população têm a mesma probabilidade de pertencer à amostra.

Imagine que você queira amostrar um número de pessoas que estão fazendo um determinado concurso com  $N = 10.000$  inscritos. Como a população é finita, devemos enumerar cada um dos  $N$  candidatos e sortear  $n = 1.000$  deles.



### Amostragem Sistemática

Em algumas situações como amostrar pessoas que ficam em uma fila, é conveniente retirar da população os elementos que irão compor a amostra de forma cíclica (em períodos), por exemplo, quando os elementos da população se apresentam ordenados. Porém, é de fundamental importância que a variável de interesse não apresente ciclos de variação coincidente com os ciclos de retirada, pois esse fato tornará a amostragem não representativa. Essa técnica de amostragem é o que denominamos de amostragem sistemática.

Para entender melhor, vamos imaginar que você queira retirar uma amostra dos currículos apresentados pelos candidatos em um processo seletivo, e a variável de interesse corresponde à idade deles. Pode ocorrer que pessoas de uma determinada faixa etária deixem para entregar o currículo no último dia. Então, se pegássemos de forma aleatória, poderíamos estar subestimando ou superestimando a idade média. Nessa situação, foram recebidos 500 currículos ordenados por ordem de entrega. Considerando que amostrar 50 currículos é o suficiente para estimar a idade média dos candidatos, utilizamos a técnica de amostragem sistemática, pois pode ocorrer que um grupo de pessoas da mesma faixa etária tenha feito a inscrição em grupo e,

assim, na ordem de inscrição, teremos diversas pessoas com a mesma idade. Devemos considerar então que as idades estejam aleatoriamente distribuídas na população, ou seja, sem qualquer ciclo de repetição.

Para tanto, é necessário, antes, que enumeremos a população de 1 a 500 e calcularemos uma constante ( $K$ ) que servirá como fator de ciclo para retirada dos currículos amostrados. Assim, podemos dividir os 500 currículos pelo tamanho da amostra (50) que desejamos trabalhar e, então, teremos uma constante igual a 10 e os elementos serão amostrados a cada 10 elementos. Generalizando, teremos que a constante ( $K$ ) será dada por  $K = N/n$ , em que  $N$  é o tamanho da população e  $n$  o tamanho da amostra.

Após a definição do valor de  $K$ , fazemos o sorteio de um ponto inicial da amostragem (PIA), ou seja, um dos elementos do primeiro intervalo constituído pelos elementos populacionais numerados de 1 até 10. Na sequência, devemos escolher o próximo que será o elemento de ordem  $(i + K)$  e assim por diante, sempre somando  $K$  à ordem do elemento anterior até completar a escolha dos  $n$  elementos que irão compor a amostra. Um esquema é apresentado na Figura 3 no caso em que  $K = 5$ .

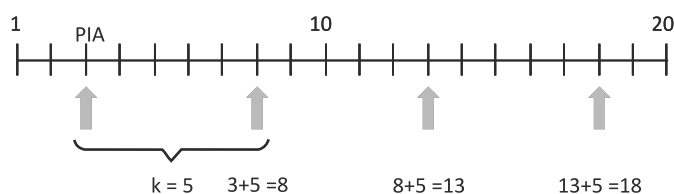


Figura 4: Exemplo de amostra sistemática

Fonte: Elaborada pelo autor

Para fixar os conceitos de amostragem sistemática, vamos fazer, juntos, um esquema de amostragem para saber a opinião dos usuários de um banco, em relação ao tempo de atendimento.

*Imagine um Banco X com uma listagem de 33.400 clientes em uma determinada cidade. A pesquisa será feita por telefone, utilizando uma estrutura de call center. Desejando que seja trabalhado com uma amostra de 300 clientes. Como seria organizada a amostragem sistemática?*

Antes, você deve dividir o número total de clientes, 33.400, por 300, que é o tamanho da amostra.

$$K = \frac{N}{n} = \frac{33400}{300} = 111,33$$

Como encontramos um valor com casas decimais, então, você irá utilizar um K de aproximadamente 111.

Agora, do primeiro cliente da lista até o de número 111, você irá sortear um número. Vamos considerar que sorteamos o cliente número 10.

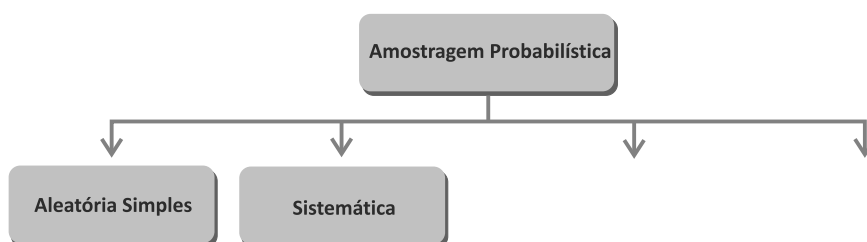
Logo, esse será o primeiro elemento da amostra.

O próximo elemento da amostra será dado pela soma do primeiro sorteado (10º cliente) ao valor de K (111).

Então, o próximo cliente sorteado será o 121º cliente (10 + 111).

Para o sorteio do próximo cliente que irá compor a amostra, teremos o 121º cliente mais o valor de K = 111, ou seja, o 232º cliente.

E, desse modo, você continua até que obtenha todos os elementos da amostra (n = 300 clientes).



### Amostragem Estratificada

Quando a variável de interesse apresenta uma heterogeneidade na população e essa heterogeneidade permite a identificação de grupos homogêneos, você pode dividir a população em grupos (estratos) e fazer uma amostragem dentro de cada estrato, garantindo, assim, a representatividade de cada estrato na amostra.

Podemos verificar que pesquisas eleitorais apresentam uma grande heterogeneidade em relação à intenção de votos quando

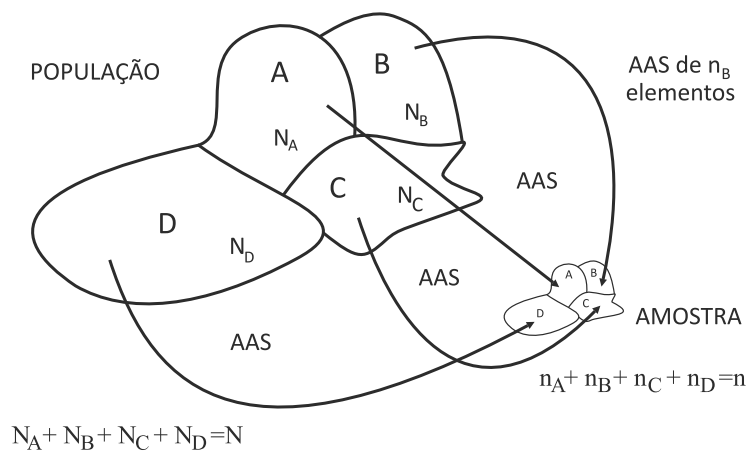
consideramos, por exemplo, a faixa salarial ou o nível de escolaridade. Então, se fizéssemos uma AAS, poderíamos incluir na amostra maior quantidade de elementos de um grupo, embora, proporcionalmente, esse grupo seja pequeno em relação à população. Dessa forma, não teríamos uma amostra representativa da população a ser estudada. Portanto, podemos dividir a população em grupos (estratos) que são homogêneos para a característica que estamos avaliando, ou seja, nesse caso a intenção de votos.

Como estamos dividindo a população em estratos (grupos) que são homogêneos dentro de si, podemos caracterizar a amostragem estratificada. Para efetuarmos a amostragem estratificada de forma proporcional, precisamos, primeiramente, definir a **proporção do estrato em relação à população**.

A proporção do estrato  $h$  será igual ao número de elementos presentes nesse estrato ( $N_h$ ) dividido pelo tamanho da população ( $N$ )  $\rightarrow (N_h/N)$ .

Após você obter essa proporção do estrato em relação à população, deve multiplicar o tamanho total da amostra ( $n$ ) pela proporção de cada estrato na população ( $N_h/N$ ).

Dessa maneira, teremos um tamanho de amostra em cada estrato proporcional ao tamanho do estrato em relação à população. A Figura 4 mostra uma população dividida em estratos (grupos) e como é feita a escolha dos elementos de cada estrato (A, B, C, D). Logo, dentro de cada um dos estratos, você pode fazer amostragem usando AAS devido aos estratos serem homogêneos individualmente, considerando a variável de interesse.



Perceba que a quantidade de elementos que iremos sortear dentro de cada estrato é proporcional ao tamanho de cada estrato na população, pois o desenho da amostra é o mesmo da população, porém menor, já que você irá pegar somente uma parte de cada estrato para compor a amostra final.

Para você fixar melhor os conceitos de amostragem estratificada, vamos resolver juntos a seguinte situação: imagine que o governo federal deseja fazer uma pesquisa de satisfação das pessoas em relação a serviços prestados por prefeituras. Estudos anteriores mostram uma relação entre a satisfação das pessoas e o tamanho da cidade. A população a ser considerada diz respeito às cidades de um determinado Estado. Essas cidades foram divididas em três grupos (estratos) levando em conta o tamanho da cidade (pequena, média e de grande porte). Considere que vamos trabalhar com uma amostra de tamanho  $n = 200$  cidades e, com as informações a seguir, faça o esquema de uma amostragem estratificada.

ESTRATOS	TAMANHO DO ESTRATO ( $N^\circ$ DE CIDADES)
Pequeno porte	$N_1 = 700$
Médio porte	$N_2 = 100$
Grande porte	$N_3 = 27$

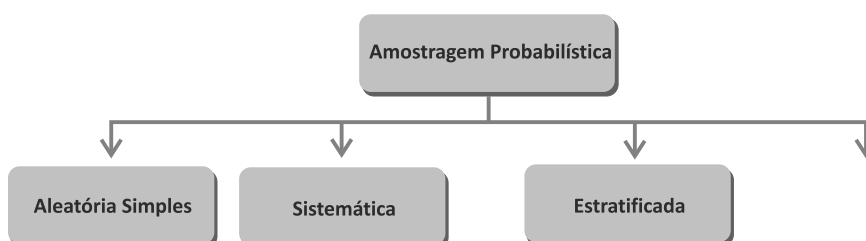
Calcule, antes, a proporção de cada estrato na população, dividindo o tamanho do estrato pelo tamanho da população ( $700 + 100 + 27 = 827$ ).

ESTRATOS	Nº DE CIDADES	PROPORÇÃO
Pequeno porte	$N_1 = 700$	$\frac{N_1}{N} = \frac{700}{827} = 0,8464$
Médio porte	$N_2 = 100$	$\frac{N_2}{N} = \frac{100}{827} = 0,1209$
Grande porte	$N_3 = 27$	$\frac{N_3}{N} = \frac{27}{827} = 0,0326$

A obtenção da quantidade de cidades que será amostrada na população será dada por meio da proporção de cada estrato multiplicado pelo tamanho total da amostra ( $n=200$ ), como é mostrado a seguir:

ESTRATOS	Nº DE CIDADES	PROPORÇÃO	Nº DE CIDADES AMOSTRADA EM CADA ESTRATO
Pequeno porte	$N_1 = 700$	$\frac{N_1}{N} = \frac{700}{827} = 0,8464$	$n_1 = 0,8464.200 = 169,28 \cong 170$
Médio porte	$N_2 = 100$	$\frac{N_2}{N} = \frac{100}{827} = 0,1209$	$n_2 = 0,1209.200 = 24,18 \cong 24$
Grande porte	$N_3 = 27$	$\frac{N_3}{N} = \frac{27}{827} = 0,0326$	$n_3 = 0,0326.200 = 6,52 \cong 6$

Então, na nossa amostra, teremos 170 cidades de porte pequeno, 24 cidades de porte médio e 6 cidades de grande porte.



### Amostragem por Conglomerados

Apesar de a amostragem estratificada apresentar resultados satisfatórios, a sua implementação é dificultada pela falta de informações sobre a população para fazer a estratificação.

Para poder contornar esse problema, podemos trabalhar com o esquema de amostragem chamado amostragem por conglomerados.

Os conglomerados são definidos em razão da experiência do gestor ou do pesquisador. Geralmente, podemos definir os conglomerados por fatores geográficos, como bairros e quarteirões. A utilização da amostragem por conglomerados possibilita uma redução significativa do custo no processo de amostragem. Portanto, um conglomerado é um subgrupo da população que, individualmente, reproduz a população. Esse tipo de amostragem é muito útil quando a população é grande, por exemplo, no caso de uma pesquisa em nível nacional.

*Você pode estar se perguntando como: realizar uma amostragem por conglomerados?*

Apesar de a amostragem por conglomerados, nesse tipo de amostragem, ser utilizada para uma população grande, é simples de calculá-la. Primeiramente, definimos o conglomerado e, assim, dividimos a população nele. Sorteamos os conglomerados por meio de um processo aleatório e avaliamos todos os indivíduos presentes neles, isso é chamado de amostragem por conglomerados em um estágio. Caso façamos um sorteio de elementos dentro de cada conglomerado, teremos uma amostragem por conglomerados em dois estágios. Para entender melhor esse cálculo, observe a Figura 5, que mostra uma amostragem por conglomerados em um único estágio. Cada quadrado corresponde a uma residência. Analise.



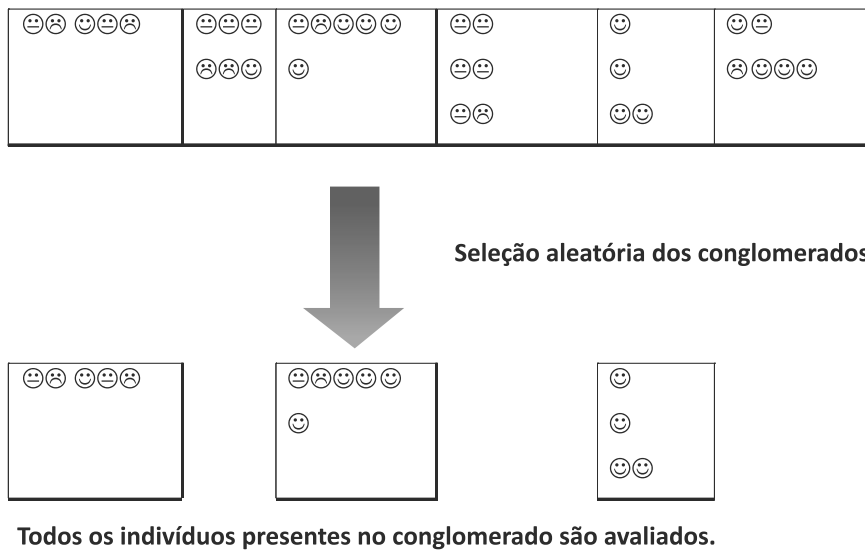
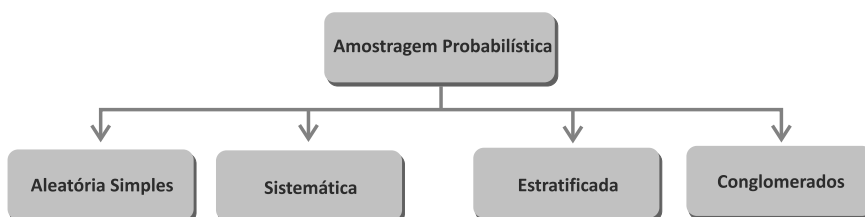


Figura 5: Amostra por conglomerados  
Fonte: Elaborada pelo autor

Um exemplo prático de utilização dessa amostra é a Pesquisa Nacional por Amostra de Domicílios (PNAD) do Instituto Brasileiro de Geografia e Estatística (IBGE), feita por conglomerados em três estágios.

Para saber mais sobre essa pesquisa acesse [www.ibge.com.br](http://www.ibge.com.br).  
Acesso em: 18 nov. 2010.



O cálculo do tamanho amostral será visto em conjunto com a parte de intervalos de confiança na Unidade 5.

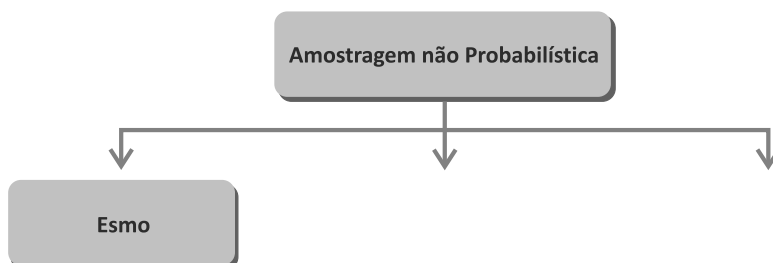
## AMOSTRAGEM NÃO PROBABILÍSTICA

Quando trabalhamos com a amostragem não probabilística, não conhecemos *a priori* a probabilidade que um elemento da população tem de pertencer à amostra. Nesse caso, não é possível calcular o erro decorrente da generalização dos resultados das análises estatísticas da amostra para a população de onde a amostra foi retirada. Então, utilizamos geralmente a amostragem não probabilística, por simplicidade ou por impossibilidade de se obter uma amostra probabilística como seria desejável.

Os principais tipos de amostragem não probabilística que temos são: amostragem sem norma, ou a esmo; intencional; e por cotas.

### Amostragem a Esmo

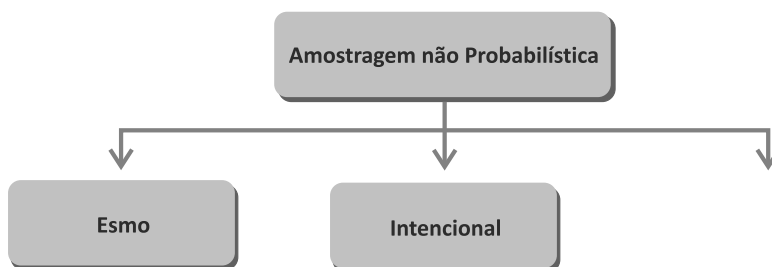
Imagine uma caixa com 1.000 parafusos. A enumeração desses parafusos ficaria muito difícil e a AAS tornar-se-ia inviável. Então, em situações desse tipo, supondo que a população de parafusos seja homogênea, escolhemos a esmo a quantidade relativa ao tamanho da amostra. Quanto mais homogênea for a população, mais podemos supor a equivalência com uma AAS. Dessa forma, os parafusos serão escolhidos para compor a amostra de um determinado tamanho sem nenhuma norma ou a esmo. Daí vem o nome desse tipo de amostragem.



## Amostragem Intencional

A amostragem intencional corresponde àquela em que o amostrador deliberadamente escolhe certos elementos para pertencer à amostra por julgar tais elementos bem representativos da população.

Um exemplo desse tipo de amostragem corresponde à situação em que desejamos saber a aceitação de uma nova marca de *whisky* a ser inserida no mercado de uma cidade. Somente entrarão para compor a amostra pessoas que façam uso da bebida e que tenham condições financeiras de comprar essa nova marca (classe social de maior poder aquisitivo).

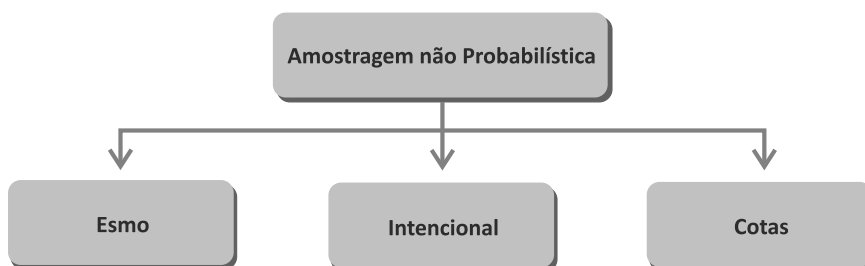


## Amostragem por Cotas

Nesse tipo de amostragem, a população é dividida em grupos e, na sequência, é determinada uma cota proporcional ao tamanho de cada grupo. Entretanto, dentro de cada grupo não é feito sorteio, mas sim os elementos são procurados até que a cota de cada grupo seja cumprida.

Encontramos esse tipo de amostra em pesquisas eleitorais quando a **divisão de uma população** ocorre em grupos; considerando, por exemplo, o sexo, o nível de escolaridade, a faixa etária e a renda, que podem servir de base para a definição dos grupos, partindo da suposição de que essas variáveis definem grupos com comportamentos diferenciados no processo eleitoral.

Para termos uma ideia do tamanho desses grupos, podemos recorrer a pesquisas feitas anteriormente pelo IBGE.



Juntando todos os desenhos dos vários tipos de amostragem que fizemos, teremos, então, a Figura 6:

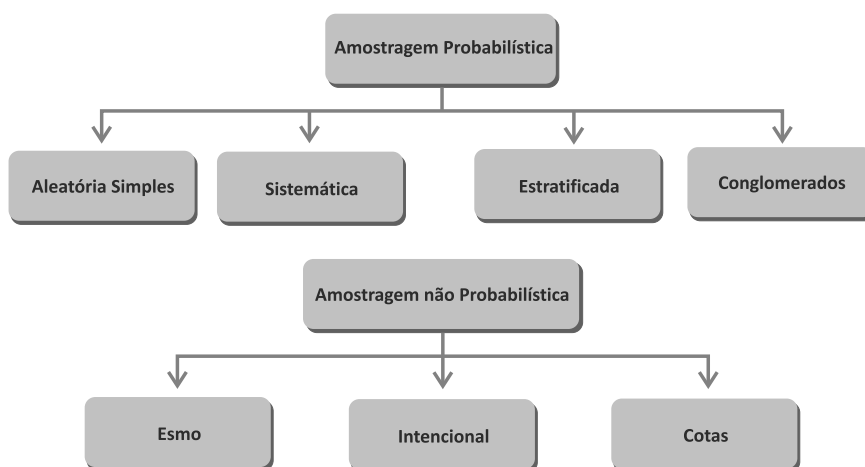


Figura 6: Tipos de amostragem

Fonte: Elaborada pelo autor

### Complementando...

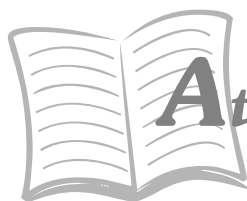
Lembre-se de que a construção do conhecimento é um processo que deve ser cíclico e renovado a cada dia; para tanto, procure descobrir mais acerca desse mundo estatístico seguindo esta orientação:

- 📌 *Programa estatístico Bioestat.* Disponível em: <<http://www.mamiraua.org.br/download/Default.aspx?dirpath=e:\home\mamiraua\Web\download\BioEstat 5 Portugues&tipo=diretorio>>. Acesso em: 18 nov. 2010. Esse programa permite que você realize os métodos de amostragem, apresentados aqui, computacionalmente.

# *Resumindo*



Nesta Unidade, você conheceu conceitos básicos relacionados à estatística e aprendeu a retirar amostras de populações. Esses conceitos serão importantes para a compreensão de novas informações contidas nas Unidades posteriores.



## Atividades de aprendizagem

Depois de ter visto todos os conceitos das fases do método estatístico, a classificação de variáveis e os diferentes planos amostrais, resolva as atividades a seguir. Lembre-se de que as respostas de todas as atividades de aprendizagem estão no final do livro. Em caso de dúvidas, você deve consultar seu tutor.

1. Imagine a situação de um pesquisador que deseje estudar o uso semanal da internet por estudantes de uma escola do Ensino Fundamental. Diferentes perguntas poderiam ser feitas, leia os exemplos e classifique-os em qualitativa nominal ou ordinal e quantitativa discreta ou contínua.
  - a) Você usa internet durante a semana? (sim ou não).
  - b) Qual a intensidade de uso da internet durante a semana? (nenhuma, pequena, média ou grande).
  - c) Quantas vezes você usa a internet durante a semana?
  - d) Por quantas horas você usa a internet durante a semana?
2. Identifique o tipo de amostragem utilizada nas situações a seguir:
  - a) Uma empresa seleciona uma a cada 300 pilhas produzidas em sua linha de produção para a realização de testes de qualidade a fim de conseguir vencer uma licitação pública.

- b) Um pesquisador de empresa aérea seleciona aleatoriamente dez voos para entrevistar todos os passageiros desses voos.
- c) Uma prefeitura testa uma nova estratégia de cobrança selecionando aleatoriamente 250 consumidores com renda inferior a R\$ 300,00 e 250 consumidores com renda de ao menos R\$ 300,00.
- d) Um eleitor indeciso resolve escolher seu candidato da seguinte forma: escreve o nome de cada um deles em cartões separados, mistura-os e extrai um nome, no qual irá votar.
- e) Um pesquisador ficou em um ponto de checagem da polícia (esquina), onde, a cada cinco carros que passavam, era feito um teste de bafômetro para checar a sobriedade do motorista.
- f) Em uma pesquisa com 1.000 pessoas, estas foram selecionadas usando-se um computador para gerar números de telefones para os quais eram, então, discados.
- g) Uma prefeitura, para não perder uma fábrica de montagem de carros, auxiliou em uma pesquisa na qual a empresa dividiu seus carros em cinco categorias: subcompacto, compacto, médio, intermediário e grande; e está entrevistando 200 proprietários de cada categoria para saber da satisfação desses clientes e, assim, ajudar a melhorar as vendas.
- h) Motivada pelo fato de um estudante ter morrido por excesso de bebida, uma universidade fez um estudo sobre o hábito de beber dos estudantes e, para isso, selecionou dez salas de aula e entrevistou os estudantes que lá estavam.





# UNIDADE 2

## DISTRIBUIÇÕES DE FREQUÊNCIAS E REPRESENTAÇÃO GRÁFICA

### OBJETIVOS ESPECÍFICOS DE APRENDIZAGEM

Ao finalizar esta Unidade, você deverá ser capaz de:

- ▶ Descrever e apresentar os resultados de um conjunto de observações a partir de uma distribuição de frequências;
- ▶ Compreender os tipos de gráficos existentes;
- ▶ Utilizar os gráficos de forma adequada; e
- ▶ Interpretar os resultados apresentados em um gráfico de forma clara, objetiva e passando o máximo de informações possíveis.



## DISTRIBUIÇÕES DE FREQUÊNCIAS

Caro estudante,

Vamos dar início a segunda Unidade de nossa disciplina e, nela, você encontrará conceitos relacionados à distribuição de frequências e à representação gráfica que lhe permitirão sintetizar uma grande quantidade de dados em tabelas e em gráficos representativos.

Quando coletamos informações, sejam de populações ou de amostras, como vimos na Unidade anterior, geralmente trabalhamos com uma quantidade grande de observações. Mas, como vamos apresentar esses resultados? Precisamos, então, aprender como sintetizar esses dados e colocá-los de forma que as pessoas possam entender as informações obtidas.

Uma forma de fazermos isso é utilizando distribuições de frequências e análises gráficas, as quais aprenderemos a partir de agora, já que entraremos no mundo da estatística, que se preocupa com a forma de apresentação dos dados. Vamos começar?

Quando coletamos os dados para uma pesquisa, as observações realizadas são chamadas de **dados brutos**\*.

Um exemplo de dados brutos corresponde ao percentual dos trabalhadores que contribuem com o Instituto Nacional de Seguro Social (INSS) em 20 cidades de uma determinada região do Brasil no ano de 2008 (dados simulados pelo autor a partir de um caso real). Os dados são apresentados na Tabela 1 na forma em que foram coletados, por esse motivo são denominados **dados brutos**. Geralmente, esse tipo de dado traz pouca ou nenhuma informação ao leitor, sendo necessário organizá-lo, com o intuito de aumentar sua capacidade de informação.

\*Dados brutos – dados na forma em que foram coletados, sem nenhum tratamento. Fonte: Elaborado pelo autor.

Tabela 1: Percentual dos trabalhadores que contribuem com o INSS em 20 cidades de uma determinada região do Brasil no ano de 2008

45	51	50	58
50	44	46	57
42	41	60	58
41	50	54	60
52	46	52	51

Fonte: Elaborada pelo autor

Se fizermos uma ordenação desse conjunto de dados brutos (do menor para o maior), teremos dados elaborados como mostra a Tabela 2.

Tabela 2: Percentual ordenado dos trabalhadores que contribuem com o INSS em 20 cidades de uma determinada região do Brasil, no ano de 2008

41	46	51	57
41	46	51	58
42	50	52	58
44	50	52	60
45	50	54	60

Fonte: Elaborada pelo autor

Com base nessa tabela, podemos observar que a simples organização dos dados em um **rol**\* aumenta muito o nível de informação destes. Na Tabela 2, você pode verificar ainda que o menor percentual foi 41% e o maior 60%, o que nos fornece uma **amplitude total**\* da ordem de 19%.

Outra informação que podemos obter dos dados por meio da Tabela 2 (organizada em rol crescente) é que nas cidades avaliadas, o valor 50, correspondente à percentagem de trabalhadores que contribuem para o INSS, ocorre com maior frequência, ou seja, é o que mais se repete.

\***Rol** – dados classificados em forma crescente ou decrescente. Fonte: Elaborado pelo autor.

\***Amplitude total** – diferença entre o maior e o menor valor observado. Fonte: Elaborado pelo autor.

*Com base nessa nossa discussão, reflita: como organizar os dados de uma variável quantitativa contínua de forma mais eficiente, na qual se possa apresentar uma quantidade maior de informações?*

A resposta a essa pergunta será apresentada na próxima seção. Fique atento e, em caso de dúvidas, lembre-se de que você não está sozinho, basta solicitar o auxílio de seu tutor.

## DISTRIBUIÇÃO DE FREQUÊNCIAS DE UMA VARIÁVEL QUANTITATIVA CONTÍNUA

Uma maneira de organizar os dados de uma variável quantitativa contínua, de modo que você melhor possa representá-la, é por meio de uma **tabela de distribuição de frequências**, que corresponde a uma tabela em que são apresentadas as frequências de cada uma das classes.

Distribuindo os dados observados em **classes\*** e contando o número de observações contidas em cada classe, obtemos a **frequência de classe**. Sendo que a disposição tabular dos dados agrupados em classes, juntamente com as frequências correspondentes, é o que denominamos de distribuição de frequência.

Sendo assim, para identificarmos uma classe, devemos conhecer os valores dos **limites inferior e superior da classe** que delimitam o **intervalo de classe**.

\*Classes – intervalos nos quais os valores da variável analisada são agrupados. Fonte: Elaborado pelo autor.

*Você pode estar se perguntando: como se constituem esses intervalos?*

Vimos, no início do curso, os tipos de intervalos na Unidade 1 da disciplina *Matemática Básica*. Vamos relembrar rapidamente como é essa classificação dos intervalos:

- ▶ **Intervalos abertos:** os limites da classe (inferior e superior) não pertencem a mesma.
- ▶ **Intervalos fechados:** os limites da classe (superior e inferior) pertencem à classe em questão.
- ▶ **Intervalos mistos:** um dos limites pertence à classe e o outro não.

Você pode utilizar qualquer um deles. Porém, os intervalos mais utilizados e que usaremos como padrão na resolução dos problemas, é o intervalo misto, o qual é apresentado da seguinte forma:

$$43,5 \vdash 48,5$$

(o 43,5 está incluído e o 48,5 não está incluído no intervalo)

Esses valores de 43,5 e 48,5 foram escolhidos aleatoriamente, somente para demonstrar o formato do intervalo.

Para você entender melhor, acompanhe o exemplo a seguir, a partir dos dados da porcentagem de trabalhadores que contribuem com o INSS. Com esses dados iremos construir uma distribuição de frequência e, ao longo desse exemplo, identificar, também, os conceitos presentes nessa distribuição.

Para darmos início a esse entendimento, é importante, antes, considerarmos que existem diversos critérios para a construção das classes das distribuições de frequências apresentados na literatura. No nosso caso, utilizaremos os critérios apresentados a seguir.

Para elaborar uma distribuição de frequência, é necessário, inicialmente, determinar o **número de classes** (k) em que os dados serão agrupados. Por questões de ordem prática e estética, sugerimos utilizar de 5 a 20 classes. O número de classes (k) a ser utilizado, pode ser calculado em função do número de observações (n), conforme é mostrado para você a seguir:

$$k = \sqrt{n}, \text{ para } n \leq 100$$

$$k = 5 \log n, \text{ para } n > 100$$

Considerando que nessa pesquisa  $n = 20$  consumidores; temos, então, o número de classes definido por  $k = \sqrt{n} = \sqrt{20} = 4,47$ ; e, como o número de classes é inteiro, usaremos 5 classes. O arredondamento utilizado nesse material é o padrão de algarismos significativos (como foi aprendido no segundo grau). O número de classes pode também ser definido de uma forma arbitrária, sem o uso dessa regra.

Após determinarmos o número de classes ( $k$ ) em que os dados serão agrupados, determinamos a **amplitude do intervalo de classe ( $c$ )**. E, para calcularmos a amplitude do intervalo de classe, vamos, primeiramente, calcular a **amplitude total dos dados ( $A$ )**, que corresponde à diferença entre o maior valor observado e o menor valor observado.

No nosso caso (usando dados da Tabela 2), teremos  $A = 60 - 41 = 19\%$ .

Com base nesse valor da amplitude total ( $A$ ) calculado, iremos obter a amplitude do intervalo de classe ( $c$ ), como é mostrado a seguir:

$$c = \frac{A}{k - 1}$$

Onde:

$c$  = amplitude de classe;

$A$  = amplitude total; e

$k$  = número de classes.

Substituindo os valores já encontrados nessa expressão e considerando o caso do exemplo que estamos resolvendo, teremos:

$$c = \frac{19}{5 - 1} = 4,75 \cong 5\%$$

Mas atenção: existem outros procedimentos para a determinação da amplitude do intervalo de classe que podem ser encontrados na literatura.

Conhecida a amplitude de classes, você deve determinar os intervalos de classe. O limite inferior e superior das classes deve ser escolhido de modo que o menor valor observado esteja localizado no **ponto médio (PM)** da primeira classe. O ponto médio da classe corresponde à soma dos limites inferior e superior dividido por dois.

Partindo desse raciocínio, o limite inferior da primeira classe será:

$$\text{Limite inf. 1ª classe} = \text{menor valor} - \frac{c}{2}.$$

No nosso caso, substituindo os valores que você encontrou anteriormente, teremos:

$$\text{Limite inf. 1ª classe} = 41\% - \frac{5\%}{2} = 38,5 \%$$

Definindo, então, o limite inferior da primeira classe basta, para você obter as classes da nossa distribuição, somar a amplitude do intervalo de classe ( $c = 5$ ) a cada limite inferior.

Assim, você terá:

- 38,5 + 43,5 → primeira classe
- 43,5 + 48,5 → segunda classe
- 48,5 + 53,5 → terceira classe
- 53,5 + 58,5 → quarta classe
- 58,5 + 63,5 → quinta classe

Com base nesse cálculo, você pode obter uma organização dos dados conforme mostra a Tabela 3, a seguir:

Tabela 3: Distribuição de frequências do percentual dos trabalhadores que contribuem com o INSS em 20 cidades de uma determinada região do Brasil no ano de 2008

CLASSES (%)	FREQUÊNCIA
38,5 + 43,5	?
43,5 + 48,5	?
48,5 + 53,5	?
53,5 + 58,5	?
58,5 + 63,5	?
Total	

Fonte: Elaborada pelo autor



Na Tabela 3 aparece uma nova denominação chamada “frequência”, em que abaixo dela há uma coluna repleta de interrogações (?). Vamos aprender a calcular valores no lugar dessas interrogações. Podemos obter frequências chamadas de **frequência absoluta** (fa), **frequência relativa** (fr) e frequência acumulada.

A **frequência absoluta (fa)** corresponde ao número de observações que temos em uma determinada classe ou em um determinado atributo de uma variável qualitativa. A **frequência relativa (fr)** corresponde à proporção do número de observações em uma determinada classe em relação ao total de observações que temos. Essa frequência pode ser expressa em termos percentuais. Para isso, basta multiplicar a frequência relativa obtida por 100.

O cálculo da frequência relativa é obtido por meio da seguinte expressão:

$$fr_i = \frac{fa_i}{\sum_{i=1}^n fa_i}$$

Sendo:

$fa_i$  = frequência absoluta da classe i.

$\sum_{i=1}^n fa_i$ : somatório das frequências absolutas para i variando de 1 até n classes, ou seja, somar as frequências de cada uma das classes ( $fa_1 + fa_2 + fa_3 + \dots + fa_n$ ).

Apresentando os dados na forma de distribuição de frequência, você consegue sintetizar as informações contidas neles, além de facilitar sua visualização. Considerando essa discussão, elaboramos a Tabela 4, que traz as frequências (fa e fr) relacionadas à variável analisada.

Tabela 4: Distribuição de frequências do percentual dos trabalhadores que contribuem com o INSS em 20 cidades de uma determinada região do Brasil, no ano de 2008

CLASSES (%)	FA (CIDADES)	FR (PROPORÇÃO DE CIDADES)
38,5 ┤ 43,5	3	0,15
43,5 ┤ 48,5	4	0,20
48,5 ┤ 53,5	7	0,35
53,5 ┤ 58,5	4	0,20
58,5 ┤ 63,5	2	0,10
Total	20	1,00

Fonte: Elaborada pelo autor

Para calcularmos a primeira proporção de 0,15, precisamos dividir a frequência da primeira classe (3) pelo total de observações (20). De forma similar, é calculada as proporções das outras classes.

*Então, como ficaria a interpretação da distribuição de frequências?*

Se considerarmos ainda a Tabela 4, podemos dizer que a porcentagem de trabalhadores que contribuem com o INSS entre 43,5% e 58,5%, dos 20 municípios avaliados em questão, está concentrada nas classes segunda, terceira e quarta, decrescendo em direção às classes do início e fim da tabela. A apresentação dos dados em forma de distribuição de frequência facilita o cálculo manual de várias medidas estatísticas de interesse e facilita, também, a apresentação gráfica dos dados.

Além das frequências absolutas e relativas, muitas vezes podemos estar interessados na quantidade de observações que existe acima ou abaixo de um determinado ponto na distribuição.

Dessa forma, você poderá trabalhar com a **frequência acumulada**, como sugere a Tabela 5, que apresenta as frequências

acumuladas da percentagem de trabalhadores que contribuem com o INSS nas 20 cidades avaliadas.

A **frequência acumulada** corresponde à soma da frequência de uma classe às frequências de todas as classes abaixo dela.

A frequência acumulada apresentada na Tabela 5 pode ser obtida da seguinte forma: abaixo do limite superior da primeira classe (43,5), temos três pessoas presentes nela, como vimos na Tabela 3 da distribuição de frequências absoluta. Quando consideramos a segunda classe (43,5 – 48,5), a frequência acumulada corresponde ao número de pessoas que temos abaixo do limite superior dessa classe (48,5), ou seja, as quatro cidades da segunda classe mais as três cidades da primeira classe, totalizando sete cidades abaixo de 48,5%. Para as outras classes, o raciocínio é semelhante.

Tabela 5: Distribuição de frequência acumulada dos trabalhadores que contribuem com o INSS em 20 cidades de uma determinada região do Brasil no ano de 2008

CLASSES (%)	FREQ. ACUMULADA	FREQ. ACUMULADA (RELATIVA)
38,5 – 43,5	3	0,15
43,5 – 48,5	7	0,35
48,5 – 53,5	14	0,70
53,5 – 58,5	18	0,90
58,5 – 63,5	20	1,00
Total		

Fonte: Elaborada pelo autor

Já o valor da frequência acumulada relativa da segunda classe (0,35) é dado pela **soma** da frequência relativa da primeira classe (0,15) e da frequência relativa da segunda classe (0,20), dando um valor acumulado para a segunda classe de 0,35.

Os valores das frequências que você usou para somar estão na Tabela 3, em caso de dúvida, retorne à tabela.



## DISTRIBUIÇÃO DE FREQUÊNCIAS DE UMA VARIÁVEL QUALITATIVA

Quando você trabalha com variáveis qualitativas, os atributos são as variações nominativas da variável. A construção da tabela consiste em contar as ocorrências dos níveis de cada atributo. O resultado da contagem define a frequência absoluta do atributo. Para podermos entender isso, tomemos como exemplo uma pesquisa na qual se procurou avaliar as frequências de cada gênero (homem ou mulher) de uma determinada cidade, que considera os serviços prestados pela prefeitura como satisfatórios, em uma amostra de 50 pessoas. Esses resultados são apresentados na Tabela 6.

Tabela 6: Distribuição de frequências do gênero de pessoas que consideram os serviços prestados pela prefeitura como satisfatórios

GÊNERO	FA	FR
Masculino	20	0,40
Feminino	30	0,60
Total	50	1,00

Fonte: Elaborada pelo autor

## DISTRIBUIÇÃO DE FREQUÊNCIAS DE UMA VARIÁVEL QUANTITATIVA DISCRETA



Vimos esse conceito na Unidade 1. Em caso de dúvida, retorne e faça uma releitura atenciosa.

Tomando-se como exemplo o caso de uma **variável aleatória discreta** (v.a), realizou-se uma pesquisa durante 30 dias em um determinado mês com relação ao número de reclamações (N.R.) no setor de tributos de uma prefeitura considerada como modelo de gestão em tributos. Os resultados encontrados você pode acompanhar na Tabela 7, a seguir:

Tabela 7: Dados referentes ao número de reclamações (NR) por dia no setor de tributos de uma prefeitura ao longo de 30 dias

DIA	N.R.	DIA	N.R.	DIA	N.R.	DIA	N.R.	DIA	N.R.
1	0	7	1	13	0	19	1	25	0
2	2	8	2	14	0	20	0	26	3
3	1	9	2	15	1	21	0	27	4
4	5	10	3	16	2	22	2	28	0
5	3	11	0	17	3	23	0	29	2
6	2	12	3	18	5	24	4	30	1

Fonte: Elaborada pelo autor

Dispondo esses dados em um rol (crescente) temos:

0 0 0 0 0 0 0 0 0 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 4 4 5 5

Podemos apresentar, a seguir, esses dados em uma distribuição de frequências. Nesse caso, não é necessário definir intervalos de classes porque a variação dos valores é pequena (varia de 0 a 5) e a variável é discreta.

Quando a variável é discreta, mas você tem uma quantidade muito grande de valores que ocorrem na amostra, então, você irá trabalhar com uma distribuição de frequências em classes. Na Tabela 8, você pode visualizar a distribuição de frequências do número de reclamações. Os cálculos das frequências absoluta e relativa são obtidos de forma semelhantes ao que foi visto anteriormente.

Tabela 8: Número de reclamações ocorridas diariamente durante certo mês

NÚMERO DE RECLAMAÇÕES POR DIA	NÚMERO DE DIAS (f <sub>A</sub> )	FREQ. RELATIVA
0	9	0.3
1	5	0.17
2	7	0.23
3	5	0.17
4	2	0.07
5	2	0.07
Total	30	1

Fonte: Elaborada pelo autor

Observe que esses valores da variável discreta correspondem a cada uma das classes.

*Você pode estar se perguntando: as tabelas de distribuição de frequências são a única forma que você tem de apresentar um conjunto de dados?*

Para descobrir a resposta a sua curiosidade, continue lendo o livro que a responderemos na seção seguinte.

## REPRESENTAÇÃO GRÁFICA

Na tentativa de responder ao seu questionamento anterior, vamos falar um pouco sobre algumas formas de **representação gráfica** de tabelas de frequência. Logicamente, dependendo do tipo de variável, temos um gráfico mais adequado. Os diferentes tipos de gráficos (histogramas, polígonos de frequência, ogivas, gráficos de setores, pictogramas e outros) permitem melhor visualização de resultados. Esses gráficos podem ser obtidos utilizando planilhas eletrônicas, como o Excel ou a planilha **CALC** do OpenOffice.

Os **histogramas** são gráficos constituídos de um conjunto de retângulos com as bases assentadas sobre um eixo horizontal, tendo o centro delas no ponto médio da classe que as representa e cuja altura é proporcional à frequência da classe. Esses gráficos são utilizados para representar tabelas intervalares.

Na Figura 7, temos o histograma da porcentagem de trabalhadores que contribuem com o INSS em cada uma das 20 cidades analisadas. Os dados utilizados, nesse gráfico, foram os da distribuição de frequências apresentados na Tabela 5, que indica o percentual de trabalhadores que contribuem com o INSS em 20 cidades de uma determinada região do Brasil em 2008.

Para você construir gráficos e distribuições de frequência, baixe o programa estatístico Bioestat que, além de ser gratuito, traz um livro na opção “ajuda”. Para isso, visite o site: <[http://www.mamiraua.org.br/download/Default.aspx?dirpath=e:\home\mamiraua\Web\download\BioEstat5 Portugues&tipo=directorio](http://www.mamiraua.org.br/download/Default.aspx?dirpath=e:\home\mamiraua\Web\download\BioEstat5%20Portugues&tipo=directorio)>. Acesso em: 19 nov. 2010.

Para saber como utilizar a planilha calc do pacote OpenOffice nas distribuições de frequências e de gráficos, acesse o site: <<http://www2.ufpa.br/dicas/open/oo-ind.htm>>. Acesso em: 19 nov. 2010.

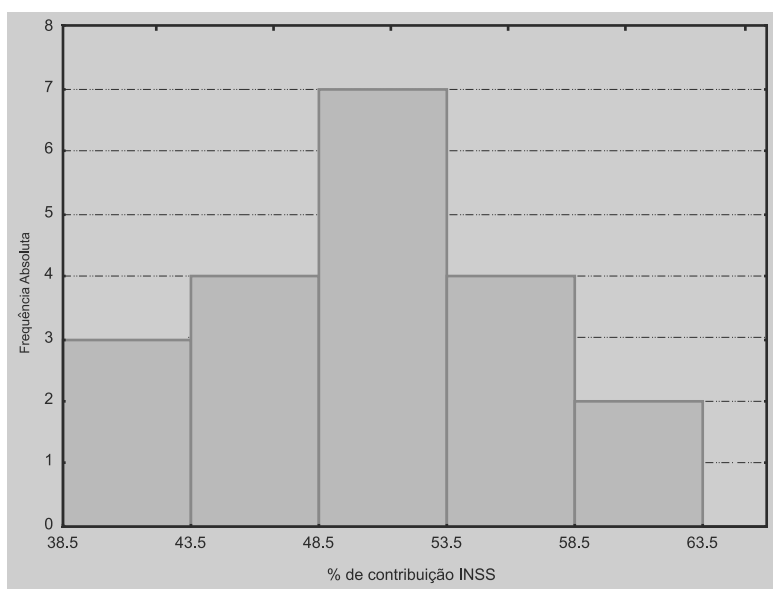


Figura 7: Histograma representativo da distribuição de frequências do percentual dos trabalhadores que contribuíram com o INSS em 2008  
Fonte: Elaborada pelo autor

Quanto ao **polígono de frequência**, você pode obtê-lo pela simples união dos pontos médios dos topos dos retângulos de um histograma. Para completar o polígono é necessário unir as extremidades da linha, que une os pontos representativos das frequências de classe, aos pontos médios das classes imediatamente anteriores e posteriores às classes extremas, que têm frequência nula.

A Figura 8 mostra o polígono de frequências do percentual dos trabalhadores que contribuíram com o INSS em 20 cidades de uma determinada região do Brasil em 2008.



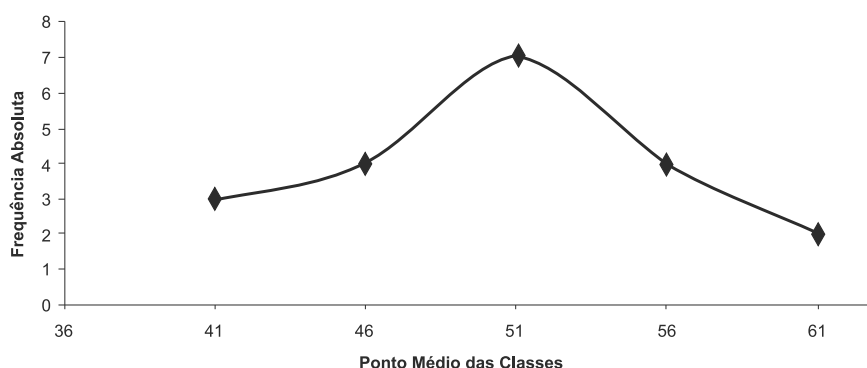


Figura 8: Polígono de frequências do percentual dos trabalhadores que contribuíram com o INSS em 2008

Fonte: Elaborada pelo autor

Quando você tem uma tabela que é trabalhada com uma variável qualitativa, o tipo de gráfico adequado para apresentar os resultados corresponde ao gráfico de setores, também popularmente conhecido como gráfico tipo pizza, como demonstra a Figura 9. Sua construção é simples: sabemos que o ângulo de  $360^\circ$  equivale a 100% da área da circunferência; assim, para obtermos o ângulo do setor cuja área representa uma determinada frequência, basta resolvermos uma regra de três simples, como a apresentada a seguir:

$$\begin{array}{rcl} 360^\circ & \text{-----} & 100\% \\ x^\circ & \text{-----} & \text{Frequência relativa (percentual)} \end{array}$$

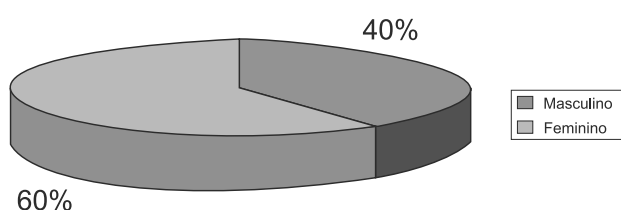


Figura 9: Gráfico do gênero de pessoas que consideram os serviços da prefeitura satisfatórios

Fonte: Elaborada pelo autor

Com respeito aos gráficos chamados de **ogivas**, estes correspondem a um polígono de frequências acumuladas, no qual as frequências acumuladas são localizadas sobre perpendiculares

levantadas nos limites superiores das classes, sendo os pontos unidos para formar o polígono que representa as frequências acumuladas. Observe o modelo apresentado na Figura 10.

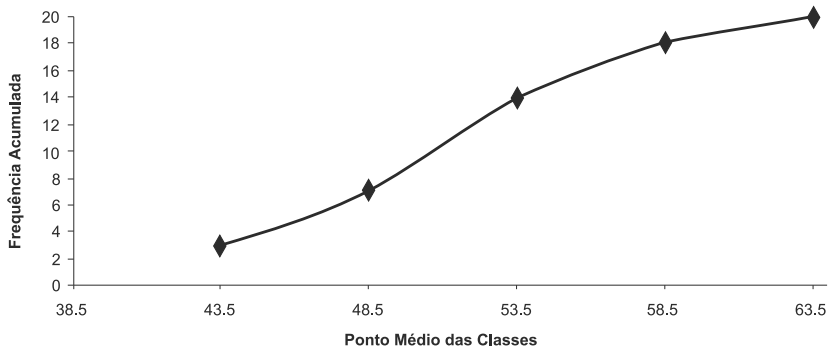


Figura 10: Ogiva “abaixo de” do percentual dos trabalhadores que contribuem com o INSS em 20 cidades de uma determinada região do Brasil em 2008  
Fonte: Elaborada pelo autor

Após o estudo da construção de distribuições de frequências e gráficos, você deve ser capaz de organizar um conjunto de dados, por meio de uma **distribuição de frequências** (absoluta, relativa, e acumuladas), e representá-lo graficamente. Para tanto, propomos a você um exemplo comentado para melhor fixar os conhecimentos adquiridos.

### Exemplo

Uma amostra de valores de IPTU de uma determinada região da cidade de Arapongas, no ano passado, revelou valores iguais a: {68,98; 72,92; 89,19; 98,57; 123,34; 134,80; 141,34; 153,59; 158,59; 165,92; 169,21; 175,76; 177,79; 178,07; 180,38; 181,99; 185,95; 188,83; 194,88; 208,09; 214,66; 251,94; 265,70; 271,90; 276,59; 280,56; 303,99; 318,33}. Com base nos dados fornecidos, vamos construir a tabela de distribuição de frequência.

Bem, para construirmos uma tabela de frequência, primeiro precisamos encontrar: o número de classes, a amplitude total, a amplitude de classe e o limite inferior da primeira classe.

Para você fazer cálculos de distribuições de frequências e gráficos, utilize a planilha Calc do pacote OpenOffice disponível no site: <<http://www2.ufpa.br/dicas/open/oo-ind.htm>>. Acesso em: 10 mar. 2011.

O número de classes é dado por:  $k = \sqrt{n}$ , pois o tamanho da amostra é menor ou igual a 100. Como  $n = 28$ , temos:

$$k = \sqrt{28} \cong 6$$

Nesse caso, aproximamos para seis classes e não para cinco, pois com cinco teremos valores superiores que podem ficar sem classe.

Amplitude total (A) é a diferença entre o maior valor observado e o menor valor observado. Substituindo os valores, encontraremos:

$$A = 318,33 - 68,98 = 249,35$$

Sendo assim, a amplitude de classe será:

$$c = \frac{A}{k-1} \text{ e, substituindo os valores correspondentes, teremos:}$$

$$c = \frac{249,35}{6-1} = 49,87$$

Logo, o limite inferior da primeira classe é dado por:

$$LI_{1^a} = \text{menor valor} - \frac{c}{2}$$

$$LI_{1^a} = 68,98 - \frac{49,87}{2} = 44,04 \text{ (esse é o primeiro valor a ser colocado na tabela).}$$

Agora, a partir desse limite inferior, podemos construir a tabela de distribuição de frequência. Para preencher a coluna classes, começamos com o limite inferior da primeira classe, lembrando que para encontrar o limite superior das classes basta somar a amplitude de classe (c) ao limite inferior. **Agora é com você.** Termine de calcular os limites de cada uma das classes.

$$44,04 + 49,87 = 93,91$$

$$93,91 + 49,87 = 143,78$$

↓

$$293,39 + 49,87 = 343,26$$

Após esse cálculo, vamos encontrar os valores da coluna frequência absoluta ( $F_a$ ) e, para tanto, temos que contar quantos elementos da amostra pertencem a cada classe que acabamos de construir, vamos lá:

- **Primeira classe:** 44,04 (inclusive) a 93,91 (exclusive), então, do conjunto de dados, os valores que pertencem a esse intervalo são: 68,98; 72,92; 89,19; ou seja, três valores.
- **Segunda classe:** 93,91 (inclusive) a 143,78 (exclusive), então, do conjunto de dados, os valores que pertencem a esse intervalo são: 98,57; 123,34; 134,80; 141,34; ou seja, quatro valores.

E, assim, procedemos até encontrarmos as frequências das seis classes. Feita essa operação, é hora de calcularmos a coluna da frequência relativa da classe  $i$  ( $F_{ri}$ ), onde temos:

$$F_{ri} = \frac{F_i}{n}$$

$$F_{r1} = \frac{3}{28} \cong 0,11$$

$$F_{r2} = \frac{4}{28} = 0,14$$

Você deve proceder da mesma forma até a última classe e, após todos os cálculos, deve terminar de completar os valores para a montagem final da distribuição de frequências. Lembre-se de que o preenchimento da coluna frequência acumulada ( $F_{ac}$ ) corresponde à soma da frequência daquela classe às frequências de todas as classes anteriores a ela. Observe a Tabela 9.

Tabela 9: Distribuição de frequências de valores de IPTU de uma determinada região da cidade de Arapongas

CLASSES	$F_A$	$F_{R_i}$	$F_{AC}$
44,04 ┆ 93,91	3	0,11	3
93,91 ┆ 143,78	4	0,14	7
143,78 ┆ 193,65	-	-	-
193,65 ┆ 243,52	-	-	-
243,52 ┆ 293,39	-	-	-
293,39 ┆ 343,26	-	-	-
Total	28	1,0	

Fonte: Elaborada pelo autor

### Exemplo

Imagine que a área de supervisão de atendimento de controle de uma prefeitura verificou a quantidade de materiais que foram rejeitados em quilograma (kg) da fábrica Manda Brasa S.A., que havia vencido uma licitação conforme os resultados apresentados na Tabela 10.

Tabela 10: Frequência dos materiais rejeitados pela fábrica Manda Brasa S.A.

REJEITOS (EM KG)	$F_i$
2 ┆ 8	3
8 ┆ 14	7
14 ┆ 20	18
20 ┆ 26	15
26 ┆ 32	4
32 ┆ 38	3
Total	50

Fonte: Elaborada pelo autor

Com base nos dados, vamos construir o histograma para as frequências apresentadas. Para tanto, basta colocar no eixo x os intervalos de classe e no eixo y as frequências, como mostra a Figura 11.

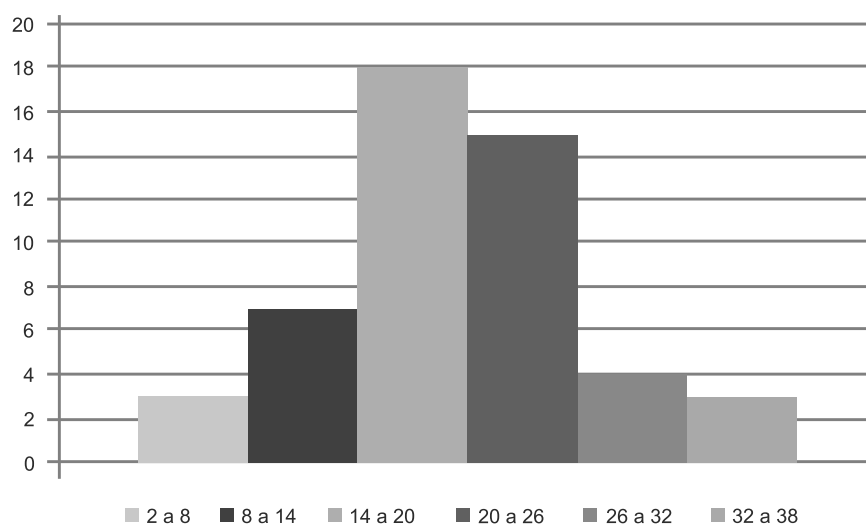


Figura 11: Histograma da frequência de materiais rejeitados na fábrica Manda Brasa S.A.  
Fonte: Elaborada pelo autor

# *Resumindo*



Nesta Unidade, você aprendeu a representar um conjunto de observações e resumi-lo em tabelas e gráficos. Esses conceitos serão importantes na compreensão e no entendimento de um conjunto de dados.



## Atividades de aprendizagem

Agora que você já viu os conceitos relacionados a distribuições de frequências e a representação gráfica de um conjunto de observações, faça a atividade proposta a seguir. Em caso de dúvida, lembre-se de que você tem um tutor pronto a lhe auxiliar.

1. Dado o tempo, em minutos, de reuniões em um setor de uma prefeitura, conforme mostra a tabela, responda as questões a seguir:

60	55	42	57
40	28	44	28
40	30	55	35
25	55	40	38
50	55	40	60

- a) Construa a distribuição de frequências absoluta, relativa e acumuladas.
- b) Faça o histograma e o polígono de frequência da distribuição.



# UNIDADE 3

## MEDIDAS DE POSIÇÃO E DISPERSÃO

### OBJETIVOS ESPECÍFICOS DE APRENDIZAGEM

Ao finalizar esta Unidade, você deverá ser capaz de:

- ▶ Calcular e interpretar as medidas de posição média, moda, mediana;
- ▶ Entender como as medidas de posição influenciam na forma da distribuição dos dados;
- ▶ Calcular e interpretar as medidas de dispersão amplitude total, variância, desvio padrão e coeficiente de variação;
- ▶ Entender as propriedades da média e o desvio padrão; e
- ▶ Calcular e interpretar resultados de medidas separatrizes.



# MEDIDAS DE POSIÇÃO

Caro estudante,

A partir de agora, iremos conhecer uma nova forma de caracterizar um conjunto de observações. Para isso, você irá aprender novos conceitos de medidas de posição e de dispersão.

Para o entendimento dessas medidas de posição e de dispersão, serão utilizadas as duas situações apresentadas a seguir. Sempre que mencionarmos as situações, você deve vir até esta página para entender como estão sendo realizados os cálculos.

Preparado para mais esse desafio?

Então, vamos lá!

Vamos iniciar nossa discussão pelas duas situações que utilizaremos como base.

- Para facilitar um projeto de aplicação da rede de esgoto de certa região de uma cidade, os engenheiros da Prefeitura Municipal tomaram uma amostra de 52 ruas, (tamanho total da amostra ou a soma de todas as frequências absolutas) contando o número de casas por rua. Os dados referentes a uma pesquisa de mercado foram agrupados como segue na Tabela 11:

Tabela 11: Distribuição de frequências do número de casas por rua de certa região de uma cidade

NÚMERO DE CASAS POR RUA	FREQUÊNCIA ABSOLUTA
0  — 2	5
2  — 4	7
4  — 8	11
8  — 12	16
12  — 16	8
16  — 20	5

Fonte: Elaborada pelo autor

- Taxa de efetivação da cobrança de um determinado tributo que se apresentava atrasado em uma prefeitura após uma campanha realizada para que ele fosse saldado. Esses resultados são diários, conforme mostra a Tabela 12.

Tabela 12: Taxa de efetivação da cobrança

44	46	51	54	54	55	56	56	56
58	59	60	61	61	61	62	63	63

Fonte: Elaborada pelo autor

É importante destacarmos ainda que as **medidas de posição** ou de tendência central constituem uma forma mais sintética de apresentar os resultados contidos nos dados observados, pois representam um valor central, em torno do qual os dados se concentram. As medidas de tendência central mais empregadas são a média, a mediana e a moda. A seguir, veremos cada uma delas.

## MÉDIA

Das três medidas de posição mencionadas, a **média aritmética** é a mais usada por ser a mais comum e compreensível delas e pela relativa simplicidade do seu cálculo, além de prestar-se bem ao tratamento algébrico.

Para você fazer cálculos de medidas de posição e de dispersão, utilize o programa estatístico Bioestat 5.0 e, também, planilhas eletrônicas visitando o site: <<http://www.juliobattisti.com.br/tutoriais/celsonunes/openoffice007.asp>>.

Acesso em: 19 nov. 2010.

É importante termos claro que a **média aritmética** ou simplesmente média de um conjunto de  $n$  observações,  $x_1, x_2, \dots, x_n$ , é definida por:

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

Onde o somatório ( $\Sigma$ ) corresponde à soma de todos os valores obtidos. Por exemplo, considerando o caso da taxa de efetivação (%) da cobrança de um determinado tributo que está atrasado em uma prefeitura (ver Tabela 12), se você somar todos os valores do número das taxas e dividi-los pelo total de dias avaliados, você terá, então, a **média aritmética** ( $\bar{x}$ ), a taxa de efetivações de cobrança por dia. Logo, o valor obtido será:  $\bar{x} = 56,67\%$ .

*Como podemos, então, fazer a interpretação da média?*

Podemos interpretar o resultado da média como sendo o número de efetivações diárias que é de 56,67%, podendo ocorrer taxas maiores, menores ou até iguais ao valor médio encontrado.

Portanto, de uma forma mais geral, podemos interpretar a média como sendo um valor típico do conjunto de dados que pode assumir um valor que não pertence ao conjunto de dados, pois como nos dados utilizados para cálculo (exemplo anterior) não existe um taxa de efetivação diária de 56,67%.

Todavia, se os dados estiverem agrupados na forma de uma distribuição de frequência em classes, lança-se mão da **Hipótese Tabular Básica\*** para o cálculo da média.

Então, você irá calcular a média por meio da seguinte expressão:

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

\*Hipótese Tabular Básica – todas as observações contidas em uma classe são consideradas iguais ao ponto médio da classe. Fonte: Elaborado pelo autor.

Onde:

$x_i$  é o ponto médio da classe  $i$ ;

$fa_i$  representa frequência absoluta da classe  $i$ ; e

$fr_i$  é a frequência relativa da classe  $i$ .

Considerando a situação do número de casas na rua (Tabela 11), a média será dada por:

$$\bar{x} = \frac{\sum_{i=1}^n x_i fa_i}{\sum_{i=1}^n fa_i} = \frac{(1 \times 5) + (3 \times 7) + \dots + (18 \times 5)}{5 + 7 + \dots + 5} = 8,73 \text{ casas}$$

O valor de 1, apresentado na expressão, corresponde ao ponto médio da primeira classe, que foi obtido pela soma dos limites superior e inferior ( $0 + 2$ ) divididos por dois, ou seja, a média aritmética. Os pontos médios das outras classes são obtidos de forma similar.

Antes de darmos continuidade, é muito importante você saber que, em relação à notação matemática, quando calculamos a média a partir dos dados de uma população, devemos utilizar a letra  $\mu$  para designar a média populacional e para média amostral a notação a ser utilizada é  $\bar{x}$ . Na grande maioria dos casos, iremos trabalhar com amostras. A forma de cálculo é a mesma nas duas situações, mas as notações são diferentes, ou seja:

Média populacional  $\rightarrow \mu$

Média amostral  $\rightarrow \bar{x}$

As médias são comumente utilizadas e apresentam propriedades específicas. As principais propriedades são:

- A soma dos **desvios\*** de um conjunto de dados em relação a sua média é nula, ou seja, igual a zero. Para entender essa propriedade, tomemos como exemplo a quantidade consumida de arroz do tipo A em um refeitório de uma prefeitura: 10, 14, 13, 15, 16, 18,

\*Desvios – diferenças entre cada valor e um valor padrão, que pode ser a média. Fonte: Elaborado pelo autor.

12 quilos, no qual o consumo médio diário encontrado foi de 14 quilogramas (Kg).

A soma desvios será:

$$(10 - 14) + (14 - 14) + (13 - 14) + (15 - 14) + (16 - 14) + (18 - 14) + (12 - 14) = 0$$

- A soma ou a subtração de uma constante (c) a todos os valores de uma variável, a média do conjunto fica aumentada ou diminuída dessa constante. Assim, voltando ao caso do consumo de arroz, apresentado no tópico anterior, se somarmos 2 a cada um dos valores, teremos:

$$Y = (12 + 16 + 15 + 17 + 18 + 20 + 14) / 7 = 16 \text{ kg ou} \\ Y = 14 + 2 = 16 \text{ kg}$$

- Na multiplicação ou na divisão de todos os valores de uma variável por uma constante (c), a média do conjunto fica multiplicada ou dividida por essa constante. Novamente pensando no caso do consumo de arroz, se multiplicarmos 3 a cada um dos valores, teremos:

$$Y = (30 + 42 + 39 + 45 + 48 + 54 + 36) / 7 = 42 \text{ kg ou} \\ Y = 14 \cdot 3 = 42 \text{ kg}$$

Existem outros tipos de médias que podemos utilizar, por exemplo, média ponderada (utilizada quando existe algum fator de ponderação); média geométrica (quando os dados apresentam uma distribuição que não é simétrica); entre outras.

Às vezes, podemos, ainda, associar às observações  $X_1, X_2, \dots, X_n$  determinadas ponderações, ou pesos,  $W_1, W_2, \dots, W_n$  que dependem da importância atribuída a cada uma das observações, nesse caso, a média ponderada será dada por:

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

Para entender melhor, imagine um processo de avaliação de funcionários públicos que foi dividido em três etapas. Nessa avaliação, suponha que um dos colaboradores apresentou as seguintes notas durante a avaliação: 1ª etapa = 90; 2ª etapa = 70; 3ª etapa = 85; e os pesos de cada etapa são: 1, 1 e 3, respectivamente. Qual o escore médio final do funcionário público?

$$\bar{x} = \frac{\sum_{i=1}^n x_i w_i}{\sum_{i=1}^n w_i} = \frac{(1 \times 70) + (1 \times 90) + (3 \times 85)}{1 + 1 + 3} = \frac{415}{5} = 83$$

Este tipo de média você irá utilizar na disciplina *Matemática Financeira* que trabalharemos no próximo módulo.

Outro tipo de média corresponde à geométrica (Mg), calculada pela raiz n-ésima do produto de um conjunto de n observações,  $X_1, X_2, \dots, X_n$ , associadas às frequências absolutas  $f_1, f_2, \dots, f_n$  (número de vezes que aquele valor acontece) e respectivamente dada por:

$$Mg = \sqrt[n]{x_1^{f_1} \times x_2^{f_2} \times \dots \times x_n^{f_n}}$$

Sendo assim, considerando o caso da taxa de efetivação para pagamento do tributo atrasado (exemplo apresentado anteriormente), teremos:

$$Mg = \sqrt[18]{44^1 \times 46^1 \times 51^1 \times 54^2 \times \dots \times 61^3 \times 62^1 \times 63^2} = 56,40\%$$

## MODA

Em algumas situações, você verá que é necessária a informação do número de observações que mais ocorre em um conjunto de dados. No caso da taxa de efetivação da cobrança, verificamos que a taxa que mais ocorre é 56 e 61. Assim, podemos definir a **moda (Mo) como sendo o valor em um conjunto**



**de dados que ocorre com maior frequência.** Um conjunto de dados pode ser em relação à moda:

- ▶ unimodal → possui apenas uma moda;
- ▶ amodal → não possui moda, pois não existe nenhum valor que ocorre com maior frequência; e
- ▶ multimodal → possui mais de uma moda.

Na situação comentada anteriormente, a distribuição é multimodal ou bimodal, pois apresenta duas modas, ou seja, dois valores com maior frequência, 56 e 61.

Quando os dados não estão em intervalos de classes, basta olhar o valor que ocorre com maior frequência.

Para dados agrupados em intervalos de classes, você pode calcular a moda por meio do método de Czuber, que se baseia na influência das classes adjacente na moda deslocando-se no sentido da classe de maior frequência. A expressão que você utilizará é:

$$Mo = L_i + \frac{d_1}{d_1 + d_2} \times c$$

Onde:

$L_i$  : limite inferior da classe modal;

$d_1$  : diferença entre a frequência da classe modal e a imediatamente anterior;

$d_2$  : diferença entre a frequência da classe modal e a imediatamente posterior; e

$c$  : amplitude da classe modal.

No caso em que, para facilitar um projeto de aplicação da rede de esgoto de certa região de uma cidade, os engenheiros da Prefeitura Municipal tomaram uma amostra de 52 ruas, contando o número de casas (Tabela 11), teremos que a classe modal é a **quarta**, pois apresenta maior frequência (valor igual a 16). Utilizando a expressão mostrada anteriormente, teremos:

$$Mo = L_i + \frac{d_1}{d_1 + d_2} \times c = 8 + \frac{5}{5 + 8} \times 4 = 9,54 \text{ casas}$$

Uma característica importante da moda é que ela não é afetada pelos valores extremos da distribuição, desde que esses valores não constituam a classe modal.

Dessa forma, a moda deve ser utilizada quando desejamos obter uma medida rápida e aproximada de posição ou quando a medida deva ser o valor mais frequente da distribuição.

## MEDIANA

Outra medida de posição que você pode utilizar é a **mediana (Md)**, que consiste em um conjunto de valores dispostos segundo uma ordem (crescente ou decrescente). A mediana é o valor situado de tal forma no conjunto ordenado que o separa em dois subconjuntos de mesmo número de elementos, ou seja, 50% dos dados são superiores à mediana e 50% são inferiores.

O símbolo da mediana é dado por Md ou  $\tilde{x}$ , e a sua posição é dada por meio da expressão:

$$E (\text{elemento central}) = (n+1) / 2$$

Considerando um conjunto de dados com número ímpar de elementos (1, 2, 5, 9, 10, 12, 13), a posição da mediana será dada por  $(7 + 1)/2 = 4^{\text{a}}$  posição. Portanto, a partir dos dados ordenados, o número que se encontra na  $4^{\text{a}}$  posição é o 9 e, assim, a mediana será igual a 9 (temos três valores abaixo e três valores acima, ou 50% acima da mediana e 50% abaixo).

E, caso o número de elementos do conjunto de dados seja par, por exemplo, (1, 2, 6, 8, 9, 12, 11, 13) a posição da mediana será:

$$E = (8 + 1)/2 = 4,5^{\text{a}} \text{ posição}$$

Como a posição 4,5 está entre a 4ª e a 5ª posição, calculamos a média entre os valores que ocupam essas posições.

O valor encontrado de 8,5, (vem de  $(8 + 9) / 2$ ), corresponde à mediana.

Quando os dados estão agrupados na mediana, devemos encontrar a classe mediana. Se os dados estão agrupados em intervalos de classe, como no caso do número de casa por rua, utilizaremos a seguinte expressão:

$$Md = li + \left( \frac{(n / 2) - f_{antac}}{f_{med}} \right) \times c$$

Onde:

li : limite inferior da classe mediana;

n : número total de elementos;

$f_{antac}$  : frequência acumulada anterior à classe mediana;

$f_{med}$  : frequência absoluta da classe mediana; e

c: amplitude da classe mediana.

Portanto, resolvendo o caso em que, para facilitar um projeto de aplicação da rede de esgoto de certa região de uma cidade, os engenheiros da Prefeitura Municipal tomaram uma amostra de 52 ruas, contando o número de casas por rua; você verá que a posição da mediana será dada por:

$E = (52 + 1) / 2 = 26,5^\circ$  elemento, o qual está na quarta classe  $(8 \text{ a } 12)$ , que corresponde à classe mediana.

$$Md = li + \left( \frac{(n / 2) - f_{antac}}{f_{med}} \right) \times c = 8 + \left( \frac{(52 / 2) - 23}{16} \right) \times 4 = 8,75 \text{ casas}$$

Em um conjunto de dados, a mediana, a moda e a média não necessariamente devem apresentar o mesmo valor. Uma informação importante é que a mediana não é influenciada pelos valores extremos. Comparando os resultados encontrados para uma amostra em relação às medidas de posição estudadas e verificando a inter-relação entre elas, você pode concluir que seus valores podem

nos dar um indicativo da natureza da distribuição dos dados, em função das regras definidas pela Figura 12:

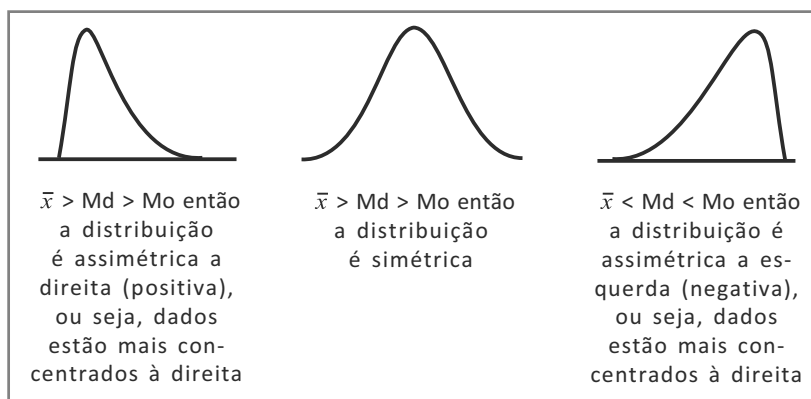


Figura 12: Natureza de distribuição de dados

Fonte: Elaborada pelo autor

## SEPARATRIZES

A principal característica das medidas separatrizes consiste na separação da série de dados ordenados em partes iguais que apresentam o mesmo número de valores. As principais são os quartis, os decis e os percentis.

Os **quartis** são valores que dividem um conjunto de dados ordenados em quatro partes iguais. São necessários, portanto, três quartis ( $Q_1$ ,  $Q_2$  e  $Q_3$ ) para dividir um conjunto de dados ordenados em quatro partes iguais.

$Q_1$  :deixa 25% dos elementos abaixo dele.

$Q_2$  :deixa 50% dos elementos abaixo dele e coincide com a mediana.

$Q_3$  :deixa 75% dos elementos abaixo dele.

A Figura 13 mostra bem a divisão dos quartis, observe.

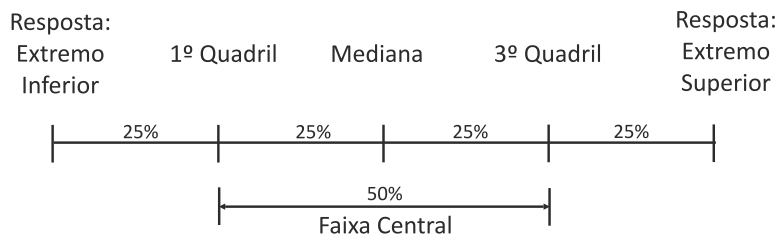


Figura 13: Representação dos quartis

Fonte: Elaborada pelo autor

Se considerarmos a situação da taxa de efetivação da cobrança de um determinado tributo, que estava atrasado em uma prefeitura, após uma campanha realizada para que ele fosse saldado, teremos, de forma semelhante à Figura 13, a Figura 14:

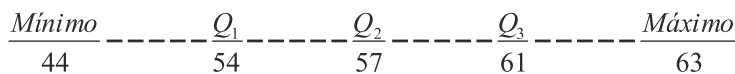


Figura 14: Quartis da taxa de efetivação da cobrança de um determinado tributo

Fonte: Elaborada pelo autor

Sendo assim, temos o cálculo da posição do elemento quartil dado por:

$$EQ_i = \text{in}/4 \quad (i = 1, 2, 3)$$

A regra para obtenção dos valores dos quartis, a partir da posição encontrada, será dada por:

- quando  $n$  é ímpar, o arredondamento deve ser para cima da posição encontrada; e
- quando  $n$  é par, devemos fazer a média do valor encontrado e do subsequente.

Para melhor entendimento, elaboramos um exemplo para realizarmos juntos. Para tanto, considere a seguinte sequência de números para cálculo dos quartis: (5, 2, 6, 9, 10, 13, 15).

Agora, precisamos ordenar o conjunto de dados e, então, temos: (2, 5, 6, 9, 10, 13, 15).

Sendo assim, obtemos a posição e, olhando no conjunto ordenado de dados, encontramos os valores dos quartis, conforme você pode observar a seguir.

$$EQ1 = 1.7/4 = 1,75 \cong 2^a \text{ posição} \Rightarrow Q1 = 5$$

$$EQ2 = 2.7/4 = 3,5 \cong 4^a \text{ posição} \Rightarrow Q2 = 9$$

$$EQ3 = 3.7/4 = 5,25 \cong 6^a \text{ posição} \Rightarrow Q3 = 13$$

Agora, vamos a outro exemplo, para tanto, considere um conjunto de dados com uma quantidade par de observações, a saber: (1, 1, 2, 3, 5, 5, 6, 7, 9, 9, 10, 13)  $\Rightarrow$  já ordenados. Então, temos:

$$EQ1 = 1.12/4 = 3^a \text{ posição} \Rightarrow Q1 = (2 + 3) / 2 = 2,5$$

$$EQ2 = 2.12/4 = 6^a \text{ posição} \Rightarrow Q2 = (5 + 6) / 2 = 5,5$$

$$EQ3 = 3.12/4 = 9^a \text{ posição} \Rightarrow Q3 = (9 + 9) / 2 = 9$$

Os **decis** são valores que dividem um conjunto de dados ordenados em dez partes iguais.

O cálculo de cada decil será obtido de forma semelhante aos quartis, sendo diferente apenas a expressão de sua obtenção, que será dada por:

Posição do elemento decil  $\rightarrow EDi = in/10 \quad (i = 1, 2, \dots, 9)$

Os **percentis** são valores que dividem um conjunto de dados ordenados em 100 partes iguais.

A posição de cada percentil será dada pela expressão a seguir que é semelhante aos quartis e aos decis:

Posição do elemento percentil  $\rightarrow EPi = in/100 \quad (i = 1, 2, \dots, 99)$

Essas medidas separatrizes são importantes quando queremos dividir um conjunto de dados em parte iguais; por exemplo, em quatro partes; e, assim, você terá os quartis. Essa separação permite uma formação de grupos que podem apresentar um mesmo padrão, quando, então, poderemos identificar perfis importantes para serem utilizados em diversas áreas da Administração.

## MEDIDAS DE DISPERSÃO

Como vimos anteriormente, é possível sintetizar um conjunto de observações em alguns valores representativos, como média, mediana, moda e separatrizes. Em várias situações, é necessário visualizar como os dados estão dispersos.

*Tomando como exemplo algumas funções da área de Administração Pública que apresentem salários médios iguais, podemos concluir que sua contribuição social (% do salário) será a mesma?*

A resposta é sim somente com base no salário médio; mas estaríamos chegando a uma conclusão errada, pois a variação em termos de faixas salariais pode ser diferente, apesar de apresentarem a mesma média.

Suponhamos três cidades: A, B e C, que foram avaliadas durante cinco anos quanto ao número de declarantes na distribuição de patrimônio na faixa de renda mensal de 8 a 10 mil reais. Esses valores estão em milhares de pessoas.

$$A = \{120, 122, 118, 124, 121\}$$

$$B = \{121, 121, 121, 121, 121\}$$

$$C = \{116, 125, 124, 120, 120\}$$

Se nós calcularmos a média de cada cidade, teremos:

$$A \rightarrow \bar{x} = 121 \text{ mil pessoas}$$

$$B \rightarrow \bar{x} = 121 \text{ mil pessoas}$$

$$C \rightarrow \bar{x} = 121 \text{ mil pessoas}$$

Note que as três cidades (A, B, C) apresentam médias iguais, apesar de elas serem bem diferentes entre si, pois enquanto na cidade B os dados são todos iguais, os das demais cidades apresentam certa variação, que é maior no conjunto C. Portanto, devemos associar medidas de posição e de dispersão para obtermos informações mais precisas de um conjunto de dados, ou seja, observar como esses dados se comportam em torno da medida de posição em questão.

## AMPLITUDE TOTAL

A amplitude total é a diferença entre o maior e o menor valor observado, como vimos na Unidade 2.

Sendo assim, retomando nossos exemplos das cidades A, B e C, temos:

$$A_A = 124 - 118 = 6 \text{ mil pessoas}$$

$$A_B = 121 - 121 = 0 \text{ mil pessoas}$$

$$A_C = 125 - 116 = 9 \text{ mil pessoas}$$

Desse modo, podemos identificar que a amplitude do conjunto C é bem maior do que nos demais e o conjunto B apresenta amplitude igual a zero.

Essa medida apresenta a vantagem de ser facilmente calculada. Entretanto, o seu inconveniente é que ela é muito afetada pelos valores extremos, pois no seu cálculo não são consideradas todas as observações.



## VARIÂNCIA

Uma boa medida de dispersão deve ter as seguintes características:

- ▶ estar baseada em todos os dados;
- ▶ ser facilmente calculada;
- ▶ ser compreensível; e
- ▶ servir bem ao tratamento algébrico.

Portanto, podemos afirmar que uma medida de dispersão deve utilizar todas as observações considerando os desvios de cada observação em relação à média (chamados erros):

$$e_i = x_i - \bar{x}$$

Para obter um único número que represente a dispersão dos dados, pensamos, inicialmente, em obter a média desses desvios, mas devemos lembrar de que a soma dos desvios de um conjunto de dados em relação a sua média é nula.

Para resolver esse problema, utilizamos a soma dos quadrados dos desvios, pois, ao elevarmos cada desvio ao quadrado, eliminamos o sinal negativo que estava trazendo complicações.

Posteriormente, dividimos a soma dos quadrados dos desvios pelo número de observações para obtermos a variância populacional, chamada de  $\sigma^2$ , que é uma medida quantitativa da dispersão de um conjunto de dados entorno da sua média, além do fato de essa soma de quadrados de desvios ser mínima.

Sendo assim, temos a expressão para cálculo da variância populacional, conforme mostrada a seguir:

$$V(x) = \sigma^2 = \frac{SQD}{N} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

E não para por aí! Na maioria das vezes, trabalhamos com amostras e, nesse caso, a variância amostral ( $s^2$ ) será obtida pela expressão:

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

Veja que nesse caso a soma do quadrado dos desvios é dividida por  $n - 1$ , onde  $n$  corresponde ao tamanho da amostra. Esse valor  $n - 1$  (número de observações menos um) é denominado de **grau de liberdade\***.

Então, o grau de liberdade é um estimador do número de categorias independentes em um teste particular ou experiência estatística. Assim, no caso das cidades teremos:

$$s_A^2 = \frac{(120-121)^2 + (122-121)^2 + \dots + (121-121)^2}{4} = 5 \text{ mil pessoas}^2$$

$$s_B^2 = \frac{(121-121)^2 + (121-121)^2 + \dots + (121-121)^2}{4} = 0 \text{ mil pessoas}^2$$

$$s_C^2 = \frac{(116-121)^2 + (125-121)^2 + \dots + (120-121)^2}{4} = 13 \text{ mil pessoas}^2$$

Para que você entenda melhor, veja a seguir algumas das principais propriedades da variância:

- A variância de uma constante  $k$  é nula.  
 $V(k) = 0$ ,  $k = \text{constante}$ .
- Ao somar ou ao subtrair uma constante  $k$  a todos os dados, a variância não se altera.  
 $x' = x \pm k$   
 $V(x') = V(x)$
- Multiplicando todos os dados por uma constante  $k$ , a variância é multiplicada por  $k^2$ .  
 $x' = x \cdot k$   
 $V(x') = k^2 \cdot V(x)$

\*Grau de liberdade – é o número de determinações independentes (dimensão da amostra) menos o número de parâmetros estatísticos a serem avaliados na população. Fonte: Elaborado pelo autor.

## DESVIO PADRÃO

Um inconveniente da variância é que ela é expressa em unidades ao quadrado, ou seja, caso esteja trabalhando com milhares de reais, o resultado será expresso em milhares de reais<sup>2</sup>, o que causa algumas dificuldades de interpretação.

Para resolver esse problema, podemos nos utilizar do desvio padrão que é definido como a raiz quadrada positiva da variância, sendo expresso na mesma unidade em que os dados foram coletados.

$$\sigma = \sqrt{\sigma^2} \quad (\text{desvio padrão populacional})$$

$$s = \sqrt{s^2} \quad (\text{desvio padrão amostral})$$

Para o exemplo em questão, temos:

$$s_A^2 = \sqrt{\frac{(120-121)^2 + (122-121)^2 + \dots + (121-121)^2}{4}} = 2,24 \text{ mil pessoas}$$

$$s_B^2 = \sqrt{\frac{(121-121)^2 + (121+21)^2 + \dots + (121-121)^2}{4}} = 0 \text{ mil pessoas}$$

$$s_C^2 = \sqrt{\frac{(116-121)^2 + (125-121)^2 + \dots + (120-121)^2}{4}} = 3,60 \text{ mil pessoas}$$

Interpretando, temos que: o desvio padrão de 3,60 mil pessoas nos indica a variação dos dados em torno da média, que é de 121 mil pessoas. Quanto menor for o desvio padrão, menor será a variabilidade, ou a variação.

No caso de dados agrupados em classes, a expressão utilizada para cálculo do desvio padrão será:

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

Para entender melhor, vamos imaginar uma situação em que, para facilitar um projeto de aplicação da rede de esgoto de certa região de uma cidade, os engenheiros da Prefeitura Municipal

tomaram uma amostra de 52 ruas (Tabela 11), contando o número de casas por rua, na qual os dados estão agrupados em classes, iremos calcular o desvio padrão da seguinte maneira:

$$s^2 = \sqrt{\frac{1}{52-1} ((1-8,73)^2 \cdot 5 + (3-8,73)^2 \cdot 7 + \dots + (18-8,73)^2 \cdot 5)} = 5,075 \text{ casas}$$

Com base nessa resolução, os números 1, 3 e 18 correspondem aos pontos médios das classes primeira, segunda e última, respectivamente. Já os valores de 5 e 7 correspondem às frequências absolutas das classes. E o número 52 corresponde ao tamanho da amostra.

Existem algumas propriedades que precisamos saber sobre desvio padrão. São elas:

- ▶ Ao somar ou ao subtrair uma constante  $k$  a todos os dados, o desvio padrão não se altera.

$$x' = x \pm k$$

$$\sigma(x') = \sigma(x)$$

- ▶ Multiplicando todos os dados por uma constante  $k$ , o desvio padrão fica multiplicado por  $k$

$$x' = x \cdot k$$

$$\sigma(x') = k \cdot \sigma(x)$$

## COEFICIENTE DE VARIAÇÃO

A variância e o desvio padrão são medidas de dispersão absolutas, desse modo, apenas podem ser utilizados para comparar a variabilidade de dois ou mais conjuntos de dados quando estes apresentarem:

- ▶ mesma média;
- ▶ mesmo número de observações; e
- ▶ estiverem expressos nas mesmas unidades.

Então, para você comparar qualquer conjunto de dados quanto à sua variabilidade quando, pelo menos, uma dessas condições não é satisfeita, é necessário lançar mão de uma medida de dispersão relativa como o **coeficiente de variação** (CV), que expressa a variabilidade dos dados em relação a sua média de forma percentual. Sua expressão será dada por:

$$CV = \frac{S}{\bar{x}} \cdot 100$$

*Para melhor entendimento, vamos elaborar um exemplo para você.*

### Exemplo

Imagine uma situação referente ao número de documentos falsificados que aparecem em um determinado setor da prefeitura e o valor arrecadado por hora de um tipo de multa em reais. Em qual das duas variáveis ocorre maior variabilidade, ou variação?

	DOCUMENTOS FALSIFICADOS (Nº)	MULTA (REAIS)
Média	22	800
Desvio padrão	5	100

Utilizando o desvio padrão para comparar a variabilidade, você pode, a princípio, considerar que a multa apresenta maior variabilidade, já que tem maior desvio padrão. Entretanto, se verificarmos as condições de se utilizar o desvio padrão para comparar a variabilidade entre amostras, você vai perceber que as médias são diferentes e as unidades também são diferentes.

Calculando, então, o coeficiente de variação, teremos os valores apresentados, a seguir:

$$CV_{DOC} = \frac{S}{\bar{x}} \cdot 100 = \frac{5}{22} \cdot 100 = 22,7\%$$

$$CV_{MULTA} = \frac{S}{\bar{x}} \cdot 100 = \frac{100}{800} \cdot 100 = 12,5\%$$

Perceba, então, que estávamos concluindo erroneamente que a multa é mais variável do que o número de documentos falsificados, além de termos cometido o disparate de comparar numericamente duas variáveis expressas em unidades diferentes.

Portanto, o número de documentos falsificados apresentou maior dispersão do que a multa, já que seu coeficiente de variação foi maior, mudando assim a conclusão anterior.

Vamos ver agora outros exemplos de situações com a resolução comentada para você fixar melhor os conceitos desta Unidade.

### Exemplo 1

Considere as idades dos funcionários do programa *Jovens que aprendem uma profissão* de duas prefeituras, apresentadas a seguir.

Prefeitura A: {16; 15; 18; 15; 16; 16; 17; 18; 19; 17; 16}

Prefeitura B: {15; 17; 19; 19; 17; 18; 19; 18; 18; 17; 16}

Encontre a média, moda e mediana de cada prefeitura e identifique qual das prefeituras apresenta maior variabilidade na idade de seus jovens aprendizes.

#### Prefeitura A

► Média:  $\bar{x} = \frac{\sum x_i}{n} = \frac{16+15+...+16}{11} = 16,64$

► Mediana: Md = 16, lembrando que, para encontrar a mediana, necessariamente os dados devem estar ordenados.

- Moda:  $Mo = 16$ , valor que aparece com maior frequência.

### **Prefeitura B**

- Média:  $\bar{x} = \frac{\sum x_i}{n} = \frac{15+17+\dots+16}{11} = 17,54$
- Mediana:  $Md = 18$ , lembrando que, para encontrar a mediana, necessariamente os dados devem estar ordenados.
- Moda:  $Mo = 17, 18$  e  $19$  (distribuição multimodal, pois apresenta mais de duas modas).

Para sabermos quem tem maior variabilidade, temos de calcular o coeficiente de variação, pois, como o valor das médias são diferentes, não podemos usar o desvio padrão para comparar a variabilidade. Para encontrarmos o desvio padrão, precisamos primeiramente encontrar a variância usando a seguinte fórmula:

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

### **Prefeitura A**

- Variância:
 
$$s^2 = \frac{(16-16,64)^2 + (15-16,64)^2 + \dots + (16-16,64)^2}{11-1} = 1,654$$
- Desvio padrão:  $s = \sqrt{1,654} = 1,2862$
- Coeficiente de variação:  $CV = \frac{s}{\bar{x}} \cdot 100 = \frac{1,2862}{16,64} \cdot 100 = 7,7\%$

### **Prefeitura B**

- Variância:
 
$$s^2 = \frac{(15-17,54)^2 + (17-17,54)^2 + \dots + (16-17,54)^2}{11-1} = 1,6726$$
- Desvio padrão:  $s = \sqrt{1,6726} = 1,2933$
- Coeficiente de variação:  $CV = \frac{s}{\bar{x}} \cdot 100 = \frac{1,2933}{17,54} \cdot 100 = 7,3\%$

Sendo assim, como os coeficientes apresentam valores muito próximos, podemos concluir que a variabilidade na idade das duas prefeituras é praticamente a mesma.

**Exemplo 2**

Considerando os dados apresentados a seguir, que são referentes ao percentual de gastos com planejamento e com administração em cidades de diferentes portes, identifique as medidas de posição e de dispersão dos dados.

Gasto	Frequência (F <sub>i</sub> )
5 – 15	2
15 – 25	7
25 – 35	20
35 – 45	5
45 – 55	4
55 – 65	2
Soma	40

Primeiramente, temos de encontrar os valores de x<sub>i</sub> (ponto médio), pois ele é indispensável no cálculo da média, variância etc. Logo, temos:

X<sub>i</sub> = 10; 20; 30; 40; 50; 60 (soma: limite inferior + limite superior dividido por 2).

Feita essa conta, vamos calcular a frequência acumulada, como você pode acompanhar a seguir:

$F_{ac} = 2; 9; 29; 34; 38; 40$

E, na sequência, com os valores do ponto médio, podemos calcular a média:

$$\bar{x} = \frac{\sum x_i \times f_i}{\sum fa} = \frac{10.2 + 20.7 + 30.20 + \dots + 60.2}{40} = 32$$



Para encontrar a mediana, primeiramente temos de encontrar a classe mediana. Como  $n$  é par:  $x_{n/2} = x_{40/2} = x_{20}$ , a qual classe pertence o elemento de posição 20º (3ª classe)?

$$Md = Li + \left( \frac{\frac{n}{2} - f_{antac}}{f_{md}} \right) \cdot c = 25 + \left( \frac{\frac{40}{2} - 9}{20} \right) \cdot 10 = 30,5$$

Vamos, agora, calcular a moda e, para tanto, precisamos encontrar a **classe modal**, aquela com maior frequência absoluta (3ª classe).

$$Mo = LI_{mo} + \left( \frac{d_1}{d_1 + d_2} \right) \cdot c = 25 + \left( \frac{13}{13 + 15} \right) \cdot 10 = 29,6$$

E, por fim, devemos fazer o cálculo das medidas de dispersão, como você pode acompanhar a seguir:

$$S^2 = \frac{\sum (x_i - \bar{x})^2 \times f_i}{n - 1} = \frac{(10 - 32)^2 \times 2 + \dots + (60 - 32)^2 \times 2}{40 - 1} = \frac{5240}{39} = 134,3590$$

$$S = \sqrt{S^2} = \sqrt{134,3590} = 11,5913$$

$$CV = \frac{S}{\bar{x}} \times 100 = 36,22\%$$

Observe que, com as medidas de dispersão calculadas, podemos verificar que a dispersão obtida foi média (36,22% em torno da média), ou seja, tanto para cima quanto para baixo. Se esse valor fosse bem menor, poderíamos considerar que os gastos com planejamento e com transportes seriam mais uniformes.

### Exemplo 3

Considerando as séries de dados apresentadas pelos gastos com transportes em relação ao total gasto em várias prefeituras, conforme descrição a seguir, faça o seguinte: imagine que você precise efetuar uma estimativa com base nesses dados. Sobre qual série é mais fácil fazer estimativas precisas? Por quê?

Fique atento, pois as classes mediana e modal não necessariamente vão pertencer a mesma classe.

Série A: {3,96; 3,17; 3,55; 3,61; 4,11; 4,57; 4,97; 5,91; 5,99; 5,74}

Série B: {1,46; 2,09; 3,04; 5,12; 7,80; 8,25; 9,95; 15,24; 17,40; 21,74}

### Série A

► Média:  $\bar{x} = \frac{\sum x_i}{n} = \frac{3,96 + 3,17 + \dots + 5,74}{10} = 4,558$

► Variância:

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{(3,96 - 4,558)^2 + \dots + (5,74 - 4,558)^2}{10 - 1} = 1,0939$$

► Desvio padrão:  $S = \sqrt{S^2} = \sqrt{1,0939} = 1,0459$

► Coeficiente de variabilidade:  $CV = \frac{S}{\bar{x}} \times 100 = 22,9\%$

### Série B

► Média:  $\bar{x} = \frac{\sum x_i}{n} = \frac{1,46 + 2,09 + \dots + 21,74}{10} = 9,206$

► Variância:

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{(1,46 - 9,206)^2 + \dots + (21,74 - 9,206)^2}{10 - 1} = 47,748$$

► Desvio padrão:  $S = \sqrt{S^2} = \sqrt{47,748} = 6,91$

► Coeficiente de variabilidade:  $CV = \frac{S}{\bar{x}} \times 100 = 75\%$

Observe que na série A é mais fácil fazermos estimativas precisas, pois ela apresenta menor dispersão.

# *Resumindo*



Nesta Unidade, você aprendeu conceitos básicos sobre as medidas de posição e de dispersão e, desse modo, sabe, agora, caracterizar um conjunto de observações. Esses conceitos são de extrema importância para as inferências estatísticas, para os testes de hipóteses e para as informações contidas nas Unidades posteriores dessa disciplina.



# Atividades de aprendizagem

Agora que você já sabe como calcular e como utilizar as principais medidas de posição e de dispersão, exercite-as fazendo as atividades, a seguir, que serão importantes na consolidação dos conhecimentos adquiridos. Em caso de dúvida, lembre-se de consultar seu tutor por meio do AVEA.

1. Considere a sequência numérica apresentada, a seguir, que mostra as idades de motociclistas e de seus caronas na época em que morreram em acidentes fatais de trânsito.

7	38	27	14	18	34	16
42	28	24	40	20	23	31
37	21	30	25	17	28	33
25	23	19	51	18	29	

Calcule a média moda, a mediana, a variância, o desvio padrão e o coeficiente de variabilidade para os dados não agrupados.

2. Imagine um determinado setor de uma prefeitura que vem apresentando problemas com o afastamento de funcionários por motivos de saúde, por período muito longo. Uma amostra de dez apresentou os seguintes números de dias afastados em um semestre:

23, 21, 10, 14, 16, 12, 39, 45, 10 e 20

Calcule as medidas de posição e de dispersão em relação ao número de dias em que eles ficaram afastados.

# UNIDADE 4

## PROBABILIDADE

### OBJETIVOS ESPECÍFICOS DE APRENDIZAGEM

Ao finalizar esta Unidade, você deverá ser capaz de:

- ▶ Definir o termo probabilidade;
- ▶ Descrever as abordagens clássicas das frequências relativa e subjetiva da probabilidade;
- ▶ Entender os termos experimento, espaço amostral e evento;
- ▶ Definir os termos probabilidade condicional e probabilidade conjunta; e
- ▶ Calcular probabilidades aplicando as regras da adição e da multiplicação.



## INTRODUÇÃO

Caro estudante,  
Vamos iniciar mais uma Unidade e nela veremos os conceitos de probabilidade. É importante que você esteja atento aos exercícios resolvidos e, à medida que for avançando, relembre os conceitos aprendidos anteriormente. Preparado para mais esse desafio?  
Então, vamos juntos!

A origem da Teoria das Probabilidades está relacionada aos jogos de azar desde o século XVII, pois surgiu da necessidade de um método racional para calcular os riscos dos jogadores em jogos de cartas, de dados etc.

Posteriormente, passou a auxiliar governos, empresas e organizações profissionais em seus processos de decisões, ajudando a desenvolver estratégias. Na área da Gestão, passou a ser uma ferramenta para tomada de decisões e para análise de chances e de riscos. Para decidir por um ou por outro procedimento, é essencial conhecermos as chances de cada um dar certo e, também, decidirmos sobre um sistema de gestão. Também, para sabermos os riscos de uma exposição poder afetar a imagem de um administrador, temos de conhecer a probabilidade de ela causar dano ou não.

Para que você possa entender melhor os principais conceitos de probabilidade, destacamos a seguir dois tipos de fenômenos:

- ▶ **Fenômenos determinísticos:** aqueles que invariavelmente dão o mesmo resultado se repetidos sob condições específicas. Um exemplo é a aceleração da gravidade na ausência de ar (vácuo). Nesse caso, o resultado sempre será o mesmo, pois não temos variações que venham a influenciar o resultado.
- ▶ **Fenômenos aleatórios:** aqueles que, mesmo realizados sob as mesmas condições, apresentam variações nos resultados de diferentes observações. Pense na reação de um contribuinte quando ele é atendido ou no lançamento de um dado. Em cada uma dessas situações, os resultados nem sempre serão os mesmos. Por isso são aleatórios, ou seja, ocorrem de forma aleatória, sem resultado previsível.

São nos fenômenos aleatórios que a Teoria das Probabilidades auxilia na análise e na previsão de um resultado futuro. Quando você pensa em probabilidade, vai querer identificar a chance de ocorrência de um determinado resultado de interesse em situações nas quais não é possível calcular com exatidão o valor real do **evento (fenômeno aleatório)**. Dessa forma, trabalhamos com chances ou probabilidades.

*Uma situação que exemplifica esse fato está associada à seguinte pergunta: um funcionário público poderá cumprir sua meta de trabalho na semana que vem?*

Para responder a essa e a outras perguntas, você poderá aplicar alguns conceitos apresentados, a seguir.



## EXPERIMENTO ALEATÓRIO

Para você calcular uma probabilidade, é necessário ter um experimento aleatório, ou seja, qualquer processo que venha a gerar um resultado incerto ou casual.

Para que um processo possa ser considerado um experimento aleatório, ele deve ter as seguintes características:

- ▶ cada experimento pode ser repetido indefinidamente sob as mesmas condições ( $n$ );
- ▶ não se conhece *a priori* o resultado do experimento, mas pode-se descrever todos os possíveis resultados; e
- ▶ quando o experimento for repetido inúmeras vezes, surgirá uma regularidade do resultado, isto é, haverá uma estabilidade da fração  $f = \frac{r}{n}$  (frequência relativa) da ocorrência de um particular resultado, em que  $r$  corresponde ao número de vezes que um determinado resultado aconteceu.

Sendo assim, podemos considerar que um processo aleatório corresponde, para ilustrar, ao lançamento de uma moeda jogada inúmeras vezes, já que pode ser repetido indefinidamente. Não conhecemos o resultado, mas podemos descrever os possíveis resultados (cara ou coroa). Além disso, quando você lança a moeda três mil vezes, por exemplo, ocorre uma estabilização da frequência relativa ou probabilidade em 0,5. A Figura 14 nos mostra que no início a frequência relativa não é tão próxima de 0,5, como acontece após 1.000 jogadas.

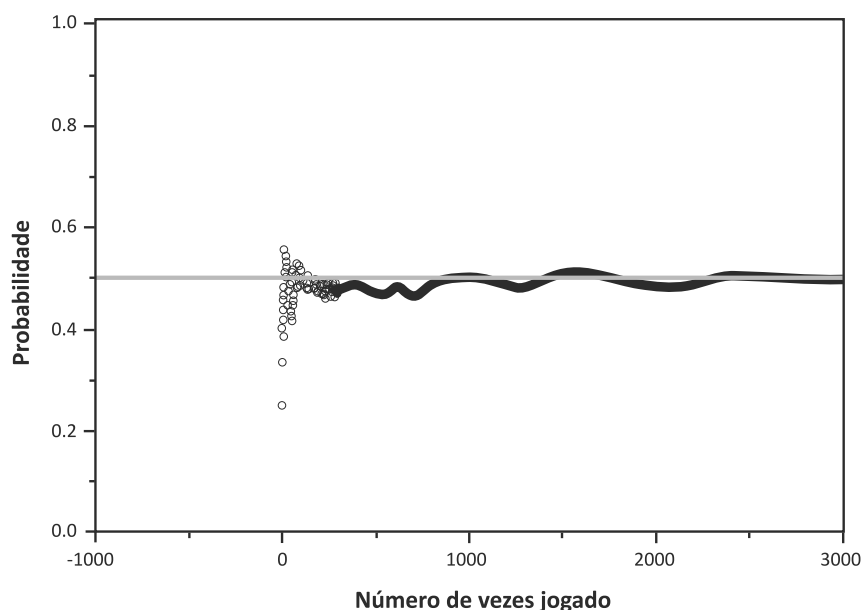


Figura 14: Experimento aleatório

Fonte: Elaborada pelo autor

A incerteza está associada à chance de ocorrência que atribuímos ao resultado de interesse.

Perceba, com base nos experimentos e nas situações mencionadas, que a **incerteza** sempre está presente, o que quer dizer que, se esses experimentos forem repetidos em idênticas condições, não se pode determinar qual resultado ocorrerá.

Para entender melhor esse conceito, vamos considerar como exemplo o setor de atendimento de uma determinada prefeitura que conta com seis funcionários. Um experimento ao acaso seria a escolha aleatória de um dos funcionários. Podemos considerar o gênero do funcionário escolhido como o que queremos avaliar. Você, então, vai aplicar os conceitos vistos de experimento aleatório. Veja que este corresponde a um experimento aleatório, pois sabemos quais resultados podem ocorrer, ou seja, um dos seis funcionários será o avaliado. Entretanto, não podemos dizer que resultado (pessoa) sairá nesse sorteio.

*Agora que você entendeu o que é experimento aleatório, você irá compreender outro conceito importante: o de espaço amostral.*

## ESPAÇO AMOSTRAL ( $\Omega$ )

Vamos considerar a situação em que um funcionário público consegue ou não atingir sua meta de produtividade.

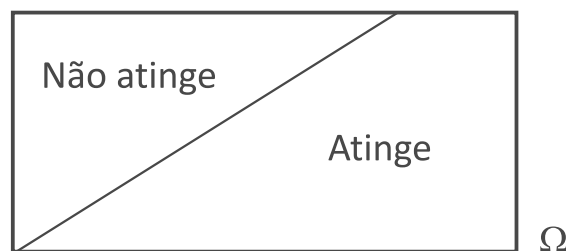
*Nesse caso, quais os possíveis resultados que você pode ter?*

O funcionário poderá atingir ou não a meta. Então, temos apenas dois resultados possíveis. O conjunto desses resultados possíveis, que poderiam ser mais de dois também, no caso de outras situações, é definido como **espaço amostral\*** e pode ser simbolizado por S ou  $\Omega$  (omega).

\*Espaço amostral – conjunto de todos os resultados possíveis. Fonte: Elaborado pelo autor.

No nosso caso, teremos  $\Omega = \{\text{atinge; não atinge}\}$

Lembrando do **Diagrama de Venn**, que você estudou na disciplina *Matemática para Administradores*, podemos representar o espaço amostral conforme indica a Figura 15:



Os Diagramas de Venn são úteis para mostrar a relação entre os elementos de um conjunto.

Figura 15: Representação do espaço amostral  
Fonte: Elaborada pelo autor

A definição do espaço amostral é de fundamental importância, pois, muitas vezes, a partir dele, você pode calcular probabilidades. Veremos isso um pouco mais a frente.

*Nesse caso, se todos os resultados possíveis constituem o nosso espaço amostral, o que será cada resultado em particular?*

Com intuito de responder a essa proposição, daremos continuidade ao nosso estudo. Vamos à próxima seção.

## EVENTO

Qualquer subconjunto do espaço amostral ( $\Omega$ ) associado ao experimento aleatório é chamado de evento, ou seja, um determinado resultado que ocorra dentro do espaço amostral. Então, em nosso exemplo, teremos que o funcionário público que cumprir a meta será considerado como um dos eventos que compõem o espaço amostral. Nesse caso, o nosso espaço amostral apresenta dois eventos apenas (cumprir ou não cumprir a meta).

Geralmente, calculamos as chamadas probabilidades desses eventos associados ao nosso espaço amostral. Por isso a importância de você ter esse conceito bem definido em sua mente!

Imagine que algumas secretarias municipais oferecem, por cortesia, cadeiras suficientes em determinado setor para que os contribuintes possam esperar confortavelmente; e, outras secretarias, não oferecem essa cortesia. Vamos ver como esse problema pode ser formulado dentro do contexto de experimento aleatório, espaço amostral e eventos.

O **experimento** é a seleção de uma secretaria e a observação do fato dessa secretaria oferecer ou não a cortesia. Há dois pontos amostrais no espaço correspondente a esse experimento:

S:{a secretaria oferece a cortesia}

N:{a cortesia de cadeira não é oferecida pela secretaria}

Um ponto importante a ser considerado é o de que nem sempre as chances de ocorrência dos eventos são iguais a 50%, como no caso do lançamento de uma moeda. Nessa situação, provavelmente a chance da secretaria oferecer a cortesia de assentos (S) poderá ser bem maior do que a de não oferecer (N).

## DEFINIÇÕES DE PROBABILIDADES

Até agora vimos diferentes e importantes conceitos relacionados à estatística. Vamos agora definir o que vem a ser probabilidade. Para o bom entendimento desse conceito, imagine as seguintes situações:

- ▶ 50% de chance de um projeto dar certo;
- ▶ 95% de certeza de que um determinado serviço será realizado por uma prefeitura em tempo hábil; e
- ▶ 1 em cada 10 servidores públicos não tem ido trabalhar pelo menos um dia na semana.

Como você pode ver, estamos falando das chances acerca de que algo venha a acontecer. Então, probabilidade pode ser considerada a chance de que um determinado evento venha a ocorrer.

*As probabilidades apresentam diferentes visões. As principais são mostradas a seguir, acompanhe!*

A Probabilidade Objetiva nasceu no século XVII por interesse comum de Fermat e Pascal.



### **Saiba mais** Pierre Fermat (1601-1665)

Matemático francês que passou parte de sua vida como conselheiro do parlamento de Toulouse. Seu campo predileto de estudos foi o da teoria dos números, na qual se consagra. Fermat dá considerável impulso à aritmética superior moderna, exercendo grande influência sobre o desenvolvimento da álgebra. Fermat se sobressai, ainda, no terreno do cálculo de probabilidades. Fonte: <<http://ecalculo.if.usp.br/historia/fermat.htm>>. Acesso em: 24 nov. 2010.

### **Blaise Pascal (1623-1662)**

Com apenas três anos, perdeu a mãe. O pai encarregou-se diretamente da sua educação, desenvolvendo um método singular de educação com exercícios e jogos de disciplinas, como Geografia, História e Filosofia. Contudo, seu pai acreditava que a Matemática somente deveria ser ensinada ao filho quando este fosse mais velho. Porém, Pascal descobriu as maravilhas da ciência dos números. Aos 12 anos, mesmo sem professor, ele descobre que a soma dos ângulos de um triângulo é igual a dois ângulos retos. Fonte: <<http://tinyurl.com/285crno>>. Acesso em: 24 nov. 2010.

\*Mutuamente excludentes – a ocorrência de um evento exclui a ocorrência do outro. Fonte: Elaborado pelo autor.

\*Igualmente prováveis – ocorrem com a mesma probabilidade. Fonte: Elaborado pelo autor.

Se um **evento** pode ocorrer em  $n$  maneiras **mutuamente excludentes\*** e **igualmente prováveis\***, e, se  $m$  dessas ocorrências tem uma característica  $E$ , então, a probabilidade de ocorrência de  $E$  é:

$$P(E) = \frac{m}{N}$$

Onde:

$m$ : número de eventos favoráveis à probabilidade  $E$  que se deseja calcular, ou seja, o número de vezes que  $E$  acontece; e

$N$ : número total de ocorrências dentro do espaço amostral.

Vejamos exemplos de probabilidades a serem obtidas:

- ▶ Um dado homogêneo tem probabilidade  $1/6$  de cair com a face 2 para cima.
- ▶ Em um conjunto de cartas (sem os coringas) bem embaralhadas, a probabilidade de sortearmos uma carta de copas é de  $13/52$ .

\* Reprodutibilidade – ocorrência de diversas vezes de um mesmo evento. Fonte: Elaborado pelo autor.

A visão da frequência relativa depende da **reprodutibilidade\*** do mesmo processo e da habilidade de contarmos o número de repetições.

Sendo assim, se algum processo é repetido um grande número de vezes,  $n$ , e se algum evento com característica  $E$  ocorre  $m$  vezes, a frequência relativa  $m/n$  é aproximadamente igual à probabilidade de  $E$ :

$$P(E) \approx m/n$$

Contudo, observe que  $m/n$  é apenas uma estimativa de  $P(E)$ .

A visão da probabilidade subjetiva é uma medida da “confiança” que temos sobre a verdade de certa proposição, apesar de não termos cálculos precisos sobre esse valor. Imagine

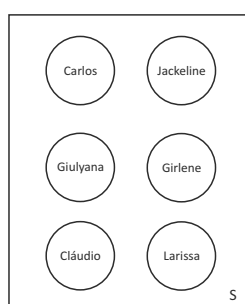
proposições sobre a probabilidade de que em três anos teremos um modelo eficiente de gestão pública ou que as capacidades do processamento computacional se igualarão à capacidade do cérebro humano em 30 anos. Ambas são apenas estimativas que não se baseiam em cálculos.

*Para que você entenda melhor algumas das definições de probabilidade, veja a descrição que preparamos ao longo de uma situação.*

Imagine que em um determinado setor de uma prefeitura temos os seguintes funcionários: Carlos, Jackeline, Giulyana, Girlene, Cláudio e Larissa. Então, você pode verificar que temos seis funcionários. Vamos pensar agora: qual a probabilidade de se escolher um funcionário ao acaso e ele ser do gênero masculino?

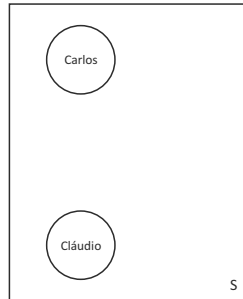
Para obtermos as respostas, vamos definir o espaço amostral e o evento desejado. Consideremos espaço amostral ou conjunto de possibilidades todos os funcionários públicos do setor.

$$S = \{\text{Carlos, Jackeline, Giulyana, Girlene, Cláudio, Larissa}\}$$



E, para definir o evento favorável, precisamos considerar este o conjunto de possibilidades favoráveis que nos interessa, ou seja, os funcionários do gênero masculino.

**Evento = {Carlos, Cláudio}**



Então, a probabilidade que estamos procurando, ou seja, a de escolher um funcionário ao acaso e ele ser do gênero masculino, pode ser apresentada conforme descrição, a seguir:

$$P(\text{funcionário público gênero masculino}) = \frac{2}{6} = \frac{\text{número de funcionários do sexo masculino}}{\text{número total de funcionários}}$$

Logo, considerando três eventos relativos aos funcionários da prefeitura, conforme descrevemos anteriormente, temos:

- ▶ A (funcionário ser do sexo feminino).
- ▶ B (seu nome começar com a letra G).
- ▶ C (seu nome começar com a letra C).

Então, poderemos definir os eventos mencionados anteriormente como:

- ▶ A = {Jackeline, Giulyana, Girlene, Larissa}.
- ▶ B = {Giulyana, Girlene}.
- ▶ C = {Carlos, Cláudio}.

Você pode definir a probabilidade como uma função que atribui um número real aos eventos do  $\Omega$  (se A é um evento do  $\Omega$ ,  $P(A)$  é a probabilidade de A), a qual satisfaz:

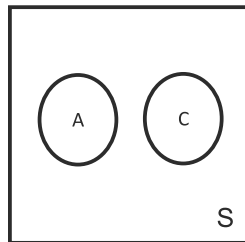
- ▶  $P(\emptyset) = 0$  (probabilidade de vazio é igual a zero).
- ▶  $P(\Omega) = 1$  (probabilidade de acontecer todo o espaço amostral é igual a um).



- $0 \leq P(A) \leq 1$  (a probabilidade de um determinado evento, sempre estará entre zero e um).

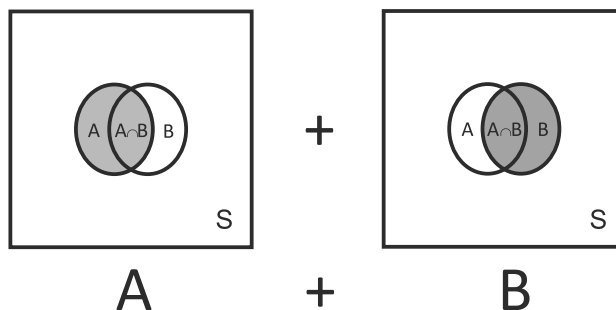
Você pode ainda utilizar a regra da soma, em que dados dois **eventos mutuamente exclusivos\***, A e C de  $\Omega$ , temos:

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



\* Eventos mutuamente exclusivos – são aqueles que não podem acontecer simultaneamente.  
Fonte: Elaborado pelo autor.

Já no caso a seguir, em que os eventos não são mutuamente exclusivos e podem ocorrer simultaneamente, na regra da soma, devemos considerar que a intersecção (área) será contada duas vezes.



Nesse caso, devemos retirar uma vez a área de  $(A \cap B)$  na regra da soma, pois, como você pode ver nos desenhos anteriores, a intersecção  $(A \cap B)$  é contada duas vezes.

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

Lembre-se de que o símbolo  $\cap$  corresponde à intersecção e  $\cup$  corresponde à união.

Considerando os eventos A, B e C, temos as seguintes situações:

- $A \cup C$  é o evento em que A ocorre ou C ocorre ou, ainda, ambos ocorrem  $\rightarrow \{\text{Carlos, Jackeline, Giulyana, Girlene, Cláudio, Larissa}\}$ .

E a chance de acontecerem dois eventos simultaneamente corresponde à chance de os eventos acontecerem ao mesmo tempo, como você pode observar na descrição, a seguir:

- $A \cap B$  é o evento em que A e B ocorrem simultaneamente  $\rightarrow \{\text{Giulyana, Girlene}\}$ .

Em muitas situações o que nos interessa é aquilo que pertence ao espaço amostral e não pertence ao evento de interesse. A Figura 16 mostra bem isso:

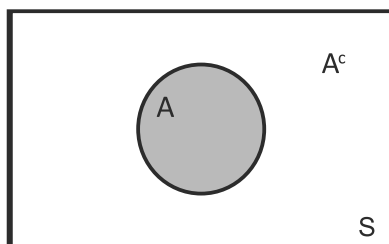


Figura 16: Espaço amostral  
Fonte: Elaborada pelo autor

$\bar{A}$  ou  $A^c$  é o evento em que A não ocorre (complementar de A). Em nosso exemplo, consideramos que o complementar de A (funcionário ser do gênero feminino) corresponde a todas as pessoas do gênero masculino, ou seja:

$$\bar{A} \text{ ou } A^c = \{\text{Carlos, Claudio}\}$$

## PROBABILIDADE CONDICIONAL

A partir de agora veremos outros conceitos de probabilidade e para tanto você deve considerar os dados, a seguir, referentes a uma prefeitura, em que foram selecionados, a partir de uma amostragem estratificada (vista anteriormente), 101.850 contribuintes das classes **média-baixa** e **alta**. Posteriormente, foi feita a verificação do número de contribuintes, de cada classe social, que pagaram um determinado tributo em dia (evento: pagaram) e também o número de contribuintes das classes estudadas que não pagaram em dia o tributo (evento: não pagaram). Para compreender essa descrição, observe os resultados descritos na Tabela 13:

Tabela 13: Contribuintes pagantes e não pagantes

	MÉDIA-BAIXA	ALTA	TOTAL
Pagaram (P)	39.577	8.672	48.249
Não Pagaram (NP)	46.304	7.297	53.601
Total	85.881	15.969	101.850

Fonte: Elaborada pelo autor

De acordo com os dados apresentados, podemos considerar então que o nosso espaço amostral ( $\Omega$ ) corresponderá ao conjunto de 101.850 contribuintes.

Agora, para ampliarmos essa discussão juntos, você vai considerar os eventos apresentados, a seguir, para que possamos trabalhar com eles.

- P = contribuintes que **pagaram** o tributo em dia.
- NP = contribuintes que **não pagaram** o tributo em dia.

- $MB$  = contribuintes da classe **média-baixa**.
- $P \cap MB$  = contribuintes que **pagaram (P)** o tributo em dia e **ao mesmo tempo** são da classe **média-baixa (MB)**.
- $P \cup MB$  = contribuintes que **pagaram (P)** o tributo em dia ou são da classe **média-baixa (MB)**.

Você pode obter, então, as probabilidades de alguns eventos considerados anteriormente, por exemplo:

$$P(MB) = \frac{\text{n}^\circ \text{ de contribuintes que são da classe MÉDIA-BAIXA}}{\text{n}^\circ \text{ total de contribuintes}} = \frac{85.881}{101.850} = 0,843$$

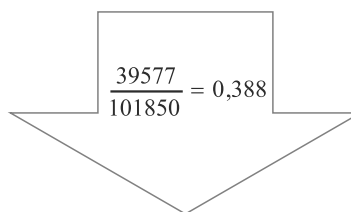
$$P(P) = \frac{\text{n}^\circ \text{ de contribuintes que pagaram em dia}}{\text{n}^\circ \text{ total de contribuintes}} = \frac{48.249}{101.850} = 0,473$$

Considerando os contribuintes que pagam e os que não pagam em dia, temos apenas estes dois resultados possíveis. E, para obtermos a probabilidade de contribuintes que não pagaram em dia, teremos a probabilidade de todo o espaço amostral (101.850), que é igual a 1 menos a probabilidade de contribuintes que pagaram em dia ( $P$ ). Nesse caso, estamos usando o conceito de eventos complementares. Este cálculo é mostrado para você a seguir:

$NP = \bar{P}$  (não pagaram (NP ou  $\bar{P}$ ) é o complementar dos que pagaram ( $P$ ))

$$\text{ou seja, } P(NP) = P(\bar{P}) = 1 - P(P) = 1 - 0,473 = 0,527$$

Com base nesse conhecimento, podemos calcular a probabilidade de escolher um contribuinte aleatoriamente e este ser da classe média-baixa ou ser quem paga em dia o tributo. Veja que, nesse caso, os eventos não são mutuamente exclusivos, ou seja, existem contribuintes que são comuns nas duas situações ao mesmo tempo. Assim, a probabilidade procurada será dada por:



$$\frac{39577}{101850} = 0,388$$

$$P(P \cup MB) = P(P) + P(MB) - P(P \cap MB)$$

$$P(P \cup MB) = 0,473 + 0,843 - 0,388$$

$$P(P \cup MB) = 0,928$$

*Vamos considerar ainda o exemplo anterior. Se você souber que um contribuinte sorteado paga em dia o tributo, qual a probabilidade de que ele seja da classe média-baixa?*

Agora, temos uma informação parcial e importante: o contribuinte selecionado paga em dia. Vamos então designar a probabilidade de P quando se sabe que o contribuinte selecionado paga em dia o tributo e MB quando o contribuinte é da classe social média-baixa.

Assim, a probabilidade que chamaremos de  $P(MB/P)$  é denominada de **probabilidade (condicional) de MB dado P (lembre-se que o símbolo / não corresponde a uma divisão e sim a uma condição de que outro evento já aconteceu)**. Então, nesse caso, temos o que chamamos de probabilidade condicionada, ou seja, a probabilidade de um evento **acontecer dado que, sabendo que**, outro evento já aconteceu.

Sendo assim, é natural atribuímos:

$$\begin{aligned} P(MB/P) &= \frac{\text{nº de contribuintes que são da classe MÉDIA-BAIXA e pagam em dia}}{\text{nº total contribuintes que pagam em dia}} = \\ &= \frac{39.577}{48.249} = 0,820 \end{aligned}$$

Veja que, nesse caso, ocorreu uma redução no espaço amostral, já que tínhamos a informação anterior de que o

contribuinte selecionado pagava em dia. Dessa forma, do espaço amostral total que tínhamos (101.850), ele foi reduzido para 48.249 e, destes, interessavam-nos os que eram da classe social média-baixa. Sendo assim:

$$P(MB/P) = \frac{\frac{\text{nº de contribuintes da classe MÉDIA-BAIXA e que pagam em dia}}{\text{nº total de contribuintes}}}{\frac{\text{nº de contribuintes que pagam em dia}}{\text{nº total de contribuintes}}}$$

$$P(MB/P) = \frac{P(MB \cap P)}{P(P)}$$

Portanto, você pode generalizar para dois eventos A e B quaisquer de um experimento aleatório. Dessa forma, podemos dizer que a probabilidade condicional de A dado B (nota-se por  $P(A/B)$ ) é definida como:

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

De posse desse conhecimento, podemos definir, a partir de agora, a regra do produto; conforme discutiremos na próxima seção.

## REGRA DO PRODUTO E EVENTOS INDEPENDENTES

A partir da probabilidade condicionada definida anteriormente, obteremos a chamada **regra do produto** para a probabilidade da interseção de dois eventos A e B de um espaço amostral:

Passe a probabilidade de ocorrência de B na probabilidade condicionada e multiplique pela probabilidade de ocorrência de A sabendo que B já aconteceu.

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

Logo, se dois eventos A e B são independentes, então  $P\{A/B\} = P\{A\}$  ou  $P\{B/A\} = P\{B\}$ , já que um evento não interfere no outro, ou seja, eles são independentes.

Desse modo, se A e B forem independentes, você pode verificar que:

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

Então, para que dois eventos A e B quaisquer sejam considerados independentes é necessário fazer a seguinte relação:

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

Para compreender melhor essa nossa discussão, analise outra situação na qual utilizaremos os conceitos aprendidos de probabilidade. Para tanto, considere os dados a seguir, representativos da distribuição da renda anual de funcionários públicos de dois setores (A e B), apresentados na Tabela 14.

Tabela 14: Distribuição de renda anual do funcionário público

FAIXA DE RENDA ANUAL (EM R\$1.000,00)	SETOR		TOTAL
	A	B	
15 a 20 (R1)	70	40	110
20 a 25 (R2)	15	15	30
25 a 30 (R3)	10	20	30
30 a 35 (R4)	20	10	30
Total	115	85	200

Fonte: Elaborada pelo autor

Observando os dados descritos na Tabela 14, podemos identificar claramente que a probabilidade de um funcionário aleatoriamente escolhido ser:

- a) do setor A  $\rightarrow P(A) = 115/200 = 0,575$  (temos 115 funcionários do setor A em um total de 200 funcionários);
- b) do setor B  $\rightarrow P(B) = 85/200 = 0,425$  (temos 115 funcionários do setor A em um total de 200 funcionários);
- c) de ter renda entre R\$ 15.000,00 e R\$ 20.000,00  $\rightarrow P(R1) = 110/200 = 0,550$  (110 funcionários correspondem aos que têm a faixa de renda solicitada);
- d) do setor B e ter renda entre R\$ 15.000,00 e R\$ 20.000,00  $\rightarrow$  (intersecção), ou seja,  $P(B \cap R1) = 40/200 = 0,20$  (temos 40 funcionários que correspondem aos que têm a faixa de renda solicitada e ao mesmo tempo são do setor B); e
- e) ter renda entre R\$ 15.000,00 e R\$ 20.000,00, dado que é do setor B  $\rightarrow$

$$P(R1/B) = \frac{P(R1 \cap B)}{P(B)} = \frac{0,20}{0,425} = 0,4706$$

Sabendo que o funcionário é do setor B (temos 85 funcionários agora), houve uma redução no espaço amostral de 200 para 85 que será utilizado no denominador. Logo, perguntamos: qual a chance de estar na faixa de renda solicitada?

Como  $P(R1) \neq P(R1/B)$ , podemos concluir que os eventos setor e renda são dependentes. Podemos visualizar um exemplo de aplicação dos conceitos de **independência** de eventos por meio do lançamento de uma moeda não viciada (não existe preferência para cara ou coroa) três vezes. Considere os seguintes eventos:

A = no primeiro lançamento da moeda sai cara; e

B = no segundo lançamento da moeda sai cara.

Para que sejam considerados independentes, a relação de independência deve ser válida para todas as intersecções presentes na Tabela 14.



**Considere C = cara e R = coroa**

Verifique se é verdadeira a hipótese de que os eventos A e B são independentes. O espaço amostral e os eventos são apresentados, a seguir:

$$\Omega = \{CCC, CCR, CRC, CRR, RCC, RCR, RRC, RRR\}$$

$$(A) = \{\mathbf{CCC}, \mathbf{CCR}, CRC, CRR\}$$

$$(B) = \{\mathbf{CCC}, \mathbf{CCR}, RCC, RCR\}$$

$$P(A \cap B) = 2/8 = 1/4$$

$$P(A) = 4/8 = 1/2$$

$$P(B) = 4/8 = 1/2$$

$$\text{Portanto, } P(A \cap B) = P(A) \cdot P(B) \Rightarrow 1/4 = 1/2 \cdot 1/2 \text{ ou}$$

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{1/2} = \frac{2}{4} = \frac{1}{2} \Rightarrow P(A/B) = P(A) \Rightarrow 1/2 = 1/2$$

Sendo assim, percebe-se que os eventos são independentes, pois  $P(A \cap B) = P(A) \cdot P(B)$  ou  $P(A/B) = P(A)$ .

*Vamos ver outros exemplos relacionados a probabilidades para compreendermos melhor o que vimos.*

**Exemplo**

Um estudante chega atrasado em 40% das aulas e esquece o material didático em 18% das aulas. Supondo eventos independentes, calcule a probabilidade de:

- O estudante chegar na hora e com material.
- Não chegar na hora e ainda sem material.

Como o exercício afirma que o estudante chega atrasado em 40% das aulas, entendemos que  $40\% = 0,40$ , ou seja, ele não chegar atrasado =  $60\% = 0,6$ . O exercício afirma também que ele esquece



Os resultados que estão em negrito ocorrem no espaço amostral (8) somente duas vezes.

o material didático em 18% da aula, isto é, ele esquece o material = 18% = 0,18 e ele não esquecer o material = 82% = 0,82.

Logo, para resolver a alternativa do exemplo, probabilidade de o estudante chegar na hora e com material, considerando que os eventos são independentes, temos:

$$P(\text{chegar na hora e com material}) = P(\text{chegar na hora} \cap \text{c/ material}) = \\ P(\text{chegar na hora}) \cdot P(\text{c/ material}) = 0,60 \cdot 0,82 = 0,492 \text{ ou } 49,2\%$$

Já para resolvermos a alternativa b, vamos considerar que:

$$P(\text{não chegar na hora e sem material}) = P(\bar{n} \text{ chegar na hora} \cap \text{s/ material}) = \\ P(\bar{n} \text{ chegar na hora}) \cdot P(\text{s/ material}) = 0,40 \cdot 0,18 = 0,072 \text{ ou } 7,2\%$$

### Exemplo:

Vamos considerar um pesquisador que estudou o comportamento de consumo de bebidas lácteas no Brasil. Após análise da classe econômica do consumidor e o principal aspecto determinante da escolha da marca, o pesquisador tabulou os dados conforme disposto a seguir.

CLASSE/ASPECTO	PREÇO	QUALIDADE	SOMA
Alta	42	56	98
Média	37	21	58
Baixa	13	97	110
Total	92	174	266

Considerando esses dados, qual a probabilidade de um consumidor escolhido:

- Priorizar o preço, dado que é da classe alta.
- Priorizar a qualidade, dado que é da classe média.
- Ser da classe baixa, dado que atribui maior importância ao fator qualidade.

Com base nos dados da tabela desse exemplo, para priorizar o preço, dado que é da classe alta, temos uma probabilidade condicional igual:

$$P(\text{preço}/\text{classe alta}) = \frac{P(\text{preço} \cap \text{classe alta})}{P(\text{classe alta})} = \frac{42}{98} = 0,4286 \text{ ou } 42,86\%$$

Já para priorizar a qualidade, dado que é da classe média, temos uma probabilidade condicional dada por:

$$P(\text{qualidade}/\text{classe média}) = \frac{P(\text{qualidade} \cap \text{classe média})}{P(\text{classe média})} = \frac{21}{58} = 0,3621 \text{ ou } 36,21\%$$

Por fim, para ser da classe baixa, dado que atribuiu maior importância ao fator qualidade, o cálculo é feito por:

$$P(\text{classe baixa}/\text{qualidade}) = \frac{P(\text{classe baixa} \cap \text{qualidade})}{P(\text{qualidade})} = \frac{97}{174} = 0,5575 \text{ ou } 55,75\%$$

# ALGUMAS REGRAS BÁSICAS DE PROBABILIDADE

Para que você possa aplicar todos os conceitos de probabilidade aprendidos até aqui, apresentaremos, por meio da Figura 17, algumas regras básicas que irão ajudá-lo. Observe com atenção:

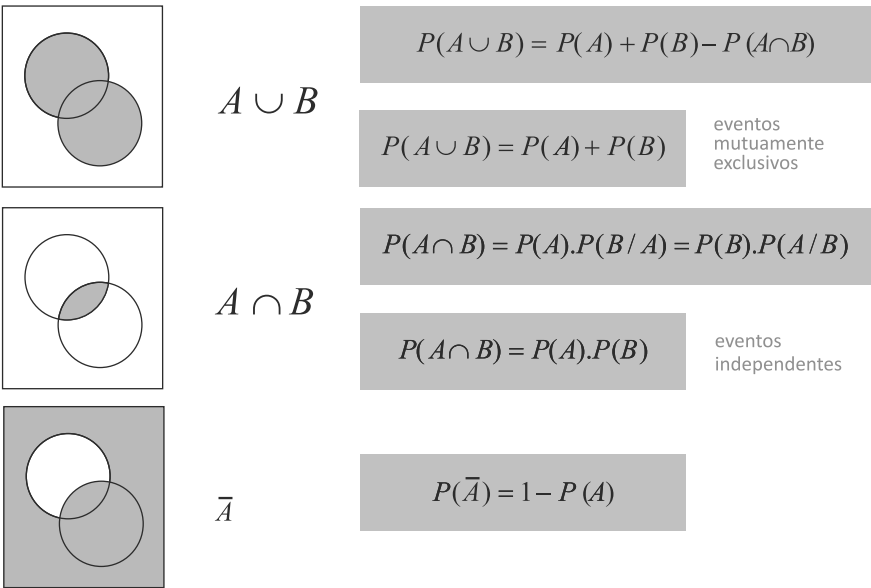


Figura 17: Regras gerais da probabilidade  
Fonte: Elaborada pelo autor

Outra questão que merece destaque quando falamos de probabilidade é que a probabilidade condicional de A dado B é definida por:

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

# Resumindo

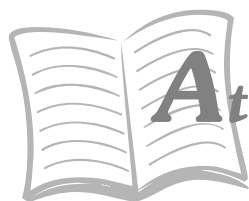


Nesta Unidade, você ampliou o seu conhecimento quanto ao termo probabilidade.

Descrevemos as abordagens clássicas das frequências relativa e subjetiva da probabilidade e entendemos os termos experimento, espaço amostral e evento.

Vimos a definição dos termos probabilidade condicional e probabilidade conjunta, além de aprendermos a calcular as probabilidades aplicando as regras da adição e da multiplicação. Para intensificar nosso estudo, vimos esses conceitos aplicados a partir da apresentação de exemplos.

Caso algum conceito não tenha ficado claro, retome a leitura, pois eles serão importantes para a compreensão de novas informações contidas nas Unidades posteriores.



## Atividades de aprendizagem

Agora que você já entendeu todos os conceitos relacionados aos cálculos de probabilidade apresentados, resolva as atividades apresentadas, a seguir, e, em caso de dúvidas, não hesite em consultar o seu tutor.

1. Considerando as probabilidades de três fiscais A, B e C, que trabalham independentemente, efetivarem uma autuação quando abordam uma obra são  $\frac{2}{3}$ ,  $\frac{4}{5}$  e  $\frac{7}{10}$ , respectivamente. Se cada um abordar uma obra, qual a probabilidade de que pelo menos um efetive a multa?
2. Sendo A e B dois mestres que já estão suficientemente treinados em partidas de xadrez e jogam 120 partidas, das quais A ganha 60, B ganha 40 e 20 terminam empatadas; A e B concordam em jogar três partidas. Determine a probabilidade de:
  - a) A ganhar todas as partidas.
  - b) Duas partidas terminarem empatadas.
  - c) A e B ganharem alternadamente.
3. Em um período de um mês, 100 funcionários de uma prefeitura que trabalham com resíduos tóxicos, sofrendo de determinada doença, foram tratados. As informações sobre o método de tratamento aplicado a cada funcionário e o resultado final obtido estão na tabela a seguir:

		TRATAMENTO	
		A	B
Resultado	Cura Total	24	16
	Cura Parcial	24	16
	Morte	12	8

Sorteando-se aleatoriamente um desses funcionários, determine a probabilidade de o funcionário escolhido ter sido:

- Submetido ao tratamento A.
- Totalmente curado.
- Submetido ao tratamento A e ter sido parcialmente curado.
- Submetido ao tratamento A ou ter sido parcialmente curado.



# UNIDADE 5

## DISTRIBUIÇÃO DE PROBABILIDADES DISCRETAS E CONTÍNUAS

### OBJETIVOS ESPECÍFICOS DE APRENDIZAGEM

Ao finalizar esta Unidade, você deverá ser capaz de:

- ▶ Identificar e aplicar modelos probabilísticos discretos;
- ▶ Identificar e aplicar modelos probabilísticos contínuos (distribuição normal);
- ▶ Saber quando e como utilizar as distribuições amostrais;
- ▶ Calcular e interpretar intervalos de confiança; e
- ▶ Dimensionar amostras para serem utilizadas em pesquisas e projetos.



## INTRODUÇÃO

Caro estudante,

Como você progrediu nos conhecimentos básicos de probabilidade, agora iremos trabalhar com as chamadas distribuições de probabilidades. Essas distribuições auxiliam no cálculo de probabilidades e, ainda, nos processos de estimação e de decisão, conforme veremos na próxima Unidade. Estudaremos as distribuições de amostragem e dimensionamento de amostras que, também, serão vistas nesta Unidade.

Bons estudos e conte conosco para auxiliá-lo sempre que necessário.

Vamos começar com alguns conceitos preliminares.

Para que você tenha condições de entender as distribuições, é necessário conhecer bem o que é uma **variável aleatória\***, que pode ser discreta ou contínua.

Um exemplo de uma variável aleatória discreta (v.a.) é a quantidade de ações que tiveram queda em um determinado dia, em uma carteira composta por cinco ações diferentes. A função será dada por:

$X$  = “quantidade de ações que tiveram queda em um determinado dia” define uma variável aleatória discreta, que pode assumir os valores 0, 1, 2, 3, 4, 5.

Vamos considerar agora uma situação na qual se verificou o tempo gasto por um funcionário público para atender um contribuinte. A função será:

$Y$  = “tempo gasto por um funcionário público para atender um contribuinte” define uma variável aleatória contínua, que pode assumir infinitos valores.

\* Variável aleatória – função que associa valores reais aos eventos de um espaço amostral. Fonte: Elaborado pelo autor.

Vamos trabalhar aqui principalmente com as variáveis aleatórias discretas. Se uma variável aleatória  $X$  pode assumir os valores  $x_1, x_2, \dots, x_n$  com probabilidades respectivamente iguais a

$p_1, p_2, \dots, p_n$ , e  $\sum_{i=1}^n p_i = 1$ , temos então definida uma **distribuição**

**de probabilidade\***.

\* Distribuição de probabilidade – é um tipo de distribuição que descreve a chance que uma variável pode assumir ao longo de um espaço de valores. Fonte: Elaborado pelo autor.

É importante ressaltarmos que a variável aleatória tem notação de letra maiúscula e seus possíveis valores minúsculos, como utilizamos anteriormente.

Se a variável  $X$  em questão for discreta, sua distribuição é caracterizada por uma **função de probabilidade ( $P(X=x)$ )**, que associa probabilidades não nulas aos possíveis valores da variável aleatória.

## DISTRIBUIÇÕES DISCRETAS

Imagine uma situação na qual somente podem ocorrer dois possíveis resultados, “sucesso” e “fracasso”. Veja alguns exemplos:

- ▶ uma venda é efetuada ou não em uma ligação de *call center*;
- ▶ um contribuinte pode ser adimplente ou inadimplente;
- ▶ uma guia recolhida pode ter seu preenchimento ocorrido de forma correta ou incorreta; e
- ▶ um consumidor que entra em uma loja pode comprar ou não comprar um produto.

Essas situações correspondem à Distribuição de Bernoulli. Ou seja, se associarmos uma variável aleatória  $x$  aos possíveis resultados do experimento de forma que  $X=1$  se o resultado for “sucesso” e  $X=0$  se o resultado for “fracasso”, então, a variável aleatória  $X$ , assim definida, tem Distribuição de Bernoulli, com  $p$  sendo a probabilidade de ocorrer “**sucesso**” e  $q = (1-p)$  a probabilidade de ocorrer “**fracasso**”.

Neste momento, você deve saber que quando estamos falando de sucesso, devemos relacioná-lo com o objetivo do exercício ou do problema a ser resolvido, que, muitas vezes, pode não ser algo bom.

Ampliando nossa discussão, é importante mencionarmos ainda que a função de probabilidade da Distribuição de Bernoulli é dada por:

$$P(X = x) = \begin{cases} p & \text{para } x = 1, \\ q = 1 - p & \text{para } x = 0 \\ 0 & \text{para } x \text{ diferente de 0 ou 1} \end{cases}$$

Sendo assim, a média e a variância serão obtidas por:

- Média =  $p$  (onde  $p$  corresponde à probabilidade de sucesso).
- Variância =  $p \cdot q$  (onde  $q$  corresponde à probabilidade de fracasso).

Essa obtenção da estimativa de média e desvio padrão é importante, pois tais medidas podem ser usadas para caracterizar a situação e também para a definição da média e do desvio padrão da distribuição binomial que iremos ver posteriormente.

Contextualizando a Distribuição de Bernoulli, temos a seguinte situação: a experiência tem mostrado que até fevereiro o motorista que é parado em uma *blitz* tem 60% de chance de estar adimplente em relação ao Imposto sobre a Propriedade de Veículos Automotores (IPVA). Temos, portanto, uma probabilidade de sucesso (**o motorista não estar devendo o IPVA**) de 0,6 e uma probabilidade de estar devendo de 0,4 (**vem da diferença  $q = 1 - 0,6$** ).

## DISTRIBUIÇÃO BINOMIAL

Para que uma situação possa se enquadrar em uma distribuição binomial, deve atender as seguintes condições:

- são realizadas  $n$  repetições (tentativas) independentes;
- cada tentativa é uma prova de Bernoulli (somente podem ocorrer dois possíveis resultados); e

- a probabilidade  $p$  de sucesso em cada prova é constante.

Se uma situação atende a todas as condições anteriores, então a variável aleatória  $X$  = número de sucessos obtidos nas  $n$  tentativas terá uma distribuição binomial com  $n$  tentativas e  $p$  probabilidades de sucesso.

*Agora você deve parar a sua leitura e lançar uma moeda 30 vezes para cima. Após fazer isso e anotar os resultados, veja se o experimento que acabou de fazer se encaixa em uma distribuição binomial (condições apresentadas anteriormente).*

Simbolicamente, temos:  $X \sim B(n, p)$  com a interpretação:

A variável aleatória  $X$  tem distribuição binomial (B) com  $n$  ensaios e uma probabilidade  $p$  de sucesso (em cada ensaio).

A função de probabilidade utilizada para cálculo de probabilidades, quando a situação se enquadra na distribuição binomial, será dada por meio da seguinte expressão:

$$P(X = x) = C_n^x p^x q^{n-x} \text{ onde:}$$

$p$  é probabilidade de “sucesso” em cada ensaio;

$q = 1-p$  é a probabilidade de “fracasso” em cada ensaio;

$C_n^x = \frac{n!}{x!(n-x)!}$ , onde  $n!$  é o fatorial de  $n$ , é combinação de  $n$  valores tomados  $x$  a  $x$

Lembre-se dos conceitos de análise combinatória vistos no segundo grau!



### Exemplo

Vamos considerar que algumas pessoas entram em uma loja no período próximo ao dia das mães. Sabemos que a probabilidade de uma pessoa do gênero masculino comprar um presente é de  $1/3$ . Se entrarem quatro pessoas do gênero masculino na tal loja, qual a probabilidade de que duas venham a comprar presentes?

Se essas quatro pessoas entram na loja e duas delas compram, podemos colocar as possibilidades da seguinte forma (C → compra e não-C → não compra). O espaço amostral associado ao experimento é:

C, C, não-C, não-C **ou** C, não-C, não-C, C **ou** C, não-C, C, não-C **ou** não-C, não-C, C, C **ou** não-C, C, não-C, C **ou** não-C, C, C, não-C

Logo, calculando as probabilidades usando as regras do “e” (multiplicação, pois são independentes) e do “ou” (soma), a **probabilidade de 2 clientes do gênero masculino comprarem presentes é:**

$$p = \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} + \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} + \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} + \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} + \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3}$$

$$p = 6 \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{2}{3}$$

$$p = 6 \cdot \left(\frac{1}{3}\right)^2 \cdot \left(\frac{2}{3}\right)^2$$

$$p = \frac{24}{81} \cong 29,63\%$$

Agora, vamos calcular utilizando a função de probabilidade apresentada anteriormente e verificar que o resultado será o mesmo.

$$P(X=2) = C_4^2 \cdot \left(\frac{1}{3}\right)^2 \cdot \left(\frac{2}{3}\right)^2 = \frac{4!}{2! \cdot (4-2)!} \cdot \frac{1}{9} \cdot \frac{4}{9} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} \cdot \frac{4}{81} = \frac{24}{81} \cong 0,2963 \text{ ou } 29,63\%$$



Os valores da média e da variância da distribuição binomial são:

$$\text{Média} = n \cdot p$$

$$\text{Variância} = n \cdot p \cdot q$$

### Exemplo

Em uma determinada repartição pública, 10% das guias preenchidas estão incorretas. Essas guias correspondem a uma liberação na qual cinco guias devem estar preenchidas conjuntamente. Considere que cada guia tem a mesma probabilidade de ser preenchida incorretamente (como se houvesse repetição no experimento de retirar guias).

a) Qual a probabilidade de haver exatamente três guias incorretas nas cinco guias para liberação?

O **sucesso** é a ocorrência de guias preenchidas incorretamente.

$$p = 0,1 \quad n = 5$$

$$P(X = 3) = C_5^3 \cdot 0,1^3 \cdot 0,9^2 = 0,0081$$

b) Qual a probabilidade de haver duas ou mais guias incorretas nas cinco guias para liberação?

$$\begin{aligned} P(X \geq 2) &= P(X=2) + P(X=3) + P(X=4) + P(X=5) \\ &= 1 - [P(X=0) + P(X=1)] = 0,0815 \end{aligned}$$

c) Qual a probabilidade de um conjunto de cinco guias não apresentar nenhuma guia incorreta?

$$P(X = 0) = C_5^0 \cdot 0,1^0 \cdot 0,9^5 = 0,5905$$

*Antes de prosseguir, desta vez com o estudo da Distribuição de Poisson, você deve realizar as Atividades 1 e 2, ao final desta Unidade, para aplicar os conhecimentos já adquiridos sobre a distribuição binomial. Lembre-se de que as respostas se encontram no final do livro.*



Como na binomial são n ensaios de Bernoulli e a distribuição tem média p, a média da binomial será n.p. Raciocínio semelhante é feito para a variância.

## DISTRIBUIÇÃO DE POISSON

Você pode empregar a Distribuição de Poisson em situações nas quais não se está interessado no número de sucessos obtidos em  $n$  tentativas, como ocorre no caso da distribuição binomial, entretanto, esse número de sucessos deve estar dentro de um intervalo contínuo, ou seja, **o número de sucessos ocorridos durante um intervalo contínuo**, que pode ser um intervalo de tempo, espaço etc.

Imagine que você queira estudar o número de suicídios ocorridos em uma cidade durante um ano ou o número de acidentes automobilísticos ocorridos em uma rodovia em um mês ou o número de defeitos encontrados em um rolo de arame ovalado de 500m. Essas situações são exemplos daquelas que se enquadram na Distribuição de Poisson.

Note que nos exemplos anteriores não há como você determinar a probabilidade de ocorrência de um sucesso, mas sim a frequência média de sua ocorrência, como dois suicídios por ano, que denominaremos  $\lambda$ .

Em uma situação com essas características, a variável aleatória  $X$  = número de sucessos em um intervalo contínuo, terá uma Distribuição Poisson, com  $\lambda$  (frequência média de sucesso). Simbolicamente, podemos utilizar a notação  $\mathbf{X} \sim \mathbf{P}(\lambda)$ .

A variável aleatória  $X$  tem uma Distribuição de Poisson ( $P$ ) com uma frequência média de sucesso  $\lambda$ .

A função de probabilidade da Distribuição de Poisson será dada por meio da seguinte expressão:

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

Onde:

$e = 2,7182$  (base dos logaritmos neperianos); e

$\lambda$  corresponde a frequência média de sucesso no intervalo contínuo que se deseja calcular a probabilidade.

### Exemplo

A análise dos dados dos últimos anos de uma empresa de energia elétrica forneceu o valor médio de um blecaute por ano. Pense na probabilidade de isso ocorrer no próximo ano:

- a) Nenhum blecaute.
- b) De 2 a 4 blecautes.
- c) No máximo 2 blecautes.

Note que o exemplo afirma que a cada ano acontece em média um blecaute, ou seja, o **número de sucesso ocorrido em um intervalo contínuo**. Verificamos que a variável tem Distribuição Poisson:

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

Veja que aqui não é necessário fazer regra de três, pois as perguntas são no intervalo de um ano. Então:  $\lambda = 1$ :

$$a) P(x = 0) = \frac{e^{-1} \cdot 1^0}{0!} = \frac{0,3679 \cdot 1}{1} = 0,3679 \text{ ou } 36,79\%$$

$$\begin{aligned} b) P(x = 2) + P(x = 3) + P(x = 4) &= \frac{e^{-1} \cdot 1^2}{2!} + \frac{e^{-1} \cdot 1^3}{3!} + \frac{e^{-1} \cdot 1^4}{4!} = \\ &= 0,1839 + 0,061 + 0,015 \\ &= 0,2599 \text{ ou } 25,99\% \end{aligned}$$

- c) Como já temos os valores de  $x = 0$  e  $x = 2$  basta calcularmos para  $x = 1$  e somarmos os resultados.

$$P(x = 1) = \frac{e^{-1} \cdot 1^1}{1!} = \frac{0,3679 \cdot 1}{1} = 0,3679 \text{ ou } 36,79\%$$

$$\begin{aligned} P(x \leq 2) &= P(x=0) + P(x=1) + P(x=2) = \\ &= 0,3679 + 0,3679 + 0,1839 = 0,9197 \text{ ou } 91,97\% \end{aligned}$$

Vejamos uma aplicação da Distribuição de Poisson considerando que o Corpo de Bombeiros de uma determinada cidade recebe, em média, três chamadas por dia. Queremos saber, então, qual a probabilidade do Corpo de Bombeiros receber:

- a) **4 chamadas em um dia:** verificamos que a variável tem Distribuição Poisson, pois temos número de chamadas (variável discreta) por dia (**intervalo contínuo**). A probabilidade será calculada por meio da expressão:

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

Como não é necessário fazer regra de três, pois as perguntas são no intervalo de um dia, então:  $\lambda = 3$ . Substituindo na expressão, teremos:

$$P(X = 4) = e^{-3} \frac{3^4}{4!} = 0,1680$$

- b) **Nenhuma chamada em um dia:** nesse caso, o intervalo continua sendo um dia. Logo, o  $\lambda$  continua sendo o mesmo, ou seja,  $\lambda = 3$ . Substituindo então na expressão, teremos:

$$P(X = 0) = e^{-3} \frac{3^0}{0!} = 0,0498$$

- c) **20 chamadas em uma semana:** nesse caso o intervalo em que se deseja calcular a probabilidade é de uma semana, ou seja, sete dias. Então, em uma semana, a frequência média de chamadas será de 7 dias vezes 3 chamadas/dia:

$$\lambda = 21 \text{ chamadas por semana}$$

Substituindo os valores, teremos a seguinte probabilidade:

$$P(X = 20) = e^{-21} \frac{21^{20}}{20!} = 0,0867$$

Como o intervalo em que se deseja calcular a probabilidade é um dia, o  $\lambda$  será igual a 3.

Uma característica da Distribuição de Poisson é que as estatísticas da distribuição (média e variância) apresentam o mesmo valor, ou seja, são iguais a  $\lambda$ . Então, teremos:

$$\text{Média} = \text{Variância} = \lambda$$

*Antes de discutir as distribuições contínuas, vamos aplicar os conhecimentos relacionados à Distribuição de Poisson realizando a Atividade 3 ao final desta Unidade. É importante salientarmos que nesta Unidade a resolução das atividades de aprendizagem serão solicitadas ao longo do texto para facilitar a sua compreensão dos conceitos e de como utilizá-los.*

## DISTRIBUIÇÕES CONTÍNUAS

Dentre as várias distribuições de probabilidade contínuas, abordaremos aqui apenas a distribuição normal, pois ela apresenta grande aplicação em pesquisas científicas e tecnológicas. Grande parte das variáveis contínuas de interesse prático segue essa distribuição, aliada ao Teorema do Limite Central (TLC), que é a base das estimativas e dos testes de hipóteses realizados sobre a média de uma população qualquer, e garante que a distribuição amostral das médias segue uma distribuição normal, independentemente da distribuição da variável em estudo, como será visto mais adiante.

### DISTRIBUIÇÃO NORMAL

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

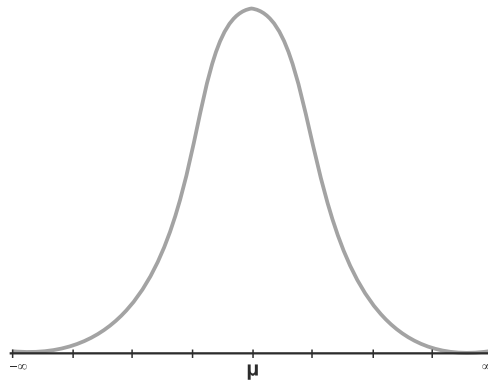
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, x \in R.$$

Onde:

$\mu$  e  $\sigma$  são a média e o desvio padrão, respectivamente, da distribuição de probabilidade.

$\pi$  corresponde a 3,1415 e  $\exp$  a uma função exponencial.

O gráfico da distribuição normal, utilizando a função mostrada anteriormente e os conceitos vistos nas disciplinas *Matemática Básica* e *Matemática para Administradores*, é dado por:

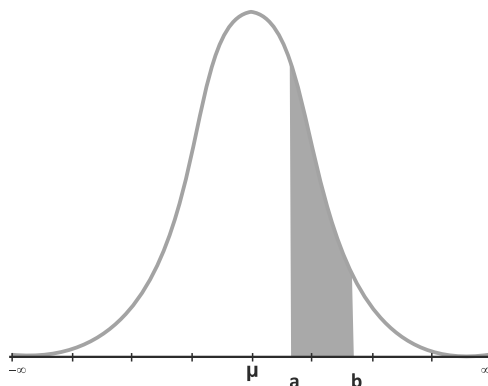


Você encontrará a seguir as principais propriedades da distribuição normal:

- ▶ é simétrica em relação ao ponto  $x = \mu$  (50% abaixo e 50% acima da média);
- ▶ tem forma **campanular\***;
- ▶ as três medidas de posição – média, mediana e moda – se confundem no ponto máximo da curva ( $x = \mu$ );
- ▶ fica perfeitamente definida conhecendo-se a média e o desvio padrão, pois outros termos da função são constantes; e
- ▶ toda a área compreendida entre a curva e o eixo  $x$  é igual a 1 (conceito de probabilidades).

\* Campanular – relativo à campânula; objeto em forma de sino. Fonte: Houaiss (2009).

Portanto, a área sob a curva entre os pontos  $a$  e  $b$ , em que  $a < b$ , representa a probabilidade da variável  $X$  assumir um valor entre  $a$  e  $b$  (área escura), como observamos a seguir.



Desse modo, você pode associar que, no caso das distribuições contínuas, a área do gráfico corresponde a probabilidades.

Então, veja a notação utilizada para a distribuição normal:

$$X \sim N(\mu, \sigma^2)$$

A variável  $x$  tem distribuição normal com média  $\mu$  e variância  $\sigma^2$ .

Para calcularmos as probabilidades via distribuição normal, é necessário o conhecimento de cálculo integral. Assim, procuramos tabular os valores de probabilidade que seriam obtidos por meio da integração da função densidade de probabilidade normal em um determinado intervalo.

A dificuldade para se processar esse tabelamento se prendeu na infinidade de valores que  $\mu$  (média) e  $\sigma$  (desvio padrão) poderiam assumir. Nessas condições, teríamos que dispor de uma tabela para cada uma das infinitas combinações de  $\mu$  e  $\sigma$ , ou seja, em cada situação que se quisesse calcular uma probabilidade.

Para resolver esse problema, podemos obter uma nova forma para a distribuição normal, que não seja influenciada por  $\mu$  e  $\sigma$ . O problema foi solucionado mediante emprego de uma nova variável definida por:

$$Z = \frac{X - \mu}{\sigma}$$



Essa variável transforma todas as distribuições normais em uma distribuição normal reduzida ou padronizada, de média zero e desvio padrão um. Então, temos:  $Z \sim N(0,1)$ .

Assim, utilizamos apenas uma tabela para o cálculo de probabilidades para qualquer que seja a curva correspondente a uma distribuição normal.

Portanto, para um valor de  $x = \mu$  em uma distribuição normal qualquer, corresponde o valor:

$$z = \frac{x - \mu}{\sigma} = \frac{\mu - \mu}{\sigma} = 0 \text{ na distribuição normal reduzida.}$$

Para  $x = \mu + \sigma$ , temos:

$$z = \frac{x - \mu}{\sigma} = \frac{\mu + \sigma - \mu}{\sigma} = \frac{\sigma}{\sigma} = 1 \text{ e, assim por diante.}$$

Podemos definir a distribuição normal reduzida ou padronizada como sendo uma distribuição da variável  $Z$  que apresenta distribuição normal com média zero e variância 1 ( $Z \sim N(0;1)$ ).

Na Tabela 15, que apresenta a distribuição normal padronizada, as áreas ou probabilidades fornecidas estão entre zero e o valor de  $Z$ , como vemos a seguir.

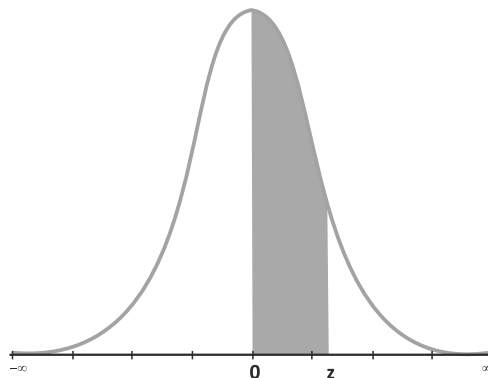


Tabela 15: Área sob a curva normal padronizada compreendida entre os valores 0 e Z

Z	0	1	2	3	4	5	6	7	8	9
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

Fonte: Elaborada pelo autor

Veja que na Tabela 15 os valores apresentados na primeira coluna correspondem a parte inteira e decimal do valor de Z (por exemplo 1,5), enquanto os valores da primeira linha correspondem a parte centesimal (por exemplo 8). Assim, teremos o valor de  $Z = 1,58$ . Já os valores encontrados no meio da tabela correspondem às probabilidades dos respectivos valores compreendidos entre zero e Z.

Para que você possa entender a utilização da distribuição normal, vamos considerar a arrecadação como um tributo de uma pequena cidade. Verificamos que essa arrecadação seguia uma distribuição normal com duração média de R\$ 60.000,00 e desvio padrão de R\$ 10.000,00. Procuramos, então, responder os seguintes questionamentos:

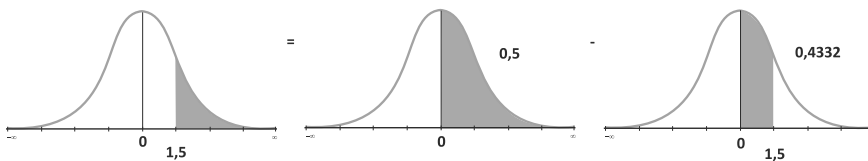
- a) Qual a probabilidade de uma arrecadação ser maior do que R\$ 75.000,00?

Como a variável arrecadação apresenta distribuição aproximadamente normal com média 60000 e variância de  $10000^2$  [ $X \sim N(60000; 10000^2)$ ] e procura-se calcular a  $P(X > 75000) = ?$

Primeiramente, precisamos transformar a variável  $X$  em  $Z$  e, depois, substituindo na expressão, teremos:

$$z = \frac{x - \mu}{\sigma} = \frac{75000 - 60000}{10000} = 1,50$$

Olhando esse valor na Tabela 15,  $z = 1,50$  (1,5 na primeira coluna e o zero na primeira linha), encontraremos no meio da tabela o valor de 0,4332 que corresponde à probabilidade de  $z$  estar entre zero e 1,5, como você pode observar a seguir.



A área escura da figura corresponde a  $P(X > 75000)$ , que é a mesma coisa que:  $P(z > 1,50)$ . Então:

$$P(z > 1,50) \text{ [Figura 1]} = P(0 < z < +\infty) \text{ [Figura 2]} -$$

$$P(0 < z < 1,50) \text{ [Figura 3]} = 0,5 - 0,4332 = 0,0668.$$

Retirou-se a probabilidade encontrada de 0,5, pois esse valor corresponde à probabilidade de zero até o infinito.

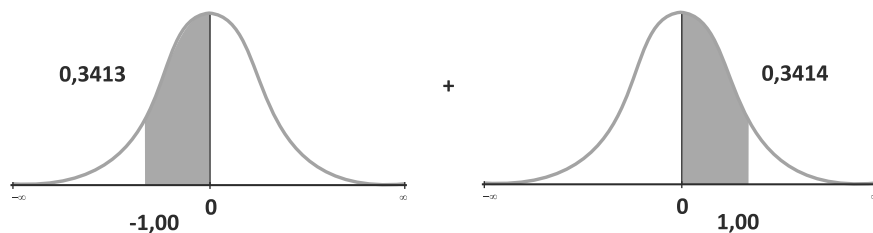
- b) Qual a probabilidade da arrecadação estar entre R\$ 50.000,00 e R\$ 70.000,00?

$$P(50000 < X < 70000) = ?$$

Primeiramente, precisamos transformar a variável  $X$  em  $Z$  e, depois, substituindo na expressão de  $Z$ , teremos valores de  $Z_1$  e  $Z_2$ , relacionados aos valores de  $X_1=50000$  e  $X_2=70000$ :

$$z_1 = \frac{x - \mu}{\sigma} = \frac{50000 - 60000}{10000} = -1,00$$

$$z_2 = \frac{x - \mu}{\sigma} = \frac{70000 - 60000}{10000} = 1,00$$



Podemos verificar que:

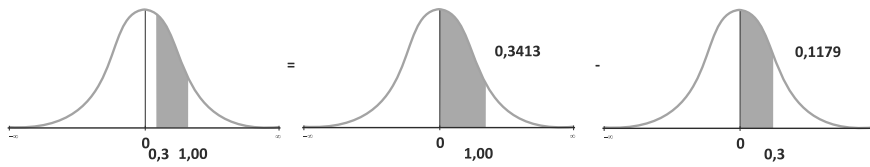
$$P(50000 < X < 70000) = P(-1,00 < z < 1,00) = 0,3413 + 0,3413 = 0,6826$$

- c) Qual a probabilidade da arrecadação estar entre R\$ 63.000,00 e R\$ 70.000,00?

$$P(63000 < X < 70000) = ?$$

$$z_1 = \frac{x - \mu}{\sigma} = \frac{63000 - 60000}{10000} = 0,30$$

$$z_2 = \frac{x - \mu}{\sigma} = \frac{70000 - 60000}{10000} = 1,00$$



$$P(63000 < X < 70000) = P(0.30 < z < 1.00) = 0.3413 - 0.1179 = 0.2234$$

Destacamos que existem outras distribuições tanto discretas quanto contínuas que não foram abordadas neste livro. Portanto, recomendamos que você procure outras fontes de conhecimento, a começar por fazer uma pesquisa na internet sobre essas distribuições.

*Antes de prosseguir, você deve realizar as Atividades 4 e 5 ao final desta Unidade, na qual você terá a oportunidade de verificar o seu grau de compreensão sobre a distribuição normal.*

## DISTRIBUIÇÕES AMOSTRAIS

Com as distribuições amostrais, você pode inferir propriedades de um agregado maior (a população) a partir de um conjunto menor (a amostra), ou seja, inferir sobre parâmetros populacionais, dispondo apenas de estatísticas amostrais. Portanto, torna-se necessário um estudo detalhado das distribuições amostrais, que são base para intervalos de confiança e testes de hipóteses.

Para que você tenha condições de fazer afirmações sobre um determinado parâmetro populacional (ex:  $\mu$ ), baseados na estimativa  $\bar{x}$ , obtida a partir dos dados amostrais, é necessário conhecer a relação existente entre  $\bar{x}$  e  $\mu$ , isto é, o comportamento de  $\bar{x}$ , quando se extraem todas as amostras possíveis da população, ou seja, sua distribuição amostral.

Para obtermos a distribuição amostral de um estimador, é necessário conhecer o processo pelo qual as amostras foram retiradas, isto é, se amostras foram retiradas **com reposição** ou **sem reposição**. Neste material, iremos considerar apenas as situações de amostragens com reposição.

Dessa forma, a partir do comportamento da estatística amostral, podemos aplicar um teorema muito conhecido na estatística como Teorema do Limite Central (TLC). Esse teorema propõe que, se retirarmos todas as possíveis amostras de tamanho  $n$  de uma população, independente de sua distribuição, e verificarmos como as estatísticas amostrais obtidas se distribuem, teremos uma distribuição **aproximadamente normal**, com  $\mu_{\bar{x}} = \mu$  (**média das médias amostrais igual à média populacional**) e variância das médias  $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$  (**variância das médias**

**mostrais igual à variância da população dividida pelo tamanho da amostra**), independentemente da distribuição da variável em questão.

Portanto, considerando a distribuição **amostral** de médias, quando se conhece a variância populacional ou a amostra é grande ( $n > 30$ ), utilizamos a estatística  $z$  da distribuição normal vista anteriormente, independentemente da distribuição da população.

Então, por meio do TLC, a estatística será dada por:  $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ .



Confira a indicação de um programa para cálculo amostral na seção Complementando ao final desta Unidade.

## DISTRIBUIÇÃO $t$ DE *STUDENT*

Na prática, muitas vezes não se conhece  $\sigma^2$  e trabalha-se com amostras pequenas, ou seja, menor ou igual a 30. Assim, você conhece apenas sua estimativa  $s$  (desvio padrão amostral). Substituindo  $\sigma$  por seu estimador  $s$ , na expressão da variável padronizada, obtemos a seguinte variável:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \quad (\text{expressão semelhante a } Z)$$

Essa variável segue uma distribuição  $t$  de *Student* com  $(n - 1)$  **graus de liberdade**\*.

O  $n - 1$  corresponde ao divisor do cálculo da variância amostral, ou seja, o número de variáveis na amostra que variam livremente na definição da estatística.

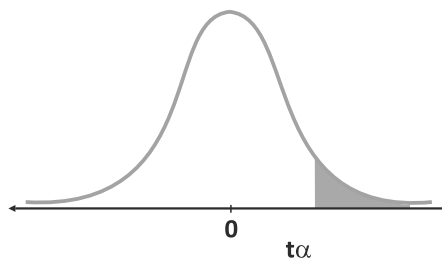
A distribuição  $t$  de *Student* apresenta as seguintes características:

- ▶ é simétrica em relação à média, que é zero;
- ▶ tem forma campanular (semelhante a normal);

\* Graus de liberdade (GL)  
 - é o número de determinações independentes (dimensão da amostra) menos o número de parâmetros estatísticos a serem avaliados na população. Fonte: Elaborado pelo autor.

- ▶ quando  $n$  tende para infinito, a distribuição  $t$  tende para a distribuição normal. Na prática, a aproximação é considerada boa quando  $n > 30$ ; e
- ▶ possui  $n-1$  graus de liberdade.

Vamos aprender a utilizar a tabela da distribuição de  $t$  de *Student*. Na tabela  $t$  de *Student*, na primeira linha, temos o valor de  **$\alpha$  que corresponde à probabilidade (área) acima de um determinado valor da tabela**. Veja a seguir o conceito de  **$\alpha$  (área mais escura)**.



Observe que na Tabela 16, a seguir, temos, na primeira coluna, os graus de liberdade (GL) e, no centro da tabela, os valores da **estatística  $t$  de *Student***. Na primeira linha, temos os valores de  $\alpha$ .



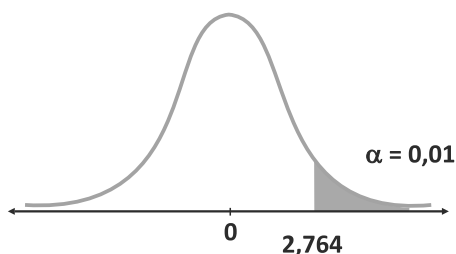
Tabela 16: Limites unilaterais da distribuição t de *Student*  
ao nível  $\alpha$  de probabilidade

GL	$\alpha$								
	0.250	0.200	0.150	0.100	0.050	0.025	0.010	0.005	0.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.656	318.289
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.328
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.214
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.894
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.261
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.160
240	0.676	0.843	1.039	1.285	1.651	1.970	2.342	2.596	3.125
480	0.675	0.842	1.038	1.283	1.648	1.965	2.334	2.586	3.107
700	0.675	0.842	1.037	1.283	1.647	1.963	2.332	2.583	3.102
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098

Fonte: Elaborada pelo autor

Para exemplificar o uso da tabela, considere que desejamos encontrar a probabilidade ser maior do que um valor de t igual a 2,764 trabalhando com uma amostra de tamanho  $n = 11$ . Portanto, teremos 10 graus de liberdade e, nessa linha, procuremos o valor que desejamos encontrar: 2,764. Subindo na tabela em direção

ao  $\alpha$ , encontraremos um valor de 0,01 na primeira linha, ou seja, essa é a probabilidade de ser maior do que 2,764 com 10 graus de liberdade.

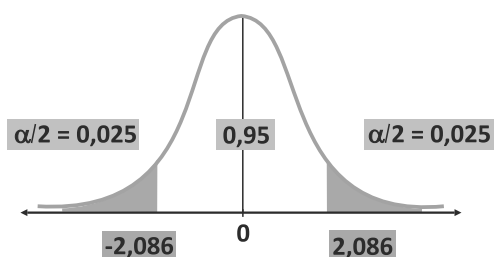


Vamos resolver outro exemplo:

Encontre o valor de  $t$  tal que a probabilidade de  $t$  (distribuição) esteja entre  $-t$  e  $t$  e seja igual a 0,95 com 20 graus de liberdade. Isso pode ser representado da forma a seguir:

$$t / P(-t < t < t) = 0,95 \text{ com } 20 \text{ gl}$$

A área do meio corresponde a uma probabilidade de 0,95. Então, como a probabilidade total é igual a 1, sobrou 0,05 de probabilidade para ser dividida pelas áreas do lado direito e esquerdo. Observando o valor de  $\alpha/2 = 0,025$  (área à direita do valor tabelado) na tabela de  $t$  de *Student* e com 20 graus de liberdade, encontraremos o valor de 2,086. Do outro lado, teremos um valor negativo, pois ele está à esquerda da média igual a zero, como você pode ver a seguir.

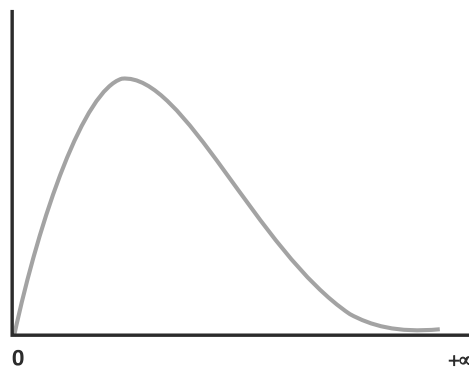


## DISTRIBUIÇÃO DE QUI-QUADRADO

Retirando uma amostra de  $n$  elementos de uma população normal com média  $\mu$  e variância  $\sigma^2$ , podemos demonstrar que a distribuição amostral da variância amostral segue uma **distribuição de  $\chi^2$  (qui-quadrado)** com  $n-1$  graus de liberdade. A variável da estatística de qui-quadrado será dada por:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \text{ tem distribuição } \chi^2 \text{ com } n-1 \text{ graus de liberdade.}$$

Essa distribuição é sempre positiva, o que pode ser comprovado pela própria definição da variável. É, ainda, assimétrica à direita, como você pode ver no gráfico da distribuição, a seguir.



Por meio da Tabela 17, você pode ver como é feita a utilização da distribuição de qui-quadrado com graus de liberdade (GL).

Tabela 17: Limites unilaterais da distribuição de  $\chi^2$  ao nível  $\alpha$  de probabilidade

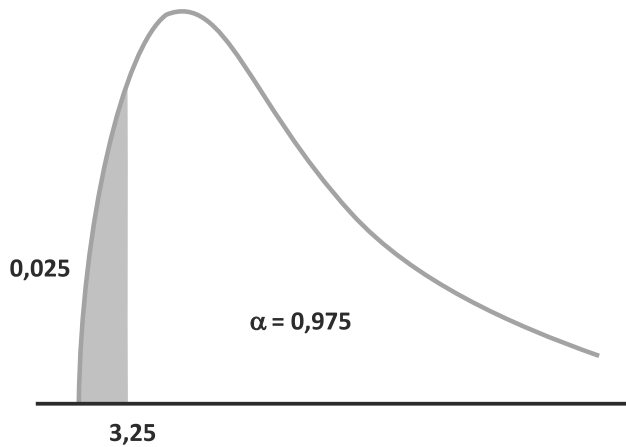
GL	$\alpha$												
	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005
1	0.0000	0.0002	0.0010	0.0039	0.0158	0.1015	0.4549	1.3233	2.7055	3.8415	5.0239	6.6349	7.8794
2	0.0100	0.0201	0.0506	0.1026	0.2107	0.5754	1.3863	2.7726	4.6052	5.9915	7.3778	9.2104	10.5965
3	0.0717	0.1148	0.2158	0.3518	0.5844	1.2125	2.3660	4.1083	6.2514	7.8147	9.3484	11.3449	12.8381
4	0.2070	0.2971	0.4844	0.7107	1.0636	1.9226	3.3567	5.3853	7.7794	9.4877	11.1433	13.2767	14.8602
5	0.4118	0.5543	0.8312	1.1455	1.6103	2.6746	4.3515	6.6257	9.2363	11.0705	12.8325	15.0863	16.7496
6	0.6757	0.8721	1.2373	1.6354	2.2041	3.4546	5.3481	7.8408	10.6446	12.5916	14.4494	16.8119	18.5475
7	0.9893	1.2390	1.6899	2.1673	2.8331	4.2549	6.3458	9.0371	12.0170	14.0671	16.0128	18.4753	20.2777
8	1.3444	1.6465	2.1797	2.7326	3.4895	5.0706	7.3441	10.2189	13.3616	15.5073	17.5345	20.0902	21.9549
9	1.7349	2.0879	2.7004	3.3251	4.1682	5.8988	8.3428	11.3887	14.6837	16.9190	19.0228	21.6660	23.5893
10	2.1558	2.5582	3.2470	3.9403	4.8652	6.7372	9.3418	12.5489	15.9872	18.3070	20.4832	23.2093	25.1881
11	2.6032	3.0535	3.8157	4.5748	5.5778	7.5841	10.3410	13.7007	17.2750	19.6752	21.9200	24.7250	26.7569
12	3.0738	3.5706	4.4038	5.2260	6.3038	8.4384	11.3403	14.8454	18.5493	21.0261	23.3367	26.2170	28.2997
13	3.5650	4.1069	5.0087	5.8919	7.0415	9.2991	12.3398	15.9839	19.8119	22.3620	24.7356	27.6882	29.8193
14	4.0747	4.6604	5.6287	6.5706	7.7895	10.1653	13.3393	17.1169	21.0641	23.6848	26.1189	29.1412	31.3194
15	4.6009	5.2294	6.2621	7.2609	8.5468	11.0365	14.3389	18.2451	22.3071	24.9958	27.4884	30.5780	32.8015
16	5.1422	5.8122	6.9077	7.9616	9.3122	11.9122	15.3385	19.3689	23.5418	26.2962	28.8453	31.9999	34.2671
17	5.6973	6.4077	7.5642	8.6718	10.0852	12.7919	16.3382	20.4887	24.7690	27.5871	30.1910	33.4087	35.7184
18	6.2648	7.0149	8.2307	9.3904	10.8649	13.6753	17.3379	21.6049	25.9894	28.8693	31.5264	34.8052	37.1564
19	6.8439	7.6327	8.9065	10.1170	11.6509	14.5620	18.3376	22.7178	27.2036	30.1435	32.8523	36.1908	38.5821
20	7.4338	8.2604	9.5908	10.8508	12.4426	15.4518	19.3374	23.8277	28.4120	31.4104	34.1696	37.5663	39.9969
21	8.0336	8.8972	10.2829	11.5913	13.2396	16.3444	20.3372	24.9348	29.6151	32.6706	35.4789	38.9322	41.4009
22	8.6427	9.5425	10.9823	12.3380	14.0415	17.2396	21.3370	26.0393	30.8133	33.9245	36.7807	40.2894	42.7957
23	9.2604	10.1957	11.6885	13.0905	14.8480	18.1373	22.3369	27.1413	32.0069	35.1725	38.0756	41.6383	44.1814
24	9.8862	10.8563	12.4011	13.8484	15.6587	19.0373	23.3367	28.2412	33.1962	36.4150	39.3641	42.9798	45.5584
25	10.5196	11.5240	13.1197	14.6114	16.4734	19.9393	24.3366	29.3388	34.3816	37.6525	40.6465	44.3140	46.9280
26	11.1602	12.1982	13.8439	15.3792	17.2919	20.8434	25.3365	30.4346	35.5632	38.8851	41.9231	45.6416	48.2898
27	11.8077	12.8785	14.5734	16.1514	18.1139	21.7494	26.3363	31.5284	36.7412	40.1133	43.1945	46.9628	49.6450
28	12.4613	13.5647	15.3079	16.9279	18.9392	22.6572	27.3362	32.6205	37.9159	41.3372	44.4608	48.2782	50.9936
29	13.1211	14.2564	16.0471	17.7084	19.7677	23.5666	28.3361	33.7109	39.0875	42.5569	45.7223	49.5878	52.3355
30	13.7867	14.9535	16.7908	18.4927	20.5992	24.4776	29.3360	34.7997	40.2560	43.7730	46.9792	50.8922	53.6719
40	20.7066	22.1642	24.4331	26.5093	29.0505	33.6603	39.3353	45.6160	51.8050	55.7585	59.3417	63.6908	66.7660
50	27.9908	29.7067	32.3574	34.7642	37.6886	42.9421	49.3349	56.3336	63.1671	67.5048	71.4202	76.1538	79.4898
60	35.5344	37.4848	40.4817	43.1880	46.4589	52.2938	59.3347	66.9815	74.3970	79.0820	83.2977	88.3794	91.9518
100	67.3275	70.0650	74.2219	77.9294	82.3581	90.1332	99.3341	109.1412	118.4980	124.3421	129.5613	135.8069	140.1697
120	83.8517	86.9233	91.5726	95.7046	100.6236	109.2197	119.3340	130.0546	140.2326	146.5673	152.2113	158.9500	163.6485

Fonte: Elaborada pelo autor

Para obter probabilidades ou o valor da estatística de qui-quadrado, você irá proceder do mesmo modo que na tabela da distribuição t de *Student*. Na primeira linha, temos os valores de  $\alpha$ , na primeira coluna temos os graus de liberdade e no meio da tabela temos os valores da estatística de qui-quadrado.

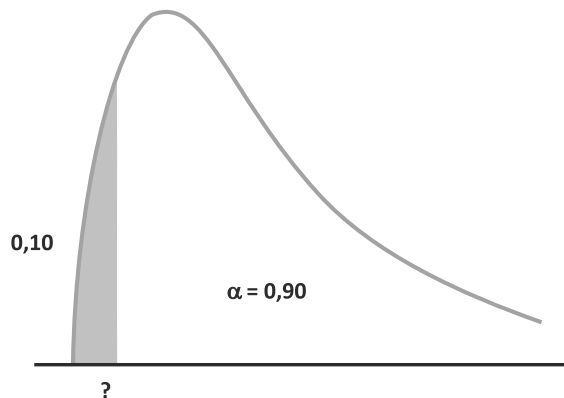
*Vamos, então, aprender a olhar a tabela de qui-quadrado.*

Encontre a probabilidade de o valor de qui-quadrado ser maior do que 3,25 com 10 graus de liberdade, ou seja,  $P(\chi^2 > 3,25) = ?$



Para 10 graus de liberdade e um valor de 3,25 (valor aproximado) na tabela, encontraremos na parte superior um valor de  $\alpha = 0,975$ , que corresponde à probabilidade procurada.

Agora, sabemos que a probabilidade de ser maior que um determinado valor de qui-quadrado é igual a 0,90 ( $P(x^2 > ?) = 0,9$  com 15 graus de liberdade. Então, o valor da interrogação (?) será obtido na tabela de qui-quadrado.



Observando a tabela de qui-quadrado com 15 graus de liberdade e um valor de  $\alpha = 0,90$ , encontraremos no meio da tabela um valor de 8,55, que será o valor de qui-quadrado, cuja probabilidade de ser maior do que ele é de 0,90 ( $\alpha$ ).

## DISTRIBUIÇÃO DE F

A **distribuição de F de Fischer-Snedecor** corresponde à distribuição da razão de duas variâncias. Temos, então, duas populações que apresentam variâncias populacionais e delas são retiradas amostras nas quais são calculadas variâncias amostrais. A relação entre essas variâncias é que nos dá a distribuição de F. A estatística da distribuição é apresentada a seguir:

$$F = \frac{\frac{s_A^2}{\sigma_A^2}}{\frac{s_B^2}{\sigma_B^2}}$$

Segue uma distribuição F com  $v_1 = n_1 - 1$  e  $v_2 = n_2 - 1$  graus de liberdade para o numerador e o denominador, respectivamente.

Uma das tabelas de F de Snedecor é apresentada a seguir:

Tabela 18: Limites unilaterais da distribuição F de Fischer–Snedecor ao nível de 10% de probabilidade

GL	v1																							
v2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	40	60	120	240				
1	39.864	49.500	53.593	55.833	57.240	58.204	58.906	59.439	59.857	60.195	60.473	60.705	60.902	61.073	61.220	61.740	62.529	62.794	63.061	63.194				
2	8.526	9.000	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392	9.401	9.408	9.415	9.420	9.425	9.441	9.466	9.475	9.483	9.487				
3	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.240	5.230	5.222	5.216	5.210	5.205	5.200	5.184	5.160	5.151	5.143	5.138				
4	4.545	4.325	4.191	4.107	4.051	4.010	3.979	3.955	3.936	3.920	3.907	3.896	3.886	3.878	3.870	3.844	3.804	3.790	3.775	3.768				
5	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316	3.297	3.282	3.268	3.257	3.247	3.238	3.207	3.157	3.140	3.123	3.114				
6	3.776	3.463	3.289	3.181	3.108	3.055	3.014	2.983	2.958	2.937	2.920	2.905	2.892	2.881	2.871	2.836	2.781	2.762	2.742	2.732				
7	3.589	3.257	3.074	2.961	2.883	2.827	2.785	2.752	2.725	2.703	2.684	2.668	2.654	2.643	2.632	2.595	2.535	2.514	2.493	2.482				
8	3.458	3.113	2.924	2.806	2.726	2.668	2.624	2.589	2.561	2.538	2.519	2.502	2.488	2.475	2.464	2.425	2.361	2.339	2.316	2.304				
9	3.360	3.006	2.813	2.693	2.611	2.551	2.505	2.469	2.440	2.416	2.396	2.379	2.364	2.351	2.340	2.298	2.232	2.208	2.184	2.172				
10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323	2.302	2.284	2.269	2.255	2.244	2.201	2.132	2.107	2.082	2.069				
11	3.225	2.860	2.660	2.536	2.451	2.389	2.342	2.304	2.274	2.248	2.227	2.209	2.193	2.179	2.167	2.123	2.052	2.026	2.000	1.986				
12	3.177	2.807	2.606	2.480	2.394	2.331	2.283	2.245	2.214	2.188	2.166	2.147	2.131	2.117	2.105	2.060	1.986	1.960	1.932	1.918				
13	3.136	2.763	2.560	2.434	2.347	2.283	2.234	2.195	2.164	2.138	2.116	2.097	2.080	2.066	2.053	2.007	1.931	1.904	1.876	1.861				
14	3.102	2.726	2.522	2.395	2.307	2.243	2.193	2.154	2.122	2.095	2.073	2.054	2.037	2.022	2.010	1.962	1.885	1.857	1.828	1.813				
15	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086	2.059	2.037	2.017	2.000	1.985	1.972	1.924	1.845	1.817	1.787	1.771				
16	3.048	2.668	2.462	2.333	2.244	2.178	2.128	2.088	2.055	2.028	2.005	1.985	1.968	1.953	1.940	1.891	1.811	1.782	1.751	1.735				
17	3.026	2.645	2.437	2.308	2.218	2.152	2.102	2.061	2.028	2.001	1.978	1.958	1.940	1.925	1.912	1.862	1.781	1.751	1.719	1.703				
18	3.007	2.624	2.416	2.286	2.196	2.130	2.079	2.038	2.005	1.977	1.954	1.933	1.916	1.900	1.887	1.837	1.754	1.723	1.691	1.674				
19	2.990	2.606	2.397	2.266	2.176	2.109	2.058	2.017	1.984	1.956	1.932	1.912	1.894	1.878	1.865	1.814	1.730	1.699	1.666	1.649				
20	2.975	2.589	2.380	2.249	2.158	2.091	2.040	1.999	1.965	1.937	1.913	1.892	1.875	1.859	1.845	1.794	1.708	1.677	1.643	1.626				
21	2.961	2.575	2.365	2.233	2.142	2.075	2.023	1.982	1.948	1.920	1.896	1.875	1.857	1.841	1.827	1.776	1.689	1.657	1.623	1.605				
22	2.949	2.561	2.351	2.219	2.128	2.060	2.008	1.967	1.933	1.904	1.880	1.859	1.841	1.825	1.811	1.759	1.671	1.639	1.604	1.586				
23	2.937	2.549	2.339	2.207	2.115	2.047	1.995	1.953	1.919	1.890	1.866	1.845	1.827	1.811	1.796	1.744	1.655	1.622	1.587	1.568				
24	2.927	2.538	2.327	2.195	2.103	2.035	1.983	1.941	1.906	1.877	1.853	1.832	1.814	1.797	1.783	1.730	1.641	1.607	1.571	1.552				
25	2.918	2.528	2.317	2.184	2.092	2.024	1.971	1.929	1.895	1.866	1.841	1.820	1.802	1.785	1.771	1.718	1.627	1.593	1.557	1.538				
26	2.909	2.519	2.307	2.174	2.082	2.014	1.961	1.919	1.884	1.855	1.830	1.809	1.790	1.774	1.760	1.706	1.615	1.581	1.544	1.524				
27	2.901	2.511	2.299	2.165	2.073	2.005	1.952	1.909	1.874	1.845	1.820	1.799	1.780	1.764	1.749	1.695	1.603	1.569	1.531	1.511				
28	2.894	2.503	2.291	2.157	2.064	1.996	1.943	1.900	1.865	1.836	1.811	1.790	1.771	1.754	1.740	1.685	1.592	1.558	1.520	1.500				
29	2.887	2.495	2.283	2.149	2.057	1.988	1.935	1.892	1.857	1.827	1.802	1.781	1.762	1.745	1.731	1.676	1.583	1.547	1.509	1.489				
30	2.881	2.489	2.276	2.142	2.049	1.980	1.927	1.884	1.849	1.819	1.794	1.773	1.754	1.737	1.722	1.667	1.573	1.538	1.499	1.478				
40	2.835	2.440	2.226	2.091	1.997	1.927	1.873	1.829	1.793	1.763	1.737	1.715	1.695	1.678	1.662	1.605	1.506	1.467	1.425	1.402				
50	2.809	2.412	2.197	2.061	1.966	1.895	1.840	1.796	1.760	1.729	1.703	1.680	1.660	1.643	1.627	1.568	1.465	1.424	1.379	1.354				
60	2.791	2.393	2.177	2.041	1.946	1.875	1.819	1.775	1.738	1.707	1.680	1.657	1.637	1.619	1.603	1.543	1.437	1.395	1.348	1.321				
80	2.769	2.370	2.154	2.016	1.921	1.849	1.793	1.748	1.711	1.680	1.653	1.629	1.609	1.590	1.574	1.513	1.403	1.358	1.307	1.278				
100	2.756	2.356	2.139	2.002	1.906	1.834	1.778	1.732	1.695	1.663	1.636	1.612	1.592	1.573	1.557	1.494	1.382	1.336	1.282	1.250				
120	2.748	2.347	2.130	1.992	1.896	1.824	1.767	1.722	1.684	1.652	1.625	1.601	1.580	1.562	1.545	1.482	1.368	1.320	1.265	1.232				
240	2.727	2.325	2.107	1.968	1.871	1.799	1.742	1.696	1.658	1.625	1.598	1.573	1.552	1.533	1.516	1.451	1.332	1.281	1.219	1.180				

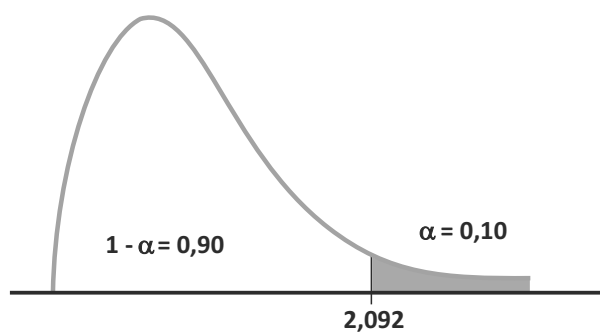
Fonte: Elaborada pelo autor

Note que, no caso da tabela de F, o valor de  $\alpha$  que corresponde à área extrema à direita da curva é apresentado no título da tabela, pois, para cada valor de  $\alpha$ , temos uma tabela diferente.

Encontramos uma aplicação prática da distribuição de F na verificação da homogeneidade das variâncias provenientes de duas populações normais e independentes. Então, encontre o valor de F1 cuja probabilidade de ser maior do que ele é 0,10 com 5 e 25 graus de liberdade, ou seja,  $P(F > F1) = 0,10$  com  $v_1 = 5$  e  $v_2 = 25$  gl.

Como temos a probabilidade do resultado ser maior do que um valor de F, esse valor corresponde ao valor de  $\alpha$ . Precisaremos, então, trabalhar com a tabela que apresenta 10% de probabilidade no título, como a Tabela 18.

Observando  $v_1 = 5$  e  $v_2 = 25$ , encontraremos um valor de F igual a 2,092.





## NOÇÕES DE ESTIMAÇÃO

Um dos principais objetivos da estatística inferencial consiste em estimar os valores de parâmetros populacionais desconhecidos (estimação de parâmetros) utilizando dados amostrais. Então, qualquer característica de uma população pode ser estimada a partir de uma amostra aleatória, desde que esta amostra represente bem a **população**.

A estatística inferencial apresenta uma relevância alta, já que a utilização de dados amostrais está associada à maioria das decisões que um gestor ou um pesquisador deve tomar. Consiste em tirar conclusões de uma população a partir de amostra representativa dessa população, tendo isso grande importância em muitas áreas do conhecimento.

A partir de uma amostra de 800 clientes (escolhidos aleatoriamente entre todos os clientes que abasteceram na primeira quinzena de um determinado mês) de um posto de gasolina que possuem carros populares, verificou-se que o consumo médio de gasolina foi de R\$ 200,00 por quinzena.

*Refleta sobre a afirmação a seguir:*

Podemos inferir que o consumo médio da população de clientes da primeira quinzena do mês em estudo, proprietários de carros populares que abastecem nesse posto de gasolina, é de R\$ 200,00.



Os parâmetros populacionais mais comuns a serem estimados são: a média, o desvio padrão e a proporção.

Esta é uma estimativa que chamamos de **pontual**, ou seja, inferimos sobre a população considerando apenas o valor da estimativa. Essas estimativas por ponto não nos dão uma ideia sobre confiança e sobre as margens de erro que deveriam ser aplicadas ao resultado. Tudo que nós sabemos, por exemplo, é que o consumo médio de gasolina foi estimado em R\$ 200,00 por quinzena, independentemente do tamanho da amostra e da variabilidade inerente aos dados. Se fosse usado um tamanho grande de amostra e houvesse pouca variabilidade, teríamos grandes razões para acreditar no resultado; mas não sabemos nada quando temos apenas uma estimativa por ponto.

Entretanto, podemos estimar ou fazer inferências sobre os valores da população usando uma segunda abordagem chamada de **estimativas por intervalos** ou **intervalos de confiança**, que dão o intervalo dentro do qual se espera que esteja o valor da população, com uma dada probabilidade ou um nível de confiança. Nesse caso, poderíamos inferir, por exemplo, que o consumo de carros populares que abastecem no posto de gasolina está no intervalo de R\$180,00 a R\$ 220,00 e, ainda, afirmaríamos isso com, por exemplo, 95% de certeza.

Como a estimativa por intervalos nos fornece uma informação mais precisa em relação ao parâmetro, esta é a melhor forma de se estimar o parâmetro populacional. Então, para você estimar parâmetros populacionais por meio de dados amostrais, é necessário o conhecimento da **distribuição amostral** da estatística que está sendo usada como estimador.

Na seção Distribuições Amostrais, abordamos esse assunto. Se julgar necessário, retome o conteúdo.

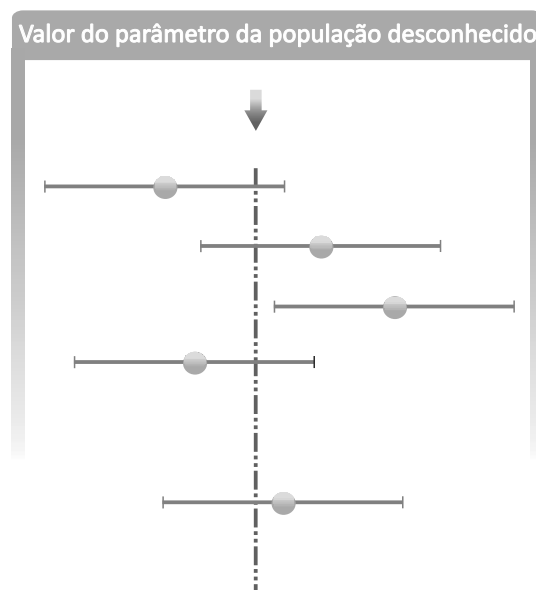
Em resumo, podemos dizer que a estimativa pontual fornece uma estimativa única de um parâmetro e que a estimativa intervalar nos dá um intervalo de valores possíveis, no qual se admite que esteja o parâmetro populacional com uma probabilidade conhecida.

## ESTIMAÇÃO POR INTERVALOS

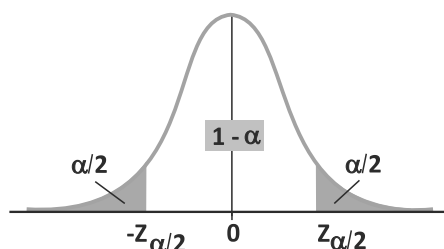
Você irá ver agora que um **intervalo de confiança** dá um intervalo de valores, centrado na estatística amostral, no qual julgamos, com um risco conhecido de erro, estar o parâmetro da população.

É o nível de significância que nos dá a medida da **incerteza** dessa inferência. O  $\alpha$  geralmente assume valores entre 1 e 10%.

*Então, a partir de informações de amostras, devemos calcular os limites de um intervalo, valores críticos, que em  $(1-\alpha)\%$  dos casos inclua o valor do parâmetro a estimar e em  $\alpha\%$  dos casos não inclua o valor do parâmetro, como podemos ver no desenho abaixo.*



O nível de confiança  $1 - \alpha$  é a probabilidade de o intervalo de confiança conter o parâmetro estimado. Em termos de variável normal padrão  $Z$ , isso representa a área central sob a curva normal entre os pontos  $-Z$  e  $Z$ .



Você pode observar que a área total sob a curva normal é unitária. Se a área central é  $1 - \alpha$ , o ponto  $-z$  representa o valor de  $Z$ , que deixa à sua esquerda a área  $\alpha/2$ , e o ponto  $z$  representa o valor de  $Z$ , que deixa à sua direita a área  $\alpha/2$ .

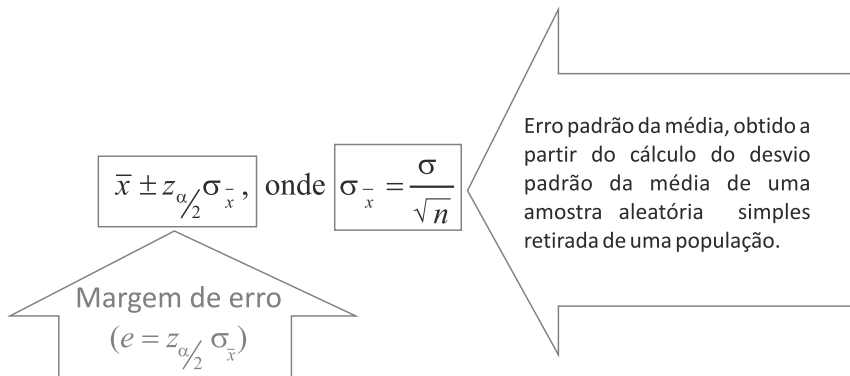
*Vamos aprender agora a construir o intervalo de confiança para uma média quando o desvio padrão populacional é conhecido ou a amostra é grande.*

**Vamos imaginar a seguinte situação:** o Departamento de Recursos Humanos de uma prefeitura informa que o tempo de execução de tarefas que envolvem participação manual varia de tarefa para tarefa, mas que o desvio padrão permanece aproximadamente constante, em 3 minutos. Novas tarefas estão sendo implantadas na prefeitura. Uma amostra aleatória do tempo de execução de 50 das novas tarefas forneceu o valor médio de 15 minutos. Determine um intervalo de confiança de 95% para o tempo médio de execução de uma dessas novas tarefas.

Primeiramente, você precisa identificar que o desvio padrão populacional é conhecido e também a amostra é considerada grande ( $n > 30$ ). Então, a construção do intervalo de confiança será feita utilizando a média amostral. Utilizaremos, para a obtenção dos limites de confiança, a curva normal padrão  $Z$ .

Como os limites são dados por meio da estatística calculada a partir dos dados amostrais e da margem de erro (fornecido pela estatística da distribuição multiplicada pelo desvio padrão da

distribuição amostral), teremos, nessa situação, os limites calculados por meio da seguinte expressão:



Logo, o intervalo de confiança tem centro na média amostral:

Calculando, teremos:

$$1 - \alpha = 0,95 \Rightarrow \alpha = 0,05 \Rightarrow \alpha/2 = 0,025$$

Olhando na tabela de Z, você encontrará  $Z_{\alpha/2} = 1,96$

$$e = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1,96 \cdot \frac{3}{\sqrt{50}} = 0,8315$$

$$P(\bar{x} - e < \mu < \bar{x} + e) = (1 - \alpha)$$

$$P(15 - 0,8315 < \mu < 15 + 0,8315) = 0,95$$

$$P(14,168 < \mu < 15,831) = 0,95$$

**Interpretação do resultado:** em cada grupo de 100 amostras retiradas de 50 pessoas, espera-se que, em 95 delas, a média esteja dentro do intervalo de 14,168 a 15,831.

*Antes de continuar a leitura, você deve realizar, ao final desta Unidade, a Atividade 6, na qual irá aplicar os conhecimentos relacionados à amostra e ao intervalo de confiança. Em caso de dúvida, faça contato com seu tutor.*

## DIMENSIONAMENTO DE AMOSTRAS

Desenvolvendo a expressão de erro mostrada anteriormente, obteremos o tamanho de amostra para estimação da média populacional quando o desvio padrão populacional for conhecido, como mostramos a seguir:

$$e = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow \sqrt{n} = z_{\alpha/2} \cdot \frac{\sigma}{e} \Rightarrow n = \left( \frac{z_{\alpha/2} \cdot \sigma}{e} \right)^2 \Rightarrow n = \frac{(z_{\alpha/2})^2 \cdot \sigma^2}{e^2}$$

**Imagine a seguinte situação:** que tamanho de amostra será necessário para produzir um intervalo de 95% de confiança para a verdadeira média populacional, com erro de 1,0, se o desvio padrão da população é 10,0?

Substituindo esses valores na expressão, teremos:

$$n_o = \frac{Z_{\alpha/2}^2 \cdot \sigma^2}{e^2} = \frac{1,96^2 \cdot 10^2}{1^2} = 384,16 \cong 385$$

Você pode alterar a confiança que teremos um diferente valor de Z e também o erro. Isso irá depender da precisão que você irá desejar nas suas estimativas.

Quando trabalhamos com proporção de sucesso, podemos substituir a variância por p.q (proporção de sucesso vezes a proporção de fracasso) da Distribuição de Bernoulli.

$$n = \frac{Z_{\alpha/2}^2 \cdot \hat{p}\hat{q}}{e^2}$$

Onde  $\hat{p}$  e  $\hat{q}$  correspondem às estimativas de sucesso e de fracasso, respectivamente, obtidos a partir de resultados amostrais.

*Vamos ver uma aplicação?*

Um setor da prefeitura que cuida da documentação de imóveis está interessado em **estimar** a proporção de pessoas que comprem novos imóveis na cidade para melhor dimensionar o setor de atendimento. Para isso, amostrou 80 pessoas do seu cadastro, verificando que 30 delas teriam comprado imóvel no último ano. Determine o tamanho da amostra necessário para estimar com 95% de confiança essa proporção e com erro máximo de 4%.


Substituindo os valores, teremos:

$$\hat{p} = \frac{30}{80} = 0,375 \quad e \quad \hat{q} = 1 - \hat{p} = 1 - 0,375 = 0,625$$

$$n = \frac{Z_{\alpha/2}^2 \cdot \hat{p}\hat{q}}{e^2} = \frac{1,96^2 \cdot 0,375 \cdot 0,625}{0,04^2} = 562,73 \cong 563$$

### Complementando...

Através do link que apresentamos a seguir, você poderá fazer cálculos das distribuições de probabilidade discretas ou contínuas, de dimensionamento de amostras e de intervalos de confiança.

 *Programa estatístico Bioestat.* Disponível em: <[http://www.mamiraua.org.br/download/Default.aspx?dirpath=e:\home\mamiraua\Web\download\BioEstat 5 Portugues&tipo=diretorio](http://www.mamiraua.org.br/download/Default.aspx?dirpath=e:\home\mamiraua\Web\download\BioEstat%20Portugues&tipo=diretorio)>. Acesso em: 19 nov. 2010.

# *Resumindo*



Nesta Unidade, você aprendeu sobre as principais distribuições de probabilidade, sejam elas discretas ou contínuas, e como utilizá-las. Também conheceu as distribuições de amostragem e, quando utilizá-las, e noções básicas de estimação (intervalos de confiança) e dimensionamento de amostras. Essas informações serão muito importantes para a compreensão da próxima Unidade.





## Atividades de aprendizagem

Para verificar se você está acompanhando o que apresentamos nesta Unidade, procure responder às atividades propostas, a seguir. Se tiver dificuldades para resolvê-las, consulte seu tutor.

1. No Brasil, a proporção de microempresas que fecham em até um ano de atividade é de 10%. Em uma amostra aleatória de 20 microempresas, qual a probabilidade de 5 terem fechado em até um ano de sua criação?
2. Entre 2.000 famílias de baixa renda e com quatro crianças, considerando-se que a chance de nascer uma criança do sexo masculino é igual a do sexo feminino, em quantas famílias se esperaria que tivessem:
  - a) Dois filhos do sexo masculino.
  - b) Um ou dois filhos do sexo masculino.
  - c) Nenhum filho do sexo feminino.
3. A ouvidoria de uma prefeitura recebe em média 2,8 reclamações/hora, segundo uma Distribuição de Poisson. Determine a probabilidade de chegarem duas ou mais reclamações em um período de:
  - a) 30 minutos.
  - b) 1 hora.
  - c) 2 horas.

4. As rendas mensais de funcionários do setor de arrecadação de uma prefeitura são normalmente distribuídas com uma média de R\$ 2.000,00 e um desvio padrão de R\$ 200,00. Qual é o valor de  $Z$  para uma renda  $X$  de R\$ 2.200,00 e de R\$ 1.700,00?
5. O uso diário de água por pessoa em uma determinada cidade é normalmente distribuído com média  $\mu$  igual a 20 litros e desvio padrão  $\sigma$  igual a 5 litros.
  - a) Que percentagem da população usa entre 20 e 24 litros por dia?
  - b) Que percentagem usa entre 16 e 20 litros?
  - c) Qual é a probabilidade de que uma pessoa selecionada ao acaso use mais do que 28 litros?
6. Considere que as despesas mensais com alimentação em restaurantes de comida a quilo para um casal são normalmente distribuídas com desvio padrão de R\$ 3,00. Uma amostra de 100 casais revelou uma despesa média de R\$ 27,00. Determine o intervalo de confiança de 95% para a despesa com alimentação de casais.

# UNIDADE 6

## TESTES DE HIPÓTESES

### OBJETIVOS ESPECÍFICOS DE APRENDIZAGEM

Ao finalizar esta Unidade, você deverá ser capaz de:

- ▶ Escolher o teste de hipótese adequado;
- ▶ Formular um teste de hipótese;
- ▶ Chegar a uma conclusão sobre uma população a partir dos resultados amostrais; e
- ▶ Interpretar os passos e os resultados de um teste de hipótese.



## INTRODUÇÃO

Caro estudante,  
Vamos conhecer agora os principais testes de hipóteses utilizados na inferência estatística.  
Você, como gestor, muitas vezes terá de tomar decisões baseadas na análise de dados a partir de um teste de hipótese. Portanto, esteja atento ao conteúdo que iremos apresentar a você nesta última Unidade, pois ao longo da leitura você certamente perceberá a importância desse assunto quando tratamos de Estatística Aplicada à Administração. Bom estudo!

Na teoria de decisão estatística, os testes de hipóteses assumem uma importância fundamental, já que nos permitem dizer, por exemplo, se duas populações são, de fato, iguais ou diferentes utilizando, para isso, amostras dessas populações. Sendo assim, a tomada de decisão de um gestor público deve estar baseada na análise de dados a partir de um teste de hipótese.

Você pode definir as hipóteses a serem testadas, retirar as amostras das populações a serem estudadas, calcular as estatísticas delas e, por fim, determinar o grau de aceitação de hipóteses baseadas na teoria de decisão, ou seja, se uma determinada hipótese será validada ou não.

Para você decidir se uma hipótese é verdadeira ou falsa, ou seja, se ela deve ser aceita ou rejeitada, considerando-se uma determinada amostra, precisamos seguir uma série de passos que são:

1. Definir a hipótese de igualdade ( $H_0$ ) e a hipótese alternativa ( $H_1$ ) para tentar rejeitar  $H_0$  (possíveis erros associados à tomada de decisão).
2. Definir o nível de significância ( $\alpha$ ).
3. Definir a distribuição amostral a ser utilizada.
4. Definir os limites da região de rejeição e de aceitação.
5. Calcular a estatística da distribuição escolhida a partir dos valores amostrais obtidos e tomar a decisão.

Você deve tomar a decisão baseado na seguinte regra: se o valor da estatística da distribuição calculado estiver na região de rejeição, rejeite a hipótese nula. Caso contrário, se o valor da estatística calculado caiu na região de aceitação, a decisão será que a hipótese nula não poderá ser rejeitada ao nível de significância determinado.

*Agora, você terá o detalhamento dos passos na formulação de um teste de hipótese. Esteja bem atento!*

## ESTRUTURA DOS TESTES DE HIPÓTESES

Diversos conceitos serão apresentados ao longo do detalhamento dos passos a serem seguidos na formulação de um teste de hipótese.

### 1) Formular as hipóteses ( $H_0$ e $H_1$ ).

Primeiramente, vamos estabelecer as **hipóteses nula e alternativa**. Para exemplificar, você deve considerar um teste de hipótese para uma média. Então, a hipótese de igualdade é chamada de **hipótese de nulidade ou  $H_0$** .

Suponha que você queira testar a hipótese de que o tempo médio de atendimento na retirada de uma guia, em uma prefeitura considerada modelo de atendimento, é igual a 50 segundos. Essa hipótese será simbolizada da seguinte maneira:

$$H_0: \mu = 50 \text{ (hipótese de nulidade)}$$

Essa hipótese, na maioria dos casos, será de igualdade.

Se você rejeitar essa hipótese, irá aceitar, nesse caso, outra hipótese, que chamamos de **hipótese alternativa**. Esse tipo de hipótese é simbolizado por  **$H_1$  ou  $H_a$** .

A partir do nosso exemplo, as hipóteses alternativas mais comuns são as apresentadas a seguir:

- $H_1: \mu > 50$  (teste unilateral ou unicaudal à direita).

O tempo médio de retirada da guia é superior a 50 segundos ( $>$ ).

- $H_1: \mu < 50$  (teste unilateral ou unicaudal à esquerda).

O tempo médio de retirada da guia é inferior a 50 segundos ( $<$ ).

- $H_1: \mu \neq 50$  (teste bilateral ou bicaudal).

O tempo médio de retirada da guia pode ser superior ou inferior a 50 segundos.

*Surge uma dúvida. Qual hipótese alternativa você utilizará?*

*A resposta é bem simples.*

A hipótese alternativa será definida por você em razão do tipo de decisão que deseja tomar.

Veja o seguinte exemplo: você inspeciona uma amostra, relativa a uma grande remessa que chega a uma prefeitura, e constata que 8% dela está defeituosa. O fornecedor garante que não haverá mais de 6% de peças defeituosas em cada remessa. O que devemos responder, com auxílio dos testes de significância, é se a afirmação do fornecedor é verdadeira.

### **As hipóteses que você vai formular são:**

$$H_0: p = 0,06;$$

$$H_1: p > 0,06.$$

É importante ressaltar que o sinal de igual para a hipótese  $H_0$  corresponde a um sinal de menor ou igual (nesse exemplo), pois o teste é unilateral à direita ( $p > 0,06$ ). Portanto, sempre que o teste for unilateral, deve ser feita essa consideração.

### **2) Definir o nível de significância.**

O nível de significância de um teste é dado pela probabilidade de se cometer erro do tipo I (**ocorre quando você rejeita a hipótese  $H_0$  e essa hipótese é verdadeira**). Com o valor dessa



A hipótese alternativa somente pode ser maior, pois o fornecedor garante que não haverá mais de 6%.



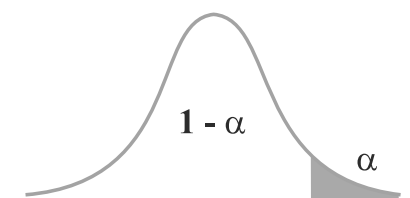
probabilidade fixada, você pode determinar o chamado **valor crítico**, que separa a chamada **região de rejeição** da hipótese  $H_0$ , da região de aceitação da hipótese  $H_0$ .

No desenho, a seguir, as áreas escuras correspondem à significância do teste, ou seja, a probabilidade de se cometer o chamado erro tipo I (rejeitar  $H_0$  quando ela é verdadeira). Essa probabilidade é chamada de  $\alpha$  e geralmente os valores mais utilizados são 0,01 e 0,05. O complementar do nível de significância é chamado de nível de confiança (área clara dos gráficos) e é dado por  $1 - \alpha$ .

Unilateral à direita:

$H_0: \mu = 50$

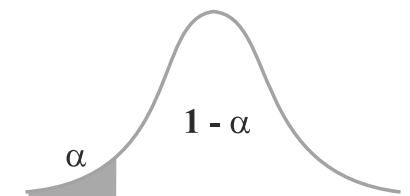
$H_1: \mu > 50$



Unilateral à esquerda:

$H_0: \mu = 50$

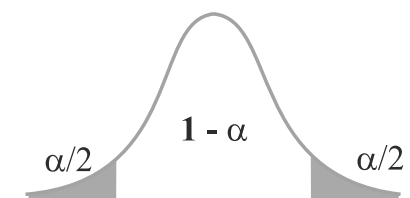
$H_1: \mu < 50$



Bilateral:

$H_0: \mu = 50$

$H_1: \mu \neq 50$



### 3) Definir a distribuição amostral a ser utilizada.

Você definirá a estatística a ser utilizada no teste em razão da distribuição amostral a qual os dados seguem. Se você fizer um teste de hipótese para uma média ou diferença entre médias, utilize a **distribuição de Z ou t de Student**.

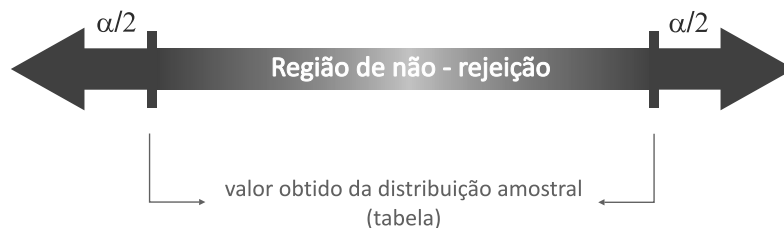
Note que o conhecimento das distribuições amostrais vistas na Unidade 5 é muito importante. Caso ainda tenha alguma dúvida, volte e relembre os conceitos das distribuições de t, qui-quadrado e F, e como utilizar as tabelas.



Outro exemplo é se você quiser comparar a variância de duas populações; para tal, deverá trabalhar então com a distribuição F, ou seja, da razão de duas variâncias.

#### 4) Definir os limites da região de rejeição.

Os limites entre as regiões de rejeição e de aceitação da hipótese  $H_0$  você definirá em razão do tipo de hipótese  $H_1$ , do valor de  $\alpha$  (nível de significância) e da distribuição amostral utilizada. Considerando um teste bilateral, você terá a região de aceitação (não rejeição) com uma probabilidade de  $1-\alpha$ , e uma região de rejeição com probabilidade  $\alpha$  ( $\alpha/2 + \alpha/2$ ).



Por meio da amostra obtida, você deve calcular a estimativa que servirá para aceitar ou para rejeitar a hipótese nula. Neste momento, você deve estar se perguntando: **como irei calcular a estimativa, ou seja, o valor da estatística a partir dos dados amostrais?** A resposta será dada no próximo item.

#### 5) Tomar a decisão

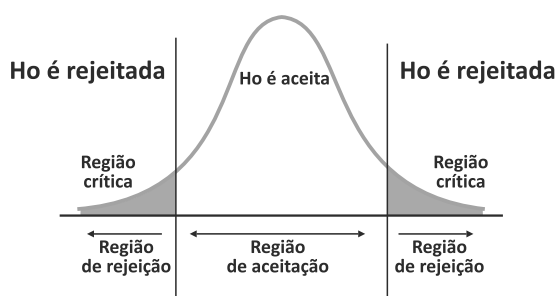
Para tomar a decisão, você deve calcular a estimativa do teste estatístico que será utilizada para rejeitar ou não a hipótese  $H_0$ . A estrutura desse cálculo para a média de forma generalista é dada por:

$$\text{Estatística da distribuição} = \frac{(\text{estimativa} - \text{parâmetro})}{\text{erro padrão da estimativa}}$$

Podemos exemplificar pela distribuição de Z, que será:

$$\text{Estatística do teste} \rightarrow Z_{cal} = \frac{(\bar{x} - \mu)}{(\sigma/\sqrt{n})} \leftarrow \text{Variabilidade das médias}$$

Se o valor da estatística estiver na região crítica (de rejeição), você vai rejeitar  $H_0$ , caso contrário, aceite  $H_0$ . O esquema a seguir mostra bem a situação de decisão.



## TESTE DE HIPÓTESE PARA UMA MÉDIA

Quando você retira uma amostra de uma população e calcula a média dessa amostra, é possível verificar se a afirmação sobre a média populacional é verdadeira. Para tanto, basta verificar se a estatística do teste estará na região de aceitação ou de rejeição da hipótese  $H_0$ .

Aqui, você tem duas situações distintas:

**Primeira situação:** se o desvio padrão da população é conhecido ou a amostra é considerada grande ( $n > 30$ ): a distribuição amostral a ser utilizada será da Normal ou Z e a estatística teste que você utilizará será:

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

Onde:

$\bar{x}$  : média amostral;

$\mu$  : média populacional;

$\sigma$  : desvio padrão populacional; e

$n$ : tamanho da amostra.

**Imagine a seguinte situação:** um gestor público sabe que, para montar um determinado negócio em um bairro de Curitiba, é necessário que nele circulem, no mínimo, 1.500 pessoas por dia. Para o tipo de bairro em questão, é possível supor o desvio padrão como sendo igual a 200 pessoas. Uma amostra aleatória formada por 12 observações revelou que passariam pelo local

escolhido 1.400 pessoas por dia, em média. O negócio pode ser montado ou não? Assuma  $\alpha = 5\%$  e suponha população normalmente distribuída.

### Resolução:

Sempre, em um exercício de tomada de decisão, precisamos da formulação de um teste de hipótese, seguindo os passos apresentados:

1. Formular as hipóteses.
2. Definir o nível de significância.
3. Definir a distribuição amostral a ser utilizada.
4. Definir os limites da região de rejeição (gráfico).
5. Tomar a decisão.

Vamos primeiramente retirar os dados do problema:

$$n = 12; \bar{x} = 1400 \text{ e } \sigma = 200$$

Vamos estabelecer as hipóteses com base no exercício:

$$H_0: \mu = 1500$$

$$H_1: \mu < 1500$$

Caso tenhamos uma média igual a 1.500 pessoas, podemos montar o negócio. Mas se aceitarmos a hipótese  $H_1$ , não devemos indicar a montagem do negócio.

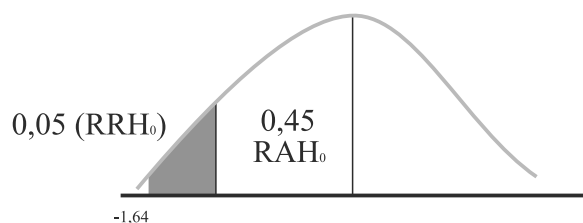
$$\alpha = 0,05$$

A estatística escolhida é  $Z$ . Substituindo os valores da amostra e o da hipótese  $H_0$  na estatística de  $Z$ , teremos:

$$Z_c = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1400 - 1500}{\frac{200}{\sqrt{12}}} = \frac{-100}{57,73} = -1,73$$

Denominamos esse desvio como populacional, pois, baseados nas características do bairro (conhecimento prévio), podemos supor o valor do desvio.

Veja que, mesmo com  $n \leq 30$ , o desvio padrão populacional foi informado. Quando temos essa situação, devemos sempre usar  $Z$ .



Valor mais próximo de 0,45, pois este não existe na tabela.

O valor  $Z_t = -1,64$ , que divide a  $RRH_0$  e  $RAH_0$ , foi encontrado na tabela Z procurando em seu **interior** o valor 0,4495. Como Z calculado é menor que Z tabelado, ou seja,  $-1,73$  pertence a  $RRH_0$ , podemos afirmar com 95% de certeza que transitam menos de 1.500 pessoas por dia no bairro e, assim, verificamos que não é viável montar o negócio no bairro.

*Agora, antes de prosseguir, você deve resolver a Atividade 1, ao final desta Unidade. Caso tenha alguma dúvida, retorne a situação anterior, aquela que resolvemos juntos.*

**Segunda situação:** se você não conhecer o desvio padrão populacional e a amostra for pequena ( $n > 30$ ), a distribuição amostral a ser utilizada será a t de *Student* e a estatística teste será:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Onde:

$\bar{x}$  : média amostral;

$\mu$  : média populacional;

s : desvio padrão amostral; e

n: tamanho da amostra.

Uma observação importante: quando trabalhamos com amostras grandes, ou seja,  $n \geq 30$ , a distribuição de Z e t de *Student* apresentam comportamentos e valores da estatística próximos.

Neste momento, releia os passos anteriores para que não fique nenhuma dúvida em relação à estrutura de um teste de hipótese, pois iremos trabalhar juntos em situações nas quais iremos aplicar os diferentes testes de hipóteses para uma média.

*Então, após a releitura do conteúdo apresentado, vamos analisar as situações.*

Veja, abaixo, a primeira situação em que utilizaremos o teste de hipótese para uma média usando a estatística de Z (amostras grandes ou variância populacional conhecida). Para resolver essa situação, utilizaremos o teste de hipótese para uma média usando a estatística de t de *Student* (amostra pequena e variância populacional desconhecida).

A Construtora Estrada Forte Ltda. alega ser capaz de produzir concreto com, no máximo, 15 kg de impurezas para cada tonelada fabricada. Mas, segundo a legislação municipal, caso essa quantidade seja maior do que 15 kg, a obra deve ser embargada pela prefeitura. Dezenove amostras de uma tonelada cada uma revelaram possuir impurezas com média amostral igual a 23 kg e desvio padrão igual a 9 kg. Assumindo  $\alpha = 5\%$  e população normalmente distribuída, a obra deve ser embargada ou não?

### **Resolução:**

Retirando os dados do problema:

$n = 19$ ;  $\bar{x} = 23$ ;  $s = 9$ ;  $\alpha = 0,05$ . Vamos estabelecer as hipóteses baseando-nos na afirmação do exercício:

$$H_0: \mu = 15$$

$$H_1: \mu > 15 \quad \text{Caso a hipótese } H_0 \text{ seja aceita, a obra não será}$$

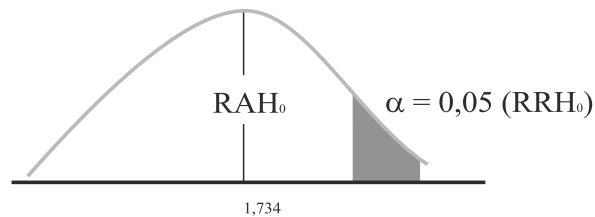
embargada, pois ela está de acordo com a lei. Caso contrário, a prefeitura embarga a obra.

$$\alpha = 0,05$$

A estatística escolhida é **t de Student**.

Substituindo os valores do problema na expressão, teremos:

$$t_c = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{23 - 15}{\frac{9}{\sqrt{19}}} = \frac{8}{2,06} = 3,87$$



O valor  $t_t = 1,734$  que divide a  $RRH_0$  e  $RAH_0$  foi encontrado na tabela t procurando grau de liberdade 18 ( $gl = n - 1 = 19 - 1 = 18$ ) e  $\alpha = 0,05$ . Como  $t$  calculado é maior do que  $t$  tabelado, ou seja,  $1,734$  pertence a  $RRH_0$ , podemos afirmar com 95% de certeza que a alegação da construtora não é verdadeira. Eles não são capazes de produzir concreto com, no máximo, 15 kg de impurezas para cada tonelada fabricada. Então, concluímos que a obra deve ser embargada pela prefeitura.

Veja que o  $n$  foi menor ou igual a 30 ( $n \leq 30$ ), foi informado o desvio padrão amostral e não foi apresentado o desvio padrão populacional. Nessas condições, devemos sempre usar distribuição t de Student.



## TESTE DE HIPÓTESE PARA A RAZÃO DE DUAS VARIÂNCIAS

Esse teste de hipótese é utilizado para saber se duas variâncias populacionais são estatisticamente iguais ou se uma é maior do que a outra. Utilizando a distribuição F, poderemos formular o teste de hipótese da razão entre duas variâncias e chegar à conclusão baseados apenas nas estimativas calculadas a partir das amostras.

As hipóteses  $H_0$  e  $H_1$  serão:

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ (variâncias das duas populações são iguais)}$$

$$H_1 : \sigma_1^2 > \sigma_2^2 \text{ (variância da população 1 é maior do que a da população 2).}$$

Como estamos utilizando um teste unilateral à direita, por questões didáticas, então, no cálculo da estatística de F, teremos a maior variância dividida pela menor variância.

A maior variância amostral encontrada será chamada de  $s_1^2$  (proveniente de uma amostra de tamanho  $n_1$ ) e a menor variância amostral será chamada  $s_2^2$  (proveniente de amostra de tamanho  $n_2$ ).

Vamos considerar duas amostras provenientes de duas populações. Desejamos saber se as variâncias das populações são estatisticamente iguais ou se uma é maior do que a outra. Considere uma significância de 2,5%. Os resultados amostrais são apresentados a seguir:

$$s_1^2 = 0,5184 \quad \text{com} \quad n_1 = 14$$

$$s_2^2 = 0,2025 \quad \text{com} \quad n_2 = 21$$

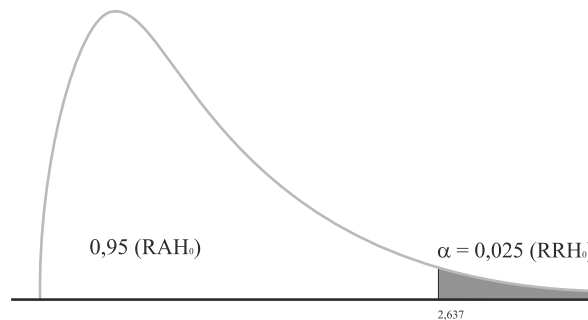
Então, a variável de teste do teste F será:

$$F = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

Como em  $H_0$  estamos considerando que as variâncias populacionais são iguais, então, na expressão acima, as duas variâncias populacionais irão se cancelar. No nosso exemplo, teremos:

$$F = \frac{s_1^2}{s_2^2} = \frac{0,5184}{0,2025} = 2,56$$

O valor tabelado (crítico) da distribuição de F será obtido na tabela da distribuição com uma significância de 2,5%. Considerando como graus de liberdade iguais a 13 ( $n_1 - 1$ ) para o numerador ( $v_1$ ) e 20 ( $n_2 - 1$ ) para o denominador ( $v_2$ ), chegaremos ao seguinte resultado: valor tabelado igual a 2,637.



O valor calculado da estatística (2,56) foi menor do que o tabelado (2,637), então, o valor calculado caiu na região de aceitação de  $H_0$ . Assim, aceitamos  $H_0$  e consideramos que a variância da população 1 é estatisticamente igual à variância da população 2, ou seja, não ocorre uma diferença entre elas.

Esse teste servirá de base para a escolha do próximo teste (diferença entre médias para amostras independentes), ou seja, escolher o tipo de teste a ser utilizado.

## TESTE DE HIPÓTESE PARA A DIFERENÇA ENTRE MÉDIAS

Quando queremos comparar a média de duas populações, retiramos amostras das duas populações, que podem apresentar tamanhos diferentes. Vamos considerar as situações de amostras independentes (as populações não apresentam nenhuma relação entre si) e de amostras dependentes (uma população sofre uma intervenção e avalia-se antes e depois da intervenção para saber se a intervenção resultou algum efeito).

**1º caso:** amostras independentes e grandes ( $n > 30$ ) ou variâncias populacionais conhecidas.

**2º caso:** amostras independentes e pequenas ( $n \leq 30$ ), mas que apresentam variâncias populacionais desconhecidas e estatisticamente iguais.

**3º caso:** amostras independentes e pequenas ( $n \leq 30$ ), mas que apresentam variâncias populacionais desconhecidas e estatisticamente desiguais.

**4º caso:** amostras dependentes.

*Vamos analisar cada uma dessas situações. Lembre-se de que as considerações anteriores em relação aos passos para formulação dos testes de hipóteses permanecem os mesmos.*

A grande diferença, como você verá, ocorrerá somente na determinação das hipóteses a serem testadas. A hipótese  $H_0$  será:

$$H_0: \mu_1 - \mu_2 = d_0$$

Onde:

$\mu_1$ : média da população 1;

$\mu_2$ : média da população 2; e

$d_0$  corresponde a uma diferença qualquer que você deseja testar.

Geralmente, quando queremos saber se as médias das duas populações são estatisticamente iguais, utilizamos o valor de  $d_0$  igual a zero.

As hipóteses alternativas seguem a mesma linha de raciocínio, como você pode visualizar a seguir.

$H_0$	$H_1$
$\mu_1 - \mu_2 = d_0$	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$

É importante ressaltar que, se as hipóteses alternativas forem unilaterais, o sinal da hipótese  $H_0$  será **menor ou igual** ou **maior ou igual** dependendo da hipótese alternativa, apesar de utilizarmos a notação de igual (conforme comentado anteriormente).

Todas as outras considerações em relação aos testes de hipótese permanecem as mesmas. Vamos, então, procurar entender cada situação para os testes de hipóteses para diferença entre médias.

**1ª caso:** amostras independentes e grandes ( $n > 30$ ) ou variâncias populacionais conhecidas: como estamos trabalhando aqui com amostras grandes ou com desvios padrão populacionais conhecidos, devemos trabalhar com a distribuição amostral de Z (raciocínio semelhante ao utilizado no teste de hipótese para uma média). Portanto, a estatística do teste será dada por:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$

Onde:

$\bar{X}_1$ : média da amostra 1;  $\bar{X}_2$ : média da amostra 2;  $\mu_1$ : média da população 1;  $\mu_2$ : média da população 2;  $\sigma_1^2$ : variância da população 1;  $\sigma_2^2$ : variância da população 2;  $n_1$ : tamanho da amostra 1 e  $n_2$  tamanho da amostra 2.

Se trabalharmos com amostras grandes, poderemos substituir as variâncias populacionais pelas variâncias amostrais sem nenhum problema.

Vamos, então, ver como podemos aplicar o teste de hipótese para a diferença entre médias nesta situação:

Foram retiradas amostras do valor recebido em milhares de reais de um determinado imposto de duas prefeituras (A e B) de mesmo porte. Os resultados são apresentados no quadro, a seguir. Verifique se as duas prefeituras têm o mesmo recebimento ou se são diferentes, com uma significância de 0,05.

MARCAS	A	B
Média	1160	1140
Desvio padrão	90	80
Tamanho amostra	100	100

### Como fazer:

Vamos retirar os dados apresentados em nossa situação:

Amostra A:  $n = 100$ ;  $\bar{x} = 1160$ ;  $s = 90$

Amostra B:  $n = 100$ ;  $\bar{x} = 1140$ ;  $s = 80$

As hipóteses a serem formuladas são:

$$H_0: \mu_a = \mu_b \rightarrow \mu_a - \mu_b = 0$$

$$H_1: \mu_a \neq \mu_b$$

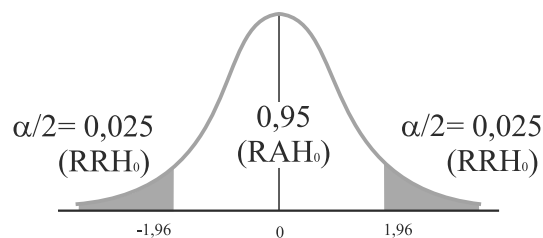
O teste t deve ser bilateral, já que a preocupação está na verificação do fato da média da prefeitura A ser diferente da média da prefeitura B.

$$\alpha = 0,05$$

A estatística usada será Z, pois as amostras são grandes ( $n > 30$ ), apesar de não termos os desvios padrão populacionais. Sendo assim, nessa situação, ainda utilizamos a estatística de Z.

Substituindo os valores na estatística, teremos:

$$Z_c = \frac{(\bar{X}_a - \bar{X}_b) - (\mu_a - \mu_b)}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}} = \frac{(1160 - 1140) - (0)}{\sqrt{\frac{90^2}{100} + \frac{80^2}{100}}} = 1,67$$



Como o valor calculado  $Z_c = 1,67$  está entre os valores de  $-1,96$  e  $1,96$ , valores que dividem a  $RRH_0$  da  $RAH_0$ , verificamos que o valor calculado  $Z_c = 1,67$  pertence a  $RAH_0$  e podemos afirmar, com 95% de certeza, que os valores recebidos pelas duas prefeituras são estatisticamente iguais, ou seja, aquela diferença encontrada entre as amostras foi fruto do acaso.

**2ª caso:** amostras independentes e pequenas, mas que apresentam variâncias populacionais estatisticamente iguais e desconhecidas: você deve trabalhar com a distribuição t de *Student*, uma vez que as amostras que estamos trabalhando são pequenas, e as variâncias populacionais desconhecidas.

Aqui, estaremos considerando que as variâncias populacionais são estatisticamente iguais, pois essa situação influenciará nos cálculos e, conseqüentemente, no processo decisório.

Para saber se as variâncias podem ser consideradas iguais, você deve fazer um teste da razão de duas variâncias (teste F), apresentado anteriormente.

A estatística do teste será dada por:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{1/n_1 + 1/n_2}}$$

Aqui, aparece um termo novo ( $s_p$ ). Ele corresponde ao desvio padrão ponderado pelos graus de liberdade, ou seja, calculamos um novo desvio padrão cujo fator de ponderação corresponde ao grau de liberdade de cada amostra. Veja a seguir:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Para você encontrar o valor tabelado que limita as regiões de aceitação e de rejeição na **tabela t de Student**, o número de graus de liberdade ( $v$ ) a ser usado na tabela será dado por:

$$v = n_1 + n_2 - 2$$

Onde:

$n_1$  e  $n_2$  correspondem aos tamanhos de amostras utilizados.

Lembre-se de que você pode voltar à tabela t de Student quando desejar, ela se encontra na Unidade 5.

*Vamos resolver, agora, uma situação na qual temos a comparação entre médias de amostras pequenas e variâncias populacionais desconhecidas e estatisticamente iguais.*

**Situação:** em uma comparação de aprovação no vestibular de uma importante universidade, seis estudantes do sexo masculino de colégios da rede pública (amostra A) preencheram o gabarito no tempo médio de 6,4 minutos e desvio padrão de 60 segundos. Outra amostra foi formada por cinco estudantes do sexo feminino selecionados aleatoriamente do mesmo universo (amostra B), com os resultados de tempo de preenchimento do gabarito, de um tempo médio de 5,9 minutos e com desvio padrão de 60 segundos (assuma variâncias populacionais iguais). A Secretaria Municipal de Educação deseja saber se existe diferença ou não entre o sexo dos estudantes para definir se há necessidade de se fazer treinamentos específicos para cada sexo ou um mesmo treinamento para os dois sexos; para, assim, poder reduzir esse tempo e melhorar a *performance* dos estudantes da rede pública no vestibular.

### Resolução:

Retirando os dados do nosso exemplo, teremos:

Amostra A:  $n = 6$ ;  $\bar{x} = 6,4$ ;  $s = 1$

Amostra B:  $n = 5$ ;  $\bar{x} = 5,9$ ;  $s = 1$

As hipóteses a serem formuladas são:

$$H_0: \mu_a = \mu_b \rightarrow \mu_a - \mu_b = 0$$

$$H_1: \mu_a \neq \mu_b$$

O teste t deve ser bilateral, já que a atenção está voltada para a preocupação em se constatar se, de fato, ocorre diferença entre os estudantes do sexo masculino ou feminino.

$$\alpha = 0,05$$

A estatística usada será t, pois as amostras são menores ou iguais a 30 ( $n \leq 30$ ) e a variância populacional é desconhecida. Além disso, consideramos que as variâncias populacionais são estatisticamente iguais, **informação** que é dada no problema analisado.

Caso isso não seja informado no problema, você deve fazer um teste de hipótese para comparar as variâncias populacionais com base nas variâncias amostrais, como vimos anteriormente.



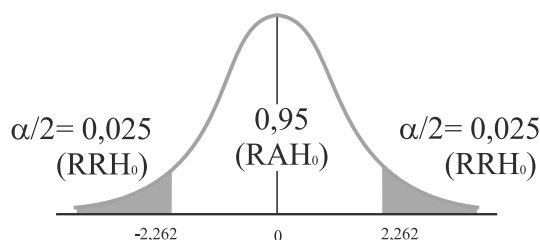


Substituindo os valores nas expressões, teremos:

$$Sp = \sqrt{\frac{(n_a - 1) \cdot s_a^2 + (n_b - 1) \cdot s_b^2}{n_a + n_b - 2}} = \sqrt{\frac{5 \cdot 1^2 + 4 \cdot 1^2}{6 + 5 - 2}} = 1$$

$$t_c = \frac{(\bar{x}_a - \bar{x}_b) - (\mu_a - \mu_b)}{Sp \sqrt{\frac{1}{n_a} + \frac{1}{n_b}}} = \frac{(6,4 - 5,9) - (0)}{1 \cdot \sqrt{\frac{1}{6} + \frac{1}{5}}} = \frac{0,5}{0,6055} = 0,82$$

$$v = n_a + n_b - 2 = 6 + 5 - 2 = 9 \text{ (grau de liberdade)}$$



O valor  $t_t = 2,262$  que divide a  $RRH_0$  e  $RAH_0$  foi encontrado na tabela t procurando grau de liberdade 9 e  $\alpha = 0,025$ . Como  $t$  calculado está entre os valores que dividem a região de aceitação de  $H_0$ , ou seja, 0,82 pertence a  $RAH_0$ , podemos afirmar com 95% de certeza que o tempo de preenchimento dos estudantes e das estudantes é o mesmo. Então, a prefeitura deve fazer o treinamento independentemente do sexo dos estudantes, ou seja, o mesmo treinamento para todos os estudantes.

*Antes de analisar o terceiro caso, realize a Atividade 2, ao final desta Unidade.*

**3º caso:** amostras independentes e pequenas, mas que apresentam variâncias populacionais estatisticamente desiguais e desconhecidas: a diferença dessa situação para a anterior é que você agora considera que as populações apresentam variâncias estatisticamente desiguais. Para saber se elas são estatisticamente

desiguais ou diferentes, você deve fazer um teste de hipótese para a razão de duas variâncias, visto anteriormente nesta Unidade. Também utilizaremos a estatística do teste a partir da distribuição *t* de *Student*. A estatística do teste será dada por:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2 / n_1 + s_2^2 / n_2}}$$

Outra diferença está no cálculo do número de graus de liberdade, pois, nessa situação, utilizaremos uma aproximação que é dada pela expressão a seguir:

$$v = gl = \frac{(s_1^2 / n_1 + s_2^2 / n_2)^2}{\frac{(s_1^2 / n_1)^2}{n_1 - 1} + \frac{(s_2^2 / n_2)^2}{n_2 - 1}}$$

Se esse valor calculado apresentar valores decimais, você deve fazer o arredondamento para um número inteiro.

*Vamos resolver, a seguir, outra situação.*

**Situação:** uma prefeitura deseja reduzir seus custos com combustíveis. Para isso, deseja saber se duas marcas de carro apresentam o mesmo consumo ou se um dos fabricantes apresenta menor consumo. Não confiando nas especificações do fabricante, já que as condições de uso dos veículos pela prefeitura não são ideais. Para tomar a decisão acerca de qual comprar, foi analisada uma amostra de 22 automóveis das duas marcas, obtendo o resultado apresentado, a seguir. Seria possível afirmar que o carro Andaluz é mais econômico, isto é, que apresenta uma média populacional inferior que a do Reluzente? Assuma  $\alpha = 5\%$  e população normalmente distribuída.

AUTOMÓVEL	TAMANHO DA AMOSTRA	MÉDIA DE CONSUMO	DESVIO PADRÃO
Andaluz	12 unidades	14 km/l	2 km/l
Reluzente	10 unidades	15 km/l	4 km/l

### Resolução:

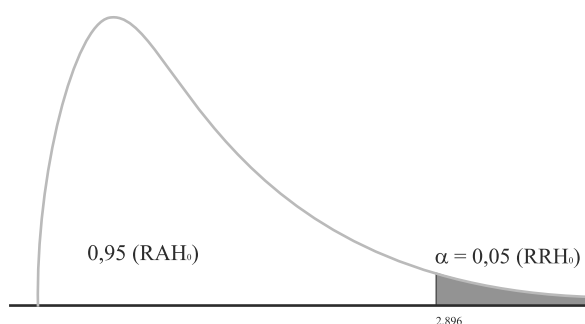
Nessa situação, faremos um teste de hipótese para diferença entre médias populacionais. Como as amostras são pequenas, precisamos saber se as variâncias são estatisticamente iguais ou não. Para isso, vamos testar se as variâncias populacionais são estatisticamente iguais ou não por meio de teste de F. As hipóteses são:

$$H_0 : \sigma_R^2 = \sigma_A^2$$

$$H_1 : \sigma_R^2 > \sigma_A^2 \quad \alpha = 0,05$$

$$F_c = \frac{s_R^2}{s_A^2} = \frac{16}{4} = 4$$

Como estabelecemos utilizar o teste unilateral no cálculo de F, teremos, então, a maior variância dividida pela menor variância. As variâncias populacionais não estão presentes na fórmula, devida, na hipótese  $H_0$ , serem consideradas iguais e, assim, elas se cancelam.



O valor 2,896 foi encontrado na tabela F de 5% com grau de liberdade 9 para o numerador e 11 para o denominador. Como  $F_c > 2,896$ , rejeita-se  $H_0$  e, portanto, as variâncias populacionais são estatisticamente desiguais, ou seja, uma é maior do que a outra.

Agora, vamos testar as médias populacionais:

$$H_0: \mu_{\text{andaluz}} = \mu_{\text{reluzente}} \rightarrow \mu_{\text{andaluz}} - \mu_{\text{reluzente}} = 0$$

$$H_1: \mu_{\text{andaluz}} < \mu_{\text{reluzente}}$$

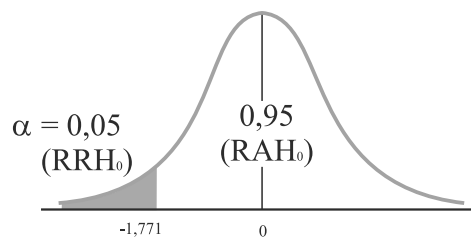
$$\alpha = 0,05$$

Como as amostras são independentes, pequenas e com variâncias populacionais estatisticamente desiguais, usaremos a estatística t.

Vamos encontrar o grau de liberdade:

$$V = \frac{\left( \frac{s_A^2}{n_A} + \frac{s_R^2}{n_R} \right)^2}{\frac{\left( \frac{s_A^2}{n_A} \right)^2}{n_A - 1} + \frac{\left( \frac{s_R^2}{n_R} \right)^2}{n_R - 1}} = \frac{\left( \frac{4}{12} + \frac{16}{10} \right)^2}{\frac{\left( \frac{4}{12} \right)^2}{11} + \frac{\left( \frac{16}{10} \right)^2}{9}} = \frac{3,74}{0,01 + 0,28} = \frac{3,74}{0,29} = 12,89 \cong 13$$

$$t = \frac{(\bar{x}_A - \bar{x}_R) - (\mu_{\text{andaluz}} - \mu_{\text{reluzente}})}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_R^2}{n_R}}} = \frac{(14 - 15) - (0)}{\sqrt{\frac{4}{12} + \frac{16}{10}}} = \frac{-1}{1,39} = -0,72$$



O valor  $t_t = -1,771$ , que divide a  $RRH_0$  e  $RAH_0$ , foi encontrado na tabela t procurando grau de liberdade 13 e  $\alpha = 0,05$ . Como t calculado ( $t = 0,72$ ) pertence a  $RAH_0$ , podemos afirmar, com 95% de certeza, que o consumo dos carros Andaluz e Reluzente é o mesmo, ou seja, tanto faz a prefeitura comprar uma marca ou outra que o consumo será o mesmo.

*Antes de passarmos ao estudo do quarto caso, resolva a Atividade 3, ao final desta Unidade. Dessa forma, você poderá aplicar os conhecimentos sobre a diferença entre médias.*

**4ª caso:** amostras dependentes: sabemos que amostras dependentes ocorrem quando fazemos uma intervenção e desejamos saber se os resultados antes da intervenção são iguais aos resultados depois da intervenção. Um ponto importante, nessa situação, é que são calculadas, primeiramente, as diferenças de antes e de depois. Essa diferença é chamada de  $d_i$ .

Então, você pode ver que:

$$d_i = \text{valor antes} - \text{valor depois}$$

Com base nessas diferenças ( $d_i$ ), você irá calcular a média ( $\bar{D}$ ) e o desvio padrão dessas diferenças ( $S_D$ ).

$$\bar{D} = \frac{\sum_{i=1}^n d_i}{n} \quad e \quad S_D = \frac{\sum_{i=1}^n d_i^2 - \frac{\left(\sum_{i=1}^n d_i\right)^2}{n}}{n-1}$$

Veja que essas fórmulas são iguais às de cálculo da média e do desvio padrão apresentados anteriormente. Nesse caso, no lugar da variável  $x$  são utilizados os valores de  $d_i$  (diferenças).

Com esses valores, a estatística teste será dada por:

$$t = \frac{\bar{D} - d_o}{S_D / \sqrt{n}}$$

O valor de  $n$  corresponde ao número de diferenças calculadas e o grau de liberdade para ser olhado na tabela  $t$  de *Student* será dado por  $n - 1$ .

*Vamos resolver uma situação em que trabalharemos com o caso de amostras dependentes.*

**Situação:** em um estudo procurou-se investigar se a redução em uma gratificação no salário iria diminuir a produtividade dos funcionários de uma prefeitura, considerando uma escala de produtividade de 0 a 12. A tabela a seguir dá os resultados de pessoas selecionadas anteriormente. No nível de 5% de significância, teste a afirmação de que a redução da gratificação reduziu a produtividade, ou seja, que a diferença entre antes e depois deve ser maior do que zero.

PESSOA	A	B	C	D	E	F	G	H
Antes	6,6	6,5	9,0	10,3	11,3	8,1	6,3	11,6
Depois	6,8	2,4	7,4	8,5	8,1	6,1	3,4	2,0

Primeiramente, vamos montar as nossas hipóteses:

$$H_0: \mu_D = 0$$

$$H_1: \mu_D < 0$$

Veja que as escolhas dessas hipóteses estão associadas ao que queremos testar. No caso da hipótese  $H_0: \mu_D = 0$ , estamos testando que as médias das diferenças de antes menos depois são iguais a zero, ou seja, que a redução na gratificação não interferiu na produtividade (a produtividade foi a mesma), já que estamos avaliando os mesmos indivíduos. No caso da hipótese  $H_1: \mu_D > 0$ , estamos testando que os valores de antes eram maiores do que os valores de depois da redução da gratificação, ou seja, se esta diferença de antes menos de depois for maior do que zero, indica que antes da intervenção os funcionários tinham uma produtividade maior antes do que depois.

Poderíamos testar também, dependendo do caso, as hipóteses  $H_1: \mu_D < 0$  ou  $H_1: \mu_D \neq 0$ .

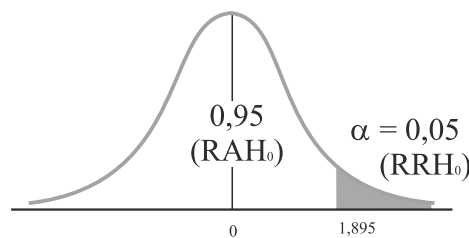
Consideramos um  $\alpha = 0,05$ .

Para calcularmos os valores de  $\bar{D}$  e  $S_D$ , devemos, primeiramente, calcular as diferenças entre os valores de antes menos de depois de cada indivíduo e com essas diferenças calcular a média das diferenças ( $\bar{D}$ ) e o desvio padrão das diferenças ( $S_D$ ) para utilizá-las na expressão de t para amostras dependentes. Os resultados das diferenças são apresentados a seguir:

PESSOA	A	B	C	D	E	F	G	H
Antes	6,6	6,5	9	10,3	11,3	8,1	6,3	11,6
Depois	6,8	2,4	7,4	8,5	8,1	6,1	3,4	2
Diferença (antes – depois)	-0,2	4,1	1,6	1,8	3,2	2	2,9	9,6

Como as amostras são dependentes, usaremos a estatística t da seguinte forma:

$$t = \frac{\bar{D} - d_0}{S_D / \sqrt{n}} = \frac{3,125 - 0}{2,9114 / \sqrt{8}} = 3,03$$



O valor  $t_t = 1,895$ , que divide a  $RRH_0$  e  $RAH_0$ , foi encontrado na tabela t quando procurávamos o grau de liberdade, 7 graus de liberdade ( $n - 1$ , onde  $n$  é o número de indivíduos avaliados) e  $\alpha = 0,05$ . Como  $t$  calculado ( $t = 3,03$ ) pertence a  $RRH_0$ , podemos considerar que os valores de produtividade eram maiores antes e, assim, a redução na gratificação influenciou na produtividade dos funcionários da prefeitura.

## TESTE DE HIPÓTESE PARA A DIFERENÇA ENTRE PROPORÇÕES

Em diversas situações, o que nos interessa é saber se a proporção de sucessos (evento de interesse) em duas populações apresenta a mesma proporção ou não. Nesse caso, os dados seguem uma Distribuição de proporção **Bernoulli** com média  $p$  e variância  $pq$ . Portanto, a expressão da estatística teste (no caso utilizaremos a distribuição de  $Z$ ) será dada por:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}$$

Onde:

$\hat{p}_1$  e  $\hat{p}_2$ : correspondem à proporção de sucesso nas amostras 1 e 2, respectivamente; e

$p_1$  e  $p_2$ : correspondem à **proporção de sucesso** nas populações 1 e 2, respectivamente.

*Vejamos como aplicar o teste da diferença de proporções.*

**Situação:** uma empresa de pesquisa de opinião pública selecionou, aleatoriamente, 500 eleitores da Bahia e 600 de Pernambuco, e perguntou a cada um se votaria ou não no candidato Honesto Certo nas próximas eleições presidenciais. Responderam afirmativamente 80 eleitores da Bahia e 150 eleitores de

Vimos sobre a Distribuição de Bernoulli na Unidade 5. Você pode retomar esse conceito.

Você deve se lembrar de que a proporção de fracasso ( $q$ ) é dada por um menos a proporção de sucesso.



Pernambuco. Existe alguma diferença significativa entre as proporções de eleitores a favor do candidato nos dois Estados? Use nível de significância igual a 6%.

**Como fazer:**

$$\text{Bahia: } n = 500; \hat{p} = \frac{80}{500} = 0,16; \hat{q} = 0,84$$

$$\text{Pernambuco: } n = 600; \hat{p} = \frac{150}{600} = 0,25; \hat{q} = 0,75$$

Vamos estabelecer as hipóteses:

$$H_0: p_B = p_P \rightarrow p_B - p_P = 0$$

$$H_1: p_B \neq p_P \rightarrow p_B - p_P \neq 0$$

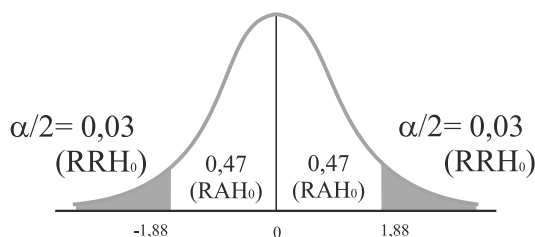
Aqui, seguem as mesmas considerações vistas anteriormente para a formulação das hipóteses.

$$\alpha = 0,06$$


A estatística usada será Z.

$$Z_c = \frac{(\hat{p}_B - \hat{p}_P) - (p_B - p_P)}{\sqrt{\left(\frac{\hat{p}_B \cdot \hat{q}_B}{n_B}\right) + \left(\frac{\hat{p}_P \cdot \hat{q}_P}{n_P}\right)}} = \frac{(0,16 - 0,25) - (0)}{\sqrt{\left(\frac{0,16 \cdot 0,84}{500}\right) + \left(\frac{0,25 \cdot 0,75}{600}\right)}} =$$

$$\frac{-0,09}{\sqrt{0,0002688 + 0,0003125}} = \frac{-0,09}{0,024} = -3,73$$



O valor 1,88 foi encontrado no **interior** da tabela Z procurando 0,4699.



Veja que 0,47 não existe na tabela, então, optamos pelo valor mais próximo.

Como Z calculado está na região de rejeição de  $H_0$  (menor que  $-1,88$ ), rejeitamos  $H_0$  e, portanto, podemos afirmar com 94% de certeza que existe diferença significativa entre as proporções de eleitores a favor do candidato nos dois Estados.

## TESTE DO QUI-QUADRADO DE INDEPENDÊNCIA

O teste do qui-quadrado de independência está associado a duas variáveis qualitativas, ou seja, a uma análise bidimensional. Muitas vezes, queremos verificar a relação de dependência entre as duas variáveis qualitativas a serem analisadas.

Nesse caso, procuramos calcular a frequência de ocorrência das características dos eventos a serem estudados. Por exemplo, podemos estudar a relação entre o sexo de pessoas (masculino e feminino) e o grau de aceitação do governo estadual (ruim, médio e bom). Então, obteremos, por exemplo, o número de pessoas (frequência) que são do sexo feminino e que acham o governo bom. Todos os cruzamentos das duas variáveis são calculados.

Vamos apresentar a você, como exemplo, os possíveis resultados da situação sugerida anteriormente (dados simulados).

GRAU DE ACEITAÇÃO				
SEXO	RUIM	MÉDIO	BOM	TOTAL
Masculino	157	27	74	258
Feminino	206	0	10	216
Total	363	27	84	474

Podemos determinar o grau de associação entre essas duas variáveis, ou seja, determinar se o grau de aceitação do governo depende do sexo ou se existe uma relação de dependência.

As hipóteses a serem testadas são:

$H_0$ : variável linha independe da variável coluna (no exemplo anterior, o grau de aceitação independe do sexo das pessoas).

$H_1$ : variável linha está associada à variável coluna (no exemplo anterior, o grau de aceitação depende do sexo das pessoas).

A estatística de qui-quadrado será dada por meio da seguinte expressão:

$$\chi^2_c = \sum_{i=1}^k \frac{(fo_i - fe_i)^2}{fe_i} = \frac{(fo_1 - fe_1)^2}{fe_1} + \frac{(fo_2 - fe_2)^2}{fe_2} + \dots + \frac{(fo_k - fe_k)^2}{fe_k}$$

Onde:

$k$  corresponde ao número de classes (frequências encontradas). Você pode verificar que **fo** corresponde à frequência observada, ou seja, ao valor encontrado na tabela de contingência.

Já **fe** corresponde à frequência esperada caso as variáveis não tenham nenhuma relação de dependência, ou seja, caso as duas variáveis sejam independentes. Por causa dessa definição, o cálculo da frequência esperada (**fe**) será obtido por:

$$fe = \frac{(total\ linha) \cdot (total\ coluna)}{total\ geral}$$

Nesse caso, os graus de liberdade ( $v$ ), para que possamos olhar a tabela de qui-quadrado, são dados por:

$$v = (h-1) (k-1) \text{ nas tabelas com } h \text{ linhas e } k \text{ colunas}$$

$$(\text{no exemplo anterior: } v = (2-1) \times (3-1) = 2 \text{ graus de liberdade})$$

Então, para cada célula da tabela de contingências, você irá calcular a diferença entre **fe** e **fo**. Essa diferença é elevada ao quadrado para evitar que as diferenças positivas e negativas se anulem. A divisão pela frequência esperada é feita para obtermos diferenças em termos relativos.

*Vamos entender melhor o teste de qui-quadrado do tipo independência por meio da análise de outra situação.*

**Situação:** o gestor de uma prefeitura deseja saber como seus funcionários atuam no uso do MSN durante o trabalho. Para realizar um programa de conscientização, os gestores públicos precisam saber se o fato de os funcionários usarem pouco ou muito o MSN durante o trabalho depende do sexo das pessoas. Mediante essa informação, a gestão pode definir se fará programas de conscientização para homens e mulheres de forma separada ou em conjunto (um único programa). Para testar essa hipótese, foram selecionados, ao acaso, 96 funcionários de ambos os sexos que usavam pouco ou muito o MSN em razão dessas características na população. Verifique, com uma significância de 5%, a hipótese do gestor público.

USO DO MSN SEXO		
	POUCO	MUITO
Homem	8	32
Mulher	16	40

**Resolução:**

Definindo primeiro as hipóteses  $H_0$  e  $H_1$ .

$H_0$ : uso do MSN independe do sexo.

$H_1$ : uso do MSN depende do sexo.

Agora, iremos calcular as frequências esperadas, que são os valores que estão entre parênteses. Confira os cálculos das outras frequências esperadas cujos valores (**fe**) aparecem entre parênteses.

USO DO MSN			
SEXO	POUCO	MUITO	
Homem	8 (10)	32 (30)	40
Mulher	16 (14)	40 (42)	56
	24	72	96

$$\frac{56 \cdot 24}{96} = 14$$

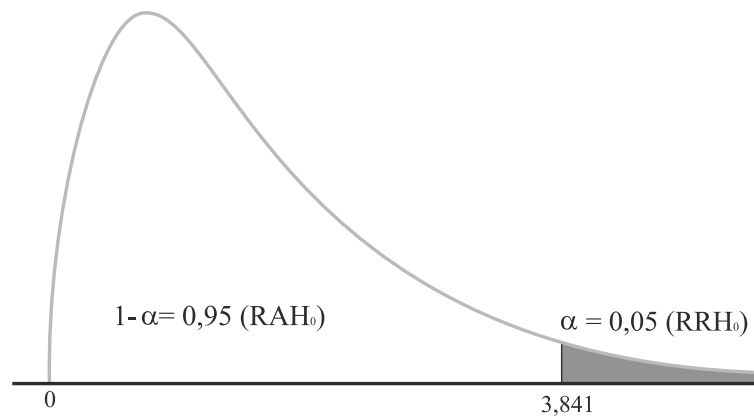
Agora, basta substituir os valores das frequências esperadas e observadas de todas as classes.

$$\chi^2_c = \sum_{i=1}^k \frac{(fo_i - fe_i)^2}{fe_i} = \frac{(8-10)^2}{10} + \dots + \frac{(40-42)^2}{42} = 0,914$$

O valor do grau de liberdade é apresentado a seguir:

$$v = (2 - 1) \cdot (2 - 1) = 1 \text{ gl}$$

Considerando um  $\alpha = 0,05$  e olhando na tabela de qui-quadrado para 1 grau de liberdade, teremos:

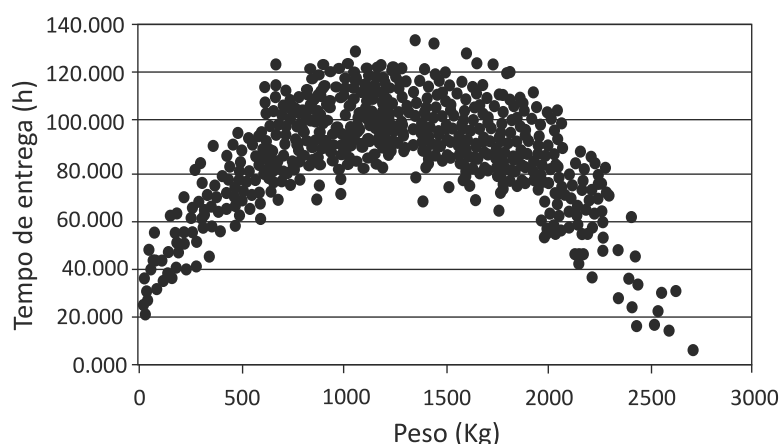


Como o valor calculado (0,914) foi menor do que o tabelado (3,841), então o calculado caiu na região de aceitação de  $H_0$ . Portanto, não temos indícios para rejeitar a hipótese  $H_0$ , ou seja, o uso do MSN independe do sexo dos funcionários. Dessa forma, o gestor pode fazer um único programa de conscientização tanto para homens quanto para mulheres.

## ASSOCIAÇÃO ENTRE VARIÁVEIS

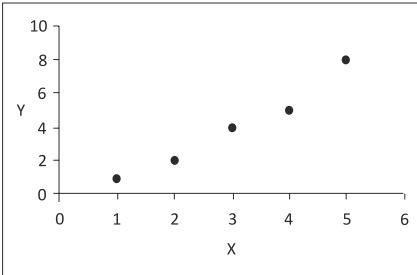
Para verificar o grau de relacionamento entre duas variáveis, ou seja, o grau de associação entre elas, devemos estudar um coeficiente chamado de coeficiente de correlação. Existem vários coeficientes de correlação e, cada um deles, aplicado em casos específicos. Aqui, iremos estudar o coeficiente de correlação de Pearson ( $r$ ).

Para que possamos ter uma ideia da associação entre as variáveis que estamos estudando, iremos utilizar um gráfico de dispersão como o apresentado, a seguir, pelo qual podemos constatar a relação entre as variáveis: o peso de um pacote e o seu tempo de entrega.

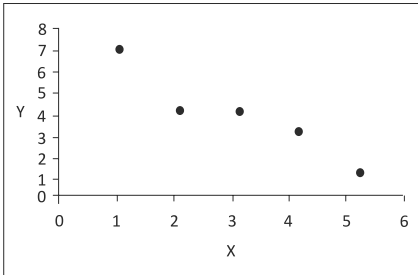


As estimativas de correlação podem ser positivas (à medida que a variável  $x$  aumenta a variável  $y$  também aumenta) ou negativas (à medida que a variável  $x$  aumenta a variável  $y$  diminui), como você pode ver a partir dos dados e dos gráficos a seguir:

Positiva		Negativa	
1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5



Correlação Positiva



Correlação Negativa

O coeficiente de correlação de Pearson ( $r$ ) nos dá uma ideia da variação conjunta das variáveis analisadas e pode assumir valores de  $-1$  a  $+1$ .

Veja a expressão por meio da qual podemos obter o coeficiente de correlação de Pearson:

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \cdot \sum y_i}{n}}{\left[ \sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] \cdot \left[ \sum y_i^2 - \frac{(\sum y_i)^2}{n} \right]}$$

A ocorrência de um valor de  $r = 0$  ou próximo de zero indica apenas que não há correlação **linear** entre as variáveis, porque pode existir uma forte relação não linear entre as variáveis, como no gráfico de dispersão do peso do pacote e o tempo de entrega, na qual temos uma relação não linear.

Vejam as características que o coeficiente de correlação de Pearson pode apresentar:

- ▶ seus valores estão compreendidos entre  $-1$  e  $1$ ;
- ▶ se o coeficiente for positivo, as duas características estudadas tendem a variar no mesmo sentido.

No exemplo que iremos trazer mais adiante, você encontrará a explicação dos somatórios dessa expressão. Não se preocupe!



- ▶ se o sinal for negativo, as duas características estudadas tendem a variar em sentido contrário;
- ▶ a relação entre duas variáveis é tanto mais estreita quanto mais o coeficiente se aproxima de 1 ou -1; e
- ▶ o valor de  $r$  é uma estimativa do parâmetro  $\rho$  (**rho**), da mesma forma que a média  $\bar{x}$  é uma estimativa de  $\mu$ . Para testar se o valor de  $r$  é estatisticamente igual ao parâmetro de uma população em que  $\rho$  (**rho**) = 0, podemos empregar o teste  $t$  definido por:

$$t_c = \frac{r - \rho}{\sqrt{1 - r^2}} \cdot \sqrt{n - 2}$$

onde:

$n$  : número total de pares;

$r^2$  : coeficiente de correlação ao quadrado; e

$\rho$  : parâmetro da correlação populacional (considerado igual a zero).

A hipótese  $H_0$  será de que  $\rho$  (**rho**) = 0 e a hipótese  $H_1$ , que iremos utilizar, será de que  $\rho$  (**rho**)  $\neq$  0.

*Vamos analisar a situação, a seguir, para entender melhor esse coeficiente.*

### Situação:

Vamos determinar o coeficiente de correlação entre a porcentagem de aplicação do total de recursos com Educação em uma prefeitura ( $x$ ) e o grau de conhecimento médio da população da cidade ( $y$ ). Para isso, foram avaliadas dez cidades.

PORCENTAGEM DE APLICAÇÃO DO TOTAL DE RECURSOS COM EDUCAÇÃO EM UMA PREFEITURA	GRAU DE CONHECIMENTO MÉDIO DA POPULAÇÃO DA CIDADE
5	70
10	40
20	27
30	22
40	18
50	16
60	15
70	14
80	13
90	12

Para obtermos a estimativa de correlação, precisamos calcular todos os somatórios presentes na expressão:

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \cdot \sum y_i}{n}}{\sqrt{\left[ \sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] \cdot \left[ \sum y_i^2 - \frac{(\sum y_i)^2}{n} \right]}}$$

Calculando os somatórios, teremos:

Somatório de todos os valores de x:

$$\sum x_i = x_1 + x_2 + ..... + x_{10} = 5 + 10 + ..... + 90 = 455$$

Somatório de todos os valores de x elevados ao quadrado:

$$\sum x_i^2 = x_1^2 + x_2^2 + ..... + x_{10}^2 = 5^2 + 10^2 + ..... + 90^2 = 28525$$

Somatório de todos os valores de y:

$$\sum y_i = y_1 + y_2 + ..... + y_{10} = 70 + 40 + ..... + 12 = 247$$

Somatório de todos os valores de y elevados ao quadrado:

$$\sum y_i^2 = y_1^2 + y_2^2 + ..... + y_{10}^2 = 70^2 + 40^2 + ..... + 12^2 = 9027$$

Somatório de todos os valores obtidos por meio do produto dos valores de x e y de cada cidade:

$$\sum x_i y_i = x_1 y_1 + x_2 y_2 + \dots + x_{10} y_{10} =$$

$$\sum x_i y_i = 5 \cdot 70 + 10 \cdot 40 + \dots + 90 \cdot 12 = 7470$$

Substituindo esses valores na expressão, teremos:

$$r = \frac{7470 - \frac{455 \cdot 247}{10}}{\sqrt{\left[28525 - \frac{(455)^2}{10}\right] \cdot \left[9027 - \frac{(247)^2}{10}\right]}} = \frac{-3768,5}{4784,28} = -0,7877$$

O valor de  $r = -0,7877$  indica que existe uma associação inversa (negativa) e de média magnitude entre a variação da porcentagem de aplicação do total de recursos com educação em uma prefeitura e o grau de conhecimento médio da população da cidade, ou seja, nesta população de cidades, provavelmente os recursos da educação não estão sendo bem empregados, já que a relação foi negativa quando se esperava uma relação positiva.

Para verificarmos se esse resultado é significativo, vamos fazer o seguinte teste de hipótese:

$$H_0: \rho (\mathbf{rho}) = 0$$

$$H_1: \rho (\mathbf{rho}) \neq 0.$$

Iremos calcular a estatística por meio da expressão:

$$t_c = \frac{r - \rho}{\sqrt{1 - r^2}} \cdot \sqrt{n - 2}$$

Substituindo os valores na expressão, teremos:

$$t_c = \frac{-0,78770 - 0}{\sqrt{1 - 0,7877^2}} \cdot \sqrt{10 - 2} = -1,25 \cdot 2,82 = 3,525$$


Olhando na tabela de t para 8 graus de liberdade (10-2) e um  $\alpha=0,025$ , já que estamos considerando uma significância de 0,05 e o nosso teste é bilateral, teremos um valor tabelado de 2,306. Verificamos que o valor calculado de 3,525 está na região de rejeição da hipótese  $H_0$  e, portanto, iremos aceitar a hipótese

$H_1$ , ou seja, de que  $\rho$  (**rho**)  $\neq 0$ . Então, o resultado encontrado na amostra ( $r$ ) não foi fruto do acaso, considerando uma significância de 5%.

Devemos ter cuidado na interpretação do coeficiente de correlação, pois este não implica necessariamente uma medida de causa e efeito. É mais seguro interpretar o coeficiente de correlação como uma medida de associação. Por exemplo, podemos encontrar uma correlação muito alta entre o aumento dos salários dos professores e o consumo de bebidas alcoólicas através de uma série de anos em uma região. Esse valor de  $r$  encontrado foi alto apenas porque pode ser que ambas as variáveis tenham sido afetadas por uma causa comum, ou seja, a elevação do padrão de vida de uma região.

### Complementando...

Através do link que apresentamos a seguir, você poderá fazer os testes de hipóteses e de estimativas de correlação de Pearson.

 *Programa estatístico Bioestat*. Disponível em: <[http://www.mamiraua.org.br/download/Default.aspx?dirpath=e:\home\mamiraua\Web\download\BioEstat 5 Portugues&tipo=diretorio](http://www.mamiraua.org.br/download/Default.aspx?dirpath=e:\home\mamiraua\Web\download\BioEstat%20Portugues&tipo=diretorio)>. Acesso em: 29 nov. 2010.

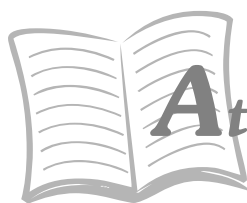
# Resumindo



Nesta Unidade, conhecemos os principais testes de hipóteses e vimos suas aplicações no dia a dia da gestão de empresas públicas.

Apresentamos a estrutura de um teste de hipótese, de testes de hipóteses para médias, de diferença entre médias e de diferença entre proporções.

Verificamos que o teste de qui-quadrado pode ser utilizado para medir a dependência entre variáveis qualitativas. Dessa forma, você terá plenas condições de aplicar e de interpretar um teste estatístico de maneira correta.



## Atividades de aprendizagem

Chegou o momento de analisarmos se você entendeu o que estudamos até aqui! Para saber, procure, resolver as atividades propostas a seguir. Lembre-se: você pode contar com o auxílio de seu tutor.

1. Um fabricante afirma que seus pneus radiais suportam em média uma quilometragem superior a 40.000 km. Uma prefeitura compra os pneus desse fabricante. Existe uma dúvida no setor de compras da prefeitura: “A afirmação do fabricante está correta?”. Para testar essa afirmação, a prefeitura selecionou uma amostra de 49 pneus. Os testes, nessa amostra, forneceram uma média de 43.000 km. Sabe-se que a quilometragem de todos os pneus tem desvio padrão de 6.500 km. Se o comprador (gestor público) testar essa afirmação ao nível de significância de 5%, qual será sua conclusão?
2. Duas técnicas de cobrança de impostos são aplicadas em dois grupos de funcionários do setor de cobrança de uma prefeitura. A técnica A foi aplicada em um grupo de 12 funcionários, resultando em uma efetivação média de pagamento de 76% e uma variância de 50%. Já a técnica B foi aplicada em um grupo de 15 funcionários, resultando em uma efetivação média de 68% e uma variância de 75%. Considerando as variâncias estatisticamente iguais e com uma significância de 0,05, verifique se as efetivações de pagamento são estatisticamente iguais.

3. Um secretário da Educação de uma prefeitura deseja saber se há, no futuro, profissionais promissores em escolas de regiões pobres e de regiões ricas. Uma amostra de 16 estudantes de uma zona pobre resultou, em um teste específico, uma média de 107 pontos e um desvio padrão de 10 pontos. Já 14 estudantes de uma região rica apresentaram uma média de 112 pontos e um desvio padrão de 8 pontos. Você deve verificar se a média dos pontos dos dois grupos é diferente ou igual a fim de que o empresário possa saber se ele deve investir em qualquer uma das áreas ou se uma delas é mais promissora (primeiro verifique se as variâncias são estatisticamente iguais ou diferentes).

## *Respostas das Atividades de aprendizagem*

### Unidade 1

1. a) Qualitativa Nominal.  
b) Qualitativa Ordinal.  
c) Quantitativa Discreta.  
d) Quantitativa Contínua.
2. a) Amostragem Sistemática.  
b) Amostragem por Conglomerado.  
c) Amostragem Estratificada.  
d) Amostragem Aleatória Simples.  
e) Amostragem Sistemática.  
f) Amostragem Aleatória Simples.  
g) Amostragem Estratificada.  
h) Amostragem por Conglomerado.

### Unidade 2

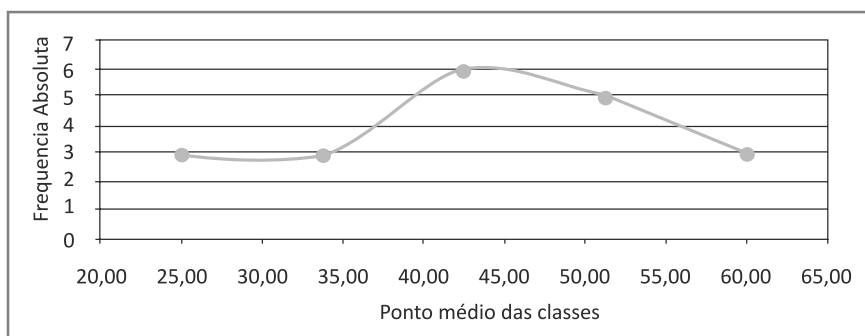
1. a)  $n = 20$ ,  $A = 35$ ,  $k = 5$  (aproximadamente),  $c = 8,75$ ,  $Li_1a = 20,925$ .

CLASSES	FREQUÊNCIAS ABSOLUTAS
20,625  — 29,375	3
29,375  — 38,125	3
38,125  — 46,875	6
46,875  — 55,625	5
55,625  — 64,375	3
Total	20



CLASSES	FREQUÊNCIAS ACUMULADA
20,625  — 29,375	3
29,375  — 38,125	6
38,125  — 46,875	12
46,875  — 55,625	17
55,625  — 64,375	20

b)



### Unidade 3

$$1. \bar{x} = \frac{\sum x_i}{n} = \frac{7 + 42 + 37 + 25 + 38 + \dots + 33}{27} = 26,6$$

$$Md = X_{\left(\frac{n+1}{2}\right)} = X_{\left(\frac{27+1}{2}\right)} = X_{14} = 25 \text{ (elemento de posição 14º)}$$

Mo = 18,23,25 e 28, todos esses valores tem frequência 2 (multimodal)

$$\text{Variância: } S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(7-26,6)^2 + \dots + (33-26,6)^2}{27-1} = 94,33$$

$$\text{Desvio Padrão: } S = \sqrt{S^2} = \sqrt{94,33} = 9,7$$

Coefficiente de Variabilidade:

$$CV = \frac{S}{\bar{x}} \cdot 100 = \frac{9,7}{26,6} \cdot 100 = 36,47\%$$

2. Média	21.0
Mediana	18.0
Moda	10.0
Desvio padrão	12.0
Coeficiente de Variação	57.3

#### Unidade 4

1. R:  $1 - (1/3 * 1/5 * 3/10) = 0,98$ .

2. a) R: 0,125.

b) R: 0,0694.

c) R: 0,1388.

3. a) R: 60/100.

b) R: 40/100.

c) R: 24/100.

d) R: 76/100.

#### Unidade 5

1. R:  $P(X = 5) = C_{20}^5 0,1^5 0,9^{15} = 0,03192$ .

2. Distribuição binomial com  $n = 4$  e  $p = \frac{1}{2}$

a) R:  $P(x=2) \cdot 2000 = 0,3750 \cdot 2000 = 750$  famílias.

b) R:  $[P(1) + P(2)] \cdot 2000 = (0,25 + 0,375) \cdot 2000 = 1250$  famílias.

c) R:  $P(0) \cdot 2000 = 0,0625 \cdot 2000 = 125$  famílias.

3. R:  $1 - [P(0) + P(1)]$ , em que a distribuição de probabilidade é uma Poisson com parâmetro lambda.

a)  $\lambda = 1,4$  R= 0,40817

b)  $\lambda = 2,8$  R=0,76892

c)  $\lambda = 5,6$  R=0,97559

$$4. \text{ Para } X = 2200 \rightarrow Z = \frac{X - \mu}{\sigma} = \frac{2200 - 2000}{200} = 1,00$$

$$\text{Para } X = 1700 \rightarrow Z = \frac{X - \mu}{\sigma} = \frac{1700 - 2000}{200} = -1,50$$

$$5. a) X = 20 \rightarrow Z = 0$$

$$X = 24 \rightarrow Z = \frac{24 - 20}{5} = 0,8$$

$$P(20 < X < 24) = P(0 < Z < 0,8) = 0,2881 \text{ (28,81 \%)}.$$

$$b) X = 16 \rightarrow Z = \frac{16 - 20}{5} = -0,8$$

$$X = 20 \rightarrow Z = 0$$

$$P(16 < X < 20) = P(-0,8 < Z < 0) = P(0 < Z < 0,8) = 0,2881 = 28,81$$

$$c) X = 28 \rightarrow Z = (28 - 20) / 5 = 1,6$$

$$P(X > 28) = P(Z > 1,6) = 0,5 - 0,4452 = 0,0548$$

$$6. 1 - \alpha = 0,95 \rightarrow \alpha = 0,05 \rightarrow \alpha/2 = 0,025$$

$$e = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1,96 \cdot \frac{3}{\sqrt{100}} = 0,588$$

$$P(26,412 < \mu < 27,588) = 0,95$$

## Unidade 6

1. Sugestão: siga os passos para realizar um teste de hipótese:

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{43000 - 40000}{6500 / \sqrt{49}} = 3,23 \quad Z_{\alpha} = Z_{0,05} = 1,64$$

Conclusão: como o valor calculado foi maior do que o tabelado (1,64), ele caiu na região de rejeição de  $H_0$ .

$$2. H_0: \mu_A - \mu_B = 0 \quad H_0: \mu_1 - \mu_2 \neq 0$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{1/n_1 + 1/n_2}} = \frac{(76 - 68) - 0}{8 \sqrt{1/12 + 1/15}} = 2,56$$

$$t_{0,025} = 2,060$$

Conclusão: como o valor calculado foi maior do que o tabelado (2,060), ele caiu na região de rejeição de  $H_0$ .

$$3. H_0: \mu_1 - \mu_2 = 0 \quad H_0: \mu_1 - \mu_2 \neq 0$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{(112 - 107) - 0}{\sqrt{8^2/14 + 10^2/16}} = -1,52$$

$v = 29,7425 = 30$  (graus de liberdade obtido pela aproximação).

$$t_{0,025} = 2,042 \text{ (com 30 gl)}$$

Conclusão: como o valor calculado caiu na região de aceitação, as médias são estatisticamente iguais, o que indica que as duas regiões apresentam o mesmo potencial.

## CONSIDERAÇÕES FINAIS

Com os conhecimentos de estatística adquiridos ao longo deste livro, você agora já pode imaginar quantas análises estatísticas podem ser feitas. A análise de dados está presente até em uma simples ligação telefônica que uma empresa de crédito faz para você. A empresa cruza informações como sexo, renda mensal e hábitos de consumo para oferecer um produto na medida certa.

Para fazer tudo isso, é necessário, entretanto, conhecimento básico de estatística para que empresas de Gestão Pública ou não venham a descobrir como transformar quantidades de números e de gráficos em informações que servirão para reduzir os custos e aumentar os lucros. O problema é que falta gente qualificada e com conhecimento de mercado para realizar as análises de dados. Para você trabalhar com conceitos estatísticos em qualquer setor, é necessário desenvolver um raciocínio lógico e, também, administrar informações, além de procurar entender como e por que as coisas acontecem.

Para decidir algo importante, é necessário avaliar os riscos e as oportunidades. Para que isso seja feito com muita precisão, é necessária a estatística!

Assim, você poderá aplicar os conhecimentos de estatística aprendidos em áreas, como a de Recursos Humanos, de Produção, Financeira e muitas outras que você irá identificar à medida que seus conhecimentos na área de Administração forem aumentando.

Espero que você tenha gostado de trabalhar com Estatística e que ela seja uma importante ferramenta a ser utilizada em seu dia a dia.

Um grande abraço e sucesso em sua vida profissional, com bastante estatística, é o que desejamos a você.

*Professor Marcelo Tavares*

## Referências



- ARANGO, Hector G. *Bioestatística: teórica e computacional*. Rio de Janeiro: Guanabara Koogan, 2001.
- BARBETTA, Pedro Alberto. *Estatística Aplicada às Ciências Sociais*. 4. ed. Florianópolis: Editora da UFSC, 2002.
- BEIGUELMAN, Bernardo. *Curso Prático de bioestatística*. Ribeirão Preto: Revista Brasileira de Genética, 1996.
- BRAULE, Ricardo. *Estatística Aplicada com Excel: para cursos de administração e economia*. Rio de Janeiro: Campus, 2001.
- BUSSAB, Wilton O.; MORETTIN, Pedro. *Estatística Básica*. São Paulo: Atual, 2002.
- COSTA NETO, Pedro Luiz de Oliveira. *Estatística*. São Paulo: Edgard Blucher, 2002.
- DOWNING, D.; CLARK, J. *Estatística Aplicada*. São Paulo: Saraiva, 2000.
- FONSECA, Jairo Simon da; MARTINS, Gilberto de Andrade. *Curso de Estatística*. Rio de Janeiro: LTC, 1982.
- FREUD, Jonh E.; SIMON, Gary A. *Estatística aplicada*. Bookman, 2000.
- HOUAISS, Instituto Antônio Houaiss. *Dicionário eletrônico Houaiss da Língua Portuguesa*. Versão monousuário, 3.0. Objetiva: junho de 2009. CD-ROM.
- LEVINE, David M.; BERENSON, Mark L.; STEPHAN, David F. *Estatística: teoria e aplicações usando o Microsoft Excel em português*. Rio de Janeiro: LTC, 2000.
- MORETTIN, Luiz Gonzaga. *Estatística Básica – Probabilidade*. São Paulo: Makron Books, 1999. 1 v.

\_\_\_\_\_. *Estatística Básica – Inferência*. São Paulo: Makron Books, 1999.  
2 v.

SOARES, José F.; FARIAS, Alfredo A.; CESAR, Cibele C. *Introdução à Estatística*. Rio de Janeiro: LTC, 1991.

SPIEGEL, Murray R. *Probabilidade e Estatística*. São Paulo: Mc Graw Hill, 1993.

STEVENSON, William J. *Estatística Aplicada à Administração*. São Paulo: Harper, 1981.

TRIOLA, Mário F. *Introdução à Estatística*. Rio de Janeiro: LTC, 1999.

WONNACOTT, T. H., WONNACOTT, R. J. *Estatística Aplicada à Economia e à Administração*. Rio de Janeiro: LTC, 1981.

## MINICURRÍCULO

### Marcelo Tavares

Possui Graduação (1989) e Mestrado (1993) pela Universidade Federal de Lavras, e Doutorado pela Escola Superior de Agricultura Luiz de Queiroz/USP (1998). Atualmente, é professor Associado II da Universidade Federal de Uberlândia (UFU). Tem experiência na área de Estatística Aplicada e atua, principalmente, nos seguintes temas: modelagem estatística, estatística, amostragem, controle de qualidade e estatística multivariada. Também foi coordenador do Curso de Especialização em Estatística Empresarial do Núcleo de Estudos Estatísticos e Biométricos da Faculdade de Matemática e, atualmente, é Coordenador da Universidade Federal de Uberlândia na Universidade Aberta do Brasil (UAB) e ministro das disciplinas de Estatística para o Curso de Administração da UFU.







ISBN 978-85-7988-099-5



**UENF**  
Universidade Estadual  
do Norte Fluminense



Universidade Federal Fluminense



**GOVERNO DO  
Rio de Janeiro**

SECRETARIA DE  
CIÊNCIA E TECNOLOGIA

**UNIVERSIDADE  
ABERTA DO BRASIL**

Ministério da  
Educação

GOVERNO FEDERAL  
**BRASIL**  
PAÍS RICO E PAÍS SEM POBREZA