José Evandeilton lopes

R NA PRÁTICA: Estatística Descomplicada com R

To my son,

without whom I should have finished this book two years earlier

Conteúdo

Lista de Tabelas

Lista de Figuras

.reveal section img{ border-color:#000; }

Introdução

Bem vindo ao módulo I de Estatística descomplicada com o R NA PRÁTICA. Este é o primeiro módulo desta série de quatro módulos e contem pará os principais conceitos estatísticos até análise descritiva. O objetivo maior desta parte é revisar os conceitos mais importantes do inicio dos estudos estatísticos. Abordaremos a parte conceitual com algumas definições e termos estatísticos, tabelas de frequências, bem como as principais medidas descritivas (média, mediana e outras). Veremos por fim, os principais gráficos estatísticos mais utilizados. Para reforçar os conhecimentos, faremos muitos exercícios práticos com apoio do R com foco em bases de dados reais do INEP[1] - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira do ano de 2017.

No final deste módulo você será capaz de:

- Definir estatística;
- Compreender os principais tipos de variáveis;
- Entender o que é e como utilizar tabelas de frequências;
- Trabalhar com as principais medidas estatísticas (media, mediana, desvio padrão, etc.);
- Entender correlação e covariância;
- Construir gráficos estatísticos para os tipos corretas de dados.

[1] http://portal.inep.gov.br/web/guest/dados

Motivação

A Estatística está em tudo! É atribuída a H.G. Wells[2] a seguinte frase:

> "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write. — H.G. Wells"

Que em português se traduz em algo como:

> "O conhecimento estatístico um dia será tão necessário para uma cidadania eficiente quanto a habilidade de ler e escrever."

Veja alguns exemplos de onde a estatística pode ser aplicada:

- **Na sua lista de compras:** Sim, é importante medir as diferenças de preços dos produtos e o percentual de seu salário gasto com alimentos pra não ficar sem grana por aluguel;
- **Na previsão do tempo:** Sair sem guarda-chuva ou de bicicleta num dia de chuva, nem pensar. Mas você não saberia o que fazer pela manhã se observasse o dia claro e sem sinais de chuva. Somente uma previsão estatística confiável seria capaz de te ajudar a decidir antecipadamente;
- **Na pesquisa científica:** Toda pesquisa científica usa estatística. Se não utilizar é porque não é científica mas sim, especulativa;

[2] https://www.sciencedirect.com/science/article/pii/0315086079901010

- **Em seguros:** Seguros de vida, de veículos, de saúde, dental, daquele bumbum siliconado[3]. Tudo envolve estatística. É através dela que os estatísticos constroem os melhores modelos com base em variáveis históricas e nas segmentações de perfis de cada grupo de indivíduos;
- **Em pesquisas médicas:** Controle de epidemias, vacinas, taxas de sobrevida, taxas de recuperação de tecidos e ossos, regressão de doenças infecciosas, testes de grupos de risco e uma infinidade de situações utilizam estatística para prever, aumentar e melhorar processos que trazem ganho para a saúde e vida das pessoas;
- **Em testes de qualidade:** A estatística bate forte na hora de uma empresa obter uma certificação ISO. Seja no controle estatístico de processos, gráficos de controle e/ou gestão da qualidade, sem esta ciência estes nichos não teriam os mesmos resultados;
- **No monitoramento de ataques de vírus:** A empresa Kaspesrki por exemplo, coleta dados globais de vírus e sintetiza em um mapa do globo virtual denominado cybermap[4]. Através dele o usuário pode navegar e obter estatísticas relevantes sobre atividades de vírus nos países pelo mundo.
- **No IDH e expectativa de vida:** O IDH (Índice de Desenvolvimento Humano) de um país é um bom indicador de desenvolvimento do mesmo. Não tem como calcular IDH sem modelos estatísticos.

Prerequisitos

Para fazer o melhor proveito deste material, temos as seguintes recomendações:

[3] https://www.moneytips.com/how-much-is-my-butt-worth

[4] https://cybermap.kaspersky.com/pt

- Tenha um computador com acesso á internet para poder assistir aos vídeos, fazer pesquisa e download dos materiais do curso;
- Dedique pelo menos 2 horas da sua semana para ler o material e resolver os exercícios propostos;
- Sempre que alguma dúvida surgir e não conseguir resolver com ajuda do material, contatar o professor via plataforma para obter suporte adicional;
- Algum conhecimento prévio da linguagem R ou lógica de programação.

Ambiente de trabalho

Este curso não tem a intenção de ensinar detalhadamente programação em R, pois o foco é em estatística. Contudo, sempre que necessário explicaremos algumas funções e comandos utilizados nos scripts. Além disso, para todos os exemplos e exercícios, deixaremos os códigos gerados como forma de estudo e revisão para que tudo possa ser replicado pelo aluno.

Se você sentir dificuldade em compreender algum conceito em linguagem R, recomendamos fazer o nosso curso R NA PRÁTICA: Data Wrangling com R para Ciência de Dados[5] ou qualquer outro do seu interesse para obter mais conhecimentos dobre a linguagem R.

Pacotes

Trabalharemos sempre que possível com o operador %>% (pipe) e com funções dos pacotes do tidyverse[6] e outros relacionados.

Trabalharemos com o software R[7] e com a IDE (Ambiente de Desenvolvimento Integrado, traduzindo do inglês) de desenvolvi-

[5] https://www.udemy.com/r-na-pratica-ciencia-de-dados/learn/v4/overview

[6] https://www.tidyverse.org/

[7] https://cloud.r-project.org/

mento em R RStudio Desktop[8]. Nos links você poderá baixar os dois programas e configurar de acordo com o seu sistema operacional.

Especialmente para o R NA PRÁTICA, desenvolvemos o pacote `rnp` que poderá ser baixado direto to `github`. Este pacote possui recursos deste livro e também de outros materiais do R NA PRÁTICA. Rode o comando abaixo para instalar, caso ainda não tenha o pacote.

```
require(devtools, quietly = TRUE)
if(!require("rnp")) {
 devtools::install_github("evandeilton/rnp")
}
```

Não nos restringimos apenas a este, remendamos caso ainda não tenha, instalar os pacotes listados a seguir:

```
pacotes <- c("tidyverse","lubridate", "magrittr","broom",
        "stringr", "plotly", "ggplot2","data.table")
for(i in pacotes){
 if(!require(package = i, character.only = TRUE)){
  install.packages(i, dependencies = TRUE)
 }
}
```

Conjuntos de dados

Trabalharemos com conjuntos de dados do Censo do Ensino Superior no Brasil feito anualmente pelo INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira)[9], especificamente para o ano de 2017. Os dados podem ser baixados diretamente na seção de microdados do site neste link Censo da Educação Superior de 2017[10], ou se preferir, baixar diretamente com o

[8] https://www.rstudio.com/products/rstudio/download/#download

[9] www.inep.gov.br/

[10] http://portal.inep.gov.br/microdados

comando abaixo. Em seguinda descompactar os dados utilizando seu descompactador favorito.

Note que o argumento `salvar` na função `rnp_get_inep_censo()` dever conter o caminho para a pasta onde os dados baixados serão salvos. Se não for passada uma pasta ou seja, `salvar = NULL` o R baixará os dados na pasta da seção onde o R foi carregagado. Para saber qual é este local digite `getwd()` no console do R.

```
# Informa o ano dos dados e o local completo de onde deseja salvar os dados
# conforme exemplos abaixo. Recomendamos criar uma pasta para salvar os
# dados e tornar mais acessíveis nas aulas.
rnp::rnp_get_inep_censo(ano = 2017, salvar = "Dados/INEP/")
```

Você poderá também ler os dados sem decompactar com o pacote `readr`, mas recomendamos descompartar para leitura mais rápida com a função `rnp_read()`.

```
base_curso <- readr::read_delim("Dados/INEP/DADOS/DM_CURSO.zip",
                delim = "|", locale = locale(encoding = "Latin1"))
```

1

A Estatística

Estatística é um ciência vasta e para tentar entender melhor do que ela trata, vamos conferir algumas definições formais e outras não tão formais.

1.1 O que é estatística?

Segundo o dicionário Aurélio:

> "Ramo das matemáticas aplicadas cujos princípios decorrem da teoria das probabilidades e que tem por objeto o estudo, bem como o agrupamento metódico, de séries de fatos ou de dados numéricos. — Dicionário Aurélio"

FIGURA 1.1: Tenho uma dúvida

Calma meu jovem. Vamos ver outras definições então!

> A Estatística é a ciência que fornece métodos para a coleta, organização, descrição ou apresentação, análise e interpretação de dados para suportar a tomada de decisão. — Adaptação

---

O dicionário do Google também define como:

> Ramo da matemática que trata da coleta, da análise, da interpretação e da apresentação de massas de dados numéricos. — Google

---

Temos mais algumas frases a seguir!

> A estatística é a arte de nunca ter que dizer que você está errado. — C. J. Bradfield

---

Ou ainda minha favorita!

> Estatística é a arte de torturar os números até que eles confessem. — Desconhecido

---

FIGURA 1.2: Não tenho mais dúvidas

Nos dias atuais, cada vez mais as pessoas, empresas e governos estão tomando decisões com base em dado e como podemos ver das definições acima, a ciência estatística possui relação direta com a descoberta de conhecimento através de dados. Isso faz com que esta ciência seja de total relevância para quem quer descobrir e entender o que seus dados estão tentando esconder.

O processo de coleta, organização, descrição dos dados, cálculo e interpretação de estatísticas pertencem à **Estatística descritiva**, enquanto a análise e a interpretação dos dados, associado a uma margem de incerteza, geralmente associada a uma **amostra** ficam por conta da **Estatística inferencial ou indutiva** fortemente fundamentada na teoria das probabilidades que veremos em módulo específico.

1.2 Fases do trabalho estatístico

No contexto das empresas, o planejamento estatístico não existe ou não é feito de forma rigorosa e isso ocorre muitas vezes devido ao fato de a empresa gerar dados de acordo com seus próprios produtos e processos sem se dar conta de que os dados poderão gerar valor. Em muitos casos, o acúmulo desordenado de dados dificulta

a obtenção de conhecimento. Empresas competitivas entendem o valor dos dados e investem esforço, tempo e dinheiro para gerar bases de dados robustas e orientadas para o trabalho de *Data Science*. Por isso, sempre que possível devemos nos esforçar para blindar o trabalho estatístico seguindo das recomendações científicas, pois isso nos garantirá que as melhores ferramentas e técnicas serão empregadas desde o início da pesquisa/estudo até a comunicação dos resultados.

Como planejar pesquisas estatísticas e experimentos são temas profundos e que não cobriremos neste texto, mas para os leitores interessados sugerimos os textos de (Gil, 2008) para Métodos e Técnicas de Pesquisa; Fundamentos de metodologia científica do (Köche, 2016). Para planejamento de experimentos, sugerimos o texto de (Montgomery, 2017).

- **Definição do problema**: A figura 1.3 exibe um *roadmap* do processo estatístico. Este processo inicia sempre pela **Definição do problema**. Nesta fase inicial é preciso delimitar muito bem o problema de pesquisa. Seja ele simples ou complexo, se esta etapa for mal pensada, poderá conduzir a resultados inesperados ou questões não respondidas;

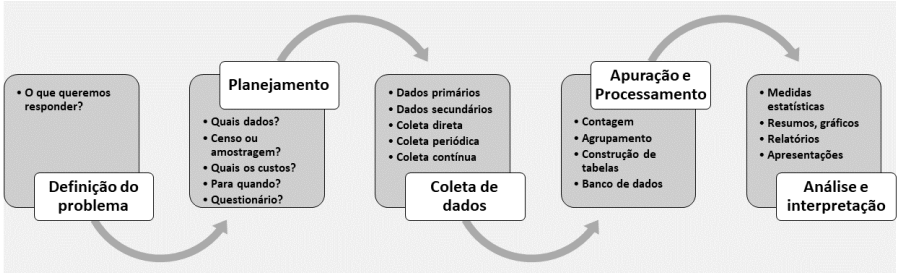


FIGURA 1.3: Fases do trabalho estatístico

- **Planejamento:** Nesta etapa é preciso responder muitos porquês, afinal não podem restar dúvidas que comprometam futuramente os trabalhos. Deve-se responder questões sobre quais dados utilizar, tamanho e tipo das amostras, custos do projeto, tempo de execução, ferramentas, pessoal qualificado e uma série de questões relacionadas ao projeto.

- **Coleta dos dados:** Na etapa de coleta, o pesquisador deve estar atento ás fontes de dados e à qualidade dos mesmos. Seja através de questionários ou de bases já montadas, os dados precisam ser confiáveis e consistentes. Os dados devem ser catalogados respeitando-se o tipo de coleta, se periódica, se contínua, se os dados são primários (gerados pela própria empresa ou pesquisador) ou secundários (gerados por terceiros).
- **Apuração:** Esta etapa serve para qualificar os dados e nela são feitas contagens, tabulações, agrupamentos e inserção em bancos de dados para o trabalho de análise.
- **Análise e interpretação:** Nesta fase todo o ferramental estatístico entra em ação para analisar e descobrir as relações entre as variáveis buscando responder às hipóteses do problema de pesquisa. A geração de relatórios, apresentações e painéis (*dashboards*) fazem parte desta etapa e auxiliarão na tomada de decisão e na geração de conhecimento.

1.3 Como utilizar estatística?

Esta é uma pergunta vastas, mas com base em nossa experiência e na comunidade estatística, segue nove boas práticas para direcionar o pesquisador.

- **Estatística deve ser utilizada para ajudar a responder perguntas científicas:** É importante inserir estatística desde o planejamento do experimento até a condução das análises e por fim a compilação dos conhecimentos adquiridos com base nos dados gerados pela pesquisa.
- **Pessoas se enganam. Dados não:** Isso é verdade, mas tome cuidado, pois sinais sempre vêm com ruído. Por isso é fundamental entender muito bem o problema de pesquisa e conhecer seus dados para saber diferenciar dado bom de ruído. Questione as fontes e as formas que te foram apresentadas.
- **Planejamento com foco no presente e no futuro:** Este

é um dos princípios mais violados nas ciências, portanto fique alerta. Fazer as perguntas certas e obter as respostas adequadas pode evitar perda de tempo, dinheiro e dores de cabeça na hora de analisar os dados obtidos de experimento mal delineado.

- **Atenção à qualidade dos dados:** Se você tiver acesso ao processo de planejamento e coleta de dados seja cuidadoso e tenha em mente os impactos futuros na condução das análises. Dados ruim pode arruinar um estudo ou conduzir a resultados errados.

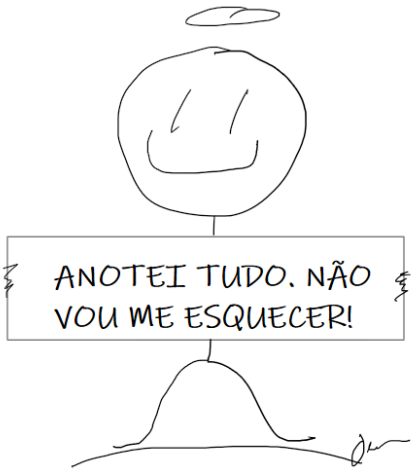


FIGURA 1.4: Anotei tudo

- **Estatística não é só técnica:** Análise estatística é mais que um software. O software estatístico fornece ferramentas para auxiliar a análise, não para defini-las. O contexto científico é crítico, e a chave para a análise estatística baseada em princípios é aproximar os métodos analíticos das questões científicas e de negócio.
- **Busque a simplicidade:** As pessoas não gostam de complexidade. Uma boa parte dos modelos estatísticos exige formulação simples. Em muitos casos, uma simples análise descritiva resolve o problema. Tenha em mente que um grande número de medições, dados ausentes, erros, vieses de amostragem e outros

fatores podem aumentar a complexidade do modelo e tornar o estudo impraticável.

- **Calcule a variabilidade:** faz parte da análise estatística justamente ajudar a avaliar a incerteza, muitas vezes na forma de um erro padrão ou intervalo de confiança, e um dos grandes sucessos da modelagem estatística e inferência é que ela pode fornecer estimativas de erros padrão dos mesmos dados. Ao apresentar resultados, é essencial fornecer alguma noção de incerteza estatística envolvida em seu estudo.
- **Verifique as suposições das suas técnicas:** É importante entender as suposições por trás dos métodos estatísticos e fazer o que for possível para entender e avaliar essas suposições. Não deixe que o software faça o papel do analista, ele apenas deve auxiliá-lo no processamento dos dados e nos cálculos. A validação das técnicas e as interpretações são sempre por conta do analista.
- **Torne seu trabalho reprodutível:** Resultados replicáveis são fundamentais para que outros pesquisadores/analistas possam revisitar e reprocessar seus achados. Em muitos contextos, a replicação completa é muito difícil ou impossível, como em experimentos de larga escala, como ensaios clínicos multicêntricos, porém é sempre bom perseguir esta meta. Quando possível, forneça o conjunto de dados, juntamente com uma descrição completa da análise. Com isso deve ser possível reproduzir as tabelas, figuras e inferências estatísticas. Melhore drasticamente a capacidade de reproduzir descobertas sendo muito sistemático sobre as etapas da análise, compartilhando os dados e o código usados para produzir os resultados e seguindo as práticas recomendadas de estatística aceitas.

1.4 Como não utilizar estatística?

Aqui listamos também nove pontos de atenção para evitar mal uso da estatística.

- **Não minta com estatística:** Alguns pesquisadores podem ser tentados a maquiar algum dados para seu benefício ou de outros. É sempre bom ter em mente que sua reputação e carreira podem estar em jogo ao apresentar falsos resultados. Sugerimos aqui uma leitura extra do livro do Darrell Huff[1], pois para não mentir é importante saber como se mente com estatística.
- **Resista aos mau intencionados:** Se alguém te pediu pra fazer algo estatisticamente ilegal ou aplicar uma técnica inadequada, seja resistente e questione. Nem sempre o problema é simples, então é sempre bom ter uma compreensão da situação como um todo. Aprenda a dizer não para evitar problemas futuros.

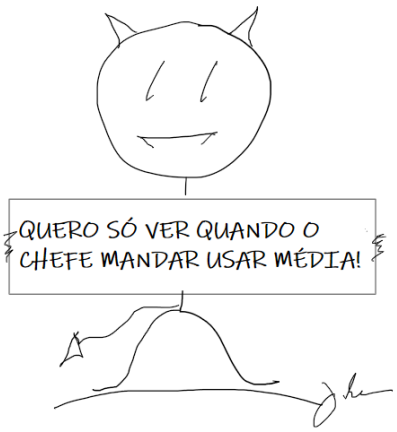


FIGURA 1.5: Use a média

- **Cuidado com as suposições:** É melhor assumir que não tem a resposta no momento e que a traz em outro momento do que inventar suposições incorretas só pra não sair por baixo em uma conversa. Ou pior, realizar um estudo/projeto apoiado por suposições incorretas sobre uma técnica estatística. Cedo ou tarde e você poderá se por em uma saia justa e ter de voltar atrás.
- **Evite ambiguidades:** A comunicação estatística precisa ser

[1] https://www.amazon.com.br/Como-Mentir-Estat%C3%ADstica-Darrell-Huff/dp/858057952X

clara. Em termos estatísticos não há espaço para meias verdades, pois são os dados falando.

- **Tenha certeza do que está falando:** Não subestime seu público. É verdade que algumas pessoas não falam ou entendem bem o *estatiquês*, mas fique alerta, pois muitos são conhecedores desta ciência, então por vias das dúvidas é melhor saber o que você vai comunicar para evitar constrangimentos.
- **Não seja complexo demais nas suas análises e comunicações:** Neste ponto seja ponderado, pois nem tudo que é simples é fácil e nem tudo que é difícil é complexo. Do planejamento à entrega é sempre bom ter seu trabalho revisado / acompanhado por outra pessoa de forma a identificar pontos de melhoria. A linguagem, sempre que possível deve ser de simples compreensão.
- **Maria vai com as outras:** Muito cuidado. Não é porque todo mundo faz algo de um jeito que você pode assumir que é certo. Censo crítico faz toda a diferença na identificação deste tipo de fenômeno.
- **Cuidado como modelos automáticos:** Modelos automatizados ou semi-automatizados podem às vezes gerar saídas inesperadas. *".. todos os modelos são errados, mas alguns são úteis. — George E. P. Box"*. Com a popularização do *machine learning* muitas pessoas tendem a pensar que se colocar os dados no computador e passar o algoritmo tudo ficará pronto. Temos casos recentes de que isso é possível. Não vamos entrar no mérito, mas como já comentamos antes, a inteligência é do analista e saídas imprevistas podem ocorrer.
- **Não acredite apenas na estatística:** Pois é. A estatística é fundamental e sem ela nada podemos fazer com os dados, mas esteja sempre atento ao negócio ou qualquer evento externo que possa influenciar seus resultados.

1.5 Estatística, *Data Science* e *Big Data*.

Desde que o conceito de dado passou a existir, a estatística se faz presente. A seguir um pequeno texto que relaciona a estatística com *Data Science* e *Big Data*.

Estatística e *Data Science* (Ciência de Dados) são partes inseparáveis. De certa forma, a estatística é um subconjunto da ciência de dados. Mas o que é *Data Science*? É difícil definir um conceito tão amplo porém, a *Wikipedia* Norte Americana coloca da seguinte forma:

> *Data science, also known as data-driven science, is an interdisciplinary field about scientific methods, processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to Knowledge Discovery in Databases (KDD).* Fonte: https://www.wikiwand.com/en/Data_science, acesso em "03/03/2019"

No texto acima podemos isolar o termo *interdisciplinary* que remete a um conjunto de muitas áreas do conhecimento. Podemos marcar desta definição aos conceitos:

- **Métodos Científicos** - Métodos estatísticos e computacionais, por exemplo;
- **Processos** - Organização de passos para atingir um objetivo;
- **Sistemas** - Sistemas informatizados, por exemplo, são ferramentas superpoderosas para realizar *Data Science*.

Tudo com objetivo de obter conhecimento com base em dados, sejam eles estruturados ou não estruturados. Expandindo mais um pouco, pode-se incluir o termo *Big Data* que envolve dados estruturados ou não estruturados em grandes volumes. O conceito de

Big Data também é complexo e discutível. (De Mauro et al., 2016) propõem uma definição deste termo como segue.

> *Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value.*

Deixamos ao leitor interessado consultar (De Mauro et al., 2016) para mais definições sobre *Big Data*.

Podemos isolar nesta definição os termos:

- **Volume Alto** - O termos alto pode ser relativo e depende do tipo de computador que está suportando a análise. Nós entendemos como volume alto qualquer base de dados que não pode ser processada por um sistema de tabulação como *LibreOffice calc* e seu concorrente da *Microsoft*;
- **Velocidade** - Grandes bases exigem processamento rápido e algorítimos rápidos muitas vezes são mais relevantes que um bom *Hardware*;
- **Variedade** - Bancos de dados de fontes diversas podem ser relacionados para obter conhecimento.

Estes termos chave estão direcionados diretamente com tecnologia e métodos estatísticos para transformar dados o obter valor ou seja, conhecimento. Aqui percebemos que *Data Science*, *Big Data* possuem muito em comum. Ambas utilizam dados como combustível visando descobrir novos conhecimentos. Ou seja, tudo isso precisa de estatística para fazer sentido.

Com base nos conceitos até aqui, poderíamos sugerir nossa própria definição de *Data Science*.

"Data Science é uma ciência ampla que une métodos científicos

> como a estatística, processos e sistemas tecnológicos buscando, através da análise de dados, sejam eles simples ou *Big Data* estruturados ou não, obter conhecimento acerca de fenômenos e processos variados".

1.6 Conceitos e definições

A partir de agora iniciamos os estudos de estatística e para melhor compreender alguns exemplos, vamos trabalhar com o conjunto de dados da IES do Censo da Educação Superior no Brasil de 2017.

```
# Carregando funções extras do rnp
require(rnp, quietly = TRUE)

# Classes para base de dados das IES
dicionario <- "Dados/INEP/ANEXOS/ANEXO I - Dicionario de Dados e Tabelas Auxiliares/Dicionario_de_Dados.xlsx"
classes <- rnp::rnp_get_classes_inep(caminho = dicionario,
                                     aba = "DM_IES",
                                     retorna_lista = FALSE)

# Base de dados
base <- rnp::rnp_read(base = paste0("Dados/INEP/DADOS/DM_IES.CSV"),
                      sep = "|",
                      dec = ".",
                      header = TRUE,
                      encoding = "Latin-1",
                      verbose = FALSE,
                      showProgress = FALSE)
# Aplica classes
base_ies <- rnp::rnp_aplica_classes(base = base, classes = classes) %>%
  dplyr::mutate(Sigla = SG_IES,
                TotalTecnicos = QT_TEC_TOTAL,
                ReceitaPropria = VL_RECEITA_PROPRIA,
```

```
        DespesaPesquisa = VL_DESPESA_PESQUISA)

base_ies %>%
 dplyr::select(Sigla, TotalTecnicos, ReceitaPropria,DespesaPesquisa) %>%
 head(n = 10) %>%
 knitr::kable(digits = 2, align = "llrrr",
       booktabs = TRUE, format = tb_formata,
       caption = "Dez primeiras observações da base de IES censo 2017") %>%
 kableExtra::kable_styling(latex_options = "hold_position")
```

TABELA 1.1: Dez primeiras observações da base de IES censo 2017

Sigla	TotalTecnicos	ReceitaPropria	DespesaPesquisa
UFMT	1574	6913132	5807924
UNB	3206	63239902	5772538
UFS	1429	3161937	3033336
UFAM	1721	4136205	1490991
UFOP	786	3458763	748715
PUCPR	1275	616991005	11948066
UNICAP	441	142785305	560400
UCS	986	338035351	3380182
UNISINOS	1062	487902230	6687713
UCPEL	246	100089191	3540465

Iniciamos aqui a revisão de alguns dos conceitos utilizados na linguagem estatística de forma simplificada. Isso nos ajudará a falar os termos e alguns jargões da linguagem estatística para enriquecer o diálogo.

1.6.1 População, amostra, censo

- **População:** É o conjunto de todos os elementos (pessoas, animais, plantas, etc.) que possuam alguma característica de interesse.

- **Amostra:** É um pedaço ou subconjunto da população e, a partir dela, faz-se inferência sobre as características da população. A ideia de amostra significativa vem do fato de que para representar a população, a amostra precisa preservar suas características.

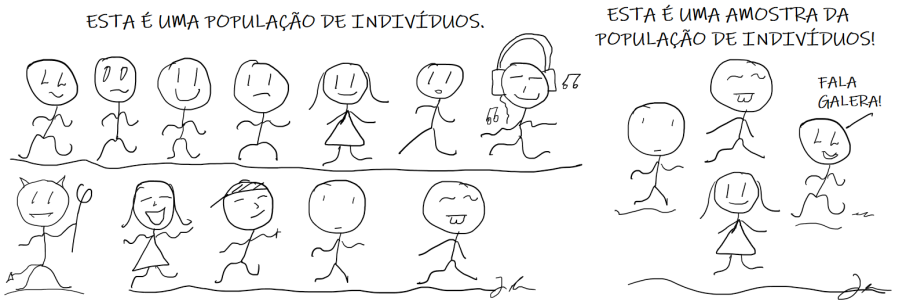


FIGURA 1.6: População e amostra

- **Censo:** É o processo utilizado para coletar dados abordando todos os elementos de uma população. Neste caso não há necessidade de amostras, pois toda a população é considerada.
- **Parâmetro:** É qualquer medida numérica que serve para descrever uma **característica de uma população**. São geralmente representados por letras gregas. Por exemplo: média (μ), variância (σ^2) e desvio padrão (σ).
- **Estatística:** Tem a mesma função do parâmetro, porém é medido na amostra. São geralmente representados por letras latinas com acentos. Por exemplo: média amostral ($\bar{X}$), variância (S^2) e desvio padrão (S) amostrais.

1.6.2 Dados, informação e conhecimento

- **Dados:** Dados são resultados de medições ou qualquer fonte relato documentado. Por si só dados não tem significado concreto porém, a disponibilidade dos mesmos é matéria prima para a obtenção de informações. Dados podem ser obtidos pela percepção através dos sentidos, pela execução de um processo de medição, por sensores, pesquisas, censos, etc.

- **Informação:** O processamento e análise dos dados gera informação que auxilia a tomada de decisão seja nos negócios, pela ciência e na vida cotidiana em geral.
- **Conhecimento:** O conhecimento extrapola a informação, ele produz ideias e experiências. Através do conhecimento é possível a abstração e a evolução de ideias e conceitos novos com base em todo tipo de experiência, sejam elas oriundas das informações ou vivências.

1.6.3 Variáveis

Variáveis são características medidas em cada elemento da amostra ou população. Elas podem ter valores numéricos ou não numéricos e seus valores podem varia de elemento para elemento.

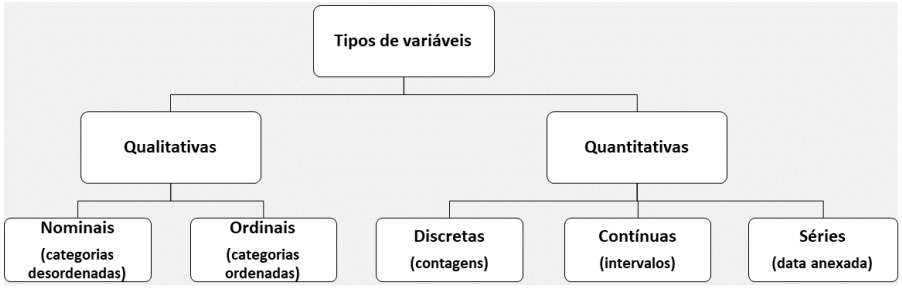


FIGURA 1.7: Tipos de variáveis

A figura 1.7 trás um resumo dos tipos mais comuns de variáveis. Detalhamos um pouco mais a seguir.

- **Variáveis numéricas:** As variáveis numéricas ou quantitativas se classificam em discretas e contínuas. As discretas persentam contagens e as contínuas, medidas. Na tabela 1.1, a coluna TotalTecnicos que representa o total de técnicos das IES é um bom exemplo de variável quantitativa discreta. Já as colunas ReceitaPropria e DespesaPesquisa são variáveis contínuas, pois são medidas que representam valores de receita.
- **Séries temporais:** As variáveis quantitativas quando indexadas de uma variável de tempo, como por exemplo hora, dia,

mês, ano, etc. são classificadas como séries temporais. Séries temporais possuem um grande valor e por isso, a estatística reserva um campo completo de estudos para este tipo de dado.

- **Variáveis categóricas:** As variáveis categóricas ou qualitativas descrevem caraterísticas dos indivíduos da amostra ou população. Elas podem ser nominais, quando descrevem características arbitrárias sem efeito de ordenação ou ordinais, quando descrevem relações de ordenação. A coluna Sigla da tabela 1.1 representa uma variável qualitativa nominal. Exemplos de ordinais são classe social e escolaridade

1.6.4 Análise univariada, bivariada e multivariada

- **A análise univariada:** Busca-se descrever a população ou amostra analisando cada variável de forma isolada. É a maneira mais simples de obter informação e de fazer a estimativa estatística. Exemplos de análises univariadas são as médias, medianas e quatis.
- **A análise bivariada:** Analisa relações existentes entre pares de variáveis para fins de explicação e/ou previsão. Na análise bivariada, a formulação de uma hipótese precisar ser feita e a estatística permitirá inferir ou confirmar esta hipótese. Análise de correção entre duas variáveis e tabelas de frequência de dupla entrada são exemplos de análise bivariada
- **A análise multivariada:** Na análise multivariada a estatística dispões de uma série de técnicas que analisam de forma conjunta as múltiplas relações das variáveis. Regressão múltipla, análise de cluster e fatorial são exemplos de análises multivariadas.

1.6.5 *tidydata*

Em seu artigo, (Wickham et al., 2014) discutem e mostram uma forma repensada de organizar tabelas estruturadas para análise de dados. De forma simples, um significado para *tidy data* é: dados arrumados, organizados. Um conjunto de dados para ser tidy precisa ter três ingredientes:

- Cada observação é uma linha;
- Cada variável é uma coluna;

- Cada valor está em uma célula (linha x coluna);

Dados neste formato ajudam a tornar a análise mais rápida, principalmente com as ferramentas do tidyverso (Wickham, 2017). A Tabela 1.1 mostra uma configuração de dados tidy, onde cada linha representa uma observação ou indivíduo que no caso é uma IES; cada coluna representa uma variável ou característica, por exemplo Sigla ou ReceitaPropria da IES e o cruzamento entre linhas e colunas são os valores correspondentes.

Observação: Nas empresas o conceito de tidy data geralmente é atribuído ao que chamam de ABT (Analytics Base Table). Este é o estado da arte que nem sempre é fácil de se chagar, mas uma vez lá, os dados estarão prontos para todo tipo de análise de modelagem.

```
# Exemplos de tidydasets
knitr::kable(
 list(head(mtcars[,1:4], 10), head(dplyr::starwars[,1:4], 10)),
 booktabs = TRUE, format = tb_formata,
 caption = "Amostra de outras tabelas de dados organizados") %>%
 kableExtra::kable_styling(latex_options = "hold_position")
```

TABELA 1.2: Amostra de outras tabelas de dados organizados

	mpg	cyl	disp	hp
Mazda RX4	21.0	6	160.0	110
Mazda RX4 Wag	21.0	6	160.0	110
Datsun 710	22.8	4	108.0	93
Hornet 4 Drive	21.4	6	258.0	110
Hornet Sportabout	18.7	8	360.0	175
Valiant	18.1	6	225.0	105
Duster 360	14.3	8	360.0	245
Merc 240D	24.4	4	146.7	62
Merc 230	22.8	4	140.8	95
Merc 280	19.2	6	167.6	123

name	height	mass	hair_color
Luke Skywalker	172	77	blond
C-3PO	167	75	NA
R2-D2	96	32	NA
Darth Vader	202	136	none
Leia Organa	150	49	brown
Owen Lars	178	120	brown, grey
Beru Whitesun lars	165	75	brown
R5-D4	97	32	NA
Biggs Darklighter	183	84	black
Obi-Wan Kenobi	182	77	auburn, white

.

1.6.6 Exercícios resolvidos

Com apoio da base de IES, cuja amostra foi mostrada na tabela 1.1 vamos exercitar os conceitos vistos até agora.

- **População e amostra**

Exercício 1.1. Quem é a população de IES?

Solução. Neste caso, a população se confunde com um censo, pois a base das IES faz parte do levantamento censitário de 2017 e contempla, a princípio todas as IES do Brasil

```
paste(nrow(base_ies), "Universidades.")
```

```
## [1] "2448 Universidades."
```

Exercício 1.2. O que poderia ser uma amostra da base de IES?

Solução. Uma amostra poderia ser todas as IES grandes com mais de 2000 técnicos.

```
paste("A base tem", base_ies %>% filter(TotalTecnicos > 2000) %>% nrow(), "IES com mais de 2000 ténicos")
```

```
## [1] "A base tem 29 IES com mais de 2000 ténicos"
```

Exercício 1.3. Defina um parâmetro desta base.

Solução. Neste caso poderíamos definir $\mu_{receita}$ como a receita própria média das IES brasileiras.

Exercício 1.4. Qual seria uma estatística?

Solução. A receita média média da amostra de SP pode ser definida uma estatística dada por $\bar{X}_{receita}$

- **Dados, informação e conhecimento**

Exercício 1.5. A tabela de IES possui muitos registros numéricos e também textuais. Ela representa melhor, dados, informações ou conhecimento?

Solução. Valores brutos em uma tabela são exemplos de dados que por sí só não são informação. Assim a tabela representa um conjunto de **dados**

Exercício 1.6. A frase, 'O estado de São Paulo possui o maior número de IES do Brasil' é um exemplo de dado, informação ou conhecimento?

Solução. Estas informações podem ser obtidas dos conjuntos de dados das IES através de uma contagem de toda as IES do estado de São Paulo

Exercício 1.7. "Um estado que possui muitas universidade possui indicadores de educação superiores." Esta frase é um exemplos de...

Solução. Conhecimento abstraído do fato de que muitas universidades em um estado podem representar alto índice de formação de seu povo.

- **Variáveis e tipos de análises**

Exercício 1.8. No conjunto de IES temos quantas variáveis e quantas observações?

Solução. A base IES possui 2448 linhas (observações) e 4 colunas (variáveis). Além disso, são uma variável caractere e três numéricas conforme resultados abaixo, obtidos com a função extra rnp_atributos().

```
# Listando os atributos da base de ies para as 10 primeiras colunas
rnp::rnp_atributos(base_ies) %>%
 head(n = 10) %>%
 knitr::kable(booktabs = TRUE, format = tb_formata) %>%
 kableExtra::kable_styling(latex_options = "hold_position")
```

Exercício 1.9. A mediana da receita própria anual das IES brasileiras é R$ 7.429.221,00. Que tipo de análise é esta?

Solução. Olhando apenas para uma variável, temos uma análise univariada.

classeBase	comprimento	variaveis	classeVars
tbl_df	2448 linhas e 62 colunas	NU_ANO_CENSO	integer
tbl_df	2448 linhas e 62 colunas	CO_IES	integer
tbl_df	2448 linhas e 62 colunas	NO_IES	character
tbl_df	2448 linhas e 62 colunas	SG_IES	character
tbl_df	2448 linhas e 62 colunas	CO_MANTENEDORA	integer
tbl_df	2448 linhas e 62 colunas	NO_MANTENEDORA	character
tbl_df	2448 linhas e 62 colunas	CO_REGIAO	integer
tbl_df	2448 linhas e 62 colunas	CO_UF	integer
tbl_df	2448 linhas e 62 colunas	CO_MUNICIPIO	integer
tbl_df	2448 linhas e 62 colunas	QT_TEC_TOTAL	integer

Exercício 1.10. A correlação entre o total de técnicos e a receita própria é muito baixa (0.0577). Que tipo de análise é esta?

Solução. Neste caso relaciona-se duas variáveis, então temos uma análise bivariada.

Exercício 1.11. Conjuntamente, as variáveis Receita propria e Depesa com pesquisa não explicam o Total de tecnicos. Que tipo de análise é esta?

Solução. Como temos o relacionamento de três variáveis, neste caso temos uma análise multivariada.

.

2

Estatística descritiva

Faz parte da Estatística descritiva apresentar técnicas para descrever e sumarizar conjuntos de dados dados de natureza diversa. Entre estas técnicas estão tabelas de frequências, resumos numéricos e gráficos de acordo com o tipo de variável envolvida. Neste capítulo trataremos destas técnicas sempre com foco em conjuntos de dados do Censo de Educação Superior do INEP.

2.1 Variáveis categóricas

No ramo esquerdo da figura 1.7 temos as variáveis qualitativas. Elas são em geral, variáveis em formato de texto ou números inteiros que representam atributos nominais ou ordinais de determinada observação ou indivíduo de uma base de dados. A seguir vamos apresentar algumas técnicas descritivas para este tipo de variável e para melhor exemplificar vamos trabalhar com mais algumas variáveis do conjunto de dados dos docentes do ensino superior do censo do INEP de 2017. Para mais informações sobre as variáveis consulte o dicionário de dados da base. As melhores técnicas utilizadas são para variáveis categóricas são **contagens/frequências, proporções/percentuais e gráficos**.

```
# Classes dos dados
classes <- rnp::rnp_get_classes_inep(caminho = dicionario,
                                     aba = "DM_DOCENTE",
                                     retorna_lista = FALSE)
# Base de dados
```

```
base <- rnp::rnp_read(base = paste0("Dados/INEP/DADOS/DM_DOCENTE.CSV"),
                      sep = "|",
                      dec = ".",
                      header = TRUE,
                      encoding = "Latin-1",
                      verbose = FALSE,
                      showProgress = FALSE)
# Aplica classes
base_docentes <- rnp::rnp_aplica_classes(base = base,
                                         classes = classes)

# Variáveis importantes da base de Docentes
vars_doc <- c("CO_IES", "DESC_TP_CATEGORIA_ADMINISTRATIVA", "DESC_TP_SEXO",
              "NU_IDADE", "DESC_TP_ESCOLARIDADE","DESC_TP_REGIME_TRABALHO",
              "CO_DOCENTE","DESC_TP_SITUACAO")

base_docentes <- base_docentes %>%
 dplyr::select(vars_doc) %>%
 dplyr::mutate(faixaIdade = if_else(NU_IDADE <= 30, "01.Até 30 anos",
                       if_else(NU_IDADE <= 40, "02.Entre 30 e 40 anos",
                       if_else(NU_IDADE <= 50, "03.Entre 40 e 50 anos",
                       if_else(NU_IDADE <= 60, "04.Entre 50 e 60 anos",
                       "05.Acima de 60 anos"))))
               ) %>%
 dplyr::rename(cdDocente = CO_DOCENTE,
               cdIES = CO_IES,
               catAdm = DESC_TP_CATEGORIA_ADMINISTRATIVA,
               situacao = DESC_TP_SITUACAO,
               escolaridade = DESC_TP_ESCOLARIDADE,
               faixaIdade = faixaIdade,
               regimeTrabalho = DESC_TP_REGIME_TRABALHO,
               sexo = DESC_TP_SEXO,
               idade = NU_IDADE) %>%
 dplyr::arrange(faixaIdade)

rnp::rnp_atributos(base_docentes) %>%
```

```
knitr::kable(booktabs = TRUE, format = tb_formata,
        caption = "Atributos da base dos docentes") %>%
kableExtra::kable_styling(latex_options = "hold_position")
```

TABELA 2.1: Atributos da base dos docentes

classeBase	comprimento	variaveis	classeVars
tbl_df	392036 linhas e 9 colunas	cdIES	integer
tbl_df	392036 linhas e 9 colunas	catAdm	character
tbl_df	392036 linhas e 9 colunas	sexo	character
tbl_df	392036 linhas e 9 colunas	idade	integer
tbl_df	392036 linhas e 9 colunas	escolaridade	character
tbl_df	392036 linhas e 9 colunas	regimeTrabalho	character
tbl_df	392036 linhas e 9 colunas	cdDocente	integer64
tbl_df	392036 linhas e 9 colunas	situacao	character
tbl_df	392036 linhas e 9 colunas	faixaIdade	character

A tabela 2.1 mostra uma parte dos dados dos decentes contendo algumas variáveis que exploraremos mais adiante. Temos 392036 observações ou docentes de ensino superior no censo de 2017.

2.1.1 Tabelas de frequências

As tabelas de frequência são muito úteis para analisar a distribuição dos dados de uma variável segundo suas categorias ou classes, com elas podemos analisar contagens e proporções de cada categoria da variável. Para isso, entra em cena os seguintes conceitos:

Classe: É a descrição da categoria ou nível da variável;

Frequência absoluta (f_a): trata-se da contagem de observações pertencentes a uma dada categoria da variável;

Frequência relativa (f_r): trata-se da contagem de observações pertencentes a uma dada categoria da variável dividida pelo total N de observações. É a representação percentual da f_a;

Frequência absoluta acumulada (F_{aa}): é dada pelo soma acumulada das f_a;

Frequência relativa acumulada (F_{ra}): é dada pelo soma acumulada das f_r;

OBS.: Adotamos letras maiúsculas para definir frequências acumuladas e minúscula para simples.

A junção destas estatísticas constitui uma tabela de frequências relativas que podem ser:

2.1.1.1 Tabela de frequências simples

São tabelas simples para apenas uma variável. Construímos uma função rnp_freq() para realizar a tabulação dos dados. Esta função e muitas outras podem ser encontradas no script de apoio.

```
rnp::rnp_freq(x = base_docentes$escolaridade,
        sortd = FALSE, digits = 3) %>%
knitr::kable(booktabs = TRUE, format = tb_formata,
        caption = "Frequências simples para escolaridade do docente") %>%
kableExtra::kable_styling(latex_options = "hold_position")
```

TABELA 2.2: Frequências simples para escolaridade do docente

classe	fa	fr	Faa	Fra
1. Sem graduação	10	0.000	10	0.000
2. Graduação	4613	0.012	4623	0.012
3. Especialização	72301	0.184	76924	0.196
4. Mestrado	154285	0.394	231209	0.590
5. Doutorado	160827	0.410	392036	1.000

A tabela 2.2 exemplifica uma tabela de frequência simples onde podemos analisar diretamente os dados da variável escolaridade dos docentes de ensino superior no Brasil, segundo os dados do censo do INEP de 2017. Podemos ver que há 397.611 docentes e que destes, 39,58% possuem mestrado e 38,48% doutorado. Juntas,

estas duas categorias representam 78,07% da base. Se somarmos os especialistas, temos um total de 98,52%.

2.1.1.2 Tabela de frequências de dupla entrada

Quando desejamos analisar a relação entre duas variáveis categóricas, podemos aplicar a mesma ideia da tabela de frequências simples, porém o resultado para cada estatística fica um pouco trabalhosa e em geral, é melhor analisar uma das estatísticas $f_a e f_r$. Como a relação é bivariada, as estatísticas para F_{aa}, F_{ra} não são tão simples. Neste caso, o ideal é analisar cada variável separadamente gerando tabelas simples.

```
rnp::rnp_2freq(x = base_docentes$escolaridade,
        y = base_docentes$sexo,
        digits = 3, percents = TRUE) %>%
 knitr::kable(booktabs = TRUE, format = tb_formata,
       caption = "Frequências cruzadas da escolaridade por sexo do docente") %>%
kableExtra::kable_styling(latex_options = "hold_position")
```

A tabela 2.3 exemplifica a utilização da tabela de frequências gerada para as variáveis escolaridade e sexo do docente. A primeira coluna da tabela representa o tipo de estatística e a visão que ela foi calculada. Por exemplo, a frequência absoluta de doutores em relação ao total da base é $f_{a(total)} = 160.827$ que equivale à relativa $f_{r(total)} = \frac{160.827}{392.036} = 0.410$.

Exercício 2.1. Qual a frequência relativa docentes doutores do sexo feminino?

Solução. Se olharmos a frequência dos doutores de sexo feminino, temos que $f_{r(doutores/feminino)} = \frac{73.812}{160.827} = 0.459$ que equivale à $f_{r(linha)} = 0.459$.

Exercício 2.2. Qual a frequência relativa docentes feminino que são doutores?

Solução. Seguindo a mesma lógica, mas agora olhando para a coluna de sexo, temos que $f_{r(feminino/doutor)} = \frac{73.812}{179.856} = 0.4104$.

TABELA 2.3: Frequências cruzadas da escolaridade por sexo do docente

Tipo	Classe X/Y	1. Feminino	2. Masculino	Total
fa	1. Sem graduação	3.000	7.000	10.000
fr	1. Sem graduação	0.000	0.000	0.000
fr_col	1. Sem graduação	0.000	0.000	0.000
fr_lin	1. Sem graduação	0.300	0.700	1.000
fa	2. Graduação	1848.000	2765.000	4613.000
fr	2. Graduação	0.005	0.007	0.012
fr_col	2. Graduação	0.010	0.013	0.012
fr_lin	2. Graduação	0.401	0.599	1.000
fa	3. Especialização	30579.000	41722.000	72301.000
fr	3. Especialização	0.078	0.106	0.184
fr_col	3. Especialização	0.170	0.197	0.184
fr_lin	3. Especialização	0.423	0.577	1.000
fa	4. Mestrado	73614.000	80671.000	154285.000
fr	4. Mestrado	0.188	0.206	0.394
fr_col	4. Mestrado	0.409	0.380	0.394
fr_lin	4. Mestrado	0.477	0.523	1.000
fa	5. Doutorado	73812.000	87015.000	160827.000
fr	5. Doutorado	0.188	0.222	0.410
fr_col	5. Doutorado	0.410	0.410	0.410
fr_lin	5. Doutorado	0.459	0.541	1.000
fa	Total	179856.000	212180.000	392036.000
fr	Total	0.459	0.541	1.000

Este tipo de análise também é chamada de análise marginal, pois estamos olhando as margens. Sempre que olhamos as margens estamos na verdade tomando uma das variáveis como referência e fazendo verificações sobre ela em relação à outra. É normal acontecer alguma confusão em relação às frequências. Neste caso, recomendamos analisar cada variável separadamente ou fazer as con-

tas olhando para a tabela de frequências absolutas de dupla entrada.

2.1.2 Gráficos para uma variável categórica

Além das tabelas de frequência simples também é possível complementar a análise de variáveis categóricas através de gráficos. Os mais conhecidos são os gráficos de **setores (ou pizza)** e os de **barras**. Através destes gráficos a informação fica facilmente visível e a obtenção de informações valiosas fica evidente.

2.1.2.1 Setores

```
tb <- rnp::rnp_freq(base_docentes$faixaIdade, sortd = FALSE)
p <- ggplot2::ggplot(tb, aes("", fr, fill = classe))
p + ggplot2::geom_bar(width = 1, size = 1, color = "white", stat = "identity") +
 ggplot2::coord_polar("y") +
 ggplot2::geom_text(aes(label = paste0(round(100*fr, 1), "%")),
            position = position_stack(vjust = 0.5), size=3) +
 ggplot2::labs(x = NULL, y = NULL, fill = NULL, title = "") +
 ggplot2::guides(fill = guide_legend(reverse = TRUE)) +
 ggplot2::theme_gray() +
 ggplot2::theme(axis.line = element_blank(),
          axis.text = element_blank(),
          axis.ticks = element_blank(),
          legend.position="right", legend.text = element_text("Classe"))
```

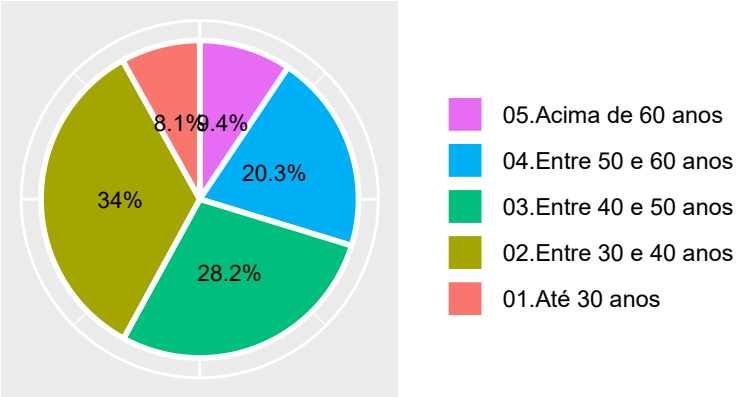


FIGURA 2.1: Gráfico de setores para faixa de idade dos docentes

Como podemos notar, a figura 2.1 é muito intuitiva para representar visualmente a distribuição de frequências das classes de uma variável categórica. Combinado com as cores de cada fatia, fica claro e objetivo a parcela de cada categoria para explicar o todo que por sua vez representa 100%.

2.1.2.2 Barras

Gráficos de barras também são intuitivos e geralmente são preferíveis em relação aos gráficos de setores. Isso ocorre porque o olho humano é mais sensível a linhas do que círculos e formas em 3D e prefere analisar figuras mais limpas. Para expandir seus conhecimentos sobre análise visual, sugerimos a (Tufte and Graves-Morris, 2014).

```
p <- ggplot2::ggplot(tb, aes(classe, fr, fill = classe))
p + ggplot2::geom_bar(width = 1, size = 1, color = "white",
                      stat = "identity", show.legend = FALSE) +
 ggplot2::geom_text(aes(label = paste0(round(100*fr, 1), "%")),
                    position = position_stack(vjust = 1), size=3) +
 ggplot2::labs(x = NULL, y = NULL, fill = NULL, title = "") +
```

```
ggplot2::theme_gray() + coord_flip() +
ggplot2::theme(axis.line = element_blank())
```

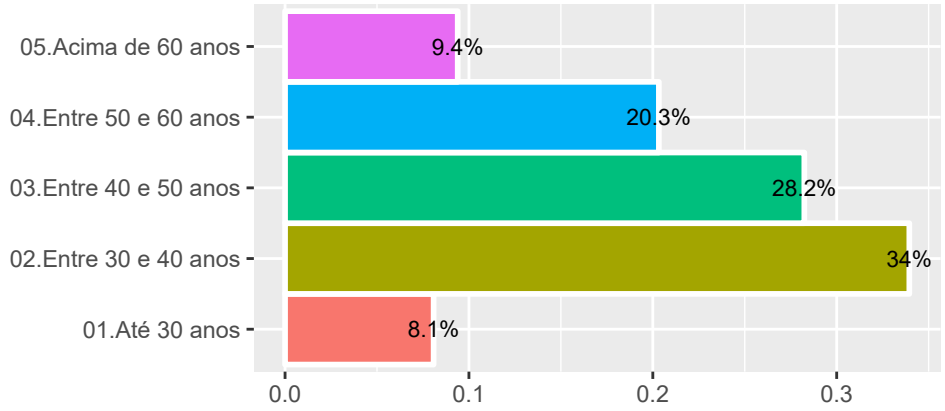


FIGURA 2.2: Gráfico de barras para faixa de idade dos docentes

A figura 2.2 mostra um gráfico de barras para as faixas de idade dos docentes, nele pode-se notar que visualmente as diferenças de patemares ficam evidentes mesmo sem ter a informação em percentual em cada barra. Com isso a leitura fica mais direta e é possível comparar todas as barras simultaneamente. O eixo y contém as proporções e o eixo x a descrição de cada categoria.

2.1.3 Gráfico para duas variáveis categóricas

Vimos nas tabelas de frequência que as tabelas de dupla entrada são boas ferramentas para analisar conjuntamente a relação entre duas variáveis, mas isso também pode ser feito de forma visual.

2.1.3.1 *mosaicplot*

Podemos visualizar a relação entre duas variáveis categóricas ou numéricas de poucas classes, através do **Gráfico de mosaico (*mosaic plot*)**. O pacote ggmosaic expande o ggplot2 para produzir este tipo de gráfico.

```
p <- ggplot2::ggplot(base_docentes)
p + ggplot2::theme_gray() +
 ggmosaic::geom_mosaic(aes(x = product(sexo), fill = faixaIdade),
                       show.legend = FALSE) +
 ggplot2::labs(x = NULL, y = NULL, fill = NULL, title = "") +
 ggplot2::theme(axis.line = element_blank())
```

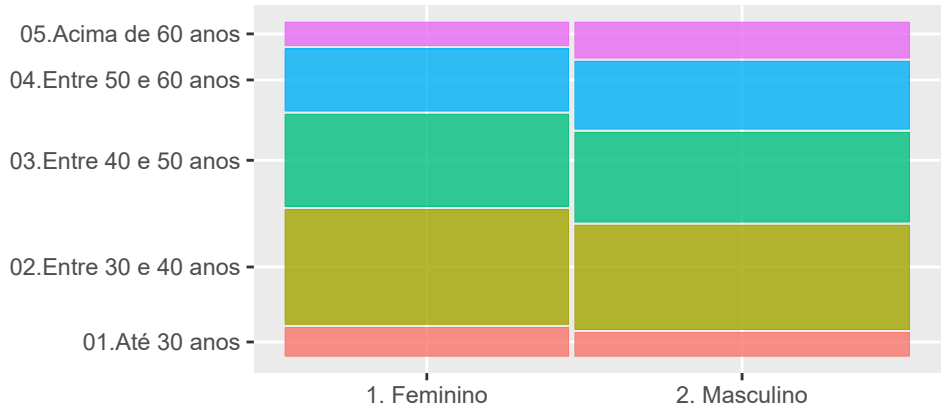


FIGURA 2.3: Gráfico de mosaico para faixa de idade dos docentes por sexo

Cada coluna da figura 2.3 representa uma classe da variável sexo do docente e cada linha representa a uma classe da variável faixa de idade. Note que o cruzamento dente linhas e colunas geram retângulos que estimam a frequência de de cada cruzamento, sendo maior nos casos em que existem mais dados. Por exemplo, proporcionalmente docentes do sexo masculino acima de 60 anos são maioria nesta faixa de idade. Aliás, gráfico mostra que a proporção de docentes do sexo masculino é maior do que o feminino em todas as faixas de idade.

2.1.4 Exercícios

Para resolver os exercícios desta seção, utilize o conjunto de dados DM_CURSO.CSV presente na pasta de dados ou pacote ?rnp::dm_curso. Esta base de dados possui informações sobre os cursos das IES no censo de 2017 do INEP. Mais informações sobre as variáveis podem ser obtidas no dicionario de dados presente na pasta AJUDA/ANEXOS ou no site do INEP.

```
classes <- rnp::rnp_get_classes_inep(caminho = dicionario,
                        aba = "DM_CURSO",
                        retorna_lista = FALSE)
# Base de dados
base <- rnp::rnp_read(base = "Dados/INEP/DADOS/DM_CURSO.CSV",
              sep = "|",
              dec = ".",
              header = TRUE,
              encoding = "Latin-1",
              verbose = FALSE,
              showProgress = FALSE)
# Aplica classes
base_curso <- rnp::rnp_aplica_classes(base = base, classes = classes)
# Verifica algumas propriedades das 10 primeiras colunas.
dplyr::glimpse(base_curso[,1:10])
```

```
## Observations: 35,693
## Variables: 10
## $ NU_ANO_CENSO         <int> 2017, 2017, 2017, ...
## $ CO_IES               <int> 789, 4567, 2341, 6...
## $ CO_LOCAL_OFERTA      <int> 1033528, 659871, 1...
## $ CO_UF                <int> 14, 51, 35, 35, 53...
## $ CO_MUNICIPIO         <int> 1400100, 5107925, ...
## $ CO_CURSO             <int> 1259131, 1258115, ...
## $ NO_CURSO             <chr> "MÚSICA", "GESTÃO ...
## $ CO_OCDE_AREA_GERAL   <int> 1, 3, 3, 5, 3, 7, ...
## $ CO_OCDE_AREA_ESPECIFICA <int> 14, 34, 38, 52, 38...
## $ CO_OCDE_AREA_DETALHADA  <int> 146, 345, 380, 522...
```

Exercício 2.3. Faça uma análise de frequências da variável DESC_TP_CATEGORIA_ADMINISTRATIVA e responda qual a taxa de cursos por IES tipo pública.

Solução. Analisando a tabela de frequências abaixo temos que as três categorias que definem cursos de IES públicas representam 29,7% medidos pela frequência acumulada relativa (Fra). Aqui o resultado refere-se à soma das classes 1, 2 e 3 na ordem em que aparecem.

```
rnp::rnp_freq(base_curso$DESC_TP_CATEGORIA_ADMINISTRATIVA, digits = 3) %>%
 knitr::kable(booktabs = TRUE, format = tb_formata,
        caption = "ex: Frequências para categoria administrativa (cursos)") %>%
 kableExtra::kable_styling(latex_options = "hold_position")
```

TABELA 2.4: ex: Frequências para categoria administrativa (cursos)

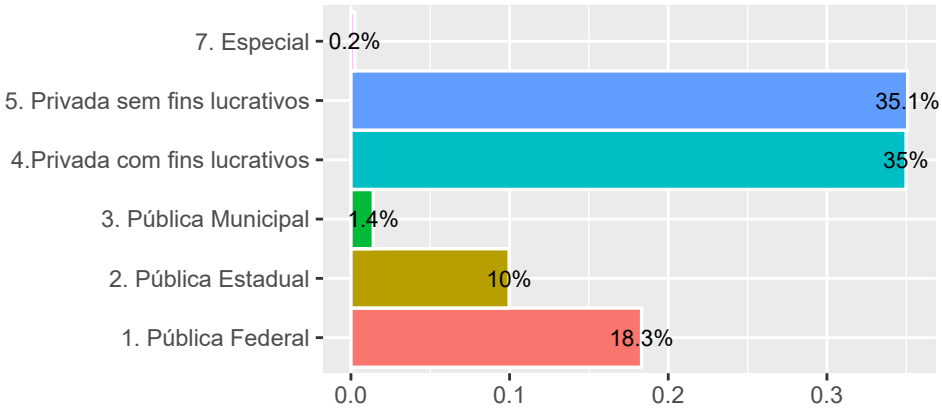
classe	fa	fr	Faa	Fra
1. Pública Federal	6538	0.183	6538	0.183
2. Pública Estadual	3558	0.100	10096	0.283
3. Pública Municipal	502	0.014	10598	0.297
4.Privada com fins lucrativos	12488	0.350	23086	0.647
5. Privada sem fins lucrativos	12523	0.351	35609	0.998
7. Especial	84	0.002	35693	1.000

Exercício 2.4. Represente graficamente a variável DESC_TP_CATEGORIA_ADMINISTRATIVA e interprete quais as categorias de maior e menor influência.

Solução. Para dados categóricos, os gráficos mais utilizados são de barras e setores. Vamos fazer um gráfico de barras com ggplot2.

```
# Primeiro fazemos as frequências, depois o gráfico
tb <- rnp::rnp_freq(base_curso$DESC_TP_CATEGORIA_ADMINISTRATIVA)

p <- ggplot2::ggplot(tb, aes(classe, fr, fill = classe))
p + ggplot2::geom_bar(width = 1, size = 0.6, color = "white",
                      stat = "identity", show.legend = FALSE) +
 ggplot2::geom_text(aes(label = paste0(round(100*fr, 1), "%")),
                    position = position_stack(vjust = 1), size=3) +
 ggplot2::labs(x = NULL, y = NULL, fill = NULL, title = "") +
 ggplot2::theme_gray() + coord_flip() +
 ggplot2::theme(axis.line = element_blank())
```



As categorias privada com e sem fins lucrativos representam 70,1% dos dados, sendo as mais representativas. Públicas municipais são minoria (1,4%) e especial apenas (0.2%). Acho que políticas públicas para aumentar a participação de cursos especiais devem ser melhoradas.

Exercício 2.5. Olhando para a variável DESC_TP_GRAU_ACADEMICO responda quais as categorias de maior e menor ocorrências de cursos. Desconsidere a categoria missing.

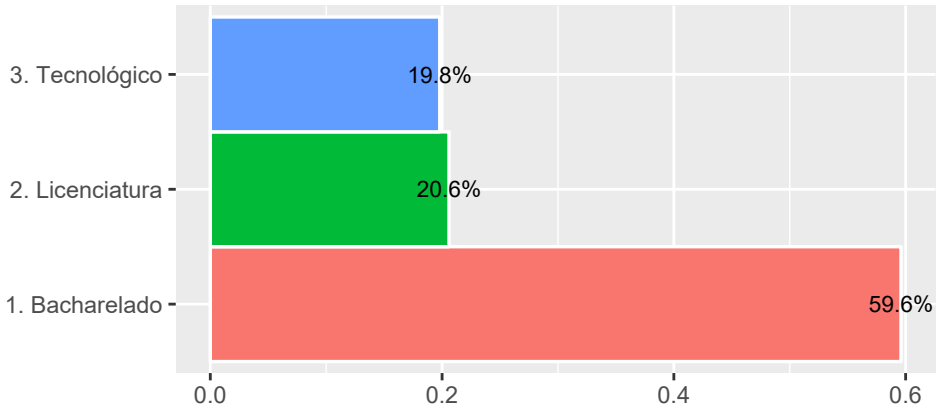
Solução. Vamos gerar uma tabela de frequências simples para

avaliar as categorias e depois eliminar os dados da categoria missing.

```
# As categorias diferentes de missing são: 1. Bacharelado, 2. Licenciatura e 3. Tecnológico
# missing representa 0.09% dos dados.
GrauAcad <- base_curso %>%
 dplyr::transmute(GrauAcad = DESC_TP_GRAU_ACADEMICO) %>%
 dplyr::filter(GrauAcad %in% c("1. Bacharelado", "2. Licenciatura", "3. Tecnológico"))

tb <- rnp::rnp_freq(GrauAcad$GrauAcad, digits = 3)

p <- ggplot2::ggplot(tb, aes(classe, fr, fill = classe))
p + ggplot2::geom_bar(width = 1, size = 0.6, color = "white",
                      stat = "identity", show.legend = FALSE) +
 ggplot2::geom_text(aes(label = paste0(round(100*fr, 1), "%")),
                    position = position_stack(vjust = 1), size=3) +
 ggplot2::labs(x = NULL, y = NULL, fill = NULL, title = "") +
 ggplot2::theme_gray() + coord_flip() +
 ggplot2::theme(axis.line = element_blank())
```



Como o gráfico mostra, 59,6% dos cursos são de bacharelado frente a 20,6% de licenciatura. Estamos formando poucos professores?

Exercício 2.6. Represnete a variável DESC_TP_GRAU_ACADEMICO como um gráfico de barras e interprete os reusltados para cursos tecnologicos em relação aos cursos de licneciatura.

Exercício 2.7. A variável DESC_TP_ORGANIZACAO_ACADEMICA descreve características do tipo de organização, estude esta variável graficamente.

Exercício 2.8. Analise conjuntamente as variáveis DESC_IN_POSSUI_LABORATORIO e DESC_TP_CATEGORIA_ADMINISTRATIVA para verificar a existência de laboratórios nos cursos públicos e privados.

2.2 Variáveis numéricas

Variáveis numéricas são se longe o tipo mais comum e analisável de dados, pois contemplam medidas de todo tipo de processo. Por exemplo, idade de uma pessoa em dias, peso, altura, total de pessoas em um metrô, em uma fila de cinema e por aí vai. Por permitir cálculos matemáticos, este tipo de variável tem sido estudado há milênios e portanto, boa parte das técnicas estatísticas atuais de baseiam em dados numéricos. Neste tópico abordaremos as principais medidas estatísticas cobrindo centralidade, dispersão entre outras e os principais gráficos que podem ser empregados.

2.2.1 Medidas estatísticas de centraliade

Quando olhamos um conjunto de dados de uma variável numérica logo pensamos em alguma forma de resumir estes dados para gerar algum tipo de informação. As medidas estatísticas de centralidade representam resumos numéricos que apontam para o centro do conjunto de dados. A figura 1.7 ilustra bem a ideia e também indica uma vulnerabilidade da **média** que são os extremos. Pontos extremos podem inserir viés no valor e na interpretação de uma média, mas para complementar a média temos a **mediana**, **moda** e **quartis** que veremos nos tópicos a seguir.

FIGURA 2.4: Na média, tudo bem

As medidas estatísticas descritas nesta seção são aplicadas a conjuntos de dados (amostras e populações) e também a distribuições de probabilidade. Para manter o tero prático deste texto, vamos focar em dados. No capitulo sobre probabilidade retornaremos o assunto no contexto das distribuições de probabilidade.

2.2.1.1 Média ($\bar{X}, \mu$)

Existem muitos tipos de média como por exemplo aritmética, ponderada, geométrica e a harmônica. Em qualquer uma delas, o intuito é resumir a centralidade dos dados em relação a seus extremos dando mais ou menos peso para cada observação. Representamos a média de uma amostra pela letra latina $\bar{X}$ (xis-barra) e a média populacional pela letra grega μ e calculamos com a mesma expressão matemática.

Veremos no capítulo sobre teoria das probabilidades que a média de uma veriável aleatória X é chamada de **valor esperado ou esperança** ($E[X]$).

Média aritmética: é a soma de todos os valores e dividido pelo total deles. Ou seja, o resultado dessa divisão equivale a um valor médio entre todos os valores e é calculada por:

$$\bar{X} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}, i = 1, 2, 3, \ldots, n$$

onde $x_1, x_2, x_3, \ldots, x_n$ representam cada valor correspondente a um elemento i da amostra e n o total de elementos.

> Este tipo de média é aplicado preferecialmente quando cada elemento tem peso igual a uma unidade, ou seja, quando não houver muita repetição.

Exemplo 2.1. Com base no conjunto de dados de IES, vamos calcular algumas médias.

```
base_ies %>%
 dplyr::summarise(`Total técnicos` = mean(TotalTecnicos),
           `Receita própria` = mean(ReceitaPropria),
           `Depesa pesquisa` = mean(DespesaPesquisa)) %>%
 knitr::kable(caption = "Média anual para dados de IES", format = tb_formata) %>%
 kableExtra::kable_styling(latex_options = "hold_position")
```

TABELA 2.5: Média anual para dados de IES

Total técnicos	Receita própria	Depesa pesquisa
168.1	145378180	886404

A tabela 2.5 mostra que a média da receita própria anual das IES brasileiras, segundo o censo de 2017, foi de R$ 143.468.742.

Este valor parece um pouco suspeito, pois representa 85,76% do PIB (Produto Interno Bruto) do estado de pernambuco tendo 2018 como ano base, segundo dados do IBGE. Mas vamos entender adiante como investigar isso melhor.

Média aritmética ponderada: Neste caso, é assumido que cada elemento amostral tem um peso, então a média aritmética ponderada é calculada multiplicando cada valor do conjunto de dados pelo seu peso, somando tudo e dividindo pela soma de todos os pesos. Na verdade a média aritmética simples é um caso especial da ponderada quando cada peso vale 1. A média ponderada é dada por:

$$\bar{X}_p = \frac{p_1 \times x_1 + p_2 \times x_2 + p_3 \times x_3 + \cdots + p_n \times x_n}{p_1 + p_2 + p_3 + \cdots + p_n}, i = 1, 2, 3 \ldots, n$$

sendo $x_1, x_2, x_3, \ldots, x_n$ cada valor associado a um $i-simo$ elemento da amostra e $p_1 + p_2 + p_3 + \cdots + p_n$ cada peso relacionado com cada elemento da amostra.

Quando a amostra possuir muitas repetições ou precisar ser balizada por algum peso, esta média é mais recomendada.

Exemplo 2.2. Com base no conjunto de dados de IES, vamos calcular a média da receita própria ponderada pelo total de técnicos.

```
base_ies %>%
 dplyr::summarise(`Receita própria` = mean(ReceitaPropria),
       `Receita própria ponderada` = weighted.mean(x = ReceitaPropria,
                                   w = TotalTecnicos)) %>%
 knitr::kable(format = tb_formata,
        caption = "Média receita própria ponderada pelo total de técnicos") %>%
 kableExtra::kable_styling(latex_options = "hold_position")
```

TABELA 2.6: Média receita própria ponderada pelo total de técnicos

Receita própria	Receita própria ponderada
145378180	229043797

Note da tabela 2.6 que a média ponderada é maior que a aritmética da tabela 2.5 e isso ocorre porque IES maiores possuem mais receita e sendo esta ponderada por um volume maior de técnicos faz com que a média suba.

A médias geométrica é mais rara, porém existem aplicações em áreas como ciências sociais como formas de estivar a expectativa de vida ao nascer, na economia como indicadores financeiros e na geometria. Assim como a média geométrica, a harmônica também é rara e possui aplicações na área da física em situações que envolvem taxas. Fica ao cargo do leitor interessado pesquisar mais sobre estas médias.

2.2.1.2 Mediana (M_d)

Para compreender melhor o conceito de mediana é importante saber que ela depende da ordenação de forma crescente dos dados da variável numérica. Como se trata de números, ao ordenar os dados podemos trazer a ideia de centro. A mediana é uma medida estatística que calcula o valor central dos dados de forma que se tenha metade dos valores abaixo e metade acima da mediana. Resumindo, mediana é o valor do meio do conjunto de dados.

Quando o conjunto de dados tiver um número impar de observações, a media será o valor central e quando o comprimento for par, a mediana será a média dos dois elementos centrais. A mediana também é conhecida como uma medida resistente a pontos discrepantes.

Exemplo 2.3. No conjunto de dados `c(0, 1, 2, 3, 4, 5, 6)` qual é a mediana?

Neste caso a mediana é 3, porque a amostra tem tamanho 7 e 3 é o elemento que separa os dados 50% / 50%. No R utilizamos a função `median()` para calcular a mediana.

```
x <- c(0, 1, 2, 3, 4, 5, 6)
median(x, na.rm = TRUE)
```

```
## [1] 3
```

```
# OBS: na.rm remove elementos nulos da amostra, quando existirem
```

Exemplo 2.4. No conjunto de dados `c(2, 3, 4, 5, 6, 7)` qual é a mediana?

Neste caso a mediana é $\frac{4+5}{2} = 4.5$, porque a amostra tem tamanho 6 e sendo que 4 e 5 são os elementos que estão no centro.

```
x <- c(2, 3, 4, 5, 6, 7)
median(x, na.rm = TRUE)
```

```
## [1] 4.5
```

2.2.1.3 Moda (M_o)

A moda é uma medida estatística que aponta quem são os valores mais frequentes numa amostra com elementos repetidos sendo ela o valor que ocorre com maior frequência ou o valor mais comum. Quando os dados são numéricos e já estão agrupados em classes, chamamos a classe com maior frequência de **classe modal** e seu valor é determinado pela média dos seus extremos.

Diferentemente da média e mediana já vistas, a moda também se

aplica a variáveis categóricas, uma vez que serve para identificar as classes ou valores mais frequentes.

Quando uma amostra possui apenas uma moda diz-se que ele é *unimodal*, sem tem duas é *bimodal* e se tem três ou mais é dita *multimodal*.

```
knitr::kable(rnp::rnp_freq(base_docentes$idade),
       digits = 3,
       booktabs = TRUE,
       format = tb_formata,
       caption = "Frequências para idade dos docentes das IES") %>%
 kableExtra::kable_styling(latex_options = "hold_position")
```

TABELA 2.7: Frequências para idade dos docentes das IES

classe	fa	fr	Faa	Fra
19–36	107295	0.274	107295	0.274
36–43	95550	0.244	202845	0.517
43–52	92360	0.236	295205	0.753
52–99	96831	0.247	392036	1.000

Na tabela 2.7 vemos que a classe modal é a que contém idades entre 20 e 36 anos, pois ela representa 28,6% da amostra. Com base na classe modal, temos que $M_o = \frac{36+20}{2} = 28$ anos.

Exemplo 2.5. Dadas as amostras genéricas `x = {2, 5, 3, 4, 4}, y = {5,5,7,7,6,1,2,1}, z = {9,1,7,8,4}` determine, quando existir e moda e sua classificação.

Uma forma de verificar se um conjunto de dados possui moda é verificar se tem algum valor que se repete ao longo da amostra. Isso pode ser feito através das funções `table()` e `duplicated()`. Enquanto a primeira faz uma tabulação dos valores ou das classes, a segunda

varre a variável buscando quem são os valores que ocorrem mais de uma veze retornando TRUE, caso algum se repita. Vejamos a solução.

```
# Preparando os dados como vetores através da função c() e ordenando com sort()
x <- sort(c(2, 5, 3, 4, 4))
paste("x possui", x[duplicated(x)], "como moda")
```

```
## [1] "x possui 4 como moda"
```

```
# x é unimodal

y <- sort(c(5,5,7,7,6,1,2,1))
paste("y possui", y[duplicated(y)], "como moda")
```

```
## [1] "y possui 1 como moda" "y possui 5 como moda"
## [3] "y possui 7 como moda"
```

```
# y é multimodal

z <- sort(c(9,1,7,8,4))
paste("z possui", z[duplicated(z)], "como moda")
```

```
## [1] "z possui  como moda"
```

```
# z não possui moda.
```

As medidas de centralidade de forma geral sempre buscarão representar quais dados estão no centro ou apontando para mesmo no conjunto de dados. A figura 2.5 exemplifica um conjunto de medidas em uma linha onde a média representa o centro e os pontos as possíveis medidas realizadas.

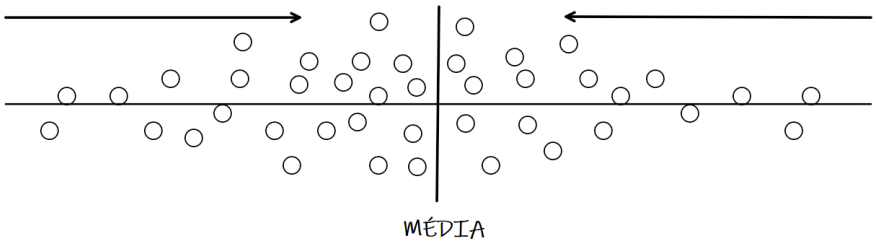


FIGURA 2.5: Centralidade

2.2.1.4 Exercícios

Utilize a base de dados das IES para responder à questão a seguir.

Exercício 2.9. Interprete a média e mediana para as seguintes variáveis: QT_TEC_MEDIO_FEM, QT_TEC_MEDIO_MASC, QT_TEC_SUPERIOR_FEM e QT_TEC_SUPERIOR_MASC.

Solução. Aplicando as funções mean() e median().

```
variaveis <- c("QT_TEC_MEDIO_FEM", "QT_TEC_MEDIO_MASC",
         "QT_TEC_SUPERIOR_FEM", "QT_TEC_SUPERIOR_MASC")
tb <- sapply(variaveis, function(i) {
 base_ies %>%
  select(i) %>%
  summarise("Média"   = mean(.[[i]], na.rm = TRUE),
        "Mediana" = median(.[[i]], na.rm = TRUE))
}) %>%t()

knitr::kable(tb, booktabs = TRUE, format = tb_formata, digits = 3,
       caption = "ex: Média e mediana técnicos (IES)") %>%
 kableExtra::kable_styling(latex_options = "hold_position")
```

A tabela mostra quem média as IES possuem 27 técnicos de nível médio, masculino ou feminino com mediana de 6 para feminino e 4 para masculino. Isso significa que cerca de 50% das IES possuem

TABELA 2.8: ex: Média e mediana técnicos (IES)

	Média	Mediana
QT_TEC_MEDIO_FEM	27.2561274509804	6
QT_TEC_MEDIO_MASC	27.5065359477124	4
QT_TEC_SUPERIOR_FEM	27.8692810457516	6
QT_TEC_SUPERIOR_MASC	20.3729575163399	3

até 10 (6 + 4) técnicos de nível médio. Para os técnicos de nível superior, a média feminina é superior em quase 8, sendo de 27,87 contra 20,37 dos masculinos. A mediana de mulheres se iguala àquelas com curso técnico.

2.2.2 Medidas estatísticas de dispersão

De forma simples, podemos entender medidas de dispersão como estatísticas que medem o quanto os dados estão espalhados. Desta forma, este tipo de medida é zero se os dados são todos iguais e vai aumentando à medida em que a diversidade dos dados aumenta. Estatísticas de dispersão são muito aplicadas em área como física ajudando a medir a variabilidade de medições feitas em experimentos; nas ciências biológicas estimando a variabilidade **interindivíduos** (membros distintos da mesma amostra são diferentes uns dos outros) e **intraindivíduos** (um mesmo individuo submetido a algum teste em condições distintas produzem resultados diferentes) e em muitos ramos das ciências como economia, medicina e engenharia. As principais medidas estatísticas de dispersão são **Desvio padrão (S, σ) e Variância (S^2, σ^2), Amplitude, Desvio absoluto e Coeficiente de variação**

2.2.2.1 Desvio padrão (S, σ) e variância (S^2, σ^2)

Desvio padrão e variância são medidas que buscam estimar a dispersão dos dados em torno da sua média. Quando estamos falando de população temos o **desvio padrão ou variância populacional**, representados pela letra grega minúscula σ para desvio padrão e σ^2 para variância. No caso de amostra, representamos pela letra latina S para o primeiro e S^2 para o segundo caso.

O desvio padrão populacional é dado por:

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

em que $X_i, i = 1, 2, ..., N$ são os elementos da população e μ é a média populacional.

O desvio padrão amostral é dado por:

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

em que $X_i, i = 1, 2, ..., n$ são os elementos da amostra e $\bar{X}$ é a média amostral.

O denominador do desvio padrão amostral é $n - 1$ em vez de n. Este fator de correção é conhecido como correção de Bessel (Reichmann, 1961) e é aplicado porque no cálculo da média a partir da amostra, perde-se um **grau de liberdade**. Grau de liberdade refere-se ao total de elementos da amostra supondo que cada um é independente do outro. Como S utiliza $\bar{X}$ que por sua vez está ligada com cada elemento da amostra, há epenas $n - 1$ elementos independentes após $\bar{X}$ ser calculado.

A variância é o quadrado do desvio padrão. Assim:

Variância populacional é dada por

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

e a variância amostral por:

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

Em R calculamos o desvio padrão de uma variável com a função sd() e a variância com a função var().

Exemplo 2.6. Calcule a média, desvio padrão e a variância da idade dos docentes das IES.

```
base_docentes %>%
 dplyr::summarise(`Média` = mean(idade),
           `Desvio padrão` = sd(idade),
           `Variância` = var(idade))
```

```
## # A tibble: 1 x 3
##   Média `Desvio padrão` Variância
##   <dbl>           <dbl>     <dbl>
## 1  44.5            11.0      120.
```

ou diretamente com.

```
paste("Desvio padrão =", round(sd(base_docentes$idade), 3))
```

```
## [1] "Desvio padrão = 10.963"
```

```
paste("Variância =", round(var(base_docentes$idade), 3))
```

```
## [1] "Variância = 120.181"
```

- **Interpretação do desvio padrão**: No exemplo acima vemos que a média de idade dos docentes é de 44,2 anos com desvio padrão de 11 e variância de 120, mas o que isso significa? - A variância de idade é uma medida cuja unidade de media é ano^2. Ano ao quadrado não tem interpretação direta então utilizamos o desvio padrão. Em geral quanto maior o desvio padrão mais espalhados estão os dados em relação à media. Não é conhecida

uma regra generalizada para dizer se um desvio padrão é menor ou maior, porém, com base na teoria das probabilidades temos uma regra de ouro que é aplicada sempre que a curva dos dados segue uma distribuição Normal (por hora, epenas aceite, veremos ela mais adiante!).

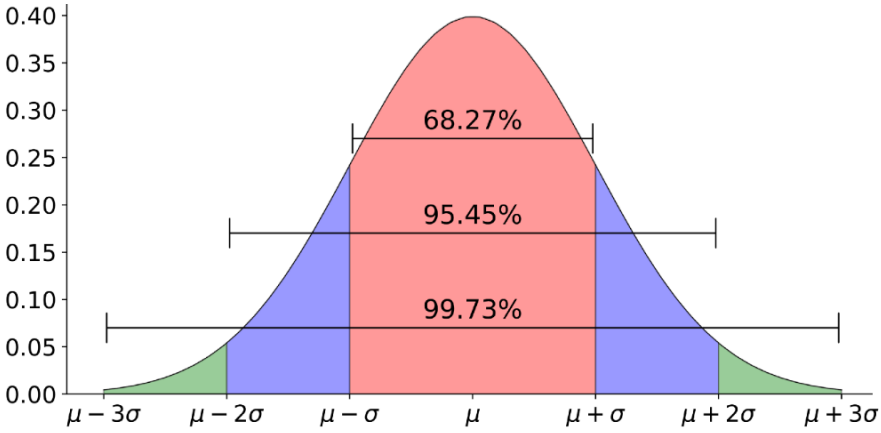


FIGURA 2.6: Centralidade

Conforme vemos na figura 2.6, em torno da média mais ou menos um desvio padrão devem estar 68,27% dos dados, já entre a média mais ou menos 2 desvios padrão deve estar 95,45% dos dados. Seguindo esta lógica, a interpretação deve levar em conta a distribuição dos dados e a precisão que o experimento ou estudo exige. Assim, sendo no nosso exemplo a média de idade dos docentes é 44,2 então entre $44,2 \pm 11 = (33.4 - 55.4)$ devem estar 68,27% dos docentes.

2.2.2.2 Amplitude

A amplitude de um conjunto de dados ordenado é a distância entre o menor e o maior valor. Na figura 2.5 se seus valores estiverem ordenados do maior para o menor, a amplitude seria o ponto mais à direita **Máximo** menos o ponto mais à esquerda **Mínimo**.

Representamos um conjunto de dados ordenado da seguinte forma.

$$X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}$$

Assim, sendo podemos expressar a amplitude ou range por

$$R = X_{(n)} - X_{(1)} = Max(X) - Min(X)$$

Exemplo 2.7. Qual a amplitude da idade dos docentes?

```
# Ordenando os dados do menor para o maior
idade <- sort(base_docentes$idade, decreasing = FALSE)
# obtendo o menor e o maior valor
menor <- idade[1]
maior <- idade[length(idade)]
R1 <- maior - menor
# ou pelo minimo e máximo dos dados
R2 <- max(idade) - min(idade)
paste("As duas medidas são iguais?", all.equal(R1,R2))
```

```
## [1] "As duas medidas são iguais? TRUE"
```

```
c(R1,R2)
```

```
## [1] 80 80
```

2.2.2.3 Coeficiente de variação cv

Esta medida estatística muitas vezes é chamada de desvio padrão relativo e é uma medida padronizada de dispersão. Em alguns contextos é possível optar pelo cv ao invés do desvio padrão. Quanto maior for o coeficiente de variação, maior será a dispersão nos dados em torno da média. O cv é divisão entre o desvio padrão e a média e pode ser calculado pela seguinte expressão.

$$cv_{amostral} = \frac{S}{\bar{X}}$$

$$cv_{populacional} = \frac{\sigma}{\mu}$$

Vale salientar que o cv e o desvio padrão se aplicam a dados estritamente positivos.

- **Interpretação**: por ser uma medida adimensional, o cv é uma medida prática para interpretar a variabilidade entre dois conjuntos de dados de tipos diferentes e pode ser interpretada em termos percentuais. Veja o exemplo a seguir.

Exemplo 2.8. Vamos determinar e interpretar o coefiente de variação da receita das IES em relação ao total de técnicos.

```
base_ies %>%
 dplyr::summarise(`Média receita` = mean(ReceitaPropria),
                  `Média técnicos` = mean(TotalTecnicos),
                  `CV receita` = sd(ReceitaPropria) / mean(ReceitaPropria),
                  `CV técnicos` = sd(TotalTecnicos) / mean(TotalTecnicos)) %>%
 round(., digits = 3)
```

```
## # A tibble: 1 x 4
##   `Média receita` `Média técnicos` `CV receita`
##             <dbl>            <dbl>        <dbl>
## 1     145378180.              168.         2.87
## # ... with 1 more variable: `CV técnicos` <dbl>
```

Os cálculos mostram uma enorme variabilidade dos dados das IES, pois o cv para receita é 286% e para total de técnicos é 347%. Neste caso, temos indicativos de que a dispersão dos dados é grande. Isso pode ser explicado pelo tamanho das IES. Por exemplo, as federais são minoria na base de dados, mas possuem grande quantidade de técnicos e alto aporte de receita, enquanto as IES menores, geralmente privadas possuem menor número de técnicos e menor receita. Estados como São Paulo apresentam número muito grandes em relação ao restante do país.

2.2.3 Outras medidas

Existem muitas estatísticas úteis para analisar dados numéricos que nem sempre são exploradas, entre elas temos os *quartis, decis, percentis e amplitude interquartis.*

2.2.3.1 Quartis, decis e percentis

- **Quartis**: Chamamos de quartil qualquer uma das três medidas que separam um conjunto de dados ordenado em q partes iguais. Quartil vem de 1/4 (um quarto dos dados). A mediana que já vimos representa o segundo quartil. Costumamos representar os quatis pela letra Q seguida de um número tais como:
  - $Q1$: primeiro quartil representa 25% da amostra ordenada;
  - $Q2$: segundo quartil ou mediana representa 50% da amostra ordenada;
  - $Q3$: terceiro quartil representa 75% da amostra ordenada;
- **Decil**: O raciocínio é o mesmo dos quatis. Decis são medidas que dividem o conjunto de dados em 10 partes iguais. O primeiro decil representa 10% dos dados, o segundo 20% e assim por diante.
- **Percentil**: Percentis semelhante aos decis, os percentis dividem o conjunto de dados em 100 partes iguais.

Para obter estas estatísticas seguimos o mesmo racional da mediana, dividindo os dados em partes iguais e identificando os elementos do centro e borda. No R os podemos calcular facilmente estas estatísticas pela função `quantile()` para qualquer tamanho de faixa e por `summary()` para os quartis. Sempre que desejar fazer um raio-x dos dados é sugerido fazer uma análise de quartil, decil ou percentil, pois desta forma ficará evidente qualquer anomalia nos dados.

2.2.3.2 Amplitude interquartil

A amplitude interquartil ou do inglês *InterQuartile Range (IQR)* é a medida de distância ou range entre o primeiro quartil Q_1 e o terceiro Q_3, sua importância reside no fato de que ela representa os 50% dos dados centrais do conjunto de dados.

$$IQR = Q_3 - Q_1$$

Junto com esta estatística surge também dois conceitos importantes que são os limites superiores LS e inferiores LI para decidir se determinado ponto é **discrepante** ou não. Uma dada medida é dita discrepante ou *outlier* quando ela está muito diferente da maioria das medidas realizadas. É demonstrado que no intervalo determinado por $LI = Q_1 - 1.5 \times IQR$ e $LS = Q_3 + 1.5 \times IQR$ temos 99% dos dados, assim qualquer valor que cair fora deste intervalo em geral, pode ser chamado de *outlier*.

Exemplo 2.9. Vamos determinar se existe algum outlier no conjunto de dados das idades dos docentes das IES.

```
base_docentes %>%
  dplyr::summarise(Q1 = quantile(idade, probs = 0.25),
                   Q2 = quantile(idade, probs = 0.50),
                   Q3 = quantile(idade, probs = 0.75),
                   IQR = Q3 - Q1,
                   LI = Q1 - 1.5*IQR,
                   LS = Q3 + 1.5*IQR,
                   Noutliers = sum(idade > LS),
                   Ntotal = n(),
                   Pct = Noutliers / Ntotal) %>%
  round(., digits = 3) %>%
  print(options(tibble.width = Inf))
```

```
## # A tibble: 1 x 9
##      Q1    Q2    Q3   IQR    LI    LS Noutliers Ntotal
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>  <dbl>
## 1    36    43    52    16    12    76      951 392036
## # ... with 1 more variable: Pct <dbl>
```

Conforme a analise acima, vemos que apenas $\frac{951}{392036} = 0,24\%$ dos docentes são outliers possuindo idade acima de 76 anos.

Outliers possuem grande importância na estatística e nunca devem ser negligenciados, pois podem trazer informação valiosa para a análise. Existem muitas técnicas de detecção de outiers mais robustas que esta que vimos a partir dos quartis. Ao leitor interessado ver (Barnett and Lewis, 1974) e para uma visão baseada em R ver (Komsta, 2011).

2.2.3.3 Os cinco números

Os cinco números são um conjunto de estatísticas composto por Min, Q_1, Q_2, Q_3 e Max, estas cinco estatísticas costumam ser suficientes para analisar a distribuição dos dados pois junta as estatísticas mais importantes, a mediana representando uma medida de centralidade, os quartis $Q1, Q3$ representado medidas de dispersão e o mínimo e máximo que representam o range dos dados.

É comum em estatística, juntarmos em uma tabela as principais estatísticas de uma variável numérica para interpretar sua relevância no contexto do estudo ou experimento em questão. Além do resumo dos cinco números, podemos acionar outras estatísticas de nosso interesse. A função `rnp_summary()` em conjunto com `rnp_freq()` nos auxiliarão em muitas análises no curso deste livro.

2.2.3.4 Exercícios

Ainda na base das IES, analise através de estatísticas de dispersão as questões a seguir.

Exercício 2.10. A despesa anual com docentes é representada pela variáve `VL_DESPESA_PESSOAL_DOCENTE`. Interprete estatísticas de dispersão para esta variável.

```
tb <- rnp::rnp_summary(base_ies$VL_DESPESA_PESSOAL_DOCENTE)[-c(1:3)]
knitr::kable(t(tb), booktabs = TRUE, format = tb_formata, digits = 3,
       caption = "ex: Média e mediana técnicos (IES)") %>%
 kableExtra::kable_styling(latex_options = "hold_position")
```

Solução. Como podemos ver na tabela acima, o desvio padrão da

TABELA 2.9: ex: Média e mediana técnicos (IES)

Min	Q1	Media	Mediana	Q3	Max	DevPad	cv
1	696484	66328326	3295507	37257632	7600039210	242807546	3.661

depesa com pessoal é muito grande e isso de deve ao fato de que existe grande variabilidade nos dados, escialmente por conta das IES federais em contraste com as IES particulares menores. O coeficiente de variação (cv) mostra que esta variabilidade em torno da média é de 366,1 por cento. Temos ainda que 75 por cento (Q_3) das IES gastam com docentes anualmente até R$37.257.632.

Exercício 2.11. Qual o top 10 IES com maior receita própria? Dica: Para resolver este problema, isolamos as variáveis NO_IES e VL_RECEITA_PROPRIA ordenada da maior para a menor.

```
base_ies %>%
 dplyr::transmute(NomeIES = NO_IES, ReceitaPropria = VL_RECEITA_PROPRIA) %>%
 dplyr::arrange(desc(ReceitaPropria)) %>%
 head(n = 10)
```

```
## # A tibble: 10 x 2
##   NomeIES                       ReceitaPropria
##   <chr>                          <dbl>
##  1 FACULDADE DE ADMINISTRAÇÃO, CIÊNCIAS~   6248050290
##  2 FACULDADE MAURÍCIO DE NASSAU DE MOSS~   4434457964.
##  3 FACULDADE MAURÍCIO DE NASSAU DE MACE~   4024118925
##  4 Faculdade Fernanda Bicchieri            4000452319.
##  5 FACULDADE ESTÁCIO DE SÁ DE VILA VELHA   3131607366.
##  6 FACULDADE ESTÁCIO DE SÁ DE CAMPO GRA~   3131607366.
##  7 FACULDADE ESTÁCIO DE SÁ DE GOIÁS        3131607366.
##  8 CENTRO UNIVERSITÁRIO ESTÁCIO JUIZ DE~   3131607366.
##  9 UNIVERSIDADE ESTÁCIO DE SÁ              3131607366.
## 10 FACULDADE ESTÁCIO DE SÁ DE OURINHOS     3131607366.
```

2.2.4 Gráficos para uma variável numérica

Existem muitos tipos de gráficos, porém para uma variável listamos os três que consideramos mais importantes.

2.2.4.1 Histogramas

Os histogramas são um tipo de gráfico de barras para variáveis numéricas e servem principalmente para analisar visualmente a centralidade e dispersão dos dados. No processo de construção do histograma, os dados são categorizados em classes e as frequências são contadas. No eixo horizontal geralmente são mostradas as classes e eixo vertical as frequências que podem ser absolutas ou relativas.

Exemplo 2.10. Vamos criar um histograma para a variável idade dos docentes.

```
p <- ggplot2::ggplot(base_docentes, aes(x = idade))
p + ggplot2::theme_gray() +
 ggplot2::geom_histogram(colour='white', bins = 10) +
 ggplot2::labs(y = "Frequência", x = "Faixa de idade", fill = NULL, title = "") +
  ggplot2::scale_x_continuous(
   breaks=seq(10, 90, 10),
   labels = seq(10, 90, 10)
  ) +
 ggplot2::theme(axis.line = element_blank())
```

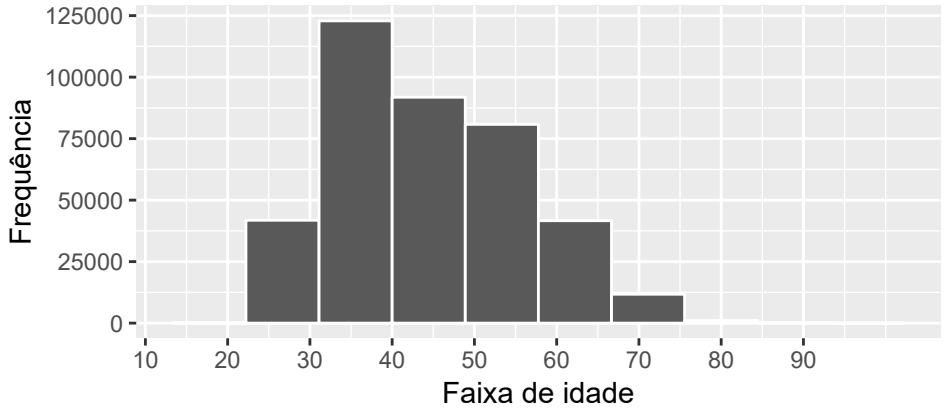


FIGURA 2.7: Histograma idade do docente

Veja que a figura 2.7 evidencia que as maiores concentrações de docentes estão nas faixas de idade entre 30 e 50 anos.

2.2.4.2 Densidade

Gráficos de densidade possuem aplicação semelhante aos histogramas, porém são mais indicados pata amostras grandes. Ele evidenciam a melhor curva que representam os dados. Este tipo de gráfico nos ajuda também a verificar a distribuição de probabilidade aproximada que os dados podem seguir.

Exemplo 2.11. Vamos criar agora um gráfico de densidade para a variável idade dos docentes.

```
p <- ggplot2::ggplot(base_docentes, aes(x = idade))
p + ggplot2::theme_gray() +
 ggplot2::geom_density(adjust = 1) +
 ggplot2::labs(y = "Densidade", x = "Idade", fill = NULL, title = "") +
 ggplot2::scale_x_continuous(
  breaks=seq(10, 90, 10),
  labels = seq(10, 90, 10)) +
 ggplot2::theme(axis.line = element_blank())
```

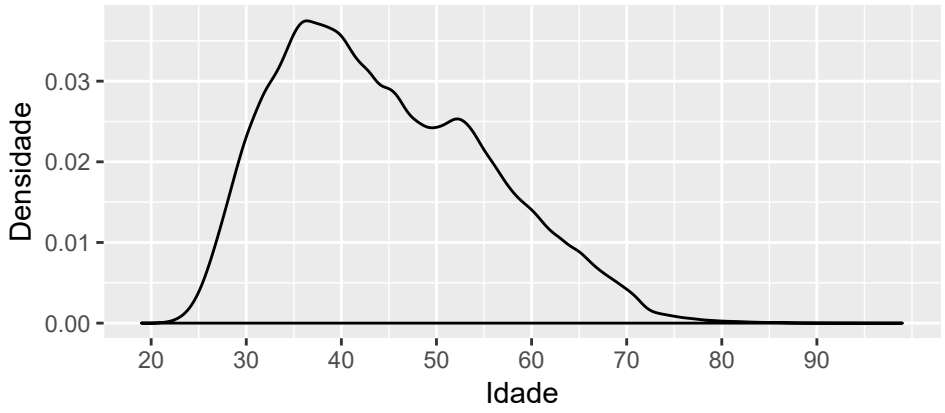


FIGURA 2.8: Densidade idade do docente

Perceba na figura 2.7 como esperado, que a densidade dos dados está concentrada idade entre 30 e 50 anos. Porém, ela aponta uma elevação próxima a 50 anos apontando comportamento bimodal nos dados.

2.2.4.3 Box-Plot

De longe o gráfico Box-plot ou para muitos, diagrama de caixa é o tipo mais completo de gráfico para variável numérica.

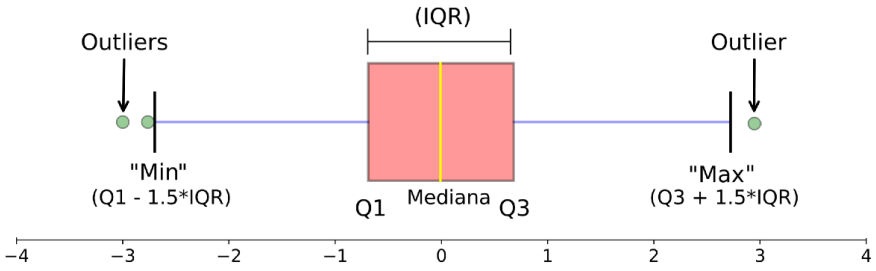


FIGURA 2.9: Definindo um Box-plot

A figura 2.9 ilustra os elementos que compõem um Box-plot.

Perceba que visualmente ele contempla as estatísticas Q_1, Q_2 (mediana), Q_3, IQR, LI, LS e *outliers*. Com base nestas estatísticas, uma variável numéricas estará bem caracterizada.

Exemplo 2.12. Ainda para os dados de idade, vamos montar um Box-plot

```
p <- ggplot2::ggplot(base_docentes, aes(y = idade))
p + ggplot2::theme_gray() +
 ggplot2::geom_boxplot(adjust = 1) +
 ggplot2::labs(x = "", y = "Idade", fill = NULL, title = "") +
 ggplot2::coord_flip() +
 ggplot2::theme(axis.line.x = element_line(),
          axis.text.x = NULL,
          axis.line.y = element_blank())
```

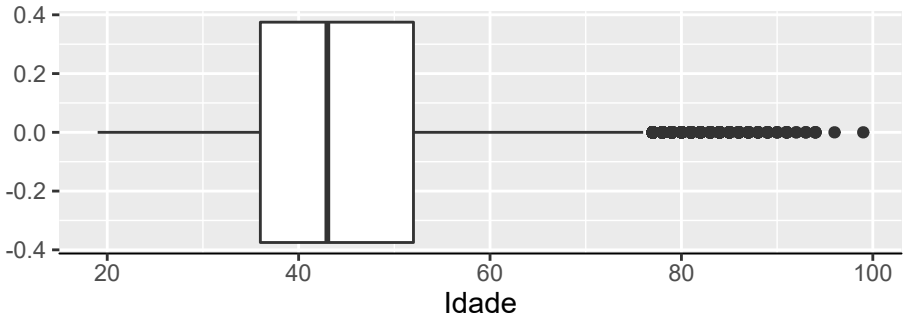


FIGURA 2.10: Box-plot idade do docente

Conforme vimos antes, idades acima de 76 anos são pontos atípicos na base de docentes, por isso no Box-plot estes pontos aparecem fora do limite superior de outliers na figura 2.19.

2.2.5 Gráficos para duas variáveis numéricas

Os principais gráficos para analisar a relação entre duas variáveis são o gráfico de pontos ou *scatterplot* e o gráfico de linhas, através

deles é possível analisar a relação conjunta entre as variáveis e determinar se uma influencia a outra de alguma forma. além destes, também podemos traçar gráficos de densidade para comparar as duas curvas.

2.2.5.1 *scatterplot*

O gráfico de pontos é um gráfico bidimensional onde cada eixo representa os valores de uma variável. Este tipo de gráfico é ótimo para analisar a correlação de duas variáveis bem como sua dispersão, pois cada ponto representa a ligação dos elementos das duas variáveis.

Exemplo 2.13. Para lustrar vamos traçar um gráfico de pontos para a receita próprias das IES pelo total de técnicos na base das IES. Obs.: como temos muitos autliers na variável de receita própria, vamos limitar a 10.000.000 (dez milhões de reais por ano)

```
p <- base_ies %>%
 dplyr::filter(ReceitaPropria <= 10000000) %>%
 ggplot2::ggplot(aes(x = TotalTecnicos, y = ReceitaPropria))
p + ggplot2::theme_gray() +
 ggplot2::geom_point() +
 ggplot2::labs(x = "Total técnicos", y = "Receita própria", fill = NULL, title = "") +
 ggplot2::theme(axis.line.x = element_line(),
          axis.text.x = NULL,
          axis.line.y = element_blank())
```

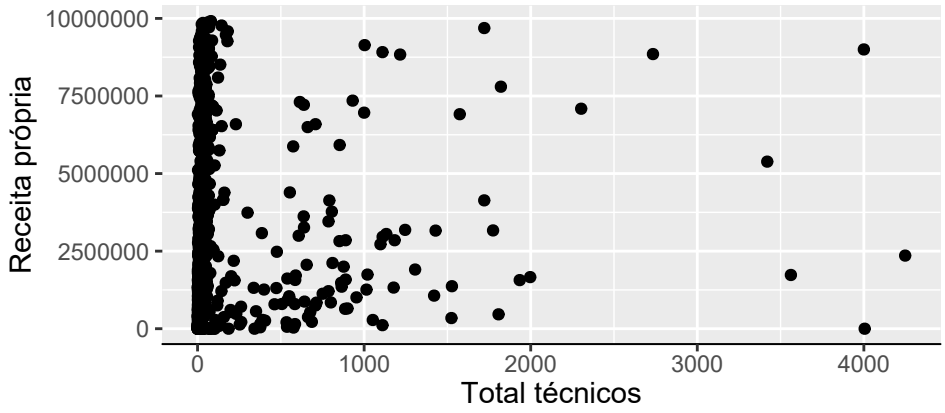


FIGURA 2.11: Gráfico de pontos total de técnicos versus receita própria

2.2.5.2 Grafico de linhas

O gráfico de linhas possui aplicação para duas variáveis contínuas e também para séries temporais, onde um dos eixos é uma variável numérica de data.

```
p <- base_ies %>%
 dplyr::filter(TotalTecnicos < 100) %>%
 ggplot2::ggplot(aes(x = TotalTecnicos, y = ReceitaPropria))
p + ggplot2::theme_gray() +
 ggplot2::geom_line() +
 ggplot2::labs(x = "Total técnicos", y = "Receita própria",
         fill = NULL, title = "") +
 ggplot2::theme(axis.line.x = element_line(),
          axis.text.x = NULL,
          axis.line.y = element_blank())
```

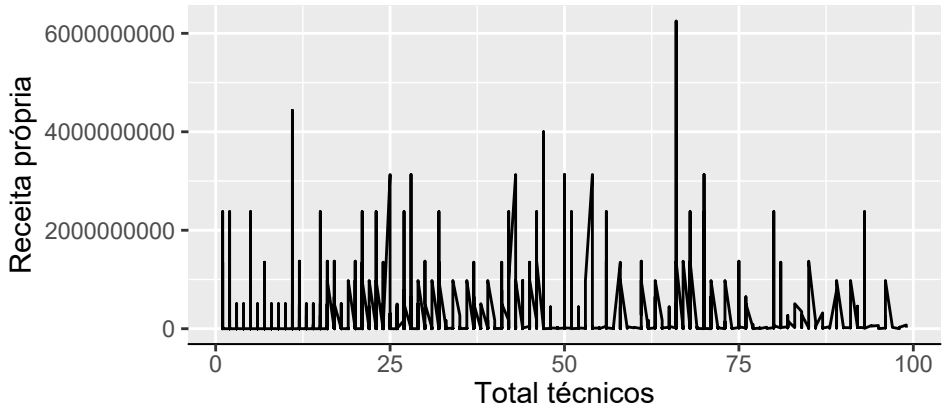


FIGURA 2.12: Gráfico de pontos total de técnicos (<50) versus receita própria

2.3 Variáveis categóricas versus numéricas

O trabalho estatístico muitas vezes exige que alguma variável numérica seja categorizada ou analisada em conjunto com alguma variável categórica. Todas as técnicas vistas até agora, tanto as medidas estatísticas quanto os gráficos podem ser analisados em conjunto para gerar informação.

2.3.1 Categorizando variáveis numéricas.

Em R podemos agrupar uma variável numérica de muitas formas, umas delas é através das estatísticas de quartis, decis ou percentis dependendo do tamanho da base, com a função quantile() combinada com a função cut(). Outra forma é através do conhecimento próprio do analista com os operadores relacionais do R: <, >, >=, <=, ==, !=, %in% combinadas com ifelse().

Exemplo 2.14. Vamos categorizar a variável receita próprias das

IES de duas formas distintas: por quartis e por operadores relacionais com quatro faixas.

```
## Categorizando por quartis
fx_receita_q <- cut(base_ies$ReceitaPropria,
          breaks = round(quantile(base_ies$ReceitaPropria)),
          dig.lab = 10, include.lowest = TRUE)
rnp::rnp_freq(fx_receita_q, sortd = FALSE) %>%
knitr::kable(digits = 3,
       booktabs = TRUE, format = tb_formata,
       caption = "Categorização por quartis") %>%
kableExtra::kable_styling(latex_options = "hold_position")
```

TABELA 2.10: Categorização por quartis

classe	fa	fr	Faa	Fra
0–1389966	612	0.25	612	0.25
1389966–7429221	612	0.25	1224	0.50
7429221–61258594	612	0.25	1836	0.75
61258594–6248050290	612	0.25	2448	1.00

```
## Categorização por ifelse com operadores relacionais
fx_receita_r <- ifelse(base_ies$ReceitaPropria <= 1200000, "A.-1200000",
         ifelse(base_ies$ReceitaPropria <= 5000000, "B.1200000 a 5000000",
         ifelse(base_ies$ReceitaPropria <=35000000, "C.5000000 a 35000000","D.35000000+")))

rnp::rnp_freq(fx_receita_r, sortd = FALSE) %>%
knitr::kable(digits = 3,
       booktabs = TRUE, format = tb_formata,
       caption = "Categorização por operadores relacionais") %>%
kableExtra::kable_styling(latex_options = "hold_position")
```

No primeiro caso, criamos os cortes utilizando os quartis da variável e em seguida passamos estes cortes para a função cut() que por sua vez particionou a variável de acor com as partes informadas

TABELA 2.11: Categorização por operadores relacionais

classe	fa	fr	Faa	Fra
A.-1200000	571	0.233	571	0.233
B.1200000 a 5000000	500	0.204	1071	0.438
C.5000000 a 35000000	635	0.259	1706	0.697
D.35000000+	742	0.303	2448	1.000

pelo argumento breaks. Desta forma fica mais rápido a categorização, mas o analista não tem como personalizar as faixas. Para atender a esta limitação o segundo método ajuda a customizar as faixas de acordo com a preferência ou necessidade. Embora exija um pouco mais de código, esta última opção é mais flexível.

2.3.2 Medidas estatísticas por agrupamento

Em muitas situações da análise de dados, estamos interessados em analisar a influência de uma variável categórica sob uma ou mais variáveis numéricas. Felizmente, a grande maioria das medidas estatísticas aprendidas até agora se aplicam a este tipo de análise que exige estatísticas agrupadas.

Exemplo 2.15. Vamos calcular estatísticas descritivas de idade dos docentes (em anos) por escolaridade.

```
base_docentes %>%
 dplyr::group_by(escolaridade) %>%
 dplyr::summarise(N   = n(),
          Min  = min(idade),
          Q1   = unname(quantile(idade, probs = 0.25)),
          Me   = mean(idade),
          Md   = median(idade),
          Q3   = unname(quantile(idade, probs = 0.75)),
          Max  = max(idade)
          #Dp   = sd(idade),
          #cv   = sd(idade)/mean(idade)
```

```
        ) %>%
 knitr::kable(digits = 2,
       booktabs = TRUE, format = tb_formata,
       caption = "Estatísticas descritivas idade vs escolaridade") %>%
kableExtra::kable_styling(latex_options = "hold_position")
```

TABELA 2.12: Estatísticas descritivas idade vs escolaridade

escolaridade	N	Min	Q1	Me	Md	Q3	Max
1. Sem graduação	10	23	47.25	52.60	54	63	66
2. Graduação	4613	20	27.00	37.78	33	45	94
3. Especialização	72301	19	35.00	43.01	41	50	93
4. Mestrado	154285	22	34.00	42.81	41	50	90
5. Doutorado	160827	19	38.00	47.06	46	54	99

Em $Q3$ temos que 75% dos docentes com Doutorado possuem idade até 54 anos. Vemos também alguns doutores excepcionais com idade mínima de 19 anos. A idade máxima registrada foi de 99 anos presente no grupos dos doutores. A categoria sem graduação tem dez indivíduos e é pouco representativa.

```
# Como nossa rnp_summary, poderiamos fazer assim
aggregate(idade ~ escolaridade,
     data = base_docentes,
     FUN = function(i) rnp::rnp_summary(i))
```

2.3.3 Gráficos para categóricas vs numéricas

Com apoio do pacote ggplot podemos combinar a maioria dos gráficos vistos até agora para analisar dados por grupo ou classes. Como citamos na seção 1.6 uma das ferramentas estatísticas é a análise bivariada e ela pode ser feita entre duas variáveis podendo ser de mesmo tipo ou de tipos diferentes

- **Box-plot agrupado**: Box-plots por categoria são muito infor-

mativos uma vez que resumem sete estatísticas fundamentais de uma variável continua conforma já vimos.

Exemplo 2.16. Represente visualmente a análise da tabela 2.12 através de Box-plots.

```
p <- base_docentes %>%
 ggplot2::ggplot(aes(x = escolaridade, y = idade, fill = escolaridade))
p + ggplot2::theme_gray() +
 ggplot2::geom_boxplot(show.legend = FALSE) +
 ggplot2::coord_flip()+
 ggplot2::theme(axis.line.x = element_line(),
          axis.text.x = NULL,
          axis.line.y = element_blank())
```

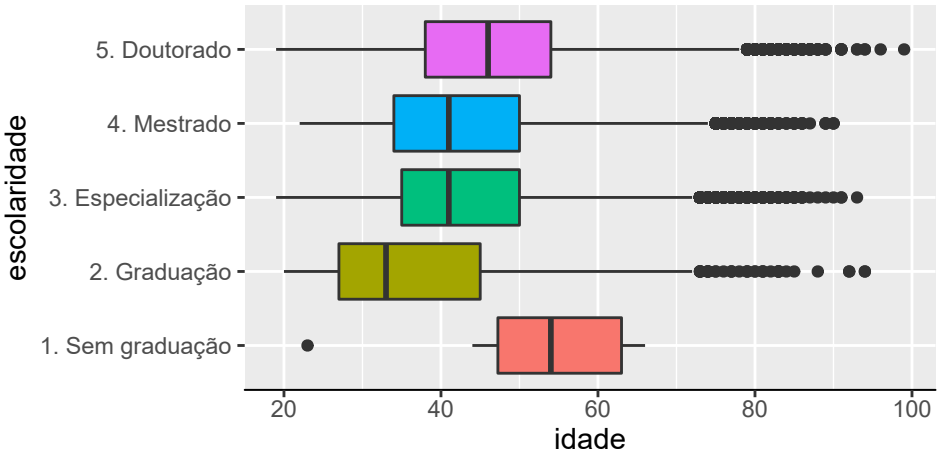


FIGURA 2.13: Box-plot de idade vs escolaridade

No Box-plot conseguimos ver que docentes com graduação apenas são minoria como visto na tabela e doutores são maioria. Notamos também que, com exceção daqueles sem graduação todos os grupos possuem dispersão parecida uma vez que a distância entre Q_1 e Q_3 é pequena.

Para todos os graficos agrupados também é possível quebrar a visualização em um gráfico por categoria adicionando uma terceira variável com facet_wrap() .

```
p <- base_docentes %>%
 ggplot2::ggplot(aes(x = sexo, y = idade, fill = escolaridade))
p + ggplot2::theme_gray() +
 ggplot2::geom_boxplot(show.legend = FALSE) +
 ggplot2::coord_flip()+
 ggplot2::facet_wrap(escolaridade~.)+
 ggplot2::theme(axis.line.x = element_line(),
          axis.text.x = NULL,
          axis.line.y = element_blank())
```

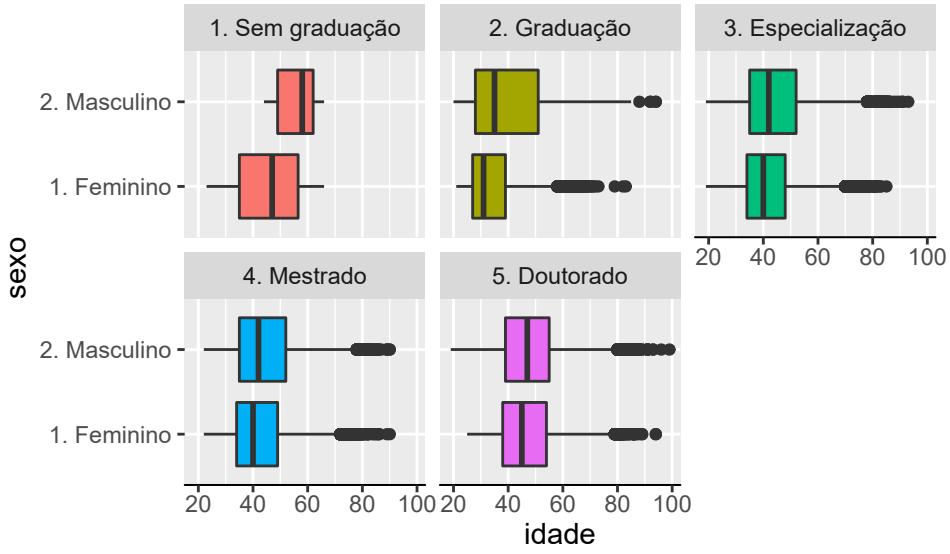


FIGURA 2.14: Box-plot de idade vs escolaridade por sexo

- **Gráfico de densidade agrupado**: é possível comparara

várias curvas simultaneamente no mesmo gráfico para analisar a distribuição dos dados.

Exemplo 2.17. Represente visualmente a análise da tabela 2.12 desta vez utilizando *densityplot*.

```
p <- base_docentes %>%
 ggplot2::ggplot(aes(x = idade, fill = escolaridade))
p + ggplot2::theme_gray() +
 ggplot2::geom_density(alpha = 0.6) +
 ggplot2::theme(legend.position = "right") +
 ggplot2::theme(axis.line.x = element_line(),
          axis.text.x = NULL, axis.line.y = element_blank()) +
 ggplot2::labs(x = "Idade", y = "Densidade", fill = NULL, title = "")
```

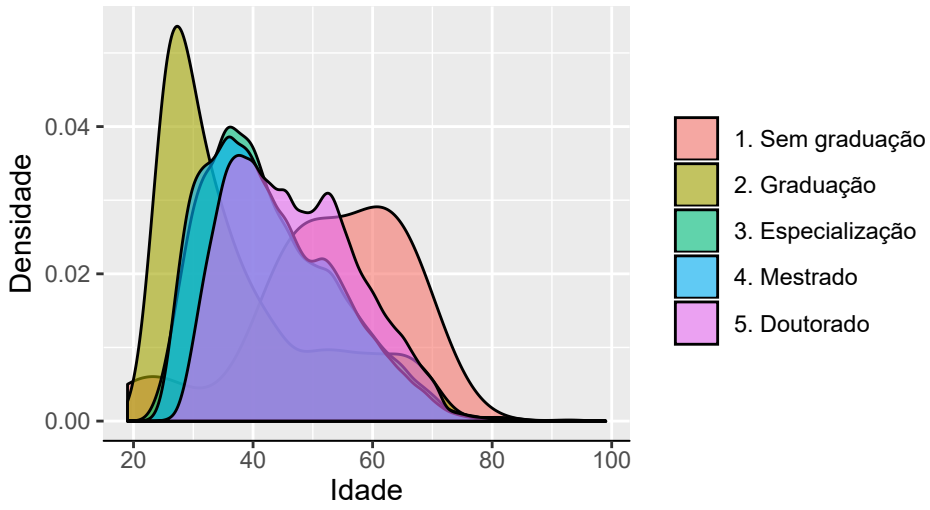


FIGURA 2.15: Densidade de idade vs escolaridade

O resultado é um gráfico elegante e que indica precisamente a forma da distribuição dos dados.

Agora quebrado por sexo.

```
p <- base_docentes %>%
 ggplot2::ggplot(aes(x = idade, color = sexo, fill = escolaridade))
p + ggplot2::theme_gray() +
 ggplot2::geom_density(alpha = 0.6, show.legend = FALSE) +
 ggplot2::facet_wrap(~escolaridade)+
 ggplot2::theme(axis.line.x = element_line(),
          axis.text.x = NULL,
          axis.line.y = element_blank()) +
 ggplot2::labs(x = "Idade", y = "Densidade", fill = NULL, title = "")
```

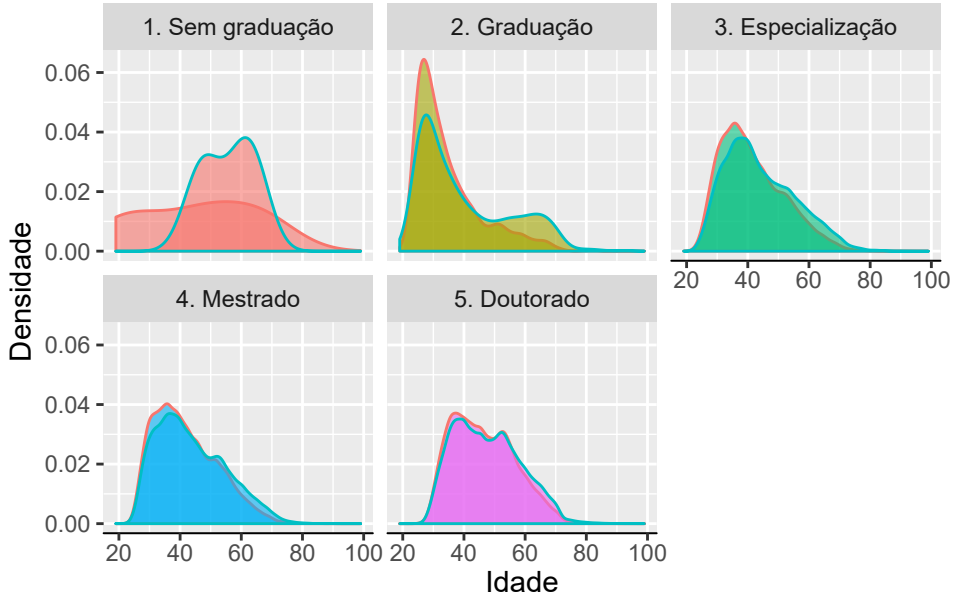


FIGURA 2.16: Densidade de idade vs escolaridade por sexo

- **Gráfico de colunas agrupadas**: é possível assim, como nos dois últimos exemplos, gerar gráficos e coluna agrupadas e fazer a quebra adicionando uma terceira variável.

```
p <- base_docentes %>%
 ggplot2::ggplot(aes(x = escolaridade, y = idade, fill = sexo))
p + ggplot2::theme_gray() +
 ggplot2::geom_col(show.legend = TRUE, position = "dodge") +
 ggplot2::theme(legend.position = "bottom") +
 ggplot2::coord_flip()+
 ggplot2::theme(axis.line.x = element_line(),
          axis.text.x = NULL,
          axis.line.y = element_blank())
```

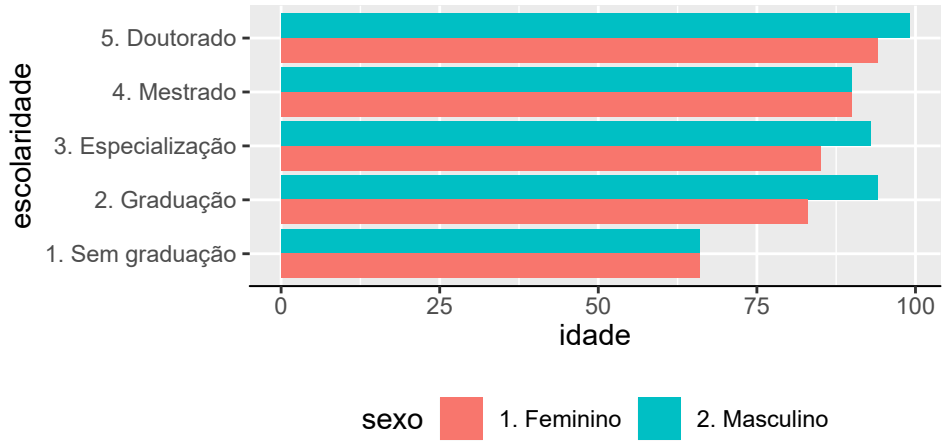


FIGURA 2.17: Colunas de idade vs escolaridade por sexo

- **Histogramas agrupados**: Histogramas também são agrupáveis segundo as categorias da variável.

```
p <- base_docentes %>%
 ggplot2::ggplot(aes(x = idade, fill = escolaridade))
p + ggplot2::theme_gray() +
 ggplot2::geom_histogram(bins = 10, show.legend = FALSE) +
 ggplot2::facet_wrap(.~escolaridade)+
 ggplot2::theme(axis.line.x = element_line(),
```

```
        axis.text.x = NULL,
        axis.line.y = element_blank())
```

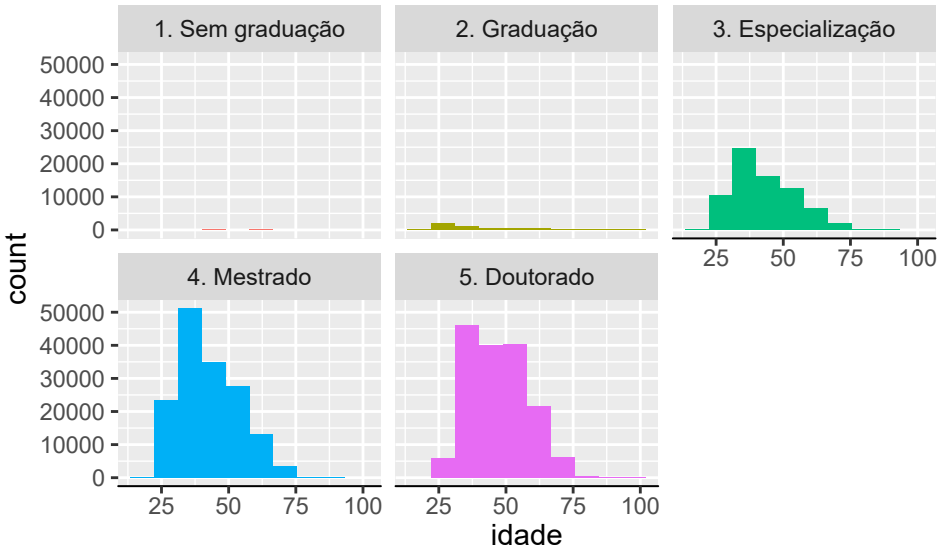


FIGURA 2.18: Histograma de idade vs escolaridade

2.4 Covariância e correlação

Na análise bivariada de variáveis numéricas a covariância e a correlação tem papel fundamental. Ambas se assemelham pelo fato de que medem a direção da relação entre duas variáveis ao longo de seus pontos. As relações mais comuns são *ambas as variáveis crescem, ambas decrescem, uma decresce e a outra cresce*. A diferença mais significativa entre a covariância e a correlação é que a primeira oferece um valor absoluto variando de acordo com os dados, graças a isso não tem como estimar a força de uma relação linear. É aí que entra a correlação. Em R a covariância pode ser calculada por `var()` e `cov()` e a correção por `cor()` e se aplica a bases com mais de 2 variáveis numéricas.

2.4.1 Covariância

O conceito de covariância pode ser aplicado tanto a conjunto de dados como a variáveis aleatórias. Quando aplicada a duas variáveis X e Y é expressa por:

$$Cov(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N-1}, i = 1, .., N$$

Sendo que $\bar{X}$ e $\bar{Y}$ são as médias das variáveis X e Y.

Exemplo 2.18. Vamos calcular a covariância para TotalTecnicos e ReceitaProprianos dados das IES

```
base_ies %>%
 dplyr::select(TotalTecnicos, ReceitaPropria) %>%
 cov() %>%
 knitr::kable(booktabs = TRUE, format = tb_formata,
        caption = "Análise de covariância") %>%
 kableExtra::kable_styling(latex_options = "hold_position")
```

TABELA 2.13: Análise de covariância

	TotalTecnicos	ReceitaPropria
TotalTecnicos	341089	14066046853
ReceitaPropria	14066046853	173738053485884832

Note que os números são extremamente grandes e tudo que podemos tirar é que a covariância é positiva entre as duas variáveis, mas não conseguimos estimar a força desta relação.

Os elementos da diagonal principal da matriz de covariâncias representam a variância de cada variável e os valores fora da diagonal são covariâncais. A matriz é espelhada e para interpretá-la basta ler acima ou abaixo da diagonal principal.

2.4.2 Correlação

A correlação linear ou de Pearson é uma estatística padronizada que varia de -1 a 1 e expressa a relação linear entre duas variáveis numéricas. A correlação é expressa pela letra grega ρ e é dada pela seguinte expressão matemática:

$$\rho = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2} \times \sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}} = \frac{Cov(X,Y)}{\sqrt{S^2(X) \times S^2(Y)}}, i = 1, .., N$$

Em resumo, esta formula aparentemente complicada diz que a correlação é resultado da covariância dividida pela raiz quadrada do produto das variâncias de cada variável.

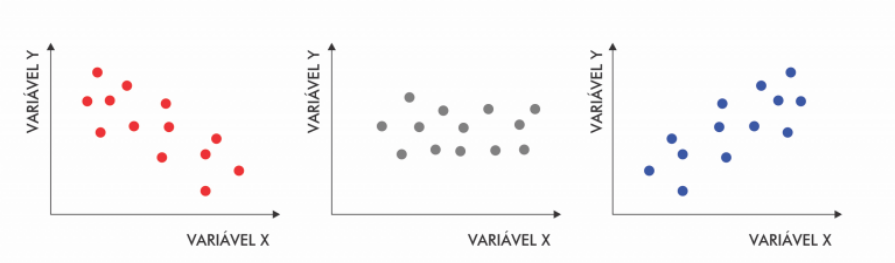


FIGURA 2.19: Tendência da correlação

A figura 2.19 ilustra a relação entre duas variáveis X e Y. Da esquerda para a direita temos **correlação linear negativa forte** (tende a -1), **correlação fraca** (tende a 0) e **correlação linear positiva forte** (tende a $+1$).

Como regra geral, costuma-se considerar a distribuição expressa na tabela 2.14 para interpretar a correlação. A mesma se aplica para negativa ou positiva bastando apenas dizer estar alerta ao sinal de ρ.

Exemplo 2.19. Para o exemplo anterior, determinar e interpretar a correlação entre as duas variáveis. Incluia a variével DespesaPesquisa.

TABELA 2.14: Interpretação da correlação

Faixa	Interprecao
0.00–0.19	muito fraca
0.20–0.39	fraca
0.40–0.69	moderada
0.70–0.89	forte
0.90–1.00	muito forte

```
base_ies %>%
 dplyr::select(TotalTecnicos, ReceitaPropria, DespesaPesquisa) %>%
 cor(method = "pearson") %>%
 round(digits = 3) %>%
 knitr::kable(booktabs = TRUE, format = tb_formata,
       caption = "Análise de correlação") %>%
 kableExtra::kable_styling(latex_options = "hold_position")
```

TABELA 2.15: Análise de correlação

	TotalTecnicos	ReceitaPropria	DespesaPesquisa
TotalTecnicos	1.000	0.058	0.292
ReceitaPropria	0.058	1.000	0.064
DespesaPesquisa	0.292	0.064	1.000

Note agora que faz mais sentido estimar a força da relação linear entre as variáveis utilizando a correlação. No nosso exemplo, uma correlação linear de 0.047 entre TotalTecnicos e ReceitaPropria nos diz que a relação linear entre estas duas variáveis é muito fraca. O mesmo se aplica à variável DespesaPesquisa em relação à variável TotalTecnicos com 0.190. Estas três variáveis não possuem correlação forte entre si.

Bibliografia

Barnett, V. and Lewis, T. (1974). *Outliers in statistical data*. Wiley.

De Mauro, A., Greco, M., and Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review*, 65(3):122–135.

Gil, A. C. (2008). *Métodos e técnicas de pesquisa social*. 6. ed. Ediitora Atlas SA.

Köche, J. C. (2016). *Fundamentos de metodologia científica*. Editora Vozes.

Komsta, L. (2011). *outliers: Tests for outliers*. R package version 0.14.

Montgomery, D. C. (2017). *Design and analysis of experiments*. John wiley & sons.

Reichmann, W. (1961). Use and abuse of statistics: Methuen. *1961*.

Tufte, E. and Graves-Morris, P. (2014). The visual display of quantitative information.; 1983.

Wickham, H. (2017). Tidyverse: Easily install and load'tidyverse'packages. *R package version*, 1(1).

Wickham, H. et al. (2014). Tidy data. *Journal of Statistical Software*, 59(10):1–23.