O manual do Cientista de Dados

Wlademir Prates

2025-01-08

Table of contents

Prefácio

Seja Bem vindo (a) ao livro de Data Science da Mentoria de Ciência de Dados da Data Mundo.

To learn more about Quarto books visit https://quarto.org/docs/books.

1 Introduction

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

```
1 + 1
```

```
[1] 2
```

2 O que é Ciência de Dados (Data Science)?

Ciência de dados é um campo que tem se destacado pela sua **capacidade de transformar dados em resultados para as organizações**, utilizando técnicas e ferramentas que vão além das tradicionais planilhas eletrônicas e dos *dashboards* descritivos de *business intelligence* (BI). Veja neste artigo o que é ciência de dados, bem como exemplos práticos de aplicação da área e também uma descrição acerca de quem é e o que se espera do profissional desta área, o cientista de dados.

2.1 O que é ciência de dados (*data science*)?

Para auxiliar na definição do termo, vamos ver o que alguns autores e sites especializados dizem a respeito de ciência de dados:

- "A ciência de dados é uma disciplina multifacetada, que abrange [aprendizado de máquina] e outros processos analíticos, estatísticas e ramos relacionados da matemática. Cada vez mais se utiliza da computação científica de alto desempenho, tudo isso para extrair informações dos dados e usar essas informações encontradas para contar histórias." (Matthew Mayo, KDnuggets).
- "*Data science* envolve princípios, processos e técnicas para compreender fenômenos por meio da análise (automatizada) de dados." (PROVOST e FAWCETT, p. 4, 2016).
- "A ciência de dados é um conjunto multidisciplinar de inferência de dados, desenvolvimento de algoritmos e tecnologia para resolver problemas analiticamente complexos. No centro estão os dados: um grande número de informações brutas, transmitidas e armazenadas em data warehouses corporativos. [...] A ciência de dados é basicamente o uso desses dados de maneiras criativas para gerar valor aos negócios." (Divya Singh, Data Science Central).

Definição que acredito:

> Ciência de dados é uma **área multidisciplinar** que se utiliza principalmente, mas não apenas, de **método científico, estatística, conhecimento de negócio e ciência da computação** para gerar valor aos negócios.

A forma como cada uma das subáreas contribui para o campo de ciência de dados é, de maneira geral, a seguinte:

- **Método científico**: auxilia na estrutura do projeto de *data science*, que contempla a definição de um problema de negócio (análogo ao problema de pesquisa dos trabalhos acadêmicos); definição de objetivos geral e específicos; discussão e apresentação de resultados; conclusões e procedimentos futuros.
- **Estatística**: formas de resumir e visualizar dados; testes de hipóteses; técnicas de análise preditiva.
- **Negócio**: necessário para ser capaz de definir hipóteses de negócio a serem posteriormente transformadas em hipóteses estatísticas; fundamental para escolher o **problema de negócio de fato mais relevante** para se investir tempo e dinheiro com análise de dados.
- **Ciência da computação**: contribui com métodos que impulsionam as técnicas estatíticas (*machine learning*), utilizando poder computacional, linguagens de programação, computação na nuvem, bancos de dados, entre outros.

Um dos pontos principais para um bom funcionamento de um projeto de ciência de dados na prática é um perfeito alinhamento entre a equipe técnica (os cientistas de dados) e a área de negócio do cliente (interno ou externo). É comum que sejam realizadas sessões de *Design Thinking* e utilizadas adaptações do *Business Canvas* para identificar o problema de negócio e também gerar *insights* relevantes à equipe técnica, o que deverá conduzir a entregas de melhor qualidade.

2.2 Surgimento da ciência de dados

A área de ciência de dados é bastante nova, tendo sido assim chamada pela primeira vez a princípio em 2001. Porém, foi a partir de aproximadamente o ano de 2010 que a área começou a ganhar notoriedade, principalmente devido à onda de *big data*.

A razão para as empresas passarem a gerar e armazenar grandes volumes de dados (*big data*) se deu com o avanço da tecnologia a partir da bolha "ponto com", e principalmente com o barateamento dos discos rígidos.

A partir disso, grandes companhias como Google e Amazon desenvolveram novas arquiteturas computacionais, que chamamos hoje de computação nas nuvens (*cloud computing*).

Com todo este cenário ficou fácil e barato para as empresas armazenarem diversos tipos de dados, muitos que até então eram ignorados.

A grande pergunta que surgiu foi **"o que fazer com todos estes dados, e como gerar valor de negócio a partir deles?"**.

Sendo assim, o termo *big data* saiu um pouco de enfoque, dando lugar à multidisciplinariedade da **ciência de dados**, em que *big data* é apenas uma parte de um todo.

2.3 Perfil e papel do cientista de dados

O objetivo aqui é de trazer uma base do perfil e também do papel do cientista de dados nas organizações. Claro que dependendo da companhia a opinião pode divergir em alguns aspectos. Por isso, meu objetivo aqui não é focar em habilidades técnicas exigidas, nem tampouco em linguagens de programação ou tecnologias.

Inicialmente, apresento uma citação de Provost e Fawcett (2016, p. 333) que resume muito bem o processo de ciência de dados no ponto de vista do cientista de dados:

> A prática de *data science* pode ser melhor descrita como uma combinação de **engenharia analítica e exploração**. O negócio apresenta um problema que gostaríamos de resolver. Raramente, o problema de negócio é, de modo direto, uma de nossas tarefas básicas de mineração de dados. **Decompomos o problema em subtarefas** que achamos que podemos resolver, geralmente, começando com as ferramentas existentes. Para algumas dessas tarefas podemos não saber o quão bem podemos resolvê-las, por isso, temos que explorar os dados e fazer uma avaliação para verificar. Se isso não for possível, **poderemos ter de tentar algo completamente diferente**. No processo, **podemos descobrir o conhecimento** que vai nos ajudar a resolver o problema que queremos ou podemos descobrir algo inesperado que nos leva a outros sucessos importantes.

Este parágrafo acima é excelente, pois resume algumas das principais capacidades que um cientista de dados deve ter:

- Criatividade.
- Capacidade de explorar possibilidades de soluções de problemas até então desconhecidas.
- Pensamento crítico para ser capaz de extrair conclusões importantes, resultantes dos processos de análise de dados, mas que não faziam parte diretamente da questão inicial levantada. Claro que sobre este ponto vale ressaltar que o cientista de dados precisa ter cuidado para não perder o foco do projeto. É preciso ter bom senso.

Um dos pontos que me chama atenção no perfil do profissional cientista de dados é que ter conhecimento do método científico contribui muito para a execução prática dos projetos de ciência de dados. Isto faz com que a área de *data science* seja capaz de interligar a "teoria" do mundo acadêmico com a "prática" do mundo dos negócios.

Na minha opinião, sempre achei que a academia e o meio corporativo têm muito a aprender um com o outro. O campo de ciência de dados é um exemplo de sucesso desta mescla de abordagens.

2.4 Ferramentas do cientista de dados

Não há necessariamente um conjunto de ferramentas padrão para trabalhar com ciência de dados. Porém, devido à característica do trabalho, alguns pontos importantes são:

- Ter conhecimento de alguma linguagem de programação com alto poder de aplicação analítica. Hoje em dia se destacam as linguagens R e Python.
- Ter conhecimentos intermediários em métodos estatísticos.
- Capacidade didática de explicar os resultados encontrados, principalmente de forma escrita.
- Ter conhecimentos básicos de computação na nuvem.
- Saber trabalhar com versionamento de códigos (basicamente Git).
- Outros conhecimentos são importantes, mas que talvez não sejam tão essenciais quanto os demais, que são: bancos de dados, html, javascript.

Como já mencionado, o trabalho do cientista de dados é muito versátil e dinâmico. Por isso, quanto mais conhecimentos o indivíduo tiver melhor, mas as linguagens R e Python são tão poderosas para fins analíticos e tão integradoras de outras tecnologias que geralmente não é necessário ter profundos conhecimentos além delas para executar bons projetos de ciência de dados.

2.5 Aplicações: ciência de dados para negócios

As aplicações são diversas, mas a seguir tento exemplificar algumas possibilidades de aplicação de ciência de dados na solução prática de problemas de negócio reais:

2.5.1 Recursos humanos

- **Turnover**: identificação das causas principais que levam um colaborador a pedir demissão da empresa, bem como aplicação de modelo preditivo para gerar uma lista com os colaboradores com maior probabilidade de pedirem para sair.
- **Recrutamento**: identificar os perfis de candidatos (internos ou externos) a vagas dentro da empresa que melhor se enquadram, utilizando dados de perfil, CV, experiências anteriores, aspectos demográficos, entre outros.

2.5.2 Logística

- **Falhas em entregas**: identificação, por meio de modelo preditivo, dos casos em que há maior probabilidade de uma entrega não ser efetivida.

2.5.3 Finanças

- **Gestão de carteiras**: identificação dos ativos com maior probabilidade de apresentarem bom desempenho no futuro com base em diversos dados históricos. É possível mesclar dados contábeis/fundamentalistas com indicadores técnicos e também variáveis categóricas, como setor ou níveis de governança corporativa, por exemplo.

2.5.4 Marketing

- **Identificação de *leads***: utilizar modelo preditivo para encontrar *leads* com maior probabilidade de se tornarem clientes.
- **Redução de *churn*** (clientes que cancelam assinaturas): abordagem muito próxima a utilizada no caso de *turnover* (colaboradores que pedem demissão).

2.6 Ciclo de Vida de um Projeto de Ciência de Dados

A ciência de dados se tornou um componente crucial em muitos setores, fornecendo insights valiosos e informando decisões baseadas em dados. Compreender as fases do ciclo de vida de um projeto de ciência de dados é essencial para obter os melhores resultados.

Um projeto de ciência de dados é uma série de etapas inter-relacionadas. Cada etapa desempenha um papel importante na obtenção de insights significativos dos dados. Vamos entender melhor cada uma dessas fases.

2.6.1 Definição do Problema

A fase inicial de qualquer projeto de ciência de dados é a **definição do problema**. É aqui que identificamos a questão ou desafio que queremos resolver. Esta fase define o rumo do projeto, garantindo que as soluções geradas sejam relevantes e impactantes.

Uma clara definição do problema serve como guia para as fases subsequentes do projeto, incluindo a coleta e análise de dados. Garantir que o problema esteja bem definido desde o início é vital para o sucesso de qualquer projeto de ciência de dados.

2.6.2 Coleta de Dados

Depois de definir claramente o problema, o próximo passo é a **coleta de dados**. Os dados podem ser obtidos de diversas fontes, como bancos de dados internos, APIs da web ou fontes de terceiros. A escolha das fontes de dados depende da natureza do problema que estamos tentando resolver.

Os dados coletados formam a base do projeto. Portanto, é essencial garantir que os dados sejam relevantes para o problema e de alta qualidade. Dados de baixa qualidade ou irrelevantes podem levar a insights imprecisos e soluções ineficazes.

2.6.3 Preparação de Dados

Uma vez que os dados foram coletados, eles precisam ser preparados para análise. A **preparação de dados** inclui a limpeza dos dados, como tratar valores ausentes e remover outliers. Mas também abrange a transformação de dados e a criação de novas variáveis, que são partes fundamentais da preparação dos dados.

Essa etapa é vital, pois a qualidade dos dados afeta a qualidade dos insights e soluções geradas. Sem um adequado preparo dos dados, corremos o risco de tirar conclusões erradas e propor soluções que não resolvam efetivamente o problema.

2.6.4 Modelagem de Dados

A **modelagem de dados** é onde aplicamos técnicas e algoritmos de aprendizado de máquina aos nossos dados preparados. A escolha do modelo a ser usado depende do problema que estamos tentando resolver. Podemos empregar desde modelos mais simples, como regressões, até abordagens mais complexas, como redes neurais.

Esta fase é a essência da ciência de dados, onde os dados são transformados em insights valiosos. Um bom modelo pode extrair informações significativas dos dados, proporcionando soluções eficazes para o problema definido.

2.6.5 Avaliação do Modelo de Dados

A **avaliação do modelo** é a última fase do ciclo de vida de um projeto de ciência de dados. Aqui, testamos o desempenho do nosso modelo. Verificamos se o modelo é capaz de fornecer insights precisos e úteis para o problema.

As métricas de avaliação variam dependendo do problema e do tipo de modelo usado. O objetivo é garantir que o modelo seja não só preciso, mas também relevante e útil para resolver o problema que foi definido no início do projeto.

Cada projeto de ciência de dados é único e pode exigir abordagens diferentes. No entanto, as fases básicas descritas aqui proporcionam uma estrutura sólida que pode ser adaptada conforme necessário. Dominar essas fases será um trampolim para qualquer projeto de ciência de dados bem-sucedido.

2.7 Metodologia Ágil em Ciência de Dados

A aplicação de metodologias ágeis em projetos de ciência de dados tem se mostrado extremamente eficaz, especialmente considerando a natureza iterativa e experimental destes projetos. O Scrum, em particular, tem sido amplamente adotado, com algumas adaptações específicas para projetos de dados.

2.7.1 Scrum em Projetos de Data Science

O Scrum em ciência de dados mantém seus princípios fundamentais, mas adapta-se às particularidades da área:

- **Sprints**: Geralmente de 2 a 4 semanas, com objetivos específicos como:
  - Sprint 1: Definição do problema e coleta inicial de dados
  - Sprint 2: Limpeza e preparação dos dados
  - Sprint 3: Desenvolvimento do primeiro modelo (MVP)
  - Sprints subsequentes: Iterações e melhorias do modelo

2.7.2 Cerimônias Essenciais

- **Daily Scrum**: Reuniões diárias de 15 minutos onde a equipe discute:
  - Progresso na preparação dos dados
  - Resultados preliminares dos modelos
  - Bloqueios técnicos encontrados
- **Sprint Planning**: Define os objetivos da sprint, como:
  - Métricas a serem alcançadas
  - Quantidade de dados a ser processada
  - Features a serem desenvolvidas
- **Sprint Review**: Apresentação dos resultados para stakeholders:
  - Demonstração dos modelos desenvolvidos
  - Apresentação de métricas alcançadas
  - Visualizações de dados relevantes

- **Sprint Retrospective**: Reflexão sobre o processo:
  - O que funcionou bem no desenvolvimento dos modelos
  - Desafios na coleta ou processamento de dados
  - Ajustes necessários para a próxima sprint

2.7.3 Artefatos Adaptados

- **Product Backlog**: Lista priorizada incluindo:
  - Features do modelo a serem desenvolvidas
  - Conjuntos de dados a serem incorporados
  - Métricas de performance a serem alcançadas
- **Sprint Backlog**: Tarefas específicas como:
  - Limpeza de determinado conjunto de dados
  - Implementação de algoritmos específicos
  - Desenvolvimento de visualizações
- **Quadro Kanban**: Adaptado com colunas como:
  - Coleta de Dados
  - Preparação
  - Modelagem
  - Validação
  - Produção

2.7.4 Papéis Principais

- **Product Owner**: Foca em:
  - Definição clara dos objetivos de negócio
  - Priorização de features do modelo
  - Validação dos resultados do ponto de vista do negócio
- **Scrum Master**: Auxilia removendo impedimentos como:
  - Acesso a dados necessários
  - Recursos computacionais adequados
  - Comunicação com áreas de negócio
- **Time de Data Science**: Composto por:
  - Cientistas de dados
  - Engenheiros de dados

– Analistas de negócio

Esta estrutura ágil permite que projetos de ciência de dados mantenham o foco na entrega de valor, enquanto permanecem flexíveis para incorporar novos insights e requisitos que surgem durante o desenvolvimento dos modelos.

2.8 Conclusões

A área de ciência de dados muito se desenvolveu, e hoje as empresas em geral já vêem valor tanto em contratar projetos quanto em construir áreas de *data science*. A maior parte das empresas grandes, na verdade, já possuem áreas de ciência de dados constituídas. Contudo, ainda há muito que se consolidar em termos metodológicos e também de quais são os tipos de entregas mais adequadas.

Vale lembrar, tanto para gestores das áreas de ciência de dados, quanto para cientistas de dados, que o papel mais importante de um projeto de *data science* não está no modelo de *machine learning* utilizado, mas sim na capacidade de impactar positivamente a organização em algum KPI de negócio.

2.9 Referências

PROVOST, F., FAWCETT, T. (2016). Data science para negócios: o que você precisa saber sobre mineração de dados e pensamento analítico de dados. Rio de Janeiro: Alta Books.

3 Análise de Hipóteses

Esta seção apresenta detalhes sobre os testes estatísticos aplicados para avaliar hipóteses de negócio levantadas a partir de interação com o cliente.

A estatística possui uma área chamada de **teste de hipóteses**, que viabiliza verificar ou refutar uma hipótese existente acerca do comportamento dos dados. Os testes de hipóteses são muito utilizados no **meio acadêmico**, mas também possuem grande valor se usados de forma adequada no **contexto corporativo**.

Nas empresas existem diversas crenças acerca de como as coisas acontecem e se relacionam. Formalizar estas crenças em **hipóteses de negócio** permite que sejam formuladas **hipóteses estatísticas**, as quais podem ser testadas e validadas (ou não) com o uso de testes de hipóteses.

Porém, para que os resultados dos testes sejam corretos, é preciso conhecer os principais tipos de testes existentes e alguns dos seus conceitos. Assim é possível direcionar para a aplicação correta a ser utilizado em cada hipótese.

Basicamente os **passos para uma boa execução** de testes de hipóteses são:

1. Definir uma **hipótese de negócio**;
2. Transformá-la em uma **hipótese estatística**;
3. **Escolher o teste correto** baseado nas características dos dados e no resultado esperado;
4. **Transformar os dados** para que estejam de acordo com o *input* exigido pelo teste escolhido;
5. **Aplicar** o teste de hipótese.

Por que Utilizar um Teste de Hipótese?

Quando as tabelas dinâmicas do Excel já não são suficientes para que os resultados das análises sejam conclusivas é um bom indicativo de que já passou da hora de utilizar testes de hipóteses.

Um teste de hipótese auxilia a eliminar a incerteza que permanece mesmo após procedimentos de sumarização dos dados, como ocorre nas tabelas dinâmicas dos *softwares* de planilha eletrônica.

Por exemplo, uma organização quer saber se os homens possuem salário superior às mulheres em um determinado cargo. A tabela dinâmica vai trazer o resultado, que pode ser expresso através da média dos salários daquele cargo em cada gênero (masculino ou feminino). Porém, o resultado da média pode mostrar apenas uma diferença pequena entre os salários de homens e mulheres, o que acaba deixando no ar a dúvida inicial e nenhuma conclusão pode ser tomada.

Um teste de hipótese estatística é capaz de dizer (quando bem aplicado) que o salário das mulheres é de fato menor ou maior naquele caso, e que o resultado mostrado na média não é uma simples obra do acaso.

3.1 Conceitos Importantes sobre os Testes de Hipóteses

Uma hipótese estatística, formalmente, é uma afirmação sobre alguma característica da população. Um teste de hipótese, por sua vez, **é um procedimento estatístico para dizer se uma afirmação sobre a população é verdadeira**.

Se a probabilidade de ocorrência de um evento atrelado à hipótese for baixa, então a hipótese é assumida como não verdadeira.

Hipótese Nula e Hipótese Alternativa

Todo teste de hipóteses precisa de uma **hipótese nula** (H_0) e uma **hipótese alternativa** (H_1).

A H_0 é uma afirmação que **sempre representará uma igualdade**. Por exemplo: "a média salarial de homens **é igual** a média salarial de mulheres"; "colaboradores com maior grade **possuem o mesmo** tempo de empresa que colaboradores com menor grade". Veja que a H_0 pode ser dividia em duas declarações, que chamaremos de X e Y. Nos exemplos citados, as declarações que representam X seriam "média salarial de homens" e "colaboradores com maior grade"; enquanto Y seriam "média salarial de mulheres" e "colaboradores com menor grade".

Já quando olhamos para H_1, temos três possibilidades de configurações: existe diferença, é maior ou é menor. Olhando para as declarações de X e Y, poderíamos ter como hipótese alternativa à H_0, citada acima, que "a média salarial de homens **é maior** que a média salarial de mulheres". Ou ainda, quando não temos um palpite definido (se maior ou menor), podemos definir H_1 como "colaboradores com maior grade possuem tempo de empresa **diferentes** que colaboradores com menor grade".

Resumindo, para definirmos as hipóteses nula e alternativa teremos:

$H_0 : X = Y$

$H_1 : X \neq Y$ ou $X < Y$ ou $X > Y$

Qual a Decisão em um Teste de Hipótese?

Em um teste de hipóteses a decisão será sempre **rejeitar a hipótese nula** ou **não rejeitar a hipótese nula**.

Para isso, os testes geram um *p-valor* para representar a **probabilidade de significância** estatística (métrica que vai de 0 a 1). **Quanto mais próximo de 0 o *p-valor*, mais significativa é a diferença testada**.

> Um **baixo p-valor** indica **forte evidência** contra a hipótese nula, o que leva à **rejeição** de H_0

De forma geral, recomenda-se **rejeitar** H_0 com ***p-valor* menor que** $0,05$ (ou 5%). Em alguns casos específicos, o nível de significância **pode ser de** 10%.

Testes Paramétricos e não Paramétricos?

Testes paramétricos assumem que os dados são distribuídos aleatoriamente a partir da população e que seguem uma **distribuição normal**.

Testes não paramétricos também assumem que os dados são aleatoriamente distribuídos a partir da população, mas **não exigem que sigam uma distribuição normal**.

Os testes não paramétricos, além de não exigirem "normalidade" na distribuição dos dados, também apresentam resultados melhores quando aplicados em **amostras pequenas**. Por esses motivos, em casos reais normalmente são os preferidos e mais adequados em testes de hipóteses.

A questão que permanece é **como saber se o teste a ser escolhido é paramétrico ou não paramétrico**?

A resposta é simples: se o dado a ser testado segue uma distribuição normal, então recomenda-se utilizar testes paramétricos; caso o dado não possua distribuição normal então utiliza-se testes não paramétricos.

Quando o dado é normalmente distribuído então alguns testes podem trazer inferências sobre intervalos, pois já se conhece a distribuição. Já os testes não paramétricos não se baseiam na distribuição dos dados.

Testar a normalidade de uma série de dado é bastante simples pela linguagem R. Uma das possibilidades é utilizar o teste Shapiro-Wilk por meio da função `shapiro.test()`. O teste possui hipótese nula (H_0) de que o dado é normalmente distribuído, a qual recomenda-se ser rejeitada a um p-valor menor que 0,05. Em outras palavras, se o p-valor for maior que 0,05 então assumimos que o dado é normalmente distribuído e seguimos com testes paramétricos; caso contrário procuramos um teste adequado entre os não paramétricos.

Testes Pareados e não Pareados

Em testes para **dados pareados** as amostras são **dependentes**. Aplicam-se, por exemplo, no caso das duas métricas que serão comparadas serem obtidas a partir do **mesmo indivíduo**, antes e após um tratamento.

Exemplo de teste pareado

Após um colaborador receber movimentação salarial por mérito, sua produtividade melhora.

Para validar esta hipótese um teste pareado sobre as observações de produtividade deveria ser executado com duas amostras do mesmo colaborador: uma antes e outra depois da movimentação salarial por mérito.

Em testes para **dados não pareados** os dados são coletados de **indivíduos distintos** e que pertencem a grupos também distintos. As amostras a serem testadas são **independentes**.

Exemplo de teste não pareado

As avaliações de lideranças na área de TI são inferiores às avaliações na área de Recursos Humanos .

Para validar esta hipótese é necessário um teste não pareado sobre as notas dos líderes em cada área.

Vale ainda ressaltar que, em testes para dados pareados, obrigatoriamente o tamanho das amostras deve ser igual (afinal, as amostras devem ser "pares"). Já em testes não pareados os tamanhos das amostras podem ser diferentes.

3.2 Tipos de Testes

Existem diversos testes de hipóteses, sendo que cada um é mais adequado para uma situação específica. A seguir são apresentados alguns testes (não todos, pois existem diversos) que servem para a maior parte das situações que envolvem testes de hipóteses.

3.2.1 Testes de Proporções

Os testes de proporções são adequados quando se têm **variáveis binárias ou categóricas** (ou numéricas divididas em faixas, como renda, idade ou número de funcionários), e se deseja saber se determinada característica é mais ou menos presente em um certo tratamento. Alguns exemplos de hipóteses alternativas a serem verificadas com testes de proporções:

- O *turnover* voluntário é maior em colaboradores do gênero masculino do que feminino;
- Há um maior índice de *turnover* voluntário em colaboradores cuja frequência de viagem a trabalho é maior.

Veja, *turnover* é uma variável binária, que indica quando um colaborador é desligado ou não, e o que normalmente se deseja testar com esta variável são casos em que há maior ou menor índice de *turnover*. Este é um tipo de situação em que é adequado aplicar testes de proporções.

Teste Z

É um teste **paramétrico** utilizado para comparar diferenças de proporções entre duas amostras independentes. É idêntico ao teste Qui-Quadrado para diferença de proporções (apresentado a seguir), exceto que este permite estimar o desvio-padrão pela distribuição normal. Um cuidado que deve-se tomar com este teste é relacionado a sua aplicação em amostras que não são independentes. A equação implementada na linguagem R para teste de proporções não contempla o teste Z, apenas o Qui-Quadrado.

Usos e mau usos do Teste Z

Teste Qui-Quadrado (X^2)

É uma alternativa **não paramétrica** ao teste Z. O teste *Chi-Squared* para proporções é um dos testes estatísticos mais utilizados. É mais adequado para amostras pequenas que o teste Z. Um dos seus principais usos incorretos está atrelado também a não independência entre as amostras.

Usos e mau usos do Person's Chi-Squared Test

Na linguagem R há uma função nativa para este teste, `prop.test()`, que necessita como input (i) a quantidade de ocorrências para cada evento e (ii) o total de casos. Assim, o próprio teste calcula as proporções. Além disso, é possível aplicar o teste também para verificar se há diferença entre **mais de duas amostras**, e também verificar se há **tendência nas proporções** entre os grupos, por meio da função `prop.trend.test()`.

Teste de Fisher

O teste exato de Fisher é um teste não paramétrico que tem o objetivo de testar a independência de duas ou mais variáveis categóricas. Ele é uma alternativa ao teste Qui-Quadrado e normalmente é utilizado para a tabela de contingência 2×2, ou quando as frequências esperadas de uma das células da tabela de contingência são menores do que 5.

A lógica do teste de Fisher é a mesma apresentada no teste Qui-Quadrado: identificar se as variáveis categóricas são independentes (o `H0`, hipótese nula do teste) ou se exsite alguma relação entre elas (o `H1`, hipótese alternativa do teste).

Na linguagem R, o teste de Fisher pode ser facilmente implementado utilizando a função `fisher.test`, que necessita como input uma tabela de continguência ou a especificação de

duas variáveis categorias de uma base de dados, que serão então utilizadas pelo teste para identificar a existência ou não de associações entre elas.

3.2.2 Testes para Diferenças com uma Amostra (*one sample*)

São testes aplicados para **variáveis contínuas** em caso de se ter apenas uma amostra de dados e desejar testar se há diferença desta amostra contra parâmetros hipotéticos. Por exemplo:

H_0: os colaboradores recém promovidos recebem em torno de 100% da faixa.

H_1: os colaboradores recém promovidos recebem mais de 100% da faixa.

Neste caso, não se compara duas ou mais amostras, mas realiza-se o teste com base em um valor que seria esperado por alguma razão, por exemplo, por uma política de recursos humanos da organização.

Teste T One Sample

O teste T para uma amostra é um teste **paramétrico** que permite verificar se a média de uma série de dados é diferente de uma média hipotética que se deseja testar. Para implementar o teste em R, utiliza-se `t.test(x, mu = 0)`, em que `x` representa o vetor com a amostra a ser testada, e `mu` é o parâmetro para definir a média esperada.

Wilcoxon Signed Rank

É um teste **não paramétrico** para uma amostra contra uma mediana hipotética. Foi proposto no mesmo artigo que o teste Wilcoxon Rank Sum, aplicável para duas amostras. Na linguagem R, utiliza-se a função `wilcox.test()`.

3.2.3 Testes para Diferenças entre Dois Grupos (*two sample*)

Quando se têm **variáveis contínuas** e se deseja verificar uma possível diferença entre duas amostras utiliza-se um teste para verificar se há diferenças entre as distribuições.

Teste T Two Sample

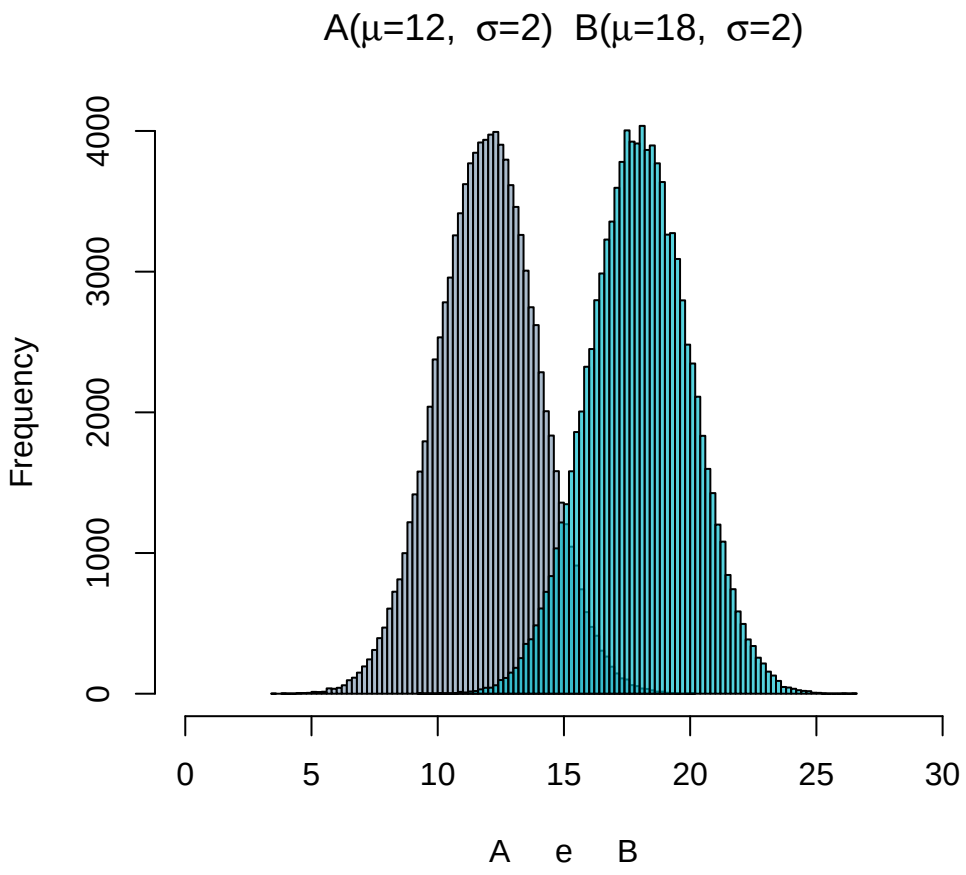
É um teste **paramétrico**, capaz de verificar se duas populações possuem **médias** iguais para uma determinada variável. A versão do teste implementada na linguagem R (Welch's t-test) é mais confiável do que o tradicional Teste t de Student quando as amostras não possuem a mesma variância e/ou o tamanho das amostras é desigual.

O gráfico ao lado é um exemplo em que pode se aplicar o seguinte teste:

H_0 : A = B.

H_1 : A < B.

O gráfico foi construído com uma amostra aleatória que segue a distribuição normal. As médias (μ) e desvios-padrão (σ) simuladas estão expressas no gráfico. De fato H_0 é rejeitada pelo teste T a um p-valor < 0,05 neste exemplo.



Para implementar o teste em R, utiliza-se `t.test(x,y)`, em que `x` e `y` representam vetores com as duas amostras a serem testadas. Para teste pareado, basta utilizar o parâmetro `paired = TRUE` dentro da função.

Wilcoxon Rank Sum

Este é um teste que serve como alternativa **não paramétrica** ao teste T para duas amostras, sendo também chamado de **Teste U de Mann-Whitney**. Muito utilizado para testar diferenças entre duas amostras, pois não é preciso cumprir a premissa de normalidade. Essa característica faz com que este seja um teste mais abrangente que o teste T, servindo para variadas situações do dia a dia.

É um teste baseado apenas na ordem em que as observações das duas amostras aparecem. Um caso de uso interessante ocorre **quando a amostra é pequena** demais a ponto de não ser possível dizer se a distribuição é normal ou não.

O teste Rank Sum de Wilcoxon baseia-se na classificação das observações das duas amostras sendo testadas. À cada observação é atribuída uma classificação, sendo que a menor tem classificação 1, a segunda menor classificação 2, e assim por diante. A estatística de teste é calculada com base na soma das classificações de cada uma das amostras. Dessa forma o teste consegue dizer se a soma dos *rankings* associados a uma amostra é menor, igual ou maior que da outra, apontando se há diferenças nas amostras e também o sentido desta diferença. O resultado é similar ao do teste T, mas por utilizar um sistema de *ranking* não se presume nada acerca de como o dado é distribuído.

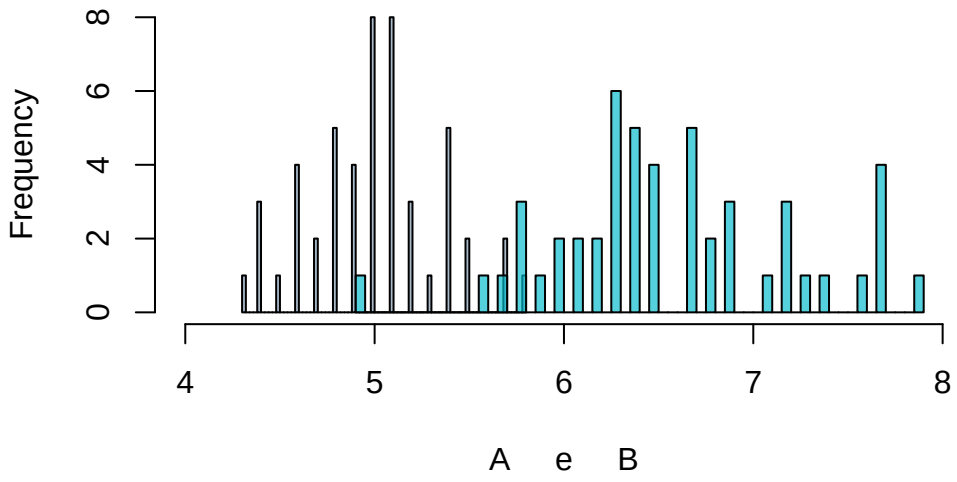
Para aplicar o teste na linguagem R utiliza-se a função `wilcox.test(x, y)`, em que `x` e `y` representam vetores com as duas amostras a serem testadas. Para teste pareado basta utilizar o parâmetro `paired = TRUE`.

O gráfico ao lado é um exemplo em que pode se aplicar o seguinte teste:

H_0 : A = B.

H_1 : A < B.

O gráfico foi construído com duas amostras de 50 observações cada, com dados que não possuem nitidamente o formato de sino da distribuição normal, apesar de o teste de normalidade não ter fornecido indícios de que o dado não segue uma distribuição normal. Pelo teste Wilcoxon H_0 é rejeitada a um p-valor $< 0{,}05$.



- Veja mais sobre o Wilcoxon Rank Sum.
- Artigo comparativo entre teste T e Wilcoxon Rank Sum

3.2.4 Testes para Diferenças entre mais de Dois Grupos

ANOVA

Análise de Variância (ANOVA, do inglês *Analisys of Variance*), compreende uma família de testes que permitem verificar diferenças entre séries de dados. A ideia central da ANOVA é testar se há **diferença entre as médias** (a ANOVA pertence aos **testes paramétricos**) das amostras analisadas, permitindo trabalhar com **3 ou mais amostras**. A Análise de Variância é muito utilizada em ambientes experimentais, para verificar a existência de diferenças entre variados tratamentos aplicados em grupos distintos.

É, portanto, uma alternativa paramétrica ao teste T, para os casos com mais de duas amostras. Porém, por ser paramétrico possui seu uso restrito.

Kruskal-Wallis

Este é um teste **não paramétrico** que serve como alternativa à ANOVA, caso os critérios de normalidade e homoscedasticidade (igual variância entre as amostras) não sejam cumpridos. Kruskal-Wallis é uma extensão do teste Wilcoxon Rank Sum, sendo aplicável para casos com

mais de duas amostras a serem comparadas. O resultado do teste indica se há diferença entre pelo menos duas das amostras testadas.

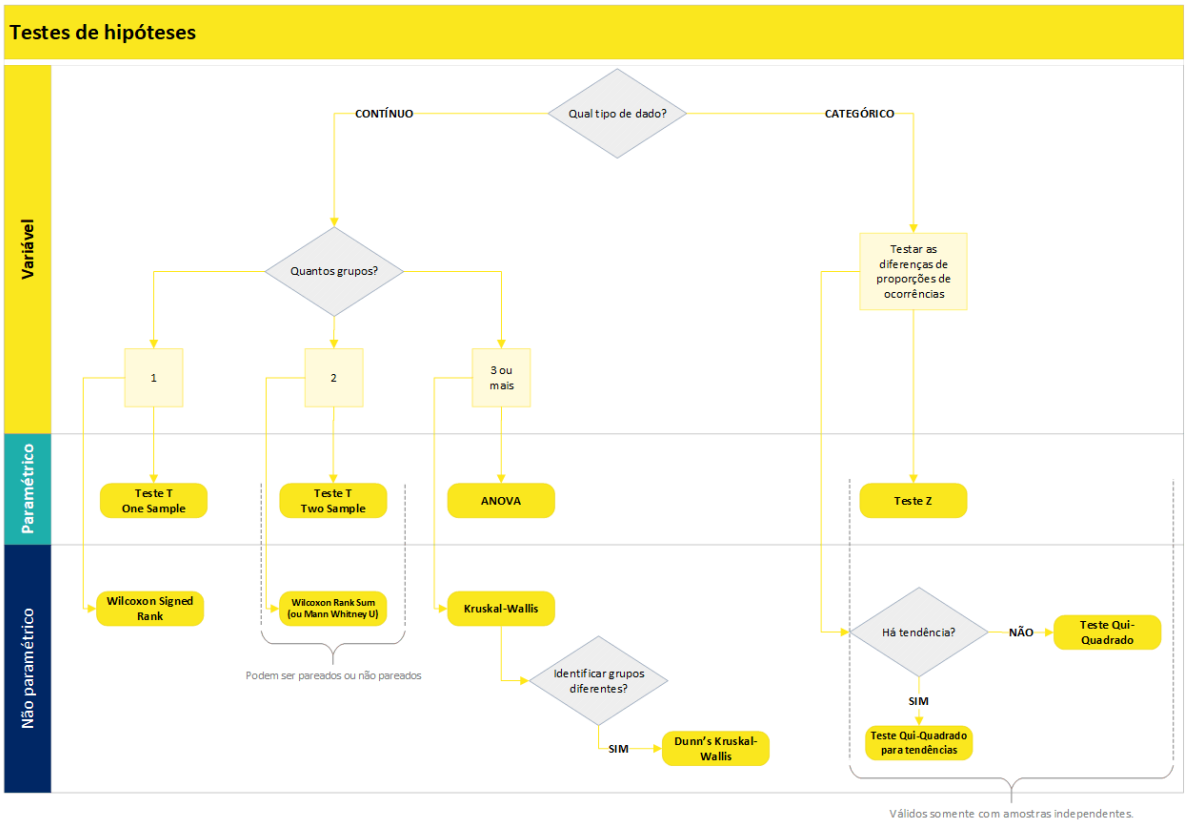
O teste pode ser executado com amostras extremamente pequenas (a partir de 6 observações, sendo pelo menos duas para cada grupo). Também é um teste aplicável para amostras não balanceadas (tamanhos diferentes).

Kruskal-Wallis Dunn Test

Quando a hipótese nula do teste Kruskal-Wallis é rejeitada, então sabe-se que pelo menos duas amostras são distintas (das 3 ou mais que foram testadas). Porém, na maioria dos casos ainda se deseja saber mais. A conclusão que de fato vai agregar valor é quais as combinações de grupos que diferem, e quais são os sinais destas diferenças. É para esta situação que utiliza-se o Dunn Test.

3.3 Qual teste utilizar em cada caso?

Dado que existem diversos testes de hipótese e que cada um é mais adequado para um determinado tipo de situação, a seguir é apresentado um fluxograma que auxilia na escolha do teste.



Por fim, a tabela a seguir sintetiza o que foi abordado neste documento acerca de testes de hipóteses. Existem diversos outros testes, porém, os que são apresentados na sequência viabilizam a verificação estatística da maior parte das hipóteses de negócio levantas em projetos de análise de dados.

Teste	Testa diferenças entre:	Paramétrico ou não?	Qtde de amostras que compara	Função no R / pacote	Observações
Z	Proporções	Paramétrico	2	Necessário criar	
Chi-Squared	Proporções	Não paramétrico	2	prop.test() / Default	
T one sample	Médias	Paramétrico	2	t.test() / Default	Pareado
T two sample	Médias	Paramétrico	2	t.test() / Default	Não pareado

Teste	Testa diferenças entre:	Paramétrico ou não?	Qtde de amostras que compara	Função no R / pacote	Observações
Wilcoxon one sample	Distribuições	Não paramétrico		wilcox.test() / Default	Pareado, análogo ao teste T one sample.
Wilcoxon rank sum test (ou Mann Whitney U Test)	Distribuições	Não paramétrico	2	wilcox.test() / Default	Não pareado, análogo ao teste T two sample.
Kruskal-Wallis	Distribuições	Não paramétrico	3 ou mais	kruskal.test / Default	Não pareado. Extensão do Wilcoxon rank sum. Se pelo menos dois grupos apresentarem diferença, então o p-valor é significativo.
Dunn's Kruskal-Wallis	Distribuições	Não paramétrico	Testa todas as combinações em pares	dunnTest() / FSA	Quando o Kruskal-Wallis é significativo, o Dunn's test verifica em quais combinações há diferença.

References

Knuth, Donald E. 1984. "Literate Programming." *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.