



**CEFET/RJ - CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA CELSO
SUCKOW DA FONSECA**
Campus Nova Friburgo
Bacharelado em Sistemas de Informação
5º Período

Gestão do Conhecimento e da Informação

Compilação de Materiais

Projeto de Banco de Dados e Inteligência de Negócios Operacional
Universidade da Califórnia - Irvine

Análise de Dados
Universidade Wesleyana

Índice

1.	A Natureza dos dados e o Projeto de Banco de Dados Relacionais	3
1.1.	Business Intelligence, Business Analytics e Data Science	3
1.2.	OLTP versus OLAP.....	4
1.3.	Data Warehousing para BI	5
1.4.	Definindo Bancos de Dados Relacionais.....	6
1.4.1.	Diagrama Entidade-Relacionamento (ERD)	6
1.4.2.	Normalização e Desnormalização	8
2.	Data Warehousing e Business Intelligence.....	11
2.1.	Necessidade de armazenamento de dados	11
2.1.1.	Arquiteturas de armazenamento de dados	11
2.1.2.	Extração, transformação e carga (ETL).....	12
2.1.3.	Data Marts	12
2.1.4.	Armazenamentos de dados operacionais	12
2.1.5.	Armazenamento de dados na nuvem	13
2.2.	Modelagem de dados para Data Warehouse.....	13
2.2.1.	Modelagem de dados multidimensionais	14
2.2.2.	NoSQL, Big Data, Data Lakes e Data Warehousing.....	15
2.3.	O Processo de Preparação de Dados.....	15
3.	A Natureza dos Dados	17
3.1.	Análise de dados	17
3.2.	Dados e Tipos de Dados	18
3.3.	Datasets e Codebooks	20
3.4.	Desenvolvendo uma questão de pesquisa.....	22
4.	Estatística	23
4.1.	Estatísticas descritivas	23
4.1.1.	Análise Exploratória de Dados	23
4.1.2.	Examinando a distribuição de frequência.....	24
4.1.3.	Plotando as distribuições	24
4.1.4.	Medidas de Centralidade e Dispersão	29
4.2.	Estatística inferencial	33
4.2.1.	Da amostra à população.....	34
4.2.2.	Teste de Hipótese	36
4.2.3.	Valor-p e Intervalo de Confiança	38
4.2.4.	Escolhendo testes estatísticos	39
4.2.5.	Análise de Variância - ANOVA	40
4.2.6.	Teste de Independência Qui-Quadrado	45
4.2.7.	Teste de Correlação de Pearson.....	48
5.	Data Mining para Predição e Explicação.....	Erro! Indicador não definido.
5.1.	Visão geral da mineração de dados para BI	Erro! Indicador não definido.
5.2.	Processo de mineração de dados	Erro! Indicador não definido.
5.3.	Métodos de mineração de dados.....	Erro! Indicador não definido.
5.4.	Algoritmos de mineração de dados para modelagem preditiva.....	Erro! Indicador não definido.
5.5.	Data Mining para Agrupamento e Associação	Erro! Indicador não definido.
5.5.1.	Análise de Associação e Cesta de Mercado	Erro! Indicador não definido.
6.	Business Intelligence e Visual Analytics.....	51
6.1.	Visualização de dados e Visual Analytics	51
6.2.	Qual visualização é boa para que propósito?.....	51
6.3.	Uma visão geral do SAS Viya	53

1. A Natureza dos dados e o Projeto de Banco de Dados Relacionais

1.1. Business Intelligence, Business Analytics e Data Science

Decifrando a confusão de nomes

Inteligência de negócios, análise de negócios e ciência de dados são todos usados como termos abrangentes para campos relacionados, e essas semelhanças geralmente podem levar à confusão ao tentar entender o que significam. Embora esses conceitos estejam de fato relacionados, eles também são distintamente diferentes.

- **Inteligência de negócios:** Business Intelligence (BI) é um processo bem definido de análise e processamento de dados para fins de visualização e aplicação de informações acionáveis. O conceito de business intelligence evoluiu ao longo de várias décadas e é frequentemente usado como um termo abrangente. Em última análise, a inteligência de negócios adiciona “contexto” aos dados para produzir informações acionáveis, ou seja, aquelas que auxiliam no suporte à decisão. Um dos principais objetivos do BI é colocar o poder da visualização nas mãos dos usuários finais e permitir a tomada de decisões orientada por dados. Existem muitas ferramentas e aplicativos no mercado atual para dar suporte ao BI e impulsionar as soluções de negócios.
- **Análise de negócios:** A análise de negócios usa matemática e estatística para analisar os dados de uma organização. A análise de negócios oferece suporte direto ao BI para permitir a tomada de decisões orientada por dados e obter insights para suporte à decisão. Os principais componentes da análise de negócios são qualidade de dados, análise precisa e profunda, aplicação eficiente de ferramentas e modelos preditivos e automação. Os dados podem ser coletados de muitas fontes diferentes, incluindo sistemas transacionais, data warehouses e até mesmo fontes de dados não estruturadas. A análise de negócios geralmente é categorizada como descritiva, preditiva ou prescritiva, e essas categorias aumentam em valor de negócios (e complexidade) à medida que você passa de uma para outra.
 - **A análise descritiva** é usada para rastrear os principais indicadores de desempenho (KPIs) e para entender e descrever o estado atual. A inteligência de negócios tradicional usa análises descritivas para analisar as operações de negócios existentes e gerar uma imagem atual dos negócios.
 - **A análise preditiva** é usada para realizar análises de tendências e tentar identificar resultados futuros.
 - **A análise prescritiva** usa dados de desempenho anteriores para gerar recomendações para situações futuras com entradas semelhantes.
- **Ciência de dados:** A ciência de dados é um campo avançado que abrange áreas como mineração de dados, aprendizado de máquina e estatística. Essas áreas geralmente exigem níveis profundos de codificação personalizada para explorar perguntas abertas. Os cientistas de dados empregam métodos estatísticos avançados para explorar e descobrir padrões e novos insights por meio de análises. Os objetivos da ciência de dados incluem aumentar a eficiência operacional, encontrar oportunidades e fornecer vantagens competitivas. A ciência de dados também é essencial para alavancar o poder de processamento computacional para suporte a decisões, modelagem preditiva, simulação avançada e muitos outros aplicativos de negócios.

Ter uma melhor compreensão das distinções desses termos (inteligência de negócios, análise de negócios e ciência de dados) nos ajudará a explorar outros conceitos relacionados neste e nos módulos futuros.

Sistemas de Suporte a Decisão

Um sistema de suporte à decisão (DSS) é um sistema de informação que permite e suporta diretamente a tomada de decisões orientada por dados. Os gerentes e líderes organizacionais tradicionalmente empregam esses sistemas para fornecer uma imagem de “verdade básica” de uma determinada situação. O DSS permite a análise rápida de grandes quantidades de dados para resolver desafios complexos. O poder do DSS vem por meio de

relatórios em tempo real, que fornecem dados constantemente atualizados para dar suporte a decisões críticas em um ambiente de negócios complexo. Um exemplo bem conhecido, mas direto de um DSS é o planejamento de destino/rota usando GPS. O sistema de informação GPS gera várias rotas disponíveis para o usuário e recomenda uma rota com base em variáveis como tráfego, interdições de estradas, pedágios, etc. dados relacionados e fazer recomendações.

Atores de BI e Análise

Existem várias e amplas preocupações que impulsionam a necessidade de análise de negócios. Alguns dos fatores mais comuns incluem o enorme volume de dados coletados, os requisitos de disponibilidade e segurança de dados e a necessidade de tomar decisões de negócios melhores e mais rápidas. À medida que as organizações coletam mais volumes de dados em velocidades cada vez maiores, a necessidade de organizar e analisar esses dados com eficiência também aumenta. Além disso, a natureza móvel dos negócios exige disponibilidade consistente de dados para dar suporte à tomada de decisões em tempo real, independentemente da localização.

Embora a disponibilidade de dados seja fundamental para a implementação eficaz do BI, a segurança dos dados também é um foco principal e continuaremos a discutir ao longo deste programa. E, finalmente, o ambiente de negócios em rápida mudança de hoje exige decisões melhores e mais rápidas, e a análise de dados pode capacitar e apoiar os líderes na tomada de decisões orientadas por dados. À medida que grandes volumes de dados são coletados, é crucial ter uma estratégia de dados clara para uma análise adequada e esforços de implementação. O foco precisa estar na conversão de dados em informações açãoáveis.

Uma taxonomia simples para análise

Desenvolver uma taxonomia simples e aceitável para análise de negócios é essencial, pois os conceitos e as tecnologias mudam tão rapidamente. As partes interessadas podem maximizar o valor e garantir clareza se puderem falar a partir de um contexto compartilhado e entender a terminologia de análise de negócios. Várias empresas e instituições acadêmicas tentaram alinhar o contexto, a compreensão e a terminologia, e seu trabalho acabou produzindo uma versão da taxonomia vista na tabela de análise de negócios, vinculada aqui (adaptado de Delen, 2020).

Referências

Delen, D. (2020). Prescriptive analytics: The final frontier for evidence-based management and optimal decision making. << <https://www.pearson.com/us/higher-education/program/Delen-Prescriptive-Analytics-The-Final-Frontier-for-Evidence-Based-Management-and-Optimal-Decision-Making/PGM239919.html> >>

1.2. OLTP versus OLAP

OLTP e OLAP são ambos sistemas de processamento online. A distinção entre esses sistemas está no que está sendo processado, ou seja, transações ou consultas analíticas.

- OLTP = Processamento de Transações Online
- OLAP = Processamento analítico online

Processamento de transações on-line (OLTP)

O OLTP é utilizado para processar sistemas transacionais e normalmente envolve a modificação de um sistema de banco de dados online. Um exemplo simples é um site de comércio eletrônico. Cada vez que um pedido é feito, um banco de dados (ou vários bancos de dados) são modificados para armazenar detalhes do cliente e do pedido (entre outros dados). Essa transação é processada pelo OLTP, que lida com inserções, atualizações e exclusões. Os bancos de dados OLTP são atualizados com frequência e geralmente são chamados de sistemas transacionais ou operacionais.

Processamento analítico online (OLAP)

O OLAP lida com a consulta de um sistema de banco de dados online. Os bancos de dados OLAP armazenam dados históricos para relatórios e análises em suporte direto à tomada de decisões orientada por dados. O mesmo site de comércio eletrônico pode relatar contagens de estoque atuais ou gerar relatórios de vendas. Nesse caso, o OLAP extrai dados do sistema de banco de dados para suporte à decisão.

1.3. Data Warehousing para BI

O data warehousing (DW) emprega um processo de extração, transformação e carregamento (ETL) para coletar dados de sistemas transacionais distintos (OLTP) e armazenar esses dados para fins históricos, analíticos e de relatórios. Os data warehouses são imutáveis, integrados, granulares e históricos por natureza. Eles são frequentemente considerados a “fonte única da verdade” devido à sua natureza imutável; ou seja, uma vez que os dados passaram pelo processo de ETL, eles não são alterados novamente.

O processo ETL limpa, normaliza, alinha e carrega dados no data warehouse para permitir análises e relatórios eficientes e eficazes por meio do data warehouse (OLAP). O DW fornece contexto histórico e um conjunto de dados normalizado a partir do qual relatórios e análises podem ser conduzidos. O data warehouse geralmente agrupa e calcula cálculos comuns normalmente incluídos nos relatórios e visualizações organizacionais. Essas etapas reduzirão o tempo de processamento computacional ao executar análises e gerar relatórios *ad hoc*.

Exercício

1. Business Intelligence (BI) adiciona _____ aos dados para produzir informações acionáveis.
 - a) Visualizações
 - b) Tecnologia
 - c) Contexto
2. Quais dos seguintes são objetivos do BI?
 - a) Coloque o poder da visualização nas mãos dos usuários finais
 - b) Habilite a tomada de decisões orientada por dados
 - c) Todas essas opções estão corretas.
3. _____ inclui qualidade de dados, análise precisa e profunda, aplicação eficiente de ferramentas e modelos preditivos e automação.
 - a) Analista de negócios
 - b) Inteligência de negócios
 - c) Ciência de dados
4. Que tipo de análise de negócios é usada para conduzir a análise de tendências?
 - a) Descritivo
 - b) Preditivo
 - c) Prescritivo
5. A ciência de dados mergulha profundamente em _____.
 - a) Mineração de dados
 - b) Aprendizado de máquina
 - c) Todas essas opções estão corretas.
6. _____ é um sistema de informação que suporta e permite a tomada de decisões orientada por dados, fornecendo uma imagem da verdade.
 - a) Sistema de Informação Geoespacial
 - b) Sistema de Apoio à Decisão
 - c) Sistema de Gestão de Relacionamento com o Cliente
7. Os drivers de análise reduzem a clareza das implementações de BI e causam confusão sobre dados críticos.
 - a) Verdadeiro
 - b) Falso
8. Em qual categoria de análise de negócios normalmente pertence o Business Intelligence?
 - a) Descritivo
 - b) Prescritivo
 - c) Preditivo

9. Qual das opções a seguir lida com a consulta de um sistema de banco de dados online?
- a) OLTP
 - b) OLAP
 - c) Todas essas opções estão corretas.
10. O processo ETL limpa, normaliza, alinha e carrega dados no data warehouse.
- a) Verdadeiro
 - b) Falso

Leituras recomendadas:

Article: Yellowfin Team. (nd). [Business Intelligence: Drivers, Challenges, Benefits and ROI](https://www.yellowfinbi.com/blog/2011/04/yfcommunitynews-business-intelligence-drivers-challenges-benefits-and-roi-103783). (5 min)
<<<https://www.yellowfinbi.com/blog/2011/04/yfcommunitynews-business-intelligence-drivers-challenges-benefits-and-roi-103783>>>

Article: Glen, S. (2020). [Business Intelligence vs Business Analytics](https://www.datasciencecentral.com/profiles/blogs/business-intelligence-vs-business-analytics-vs-data-analytics). (5 min)
<<<https://www.datasciencecentral.com/profiles/blogs/business-intelligence-vs-business-analytics-vs-data-analytics>>>

1.4. Definindo Bancos de Dados Relacionais

Um banco de dados relacional é uma coleção de dados relacionados armazenados em um local ou repositório centralizado. Os dados armazenados são organizados em tabelas que abrigam informações sobre vários objetos armazenados no banco de dados. Os bancos de dados relacionais fornecem uma maneira eficiente, flexível e escalável de armazenar e acessar informações estruturadas.

Os bancos de dados relacionais geralmente são hospedados e gerenciados usando um sistema de gerenciamento de banco de dados relacional (RDBMS). O RDBMS emprega Structured Query Language (SQL) para permitir a recuperação e interação com dados em várias tabelas. Esses sistemas também geralmente implantam autenticação, autorização, ajuste de desempenho e muitos outros recursos.

Bancos de dados relacionais são organizados por agrupamentos de objetos que possuem um identificador único ou chave primária. A chave primária identifica a linha em uma tabela que corresponde a um registro individual e seus dados associados. A chave primária também pode ser usada como chave estrangeira em outra tabela para indicar relacionamento. Chaves estrangeiras criam conexões lógicas entre tabelas e estabelecem relacionamentos.

1.4.1. Diagrama Entidade-Relacionamento (ERD)

Um diagrama entidade-relacionamento (ERD) é uma representação gráfica de um projeto de banco de dados. Os diagramas de exemplo abaixo (Figuras 1-3) ilustram um ERD simples que descreve o design geral e estabelece a base e os requisitos para implementação em um RDBMS. O ERD também estabelece relacionamentos entre objetos e serve como documentação para o sistema de banco de dados.

O Processo de Projetar Bancos de Dados

A modelagem de dados é o processo de projetar bancos de dados e existem três modelos de dados: dados conceituais, dados lógicos e dados físicos.

Projeto conceitual

O projeto conceitual estabelece entidades, atributos e relacionamentos. O objetivo de um modelo de dados conceitual é apresentar uma imagem de alto nível do sistema a ser implementado com foco nos objetos de negócios envolvidos no sistema. As tabelas de banco de dados não são projetadas no nível conceitual.

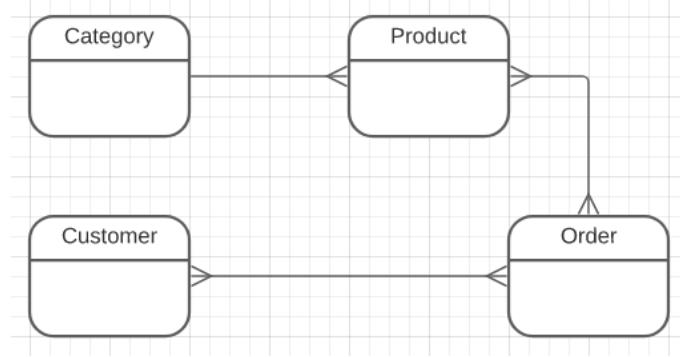


Figura 1 Objetos de Negócio da Entidade (design conceitual)

A Figura 1 descreve os objetos de negócios da entidade que interagem ou fazem parte de um sistema de informações. Neste exemplo, temos clientes solicitando produtos. As relações de base são identificadas usando a notação pé de galinha. Uma única linha indica um único relacionamento (ou seja, um produto só pode estar em uma categoria), e um pé de galinha de três linhas indica um relacionamento do tipo “muitos” (ou seja, uma categoria pode ter muitos produtos).

Projeto Lógico

O design lógico define a estrutura dos elementos de dados e estabelece relacionamentos entre os elementos de dados. O modelo de dados lógico adiciona uma camada de detalhes ao projeto conceitual, definindo as colunas de dados que precisam ser incluídas em cada entidade, como visto na Figura 2. Nesta fase do projeto, ainda não há consideração por um sistema de banco de dados específico já que o foco está na estrutura e no relacionamento.

Projeto conceitual de objetos de negócios de entidade com atributos.

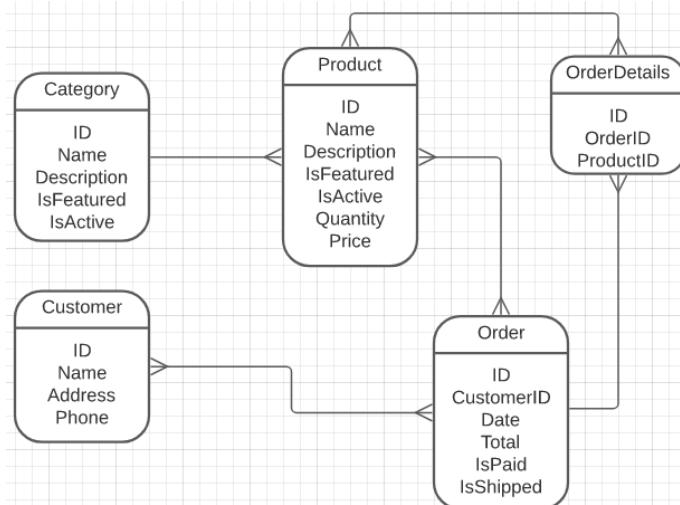


Figura 2 Objetos de Negócio da Entidade (design conceitual com atributos)

Cada objeto de negócios ou entidade agora inclui atributos ou colunas que descreverão registros individuais dentro da eventual tabela do banco de dados. Esses atributos começam a detalhar as informações que compõem um único registro (ou linha) dentro de uma eventual tabela de banco de dados.

Projeto Físico

O design físico descreve detalhes de implementação específicos do banco de dados e fornece um plano para o banco de dados relacional. O modelo de dados físico inclui detalhes adicionais sobre cada coluna dentro

de uma entidade. Nesta fase do projeto, é importante operar dentro das construções de um RDBMS específico, pois as estruturas, convenções e restrições podem variar.

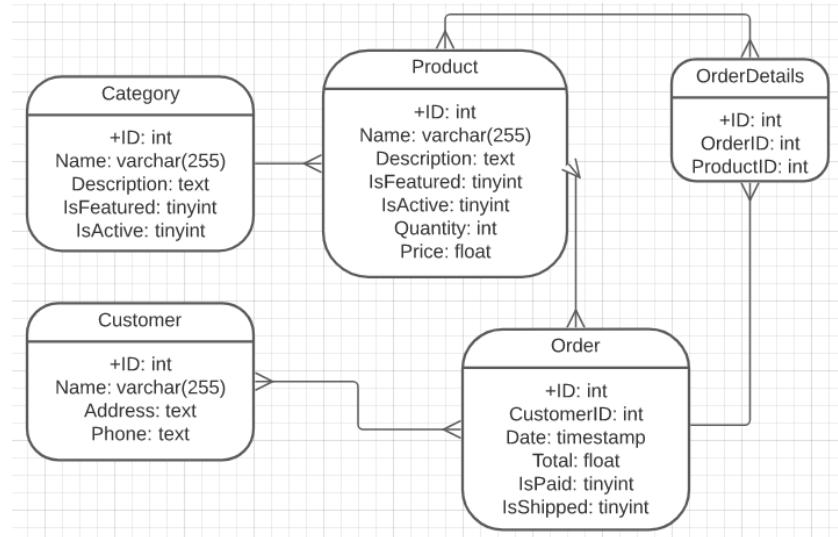


Figura 3 Objetos de Negócio da Entidade (modelo de dados físico)

Como mostra a Figura 3, agora temos um design de banco de dados totalmente definido que está pronto para implementação em nosso RDBMS selecionado. As chaves primárias para cada tabela são marcadas com um símbolo “+”, e os tipos de dados para cada coluna são identificados e seguem os tipos de dados aceitáveis para o MySQL RDBMS.

1.4.2. Normalização e Desnormalização

Os conceitos de normalização e desnormalização descrevem a organização do conteúdo de um banco de dados. A normalização envolve a separação de dados em objetos bem definidos para limitar a redundância de dados. Na normalização, há um grande foco nos relacionamentos entre tabelas, e cada tabela contém informações exclusivas que são necessárias para descrever um registro ou entidade individual. Para recuperar todos os dados associados sobre um determinado registro, o usuário precisaria executar muitas junções (exploradas mais adiante), o que pode causar problemas de desempenho.

A desnormalização combina dados em uma única tabela para remover relacionamentos e dependências externas. Embora essa abordagem possa acelerar as consultas SQL, muitas vezes também resulta em dados redundantes ou duplicados em todo o banco de dados. A tabela de normalização e desnormalização vinculada aqui contém mais detalhes sobre cada um desses conceitos.

Aplicação do Diagrama Entidade-Relacionamento

Selecione um sistema e crie um ERD com progressão do projeto conceitual para o lógico e o físico.

Material de apoio

A seguir está uma lista de recursos opcionais que você pode achar úteis para melhorar sua compreensão dos tópicos deste módulo.

Vídeo: Lucidchart. (2018). [The Basics of Relational Database Design.](#) (5 min)

Vídeo: CBT Nuggets. (2019). [How to Normalize Databases.](#) (7 min)

Artigo: Guru99. (n.d.). [What is Normalization? 1NF, 2NF, 3NF, BCNF Database Example.](#) (10 min)

A seguir está uma lista de recursos opcionais que você pode achar úteis para melhorar sua compreensão sobre SQL.

Video: Guru99. (2013). [What is Database & SQL? \(6 min\)](#)

Video: Socratica. (2019). [SQL SELECT Tutorial ||| SQL Tutorial ||| SQL for Beginners. \(9 min\)](#)

Article: Menshov, S. (2019). [Tutorial on SQL \(DDL, DML\) on the Example of MS SQL Server Dialect. \(30 min\)](#)

Article: W3Schools. (n.d.). [SQL Tutorial. \(30 min\)](#)

Exercício

1. Um _____ é uma coleção de dados relacionados armazenados em um local ou repositório centralizado.

- a) Sistema de gerenciamento de banco de dados relacional
- b) Banco de dados relacional
- c) Diagrama de Entidade-Relacionamento

2. O que é SQL?

- a) Structured Query Language
- b) Simple Question Location
- c) Simplified Query Language

3. Um _____ identifica a linha em uma tabela que corresponde a um registro individual e seus dados associados.

- a) Chave primária
- b) ERD
- c) Chave estrangeira

4. _____ criar conexões lógicas entre tabelas e estabelecer relacionamento.

- a) Chaves primárias
- b) Chaves estrangeiras
- c) Nenhuma dessas opções está correta.

5. _____ é o processo de projetar bancos de dados.

- a) ERD
- b) Linguagem de consulta estruturada
- c) Modelagem de dados

6. _____ estabelece entidades, atributos e relacionamentos.

- a) Projeto conceitual
- b) Projeto físico
- c) Projeto lógico

7. O modelo _____ adiciona uma camada de detalhes ao projeto conceitual definindo as colunas de dados que precisam ser incluídas em cada entidade.

- a) Projeto físico
- b) Projeto conceptual
- c) Projeto lógico

8. O modelo de dados _____ inclui detalhes adicionais sobre cada coluna dentro de uma entidade.

- a) Lógico
- b) Físico
- c) Conceptual

9. O _____ estabelece relacionamentos entre objetos e serve como documentação para o sistema de banco de dados.

- a) Modelo de dados conceitual
- b) ERD
- c) RDBMS

10. Qual das opções a seguir reduz a redundância e a inconsistência de dados?

- a) Desnormalização
- b) Normalização
- c) Modelagem de dados

Exercício de SQL

1. _____ é uma linguagem de programação de banco de dados que permite interagir com um banco de dados para executar operações como SELECT, INSERT, UPDATE e DELETE.

- a) PHP
- b) SQL
- c) RDBMS

2. Qual dos seguintes não faz parte do DDL?

- a) SELECT
- b) ALTER
- c) CREATE

3. Qual dos seguintes não faz parte da DML?

- a) DELETE
- b) ALTER
- c) INSERT

4. DCL inclui todos os itens a seguir, exceto:

- a) REVOKE
- b) GRANT
- c) Todas essas opções estão corretas

5. Quais dos seguintes não fazem parte do TCL?

- a) COMMIT
- b) Todas essas opções fazem parte do TCL
- c) ROLLBACK

6. _____ retorna linhas e nos permite coletar dados de tabelas normalizadas.

- a) Subqueries
- b) Inserts
- c) Joins

7. Um uso comum para um _____ pode ser calcular o total de todos os produtos em nosso pedido ou um preço médio de nossos produtos.

- a) DDL
- b) Join
- c) Subquery

8. Qual palavra-chave do MySQL gerencia a atribuição de um valor de chave primária sem intervenção do usuário?

- a) AUTO_INCREMENT
- b) PRIMARY KRY
- c) NOT NULL

9. Qual tipo de dados MySQL permite que uma coluna não contenha mais de 255 caracteres?

- a) FLOAT
- b) TEXT
- c) VARCHAR(255)

10. Qual palavra-chave do MySQL define um valor padrão para uma coluna da tabela de banco de dados quando o usuário não fornece um valor?

- a) INSERT
- b) NOT NULL
- c) DEFAULT

2. Data Warehousing e Business Intelligence

2.1. Necessidade de armazenamento de dados

Um data warehouse (DW) é um repositório que armazena dados relacionais organizados, limpos e padronizados para uso corporativo. Um data warehouse é organizado por bancos de dados orientados a assunto e não é volátil no suporte direto à funcionalidade do sistema de suporte à decisão (DSS). Ao fazer isso, um data warehouse inclui dados estrategicamente selecionados que são importantes para uma empresa para rastreamento histórico, relatórios e análises.

Um data warehouse tem as seguintes características:

- **Orientado a assunto:** os dados são baseados em tema ou objeto (ou seja, cliente, produto, vendas, etc.)
- **Integrado:** dados díspares são combinados e normalizados a partir de sistemas de origem
- **Variante de tempo:** os dados são organizados por vários intervalos de tempo para relatórios históricos e preservação (ou seja, semana, mês, trimestre, ano)
- **Não volátil:** os dados nunca são alterados ou excluídos; os dados são somente leitura e atualizados em intervalos de tempo bem definidos
- **Resumido:** os dados geralmente são agregados para otimização dos relatórios

Um data warehouse deve incluir metadados, que são “dados que descrevem dados”. Metadados geralmente incluem localização de dados, estrutura de dados e parâmetros de valores válidos. Essencialmente, os metadados atuam como “um dicionário vivo” e documentação para o data warehouse.

A necessidade de armazenamento de dados (data warehousing) torna-se evidente quando entendemos que os dados estão em toda parte. Muitas organizações utilizam meios e sistemas diferentes para coletar dados. Um data warehouse extrai dados desses sistemas de origem díspares, que podem incluir ponto de venda (SPT), planejamento de recursos empresariais (ERP), gerenciamento de relacionamento com o cliente (CRM), etc. O processo de extração, transformação, carregamento (ETL), que será discutido posteriormente neste módulo, prepara e normaliza os dados extraídos para análise e relatório. Além disso, um data warehouse permite o rastreamento e a manutenção de informações históricas e fornece uma única fonte de verdade.

2.1.1. Arquiteturas de armazenamento de dados

Os data warehouses podem ser arquitetados usando abordagens variadas. Existem duas abordagens principais: a abordagem dimensional (popularizada por Ralph Kimball) e a abordagem normalizada (popularizada por Bill Inmon).

Abordagem Dimensional

A abordagem de Kimball descreve um data warehouse por meio de um modelo dimensional (esquema em estrela ou floco de neve). A abordagem dimensional usa um design bottom-up “de baixo para cima”, no qual data marts individuais são criados em nível departamental ou organizacional (ou seja, vendas, recursos humanos, finanças, etc.) e construído para um armazém de dados corporativo (Enterprise Data Warehouse - EDW). Hoje, a abordagem de Kimball é mais popular porque os usuários de negócios podem rapidamente ganhar utilidade com ela.

Abordagem Normalizada

A Inmon, por outro lado, utilizou uma abordagem Top-Down “de cima para baixo” para normalizar um data warehouse. O modelo de dados corporativos normalizado cria um repositório central ou data warehouse

corporativo. Data marts dimensionais para departamentos ou unidades organizacionais específicas podem ser criados a partir do data warehouse corporativo mestre.

2.1.2. Extração, transformação e carga (ETL)

Extração, transformação e carga (ETL) é o processo de integração de dados de sistemas operacionais ou transacionais de origem para combinar dados diferentes em um único formato em um repositório central. Os dados de origem são extraídos de sistemas transacionais; transformado para normalização, formatação e correção de erros; e carregado no data warehouse para análise e relatórios (como visto na Figura 4).

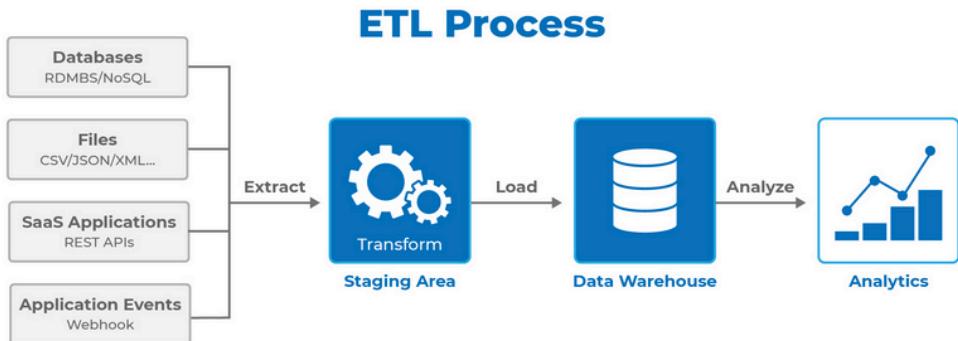


Figura 4 O processo ETL

2.1.3. Data Marts

Um data mart é um subconjunto de um data warehouse corporativo e geralmente é chamado de "data warehouse departamental". Um data mart contém o mesmo tipo de informação que existe em um data warehouse corporativo, mas os dados são organizados e otimizados para um departamento específico ou unidade organizacional. O diagrama na Figura 5 fornece uma arquitetura de alto nível de data warehousing e mostra como os data marts se encaixam nessa arquitetura.

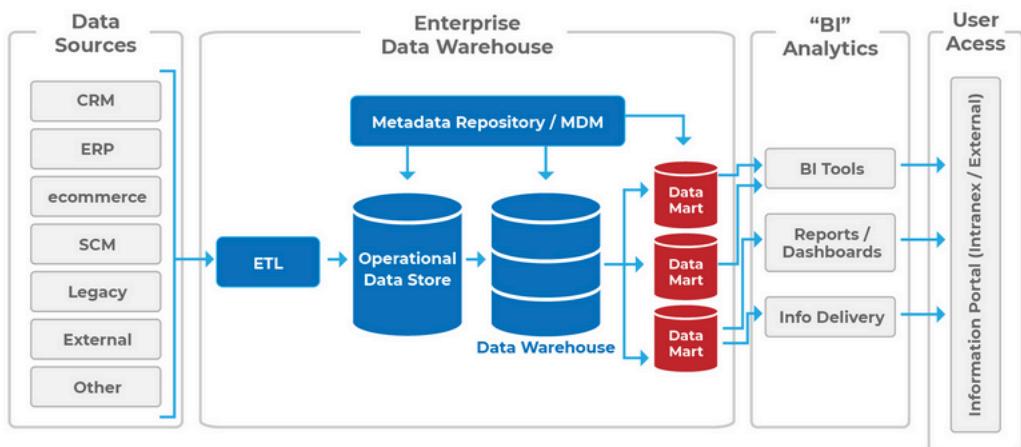


Figura 5 Data marts em uma arquitetura de data warehousing

2.1.4. Armazenamentos de dados operacionais

Um armazenamento de dados operacionais (ODS) utiliza snapshots de dados de sistemas operacionais ou transacionais para fornecer relatórios operacionais de negócios. O ODS difere de um data warehouse porque os dados são acessados diretamente dos bancos de dados do sistema transacional e o armazenamento de dados operacional pode gravar dados de volta nos sistemas de origem. Um objetivo principal de um armazenamento de dados operacional é lidar com as complexidades de manter dados atualizados no data warehouse. Assim, o ODS pode ser visto como uma abordagem menos dispendiosa para relatórios de dados em tempo real.

2.1.5. Armazenamento de dados na nuvem

Os data warehouses tradicionalmente existem dentro da infraestrutura local de uma organização (on-premises), onde a responsabilidade pela configuração e manutenção recai exclusivamente sobre a equipe de tecnologia da informação (TI) da organização. O armazenamento de dados na nuvem transfere grande parte da responsabilidade de hardware, rede, segurança e manutenção para terceiros, o que permite que a organização se concentre mais nas metas e objetivos de negócios. Essa abordagem também permite aos usuários (que geralmente são remotos ou móveis) um nível mais alto e mais consistente de disponibilidade de data warehouse.

Exercício

1. Um _____ é um repositório que armazena dados relacionais organizados, limpos e padronizados para uso corporativo.
 - a) Base de dados
 - b) Sistema de gerenciamento de banco de dados
 - c) Data Warehouse
 2. Qual das seguintes características descreve um DW como sendo organizado por intervalos de tempo?
 - a) Não volátil
 - b) Tempo variável
 - c) Integrado
 3. Metadados são dados sobre dados.
 - a) Falso
 - b) Verdadeiro
 4. Qual abordagem de arquitetura de data warehousing utiliza um design de bottom-up?
 - a) Desnormalizado
 - b) Dimensional
 - c) Normalizado
 5. A abordagem top-down de Inmon para a arquitetura DW cria um repositório central normalizado ou _____.
 - a) Armazenamento de dados operacionais
 - b) Enterprise Data Warehouse
 - c) Data Mart
 6. O processo _____ combina dados díspares em um repositório central.
- a) Extração
 - b) Transformação
 - c) Extrair, transformar, carregar (ETL)
7. Qual das opções a seguir é um subconjunto de um data warehouse e geralmente é focado no departamento?
 - a) Data Mart
 - b) Armazenamento de dados operacionais
 - c) Banco de dados transacional
 8. Qual dos seguintes usa instantâneos de sistemas transacionais para fornecer relatórios operacionais de negócios?
 - a) Armazenamento de dados operacionais (ODS)
 - b) Data Mart
 - c) Enterprise Data Warehouse
 9. Qual das opções a seguir é um exemplo de uma fonte de dados transacional?
 - a) CRM
 - b) ERP
 - c) Todas essas opções estão corretas
 10. O data warehouse baseado em nuvem transfere grande parte da responsabilidade de hardware, rede, segurança e manutenção para terceiros.
 - a) Falso
 - b) Verdadeiro

Material Complementar

https://www.youtube.com/watch?v=Tff34jj_V-0

2.2. Modelagem de dados para Data Warehouse

Anteriormente, vimos a importância da modelagem de dados no projeto e implementação de banco de dados. Isso também se aplica ao Data Warehouse. O processo de modelagem de dados permanece o mesmo, sendo o objetivo “a organização e armazenamento de dados de longo prazo para análise e relatórios”. O modelo de dados precisa suportar as características básicas de um data warehouse, ou seja, ser orientado por assunto, integrado, variante no tempo, não volátil e resumido. O processo de modelagem de dados para data warehousing

ainda segue o processo de design - do conceitual ao lógico e aos ERDs físicos (diagramas de entidade-relacionamento). Outra área a ser considerada é a arquitetura de data warehouse selecionada (ou seja, dimensional ou normalizada) e se os data marts serão incorporados à arquitetura.

2.2.1. Modelagem de dados multidimensionais

Os modelos de dados multidimensionais representam estruturas de dados complexas (geralmente em formato de cubo) em oposição a uma única dimensão (geralmente representada por uma lista). Modelos de dados bidimensionais e tridimensionais são frequentemente utilizados em data warehouse. Esses modelos permitem uma estrutura e organização de dados bem definida. As etapas gerais na construção de um modelo de dados multidimensional incluem:

- Coletando os requisitos do usuário
- Categorizando os módulos do sistema
- Identificando dimensões para organizar dados em torno de objetos e funções
- Esboçar as dimensões em tempo real e as propriedades correspondentes
- Descobrindo os fatos a partir das dimensões e suas propriedades
- Construindo o esquema para armazenamento de dados

Esquema em estrela (Star Schema)

Um esquema em estrela é um modelo que descreve dados em uma forma semelhante à de uma estrela. Uma tabela de **fatos** existe no centro da estrela e contém chaves primárias e estrangeiras para tabelas de **dimensões associadas**, bem como dados agregados dos sistemas operacionais ou transacionais. As tabelas de dimensão descrevem os dados e são incluídas com base nas necessidades de negócios. Um esquema em estrela não é normalizado e fornece modelagem simples sem a necessidade de junções complexas.

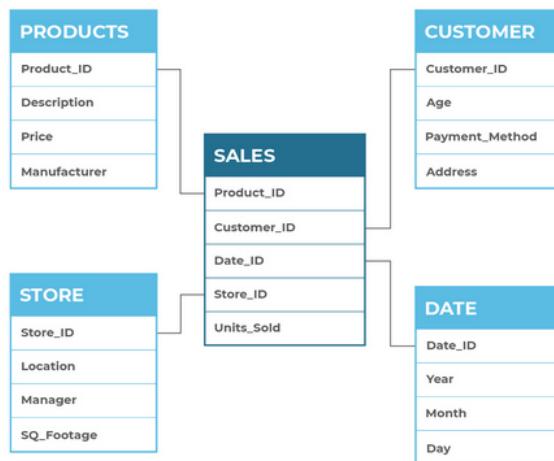


Figura 6 Exemplo de um esquema estrela

Esquema de floco de neve (Snowflake schema)

O design do esquema floco de neve contém os mesmos dados que existiriam em um esquema em estrela, e a tabela de fatos e as tabelas de dimensões têm a mesma aparência. A principal diferença entre os dois é que o esquema floco de neve é normalizado. O processo de normalização do projeto é conhecido como floco de neve. O esquema floco de neve também requer menos trabalho para adicionar mais dados às dimensões existentes e requer menos armazenamento devido à falta de redundância no processo de normalização. A Figura 7 exibe um exemplo de um esquema de floco de neve.

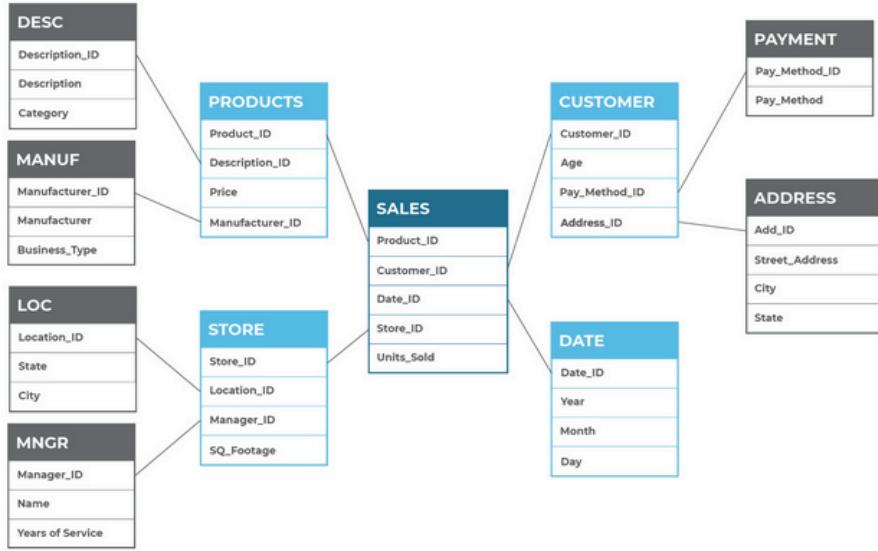


Figura 7 Exemplo de um esquema floco de neve (snowflake)

2.2.2. NoSQL, Big Data, Data Lakes e Data Warehousing

Ao contrário da abordagem tradicional de banco de dados relacional para armazenamento de dados, o NoSQL é uma abordagem alternativa que utiliza bancos de dados não relacionais e não estruturados. O NoSQL pode armazenar dados de qualquer forma porque não é limitado pelas estruturas estritamente definidas dos bancos de dados relacionais. Devido à falta de clareza e requisitos em torno da estrutura dos dados, muitas vezes não é possível desenvolver um esquema. Assim, os bancos de dados NoSQL permitem a flexibilidade de armazenar e consultar dados não estruturados. Isso é realizado por meio de uma organização orientada a documentos, em vez da organização orientada a tabelas de bancos de dados SQL estruturados. No entanto, é importante observar que esse tipo de armazenamento de dados também requer processamento e armazenamento adicionais.

Big data é um conceito para lidar com grandes quantidades de dados brutos e não estruturados em vários tipos e formatos. Torna-se rapidamente difícil para um data warehouse gerenciar esse tipo de estratégia de dados e o modelo de big data tenta resolver o problema. Devido ao tamanho, complexidade e natureza dinâmica do big data, os dados geralmente são transformados durante a análise e requerem poder de processamento significativo.

O conceito relativamente novo de data lakes oferece uma abordagem descentralizada para armazenamento e análise de dados, em vez da abordagem centralizada empregada por data warehouses tradicionais. Um data lake prefere ter repositórios de dados brutos de sistemas operacionais ou transacionais de origem disponíveis para analistas e cientistas de dados, em vez de transformar e carregar todos os dados em um repositório centralizado. Esse conceito fornece uma estratégia de armazenamento de dados e limita o pré-processamento e a governança rígida, o que certamente pode trazer benefícios e desafios para a organização. Após a análise e processamento de dados, os dados em um data lake podem ser incorporados a um data warehouse para armazenamento de longo prazo e análise futura, embora um data lake não seja necessariamente um substituto para um data warehouse.

2.3. O Processo de Preparação de Dados

A preparação de dados garante a prontidão de um conjunto de dados para análise. Em geral, esse processo consiste em preparar dados brutos para ingestão em uma ferramenta ou serviço de análise de dados. Como

consideramos brevemente no módulo anterior, os dados devem passar por um processo definido chamado extrair, transformar, carregar (ETL).

- **Extração:** os dados são extraídos de sistemas de origem, repositórios e ferramentas.
- **Transformação:** os dados são limpos, normalizados e agregados para facilitar a análise.
- **Carga (load):** os dados são carregados em um banco de dados comum, data warehouse, etc. para facilitar o acesso comum e uma única fonte de verdade para análise.

Embora o ETL descreva o processo geral, há muitas etapas detalhadas que geralmente estão envolvidas nas fases preparatórias da análise. Isso inclui agregação, combinação ou separação de campos, normalização do formato de um ponto de dados, codificação, transcrição, tratamento de valores nulos ou ausentes, verificação de erros, etc.

Material Complementar

Article: The Kimball Group. (2016). [Dimensional Modeling Techniques](https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/). (20 min) << <https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/> >>

Exercício

1. _____ é uma etapa crítica para data warehousing e produz projetos em suporte às características de DW.
 - a) Análise de dados
 - b) ETL
 - c) Modelagem de dados
 2. As considerações da arquitetura de data warehouse devem ser incluídas na fase de projeto de modelagem de dados.
 - a) Falso
 - b) Verdadeiro
 3. _____ são usados para representar estruturas de dados complexas, geralmente em formato de cubo.
 - a) Diagramas de Relacionamento de Entidade (ERDs)
 - b) Modelos de dados multidimensionais
 - c) Modelos de dados unidimensionais
 4. Qual das alternativas a seguir não é uma etapa na construção de um modelo de dados multidimensional?
 - a) Coletando os requisitos do usuário
 - b) Identificando dimensões para organizar dados em torno de objetos e funções
 - c) Todas essas opções estão corretas
 5. Um _____ é um modelo que descreve os dados em uma forma semelhante a uma estrela.
 - a) Modelo de dados multidimensional
 - b) Star Schema
- c) Snowflake Schema
 6. O esquema floco de neve fornece um design normalizado.
 - a) Verdadeiro
 - b) Falso
 7. As tabelas _____ contêm chaves primárias e estrangeiras para atributos associados de um modelo de dados.
 - a) Base de dados
 - b) Dimensão
 - c) Fato
 8. O esquema _____ requer menos armazenamento e tem menos _____ no processo de normalização.
 - a) Snowflake; Redundância
 - b) Estrela; Redundância
 - c) Snowflake; Confiança
 9. _____ é uma abordagem de banco de dados alternativa que utiliza bancos de dados não relacionais e não estruturados.
 - a) PSQL
 - b) NoSQL
 - c) MySQL
 10. _____ fornece uma abordagem descentralizada para armazenamento e análise de dados.
 - a) Data Warehouse
 - b) Data Lake
 - c) Big Dat

3. A Natureza dos Dados

3.1. Análise de dados

A estatística desempenha um papel significativo nas ciências físicas e sociais. É sem dúvida o ponto mais saliente da interseção entre diversas disciplinas. É a linguagem comum da ciência. Cientistas usam estatísticas para converter dados em informações úteis. Estatística existe para um processo, onde estamos coletando dados, resumindo dados e interpretando dados. O processo de estatística começa quando identificamos qual grupo queremos estudar ou aprender algo. Chamamos este grupo de população.

A palavra população não é apenas usada para se referir a pessoas. É usado em um sentido estatístico mais amplo. Onde a população se refere a um grupo inteiro no qual você deseja se concentrar. Pode ser um grupo inteiro de pessoas ou animais ou insetos, ou objetos inanimados como prédios de apartamentos ou crateras em Marte. Por exemplo, podemos estar interessados nas opiniões da população adulta dos EUA sobre a pena de morte. Como a população de ratos reage a um determinado produto químico. O preço médio da população de todos os apartamentos de um quarto em uma determinada cidade. População, então, é todo o grupo que é o alvo de nosso interesse. Na maioria dos casos, a população é tão grande, que por mais que quisermos, não há absolutamente nenhuma maneira de podermos estudar tudo isso.

Uma abordagem mais prática seria examinar e coletar dados apenas de um subgrupo da população, que chamamos de amostra. Chamamos este primeiro passo que envolve a escolha de uma amostra e coleta de dados dele, produzindo dados. Uma vez que, por razões práticas, precisamos comprometer e examinar apenas um subgrupo da população em vez de toda a população, devemos fazer um esforço para escolher uma amostra de tal forma que ela representará a população também.

James Thompson tem estudado o sucesso da polinização de uma flor obscura que floresce em altas altitudes, o lírio glaciado. Estamos olhando para uma espécie de alta altitude. E como as pessoas interessadas em mudanças globais e climas perceberam que é aqui que podemos primeiro ver coisas como espécies fazendo mal porque as situações mudaram, e, de fato, há muito foco na pesquisa de alta elevação em o contexto geral da mudança climática. Existem milhões destas flores em toda a Rockys e o professor Thompson não pode estudá-las todas, ele tem que escolher uma amostra. Se eu estou dizendo alguma coisa precisa esses dados têm que realmente refletir o que está acontecendo aqui. Então, por exemplo, quando eu olho para uma amostra de flores eu tenho que estar pensando o tempo todo sobre se eu estou selecionando um conjunto de flores que é adequadamente representativo de todo o que eu quero falar.

Isto é verdade se estamos estudando flores, crateras em Marte, ou as opiniões de adultos norte-americanos sobre a pena de morte. Nossa amostra não representaria adultos dos EUA, se perguntassemos apenas aos republicanos ou apenas perguntassem aos democratas. Tal amostra não representaria a população. Os conjuntos de dados podem ser muito diferentes dependendo do que está sendo estudado. Estes dados podem assumir a forma de respostas a perguntas de pesquisa, tabelas de números, como detalhes da cratera, ou, no caso de lírios glaciares, observações coletadas ao longo de muitos anos. Essencialmente, eu fiz uma pergunta muito simples. O que faz com que uma flor seja um sucesso? Será que ele é polinizado? Será que ela define uma fruta? Faz sementes? Quantas sementes faz?

Para dar sentido a esses dados, eles precisam ser resumidos de forma significativa. Isso é chamado de análise exploratória de dados. Análise exploratória de dados muitas vezes revela novas maneiras de pensar sobre os dados. Como acontece frequentemente na ciência, quanto mais cuidadosamente eu olhava, mais coisas eu via para me interessar. A análise exploratória de dados ajuda os cientistas a refinar suas perguntas. E às vezes até revelam perguntas inteiramente novas.

Se os climas estão mudando, as relações entre plantas e polinizadores e outras relações mutuamente benéficas, essas relações podem ser interrompidas. Cientistas que estudam a mudança climática têm muitas vezes se perguntado que efeito um clima de aquecimento terá sobre a relação entre plantas e animais. É possível que pequenas mudanças no clima possam ter um grande impacto nessas relações? Análise exploratória de dados sugere que essa é uma pergunta que o Professor Thompson pode ser capaz de responder usando 30 anos de dados.

Isso leva até a etapa final, a inferência. O que podemos inferir sobre a população como um todo a partir dos dados em nossa amostra? Lembre-se, após análise exploratória de dados, somos capazes de fazer perguntas específicas sobre nossos dados. Inferência é onde chegamos círculo completo com a esperança de revelar novos conhecimentos sobre a população.

Então, o que o Professor Thompson pode inferir sobre Lírios Glaciares? O que seus dados revelaram é que os lírios glaciares e as abelhas que os polinizam estão se separando no tempo. À medida que o clima aquece, os lírios florescem mais cedo antes das abelhas chegarem:

“Meu artigo sobre lírios glaciares, tanto quanto posso dizer, é a primeira demonstração dele ou a primeira demonstração mesmo plausível dele. Não é uma coisa fácil de mostrar. É uma coisa fácil de dizer, ei, isso pode acontecer. Meus conjuntos de dados de longo prazo me permitiram fazer é dizer, sim, e parece que aconteceu.”

James Thompson foi capaz de explorar seus dados para mostrar como mudanças climáticas faz com que plantas e animais se desconectem no tempo. Você também estará olhando para grandes conjuntos de dados e fazendo novas perguntas de interesse para você. Você não criará novos dados, mas criará novos conhecimentos através da análise de dados exploratória e análise de dados inferenciais. A educação estatística é mais frequentemente conduzida dentro de um contexto específico de disciplina ou como treinamento matemático genérico.

3.2. Dados e Tipos de Dados

O que realmente queremos dizer com dados? Simplificando, dados são pedaços de informação sobre indivíduos organizados em variáveis. Por indivíduo, queremos dizer uma unidade de observação. Uma observação ou unidade de observação refere-se a uma determinada pessoa ou um objeto específico, qualquer unidade específica de observação dentro de sua amostra de estudo. Os dados fornecem a base para inteligência de negócios, análise de negócios e ciência de dados. Como tal, é importante entender os vários tipos de dados que podem ser coletados, explorados, analisados e visualizados.

Por uma variável precisamos de uma característica particular da unidade de observação. No nível da pessoa, podemos coletar dados sobre Altura, Peso, Sexo, Corrida etc. Se estamos coletando dados em uma amostra de carros, podemos medir variáveis como Cor, Tamanho do Pneu, Quilometragem, Modelo e Número de assentos etc. Se nossa amostra incluir cidades, podemos medir variáveis como Tamanho da população, Receita Fiscal, Consumo de Energia, Número de Hospitais e assim por diante.

Um conjunto de dados é composto de observações e variáveis individuais. Os conjuntos de dados são normalmente exibidos em tabelas nas quais as linhas representam indivíduos, ou unidades de observação, e as colunas representam variáveis. Aqui está um conjunto de dados que mostra registros médicos de uma pesquisa. Neste exemplo, as unidades de observação são pacientes e as variáveis são Sexo, Idade, Altura, Peso, Fumar e Raça. Cada linha nos dá todas as informações sobre uma observação específica. Neste caso, um paciente. E cada coluna nos dá informações sobre uma característica particular de todos os pacientes.

Dados estruturados vs. não estruturados

Dados estruturados são dados bem definidos com padrões facilmente identificáveis. Alguns exemplos familiares são números de telefone e endereços de correspondência. Você pode discernir facilmente as partes dessas informações (dados) porque entende seu padrão e formato distintos. A natureza organizada e estruturada desses dados também os torna facilmente pesquisáveis. Os dados estruturados normalmente estão presentes em um sistema de gerenciamento de banco de dados (relacional) (RDBMS ou DBMS), que discutiremos mais no próximo módulo.

Dados não estruturados são entendidos como “todo o resto”, ou seja, dados em que os padrões não surgem facilmente e nem sempre podem se encaixar em um formato padrão. Exemplos típicos incluem arquivos de áudio, arquivos de vídeo e postagens de mídia social. Embora os dados não estruturados possam ser armazenados em vários formatos em um RDBMS, geralmente é mais comum encontrar dados não estruturados em um banco de dados não relacional ou armazenado em um sistema de arquivos.

A análise de dados estruturados é um processo bem definido e maduro, enquanto a análise de dados não estruturados é fortemente investida em pesquisa e desenvolvimento e na descoberta de novas tecnologias para analisar tipos de dados complexos com mais eficiência. Devido às complexidades inerentes à análise de dados não estruturados, essa análise requer muito mais tempo e poder de processamento. Consulte a tabela de comparação de dados estruturados versus não estruturados vinculada aqui para obter detalhes adicionais.

Dados Estruturados e Dados Não-estruturados		
	Estruturado	Não-estruturado
Características	<ul style="list-style-type: none"> Modelo de dados pré-definidos Tipicamente textual Facilmente pesquisável Facilmente identificável por padrões 	<ul style="list-style-type: none"> Modelo de dados não estabelecido Pode ser texto, imagem, som, vídeo, etc Difícil de pesquisar Difícil de identificar padrão
Reside em	<ul style="list-style-type: none"> Bancos de dados relacionais Data Warehouses 	<ul style="list-style-type: none"> Aplicações Bancos de dados NoSQL Data Warehouse Data Lakes
Exemplos	<ul style="list-style-type: none"> Número de telefone Endereço de e-mail Número do CPF Informação de transação 	<ul style="list-style-type: none"> Imagens Audio Vídeo Web e mídia social

Dados Quantitativos x Qualitativos

Agora, vamos considerar algumas das diferenças entre dados quantitativos e qualitativos.

Variáveis também podem ser classificadas em um dos dois tipos, Quantitativo ou Categórico (ou qualitativos). Variáveis quantitativas tomam valores numéricos e representam algum tipo de medição. Variáveis categóricas, por outro lado, tomam valores de categoria ou e colocam uma observação ou indivíduo em um dos vários grupos. Neste exemplo, existem várias variáveis de cada tipo. Idade, peso e altura são variáveis quantitativas. Raça, Sexo e Fumar são variáveis categóricas.

Quantitativo: Discreto e Contínuo

Os dados quantitativos são estruturados e estatísticos e, portanto, podem ser contados, medidos e expressos usando números e cálculos. Esse requisito permite a facilidade de computação, agregação e análise.

Dois tipos principais de dados quantitativos são dados discretos e contínuos.

- **Dados discretos** são dados que não podem ser divididos em partes menores. Portanto, existe um conjunto finito de valores que podem ser aplicados. Dados discretos normalmente incluem números inteiros ou inteiros.
- Os **dados contínuos** podem ser divididos em partes menores e têm o potencial de flutuar continuamente.

Qualitativo: Nominal & Ordinal

Os dados qualitativos (ou categóricos) são de natureza descritiva e conceitual. É não estatístico e normalmente não estruturado ou semiestruturado. Os dados qualitativos são frequentemente categorizados usando traços e características. Geralmente é aberto e pode ajudar a responder à pergunta “Por quê?” No entanto, para fins de análise, os valores qualitativos geralmente precisam ser convertidos ou mapeados em dados numéricos.

Dois tipos principais de dados qualitativos são dados nominais e ordinais.

- Os **dados nominais** consistem em valores que não possuem ordem natural. Por exemplo, o gênero de uma pessoa não pode ser classificado como superior ou inferior a qualquer outro gênero.
- Os **dados ordinais** têm uma ordem natural e podem ser categorizados por agrupamentos de ordem. Os tamanhos das camisas são um ótimo exemplo de ordem em que grande > médio > pequeno.

Observe que os valores da variável categórica FUMANTE podem ser codificados como zero ou um. É bastante comum codificar os valores de uma variável categórica como números. Mas você deve sempre lembrar que estes são apenas códigos. Muitas vezes referido como *Códigos Dummy* (códigos fictícios) porque eles não têm significado aritmético. Ou seja, não faz sentido adicioná-los, subtraí-los, multiplicá-los ou dividi-los. Ou até mesmo comparar a magnitude desses valores.

IDs

Finalmente, um identificador exclusivo é uma variável que se destina a distinguir cada uma das unidades de observação do seu conjunto de dados. Exemplos podem incluir números de série para dados sobre um determinado produto, números de segurança social para dados sobre uma pessoa individual. Ou talvez números aleatórios gerados para qualquer tipo de observação. Para nos ajudar a organizar nossos dados, cada conjunto de dados deve ter uma variável que identifique exclusivamente as observações. Esta variável é particularmente útil se você precisar mesclar informações em diferentes conjuntos de dados.

3.3. Datasets e Codebooks

Alguns dos conjuntos de dados disponíveis para o curso incluem o Estudo Longitudinal Nacional de Saúde de Adolescentes, comumente conhecido como Add Health. Esta é uma pesquisa nacional representativa baseada na escola. A onda um da pesquisa incluiu adolescência nos graus 7 a 12 em os Estados Unidos. O Add Health inclui dados de pesquisa sobre bem-estar social, econômico, psicológico e de adolescentes. Em seguida é o estudo das crateras de Marte. Como você deve saber, o planeta Marte tem terreno fortemente craterizado. Estas crateras foram criadas há cerca de 4 bilhões de anos durante um período de bombardeamento pesado de asteroides, protoplanetas e cometas. Disponibilizado por pesquisadores da Universidade do Colorado Boulder, este conjunto de dados inclui características de mais de 350.000 dessas crateras de Marte. Também está disponível uma parte do Wave 1, Estudo Epidemiológico Nacional de Álcool e Condições Relacionadas, comumente conhecido como NESARC. Esta é uma amostra representativa da população adulta dos EUA com idade igual ou superior a 18 anos. E inclui dados sobre saúde mental e distúrbios do uso de substâncias que são experimentados por adultos. Outro conjunto de dados é o conjunto de dados GapMinder, que é disponibilizado por gapminder.org. Inclui numerosas medidas de 195 países. Os dados foram coletados de várias fontes, incluindo a Organização Mundial de Saúde, a Agência Internacional para Research on Cancer, as Nações Unidas e o Banco Mundial.

Para ajudá-lo a aprender mais sobre esses conjuntos de dados e em qual deles você está mais interessado, você estará revisando os códigos disponíveis desses conjuntos de dados. Às vezes chamados de dicionários de dados, os codebooks geralmente oferecem informações completas sobre o conjunto de dados. Isso é tópicos gerais abordados, perguntas e/ou medidas usadas para registrar cada uma das variáveis. E em alguns casos, a frequência de respostas ou valores de cada uma das variáveis.

Rever um livro de códigos é sempre o primeiro passo na pesquisa com base em dados existentes. Primeiro de tudo, os livros de código podem ser usados para gerar perguntas de pesquisa. Em segundo lugar, os dados são muitas vezes inúteis e completamente impossível de interpretá-los sem eles.

O livro de códigos descreve como os dados são organizados no arquivo do computador. O que significam os vários números e letras, e quaisquer instruções especiais sobre como usar os dados corretamente. Como qualquer outro livro, alguns codebooks são melhores do que outros. No livro de códigos Add Health, cada variável tem uma descrição do que é medido. Neste caso, é a questão de qual nota você está.

Um livro de código também incluirá as várias opções de medição ou resposta. Para esta variável, possíveis opções de resposta incluem 7^a a 12^a série, recusa em responder à pergunta, um salto legítimo para aqueles que não estão na escola, não sabem e a escola não tem os níveis da série, ou a pergunta não é aplicável. Além de incluir uma listagem ou descrição das opções de resposta para a variável, o livro de códigos também incluirá valores correspondentes que podem ser encontrados no conjunto de dados.

Como vimos anteriormente com o exemplo do conjunto de dados de registros médicos, conjuntos de dados normalmente incluem números em vez de palavras. Assim, para variáveis categóricas, como nível de grau, cada uma das opções de resposta tem um valor numérico correspondente. É esse valor numérico que pode ser encontrado no conjunto de dados. Você pode ver que os alunos da 7^a a 12^a série são logicamente codificados como os números 7 a 12.

- 96 indica que o adolescente **se recusou a responder**.
- 97 indica um salto legítimo para os adolescentes que **não estão atualmente na escola**.
- 98 indica que **não sei**.
- 99 é gravado em um caso em que **a escola não tem níveis de série**.

Estes valores numéricos são conhecidos como códigos fictícios, como estão incluídos no conjunto de dados, mas não têm significado numérico direto. No livro de código das crateras de Marte, encontramos uma descrição das variáveis para nome, latitude, longitude e diâmetro da cratera. Também a profundidade da borda da cratera, bem como o nome da variável no conjunto de dados. Como a maioria dessas variáveis são quantitativas, em vez de listar uma opção de resposta, o livro de códigos inclui uma descrição de como a variável é medida. Por exemplo, a latitude é medida em graus decimais Norte, longitude é medida em graus decimais Leste e o diâmetro e a profundidade são medidos em quilômetros.

Do dataset Gapminder, vemos uma aparência ligeiramente diferente do livro de código, mas características muito semelhantes. Você pode ver que a coluna do meio descreve cada uma das variáveis. A coluna à esquerda indica o nome da variável usada no conjunto de dados. E, finalmente, a coluna da direita lista a fonte de dados. Novamente, estas são variáveis quantitativas. O livro de códigos inclui informações sobre como cada uma dessas variáveis foi medida. Você pode ver que a renda por pessoa é medida em dólares americanos. O consumo de álcool é medido em litros de álcool puro. Forças de trabalho é medida como a porcentagem da força de trabalho total, e taxa de câncer de mama é medida como novos casos por 100.000 mulheres.

3.4. Desenvolvendo uma questão de pesquisa

Uma vez que você tenha uma compreensão geral acerca dos conjuntos de dados, tipos de variáveis e livros de código, o próximo passo é selecionar um conjunto de dados. Selecione um conjunto de dados que inclua variáveis em uma área que lhe interessa.

Depois de selecionar os conjuntos de dados, identifique um tópico específico de interesse e imprima as páginas do livro de códigos que incluem a variável ou as variáveis que medem o tópico selecionado. Note que muitos livros de código são muito grandes para imprimir, por isso é muito importante criar o seu próprio livro de código pessoal com apenas as páginas que incluem as variáveis que você gostaria de examinar.

Nosso exemplo vem do conjunto de dados NESARC, e nosso tópico escolhido é a dependência da nicotina. Existem várias variáveis relacionadas à Dependência de Nicotina, e podemos ver 2 aqui: dependência de nicotina ao longo da vida e dependência de nicotina nos últimos 12 meses. Um valor de zero para estas variáveis indica que não há Dependência de Nicotina, e um valor de 1 indica a presença de Dependência de Nicotina. O nome dessas variáveis são TAB12MDX e TABLIFEDX. Usaremos esses nomes de variáveis quando começarmos a trabalhar com os dados.

Não estamos sugerindo que este tópico seja mais ou menos interessante, ou mais ou menos importante do que qualquer outro. O que é importante é que você escolha um tópico que é de seu interesse. Escolhemos analisar a dependência da nicotina.

Depois de ter um tópico e ter impresso as páginas do livro de código que medem esse tópico, é hora de criar uma pergunta de pesquisa. Uma das perguntas de pesquisa mais simples que podem ser feitas é se dois tópicos estão associados um ao outro. Por exemplo, a procura de tratamento médico está associada à renda? A profundidade da cratera está associada ao diâmetro da cratera? A fluoração da água está associada ao número de cavidades durante visitas ao dentista? Esses conjuntos de dados são vastos, portanto, há muitas associações potenciais para explorar. Vamos olhar para o nosso exemplo escolhido: dependência de nicotina.

Primeiro eu preciso determinar o que é sobre a dependência de nicotina que me interessa. Parece-me que amigos e conhecidos que eu conheci ao longo dos anos, que ficou viciado em cigarros o fizeram em períodos muito diferentes de tempo. Alguns pareciam ser dependentes de fumar fortemente logo após sua primeira experiência com um cigarro, e outros depois de muitos anos de comportamento geralmente irregular de fumar.

Decidimos que estamos mais interessados em explorar a associação entre o comportamento do tabagismo e a dependência da nicotina. Acreditamos que eles estão positivamente associados. Ou seja, quanto mais um indivíduo fuma, mais provável é que seja dependente da nicotina. Também estamos nos perguntando o quanto uma pessoa precisa fumar para ser dependente da nicotina.

Continuamos a ler o livro de códigos NESARC e descobrimos que o comportamento de fumar também foi medido nesta amostra. Então, em seguida, eu dou um passo semelhante a um que eu acabei de tomar ao escolher a dependência de nicotina. Ou seja, identifique as variáveis que medem o segundo tópico, comportamento de tabagismo, no meu conjunto de dados. As variáveis que escolho incluem status de tabagismo, frequência usual, e quantidade usual.

Durante sua segunda revisão do livro de códigos para o conjunto de dados que você selecionou, você também deve identificar um segundo tópico que você gostaria de explorar em termos de associação com seu tópico original. E, novamente, imprima as páginas do livro de códigos que incluem a variável, ou variáveis, que medem o segundo tópico selecionado.

4. Estatística

Em sua essência, a estatística é uma análise técnica de dados baseada em matemática usando vários testes e análises. Embora não possamos aprofundar muito neste curso, é importante considerarmos as metodologias estatísticas que são usadas na análise de dados. Para os propósitos deste curso, exploraremos brevemente dois métodos principais: estatística descritiva e estatística inferencial.

4.1. Estatística descritiva

A estatística descritiva permite a sumarização e a representação gráfica de um conjunto de dados. A natureza descritiva das informações resultantes permite que um analista descreva uma amostra da população do conjunto de dados. (Observe que isso não nos permite generalizar uma população inteira ou inferir atributos ou propriedades da população.)

Geralmente usamos estatística descritiva para explorar:

- **Tendência central** (média): média, mediana ou moda para explicar as médias de um ponto de dados
- **Dispersão**: intervalo e desvio padrão para descrever a distância da média ou distância entre os valores de dados mais altos e mais baixos
- **Skewness (distorção)**: descreve a natureza simétrica ou assimétrica do conjunto de dados
- **Correlação**: explora as relações entre as variáveis no conjunto de dados de amostra

4.1.1. Análise Exploratória de Dados

Os dados brutos consistem em longas listas de números e rótulos que não parecem ser muito informativos. Dados brutos carece de contexto. Análise exploratória de dados é o que você usa para entender os dados. Você faz isso convertendo dados de sua forma bruta, em um formulário que faz sentido, que tem contexto, que conta a história que você quer contar.

Basicamente, a análise exploratória de dados consiste em organizar e resumindo dados brutos, procurando características e padrões importantes em os dados, procurando quaisquer desvios marcantes desses padrões, e interpretando suas descobertas no contexto do problema ou questão de pesquisa. Começaremos a análise exploratória de dados analisando uma variável de cada vez, também chamada de análise univariada.

Para converter dados brutos em informações úteis, precisamos resumir e, em seguida, examinar a distribuição de quaisquer variáveis de interesse. Por distribuição de uma variável, queremos dizer quais valores a variável toma, e com que frequência a variável leva esses valores.

Se estivéssemos estudando um pequeno número de observações, poderíamos fazer isso com um lápis e papel, uma calculadora, ou mesmo em nossas cabeças. Os conjuntos de dados com os quais você está trabalhando, muitas vezes têm milhares de observações. Trabalhar com amostras tão grandes só é possível se usarmos software estatístico. Esses programas de software exigem o uso de sintaxe ou código formal para recuperar, analisar e manipular dados. Aprender a escrever código, aprender o uso adequado da sintaxe pode realmente expandir sua capacidade de se envolver em aplicativos estatísticos. Essa habilidade também expandirá muito sua capacidade de se engajar em níveis mais profundos de raciocínio quantitativo sobre dados.

Para este curso, você estará usando Python. Python é uma linguagem de uso geral amplamente utilizada que é projetado para ser mais legível. Ou seja, o código é mais fácil de ler e escrever do que em outras linguagens de uso geral, como C++ ou Java. Embora o Python não tenha sido desenvolvido especificamente para análise de dados pandas e outras bibliotecas fornecem ferramentas de análise de dados para uso com a linguagem Python. Olhando para todas as janelas, opções, menus e recursos embora, pode ser bastante assustador. Portanto, é importante para você perceber, este curso irá apresentá-lo ao básico. Você aprenderá o que precisa saber para começar a perguntar e respondendo perguntas interessantes sobre dados. >> No início, você pode sentir que está

aprendendo outro idioma. Basicamente, é. À medida que você trabalha em seu projeto, você deve começar a se sentir mais confortável implementando as várias decisões que você vai tomar sobre os dados.

4.1.2. Examinando a distribuição de frequência

A análise exploratória de dados começa olhando em uma variável de cada vez. Isso é chamado de univariado ou análise descritiva. Para converter dados brutos em informações úteis, precisamos resumir e examinar a distribuição de qualquer variável de interesse. As variáveis de interesse são as variáveis de interesse para você, pesquisador. Ao responder suas perguntas de pesquisa, abordando seu problema de pesquisa e contando a história que você deseja contar com sua pesquisa. Por distribuição de uma variável, queremos dizer quais valores a variável leva e com que frequência a variável leva esses valores.

Aqui está um exemplo. Em uma amostra aleatória de 1.200 estudantes universitários dos EUA convidados a responder as seguintes perguntas como parte de uma pesquisa maior: Qual é a sua percepção do seu próprio corpo? Você sente isso acima do peso? Razoável? Ou abaixo do peso? Esta tabela mostra parte dos dados, cinco das 1.200 observações.

Informações que seriam interessantes obter a partir desses dados inclui que porcentagem dos alunos da amostra se enquadram em cada categoria ou como os alunos são divididos ao longo dos três tipos de imagens? Eles estão igualmente divididos? Se não, faça as percentagens seguir algum tipo de padrão? Não há como responder a essas perguntas por olhando para os dados brutos, que estão na forma de uma longa lista de 1.200 respostas.

Isso não é muito útil. No entanto, todas essas perguntas serão facilmente respondidas quando resumirmos e observarmos a distribuição de frequência da imagem corporal variável. Isto é, uma vez que resumimos com que frequência cada uma das categorias ocorre. Para resumir a distribuição de um variável categórica, primeiro criamos uma tabela dos diferentes valores ou categorias que a variável assume.

Quantas vezes cada ocorre a variável, que é a contagem, e, mais importante, com que frequência cada variável ocorre, o que é expresso convertendo as contagens em percentagens. Agora que resumimos a distribuição da variável de imagem corporal, vamos voltar e interpretar os resultados no contexto das perguntas que postamos.

Qual porcentagem de os alunos da amostra se enquadram em cada categoria? Como os alunos são divididos em três corpos categorias de imagens, e elas estão igualmente divididas? Você pode ver isso a maioria das amostras, ou seja, 71,3% sentida que seu peso estava quase certo e que uma pequena porcentagem sentiu-se abaixo do peso em 9,2%. A categoria sobre peso foi de 19,6%.

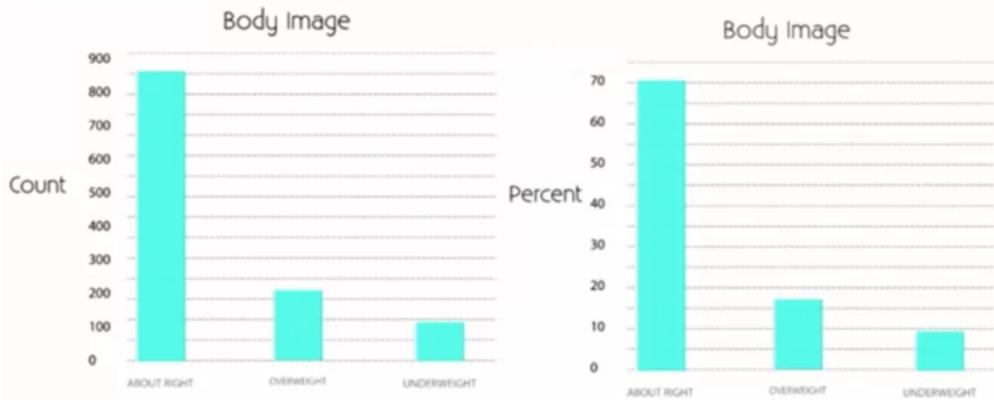
4.1.3. Plotando as distribuições

Ferramentas de visualização são importantes meios para ampliar a compreensão acerca do comportamento dos dados. Para começar a visualizar nossas variáveis com gráfico, iniciaremos com gráficos com uma variável de cada vez, usaremos isso como um trampolim para visualizar várias variáveis simultaneamente com gráficos internos.

Acompanhe o exemplo abaixo através do script disponibilizado “**plotando_distrib**”.

Os gráficos de barras são mais comumente usados examinar a distribuição de variáveis individuais. Considere uma distribuição para a amostra aleatória de 1.200 estudantes universitários americanos que foram questionados sobre o que é a sua percepção do seu próprio corpo. Neste gráfico de barras, o eixo X ou horizontal inclui as três categorias de resposta. Abaixo do peso, acima do peso e quase certo. No primeiro gráfico de barras, a altura das barras é medida no eixo Y, ou vertical, como o número ou contagem de estudantes universitários dando cada resposta. O segundo gráfico de barras mostra os mesmos dados, mas como uma porcentagem da

amostra total. Um gráfico de barras nos ajuda a exibir a distribuição de uma variável categórica, por exemplo, porcentagem de observações em cada categoria.

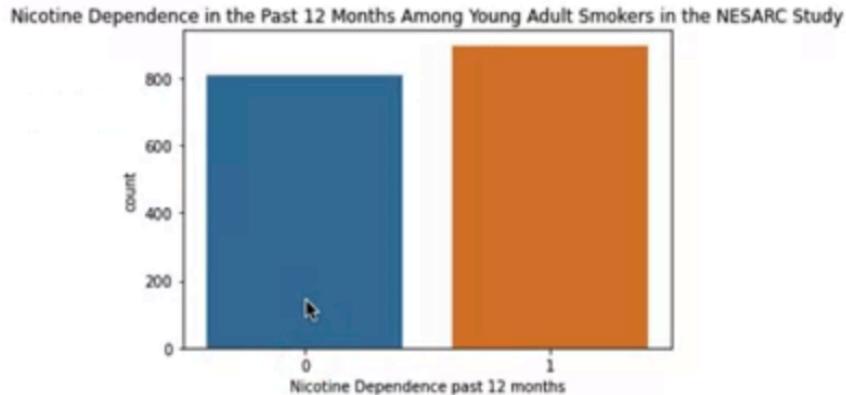


Para ilustrar, usaremos o dataset NESARC, buscando interpretar as relações entre a dependência de nicotina nos últimos meses (representado no dataset pela variável TAB12MDX) e a estimativa de cigarros fumados por mês (representado por NUMCIG_EST). Vamos executar distribuições de frequência para cada uma dessas variáveis, incluindo contagens e percentagens. Vou usar a função groupby para isso que também apresentamos quando introduzindo distribuições de frequência. Além das distribuições de frequência, também queremos examinar os gráficos de barras correspondentes para essas duas variáveis também. O gráfico de barras é uma das mais visualizações gráficas frequentemente usadas.

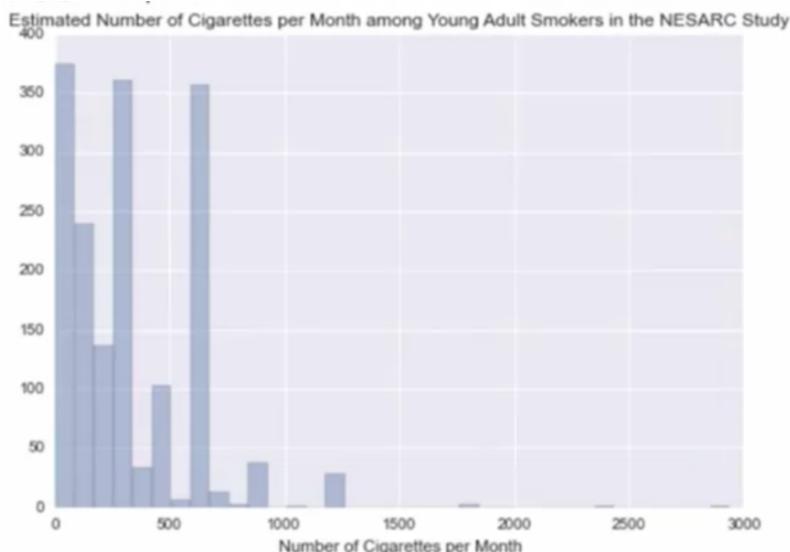
Ao visualizar dados em Python, precisaremos importar bibliotecas em nosso programa. Primeiro, vamos importar o seaborn pacote com a sintaxe **import seaborn**. Também precisamos importar a biblioteca **matplotlib.pyplot** porque o pacote seaborn é dependente neste pacote para criar gráficos. Porque o nome deste pacote é tão longo, daremos a ele o apelido **plt**, que pode ser usado no lugar de o nome completo do pacote quando escrevemos o código chamando isso pacote em nosso programa. Nós vamos mantê-lo simples. Usaremos o código Python para gerar gráficos que nos ajudam a aprender mais sobre nossos dados e a tomar decisões sobre próximos passos de nossa pesquisa.

Estamos focando na função de visualizações gráficas em vez de produzir imagens polidas e prontas para apresentações gráficas neste momento. Variáveis categóricas podem ser visualizadas um de cada vez com os gráficos univariados, ou seja, com gráficos de barras de variável única. Em primeiro lugar, a fim de categórico variáveis sejam ordenadas corretamente no eixo horizontal ou X de uma variável univariada gráfico, você deve converter suas variáveis categóricas, que geralmente são formatados como variáveis numéricas, em um formato que o Python reconhece como categórico. Aqui está o código. Aqui estou usando o **astype** função para converter TAB12MDX em uma variável categórica, mantendo o nome da variável original como está.

O código básico para um gráfico univariado de uma variável categórica é a seguinte. Com a função de gráfico de contagem, nomeamos a variável categórica para o eixo X e para encontrar o quadro de dados aqui, sub2. Com a função **xlabel**, podemos rotular o eixo X, e com a função **title**, fornecer o gráfico de barras com um título. Aqui está o código do gráfico de barras univariado inserido em nosso programa de exemplo, e salvamos e executamos o programa para gerar o gráfico de barras solicitado. Podemos visualizar o gráfico clicando em a guia de plotagens para abrir o painel de plotagens. Isto mostrará o número de jovens adultos fumantes com dependência de nicotina, 896, indicado por um código de resposta de 1. E aqueles sem dependência de nicotina, 810, indicado por um 0.



Agora vamos exibir graficamente a distribuição de frequência para uma de nossas variáveis de tabagismo gerenciadas por dados, ou seja, o número estimado de cigarros fumava por mês, **NUMCIGMO_EST**. Porque **NUMCIGMO_EST** é na verdade uma variável quantitativa, a sintaxe que usamos no Python programa é um pouco diferente. Para visualizar uma variável quantitativa, você usaria a seguinte sintaxe. Com a função de plotagem de distribuição, ou **distplot**, nomeamos a variável quantitativa para o eixo X e peça ao Python para descartar os dados ausentes. Isso é as NaNs. Também incluímos a opção **kde=False**. Novamente da biblioteca matplotlib.pyplot, que estamos chamando de plt, usamos o rótulo X para rotular o eixo X com direito a fornecer o gráfico com o título. Ao executar isso, você verá que o programa gera uma distribuição gráfica da variável quantitativa. Gera um histograma. Em um histograma, intervalos de valores são plotados no eixo X em vez de valores discretos ou separados. Das barras aqui, você pode ver que o que é exibido é o ponto médio dos intervalos.



Vamos olhar para um exemplo mais básico de como um histograma pode ser construído, e então usar isso como um trampolim para falar sobre estatísticas descritivas adicionais que podem ser geradas para variáveis quantitativas. Neste exemplo, temos as notas de exame de 15 alunos.

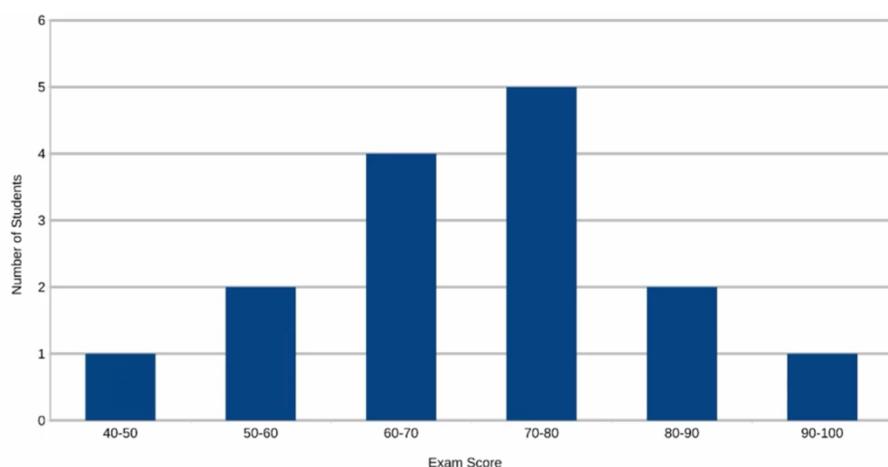
88, 48, 60, 51, 57, 85, 69, 75, 97, 72, 71, 79, 65, 63, 73

Primeiro precisamos dividir o intervalo de valores em intervalos. Também chamado de compartimentos, grupos ou classes. Neste caso, uma vez que o nosso conjunto de dados consiste em pontuações de exames, fará sentido escolher intervalos que normalmente correspondam ao intervalo de notas de letra. Então dez pontos de largura, 40 a 50, 50 a 60, etc. Ao contar quantos das 15 observações caem em cada um dos intervalos, obtemos esta tabela.

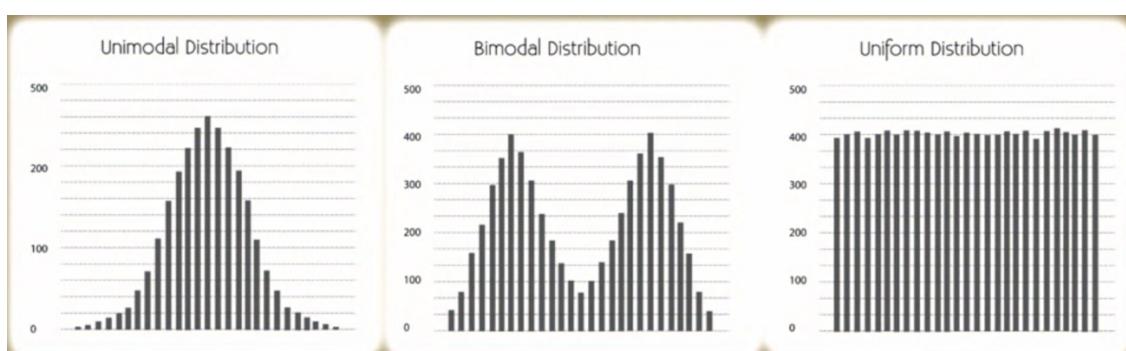
Para construir o histograma a partir desta tabela, os intervalos são plotados no eixo X e mostram o número de observações em cada intervalo, ou a porcentagem de observações em cada intervalo no eixo Y, que é representada pela altura da barra localizada acima do intervalo.

Pontuação	Ocorrências
40-50	1
50-60	2
60-70	4
70-80	5
80-90	2
90-100	1

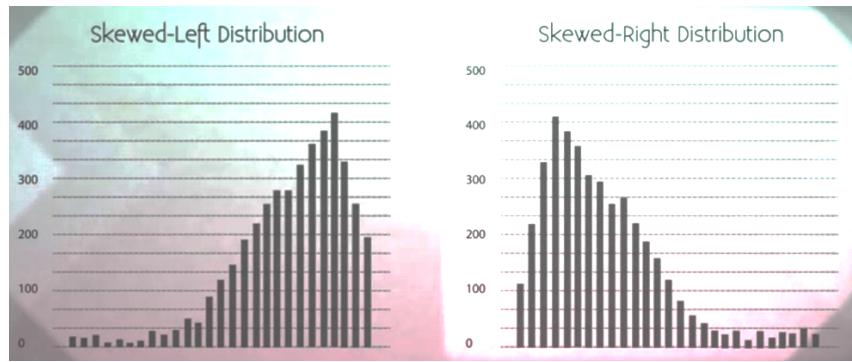
Uma vez que a distribuição tenha sido exibida graficamente como um histograma, podemos descrever o padrão geral da distribuição e mencionar quaisquer desvios marcantes desse padrão. Mais especificamente, devemos considerar os seguintes recursos. Teremos uma noção do padrão geral dos dados do centro dos histogramas, da dispersão e da forma, enquanto os outliers destacarão desvios desse padrão.



Ao descrever a forma de uma distribuição, devemos considerar simetria ou assimetria da distribuição e pico ou modalidade. Ou seja, o número de picos ou modos que a distribuição tem. Aqui, todas as três distribuições seriam referidas como simétricas. Mas eles são diferentes em sua modalidade ou pico. A primeira distribuição é unimodal. Ele tem um modo, aproximadamente em 10, em torno do qual as observações estão concentradas. A segunda distribuição é bimodal. Tem dois modos, aproximadamente em 10 e 20, em torno dos quais as observações estão concentradas. A terceira distribuição é tipo plana ou uniforme. A distribuição não tem modos, ou nenhum valor em torno do qual as observações estão concentradas. Em vez disso, as observações são distribuídas aproximadamente uniformemente entre os diferentes valores.

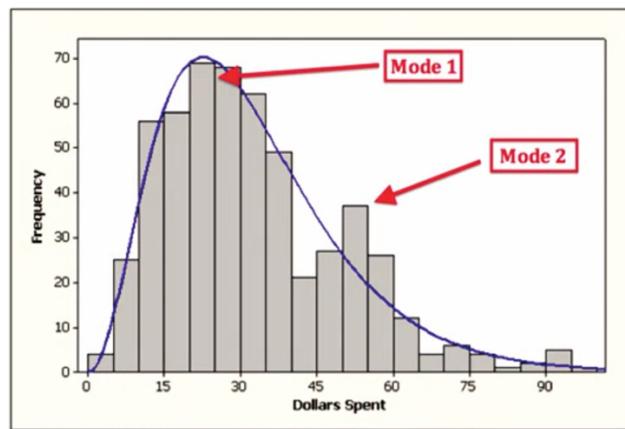


Uma distribuição é chamada skewed-right (distorcida à direita). Com a cauda direita, os valores maiores são muito mais longos do que a cauda esquerda, ou valores menores. Note que em uma distribuição assimétrica à direita, como você pode ver aqui à direita. A maior parte das observações é pequena a média, com algumas observações que são muito maiores do que o resto. Um exemplo de uma variável da vida real que tem uma distribuição assimétrica à direita é o salário. A maioria das pessoas ganha na faixa baixa a média de salários com algumas exceções, como CEOs, atletas profissionais, etc. Que são distribuídos ao longo de uma ampla gama, que é a longa cauda de valores mais elevados.



Uma distribuição é chamada skewed-left (distorcida à esquerda) se a cauda esquerda ou valores menores forem muito maiores do que a cauda direita ou valores maiores. Não que em uma distribuição assimétrica à esquerda, a maior parte das observações seja de média a grande, com algumas observações que são muito menores do que as restantes. Um exemplo de uma variável da vida real que tem uma distribuição distorcida à esquerda é a idade da morte por causas naturais. A maioria das mortes por causas naturais ocorre em idades mais velhas, com menos casos acontecendo em idades mais jovens.

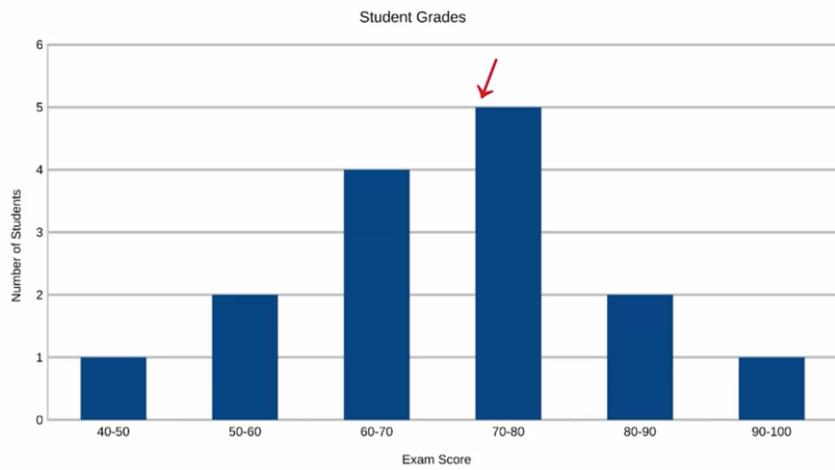
Distribuições distorcidas também podem ser bimodais. Aqui está um exemplo, um bairro de tamanho médio 24 horas loja de conveniência coletou dados de 537 clientes sobre a quantidade de dinheiro que gastaram em uma única visita à loja. Observe o histograma abaixo. Você pode ver que a quantidade de dinheiro gasto é concentrada em torno de US \$20, e, em seguida, concentrado novamente em torno de US \$50.



A moda de uma variável são os valores que ocorrem com mais frequência. E saber disso pode ajudá-lo a tomar melhores decisões. A moda, por exemplo, tem aplicativos na publicação de livros. Não surpreendentemente, é importante para a editora imprimir mais dos livros mais populares, porque imprimir livros diferentes em números iguais causaria uma escassez de alguns livros e um excesso de oferta de outros. Da mesma forma, o modo tem aplicações na fabricação. Por exemplo, também é importante fabricar mais dos sapatos e tamanhos de sapato mais populares.

A moda nem sempre está no centro. O centro da distribuição é o seu ponto médio, o valor que divide as distribuições de modo que aproximadamente metade das observações leve valores menores e aproximadamente

metade leva valores maiores. Como você pode ver no histograma, o centro da distribuição de graus é aproximadamente 70. Podemos obter apenas uma estimativa aproximada para o centro da distribuição. Sete alunos marcaram menos de 70, e oito alunos pontuaram acima de 70. Estimativas geralmente podem ser feitas a partir do exame de um histograma.



Então, e quanto a dispersão? A dispersão da distribuição, também chamada de variabilidade, pode ser descrita pelo intervalo aproximado coberto pelos dados. De olhar para o histograma, podemos aproximar a menor observação, ou mínimo, e a maior observação, ou máximo, e assim aproximar o intervalo. Em nosso exemplo de pontuação de exame, você pode ver que o mínimo aproximado é 45, que é o meio do menor intervalo de pontuações. O máximo aproximado é 95, o meio do maior intervalo de pontuações. Então, nosso alcance aproximado é de cerca de 50 pontos. 95 menos 45. O padrão geral da distribuição da variável quantitativa é descrito por sua forma, centro e dispersão. Ao inspecionar o histograma, podemos descrever a forma da distribuição, mas como vimos, só podemos obter uma estimativa aproximada do centro e dispersão.

4.1.4. Medidas de Centralidade e Dispersão

Para descrever a distribuição de uma variável quantitativa, você também precisa de descrições numéricas precisas do centro e da dispersão. A moda é um tipo de média. Há três tipos de média e cada um nos diz algo diferente. Portanto, precisamos ter certeza de que entendemos o que cada média significa. Quando usamos o termo média, queremos dizer uma das três coisas geralmente, ou queremos dizer a média aritmética, a moda ou mediana.

É muito fácil entender a diferença entre estes, especialmente se você já jogou dardos antes. Depois de dois lotes de três dardos e no meu sexto lançamento marquei um 2, 3, 3, 12 e 13. Agora vamos ver se podemos descobrir a média aritmética, a mediana e a moda. Primeiro de tudo a média aritmética. Tomamos o total de todas as seis pontuações e dividimos pelo número de observações, e essa é a média. Se quisermos a pontuação modal simplesmente procuramos a pontuação mais comum, o número mais comum de observações. Se quisermos a pontuação mediana, escrevemos as pontuações em ordem crescente e, em seguida, procuramos o valor do meio.

Há um pequeno problema aqui que temos um número par de observações, então pegamos os dois valores médios, e calculamos a média desses dois. Então, para o meu dardo não muito bom jogando, as pontuações foram 2, 3, 3, 3, 12, 13. A média é $2+3+3+12+13$ dividido por 6, $36/6 = 6$. A moda é 3. A mediana desde que temos um número par de observações, é $3 + 3$, o meio duas observações, dividido por 2, que é igual a 3.

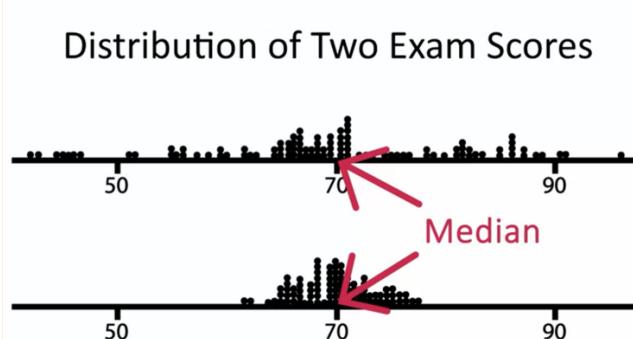
Aviso, se o jogador de dardo tivesse marcado, 19 em vez de 13. A média aumenta para 7, mas a moda e a pontuação mediana permanecem inalterados.

Então vamos rever brevemente as medidas numéricas centrais. Intuitivamente falando, a medida aritmética do centro está nos dizendo o que é um valor típico da distribuição de uma variável. As três principais medidas numéricas do centro da distribuição são a moda, a mediana e a média aritmética. Até agora, quando olhamos para a forma da distribuição, identificamos a moda como o valor em que a distribuição tem um pico. E vimos exemplos quando as distribuições têm uma moda, que é uma distribuição unimodal, ou duas modas, uma distribuição bimodal. Em outras palavras, até agora identificamos a moda visualmente a partir do histograma. Olhando para os nossos histogramas novamente, podemos facilmente ver a moda. É o valor que ocorre mais comum na distribuição.

A mediana, que é o ponto médio da distribuição, é o número tal que metade das observações cai acima e metade cai abaixo? Encontramos a mediana ordenando os dados do menor para o maior. Considere quando N, o número de observações é par ou ímpar. Se N for ímpar a mediana é a observação central na lista ordenada, quando o número de observações é mesmo a mediana é a média ou média do valor das duas observações centrais.

A média, é claro, pode ser calculada adicionando os valores para todas as observações e dividindo pelo número de observações para gerar uma média aritmética. Nossa objetivo aqui é descrever a distribuição. Como você descreveria essas duas distribuições de escores de exames? Ambas as distribuições estão centradas em 70. A média de ambas as distribuições é de aproximadamente 70. Mas as distribuições são realmente muito diferentes. A primeira distribuição tem variabilidade muito maior e pontuações em comparação com a segunda.

Para descrever uma distribuição, precisamos complementar a exibição gráfica, não só com a medida do centro, mas também com a medida da variabilidade ou dispersão da distribuição. Existem várias maneiras de descrever a dispersão. Uma medida comumente usada é o desvio padrão. A ideia por trás do desvio padrão é quantificar a propagação de a distribuição medindo o quanto longe as observações estão de sua média.



O desvio padrão dá a média ou distância típica entre um ponto de dados e a média. Para entender melhor o desvio padrão, seria útil ver um exemplo de como ele é calculado. Na prática, é claro, o software estará fazendo esses cálculos para nós.

Empresas de serviços médicos de emergência gostariam de estimar quantas tripulações de ambulância devem manter em espera. Aqui está o número de chamadas de ambulância durante um período de oito horas.

7, 9, 5, 13, 3, 11, 15, 9

$$\text{Média} \Rightarrow \bar{X} = (7 + 9 + 5 + 13 + 3 + 11 + 15 + 9) / 8 = 9$$

Para encontrar o desvio padrão do número de chamadas por hora, primeiro encontrariam a média dos nossos dados. Em seguida, precisaríamos encontrar os desvios da média. Essa é a diferença entre cada observação na média. Como nossa média é 9, subtraíramos 9 de cada uma de nossas observações.

$$\begin{array}{r}
 7 & 9 & 5 & 13 & 3 & 11 & 15 & 9 \\
 - & - & - & - & - & - & - & - \\
 \frac{9}{-2} & \frac{9}{0} & \frac{9}{-4} & \frac{9}{4} & \frac{9}{-6} & \frac{9}{2} & \frac{9}{6} & \frac{9}{0}
 \end{array}$$

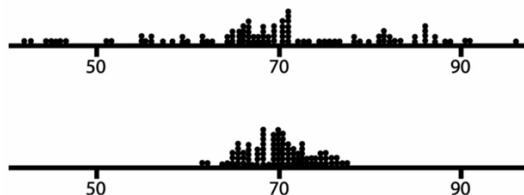
Como um terceiro passo, nós elevaríamos ao quadrado cada um desses desvios. Em seguida, medimos os desvios quadrados adicionando-os e dividindo-os por N-1, que é um a menos do que o tamanho amostral, esta média dos desvios quadrados é chamada de variância. O desvio padrão de sua variável é a raiz quadrada dessa variância.

$$\frac{(4 + 0 + 16 + 16 + 36 + 4 + 36 + 0)}{(8 - 1)} = \frac{112}{7} = \sqrt{16} = 4$$

Então, por que tomamos raiz quadrada? Note que 16 é a média dos desvios quadrados e, portanto, tem diferentes unidades de medida. Neste caso, 16 é medido em número quadrado de chamadas de ambulância, que obviamente não pode ser interpretado. Nós, portanto, tomamos a raiz quadrada para compensar o fato de que nós temos quadrado todos os nossos desvios e também para voltar para a unidade de medida original. Lembre-se de que o número médio de chamadas de emergência em uma hora é 9. A interpretação do desvio padrão igual a 4 é que, em média, o número real de chamadas de emergência a cada hora está a 4 de distância de 9. Outra maneira de dizer isso é que há uma média de chamadas de ambulância/hora = 9 ± 4 .

Desde que estamos trabalhando com um grande número de observações cálculos de mão de desvio padrão realmente não são viáveis. Python fará todos esses cálculos para você, mas é importante saber como calcular desvios padrão para que você possa entender sua variabilidade. Por exemplo, olhando para uma distribuição de variáveis em duas amostras diferentes, você deve ser capaz de dizer qual tem maior variabilidade, ou seja, um desvio padrão maior. Para calcular o desvio padrão e gerar outras estatísticas descritivas para uma variável quantitativa, muitas vezes usamos a função de descrição do Python.

Variable Distribution in Two Samples



Aqui está a sintaxe para descrever NUMCIGMO_EST como a variável quantitativa.

```

61 seaborn.distplot(sub2["NUMCIGMO_EST"].dropna(), kde=False);
62 plt.xlabel('Number of Cigarettes per Month')
63 plt.title('Estimated Number of Cigarettes per Month among Young Adult Smokers in the NESARC Study')
64
65 # standard deviation and other descriptive statistics for quantitative variables
66 print('describe number of cigarettes smoked per month')
67 desc1 = sub2['NUMCIGMO_EST'].describe()
68 print(desc1)
69

```

desc1 é o nome dado ao objeto que armazenará esses cálculos, igual a NUMCIGMO_EST. Antes, há um título da saída e então mandamos Python imprimir os resultados. Isso fornece uma contagem, média, desvio padrão, valores mínimos e máximos e os valores de percentil 25, 50 e 70. Então você pode ver que descrever é extremamente útil na melhor compreensão das características importantes desta variável cigarros fumados por mês.

```

describe number of cigarettes smoked per month
count    1706.000000
mean      0.525205
std       .499751
min      0.000000
25%     0.000000
50%     1.000000
75%     1.000000
max      1.000000
Name: NUMCIGMO_EST, dtype: float64

```

Sabemos agora que os jovens fumantes adultos em nossa amostra fumam em média 320 cigarros por mês. Em que o desvio padrão é de cerca de 274, podemos dizer que, em média, jovens fumantes adultos fumaram 320 por mês \pm 274 cigarros. Então, como você pode ver, há uma gama extremamente grande em termos de cigarros fumados, e muita variabilidade nesta variável. Código muito semelhante pode ser usado para calcular muitas dessas estatísticas individualmente ou para gerar estatísticas descritivas adicionais. Aqui está o código adicional para gerar a média, desvio padrão, mínimo e máximo, mediana e moda de uma variável quantitativa.

Note que a contagem para esta variável é 1697 em vez de o tamanho da nossa amostra de jovens fumantes adultos que foi 1706. Isso ocorre porque o Python não inclui os casos com dados ausentes ou NaN nesses cálculos. Mas se incluirmos uma variável categórica ao empregar a função describe? Como definimos anteriormente TAB12MDX, nossa variável de dependência de nicotina é categórica. Adicionando a sintaxe de descrição nos fornece estatísticas descritivas apropriadas para dados categóricos. Isto é count, número de valores exclusivos, o valor superior ou mais alto e a frequência desse valor superior.

```

print('mean')
mean1 = sub2['NUMCIGMO_EST'].mean()
print(mean1)

print('std')
std1 = sub2['NUMCIGMO_EST'].std()
print(std1)

print('min')
min1 = sub2['NUMCIGMO_EST'].min()
print(min1)

print('max')
max1 = sub2['NUMCIGMO_EST'].max()
print(max1)

print('median')
median1 = sub2['NUMCIGMO_EST'].median()
print(median1)

print('mode')
mode1 = sub2['NUMCIGMO_EST'].mode()
print(mode1)

```

```

69
70 print('describe nicotine dependence')
71 desc2 = sub2['TAB12MDX'].describe()
72 print(desc2)
73

```

	describe nicotine dependence
count	1706
unique	2
top	1
freq	896
Name:	TAB12MDX, dtype: int64

Se você não tivesse descrito esta variável como categórica, Python ainda geraria estatísticas descritivas. No entanto, muitos não fariam nenhum sentido. Se você se lembrar da variável de dependência de nicotina representada com códigos fictícios. Ou seja, sim é indicado com um 1 e não indicado com um 0. Como você pode ver aqui temos um desvio padrão baseado em códigos fictícios de 1 e 0. Além disso, os percentis são listados representando sim e não em vez de quantidades reais.

```

describe number of cigarettes smoked per month
count    1697.000000
mean     320.304361
std      274.436777
min      1.000000
25%     90.000000
50%     300.000000
75%     600.000000
max     2940.000000
Name: NUMCIGMO_EST, dtype: float64

```

Então, novamente, é muito importante lembrar de usar as estatísticas descritivas apropriadas para variáveis quantitativas e categóricas. Para variáveis quantitativas, é melhor examinar histogramas e, em seguida, complementá-los com medidas exatas de forma, centro e propagação. Variáveis categóricas podem ser descritas com frequência distribuições ou com um gráfico de barras.

Tarefa - Criar gráficos sobre seus dados

Há uma variedade de maneiras convencionais de visualizar dados - tabelas, histogramas, gráficos de barras, etc. Agora que seus dados foram gerenciados, é hora de representar graficamente suas variáveis. Essa parte do projeto é vital, pois fornecerá aos leitores representações visuais de seus dados e ajudará você a exibir melhor suas descobertas.

Pontuação

Sua avaliação será baseada nas evidências fornecidas por você de que concluiu todas as etapas. Quando relevante, a pontuação deverá recompensar a clareza (por exemplo, você receberá um ponto por enviar gráficos que não representam seus dados com precisão, mas dois pontos se os dados forem representados com precisão).

Você será avaliado igualmente em sua descrição de suas distribuições de frequência. Os itens específicos e seus valores de pontos são os seguintes:

1. Foi criado um gráfico univariado para cada uma das variáveis selecionadas? (2 pontos)
2. Foi criado um gráfico bivariado para as variáveis selecionadas? (2 pontos)
3. O resumo descreveu o que os gráficos revelaram em termos de variáveis individuais e a relação entre elas? (2 pontos)

Instruções

Continue com o programa que você executou com sucesso.

PASSO 1: Crie gráficos de suas variáveis uma de cada vez (gráficos univariados). Examine as medidas centrais e a dispersão.

PASSO 2: Crie um gráfico mostrando a associação entre suas variáveis explicativas e de resposta (gráfico bivariado). Sua saída deve ser interpretável (ou seja, organizada e rotulada).

O QUE APRESENTAR: Depois de escrever um programa bem-sucedido que cria gráficos univariados e bivariados, crie uma entrada de blog onde você publica seu programa e os gráficos que criou. Escreva algumas frases descrevendo o que seus gráficos revelam em termos de suas variáveis individuais e a relação entre elas.

4.2. Estatística inferencial

Até agora, demos os primeiros passos em um quadro maior de pesquisa estatística. Você identificou um conjunto de dados e usou a análise exploratória de dados para organizar e resumir os dados brutos de forma significativa e informativa. As ferramentas de análise exploratória de dados, incluindo avaliação de frequência de distribuição, representações gráficas de suas variáveis de interesse, e cálculos centrais e de dispersão, que nos ajudam a descobrir características importantes e padrões nos dados e quaisquer desvios marcantes desses padrões. Tudo isso se enquadra em Estatística Descritiva. A Estatística Descritiva visa descrever quantitativamente ou resumir uma amostra de dados.

Agora você será apresentado às Estatísticas Inferenciais, que é o nosso objetivo final. A estatística inferencial é usada para fazer inferências sobre uma população a partir da análise de uma amostra dessa população. Isso tem o objetivo expresso de chegar a conclusões generalizadas que se aplicam a toda a população. Normalmente, uma amostra aleatória da população é selecionada com base na média. Algumas das análises mais comuns em estatística inferencial incluem testes de hipóteses, intervalos de confiança e análise de regressão.

O teste de hipóteses é uma das ferramentas inferenciais mais importantes na aplicação de estatísticas para problemas da vida real. É usado quando precisamos tomar decisões sobre populações, com base em apenas uma amostra. Teste de Hipótese Estatística é definido como a avaliação de evidências fornecidas pelos dados a favor ou contra cada hipótese sobre a população. O teste de hipóteses usa métodos estatísticos para gerar evidências e tirar conclusões sobre populações inteiras. Esse teste usa teorias mutuamente exclusivas dentro do conjunto de dados da amostra, operando dentro da taxa de erro da amostra, para determinar qual hipótese tem o suporte dos dados. Os intervalos de confiança (ICs) incorporam incerteza e taxas de erro de amostra para criar uma faixa viável de valores para um valor desconhecido em toda a população. Já a análise de regressão explica a relação

entre várias variáveis independentes e uma variável dependente. Os modelos de regressão permitem que os analistas façam previsões com base nos valores presentes em um conjunto de dados de amostra.

Para realmente entender como a inferência funciona, primeiro precisamos falar sobre Probabilidade. Porque é a base subjacente de todos os métodos estatísticos. Aqui está a ideia básica. Como você sabe, as estatísticas usam uma amostra para aprender sobre a população maior da qual a amostra foi desenhada. Idealmente, a amostra deve ser aleatória para que possa representar melhor toda a população. É muito importante reconhecer embora que isso não significa que todas as amostras aleatórias são ideais. Nenhuma amostra aleatória será exatamente a mesma que qualquer outra.

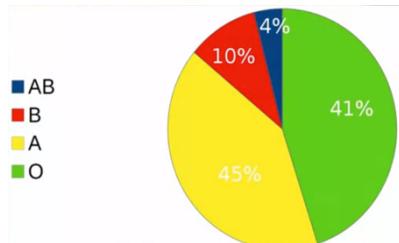
Uma amostra aleatória pode ser uma representação bastante precisa da população maior, enquanto outra amostra aleatória pode não ser precisa, puramente devido ao acaso. Infelizmente, ao olhar para uma amostra aleatória específica, que é o que acontece nas estatísticas, nunca saberemos o quanto que amostra aleatória difere da população. Esta incerteza é onde a probabilidade entra na imagem. Usamos probabilidade para quantificar o quanto esperamos que amostras aleatórias variem. Isso nos dá uma maneira de tirar conclusões sobre a população em face da incerteza que é gerada pelo uso de uma amostra aleatória.

Como exemplo, vamos supor que estamos interessados em estimar a porcentagem de adultos norte-americanos que favorecem a pena de morte. Para fazer isso, escolhemos uma amostra aleatória de 1.200 adultos norte-americanos e pedir sua opinião a favor ou contra a pena de morte. Descobrimos que 744 dos 1200, ou 62% são a favor. Aqui está uma imagem que ilustra o que fizemos e encontramos em nosso exemplo. Nosso objetivo aqui é inferir, tirar conclusões sobre as opiniões de toda a população de adultos norte-americanos sobre a pena de morte, com base nas opiniões de apenas 1200 deles, podemos concluir absolutamente que 62% da população favorece a pena de morte?

Outra amostra aleatória poderia dar um resultado muito diferente, então estamos incertos. Como nossa amostra é aleatória, sabemos que nossa incerteza é devido ao acaso. Não se deve a problemas de como a amostra foi coletada. Portanto, podemos usar a probabilidade para descrever a probabilidade de que nossa amostra esteja dentro de um nível desejado de precisão. Por exemplo, probabilidade pode responder à pergunta, quão provável é que nossa estimativa amostral esteja dentro de 3% da porcentagem REAL de TODOS os adultos norte-americanos que são a favor da pena de morte. A resposta a esta pergunta, que encontramos usando a probabilidade obviamente terá um impacto importante na confiança que podemos anexar ao passo de inferência. Em particular, se acharmos bastante improvável que a porcentagem da amostra seja muito diferente da porcentagem da população, então temos boa confiança de que podemos tirar conclusões sobre a população com base na amostra. Então vamos definir probabilidade um pouco mais cuidadosamente.

4.2.1. Da amostra à população

Para entender melhor a relação entre amostra e população, vamos considerar dois exemplos simples. Aqui estão as distribuições de tipos sanguíneos na população dos EUA. Você pode ver os tipos de sangue comuns incluem Tipo A e Tipo O, com tipos de sangue menos comuns, incluindo AB e B. Vamos supor agora que tomamos uma amostra de 500 pessoas nos Estados Unidos, registramos seu tipo sanguíneo e exibir os resultados da amostra.



Se você olhar com cuidado, notará que as percentagens de cada tipo sanguíneo de nossa amostra são ligeiramente diferentes das percentagens da população. Mas tenho certeza de que isso não te surpreende, certo?

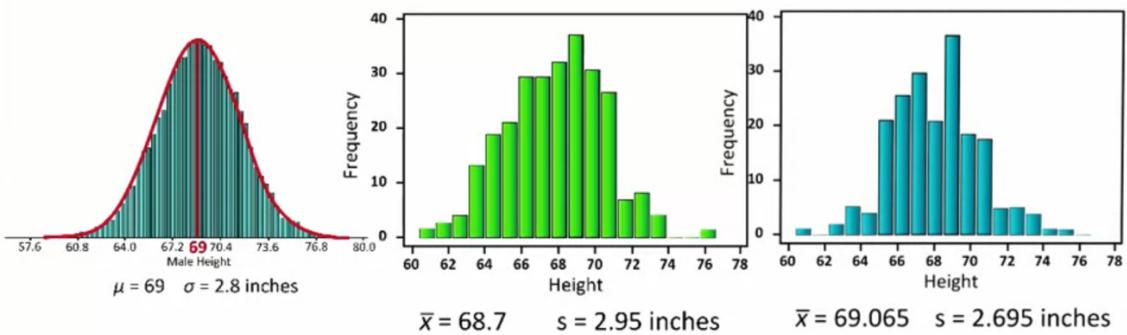
Quero dizer, já que pegamos uma amostra de apenas 500 indivíduos, não podemos esperar que nossa amostra se comporte exatamente como a população. Mas se a amostra é aleatória, e este foi, esperamos obter resultados que não são tão diferentes dos resultados de toda a população e isso é o que encontramos. Mais uma amostra aleatória de 500 indivíduos, revela resultados que são ligeiramente diferentes das figuras populacionais e também de que temos na primeira amostra.

Esta ideia muito intuitiva de que os resultados da amostra mudam de amostra para amostra, é chamada de variabilidade de amostragem. Aqui está outro exemplo para ajudar a entender melhor a relação entre a população de amostragem. Este exemplo é baseado nas alturas entre a população dos EUA de **todos** os homens adultos. Como você pode ver, segue uma distribuição normal com uma média de 69 polegadas e um desvio padrão de 2,8 polegadas.



Digamos que uma amostra de 200 homens foi escolhida e suas alturas foram registradas. Estes são os resultados da amostra 2. A média da amostra é de 68,7 polegadas, e o desvio padrão da amostra é de 2,95 polegadas. Novamente, observe que os resultados da amostra são ligeiramente diferentes dos resultados da população.

O histograma que criamos para a primeira amostra, se assemelha à distribuição normal da população. No entanto, a média da amostra no desvio padrão é ligeiramente diferente da média da população no desvio padrão. Vamos tirar outra amostra de duzentos homens exibidos aqui na amostra dois. A média da amostra é de 69,065 polegadas e o desvio padrão da amostra é de 2,659 polegadas. Este exemplo, novamente, demonstra a variabilidade da amostragem. Embora os resultados da amostra estejam muito próximos dos resultados da população, eles são ligeiramente diferentes dos resultados encontrados na primeira amostra.



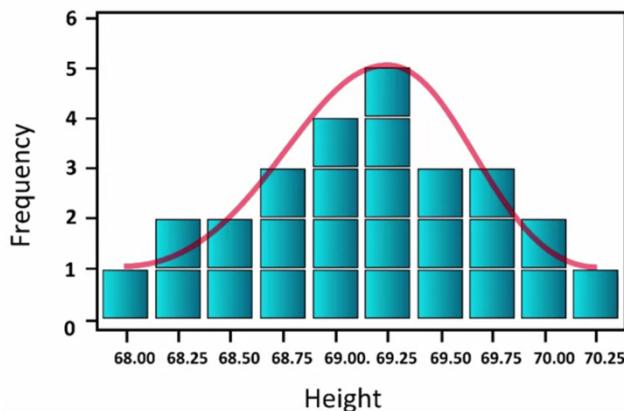
Em ambos os exemplos, temos números que descrevem a população e números que descrevem a amostra.

Um parâmetro é um número que descreve a população e uma estatística é um número calculado a partir de uma amostra. Os parâmetros são tipicamente desconhecidos, porque é impraticável ou até mesmo impossível saber exatamente quais valores uma variável leva para cada membro de uma população muito grande. As estatísticas são calculadas a partir de amostras, e cada amostra de uma população terá estatísticas diferentes. As estatísticas de diferentes amostras de uma população variam. Isto é devido à variabilidade da amostragem.

Até agora, temos feito distribuições baseadas em variáveis individuais. Teoricamente, podemos criar distribuições a partir de médias ou proporções tiradas de várias amostras aleatórias extraídas de uma população. Esta é a grande ideia por trás das estatísticas inferenciais.

Como exemplo, suponha que selecionamos 30 amostras aleatórias separadas em vez de apenas duas. E cada uma das 30 amostras aleatórias tem 500 indivíduos retirados da população de adultos norte-americanos. A primeira amostra tem uma altura média de 69 polegadas. Poderíamos criar um gráfico de barras e plotar essa média para nossa primeira amostra no gráfico. Se nossa segunda amostra tivesse uma altura média de 68,5 polegadas, adicionaríamos isso ao gráfico. À medida que continuamos a traçar a altura média de cada amostra aleatória, um padrão começaria a surgir. Observe como há mais meios de amostra a 69,25 polegadas do que em qualquer outro comprimento.

Observe também como, à medida que o comprimento se torna maior ou menor, há cada vez menos meios de amostra. Esta é uma característica da distribuição amostral, se estamos medindo a média de uma variável quantitativa ou a proporção de variável categórica ou qualquer outra estatística amostral. Ou seja, à medida que desenhamos mais e mais amostras, a distribuição da amostra estatística se tornará cada vez mais normalmente distribuída.



Este resultado é conhecido como o **Teorema do Limite Central**, que afirma que, enquanto amostras adequadamente grandes e um número suficientemente grande de amostras são extraídas de uma população, a distribuição das estatísticas das amostras, seja de média, proporção, desvio padrão ou qualquer outra estatística, será normalmente distribuída. Nossos projetos dependem de apenas uma amostra. No entanto, se essa amostra é representativa de uma população maior, os testes estatísticos inferenciais nos permitem estimar com diferentes níveis de parâmetros de certeza para toda a população. Esta ideia é a base para cada uma das ferramentas inferenciais que você usará para responder à sua pergunta de pesquisa.

4.2.2. Teste de Hipótese

Teste de hipóteses é uma das ferramentas inferenciais mais importantes quando se trata de para a aplicação de estatísticas para problemas da vida real. Teste de hipóteses é usado quando precisamos tomar decisões sobre populações com base apenas em informações de amostra. Uma variedade de testes estatísticos é usada para ajudar a chegar a essas decisões. Por exemplo, a análise do teste de variância, ANOVA. E o Qui Quadrado Teste da Independência, para citar alguns. Mas todos eles incluem os mesmos passos básicos.

Passos envolvidos no teste de hipóteses, incluem especificar a hipótese nula H_0 , e a hipótese alternativa, H_a . Escolhendo uma amostra, avaliando a evidência e tirando conclusões. Teste de hipóteses estatísticas é definido como a avaliação de evidências fornecidas pelos dados a favor ou contra cada hipótese sobre a população.

Para fornecer um exemplo de teste de hipótese, vamos usar o conjunto de dados NESARC. Uma amostra representativa de 43.093 adultos nos Estados Unidos. Vamos avaliar se existe ou não uma associação entre um

diagnóstico de depressão maior e o quanto uma pessoa fuma. Vamos trabalhar através do exemplo usando as quatro etapas.

1. Especificar a hipótese nula e alternativa,
2. Escolher uma amostra
3. Avaliar a evidência e
4. Tirar conclusões

Primeiro, há duas hipóteses opostas para questionar. A hipótese nula, comumente mostrada como H_0 , é que não há diferença na quantidade de tabagismo entre pessoas com e sem depressão. A hipótese alternativa, mostrada como H_a ou às vezes mostrado como H_1 , é que existe uma diferença na quantidade de tabagismo entre pessoas com e sem depressão.

A hipótese nula basicamente, diz que nada de especial está acontecendo entre depressão e tabagismo. Em outras palavras, que eles não estão relacionados uns com os outros. A hipótese alternativa diz que existe uma relação e permite que a diferença no tabagismo naqueles indivíduos com e sem depressão possa ser positiva ou negativa. Ou seja, indivíduos com depressão podem fumar mais do que indivíduos sem depressão, ou podem fumar menos.

Depois de declarar a hipótese nula e alternativa, precisamos escolher uma amostra. Nós vamos usar o conjunto de dados NESARC, e nós só vamos avaliar essas hipóteses entre indivíduos que são fumantes e que são mais jovens, em vez de adultos mais velhos. Restringimos os dados NESARC para indivíduos que são: 1. fumantes diários atuais, ou seja, eles fumaram todos os dias no mês anterior ao questionário. E, 2. tem idades entre 18 a 25 anos.

Esta amostra, $N = 1320$, mostrou o seguinte. Jovens adultos fumantes diários com depressão fumavam uma média de 13,9 cigarros por dia com um desvio padrão de 9,2 cigarros. Jovens adultos fumantes diários sem depressão fumavam uma média de 13,2 cigarros por dia com um desvio padrão de 8,5 cigarros. Embora seja verdade que 13,9 cigarros por dia são mais de 13,2 cigarros por dia, não é de todo claro que este é uma diferença grande o suficiente para rejeitar a hipótese nula. Ou dizer que os fumantes com depressão fumam significativamente mais do que os fumantes sem depressão.

Embora seja verdade que 13,9 cigarros por dia são mais de 13,2 cigarros por dia, não é de todo claro que esta é uma diferença grande o suficiente para rejeitar a hipótese nula. Ou dizer que os fumantes com depressão fumam significativamente mais do que os fumantes sem depressão. Portanto, precisamos avaliar a evidência, a fim de determinar se os dados fornecem evidência forte o suficiente contra a hipótese nula. Ou seja, contra a alegação de que não há relação entre fumar e depressão. Nós realmente precisamos nos perguntar, quão surpreendente ou raro é para obter uma diferença de 0,7 cigarros fumaça por dia entre nossos dois grupos? Ou seja, aqueles com depressão, e aqueles sem, assumindo que a hipótese nula é verdadeira, que não há relação entre fumar e depressão.

Esta é uma etapa onde calculamos a probabilidade de obter dados como este quando a hipótese nula é verdadeira. Em certo sentido, este é realmente o coração do processo, uma vez que tiramos nossas conclusões com base na estimativa de probabilidade. A hipótese nula é geralmente assumida como verdadeira até que a evidência indique o contrário. A probabilidade de obtermos uma diferença desse tamanho no número médio de cigarros fumados em uma amostra aleatória de 1.320 participantes é de aproximadamente 0,17 ou 17%.

Vamos falar sobre como isso é calculado para os diferentes testes estatísticos mais tarde. O ponto importante nesta fase é que é esse tipo de evidência que seremos considerando cada vez que decidirmos aceitar ou rejeitar a hipótese nula. Então, como exatamente usamos essa probabilidade para chegar a uma conclusão sobre a hipótese nula? Lembre-se, se a hipótese nula for verdadeira, não há associação. Há uma probabilidade de 0,17 ou 17% de observar esse tamanho de diferença entre fumantes com e sem depressão.

A tradução desta probabilidade de 17% é que se tirássemos 100 amostras aleatórias de nossa população, estariamos errados 17 de 100 vezes se rejeitássemos a hipótese nula e dissemos que havia uma diferença na quantidade de tabagismo para fumantes com e sem depressão. Agora temos que decidir se ou não isso é algo que nos sentimos confortáveis. Importa-se de cometer um erro e dizendo que há uma diferença na quantidade de fumar 17 em cada 100 vezes?

Essa probabilidade de 0,17 torna o que estamos observando raro o suficiente para nos fazer sentir confiantes em rejeitar a hipótese nula?

Provavelmente todos concordamos que uma probabilidade de 0,50 certamente não nos daria confiança suficiente para rejeitar a hipótese nula. Porque 0,50, ou 50%, significa que estariamos certos 50 em 100 vezes, e errado 50 em 100 vezes. Não é melhor do que tomar decisões baseadas no lançar de uma moeda.

Estar errado 17 de 100 vezes nos faria muito menos propensos a ser errados ao rejeitar a hipótese nula, mas ainda estariamos menos certos do que se a probabilidade fosse ainda menor, digamos 10 ou até 5%. Basicamente, esta é a nossa decisão ao testar hipóteses. Para tomar essa decisão, seria bom ter algum tipo de diretriz ou padrão. Que probabilidade nos daria confiança em rejeitar uma hipótese nula?

4.2.3. Valor-p e Intervalo de Confiança

A razão para usar um Teste Inferencial é obter um valor de probabilidade, comumente chamado valor-p. O valor de p fornece uma estimativa de quantas vezes nós iríamos obter o resultado obtido por acaso se de fato, a hipótese nula é verdadeira. Em estatística, um resultado é chamado de estatisticamente significativo se é improvável que tenha ocorrido apenas por acaso.

O padrão ou corte mais comumente usado é 0,05 ou 5%. Porque este padrão, ou corte é tão importante que tem um nome especial. É chamado de nível de significância de um teste, e é geralmente denotado pela letra grega alfa, então alfa é igual a 0,05.

Se o valor de p for pequeno, menor que 0,05, isso sugere que é mais de 95% provável que a associação de interesse esteja presente após amostras repetidas tiradas da população, em outras palavras, uma distribuição de amostragem. Se o valor de p for menor que alfa, que geralmente é 0,05, então os dados que obtivemos são considerados raros ou surpreendentes o suficiente quando a hipótese nula, H_0 é verdadeira. E dizemos, que os dados fornecem evidências significativas contra a hipótese nula. Então, rejeitamos a hipótese nula e aceitamos a hipótese alternativa, H_a .

Se o valor-p for maior que alfa, então os dados não são considerados surpreendentes o suficiente quando a hipótese nula é verdadeira. E dizemos, que nossos dados não fornecem evidências suficientes para rejeitar a hipótese nula. Ou equivalentemente, que os dados não fornecem evidências suficientes para aceitar a hipótese alternativa.

Assim, encontrar um valor de p menor que ou igual a 0,05 significa que o achado é estatisticamente significativo, e podemos rejeitar a hipótese nula e aceitar a hipótese alternativa. Este valor-p também é conhecido como **Taxa de Erro do Tipo Um**, uma vez que denota o número de vezes que estariamos errados ao rejeitar a hipótese nula quando era verdadeira.

Rejeitar a hipótese nula quando é verdadeira também é chamado de Erro do Tipo Um.

Olhando para o valor de p em nosso exemplo, vemos que não há evidência adequada para rejeitar a hipótese nula porque o valor de p foi 0,17, que é definitivamente maior que 0,05. Em outras palavras, não foi rejeitada a hipótese nula de que não há associação entre depressão e número de cigarros fumados entre jovens fumantes diários. Aceitamos a hipótese nula. Não há associação entre tabagismo e depressão, porque os dados não fornecem evidências suficientes para aceitar a hipótese alternativa, de que existe associação entre tabagismo e depressão.

Vamos mudar ligeiramente a questão da pesquisa para demonstrar que as decisões que você toma sobre sua amostra e suas variáveis podem afetar suas descobertas e as conclusões que você tira. Usando nosso exemplo, ainda estamos interessados na associação entre depressão e tabagismo. No entanto, decidimos não nos limitar a considerar apenas indivíduos que fumam diariamente. Vamos olhar para uma população mais ampla de jovens adultos, e considerar aqueles que já fumaram no ano passado, seja diariamente ou mais irregularmente.

O tamanho da amostra no conjunto de dados NESARC é 1.706. Com esta amostra, descobrimos que jovens adultos com depressão fumavam uma média de 351,7 cigarros por mês com um desvio padrão de 300 cigarros. Jovens adultos sem depressão fumavam uma média de 313,5 cigarros por mês, com um desvio padrão de 268,2 cigarros. Assim, a diferença entre a quantidade de cigarros fumados entre os jovens adultos que fumou no ano passado com e sem depressão é de 38,2 cigarros por mês, quase 2 pacotes.

O valor de p deste cenário revisado é 0,0285, obviamente inferior a 0,05. Isso significa que a probabilidade de obtermos uma diferença desse tamanho em o número médio de cigarros fumados em uma amostra aleatória de 1.706 participantes é menor que 3%, que é um valor-p inferior a 0,05. Então, neste caso, podemos rejeitar a hipótese nula, e dizem que jovens fumantes adultos com depressão fumam significativamente mais cigarros por mês do que jovens fumantes adultos sem depressão.

Se olharmos novamente para a linha numérica de probabilidades, podemos traduzir esta descoberta da seguinte maneira. Se rejeitarmos a hipótese nula e dissermos que há uma diferença entre o número médio de cigarros fumados por mês entre os jovens adultos, com e sem depressão, estariámos errados menos de 3 em cada 100 vezes. Estariámos corretos mais de 97% do tempo. Baseado nos padrões da ciência, este é um nível de certeza que nos dá confiança em dizer que há uma associação significativa entre fumar e depressão entre jovens fumantes adultos atuais.

4.2.4. Escolhendo testes estatísticos

Você foi apresentado ao processo geral de testes de hipóteses. É hora de aprender a testar sua própria hipótese. Você sempre estará interpretando valores p, independentemente do teste inferencial que você usa.

O teste estatístico específico que você usa para avaliar suas hipóteses, dependerá do tipo de variáveis explicativas e de resposta que você escolheu.

- Se você tiver uma variável explicativa categórica e uma variável de resposta quantitativa, você usaria uma Análise de Variância, ANOVA como teste inferencial.
- Se você tem uma variável explicativa categórica, e sua variável de resposta também é uma variável categórica, você usaria o Teste de Independência Qui-Quadrado como seu teste inferencial.
- Se ambas as variáveis explicativas e de resposta forem quantitativas, você usaria um coeficiente de correlação como teste inferencial.
- Se sua variável explicativa for quantitativa e sua variável de resposta for categórica, você categorizaria sua variável explicativa com apenas dois níveis e, em seguida, use o Teste Qui-Quadrado da Independência como seu teste inferencial.

		Resposta (dependente)	
		Categórica	Quantitativa
Explanatória (independente)	Categórica	C -> C Teste de Independência Qui-quadrado	C -> Q Análise de Variância (ANOVA)
	Quantitativa	Q -> C Qui-quadrado ajustado	Q -> Q Correlação de Pearson

4.2.5. Análise de Variância - ANOVA

Então, finalmente, estamos prontos para começar a testar nossas perguntas de pesquisa estatisticamente. Embora tenhamos demorado algum tempo para chegar aqui, nossos passos anteriores nunca devem ser evitados. Ou seja, não importa o quão sofisticado você possa se tornar como um pesquisador quantitativo, você sempre precisará examinar seu livro de códigos, gerenciar seus dados e examinar estatísticas descritivas para as variáveis de interesse.

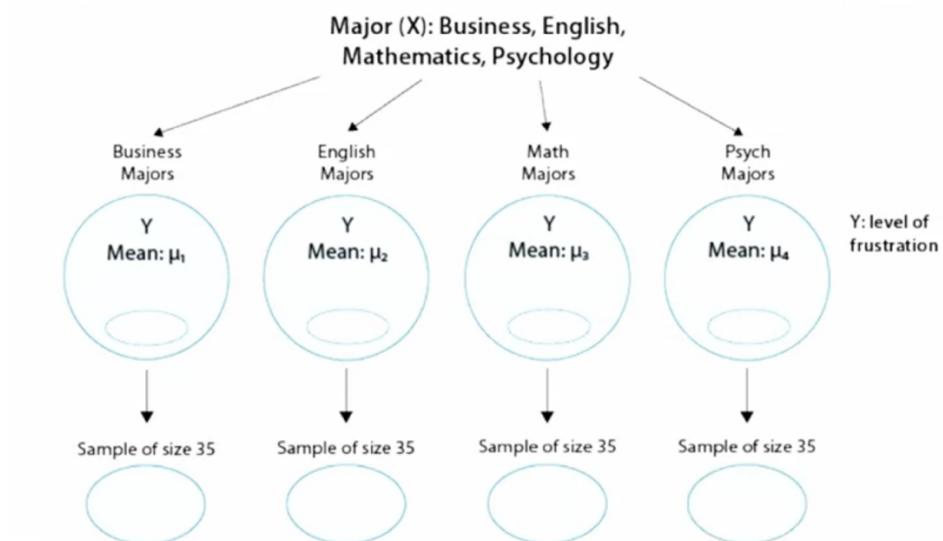
Na descrição do teste de hipóteses, quando analisamos a associação entre depressão e tabagismo, estávamos trabalhando com uma variável explicativa categórica, a presença ou ausência de depressão, e uma variável de resposta quantitativa, o número de cigarros fumados por mês. Quando você está testando hipótese com a variável explicativa categórica e uma variável de resposta quantitativa, a ferramenta que você deve usar é Análise de Variância, também chamada ANOVA.

Agora que você entende em quais situações você usaria ANOVA, estamos prontos para aprender como ela funciona ou mais especificamente o que a ideia está por trás da comparação de médias. O teste que você usará chama-se ANOVA F-test. Então vamos usar outra questão de pesquisa categórica para quantitativa.

A frustração acadêmica está relacionada à área cursada?

Neste exemplo, um reitor da faculdade acredita que estudantes com diferentes cursos podem experimentar diferentes níveis de frustração acadêmica. Amostras aleatórias de 35 indivíduos, cada um dos cursos de Negócios, Inglês, Matemática, e Psicologia foram convidados a avaliar seu nível de frustração acadêmica, em uma escala de um, o mais baixo, para vinte, o mais alto.

Esta figura destaca que estaremos examinando a relação entre major, nossa variável explicativa ou X, e o nível de frustração, nossa resposta, ou variável Y para comparar os diferentes meios de níveis de frustração entre os quatro principais definidos por X.



As alegações de hipótese nula que não há relação entre as variáveis de resposta explicativa e, x e y. Uma vez que a relação é examinada comparando as médias de y nas populações, definidas pelos valores de x, nenhuma relação significaria que todas as médias são iguais. Portanto, a hipótese nula do teste f é média da população 1 igual à média da população 2 é igual a média da população 3 igual à média da população 4.

Aqui temos apenas uma hipótese alternativa que afirma que há uma relação entre x e y. A variável independente e dependente. Em termos dos meios, ele simplesmente diz o contrário, que nem todos os meios são iguais e simplesmente escrevemos h1, nem todas as médias da população são iguais. Há muitas maneiras para a população significar não ser igual. Falaremos sobre isso mais tarde.

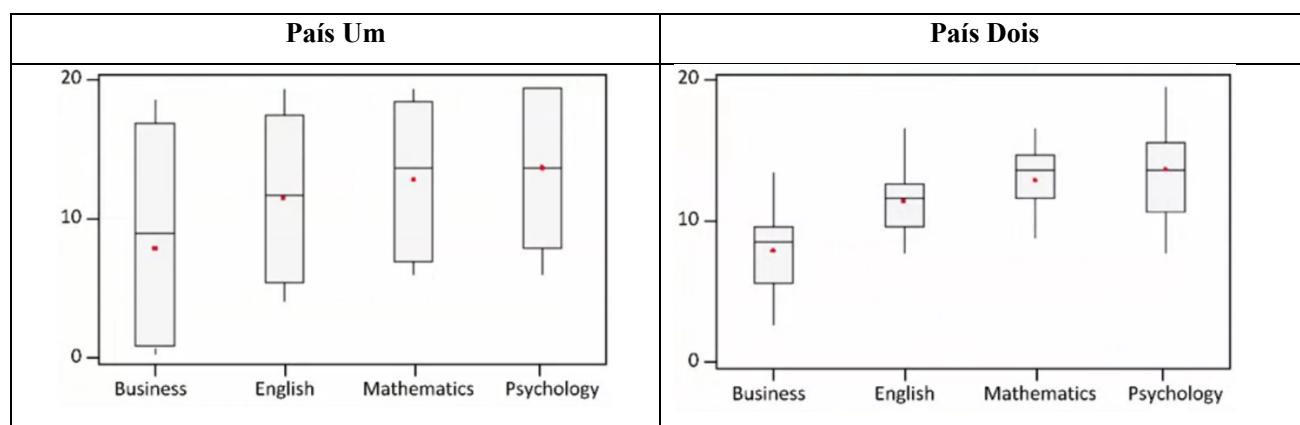
$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_a: \text{not all the } \mu \text{ are equal}$$

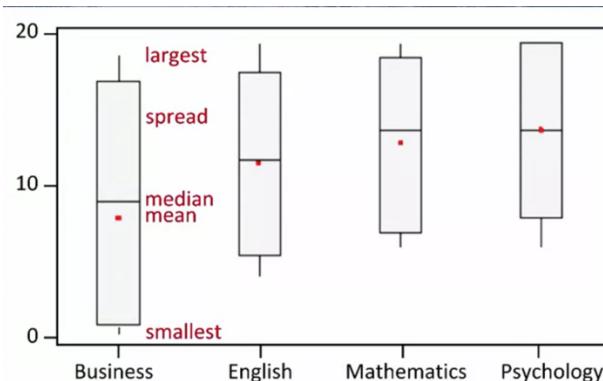
Por enquanto, vamos pensar sobre como iríamos testar se a população significa são iguais. Poderíamos calcular o nível médio de frustração para cada major e ver quão distantes essas médias de amostra estão. Ou, em outras palavras, meça a variação entre as médias da amostra. Se descobrirmos que as quatro médias da amostra não estão todas juntas, diremos que temos evidências contra a hipótese nula. E caso contrário, se eles estão próximos, diremos que não temos evidências contra a hipótese nula. Isso parece bastante simples, mas isso é suficiente?

- A pontuação média de frustração da amostra dos 35 alunos da área de negócios é: $y_1 = 7,3$.
- A pontuação média de frustração da amostra para os 35 alunos de inglês é: $y_2 = 11,8$.
- A pontuação média de frustração da amostra para os 35 alunos de matemática é: $y_3 = 13,2$.
- E a pontuação média de frustração da amostra para os 35 alunos de Psicologia é: $y_4 = 14,0$.

Aqui está uma representação gráfica de dois conjuntos de dados hipotéticos tirados de duas diferentes populações. Por exemplo, estudantes no País Um e estudantes no País Dois. Em nossas amostras hipotéticas, os meios são os mesmos, mas eles aparecem neste boxplot de forma muito diferente.



Um boxplot é uma maneira conveniente de descrever graficamente grupos de dados numéricos incluindo informações descritivas como a menor observação do grupo, a média e a mediana, a maior observação e a dispersão ou variabilidade dos valores. A parte superior da linha que se destaca do topo do gráfico de caixa e a parte inferior da linha que se destaca da parte inferior do gráfico de caixa são os valores mais altos e mais baixos. O ponto vermelho é a média. A linha horizontal do meio é a mediana.



Você pode ver que cada conjunto de dados tem o mesmo conjunto de médias e, portanto, as mesmas diferenças entre eles. Ou seja, estudantes no País Um e estudantes no País Dois. Ambos mostram dados para

quatro grupos com uma média de amostra de 7.3, 11.8, 13.2 e 14.0 indicada com marcas vermelhas. A diferença importante entre os dois conjuntos de dados é que o primeiro representa os dados com uma grande quantidade de variação dentro de cada um dos quatro grupos. O segundo representa dados com uma pequena quantidade de variação dentro de cada um dos quatro grupos.

Boxplots para País Um mostram muita sobreposição entre os quatro grupos devido à grande quantidade de variação nas pontuações de frustração dentro dos grupos. Pode-se imaginar os dados decorrentes de quatro amostras aleatórias tiradas de quatro populações, todas com a mesma média de cerca de 11 ou 12. O primeiro grupo de valores pode ter sido um pouco no lado baixo e os outros três um pouco no lado alto. Mas tais diferenças poderiam ter surgido por acaso. Este seria o caso se a hipótese nula alegando que médias de população iguais fossem verdadeiras.

Boxplots para o País Dois mostram muito pouca sobreposição por causa da pequena quantidade de pontuação de variação e frustração dentro dos grupos. Seria muito difícil acreditar que estamos amostrando de quatro grupos que têm necessidades populacionais iguais. Este caso é um exemplo de quando a hipótese nula alegando que a população igual precisa ser falsa.

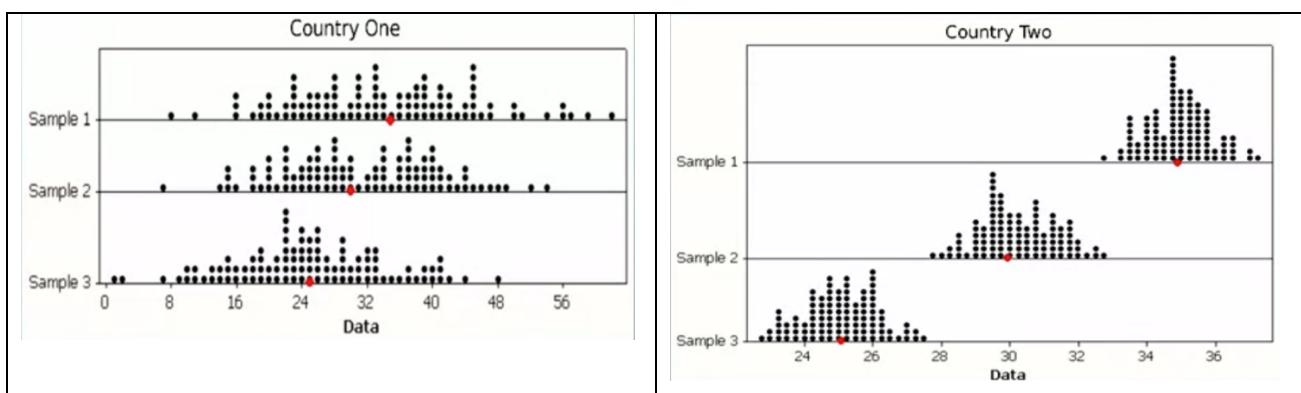
A pergunta que precisamos responder com o Teste ANOVA F é, as diferenças entre as médias da amostra devido a verdadeiras diferenças entre as médias da população, ou meramente devido à variabilidade amostral? Para responder a esta pergunta, usando nossos dados, obviamente precisamos olhar para a variação entre as médias de amostra. Mas isso não é suficiente. Também precisamos olhar para a variação entre as médias de amostra em relação à variação dentro dos grupos.

Então F é a variação entre as médias da amostra dividida pela variação dentro dos grupos. Em outras palavras, precisamos olhar para a quantidade, variação entre as médias de amostra, dividido por variação dentro de grupos. Que mede até que ponto a diferença entre os grupos amostrais, significa, domina sobre a variação usual dentro dos grupos amostrais. Que reflete diferenças em indivíduos que são típicos em amostras aleatórias.

F = Variação entre as médias das amostras

Variação dentro dos Grupos

Quando a variação dentro dos grupos é grande, como no País Um, as diferenças ou variação entre as médias da amostra podem se tornar insignificantes. E os dados fornecem muito pouca evidência contra a hipótese nula. Quando a variação dentro de grupos é pequena, como no País Dois, a variação entre as médias da amostra domina. E os dados têm evidências mais fortes contra a hipótese nula. Olhando para a proporção de variações é a ideia por trás das comparações e significa, portanto, a análise do nome da variância.



Aqui estão os resultados da análise de variância para o País Dois. Testando a relação entre pontuação maior e frustração. A estatística F circulada em vermelho é 46,60. Como sabemos que esta é a variabilidade entre as médias de amostra divididas pela variabilidade dentro dos grupos, esse grande número sugere que a variabilidade entre as médias amostrais é muito maior do que a dos grupos amostrais.

O valor P do Teste ANOVA F é a probabilidade de obter uma estatística F como maior que obtivemos ou mesmo maior se a hipótese nula fosse verdadeira. Ou seja, se a população significa ser igual. Em outras palavras, ele nos diz como é surpreendente encontrar dados como os observados, assumindo que não há diferença entre os meios populacionais. Este valor P é praticamente 0, dizendo-nos que seria quase impossível obter dados como aqueles observados se o nível médio de frustração dos quatro cursos fosse o mesmo que as alegações da hipótese nula.

One-Way ANOVA: Frustration Score Versus Major					
Source	DF	SS	MS	F	P
Major	3	939.85	313.28	46.60	0.0001
Error	136	914.29	6.72		
Total	139	1854.14			
$S = 2.593 \quad R-Sq = 50.69\% \quad R-Sq = (adj) = 49.60\%$					
Level	N	Mean	StDev		
Business	35	7.314	2.898		
English	35	11.771	2.088		
Mathematics	35	13.200	2.153		
Psychology	35	14.029	3.080		

O valor P 0,0001 sugere que vamos rejeitar incorretamente a hipótese nula uma em dez mil vezes. E estaremos corretos em aceitar a hipótese alternativa 9999 vezes em 10.000 vezes. Assim, podemos concluir com confiança que os meios de nível de frustração dos quatro cursos não são todos iguais. Ou em outras palavras, há uma associação significativa entre nível de frustração e maior. Então aceitamos a hipótese alternativa e rejeitamos a hipótese nula. Agora que você tem uma sensação de análise de variância, vamos executar o teste usando SAS. Usaremos um exemplo descrito pela primeira vez no teste de hipóteses.

Teste de Post Hoc com Anova

Quando a variável explicativa (independente) representa mais de dois grupos, um teste ANOVA significativo não nos diz quais grupos são diferentes dos outros. Para determinar quais grupos são diferentes dos outros, precisaríamos realizar um teste post hoc. Um teste post hoc conduz comparações emparelhadas post hoc. Post hoc significa depois do fato. E essas comparações emparelhadas post hoc devem ser conduzidas de uma maneira específica, a fim de evitar erros excessivos do tipo 1.

Erro do Tipo 1, ocorre quando você toma uma decisão incorreta sobre a hipótese nula. Ou seja, você rejeita a hipótese nula quando a hipótese nula for verdadeira. Por que não podemos simplesmente executar vários ANOVAs? Ou seja, por que não podemos apenas subdefinir nossas observações e levar duas de cada vez?

Como você sabe, aceitamos significância e rejeitamos a hipótese nula em p menor ou igual a 0,05. Uma chance de 5% de estarmos errados e cometermos um erro de tipo 1. Na verdade, há 5% de chance de fazer um erro de tipo 1 para cada análise de variância que realizamos nesta questão. Portanto, realizar vários testes significa que nossa chance geral de cometer erro tipo 1, pode ser muito maior do que 5%. Veja como funciona.

# Tests	Comparison α	Family-wise α
1	.05	.05
3	.05	.14
6	.05	.26
10	.05	.40
15	.05	.54

$$\alpha_{FW} = 1 - (1 - \alpha_{PC})^c$$

Where c = # of comparisons, α =normal Type 1 Error (.05)

Usando a fórmula exibida sob esta tabela, você pode ver que, enquanto um teste tem uma Taxa de Erro Tipo 1 de 0,05, no momento em que realizamos dez testes sobre esta questão, nossa chance de rejeitar a hipótese nula quando a hipótese nula for verdadeira é de até 40%. Este aumento na taxa de erro Tipo 1 é chamado de taxa de erro familiar e é a taxa de erro para o grupo de comparação de pares.

Os testes post-hoc são projetados para avaliar a diferença entre pares de médias enquanto protegem contra a inflação de erros de Tipo 1. E há muitos testes post hoc para escolher, quando se trata de análise de variância. Há o Sidak, o teste T Holm. e Teste de diferença menos significativa de Fisher. Teste de diferença honestamente significativa de Tukey, teste de Scheffe, teste de Newman-Keuls, teste de Comparação Múltipla de Dunnett, teste de alcance múltiplo Duncan e o Procedimento Bonferroni. É o suficiente para fazer sua cabeça nadar.

Embora haja certamente diferenças em quanto conservador cada teste é em termos de proteção contra erro do tipo um, em muitos casos é muito menos importante qual teste post hoc você conduz e muito mais importante que você conduza um.

Para realizar comparações emparelhadas post hoc no contexto da minha ANOVA, examinando a associação entre etnia e número de cigarros fumados por mês, vou usar o Tukey HSDT, ou Honestamente Significativa Diferença Test. Para fazer isso, vou primeiro adicionar uma instrução import para a biblioteca statsmodels.stats.multicomp no meu script python como multi, o termo que usarei para me referir à biblioteca mais tarde no meu programa. Em seguida, adicionarei o seguinte código ao final do meu programa. Estou chamando o objeto que irá armazenar minhas múltiplas comparações MC1 e usar a função multicomp da biblioteca multicomp de estatísticas de modelos multicomp, que eu importei como multi acima. Depois, incluo nesta declaração a variável de resposta quantitativa e a variável explicativa categórica entre parênteses. res1 é o nome que estou dando ao objeto que armazenará meus resultados post hoc. Em seguida, eu defini que igual ao meu objeto de comparações múltiplas, e eu solicito o teste hsd tukey. Finalmente, peço ao Python para imprimir esses resultados com a função de resumo.

```
03
70 mc1 = multi.MultiComparison(sub3['NUMCIGMO_EST'], sub3['ETHRACE2A'])
71 res1 = mc1.tukeyhsd()
72 print(res1.summary())
73
```

Aqui vemos uma tabela exibindo as comparações emparelhadas post hoc Tukey. Ou seja, diferenças na quantidade de tabagismo para cada par de grupos étnicos. Na primeira linha da tabela, vemos a comparação entre o grupo étnico um e dois. Indivíduos endossando etnia branca versus aqueles que endossam etnia negra. Assim como diferenças médias no número de cigarros fumados entre estes dois grupos. Python calculou um valor P, embora não seja exibido, que leva as múltiplas comparações em consideração e nos protege de inflar nosso erro tipo 1 e rejeitar a hipótese nula quando a hipótese nula é verdadeira. Na última coluna, podemos determinar quais grupos étnicos fumam significativamente diferente do número médio de cigarros que os outros identificando as comparações nas quais podemos rejeitar a hipótese nula, isto é, em que rejeitar é igual a verdadeiro. Assim, podemos ver que o grupo étnico um é significativamente diferente dos grupos étnicos dois, quatro e cinco. E quando examinamos novamente meios de grupo, podemos dizer que indivíduos endossando etnia branca, grupo um, fumam significativamente mais cigarros por mês, do que indivíduos endossando etnia negra, asiática e hispânica. Grupos dois, quatro e cinco.

Multiple Comparison of Means - Tukey HSD, FWER=0.05						
group1	group2	meandiff	lower	upper	reject	
1	2	-109.5127	-164.6441	-54.3814	True	
1	3	-57.7984	-172.5914	56.9945	False	
1	4	-124.5279	-222.9229	-26.1329	True	
1	5	-149.0283	-194.89	-103.1665	True	
2	3	51.7143	-71.6021	175.0307	False	
2	4	-15.0152	-123.233	93.2026	False	
2	5	-39.5156	-103.8025	24.7714	False	
3	4	-66.7295	-214.5437	81.0848	False	
3	5	-91.2298	-210.6902	28.2305	False	
4	5	-24.5004	-128.3027	79.302	False	

4.2.6. Teste de Independência Qui-Quadrado

A análise de variância envolveu examinar a relação entre uma variável explicativa categórica e a variável Resposta quantitativa. Em seguida, vamos considerar inferências sobre as relações entre duas variáveis categóricas. O teste estatístico que responderá a esta pergunta é chamado de Teste de Independência Qui-Quadrado. Chi é uma letra grega que se parece com um grande X. Então, às vezes, você verá este teste denotado com um X ao quadrado.

Para esta ferramenta estatística, vamos começar com um novo exemplo. No início da década de 1970, um jovem desafiou uma Lei Estadual de Oklahoma que proibia a venda de 3,2 cerveja, homens com menos de 21 anos de idade. Mas permitiu que fosse venda a mulheres na mesma faixa etária. O caso foi finalmente ouvido pelo Supremo Tribunal dos EUA. A principal justificativa fornecida por Oklahoma para a lei era a segurança do trânsito. Uma das três principais peças de dados apresentadas ao tribunal foi o resultado de uma pesquisa aleatória na estrada que registrou informações sobre gênero. E se o motorista estava ou não bebendo álcool em nas duas horas anteriores. Houve um total de 619 motoristas com menos de 20 anos de idade incluídos na pesquisa. Abaixo representamos uma tabela bidirecional resumindo os relatos observados na pesquisa na estrada.

Gênero	Sim	Não	Total
Masculino	77	404	481
Feminino	16	122	138
Total	93	526	619

Nossa tarefa é abordar se esses resultados fornecem evidências de uma significativa ou estatisticamente significativa relação entre gênero e direção embriagada. Ambas as variáveis são duas variáveis categóricas valorizadas e, portanto, nossa tabela de duas vias de contagens observadas é um dois por dois.

O procedimento Qui-Quadrado não se limita a duas situações. Ele também pode ser usado para um número maior de categorias explicativas. A chave para relatar resumos apropriados para uma tabela bidirecional é decidir qual das duas variáveis categóricas desempenha o papel da variável explicativa. E, em seguida, calculando as percentagens condicionais separadamente. Ou seja, as percentagens da variável de resposta para cada valor da variável explicativa.

Neste caso, uma vez que a variável explicativa é gênero, calculamos a porcentagem de motoristas que beberam e não beberam álcool para machos e para fêmeas separadamente. Aqui está a tabela das percentagens condicionais. Para os 619 motoristas da amostra, verificou-se que um percentual maior de homens era embriagado do que as mulheres, 16% versus 11,6%. Nossos dados em outras palavras, fornece algumas evidências de que a condução embriagada está relacionada ao gênero. No entanto, isso por si só não é suficiente para concluir que tal relação existe em uma população maior de motoristas com menos de 20 anos.

Precisamos investigar mais os dados e decidir entre os dois pontos de vista a seguir. Que não há diferença na taxa de condução embriagada entre homens e mulheres com menos de 20 anos, nossa hipótese nula. Ou que há uma diferença na taxa de condução embriagada entre homens e mulheres com menos de 20 anos, nossa hipótese alternativa. Em outras palavras, é a evidência fornecida pela pesquisa na estrada, 16% versus 11,6%, forte o suficiente para concluir além de uma dúvida razoável que deve ser devido a uma relação entre dirigir bêbado e gênero na população de motoristas menores de 20 anos. Ou a evidência fornecida pelo inquérito à beira da estrada não é suficientemente forte para chegar a essa conclusão? E isso poderia ter acontecido por acaso?

Isso se deve à variabilidade da amostragem e não necessariamente porque existe uma relação na população. Estas são as hipóteses alternativas nulas e para o teste de independência qui-quadrado. Aqui estão outras maneiras que a hipótese nula e alternativa pode ser declarada para um teste qui-quadrado de independência. Não há relação entre as duas variáveis categóricas. Eles são independentes. Ou, há uma relação entre as duas variáveis categóricas. Eles não são independentes.

Algebricamente, a independência entre gênero e dirigir bêbado equivale a ter proporções iguais de quem bebeu ou não bebeu para homens versus mulheres. Na verdade, a hipótese nula e alternativa poderia ser reformulada, já que a proporção de motoristas homens bêbados é igual à proporção de motoristas mulheres bêbadas. Ou a proporção de motoristas bêbados masculinos não é igual à proporção de mulheres motoristas bêbadas.

A ideia por trás do teste de independência qui-quadrado, muito parecido com a análise da variância é medir o quanto longe os dados estão do que é reivindicado na hipótese nula. Quanto mais longe os dados estiverem da hipótese nula, mais evidências os dados apresentam contra ela. Aqui, os dados de gênero e condução embriagada são representados pelas contagens observadas. Para representar a hipótese nula, vamos calcular outro conjunto de contagens. As contagens que esperaríamos ver, em vez das observadas.

Se dirigir bêbado e sexo eram realmente independentes. Ou seja, se a hipótese nula fosse verdadeira. Por exemplo, nós realmente observamos 77 homens que dirigiam bêbados. Se dirigir bêbado e sexo fossem realmente independentes, se a hipótese nula fosse verdadeira, quantos motoristas bêbados do sexo masculino esperaríamos ver em vez de 77?

Também faremos o mesmo tipo de pergunta sobre as outras três células em nossa tabela. Se a hipótese nula fosse verdadeira, quantas motoristas bêbadas esperaríamos ver em vez de 16? Quantos não bêbados dirigindo machos esperaríamos ver em vez de 404? Quantas mulheres dirigindo não bêbadas esperaríamos ver em vez de 122?

Em outras palavras, teremos dois conjuntos de contagens. As contagens observadas, que são os dados. E as Contagens Esperadas, se a hipótese nula fosse verdadeira. Vamos medir o quanto longe estão as contagens observadas das esperadas. Basearemos nossa decisão no tamanho da discrepância entre o que observamos e o que esperaríamos observar, se a hipótese nula fosse verdadeira. Como as contagens esperadas foram calculadas?

Se os eventos A e B forem independentes, a probabilidade de A e B é igual à probabilidade de A vezes a probabilidade de B. Usamos esta regra para calcular contagens esperadas uma célula de cada vez. Aplicando a regra à primeira célula superior esquerda. Se dirigir bêbado e gênero são independentes, então a probabilidade de um homem ter bebido é igual à probabilidade de ser bêbado vezes a probabilidade de ser homem. Ao dividir as contagens em nossa tabela, vemos que a probabilidade de ter bebido é igual a 93 dividido por 619. E a probabilidade de ser homem é 481 dividida por 619. Então a probabilidade de estar bêbado e ser homem é 93 dividido por 619 vezes 481 dividido por 619. Portanto, uma vez que há um total de 619 motoristas. Se a

condução bêbada e o sexo fossem independentes, a contagem de motoristas bêbados do sexo masculino que esperaríamos ver são os seguintes.

$$P(A \text{ AND } B) = P(A) * P(B)$$

$$P(\text{DRUNK AND MALE}) = P(\text{DRUNK}) * P(\text{MALE})$$

$$P(\text{DRUNK}) = 93/619$$

$$P(\text{MALE}) = 481/619$$

$$P(\text{DRUNK AND MALE}) = (93/619) * (481/619)$$

Portanto, a fórmula para calcular Contagens Esperadas é Total da Coluna vezes Total da Linha dividido pelo Total da Tabela. Seguindo esta fórmula, aqui estão as tabelas completas de Contagens Esperadas e Observadas.

Drank Alcohol in the Last 2 Hours				
Gender (x)	Yes	No	Total	
Male	77	404	408.7	481
Female	16	122	117.3	138
Total	93	526	619	

Importante, o único número que resume a diferença geral entre Observadas e Contagens Esperadas é a estatística qui-quadrado denotada como chi ou X^2 . O que nos diz de forma padronizada, o quanto longe o que observamos, que é os dados são. Do que esperaríamos observar, se a hipótese nula fosse verdadeira. Aqui está a fórmula.

$$\chi^2 = \sum_{\text{all cells}} \frac{(Observed Count - Expected Count)^2}{Expected Count}$$

Para cada célula, tomamos a Contagem Observada, subtraímos a Contagem Esperada e elevamos ao quadrado esse valor. Este valor é dividido pela contagem esperada e, em seguida, este número é somado para todas as células na tabela. Uma vez que a estatística qui-quadrado tenha sido calculada, podemos ter uma sensação de seu tamanho. No nosso caso, o valor de $X^2 = 1,62$. Existe uma diferença relativamente grande entre o que observamos e o que a hipótese nula afirma? Ou relativamente pequena? Acontece que para dois casos como o nosso, estamos inclinados a chamar a estatística qui-quadrado grande se for maior que 3,84. Portanto, nossa estatística de teste não é grande, indicando que os dados não são diferentes o suficiente da hipótese nula para nós rejeitá-la. Para casos diferentes de dois por dois, há cortes diferentes para o que é considerado grande, que são determinados pela distribuição nula nesse caso. Assim, vamos confiar apenas no valor p para conclusões.

Mesmo que não possamos realmente usar a estatística qui-quadrado, foi importante aprender sobre isso, já que engloba a ideia por trás do teste. O valor de p para o teste de independência do qui-quadrado é a probabilidade de obter contagens como as observadas, assumindo que as duas variáveis não estão relacionadas. Que é o que é reivindicado pela hipótese nula. Quanto menor o valor p, mais surpreendente seria obter contagens como fizemos, se a hipótese nula fosse verdadeira. Tecnicamente, o valor p é a probabilidade de observar um qui-quadrado pelo menos tão grande quanto o observado. Usando nosso software estatístico, descobriremos que o valor de p para este teste é 0,201. O valor de p de 0,201 não é pequeno.

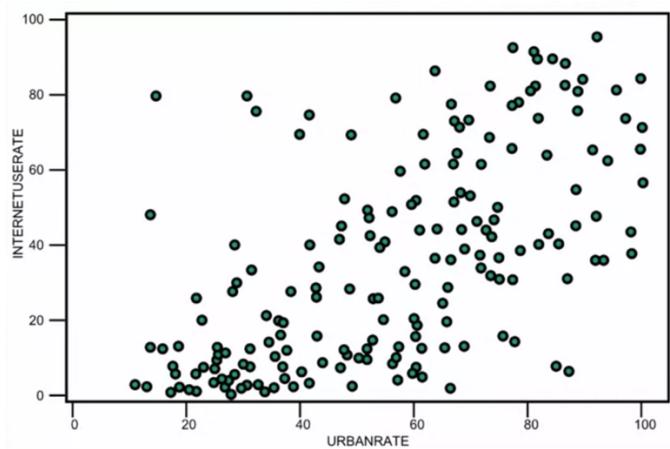
Não há evidência estatística convincente para rejeitar a hipótese nula. E assim continuaremos a assumir que pode ser verdade. Gênero e condução embriagada podem ser independentes. E assim os dados sugerem que

uma lei que proíbe a venda de 3,2% de cerveja a homens e permite às mulheres é injustificada. Na verdade, o Supremo Tribunal, por um voto de sete a dois maioria derrubou a Lei de Oklahoma como discriminatória e injustificada.

4.2.7. Teste de Correlação de Pearson

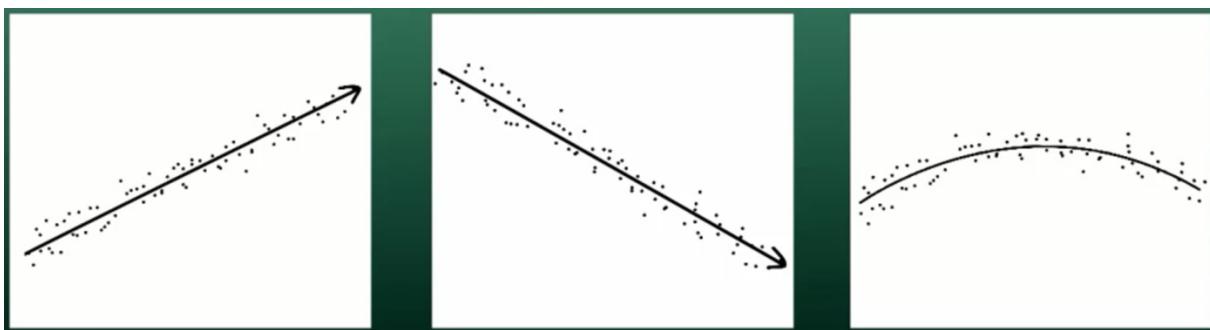
Análise de Variância examina a relação entre uma variável explicativa categórica e uma variável de resposta quantitativa, na qual analisamos a primeira ferramenta inferencial. O teste de independência Qui Quadrado é uma ferramenta inferencial que examina a relação entre dois valores categóricos. Se você tem uma variável explicativa quantitativa e uma variável de resposta categórica, para o propósito deste curso eu encorajo você a categorizar a variável explicativa quantitativa e usar este teste de independência de qui quadrado para examinar este tipo de associação.

A próxima ferramenta inferencial que vamos olhar é usada para examinar a associação entre duas variáveis quantitativas. A Correlação de Pearson. Já discutimos anteriormente que um gráfico de dispersão é a maneira apropriada de gráfico ou visualizar duas variáveis quantitativas quando você deseja examinar a relação entre elas. Vamos primeiro rever brevemente Scatterplots e como interpretá-los. Para criar um gráfico de dispersão, cada par de valores é plotado de modo que o valor da variável explicativa x, seja plotado no eixo horizontal e o valor da variável de resposta y, seja plotado no eixo vertical. Em outras palavras, cada indivíduo aparece no gráfico de dispersão como um único ponto cuja coordenada x é o valor da variável explicativa para esse indivíduo, e cuja coordenada y é o valor da variável de resposta. Ao descrever o padrão geral do relacionamento, vamos olhar para sua direção, forma e força.

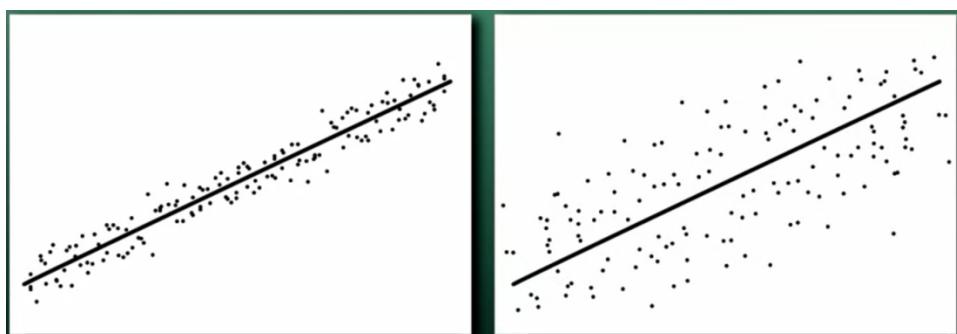


A direção do relacionamento pode ser positiva, negativa ou nenhuma delas. Um relacionamento positivo, ou crescente, significa que um aumento em uma das variáveis está associado a um aumento na outra. Um negativo, ou diminuição no relacionamento significa que um aumento em uma das variáveis está associado a uma diminuição na outra. Nem todos os relacionamentos podem ser classificados como positivos ou negativos. A forma da relação é a sua forma geral. Ao identificar o formulário, tentamos encontrar a maneira mais simples de descrever a forma do gráfico de dispersão. Existem muitas formas possíveis. Aqui estão alguns que são bastante comuns.

Relacionamentos com uma forma linear são mais simplesmente descritos como pontos espalhados sobre uma linha. Relacionamentos com uma forma curvilínea são mais simplesmente descritos como pontos dispersos em torno da mesma linha curva. Por definição, o coeficiente de correlação mede uma relação linear entre duas variáveis quantitativas. Portanto, neste momento, não nos preocuparemos com curvilíneos ou quaisquer outras formas possíveis que um gráfico de dispersão possa tomar. A força do relacionamento é determinada pela proximidade com que os dados seguem a forma do relacionamento.



Esses dois gráficos de dispersão abaixo exibem relações lineares positivas. A força do relacionamento é determinada pela proximidade com que os pontos de dados seguem o formulário. Pontos de dados no gráfico de dispersão à esquerda seguem o padrão linear bastante de perto. Este é um exemplo de uma relação forte. Pontos de dados no gráfico de dispersão à direita também seguem o padrão linear, mas muito menos de perto. Portanto, podemos dizer que o relacionamento é mais fraco em geral. Embora avaliar a força de um relacionamento apenas olhando para o gráfico de dispersão é bastante problemático. Precisamos de uma medida numérica para nos ajudar com isso.



A medida numérica que mede a força de uma relação linear entre duas variáveis quantitativas é chamada de coeficiente de correlação. E é denotado por um r minúsculo. O valor de r varia de -1 a +1. Não surpreendentemente valores negativos de r indicam uma direção negativa para uma relação linear entre as duas variáveis. E valores positivos indicam uma direção positiva para a relação linear. Valores próximos a 0, sejam negativos ou positivos. Indique uma relação linear fraca. E valores próximos a -1 ou próximos a +1 indicam uma forte relação linear. Negativo ou positivo.

Material complementar

Blog: Mehta, A. (2019). [Descriptive Vs Inferential Statistics: Which Is Better & Why.](https://www.digitalvidya.com/blog/descriptive-vs-inferential-statistics/) (7 min)

<https://www.digitalvidya.com/blog/descriptive-vs-inferential-statistics/>

Article: Laerd Statistics. (n.d.). [Descriptive and Inferential Statistics.](https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php) (5 min)

<https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php>

Video: The Organic Chemistry Tutor. (2019). [Descriptive Statistics vs Inferential Statistics.](https://www.youtube.com/watch?v=VHYOuWu9jQI) (7 min)

<https://www.youtube.com/watch?v=VHYOuWu9jQI>

Exercício

1. Os dados _____ incluem dados bem definidos com padrões facilmente identificáveis.

- a) Não estruturado
- b) Estruturada
- c) Semi-estruturado

2. Qual das opções a seguir é um exemplo de dados não estruturados?

- a) Arquivos de imagem
- b) Informações da transação
- c) Números de segurança social

3. Qual dos seguintes é considerado dados quantitativos?
- a) Nominal
 - b) Discreto
 - c) Ordinal
4. Qual dos seguintes é considerado dados qualitativos?
- a) Contínuo
 - b) Discreto
 - c) Ordinal
5. A preparação de dados inclui todas as etapas a seguir, exceto:
- a) Extrair
 - b) Transformar
 - c) Carregar
 - d) Todas essas opções são etapas adequadas na preparação de dados
6. As estatísticas _____ permitem a sumarização e a representação gráfica de um conjunto de dados.
- a) Descritivo
 - b) Inferencial
 - c) Regressão
7. _____ é usado para explicar as médias de um ponto de dados.
- a) distorção
 - b) Correlação
 - c) Tendência Central
8. A estatística descritiva permite que um analista generalize os resultados de uma amostra para uma população inteira.
- a) Verdadeiro
 - b) Falso
9. O objetivo da estatística _____ é inferir e generalizar conclusões de uma amostra para uma população inteira.
- a) Descritivo
 - b) Regressão
 - c) Inferencial
10. Qual dos seguintes é usado para fazer previsões com base em valores dentro de um conjunto de dados de amostra?
- a) Teste de hipóteses
 - b) Correlação
 - c) Modelos de regressão

5. Business Intelligence e Visual Analytics

5.1. Visualização e Análise de Dados

A necessidade de visualização de dados para relatórios de negócios

Existe um ditado bem estabelecido: “Uma imagem vale mais que mil palavras”. Agora, imagine que você está percorrendo milhares de linhas de dados tabulares para coletar informações pertinentes aos negócios para tomar uma decisão. Esta tarefa cansativa pode levar horas! E então o que você faz quando os dados são atualizados? Recomeçar? Este é um exemplo de porque a visualização de dados pode ser tão importante e impactante.

A visualização de dados fornece uma imagem que descreve os dados, permitindo que você tome decisões mais rápidas e precisas. Padrões claros geralmente surgem e podem ser reconhecidos mais facilmente por meio da visualização de dados, e muitas vezes você pode reter e explicar melhor a saída por meio de uma visualização pictórica das informações. A visualização de dados permite criar uma representação visual (ou imagem) de informações para um conjunto de dados ou coleção de fontes de dados. Esse processo traz clareza, envolvimento do usuário, insights eficazes e tomada de decisão informada aos dados. Muitas ferramentas de inteligência de negócios existem hoje para aprimorar e permitir a criação eficiente de visualizações de dados.

Tipos de visualização de dados

A visualização eficaz de dados é tanto “arte” quanto “ciência”. Existem inúmeras visualizações que são usadas para representar dados. Um dos principais desafios é selecionar o tipo adequado de visualização para comunicar efetivamente a história que está sendo contada por meio dos dados.

As visualizações de dados geralmente podem ser categorizadas em sete tipos:

- **Linear** (1-dimensional): listas de itens
- **Planar** (2-dimensional): mapas geoespaciais
- **Volumétrico** (3-dimensional): renderizações de superfície e volume, simulações de computador, modelos 3D, etc.
- **Temporais**: linhas do tempo, gráficos de séries temporais, gráficos de Gantt, etc.
- **Multidimensional**: gráficos de pizza, histogramas, nuvens de tags, gráficos de barras, gráficos de linhas, gráficos de dispersão, etc.
- **Hierárquico**: árvores
- **Rede**: matrizes, diagramas nó-link, etc.

Cada tipo de visualização de dados tem sua finalidade exclusiva e caso de melhor uso. Algumas categorias, como temporal e multidimensional, são muito mais comumente encontradas em visualizações de dados, painéis e infográficos hoje. Outros são bastante complexos e normalmente usados em domínios altamente científicos.

5.2. Qual visualização é boa para que propósito?

Com tantos tipos de visualizações disponíveis, como você decide qual visualização é uma representação eficaz para seu propósito específico? Embora existam ferramentas disponíveis para ajudar a orientar sua seleção, parte da decisão dependerá de sua experiência, dos requisitos do trabalho e até de tentativa e erro. É aí que entra a “arte” da visualização de dados. É crucial selecionar uma visualização que contribua para a história dos dados, seja rapidamente compreendida pelo público e mostre uma imagem clara e precisa dos dados.

Uma ferramenta eficaz que ajuda a determinar o tipo de gráfico, gráfico ou visualização a ser exibido é o Catálogo de Visualização de Dados¹. Esta ferramenta organiza visualizações por várias funções, por exemplo,

¹ Catálogo de Visualização de Dados - <https://datavizcatalogue.com/search.html>

comparação, relacionamento, distribuição, dados ao longo do tempo, etc. Por exemplo, se você precisa visualizar diferenças ou semelhanças entre valores em um conjunto de dados, você pode selecionar a função Comparações na Visualização de Dados Catálogo. Fazê-lo apresenta dois grupos de visualizações: “Com eixo” ou “Sem eixo”. Se você deseja visualizar dados quantitativos em um período de tempo, pode selecionar a opção Gráfico de linhas na categoria “Com um eixo”. Uma vez selecionada, a ferramenta fornece informações descritivas sobre o gráfico/tipo de gráfico selecionado, bem como seleções de gráficos adicionais que podem ser adequadas para seus dados.

Visão geral de Análise Visual

A análise visual emprega visualizações de dados para apoiar o raciocínio analítico e o desenvolvimento de ferramentas e processos para analisar conjuntos de dados. A análise visual geralmente produz padrões e insights que podem não surgir tão facilmente por outros meios analíticos. As visualizações de dados geralmente respondem a perguntas sobre “o quê”, enquanto a análise visual mergulha no “porquê” mais profundo da exploração de dados. Essa abordagem se presta ao aprendizado profundo sobre o conjunto de dados e à compreensão dos padrões emergentes, anomalias e relacionamentos intrincados entre os pontos de dados. A análise visual agrega valor ao permitir que o usuário altere parâmetros rapidamente, explore visualizações de dados para explorar por que um gráfico se parece com ele ou forneça visualizações alternativas de visualizações de dados com o mínimo de esforço. A análise visual é poderosa por causa de sua flexibilidade, capacidade de atualizações em tempo real e eficiência na exploração de dados, o que permite descobrir padrões inesperados que impulsionam os “porquês” por trás dos dados.

O cenário das ferramentas de análise visual

Muitas ferramentas de análise visual existem hoje no mercado, e a demanda por essas ferramentas está aumentando exponencialmente à medida que as organizações descobrem a necessidade de explorar e aprender profundamente com os dados que coletam há muitos anos. Ferramentas poderosas que anteriormente exigiam investimentos significativos em hardware, redes e infraestrutura de TI agora estão disponíveis por meio de soluções baseadas em nuvem, navegadores da Web e dispositivos móveis. Toda essa inovação ajuda a aliviar o fardo da análise, colocando o poder da análise visual nas mãos de cada usuário e levando a soluções mais fáceis de usar e econômicas.

Uma das plataformas de análise visual mais populares e poderosas do mercado atualmente é o SAS Visual Analytics e o SAS Viya². Essa solução baseada em nuvem fornece análises visuais ao usuário, ajudando-o a criar insights poderosos sobre os dados, recursos de relatório de dados e ferramentas de exploração de dados. Passaremos o restante deste curso nos familiarizando com essa plataforma e ganhando experiência prática no desenvolvimento de soluções analíticas.

Atividade: Seleção de visualização do conjunto de dados

Usando uma ferramenta como o kaggle, selecione um conjunto de dados de seu interesse e descreva uma visualização eficaz que derivaria valor e insights com base nos pontos de dados disponíveis. Certifique-se de aplicar as leituras do módulo à sua avaliação e escolha de seleção.³

Exercício

1. _____ fornece uma imagem que descreve os dados, permitindo que você tome decisões mais rápidas e precisas.
 - a) Data Analysis
 - b) Data Visualization
- c) Statistics
2. Qual das opções a seguir é um benefício da visualização de dados?
 - a) Insights eficazes
 - b) Tomada de decisão informada

² SAS Visual Analytics and SAS Viya - https://www.sas.com/en_us/software/visual-analytics.html

³ Kaggle - <https://kaggle.com/datasets>

- c) Todas essas opções estão corretas
3. Um dos principais desafios da visualização de dados é selecionar o tipo adequado de visualização para comunicar efetivamente _____.
a) Histórias
b) Finanças
c) Erros nos dados
4. Qual tipo de visualização de dados inclui histogramas e gráficos de dispersão?
a) Multidimensional
b) Planar
c) Temporal
5. A seleção da visualização mais eficaz depende significativamente da experiência, requisitos do trabalho e tentativa e erro.
a) Falso
b) Verdadeiro
6. A seleção de uma visualização que contribui para a história dos dados diminui a probabilidade de o público entender rapidamente os dados.
a) Verdadeiro
b) Falso
7. _____ emprega visualizações de dados para apoiar o raciocínio analítico e o desenvolvimento de ferramentas e processos para analisar conjuntos de dados.

- a) Visual Analytics
b) Business Intelligence
c) Data Visualization
8. Um objetivo principal da análise visual é descobrir _____ e intrincadas relações entre os pontos de dados.
a) Falso-positivo
b) Padrões emergentes
c) Dados incorretos
9. Muitas ferramentas de análise visual estão agora disponíveis como soluções baseadas em nuvem para aumentar a disponibilidade e oferecer poder adicional aos usuários finais.
a) Verdadeiro
b) Falso
10. O SAS Viya é uma plataforma em nuvem que fornece _____ aos usuários, ajudando-os a criar insights poderosos sobre os dados, recursos de relatórios de dados e ferramentas de exploração de dados.
a) Visual Analytics
b) Data visualizations
c) Data Warehousing

Material complementar

Book: SAS Institute. (2019). Exploring SAS Viya: Visual Analytics, Statistics, and Investigations. (1 hour)
<< <https://support.sas.com/content/dam/SAS/support/en/books/free-books/exploring-sas-viya-va-statistics-investigations.pdf> >>

5.3. Uma visão geral do Public Tableau ou Power BI???