“R” Com Data Science

Profª Gabrielle Gomes dos Santos Ribeiro

2023

ESTATÍSTICA

A estatística é um ramo da Matemática que fornece métodos para coleta, organização, descrição, análise e interpretação de dados, para utilização dos mesmos na **tomada de decisões**.

A estatística está presente nos fenômenos e fatos do nosso dia a dia, mais do que imaginamos. Praticamente todas as **informações divulgadas pelos meios de comunicação** provêm de alguma forma de pesquisas e estudos estatísticos. O crescimento populacional, os índices de inflação, emprego e desemprego, o custo da cesta básica, os índices de desenvolvimento humano são alguns exemplos de pesquisas divulgadas pelos meios de comunicação e que se utilizam dos métodos estatísticos.

A teoria das probabilidades

- Recomendação: Filme **Quebrando a Banca**
- *É inspirado na história verídica de jovens brilhantes dos Estados Unidos - e de como eles ganharam milhões em Las Vegas de acordo como o relatado em Bringing Down the House, livro best-seller de Ben Mezrich.*

A teoria das probabilidades

A ESTATÍSTICA

A Estatística tem duas grandes áreas:

- **Estatística Descritiva**: envolve o resumo e apresentação de dados (utilizando gráficos e tabelas, por exemplo);

- **Estatística Inferencial**: ajuda a concluir sobre conjuntos maiores de dados (populações) quando apenas partes desses conjuntos (as amostras) foram estudadas

Análise de Dados

- **Importância??**

Produzimos um fluxo constante e exaustivo de dados digitais. Estima-se que 90% dos dados armazenados no mundo foram produzidos apenas nos últimos dois anos e os rastros desses dados continuam duplicando a cada ano.

Metodologias estatísticas podem ser aplicadas para que os dados sejam transformados em informações, e essas informações/resultados sejam utilizados para tomar decisões inteligentes a favor de uma empresa/negócio.

EM QUE MOMENTO ENTRA A ESTATÍSTICA NO MEU TRABALHO?

MÉTODO ESTATÍSTICO

- DEFINIÇÃO DO PROBLEMA
- PLANEJAMENTO DA PESQUISA
- COLETA DE DADOS
- APURAÇÃO DOS DADOS
- CRÍTICA DOS DADOS
- APRESENTAÇÃO DOS DADOS
- ANÁLISE DE RESULTADOS

FASES DO MÉTODO ESTATÍSTICO

1. DEFINIÇÃO DO PROBLEMA

O que pesquisar?

Nesta etapa você deve conhecer o problema a ser pesquisado e fazer as perguntas que você quer responder com sua pesquisa.

2. PLANEJAMENTO DA PESQUISA

Como pesquisar?

Nessa fase é essencial que você tenha clareza de como a pesquisa será feita, quais serão os procedimentos metodológicos.

FASES DO MÉTODO ESTATÍSTICO

3. COLETA DE DADOS

Nessa etapa você vai obter as informações (coletar os dados) de acordo com o que foi planejado na etapa anterior.

Você ter claro qual a sua população de interesse, o tamanho da amostra que você irá coletar e qual método de coleta irá utilizar (Simples, Sistemático, Estratificado, etc..)

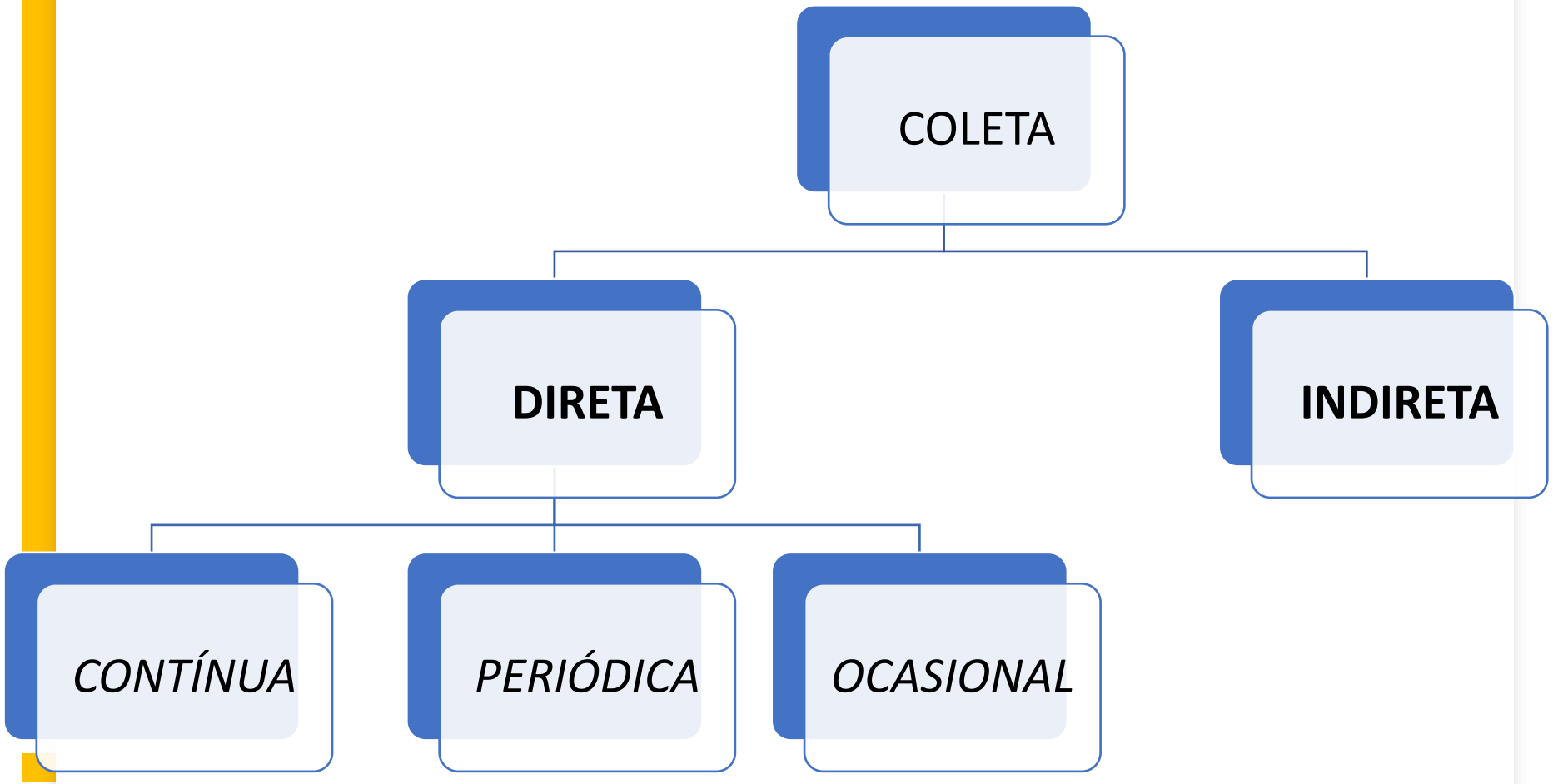
COLETA DE DADOS

Lembrando que, **População** é o conjunto de elementos (pessoas, objetos, animais) que tem pelo menos uma característica (de interesse) observável em comum.

POPULAÇÃO  **AMOSTRA**

- **Amostra -** subconjunto ou parte da população escolhido segundo determinadas regras e critérios
- A partir das informações retiradas da amostra são feitas afirmações sobre a população!
- **Principais vantagens: economia de tempo, redução dos custos e a obtenção de resultados menos propícios ao erro.**

COLETA DE DADOS



COLETA DE DADOS DIRETA

- **<u>CONTÍNUA</u>**: quando feita continuamente,
- ***Exemplos:*** Nascimentos e óbitos, frequência dos alunos em aula

- **<u>PERIÓDICA:</u>** quando feita em intervalos constantes de tempo,
- ***Exemplos***: como os censos e avaliações bimestrais de alunos

- **<u>OCASIONAL:</u>** feita extemporaneamente, a fim de atender a uma conjuntura ou a uma emergência
- Exemplos: epidemias

COLETA DE DADOS INDIRETA

- Quando é feita com base em elementos já pesquisados
- *Exemplos*: base de jornais, sites, revistas, etc.

FASES DO MÉTODO ESTATÍSTICO

4. APURAÇÃO DOS DADOS

- Nada mais é do que a soma e o processamento dos dados obtidos e a disposição mediante critérios de classificação.
- Normalmente feita de forma eletrônica, por meio de criação de banco eletrônico (Excel).
- FASE DE EXTREMO CUIDADO (uso de virgulas, pontos, multiplicação de planilhas, cores)

FASES DO MÉTODO ESTATÍSTICO

5. CRÍTICA DOS DADOS

Os dados estão coerentes?

- Dados devem cuidadosamente criticados, a procura de irregularidades, falhas e imperfeições, que possam influir sensivelmente nos resultados.

- Assim, se detectar algum erro, poderá evitar que seja repetido nas coletas futuras e corrigir para que conclusões erradas sejam feitas sobre a população de estudo.

FASES DO MÉTODO ESTATÍSTICO

6. APRESENTAÇÃO DOS DADOS

- Após organizar os dados em planilhas eletrônicas, você deve apresentá-los.

- É nesta etapa que você irá **resumir e organizar** os dados coletados por meio de gráficos, tabelas ou medidas numéricas e a partir deste resumo, examinar a variável que está sendo objeto de estudo e procurar algum padrão nas observações.

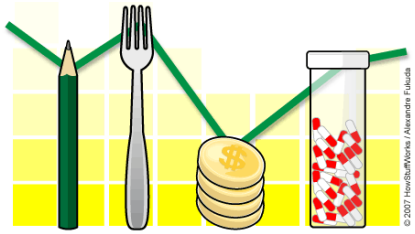
FASES DO MÉTODO ESTATÍSTICO

7. ANÁLISE DOS RESULTADOS

- É a parte final do método estatístico.
- Você deve descrever e analisar os dados pesquisados e, chegar a uma conclusão, ou seja, responder à sua pergunta inicial.

Na Atualidade

• A estatística está presente em diversos fenômenos e fatos do nosso dia a dia, mais do que imaginamos. Praticamente todas as **informações divulgadas pelos meios de comunicação** provêm de alguma forma de pesquisas e estudos estatísticos. O crescimento populacional, os índices de inflação, emprego e desemprego, o custo da cesta básica, os índices de desenvolvimento humano são alguns exemplos de pesquisas divulgadas pelos meios de comunicação e que se utilizam dos métodos estatísticos.

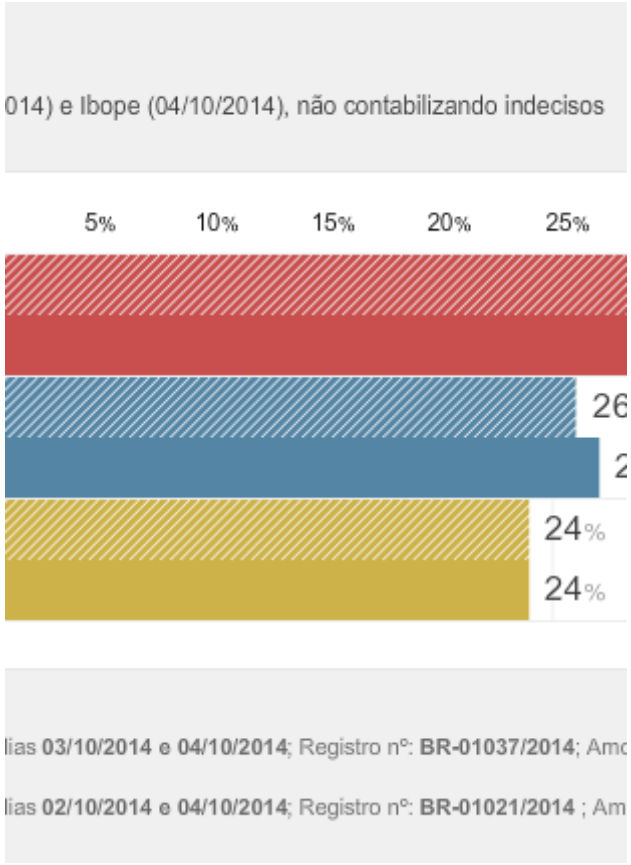


Na atualidade

- Em algumas situações está evidente, em outras nem tanto.
- Ex: A metorologia utiliza Estatística para a **previsão do tempo**.
- De acordo com o meteorologista Gustavo Escobar, coordenador do Grupo de Previsão de Tempo do Cptec/Inpe
- *"A previsão do tempo começa com um diagnóstico da situação meteorológica, que é observar a condição da atmosfera em um determinado momento. Essa observação só é possível através da análise de dados coletados por estações meteorológicas espalhadas pelo país. Os dados são justamente o que chamamos de variáveis. E a partir dessa avaliação é possível identificar os sistemas meteorológicos que estão atuando sobre a atmosfera, como ciclones, anticiclones, frentes frias ou frentes quentes, por exemplo"*, explica o especialista.

Na atualidade

- Campanhas políticas: Você provavelmente já ouviu falar dos estudos amostrais, a intenção de voto, e as margens de erro (noticiário). Os modelos estatísticos são capazes de prever qual candidato tem mais chance de ganhar, e em quais lugares.
- Nos Estados Unidos, o professor Allan Lichtman ganhou fama ao acertar o resultado de todas as eleições presidenciais, desde 1984, inclusive a última, de Donald Trump, com base em dados históricos e variáveis, como o carisma e o contexto econômico do país ou o êxito das operações militares em curso.



Na atualidade

- **Seguro do seu carro**: O valor que você paga é precificado baseado em estatísticas de outros clientes.
- A Seguradora se baseia em estatísticas de idade, estado civil, cidade, modelo do veículo, local onde mora e trabalha, estacionamento, e muitas outras variáveis, que geram resultados com probabilidades de acontecer.
- Ex: Homens e jovens pagam mais caro pelos seguros, pois se envolvem mais em acidentes.
- Motoristas que já se envolveram em acidentes também pagam mais.

Na atualidade

- **Avaliações do Sistema Educacional**
- Os sistemas de avaliação do Enem e Enade utilizam Estatística para medir as habilidades dos estudantes.
- Você já se perguntou, por exemplo, por quê alunos que acertam a mesma quantidade de questões no Enem, ficam com notas diferentes?

ALUNOS	DIFICULDADE BAIXA	DIFICULDADE MÉDIA	DIFICULDADE ALTA	NOTA
	Luiza acertou os 20 itens mais fáceis, ou seja, obedeceu a um **comportamento coerente** com a régua do Enem, pois é esperado que o aluno acerte os itens mais fáceis e não consiga superar os itens a partir de um determinado nível de dificuldade.			615,8
	Um comportamento próximo do real foi o de Raquel, que dos 20 itens que acertou, a maioria era de menor dificuldade. Esse **comportamento é razoavelmente coerente** e, por isso, sua nota, de acordo com a TRI, ficou maior que as de Rafael.			587,1
	Rafael acertou os 20 itens mais difíceis, um **comportamento muito incoerente**, e, por isso, sua nota foi bastante baixa. De acordo com a TRI, como ele não tem conhecimento para resolver os fáceis, os acertos dos difíceis são interpretados como "chutes".			301,5

Na atualidade

- **Testes de medicamentos**: qualquer droga que esteja à venda em farmácias e drogarias, já foi testada estatisticamente, e validada a sua eficácia. Portanto, se você toma ou já tomou algum medicamento, a estatística já influencia sua vida.

Na atualidade

- **Consumo de produtos**: um supermercado que controla seu estoque com uso de estatísticas, é capaz de calcular o tempo certo de quando e quanto comprar. E até mesmo de escolher um determinado local para colocar seu produto, onde aumente a probabilidade de venda.
- Você já ouviu a história de um supermercado que colocou cervejas do lado de fraldas? Quando as mães pediam para seus maridos comprarem fraldas para os filhos, eles sempre voltavam com cervejas. Genial!

Na atualidade

- **Mercado de ações**: se você souber usar a estatística, a ponto de construir modelos, eles podem ajudar você a prever a economia, e quem sabe ser mais assertivo nas suas compras e vendas de ações daquelas empresas que você nunca sabe o que fazer com elas.

Na atualidade

- **Controle estatístico de processos (Controle de qualidade)**:
- Ferramenta imprescindível como estratégia para prevenção de defeitos, melhoria da qualidade de produtos e serviços e redução de custos.
- Suponha que você trabalha numa fábrica de água sanitária. A legislação da Anvisa especifica os valores máximos para pH da água sanitária, como demonstrado a seguir:

Produto Puro	**Produto Diluído a 1 % (p/p)**
13,5	11,5

- Assim, uma amostra com pH 13,77 seria considerada NÃO CONFORME.
- É necessário garantir que toda uma linha de produção esteja em conformidade com as especificações, com base nas amostras coletadas.

Na atualidade

• O uso de ferramentas estatísticas nas ciências forenses é de extrema importância nos meios judiciais. Os cientistas forenses podem avaliar e interpretar as evidências que incluem elementos de incerteza. Eles, cada vez mais, necessitam do apoio da ciência Estatística nos seus mais diversos ramos.

NA SAÚDE

- A aplicação de estatística na saúde está diretamente relacionada com a **epidemiologia**, que estuda os fatores que determinam a frequência e a distribuição das doenças em grupos de pessoas.

- Também é utilizada para elencar e selecionar **novas tecnologias e soluções inovadoras** relacionadas ao processo saúde-doença, tais como a formulação de diferentes procedimentos cirúrgicos.

- Segundo David Goggon (2015), os procedimentos estatísticos têm contribuído em muitos dos sucessos da medicina moderna, salvando muitas vidas. Por isso, que o máximo de profissionais de saúde deveriam conhecer, pelo menos, os mecanismos básicos relevantes.

- É um fato que os estudos desenvolvidos no âmbito da bioestatística e da epidemiologia são fundamentais para aperfeiçoar a gestão de serviços em saúde de maneira contínua. No entanto, muitos estudantes e profissionais ainda têm a perspectiva de que a sua aplicabilidade e compreensão são complexas.

	FINALIZAÇÕES CERTAS	FINALIZAÇÕES ERRADAS	PASSES ERRADOS	FALTAS COMETIDAS
CORITIBA FOOT BALL CLUB CFC PARANA	6	5	28	13
CLUBE ATLÉTICO PARANAENSE	3	6	21	15

Na atualidade

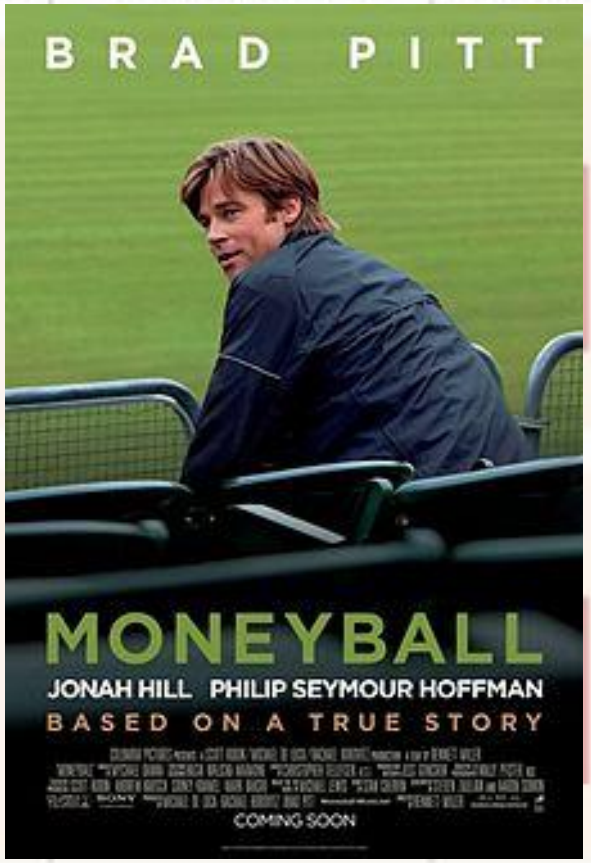
- O uso da Estatística no esporte é muito mais do que o percentual de posse de bola em um jogo de futebol ou o índice de acertos dos arremessos de três pontos em um jogo de basquete.
- A Estatística tem sido usada para a contratação de jogadores de futebol, e até na construção de modelos probabilísticos que preveem os rumos de um campeonato.

Na atualidade

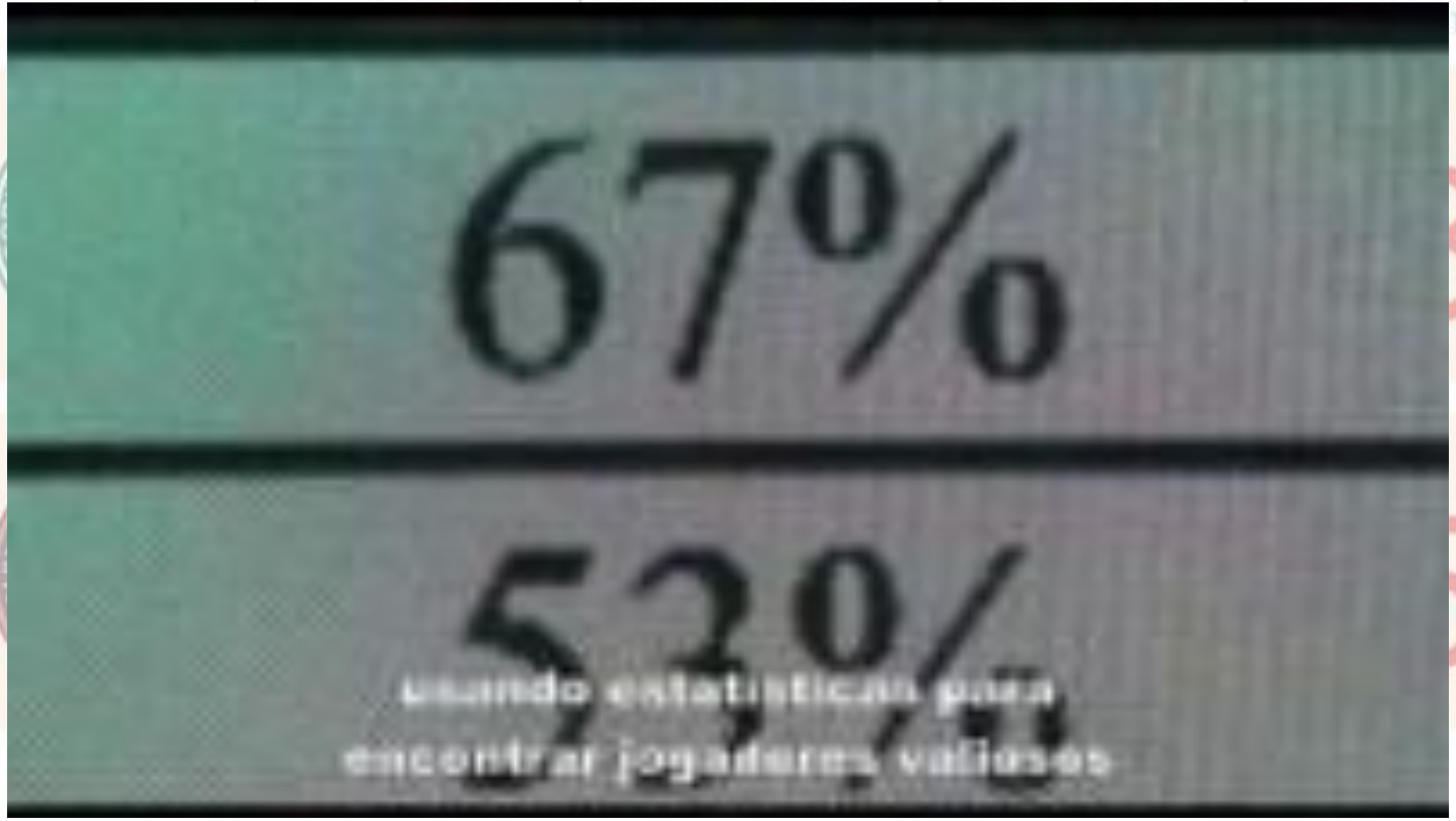
Moneyball (“O homem que mudou o jogo”) é baseado na história verdadeira de Billy Beane, gerente geral do time de basebol do Oakland Athletics.

O filme foca nas tentativas de Beane de criar um time competitivo para a temporada de 2002 de Oakland, apesar da situação financeira desfavorável da equipe, usando uma sofisticada análise estatística dos jogadores.

Na história, o gerente conhece Peter Brand, um jovem economista que tem levantamentos de dados detalhados e peculiares dos jogadores. Ele vira assistente de Billy Beane e, baseado nas estatísticas, o time começa a contratar atletas desacreditados por toda a mídia e torcedores. A tentativa funciona, e o Oakland Athletics vence 20 partidas consecutivas, um recorde da Liga.



Estatística



Na atualidade


"Moneyball": São Paulo se inspira em filme estrelado por Brad Pitt para traçar perfil de reforços

Tricolor foge de badalação no mercado e aposta em estatísticas para montar elenco

- O movimento no São Paulo na última temporada teve algumas semelhanças. Do time campeão da Copa do Brasil, por exemplo, três jogadores tinham sido rebaixados com suas equipes antes de chegar ao Tricolor.

- *"A gente continua apostando no projeto de quando chegamos aqui. O São Paulo não está atrás de grandes nomes, está atrás de grandes jogadores, que possam agregar".*

- Em junho, o São Paulo contratou o **analista de scout João Marcos** e deve ter novas contratações de profissionais.

Algoritmo de recomendações de filmes e séries com base na análise estatística de dados do cliente

Sempre que você acessa o serviço Netflix, o sistema de recomendações tenta ajudar você a encontrar uma série ou filme de forma fácil. A plataforma estima a probabilidade de você assistir a um título em particular do catálogo com base em um número de fatores.

DADOS COLETADOS

- **Sobre o que você assiste:** qual gênero, por quanto tempo, em que dispositivos, datas, locais e horários, como você classifica os conteúdos, com que frequência você pausa, em que momentos para de assistir ou mesmo abandona de vez uma produção.
- **Sobre como você navega:** que termos pesquisa, que trailers vê dentro da plataforma, como utiliza a barra de rolagem, em que você clica, quanto tempo leva para selecionar um filme ou série, o histórico e os cookies do seu browser, tipo de dispositivo usado, comportamento de navegação
- **Sobre quem você é:** seu nome, gênero, e-mail, endereço, telefone, método de pagamento e região, além de dados demográficos comumente providos por fornecedores terceirizados.

Se já notaram que o mesmo filme ou série por vezes aparece com imagens de capa diferentes, isso não está a acontecer por acidente. A Netflix utiliza toda a informação que vai recolhendo sobre os programas que vemos, para tentar determinar que géneros mais apreciamos, apresentando uma imagem "à medida" com estilo mais apelativo para cada cliente. Uma mesma série que para um cliente apreciador de thrillers seja representada num tom sombrio e misterioso, pode para um cliente apreciador de romances aparecer como parecendo ser uma comédia divertida e jovial.

Na atualidade

- A **pandemia de Covid-19** trouxe a atenção à importância do método científico com o uso da Estatística, tanto para a compreensão da dinâmica epidemiológica dos casos/óbitos no país e no mundo, como para a validação de medicamentos e vacinas propostos no combate à doença.
- A área de ANÁLISE DESCRITIVA na estatística é conhecida por ser usada em situações em que são encontradas uma grande quantidade de informações, sendo necessária torná-las compactas para conseguir trabalhar com os dados. (PEREIRA, 2019).

- A estatística ajudou os profissionais da saúde, para o enfrentamento de tomadas de decisões, as análises estatísticas apresentaram que algumas medidas usadas para diminuir o número de contagio do coronavírus tiveram resultado na redução de casos, que por exemplo, o mais adequado foram as políticas de afastamento social. Foram elas que fizeram os governos adotarem essa medida de restrição, que por sua vez, mostrou um melhor diagnóstico e alguns dos países que adotaram essa medida foram Itália e Reino Unido.
- É com base nos dados estatísticos, que os governos formaram discussões em busca de estratégias eficientes de combate à propagação do vírus.
- Foram utilizados modelos matemáticos para estimar o número de casos em diversos cenários, auxiliando os tomadores de decisão, por exemplo, a determinar o número de vagas em Unidades de Terapia Intensiva (UTI) necessárias em determinadas regiões. Ao observar a tendência da quantidade de casos, podemos avaliar o melhor momento para o relaxamento da política de afastamento social.

Gráfico Média Móvel

Data Wrangling

- O conceito *Data Wrangling* é um tanto genérico em nosso português brasileiro, mas poderia ser traduzido como disputa/briga/luta de dados. Esta disputa está intimamente ligada ao processo de **transformação de dados** e isso inclui: obtenção, transformação, limpeza, agregação, visualização e criação de bases limpas para fins Analytics na Ciência de Dados.

Data Wrangling

- Os dados nem sempre chegam ao pesquisador num formato que permita a aplicação direta dos recursos computacionais. Portanto, colocar os dados em formato adequado muitas vezes pode ser tão importante quanto (e até mais trabalhoso) do que a análise em si.

- Usualmente, o cientista de dados parte de uma base “crua” e a transforma até obter uma base de dados analítica.

- A base analítica é necessariamente estruturada e, a menos de transformações simples, está preparada para passar por análises estatísticas.

Data Wrangling

Um conceito importante para obtenção de uma base analítica é o ***data tidying***, ou arrumação de dados. Uma base é considerada ***tidy*** se:

1. Cada linha da base representa uma observação.
2. Cada coluna da base representa uma variável.

Data Wrangling

- A base de dados analítica é estruturada de tal forma que pode ser colocada diretamente em ambientes de modelagem estatística ou de visualização. Nem sempre uma base de dados analítica está no formato ***tidy***, mas usualmente são necessários poucos passos para migrar de uma para outra.

- **Os principais pacotes encarregados da tarefa de estruturar os dados são o dplyr e o tidyr**

Pacote *dplyr*

install.packages("dplyr")
library(dplyr)

- O ***dplyr*** é o pacote mais útil para realizar transformação de dados, aliando simplicidade e eficiência de uma forma elegante.
- O pacote ***dplyr*** fornece ferramentas convenientes para as tarefas mais comuns de manipulação de dados.
- A utilização é facilitada com o emprego do operador %>%.

As principais funções do ***dplyr*** são:

- filter() - filtra linhas
- summarise() - sumariza a base
- select() - seleciona colunas
- mutate() - cria/modifica colunas
- arrange() - ordena a base

Pacote *dplyr*

- **Exemplo: Dados “Milsa.txt”**

dados<-read.table("Milsa.txt", header=T)

- **filter()** - filtra linhas

filter(dados, filhos==1)
Ou
dados %>%
filter(filhos==1)

- **select()** - seleciona colunas

select(dados, filhos, salario)
select(dados, salario:mes)

Pacote *dplyr*

- **mutate()** - cria/modifica colunas. Ela é equivalente à função transform(), mas aceita várias novas colunas iterativamente.

```
dados %>%
select(civil, ano, salario) %>%
mutate(salario2 = salario * 958,
razao=salario/ano)
```

- **arrange()** - ordena a base. O argumento desc= pode ser utilizado para gerar uma ordem decrescente.

```
dados %>%
arrange(desc(ano))
```

Pacote *dplyr*

- **summarise()** - sumariza a base. Ela aplica uma função às variáveis, retornando um vetor de tamanho 1. Geralmente ela é utilizada em conjunto da função *group_by().*

```
dados %>%
group_by(regiao, instr) %>%
summarise (n=n()) %>%
arrange(regiao)
```

	regiao <int>	instr <int>	n <int>
1	1	1	3
2	1	2	7
3	1	3	3
4	2	1	4
5	2	2	6
6	2	3	3
7	3	1	6
8	3	2	6
9	3	3	2

A função n() costuma ser bastante utilizada com a função summarise().

A função count() também pode ser usada para sumarizar em relação à frequência.

```
dados %>%
filter(civil=="1") %>%
count(regiao)
```

	regiao	n
1	1	4
2	2	6
3	3	8

Pacote *tidyr*

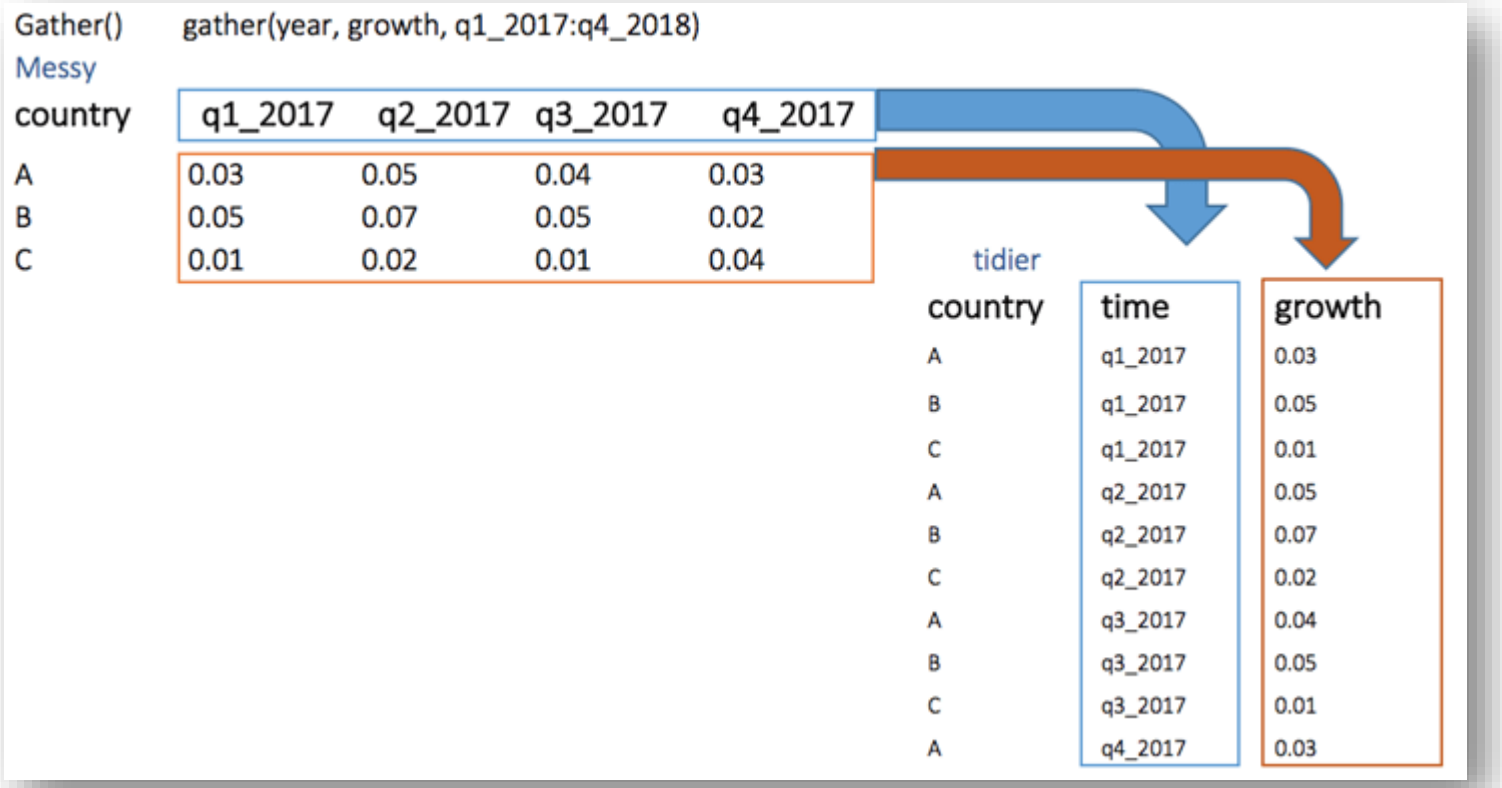
install.packages("tidyr")

- O pacote tidyr dispõe de funções úteis para deixar os seus dados no formato que você precisa para a análise. Na maioria das vezes, utilizamos para deixá-los ***tidy***. Outras, precisamos “bagunçá-los” um pouco para poder aplicar alguma função.

- As principais funções são a gather() e a spread()

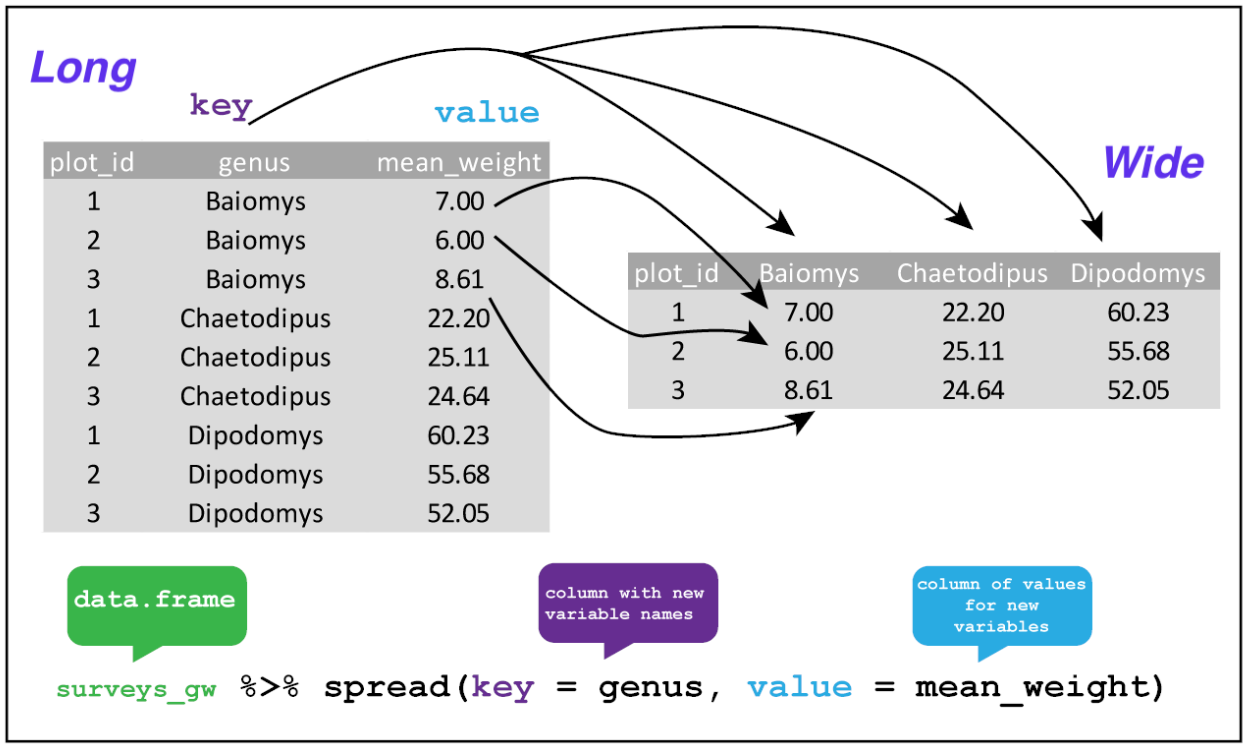
Pacote *tidyr*

- **gather()** - “empilha” o banco de dados. Ela é utilizada principalmente quando as colunas da base não representam nomes de variáveis, mas sim seus valores.



Pacote *tidyr*

- **spread()** - "Joga" uma variável nas colunas. É essencialmente a função inversa de gather



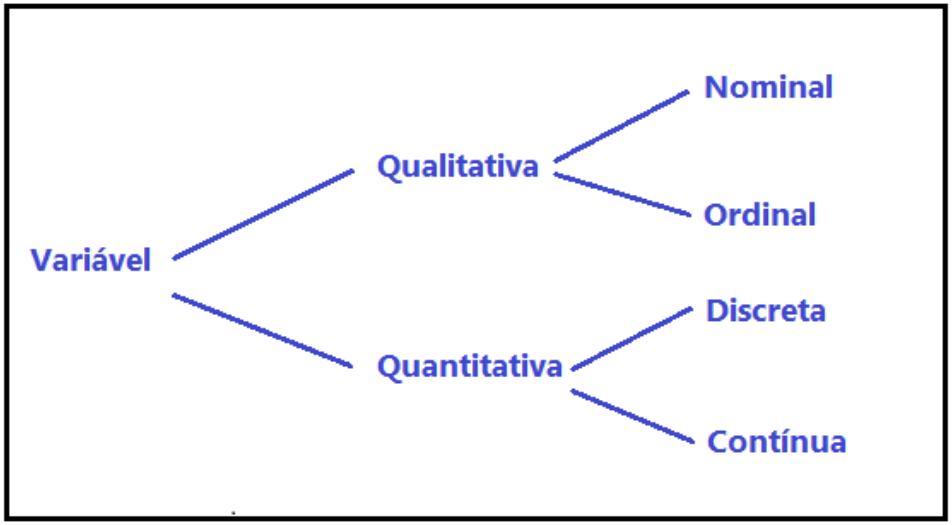
ANÁLISE DOS DADOS

Estatística Descritiva

- Fase inicial do processo de análise dos dados;
- É um conjunto de ferramentas utilizados para organizar, resumir e descrever as principais características observadas em um conjunto de dados, permitindo ao pesquisador melhor compreensão do comportamento dos dados;
- As ferramentas descritivas são:
  - Tabelas
  - Gráficos
  - Medidas-resumo: porcentagens, média, moda, mediana, variância, desvio-padrão coeficiente de variação, entre outras.

Classificação de variáveis

Para escolher o gráfico ou a medida mais adequada para representar uma variável, é importante verificar primeiro o tipo da variável:



Classificação de variáveis

QUANTITATIVAS → descrevem **quantidades** e têm seus valores expressos **numericamente**

Discretas: assumem valores pertencentes a um conjunto finito ou enumerável. Geralmente elas são resultados de **contagens** e, por isso, somente fazem sentido **números inteiros**.

Contínuas: podem assumir qualquer valor num determinado intervalo de variação (reta real). São resultantes de **mensurações e medições** e, por isso, seus valores têm **casas decimais**.

Classificação de variáveis

QUANTITATIVAS → descrevem **quantidades** e têm seus valores expressos **numericamente**

Exemplos:

- **Variáveis Quantitativas Discretas:**

Nº de pessoas na família

Nº de acidentes na BR101 em 2009

Nº de alunos em cada sala de aula

- **Variáveis Quantitativas Contínuas:**

Peso (Kg)

Idade (anos)

Duração do efeito da medicação (horas)

Classificação de variáveis

QUALITATIVAS → descrevem uma **qualidade** ou atributo dos indivíduos da amostra.

Nominais : não apresentam uma ordem natural de ocorrência, ou seja, não existe nenhuma ordenação nos possíveis resultados.

Ordinais: apresentam uma ordem nos seus resultados, uma hierarquia em seus próprios valores, por exemplo:

Classificação de variáveis

QUALITATIVAS → descrevem uma **qualidade** ou atributo dos indivíduos da amostra.

Exemplos:

- **Variáveis Qualitativas Nominais:**

Sexo: Feminino, Masculino

Cor de olhos: Pretos, Castanhos, Azuis, Verdes

Tipo de Veículo: Carro, Moto, Ônibus

- **Variáveis Qualitativas Ordinais:**

Estado de Saúde: Ruim, Regular, Bom

Tipo de Acidente: Leve, Moderado, Grave

Cargo na empresa: diretor, vice-presidente, presidente

VARIÁVEL

- **Para que conhecer o tipo de variável?**

Conhecer a classificação da variável estudada é uma das premissas básicas para a escolha do melhor gráfico para representa-la e do melhor teste estatístico a ser utilizado na análise dos dados.

VARIÁVEL

OBSERVAÇÕES:

- Uma variável originalmente quantitativa pode ser coletada de forma qualitativa.
- Por exemplo, o peso dos lutadores de boxe é uma variável quantitativa (contínua) se trabalhamos com o valor obtido na balança, mas qualitativa (ordinal) se o classificarmos nas categorias do boxe (peso-pena, peso-leve, peso-pesado, etc.).
- Outro ponto importante é que nem sempre uma variável representada por números é quantitativa. O número do telefone de uma pessoa, o número da casa, o número de sua identidade. Às vezes o sexo do indivíduo é registrado na planilha de dados como 1 se macho e 2 se fêmea, por exemplo. Isto não significa que a variável sexo passou a ser quantitativa!

Defina as variáveis abaixo em quantitativa (especifique em contínuas e discretas) ou qualitativas (especifique nominais ou ordinais):

- **Cor da pele das alunas:**
  qualitativa nominal
- **Número de filhos:**
  quantitativa discreta
- **Tempo que os atletas demoram para realizar uma prova de atletismo:**
  quantitativa contínua
- **Número de defeitos em aparelhos de TV:**
  quantitativa discreta
- **Comprimento dos pregos produzidos por uma empresa:**
  quantitativa contínua
- **Estágio da doença dos pacientes:**
  qualitativa ordinal (inicial, intermediário e terminal)
- **Classificação dos funcionários entre fumante e não fumante:**
  qualitativa nominal
- **O ponto obtido em cada jogada de um dado:**
  quantitativa discreta
- **Mês que o aluno faz aniversário:**
  qualitativa ordinal (jan., fev., março, ...)

Tabela

- TABELA: É a disposição de um conjunto de dados/informações em linhas e colunas de maneira sistemática.

- Os elementos fundamentais da tabela são: título, cabeçalho, coluna indicadora e corpo.

➢ Título: Indicação que precede a tabela e que contém a designação do fato observado, o local e a época em foi registrado;

➢ Cabeçalho: Parte superior da tabela que especifica o conteúdo das colunas;

➢ Coluna Indicadora: Geralmente a primeira coluna, que especifica o conteúdo das linhas.

➢ Corpo da Tabela: Conjunto de colunas e linhas que contém as informações sobre a variável em estudo.

Tabela (dados brutos)

Candidato	Tema	Escolaridade	Região	Ocupação	Est. Civil	Renda
Malvina Selva	Geração de empregos	Médio completo	Centro Oeste	Assalariado/autônomo	Casado	38,7
Tomé Terra	Geração de empregos	Médio completo	Centro Oeste	Funcionário público	Casado	5,2
Tomé Terra	Geração de empregos	Médio completo	Centro Oeste	Assalariado/autônomo	Solteiro	6,2
Tomé Terra	Honestidade	Médio completo	Centro Oeste	Assalariado/autônomo	Casado	9,2
Malvina Selva	Geração de empregos	Médio completo	Centro Oeste	Assalariado/autônomo	Divorciado	8,6
Vilma Kasseb	Estabilidade econômica	Fundamental completo	Centro Oeste	Desempregado	Solteiro	3,3

CABEÇALHO

CORPO

COLUNA INDICADORA

Obs: - O lado direito e esquerdo de uma tabela oficial deve ser aberto.
- É conveniente que o número de casas decimais seja padronizado.

Tabela (Distribuição de frequências)

Quantidade de vezes que apareceu aquela resposta

Estado Civil	Frequência absoluta (f_i)	Frequência relativa (fr_i)
Solteiro	23	0,371
Casado	32	0,516
Divorciado	7	0,113
Total	62	1

Quantidade de vezes que apareceu aquela resposta **em relação ao total**.
Se multiplicarmos por 100, obtemos a porcentagem.

- Uma *distribuição de frequências* é uma tabela na qual dispomos as possíveis realizações (resultados) de uma variável e verificamos com qual frequência cada uma destas apareceu no conjunto de dados.

Tabela de Frequência no R

- **Exemplo: Dados “Milsa.txt”**

dados<-read.table("Milsa.txt", header=TRUE)

1. Precisamos informar para o programa que as variáveis **civil, instrução e região** NÃO são numéricas e sim categóricas, usando o comando *factor().*

- Primeiro, redefinimos a variável **civil** com os rótulos (labels) solteiro e casado associados aos níveis (levels) 1 e 2.

dados$civil <- factor(dados$civil, label = c("solteiro", "casado"), levels = 1:2)

- Para a variável **instrução**, usamos o argumento adicional ordered = TRUE para indicar que é uma variável ordinal.

dados$instr <- factor(dados$instr, label = c("1º Grau", "2º Grau", "Superior"), lev = 1:3, ord= T)

Tabela de Frequência no R

- Na variável **região**, codificamos assim: 2=capital, 1=interior, 3=outro.

dados$regiao <- factor(dados$regiao, label = c("capital", "interior", "outro"), lev = c(2, 1, 3))

2. Vamos agora, definir uma nova variável denominada **idade**, em anos, a partir das variáveis ano e mês.

dados$idade <- dados$ano + dados$mes/12

head(dados)

Tabela de Frequência no R

- **Exemplo: Dados “Milsa.txt”**

Primeiro, vamos listar os dados da var. Civil e checar se estão na forma de um fator, que é adequada para variáveis desse tipo.

dados$civil

is.factor(dados$civil)

Frequência absoluta: *table(dados$civil)*

Frequência relativa: *prop.table(table(dados$civil))*

Frequência percentual: *prop.table(table(dados$civil))*100*

Tabela de Frequência no R

- **Exemplo: Dados “Milsa.txt”**

Frequência absoluta: *table(dados$civil)*

Frequência relativa: *prop.table(table(dados$civil))*

Frequência percentual: *prop.table(table(dados$civil))*100*

Para juntar tudo em uma tabela:

*tabela<-matrix(c(table(dados$civil), prop.table(table(dados$civil)), prop.table(table(dados$civil))*100), nrow=2, ncol=3)*

colnames(tabela)<-c("freq.absoluta", "freq.relativa", "freq.percentual")

rownames(tabela)<-c("solteiro", "casado")