André Luiz Morais Lo Feudo
Eliana de Aquino Bonilha
Renan Ishikawa Gasparini

Estatística descritiva aplicada à Epidemiologia.

CENTRO UNIVERSITÁRIO
SÃOCAMILO

Estatística descritiva aplicada à Epidemiologia.

Centro Universitário São Camilo

REITOR
João Batista Gomes de Lima
VICE-REITOR e PRÓ-REITOR ADMINISTRATIVO
Francisco de Lélis Maciel
PRÓ-REITOR ACADÊMICO
Carlos Ferrara Junior

Produção editorial
Coordenadora Editorial
Bruna San Gregório
Analista Editorial
Cintia Machado dos Santos
Assistente Editorial
Bruna Diseró

Organizadora
Tania Leiko Tanaka

Autores
André Luiz Morais Lo Feudo
Eliana de Aquino Bonilha
Renan Ishikawa Gasparini

E82

Estatística descritiva aplicada à epidemiologia / André Luiz Morais Lo Feudo (Org.). - São Paulo: Setor de Publicações - Centro Universitário São Camilo, 2024
79 p.

Vários autores
ISBN 978-65-86702-82-8

1. Estatística descritiva 2. Análise de dados 3. Epidemiologia urbana I. Lo Feudo, André Luiz Morais II. Título

CDD: 614.4072

Ficha Catalográfica elaborada pela Bibliotecária Ana Lucia Pitta
CRB 8/9316

apresentação

Este livro é destinado a estudantes e profissionais da saúde, gestores de saúde, jornalistas e outras áreas. Com uma estrutura objetiva, o e-book aborda de maneira prática os conceitos básicos e aplicações da Estatística Descritiva, apresentando aos leitores diferentes tipos de dados e como podem ser medidos, organizados, interpretados e analisados.
A análise, interpretação e visualização de dados na área da saúde podem ajudar a transformar dados rudimentares em delineamentos numéricos que recontam uma história. Isso pode ajudar a melhorar a precisão e a investigação dos padrões e das tendências de saúde, com diagnósticos mais precisos da situação de saúde de grupos populacionais.
O diferencial do e-book é a inclusão de um apêndice dedicado ao estudo epidemiológico no município de São Paulo, com a aplicação dos conceitos apresentados para compreensão da distribuição e das tendências da mortalidade no município de São Paulo a partir de dados públicos.

Agradecimentos

Agradeço à professora Ma. Eliana A. Bonilha pela grande contribuição no desenvolvimento do conteúdo do apêndice 1 sobre o uso do Tabnet e dos conceitos/cálculos dos coeficientes de mortalidade, e ao monitor de Estatística, Renan Ishikawa Gasparini, pela revisão do conteúdo e elaboração de partes importantes deste e-book.

Conteúdo

1. estatística e seus principais campos de atuação

Estatística é uma área da matemática que envolve a **coleta, organização, análise, interpretação e apresentação de dados.** Ela tem como objetivo fornecer métodos e técnicas para lidar com a incerteza, obter informações significativas a partir dos dados e, assim, tomar decisões.

A estatística pode ser dividida em diferentes ramos ou subcampos, cada um com suas próprias técnicas e aplicações específicas. Os três principais ramos da estatística são:

- **Estatística descritiva:** envolve a organização, o resumo e a apresentação dos dados de forma compreensível, utilizando medidas descritivas, como médias, medianas, desvios-padrão, gráficos, tabelas, entre outros métodos. Este e-book foca nessa parte da estatística.

- **Probabilidade:** é o estudo matemático da incerteza e da aleatoriedade. A probabilidade é usada para modelar eventos aleatórios e calcular a chance de que um determinado evento ocorra. Portanto, é fundamental para a inferência estatística. Ao calcularmos as frequências nas tabelas de frequência, estamos automaticamente calculando probabilidades.

- **Inferência estatística:** refere-se ao processo de tirar conclusões ou fazer inferências sobre uma população com base em uma amostra dos dados. Envolve a estimativa de parâmetros populacionais desconhecidos, teste de hipóteses, construção de intervalos de confiança e análise de regressão, entre outros métodos. Esse ramo da estatística ficará para o próximo e-book.

Além desses ramos principais, a estatística também se relaciona com outras disciplinas, como estatística computacional, análise de dados, estatística experimental, séries temporais, entre outros, que têm suas próprias abordagens e técnicas específicas para lidar com diferentes tipos de dados e situações. E, já que falamos em dados, vamos entender a diferença entre **dados, variáveis** e **elementos** dentro da estatística.

2. elementos, variáveis e dados

Os **elementos** são as unidades individuais ou objetos sobre os quais as observações são feitas. Em termos mais simples, eles são as **entidades que estão sendo estudadas**. Os elementos podem ser pessoas, animais, plantas, produtos, empresas ou qualquer outra unidade de interesse em um estudo estatístico. Por exemplo, se estivermos estudando a **altura** de estudantes em uma sala de aula, **cada estudante individual seria um elemento.**

As **variáveis** são características ou propriedades que estão sendo medidas ou observadas em relação a cada elemento da amostra. Elas representam as diferentes quantidades, atributos ou características dos elementos. As variáveis podem ser quantitativas ou qualitativas.

- Variáveis qualitativas: são aquelas que representam categorias ou atributos e que não possuem uma ordenação ou quantificação numérica. Exemplos incluem gênero, cor dos olhos, estado civil, tipo sanguíneo etc.
- Variáveis quantitativas: são aquelas que têm valores numéricos e expressam uma quantidade ou uma medida. Exemplos incluem **altura**, peso, idade, renda, temperatura etc.

Já os **dados** referem-se às observações ou medidas obtidas para as variáveis em relação aos elementos. Eles são informações coletadas sobre os elementos de interesse em um estudo estatístico. Os dados podem ser obtidos através de pesquisas, experimentos, registros, questionários, medições, entre outros métodos. Eles podem ser representados em forma de números, palavras, códigos ou símbolos, dependendo do tipo de variável e do método de coleta.

Exemplo 1: Siga as instruções para as duas tabelas fornecidas abaixo.

Observe a Tabela 1 e identifique qual é o elemento, qual é a variável de estudo e quais são os dados.

Tabela 1 – Peso de nascidos vivos no Brasil

Peso ao nascer	**2016**	**2017**	**2018**	**2019**	**2020**
Menos de 500g	4015	4065	3926	4127	3928
500 a 999g	13796	14541	14320	14217	13618
1000 a 1499 g	21462	22460	22259	22109	20914
1500 a 2499 g	203205	207288	209913	207403	195865
2500 a 2999 g	651109	655175	659310	645303	604473
3000 a 3999 g	1818250	1865156	1879622	1810092	1743061
4000g e mais	144940	153857	154637	145149	147649
Ignorado	1023	993	945	746	637
Total	**2857800**	**2923535**	**2944932**	**2849146**	**2730145**

Fonte: MS/SVS/DASIS - Sistema de Informações sobre Nascidos Vivos – SINASC

Observe a Tabela 2 e identifique qual é o elemento, qual é a variável de estudo e quais são os dados.

Tabela 2 – Informações de alunos (fictícios).

Aluno	Número de *pets*	Idade	Nível escolar	Estado civil
Joca	2	30	EM	casada
Juca	2	19	EM	solteira
Elenna	2	22	EM	solteira
Beto	1	19	EM	solteira
Beta	1	32	ES	solteira
Anna	1	20	EM	solteira
Rachel	2	19	EM	solteira
Harvey	1	24	EM	solteira
Loui	1	29	ES	solteiro

Fonte: criada pelos autores.

Confira abaixo a resposta correta para as tabelas 1 e 2:

Respostas da Tabela 1:

Os elementos são todos os recém-nascidos vivos entre 2016 e 2020 no Brasil, a variável de estudos é o peso ao nascer em "g" (que é quantitativa) e os dados são todos os pesos medidos de cada elemento.

Respostas da Tabela 2:

Os elementos são todos os 13 alunos. Aqui, na verdade, temos várias variáveis: o número de *pets* dos alunos, as idades dos alunos, o nível escolar deles e o estado civil de cada um (essas duas últimas são do tipo qualitativa). Os dados são todas as medidas e valores de cada aluno relativo a cada variável de estudo.

3. classificação e subclassificação das variáveis

Já vimos que as variáveis podem ser classificadas em dois principais tipos: variáveis **quantitativas** e variáveis **qualitativas**. Vamos explorar agora as subclassificações de cada uma delas.

Variáveis quantitativas: são aquelas que têm valores numéricos e expressam uma quantidade ou uma medida. Elas podem ser subdivididas em dois subtipos:

a) Variáveis quantitativas contínuas: são variáveis que podem assumir qualquer valor dentro de um intervalo. Elas são representadas pelos números reais. Exemplos comuns incluem altura, peso, temperatura, renda, tempo de reação, entre outros. Essas variáveis podem ser medidas com precisão e podem ter uma infinidade de valores possíveis entre quaisquer dois valores inteiros.

b) Variáveis quantitativas discretas: são variáveis que assumem apenas valores **inteiros** e, por isso, são representadas por números inteiros ou contáveis. Exemplos incluem o número de filhos, o número de erros em um teste, o número de visitantes em um evento, entre outros.

Variáveis qualitativas: as variáveis qualitativas, também conhecidas como variáveis categóricas, são aquelas que representam categorias ou atributos e não possuem um valor numérico. Elas são subdivididas em dois subtipos:

a) Variáveis qualitativas nominais: são variáveis em que as categorias não possuem uma ordem natural. Cada categoria é única e exclusiva, sem relação de precedência entre elas. Exemplos incluem gênero (masculino/feminino), cor dos olhos (azul/verde/castanho), nacionalidade (brasileiro/americano/francês), entre outros.

b) Variáveis qualitativas ordinais: são variáveis em que as categorias possuem uma ordem ou hierarquia natural. Embora as categorias ainda sejam distintas, existe uma relação de ordem entre elas. Exemplos incluem classificação em uma escala de satisfação (muito insatisfeito/insatisfeito/neutro/satisfeito/muito satisfeito), níveis de concordância (discordo totalmente/discordo parcialmente/neutro/concordo parcialmente/concordo totalmente), entre outros.

A distinção entre variáveis quantitativas e qualitativas e seus subtipos é importante na análise estatística, pois os métodos e técnicas estatísticas apropriados variam de acordo com o tipo de variável. Por exemplo: os **dados** de variáveis **quantitativas discretas** podem ser organizados em quase todo tipo de gráfico (linha, ou coluna, ou setor, ou barras etc.), porém os **dados** das **quantitativas contínuas** só podem ser apresentados em dois tipos de gráficos que são o histograma e os polígonos de frequências. Vamos treinar?

Exercício 1: Classifique cada variável identificada nas tabelas 1 e 2 do exemplo anterior.

A. Peso ao nascer em "g"

B. Número de *pets* dos alunos

C. Idades dos alunos

D. Nível escolar

E. Estado civil

4. população e amostra

Além dos tipos de variáveis e suas subclassificações, outro assunto importante de ser compreendido é o conceito de população e amostra. Na estatística, "**população**" refere-se ao conjunto completo de elementos que compartilham uma característica comum e são de interesse para um estudo. Esses elementos podem ser indivíduos, objetos, eventos, medidas ou qualquer outra unidade que seja relevante para a pesquisa em questão. Em outras palavras, a população é o conjunto total de todos os elementos que se deseja estudar e sobre os quais se deseja organizar, medir e fazer inferências.

Por outro lado, uma "**amostra**" é um subconjunto selecionado da população. É um grupo menor de elementos retirados da população com o objetivo de estudá-los e fazer **inferências** sobre a população. A amostra quando representa bem a população pode ser usada para obter informações sobre os parâmetros da população, como médias, proporções, desvios-padrão, entre outros.

A principal razão para usar uma amostra em vez da população inteira é que, na maioria dos casos, é mais prático e viável coletar dados de uma amostra menor em termos de tempo, custo e recursos disponíveis. Se a amostra for **bem escolhida e representativa** da população, as conclusões obtidas a partir da análise da amostra podem ser generalizadas para a população como um todo.

É importante ressaltar que a seleção adequada da amostra e a aplicação de métodos estatísticos apropriados são essenciais para garantir que as inferências feitas a partir de uma amostra sejam válidas e precisas para a população de interesse.

Vejam, a seguir, alguns exemplos nos quais pedimos para definir quem é a população e quem é a amostra a partir de um contexto dado.

Exemplo 2: Leia as três situações abaixo e defina qual é a população e qual é a amostra. Depois de responder, compare sua resposta com o gabarito.

Situação 1: Para estimar o % de nascidos abaixo do peso desde 2010 até hoje no Brasil, foram levantados os pesos de todos os nascidos vivos entre 2012 e 2016. Qual é a população e qual é a amostra?

Situação 2: Com o objetivo de estimar o % de nascidos vivos acima do peso no estado de São Paulo, foram selecionadas algumas cidades para levantar os pesos de todos os nascidos nessas cidades. Qual é a população e qual é a amostra?

Situação 3: Calculou-se a porcentagem entre os partos por cesárea e por via vaginal dos nascidos vivos na região Sudeste do Brasil para fazer uma estimativa a nível nacional. Qual é a população e qual é a amostra?

Confira abaixo a resposta correta dos três exemplos dados:

Situação 1: A população é composta com os pesos de todos os nascidos vivos no Brasil desde 2010 até hoje, enquanto a amostra é composta pelos pesos levantados dos nascidos vivos entre 2012 e 2016.

Situação 2: A população é composta com os pesos de todos os nascidos vivos no Brasil. A amostra é composta pelos pesos dos nascidos vivos nas cidades selecionadas.

Situação 3: A população é composta com todos os nascidos vivos por cesárea e parto vaginal no Brasil. A amostra é composta com os nascidos vivos por cesárea e por parto vaginal na região Sudeste.

5. estatística descritiva

Já dissemos que a estatística descritiva envolve a organização, o resumo e a apresentação dos dados de forma compreensível, utilizando vários tipos de **medidas**, **gráficos** e **tabelas**. Vamos ver, agora, por meio de exemplos, cada um desses tipos e como a classificação das variáveis do item 3 influenciam na sua escolha.

5.1. Tabelas de frequência e gráficos

Antes de entrarmos nos tipos de tabelas de frequência (TDF) e gráficos, precisamos entender os conceitos de dados agrupados e não agrupados.

Dados **não agrupados** referem-se a um conjunto de observações individuais sem categorização ou agrupamento. Cada observação é considerada como um valor único e independente. Por exemplo, uma lista de idades de um grupo de pessoas, como 25, 32, 38, 41 e 27, seria um conjunto de dados não agrupados.

Por outro lado, **dados agrupados** envolvem a organização das observações em categorias ou intervalos. As observações são agrupadas em intervalos específicos, e a frequência de ocorrência de cada intervalo é registrada. Por exemplo, ao agrupar as idades em intervalos de 10 anos (20-29, 30-39, 40-49 etc.), contando o número de pessoas em cada intervalo, teríamos um conjunto de dados agrupados.

O agrupamento de dados pode ser útil quando temos um grande conjunto de dados e queremos resumir as informações de maneira mais compacta. Ele permite visualizar a distribuição dos dados em intervalos e fornecer uma visão geral dos padrões e tendências.

Uma importante ferramenta para AGRUPAR dados são as **TABELAS**. Uma tabela consiste em colunas e linhas (corpo da tabela), título e fontes ou notas. Tabelas são amplamente utilizadas na estatística, na pesquisa, na ciência e em muitos outros campos para organizar, resumir e apresentar informações de forma clara e concisa. Elas fornecem uma estrutura visualmente organizada que permite comparar e analisar os dados de maneira sistemática. **Planilhas**, como o Excel da Microsoft ou as planilhas Google, são tabelas eletrônicas muito poderosas para agrupar dados em tabelas e gráficos.

O quadro 1 relaciona quais as tabelas e gráficos **mais adequados** em função da classificação das variáveis. Notem que o tipo de TDF determina quais as possibilidades de gráficos que podem ser utilizados para representar seus dados.

Quadro 1 - Relação entre variáveis, tabelas e gráficos.

Classificação das variáveis	Tabelas de frequência (TDF)	Gráficos
Nominais, ordinais ou Discretas	dados agrupados SEM CLASSES	Linha
		Colunas Verticais ou Horizontais
		Setores (ou pizza)
Quantitativas Discretas	dados agrupados SEM CLASSES	BoxPlot
Quantitativas Contínuas	dados agrupados COM CLASSES	Histograma
		Polígono de frequências

Fonte: Elaborado pelos autores

Mas o que são DADOS agrupados SEM CLASSE e COM CLASSE? CLASSES são FAIXAS ou INTERVALOS de valores. Agrupar dados COM CLASSES significa agrupar os dados dentro de faixas de valores.

ESTATURAS DE 40 ALUNOS DA FACULDADE A - 2007	
ESTATURAS (cm)	FREQUÊNCIA
150 ⊢ 154	4
154 ⊢ 158	9
158 ⊢ 162	11
162 ⊢ 166	8
166 ⊢ 170	5
170 ⊢ 174	3
total	40
Dados fictícios.	

Tabela 3 – Alturas de 40 alunos.

Notem na tabela 3 que cada CLASSE tem: o tamanho de 4 cm, um limite inferior e um limite superior, e temos nessa tabela seis classes. É importante entender que não há uma fórmula única para o cálculo do nº de classes (k). As mais usadas são:

1ª) $k=\sqrt{n}$, aproximando o resultado para o menor ou maior valor. Sendo "n" o total de dados. Na tabela 3, temos as alturas de 40 alunos, logo n = 40 e $k=\sqrt{n}=\sqrt{40}=6{,}32$ que aproximamos para **6 classes**.

2ª) Fórmula de Sturges: $K \cong 1 + 3{,}22 \times \log n$

Para as 40 alturas da tabela 3, teríamos: k≅ 1 + 3,22 x log(40).
Fonte: elaborado pelos autores.

$$k \cong 1 + 3{,}22 \times 1{,}60206 \qquad k \cong 1 + 5{,}1586 \rightarrow \qquad k \cong 6{,}1586 \rightarrow$$

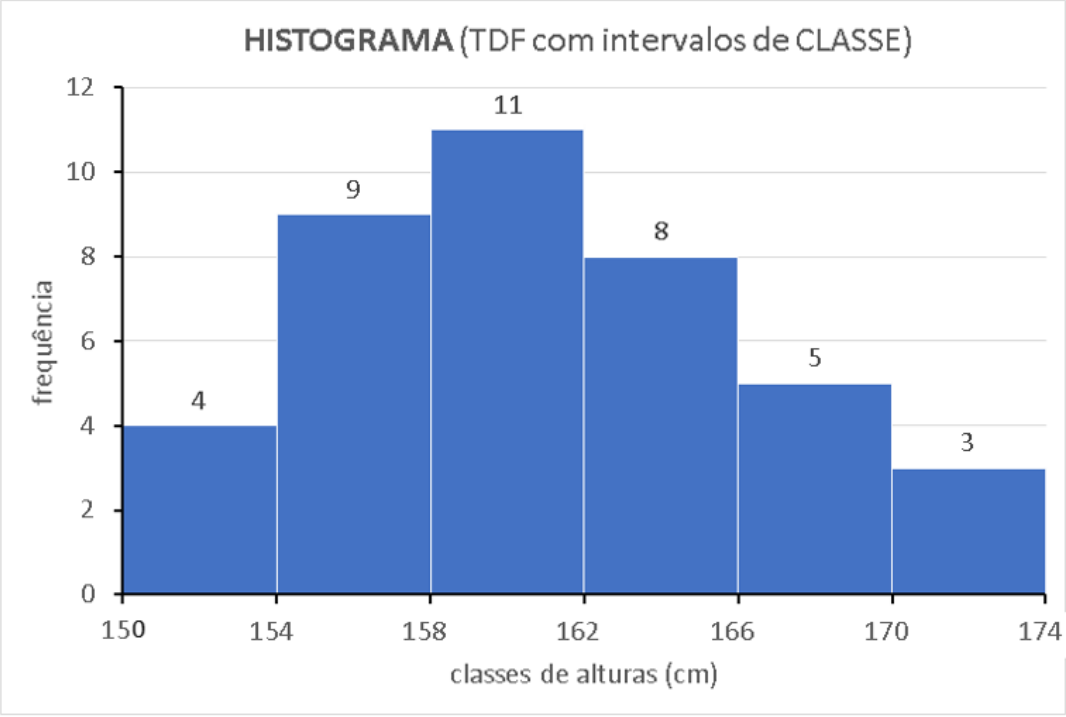
E chegamos nas mesmas **6 classes**.

Definido o " k" (nº de classes), calcula-se, em seguida, o tamanho de cada classe (h) com a seguinte fórmula:

$$h = \frac{(\text{maior valor} - \text{menor valor})}{k}$$

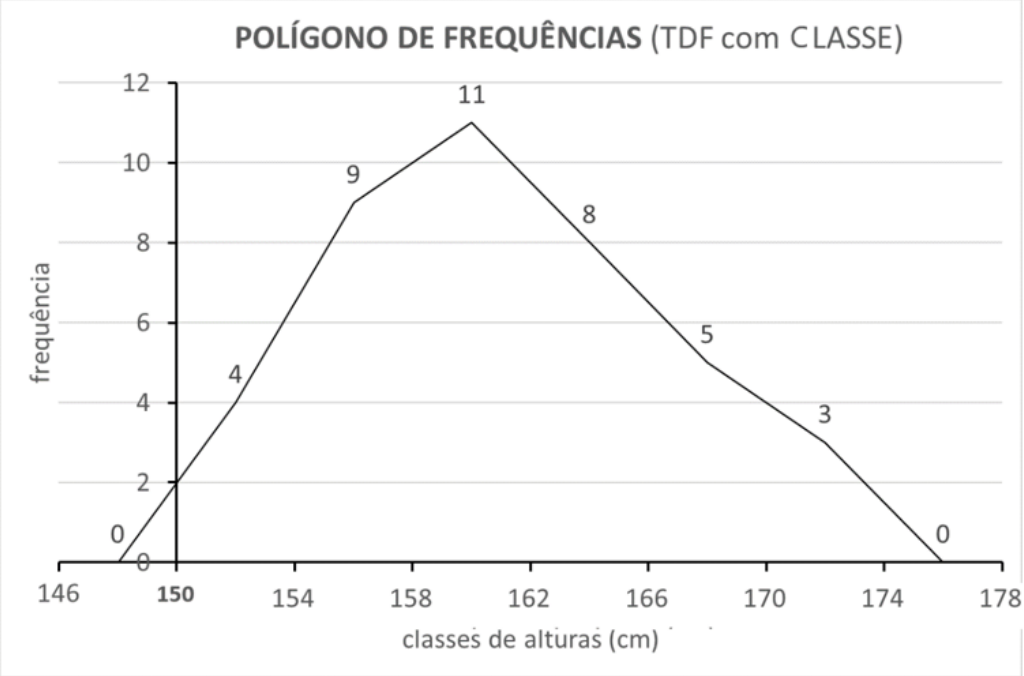
As **TDF com (intervalos de) CLASSE** podem ser representadas por dois tipos de gráficos: **histogramas** e **polígonos de frequência**. Vejam como ficam esses gráficos usando os dados da TDF com classes como ESTATURAS DE 40 ALUNOS DA FACULDADE A – 2007.

Gráfico 1 – Histograma (TDF com intervalos de CLASSE).



Fonte: Elaborado pelos autores

Gráfico 2 – Polígono de Frequências (TDF com CLASSE).



Fonte: Elaborado pelos autores

Percebam que a escala do eixo "x" (horizontal) é composta pelos próprios intervalos de classe da tabela 3, e que, no polígono de frequências, o gráfico inicia num limite inferior (146) anterior ao menor limite da tabela 3 (que é 150) e ***fecha*** num limite superior acima (178) do maior limite da tabela 3 (que é o limite 174).

AGRUPAR DADOS SEM CLASSE significa agrupar os dados **SEM** intervalos de valores. Logo, nas tabelas de dados agrupados SEM classe, **não se usa** o histograma, nem o polígono de frequências.

Vejam as alturas (fictícias) que deram origem a **TDF com CLASSE** anterior (tabela 3) desagrupadas, mas já em ordem crescente:

150	151	152	153	154	154	155	155	156	156	157	157
157	158	158	158	159	159	160	160	161	161	161	161
162	162	162	163	163	164	165	165	167	168	168	169
169	170	173	173								

Tabela 4 - Alturas de 40 alunos.

Alturas de 40 alunos	
Alturas	Frequência
150	1
151	1
152	1
153	1
154	2
155	2
156	2
157	3
158	3
159	2
160	2
161	4
162	3
163	2
164	1
165	2
167	1
168	2
169	2
170	1
173	2
total	40

Dados fictícios.

Vamos, agora, agrupar essas alturas SEM INTERVALOS DE CLASSES (ou simplesmente SEM CLASSES).

A desvantagem de agrupar dados de variável contínua em **TDF SEM CLASSES** (ver tabela 4), é que a tabela geralmente fica muito grande, com um nº exagerado de linhas, principalmente quando não há muita repetição dos valores (as frequências dos valores são, na maioria das vezes, baixas). Por essa razão, no quadro 1, orientamos a agrupar dados de variável contínua em TDF COM CLASSES.

Os gráficos 3 e 4 de colunas (horizontais ou verticais) são os mais comuns para representar as frequências de uma TDF SEM CLASSES.

Gráfico 3 – Colunas verticais.

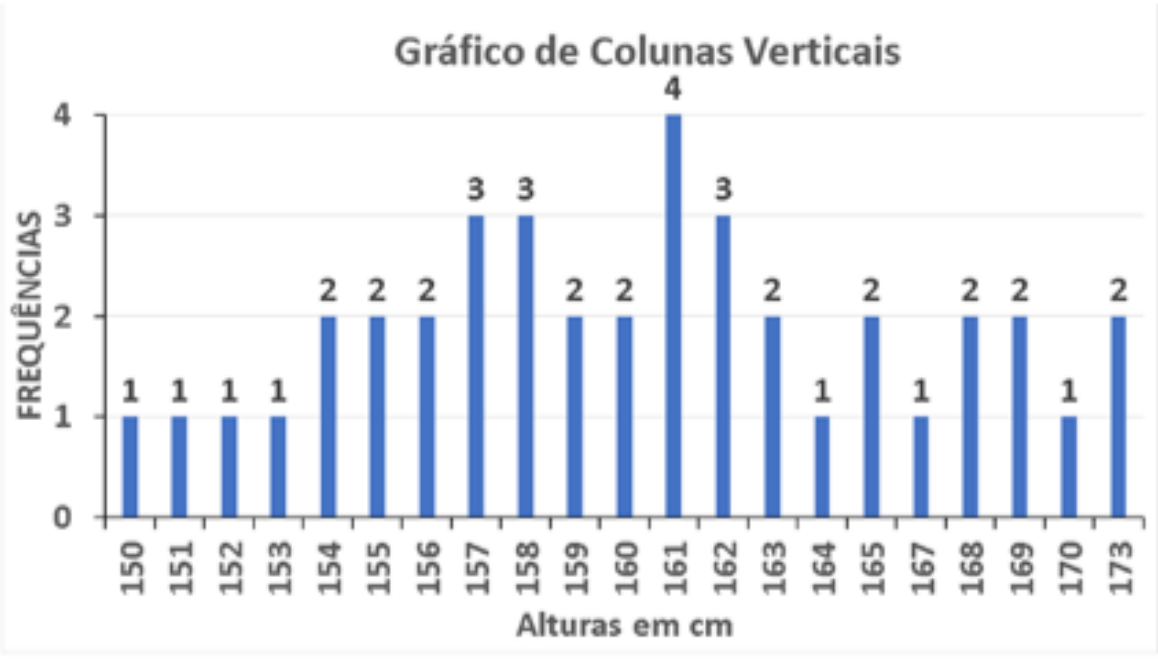
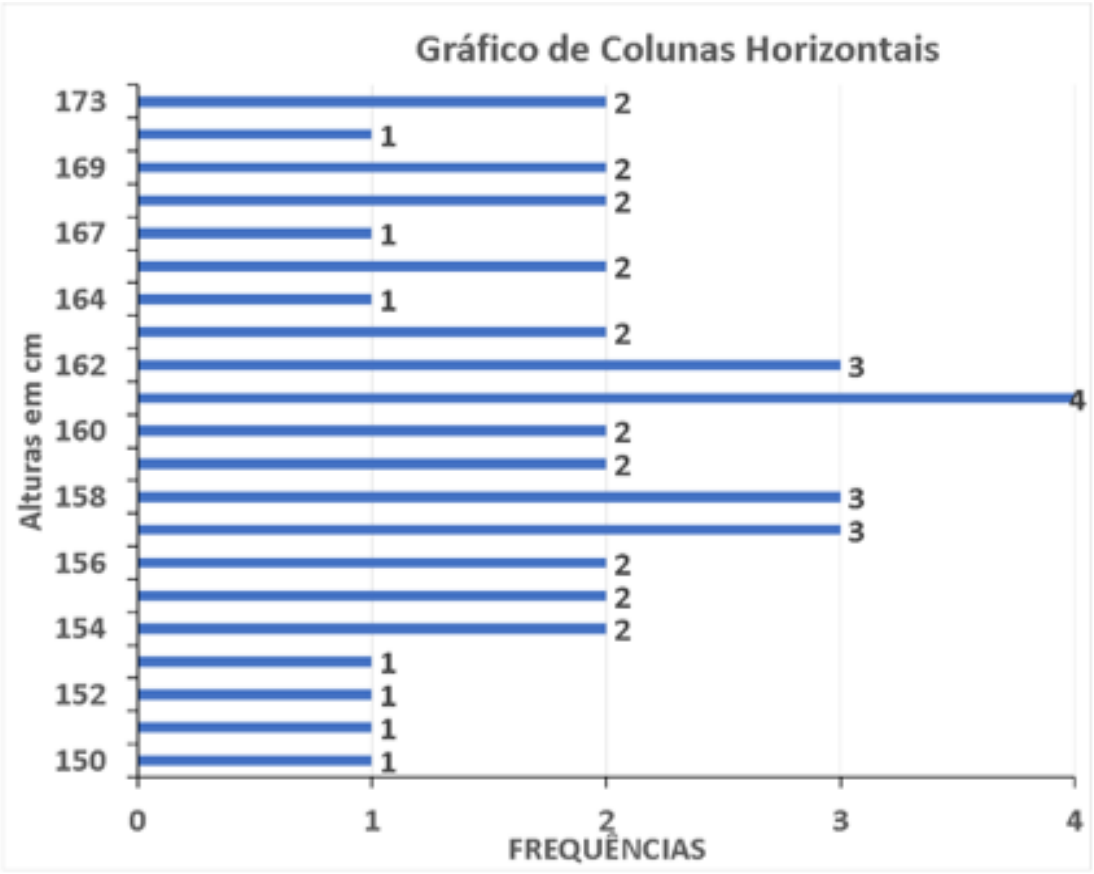


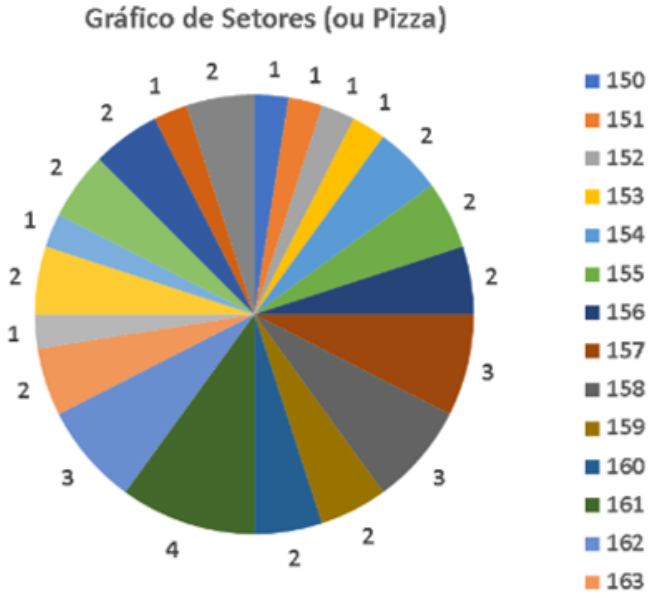
Gráfico 4 – Colunas Horizontais.



Fonte: Elaborado pelos autores

Gráfico 5 – Setores.

O gráfico 5 de setores, ou de pizza, também pode ser utilizado, mas sua visualização é prejudicada quando há mais de 6 ou 7 valores diferentes, como no caso deste exemplo.



Fonte: Gerado pelos autores usando o Excel

As planilhas eletrônicas possuem muitas opções de gráficos. Basta clicar em INSERIR e depois selecionar entre os tipos disponíveis para escolher o que melhor se adéqua aos seus objetivos.

Figura 1 - Opções de gráficos em planilhas.

Fonte: print da tela do Excel da Microsoft.

Para finalizar o tópico 5.1 sobre TDF e gráficos, precisamos revelar que não existe apenas um tipo de frequência. Nas duas TDF das alturas (com e sem classes), que acabamos de mostrar (tabelas 3 e 4), só calculamos a frequência SIMPLES. Além dela, ainda podemos calcular as:

Frequências acumuladas crescentes (fa ou fac), que são o acúmulo das frequências simples (f).

Frequências relativas simples (fr), que são cada f dividida pelo total de valores.

Frequências relativas acumuladas (fra), que são o acúmulo das fr.

Vejam, abaixo, como as TDF que já vimos ficam com todas as frequências calculadas.

Tabela 5 - Frequências adicionadas à tabela 3 (fa, fr e fra).

TDF COM DADOS AGRUPADOS COM CLASSES

ESTATURAS	f frequência simples	fa f acumulada	fr frequência relativa	fra fr acumulada
150 I— 154	4	4	10%	10%
154 I— 158	9	13	23%	33%
158 I— 162	11	24	28%	60%
162 I— 166	8	32	20%	80%
166 I— 170	5	37	13%	93%
170 I— 174	3	40	8%	100%
total	40		100%	

Fonte: gerada pelos autores.

Tabela 6 - Frequências adicionadas à tabela 4 (fa, fr e fra).

TDF COM DADOS AGRUPADOS SEM CLASSES

Alturas	f frequência simples	fa f acumulada	fr frequência relativa	fra fr acumulada
150	1	1	2,5%	3%
151	1	2	2,5%	5%
152	1	3	2,5%	8%
153	1	4	2,5%	10%
154	2	6	5,0%	15%
155	2	8	5,0%	20%
156	2	10	5,0%	25%
157	3	13	7,5%	33%
158	3	16	7,5%	40%
159	2	18	5,0%	45%
160	2	20	5,0%	50%
161	4	24	10,0%	60%
162	3	27	7,5%	68%
163	2	29	5,0%	73%
164	1	30	2,5%	75%
165	2	32	5,0%	80%
167	1	33	2,5%	83%
168	2	35	5,0%	88%
169	2	37	5,0%	93%
170	1	38	2,5%	95%
173	2	40	5,0%	100%
totais	40		100,0%	

Fonte: gerada pelos autores.

Exercício 2: Vinte pacientes receberam uma nova medicação e tiveram seu tempo de reação medidos. Os dados foram os seguintes (em min.): 2,9; 3,4; 3,5; 4,1; 4,6; 4,7; 4,5; 3,8; 5,3; 4,9; 4,8; 5,7; 5,8; 5,0; 3,4; 5,9; 6,3; 4,6; 5,5 e 6,2. Com base nestas informações, organize esses dados na TDF e em gráficos mais adequados.

Exercício 3: As idades de 20 pacientes com cisto no pâncreas seguem abaixo (já ordenadas).

21, 21, 21, 21, 21, 21, 21, 21, 22, 22, 22, 22, 23, 23, 24, 24, 25, 26, 27 e 28

Organize os dados em uma distribuição de frequências adequada.

5.2. Séries estatísticas

As TDF e gráficos que acabamos de estudar no item 5.1 organizam e representam dados de apenas uma variável em estudo. No entanto, em muitas situações, estamos interessados em organizar dados de mais de uma variável. Nessas situações, entra o conceito de séries estatísticas.

Segundo Crespo (2009), "denominamos série estatística toda tabela que apresenta a distribuição de um conjunto de dados estatísticos em função da época, do local ou da espécie".

Crespo (2009) continua explicando que, a partir dessa definição, podemos ter um dos três elementos em uma série: tempo para **séries temporais**, espaço para **séries geográficas**, ou espécie para **séries específicas**. Por fim, também temos as **tabelas de dupla entrada** (ou conjugadas), utilizadas quando queremos mostrar em uma única tabela a variação de **duas ou mais variáveis**, como tempo com região ou tempo com uma espécie.

Vejam, a seguir, exemplos de cada uma dessas tabelas elaboradas pelos autores:

Tabela 7 – Exemplo de tabela de série específica.

EXEMPLO DE SÉRIE ESPECÍFICA

Óbitos residentes no MSP em 2022 por tipos de câncer

Causas específicas	Óbitos
CA Pâncreas	1058
CA Pulmão	1903
CA Mama	1365
CA Ovário	338

Fonte: SIM/PRO-AIM – CEInfo –SMS-SP

Tabela 8 – Exemplo de tabela de série temporal.

EXEMPLO DE SÉRIE TEMPORAL

Óbitos por Insuficiência Cardíaca por ANO

ANO	Óbitos Residentes no MSP
2017	1296
2018	1426
2019	1351

Fonte: elaborado pelos autores, com base nos dados do SIM/PRO-AIM – CEinfo – SMS – SP.

Tabela 9 – Exemplo de tabela de série.

EXEMPLO DE SÉRIE GEOGRÁFICA

Óbitos por Insuficiência Cardíaca por REGIÃO de 2017 a 2019

Coord Regional de saúde	Óbitos Residentes no MSP
Centro	165
Leste	892
Oeste	379
Norte	903
Sul	553
Sudeste	1181

Fonte: elaborado pelos autores, com base nos dados do SIM/PRO-AIM – CEinfo – SMS – SP.

Tabela 10 – Exemplo de tabela de dupla entrada (Tempo x Geográfica).

EXEMPLO DE SÉRIE DE DUPLA ENTRADA (Tempo x Geográfica)

Causa dos óbitos por Insuficiência Cardíaca

	Coord REGIONAL de Saúde resid						
ANO	Centro	Leste	Oeste	Norte	Sul	Sudeste	Total
2017	51	251	117	334	168	375	1296
2018	62	341	125	298	206	394	1426
2019	52	300	137	271	179	412	1351

Fonte: elaborado pelos autores, com base nos dados do SIM/PRO-AIM – CEinfo – SMS – SP.

5.3. Medidas de tendência central e de dispersão

Vimos até aqui que podemos organizar e resumir os dados usando **tabelas** (TDFs sem ou com classe) ou **gráficos**. Mas essas não são as duas únicas possibilidades.

Uma outra forma de representar um conjunto de dados é por meio de uma única **medida de tendência central** acompanhada, sempre que possível, de uma **medida de dispersão**, que tem a função de representar quanto os valores estão distantes entre si.

Exemplo 3: Como representar, com uma única medida, o cargo de 50 funcionários de uma empresa, sabendo que 40 são analistas, 5 são chefes, 3 são gerentes e 2 são diretores? E qual é o seu valor?

Como CARGO é uma variável QUALITATIVA (ordinal), uma opção é usar a medida de tendência central denominada **MODA** (ou M_o).

A MODA é a medida de tendência central que mais se repete (maior frequência), o que, nesse caso, é o cargo "ANALISTA", pois é o que mais aparece (f = 40).

Resposta deste exemplo: A medida é a MODA ou M_o = ANALISTA.

O quadro 2 elaborado pelos autores relaciona qual a medida de tendência central **mais adequada** em função da **classificação das variáveis** e qual sua respectiva **medida de dispersão**.

Quadro 2 - Relação entre variáveis com as medidas de posição e dispersão.

Classificação das variáveis	Medida de Tendência Central (ou de Posição)	Medidas de Dispersão
Quantitativas (discretas ou contínuas)	**Média (simples ou Ponderada):** *soma de todos os valores dividida pela qtde de valores*	Amplitude
		Variância
		Desvio Padrão
		Coeficiente de variação
		z-score
Quantitativas ou Qualitativas ORDINAIS	**Mediana:** *valor que fica exatamente no meio de todos*	Intervalo interquartil
Qualitativas ou Quantitativas	**Moda:** *o valor de maior frequência*	Não possui

Fonte: Elaborado pelos autores

Notem que a **média** possui quatro possibilidades de medidas de dispersão para indicar o quanto os dados estão longe dela, a mediana só tem uma **medida** de dispersão, e a **moda** nenhuma.

Interpretando o quadro 2:

> Se você quer uma medida de tendência central de rápida e fácil obtenção, a MODA é a melhor opção, pois basta identificar qual o valor que mais se repete para ter seu resultado. Porém, a MODA não possui o apoio de uma dispersão dos dados.
>
> Precisa-se de uma medida de tendência central ROBUSTA, que não é afetada por valores muito extremos que costumam distorcer o seu resultado e que ainda tem o intervalo interquartil para representar a dispersão dos dados? A mediana é a melhor opção, porém essa medida não se aplica a dados qualitativos NOMINAIS, já que a premissa para uso da MEDIANA é colocar os dados ou valores em ordem crescente ou decrescente de importância.
>
> Seu conjunto de dados quantitativos (discretos ou contínuos) tem uma distribuição de frequência próxima da **simetria**[1] (veja a forma do gráfico de colunas) com o valor da M_o bem próxima do valor da média? Use a média com suas várias opções para medir a dispersão de dados, sabendo que ela não serve para dados qualitativos.

Na prática, as medidas mais utilizadas são a média e a mediana. Mas, dependendo da situação, é bom calcular as três medidas para dar aos leitores a opção de escolher qual melhor irá lhe atender.

Os valores da "tabela de dados de 22 pacientes com cisto no pâncreas" serão utilizados para definir cada medida de tendência central e dispersão.

A tabela está ordenada apenas em função dos elementos da pesquisa, que são os 22 pacientes. Em relação a cada paciente temos dados relativos a cinco variáveis diferentes (sexo, idade, nível escolar, tamanho do cisto e localização). Utilizando a classificação das variáveis (quadro 1) em conjunto com o quadro 2, vamos aplicar as medidas centrais e de dispersão.

[1]Veremos o conceito de simetria ao estudar a média.

Tabela 11 - 22 pacientes com cisto no pâncreas

paciente	sexo	idade	Nível escolar	tamanho do cisto (cm)	Localização do cisto no pâncreas
1	F	49	Fund	6	cabeça
2	F	61	Médio	10	cabeça
3	M	34	Superior	8,2	cauda
4	F	73	Médio	3	colo
5	M	47	Superior	3,6	cabeça
6	M	58	Médio	10	colo
7	M	43	Superior	1	cabeça
8	M	71	Fund	1	cabeça
9	M	32	Superior	7	cauda
10	M	56	Superior	1	cabeça
11	M	61	Médio	6,6	corpo
12	F	49	Superior	4	cabeça
13	M	80	Médio	3,1	cauda
14	M	72	Médio	2,3	cabeça
15	M	47	Superior	10,5	cabeça
16	F	48	Superior	6,5	corpo
17	F	37	Superior	13	corpo
18	M	71	Médio	1	colo
19	M	74	Fund	7	cabeça
20	F	21	Médio	12	corpo
21	F	45	Médio	8,5	corpo
22	M	38	Superior	10	colo

Fonte: https://docs.ufpr.br/~jomarc/exerciciosestatistica2.pdf

5.3.1. Moda (M_o)

A medida de tendência central (MTC), **moda**, é usada na estatística para descrever o valor ou os valores mais frequentes em um conjunto de dados. Em outras palavras, a moda representa o ponto de maior densidade de ocorrência dos dados.

Conforme o quadro 2, a $\mathbf{M_o}$ pode ser usada para representar todo tipo de variável, ou seja, serve para dados QUALITATIVOS e QUANTITATIVOS. Porém, como não há uma medida de dispersão para acompanhar a MODA, ela acaba sendo mais útil quando se necessita de uma medida fácil e rápida de medir, já que seu cálculo é muito fácil (basta achar o valor que mais se repete).

Por exemplo, se estivermos analisando a cor favorita de um grupo de pessoas, a moda nos dirá qual cor é a mais comum no conjunto de dados. Vamos dar outro exemplo usando a variável qualitativa SEXO dos 22 pacientes com cisto no pâncreas.

Isolando os dados da variável SEXO...

Sexo	F	F	M	F	M	M	M	M	M	M	M	F	M	M	M	F	F	M	M	F	F	M

... e colocando em ordem alfabética...

Sexo	F	F	F	F	F	F	F	F	M	M	M	M	M	M	M	M	M	M	M	M	M	M

... fica fácil visualizar que o valor mais frequente é o sexo MASCULINO, portanto $\mathbf{M_o = M.}$

Isolando, agora, os dados da variável **NÍVEL ESCOLAR**... (F = fundamental, M = médio e S = superior)

Nível	F	M	S	M	S	M	S	F	S	S	M	S	M	M	S	S	S	M	F	M	M	S

... e colocando em ordem alfabética...

Nível	F	F	F	M	M	M	M	M	M	M	M	M	S	S	S	S	S	S	S	S	S	S

... não é fácil ver qual nível escolar é mais frequente. Nesse caso, o jeito é contarmos a frequência de cada um. "M" aparece 9 vezes, "S" aparece 10 vezes e "F" apenas 3 vezes. Portanto, a moda dessa variável é o Nível Superior. $\mathbf{M_o = Superior}$.

Uma forma de contar valores para descobrir sua frequência (nº de repetições) é usar a função **=CONT.SE(intervalo; critério)** do Excel.

É importante observar que, em algumas distribuições, pode haver mais de uma moda, o que significa que há múltiplos valores com a mesma frequência máxima.

A melhor maneira de calcular a(s) MODA(s) de uma grande quantidade de dados no **Excel** é usando a fórmula = MODO.MULT(**intervalo**) como matriz. Para isso, você deve selecionar com o *mouse* umas três ou quatro células ANTES de colocar a fórmula, depois clique no *prompt* para digitar a fórmula e, ao final, após fechar os parênteses, clique em CTRL+SHIFT+ENTER (não dê somente **ENTER** para inserir a fórmula).

Assista ao vídeo no link abaixo para ver os exemplos de uso dessa função do Excel.

youtube.com/watch?v=bh-yGUakKlI

Obs.: também poderíamos "calcular" o valor modal para as variáveis QUANTITATIVAS DISCRETAS, desde que houvesse números que se repetem, porém não teríamos a informação de quão dispersos os valores estão.

5.3.2. Mediana (M_d) e intervalo interquartil (IQ)

A medida de tendência central mais robusta é a mediana. Ela é usada na estatística para descrever o valor que se encontra exatamente na posição central de um conjunto de valores, estando estes ordenados segundo uma ordem de grandeza (crescente ou decrescente).

Conforme o quadro 2, a **mediana** é útil quando se lida com dados quantitativos ou qualitativos ordinais. Por exemplo, se estivermos analisando as idades dos 22 pacientes com cisto no pâncreas, a mediana será a idade da posição 11,5º (entre a 11º e 12º), já que ela divide as 22 idades em dois pedaços iguais. Vejamos o passo a passo de como achar a mediana a partir dos dados abaixo:

Idade	49	61	34	73	47	58	43	71	32	56	61	49	80	72	47	48	37	71	74	21	45	38

De acordo com a definição de mediana, o primeiro passo a ser dado é o da ordenação (crescente ou decrescente) dos valores:

Idade	21	32	34	37	38	43	45	47	47	48	49	49	56	58	61	61	71	71	72	73	74	80

A mediana é aquele valor central que deixa o mesmo número de elementos à direita e à esquerda. Se a série dada tiver um número par de termos, a mediana será o ponto médio dos dois valores centrais da série. Como no exemplo das 22 idades, em que temos 11 idades de cada lado...

Idade	21	32	34	37	38	43	45	47	47	48	**49**	**49**	56	58	61	61	71	71	72	73	74	80

... a mediana será a média aritmética entre os dois valores mais próximos do meio, que são "49" e "49". (49+49)/2 = 49 Mediana igual a "49"

O valor da mediana pode coincidir ou não com um elemento da série. Quando o número de elementos da série é ímpar, há coincidência, porém, quando esse número é par, pode ou não coincidir.

O **intervalo interquartil** (IQ) é uma medida de dispersão usada na estatística para análise de um conjunto de dados e é definida como a diferença entre o quartil superior Q3 (os 25% no topo) e o quartil inferior Q1 (os 25% na base) de um conjunto de dados. Os quartis são valores que dividem um conjunto de dados ordenados em QUATRO PARTES iguais. O Q2 é a própria MEDIANA. Q1 e Q3 são as medianas das duas metades restantes. Veja:

Dada uma série de valores, uma forma rápida de achar o Q1 e o Q3 é achando a mediana das duas metades dos dados. Observe, no exemplo abaixo, como achar Q1, Q3 e depois o IQ:

Idade	21	32	34	37	38	43	45	47	47	48	49	49	56	58	61	61	71	71	72	73	74	80

Ache a mediana e divida o conjunto de dados em duas metades.

Posição	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Idade	21	32	34	37	38	43	45	47	47	48	49	49	56	58	61	61	71	71	72	73	74	80

Calcule a mediana das metades superior e inferior.

Posição	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Idade	21	32	34	37	38	**43**	45	47	47	48	49	49	56	58	61	61	**71**	71	72	73	74	80

Mediana inferior = Q1 = “43”

Mediana superior = Q3 = “71”

IQ = Q3 – Q1 = 71 – 43 = 28

Intervalo interquartil é igual a “28”.

Exemplo de análise: se uma prova tem nota máxima de 10, e o IQ de todas as notas for igual a 2, pode-se concluir que pelo menos 50% dos alunos que realizaram a prova tiveram um nível de conhecimento semelhante, uma vez que a amplitude superior-inferior não é tão grande. Se o IQ for igual a 5, por outro lado, pode-se perceber um desempenho muito elevado de alguns em comparação a outros.

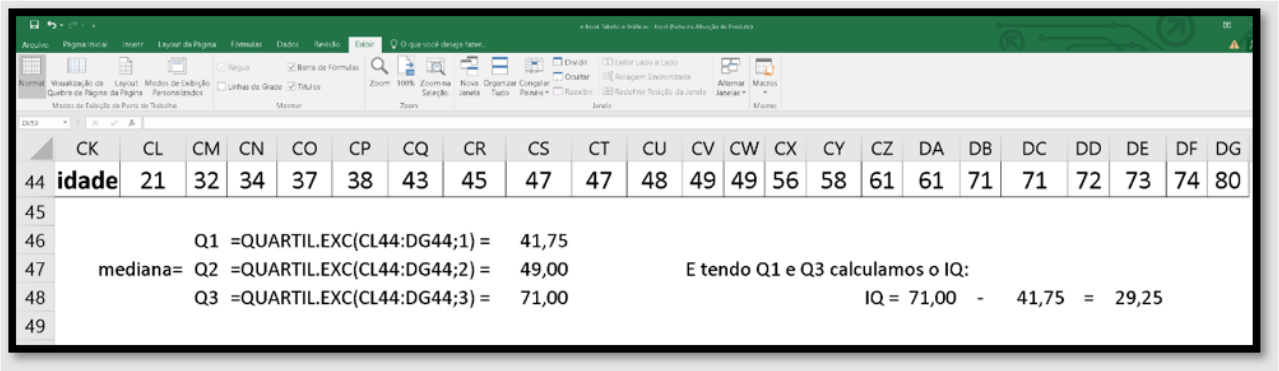
Uma ótima maneira de visualizar a mediana, os três quartis e o IQ é montando com o Excel um gráfico de colunas especial chamado de **BOXPLOT**. Neste caso, o Excel adota duas fórmulas diferentes para calcular o Q1 e o Q3 usando interpolação linear:

=QUARTIL.INC(intervalo;quartil) (aqui se inclui o valor da mediana no cálculo de Q1 e Q3).

=QUARTIL.EXC(intervalo;quartil) (aqui se exclui o valor da mediana no cálculo de Q1 e Q3, que é o método mais seguro para construção do Box-Plot, pois o IQ, nesse caso, é sempre maior que o IQ calculado na primeira fórmula).

Antes de mostrar o BoxPlot criado com o Excel, vamos mostrar como calcular Q1 e Q3 com a fórmula do Excel que exclui a mediana (Q2). Veja o print da tela do Excel com a demonstração abaixo:

Figura 2 - Como calcular Q1 e Q3 com a fórmula do Excel que exclui a mediana (Q2).



	CK	CL	CM	CN	CO	CP	CQ	CR	CS	CT	CU	CV	CW	CX	CY	CZ	DA	DB	DC	DD	DE	DF	DG
44	idade	21	32	34	37	38	43	45	47	47	48	49	49	56	58	61	61	71	71	72	73	74	80
45																							
46					Q1	=QUARTIL.EXC(CL44:DG44;1) =			41,75														
47			mediana=		Q2	=QUARTIL.EXC(CL44:DG44;2) =			49,00							E tendo Q1 e Q3 calculamos o IQ:							
48					Q3	=QUARTIL.EXC(CL44:DG44;3) =			71,00							IQ = 71,00 - 41,75 = 29,25							
49																							

Fonte: Elaborado pelos autores,utilizando a planilha eletrônica Excel

Compare os valores de Q1 e Q3 de antes com o de agora:

Q1: mudou de 43 para 41,75

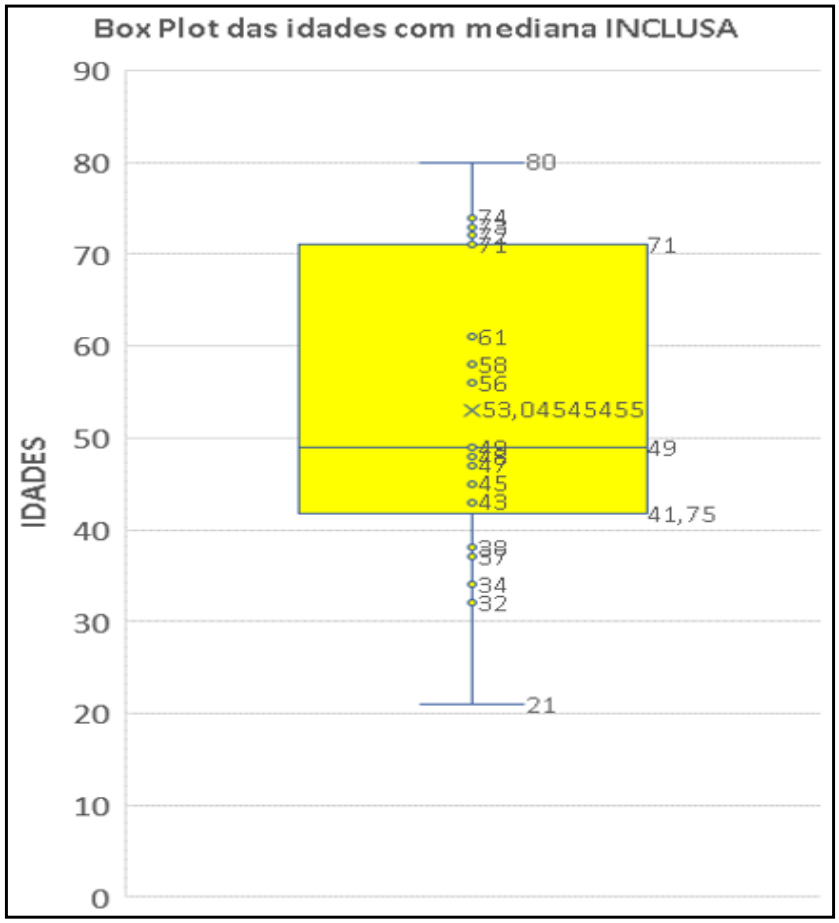
Q3: manteve-se em 71

Agora, veja abaixo como fica o BoxPlot dos valores das idades, considerando os cálculos dos quartis com a mediana exclusa:

A caixa amarela (box) contém 50% das idades dos pacientes, e dentro deste box há pouca variação de idades entre Q1 e Q2: Q2-Q1 = 7,25 anos.

Já a variação entre as idades de Q2 e Q3 é maior: Q3 - Q2 = 22 anos.

Gráfico 6 – BoxPlot.



Fonte: elaborado pelos autores.

E a variação total das idades comparando o limite inferior com o limite superior é de 80 – 21 = 59 anos e bem alta.

Obs.: 53,05 anos é a idade média dos pacientes, que será a próxima medida de posição que veremos.

É possível construir vários BoxPlots para diversas séries de dados. Assista ao vídeo abaixo como exemplo.

youtube.com/watch?v=lKoV5sS2Gqc

Exercício 4: Utilizando no Excel a fórmula dos quartis que exclui a mediana, calcule os quartis (Q1, Q2 e Q3) e monte o BoxPlot dos tamanhos dos cistos no pâncreas dos 22 pacientes utilizando o Excel ou similar. Tamanho do cisto (cm): 6; 10; 8,2; 3; 3,6; 10; 1; 1; 7; 1; 6,6; 4; 3,1; 2,3; 11; 6,5; 13; 1; 7; 12; 8,5; 10.

Em seguida, responda às seguintes questões:

1)Quantos dados estão entre o Q1 e o Q3? E isso dá quantos %?

2)Quando a diferença entre Q1 e Q2 é muito menor que a diferença entre Q2 e Q3, isso significa o quê? Explique.

5.3.3. Média e suas medidas de dispersão

A medida de tendência central, **média aritmética**, é calculada somando todos os valores do conjunto de dados e, em seguida, dividindo essa soma pelo número total de valores. Na prática, a média pode ser considerada como o centro de gravidade dos dados.

Conforme o quadro 2, a média só pode ser usada com dados quantitativos desde que sua distribuição seja **simétrica (ou próxima da simetria)**. Agora, vocês podem perguntar: "... mas o que significa simetria?".

Na estatística, quando um conjunto de dados é **perfeitamente simétrico**, os valores se distribuem de maneira que uma metade do gráfico tem formato idêntico à outra metade.

Exemplo 4: Curioso em constatar se o nº de filhos por família depende da classe social, um pesquisador selecionou 51 famílias de diferentes classes sociais. Dessas 51 famílias, 17 eram de famílias de classe vulnerável, 17 de classe média e 17 de classe rica. O nº de filhos para cada tipo de família apurado foi organizado em TDF sem classes e, a partir delas, montou-se um gráfico de colunas para cada tabela. O resultado segue abaixo:

Gráfico 7

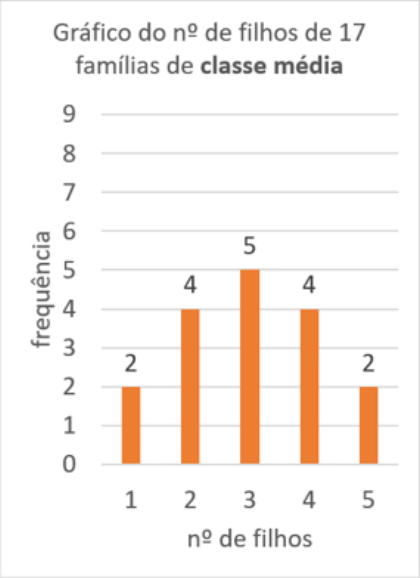


Gráfico 8

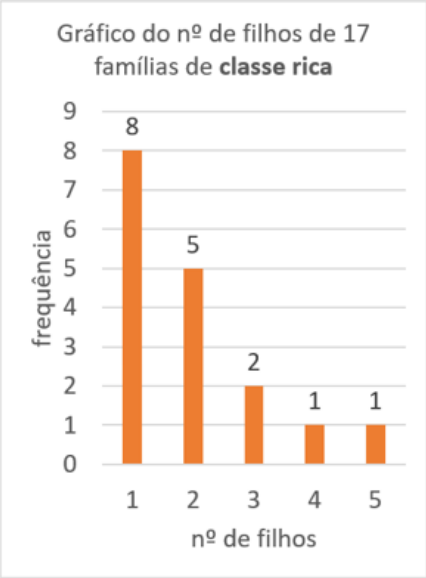
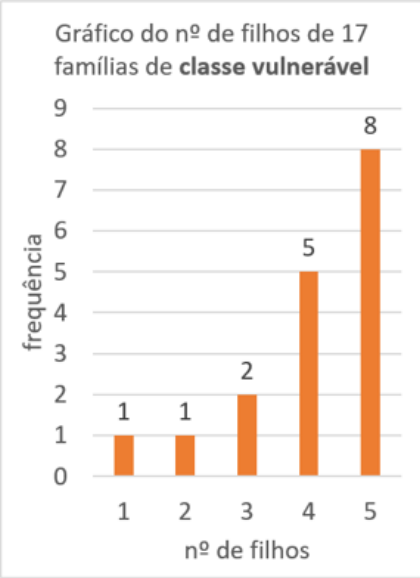


Gráfico 9



Fonte: elaborado pelos autores, com uso do Excel e dados fictícios.

Observem que o resultado foi como esperado: famílias ricas tendem a ter menos filhos por casal, as vulneráveis tendem a ter mais filhos por casal e as famílias de classe média um nº de filhos sem grandes extremos (não continue a leitura sem constatar essas informações interpretando os gráficos acima).

Porém, somente um desses três gráficos é simétrico, ou seja, pode ser dividido em duas metades IDÊNTICAS. Qual você acha que é? Se respondeu que é o gráfico dos nº de filhos de famílias de classe média, você acertou, pois as duas metades são iguais.

Gráfico 10

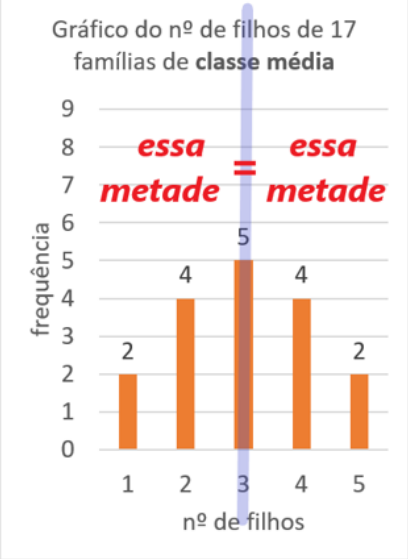


Gráfico 11

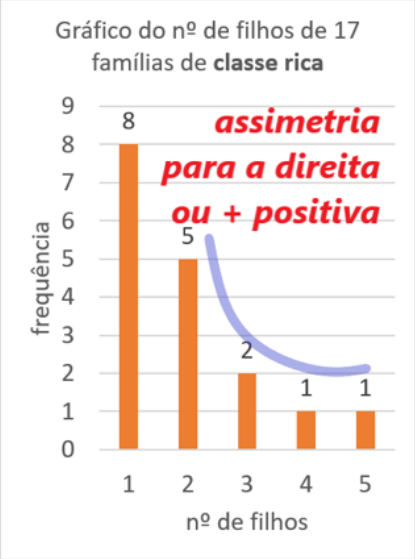
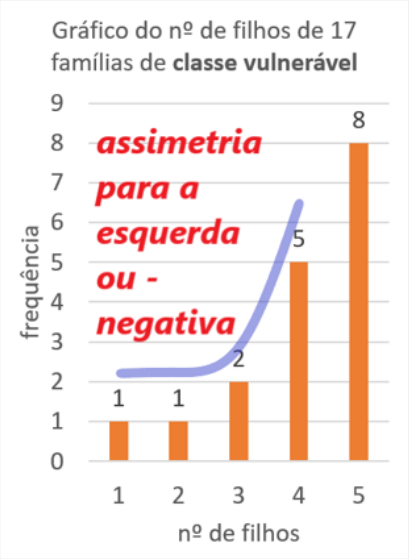


Gráfico 12



Fonte: elaborado pelos autores, com uso do Excel e dados fictícios.

Temos, assim, três possibilidades básicas para a simetria de dados: SIMÉTRICO, ASSIMÉTRICO POSITIVO (tem uma cauda caindo para a direita), ou ASSIMÉTRICO NEGATIVO (tem uma cauda caindo para a esquerda).

Obs.: Ainda há uma 4ª possibilidade quando um gráfico apresenta mais de um PICO, na qual, nesse caso, aquele gráfico seria classificado simplesmente como ASSIMÉTRICO.

Entendido o conceito de distribuição simétrica ou assimétrica (+ ou -), por que a média representa melhor dados com distribuição simétrica (ou pouco assimétrica)?

Porque, diferentemente da MEDIANA, que é uma medida robusta quanto a sua localização (ela é sempre o valor que está no meio de todos os valores ordenados), a MÉDIA já é uma medida facilmente afetada pelos valores muito afastados dos demais.

Imagine que a média da idade de 5 crianças é:

$$\frac{2+2+2+4+4+4+5+5}{8} = \frac{28}{8} = 3{,}5 \text{ anos}$$

A mediana dessas mesmas idades é o valor do meio, que, no caso, é o (4+4)/2 = 4 anos

Imagine, agora, que **substituíssemos** a última idade de 5 anos por uma pessoa com 25 anos.

$$\frac{2+2+2+4+4+4+5+25}{8} = \frac{28}{8} = 6{,}0 \text{ anos}$$

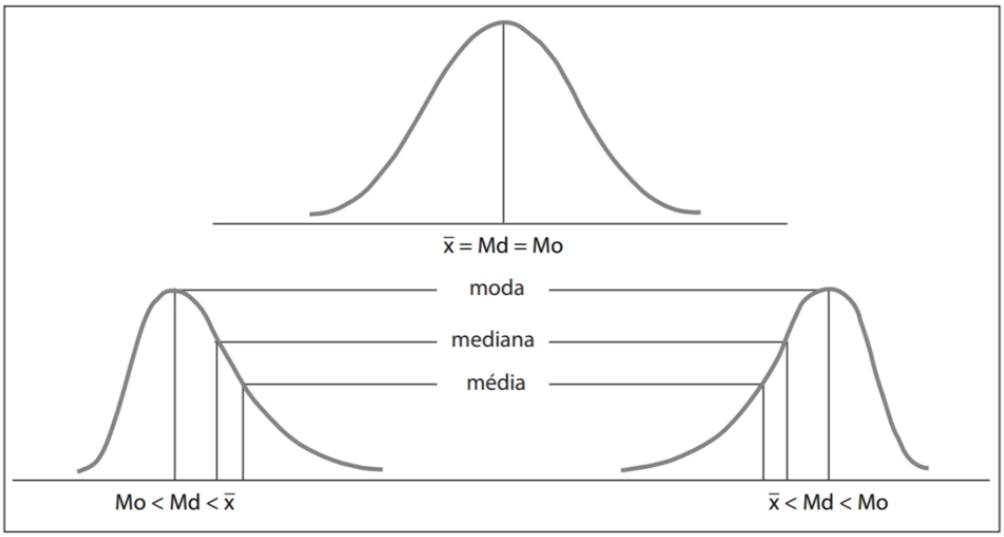
Notem que a média foi PUXADA para cima, chegando a 6,0 anos de idade média. Porém, a MEDIANA não foi afetada pela idade de 25 anos, permanecendo os mesmos (4 + 4)/2 = 4 anos.

É nesse sentido que dizemos que a média é muito útil para representar o centro de dados simétricos ou com baixa assimetria e, quando há assimetria é alta, é melhor usar a MEDIANA com o IQ.

As **três possibilidades básicas para a simetria** de dados também mostraram aos estatísticos que, quando uma distribuição de dados é simétrica, os valores da MÉDIA, MODA e MEDIANA são iguais. Sendo a distribuição assimétrica à esquerda ou negativa, a MÉDIA < MEDIANA < MODA. Sendo assimétrica à direita ou positiva, a MÉDIA > MEDIANA > MODA.

A figura 3 do livro de Crespo (2020, p. 127) ajuda na visualização destes conceitos:

Figura 3 – Representação das medidas de tendência central.



Fonte: Crespo (2020, p. 127).

A partir dessa figura X, Crespo (2020, p. 128) explica que, baseado na relação entre a média e a moda, é possível determinar o tipo de assimetria, calculando a diferença entre elas:

$\underline{X} - M_o = 0 \rightarrow$ assimetria nula ou distribuição simétrica

$\underline{X} - M_o < 0 \rightarrow$ assimetria negativa ou à esquerda

$\underline{X} - M_o > 0 \rightarrow$ assimetria positiva ou à direita

Vejamos mais dois exemplos (adaptados) antes de avançarmos no cálculo da média e suas medidas de dispersão.

Exemplo 5: Levin, Fox e Forde (2012, p. 88 e 89) mostraram a vantagem da mediana sobre as demais medidas em uma distribuição assimétrica de salários anuais de secretárias que trabalham em uma universidade. Vamos à tabela de salários para calcular as possíveis medidas de tendência central.

Salários

$ 120.000
$ 60.000 $\underline{x} = 50.000$
$ 40.000 $M_d = 40.000$
$ 40.000 $M_o = 30.000$
$ 30.000 $\underline{x} - M_o > > 0$ temos assimetria positiva forte
$ 30.000
$ 30.000

Se quiséssemos promover a empresa como uma organização que oferece ótimos salários, provavelmente iríamos calcular a média para mostrar que o empregado médio ganha $50.000 por ano.

Porém, caso fôssemos representantes do sindicato buscando argumentos para aumentar o salário das secretárias, provavelmente usaríamos a moda para mostrar que o salário mais comum é de $30.000 por ano.

E, finalmente, caso fôssemos pesquisadores sociais buscando divulgar de maneira precisa o salário médio das secretárias, corretamente usaríamos o salário MEDIANO de $40.000 por ano, pois ele se localiza entre a média e a moda, fornecendo uma informação **mais equilibrada da estrutura salarial** da universidade.

Exemplo 6: Levin, Fox e Forde (2012, p. 89) destacam, também, que há distribuições que podem ser BIMODAIS (possuem duas modas), possuindo, assim, dois pontos de frequência máxima. Nesse caso, a média vai ficar entre esses dois valores MODAIS e pode ocorrer de termos uma distribuição SIMÉTRICA, mas na qual $M_o \neq$ mediana e da média. Imagine um pesquisador social ou da saúde que entrevistou 20 pessoas de baixa renda perguntando "qual o tamanho de família ideal para você, considerando o casal e filhos?" As respostas seguem abaixo ordenadas em ordem crescente.

1 2 2 2 3 3 3 3 4 4 5 6 6 7 7 7 7 8 8 9

Moda =3 Moda = 7

$\underline{x} = 4{,}85$

Usando a média ou a mediana $M_d = 4{,}5$, poderíamos concluir que o tamanho ideal de família do entrevistado médio fica entre 4 e 5 pessoas, quando, na verdade, somente três pessoas responderam 4 ou 5.

Sabendo que a distribuição é BIMODAL, sabemos que houve duas preferências para tamanho de família ideal nesse grupo de 20 entrevistados: parte prefere famílias pequenas com $M_o = 3$ pessoas por família, enquanto outra parte prefere famílias grandes com $M_o = 7$ pessoas por família.

Vamos, agora, ilustrar com um exemplo uma das inúmeras aplicações de uso da média e suas medidas de dispersão.

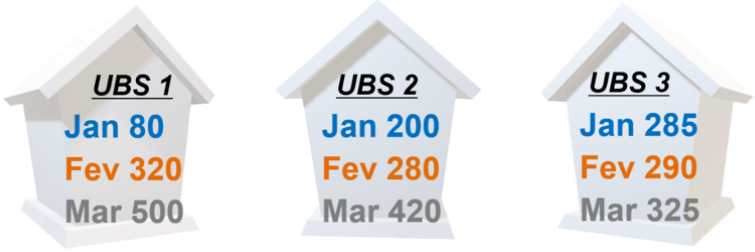
Exemplo 7: Após os três primeiros meses de 2022, as UBSs 1, 2 e 3 de um bairro de São Paulo apresentaram o mesmo consumo médio de luvas de proteção.

Figura 4 – Consumo médio de 300 luvas/mês.

Fonte: elaborada pelos autores.

Olhando somente para o valor médio idêntico de consumo de luvas nas três UBSs, podemos ter a falsa impressão de que para o 4º mês (abril) serão consumidas mais 300 luvas em cada UBS. Porém, ao olharmos os dados de consumo mensais, podemos ter uma surpresa...

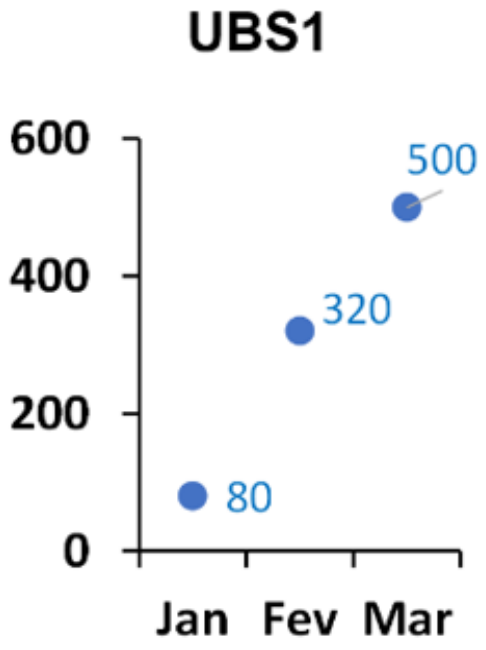
Figura 5 – Comparativo de consumo mensais.



Fonte: elaborada pelos autores.

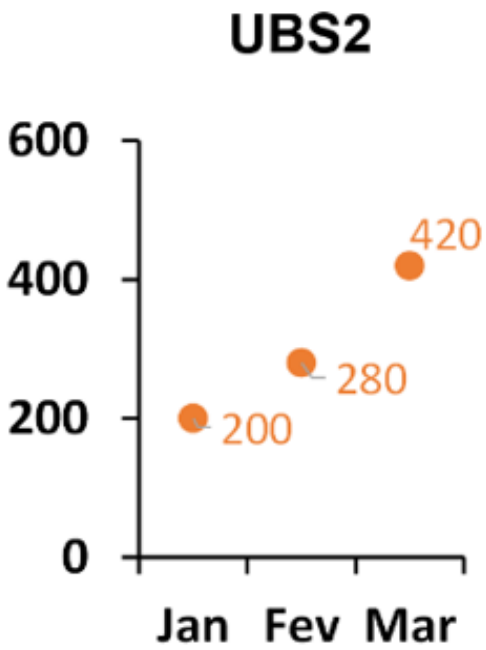
Percebam que o consumo de luvas das UBSs **1** e **2** estão **muito mais dispersos** do que na UBS **3**, além do fato de que a tendência de aumento de consumo da UBS 3 é bem menor do que a tendência de consumo de luvas nas outras duas UBSs (1 e 2). Vejam todos esses dados de consumo nos gráficos abaixo.

Gráfico 13 – Dados de consumo.



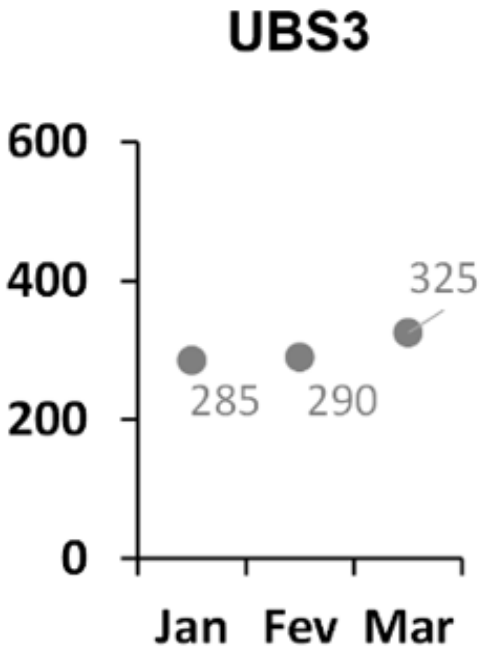
Fonte: elaborado pelos autores.

Gráfico 14 – Dados de consumo.



Fonte: elaborado pelos autores.

Gráfico 15 – Dados de consumo.



Fonte: elaborado pelos autores.

Notaram como, nas UBSs 1 e 2, haverá um alto risco de falta de material em abril caso se decida comprar apenas 300 luvas para cada uma delas? Ao contrário da UBS 3, em que o risco de falta de luvas é bem menor caso se compre as 300 luvas para abril.

Perceberam também que o consumo mensal na USB 1 muda muito mais que na UBS 2 e que, na UBS 3, a mudança ou variação de consumo é bem menor? Se sim, vocês acabaram de compreender o conceito de DISPERSÃO (ou VARIAÇÃO DOS DADOS).

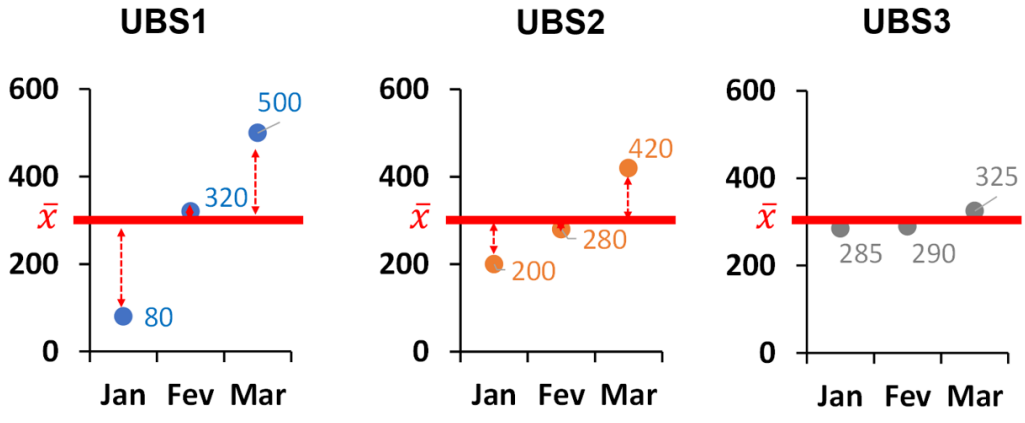
O consumo de luvas na UBS 1 está muito mais DISPERSO do que nas outras duas. O consumo na UBS 2 está muito mais DISPERSO do que na UBS 1. Já o consumo de luvas da UBS 1 é o MENOR, pois seus valores estão muito próximos de sua média de 300 luvas/mês.

Inserindo a média de consumo de 300 luvas/mês nos gráficos das três UBSs, podemos ter uma visão mais clara da dispersão:

Gráfico 16 - Média de consumo de 300 luvas/mês.

Gráfico 17 - Média de consumo de 300 luvas/mês.

Gráfico 18 - Média de consumo de 300 luvas/mês.



Fonte: elaborado pelos autores.

Conclusão inicial: quanto mais longe os dados estão de sua MÉDIA, maior é sua DISPERSÃO. E quanto mais próximos os dados estiverem de sua média, menor será a DISPERSÃO.

> "**A média aritmética sempre** deve estar acompanhada de uma **medida de dispersão"**

Como apresentado no quadro 2, há várias possibilidades de escolha de medida de dispersão para acompanhar a média: AMPLITUDE, VARIÂNCIA, DESVIO-PADRÃO, COEFICIENTE DE VARIAÇÃO (CV) E Z-SCORE.

No quadro 3, apresentamos um resumo de cada medida de dispersão da média, com suas fórmulas matemáticas, comandos de planilha eletrônica e indicações de uso.

Quadro 3 – Medidas de dispersão para a MÉDIA.

MEDIDA DE DISPERSÃO	COMO CALCULAR	QUANDO USAR
Amplitude (A)	A = Maior - Menor	Com dados ***quantitativos*** ***Muito*** sensível aos valores discrepantes (outliers) Só dá uma ideia da ***dispersão entre os extremos***
Desvio Padrão Populacionnal ("σ")	$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{n}}$ $\sigma = \sqrt{\frac{\sum x^2}{n} - \mu^2}$ ou =DESVPAD.P(dados)	Com dados ***quantitativos*** É mais precisa que a **"A"**, pois determina a ***dispersão de todos os dados em relação a média desses dados*** . Difícil e trabalhosa de calcular. Melhor calcular com planilha ou outros recursos eletrônicos
Desvio Padrão Amostral ("s")	$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ $s = \frac{\sqrt{n}}{\sqrt{n-1}} \cdot \sqrt{\frac{\sum x^2}{n} - (\bar{x})^2}$ ou '=DESVPAD.A(dados)	
Coeficiente de variação ("cv")	$cv = \frac{\sigma}{\mu} * 100\%$ cv da população $cv = \frac{s}{\bar{x}} * 100\%$ cv da amostra	Necessário saber a relação entre o desvio padrão e a média. 'Com ***médias*** muito diferentes ou de ***unidades de medida*** distintas
Escores-z	$z = \frac{x-\mu}{\sigma}$	Para saber quantos desvios um valor está distante da média.

Fonte: Griffths, Dawn. Use a cabeça Estatística, 2009.

5.3.3.1. Exemplos das medidas de dispersão da média

Continuaremos a usar os dados sobre o consumo de luvas nas três UBSs (exemplo 7) para calcular as várias medidas de dispersão que podem acompanhar a média. Já sabemos os valores dos três primeiros meses de cada UBS do exemplo 7, e que o consumo médio de luvas para as três UBSs é o mesmo: 300 luvas por mês. Vejamos como calcular e interpretar as dispersões de cada UBS:

AMPLITUDE: A = MAIOR – MENOR		
UBS 1	UBS 2	UBS 3
A = 500 – 80 = 420 luvas	A = 420 – 200 = 220 luvas	A = 325 – 285 = 40 luvas
Conclusão: quanto maior a amplitude, maior a dispersão dos dados. Em outras palavras: quanto maior a amplitude, mais alguns valores estarão MUITO longe de sua média de 300 luvas, e VICE-VERSA.		

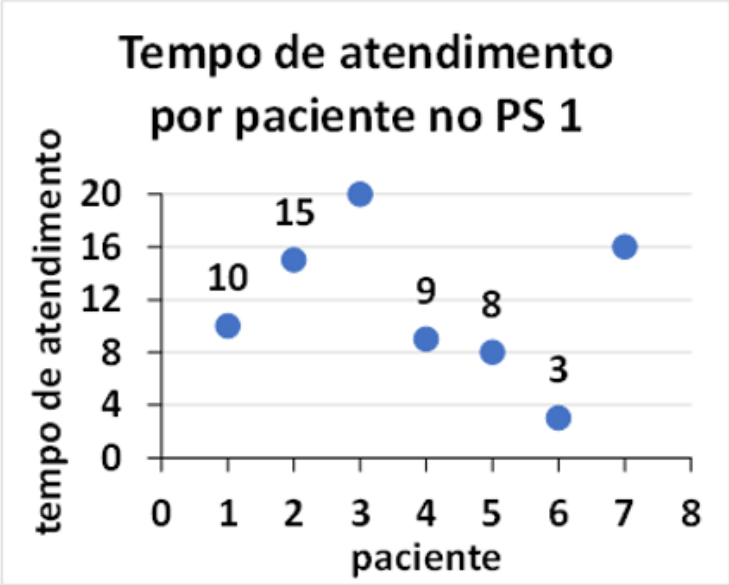
DESVIO-PADRÃO AMOSTRAL: S = DESVPAD.A(selecione dados)		
UBS 1	UBS 2	UBS 3
S = 210,71 luvas	S = 111,36 luvas	S = 21,79 luvas
Conclusão: quanto maior o "S", maior a dispersão dos dados. E quanto menor o "S", menor a dispersão dos dados. Em outras palavras: quanto maior o "S", mais alguns valores estarão MUITO longe de sua média de 300 luvas, e VICE-VERSA.		

COEFICIENTE DE VARIAÇÃO: CV = (S/MÉDIA) *100		
UBS 1	UBS 2	UBS 3
CV= 210,71/300 = 0,70 = **70%**	CV= 111,36/300= 0,371= **37%**	CV= 21,79/300 = 0,073 = **7%**
Conclusão: quanto maior o "CV", maior a dispersão dos dados. E quanto menor o "CV", menor a dispersão dos dados. CV>10% indica alta dispersão. CV<10% indica baixa dispersão. Em outras palavras: quanto maior o "CV", mais alguns valores estarão MUITO longe de sua média de 300 luvas, e VICE-VERSA.		

Z-SCORE: Z = (X - MÉDIA)/S		
UBS 1	UBS 2	UBS 3
Calcule quantos "Zs" o consumo de JANEIRO está longe da média	Calcule quantos "Zs" o consumo de MARÇO está longe da média	Calcule quantos "Zs" o consumo de MARÇO está longe da média
Para tanto, precisaremos: S = 210,71 luvas Média = 300 luvas Consumo de JAN: X = 80 luvas Z = (80 – 300)/210,71 Z = -220/210,74 = **-1,04**	Para tanto, precisaremos: S = 111,36 luvas Média = 300 luvas Consumo MAR: X = 420 luvas Z = (420– 300)/210,74 Z = +120/111,37= **+1,08**	Para tanto, precisaremos: S = 21,79 luvas Média = 300 luvas Consumo MAR: X = 325 luvas Z = (325– 300)/21,79 Z = +25/21,79 = **+1,15**
Interpretação: 80 está pouco mais de **1S** (um S) à **esquerda (Z deu negativo)** da média de 300 luvas.	Interpretação: 420 está a pouco mais de **1S** à **direita (Z deu positivo)** da média de 300 luvas.	Interpretação: 325 está a pouco mais de **1S** à **direita (Z deu positivo)** da média de 300 luvas.
Utilização do Z-score: são inúmeras, e uma delas é dizer se um valor é ou não um **outlier** (difere muito dos demais). Quando o (3 vezes o desvio-padrão), dizemos que esse valor X é um outlier, por estar muito longe da maioria dos dados. Neste caso, pode ser que esse valor X em questão possa ser desconsiderado no cálculo da média.		

Exercício 5: O tempo (min.) de atendimento de 7 pacientes no posto de saúde 1 (PS 1) é dado no gráfico abaixo:

Gráfico 19 – Tempo de atendimento por paciente.



Fonte: elaborado pelos autores.

a. Qual o tempo médio?

b. Qual o desvio-padrão amostral?

c. Qual o CV?

d. Qual o z-escore do maior e do menor valor?

Finalizamos, aqui, o que pretendíamos abordar sobre estatística descritiva. Vamos, agora, aplicar o conteúdo visto fazendo um estudo epidemiológico, com base nos óbitos da população residente no município de São Paulo, devido ao câncer de estômago entre 2013 e 2022. Veja o passo a passo no Apêndice 1.

bibliografia

ANDERSON, David R.; SWEENEY, Dennis J.; WILLIAMS, Thomas A.; CAMM, Jeffrey D.; COCHRAN, James J. Estatística aplicada a administração e economia – Tradução da 8ª edição norte-americana. São Paulo: Cengage Learning Brasil, 2019. E-book. ISBN 9788522128006. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788522128006/. Acesso em: 16 set. 2024.

COCHRAN, James J. **Estatística apli-cada à administração e economia** – Tradu-ção da 8ª edição norte-americana. São Pau-lo: Cengage Learning Brasil, 2019. E-book.

GRIFFITHS, Dawn. **Use A Cabeça! Estatística**. Rio de Janeiro: Alta Books, 2009.

ISBN 9788522128006. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788522128006/. Acesso em: 25 ago. 2023.

CRESPO, Antônio A. **Estatística** (Série EM FOCO) 2ª . São Paulo: Editora Saraiva, 2020. E-book. ISBN 9788571440821. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788571440821/. Acesso em: 25 ago. 2023.

LEVIN, Jack; FOX, James Alan; FORDE, David. **Estatística para ciências humanas**. 11. ed. São Paulo: Pearson, 2012. 458 p.

https://www.prefeitura.sp.gov.br/cidade/secretarias/saude/tabnet/

apêndices

Estudo epidemiológico

O objetivo é realizar um estudo série temporal (de 2013 a 2022) e descritivo sobre óbitos, por SEXO, devido ao câncer de estômago na população do município de São Paulo (MSP). Para tanto, iremos precisar entender o conceito de Coeficiente de Mortalidade Específico (CME), o qual é representado pela razão abaixo:

$$\text{CME} = \frac{\text{nº de óbitos devido a uma causa específica}}{\text{população de risco}}$$

Nesse estudo, a causa específica escolhida foi o câncer de estomago no MSP, portanto o CME de cada ano para cada sexo será:

$$\text{CME} = \frac{\text{nº de óbitos devido a câncer de estômago}}{\text{população do MSP}}$$

Os dados devem ser baixados do TabNet do MSP (item 6.1). Após baixar os dados, vamos:

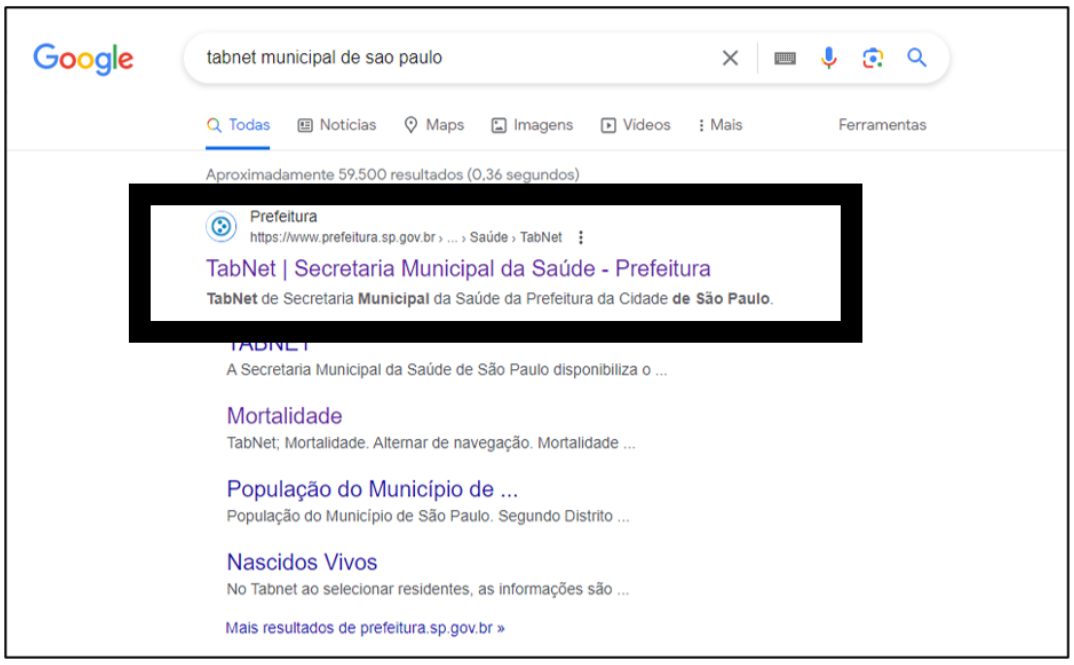
- Organizar os dados em uma tabela (planilha) de série temporal, calculando os CMEs.
- Calcular as diferenças percentuais (ou variações percentuais) entre os valores do início de cada série (2013) e final (2022). Exemplo: variação percentual entre o CME de 2022 em relação ao CME de 2013 de óbitos masculinos da doença X.
- Calcular e comparar as médias, desvios-padrões e coeficientes de variação (CV) dos CMEs.
- Analisar como serão os gráficos dos **CME** com base na dispersão dos dados apontada pelo CV, explicando por que a dispersão será alta ou não.
- Montar um gráfico de linhas de 2013 a 2022, corroborando a conclusão da análise da dispersão.

1. Baixando os dados do TabNet do MSP

O TabNet é uma ferramenta de tabulação desenvolvida pelo DATASUS que permite tabulações online de dados e geração de planilhas, com rapidez e objetividade, da base de dados do SUS, ou seja, dados de receitas totais e despesas com ações e serviços públicos de saúde (ASPS) dos entes federados declarados no SIOPS (SIOPS – Manual TabNet, 2012).

Como acessar:

Digitar em um buscador, como o Google, "TabNet municipal de São Paulo" e clicar no destaque abaixo:

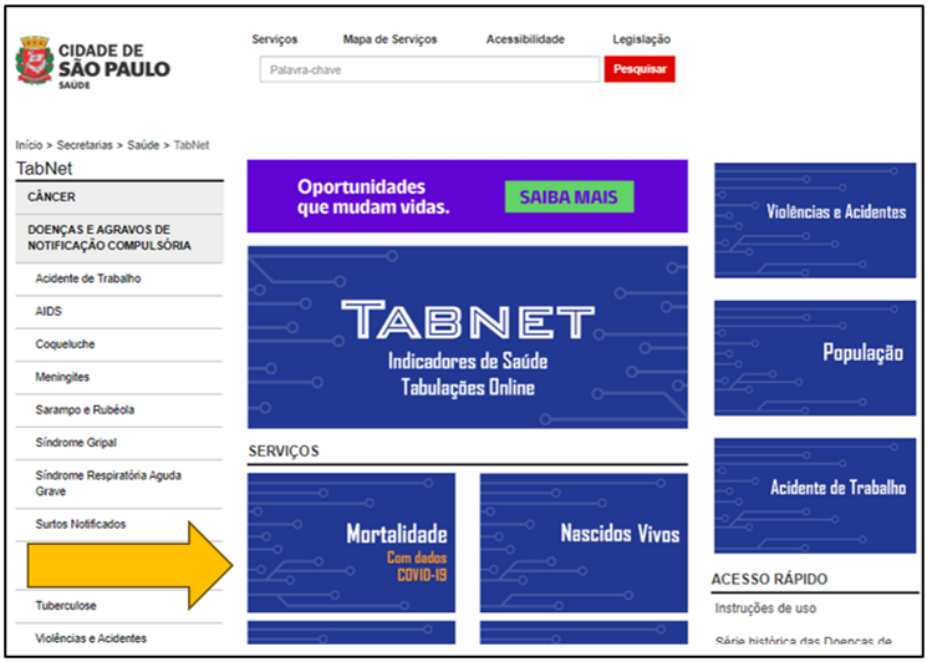


Fonte: cópia modificada da tela do Google ao pesquisar sobre o TabNet do MSP.

1.1. Baixando os óbitos de 2013 a 2019

Como exemplo, vamos baixar no TabNet uma tabela com o número de óbitos por câncer de estômago, por sexo, de 2013 a 2019 no MSP. Como baixar as planilhas: clique em Mortalidade com dados COVID-19.

Tela do TabNet para acessar a mortalidade



Fonte: cópia modificada da tela do TabNet.

Em seguida clicar em Mortalidade Geral:

Tela do TabNet para acessar Mortalidade Geral

Fonte: cópia modificada da tela do TabNet.

Aparecerá uma tela (próxima página) com vários filtros que precisam ser definidos.

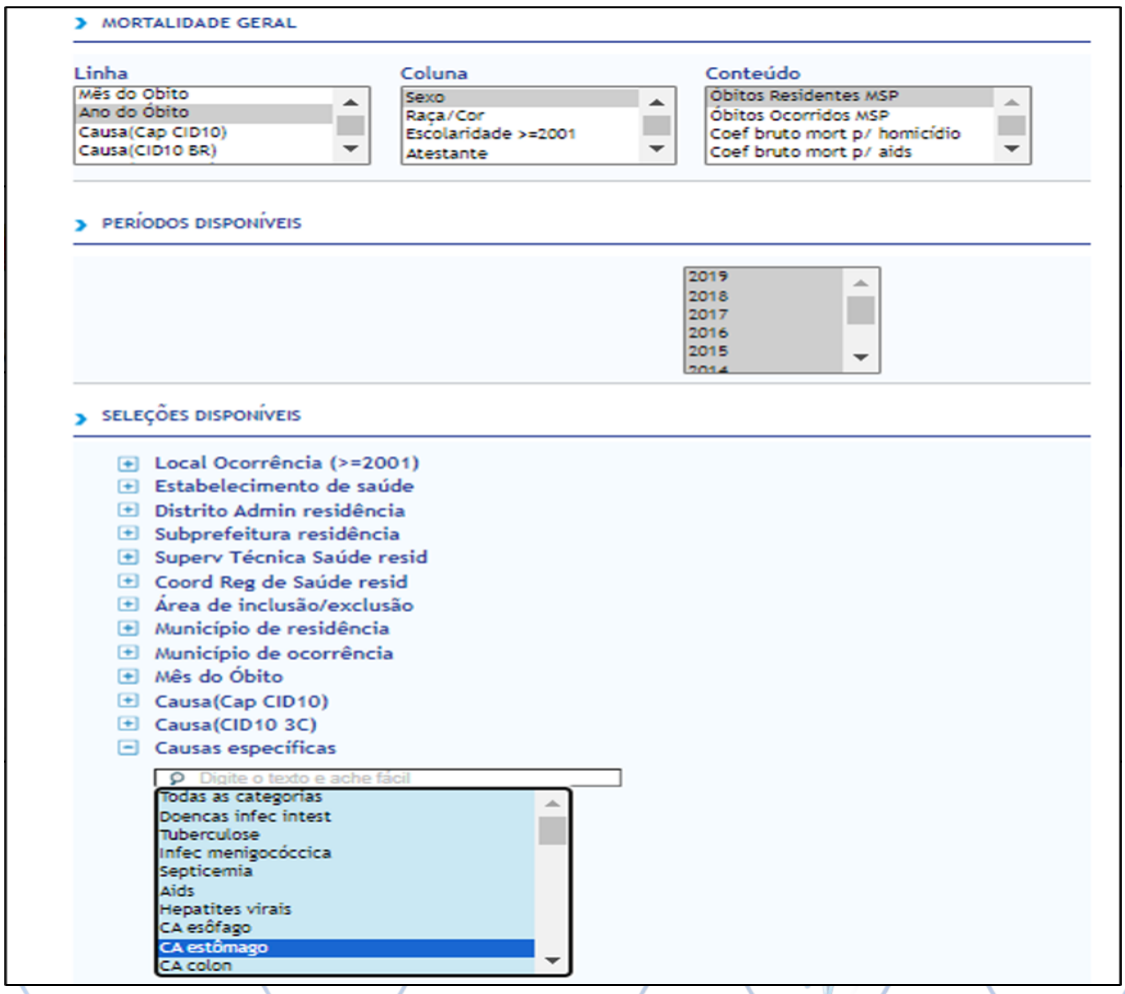
Para as tabelas terem as informações desejadas, é necessário alterar os tópicos de Linha, Coluna e Conteúdo. Assim como os anos (Períodos Disponíveis) e tema (Seleções Disponíveis).

Selecione os seguintes **tópicos**:

- **Linha**: ano do óbito
- **Coluna**: sexo
- **Conteúdo**: óbitos residentes
- **Períodos disponíveis:** 2013 a 2019
- **Seleções disponíveis:** escolher em **Causa** (CID 10 3C) o código da doença que será estudada ou digitar o nome da doença em **CAUSAS ESPECÍFICAS**, como fizemos.

Fonte: cópia modificada da tela do TabNet.

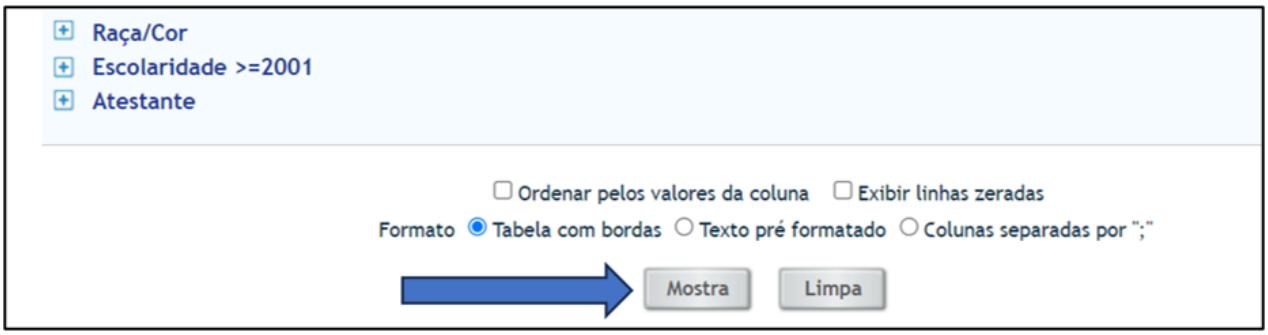
Abaixo apresentamos uma cópia da tela do TabNet com as definições dos tópicos explicados acima.



Fonte: cópia modificada da tela do TabNet.

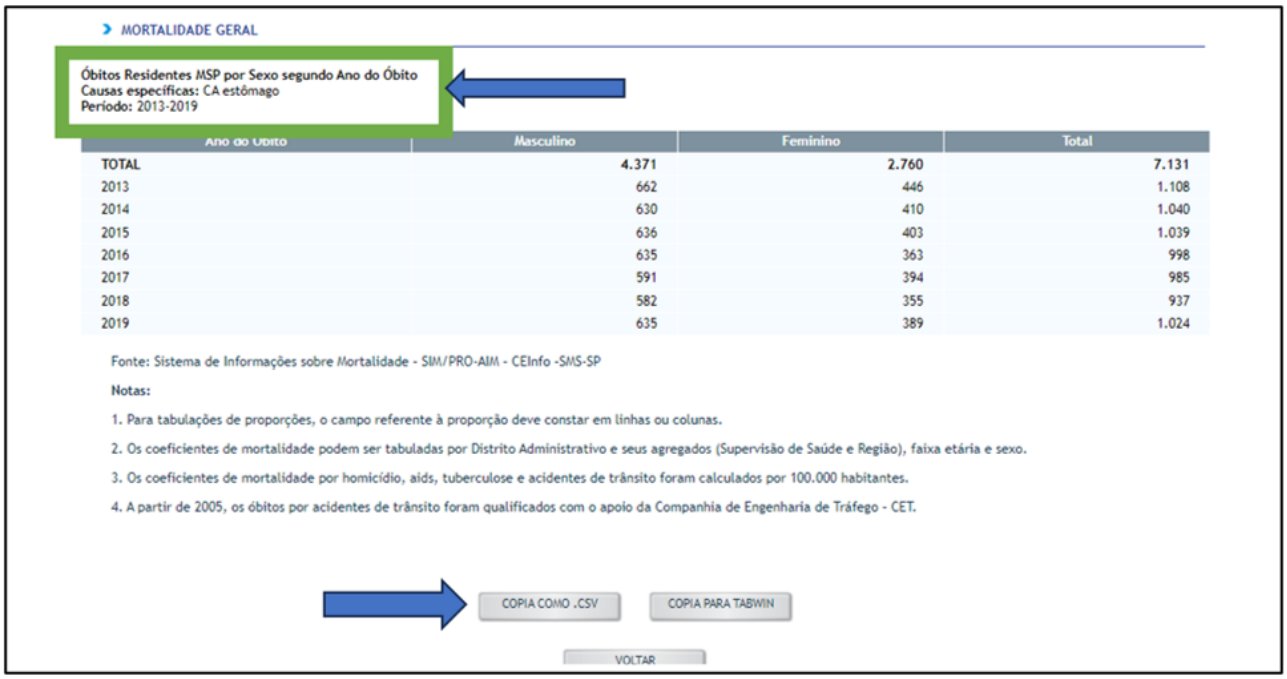
Ao terminar de definir cada tópico (Linha, Coluna, Conteúdo etc.), clique no final da página no botão mostra para obter os primeiros dados populacionais:

Botão "mostra" na tela do TabNet para mostrar resultado da pesquisa



Fonte: cópia modificada da tela do TabNet.

Assim, o TabNet irá gerar o **nº óbitos por sexo, devido ao CA de estômago de 2013 a 2019**, no MSP. É importante conferir se a tabela foi gerada com os dados solicitados, observando o que está no destaque em VERDE abaixo:



Fonte: cópia modificada da tela do TabNet.

Para baixar e salvar esses dados, role a tela para baixo e clique em CÓPIA COMO .CSV

Abrindo o arquivo salvo, veremos uma planilha Excel com dados no formato de texto.

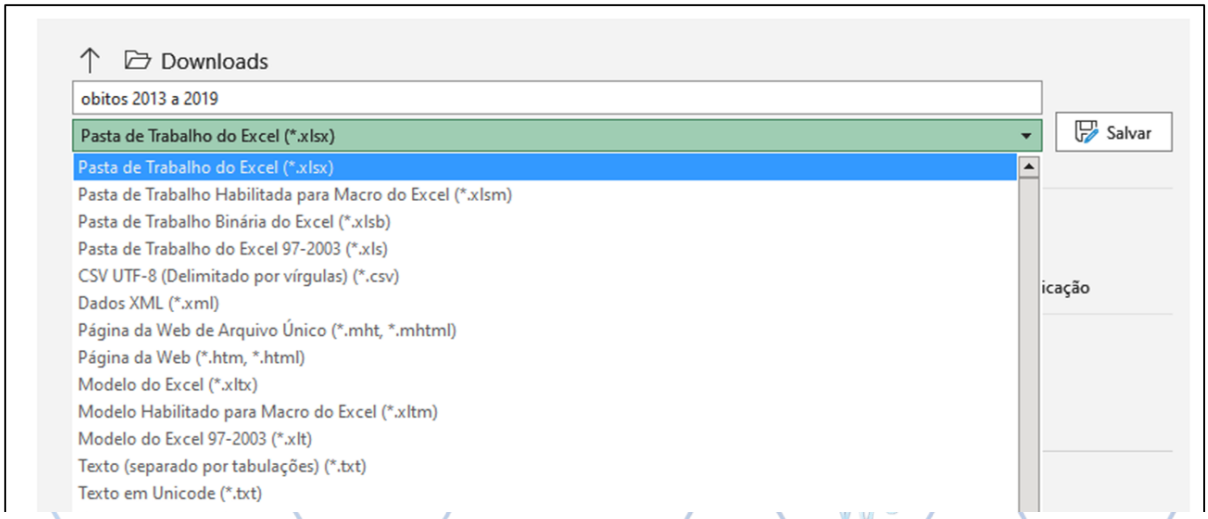
O CSV (Comma Separated Values) é um tipo de arquivo de texto que pode ser aberto pelo Excel (XLS = Excel Spreadsheet), porém, como os dados continuam no formato de texto, o Excel não consegue salvar os gráficos que iremos precisar criar.

Arquivo .CSV aberto com o Excel

	A	B	C	D	E	F	G
1	Mortalidade Geral						
2	Óbitos Residentes MSP por Ano do Óbito e Sexo						
3	Causas específicas: CA estômago						
4	Período:2013-2019						
5	Ano do Óbito	Masculino	Feminino	Total			
6	2013	662	446	1108			
7	2014	630	410	1040			
8	2015	636	403	1039			
9	2016	635	363	998			
10	2017	591	394	985			
11	2018	582	355	937			
12	2019	635	389	1024			
13	Total	4371	2760	7131			
14	Fonte: Sistema de Informações sobre Mortalidade – SIM/PRO-AIM – CEInfo –SMS-SP						
15	Notas:						
16	1. Para tabulações de proporções, o campo referente à proporção deve constar em linhas ou colun						

Fonte: print do arquivo CSV aberto com o Excel.

Para poder criar gráficos, é necessário salvar o arquivo .CSV como .XLS. Para tanto, clique em "arquivo", depois em "Salvar como" e em "tipo:" selecione **Pasta de Trabalho do Excel (*.xlsx).**



Fonte: print da tela do Excel.

1.2. Baixando os óbitos de 2020 a 2022

Com a chegada da COVID, no início de 2020, foi necessário separar a mortalidade geral devido a causas externas (homicídios, suicídios, acidentes e outras mortes devido a fatores externos) das demais mortalidades, para não atrasar o lançamento dos dados dos óbitos devido à COVID-19.

Sendo assim, para baixar os dados de 2020 a 2022, deve-se entrar de novo no TabNet do município de São Paulo (como feito no 1.1), clicar em **Mortalidade** como mostrado abaixo.

Tela do TabNet indicando onde baixar mortalidade

Fonte: cópia modificada da tela do TabNet.

E, em seguida, clicar em Mortalidade Geral exceto causas externas (2019 a 2023), pois esse link possui dados posteriores a 2019.

Tela do TabNet mostrando onde baixar mortalidade por causas externas

Fonte: cópia modificada da tela do TabNet.

Em seguida, teremos mais uma janela onde devemos clicar em Mortalidade Geral – exceto causas externas (com dados COVID 2019).

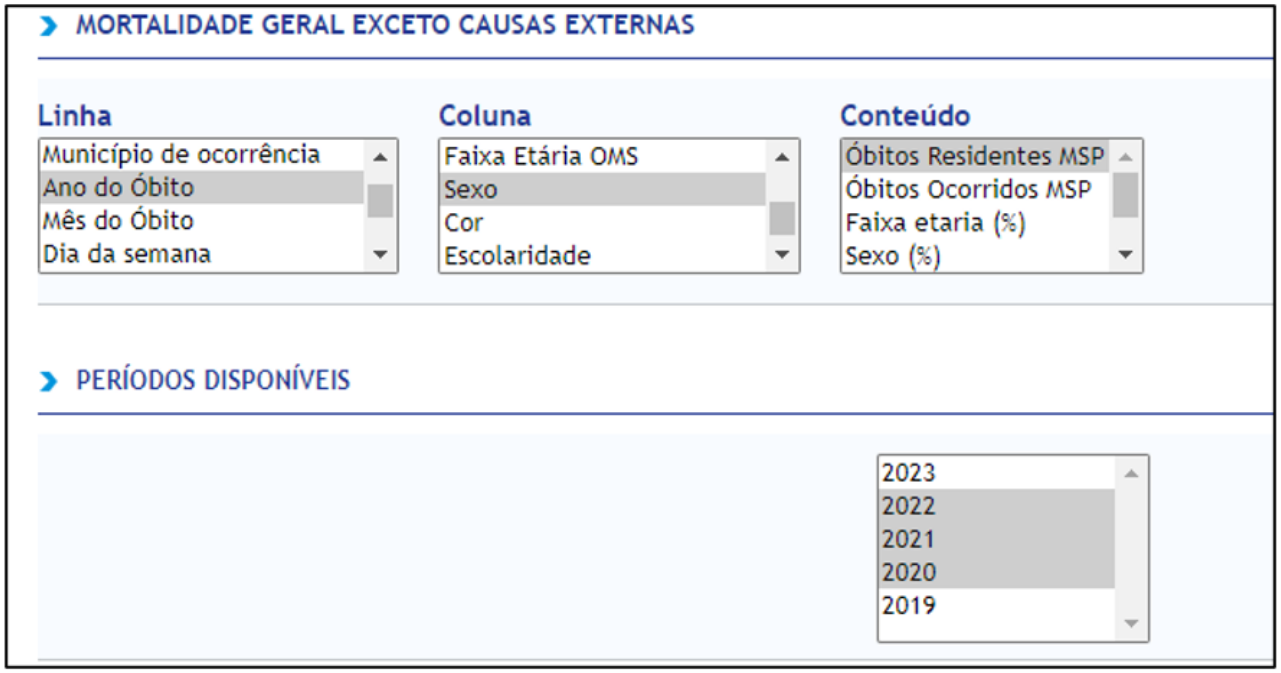
Tela do TabNet mostrando onde baixar mortalidade por causas externas

Fonte: cópia modificada da tela do TabNet.

As etapas seguintes são iguais às da tabela anterior (2013 a 2019). Selecione os seguintes tópicos:

- Linha: ano de óbito
- Coluna: sexo
- Conteúdo: óbitos residentes
- Períodos disponíveis: 2020 a 2022
- Seleções disponíveis: escolher em Causa (CID 10 3C) o código da doença que será estudada ou em CAUSAS ESPECÍFICAS como fizemos em 6.1.1.

Tela do TabNet onde filtramos as informações dos anos faltantes



Fonte: cópia modificada da tela do TabNet.

Baixar a planilha como .CSV e salvar como Excel, como explicado anteriormente.

1.3. Juntando os óbitos de 2013 a 2022

Para juntar os dados de óbitos das duas planilhas, selecione e copie os dados da tabela de 2020 a 2022 para, em seguida, colá-los na tabela de 2013 a 2019:

Arquivo Excel mostrando onde adicionar as mortalidades de 2020 a 2022

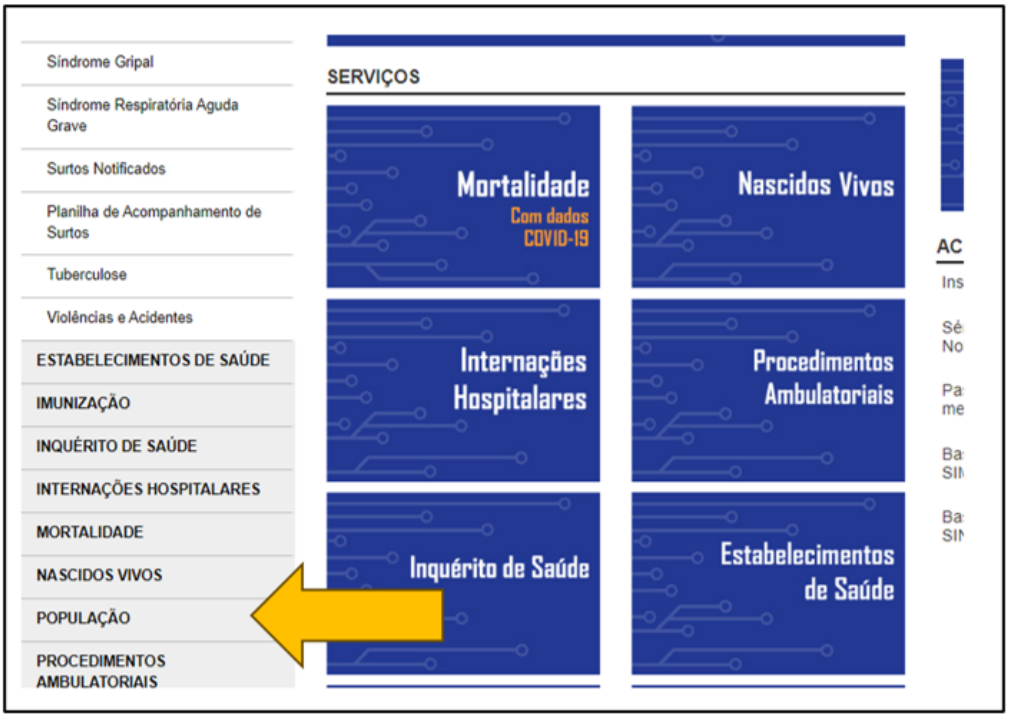
	A	B	C	D	E	F
1	Mortalidade Geral					
2	Óbitos Residentes MSP por Ano do Óbito e Sexo					
3	Causas específicas: CA estômago					
4	Período:2013-2019					
5	Ano do Óbito	Masculino	Feminino	Total		
6	2013	662	446	1108		
7	2014	630	410	1040		
8	2015	636	403	1039		
9	2016	635	363	998		
10	2017	591	394	985		
11	2018	582	355	937		
12	2019	635	389	1024		
13	2020	500	334	834		
14	2021	521	387	908		
15	2022	508	340	848		
16	TOTAL	5900	3821	9721		

Fonte: cópia da tela do Excel.

Neste ponto, já temos todos os óbitos por sexo e ano de 2019 a 2022 na mesma planilha. Só falta agora baixar os dados da população do MSP.

1.4. Baixando a população de 2013 a 2022

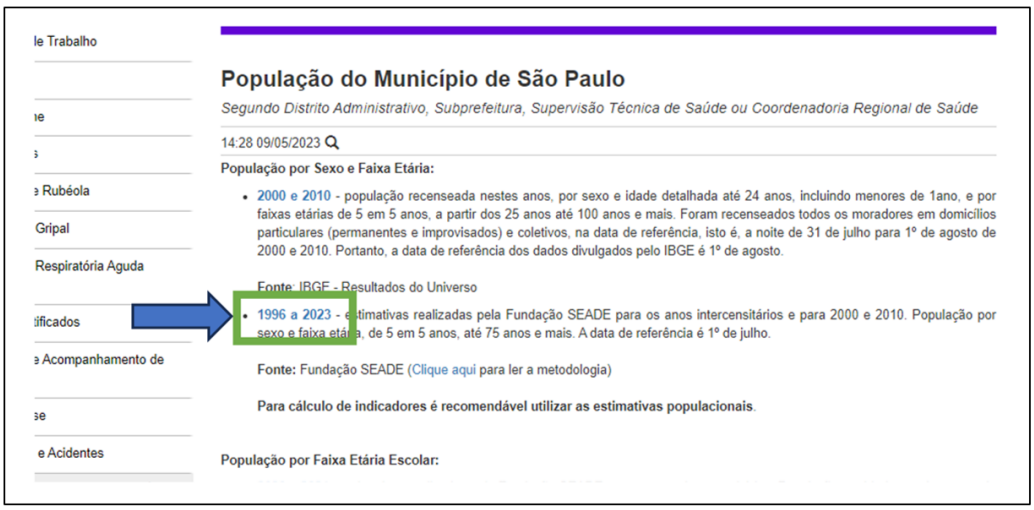
Na página inicial do site da TabNet, selecione População:



Fonte: cópia modificada da tela do TabNet.

Em seguida, selecione (clique) na população 1996 a 2023:

Tela do TabNet destacando onde clicar para baixar população do período da pesquisa



Fonte: cópia modificada da tela do Tabnet

Para montar a tabela de população, selecione os seguintes tópicos:

- Linha: ano
- Coluna: sexo
- Conteúdo: população
- Períodos disponíveis: 2013 a 2022

Tela do TabNet onde filtramos as informações para baixar a população

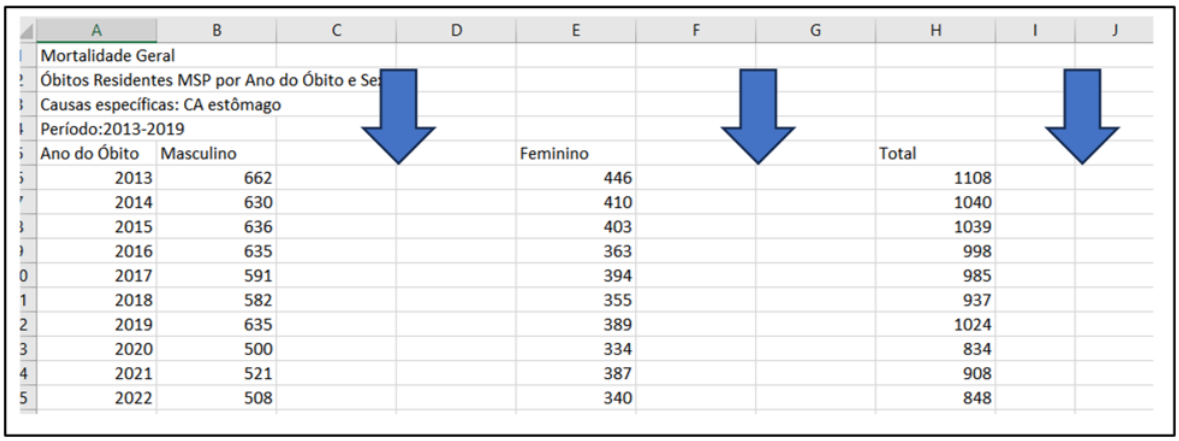
Fonte: cópia modificada da tela do TabNet.

Clique em Mostra, baixe a planilha e salve como Excel, como já explicado anteriormente.

2. Montando a tabela para cálculo dos CMEs

Para construir a tabela para calcular os CMEs por ano no arquivo Excel, acrescente duas colunas à direita de cada sexo.

Indicações de onde inserir colunas no Excel para lançar dados populacionais e calcular os CMEs

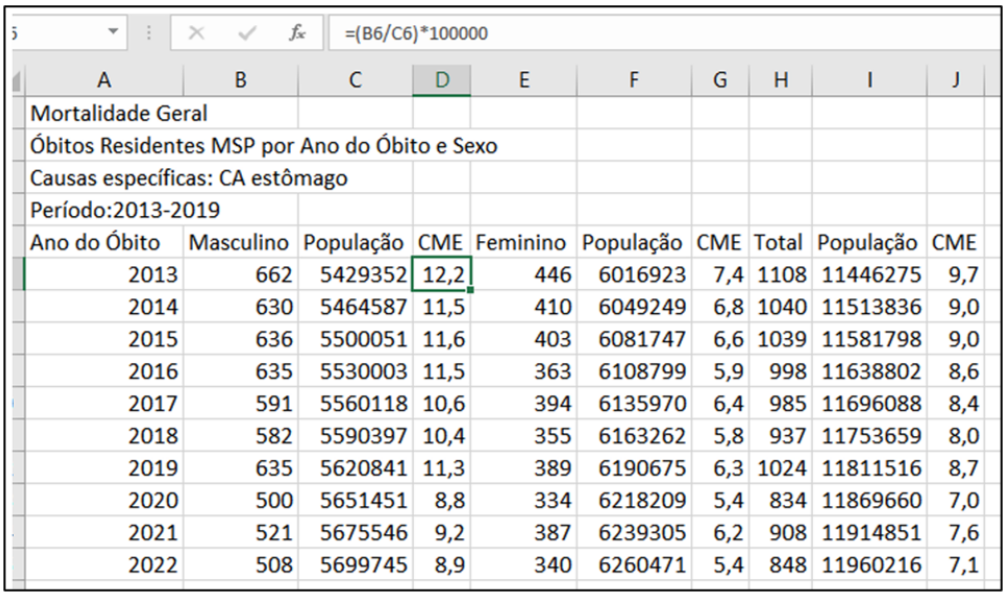


	A	B	C	D	E	F	G	H	I	J
1	Mortalidade Geral									
2	Óbitos Residentes MSP por Ano do Óbito e Se									
3	Causas específicas: CA estômago									
4	Período:2013-2019									
5	Ano do Óbito	Masculino			Feminino			Total		
6	2013	662			446			1108		
7	2014	630			410			1040		
8	2015	636			403			1039		
9	2016	635			363			998		
10	2017	591			394			985		
11	2018	582			355			937		
12	2019	635			389			1024		
13	2020	500			334			834		
14	2021	521			387			908		
15	2022	508			340			848		

Fonte: cópia da tela do Excel.

Copie a população de cada ano e cole na coluna vazia (masculino, feminino e total) e, em seguida, calcule o CME de cada ano, multiplicando por 100.000, utilizando a fórmula do Excel, conforme exemplo abaixo feito para a célula D6, onde colocamos =(B6/C6)*100000.

Tela da planilha Excel com dados populacionais e fórmula CME



=(B6/C6)*100000

	A	B	C	D	E	F	G	H	I	J
1	Mortalidade Geral									
2	Óbitos Residentes MSP por Ano do Óbito e Sexo									
3	Causas específicas: CA estômago									
4	Período:2013-2019									
5	Ano do Óbito	Masculino	População	CME	Feminino	População	CME	Total	População	CME
6	2013	662	5429352	12,2	446	6016923	7,4	1108	11446275	9,7
7	2014	630	5464587	11,5	410	6049249	6,8	1040	11513836	9,0
8	2015	636	5500051	11,6	403	6081747	6,6	1039	11581798	9,0
9	2016	635	5530003	11,5	363	6108799	5,9	998	11638802	8,6
10	2017	591	5560118	10,6	394	6135970	6,4	985	11696088	8,4
11	2018	582	5590397	10,4	355	6163262	5,8	937	11753659	8,0
12	2019	635	5620841	11,3	389	6190675	6,3	1024	11811516	8,7
13	2020	500	5651451	8,8	334	6218209	5,4	834	11869660	7,0
14	2021	521	5675546	9,2	387	6239305	6,2	908	11914851	7,6
15	2022	508	5699745	8,9	340	6260471	5,4	848	11960216	7,1

Fonte: cópia da tela do Excel.

A tabela final já formatada ficará assim:

Tela do Excel com a tabela final com CME

Ano do	Masculino			Feminino			Total		
Óbito	Óbitos	População	CME	Óbitos	População	CME	Óbitos	População	CME
2013	662	5429352	12,2	446	6016923	7,4	1108	11446275	9,7
2014	630	5464587	11,5	410	6049249	6,8	1040	11513836	9,0
2015	636	5500051	11,6	403	6081747	6,6	1039	11581798	9,0
2016	635	5530003	11,5	363	6108799	5,9	998	11638802	8,6
2017	591	5560118	10,6	394	6135970	6,4	985	11696088	8,4
2018	582	5590397	10,4	355	6163262	5,8	937	11753659	8,0
2019	635	5620841	11,3	389	6190675	6,3	1024	11811516	8,7
2020	500	5651451	8,8	334	6218209	5,4	834	11869660	7,0
2021	521	5675546	9,2	387	6239305	6,2	908	11914851	7,6
2022	508	5699745	8,9	340	6260471	5,4	848	11960216	7,1

Fonte: cópia da tela do Excel.

Notem que, para atender às normas da ABNT, ainda será necessário colocarmos o título da tabela e a fonte de informações.

2.1. Calculando e comparando as diferenças % entre 2013 e 2022

Para calcular as diferenças % entre 2013 e 2022, deve-se aplicar a seguinte fórmula: (Valor de 2022 / Valor de 2013) – 1, e depois formatar como porcentagem, como demonstrado abaixo:

Tabela do Excel mostrando como calcular as diferenças %

	A	B	C	D	E	F	G	H	I	J
5										
6	Ano do		Masculino			Feminino			Total	
7	Óbito	Óbitos	População	CME	Óbitos	População	CME	Óbitos	População	CME
8	2013	662	5429352	12,2	446	6016923	7,4	1108	11446275	9,7
9	2014	630	5464587	11,5	410	6049249	6,8	1040	11513836	9,0
10	2015	636	5500051	11,6	403	6081747	6,6	1039	11581798	9,0
11	2016	635	5530003	11,5	363	6108799	5,9	998	11638802	8,6
12	2017	591	5560118	10,6	394	6135970	6,4	985	11696088	8,4
13	2018	582	5590397	10,4	355	6163262	5,8	937	11753659	8,0
14	2019	635	5620841	11,3	389	6190675	6,3	1024	11811516	8,7
15	2020	500	5651451	8,8	334	6218209	5,4	834	11869660	7,0
16	2021	521	5675546	9,2	387	6239305	6,2	908	11914851	7,6
17	2022	508	5699745	8,9	340	6260471	5,4	848	11960216	7,1
18	Diferença 2013 e 2022	=(B17/B8)-1	5%	-27%	-24%	4%	-27%	-23%	4%	-27%

Fonte: cópia da tela do Excel.

Tabela no Excel com todos os cálculos das diferenças %

Ano do Óbito	Masculino			Feminino			Total		
	Óbitos	População	CME	Óbitos	População	CME	Óbitos	População	CME
2013	662	5429352	12,2	446	6016923	7,4	1108	11446275	9,7
2014	630	5464587	11,5	410	6049249	6,8	1040	11513836	9,0
2015	636	5500051	11,6	403	6081747	6,6	1039	11581798	9,0
2016	635	5530003	11,5	363	6108799	5,9	998	11638802	8,6
2017	591	5560118	10,6	394	6135970	6,4	985	11696088	8,4
2018	582	5590397	10,4	355	6163262	5,8	937	11753659	8,0
2019	635	5620841	11,3	389	6190675	6,3	1024	11811516	8,7
2020	500	5651451	8,8	334	6218209	5,4	834	11869660	7,0
2021	521	5675546	9,2	387	6239305	6,2	908	11914851	7,6
2022	508	5699745	8,9	340	6260471	5,4	848	11960216	7,1
Diferença 2013 e 2022	-23%	5%	-27%	-24%	4%	-27%	-23%	4%	-27%

Fonte: cópia da tela do Excel.

Percebam que, comparando as diferenças % de 2022 com 2013, concluímos que:

- A população masculina aumentou um pouco a mais que a feminina (5% *versus* 4%);
- O CME masculino (-26,9%) e o feminino são praticamente idênticos (-26,7%);
- Portanto, o CME do total é igual a -26,8% (=26%).

Olhando agora o nº absoluto de óbitos e de população, podemos concluir que:

- Todos os anos morreram mais homens que mulheres devido ao CA de estômago;
- Em 2016, o nº de óbitos masculinos foi 75% maior que o feminino (635/363-1=0,75=75%), sendo esta a **maior diferença** de todos esses anos;
- Em 2021, tivemos a **menor diferença** com 35% mais óbitos de homens do que de mulheres (521/387 -1 = 0,35);

- Calculando para cada ano o [(nº de óbitos masc / nº de óbitos fem) -1] e tirando a média dessas %, teremos que, em média, houve 55% mais óbitos de homens do que de mulheres (deixamos para você fazer esses cálculos com o Excel);
- A proporção da população de homens e mulheres praticamente se manteve constante de 2013 a 2022, com as mulheres representando entre 52% e 53% do total do MSP (chequem com o Excel fazendo para ano [população de mulheres/população total]*100 ou formatando como %).

2.2. Calculando e comparando as medidas estatísticas

Para calcular as médias, utiliza-se a fórmula =**MÉDIA(valores)**, como exemplificado abaixo:

Tabela do Excel mostrando como calcular as médias aritméticas usando a função MÉDIA

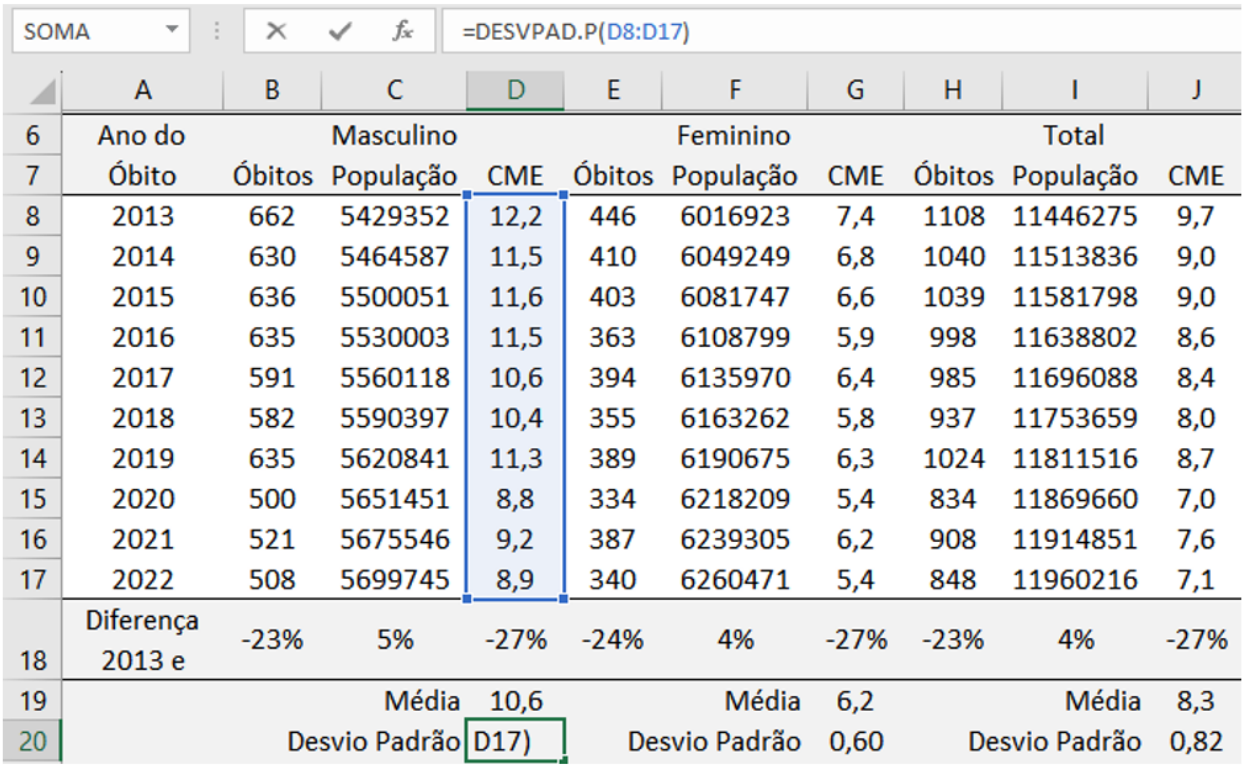
SOMA | =MÉDIA(D8:D17)

	A	B	C	D	E	F	G	H	I	J
6	Ano do		Masculino			Feminino			Total	
7	Óbito	Óbitos	População	CME	Óbitos	População	CME	Óbitos	População	CME
8	2013	662	5429352	12,2	446	6016923	7,4	1108	11446275	9,7
9	2014	630	5464587	11,5	410	6049249	6,8	1040	11513836	9,0
10	2015	636	5500051	11,6	403	6081747	6,6	1039	11581798	9,0
11	2016	635	5530003	11,5	363	6108799	5,9	998	11638802	8,6
12	2017	591	5560118	10,6	394	6135970	6,4	985	11696088	8,4
13	2018	582	5590397	10,4	355	6163262	5,8	937	11753659	8,0
14	2019	635	5620841	11,3	389	6190675	6,3	1024	11811516	8,7
15	2020	500	5651451	8,8	334	6218209	5,4	834	11869660	7,0
16	2021	521	5675546	9,2	387	6239305	6,2	908	11914851	7,6
17	2022	508	5699745	8,9	340	6260471	5,4	848	11960216	7,1
18	Diferença 2013 e	-23%	5%	-27%	-24%	4%	-27%	-23%	4%	-27%
19			Média	D17)		Média	6,2		Média	8,3

Fonte: cópia da tela do Excel.

Para calcular o desvio-padrão populacional, utiliza-se a fórmula = DESVPAD.P(valores), como exemplificado abaixo na célula D20:

Tabela do Excel mostrando como calcular os desvios-padrões usando a função DESVPAD.P

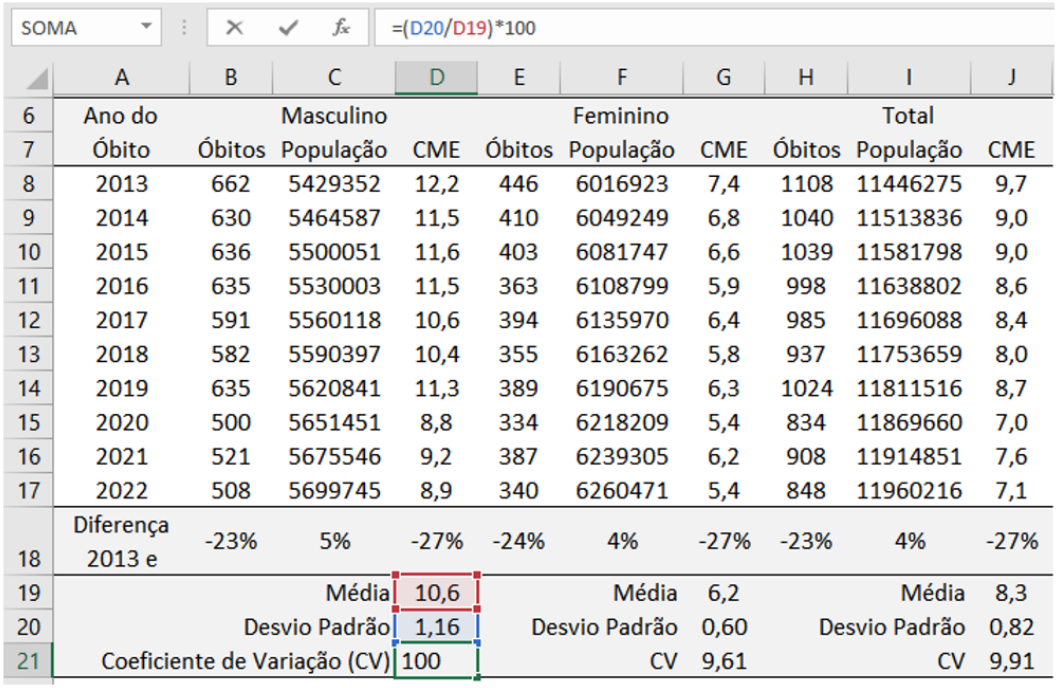


	A	B	C	D	E	F	G	H	I	J
6	Ano do		Masculino			Feminino			Total	
7	Óbito	Óbitos	População	CME	Óbitos	População	CME	Óbitos	População	CME
8	2013	662	5429352	12,2	446	6016923	7,4	1108	11446275	9,7
9	2014	630	5464587	11,5	410	6049249	6,8	1040	11513836	9,0
10	2015	636	5500051	11,6	403	6081747	6,6	1039	11581798	9,0
11	2016	635	5530003	11,5	363	6108799	5,9	998	11638802	8,6
12	2017	591	5560118	10,6	394	6135970	6,4	985	11696088	8,4
13	2018	582	5590397	10,4	355	6163262	5,8	937	11753659	8,0
14	2019	635	5620841	11,3	389	6190675	6,3	1024	11811516	8,7
15	2020	500	5651451	8,8	334	6218209	5,4	834	11869660	7,0
16	2021	521	5675546	9,2	387	6239305	6,2	908	11914851	7,6
17	2022	508	5699745	8,9	340	6260471	5,4	848	11960216	7,1
18	Diferença 2013 e	-23%	5%	-27%	-24%	4%	-27%	-23%	4%	-27%
19			Média	10,6		Média	6,2		Média	8,3
20			Desvio Padrão	D17)		Desvio Padrão	0,60		Desvio Padrão	0,82

Fonte: cópia da tela do Excel.

Para calcular o coeficiente de variação, utiliza-se a fórmula = (desvio-padrão / média)*100, como exemplificado abaixo na célula D21:

Tabela do Excel mostrando como calcular o coeficiente de variação (CV)



	A	B	C	D	E	F	G	H	I	J
6	Ano do		Masculino			Feminino			Total	
7	Óbito	Óbitos	População	CME	Óbitos	População	CME	Óbitos	População	CME
8	2013	662	5429352	12,2	446	6016923	7,4	1108	11446275	9,7
9	2014	630	5464587	11,5	410	6049249	6,8	1040	11513836	9,0
10	2015	636	5500051	11,6	403	6081747	6,6	1039	11581798	9,0
11	2016	635	5530003	11,5	363	6108799	5,9	998	11638802	8,6
12	2017	591	5560118	10,6	394	6135970	6,4	985	11696088	8,4
13	2018	582	5590397	10,4	355	6163262	5,8	937	11753659	8,0
14	2019	635	5620841	11,3	389	6190675	6,3	1024	11811516	8,7
15	2020	500	5651451	8,8	334	6218209	5,4	834	11869660	7,0
16	2021	521	5675546	9,2	387	6239305	6,2	908	11914851	7,6
17	2022	508	5699745	8,9	340	6260471	5,4	848	11960216	7,1
18	Diferença 2013 e	-23%	5%	-27%	-24%	4%	-27%	-23%	4%	-27%
19			Média	10,6		Média	6,2		Média	8,3
20			Desvio Padrão	1,16		Desvio Padrão	0,60		Desvio Padrão	0,82
21			Coeficiente de Variação (CV)	100		CV	9,61		CV	9,91

Fonte: cópia da tela do Excel.

A tabela do CME devido ao CA de estômago pronta fica assim:

CME (por 100 mil) devido a CA estômago de residentes no MSP de 2013 a 2022

	Masculino			Feminino			Total		
Ano do Óbito	Óbitos	População	CME Masc	Óbitos	População	CME Fem	Óbitos	População	CME Geral
2013	662	5429352	12,2	446	6016923	7,4	1108	11446275	9,7
2014	630	5464587	11,5	410	6049249	6,8	1040	11513836	9,0
2015	636	5500051	11,6	403	6081747	6,6	1039	11581798	9,0
2016	635	5530003	11,5	363	6108799	5,9	998	11638802	8,6
2017	591	5560118	10,6	394	6135970	6,4	985	11696088	8,4
2018	582	5590397	10,4	355	6163262	5,8	937	11753659	8,0
2019	635	5620841	11,3	389	6190675	6,3	1024	11811516	8,7
2020	500	5651451	8,8	334	6218209	5,4	834	11869660	7,0
2021	521	5675546	9,2	387	6239305	6,2	908	11914851	7,6
2022	508	5699745	8,9	340	6260471	5,4	848	11960216	7,1
Diferença %			-26,9%			-26,7%			-26,8%
		Média	10,6			6,2			8,3
		Desvio Padrão	1,16			0,60			0,82
		Coeficiente de variação (CV)	10,98%			9,61%			9,91%

Fonte: cópia da tela do Excel.

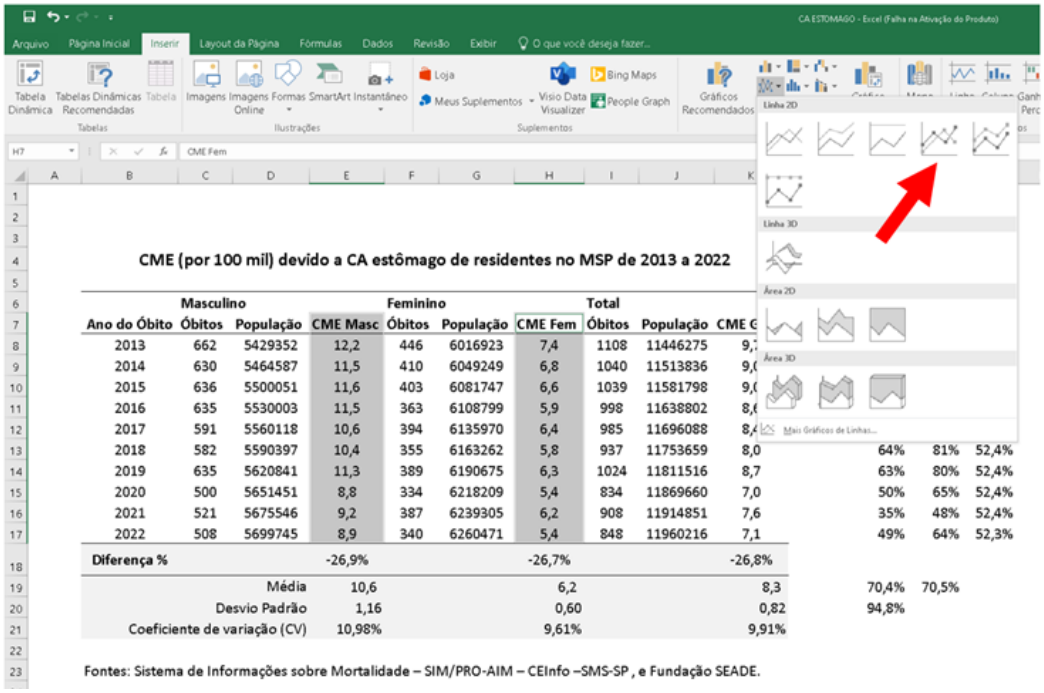
Percebam a presença obrigatória do título e da fonte dos dados. Notem, também, que as laterais da tabela não devem ser fechadas com bordas, e que o corpo da tabela deve estar em preto e branco.

Quanto à análise das medidas estatísticas calculadas, podemos concluir que:

- Em média, o CME masculino é 71% maior que o feminino (10,6/6,2-1 = 0,709... = 71%);
- A dispersão absoluta (o desvio-padrão) dos homens (1,16) é 93% maior que das mulheres (0,6); mas, como sabemos que não devemos comparar dispersões de média diferentes (10,6 *versus* 6,2) usando o desvio-padrão (ver quadro 3), o correto é comparar os CVs.
- A dispersão relativa (o coeficiente de variação) dos homens (10,98% =11%) é bem próximo do CV das mulheres (9,61% = 10%). Como esses valores de CV estão próximos de 10%, teremos gráficos de CME sem grandes dispersões dos dados em relação às suas respectivas médias de CME.

2.3. Montando os gráficos de linha dos CMEs

Para montar o gráfico de linhas dos CME masculino e feminino, primeiro, segurando CTRL, selecionamos as colunas CME Masc e CME Fem, depois clicamos em **Inserir** e selecionamos o **gráfico de linhas com marcador:**



CME (por 100 mil) devido a CA estômago de residentes no MSP de 2013 a 2022

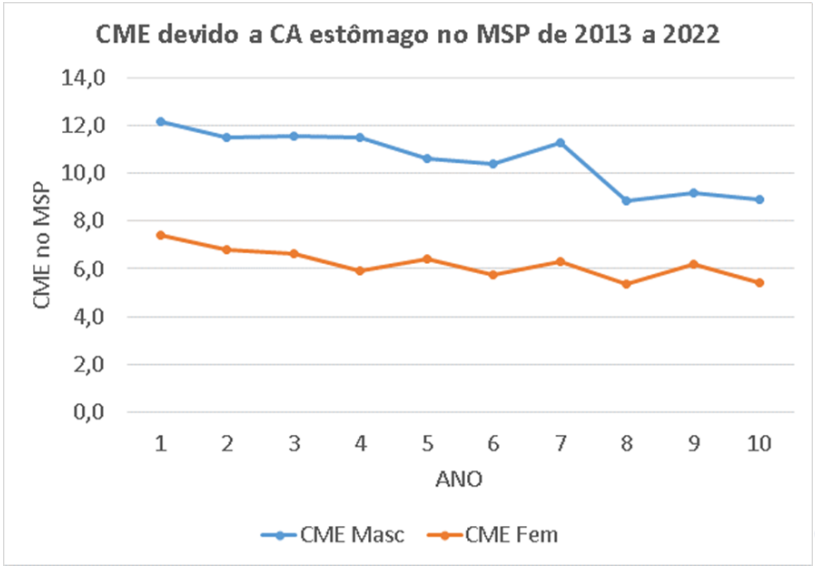
	Masculino			Feminino			Total					
Ano do Óbito	Óbitos	População	CME Masc	Óbitos	População	CME Fem	Óbitos	População	CME G			
2013	662	5429352	12,2	446	6016923	7,4	1108	11446275	9,7			
2014	630	5464587	11,5	410	6049249	6,8	1040	11513836	9,0			
2015	636	5500051	11,6	403	6081747	6,6	1039	11581798	9,0			
2016	635	5530003	11,5	363	6108799	5,9	998	11638802	8,6			
2017	591	5560118	10,6	394	6135970	6,4	985	11696088	8,4			
2018	582	5590397	10,4	355	6163262	5,8	937	11753659	8,0	64%	81%	52,4%
2019	635	5620841	11,3	389	6190675	6,3	1024	11811516	8,7	63%	80%	52,4%
2020	500	5651451	8,8	334	6218209	5,4	834	11869660	7,0	50%	65%	52,4%
2021	521	5675546	9,2	387	6239305	6,2	908	11914851	7,6	35%	48%	52,4%
2022	508	5699745	8,9	340	6260471	5,4	848	11960216	7,1	49%	64%	52,3%
Diferença %			-26,9%			-26,7%			-26,8%			
		Média	10,6			6,2			8,3	70,4%	70,5%	
		Desvio Padrão	1,16			0,60			0,82	94,8%		
		Coeficiente de variação (CV)	10,98%			9,61%			9,91%			

Fontes: Sistema de Informações sobre Mortalidade – SIM/PRO-AIM – CEInfo –SMS-SP , e Fundação SEADE.

Fonte: cópia da tela do Excel.

O gráfico que obtemos já tem legenda e escala vertical (eixo Y), mas ainda precisaremos colocar os anos na escala horizontal (eixo X), o nome dos eixos e o título do gráfico.

Descrição de gráfico sem títulos

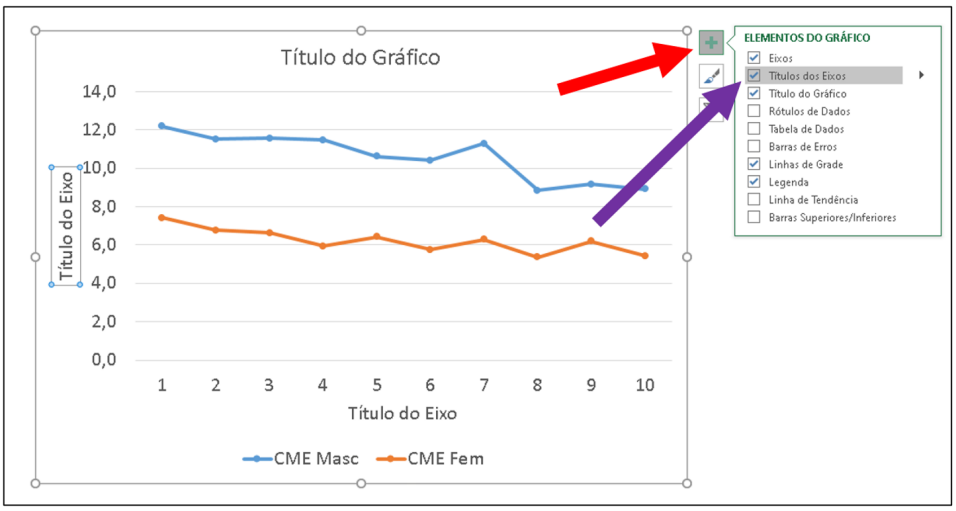


Fonte: cópia da tela do Excel.

Para inserir o título, basta clicar no Título do Gráfico no próprio gráfico no Excel e digitar o título desejado.

Para inserir os títulos dos eixos X e Y, basta clicar num espaço em branco dentro do gráfico, depois clicar no símbolo **+** que aparece à direita, e selecionar **Título dos Eixos** (ver abaixo).

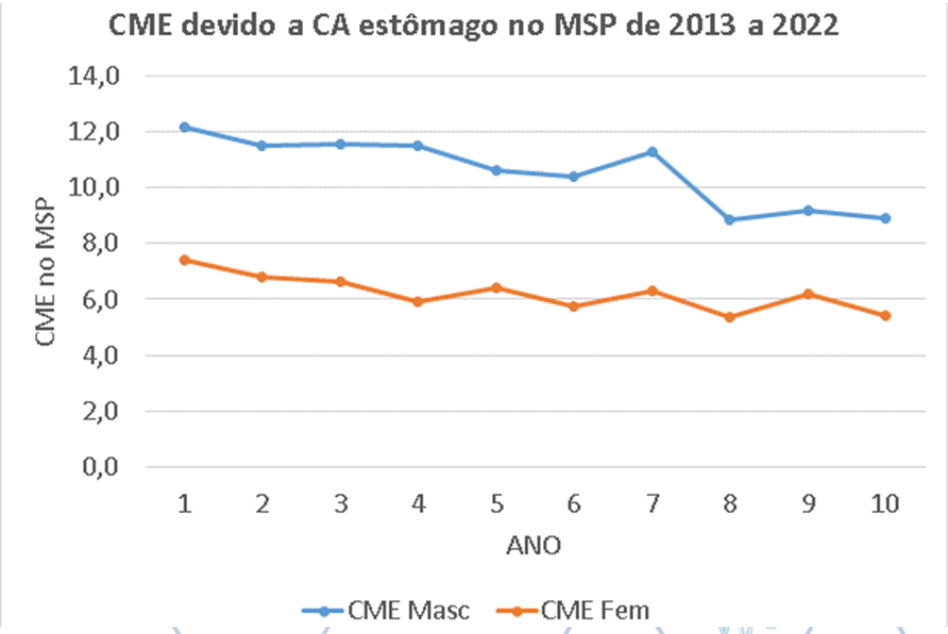
Indicação onde clicar para inserir títulos no gráfico



Fonte: cópia da tela do Excel.

Em seguida, digite os títulos dos eixos. Após inserir "ANO" no título do eixo X e digitar "CME no MSP" como título do eixo Y, obteremos o seguinte gráfico:

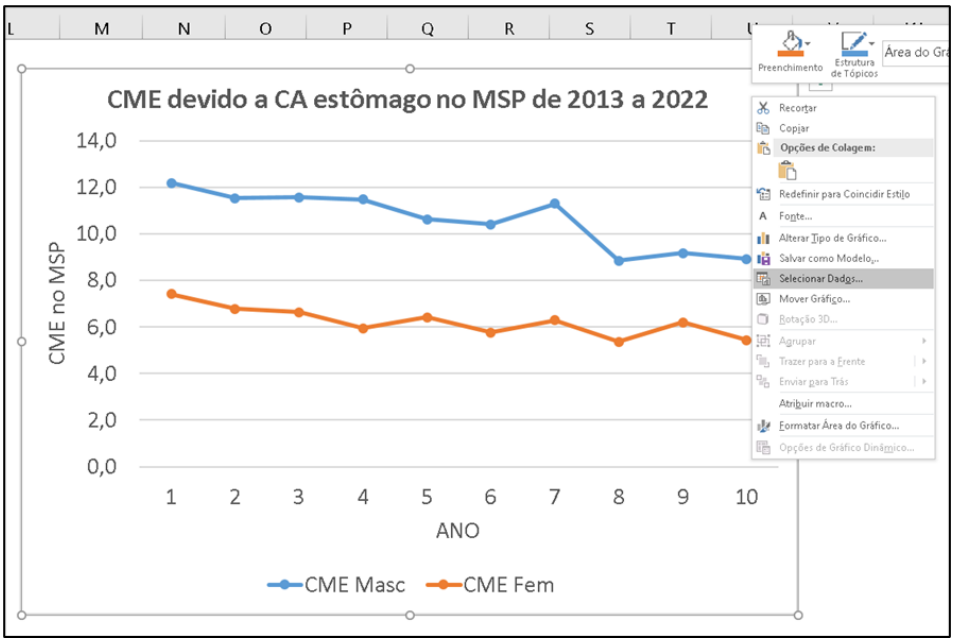
Gráfico do CME já com títulos e sem os anos



Fonte: cópia da tela do Excel.

Para inserir os anos de 2013 a 2022 no eixo X, clique com o botão direito do *mouse* em um espaço em branco dentro do gráfico, clique na janela que aparecer em Selecionar dados:

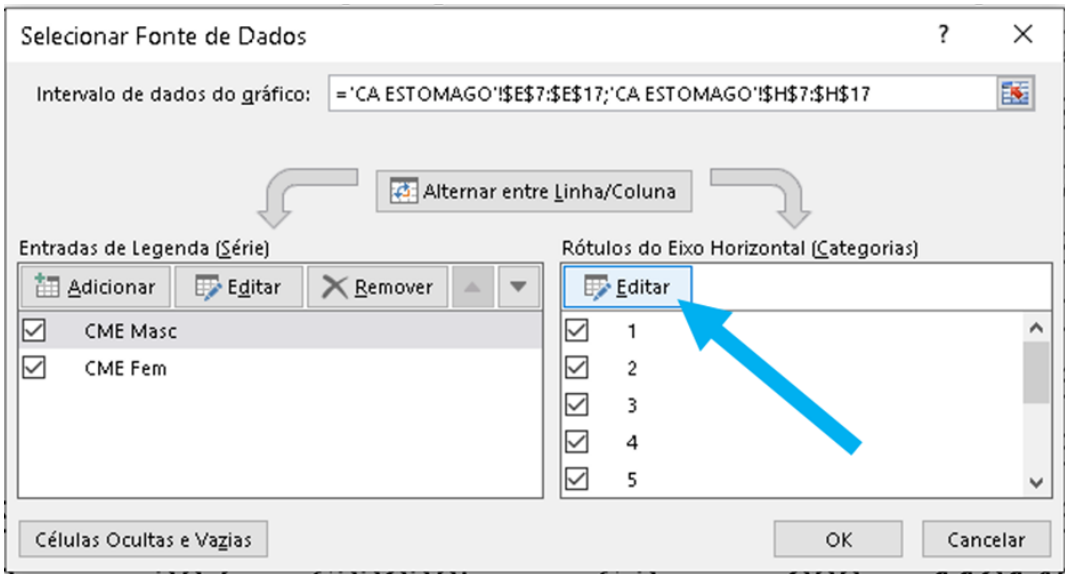
Tela do Excel indicando onde clicar para selecionar dados



Fonte: cópia da tela do Excel.

Aparecerá a janela **Selecionar Fonte de Dados**, clique no EDITAR de Rótulos do eixo horizontal (Categorias).

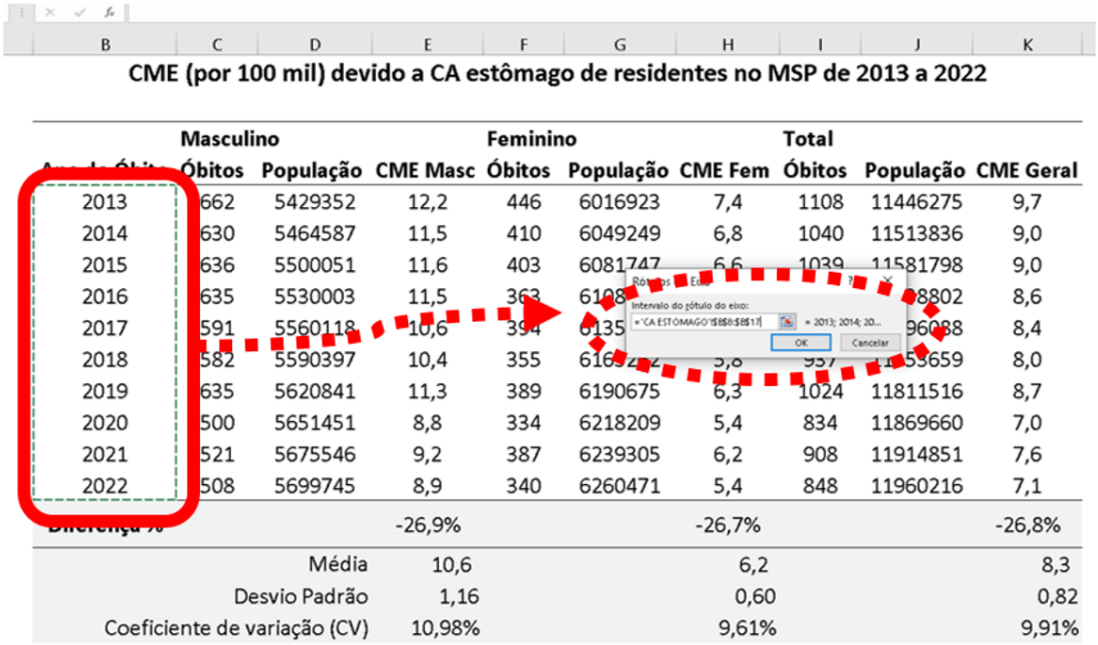
Tela do Excel indicando onde clicar para inserir anos no eixo horizontal



Fonte: cópia da tela do Excel.

Selecione os ANOS diretamente da tabela de dados criada.

Tela do Excel indicando onde selecionar os anos do eixo horizontal



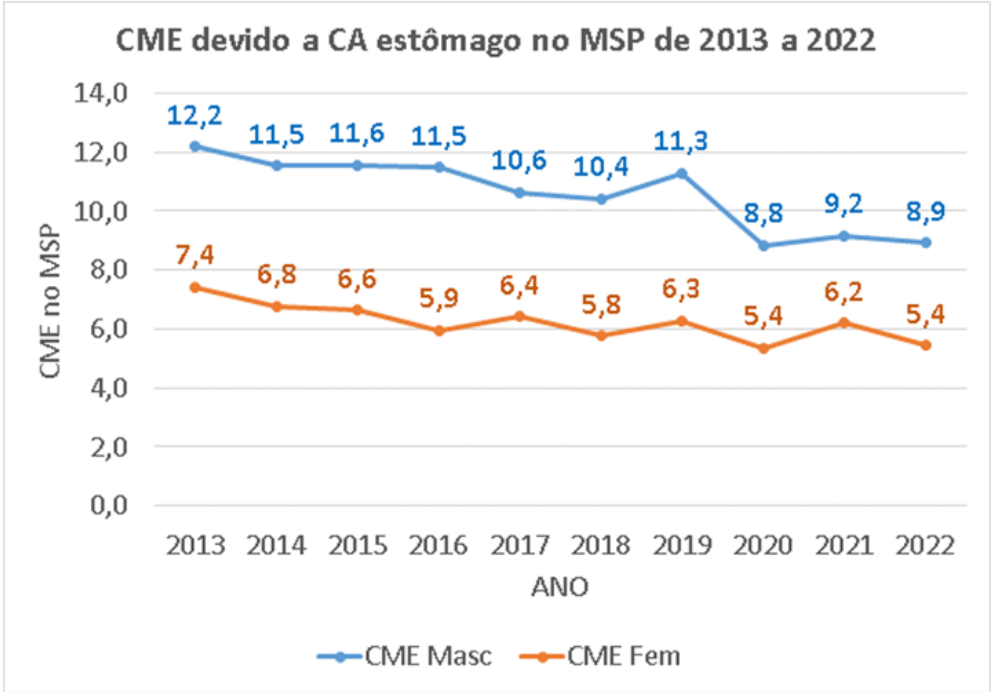
CME (por 100 mil) devido a CA estômago de residentes no MSP de 2013 a 2022

	Masculino			Feminino			Total		
Ano do Óbito	Óbitos	População	CME Masc	Óbitos	População	CME Fem	Óbitos	População	CME Geral
2013	662	5429352	12,2	446	6016923	7,4	1108	11446275	9,7
2014	630	5464587	11,5	410	6049249	6,8	1040	11513836	9,0
2015	636	5500051	11,6	403	6081747	6,6	1039	11581798	9,0
2016	635	5530003	11,5	363	6108			8802	8,6
2017	591	5560118	10,6	394	6135			96088	8,4
2018	582	5590397	10,4	355	6169242	5,8	937	11753659	8,0
2019	635	5620841	11,3	389	6190675	6,3	1024	11811516	8,7
2020	500	5651451	8,8	334	6218209	5,4	834	11869660	7,0
2021	521	5675546	9,2	387	6239305	6,2	908	11914851	7,6
2022	508	5699745	8,9	340	6260471	5,4	848	11960216	7,1
Diferença %			-26,9%			-26,7%			-26,8%
		Média	10,6			6,2			8,3
		Desvio Padrão	1,16			0,60			0,82
		Coeficiente de variação (CV)	10,98%			9,61%			9,91%

Fonte: cópia da tela do Excel.

Depois de selecionar os ANOS sem o título da coluna, clique no OK da janela, depois clique OK de novo e, pronto, os anos foram inseridos.

Tela do Excel mostrando gráfico do CME por ano completo



Fonte: cópia da tela do Excel.

Notem, no gráfico acima, que depois foram inseridos os rótulos dos dados. Para tanto, clique na linha que quer inserir os rótulos com o botão direito do *mouse* e selecione "adicionar rótulo de dados".

Depois, clique nos rótulos e formate como achar melhor.

Analisando o gráfico, concluímos que:

- Ambos os CME apresentam uma tendência de queda, porém essa tendência foi pontualmente interrompida em 2019 no CME Masc, e em 2017, 2019 e 2021 no CME Fem;
- Em 2016, houve a maior diferença % entre os CME Masc e CME Fem: [(11,5/5,9)-1]*100 = 95%;
- A partir de 2020 essa diferença entre os CME foi reduzida, já que a distância entre as linhas está menor, sendo que o CME Masc ficou maior que o CME Fem em 63% (2020), 48% (2021) e 65% (2022).

gabarito

Exercício 1

Quantitativa discreta: Idade e número de *pets*
Quantitativa contínua: Peso
Qualitativa nominal: Estado civil
Qualitativa ordinal: Nível escolar

Exercício 2

A variável do estudo é o tempo de reação, que é do tipo QUANTITATIVA CONTÍNUA, pois são valores medidos. Nesse caso, seguindo o Quadro 1, a TDF COM CLASSES com o histograma ou o polígono de frequências são os mais adequados para organizar esses dados.

$k=\sqrt{20}=4{,}47$, e podemos usar 4 ou 5 classes. Adotamos 5 classes.

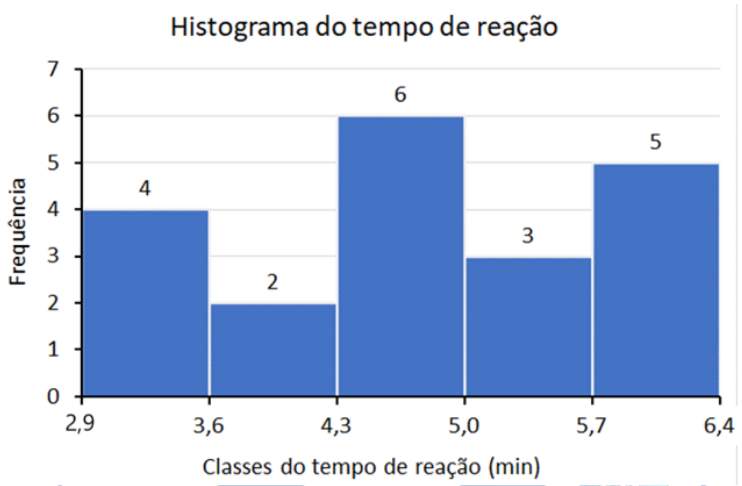
$$h = \frac{(6{,}3-2{,}9)}{5} = \frac{3{,}4}{5} = 0{,}68 \cong 0{,}7$$

Solução da TDF completa do exercício 2

TDF COM DADOS AGRUPADOS COM CLASSES

Tempo de reação	f frequência simples	fa f acumulada	fr frequência relativa	fra fr acumulada
2,9 ⊢ 3,6	4	4	20%	20%
3,6 ⊢ 4,3	2	6	10%	30%
4,3 ⊢ 5,0	6	12	30%	60%
5,0 ⊢ 5,7	3	15	15%	75%
5,7 ⊢ 6,4	5	20	25%	100%
total	20		100%	

Solução gráfica do exercício 2



Fonte: elaborado pelos autores.

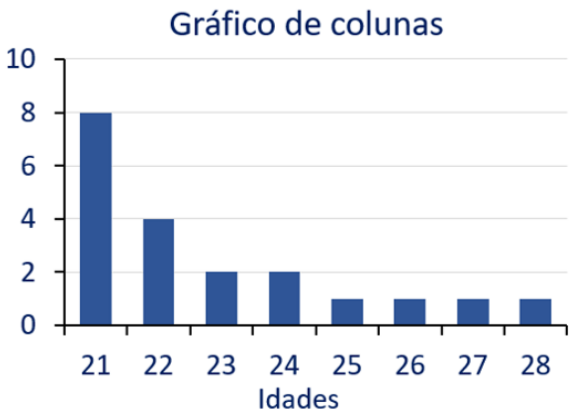
Exercício 3

A variável do estudo idade é do tipo QUANTITATIVA DISCRETA. Nesse caso, seguindo o Quadro 1, a TDF SEM CLASSES é a mais adequada para organizar esses dados.

Solução da tabela de frequência (TDF) e gráfico de colunas do exercício 3

TDF COM DADOS AGRUPADOS SEM CLASSES

Idades	f freq simples	fa f acumulada	fr freq relativa	fra fr acumulada
21	8	8	40%	40%
22	4	12	20%	60%
23	2	14	10%	70%
24	2	16	10%	80%
25	1	17	5%	85%
26	1	18	5%	90%
27	1	19	5%	95%
28	1	20	5%	100%
totais	20		100,0%	

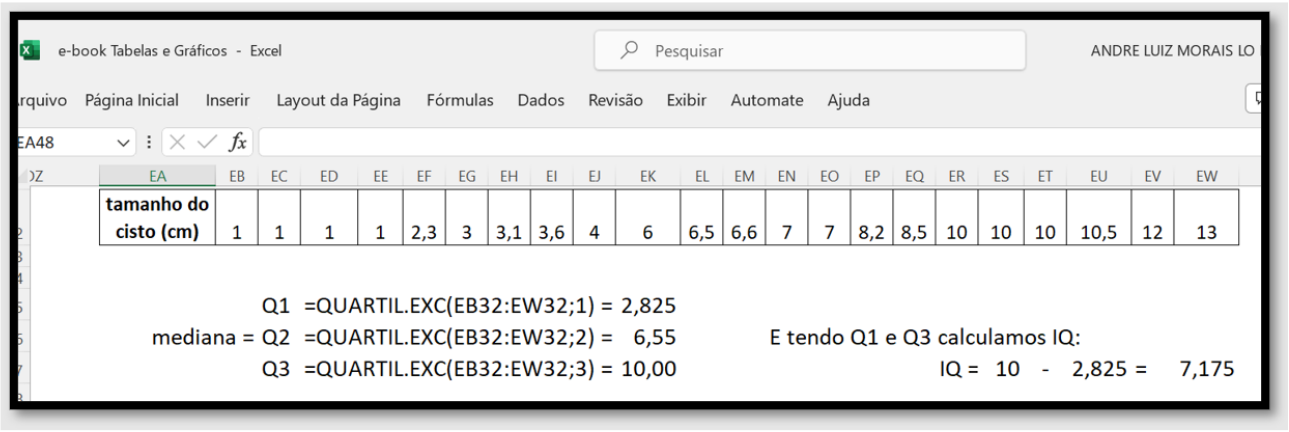


Fonte: elaborado pelos autores.

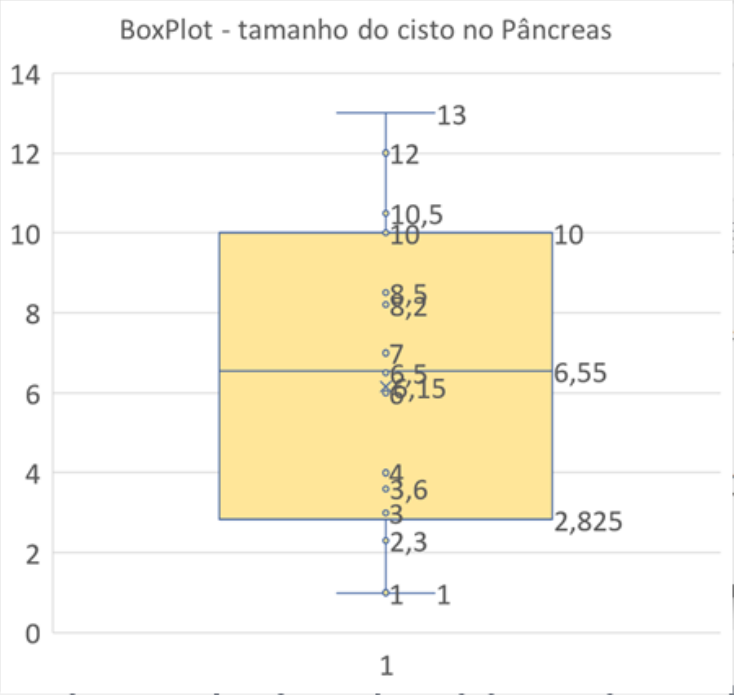
Exercício 4

Solução exercício sobre os quartis e o gráfico BoxPlot:

Tela do Excel apresentando as funções do Excel para cálculo dos quartis



Fonte: Elaborado por Lo Feudo utilizando a planilha eletrônica Excel



1) Quantos dados estão entre o Q1 e o Q3? E isso dá quantos %?

Como temos 22 valores e como os quartis separam os dados em quatro partes de mesma quantidade de valores, entre Q1 e Q3 temos 50% de 22, o que nos dá 11 valores.

2) Quando a diferença entre Q1 e Q2 é muito menor que a diferença entre Q2 e Q3, isso significa o quê? Explique.

Significa que os 25% dos valores que estão entre Q1 e Q2 são muito mais próximos entre si do que entre Q2 e Q3.

Exercício 5

a) Tempo médio:

$$\sum x = 10+15+20+9+8+3+16 = 81 \qquad \underline{x} = \frac{\sum x}{n}$$

$$\underline{x} = \frac{\sum x}{n} = \frac{10+15+20+9+8+3+16}{7} = \frac{81}{7}$$

b) Desvio-padrão:

USANDO O EXCEL ACHAMOS O “s”

= DESVPAD.A(dados) = 5,74 minutos

c) Coeficiente de variação:

$$cv = \frac{s}{x} .100$$

$$cv = \frac{5,74}{11,57} .100 = 49,6\% = 50\% > 10\% \text{ indicando alta dispersão}$$

d) Z-escore do maior e do menor valor:

Menor valor: X= 3

$$z = \frac{x - \bar{x}}{s}$$

$$z = \frac{3 - 11,57}{5,74}$$

$$z = \frac{-8,57}{5,74} \approx -1,49$$

Maior valor: X = 20

$$z = \frac{x - \bar{x}}{s}$$

$$z = \frac{20 - 11,57}{5,74}$$

$$z = \frac{-8,43}{5,74} \approx 1,47$$

Interpretação: O X = 3 está a 1,49 desvios-padrões à esquerda da média amostral. O X = 20 está a 1,47 desvios-padrões à direita da média amostral. Se o Z > 3 ou Z < -3 significa que o valor "X" é um *outlier*, ou seja, um ponto muito longe da média.

Resumo da solução do exercício 5:

Tela do Excel com resumo da solução do exercício 5

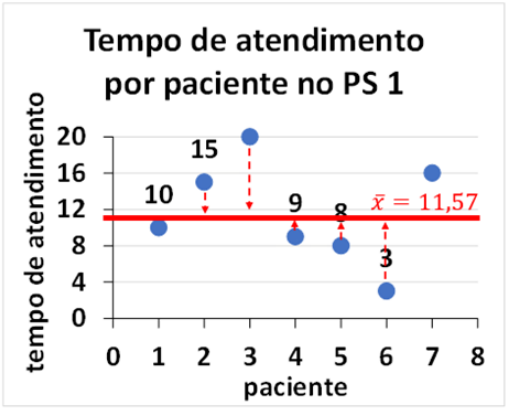
$\bar{x} = 11,57 \pm 5,74$ min

$cv = 50\%$

cv>10% indica valores ***Longe da média: alta dispersão***

$z_{x=3} \approx -1,49$

$z_{x=20} \approx 1,47$



Fonte: elaborado pelos autores.

CENTRO UNIVERSITÁRIO
SÃOCAMILO