Estratégia Concursos

Aula 00

CAESB (Analista de Suporte ao Negócio - Estatístico) Conhecimentos Específicos - 2024 (Pós-Edital)

Autor:
Equipe Exatas Estratégia Concursos

26 de Novembro de 2024

Índice

INTRODUÇÃO À ESTATÍSTICA

A origem da estatística remonta às civilizações antigas, em que vários povos coletavam e registravam dados populacionais e econômicos de interesse do Estado. Nessa época, também eram realizadas estimativas das riquezas individuais e familiares, as quais eram utilizadas para determinar o montante de impostos a serem pagos pela população.

O termo estatística se originou da palavra *status*, que significa Estado em latim. O termo era utilizado para designar um conjunto de dados, relativos aos Estados, que os governantes utilizavam com a finalidade de controle fiscal e segurança nacional. O primeiro a utilizar a palavra foi Schmeitzel, ainda no século XVII, em latim. Depois, foi adotada pelo acadêmico alemão Godofredo Achenwall.

A **Estatística** pode ser definida como **a ciência que estuda os processos de coleta, organização, análise e interpretação de dados numéricos variáveis referentes a qualquer fenômeno**. Ou ainda, podemos conceituá-la como **um conjunto de técnicas de coleta, organização, análise e interpretação de dados, aplicáveis a várias áreas do conhecimento, que auxiliam no processo de tomada de decisão.**

Os avanços computacionais tornaram a Estatística mais acessível e permitiram aplicações mais sofisticadas em diferentes áreas do conhecimento. Nesse cenário, os softwares estatísticos passaram a disponibilizar ferramentas antes inimagináveis, voltadas para planejamento de experimentos, teste de hipóteses, cálculos de confiabilidade, criação de gráficos complexos e elaboração de modelos preditivos.

A Estatística pode ser dividida em três grandes ramos: Estatística Descritiva (ou dedutiva), Estatística Probabilística e Estatística Inferencial (ou indutiva). Alguns autores, porém, consideram a Estatística Probabilística como parte da Estatística Inferencial.

A **Estatística Descritiva (ou Dedutiva)** é **responsável pela coleta, organização, descrição e resumo dos dados observados**. A partir de um determinado conjunto de dados, a Estatística Descritiva busca organizá-los em tabelas (ou gráficos) e estabelecer um sumário por meio de medidas descritivas como a média, os valores mínimo e máximo, o desvio padrão, entre outras.

A **Estatística Probabilística** é **responsável por estabelecer o modelo matemático probabilístico adotado para explicar os fenômenos aleatórios investigados pela Estatística**. Os resultados desses fenômenos aleatórios podem variar de uma observação para outra, o que dificulta muito a previsão de um resultado futuro. Por isso, a Teoria da Probabilidade é usada para medir a chance de ocorrência de determinados eventos.

A **Estatística Inferencial (ou Indutiva)** é **responsável pela análise e interpretação dos dados**. A partir da análise de dados de uma amostra, a Estatística Indutiva estabelece inferências e previsões sobre a população, auxiliando na tomada decisões. Além disso, busca generalizar conclusões a respeito da população a partir de uma amostra, analisando a representatividade, a significância e a confiabilidade dos resultados obtidos.

ESTATÍSTICA DESCRITIVA	ESTATÍSTICA PROBABILÍSTICA	ESTATÍSTICA INFERENCIAL
É responsável pela coleta, organização, descrição e resumo dos dados observados.	É responsável por estabelecer o modelo matemático adotado para explicar fenômenos aleatórios.	É responsável pela análise e interpretação dos dados.

(VUNESP/Câmara Municipal de São Carlos/2013) Trata-se da estatística que tem as atribuições de obtenção, organização, redução e representação de dados estatísticos, bem como a obtenção de algumas informações que auxiliam na descrição de um fenômeno observado. Esta estatística é denominada

a) Coletora.

b) Celetista.

c) Populacional.

d) Amostral.

e) Descritiva.

Comentários:

A Estatística está dividida em três grandes ramos, a saber:

a) **Estatística Descritiva (ou Dedutiva)** - responsável pela coleta (ou obtenção), organização, redução e representação de dados estatísticos;

b) **Estatística Probabilística** - responsável por estabelecer o modelo matemático probabilístico adotado para explicar os fenômenos aleatórios investigados pela Estatística; e

c) **Estatística Inferencial (ou Indutiva) -** responsável pela análise e interpretação dos dados.

Portanto, a alternativa correta é a letra E.

Gabarito: E.

CONCEITOS INICIAIS

Neste tópico, apresentaremos alguns conceitos iniciais da estatística que costumam ser abordados em provas de concursos públicos, dentre os quais podemos citar: população, amostra, censo, amostragem, parâmetros e estatísticas.

População

Uma POPULAÇÃO é um conjunto que contém TODOS OS INDIVÍDUOS, OBJETOS OU ELEMENTOS a serem estudados, que apresentam uma ou mais características em comum. A população pode ser finita, quando apresenta um número pequeno ou limitado de observações; ou infinita, quando apresenta um número muito grande ou ilimitado de observações.

Amostra

Uma AMOSTRA é um SUBCONJUNTO EXTRAÍDO DA POPULAÇÃO para análise, devendo ser representativo daquele grupo. A partir das informações colhidas da amostra, os resultados obtidos podem ser utilizados para generalizar, inferir ou tirar conclusões acerca da população. Como exemplo, podemos citar as pesquisas eleitorais, em que uma amostra de eleitores deve ser extraída de acordo com a proporcionalidade de gênero, idade, grau de instrução e classe social.

Censo

O CENSO, ou recenseamento, é um estudo dos dados relativos a TODOS os elementos de uma população. O censo pode custar muito caro e demandar um tempo considerável, de forma que um estudo considerando apenas uma parcela da população pode ser uma alternativa mais simples, rápida e menos onerosa. Como exemplos, podemos citar a pesquisa sobre o grau de escolaridade dos habitantes brasileiros, o estudo sobre a renda dos brasileiros e a pesquisa de emprego.

Amostragem

A AMOSTRAGEM é um processo que consiste na SELEÇÃO CRITERIOSA dos elementos a serem submetidos à investigação. Se forem cometidos erros no processo de seleção da amostra, muito provavelmente, o estudo ficará comprometido e os resultados serão tendenciosos. Portanto, devemos garantir que a amostra seja representativa da população. Isso significa que, com exceção de pequenas discrepâncias inerentes à

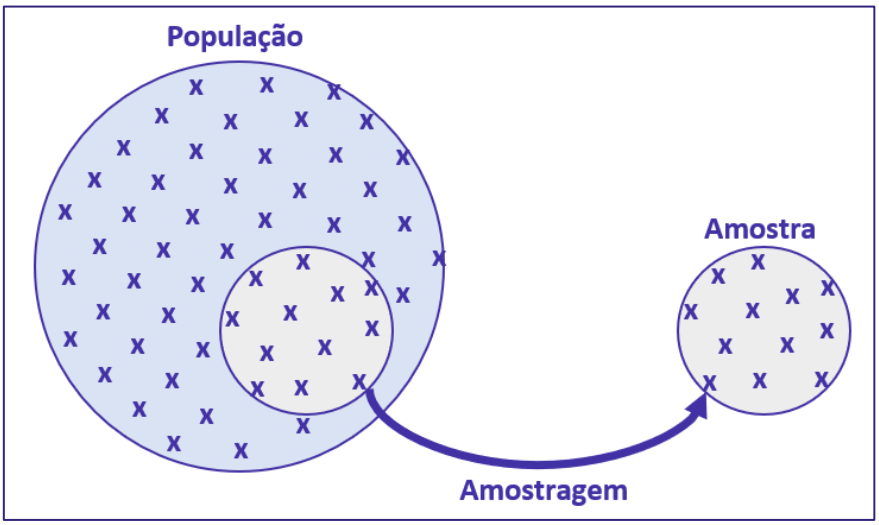
aleatoriedade existente no processo de amostragem, uma amostra deve possuir as mesmas características básicas da população, no que diz respeito às variáveis que desejamos pesquisar.

Parâmetros

Os PARÂMETROS são DESCRIÇÕES NUMÉRICAS de CARACTERÍSTICAS POPULACIONAIS que raramente são conhecidas. Em geral, é muito caro ou demorado obter os dados da população inteira. Assim, **algumas medidas precisam ser estimadas a partir de critérios ou métodos definidos pelo pesquisador**, **para representar características desconhecidas de uma população (por exemplo, a proporção de homens e mulheres na população brasileira)**. Normalmente, os parâmetros populacionais são constantes para uma população.

Estatística (ou estimador)

As ESTATÍSTICAS são MEDIDAS NUMÉRICAS OBTIDAS DE AMOSTRAS representativas extraídas da população. A partir das informações colhidas da amostra, as **estatísticas amostrais obtidas podem ser utilizadas para inferir ou tirar conclusões acerca dos parâmetros populacionais**, como a proporção de homens e mulheres na população brasileira. De forma resumida, as estatísticas (ou estimadores) são **descrições numéricas de características amostrais.** Normalmente, as estatísticas amostrais diferem de uma amostra para outra.



ESQUEMATIZANDO

POPULAÇÃO CONJUNTO de TODOS os elementos a serem estudados, que apresentam uma ou mais características em comum.	**CENSO** ESTUDO dos dados relativos a TODOS os ELEMENTOS de uma população.
AMOSTRA SUBCONJUNTO extraído DA POPULAÇÃO para análise, devendo ser representativo daquele grupo	**AMOSTRAGEM** PROCESSO que consiste na seleção criteriosa dos elementos a serem submetidos à investigação.
PARÂMETROS Descrições numéricas de características da POPULAÇÃO, que normalmente precisam ser estimadas.	**ESTATÍSTICAS** Medidas numéricas extraídas de AMOSTRAS representativas extraídas da população.

HORA DE PRATICAR!

(VUNESP/Pref. Valinhos/2019) O grupo completo de unidades elementares de pessoas, objetos ou coisas é denominado, para a estatística, de

a) Amostra.

b) Unidades.

c) Censo.

d) População.

e) Variáveis.

Comentários:

Para resolvermos a questão vamos conceituar cada uma das alternativas:

- Letra A: amostra é um subconjunto de uma população. Os dados coletados e/ou selecionados de uma população estatística por um procedimento definido, denotado de amostragem.
- Letra B: unidades são elementos da amostra (unidades amostrais) ou da população (unidade populacional).
- Letra C: censo consiste na análise de todos os elementos da população.
- Letra D: população é o grupo completo de unidades elementares de pessoas, objetos ou coisas, ou seja, constitui todo o universo de informações de que se necessita.
- Letra E: variável é a característica de relevância que é medida em cada componente da amostra ou população. As variáveis podem ter valores numéricos (variáveis quantitativas) ou não numéricos (variáveis qualitativas).

Gabarito: D.

(CESPE/DEPEN/2015) O diretor de um sistema penitenciário, com o propósito de estimar o percentual de detentos que possuem filhos, entregou a um analista um cadastro com os nomes de 500 detentos da instituição para que esse profissional realizasse entrevistas com os indivíduos selecionados.

A partir dessa situação hipotética e dos múltiplos aspectos a ela relacionados, julgue o item, referente a técnicas de amostragem.

A diferença entre um censo e uma amostra consiste no fato de esta última exigir a realização de um número maior de entrevistas.

Comentários:

Um censo é a análise de todos os elementos de determinada população. Trabalhamos com uma amostra quando obtemos informações de apenas uma parte dessa população. Portanto, o censo requer a realização de um número maior de entrevistas.

Gabarito: Errado.

MÉTODO EXPERIMENTAL X MÉTODO ESTATÍSTICO

Para a investigação de um fenômeno, temos a nossa disposição dois métodos: **experimental** e **estatístico**. De forma resumida, o **MÉTODO EXPERIMENTAL consiste em manter constantes as causas (fatores), com exceção de uma, que é variada para que seus efeitos sejam descobertos**. Contudo, nem sempre poderemos aplicar o método experimental, pois os fatores que afetam um fenômeno podem não permanecer constantes enquanto variamos a causa que nos interessa.

Por exemplo, para analisarmos uma queda nas vendas de uma empresa nacional que produz chocolates finos, teríamos que considerar vários fatores que não necessariamente permanecerão constantes durante toda a investigação do fenômeno, tais como a região, o fluxo de turistas na localidade; a temperatura média; o preço do concorrente; o mês de férias; etc.

Assim, diante da impossibilidade de manter as causas ou fatores constantes, **o MÉTODO ESTATÍSTICO admite e registra todas as possíveis variações das causas presentes, procurando determinar a influência de cada fator no resultado final**. Dessa forma, o método estatístico busca descobrir relações entre os fatores, como, por exemplo, a influência da temperatura média e do fluxo de turistas na venda de chocolates finos.

ESQUEMATIZANDO

MÉTODO EXPERIMENTAL	MÉTODO ESTATÍSTICO
As CAUSAS são mantidas CONSTANTES, COM EXCEÇÃO DE UMA, que é VARIADA para que seus efeitos sejam descobertos.	Admite e REGISTRA TODAS AS POSSÍVEIS VARIAÇÕES DAS CAUSAS PRESENTES, procurando determinar a influência de cada fator no resultado.

DADOS ESTATÍSTICOS

Os **dados estatísticos** constituem os valores resultantes da **coleta de dados**. Os dados referem-se a um **conjunto de valores**, os quais são organizados por meio de **variáveis** (a característica está sendo medida) e **observações** (elementos da amostra/população). É o caso, por exemplo, dos valores obtidos na pesquisa de peso, altura, idade e sexo de uma determinada amostra de indivíduos/população.

Com relação ao número de observações coletadas, os dados são classificados em **univariados**, **bivariados** ou **multivariados**:

a) **dados univariados**: quando uma única observação de cada indivíduo é registrada. Por exemplo: peso;
b) **dados bivariados**: quando duas observações de cada indivíduo são registradas. Por exemplo: peso e altura;
c) **dados multivariados**: quando mais duas observações acerca de cada indivíduo são registradas. Por exemplo: peso, altura, sexo e idade.

Quanto à forma de apresentação, os dados podem ser classificados em **dados brutos** ou **rol**.

Dados Brutos

Os **dados brutos** são aqueles que **não foram numericamente organizados em ordem crescente ou decrescente**, ou seja, **estão na forma como foram coletados**. Como exemplo de dados brutos, podemos citar uma relação dos tempos médios de estudo diário, em minutos, de 50 alunos do Estratégia, em que a seleção dos alunos ocorreu de forma aleatória, não havendo qualquer ordenação de valores.

Aluno	Tempo	Aluno	Tempo	Aluno	Tempo	Aluno	Tempo	Aluno	Tempo
1	**143**	11	**113**	21	**170**	31	**124**	41	**105**
2	**142**	12	**143**	22	**158**	32	**137**	42	**154**
3	**161**	13	**159**	23	**123**	33	**153**	43	**99**
4	**126**	14	**168**	24	**96**	34	**129**	44	**114**
5	**134**	15	**123**	25	**98**	35	**148**	45	**161**
6	**137**	16	**135**	26	**135**	36	**173**	46	**128**
7	**171**	17	**135**	27	**129**	37	**126**	47	**175**
8	**85**	18	**175**	28	**126**	38	**104**	48	**137**
9	**155**	19	**115**	29	**103**	39	**157**	49	**165**
10	**171**	20	**89**	30	**171**	40	**127**	50	**115**

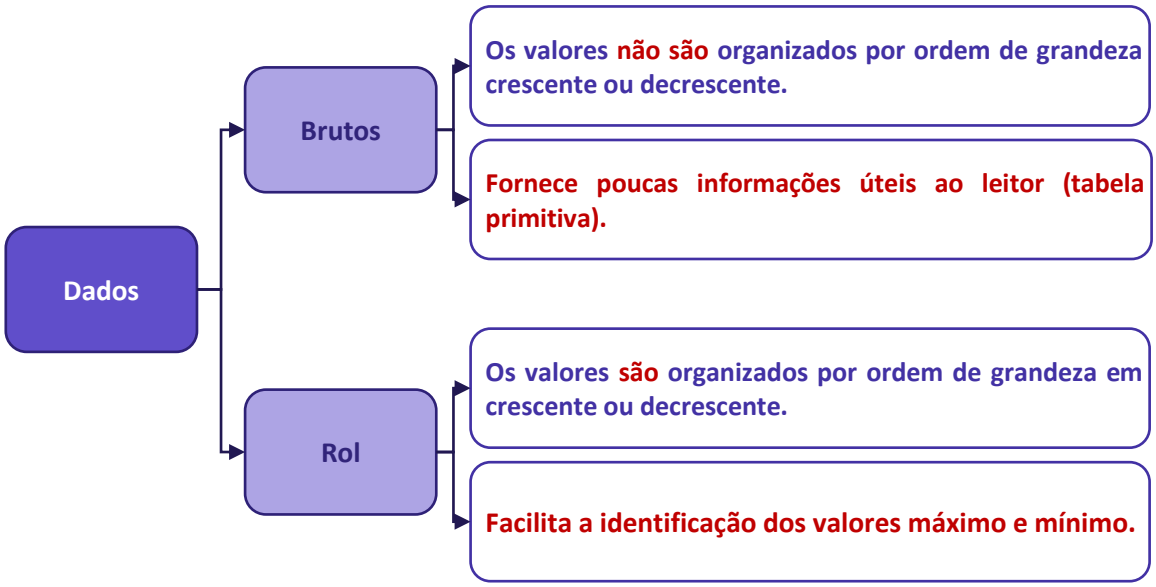
Esse tipo de tabela, em que os elementos não aparecem numericamente ordenados, é denominada de **tabela primitiva**. A **tabela primitiva**, em geral, **oferece pouca ou nenhuma informação ao leitor**, sendo necessário haver uma organização dos dados, a fim de torná-los mais expressivos.

Rol

O **rol** é a organização dos dados brutos em ordem de grandeza crescente ou decrescente. Com os dados organizados em rol, podemos saber, com facilidade, qual o menor e o maior elemento de um conjunto de dados. Os dados do nosso exemplo, isto é, os tempos médios de estudo diário, podem ser organizados em ordem crescente ou decrescente:

Rol (em ordem crescente)				
85	115	129	143	161
89	115	129	143	165
96	123	134	148	168
98	123	135	153	170
99	124	135	154	171
103	126	135	155	171
104	126	137	157	171
105	126	137	158	173
113	127	137	159	175
114	128	142	161	**175**

Rol (em ordem decrescente)				
175	161	142	128	114
175	159	137	127	113
173	158	137	126	105
171	157	137	126	104
171	155	135	126	103
171	154	135	124	99
170	153	135	123	98
168	148	134	123	96
165	143	129	115	89
161	143	129	115	**85**



VARIÁVEIS ESTATÍSTICAS

A **variável estatística** consiste no **conjunto de características que desejamos averiguar estatisticamente**. Ela também pode ser definida como o **objeto da pesquisa estatística**. Por exemplo, se nosso interesse é conhecer quantas horas os alunos do Estratégia estudam diariamente, então nossa variável é o número de horas estudadas por dia. As variáveis estatísticas podem ser classificadas, inicialmente, em duas categorias: **qualitativas** e **quantitativas**.

Variáveis Qualitativas

As **variáveis qualitativas** são as características que **não podem ser descritas de forma numérica**, mas que podem ser definidas por meio de **qualidades** (**atributos** ou **categorias**) **do indivíduo pesquisado**. Elas podem ser classificadas em **nominais** ou **ordinais**:

a) **variável qualitativa nominal (ou categórica)**, as possíveis categorias **não podem** ser ordenadas. Por exemplo, a cor dos olhos dos moradores de uma determinada cidade (pretos, castanhos, azuis e verdes);

b) **variável qualitativa ordinal**, as possíveis categorias **podem** ser ordenadas de alguma forma. Por exemplo, o grau de instrução dos funcionários de um determinado órgão (fundamental, médio, superior).

Variáveis Quantitativas

As **variáveis quantitativas** são características que podem ser descritas em termos de quantidades (valores numéricos), obtidas por meio de contagem ou mensuração. Elas podem ser classificadas em **discretas** e **contínuas**:

a) **variáveis quantitativas discretas**, os possíveis valores formam um conjunto finito ou enumerável de números e, geralmente, **resultam de um processo de contagem**. O número de ocorrências da característica em análise pode ser contado. Por exemplo, o número de leitos abertos em hospitais de uma determinada cidade;

b) **variáveis quantitativas contínuas**, os possíveis valores formam um intervalo de números reais e, normalmente, **resultam de um processo de mensuração**. A característica pode ser medida em uma escala contínua, a qual podem ser associados um número infinito de possíveis valores, de modo a não haver lacunas ou interrupções. Por exemplo, a altura dos moradores de uma determinada cidade.

- **Variáveis**
  - **Qualitativas**
    - **Nominais**
      - Os resultados obtidos não podem ser ordenados/hierarquizados.
      - Ex: cor dos olhos; esporte praticado.
    - **Ordinais**
      - Os resultados obtidos podem ser ordenados/hierarquizados.
      - Ex: nível de escolaridade.
  - **Quantitativas**
    - **Discretas**
      - Os possíveis valores formam um conjunto finito ou enumerável; resultam de contagem.
      - Ex: número de leitos por cidade; idade.
    - **Contínuas**
      - Os possíveis valores formam um intervalo de números reais; resultam de mensuração.
      - Ex: peso; altura.

HORA DE PRATICAR!

(VUNESP/CM São Joaquim Barra/2018) A Estatística descritiva faz uso de variáveis, que são classificadas como quantitativas ou qualitativas. Assinale a alternativa correta em relação a essas variáveis.

a) Quantitativas referem-se às variáveis ordinal ou discreta; as qualitativas referem-se às variáveis nominal ou contínua.

b) Quantitativas referem-se às variáveis ordinal ou contínua; as qualitativas referem-se às variáveis nominal ou discreta.

c) Quantitativas referem-se às variáveis nominal ou contínua; as qualitativas referem-se às variáveis discreta ou ordinal.

d) Quantitativas referem-se às variáveis contínua ou discreta; as qualitativas referem-se às variáveis nominal ou ordinal.

e) Quantitativas referem-se às variáveis nominal ou ordinal; as qualitativas referem-se às variáveis contínua ou discreta.

Comentários:

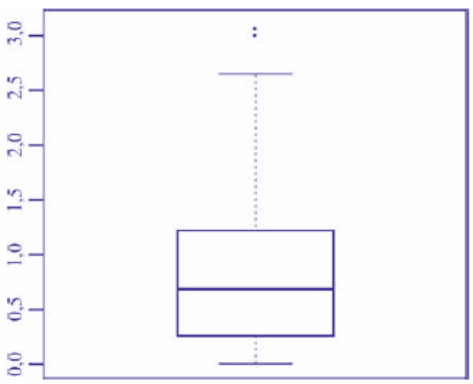
As variáveis podem ser classificadas em:

a) variáveis quantitativas: são as características que podem ser medidas em uma escala quantitativa, isto é, numérica. Elas podem ser **contínuas** ou **discretas**.

b) variáveis qualitativas (ou categóricas): são as características que não possuem valores quantitativos. Nesse caso, são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos. Elas podem ser **nominais** ou **ordinais**.

Gabarito: D.

(CESPE/TCE-PA/2016)



média amostral	0,80
desvio padrão amostral	0,70
primeiro quartil	0,25
mediana	0,70
terceiro quartil	1,20
mínimo	0
máximo	3,10

Um indicador de desempenho X permite avaliar a qualidade dos processos de governança de instituições públicas. A figura mostra, esquematicamente, a sua distribuição, obtida mediante estudo amostral feito por determinada agência de pesquisa. A tabela apresenta estatísticas descritivas referentes a essa distribuição.

Com base nessas informações, julgue o item a seguir.

X representa uma variável qualitativa ordinal.

Comentários:

As variáveis podem ser classificadas em:

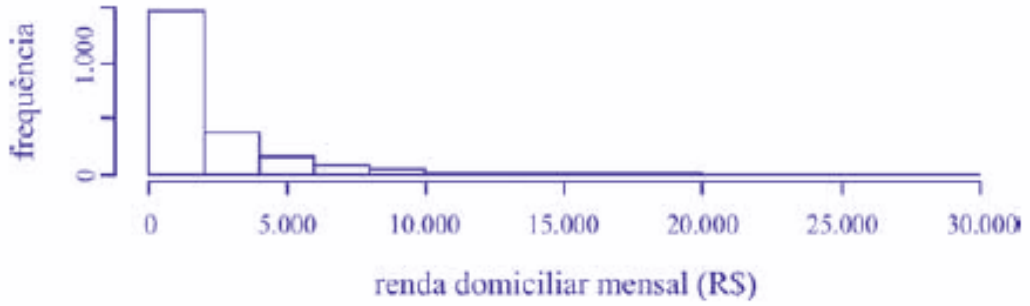
a) variáveis quantitativas: são as características que podem ser representadas numericamente, isto é, podem ser medidas em uma escala quantitativa, podendo ser contínuas ou discretas.

b) variáveis qualitativas (ou categóricas): são as características que não podem ser expressas numericamente, ou seja, não possuem valores quantitativos. Elas são definidas por categorias que classificam os indivíduos, podendo ser nominais ou ordinais.

Analisando as informações dadas na questão, percebemos que X pode adotar valores que vão de 0 (mínimo) a 3,10 (máximo). Logo, como X assume valores numéricos, dizemos que ela é uma variável quantitativa.

Gabarito: Errado.

(CESPE/TELEBRAS/2015)



Uma empresa coletou e armazenou em um banco de dados diversas informações sobre seus clientes, entre as quais estavam o valor da última fatura vencida e o pagamento ou não dessa fatura. Analisando essas informações, a empresa concluiu que 15% de seus clientes estavam inadimplentes. A empresa recolheu ainda dados como a unidade da Federação (UF) e o CEP da localidade em que estão os clientes. Do conjunto de todos os clientes, uma amostra aleatória simples constituída por 2.175 indivíduos prestou também informações sobre sua renda domiciliar mensal, o que gerou o histograma apresentado. Com base nessas informações e no histograma, julgue o item a seguir.

O CEP da localidade dos clientes e o valor da última fatura vencida são variáveis quantitativas.

Comentários:

O CEP, embora possua uma representação numérica, não exprime quantidades e, portanto, não é uma variável quantitativa. Esse código número apenas qualifica as ruas de um determinado endereço, podendo ser classificada como uma variável qualitativa.

O valor da última fatura vencida, por sua vez, é sim uma variável quantitativa.

Gabarito: Errado.

(CESPE/TELEBRAS/2015) Roberto comprou, por R$ 2.800,00, rodas de liga leve para seu carro, e, ao estacionar no shopping, ficou indeciso sobre onde deixar o carro, pois, caso o coloque no estacionamento público, correrá o risco de lhe roubarem as rodas, ao passo que, caso o coloque no estacionamento privado, terá de pagar R$ 70,00, com a garantia de que eventuais prejuízos serão ressarcidos pela empresa administradora.

Considerando que p seja a probabilidade de as rodas serem roubadas no estacionamento público, que X seja a variável aleatória que representa o prejuízo, em reais, ao deixar o carro no estacionamento público, e que Y seja a variável aleatória que representa o valor, em reais, desembolsado por Roberto ao deixar o carro no estacionamento pago, julgue o item subsequente.

A variável aleatória Y é contínua.

Comentários:

As variáveis quantitativas podem ser classificadas em:

a) variáveis contínuas: características mensuráveis que assumem valores em uma escala contínua, nesse caso, valores fracionários fazem sentido;

b) variáveis discretas: características mensuráveis que podem assumir apenas um número finito ou infinito contável de valores e, assim, somente fazem sentido valores inteiros.

No problema em questão, a variável aleatória Y pode assumir o valor R$ 70,00, se Roberto deixar o carro no estacionamento pago; ou o valor R$ 0,00, se Roberto não deixar o carro no estacionamento pago. Portanto, Y é uma variável aleatória discreta.

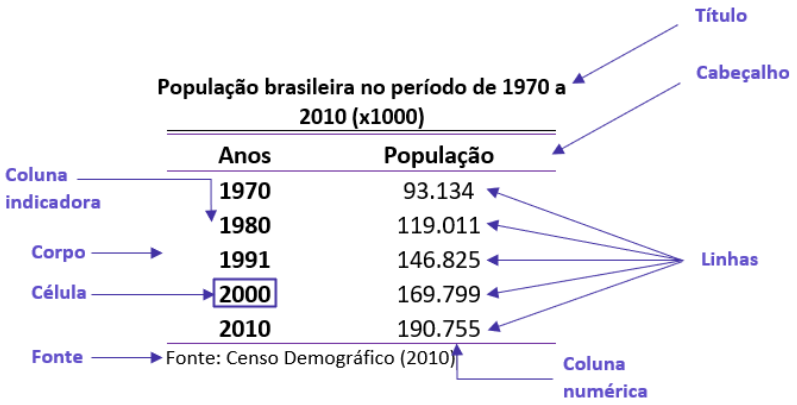
Gabarito: Errado.

SÉRIES ESTATÍSTICAS

Uma **série estatística** consiste em um conjunto de dados organizado com base em uma característica comum, ou seja, uma mesma variável. Normalmente, uma série estatística é representada por meio de uma **tabela** ou de um **gráfico**, conforme ficar melhor representado, a fim de sintetizar os dados estatísticos observados e torná-los mais compreensivos.

Uma **tabela** é, basicamente, um quadro que resume um conjunto de observações, sendo composta de:

a) corpo – conjunto de linhas e colunas com as informações sobre a variável em estudo;
b) cabeçalho – parte superior que especifica o conteúdo das colunas;
c) coluna indicadora – parte que indica o conteúdo das linhas;
d) linhas – traços que facilitam a leitura dos dados;
e) célula – espaço onde os dados são armazenados;
f) título – identificação da tabela, contendo as informações sobre seu conteúdo;
g) fonte – referência de onde os dados foram obtidos, localizada no rodapé.



Um **gráfico** é uma **forma clara e objetiva** de apresentar uma **série estatística**. O objetivo é **proporcionar uma compreensão mais rápida do fenômeno em estudo**. Para isso, o gráfico deve ser destituído de detalhes sem importância (**ser simples**); permitir a correta intepretação dos valores representativos do fenômeno (**ser claro**); e transmitir a verdade sobre o fenômeno (**ser verossímil**). A série estatística apresentada na tabela anterior pode ser representada graficamente da seguinte forma:

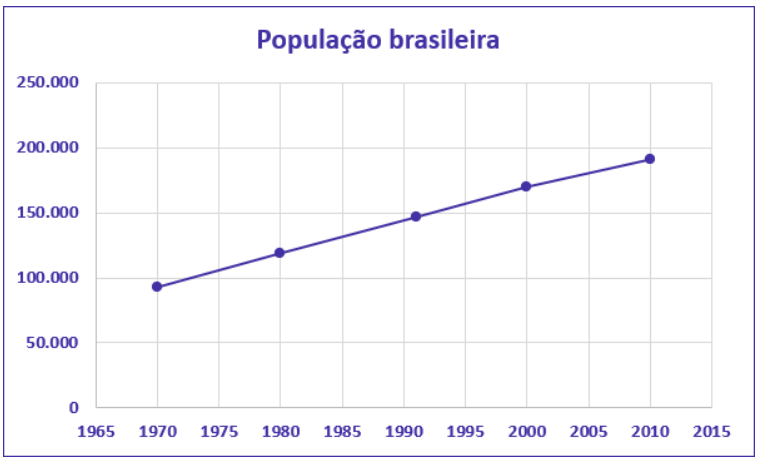


Tabela	Gráfico
• Quadro que resume um conjunto de observações. • Composta de cabeçalho, corpo, coluna indicadora, linhas, células, título e fonte.	• Forma simples e clara de apresentar uma série estatística. • Proporcionar uma compreensão mais rápida do fenômeno em estudo. • Deve ser simples, claro e verossímil.

Finalmente, podemos verificar a presença de **três elementos** nas séries estatística: o **tempo**, o **espaço** e a **espécie**. Conforme os elementos variem, a série pode ser classificada em **temporal** (ou cronológica), **geográfica** (ou territorial) e **específica**.

Séries Temporais (ou Cronológicas)

É a série cujos dados são dispostos segundo a época de ocorrência. Isto é, enquanto o tempo varia, o fato e o local permanecem constantes. Também são chamadas de séries **históricas** ou **evolutivas**. A principal característica é o **fator cronológico variável**.

A seguir temos a série histórica da população residente no Brasil no período de 1970 a 2010, com frequência decenal. Percebam que o **tempo varia**; contudo, o fato que está sendo analisado (quantidade populacional) e o local alvo da pesquisa (Brasil), continuam iguais.

População brasileira no período de 1970 a 2010 (x1000)

Anos	População
1970	93.134
1980	119.011
1991	146.825
2000	169.799
2010	190.755

Fonte: Censo Demográfico (2010)

Séries Geográficas (ou Territoriais)

É a série cujos dados são dispostos segundo a localidade de ocorrência. Isto é, enquanto o local varia, o fato e o tempo permanecem constantes. Também são chamadas de séries **espaciais** ou de **localização**. A principal característica é o **fator geográfico variável**.

A seguir temos a série geográfica da população urbana residente em cada uma das regiões brasileiras no ano de 2010. Percebam que o **local (região) varia**; contudo, o fato que está sendo analisado (quantidade populacional) e o tempo (ano de 2010) permanecem constantes.

População Urbana em 2010 (x1000)

Região	População
Norte	11.664
Nordeste	38.821
Sudeste	74.696
Sul	23.260
Centro-Oeste	12.482

Fonte: Censo Demográfico (2010)

Séries Específicas

É a série cujos dados são dispostos segundo a modalidade de ocorrência. Isto é, enquanto o fato varia, a época e o local permanecem constantes. Também são chamadas de séries **categóricas**. A principal característica é o **fator especificativo variável**.

A seguir temos uma série específica das populações urbana e rural residentes no Brasil no ano de 2010. Percebam que os fatos analisados variam (população urbana x população rural); contudo, o tempo (2010) e o local de análise (Brasil) são constantes.

População Urbana e Rural em 2010 (x1000)

Zona	População
Urbana	93.134
Rural	119.011
Total	**190.755**

Fonte: Censo Demográfico (2010)

Séries Mistas (ou Compostas)

Muitas vezes, podemos ter a necessidade de apresentar, em uma única tabela, a variação de valores de mais de uma variável, isto é, combinar duas ou mais séries. As séries resultantes desse processo de combinação são chamadas de **séries mistas (ou compostas)** e apresentadas por meio de **tabelas de dupla entrada**.

O nome da nova série deve levar em consideração pelo menos dois elementos. Assim, **se for uma série mista de fato e tempo, denominaremos de série específico-temporal**. A seguir temos uma série específico-temporal representando as populações de homens e mulheres residentes no brasil, no período de 1970 a 2010, com variação decenal.

População do Brasil por Sexo de 1970 a 2010 (x1000)

Anos	Sexo	
	Homens	Mulheres
1970	46.327	46.807
1980	59.142	59.868
1991	72.485	74.340
2000	83.602	86.270
2010	93.406	97.348

Fonte: Censo Demográfico (2010)

Por sua vez, **se tivermos uma série mista de local e tempo, denominaremos de série geográfica-temporal**. A seguir temos uma série geográfico-temporal representando as populações residentes em cada região brasileira, no período de 1970 a 2010, com variação decenal.

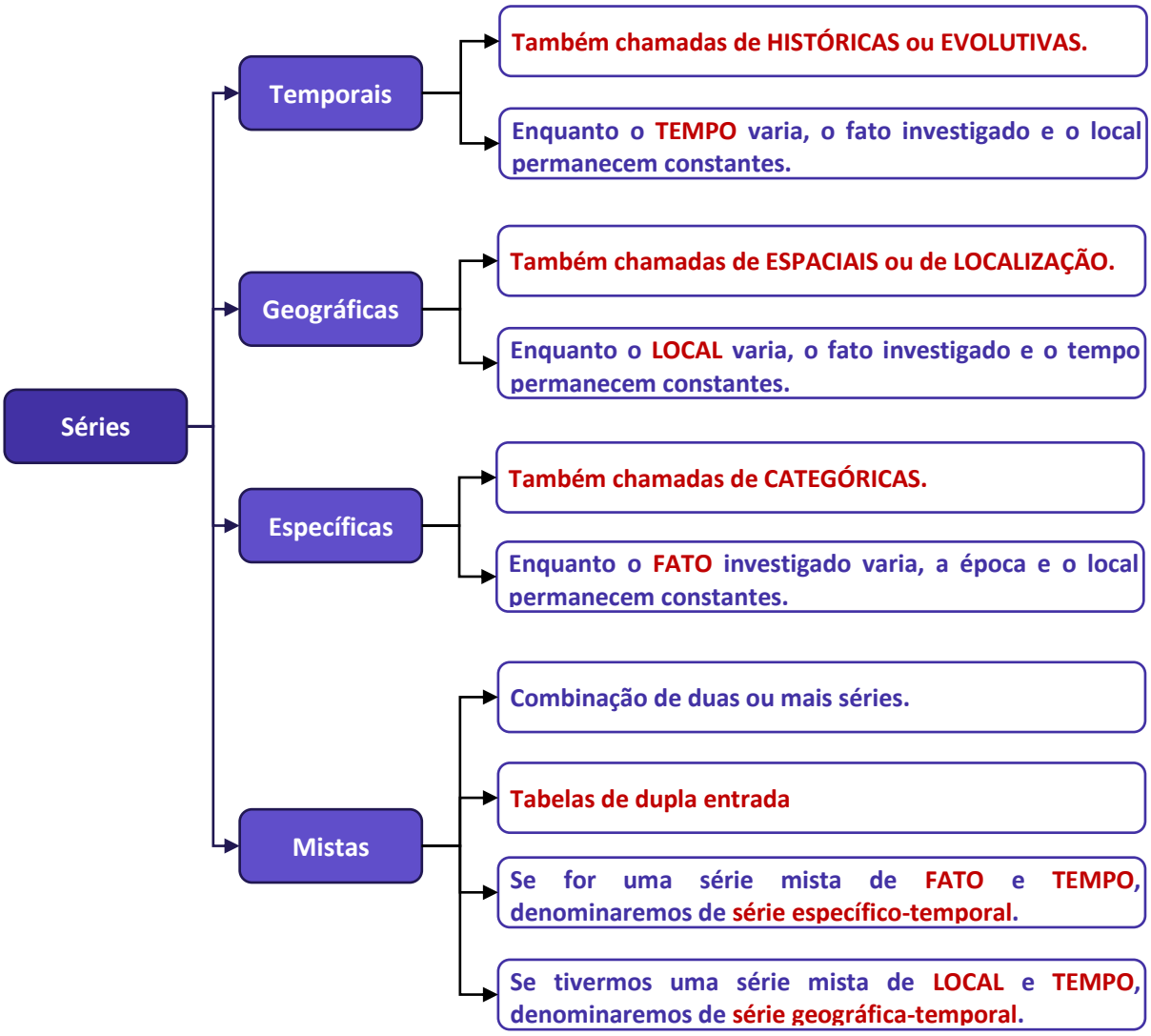
População do Brasil por Região de 1970 a 2010 (x1000)

Anos	Regiões				
	N	NE	SE	S	CO
1970	3.603	28.111	39.850	16.496	5.072
1980	5.880	34.815	51.737	19.031	7.545
1991	10.030	42.497	62.740	22.129	9.427
2000	12.900	47.741	72.412	25.107	11.636
2010	15.864	53.081	80.364	27.386	14.058

Fonte: Censo Demográfico (2010)

Por fim, devemos notar que podem existir séries compostas de três ou mais entradas, embora isso raramente aconteça, por conta da dificuldade de representação.


RESUMINDO


Séries
Temporais
Também chamadas de HISTÓRICAS ou EVOLUTIVAS.
Enquanto o TEMPO varia, o fato investigado e o local permanecem constantes.
Geográficas
Também chamadas de ESPACIAIS ou de LOCALIZAÇÃO.
Enquanto o LOCAL varia, o fato investigado e o tempo permanecem constantes.
Específicas
Também chamadas de CATEGÓRICAS.
Enquanto o FATO investigado varia, a época e o local permanecem constantes.
Mistas
Combinação de duas ou mais séries.
Tabelas de dupla entrada
Se for uma série mista de FATO e TEMPO, denominaremos de série específico-temporal.
Se tivermos uma série mista de LOCAL e TEMPO, denominaremos de série geográfica-temporal.

Distribuição de Frequências

Vimos anteriormente que, logo após a coleta de dados, temos o que denominamos de dados brutos. Como exemplo de dados brutos, citamos uma pesquisa de tempo médio de estudo diário, em minutos, envolvendo 50 alunos do Estratégia, em que os alunos foram escolhidos de maneira aleatória, não havendo qualquer organização dos valores observados. Por terem sido apresentados na forma em que foram coletados, são denominados de dados brutos.

Aluno	Tempo	Aluno	Tempo	Aluno	Tempo	Aluno	Tempo	Aluno	Tempo
1	143	11	113	21	170	31	124	41	105
2	142	12	143	22	158	32	137	42	154
3	161	13	159	23	123	33	153	43	99
4	126	14	168	24	96	34	129	44	114
5	134	15	123	25	98	35	148	45	161
6	137	16	135	26	135	36	173	46	128
7	171	17	135	27	129	37	126	47	175
8	85	18	175	28	126	38	104	48	137
9	155	19	115	29	103	39	157	49	165
10	171	20	89	30	171	40	127	50	115

Normalmente, esses dados fornecem pouca informação ao leitor, sendo necessário organizá-los, com o propósito de aumentar sua capacidade informativa. A simples organização dos dados em um rol crescente já ajuda bastante nesse sentido. Com os dados organizados em rol, conseguimos verificar que o menor tempo observado foi de 85 minutos, e o maior, de 175 minutos, o que nos fornece uma amplitude total (AT = 175 - 85 = 90) de variação da ordem de 90 minutos.

Rol Crescente				
85	115	129	143	161
89	115	129	143	165
96	123	134	148	168
98	123	135	153	170
99	124	135	154	171
103	126	135	155	171
104	126	137	157	171

105	126	137	158	173
113	127	137	159	175
114	128	142	161	175

Outra informação que conseguimos extrair dos dados organizados em rol crescente é que alguns tempos, como 126 min, 135 min, 137 min e 171 min, foram mais frequentes, ou seja, apareceram mais vezes durante a pesquisa.

Uma maneira mais concisa de mostrar os dados do rol é apresentar cada valor juntamente com o número de vezes em que ocorre, em vez de repeti-los. O número de ocorrências de um determinado valor recebe o nome de frequência. A tabela que contém todos os valores com suas respectivas frequências é denominada de distribuição de frequências.

Uma distribuição de frequências também pode ser definida como uma série estatística na qual permanecem constantes o fato, o local e a época. Ela pode ser classificada em dois tipos: distribuição de frequências pontual (ou discreta) e distribuição de frequências intervalar (ou contínua).

Na distribuição de frequências pontual, são apresentados todos os dados coletados juntamente com suas respectivas frequências, não havendo perda de valores. Contudo, esse processo pode exigir muito espaço, especialmente quando o número de valores da variável tende a aumentar.

Tempo (min)	Freq.	Tempo (min)	Freq.	Tempo (min)	Freq.	Tempo (min)	Freq.
85	1	114	1	135	3	158	1
89	1	115	2	137	3	159	1
96	1	123	2	142	1	161	2
98	1	124	1	143	2	165	1
99	1	126	3	148	1	168	1
103	1	127	1	153	1	170	1
104	1	128	1	154	1	171	3
105	1	129	2	155	1	173	1
113	1	134	1	157	1	175	2

Nesse caso, quando a variável é contínua, o mais recomendável é agrupar os valores por intervalos de classe. Desse modo, em vez de listar cada um dos valores que ocorrem, utilizamos uma distribuição de frequências intervalar, listando os intervalos de classe e as frequências correspondentes.

Tempo médio (X_i)	Frequência (f_i)
$85 \leq x < 100$	5

$100 \leq x < 115$	5
$115 \leq x < 130$	12
$130 \leq x < 145$	10
$145 \leq x < 160$	7
$160 \leq x < 175$	9
$175 \leq x < 190$	2

Procedendo dessa forma, perdemos a informação detalhada dos tempos médios, mas ganhamos em termos de praticidade, o que simplifica o processo de análise de dados. Examinando a tabela acima, percebemos facilmente que a maioria dos alunos estuda diariamente entre 115 e 130 minutos, enquanto uma minoria alcança entre 175 e 190 minutos.

Para identificar uma classe, temos que conhecer os valores dos limites inferior e superior da classe, que delimitam um intervalo de classe. Desse modo, precisamos definir a natureza do intervalo de classe, se aberto ou fechado. Portanto, temos as seguintes notações para os diferentes tipos de intervalos:

Tipo de Intervalo	Notação matemática	Notação estatística	Significado
Intervalo aberto	$a < x < b$	$a - b$	Engloba todos os elementos entre a e b, mas não engloba a nem b.
Intervalo fechado à esquerda e aberto à direita	$a \leq x < b$	$a \vdash b$	Engloba todos os elementos entre a e b, inclusive a mas não b.
Intervalo aberto à esquerda e fechado à direita	$a < x \leq b$	$a \dashv b$	Engloba todos os elementos entre a e b, inclusive b mas não a.
Intervalo fechado	$a \leq x \leq b$	$a \vdash\!\dashv b$	Engloba todos os elementos entre a e b, inclusive a e b.

Por fim, é importante salientarmos que, em análises estatísticas, constantemente encontramos distribuições de frequências intervalares, pois o objetivo da estatística é justamente fazer um apanhado geral das características de um conjunto de dados, sem adentrar em detalhes de casos particulares.

Elementos de uma Distribuição de Frequências

Agora, analisaremos de forma detalhada cada elemento de uma distribuição de frequências. Tomaremos como referência a tabela apresentada anteriormente:

Tempo médio (X_i)	Frequência (f_i)
$85 \leq x < 100$	5
$100 \leq x < 115$	5
$115 \leq x < 130$	12
$130 \leq x < 145$	10
$145 \leq x < 160$	7
$160 \leq x < 175$	9
$175 \leq x < 190$	2

Classe

As classes são os intervalos nos quais o fenômeno é subdividido. Podemos dizer que as classes são os intervalos ou subdivisões dos elementos que compõem um conjunto de dados. Na tabela anterior, temos as seguintes classes:

Classe	Intervalo	Frequência (f_i)
1ª Classe	$85 \leq x < 100$	5
2ª Classe	$100 \leq x < 115$	5
3ª Classe	$115 \leq x < 130$	12
4ª Classe	$130 \leq x < 145$	10
5ª Classe	$145 \leq x < 160$	7
6ª Classe	$160 \leq x < 175$	9
7ª Classe	$175 \leq x < 190$	2
		$n = 50$

Existem duas maneiras de determinar o número "ideal" de classes, k, em função do número de dados da tabela, n. A primeira consiste em utilizar a fórmula de Sturges:

$$k = 1 + 3,3 \times log\, n$$

Outra abordagem, utilizada quando o número de dados é menor ou igual a 50, é por meio da fórmula:

$$k = \sqrt{n}$$

Vamos aproveitar para calcular o número de classes do nosso exemplo:

a) pela fórmula de Sturges:

$$k = 1 + 3,3 \times log\, n$$

$$k = 1 + 3,3 \times log\, 50$$

$$k = 1 + 3,3 \times 1,7$$

$$k = 1 + 5,61 = 6,61$$

b) pela outra fórmula:

$$k = \sqrt{n}$$

$$k = \sqrt{50} = 7,07$$

Limite de Classe

Cada classe tem um **limite inferior de classe** (l_{inf}), que é o menor número que pode pertencer à classe, e um **limite superior de classe** (l_{sup}), que é o maior número que pode pertencer à classe. Os limites de uma classe são seus valores extremos.

Vamos identificar os limites inferiores e superiores do nosso exemplo:

Classes	Limite Inferior (l_{inf})	Limite Superior (l_{sup})
$85 \leq x < 100$	85	100
$100 \leq x < 115$	100	115
$115 \leq x < 130$	115	130
$130 \leq x < 145$	130	145
$145 \leq x < 160$	145	160
$160 \leq x < 175$	160	175
$175 \leq x < 190$	175	190

Amplitude de um Intervalo de Classe

A amplitude de um intervalo de classe, ou simplesmente intervalo de classe, é a distância entre os limites inferiores (ou superiores) de classes consecutivas. Ela é obtida pela diferença entre dois limites inferiores (ou superiores) consecutivos:

$$h = l_{sup} - l_{inf}$$

em que l_{inf} é o limite inferior do intervalo de classe e l_{sup} é o limite superior do intervalo de classe.

Dando continuidade ao nosso exemplo, vejamos como calcular as amplitudes dos intervalos:

Classes	Limite Inferior (l_{inf})	Limite Superior (l_{sup})	Amplitude de Classe ($h = l_{sup} - l_{inf}$)
$85 \leq x < 100$	85	100	100 - 85 = 15
$100 \leq x < 115$	100	115	115 - 100 = 15
$115 \leq x < 130$	115	130	130 - 115 = 15
$130 \leq x < 145$	130	145	145 - 130 = 15
$145 \leq x < 160$	145	160	160 - 145 = 15
$160 \leq x < 175$	160	175	175 - 160 = 15
$175 \leq x < 190$	175	190	190 - 175 = 15

Embora seja desejável, a amplitude do intervalo de classe nem sempre será constante ao longo de toda a distribuição de frequências intervalar.

Amplitude Total

A amplitude total é a diferença entre o limite superior da última classe (limite superior máximo) e o limite inferior da primeira classe (limite inferior mínimo). Portanto, corresponde à diferença entre o último e o primeiro elemento de um conjunto de dados ordenado de forma crescente:

$$AT = l_{máx} - l_{mín}$$

Note que, quando todas as classes possuem a mesma amplitude, também podemos determinar o valor da amplitude total multiplicando o valor do intervalo de classe (h) pela quantidade de classes da distribuição (k):

$$AT = h \times k$$

Em nosso exemplo, a amplitude total é calculada da seguinte maneira:

$$AT = l_{máx} - l_{mín}$$

$$AT = 190 - 85 = 105$$

Ponto médio de classe

O ponto médio é a média aritmética simples dos valores extremos de uma classe, ou seja, a soma dos limites inferior e superior dividida por dois. Esse ponto divide a classe em duas partes iguais. Também costuma ser chamado de marca ou representante da classe.

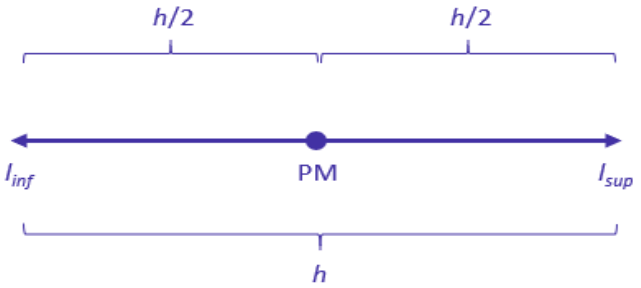
$$PM = \frac{(l_{inf} + l_{sup})}{2}$$

Para praticar, vamos calcular os pontos médios de nossa distribuição de frequências:

Tempo médio (X_i)	Ponto médio (PM_i)	Frequência (f_i)
$85 \leq x < 100$	$\frac{85+100}{2} = 92,5$	5
$100 \leq x < 115$	$\frac{100+115}{2} = 107,5$	5
$115 \leq x < 130$	$\frac{115+130}{2} = 122,5$	12
$130 \leq x < 145$	$\frac{130+145}{2} = 137,5$	10
$145 \leq x < 160$	$\frac{145+160}{2} = 152,5$	7
$160 \leq x < 175$	$\frac{160+175}{2} = 167,5$	9
$175 \leq x < 190$	$\frac{175+190}{2} = 182,5$	2

Veja que os pontos médios formaram uma progressão aritmética, pois a diferença entre dois pontos médios consecutivos foi constante e igual a 15. Isso ocorreu porque o intervalo de classe, h, também foi constante (e igual a 15) em toda a distribuição. Assim, quando o intervalo de classes é constante, a diferença entre os pontos médios também será constante e igual ao intervalo de classe.

Adicionalmente, sabendo que o ponto médio divide a classe em duas partes iguais, podemos derivar outras relações envolvendo o próprio ponto médio, a amplitude de classe e os limites inferior e superior.



Dada a figura anterior, podemos obter os limites de uma classe por meio das seguintes expressões:

$$l_{inf} = PM - \frac{h}{2}$$

$$l_{sup} = PM + \frac{h}{2}$$

Além disso, podemos encontrar o ponto médio de uma classe a partir das seguintes relações:

$$PM = l_{inf} + \frac{h}{2}$$

$$PM = l_{sup} - \frac{h}{2}$$

Frequência

Ao longo dessa aula, em várias oportunidades abordamos conceitos relacionados à frequência, isto é, ao número de ocorrências de um determinado valor ou de uma certa classe. Esse conceito é de grande relevância para a estatística descritiva e deve ser estudado de forma mais aprofundada. Nesse contexto, é importante sabermos que existem quatro tipos de frequência, os quais serão analisados nas subseções seguintes:

a) frequência absoluta simples (f_i);
b) frequência absoluta acumulada (f_{ac});
c) frequência relativa simples (F_i);
d)frequência relativa acumulada (F_{ac}).

Frequência Absoluta Simples

A frequência absoluta simples corresponde ao número de observações correspondentes a uma determinada classe ou a um determinado valor.

i	Tempos (min)	Frequência (f_i)
1	$85 \vdash 100$	5
2	$100 \vdash 115$	5
3	$115 \vdash 130$	12
4	$130 \vdash 145$	10
5	$145 \vdash 160$	7
6	$160 \vdash 175$	9

7	$175 \vdash 190$	2

A frequência simples é simbolizada por f_i. No exemplo anterior, temos: $f_1 = 5$, $f_2 = 5$, $f_3 = 12$, $f_4 = 10$, $f_5 = 7$, $f_6 = 9$ e $f_7 = 2$.

A soma de todas as frequências é igual ao número total de dados analisados:

$$\sum_{i=1}^{k} f_i = n$$

em que a notação $\sum_{i=1}^{k} f_i$ representa o somatório das frequências de cada uma das k classes.

Para a distribuição em análise, temos:

$$\sum_{i=1}^{7} f_i = f_1 + f_2 + f_3 + f_4 + f_5 + f_6 + f_7 = 5 + 5 + 12 + 10 + 7 + 9 + 2 = 50$$

Agora, podemos incluir essa informação na representação tabular:

i	Tempos (min)	Frequência (f_i)
1	$85 \vdash 100$	5
2	$100 \vdash 115$	5
3	$115 \vdash 130$	12
4	$130 \vdash 145$	10
5	$145 \vdash 160$	7
6	$160 \vdash 175$	9
7	$175 \vdash 190$	2
		$\sum_{i=1}^{7} f_i = 50$

Frequência Absoluta Acumulada

A frequência absoluta acumulada crescente (f_{ac}) é a soma das frequências de todos os valores inferiores ao limite superior do intervalo de uma determinada classe. No exemplo apresentado anteriormente, a frequência acumulada correspondente à quarta classe é:

$f_{ac_4} = f_1 + f_2 + f_3 + f_4 = 5 + 5 + 12 + 10 = 32$, significando que 32 alunos estudam por um período igual ou superior a 85 minutos e inferior a 145 minutos (limite superior da quarta classe).

$$f_{ac_i} = f_1 + f_2 + f_3 + \cdots + f_i$$

A frequência absoluta acumulada crescente é calculada de cima para baixo, da seguinte forma:
1) repetimos a frequência absoluta da primeira classe;

2) os demais valores da frequência absoluta são obtidos a partir da soma da frequência acumulada anterior com a frequência absoluta da classe correspondente;

3) a frequência acumulada crescente sempre termina com o valor de n.

i	Tempos (min)	Frequência (f_i)	Frequência Acumulada
1	85 ⊢ 100	5	5
2	100 ⊢ 115	5	10
3	115 ⊢ 130	12	22
4	130 ⊢ 145	10	32
5	145 ⊢ 160	7	39
6	160 ⊢ 175	9	48
7	175 ⊢ 190	2	50
		$\sum_{i=1}^{7} f_i = 50$	

A frequência absoluta acumulada decrescente (f_{ad}) é a soma das frequências de todos os valores superiores ao limite inferior do intervalo de uma determinada classe. No exemplo apresentado anteriormente, a frequência acumulada correspondente à quarta classe é: $f_{ad_4} = f_4 + f_5 + f_6 + f_7 = 10 + 7 + 9 + 2 = 28$, significando que 28 alunos estudam por um período igual ou superior a 130 minutos (limite inferior da quarta classe) e inferior a 190 minutos.

$$f_{ad_i} = f_i + f_{i+1} + f_{i+2} + \cdots + f_k$$

A frequência absoluta acumulada decrescente é calculada de baixo para cima, da seguinte forma:
1) repetimos a frequência absoluta da última classe;

2) os demais valores da frequência acumulada decrescente são obtidos a partir da soma da frequência acumulada anterior com a frequência absoluta da classe correspondente;

3) a frequência acumulada decrescente sempre termina com o valor de n.

i	Tempos (min)	Frequência (f_i)	Frequência Acumulada
1	$85 \vdash 100$	5	50
2	$100 \vdash 115$	5	45
3	$115 \vdash 130$	12	40
4	$130 \vdash 145$	10	28
5	$145 \vdash 160$	7	18
6	$160 \vdash 175$	9	11
7	$175 \vdash 190$	2	2
		$\sum_{i=1}^{7} f_i = 50$	

Frequência Relativa Simples

A frequência relativa simples corresponde à proporção de dados existentes em uma determinada classe. Para calcular a frequência relativa de uma classe, dividimos a frequência absoluta simples f_i pela frequência total (isto é, dividimos a parte pelo todo):

$$F_i = \frac{f_i}{\Sigma f_i} = \frac{f_i}{n}$$

Em nosso exemplo, as frequências relativas são:

i	Tempos (min)	Frequência Absoluta (f_i)	Frequência Relativa (F_i)
1	$85 \vdash 100$	5	$F_1 = \frac{5}{50} = 0,10$
2	$100 \vdash 115$	5	$F_2 = \frac{5}{50} = 0,10$
3	$115 \vdash 130$	12	$F_3 = \frac{12}{50} = 0,24$
4	$130 \vdash 145$	10	$F_4 = \frac{10}{50} = 0,20$
5	$145 \vdash 160$	7	$F_5 = \frac{7}{50} = 0,14$
6	$160 \vdash 175$	9	$F_6 = \frac{9}{50} = 0,18$
7	$175 \vdash 190$	2	$F_7 = \frac{2}{50} = 0,04$

Para representar esses valores em termos de porcentagem, basta multiplicarmos por 100%. A tabela ficaria assim:

i	Tempos (min)	Frequência Absoluta (f_i)	Frequência Relativa (F_i)
1	$85 \vdash 100$	5	$F_1 = \frac{5}{50} \times 100\% = 10\%$
2	$100 \vdash 115$	5	$F_2 = \frac{5}{50} \times 100\% = 10\%$
3	$115 \vdash 130$	12	$F_3 = \frac{12}{50} \times 100\% = 24\%$
4	$130 \vdash 145$	10	$F_4 = \frac{10}{50} \times 100\% = 20\%$
5	$145 \vdash 160$	7	$F_5 = \frac{7}{50} \times 100\% = 14\%$
6	$160 \vdash 175$	9	$F_6 = \frac{9}{50} \times 100\% = 18\%$
7	$175 \vdash 190$	2	$F_7 = \frac{2}{50} \times 100\% = 4\%$

O propósito das frequências relativas é facilitar a realização de comparações de classes individuais com o total das observações. Na tabela anterior, por exemplo, conseguimos verificar facilmente que 20% das observações pertencem à quarta classe e que 18% das observações pertencem à sexta classe.

Repare que a soma de todas as frequências relativas deve ser igual a 100%:

$$\sum_{i=1}^{k} F_i = 100\%$$

Podemos incluir essa informação na representação tabular:

i	Tempos (min)	Frequência Absoluta (f_i)	Frequência Relativa (F_i)
1	$85 \vdash 100$	5	10%
2	$100 \vdash 115$	5	10%
3	$115 \vdash 130$	12	24%
4	$130 \vdash 145$	10	20%
5	$145 \vdash 160$	7	14%
6	$160 \vdash 175$	9	18%
7	$175 \vdash 190$	2	4%
		$\sum_{i=1}^{7} f_i = 50$	$\sum_{i=1}^{7} F_i = 100\%$

Frequência Relativa Acumulada

A frequência relativa acumulada crescente (F_{ac}) é a proporção de valores inferiores ao limite superior do intervalo de uma dada classe. No exemplo apresentado anteriormente, a frequência acumulada correspondente à quarta classe é: $F_{ac_4} = F_1 + F_2 + F_3 + F_4 = 10\% + 10\% + 24\% + 20\% = 64\%$, significando que 64% dos alunos estudam por um período igual ou superior a 85 minutos e inferior a 145 minutos (limite superior da quarta classe).

$$F_{ac_i} = F_1 + F_2 + F_3 + \cdots + F_i$$

A frequência relativa acumulada crescente é calculada de cima para baixo, da seguinte forma:
1) repetimos a frequência relativa da primeira classe;

2) os demais valores são obtidos a partir da soma da frequência relativa acumulada anterior com a frequência relativa da classe correspondente;

3) a frequência relativa acumulada sempre termina com o valor de 100%.

i	Tempos (min)	Frequência Absoluta (f_i)	Frequência Relativa (F_i)	Frequência Acumulada (F_{ac})
1	85 ⊢ 100	5	10%	10%
2	100 ⊢ 115	5	10%	20%
3	115 ⊢ 130	12	24%	44%
4	130 ⊢ 145	10	20%	64%
5	145 ⊢ 160	7	14%	78%
6	160 ⊢ 175	9	18%	96%
7	175 ⊢ 190	2	4%	100%
		$\sum_{i=1}^{7} f_i = 50$	$\sum_{i=1}^{7} F_i = 100\%$	

A frequência relativa acumulada decrescente (F_{ad}) a proporção de valores superiores ao limite inferior do intervalo de uma dada classe. No exemplo apresentado anteriormente, a frequência acumulada correspondente à quarta classe é: $F_{ad_4} = F_4 + F_5 + F_6 + F_7 = 20\% + 14\% + 18\% + 4\% = 56\%$, significando que 56% dos alunos estudam por um período igual ou superior a 130 minutos (limite inferior da quarta classe) e inferior a 190 minutos.

$$F_{ad_i} = F_i + F_{i+1} + F_{i+2} + \cdots + F_k$$

A frequência relativa acumulada decrescente é calculada de baixo para cima, da seguinte forma:
1) repetimos a frequência relativa da última classe;

2) os demais valores são obtidos a partir da soma da frequência acumulada anterior com a frequência relativa da classe correspondente;

3) a frequência acumulada sempre termina com o valor de 100%.

i	Tempos (min)	Frequência Absoluta (f_i)	Frequência Relativa (F_i)	Frequência Acumulada (F_{ad})
1	85 ⊢ 100	5	10%	100%
2	100 ⊢ 115	5	10%	90%
3	115 ⊢ 130	12	24%	80%
4	130 ⊢ 145	10	20%	56%
5	145 ⊢ 160	7	14%	36%
6	160 ⊢ 175	9	18%	22%
7	175 ⊢ 190	2	4%	4%
		$\sum_{i=1}^{7} f_i = 50$	$\sum_{i=1}^{7} F_i = 100\%$	

Densidade de Frequência

A densidade de frequência de uma classe consiste no quociente entre a frequência da classe (absoluta ou relativa) e sua amplitude:

$$densidade = \frac{frequência\ absoluta\ ou\ relativa}{amplitude}$$

$$d_i = \frac{f_i}{h_i}$$

Para o nosso exemplo, as densidades de frequência são:

i	Tempos (min)	Frequência Absoluta (f_i)	Densidade de Frequência (d)
1	85⊢100	5	$d_1 = \frac{f_1}{h_1} = \frac{5}{15} = 0,33$
2	100⊢115	5	$d_2 = \frac{f_2}{h_2} = \frac{5}{15} = 0,33$
3	115⊢130	12	$d_3 = \frac{f_3}{h_3} = \frac{12}{15} = 0,80$
4	130⊢145	10	$d_4 = \frac{f_4}{h_4} = \frac{10}{15} = 0,66$
5	145⊢160	7	$d_5 = \frac{f_5}{h_5} = \frac{7}{15} = 0,46$

6	160⊢175	9	$d_6 = \frac{f_6}{h_6} = \frac{9}{15} = 0,60$
7	175⊢190	2	$d_7 = \frac{f_7}{h_7} = \frac{2}{15} = 0,13$

Item	Definição	Símbolos e Fórmulas
Número de Classes	As classes são os intervalos nos quais o fenômeno é subdividido.	$k = 1 + 3,3 \times log\ n$ ou $k = \sqrt{n}$
Limites de Classe	Correspondem aos valores extremos.	l_{inf} e l_{sup}
Amplitude de um Intervalo de Classe	Distância entre os limites inferiores (ou superiores) de classes consecutivas.	$h = l_{sup} - l_{inf}$
Amplitude total	Diferença entre o limite superior da última classe (limite superior máximo) e o limite inferior da primeira classe (limite inferior mínimo).	$AT = l_{máx} - l_{mín}$ $AT = h \times k$
Ponto Médio	Média aritmética simples dos valores extremos de uma classe.	$PM = \frac{(l_{inf} + l_{sup})}{2}$ $PM = l_{inf} + \frac{h}{2}$ $PM = l_{sup} - \frac{h}{2}$
Frequência Absoluta Simples	Número de observações correspondentes a uma determinada classe ou a um determinado valor.	f_i
Frequência Absoluta Acumulada	Total das frequências de todos os valores inferiores ao limite superior do intervalo de uma dada classe	$f_{ac_i} = f_1 + f_2 + f_3 + \cdots + f_i$
Frequência Relativa Simples	Proporção de dados existentes em uma determinada classe.	$F_i = \frac{f_i}{\Sigma f_i} = \frac{f_i}{n}$
Frequência Relativa Acumulada	Proporção de valores inferiores ao limite superior do intervalo de uma dada classe.	$F_{ac_i} = F_1 + F_2 + F_3 + \cdots + F_i$

Densidade de Frequência	Quociente entre a frequência da classe (absoluta ou relativa) e sua amplitude	$d = \frac{f}{h}$

(VUNESP/IPSM-SJC/2018) Considere as informações a seguir para construir uma distribuição de frequência sem intervalo de classe e responder a questão.

Um dado foi lançado 50 vezes e foram registrados os seguintes resultados:

5 4 6 1 2 5 3 1 3 3

4 4 1 5 5 6 1 2 5 1

3 4 5 1 1 6 6 2 1 1

4 4 4 3 4 3 2 2 2 3

6 6 3 2 4 2 6 6 2 1

A frequência simples relativa do primeiro elemento é

a) 10%.

b) 20%.

c) 50.

d) 7.

e) 20.

Comentários:

A frequência absoluta é o número de vezes em que uma determinada variável assume um valor, enquanto a frequência relativa é o quociente entre a frequência absoluta da classe correspondente e a soma das frequências (total observado).

Para os dados apresentados, temos que :

Valor obtido no lançamento	Frequência absoluta	Frequência relativa
1	10	10/50 = 0,2 = 20%
2	9	9/50 = 0,18 = 18%
3	8	8/50 = 0,16 = 16%
4	9	9/50 = 0,18 = 18%
5	6	6/50 = 0,12 = 12%
6	8	8/50 = 0,16 = 16%
Total	50	100%

Portanto, a frequência simples absoluta do primeiro elemento é igual a 20%.

Gabarito: B.

(CESPE/Polícia Federal/2018)

	DIA				
	1	2	3	4	5
X (quantidade diária de drogas apreendidas, em kg)	10	22	18	22	28

Tendo em vista que, diariamente, a Polícia Federal apreende uma quantidade X, em kg, de drogas em determinado aeroporto do Brasil, e considerando os dados hipotéticos da tabela precedente, que apresenta os valores observados da variável X em uma amostra aleatória de 5 dias de apreensões no citado aeroporto, julgue o item.

A tabela em questão descreve a distribuição de frequências da quantidade de drogas apreendidas nos cinco dias que constituem a amostra.

Comentários:

A tabela apenas registra as quantidades de drogas apreendidas por dia. Para que ela configurasse uma distribuição de frequências da quantidade de drogas apreendidas nos cinco dias, era necessário que cada valor estivesse associado a uma frequência, o que não ocorreu. Uma distribuição de frequência dos dados acima seria assim:

Valor	Frequência
10	1
18	1
22	2
28	1

Gabarito: Errado.

(CESPE/BACEN/2013)

2 4 8 4 8 1 2 32 12 1 5 7 5 5 3 4 24 19 4 14

Os dados mostrados acima representam uma amostra, em minutos, do tempo utilizado na armazenagem de formulários no almoxarifado central de certa instituição por diversos funcionários.

Com base nesses dados, julgue o próximo item.

A distribuição de frequência acumulada para tempo de armazenagem observado na amostra inferior a 8 minutos é igual a 13, o que corresponde a uma frequência relativa superior a 0,60.

Comentários:

Não precisamos construir a distribuição de frequências para responder à questão. Sabemos que o total de observações é 20, pois é esse o número de termos.

Vamos destacar as observações que são inferiores a 8:

2 4 8 4 8 1 2 32 12 1 5 7 5 5 3 4 24 19 4 14

São 13 números inferiores a 8. Portanto, a frequência acumulada para um tempo menor do que 8 é 13.

Para calcular a frequência relativa, basta dividir a frequência absoluta pelo total de observações:

$$\frac{13}{20} = 0,65$$

Logo, a frequência relativa acumulada é maior do que 0,60

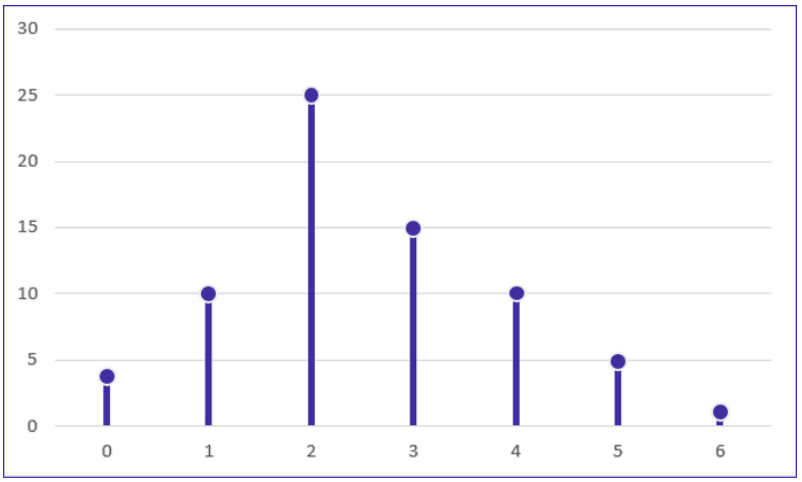
Gabarito: Certo.

Representações Gráficas das Distribuições de Frequência

Gráfico de Hastes ou Bastões

O gráfico de hastes ou bastões é muito utilizado para representar dados não agrupados em classes, o que normalmente ocorre com dados discretos. Nesse caso, não há perda de informação pois os valores da variável aparecem individualmente, conforme constam da amostra. Com relação a sua construção, basta representarmos as frequências simples absolutas ou relativas de cada elemento do conjunto de dados.

X_i	Frequência (f_i)
0	4
1	10
2	25
3	15
4	10
5	5
6	1



Repare que podemos reconstruir facilmente a tabela de frequências a partir do gráfico de hastes. De igual modo, conhecendo a tabela de frequências, podemos construir rapidamente o gráfico de hastes.

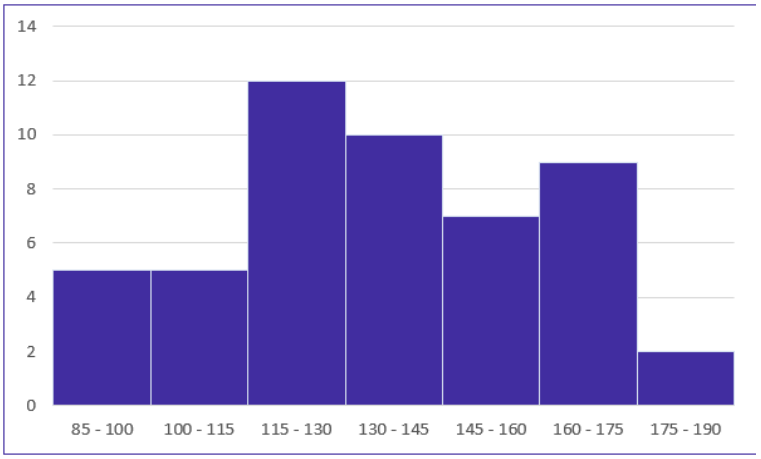
Histogramas

O histograma é um gráfico destinado a representar dados agrupados em classe, sendo composto por um conjunto de retângulos contíguos (justapostos) cujas bases ficam localizadas sobre o eixo horizontal (eixo x), de forma que os seus pontos médios devem coincidir com os pontos médios dos intervalos de classe e seus limites devem coincidir com os limites da classe.

A quantidade de retângulos em um histograma é equivalente ao número de intervalos de classe. A largura de cada retângulo deve ser igual à amplitude do intervalo de classe, enquanto a altura

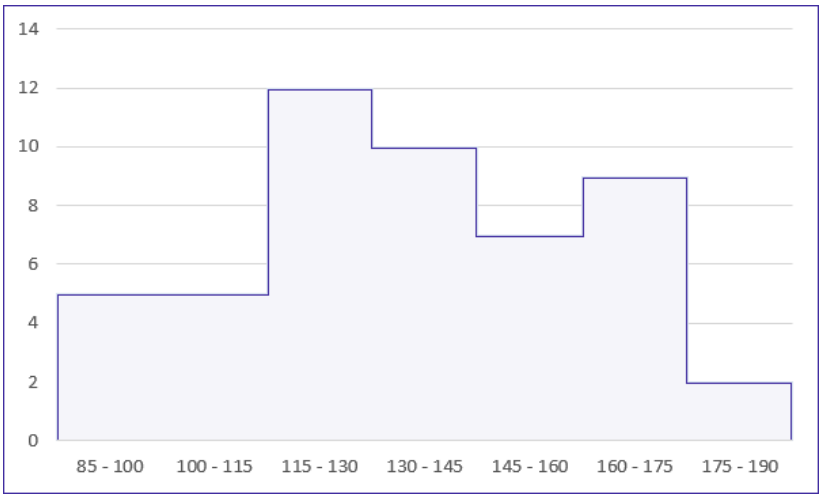
precisa ser proporcional à frequência do intervalo de classe. Além disso, a área do histograma é proporcional ao somatório das frequências.

Tempo médio (X_i)	Frequência (f_i)
$85 \leq x < 100$	5
$100 \leq x < 115$	5
$115 \leq x < 130$	12
$130 \leq x < 145$	10
$145 \leq x < 160$	7
$160 \leq x < 175$	9
$175 \leq x < 190$	2



A diferença básica entre um histograma e um gráfico de colunas (estudaremos na próxima seção) é a separação entre os retângulos adjacentes. Veja que não existe separação entre os retângulos no caso do histograma.

Dito isso, é importante mencionarmos a existência do gráfico denominado de poligonal característica, que construímos utilizando apenas os contornos do histograma.



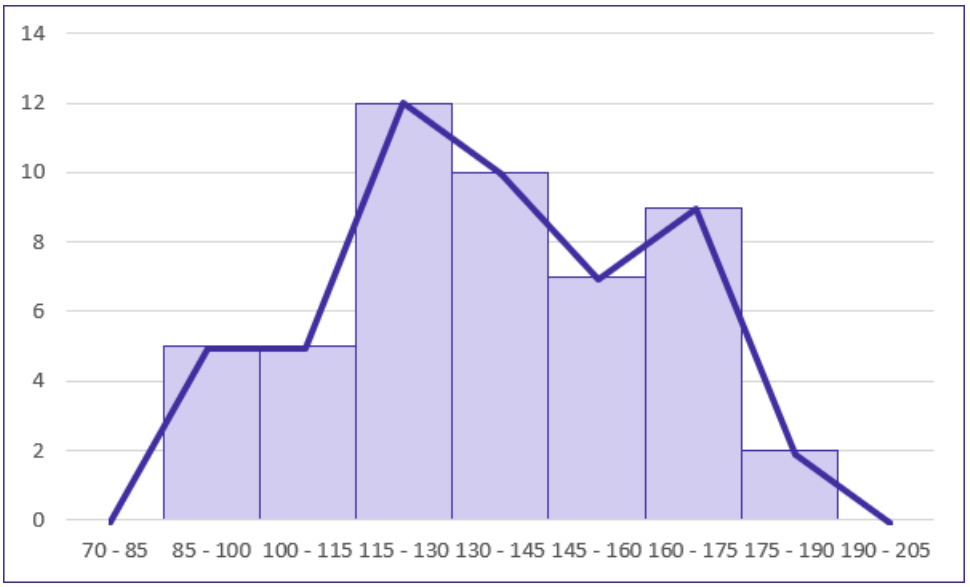
Por fim, o histograma pode ocasionar um certo nível de perda de informações, pois os elementos da distribuição de frequência não são representados de forma individualizada, mas sim por meio de suas classes.

Polígono de Frequências

O polígono de frequências é um gráfico em linha obtido por meio da ligação, por segmentos de reta, dos pontos médios das bases superiores dos retângulos de um histograma. Também é necessário considerar a existência de uma classe anterior à primeira e outra posterior à última, ambas com a frequência nula.

Assim como o histograma, o polígono de frequências apresenta área proporcional ao somatório das frequências.

Tempo médio (X_i)	Ponto médio (PM_i)	Frequência (f_i)
$70 \leq x < 85$	77,5	0
$85 \leq x < 100$	92,5	5
$100 \leq x < 115$	107,5	5
$115 \leq x < 130$	122,5	12
$130 \leq x < 145$	137,5	10
$145 \leq x < 160$	152,5	7
$160 \leq x < 175$	167,5	9
$175 \leq x < 190$	182,5	2
$190 \leq x < 205$	197,5	0



Curva de Frequências

A curva de frequências é obtida a partir do polimento de um polígono de frequências. Em sentido geométrico, o polimento corresponde à eliminação dos vértices (cantos) da linha poligonal. Esse processo suaviza os contornos do polígono de frequências, o que evidencia a verdadeira natureza dos dados em análise.

O polígono de frequências fornece a imagem real do fenômeno investigado, enquanto a curva de frequência mostra sua tendência. Naturalmente, quando o conjunto de dados é grande, a linha poligonal se torna curva. Por isso, podemos afirmar que a curva de frequência antecipa o comportamento da distribuição para um número maior de dados.

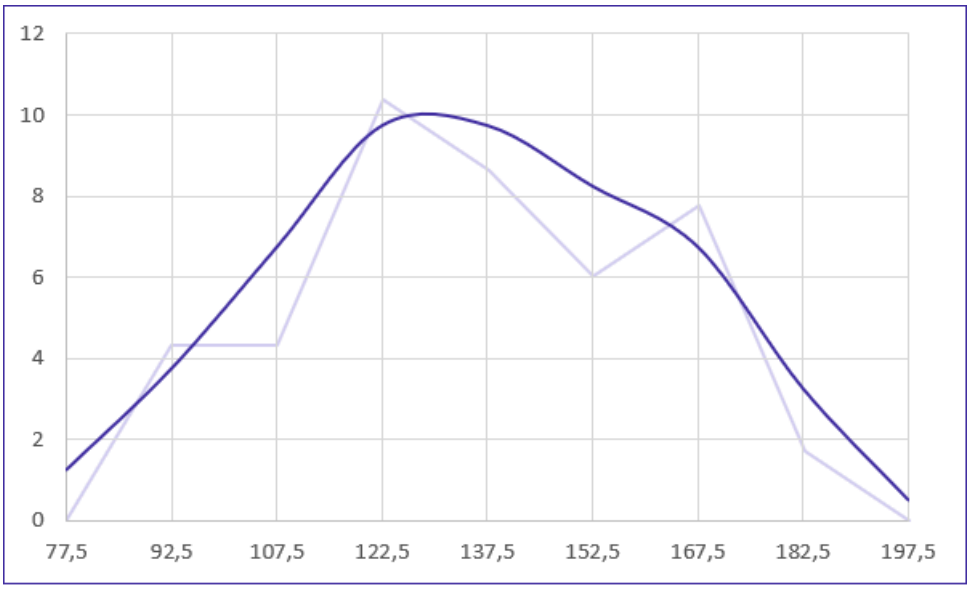
O processo de polimento é realizado por meio da seguinte fórmula:

$$fc_i = \frac{f_{ant}+2\times f_i+f_{post}}{4}$$

em que fc_i é a frequência calculada da classe considerada (freq. polida); f_i é a frequência simples da classe considerada; f_{ant} é a frequência simples da classe anterior à da classe considerada; e f_{post} é a frequência simples da classe posterior à da classe considerada.

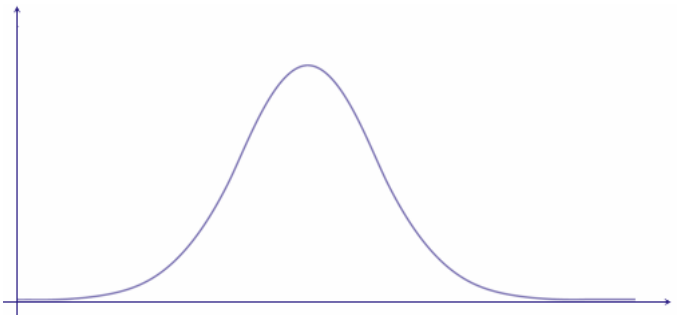
Vejamos como utilizá-la:

Tempo médio (X_i)	Ponto médio (PM_i)	Frequência (f_i)	Frequência Calculada (fc_i)
$70 \leq x < 85$	77,5	0	$fc_0 = \frac{0+2\times0+5}{4} = 1,25$
$85 \leq x < 100$	92,5	5	$fc_1 = \frac{0+2\times5+5}{4} = 3,75$
$100 \leq x < 115$	107,5	5	$fc_2 = \frac{5+2\times5+12}{4} = 6,75$
$115 \leq x < 130$	122,5	12	$fc_3 = \frac{5+2\times12+10}{4} = 9,75$
$130 \leq x < 145$	137,5	10	$fc_4 = \frac{12+2\times10+7}{4} = 9,75$
$145 \leq x < 160$	152,5	7	$fc_5 = \frac{10+2\times7+9}{4} = 8,75$
$160 \leq x < 175$	167,5	9	$fc_6 = \frac{7+2\times9+2}{4} = 6,75$
$175 \leq x < 190$	182,5	2	$fc_7 = \frac{9+2\times2+0}{4} = 3,25$
$190 \leq x < 205$	197,5	0	$fc_8 = \frac{2+2\times0+0}{4} = 0,50$

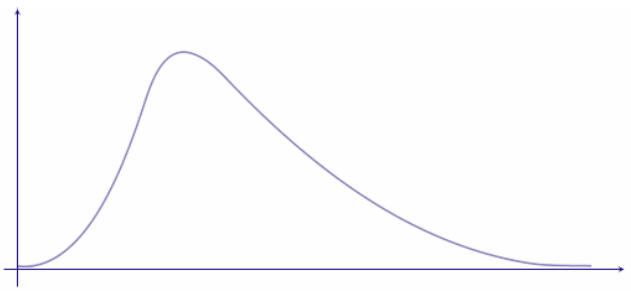


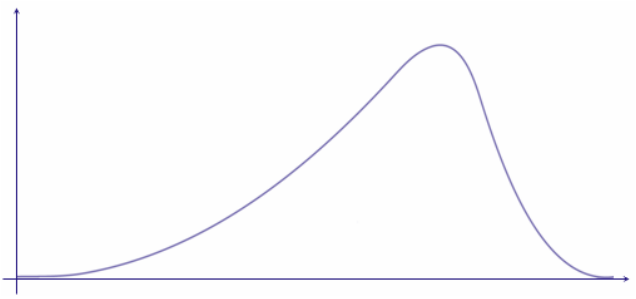
As curvas de frequências podem assumir as seguintes formas características:

a) curvas em forma de sino: são curvas que apresentam concentração de valores em torno da região central da distribuição. Tais curvas podem ser simétricas ou assimétricas. Quando assimétricas, as curvas ainda podem apresentar uma cauda mais alongada à esquerda (assimetria à esquerda) ou mais alongada à direita (assimetria à direita). Assim, as possíveis configurações para as curvas em formas de sino são:



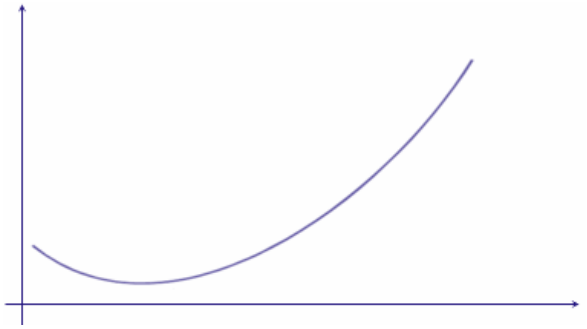
Curva Simétrica.



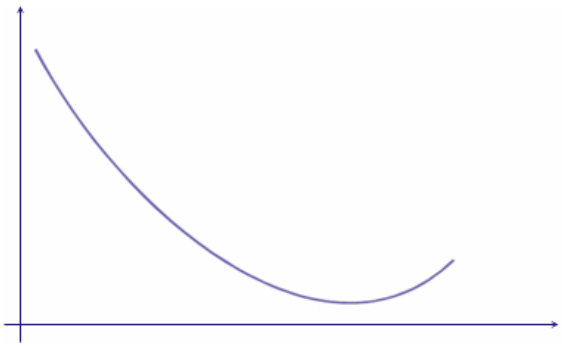


Curva assimétrica à direita. **Curva assimétrica à esquerda.**

b) curvas em forma de jota: são curvas que apresentam o ponto de ordenada máxima em uma das extremidades, representando distribuições extremamente assimétricas. As possíveis configurações são:

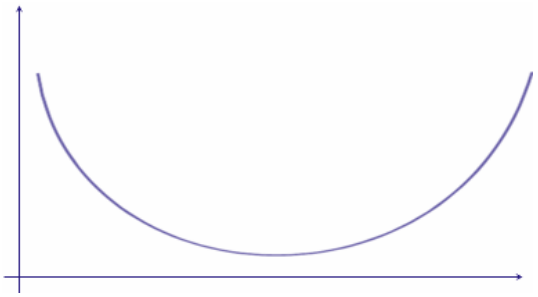


Curva em formato de J.



Curva em formato de J invertido.

c) curvas em forma de U: são curvas que apresentam as ordenadas máximas em ambas as extremidades.



Curva em formato de U.

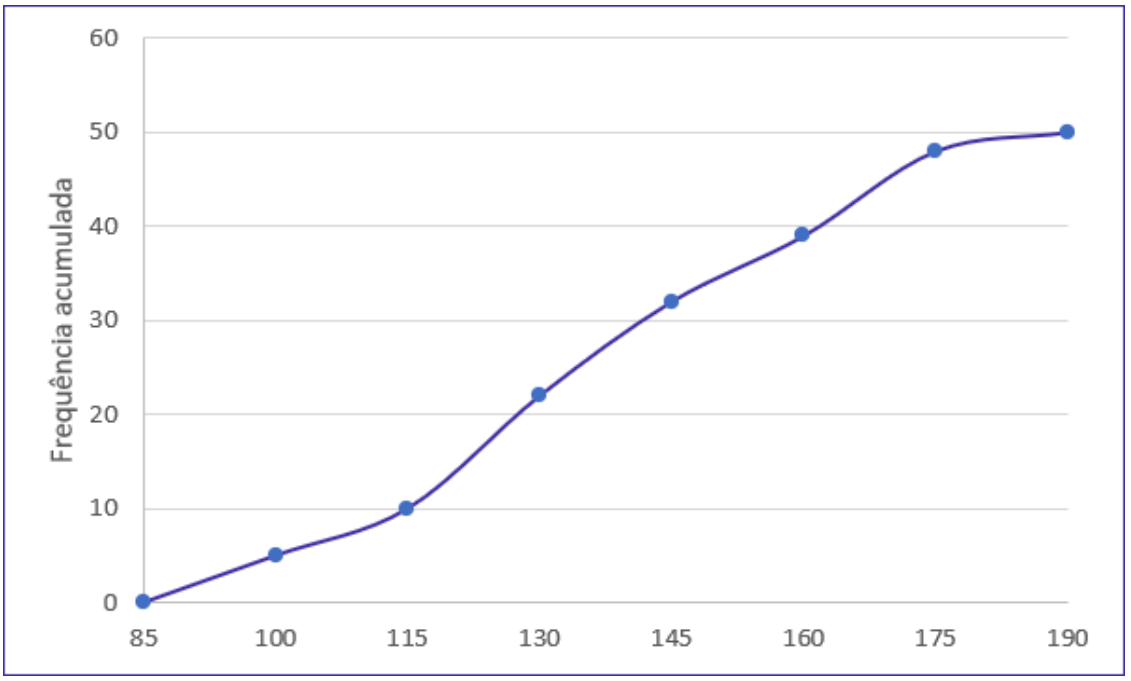
Ogiva (Polígono de Frequências Acumuladas)

O gráfico de ogiva corresponde a um polígono de frequências acumuladas. Esse gráfico é empregado na representação de distribuições de frequências acumuladas, sejam elas crescentes

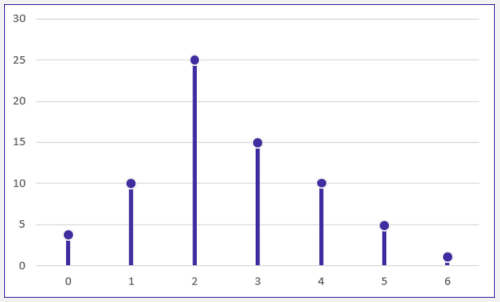
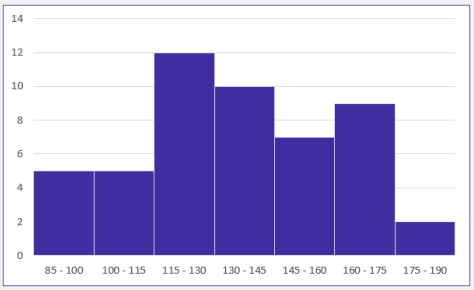
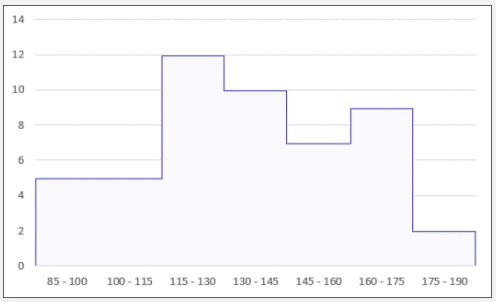
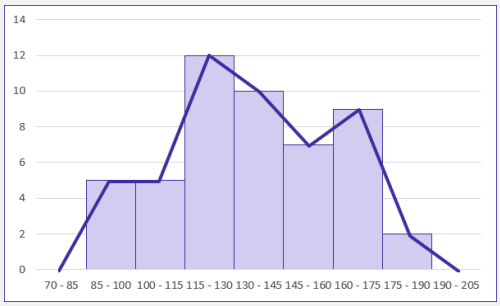
ou decrescentes. No eixo horizontal, colocamos as extremidades de cada classe e no eixo vertical as frequências acumuladas.

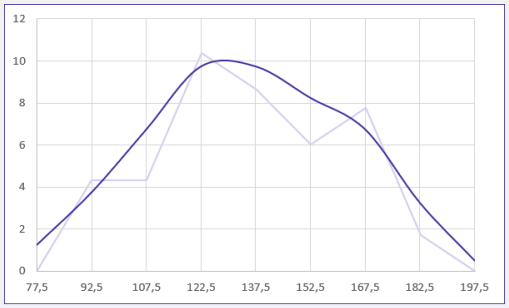
Ao contrário do polígono de frequências, a ogiva utiliza os pontos extremos das classes, e não os pontos médios. Na construção do polígono de frequências acumuladas, devemos considerar a existência de uma classe anterior à primeira, com frequência nula.

Tempo médio (X_i)	Frequência (f_i)	Frequência Acumulada (f_{ac})
$70 \leq x < 85$	0	0
$85 \leq x < 100$	5	5
$100 \leq x < 115$	5	10
$115 \leq x < 130$	12	22
$130 \leq x < 145$	10	32
$145 \leq x < 160$	7	39
$160 \leq x < 175$	9	48
$175 \leq x < 190$	2	50

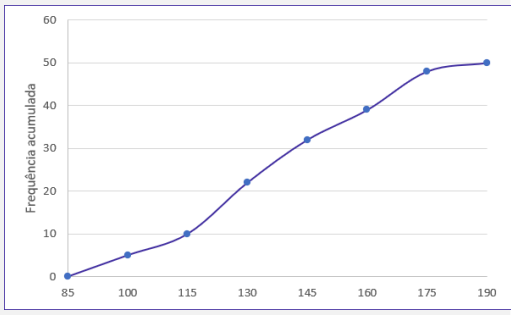


RESUMINDO

Gráfico	Definição
	O **gráfico de hastes ou bastões** é muito utilizado para representar dados não agrupados em classes, o que normalmente ocorre com **dados discretos**.
	O **histograma** é um gráfico destinado a representar **dados agrupados em classe**, sendo composto por um conjunto de **retângulos contíguos (justapostos).**
	A **poligonal característica** é construída utilizando apenas os **contornos do histograma**.
	O **polígono de frequências** é um gráfico em linha **obtido por meio da ligação, por segmentos de reta, dos pontos médios** das bases superiores dos retângulos de um histograma.

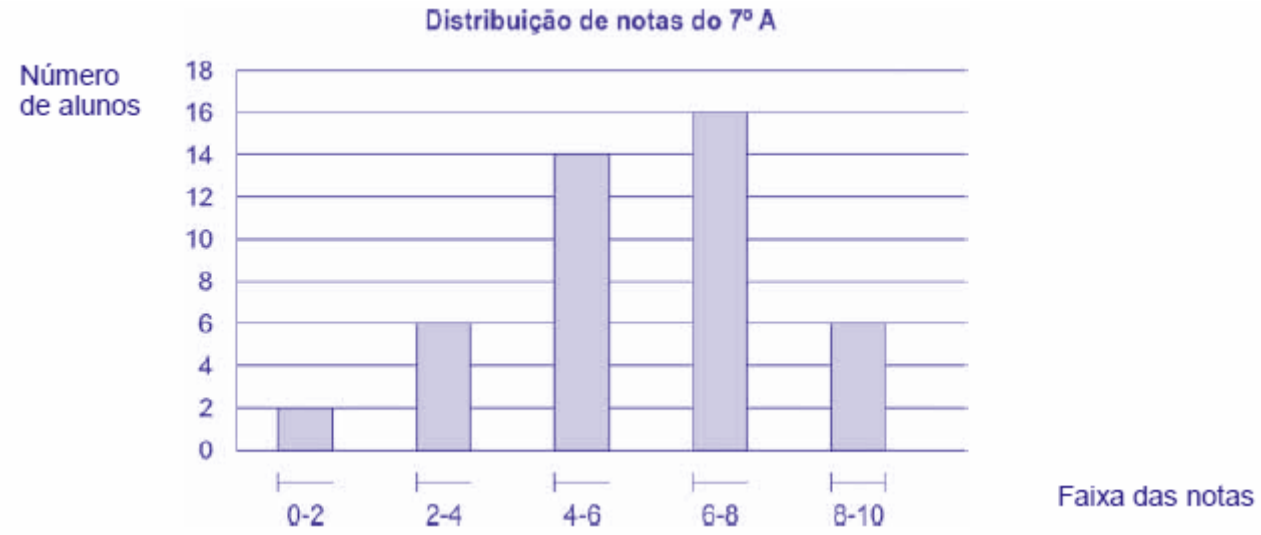


A **curva de frequências** é obtida a partir do **polimento de um polígono de frequências.**



O **gráfico de ogiva** é empregado na **representação de distribuições de frequências acumuladas**, sejam elas crescentes ou decrescentes

(FCC/SEDU-ES/2018) Depois da aplicação de uma prova para todos os alunos do 7º A, a professora Marli fez um histograma com a distribuição das notas, conforme indicado abaixo.



Se a média de notas da sala foi igual a seis, a porcentagem dos alunos que tiraram nota maior ou igual a média da sala nessa prova foi de

a) 54%.

b) 50%.

c) 56%.

d) 60%.

e) 52%.

Comentários:

Vamos, incialmente, calcular o total de alunos:

$$2 + 6 + 14 + 16 + 6 = 44$$

Reparem na notação intervalar adotada para as colunas, vamos relembrar o significado dela:

Tipo de Intervalo	Notação matemática	Notação estatística	Significado
Intervalo fechado à esquerda e aberto à direita	$a \leq x < b$	$a \vdash b$	Engloba todos os elementos entre a e b, inclusive a mas não b.
Intervalo fechado	$a \leq x \leq b$	$a \dashv\vdash b$	Engloba todos os elementos entre a e b, inclusive a e b.

Agora, com base nas informações do gráfico, percebemos que existem 22 pessoas com média maior ou igual a 6, representadas pelas duas últimas colunas do gráfico. Assim:

$$16 + 6 = 22$$

Logo, a porcentagem será de:

$$\frac{22}{44} = 0,5 = 50\%$$

Gabarito: B.

(FCC/TRT 14ª Região/2018) De um histograma e uma tabela de frequências absolutas, elaborados para analisar a distribuição dos salários dos empregados em uma empresa, obtém-se a informação que 24 empregados ganham salários com valores pertencentes ao intervalo (2.000; 4.000], em reais, que apresenta uma densidade de frequência de $0,75 \times 10^{-4} (R\$)^{-1}$

Densidade de frequência de um intervalo é o resultado da divisão da respectiva frequência relativa pela amplitude deste intervalo. Em um intervalo do histograma que está sendo analisado, com uma amplitude de R$ 3.000,00 e uma densidade de frequência de $1 \times 10^{-4} (R\$)^{-1}$, tem-se que o correspondente número de empregados é igual a

a) 40.

b) 36.

c) 30.

d) 48.

e) 42.

Comentários:

Vamos destacar os dados da questão:

- densidade de frequência: $0,75\times10^{-4}(R\$)^{-1}$;
- amplitude da classe: $4000 - 2000 = 2000$;
- número de funcionários na classe: 24; e
- total de funcionários na classe: T.

A frequência relativa é calculada pela fórmula a seguir:

$$Frequência\ relativa = \frac{Nº\ de\ funcionários\ na\ classe}{Total\ de\ funcionários} = \frac{24}{T}$$

Então, pela fórmula da densidade de frequência, teremos:

$$Densidade\ de\ frequência = \frac{Frequência\ relativa}{Amplitude}$$

$$0,75\times10^{-4} = \frac{\frac{24}{T}}{2000}$$

$$0,15 = \frac{24}{T}$$

$$T = 160$$

Agora, vamos aplicar a mesma metodologia para a classe do histograma com amplitude de 3000 e densidade de frequência de 1×10^{-4}. Considerando o número de funcionários na classe igual a N, teremos a seguinte frequência relativa:

$$Frequência\ relativa = \frac{Nº\ de\ funcionários\ na\ classe}{Total\ de\ funcionários} = \frac{N}{160}$$

Então, pela fórmula da densidade de frequência, teremos:

$$Densidade\ de\ frequência = \frac{Frequência\ relativa}{Amplitude}$$

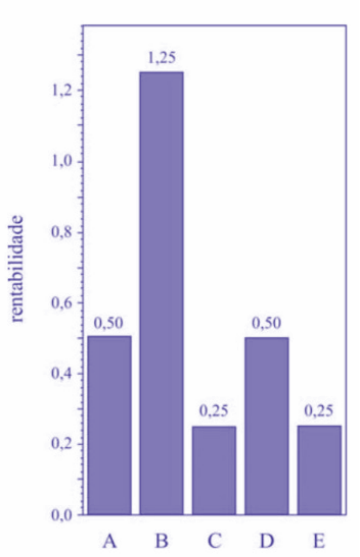
$$10^{-4} = \frac{\frac{N}{160}}{3000}$$

$$0,3 = \frac{N}{160}$$

$$N = 48$$

Gabarito: D.

(CESPE/FUNPRESP/2016)



O gráfico ilustra cinco possibilidades de fundos de investimento com suas respectivas rentabilidades. Considerando que as probabilidades de investimento para os fundos A, B, C e D sejam, respectivamente, P(A) = 0,182; P(B) = 0,454; P(C)= 0,091; e P(D)= 0,182, julgue o item subsequente.

O gráfico apresentado é um histograma.

Comentários:

O gráfico não é um histograma por dois motivos: primeiro porque há uma separação entre as colunas, o que não ocorre em um histograma, e sim em um gráfico de colunas; segundo porque um histograma representa dados que estão agrupados em intervalos de classe, e não em categorias, como é o caso.

Gabarito: Errado.

Outros Gráficos e Representações

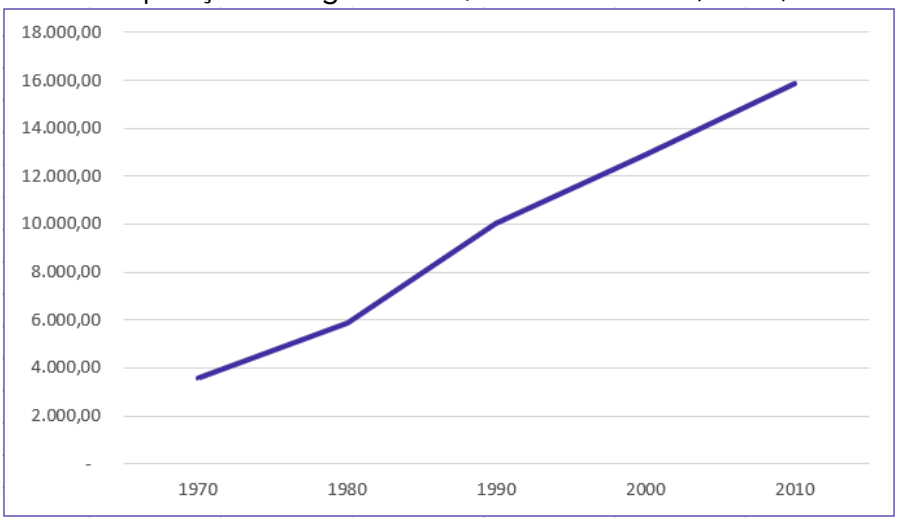
O principal objetivo dos gráficos estatísticos é proporcionar uma visualização mais rápida dos dados estatísticos ou do fenômeno sob investigação. A seguir vamos ver as principais formas de representação de dados estatísticos.

Gráficos em Linhas

Os gráficos em linha normalmente são usados para representar dados de séries temporais, com a finalidade de mostrar a variação dos valores de uma variável ao longo do tempo. Esse tipo de gráfico permite-nos comparar duas variáveis: uma é traçada no eixo x (horizontal) e a outra no eixo y (vertical). O eixo y geralmente indica uma quantidade, enquanto o eixo x representa uma unidade de tempo.

Por exemplo, o gráfico a seguir mostra a evolução da população residente na Região Norte do Brasil no período de 1970 a 2010. Veja que o eixo horizontal está indicando o decurso do tempo, enquanto o eixo vertical apresenta a população residente na região.

População da região Norte, de 1970 a 2010 (x1000)

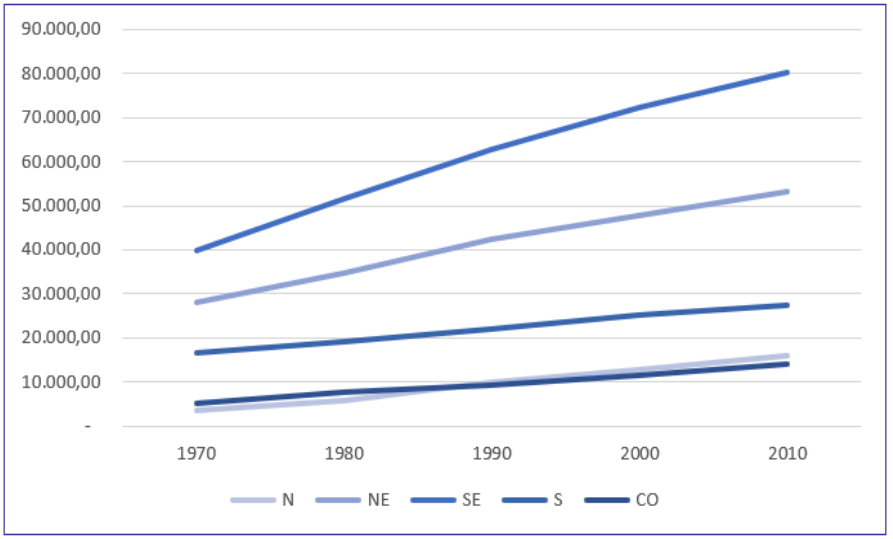


Fonte: Censo Demográfico 1970/2010 (IBGE)

TOME NOTA!

Também podemos elaborar um gráfico de linhas múltiplas para comparar a evolução da população residente nas Grandes Regiões do Brasil em diferentes períodos:

População brasileira, por Grandes Regiões, de 1970 a 2010 (x1000)



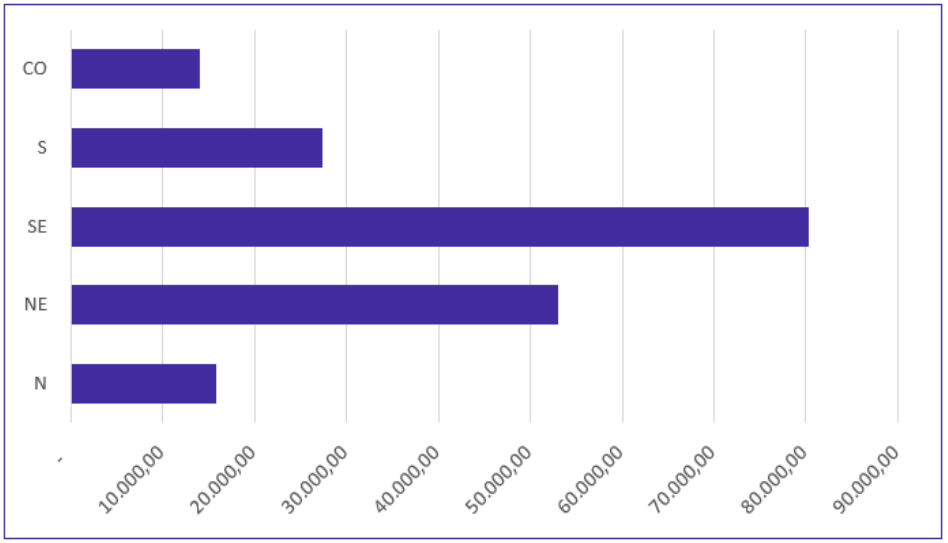
Fonte: Censo Demográfico 1970/2010 (IBGE)

Gráficos em Barras

Os gráficos em barra normalmente são usados para representar distribuições de dados categóricos ou qualitativos. Uma série estatística é representada por um conjunto de retângulos dispostos horizontalmente, cada um indicando uma categoria particular, os quais possuem a mesma altura e comprimentos proporcionais aos respectivos dados.

Por exemplo, a distribuição da população residente em cada região brasileira, no ano de 2010, pode ser representada por meio do seguinte gráfico:

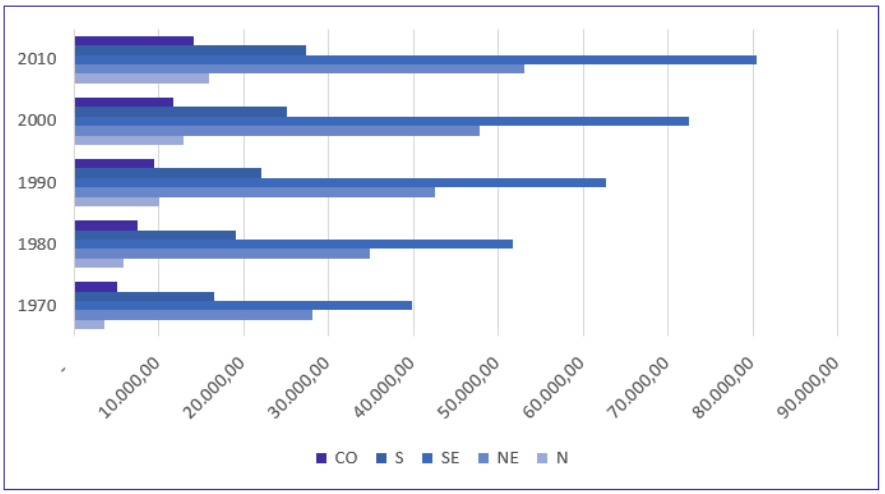
População brasileira, por região, em 2010 (x1000)



Fonte: Censo Demográfico 1970/2010 (IBGE)

Adicionalmente, podemos utilizar um gráfico de barras justapostas para representar a evolução da população residente em cada região brasileira, como o mostrado a seguir:

População brasileira, por região, de 1970 a 2010 (x1000)



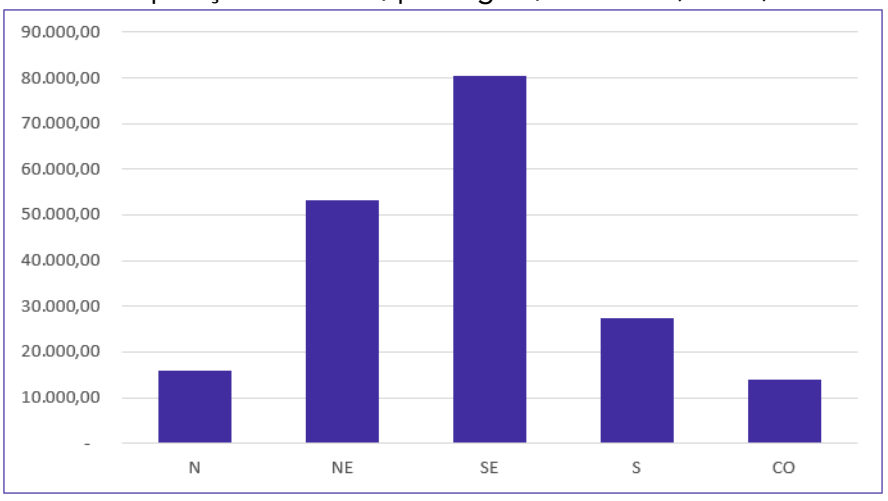
Fonte: Censo Demográfico 1970/2010 (IBGE)

Gráficos em Colunas

Os gráficos em coluna também são usados para distribuições de dados categóricos ou qualitativos. A diferença básica é que, agora, uma série estatística é representada por um conjunto de retângulos dispostos verticalmente, cada um indicando uma categoria particular, todos com a mesma largura e alturas proporcionais aos respectivos dados.

Por exemplo, a distribuição da população residente em cada região brasileira, no ano de 2010, pode ser representada por meio do seguinte gráfico:

População brasileira, por região, em 2010 (x1000)

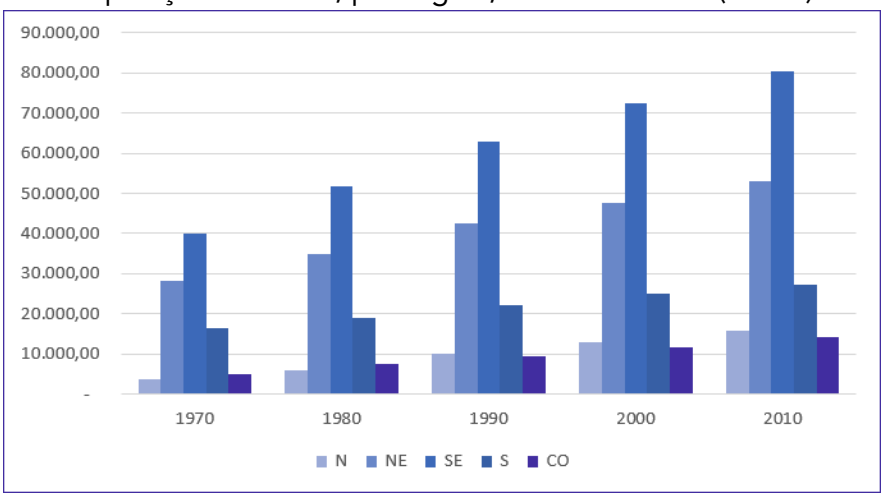


Fonte: Censo Demográfico 1970/2010 (IBGE)

Também podemos utilizar um gráfico de colunas justapostas para representar a evolução da população residente em cada região brasileira, no período de 1970 a 2010. Dessa maneira,

conseguimos apresentar mais informações em um espaço consideravelmente menor. Vejamos o gráfico a seguir:

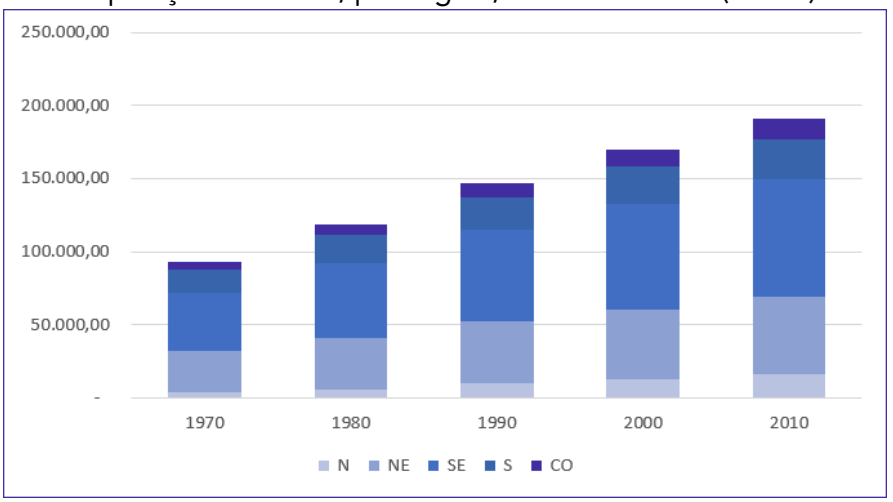
População brasileira, por região, de 1970 a 2010 (x1000)



Fonte: Censo Demográfico 1970/2010 (IBGE)

Adicionalmente, essas informações também podem ser representadas por meio de um gráfico de colunas sobrepostas (ou gráfico de colunas empilhadas). Esse tipo de gráfico é considerado uma extensão do formato tradicional, pois permite analisarmos duas dimensões de uma variável categórica, em vez de apenas uma. Cada coluna é dividida em várias partes que ficam empilhadas umas sobre as outras, cada uma correspondendo a um nível da segunda variável categórica.

População brasileira, por região, de 1970 a 2010 (x1000)



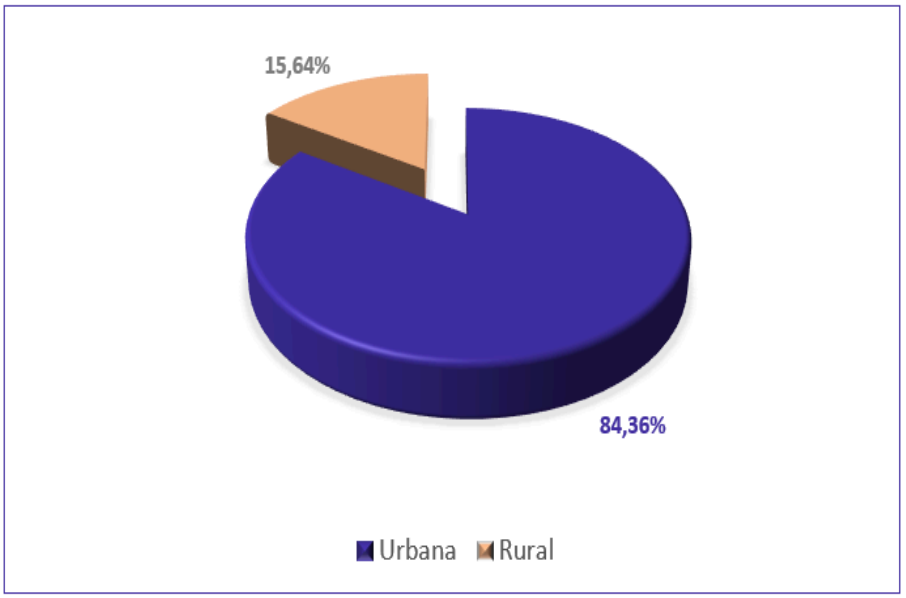
Fonte: Censo Demográfico 1970/2010 (IBGE)

Gráfico em Setores

O gráfico em setores (também conhecido como gráfico de pizza) é usado para representar a frequência relativa (porcentagem) de uma variável categórica. Ele é formado por um círculo dividido em setores circulares, cada um representando uma categoria, cujos ângulos centrais são proporcionais às frequências relativas da categoria.

Para a construção do gráfico de setores, utilizaremos uma regra de três simples, em que as frequências relativas de cada categoria correspondem ao ângulo central que desejamos representar em relação à frequência total, que corresponde ao ângulo de 360°.

População Urbana e Rural do Brasil em 2010

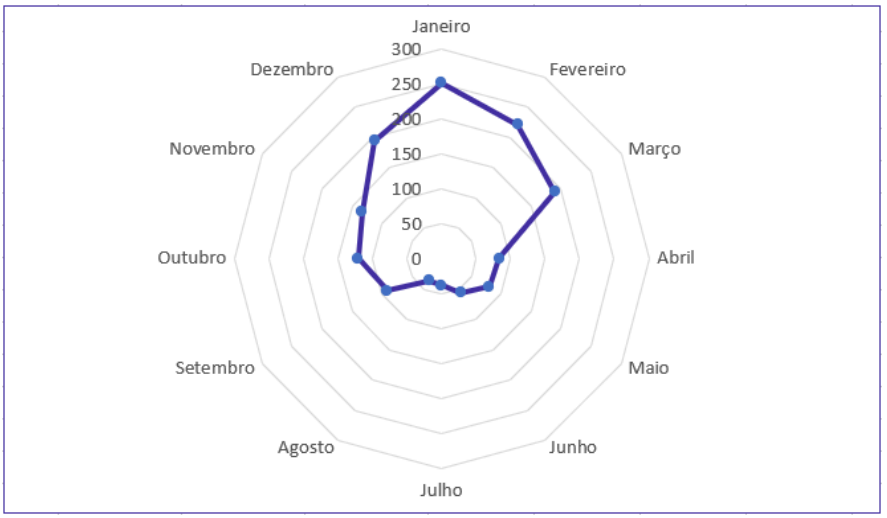


Fonte: Sinopse do Censo Demográfico de 2010 (IBGE)

Gráfico Polar

O gráfico polar consiste em uma sequência de eixos igualmente espaçados (ângulos iguais), cada um representando uma das variáveis. Uma linha é desenhada ligando os valores de cada eixo. Esse tipo de gráfico é usado para representar séries temporais cíclicas, que apresentam uma determinada periodicidade, como é o caso da precipitação pluviométrica mensal média:

Precipitação pluviométrica média do município de São Paulo

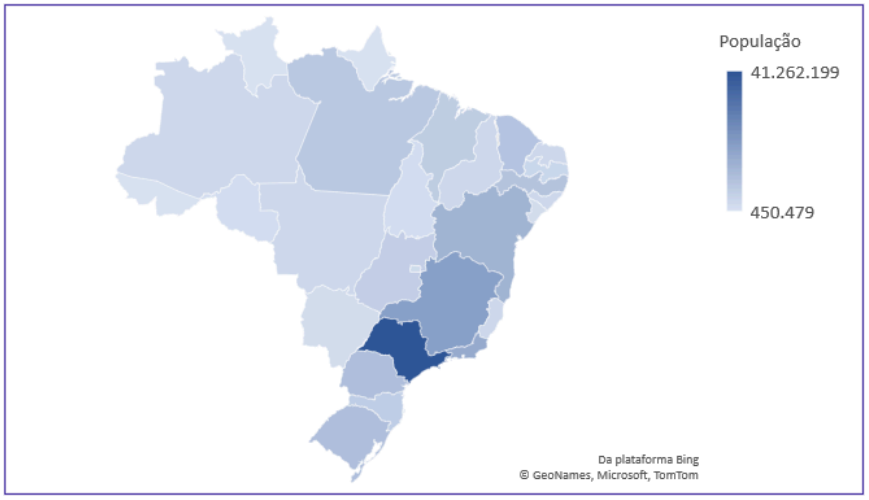


Fonte: Atlas Pluviométrico do Brasil (Serviço Geológico do Brasil)

Cartograma

O cartograma é empregado com a finalidade de apresentar dados estatísticos diretamente relacionados com áreas geográficas. As áreas do cartograma podem ser preenchidas por pontos, hachuras ou cores. O significado do preenchimento será indicado em uma legenda. Vejamos a população residente em cada Estado brasileiro:

População residente no Brasil por Estado em 2010 (x1000)

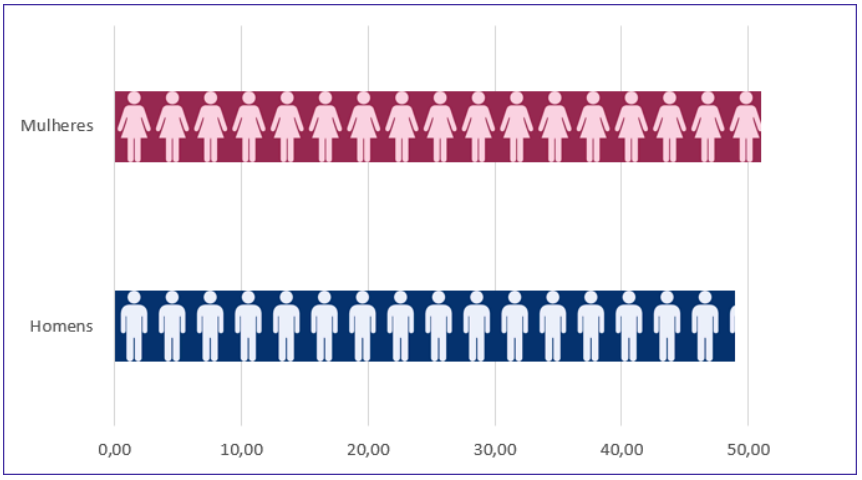


Fonte: Sinopse do Censo Demográfico de 2010 (IBGE)

Pictograma

O pictograma substitui valores por ícones, tornando os dados mais atraentes e facilitando o entendimento acerca de um determinado fenômeno. Normalmente, uma legenda é utilizada para indicar o que cada ícone representa. Os ícones devem possuir o mesmo tamanho, mas podem aparecer fracionados para mostrar a respectiva fração de uma determinada quantidade. A proporção de homens e mulheres na população brasileira é apresentada no pictograma a seguir:

População residente no Brasil por Estado em 2010 (x1000)

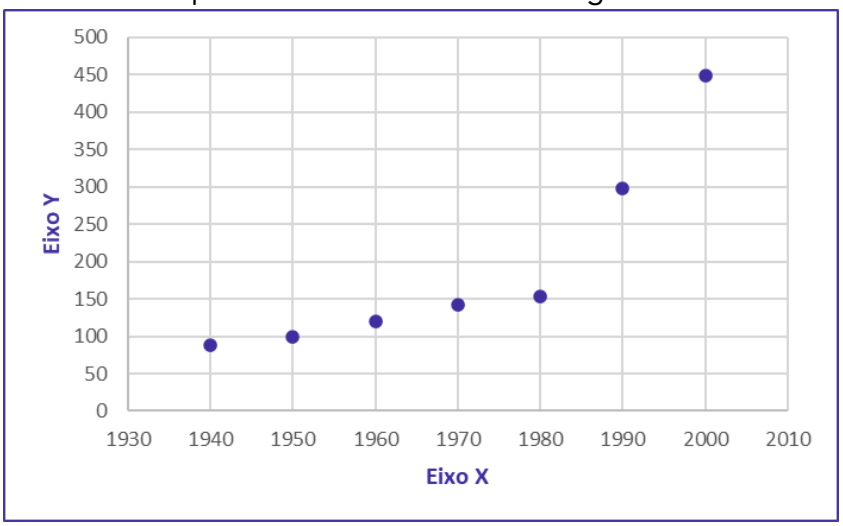


Fonte: Sinopse do Censo Demográfico de 2010 (IBGE)

Gráfico de Dispersão

O gráfico de dispersão é uma representação de pares ordenados em um plano cartesiano, composto por um eixo vertical (ordenada) e um eixo horizontal (abcissa). Os dados são representados como uma coleção de pontos, cada um com o valor de uma variável determinando a posição no eixo horizontal e o valor da outra variável determinando a posição no eixo vertical. É uma ferramenta poderosa para estudar a relação entre duas variáveis.

Municípios criados e instalados na região Norte.



Fonte: Sinopse do Censo Demográfico de 2010 (IBGE)

Diagrama de Ramos e Folhas

O diagrama de ramos e folhas fornece uma maneira rápida de representar graficamente a distribuição dos dados. Nesse diagrama, cada número é separado em duas partes. Em geral, de um lado ficam as unidades do número e do outro lado fica o restante desse número.

Consideremos o seguinte rol crescente:

85, 89, 96, 98, 99, 103, 104, 105, 113, 114, 115, 115, 123, 123, 124, 126, 126, 126, 127, 128, 129, 129, 134, 135, 135, 135, 137, 137, 137, 142, 143, 143, 148, 153, 154, 155, 157, 158, 159, 161, 161, 165, 168, 170, 171, 171, 171, 173, 175, 175

A representação utilizando um diagrama de ramos e folhas ficaria assim:

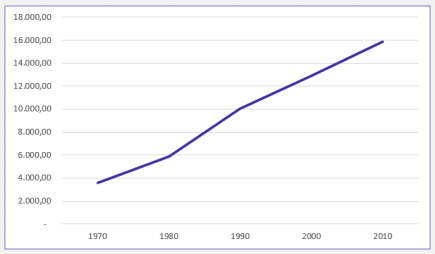
8	5 9	Chave: 8 | 5 = 85
9	6 8 9	
10	3 4 5	
11	3 4 5 5	
12	3 3 4 6 6 6 7 8 9 9	
13	4 5 5 5 7 7 7	
14	2 3 3 8	
15	3 4 5 7 8 9	
16	1 1 5 8	
17	0 1 1 1 3 5 5	

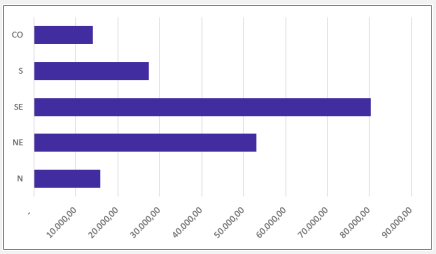
Repare que, no lado esquerdo, temos as centenas e as dezenas representando os ramos. Por sua vez, no lado direito, temos as unidades representando as folhas. As folhas, portanto, estão vinculadas aos ramos. Dessa maneira, a chave "9 | 6 8 9" significa que, no rol original, havia os números 96, 98 e 99.

Também é comum encontrarmos diagramas de ramos e folhas em que as unidades são separadas em dois grupos: de 0 a 4 e de 5 a 9. Nesse caso, teríamos o seguinte diagrama:

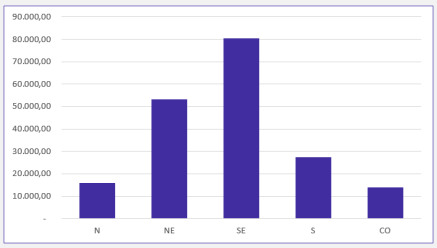
8	Chave: 8 | 5 = 85
8	5 9
9	
9	6 8 9
10	3 4
10	5
11	3 4
11	5 5
12	3 3 4
12	6 6 6 7 8 9 9
13	4
13	5 5 5 7 7 7
14	2 3 3
14	8
15	3 4
15	5 7 8 9
16	1 1
16	5 8
17	0 1 1 1 3
17	5 5

Por fim, é importante observarmos que não existe uma regra única para a construção do diagrama de ramos e folhas. O formato mais comumente encontrado é o que separa o número em duas partes, porém, a depender da chave escolhida, o número pode ser separado em mais partes. Assim, ao resolver questões que envolvem esse tipo de diagrama, devemos observar a chave adotada.

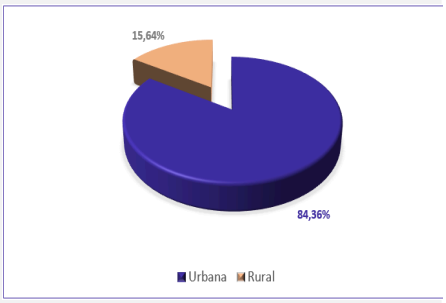
Gráfico	Definição
	Os **gráficos em linha** normalmente são usados para representar a variação dos valores de uma variável ao longo do tempo. Esse tipo de gráfico permite-nos comparar duas variáveis: uma é traçada no eixo x (horizontal) e a outra no eixo y (vertical).



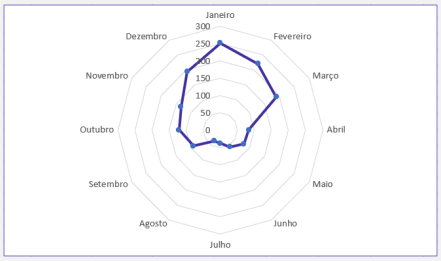
Os **gráficos em barra** normalmente são usados para representar distribuições de dados categóricos ou qualitativos. Uma série estatística é representada por um conjunto de retângulos dispostos **horizontalmente**, cada um indicando uma categoria particular.



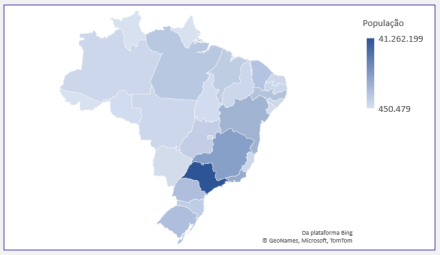
Os **gráficos em coluna** também são usados para distribuições de dados categóricos ou qualitativos. A diferença básica é que, agora, uma série estatística é representada por um conjunto de retângulos dispostos **verticalmente**, cada um indicando uma categoria particular.



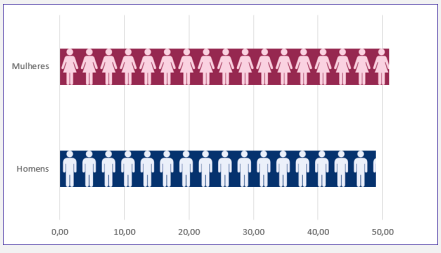
O **gráfico em setores** é usado para representar a frequência relativa (porcentagem) de uma variável categórica, sendo formado por um círculo dividido em setores circulares, cada um representando uma categoria.



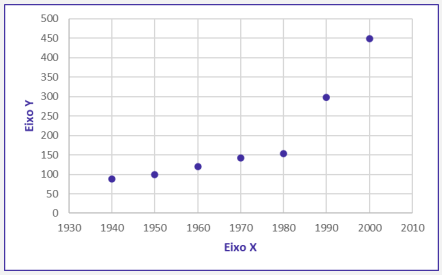
O **gráfico polar** consiste em uma sequência de eixos igualmente espaçados (ângulos iguais), cada um representando uma das variáveis. Uma linha é desenhada ligando os valores de cada eixo.



O **cartograma** é empregado com a finalidade de apresentar dados estatísticos diretamente relacionados com áreas geográficas.



O **pictograma** substitui valores por ícones, tornando os dados mais atraentes e facilitando o entendimento acerca de um determinado fenômeno.



O **gráfico de dispersão** é uma representação de pares ordenados em um plano cartesiano, composto por um eixo vertical (ordenada) e um eixo horizontal (abcissa). Os dados são representados como uma coleção de pontos.

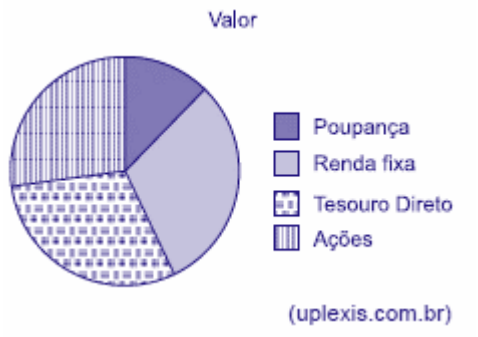
8	5	9								Chave: 8 | 5 = 85
9	6	8	9							
10	3	4	5							
11	3	4	5	5						
12	3	3	4	6	6	6	7	8	9	9
13	4	5	5	5	7	7	7			
14	2	3	3	8						
15	3	4	5	7	8	9				

O **diagrama de ramos e folhas** fornece uma maneira rápida de representar graficamente a distribuição dos dados. Nele, cada número é separado em duas partes. Em geral, de um lado ficam as unidades do número e do outro lado fica o restante desse número.

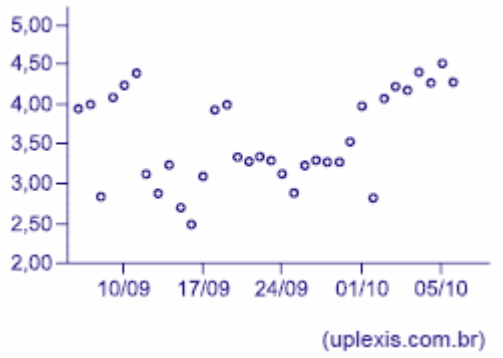
HORA DE PRATICAR!

(VUNESP/UNICAMP/2019) Assinale dentre os exemplos a seguir, o gráfico de dispersão.

a)

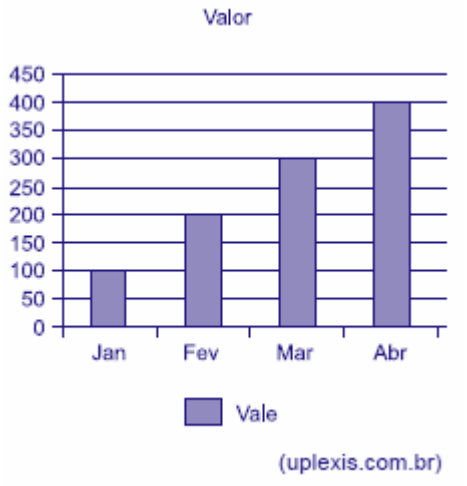


b)

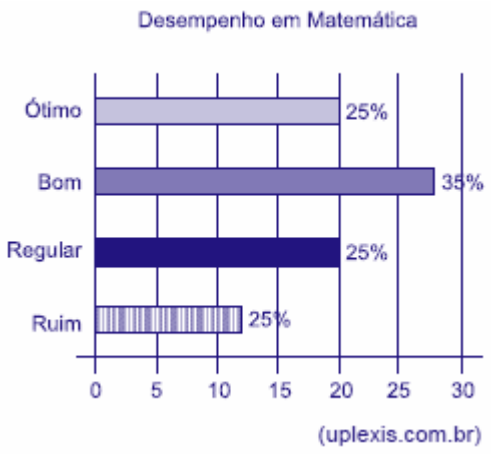


c)

d)



e)



Comentários:

Analisando cada gráfico, temos:

Letra A: Alternativa Errada. O gráfico de setores é formado por um círculo dividido em setores circulares, cada um representando uma categoria, cujos ângulos centrais são proporcionais às frequências relativas da categoria.

Letra B: Alternativa Correta. O gráfico de dispersão é uma representação gráfica que analisa a relação entre duas variáveis quantitativas — uma de causa e outra de efeito, chamadas de variáveis independente e dependente, respectivamente. Esse tipo de diagrama traz números simultâneos das duas variáveis, deixando visível se o que acontece em uma variável interfere na outra.

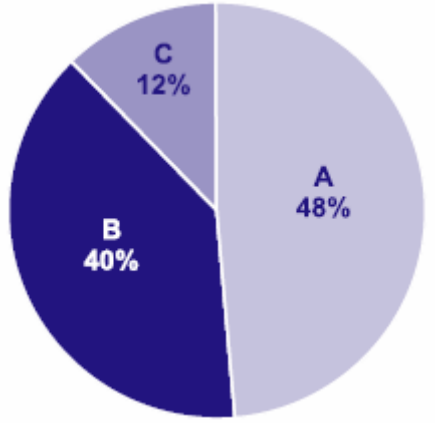
Letra C: Alternativa Errada. O gráfico em linhas é o gráfico em que os pontos são geralmente usados para controlar alterações ao longo do tempo e para facilitar a identificação de tendências ou de anomalias.

Letra D: Alternativa Errada. Os gráficos em colunas assim como os gráficos de barras são usados para distribuições de dados categóricos ou qualitativos. A diferença básica é que, no primeiro, uma série estatística é representada por um conjunto de retângulos dispostos verticalmente, cada um indicando uma categoria particular, todos com a mesma largura e alturas proporcionais aos respectivos.

Letra E: Alternativa Errada. Os gráficos em barras normalmente são usados para representar distribuições de dados categóricos ou qualitativos. Uma série estatística é representada por um conjunto de retângulos dispostos horizontalmente, cada um indicando uma categoria particular, os quais possuem a mesma altura e comprimentos proporcionais aos respectivos dados.

Gabarito: B.

(VUNESP/Pref. Campinas/2019) Uma empresa atua em três segmentos de mercado, A, B e C. O gráfico de setores mostra a distribuição percentual, por segmento, da receita total obtida por essa empresa em 2018.



Sabendo-se que a receita obtida no segmento A superou a receita obtida no segmento B em R$ 64 milhões, é correto afirmar que a receita obtida no segmento C foi igual a

a) R$ 98 milhões.

b) R$ 96 milhões.

c) R$ 94 milhões.

d) R$ 88 milhões.

e) R$ 86 milhões.

Comentários:

Consideremos T o total de receitas obtidas pela empresa em 2018. Pelas informações apresentadas no gráfico de pizza, temos:

$$A = 0,48 \times T$$

$$B = 0,4 \times T$$

$$C = 0,12 \times T$$

Sabendo que a receita obtida no segmento A superou a receita obtida no segmento B em R$ 64 milhões, então:

$$A = B + 64$$

$$0,48 \times T = 0,4 \times T + 64$$

$$0,8 \times T = 64$$

$$T = 800\ milhões$$

Logo,

$$C = 0,12 \times T$$

$$C = 12 \times 800$$

$$C = 96\ milhões$$

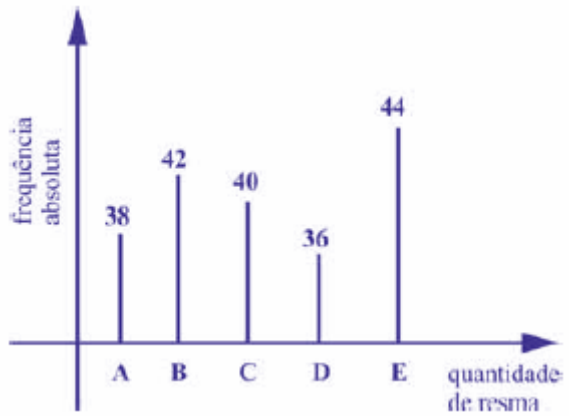
Gabarito: B.

(CESPE/CBM-AL/2017) Na tabela a seguir, A, B, C, D e E são as quantidades de resmas de papel A4 consumidas, em quatro meses, pelas seções administrativas I, II, III, IV e V, respectivamente. Apesar de não mostrar explicitamente essas quantidades, a tabela apresenta as frequências absolutas e (ou) relativas de algumas dessas quantidades.

Seção	Quantidades de Resmas	Frequência Absoluta	Frequência Relativa
I	A	38	19%
II	B		
III	C		20%
IV	D	36	
V	E	44	
	Total		**100%**

Considerando que cada uma dessas resmas, juntamente com a embalagem, tem forma de um paralelepípedo retângulo reto que mede 5 cm × 21 cm × 30 cm, julgue o item seguinte.

O gráfico de barras verticais a seguir apresenta as frequências absolutas de resmas consumidas pelas cinco seções.



Comentários:

Podemos usar uma regra de três simples para completar a tabela. Primeiro, vamos encontrar o valor que representa 100%:

$$38 - 0,19$$

$$Q - 1,00$$

$$Q = \frac{38 \times 1,00}{0,19} = 200$$

Encontrando o valor de 20%:

$$200 - 1,00$$

$$C - 0,20$$

$$C = \frac{200 \times 0,2}{1,00} = 40$$

Agora, basta sabermos a frequência absoluta na seção II. Como foram consumidas 200 resmas, então o número de resmas consumidas pela seção II foi:

$$200 - 38 - 40 - 36 - 44 = 42$$

Dessa forma, nas seções I, II, III, IV e V foram consumidas, respectivamente 38, 42, 40, 36 e 44 resmas de papel A4, conforme mostra o gráfico.

Gabarito: Certo.

Resumo da Aula

INTRODUÇÃO À ESTATÍSTICA

A Estatística pode ser dividida em três grandes ramos: Estatística Descritiva (ou dedutiva), Estatística Probabilística e Estatística Inferencial (ou indutiva).

ESTATÍSTICA DESCRITIVA	ESTATÍSTICA PROBABILÍSTICA	ESTATÍSTICA INFERENCIAL
É responsável pela coleta, organização, descrição e resumo dos dados observados.	É responsável por estabelecer o modelo matemático adotado para explicar fenômenos aleatórios.	É responsável pela análise e interpretação dos dados.

CONCEITOS INICIAIS

POPULAÇÃO CONJUNTO de TODOS os elementos a serem estudados, que apresentam uma ou mais características em comum.	**CENSO** ESTUDO dos dados relativos a TODOS os ELEMENTOS de uma população.
AMOSTRA SUBCONJUNTO extraído DA POPULAÇÃO para análise, devendo ser representativo daquele grupo	**AMOSTRAGEM** PROCESSO que consiste na seleção criteriosa dos elementos a serem submetidos à investigação.
PARÂMETROS Descrições numéricas de características da POPULAÇÃO, que normalmente precisam ser estimadas.	**ESTATÍSTICAS** Medidas numéricas extraídas de AMOSTRAS representativas extraídas da população.

MÉTODO EXPERIMENTAL X MÉTODO ESTATÍSTICO

MÉTODO EXPERIMENTAL	MÉTODO ESTATÍSTICO
As CAUSAS são mantidas CONSTANTES, COM EXCEÇÃO DE UMA, que é VARIADA para que seus efeitos sejam descobertos.	Admite e REGISTRA TODAS AS POSSÍVEIS VARIAÇÕES DAS CAUSAS PRESENTES, procurando determinar a influência de cada fator no resultado.

DADOS ESTATÍSTICOS

Com relação ao número de observações coletadas, os dados são classificados em univariados, bivariados ou multivariados:

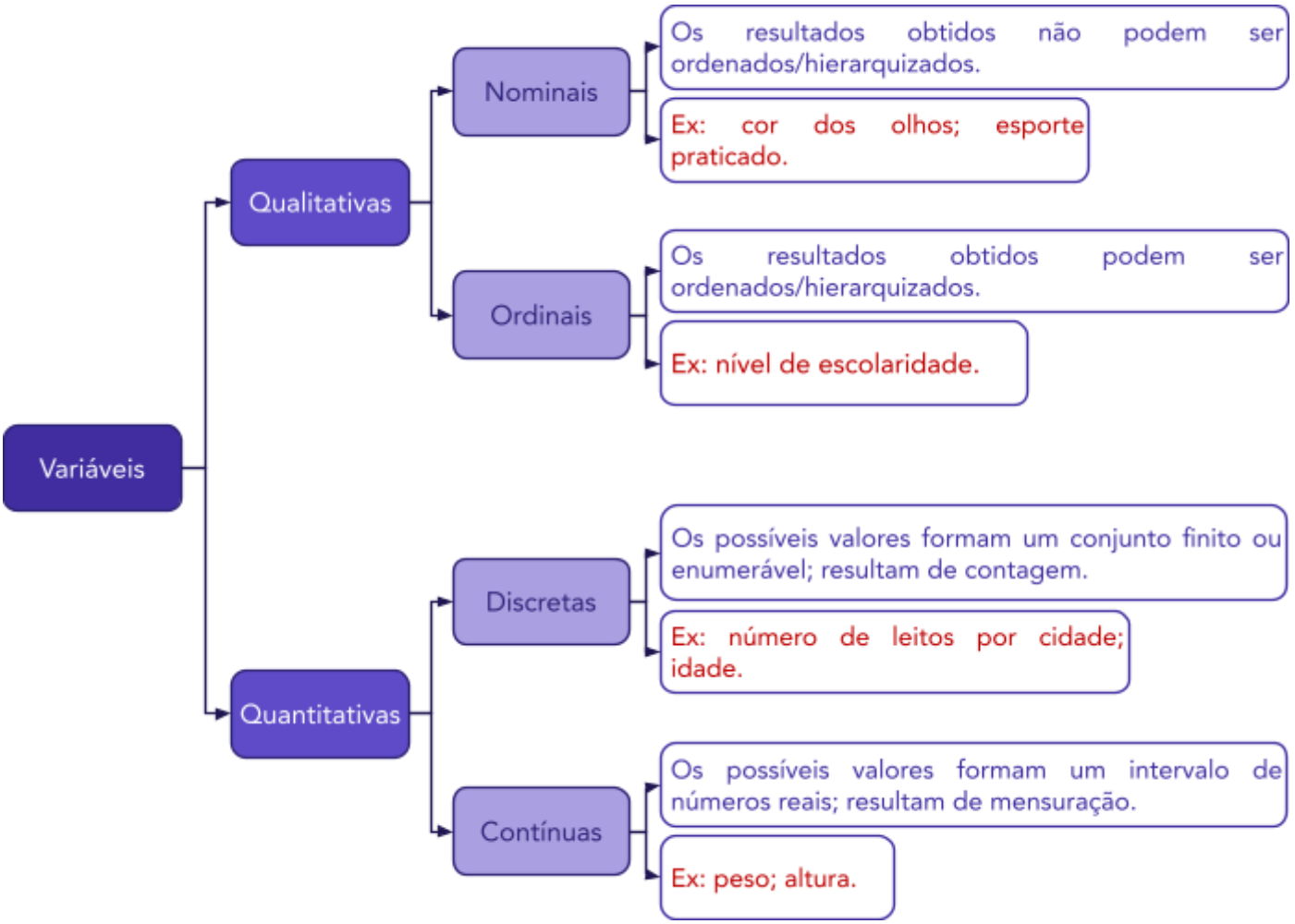
DADOS UNIVARIADOS	DADOS BIVARIADOS	DADOS MULTIVARIADOS
É quando uma única observação de cada indivíduo é registrada.	É quando duas observações de cada indivíduo são registradas.	É quando mais duas observações acerca de cada indivíduo são registradas

Quanto à forma de apresentação, os dados podem ser classificados em dados brutos ou rol.

VARIÁVEIS ESTATÍSTICAS

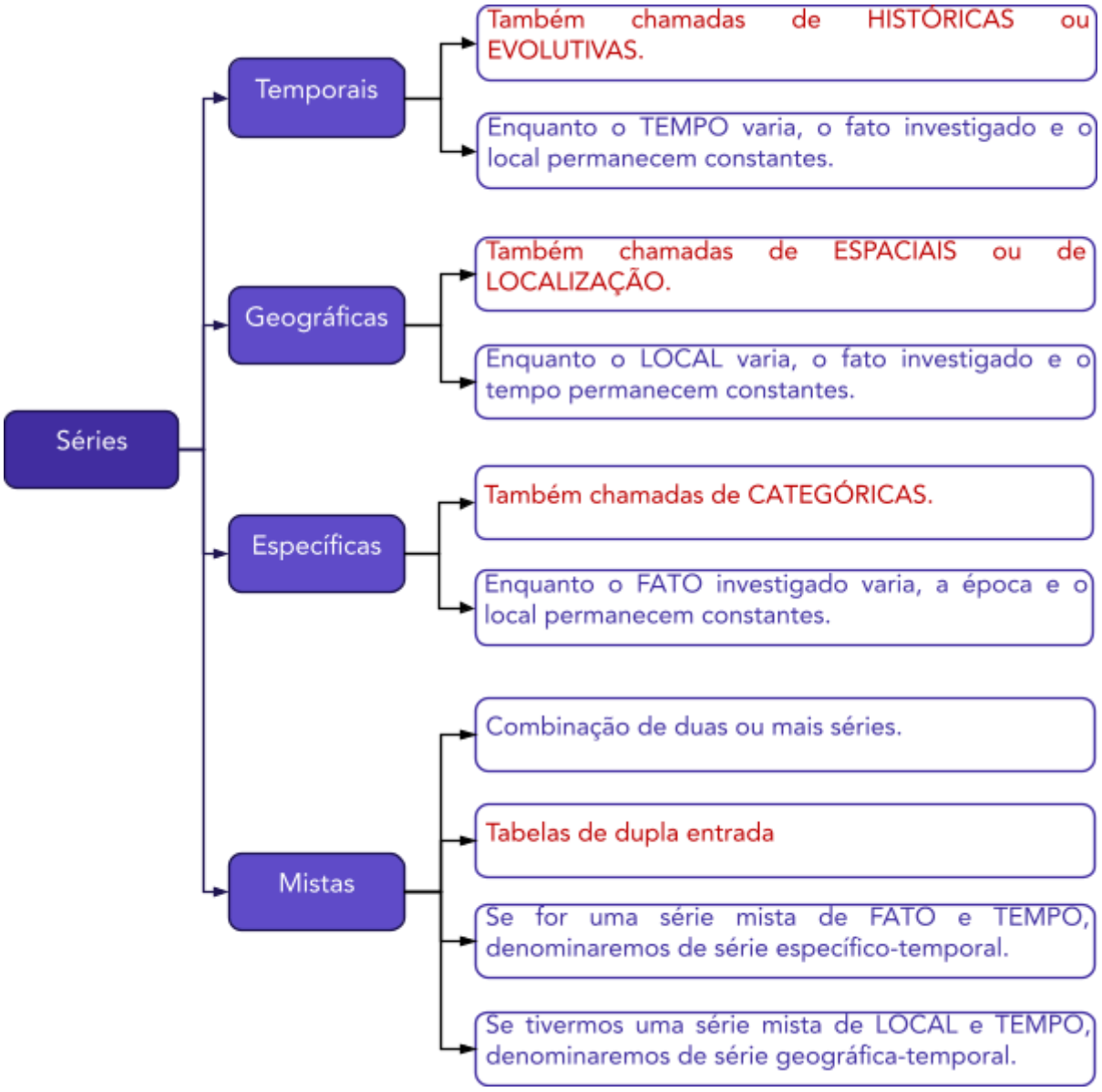
As variáveis estatísticas podem ser classificadas, inicialmente, em duas categorias: qualitativas e quantitativas.

FIQUE ATENTO!



SÉRIES ESTATÍSTICAS

As séries estatísticas podem ser classificadas em: temporais, geográficas, específicas ou mistas:



DISTRIBUIÇÃO DE FREQUÊNCIAS

As distribuições de frequências podem ser classificadas em dois tipos: distribuição de frequências pontual (ou discreta) e distribuição de frequências intervalar (ou contínua).

DISTRIBUIÇÃO DE FREQUÊNCIAS PONTUAL

São apresentados todos os dados coletados juntamente com suas respectivas frequências, não havendo perda de valores.

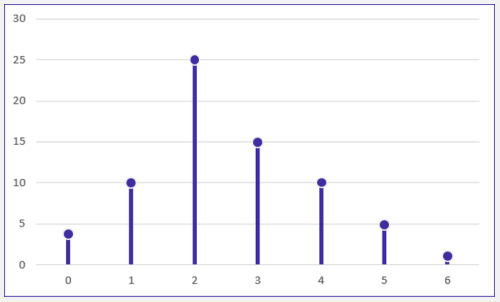
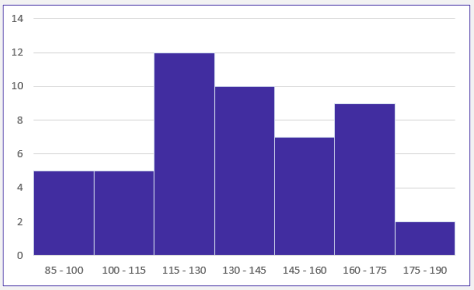
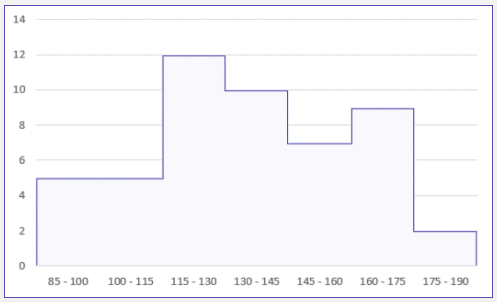
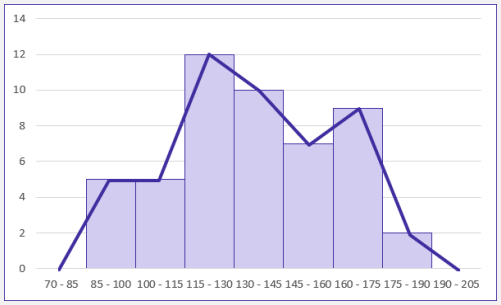
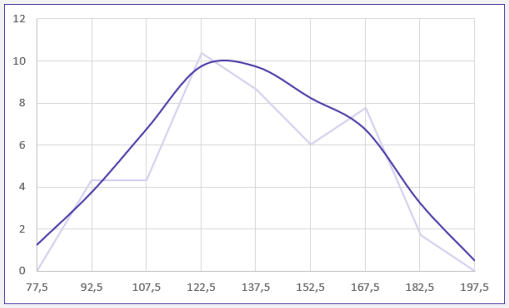
DISTRIBUIÇÃO DE FREQUÊNCIAS INTERVALAR

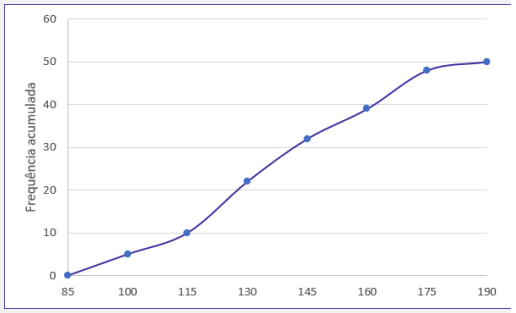
É agrupamento os valores por intervalos de classe.

Elementos de uma Distribuição de Frequências

Item	Definição	Símbolos e Fórmulas
Número de Classes	As classes são os intervalos nos quais o fenômeno é subdividido.	$k = 1 + 3,3 \times log\, n$ $k = \sqrt{n}$
Limites de Classe	Correspondem aos valores extremos.	l_{inf} e l_{sup}
Amplitude de um Intervalo de Classe	Distância entre os limites inferiores (ou superiores) de classes consecutivas.	$h = l_{sup} - l_{inf}$
Amplitude total	Diferença entre o limite superior da última classe (limite superior máximo) e o limite inferior da primeira classe (limite inferior mínimo).	$AT = l_{máx} - l_{mín}$ $AT = h \times k$
Ponto Médio	Média aritmética simples dos valores extremos de uma classe.	$PM = \frac{\left(l_{inf} + l_{sup}\right)}{2}$ $PM = l_{inf} + \frac{h}{2}$ $PM = l_{sup} - \frac{h}{2}$
Frequência Absoluta Simples	Número de observações correspondentes a uma determinada classe ou a um determinado valor.	f_i
Frequência Absoluta Acumulada	Total das frequências de todos os valores inferiores ao limite superior do intervalo de uma dada classe	$f_{ac_i} = f_1 + f_2 + f_3 + \cdots + f_i$
Frequência Relativa Simples	Proporção de dados existentes em uma determinada classe.	$F_i = \frac{f_i}{\Sigma f_i} = \frac{f_i}{n}$
Frequência Relativa Acumulada	Proporção de valores inferiores ao limite superior do intervalo de uma dada classe.	$F_{ac_i} = F_1 + F_2 + F_3 + \cdots + F_i$
Densidade de Frequência	Quociente entre a frequência da classe (absoluta ou relativa) e sua amplitude	$d = \frac{f}{h}$

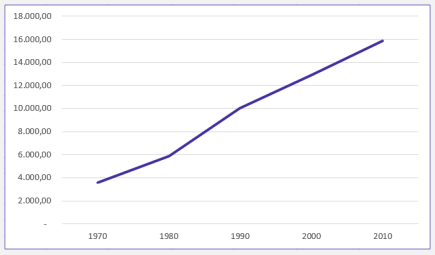
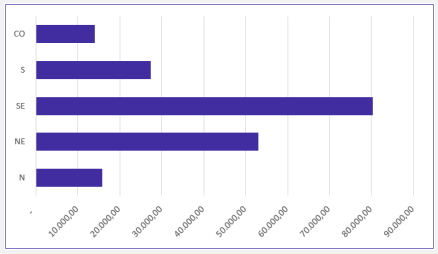
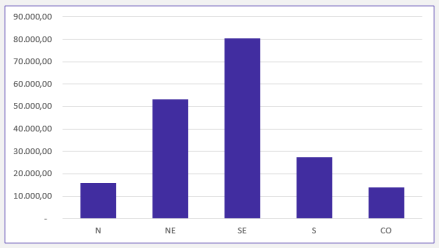
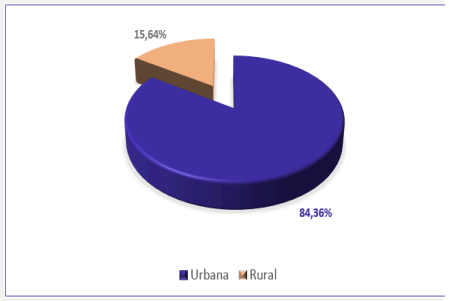
REPRESENTAÇÕES GRÁFICAS DAS DISTRIBUIÇÕES DE FREQUÊNCIA

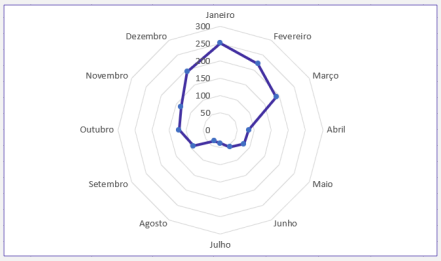
Gráfico	Definição
	O gráfico de hastes ou bastões é muito utilizado para representar dados não agrupados em classes, o que normalmente ocorre com dados discretos.
	O histograma é um gráfico destinado a representar dados agrupados em classe, sendo composto por um conjunto de retângulos contíguos (justapostos).
	A poligonal característica é construída utilizando apenas os contornos do histograma.
	O polígono de frequências é um gráfico em linha obtido por meio da ligação, por segmentos de reta, dos pontos médios das bases superiores dos retângulos de um histograma.
	A curva de frequências é obtida a partir do polimento de um polígono de frequências



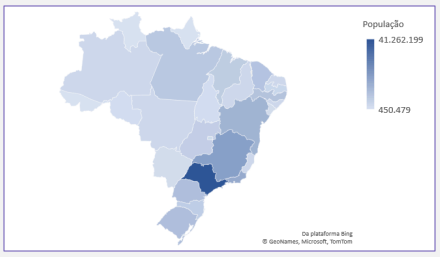
O gráfico de ogiva é empregado na representação de distribuições de frequências acumuladas, sejam elas crescentes ou decrescentes

Principais formas de representação de dados estatísticos:

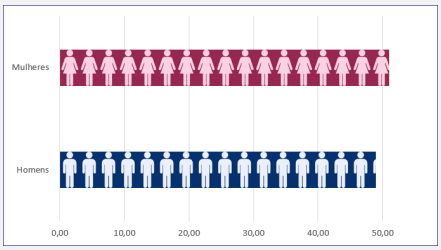
Gráfico	Definição
	Os gráficos em linha normalmente são usados para representar a variação dos valores de uma variável ao longo do tempo. Esse tipo de gráfico permite-nos comparar duas variáveis: uma é traçada no eixo x (horizontal) e a outra no eixo y (vertical).
	Os gráficos em barra normalmente são usados para representar distribuições de dados categóricos ou qualitativos. Uma série estatística é representada por um conjunto de retângulos dispostos horizontalmente, cada um indicando uma categoria particular.
	Os gráficos em coluna também são usados para distribuições de dados categóricos ou qualitativos. A diferença básica é que, agora, uma série estatística é representada por um conjunto de retângulos dispostos verticalmente, cada um indicando uma categoria particular.
	O gráfico em setores é usado para representar a frequência relativa (porcentagem) de uma variável categórica, sendo formado por um círculo dividido em setores circulares, cada um representando uma categoria.



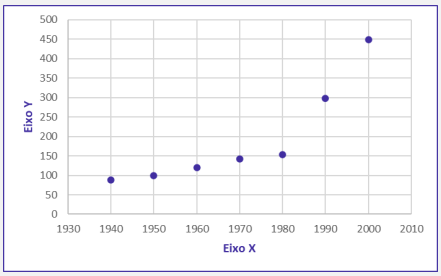
O gráfico polar consiste em uma sequência de eixos igualmente espaçados (ângulos iguais), cada um representando uma das variáveis. Uma linha é desenhada ligando os valores de cada eixo.



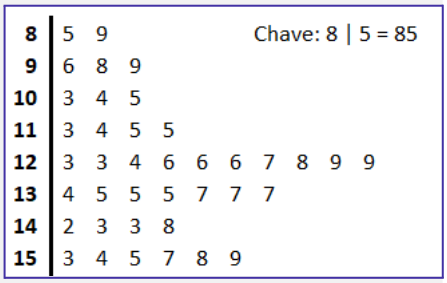
O cartograma é empregado com a finalidade de apresentar dados estatísticos diretamente relacionados com áreas geográficas.



O pictograma substitui valores por ícones, tornando os dados mais atraentes e facilitando o entendimento acerca de um determinado fenômeno.



O gráfico de dispersão é uma representação de pares ordenados em um plano cartesiano, composto por um eixo vertical (ordenada) e um eixo horizontal (abcissa). Os dados são representados como uma coleção de pontos.



		Chave: 8 | 5 = 85
8	5 9	
9	6 8 9	
10	3 4 5	
11	3 4 5 5	
12	3 3 4 6 6 6 7 8 9 9	
13	4 5 5 5 7 7 7	
14	2 3 3 8	
15	3 4 5 7 8 9	

O diagrama de ramos e folhas fornece uma maneira rápida de representar graficamente a distribuição dos dados. Nele, cada número é separado em duas partes. Em geral, de um lado ficam as unidades do número e do outro lado fica o restante desse número.

Análise Exploratória de Dados

Conceitos Básicos

A análise exploratória de dados (AED) é uma etapa que ocorre após a coleta de dados e deve ser realizada antes de qualquer tentativa modelagem estatística. Esta etapa é fundamental, pois permite ao pesquisador familiarizar-se com os dados, organizá-los e sintetizá-los de forma a obter as informações necessárias do conjunto de dados para responder as questões que estão sendo estudadas.

O objetivo da AED é usar medidas estatísticas e visualizações para entender melhor os dados e encontrar pistas sobre tendências, sobre a qualidade dos dados e formular suposições e hipóteses. O foco não é em criar visualizações sofisticadas ou mesmo esteticamente agradáveis, mas em tentar responder a perguntas relacionadas aos dados em análise.

A AED foi proposta pelo estatístico norte-americano John Tukey, como forma de incentivar outros estatísticos a explorarem dados e formularem hipóteses, levando a novas coletas de dados e experimentos. A finalidade é examinar os dados antes da aplicação de qualquer técnica estatística.

A AED faz uso de técnicas gráficas e quantitativas, visando maximizar a obtenção de informações ocultas na estrutura dos dados, identificar tendências, detectar comportamentos anômalos, testar a validade das hipóteses assumidas. A AED também é responsável por sintetizar os dados por meio das chamadas medidas estatísticas (medidas-resumo), que podem ser classificadas em quatro grupos:

- Medidas de posição, entre elas as medidas de tendência central e as separatrizes;
- Medidas de dispersão como a variância e o desvio padrão;
- Medidas de assimetria e
- Medidas de achatamento ou de curtose.

Portanto, a análise de dados consiste em métodos e técnicas estatísticas que permitem ao investigador reforçar, confirmar ou não ideias acerca de um determinado fenômeno.

Tipos de dados

Antes de começarmos a aprender sobre a exploração de dados, vamos primeiro aprender os diferentes tipos de dados ou níveis de medição. Os dados podem aparecer em vários formatos, mas podem ser classificados em dois grupos principais:

- dados estruturados: são dados em um formato padronizado, com estrutura bem definida e que obedecem a um modelo de dados e seguem uma ordem persistente e de fácil

acesso por humanos e programas. Esse tipo de dado geralmente é armazenado em um banco de dados;

- dados não estruturados: são dados que não estão organizados e não podem ser armazenados de maneira lógica. Os dados não estruturados não se encaixam em nenhuma estrutura de banco de dados. Exemplos de dados não estruturados são fotos, imagens, áudio, textos, etc.

Os dois tipos comuns de dados estruturados com os quais geralmente lidamos são dados categóricos ou dados numéricos, os quais veremos a seguir.

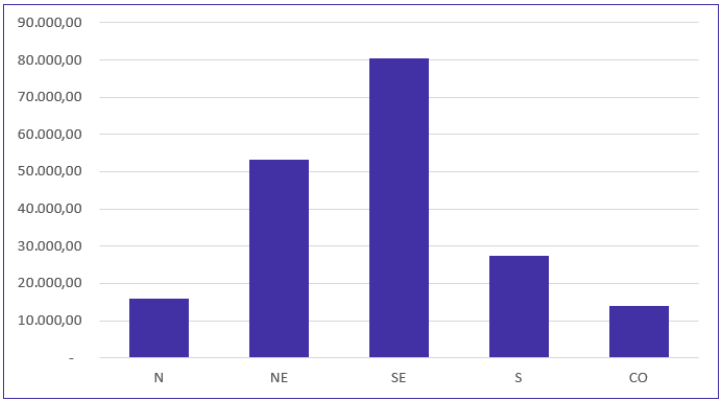
Dados Categóricos

As variáveis categóricas são as características que podem ser definidas por meio de qualidades (atributos) do indivíduo pesquisado. Elas são classificadas em nominais, quando as possíveis categorias não podem ser ordenadas; ou ordinais, quando as possíveis categorias podem ser ordenadas de alguma forma.

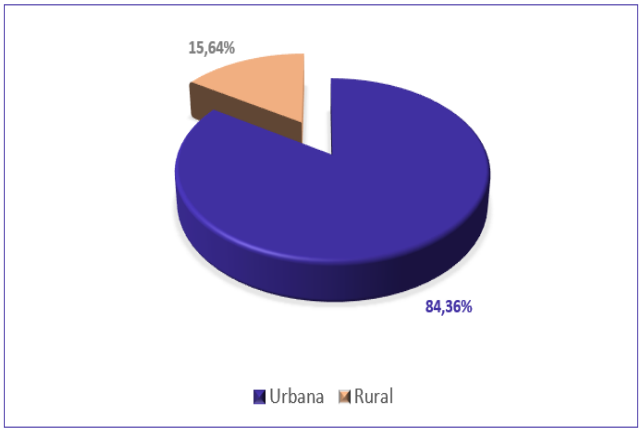
Um método comum para análise exploratória de dados categóricos consiste na construção de uma tabela de frequências contendo o número de ocorrências e a frequência relativa dos dados de cada categoria. Um exemplo de tabela de frequências é mostrado a seguir:

Escolaridade	Frequência Absoluta	Frequência Relativa
Ensino Fundamental	15	30%
Ensino Médio	20	40%
Ensino Superior	15	30%
Total	**50**	**100%**

Os gráficos em colunas podem ser utilizados para representar dados categóricos. Nessa disposição gráfica, uma série estatística é representada por um conjunto de retângulos dispostos verticalmente, cada um indicando uma categoria particular, todos com a mesma largura e alturas proporcionais aos respectivos dados.



O gráfico em setores (também conhecido como gráfico de pizza) também pode ser usado para representar a frequência relativa (porcentagem) de um dado categórico. Ele é formado por um círculo dividido em setores circulares, cada um representando uma categoria, cujos ângulos centrais são proporcionais às frequências relativas da categoria.



Dados Numéricos

As variáveis numéricas são características que podem ser descritas em termos de quantidades, obtidas por meio de contagem ou mensuração. Elas são classificadas em discretas, quando os possíveis valores formam um conjunto finito ou enumerável de números e, geralmente, resultam de um processo de contagem; ou contínuas, quando os possíveis valores formam um intervalo de números reais e, normalmente, resultam de um processo de mensuração.

Categorização (numérico para categórico)

A categorização é o processo de transformação de variáveis numéricas em categóricas. Por exemplo, a idade pode ser categorizada em 0-12 (criança), 13-19 (adolescente), 20-65 (adulto),

65+ (idoso). Ela pode ser usada como um filtro para reduzir o ruído ou a não linearidade dos dados. Esse processo também permite que os pesquisadores avaliem rapidamente valores discrepantes, valores inválidos ou ausentes para valores numéricos.

Codificação (categórico para numérico)

A **codificação** é a transformação de variáveis categóricas em variáveis numéricas (ou binárias). Um exemplo básico de codificação é a avaliação de um atendimento em péssimo (1), ruim (2), regular (3), bom (4) e ótimo (5). A codificação binária é um caso especial de codificação em que o valor é definido como 0 ou 1 para indicar ausência ou presença de uma categoria.

Análise Exploratória de Dados Univariados

Dados univariados são dados relativos a **uma única variável**. Por exemplo, se registrarmos o valor de mercado estimado de cada imóvel em uma determinada região e não registrarmos mais nada, nosso conjunto de dados será univariado.

Posto assim, é muito difícil que os conjuntos de dados do mundo real sejam univariados. Em qualquer conjunto de dados, por menor que seja, quase sempre teremos valores referentes a mais de uma variável. No entanto, a compreensão de dados univariados tem um papel fundamental na análise de dados.

Em primeiro lugar, mesmo quando temos dados relativos a dezenas ou centenas de variáveis, muitas vezes precisamos entender as variáveis individuais de maneira isolada. Em outras palavras, primeiro exploramos nossas muitas variáveis individualmente antes de examinarmos as possíveis relações entre elas.

A análise de dados univariados é uma parte importante da análise exploratória. Muitas vezes, estamos interessados em explorar, seja por meio visualizações ou usando medidas estatísticas, como uma variável é distribuída quando outras variáveis são mantidas constantes. Por exemplo, podemos estar interessados em entender como os preços dos imóveis variam em função da idade do imóvel, do número de quartos, da área construída, da vizinhança, etc.

Sumarização dos Dados por meio de Medidas Estatísticas

Podemos descrever qualquer distribuição univariada em termos de três características principais: **posição**, **dispersão** e **forma**.

As **medidas de posição** são estatísticas que caracterizam o comportamento dos elementos de uma série de dados, orientando quanto à posição da distribuição em relação ao eixo horizontal do gráfico da curva de frequência. Em outras palavras, podemos dizer que as medidas de posição indicam a tendência de concentração dos elementos de uma série, apontando o valor que melhor representa o conjunto de dados.

A **dispersão ou variabilidade** de uma distribuição nos diz o quão dispersa ou espalhada está a distribuição. Ele nos diz quanta variação existe nos valores da distribuição ou quão distantes estão os valores um do outro em média.

Por último, a **forma** de uma distribuição é qualquer coisa que não seja descrita nem pela localização nem pela dispersão. Duas das características de forma mais importantes são a assimetria e a curtose. A assimetria mede o grau de afastamento de uma distribuição em relação ao eixo de simetria. A curtose mede o grau de achatamento de uma distribuição em relação a uma distribuição padrão.

Medidas de Posição

As medidas de posição podem ser divididas em: **medidas de tendência central** e **separatrizes**. As medidas de tendência central representam o ponto central ou o valor típico de um conjunto de dados, indicando onde está localizada a maioria dos valores de uma distribuição. As medidas separatrizes: dividem (ou separam) uma série em duas ou mais partes, cada uma contendo a mesma quantidade de elementos.

Média Aritmética

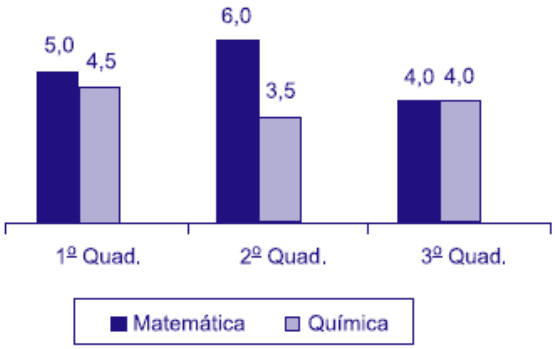
A média aritmética simples está muito presente em nosso cotidiano, seja no consumo médio de combustível, na temperatura média ou na renda per capita. Essa medida é definida como o quociente entre a soma de todos os elementos e o número deles:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Reparem que no numerador somamos todos os elementos, ao passo que no denominador temos a quantidade de elementos somados (n).

HORA DE PRATICAR!

(VUNESP/FITO/2020) O gráfico apresenta as notas de um aluno, nas disciplinas de matemática e química, nos três quadrimestres de 2019.



A média das notas de matemática desse aluno corresponde, da média das notas de química, a

a) 120%

b) 125%

c) 130%

d) 135%

e) 140%

Comentários:

A média aritmética é definida pelo quociente entre a soma dos valores de um determinado conjunto e a quantidade de valores nele existentes. Pelos valores dados no enunciado, a média das notas de matemática é:

$$\bar{x}_{mat} = \frac{5 + 6 + 4}{3}$$

$$\bar{x}_{mat} = \frac{15}{3}$$

$$\bar{x}_{mat} = 5$$

Já a média das notas de química é:

$$\bar{x}_{quím} = \frac{4{,}5 + 3{,}5 + 4}{3}$$

$$\bar{x}_{quím} = \frac{12}{3}$$

$$\bar{x}_{quím} = 4$$

Com isso, em termos percentuais, a média das notas de matemática desse aluno corresponde, da média das notas de química, a:

$$\frac{5}{4} = 1{,}25 = 125\%$$

Gabarito: B.

Mediana

A **mediana** é, simultaneamente, uma **medida de posição, de tendência central e separatriz**. Ela caracteriza **a posição central de uma série de valores**. Além disso, ela também separa uma série de valores em duas partes de tamanhos iguais, cada uma contendo o mesmo número de elementos.

Para determinação da mediana, precisamos ordenar o conjunto de dados. Feito isso, **a mediana é o elemento que ocupa a posição central de uma série de observações ordenada segundo suas grandezas**.

É importante notarmos que, quando uma série possui um número ímpar de elementos, a mediana sempre coincide com o elemento central do conjunto de dados. Contudo, **se porventura a série tivesse um número par de elementos, por convenção, a mediana seria a média aritmética dos dois termos centrais**.

Note que, quando o número é ímpar, o termo central sempre ocupa a posição $\frac{n+1}{2}$. Por outro lado, quando o número de termos é par, existem dois termos centrais, sendo que o primeiro ocupa a posição $\frac{n}{2}$; e o segundo ocupa a posição imediatamente seguinte, ou seja, $\frac{n}{2} + 1$.

A mediana depende da apenas posição e não dos valores dos elementos de uma série ordenada. Essa é uma das principais diferenças entre a média e a mediana, pois a primeira é muito impactada pela presença de valores extremos enquanto a última não. Por isso, a mediana é considerada uma **medida robusta**!!

(CESPE/FUB/2018) A tabela seguinte mostra as quantidades de livros de uma biblioteca que foram emprestados em cada um dos seis primeiros meses de 2017.

	Mês					
	1	2	3	4	5	6
Quantidade	50	150	250	250	300	200

A partir dessa tabela, julgue o próximo item.

A mediana dos números correspondentes às quantidades de livros emprestados no primeiro semestre de 2017 é igual a 200.

Comentários:

A mediana é o termo central de uma amostra ou população. Se temos 6 meses, então a mediana poderá ser encontrada pela média dos termos que ocupam as posições 3 e 4, pois, nesse caso, não há apenas um termo central. Organizando os dados da tabela em ordem crescente (isto é, em rol crescente), temos:

$$50 \quad 150 \quad \underbrace{200 \quad 250}_{termos\ centrais} \quad 250 \quad 300$$

Encontrando a média dos termos nas posições 3 e 4:

$$M_d = \frac{200 + 250}{2} = 225$$

Gabarito: Errado.

Moda

A moda é uma medida de posição e de tendência central que descreve o valor mais frequente de um conjunto de observações, ou seja, o valor de maior ocorrência dentre os valores observados. Um conjunto de dados pode ser unimodal, bimodal, trimodal ou plurimodal, de acordo com o número de modas que apresenta. A ausência de uma moda caracteriza o conjunto como amodal.

Em geral, a moda é utilizada em distribuições nas quais o valor mais frequente é o mais importante da distribuição. A moda também é útil para a determinação da medida de posição de variáveis qualitativas nominais, ou seja, variáveis não numéricas que não podem ser ordenadas.

(CESPE/IPHAN/2018) Define-se estatística descritiva como a etapa inicial da análise utilizada para descrever e resumir dados. Em relação às medidas descritivas, julgue o item a seguir.

A moda é o valor que apresenta a maior frequência da variável entre os valores observados.

Comentários:

A moda pode ser definida como o valor (ou os valores) que mais se repete(m) em uma amostra ou conjunto. Ou seja, que aparece(m) com maior frequência. Uma amostra pode apresentar mais de uma moda, sendo classificada como plurimodal; ou apenas uma moda, recebendo a denominação de unimodal; ou ainda amodal, quando todos os valores das variáveis em estudo apresentarem uma mesma frequência.

Gabarito: Certo.

Medidas de Dispersão

As medidas de dispersão (ou variabilidade) são métricas que mostram a variação dos dados de um conjunto, indicando o grau de homogeneidade ou heterogeneidade existente entre os valores que compõem o conjunto.

Desvio Padrão e Variância

A **variância populacional** é simbolizada pela letra grega σ (sigma), sendo **calculada usando todos os elementos da população**, pela seguinte fórmula:

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

em que: x_i é o valor de ordem i assumido pela variável; μ é a média populacional de x; σ^2 é a variância populacional; e n é o número de dados da população.

A variância amostral é simbolizada pela letra s, sendo calculada a partir de uma amostra da população, pela seguinte fórmula:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

em que: x_i é o valor de ordem i assumido pela variável; $\bar{x}$ é a média amostral de x; s^2 é a variância amostral; e n é o número de dados da amostra.

Normalmente, uma população possui uma grande quantidade de elementos, o que inviabiliza a realização de um estudo aprofundado de suas medidas, chamadas de parâmetros populacionais. Por isso, recorremos ao estudo de amostras representativas dessa população, buscando obter indícios do valor correto do parâmetro populacional desconhecido. Esse valor amostral é denominado de estimador do parâmetro populacional.

Reparem que, quando a variância representa uma descrição da amostra e não da população, caso mais frequente em estatística, o denominador das expressões deve ser $n-1$, em vez de n. Isso ocorre porque a utilização do divisor $(n-1)$ resulta em uma melhor estimativa do parâmetro populacional.

O desvio padrão, por sua vez, é simplesmente a raiz quadrada da variância. Portanto, possui as mesmas unidades dos dados originais, ajudando a torná-los mais interpretáveis. O desvio padrão da amostra geralmente é representado pelo símbolo s.

(VUNESP/TJ-SP/2015) Dados os valores de uma variável: 5, 10, 15, 20, 25, as variâncias amostral e populacional são, respectivamente,

a) 14,7 e 15.

b) 125 e 250.

c) 62,5 e 50.

d) 29,4 e 30,8.

e) 83,3 e 85.

Comentários:

Vamos começar calculando a média:

$$\frac{5+10+15+20+25}{5}=15$$

Agora, vamos encontrar os desvios em relação à média:

$$d_1 = 5-15=-10$$
$$d_2 = 10-15=-5$$
$$d_3 = 15-15=0$$
$$d_4 = 20-15=5$$
$$d_5 = 25-15=10$$

Para calcular a variância (populacional ou amostral), precisamos calcular a soma dos quadrados dos desvios, isto é:

$$\sum d_i^2 = (-10)^2+(-5)^2+0^2+5^2+10^2$$
$$\sum d_i^2 = 250$$

Nesse momento, dividiremos esse valor por n para encontrar a variância populacional e por $n-1$ para encontrar a variância amostral:

$$s^2 = \frac{\sum d_i^2}{n-1}=\frac{250}{5-1}=\frac{250}{4}=62{,}5 \text{ (variância amostral)}$$

$$\sigma^2 = \frac{\sum d_i^2}{n}=\frac{250}{5}=50 \text{ (variância populacional)}$$

Gabarito: C.

Amplitude Interquartílica

A **amplitude interquartílica** (ou distância interquartílica, ou intervalo interquartílico) é o resultado da subtração entre o terceiro quartil e o primeiro quartil:

$$\boldsymbol{DIQ = Q_3 - Q_1}$$

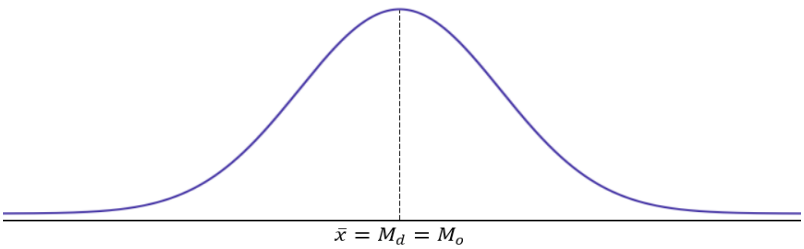
Os quartis são os valores que dividem uma série em quatro partes iguais, isto é, quatro partes contendo o mesmo número de elementos (25%). Temos, então, 3 quartis (Q_1 , Q_2 e Q_3) para dividir uma série em quatro partes iguais:

- Q_1: o primeiro quartil corresponde à separação dos primeiros 25% de elementos da série;
- Q_2: o segundo quartil corresponde à separação de metade dos elementos da série, coincidindo com a mediana ($Q_2 = M_d$);
- Q_3: o terceiro quartil corresponde à separação dos primeiros 75% de elementos da série, ou dos últimos 25% de elementos da série.

Medidas de Assimetria e Curtose

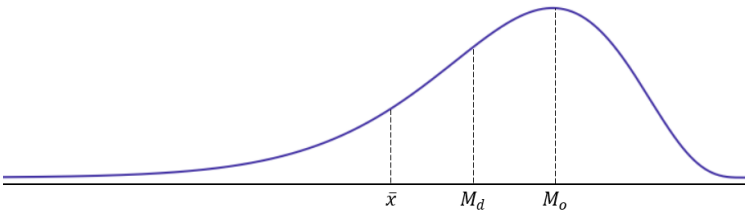
Assimetria

A assimetria mede o grau de afastamento de uma distribuição em relação ao eixo de simetria. Uma distribuição é simétrica quando possui um único valor para a moda, a média e a mediana. Ela tem associada a si uma curva de frequências *unimodal* que, em relação à linha vertical que passa pelo seu ponto mais alto (eixo de simetria), apresenta duas "caudas" simétricas. Nesse caso, as medidas estão localizadas no ponto central da distribuição.



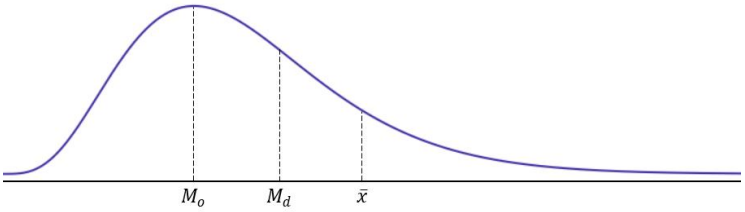
Uma distribuição é assimétrica quando não possui um único valor para a moda, a média e a mediana. Ela tem associada a si uma curva de frequências *unimodal* que, em relação à linha vertical que passa pelo seu ponto mais alto, apresenta ou uma "cauda" mais longa para a esquerda (assimetria negativa), ou uma "cauda" mais longa para a direita (assimetria positiva). Nesse caso, a mediana sempre está localizada entre a moda, que corresponde ao ponto mais alto da curva; e a média, situada perto "cauda".

A assimetria é negativa quando os valores mais baixos das observações são predominantes, isto é, a curva de frequência tem uma "cauda" mais longa à esquerda do ponto que corresponde à frequência máxima. Nesse caso, o valor da média será menor que o da mediana que, por sua vez, será menor que o da moda:



$$\bar{x} < M_d < M_o$$

Dizemos que **a assimetria é positiva quando os valores mais altos das observações são predominantes**, isto é, a curva de frequência tem uma "cauda" mais longa à direita do ponto que corresponde à frequência máxima. Nesse caso, o valor da média será maior que o da mediana que, por sua vez, será maior que o da moda:

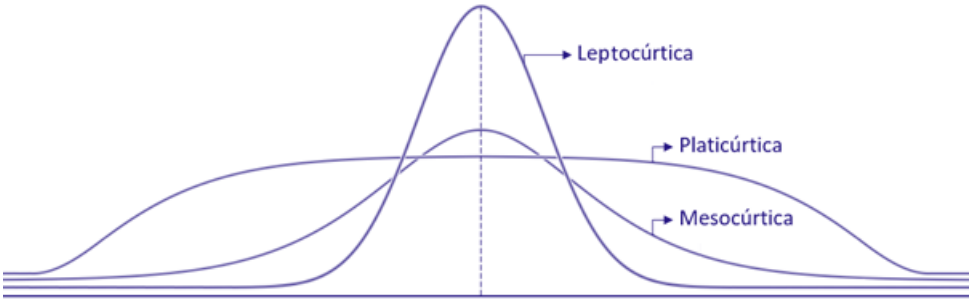


$$\bar{x} > M_d > M_o$$

Curtose

A curtose **mede o grau de achatamento (ou afilamento) de uma distribuição em relação a uma distribuição padrão (chamada curva normal padrão)**. De acordo com o grau da curtose, podemos classificar as curvas de frequência em três tipos:

- **mesocúrtica**: é a curva chamada de curva normal padrão;
- **leptocúrtica**: a medida de curtose é maior do que a da distribuição normal. A curva é mais alta e mais fechada (ou mais afilada) que a curva da distribuição normal;
- **platicúrtica**: a medida de curtose é menor do que a da distribuição normal. A curva é mais aberta (ou mais achatada) que a curva da distribuição normal.



(FCC/TRT 14ª Região/2018) Analisando uma curva de frequência de uma distribuição estatística, observa-se que ela:

I) Unimodal.

II) Apresenta a moda menor que a mediana e a mediana menor que a média.

III) Possui os dados da distribuição fortemente concentrados em torno da moda.

Então, essa distribuição

a) É assimétrica à esquerda e caracteriza-se como platicúrtica.

b) É assimétrica à direita e caracteriza-se como leptocúrtica.

c) Apresenta uma assimetria negativa e caracteriza-se como platicúrtica.

d) É assimétrica à esquerda e caracteriza-se como leptocúrtica.

e) É assimétrica à direita e caracteriza-se como platicúrtica.

Comentários:

Podemos classificar uma distribuição quanto à assimetria com base em três critérios:

- se $\bar{x} = M_d = M_o$, a curva da distribuição é simétrica;

- se $\bar{x} > M_d > M_o$, a curva da distribuição tem assimetria positiva ou à direita;

- se $\bar{x} < M_d < M_o$, a curva da distribuição tem assimetria negativa ou à esquerda.

Com relação à curtose, a distribuição pode ser classificada em:

- mesocúrtica – quando apresenta uma medida de curtose igual à da distribuição normal;

- platicúrtica – quando apresenta uma medida de curtose menor que a da distribuição normal. Tem aparência mais achata;

- leptocúrtica - quando apresenta uma medida de curtose maior que a da distribuição normal. Tem aparência mais afilada.

Analisando enunciado, concluímos que a distribuição é assimétrica positiva ou à direita, caracterizada como leptocúrtica.

Gabarito: B.

Visualização de Dados Univariados

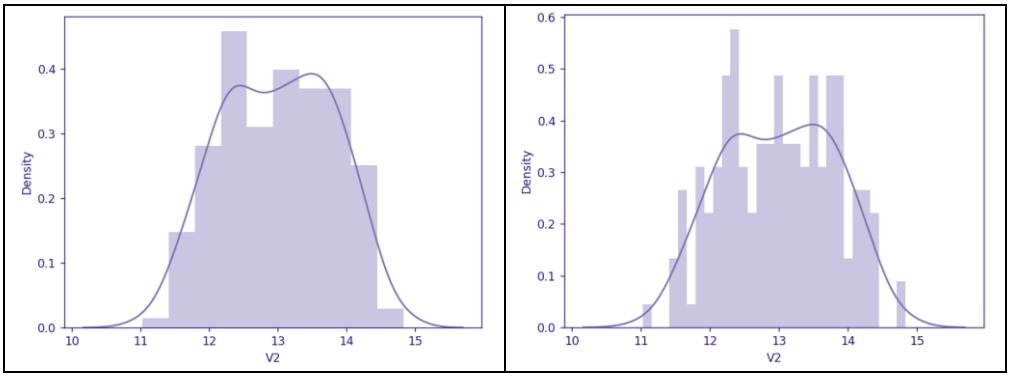
O principal objetivo dos gráficos estatísticos é proporcionar uma visualização mais rápida dos dados estatísticos ou do fenômeno sob investigação. A seguir vamos ver as principais formas de representação de dados univariados.

Histograma

O **histograma** é um gráfico destinado a representar **dados agrupados em classes**, sendo composto por um conjunto de **retângulos contíguos (justapostos)** cujas bases estão situadas sobre o eixo horizontal (eixo x), de forma que os seus pontos médios devem coincidir com os pontos médios dos intervalos de classe e seus limites devem coincidir com os limites da classe.

A quantidade de retângulos em um histograma é equivalente ao número de intervalos de classe. A largura de cada retângulo deve ser igual à amplitude do intervalo de classe, enquanto a altura precisa ser proporcional à frequência do intervalo de classe. Além disso, **a área do histograma é proporcional ao somatório das frequências**.

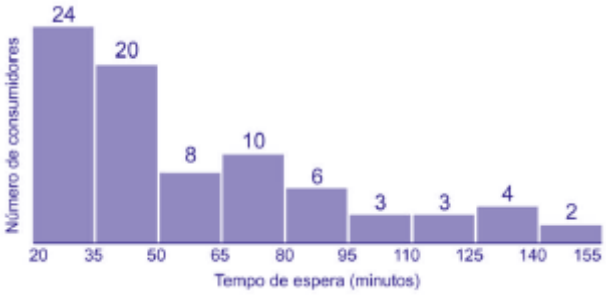
Cada retângulo representa a frequência absoluta (número de ocorrências) ou relativa (número de ocorrências dividido pelo total) de casos para um intervalo de valores. **Os histogramas fornecem uma impressão imediata da forma da distribuição (simétrica, uni/plurimodal, assimétrica, presença de valores discrepantes)**. O número de colunas influencia fortemente o aspecto final do histograma, como vemos a seguir.



Por fim, **o histograma pode ocasionar um certo nível de perda de informações, pois os elementos da distribuição de frequência não são representados de forma individualizada, mas sim por meio de suas classes**.

HORA DE PRATICAR!

(VUNESP/CMSJC/2022) Os tempos de espera, em minutos, para o atendimento de 80 consumidores em um centro de atendimento ao consumidor estão registrados no gráfico a seguir.



De acordo com o gráfico, é correto afirmar que o tempo de espera de

a) mais da metade dos consumidores foi superior a 1 hora.

b) 12,5% dos consumidores foi entre 1 h e 35 min e 2 h e 20 min.

c) 65% dos consumidores foi inferior a 1 hora.

d) 24 consumidores foi entre 50 min e 80 min.

e) no mínimo 2 pessoas, foi superior a 2 h e 30 min.

Comentários:

Primeiro, podemos montar a tabela de frequências acumuladas para simplificar a análise das alternativas:

Classes	Frequência Absoluta	Frequência Absoluta Acumulada	Frequência Relativa	Frequência Relativa Acumulada
20 ⊢ 35	24	24	30,00%	30,00%
35 ⊢ 50	20	44	25,00%	55,00%
50 ⊢ 65	8	52	10,00%	65,00%
65 ⊢ 80	10	62	12,50%	77,50%
80 ⊢ 95	6	68	7,50%	85,00%
95 ⊢ 110	3	71	3,75%	88,75%
110 ⊢ 125	3	74	3,75%	92,50%
125 ⊢ 140	4	78	5,00%	97,50%
140 ⊢ 155	2	80	2,50%	100,00%
Total	**80**	**80**	**100%**	**100%**

Agora, analisando as alternativas, podemos afirmar que:

Alternativa A: Incorreta. O tempo de espera foi inferior a 50 minutos para mais da metade (55%) das pessoas.

Alternativa B: Correta. Para 12,5% dos consumidores, o tempo de espera foi entre 1h e 35 min e 2h e 20min. Somando as frequências relativas referentes às classes 95 ⊢ 110, 110 ⊢ 125 e 125 ⊢ 140, temos:

$$3{,}75\% + 3{,}75\% + 5{,}00\% = 12{,}5\%$$

Alternativa C: Incorreta. Não podemos afirmar isso, pois a terceira classe termina em 65 min. O que podemos afirmar é que para 65% das pessoas o tempo de espera foi inferior a 1h e 5 minutos.

Alternativa D: Incorreta. O tempo de espera foi entre 50 min e 80 min para 18 pessoas.

Alternativa E: Incorreta. O que se pode afirmar é que, para no mínimo duas pessoas, o tempo de espera foi superior a 2h e 20 min.

Gabarito: B.

Diagrama de Ramos e Folhas

O **diagrama de ramos e folhas** fornece uma maneira rápida de representar graficamente a distribuição dos dados. Nesse diagrama, cada número é separado em duas partes. Em geral, de um lado ficam as unidades do número e do outro lado fica o restante desse número. Consideremos o seguinte conjunto de dados:

85, 89, 96, 98, 99, 103, 104, 105, 113, 114, 115, 115, 123, 123, 124, 126, 126, 126, 127, 128, 129, 129, 134, 135, 135, 135, 137, 137, 137, 142, 143, 143, 148, 153, 154, 155, 157, 158, 159, 161, 161, 165, 168, 170, 171, 171, 171, 173, 175, 175

A representação utilizando um diagrama de ramos e folhas ficaria assim:

8	5 9	Chave: 8 | 5 = 85
9	6 8 9	
10	3 4 5	
11	3 4 5 5	
12	3 3 4 6 6 6 7 8 9 9	
13	4 5 5 5 7 7 7	
14	2 3 3 8	
15	3 4 5 7 8 9	
16	1 1 5 8	
17	0 1 1 1 3 5 5	

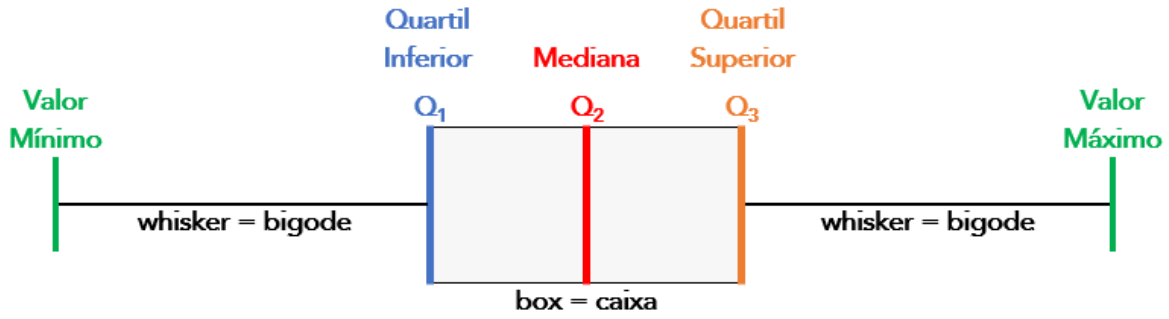
No lado esquerdo, temos as centenas e as dezenas representando os ramos. No lado direito, temos as unidades representando as folhas. As folhas, portanto, estão vinculadas aos ramos. Dessa maneira, a chave "**9** | 6 8 9" significa que, no rol original, estão presentes os números **9**6, **9**8 e **9**9.

Diagrama de Boxplot

Um boxplot é uma ferramenta gráfica frequentemente utilizada na análise exploratória de dados que permite visualizar a distribuição dos dados e os valores discrepantes (outliers), assim como a distância dos valores extremos em relação à maioria dos dados. Essa ferramenta resume cinco medidas descritivas de um conjunto de dados, incluindo: o valor mínimo, o primeiro quartil, a mediana, o terceiro quartil e o valor máximo.

Para construir um gráfico de *boxplot*, usamos uma haste horizontal ou vertical e uma caixa retangular (*box*). O local em que a haste começa (da esquerda para a direita) indica o valor mínimo e o ponto em que a haste termina indica o valor máximo.

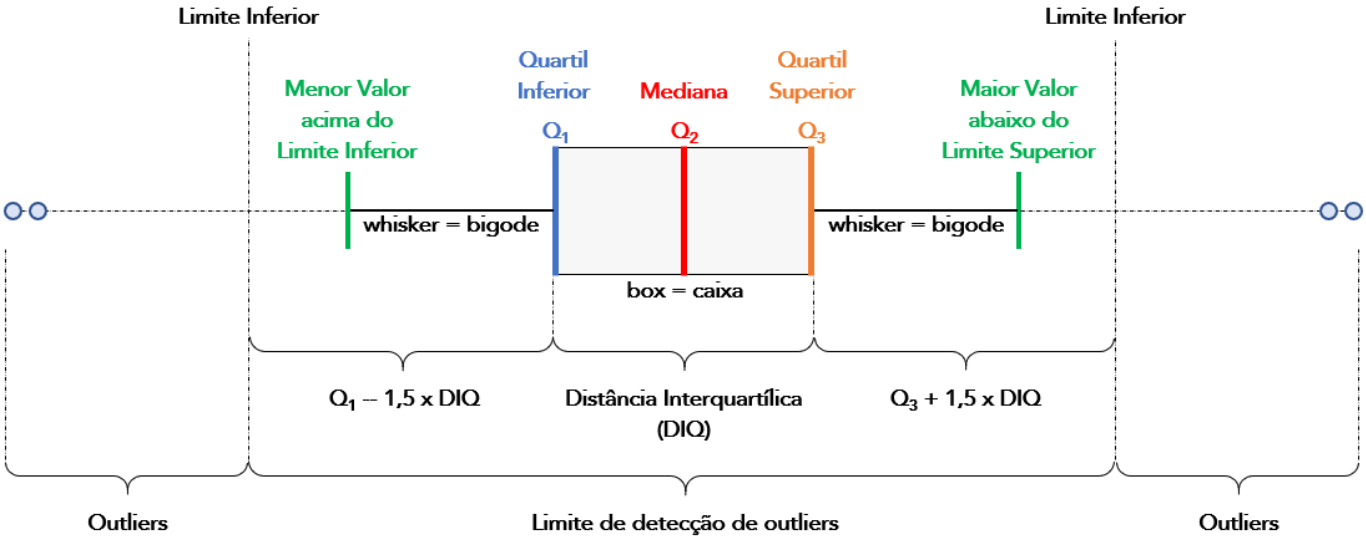
A caixa retangular, localizada no meio da haste, em geral, possui três linhas. A primeira linha, na extremidade esquerda da caixa, indica o primeiro quartil. A terceira linha, na extremidade direita, indica o terceiro quartil. A linha do meio, no interior da caixa, indica o segundo quartil ou a mediana. O segundo quartil pode estar entre o primeiro e o terceiro quartis, ou pode coincidir com um, ou outro, ou ambos.



Além disso, há dois traços, chamados de *whiskers* (ou bigodes), ligando o valor mínimo à extremidade esquerda da caixa e o valor máximo à extremidade direita da caixa. Cada um desses traços comporta, aproximadamente, 25% dos dados. O restante, cerca de 50%, está distribuído no interior da caixa.

Também podemos encontrar gráficos de *boxplot* com pontos ou asteriscos marcando valores discrepantes (*outliers*). Nesses casos, os *whiskers* não se estendem aos valores mínimo e máximo do conjunto de dados, mas ficam limitados a um comprimento máximo de $1{,}5 \times DIQ$, em que $\boldsymbol{DIQ}$ é a distância interquartílica, calculada pela fórmula:

$$\boldsymbol{DIQ = Q_3 - Q_1}$$



Dessa forma, valores menores que $Q_1 - 1,5 \times DIQ$ ou maiores que $Q_3 + 1,5 \times DIQ$ são considerados valores discrepantes (*outliers*) e representados por pontos ou asteriscos.

HORA DE PRATICAR!

(FGV/MPE-BA/2017) Em uma amostra desconfia-se de que três valores sejam, na verdade, "outliers" e que deveriam ser descartados. Para tal avaliação o estatístico dispõe apenas dos valores dos 1° e 3° quartil da distribuição. Os números são os seguintes:

$Q_1(X) = 21, Q_3(X) = 33, X_1 = 58, X_2 = 2$ e $X_3 = 43$

Onde Q_{is} são os quartis e os X_{is} os valores em análise.

Assim, é correto afirmar que:

a) Todos os valores são "outliers";

b) Os valores X_1 e X_3 são "outliers";

c) Nenhum dos valores é "outliers";

d) Apenas o valor X_2 é "outlier";

e) Os valores X_1 e X_2 são "outliers".

Comentários:

Para resolvermos a questão, precisamos calcular os limites inferior e superior da amostra. Assim:

$$l_{inf} = Q_1 - 1,5 \times (Q_3 - Q_1)$$

$$l_{inf} = 21 - 1,5 \times (33 - 21)$$

$$l_{inf} = 21 - 18 = 3$$

$$l_{sup} = Q_3 + 1{,}5 \times (Q_3 - Q_1)$$
$$l_{sup} = 33 + 1{,}5 \times (33 - 21)$$
$$l_{sup} = 33 + 18 = 51$$

Como $2 < 3$ e $58 > 51$, então podemos afirmar que os valores X_1 e X_2 são *outliers*.

Gabarito: E.

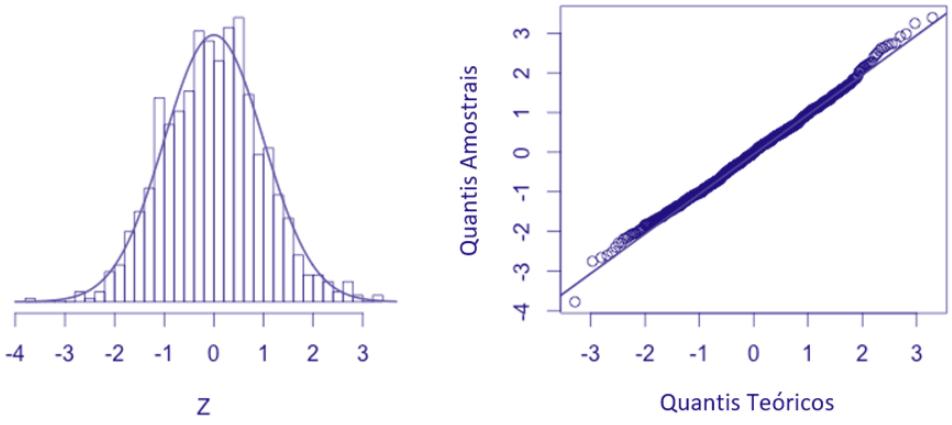
Gráficos de Probabilidade

Os gráficos de probabilidade são usados para avaliar se os dados seguem uma distribuição específica. Eles são usados com mais frequência para testar a normalidade de um conjunto de dados, pois muitos testes estatísticos assumem que as variáveis de interesse são normalmente distribuídas.

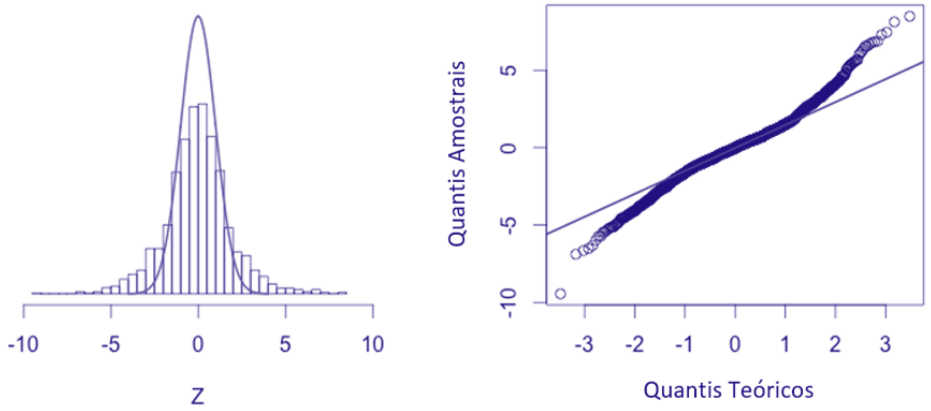
No gráfico de probabilidade normal, os resíduos são comparados com a distribuição normal. A distribuição normal forma uma reta diagonal. O gráfico de probabilidade normal utilizado é o quantil-quantil normal (gráfico Q-Q). Também pode ser usado o gráfico percentil-percentil normal (P-P).

Nesses gráficos, para que uma distribuição de resíduos possa ser classificada como normal, os resíduos observados devem estar ao redor da linha diagonal. A seguir, temos exemplos de gráficos quantil-quantil para vários tipos de distribuições de resíduos:

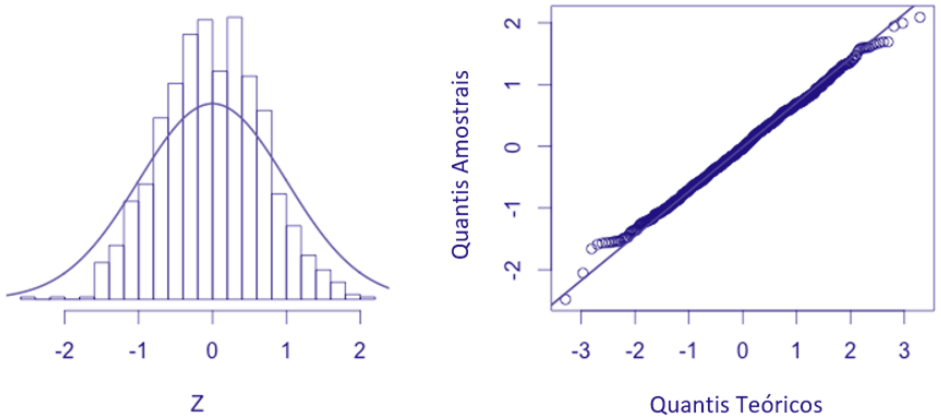
a) distribuição simétrica:



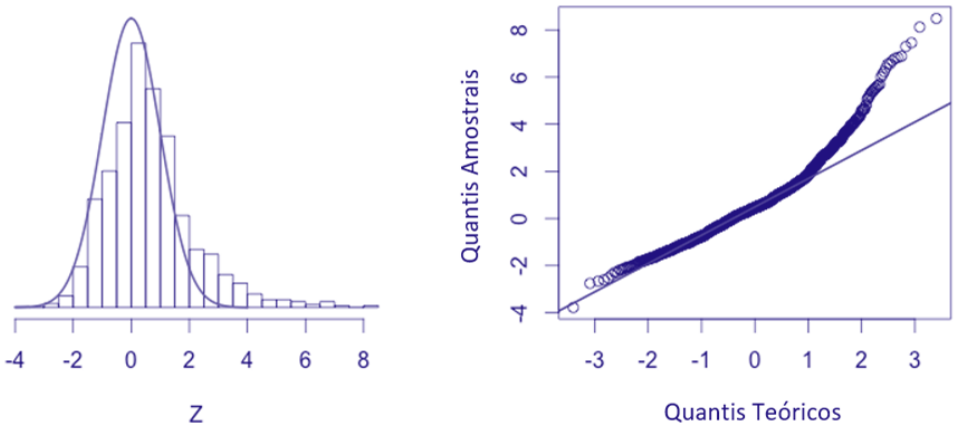
b) distribuição simétrica com caudas pesadas:



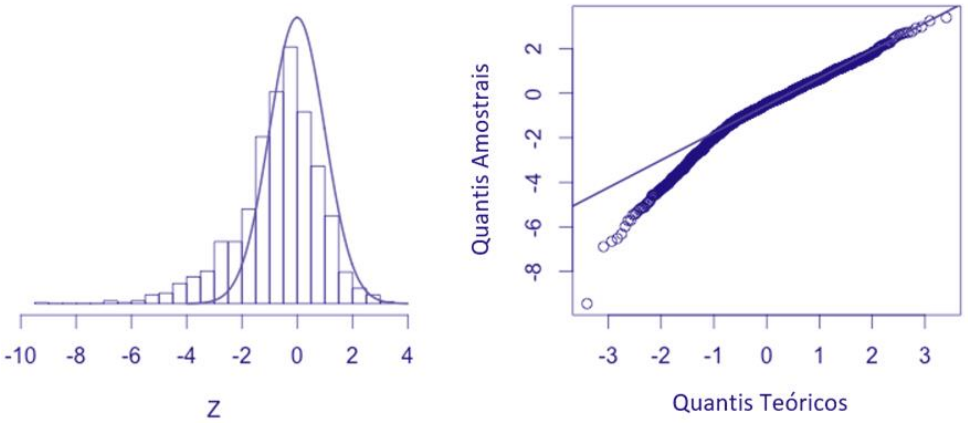
c) distribuição simétrica com caudas leves:



d) distribuição assimétrica à direita:



d) distribuição assimétrica à esquerda:



Além dos gráficos de probabilidade, existem muitos testes estatísticos quantitativos (não gráficos) para testar a normalidade, como o teste qui-quadrado de Pearson, Shapiro-Wilk e Kolmogorov-Smirnov.

Análise Exploratória de Dados Multivariados

Este tipo de análise envolve a investigação de duas ou mais variáveis, buscando identificar possíveis relações. Algumas ferramentas utilizadas são a tabela de contingência, a covariância, a correlação de Pearson, a correlação de Spearman (para dados categóricos) e os gráficos de dispersão.

Tabelas de dupla entrada

As tabelas de dupla entrada (ou tabelas de contingência) exibem o relacionamento entre duas ou mais variáveis categóricas (nominais ou ordinais). O tamanho da tabela é determinado pelo número de valores distintos para cada variável, com cada célula na tabela representando uma combinação exclusiva de valores. Vejamos como seria uma tabela desse tipo:

Área/Sexo	Masculino	Feminino	Total
Jurídica	50	70	120
Controle	60	20	80
Total	110	90	200

Cada entrada dessa tabela representa quantas vezes ocorre cada realização conjunta. O que isso significa? Vejamos a primeira célula da tabela, que tem o valor de 50:

Área/Sexo	Masculino	Feminino	Total
Jurídica	50	70	120
Controle	60	20	80
Total	110	90	200

Essa célula está nos dizendo, basicamente, que 50 homens se interessam pela área jurídica, ou seja, está nos informando a realização simultânea de (sexo = masculino) e (área = jurídica).

Como faríamos para descobrir a quantidade de alunos interessados apenas pela área jurídica, independentemente do sexo? Ora, poderíamos simplesmente somar as observações referentes à área jurídica:

Área/Sexo	Masculino	Feminino	Total
Jurídica	50	70	120
Controle	60	20	80
Total	110	90	200

Nessa amostra, há 120 alunos interessados na área jurídica, sendo 50 homens e 70 mulheres. Esse valor, que nos **informa o total de realizações de uma variável qualitativa** (independentemente das outras variáveis qualitativas), denominamos de **distribuição marginal**.

Em nossa tabela, esses valores são representados pelas células referentes aos totais de homens (110), de mulheres (90), de pessoas interessadas pela área jurídica (120); e de pessoas interessadas pela área de controle (90):

Área/Sexo	Masculino	Feminino	Total
Jurídica	65	55	120
Controle	45	35	80
Total	110	90	200

Muitas vezes, podemos utilizar **frequências relativas** para facilitar a visualização de possíveis interações entre as variáveis, em vez de **frequências absolutas**, como estávamos fazendo. Para calcular as frequências relativas, basta dividirmos as células pelas suas **distribuições marginais**.

Nesse ponto, podemos questionar: devemos utilizar as distribuições marginais das linhas ou das colunas? Isso vai depender do que se quer avaliar. Em nosso caso, fixaremos os totais de cada sexo como 100% e, com base nisso, encontraremos o quanto cada área de interesse representa de cada sexo. Vejamos como ficaria:

Área/Sexo	Masculino	Feminino	Total
Jurídica	59,09%	61,11%	60,00%
Controle	40,91%	38,89%	40,00%
Total	100%	100%	100%

Observem que cada célula foi dividida pelo total da coluna e multiplicada por 100%. Por exemplo, na realização simultânea de (sexo = masculino) e (área = jurídica), fizemos a divisão de 65 por 110, que resulta em, aproximadamente, 59,09%.

Agora, podemos nos perguntar: existe alguma relação entre o sexo e a área de interesse do aluno?

Ao calcular as frequências relativas, estamos encontrando, nas duas primeiras colunas, o percentual de cada sexo que se interessa por uma determinada área, ao passo que, na última coluna, estamos determinando o percentual de pessoas que se interessam por uma determinada área, independentemente do sexo.

Em nosso exemplo, 60% das pessoas da amostra se interessam por pela área jurídica e 40% pela área de controle, independentemente do sexo. Também podemos observar que as proporções do sexo masculino (59,09% e 40,91%) e feminino (61,11% e 38,89%) são muito próximas das marginais (60% e 40%).

Esse resultado indica não haver dependência entre as duas variáveis, pois as frequências relativas de cada categoria não são muito diferentes dos valores marginais. A partir dessas informações, podemos inferir que, muito provavelmente, o sexo de uma pessoa tem pouca influência na escolha entre as áreas jurídica e de controle.

Vejamos outro exemplo:

Área/Sexo	Masculino	Feminino	Total
Policial	100 (66,66%)	20 (40,00%)	120 (60,00%)
Fiscal	50 (33,33%)	30 (60,00%)	80 (40,00%)
Total	150 (100%)	50 (100%)	200 (100%)

Reparem que agora estamos diante de uma situação diferente: as frequências relativas de interesse pelas áreas policial e fiscal, por parte do sexo masculino (66,66% e 33,33%, respectivamente) e feminino (40% e 60%, respectivamente) divergem bastante das proporções marginais (60% e 40%, respectivamente).

Ao dividirmos os indivíduos por sexo, fizemos com que a distribuição de pessoas por áreas de interesse se distanciasse muito em relação valores marginais. Quanto maior é o distanciamento das frequências relativas em relação aos valores marginais, mais forte é a evidência de associação entre as variáveis.

Em essência, o que fizemos foi comparar a distribuição marginal de cada área de interesse com relação às suas respectivas proporções associadas a cada sexo. Assim, caso as variáveis não tivessem nenhuma associação, esperaríamos a seguinte distribuição:

Área/Sexo	Masculino	Feminino	Total
Policial	90 (60,00%)	30 (60,00%)	120 (60,00%)
Fiscal	60 (40,00%)	20 (40,00%)	80 (40,00%)
Total	150 (100%)	50 (100%)	200 (100%)

Ou seja, se as variáveis não são associadas, podemos esperar que 60% dos alunos tenham interesse na área policial e 40% dos alunos tenham interesse na área fiscal, independentemente do sexo. Se isso for verdade, bastaria aplicarmos esses percentuais ao total de cada coluna que encontraríamos os valores esperados de cada célula.

Se compararmos o valor real (valor observado) de cada célula com seu valor esperado, teremos a seguinte distribuição:

Área/Sexo	Masculino	Feminino	Total
Policial	100 - 90 = 10	20 - 30 = -10	0
Fiscal	50 - 60 = -10	30 - 20 = 10	0
Total	0	0	0

Notem que recaímos no mesmo problema de quando estudamos a média, pois a soma dos desvios sempre vai igualar a zero. O desvio, portanto, não é suficiente para caracterizar o grau de associação entre duas variáveis, e precisamos recorrer a outras estratégias.

De modo geral, essa medição é feita pelos chamados coeficientes de associação ou correlação. Essas medidas descrevem, por meio de um único número, a associação (ou dependência) entre

duas variáveis, usualmente variando entre –1 e +1, sendo a proximidade de 0 (zero) um indicativo da inexistência de associação.

Coeficiente de Correlação de Spearman

O coeficiente de correlação de Spearman (ou coeficiente de correlação de postos de Spearman) avalia com que força a relação entre duas variáveis pode ser descrita como uma função monótona, sendo ela linear ou não. A função monótona é aquela que ou apenas cresce ou apenas decresce, preservando a relação de ordem.

Esse coeficiente permite, inclusive, que variáveis de naturezas distintas sejam analisadas, ou seja, uma variável pode ser qualitativa ordinal e outra quantitativa. As variáveis qualitativas ordinais definem características que permitem ordenar um conjunto: péssimo (1), ruim (2), regular (3), bom (4) e ótimo (5).

Para calcular correlação de Spearman é necessário ordenar os elementos de forma crescente e cada elemento ordenado corresponderá a um posto de cada variável. O coeficiente de Spearman, indicado por r_s, é definido como o coeficiente de Pearson (a ser abordado no tópico de análise quantitativa) entre variáveis X, Y classificadas em postos rg_X e rg_Y:

$$r_s = \frac{Cov(rg_X, rg_Y)}{\sigma_{rg_X} \cdot \sigma_{rg_Y}}$$

Em consequência dessa definição, o coeficiente r_s varia entre -1 e 1, assumindo o valor de 0 quando as variáveis forem independentes. Quando todos os postos são distintos, então o coeficiente pode ser calculado a partir da seguinte fórmula:

$$r_s = 1 - \frac{6 . \sum_{i=1}^{n} d_i^2}{n^3 - n}$$

Nessa fórmula, $d_i = rg(X_i) - rg(Y_i)$ é a diferença entre os dois postos de cada observação e n é o número de observações.

EXEMPLIFICANDO

Para exemplificar o coeficiente de Spearman, vamos considerar a seguinte tabela, que relaciona o número de horas diárias de estudo e o desempenho de alguns alunos. A variável desempenho pode ser ordenada em péssimo (1), ruim (2), regular (3), bom (4) e ótimo (5).

Aluno	X: Horas de Estudo	Y: Desempenho
André	1	Bom
Bruno	3	Ruim
Carlos	8	Péssimo
Diego	9	Ótimo
Eduardo	12	Regular

Agora, ordenamos os elementos de forma crescente e atribuímos postos (valores, de 1 a 5 por exemplo) a eles, para, então, calcularmos os valores de $d_i = rg(X_i) - rg(Y_i)$ para cada observação:

Aluno	X	Y	$rg(X_i)$	$rg(Y_i)$	d_i	${d_i}^2$
André	1	Bom	1	4	-3	9
Bruno	3	Ruim	2	2	0	0
Carlos	8	Péssimo	3	1	2	4
Diego	9	Ótimo	4	5	-1	1
Eduardo	12	Regular	5	3	2	4
$\sum_{i=1}^{n} {d_i}^2$						**18**

Em seguida, aplicamos a fórmula para encontrar o valor do coeficiente de Spearman:

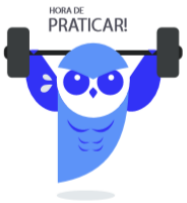
$$r_s = 1 - \frac{6 \times \sum_{i=1}^{n} {d_i}^2}{n^3 - n}$$

$$r_s = 1 - \frac{6 \times 18}{5^3 - 5}$$

$$r_s = 1 - \frac{108}{120}$$

$$r_s = 1 - 0{,}9 = 0{,}1$$

Em nosso exemplo, portanto, constatamos haver pouca relação entre as variáveis X e Y.



(Instituto AOCP/ADAF/2018) O desempenho de estudantes, disposto em ordem alfabética, nas aulas de teoria e prática da disciplina de estatística, foi observado e está na seguinte tabela.

Teoria	8	3	9	2	7	10	4	6	1	5
Prática	9	5	10	1	8	7	3	4	2	6

Qual é o coeficiente de correlação de postos? $\left(r_s = 1 - \frac{6\sum D^2}{n(n^2-1)}\right)$

a) 0,1455

b) 0,9757

c) 0,9819

d) 0,8545

e) 0,0180

Comentários:

Vamos reescrever a tabela calculando as diferenças dos postos:

Posto de x	Posto de y	d_i	d_i^2
8	9	-1	1
3	5	-2	4
9	10	-1	1
2	1	1	1
7	8	-1	1
10	7	3	9
4	3	1	1
6	4	2	4
1	2	-1	1
5	6	-1	1
Soma dos quadrados			**24**

Agora, basta aplicarmos a fórmula dada no enunciado:

$$r_s = 1 - \frac{6 \times \sum d_i^2}{n(n^2-1)}$$

$$r_s = 1 - \frac{6 \times 24}{10 \times (10^2-1)}$$

$$r_s = 1 - \frac{144}{10 \times (100 - 1)}$$

$$r_s = 1 - \frac{144}{10 \times 99}$$

$$r_s = 1 - \frac{144}{990}$$

$$r_s = 0{,}8545$$

Gabarito: D.

Coeficiente de Contingência C

O **coeficiente de contingência C** é uma medida de associação entre dois conjuntos de atributos empregada quando se dispõe apenas de dados apresentados em escala nominal em um ou nos dois atributos. Essa medida possibilita, a partir de observações amostrais, tirar conclusões a respeito das características populacionais, mediante o confronto de uma característica em relação a outra.

Para calcular o coeficiente C, os dados devem ser apresentados em uma **tabela de contingência**, podendo estar dispostos em qualquer ordem e divididos em qualquer número de categorias, isto é, a tabela pode ter l linhas e c colunas. Vejamos a tabela de contingência ilustrada a seguir:

	A_1	A_2	$\cdots$	A_c	Total
B_1	$O_{1,1}$	O_{12}	$\cdots$	O_{1c}	L_1
B_2	$O_{2,1}$	O_{22}	$\cdots$	O_{2c}	L_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
B_l	$O_{l,1}$	O_{l2}	$\cdots$	O_{lc}	L_l
Total	C_1	C_2	$\cdots$	C_c	T

De posse da tabela de contingência, teremos que comparar os respectivos valores **observados** O_{ij}, também chamados de **frequências observadas**, com os valores **esperados** E_{ij}, também chamados de **frequências esperadas**, para cada linha i e coluna j. Nessa etapa, as frequências esperadas são dadas pela expressão:

$$\boldsymbol{E_{ij} = \frac{L_i \times C_j}{T}.}$$

O próximo passo é calcular a estatística qui-quadrado da distribuição conjunta. Para isso, calculamos as diferenças entre as frequências observadas e esperadas, $(O_{ij} - E_{ij})$; elevamos ao quadrado, $(O_{ij} - E_{ij})^2$; dividimos esse valor pelo esperado E_{ij}; e somamos os quocientes para todas as linhas e colunas:

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{l} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

O coeficiente de contingência é, então, obtido da seguinte forma:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

Assim como os demais coeficientes de correlação, o coeficiente de contingência assume o valor **zero quando as variáveis não são relacionadas**. Porém, não assume o valor 1 quando as variáveis são completamente relacionadas, o que podemos considerar uma limitação desse coeficiente.

O valor máximo do coeficiente depende do número l de linhas e c de colunas da tabela de contingência, o que permite **comparações** somente entre tabelas com as mesmas dimensões. Quando uma tabela possui o mesmo número de linhas e colunas, $l = c$, o **valor máximo** do coeficiente C é:

$$C_{máx} = \sqrt{\frac{l-1}{l}}$$

Para evitar esse inconveniente, podemos fazer uso do Coeficiente de Contingência Modificado, C^*, que satisfaz a relação $0 \leq C^* \leq 1$. O coeficiente modificado é definido pela expressão:

$$C^* = \sqrt{\frac{(k \times \chi^2)}{[(k-1) \times (n + \chi^2)]}} = C \times \sqrt{\frac{k}{k-1}}$$

em que k é o menor valor entre o número de linhas e o número de colunas da tabela.

É importante observarmos que o coeficiente de contingência se fundamenta em conceitos distintos dos demais coeficientes de correlação, apresentando, inclusive, valores sempre maiores ou iguais a zero. Isso, por si só, inviabiliza a comparação de seus resultados com os de qualquer medida de correlação.

Também devemos notar que, por depender do cálculo de $\mathcal{X}^2$, o coeficiente apresenta as **mesmas limitações da estatística qui-quadrado**. Dentre elas, vale mencionar que esse cálculo **não deve ser utilizado quando a frequência esperada em uma das categorias é pequena** (valores inferiores a 5 resultam em distorções).

Por fim, a expressão do qui-quadrado considera que a **característica populacional em estudo é representada por uma variável contínua**. Quando a característica populacional é representada por uma **variável discreta**, precisamos corrigir a **falta de continuidade** na fórmula do qui-quadrado.

EXEMPLIFICANDO

Calcular a estatística qui-quadrado e o coeficiente de contingência da tabela de dupla entrada apresentada a seguir:

Curso/Sexo	Masculino	Feminino	Total
Matemática	40	10	**50**
Medicina	45	45	**90**
Direito	25	35	**60**
Total	**110**	**90**	**200**

Primeiro, vamos calcular as frequências esperadas:

$$E_{1,1} = \frac{50 \times 110}{200} = 27{,}5$$

$$E_{1,2} = \frac{50 \times 90}{200} = 22{,}5$$

$$E_{2,1} = \frac{90 \times 110}{200} = 49{,}5$$

$$E_{2,2} = \frac{90 \times 90}{200} = 40{,}5$$

$$E_{3,1} = \frac{60 \times 110}{200} = 33$$

$$E_{3,2} = \frac{60 \times 90}{200} = 27$$

Calculando as diferenças e elevando ao quadrado:

$$(O_{1,1} - E_{1,1}) = 40 - 27{,}5 = 12{,}5 \implies (O_{1,1} - E_{1,1})^2 = (12{,}5)^2 = 156{,}25$$
$$(O_{1,2} - E_{1,2}) = 10 - 22{,}5 = -12{,}5 \implies (O_{1,1} - E_{1,1})^2 = (-12{,}5)^2 = 156{,}25$$
$$(O_{2,1} - E_{2,1}) = 45 - 49{,}5 = -4{,}5 \implies (O_{1,1} - E_{1,1})^2 = (-4{,}5\ )^2 = 20{,}25$$
$$(O_{2,2} - E_{2,2}) = 45 - 40{,}5 = 4{,}5 \implies (O_{1,1} - E_{1,1})^2 = (4{,}5)^2 = 20{,}25$$
$$(O_{3,1} - E_{3,1}) = 25 - 33{,}0 = -8{,}0 \implies (O_{1,1} - E_{1,1})^2 = (-8{,}0)^2 = 64{,}00$$
$$(O_{3,2} - E_{3,2}) = 35 - 27{,}0 = 8{,}0 \implies (O_{1,1} - E_{1,1})^2 = (8{,}0)^2 = 64{,}00$$

A estatística qui-quadrado é dada pela seguinte expressão:

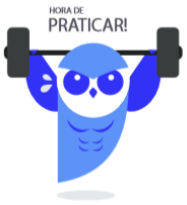
$$\mathcal{X}^2 = \sum_{j=1}^{c}\sum_{i=1}^{l}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{156{,}25}{27{,}5} + \frac{156{,}25}{22{,}5} + \frac{20{,}25}{49{,}5} + \frac{20{,}25}{40{,}5} + \frac{64{,}00}{33{,}0} + \frac{64{,}00}{27{,}0} \cong 17{,}84$$

O coeficiente de contingência é expresso por:

$$C = \sqrt{\frac{\mathcal{X}^2}{n + \mathcal{X}^2}} = \sqrt{\frac{17{,}84}{200 + 17{,}84}} = \sqrt{0{,}082} \cong 0{,}286$$

Por fim, o coeficiente de contingência modificado é definido por:

$$C^* = \sqrt{\frac{k \times \mathcal{X}^2}{[(k-1) \times (n + \mathcal{X}^2)]}} = \sqrt{\frac{2 \times 17{,}84}{[(2-1) \times (200 + 17{,}84)]}} = \sqrt{\frac{35{,}68}{217{,}84}} = \sqrt{0{,}164} \cong 0{,}405$$



(CESPE/ME/2020)

		A			
		-1	0	1	total (%)
B	-1	10	5	5	20
	0	0	60	0	60
	1	10	5	5	20
	total(%)	20	70	10	100

Considerando que a tabela precedente mostra o cruzamento de duas variáveis categorizadas A e B, que foram codificadas em três níveis numéricos de resposta: −1, 0 e 1, julgue o item que se segue.

O coeficiente de contingência é nulo.

Comentários:

O coeficiente de contingência é dado por:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Em que χ^2 representa o qui-quadrado que é dado por:

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{l} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Em que:

$O_{ij} \rightarrow frequências\ observadas$;

$E_{ij} \rightarrow frequências\ esperadas$.

Para calcular as frequências esperadas da tabela de contingência, basta multiplicarmos o total da linha pelo total da coluna e dividirmos pelo total geral. Assim, para a primeira linha temos as seguintes frequências esperadas:

$$E_{11} = \frac{L_1 \times C_1}{T} = \frac{20 \times 20}{100} = 4$$

$$E_{12} = \frac{L_1 \times C_2}{T} = \frac{20 \times 70}{100} = 14$$

$$E_{13} = \frac{L_1 \times C_3}{T} = \frac{20 \times 10}{100} = 2$$

Para a segunda linha, temos:

$$E_{21} = \frac{L_2 \times C_1}{T} = \frac{60 \times 20}{100} = 12$$

$$E_{22} = \frac{L_2 \times C_2}{T} = \frac{60 \times 70}{100} = 42$$

$$E_{23} = \frac{L_2 \times C_3}{T} = \frac{60 \times 10}{100} = 6$$

Para a terceira linha, temos:

$$E_{31} = \frac{L_3 \times C_1}{T} = \frac{20 \times 20}{100} = 4$$

$$E_{32} = \frac{L_3 \times C_2}{T} = \frac{20 \times 70}{100} = 14$$

$$E_{33} = \frac{L_3 \times C_3}{T} = \frac{20 \times 10}{100} = 2$$

Recriando a tabela, temos as

		A			
		-1	0	1	total (%)
B	-1	10 **(4)**	5 **(14)**	5 **(2)**	20
	0	0 **(12)**	60 **(42)**	0 **(6)**	60
	1	10 **(4)**	5 **(14)**	5 **(2)**	20
	total(%)	20	70	10	100

Vamos calcular $\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ para cada célula:

$$\frac{(O_{11} - E_{11})^2}{E_{11}} = \frac{(10-4)^2}{4} = 9$$

$$\frac{(O_{12} - E_{12})^2}{E_{12}} = \frac{(5-14)^2}{14} \cong 5{,}78$$

$$\frac{(O_{13} - E_{13})^2}{E_{13}} = \frac{(5-2)^2}{2} = 4{,}5$$

Para a segunda linha:

$$\frac{(O_{21} - E_{21})^2}{E_{21}} = \frac{(0-12)^2}{12} = 12$$

$$\frac{(O_{22} - E_{22})^2}{E_{22}} = \frac{(60-42)^2}{42} \cong 7{,}71$$

$$\frac{(O_{23} - E_{23})^2}{E_{23}} = \frac{(0-6)^2}{6} = 6$$

Para a segunda linha:

$$\frac{(O_{31} - E_{31})^2}{E_{31}} = \frac{(10-4)^2}{5} = 9$$

$$\frac{(O_{32} - E_{32})^2}{E_{32}} = \frac{(5-14)^2}{14} \cong 5{,}78$$

$$\frac{(O_{33} - E_{33})^2}{E_{33}} = \frac{(5-2)^2}{2} = 4{,}5$$

Agora, vamos calcular a estatística qui-quadrado dessa tabela de contingência:

$$\chi^2 = \sum_{j=1}^{c}\sum_{i=1}^{l}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 9 + 5{,}78 + 4{,}5 + 12 + 7{,}71 + 6 + 9 + 5{,}78 + 4{,}5 = 64{,}28$$

Calculando o coeficiente de contingência:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{64{,}28}{64{,}28 + 100}} = 0{,}624$$

Poderíamos ter parado a resolução da questão quando calculamos a primeira frequência esperada diferente de um valor observado, $O_{11} \neq E_{11}$, pois o coeficiente de contingência somente é nulo se cada frequência esperada é igual ao seu respectivo valor observado.

Gabarito: Errado.

Covariância e Correlação

Covariância

A covariância é uma medida de variação que caracteriza tanto a **força** da relação entre duas variáveis, quanto a sua **orientação** (se variam no **mesmo sentido** ou em **sentidos opostos**).

Quando os **maiores** valores de uma variável correspondem principalmente aos **maiores** valores da outra, e quando os **menores** valores de uma variável se relacionam com os **menores** valores da outra, dizemos que as variáveis tendem a apresentar comportamento semelhante, por isso a **covariância** é **positiva**.

Entretanto, quando os **maiores** valores de uma variável correspondem principalmente aos **menores** valores da outra variável, e quando os **menores** valores de uma se relacionam com os **maiores** valores da outra variável, dizemos que as variáveis tendem a apresentar comportamento oposto, por isso a **covariância** é **negativa**.

Assim, o sinal mostrará a tendência na relação linear entre as variáveis. Se o sinal for **negativo**, significa dizer que as variáveis têm **relação negativa**, isto é, quando uma aumenta, a outra diminui. Já se o sinal for **positivo**, significa que as variáveis têm **relação positiva**, isto é, quando uma aumenta, a outra também aumenta.

Quando a covariância é **positiva**, duas variáveis **tendem a variar na mesma direção**; isto é, se uma **aumenta**, a outra também tende a **aumentar** e vice-versa. Quando a covariância é **negativa**, duas variáveis **tendem a variar em direções opostas**; isto é, se uma **aumenta**, a outra tende a **diminuir** e vice-versa.

Devemos ter em mente que a covariância é uma medida de difícil interpretação, visto que seu resultado sofre muita influência da magnitude das variáveis e não apresenta um valor normalizado (padronizado), isto é, não possui uma base que possa ser utilizada em comparações futuras. Essa deficiência da variância é superada pelo coeficiente de correlação linear, que fornece um valor normalizado no intervalo entre -1 e 1.

Correlação Linear

A correlação linear também mede a força e a orientação com que duas variáveis se relacionam linearmente. A análise da correlação linear busca identificar se existe alguma relação entre duas variáveis, ou seja, se as alterações nas variáveis estão associadas umas com as outras.

Para avaliar a existência de correlação linear, recorremos a uma forma de representação gráfica bem simples, denominada de gráfico de dispersão. Basicamente, é uma representação de pares ordenados em um plano cartesiano, composto por um eixo vertical (ordenada) e um eixo horizontal (abcissa). A correlação linear pode ser:

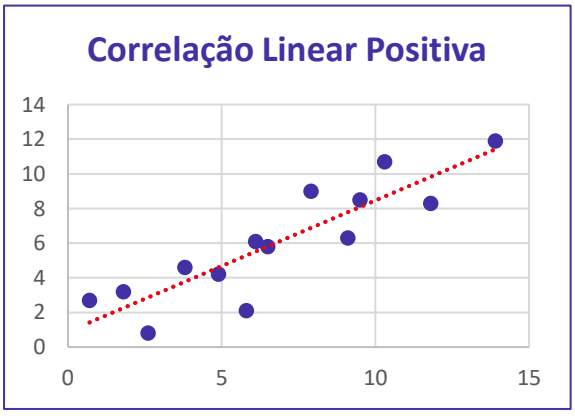
a) direta ou positiva – quando temos dois fenômenos que variam no mesmo sentido. Se aumentarmos ou diminuirmos um deles, o outro também aumentará ou diminuirá;

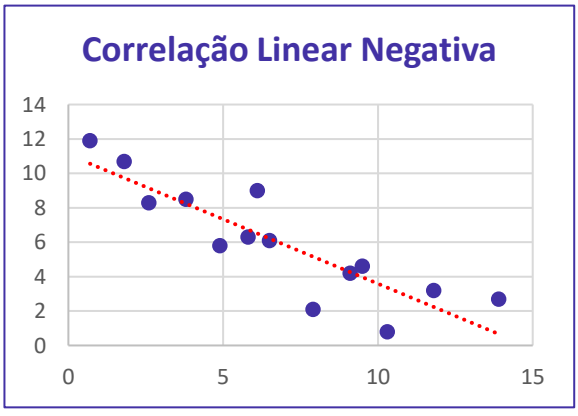
b) inversa ou negativa – quando temos dois fenômenos que variam em sentido contrário. Se aumentarmos ou diminuirmos um deles, acontecerá o contrário com o outro, no caso, diminuirá ou aumentará;

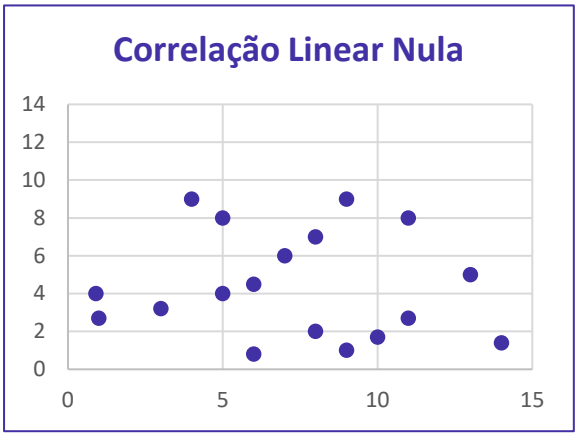
c) inexistente ou nula – quando não existe correlação ou dependência entre os dois fenômenos.

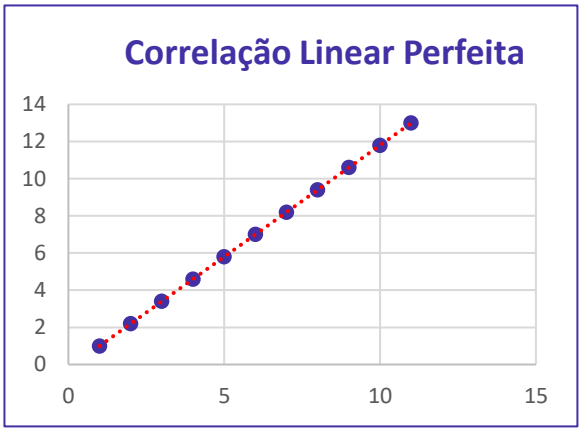
d) perfeita – quando os fenômenos se ajustam perfeitamente a uma reta.

As figuras a seguir ilustram essas quatro situações:

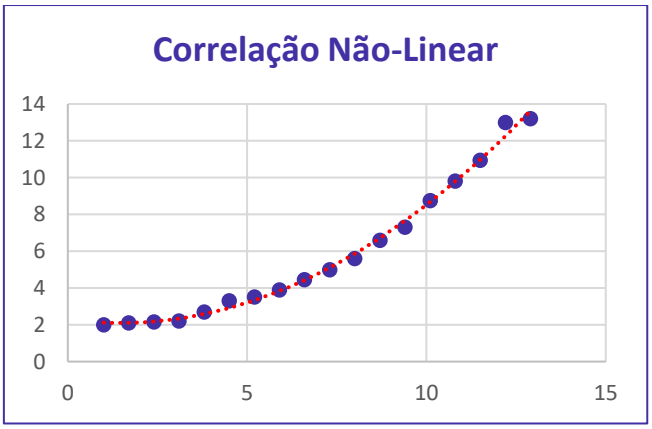








A construção do diagrama de dispersão é muito importante porque, em determinadas situações, pode acontecer de não existir uma relação linear entre as variáveis, mas, ainda assim, as variáveis estarem associadas de forma não-linear, como mostrado abaixo:



Assim, sempre que duas variáveis quantitativas apresentarem uma relação linear entre si, adotaremos o coeficiente de correlação linear de Pearson, definido pela letra grega ρ (rô), para medir a força dessa interação. Por outro lado, quando a relação for não linear, utilizaremos a correlação de Spearman e o Coeficiente de Contingência.

O coeficiente de correlação linear de Pearson é calculado por meio da seguinte expressão:

$$\rho = \frac{\sum_{i=1}^{n}[(X_i - \overline{X}) \times (Y_i - \overline{Y})]}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \times \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

O coeficiente de Pearson pode assumir quaisquer valores entre 1 e -1, ou seja:

$$-1 \leq \rho \leq 1$$

Assim, quanto mais próximo ρ estiver de 0, menor será a relação linear entre as duas variáveis. Por sua vez, quanto mais próximo ρ estiver de (1 ou -1), maior será a relação linear entre as duas variáveis.

O coeficiente de correlação linear também pode ser definido por meio das seguintes expressões:

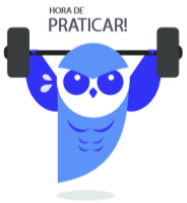
$$\boldsymbol{\rho = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}}}$$

Em que $S_{XY} = \sum_{i=1}^{n}[(X_i - \bar{X}) \times (Y_i - \bar{Y})]$; $S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2$ e $S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

Também pode aparecer na seguinte forma:

$$\boldsymbol{\rho = \frac{Cov(X,Y)}{\sigma_X \times \sigma_Y}}$$

Em que $Cov(X,Y)$ representa a covariância das variáveis X e Y; σ_X e σ_Y representam o desvio padrão, respectivamente, das variáveis X e Y.



(FUNIVERSA/PC-DF/2015 - ADAPTADA) Considerando que X e Y sejam variáveis aleatórias contínuas, com variâncias iguais a 16 e 4, respectivamente, e que a covariância entre X e Y seja igual a 6, a correlação linear de Pearson entre X e Y é igual a

a) 0,90.

b) 0,85.

c) 0,60.

d) 0,75.

e) -0,25.

Comentários:

O coeficiente de correlação entre duas variáveis X e Y é expresso por:

$$\rho = \frac{Cov(X,Y)}{\sigma_X \times \sigma_Y}.$$

Logo, basta dividir a covariância pelo produto dos desvios padrão, os quais podem ser calculados pela raiz quadrada da variância. Se a variância de X vale 16, seu desvio padrão é $\sqrt{16}$. Se a variância de Y vale 4, seu desvio padrão é $\sqrt{4}$:

$$\rho = \frac{6}{\sqrt{16} \times \sqrt{4}}$$

$$\rho = \frac{6}{4 \times 2}$$

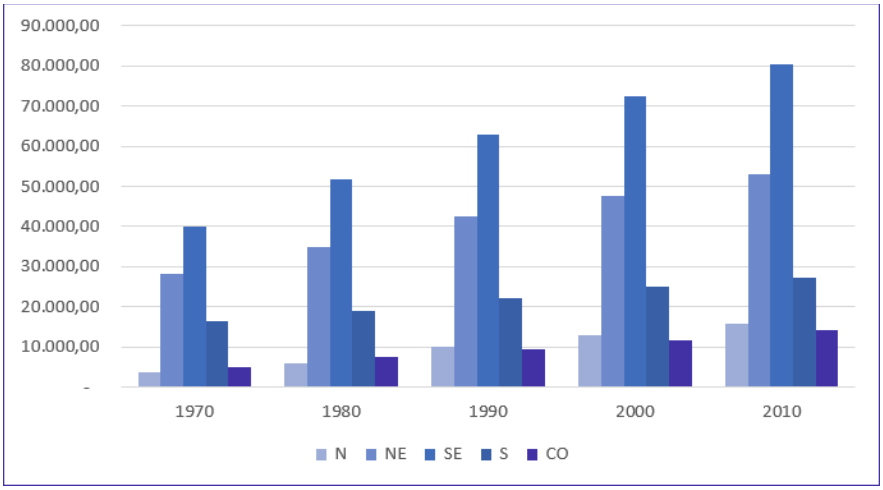
$$\rho = \frac{6}{8} = 0{,}75$$

Gabarito: D.

Visualização de Dados Multivariados

Gráficos de Colunas Justapostas/Empilhadas

Podemos utilizar o **gráfico de colunas justapostas** para analisar duas dimensões de uma variável categórica. Dessa maneira, conseguimos apresentar mais informações em um espaço consideravelmente menor. Vejamos o gráfico a seguir:



Adicionalmente, essas informações também podem ser representadas por meio de um **gráfico de colunas sobrepostas** (ou **gráfico de colunas empilhadas**). Cada coluna é dividida em várias partes que ficam empilhadas umas sobre as outras, cada uma correspondendo a um nível da segunda variável categórica.

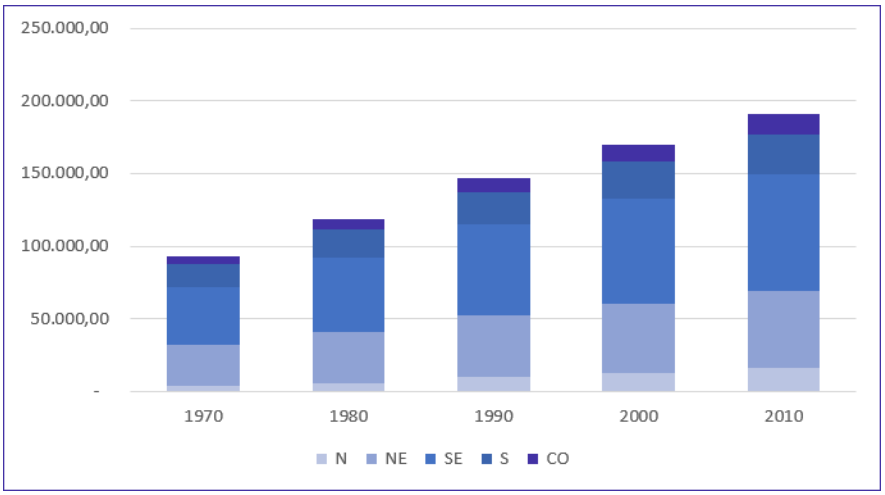


Gráfico em Linha

Os gráficos em linha normalmente são usados para representar dados de séries temporais, com a finalidade de mostrar a variação dos valores de uma variável ao longo do tempo. Esse tipo de gráfico permite-nos comparar duas variáveis: uma é traçada no eixo x (horizontal) e a outra no eixo y (vertical). O eixo y geralmente indica uma quantidade, enquanto o eixo x representa uma unidade de tempo.

Histogramas Comparativos

Histogramas podem ser utilizados para comparar o comportamento de uma variável com relação a diferentes categorias de uma outra variável, como mostra a figura a seguir.

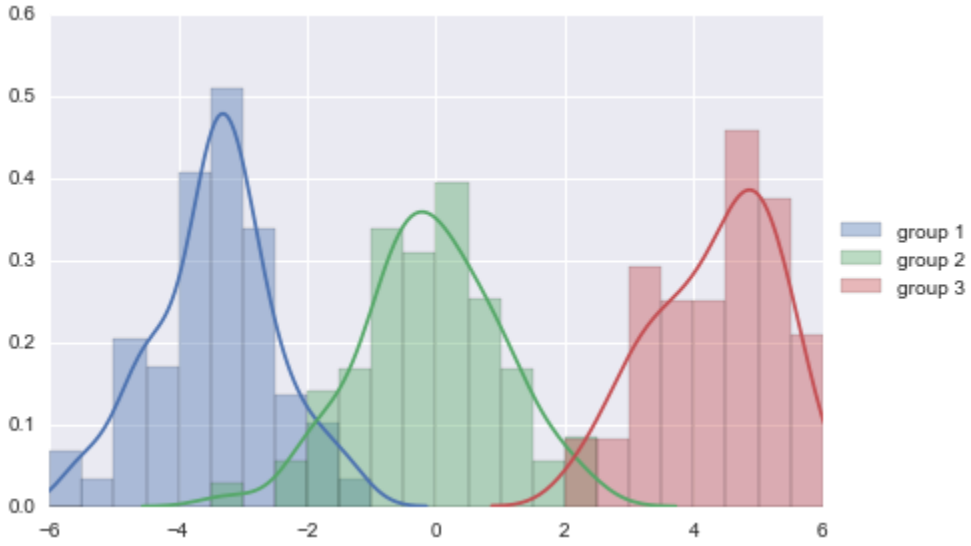
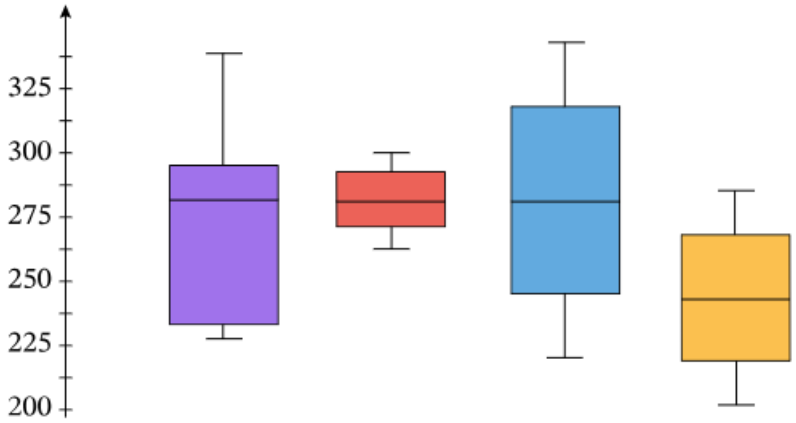


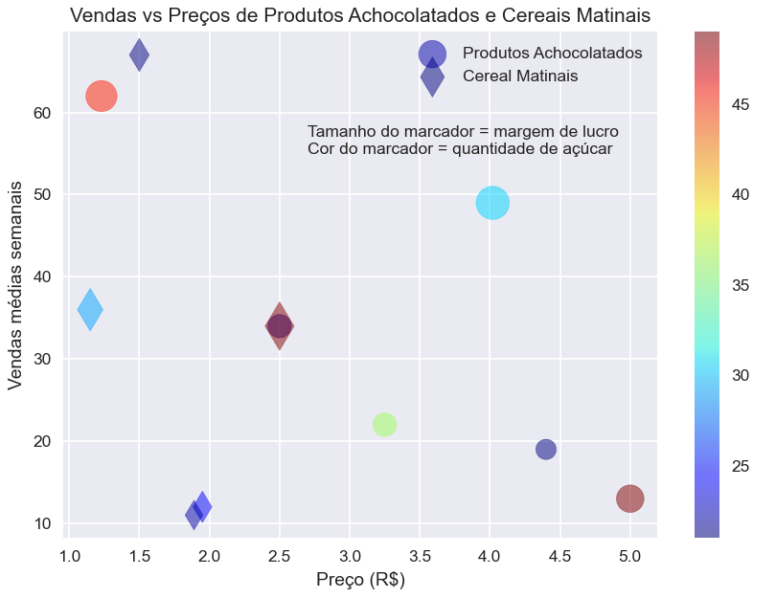
Diagrama de Box-Plot Comparativo

A representação de vários diagramas de box-plot lado a lado permite a comparação das características de várias categorias ao mesmo tempo.



Gráficos de Dispersão

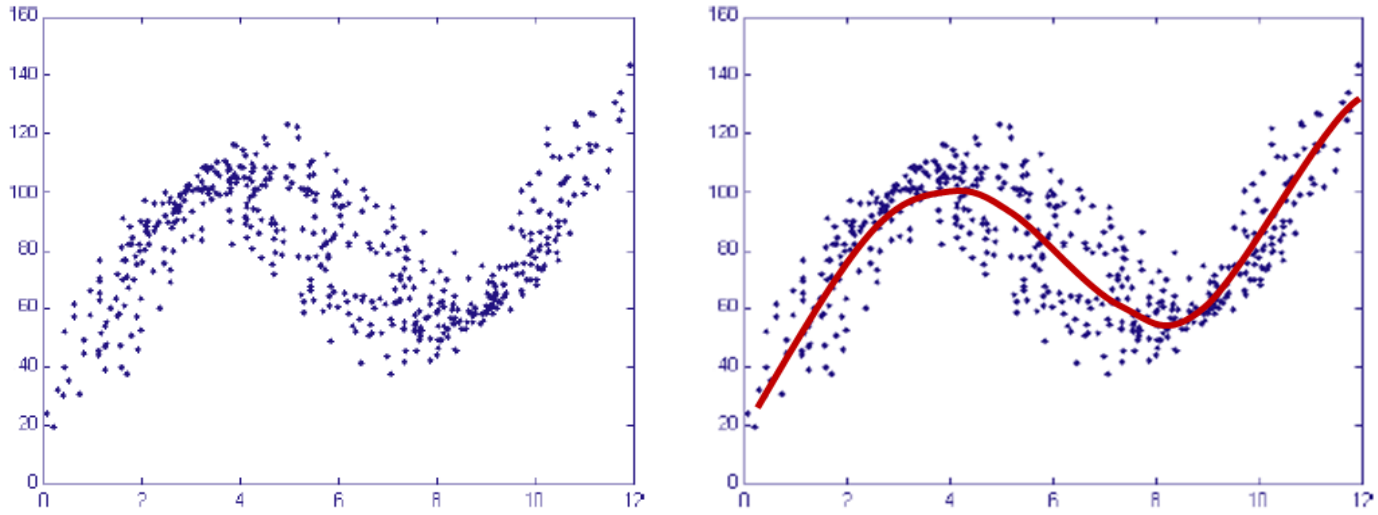
Os gráficos de dispersão são construídos usando duas variáveis quantitativas contínuas, ordinais ou discretas. A coordenada de cada ponto de dados corresponde a uma variável. Eles podem ser codificados em até cinco dimensões usando outras variáveis, diferenciando o tamanho, a forma ou a cor dos pontos de dados.



Curvas de Ajuste

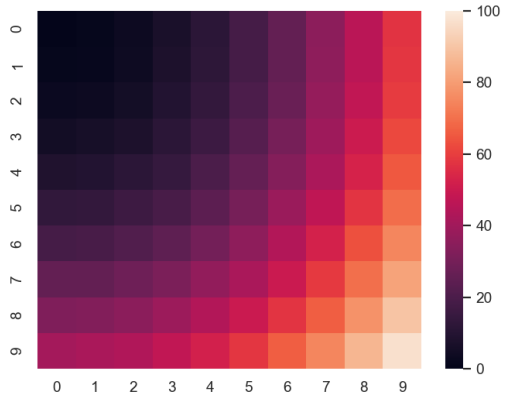
O ajuste de curvas é uma maneira de quantificar a relação entre duas variáveis ou a mudança nos valores ao longo do tempo. O método mais comum para o ajuste se baseia em minimizar a soma dos quadrados médios dos erros entre os dados e a função ajustada. Existe um conjunto cada vez maior de métodos para lidar com o ajuste:

- adicionar variáveis explicativas transformadas, por exemplo, adicionar x^2 ou x^3 ao modelo;
- usar outros métodos para lidar com relações mais complexas entre variáveis (por exemplo, regressão polinomial, máquinas de vetor de suporte, etc.).

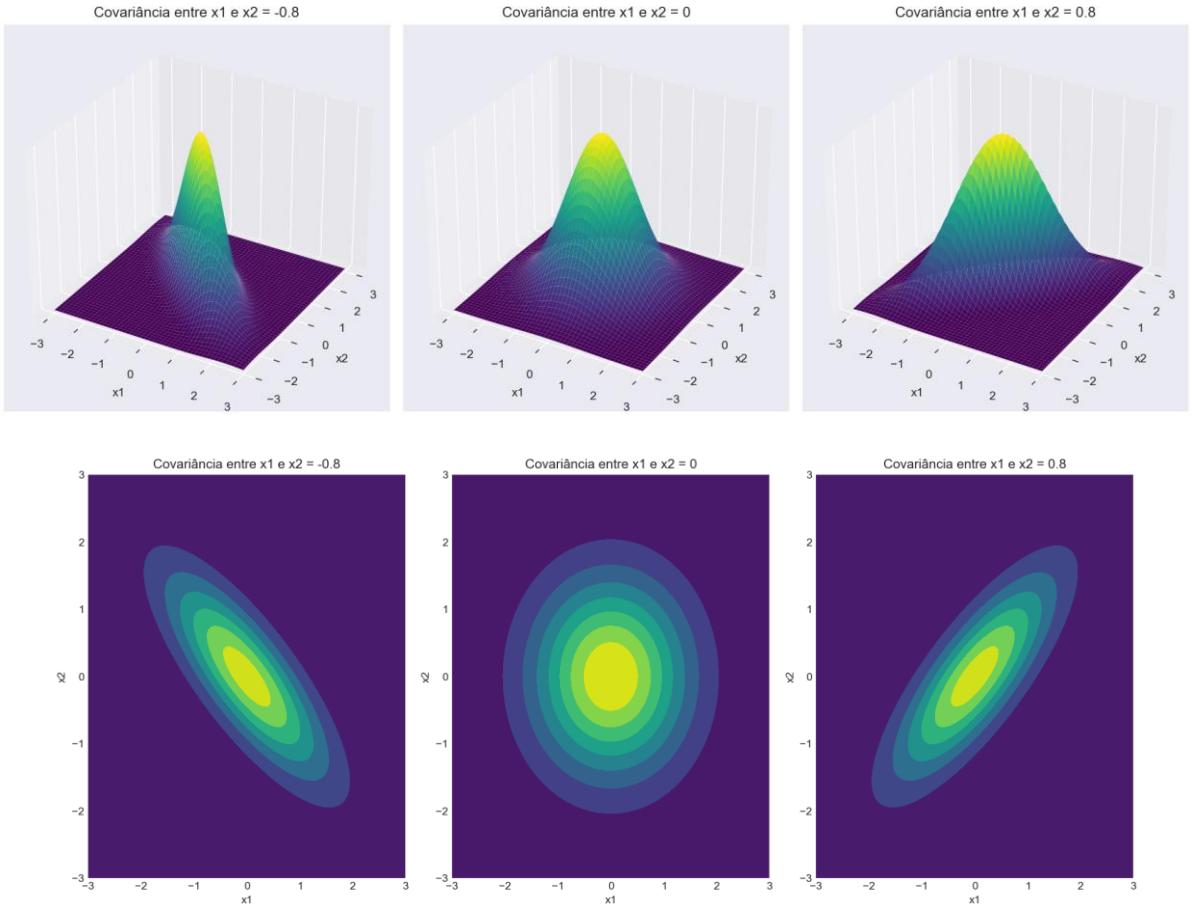


Mapas de Calor e Gráficos de Superfície 3D

Por fim, os mapas de calor são simplesmente uma matriz (duas dimensões), cuja cor depende dos valores de duas variáveis, representadas nas linhas e colunas. Essa técnica é útil quando desejamos representar a mudança de uma variável em função de duas outras variáveis. As cores podem ser personalizadas (por exemplo, arco-íris ou tons de cinza).



O equivalente em 3 dimensões são os gráficos de malha ou curvas de superfície.



QUESTÕES COMENTADAS – CEBRASPE

Conceitos Iniciais

1. (CESPE/PGDF/2021) Certa empresa desejava conhecer as opiniões de seus 20.000 funcionários acerca da confiança que eles têm no canal interno de denúncias. Para tanto, elaborou-se um questionário eletrônico que foi remetido, por email, para todos os endereços eletrônicos cadastrados, tendo sido desenvolvidos mecanismos para evitar que uma pessoa respondesse em lugar de outra, ou que uma mesma pessoa respondesse mais de uma vez. O questionário foi respondido por 400 pessoas, das quais 68% disseram confiar no processo de apuração de denúncias e 32% disseram ter reservas quanto ao processo. Verificou-se ainda que cerca de 500 mensagens retornaram por falha no cadastro dos endereços eletrônicos (erros de digitação), e que algumas respostas foram atribuídas a pessoas que não são mais funcionários; ainda, os endereços eletrônicos de alguns funcionários recém contratados não constavam do cadastro.

Com relação a essa situação hipotética, julgue o item a seguir.

As informações apresentadas permitem afirmar que a população- alvo da pesquisa difere da população referenciada.

Comentários:

Vamos analisar as informações contidas no enunciado da questão:

i) a empresa desejava conhecer a opinião dos seus 20.000 funcionários, logo, essa é a população-alvo da pesquisa;

ii) o questionário foi enviado por e-mail aos funcionários cadastrados, portanto, essa é a população referenciada da pesquisa;

iii) apenas 400 (quatrocentas) pessoas responderam ao questionário, tendo havido falha na entrega de 500 mensagens;

iv) da população referenciada, alguns funcionários não trabalhavam mais na empresa; e

v) da população-alvo, alguns funcionários novos não estavam cadastrados no e-mail.

Assim, podemos concluir que a população-alvo não foi alcançada, visto que era diferente da população referenciada, isto é, das pessoas cadastradas no serviço de e-mail.

Gabarito: Certo.

2. (CESPE/DEPEN/2015) O diretor de um sistema penitenciário, com o propósito de estimar o percentual de detentos que possuem filhos, entregou a um analista um cadastro com os nomes de 500 detentos da instituição para que esse profissional realizasse entrevistas com os indivíduos selecionados.

A partir dessa situação hipotética e dos múltiplos aspectos a ela relacionados, julgue o item, referente a técnicas de amostragem.

A diferença entre um censo e uma amostra consiste no fato de esta última exigir a realização de um número maior de entrevistas.

Comentários:

Um censo é a análise de todos os elementos de determinada população. Trabalhamos com uma amostra quando obtemos informações de apenas uma parte dessa população. Portanto, o censo requer a realização de um número maior de entrevistas.

Gabarito: Errado.

3. (CESPE/SEFAZ-AL/2002) Julgue os seguintes itens.

Um censo consiste no estudo de todos os indivíduos da população considerada.

Comentários:

O conceito apresentado está correto. O censo consiste na obtenção dos dados de todos os elementos da população considerada.

Gabarito: Certo.

4. (CESPE/SEFAZ-AL/2002) Julgue os seguintes itens.

Como a realização de um censo tipicamente é muito onerosa e (ou) demorada, muitas vezes é conveniente estudar um subconjunto próprio da população, denominado amostra.

Comentários:

O item está correto. Como normalmente é muito caro e demorado obter os dados de toda a população, trabalhamos apenas com uma parte dela, que chamamos de amostra.

Gabarito: Certo.

QUESTÕES COMENTADAS – CEBRASPE

Variáveis Estatísticas

1. (CESPE/DPE-RO/2022)

Variável	Valores
estado civil	casado, solteiro, divorciado
quantidade de filhos	0, 1, 2, 3 ...
salário	6.510,25; 7.915,68
idade	22, 23, 27

Com relação às variáveis apresentadas na tabela anterior, julgue os itens a seguir.

I. A variável estado civil é qualitativa nominal.

II. A variável quantidade de filhos é quantitativa discreta.

III. As variáveis salário e estado civil são quantitativas discretas.

IV. As variáveis idade e quantidade de filhos são qualitativas nominais.

Estão certos apenas os itens

a) I e II.

b) II e III.

c) III e IV.

d) I, II e IV.

e) I, III e IV.

Comentários:

Vamos analisar cada variável da tabela:

A variável "estado civil" é qualitativa nominal. Dizemos que a variável "estado civil" é qualitativa pois ela não possui um sentido numérico ou quantitativo. Também podemos afirmar que é nominal por não possuir significado matemático, logo, as possíveis categorias não podem ser ordenadas.

A variável "quantidade de filhos" é quantitativa e discreta. É quantitativa porque possui valor quantitativo, exprime uma quantidade. Também podemos afirmar que a variável é discreta porque ela possui uma

quantidade finita de valores. Os possíveis valores formam um conjunto finito ou enumerável de números e, em geral, resultam de um processo de contagem.

A variável "salário" é quantitativa e discreta. É quantitativa pois possui valor quantitativo, exprimindo uma quantidade. Também podemos afirmar que é discreta porque possui uma quantidade finita de valores.

A variável "idade" é quantitativa e discreta. É quantitativa pois possui valor quantitativo, exprime uma quantidade. Também podemos afirmar que é discreta porque possui uma quantidade finita de valores.

Portanto, analisando os itens, podemos constatar que os únicos corretos são os itens I e II.

Gabarito: A.

2. (CESPE/DPE-RO/2022) O valor de um atributo de um dado objeto é uma medida da quantidade daquele atributo, a qual pode ser numérica ou categórica.

Nesse caso, estado civil e sexo são classificados como atributo

a) binário.

b) nominal.

c) ordinal.

d) ausente.

e) razão.

Comentários:

As variáveis qualitativas representam as características que não podem ser descritas de forma numérica, mas que podem ser definidas por meio de qualidades (atributos ou categorias) do indivíduo pesquisado. No caso, as variáveis "estado civil" e "sexo" são qualitativas nominais (ou categóricas), pois as categorias não podem ser ordenadas nem possuem significado matemático.

Gabarito: B.

3. (CESPE/IPHAN/2018). Julgue o item subsequente, referente à análise exploratória de dados.

O gráfico de barras é adequado para a análise de variáveis qualitativas ordinais ou quantitativas discretas, pois permite investigar a presença de tendência nos dados.

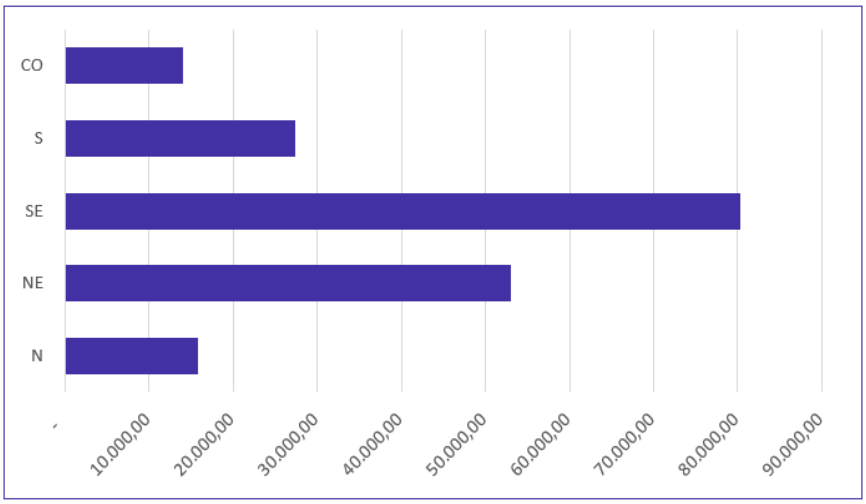
Comentários:

O gráfico em barras normalmente é usado para representar distribuições de dados categóricos ou qualitativos. Por meio desse gráfico, uma série estatística é representada por um conjunto de retângulos

dispostos horizontalmente, cada um indicando uma categoria em particular, os quais possuem a mesma altura e comprimentos proporcionais aos respectivos dados.

Para ficar claro, apresentamos um gráfico de barras como exemplo:

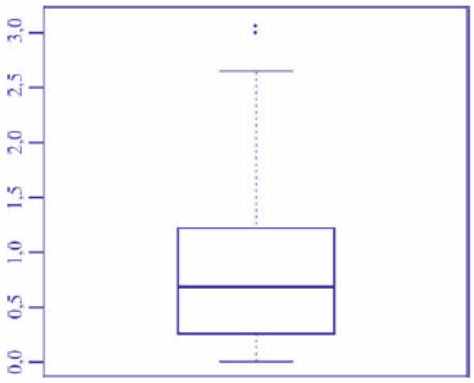
População brasileira, por Grandes Regiões, em 2010 (x1000)



Fonte: Censo Demográfico 1970/2010 (IBGE)

Gabarito: Certo.

4. (CESPE/TCE-PA/2016)



média amostral	0,80
desvio padrão amostral	0,70
primeiro quartil	0,25
mediana	0,70
terceiro quartil	1,20
mínimo	0
máximo	3,10

Um indicador de desempenho X permite avaliar a qualidade dos processos de governança de instituições públicas. A figura mostra, esquematicamente, a sua distribuição, obtida mediante estudo amostral feito por determinada agência de pesquisa. A tabela apresenta estatísticas descritivas referentes a essa distribuição.

Com base nessas informações, julgue o item a seguir.

X representa uma variável qualitativa ordinal.

Comentários:

As variáveis podem ser classificadas em:

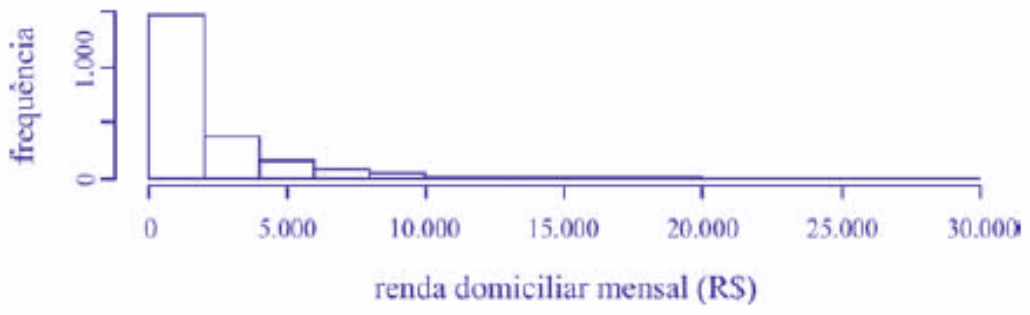
a) variáveis quantitativas: são as características que podem ser medidas em uma escala quantitativa, isto é, numérica, podendo ser contínuas ou discretas.

b) variáveis qualitativas (ou categóricas): são as características que não possuem valores quantitativos. Nesse caso, são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos. Podem ser nominais ou ordinais.

Analisando as informações dadas na questão, percebemos que X pode adotar valores que vão de 0 (mínimo) a 3,10 (máximo). Logo, como X assume valores numéricos, dizemos que ela é uma variável quantitativa.

Gabarito: Errado.

5. (CESPE/TELEBRAS/2015)



Uma empresa coletou e armazenou em um banco de dados diversas informações sobre seus clientes, entre as quais estavam o valor da última fatura vencida e o pagamento ou não dessa fatura. Analisando essas informações, a empresa concluiu que 15% de seus clientes estavam inadimplentes. A empresa recolheu ainda dados como a unidade da Federação (UF) e o CEP da localidade em que estão os clientes. Do conjunto de todos os clientes, uma amostra aleatória simples constituída por 2.175 indivíduos prestou também informações sobre sua renda domiciliar mensal, o que gerou o histograma apresentado. Com base nessas informações e no histograma, julgue o item a seguir.

O CEP da localidade dos clientes e o valor da última fatura vencida são variáveis quantitativas.

Comentários:

O CEP, embora possua uma representação numérica, não exprime quantidades e, portanto, não é uma variável quantitativa. Esse código número apenas qualifica as ruas de um determinado endereço, podendo ser classificada como uma variável qualitativa.

O valor da última fatura vencida, por sua vez, é sim uma variável quantitativa.

Gabarito: Errado.

6. (CESPE/TELEBRAS/2015) Roberto comprou, por R$ 2.800,00, rodas de liga leve para seu carro, e, ao estacionar no shopping, ficou indeciso sobre onde deixar o carro, pois, caso o coloque no estacionamento público, correrá o risco de lhe roubarem as rodas, ao passo que, caso o coloque no estacionamento privado, terá de pagar R$ 70,00, com a garantia de que eventuais prejuízos serão ressarcidos pela empresa administradora.

Considerando que p seja a probabilidade de as rodas serem roubadas no estacionamento público, que X seja a variável aleatória que representa o prejuízo, em reais, ao deixar o carro no estacionamento público, e que Y seja a variável aleatória que representa o valor, em reais, desembolsado por Roberto ao deixar o carro no estacionamento pago, julgue o item subsequente.

A variável aleatória Y é contínua.

Comentários:

As variáveis quantitativas podem ser classificadas em:

a) variáveis contínuas: características mensuráveis que assumem valores em uma escala contínua, nesse caso, valores fracionários fazem sentido;

b) variáveis discretas: características mensuráveis que podem assumir apenas um número finito ou infinito contável de valores e, assim, somente fazem sentido valores inteiros.

No problema em questão, a variável aleatória Y pode assumir o valor R$ 70,00, se Roberto deixar o carro no estacionamento pago; ou o valor R$ 0,00, se Roberto não deixar o carro no estacionamento pago. Portanto, Y é uma variável aleatória discreta.

Gabarito: Errado.

7. (CESPE/TJ-SE/2014)

Quantidade	São Paulo (j =1)	Rio de Janeiro (j =2)	Minas Gerais (j =3)	Rio Grande do Sul (j =4)	Total
Casos novos (X, em milhões)	5	2	1	2	**18**
Casos pendentes (Y, em milhões)	16	8	3	2	**48**
Processos baixados (Z, em milhões)	5	2	2	2	**18**

Sentenças e Decisões (W, em milhões)	4	3	1	1	**16**

O quadro acima mostra uma síntese da movimentação processual dos tribunais de justiça dos estados de São Paulo, Rio de Janeiro, Minas Gerais, Rio Grande do Sul e do total da justiça estadual no Brasil em 2010. Considere que o estoque de processos em andamento no estado j (Ej), no final de 2010, seja um indicador que se define como $E_j = X_j + Y_j - Z_j - W_j$, em que j = 1, 2, ..., 27; X_j representa o número de casos novos registrados em 2010 no estado j; Yj seja a quantidade de casos pendentes no estado j (i.e., casos anteriores que não foram solucionados até o final de 2010); Z_j denota o total de processos baixados (arquivados) no estado j durante 2010 e W_j seja o número de sentenças e decisões proferidas no estado j até o final de 2010. Considere, por fim, que, para todos os efeitos, o Distrito Federal seja um estado. Com base nessas informações e no quadro acima, julgue o item que se segue.

O quadro apresentado é uma tabela de contingência que mostra o cruzamento entre uma variável qualitativa nominal com 4 níveis de resposta (estados) e outra variável qualitativa com quatro níveis de resposta (casos novos, pendentes, baixados e resolvidos).

Comentários:

As variáveis podem ser classificadas em:

a) variáveis quantitativas: são as características que podem ser medidas em uma escala quantitativa, isto é, numérica. Elas podem ser contínuas ou discretas:

i) variáveis discretas: características mensuráveis que podem assumir apenas um número finito ou infinito contável de valores e, assim, somente fazem sentido valores inteiros. Geralmente são o resultado de contagens. Exemplos: número de membros de uma família, número de carros produzidos por dia.

ii) variáveis contínuas: características mensuráveis que assumem valores em uma escala contínua (na reta real), nesse caso, valores fracionários fazem sentido. Geralmente são medidas por meio de algum instrumento. Exemplos: peso e altura.

b) variáveis qualitativas (ou categóricas): são as características que não possuem valores quantitativos. Nesse caso, são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos. Elas podem ser nominais ou ordinais:

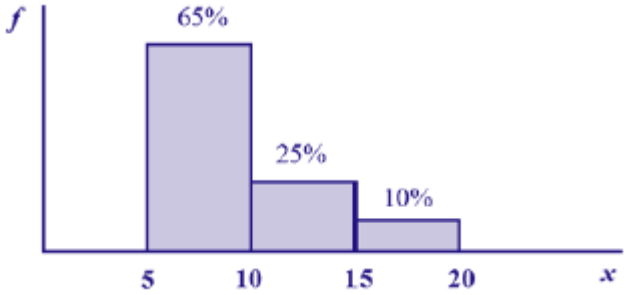
i) variáveis nominais: não existe uma ordenação entre as categorias. Exemplos: sexo, estado civil, cor dos olhos.

ii) variáveis ordinais: existe uma ordenação entre as categorias. Exemplos: grau de instrução (fundamental, médio, superior), nível de aprendizado (básico, intermediário, avançado), nível de queimadura (1°, 2°, 3° graus).

No problema em questão, a tabela não representa uma variável qualitativa, mas sim quantitativa, vez que podemos atribuir valores às variáveis X, Y, Z e W.

Gabarito: Errado.

8. (CESPE/STF/2013)



Com referência à figura acima, que mostra a distribuição da renda mensal — x, em quantidades de salários mínimos (sm) — das pessoas que residem em determinada região, julgue o item subsequente.

A variável x, por possuir quatro níveis de respostas, é do tipo qualitativa ordinal.

Comentários:

O gráfico representa um histograma, em que as pessoas foram classificadas de acordo com as quantidades de salários-mínimos recebidos. Reparem na existência de três intervalos de classe (5 a 10; 10 a 15; e 15 a 20). Portanto, a variável x exprime a quantidade de salários-mínimos, sendo classificada como uma variável quantitativa.

Gabarito: Errado.

9. (CESPE/TRE-ES/2011)

Quantidade de eleitores	Quantidade de municípios
$0 \vdash 2.000$	364
$2.000 \vdash 4.000$	1.000
$4.000 \vdash 6.000$	3.000
$6.000 \vdash 8.000$	1.000
$8.000 \vdash 10.000$	200

Total	**5.564**

A tabela acima apresenta uma distribuição hipotética das quantidades de eleitores que não votaram no segundo turno da eleição para presidente da República bem como os números de municípios em que essas quantidades ocorreram. Com base nessa tabela, julgue o item seguinte, relativo à análise exploratória de dados.

Na tabela de frequências, o uso de intervalos de classe permite concluir que a variável em questão é contínua.

Comentários:

As variáveis quantitativas podem ser classificadas em:

a) variáveis contínuas: características mensuráveis que assumem valores em uma escala contínua, nesse caso, valores fracionários fazem sentido;

b) variáveis discretas: características mensuráveis que podem assumir apenas um número finito ou infinito contável de valores e, assim, somente fazem sentido valores inteiros.

A questão apenas agrupou a quantidade de eleitores em classes, mas continua sendo uma variável discreta.

Gabarito: Errado.

10. (CESPE/TRE-ES/2011)

Cargo	Candidatos	Candidatos aptos	Eleitos
Presidente da República	9	9	1
Governador de Estado	170	156	27
Senador	272	234	54
Deputado Federal	6.021	5.058	513
Deputado Estadual/Distrital	15.268	13.076	1.059
Total	21.640	18.533	1.658

Internet: <www.tse.gov > (com adaptações).

Com base na tabela acima, referente às eleições de 2010, que apresenta a quantidade de candidatos para os cargos de presidente da República, governador de estado, senador, deputado federal e deputado

estadual/distrital, bem como a quantidade de candidatos considerados aptos pela justiça eleitoral e o total de eleitos para cada cargo pretendido, julgue o item a seguir.

A variável "cargo" classifica-se como uma variável qualitativa ordinal.

Comentários:

As variáveis podem ser classificadas em:

a) variáveis quantitativas: são as características que podem ser medidas em uma escala quantitativa, isto é, numérica. Elas podem ser contínuas ou discretas:

i) variáveis discretas: características mensuráveis que podem assumir apenas um número finito ou infinito contável de valores e, assim, somente fazem sentido valores inteiros. Geralmente são o resultado de contagens. Exemplos: número de membros de uma família, número de carros produzidos por dia;

ii) variáveis contínuas: características mensuráveis que assumem valores em uma escala contínua (na reta real), nesse caso, valores fracionários fazem sentido. Geralmente são medidas por meio de algum instrumento. Exemplos: peso e altura.

b) variáveis qualitativas (ou categóricas): são as características que não possuem valores quantitativos. Nesse caso, são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos. Elas podem ser nominais ou ordinais:

i) variáveis nominais: não existe uma ordenação entre as categorias. Exemplos: sexo, estado civil, cor dos olhos;

ii) variáveis ordinais: existe uma ordenação entre as categorias. Exemplos: grau de instrução (fundamental, médio, superior), nível de aprendizado (básico, intermediário, avançado), nível de queimadura (1°, 2°, 3° graus).

Com base nesses conceitos, podemos afirmar que o cargo é realmente uma variável qualitativa, pois classifica os indivíduos, e ordinal, pois estabelece uma ordem para os cargos (a tabela está ordenada pela coluna número de eleitos).

Gabarito: Certo.

11. (CESPE/TCU/2008) Uma agência de desenvolvimento urbano divulgou os dados apresentados na tabela a seguir, acerca dos números de imóveis ofertados (X) e vendidos (Y) em determinado município, nos anos de 2005 a 2007.

Ano	Número de imóveis	
	Ofertados (X)	Vendidos (Y)

2005	1.500	100
2006	1.750	400
2007	2.000	700

Correios Braziliense, 29/4/2008, p.17 (com adaptações)

Com respeito ao texto, considere que cada imóvel ofertado em determinado ano seja classificado como vendido ou não-vendido, e, a um imóvel e classificado como vendido seja atribuído um valor Z = 1, e, ao imóvel classificado como não-vendido, seja atribuído um valor Z = 0. Supondo-se que as classificações dos imóveis como vendido ou não-vendido em um dado ano possam ser consideradas como sendo realizações de uma amostragem aleatória simples, julgue os itens a seguir.

A variável Z é classificada como variável qualitativa nominal, pois representa o atributo do imóvel como vendido ou não-vendido.

Comentários:

Analisando os dados fornecidos no enunciado, observamos que a variável Z é uma variável qualitativa que assume os valores nominais “vendido” ou “não vendido”. No entanto, a questão atribuiu valores numéricos à variável, 1 para vendido e 0 para não vendido. Diante disso, a variável Z também passou a ser classificada como variável quantitativa.

Gabarito: Errado.

QUESTÕES COMENTADAS – CEBRASPE

Séries Estatísticas

1. (CESPE/ANTAQ/2009)

	Variável	2003	2004	2005	2006	2007
Exportação	X	40	46	50	52	54
Importação	Y	20	21	22	24	27
Total	X+Y	60	67	72	76	81

Internet: <www.portodesantos.com> (com adaptações).

Considerando a tabela acima, que apresenta a movimentação anual de cargas no porto de Santos de 2003 a 2007, em milhões de toneladas/ ano e associa as quantidades de carga movimentadas para exportação e importação às variáveis X e Y, respectivamente, julgue o item subsequente.

As séries estatísticas apresentadas na tabela formam três séries temporais.

Comentários:

Sabemos que uma série temporal é uma sequência de dados observados em um período de tempo. No caso da tabela, temos três séries cronológicas correspondentes às variáveis X (exportação); Y (importação); e X+Y (total).

Gabarito: Certo.

2. (CESPE/TCU/2008) Uma agência de desenvolvimento urbano divulgou os dados apresentados na tabela a seguir, acerca dos números de imóveis ofertados (X) e vendidos (Y) em determinado município, nos anos de 2005 a 2007.

Ano	Número de imóveis	
	Ofertados (X)	Vendidos (Y)
2005	1.500	100
2006	1.750	400

2007	2.000	700

Correios Braziliense, 29/4/2008, p.17 (com adaptações)

Considerando as informações do texto, julgue os itens subsequentes.

A variável X forma uma série estatística denominada série temporal.

Comentários:

Vamos analisar os elementos da tabela. Temos uma série em que o fato (número de imóveis) e o local (um determinado município) permanecem constantes, mas há variação do tempo (2005, 2006, 2007). Portanto, temos todos os elementos para classificar a série em temporal ou cronológica.

Gabarito: Certo.

QUESTÕES COMENTADAS – CEBRASPE

Distribuições de Frequência

1. (CEBRASPE/BACEN/2024)

x	$P(X = x)$
-2	$5c$
-1	c
0	$2c$
+1	$3c$
+2	$4c$

Considerando que X representa uma variável aleatória com suporte $x \in \{-2, -1, 0, +1, +2\}$, cuja função de distribuição de probabilidade é dada no quadro acima, na qual c é uma constante real positiva, julgue os próximos itens.

X segue uma distribuição contínua, pois C é uma constante real positiva.

Comentários:

Não, o fato de c ser uma constante real positiva não implica que X siga uma distribuição contínua. Na verdade, a distribuição apresentada é uma distribuição discreta.

Uma variável aleatória discreta toma valores em um conjunto finito ou contável de pontos, e cada valor tem uma probabilidade associada. No exemplo, X pode assumir os valores −2, −1, 0, +1 e +2, que são um conjunto finito de valores.

As probabilidades associadas a cada valor de X são $P(X = -2) = 5c$, $P(X = -1) = c$, e assim por diante, o que caracteriza uma distribuição discreta.

Gabarito: Errado.

2. (CEBRASPE/BACEN/2024)

x	$P(X = x)$
-2	$5c$
-1	c
0	$2c$
+1	$3c$
+2	$4c$

Considerando que X representa uma variável aleatória com suporte $x \in \{-2, -1, 0, +1, +2\}$, cuja função de distribuição de probabilidade é dada no quadro acima, na qual c é uma constante real positiva, julgue os próximos itens.

$P(X = +1) = 0{,}2$.

Comentários:

Para determinar o valor de $P(X = +1) = 0{,}2$, precisamos encontrar o valor da constante c que satisfaz as condições da distribuição de probabilidade. A soma das probabilidades para todas as possíveis realizações de X deve ser igual a 1, ou seja:

$$P(X = -2) + P(X = -1) + P(X = 0) + P(X = +1) + P(X = +2) = 1$$

Substituindo os valores dados:

$$5c + c + 2c + 3c + 4c = 1$$

Somando os termos:

$$15c = 1$$

$$c = \frac{1}{15}$$

Agora que sabemos o valor de c, podemos calcular as probabilidades $P(X = +1)$:

$$P(X = +1) = 3c = 3 \times \frac{1}{15} = \frac{3}{15} = \frac{1}{5} = 0{,}2$$

Portanto, a probabilidade $P(X = +1) = 0{,}2$.

Gabarito: Correto.

3. (CEBRASPE/CNJ/2024)

X	frequência absoluta acumulada
0	120
1	180
2	220
3	240
4	250

A tabela precedente mostra a distribuição de frequências do número diário (X) de denúncias recebidas pela ouvidoria de um tribunal de justiça.

Com base nos dados apresentados na tabela, julgue os itens que se seguem.

O tamanho da amostra é igual a 1.010.

Comentários:

O tamanho da amostra é a soma de todas as frequências simples:

$$120 + 60 + 40 + 20 + 10 = 250$$

Portanto, o tamanho da amostra não é 1.010, mas sim 250.

Gabarito: Errado.

4. (CEBRASPE/TC-DF/2023)

X	Frequência Absoluta	Frequência Relativa
0	3	0,10
5	6	0,20
10	15	0,50
15	6	0,20

Considerando que, em um levantamento estatístico realizado por amostragem aleatória simples, tenha sido produzida a tabela de frequências apresentada anteriormente, na qual X denota uma variável de interesse, julgue os seguintes itens.

O tamanho da amostra é igual ou superior a 16.

Comentários:

Para encontrarmos o tamanho da amostra, teremos que somar todos os valores da coluna de frequências absolutas:

$$3 + 6 + 15 + 6 = 30$$

Portanto, a amostra é composta por 30 elementos.

Gabarito: Certo.

5. (CEBRASPE/FUB/2022) A tabela de frequência a seguir mostra dados coletados em uma pesquisa para se verificar o número de disciplinas que os estudantes de determinada universidade estão cursando por semestre.

Disciplinas	2	3	4	5	6	7	8
Estudantes	10	15	40	35	28	10	4

Considerando essas informações, julgue o item seguinte.

A proporção de alunos que cursam mais de 6 disciplinas é maior que a proporção de alunos que cursam 3 disciplinas.

Comentários:

Calculando as proporções apresentadas, verificamos que 15 alunos cursam 3 disciplinas e que 14 alunos cursam 7 e 8 disciplinas. Temos ainda um total de 142 alunos. Portanto, segue:

Para 3 disciplinas, a proporção de alunos é de:

$$\frac{15}{142} = 0{,}10 = 10{,}56\%$$

Para mais de 6 disciplinas, a proporção de alunos é de:

$$\frac{10}{142} = 0{,}07 = 7\%$$

$$\frac{4}{142} = 0{,}028 = 2{,}8\%$$

E somando as proporções de 7 e 8 disciplinas, temos:

$$0{,}07 + 0{,}028 = 0{,}098 = 9{,}8\%$$

Logo, a proporção de alunos que cursam mais de 6 disciplinas é **menor** que a proporção de alunos que cursam 3 disciplinas.

Gabarito: Errado.

6. (CEBRASPE/FUB/2022) Uma universidade está fazendo um estudo para verificar a distribuição dos tempos que os alunos do curso de mestrado levam até a defesa da dissertação. Os dados a seguir mostram a função de probabilidade desses tempos, em meses.

Tempo de Defesa (meses)	Probabilidade
12	0,01

15	0,02
18	0,04
20	0,10
22	0,22
24	0,31
25	0,18
26	0,04
28	0,03
30	0,05

Considerando essas informações, julgue o item subsequente.

Se o prazo máximo recomendado para a defesa da dissertação de mestrado é de 24 meses, então a probabilidade de um aluno defender sua dissertação até 2 meses antes desse prazo é igual à probabilidade de um aluno defendê-la até 2 meses depois.

Comentários:

Analisando os dados da tabela, observamos que a probabilidade de defesa em 24 meses é de 0,31, que corresponde a 31% das chances totais. Percebam que a tabela não traz a informação de defesa em 23 meses, porém, a probabilidade de defesa em 22 meses é de 0,22, que corresponde a 22% das chances totais.

Assim, para defesa em até 2 meses antes do prazo máximo, a probabilidade é de 22%.

Agora, vejam na tabela que as probabilidades de defesa nos meses 25 e 26 são respectivamente 0,18 e 0,04, ou seja, 18% e 4%. Ao somarmos esses dois meses, também temos a probabilidade de 22%.

Logo, para a defesa em até dois meses após o prazo máximo, a probabilidade também é de 22%.

Portanto, a probabilidade de um aluno defender sua dissertação em até 2 meses antes do prazo máximo, 24 meses, é igual à probabilidade de um aluno defendê-la em até 2 meses depois.

Gabarito: Certo.

7. (CEBRASPE/FUB/2022) Uma universidade está fazendo um estudo para verificar a distribuição dos tempos que os alunos do curso de mestrado levam até a defesa da dissertação. Os dados a seguir mostram a função de probabilidade desses tempos, em meses.

Tempo de Defesa (meses)	Probabilidade
12	0,01
15	0,02
18	0,04
20	0,10
22	0,22
24	0,31
25	0,18
26	0,04
28	0,03
30	0,05

Considerando essas informações, julgue o item subsequente.

Os valores da probabilidade de um aluno defender a dissertação em 13, 14, 16, 19, 21, 23, 27 ou 29 meses, somados, é igual à probabilidade de um aluno defender a dissertação em exatamente 31 meses.

Comentários:

A partir dos dados apresentados na tabela, podemos perceber que a probabilidade de um aluno defender a dissertação em 13, 14, 16, 19, 21, 23, 27 ou 29 meses, somados, é igual a zero, posto que não foram discriminadas as frequências relativas para os referidos meses. Igualmente, não foi informada a frequência com que os alunos defendem a dissertação com 31 meses, sendo a probabilidade, portanto, igual a zero.

Gabarito: Certo.

8. (CEBRASPE/FUB/2022) Uma universidade está fazendo um estudo para verificar a distribuição dos tempos que os alunos do curso de mestrado levam até a defesa da dissertação. Os dados a seguir mostram a função de probabilidade desses tempos, em meses.

Tempo de Defesa (meses)	Probabilidade
12	0,01

15	0,02
18	0,04
20	0,10
22	0,22
24	0,31
25	0,18
26	0,04
28	0,03
30	0,05

Considerando essas informações, julgue o item subsequente.

Se o prazo máximo de defesa recomendado é de 24 meses, então a probabilidade de um aluno defender sua dissertação no prazo é superior a 70%.

Comentários:

Com base nos dados da tabela, a probabilidade de um aluno defender a dissertação em um prazo superior a 24 meses é igual a:

$$0{,}18 + 0{,}04 + 0{,}03 + 0{,}05 = 0{,}30$$

Portanto, 30% dos alunos defendem sua dissertação em prazo superior a 24 meses. Assim, a probabilidade de um aluno defender sua dissertação no prazo é de até 70%, não mais que isso.

Gabarito: Errado.

9. (CEBRASPE/FUB/2022) A tabela de frequência a seguir mostra dados coletados em uma pesquisa para se verificar o número de disciplinas que os estudantes de determinada universidade estão cursando por semestre.

Disciplinas	2	3	4	5	6	7	8
Estudantes	10	15	40	35	28	10	4

Considerando essas informações, julgue o item seguinte.

Na pesquisa foram entrevistados 142 alunos.

Comentários:

Para identificarmos o número de entrevistados, basta somarmos as quantidades de estudantes inscritos em cada disciplina:

$$10 + 15 + 40 + 35 + 28 + 10 + 4 = 142$$

Gabarito: Certo.

10. (CEBRASPE/Polícia Federal/2018)

	DIA				
	1	2	3	4	5
X (quantidade diária de drogas apreendidas, em kg)	10	22	18	22	28

Tendo em vista que, diariamente, a Polícia Federal apreende uma quantidade X, em kg, de drogas em determinado aeroporto do Brasil, e considerando os dados hipotéticos da tabela precedente, que apresenta os valores observados da variável X em uma amostra aleatória de 5 dias de apreensões no citado aeroporto, julgue o item.

A tabela em questão descreve a distribuição de frequências da quantidade de drogas apreendidas nos cinco dias que constituem a amostra.

Comentários:

A tabela apenas registra as quantidades de drogas apreendidas por dia. Para que ela configurasse uma distribuição de frequências da quantidade de drogas apreendidas nos cinco dias, era necessário que cada valor estivesse associado a uma frequência, o que não ocorreu. Uma distribuição de frequência dos dados acima seria assim:

Valor	Frequência
10	1
18	1
22	2
28	1

Gabarito: Errado.

11. (CEBRASPE/IPHAN/2018). A tabela a seguir mostra as quantidades de bibliotecas públicas presentes em 20 microrregiões brasileiras.

90	**66**	**78**	**82**
77	**60**	**64**	**90**
87	**85**	**67**	**91**
82	**70**	**81**	**80**
69	**78**	**90**	**67**

A partir desses dados, pretende-se construir um gráfico de distribuição de frequências com quatro classes de igual amplitude. Os valores mínimo e máximo de cada classe devem ser números inteiros.

Considerando essas informações, julgue o item subsequente, relativo ao gráfico de distribuição a ser apresentado.

A amplitude de cada classe deverá ser superior a 6.

Comentários:

A questão forneceu um conjunto de dados e disse que se pretende construir um gráfico de distribuição de frequências com quatro classes de igual amplitude. Para encontrar a amplitude de cada classe, podemos utilizar a fórmula da amplitude total:

$$AT = h \times k \Longrightarrow h = \frac{AT}{k}$$

Analisando a tabela, identificamos os limites mínimo e máximo:

$$l_{máx} = 91$$
$$l_{mín} = 60$$

Calculando a amplitude total, temos:

$$AT = l_{máx} - l_{mín}$$

$$AT = 91 - 60$$

$$AT = 31$$

Aplicando a fórmula anterior, temos:

$$h = \frac{AT}{k} = \frac{31}{4} = 7{,}75$$

O número inteiro mais próximo deste resultado é 8 (superior a 6).

Logo, o item está correto.

Gabarito: Certo.

12. (CEBRASPE/IPHAN/2018) A tabela a seguir mostra as quantidades de bibliotecas públicas presentes em 20 microrregiões brasileiras.

90	66	78	82
77	60	64	90
87	85	67	91
82	70	81	80
69	78	90	67

A partir desses dados, pretende-se construir um gráfico de distribuição de frequências com quatro classes de igual amplitude. Os valores mínimo e máximo de cada classe devem ser números inteiros. Considerando essas informações, julgue o item subsequente, relativo ao gráfico de distribuição a ser apresentado.

A última classe deverá variar de 84 a 91.

Comentários:

A questão não informou qual o limite inferior do primeiro intervalo, portanto, poderíamos ter adotado qualquer valor. Isso alteraria totalmente o resultado esperado pelo avaliador. Dessa forma, considero que a questão deveria ter sido anulada. De todo modo, vamos fazer a questão como a banca esperava que ela fosse resolvida.

Para encontrar a amplitude de cada classe, podemos utilizar a fórmula da amplitude total:

$$AT = h \times k \Longrightarrow h = \frac{AT}{k}$$

Analisando a tabela, identificamos os limites mínimo e máximo:

$$l_{máx} = 91$$
$$l_{mín} = 60$$

Calculando a amplitude total, temos:

$$AT = l_{máx} - l_{mín}$$
$$AT = 91 - 60$$
$$AT = 31$$

Aplicando a fórmula anterior, temos:

$$h = \frac{AT}{k} = \frac{31}{4} = 7{,}75$$

Para o número inteiro mais próximo deste resultado pode ser 8 ou 7. Vamos iniciar calculando para o menor valor, 7:

$$60 \vdash\dashv 67$$
$$68 \vdash\dashv 75$$
$$76 \vdash\dashv 83$$
$$84 \vdash\dashv 91$$

Assim, concluímos que o examinador optou por esse valor para o cálculo. Novamente, a questão não explicitou que o valor mínimo deveria ser 60, logo, poderíamos ter adotado qualquer valor, o que alteraria esse resultado.

Gabarito: Certo.

13. (CEBRASPE/IPHAN/2018). Julgue o item subsequente, referente à análise exploratória de dados.

O histograma é um diagrama de retângulos contíguos com base na curtose das faixas de valores da variável e com área igual à diferença da frequência absoluta da respectiva faixa.

Comentários:

Um histograma é uma representação gráfica (um gráfico de barras verticais ou barras horizontais) da distribuição de frequências de um conjunto de dados quantitativos contínuos. Ele pode ser construído com base em valores absolutos ou frequência relativa ou densidade.

Em se tratando de densidade, a frequência relativa do intervalo *i*, (f_{ri}), é representada pela área de um retângulo, colocado acima do ponto médio da classe *i*. Consequentemente, a área total do histograma (igual a soma das áreas de todos os retângulos) será igual a 1.

Assim, ao construir o histograma, cada retângulo deverá ter área proporcional à frequência relativa (ou à frequência absoluta) correspondente. No caso em que os intervalos são de tamanhos iguais, as alturas dos retângulos serão iguais às frequências relativas (ou iguais às frequências absolutas) dos intervalos correspondentes.

Desse modo, a área não tem relação com a diferença de frequência absoluta de um certo intervalo. A área de cada retângulo é proporcional à frequência da classe, sem qualquer operação de diferença envolvida.

Outro ponto relevante é que o histograma não tem relação com a curtose dos intervalos. A curtose é uma medida que indica o grau de achatamento de uma distribuição de frequências.

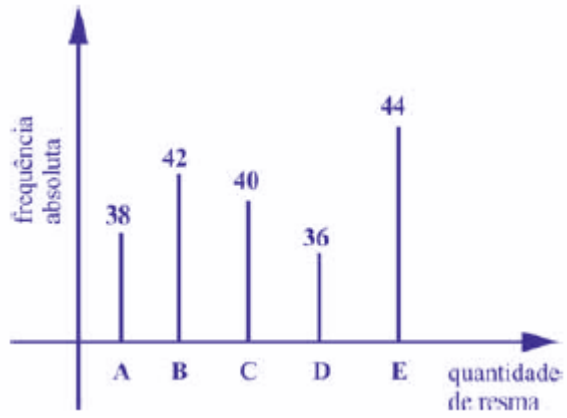
Gabarito: Errado.

14. (CEBRASPE/CBM-AL/2017) Na tabela a seguir, A, B, C, D e E são as quantidades de resmas de papel A4 consumidas, em quatro meses, pelas seções administrativas I, II, III, IV e V, respectivamente. Apesar de não mostrar explicitamente essas quantidades, a tabela apresenta as frequências absolutas e (ou) relativas de algumas dessas quantidades.

Seção	Quantidades de Resmas	Frequência Absoluta	Frequência Relativa
I	A	38	19%
II	B		
III	C		20%
IV	D	36	
V	E	44	
	Total		**100%**

Considerando que cada uma dessas resmas, juntamente com a embalagem, tem forma de um paralelepípedo retângulo reto que mede 5 cm × 21 cm × 30 cm, julgue o item seguinte.

O gráfico de barras verticais a seguir apresenta as frequências absolutas de resmas consumidas pelas cinco seções.



Comentários:

Podemos usar uma regra de três simples para completar a tabela. Primeiro, vamos encontrar o valor que representa 100%:

$$38 - 0{,}19$$

$$Q - 1{,}00$$

$$Q = \frac{38 \times 1{,}00}{0{,}19} = 200$$

Encontrando o valor de 20%:

$$200 - 1{,}00$$

$$C - 0{,}20$$

$$C = \frac{200 \times 0{,}2}{1{,}00} = 40$$

Agora, basta sabermos a frequência absoluta na seção II. Como foram consumidas 200 resmas, então o número de resmas consumidas pela seção II foi:

$$200 - 38 - 40 - 36 - 44 = 42$$

Dessa forma, nas seções I, II, III, IV e V foram consumidas, respectivamente 38, 42, 40, 36 e 44 resmas de papel A4, conforme mostra o gráfico.

Gabarito: Certo.

15. (CEBRASPE/TCE-PA/2016)

Número diário de denúncias registradas (X)	Frequência Relativa
0	0,3
1	0,1
2	0,2
3	0,1
4	0,3
Total	**1,0**

A tabela precedente apresenta a distribuição de frequências relativas da variável X, que representa o número diário de denúncias registradas na ouvidoria de determinada instituição pública. A partir das informações dessa tabela, julgue o item seguinte.

A variável X é do tipo qualitativo nominal.

Comentários:

De acordo com a tabela X corresponde à quantidade diária de denúncias. Logo, X é uma variável quantitativa discreta, pois assume valores inteiros.

Gabarito: Errado.

16. (CEBRASPE/DEPEN/2015)

Região	Quantidade de detentos no sistema penitenciário brasileiro (mil pessoas)	Déficit de vagas no sistema penitenciário (mil vagas)	População brasileira (milhões de habitantes)
Norte	37	13	17
Centro-Oeste	51	24	15
Nordeste	94	42	55
Sudeste	306	120	85
Sul	67	16	28
Total	555	215	200

Ministério da Justiça — Departamento Penitenciário Nacional — Sistema Integrado de Informações Penitenciárias – InfoPen, Relatório Estatístico Sintético do Sistema Prisional Brasileiro, dez./2013

Internet: <www.justica.gov.br> (com adaptações).

A tabela mostrada apresenta a quantidade de detentos no sistema penitenciário brasileiro por região em 2013. Nesse ano, o déficit relativo de vagas — que se define pela razão entre o déficit de vagas no sistema penitenciário e a quantidade de detentos no sistema penitenciário — registrado em todo o Brasil foi superior a 38,7%, e, na média nacional, havia 277,5 detentos por 100 mil habitantes.

Com base nessas informações e na tabela apresentada, julgue o item a seguir.

Em 2013, mais de 55% da população carcerária no Brasil se encontrava na região Sudeste.

Comentários:

Vamos analisar os dados da região sudeste. Sabemos que o total da população carcerária é 555 mil. Se a região sudeste contém 306 mil, basta calcularmos o percentual desse valor com relação ao total. Assim:

$$X = \frac{306}{555} \times 100\% = 55{,}13\%$$

Gabarito: Certo.

17. (CEBRASPE/DEPEN/2015)

Região	Quantidade de detentos no sistema penitenciário brasileiro (mil pessoas)	Déficit de vagas no sistema penitenciário (mil vagas)	População brasileira (milhões de habitantes)
Norte	37	13	17
Centro-Oeste	51	24	15
Nordeste	94	42	55
Sudeste	306	120	85
Sul	67	16	28
Total	555	215	200

Ministério da Justiça — Departamento Penitenciário Nacional — Sistema Integrado de Informações Penitenciárias – InfoPen, Relatório Estatístico Sintético do Sistema Prisional Brasileiro, dez./2013

Internet: <www.justica.gov.br> (com adaptações).

A tabela mostrada apresenta a quantidade de detentos no sistema penitenciário brasileiro por região em 2013. Nesse ano, o déficit relativo de vagas — que se define pela razão entre o déficit de vagas no sistema penitenciário e a quantidade de detentos no sistema penitenciário — registrado em todo o Brasil foi superior a 38,7%, e, na média nacional, havia 277,5 detentos por 100 mil habitantes.

Com base nessas informações e na tabela apresentada, julgue o item a seguir.

A quantidade total de vagas existentes no sistema penitenciário brasileiro em 2013 era de 340 mil vagas.

Comentários:

Analisando a tabela, temos a informação do total de 555 mil detentos, e a informação do déficit de vagas no sistema penitenciário 215 mil.

Ora, para sabermos a quantidade real de vagas, basta subtrairmos o déficit do total.

$$555 - 215 = 340$$

Gabarito: Certo.

18. (CEBRASPE/DEPEN/2015)

Região	Quantidade de detentos no sistema penitenciário brasileiro (mil pessoas)	Déficit de vagas no sistema penitenciário (mil vagas)	População brasileira (milhões de habitantes)
Norte	37	13	17
Centro-Oeste	51	24	15
Nordeste	94	42	55
Sudeste	306	120	85
Sul	67	16	28
Total	555	215	200

Ministério da Justiça — Departamento Penitenciário Nacional — Sistema Integrado de Informações Penitenciárias – InfoPen, Relatório Estatístico Sintético do Sistema Prisional Brasileiro, dez./2013

Internet: <www.justica.gov.br> (com adaptações).

A tabela mostrada apresenta a quantidade de detentos no sistema penitenciário brasileiro por região em 2013. Nesse ano, o déficit relativo de vagas — que se define pela razão entre o déficit de vagas no sistema penitenciário e a quantidade de detentos no sistema penitenciário — registrado em todo o Brasil foi superior a 38,7%, e, na média nacional, havia 277,5 detentos por 100 mil habitantes.

Com base nessas informações e na tabela apresentada, julgue o item a seguir.

O déficit relativo de vagas observado na região Sudeste, em 2013, foi superior ao déficit relativo de vagas registrado na região Centro-oeste no mesmo período.

Comentários:

Precisamos calcular o percentual do déficit das regiões sudeste e centro-oeste. Para isso, basta dividirmos o déficit de vagas da região pela respectiva quantidade de detentos dessa região. Assim:

$$Sudeste = \frac{120}{306} = 39{,}21\%$$

$$Centro - Oeste = \frac{24}{51} = 47{,}05\%$$

Portanto, o déficit é maior na região Centro-Oeste.

Gabarito: Errado.

19. (CEBRASPE/DEPEN/2015)

Região	Quantidade de detentos no sistema penitenciário brasileiro (mil pessoas)	Déficit de vagas no sistema penitenciário (mil vagas)	População brasileira (milhões de habitantes)
Norte	37	13	17
Centro-Oeste	51	24	15
Nordeste	94	42	55
Sudeste	306	120	85
Sul	67	16	28
Total	555	215	200

Ministério da Justiça — Departamento Penitenciário Nacional — Sistema Integrado de Informações Penitenciárias – InfoPen, Relatório Estatístico Sintético do Sistema Prisional Brasileiro, dez./2013

Internet: <www.justica.gov.br> (com adaptações).

A tabela mostrada apresenta a quantidade de detentos no sistema penitenciário brasileiro por região em 2013. Nesse ano, o déficit relativo de vagas — que se define pela razão entre o déficit de vagas no sistema penitenciário e a quantidade de detentos no sistema penitenciário — registrado em todo o Brasil foi superior a 38,7%, e, na média nacional, havia 277,5 detentos por 100 mil habitantes.

Com base nessas informações e na tabela apresentada, julgue o item a seguir.

No ano considerado, a quantidade média de detentos por 100 mil habitantes na região Nordeste foi superior ao número médio de detentos por 100 mil habitantes na região Centro-oeste.

Comentários:

Precisamos calcular a razão entre a quantidade de detentos e a população das respectivas regiões. Assim, para as regiões Nordeste e Centro-Oeste, temos que:

$$Nordeste = \frac{94 \times 10^3}{55 \times 10^6} = 1{,}7 \times 10^{-3}$$

$$Centro - Oeste = \frac{51 \times 10^3}{15 \times 10^6} = 3{,}4 \times 10^{-3}$$

Portanto, a quantidade é maior na região Centro-Oeste.

Gabarito: Errado.

20. (CESPE/BACEN/2013)

2 4 8 4 8 1 2 32 12 1 5 7 5 5 3 4 24 19 4 14

Os dados mostrados acima representam uma amostra, em minutos, do tempo utilizado na armazenagem de formulários no almoxarifado central de certa instituição por diversos funcionários.

Com base nesses dados, julgue o próximo item.

A distribuição de frequência acumulada para tempo de armazenagem observado na amostra inferior a 8 minutos é igual a 13, o que corresponde a uma frequência relativa superior a 0,60.

Comentários:

Não precisamos construir a distribuição de frequências para responder à questão. Sabemos que o total de observações é 20, pois é esse o número de termos.

Vamos destacar as observações que são inferiores a 8:

2 4 8 **4** 8 **1 2** 32 12 **1 5 7 5 5 3 4** 24 19 **4** 14

São 13 números inferiores a 8. Portanto, a frequência acumulada para um tempo menor do que 8 é 13.

Para calcular a frequência relativa, basta dividir a frequência absoluta pelo total de observações:

$$\frac{13}{20} = 0{,}65$$

Logo, a frequência relativa acumulada é maior do que 0,60.

Gabarito: Certo.

21. (CESPE/ALECE/2011)

X	0	1	2	3	4	5
Frequência Absoluta	80	47	30	20	6	1

Um levantamento foi realizado para se avaliar, por município, a quantidade X de obras que estão sob suspeita de irregularidade. Com base em uma amostra de municípios, foi obtida a distribuição de frequências mostrada na tabela acima. Com base nessas informações, julgue o item a seguir.

Os extremos mínimo e máximo da variável X foram, respectivamente, iguais a 1 e 80.

Comentários:

Na tabela temos que a variável X assume valores que vão de 0 a 5. As frequências absolutas correspondem a quantas vezes esses valores se repetem. Portanto, concluímos que os valores mínimo e máximo de X são, respectivamente, 0 e 5.

Gabarito: Errado

22. (CESPE/ALECE/2011)

X	0	1	2	3	4	5
Frequência Absoluta	80	47	30	20	6	1

Um levantamento foi realizado para se avaliar, por município, a quantidade X de obras que estão sob suspeita de irregularidade. Com base em uma amostra de municípios, foi obtida a distribuição de frequências mostrada na tabela acima. Com base nessas informações, julgue o item a seguir.

O número de municípios que têm obras sob suspeita de irregularidades é superior a 120.

Comentários:

Para respondermos a questão, basta entendermos que precisamos fazer um somatório de todas as obras que estão sob suspeita de irregularidades. Se $X = 0$ se refere ao fato de nenhuma obra estar sob suspeita de irregularidade, então partiremos da soma $X = 1$ a $X = 5$. Assim, somaremos todas as frequências absolutas:

$$47 + 30 + 20 + 6 + 1 = 104$$

Gabarito: Errado.

23. (CESPE/ALECE/2011)

X	0	1	2	3	4	5
Frequência Absoluta	80	47	30	20	6	1

Um levantamento foi realizado para se avaliar, por município, a quantidade X de obras que estão sob suspeita de irregularidade. Com base em uma amostra de municípios, foi obtida a distribuição de frequências mostrada na tabela acima. Com base nessas informações, julgue o item a seguir.

O total de municípios considerado no levantamento foi superior a 180.

Comentários:

Para sabermos o total de municípios, basta somarmos todas as frequências absolutas:

$$80 + 47 + 30 + 20 + 6 + 1 = 184$$

Gabarito: Certo.

24. (CESPE/SEFAZ-MT/2004) Considere a seguinte situação hipotética.

Um órgão do governo recebeu pela Internet denúncias de sonegação de impostos estaduais contra 600 pequenas empresas. Denúncias contra outras 200 pequenas empresas foram encaminhadas pessoalmente para esse órgão. Para a apuração das denúncias, foram realizadas auditorias nas 800 empresas denunciadas. Como resultado dessas auditorias, foi elaborada a tabela abaixo, que apresenta um quadro das empresas denunciadas e os correspondentes débitos fiscais ao governo. Das empresas denunciadas, observou-se que apenas 430 tinham débitos fiscais.

	Valor do débito fiscal (VDF), em R$ mil, apurado após auditoria na empresa denunciada				
Forma de recebimento da denúncia	$0 < VDF < 1$	$1 \leq VDF < 2$	$2 \leq VDF < 3$	$3 \leq VDF \leq 4$	Total
Pela internet	60	100	50	30	240
Pessoalmente	20	120	40	10	190
Total	80	220	90	40	430*

Nota: *Para as demais empresas, VDF=0

Com base na situação hipotética acima e de acordo com as informações apresentadas, julgue o item que se segue.

O valor total dos débitos fiscais apurados após as auditorias feitas nas empresas denunciadas é inferior a R$ 500 mil.

Comentários:

Para responder essa questão, trabalharemos com os limites inferiores das classes, buscando minimizar os valores dos débitos, para ver se é possível que o total seja menor de R$ 500 mil.

Débito Mínimo	Frequência	Produto
0	80	0

1	220	220
2	90	180
3	40	120
Total		**520**

Portanto, o débito total é de, no mínimo, R$ 520 mil. Logo, é impossível que o valor total dos débitos fiscais seja inferior a R$ 500 mil.

Gabarito: Errado.

25. (CESPE/SEFAZ-AL/2002) Julgue o seguinte item.

Em uma distribuição de frequências para um conjunto de n indivíduos, pode-se calcular as frequências relativas, dividindo-se cada frequência absoluta pela amplitude da correspondente classe ou do intervalo.

Comentários:

A frequência relativa é calculada pela divisão entre a frequência absoluta simples e a frequência total. Na realidade, a questão fez referência ao cálculo da densidade de frequência.

Gabarito: Errado.

QUESTÕES COMENTADAS – CEBRASPE

Representação Gráfica das Distribuições de Frequências

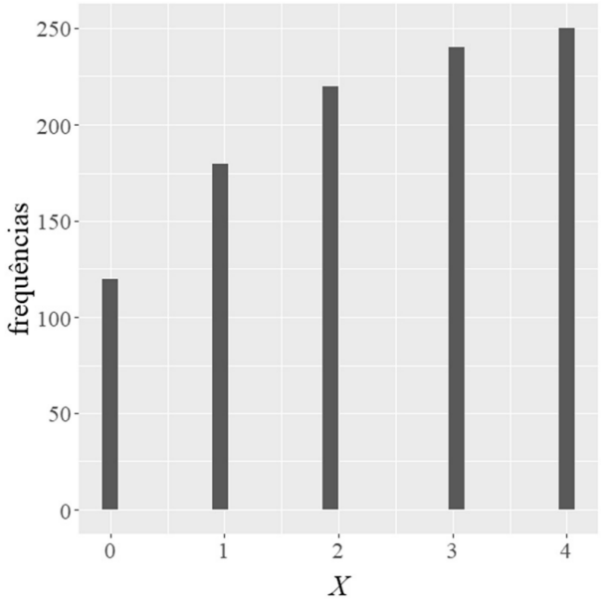
1. (CESPE/CNJ/2024)

X	frequência absoluta acumulada
0	120
1	180
2	220
3	240
4	250

A tabela precedente mostra a distribuição de frequências do número diário (X) de denúncias recebidas pela ouvidoria de um tribunal de justiça.

Com base nos dados apresentados na tabela, julgue os itens que se seguem.

O histograma a seguir representa corretamente a distribuição de frequências da variável X.



Comentários:

Primeiramente, é necessário converter a frequência acumulada em frequência simples para cada valor de X. Calculando a frequência simples a partir da frequência acumulada, temos:

$$Frequência\ para\ X = 0: 120$$

$$Frequência\ para\ X = 1: 180 - 120 = 60$$

$$Frequência\ para\ X = 2: 220 - 180 = 40$$

$$Frequência\ para\ X = 3: 240 - 220 = 20$$

$$Frequência\ para\ X = 4: 250 - 240 = 10$$

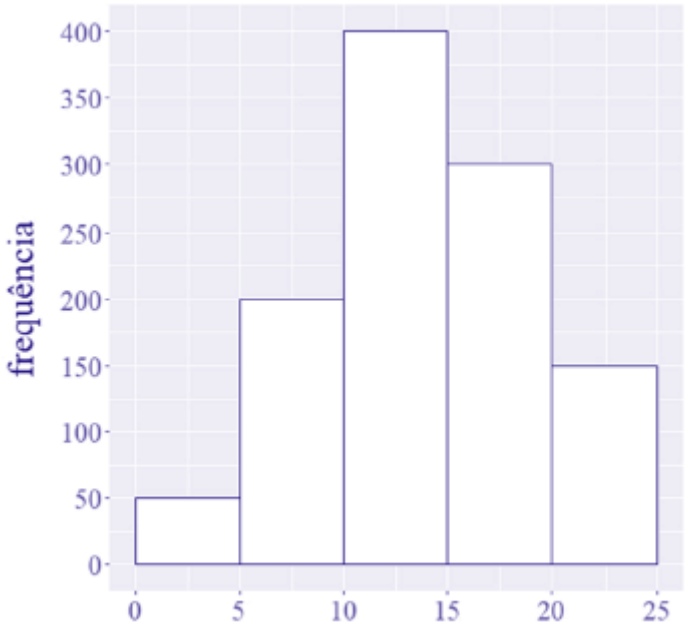
Assim, a tabela de frequência simples fica:

X	Frequência Simples
0	120
1	60
2	40
3	20
4	10

O histograma mostrado tem barras com frequências aproximadamente iguais, com a frequência máxima em 250 e a mínima em 200, o que não corresponde às frequências reais. Portanto, o histograma **não representa corretamente** a distribuição de frequências da variável X.

Gabarito: Errado.

2. (CESPE/TELEBRAS/2022)



Considerando que o histograma apresentado descreve a distribuição de uma variável quantitativa X por meio de frequências absolutas, julgue o item que se segue.

O número de observações que constituem a variável X é igual a 1.000.

Comentários:

O número de observações consiste na soma de todas as frequências representadas no histograma. Assim, com base no histograma, temos que a soma das frequências é igual a:

$$400 + 300 + 200 + 150 + 50 = 1100$$

Gabarito: Errado.

3. (CESPE/SEDUC AL/2021) Com base em estatística, julgue o item a seguir.

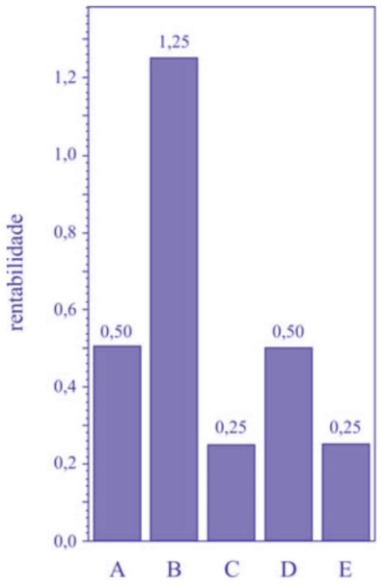
Suponha que o histograma a seguir represente a frequência relativa de alunos, distribuída por faixa etária, que ingressaram no ensino superior no estado de Alagoas em 2020. Com base nas informações desse gráfico, é correto afirmar que mais de 50% dos novos alunos têm idade superior a 22 anos.

Comentários:

Observando o histograma, percebemos que as duas primeiras classes correspondem a mais de 50% dos dados, pois possuem frequências superiores a 30% e 20%, respectivamente. Como as idades nesse intervalo variam de 18 a 22 anos, menos de 50% terão idades superiores a 22 anos.

Gabarito: Errado.

4. (CESPE/FUNPRESP/2016)



O gráfico ilustra cinco possibilidades de fundos de investimento com suas respectivas rentabilidades. Considerando que as probabilidades de investimento para os fundos A, B, C e D sejam, respectivamente, P(A) = 0,182; P(B) = 0,454; P(C)= 0,091; e P(D)= 0,182, julgue o item subsequente.

O gráfico apresentado é um histograma.

Comentários:

O gráfico não é um histograma por dois motivos: primeiro porque há uma separação entre as colunas, o que não ocorre em um histograma, e sim em um gráfico de colunas; segundo porque um histograma representa dados que estão agrupados em intervalos de classe, e não em categorias, como é o caso.

Gabarito: Errado.

5. (CESPE/DEPEN/2015)

Região	Quantidade de detentos no sistema penitenciário brasileiro (mil pessoas)	Déficit de vagas no sistema penitenciário (mil vagas)	População brasileira (milhões de habitantes)
Norte	37	13	17
Centro-Oeste	51	24	15
Nordeste	94	42	55
Sudeste	306	120	85
Sul	67	16	28
Total	555	215	200

Ministério da Justiça — Departamento Penitenciário Nacional — Sistema Integrado de Informações Penitenciárias – InfoPen, Relatório Estatístico Sintético do Sistema Prisional Brasileiro, dez./2013

Internet: <www.justica.gov.br> (com adaptações).

A tabela mostrada apresenta a quantidade de detentos no sistema penitenciário brasileiro por região em 2013. Nesse ano, o déficit relativo de vagas — que se define pela razão entre o déficit de vagas no sistema penitenciário e a quantidade de detentos no sistema penitenciário — registrado em todo o Brasil foi superior a 38,7%, e, na média nacional, havia 277,5 detentos por 100 mil habitantes.

Com base nessas informações e na tabela apresentada, julgue o item a seguir.

Na análise exploratória, o histograma é um gráfico adequado para descrever a distribuição da quantidade de detentos por região em 2013.

Comentários:

O histograma é destinado a representar dados agrupados em classes, mas o enunciado apresenta dados com valores individualizados. Como a informação não está agrupada em classes, o histograma não representaria adequadamente os dados da tabela.

Gabarito: Errado.

6. (CESPE/BACEN/2013)

2 4 8 4 8 1 2 32 12 1 5 7 5 5 3 4 24 19 4 14

Os dados mostrados acima representam uma amostra, em minutos, do tempo utilizado na armazenagem de formulários no almoxarifado central de certa instituição por diversos funcionários.

Com base nesses dados, julgue o próximo item.

É inviável a elaboração de um histograma em decorrência do fato de ser este um conjunto de dados quantitativos discretos; dessa forma, apenas por meio de um gráfico de barras pode ser realizada a representação gráfica.

Comentários:

O item está errado. É sim possível elaborar um histograma para os dados apresentados, bastaria, para tanto, organizá-los em intervalos de classe.

Gabarito: Errado.

QUESTÕES COMENTADAS – CEBRASPE

Outros Gráficos e Representações

1. (CESPE/FUB/2022) Uma universidade está fazendo um estudo para verificar a distribuição dos tempos que os alunos do curso de mestrado levam até a defesa da dissertação. Os dados a seguir mostram a função de probabilidade desses tempos, em meses.

Tempo de Defesa (meses)	Probabilidade
12	0,01
15	0,02
18	0,04
20	0,10
22	0,22
24	0,31
25	0,18
26	0,04
28	0,03
30	0,05

Considerando essas informações, julgue o item subsequente.

O gráfico de setores é adequado para representar a distribuição em questão.

Comentários:

O gráfico em setores (também conhecido como gráfico de pizza) é usado para representar a frequência relativa (porcentagem) de uma variável categórica. Ele é formado por um círculo dividido em setores circulares, cada um representando uma categoria, cujos ângulos centrais são proporcionais às frequências relativas de cada categoria. Na tabela em questão, o tempo de defesa é representado de forma numérica, o que não é condizente com uma variável categórica. Portanto, o gráfico em setores não é adequado para representar a distribuição apresentada na tabela.

Gabarito: Errado.

2. (CESPE/FUB/2022) A tabela de frequência a seguir mostra dados coletados em uma pesquisa para se verificar o número de disciplinas que os estudantes de determinada universidade estão cursando por semestre.

Disciplinas	2	3	4	5	6	7	8
Estudantes	10	15	40	35	28	10	4

Considerando essas informações, julgue o item seguinte.

O gráfico do tipo pizza é o mais apropriado para representar os dados apresentados na tabela, visto que a variável analisada é qualitativa ordinal.

Comentários:

O gráfico em setores (também conhecido como gráfico de pizza) é usado para representar a frequência relativa (porcentagem) de uma variável categórica. Ele é formado por um círculo dividido em setores circulares, cada um representando uma categoria, cujos ângulos centrais são proporcionais às frequências relativas da categoria.

As variáveis qualitativas são as características que não podem ser descritas de forma numérica, mas que podem ser definidas por meio de qualidades (atributos ou categorias) do indivíduo pesquisado. Elas podem ser classificadas em nominais ou ordinais:

a) variável qualitativa nominal (ou categórica), as possíveis categorias não podem ser ordenadas. Por exemplo, a cor dos olhos dos moradores de uma determinada cidade (pretos, castanhos, azuis e verdes);

b) variável qualitativa ordinal, as possíveis categorias podem ser ordenadas de alguma forma. Por exemplo, o grau de instrução dos funcionários de um determinado órgão (fundamental, médio, superior).

Portanto, analisando os dados apresentados na tabela, verificamos tratar-se de uma variável quantitativa, a qual podemos atribuir valores numéricos, e não uma variável qualitativa. Assim, a afirmativa da questão está incorreta.

Gabarito: Errado.

3. (CESPE/IPHAN/2018) Julgue o item subsequente, referente à análise exploratória de dados.

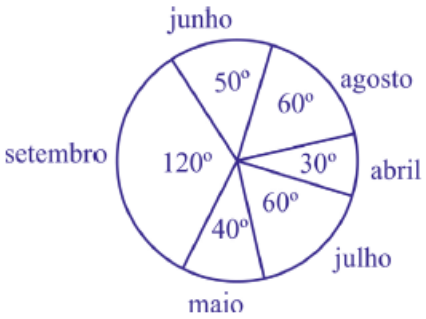
A representação de diagramas de barras, de linha e de pizza possui escala de medida nominal e tem a moda como medida de tendência central.

Comentários:

Os diagramas de barras, de linha e de pizza são geralmente usados para representar dados nominais ou categóricos. Como veremos na aula de medidas de posição, a única medida de tendência central que podemos utilizar no caso de dados categóricos é a moda, pois ela apenas registra a categoria que mais se repete em uma amostra, sem envolver outras operações matemáticas.

Gabarito: Certo.

4. (CESPE/CBM-AL/2017) O gráfico de setores a seguir mostra a distribuição das quantidades de incêndios em determinada região, nos meses de abril a setembro de determinado ano.



Sabendo-se que nesses meses ocorreram 1.548 incêndios nessa região, julgue o item que se segue.

A frequência relativa à classe “incêndios no mês de setembro” é superior a 30%.

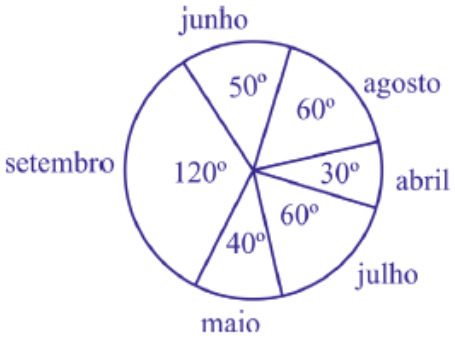
Comentários:

Para resolvermos essa questão, basta aplicarmos uma regra de três simples. Se uma circunferência total tem 360° e corresponde a 100%, então o ângulo de 120° corresponderá a:

$$Set. = \frac{120}{360} = \frac{1}{3} \approx 33{,}33\%$$

Gabarito: Certo.

5. (CESPE/CBM-AL/2017) O gráfico de setores a seguir mostra a distribuição das quantidades de incêndios em determinada região, nos meses de abril a setembro de determinado ano.



Sabendo-se que nesses meses ocorreram 1.548 incêndios nessa região, julgue o item que se segue.

Nos meses de maio e junho ocorreram mais de 400 incêndios nessa região.

Comentários:

Para resolver a questão, devemos aplicar uma regra de três simples. Isto é, se a circunferência total possui 360° e corresponde a 100%, então 360° também equivale ao número total de incêndios, 1548.

Podemos verificar também que os meses de maio e junho somam um total de 90°. Calculando 90°, temos:

$$360 - 1548$$

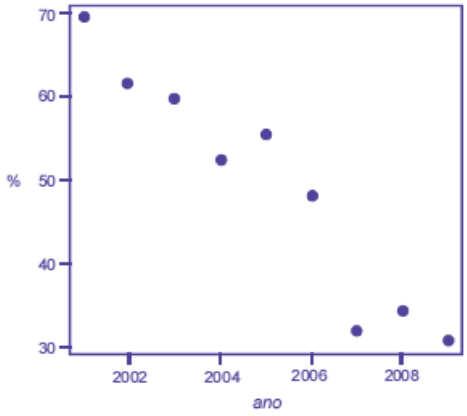
$$90 - X$$

$$X = \frac{139.320}{360} = 387$$

Logo, a quantidade de incêndios durante os meses de maio e junho não foi superior a 400, como afirma o item.

Gabarito: Errado.

6. (CESPE/DEPEN/2015)

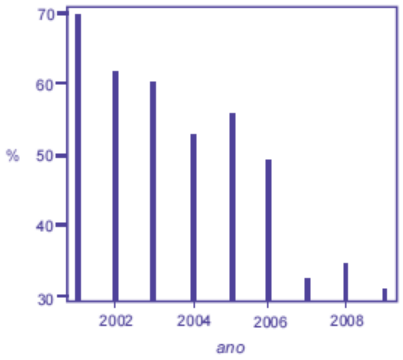


Dado que a participação dos presidiários em cursos de qualificação profissional é um aspecto importante para a reintegração do egresso do sistema prisional à sociedade, foram realizados levantamentos

estatísticos, nos anos de 2001 a 2009, a respeito do valor da educação e do trabalho em ambientes prisionais. Cada um desses levantamentos, cujos resultados são apresentados no gráfico, produziu uma estimativa anual do percentual P de indivíduos que participaram de um curso de qualificação profissional de curta duração, mas que não receberam o diploma por motivos diversos. Em 2001, 69,4% dos presidiários que participaram de um curso de qualificação profissional não receberam o diploma. No ano seguinte, 2002, esse percentual foi reduzido para 61,5%, caindo, em 2009, para 30,9%.

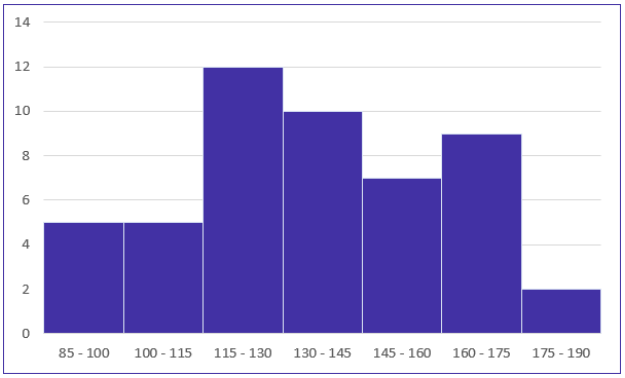
A partir das informações e do gráfico apresentados, julgue o item que se segue.

Se os percentuais forem representados por barras verticais, conforme o gráfico a seguir, então o resultado será denominado histograma.



Comentários:

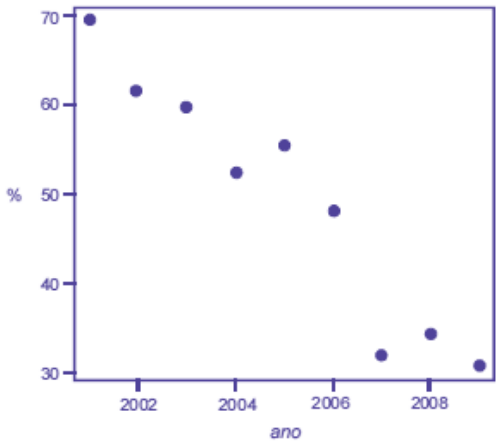
O histograma é destinado a representar dados contínuos agrupados em classes (em intervalos), como o mostrado na figura a seguir.



Em realidade, a assertiva apresentou um gráfico de colunas (barras verticais) que representa dados agrupados por valores.

Gabarito: Errado.

7. (CESPE/DEPEN/2015)



Dado que a participação dos presidiários em cursos de qualificação profissional é um aspecto importante para a reintegração do egresso do sistema prisional à sociedade, foram realizados levantamentos estatísticos, nos anos de 2001 a 2009, a respeito do valor da educação e do trabalho em ambientes prisionais. Cada um desses levantamentos, cujos resultados são apresentados no gráfico, produziu uma estimativa anual do percentual P de indivíduos que participaram de um curso de qualificação profissional de curta duração, mas que não receberam o diploma por motivos diversos. Em 2001, 69,4% dos presidiários que participaram de um curso de qualificação profissional não receberam o diploma. No ano seguinte, 2002, esse percentual foi reduzido para 61,5%, caindo, em 2009, para 30,9%.

A partir das informações e do gráfico apresentados, julgue o item que se segue.

Os dados apresentados são suficientes para que se possa afirmar que o total de presidiários que participaram de um curso de qualificação profissional de curta duração e que não receberam o diploma em 2008 foi superior ao total referente ao ano de 2007.

Comentários:

As informações presentes no gráfico foram dispostas em termos percentuais e não por meio de valores absolutos, portanto, não podemos afirmar que o número de presidiários que participaram do curso e não receberam diploma foi maior em um ano do que em outro. Apenas podemos afirmar que o percentual de presidiários que participaram dos cursos e não receberam o diploma foi maior em 2008 que o percentual em 2007.

Gabarito: Errado.

8. (CESPE/TRE-ES/2011)

Cargo	Candidatos	Candidatos aptos	Eleitos
Presidente da República	9	9	1

Governador de Estado	170	156	27
Senador	272	234	54
Deputado Federal	6.021	5.058	513
Deputado Estadual/Distrital	15.268	13.076	1.059
Total	21.640	18.533	1.658

Internet: <www.tse.gov > (com adaptações).

Com base na tabela acima, referente às eleições de 2010, que apresenta a quantidade de candidatos para os cargos de presidente da República, governador de estado, senador, deputado federal e deputado estadual/distrital, bem como a quantidade de candidatos considerados aptos pela justiça eleitoral e o total de eleitos para cada cargo pretendido, julgue o item a seguir.

Considerando-se a representação das quantidades de eleitos para cada cargo em um gráfico de pizza, a fatia desse gráfico correspondente ao cargo de deputado federal terá ângulo superior a 120°.

Comentários:

Inicialmente, precisamos descobrir o quanto o ângulo de 120° corresponde do valor total. Sabemos que 360° é o ângulo total, logo, 120° corresponde a 1/3 de 360°. Portanto, para afirmarmos que o cargo de deputado federal corresponde ao ângulo de 120°, ele deve necessariamente corresponder a 1/3 do total de candidatos eleitos. Vejamos:

$$\frac{1658}{3} \cong 552{,}66$$

Logo, para que o cargo de deputado federal tivesse um ângulo superior a 120°, deveríamos ter 553 candidatos eleitos. Não é o caso, pois, de acordo com a tabela, foram eleitos 513 candidatos.

Gabarito: Errado.

QUESTÕES COMENTADAS – CEBRASPE

Análise Exploratória de Dados

1. (CESPE/FUB/2022) Uma universidade está fazendo um estudo para verificar a distribuição dos tempos que os alunos do curso de mestrado levam até a defesa da dissertação. Os dados a seguir mostram a função de probabilidade desses tempos, em meses.

Tempo de Defesa (meses)	Probabilidade
12	0,01
15	0,02
18	0,04
20	0,10
22	0,22
24	0,31
25	0,18
26	0,04
28	0,03
30	0,05

Considerando essas informações, julgue o item subsequente.

O gráfico de setores é adequado para representar a distribuição em questão.

Comentários:

O gráfico em setores (também conhecido como gráfico de pizza) é usado para representar a frequência relativa (porcentagem) de uma variável categórica. Ele é formado por um círculo dividido em setores circulares, cada um representando uma categoria, cujos ângulos centrais são proporcionais às frequências relativas de cada categoria. Na tabela em questão, o tempo de defesa é representado de forma numérica, o que não é condizente com uma variável categórica. Portanto, o gráfico em setores não é adequado para representar a distribuição apresentada na tabela.

Gabarito: Errado.

2. (CESPE/FUB/2022) A tabela de frequência a seguir mostra dados coletados em uma pesquisa para se verificar o número de disciplinas que os estudantes de determinada universidade estão cursando por semestre.

Disciplinas	2	3	4	5	6	7	8
Estudantes	10	15	40	35	28	10	4

Considerando essas informações, julgue o item seguinte.

O gráfico do tipo pizza é o mais apropriado para representar os dados apresentados na tabela, visto que a variável analisada é qualitativa ordinal.

Comentários:

O gráfico em setores (também conhecido como gráfico de pizza) é usado para representar a frequência relativa (porcentagem) de uma variável categórica. Ele é formado por um círculo dividido em setores circulares, cada um representando uma categoria, cujos ângulos centrais são proporcionais às frequências relativas da categoria.

As variáveis qualitativas são as características que não podem ser descritas de forma numérica, mas que podem ser definidas por meio de qualidades (atributos ou categorias) do indivíduo pesquisado. Elas podem ser classificadas em nominais ou ordinais:

a) variável qualitativa nominal (ou categórica), as possíveis categorias não podem ser ordenadas. Por exemplo, a cor dos olhos dos moradores de uma determinada cidade (pretos, castanhos, azuis e verdes);

b) variável qualitativa ordinal, as possíveis categorias podem ser ordenadas de alguma forma. Por exemplo, o grau de instrução dos funcionários de um determinado órgão (fundamental, médio, superior).

Portanto, analisando os dados apresentados na tabela, verificamos tratar-se de uma variável quantitativa, a qual podemos atribuir valores numéricos, e não uma variável qualitativa. Assim, a afirmativa da questão está incorreta.

Gabarito: Errado.

3. (CESPE/DPE-RO/2022) O valor de um atributo de um dado objeto é uma medida da quantidade daquele atributo, a qual pode ser numérica ou categórica.

Nesse caso, estado civil e sexo são classificados como atributo

a) binário.

b) nominal.

c) ordinal.

d) ausente.

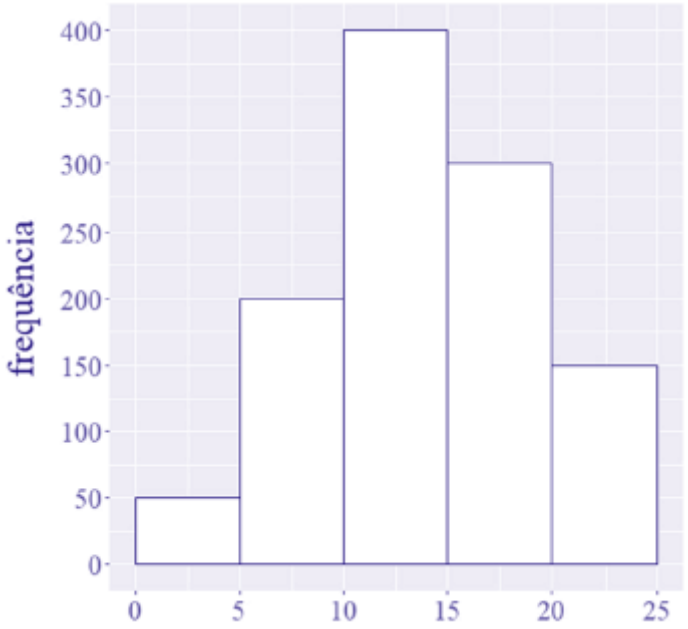
e) razão.

Comentários:

As variáveis qualitativas representam as características que não podem ser descritas de forma numérica, mas que podem ser definidas por meio de qualidades (atributos ou categorias) do indivíduo pesquisado. No caso, as variáveis "estado civil" e "sexo" são qualitativas nominais (ou categóricas), pois as categorias não podem ser ordenadas nem possuem significado matemático.

Gabarito: B.

4. (CESPE/TELEBRAS/2022)



Considerando que o histograma apresentado descreve a distribuição de uma variável quantitativa X por meio de frequências absolutas, julgue o item que se segue.

O número de observações que constituem a variável X é igual a 1.000.

Comentários:

O número de observações consiste na soma de todas as frequências representadas no histograma. Assim, com base no histograma, temos que a soma das frequências é igual a:

$$400 + 300 + 200 + 150 + 50 = 1100$$

Gabarito: Errado.

5. (CESPE/SEDUC AL/2021) Com base em estatística, julgue o item a seguir.

Suponha que o histograma a seguir represente a frequência relativa de alunos, distribuída por faixa etária, que ingressaram no ensino superior no estado de Alagoas em 2020. Com base nas informações desse gráfico, é correto afirmar que mais de 50% dos novos alunos têm idade superior a 22 anos.

Comentários:

Observando o histograma, percebemos que as duas primeiras classes correspondem a mais de 50% dos dados, pois possuem frequências superiores a 30% e 20%, respectivamente. Como as idades nesse intervalo variam de 18 a 22 anos, menos de 50% terão idades superiores a 22 anos.

Gabarito: Errado.

6. (CESPE/IPHAN/2018) Julgue o item subsequente, referente à análise exploratória de dados.

A representação de diagramas de barras, de linha e de pizza possui escala de medida nominal e tem a moda como medida de tendência central.

Comentários:

Os diagramas de barras, de linha e de pizza são geralmente usados para representar dados nominais ou categóricos. Como veremos na aula de medidas de posição, a única medida de tendência central que podemos utilizar no caso de dados categóricos é a moda, pois ela apenas registra a categoria que mais se repete em uma amostra, sem envolver outras operações matemáticas.

Gabarito: Certo.

7. (CESPE/IPHAN/2018). Julgue o item subsequente, referente à análise exploratória de dados.

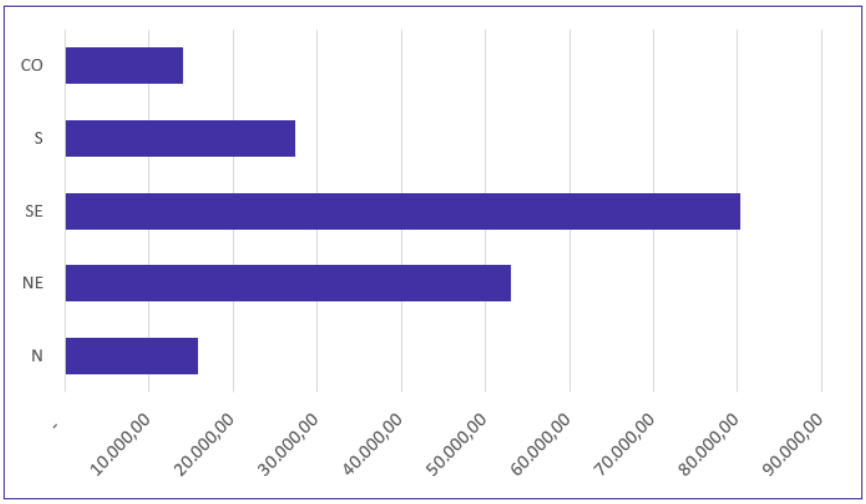
O gráfico de barras é adequado para a análise de variáveis qualitativas ordinais ou quantitativas discretas, pois permite investigar a presença de tendência nos dados.

Comentários:

O gráfico em barras normalmente é usado para representar distribuições de dados categóricos ou qualitativos. Por meio desse gráfico, uma série estatística é representada por um conjunto de retângulos dispostos horizontalmente, cada um indicando uma categoria em particular, os quais possuem a mesma altura e comprimentos proporcionais aos respectivos dados.

Para ficar claro, apresentamos um gráfico de barras como exemplo:

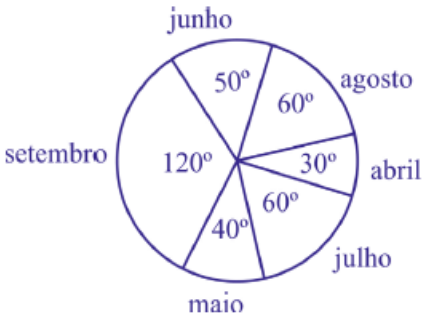
População brasileira, por Grandes Regiões, em 2010 (x1000)



Fonte: Censo Demográfico 1970/2010 (IBGE)

Gabarito: Certo.

8. (CESPE/CBM-AL/2017) O gráfico de setores a seguir mostra a distribuição das quantidades de incêndios em determinada região, nos meses de abril a setembro de determinado ano.



Sabendo-se que nesses meses ocorreram 1.548 incêndios nessa região, julgue o item que se segue.

A frequência relativa à classe “incêndios no mês de setembro” é superior a 30%.

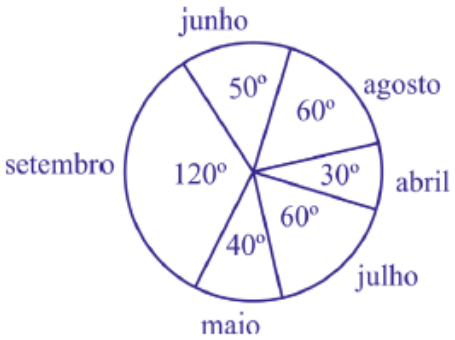
Comentários:

Para resolvermos essa questão, basta aplicarmos uma regra de três simples. Se uma circunferência total tem 360° e corresponde a 100%, então o ângulo de 120° corresponderá a:

$$Set. = \frac{120}{360} = \frac{1}{3} \approx 33{,}33\%$$

Gabarito: Certo.

9. (CESPE/CBM-AL/2017) O gráfico de setores a seguir mostra a distribuição das quantidades de incêndios em determinada região, nos meses de abril a setembro de determinado ano.



Sabendo-se que nesses meses ocorreram 1.548 incêndios nessa região, julgue o item que se segue.

Nos meses de maio e junho ocorreram mais de 400 incêndios nessa região.

Comentários:

Para resolver a questão, devemos aplicar uma regra de três simples. Isto é, se a circunferência total possui 360° e corresponde a 100%, então 360° também equivale ao número total de incêndios, 1548.

Podemos verificar também que os meses de maio e junho somam um total de 90°. Calculando 90°, temos:

$$360 - 1548$$

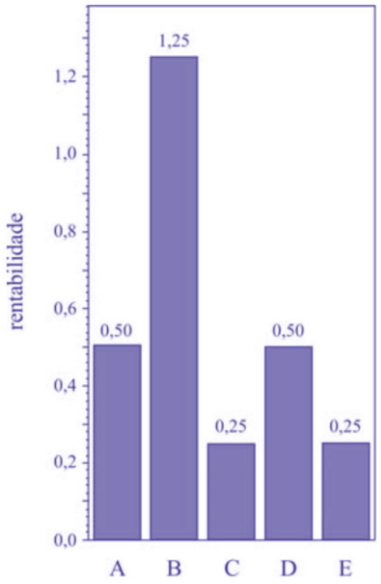
$$90 - X$$

$$X = \frac{139.320}{360} = 387$$

Logo, a quantidade de incêndios durante os meses de maio e junho não foi superior a 400, como afirma o item.

Gabarito: Errado.

10. (CESPE/FUNPRESP/2016)



O gráfico ilustra cinco possibilidades de fundos de investimento com suas respectivas rentabilidades. Considerando que as probabilidades de investimento para os fundos A, B, C e D sejam, respectivamente, P(A) = 0,182; P(B) = 0,454; P(C)= 0,091; e P(D)= 0,182, julgue o item subsequente.

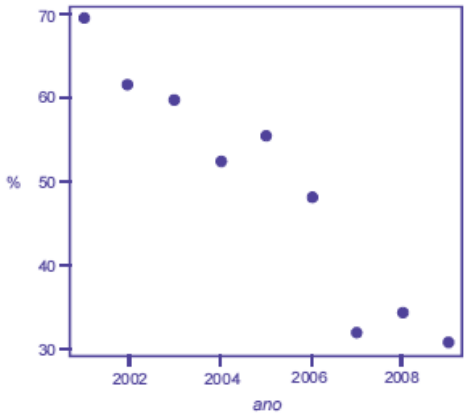
O gráfico apresentado é um histograma.

Comentários:

O gráfico não é um histograma por dois motivos: primeiro porque há uma separação entre as colunas, o que não ocorre em um histograma, e sim em um gráfico de colunas; segundo porque um histograma representa dados que estão agrupados em intervalos de classe, e não em categorias, como é o caso.

Gabarito: Errado.

11. (CESPE/DEPEN/2015)

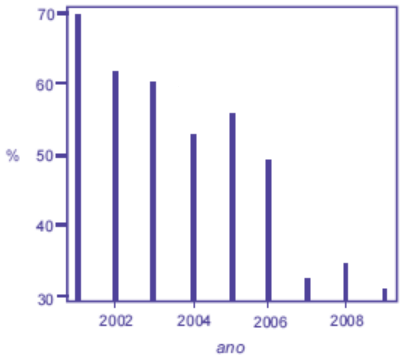


Dado que a participação dos presidiários em cursos de qualificação profissional é um aspecto importante para a reintegração do egresso do sistema prisional à sociedade, foram realizados levantamentos

estatísticos, nos anos de 2001 a 2009, a respeito do valor da educação e do trabalho em ambientes prisionais. Cada um desses levantamentos, cujos resultados são apresentados no gráfico, produziu uma estimativa anual do percentual P de indivíduos que participaram de um curso de qualificação profissional de curta duração, mas que não receberam o diploma por motivos diversos. Em 2001, 69,4% dos presidiários que participaram de um curso de qualificação profissional não receberam o diploma. No ano seguinte, 2002, esse percentual foi reduzido para 61,5%, caindo, em 2009, para 30,9%.

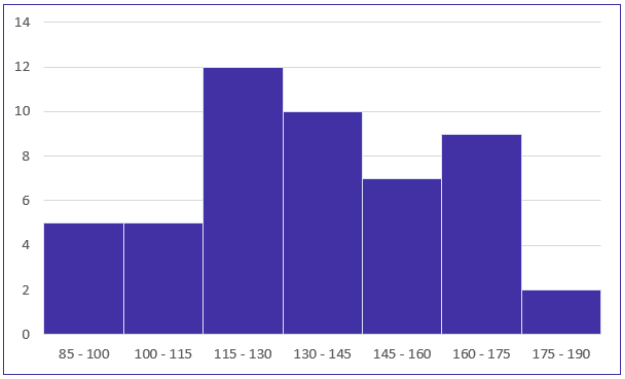
A partir das informações e do gráfico apresentados, julgue o item que se segue.

Se os percentuais forem representados por barras verticais, conforme o gráfico a seguir, então o resultado será denominado histograma.



Comentários:

O histograma é destinado a representar dados contínuos agrupados em classes (em intervalos), como o mostrado na figura a seguir.



Em realidade, a assertiva apresentou um gráfico de colunas (barras verticais) que representa dados agrupados por valores.

Gabarito: Errado.

12. (CESPE/DEPEN/2015)

Região	Quantidade de detentos no sistema penitenciário brasileiro (mil pessoas)	Déficit de vagas no sistema penitenciário (mil vagas)	População brasileira (milhões de habitantes)
Norte	37	13	17
Centro-Oeste	51	24	15
Nordeste	94	42	55
Sudeste	306	120	85
Sul	67	16	28
Total	555	215	200

Ministério da Justiça — Departamento Penitenciário Nacional — Sistema Integrado de Informações Penitenciárias – InfoPen, Relatório Estatístico Sintético do Sistema Prisional Brasileiro, dez./2013

Internet: <www.justica.gov.br> (com adaptações).

A tabela mostrada apresenta a quantidade de detentos no sistema penitenciário brasileiro por região em 2013. Nesse ano, o déficit relativo de vagas — que se define pela razão entre o déficit de vagas no sistema penitenciário e a quantidade de detentos no sistema penitenciário — registrado em todo o Brasil foi superior a 38,7%, e, na média nacional, havia 277,5 detentos por 100 mil habitantes.

Com base nessas informações e na tabela apresentada, julgue o item a seguir.

Na análise exploratória, o histograma é um gráfico adequado para descrever a distribuição da quantidade de detentos por região em 2013.

Comentários:

O histograma é destinado a representar dados agrupados em classes, mas o enunciado apresenta dados com valores individualizados. Como a informação não está agrupada em classes, o histograma não representaria adequadamente os dados da tabela.

Gabarito: Errado.

13. (CESPE/BACEN/2013)

2 4 8 4 8 1 2 32 12 1 5 7 5 5 3 4 24 19 4 14

Os dados mostrados acima representam uma amostra, em minutos, do tempo utilizado na armazenagem de formulários no almoxarifado central de certa instituição por diversos funcionários.

Com base nesses dados, julgue o próximo item.

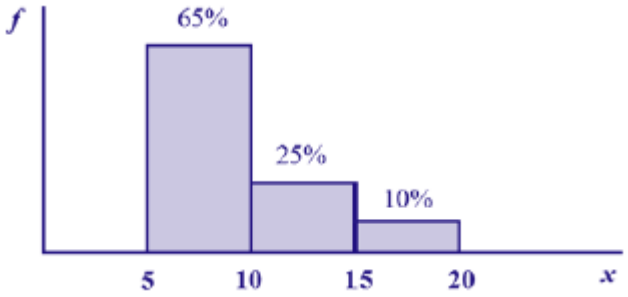
É inviável a elaboração de um histograma em decorrência do fato de ser este um conjunto de dados quantitativos discretos; dessa forma, apenas por meio de um gráfico de barras pode ser realizada a representação gráfica.

Comentários:

O item está **errado**. É sim possível elaborar um histograma para os dados apresentados, bastaria, para tanto, organizá-los em intervalos de classe.

Gabarito: Errado.

14. (CESPE/STF/2013)



Com referência à figura acima, que mostra a distribuição da renda mensal — x, em quantidades de salários mínimos (sm) — das pessoas que residem em determinada região, julgue o item subsequente.

A variável x, por possuir quatro níveis de respostas, é do tipo qualitativa ordinal.

Comentários:

O gráfico representa um histograma, em que as pessoas foram classificadas de acordo com as quantidades de salários-mínimos recebidos. Reparem na existência de três intervalos de classe (5 a 10; 10 a 15; e 15 a 20). Portanto, a variável x exprime a quantidade de salários-mínimos, sendo classificada como uma variável quantitativa.

Gabarito: Errado.

LISTA DE QUESTÕES – CEBRASPE

Conceitos Iniciais

1. (CESPE/PGDF/2021) Certa empresa desejava conhecer as opiniões de seus 20.000 funcionários acerca da confiança que eles têm no canal interno de denúncias. Para tanto, elaborou-se um questionário eletrônico que foi remetido, por email, para todos os endereços eletrônicos cadastrados, tendo sido desenvolvidos mecanismos para evitar que uma pessoa respondesse em lugar de outra, ou que uma mesma pessoa respondesse mais de uma vez. O questionário foi respondido por 400 pessoas, das quais 68% disseram confiar no processo de apuração de denúncias e 32% disseram ter reservas quanto ao processo. Verificou-se ainda que cerca de 500 mensagens retornaram por falha no cadastro dos endereços eletrônicos (erros de digitação), e que algumas respostas foram atribuídas a pessoas que não são mais funcionários; ainda, os endereços eletrônicos de alguns funcionários recém contratados não constavam do cadastro.

Com relação a essa situação hipotética, julgue o item a seguir.

As informações apresentadas permitem afirmar que a população- alvo da pesquisa difere da população referenciada.

2. (CESPE/DEPEN/2015) O diretor de um sistema penitenciário, com o propósito de estimar o percentual de detentos que possuem filhos, entregou a um analista um cadastro com os nomes de 500 detentos da instituição para que esse profissional realizasse entrevistas com os indivíduos selecionados.

A partir dessa situação hipotética e dos múltiplos aspectos a ela relacionados, julgue o item, referente a técnicas de amostragem.

A diferença entre um censo e uma amostra consiste no fato de esta última exigir a realização de um número maior de entrevistas.

3. (CESPE/SEFAZ-AL/2002) Julgue os seguintes itens.

Um censo consiste no estudo de todos os indivíduos da população considerada.

4. (CESPE/SEFAZ-AL/2002) Julgue os seguintes itens.

Como a realização de um censo tipicamente é muito onerosa e (ou) demorada, muitas vezes é conveniente estudar um subconjunto próprio da população, denominado amostra.

GABARITO – CEBRASPE

Conceitos Iniciais

1. CERTO
2. ERRADO
3. CERTO
4. CERTO

LISTA DE QUESTÕES – CEBRASPE

Variáveis Estatísticas

1. (CESPE/DPE-RO/2022)

Variável	Valores
estado civil	casado, solteiro, divorciado
quantidade de filhos	0, 1, 2, 3 ...
salário	6.510,25; 7.915,68
idade	22, 23, 27

Com relação às variáveis apresentadas na tabela anterior, julgue os itens a seguir.

I. A variável estado civil é qualitativa nominal.

II. A variável quantidade de filhos é quantitativa discreta.

III. As variáveis salário e estado civil são quantitativas discretas.

IV. As variáveis idade e quantidade de filhos são qualitativas nominais.

Estão certos apenas os itens

a) I e II.

b) II e III.

c) III e IV.

d) I, II e IV.

e) I, III e IV.

2. (CESPE/DPE-RO/2022) O valor de um atributo de um dado objeto é uma medida da quantidade daquele atributo, a qual pode ser numérica ou categórica.

Nesse caso, estado civil e sexo são classificados como atributo

a) binário.

b) nominal.

c) ordinal.

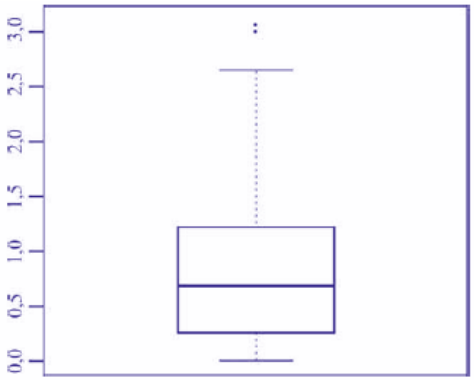
d) ausente.

e) razão.

3. (CESPE/IPHAN/2018). Julgue o item subsequente, referente à análise exploratória de dados.

O gráfico de barras é adequado para a análise de variáveis qualitativas ordinais ou quantitativas discretas, pois permite investigar a presença de tendência nos dados.

4. (CESPE/TCE-PA/2016)



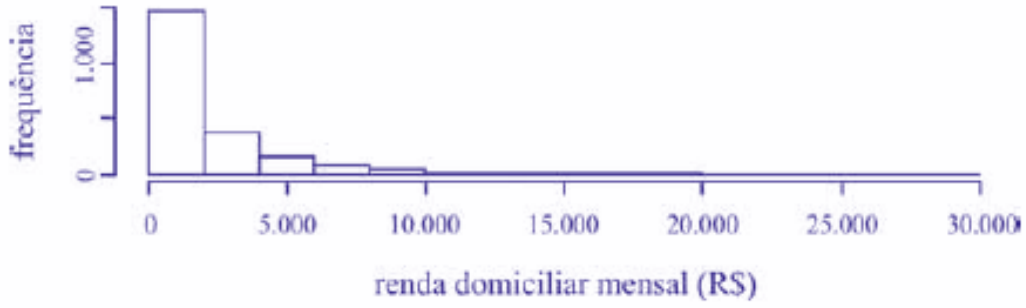
média amostral	0,80
desvio padrão amostral	0,70
primeiro quartil	0,25
mediana	0,70
terceiro quartil	1,20
mínimo	0
máximo	3,10

Um indicador de desempenho X permite avaliar a qualidade dos processos de governança de instituições públicas. A figura mostra, esquematicamente, a sua distribuição, obtida mediante estudo amostral feito por determinada agência de pesquisa. A tabela apresenta estatísticas descritivas referentes a essa distribuição.

Com base nessas informações, julgue o item a seguir.

X representa uma variável qualitativa ordinal.

5. (CESPE/TELEBRAS/2015)



Uma empresa coletou e armazenou em um banco de dados diversas informações sobre seus clientes, entre as quais estavam o valor da última fatura vencida e o pagamento ou não dessa fatura. Analisando essas informações, a empresa concluiu que 15% de seus clientes estavam inadimplentes. A empresa recolheu ainda dados como a unidade da Federação (UF) e o CEP da localidade em que estão os clientes. Do conjunto de todos os clientes, uma amostra aleatória simples constituída por 2.175 indivíduos prestou também

informações sobre sua renda domiciliar mensal, o que gerou o histograma apresentado. Com base nessas informações e no histograma, julgue o item a seguir.

O CEP da localidade dos clientes e o valor da última fatura vencida são variáveis quantitativas.

6. (CESPE/TELEBRAS/2015) Roberto comprou, por R$ 2.800,00, rodas de liga leve para seu carro, e, ao estacionar no shopping, ficou indeciso sobre onde deixar o carro, pois, caso o coloque no estacionamento público, correrá o risco de lhe roubarem as rodas, ao passo que, caso o coloque no estacionamento privado, terá de pagar R$ 70,00, com a garantia de que eventuais prejuízos serão ressarcidos pela empresa administradora.

Considerando que p seja a probabilidade de as rodas serem roubadas no estacionamento público, que X seja a variável aleatória que representa o prejuízo, em reais, ao deixar o carro no estacionamento público, e que Y seja a variável aleatória que representa o valor, em reais, desembolsado por Roberto ao deixar o carro no estacionamento pago, julgue o item subsequente.

A variável aleatória Y é contínua.

7. (CESPE/TJ-SE/2014)

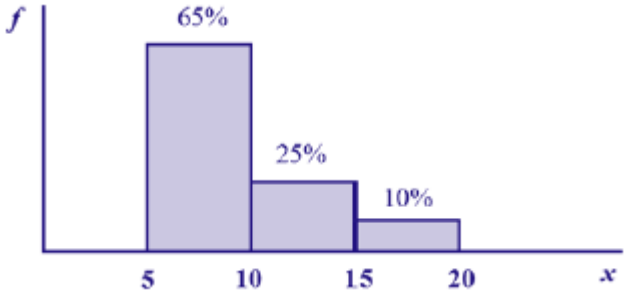
Quantidade	São Paulo (j =1)	Rio de Janeiro (j =2)	Minas Gerais (j =3)	Rio Grande do Sul (j =4)	Total
Casos novos (X, em milhões)	5	2	1	2	**18**
Casos pendentes (Y, em milhões)	16	8	3	2	**48**
Processos baixados (Z, em milhões)	5	2	2	2	**18**
Sentenças e Decisões (W, em milhões)	4	3	1	1	**16**

O quadro acima mostra uma síntese da movimentação processual dos tribunais de justiça dos estados de São Paulo, Rio de Janeiro, Minas Gerais, Rio Grande do Sul e do total da justiça estadual no Brasil em 2010. Considere que o estoque de processos em andamento no estado j (Ej), no final de 2010, seja um indicador que se define como $E_j = X_j + Y_j - Z_j - W_j$, em que j = 1, 2, ..., 27; X_j representa o número de casos novos registrados em 2010 no estado j; Yj seja a quantidade de casos pendentes no estado j (i.e., casos

anteriores que não foram solucionados até o final de 2010); Z_j denota o total de processos baixados (arquivados) no estado j durante 2010 e W_j seja o número de sentenças e decisões proferidas no estado j até o final de 2010. Considere, por fim, que, para todos os efeitos, o Distrito Federal seja um estado. Com base nessas informações e no quadro acima, julgue o item que se segue.

O quadro apresentado é uma tabela de contingência que mostra o cruzamento entre uma variável qualitativa nominal com 4 níveis de resposta (estados) e outra variável qualitativa com quatro níveis de resposta (casos novos, pendentes, baixados e resolvidos).

8. (CESPE/STF/2013)



Com referência à figura acima, que mostra a distribuição da renda mensal — x, em quantidades de salários mínimos (sm) — das pessoas que residem em determinada região, julgue o item subsequente.

A variável x, por possuir quatro níveis de respostas, é do tipo qualitativa ordinal.

9. (CESPE/TRE-ES/2011)

Quantidade de eleitores	Quantidade de municípios
$0 \vdash 2.000$	364
$2.000 \vdash 4.000$	1.000
$4.000 \vdash 6.000$	3.000
$6.000 \vdash 8.000$	1.000
$8.000 \vdash 10.000$	200
Total	**5.564**

A tabela acima apresenta uma distribuição hipotética das quantidades de eleitores que não votaram no segundo turno da eleição para presidente da República bem como os números de municípios em que essas

quantidades ocorreram. Com base nessa tabela, julgue o item seguinte, relativo à análise exploratória de dados.

Na tabela de frequências, o uso de intervalos de classe permite concluir que a variável em questão é contínua.

10. (CESPE/TRE-ES/2011)

Cargo	Candidatos	Candidatos aptos	Eleitos
Presidente da República	9	9	1
Governador de Estado	170	156	27
Senador	272	234	54
Deputado Federal	6.021	5.058	513
Deputado Estadual/Distrital	15.268	13.076	1.059
Total	21.640	18.533	1.658

Internet: <www.tse.gov > (com adaptações).

Com base na tabela acima, referente às eleições de 2010, que apresenta a quantidade de candidatos para os cargos de presidente da República, governador de estado, senador, deputado federal e deputado estadual/distrital, bem como a quantidade de candidatos considerados aptos pela justiça eleitoral e o total de eleitos para cada cargo pretendido, julgue o item a seguir.

A variável "cargo" classifica-se como uma variável qualitativa ordinal.

11. (CESPE/TCU/2008) Uma agência de desenvolvimento urbano divulgou os dados apresentados na tabela a seguir, acerca dos números de imóveis ofertados (X) e vendidos (Y) em determinado município, nos anos de 2005 a 2007.

Ano	Número de imóveis	
	Ofertados (X)	Vendidos (Y)
2005	1.500	100
2006	1.750	400

2007	2.000	700

Correios Braziliense, 29/4/2008, p.17 (com adaptações)

Com respeito ao texto, considere que cada imóvel ofertado em determinado ano seja classificado como vendido ou não-vendido, e, a um imóvel e classificado como vendido seja atribuído um valor Z = 1, e, ao imóvel classificado como não-vendido, seja atribuído um valor Z = 0. Supondo-se que as classificações dos imóveis como vendido ou não-vendido em um dado ano possam ser consideradas como sendo realizações de uma amostragem aleatória simples, julgue os itens a seguir.

A variável Z é classificada como variável qualitativa nominal, pois representa o atributo do imóvel como vendido ou não-vendido.

GABARITO – CEBRASPE

Variáveis Estatísticas

1. LETRA A
2. LETRA B
3. CERTO
4. ERRADO
5. ERRADO
6. ERRADO
7. ERRADO
8. ERRADO
9. ERRADO
10. CERTO
11. ERRADO

LISTA DE QUESTÕES – CEBRASPE

Séries Estatísticas

1. (CESPE/ANTAQ/2009)

	Variável	2003	2004	2005	2006	2007
Exportação	X	40	46	50	52	54
Importação	Y	20	21	22	24	27
Total	X+Y	60	67	72	76	81

Internet: <www.portodesantos.com> (com adaptações).

Considerando a tabela acima, que apresenta a movimentação anual de cargas no porto de Santos de 2003 a 2007, em milhões de toneladas/ ano e associa as quantidades de carga movimentadas para exportação e importação às variáveis X e Y, respectivamente, julgue o item subsequente.

As séries estatísticas apresentadas na tabela formam três séries temporais.

2. (CESPE/TCU/2008) Uma agência de desenvolvimento urbano divulgou os dados apresentados na tabela a seguir, acerca dos números de imóveis ofertados (X) e vendidos (Y) em determinado município, nos anos de 2005 a 2007.

Ano	Número de imóveis	
	Ofertados (X)	Vendidos (Y)
2005	1.500	100
2006	1.750	400
2007	2.000	700

Correios Braziliense, 29/4/2008, p.17 (com adaptações)

Considerando as informações do texto, julgue os itens subsequentes.

A variável X forma uma série estatística denominada série temporal.

GABARITO – CEBRASPE

Séries Estatísticas

1. CERTO

2. CERTO

LISTA DE QUESTÕES – CEBRASPE

Distribuições de Frequência

1. (CEBRASPE/BACEN/2024)

x	$P(X = x)$
-2	$5c$
-1	c
0	$2c$
+1	$3c$
+2	$4c$

Considerando que X representa uma variável aleatória com suporte $x \in \{-2, -1, 0, +1, +2\}$, cuja função de distribuição de probabilidade é dada no quadro acima, na qual c é uma constante real positiva, julgue os próximos itens.

X segue uma distribuição contínua, pois C é uma constante real positiva.

2. (CEBRASPE/BACEN/2024)

x	$P(X = x)$
-2	$5c$
-1	c
0	$2c$
+1	$3c$
+2	$4c$

Considerando que X representa uma variável aleatória com suporte $x \in \{-2, -1, 0, +1, +2\}$, cuja função de distribuição de probabilidade é dada no quadro acima, na qual c é uma constante real positiva, julgue os próximos itens.

$P(X = +1) = 0{,}2$.

3. (CEBRASPE/CNJ/2024)

X	frequência absoluta acumulada
0	120
1	180
2	220
3	240
4	250

A tabela precedente mostra a distribuição de frequências do número diário (X) de denúncias recebidas pela ouvidoria de um tribunal de justiça.

Com base nos dados apresentados na tabela, julgue os itens que se seguem.

O tamanho da amostra é igual a 1.010.

4. (CEBRASPE/TC-DF/2023)

X	Frequência Absoluta	Frequência Relativa
0	3	0,10
5	6	0,20
10	15	0,50
15	6	0,20

Considerando que, em um levantamento estatístico realizado por amostragem aleatória simples, tenha sido produzida a tabela de frequências apresentada anteriormente, na qual X denota uma variável de interesse, julgue os seguintes itens.

O tamanho da amostra é igual ou superior a 16.

5. (CEBRASPE/FUB/2022) A tabela de frequência a seguir mostra dados coletados em uma pesquisa para se verificar o número de disciplinas que os estudantes de determinada universidade estão cursando por semestre.

Disciplinas	2	3	4	5	6	7	8
Estudantes	10	15	40	35	28	10	4

Considerando essas informações, julgue o item seguinte.

A proporção de alunos que cursam mais de 6 disciplinas é maior que a proporção de alunos que cursam 3 disciplinas.

6. (CEBRASPE/FUB/2022) Uma universidade está fazendo um estudo para verificar a distribuição dos tempos que os alunos do curso de mestrado levam até a defesa da dissertação. Os dados a seguir mostram a função de probabilidade desses tempos, em meses.

Tempo de Defesa (meses)	Probabilidade
12	0,01
15	0,02
18	0,04
20	0,10
22	0,22
24	0,31
25	0,18
26	0,04
28	0,03
30	0,05

Considerando essas informações, julgue o item subsequente.

Se o prazo máximo recomendado para a defesa da dissertação de mestrado é de 24 meses, então a probabilidade de um aluno defender sua dissertação até 2 meses antes desse prazo é igual à probabilidade de um aluno defendê-la até 2 meses depois.

7. (CEBRASPE/FUB/2022) Uma universidade está fazendo um estudo para verificar a distribuição dos tempos que os alunos do curso de mestrado levam até a defesa da dissertação. Os dados a seguir mostram a função de probabilidade desses tempos, em meses.

Tempo de Defesa (meses)	Probabilidade

12	0,01
15	0,02
18	0,04
20	0,10
22	0,22
24	0,31
25	0,18
26	0,04
28	0,03
30	0,05

Considerando essas informações, julgue o item subsequente.

Os valores da probabilidade de um aluno defender a dissertação em 13, 14, 16, 19, 21, 23, 27 ou 29 meses, somados, é igual à probabilidade de um aluno defender a dissertação em exatamente 31 meses.

8. (CEBRASPE/FUB/2022) Uma universidade está fazendo um estudo para verificar a distribuição dos tempos que os alunos do curso de mestrado levam até a defesa da dissertação. Os dados a seguir mostram a função de probabilidade desses tempos, em meses.

Tempo de Defesa (meses)	Probabilidade
12	0,01
15	0,02
18	0,04
20	0,10
22	0,22
24	0,31

25	0,18
26	0,04
28	0,03
30	0,05

Considerando essas informações, julgue o item subsequente.

Se o prazo máximo de defesa recomendado é de 24 meses, então a probabilidade de um aluno defender sua dissertação no prazo é superior a 70%.

9. (CEBRASPE/FUB/2022) A tabela de frequência a seguir mostra dados coletados em uma pesquisa para se verificar o número de disciplinas que os estudantes de determinada universidade estão cursando por semestre.

Disciplinas	2	3	4	5	6	7	8
Estudantes	10	15	40	35	28	10	4

Considerando essas informações, julgue o item seguinte.

Na pesquisa foram entrevistados 142 alunos.

10. (CEBRASPE/Polícia Federal/2018)

	DIA				
	1	2	3	4	5
X (quantidade diária de drogas apreendidas, em kg)	10	22	18	22	28

Tendo em vista que, diariamente, a Polícia Federal apreende uma quantidade X, em kg, de drogas em determinado aeroporto do Brasil, e considerando os dados hipotéticos da tabela precedente, que apresenta os valores observados da variável X em uma amostra aleatória de 5 dias de apreensões no citado aeroporto, julgue o item.

A tabela em questão descreve a distribuição de frequências da quantidade de drogas apreendidas nos cinco dias que constituem a amostra.

11. (CEBRASPE/IPHAN/2018). A tabela a seguir mostra as quantidades de bibliotecas públicas presentes em 20 microrregiões brasileiras.

90	66	78	82
77	60	64	90
87	85	67	91
82	70	81	80
69	78	90	67

A partir desses dados, pretende-se construir um gráfico de distribuição de frequências com quatro classes de igual amplitude. Os valores mínimo e máximo de cada classe devem ser números inteiros.

Considerando essas informações, julgue o item subsequente, relativo ao gráfico de distribuição a ser apresentado.

A amplitude de cada classe deverá ser superior a 6.

12. (CEBRASPE/IPHAN/2018) A tabela a seguir mostra as quantidades de bibliotecas públicas presentes em 20 microrregiões brasileiras.

90	66	78	82
77	60	64	90
87	85	67	91
82	70	81	80
69	78	90	67

A partir desses dados, pretende-se construir um gráfico de distribuição de frequências com quatro classes de igual amplitude. Os valores mínimo e máximo de cada classe devem ser números inteiros. Considerando essas informações, julgue o item subsequente, relativo ao gráfico de distribuição a ser apresentado.

A última classe deverá variar de 84 a 91.

13. (CEBRASPE/IPHAN/2018). Julgue o item subsequente, referente à análise exploratória de dados.

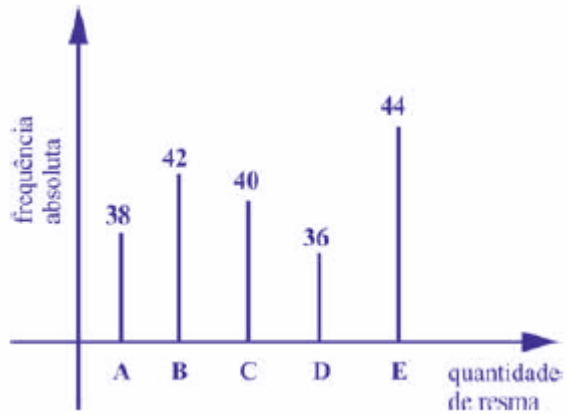
O histograma é um diagrama de retângulos contíguos com base na curtose das faixas de valores da variável e com área igual à diferença da frequência absoluta da respectiva faixa.

14. (CEBRASPE/CBM-AL/2017) Na tabela a seguir, A, B, C, D e E são as quantidades de resmas de papel A4 consumidas, em quatro meses, pelas seções administrativas I, II, III, IV e V, respectivamente. Apesar de não mostrar explicitamente essas quantidades, a tabela apresenta as frequências absolutas e (ou) relativas de algumas dessas quantidades.

Seção	Quantidades de Resmas	Frequência Absoluta	Frequência Relativa
I	A	38	19%
II	B		
III	C		20%
IV	D	36	
V	E	44	
	Total		**100%**

Considerando que cada uma dessas resmas, juntamente com a embalagem, tem forma de um paralelepípedo retângulo reto que mede 5 cm × 21 cm × 30 cm, julgue o item seguinte.

O gráfico de barras verticais a seguir apresenta as frequências absolutas de resmas consumidas pelas cinco seções.



15. (CEBRASPE/TCE-PA/2016)

Número diário de denúncias registradas (X)	Frequência Relativa
0	0,3
1	0,1
2	0,2
3	0,1
4	0,3
Total	**1,0**

A tabela precedente apresenta a distribuição de frequências relativas da variável X, que representa o número diário de denúncias registradas na ouvidoria de determinada instituição pública. A partir das informações dessa tabela, julgue o item seguinte.

A variável X é do tipo qualitativo nominal.

16. (CEBRASPE/DEPEN/2015)

Região	Quantidade de detentos no sistema penitenciário brasileiro (mil pessoas)	Déficit de vagas no sistema penitenciário (mil vagas)	População brasileira (milhões de habitantes)
Norte	37	13	17
Centro-Oeste	51	24	15
Nordeste	94	42	55
Sudeste	306	120	85
Sul	67	16	28
Total	555	215	200

Ministério da Justiça — Departamento Penitenciário Nacional — Sistema Integrado de Informações Penitenciárias – InfoPen, Relatório Estatístico Sintético do Sistema Prisional Brasileiro, dez./2013

Internet: <www.justica.gov.br> (com adaptações).

A tabela mostrada apresenta a quantidade de detentos no sistema penitenciário brasileiro por região em 2013. Nesse ano, o déficit relativo de vagas — que se define pela razão entre o déficit de vagas no sistema penitenciário e a quantidade de detentos no sistema penitenciário — registrado em todo o Brasil foi superior a 38,7%, e, na média nacional, havia 277,5 detentos por 100 mil habitantes.

Com base nessas informações e na tabela apresentada, julgue o item a seguir.

Em 2013, mais de 55% da população carcerária no Brasil se encontrava na região Sudeste.

17. (CEBRASPE/DEPEN/2015)

Região	Quantidade de detentos no sistema penitenciário brasileiro (mil pessoas)	Déficit de vagas no sistema penitenciário (mil vagas)	População brasileira (milhões de habitantes)
Norte	37	13	17
Centro-Oeste	51	24	15
Nordeste	94	42	55
Sudeste	306	120	85
Sul	67	16	28
Total	555	215	200

Ministério da Justiça — Departamento Penitenciário Nacional — Sistema Integrado de Informações Penitenciárias – InfoPen, Relatório Estatístico Sintético do Sistema Prisional Brasileiro, dez./2013

Internet: <www.justica.gov.br> (com adaptações).

A tabela mostrada apresenta a quantidade de detentos no sistema penitenciário brasileiro por região em 2013. Nesse ano, o déficit relativo de vagas — que se define pela razão entre o déficit de vagas no sistema penitenciário e a quantidade de detentos no sistema penitenciário — registrado em todo o Brasil foi superior a 38,7%, e, na média nacional, havia 277,5 detentos por 100 mil habitantes.

Com base nessas informações e na tabela apresentada, julgue o item a seguir.

A quantidade total de vagas existentes no sistema penitenciário brasileiro em 2013 era de 340 mil vagas.

18. (CEBRASPE/DEPEN/2015)

Região	Quantidade de detentos no sistema penitenciário brasileiro (mil pessoas)	Déficit de vagas no sistema penitenciário (mil vagas)	População brasileira (milhões de habitantes)
Norte	37	13	17
Centro-Oeste	51	24	15
Nordeste	94	42	55
Sudeste	306	120	85
Sul	67	16	28
Total	555	215	200

Ministério da Justiça — Departamento Penitenciário Nacional — Sistema Integrado de Informações Penitenciárias – InfoPen, Relatório Estatístico Sintético do Sistema Prisional Brasileiro, dez./2013

Internet: <www.justica.gov.br> (com adaptações).

A tabela mostrada apresenta a quantidade de detentos no sistema penitenciário brasileiro por região em 2013. Nesse ano, o déficit relativo de vagas — que se define pela razão entre o déficit de vagas no sistema penitenciário e a quantidade de detentos no sistema penitenciário — registrado em todo o Brasil foi superior a 38,7%, e, na média nacional, havia 277,5 detentos por 100 mil habitantes.

Com base nessas informações e na tabela apresentada, julgue o item a seguir.

O déficit relativo de vagas observado na região Sudeste, em 2013, foi superior ao déficit relativo de vagas registrado na região Centro-oeste no mesmo período.

19. (CEBRASPE/DEPEN/2015)

Região	Quantidade de detentos no sistema penitenciário brasileiro (mil pessoas)	Déficit de vagas no sistema penitenciário (mil vagas)	População brasileira (milhões de habitantes)
Norte	37	13	17
Centro-Oeste	51	24	15
Nordeste	94	42	55
Sudeste	306	120	85

Sul	67	16	28
Total	555	215	200

Ministério da Justiça — Departamento Penitenciário Nacional — Sistema Integrado de Informações Penitenciárias – InfoPen, Relatório Estatístico Sintético do Sistema Prisional Brasileiro, dez./2013

Internet: <www.justica.gov.br> (com adaptações).

A tabela mostrada apresenta a quantidade de detentos no sistema penitenciário brasileiro por região em 2013. Nesse ano, o déficit relativo de vagas — que se define pela razão entre o déficit de vagas no sistema penitenciário e a quantidade de detentos no sistema penitenciário — registrado em todo o Brasil foi superior a 38,7%, e, na média nacional, havia 277,5 detentos por 100 mil habitantes.

Com base nessas informações e na tabela apresentada, julgue o item a seguir.

No ano considerado, a quantidade média de detentos por 100 mil habitantes na região Nordeste foi superior ao número médio de detentos por 100 mil habitantes na região Centro-oeste.

20. (CESPE/BACEN/2013)

2 4 8 4 8 1 2 32 12 1 5 7 5 5 3 4 24 19 4 14

Os dados mostrados acima representam uma amostra, em minutos, do tempo utilizado na armazenagem de formulários no almoxarifado central de certa instituição por diversos funcionários.

Com base nesses dados, julgue o próximo item.

A distribuição de frequência acumulada para tempo de armazenagem observado na amostra inferior a 8 minutos é igual a 13, o que corresponde a uma frequência relativa superior a 0,60.

21. (CESPE/ALECE/2011)

X	0	1	2	3	4	5
Frequência Absoluta	80	47	30	20	6	1

Um levantamento foi realizado para se avaliar, por município, a quantidade X de obras que estão sob suspeita de irregularidade. Com base em uma amostra de municípios, foi obtida a distribuição de frequências mostrada na tabela acima. Com base nessas informações, julgue o item a seguir.

Os extremos mínimo e máximo da variável X foram, respectivamente, iguais a 1 e 80.

22. (CESPE/ALECE/2011)

X	0	1	2	3	4	5
Frequência Absoluta	80	47	30	20	6	1

Um levantamento foi realizado para se avaliar, por município, a quantidade X de obras que estão sob suspeita de irregularidade. Com base em uma amostra de municípios, foi obtida a distribuição de frequências mostrada na tabela acima. Com base nessas informações, julgue o item a seguir.

O número de municípios que têm obras sob suspeita de irregularidades é superior a 120.

23. (CESPE/ALECE/2011)

X	0	1	2	3	4	5
Frequência Absoluta	80	47	30	20	6	1

Um levantamento foi realizado para se avaliar, por município, a quantidade X de obras que estão sob suspeita de irregularidade. Com base em uma amostra de municípios, foi obtida a distribuição de frequências mostrada na tabela acima. Com base nessas informações, julgue o item a seguir.

O total de municípios considerado no levantamento foi superior a 180.

24. (CESPE/SEFAZ-MT/2004) Considere a seguinte situação hipotética.

Um órgão do governo recebeu pela Internet denúncias de sonegação de impostos estaduais contra 600 pequenas empresas. Denúncias contra outras 200 pequenas empresas foram encaminhadas pessoalmente para esse órgão. Para a apuração das denúncias, foram realizadas auditorias nas 800 empresas denunciadas. Como resultado dessas auditorias, foi elaborada a tabela abaixo, que apresenta um quadro das empresas denunciadas e os correspondentes débitos fiscais ao governo. Das empresas denunciadas, observou-se que apenas 430 tinham débitos fiscais.

	Valor do débito fiscal (VDF), em R$ mil, apurado após auditoria na empresa denunciada				
Forma de recebimento da denúncia	$0 < VDF < 1$	$1 \leq VDF < 2$	$2 \leq VDF < 3$	$3 \leq VDF \leq 4$	Total
Pela internet	60	100	50	30	240
Pessoalmente	20	120	40	10	190
Total	80	220	90	40	430*

Nota: *Para as demais empresas, VDF=0

Com base na situação hipotética acima e de acordo com as informações apresentadas, julgue o item que se segue.

O valor total dos débitos fiscais apurados após as auditorias feitas nas empresas denunciadas é inferior a R$ 500 mil.

25. (CESPE/SEFAZ-AL/2002) Julgue o seguinte item.

Em uma distribuição de frequências para um conjunto de n indivíduos, pode-se calcular as frequências relativas, dividindo-se cada frequência absoluta pela amplitude da correspondente classe ou do intervalo.

GABARITO – CEBRASPE

Distribuições de Frequência

1. ERRADO
2. CORRETO
3. ERRADO
4. CERTO
5. ERRADO
6. CERTO
7. CERTO
8. ERRADO
9. CERTO
10. ERRADO
11. CERTO
12. CERTO
13. ERRADO
14. CERTO
15. ERRADO
16. CERTO
17. CERTO
18. ERRADO
19. ERRADO
20. CERTO
21. ERRADO
22. ERRADO
23. CERTO
24. ERRADO
25. ERRADO

LISTA DE QUESTÕES – CEBRASPE

Representação Gráfica das Distribuições de Frequências

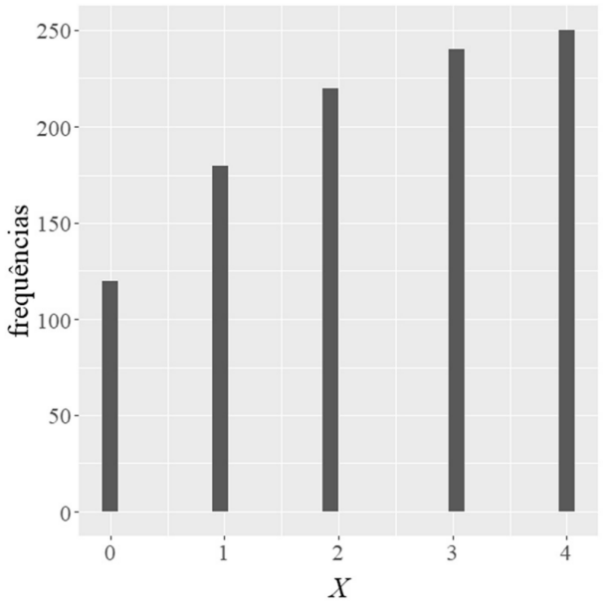
1. (CESPE/CNJ/2024)

X	frequência absoluta acumulada
0	120
1	180
2	220
3	240
4	250

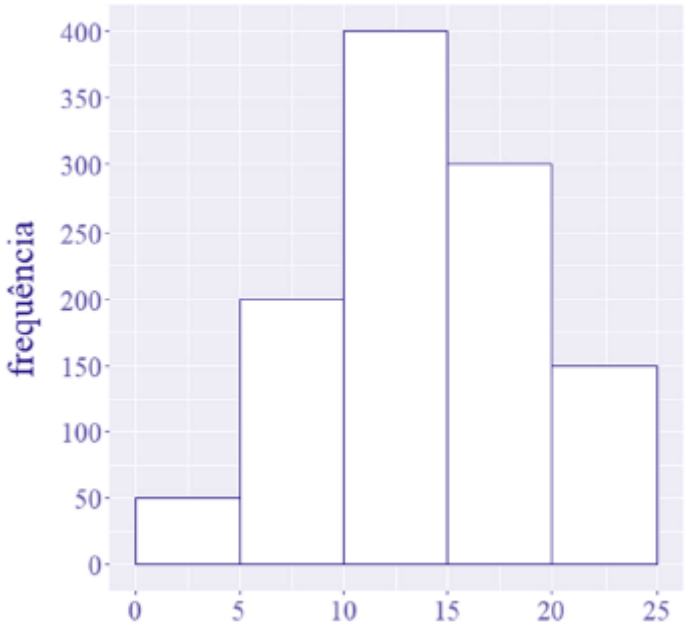
A tabela precedente mostra a distribuição de frequências do número diário (X) de denúncias recebidas pela ouvidoria de um tribunal de justiça.

Com base nos dados apresentados na tabela, julgue os itens que se seguem.

O histograma a seguir representa corretamente a distribuição de frequências da variável X.



2. (CESPE/TELEBRAS/2022)

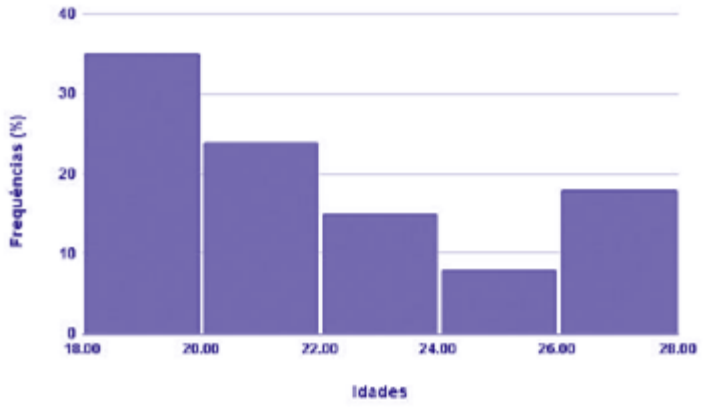


Considerando que o histograma apresentado descreve a distribuição de uma variável quantitativa X por meio de frequências absolutas, julgue o item que se segue.

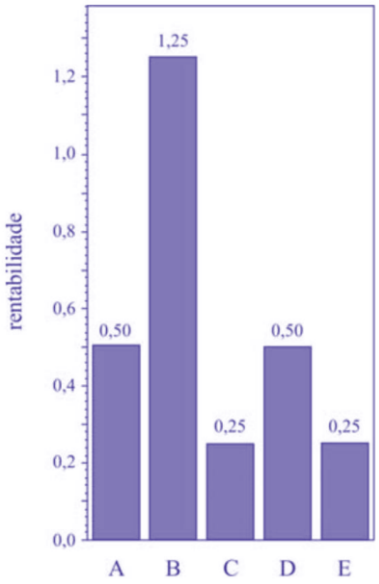
O número de observações que constituem a variável X é igual a 1.000.

3. (CESPE/SEDUC AL/2021) Com base em estatística, julgue o item a seguir.

Suponha que o histograma a seguir represente a frequência relativa de alunos, distribuída por faixa etária, que ingressaram no ensino superior no estado de Alagoas em 2020. Com base nas informações desse gráfico, é correto afirmar que mais de 50% dos novos alunos têm idade superior a 22 anos.



4. (CESPE/FUNPRESP/2016)



O gráfico ilustra cinco possibilidades de fundos de investimento com suas respectivas rentabilidades. Considerando que as probabilidades de investimento para os fundos A, B, C e D sejam, respectivamente, P(A) = 0,182; P(B) = 0,454; P(C)= 0,091; e P(D)= 0,182, julgue o item subsequente.

O gráfico apresentado é um histograma.

5. (CESPE/DEPEN/2015)

Região	Quantidade de detentos no sistema penitenciário brasileiro (mil pessoas)	Déficit de vagas no sistema penitenciário (mil vagas)	População brasileira (milhões de habitantes)
Norte	37	13	17
Centro-Oeste	51	24	15
Nordeste	94	42	55
Sudeste	306	120	85
Sul	67	16	28
Total	555	215	200

Ministério da Justiça — Departamento Penitenciário Nacional — Sistema Integrado de Informações Penitenciárias – InfoPen, Relatório Estatístico Sintético do Sistema Prisional Brasileiro, dez./2013

Internet: <www.justica.gov.br> (com adaptações).

A tabela mostrada apresenta a quantidade de detentos no sistema penitenciário brasileiro por região em 2013. Nesse ano, o déficit relativo de vagas — que se define pela razão entre o déficit de vagas no sistema penitenciário e a quantidade de detentos no sistema penitenciário — registrado em todo o Brasil foi superior a 38,7%, e, na média nacional, havia 277,5 detentos por 100 mil habitantes.

Com base nessas informações e na tabela apresentada, julgue o item a seguir.

Na análise exploratória, o histograma é um gráfico adequado para descrever a distribuição da quantidade de detentos por região em 2013.

6. (CESPE/BACEN/2013)

2 4 8 4 8 1 2 32 12 1 5 7 5 5 3 4 24 19 4 14

Os dados mostrados acima representam uma amostra, em minutos, do tempo utilizado na armazenagem de formulários no almoxarifado central de certa instituição por diversos funcionários.

Com base nesses dados, julgue o próximo item.

É inviável a elaboração de um histograma em decorrência do fato de ser este um conjunto de dados quantitativos discretos; dessa forma, apenas por meio de um gráfico de barras pode ser realizada a representação gráfica.

GABARITO – CEBRASPE

Representação Gráfica das Distribuições de Frequências

1. ERRADO
2. ERRADO
3. ERRADO
4. ERRADO
5. ERRADO
6. ERRADO

LISTA DE QUESTÕES – CEBRASPE

Outros Gráficos e Representações

1. (CESPE/FUB/2022) Uma universidade está fazendo um estudo para verificar a distribuição dos tempos que os alunos do curso de mestrado levam até a defesa da dissertação. Os dados a seguir mostram a função de probabilidade desses tempos, em meses.

Tempo de Defesa (meses)	Probabilidade
12	0,01
15	0,02
18	0,04
20	0,10
22	0,22
24	0,31
25	0,18
26	0,04
28	0,03
30	0,05

Considerando essas informações, julgue o item subsequente.

O gráfico de setores é adequado para representar a distribuição em questão.

2. (CESPE/FUB/2022) A tabela de frequência a seguir mostra dados coletados em uma pesquisa para se verificar o número de disciplinas que os estudantes de determinada universidade estão cursando por semestre.

Disciplinas	2	3	4	5	6	7	8
Estudantes	10	15	40	35	28	10	4

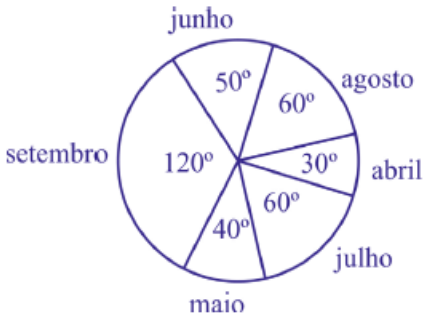
Considerando essas informações, julgue o item seguinte.

O gráfico do tipo pizza é o mais apropriado para representar os dados apresentados na tabela, visto que a variável analisada é qualitativa ordinal.

3. (CESPE/IPHAN/2018) Julgue o item subsequente, referente à análise exploratória de dados.

A representação de diagramas de barras, de linha e de pizza possui escala de medida nominal e tem a moda como medida de tendência central.

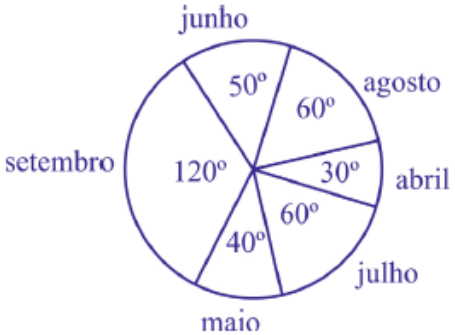
4. (CESPE/CBM-AL/2017) O gráfico de setores a seguir mostra a distribuição das quantidades de incêndios em determinada região, nos meses de abril a setembro de determinado ano.



Sabendo-se que nesses meses ocorreram 1.548 incêndios nessa região, julgue o item que se segue.

A frequência relativa à classe “incêndios no mês de setembro” é superior a 30%.

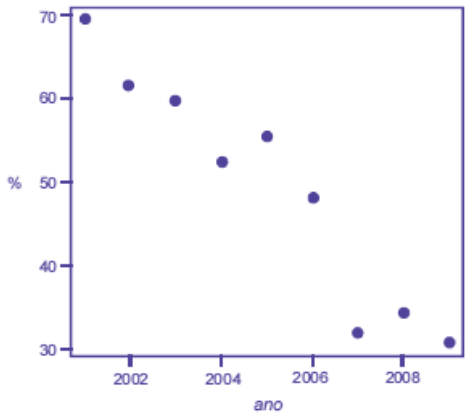
5. (CESPE/CBM-AL/2017) O gráfico de setores a seguir mostra a distribuição das quantidades de incêndios em determinada região, nos meses de abril a setembro de determinado ano.



Sabendo-se que nesses meses ocorreram 1.548 incêndios nessa região, julgue o item que se segue.

Nos meses de maio e junho ocorreram mais de 400 incêndios nessa região.

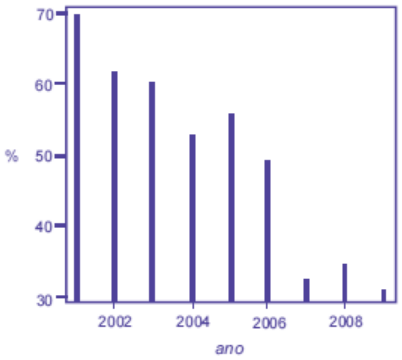
6. (CESPE/DEPEN/2015)



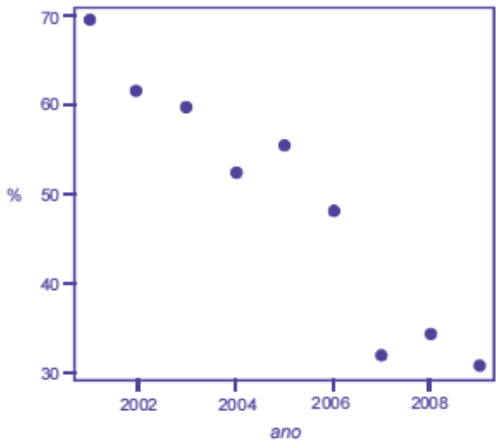
Dado que a participação dos presidiários em cursos de qualificação profissional é um aspecto importante para a reintegração do egresso do sistema prisional à sociedade, foram realizados levantamentos estatísticos, nos anos de 2001 a 2009, a respeito do valor da educação e do trabalho em ambientes prisionais. Cada um desses levantamentos, cujos resultados são apresentados no gráfico, produziu uma estimativa anual do percentual P de indivíduos que participaram de um curso de qualificação profissional de curta duração, mas que não receberam o diploma por motivos diversos. Em 2001, 69,4% dos presidiários que participaram de um curso de qualificação profissional não receberam o diploma. No ano seguinte, 2002, esse percentual foi reduzido para 61,5%, caindo, em 2009, para 30,9%.

A partir das informações e do gráfico apresentados, julgue o item que se segue.

Se os percentuais forem representados por barras verticais, conforme o gráfico a seguir, então o resultado será denominado histograma.



7. (CESPE/DEPEN/2015)



Dado que a participação dos presidiários em cursos de qualificação profissional é um aspecto importante para a reintegração do egresso do sistema prisional à sociedade, foram realizados levantamentos estatísticos, nos anos de 2001 a 2009, a respeito do valor da educação e do trabalho em ambientes prisionais. Cada um desses levantamentos, cujos resultados são apresentados no gráfico, produziu uma estimativa anual do percentual P de indivíduos que participaram de um curso de qualificação profissional de curta duração, mas que não receberam o diploma por motivos diversos. Em 2001, 69,4% dos presidiários que participaram de um curso de qualificação profissional não receberam o diploma. No ano seguinte, 2002, esse percentual foi reduzido para 61,5%, caindo, em 2009, para 30,9%.

A partir das informações e do gráfico apresentados, julgue o item que se segue.

Os dados apresentados são suficientes para que se possa afirmar que o total de presidiários que participaram de um curso de qualificação profissional de curta duração e que não receberam o diploma em 2008 foi superior ao total referente ao ano de 2007.

8. (CESPE/TRE-ES/2011)

Cargo	Candidatos	Candidatos aptos	Eleitos
Presidente da República	9	9	1
Governador de Estado	170	156	27
Senador	272	234	54
Deputado Federal	6.021	5.058	513
Deputado Estadual/Distrital	15.268	13.076	1.059
Total	21.640	18.533	1.658

Internet: <www.tse.gov > (com adaptações).

Com base na tabela acima, referente às eleições de 2010, que apresenta a quantidade de candidatos para os cargos de presidente da República, governador de estado, senador, deputado federal e deputado estadual/distrital, bem como a quantidade de candidatos considerados aptos pela justiça eleitoral e o total de eleitos para cada cargo pretendido, julgue o item a seguir.

Considerando-se a representação das quantidades de eleitos para cada cargo em um gráfico de pizza, a fatia desse gráfico correspondente ao cargo de deputado federal terá ângulo superior a 120°.

GABARITO – CEBRASPE

Outros Gráficos e Representações

1. ERRADO
2. ERRADO
3. CERTO
4. CERTO
5. ERRADO
6. ERRADO
7. ERRADO
8. ERRADO

LISTA DE QUESTÕES – CEBRASPE

Análise Exploratória de Dados

1. (CESPE/FUB/2022) Uma universidade está fazendo um estudo para verificar a distribuição dos tempos que os alunos do curso de mestrado levam até a defesa da dissertação. Os dados a seguir mostram a função de probabilidade desses tempos, em meses.

Tempo de Defesa (meses)	Probabilidade
12	0,01
15	0,02
18	0,04
20	0,10
22	0,22
24	0,31
25	0,18
26	0,04
28	0,03
30	0,05

Considerando essas informações, julgue o item subsequente.

O gráfico de setores é adequado para representar a distribuição em questão.

2. (CESPE/FUB/2022) A tabela de frequência a seguir mostra dados coletados em uma pesquisa para se verificar o número de disciplinas que os estudantes de determinada universidade estão cursando por semestre.

Disciplinas	2	3	4	5	6	7	8
Estudantes	10	15	40	35	28	10	4

Considerando essas informações, julgue o item seguinte.

O gráfico do tipo pizza é o mais apropriado para representar os dados apresentados na tabela, visto que a variável analisada é qualitativa ordinal.

3. (CESPE/DPE-RO/2022) O valor de um atributo de um dado objeto é uma medida da quantidade daquele atributo, a qual pode ser numérica ou categórica.

Nesse caso, estado civil e sexo são classificados como atributo

a) binário.

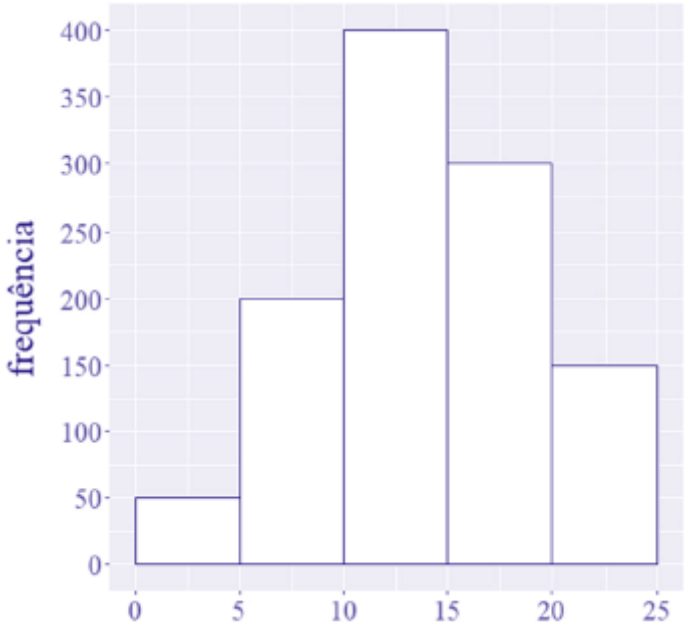
b) nominal.

c) ordinal.

d) ausente.

e) razão.

4. (CESPE/TELEBRAS/2022)

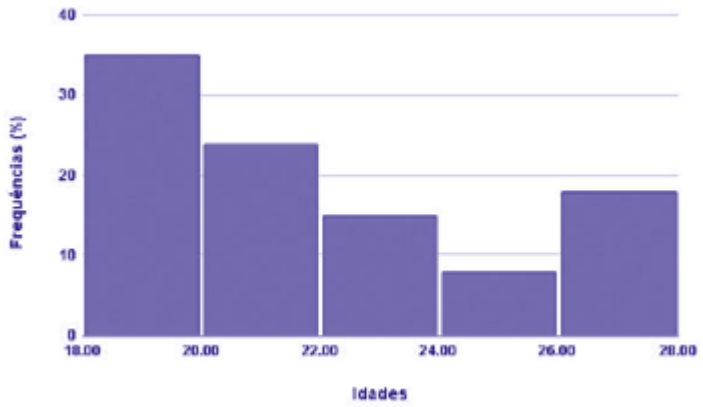


Considerando que o histograma apresentado descreve a distribuição de uma variável quantitativa X por meio de frequências absolutas, julgue o item que se segue.

O número de observações que constituem a variável X é igual a 1.000.

5. (CESPE/SEDUC AL/2021) Com base em estatística, julgue o item a seguir.

Suponha que o histograma a seguir represente a frequência relativa de alunos, distribuída por faixa etária, que ingressaram no ensino superior no estado de Alagoas em 2020. Com base nas informações desse gráfico, é correto afirmar que mais de 50% dos novos alunos têm idade superior a 22 anos.



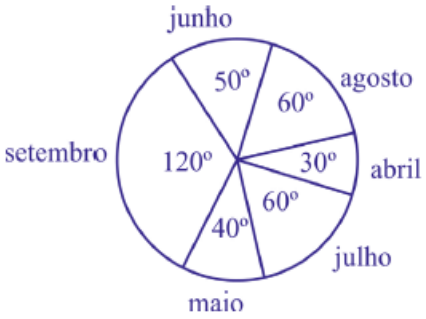
6. (CESPE/IPHAN/2018) Julgue o item subsequente, referente à análise exploratória de dados.

A representação de diagramas de barras, de linha e de pizza possui escala de medida nominal e tem a moda como medida de tendência central.

7. (CESPE/IPHAN/2018). Julgue o item subsequente, referente à análise exploratória de dados.

O gráfico de barras é adequado para a análise de variáveis qualitativas ordinais ou quantitativas discretas, pois permite investigar a presença de tendência nos dados.

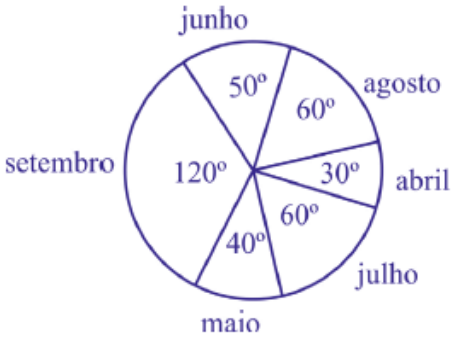
8. (CESPE/CBM-AL/2017) O gráfico de setores a seguir mostra a distribuição das quantidades de incêndios em determinada região, nos meses de abril a setembro de determinado ano.



Sabendo-se que nesses meses ocorreram 1.548 incêndios nessa região, julgue o item que se segue.

A frequência relativa à classe “incêndios no mês de setembro” é superior a 30%.

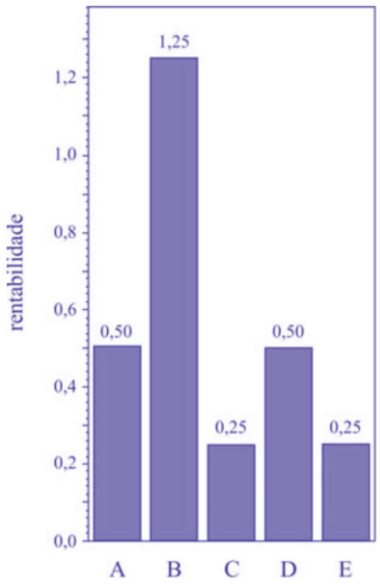
9. (CESPE/CBM-AL/2017) O gráfico de setores a seguir mostra a distribuição das quantidades de incêndios em determinada região, nos meses de abril a setembro de determinado ano.



Sabendo-se que nesses meses ocorreram 1.548 incêndios nessa região, julgue o item que se segue.

Nos meses de maio e junho ocorreram mais de 400 incêndios nessa região.

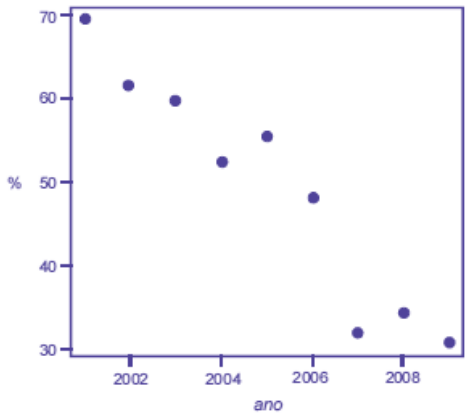
10. (CESPE/FUNPRESP/2016)



O gráfico ilustra cinco possibilidades de fundos de investimento com suas respectivas rentabilidades. Considerando que as probabilidades de investimento para os fundos A, B, C e D sejam, respectivamente, P(A) = 0,182; P(B) = 0,454; P(C)= 0,091; e P(D)= 0,182, julgue o item subsequente.

O gráfico apresentado é um histograma.

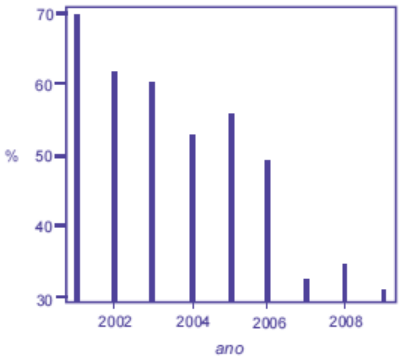
11. (CESPE/DEPEN/2015)



Dado que a participação dos presidiários em cursos de qualificação profissional é um aspecto importante para a reintegração do egresso do sistema prisional à sociedade, foram realizados levantamentos estatísticos, nos anos de 2001 a 2009, a respeito do valor da educação e do trabalho em ambientes prisionais. Cada um desses levantamentos, cujos resultados são apresentados no gráfico, produziu uma estimativa anual do percentual P de indivíduos que participaram de um curso de qualificação profissional de curta duração, mas que não receberam o diploma por motivos diversos. Em 2001, 69,4% dos presidiários que participaram de um curso de qualificação profissional não receberam o diploma. No ano seguinte, 2002, esse percentual foi reduzido para 61,5%, caindo, em 2009, para 30,9%.

A partir das informações e do gráfico apresentados, julgue o item que se segue.

Se os percentuais forem representados por barras verticais, conforme o gráfico a seguir, então o resultado será denominado histograma.



12. (CESPE/DEPEN/2015)

Região	Quantidade de detentos no sistema penitenciário brasileiro (mil pessoas)	Déficit de vagas no sistema penitenciário (mil vagas)	População brasileira (milhões de habitantes)
Norte	37	13	17

Centro-Oeste	51	24	15
Nordeste	94	42	55
Sudeste	306	120	85
Sul	67	16	28
Total	555	215	200

Ministério da Justiça — Departamento Penitenciário Nacional — Sistema Integrado de Informações Penitenciárias – InfoPen, Relatório Estatístico Sintético do Sistema Prisional Brasileiro, dez./2013

Internet: <www.justica.gov.br> (com adaptações).

A tabela mostrada apresenta a quantidade de detentos no sistema penitenciário brasileiro por região em 2013. Nesse ano, o déficit relativo de vagas — que se define pela razão entre o déficit de vagas no sistema penitenciário e a quantidade de detentos no sistema penitenciário — registrado em todo o Brasil foi superior a 38,7%, e, na média nacional, havia 277,5 detentos por 100 mil habitantes.

Com base nessas informações e na tabela apresentada, julgue o item a seguir.

Na análise exploratória, o histograma é um gráfico adequado para descrever a distribuição da quantidade de detentos por região em 2013.

13. (CESPE/BACEN/2013)

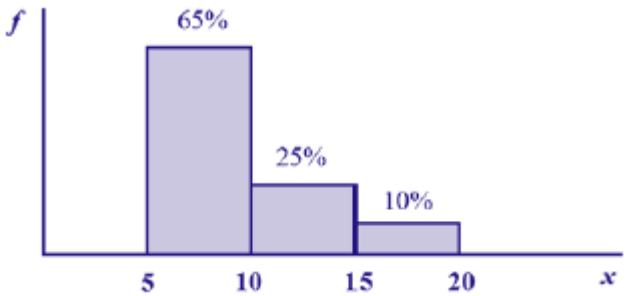
2 4 8 4 8 1 2 32 12 1 5 7 5 5 3 4 24 19 4 14

Os dados mostrados acima representam uma amostra, em minutos, do tempo utilizado na armazenagem de formulários no almoxarifado central de certa instituição por diversos funcionários.

Com base nesses dados, julgue o próximo item.

É inviável a elaboração de um histograma em decorrência do fato de ser este um conjunto de dados quantitativos discretos; dessa forma, apenas por meio de um gráfico de barras pode ser realizada a representação gráfica.

14. (CESPE/STF/2013)



Com referência à figura acima, que mostra a distribuição da renda mensal — x, em quantidades de salários mínimos (sm) — das pessoas que residem em determinada região, julgue o item subsequente.

A variável x, por possuir quatro níveis de respostas, é do tipo qualitativa ordinal.

GABARITO – CEBRASPE

Análise Exploratória de Dados

1. ERRADO
2. ERRADO
3. LETRA B
4. ERRADO
5. ERRADO
6. CERTO
7. CERTO
8. CERTO
9. ERRADO
10. ERRADO
11. ERRADO
12. ERRADO
13. ERRADO
14. ERRADO

ESSA LEI TODO MUNDO CONHECE:
PIRATARIA É CRIME.

Mas é sempre bom revisar o porquê e como você pode ser prejudicado com essa prática.

1. 

Professor investe seu tempo para elaborar os cursos e o site os coloca à venda.

2. 

Pirata divulga ilicitamente (grupos de rateio), utilizando-se do anonimato, nomes falsos ou laranjas (geralmente o pirata se anuncia como formador de "grupos solidários" de rateio que não visam lucro).

3. 

Pirata cria alunos fake praticando falsidade ideológica, comprando cursos do site em nome de pessoas aleatórias (usando nome, CPF, endereço e telefone de terceiros sem autorização).

4. 

Pirata compra, muitas vezes, clonando cartões de crédito (por vezes o sistema anti-fraude não consegue identificar o golpe a tempo).

5. 

Pirata fere os Termos de Uso, adultera as aulas e retira a identificação dos arquivos PDF (justamente porque a atividade é ilegal e ele não quer que seus fakes sejam identificados).

6. 

Pirata revende as aulas protegidas por direitos autorais, praticando concorrência desleal e em flagrante desrespeito à Lei de Direitos Autorais (Lei 9.610/98).

7. 

Concurseiro(a) desinformado participa de rateio, achando que nada disso está acontecendo e esperando se tornar servidor público para exigir o cumprimento das leis.

8. 

O professor que elaborou o curso não ganha nada, o site não recebe nada, e a pessoa que praticou todos os ilícitos anteriores (pirata) fica com o lucro.

Deixando de lado esse mar de sujeira, aproveitamos para agradecer a todos que adquirem os cursos honestamente e permitem que o site continue existindo.