PROBABILIDADE E ESTATÍSTICA

REPRESENTAÇÃO TABULAR E GRÁFICA

CONCEITOS

A **estatística** é a parte da matemática que se dedica à análise, apresentação e interpretação de dados coletados. Esses dados são coletados dentro de uma **população**, que é o conjunto total dos elementos a serem estudados (podem ser pessoas, objetos etc.). Dessa população, podemos coletar os dados de duas maneiras:

- **Censo:** quando são coletados os dados de toda a população; e
- **Mostra**: é um subconjunto da população, da qual são coletados dados para, posteriormente, fazer uma inferência sobre a população (inferência estatística).

Ex.: na eleição, todas as pessoas aptas a votar são a população. Quando uma empresa é contratada para fazer uma pesquisa de intenção de voto, eles selecionam uma amostra dessa população, fazem as perguntas predeterminadas pela pesquisa, e com os dados das respostas fazem uma inferência de como a população toda irá votar.

Um dos ramos da estatística é a **estatística descritiva**, onde estudaremos 4 tipos de medidas descritivas:

- As **medidas de tendência central** que são medidas que indicam a posição dos dados, como média, mediana, moda e quartis (também chamadas de **medidas de posição**);
- As **medidas de dispersão** que medem o grau de variabilidade dos elementos de um conjunto, como desvio-padrão, variância, amplitude;
- A **assimetria** da curva; e
- O achatamento da curva, chamado de **curtose**.

Importante!

Essas duas últimas (assimetria e curtose) também são conhecidas como **medidas de distribuição.**

Os dados de uma amostra podem ser **qualitativos,** que são aqueles dados não numéricos, como sexo, nacionalidade, avaliação nominal (bom, regular, ruim) etc., ou **quantitativos,** que são dados expressos em números, que podem ser objeto de contagens, medições como altura, peso etc.

Cuidado, não necessariamente dados expressos em números serão quantitativos, eles podem também ser qualitativos, como RG, CPF, CNPJ, CEP, CNAE (classificação nacional de atividades econômicas), geralmente esse tipo de código ou classificação é feito em números. Ex: queremos saber a quantidade de empresas que atuam em cada setor, podemos usar a CNAE, que é um número, para separar a quantidade de mercados, farmácias, postos de combustíveis etc.

Os dados qualitativos podem ser:

- **Ordinais:** são aqueles que podem ser ordenados, como mês, nível de escolaridade, tamanho de roupa (P, M, G) entre outros. Ex: o nível de escolaridade pode ser dividido em ensino fundamental, ensino médio, ensino superior, pós-graduação. Por mais que seja um dado qualitativo, a gente consegue colocar isso em ordem, pois sabemos que primeiro vem o ensino fundamental, depois o ensino médio e assim por diante. Da mesma forma o mês, sabemos que primeiro vem janeiro, depois fevereiro até chegar em dezembro;
- **Nominais:** são aqueles que não podem ser ordenados, como sexo, estado civil entre outros. Ex: podemos dividir os estados civis em casado, união estável, solteiro, viúvo... claramente não temos uma ordem entre essas opções.

Os dados quantitativos podem ser:

- **Discretos:** são aqueles dados que possuem um conjunto finito de valores, como a quantidade de acertos em uma prova de múltipla escolha, a quantidade será apenas números inteiros, 0, 1, 2, 3 e assim por diante, ou
- **Contínuos:** são aqueles que possuem uma escala contínua de valor como tempo, comprimento etc. Ex: vamos considerar a variável altura. Entre os dados 1,70m e 1,71m existe uma infinidade de números.

NORMAS DE APRESENTAÇÃO TABULAR

Modelo de uma Tabela

Para que serve, e como montar uma tabela?

Uma tabela deve ser composta por diversas linhas e colunas, sendo que devemos ter um título (normalmente na primeira linha da tabela), e vários dados organizados nas linhas e colunas seguintes.

Geralmente, na primeira linha, depois do título, teremos as classes que serão retratadas nas linhas, ex.: Estados, Siglas, População, Área etc. Nas linhas seguintes teremos os dados da tabela, onde teremos em uma mesma linha o Estado, relacionando sua sigla, sua população, sua área etc. Vejamos o exemplo dessa tabela citada.

INFORMAÇÕES DOS ESTADOS DA REGIÃO SUDESTE			
ESTADO	SIGLA	POPULAÇÃO	ÁREA (KM²)
Minas Gerais	MG	21.292.666	586.522
Espírito Santo	ES	4.064.052	46.095
Rio de Janeiro	RJ	17.366.189	43.780
São Paulo	SP	45.919.049	248.222

Analisando a tabela, podemos concluir que Minas Gerais (MG) é o maior estado da Região Sudeste, pois tem a maior área, mas que o estado de São Paulo é o mais populoso, por ter uma população maior que os outros.

O importante é olhar uma tabela e entender quais dados podemos extrair com o que está apresentado nela.

As tabelas mais utilizadas na estatística são as tabelas de frequência, conforme apresentamos no item de média aritmética para dados agrupados.

Tipos de Séries Estatísticas

As séries estatísticas são as diversas maneiras de apresentar os dados desejados em forma de tabela, o objetivo das séries estatísticas é organizar os dados observados e mostrá-los de maneira organizada, facilitando sua compreensão.

Temos vários tipos séries estatísticas, mas vamos destacar algumas mais importantes:

- **Séries Temporais:** é um conjunto de observações de uma variável ao longo do tempo, ou seja, uma sequência de dados numéricos em ordem sucessiva. Nesse tipo de série o que varia é o tempo, mas o fato e o local de observação são fixos;
- **Séries Geográficas:** é um conjunto de observações de uma variável em diferentes locais. Nesse tipo de série o que varia é o local (região) da observação, mas o tempo e o fato observado são fixos. Ex.:

POPULAÇÃO DOS ESTADOS DA REGIÃO SUDESTE	
ESTADO	**POPULAÇÃO**
Minas Gerais	21.292.666
Espírito Santo	4.064.052
Rio de Janeiro	17.366.189
São Paulo	45.919.049

- **Séries Específicas:** é um conjunto de observações de uma variável com diferentes categorias (espécies). Nesse tipo de variável o que varia são as categorias observadas, mas o tempo e o local são fixos. Ex.: queremos analisar a quantidade de animais diferentes que habitam uma certa região de proteção florestal. Nesse caso a tabela será classificada pelas espécies observadas na região de interesse em um mesmo intervalo de tempo.

ESPÉCIES QUE HABITAM A REGIÃO EM 2020	
ESPÉCIE	**QUANTIDADE OBSERVADA**
Onça	45
Tamanduá	75
Lobo	107
Anta	90

- **Séries Conjugadas (Mistas):** nesse tipo de séries vamos conjugar dois tipos de séries em uma mesma tabela. Ex.: vamos conjugar a tabela específica acima, com uma série temporal, mostrando a quantidade de cada espécie observada ao longo dos últimos 3 anos.

ESPÉCIES QUE HABITAM A REGIÃO – ÚLTIMOS 3 ANOS			
ESPÉCIE	**QUANTIDADE OBSERVADA**		
	2018	**2019**	**2020**
Onça	30	38	45
Tamanduá	60	65	75
Lobo	30	90	107
Anta	50	80	90

Séries Temporais

De todas as séries, uma das mais importantes é a série temporal, que corresponde a um conjunto de observações de uma variável ao longo do tempo, ou seja, uma sequência de dados numéricos em ordem sucessiva, que geralmente (mas não necessariamente) ocorre em intervalos uniformes. Ex.: uma série que mostra a quantidade de picolés vendidos por uma sorveteria mensalmente ao longo de um ano.

Podemos definir exemplos de séries temporais e de séries não temporais:

SÉRIES TEMPORAIS	SÉRIES NÃO TEMPORAIS
Série diária da temperatura na cidade de São Paulo ao longo de um ano.	Temperaturas de várias cidades em um mesmo dia, ou períodos diferentes.
Quantidade de furtos anuais em Cuiabá.	Quantidade de furtos no ano de 2018 nas diferentes capitais do país.
Salário de um funcionário ao longo do ano.	Salários dos funcionários de uma empresa no mesmo mês.

As séries temporais são importantes para identificar padrões de variáveis no tempo, para que se possa tentar prever possíveis danos no futuro. Para descrever séries temporais, são utilizados modelos de processos estocásticos, que são processos controlados por leis probabilísticas.

Uma série temporal pode ser decomposta em três séries temporais: tendência, sazonalidade e uma componente aleatória (nível).

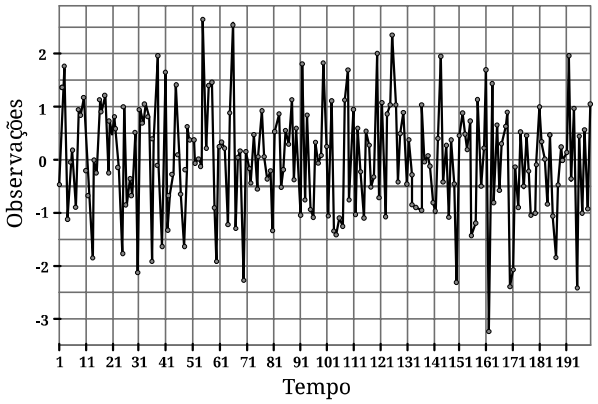
- **Tendência:** é o comportamento de longo prazo na série, podendo ser determinístico (quando os valores da série podem ser descritos por uma função matemática) e podendo ser estocástico (aquele cujo estado é indeterminado, com origem em eventos aleatórios). Ex.: Quando analisamos a população brasileira ao longo dos anos (vamos supor uma série com 100 anos), vemos que a cada ano esse valor aumenta, nem sempre em uma mesma proporção, mas podemos notar uma tendência de crescimento;
- **Sazonalidade:** é um padrão regular que ocorre na série temporal. Uma série temporal é sazonal quando fenômenos que ocorrem durante o tempo se repetem em um mesmo período de tempo, ou seja, ocorrem sempre em uma mesma hora, todos os dias, ou em um mesmo mês, todos os anos. Ex.: um aumento nas vendas de uma loja de roupas em dezembro em todos os anos (período do Natal);

- **Aleatório:** é o que não pode ser explicado pela tendência e sazonalidade, ou seja, é o resíduo, sendo que a aleatoriedade não pode ser determinística (descrito por uma função matemática) e será sempre estocástica;
- **Ciclo:** Longas ondas, mais ou menos regulares, em torno de uma linha de tendência.

Outro ponto importante é a estacionariedade da série temporal.

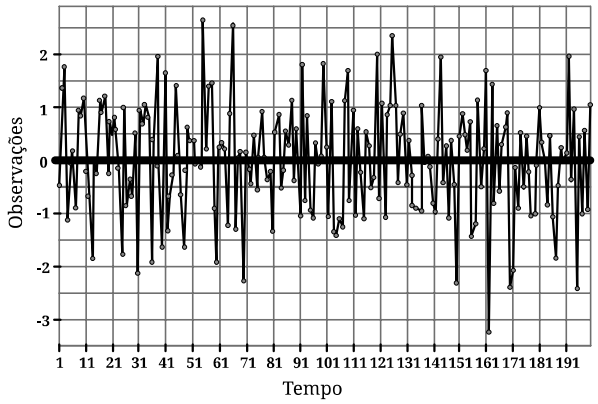
Dizemos que uma série temporal é estacionária quando suas observações no tempo se posicionam aleatoriamente ao redor de uma média constante, transparecendo um equilíbrio estável.

Série Temporal Estacionária e Observações



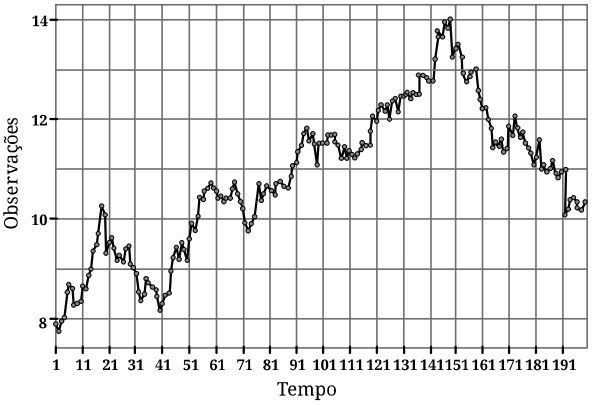
Podemos notar que uma série estacionária se mantém sempre em torno de uma média, que representamos pela linha preta central na figura a seguir.

Série Temporal Estacionária



Uma série pode ser estacionária por um período curto, ou por um período longo, o modelo ARIMA pode descrever séries estacionárias e séries não estacionárias que não apresentam um comportamento totalmente aleatório, ou seja, uma não estacionariedade homogênea, que é quando uma série é estacionária, flutuando ao redor de um nível, por um certo tempo, depois muda de nível e flutua ao redor desse novo nível, e depois muda novamente de nível e assim por diante.

Série Temporal Estacionária e Observações



A seguir, algumas questões comentadas de bancas diversas para que você veja como o assunto estudado já foi abordado em provas:

1. (FCC – 2018) Em séries temporais, as oscilações aproximadamente regulares em torno da tendência

a) são típicas de séries muito curtas, como dados dentro de um mês.
b) dão a direção global dos dados.
c) podem ser decorrentes de fenômenos naturais e socioeconômicos.
d) caracterizam uma série sem variável residual.
e) determinam o componente não sistemático.

Vamos analisar cada alternativa, mas iremos ver que 4 são totalmente incorretas, e uma delas é um pouco vaga, mas, mesmo assim correta, é mais fácil chegar na resposta por eliminação:
Alternativa a) Essas oscilações em torno da tendência são características de séries temporais, e acontecem tanto para séries longas quanto curtas. Alternativa incorreta.
Alternativa b) Essa direção global é exatamente a tendência, e não as oscilações em torno dela. Alternativa incorreta.
Alternativa c) Exatamente, a tendência é o comportamento a longo prazo, entretanto os valores não serão exatamente os considerados em uma certa função matemática, tendo sempre uma pequena variação, que são essas oscilações em torno da tendência. Essas variações podem ocorrer por inúmeros fatores, dentre eles fenômenos naturais e socioeconômicos. Alternativa correta.
Alternativa d) Essas oscilações são caracterizadas pelos resíduos, ou seja, variáveis residuais que fazem com que os valores observados sejam próximo, mas não idênticos aos valores previstos, portanto essas oscilações. Alternativa incorreta.
Alternativa e) Essas oscilações são naturais das observações, o comportamento sistemático de uma observação se caracteriza pelo erro sistemático de observações, por exemplo, quando anotamos o horário de certas observações, mas o relógio utilizado está com 1 hora de atraso. Alternativa incorreta. Resposta: Letra C.

2. (IBADE – 2017) Considerando uma série temporal, é correto afirmar que a tendência indica:

a) comportamento sazonal a curto prazo.
b) ciclos de altas e quedas periódicas de valores a curto prazo.

c) comportamento independente dos dados a longo, curto e médio prazo.
d) comportamento a longo prazo.
e) somente se há um outlier conhecido como ponto influente.

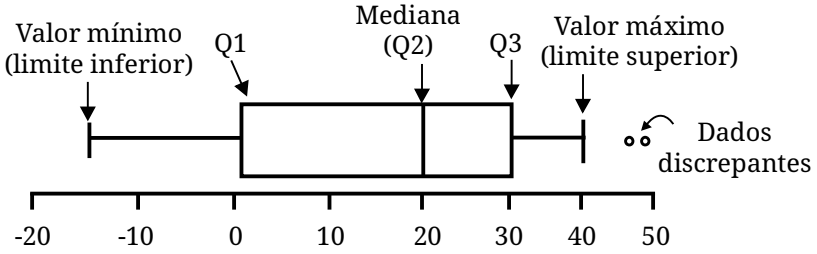
Por definição sabemos que a tendência caracteriza o comportamento a longo prazo. Resposta: Letra D.

GRÁFICOS E DIAGRAMAS

Para representar os dados coletados existem vários tipos de gráficos usados na estatística. Muitos deles são conhecidos e sempre aparecem em reportagens, jornais etc. Vamos mostrar alguns dos principais tipos de gráficos:

Caixa *Box-plot*

A caixa *box-plot* (muito cobrado) é um gráfico que nos mostra a distribuição de frequência, e é formado pelos quartis e pelos valores extremos.



Vemos que a caixa é formada pelos 3 quartis, que no caso seriam aproximadamente: Q1 = 0, Q2 = 20 e Q3 = 30. Os dados extremos seriam aproximadamente -15 e 40, portanto amplitude seria: 40 – (–15) = 40 + 15 = 55.

As duas bolinhas significam dados discrepantes (*outliers*), que seriam dados que fogem do padrão do restante dos dados. Os dados discrepantes são aqueles que superam em 1,5 vezes o intervalo interquartílico (diferença entre Q3 e Q1).

No nosso exemplo:

$$Q3 - Q1 = 30 - 0 = 30$$
$$1{,}5 \cdot 30 = 45$$

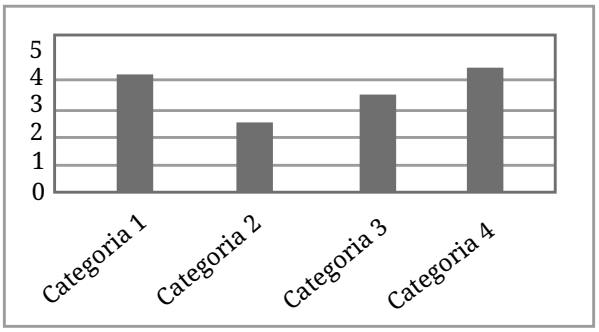
Os dados discrepantes serão aqueles que forem inferiores a 45 unidades de Q1, ou superiores a 45 unidades de Q3.

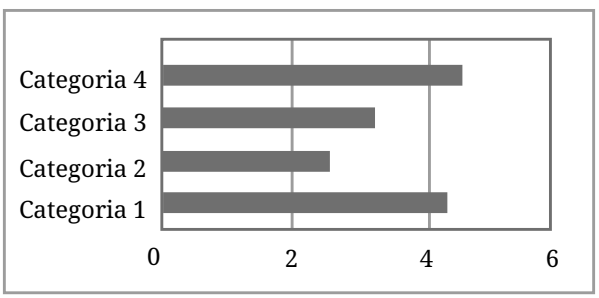
Limite inferior: 0 – 45 = - 45
Limite superior: 30 + 45 = 75

Portanto, serão dados discrepantes os que tiverem valores inferiores a – 45 ou os superiores a 75.

Gráfico de Barras e Colunas

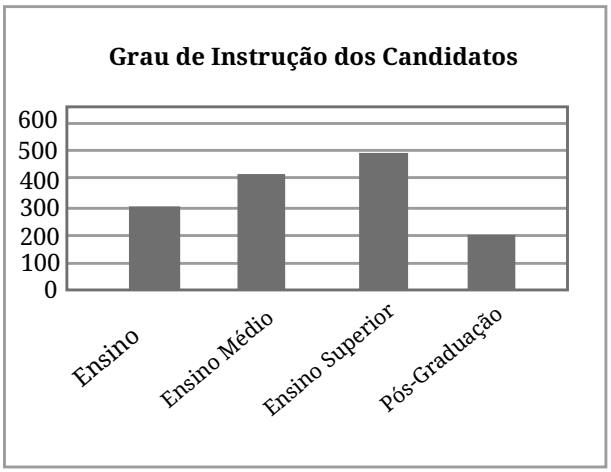
Esses dois tipos de gráficos na verdade são basicamente os mesmos, a diferença é que no gráfico de barras, as barras/colunas são horizontais e no gráfico de colunas, as barras/colunas são verticais. Normalmente esses gráficos são usados para representar dados qualitativos ordinais e quantitativos discretos.





No gráfico de colunas o eixo horizontal traz os dados qualitativos, ou quantitativos discretos, e o eixo vertical traz as frequências (quantidades) de cada categoria. Já no gráfico de barras o eixo vertical traz os dados qualitativos ou quantitativos discretos, e o eixo horizontal traz as frequências de cada categoria.

Vamos supor um gráfico de colunas, no qual queremos saber a quantidade de pessoas com diferentes graus de escolaridade em um determinado concurso.



Nesse tipo de gráfico vemos que não é possível falar em média ou mediana, mas podemos usar como medida de tendência central a moda, que nesse caso seria o Ensino Superior, que é o grau de instrução que mais se repete no gráfico.

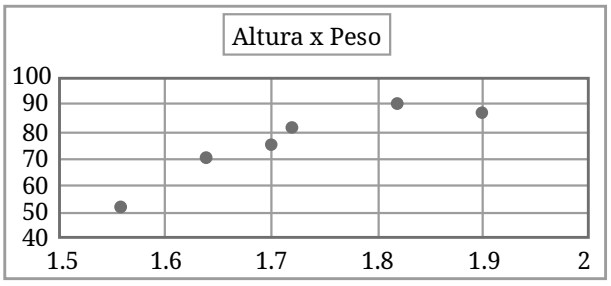
Gráfico de Dispersão

Estudaremos melhor esse tipo de gráfico quando falarmos de correlação, mas, irei apresentar como ele é feito, e quando é usado.

O gráfico de dispersão, ou diagrama de dispersão, é uma associação entre pares de dados quantitativos, normalmente são usados para entender a correlação entre duas variáveis. Nele, vamos verificar as duas variáveis e colocar esses pontos em um plano cartesiano.

Vamos fazer um gráfico de dispersão com as variáveis peso e altura de um grupo de 6 pessoas:

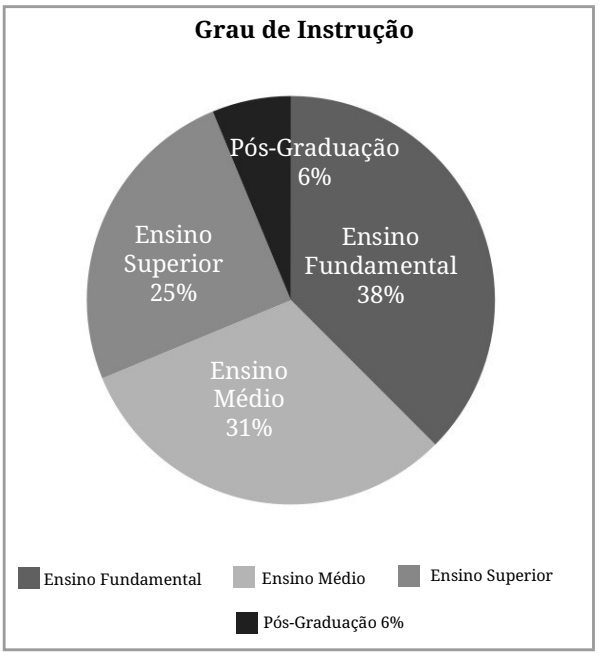
ALTURA	1,7	1,72	1,9	1,56	1,64	1,82
PESO	75	81	87	52	70	90



Podemos ver que cada ponto representa o par peso/altura de um indivíduo, mas discutiremos isso mais profundamente posteriormente.

Gráfico de Setores (Pizza)

O gráfico de setores normalmente é usado para apresentar dados qualitativos como setores de um círculo (pedaços de uma pizza). Normalmente os dados são apresentados em percentuais, sendo que o total é 100% (360º).

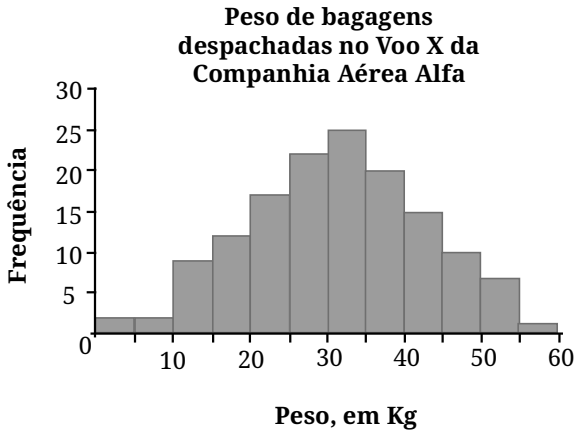


Desse gráfico, podemos ver que, na pesquisa, o grau de instrução predominante é o Ensino Fundamental, com 38%, e o mais raro é a Pós-Graduação, com 6%. Também é um gráfico em que temos a moda como medida de tendência central.

Histograma

O histograma é um dos principais gráficos da estatística, ele é muito utilizado para demonstrar a distribuição de frequência de dados quantitativos contínuos.

O histograma é formado por colunas ou barras (retângulos) de um conjunto de dados e dividido em classes uniformes ou não uniformes. Sendo que a base do retângulo representa uma classe e a altura do retângulo representa a quantidade ou a frequência absoluta com que o valor da classe ocorre no conjunto de dados.



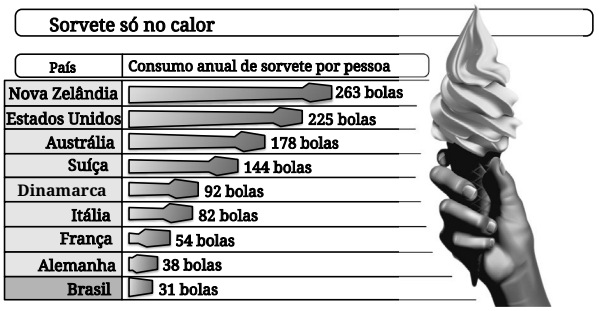
A tabela nos mostra que a maior parte das bagagens possui entre 30 e 35 kg, mas existem malas desde 0 a 60 kg.

Gráfico Pictórico

O gráfico pictórico não é propriamente um tipo específico de gráfico. Na verdade, o gráfico pictórico pode ser representado por um gráfico de barras, de

colunas, de linhas, ente outros. A grande característica desse gráfico é que, para chamar mais a atenção, são usadas figuras (imagens) na composição do gráfico.

Esse tipo de gráfico é muito usado em jornais e telejornais, pois proporciona uma comunicação mais rápida e com precisão de entendimento. Nesse tipo de gráfico as figuras são, ao mesmo tempo, os dados estatísticos e indicam a proporcionalidade desses dados.



Podemos notar que esse gráfico nada mais é que um gráfico de barras, mas, pelo fato de o gráfico mostrar a quantidade de consumo anual de sorvete em diferentes países, ao invés de se colocar as barras, foi colocada uma espátula de tomar sorvete, em um tamanho proporcional ao da barra, e ao lado do gráfico foi colocada uma imagem de sorvetes.

MEDIDAS DE TENDÊNCIA CENTRAL (MÉDIA, MEDIANA, MODA, MEDIDAS DE POSIÇÃO, MÍNIMO E MÁXIMO)

MÉDIA

Existem 3 tipos de média: a mais cobrada é a **média aritmética (simples ou ponderada)**, mas temos também a **média geométrica** e a **média harmônica**.

Como veremos adiante, a média é o **primeiro momento** de uma distribuição.

Temos duas formas de apresentação de dados para que seja calculada a média, a mais comum é a apresentação de vários **dados não agrupados**, onde são listados vários valores. Por exemplo, em uma prova a lista de notas de todos os alunos separadamente são dados não agrupados. A outra forma de apresentação é através de **dados agrupados**, nesse caso, vamos pensar no mesmo exemplo, mas, ao invés de termos todas as notas diretamente, receberemos uma lista dizendo que, 5 pessoas tiraram de 0 a 2, outras 7 tiraram de 2 a 4, mais 15 tiraram de 4 a 6, outras 10 de 6 a 8 e os 3 restantes tiraram de 8 a 10, normalmente isso é apresentado em uma tabela.

NOTA	QUANTIDADE DE ALUNOS
0 – 2	5
2 – 4	7
4 – 6	15
6 – 8	10
8 – 10	3

Quando estamos tratando de **média** de dados de uma **população** representamos pela letra grega **μ** (mi), já quando estamos tratando de **média** de dados de uma **amostra,** representamos por $\bar{x}$.

Média Aritmética Simples

A **média aritmética simples** é a que estamos mais acostumados no dia a dia, ela é dada pela soma dos valores dos dados que queremos saber, dividido pela quantidade desses dados.

$$\text{Média} = \frac{\text{Soma}}{\text{Quantidade}}$$

Soma = Média x Quantidade

Em linguagem matemática isso é dado por:

$\bar{x} = \frac{\sum xi}{N}$ para dados de uma **amostra**.

Ou:

$\mu = \frac{\sum xi}{N}$ para dados de uma **população**.

Onde: $\sum Xi$ é o somatório dos dados X_1, X_2, X_3...X_N, e N é a quantidade de dados da amostra/população. Cada dado é representado por Xi.

Vamos supor que, na faculdade, teremos 4 provas da disciplina de Estatística. Para calcular a média final basta somar as 4 notas e dividir por 4.

Supondo que as notas tenham sido na ordem: 6,0; 7,0; 5,0; 8,0.

$$\text{Média} = \frac{\text{Soma}}{\text{Quantidade}}$$

$$\text{Média} = \frac{6 + 7 + 5 + 8}{4}$$

$$\text{Média} = \frac{26}{4}$$

Média = 6,5

Ex: Em uma faculdade a quantidade de alunos matriculados em cada curso está apresentada na tabela a seguir:

CURSO	QUANTIDADE DE ALUNOS
Direito	55
Contabilidade	24
Estatística	35
Física	?

A média do número de alunos matriculados por curso é 38,5. Nesse caso, qual a quantidade de alunos matriculados em Física?

A média de alunos matriculados por curso é dada pela soma dos alunos de cada curso dividido pela quantidade de cursos, onde a média foi dada, e é 38,5, e o total de cursos é 4.

Sabemos que a fórmula da média é:

$$\text{Média} = \frac{\text{Soma}}{\text{Quantidade}}$$

$$38{,}5 = \frac{\text{Soma}}{4}$$

$\text{Soma} = 38{,}5 \cdot 4$

$\text{Soma} = 154$

A soma dos alunos matriculados é 154, e vamos chamar o número de alunos matriculados em Física de X.

$$55 + 24 + 35 + X = 154$$
$$114 + X = 154$$
$$X = 154 - 114$$
$$X = 40$$

Portanto, são 40 alunos matriculados em Física.

Média Aritmética Ponderada

A **média ponderada** é muito parecida com a média simples, mas nela são colocados pesos diferentes para alguns dados. Essa média é muito utilizada em provas.

Vamos pensar no mesmo exemplo da média simples. Supondo que, na faculdade, teremos 4 provas de Estatística, mas a prova 3 tem peso 2 e a prova 4 tem peso 3. Para calcular a média ponderada temos que somar todas as notas, mas, as notas que têm pesos devem ser somadas conforme seu peso, ou seja, se for peso 2 temos que multiplicar essa nota por 2, se for peso 3 temos que multiplicar essa nota por 3. E na quantidade temos que considerar o peso também.

Supondo que as notas tenham sido na ordem: 6,0; 7,0; 5,0 (peso 2); 8,0 (peso 3). Nesse caso, a quantidade de notas que vamos considerar deve ser 7, pois as provas 1 e 2 têm peso 1, a prova 3 é peso 2, e a prova 4 é peso 3. Portanto: 1 + 1 + 2 + 3 = 7.

$$\text{Média} = \frac{\text{Soma}}{\text{Quantidade}}$$

$$\text{Média} = \frac{6 + 7 + 2 \cdot 5 + 8 \cdot 3}{7}$$

$$\text{Média} = \frac{6 + 7 + 10 + 24}{7}$$

$$\text{Média} = \frac{47}{7}$$

$\text{Média} = 6{,}71$

Média Aritmética para Dados Agrupados

Podemos ter duas formas de apresentação de dados, uma que nos traz o dado e a frequência que esse dado ocorre, e outra que dá um intervalo (que chamamos de classe) e a frequência de dados dessa classe, ambas apresentadas no formato de tabela.

Vamos mostrar um exemplo para cada tipo de apresentação de dados:

- Ex: Em um evento foram compradas 4 marcas diferentes de água, conforme a tabela.

TIPOS	QUANTIDADE DE GARRAFAS	PREÇO UNITÁRIO
A	5	R$ 3,80
B	8	R$ 6,00
C	15	?
D	6	R$ 5,00

O preço médio do total de garrafas compradas foi de R$ 5,50. Qual o valor unitário da garrafa da marca C?

Para achar a média, temos que somar o valor de todas as garrafas, sendo que para cada marca foi comprada uma quantidade diferente de garrafas. Para resolver esse tipo de questão vamos incluir mais uma coluna na tabela, que é a multiplicação do valor unitário com a quantidade de garrafas.

TIPOS	QUANTIDADE DE GARRAFAS	PREÇO UNITÁRIO	Q X P
A	5	R$ 3,80	$5 \cdot 3{,}8 = 19$
B	8	R$ 6,00	$8 \cdot 6 = 48$
C	15	?	$15 \cdot X$
D	6	R$ 5,00	$6 \,.\, 5 = 30$

$$\text{Média} = \frac{19 + 48 + 15 \cdot x + 30}{5 + 8 + 15 + 6}$$

Sabemos que a média dos valores foi R$ 5,50, portanto:

$$5{,}5 = \frac{97 + 15 \cdot x}{34}$$

$5{,}5 \cdot 34 = 97 + 15 \cdot X$

$187 = 97 + 15 \cdot X$

$15 \cdot X = 187 - 97$

$15 \cdot X = 90$

$$X = \frac{90}{15}$$

$X = 6$

Logo o Preço Unitário da Marca C é R$ 6,00.

Agora vamos ver um exemplo em que os dados são dados em classes. Vamos pensar no exemplo que demos anteriormente, das notas dos alunos, sendo que 5 pessoas tiraram de 0 a 2, outras 7 tiraram de 2 a 4, mais 15 tiraram de 4 a 6, outras 10 de 6 a 8 e os 3 restantes tiraram de 8 a 10.

NOTA	QUANTIDADE DE ALUNOS
0 |– 2	5
2 |– 4	7
4 |– 6	15
6 |– 8	10
8 |– 10	3

A barra (|) normalmente mostra que o dado limite entre as classes pertence àquela classe, ou seja, a nota 2 estaria no limite entre a 1ª e a 2ª classe, nesse caso por conta da barra vamos considerar que a nota 2 pertencente à 2ª classe.

Para achar a média nesse caso vamos fazer praticamente a mesma coisa do exemplo anterior, mas o valor da classe que iremos usar é o ponto médio de cada classe, ou seja, basta fazer a média da classe.

Cada classe tem uma variação de 2 pontos, portanto a média de cada classe será:

$$\text{Classe 1: } \frac{0 + 2}{2} = 1$$

Classe 2: $\frac{2+4}{2} = 3$

Classe 3: $\frac{4+6}{2} = 5$

Classe 4: $\frac{6+8}{2} = 7$

Classe 5: $\frac{8+10}{2} = 9$

NOTA	QUANTIDADE DE ALUNOS	PONTO MÉDIO DA CLASSE
0 |– 2	5	1
2 |– 4	7	3
4 |– 6	15	5
6 |– 8	10	7
8 |– 10	3	9

Agora basta fazer a coluna da multiplicação da frequência (quantidade de alunos) pelo ponto médio da classe.

NOTA	QUANTIDADE DE ALUNOS	PONTO MÉDIO DA CLASSE	QUANT. X PONTO MÉDIO
0 |– 2	5	1	5 x 1 = 5
2 |– 4	7	3	7 x 3 = 21
4 |– 6	15	5	15 x 5 = 75
6 |– 8	10	7	10 x 7 = 70
8 |– 10	3	9	3 x 9 = 27

Como temos toda a sala, estamos falando de população, e a média é dada por μ.

$$\mu = \frac{5 + 21 + 75 + 70 + 27}{5 + 7 + 15 + 10 + 3}$$

$$\mu = \frac{198}{40}$$

$\mu = 4{,}95$

Portanto, a nota média da classe foi 4,95.

A frequência apresentada em um exercício pode ser absoluta ou relativa. A **frequência absoluta** é a quantidade total, no nosso exemplo é a quantidade de alunos que tiraram tais notas. A **frequência relativa** é a razão ou percentual entre a quantidade daquele dado e o total, portanto a soma da frequência relativa de todas as classes resulta sempre em 1.

NOTA	QUANTIDADE DE ALUNOS	FREQ. RELATIVA
0 |– 2	5	5÷40 = 0,125 = 12,5%
2 |– 4	7	7÷40 = 0,175 = 17,5%
4 |– 6	15	15÷40 = 0,375 = 37,5%
6 |– 8	10	10÷40 = 0,250 = 25%
8 |– 10	3	3÷40 = 0,075 = 7,5%
TOTAL	40	1 = 100%

Média Geométrica

A média geométrica, μ_G, é dada por:

$$\mu_G = \sqrt[N]{X_1 \cdot X_2 \cdot X_3 \ldots\ldots X_N}$$

Fazendo um paralelo com a média aritmética, na média geométrica ao invés de somarmos os dados, vamos multiplicar, e ao invés de dividirmos pela quantidade de dados observados (N), vamos fazer a raiz enésima dessa quantidade.

Média Harmônica

A média harmônica, μ_H, é dada por:

$$\mu_H = \frac{N}{\sum \frac{1}{X1}}$$, ou seja:

$$\mu_H = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3} + \ldots + \frac{1}{X_N}}$$

Fazendo um paralelo com a média aritmética, na média harmônica vamos inverter os termos, a quantidade de dados observados (N) fica no numerador (em cima) e a soma fica no denominador (embaixo). A diferença é que a soma não é dos dados propriamente, e sim, do inverso dos dados, ou seja, um sobre o valor observado.

Dificilmente vemos uma questão cobrar o cálculo das médias geométrica e harmônica, mas, uma propriedade que é mais cobrada é a comparação das médias aritmética, geométrica e harmônica. Precisamos saber que a média aritmética é a maior, depois a média geométrica e, por fim, a média harmônica que é a menor.

$$\mu_A > \mu_G > \mu_H$$

Hora de treinar o que aprendemos na teoria com exercícios comentados de diversas bancas. Vamos lá!

1. **(CESPE-CEBRASPE – 2018-ADAPTADA)** A tabela a seguir mostra a distribuição das idades dos 30 alunos de uma sala de aula.

Idade (em anos)	10	11	12	13	14
Número de alunos	14	8	3	4	1

Nesse caso, a média de idade dos alunos dessa sala é igual a 11 anos.

( ) CERTO ( ) ERRADO

Na primeira linha temos a idade e na segunda temos a quantidade de alunos com cada idade (frequência). Veja que normalmente fazemos esse tipo de questão com os dados dispostos em colunas, e aqui estão em linhas, portanto basta acrescentar uma linha da freq. x idade.

IDADE (EM ANOS)	NÚMERO DE ALUNOS	IDADE X FREQ
10	14	10x14=140
11	8	11x8=88
12	3	12x3=36
13	4	13x4=52
14	1	14x1=14

$$Média = \frac{140 + 88 + 36 + 52 + 14}{14 + 8 + 3 + 4 + 1}$$

$$Média = \frac{330}{30}$$

Média = 11 anos
Resposta: Certo.

2. (CESPE-CEBRASPE – 2018) 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 a seguir, que apresenta os valores observados da variável X em uma amostra aleatória de 5 dias de apreensões no citado aeroporto, julgue o item 2.

	DIA				
X (QUANTIDADE DIÁRIA DE DROGAS APREENDIDAS, EM KG)	1	2	3	4	5
	10	22	18	22	28

A mediana das quantidades X observadas na amostra em questão foi igual a 18 kg.

( ) CERTO ( ) ERRADO

A mediana é a medida central da nossa amostra, entretanto, para achar esse valor os dados precisam estar em ordem crescente. Vamos ordenar nossos dados para poder achar a mediana, sendo que o total de dados são 5.
10, 18, 22, 22, 28
A mediana será o 3º elemento, ou seja, 22. Resposta: Errado.

MODA, MEDIANA E QUARTIS

Moda

Quando dizemos que uma roupa está na moda, isso quer dizer que muita gente está usando esse tipo de roupa. A **moda** (Mo) na estatística é exatamente isso, é aquele dado que mais se repete na população ou na amostra, ou seja, o dado com **maior frequência**.

Ex: Dado os valores observados: 3, 4, 5, 6, 6, 8, 8, 8, 9, 10. A moda é o 8 pois aparece 3 vezes.

Para dados não agrupados é tranquilo de achar a moda, mas quando temos dados agrupados com intervalos de classes não sabemos exatamente qual valor dentro daquela classe mais se repete, para isso temos 3 métodos para achar a moda: moda bruta, moda de Pearson e moda de Czuber.

A **moda bruta** nada mais é que o valor médio da classe modal.

Ex:

NOTA	QUANTIDADE DE ALUNOS
0 |– 2	5
2 |– 4	7
4 |– 6	15
6 |– 8	10
8 |- 10	3

A classe modal é a que tem a maior frequência, ou seja, a que possui 15 alunos, que é a classe que vai de 4 a 6. Assim, pelo método da moda bruta, a moda será a média da classe, ou seja, a média entre os limites da classe (4 e 6):

$$Mo = \frac{4 + 6}{2}$$

$$Mo = \frac{10}{2}$$

$$Mo = 5$$

A **moda de Pearson** é dada pela fórmula:

$$Mo = 3 \cdot Md - 2 \cdot \overline{x}$$

Portanto, para achar a moda pelo método de Pearson, é preciso achar a mediana (Md) e a média aritmética ($\overline{x}$).

A **moda de Czuber** é dada pela Fórmula de Czuber:

$$Mo = L + h \cdot \frac{d_1}{d_1 + d_2}$$

Onde: L é o limite inferior da classe modal (classe com maior frequência);

h é a amplitude da classe (diferença entre os extremos da classe);

d1 é a diferença entre a frequência da classe modal com a classe anterior;

d2 é a diferença entre a frequência da classe modal com a classe posterior.

Veja que na fórmula de Czuber a classe anterior é mais importante que a classe posterior à classe modal, pois d1 aparece embaixo e em cima, e o L é o limite inferior da classe modal. Vamos entender com um exemplo:

Dada a distribuição de frequência.

NOTA	QUANTIDADE DE ALUNOS
0 |– 2	5
2 |– 4	7
4 |– 6	15
6 |– 8	10
8 |- 10	3

Para achar a moda vamos usar a Fórmula de Czuber. A classe modal é a classe com maior frequência, portanto das notas entre 4 e 6, que tem 15 alunos.

Portanto:

L = 4 (limite inferior)
h = 2 (intervalo da classe: 6 – 4 = 2)
d1 = 8 (diferença de frequência entre a classe modal e a classe anterior: 15-7 = 8)
d2 = 5 (diferença de frequência entre a classe modal e a classe posterior: 15-10 = 5)

$$Mo = L + h \cdot \frac{d_1}{d_1 + d_2}$$

$$Mo = 4 + 2 \cdot \frac{8}{8 + 5}$$

$$Mo = 4 + 2 \cdot \frac{8}{13}$$

$$Mo = 4 + \frac{16}{13}$$

Mo = 4 + 1,23
Mo = 5,23

Mediana

A **mediana** (Md) é o valor que divide os dados em duas partes iguais, ou seja, ao relacionar os dados em ordem crescente é o dado que fica na posição central.

Ex: Dado os valores observados: 3, 4, 5, 6, 6, 8, 8, 8, 9, 10, 14.

Temos 11 dados nessa amostra, portanto o 6º valor é o valor central, pois, temos 5 valores antes dele e 5 depois, portanto a mediana é 8.

3, 4, 5, 6, 6, **8**, 8, 8, 9, 10, 14

Quando temos número ímpar de observações, a mediana será o próprio valor, mas quando tivermos uma quantidade **par** de dados, não teremos um valor central, nesse caso vamos fazer a média entre os dois valores centrais.

Vamos pensar nos seguintes dados: 3, 4, 5, 6, 6, 8, 8, 8, 9, 10

Aqui temos 10 dados, e para dividirmos nossa amostra em duas partes iguais vamos separar entre o 6 e o 8, restando 5 dados para cada lado. Portanto, a mediana será a média entre 6 e 8.

3, 4, 5, 6, 6, / 8, 8, 8, 9, 10

$$Md = \frac{6 + 8}{2} = \frac{14}{2} = 7$$

Para dados não agrupados é tranquilo de achar a mediana, mas quando temos dados agrupados com intervalos de classes não sabemos exatamente qual valor dentro daquela classe será o central, para isso vamos fazer uma regra de três para achar esse valor, esse método é chamado de **interpolação linear**.

Dada a distribuição de frequência:

NOTA	QUANTIDADE DE ALUNOS
0 |– 2	5
2 |– 4	7
4 |– 6	15
6 |– 8	10
8 |- 10	3

O total de observações nessa população é 40. Portanto, a mediana estará entre o 20 e o 21, mas para fazer a conta vamos usar o termo inferior, portanto o 20º.

A mediana estará na terceira classe, pois até a segunda temos 12 alunos (5 + 7), e ao final da terceira já teremos 27 alunos, portanto o 20º termo estará dentro da 3ª classe (4 a 6).

Como disse anteriormente, até a classe anterior temos 12 alunos, para chegar a nossa mediana, que é o 20º termo, faltam 8 alunos. Portanto, vamos fazer uma regra de 3, onde a classe mediana tem 15 alunos em uma variação de 2 pontos, portanto para 8 alunos a variação de pontos será X.

Frequência (alunos)	Variação dentro da classe
15	2
8	X

15 · X = 2 · 8
15 · X = 16

$$X = \frac{16}{15}$$

X = 1,06

Portanto, o 8º termo da 3ª classe (que é a nossa mediana) está 1,06 acima do limite inferior da classe mediana, que é 4.

Md = 4 + 1,06 = 5,06

Quartis

Assim como a mediana divide os dados em duas partes iguais, os quartis dividem em 4 partes iguais. Portanto, teremos 3 quartis (Q1, Q2 e Q3).

O primeiro quartil (Q1) divide 25% dos dados antes dele e 75% dos dados depois, o segundo quartil (Q2) divide 50% dos dados de um lado e 50% de outro, e o terceiro quartil (Q3) divide 75% dos dados antes dele e 25% depois. Entre cada quartil temos 25% dos dados.

Alguém mais atento pode me falar, “mas se o Q2 divide os dados restando 50% para cada lado ele é a mediana!”. Exatamente isso, o Q2 é a própria mediana.

Vamos supor uma amostra com 20 dados:

(5) (5) (5) (5)
1, 2, 3, 4, 5, | 6, 7, 8, 9, 10, | 11, 12, 13, 14, 15, | 16, 17, 18, 19, 20
Q1 Q2 Q3
5,5 10,5 15,5

Existem ainda os:

- Decis, que têm essa mesma ideia dos quartis, mas dividem a população/amostra em 10 partes iguais, portanto teremos 9 decis, ficando 10% da amostra em cada parte, e ;
- Percentis, que dividem a população/amostra em 100 partes iguais, portanto teremos 99 percentis, ficando 1% da amostra em cada parte. Por exemplo, o P10 significa que temos 10% dos dados a esquerda, e o restante (90%) a direita.

DISPERSÃO (AMPLITUDE, AMPLITUDE INTERQUARTIL, VARIÂNCIA, DESVIO PADRÃO E COEFICIENTE DE VARIAÇÃO)

AMPLITUDE

A **amplitude** é a diferença entre o maior valor e o menor valor da lista de dados, vamos supor a lista: **3**, 6, 10, 12, 14, **20**. A amplitude será:

$Ampl = X_{máx} - X_{mín}$
$Ampl = 20 – 3 = 17$

DESVIO QUARTÍLICO

O desvio quartílico, que também podemos chamar de amplitude semi-interquartílica, é a diferença entre o 3º quartil e o 1º quartil dividido por 2.

$$DQ = \frac{Q_3 - Q_1}{2}$$

Intervalo Quartílico

O intervalo quartílico é diferente do desvio quartílico (intervalo semi-interquartílico), o intervalo quartílico é apenas a diferença entre o quartil 3 e o quartil 1.

$$IQ = Q3 – Q1$$

DESVIO

Desvio é a diferença entre um dado qualquer de uma amostra/população (X_i) e a média desses dados ($\mu / \overline{x}$).

$$D = X_i - \overline{x}$$

O desvio em si não diz muita coisa e é dificilmente cobrado, mas é importante para entendermos o desvio médio e o desvio padrão (esse sim muito cobrado).

Desvio Médio

A soma de todos os desvios de uma amostra resulta em 0, vejamos o exemplo a seguir, de uma amostra com 4 dados: 4, 6, 8, 10

$$\overline{x} = \frac{4 + 6 + 8 + 10}{4} = \frac{28}{4}$$

$\overline{x} = 7$

Portanto os desvios são:

4: D = 4 – 7 = -3
6: D = 6 – 7 = -1
8: D = 8 – 7 = 1
10: D = 10 – 7 = 3

Soma dos desvios: - 3 – 1 + 1 + 3 = 0

Portanto, para acharmos o **desvio médio** vamos utilizar o somatório dos **módulos** dos desvios e dividir pela quantidade de dados. Utilizamos os módulos porque se usássemos os valores normais a soma daria 0.

$DM = \frac{\sum|Xi - X|}{n}$ para dados de uma **amostra**.

Ou:

$DM = \frac{\sum|Xi - \mu|}{n}$ para dados de uma **população**.

Ex: Vamos usar o mesmo exemplo:

4: D = 4 – 7 = -3
6: D = 6 – 7 = -1
8: D = 8 – 7 = 1
10: D = 10–7 = 3

Os módulos serão: 3, 1, 1 e 3.

$$DM = \frac{3 + 1 + 1 + 3}{4}$$

$$DM = \frac{8}{4}$$

$DM = 2$

Variância

Falamos que a média é o primeiro momento de uma distribuição, já a variância é chamada do segundo momento de uma distribuição.

A variância é uma medida de dispersão que mostra o quão distante cada valor desse conjunto está da média.

Na variância teremos pela primeira vez uma diferença na fórmula quando consideramos uma população e uma amostra. Até agora, as fórmulas eram sempre iguais, mudávamos apenas a representação da média (de μ para $\overline{x}$).

Em uma população a variância é representada por σ^2 (sigma ao quadrado), já para uma amostra a variância é representada por S^2.

Para uma amostra:

$$S^2 = \frac{\sum(Xi - \overline{X})^2}{n-1}$$

Para uma população:

$$\sigma^2 = \frac{\sum(Xi - \mu)^2}{n}$$

Podemos perceber que para uma amostra vamos dividir por "n-1" e para uma população vamos dividir por N. Usamos N (maiúsculo e minúsculo) para ilustrar que em uma população a quantidade de dados é maior que na amostra, mas os dois "enes" querem dizer a mesma coisa, ou seja, a quantidade de dados.

Quando temos a variância populacional, podemos calcular a variância amostral aplicando um fator de correção, que é:

$$\frac{n}{n-1}$$

$$S^2 = \sigma^2 \cdot \frac{n}{n-1}$$

Quando tivermos dados agrupados, basta multiplicar cada parênteses pela frequência da classe.

$$S^2 = \frac{\sum f(Xi - \overline{X})^2}{n-1}$$

ou

$$\sigma^2 = \frac{\sum f(Xi - \mu)^2}{n-1}$$

Existe uma outra forma de calcular a variância, que em boa parte dos casos é mais rápida, que é **a diferença entre a média dos quadrados e o quadrado da média**.

$$\sigma^2 = \overline{x}^2 - (\overline{x})^2$$

Ou seja, basta elevar todos os dados ao quadrado e calcular a média desses valores, e subtrair disso a média dos dados ao quadrado.

Desvio Padrão

O desvio padrão nos dá a noção de quão dispersos são os dados da amostra, quanto menor o desvio padrão, mais concentrado em torno da média aritmética estão os dados, quanto maior o desvio padrão, mais dispersos esses dados.

O desvio padrão é a raiz da variância, portanto a fórmula e as considerações sobre população e amostra são as mesmas, apenas vamos fazer a raiz do resultado.

$$\text{Desvio padrão} = \sqrt{\text{Variância}}$$

Em uma população, o desvio padrão é representada por σ (sigma), já para uma amostra o desvio padrão é representado por S.

Para uma amostra: $s = \sqrt{\frac{\sum(Xi - \overline{X})^2}{n-1}}$

Para uma população: $\sigma = \sqrt{\frac{\sum(Xi - \mu)^2}{n}}$

Quando tivermos dados agrupados, basta multiplicar cada parênteses pela frequência da classe.

$$s = \sqrt{\frac{\sum f(Xi - \overline{X})^2}{n-1}} \quad \text{ou} \quad \sigma = \sqrt{\frac{\sum f(Xi - \mu)^2}{n}}$$

Coeficiente de Variação

O coeficiente de variação (ou coeficiente de variação de Pearson) é a divisão do desvio padrão pela média.

Para uma amostra: $CV = \frac{S}{\overline{X}}$

Para uma população: $CV = \frac{\sigma}{\mu}$

Propriedades

Vamos pensar num conjunto de dados do salário de uma determinada empresa. Com esses valores podemos calcular a média, mediana, desvio padrão, variância etc.

Em determinado momento o dono da empresa resolve dar um aumento de 10% para todos os funcionários, ou seja, multiplicando o salário de todos por 1,1. Nesse caso, o que acontecerá com esses dados estatísticos? Ou então, o dono resolve aumentar R$ 100,00 no salário de cada um, ou seja, somar 100,00 em todos os dados, o que acontecerá com esses dados estatísticos?

Quando multiplicamos todos os salários por uma constante qualquer (1,1 por exemplo), a média será alterada na mesma ordem, ou seja, para achar a nova média multiplicaremos a média anterior pela mesma constante (1,1). A mesma coisa ocorre para a mediana, a moda e o desvio padrão. A única grandeza diferente é a variância, que será multiplicada pelo quadrado dessa constante, pois lembremos que ela é o quadrado do desvio padrão.

Em resumo, ao multiplicar os dados por uma constante k, os dados estatísticos serão alterados da seguinte maneira:

DADO ESTATÍSTICO	FORMA DE CHEGAR AO NOVO VALOR
Média (μ)	μ . k
Mediana (Md)	Md . k
Moda (Mo)	Mo . k
Desvio padrão (σ)	σ. k
Variância (σ^2)	$\sigma^2 . k^2$

Quando somamos todos os salários por uma constante qualquer (100 por exemplo), a média será alterada na mesma ordem, ou seja, para achar a nova média somaremos a média anterior pela mesma constante. A mesma coisa ocorre para a mediana e a moda. Entretanto as medidas de dispersão, desvio padrão e variância não serão alteradas pela soma de uma constante.

Assim o resumo dos valores, após a soma dos dados pela constante k, é:

DADO ESTATÍSTICO	FORMA DE CHEGAR AO NOVO VALOR
Média (μ)	μ + k
Mediana (Md)	Md + k
Moda (Mo)	Mo + k
Desvio padrão (σ)	Não muda
Variância (σ^2)	Não muda

Resumindo, quando somamos k a todos os elementos de uma série de dados, e/ou quando multiplicamos por k todos os elementos de uma série de dados:

	SOMA DE UMA CONSTANTE K	MULTIPLICAÇÃO DE UMA CONSTANTE K
NOVA MÉDIA/ MEDIANA/ MODA	Soma k	Multiplica por k
NOVO DESVIO PADRÃO	Não altera	Multiplica pelo módulo de k
NOVA VARIÂNCIA	Não altera	Multiplica por k^2

Vantagens e Desvantagens da Média Aritmética

A média aritmética é, com certeza, o parâmetro estatístico mais conhecido pela grande maioria das pessoas, pois é muito usada desde os primeiros anos de escola, e em muitos casos no dia a dia da população, até pela facilidade de ser calculada.

Entretanto, levando em conta que a média aritmética é uma medida estatística (medida de tendência central), assim como a mediana e a moda, por exemplo, ela tem suas vantagens e desvantagens, ao compararmos com essas outras duas medidas.

VANTAGENS	DESVANTAGENS
Extremamente indicada para dados em que os valores sejam simétricos em relação ao valor central (Curva Normal)	É muita influenciada pelos valores extremos, podendo causar uma distorção caso a distribuição seja assimétrica.
Indicada para comparar conjuntos de dados que guardem semelhança entre si.	Não traz informações sobre a quantidade de dados acima ou abaixo da média, bem como não é possível saber o valor mais frequente.

CÁLCULO DE PROBABILIDADE E TEOREMA DE BAYES

CONCEITOS

A probabilidade é a parte da Matemática que calcula a "chance" (probabilidade) de que algo aconteça, como, por exemplo, jogar na Mega-Sena e ganhar, ou então, de chutar uma questão no concurso e acertar.

Para iniciar essa teoria, vamos partir de alguns conceitos importantíssimos para a probabilidade:

- **Experimento aleatório:** é o evento que pode ter diferentes resultados, quando repetidos nas mesmas condições, ou seja, não sabemos o resultado, mas podemos saber quais são os resultados possíveis de serem obtidos. Ex.: o lançamento de um dado é um experimento aleatório, sabemos que pode cair 1,2, 3, 4, 5 ou 6, mas não sabemos qual exatamente irá cair naquele lançamento específico;
- **Ponto amostral:** é qualquer um dos resultados possíveis no experimento aleatório. Ex.: No lançamento do dado, se o dado cair com a face 1 para cima é um ponto amostral, assim como todas as outras faces, 2, 3, 4, 5 e 6;
- **Espaço amostral**: é o **conjunto** de todos os resultados possíveis do experimento aleatório, é dado pela letra S. Ex.: No lançamento do dado, o espaço amostral é: S = {1, 2, 3, 4, 5, 6}. Nesse caso, dizemos que o número de elementos do espaço amostral é dado por n(S), portanto, nesse exemplo n(S) = 6. A definição do espaço amostral é um passo muito importante na resolução dos exercícios de probabilidade. Em muitas questões, para definirmos o espaço amostral será usada a análise combinatória;
- **Evento:** é um subconjunto do espaço amostral. O evento será representado por uma letra maiúscula, A, e o número de elementos do evento é dado por n(A). Ex.: No lançamento do dado, se quisermos um resultado ímpar, temos 3 opções: 1, 3 e 5. Portanto, representamos isso da seguinte maneira: A = {1, 3, 5} e n(A) = 3, pois o conjunto A possui 3 elementos. Quando temos um subconjunto vazio (), dizemos que é um **evento impossível**, já que ele nunca poderá ocorrer. Já quando temos o subconjunto S, ele é igual ao **espaço amostral** e chamamos de **evento certo**, pois ele, com certeza, irá ocorrer.

Resumindo:

- Experimento aleatório: lançamento do dado;
- Ponto amostral: a face que caiu virada pra cima;
- Espaço amostral: S = {1, 2, 3, 4, 5, 6} e n(S) = 6;
- Evento A: obter um resultado maior que 4: A = {5, 6} e n(A) = 2;
- Evento impossível: obter o resultado 7. B = {} e n(B)=0;
- Evento certo: obter resultado maior que 0 e menor que 7; C = {1, 2, 3, 4, 5, 6} e n(C)=6.

A probabilidade de ocorrer o evento A, representada por P(A), em um experimento aleatório é a divisão entre número de elementos de A, n(A), pelo número de elementos do espaço amostral n(S). Em outras palavras, dizemos que é a divisão do número de casos favoráveis, n(A), pelo número de casos possíveis, n(S).

$$P(A) = \frac{n(A)}{n(S)}$$

Onde:

P(A) é a probabilidade de acontecer o evento A;

n(A): é o número de elementos favoráveis que fazem parte do evento A;

n(S): é o número de elementos do espaço amostral S.

Agora, vamos conhecer uma questão já cobrada em concursos públicos e que ilustra bem o assunto:

1. (CESPE-CEBRASPE – 2016) Considere a seguinte informação para responder à questão: a Prefeitura do Município de São Paulo (PMSP) é subdividida em 32 subprefeituras e cada uma dessas subprefeituras administra vários distritos.

A tabela a seguir, relativa ao ano de 2010, mostra as populações dos quatro distritos que formam certa região administrativa do município de São Paulo.

Distrito	População (em 2010)
Alto de Pinheiros	43.000
Itaim Bibi	92.500
Jardim Paulista	89.000
Pinheiros	65.500
Total	**290.000**

Considerando-se a tabela apresentada, é correto afirmar que, se, em 2010, um habitante dessa região administrativa tivesse sido selecionado ao acaso, a chance de esse habitante ser morador do distrito Jardim Paulista seria superior a 29% e inferior a 33%.

( ) CERTO ( ) ERRADO

A probabilidade é a divisão do número total de casos favoráveis pelo total de casos possíveis. Os casos favoráveis são os moradores do Jardim Paulista (89.000) e o total de casos possíveis é 290.000, ou seja, o total de habitantes nessa população. Logo, a probabilidade de selecionar um habitante do Jardim Paulista "P(JP)" é:

$$P(JP) = \frac{89.000}{290.000}$$

P(JP) = 0,3068 = 30,68%. Resposta: Certo.

AXIOMAS DA PROBABILIDADE

Axiomas são verdades inquestionáveis, ou seja, conceitos fundamentais da probabilidade, também conhecido como axiomas de *Kolmogorov*.

Os axiomas da probabilidade são 3:

- A probabilidade de um evento acontecer está entre 1 (evento certo – 100%) e 0 (evento impossível – 0%); $\leq P(A) \leq 1$;
- A soma das probabilidades de todos os eventos do espaço amostral é igual a 1: P(S) = P(A) + P(B) + P(C) = 1 para um espaço amostral onde os únicos eventos possíveis são A, B e C;
  Ex.: No lançamento de uma moeda, são dois eventos possíveis, podendo sair cara ou sair coroa. A probabilidade de cada um dos eventos é ½, portanto, a soma entre eles é ½ + ½ = 1;
- Se A e B forem eventos mutuamente excludentes, ou seja, se um ocorre o outro não ocorre, não havendo intersecção entre eles, a soma das probabilidades de cada evento é a probabilidade da união dos dois eventos, que será 1.

  P(AB) = P(A) + P(B), quando AB é vazio (θ).

  Ex.: Quando temos dois eventos complementares, isso sempre acontece, pois a probabilidade de um evento acontecer somado com a probabilidade desse evento não acontecer será sempre 1. Você fez uma aposta na vitória do seu time, temos apenas dois eventos possíveis, ou seu time ganha, ou seu time não ganha (perde ou empata), é impossível acontecer as duas coisas. Esses eventos, portanto, são complementares.

Intersecção de Eventos (regra do E)

Quando um exercício pede a probabilidade de acontecer um evento **E** um outro evento qualquer, temos a probabilidade da intersecção de dois eventos P(A∩B). A fórmula é dada por:

$$\mathbf{P(A \text{ e } B) = P(A) \cdot P(B|A)}$$

Onde:

P(B|A) significa a probabilidade de ocorrer B, sendo que A já ocorreu.

Ex.: Qual a probabilidade de retirar dois ases de um baralho de cartas, sem que haja reposição das cartas.

Apesar de o exercício não colocar o "e", temos uma intersecção de eventos aí, já que queremos retirar uma ás (A1), **E** depois, mais um Ás (A2). O total de cartas do baralho é 52 e o total de ases é 4.

A probabilidade de sair o primeiro ás é: $\frac{4}{52}$,

pois o total de eventos possíveis é 52, e o total de eventos favoráveis é 4.

A probabilidade de sair o segundo ás, dado que já saiu um ás, é $\frac{3}{51}$, pois agora temos apenas 3 ases e 51 cartas.

Portanto, a probabilidade de sortear dois ases é

$$P(A1 \text{ e } A2) = P(A1) \cdot P(A2|A1)$$

$$P(A1 \text{ e } A2) = \frac{4}{52} \cdot \frac{3}{51}$$ (simplificando o 4 com o 52, e o 3 com 51)

$$P(A1 \text{ e } A2) = \frac{1}{13} \cdot \frac{1}{17}$$

$$P(A1 \text{ e } A2) = \frac{1}{221}$$

Logo, a probabilidade de retirar dois ases é $\frac{1}{221}$.

Eventos Independentes

Dizemos que dois eventos são independentes quando a ocorrência de um não interfere na probabilidade do outro.

Um exemplo fácil de assimilar é no lançamento sucessivo de um dado. Se no primeiro lançamento eu tiro 3, quando vou lançar o segundo dado, a probabilidade de tirar qualquer coisa foi alterada? Não. A probabilidade de tirar qualquer número específico será sempre 1/6, logo, podemos dizer que esses eventos são **independentes**.

Agora, vamos pensar no nosso exemplo do tópico anterior, retirando dois ases de um baralho.

- Se tirarmos duas cartas em sequência, sem reposição da carta retirada, a probabilidade de retirar um ás na segunda tentativa é diferente da primeira, uma vez que já foi retirada uma carta, mudando o nosso espaço amostral e mudando o número total de eventos favoráveis, portanto, esses eventos são **dependentes;**
- Se retirarmos duas cartas em sequência, com reposição da carta retirada, a probabilidade de retirar um ás na segunda tentativa é a mesma que na primeira, pois o total de cartas não foi alterado nem o total de ases, portanto, esses eventos são **independentes**.

Na maioria dos casos conseguimos concluir se os eventos são independentes ou dependentes, mas em alguns casos, pelo relato, não é possível sabermos. Nesse caso, sabemos que quando os eventos são independentes, a probabilidade de B ocorrer, sabendo que A já ocorreu "P(B|A)", é igual a probabilidade de B ocorrer: "P(B)".

$$P(B|A) = P(B).$$

Da mesma forma, P(A|B) = P(A).
Portanto, podemos dizer que os eventos A e B são independentes se, e somente se, ocorrer:

$$\mathbf{P(A \text{ e } B) = P(A) \cdot P(B)}$$

Eventos Mutuamente Exclusivos

Dizemos que dois eventos são mutuamente exclusivos quando não podem ocorrer ao mesmo tempo, ou seja, quando um ocorre, o outro não ocorre.
Portanto, quando dois eventos são mutuamente exclusivos:

- P(A|B) = 0; A probabilidade de "A" ocorrer, sabendo que "B" já ocorreu, é zero;
- P(B|A) = 0; A probabilidade de "B" ocorrer, sabendo que "A" já ocorreu, é zero;
- P(A e B) = 0; A probabilidade de A e B ocorrerem ao mesmo tempo é zero.

Os eventos complementares são sempre mutuamente exclusivos, mas os eventos mutuamente exclusivos nem sempre são complementares, apesar de ambos não acontecerem ao mesmo tempo. Para os eventos serem mutuamente exclusivos, basta que eles não possam ocorrer ao mesmo tempo. Para ser complementar, quando um não ocorre, o outro obrigatoriamente tem que ocorrer. A probabilidade de ocorrer o evento A é P(A), e seu complementar é não A, representado por $\bar{A}$, sendo que a probabilidade de acontecer $\bar{A}$ é $P(\bar{A})$.

Exemplo 1: No lançamento de um dado

Evento A: sair um número par;
Evento B: sair um número ímpar.

Esses dois eventos são mutuamente exclusivos e complementares, pois os dois não podem acontecer ao mesmo tempo, e se o resultado não for ímpar, tem que ser par.

Exemplo 2: Jogo do Palmeiras

Evento A: O Palmeiras ganhou o jogo;
Evento B: O Palmeiras perdeu o jogo.

Esses dois eventos são mutuamente exclusivos, pois os dois eventos não podem ocorrer ao mesmo tempo, mas não são complementares, pois pode acontecer um empate também. Portanto, se o Palmeiras não ganhou, não podemos dizer necessariamente que ele perdeu.
Obviamente, dois eventos complementares ou dois eventos mutuamente excludentes são **dependentes**, pois um evento altera a probabilidade do outro.

União de Dois Eventos (Regra do OU)

Quando um evento pede a probabilidade de acontecer um evento **OU** outro, temos a probabilidade da união de dois eventos P(A∪B). A fórmula é dada por:

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B)$$

Ou, usando os símbolos:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Quando dois eventos são **dependentes**, sabemos que:

  $P(A \cap B) = P(A) \cdot P(B|A)$
  $P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B|A)$

- Quando temos dois eventos **independentes**, sabemos que:

  $P(A \cap B) = P(A) \cdot P(B)$
  $P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$

- Quando temos dois eventos **mutuamente exclusivos**, sabemos que:

  $P(A \cap B) = 0$

  $P(A \cup B) = P(A) + P(B)$

Importante!

Eventos **dependentes**: $P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B|A)$
Eventos **independentes**: $P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$
Eventos **mutuamente exclusivos**: $P(A \cup B) = P(A) + P(B)$

Ex.: Vamos considerar os eventos A e B quaisquer, pertencentes a um mesmo espaço amostral, em um experimento aleatório. Considerando que P(A) = 0,30, vamos julgar essas 4 afirmações:

- **P(B) = 0,6 e $P(A \cap B)$ = 0.30, se A e B forem independentes;**

  Para os eventos serem independentes: $P(A \cap B) = P(A) \cdot P(B)$
  P(A) = 0,3; P(B) = 0,6; $P(A \cap B)$ = 0,18
  $P(A \cap B) = P(A) \cdot P(B)$
  0,18 = 0,3 . 0,6
  0,18 = 0,18 (CERTO)

- **P(B) = 0,6 e $P(A \cup B)$ = 0,7, se A e B forem independentes;**

  Para os eventos serem independentes: $P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$
  P(A) = 0,3; P(B) = 0,6; $P(A \cup B)$ = 0,7

$P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$
$0,7 = 0,3 + 0,6 - 0,3.0,6$
$0,7 = 0,9 – 0,18$
$0,7 = 0,72$ (ERRADO)

- **P(B) = 0,75, se A e B forem mutuamente exclusivos;**

Para os eventos serem mutuamente exclusivos: $P(A \cup B) = P(A) + P(B)$
$P(A) = 0,3$; $P(B) = 0,75$;
$P(A \cup B) = P(A) + P(B)$
$P(A \cup B) = 0,3 + 0,75$
$P(A \cup B) = 1,05$

A probabilidade nunca pode ser maior que 1, portanto, item **errado**.

- **P(A∪B) = 0,15, se A e B forem mutuamente exclusivos;**

Para os eventos serem mutuamente exclusivos: $P(A \cup B) = P(A) + P(B)$
$P(A) = 0,3$; $P(A \cup B) = 0,15$;
$P(A \cup B) = P(A) + P(B)$
$0,15 = 0,3 + P(B)$ (passando o 0,3 para o outro lado do igual com sinal trocado).
$P(B) = 0,15 – 0,3$
$P(B) = - 0,15$
A probabilidade nunca pode ser negativa, portanto, item errado.

PROBABILIDADE – TEOREMA DE BAYES

Usamos o Teorema de Bayes quando conhecemos as probabilidades condicionais da forma $P(B|A)$ e queremos calcular a probabilidade condicional da forma $P(A|B)$, ou seja, conhecemos as probabilidades com o evento B, a *posteriori*, e desejamos conhecer as probabilidades para esse evento a *priori*. De maneira geral, com n eventos A_i e conhecendo $P(B|A_i)$, a probabilidade de algum evento A_m, condicionada ao evento B, $P(A_m|B)$, é:

$$P(A_m|B) = \frac{P(B|A_m) \cdot P(A_m)}{P(B|A_1) \cdot P(A_1) + P(B|A_2) \cdot P(A_2) + \cdots + P(B|A_n) \cdot P(A_n)}$$

Vejamos um exemplo para facilitar o entendimento:

Uma cidade sede do interior possui três varas trabalhistas. A 1ª Vara, comporta 50% das ações trabalhistas, a 2ª Vara comporta 30% e a 3ª Vara as 20% restantes. As porcentagens de ações trabalhistas oriundas da atividade agropecuária são 3%, 4% e 5% para a 1ª, 2ª e 3ª Varas, respectivamente. Escolhe-se uma ação trabalhista aleatoriamente e constata-se ser originária da atividade agropecuária. A probabilidade dessa ação ser da 1ª Vara trabalhista é, aproximadamente:

a) 0,5312.
b) 0,3332.
c) 0,1241.
d) 0,4909.
e) 0,4054.

Para resolvermos a questão, é imperativo notar que as proporções das ações de atividade agropecuária para cada uma das varas são a probabilidade a *posteriori* e que a questão quer saber a probabilidade de uma ação ser da 1ª Vara, ou seja, um evento a *priori*.

Utilizando o Teorema de Bayes, $P(V_1|A)$ é dado por:

$$P(V_1|A) = \frac{P(A|V_1) \cdot P(V_1)}{P(A|V_1) \cdot P(V_1) + P(A|V_2) \cdot P(V_2) + P(A|V_3) \cdot P(V_3)}$$

Temos os seguintes dados extraídos da questão:

- A 1ª Vara comporta 50% das ações: $P(V_1) = 0,5$;
- A 2ª Vara comporta 30% das ações: $P(V_2) = 0,3$;
- A 3ª Vara comporta 20% das ações: $P(V_3) = 0,2$;
- As percentagens: $V_1 = 3\%$, $V_2 = 4\%$ e $V_3 = 5\%$

Logo,

$P(A|V_1) = 0,03$, $P(A|V_2) = 0,04$ e $P(A|V_3) = 0,05$.

Substituindo esses valores na fórmula do Teorema de Bayes, temos:

$$P(V_1|A) = \frac{0,03 \cdot 0,5}{0,03 \cdot 0,5 + 0,04 \cdot 0,3 + 0,05 \cdot 0,2} = \frac{0,015}{0,015 + 0,012 + 0,01} = \frac{0,015}{0,037} \cong 0,4054$$

Podemos concluir, portanto, que a resposta é a alternativa E.

Dica

O Teorema de Bayes está interligado à probabilidade condicional de dois eventos, ou seja, mostra a relação entre uma probabilidade condicional e a sua inversa.

PROBABILIDADE CONDICIONAL

Já discutimos um pouco sobre isso, mas existe uma fórmula muito comum nas questões de concurso, que pode ser usada para questões com Probabilidade.

Lembre-se de que a probabilidade condicional é aquela em que queremos a probabilidade de ocorrer um evento, dado que um outro evento já ocorreu.

Por exemplo, vamos supor que eu retirei uma carta de um baralho (que tem 52 cartas).

A probabilidade de eu retirar uma dama (Q) é: 4/52, uma vez que temos 4 damas em 52 cartas possíveis.

$$P(Q) = 4/52 = 0,076 = 7,6\%$$

Agora, a probabilidade de eu retirar uma carta de copas (♥) é: 13/52, uma vez que temos 13 cartas de copas e 52 cartas possíveis.

$$P(♥) = 13/52 = 0,25 = 25\%$$

Indo mais além, a probabilidade de eu retirar uma dama de copas, ou seja, retirar uma dama e que seja de copas, será 1/52, pois só temos uma dama de copas no baralho.

$$P(Q♥) = 1/52 = 0,019 = 1,9\%$$

Agora sim, vamos chegar à fórmula muito usada da Probabilidade Condicional. Vamos supor que eu queira saber a probabilidade de tirar uma dama, dado que a carta retirada é de copas. Ou seja, existe uma condição, eu já sei que a carta é de copas, portanto, meu espaço amostral, nesse caso, será apenas as cartas de copas do baralho, ou seja, 13. Pensando diretamente, a probabilidade é 1/13 = 7,7%, mas muitas vezes não é tão fácil de enxergar como nesse caso, portanto, existe uma fórmula para resolver essa questão.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Ou seja, a probabilidade de ocorrer A, uma vez que B já ocorreu, é a probabilidade da intersecção entre A e B, dividido pela probabilidade de B.

Nesse nosso caso, iríamos dividir a probabilidade de ocorrer uma dama de copas: P(Q♥) = 1,9%, pela probabilidade de sair uma carta de copas: P(♥) = 25%.

$$P(Q|♥) = \frac{1,9}{25} = 7,6\%$$

Agora, treine o que aprendeu com questões comentadas já cobradas em concursos públicos e que tratam do assunto:

1. **(CESPE-CEBRASPE – 2019)** Em um espaço de probabilidades, as probabilidades de ocorrerem os eventos independentes A e B são, respectivamente, P(A) = 0,3 e P(B) = 0,5. Nesse caso, P(B/A) = 0,2.

( ) CERTO ( ) ERRADO

Queremos saber a probabilidade de ocorrer B, dado que A já ocorreu. Portanto, vamos usar a fórmula da probabilidade condicional. Nesse caso, o evento que já ocorreu é o A, portanto, ele vem depois da barra (|) e será o denominador:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Pelo fato dos eventos A e B serem independentes, sabemos que:
$P(B \cap A) = P(A) \cdot P(B)$ – valores dados no enunciado.
$P(B \cap A) = 0,3 \cdot 0,5$
Portanto, a probabilidade condicional é:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

$$P(B|A) = \frac{0,3 \cdot 0,5}{0,3}$$

$P(B|A) = 0,5$. Resposta: Errado.

2. (CESPE-CEBRASPE – 2018) Os indivíduos S1, S2, S3 e S4, suspeitos da prática de um ilícito penal, foram interrogados, isoladamente, nessa mesma ordem. No depoimento, com relação à responsabilização pela prática do ilícito, S1 disse que S2 mentiria; S2 disse que S3 mentiria; S3 disse que S4 mentiria. A partir dessa situação, julgue o item a seguir. Considerando que a conclusão ao final do interrogatório tenha sido a de que apenas dois deles mentiram, mas que não fora possível identificá-los, escolhendo-se ao acaso dois entre os quatro para novos depoimentos, a probabilidade de apenas um deles ter mentido no primeiro interrogatório é superior a 0,5.

( ) CERTO ( ) ERRADO

Vamos ver duas formas de resolver esse tipo de exercício.
São 4 suspeitos, sendo que, 2 mentem e 2 não mentem. Vamos escolher 2 ao acaso, e queremos a possibilidade de escolher 1 que mente e 1 que não mente. Podemos escolher isso de duas formas: 1º mente (A) E 2º não mente (B) OU 1º não mente (B) E 2º mente (A). Resposta: Certo.

1º mente (A) E 2º não mente (B)

A probabilidade de escolher um que mente, com os 4 à disposição, é: P(A) = 2/4.
Saindo um que mente, restam 3 à disposição, sendo que 2 falam a verdade, a probabilidade de escolher um que fala a verdade é: P(B|A) = 2/3.
1º mente (A) E 2º não mente (B):

$$P(A) \,.\, P(B|A) = \frac{2}{4} \cdot \frac{2}{3} = \frac{4}{12} = \frac{1}{3}$$

1º não mente (B) e 2º mente (A)

A probabilidade de escolher um que não mente, com os 4 à disposição, é: P(B) = 2/4.
Saindo um que não mente, restam 3 à disposição, sendo que 2 mentem. A probabilidade de escolher um que mente é: P(A|B) = 2/3.
1º não mente (B) e 2º mente (A):

$$P(B) \,.\, P(A|B) = \frac{2}{4} \cdot \frac{2}{3} = \frac{4}{12} = \frac{1}{3}$$

1º mente (A) E 2º não mente (B) OU 1º não mente (B) E 2º mente (A)

Agora, basta somar as duas probabilidades, pois, irá acontecer uma coisa ou outra, ou seja, se acontecer uma coisa com certeza não acontecerá a outra (eventos mutuamente exclusivos).

$$\frac{1}{3} + \frac{1}{3} = \frac{2}{3} = 0,666 = 66,6\%$$

A segunda forma de resolver é:
De 4 pessoas, temos que escolher 2, portanto, o total de possibilidades é dado pela combinação de 4 pegando 2.

$$C_{4,2} = \frac{4!}{2! \cdot 2!}$$

$$C_{4,2} = \frac{4 \cdot 3 \cdot 2!}{2 \cdot 2!}$$

$$C_{4,2} = 2 \,.\, 3$$
$$C_{4,2} = 6$$

Das 6 opções, as favoráveis são aquelas em que são escolhidos 1 que mente e 1 que não mente, portanto, não serão favoráveis apenas quando forem escolhidos juntos os 2 que mentem, ou juntos os 2 que não mentem. Portanto, o total de casos desfavoráveis é igual a 2, logo, o número de casos favoráveis é 4.

$$P(A) = \frac{4}{6} = \frac{2}{3} = 0,666 = 66,6\%.$$ *Resposta: Certo.*

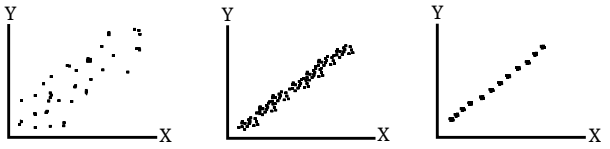
CORRELAÇÃO LINEAR SIMPLES

CORRELAÇÃO DE DUAS VARIÁVEIS

Considerando duas variáveis quaisquer, essas variáveis podem ter um grau de correlação entre elas, que chamamos de **coeficiente de correlação linear**, que pode ser representado por ρ(rho) ou R.

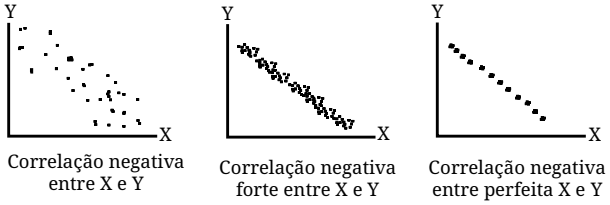
O coeficiente de correlação pode variar entre -1 e 1. E ele pode ser representado por um gráfico de dispersão, onde um eixo é a variável X e o outro a variável Y.

$0 < R \leq 1$: Correlação positiva – enquanto uma variável aumenta a outra também aumenta. Ex: Vamos considerar as variáveis altura e peso. Normalmente, quanto maior a altura, maior o peso. Podemos dizer que, quanto mais próximo de 1, mais forte será a correlação positiva, e, quanto mais próximo de zero, menos relacionada estarão as variáveis. Chamamos a correlação igual a 1 de perfeita, cujos dados resultam em uma reta.



Disponível em: <https://repositorio.utfpr.edu.br/jspui/bitstream/1/2460/2/PG_PPGECT_M_Lima%2C%20Sabrina%20Anne% 20de_2015_1.pdf>. Acesso em: 10 set. 2021

-1 ≤ R < 0: Correlação negativa – enquanto uma variável aumenta, a outra diminui. Ex.: Considerando a variável temperatura e a variável venda de chocolate quente. Quanto maior a temperatura do ambiente menor será a venda de chocolate quente. Podemos dizer que, quanto mais próximo de -1, mais forte será a correlação negativa, e, quanto mais próximo de zero, menos relacionada estarão as variáveis. Assim como na positiva, temos uma correlação negativa perfeita quando R = -1.



Disponível em: <https://repositorio.utfpr.edu.br/jspui/bitstream/1/2460/2/PG_PPGECT_M_Lima%2C%20Sabrina%20Anne% 20de_2015_1.pdf>. Acesso em: 10 set. 2021.

- **R = 0: Correlação nula** – as variáveis não são correlacionadas.

Disponível em: <https://repositorio.utfpr.edu.br/jspui/bitstream/1/2460/2/PG_PPGECT_M_Lima%2C%20Sabrina%20Anne% 20de_2015_1.pdf>. Acesso em: 10 set. 2021.

O coeficiente de correlação linear é dado pela razão entre a covariância de X e Y e o produto dos desvios-padrão de cada variável.

$$R = \frac{Cov(X, Y)}{\sigma X \cdot \sigma Y}$$

REGRESSÃO LINEAR

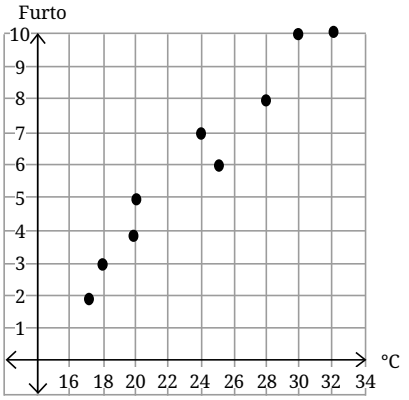
O coeficiente de correlação de Pearson (R) é o coeficiente que mostra o grau de dependência entre duas variáveis, também podemos achar o R através dos pares ordenados formados por cada ponto das duas variáveis.

Vamos imaginar que queremos saber se duas variáveis aleatórias quaisquer estão correlacionadas, como por exemplo o aumento da temperatura e a quantidade de furtos na Baixada Santista aos finais de semana. Para isso são coletados a quantidade de furtos registrados em 10 dias (5 finais de semanas) e a temperatura média em cada um desses dias, e chegamos na seguinte tabela:

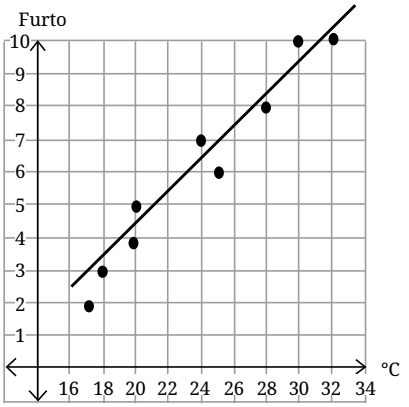
TEMPERATURA	Nº DE FURTOS
17°	2
18°	3
25°	6
24°	7
20°	5
30°	10
28°	8
20°	4
32°	10
33°	12

*Dados não verídicos

Olhando os dados podemos até notar que nos dias mais quentes o número de furtos aumenta, mas vamos colocar isso em um gráfico de coordenadas, no eixo x vamos colocar a temperatura, e no eixo y os furtos, para visualizar melhor. Cada coluna da nossa tabela é um par ordenado: (17,2), (18,3), (25,6), (24,7) e assim por diante. Vamos colocar um ponto no gráfico para cada par ordenado para visualizar essa correlação.



Podemos ver que existe uma correlação entre as variáveis, pois pelo gráfico vemos que conforme uma variável aumenta a outra aumenta também, mas podemos ver que não é algo exato, não forma uma linha perfeita (casos em que R = 1). A regressão linear tem o intuito de definir uma função que defina essa correlação entre as variáveis e a regressão mais usada é a linear, a qual definiremos uma função linear para explicar essa correlação. Basicamente faremos uma linha reta resumindo esses dados:



O coeficiente de correlação pode ser calculado pela seguinte fórmula, sendo X e Y as duas variáveis aleatórias que queremos analisar a correlação:

$$R = \frac{n \cdot \sum X \cdot Y - \sum X \cdot \sum Y}{\sqrt{n \cdot (\sum X^2) - (\sum X)^2} \cdot \sqrt{n \cdot (\sum Y^2) - (\sum Y)^2}}$$

Vamos achar o R do nosso exemplo, sendo temperatura X e nº de furtos Y e o número de dados n = 10 (10 dias de observação):

	X	Y	X.Y	X2	Y2
	17	2	34	289	4
	18	3	54	324	9
	25	6	150	625	36
	24	7	168	576	49
	20	5	100	400	25
	30	10	300	900	100
	28	8	224	784	64
	20	4	80	400	16
	32	10	320	1024	100
	33	12	396	1089	144
TOTAL (Σ)	**247**	**67**	**1826**	**6411**	**547**

$$R = \frac{10 \cdot 1826 - 247 \cdot 67}{\sqrt{10 \cdot 6411 - 247^2} \cdot \sqrt{10 \cdot 547 - 67^2}}$$

$$R = \frac{18.260 - 16.549}{\sqrt{64.110 - 61.009} \cdot \sqrt{5.470 - 4.489}}$$

$$R = \frac{1.711}{\sqrt{3.101} \cdot \sqrt{981}}$$

$$R = \frac{1.711}{55{,}69 \cdot 31{,}32}$$

$$R = \frac{1.711}{1.744{,}21}$$

$$R = 0{,}98$$

Com isso, vemos que o R está muito próximo de 1, o que caracteriza uma correlação muito forte entre as variáveis.

Temos que fazer uma observação importante. A correlação nem sempre tem efeito de causa e consequência, pois, o fato de aumentar a temperatura não faz com que os bandidos queiram roubar mais, na verdade o aumento da temperatura faz com que mais pessoas queiram ir para a praia se refrescar, e nesse momento os bandidos enxergam uma melhor oportunidade para praticar seus delitos, por isso o aumento. Portanto apesar das variáveis estarem correlacionadas elas não são causa e consequência.

Do coeficiente de correlação de Pearson podemos achar o **Coeficiente de Determinação (CD) ou coeficiente de explicação**, que mede o percentual da variação de Y que é explicado pela variação de X.

$CD = R^2$

No nosso exemplo:

$CD = 0{,}98^2 = 0{,}96$

Portanto 96% da variação de X é explicada pela variação de Y e os outros 4% de variação têm outros motivos.

Importante!

Correlação: mede o grau de relação ($-1 \leq R \leq 1$);
Regressão: dá uma função que relaciona as variáveis.

A regressão linear é a definição de uma função que relacione as duas variáveis X e Y. Vamos padronizar que Y é a variável dependente, e X é a variável independente. Como no nosso exemplo, a temperatura (X) é uma variável independente pois ela não sofre influência do número de furtos, e o número de furtos (Y) é a variável dependente, pois ela sim sofrerá influência da temperatura (X). Por ser uma função linear ela terá essa cara:

$$Y = a \,.\, X + b$$

Y é a variável dependente, que varia de acordo com o X, X é a variável independente, "a" é o coeficiente angular que acompanha o X, e "b" é o coeficiente linear (intercepto), que é o termo independente, que vem sozinho.

Para achar essa função vamos utilizar o **método dos mínimos quadrados**.

Para achar o valor do coeficiente angular ("a") vamos usar a fórmula:

$$a = \frac{\text{cov}(X, Y)}{\text{Var}(x)}$$

Que podemos extrapolar para:

$$a = \frac{n \cdot \sum X \cdot Y - \sum X \cdot \sum Y}{n \cdot (\sum X^2) - (\sum X)^2}$$

ou

$$a = \frac{n \cdot \sum X \cdot Y - n\,\overline{X} \cdot \overline{Y}}{n \cdot \sum X^2 - n(\overline{X})^2}$$

Onde $\overline{X}$ é a média dos dados de X e $\overline{Y}$ é a média dos dados de Y.

Para achar o b (coeficiente linear), vamos usar as médias de X e Y.

$\overline{Y} = a.\,\overline{X} + b$

Muitas vezes, podemos ver $Y = a + bx$, isso é a mesma coisa, só temos que tomar cuidado para não confundir, pois, nesse caso, usaremos a fórmula para achar o b. A fórmula é usada para achar o coeficiente angular (que acompanha o X), independente do nome dado a ele.

ANÁLISE DE VARIÂNCIA (ANOVA) E ANÁLISE DE RESÍDUOS

A análise de variância (da sigla em inglês ANOVA) é usada para comparar médias de populações diferentes, com a finalidade de descobrir se as médias de certo parâmetro entre populações são iguais ou não, possibilitando comparar 3 ou mais grupos de uma só vez.

A ANOVA nos mostra se as diferenças amostrais observadas são reais, ou seja, causadas por uma diferença da população observada, ou casuais, ou seja, causadas pelo acaso por conta da aleatoriedade na escolha da amostra.

Para aplicarmos a ANOVA, as amostras devem ser aleatórias e independentes, as populações devem possuir distribuição normal, e as variâncias populacionais devem ser iguais (não rigorosamente iguais, mas próximas).

Para entender melhor, vamos utilizar um exemplo simples. Vamos supor que queremos testar o tempo de resposta de 3 drogas diferentes para o tratamento de uma doença (remédios A, B e C). Para isso, vamos aplicar cada droga em 5 pacientes diferentes, e cronometrar o tempo de resposta, ou seja, o tempo necessário para a eliminação dos sintomas.

Ao realizarmos a ANOVA iremos comparar a variância **entre** os grupos e a variância **dentro** de cada grupo. Queremos saber se os tempos médios de respostas para cada droga são iguais, portanto, a hipótese nula será:

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

H_1: A hipótese alternativa indicará que pelo menos uma das médias é diferente.

Realizadas as medições dos tempos de respostas, em minutos, temos:

	A	B	C
PACIENTE 1	9	10	9
PACIENTE 2	12	14	13
PACIENTE 3	18	19	17
PACIENTE 4	24	23	26
PACIENTE 5	36	38	35
MÉDIA	19,8	20,8	20

A média geral das 15 observações é dada por: $\overline{X} = 20,2$.

Podemos notar que as médias amostrais para cada remédio são diferentes umas das outras. O que a ANOVA quer analisar é se essas médias populacionais são iguais ou não, ou seja, se essa diferença nas médias amostrais se deve apenas aos fatores aleatórios da amostra selecionada ou se essa diferença se deve à qualidade do remédio mesmo.

Analisando os dados, vemos que a variação dentro de cada grupo é grande, e a variação entre os grupos não tem muita diferença, portanto, nesse caso, aceitaríamos H_0 e concluiríamos que a variação é causada pela diferença da amostra mesmo.

Por outro lado, podemos ter os seguintes dados das medições:

	A	B	C
PACIENTE 1	29	17	11
PACIENTE 2	28	18	12
PACIENTE 3	28	18	11
PACIENTE 4	31	19	12
PACIENTE 5	32	20	14

Analisando os dados, vemos que a variação dentro de cada grupo é pequena e a variação entre os grupos é grande, portanto, nesse caso, rejeitaríamos H_0 e concluiríamos que a variação realmente é causada pela diferença da droga.

Olhando essa tabela com poucos dados e uma diferença gritante, é fácil de deduzirmos isso, mas nem sempre é assim fácil de enxergar, por isso são usados vários cálculos e índices para se chegar a essa conclusão. Lembrando que a ANOVA só irá nos mostrar se existe diferença ou não entre os grupos, mas não irá mostrar qual dos grupos (drogas, no nosso exemplo) diverge dos demais.

Para concluirmos sobre a hipótese nula vamos usar o teste F, que é a razão entre a variância **entre** os grupos pela variância **dentro** dos grupos:

$$F = \frac{S^2_{\text{entre}}}{S^2_{\text{dentro}}} = \frac{\text{(variância entre os grupos)}}{\text{(variância dentro dos grupos)}}$$

Para calcularmos as variâncias vamos usar a soma de quadrados de desvios. São 3 somas que temos que saber: a soma de quadrados total (SQT), a soma de quadrados **entre** grupos ou soma de quadrados de **tratamentos** ou soma dos quadrados do modelo (SQE, SQ_{Trat} ou SQM) e a soma de quadrados **dentro** dos grupos ou soma de quadrados dos **resíduos** ou soma de quadrados dos **erros** (SQD, SQ_{Res} ou SQ_{Erro}). Muito cuidado com essas siglas, em cada lugar elas aparecem de uma maneira, por isso, tenham em mente que a "entre grupos" – SQE – é a do modelo, ou da regressão ou outro nome que remeta ao modelo da regressão linear, já a "dentro dos grupos" – SQD - é o erro, ou resíduos, ou algo relacionado a isso, já o SQT é o total.

Uma boa notícia é que geralmente os exercícios fornecem os valores de SQT, SQE e SQD. A SQ_{total} é a soma das outras duas:

$$SQ_{total} = SQ_{entre} + SQ_{dentro}$$

Para achar as variâncias, que também chamamos nesse caso de **quadrados médios (QM)**, temos que dividir o SQ pelos graus de liberdade. Cada SQ terá uma quantidade de graus de liberdade diferente.

Os graus de liberdade são dados por:

Entre os grupos: gl_1 ou $gl_{entre} = k - 1$
Dentro dos grupos: gl_2 ou $gl_{dentro} = n - k$
Total: $gl_{total} = n - 1$

Sendo que: k é a quantidade de grupos; e n é a quantidade de observações totais (de todos os grupos)

Assim como vimos nas SQs, o grau de liberdade total é dado pela soma de gl_1 e gl_2:

$gl_{total} = n - k + (k - 1)$
$gl_{total} = n - 1$

A quantidade de graus de liberdade do modelo (entre) é a quantidade de variáveis explicativas do modelo.

No nosso exemplo temos 3 grupos (remédios A, B e C), portanto k = 3, e para cada grupo temos 5 pacientes, portanto o total de amostras é 15, logo: n = 15.

Assim:

gl_1 ou $gl_{entre} = 3 - 1 = 2$
gl_2 ou $gl_{dentro} = 15 - 3 = 12$
$gl_{total} = 15 - 1 = 14$

Podemos ver que a soma de gl_1 e gl_2 resulta no gl_{total}.

Portanto, as variâncias ou quadrados médios (QM) são dados por:

$$QMT \text{ ou } S^2_{Total} = \frac{SQT}{n - 1}$$

$$QME \text{ ou } S^2_{Entre} = \frac{SQE}{k - 1}$$

$$QMD \text{ ou } S^2_{Dentro} = \frac{SQD}{n - k}$$

Assim podemos achar o valor do teste F.

$$F = \frac{S^2_{entre}}{S^2_{dentro}} = \frac{QME}{QMD}$$

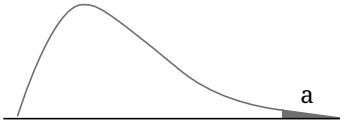
O modo de usar a tabela F não é muito diferente das outras que já vimos, mas temos uma tabela para cada nível de significância (1 - α), e para usá-la vamos precisar da quantidade de **graus de liberdade entre e dentro**.

Vamos ver a tabela do teste F para um nível de significância de 95%. Nas colunas temos a quantidade de graus de liberdade **entre** gl_1, que na tabela está como v_1, e nas linhas temos a quantidade de graus de liberdade **dentro** (gl_2, que na tabela está como v_2). Portanto, vamos olhar o valor na coluna 2 e na linha 12.

	A = 0,05										
	V1	1	2	3	4	5	10	15	25	30	60
V2	1	161,5	**199,5**	215,7	224,6	230,2	241,9	246	249,3	250,1	252,2
	2	18,51	**19**	19,16	19,25	19,3	19,4	19,43	19,46	19,46	19,48
	3	10,13	**9,55**	9,28	9,12	9,01	8,79	8,70	8,63	8,62	8,57
	...	...	**...**	...	...	...	...	...	...	...	...
	10	4,96	**4,10**	3,71	3,48	3,33	2,98	2,85	2,73	2,7	2,62
	12	4,84	**3,98**	3,59	3,36	3,2	2,85	2,72	2,6	2,57	2,49
	12	**4,75**	**3,89**	**3,49**	**3,26**	**3,11**	**2,75**	**2,62**	**2,5**	**2,47**	**2,38**
	...	...	**...**	...	...	...	...	...	...	...	...
	60	4	**3,15**	2,76	2,53	2,37	1,99	1,84	1,69	1,65	1,53
	120	3,92	**3,07**	2,68	2,45	2,29	1,91	1,75	1,60	1,55	1,43
	∞	3,84	**3**	2,61	2,37	2,21	1,83	1,67	1,51	1,46	1,32

Portanto, para um nível de significância de 95% temos: $F_{(2,12)} = 3,89$.

Esse é o valor do F crítico, se o F encontrado for maior que 3,89, ele está localizado depois do $F_{crítico}$, portanto rejeitamos H_0, e concluímos que pelo menos uma das médias é diferente. Ou seja, quando o F encontrado é maior que o $F_{crítico}$ rejeitamos H_0, pois o valor estará na área cinza do gráfico. Já quando o F encontrado é menor que o $F_{crítico}$ aceitamos H_0, pois o valor estará na área branca do gráfico.



F_a (v1 , v2)

Com os valores da ANOVA podemos achar também o **Coeficiente de Determinação (CD), ou R^2,** que é dado por:

$$R^2 = \frac{SQE}{SQT}$$

Outro valor da regressão que podemos encontrar utilizando os dados da tabela da ANOVA é o coeficiente angular do modelo de regressão (que chamamos de "a" anteriormente). Esse valor é dado por:

$$a^2 = \frac{SQE}{\sum(Xi - \bar{X})^2}$$

Onde SQE é a soma dos quadrados entre as amostras, ou do modelo, e $\sum (X_i - \bar{X})^2$ é a soma dos quadrados das diferenças entres os valores da amostra e a média, termo que aparece na definição de variância da amostra no começo deste material (estatística descritiva).

Hora de praticar o que aprendeu com questões comentadas de bancas variadas o assunto estudado:

1. **(CESPE-CEBRASPE – 2018)** Ao avaliar o efeito das variações de uma grandeza X sobre outra grandeza Y por meio de uma regressão linear da forma $\hat{P} = \hat{\alpha} + \hat{\beta} X$, um analista, usando o método dos mínimos quadrados, encontrou, a partir de 20 amostras, os seguintes somatórios (calculados sobre os vinte valores de cada variável):
$\sum X = 300$; $\sum Y = 400$; $\sum X^2 = 6.000$; $\sum Y^2 = 12.800$ e $\sum (XY) = 8.400$
A partir desses resultados, julgue o item a seguir.
Para X = 10, a estimativa de Y é $\hat{Y} = 12$.

( ) CERTO ( ) ERRADO

Levando em consideração que temos 20 amostras, podemos achar a média de $\hat{Y}$ e a média de X, tendo em vista os somatórios dados no enunciado.

$\overline{Y} = \frac{400}{20}$
$\overline{Y} = 20$
Média de X.
$\overline{X} = \frac{300}{20}$
$\overline{X} = 15$
Vamos achar o valor do coeficiente angular , usando os valores dados no enunciado.

$$\hat{\beta} = \frac{n \cdot \sum X \cdot Y - \sum X \cdot \sum Y}{n \cdot \sum X^2 - (\sum X)^2}$$

$$\hat{\beta} = \frac{20 \cdot 8400 - 300 \cdot 400}{20 \cdot 6000 - (300)^2}$$

$$\hat{\beta} = \frac{168.000 - 120.000}{120.000 - 90.000}$$

$$\hat{\beta} = \frac{48.000}{30.000}$$

$\hat{\beta} = 1,6$

Assim podemos achar o valor do coeficiente linear, $\hat{\alpha}$.
$\overline{Y} = \hat{\alpha} + \hat{\beta}$
$20 = \hat{\alpha} + 1,6 \,.\, 15$
$20 = \hat{\alpha} + 24$
$\hat{\alpha} = 20 - 24$
$\hat{\alpha} = -4$
Portanto a regressão fica: $= -4 + 1,6.X$
Se X = 10;
$\overline{Y} = -4 + 1,6 \,.\, 10$
$\overline{Y} = -4 + 16$
$\overline{Y} = 12$
Resposta: Certo.

2. (CESPE-CEBRASPE – 2019) Considerando-se que, em uma regressão múltipla de dados estatísticos, a soma dos quadrados da regressão seja igual a 60.000 e a soma dos quadrados dos erros seja igual a 15.000, é correto afirmar que o coeficiente de determinação — R^2 — é igual a

a) 0,75.
b) 0,25.
c) 0,50.
d) 0,20.
e) 0,80.

Sabemos que R^2 é dado por:

$$R^2 = \frac{SQE}{SQT}$$

O SQE é a soma dos quadrados da regressão ou do modelo (Entre amostras), e o SQT é a soma do SQE com o SQD, soma dos quadrados dos erros (Dentro da amostra).
SQT = 60.000 + 15.000
SQT = 75.000

$$R^2 = \frac{60.000}{75.000}$$

$$R^2 = \frac{60}{75}$$

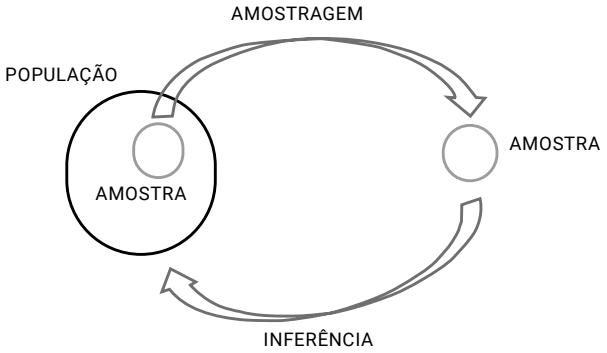
$R^2 = 0,8$
Resposta: Letra E.

POPULAÇÃO E AMOSTRA

AMOSTRA

É comum ouvirmos por aí pessoas que dizem não acreditar em pesquisas como Ibope, Datafolha entre outras, com o argumento de que nunca foi entrevistado por esses órgãos e que não conhecem nenhuma pessoa que tenha sido consultada também. Ora, consultar a população toda acerca de algum assunto é muito custoso e demandaria um tempo muito grande. Por isso, para se chegar a essa estimativa, esse tipo de pesquisa é realizado em parte da população.

Para analisar parâmetros estatísticos de uma população é necessário selecionar uma amostra a partir de técnicas de amostragem. Dessa amostra são extraídos dados estatísticos, que serão estudados e analisados, e então, generalizados para toda a população por meio da **Inferência Estatística.**



Quando vamos tratar de inferência estatística, o nosso conjunto de dados a ser analisado são os dados extraídos das amostras da população. Para isso, a amostra selecionada deve ser representativa da população.

Uma amostra **não representativa** é uma **amostra viciada**, que chamamos de **vício de amostragem**. Utilizando esse tipo de amostra chegaríamos a um resultado que não corresponde à realidade.

Ex.: Queremos saber a quantidade de torcedores de cada time na cidade de São Paulo, portanto nossa população é o total de habitantes de São Paulo. Vamos supor que a amostra selecionada para responder fosse feita na porta do estádio do Palmeiras em dia de jogo. Nesse caso, praticamente 100% dos entrevistados se declarariam palmeirenses, o que, claramente, está incorreto, pois a escolha da amostra foi realizada de forma incorreta.

Conceitos importante:

- **População objeto:** é a população total que temos o interesse de obter informações;
- **Unidade elementar (unidade amostral):** é a unidade (indivíduo) da população que será observada (analisada), pode ser uma pessoa, uma peça de uma produção etc.
- **Amostra:** é o subconjunto da população.
- **Censo:** análise de todos os elementos da população.
- **Erro amostral:** é a diferença entre o resultado amostral para o resultado populacional.

Existem várias formas de se determinar uma amostra da população a ser estudada, que se dividem em dois grandes grupos:

PROBABILÍSTICAS	NÃO-PROBABILÍSTICAS
Quando todos os elementos da população possuem a mesma probabilidade de serem selecionados. Só nas amostras probabilísticas podem ser utilizados os métodos de inferência. Alguns exemplos são a amostra aleatória simples, a sistemática, a estratificada e amostra por conglomerado.	Quando os elementos da população não possuem a mesma probabilidade de serem escolhidos, pois dependem do critério do pesquisador. Esse método tem uma desvantagem por não ser possível fazer inferências sobre a população. Apesar disso, ainda é um método muito utilizado. Alguns exemplos são a amostra por conveniência, intencional, amostra por cotas e voluntários.

AMOSTRA PROBABILÍSTICAS

Aleatória Simples

É o método mais utilizado, que consiste na escolha aleatória dos indivíduos de uma determinada população que formarão a amostra. Um detalhe importante desse método é que é preciso conhecer os dados de toda a população para, a partir de uma listagem, por exemplo, sejam definidos aleatoriamente os indivíduos estudados.

Ex.: Em uma sala de aula com 100 alunos queremos escolher 15 para responderem a uma pesquisa de satisfação dos professores. Portanto, a partir da lista dos 100 alunos da sala, selecionamos 15 aleatoriamente, enumerando cada aluno e fazendo um sorteio, como um bingo, por exemplo, ou até por algum sistema específico randômico, como a funcionalidade *Random* do Excel.

Sistemática

É uma variação da amostragem aleatória simples, na qual a partir da população ordenada (lista) escolhemos um critério para fazer a seleção.

Ex.: Em uma fábrica que produz um certo objeto, a cada 10 peças produzidas retiramos uma para compor nossa amostra, a fim de detectar possíveis erros de produção.

Estratificada

Consiste em dividir a população em grupos menores homogêneos, que chamamos de estratos, de forma que os elementos de cada estrato sejam mais homogêneos (pouca variabilidade) e os estratos sejam mais heterogêneos (grande variabilidade). Os estratos devem ser mutuamente exclusivos, ou seja, cada indivíduo deverá estar em apenas 1 estrato. Esses estratos podem ser divididos em idade, renda mensal, classe social, sexo, entre outros.

Dividida a população em estratos, é selecionada de forma aleatória simples parte dos elementos de cada estrato.

Essa escolha dos elementos de cada estrato pode ser feita de maneira:

- **Uniforme:** escolhendo, por exemplo, 10 pessoas de cada estrato, ou;
- **Proporcional:** obedecendo a representatividade de cada estrato na população, por exemplo, dividindo os estratos entre homem e mulher, se 30% da minha população é mulher e 70% de homens, minha amostra será formada por 30% de mulheres e 70% de homens.

Conglomerado

Consiste em dividir uma população em conglomerado (subgrupos) como: quarteirões, famílias, organizações, agências, edifícios etc., selecionar de forma aleatória simples alguns dos conglomerados e analisar todos os elementos do subgrupo escolhido. Nesse tipo de amostra as pessoas de cada conglomerado não necessariamente precisam ter características comuns, como na estratificada.

Importante!

As bancas geralmente tentam confundir as amostras estratificada e por conglomerado, mas é bem fácil de diferenciar as duas.
Na amostra estratificada, a população é dividida em grupos e, de **todos os grupos**, são selecionados **alguns elementos** de forma aleatória.
Já na amostra por conglomerado, a população é dividida em grupos, e são selecionados **alguns grupos**, que terão **todos os elementos** analisados.

AMOSTRA NÃO-PROBABILÍSTICAS

Aleatória Acidental ou por Conveniência

Esse tipo de amostra é a menos rigorosa de todas, usado nas pesquisas de opinião, quando os entrevistadores ficam em lugares com grande fluxo de pessoas, como mercado, metrô e calçadões, para fazer teste de produtos, entrevistando pessoas ao acaso de forma acidental.

Intencional

São selecionadas pessoas que os pesquisadores julgam ser relevantes para a pesquisa, como no caso da amostra aleatória, mas com alguns filtros. Ex: Em uma pesquisa sobre conforto de bonés, o entrevistador conversa apenas com pessoas que usam boné na rua.

Por Cotas

Essa amostra é parecida com as probabilísticas estratificada e conglomerado, pois é necessário dividir a população em grupos e determinar a proporção da população de cada grupo, e fixar cotas que serão extraídas de cada grupo proporcionalmente às classes consideradas. A diferença é que usamos esse tipo de amostra quando não existe uma listagem da população alvo.

Ex: pesquisa eleitoral, no caso definem o percentual de pessoas em cada faixa de idade a ser entrevistada, e escolhem as pessoas de forma acidental até completar cada grupo de idade.

Voluntários

São pessoas que se voluntariam a participar da pesquisa. Ex.: teste de novos remédios, nos quais pessoas se voluntariam para serem cobaias.

DISTRIBUIÇÕES AMOSTRAIS

Após a realização das amostragens, e das pesquisas que queremos realizar, teremos amostras com dados sobre os assuntos de interesse da pesquisa. Com esses dados das amostras em mãos, faremos a inferência sobre os dados da população. Os dados extraídos das pesquisas podem ser: média, desvio-padrão, variância, proporção dentre outros valores.

Conforme vimos nas distribuições de probabilidades das variáveis aleatórias (discretas e contínuas), esses modelos probabilísticos são baseados em alguns parâmetros (como exemplo, o λ na distribuição de Poisson) que na prática são desconhecidos, assim precisamos estimar esses parâmetros com base nessas amostras aleatórias.

Vamos considerar uma população, de tamanho N (que pode ser uma população infinita), representada por uma variável X. Dessa população será sorteada uma amostra aleatória simples de tamanho "n". Cada sorteio de um elemento dá origem a uma variável aleatória que chamaremos de X_i, sendo que todos os sorteios são independentes e possuem a mesma distribuição de X.

Portanto, uma amostra aleatória simples da variável X é um conjunto de variáveis aleatórias X_1, X_2, ..., X_n, identicamente distribuídas.

Os valores da **população são chamados de parâmetros**, já os valores encontrados nas **amostras são chamados de estimadores**. Portanto, quando falarmos de um estimador, estaremos nos referindo a um valor amostral, já quando falarmos em parâmetros, estaremos nos referindo a um valor populacional.

Os parâmetros da população podem ser vários, mas, os mais comuns são a média, a variância, o desvio-padrão e a proporção. Vamos simbolizar um parâmetro qualquer por θ. Já os estimadores são as estimativas para esses parâmetros, e, para simbolizar que estamos tratando de um estimador, colocamos um ^ sobre o símbolo do parâmetro ($\hat{\theta}$).

Quando vamos usar os valores encontrados em uma amostra aleatória, podemos ter erros em relação ao parâmetro populacional, portanto definimos esse erro como a diferença do estimador para o parâmetro real da população.

$$e = \hat{\theta} - \theta$$

Chamamos esse erro de viés, ou vício, portanto, quando temos um **estimador não-viesado,** esse erro é 0, portanto: $\hat{\theta} = \theta$. Isso significa que um estimador não-viesado é igual ao parâmetro populacional.

Nas distribuições amostrais, assim como nas estimativas de máxima verossimilhança, **a média populacional** (parâmetro) pode ser dada pela **média amostral** (estimador), isso está associado ao Teorema Central do Limite, ou seja, podemos dizer que esses parâmetros são não-viesados. Assim como a proporção amostral também é um parâmetro não-viesado, e podem ser usados como proporção populacional. Portanto:

$$E(\bar{X}) = \mu$$

Já a variância do estimador é dada pela variância populacional dividido pelo tamanho da amostra.

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

Sendo uma amostra aleatória simples (X_1, X_2, ..., X_n), de tamanho n, de uma população representada por uma variável aleatória normal X (ou seja, com média μ e variância σ^2, simbolizada por $N(\mu,\sigma^2)$), então a distribuição amostral da média amostral X também será normal, e sua média será μ e sua variância será $\frac{\sigma^2}{n}$.

Simbolicamente temos: $X \sim N(\mu;\sigma^2) \Rightarrow \bar{X} \sim N(\mu; \frac{\sigma^2}{n})$

O desvio padrão da distribuição amostral é chamado de erro padrão ($EP(\bar{X})$ ou s), e o erro padrão da média é dado por:

$$EP(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Já o erro padrão da proporção amostral é dado por:

$$s = \sqrt{\frac{p.q}{n}}$$

A estimativa da variância da proporção amostral é dada por:

$$s^2 = \frac{p \,.\, q}{n}$$

Outra coisa muito cobrada é a fração amostral (f), que nada mais é que o tamanho da amostra (n) dividido pelo tamanho da população (N).

$$f = \frac{n}{N}$$

Não podemos esquecer que quando temos uma população finita e uma amostra sem reposição, temos que multiplicar o estimador pelo fator de correção: $\sqrt{\frac{N-n}{N-1}}$. Já nos casos de população finita com reposição, ou nos casos de população infinita não é necessário aplicar o fator de correção.

Os bons estimadores possuem quatro propriedades características:

- Suficiência – o estimador é suficiente se a informação tirada da amostra consegue mostrar tudo o que for possível sobre o parâmetro desconhecido, ou seja, qualquer outra informação não acarretará uma melhora da informação sobre o parâmetro.
- Não viés – esse já falamos anteriormente, é quando o erro na medida é nulo: $e = \hat{\theta} - \theta$.
- Consistência – considera que aumentando o tamanho da amostra, as estimativas tenderão ao valor desconhecido do parâmetro.

- Eficiência – o estimador será mais eficiente quando o erro quadrático médio (variância do erro) for o menor possível.

Pratiquemos, por fim, com exercícios comentados de diversas bancas que, em suas provas, já cobraram do conteúdo estudado.

1. **(CESPE-CEBRASPE – 2018)** Um estudo acerca do tempo (x, em anos) de guarda de autos findos em determinada seção judiciária considerou uma amostragem aleatória estratificada. A população consiste de uma listagem de autos findos, que foi segmentada em quatro estratos, segundo a classe de cada processo (as classes foram estabelecidas por resolução de autoridade judiciária.) A tabela a seguir mostra os tamanhos populacionais (N) e amostrais (n), a média amostral ($\hat{x}$) e a variância amostral dos tempos (s^2) correspondentes a cada estrato.

ESTRATOS	TAMANHOS POPULACIONAIS (N)	TAMANHOS AMOSTRAIS (N)	$\hat{x}$	S2
A	30.000	300	20	3
B	40.000	400	15	16
C	50.000	500	10	5
D	80.000	800	5	8
total	200.000	2.000	-	-

Considerando que o objetivo do estudo seja estimar o tempo médio populacional (em anos) de guarda dos autos findos, julgue o item a seguir.
A estimativa do tempo médio populacional da guarda dos autos findos é maior ou igual a 12 anos.

( ) CERTO ( ) ERRADO

Para achar o tempo médio temos que achar a média dentre todos os grupos, para isso vamos precisar multiplicar a coluna N pela coluna $\hat{x}$, para encontrar a soma total, e dividir pelo tamanho populacional todo.

ESTRATOS	TAMANHOS POPULACIONAIS (N)	TAMANHOS AMOSTRAIS (N)	$\hat{x}$	$N \cdot \hat{x}$	S2
A	30.000	300	20	20 x 30.000 = 600.000	3
B	40.000	400	15	15 . 40.000 = 600.000	16
C	50.000	500	10	10 . 50.000 = 500.000	5
D	80.000	800	5	5 . 80.000 = 400.000	8
Total	200.000	2.000	-	2.100.000	-

$E(X) = \frac{2.100.000}{200.000}$

$E(X) = \frac{21}{2}$

E(X) = 10,5 anos Resposta: Errado.

2. **(CESPE-CEBRASPE – 2018)** Uma pesquisa realizada com passageiros estrangeiros que se encontravam em determinado aeroporto durante um grande evento esportivo no país teve como finalidade investigar a sensação de segurança nos voos internacionais. Foram entrevistados 1.000 passageiros, alocando-se a amostra de acordo com o continente de origem de cada um — África, América do Norte (AN), América do Sul (AS), Ásia/Oceania (A/O) ou Europa. Na tabela seguinte, N é o tamanho populacional de passageiros em voos internacionais no período de interesse da pesquisa; n é o tamanho da amostra por origem; P é o percentual dos passageiros entrevistados que se manifestaram satisfeitos no que se refere à sensação de segurança.

ORIGEM	N	N	P
África	100.000	100	80
NA	300.000	300	70
AS	100.000	100	90
A/O	300.000	300	80
Europa	200.000	200	80
Total	**1.000.000**	**1.000**	$\mathbf{P_{pop}}$

Em cada grupo de origem, os passageiros entrevistados foram selecionados por amostragem aleatória simples. A última linha da tabela mostra o total populacional no período da pesquisa, o tamanho total da amostra e P_{pop} representa o percentual populacional de passageiros satisfeitos. A partir dessas informações, julgue o item a seguir.
Na situação apresentada, o desenho amostral é conhecido como amostragem aleatória por conglomerados, visto que a população de passageiros foi dividida por grupos de origem.

( ) CERTO ( ) ERRADO

Na amostragem por conglomerado a população é dividida em grupos menores, que são selecionados de forma aleatória simples e todos os elementos do grupo selecionado são analisados.
Nesse caso foram selecionados estratos (continente de origem) e selecionados de forma aleatória simples os elementos entrevistados, portanto uma amostragem estratificada. Resposta: Errado.

INTERVALO DE CONFIANÇA

Sempre que selecionamos uma amostra para calcular estimadores e obter parâmetros da população (média, variância, proporção etc.), teremos um erro embutido nesse valor.

Imaginemos uma população que tem uma média de altura μ, e, ao analisarmos uma amostra dessa população, encontremos uma média . Essa média amostral é um bom parâmetro para definirmos a média populacional, mas a média amostral encontrada não será totalmente idêntica à populacional, sempre teremos um erro aleatório. Nesse caso temos que quantificar o erro construindo um intervalo de confiança.

$$\bar{x} = \mu \pm e$$

Encontrando esse intervalo de confiança definiremos um grau de incerteza associado a nossa estimativa.

Para achar os intervalos de confiança, vamos considerar variáveis que possuem distribuição normal, e vamos utilizar as tabelas da normal padrão (Z), assim como t de *Student* e qui-quadrado. Voltaremos a usar o nível de significância, que aqui pode ser chamado de **coeficiente de confiança**.

O erro aleatório dos estimadores encontrado na amostra, em geral, estará atrelado ao grau de confiança (1 - α) que queremos e do tamanho da amostra selecionada.

Vamos supor que uma população tenha média de idade de 43 anos, e ao selecionarmos uma amostra vamos querer um erro de no máximo 1 ano, assim teremos que ter uma amostra de **n** elementos para isso.

Definido o erro, a média da amostra terá que resultar em: 43 ± 1, ou seja, um intervalo limitado a: [42 ; 44].

Nesse caso temos que selecionar uma amostra que nos dê uma média amostral entre 42 e 44 anos.

Nesse intervalo o erro é de 1 ano, pois estamos variando 1 ano acima da média e 1 ano abaixo da média. Por outro lado, a amplitude desse intervalo é de 2 anos, pois 44 – 42 = 2. Assim podemos dizer que a amplitude de um intervalo de confiança é o dobro do erro.

$$\mathbf{A = 2.e}$$

INTERVALO DE CONFIANÇA PARA MÉDIA

Quando sabemos o valor do desvio padrão populacional (σ), o erro é dado por:

$$\mathbf{e = Z_0 \cdot \frac{\sigma}{\sqrt{n}}}$$

Sendo que: "e" é o erro; "σ" é o desvio padrão populacional; "n" é o tamanho da amostra; "Z_0" também chamado de $Z_{\alpha/2}$, que é o valor na curva normal padronizada, atrelada ao nível de confiança da amostra.

Obs.: O valor mais clássico, que algumas bancas até exigem que você saiba de cabeça, é o coeficiente de confiança de 95%, ou seja,

$1 - \alpha = 0{,}95$
$\alpha = 0{,}05 = 5\%$

Portanto ao olharmos na tabela (aquela mesma da curva normal padronizada) o valor de $Z_{0{,}005/2} = Z_{0{,}025} = 1{,}96$, pois $P(-1{,}96<Z<1{,}96) = 0{,}95$, ou 95%.

Logo, quando tivermos um nível de confiança de 95% podemos usar diretamente o valor de $Z_0 = 1{,}96$, mas, caso o exercício forneça os dados, é sempre bom conferir.

Da fórmula do erro, podemos multiplicar os lados por 2, assim teremos:

$$\mathbf{2.e = 2 \cdot Z_0 \cdot \frac{\sigma}{\sqrt{n}}}$$

Sabemos que "2e" é a amplitude do intervalor de confiança, portanto:

$$\mathbf{A = 2 \cdot Z_0 \cdot \frac{\sigma}{\sqrt{n}}}$$

Com base na fórmula do erro, vamos chegar que o intervalo de confiança para a média é dado por:

$$\mathbf{\overline{X} \pm Z_0 \cdot \frac{\sigma}{\sqrt{n}}}$$

Sendo que:
$\overline{X}$ é a média amostral

Quando não sabemos o valor do desvio padrão populacional e tivermos uma amostra pequena (geralmente até 30 observações), podemos usar a tabela t de *Student*, nesse caso a fórmula será um pouco diferente.

$$\mathbf{\overline{X} \pm t_0 \cdot \frac{s}{\sqrt{n}}}$$

Nesse caso: "$\overline{X}$" é a média amostral; "S" é o desvio padrão amostral; "n" é o tamanho da amostra; "t_0" também chamado de $t_{n-1,\,\alpha/2}$, que é o valor na tabela t de *Student*, atrelada ao nível de confiança, e ao número de graus de liberdade (n – 1) da amostra.

Quando temos uma população, com desvio padrão constante, e para um mesmo grau de confiança, o erro será menor quanto maior for o tamanho da amostra. Considerando duas amostras diferentes (n_1 e n_2) de uma mesma população (mesmo desvio padrão populacional), e um grau de confiança igual para as amostras (mesmo Z_0) temos que:

$$\mathbf{e_1 \cdot \sqrt{n_1} = e_2 \cdot \sqrt{n_2}}$$

INTERVALO DE CONFIANÇA PARA PROPORÇÃO

Assim como construímos o intervalo de confiança para a média, podemos construir um intervalo de confiança para a proporção populacional, com base na proporção da amostra. Na proporção temos uma variável com distribuição binomial, onde temos apenas duas opções: ou um evento ocorre (sucesso) ou não ocorre (fracasso).

Da mesma forma que vimos anteriormente, a proporção amostral terá um erro aleatório, que nesse caso é dado por:

$$\mathbf{e = Z_0 \cdot \sqrt{\frac{p \cdot q}{n}}}$$

Sendo que: "e" é o erro; "p" é a proporção amostral para o caso favorável; "q" é "1 – p", ou seja, a proporção dos casos desfavoráveis; "n" é o tamanho da amostra; "Z_0" também chamado de $Z_{\alpha/2}$, que é o valor na curva normal padronizada, atrelada ao nível de confiança da amostra.

Multiplicando os lados por 2, teremos:

$$\mathbf{A = 2 \cdot Z_0 \cdot \sqrt{\frac{p \cdot q}{n}}}$$

Com base na fórmula do erro, vamos chegar que o intervalo de confiança para a proporção é dado por:

$$\mathbf{p \pm Z_0 \cdot \sqrt{\frac{p \cdot q}{n}}}$$

Para qualquer um dos intervalos de confiança (média e proporção) são dois tipos de exercício que se repetem mais, sendo um para achar o intervalo de confiança e outro para achar o tamanho da amostra. Para achar o tamanho da amostra, vamos usar a fórmula do erro para média ou proporção, e para achar o intervalo de confiança vamos usar a fórmula do intervalo de confiança da média ou da proporção.

TESTE DE HIPÓTESE

Um dos métodos de avaliação da inferência estatística é chamado de Teste de Hipótese. Consiste em considerar uma hipótese para um parâmetro **Populacional** (pode ser média, proporção etc.) e a partir dos dados recolhidos da amostra, testar se essa hipótese deve ser aceita ou rejeitada.

Por se tratar de uma decisão com base na análise de uma amostra, a decisão nunca terá 100% de certeza, sempre teremos um percentual de erro estatístico embutido nessa decisão, que é dado em um valor percentual.

Ex.: temos uma população e queremos saber a média de idade (μ). Vamos supor que, por algum motivo, a gente acredite que a média seja de 35 anos. Para testar se esse é realmente o valor, vamos tirar uma amostra da população e achar a média da amostra, ou **média amostral** ($\overline{X}$ ou $\mu_{\overline{X}}$), e, suponhamos que a média amostral encontrada foi de 33. Assim vamos fazer um teste de hipótese com base no valor da média amostral para testar se a média populacional realmente pode ser 35.

As hipóteses podem ser simples, quando ela especifica completamente a distribuição da população (exemplo: H: μ = 0), ou composta, quando ela não especifica completamente a distribuição da população (exemplo: H: μ > 0, ou H: μ < 0).

Para calcular o desvio padrão amostral ($\sigma_{\overline{X}}$) e a variância amostral ($\sigma_{\overline{X}}$)² a partir do valor populacional usamos:

$$\text{Variância: } (\sigma_{\overline{X}})^2 = \frac{\sigma^2}{n} \qquad \text{Desvio-padrão: } \sigma_{\overline{X}} = \frac{\sigma}{n}$$

A hipótese criada será sempre para a população e nunca para a amostra, os principais parâmetros usados nas hipóteses são:

μ: média populacional
p: proporção populacional
σ: desvio-padrão populacional
σ^2: variância populacional

Quando vamos fazer o teste de hipótese temos que criar duas hipóteses: a **hipótese nula (H_0),** que geralmente é uma hipótese simples**,** e a **hipótese alternativa (H_1 ou H_A),** que geralmente é uma hipótese composta. Essas duas hipóteses são complementares, ou seja, se uma é verdadeira a outra será falsa, e elas não podem possuir elementos em comum.

HIPÓTESE NULA	HIPÓTESE ALTERNATIVA
É a base do nosso teste e ela normalmente terá um sinal de igualdade, podendo ser: =, ≤ ou ≥.	A hipótese alternativa é o complemento da hipótese nula, e geralmente não tem o sinal de igual, podendo ser: ≠, < ou >.

Dito isso, vamos citar alguns exemplos de formação de hipóteses, considerando que k é um número real (k ∈ R):

$$\left.\begin{array}{l} H_0: \mu = k \\ H_1: \mu \neq k \end{array}\right\} \textbf{ bilateral}$$

$$\left.\begin{array}{llll} H_0: \mu = k & H_0: \mu = k & H_0: \mu \leq k & H_0: \mu \leq k \\ \text{ou} & \text{ou} & \text{ou} & \\ H_1: \mu > k & H_1: \mu < k & H_1: \mu > k & H_1: \mu < k \end{array}\right\} \textbf{ unilateral}$$

Dos conceitos ligados à lateralidade falarei mais para frente, mas já adianto aqui que, quando a hipótese alternativa tiver sinal de ≠, o teste será bilateral, e quando a hipótese alternativa tiver sinal de < ou >, o teste será unilateral.

Como estamos partindo de uma amostra para estimar os dados de uma população, podemos incorrer em dois tipos de erros, os quais chamamos de erro tipo I e erro tipo II. Como falei anteriormente, o teste é baseado na hipótese nula (H_0), portanto os **dois tipos de erros** serão baseados em H_0.

ERRO TIPO I	ERRO TIPO II
Quando rejeitamos H_0 (e consequentemente aceitamos H_1) sendo que a H_0 é verdade na realidade.	Quando aceitamos H_0 (e consequentemente rejeitamos H_1) sendo que H_0 é falsa na realidade.

Ex.: Em uma linha de produção, a média de produtos com defeito fabricados a cada dia é 1,5. Esse é um parâmetro populacional, mas, vamos considerá-lo desconhecido (como normalmente é). Vamos fazer um teste de hipótese, no qual nossa hipótese nula será que a média populacional é no máximo 2 (o que mostra que na realidade o teste deveria ser aceito, pois a média populacional é 1,5). Assim:

$H_0: \mu \leq 2$
$H_1: \mu > 2$

Supondo que ao realizar o teste fosse constatado que o mesmo foi rejeitado, o que significa que a hipótese nula foi rejeitada. Assim, o teste foi rejeitado quando na verdade deveria ter sido aceito, pois a hipótese nula era verdadeira. Nesse caso realizamos o erro tipo I.

Agora imagine que a média populacional fosse na verdade 3, e, ao realizar o teste, como realizado anteriormente, ele fosse aceito. Nesse caso o teste deveria ter sido rejeitado uma vez que a média populacional é maior que 2. Assim teremos cometido o erro tipo II.

A probabilidade de cometer o erro tipo I é dada por α, que chamamos de **nível de significância** (que normalmente é dado no exercício). A probabilidade de cometer o erro tipo II é dada por β, e associado a isso temos o **poder do teste**, que é "1 – β", ou seja, a probabilidade de rejeitar H_0 quando ela realmente é falsa.

Fique atento à tabela a seguir:

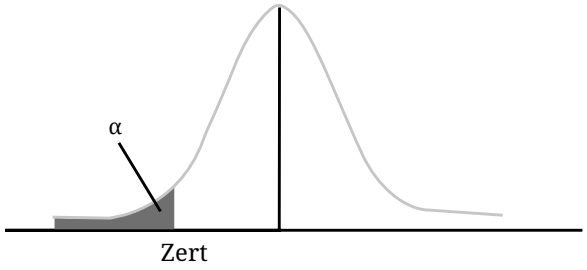
		REALIDADE	
		H_0 VERDADEIRA	H_0 FALSA
DECISÃO	ACEITAR H_0	Decisão correta (1 – α)	Erro do tipo II (β)
	REJEITAR H_0	Erro do tipo I (α) (nível de significância)	Decisão correta (1 – β) (poder do teste)

Em um teste, quando queremos diminuir a probabilidade de cometer o erro tipo I, estaremos aumentando a probabilidade de cometer o erro tipo II. A única forma de minimizar os dois erros é aumentando o tamanho da amostra. Podemos concluir que quanto maior a amostra menor a chance de erro, mas ele nunca será zero, a não ser que a nossa amostra seja grande o suficiente, ou seja, a própria população.

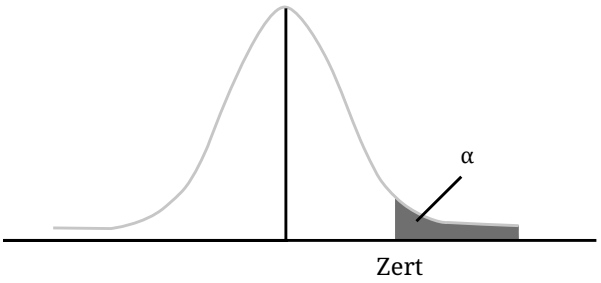
O nível de significância é o que define a aceitação ou rejeição do teste. Para poder analisar esse valor vamos usar a distribuição normal. Vamos supor que $\alpha = 1\%$, ou seja, a probabilidade de cometer o erro tipo I é de 1%.

Temos 3 opções diferentes de forma de montar a distribuição.

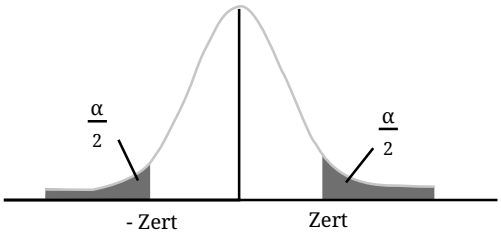
- Teste unilateral à esquerda:



- Teste unilateral à direita:



- Teste bilateral:



As regiões cinzas são chamadas de **Região Crítica (RC)** ou região de rejeição, e a parte branca é chamada de região de aceitação.

Para resolver uma questão de teste de hipótese vamos seguir 6 passos:

- **Escrever as hipóteses:**

  Hipótese nula: H_0. (utilizar os sinais =, ≥ ou ≤)
  Hipótese alternativa: H_1. (utilizar os sinais ≠, > ou <)

- **Calcular o valor observado ou estatística do teste** (Z_{obs}, t_{obs}), que é o valor que iremos usar ao final para fazer a comparação com a região crítica.

Para calcular esse valor vamos utilizar a tabela da normal padrão para o Z, ou a tabela t de *Student.*

As fórmulas para média são:

$$Z_{obs} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Onde σ é o desvio-padrão populacional.

$$t_{obs} = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Onde S é o desvio-padrão amostral.
Podemos calcular também o Z_{obs} para proporção:

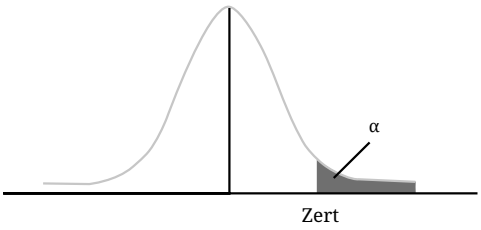
$$Z_{obs} = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}}$$

Em que: p é a proporção populacional e $\hat{p}$ é a proporção amostral.

A probabilidade de encontrarmos valores da amostra abaixo do Z_{obs} é chamado de P-valor.

Ex.: caso o valor encontrado para Z_{obs} seja -1, a probabilidade de encontrarmos um valor menor que -1 é dada pela tabela da normal padrão, e essa probabilidade é chamada de P-valor.

- **Fazer o gráfico** (unilateral esquerda ou direita ou bilateral). Para saber qual gráfico usaremos vamos considerar o sinal da hipótese alternativa.



- **Achar o valor crítico**, que pode ser observado na tabela da normal padrão (ou dado no exercício). E vamos encontrar o valor de Z que deixe a probabilidade do nível de significância na região crítica.
- Marcar o Z_{obs} no gráfico.
- Avaliar a aceitação e rejeição de H_0.

  Se Z_{obs} está dentro da Região Crítica, então rejeitamos H_0.
  Se Z_{obs} está fora da Região Crítica, então aceitamos H_0.

Algumas dicas importantes:

Quanto menor for o P-valor mais chance de rejeitar o H_0.

Se p-valor ≤ α, rejeitamos H_0 em favor de H_1.

Se p-valor > α, não rejeitamos H_0 em favor de H_1.

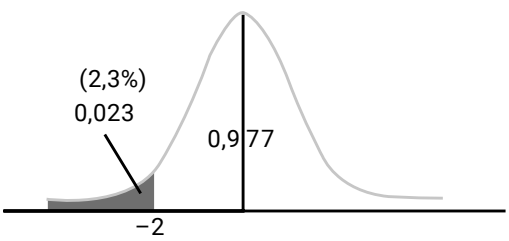
Quando **não** sabemos o valor do desvio-padrão populacional, vamos usar o teste t de Student. Para isso temos que saber apenas duas coisas: usar a tabela e que a quantidade de graus de liberdade é: gl = n – 1 (como vimos nas distribuições).

Hora de, com os exercícios comentados de bancas diversas, por a teoria em prática. Vamos lá!

1. (CESPE-CEBRASPE – 2016) Por meio de uma pesquisa, estimou-se que, em uma população, o percentual p de famílias endividadas era de 57%. Esse resultado foi observado com base em uma amostra aleatória simples de 600 famílias.
Nessa situação, considerando a hipótese nula H_0: p ≥ 60%, a hipótese alternativa H_1: p < 60% e P(Z ≤ 2) = 0,977, em que Z representa a distribuição normal padrão, bem como sabendo que o teste se baseia na aproximação normal, assinale a opção correta, a respeito desse teste de hipóteses.

a) O erro do tipo II representa a probabilidade de se rejeitar a hipótese nula, uma vez que, na realidade, p = 60%.
b) Com nível de significância α = 2,3%, a regra de decisão do teste é rejeitar a hipótese nula caso o percentual de famílias endividadas na amostra seja inferior a 56%.
c) Trata-se de um teste unilateral à direita.
d) A estatística do teste foi igual ou superior a 1.
e) A hipótese nula deve ser rejeitada caso a probabilidade de ocorrência de erro do tipo I seja igual ou inferior a 0,01.

Vamos escrever os dados apresentados e depois analisar cada alternativa.
$\hat{p}$ = 57%.
n = 600
H_0: p ≥ 60%
H_1: p < 60%
Alternativa a) Como a proporção na realidade é 60% e a nossa hipótese nula diz que p ≥ 60%, a hipótese nula é na realidade verdadeira. Caso no teste essa hipótese seja rejeitada, estaremos cometendo o erro tipo I, que é quando rejeitamos H_0 quando ele na verdade é verdadeiro. O erro tipo II é quando aceitamos H_0 quando ele é falso. Alternativa incorreta.
Alternativa b) Vamos fazer o teste de hipótese seguindo os passos.
1- H_0: p ≥ 60%
H_1: p < 60%
O 2º passo não faremos, pois não queremos o Z_{obs}, mas sim o valor da média amostral que está na fórmula do Z.
3 - Para saber o gráfico que vamos usar, precisamos olhar o sinal do H_1, que é p < 60%. Como o sinal é <, vamos usar o gráfico unilateral à esquerda.
4 - O nível de significância é 2,3%, portanto a região crítica tem 2,3%, ou seja, 0,023, na região branca temos 1 – 0,023 = 0,977. O exercício informa que P(Z ≤ 2) = 0,977, por simetria, sabemos que P(Z ≥ -2) = 0,977, ou seja:



Portanto, Z_{crit} = -2
5 – O valor que delimita a RC é -2, voltando ao passo 2, vamos achar qual o valor da média amostral que resulta em Z = -2.

$$Z_{obs} = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}}$$

$$-2 = \frac{\hat{p} - 0{,}6}{\sqrt{\frac{0{,}6 \cdot (1-0{,}6)}{600}}}$$

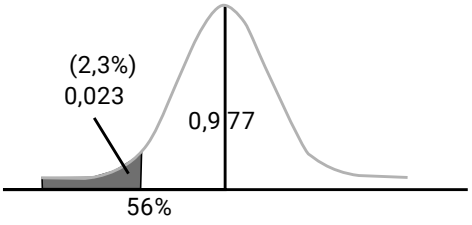
$$-2 = \frac{\hat{p} - 0{,}6}{\sqrt{\frac{0{,}6 \cdot 0{,}4}{600}}}$$

$$-2 = \frac{\hat{p} - 0{,}6}{\sqrt{\frac{0{,}24}{600}}}$$

$$-2 = \frac{\hat{p} - 0{,}6}{\sqrt{0{,}0004}}$$

$$-2 = \frac{\hat{p} - 0{,}6}{0{,}02}$$

$\hat{p}$ - 0,6 = - 0,04
$\hat{p}$ = 0,6 – 0,04
$\hat{p}$ = 0,56
$\hat{p}$ = 56%



Portanto, para p < 56% estaremos na Região Crítica da curva, portanto temos que rejeitar H_0. Alternativa correta.
Alternativa c) O teste é unilateral a esquerda. Alternativa incorreta.
Alternativa d) A estatística do teste (Z_{obs}) é dada por:

$$Z_{obs} = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}}$$

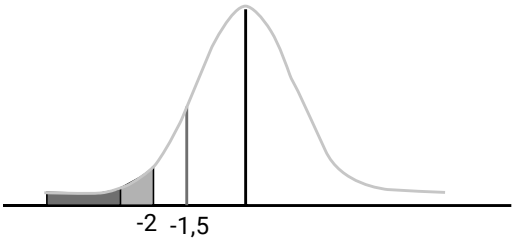
$$Z_{obs} = \frac{0{,}57 - 0{,}6}{\sqrt{\frac{0{,}6 \cdot (1-0{,}6)}{600}}}$$

$$Z_{obs} = \frac{-0{,}03}{0{,}02}$$

Z_{obs} = - 1,5

O valor é menor que 1. Alternativa incorreta.
Alternativa e) Sabemos que a RC para Z = -2 é de 2,3%, ou seja, o nível de significância.
Caso a o nível de significância fosse 1%, a região seria menor ainda. Pelo gráfico abaixo, a região cinza escuro seria = 1%, e a região compreendida pela soma das áreas cinzas seria = 2,3%. Como o Z_{obs} é

-1,5, ele fica à direita do -2. Considerado apenas a área cinza escura o ponto -1,5 está fora da RC, portanto temos que aceitar H_0. Alternativa incorreta.



Resposta: Letra B.

2. (CESPE-CEBRASPE – 2018) Em uma fábrica de ferragens, o departamento de controle de qualidade realizou testes na linha de produção de parafusos. Os testes ocorreram em dois campos: comprimento dos parafusos e frequência com que esse comprimento fugia da medida padrão. Historicamente, o comprimento médio desses parafusos é 3 cm, e o desvio padrão observado é 0,3 cm. Foram avaliados 10.000 parafusos durante uma semana. Desses, 1.000 fugiram às especificações técnicas da gerência: o comprimento do parafuso deveria variar de 2,4 cm a 3,6 cm. O chefe da linha de produção, porém, insiste em afirmar que, em média, 4% da produção de parafusos fogem às especificações. O departamento de controle de qualidade assume que os comprimentos dos parafusos têm distribuição normal.

A partir dessa situação hipotética, julgue o item subsequente, considerando que Φ(1) = 0,841, Φ(1,65) = 0,95, Φ(2) = 0,975 e Φ(2,5) = 0,994, em que Φ(z) é a função distribuição normal padronizada acumulada, e que 0,002 seja valor aproximado para $\sqrt{\frac{0,0384}{10.000}}$.

Com base nos dados apresentados, pode-se rejeitar, com significância de 5%, a afirmação do chefe da linha de produção.

( ) CERTO ( ) ERRADO

Na amostra de 10.000 parafusos, 1.000 fugiram às especificações, o que corresponde a 10%, portanto, 10% dos parafusos estão fora das especificações. O chefe diz que, em média, apenas 4% (0,04) fogem às especificações. Assim temos os seguintes dados:
$\hat{p}$ = 0,1 (proporção amostral)
n = 10.000
Passo 1 – Escrever hipóteses:
As hipóteses são:
H_0: p = 0,04
H_1: p ≠ 0,04
Temos um teste bilateral.
Passo 2 – calcular o Z_{obs}:
Vamos calcular o Z_{obs}.

$$Z_{obs} = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}}$$

$$Z_{obs} = \frac{0,1 - 0,04}{\sqrt{\frac{0,04 \cdot 0,96}{10,000}}}$$

$$Z_{obs} = \frac{0,06}{\sqrt{\frac{0,04 \cdot 0,96}{10,000}}}$$

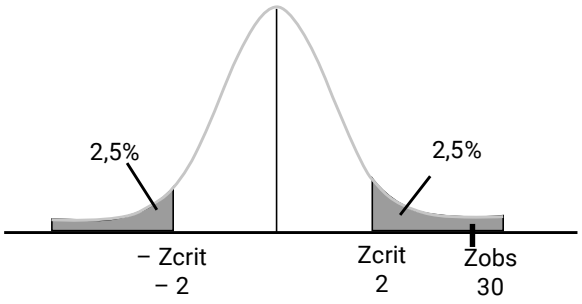
$$Z_{obs} = \frac{0,06}{\sqrt{\frac{0,0384}{10,000}}} \text{ (essa raiz foi dada no enunciado)}$$

$$Z_{obs} = \frac{0,06}{0,002}$$

$$Z_{obs} = 30$$

Passos 3, 4 e 5 – Desenhar o gráfico, encontrar o Z_{crit} e marcar o Z_{obs}:

Temos um teste bilateral com nível de significância de 5%, portanto em cada lado teremos 2,5%, assim vamos considerar Φ(2) = 0,975, sendo o Z_{crit} o -2 e o 2.



Como o Z_{obs} está na região crítica, vamos rejeitar H_0.
Resposta: Certo.

HORA DE PRATICAR!

1. (CESGRANRIO – 2013) A tabela a seguir apresenta a distribuição dos clientes de uma determinada agência bancária classificados segundo o perfil do investidor em: conservadores, moderados e arrojados.

CLASSIFICAÇÃO DOS CLIENTES	FREQUÊNCIA ABSOLUTA
Total	11.000
Conservadores	3.300
Moderados	5.400
Arrojados	2.300

Considere as medidas estatísticas: média, moda, mediana, variância e desvio padrão.

Para análise da classificação dos clientes, é possível determinar a:

a) Moda, apenas.
b) Média e a mediana, apenas.
c) Média, a moda e a mediana, apenas.
d) Média, a variância e o desvio padrão, apenas.
e) Média, a moda, a mediana, a variância e o desvio padrão.

2. (CESGRANRIO – 2014) Observe as afirmações a seguir relativas a histograma e a gráfico de ramo e folha.

I. Histogramas serão mais úteis do que gráfico de ramo e folha para mostrar quaisquer observações que estejam bem afastadas da maioria dos dados, se os gráficos forem construídos com um número suficiente de intervalos de classe.
II. Se um gráfico de ramo e folha ou um histograma utilizar uma escala muito expandida, apresentará o comportamento de um gráfico de pontos, em vez de mostrar as densidades relativas dos dados.
III. Na construção de um modelo estatístico para o processo que descreve os dados, o histograma pode sugerir uma função matemática cuja curva se ajusta bem ao histograma.

Está correto **apenas** o que se afirma em

a) **I**.
b) **II**.
c) **III**.
d) **I** e **III**.
e) **II** e **III**.

3. (CESGRANRIO – 2018) Em uma avaliação na qual é atribuído grau de zero a dez, um hotel obteve média 8 em quarenta e nove avaliações. O avaliador seguinte atribuiu ao hotel nota zero. Para que a média de notas do hotel passe a ser maior que 8, será necessário, no mínimo, a avaliação de mais quantos hóspedes?

a) 1.
b) 2.
c) 3.
d) 4.
e) 5.

4. (CESGRANRIO – 2014) A seguir são apresentadas estatísticas das notas brutas obtidas pelos candidatos em um concurso público:

Média aritmética: 78

Variância: 100

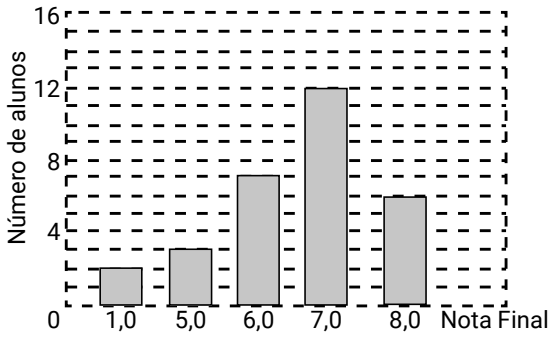
A nota de cada candidato foi transformada em nota padronizada, calculada considerando-se a seguinte fórmula:

$$\textbf{Nota padronizada} = 50 + 5 \times \frac{\text{Nota bruta do candidato} - \text{Média aritmética das notas brutas}}{\text{Desvio padrão das notas brutas}}$$

A média das notas padronizadas é

a) 0.
b) 28.
c) 50.
d) 55.
e) 78.

5. (CESGRANRIO – 2010) As notas finais dos alunos de determinado curso estão representadas no gráfico a seguir.



A média da turma foi, aproximadamente:

a) 5,8.
b) 6,0.
c) 6,2.
d) 6,4.
e) 6,6.

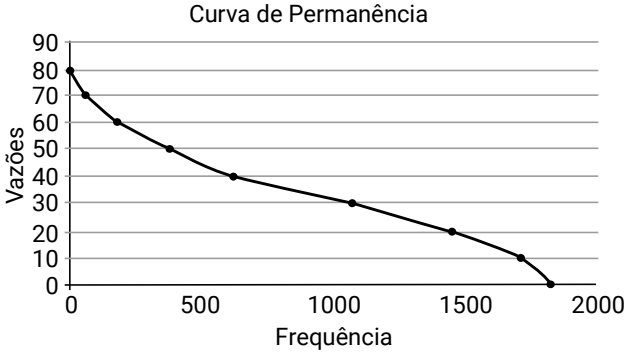
6. **(CESGRANRIO – 2018)** Sabe-se que 30% dos clientes de um banco são do sexo masculino e os 70% restantes são do sexo feminino. Entre os clientes do sexo masculino, a média do tempo de vínculo com o banco é igual a 4 anos e, entre os clientes do sexo feminino, é igual a 6 anos.
Considerando-se todos os clientes, de ambos os sexos, qual é a média do tempo de vínculo de cada um com o banco?

a) 5 anos.
b) 5,3 anos.
c) 6 anos.
d) 5,4 anos.
e) 5,7 anos.

7. **(CESGRANRIO – 2010)** Considere que tenha sido realizado um levantamento do tempo gasto para o abastecimento dos carros em um posto de combustíveis. Foi escolhida aleatoriamente uma amostra de 4 carros em um determinado posto e observado o tempo que gastavam para abastecer. O resultado, em minutos, foi o seguinte: 5; 2; 10 e 5. Qual a média harmônica do tempo gasto para o abastecimento dos carros neste posto?

a) 0,05.
b) 0,25.
c) 1.
d) 4.
e) 5,5.

8. **(CESGRANRIO – 2012)** A figura a seguir apresenta a curva de permanência de vazões decrescentes construída com os dados de vazões observados na foz de uma bacia hidrográfica. A tabela mostra os dados usados para obtenção da curva de permanência. Baseando-se nesses dados, qual o valor, em m3/s, da vazão modal?



INTERVALO DE VAZÃO M3/S	FREQUÊNCIA ACUMULADA
0-10	1822
0-20	1712
20-30	1450
30-40	1070
40-50	620
50-60	380
60-70	180
70-80	60

a) 15.
b) 25.
c) 35.
d) 45.
e) 55.

9. **(CESGRANRIO – 2013)** Seja X uma variável aleatória com distribuição normal cuja média é μ e o desvio padrão é σ.
Se Y = 2X - 1 tem distribuição normal com média 5 e variância 20, o coeficiente de variação populacional $\frac{\sigma}{\mu}$:

a) $\frac{\sqrt{42}}{6}$
b) $\frac{\sqrt{21}}{6}$
c) $\frac{\sqrt{5}}{3}$
d) $\frac{\sqrt{39}}{9}$
e) $\frac{4\sqrt{5}}{9}$

10. **(CESGRANRIO – 2018)** Há dez anos a média das idades, em anos completos, de um grupo de 526 pessoas era de 30 anos, com desvio padrão de 8 anos.
Considerando-se que todas as pessoas desse grupo estão vivas, o quociente entre o desvio padrão e a média das idades, em anos completos, hoje, é:

a) 0,45.
b) 0,42.
c) 0,20.
d) 0,27.
e) 0,34.

11. **(CESGRANRIO – 2012)** Numa distribuição assimétrica positiva, os valores da média, da moda e da mediana são tais que:

a) Moda < mediana < média.
b) Moda < média < mediana.
c) Média < moda < mediana.
d) Média < mediana < moda.
e) Mediana < média < moda.

12. **(CESGRANRIO – 2018)** Para montar uma fração, deve-se escolher, aleatoriamente, o numerador no conjunto N = {1,3,7,10} e o denominador no conjunto D = {2,5,6,35}. Qual a probabilidade de que essa fração represente um número menor do que 1(um)?

a) 50%
b) 56,25%
c) 25%
d) 75%
e) 87,5%

13. **(CESGRANRIO – 2018)** Em uma fábrica existem três máquinas (M1, M2 e M3) que produzem chips. As máquinas são responsáveis pela produção de 20%, 30% e 50% dos chips, respectivamente. Os percentuais de chips defeituosos produzidos pelas máquinas M1, M2 e M3 são 5%, 4% e 2%, respectivamente. Ao se retirar aleatoriamente um chip, constata-se que ele é defeituoso; então, a probabilidade de ele ter sido produzido pela máquina M1 é de, aproximadamente:

a) 0,025.
b) 0,032.

c) 0,31.
d) 0,55.
e) 0,78.

14. (CESGRANRIO – 2011) Utilize as informações da reportagem a seguir para responder à questão.
SÃO PAULO. Quatro entre nove brasileiros já têm computador em casa ou no trabalho. (...) É o que revela a 22ª Pesquisa do Centro de Tecnologia de Informação Aplicada da Fundação Getúlio Vargas (...). De acordo com o levantamento, existem 85 milhões de computadores no Brasil. No ano passado, foram vendidos 14,6 milhões de unidades. (...)

Jornal O Globo, Rio de Janeiro, p. 27, 20 abr. 2011.

Considere que a pesquisa da Fundação Getúlio Vargas foi feita entrevistando pessoas e perguntando se possuíam, ou não, computador. Suponha que, dentre os entrevistados que declararam ainda não ter computador, três em cada cinco tenham a intenção de adquiri-lo nos próximos 12 meses.

Escolhendo-se, ao acaso, uma das pessoas que participaram da pesquisa, a probabilidade de que a pessoa escolhida não tenha computador mas pretenda adquirir um nos próximos 12 meses é de, aproximadamente,

a) 24%
b) 33%
c) 40%
d) 52%
e) 60%

15. (CESGRANRIO – 2013) João e Maria estão enfrentando dificuldades em algumas disciplinas do 1º ano do Ensino Médio. A probabilidade de João ser reprovado é de 20%, e a de Maria é de 40%.
Considerando-se que João e Maria são independentes, qual é a probabilidade de que um ou outro seja reprovado?

a) 0.
b) 0,2.
c) 0,4.
d) 0,52.
e) 0,6.

✓ GABARITO

1	A
2	E
3	E
4	C
5	E
6	D
7	D
8	C
9	C
10	C
11	A
12	B
13	C
14	B
15	D

ANOTAÇÕES