



ESTADÍSTICA

UVA●

UNIJORGE



ESTATÍSTICA

Copyright © UVA 2020

Nenhuma parte desta publicação pode ser reproduzida por qualquer meio sem a prévia autorização desta instituição.

Texto de acordo com as normas do Novo Acordo Ortográfico da Língua Portuguesa.

AUTORIA DO CONTEÚDO

Adriana Maria Balena Tostes

PROJETO GRÁFICO

UVA

REVISÃO

Janaina Vieira

Lydianna Lima

DIAGRAMAÇÃO

UVA

T716 Tostes, Adriana Maria Balena.
Estatística [recurso eletrônico] / Adriana Maria Balena Tostes. – Rio de Janeiro: UVA, 2021.

1 recurso digital (3621 KB)

Formato: PDF

ISBN 978-65-5700-089-2

1. Estatística. 2. Probabilidades. I. Universidade Veiga de Almeida. II. Título.

CDD – 519.5

Bibliotecária Adriana R. C. de Sá CRB 7 – 4049.
Ficha Catalográfica elaborada pelo Sistema de Bibliotecas da UVA.

SUMÁRIO

Apresentação **6**

Autor **7**

UNIDADE 1

Introdução ao estudo estatístico **8**

- A estatística no Brasil e no mundo
- Elaboração e análise de gráficos e tabelas
- Medidas de posição e de dispersão

UNIDADE 2

Probabilidade **93**

- Introdução à probabilidade
- Conceitos de probabilidade
- Probabilidade condicional

SUMÁRIO

UNIDADE 3

Distribuições de Probabilidade **166**

- Variáveis aleatórias e distribuições de probabilidade
- Distribuições de probabilidades discretas
- Distribuição de probabilidade normal

UNIDADE 4

Intervalos de confiança e Relação entre variáveis **228**

- Intervalos de confiança
- Correlação
- Regressão linear

APRESENTAÇÃO

O mundo em que vivemos está repleto de uma quantidade extraordinária de informações. Sozinhas, essas informações podem gerar grandes confusões ou simplesmente podem não ter significado algum.

Para coletar, organizar, analisar, apresentar, interpretar e tirar conclusões sobre elas precisamos de um ramo poderoso da Matemática Aplicada, que denominamos de Estatística.

A Estatística é ferramenta essencial para fazer o tratamento de informações, que denominamos de tratamento dos dados. É considerada disciplina-chave da Teoria de Informação, tendo como objetivo principal subsidiar o processo decisório. Ela é indispensável para qualquer profissional que necessita analisar informações em suas tomadas de decisão diárias, seja no campo profissional ou na vida pessoal.

Não podemos escapar dos dados, assim como não podemos evitar o uso de palavras. Tal como palavras os dados não se interpretam a si mesmos, mas devem ser lidos com entendimento. Da mesma maneira que um escritor pode dispor as palavras em argumentos convincentes ou frases sem sentido, assim também os dados podem ser convincentes, enganosos ou simplesmente inócuos. A instrução numérica, a capacidade de acompanhar e compreender argumentos baseados em dados, é importante para qualquer um de nós. O estudo da estatística é parte essencial de uma formação sólida. (MOORE, 2000)

Seja qual for a área ou o objeto de estudo do pesquisador, este poderá utilizar conceitos de Estatística. É difícil encontrar uma situação real em que não se utilize a Estatística para solucionar problemas, gerenciar qualidade, comparar metodologias, testar hipóteses, evidenciar tendências, diminuir riscos e, principalmente, auxiliar na tomada de decisões.

ADRIANA MARIA BALENA TOSTES

Graduada em Engenharia Civil pela Universidade Federal de Juiz de Fora – UFJF, licenciada em Matemática pela Universidade Federal de Santa Catarina – UFSC e mestra em Educação Matemática pela Universidade Severino Sombra – USS. Iniciou sua trajetória docente no Ensino Superior em 1991, ingressando como professora da Universidade Veiga de Almeida – UVA em 2010. Ministra aulas de Cálculo Elementar, Cálculo I, Cálculo II, Matemática I, Matemática II, Matemática Financeira e Estatística, para os cursos de Engenharias, Administração e Ciências Contábeis.

UNIDADE 1

Introdução ao estudo estatístico

INTRODUÇÃO

Nesta unidade apresentaremos os principais conceitos da Estatística, bem como os planejamentos para coleta de dados, organização por meio de tabelas, apresentação a partir de gráficos e cálculo das medidas de resumo, tanto as medidas de posição quanto as medidas de dispersão, os quais compõem a estatística descritiva. O que veremos nos tópicos, a saber:

- **Tópico 1:** um breve histórico enfocando a relevância da Estatística no mundo em que vivemos e algumas situações que exemplificam o mau uso dessa área do conhecimento humano. Seu estudo se desenvolverá ancorado em conceitos básicos como dados, população e amostra, censo e amostragem, parâmetros e estatísticas, tipos de variáveis, tipos de amostragens, índices e indicadores.
- **Tópico 2:** a organização e a visualização dos conjuntos de dados, utilizando tabelas, sejam elas com ou sem intervalos de classes, e gráficos de diversos tipos, como os histogramas, os polígonos de frequências, a Ogiva de Galton, além dos gráficos de barras, colunas, setores e linhas, associando o uso de tecnologia para essa elaboração. Esses conhecimentos e habilidades são fundamentais para o desenvolvimento de competências associadas à resolução de problemas que envolvam a tomada de decisão a partir da organização, apresentação e análise de dados.
- **Tópico 3:** medidas descritivas, divididas em medidas de posição, que são a média, a moda e a mediana, e as medidas de dispersão, que são a amplitude, a variância, o desvio-padrão e o coeficiente de variação, fundamentais para resumir as informações.



OBJETIVO

Nesta unidade você será capaz de:

- Realizar a coleta, a organização, a descrição e a análise dos dados referentes a uma pesquisa.

A estatística no Brasil e no mundo

Podemos pensar que os conceitos de Estatística nasceram no mundo contemporâneo, em que valorizamos cada vez mais a rapidez e a quantidade de informações, quando o avanço tecnológico é incessante. Entretanto, sabemos que a utilização dos conceitos da Estatística como suporte para a tomada de decisões é verificada também no mundo antigo.

O primeiro registro disponível acerca de um levantamento estatístico refere-se à solicitação de Heródoto que, em 3050 a.C., no Egito, promoveu um estudo populacional com a intenção de quantificar a disponibilidade de recursos humanos e econômicos para a construção das pirâmides. Determinados autores apontam ainda que anos antes, em 5000 a.C., já havia ocorrido um levantamento para quantificar os egípcios presos de guerra.

Podemos dividir a evolução da Estatística em três fases. São elas:

1ª fase	Ocorre na Idade Média, quando as coletas de informações tinham finalidades tributárias ou bélicas.
2ª fase	Por volta do século XVI iniciam-se as primeiras análises sistemáticas dos fatos sociais, tais como casamentos, óbitos, nascimentos, migração, entre outros.
3ª fase	No século XVIII a Estatística foi batizada por Godofredo Achenwall como nova ciência, estabelecendo, assim, seus objetivos e relacionamentos com as demais ciências

Agora veremos a evolução da estatística no Brasil até o momento atual. Observe a tabela a seguir.

Destaques no Brasil

1585

Padre José de Anchieta registra os habitantes de algumas capitanias.

1872

Primeiro censo geral da população brasileira feito por José Maria da Silva Paranhos, conhecido como Visconde do Rio Branco.

1907

Criação do Conselho Superior de Estatística, com vistas à padronização de conceitos e apuração de resultados em todo o território nacional.

1934

Criação do Instituto Nacional de Estatística, que só passou a existir de fato em 1936, mudando em 1938 para Instituto Brasileiro de Geografia e Estatística – IBGE, que traz uma série de informações referentes à população, economia e uma variedade de indicadores sociais.

1937

Criação do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP (www.inep.gov.br), que apresenta uma vasta quantidade de informações referentes à área educacional, trazendo resultados de censos escolares, avaliações institucionais; sinopses estatísticas e microdados.

1940

Iniciaram os “modernos censos” decenais, não ocorrendo apenas o de 1990 (foi adiado para 1991), devido à “falta de recursos” alegada pelo governo Collor. Antes disso ocorreram os de 1872, 1890, 1900 e 1920.

1947

Primeiro curso de Inferência dado no Brasil baseado no livro de Cramer.

1953

Duas escolas iniciaram o Ensino de Estatística no Brasil: uma no Rio de Janeiro, a Escola Nacional de Ciências Estatística (ENCE) e a outra conhecida como Escola de Estatística da Bahia.

1955

O Brasil teve a honra de receber Sir Ronald Aylmer Fisher para participar do 2º Congresso Internacional de Biometria, realizado em Campinas.

1964

Criação do Instituto de Pesquisa Econômica Aplicada – IPEA (www.ipea.gov.br), que possui muitos indicadores e dados referentes à área econômica e financeira.

1970

Iniciou-se a formação de grupos de pesquisa em probabilidades, propiciando um dos grandes passos para a criação de outros cursos nessa área.

1992

Registram-se 25 universidades em todo o país com cursos de graduação e pós-graduação em Estatística.

2020

Podemos notar um aumento significativo na oferta de cursos. Existem 56 cursos de graduação, 115 pós-graduações Lato Sensu e nove cursos de mestrado e doutorado em Estatística no Brasil.

Como podemos observar os conceitos da Estatística estão presentes na vida do homem desde a Antiguidade, no entanto, atualmente a utilização da Estatística expandiu-se para muito além de suas origens. Indivíduos e organizações usam a estatística para compreender dados e tomar decisões bem informadas em ciências biológicas, ciências exatas, ciências sociais, ciências humanas, em negócios e em inúmeras áreas do conhecimento. Com o surgimento dos computadores e de softwares estatísticos específicos para trabalhar com dados, o estudo ficou muito mais dinâmico e interessante, de forma exata e precisa.

Conceitos fundamentais

Agora, vamos conhecer alguns conceitos fundamentais da Estatística.

1. Variáveis

É a característica que será observada em uma pesquisa.



Exemplo

Sexo, idade, massa corporal, grau de instrução.

2. Dados

São as observações coletadas a respeito de uma determinada variável.



Exemplo

Feminino, masculino, 20 anos, 65 Kg, superior completo.

3. População e Amostra

Em alguns casos seria impossível entrevistar todos os elementos para coletarmos os dados, pois levaria muito tempo para concluir o trabalho ou até mesmo seria inviável financeiramente. Dessa forma, com técnicas robustas de amostragem, escolhemos uma parte significativa para fazer a coleta desses dados.

Você já imaginou se, ao fazer uma pesquisa eleitoral em nosso país, tivéssemos que entrevistar todos os eleitores do Brasil?

Quando fazemos a coleta dos dados em situações em que **todos** os elementos são analisados, estamos trabalhando com uma “**população**”.

Quando fazemos a coleta dos dados em situações em que **uma parte ou subgrupo é considerado**, estamos trabalhando com uma “**amostra**”.

Entenda melhor esses conceitos.

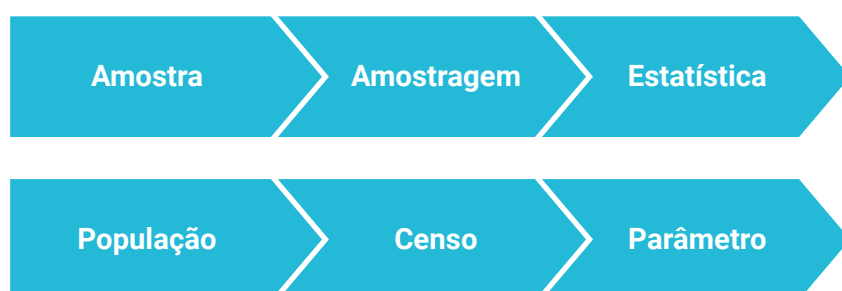
População:

É o conjunto de todos os elementos sobre o qual desejamos obter informação.

Amostra:

É um subconjunto de elementos retirados da população para obter-se a informação desejada.

Ao referirmo-nos a uma população fazemos um recenseamento e o dado estatístico é considerado parâmetro. No caso de uma amostra, fazemos uma pesquisa por amostragem e o dado estatístico é denominado estatística, como podemos observar na figura a seguir.



Ampliando o foco



O Censo é uma pesquisa realizada no Brasil pelo IBGE, em que todos os domicílios brasileiros são visitados para que se possa saber o número de brasileiros, quem são, onde vivem e como vivem.

Para saber mais sobre o recenseamento visite o Portal do IBGE.

Ao realizar uma pesquisa com amostras é importante ter muito cuidado com a representatividade delas. É importante assegurar que a parte escolhida da população seja suficientemente adequada para estimar o comportamento do todo.

Podemos minimizar a possibilidade de erro melhorando a qualidade dos instrumentos de pesquisa, determinando criteriosamente o tamanho da amostra e melhorando a habilidade dos pesquisadores.

No planejamento da pesquisa é essencial estudar as técnicas de amostragens. São elas:

1. Amostragens probabilísticas: aleatória simples, sistemática, estratificada, por conglomerado e múltipla.

2. Não probabilísticas: inacessibilidade, a esmo, por material contínuo, intencionais, por voluntários, por quotas, bola de neve.

Ainda é cedo para apresentarmos as técnicas de determinação do tamanho da amostra, mas, ao final desse curso, você já terá condições de compreender as fórmulas utilizadas para o cálculo do tamanho de uma amostra representativa.

O uso inadequado da estatística

O uso inadequado da estatística induz a erro. Isso ocorre quando:

- Utilizamos amostras tendenciosas ou de tamanho pequeno.
- Há uso ambíguo de conceitos.
- Há apresentação equivocada de variáveis.
- Há erros na representação gráfica (intervalos em escalas desiguais).
- Há omissão de dados reveladores.

Veja algumas dessas situações a seguir.

Situação 1

Afirmativa: Cerca de 68% dos entrevistados defendem a pena de morte como solução para a violência no país. Isso equivale a mais da metade dos brasileiros.

Pensamento correto: Se não for informado o número de entrevistados isso pode ser falso. Suponhamos que apenas 3 pessoas deram essa opinião, em que 2 foram favoráveis e 1 contra. O tamanho da amostra não foi suficiente para ser representativa.

Situação 2

Afirmativa: Melhoramos o nosso serviço de delivery em 100%.

Pensamento correto: Se o serviço de entrega era péssimo, melhorar em 100% significa que ele ficou pior.

Situação 3

Afirmativa: O quadro de funcionários de nossa empresa é composto de 100 funcionários com o salário médio de R\$ 5.000,00.

Pensamento correto: Essa informação não é suficiente para saber se nessa empresa todos recebem bons salários. Suponhamos que o presidente da empresa receba R\$ 401.000,00 e 99 funcionários recebam R\$ 1.000,00.

Situação 4

Afirmativa: A pesquisa revelou que 9 em cada 10 pessoas preferem o nosso produto.

Pensamento correto: Essa pesquisa pode ter sido feita baseada em amostras de 10 pessoas que já tinham afinidade com a marca e foram especialmente selecionadas para dar esse resultado.

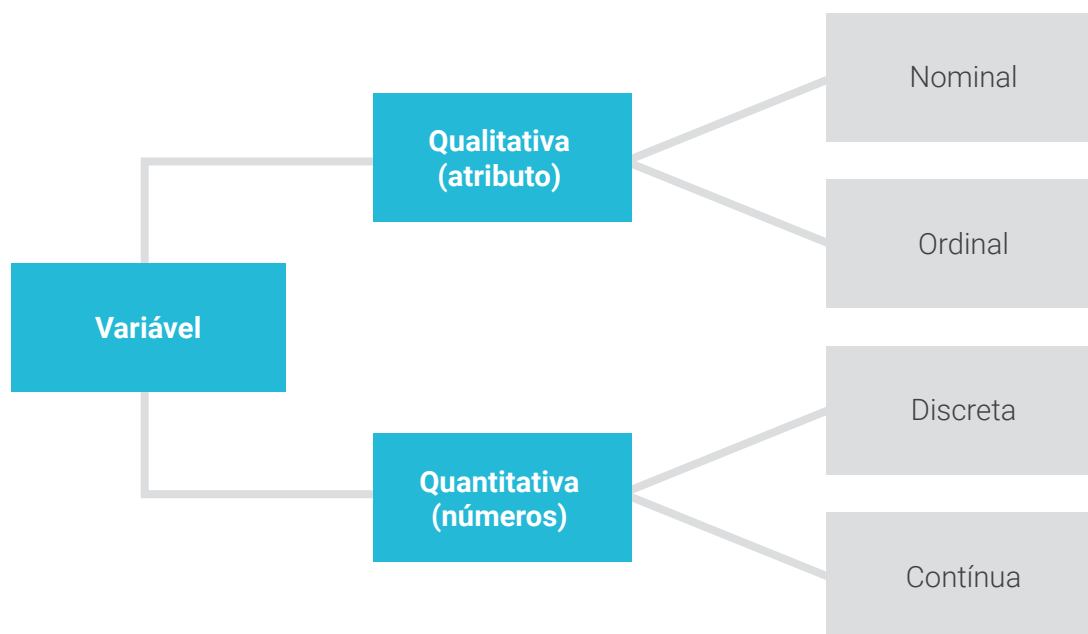
Tipos de erros

1. Erros amostrais: diferença entre o resultado da amostra e o verdadeiro valor da população. Ocorre quando as amostras são aleatórias.

2. Erros não amostrais: ocorrem quando os dados amostrais são coletados incorretamente, devido a uma amostra tendenciosa, instrumento de medida defeituoso, anotações erradas, erro do pesquisador. Não devem existir ou devem ser minimizados com cuidado, atenção e planejamento.

Tipos de variáveis

Observe o esquema a seguir, que ilustra os tipos de variáveis de maneira geral.



Agora, com base na figura anterior, vamos apresentar cada uma delas. Preste atenção!

Qualitativas

A resposta a ser anotada é uma qualidade ou atributo, ou seja, será anotada com uma palavra ou um código numérico.

As variáveis qualitativas podem dividir-se em duas categorias:

- **Qualitativa nominal:** não existe uma hierarquia ou uma ordenação nas possíveis respostas.
Exemplo: profissão, sexo, estado civil, CPF, RG, senha.
- **Qualitativa ordinal:** existe certa ordem nas possíveis respostas.
Exemplo: grau de instrução, classe social.

Quantitativas

A resposta a ser anotada é uma medição ou contagem, ou seja, será anotada com um número.

As variáveis quantitativas podem dividir-se em duas categorias:

- **Quantitativa discreta:** os números a serem anotados são inteiros.
Exemplo: número de filhos, idade.
- **Quantitativa contínua:** os números a serem anotados são reais.
Exemplo: massa corporal, estatura.

Vamos a um exemplo prático.

Exemplo

O trecho a seguir foi extraído da página do IBOPE na internet.

Inquérito domiciliar para monitorar a soroprevalência da infecção pelo vírus SARS-CoV-2 em adultos no município de São Paulo

Estudo transversal com amostragem probabilística realizado no Município de São Paulo entre os dias 15 e 24 de junho de 2020 (16 semanas após o primeiro caso registrado na cidade).

Para medir a soroprevalência no Município de São Paulo foram analisadas 1.183 coletas de sangue dos participantes em 115 setores censitários sendo 12 residências foram sorteadas em cada setor censitário.

Sumário da Metodologia: o Município de São Paulo tem uma população de 8.407.202 habitantes com 18 anos ou mais. Foram criados dois estratos na cidade: distritos com maior renda e distritos com menor renda, sendo que cada um deles corresponde a cerca de metade da população pesquisada. Após responderem um questionário, uma amostra de sangue foi colhida por punção venosa dos participantes. A quantidade de anticorpos anti-SARS-CoV-2 (IgG e IgM) foi medida usando um método de quimiluminescência.

Tabulação dos dados coletados:

Variável	n=1183 %	Prevalência %	IC 95%		valor p
Total	100	11.4	9.2	13.6	
Sexo					
Masculino	46.6	11.6	8.7	14.5	0.8192
Feminino	53.4	11.2	8.5	13.9	
Idade					
18 a 34	34.7	9.2	6.0	12.3	0.1998
35 a 44	19.8	11.0	6.3	15.6	
45 a 59	25.2	15.1	10.7	19.5	
60+	20.2	11.1	6.0	16.2	
Escolaridade					
Menos que Fundamental	19.3	22.9	15.3	30.6	< 0.0001
Fundamental ou Médio	53.2	9.0	4.6	13.4	
Superior	27.4	5.1	1.8	8.4	
Raça/cor*					
Preta	10.5	19.7	10.7	28.7	0.0008
Parda	34.6	14.0	10.6	17.3	
Branca	54.9	7.9	5.6	10.2	
Pessoas no domicílio					
1 a 2	26.2	8.1	3.9	12.3	0.0371
3 a 4	42.6	10.2	7.2	13.3	
5 +	31.2	15.8	11.2	20.4	

Fonte: ibopeinteligencia.com

Observe a metodologia da pesquisa e poderemos verificar que:

População e Amostra

População: 8.407.202 habitantes com 18 anos ou mais do município de São Paulo.

Amostra: 1.183 indivíduos com 18 anos ou mais moradores no município de São Paulo.

Variáveis pesquisadas, classificando-as como qualitativas nominais, qualitativas ordinais, quantitativas discretas, quantitativas contínuas

Sexo: qualitativa nominal.

Idade: quantitativa discreta.

Escolaridade: qualitativa ordinal.

Raça/cor: qualitativa nominal.

Nº de pessoas no domicílio: quantitativa discreta.

Considerações gerais:

- As diferentes prevalências de anticorpos anti-SARS-CoV-2 refletem a desigualdade social do município de São Paulo.
- A soroprevalência diminui com o aumento do nível educacional, sendo 4,5 vezes maior entre os indivíduos que não completaram o ensino fundamental quando comparada com os que terminaram o ensino superior (22,9% versus 5,1%),
- A soroprevalência é 2,5 vezes maior entre os participantes que se identificam como pretos do que nos brancos (19,7% versus 7,9%),
- Participantes que vivem em habitações com 5 ou mais indivíduos apresentam uma soroprevalência quase 2 vezes maior do que aqueles que habitam com 1 ou dois indivíduos (15,8% versus 8,1%),

Exemplo 2

Observe a tabela com as informações e veja o posicionamento das variáveis envolvidas no esquema apresentado, de acordo com a classificação a que ela pertence.

Informações sobre nome, estado civil, grau de instrução, número de filhos, salário, idade, procedência, altura e massa corporal de 10 funcionários da Universidade Veiga de Almeida.

Nome	Estado Civil	Grau de Instrução	Nº de filhos	Salário R\$	Idade (anos)	Procedência	Altura (m)	Massa (kg)
Eduardo	Casado	Superior completo	3	12.534,72	30	capital	1,87	85
Cláudia	Casada	Superior completo	2	18.145,78	56	interior	1,65	62
Carlos	Solteiro	Ensino médio	0	3.456,78	18	capital	1,91	95
Paola	Solteira	Superior Incompleto	0	4.567,89	20	Interior	1,68	55
Maria	Casada	Superior completo	1	6.789,10	48	capital	1,60	65

Fonte: Elaborada pela autora. Dados fictícios.



Exemplo 3

Observe possíveis respostas de uma pesquisa sobre os tipos de amostragens, tanto probabilísticas quanto não probabilísticas, com um exemplo para cada tipo pesquisado.

Tipo	Amostragem	Descrição	Exemplos
Amostragens Probabilísticas	Aleatória Simples (ou casual)	Processo mais elementar e muito utilizado. Equivale a um sorteio. Numera-se a população de 1 a n e, em seguida, realiza-se um sorteio aleatório de elementos dessa numeração que serão os elementos pertencentes à amostra.	Consideremos uma população de 1.000 elementos. Assim, podemos numerá-la de 000 a 999. Em seguida, escolhemos um ponto de uma tabela de números aleatórios e agrupamos de 3 em 3; assim, vamos selecionando os elementos da amostra. Considerando a hipotética linha 123456789012345678901, selecionaríamos os elementos: 123 – 456 – 789 – 012 – 345 – 678 – 901.
	Estratificada (ou proporcional)	Classifica a população em pelo menos duas subpopulações (ou estratos) que possuem as mesmas características, para, assim, extrair proporcionalmente amostras de cada estrato.	Seja uma população formada por 100 estudantes, sendo 40% do sexo feminino e 60% do sexo masculino. Resolve-se selecionar uma amostra de 10 elementos. De acordo com essa técnica devemos manter a mesma proporção de representatividade de cada grupo, portanto a amostra deve conter 4 mulheres e 6 homens.

Amostragens Probabilísticas	Sistemática	Quando a população em estudo já encontra-se ordenada de acordo com algum critério específico, como fichários, prontuários, listas, calculamos um intervalo de amostragem, também conhecido com intervalo de seleção, que definirá quais elementos participarão da pesquisa. O cálculo dele é dado pelo tamanho da população dividido pelo tamanho da amostra.	Deseja-se obter uma amostra composta de 80 famílias residentes em determinada rua, que conta com 2.000 casas. Assim, nosso intervalo seria dado por $2.000/80$, que resulta em 25. Considerando os critérios aleatórios, devemos escolher um número entre as casas 0 e 25 e, a partir daí, contar de 25 em 25 casas para inclusão na amostra. Considerando que o número sorteado foi 5, selecionaremos a família da referida casa para participar da amostra, e depois aqueles residentes nas casas 5, 30, 55, 80, 105, até atingirmos 80 casas.
	Por conglomerado	Divide-se a área da população em seções (ou conglomerados), depois selecionam-se aleatoriamente algumas dessas seções e, finalmente, consideram-se todos os elementos dos conglomerados selecionados.	Ex.: pesquisa pré-eleitoral. Selecionam-se aleatoriamente 30 zonas eleitorais e realiza-se a pesquisa com todos os elementos das zonas selecionadas.
Amostragens não probabilísticas	Acidental ou casual	Os indivíduos escolhidos nessa pesquisa estão prontamente disponíveis e não porque foram selecionados por meio de um critério estatístico.	Pesquisas de opinião em supermercados.
	Intencional	Intencionalmente é escolhido o grupo de elementos que irá compor a amostra.	Para um estudo sobre automóveis, o pesquisador procura apenas oficinas.
	Por quotas	Assemelha-se ao método de amostragem estratificada, diferindo apenas pelo fato de a escolha não seguir critérios de aleatoriedade.	Deseja-se entrevistar apenas indivíduos da classe A, que representam 12% da população. Esta será a quota para o trabalho.

Amostragens não probabilísticas	Bola de neve	Cada indivíduo entrevistado indica uma ou mais pessoas para a próxima entrevista.	Ao entrevistar imigrantes de um país específico, pode-se iniciar buscando pessoas conhecidas e pedir novos contatos de contêrâneos conhecidos delas. Esse processo continua até que o pesquisador tenha todas as entrevistas de que necessita ou até que todos os contatos tenham sido atingidos.
--	--------------	---	---

Indicadores e Índices

O **indicador** é uma medida quantitativa que auxilia na tomada de decisões. É obtido relacionando-se as variáveis. Muito usado para avaliar tendências e performances. É um excelente feedback, já que permite demonstrar, de forma clara e objetiva, o desempenho de um processo em um determinado intervalo de tempo.

Índice é o valor numérico assumido por um indicador em determinado momento.



Exemplo

Para entender melhor esses conceitos vamos supor que uma organização utilize como indicador o número de colaboradores com nível superior completo dividido pelo número total de colaboradores. Se a empresa apurar esse indicador em um determinado período e alcançar 0,5 ou 50% esse será o seu índice de escolaridade de nível superior dos funcionários.

Os indicadores e índices são muito usados nas áreas de sustentabilidade, meio ambiente, economia, entre outras.

Observe a tabela a seguir com os indicadores e índices mais usados.

Indicadores mais utilizados	Índices mais utilizados
PIB – Produto Interno Bruto - É um dos indicadores mais utilizados em macroeconomia. Seu objetivo é quantificar a atividade econômica de uma região.	IDH – Índice de desenvolvimento Humano - Proposto pela ONU, calculado no nosso país a partir dos dados censitários do IBGE. Calculado como sendo a média geométrica de três índices, educação, renda e saúde. É mais representativo do que o PIB, já que este só se importa com a economia, sem preocupação com as condições da população; por isso, pode mascarar a desigualdade existente.
Tabela Fipe - Expressa preços médios de veículos anunciados por vendedores, no mercado nacional, servindo como parâmetro para negociações.	IBOVESPA – Índice da Bolsa de Valores de São Paulo – É um índice de desempenho das ações negociadas na Bovespa.
IQA – Índice de qualidade do ar - É um indicador padronizado do nível de poluição do ar em uma determinada região.	IPC – Índice de Preços ao Consumidor , apurado pela FGV (Fundação Getúlio Vargas). É um índice para avaliação do poder de compra do consumidor.
Indicador social - Usado para informar algo sobre determinado aspecto da realidade social, para fins de pesquisa ou avaliação de políticas públicas.	IPCA – Índice de Preços ao Consumidor Amplo - Medido mensalmente pelo IBGE (Instituto Brasileiro de Geografia e Estatística), foi criado com o objetivo de oferecer a variação dos preços no comércio para o público final. O IPCA é considerado o índice oficial de inflação do país.
	IGP-M – Índice Geral de Preços do Mercado - Medido pela FGV (Fundação Getúlio Vargas), é um indicador de variação dos preços na economia brasileira. Como o Brasil é um país fortemente impactado pela inflação, ele serve para verificar as mudanças quanto ao valor da moeda e nas alterações de preços.

INCC - Índice Nacional de Custo da Construção - Mede a variação de preços que faz parte do setor da construção civil na economia brasileira.



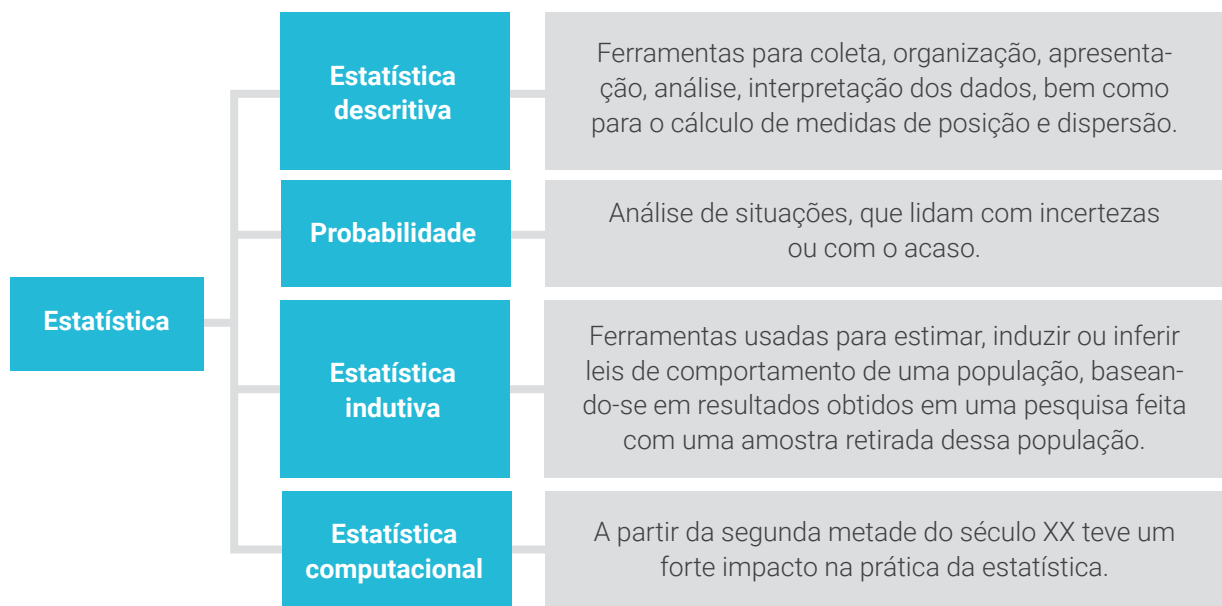
Ampliando o foco

Tenha em mente que tanto os indicadores quanto os índices são de grande utilidade para o planejamento estratégico de uma empresa e também para o nosso dia a dia. Neste material apresentaremos apenas alguns desses indicadores e índices, contudo existe uma variedade deles.

Para saber mais é importante que você pesquise sobre indicadores e índices ligados à **sua área de interesse**.

Divisões da Estatística

Observe o esquema a seguir para conhecer as divisões da Estatística.



Softwares estatísticos

Muitos softwares fazem integração entre informática e estatística. Contudo, neste material listaremos os softwares mais usados. Veja a seguir uma breve descrição sobre cada um deles.

1. SPSS – Software comercial de análises estatísticas. Fornece um conjunto robusto de recursos que permitem extrair informações dos dados. Possui versão demo.

2. MINITAB – Software comercial de análises estatísticas. Possui um ambiente completo para a análise de dados. Sua interface intuitiva permite que o usuário trabalhe, simultaneamente, com planilhas eletrônicas de dados, tabelas estatísticas, gráficos e textos em forma de “janelas”. Possui versão demo.

3. R for Windows – É um dos softwares estatísticos mais usados no mundo. Gratuito. Disponível para Windows e Unix.

4. STATISTICA – Software comercial de análises estatísticas. Possui versão demo.

5. Excel – Na planilha eletrônica da Microsoft encontramos várias funções estatísticas, de fácil utilização e amplamente comercializadas.

6. Bioestat – Software para estudantes de fácil uso para iniciantes, voltado sobretudo para a área de Ciências Biológicas, mas pode ser usado em qualquer área. É gratuito e tem versão em português.



MIDIAATECA

Na midiateca você encontra os links com os endereços para fazer o download da versão “demo” dos softwares comerciais.

Elaboração e análise de gráficos e tabelas

Um dos objetivos da Estatística é fazer a síntese e a organização dos valores que a variável de estudo pode assumir. Para alcançar esse propósito lançamos mão das tabelas e dos gráficos que viabilizarão informações rápidas e seguras a respeito da variável em estudo.

Neste tópico abordaremos como fazer a coleta dos dados, a organização em tabelas de distribuição de frequências e algumas representações gráficas.

Coleta dos dados

Dentre diversas maneiras para se coletarem dados, a amostragem é a mais frequente. Pode ser feita a partir da aplicação de um questionário, que deverá ser respondido pelo entrevistado fisicamente, por meio do telefone ou, ainda, pela internet. Podem também ser obtidos por meio de gravação de entrevistas ou por um levantamento dos dados que já tenham sido disponibilizados por outro pesquisador.

São importantes alguns cuidados para coletar os dados. São eles:

- **Identificação:** o entrevistado só deve ser identificado nos casos em que realmente seja necessário, pois sem identificação ele ficará mais à vontade para responder às questões propostas.
- **Palavras adequadas:** a comunicação deve ser feita de acordo com o público-alvo, tomando-se o cuidado com palavras muito rebuscadas ou termos vulgares.
- **Perguntas impróprias:** somente perguntas relevantes devem ser feitas ao entrevistado.
- **Perguntas com duplo sentido:** averiguar se o que foi perguntado está sendo compreendido.
- **Perguntas sem sentido estatístico:** evitar perguntas que não serão necessárias ao estudo. Questionários muito longos cansam o entrevistado e ele tende a não responder ou a responder de qualquer maneira.
- **Testar antes de aplicar:** faça um projeto piloto antes de sair para sua entrevista. Aplique o questionário a uma pequena quantidade de pessoas e certifique-se de que realmente o que você está perguntando está sendo bem compreendido. Só depois desse teste é que o pesquisador deve ir a campo.

Organização dos dados coletados em tabelas

A elaboração de tabelas obedece à Resolução nº 886, de 26 de outubro de 1966, do Conselho Nacional de Estatística, sendo que, a partir de 1993, o IBGE também passou a normatizar a apresentação de tabelas.

E quais elementos uma tabela deve conter?

Observe os elementos a seguir. São eles:

1. Título: conjunto de informações, as mais completas possíveis, respondendo às seguintes perguntas:

- O quê (referente ao fato).
- Onde (relativo ao lugar).
- Quando (corresponde à época).

2. Cabeçalho: parte superior da tabela que especifica o conteúdo das colunas.

3. Coluna indicadora: parte da tabela que especifica o conteúdo das linhas.

4. Classes: são as linhas do corpo da tabela.

5. Corpo: conjunto de linhas e colunas que contêm informações sobre a variável em estudo.

6. Casa ou célula: espaço destinado a um só número.

7. Fonte e notas: colocadas, preferencialmente, no rodapé.

Exemplo de uma tabela com os elementos.

Os 10 números mais sorteados na Mega-sena desde o concurso nº 1 (de 11/03/1996) até o concurso nº 2281 (de 18/07/2020) ← Título

← Cabeçalho	
Número	Quantidade de vezes que foi sorteado
1ª Classe → 53	261
10	260
23	254
5	254
33	248
4	248
27	245
54	244
9ª Classe → 24	244
42	244
Total	2.502

← Rodapé

Fonte: <https://www.somatematica.com.br/megasenaFrequentes.php>

← Corpo

← Casa ou célula

Considerações:

- A tabela não é delimitada à esquerda e nem à direita (deve ficar aberta lateralmente).
- A tabela é limitada superiormente e inferiormente por uma linha grossa ou dupla.
- Os elementos "fonte" e "título" são obrigatórios.
- Devemos colocar um traço horizontal (–) quando o valor é zero.
- Quando não temos os dados sinalizamos com três pontos (...).
- Um ponto de interrogação (?) deve ser usado quando temos dúvida quanto à exatidão de determinado valor.
- O zero (0) deve ser usado quando o valor é muito pequeno para ser expresso pela unidade utilizada.

Elaboração de uma tabela de distribuição de frequências

Uma tabela de distribuição de frequências é uma tabela-resumo que mostra classes, frequências absolutas, frequências relativas e frequências acumuladas.

- **Dados brutos:** são os dados originais na forma como foram coletados, sem nenhuma organização.
- **Rol estatístico:** é a organização dos dados brutos em ordem crescente, decrescente ou alfabética.

- **Frequência absoluta ou frequência simples ou simplesmente frequência:** é a quantidade de vezes que cada valor é observado.
- **Frequências relativas:** obtidas pela divisão entre a frequência e o total ou pela frequência acumulada e o total. Geralmente são expressas em %.
- **Frequências acumulada:** é a soma de cada frequência com as anteriores ou a soma de cada frequência relativa com as anteriores.

Vamos a um exemplo prático.

Exemplo

O SINCOM realizou uma pesquisa sobre os preços de um modelo de computador em 20 lojas de informática, na cidade de Juiz de Fora, em julho de 2020. Foram coletados os valores a seguir em reais:

2000	2500	2000	2600	2000	2600	2600	2500	2500	2000
2000	2000	2500	2600	2600	2600	2600	2600	2600	2600

Observe como organizam-se os dados brutos em uma tabela de distribuição de frequências.

1º passo: confeccionar o rol, ou seja, precisamos organizar os dados brutos em ordem crescente ou decrescente.

2000	2000	2000	2000	2000	2000	2500	2500	2500	2500
2600	2600	2600	2600	2600	2600	2600	2600	2600	2600

2º Passo: com o rol confeccionado iniciamos a contagem para preencher a coluna da **frequência absoluta**:

Rol

2000	2000	2000	2000	2000	2000	2500	2500	2500	2500
2600	2600	2600	2600	2600	2600	2600	2600	2600	2600

Tabela: Preço do microcomputador modelo X, em 20 lojas de JF – Julho 2020.

Preço	Frequência absoluta (f_i)	Frequência relativa ($fr_i\%$)	Frequência acumulada (F_i)	Frequência relativa acumulada ($Fr_i\%$)
2000	6			
2500	4			
2600	10			
Total	20			

Fonte: SINCOM (2020).

A **frequência relativa** será obtida dividindo-se a frequência absoluta pelo total. Para expressarmos o resultado em porcentagem temos que multiplicar o resultado por 100.

Tabela: Preço do microcomputador modelo X, em 20 lojas de JF – Julho 2020.

Preço	Frequência absoluta (f_i)	Frequência relativa ($fr_i\%$)	Frequência acumulada (F_i)	Frequência relativa acumulada ($Fr_i\%$)
2000	6	$6/20 = 0,30$ ou 30%		
2500	4	$4/20 = 0,2$ ou 20%		
2600	10	$10/20 = 0,5 = 50\%$		
Total	20	100%		

Fonte: SINCOM (2020).

A frequência acumulada será a soma de todas as frequências absolutas que estão na classe (linha) ou acima dela. Veja a tabela a seguir:

Tabela: Preço do microcomputador modelo X, em 20 lojas de JF – Julho 2020.

Preço	Frequência absoluta (f_i)	Frequência relativa ($fr_i\%$)	Frequência acumulada (F_i)	Frequência relativa acumulada ($Fr_i\%$)
2000	6	$6/20 = 0,30$ ou 30%	6	
2500	4	$4/20 = 0,2$ ou 20%	$6 + 4 = 10$	
2600	10	$10/20 = 0,5 = 50\%$	$6 + 4 + 10 = 20$	
Total	20	100%		

Fonte: SINCOM (2020).

A frequência acumulada relativa será a soma de todas as frequências relativas que estão na classe(linha) ou acima dela:

Tabela: Preço do microcomputador modelo X, em 20 lojas de JF – Julho 2020.

Preço	Frequência absoluta (f_i)	Frequência relativa ($fr_i\%$)	Frequência acumulada (F_i)	Frequência relativa acumulada ($Fr_i\%$)
2000	6	$6/20 = 0,30$ ou 30%	6	30%
2500	4	$4/20 = 0,2$ ou 20%	$6 + 4 = 10$	$30\% + 20\% = 50\%$
2600	10	$10/20 = 0,5 = 50\%$	$6 + 4 + 10 = 20$	$50\% + 50\% = 100\%$
Total	20	100%		

Fonte: SINCOM (2020).

Tabela de distribuição de frequência COM intervalos de classe

Quando os dados obtidos assumem muitos valores diferentes, é costume, em vez de listar as respostas uma a uma, apresentar as tabelas em intervalos, chamados de classes.

Veja um exemplo de uma tabela de distribuição de frequências com intervalo de classe:

Tabela: População feminina residente no Brasil por idade, em 2019.

Idade (anos)	Frequência Absoluta	Frequência Relativa	Frequência Acumulada	Frequência Acumulada Relativa
0 I--- 4	4991	4,6%	4991	4,6%
4 I--- 6	2638	2,4%	7629	7,0%
6 I--- 10	5516	5,1%	13145	12,1%
10 I--- 15	7167	6,6%	20312	18,7%
15 I--- 18	4624	4,3%	24936	23,0%
18 I--- 25	11063	10,2%	35999	33,2%
25 I--- 30	7674	7,0%	43673	40,2%
30 I--- 40	17011	15,7%	60684	55,9%
40 I--- 60	29168	26,9%	89852	82,8%
60 anos ou mais	18658	17,2%	108510	100,0%
Total	108510	100,0%		

Fonte: IBGE - Pesquisa Nacional por Amostra de Domicílios Contínua anual (2020).

Elementos de uma distribuição de frequências com intervalos de classe:

Uma distribuição de frequência com intervalos de classe apresenta alguns elementos. São eles:

Classe	Limites das classes	Intervalo das classes	Amplitude amostral
Amplitude de um intervalo de classe	Ponto médio de uma classe	Frequências de cada classe	

Agora vamos conhecer mais um pouco sobre eles.

1. Classes

São os intervalos da variável quantitativa.

Exemplo:

Na tabela 2, temos:

1ª classe: 0 |---- 4

2ª classe: 4 |---- 6

3ª classe: 6 |---- 10

4ª classe: 10 |---- 15

5ª classe: 15 |---- 18

6ª classe: 18 |---- 25

7ª classe: 25 |---- 30

8ª classe: 30 |---- 40

9ª classe: 40 |---- 60

10ª classe: 60 anos ou mais

2. Como calcular o número de classes: (k)

Temos três formas que podem ser usadas para calcular o número de classes de uma tabela de distribuição de frequências em função da quantidade de dados (n):

Sejam:

k = número de classes

n = quantidade de dados

• **Primeira forma:** Fórmula de Sturges: $k \cong 1 + 3,3 \cdot \log n$

• **Segunda forma:** $k = \sqrt{n}$

• **Terceira forma:** o bom senso do pesquisador. A familiaridade do pesquisador com os dados é que deve indicar quantas classes devem ser construídas. As fórmulas são apenas sugestões. Cabe ao pesquisador analisar o resultado e tomar a decisão de quantas classes deverão ser adotadas. A orientação é que o número de classes deve estar entre 5 e 20. Tabelas com poucas classes perdem a precisão dos dados e com muitas classes ficam cansativas.

3. Limites da classe

São os valores extremos de cada classe. O menor número é o limite inferior da classe (Li) e o maior número, o limite superior (LS) da classe.

Exemplo:

Na tabela 2 temos:

Na primeira classe: $L_i = 0$ $LS = 4$

Na quinta classe: $L_i = 15$ $LS = 18$

Na décima classe: $L_i = 60$ $LS = -$

4. Tipos de intervalos e representações mais usadas

a) Fechado à esquerda e fechado à direita: $l-----l$ ou $\bullet-----\bullet$ ou $[,]$

b) Fechado à esquerda e aberto à direita: $l-----$ ou $\bullet-----\circ$ ou $[, [$

c) Aberto à esquerda e fechado à direita: $-----l$ ou $\circ-----\bullet$ ou $] ,]$

d) Aberto à esquerda e aberto à direita: $-----$ ou $\circ-----\circ$ ou $] , [$

É muito importante entender essa simbologia para determinar corretamente as frequências absolutas.

Veja o exemplo a seguir.

Exemplo

1º) Considere o seguinte levantamento sobre o número de salários mínimos recebidos pelos funcionários de uma empresa:

Funcionário	Nº de S.M
Funcionário 1	1
Funcionário 2	4
Funcionário 3	4,2
Funcionário 4	3,9
Funcionário 5	5
Funcionário 6	2,5
Funcionário 7 (voluntário)	0
Funcionário 8	3
Funcionário 9	4,1

2º) Organize os dados em ordem crescente, ou seja, confeccione o ROL:

0; 1; 2,5; 3; 3,9; 4; 4,1; 4,2; 5

3º) Suponha que vamos preencher a coluna da frequência absoluta da seguinte tabela de distribuição:

Nº de salários	Frequência Absoluta	Explicação
$0 \text{ --- } 5$ ou $0 \bullet \text{ --- } \bullet 5$ ou $[0, 5]$	9	<p>Como o intervalo é fechado à esquerda, o limite inferior será considerado e, como o intervalo é fechado à direita, o limite superior também deve ser considerado.</p> <p>Então, fazendo a contagem no rol temos:</p> <p>0; 1; 2,5; 3; 3,9; 4; 4,1; 4,2; 5</p> <p>Portanto, 9 funcionários recebem de zero até, inclusive, cinco salários mínimos.</p>

4º) Modifique o intervalo e faça a contagem da frequência absoluta:

Nº de salários	Frequência Absoluta	Explicação
<p>0 ----5 ou 0 ●----○ 5 ou [0,5[</p>	8	<p>Como o intervalo é fechado à esquerda, o limite inferior será considerado e, como o intervalo é aberto à direita, o limite superior NÃO deve ser considerado.</p> <p>Então, fazendo a contagem no rol temos: 0; 1; 2,5; 3; 3,9; 4; 4,1; 4,2</p> <p>Portanto, 8 funcionários recebem de zero até antes de 5 salários mínimos.</p>

5º) Modifique, novamente o intervalo e faça a contagem:

Nº de salários	Frequência Absoluta	Explicação
<p>0 ---- 5 ou 0 ○----● 5 ou]0, 5]</p>	8	<p>Como o intervalo é aberto à esquerda, o limite inferior NÃO será considerado e, como o intervalo é fechado à direita, o limite superior deve ser considerado.</p> <p>Então, fazendo a contagem no rol temos: 1; 2,5; 3; 3,9; 4; 4,1; 4,2; 5</p> <p>Portanto, 8 valores satisfazem essa condição, ou seja, ganham mais de 0 salários até inclusive 5 salários mínimos.</p>

Por fim, modificando novamente o intervalo e fazendo a contagem teremos:

Nº de salários	Frequência Absoluta	Explicação
<p>0 ---- 5 ou 0 ○ ---- ○ 5 ou]0, 5[</p>	7	<p>Como o intervalo é aberto à esquerda, o limite inferior NÃO será considerado e, como o intervalo é aberto à direita, o limite superior também NÃO deve ser considerado.</p> <p>Então, fazendo a contagem no rol temos: 1; 2,5; 3; 3,9; 4; 4,1; 4,2</p> <p>Portanto, 7 funcionários satisfazem essa condição, ou seja, recebem entre zero e cinco salários mínimos.</p>

5. Amplitude amostral: (AA)

É a diferença entre o maior valor e o menor valor do rol.

$$AA = x_{\text{máximo}} - x_{\text{mínimo}}$$

6. Amplitude de um intervalo de classe: (h_i)

É a diferença entre o limite superior e o limite inferior da classe.

$$h_i = Ls_i - Li_i$$

Como vimos na tabela *População feminina residente no Brasil por idade, em 2019*, que apresenta a população feminina residente no Brasil por idade em 2019, temos:

Amplitude da primeira classe (0 I---- 4) $\rightarrow h_1 = Ls_1 - Li_1 = 4 - 0 = 4$

Amplitude da quarta classe (10 I---- 15) $\rightarrow h_4 = Ls_4 - Li_4 = 15 - 10 = 5$

Amplitude da décima classe (60 anos ou mais) $\rightarrow h_{10} = Ls_{10} - Li_{10} = ???$
(não pode ser determinada, já que o limite superior não foi especificado)

Quando a tabela ainda não foi criada, para calcular o intervalo de cada classe devemos dividir a amplitude total (AT) pelo número de classes(k). Então:

$$h = \frac{AT}{k}$$

A amplitude de um intervalo de classe pode ser igual em toda tabela (que é preferível, sempre que possível) ou diferente em algumas classes.

7. Ponto Médio de uma classe: (X_i)

É o valor que representa a classe, calculado como sendo a média aritmética entre os limites da classe.

$$X_i = \frac{Li_i + Ls_i}{2}$$

Na tabela que apresenta a população feminina residente no Brasil por idade em 2019, temos:

Ponto médio da primeira classe (0 |---- 4) $\rightarrow X_1 = \frac{Li_1 + Ls_1}{2} = \frac{0 + 4}{2} = 2$

Ponto médio da segunda classe (4 |---- 6) $\rightarrow X_2 = \frac{Li_2 + Ls_2}{2} = \frac{4 + 6}{2} = 5$

Calculamos da mesma maneira para todas as classes que possuem limite inferior e limite superior.

Ponto médio da décima classe (60 anos ou mais) \rightarrow (não pode ser determinado, já que o limite superior não foi especificado)

8. Frequências de cada classe

As frequências absolutas, relativas e acumuladas são calculadas exatamente iguais ao cálculo das frequências para tabelas de distribuição de frequência sem intervalos de classe. São elas:

- **Frequência absoluta ou frequência simples ou simplesmente frequência:** é a quantidade de vezes que cada valor é observado.
- **Frequências relativas:** obtida pela divisão entre a frequência e o total ou pela frequência acumulada e o total. Geralmente são expressas em %.
- **Frequências acumulada:** é a soma de cada frequência com as anteriores ou a soma de cada frequência relativa com as anteriores.

Vamos lembrar o cálculo das frequências relativas?

Para isso, é preciso efetuar uma divisão e o resultado deve ser arredondado.

Critério de arredondamento de dados

De acordo com a Resolução nº 886/1966 do IBGE:

I) Se o primeiro algarismo a ser abandonado for menor do que 5, ou seja quando o primeiro algarismo a ser abandonado for 0,1,2,3 ou 4, ficará inalterado o último algarismo que permanece.

Exemplo: Arredondar com 1 casa decimal os números seguintes:

53,24 passa para 53,2.

24,13 passa para 24,1.

II) Se o primeiro algarismo a ser abandonado for maior do que 5, ou seja, quando o primeiro algarismo a ser abandonado é o 6,7,8, ou 9, aumenta-se em uma unidade o algarismo que permanece.

Exemplos:

13,87 passa para 13,9.

24,08 passa para 24,1.

34,99 passa para 35,0.

III) Quando o algarismo a ser abandonado for igual a 5 existem dois critérios e você deve adotar o que achar mais conveniente, sem nenhum problema.

Se, após o 5 seguir em qualquer casa um algarismo diferente de zero, aumenta-se uma unidade ao algarismo que permanece. (Esse critério é adotado por calculadoras e computadores.)

Exemplos:

7,3⁵2 passa para 7,4.

45,6⁵01 passa para 45,7.

86,2⁵0002 passa para 86,3.

Outro critério quando o dígito a ser abandonado for o 5:

Se o 5 for o último algarismo ou após o 5 só se seguirem zeros, o último algarismo a ser conservado só será aumentando de uma unidade se for ímpar.

Exemplos:

18,⁷5 passa para 18,8

29,⁶5 passa para 29,6

37,⁷5000 passa para 37,8

49,⁸500 passa para 49,8

Observação: Nunca devemos fazer arredondamentos de sucessivos. O arredondamento deve ser feito apenas na resposta final.

Veja alguns exemplos práticos para entender melhor. Vamos lá!

Exemplo 1

Em julho de 2020 foi feita uma pesquisa sobre a idade de 50 funcionários da UVA. Os dados estão apresentados a seguir:

31	43	44	44	45	48	45	46	47	45
49	50	51	53	54	54	56	58	26	65
18	20	20	21	22	24	25	25	62	27
29	34	30	30	41	31	32	33	29	35
36	36	37	37	40	37	38	38	38	37

Como devemos construir a tabela completa de distribuição de frequências, com o intervalo fechado à esquerda?

Resolução:

Sete passos devem ser seguidos para confeccionarmos a tabela de distribuição de frequências com intervalos de classe:

- 1- Confeccionar o Rol Estatístico.
- 2- Determinar o número de classes (k).
- 3- Calcular a amplitude dos intervalos de classe (h).
- 4- Determinar o limite inferior da primeira classe (Li).
- 5- Determinar o limite superior da primeira classe (Ls).
- 6- Determinar os dos limites inferiores e superiores das demais classes.
- 7- Determinar a frequência absoluta de cada classe e, a partir daí, calcular as frequências relativas e acumuladas.

1º passo: confecção do Rol.

18	20	20	21	22	24	25	25	26	27
29	29	30	30	31	31	32	33	34	35
36	36	37	37	37	37	38	38	38	40
41	43	44	44	45	45	45	46	47	48
49	50	51	53	54	54	56	58	62	65

2º passo: determinação do número de classes (k).

$$k \cong 1 + 3,3 \cdot \log n$$

$$k \cong 1 + 3,3 \cdot \log 50$$

$$k \cong 1 + 3,3 \cdot 1,6989970004$$

$$k \cong 1 + 5,606601014$$

$$k \cong 6,606601014$$

$$k \cong 7 \text{ classes}$$

Se tivéssemos optado pelo método da raiz quadrada de determinação das classes teríamos:

$$k = \sqrt{n}$$

$$k = \sqrt{50}$$

$$k = 7,071067812 \cong 7 \text{ classes}$$

3º passo: cálculo da amplitude dos intervalos de classe (h).

$$h = \frac{AA}{k}$$

$$h = \frac{65 - 18}{7} = 6,714285714 \cong 7$$

4º passo: determinação do limite inferior da primeira classe (Li).

Normalmente, o limite inferior da primeira classe será o menor valor da amostra, ou um número menor do que ele.

No nosso caso, adotaremos $L_i = 18$

5º passo: determinar o limite superior da primeira classe (Ls).

Para determinar o limite superior da primeira classe basta adicionar a amplitude da classe ao limite inferior, ou seja:

$$1^{\text{a}} \text{ classe: } Li = 18 \quad Ls = 18 + 7 = 25$$

6º passo: determinar os dos limites inferiores e superiores das demais classes.

O limite superior da primeira classe será o limite inferior da segunda classe e assim sucessivamente, até que o maior valor pesquisado esteja contido na última classe.

$$2^{\text{a}} \text{ classe: } Li = 25 \quad Ls = 25 + 7 = 32$$

$$3^{\text{a}} \text{ classe: } Li = 32 \quad Ls = 32 + 7 = 39$$

$$4^{\text{a}} \text{ classe: } Li = 39 \quad Ls = 39 + 7 = 46$$

$$5^{\text{a}} \text{ classe: } Li = 46 \quad Ls = 46 + 7 = 53$$

$$6^{\text{a}} \text{ classe: } Li = 53 \quad Ls = 53 + 7 = 60$$

$$7^{\text{a}} \text{ classe: } Li = 60 \quad Ls = 60 + 7 = 67 \quad (67 > 65 \rightarrow \text{ok!})$$

7º passo: determinar a frequência absoluta de cada classe e, a partir daí, calcular as frequências relativas e acumuladas da mesma maneira que foram calculadas para as distribuições de frequências sem intervalos de classe.

Tabela: Idade de 50 funcionários da Universidade Veiga de Almeida em Julho de 2020.

Idade (anos)			fi	fri %	Fi	Fri%
18	----	25	6	12%	6	12%
25	----	32	10	20%	16	32%
32	----	39	13	26%	29	58%
39	----	46	8	16%	37	74%
46	----	53	6	12%	43	86%
53	----	60	5	10%	48	96%
60	----	67	2	4%	50	100%
Total			50	1		

Fonte: Elaborada pela autora. Dados fictícios.

Interpretação da tabela

Uma amostra com 50 dados, coletados em julho de 2020, informam as idades de colaboradores da Universidade Veiga de Almeida – UVA.

Após a organização desses dados pode-se observar que aproximadamente 26% dos funcionários apresentam idade igual ou superior a 32 anos e inferior a 39 anos. Em seguida, com um percentual de 20%, estão os colaboradores que apresentam idade de 25 até antes de 32 anos. Percebe-se que 4% dos empregados estão acima de 60 anos e 12% estão abaixo de 25 anos. Mais da metade dos colaboradores, aproximadamente 58%, têm idade abaixo de 39 anos.

Você pode perceber que outras informações ainda podem ser retiradas da tabela?

Exemplo 2

Os dados coletados pela Secretaria do Meio Ambiente e já ordenados referem-se aos níveis de ruído em decibéis de algumas áreas residenciais da cidade do Rio de Janeiro.

55,00	55,89	56,03	56,67	59,36	60,14	60,22	60,32
60,74	60,96	61,49	61,89	61,92	62,57	62,69	63,14
63,29	64,00	64,17	64,43	64,71	64,78	65,08	65,70
65,81	66,01	66,16	66,84	70,00	71,46	71,52	72,99

Como poderíamos organizá-los em uma distribuição de frequências com intervalos de classes?

Resposta:

Tabela: Níveis de ruído de algumas áreas residenciais da cidade do Rio de Janeiro, em julho de 2020.

Nível de ruído (dB)	Frequência absoluta	Frequência Relativa	Frequência Acumulada	Frequência Acumulada Relativa
55 ---- 58	4	12,5%	4	12,5%
58 ---- 61	6	18,8%	10	31,3%
61 ---- 64	7	21,9%	17	53,2%
64 ---- 67	11	34,3%	28	87,5%
67 ---- 70	0	0,0%	28	87,5%
70 ---- 73	4	12,5%	32	100,0%
Total	32	100,0%		

Fonte: Secretaria do Verde e do Meio Ambiente do Rio de Janeiro (2020).

Representação gráfica de uma distribuição de frequências

As tabelas de distribuição de frequências podem ser representadas graficamente. Sua finalidade principal é fornecer informações analíticas de maneira mais rápida e verificar o comportamento da distribuição. Apresentaremos três tipos de representações gráficas das distribuições: **histogramas, polígonos de frequências e polígonos de frequências acumuladas** (Ogiva de Galton).

Histograma

Antes de começarmos observe a tabela a seguir.

Tabela: Idade de 50 funcionários da Universidade Veiga de Almeida em Julho de 2020.

Idade (anos)	fi	fri %	Fi	Fri%
18 ---- 25	6	12%	6	12%
25 ---- 32	10	20%	16	32%
32 ---- 39	13	26%	29	58%
39 ---- 46	8	16%	37	74%
46 ---- 53	6	12%	43	86%
53 ---- 60	5	10%	48	96%
60 ---- 67	2	4%	50	100%
Total	50	1		

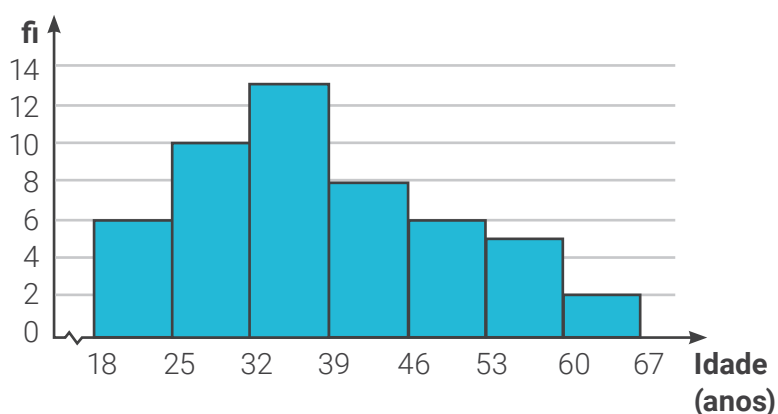
Fonte: Elaborada pela autora. Dados fictícios.

Para representá-la graficamente vamos utilizar um tipo de gráfico que se chama **Histograma**.

Para construir um polígono de frequência acumulada é necessário:

- 1- Traçar um eixo horizontal e marcar em escala os limites inferiores e superiores das classes.
- 2- Traçar um eixo horizontal e marcar as frequências (podem ser frequência absoluta, acumulada ou relativa).
- 3- Traçar um retângulo em que a base corresponde à classe e a altura seja igual à frequência.
- 4- Colocar o título no gráfico e identificar as variáveis nos eixos.

Gráfico: Idade de 50 funcionários da UVA, em Jul/2020



Fonte: Elaborado pela autora. Dados fictícios.



Ampliando o foco

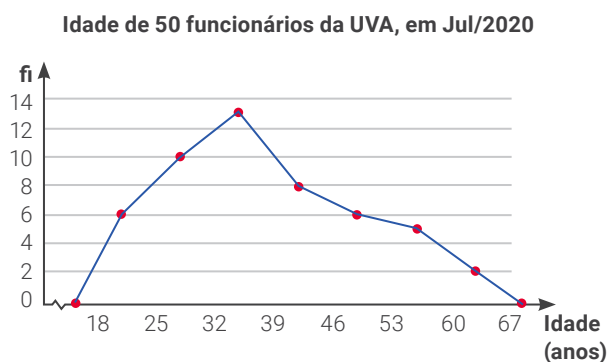
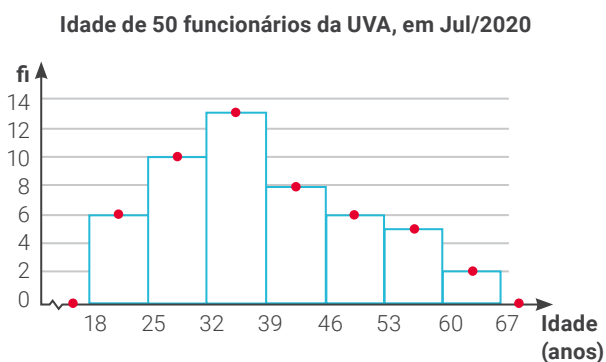
Na planilha eletrônica Excel, versão 2016, existe o comando que gera automaticamente o histograma. Basta selecionar a base de dados e clicar em **Inserir – Gráficos -Histograma**.

Polígono de frequência

Para construir um polígono de frequência é necessário:

- a. Marcar os pontos no gráfico, sendo que no eixo horizontal você marca os pontos médios das classes e no eixo vertical as frequências.

- b. Relacionar cada ponto médio com sua respectiva frequência.
- c. Marcar metade das distâncias dos pontos médios no eixo horizontal antes da primeira classe e depois da última classe.
- d. Traçar os segmentos de retas que ligam os pontos marcados.
- e. Colocar o título e o nome das variáveis nos eixos.



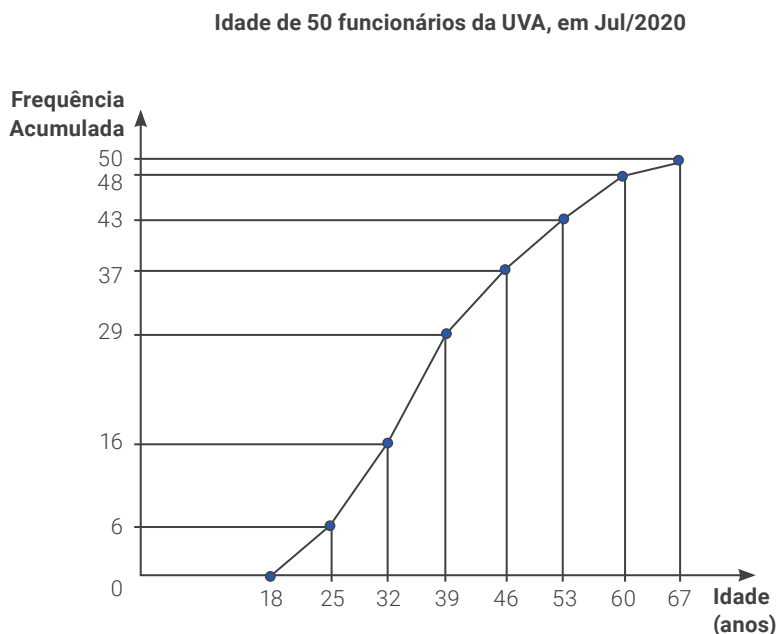
Fonte: Elaborados pela autora. Dados fictícios.

Polígono de frequência acumulada (Ogiva de Galton)

Para construir um polígono de frequência acumulada é necessário:

- 1- Marcar os limites das classes no eixo horizontal.
- 2- Marcar os pontos que correspondem aos limites superiores das classes e as respectivas frequências acumuladas de cada classe.
- 3- Traçar os segmentos de retas que ligam os pontos marcados.
- 4- Ligar os pontos marcados.
- 5- Colocar o título e o nome das variáveis nos eixos.

Figura: Ogiva de Galton da Idade de 50 funcionários da UVA, em julho de 2020.



Fonte: Elaborado pela autora. Dados fictícios.

Outros tipos de gráficos

Os gráficos têm a capacidade de comunicar os dados de forma visual eficaz e atraente. Os requisitos fundamentais para um bom gráfico devem ser simplicidade, clareza e veracidade.

Existem diversos tipos de gráficos: colunas, barras, linhas ou segmentos, setores ou pizza, pictogramas e muitos outros tipos. Com auxílio de computadores podemos representá-los em terceira dimensão e com movimentos, o que enriquece muito a visualização.

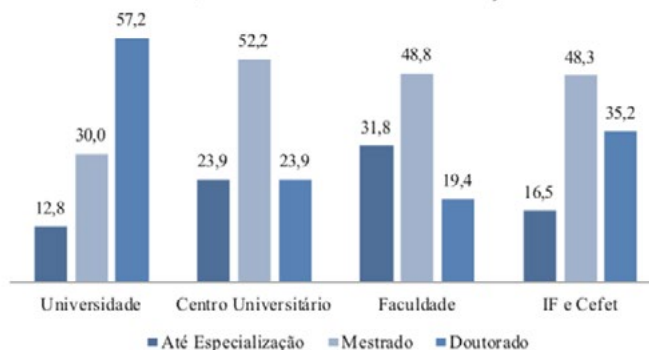
Observe na tabela a seguir os gráficos mais utilizados.

Tipo/Características	Representação																																																						
<h3>1. Gráfico de colunas</h3> <p>Representação de uma série por meio de retângulos de mesma base e alturas proporcionais aos respectivos dados.</p> <p>Utilizado quando apresentamos as categorias com palavras curtas.</p>	<p>NÚMERO DE MATRÍCULAS NA EDUCAÇÃO SUPERIOR (GRADUAÇÃO E SEQUENCIAL) – BRASIL – 2008-2018</p> <table><tr><th>Ano</th><th>Número de Matrículas</th></tr><tr><td>2008</td><td>5.843.322</td></tr><tr><td>2009</td><td>5.985.873</td></tr><tr><td>2010</td><td>6.407.733</td></tr><tr><td>2011</td><td>6.765.540</td></tr><tr><td>2012</td><td>7.058.084</td></tr><tr><td>2013</td><td>7.322.964</td></tr><tr><td>2014</td><td>7.839.765</td></tr><tr><td>2015</td><td>8.033.574</td></tr><tr><td>2016</td><td>8.052.254</td></tr><tr><td>2017</td><td>8.290.911</td></tr><tr><td>2018</td><td>8.451.748</td></tr></table> <p>Fonte: Elaboração própria com base em dados do Censo da Educação Superior 2018.</p>	Ano	Número de Matrículas	2008	5.843.322	2009	5.985.873	2010	6.407.733	2011	6.765.540	2012	7.058.084	2013	7.322.964	2014	7.839.765	2015	8.033.574	2016	8.052.254	2017	8.290.911	2018	8.451.748																														
Ano	Número de Matrículas																																																						
2008	5.843.322																																																						
2009	5.985.873																																																						
2010	6.407.733																																																						
2011	6.765.540																																																						
2012	7.058.084																																																						
2013	7.322.964																																																						
2014	7.839.765																																																						
2015	8.033.574																																																						
2016	8.052.254																																																						
2017	8.290.911																																																						
2018	8.451.748																																																						
<h3>2. Gráfico de barras</h3> <p>Representação de uma série por meio de retângulos de mesma altura e bases proporcionais aos respectivos dados.</p> <p>Utilizado quando apresentamos as categorias com palavras extensas</p>	<p>Ranking de Índice de Desenvolvimento Humano - IDH - 2015 Por país, por ordem de desenvolvimento</p> <table><tr><th>Rank</th><th>País</th><th>IDH</th></tr><tr><td>1</td><td>Noruega</td><td>0,95</td></tr><tr><td>2</td><td>Austrália</td><td>0,94</td></tr><tr><td>2</td><td>Suíça</td><td>0,94</td></tr><tr><td>4</td><td>Alemanha</td><td>0,93</td></tr><tr><td>5</td><td>Dinamarca</td><td>0,93</td></tr><tr><td>5</td><td>Cingapura</td><td>0,93</td></tr><tr><td>76</td><td>Libano</td><td>0,76</td></tr><tr><td>77</td><td>México</td><td>0,76</td></tr><tr><td>78</td><td>Azerbaijão</td><td>0,76</td></tr><tr><td>79</td><td>Brasil</td><td>0,75</td></tr><tr><td>79</td><td>Granada</td><td>0,75</td></tr><tr><td>81</td><td>Bósnia e Herzeg.</td><td>0,75</td></tr><tr><td>82</td><td>Macedônia</td><td>0,75</td></tr><tr><td>185</td><td>Burkina Faso</td><td>0,40</td></tr><tr><td>186</td><td>Chade</td><td>0,40</td></tr><tr><td>187</td><td>Níger</td><td>0,35</td></tr><tr><td>188</td><td>Rep. Centro africana</td><td>0,35</td></tr></table> <p>Fonte: Pnud</p>	Rank	País	IDH	1	Noruega	0,95	2	Austrália	0,94	2	Suíça	0,94	4	Alemanha	0,93	5	Dinamarca	0,93	5	Cingapura	0,93	76	Libano	0,76	77	México	0,76	78	Azerbaijão	0,76	79	Brasil	0,75	79	Granada	0,75	81	Bósnia e Herzeg.	0,75	82	Macedônia	0,75	185	Burkina Faso	0,40	186	Chade	0,40	187	Níger	0,35	188	Rep. Centro africana	0,35
Rank	País	IDH																																																					
1	Noruega	0,95																																																					
2	Austrália	0,94																																																					
2	Suíça	0,94																																																					
4	Alemanha	0,93																																																					
5	Dinamarca	0,93																																																					
5	Cingapura	0,93																																																					
76	Libano	0,76																																																					
77	México	0,76																																																					
78	Azerbaijão	0,76																																																					
79	Brasil	0,75																																																					
79	Granada	0,75																																																					
81	Bósnia e Herzeg.	0,75																																																					
82	Macedônia	0,75																																																					
185	Burkina Faso	0,40																																																					
186	Chade	0,40																																																					
187	Níger	0,35																																																					
188	Rep. Centro africana	0,35																																																					

3. Colunas justapostas

Descreve simultaneamente duas ou mais categorias para uma variável.

PERCENTUAL DO NÚMERO DE FUNÇÕES DOCENTES EM EXERCÍCIO, POR ORGANIZAÇÃO ACADÊMICA, SEGUNDO O GRAU DE FORMAÇÃO – BRASIL – 2017



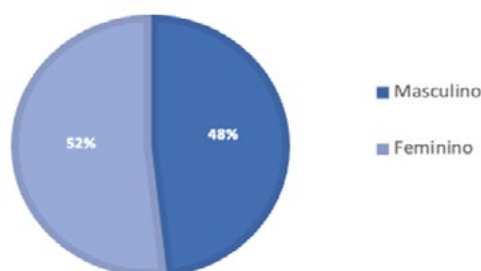
Fonte: Elaborada por Deed/Inep com base nos dados do Censo da Educação Superior.

4. Setores

Também conhecido como gráfico de pizza. É utilizado sempre que desejamos ressaltar a participação do dado no total.

Cada setor é obtido por meio de uma regra de três simples e direta, lembrando que 100% corresponde a 360°.

POPULAÇÃO BRASILEIRA, SEGUNDO O SEXO, EM 2019.

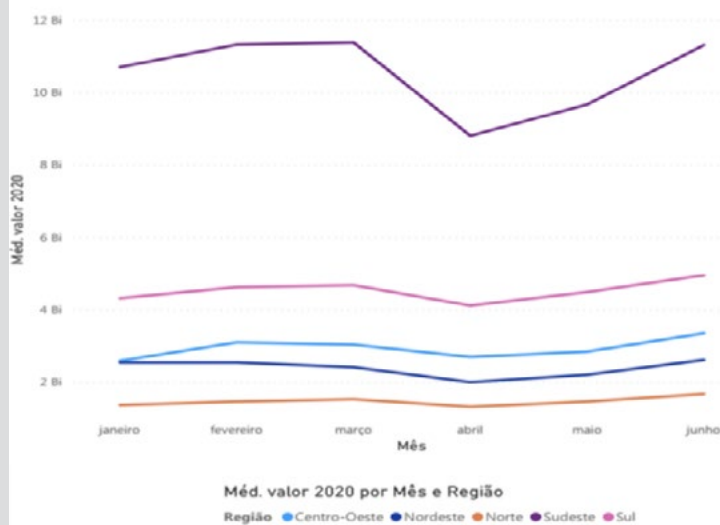


Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Trabalho e Rendimento, Pesquisa Nacional por Amostra de Domicílios Contínua 2012-2019.

5. Gráfico de linha ou gráfico de segmentos

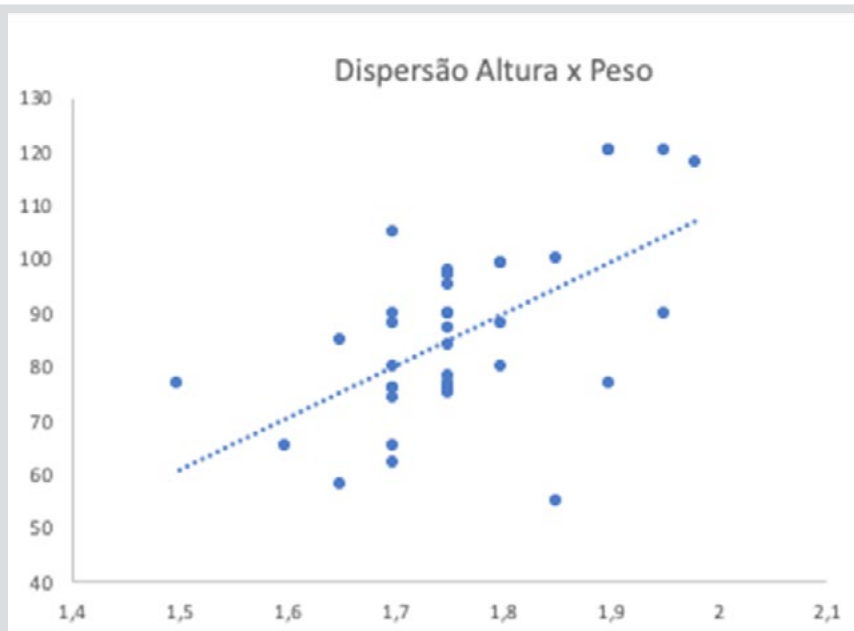
Marcamos todos os pares ordenados correspondentes à série estatística e os unimos a partir de uma linha poligonal tracejada ou contínua.

Impactos da Covid-19 no volume de vendas, por região brasileira



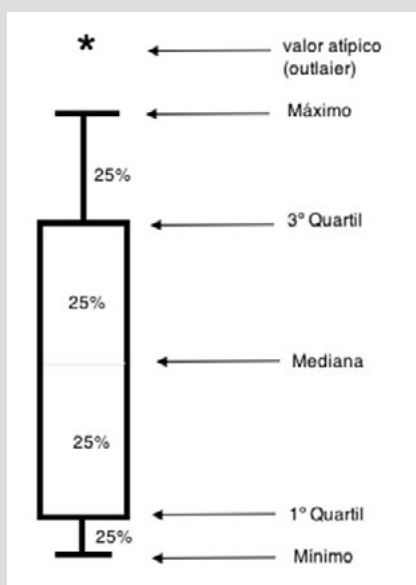
6. Diagrama de dispersão (scatterplot) e linha de tendência

O diagrama de dispersão, também conhecido como scatterplot, é a representação do relacionamento entre duas variáveis quantitativas. A linha de tendência é a linha que faz uma aproximação dessa relação.



7. Box Plot

É um gráfico que possibilita interpretar rapidamente o comportamento da distribuição em cada $\frac{1}{4}$ ou 25% (quartis). É muito útil para observarmos a existência de valores discrepantes (*outliers*) e a simetria da distribuição.



MIDIATECA

Na midiateca encontram-se alguns links que direcionam para apresentações com gráficos dinâmicos. São gráficos interessantes que se movimentam. Vale a pena conferir!

Para construir um gráfico é preciso saber que:

- Todo gráfico deve ter título e escala.
- O título deve ser escrito acima do gráfico.
- As variáveis devem ser identificadas nos eixos coordenados, inclusive com as unidades convenientes.
- As linhas auxiliares (grade) são opcionais, mas ajudam na leitura.
- A fonte é opcional nos gráficos.



Ampliando o foco

Para colocar em prática o conteúdo visto, faça os exercícios da seção 2.1 do livro LARSON, R.; FARBER, B. **Estatística aplicada**. São Paulo: Pearson, 2010, disponível na Biblioteca Virtual.

Confira as respostas que estão no final do livro e, se tiver alguma dúvida, fale com o seu tutor pelo Fórum de Dúvidas – Fale com o Tutor.

Medidas de posição e de dispersão

Neste tópico apresentaremos as medidas de posição e as medidas de dispersão, que também nos auxiliam na tarefa de resumir os dados.

Medidas de posição

São medidas usadas para resumir um conjunto de dados em um único valor. As medidas de posição mais importantes são as medidas de tendência central, que recebem esse nome pelo fato de que os dados observados tendem, em geral, a se agrupar em torno de um valor localizado no centro da distribuição de uma série de observações.

- Medidas de tendência central: a **média**, a **mediana** e a **moda**.
- Medidas de posição (**separatrizes**): os **quartis**, os **decis** e os **percentis** e a própria mediana.

Para calcularmos as medidas de posição poderemos encontrar os dados organizados de três maneiras, a saber:

Dados não agrupados.	Dados agrupados em tabelas SEM intervalos de classes.	Dados agrupados em tabelas COM intervalos de classe.
----------------------	---	--

Para cada situação acima temos uma metodologia de cálculo diferente. Por isso, antes de calcularmos as medidas de posição, primeiro temos que **identificar como os dados foram apresentados**.

Agora vamos conhecer cada metodologia.

1. Média Aritmética: (\bar{x} ou)

É a medida de posição mais conhecida e mais utilizada, embora nem sempre seja a mais representativa.

Se os dados referirem-se a uma amostra, indicamos a média por \bar{x} quando trabalhamos com a população indicamos pela letra grega μ .

Nas fórmulas estatísticas é comum representar a primeira observação da variável “x” por x_1 , o valor da segunda observação por x_2 e assim sucessivamente. Em geral o valor da i-ésima observação é indicado por x_i .

E quais as vantagens e desvantagens na utilização da média?



	
<p>Facilidade de se calcular.</p> <p>É a mais conhecida e mais utilizada.</p> <p>É usada para compararmos conjuntos semelhantes.</p>	<p>É fortemente influenciada por valores extremos e por isso só deve ser utilizada em distribuições simétricas.</p> <p>Não pode ser calculada para distribuições de frequências com limites indefinidos.</p>

Tabela: Cálculo da média aritmética.

<p>Para dados não agrupados</p>	$\bar{x} = \frac{\sum x_i}{n}$ <p>Ou</p> $\mu = \frac{\sum x_i}{N}$	<p>Em que:</p> <p>\bar{x} – média aritmética amostral. x_i – valores da variável. n – número de entradas em uma amostra ou μ – média aritmética populacional. x_i – valores da variável. N – número de entradas em uma população.</p>
<p>Para dados agrupados em tabelas sem intervalos de classes</p>	$\bar{x} = \frac{\sum x_i \cdot f_i}{n}$ <p>ou</p> $\mu = \frac{\sum x_i \cdot f_i}{N}$ <p>(média ponderada)</p>	<p>Em que:</p> <p>\bar{x} – média aritmética amostral. x_i – valores da variável. f_i – frequência da classe. n – número de entradas em uma amostra ou μ – média aritmética populacional. x_i – valores da variável. f_i – frequência da classe. N – número de entradas em uma população.</p>

Para dados agrupados em tabelas com intervalos de classes

$$\bar{X} = \frac{\sum x_i \cdot f_i}{n}$$

ou

$$\mu = \frac{\sum x_i \cdot f_i}{N}$$

Em que:

\bar{X} – média aritmética amostral.

μ – média aritmética populacional.

x_i – pontos médios das classes.

f_i – a frequência da classe.

n – número de entradas em uma amostra.

N – número de entradas em uma população.

Fonte: Elaborada pela autora (2020).

Observe alguns exemplos práticos.

Exemplo 1

Para dados não agrupados.

Suponha que a Central de Estágios da nossa universidade tenha enviado um questionário em julho de 2020, a uma amostra de 10 recém-graduados no curso de Engenharia Civil, solicitando-lhes informações acerca dos salários mensais iniciais. Os dados coletados foram os seguintes:

R\$ 5.800,00	R\$ 6.800,00	R\$ 8.000,00	R\$ 8.200,00	R\$ 8.000,00
R\$ 7.300,00	R\$ 7.300,00	R\$ 7.900,00	R\$ 7.900,00	R\$ 7.900,00

Vamos calcular a média aritmética dos salários mensais dos engenheiros.

Podemos observar que os dados não estão agrupados em tabelas. Utilize a fórmula de cálculo de média para amostras de **dados não agrupados**:

$$\bar{x} = \frac{\sum x_i}{n}$$
$$\bar{x} = \frac{5800 + 6800 + 8000 + 8200 + 8000 + 7300 + 7300 + 7900 + 7900 + 7900}{10}$$
$$\bar{x} = \frac{75100}{10} = 7510$$

Interpretação do resultado: os salários mensais dos engenheiros civis pesquisados giram em torno de R\$ 7.510,00.

Exemplo 2

Para dados agrupados em tabelas sem intervalo de classes.

Suponha que as anotações dos salários vieram organizadas em uma tabela sem intervalos de classes. Vamos calcular a média dos salários dos engenheiros.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)
5800	1
6800	1
7300	2
7900	3
8000	2
8200	1
Total	10

Fonte: Central de Estágios da UVA. Dados fictícios.

Utilizando a fórmula para cálculo da média para **dados agrupados em tabelas sem intervalo de classes** temos:

$$\bar{X} = \frac{\sum x_i \cdot f_i}{n}$$

$$\bar{X} = \frac{5800 \cdot 1 + 6800 \cdot 1 + 7300 \cdot 2 + 7900 \cdot 3 + 8000 \cdot 2 + 8200 \cdot 1}{10}$$

$$\bar{X} = \frac{5800 + 6800 + 14600 + 23700 + 16000 + 8200}{10}$$

$$\bar{X} = \frac{75100}{10} = 7510$$

Exemplo 3

Para dados agrupados em tabelas com intervalo de classes.

Agora, vamos supor que as anotações dos salários vieram organizadas em uma tabela com intervalos de classes. Vamos calcular a média dos salários dos engenheiros.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)
5000 I---- 6000	1
6000 I--- 7000	1
7000 I--- 8000	5
8000 I--- 9000	3
Total	10

Fonte: Central de Estágios da UVA. Dados fictícios.

Primeiramente, antes de calcular a média, temos que calcular o ponto médio de cada classe.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)	Ponto médio da classe (Xi)
5000 ---- 6000	1	$\frac{5000 + 6000}{2} = 5500$
6000 --- 7000	1	$\frac{6000 + 7000}{2} = 6500$
7000 --- 8000	5	$\frac{7000 + 8000}{2} = 7500$
8000 --- 9000	3	$\frac{8000 + 9000}{2} = 8500$
Total	10	

Fonte: Central de Estágios da UVA. Dados fictícios.

Agora podemos calcular a média dos salários:

$$\bar{X} = \frac{\sum x_i \cdot f_i}{n}$$

$$\bar{X} = \frac{5500 \cdot 1 + 6500 \cdot 1 + 7500 \cdot 5 + 8500 \cdot 3}{10}$$

$$\bar{X} = \frac{5500 + 6500 + 37500 + 25500}{10}$$

$$\bar{X} = \frac{75000}{10} = 7500$$

Note que, ao agruparmos os dados com intervalos de classes, perdemos um pouco a precisão das medidas. Por isso, devemos, sempre que possível, usar entre 5 e 20 classes.

Na midiateca você encontra o endereço para pesquisar mais sobre os salários médios e medianos das categorias profissionais.

1. Mediana

Muitas vezes encontramos nas distribuições assimétricas valores discrepantes (valores muito grandes ou muito pequenos) e que faz com que a média seja muito influenciada. Veja um exemplo a seguir.



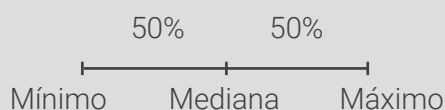
Exemplo

Suponha que as notas dos alunos sejam as seguintes: 0; 5; 5; 5.

Sem fazer nenhum cálculo podemos observar que as notas giram em torno de 5, porém se calcularmos a média obtemos como resposta o valor 3,75.

Nesses casos, em que temos alguns valores muito altos ou muito baixos em relação aos outros valores, é mais conveniente usarmos outra medida de posição, a mediana.

A mediana é o valor central do rol, ou seja, o valor que divide um grupo de valores ordenados em duas partes com o mesmo número de termos, como mostra a imagem a seguir:



E quais as vantagens e desvantagens na utilização da mediana?



Sua determinação é rápida e fácil, sem necessitar de cálculos complexos.

Não é influenciada por valores extremos.

Pode ser calculada quando temos limites de classes indefinidos, pois só depende do tamanho da amostra ou da população.



A mediana flutua mais de amostra para amostra do que a média aritmética, portanto é menos confiável.

Não utiliza todos os dados.

Não é possível para todos os dados.

Não é possível calcular a mediana de um grupo total a partir das medianas de dois subgrupos.

Não é levada em consideração na maioria dos testes estatísticos.

Tabela: Cálculo da Mediana.

Para dados não agrupados

1 – Ordenar os dados numéricos, ou seja, confeccionar o rol.

2 – Se o tamanho da amostra ou da população for um número ímpar, a mediana é o valor que estiver no centro do rol.

3 – Se o tamanho da amostra ou da população for um número par, a mediana é a média aritmética dos dois valores centrais do rol.

Para dados agrupados em tabelas sem intervalos de classes

1 – Calcular a posição ocupada mediana, que denominamos de Posto da Mediana.

a) Se a quantidade de dados for “par” usamos:

$$P = \frac{\sum f_i}{2}$$

b) Se a quantidade de dados for “ímpar” usamos:

$$P = \frac{1 + \sum f_i}{2}$$

2 – Calcula-se as frequências **acumuladas**.

3 – Identifique na coluna das frequências acumuladas, a primeira classe que contenha o posto da mediana. O valor dessa classe, será o valor da mediana.

Para dados agrupados em tabelas com intervalos de classes

1) Calcular o posto da mediana

$$P = \frac{\sum f_i}{2}$$

2) Identificar a primeira classe que contenha o posto da mediana.

3) Utilizar a fórmula de igualdade dos quocientes:

$$M_d = L_i + h \cdot \frac{\frac{\sum f_i}{2} - F_{a.ant}}{f_i}$$

Em que:

M_d = mediana.

L_i = limite inferior do Posto da Mediana.

h = amplitude da classe que contém o posto da mediana.

$\sum f_i$ = tamanho da amostra ou da população.

$F_{a.ant}$ = frequência acumulada da classe anterior a classe que contém o posto da mediana.

f_i = frequência absoluta da classe que contém o posto da mediana.

Observe alguns exemplos práticos.

Exemplo 1

Para dados não agrupados

Voltemos aos dados dos salários dos engenheiros recém-formados coletados pela Central de Estágios em julho de 2020:

R\$ 5.800,00	R\$ 6.800,00	R\$ 8.000,00	R\$ 8.200,00	R\$ 8.000,00
R\$ 7.300,00	R\$ 7.300,00	R\$ 7.900,00	R\$ 7.900,00	R\$ 7.900,00

Vamos determinar a mediana dos salários mensais dos engenheiros. Para isso:

1º) Confeccione o Rol.

5800	6800	7300	7300	7900	7900	7900	8000	8000	8200
------	------	------	------	------	------	------	------	------	------

Como o tamanho da amostra é igual a 10 (par) não encontramos um valor no centro do rol e sim dois valores. Devemos, então, fazer a média aritmética desses dois valores:

$$M_d = \frac{7900 + 7900}{2} = 7900$$

Exemplo 2

Para dados agrupados em tabelas sem intervalo de classe.

Agora vamos determinar a mediana para dados agrupados em tabelas sem intervalo de classe:

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)
5800	1
6800	1

7300	2
7900	3
8000	2
8200	1
Total	10

Fonte: Central de Estágios da UVA. Dados fictícios.

1º) Calcule o posto da mediana:

$$P = \frac{\sum f_i}{2}$$

$$P = \frac{10}{2} = 5$$

2º) Determine as frequências acumuladas:

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)	Frequência Acumulada
5800	1	1
6800	1	2
7300	2	4
7900	3	7
8000	2	9
8200	1	10
Total	10	

Fonte: Central de Estágios da UVA. Dados fictícios.

3º) Procure na coluna da frequência acumulada a primeira classe que contém o posto da mediana, que nesse exemplo foi igual a 5.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)	Frequência Acumulada (é nessa coluna que faremos a análise, mas não é aqui que estará o valor da mediana)
5800	1	1 (não contém, pois é menor do que 5)
6800	1	2 (não contém, pois é menor do que 5)
7300	2	4 não contém, pois é menor do que 5)
7900	3	7 (ok! A frequência acumulada 7 é maior do que 5) essa é a classe da mediana
8000	2	9
8200	1	10
Total	10	

Fonte: Central de Estágios da UVA. Dados fictícios.

Logo, a mediana será igual ao valor da variável que estiver na classe selecionada, a primeira classe em que encontramos o valor da frequência acumulada maior do que o valor do posto. Então: $Md = 7900$

Exemplo 3

Para os dados agrupados em tabelas com intervalos de classes.

Finalmente, vamos calcular a mediana para os dados agrupados em tabelas com intervalos de classes.

Vamos continuar trabalhando com o exemplo dos salários dos engenheiros, porém agora agrupados em uma tabela com intervalos. Observe a tabela a seguir.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)
5000 I---- 6000	1
6000 I--- 7000	1
7000 I--- 8000	5
8000 I--- 9000	3
Total	10

Fonte: Central de Estágios da UVA. Dados fictícios.

1º) Calcule o posto da mediana:

$$P = \frac{\sum f_i}{2}$$

$$P = \frac{10}{2} = 5$$

2º) Determine as frequências acumuladas:

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)	Frequência Acumulada
5000 I---- 6000	1	1
6000 I--- 7000	1	2
7000 I--- 8000	5	7
8000 I--- 9000	3	10
Total	10	

Fonte: Central de Estágios da UVA. Dados fictícios.

3º) Procure na coluna da frequência acumulada a primeira classe que contém o posto da mediana, que nesse exemplo foi igual a 5.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)	Frequência Acumulada
5000 --- 6000	1	1 (1<5)
6000 --- 7000	1	2 (2<5)
7000 --- 8000	5	7 (7>5 ok!)
8000 --- 9000	3	10
Total	10	

Fonte: Central de Estágios da UVA (dados fictícios).

Observando a classe do Posto da Mediana, temos:

Limite inferior da classe: $L_i = 7000$

Amplitude da classe: $h = 8000 - 7000 = 1000$

Frequência absoluta da classe: $f_i = 5$

Frequência acumulada da classe anterior: $F_{a.ant} = 2$

Metade dos dados:

$$\frac{\sum f_i}{2} = \frac{10}{2} = 5$$

Substituindo os valores na fórmula, temos:

$$M_d = L_i + h \cdot \frac{\frac{\sum f_i}{2} - F_{a.ant}}{f_i}$$

$$M_d = 7000 + 1000 \cdot \frac{5 - 2}{5}$$

$$M_d = 7000 + 1000 \cdot 0,6$$

$$M_d = 7000 + 600 = 7600$$



Ampliando o foco

Os **fractis** são números que dividem um conjunto de dados ordenados em partes iguais, ou seja, em partes que contêm a mesma quantidade de elementos da série.

A mediana é um fractil porque divide o rol em duas partes iguais.

Os **quartis Q_1 , Q_2 , e Q_3** dividem, aproximadamente, um conjunto de dados ordenados em quatro partes iguais.

Os **percentis** dividem, aproximadamente, um conjunto de dados ordenados em 100 partes iguais.

E assim por diante: Decis (10 partes) Quintis (5 partes).

2. MODA

É o valor, ou valores, da distribuição que ocorrem com a maior frequência, ou seja, o valor, ou valores que mais se repetem dentro de uma série de observações. De maneira geral:

- Se todos os valores repetem-se a mesma quantidade de vezes, dizemos que não há moda, ou seja, a distribuição é amodal.
- Se um valor ocorre com maior frequência, chamamos a distribuição de unimodal.
- Se dois valores repetem-se com a mesma quantidade e com maior frequência, chamamos a distribuição de bimodal.
- Três valores, então trimodal.
- Mais de três valores, dizemos que a distribuição é multimodal ou polimodal.

Vantagens e desvantagens na utilização da moda:

- ✓ Pode ser usada quando trabalhamos com variáveis qualitativas (categóricas).
- ✓ Não é influenciada por valores extremos.
- ✓ Pode ser calculada para distribuições com limites indefinidos.
- ✗ Não leva em consideração todos os valores do conjunto de dados.

Tabela: Cálculo da Moda.

Para dados não agrupados	Basta verificar qual é o valor que mais se repete.
Para dados agrupados em tabelas sem intervalos de classes	Basta observar a classe, ou classes, de maior frequência absoluta.
Para dados agrupados em tabelas com intervalos de classes (Vamos abordar quatro processos diferentes)	<div> <p>Em que:</p> <p>M_o = moda.</p> <p>Classe modal= classe de maior frequência.</p> <p>L_i = limite inferior da classe modal.</p> <p>D_1 = diferença entre a frequência absoluta da classe modal e a frequência da classe anterior a ela.</p> <p>D_2 = diferença entre a frequência absoluta da classe modal e a frequência da classe posterior a ela.</p> <p>h = amplitude da classe modal.</p> <p>f_{ant} = frequência simples da classe anterior à classe modal.</p> <p>f_{post} = frequência simples da classe posterior à classe modal.</p> <p>M_d = mediana.</p> <p>\bar{x} ou μ = média.</p> </div> <div> <p>- Moda bruta: É o ponto médio da classe de maior frequência.</p> <p>- Moda pelo processo de Czuber:</p> $M_o = L_i + \frac{D_1}{D_1 + D_2} \cdot h$ <p>- Moda pelo processo de King:</p> $M_o = L_i + \frac{f_{post}}{f_{ant} + f_{post}} \cdot h$ <p>- Moda pelo processo de Pearson</p> $M_o = 3 \cdot M_d - 2 \cdot \bar{x}$ </div>

Exemplo 1

Para dados não agrupados.

Voltemos aos dados dos salários dos engenheiros recém-formados coletados pela Central de Estágios em julho de 2020:

R\$ 5.800,00	R\$ 6.800,00	R\$ 8.000,00	R\$ 8.200,00	R\$ 8.000,00
R\$ 7.300,00	R\$ 7.300,00	R\$ 7.900,00	R\$ 7.900,00	R\$ 7.900,00

Vamos determinar a moda dos salários mensais dos engenheiros. Observamos no conjunto de dados que o valor que mais se repetiu foi R\$ 7.900,00. Portanto, esse valor é o valor da moda.

Exemplo 2

Para dados agrupados em tabelas sem intervalo de classe.

Agora vamos determinar a moda para dados agrupados em tabelas sem intervalo de classe.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)
5800	1
6800	1
7300	2
7900	3
8000	2
8200	1

Observamos que a 4ª classe apresenta a maior frequência absoluta. Logo, a média será o valor da variável nessa classe, ou seja, R\$ 7.900,00.

Exemplo 3

Para dados agrupados em tabelas com intervalo de classe

Vamos continuar trabalhando com o exemplo dos salários dos engenheiros, porém agora agrupados em uma tabela com intervalos.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)
5000 I---- 6000	1
6000 I--- 7000	1
7000 I--- 8000	5
8000 I--- 9000	3
Total	10

Fonte: Central de Estágios da UVA (dados fictícios).

Classe Modal → classe de maior frequência: 3ª classe: 7000 I--- 8000

$$L_1 = 7000$$

$$D_1 = 5 - 1 = 4$$

$$D_2 = 5 - 3 = 2$$

$$h = 8000 - 7000 = 1000$$

$$f_{ant} = 1$$

$$f_{post} = 3$$

$$M_d = 7600$$

$$\bar{x} = 7500$$

$$\text{– Moda bruta: } M_o = \frac{7000 + 8000}{2} = 7500$$

Moda pelo processo de Czuber

$$M_o = L_i + \frac{D_1}{D_1 + D_2} \cdot h$$

$$M_o = 7000 + \frac{4}{4 + 2} \cdot 1000$$

$$M_o = 7000 + 0,66667 \cdot 1000$$

$$M_o = 7000 + 666,67$$

$$M_o = 7666,67$$

Moda pelo processo de King

$$M_o = L_i + \frac{f_{\text{post}}}{f_{\text{ant}} + f_{\text{post}}} \cdot h$$

$$M_o = 7000 + \frac{3}{1 + 3} \cdot 1000$$

$$M_o = 7000 + 0,75 \cdot 1000$$

$$M_o = 7000 + 750$$

$$M_o = 7750$$

Moda pelo processo de Pearson

$$M_o = 3 \cdot M_d - 2 \cdot \bar{x}$$

$$M_o = 3 \cdot 7600 - 2 \cdot 7500$$

$$M_o = 22800 - 15000$$

$$M_o = 7800$$

Relações entre média, mediana e moda

Ao traçar o histograma e o polígono de frequências de uma distribuição podemos verificar, visualmente, qual é o comportamento dos dados.

Agora vamos relacionar a imagem dos gráficos às medidas de posição.

Se média = mediana = moda	Se média > mediana > moda	Se média < mediana < moda
Distribuição simétrica	Distribuição assimétrica positiva	Distribuição assimétrica negativa
		

Fonte: Elaborada pela autora (2020).

A forma da distribuição pode interferir na escolha da medida de tendência central. Ou seja:

- Se a distribuição for aproximadamente simétrica, a média aritmética é a mais indicada, mesmo porque esta poderá ser utilizada em estatística mais avançada e é uma medida mais estável.
- Se a pretensão for uma medida descritiva, rápida e simples ainda que grosseira, e se a distribuição for unimodal, o pesquisador deve escolher a moda.
- Se a pretensão for uma medida exata, ele poderá optar entre a média e a mediana.



Ampliando o foco

Para colocar em prática o conteúdo visto, faça os exercícios da seção 2.3 do livro LARSON, R.; FARBER, B. **Estatística aplicada**. São Paulo: Pearson, 2010, disponível na biblioteca virtual.

Confira as respostas que estão no final do livro e, se tiver alguma dúvida, fale com o seu tutor pelo Fórum de Dúvidas – Fale com o Tutor.

Medidas de dispersão ou de variabilidade

Suponha que temos três turmas diferentes, cada uma com 10 alunos, para as quais foi aplicada uma avaliação com valor de 10 pontos.

As notas foram as seguintes:

Notas Turmas	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	Total
Turma "A"	5	5	5	5	5	5	5	5	5	5	50
Turma "B"	0	0	0	5	5	5	5	10	10	10	50
Turma "C"	1	2	3	4	5	5	6	7	8	9	50

Se calcularmos a média, a moda e a mediana para cada turma teremos:

Turmas	Média	Mediana	Moda
Turma "A"	5	5	5
Turma "B"	5	5	5
Turma "C"	5	5	5

O que podemos concluir?

Analisando somente as medidas de posição, concluímos que essas três turmas são idênticas, o que seria um grande erro.

Para verificar o comportamento dos dados e o grau de dispersão em torno da média, precisamos de outras medidas estatísticas, que são as **Medidas de Dispersão** ou **Variabilidade**, ou seja:

- A Amplitude.
- O Desvio Médio.
- A Variância, o Desvio-Padrão e o Coeficiente de Variação.

Agora veremos com mais detalhes cada uma delas. Vamos lá!

Amplitude (Range)

É a diferença entre o maior e o menor valor do conjunto.

Voltando ao exemplo das notas das três turmas, temos:

- **Turma A:** $5 - 5 = 0$ (nenhuma dispersão, turma homogênea).
- **Turma B:** $10 - 0 = 10$ (dados muito dispersos, turma heterogênea).
- **Turma C:** $9 - 1 = 8$ (dados muito dispersos, turma heterogênea, porém mais homogênea do a turma B).

Para dados agrupados com intervalos de classe a amplitude total é:

A diferença entre o limite superior da última classe e o limite inferior da primeira classe.

A amplitude é muito pouco usada, pois tem uma grande desvantagem que a é de levar em conta apenas dois valores, desprezando todos os outros.

Desvio: (D)

É a diferença entre cada valor observado e a média destes valores.

$$D = x_i - \bar{x}$$

Propriedade importante: a soma dos desvios em relação à média é sempre igual a zero.

Desvio médio: (Dm)

É a média aritmética dos módulos dos desvios.

- **Para dados não agrupados:** $D_m = \frac{\sum |x_i - \bar{x}|}{n}$

- **Para dados agrupados:** $D_m = \frac{\sum |x_i - \bar{x}| \cdot f_i}{n}$

Variância: (s^2 ou σ^2)

É a média aritmética dos quadrados dos desvios em relação à média da distribuição.

Tabela: Cálculo da Variância para dados não agrupados, dados agrupados em tabelas de distribuição sem intervalos de classes e em tabelas de distribuição com intervalos de classes.

	Amostras	População	Em que:
Para dados não agrupados.	$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$	$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$	s^2 ou σ^2 = variância n = tamanho da amostra N = tamanho da população \bar{x} = média amostral μ = média populacional x_i = valor observado X_i = ponto médio da classe
Para dados agrupados em tabelas sem intervalos de classes.	$s^2 = \frac{\sum (x_i - \bar{x})^2 \cdot f_i}{n - 1}$	$\sigma^2 = \frac{\sum (x_i - \mu)^2 \cdot f_i}{N}$	
Para dados agrupados em tabelas com intervalos de classes.	$s^2 = \frac{\sum (X_i - \bar{x})^2 \cdot f_i}{n - 1}$	$\sigma^2 = \frac{\sum (X_i - \mu)^2 \cdot f_i}{N}$	

Observe:

- Os denominadores são diferentes. Para amostras usamos “ $n - 1$ ” e para população usamos “ n ”.
- Como no cálculo da variância eleva-se o desvio ao quadrado, a unidade de medida resultante sempre será a unidade de medida original elevada ao quadrado. Por exemplo, seriam reais (dinheiro) ao quadrado.
- Como esse tipo de unidade de medida não faz sentido, o que ocorre em muitos casos é extrair a raiz quadrada da variância, obtendo um valor que tem a mesma unidade de medida que os dados originais — o desvio-padrão.

Desvio-padrão: (s ou σ)

É a raiz quadrada da variância.

$$s = \sqrt{s^2} \quad \text{para amostras}$$

ou

$$\sigma = \sqrt{\sigma^2} \quad \text{para população}$$

Coeficiente de Variação

É o quociente entre o desvio-padrão e a média de uma distribuição.

Normalmente, multiplicamos o resultado por 100 para expressarmos essa medida em porcentagem.

Para saber se uma dispersão é muito grande em relação à média ou para comparar o grau de dispersão relativa entre as distribuições de duas ou mais variáveis é preciso utilizar uma medida de dispersão relativa, que se chama Coeficiente de Variação.

O coeficiente de variação é útil para comparar a variabilidade de observações com diferentes unidades de medida, como:

$$CV = \frac{s}{\bar{x}} \quad \text{para amostras}$$

ou

$$CV = \frac{\sigma}{\mu} \quad \text{para população}$$

Veja a interpretação dos valores dos coeficientes de variação. A saber:

- **Dispersão baixa:** quando CV inferior a 10%.
- **Dispersão média:** quando CV estiver entre 10% e 20%.
- **Dispersão alta:** quando CV estiver entre 20% e 30%.
- **Dispersão muito alta:** quando CV for superior a 30% (pequena representatividade da média).

Exemplo 1

Para dados não agrupados

Voltemos aos dados dos salários dos engenheiros recém-formados coletados pela Central de Estágios em julho de 2020.

R\$ 5.800,00 R\$ 6.800,00 R\$ 8.000,00 R\$ 8.200,00 R\$ 8.000,00
R\$ 7.300,00 R\$ 7.300,00 R\$ 7.900,00 R\$ 7.900,00 R\$ 7.900,00

Vamos determinar a variância, o desvio-padrão e o coeficiente de variação dos salários mensais dos engenheiros.

1º) Calcule a média aritmética da distribuição.

$$\bar{X} = \frac{\sum x_i}{n}$$

$$\bar{X} = \frac{5800 + 6800 + 8000 + 8200 + 8000 + 7300 + 7300 + 7900 + 7900 + 7900}{n}$$

$$\bar{X} = \frac{75100}{10} = 7510$$

2º) Calcule os desvios em relação à média e depois eleve cada desvio ao quadrado.

desvio	desvio ²
5800 - 7510 = -1710	2924100
6800 - 7510 = -710	504100
7300 - 7510 = -210	44100
7300 - 7510 = -210	44100

7900 - 7510 =	390	152100
7900 - 7510 =	390	152100
7900 - 7510 =	390	152100
8000 - 7510 =	490	240100
8200 - 7510 =	690	476100

3º) Some de todos os desvios elevados ao quadrado.

desvio		desvio ²
5800 - 7510 =	-1710	2924100
6800 - 7510 =	-710	504100
7300 - 7510 =	-210	44100
7300 - 7510 =	-210	44100
7900 - 7510 =	390	152100
7900 - 7510 =	390	152100
7900 - 7510 =	390	152100
8000 - 7510 =	490	240100
8000 - 7510 =	490	240100
8200 - 7510 =	690	476100
		4929000

Como estamos trabalhando com uma amostra, para calcular a variância devemos dividir esse somatório por “n-1”:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} = \frac{4929000}{10 - 1} = 547666,67$$

O desvio-padrão será a raiz quadrada da variância:

$$s = \sqrt{547666,67} = 740,05$$

O coeficiente de variação é a divisão entre o desvio-padrão e a média.

$$CV = \frac{740,05}{7510} = 0,0985$$

Expressando em porcentagem temos:

$$CV = 0,0985 \cdot 100 = 9,85\% \text{ (baixa dispersão)}$$

Exemplo 2

Para dados agrupados em tabelas sem intervalo de classe.

Agora vamos determinar as medidas de dispersão para dados agrupados em tabelas sem intervalo de classe.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)
5800	1
6800	1
7300	2
7900	3

8000	2
8200	1
Total	10

Fonte: Central de Estágios da UVA (dados fictícios).

1º) Calcule a média.

$$\bar{x} = \frac{\sum x_i \cdot f_i}{n}$$

$$\bar{x} = \frac{5800 \cdot 1 + 6800 \cdot 1 + 7300 \cdot 2 + 7900 \cdot 3 + 8000 \cdot 2 + 8200 \cdot 1}{10}$$

$$\bar{x} = \frac{5800 + 6800 + 14600 + 23700 + 16000 + 8200}{10}$$

$$\bar{x} = \frac{75100}{10} = 7510$$

2º) Determine os desvios em relação à média.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)	desvio = xi-média	
5800	1	5800 - 7510 =	-1710
6800	1	6800 - 7510 =	-710
7300	2	7300 - 7510 =	-210
7900	3	7900 - 7510 =	390

8000	2	$8000 - 7510 =$	490
8200	1	$8200 - 7510 =$	690
Total	10		

Fonte: Central de Estágios da UVA (dados fictícios).

3º) Eleve o desvio ao quadrado.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)	desvio = xi-média		desvio ²
5800	1	$5800 - 7510 =$	-1710	2924100
6800	1	$6800 - 7510 =$	-710	504100
7300	2	$7300 - 7510 =$	-210	44100
7900	3	$7900 - 7510 =$	390	152100
8000	2	$8000 - 7510 =$	490	240100
8200	1	$8200 - 7510 =$	690	476100
Total	10			

Fonte: Central de Estágios da UVA (dados fictícios).

4º) Multiplique os quadrados dos desvios pelas respectivas frequências absolutas e fazer o somatório.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)	desvio = xi-média		desvio²	desvio² . fi	
5800	1	5800 - 7510 =	-1710	2924100	2924100 . 1 =	2924100
6800	1	6800 - 7510 =	-710	504100	504100 . 1 =	504100
7300	2	7300 - 7510 =	-210	44100	44100 . 2 =	88200
7900	3	7900 - 7510 =	390	152100	152100 . 3 =	456300
8000	2	8000 - 7510 =	490	240100	240100 . 2 =	480200
8200	1	8200 - 7510 =	690	476100	476100 . 1 =	476100
Total	10					4929000

Fonte: Central de Estágios da UVA (dados fictícios).

Como estamos trabalhando com uma amostra, para calcular a variância, devemos dividir esse somatório por “n-1”:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{4929000}{10 - 1} = 547666,67$$

O desvio-padrão será a raiz quadrada da variância:

$$s = \sqrt{547666,67} = 740,05$$

O coeficiente de variação é a divisão entre o desvio-padrão e a média.

$$CV = \frac{740,05}{7510} = 0,0985$$

Expressando em porcentagem temos:

$$CV = 0,0985 \cdot 100 = 9,85\% \text{ (baixa dispersão)}$$

Exemplo 3

Para dados agrupados em uma tabela de distribuição com intervalos de classes.

Finalmente vamos calcular as medidas de dispersão para os dados agrupados em uma tabela de distribuição com intervalos de classes.

Voltemos ao exemplo.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (f_i)	Ponto médio da classe (X_i)
5000 I---- 6000	1	$\frac{5000 + 6000}{2} = 5500$
6000 I--- 7000	1	$\frac{6000 + 7000}{2} = 6500$
7000 I--- 8000	5	$\frac{7000 + 8000}{2} = 7500$
8000 I--- 9000	3	$\frac{8000 + 9000}{2} = 8500$
Total	10	

Fonte: Central de Estágios da UVA (dados fictícios).

1º) Calcule a média dos salários com a fórmula dos dados agrupados com intervalos.

$$\bar{X} = \frac{\sum X_i \cdot f_i}{n}$$

$$\bar{x} = \frac{5500 \cdot 1 + 6500 \cdot 1 + 7500 \cdot 5 + 8500 \cdot 3}{10}$$

$$\bar{x} = \frac{5500 + 6500 + 37500 + 25500}{10}$$

$$\bar{x} = \frac{75000}{10} = 7500$$

2º) Calcule o desvio de cada ponto médio em relação à média, fazendo a diferença entre o valor do ponto médio e a média.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)	Ponto médio da classe (Xi)	Desvio = = Ponto Médio - Média	
5000 I--- 6000	1	5500	5500 - 7500 =	-2000
6000 I--- 7000	1	6500	6500 - 7500 =	-1000
7000 I--- 8000	5	7500	7500 - 7500 =	0
8000 I--- 9000	3	8500	8500 - 7500 =	1000
Total	10			

Fonte: Central de Estágios da UVA (dados fictícios).

3º) Eleve os desvios ao quadrado.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)	Ponto médio da classe (Xi)	Desvio = = Ponto Médio - Média	Desvio ²
5000 I--- 6000	1	5500	5500 - 7500 = -2000	4000000
6000 I--- 7000	1	6500	6500 - 7500 = -1000	1000000

7000 I--- 8000	5	7500	7500 - 7500 =	0
8000 I--- 9000	3	8500	8500 - 7500 =	1000000
Total	10			

Fonte: Central de Estágios da UVA (dados fictícios).

4º) Multiplique o quadrado do desvio pela respectiva frequência absoluta e fazer o somatório dessa coluna.

Tabela: Salário mensal de 10 engenheiros civis em julho de 2020.

Salário	Quantidade de pessoas (fi)	Ponto médio da classe (Xi)	Desvio = Ponto Médio - Média	Desvio²	Desvio² . fi
5000 I---- 6000	1	5500	5500 - 7500 = -2000	4000000	4000000 . 1 = 4000000
6000 I--- 7000	1	6500	6500 - 7500 = -1000	1000000	1000000
7000 I--- 8000	5	7500	7500 - 7500 = 0	0	0 . 5 = 0
8000 I--- 9000	3	8500	8500 - 7500 = 1000	1000000	1000000 . 3 = 3000000
Total	10				8000000

Fonte: Central de Estágios da UVA (dados fictícios).

Como estamos trabalhando com uma amostra, para calcular a variância devemos dividir esse somatório por “n-1”:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} = \frac{8000000}{10 - 1} = 888888,89$$

O desvio-padrão será a raiz quadrada da variância:

$$s = \sqrt{888888,89} = 942,81$$

O coeficiente de variação é a divisão entre o desvio-padrão e a média.

$$CV = \frac{942,81}{7500} = 0,126$$

Expressando em porcentagem, temos:

$$CV = 0,126 \cdot 100 = 12,5\% \text{ (média dispersão)}$$

Mais uma vez notamos que os dados agrupados perdem um pouco a precisão, principalmente se forem agrupados em poucas classes.



Ampliando o foco

Para praticar o conteúdo visto, faça os exercícios da seção 2 do livro LARSON, R.; FARBER, B. **Estatística aplicada**. São Paulo: Pearson, 2010, disponível na Biblioteca Virtual.

Confira as respostas que estão no final do livro e, se tiver alguma dúvida, entre em contato com o seu tutor pelo Fórum de Dúvidas – Fale com o Tutor.

Estatísticas descritivas usando a planilha eletrônica Excel

O Excel oferece funções para calcular as medidas estatísticas. Contudo, serve apenas para dados não agrupados.

Suponha que os dados estejam digitados nas células B1 até a célula B50, como mostra a tabela a seguir.

	Comando
Frequência absoluta	=CONT.SE
Média	=MEDIA(B1:B50)

Mediana	=MED(B1:B50)
Moda	=MODO(B1:B50)
Variância amostral	=VAR.A(B1:B50)
Variância populacional	=VAR.P(B1:B50)
Desvio-padrão amostral	=DESVPAD.A(B1:B50)
Desvio-padrão populacional	=DESVPAD.P(B1:B50)



MIDIAATECA

Para ampliar o seu conhecimento veja o material complementar da Unidade 1, disponível na midiateca.



NA PRÁTICA

No portal do Conselho Regional de Estatística da 3ª Região (SP), publicado no dia 23 de maio de 2020 por Camila Soares, encontramos uma série de entrevistas com profissionais que falam sobre o uso da estatística no cotidiano e mostram como ela é fundamental na tomada de decisões, além de contribuir com outras ciências e com o mercado de trabalho dessa área.

Veja alguns trechos dessa reportagem:

Exemplos da utilização da Estatística em Customer Relationship Management CRM

O que é Customer Relationship Management (CRM)? O CRM é um termo que designa o gerenciamento do relacionamento com clientes, em geral por meio de um sistema.

Como funciona na prática? Não basta ter um banco de dados ou conjunto de dados com informações dos clientes. Também não basta fazer alguns cálculos em cima desses bancos de dados. É preciso entender a estratégia de negócios da empresa e como opera o mercado. Além disso, é necessário dispor de técnicas para analisar os dados. É nessa parte que a formação em Estatística é muito relevante quanto ao uso de técnicas para analisar esses bancos de dados. Isso ajuda a empresa a alcançar melhores resultados.

Exemplo da utilização em Analytics

O que é Analytics? Habilidade de utilizar dados, análises e raciocínio sistemático para conduzir um processo de tomada de decisão mais eficiente.

Como funciona na prática? Uma unidade hospitalar que cuida de bebês prematuros utiliza uma análise em tempo real com base em gravação de cada respiração e cada batimento cardíaco de todos os bebês em sua unidade. Em seguida, analisa os dados para identificar padrões. Com base na análise do sistema é possível prever infecções 24 horas antes de o bebê apresentar sintomas visíveis. Isso permite uma intervenção precoce e tratamento, o que é vital em bebês prematuros.

Exemplos da utilização da Estatística na área de Meio Ambiente

Como funciona na prática? Por meio dos conhecimentos é possível fazer uma avaliação do efeito do uso do ar-condicionado no consumo de combustível e nas emissões dos automóveis. Também são os estatísticos que auxiliam na criação de escore para divulgação de dados de qualidade da areia das praias.

Exemplos da utilização da Estatística na área de Qualidade

Como funciona na prática? Toda empresa, seja pequena ou grande, indústria ou serviço quer oferecer um trabalho de qualidade. Nesse ponto, o estatístico pode atuar em diversas etapas dos processos de uma empresa para garantir a qualidade e para isso é fundamental conhecer o negócio e suas metas. A empresa deve ter claro o que considera qualidade, qual o nível de qualidade esperado e a qual custo. O estatístico pode auxiliar nestas decisões por meio de indicadores de negócio, pesquisas de satisfação, entre outros. Além disso, pode medir a qualidade no processo de fabricação do produto, na qualidade dos fornecedores, na logística, nos sistemas de medição a partir de técnicas específicas para cada tipo de atividade. Dessa maneira, os gestores poderão fazer os ajustes necessários nos processos em questão. Alguns exemplos de metodologias são: CEP (controle estatístico do processo), MSA (Análise do sis-

tema de medição), modelos de otimizações na logística da empresa, os projetos de seis sigma, planejamentos e experimentos. Os testes de confiabilidade de produtos também têm papel importante na qualidade. Muitas empresas se certificam em normas técnicas de qualidade e o estatístico pode tornar-se auditor, aplicando metodologias e raciocínio estatístico para testar se a empresa está atuando conforme a norma estabelece.

Fonte: <http://www.conre3.org.br/portal/dia-do-estatistico-profissionais-falam-sobre-o-uso-da-estatistica-no-cotidiano/>. Acesso em: 5 jul. 2020. Acesso em: 5 jul. 2020.

Resumo da Unidade 1

Iniciamos o Tópico 1 com um histórico sobre a Estatística no mundo e no Brasil. A seguir apresentamos conceitos importantes para o bom entendimento da disciplina.

No Tópico 2 abordamos as formas de coletar, organizar e apresentar os dados pesquisados por meio de tabelas e gráficos. Conhecemos dois tipos importantes de tabelas estatísticas, com intervalos e sem intervalos de classes e suas representações gráficas nos histogramas, polígonos de frequência e Ogiva d Galton.

No Tópico 3 aprendemos a calcular as medidas estatísticas, tanto as medidas de posição média, mediana e moda, quanto as medidas de posição amplitude, desvio médio, variância, desvio-padrão e coeficiente de variação. Fizemos os mesmos cálculos para as três situações possíveis: dados agrupados, dados agrupados em tabelas sem intervalo de classes e dados agrupados em tabelas com intervalos de classes.

Também vimos os softwares estatísticos mais usados e algumas funções do Excel que auxiliam na Estatística descritiva.

Por fim, mostramos que tabelas e gráficos de dados podem ser encontrados com grande frequência no nosso dia a dia, relatórios anuais, artigos de jornais, revistas e estudos de pesquisa. Sendo assim, é muito importante entender como eles são organizados e, principalmente, como são interpretados.

Referências

ATIVOS DIGITAIS. **Envato Elements**. Disponível em: <https://elements.envato.com/pt-br/>. Acesso em: 28 jul. 2020.

BRASILEIRÃO de 1959 a 2019. **Esporte Interativo**. Disponível em: <https://www.esporte-interativo.com.br/futebolbrasileiro/GRAFICO-So-Paulo-lideraria-ranking-de-Brasileiro-de-todos-os-tempos-20200407-0040.html>. Acesso em: 28 jul. 2020.

BOLETIM da Receita Federal – Impactos da Covid-19. Disponível em: http://www.receita.economia.gov.br/dados/boletim-da-receita-federal_impactos-da-covid-19/boletim-da-receita-federal-impactos-da-covid-19z. Acesso em: 28 jul. 2020.

CONHEÇA o Brasil – População – Quantidade de homens e mulheres. **IBGE Educa – Jovens**. Disponível em: <https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/18320-quantidade-de-homens-e-mulheres.html>. Acesso em: 28 jul. 2020.

GLOBAL Coronavirus cases. **YouTube**. Disponível em: <https://youtu.be/Ovb2zc5M21g>. Acesso em: 30 jul. 2020.

INQUÉRITO domiciliar para monitorar a soroprevalência da infecção pelo vírus SARS-CoV-2 em adultos no município de São Paulo. **Ibobe**. Disponível em: <https://www.ibopeinteligencia.com/arquivos/Apresentação%20SoroEpi%20-%20resultados%20fase%202%20-%20FINAL.pdf>. Acesso em: 18 jul. 2020.

KOKOSKA, S. **Introdução à estatística**: uma abordagem por resolução de problemas. Rio de Janeiro: LTC, 2013.

LARSON, R.; FARBER, B. **Estatística aplicada**. São Paulo: Pearson, 2010.

MOORE, D. A **Estatística Básica e sua prática**. Rio de Janeiro: Ed. LTC, 2000.

NORMAS de apresentação tabular. Fundação Instituto Brasileiro de Geografia e Estatística. Rio de Janeiro: IBGE, 1979. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv82498.pdf>. Acesso em: 18 jul. 2020.

NÚMERO de matrículas na Educação Superior: 2008 – 2018. **Inep**. Disponível em: http://download.inep.gov.br/educacao_superior/censo_superior/documentos/2019/censo_da_educacao_superior_2018-notas_estatisticas.pdf. Acesso em: 28 jul. 2020.

NÚMEROS mais sorteados da Mega-Sena. **Só Matemática**. Disponível em: <https://www.somatematica.com.br/megasenaFrequentes.php>. Acesso em: 20 jul. 2020.

PAÍSES por números de homicídios. **YouTube**. Disponível em: <https://youtu.be/Hyuv-jXHoUi4>. Acesso em: 20 jul. 2020.

PERCENTUAL do número de funções docentes em exercício, por organização acadêmica, segundo o grau de formação. Brasil, 2017. Disponível em: http://download.inep.gov.br/educacao_superior/censo_superior/resumo_tecnico/resumo_tecnico_censo_da_educacao_superior_2017.pdf. Acesso em: 28 jul. 2020.

PORTAL do IBGE: Disponível em: <https://www.ibge.gov.br/>. Acesso em: 20 jul. 2020.

RANKING de desenvolvimento humano – 2015. **Folha de São Paulo**. Disponível em: <http://arte.folha.uol.com.br/graficos/vunXC/>. Acesso em: 28 jul. 2020.

SALÁRIOS: cargos e salários. **Salário**. Disponível em: <https://www.salario.com.br>. Acesso em: 20 jul. 2020.

SOARES, C. **Profissionais falam sobre o uso da Estatística no Cotidiano**. CONRE-3. 23/05/2020. Disponível em: <http://www.conre3.org.br/portal/dia-do-estatistico-profissionais-falam-sobre-o-uso-da-estatistica-no-cotidiano/>. Acesso em: 28 jul. 2020.

SPIEGEL, Murray R.; STEPHENS, Larry J. **Estatística**. 4. ed. Porto Alegre: Bookman, 2009

SWEENEY, Dennis J.; WILLIAMS, Thomas A.; ANDERSON, David R. **Estatística aplicada à administração e economia**. 3. ed. São Paulo: Cengage Learning, 2013.

UNIDADE 2

Probabilidade

INTRODUÇÃO

O mundo em que vivemos exige tomada de decisão, capacidade de lidar com o imprevisível e respostas para perguntas ainda desconhecidas. O estudo das teorias das probabilidades é de grande importância como suporte para tomadas de decisão, quando operamos com incertezas.

Se algo pode ter vários resultados e não sabemos ao certo qual ocorrerá, podemos usar a probabilidade para descrever a chance numérica de cada um dos resultados possíveis. Se soubermos que amanhã terá 95% de chance de chover, não vamos nos programar para passar o dia na praia. Seja na vida cotidiana ou nos negócios, ter meios de quantificar numericamente possibilidades é uma ferramenta extremamente importante para decidirmos os passos apropriados a serem tomados.

A teoria das probabilidades possui aplicação nos estudos relacionados às ciências exatas, biológicas e humanas, bem como em qualquer área que envolva o conceito de chance. Nesta unidade vamos estudar a definição e os conceitos básicos para o estudo das probabilidades.

O que veremos nos tópicos, a saber:

- **Tópico 1** – Os conceitos fundamentais, como experimentos aleatórios, experimentos determinísticos, espaço amostral e os tipos de eventos mais usuais. Também vamos recordar algumas técnicas de contagem da análise combinatória.
- **Tópico 2** – O conceito e os cálculos das probabilidades abordando três situações: a probabilidade teórica ou clássica para eventos equiprováveis, a probabilidade para eventos não equiprováveis e a visão frequencialista.
- **Tópico 3** – Os conceitos da probabilidade condicional e o notável Teorema de Bayes.

O estudo das probabilidades também será muito importante para entendermos os conceitos da Estatística Inferencial, que iremos estudar nas próximas unidades.



OBJETIVO

Nesta unidade você será capaz de:

- Utilizar o conceito de probabilidade e seus axiomas e teoremas na análise de situações de cunho prático.

Introdução à probabilidade

No dia a dia frequentemente tomamos decisões quando temos dúvidas.

Como será aceitação de um produto lançado no mercado?

Qual será a previsão do tempo na próxima semana?

Qual a chance de acertar os seis números da Mega-Sena?

Qual a possibilidade de obter bons lucros no mercado de ações?

Qual o risco que vou correr ao efetuar certo procedimento cirúrgico?

É conveniente, então, possuímos uma medida numérica. Sim, um número que revele a incerteza presente em cada um destes e de muitos outros questionamentos. Essa medida numérica é a **probabilidade**.



Ampliando o foco

A palavra probabilidade deriva do latim *probare*, que significa “testar”.

Iniciaremos o tópico apresentando os conceitos básicos que auxiliam nos cálculos das probabilidades.

Experimento

É todo o fenômeno que acontece ou toda ação que iremos realizar.

De acordo com a natureza do fenômeno que pretendemos estudar, podemos classificá-lo como aleatório ou determinístico.

1. Experimento aleatório

- Cada experimento poderá ser repetido, sob as mesmas condições, indefinidamente.
- O resultado particular de cada experimento aparecerá ao acaso, mas pode-se descrever todos os possíveis resultados.
- Quando o experimento for repetido muitas vezes, aparecerá regularidade nas respostas.



Exemplo

Sejam os experimentos:

a) Lançamento de um produto no mercado.

Podemos prever quais são os possíveis resultados: ser bem aceito, não ser bem aceito, ser indiferente.

Porém, não sabemos antecipadamente se esse produto será bem aceito ou não.

Se fizermos o lançamento desse produto no mercado inúmeras vezes, poderemos perceber regularidade na resposta.

b) Fazer um contato de vendas: possíveis resultado → Vender ou não vender.

c) Jogar uma partida de futebol: possíveis resultados → Ganhar, perder ou empatar.

d) Lançar um dado e observar a face de cima: possíveis resultados → 1, 2, 3, 4, 5, 6.

2. Experimento determinístico

- É exatamente o contrário do experimento aleatório.
- Nos experimentos determinísticos já conhecemos o resultado da experiência, pois ele será sempre o mesmo, desde que as mesmas condições e parâmetros iniciais sejam mantidos.
- Os experimentos determinísticos não são estudados pela Teoria das Probabilidades, já que não se trata de incertezas.



Exemplo

Soma dos ângulos internos de um triângulo → Sabemos que o resultado será sempre 180° . Se repetirmos a experiência de escolher vários triângulos, medirmos os três ângulos internos e efetuarmos a soma, o resultado será sempre igual a 180° .

Espaço amostral: (s ou Ω)

São todos os resultados possíveis de acontecer, quando se realiza um experimento. É representado pela letra **S** (*space*) ou pela letra grega **Ω** (ômega). Cada resultado do espaço amostral é considerado um ponto amostral.

O conceito de espaço amostral é muito importante, pois, na maioria dos casos, será a ferramenta principal para efetuarmos o cálculo das probabilidades.

Observe alguns exemplos na tabela a seguir.

Veja o experimento na coluna esquerda da tabela e especifique todos os elementos do espaço amostral. Depois confira a sua resposta na coluna à direita.

Experimento	Espaço amostral (s ou Ω) Possíveis resultados
Selecionar uma peça para ser inspecionada.	Ser defeituosa. Não ser defeituosa.
Verificar a previsão do tempo.	Céu ensolarado. Céu nublado. Ventania. Chovendo. Chuva com trovões. Granizo. Neve.
Realizar um procedimento cirúrgico.	Ser bem-sucedido. Não ser bem-sucedido.
Verificar as condições do trânsito.	Congestionado. Com lentidão. Com retenções. Boas condições de tráfego.

Jogar uma partida de futebol.	Ganhar. Perder. Empatar.
Fazer um contato de vendas.	Efetuar a venda. Não efetuar a venda.
Lançar um dado e observar a face superior.	Sair o número 1. Sair o número 2. Sair o número 3. Sair o número 4. Sair o número 5. Sair o número 6.

Evento

Todas as perguntas que formulamos a respeito do experimento são denominadas **“eventos”**, que são denotados por letras maiúsculas.

Vejamos alguns tipos de eventos. São eles:

- **Evento elementar ou evento simples:** é aquele formado por um único elemento do espaço amostral.
- **Evento composto:** é aquele formado por dois ou mais elementos do espaço amostral.
- **Evento certo:** é aquele que ocorre sempre, ou seja, os resultados favoráveis ao evento apresentam os mesmos elementos do espaço amostral.
- **Evento impossível:** é aquele que nunca ocorre, ou seja, os resultados favoráveis ao evento serão um conjunto vazio.
- **Evento soma ou união:** é aquele em que pelo menos um dos eventos ocorre, “Um ou outro”.
- **Evento produto ou interseção:** é aquele em que os eventos ocorrem simultaneamente, “Um e outro”.
- **Eventos mutuamente exclusivos, ou mutuamente excludentes:** são eventos em que “Um ou outro” não poderá acontecer.

- **Evento complementar:** são eventos mutuamente exclusivos, que unidos têm os mesmos elementos do espaço amostral. É denotado com uma barra acima do nome do evento.
- **Eventos equiprováveis:** são eventos em que todos os elementos do espaço amostral têm a mesma probabilidade de ocorrência.
- **Eventos não equiprováveis:** são eventos em que os elementos do espaço amostral não têm a mesma probabilidade de ocorrência.
- **Eventos independentes:** dois eventos são independentes quando a ocorrência de um evento não influencia a ocorrência do outro.
- **Eventos condicionados:** são aqueles em que o acontecimento de um está condicionado ao acontecimento de outro, ou seja, acontece um se o outro já aconteceu.

Agora que já conhecemos os conceitos básicos da probabilidade, apresentaremos alguns exemplos ligando esses quatro conceitos na tabela a seguir.

Experimento	Espaço Amostral	Evento	Tipo de Evento
Jogar um dado	$\Omega = \{1,2,3,4,5,6\}$	Quais são as chances, ao lançar um dado, de observarmos na face superior um número maior do que 5? A = face superior maior do que 5 (Escolhemos no espaço amostral os elementos que são favoráveis ao evento). $A = \{6\}$	Evento Elementar ou simples
	$\Omega = \{1,2,3,4,5,6\}$	Quais são as chances, ao lançar um dado, de observarmos na face superior um número primo? B = face superior ser um número primo $B = \{2,3,5\}$	Evento composto

Jogar um dado	$\Omega = \{1,2,3,4,5,6\}$	<p>Quais são as chances, ao lançar um dado, de observarmos na face superior um número menor ou igual a 6?</p> <p>C=face superior ser um número menor ou igual a 6</p> <p>$C=\{1,2,3,4,5,6\}$</p>	Evento certo
	$\Omega = \{1,2,3,4,5,6\}$	<p>Quais são as chances, ao lançar um dado, de observarmos na face superior um número maior do que 6?</p> <p>D=face superior ser um número maior do que 6</p> <p>$D=\{ \}$</p>	Evento impossível
	$\Omega = \{1,2,3,4,5,6\}$ $\Omega = \{1,2,3,4,5,6\}$	<p><i>(a partir daqui vamos omitir a pergunta e já escrever o evento)</i></p> <p>E=observar na face superior um número maior do que 4</p> <p>$E=\{5,6\}$</p> <p>ou</p> <p>F=observar na face superior um número menor do que 3</p> <p>$F=\{1,2\}$</p> <p>$G=E \cup F = \{1,2,5,6\}$</p>	Evento união

Jogar um dado	$\Omega = \{1,2,3,4,5,6\}$ $\Omega = \{1,2,3,4,5,6\}$	<p>H=observar na face superior um número maior do que 2</p> <p>$H=\{3,4,5,6\}$</p> <p>e</p> <p>I=observar na face superior um número menor do que 5</p> <p>$I=\{1,2,3,4\}$</p> <p>$J=H \cap I = \{3,4\}$</p>	Evento interseção
	$\Omega = \{1,2,3,4,5,6\}$ $\Omega = \{1,2,3,4,5,6\}$	<p>K= observar na face superior um número menor do que 3</p> <p>$K=\{1,2\}$</p> <p>L= observar na face superior uma face maior do que 3</p> <p>$L=\{4,5,6\}$</p> <p>$K \cap L = \{ \}$</p> <p>$K \cup L = \{1,2,4,5,6\}$</p>	<p>Evento mutuamente exclusivos</p> <p>(a interseção é nula, mas juntos não formam o espaço amostral).</p>
	$\Omega = \{1,2,3,4,5,6\}$ $\Omega = \{1,2,3,4,5,6\}$	<p>M= observar na face superior um número par</p> <p>$M=\{2,4,6\}$</p> <p>\overline{M}= observar na face superior uma face ímpar</p> <p>$\overline{M} = \{1,3,5\}$</p> <p>$M \cap \overline{M} = \{ \}$</p> <p>$M \cup \overline{M} = \{1,2,3,4,5,6\} = \Omega$</p>	<p>Eventos complementares</p> <p>(a interseção é nula, e juntos formam o espaço amostral).</p>

Para apresentarmos exemplos de eventos equiprováveis, não equiprováveis, independentes e condicionados precisamos primeiro estudar como calculamos as probabilidades, o que veremos a seguir.

Regras de contagem, combinações e permutações

Para calcularmos as probabilidades, muitas vezes será necessário identificar e contar o número de elementos do espaço amostral.

Vamos apresentar três modos de contagem muito usados, que são: experimentos em múltiplas etapas, arranjos, combinação e permutação.

1. Experimentos em múltiplas etapas

Se um experimento pode ser descrito como uma sequência de dois estágios sucessivos e independentes, em que o primeiro estágio pode ocorrer de “m” modos distintos, em seguida o segundo estágio pode acontecer de “n” modos distintos. Nessas condições, dizemos que “o número de maneiras distintas de ocorrer esse acontecimento é igual ao produto $m \cdot n$ ”.

Veja um exemplo:

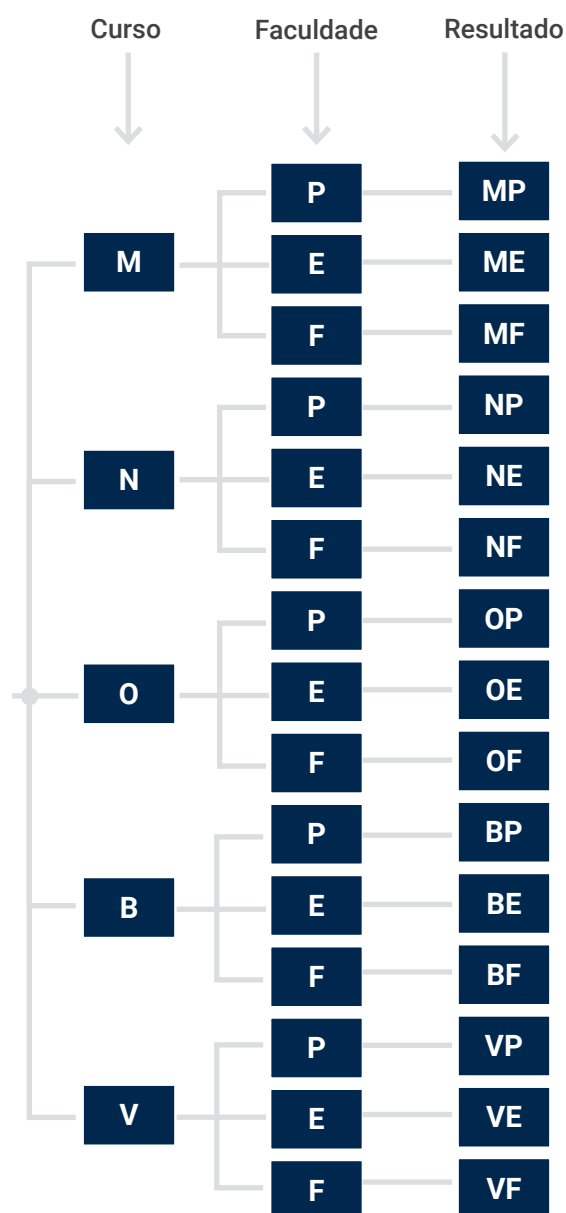
Um estudante recebeu o resultado do seu desempenho no Enem e deve escolher um curso e o tipo de faculdade que deseja cursar. Sabe-se que ele tem preferência por cinco cursos: Medicina (M), Nutrição (N), Odontologia (O), Biomedicina (B) e Veterinária (V). Cada curso pode ser feito em três tipos de faculdades: particular (P), estadual (E) ou federal (F). Qual é o número de opções que o estudante poderá escolher para fazer sua matrícula?

Resolução:

De acordo com o princípio fundamental da contagem, o número total de opções que o estudante pode fazer é $5 \cdot 3 = 15$.

Se quisermos ilustrar as 15 opções podemos lançar mão de um recurso muito útil, que se chama “árvore de possibilidades”.

Figura: Árvore das Possibilidades.



Fonte: Elaborada pela autora (2020).

Fazendo a contagem das respostas anteriores temos também as 15 possibilidades.

Podemos generalizar o princípio fundamental da contagem. Quando um acontecimento é composto por “k” estágios sucessivos e independentes, com, respectivamente, n_1 , n_2 , n_3 , n_4 , ..., n_k modos distintos de ocorrência, o número total de maneiras distintas de ocorrer esse acontecimento é $n_1 \cdot n_2 \cdot n_3 \cdot n_4 \cdot \dots \cdot n_k$.

Veja mais alguns exemplos:

Exemplo 1

Um casal planeja ter três filhos.

- De quantas maneiras diferentes a família poderá ser formada, considerando-se que nascerá um filho em cada gestação.
- Quais são essas possibilidades?

Resolução:

a) Para responder de quantas maneiras diferentes serão, basta usar o princípio fundamental da contagem:

1ª gestação: 2 opções – O filho ser do sexo masculino ou do sexo feminino.

2ª gestação: 2 opções – O filho ser do sexo masculino ou do sexo feminino.

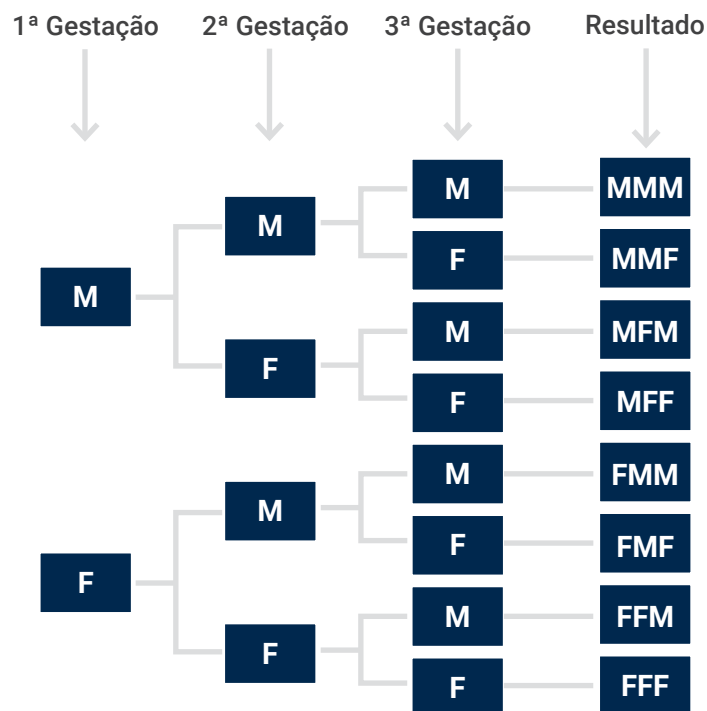
3ª gestação: 2 opções – O filho ser do sexo masculino ou do sexo feminino.

Logo: $2 \cdot 2 \cdot 2 = 8$ maneiras.

b) Para responder quais são as possibilidades vamos desenhar a árvore de possibilidades:

Vamos simbolizar: M → sexo masculino e F → sexo feminino

Figura: Árvore das Possibilidades



Fonte: Elaborada pela autora (2020).

Exemplo 2

Quantos números diferentes e de três algarismos distintos existem no sistema decimal de numeração?

Resolução:

Classes	C	D	U
	1, 2, 3, 4, 5, 6, 7, 8, 9 Lembrando que o algarismo das centenas tem que ser diferente de zero para que o número tenha três dígitos.	Supondo que escolhemos o algarismo "1" para as centenas, ele não poderá ser usado aqui, já que os algarismos são distintos. Logo, na casa das dezenas podem ser os dígitos: 0, 2, 3, 4, 5, 6, 7, 8, 9	Supondo que escolhemos o algarismo "2" para as dezenas, ele não poderá ser usado aqui, já que os algarismos são distintos. Logo, na casa das unidades podem ser os dígitos: 0, 3, 4, 5, 6, 7, 8, 9
Total de possibilidades	9	9	8

Pelo princípio fundamental da contagem temos: $9 \cdot 9 \cdot 8 = 648$ possibilidades.

Exemplo 3

Em uma indústria, uma peça deve passar por três estações para receber sua classificação.

Estação 1: classifica em ruim, bom ou excelente.

Estação 2: classifica em atrasado ou no prazo.

Estação 3: classifica em esteira 1, esteira 2, esteira 3 ou esteira 4.

Calcule o número de classificações completas que a peça pode receber.

Resolução:

Pelo princípio fundamental da contagem temos: $3 \cdot 2 \cdot 4 = 24$

Cálculos que envolvam fatorial

O fatorial de um número natural “n” é representado por $n!$ (lê-se n fatorial) e definido por:

1) $n! = n \cdot (n - 1) \cdot (n - 2) \cdot (n - 3) \dots \cdot 2 \cdot 1$, para $n \geq 2$

2) $1! = 1$

3) $0! = 1$

Para ilustrar, veja um exemplo.



Exemplo

a) $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$

b) $\frac{8!}{5! \cdot 3!} = \frac{8 \cdot 7 \cdot 6 \cdot 5!}{5! \cdot 3 \cdot 2 \cdot 1}$ como estamos multiplicando os fatores no

numerador e no denominador, podemos parar de escrever a qualquer momento, colocar o símbolo de fatorial (!) e depois efetuar a simplificação, não havendo necessidade de desenvolver até o 1.

$$\frac{8!}{5! \cdot 3!} = \frac{8 \cdot 7 \cdot \cancel{6} \cdot \cancel{5!}}{\cancel{5!} \cdot \cancel{3} \cdot \cancel{2} \cdot 1} = 56$$

2. Arranjos simples

Usados em problemas de contagem em que os agrupamentos são considerados distintos quanto a ordem e natureza de seus elementos. Por exemplo: vamos considerar o número formado pelos algarismos 1 e 2. Seja o número 12. Se invertermos a ordem o número será 21, que é diferente de 12, ou seja, formamos agrupamentos diferentes quando invertemos a ordem dos elementos.

Fórmula a ser usada:

$$A_{n,k} = \frac{n!}{(n - k)!}$$

Veja alguns exemplos a seguir:



Exemplo 1

Cinco cavalos disputam um páreo. Quantos são os possíveis resultados para as três primeiras colocações.

Resolução: Vamos formar um agrupamento com três cavalos escolhidos em um total de cinco cavalos.

Como a ordem importa nesse agrupamento, pois, se invertemos a posição dos cavalos mudamos o sentido do agrupamento, vamos usar a fórmula dos arranjos.

$$A_{n,k} = \frac{n!}{(n-k)!}$$

$$A_{5,3} = \frac{5!}{(5-3)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1} = 60$$



Exemplo 2

A senha de um cofre é composta por uma sequência de três dígitos diferentes. Calcule o número máximo de tentativas para que a senha seja descoberta.

Resolução: Vamos formar um agrupamento de três dígitos distintos, escolhidos nos 10 dígitos disponíveis no nosso sistema de numeração.

Como a ordem importa nesse agrupamento, já que ao invertemos os dígitos, as senhas ficam diferentes, vamos usar a fórmula do arranjo simples.

$$A_{n,k} = \frac{n!}{(n-k)!}$$

$$A_{10,3} = \frac{10!}{(10-3)!} = \frac{10 \cdot 9 \cdot 8 \cdot 7!}{7!} = 720$$



Importante

1 - Quando o tamanho do agrupamento é igual ao número de elementos, em vez de arranjo chamamos de Permutação.

$$A_{n,k} = \frac{n!}{(n-k)!}$$

2 - Qualquer problema que envolva permutações ou arranjo simples pode ser resolvido diretamente pelo princípio fundamental da contagem.

3. Combinação simples

Usadas em problemas de contagem em que os agrupamentos são considerados distintos quanto a ordem e natureza. A ordem dos elementos não altera o agrupamento. Vamos selecionar seis números para jogar na Mega-Sena. Se alterarmos a ordem que vamos marcar no cartão de jogos, a aposta será a mesma, ou seja, formamos agrupamentos iguais quando invertemos a ordem dos elementos.

Fórmula a ser usada:

$$C_{n,k} = \frac{n!}{(n-k)!}$$



Exemplo

Quantos cartões diferentes podemos marcar na Mega-Sena?

Sabemos que a ordem que você marca no cartão dos números escolhidos não importa nesse agrupamento. Logo, dos 60 números disponíveis, vamos escolher seis deles e calcular a combinação, já que a ordem não importará.

$$C_{n,k} = \frac{n!}{k! \cdot (n - k)!}$$

$$C_{60,6} = \frac{60!}{6! \cdot (60 - 6)!}$$

$$C_{60,6} = \frac{60!}{6! \cdot 54!} = \frac{60.59.58.57.56.55.54}{6.5.4.3.2.1.54!}$$

$$C_{60,6} = \frac{60.59.58.57.56.55}{6.5.4.3.2.1} = 50.063.860$$



Ampliando o foco

Para colocar em prática o conteúdo visto nesta unidade visite a Biblioteca Virtual, escolha um livro de Estatística, selecione e faça alguns exercícios do livro para reforçar o conhecimento.

Confira as respostas que estão no final do livro e, se tiver alguma dúvida, fale com o seu tutor pelo Fórum de Dúvidas – Fale com o Tutor.

Conceitos de probabilidade

Neste tópico vamos apresentar conceitos importantes para efetuarmos o cálculo das probabilidades.

Seja um experimento aleatório e " Ω " um espaço amostral associado a esse experimento. A cada evento " A " associamos um número real, representado por $P(A)$ e denominado probabilidade de A .

Podemos efetuar o cálculo das probabilidades considerando três cenários diferentes. São eles:

- **1º cenário:** probabilidade teórica ou clássica com eventos equiprováveis.
- **2º cenário:** probabilidade teórica com eventos não equiprováveis.
- **3º cenário:** probabilidade empírica ou estatística, usando frequência relativa.

Probabilidade teórica ou clássica com eventos equiprováveis

Na definição clássica de probabilidade trabalhamos apenas com eventos equiprováveis, que, como vimos, são os eventos em que todos os elementos do espaço amostral têm a mesma chance de ocorrer.

Para calcular a probabilidade basta efetuar a divisão entre o número de elementos dos casos **favoráveis** ao evento " A " pelo número de elementos **possíveis** representados no espaço amostral " Ω ".

$$P(A) = \frac{\text{números de casos favoráveis}}{\text{números de casos possíveis}} = \frac{n(A)}{n(\Omega)}$$

Podemos apresentar o resultado das probabilidades de três maneiras, a saber:

1. Como uma fração.
2. Como um número racional, se efetuarmos a divisão entre o numerador e o denominador da fração.
3. Como uma porcentagem, se multiplicarmos por 100 o resultado da divisão.

Exemplo 1

Em um lote de 12 peças, quatro são defeituosas. Sendo retirada uma peça para inspeção, calcule a probabilidade de essa peça ser defeituosa.

Antes de iniciar o cálculo precisamos seguir as seguintes etapas:	Respostas
1 - Qual é o experimento?	Avaliar defeitos em peças produzidas.
2 - Qual é o espaço amostral?	$\Omega = \{ D, D, D, D, \bar{D}, \bar{D}, \bar{D}, \bar{D}, \bar{D}, \bar{D}, \bar{D}, \bar{D} \}$ onde: <i>D</i> = peça defeituosa <i>D̄</i> = peça não defeituosa 12 peças, sendo 4 defeituosas e 8 perfeitas 12 elementos, então: $n(\Omega) = 12$
3 - Qual é o evento?	A = retirar uma peça defeituosa
4 - Retire do espaço amostral os elementos favoráveis ao evento.	<i>D, D, D, D</i> → 4 peças defeituosas; 4 elementos, então: $n(A) = 4$
5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência? (Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)	Sim, pois não há nenhuma informação de que alguma peça tenha maior chance de ser selecionada do que outra.
6 - Fazer o cálculo usando o método apropriado.	$P(A) = \frac{\text{números de casos favoráveis}}{\text{números de casos possíveis}} = \frac{n(A)}{n(\Omega)} = \frac{4}{12} = 0,333 = 33\%$

Exemplo 2

No lançamento de um dado “honesto”, qual a probabilidade de obter um número par na face superior?

Etapas	Respostas
1 - Qual é o experimento?	Lançamento de um dado.
2 - Qual é o espaço amostral?	As faces {1, 2, 3, 4, 5, 6}, portanto são seis elementos, então: $n(\Omega) = 6$
3 - Qual é o evento?	A = observar na face superior um número par.
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	A = {2, 4, 6}, portanto são três elementos, então: $n(A) = 3$
5 - Todos os elementos do espaço amostral têm a mesma chance de ocorrência? (Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)	Sim, pois temos a informação de que o dado é honesto, ou seja, todas as faces têm a mesma probabilidade de sair.
6 - Fazer o cálculo usando o método apropriado.	$P(A) = \frac{\text{números de casos favoráveis}}{\text{números de casos possíveis}} = \frac{n(A)}{n(\Omega)} = \frac{3}{6} = 0,5 = 50\%$

Exemplo 3

Dois dados honestos são lançados e a face superior é observada. Determine a probabilidade de a soma das duas faces superiores ser igual a oito.

Etapas	Respostas
1 - Qual é o experimento?	Lançar dois dados.
2 - Qual é o espaço amostral?	$\Omega = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6),$ $(2,1), (2,2), (2,3), (2,4), (2,5), (2,6),$ $(3,1), (3,2), (3,3), (3,4), (3,5), (3,6),$ $(4,1), (4,2), (4,3), (4,4), (4,5), (4,6),$ $(5,1), (5,2), (5,3), (5,4), (5,5), (5,6),$ $(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$ $n(\Omega) = 36$
3 - Qual é o evento?	A = soma dos números das duas faces superiores ser igual a 8.
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	A = {(2,6), (3,5), (4,4), (5,3), (6,2)}, portanto são cinco elementos, então: $n(A) = 5$
<p>5 - Todos os elementos do espaço amostral têm a mesma chance de ocorrência?</p> <p>(Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)</p>	Sim, pois os dados são honestos.
6 - Fazer o cálculo usando o método apropriado.	$P(A) = \frac{\text{números de casos favoráveis}}{\text{números de casos possíveis}} = \frac{n(A)}{n(\Omega)} = \frac{5}{36} = 0,139 = 13,9\%$

Exemplo 4

Determine a probabilidade de pelo menos uma face “cara” aparecer, na face superior, no lançamento de duas moedas “não viciadas”.

Etapas	Respostas
1 - Qual é o experimento?	Lançamento de duas moedas .
2 - Qual é o espaço amostral?	$\Omega = \{(cara, cara), (cara, coroa), (coroa, cara), (coroa, coroa)\}$ $n(\Omega) = 4$
3 - Qual é o evento?	A = observar pelo menos uma cara na face superior
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	$A = \{(cara, cara), (cara, coroa), (coroa, cara)\}$, portanto são 3 elementos, então: $n(A) = 3$
5 - Todos os elementos do espaço amostral têm a mesma chance de ocorrência? (Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)	Sim, pois temos a informação que as moedas não são “viciadas”, ou seja, as duas faces têm a mesma probabilidade de sair.
6 - Fazer o cálculo usando o método apropriado.	$P(A) = \frac{\text{números de casos favoráveis}}{\text{números de casos possíveis}} = \frac{n(A)}{n(\Omega)} = \frac{3}{4} = 0,75 = 75\%$

Exemplo 5

A equipe de professores do Departamento de Estatística é composta de 12 pessoas, sendo 9 mulheres e 3 homens. Dessas pessoas, vamos selecionar aleatoriamente duas para participar do Congresso de Estatística. Qual a probabilidade de a comissão ser formada por duas mulheres.

Etapas	Respostas
1 - Qual é o experimento?	Selecionar dois professores do sexo feminino para participar do Congresso de Estatística.
2 - Qual é o espaço amostral?	<p>Para calcular o número de elementos do espaço amostral, devemos considerar um grupo de 12 pessoas, das quais serão retirados dois elementos, sem importar a ordem, o que condiz com a combinação de 12 elementos, tomados dois a dois.</p> $C_{n,k} = \frac{n!}{k! \cdot (n - k)!}$ $C_{12,2} = \frac{12!}{2! \cdot 10!}$ $C_{12,2} = \frac{12 \cdot 11 \cdot 10!}{2 \cdot 1 \cdot 10!} = 66$
3 - Qual é o evento?	A = escolher duas mulheres
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	<p>Para calcular o número de elementos favoráveis, vamos retirar das nove mulheres possíveis, um agrupamento de dois, logo, como não importa a ordem, vamos calcular a combinação de nove tomados dois a dois.</p> $C_{9,2} = \frac{9!}{2! \cdot 7!} = \frac{9 \cdot 8 \cdot 7!}{2 \cdot 1 \cdot 7!} = 36$

<p>5 - Todos os elementos do espaço amostral têm a mesma chance de ocorrência?</p> <p>(Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)</p>	<p>Sim, pois todas as mulheres têm a mesma chance de ser selecionadas. Não há nenhuma informação sobre uma pessoa ser predileta para a escolha. → probabilidade clássica.</p>
<p>6 - Fazer o cálculo usando o método apropriado.</p>	$P(A) = \frac{\text{números de casos favoráveis}}{\text{números de casos possíveis}} = \frac{n(A)}{n(\Omega)} = \frac{36}{66} = 0,5455 = 54,55\%$

Exemplo 6

Calcule a probabilidade de um jogador acertar os seis números da Mega-Sena marcando um cartão com seis números.

Etapas	Respostas
1 - Qual é o experimento?	Jogar um cartão com 6 números na Mega-Sena .
2 - Qual é o espaço amostral?	<p>Para calcular o número de elementos do espaço amostral, devemos considerar um grupo de números, dos quais serão escolhidos seis, sem importar a ordem em que serão marcados no cartão, o que condiz com a combinação de 60 elementos, tomados seis a seis.</p> $C_{n,k} = \frac{n!}{k! \cdot (n - k)!}$ $C_{60,6} = \frac{60!}{6! \cdot 54!}$

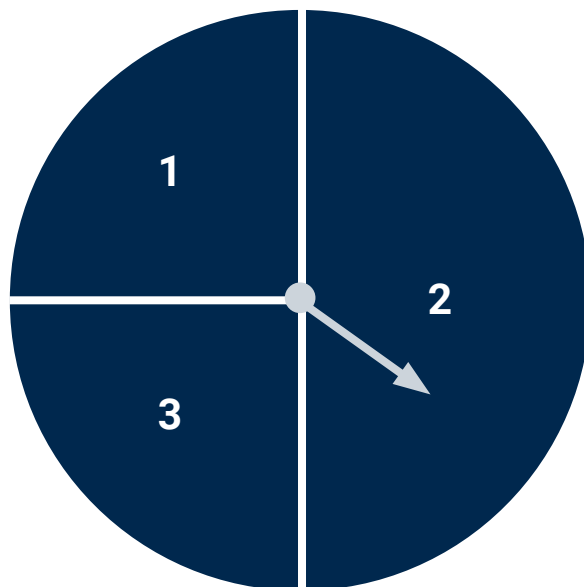
	$C_{12,2} = \frac{60.59.58.57.56.55.54!}{6.5.4.3.2.1.54!} = 50.063.860$ <p>Sim, pode acreditar! Existem 50.063.860 cartões diferentes que podem ser marcados escolhendo seis números de 00 a 60.</p>
3 - Qual é o evento?	A = marcar os seis números que serão sorteados.
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	P1 cartão.
<p>5 - Todos os elementos do espaço amostral têm a mesma chance de ocorrência?</p> <p>(Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)</p>	<p>Sim, pois todos os números têm a mesma chance de serem sorteados. → probabilidade clássica.</p>
6 - Fazer o cálculo usando o método apropriado.	$P(A) = \frac{\text{números de casos favoráveis}}{\text{números de casos possíveis}} = \frac{n(A)}{n(\Omega)} = \frac{1}{50.063.860} = 0,00000002 = 0,00$

Probabilidade teórica com eventos não equiprováveis

São usadas quando pelo menos um elemento do espaço amostral tiver diferentes probabilidades de ocorrência dos demais.

Veja um exemplo:

Considere a roleta indicada na figura.



Vamos supor que vamos dar um impulso na seta e desejamos conhecer a probabilidade de ela vir parar no setor circular com o número 2.

Podemos visualizar que a área do setor circular com os números 2 é diferente da área dos setores 1 e 3. Logo, essa probabilidade deverá ser calculada para eventos com espaços amostrais não equiprováveis.

Seja Ω o espaço amostral, então: $\Omega = \{1, 2, 3\}$. Este espaço amostral não será adequado para resolver esse tipo de problema, já que os elementos não são equiprováveis.

Uma boa estratégia para calcular a probabilidade seria redefinir um novo espaço amostral adequado, dessa vez com eventos equiprováveis.

Seja $\Omega_1 = \{1, 2, 2, 3\}$ o novo espaço amostral, desta vez com todos os elementos equiprováveis e $n(\Omega_1) = 4$ elementos

Agora sim, poderemos utilizar o conceito clássico da probabilidade para eventos equiprováveis. Então:

Seja o evento A = seta parar no número 2 $\rightarrow n(A)=2$

$$P(A) = \frac{n(A)}{n(\Omega_1)} = \frac{2}{4} = 0,5 = 50\%$$

Probabilidade empírica, ou estatística, usando frequência relativa

A frequência relativa de uma tabela de distribuição de frequências, que apresentamos na Unidade 1, pode ser usada para determinarmos a probabilidade de um evento.

Se um experimento for repetido um número muito grande de vezes, percebemos uma tendência nos resultados. A proporção de ocorrência dos resultados tenderá a probabilidade teórica. Isso é chamado de Lei dos Grandes Números.

Logo, para calcular a probabilidade de o evento “A” ocorrer temos:

$$P(A) = \frac{\text{número de vezes que o evento ocorreu}}{\text{número de repetições do experimento}}$$

Essa fórmula é exatamente o que fizemos em estatística descritiva ao organizar os dados em uma tabela de distribuição de frequências. A coluna da frequência relativa indica a probabilidade de ocorrência de cada resultado observado.

Vejamos um exemplo.

Seja um experimento de lançar um dado “honesto” várias vezes e observar a face superior. A tabela a seguir apresenta os resultados do número de vezes em que o número 6 foi observado na face superior.

Lançamento do dado		
Número de repetições	Quantidade de vezes em que apareceu o número 6 na face superior	Cálculo da frequência relativa
1	0	$\frac{0}{1} = 0\%$
2	0	$\frac{0}{2} = 0\%$
3	0	$\frac{0}{3} = 0\%$
4	1	$\frac{1}{4} = 25\%$
5	1	$\frac{1}{5} = 20\%$
6	2	$\frac{2}{6} = 33,3\%$
20	3	$\frac{3}{20} = 15\%$
100	16	$\frac{16}{100} = 16\%$
1000	165	$\frac{165}{1000} = 16,5\%$
10000	1667	$\frac{1677}{1000} = 16,67\%$

Seja o evento A=observar a face superior igual a 6 no lançamento de um dado.

Pela probabilidade clássica temos:

$$P(A) = \frac{1}{6} = 0,1667 = 16,67\%$$

Pela probabilidade empírica, observamos que, à medida que o número de repetições aumenta, a frequência relativa aproxima-se da probabilidade clássica.

Quando $n=10000 \rightarrow fr = 16,77\%$



Importante

Se o número de repetições for pequeno, haverá grandes oscilações da frequência relativa. À medida que o número de repetições aumenta consideravelmente, as oscilações tendem a ser menores e tendem a estabilizar em um valor constante, que pode ser avaliado como probabilidade. A dificuldade existente nessa definição é que o número limite real pode não existir na realidade. Sendo assim, a definição obedece à teoria matemática do limite.

Relações básicas de probabilidades

Veremos a seguir algumas relações básicas de probabilidades. Aqui, vamos estudar as probabilidades complementares e a probabilidade na união de eventos.

Probabilidade complementar

Dois eventos, A e B, são complementares se sua intersecção é vazia e sua união é o espaço amostral Ω . Isto é:

$$A \cap B = \emptyset$$

e

$$A \cup B = \Omega.$$

O complementar de um evento A é representado por \bar{A} ou A^c

• Cálculo das probabilidades complementares

Se \bar{A} é o complementar de A, então a soma das probabilidades de um evento com seu evento complementar é sempre igual a 100%, ou seja, igual a 1:

$$P(A) + P(\bar{A}) = 1 \quad \therefore \quad P(\bar{A}) = 1 - P(A) \quad \text{ou} \quad P(A) = 1 - P(\bar{A})$$

Exemplo 1

Qual a probabilidade de não sair a face 4 no lançamento de um dado?

Etapas:	Respostas
1 - Qual é o experimento?	Lançamento de um dado.
2 - Qual é o espaço amostral?	As faces $\{1, 2, 3, 4, 5, 6\}$, portanto são seis elementos. Então: $n(\Omega) = 6$
3 - Qual é o evento?	Qual a probabilidade de não sair a face 4 no lançamento de um dado? Evento A = sair a face 4 Evento \bar{A} = não sair a face 4 (Lembrando que a barra acima do nome do evento significa a negação.)
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	$A = \{4\} \rightarrow n(A) = 1$
5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência? (Se a resposta for "Sim", então vamos calcular a fórmula da probabilidade clássica. Se for "Não", iremos usar o cálculo empírico ou frequencial.)	Sim \rightarrow probabilidade clássica.
6 - Fazer o cálculo usando o método apropriado.	Usando o conceito de eventos complementares: $P(\bar{A}) = 1 - P(A) = 1 - \frac{1}{6} = \frac{5}{6} = 0,8333 = 83,33\%$ Logo, existe uma probabilidade de aproximadamente 83% de não sair a face 4 no lançamento de um dado.

Exemplo 2

A probabilidade de uma dona de casa escolher uma determinada marca de café é de 65%. Qual a probabilidade que em um determinado dia ela escolha outra marca?

Etapas:	Respostas
1 - Qual é o experimento?	Dona de casa escolher uma marca de café.
2 - Qual é o espaço amostral?	Vamos supor 100 pessoas.
3 - Qual é o evento?	Evento A= escolher determinada marca de café. Evento \bar{A} = não escolher essa determinada marca, ou seja, escolher outra marca (lembrando que a barra acima do nome do evento, significa a negação)
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	$n(A) = 65$
5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência? (Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)	Sim → probabilidade clássica.
6 - Fazer o cálculo usando o método apropriado.	Usando o conceito de eventos complementares: $P(\bar{A}) = 1 - P(A) = 1 - 0,65 = 0,35 = 35\%$

A propriedade complementar é muito usada quando precisamos calcular a probabilidade de “pelo menos” alguma ocorrência.

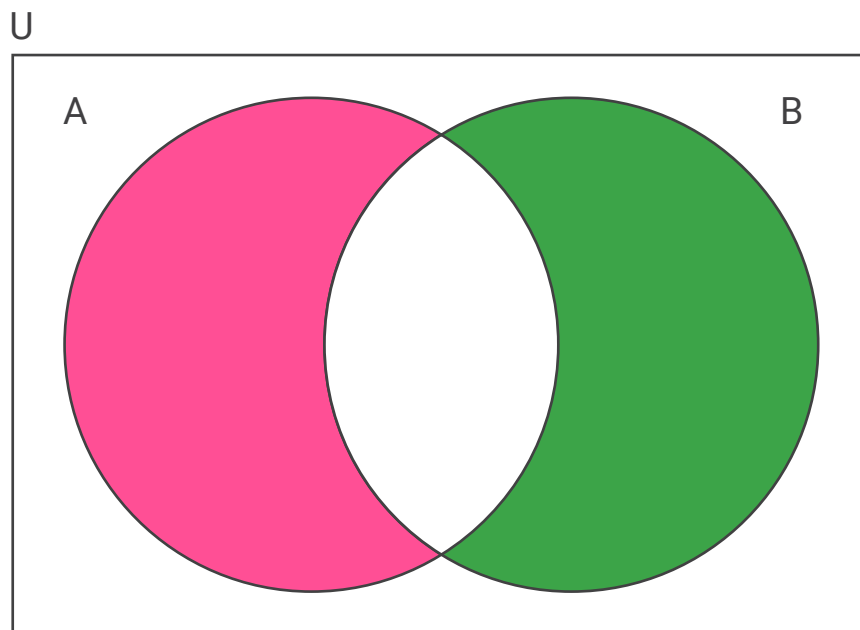
Exemplo 3

Em um lote de 12 peças, quatro são defeituosas e oito são boas. A probabilidade de retirarmos duas peças boas em uma inspeção é igual a 42%. Qual a probabilidade de pelo menos uma apresentar defeito quando retiramos duas peças para inspeção?

Etapas:	Respostas
1 - Qual é o experimento?	Retirar duas peças para inspeção
2 - Qual é o espaço amostral?	Não será preciso definir, pois vamos usar o conceito de probabilidade complementar.
3 - Qual é o evento?	A= retirar pelo menos uma peça defeituosa \bar{A} = não retirar alguma peça defeituosa, ou seja, retirar as duas peças boas. Se não retiramos duas peças boas, então pelo menos uma será defeituosa.
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	Não será preciso, pois vamos usar o conceito de probabilidade complementar.
5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência? (Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)	Sim
6 - Fazer o cálculo usando o método apropriado.	Usando o conceito de eventos complementares: $P(\bar{A}) = 1 - P(A) = 1 - 0,42 = 0,58 = 58\%$

União de eventos (μ)

Sejam dois eventos A e B:



Na união, consideramos os elementos da parte rosa, da parte verde e da parte branca.

União $\rightarrow U \rightarrow$ palavra usada OU \rightarrow operação (+)

$P(A) \cup (B) = P(A) + P(B) \rightarrow$ eventos mutuamente exclusivos

$P(A) \cup (B) = P(A) + P(B) - P(A \cap B) \rightarrow$ eventos não exclusivos

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \rightarrow$ 3 eventos não exclusivos

Exemplo 1

Retirando uma carta de um baralho comum de 52 cartas, qual é a probabilidade de ocorrer uma dama ou uma carta de copas?

Etapas:	Respostas
1 - Qual é o experimento?	Retirar uma carta de um baralho de 52 cartas.
2 - Qual é o espaço amostral?	$W = \{A\clubsuit, 2\clubsuit, 3\clubsuit, 4\clubsuit, 5\clubsuit, 6\clubsuit, 7\clubsuit, 8\clubsuit, 9\clubsuit, 10\clubsuit, J\clubsuit, Q\clubsuit, K\clubsuit, A\spadesuit, 2\spadesuit, 3\spadesuit, 4\spadesuit, 5\spadesuit, 6\spadesuit, 7\spadesuit, 8\spadesuit, 9\spadesuit, 10\spadesuit, J\spadesuit, Q\spadesuit, K\spadesuit, A\diamondsuit, 2\diamondsuit, 3\diamondsuit, 4\diamondsuit, 5\diamondsuit, 6\diamondsuit, 7\diamondsuit, 8\diamondsuit, 9\diamondsuit, 10\diamondsuit, J\diamondsuit, Q\diamondsuit, K\diamondsuit, A\heartsuit, 2\heartsuit, 3\heartsuit, 4\heartsuit, 5\heartsuit, 6\heartsuit, 7\heartsuit, 8\heartsuit, 9\heartsuit, 10\heartsuit, J\heartsuit, Q\heartsuit, K\heartsuit\}$; $n(W) = 52$
3 - Qual é o evento?	<p>Evento A: retirar uma dama.</p> <p>Evento B: retirar uma carta de copas.</p>
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	<p>$A = \{Q\clubsuit, Q\spadesuit, Q\diamondsuit, Q\heartsuit\}$; $n(A) = 4$</p> <p>$B = \{A\heartsuit, 2\heartsuit, 3\heartsuit, 4\heartsuit, 5\heartsuit, 6\heartsuit, 7\heartsuit, 8\heartsuit, 9\heartsuit, 10\heartsuit, J\heartsuit, Q\heartsuit, K\heartsuit\}$; $n(B) = 13$</p> <p>Elementos comuns em A e B $\rightarrow A \cap B = \{Q\heartsuit\}$; $n(A \cap B) = 1$</p>
<p>5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência?</p> <p>(Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)</p>	Sim

6 - Fazer o cálculo usando o método apropriado.

Calculando a probabilidade de A **ou** B temos:

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

$$P(A \cup B) = \frac{n(A)}{n(\Omega)} = \frac{n(B)}{n(\Omega)} = \frac{n(A \cap B)}{n(\Omega)}$$

$$P(A \cup B) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13} = 0,3077 = 30,77\%$$

Logo, temos aproximadamente 31% de chance de retirar uma dama de copas no baralho.

Exemplo 2

Observando a cor dos olhos e a cor do cabelo de um grupo de 100 pessoas temos:

	Cabelo louro	Cabelo castanho	Total
Olho azul	10	20	30
Olho castanho	30	40	70
Total	40	60	100

Escolhendo uma pessoa ao acaso, qual a probabilidade de ser uma pessoa com cabelo louro e olhos castanhos ou cabelo castanho e olhos azuis?

Etapas:	Respostas
1 - Qual é o experimento?	Escolher uma pessoa em um grupo de 100 pessoas.
2 - Qual é o espaço amostral?	Como o número é elevado, não precisamos enumerar cada ponto amostral, basta determinar o número de elementos do espaço amostral: $n(W) = 100$

<p>3 - Qual é o evento?</p>	<p>Evento A: escolher uma pessoa com cabelo louro e olhos castanhos.</p> <p>Evento B: escolher uma pessoa com cabelo castanho e olhos azuis.</p>
<p>4 - Retirem do espaço amostral os elementos favoráveis ao evento.</p>	<p>$n(A)=30$ $n(B)= 20$ Elementos comuns de A e B → $A \cap B = \{ \quad \}; n(A \cap B) = 0$</p>
<p>5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência?</p> <p>(Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)</p>	<p>Sim → probabilidade clássica.</p>
<p>6 - Fazer o cálculo usando o método apropriado.</p>	<p>Calculando a probabilidade de A ou B temos:</p> $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $P(A \cup B) = \frac{n(A)}{n(\Omega)} = \frac{n(B)}{n(\Omega)} - \frac{n(A \cap B)}{n(\Omega)}$ $P(A \cup B) = \frac{3}{100} + \frac{20}{100} - \frac{0}{100} = \frac{50}{100} = 0,5 = 50\%$ <p>Logo, temos 50% de chance de escolher uma pessoa com cabelo louro e olhos castanhos ou cabelo castanho e olhos azuis.</p>

Exemplo 3

Em um grupo de 60 pessoas, 10 são torcedoras do Botafogo, cinco do Fluminense e o restante torce para o Flamengo. Escolhendo ao acaso um torcedor, qual a probabilidade de ser torcedor do Botafogo ou do Fluminense? Aproveite esse resultado e determine a probabilidade de o torcedor ser flamenguista.

Etapas:	Respostas
1 - Qual é o experimento?	Escolher um torcedor em um grupo de 60 pessoas.
2 - Qual é o espaço amostral?	60 pessoas $\rightarrow n(W) = 60$
3 - Qual é o evento?	Experimento A: ser torcedor do Botafogo. Experimento B: ser torcedor do Fluminense.
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	$n(A) = 10$ $n(B) = 5$ Elementos comuns de A e B \rightarrow $A \cap B = \{ \}; n(A \cap B) = 0$
5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência? (Se a resposta for "Sim", então vamos calcular a fórmula da probabilidade clássica. Se for "Não", iremos usar o cálculo empírico ou frequencial.)	Sim

6 - Fazer o cálculo usando o método apropriado.

Calculando a probabilidade de A ou B temos:

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

$$P(A \cup B) = \frac{n(A)}{n(\Omega)} = \frac{n(B)}{n(\Omega)} = \frac{n(A \cap B)}{n(\Omega)}$$

$$P(A \cup B) = \frac{10}{60} + \frac{5}{60} - \frac{0}{60} = \frac{15}{60} = 0,25 = 25\%$$

Como os eventos dos torcedores do Flamengo, do Botafogo ou do Fluminense são complementares, para determinar a probabilidade de selecionar um torcedor do Flamengo, podemos usar o conceito de probabilidades complementares, ou seja, basta diminuir 100% - 25% = 75%.

Se quisermos confirmar podemos fazer:

C = torcedor do Flamengo $\rightarrow n(C) = 45$

$$P(C) = \frac{45}{60} = 0,75 = 75\%$$

Logo, a probabilidade de selecionar um torcedor do Botafogo ou do Fluminense é igual a 25%. Já a chance do torcedor ser flamenguista é igual a 75%.

Exemplo 4

Uma pesquisa, em uma agência de turismo, sobre o tipo de viagem nos últimos 12 meses, revelou que 45,8% dos clientes viajaram por razões comerciais, 54% por razões pessoais e 30% viajaram por razões pessoais e comerciais.

a) Qual é a probabilidade de que um viajante, selecionado aleatoriamente, tenha viajado nos últimos 12 meses por razões comerciais ou pessoais?

b) Qual é a probabilidade de que um viajante, selecionado aleatoriamente, não tenha viajado nos últimos 12 meses por razões comerciais ou pessoais?

Etapas:	Respostas
1 - Qual é o experimento?	Selecionar um cliente de uma agência de turismo que tenha viajado nos últimos 12 meses.

<p>2 - Qual é o espaço amostral?</p>	<p>Vamos considerar 100 pessoas, em que 45,8% dos clientes viajaram por razões comerciais, 54% por razões pessoais e 30% viajaram por razões pessoais e comerciais.</p>
<p>3 - Qual é o evento?</p>	<p>A = tenha viajado nos últimos 12 meses, por razões comerciais.</p> <p>B = tenha viajado nos últimos 12 meses, por razões pessoais.</p> <p>$A \cap B$ = tenha viajado nos últimos 12 meses, por razões comerciais e pessoais.</p> <p>$A \cup B$ = tenha viajado nos últimos 12 meses, por razões comerciais ou pessoais.</p> <p>$\overline{A \cup B}$ = não tenha viajado nos últimos 12 meses, por razões comerciais ou pessoais.</p>
<p>4 - Retirem do espaço amostral os elementos favoráveis ao evento.</p>	<div data-bbox="608 909 1334 1256"> </div> <p> $n(A) = 45,8$ $n(B) = 54$ $n(A \cap B) = 30$ </p>
<p>5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência?</p> <p>(Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)</p>	<p>Sim, pois o cliente foi escolhido aleatoriamente.</p>

6 - Fazer o cálculo usando o método apropriado.

a)

$$P(A \cup B) = \frac{n(A)}{n(\Omega)} = \frac{n(B)}{n(\Omega)} = \frac{n(A \cap B)}{n(\Omega)}$$

$$P(A \cup B) = \frac{45,8}{100} + \frac{54}{100} - \frac{30}{100} = \frac{69,8}{100} = 0,5 = 50\%$$

b)

$$P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - 0,698 = 30,2\%$$

Exemplo 5

Em um levantamento com 1.000 pessoas sobre a aquisição de um produto lançado no mercado, em três versões, A, B e C, constatou-se que:

470 compraram o produto A.

420 compraram o produto B.

315 compraram o produto C.

110 compraram os produtos A e B.

220 compraram os produtos A e C.

140 compraram os produtos B e C.

75 compraram os três produtos.

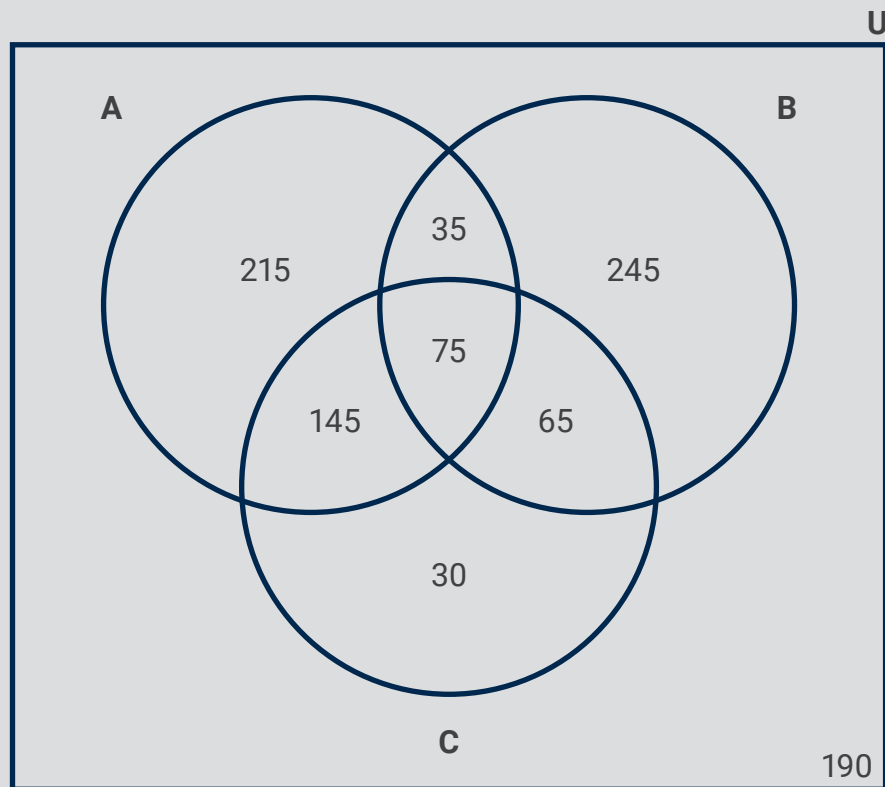
Escolhendo-se ao acaso uma pessoa entrevistada, qual é a probabilidade de que ela tenha adquirido pelo menos uma versão do produto?

Etapas:	Respostas
1 - Qual é o experimento?	Selecionar uma pessoa e perguntar qual foi a versão do produto adquirido.
2 - Qual é o espaço amostral?	1000 entrevistados $\rightarrow n(\Omega) = 1000$

<p>3 - Qual é o evento?</p>	<p>E= escolher pelo menos uma versão do produto. Podemos dividir o evento E em três partes: A = escolher o produto A. B = escolher o produto B. C = escolher o produto C. Logo, $E = A \cup B \cup C$</p>
<p>4 - Retirem do espaço amostral os elementos favoráveis ao evento.</p>	<p>$n(A)=470$ $n(B)=420$ $n(C)=315$</p>
<p>5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência?</p> <p>(Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)</p>	<p>Sim → probabilidade clássica.</p>
<p>6 - Fazer o cálculo usando o método apropriado.</p>	<p>$P(E) = P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$</p> $P(E) = P(A \cup B \cup C) = \frac{470}{1000} + \frac{420}{1000} + \frac{315}{1000} - \frac{110}{1000} - \frac{220}{1000} - \frac{140}{1000} + \frac{75}{1000} = \frac{810}{1000} = 0,81 = 81\%$ <p>Logo, haverá 81% de chance de uma pessoa comprar pelo menos uma das versões do produto lançado no mercado.</p>

Poderíamos resolver esse problema de outra maneira utilizando o **conceito da probabilidade complementar**.

Uma boa estratégia para resolver esse problema é representar as informações do exercício no **diagrama de Veen**.



Vamos começar pela interseção das 3 versões: $A \cap B \cap C = 75$

Depois, passamos para a interseção de duas versões:

$A \cap B = 110$ (como já colocamos 75 na interseção das 3 versões, restam agora $110 - 75 = 35$)

$A \cap C = 220$ (como já colocamos 75 na interseção das 3 versões, restam agora $220 - 75 = 145$)

$B \cap C = 140$ (como já colocamos 75 na interseção das 3 versões, restam agora $140 - 75 = 65$)

Agora vamos descontar as respostas que já foram posicionadas.

$$A = 470 - 35 - 75 - 145 = 215$$

$$B = 420 - 35 - 75 - 65 = 245$$

$$C = 315 - 145 - 75 - 65 = 30$$

Por fim, calculamos das 1.000 pessoas aquelas que não compraram nenhuma versão do produto:

$$1000 - 215 - 35 - 75 - 145 - 245 - 65 - 30 = 190$$

Se fizermos o complementar das pessoas que não compraram nenhuma versão do produto, vamos obter como resultado as pessoas que compraram pelo menos uma versão do produto. Logo:

C = pessoas que compraram pelo menos uma versão do produto

\bar{C} = pessoas que não compraram o produto $\rightarrow n(\bar{C}) = 190$

$$P(C) = 1 - P(\bar{C})$$

$$P(C) = 1 - \frac{190}{1000} = \frac{810}{1000} = 0,81 = 81\%$$

Exemplo 6

Um jogo de dominó é composto por 28 peças, sendo que em cada peça são representadas duas quantidades de bolinhas, que variam de zero a seis.

Se selecionarmos uma peça aleatoriamente, qual é a probabilidade de possuir ao menos um 3 ou 5 em sua face?

Etapas:	Respostas
1 - Qual é o experimento?	Retirar uma peça do dominó.
2 - Qual é o espaço amostral?	$\Omega = \{(0,0), (0,1), (0,2), (0,3), (0,4), (0,5), (0,6), (1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,2), (2,3), (2,4), (2,5), (2,6), (3,3), (3,4), (3,5), (3,6), (4,4), (4,5), (4,6), (5,5), (5,6), (6,6)\} \rightarrow n(\Omega) = 28$

3 - Qual é o evento?	<p>Evento A: Saírem 3 bolinhas.</p> <p>Evento B: Saírem 5 bolinhas.</p> <p>$A \cup B$: sair 3 OU 5 bolinhas</p>
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	<p>$A = \{(0,3), (1,3), (2,3), (3,3), (3,4), (3,5), (3,6)\} \rightarrow n(A) = 7$</p> <p>$B = \{(0,5), (1,5), (2,5), (3,5), (4,5), (5,5), (5,6)\} \rightarrow n(B) = 7$</p> <p>Verificando a existência da interseção $\rightarrow A \cap B = \{(3,5)\} \rightarrow n(A \cap B) = 1$</p>
<p>5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência?</p> <p>(Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)</p>	Sim
6 - Fazer o cálculo usando o método apropriado.	<p>Calculando a probabilidade do evento A ou B ocorrerem:</p> $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $P(A \cup B) = \frac{7}{28} + \frac{7}{28} - \frac{1}{28} = \frac{13}{28} = 0,464 = 46,4\%$ <p>Logo, haverá 46% de chance de uma peça do dominó possuir ao menos um 3 ou um 5 em sua face.</p>



Ampliando o foco

Para colocar em prática o conteúdo visto nesta unidade visite a Biblioteca Virtual, escolha um livro de Estatística, selecione e faça alguns exercícios do livro para reforçar o conhecimento.

Confira as respostas que estão no final do livro e, se tiver alguma dúvida, fale com o seu tutor pelo Fórum de Dúvidas – Fale com o Tutor.

Probabilidade condicional

Neste tópico continuaremos os estudos abordando mais algumas relações básicas das probabilidades: a probabilidade para eventos condicionados, a probabilidade para eventos independentes e a interseção de eventos. Também apresentaremos o Teorema de Bayes.

Probabilidade condicional

Dados dois eventos “A” e “B” de um espaço amostral $\Omega \neq \emptyset$, chamamos de probabilidade de “B” condicionada a “A”, a probabilidade de ocorrer “B”, sabendo que “A” já ocorreu.

Representação: $P(B|A)$ ou $P(B/A)$

Dizemos: “probabilidade de “B” ocorrer, sabendo que “A” já ocorreu.

Cálculo:

$$P(B|A) = \frac{n(A \cap B)}{n(A)} \quad \text{ou} \quad P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Exemplo1:

Para oferecer uma promoção do tipo “fale grátis de seu móvel para seu fixo”, escolhe-se aleatoriamente uma pessoa de uma base de dados com 15.000 usuários de telefonia, dos quais 10.000 possuem telefones fixos, 8.000 telefones móveis e 3.000 têm telefones fixos e móveis.

Pergunta-se:

- a) Já sabendo que a pessoa selecionada tem um telefone móvel, qual a probabilidade de ela ter telefone fixo também?

Etapas:	Respostas
1 - Qual é o experimento?	Escolher aleatoriamente uma pessoa em uma base de dados.
2 - Qual é o espaço amostral?	15.000 usuários de telefonia, dos quais 10.000 possuem telefones fixos, 8.000 telefones móveis e 3.000 têm telefones fixos e móveis.
3 - Qual é o evento?	<p>A= selecionar uma pessoa com telefone móvel.</p> <p>B= selecionar uma pessoa com telefone fixo.</p> <p>B A= selecionar uma pessoa com telefone fixo sabendo que ela tem um telefone móvel.</p>
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	<p>$n(A) = 8.000$</p> <p>$n(B) = 10.000$</p> <p>$n(A \cap B) = 3.000$</p>
<p>5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência?</p> <p>(Se a resposta for "Sim", então vamos calcular a fórmula da probabilidade clássica. Se for "Não", iremos usar o cálculo empírico ou frequencial.)</p>	<p>Sim, pois a pessoa é escolhida aleatoriamente. Temos que observar que uma condição já aconteceu nesse evento, ou seja, já sabemos que a pessoa possui uma linha de telefonia móvel, por isso vamos usar o conceito de probabilidade condicional.</p>
6 - Fazer o cálculo usando o método apropriado.	$P(B A) = \frac{n(A \cap B)}{n(A)}$ $P(B A) = \frac{3000}{8000} = 0,375 = 37,5\%$ <p>Portanto, é provável que 37,5% dos usuários de telefonia da base consultada tenham um telefone fixo, sabendo que ela já possuía um telefone móvel.</p>

b) Já sabendo que ela tem um telefone fixo, qual a probabilidade de ela ter telefone móvel também?

Etapas:	Respostas
1 - Qual é o experimento?	Escolher aleatoriamente uma pessoa em uma base de dados.
2 - Qual é o espaço amostral?	15.000 usuários de telefonia, dos quais 10.000 possuem telefones fixos, 8.000 telefones móveis e 3.000 têm telefones fixos e móveis.
3 - Qual é o evento?	A= selecionar uma pessoa com telefone móvel. B= selecionar uma pessoa com telefone fixo. A B= selecionar uma pessoa com telefone móvel sabendo que ela tem um telefone fixo.
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	$n(A) = 8.000$ $n(B) = 10.000$ $n(A \cap B) = 3.000$
5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência? (Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)	Sim, pois a pessoa é escolhida aleatoriamente. Temos que observar que uma condição já aconteceu nesse evento, ou seja, já sabemos que a pessoa possui uma linha de telefonia fixo, por isso vamos usar o conceito de probabilidade condicional.
6 - Fazer o cálculo usando o método apropriado.	$P(B A) = \frac{n(A \cap B)}{n(B)}$ $P(B A) = \frac{3000}{10000} = 0,3 = 30\%$ <p>Portanto, é provável que 30% dos usuários de telefonia da base consultada tenham um telefone móvel, sabendo que ela já possuía um telefone fixo.</p>

Exemplo 2:

Em uma escola com 100 alunos, 60 estudam Estatística, 50 estudam Inglês e 20 estudam Estatística e Inglês. Sabendo que um aluno escolhido aleatoriamente já estuda Estatística, qual a probabilidade de ele estudar Inglês?

Etapas:	Respostas
1 - Qual é o experimento?	Escolher aleatoriamente um aluno .
2 - Qual é o espaço amostral?	100 alunos, 60 estudam Estatística, 50 estudam Inglês e 20 estudam Estatística e Inglês.
3 - Qual é o evento?	<p>A= escolher um aluno que estude Estatística.</p> <p>B= escolher um aluno que estude Inglês.</p> <p>$B A$= escolher um aluno que estude Inglês, sabendo que esse aluno já estuda Estatística.</p>
<p>4 - Retirem do espaço amostral os elementos favoráveis ao evento.</p>	<div data-bbox="608 1059 1145 1317"> <p>Diagrama de Venn com dois conjuntos E e I. O conjunto E (Estatística) contém 40 elementos exclusivos e 20 elementos em comum com o conjunto I (Inglês). O conjunto I contém 30 elementos exclusivos e 20 elementos em comum com o conjunto E. O total de elementos no universo é 100.</p> </div> <p> $n(A) = 60$ $n(B) = 50$ $n(A \cap B) = 20$ </p>
<p>5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência?</p> <p>(Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)</p>	<p>Sim, pois a pessoa é escolhida aleatoriamente. Temos que observar que uma condição já aconteceu nesse evento, ou seja, já sabemos que a pessoa escolhida estuda Estatística, por isso vamos usar o conceito de probabilidade condicional.</p>

6 - Fazer o cálculo usando o método apropriado.

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

$$P(B|A) = \frac{20}{60} = 0,33 = 33\%$$

Portanto, é provável que 30% dos usuários de telefonia da base consultada tenham um telefone móvel, sabendo que ela já possuía um telefone fixo.

Probabilidade para eventos independentes

Dois eventos A e B são independentes se, e somente se:

$$P(A|B) = P(A) \quad \text{ou} \quad P(B|A) = P(B)$$

$$\text{Substituindo } P(B|A) = P(B) \text{ em: } P(B|A) = \frac{P(A \cap B)}{P(A)} \text{ temos: } P(B) = \frac{P(A \cap B)}{P(A)}$$

$$\text{Logo: } P(A \cap B) = P(A) \cdot P(B)$$

Analogamente,

$$\text{Substituindo } P(B|A) = P(A) \text{ em: } P(B|A) = \frac{P(A \cap B)}{P(B)} \text{ temos: } P(A) = \frac{P(A \cap B)}{P(B)}$$

$$\text{Logo: } P(A \cap B) = P(A) \cdot P(B)$$

Exemplo 1

O quadro a seguir apresenta o resultado de uma pesquisa sobre a primeira razão para a escolha de uma determinada Instituição de Ensino Superior – IES:

	Custo da mensalidade	Qualidade do ensino	Outros	Total
Ensino presencial	393	421	76	890
Ensino a distância	593	400	46	1039
Total	986	821	122	1929

a) Se selecionarmos um aluno aleatoriamente e observarmos que ele optou por ensino presencial, qual a probabilidade de a qualidade de ensino ser a primeira razão para a escolha da IES?

Etapas:	Respostas
1 - Qual é o experimento?	Selecionar aleatoriamente um aluno entrevistado.
2 - Qual é o espaço amostral?	Está representado no quadro no enunciado da questão.
3 - Qual é o evento?	A= aluno ser do ensino presencial. B= aluno responder que a qualidade de ensino é a primeira razão para a escolha da IES.
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	$n(A) = 890$ $n(B) = 821$ $n(A \cap B) = 421$
5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência? (Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)	Sim, pois o aluno foi escolhido aleatoriamente. Porém, temos que prestar atenção, pois uma condição já aconteceu nesse evento, ou seja, já sabemos que a pessoa escolhida frequenta o ensino presencial e por isso vamos usar o conceito de probabilidade condicional.

6 - Fazer o cálculo usando o método apropriado.	$P(B A) = \frac{n(A \cap B)}{n(A)}$ $P(B A) = \frac{421}{890} = 0,473 = 47,3\%$ <p>Logo, existe aproximadamente 47,3% de probabilidade de que o aluno que estuda presencialmente tenha escolhido a qualidade do ensino como primeira razão para escolher a IES.</p>
---	---

b) Se um aluno selecionado aleatoriamente optou por ensino a distância, qual a probabilidade de o custo da mensalidade ser a primeira razão para escolha da IES?

Etapas:	Respostas
1 - Qual é o experimento?	Selecionar aleatoriamente um aluno entrevistado.
2 - Qual é o espaço amostral?	Está representado no quadro, no enunciado da questão.
3 - Qual é o evento?	<p>A= aluno ser do ensino a distância.</p> <p>B= aluno responder que o custo da mensalidade é a primeira razão para escolha da IES.</p>
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	<p>$n(A) = 1039$</p> <p>$n(B) = 986$</p> <p>$n(A \cap B) = 593$</p>
<p>5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência?</p> <p>(Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)</p>	<p>Sim, pois o aluno foi escolhido aleatoriamente.</p> <p>Porém, temos que prestar atenção, pois uma condição já aconteceu nesse evento, ou seja, já sabemos que a pessoa escolhida é do EAD e por isso vamos usar o conceito de probabilidade condicional.</p>

6 - Fazer o cálculo usando o método apropriado.

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

$$P(B|A) = \frac{593}{1039} = 0,571 = 57,1\%$$

Logo, existe aproximadamente 57,1% de probabilidade de que o aluno que estuda a distância tenha escolhido o custo da mensalidade como primeira razão para escolher a IES.

c) Considere os eventos: A=estudante ser do ensino presencial e B=qualidade do ensino como a primeira razão de escolha da IES. Os eventos A e B são independentes?

A= ser do ensino presencial $\rightarrow n(A) = 890$

B = qualidade de ensino $\rightarrow n(B) = 821$

$n(A \cap B) = 421$

$n(\Omega) = 1929$

$$P(A) = \frac{\text{ser do presencial}}{\text{total de alunos}} = \frac{890}{1929} = 0,461 = 46,1\%$$

$$P(B) = \frac{\text{qualidade de ensino}}{\text{total de alunos}} = \frac{821}{1929} = 0,426 = 42,6\%$$

$$P(A \cap B) = \frac{\text{qualidade de ensino}}{\text{total de alunos}} = \frac{n(A \cap B)}{n(\Omega)} = \frac{421}{1929} = 0,218 = 21,8\%$$

Se os eventos forem independentes, temos: $P(A \cap B) = P(A) \cdot P(B)$

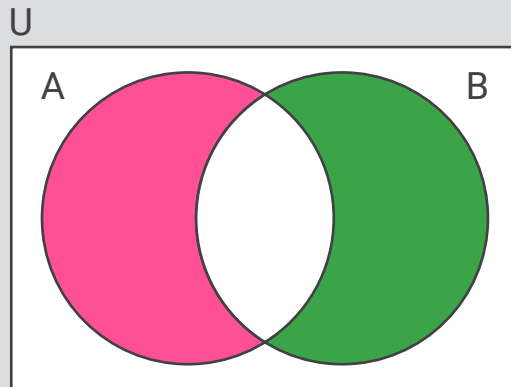
Comparando 0,218 com 0,461. 0,426

$0,218 \neq 0,196$

Como $P(A \cap B) \neq P(A) \cdot P(B)$ os eventos não são independentes.

Interseção de eventos (U)

Sejam dois eventos A e B:



Na interseção, consideramos somente os elementos da parte branca.

Interseção $\rightarrow \cap \rightarrow$ palavra usada "E" \rightarrow operação (x)

$P(A \cap B) = P(A) \cdot P(B) \rightarrow$ eventos independentes

$P(A \cap B) = P(A) \cdot P(B/A) \rightarrow$ eventos condicionados

$P(A \cap B) = P(A) \cdot P(A/B) \rightarrow$ eventos condicionados

Vejamos alguns exemplos.

Exemplo 1

Suponha que a probabilidade de a fábrica de respiradores entregar a mercadoria no prazo seja 99,4%. Suponha que duas entregas de respiradores sejam selecionadas ao acaso. Qual a probabilidade de:

- a) Ambos os respiradores serem entregues no prazo.
- b) Exatamente um dos respiradores for entregue no prazo.
- c) Ambos os respiradores forem entregues com atraso.

Resolução:

Sejam:

Evento A: entregar o respirador 1 no prazo $\rightarrow P(A) = 0,994$

Evento \bar{A} : não entregar o respirador 1 no prazo $\rightarrow P(\bar{A}) = 1 - 0,994 = 0,006$

Evento B: entregar o respirador 2 no prazo $\rightarrow P(B) = 0,994$

Evento \bar{B} : não entregar o respirador 2 no prazo $\rightarrow P(\bar{B}) = 1-0,994 = 0,006$

a) O respirador 1 foi entregue no prazo **E** o respirador 2 foi entregue no prazo.

A palavra-chave aqui é o “E”, significando a interseção (multiplicação) dos eventos. Logo:

$$P(A \cap B) = P(A) \cdot P(B)$$

$$P(A \cap B) = 0,994 \cdot 0,994 = 0,988 = 98,8\%$$

b) O respirador 1 foi entregue no prazo E o respirador 2 não foi entregue no prazo OU o respirador 1 não foi entregue no prazo E o respirador 2 foi entregue no prazo.

A palavra-chave aqui é o “E”, significando a interseção (multiplicação) dos eventos e também o “OU” significando a união (soma) dos eventos. Logo:

$$P(\bar{A} \cap B) \text{ ou } P(A \cap \bar{B}) = P(\bar{A}) \cdot P(B) + P(A) \cdot P(\bar{B}) = 0,994 \cdot 0,006 + 0,006 \cdot 0,994 = 0,012 = 1,2\%$$

c) Ambos os respiradores forem entregues com atraso, logo o respirador 1 foi entregue com atraso E o respirador 2 foi entregue com atraso.

$$P(\bar{A} \cap \bar{B}) = P(\bar{A}) \cdot P(\bar{B})$$

$$P(\bar{A} \cap \bar{B}) = 0,006 \cdot 0,006 = 0,000036 = 0,0036\%$$

Exemplo 2

Em um lote de 12 peças selecionadas para inspeção, quatro apresentam defeitos. Três peças são retiradas ao acaso, uma após a outra, sem reposição. Calcule a probabilidade de as três peças selecionadas não apresentarem defeitos.

Etapas:	Respostas
1 - Qual é o experimento?	Retirar três peças para serem inspecionadas.
2 - Qual é o espaço amostral?	12 peças, sendo quatro defeituosas e oito sem defeitos.

3 - Qual é o evento?	<p>A= selecionar para inspeção a primeira peça sem defeito.</p> <p>B= selecionar para inspeção a primeira peça sem defeito.</p> <p>C= selecionar para inspeção a primeira peça sem defeito.</p> <p>D= selecionar a primeira peça sem defeito para inspeção</p> <p>E selecionar a segunda peça sem defeito para inspeção E</p> <p>selecionar a terceira peça sem defeito para inspeção.</p>
4 - Retirem do espaço amostral os elementos favoráveis ao evento.	<p>$n(A) = 8$</p> <p>$n(B) = 7$</p> <p>$n(C) = 6$</p>
<p>5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência?</p> <p>(Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)</p>	<p>Sim, pois todas as peças têm a mesma chance de serem selecionadas para inspeção.</p>
6 - Fazer o cálculo usando o método apropriado.	<p>$P(D) = P(A) \cdot P(B) \cdot P(C)$</p> $P(D) = \frac{8}{12} \cdot \frac{7}{11} \cdot \frac{6}{10} = \frac{336}{1320} = 0,25 = 25\%$

Resumindo		
Símbolo	Palavra	Operação
\cup	ou	adição
\cap	e	Multiplicação
\bar{A}	não	$P(\bar{A}) = 1 - P(A)$

$A \cup B$	É o evento que ocorre se A ocorre, ou se B ocorrer ou se A e B ocorrerem	$P(A \cup B) = P(A) + P(B) \rightarrow$ eventos mutuamente exclusivos $P(A \cup B) = P(A) + P(B) - P(A \cap B) \rightarrow$ eventos não exclusivos
$A \cap B$	É o evento que ocorre se A e B ocorrem	$P(A \cap B) = P(A) \cdot P(B/A)$ $P(A \cap B) = P(A) \cdot P(B/A)$

Axiomas das probabilidades

A probabilidade clássica $P(A)$ satisfaz aos seguintes axiomas:

A₁) A probabilidade é um número maior ou igual a zero e menor ou igual a 1.

$$0 \leq P(A) \leq 1$$

A₂) A soma das probabilidades de todos os resultados possíveis em um espaço amostral é igual a 1 ou 100%.

$$P(\Omega) = 1$$

A₃) Se os eventos forem, dois a dois mutuamente exclusivos então:

$$P(E) = P(E_1) + P(E_2) + \dots + P(E_n)$$

Teoremas do cálculo das probabilidades para eventos equiprováveis

T₁) Se "A" e "B" são eventos equiprováveis quaisquer, então:

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

Para três eventos, temos:

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

T₂) Se \bar{A} é o complementar de A, então a soma das probabilidades de um evento com seu evento complementar é sempre igual a 100%, ou seja, igual a 1:

$$P(A) + P(\bar{A}) = 1 \quad \therefore \quad P(\bar{A}) = 1 - P(A) \quad \text{ou} \quad P(A) = 1 - P(\bar{A})$$

T₃) Se $A \subset B$, então:

$$P(A) \leq P(B).$$

Vejamos um exemplo que ilustra o conceito apresentado.

Exemplo

Um número é escolhido entre os cinco primeiros números naturais. Considere os eventos a seguir e encontre a probabilidade de que o número escolhido seja:

- a) A= Menor do que 5.
- b) B= Maior do que 5.
- c) C= Igual a zero.
- d) D= Igual a 1.
- e) E= Igual a 2.
- f) F= Igual a 3.
- g) G= Igual a 4.
- h) H= Igual a 1 ou igual a 4 ou igual a um número primo.
- i) I= Par ou primo.
- j) J= Diferente de 1.
- k) K=um número par.
- l) L=um número ímpar.
- m) M= um número par ou ímpar.

Resolução:

Primeiramente vamos definir:

Experimento: escolher um número entre os cinco primeiros números naturais

Espaço Amostral: $\Omega = \{0,1,2,3,4\}$, logo o número de elementos do espaço amostral será:
 $n(\Omega) = 5$

A seguir vamos calcular as probabilidades solicitadas:

a) A= nº menor do que 5 $\rightarrow A = \{0,1,2,3,4\}$, logo, o número de elementos do evento A será: $n(A) = 5$ então:

$$P(A) = \frac{n(A)}{n(\Omega)} = \frac{5}{5} = 1 = 100\%$$

Observamos que o número de elementos do evento A é igual ao número de elementos do espaço amostral. Esse é o maior número de elementos que podemos retirar do espaço amostral, portanto essa será a maior probabilidade que poderá ocorrer. É um exemplo do evento certo.

b) B= Maior do que 5 $\rightarrow B=\{ \}$, não existe nenhum elemento no espaço amostral maior do que 5, logo o número de elementos do evento B será: $n(B) = 0$ então:

$$P(B) = \frac{n(B)}{n(\Omega)} = \frac{0}{5} = 0 = 0\%$$

Observamos que o número de elementos do evento é o menor número de elementos que podemos escolher no espaço amostral, portanto essa será a menor probabilidade que poderá ocorrer. É um exemplo do evento impossível.

Concluimos então, observando os eventos A e B, que a probabilidade de ocorrência de um evento, é um número maior ou igual a zero e menor ou igual a 1, conforme afirma o **axioma 1**:
 $0 \leq P(A) \leq 1$

c) C= Igual a zero $\rightarrow C=\{0\}$, logo, o número de elementos do evento C será: $n(C)=1$ então:

$$P(C) = \frac{n(C)}{n(\Omega)} = \frac{1}{5} = 0,2 = 20\%$$

d) D= Igual a 1 $\rightarrow D=\{1\}$, logo, o número de elementos do evento D será: $n(D)=1$ então:

$$P(D) = \frac{n(D)}{n(\Omega)} = \frac{1}{5} = 0,2 = 20\%$$

e) E= Igual a 2 $\rightarrow E=\{2\}$, logo, o número de elementos do evento E será: $n(E)=1$ então:

$$P(E) = \frac{n(E)}{n(\Omega)} = \frac{1}{5} = 0,2 = 20\%$$

f) $F = \text{Igual a } 3 \rightarrow F = \{3\}$, logo, o número de elementos do evento F será: $n(F) = 1$ então:

$$P(F) = \frac{n(F)}{n(\Omega)} = \frac{1}{5} = 0,2 = 20\%$$

g) $G = \text{Igual a } 4 \rightarrow G = \{4\}$, logo, o número de elementos do evento G será: $n(G) = 1$ então:

$$P(G) = \frac{n(G)}{n(\Omega)} = \frac{1}{5} = 0,2 = 20\%$$

Se somarmos as probabilidades dos eventos C, D, E, F e G estamos somando as probabilidades de todos os elementos do espaço amostral. Logo, vamos obter: $0,2 + 0,2 + 0,2 + 0,2 + 0,2 = 1$ ou 100%, o que afirma o axioma 2, ou seja: a soma das probabilidades de todos os resultados possíveis de um espaço amostral é igual a 1 ou 100%.

h) $H = \text{o número igual a } 1 \text{ ou o número igual a } 4 \text{ ou um número primo.} \rightarrow H = \{1, 4, 2, 3\}$, logo, o número de elementos do evento H será: $n(H) = 4$ então:

$$P(H) = \frac{n(H)}{n(\Omega)} = \frac{4}{5} = 0,8 = 80\%$$

2ª Temos outra maneira de calcular essa probabilidade. Pode ser a seguinte:

Observamos que podemos dividir o evento H em 3 eventos: $H_1 = \text{o número igual a } 1$, $H_2 = \text{o número igual a } 4$ e $H_3 = \text{um número primo}$. Vamos buscar no espaço amostral os elementos dos eventos.

$$H_1 = \{1\}$$

$$H_2 = \{4\}$$

$$H_3 = \{2, 3\}$$

Vamos observar, dois a dois, as interseções dos elementos dos eventos:

$$H_1 \cap H_2 = \{ \}$$

$$H_1 \cap H_3 = \{ \}$$

$$H_2 \cap H_3 = \{ \}$$

Como a interseção entre os elementos dos eventos dois a dois é vazia, concluímos que são eventos mutuamente exclusivos.

Logo, para calcular a probabilidade H = o número igual a 1 ou o número igual a 4 ou um número primo, basta calcular cada probabilidade separadamente e efetuar a soma das probabilidades.

Então:

$$P(H) = \frac{n(H_1)}{n(\Omega)} = \frac{n(H_2)}{n(\Omega)} = \frac{n(H_3)}{n(\Omega)} = \frac{1}{5} + \frac{1}{5} + \frac{2}{5} = \frac{4}{5} = 0,8 = 80\%$$

Com esse item podemos verificar o que diz no **axioma 3**: se os eventos forem dois a dois mutuamente exclusivos, então: $P(E) = P(E_1) + P(E_2) + \dots + P(E_n)$

i) I =Par ou primo $\rightarrow I=\{0,2,3,4\}$, logo, o número de elementos do evento I será: $n(I)=4$ então:

$$P(H) = \frac{n(I)}{n(\Omega)} = \frac{4}{5} = 0,8 = 80\%$$

3ª Temos outra maneira de calcular essa probabilidade. Pode ser a seguinte:

Observamos que podemos dividir o evento I em 2 eventos: I_1 = número par, I_2 = número primo. Vamos buscar no espaço amostral os elementos dos eventos:

$$I_1=\{0,2,4\} \quad n(I_1) = 3$$

$$I_2=\{2,3\} \quad n(I_2) = 2$$

Vamos observar, dois a dois, as interseções dos elementos dos eventos:

$$I_1 \cap I_2 = \{2\} \therefore n(I_1 \cap I_2) = 1$$

Nesse caso, observamos um elemento na interseção dos eventos. Logo, os eventos não são mutuamente excludentes. Precisamos de, em vez de usar o axioma 3 para calcular a probabilidade, devemos usar o **teorema 1**, que diz:

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

Logo:

$$P(I_1 \cup I_2) = P(I_1) + P(I_2) - P(I_1 \cap I_2)$$

$$P(I_1 \cup I_2) = \frac{n(I_1)}{n(\Omega)} = \frac{n(I_2)}{n(\Omega)} = \frac{n(I_1 \cap I_2)}{n(\Omega)}$$

$$P(I_1 \cup I_2) = \frac{3}{5} + \frac{2}{5} - \frac{1}{5} = \frac{4}{5} = 0,8 = 80\%$$

j) J= Diferente de 1 $\rightarrow I=\{0,2,3,4\}$, logo, o número de elementos do evento I será: $n(I)=4$ então:

$$P(I) = \frac{n(I)}{n(\Omega)} = \frac{4}{5} = 0,8 = 80\%$$

4ª Temos outra maneira de calcular essa probabilidade. Pode ser a seguinte:

Outra maneira de calcular a probabilidade pedida seria utilizando o **teorema 2**, que diz: se \bar{A} é o complementar de A, então a soma das probabilidades de um evento com seu evento complementar é sempre igual a 100%, ou seja, $P(I) + P(\bar{I}) = 1 \therefore P(\bar{I}) = 1 - P(I)$ ou $P(I) = 1 - P(\bar{I})$

Se "J" é o evento de escolha do número diferente de 1, então seria escolha do número igual a 1. Nós já calculamos, na letra "d" deste exercício, que a probabilidade de retirar o número 1 cujo resultado foi igual a 0,2, ou seja, $P(\bar{I}) = 0,2$ então, a probabilidade de não retirar o número 1 seria: $P(I) = 1 - P(\bar{I}) = 1 - 0,2 = 0,8$

Teorema de Bayes

Também conhecido como Teorema da Probabilidade das Causas ou dos Antecedentes, ou ainda Teorema da Probabilidade a posteriori.

Esse teorema preocupa-se em responder à seguinte pergunta:

Supondo que o evento B já tenha ocorrido, qual a probabilidade que este evento tenha provindo do evento A_i?

O Teorema de Bayes é importante porque inverte probabilidades condicionais. Ele nos dá a probabilidade de uma causa ocorrer, desde que um evento já tenha ocorrido.

O teorema diz: "Se "n" eventos A_1, A_2, A_3, \dots , são dois a dois mutuamente exclusivos, tais que, $A_1 \cup A_2 \cup A_3 \cup \dots A_n = \Omega$, e $P(A_i) > 0$ então:

$$P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{\sum_j P(A_j) \cdot P(B|A_j)}$$

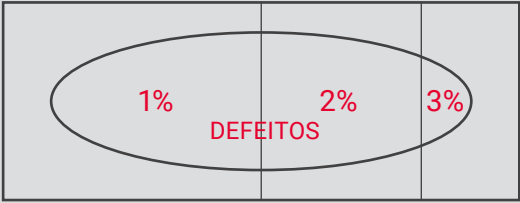
Para entendermos melhor o assunto, vejamos alguns exemplos práticos.

Exemplo 1

Em uma fábrica de produção de aparelhos celulares, as linhas de montagem I, II e III respondem respectivamente por 50, 40 e 10 por cento da produção. Alguns aparelhos saem destas linhas com defeitos. A porcentagem de celulares defeituosos é de 1%, 2% e 3%, respectivamente, para as linhas I, II e III. Para evitar que os aparelhos defeituosos saiam da empresa e cheguem ao mercado, o controle de qualidade realiza inspeções individuais em todos os celulares fabricados e os que apresentam algum defeito são enviados para uma linha especial de recuperação. Você, como responsável pelo controle de qualidade, precisa intervir na linha de montagem mais problemática. Por qual linha de montagem você iniciaria seus trabalhos?

Para ajudar nessa decisão será preciso conhecer:

- Qual a probabilidade de o celular defeituoso ter sido produzido na linha de montagem I?
- Qual a probabilidade de o celular defeituoso ter sido produzido na linha de montagem II?
- Qual a probabilidade de o celular defeituoso ter sido produzido na linha de montagem III?

Etapas:	Respostas
1 - Qual é o experimento?	Retirar um aparelho defeituoso para inspeção.
2 - Qual é o espaço amostral?	<div style="display: flex; justify-content: space-around; text-align: center;"> <div> <p>Linha de produção 1</p> <p>50%</p> </div> <div> <p>Linha de produção 2</p> <p>40%</p> </div> <div> <p>Linha de produção 3</p> <p>10%</p> </div> </div> 

3 - Qual é o evento?

Determinar a probabilidade de um aparelho celular defeituoso encontrado na inspeção final ter sido produzido na linha de produção I, na linha de produção II e na linha de produção III.

A_1 = aparelho ter sido produzido na linha de produção I.

A_2 = aparelho ter sido produzido na linha de produção II.

A_3 = aparelho ter sido produzido na linha de produção III.

B = aparelho com defeito.

$B \setminus A_1$ = aparelho defeituoso sabendo que foi produzido na linha de produção I

$B \setminus A_2$ = aparelho defeituoso sabendo que foi produzido na linha de produção II

$B \setminus A_3$ = aparelho defeituoso sabendo que foi produzido na linha de produção III

O que queremos investigar:

$A_1 \setminus B$ = aparelho ter sido produzido na linha de montagem I sabendo que ele é defeituoso

$A_2 \setminus B$ = aparelho ter sido produzido na linha de montagem II sabendo que ele é defeituoso

$A_3 \setminus B$ = aparelho ter sido produzido na linha de montagem III, sabendo que ele é defeituoso.

Ou seja, precisamos usar o teorema de Bayes para calcular a probabilidade condicional invertida.

$$P(A_i \setminus B) = \frac{P(A_i) \cdot P(B \setminus A_i)}{\sum_i P(A_i) \cdot P(B \setminus A_i)}$$

<p>4 - Retirem do espaço amostral os elementos favoráveis ao evento.</p>	<p> $P(A_1) = 50\% = 0,5$ $P(A_2) = 40\% = 0,4$ $P(A_3) = 10\% = 0,1$ $P(B A_1) = 1\% = 0,01$ $P(B A_2) = 2\% = 0,02$ $P(B A_3) = 3\% = 0,03$ $P(A_1 B) = ???$ é o que queremos determinar $P(A_2 B) = ???$ é o que queremos determinar $P(A_3 B) = ???$ é o que queremos determinar </p>
<p>5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência?</p> <p>(Se a resposta for “Sim”, então vamos calcular a fórmula da probabilidade clássica. Se for “Não”, iremos usar o cálculo empírico ou frequencial.)</p>	<p>Sim, pois todos os aparelhos têm a mesma chance de serem selecionados para inspeção.</p>
<p>6 - Fazer o cálculo usando o método apropriado.</p>	$P(A_i B) = \frac{P(A_i) \cdot P(B A_i)}{\sum_i P(A_i) \cdot P(B A_i)}$ <p>Cálculo da probabilidade da peça defeituosa ter sido produzida na linha I:</p> $P(A_1 B) = \frac{P(A_1) \cdot P(B A_1)}{P(A_1) \cdot P(B A_1) + P(A_2) \cdot P(B A_2) + P(A_3) \cdot P(B A_3)}$ $P(A_1 B) = \frac{0,5 \cdot 0,01}{0,5 \cdot 0,01 + 0,4 \cdot 0,02 + 0,1 \cdot 0,03}$ $P(A_1 B) = \frac{0,005}{0,005 + 0,008 + 0,003} = \frac{0,005}{0,016} = 0,3125 = 31,5\%$

$$P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{\sum_i P(A_i) \cdot P(B|A_i)}$$

Cálculo da probabilidade da peça defeituosa ter sido produzida na linha II:

$$P(A_2|B) = \frac{P(A_2) \cdot P(B|A_2)}{P(A_1) \cdot P(B|A_1) + P(A_2) \cdot P(B|A_2) + P(A_3) \cdot P(B|A_3)}$$

$$P(A_2|B) = \frac{0,4 \cdot 0,02}{0,5 \cdot 0,01 + 0,4 \cdot 0,02 + 0,1 \cdot 0,03}$$

$$P(A_2|B) = \frac{0,005}{0,005 + 0,008 + 0,003} = \frac{0,005}{0,016} = 0,3125 = 31,25\%$$

Cálculo da probabilidade da peça defeituosa ter sido produzida na linha III:

$$P(A_3|B) = \frac{P(A_3) \cdot P(B|A_3)}{P(A_1) \cdot P(B|A_1) + P(A_2) \cdot P(B|A_2) + P(A_3) \cdot P(B|A_3)}$$

$$P(A_3|B) = \frac{0,1 \cdot 0,03}{0,5 \cdot 0,01 + 0,4 \cdot 0,02 + 0,1 \cdot 0,03}$$

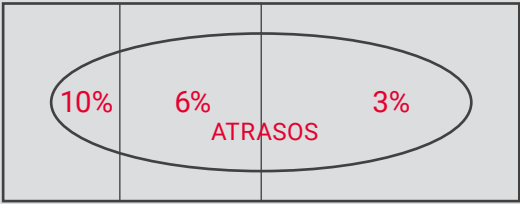
$$P(A_3|B) = \frac{0,003}{0,005 + 0,008 + 0,003} = \frac{0,003}{0,016} = 0,1875 = 18,75\%$$

Logo, você deve escolher a linha de produção II para fazer a intervenção.

Exemplo 2

Um viajante usa os serviços da companhia aérea "A" em 20% das viagens, 30% da companhia "B" e no restante do tempo ele utiliza a companhia "C". Sabe-se que a porcentagem dos voos que decolam com atraso é de 10%, 6% e 3% respectivamente para as companhias "A", "B" e "C". Suponha que o voo do viajante teve atraso.

Qual a probabilidade de que ele tenha utilizado a companhia "C" nessa viagem?

Etapas:	Respostas
1 - Qual é o experimento?	Verificar a probabilidade de um viajante ter utilizado a companhia aérea "A", sabendo que ela teve um atraso no voo.
2 - Qual é o espaço amostral?	<div style="display: flex; justify-content: space-around; margin-bottom: 10px;"> <div>CIA "A" 20%</div> <div>CIA "B" 30%</div> <div>CIA "C" 50%</div> </div> 
3 - Qual é o evento?	<p> A_1 = viajante ter escolhido a cia aérea "A" para viajar. A_2 = viajante ter escolhido a cia aérea "B" para viajar. A_3 = viajante ter escolhido a cia aérea "C" para viajar. </p> <p>B = o voo decolou atrasado.</p> <p> $B \setminus A_1$ = voo atrasado sabendo que a cia escolhida foi a "A" $B \setminus A_2$ = voo atrasado sabendo que a cia escolhida foi a "B" $B \setminus A_3$ = voo atrasado sabendo que a cia escolhida foi a "C" </p> <p>O que queremos investigar:</p> <p>$A_3 \setminus B$ = a cia escolhida foi "C" sabendo que o voo decolou com atraso.</p> <p>Ou seja, precisamos usar o teorema de Bayes para calcular a probabilidade condicional invertida.</p>

<p>4 - Retirem do espaço amostral os elementos favoráveis ao evento.</p>	$P(A_1) = 30\% = 0,2$ $P(A_2) = 20\% = 0,3$ $P(A_3) = 50\% = 0,5$ $P(B A_1) = 10\% = 0,1$ $P(B A_2) = 6\% = 0,06$ $P(B A_3) = 3\% = 0,03$ $P(A_3 B) = ??? \text{ é o que queremos determinar}$
<p>5 - Cada elemento do espaço amostral tem a mesma chance de ocorrência?</p> <p>(Se a resposta for "Sim", então vamos calcular a fórmula da probabilidade clássica. Se for "Não", iremos usar o cálculo empírico ou frequencial.)</p>	<p>Sim.</p>
<p>6 - Fazer o cálculo usando o método apropriado.</p>	$P(A_i B) = \frac{P(A_i) \cdot P(B A_i)}{\sum_i P(A_i) \cdot P(B A_i)}$ <p>Cálculo da probabilidade da peça defeituosa ter sido produzida na linha I:</p> $P(A_3 B) = \frac{P(A_3) \cdot P(B A_3)}{P(A_1) \cdot P(B A_1) + P(A_2) \cdot P(B A_2) + P(A_3) \cdot P(B A_3)}$ $P(A_3 B) = \frac{0,5 \cdot 0,03}{0,2 \cdot 0,01 + 0,3 \cdot 0,06 + 0,5 \cdot 0,03}$ $P(A_3 B) = \frac{0,015}{0,002 + 0,018 + 0,015} = \frac{0,015}{0,035} = 0,428 = 42,8\%$



Ampliando o foco

Para colocar em prática o conteúdo visto nesta unidade, visite a Biblioteca Virtual, escolha um livro de Estatística, selecione e faça alguns exercícios do livro para reforçar o conhecimento.

Confira as respostas que estão no final do livro e, se tiver alguma dúvida, fale com o seu tutor pelo Fórum de Dúvidas – Fale com o Tutor.



MIDiateca

Para ampliar o seu conhecimento veja o material complementar da Unidade 2, disponível na midiateca.



NA PRÁTICA

Foi realizada uma pesquisa de mercado com 500 consumidores com o objetivo de conhecer a aceitação de um novo produto — chamado Sampler — lançado no mercado.

Resultado da pesquisa de satisfação com 500 consumidores do produto Sampler

Sexo	Aprovou	Não aprovou	Total
Masculino	140	100	240
Feminino	200	60	260
Total	340	160	500

Fonte: Dados fictícios.

Você precisa fazer um relatório para enviar ao CEO da empresa em que trabalha dando a ele feedback sobre a aceitação do produto do mercado. Pense em como o conteúdo desta unidade pode lhe auxiliar na elaboração do relatório.

É interessante apresentar no relatório algumas probabilidades calculadas a partir dos dados pesquisados. Por exemplo:

Se uma pessoa for selecionada ao acaso qual a probabilidade de:

- a) Ser do sexo masculino?
- b) Ser do sexo feminino?
- c) Não tenha aprovado o produto.

- d) Ser do sexo feminino ou ter aprovado o produto.
 e) Sabendo que a pessoa selecionada aprovou o produto, qual a probabilidade de ser do sexo masculino?
 f) Dado que a pessoa selecionada seja do sexo feminino, qual a probabilidade de que ela não tenha aprovado o produto?

Vamos considerar:

A = a pessoa selecionada é do sexo masculino.

\bar{A} = a pessoa selecionada não é do sexo masculino, ou seja, é do sexo feminino.

B = a pessoa selecionada aprovou o produto lançado.

\bar{B} = a pessoa selecionada não aprovou o produto lançado.

$$a) P(A) = \frac{240}{500} = 0,48 = 48\%$$

$$b) P(\bar{A}) = 1 - P(A) = 1 - 0,48 = 0,52 = 52\%$$

$$c) P(\bar{B}) = \frac{160}{500} = 0,32 = 32\%$$

$$d) P(\bar{A} \cup B) = P(\bar{A}) + P(B) - P(\bar{A} \cap B) = \frac{260}{500} + \frac{340}{500} - \frac{340}{500} = \frac{400}{500} = 0,8 = 80\%$$

$$e) P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{140}{340} = 0,41 = 41\%$$

$$e) P(\bar{B}|\bar{A}) = \frac{P(\bar{B} \cap \bar{A})}{P(\bar{A})} = \frac{60}{260} = 0,23 = 23\%$$

Resumo da Unidade 2

Nesta unidade vimos os conceitos básicos da Teoria das probabilidades e determinamos situações práticas às quais ela se aplica. Também revisamos as técnicas de contagem e abordamos algumas definições importantes, necessárias para o entendimento do cálculo de probabilidades.

Apresentamos o cálculo das probabilidades considerando cenários diferentes. Estudamos as relações básicas de probabilidades, como probabilidades complementares e probabilidade na união de eventos.

Por fim, abordamos a Probabilidade Condicional e a Probabilidade na Interseção de eventos. Trataremos igualmente de alguns axiomas e teoremas de probabilidade — probabilidade condicional e teorema de Bayes — com a utilização de exemplos práticos e desenvolvidos passo a passo.

Referências

KOKOSKA, S. **Introdução à estatística**: uma abordagem por resolução de problemas. Rio de Janeiro: LTC, 2013.

LARSON, R.; FARBER, B. **Estatística aplicada**. São Paulo: Pearson, 2010.

MEGA-SENA: a loteria que paga milhões para o acertador. **CEF – Loterias**. Disponível em: <http://loterias.caixa.gov.br/wps/portal/loterias/landing/megasena>. Acesso em: 18 ago. 2020.

O PROBLEMA de Monty Hall. **YouTube**. Disponível em: < https://www.youtube.com/watch?v=6_mMLdAzUCE. Acesso em: 12 ago. 2020.

PROBABILIDADES Teóricas e Experimentais. **Khan Academy**. Disponível em: < <https://pt.khanacademy.org/math/probability/probability-geometry/probability-basics/v/comparing-theoretical-to-experimental-probabilites>. Acesso em: 12 ago. 2020.

SPIEGEL, M. R.; STEPHENS, L. J. **Estatística**. 4. ed. Porto Alegre: Bookman, 2009

SWEENEY, D. J.; WILLIAMS, T. A.; ANDERSON, D. R. **Estatística aplicada à administração e economia**. 3. ed. São Paulo: Cengage Learning, 2013.

UNIDADE 3

Distribuições de Probabilidade

INTRODUÇÃO

Nesta unidade continuaremos o estudo da teoria das probabilidades, apresentando os conceitos de variáveis aleatórias e de distribuições de probabilidades.

As distribuições teóricas de probabilidades são aquelas que foram submetidas a estudos prévios e têm propriedades conhecidas. Por isso, nos ajudam na modelagem matemática de fenômenos aleatórios. Sendo assim, é mais uma ferramenta poderosa e muito utilizada para os cálculos de probabilidades.

O que veremos nos tópicos, a saber:

- **Tópico 1:** variáveis aleatórias, discretas e contínuas e a função de distribuição de probabilidades.
- **Tópico 2:** algumas distribuições discretas de probabilidades — a distribuição binomial e a distribuição de Poisson.
- **Tópico 3:** algumas distribuições contínuas de probabilidades — a distribuição Normal e a distribuição Normal padronizada.



OBJETIVO

Nesta unidade você será capaz de:

- Utilizar as distribuições de probabilidades em situações inerentes à inferência estatística.

Variáveis aleatórias e distribuições de probabilidade

Ao determinarmos os elementos de um espaço amostral eles podem ter uma natureza **quantitativa** ou **qualitativa**, como podemos observar no exemplo a seguir.



Exemplo

Notas dos alunos de Estatística $\rightarrow \Omega = [0 ; 10] \rightarrow$ **quantitativa**

Sexo dos alunos de Estatística $\rightarrow \Omega = \{\text{feminino, masculino}\} \rightarrow$ **qualitativa**

Quando trabalharmos com variáveis aleatórias qualitativas ficamos impossibilitados de operar matematicamente em alguns casos, porque seus valores não são numéricos. As variáveis aleatórias vêm auxiliar-nos, principalmente quando necessitamos atribuir um valor **numérico** para todos os elementos do espaço amostral, mesmo que esses elementos sejam **qualitativos**.

Variável aleatória

Uma variável aleatória é uma função que associa a cada elemento do espaço amostral um único número real. Normalmente escolhemos letras maiúsculas do nosso alfabeto para representá-las, sendo as mais usuais o “X” e o “Y”.

Apesar da terminologia “variável aleatória”, ela é uma **função** cujo domínio sempre é formado pelos elementos do espaço amostral e o contradomínio o conjunto dos números reais. Tais funções são chamadas aleatórias, porque seus **valores não podem ser determinados com certeza antes que o experimento seja realizado**.

Para entendermos melhor a aplicabilidade deste conceito vejamos o exemplo a seguir.



Exemplo

A produção de aparelhos celulares de uma determinada fábrica pode ter duas classificações quanto à qualidade: Defeituoso ou Perfeito, ou seja, temos um experimento em que os elementos do espaço amostral não são números e sim atributos, que vamos representar como “D”=defeituoso e “P”=perfeito.

Exemplo

Suponha que três aparelhos sejam retirados aleatoriamente para inspeção. Seja a variável aleatória “X” definida como o **número** de aparelhos danificados.

Então:

1) O experimento consiste em observar as condições da qualidade de três aparelhos celulares.

Espaço amostral: {PPP, PPD, PDP, PDD, DPP, DPD, DDP, DDD}, lembrando que “D” = defeituoso e “P” = perfeito.

2) A variável aleatória “X” representa o número de aparelhos danificados, podendo assumir os seguintes valores, como mostra a tabela a seguir.

Resultado	X=nº de aparelhos danificados
PPP	0
PPD	1
PDP	1
PDD	2
DPP	1
DPD	2
DDP	2
DDD	3

3) O importante é que agora não estamos interessados na ordem e sim na quantidade.

Uma variável aleatória pode ser classificada como **discreta** ou **contínua**, dependendo do número de possíveis valores que a variável pode assumir.

Variável aleatória discreta

A variável aleatória é discreta quando o conjunto de todos os valores que ela pode assumir são associadas aos números naturais, 0, 1, 2, ... Normalmente, são **oriundas de contagens**.



Exemplo

Experimento	Variável aleatória	Valores possíveis
Inspecionar uma remessa de 5 celulares.	$X = \text{n}^\circ$ de celulares defeituosos.	0,1,2,3,4,5 (finito)
Classificar os 160 alunos de Estatística quanto ao sexo.	$X = \text{n}^\circ$ de alunos do sexo feminino.	0,1,2,...,160 (finito)
Observar o fluxo de carros que passam no pedágio, num intervalo de 1 hora.	$X = \text{n}^\circ$ de carros que passam no pedágio, num intervalo de 1 hora.	0,1,2,3,... (infinito numerável)

Variável aleatória contínua

A variável aleatória é contínua quando os conjuntos de todos os valores que ela pode assumir pertencem a um intervalo real ou a um conjunto de intervalos não numeráveis. Normalmente, são **oriundas de medições**.



Exemplo

Experimento	Variável aleatória	Valores possíveis
Observar o tempo de espera na fila do banco.	X =tempo, em minutos, entre a chegada e a saída do cliente na agência bancária.	$X \geq 0$
Altura dos alunos de Estatística.	X = altura, em metros, dos alunos de Estatística.	$0 \leq X \leq 3$
Pesar um carregamento de um determinado produto.	X = número de quilos do carregamento.	$X \geq 0$
Construção de um apartamento.	X = Porcentagem de conclusão do projeto depois de 1 ano.	$0 \leq X \leq 100$

Função ou Distribuição de Probabilidade

Seja " X " uma variável aleatória **discreta**, sejam $x_1, x_2, x_3, \dots, x_n$ os possíveis valores que a variável pode assumir.

Associamos a cada resultado " x " a probabilidade de ocorrência de tal forma que:

- a) $P(x_i) \geq 0$, ou seja, a probabilidade de ocorrência de resultado x_i é maior ou igual a zero.
- b) $\sum_{i=1}^n P(x_i)=1$, ou seja, a soma das probabilidades de todos os valores possíveis de " X " é igual a 1.

A Função de Probabilidade pode ser apresentada na forma de uma lista, de uma tabela, de um gráfico ou ainda de uma função matemática.

Para entendermos melhor a explicação, vamos retomar o exemplo anterior usado na variável aleatória.

Exemplo

A produção de aparelhos celulares de uma determinada fábrica pode ter duas classificações quanto à qualidade: Defeituoso ou Perfeito, ou seja, temos um experimento em que os elementos do espaço amostral não são números e sim atributos, que vamos representar como “D”=defeituoso e “P”=perfeito.

Suponha que três aparelhos sejam retirados aleatoriamente para inspeção. Seja a variável aleatória “X” definida como o **número** de aparelhos danificados.

Determine a função de probabilidade para a variável aleatória X.

Então:

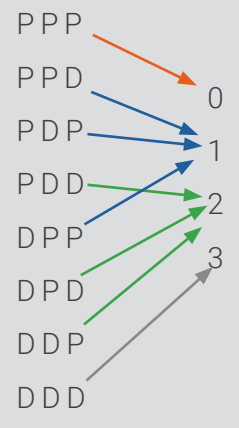
1) O experimento consiste em observar as condições da qualidade de três aparelhos celulares.

Espaço Amostral: {PPP, PPD, PDP, PDD, DPP, DPD, DDP, DDD}, sendo “D”=defeituoso e “P”=perfeito

2) A variável aleatória “X” representa o número de aparelhos danificados, podendo assumir os seguintes valores: 0,1,2 ou 3

3) Distribuição de probabilidade: associação, a cada resultado “x” a probabilidade de ocorrência.

Devemos, então, calcular a probabilidade de ocorrência de cada valor que é variável aleatória:



Número de elementos do espaço amostral $\rightarrow n(\Omega) = 8$

Se $X=0 \rightarrow 1$ ocorrência

Se $X=1 \rightarrow 3$ ocorrências

Se $X=2 \rightarrow 3$ ocorrências

Se $X=3 \rightarrow 1$ ocorrência

Calculando as probabilidades, temos:

$$P(X = 0) = \frac{1}{8}$$

$$P(X = 1) = \frac{3}{8}$$

$$P(X = 2) = \frac{3}{8}$$

$$P(X = 3) = \frac{1}{8}$$

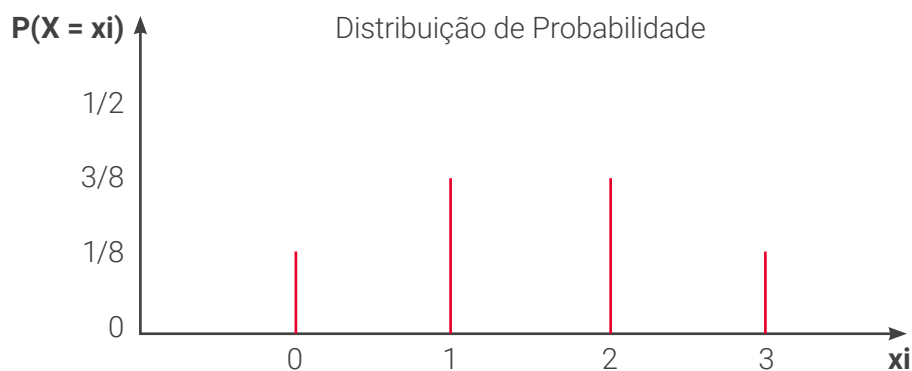
Fonte: Elaborada pela autora (2020).

Distribuição de Probabilidade, expressa por uma tabela:

Valores de "X"	0	1	2	3
Probabilidade de ocorrência $p(X=x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Fonte: Elaborada pela autora (2020).

Distribuição de Probabilidade expressa por um gráfico:



Fonte: Elaborado pela autora (2020).

Esperança Matemática ou Média: $E(X)$

A Esperança Matemática, também chamada de média, é o valor em torno do qual estão concentrados os resultados da variável aleatória em uma amostra suficientemente grande.

Seu cálculo é obtido pela fórmula:

$$E(X) = \mu = \sum x_i \cdot P(x_i)$$

Onde:

x_i = *valor da variável aleatória*

$P(x_i)$ = *respectiva probabilidade*



Ampliando o foco

Em Estatística Descritiva, o valor em torno do qual os dados estão concentrados são chamados de média e os representamos pelo símbolo \bar{X} .

Em Estatística Indutiva, o valor em torno do qual os resultados da variável aleatória estão concentrados é chamado de Esperança Matemática e os representamos pelo símbolo μ .

Variância (σ^2)

Avalia o grau de dispersão dos valores da variável aleatória em torno da média. Seu cálculo é obtido pela fórmula:

$$\sigma^2 = \sum (x_i - \mu)^2 \cdot p(x_i)$$

Em que:

x_i = valor da variável aleatória

μ = Valor esperado

$P(x_i)$ = respectiva probabilidade

Desvio-padrão (σ)

É a raiz quadrada da variância.

$$\sigma = \sqrt{\sigma^2}$$



Ampliando o foco

Em Estatística Descritiva, a variância é representada por s^2 e o desvio-padrão por "s".

Em Estatística Indutiva, a variância é representada por σ^2 e o desvio-padrão por σ .

Coeficiente de Variação (CV)

É o quociente entre o desvio-padrão e a esperança matemática. Normalmente, é expresso em %.

$$CV = \frac{\sigma}{E(X)} \cdot 100$$

Se:

$CV < 10\%$	Baixa dispersão.
$10\% \leq CV < 20\%$	Média dispersão.
$20\% \leq CV < 30\%$	Alta.
$CV \geq 30\%$	Muito alta (pequena representatividade da média).

Agora, vejamos um exemplo para entendermos a aplicação dos conceitos abordados na prática.

Exemplo

O número de chamadas telefônicas recebidas por minuto em uma empresa de telemarketing e suas respectivas probabilidades está representado na tabela a seguir:

Número de chamadas	0	1	2	3	4	5
Probabilidades	0,18	0,39	0,24	0,14	0,04	0,01

a) Qual o número esperado de chamadas recebidas em um minuto?

$$E(X) = 0 \cdot 0,18 + 1 \cdot 0,39 + 2 \cdot 0,24 + 3 \cdot 0,14 + 4 \cdot 0,04 + 5 \cdot 0,01$$

$$E(X) = 0 + 0,39 + 0,48 + 0,42 + 0,16 + 0,05 = 1,5$$

Espera-se, aproximadamente, 1,5 chamadas por minuto.

b) A média é uma medida confiável para essa distribuição?

Para responder essa questão devemos verificar o coeficiente de variação, pois, se ele for maior do que 30%, há uma alta dispersão, então não devemos confiar na média, já que ela é afetada por valores discrepantes.

- Calculando a variância:

$$\sigma^2 = \sum (x_i - \mu)^2 \cdot p(x_i)$$

Número de chamadas (xi)	0	1	2	3	4	5	Σ
Probabilidades P(xi)	0,18	0,39	0,24	0,14	0,04	0,01	
$(x_i - \mu)^2$	2,25	0,25	0,25	2,25	6,25	12,25	
$(x_i - \mu)^2 \cdot p(x_i)$	0,405	0,0975	0,06	0,315	0,25	0,1225	1,25

Como podemos observar na tabela, $(x_i - \mu)^2 \cdot p(x_i) = 1,25$, logo:

$$\sigma^2 = \sum (x_i - \mu)^2 \cdot p(x_i)$$

então, $\sigma^2 = 1,25$

- Calculando o desvio-padrão:

$$\sigma = \sqrt{\sigma^2}$$

$$\sigma = \sqrt{1,25} \cong 1,118$$

- Calculando o Coeficiente de Variação:

$$CV = \frac{\sigma}{E(X)} \cdot 100$$

$$CV = \frac{1,118}{1,5} \cdot 100 = 74,53\%$$

Como o coeficiente de variação é muito elevado, pois é maior do que 30%, os dados estão muito dispersos, então a média não é uma boa medida representativa para essa distribuição.

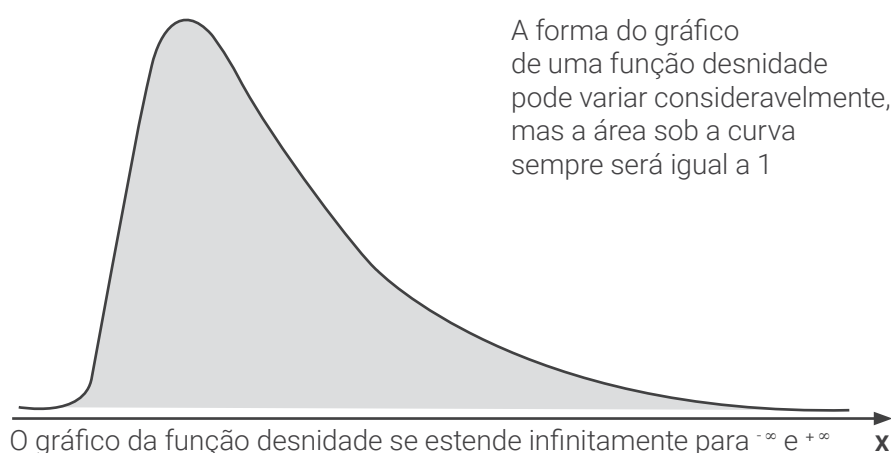
Função Densidade de Probabilidade (ou simplesmente Função Densidade)

Seja "X" uma variável aleatória contínua pertencente a um intervalo numérico.

Define-se a função densidade, a função cuja área sob a curva entre dois valores "a" e "b" nos revela a probabilidade de que a variável aleatória X assuma um valor entre "a" e "b" ($a < b$), de tal forma que:

\mathbb{R}

- a) $f(x) \geq \forall x \in \mathbb{R}$, ou seja, a função assume apenas valores positivos ou nulo, assim o gráfico fica sobre o eixo-x, ou acima dele.
- b) $\int_{-\infty}^{+\infty} f(x) dx = 1$, ou seja, a área total sob a curva de densidade é sempre igual a 1
- c) $\int_{-\infty}^{+\infty} f(x) dx = P(a \leq X \leq b) \forall a < b$, ou seja, a probabilidade de ocorrência de um evento no intervalo $[a, b]$ é a área entre "a" e "b".

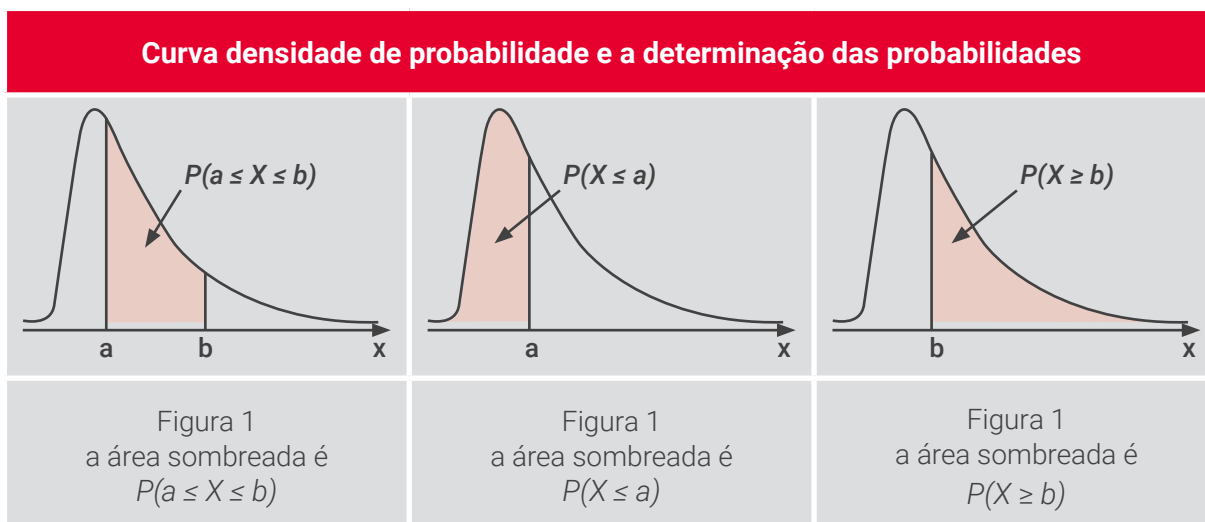


Fonte: Elaborado pela autora (2020).

No caso das variáveis aleatórias contínuas, não estaremos mais interessados em saber a probabilidade de um determinado resultado, como fazemos com as variáveis aleatórias discretas, mas sim em determinar a probabilidade de o resultado estar em um determinado intervalo.

Como não há probabilidade associada a um ponto único, é indiferente o uso dos símbolos $<$ ou \leq para menor que ou $>$ ou \geq para maior que, ou seja:

$$\int_{-\infty}^{+\infty} f(x) dx = P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$$



Fonte: Elaborado pela autora (2020).

Então, como vimos, para encontrar a probabilidade precisamos determinar a área sob uma curva e isso pode ser feito por integral, geometria, tabelas e o uso de tecnologias computacionais.

Medidas estatísticas

As medidas estatísticas Esperança Matemática, também chamada de Valor Esperado ou Média e variância, são definidas de forma semelhante ao que estudamos para as variáveis aleatórias discretas, porém substituímos o somatório por integral.

Esperança Matemática, Valor Esperado ou Média: $E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)dx$

Variância: $\sigma^2 = \int_{-\infty}^{+\infty} (x - E(X))^2 \cdot f(x)dx$



Ampliando o foco praticando

Para colocar em prática o conteúdo visto nesta unidade, visite a Biblioteca Virtual, escolha um livro de Estatística, selecione e faça alguns exercícios do livro para reforçar o conhecimento.

Confira as respostas que estão no final do livro e, se tiver alguma dúvida, fale com o seu tutor pelo Fórum de Dúvidas – Fale com o Tutor.

Distribuições de probabilidades discretas

No Tópico 1 apresentamos a definição geral das distribuições de probabilidade. Porém, existem variáveis aleatórias que apresentam certos padrões de comportamento. Para essas variáveis foram desenvolvidos modelos de distribuições específicos, com a finalidade de auxiliar no cálculo das probabilidades, bem como a esperança matemática e a variância.

Para determinarmos as **probabilidades**, a **esperança** e a **variância**, basta identificarmos qual é a **distribuição teórica que a variável aleatória pertence** e usar as fórmulas próprias da distribuição identificada.

Vamos iniciar o estudo pelos **modelos de distribuições discretas**, apresentando dois modelos entre vários existentes e que são muito utilizados: a **distribuição binomial** e a **distribuição de Poisson**.

Modelo de Distribuição Binomial

A distribuição de probabilidade binomial é uma distribuição discreta de probabilidade que tem inúmeras aplicações no mundo real. A principal característica desse modelo é que **conhecemos o número finito de repetições** do experimento.

Condições para modelagem:

1. O experimento consiste em um número finito de tentativas idênticas.
2. Cada tentativa pode resultar apenas um de dois resultados possíveis: o “sucesso” (se acontece o evento) ou um “fracasso” (se o evento não acontece).
3. A probabilidade de sucesso de cada prova é constante e igual a “p” e a probabilidade do fracasso “q=1-p”.
4. Os resultados das provas são independentes.

Exemplos:

Experimento	Variável aleatória X	Resultados possíveis
Lançar uma moeda seis vezes e observar a face superior.	X = sair a face cara	Sucesso → a face observada é cara Fracasso → a face observada é coroa

Observar o sexo de 100 partos de uma maternidade.	X= a criança ser do sexo feminino	Sucesso → a criança é menina Fracasso → a criança é menino
Perfurar uma série de poços de petróleo.	X= não encontrar petróleo	Sucesso → não se encontra petróleo Fracasso → encontra-se petróleo
Responder aleatoriamente 10 questões de uma prova objetiva.	X=acertar a resposta correta	Sucesso → o aluno acerta a questão Fracasso → o aluno erra a questão
Responder aleatoriamente 10 questões de uma prova objetiva.	X=errar a resposta	Sucesso → o aluno erra a questão Fracasso → o aluno acerta a questão
Observar 10 animais que foram submetidos a uma injeção com efeitos cancerígenos.	X= apresentar a doença	Sucesso → o animal apresenta a doença Fracasso → o animal não apresenta a doença

Parâmetros de uma distribuição binomial: $X \sim B(n,p)$

São valores definidores de uma distribuição binomial. Uma variável aleatória binomial fica completamente determinada quando conhecemos o número de tentativas e a probabilidade de um sucesso.

Em que:

n = número total de ensaios.

p = probabilidade de sucesso em cada ensaio.

Fórmulas da Distribuição Binomial

Se soubermos os parâmetros " n " e " p " seremos capazes de responder qualquer pergunta sobre a probabilidade de X , o valor esperado e a variância. Para isso, basta aplicarmos as fórmulas da distribuição binomial, que são:

Cálculo das Probabilidades	Cálculo do Valor Esperado	Cálculo da Variância
$P(X=x) = \binom{n}{x} \cdot p^x \cdot q^{(n-x)}$	$\mu = n \cdot p$	$\sigma^2 = n \cdot p \cdot q$
<p>Em que:</p> <p>X= número de eventos discretos em um número finito de repetições.</p> <p>n = número de repetições do experimento.</p> <p>x = número de sucessos.</p> <p>p = a probabilidade de sucesso em cada prova.</p> <p>q = a probabilidade de fracasso em cada prova.</p> <p>$\binom{n}{x} = \frac{n!}{x! (n-x)!}$ = número de combinações de “n” elementos tomados x a x.</p>		

Vejamos um exemplo prático.

Exemplo 1

Um professor de Estatística elaborou uma avaliação de múltipla escolha, que consiste em 10 questões, cada uma com cinco alternativas para resposta e apenas uma dessas alternativas tem a resposta correta. Suponha que os estudantes que irão a fazer a prova não frequentaram as aulas e não estudaram para a avaliação. Sendo assim, eles irão marcar a resposta aleatoriamente, o que informalmente chamamos de “chute”. Para ser aprovado, o aluno deve acertar no mínimo sete questões. Um estudante é selecionado ao acaso.

- Qual é a probabilidade de ele “chutar” as 10 questões e acertar 7?
- Qual é a probabilidade de ele ser aprovado?
- O valor esperado de acertos nessas condições.
- O que podemos concluir de um aluno que não estuda e “chuta” as questões da avaliação.

Resolução:

1º) Definir a variável aleatória, pois será essa definição que vai determinar quem terá o sucesso e o fracasso.

Seja a variável aleatória X =número de acertos na avaliação.

Logo: Sucesso \rightarrow acertar a questão

Fracasso \rightarrow errar a questão

2º) Verificar as hipóteses a serem atendidas para o uso das fórmulas da distribuição binomial. São elas:

1 - Existe um número finito de repetições?

Sim \rightarrow o aluno “chuta” 10 vezes, logo, $n=10$

2 - Só há 2 possíveis resultados para cada tentativa?

Sim \rightarrow acerto ou erro

3- Cada “chute” é independente?

Sim, pois um “chute” não influencia no resultado do outro.

4 - A probabilidade de sucesso em cada tentativa é constante?

Sim, em cada questão a probabilidade de acerto é igual a:

$$p = \frac{\text{número de opções corretas em cada questão}}{\text{número de opções em cada questão}} = \frac{1}{5} = 0,2$$

Logo, a probabilidade do fracasso é $q = 1 - p = 1 - 0,2 = 0,8$

Como todas as hipóteses foram verificadas, então: $X \sim B(10; 0,2)$, ou seja, a variável aleatória X obedece a uma distribuição binomial, com parâmetros $n=10$ e $p=0,2$.

Com isso, podemos fazer o cálculo das probabilidades pedidas, usando as fórmulas da distribuição binomial, como mostramos a seguir.

a) Um estudante é selecionado ao acaso. Qual é a probabilidade de ele “chutar” as 10 questões e acertar 7?

$$P(X=x_i) = \binom{n}{x_i} \cdot p^x \cdot q^{(n-x)}$$

$$P(X=7) = \binom{10}{7} \cdot (0,2)^7 \cdot (0,8)^{(10-7)}$$

$$P(X=7) = \frac{10!}{7! 3!} \cdot (0,2)^7 \cdot (0,8)^{(10-7)}$$

$$P(X=7) = \frac{10!}{7! (10-x)!} \cdot (0,2)^7 \cdot (0,8)^3$$

$$P(X=7) = 120 \cdot 0,0000128 \cdot 0,512 = 0,000078643 \cong 0,0786\%$$

b) Um estudante é selecionado ao acaso. Qual é a probabilidade de ele ser aprovado? Para ser aprovado, o aluno tem que acertar **pelo menos** sete questões. Logo:

$$P(X \geq 7) = P(X=7) + P(X=8) + P(X=9) + P(X=10)$$

Analogamente ao cálculo de $P(X=7)$, vamos fazer $P(X=8)$, $P(X=9)$ e $P(X=10)$, obtendo como resultado:

$$P(X \geq 7) = 0,00078643 + 0,00007373 + 0,00000410 + 0,00000010 = 0,00086436 = 0,086436\%$$

c) $E(x) = n \cdot p = 10 \cdot 0,2 = 2$ questões

Espera-se que um aluno que “chuta” 10 questões em uma avaliação, acerta 2.

Logo, com base nas respostas anteriores podemos concluir que vai ser muito difícil um estudante ser aprovado em Estatística sem assistir às aulas e sem estudar para avaliações.

Modelo de Distribuição de Poisson

Esse modelo será utilizado para modelar experimentos discretos em que o número de repetições não é conhecido.

A variável aleatória de Poisson é definida como o número de sucessos em certo intervalo fixo e contínuo, que pode ser de tempo, comprimento, área, volume, massa e outros, sendo essa a principal característica desse modelo.

Exemplos:

1. Número de carros que passam em um pedágio em uma hora.
2. Número de vezes em que o corpo de bombeiros é chamado por dia para combater incêndios em uma cidade.
3. Número de acidentes de trabalho por semana em uma empresa industrial.

4. Número de reparos em uma rodovia em um trecho de 100 km.
5. Número de automóveis que chegam ao campus entre 7 e 10 horas da manhã.
6. Número de microrganismos por cm³ de água contaminada.
7. Número de vazamentos em 100 Km de tubulação.

Condições para modelagem

1. Acertos individuais ocorrerão aleatoriamente e independentemente dentro um intervalo fixo de observação.
2. O número de ocorrências de certo evento de interesse neste intervalo é uma variável discreta com valores possíveis 0, 1, 2, 3... . Observe que x não possui limite máximo.
3. Uma distribuição de Poisson modela bem eventos “raros”. Fenômenos raros são aqueles que ocorrem com pouca frequência para qualquer intervalo de observação. Por exemplo, o número de automóveis da linha Gol que passam por um radar na estrada em um intervalo de uma hora e o número de automóveis Porsche Cayenne que passam nesse mesmo radar, nesse mesmo intervalo de uma hora. Certamente, o modelo de Poisson deverá ser mais adequado para o automóvel Porsche.

Parâmetros de uma distribuição de Poisson: $X \sim P(\mu)$

A média μ é o único parâmetro para a distribuição de Poisson.

$$\mu = \lambda \cdot t$$

Em que:

μ = média de eventos discretos em “t” unidades de medida.

λ = taxa de frequência ou coeficiente de proporcionalidade por unidade de medida.

t = unidade de medida.

Fórmulas da Distribuição de Poisson

Se soubermos a taxa de frequência por unidade de medida (λ) e a unidade de medida (t), seremos capazes de responder qualquer pergunta sobre a probabilidade de X, o valor esperado e a variância. Basta para isso aplicarmos as fórmulas da distribuição de Poisson, que são:

Cálculo das Probabilidades	Cálculo do Valor Esperado	Cálculo da Variância
$P(X=x_i) = \frac{e^{-\mu} \cdot \mu^x}{x!}$	$\mu = \lambda \cdot t$	$\sigma^2 = \lambda \cdot t$

Em que:

X : número de eventos discretos em t unidades de medida.

λ : taxa de frequência ou coeficiente de proporcionalidade por unidade de medida.

t : unidade de medida.

$\mu = \lambda t$: média de eventos discretos em t unidades de medida.

e = constante de Euler $\cong 2,71828$.

Exemplo 1

Intervalo de tempo

A administradora de pedágios de uma rodovia está interessada em conhecer o número de carros que chegam a uma determinada cabine em um intervalo de tempo de 15 minutos. Sabe-se que:

- Em qualquer período de igual duração a probabilidade de chegar um carro no pedágio é a mesma.
- O evento chegada de um carro na cabine do pedágio é independente da chegada de outro carro em qualquer período.
- Observando dados históricos da rodovia, podemos estimar o número médio de carros que chegam em uma cabine desse pedágio como 10 carros no período de 15 minutos.

Pede-se:

- a) Determine a probabilidade de chegarem exatamente 5 carros nessa cabine em um período de 15 minutos.
- b) Determine a probabilidade de chegar exatamente 1 carro nessa cabine em um período de 3 minutos.

Resolução:

Letra a)

Podemos verificar no enunciado que os pressupostos para a modelagem por Poisson foram satisfeitos. Logo, usaremos esse modelo para responder às questões propostas.

1) Definição da variável aleatória X :

Seja X = número de carros que chegam em uma cabine de pedágio de uma rodovia em um intervalo de tempo de 15 minutos.

Nesse caso, buscamos determinar a probabilidade que o número de carros que

chega em uma cabine de um pedágio de uma rodovia, em um intervalo qualquer de tempo de 15 minutos, seja exatamente igual a 5, ou seja, $X=5$

2) Intervalo de tempo: $t = 15$ minutos.

3) Frequência nesse intervalo de tempo: $\lambda = 10$ carros a cada 15 minutos $= \frac{10}{15}$

4) Média: $\mu = \lambda \cdot t = \frac{10}{15} \cdot 15 = 10$ carros

Então,

$X \sim P(10)$, ou seja, a variável aleatória X obedece à distribuição de Poisson, com média igual a 10.

Estando com o parâmetro definido, vamos substituí-lo no modelo de Poisson:

$$P(X=5) = \frac{e^{-10} \cdot 10^5}{5!} = \frac{4,539992976}{120} = 0,0378 = 3,78\%$$

Letra b)

Observamos que o intervalo de tempo agora modificou-se para três minutos. Logo:

1) Definição da variável aleatória X :

Seja X = número de carros que chegam em uma cabine de pedágio de uma rodovia em um intervalo de tempo qualquer de três minutos.

Nesse caso, buscamos determinar a probabilidade que o número de carros que chega em uma cabine de pedágio de uma rodovia, em um intervalo qualquer de tempo de três minutos seja exatamente igual a 1, ou seja, $X=1$.

2) Intervalo de tempo: $t = 3$ minutos.

3) Frequência nesse intervalo de tempo: $\lambda = 10$ carros a cada 15 minutos $= \frac{10}{15}$

4) Média: $\mu = \lambda \cdot t = \frac{10}{15} \cdot 3 = 2$ carros

5) Então: **$X \sim P(2)$**

Então,

$X \sim P(2)$, ou seja, a variável aleatória X obedece à distribuição de Poisson, com média igual a 2. Estando com o parâmetro definido, vamos substituí-lo no modelo de Poisson:

$$P(X=x_i) = \frac{e^{-\mu} \cdot \mu^x}{x!}$$

$$P(X=1) = \frac{e^{-2} \cdot 2^1}{1!} = \frac{0,2707}{1} = 0,2797 = 27,07\%$$

Portanto, a probabilidade de que exatamente cinco carros cheguem em um período de 15 minutos nessa cabine é aproximadamente igual a 3,78%, enquanto a probabilidade de que exatamente um carro chegue em um período de três minutos nessa cabine é igual a 27,07%.

Exemplo 2

Intervalo de comprimento

Suponhamos que, após um mês de obras de recapeamento de um trecho de uma rodovia, a administradora dessa rodovia esteja interessada em verificar o número de defeitos graves na pavimentação desse trecho. Sabe-se que:

- Em quaisquer dois intervalos de mesma extensão nessa rodovia a probabilidade de ocorrência de ocorrer um defeito grave é a mesma.
- Os eventos defeitos graves em uma rodovia são independentes, ou seja, a ocorrência de um defeito grave em um intervalo não tem nenhuma dependência com a existência ou não de outro defeito em outro intervalo qualquer.
- A quantidade de defeitos graves por Km pertence ao intervalo $[0, +\infty[$.
- Sabe-se que defeitos graves ocorrem um mês após as obras de recapeamento à taxa de dois defeitos por quilômetro.

Pede-se:

- a) A probabilidade de não haver nenhum defeito grave em um trecho de 3 km de rodovia.
- b) A probabilidade de existir pelo menos um defeito grave em um trecho de 3 km de rodovia.

Resolução:

Letra a)

Pelo enunciado verificamos que as hipóteses para a modelagem com a Distribuição de Poisson estão satisfeitas.

Primeiramente, vamos definir a variável aleatória de Poisson X .

X = número de defeitos graves em uma rodovia que aconteceram um mês após o recapeamento em um trecho de 3 km.

Podemos notar que estamos trabalhando em um intervalo de comprimento de 3km.

$t = 3 \text{ km}$

Agora, vamos determinar a taxa de frequência nesse intervalo de tempo.

$$\lambda = 2 \text{ defeitos a cada km} = \frac{2 \text{ defeitos}}{1 \text{ km}}$$

$$\text{Logo, a média será: } \mu = \lambda \cdot t = \frac{2}{1} \cdot 3 = 6 \text{ defeitos}$$

Para calcular a probabilidade pedida basta usar a fórmula da distribuição de Poisson.

$$P(X=x_i) = \frac{e^{-\mu} \cdot \mu^x}{x!}$$

$$P(X=1) = \frac{e^{-6} \cdot 2^0}{0!} = \frac{0,0025 \cdot 1}{1} = 0,0025 = 25\%$$

Logo, existe uma probabilidade bem pequena, em torno de 0,25%, de não se encontrar nenhum defeito grave em um trecho de 3Km de uma rodovia, um mês após o recapeamento.

Letra B)

Qual a probabilidade de existir pelo menos um defeito grave em um trecho de 3 km de rodovia?

$$P(X \geq 1) = P(X=1) + P(X=2) + P(X=3) + \dots$$

Como não sabemos o número final de defeitos, temos que determinar essa probabilidade utilizando o conceito de probabilidade complementar, ou seja:

$$P(X \geq 1) = 1 - P(X=0)$$

Sabemos que $P(X=0) = 0,0025$, pois calculamos na letra a) desse exemplo.

Logo:

$$P(X \geq 1) = 1 - 0,0025 = 0,9975 = 99,75\%$$

Então, existe uma probabilidade de 99,75% de ocorrer pelo menos um defeito grave em um trecho de 3 km da rodovia.



Ampliando o foco

Para colocar em prática o conteúdo visto nesta unidade, visite a Biblioteca Virtual, escolha um livro de Estatística, selecione e faça alguns exercícios do livro para reforçar o conhecimento.

Confira as respostas que estão no final do livro e, se tiver alguma dúvida, fale com o seu tutor pelo Fórum de Dúvidas – Fale com o Tutor.

Distribuição de probabilidade Normal

Neste tópico apresentaremos duas distribuições contínuas de probabilidades — distribuição Normal e distribuição Normal Padronizada, que nos auxiliam nos cálculos das probabilidades, oferecendo-nos, já tabelado, o cálculo das áreas sob a curva. É fundamental entendermos os conceitos que serão apresentados, pois assim teremos a alternativa de resolver os cálculos de probabilidades sem que seja preciso resolver as integrais que aparecem quando trabalhamos com as variáveis aleatórias contínuas.

Distribuição de probabilidade Normal ou Gausseana

A distribuição de probabilidade é:

- Considerada a distribuição de probabilidade mais importante em toda Estatística para descrever uma variável aleatória contínua.
- Usada na modelagem dos dados provenientes de fenômenos naturais, industriais e pesquisas em geral.
- Amplamente usada em inferência estatística.
- Desenvolvida por Moivre em 1733 e posteriormente por Laplace, usada por Gauss na derivação das equações do estudo de erros em repetições de medidas de um mesmo objeto, por isso é também conhecida como gausseana.

Temos algumas propriedades da distribuição Normal de probabilidade.

Propriedades da distribuição Normal

Modelo Matemático

A curva Normal apresenta o seguinte modelo matemático:

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

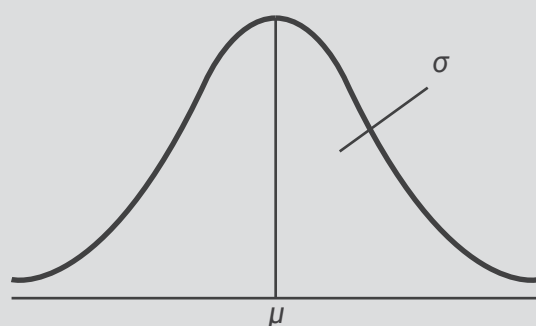
$$x \in (-\infty; +\infty) \text{ e } \sigma^2 > 0$$

Em que:

μ = média
 σ = desvio padrão
 σ^2 = variância
 $e = 2,71828...$
 $\pi = 3,14159...$

Formato da curva

O formato da curva normal assemelha-se à forma de um sino.

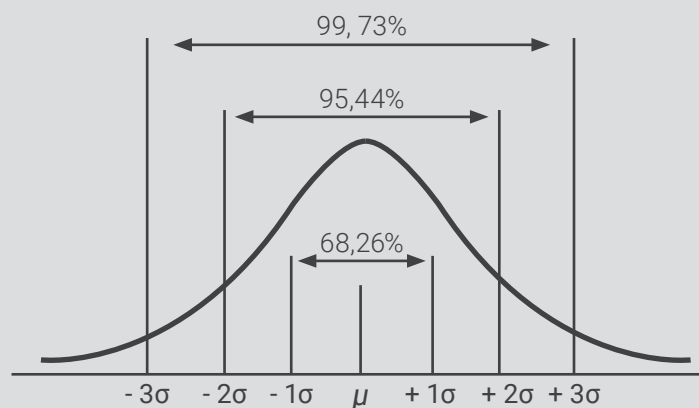


média = mediana = moda

Abrangência em função do desvio-padrão

Cerca de 68,26% da probabilidade total está entre a média e 1 desvio-padrão.
Cerca de 95,44% da probabilidade total está entre a média e 2 desvios-padrão.
Cerca de 99,73% da probabilidade total está entre a média e 3 desvios-padrão.

Observe graficamente:

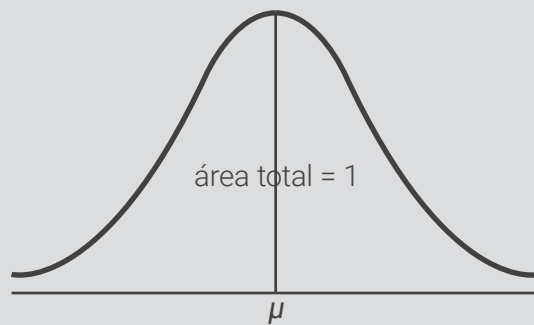


Normal

Área sob a curva

Quando calculamos a área total sob a curva normal sempre encontramos resultado igual a 1.

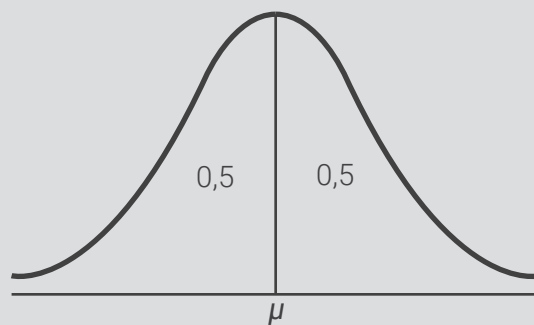
Isso significa 100% quando estamos investigando as probabilidades.



Simetria

A curva é simétrica em relação à média, portanto 50% da probabilidade está antes da média, que chamamos de ramos esquerdo, e 50% depois da média, denominado ramo direito.

Observe graficamente:



Parâmetros

Lembrando que a distribuição fica completamente definida por seus parâmetros e a distribuição Normal tem como parâmetros a média μ e a variância σ^2 .

Notação

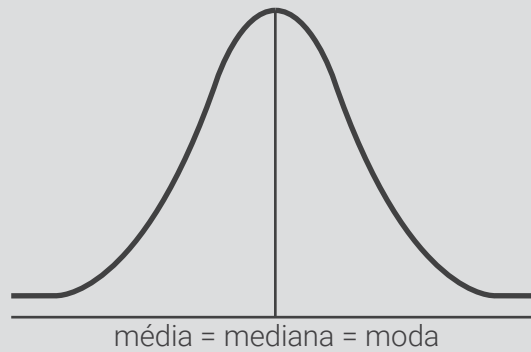
$$X \sim (N \mu \sigma^2)$$

Lê-se: A variável aleatória "X" obedece a uma distribuição Normal, com parâmetros μ e σ^2 , sendo $\mu = \text{média}$ e $\sigma^2 = \text{variância}$.

Simetria

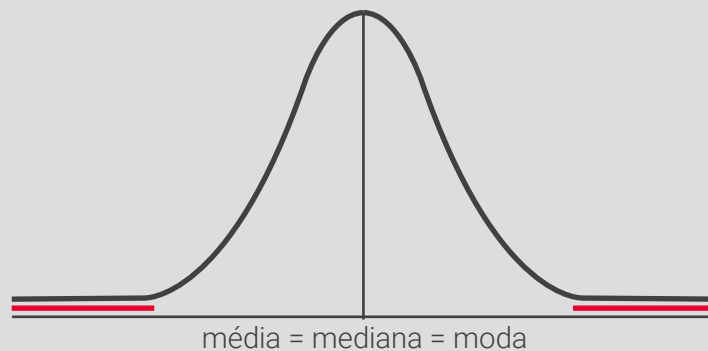
As medidas de posição, média, mediana e moda são coincidentes na curva Normal.

Observe graficamente:



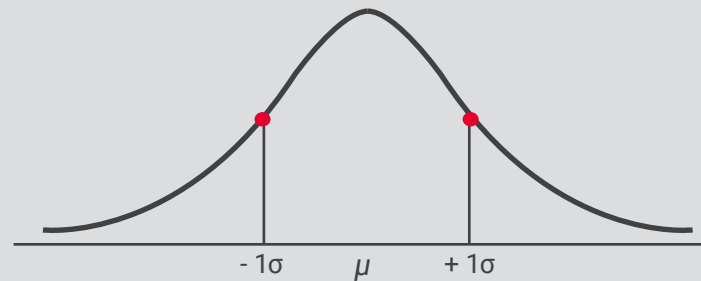
Assíntotas

A curva normal é assintótica em relação ao eixo-x, ou seja, tende a zero para valores tendendo a mais infinito ou menos infinito.



Pontos de Inflexão

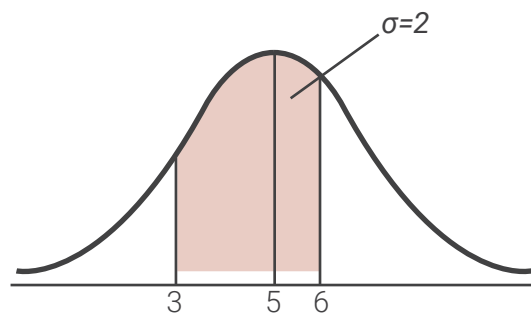
A curva normal apresenta dois pontos de inflexão, invertendo a concavidade quando $x = +1\sigma$ e $x = -1\sigma$.



Como calcular as probabilidades com o modelo de distribuição Normal?

Seja uma variável aleatória normal com média igual a 5 e a variância igual a 4. Suponhamos que desejamos calcular a probabilidade de a variável aleatória estar entre 3 e 6.

Como vimos, para determinar essa probabilidade basta determinarmos o valor da área sob a curva no intervalo considerado.



Fonte: Elaborada pela autora (2020).

Para calcular essa área e se resolvermos a integral: $\int_3^6 \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-5)^2}{2 \cdot 4}} dx$.
E coloquemos como resposta 0.532807.

Logo, $P(3 < X < 6) \cong 53,28\%$

Como o cálculo dessa integral não é nada trivial, já que ela não pode ser resolvida analiticamente, temos outra opção para determinarmos a área sob a curva, que irá fornecer-nos a probabilidade. Vamos, então, fazer uma padronização na curva Normal e determinar a

área sob a curva usando simplesmente uma tabela, em vez de calcularmos essa integral pelo desenvolvimento em série. Para isso, vamos fazer uma transformação linear da variável aleatória Normal para variável aleatória padronizada.

Distribuição Normal Padrão ou Curva Normal Padronizada

Também chamada de **Standardizada** ou **reduzida**, é a distribuição Normal com média igual a zero ($\mu = 0$), variância igual a 1, ($\sigma^2 = 1$), e consequentemente o desvio-padrão igual a 1 ($\sigma = 1$).

Seja X uma variável aleatória com distribuição N (μ, σ^2).

Consideremos a transformação linear de X para Z-score:

$$Z = \frac{X - \mu_i}{\sigma}$$

Com isso, pode-se demonstrar que a média é sempre igual a zero e a variância sempre igual a 1.

$\mu = 0$ e $\sigma^2 = 1$

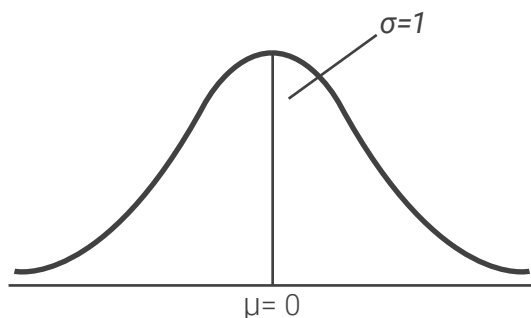
Logo, o modelo matemático fica reduzido a:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}}$$

$Z \in (-\infty + \infty)$

Sendo assim, o formato da curva:

Figura: Curva Normal Padronizada.



Fonte: Elaborada pela autora (2020).

Com o intuito de auxiliar no cálculo das áreas sob a curva, que, como já vimos, são de difícil resolução, usaremos a **distribuição Normal Padrão**. Nessa distribuição as áreas sob a curva já foram calculadas e estão apresentadas em tabelas. Serão essas as tabelas que nos auxiliarão nos cálculos das probabilidades.

Existem vários tipos de tabelas, porém as mais utilizadas são as que oferecem a área para valores de z-score até a média e também que mostram a área cumulativa. Para esse estudo escolhemos utilizar a tabela disponibilizada a seguir.

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1664	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981

2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000
4	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

Fonte: http://ead.uva.br///recurso/DEF/EST/u3_c3_r1/anexo/Tabela-da-Distribuicao-Normal.pdf



Importante

Ao escolher uma tabela precisamos estar atentos ao tipo de probabilidade que ela apresenta.

Existem tabelas que fornecem a área:

- 1- de 0 a k
- 2- de $-\infty$ a k
- 3- de k a $+\infty$

Normalmente, essa informação vem no cabeçalho da tabela.

Os nossos exercícios serão resolvidos com a tabela do tipo 1, conforme podemos ver no cabeçalho:

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k

Como devemos usar a tabela de distribuição Normal padrão?

1º Calcular o valor de Z com a fórmula

$$Z = \frac{X - \mu}{\sigma}$$

arredondando com a precisão de duas casas decimais. Algumas tabelas não apresentam valores negativos para Z, porém como a curva Normal é simétrica em relação à média, o sinal não terá nenhuma relevância e poderá ser desprezado.

- 2º Identificar a parte inteira e a primeira casa decimal, na primeira coluna da tabela.
- 3º Identificar a segunda casa decimal, na primeira linha com valores na tabela.
- 4º Fazer o cruzamento da linha com a coluna e buscar o valor da área sob a curva de zero até z.

Vejamos um exemplo a seguir.

Exemplo

Considere $X \geq 800$, $\mu = 850$, $\sigma = 45$



1 - Calculando o Z-score:

$$Z = \frac{X - \mu}{\sigma} = \frac{800 - 850}{45} = -1,111111$$

pela simetria da curva Normal = 1,111111. Com precisão de duas casas decimais vamos considerar $Z=1,11$

2 - Vamos buscar na primeira coluna da tabela o valor 1,1

Primeira
coluna

z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1664	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015

3 - Na primeira linha com valores na tabela, identificamos a segunda casa decimal, no caso 0,01

Primeira
linha
de valores

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1664	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015

Logo, temos este resultado:

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1664	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015

4 - Fazemos o cruzamento da linha com a coluna e identificamos o valor da área sob a curva de zero até z.

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1664	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015

Logo a área de zero até z=1,11 será a mesma área de zero até Z=-1,11 que é 0,3665

Como calcular a probabilidade de qualquer distribuição Normal?

Para determinar a probabilidade utilizando a distribuição Normal, devemos seguir os seguintes passos:

1º) Determine os valores da média μ e do desvio-padrão σ e nomeie a distribuição Normal: $X \sim N(\mu, \sigma^2)$

2º) Enuncie a probabilidade pedida.

3º) Transforme a variável aleatória Normal para Normal padronizada através da fórmula:

$$Z = \frac{X - \mu}{\sigma}$$

4º) Esboce o gráfico das curvas e destaque a região cuja área será investigada para determinar a probabilidade pedida.



Importante

Este passo é o mais importante, pois com o desenho não precisamos ficar decorando se devemos somar 0,5, diminuir 0,5, ou não fazer nada. Apenas comparamos o desenho da área obtida na tabela e o desenho do que queremos determinar para concluirmos o que temos que fazer.

5º) Utilize a tabela da distribuição Normal padronizada e busque o valor da área.

6º) Finalize com o valor da área sob a curva e o desenho do item 4, faça os ajustes necessários e determine a probabilidade solicitada.

Vejamos alguns exemplos com os passos apresentados. É importante observar que cada exemplo mostra uma região da curva diferente. E, em cada região, os cálculos são diferentes. Por isso, há um número grande de exemplos apresentados.

Essa parte é importante, pois a maioria dos problemas de probabilidades poderão ser resolvidos ou aproximados pela distribuição Normal.

Exemplo 1

Exemplifica o cálculo da probabilidade na qual a variável aleatória Normal padrão (z-score) é **menor ou igual a um determinado valor negativo**.

Enunciado: Em agosto de 2020, um brasileiro passou em média 77 horas conectado à internet navegando em redes sociais. Suponha que os tempos estejam normalmente distribuídos e que o desvio-padrão seja de 20 horas.

Pergunta-se:

Qual é a probabilidade de um brasileiro, escolhido aleatoriamente, ter passado menos de 50 horas conectado à internet em agosto de 2020?

Resolução:

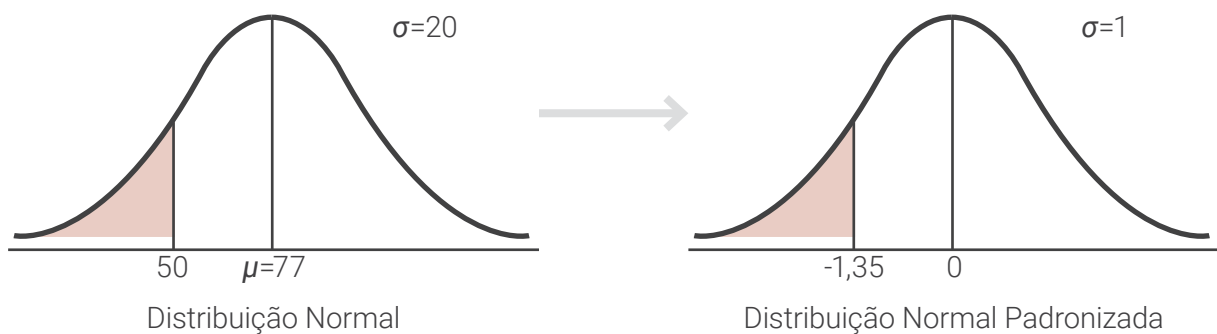
1º) Determine os valores da média μ e do desvio-padrão σ e nomeie a distribuição Normal: neste caso $\rightarrow X \sim N(77, 20^2)$.

2º) Enuncie a probabilidade pedida: $P(X \leq 50) = ?$

3º) Transforme a variável aleatória Normal para Normal padronizada a partir da fórmula:

$$Z = \frac{X - \mu}{\sigma} = \frac{50 - 77}{20} = -1,35$$

4º) Desenhe as curvas e destaque a região cuja área será investigada para determinar a probabilidade pedida.

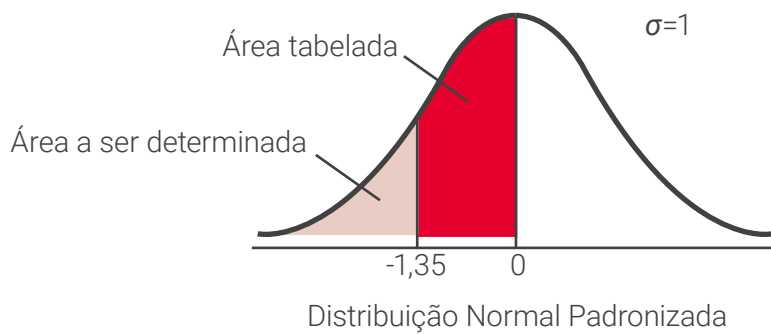


$$P(X \leq 50) = P(Z \leq -1,35)$$

Fonte: Elaborada pela autora (2020).

5º) Utilize a tabela da distribuição Normal padronizada e busque o valor da área.
Para $Z=1,35 \rightarrow A=0,4115$

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1664	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177



Fonte: Elaborada pela autora (2020).

6º) Faça alguns ajustes para chegar à resposta correta, porque a área a ser determinada é diferente da área tabelada. Como vimos, a área total sob a curva Normal é igual a 1 e, pela simetria da curva, 0,5 antes da média e 0,5 depois da média.

Verificamos que:

Área a ser determinada + área tabelada = 0,5

Área a ser determinada = 0,5 – área tabelada

Área a ser determinada = 0,5 – 0,4115 = 0,0885

Logo, a probabilidade pedida será igual à área a ser determinada:

$$P(X \leq 50) = P(Z \leq -1,35) = 0,0885 = 8,85\%$$

Podemos concluir que a probabilidade de que uma pessoa selecionada ao acaso tenha ficado menos de 50 horas nas redes sociais no mês de agosto de 2020 é igual a 8,85%.

Exemplo 2

Exemplifica a probabilidade na qual z-score é **menor ou igual a um determinado valor positivo.**

Enunciado: Suponhamos que o tempo médio que os alunos da Universidade Veiga de Almeida levam para sair de casa e chegar à universidade é igual a 58 minutos. Considere que a distribuição do tempo gasto nesse percurso seja normal, com desvio-padrão igual a 12 minutos. Determine a probabilidade de um aluno selecionado aleatoriamente ter demorado menos de 76 minutos para chegar à faculdade.

Resolução:

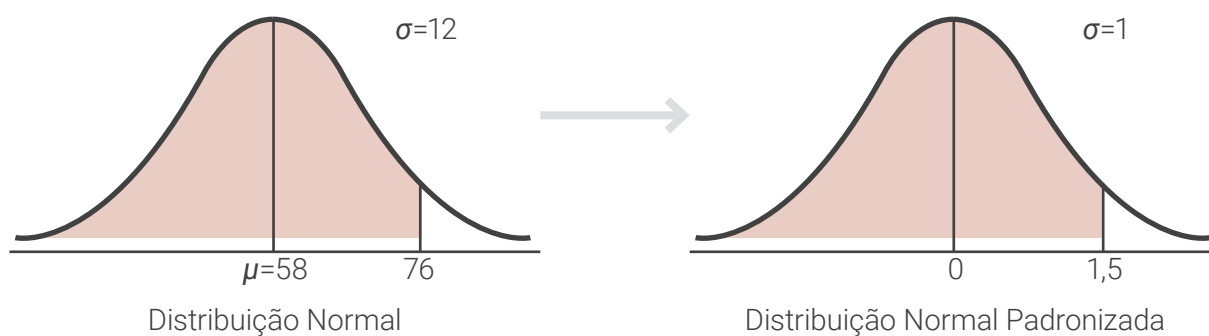
1º) Determine os valores da média μ e do desvio-padrão σ e nomeie a distribuição Normal: $X \sim N(58; 12^2)$

2º) Enuncie a probabilidade pedida: $P(X \leq 76) = ?$

3º) Transforme a variável aleatória Normal para Normal padronizada a partir da fórmula:

$$Z = \frac{X - \mu}{\sigma} = \frac{76 - 58}{12} = 1,5$$

4º) Desenhe as curvas e destaque a região cuja área será investigada para determinar a probabilidade pedida.

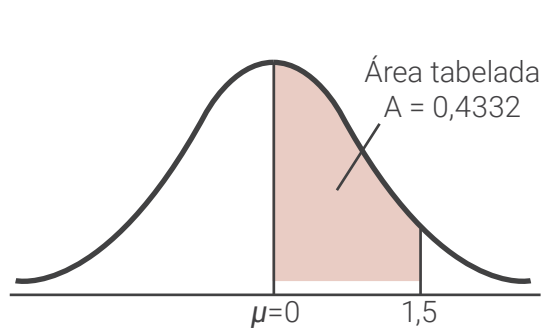


$$P(X \leq 76) = P(Z \leq 1,5)$$

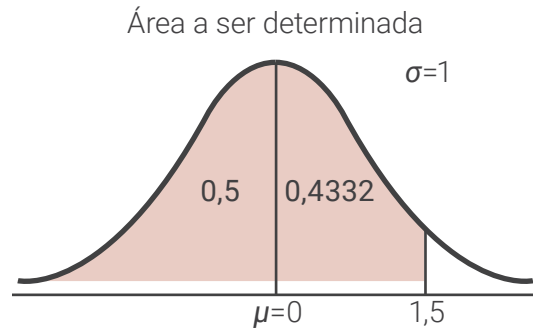
Fonte: Elaborada pela autora (2020).

5º) Utilize a tabela da distribuição Normal padronizada e busque o valor da área.
Para $Z=1,50$ temos $A=0,4332$

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1664	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633



Distribuição Normal Padronizada



Distribuição Normal Padronizada

Fonte: Elaborada pela autora (2020).

6º) Como a área a ser determinada é diferente da área tabelada precisamos fazer alguns ajustes para chegar à reposta correta.

Verificamos que:

Área a ser determinada = 0,5 + área tabelada

Área a ser determinada = 0,5 + 0,4332

Área a ser determinada = 0,9332

Logo, a probabilidade pedida será igual à área a ser determinada:

$$P(X \leq 76) = P(Z \leq 1,5) = 0,9332 = 93,32\%$$

Podemos concluir que a probabilidade de um aluno levar menos de 76 minutos para chegar na universidade é igual a 93,32%

Exemplo 3

Exemplifica o cálculo da probabilidade em que a z-score **é maior ou igual a um determinado valor negativo**.

Enunciado: Os “pesos” (massa corporal) dos frequentadores de uma academia de ginástica são normalmente distribuídos, com média de 65,3kg e desvio-padrão 5,5.

Um frequentador é selecionado ao acaso. Determine a probabilidade de seu peso ser maior do que 63,2 kg.

Resolução:

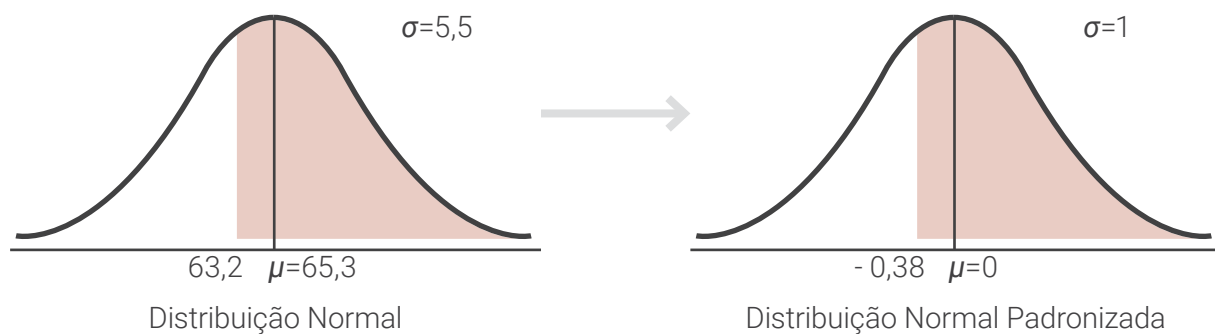
1º) Determine os valores da média μ e do desvio-padrão σ e nomeie a distribuição Normal: $X \sim N(65,3; 5,5^2)$

2º) Enuncie a probabilidade pedida: $P(X \geq 63,2) = ?$

3º) Transforme a variável aleatória Normal para Normal padronizada a partir da fórmula:

$$Z = \frac{X - \mu}{\sigma} = \frac{63,2 - 65,3}{5,5} = -0,38$$

4º) Desenhe as curvas e destaque a região cuja área será investigada para determinar a probabilidade pedida.



$$P(X \geq 63,2) = P(Z \geq -0,38)$$

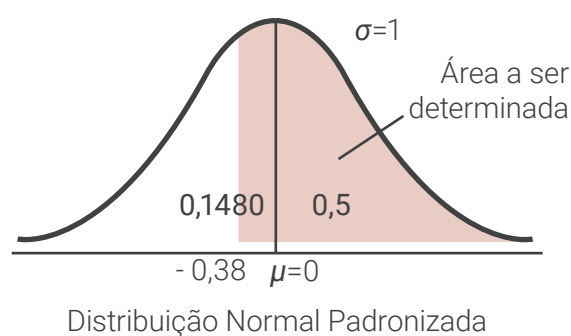
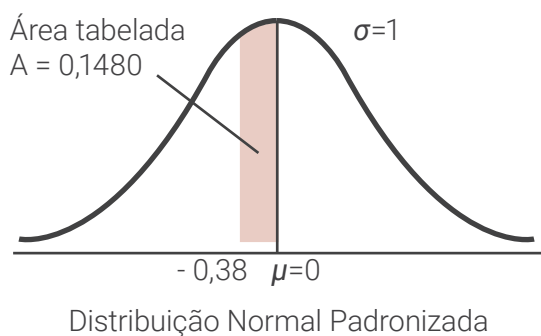
Fonte: Elaborada pela autora (2020).

5º) Utilize a tabela da distribuição Normal padronizada e busque o valor da área.

Para $Z = -0,38$ usaremos $z = 0,38$, pois as áreas são idênticas para os dois valores.

$A = 0,1480$

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517



Fonte: Elaborada pela autora (2020).

6º) Faça alguns ajustes para chegar à resposta correta, porque a área a ser determinada é diferente da área tabelada.

Verificamos que:

Área a ser determinada = 0,5 + área tabelada

Área a ser determinada = 0,5 + 0,1480

Área a ser determinada = 0,6480

Logo, a probabilidade pedida será igual à área a ser determinada:

$$P(X \geq 63,2) = P(Z \geq -0,38) = 0,6480 = 64,8\%$$

Podemos concluir que a probabilidade que um frequentador da academia selecionado ao acaso tenha seu peso superior a 63,2kg é igual a 64,8%.

Exemplo 4

Exemplifica o cálculo da probabilidade na qual z-score é **maior ou igual a um determinado valor positivo**.

Enunciado: Em 2019 o valor médio de gastos das famílias brasileiras é de R\$ 5.700,00 por mês. Suponha que os gastos mensais tenham uma distribuição normal e que o desvio-padrão seja igual a R\$ 1.500,00. Uma família é selecionada ao acaso. Determine a probabilidade dessa família ter um gasto mensal superior a R\$ 7.000,00.

Resolução:

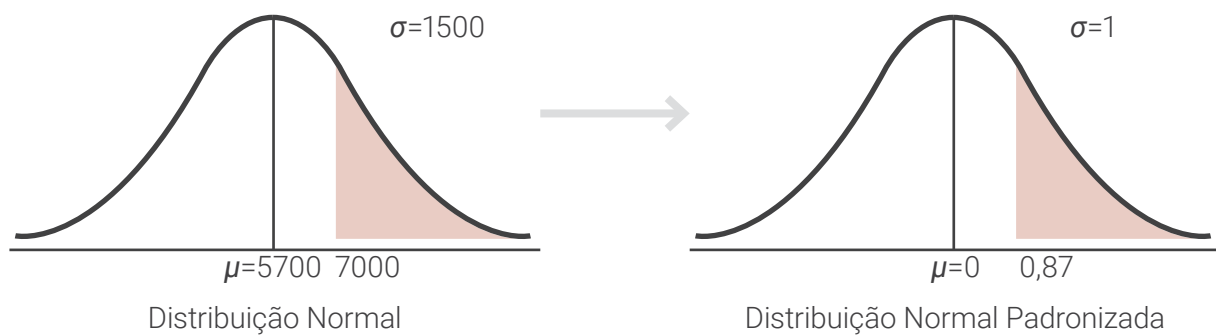
1º) Determine os valores da média μ e do desvio-padrão σ e nomeie a distribuição Normal: $X \sim N(5700; 1500^2)$

2º) Enuncie a probabilidade pedida: $P(X \geq 7000) = ?$

3º) Transforme a variável aleatória Normal para Normal padronizada a partir da fórmula:

$$Z = \frac{X - \mu}{\sigma} = \frac{7000 - 5700}{1500} = 0,87$$

4º) Desenhe as curvas e destaque a região cuja área será investigada para determinar a probabilidade pedida.



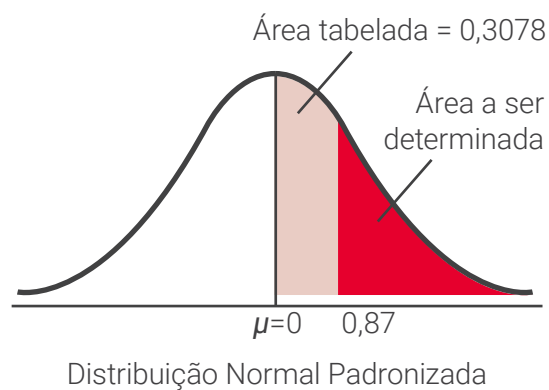
$$P(X \geq 7000) = P(Z \geq 0,87)$$

Fonte: Elaborada pela autora (2020).

5º) Utilize a tabela da distribuição Normal padronizada e busque o valor da área.

Para $Z=0,87 \rightarrow A=0,3078$

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1664	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133



Fonte: Elaborada pela autora (2020).

6º) Como a área a ser determinada é diferente da área tabelada precisamos fazer alguns ajustes para chegar à resposta correta.

Verificamos que:

Área a ser determinada + área tabelada = 0,5

Área a ser determinada = 0,5 – área tabelada

Área a ser determinada = 0,5 – 0,3078 = 0,1922

Logo, a probabilidade pedida será igual a área a ser determinada:

$$P(X \geq 7000) = P(Z \geq 0,87) = 0,1922 = 19,22\%$$

Podemos concluir que a probabilidade que uma família selecionada ao acaso tenha gasto mensal superior a R\$ 7.000,00 é igual a 19,22%.

Exemplo 5

Exemplifica o cálculo da probabilidade em que a z-score está entre dois valores determinados antes da média.

Enunciado: A quantia média anual que os moradores da cidade do Rio de Janeiro gastam por ano em aplicativos de transportes individuais R\$ 1.649,00. Suponha que a quantia gasta tenha uma distribuição normal com desvio-padrão igual a R\$ 843,00.

Determine a probabilidade de uma pessoa selecionada ao acaso gastar entre R\$ 1.042,00 e R\$ 1.568,00

Resolução:

1º) Determine os valores da média μ e do desvio-padrão σ e nomeie a distribuição Normal: $X \sim N(1649; 843^2)$

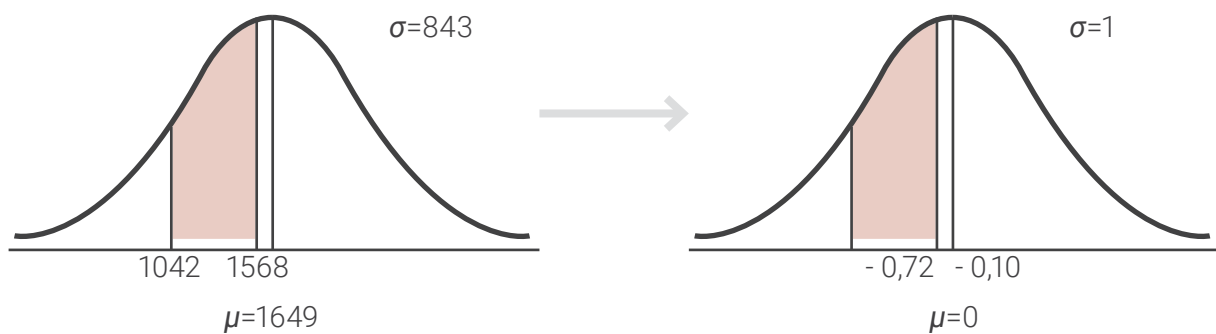
2º) Enuncie a probabilidade pedida: $P(1042 < X < 1568) = ?$

3º) Transforme a variável aleatória Normal para Normal padronizada a partir da fórmula:

$$\text{Para } X = 1042 \rightarrow Z_1 = \frac{X - \mu}{\sigma} = \frac{1042 - 1649}{843} = -0,72$$

Para $X=1568 \rightarrow Z_2 = \frac{X - \mu}{\sigma} = \frac{1568 - 1649}{843} = -0,10$

4º) Desenhe as curvas e destaque a região cuja área será investigada para determinar a probabilidade pedida.



Distribuição Normal

Distribuição Normal Padronizada

$$P(1042 < X < 1568) = P(-0,72 < Z < -0,10)$$

Fonte: Elaborada pela autora (2020).

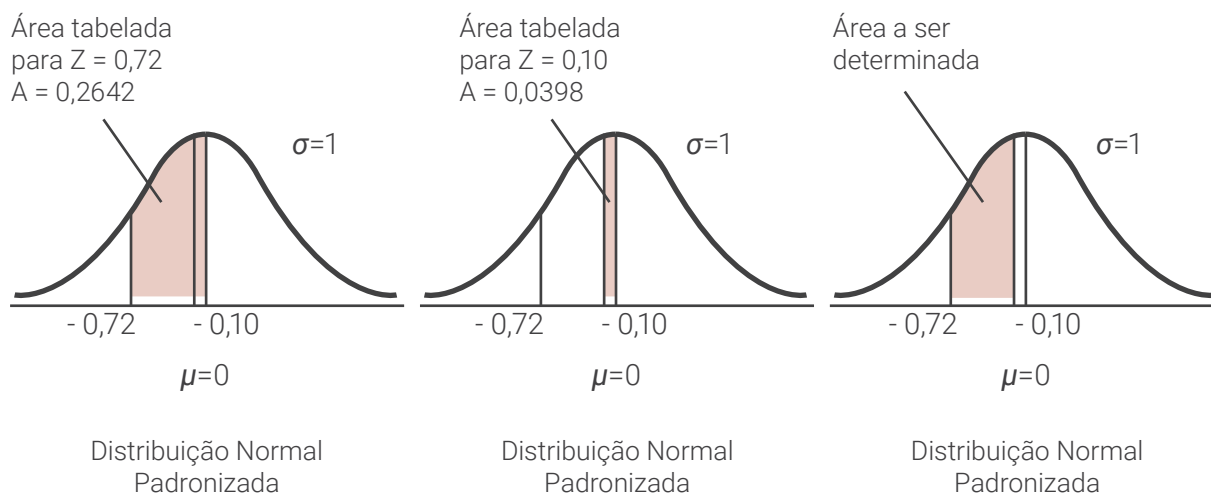
5º) Utilize a tabela da distribuição Normal padronizada e busque o valor das áreas.

Primeiramente para $Z=-0,72$ ou $Z=0,72$ já que as duas áreas são idênticas para esses valores $\rightarrow A = 0,2642$

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1664	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852

Agora para $Z = -0,10$ ou $Z = 0,10$ já que as duas áreas são idênticas para esses valores $A = 0,0398$

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753



Fonte: Elaborada pela autora (2020).

6º) Observe o colorido a área A1. Se apagarmos em A1 a área A2, vamos ficar exatamente com a área a ser determinada. Portanto:

Área a ser determinada = área tabelada 1 - área tabelada 2

Área a ser determinada = $0,2642 - 0,0398$

Área a ser determinada = $0,2244$

Logo, a probabilidade pedida será igual a área a ser determinada:

$$P(1042 < X < 1568) = P(-0,72 < Z < -0,10) = 0,2244 = 22,44\%$$

Então, a probabilidade de uma pessoa selecionada ao acaso gastar entre R\$ 1.042,00 e R\$ 1.568,00 em aplicativos de transporte é igual a 22,44%

Exemplo 6

Exemplifica o cálculo da probabilidade em que a z-score está entre dois valores determinados que contenham a média.

Vamos resolver o exemplo que apresentamos no início do estudo da distribuição Normal, que foi determinado com a resolução da integral. Agora, vamos resolvê-lo com o uso da tabela da distribuição normal padronizada.

Enunciado: Supondo que uma variável aleatória Normal apresente a média igual a 5 e a variância igual a 4, determine a probabilidade de a variável aleatória estar entre 3 e 6.

Resolução:

1º) Determine os valores da média μ e do desvio-padrão σ e nomeie a distribuição Normal: $X \sim N(5; 2^2)$

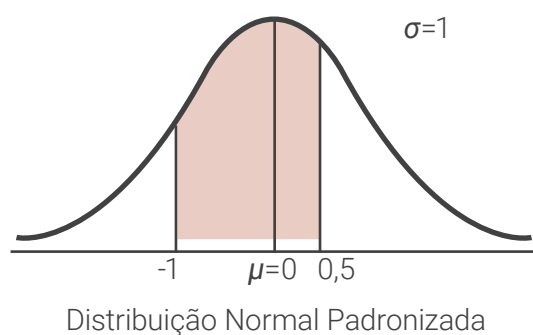
2º) Enuncie a probabilidade pedida: $P(3 \leq X \leq 6) = ?$

3º) Transforme a variável aleatória Normal para Normal padronizada a partir da fórmula:

$$\text{Para } X = 3 \rightarrow Z_1 = \frac{X - \mu}{\sigma} = \frac{3 - 5}{2} = -1$$

$$\text{Para } X = 6 \rightarrow Z_2 = \frac{X - \mu}{\sigma} = \frac{6 - 5}{2} = 0,5$$

4º) Desenhe as curvas e destaque a região cuja área será investigada para determinar a probabilidade pedida.



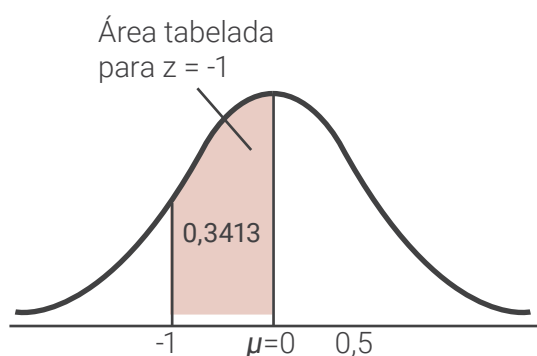
Fonte: Elaborada pela autora (2020).

5º) Utilize a tabela da distribuição Normal padronizada e busque o valor das áreas. Primeiramente para $Z = -1$ ou $Z = 1$. Com duas casas decimais $Z = 1,00 \rightarrow A = 0,3413$

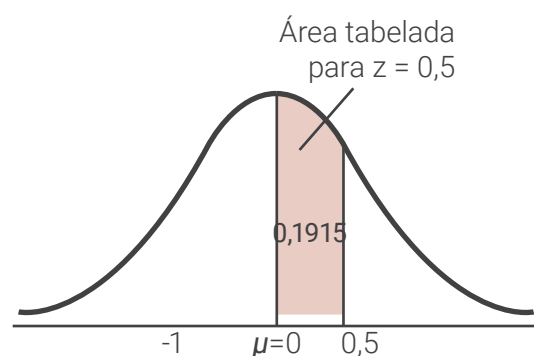
DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1664	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621

Agora para $Z=0,5$. Com duas casas decimais $Z=0,50 \rightarrow A=0,1915$

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1664	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224



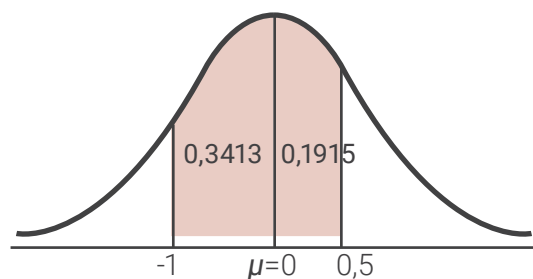
Distribuição Normal Padronizada



Distribuição Normal Padronizada

Fonte: Elaborada pela autora (2020).

Área a ser determinada



Distribuição Normal Padronizada

$$P(3 \leq X \leq 6) = P(-1 \leq Z \leq 0,5)$$

Fonte: Elaborada pela autora (2020).

6º) Como a área a ser determinada é exatamente a soma das duas áreas tabeladas temos:

Área a ser determinada = área tabelada 1 + área tabelada 2

Área a ser determinada = 0,3413 + 0,1915

Área a ser determinada = 0,5328

Logo, a probabilidade pedida será igual a área a ser determinada.

$$P(3 \leq X \leq 6) = P(-1 \leq Z \leq 0,5) = 0,5328 = 53,28\%$$

Obtivemos o mesmo valor encontrado resolvendo a integral definida,

$$\int_3^6 \frac{1}{2\sqrt{2}\pi} e^{-\frac{(x-5)^2}{2 \cdot 4}}$$

porém com o cálculo de apenas uma adição de duas áreas. As tabelas simplificam significativamente os cálculos.

Esse mesmo raciocínio do exemplo 6 será usado, caso seja preciso fazer o cálculo da probabilidade em que a z-score está **entre dois valores** determinados **depois da média**.

Exemplo 7

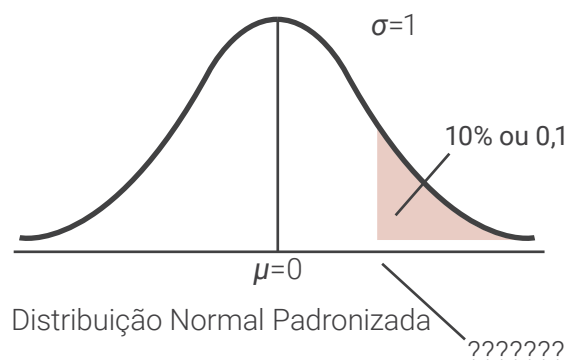
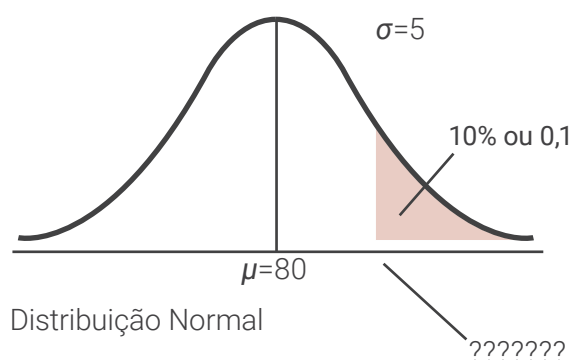
É um exemplo inverso. Nesse caso, sabemos a probabilidade pela qual queremos determinar o valor da variável aleatória.

Enunciado: As notas dos alunos de Estatística estão normalmente distribuídas com média igual a 80 pontos e desvio-padrão igual a 5 pontos. A universidade vai premiar 10% dos alunos que obtiverem as maiores notas na disciplina. Qual deve nota um aluno deve obter para que seja premiado?

Resolução:

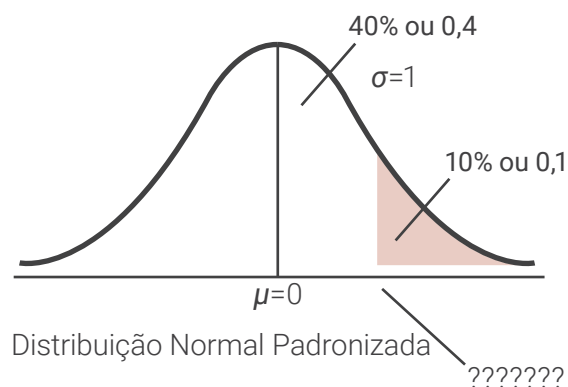
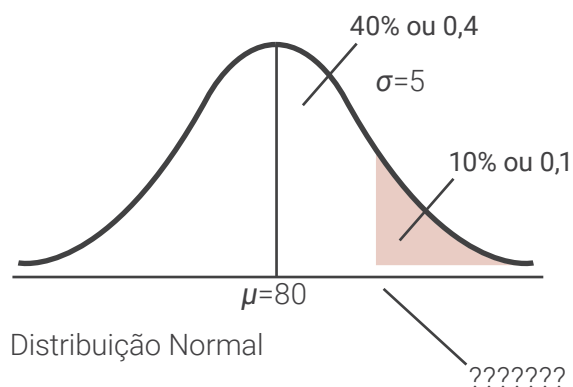
Observe que agora conhecemos a área sob a curva e precisamos determinar o z-score e consequentemente o valor da variável aleatória.

Observe o desenho da situação:



Fonte: Elaborada pela autora (2020).

Como sabemos que a curva tem 50% das probabilidades antes da média e 50% depois da média, então:



Fonte: Elaborada pela autora (2020).

Vamos buscar no corpo tabela de distribuição Normal padronizada o valor de z que gera uma área igual a 0,4.

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1664	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015

Com isso, descobrimos o valor do z-score, $Z=1,28$

Para determinarmos o valor da variável aleatória X, basta substituir os valores na fórmula:

$$Z = \frac{X - \mu}{\sigma}$$

$$1,28 = \frac{X - 80}{5}$$

$$1,28 \cdot 5 = X - 80$$

$$6,4 = X - 80$$

$$6,4 + 80 = X$$

$$86,4 = X$$

Logo, podemos concluir que os estudantes com notas iguais ou superiores a 86,4 serão premiados pela universidade.

Distribuições de Probabilidades com o Excel

A planilha eletrônica Excel tem a capacidade de calcular probabilidades de diversas distribuições de probabilidade. Apresentaremos, na tabela a seguir, os comandos das distribuições que foram estudados nesta unidade.

Distribuição	Sintaxe	Descrição
Binomial	<code>= DISTRBIONOM (x; n; p; cumulativo)</code>	<p>x: número de tentativas bem-sucedidas.</p> <p>n: número de repetições.</p> <p>p = probabilidade de sucesso em cada tentativa.</p> <p>Cumulativo: usado para acumular as probabilidades. Tem valor lógico 0 ou 1: 0 → probabilidade pontual 1 → probabilidade acumulada.</p>
Poisson	<code>= POISSON (x; μ; cumulativo)</code>	<p>x: número de eventos.</p> <p>μ: valor numérico esperado.</p> <p>Cumulativo: usado para acumular as probabilidades. Tem valor lógico 0 ou 1: 0 → probabilidade pontual. 1 → probabilidade acumulada.</p>
Normal	<code>= DIST.NORM.N (x; μ; σ; cumulativo)</code>	<p>x: valor cuja distribuição você deseja obter.</p> <p>μ: média aritmética da distribuição.</p> <p>σ = desvio-padrão da distribuição</p> <p>Cumulativo: usado para acumular as probabilidades. Tem valor lógico 0 ou 1: 0 → probabilidade não acumulada. 1 → probabilidade acumulada.</p>

Agora, com o auxílio da planilha eletrônica Excel resolveremos alguns dos exemplos práticos apresentados anteriormente.

Exemplo 1

Um professor de Estatística elaborou uma avaliação de múltipla escolha, que consiste em 10 questões, cada uma com cinco alternativas para resposta e apenas uma dessas alternativas tem a resposta correta. Suponha que os estudantes que irão a fazer a prova não frequentaram as aulas e não estudaram para a avaliação. Sendo assim, eles irão marcar a resposta aleatoriamente, o que informalmente chamamos de “chute”. Para ser aprovado, o aluno deve acertar no mínimo sete questões. Um estudante é selecionado ao acaso.

a) Qual é a probabilidade de ele “chutar” as 10 questões e acertar 7?

Sabemos que: $X \sim B(10, 0,2)$

Desejamos saber: $P(X=7)$

COMANDO	RESPOSTA
= DISTRBINOM (7;10;0,2;0)	0,00078643

b) Qual é a probabilidade de ele ser aprovado?

Desejamos saber $P(X \geq 7)$

COMANDO	RESPOSTA
= DISTRBINOM (7;10;0,2;0) + DISTRBINOM (8;10;0,2;0) + DISTRBINOM (9;10;0,2;0) + DISTRBINOM (10;10;0,2;0)	0,000864
Outra opção (usando a opção acumulada)	
= 1- DISTRBINOM (6;10;0,2;1)	0,000864

c) Qual é o valor esperado de acertos nessas condições.

COMANDO	RESPOSTA
= 10.0,2	2

d) O que podemos concluir de um aluno que não estuda e “chuta” as questões da avaliação?

COMANDO	RESPOSTA
Isso o Excel não faz. As conclusões ainda dependem da interpretação humana.	?????

Exemplo 2

Intervalo de tempo

Uma concessionária de rodovias está interessada em modelar o número de carros que chegam em uma determinada cabine de um pedágio em um período de 15 minutos. Para isso, verifica-se que a chegada dos carros satisfaz às seguintes hipóteses:

- A probabilidade de um carro chegar ao pedágio é a mesma para dois períodos quaisquer de igual duração.
- O fato de carros chegarem ou não em qualquer período independe da chegada ou não de outro carro em qualquer período.
- Segundo dados históricos o número médio de carros que passam por esse trecho da estrada é de 10 carros no período de 15 minutos.

Determine a probabilidade de:

- a) Exatamente cinco carros chegarem em um período de 15 minutos nessa cabine.
- b) Exatamente um carro chegar em um período de três minutos nessa cabine.

a) Como vimos na resolução analítica do exercício: $X \sim P(10)$
Desejamos saber $P(X=5)$

COMANDO	RESPOSTA
= POISSON (5;10;0)	0,03783327

b) Como vimos na resolução analítica do exercício: $X \sim P(2)$
Desejamos saber $P(X=1)$

COMANDO	RESPOSTA
= POISSON (1;2;0)	0,270671

Exemplo 3

Em agosto de 2020, um brasileiro passou em média 77 horas conectado à internet navegando em redes sociais. Suponha que os tempos estejam normalmente distribuídos e que o desvio-padrão seja de 20 horas.

Qual é a probabilidade de um brasileiro, escolhido aleatoriamente, ter passado menos de 50 horas conectado à internet em agosto de 2020?

Sabemos que: $X \sim N(77, 20^2)$

Desejamos saber $P(X \leq 50)$

COMANDO	RESPOSTA
= DIST.NORM.N (50;77;20;1)	0,08850799

Exemplo 4

Os “pesos” (massa corporal) dos frequentadores de uma academia de ginástica são normalmente distribuídos, com média de 65,3kg e desvio-padrão 5,5. Um frequentador é selecionado ao acaso.

Determine a probabilidade de seu peso ser maior do que 63,2 kg.

Sabemos que: $X \sim N(65,3; 5,5^2)$

Desejamos saber $P(X \geq 63,2)$

COMANDO	RESPOSTA
= 1 - DIST.NORM.N (63,2;65,3;5,5;1)	0,64870188

Exemplo 5

Supondo que uma variável aleatória Normal apresente a média igual a 5 e a variância igual a 4.

Determine a probabilidade de a variável aleatória estar entre 3 e 6.

Sabemos que $X \sim N(5; 2^2)$

Desejamos saber: $P(3 \leq X \leq 6) = ?$

COMANDO	RESPOSTA
= DIST.NORM.N (6;5;2;1) - DIST.NORM.N (3;5;2;1)	0,53280721



Ampliando o foco praticando

Para colocar em prática o conteúdo visto nesta unidade, visite a Biblioteca Virtual, escolha um livro de Estatística, selecione e faça alguns exercícios do livro para reforçar o conhecimento.

Confira as respostas que estão no final do livro e, se tiver alguma dúvida, fale com o seu tutor pelo Fórum de Dúvidas – Fale com o Tutor.



MIDIAATECA

Para ampliar o seu conhecimento veja o material complementar da Unidade 3, disponível na midiateca.



NA PRÁTICA

O problema da Grear Tire Company

Suponha que a Grear Tire Company tenha desenvolvido um pneu radial com cinturão de aço, que será vendido por meio de uma cadeia nacional de *discount stores*. Uma vez que esse tipo de pneu é um novo produto, os gerentes da Grear acreditam que a durabilidade (em termos de milhas rodadas) oferecida com o pneu será um fator importante na aceitação do produto. Antes de definirem os termos do contrato de garantia de durabilidade do pneu, os gerentes da Grear desejam obter informações de probabilidade a respeito do número de milhas que os pneus durarão.

A partir dos testes reais de estrada com os pneus, a equipe de engenharia da Grear estima que a durabilidade média dos pneus é $\mu = 36.500$ milhas (58.741 km) e que o desvio-padrão é $\sigma = 5000$ milhas. Além disso, os dados coletados indicam que a distribuição Normal é uma suposição razoável.

Qual a porcentagem de os pneus possivelmente durarem mais de 40.000 milhas (64.373 km)?

Utilizando os conceitos da distribuição Normal chegamos facilmente à resposta, ou seja 24,2% dos pneus terão durabilidade maior do que 40.000 milhas.

Imagine que a Grear esteja considerando a possibilidade de dar uma garantia que concede um desconto na troca de pneus se os originais não resistirem ao número de milhas estipulado na garantia.

Qual deve ser o número de milhas coberto pela garantia, levando-se em conta que a Grear quer que não mais do que 10% dos pneus sejam habilitados à garantia do desconto?

Utilizando os conceitos da distribuição Normal chegamos à reposta: uma garantia de 30.100 milhas cumprirá o requisito de que aproximadamente 10% dos pneus se habilitem à garantia. Talvez com essa informação, a empresa possa fixar a garantia de durabilidade de seis pneus em 30 mil milhas. Logo, constatamos o importante papel que as distribuições de probabilidade desempenham em termos de produzir informações para a tomada de decisões. Ou seja, assim que uma distribuição de probabilidade é estabelecida para uma aplicação em particular, ela pode ser usada rápida e facilmente para se obter informações a respeito do problema.

A probabilidade não determina a recomendação de uma decisão diretamente, mas fornece informações que ajudam o tomador de decisão a entender melhor os riscos e as incertezas associadas ao problema. Por fim, essas informações podem auxiliá-lo a tomar uma boa decisão.

(Texto extraído do livro citado na fonte a seguir.)

Fonte:

SWEENEY, D. J.; WILLIAMS, T. A.; ANDERSON, D. R. **Estatística aplicada à administração e economia**. 3. ed. São Paulo: Cengage Learning, 2013.

Resumo da Unidade 3

Nesta unidade ampliamos o estudo da Teoria das probabilidades. Apresentamos os modelos matemáticos que auxiliam na determinação das probabilidades de maneira fácil e metódica.

Abordamos as distribuições de probabilidades para variáveis aleatórias discretas e para variáveis aleatórias contínuas. A principal diferença conceitual entre as distribuições discretas e contínuas é que as distribuições discretas nos revelam a probabilidade de a variável aleatória assumir valores pontuais. Já as distribuições contínuas fornecem a probabilidade a partir do cálculo da área sob a curva.

Também estudamos as distribuições discretas: a Binomial e a de Poisson, enquanto as distribuições contínuas apresentadas foram a Normal e a Normal Padronizada. Vimos que a distribuição Normal tem grande importância para a Estatística e será utilizada também em inferência estatística.

Por fim, apresentamos os comandos para utilizar as distribuições de probabilidades com a planilha eletrônica Excel, que facilitam ainda mais a determinação das probabilidades.

Referências

COMO usar as tabelas da Distribuição Normal. **YouTube**. 08/10/2013. Disponível em: <https://youtu.be/ec9HWoY2kt8>. Acesso em: 6 set. 2020.

CURVA Normal. Disponível em: <https://www.geogebra.org/m/whcbqg3r> Acesso em: 3 set. 2020.

DISTRIBUIÇÃO binomial de probabilidade. **YouTube**. 25/09/2013. Disponível em: <https://www.youtube.com/watch?v=V2sfnVikFXA>. Acesso em: 6 set. 2020.

KOKOSKA, S. **Introdução à estatística**: uma abordagem por resolução de problemas. Rio de Janeiro: LTC, 2013.

LARSON, R.; FARBER, B. **Estatística aplicada**. São Paulo: Pearson, 2010.

SWEENEY, D. J.; WILLIAMS, T. A.; ANDERSON, D. R. **Estatística aplicada à administração e economia**. 3. ed. São Paulo: Cengage Learning, 2013.

TABELA da Distribuição Normal padronizada. UVA. Disponível em: http://ead.uva.br///recurso/DEF/EST/u3_c3_r1/anexo/Tabela-da-Distribuicao-Normal.pdf Acesso em: 5 set. 2020.

VARIÁVEIS aleatórias. **Khan Academy**. Disponível em: <https://pt.khanacademy.org/math/statistics-probability/random-variables-stats-library>. Acesso em: 4 set. 2020.

UNIDADE 4

Intervalos de confiança e
Relação entre variáveis

INTRODUÇÃO

Iniciamos os estudos desta disciplina falando da Estatística Descritiva, com suas técnicas para resumir um conjunto de dados por meio de tabelas, criar e apresentar gráficos e ainda para calcular as medidas de posição e dispersão. Depois, vimos a teoria da probabilidade e os modelos probabilísticos, discretos ou contínuos.

Agora, nesta unidade, esses conhecimentos serão utilizados no estudo de um significativo ramo da Estatística, denominado de **Inferência Estatística**. A Inferência tem como objetivo fornecer as ferramentas necessárias para fazermos generalizações sobre uma população, a partir de evidências fornecidas por uma amostra representativa, retirada desta população.

O que veremos nos tópicos, a saber:

Tópico 1 – Intervalos de Confiança para médias e proporções e cálculo do tamanho da amostra.

Tópico 2 – Conceitos de Correlação, medindo o grau de associação entre duas variáveis.

Tópico 3 – Concepções da Regressão Linear buscando a determinação de um modelo matemático que descreva o relacionamento entre duas variáveis aleatórias.

Com os conceitos de Intervalos de Confiança, Correlação e Regressão Linear seremos capazes de aplicar os conhecimentos adquiridos em projetos de pesquisa científica e na solução de problemas de nossa área de atuação com mais eficiência.



OBJETIVO

Nesta unidade você será capaz de:

- Utilizar os conceitos da inferência estatística para a resolução de problemas práticos.

Intervalos de confiança

Antes de apresentarmos os princípios dos Intervalos de Confiança, definiremos alguns conceitos importantes para um melhor entendimento do conteúdo, que são: parâmetros, estimadores e estimativas.

Parâmetros

São as características numéricas de uma **população**. Geralmente são representadas por letras gregas, tais como, θ , μ , σ , ρ entre outras. Em nossos estudos adotaremos " θ " para simbolizarmos parâmetros, " μ " para representarmos a média populacional, " σ " quando nos referirmos ao desvio-padrão populacional, " σ^2 " para variância populacional e " ρ " para o coeficiente de correlação populacional, entre outros.

Estimadores

Também chamados de "estatística" de um parâmetro, são as características numéricas determinada na **amostra**. Geralmente representamos por $\hat{\theta}$ (teta chapéu). São exemplos de estimadores: a média amostral (\bar{X}), a variância amostral (s^2), o desvio-padrão amostral (s), o coeficiente de correlação amostral (r), entre outros.

Entre várias características que os estimadores possuem devemos escolher estimadores que sejam:

- **Estimadores Não viciados, ou não tendenciosos ou, como dizemos em estatística, não viesados:** são aqueles em que o seu valor esperado é igual ao parâmetro estimado.
- **Estimadores consistentes:** são aqueles que à medida que o tamanho da amostra aumenta, o estimador aproxima-se cada vez mais do parâmetro da população.
- **Estimadores eficientes:** são aqueles que são não enviesados e possuem variância mínima.

Estimativa

É valor numérico assumido por um estimador. Pode ser **por ponto** (um único valor) ou **intervalar** (um intervalo de valores possíveis) de um parâmetro populacional desconhecido.

Agora que já entendemos sobre os conceitos basilares falaremos de Intervalo de Confiança.

Intervalo de confiança

É um intervalo numérico aberto, de modo que podemos estar seguramente certos ou confiantes de que o verdadeiro valor do parâmetro pertencerá a esse intervalo. Para iniciarmos a determinação de um intervalo de confiança devemos escolher qual o nível de confiança que iremos trabalhar. Temos dois níveis. São eles:

Nível de confiança	É a porcentagem de que o intervalo de confiança contenha o parâmetro em amostragens repetidas. Os níveis de confiança mais utilizados são: 90%, 95% e 99%.
Nível de significância	Representado pela letra grega α , é o complementar do nível de confiança, ou seja, se o nível de confiança for 90%, o nível de significância será 10%; para 95% de confiança, 5% de significância e 99% de confiança, 1% de significância.

Intervalos de confiança para Média

Temos dois casos a considerar quando queremos construir um intervalo de confiança para a média populacional.

1º caso: Intervalo de confiança para média populacional quando o desvio-padrão populacional é conhecido.

Para construirmos esse intervalo de confiança devemos seguir cinco passos. Vejamos:

1º) Verificar se a população obedece aos princípios de uma distribuição normal, ou assegurar que o tamanho da amostra seja suficientemente grande, ou seja, $n > 30$.

2º) Identificar o valor numérico do desvio-padrão populacional (σ); da média amostral (\bar{X}) e o tamanho da amostra (n).

3º) Calcular, então, o desvio-padrão da média com a utilização da fórmula:

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

Em que:

σ_x = desvio-padrão da média

σ = desvio-padrão populacional

n = tamanho da amostra

4º) Determinar com auxílio da tabela de distribuição normal padronizada o valor de Z-score em função do nível de confiança. Como vimos na Unidade 3, existem vários tipos de tabela normal padronizada e por isso temos que ter bastante cuidado nessa etapa. Na tabela disponibilizada em nosso curso, buscamos a **metade** no nível de confiança, no corpo da tabela, para determinar Z-score.

5º) Substituir os valores já conhecidos na fórmula do intervalo de confiança:

$$IC = \bar{X} \pm Z_{score} \cdot \sigma_x$$

Vejamos alguns exemplos práticos a seguir.

Exemplo 1:

Considere que a altura dos alunos de Estatística estejam normalmente distribuídas com o desvio-padrão populacional igual a 15 cm. Foi retirada uma amostra aleatoriamente, composta de 100 alunos, e obteve-se a média amostral igual a 1,75 cm.

Determine o intervalo com 95% de confiança que contenha a verdadeira altura média dos alunos de Estatística.

Resolução:

Vamos, então, determinar os cinco passos para construir esse intervalo de confiança, para média μ , ao nível de confiança de 95%, com o desvio-padrão populacional conhecido:

Etapas	Respostas
1º) Verificar se a população obedece aos princípios de uma distribuição normal, ou assegurar que o tamanho da amostra seja suficientemente grande, ou seja, $n > 30$;	No enunciado verificamos que a população formada pela medida da altura dos alunos de Estatística está normalmente distribuída.

2º) Identificar o valor numérico do desvio-padrão populacional (σ); da média amostral (\bar{X}) e o tamanho da amostra (n);

$$\sigma = 15 \text{ cm}$$

$$\bar{X} = 175 \text{ m}$$

$$n = 100$$

3º) Calcular o desvio-padrão da média com a utilização da fórmula:

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_x = \frac{15}{\sqrt{100}} = 1,5$$

4º) Determinar com o auxílio da tabela de distribuição normal padronizada o valor de Z-score em função do nível de confiança. Como vimos na Unidade 3, existem vários tipos de tabela normal padronizada e por isso temos que ter bastante cuidado nessa etapa. Na tabela disponibilizada em nosso curso, buscamos a metade do nível de confiança, no corpo da tabela, para determinar Z-score.

Se o **desvio-padrão populacional é conhecido e o tamanho da amostra é maior do que 30**, vamos usar a **tabela normal padronizada** para determinar o valor de Z-score.

Verificamos que o nível de confiança solicitado foi de 95%.

Como nossa tabela oferece os valores de zero até z, temos que dividir esse nível de confiança por 2, ou seja,

$$\frac{95\%}{2} = 47,5\% = \frac{47,5\%}{100} = 0,475$$

Buscamos então, **no corpo da tabela** de distribuição normal padronizada esse valor:

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1664	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817

Fonte: http://ead.uva.br///recurso/DEF/EST/u3_c3_r1/anexo/Tabela-da-Distribuicao-Normal.pdf

Depois de identificado o valor da metade do nível de confiança no corpo da tabela, determinamos o Z-score usando na primeira coluna correspondente o valor da parte inteira e da primeira casa decimal (1,9) e na primeira linha o valor da segunda casa decimal (0,06):

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k									
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761
2	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812

Logo, $Z_{score} = 1,96$

5º) Substituir os valores já conhecidos na fórmula do intervalo de confiança:

$$IC = \bar{X} \pm Z_{score} \cdot \sigma_x$$

$$IC = \bar{X} \pm Z_{score} \cdot \sigma_x$$

$$IC = 175 \pm 1,96 \cdot 1,5$$

$$IC = 175 \pm 2,94$$

$$IC = 175 - 2,94 = 172,06$$

$$IC = 175 + 2,94 = 177,94$$

Podemos afirmar, com 95% de confiança, que o intervalo]172,06;177,94[contém o verdadeiro valor da altura média dos alunos.

Exemplo 2:

Deseja-se estimar a quantia que cada cliente gasta para jantar em um famoso restaurante do Rio de Janeiro. Para isso, foram coletados dados de uma amostra de 40 clientes, cuja média de gastos foi igual a R\$ 173,88. Sabe-se que há dados históricos que nos permitem obter uma boa estimativa do desvio-padrão populacional, que, neste caso, é igual a R\$ 10,00.

Construa um intervalo de confiança de 99% para o verdadeiro valor dos gastos médios dos clientes desse famoso restaurante por jantar, sabendo que esses gastos estão normalmente distribuídos.

Etapas	Respostas
<p>1º) Verificar se a população obedece aos princípios de uma distribuição normal, ou assegurar que o tamanho da amostra seja suficientemente grande, ou seja, $n > 30$;</p>	<p>Sim, os gastos estão normalmente distribuídos. A amostra é igual a 40 clientes, logo, maior do que 30.</p>
<p>2º) Identificar o valor numérico do desvio-padrão populacional (σ); da média amostral (\bar{X}) e o tamanho da amostra (n);</p>	<p>$\sigma = 10$ $\bar{X} = 173,88$ $n = 40$</p>
<p>3º) Calcular o desvio-padrão da média com a utilização da fórmula:</p> $\sigma_x = \frac{\sigma}{\sqrt{n}}$	$\sigma_x = \frac{10}{\sqrt{40}} = 1,58113883$
<p>4º) Determinar com auxílio da tabela de distribuição normal padronizada o valor de Z-score em função do nível de confiança. Como vimos na Unidade 3, existem vários tipos de tabela normal padronizada e por isso temos que ter bastante cuidado nessa etapa. Na tabela disponibilizada em nosso curso, buscamos a metade do nível de confiança, no corpo da tabela, para determinar Z-score.</p>	<p>Se o desvio-padrão populacional é conhecido e o tamanho da amostra é maior do que 30, vamos usar a tabela normal padronizada para determinar o valor de Z-score.</p> <p>Verificamos que o nível de confiança solicitado foi de 99%.</p> <p>Como nossa tabela oferece os valores de zero até z, temos que dividir esse nível de confiança por 2, ou seja,</p> $\frac{99\%}{2} = 49,5\% = \frac{49,5\%}{100} = 0,495$

Buscamos, então, **no corpo da tabela** de distribuição normal padronizada esse valor:

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1664	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817

Fonte: http://ead.uva.br///recurso/DEF/EST/u3_c3_r1/anexo/Tabela-da-Distribuicao-Normal.pdf

Não encontramos o valor exato 0,495. Então, vamos escolher os valores mais próximos pela falta e pelo excesso, que são 0,4949 e 0,4951.

DISTRIBUIÇÃO NORMAL – ÁREA DE 0 até K => probabilidade da variável z estar entre 0 e k										
z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1664	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964

Depois de identificado o valor da metade do nível de confiança no corpo da tabela, determinamos o Z-score buscando na primeira coluna correspondente o valor da parte inteira e da primeira casa decimal

	<p>(2,5) e na primeira linha os dois valores da segunda casa decimal e calculamos a média desses dois valores. (0,08 e 0,07, cuja média é 0,075):</p> <p>Logo, $Z_{score} = 2,575$.</p>
<p>5º) Substituir os valores já conhecidos na fórmula do intervalo de confiança:</p> $IC = \bar{X} \pm Z_{score} \cdot \sigma_x$	$IC = \bar{X} \pm Z_{score} \cdot \sigma_x$ $IC = 173,88 \pm 2,575 \cdot 1,58113883$ $IC = 173,88 \pm 4,0714$ $IC = 173,88 - 4,0714 = 169,81$ $IC = 173,88 + 4,0714 = 177,95$ <p>Podemos afirmar, com 95% de confiança que o intervalo (169,81;177,95) contém o verdadeiro valor dos gastos médios dos clientes.</p>

2º caso: Intervalo de confiança para média populacional quando o desvio-padrão populacional é desconhecido.

Temos **duas situações** a considerar:

1ª Situação: se o desvio-padrão populacional não é conhecido, a população está normalmente distribuída e estamos trabalhando com **grandes amostras**, ou seja, se $n > 30$, para determinar o intervalo de confiança para a média, procedemos da mesma forma que no 1º caso.

2ª Situação: se o desvio-padrão populacional não é conhecido, a população está normalmente distribuída e estamos trabalhando com **pequenas amostras**, ou seja, se $n \leq 30$, para determinar o intervalo de confiança para a média, devemos seguir os seguintes passos:

1º) Verificar se a população obedece aos princípios de uma distribuição normal, ou assegurar que o tamanho da amostra seja suficientemente pequeno, ou seja, $n \leq 30$.

2º) Identificar o valor numérico do desvio-padrão amostral (s); da média amostral (\bar{X}) e o tamanho da amostra (n);

3º) Calcular o desvio-padrão da média com a utilização da fórmula:

$$S_x = \frac{s}{\sqrt{n}}$$

Em que:

S_x = desvio-padrão amostral da média

s = desvio-padrão amostral

n = tamanho da amostra

4º) Determinar, com auxílio da tabela de distribuição t-Student, o valor de t-score em função do nível de significância. Na tabela t-Student buscamos a **metade** no nível de **significância**, juntamente com (n-1) graus de liberdade, sendo “n” o tamanho da pequena amostra para determinar t-score.

Vejamos aqui, uma de muitas configurações da tabela t-Student.

$t \cdot \frac{\alpha}{2}$	25%	10%	5%	2,5%	1%	0,5%
Grau de liberdade						
1	1,0000	3,0777	6,3138	12,7062	31,8207	63,6574
2	0,8165	1,8856	2,9200	4,3027	6,9646	9,9248
3	0,7649	1,6377	2,3534	3,1824	4,5407	5,8409
4	0,7407	1,5332	2,1318	2,7764	3,7469	4,6041
5	0,7267	1,4759	2,0150	2,5706	3,3649	4,0322
6	0,7176	1,4398	1,9432	2,4469	3,1427	3,7074
7	0,7111	1,4149	1,8946	2,3646	2,9980	3,4995
8	0,7064	1,3968	1,8595	2,3060	2,8965	3,3554

9	0,7027	1,3830	1,8331	2,2622	2,8214	3,2498
10	0,6998	1,3722	1,8125	2,2281	2,7638	3,1693
11	0,6974	1,3634	1,7959	2,2010	2,7181	3,1058
12	0,6955	1,3562	1,7823	2,1788	2,6810	3,0545
13	0,6938	1,3502	1,7709	2,1604	2,6503	3,0123
14	0,6924	1,3450	1,7613	2,1448	2,6245	2,9768
15	0,6912	1,3406	1,7531	2,1315	2,6025	2,9467
16	0,6901	1,3368	1,7459	2,1199	2,5835	2,9208
17	0,6892	1,3334	1,7396	2,1098	2,5669	2,8982
18	0,6884	1,3304	1,7341	2,1009	2,5524	2,8784
19	0,6876	1,3277	1,7291	2,0930	2,5395	2,8609
20	0,6870	1,3253	1,7247	2,0860	2,5280	2,8453
21	0,6864	1,3232	1,7207	2,0796	2,5177	2,8314
22	0,6858	1,3212	1,7171	2,0739	2,5083	2,8188
23	0,6853	1,3195	1,7139	2,0687	2,4999	2,8073
24	0,6848	1,3178	1,7109	2,0639	2,4922	2,7969
25	0,6844	1,3163	1,7081	2,0595	2,4851	2,7874
26	0,6840	1,3163	1,7056	2,0555	2,4786	2,7787
27	0,6837	1,3137	1,7033	2,0518	2,4727	2,7707
28	0,6834	1,3125	1,7011	2,0484	2,4671	2,7633
29	0,6830	1,3114	1,6991	2,0452	2,4620	2,7564
30	0,6828	1,3104	1,6973	2,0423	2,4573	2,7500

Student é o pseudônimo do químico e matemático inglês William Sealy Gosset (1876-1937), funcionário da cervejaria irlandesa Guinness Brewing Company, em Dublin, no início do século XX, criador da Distribuição t.

Fonte: http://ead.uva.br////recurso/DEF/EST/u4_c1_r1/index.htm

Observe que, para utilizarmos essa tabela, precisamos da **metade do nível de significância** e dos **graus de liberdade**.

Para determinarmos metade do nível de significância basta fazer 100% - Nível de Confiança e dividir por 2.

Quanto aos graus de liberdade, basta diminuirmos 1 do tamanho da amostra.

5º) Substituir os valores já conhecidos na fórmula do intervalo de confiança:

$$IC = \bar{X} + t_{score} \cdot S_x$$



Importante

Chamamos de “margem de erro” o produto: $Z_{score} \cdot \sigma_x$ ou $T_{score} \cdot S_x$.

Vejamos um exemplo que aborda o conteúdo visto em questão de concurso público.

Exemplo

Concurso: Colégio Pedro II - 2017

(Fonte: <https://www.teconconcursos.com.br/questoes/720037>).

A taxa de propagação de mensagens via torpedo eletrônico é uma importante característica da velocidade de uma operadora. Suponha que a taxa de envio seja considerada uma variável aleatória com distribuição Normal. Certa operadora garante que sua taxa de transmissão média é de 54 mensagens por segundo. Para checar a validade da informação alegada pela operadora, a agência controladora de telefonia decide, então, realizar um experimento.

Para isso, ela coleta uma amostra com 25 mensagens e observa uma média de 52,4 mensagens por segundo e um desvio-padrão de 2,1 mensagens por segundo. O intervalo de confiança de 95% para a taxa média e a conclusão da agência controladora foram:

- a) [51,577;53,223] - não existem evidências de que a operadora esteja falando a verdade.
- b) [51,53;53,27] - não existem evidências de que a operadora esteja falando a verdade.
- c) [53,18;54,82] - não se pode rejeitar a afirmação da operadora.
- d) [53,13;54,87] - não se pode rejeitar a afirmação da operadora.

Etapas	Respostas
1º) Verificar se a população obedece aos princípios de uma distribuição normal, ou assegurar que o tamanho da amostra seja suficientemente grande, ou seja, $n > 30$;	Sim, no enunciado verificamos que a taxa de envio é considerada uma variável aleatória com distribuição Normal.
2º) Identificar o valor numérico do desvio-padrão populacional (σ) ou amostral (S), da média amostral (\bar{X}) e o tamanho da amostra (n);	$\sigma = \text{desconhecido}$ $S = 2,1$ $\bar{X} = 52,4$ $n = 25$ Se o desvio-padrão populacional não é conhecido e o tamanho da amostra é menor do que 30 , vamos usar a tabela t-Student para determinar o valor de t-score.
3º) Calcular o desvio-padrão da média com a utilização da fórmula: $S_x = \frac{S}{\sqrt{n}}$	$S_x = \frac{2,1}{\sqrt{25}} = 0,42$
4º) Determinar, com auxílio da tabela t-Student, o valor de t-score.	Verificamos que o nível de confiança solicitado foi de 95% . Logo, o nível de significância será: $\alpha = 100\% - 95\% = 5\%$

Então, a metade do nível de significância é igual a

$$\frac{5\%}{2} = 2,5\%$$

Para determinarmos os graus de liberdade, basta diminuir 1 no tamanho da amostra. Então: GL = 25 - 1 = 24

Procuramos a metade do nível de significância na primeira linha da tabela e os graus de liberdade na primeira coluna da tabela e fazemos o cruzamento desses dois dados, obtendo o valor de t-score.

t g 2	25%	10%	5%	2,5%	1%	0,5%
Grau de liberdade						
1	1,0000	3,0777	6,3138	12,7062	31,8207	63,6574
2	0,8165	1,8856	2,9200	4,3027	6,9646	9,9248
3	0,7649	1,6377	2,3534	3,1824	4,5407	5,8409
4	0,7407	1,5332	2,1318	2,7764	3,7469	4,6041
5	0,7267	1,4759	2,0150	2,5706	3,3649	4,0322
6	0,7176	1,4398	1,9432	2,4469	3,1427	3,7074
7	0,7111	1,4149	1,8946	2,3546	2,9980	3,4995
8	0,7064	1,3968	1,8595	2,3060	2,8965	3,3554
9	0,7027	1,3830	1,8331	2,2522	2,8214	3,2498
10	0,6998	1,3722	1,8125	2,2281	2,7638	3,1693
11	0,6974	1,3634	1,7959	2,2110	2,7181	3,1058
12	0,6955	1,3562	1,7823	2,1788	2,6810	3,0545
13	0,6938	1,3502	1,7709	2,1504	2,6503	3,0123
14	0,6924	1,3450	1,7613	2,1448	2,6245	2,9768
15	0,6912	1,3406	1,7531	2,1315	2,6025	2,9467
16	0,6901	1,3368	1,7459	2,1199	2,5835	2,9208
17	0,6892	1,3334	1,7396	2,1098	2,5669	2,8982
18	0,6884	1,3304	1,7341	2,1009	2,5524	2,8784
19	0,6876	1,3277	1,7291	2,0930	2,5395	2,8609
20	0,6870	1,3253	1,7247	2,0860	2,5280	2,8453
21	0,6864	1,3232	1,7207	2,0796	2,5177	2,8314
22	0,6858	1,3212	1,7171	2,0739	2,5083	2,8188
23	0,6853	1,3195	1,7139	2,0687	2,4999	2,8073
24	0,6848	1,3178	1,7109	2,0639	2,4922	2,7969
25	0,6844	1,3163	1,7081	2,0595	2,4851	2,7874

Logo, $t_{score} = 2,0639$.

5º) Substituir os valores já conhecidos na fórmula do intervalo de confiança:

$$IC = \bar{X} \pm t_{score} \cdot S_x$$

$$IC = \bar{X} \pm t_{score} \cdot S_x$$

$$IC = 52,4 \pm 2,0639 \cdot 0,42$$

$$IC = 52,4 - 0,866838 = 51,53$$

$$IC = 52,4 + 0,866838 = 53,27$$

[51,53;53,27] -

Podemos afirmar, com 95% de confiança, que o intervalo (51,53;53,27) contém o verdadeiro valor da taxa média de propagação das mensagens por segundo.

Verificamos que esse intervalo não contém a taxa média de 54 mensagens enviadas por segundo que foi anunciada pela operadora. Logo, não existem evidências de que essa afirmação seja verdadeira.

Portanto, a alternativa correta é a letra B.



Importante

O nível de confiança é a frequência com a qual esperamos que o intervalo observado contenha o valor correto para o parâmetro de interesse quando o experimento é repetido várias vezes.

Logo, um nível de confiança de 90% significa que 90% dos intervalos de confiança, construídos a partir das amostras aleatórias, contêm o valor verdadeiro do parâmetro, ou seja, se retirarmos inúmeras amostras da população, em 90% dessas amostras encontraremos o verdadeiro valor do parâmetro nesse intervalo.

Um intervalo de confiança de 90% **não significa** que há a probabilidade de 90% do parâmetro da população pertencer ao intervalo.

Não é somente para as médias que podemos estabelecer estimativas intervalares.

Agora, vejamos o Intervalo de Confiança para proporção:

Muito utilizados, os intervalos de confiança para proporção são construídos de maneira muito semelhante aos intervalos de confiança para médias.

As etapas continuam as mesmas, entretanto, no caso das proporções, utilizaremos sempre como referência a Distribuição Normal e o cálculo do desvio-padrão das proporções, que é dado por:

$$p_x = \sqrt{\frac{p \cdot (1 - p)}{n}}$$

Sendo, então, o intervalo de confiança:

$$IC = p \pm Z_{score} \cdot p_x$$

Em que:

p = proporção

n = tamanho da amostra

p_x = desvio-padrão da proporção

Podemos construir ainda intervalos de confiança para a variância, para o desvio-padrão e para operações como soma, subtração e quociente de medidas estatísticas. São construídos de maneira muito similar aos que foram apresentados aqui.



Ampliando o foco

Para colocar em prática o conteúdo visto nesta unidade, visite a Biblioteca Virtual, escolha um livro de Estatística, selecione e faça alguns exercícios do livro para reforçar o conhecimento sobre a construção de intervalos de confiança.

Confira as respostas que estão no final do livro e, se tiver alguma dúvida, fale com o seu tutor pelo Fórum de Dúvidas – Fale com o Tutor.

Tamanho da amostra

É importante em qualquer pesquisa **determinar o tamanho da amostra** para que ela seja representativa da população.

- **Amostra pequena:** corremos o risco de errar fortemente na inferência.
- **Amostra grande:** caímos no desperdício de tempo e dinheiro.

Como já vimos a tabela normal, agora podemos apresentar os conceitos para o cálculo do tamanho da amostra.

- Cálculo do tamanho da amostra para estimar a média de uma população infinita

Uma população é considerada infinita quando 5% do número de seus elementos for maior do que 1.000. Nesse caso, se a variável for quantitativa, o tamanho da amostra é dado por:

$$n = \left(\frac{Z_{score} \cdot \sigma}{\varepsilon} \right)^2$$

Em que:

n = tamanho da amostra aleatória simples a ser selecionada da população em estudo.
 Z_{score} = valor encontrado na tabela normal padronizada em função do nível de confiança.
 σ = desvio-padrão populacional. Que pode ser obtido, pelo menos, de quatro maneiras:

1. A partir de especificações técnicas.
2. A partir de conjecturas com base em amostras-piloto.

3. Comparação com estudos semelhantes.
4. A partir de dados históricos.

ε = erro amostral. É a máxima diferença que o pesquisador admite que ocorra entre a média populacional (μ) e a média amostral (\bar{X}). É expresso em decimais.

- Cálculo do tamanho da amostra para estimar a média de uma população finita

Uma população é considerada finita quando 5% do número de seus elementos for menor do que 1.000. Nesse caso, se a variável for quantitativa, o tamanho da amostra é dado por:

$$n = \frac{Z_{score}^2 \cdot \sigma^2 \cdot N}{\varepsilon^2 \cdot (N - 1) + Z_{score}^2 \cdot \sigma^2}$$

Em que:

n = tamanho da amostra aleatória simples a ser selecionada da população em estudo.

Z_{score} = valor encontrado na tabela normal padronizada em função do nível de confiança.

σ = desvio-padrão populacional.

ε = erro amostral.

N = tamanho da população

- Cálculo do tamanho da amostra para estimar a proporção (p) de uma população infinita

Se a variável escolhida for qualitativa, ordinal ou nominal, e a população considerada infinita, o tamanho da amostra será calculado por meio da fórmula:

$$n = \frac{Z^2 \cdot \hat{p} \cdot \hat{q}}{\varepsilon^2}$$

Em que:

n = tamanho da amostra aleatória simples a ser selecionada da população.

Z_{score} = valor encontrado na tabela normal padronizada.

ε = erro amostral. É a máxima diferença que o pesquisador admite que ocorra entre a média populacional (μ) e a média amostral (\bar{X}). É expresso em decimais.

\hat{p} = estimativa da verdadeira proporção, expresso em decimais. Caso não se tenham estimativas prévias para \hat{p} , admitimos, então, $\hat{p} = 0,5$.

$$\hat{q} = 1 - \hat{p}$$

- Cálculo do tamanho da amostra para estimar a proporção (p) de uma população finita

Se a variável escolhida for qualitativa, ordinal ou nominal e a população considerada finita, o tamanho da amostra será calculado por meio da fórmula:

$$n = \frac{Z^2 \cdot \hat{p} \cdot \hat{q} \cdot N}{\varepsilon^2 (N - 1) + Z^2 \cdot \hat{p} \cdot \hat{q}}$$

Em que:

n = tamanho da amostra aleatória simples a ser selecionada da população.

Z_{score} = valor encontrado na tabela normal padronizada .

ε = erro amostral.

\hat{p} = estimativa da verdadeira proporção, expresso em decimais. Caso não se tenham estimativas prévias para \hat{p} , admitimos, então, $\hat{p} = 0,5$

$$\hat{q} = 1 - \hat{p}$$

N = tamanho da população.

Vejamos alguns exemplos práticos.

Exemplo 1

Determine o tamanho da amostra que devemos obter para estimar a média da altura de todos os estudantes da Universidade Veiga de Almeida, com 95% de confiança e a tolerância de 2cm para mais ou para menos. Sabe-se, por estudos anteriores, que o desvio-padrão da altura dos estudantes é igual a 20cm.

Resolução:

Sabemos que:

$$n = ?$$

Z_{score} – buscando no corpo da tabela a metade do nível de confiança, ou seja, $0,95/2 = 0,475$.

Obtemos:

$$Z_{score} = 1,96$$

$$\sigma = 20 \text{ cm}$$

$$\varepsilon = 2 \text{ cm}$$

Logo, substituindo na fórmula do cálculo do tamanho da amostra para estimar a média de uma população infinita, temos:

$$n = \left(\frac{Z_{score} \cdot \sigma}{\varepsilon} \right)^2 = \left(\frac{1,96 \cdot 20}{2} \right)^2 = 385 \text{ alunos}$$

Exemplo 2

Uma empresa administradora de cartões de crédito deseja estimar a proporção de clientes que não fizeram o pagamento integral da fatura no final de um mês. Suponha que a margem de erro desejada seja de três pontos percentuais, com 98% de confiança.

Determine o tamanho da amostra: deve ser selecionado para estimar essa proporção?

Resolução:

Sabemos que:

$$n = ?$$

$$Z_{score} = \text{buscando no corpo da tabela a metade do nível de confiança, ou seja, } 0,98/2 = 0,49,$$

$$\text{obtemos, } Z_{score} \cong 2,33$$

$$\varepsilon = 3\% = 0,03$$

\hat{p} = vamos admitir $\hat{p} = 0,5$, já que não temos estimativas prévias dessa proporção.

$$\hat{q} = 1 - \hat{p} = 1 - 0,5 = 0,5$$

Logo, substituindo na fórmula do cálculo do tamanho da amostra para estimar a proporção (p) de uma população infinita, temos:

$$n = \frac{Z^2 \cdot \hat{p} \cdot \hat{q}}{\varepsilon^2} = \frac{2,33^2 \cdot 0,5 \cdot 0,5}{(0,03)^2} \cong 1509 \text{ clientes}$$

Correlação

É um termo estatístico que indica a relação entre duas ou mais variáveis. Um dos propósitos dos pesquisadores é verificar a associação entre as variáveis estudadas. A possível existência de um relacionamento entre variáveis é fundamental para análises e conclusões do estudo que está sendo realizado. Intuitivamente, muitas vezes consideramos que há uma relação de associação entre o comportamento de variáveis, como:

Tempo de estudo x nota na avaliação.
Investimento em publicidade x vendas.
Manutenção preventiva x manutenção corretiva.

Neste tópico vamos apresentar um indicador que nos revela o quão forte ou fraca é essa associação entre as variáveis, chamado de **coeficiente de correlação do produto de momentos de Pearson** ou simplesmente **coeficiente de correlação de Pearson**. Além de representar graficamente o relacionamento entre duas variáveis, por meio do diagrama de dispersão, para visualmente inferirmos sobre a correlação existente.

Coeficiente de correlação de Pearson

Representado pela letra "r" tendo como uma das suas fórmulas:

$$r = \frac{n \cdot \sum(x_i \cdot y_i) - (\sum x_i) \cdot (\sum y_i)}{\sqrt{[n \cdot \sum x_i^2 - (\sum x_i)^2] \cdot [n \cdot \sum y_i^2 - (\sum y_i)^2]}}$$

Em que:

n = tamanho da amostra.

$\sum(x_i \cdot y_i)$ = somatório do produto das variáveis.

$\sum x_i$ = somatório dos valores da variável independente x.

$\sum y_i$ = somatório dos valores da variável dependente y.

$(\sum x_i)^2$ = quadrado do somatório dos valores da variável independente x.

$(\sum y_i)^2$ = quadrado do somatório dos valores da variável dependente y.

$\sum x_i^2$ = somatório dos quadrados dos valores da variável independente x.

$\sum y_i^2$ = somatório dos quadrados dos valores da variável independente y.

Considerações sobre o coeficiente de correlação de Pearson

- O valor do coeficiente de correlação de Pearson pertence ao intervalo $[-1;1]$.
- O coeficiente de correlação de Pearson é adimensional, ou seja, não tem unidade de medida.
- Nos casos em que o coeficiente de correlação linear for igual a zero não indica ausência de correlação e sim ausência de correlação linear, mas as variáveis podem estar correlacionadas exponencialmente, polinomialmente ou logaritmicamente.

Como interpretar o coeficiente de correlação de Pearson:

- Observando o sinal de “r”, temos:

- Se $r > 0$, correlação positiva, indicando uma reta crescente, isto é, à medida que os valores de uma variável aumentam, os valores da outra também aumentam.
- Se $r < 0$, correlação negativa, indicando uma reta decrescente, isto é, à medida que os valores de uma variável aumentam, os valores da outra variável diminuem.

- Observando o valor de “r”, temos:

- $0 \leq |r| < 0,3 \rightarrow$ a correlação entre as variáveis é muito fraca.
- $0,3 < |r| < 0,6 \rightarrow$ a correlação entre as variáveis é média.
- $0,6 < |r| \leq 1 \rightarrow$ a correlação entre as variáveis é forte.



Importante

É importante que saibamos que a existência de correlação não necessariamente implica uma relação de **causalidade**. São coisas distintas, sendo responsabilidade do coeficiente de correlação apenas sinalizar o quanto associados estão os comportamentos de duas variáveis.

Vejamos:

Exemplo 1

horas de estudo X notas

Queremos verificar o quanto estão associadas as variáveis notas das avaliações em Estatística e o tempo de estudo em horas dos alunos. Para isso, coletamos as informações com uma amostra composta de 22 alunos, que estão apresentadas na tabela a seguir:

Aluno	Horas de Estudo	Notas
1	7,5	8,5
2	2	1
3	9	5
4	1,5	3
5	6	7
6	8,5	9
7	3	2
8	4,5	5
9	3,5	4
10	10	9,5
11	8	9
12	4	5
13	5,5	7
14	6,5	7
15	5,5	5,5

16	1	2
17	3	2
18	5	8
19	2	1
20	4	3
21	5	6
22	7	8

Resolução

Vamos definir:

X = número de horas de estudo.

Y= rendimento na avaliação.

Devemos, então, calcular os somatórios presentes na fórmula do coeficiente de correlação linear de Pearson:

Aluno	Horas de Estudo (X)	Notas (Y)	X . Y	X ²	Y ²
1	7,5	8,5	63,75	56,25	72,25
2	2	2	4	4	4
3	9	9	81	81	81
4	1,5	2	3	2,25	4
5	6	7	42	36	49
6	8,5	8	68	72,25	64
7	3	3	9	9	9

8	4,5	5	22,5	20,25	25
9	3,5	4	14	12,25	16
10	10	9,5	95	100	90,25
11	8	9	72	64	81
12	4	5	20	16	25
13	5,5	5	27,5	30,25	25
14	6,5	7	45,5	42,25	49
15	5,5	6	33	30,25	36
16	1	1	1	1	1
17	3	2	6	9	4
18	5	7	35	25	49
19	2	1	2	4	1
20	4	3	12	16	9
21	5	5,5	27,5	25	30,25
22	7	8	56	49	64
Total (Σ)	112	117,5	739,75	705	788,75

$$r = \frac{n \cdot \sum(x_i \cdot y_i) - (\sum x_i) \cdot (\sum y_i)}{\sqrt{[n \cdot \sum x_i^2 - (\sum x_i)^2] \cdot [n \cdot \sum y_i^2 - (\sum y_i)^2]}}$$

$$r = \frac{22 \cdot 739,75 - 112 \cdot 117,5}{\sqrt{[22 \cdot 705 - 112^2] \cdot [22 \cdot 788,75 - 117,5^2]}}$$

$$r = \frac{16274,50 - 13160}{\sqrt{2966 \cdot 3546,25}}$$

$$r = \frac{3114,5}{\sqrt{10518177,5}}$$

$$r = \frac{3114,5}{3243,173985}$$

$$r \cong 0,96$$

Exemplo 2

massa muscular X idade

Acredita-se que, com o passar dos anos, as mulheres vão perdendo massa muscular significativamente. Para avaliar essa hipótese foi feita uma pesquisa em uma amostra de 30 pacientes do sexo feminino, quando foi observada a massa muscular e a idade. O cálculo da massa muscular pode ser feito pela avaliação das dobras cutâneas e também pela fórmula da área muscular do braço. O resultado está apresentado na tabela seguinte.

Calcule o coeficiente de correlação de Pearson e verifique se realmente o índice de massa muscular está correlacionado com a idade da paciente.

Paciente	Idade	Massa muscular
1	45	97
2	56	90
3	65	74
4	58	80
5	49	105
6	62	68
7	51	83
8	53	79
9	71	68

10	69	73
11	58	95
12	73	63
13	53	100
14	45	116
15	76	65
16	69	68
17	74	65
18	67	65
19	56	82
20	63	92
21	44	102
22	72	65
23	78	77
24	68	78
25	67	88
26	39	105
27	51	102
28	64	71
29	60	87
30	40	119

Resolução

Vamos definir:

X = idade da paciente.

Y= massa muscular.

Devemos, então, calcular os somatórios presentes na fórmula do coeficiente de correlação linear de Pearson, como mostra a tabela a seguir.

Paciente	Idade	Massa muscular	x. y	x ²	y ²
1	45	97	4365	2025	9409
2	56	90	5040	3136	8100
3	65	74	4810	4225	5476
4	58	80	4640	3364	6400
5	49	105	5145	2401	11025
6	62	68	4216	3844	4624
7	51	83	4233	2601	6889
8	53	79	4187	2809	6241
9	71	68	4828	5041	4624
10	69	73	5037	4761	5329
11	58	95	5510	3364	9025
12	73	63	4599	5329	3969
13	53	100	5300	2809	10000
14	45	116	5220	2025	13456
15	76	65	4940	5776	4225

16	69	68	4692	4761	4624
17	74	65	4810	5476	4225
18	67	65	4355	4489	4225
19	56	82	4592	3136	6724
20	63	92	5796	3969	8464
21	44	102	4488	1936	10404
22	72	65	4680	5184	4225
23	78	77	6006	6084	5929
24	68	78	5304	4624	6084
25	67	88	5896	4489	7744
26	39	105	4095	1521	11025
27	51	102	5202	2601	10404
28	64	71	4544	4096	5041
29	60	87	5220	3600	7569
30	40	119	4760	1600	14161
Total (Σ)	1796	2522	146510	111076	219640

$$r = \frac{n \cdot \Sigma(x_i \cdot y_i) - (\Sigma x_i) \cdot (\Sigma y_i)}{\sqrt{[n \cdot \Sigma x_i^2 - (\Sigma x_i)^2] \cdot [n \cdot \Sigma y_i^2 - (\Sigma y_i)^2]}}$$

$$r = \frac{30 \cdot 146510 - 2522 \cdot 1796}{\sqrt{[30 \cdot 111076 - (1796)^2] \cdot [30 \cdot 219640 - (2522)^2]}}$$

$$r = \frac{-134212}{\sqrt{106664 \cdot 228716}}$$

$$r = \frac{-134212}{\sqrt{24395763424}}$$

$$r \cong -0,86$$

Como o valor de “r” foi maior do que 0,6 podemos concluir que há forte correlação linear entre as variáveis massa muscular e idade. Ainda observando o sinal negativo de “r”, podemos dizer que à medida que a idade da paciente aumenta, sua massa muscular diminui.

Exemplo 3

Pesquisas sugerem que o tamanho do cérebro de uma pessoa pode estar relacionado com seu grau de inteligência. Para investigar essa hipótese foi feita uma pesquisa com uma amostra aleatória de 40 alunos de um curso autodidata de matemática. A tabela a seguir mostra o tamanho do cérebro em cm³, obtido por ressonância magnética, que será representada pela variável independente (x) e seu respectivo escore de Quociente de Inteligência (QI), que será representado pela variável dependente (y).

Determine o coeficiente de correlação linear de Pearson e verifique se há alguma evidência que possa corroborar essa hipótese. Observe a tabela a seguir.

Observação	Volume cerebral (cm ³)	Escore de QI
1	1270	131
2	1341	138
3	1265	137
4	1233	131
5	1310	135
6	1242	97
7	1251	136

8	1337	90
9	1302	87
10	1347	131
11	1238	130
12	1339	139
13	1346	133
14	1344	138
15	1228	94
16	1340	81
17	1278	130
18	1303	98
19	1336	99
20	1326	78
21	1200	81
22	1329	95
23	1282	133
24	1348	137
25	1347	89
26	1267	139
27	1237	83
28	1318	101

29	1238	75
30	1329	128
31	1240	131
32	1308	142
33	1320	101
34	1306	88
35	1341	81
36	1280	131
37	1348	138
38	1241	86
39	1266	79
40	1299	87

Resolução

Vamos definir:

X = volume cerebral.

Y = Coeficiente de inteligência.

Devemos, então, calcular os somatórios presentes na fórmula do coeficiente de correlação linear de Pearson, como mostra a tabela a seguir.

Observação	volume cerebral (X)	QI (Y)	X . Y	X ²	Y ²
1	1270	131	166370	1612900	17161
2	1341	138	185058	1798281	19044

3	1265	137	173305	1600225	18769
4	1233	131	161523	1520289	17161
5	1310	135	176850	1716100	18225
6	1242	97	120474	1542564	9409
7	1251	136	170136	1565001	18496
8	1337	90	120330	1787569	8100
9	1302	87	113274	1695204	7569
10	1347	131	176457	1814409	17161
11	1238	130	160940	1532644	16900
12	1339	139	186121	1792921	19321
13	1346	133	179018	1811716	17689
14	1344	138	185472	1806336	19044
15	1228	94	115432	1507984	8836
16	1340	81	108540	1795600	6561
17	1278	130	166140	1633284	16900
18	1303	98	127694	1697809	9604
19	1336	99	132264	1784896	9801
20	1326	78	103428	1758276	6084
21	1200	81	97200	1440000	6561
22	1329	95	126255	1766241	9025
23	1282	133	170506	1643524	17689

24	1348	137	184676	1817104	18769
25	1347	89	119883	1814409	7921
26	1267	139	176113	1605289	19321
27	1237	83	102671	1530169	6889
28	1318	101	133118	1737124	10201
29	1238	75	92850	1532644	5625
30	1329	128	170112	1766241	16384
31	1240	131	162440	1537600	17161
32	1308	142	185736	1710864	20164
33	1320	101	133320	1742400	10201
34	1306	88	114928	1705636	7744
35	1341	81	108621	1798281	6561
36	1280	131	167680	1638400	17161
37	1348	138	186024	1817104	19044
38	1241	86	106726	1540081	7396
39	1266	79	100014	1602756	6241
40	1299	87	113013	1687401	7569
Total (Σ)	51820	4458	5780712	67207276	519462

$$r = \frac{n \cdot \sum(x_i \cdot y_i) - (\sum x_i) \cdot (\sum y_i)}{\sqrt{[n \cdot \sum x_i^2 - (\sum x_i)^2] \cdot [n \cdot \sum y_i^2 - (\sum y_i)^2]}}$$

$$r = \frac{40 \cdot 5780712 - 51820 \cdot 4458}{\sqrt{[40 \cdot 67207276 - 51820^2] \cdot [40 \cdot 519462 - 4458^2]}}$$

$$r = \frac{214920}{\sqrt{2978640 \cdot 904716}}$$

$$r = \frac{214920}{\sqrt{2694823266240}}$$

$$r = \frac{214920}{1641591,687}$$

$$r = 0,13$$

Observamos que existe uma fraca correlação entre o volume cerebral e a inteligência, pois o valor de $r < 0,3$.



Ampliando o foco

Philipp Koellinger destaca que, além da questão do tamanho, que conta apenas 2% para a inteligência, existirão vários fatores que explicam os outros 98% na variação de desempenho.

Com a colaboração de três investigadores, Gideon e Philipp incluíram também outras variáveis no estudo, como sexo, idade, altura, estatuto socioeconômico. A altura, por exemplo, está associada a melhores resultados nos testes, mas também a um cérebro maior.

A correlação estudada pelos investigadores já tem vindo a ser investigada há algum tempo, mas com amostras pequenas. Ao incluir mais participantes, ambos esperavam ir além dos estudos já feitos. Dessa forma, recorreram aos dados do Biobank, um banco de dados que contém informações de mais de meio milhão de pessoas do Reino Unido.

As conclusões acabam por ir ao encontro do senso comum: se a pessoa tem mais neurônios, isso permite-lhe ter uma memória melhor, ou realizar mais tarefas em simultâneo.

Existem, no entanto, muitos outros fatores associados, pelo que os investigadores reconhecem que são necessários mais estudos. Relativamente ao sexo, por

exemplo, parece existir uma diferença significativa no tamanho do cérebro entre homens e mulheres, mas isso não se traduz em diferenças no desempenho.

Fonte: <https://www.dn.pt/vida-e-futuro/tamanho-do-cerebro-conta-pouco-para-a-inteligencia-10267115.html>

Cálculo do coeficiente de correlação de Pearson com auxílio da planilha eletrônica Excel

=PEARSON(INÍCIO DA VARIÁVEL INDEPENDENTE : FIM DA VARIÁVEL INDEPENDENTE; INÍCIO DA VARIÁVEL DEPENDENTE : FIM DA VARIÁVEL DEPENDENTE)

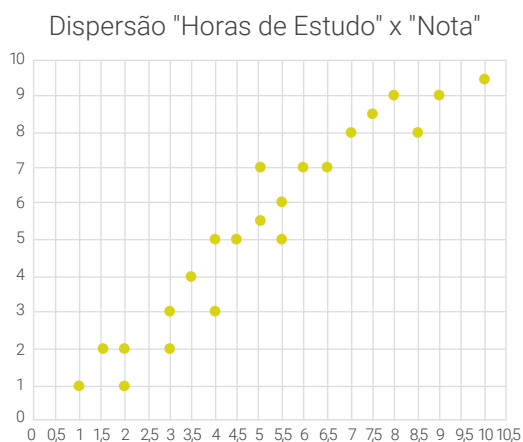
Gráfico de Dispersão (Scatter Plot)

Podemos verificar visualmente o relacionamento entre duas variáveis a partir dos gráficos de dispersão, também conhecidos como Scatter Plot.

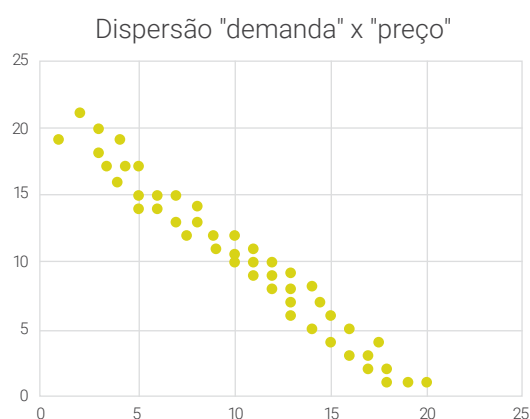
Para traçar um gráfico de dispersão, vamos marcar, no eixo das abscissas, as horas de estudo e, no eixo das ordenadas, a nota na avaliação. Marcamos um ponto cartesiano para cada aluno, relacionando suas horas de estudo com seu rendimento na avaliação.

Vejamos alguns tipos de correlação a seguir.

Correlação linear positiva



Correlação linear negativa

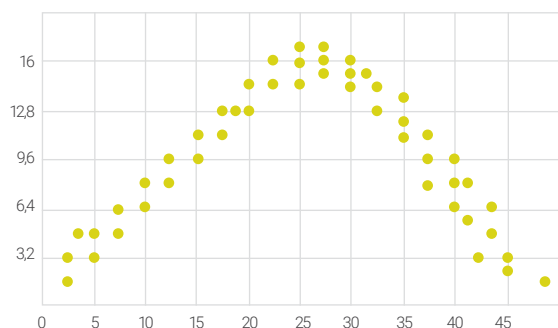


Podemos verificar que os dados apresentam uma tendência de alinhamento em torno de uma reta crescente, indicando que, à medida que o valor de uma variável aumenta, o valor da outra variável também aumenta.

Podemos verificar que os dados apresentam uma tendência de alinhamento em torno de uma reta decrescente, indicando que, à medida que o valor de uma variável aumenta, o valor da outra variável diminui.

Correlação não linear

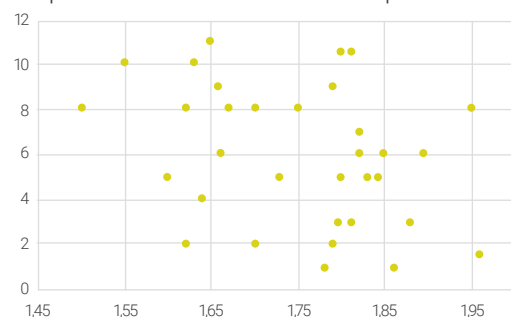
Dispersão "unidade" x "grau de compactação"



Não podemos observar uma tendência de alinhamento em torno de uma reta, mas observamos um alinhamento em torno de uma parábola, o que indica que a correlação entre as variáveis não é linear e sim parabólica.

Não há correlação

Dispersão "estatura" x "sucesso profissional"



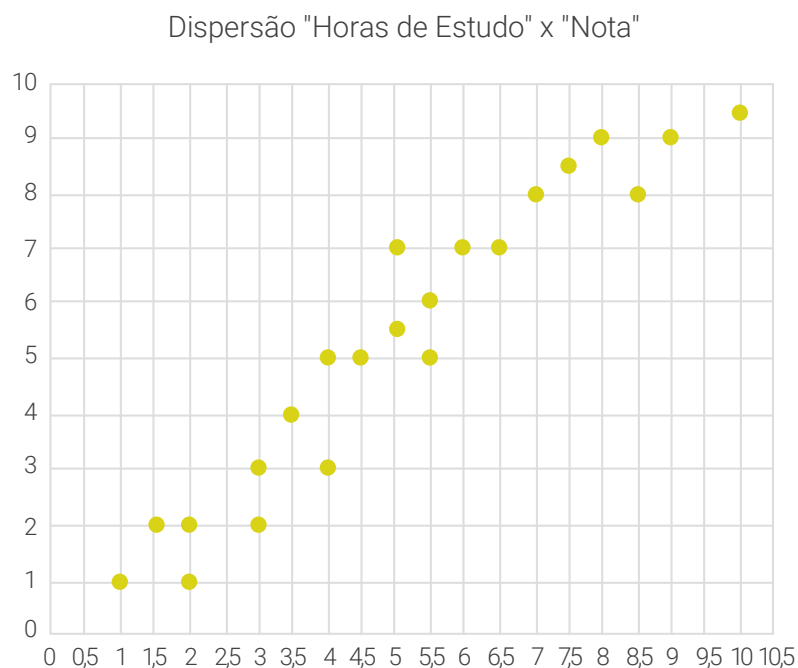
Não podemos observar nenhuma tendência de alinhamento entre os dados, sugerindo não haver correlação entre as variáveis.

Fonte: Elaborado pela autora (2020).

Agora, usando os exemplos anteriores, traçaremos os gráficos de dispersão com base no estudo de correlação.

Exemplo 1: horas de estudo x notas

Aluno	Horas de Estudo (X)	Notas
1	7,5	8,5
2	2	2
3	9	9
4	1,5	2
5	6	7
6	8,5	8
7	3	3
8	4,5	5
9	3,5	4
10	10	9,5
11	8	9
12	4	5
13	5,5	5
14	6,5	7
15	5,5	6
16	1	1
17	3	2
18	5	7
19	2	1
20	4	3
21	5	5,5
22	7	8



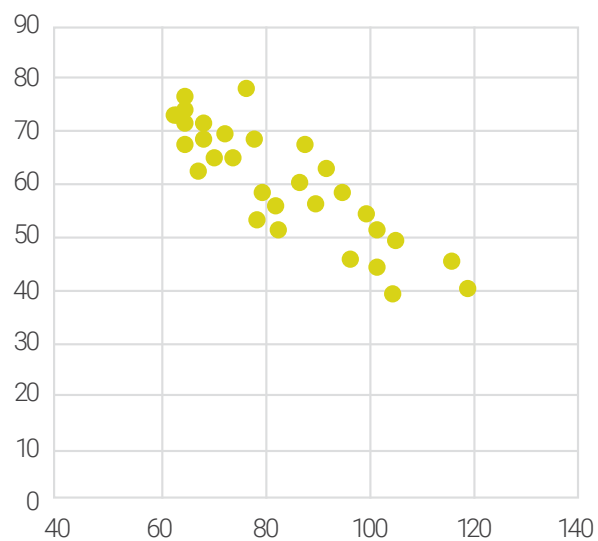
Analisando o comportamento do gráfico podemos verificar que os pontos estão praticamente alinhados em uma reta crescente, indicando que, à medida que uma variável aumenta, a outra também aumenta.

Exemplo 2: idade x massa muscular

Paciente	Idade	Massa muscular
1	45	97
2	56	90
3	65	74
4	58	80
5	49	105
6	62	68
7	51	83
8	53	79
9	71	68
10	69	73
11	58	95

12	73	63
13	53	100
14	45	116
15	76	65
16	69	68
17	74	65
18	67	65
19	56	82
20	63	92
21	44	102
22	72	65
23	78	77
24	68	78
25	67	88
26	39	105
27	51	102
28	64	71
29	60	87
30	40	119

Dispersão Idade x Massa Muscular

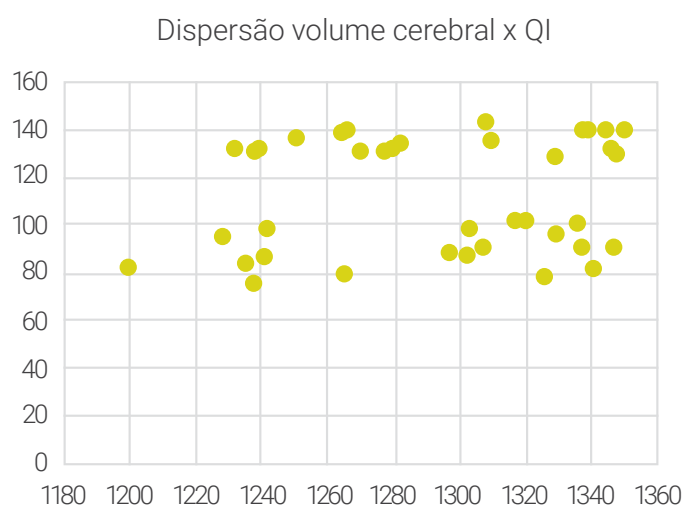


Analisando o comportamento do gráfico podemos verificar que os pontos estão alinhados em uma reta decrescente, o que indica que, à medida que uma variável aumenta, a outra irá diminuir.

Exemplo 3: volume cerebral x QI

Observação	Volume cerebral (cm³) (X)	Escore de QI
1	1270	131
2	1341	138
3	1265	137
4	1233	131
5	1310	135
6	1242	97
7	1251	136
8	1337	90
9	1302	87
10	1347	131
11	1238	130
12	1339	139
13	1346	133
14	1344	138
15	1228	94
16	1340	81
17	1278	130
18	1303	98
19	1336	99
20	1326	78
21	1200	81
22	1329	95
23	1282	133

24	1348	137
25	1347	89
26	1267	139
27	1237	83
28	1318	101
29	1238	75
30	1329	128
31	1240	131
32	1308	142
33	1320	101
34	1306	88
35	1341	81
36	1280	131
37	1348	138
38	1241	86
39	1266	79
40	1299	87



Analisando o comportamento do gráfico podemos verificar que os pontos não estão alinhados, o que indica que as variáveis não estão correlacionadas linearmente.



Ampliando o foco

Para construir um gráfico de dispersão com auxílio da planilha eletrônica Excel, você deve selecionar os valores da variável independente (x) e da variável dependente (y), nessa ordem, e simultaneamente ir em Inserir – Gráfico – Dispersão X Y.

Verifique na midiateca desta unidade um tutorial para construir um gráfico de dispersão com auxílio da planilha eletrônica Excel, adicionando a linha de tendência e o quadrado do coeficiente de correlação de Pearson.

Verificação da Normalidade dos Dados

Encontramos nos softwares estatísticos testes que verificam a hipótese de normalidade dos dados, como os testes de Kolmogorov-Sminorv, Shapiro-Wilk, Anderson-Darling, Ryan-Joiner, entre outros. Entretanto, vamos apresentar aqui três métodos simples que também nos auxiliam na verificação da normalidade dos dados. São eles:

1) Análise do histograma: basta examinar a forma da distribuição, que deve ser semelhante à forma de um sino para que a normalidade pareça razoável.

2) Regra empírica ao contrário: com a média e o desvio-padrão, calcule a proporção real das observações em cada intervalo e compare com os percentuais:

$$[\bar{X} - S ; \bar{X} + S] \rightarrow \frac{\sum f_i}{n} \cong 0,68$$

$$[\bar{X} - 2 \cdot S ; \bar{X} + 2 \cdot S] \rightarrow \frac{\sum f_i}{n} \cong 0,95 \quad \text{para amostras}$$

$$[\bar{X} - 3 \cdot S ; \bar{X} + 3 \cdot S] \rightarrow \frac{\sum f_i}{n} \cong 0,997$$

ou

$$[\mu - \sigma ; \mu + \sigma] \rightarrow \frac{\sum f_i}{n} \cong 0,68$$

$$[\mu - 2 \cdot \sigma ; \mu + 2 \cdot \sigma] \rightarrow \frac{\sum f_i}{n} \cong 0,95 \quad \text{para população}$$

$$[\mu - 3 \cdot \sigma ; \mu + 3 \cdot \sigma] \rightarrow \frac{\sum f_i}{n} \cong 0,997$$

Se os resultados estiverem próximos desses valores, então a normalidade parece razoável.

3) Gráfico de Probabilidade Normal: diagrama de dispersão de cada observação *versus* seu escore normal padronizado correspondente. Se os pontos ficarem alinhados sobre uma reta, então a normalidade parece razoável.

Vejamos um exemplo prático com base nos conceitos apresentados.

Exemplo:

Feita uma pesquisa sobre os gastos mensais com transporte com 20 alunos da Universidade Veiga de Almeida, as respostas já ordenadas estão apresentadas a seguir. Verifique se os dados obtidos obedecem a uma distribuição normal.

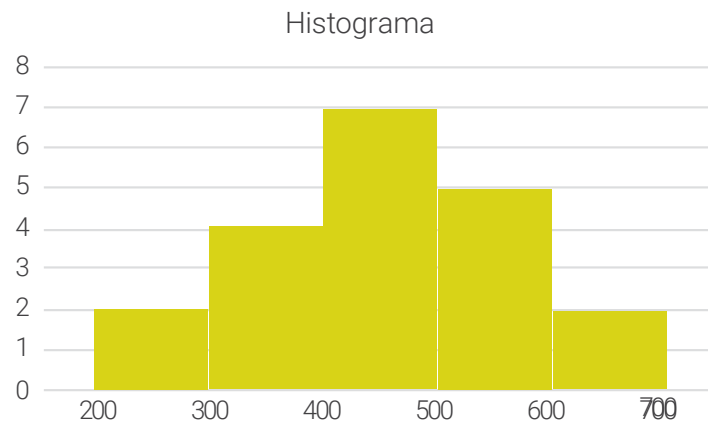
23	29	30	31	35	39	41	42	42	46	47	47	49	50	50	51	56	56	64	67
5	5	0	0	6	9	6	1	7	5	6	7	8	3	9	9	0	8	5	0

Resolução:

1) Análise do histograma:

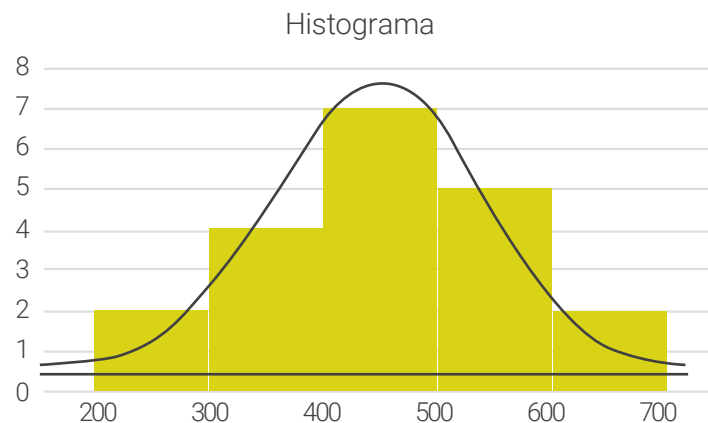
Vamos agrupar os dados em uma tabela de distribuição de frequências:

Gastos R\$	Frequência
200 --- 300	2
300 --- 400	4
400 --- 500	7
500 --- 600	5
600 --- 700	2
Total	20



Fonte: Elaborado pela autora (2020).

Observando o histograma podemos verificar que se aproxima da forma de um sino.



Fonte: Elaborado pela autora (2020).

Pelo formato do histograma parece razoável dizer que os dados seguem uma distribuição normal.

2) Regra empírica ao contrário: com a média e o desvio-padrão, calcule as diferenças: Para a amostra apresentada, podemos calcular o valor da média e do desvio-padrão amostral, obtendo como resultados: $\bar{X} = 452$ e $S = 116$. Logo,

$$[\bar{X} - S ; \bar{X} + S] \rightarrow \frac{\sum f_i}{n} \cong 0,68$$

$$[\bar{X} - 2 \cdot S ; \bar{X} + 2 \cdot S] \rightarrow \frac{\sum f_i}{n} \cong 0,95$$

$$[\bar{X} - 3 \cdot S ; \bar{X} + 3 \cdot S] \rightarrow \frac{\sum f_i}{n} \cong 0,997$$

$$[452 - 116 ; 452 + 116] = [336; 568] \rightarrow \frac{14}{20} = 0,7 \cong 0,68$$

$$[452 - 2 \cdot 116 ; 452 + 2 \cdot 116] = [220; 684] \rightarrow \frac{20}{20} = 1 \cong 0,95$$

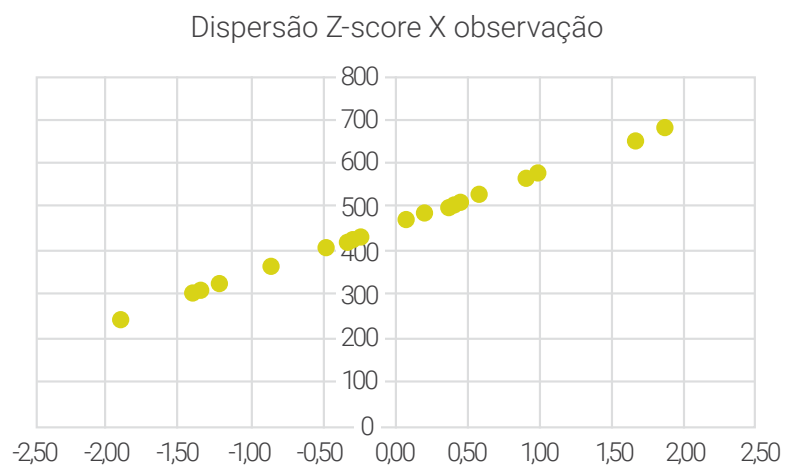
$$[452 - 3 \cdot 116 ; 452 + 3 \cdot 116] = [104; 800] \rightarrow \frac{20}{20} = 1 \cong 0,997$$

Como os quocientes estão próximos dos valores desejáveis, podemos dizer que parece razoável admitir a normalidade dos dados.

3) Gráfico de Probabilidade Normal:

Observação (eixo-y)	Zscore = $(X - \bar{X})/\sigma$ (eixo-x)
235	-1,89
295	-1,37
300	-1,32
310	-1,23
356	-0,83
399	-0,46
416	-0,31
421	-0,27
427	-0,22
465	0,11

476	0,21
477	0,22
498	0,40
503	0,44
509	0,50
519	0,58
560	0,94
568	1,01
645	1,68
670	1,90



Fonte: Elaborado pela autora (2020).

Observando o gráfico concluímos que não há evidência de não normalidade.

Podemos concluir que todos os três métodos nos mostram que não há evidências para rejeitar a normalidade. Isso não significa que podemos dizer com certeza que os dados sejam provenientes de uma distribuição normal.



Ampliando o foco

Para colocar em prática o conteúdo visto nesta unidade, visite a Biblioteca Virtual, escolha um livro de Estatística, selecione e faça alguns exercícios do livro para reforçar o conhecimento sobre o cálculo do coeficiente de correlação linear de Pearson.

Confira as respostas que estão no final do livro e se tiver alguma dúvida fale com o seu tutor pelo Fórum de Dúvidas – Fale com o Tutor.

Regressão linear

A análise de regressão é uma ferramenta estatística usada para prever os valores futuros com base no comportamento de valores passados. Para isso, vamos construir um modelo matemático capaz de descrever o relacionamento entre duas ou mais variáveis quantitativas, de maneira que possamos estimar o valor de uma variável a partir das outras.

Podemos aplicar os modelos de regressão em diversas áreas: engenharia, administração, economia e finanças, marketing, ciências físicas, biologia, ciência ambiental, geografia, saúde pública, nutrição, entre muitas outras áreas de estudo e pesquisa.

Variável independente (X) e variável dependente (Y)

Um conceito importante na análise de regressão é o de variável dependente ou variável independente.

A **variável dependente** é aquela por meio da qual desejamos realizar a previsão e que, portanto, sofre influência do comportamento da **variável independente**.

Regressão Linear Simples

O caso mais simples de regressão é quando temos **duas variáveis** e a relação entre elas pode ser representada por uma **linha reta**. Essa reta é chamada de **modelo de Regressão Linear Simples**, cujo modelo matemático será representado por:

$$y = ax + b + \varepsilon_i$$

Em que:

y = variável dependente, ou seja, o valor que queremos estimar.

x = variável independente, ou seja, aquela que influencia a previsão da variável y .

a = coeficiente angular da reta de regressão.

b = também chamado de intercepto com o eixo-y, é o coeficiente linear da reta.
 ε_i = erro aleatório de “y” para a observação i. É uma variável aleatória com valor esperado igual a zero.



Importante

Antes de determinar os coeficientes para o modelo de regressão linear, é muito conveniente traçar o diagrama de dispersão para verificar visualmente se o relacionamento entre as variáveis aproxima-se de uma reta e também fazer o cálculo do coeficiente de correlação de Pearson para garantir que as variáveis estão correlacionadas e poderão ser ajustadas por um modelo linear.

Determinação dos coeficientes “a” e “b”:

Para construirmos o modelo de regressão linear devemos **determinar os valores dos coeficientes “a” e “b”** de modo que a reta ajuste-se ao conjunto de pontos visualizados no gráfico de dispersão.

O método dos mínimos quadrados utiliza dados amostrais para determinar os valores de “a” e “b”, que minimizam a soma dos quadrados dos desvios entre os valores observados e os valores estimados da variável dependente.

Usando os conceitos do cálculo diferencial pode-se demonstrar que:

$$a = \frac{n \cdot \sum(x_i \cdot y_i) - (\sum x_i) \cdot (\sum y_i)}{n \cdot \sum x_i^2 - (\sum x_i)^2} \quad b = \bar{y} - a \cdot \bar{x} \quad \varepsilon_i = \text{vamos admitir} = 0$$

Logo, substituindo-se os valores de “a” e “b” e “ ε_i ” vamos obter o modelo de regressão linear simples: $y = ax + b$

Em que:

n = tamanho da amostra.

$\sum(x_i \cdot y_i)$ = somatório do produto das variáveis.

$\sum x_i$ = somatório dos valores da variável independente x.

$\sum y_i$ = somatório dos valores da variável dependente y.

$(\sum x_i)^2$ = quadrado do somatório dos valores da variável independente x.

$\sum x_i^2$ = somatório dos quadrados dos valores da variável independente x.

\bar{y} = média dos valores da variável dependente.

\bar{x} = média dos valores da variável independente.

Uma forma alternativa para o coeficiente angular “a”, temos:

$$a = \frac{\sum(x_i - \bar{x}) \cdot (\sum y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Como verificar se um modelo de Regressão Linear está adequado?

Algumas ferramentas estatísticas nos auxiliam na verificação da adequação de um modelo de regressão linear, como o gráfico dos resíduos e o coeficiente de determinação (r^2), que nos indica o percentual de eficácia do modelo. Quanto mais próximo de 1 (100%) mais bem ajustado poderá estar o modelo. Para maior confiança, é preciso analisar o valor de r^2 juntamente com o gráfico de resíduos.

Determinação do modelo de regressão linear com auxílio do Excel

Para determinar rapidamente o modelo de regressão linear com auxílio da planilha eletrônica Excel, basta clicar nos pontos do gráfico de dispersão; com o botão direito do mouse clicar em “adicionar linha de tendência” e no menu “formatar linha de tendência” selecionar “exibir equação no gráfico”.

Vejamos alguns exemplos práticos a seguir.

Exemplo 1

Voltemos ao Exemplo 1 usado no Tópico 2: horas de estudo X notas.

Queremos verificar o quanto estão associadas as variáveis notas das avaliações em Estatística e o tempo de estudo em horas dos alunos. Para isso, coletamos as informações com uma amostra composta de 22 alunos, que estão apresentadas na tabela a seguir:

Aluno	Horas de Estudo	Notas
1	7,5	8,5
2	2	2
3	9	9
4	1,5	2
5	6	7
6	8,5	8
7	3	3
8	4,5	5
9	3,5	4
10	10	9,5
11	8	9
12	4	5
13	5,5	5
14	6,5	7
15	5,5	6
16	1	1
17	3	2
18	5	7
19	2	1
20	4	3
21	5	5,5
22	7	8

Determine o modelo de regressão linear para as variáveis notas das avaliações em Estatística e o tempo de estudo em horas dos alunos.

Resolução

a) Como já verificamos, as variáveis estão fortemente relacionadas pois, ao calcularmos o coeficiente de correlação de Pearson, encontramos o valor de $r = 0,96$, portanto $r > 0,6$ indicando uma forte correlação entre as variáveis.

Logo, podemos determinar o modelo de regressão linear $y = ax + b$.

Precisamos, então, calcular o coeficiente angular e o coeficiente linear da reta. Pelo método dos mínimos quadrados temos:

$$a = \frac{n \cdot \sum(x_i \cdot y_i) - (\sum x_i) \cdot (\sum y_i)}{n \cdot \sum x_i^2 - (\sum x_i)^2}$$

$$b = \bar{y} - a \cdot \bar{x}$$

Calculando o coeficiente angular “a”:

Aluno	Horas de Estudo (X)	Notas (Y)	X . Y	X ²	Y ²
1	7,5	8,5	63,75	56,25	72,25
2	2	2	4	4	4
3	9	9	81	81	81
4	1,5	2	3	2,25	4
5	6	7	42	36	49
6	8,5	8	68	72,25	64
7	3	3	9	9	9
8	4,5	5	22,5	20,25	25
9	3,5	4	14	12,25	16
10	10	9,5	95	100	90,25

11	8	9	72	64	81
12	4	5	20	16	25
13	5,5	5	27,5	30,25	25
14	6,5	7	45,5	42,25	49
15	5,5	6	33	30,25	36
16	1	1	1	1	1
17	3	2	6	9	4
18	5	7	35	25	49
19	2	1	2	4	1
20	4	3	12	16	9
21	5	5,5	27,5	25	30,25
22	7	8	56	49	64
Total (Σ)	112	117,5	739,75	705	788,75

$$a = \frac{n \cdot \sum(x_i \cdot y_i) - (\sum x_i) \cdot (\sum y_i)}{n \cdot \sum x_i^2 - (\sum x_i)^2}$$

$$a = \frac{22 \cdot 739,75 - 112 \cdot 117,5}{22 \cdot 705 - 112^2} = \frac{3114,5}{2966} = 1,0501$$

Calculando o coeficiente linear “b”.

$$b = \bar{y} - a \cdot \bar{x}$$

$$b = \frac{117,5}{22} - 1,0501 \cdot \frac{112}{22}$$

$$b = 5,34090909 - 1,0501 \cdot 5,09090909$$

$$b = 5,34090909 - 5,345797734 = -0,049$$

Logo o modelo de regressão $y = ax + b$ será $y = 1,0501x - 0,049$.

Exemplo 2

Voltemos ao Exemplo 2 do Tópico 2: massa muscular X idade.

Acredita-se que, com o passar dos anos, as mulheres vão perdendo massa muscular significativamente. Para avaliar essa hipótese foi feita uma pesquisa em uma amostra de 30 pacientes do sexo feminino, quando foram observadas a massa muscular e a idade. O cálculo da massa muscular pode ser feito pela avaliação das dobras cutâneas e também pela fórmula da área muscular do braço. O resultado está apresentado na tabela a seguir.

Calcule o coeficiente de correlação de Pearson e verifique se realmente o índice de massa muscular está correlacionado com a idade da paciente.

Paciente	Idade	Massa muscular
1	45	97
2	56	90
3	65	74
4	58	80
5	49	105
6	62	68
7	51	83
8	53	79
9	71	68
10	69	73
11	58	95
12	73	63
13	53	100
14	45	116

15	76	65
16	69	68
17	74	65
18	67	65
19	56	82
20	63	92
21	44	102
22	72	65
23	78	77
24	68	78
25	67	88
26	39	105
27	51	102
28	64	71
29	60	87
30	40	119

Resolução

b) Como já verificamos, as variáveis estão fortemente relacionadas pois, ao calcularmos o coeficiente de correlação de Pearson, encontramos o valor de $r = -0,86$, portanto $r > 0,6$ indicando uma forte correlação entre as variáveis.

Logo, podemos determinar o modelo de regressão linear $y = ax + b$.

Precisamos, então, calcular o coeficiente angular e o coeficiente linear da reta. Pelo método dos mínimos quadrados temos:

$$a = \frac{n \cdot \sum(x_i \cdot y_i) - (\sum x_i) \cdot (\sum y_i)}{n \cdot \sum x_i^2 - (\sum x_i)^2}$$

$$b = \bar{y} - a \cdot \bar{x}$$

Calculando o coeficiente angular "a":

Paciente	Idade	Massa muscular	x. y	x ²	y ²
1	45	97	4365	2025	9409
2	56	90	5040	3136	8100
3	65	74	4810	4225	5476
4	58	80	4640	3364	6400
5	49	105	5145	2401	11025
6	62	68	4216	3844	4624
7	51	83	4233	2601	6889
8	53	79	4187	2809	6241
9	71	68	4828	5041	4624
10	69	73	5037	4761	5329
11	58	95	5510	3364	9025
12	73	63	4599	5329	3969
13	53	100	5300	2809	10000
14	45	116	5220	2025	13456
15	76	65	4940	5776	4225
16	69	68	4692	4761	4624
17	74	65	4810	5476	4225
18	67	65	4355	4489	4225
19	56	82	4592	3136	6724

20	63	92	5796	3969	8464
21	44	102	4488	1936	10404
22	72	65	4680	5184	4225
23	78	77	6006	6084	5929
24	68	78	5304	4624	6084
25	67	88	5896	4489	7744
26	39	105	4095	1521	11025
27	51	102	5202	2601	10404
28	64	71	4544	4096	5041
29	60	87	5220	3600	7569
30	40	119	4760	1600	14161
Total (Σ)	1796	2522	146510	111076	219640

$$a = \frac{n \cdot \sum(x_i \cdot y_i) - (\sum x_i) \cdot (\sum y_i)}{n \cdot \sum x_i^2 - (\sum x_i)^2}$$

$$a = \frac{30 \cdot 146510 - 1796 \cdot 2522}{30 \cdot 111076 - (1796)^2}$$

$$a = \frac{-134212}{106664} = -1,2583$$

Calculando o coeficiente linear “b”.

$$b = \bar{y} - a \cdot \bar{x}$$

$$b = \frac{2522}{30} - (-1,2583) \cdot \frac{1796}{30}$$

$$b = 84,0666666667 + 1,2583 \cdot 59,8666666667$$

$$b = 84,0666666667 + 75,33022667$$

$$b = 159,40$$

Logo o modelo de regressão $y = ax + b$ será: $y = -1,2583x + 159,40$.



Ampliando o foco

Para colocar em prática o conteúdo visto nesta unidade, visite a Biblioteca Virtual, escolha um livro de Estatística, selecione e faça alguns exercícios do livro para reforçar o conhecimento sobre a determinação do modelo de Regressão Linear.

Confira as respostas que estão no final do livro e se tiver alguma dúvida fale com o seu tutor pelo Fórum de Dúvidas – Fale com o Tutor.



MIDIAATECA

Para ampliar o seu conhecimento veja o material complementar da Unidade 4, disponível na midiateca.



NA PRÁTICA

Vamos refletir: Os moinhos de vento são muito barulhentos?

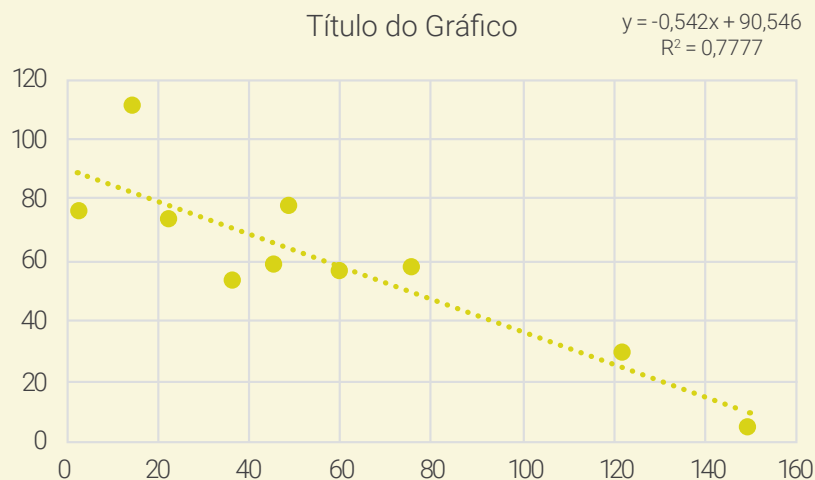
Os moinhos de vento modernos estão se tornando uma alternativa eficiente e limpa para produção de energia. Embora um moinho de vento eficaz possa utilizar vento de baixa velocidade, os locais ideais para parques eólicos são ao longo da costa oceânica ou em áreas montanhosas, onde há uma velocidade de vento consistente de pelo menos 24 km/h.

Os moinhos de vento não podem ser construídos próximos a cidades devido às leis que regulam o nível de ruído. Um moinho de vento típico produz aproximadamente 56dB a 60 metros de distância. Isso é mais suave do que o som da fala humana (que está em cerca de 70dB).

Suponha que uma pequena comunidade costeira esteja considerando a construção de um moinho de vento para gerar eletricidade para a prefeitura. Um experimento é conduzido para medir o nível de ruído do moinho de vento (em dB) a várias distâncias (em metros) do local proposto. Os dados estão resumidos na tabela a seguir.

Distância (m)	3	15	23	37	46	49	61	76	122	150
Nível de Ruído (dB)	75	110	73	52	58	77	56	57	28	4

Com as técnicas apresentadas nesta unidade podemos verificar facilmente que há uma relação linear significativa entre a variável independente “distância” e a variável dependente “nível de ruído”.



Fonte: Elaborado pela autora (2020).

Também podemos usar a análise de regressão que estudamos nesta unidade para prever um valor do nível de ruído para uma determinada distância do moinho de vento. Utilizando os dados levantados sobre o nível de ruído e a distância chegamos ao seguinte modelo matemático:

$$y = -0,542x + 90,546$$

Suponhamos que um moinho de vento esteja sendo construído a 100m de sua residência. Podemos prever, com auxílio do modelo, que, para essa distância, o nível de ruído está próximo de 36 dB (o que está entre um suave sussurro e um escritório tranquilo).

Fonte: Adaptado de Kokoska (2013).

Resumo da Unidade 4

Nesta unidade aprofundamos nossos conhecimentos na Estatística Indutiva. Iniciamos apresentando os conceitos de Intervalos de Confiança, em que a partir dos dados coletados em uma amostra foi possível inferir um intervalo sob um certo grau de confiança, que continha o verdadeiro valor do parâmetro, como se tivéssemos feito nossa pesquisa com todos os elementos da população.

Já familiarizados com o uso da tabela normal padronizada, conhecemos, então, o tão esperado cálculo do tamanho de uma amostra. Também apresentamos os conceitos do coeficiente de correlação linear de Pearson, que nos auxiliou na verificação da força com que uma variável está relacionada com a outra. Apresentamos, ainda, o modelo de Regressão Linear, por meio do qual foi possível estimar a equação de uma reta que melhor descreveria a relação entre uma variável dependente e uma variável independente.

Acreditamos, fortemente, que ao chegar até aqui você esteja familiarizado com os conceitos fundamentais, tanto da Estatística Descritiva quanto da Estatística Indutiva. Ainda temos uma vasta gama de conceitos, mas com as ferramentas adquiridas nesse curso você conseguirá facilmente assimilar os novos conhecimentos.

Esse foi o pontapé inicial para mergulharmos no fascinante mundo da Estatística.

Saudações!

Referências

KOKOSKA, S. **Introdução à estatística**: uma abordagem por resolução de problemas. Rio de Janeiro: LTC, 2013.

LARSON, R.; FARBER, B. **Estatística aplicada**. São Paulo: Pearson, 2010.

NEXO Jornal. **Como funciona uma pesquisa eleitoral**. YouTube. Disponível em: <https://youtu.be/igi7E1OY7gs>. Acesso em: 19 set. 2020.

SWEENEY, D. J.; WILLIAMS, T. A.; ANDERSON, D. R. **Estatística aplicada à administração e economia**. 3. ed. São Paulo: Cengage Learning, 2013.

TABELA de Distribuição Normal Padronizada. Disponível em: http://ead.uva.br///recurso/DEF/EST/u3_c3_r1/anexo/Tabela-da-Distribuicao-Normal.pdf. Acesso em: 5 set. 2020.

TABELA T-student. Disponível em: http://ead.uva.br///recurso/DEF/EST/u4_c1_r1/index.htm. Acesso em: 19 set. 2020.

TEC Concursos. Intervalo de confiança para a média. #720037 Pró-Reitoria GP CP2 - Estatístico (CP II)/2017. Disponível em: <https://www.tecconcursos.com.br/questoes/720037>. Acesso em: 5 out. 2020.

