KLS

Probabilidade e Estatística

Probabilidade e estatística

Thatiane Cristina dos Santos de Carvalho Ribeiro

Presidente: Rodrigo Galindo
Vice-Presidente Acadêmico de Graduação: Rui Fava
Gerente Sênior de Editoração e Disponibilização de Material Didático: Emanuel Santana
Gerente de Revisão: Cristiane Lisandra Danna
Coordenação de Produção: André Augusto de Andrade Ramos
Coordenação de Disponibilização: Daniel Roggeri Rosa
Editoração e Diagramação: eGTB Editora

Dados Internacionais de Catalogação na Publicação (CIP)

Ribeiro-Santos, Thatiane Cristina dos Santos de Carvalho
R484p Probabilidade e estatística / Thatiane Cristina dos Santos de Carvalho Ribeiro. – Londrina : Editora e Distribuidora Educacional S.A., 2015.
216 p.

ISBN 978-85-8482-225-6

1. Estatística. 2. Estatística matemática. 3. Probabilidades. I. Título.

CDD 519.5

2015
Editora e Distribuidora Educacional S.A.
Avenida Paris, 675 – Parque Residencial João Piza
CEP: 86041 -100 — Londrina — PR
e-mail: editora.educacional@kroton.com.br
Homepage: http://www.kroton.com.br/

Sumário

Palavras do autor

Caro aluno, vamos começar o estudo sobre Probabilidade e Estatística, você está preparado?

Em nossa vida cotidiana, utilizamos constantemente os conceitos de estatística e probabilidade. Você conhecerá os conceitos básicos de estatística e probabilidade aplicados às diversas áreas do conhecimento. Tal conhecimento será muito importante na sua vida profissional, no tratamento e interpretação de dados. As tomadas de decisões podem ser realizadas segundo as análises de dados coletados em uma pesquisa, por exemplo. Os relatórios construídos a partir das análises dos dados e gráficos são necessários para embasar as tomadas de decisões.

Através do autoestudo, você conhecerá conceitos sobre probabilidade e estatística e será capaz de utilizá-los para resolver problemas que enfrentamos em nosso dia a dia. Os meios de comunicação utilizam as pesquisas e as interpretações destes dados para mostrar como está, por exemplo, a economia de um país.

Na primeira unidade será apresentada a introdução à estatística, os conceitos básicos e como são importantes para o desenvolvimento do nosso estudo. Os métodos numéricos e como são tratados os dados serão detalhados nesta unidade.

Na segunda unidade serão apresentados os métodos de tratamentos das informações que foram coletadas na primeira unidade, que são os Métodos Tabulares e os Métodos Gráficos.

Na terceira unidade estudaremos os conceitos sobre Probabilidade, iniciaremos com as Distribuições de Probabilidade Discretas, que utilizam quantidades aleatórias de dados com valores finitos para resolver uma incerteza presente em uma determinada situação.

Na quarta unidade as Distribuições de Probabilidade Contínua são abordadas. As quantidades de dados tratados neste tipo de distribuição são aleatórias e contínuas, com um número infinito de valores.

Ao final do seu estudo, você terá conhecido os fundamentos de Probabilidade e Estatística que são necessários para sua formação.

Agora é com você.

Você está preparado?

Bons estudos e boa sorte na sua caminhada pelo universo da Probabilidade e Estatística!

Estatística, como ela influencia sua vida?

Convite ao estudo

Nesta unidade, veremos conteúdos que são necessários para a realização do método estatístico. Para alcançarmos a competência de fundamento de área da disciplina, que é conhecer os fundamentos estatísticos básicos necessários para a formação profissional, teremos alguns objetivos de aprendizagem nesta unidade, que o auxiliarão nesta jornada de estudo. São objetivos de aprendizagem desta unidade: trabalhar as informações coletadas de processos de pesquisa, analisar tabelas e gráficos com as informações coletadas, evidenciar a importância da Estatística na vida diária e demostrar como podemos utilizá-la de forma correta.

Os meios de comunicação utilizam com grande frequência as apresentações de estatística. Os jornais fazem uso dos gráficos e tabelas gerados a partir das pesquisas estatísticas sobre pessoas e empresas. Você já se perguntou como nossa vida é influenciada diretamente pelas informações apresentadas pelos meios de comunicação?

Uma agência de pesquisa foi contratada por uma empresa multinacional do setor de alimentos e precisa entrevistar gerentes de contratação.

O relatório que a empresa contratante necessita deverá conter os seguintes itens:

- Tabelas, cálculos de porcentagem e gráficos das respostas dos entrevistados.
- As respostas devem ser divididas em amostras e a empresa precisa saber o tipo de amostragem utilizada na pesquisa.

- Cálculos de média, mediana e moda dos resultados obtidos.
- Representação gráfica de assimetria e curtose deverão auxiliar na construção de uma conclusão sobre a pesquisa.

Você será capaz, ao final desta unidade, de elaborar o relatório, trabalhando as respostas obtidas com a pesquisa. Será possível, na conclusão, evidenciar a importância de se trabalhar os dados com o uso da Estatística.

Com o estudo da unidade, poderemos esclarecer algumas perguntas: como as pesquisas devem ser feitas? Qual a quantidade de dados necessários? Quais as conclusões que devemos chegar ao analisar os resultados da pesquisa?

Vamos lá?

Seção 1.1

Introdução à estatística

Diálogo aberto

A palavra estatística indica, à maioria das pessoas, levantamento ou recenseamento. Os governos utilizam os censos há milhares de anos, com o intuito de conhecer seus habitantes e suas características, como: economia, religião, cultura, entre outras.

Reflita

"No futuro, o pensamento estatístico será tão necessário para a cidadania eficiente como saber ler e escrever." H.G. Wells (escritor, autor de "A Guerra dos Mundos" e "A Máquina do Tempo")

Status é a origem das palavras ***estatística*** e ***estado***, é uma palavra de origem latina.

A estatística é necessária para a análise de dados derivada de quaisquer processos com variação.

Estatísticas (plural) têm denotação de qualquer quantidade de dados numéricos, formados a fim de fornecer informações acerca de uma atividade qualquer. Como exemplo, temos as estatísticas demográficas que englobam dados como nascimento, falecimento e casamentos de uma população. Já no singular, **Estatística** mostra a atividade de um corpo especializado por técnicas ou uma metodologia para a coleta, a classificação, a apresentação, a análise e a interpretação de dados quantitativos e a utilização desses dados para a tomada de decisões.

Importância da Estatística

Para solucionarmos a maioria dos problemas do mundo, precisamos de dados ou informações. Mas que tipo de informação? Qual é o número de informações? Após obtê-las, o que fazer com elas?

A Estatística associa os dados ao problema, trabalha as informações necessárias, projeta como e o que deve ser coletado e capacita o pesquisador (ou profissional ou cientista) a obter conclusões a partir dessas informações, para que outras pessoas entendam tais resultados. Portanto, o cientista social, o economista, o engenheiro, o agrônomo e muitos outros profissionais utilizam os métodos estatísticos para auxiliar na realização do seu trabalho, promovendo maior eficiência.

Uma agência de pesquisa foi contratada por uma empresa multinacional do setor de alimentos e precisa entrevistar gerentes de contratação e fazer a seguinte pergunta:

O que os empregadores procuram em um trabalhador temporário?

Você é funcionário da agência e foi designado para realizar a pesquisa. Para isso, você deve definir uma população, uma amostra, tabular os dados coletados e analisar as respostas, gerando, assim, as conclusões sobre a pesquisa.

Nesta seção, você estudará os conceitos e as fases do método estatístico que o ajudarão a trabalhar as informações necessárias para o desenvolvimento da pesquisa e entrega de conclusões. Essas são importantes, pois norteiam as tomadas de decisões dos órgãos que contrataram a pesquisa, neste caso, a empresa multinacional do setor de alimentos.

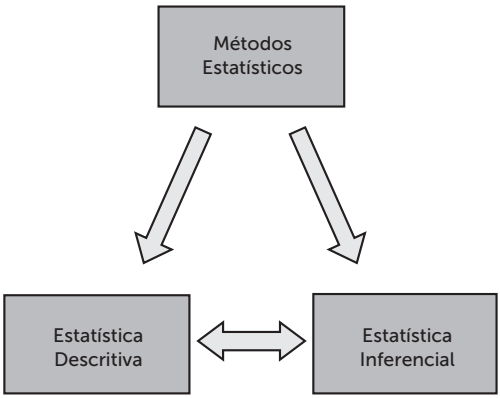
Figura 1.1 | Multinacional setor de alimentos

Fonte: Disponível em: <http://pt.wikipedia.org/wiki/Unilever#mediaviewer/File:Lever.jpg>. Acesso em: 30 maio 2015.

Assimile

A Estatística é uma parte da Matemática que fornece métodos para a coleta, organização, descrição, análise e interpretação de dados, viabilizando a utilização deles na tomada de decisões.

Figura 1.2 | Divisões dos Métodos Estatísticos



Fonte: O autor (2015).

Grandes áreas da Estatística

A Estatística pode ser dividida em duas partes:

Quadro 1.1 | Grandes áreas da estatística

Estatística Descritiva	Estatística Inferencial
Conjunto de técnicas que tem a função de coletar, organizar, apresentar, analisar e sintetizar os dados numéricos de uma população, ou amostra.	Processo de se obter informações sobre uma população a partir de resultados observados na amostra.

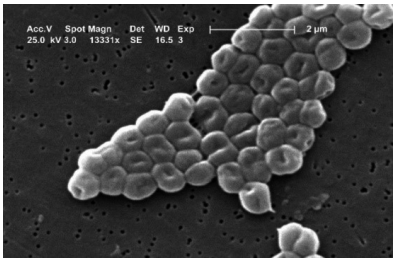
Fonte: Adaptado de Moretin (2010).

Em estatística, utilizamos extensamente os termos: população, amostra, censo, parâmetros, dados discretos, dados contínuos, dados quantitativos e dados qualitativos; e definiremos cada um deles para sua maior compreensão.

• **População** - É uma coleção completa de todos os elementos a serem estudados.

Exemplos:

Figura 1.3 | Bactérias do corpo humano. População: Todas as bactérias existentes no corpo humano



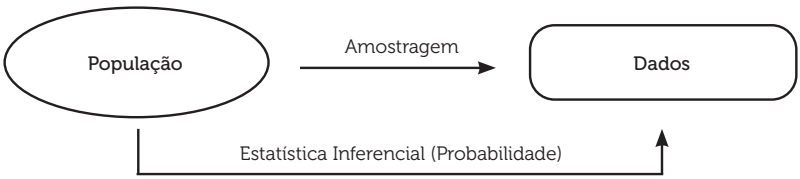
Fonte: Disponível em: http://pt.wikipedia.org/wiki/Acinetobacter_baumannii. Acesso em: 30 maio 2015.

Figura 1.4 | Comportamento das formigas de certa área. População: Todas as formigas da área em estudo

Fonte: Disponível em: http://commons.wikimedia.org/wiki/File:Meat_eater_ant_nest_swarming03.jpg. Acesso em: 30 maio 2015.

• **Amostra** - Refere-se a uma parte da população que tem pelo menos uma característica em comum, relacionada ao fato que se deseja pesquisar. A partir das amostras é possível fazer inferências que servirão de base para a tomada de decisões.

Figura 1.5 | Relação entre população e amostra



Fonte: O autor (2015).

• **Censo** - É o exame completo de toda população. Abrange todos os dados relativos a todos os elementos da população.

• **Parâmetros** - São números que descrevem características da população. A média populacional e o desvio-padrão populacional são exemplos de parâmetros.

• **Dados contínuos** - Um número infinito de valores possíveis sem que haja falhas ou lacunas entre eles; representam os dados contínuos, os quais incluem todos os valores de um intervalo numérico.

• **Dados discretos** - Um número finito de valores possíveis que representam os dados discretos. É uma quantidade "mensurável",

como os valores 0, 1, 2 e assim por diante. Exemplo: os números de ovos que as galinhas botam são dados discretos, porque representam contagens.

Figura 1.6 | Ovos de galinha – Dados Discretos

Fonte: Disponível em: <http://commons.wikimedia.org/wiki/File:Chicken_eggs.jpg>. Acesso em: 30 maio 2015.

• **Dados quantitativos** - Representam contagem ou medida. As unidades medidas são importantes quando trabalhamos com dados quantitativos.

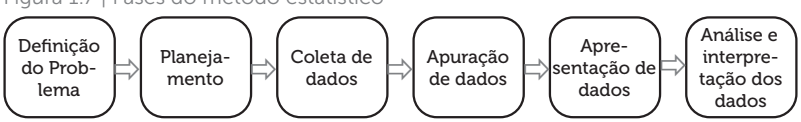
• **Dados qualitativos** - São dados de características não numéricas que podem ser separados em categorias distintas.

Os dados podem ser **absolutos**, quando são resultados de uma coleta que foi realizada diretamente na fonte. Outro tipo de dados são os **relativos**, que resultam das divisões entre dados absolutos e seu total, evidenciando ou facilitando as comparações entre quantidades e o seu total.

Fases do Método Estatístico

Quando se pretende empreender um estudo estatístico completo, existem diversas fases do trabalho que devem ser desenvolvidas para se chegar aos resultados finais de um estudo capaz de produzir resultados válidos.

Figura 1.7 | Fases do método estatístico



Fonte: Adaptado de Moretin (2010).

As fases principais são as seguintes:

Quadro 1.2 | Fases

Definição do problema	Consiste na definição ou formulação de forma correta do problema que será estudado. Definir o objeto de estudo auxilia o analista a fazer o levantamento sobre a população a ser estudada.
Planejamento	Compreende em determinar o procedimento que é necessário para a resolução do problema, como levantar o assunto ou o objeto de estudo. Nessa fase o tipo de levantamento a ser utilizado é escolhido. Há dois tipos de levantamento: levantamento censitário: considera toda a população, e levantamento por amostragem: utiliza uma parte da população ou uma amostra.
Coleta dos dados	É um passo operacional, em que é feita a coleta de informações. Nesta fase há a necessidade de estabelecer uma distinção entre os dados primários e os dados secundários. O primeiro trata de dados coletados pelo próprio pesquisador ou instituição de pesquisa. O segundo são dados oriundos de outra instituição.
Apuração dos dados	Consiste em resumir os dados através de sua contagem e agrupamento. Pode ser manual, eletromecânica ou eletrônica.
Apresentação dos dados	Os dados devem ser apresentados de forma que o exame do fenômeno estudado pelo método estatístico seja facilmente identificado. Pode ser de duas maneiras: Apresentação Tabular – apresentação numérica dos dados, disposta em uma tabela com linhas e colunas. Apresentação gráfica – apresentação dos dados numéricos de forma geométrica, permite observar o fenômeno estudado e suas variações de maneira mais rápida.
Análise e interpretação dos dados	São as conclusões que irão auxiliar o pesquisador na resolução do problema. As estatísticas apresentadas pelas outras fases evidenciam as características da população e são importantes para a tomada de decisão. As generalizações nessa fase são possíveis, contudo podem gerar certo grau de incerteza, mas não devem comprometer o resultado final. Dessa forma, a pesquisa será representativa, isto é, mostrará a realidade da população.

Fonte: Adaptado de Moretin (2010).

Complemente seus estudos

Para aprofundar seus conhecimentos sobre as fases do método estatístico, indicamos a seguinte leitura complementar: Disponível em:

<http://www.aedmoodle.ufpa.br/pluginfile.php?file=%2F61912%2Fmod_resource%2Fcontent%2F0%2FESTATISTICA_APLICADA_-_PROF._JOAO_FURTADO.pdf>. Acesso em: 30 maio 2015.

Séries Estatísticas

A série estatística é definida como todo e qualquer conjunto de dados estatísticos que se refere à classificação de dados quantitativos.

A palavra série é usada normalmente para designar um conjunto de dados dispostos de acordo com um caráter variável, residindo a qualidade serial na disposição desses valores, e não em uma disposição temporal ou espacial de indivíduos.

Tabela é um quadro que resume um conjunto de observações. A Figura 1.8 mostra uma tabela.

Figura 1.8 | Exemplo de tabela de pesquisa

TABELA 15 - INDICADORES DA PRODUÇÃO ACADÊMICA (2002-2008)

Atividades de Ensino	2002	2003	2004	2005	2006	2007	2008*
Produção Artística	227	290	375	708	567	852	637
Trabalhos Técnicos	528	465	671	784	999	1.051	833
Projetos com Financiamento	3.728	3.750	3.755	3.921	4.322	4.562	4.858
Linhas de Pesquisa	1.153	1.016	1.017	1.042	1.070	1.097	1.048
Participação em Congressos e Eventos	4.129	3.584	3.591	4.671	5.664	5.885	5.955
Promoção de Eventos	1.336	1.106	1.459	1.389	1.151	831	907

Fonte: Disponível em: <http://www.unicamp.br/unicamp/unicamp_hoje/divulgacao/gestao2005_09/cap2_pesquisa-desenvolvimento.php>. Acesso em: 30 maio 2015.

As tabelas servem para apresentar séries estatísticas. Conforme a variação de um dos elementos da série, podemos classificá-las em:

Quadro 1.3 | Séries estatísticas

Cronológicas	Quando se trata de tempo – quando ocorre o fato.
Geográficas	Local (fator espacial ou geográfico) – onde o fenômeno acontece.
Específicas	Fenômeno (espécie do fato ou fator especificativo) – o que é descrito.

Fonte: Adaptado de Moretin (2010).

Podem ser séries de dois tipos: a série variável, ou descontínua, é chamada de Homógrada. Já as séries que têm subdivisões ou graduações são chamadas Heterógradas.

A maior parte das pesquisas utiliza tais conceitos para fazer uma análise de acordo com uma característica de uma amostra. Os gráficos são utilizados para mostrar, de forma visual, os resultados da pesquisa e possibilitam nossa análise visual.

Os meios de comunicação utilizam muito as apresentações de estatística. Os jornais fazem uso dos gráficos e tabelas gerados a partir das pesquisas estatísticas sobre pessoas e empresas.

Você já se perguntou como nossa vida é influenciada diretamente por informações apresentadas por estes veículos de informação?

Vocabulário

Censo - Recenseamento da população.

Cronograma - Representação gráfica do calendário de um plano ou projeto.

Inferencial - Relativo à inferência. Ato ou efeito de inferir. Indução, conclusão.

Quantitativa - Que exprime ou determina quantidade.

Sem medo de errar

Agora vamos construir um gráfico com os números de gerentes que responderam à pergunta da Agência de Pesquisa:

$$1.\ \text{número de respostas} = \frac{porcentagem\ de\ respostas}{100} \times população$$

$$2.\ \text{experiência anterior} = \frac{20}{100} \times 1043 \cong 209\ gerentes$$

$$3.\ \text{possibilidade de trabalhar ...} = \frac{28}{100} \times 1043 \cong 292\ gerentes$$

$$4.\ \text{compromisso} \ldots = \frac{13}{100} \times 1043 \cong 135\ gerentes$$

$$5.\ \text{atitude positiva} \ldots = \frac{39}{100} \times 1043 \cong 407\ gerentes$$

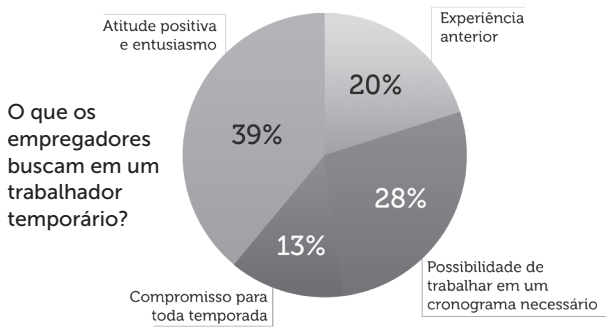
Construindo o gráfico de quantidade de gerentes, leve o Quadro 1.4 a seguir para o Excel, e escolha o gráfico tipo pizza para expressar as porcentagens de respostas levantadas com a pesquisa.

Quadro 1.4 | Pesquisa com empregados temporários

Pesquisa com Empregados Temporários	
Perguntas Realizadas	Número de Respostas
Experiência anterior	209
Possibilidade de trabalhar em um cronograma necessário	292
Compromisso para toda a temporada	135
Atitude positiva e entusiasmo	407

Fonte: adaptado de Agência de Empregos (2015).

Figura 1.9 | Gráfico sobre a pesquisa com empregados temporários



Fonte: O autor (2015).

Na sua opinião, como a pesquisa realizada pode auxiliar a empresa contratante? Como a estatística pode trabalhar os dados que foram coletados? Analise cada grupo de respostas e desenvolva uma conclusão.

Avançando na prática

Pratique mais!	
Instrução Desafiamos você a praticar o que aprendeu, transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.	
Pesquisa de Opinião sobre Educação	
1. Competência de fundamento de área	Conhecer os fundamentos estatísticos básicos necessários à formação do profissional.
2. Objetivos de aprendizagem	Trabalhar as informações obtidas através de processos de pesquisa.
3. Conteúdos relacionados	Uma empresa de educação encomendou uma pesquisa de opinião para saber qual seria a área do conhecimento mais requisitada em uma grande região metropolitana, em 2015. A pesquisa foi feita com jovens sem nenhuma formação universitária. A empresa de educação pediu que fossem entrevistadas pelo menos 5000 pessoas. 55% dos entrevistados tinha curso universitário e, por isso, não responderam à pergunta.
4. Descrição da Situação Problema	Ao restante foi feita a pergunta: Qual é a área de conhecimento que você gostaria de estudar em 2015? ( ) Exatas e Informática ( ) Humanas ( ) Saúde ( ) Gestão e Negócios As informações coletadas foram: 45% gostariam de fazer cursos de EXATAS. 20% gostariam de fazer cursos de SAÚDE. 25% gostariam de fazer cursos de GESTÃO. 10% gostariam de fazer cursos de HUMANAS. Com base nos dados fornecidos anteriormente, você pode identificar: 1. Qual é a quantidade de indivíduos que formava a população? 2. Qual é o tamanho da amostra? (número de pessoas que responderam à pergunta). 3. Qual é a característica que limitou a amostra para esta pesquisa? 4. Faça uma tabela com as quantidades de cada resposta. Faça uma conclusão sobre os dados coletados.
5. Resolução da SP	1. Qual é a quantidade de indivíduos que formava a população? A população era formada por 5000 entrevistados.

5. Resolução da SP	1. Qual é a quantidade de indivíduos que formava a população? A população era formada por 5000 entrevistados. 2. Qual é o tamanho da amostra? (número de pessoas que responderam à pergunta). As pessoas que responderam à pergunta foram 45% da população. Calcularemos 45% de 5000: amostra=45% da população amostra=0,45 x 5000 amostra=2250 entrevistados 3. Qual é a característica que limitou a amostra para esta pesquisa? A característica da amostra é que todos os indivíduos não possuíam cursos universitários. 4. Faça uma tabela com as quantidades de cada resposta. Levaremos em consideração a amostra de 2250 para realizar o cálculo de cada resposta. 45% gostariam de fazer cursos de EXATAS. 20% gostariam de fazer cursos de SAÚDE. 25% gostariam de fazer cursos de GESTÃO. 10% gostariam de fazer cursos de HUMANAS. Pesquisa sobre Educação **Respostas** — **Número de respostas** EXATAS — 1013 SAÚDE — 450 GESTÃO — 562 HUMANAS — 225 Fonte: Agência de Pesquisa 5. Faça uma conclusão sobre os dados coletados. A pesquisa demonstra que a área que é mais atrativa às pessoas sem formação universitária é a área das EXATAS. 1013 pessoas responderam à pergunta realizada pela pesquisa e isso representa 45% de todos os entrevistados. O cliente contratou a pesquisa para ter conhecimento das necessidades dos moradores da região pesquisada.

Lembre-se

Estatística - é um conjunto de técnicas e métodos de pesquisa que entre outros tópicos envolve o planejamento do experimento a ser realizado, a coleta qualificada dos dados, a inferência, o processamento, a análise e a disseminação das informações.

População - conjunto de elementos que tem pelo menos uma característica em comum.

Amostra - subconjunto de elementos de uma população, que são representativos para estudar a característica de interesse da população.

Faça você mesmo

Vamos fazer uma pesquisa com os seus colegas de trabalho (devem ser, no mínimo, 30 entrevistados).

1. Defina a população (será todos os funcionários da empresa).

2. Defina a amostra (pode ser seu setor de trabalho).

A pergunta da pesquisa é:

Esse é seu primeiro emprego na sua área de formação?

As respostas podem ser:

( ) sim ( ) não

3. Tabule as respostas. (Faça uma tabela, com todas as informações necessárias e estudadas).

4. Calcule a porcentagem de cada resposta. Faça um relatório contendo a tabela, o gráfico das respostas e a conclusão com base no levantamento das respostas.

Faça valer a pena

1. Definir uma população é necessário para que se possa fazer uma pesquisa ou o levantamento de informações. Podemos definir População, como:

a) Conjunto de dados com todas as características possíveis.

b) Conjunto de elementos que tem pelo menos uma característica em comum.

c) Conjunto de pessoas de uma cidade, estado ou país.

d) Conjunto de dados já analisados pelo pesquisador.

e) Conjunto de dados da pesquisa em questão, sem relação com os resultados esperados.

2. Em uma pesquisa realizada em uma universidade, foram identificados os seguintes indicadores:

(1) idade (2) número de anos de estudo (3) escolaridade (4) renda (5) sexo (6) local de estudo (7) quantidade de livros adquiridos.

Dos dados acima, quais são quantitativos e quais são qualitativos?

a) Quantitativos – 1, 2, 4, 7 e Qualitativos – 3, 5, 6.

b) Quantitativos – 1 e Qualitativos – 2, 3, 4, 5, 6, 7.

c) Quantitativos – 1, 2 e Qualitativos – 3, 4, 5, 6, 7.

d) Quantitativos – 7 e Qualitativos – 1, 2, 3, 4, 5, 6.

e) Quantitativos – 4, 7 e Qualitativos – 3, 5, 6.

3. Procedimento em que os dados são obtidos diretamente da fonte, como, por exemplo, empresa que realiza uma pesquisa com os próprios funcionários para saber a preferência dos consumidores pela sua marca. Esse procedimento nas fases do método estatístico se refere à:

a) Apuração dos dados.

b) Coleta Indireta.

c) Coleta Direta.

d) Apresentação dos dados.

e) Análise e interpretação dos dados.

4. Uma parte retirada da população, para que seja feita sua análise, é denominada de:

a) Universo.

b) Parte.

c) Pedaço.

d) Dados brutos.

e) Amostra.

5. A parte da estatística que se preocupa somente com a descrição de determinadas características de um grupo, sem tirar conclusões sobre um grupo maior, denomina-se:

a) Estatística de População.

b) Estatística de Amostra.

c) Estatística Inferencial.

d) Estatística Descritiva.

e) Estatística Grupal.

6. Diferencie os dados qualitativos dos dados quantitativos, de modo que fique claro como os dados podem ser utilizados nas pesquisas.

7. Uma amostra é sempre finita. Assim, é correto afirmar que, se tivermos uma pesquisa com um número muito pequeno de elementos na amostra, o estudo não será significativo?

Seção 1.2

Processos de amostragem

Diálogo aberto

Na primeira seção, vimos conceitos básicos de Estatística, como população, amostra e dados. A população é um conjunto de elementos que tem pelo menos uma característica em comum. A amostra é o subconjunto de elementos de uma população, que são representativos para estudar a característica de interesse da população. Os dados são informações da amostra que podem ter características qualitativas ou quantitativas.

Nesta seção, vamos utilizar essas definições de maneira mais prática. Os processos de amostragem são utilizados para compor uma amostra. Na prática, é impossível examinar todos os elementos de uma população estudada. Daí a importância de se trabalhar com uma amostra dessa população. Utilizamos a estatística inferencial para generalizar, com uma certa segurança, as conclusões obtidas através da amostra da população. É imprescindível garantir que a amostra seja representativa da população, isto é, a amostra deve possuir as mesmas características básicas da população no que diz respeito ao fato pesquisado.

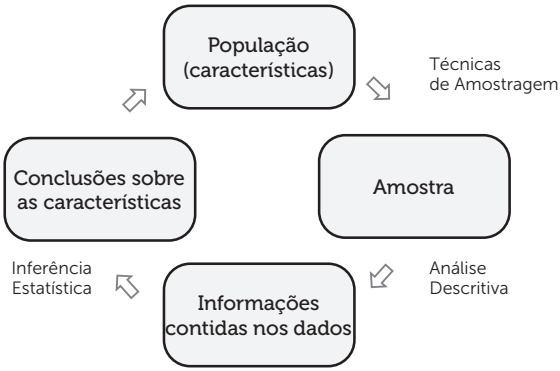
Podemos pensar que, se estudarmos todos os elementos da população, teríamos um resultado ou uma conclusão mais precisa. Mas isso não é verdade! Ao manusear um grande número de dados, estamos sujeitos a imprecisões que podem provocar erros grandes, comparadas às conclusões de uma amostra bem selecionada, que contém informações relevantes ao estudo.

Nesta seção, veremos quais as técnicas mais comuns que podemos utilizar para compor uma amostra.

Os objetivos específicos desta seção são identificar a metodologia empregada na pesquisa de um fenômeno e definir fatores que afetam a quantidade de informações de um fenômeno. A preocupação central com relação à amostra é que ela seja representativa. Os processos utilizados para a definição da amostra devem ser adequados para que as amostras apresentem essa representatividade.

A Figura 1.10 mostra a relação das técnicas de amostragem e como elas definem de forma cuidadosa as amostras, gerando informações relevantes que constroem conclusões sobre as características estudadas referentes à população.

Figura 1.10 | Relação das Técnicas de Amostragem



Fonte: O autor (2015).

Ao decidirmos a obtenção das informações pelo levantamento amostral, teremos dois tipos de problema: a definição da população de interesse de forma cuidadosa e a seleção das características que devemos pesquisar. Se os dados forem coletados de maneira descuidada, poderão ser tão inúteis e nenhum processamento estatístico conseguirá salvá-los.

Uma agência de pesquisa foi contratada por uma empresa multinacional do setor de alimentos e precisa entrevistar gerentes de contratação. A empresa decidiu analisar as características dos gerentes que responderam "**Compromisso para toda a temporada**". A porcentagem de gerentes entrevistados que deu essa resposta foi de 13%.

Qual é a amostra que será analisada?

Os estatísticos da agência de pesquisa devem utilizar qual método de amostragem para analisar as características da amostra? E por quê?

Sabe-se que na população de 1043 gerentes de contratação entrevistados, 70% eram homens e 30% eram mulheres.

Para que a amostra seja representativa, de quantos homens e de quantas mulheres que deram essa resposta serão analisados?

Assimile

A **Amostragem** consiste em procedimentos para extração de amostras que representem bem a população.

Os Riscos da Amostragem estão à margem de erro motivado pelo fato de investigarmos parcialmente (amostras) o universo (população).

Chamamos de população-alvo aquela sobre a qual vamos fazer inferências baseadas na amostra. Para que possamos fazer inferências válidas sobre ela a partir de uma amostra, é preciso que essa seja representativa.

Uma das formas de se conseguir representatividade é fazer a escolha da amostra em um processo aleatório. Além disso, a aleatoriedade permite o cálculo de estimativas dos erros envolvidos no processo de inferência.

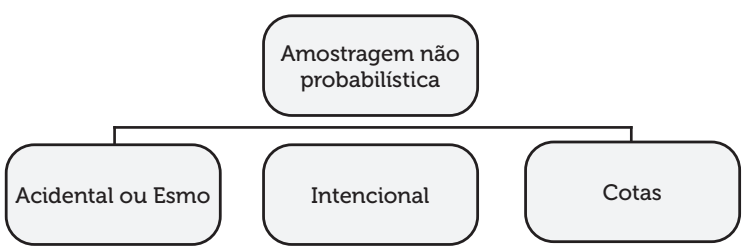
Quanto à extração dos elementos, as amostras podem ser com reposição, se um elemento sorteado puder ser sorteado novamente, ou sem reposição, se o elemento sorteado só puder figurar uma única vez na amostra. Basicamente, existem dois métodos para composição da amostra: não probabilístico (intencional) e probabilístico.

Reflita

Imagine que somos donos de uma fábrica de fósforos. Como saberemos se todos os fósforos estão funcionando? Se testarmos 100%, não teremos o que vender. Certo?

A Figura 1.11 mostra a divisão da Amostragem não probabilística.

Figura 1.11 | Tipos de amostragens não probabilísticas

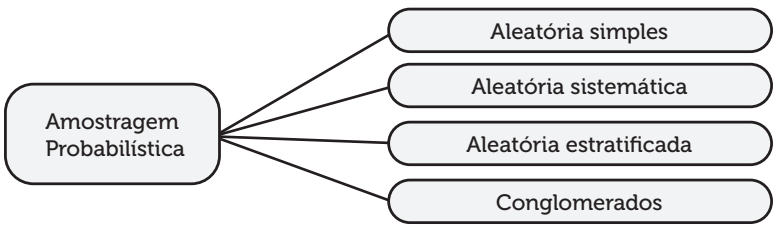


Fonte: O autor (2015).

Os métodos não probabilísticos são amostragens em que há uma escolha deliberada dos elementos que compõem a amostra. Não se pode generalizar os resultados das pesquisas para a população, uma vez que as amostras não probabilísticas não garantem a representatividade da população.

A amostragem não probabilística é dividida em: **Amostragem Acidental (Esmo)**: quando os elementos vão aparecendo até completar o número necessário de elementos para a amostra. **Amostragem Intencional:** há um critério para a escolha da amostra em um grupo de elementos. **Amostragem por Cotas:** classifica a população em termos de propriedades com uma característica relevante a ser estudada. Proporções são estipuladas para cada característica. Essas cotas para cada observador serão importantes para a troca de responsabilidade ao selecionar interlocutores ou entrevistados.

Figura 1.12 | Tipos de amostragens probabilísticas



Fonte: O autor (2015).

O método de amostragem probabilística exige que cada elemento da população possua determinada probabilidade de ser selecionado. A amostragem probabilística é dividida em: **Amostragem Aleatória Simples** – é um processo para selecionar amostras de tamanho "n" dentre as "N" unidades em que foi dividida a população. Estas amostras acidentais podem ser escolhidas por diversos métodos, inclusive por tabelas de números aleatórios (TNA) e de computadores para gerar números randômicos. Equivale a um sorteio no qual se colocam todos os números misturados dentro de uma urna. As unidades correspondentes aos números escolhidos formarão a amostra.

Exemplificando

Pesquisa da estatura de uma escola com 90 alunos (população: 90 alunos), usando uma amostra de 10% da população:

1. Numeram-se os alunos de 1 a 90.

2. Sorteiam-se 9 números (10% de 90), usando algum mecanismo aleatório ou através de uma Tabela de Números Aleatórios.

Tem-se:

14 35 30 19 66 27 77 45 38

Amostragem Aleatória Estratificada é adquirida por um processo de separação das unidades da população em grupos não superpostos chamados estratos, e selecionando-se independentemente uma amostra aleatória simples de cada estrato. A amostragem estratificada pode ser uma amostra de igual tamanho ou uma amostra proporcional (JOHNSON; KOBY, 2013).

Para a amostra estratificada de tamanho igual, é sorteado um número igual de elementos para cada estrato que será estudado. No outro caso, o processo de calcular o número de amostras por estrato é:

$$\frac{N_a}{N} = \frac{n_a}{n} \rightarrow n_a = \frac{n}{N} \cdot N_a$$

Onde:

N $\rightarrow$ Nº de unidades da população

n $\rightarrow$ Nº de unidades das amostras

$N_a \rightarrow$ Nº de unidades do estrato A

$n_a \rightarrow$ Nº de amostras de A

Exemplificando

Foram entrevistados 90 alunos, 54 meninos e 36 meninas. Vamos então obter uma amostra proporcional estratificada de 10%

Figura 1.13 | Homem e mulher

Fonte: O autor (2015).

Solução:

- São, portanto, dois estratos (sexo masculino e feminino) e queremos uma amostra de 10% da população.
- Calcula-se o número de amostras de cada estrato:

Sexo	População	10%	Número de Amostras
Masculino	54	5,4	5
Feminino	36	3,6	4
Total	90	9,0	9

- Numeramos os alunos de 01 a 90, sendo que de 01 a 54 correspondem a meninos e de 55 a 90, meninas. O próximo passo é o mesmo do exemplo anterior.

Amostragem por Conglomerado é uma amostra aleatória simples na qual cada unidade de amostragem é um grupo, ou um conglomerado de elementos. Primeiro é preciso especificar conglomerados apropriados, eles têm caractarísticas similares. O número de elementos num conglomerado deve ser pequeno em relação à população e o número de conglomerados deverá ser rasoavelmente grande (JOHNSON; KOBY. 2013).

Amostragem Aleatória Sistemática - é uma amostra quando os elementos da população já se encontram ordenados, não havendo necessidade de se construir o sistema de referência. A seleção

dos elementos que constituirão a amostra pode ser feita por um sistema imposto pelo pesquisador. Uma amostra sistemática de n elementos de uma população de tamanho N, K deve ser menor ou igual a N/n. Não é possivel determinar K, precisamente quando o tamanho da população é desconhecido, mas pode-se supor um valor de K de tal modo que seja possivel obter uma amostra de tamanho n. A amostragem sistemática é mais fácil de ser executada e, por isso, está menos sujeita a erros do entrevistador do que aqueles que acontecem na aleatória simples. A amostragem sistemática repetidamente proporciona mais informações por custo unitário do que a aleatória simples (JOHNSON; KOBY, 2013).

Diretrizes para calcular as amostras:

1° - Estabelecer o intervalo de amostragem K:	$K=\frac{N}{n}$
2° - Iniciar aleatóriamente a composição da amostra:	b → início (n° de ordem inicial sorteado da TNA). Obs: $0 < b \leq K$
3° - Composição da Amostra:	1° item → b 2° item → b + K 3° item → b + 2k

Exemplificando

Suponhamos uma rua contendo quinhentos prédios, dos quais desejamos obter uma amostra formada de vinte prédios.

Figura 1.14 | Prédios

Fonte: Disponível em: <http://upload.wikimedia.org/wikipedia/commons/f/f3/Pr%C3%A9dios_no_Setor_Bueno.jpg>. Acesso em: 30 maio 2015.

Solução:

a) Calcular K (intervalo de amostragem) K=500/20, K=25

b) b= 12 (valor encontrado na TNA)

c) vamos começar aleatoriamente pelo 12º prédio.

No final teremos as 20 amostras.

d) Composição da amostra

1º item $\rightarrow$ 12

2º item $\rightarrow$ 12 + 25 = 37

3º item $\rightarrow$ 12 + 2*25 = 62

20º item $\rightarrow$ 12 +19*25 = 487

Sem medo de errar

Qual é a amostra que será analisada?

$$amostra = \frac{13 \times 1043}{100} = 136\ entrevistados$$

O tipo de amostragem que deve ser utilizada pela agência é a amostragem aleatória estratificada. Uma amostra estratificada é obtida separando-se as unidades da população em grupos não superpostos chamados estratos, e selecionando-se independentemente uma amostra aleatória simples de cada estrato. Existem dois tipos de amostragem estratificada.

No nosso caso, o estrato que será estudado é de todos que responderam "**Compromisso para toda a temporada**".

Sabe-se que, na população de 1043 gerentes de contratação entrevistados, 70% eram homens e 30% eram mulheres. Para que a amostra seja representativa, quantos homens e quantas mulheres que deram essa resposta serão analisados?

$$\boldsymbol{homens} = \frac{70 \times 136}{100} = 95\ entrevistados$$

$$\boldsymbol{mulheres} = \frac{30 \times 136}{100} = 41\ entrevistados$$

Para que a amostra expresse a realidade da população de entrevistados, devem ser escolhidos 95 homens e 41 mulheres.

A representatividade da amostra foi preservada? A amostra separada reflete a realidade da população?

Avançando na prática

Pratique mais!	
Instrução Desafiamos você a praticar o que aprendeu, transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.	
Telespectadores do Cinema Brasileiro	
1. Competência de fundamento de área	Conhecer os fundamentos estatisticos basicos necessarios a formacao do profissional.
2. Objetivos de aprendizagem	Identificar a metodologia empregada na pesquisa de um fenômeno. Definir fatores que afetam a quantidade de informações de um fenômeno.
3. Conteúdos relacionados	Processos de Amostragem
4. Descrição da SP	Em um cinema há 200 pessoas, entre as quais 140 são homens e 60 são mulheres. Precisamos selecionar uma amostra proporcional de 10 pessoas e outra amostra proporcional de 20 pessoas, outra amostra de 30 e outra amostra de 40 pessoas para fazermos uma pesquisa de opinião sobre o espetáculo.

4. Descrição da SP	Figura 1.15 | Claquete  Fonte: Disponível em: <http://commons.wikimedia.org/wiki/File:Clap_cinema.svg>. Acesso em: 30 maio 2015.
5. Resolução da SP	Para selecionar uma amostra aleatória estratificada proporcional com 10 pessoas, devemos dividir a população em dois estratos: homens e mulheres, por exemplo. Os homens participam desta população com (140 / 100) x 200 = 70% e as mulheres com (60 / 100) x 200 = 30%. Logo, 70% da amostra deverão ser homens e 30% da amostra deverão ser mulheres. Se a amostra for de 10 pessoas, temos: Homens = (70/100) x 10 = 7 homens Mulheres = (30/100) x 10 = 3 mulheres Agora, se a amostra estratificada proporcional for de 20 pessoas, logo teremos 14 homens e 6 mulheres. Para uma amostra de 30 pessoas, logo teremos 21 homens e 9 mulheres. E, para uma amostra de 40 pessoas, teremos 28 homens e 12 mulheres. A seleção deverá ser feita por meio de sorteio, de acordo com os conceitos da amostragem aleatória simples.

Lembre-se

Amostragem aleatória simples é um processo para selecionar amostras de tamanho "n" dentre as "N" unidades em que foi dividida a população.

Amostragem estratificada é obtida separando-se as unidades da população em grupos não superpostos chamados estratos.

Amostragem por conglomerado é uma amostra aleatória simples na qual cada unidade de amostragem é um grupo, ou um conglomerado de elementos.

Amostragem Aleatória Sistemática é a seleção dos elementos que constituirão a amostra, e pode ser feita por um sistema imposto pelo pesquisador.

Amostragem Acidental (Esmo) é formada por elementos que vão aparecendo, que são possíveis de se obter até completar o número de elementos da amostra.

Amostragem Intencional é formada por elementos escolhidos por determinado critério.

Faça você mesmo

Vamos fazer uma pesquisa com os mesmos colegas de trabalho (o mínimo deve ser de 30 entrevistados).

1. Agora mostre do total entrevistado quantos são homens e quantos são mulheres.

2. Precisamos analisar 10% das respostas. Quantas respostas de homens e quantas de mulheres precisaremos analisar para que essa amostra represente a realidade da população?

3. Estratifique todas as respostas "sim", perguntando aos entrevistados a idade de cada um.

4. Faça uma tabela com os valores das idades dos entrevistados.

Faça valer a pena

1. Um pesquisador está investigando uma informação sobre a quantidade de vezes que os médicos prescrevem o remédio Oseltamivir®, que é um composto usado para tratamento contra gripe suína. Para isso, está usando dados de 15000 receituários (N) do ano anterior. A amostra que será estudada é de 10% de todos os receituários, n= 1500. Qual tipo de amostragem você sugeriria para que o pesquisador tivesse uma amostragem representativa?

a) Amostragem aleatória sistemática.

b) Amostragem acidental.

c) Amostragem por cotas.

d) Amostragem por conglomerado.

e) Amostragem intencional.

2. Na zona rural de Montevidéu – UR, foi realizada uma pesquisa de opinião pública. Os elementos na população de interesse são todos os homens e mulheres com idade acima de 30 anos. Qual é o tipo de amostragem você sugeriria para que o pesquisador tivesse uma amostragem representativa?

a) Amostragem por conglomerado.

b) Amostragem por cotas.

c) Amostragem aleatória estratificada.

d) Amostragem acidental.

e) Amostragem intencional.

3. Quais alternativas definem a amostra probabilística e a amostra não probabilística?

I. Amostragem probabilística é aquela em que todos os elementos da população possuem probabilidade não nula de participar da amostra e sua principal característica é o uso do sorteio.

II. A amostragem não probabilística é quando, por alguma razão, algum elemento ou grupo de elementos da população possuir probabilidade nula de participar da amostra ou a amostragem for feita sem sorteio.

III. Amostragem probabilística é aquela em que todos elementos da população possuem probabilidade nula de participar da amostra.

IV. A amostragem não probabilística é feita apenas por sorteio.

a) Apenas I e IV.

b) Apenas IV e III.

c) Apenas II e IV.

d) Apenas III e I.

e) Apenas I e II.

4. Uma faculdade é formada por 400 alunos do curso de biologia, 300 do curso de engenharia elétrica, 200 do curso de ciência da computação e 100 do curso de geografia. Extrair uma amostra de 100 alunos, pelo método de amostragem estratificada proporcional. Para que a amostra seja representativa, quantos devem ser os alunos de cada curso?

a) 10 alunos do Curso de Biologia, 20 do Curso de Engenharia Elétrica, 30 do Curso de Ciência da Computação e 40 do Curso de Geografia.

b) 40 alunos do Curso de Biologia, 30 do Curso de Engenharia Elétrica, 20 do Curso de Ciência da Computação e 10 do Curso de Geografia.

c) 20 alunos do Curso de Biologia, 10 do Curso de Engenharia Elétrica, 30 do Curso de Ciência da Computação e 40 do Curso de Geografia.

d) 30 alunos do Curso de Biologia, 10 do Curso de Engenharia Elétrica, 20 do Curso de Ciência da Computação e 40 do Curso de Geografia.

e) 20 alunos do Curso de Biologia, 40 do Curso de Engenharia Elétrica, 30 do Curso de Ciência da Computação e 10 do Curso de Geografia.

5. Para uma pesquisa sobre uma fragrância de perfume, foram entrevistados 50 consumidores em um magazine que tinha 800 clientes na hora da pesquisa. Foi feito um controle na entrada do magazine e sabe-se que havia 480 homens e 320 mulheres. Para que o resultado da pesquisa tenha relevância, sabemos que a amostra deve representar a realidade. Sendo assim, dos 50 consumidores entrevistados, quantos devem ser homens e quantos devem ser mulheres?

a) 10 homens e 40 mulheres.

b) 20 homens e 30 mulheres.

c) 30 homens e 20 mulheres.

d) 40 homens e 10 mulheres.

e) 0 homens e 50 mulheres.

6. Em uma empresa de pilhas do tipo AAA, necessita-se fazer uma amostragem de 10 pilhas de um lote de fabricação de 20.000 pilhas, alocadas em um depósito em caixas com 50 pilhas, sendo as caixas colocadas em um armário que comporta 10 caixas por compartimento.

Se cada caixa tem 50 pilhas e cada compartimento tem 10 caixas, então, cada compartimento tem 500 pilhas. Dessa forma, o armário deve ter 40 compartimentos para estocar as 20.000 pilhas.

Passos a serem seguidos na definição da amostra:

(i) Sorteamos um compartimento entre os 40.

(ii) Sorteamos uma caixa desse compartimento.

(iii) Sorteamos 10 pilhas na caixa que pegamos anteriormente.

Assim definimos a amostra.

Essa situação ilustra qual tipo de amostragem e por quê?

7. Uma pesquisa salarial foi realizada em março de 2015, mês em que acontece a conferência sindical dos profissionais da indústria de automóveis, e os salários de diversos profissionais de produção, vendas, departamento, marketing e recursos humanos foram levantados. Foram relacionados os salários na tabela abaixo, expressos em mil reais. Necessita-se separar as informações em uma amostra de 7 salários. Liste as amostras aleatórias sistemáticas possíveis.

4,2	4,6	5,6	5,6	4,3	4,6	4,7
3,9	5,0	5,6	4,2	4,8	4,2	3,9
4,9	5,7	4,9	4,3	4,1	4,7	4,0
4,6	4,3	4,9	4,7	4,9	4,0	4,3

Seção 1.3

Medidas de tendência central e medidas de dispersão

Diálogo aberto

Vimos, na seção 1.1, uma introdução aos conceitos básicos de Estatística, como amostra e população e como estes dados influenciam na pesquisa. São conceitos importantes para o desenvolvimento do nosso estudo. Na seção 1.2 foram apresentados os processos de amostragem, utilizados para compor uma amostra, pois, na prática, é impossível examinar todos os elementos de uma população estudada.

As medidas de tendência central e as medidas de dispersão serão abordadas nesta seção. As medidas de tendência central recebem esse nome pelo fato dos dados observados se agruparem em torno de valores centrais. As medidas de dispersão são conhecidas como variância e desvio-padrão. Os objetivos desta seção são identificar as medidas de tendência central moda, média e mediana e definir medidas de dispersão como desvio-padrão e variância, sabendo interpretá-las de forma correta.

Uma agência de pesquisa foi contratada por uma empresa Multinacional do setor de alimentos e precisa entrevistar gerentes de contratação. Conforme falamos anteriormente, a empresa decidiu analisar as características dos gerentes que responderam "Compromisso para toda a temporada". A porcentagem de gerentes que deram essa resposta foi de 13% dos entrevistados. Já sabemos o tamanho da amostra. Dessa amostra (136 funcionários), mais alguns dados foram analisados.

Foram tabuladas as idades de todos os entrevistados dessa amostra para que novas conclusões pudessem ser estabelecidas através da análise desses dados. As primeiras linhas são os dados dos homens e as demais são os dados referentes às mulheres.

Figura 1.16 Empregados homens e mulheres

Fonte: Disponível em: <http://goo.gl/ATTGWc>. Acesso em: 30 mai.o2015.

Idades dos Entrevistados que responderam: "Compromisso para toda a temporada"																	
32	39	60	45	43	33	39	59	49	32	48	30	34	48	62	51	39
43	62	41	39	33	64	33	45	50	36	56	31	40	60	65	61	51
62	52	57	50	63	31	45	32	64	39	31	43	51	46	58	54	37
38	39	30	54	32	46	33	60	44	61	52	55	36	59	64	44	61
61	59	65	30	32	42	56	60	57	54	34	30	52	57	30	39	51
41	34	43	36	42	44	63	56	53	60	47	56	44	47	49	51	61
60	57	62	32	37	58	35	49	59	55	58	47	50	53	49	37	54
55	40	37	38	35	41	45	32	45	46	41	53	48	30	44	31	45

Para fazer o relatório, é necessário que você:

Estratifique em duas amostras: homens (95) e mulheres (41) e monte duas tabelas distintas.

Coloque em ordem crescente os dados das duas tabelas. Calcule a média aritmética, mediana e moda para a idade dos homens e repita o cálculo para a idade das mulheres.

Calcule a variância e o desvio-padrão e estabeleça as conclusões com base nos resultados encontrados.

Reflita

Medidas de posição podem ser utilizadas em conjunto para auxiliar na análise dos dados, mas existem situações em que uma das medidas pode ser mais conveniente do que a outra.

Medidas de Tendência Central

O cálculo das medidas pode possibilitar a localização da maior concentração de valores de uma dada distribuição, isto é, se ela se localiza no início, no meio ou no final, ou, ainda, se há uma distribuição por igual. Essas medidas promovem comparações de séries de dados entre si pela confrontação desses números.

Média

A média pode ser obtida pelo quociente da soma de todos os dados do experimento e o número total de dados. A média (aritmética) é, de modo geral, a mais importante de todas as

medidas descritivas. Seu valor é calculado por meio da divisão dos números somados pela quantidade deles. A média possui a função de transformar um conjunto de números em um único valor, dando uma visão global dos dados.

Média aritmética simples

A média aritmética simples é, como o nome já diz, a mais simples, e a de uso mais comum. $\bar{x} = \frac{\sum_i^n x_i}{n}$

Exemplificando

Na primeira prova de Estatística, você teve a nota 8, na segunda a nota foi 7, na terceira obteve a nota 6 e na última prova sua nota foi 7. Para ser aprovado, sua média tem de ser igual ou maior que 7.

$$\bar{x} = \frac{\sum_i^n x_i}{n} = \frac{8+7+6+7}{4} = 7$$

Sua média, nesse caso, seria 7 e você estaria aprovado.

Média ponderada

Diferente da simples, a média aritmética ponderada calcula a média quando os valores possuem pesos diferentes.

$$\overline{x_p} = \frac{\sum_{i=1}^n (p_i \times x_i)}{\sum_{i=1}^n p_i}$$

Exemplificando

Usando o mesmo exemplo das notas de Estatística, imagine que cada uma das notas tem um peso distinto. A primeira prova possuía peso 2, a segunda, peso 2, a terceira, peso 3, e a quarta, peso 3. Como isso pode ser calculado? Multiplica-se o valor pelo seu peso, somando aos resultados das outras multiplicações, e, então, divide-se pela soma de todos os pesos. Confira o cálculo do exemplo:

$$\frac{8.2+7.2+6.3+7.3}{2+2+3+3} = \frac{16+14+18+21}{10} = \frac{69}{10} = 6{,}9$$

Nesse caso, a média seria 6,9.

Na média ponderada, ao contrário da média simples, a alteração da posição dos números pode ocasionar resultados errados.

Moda (Mo)

A moda é o valor que mais aparece no conjunto de dados do experimento. É o valor que ocorre com maior frequência em um conjunto de dados, sendo denominado valor modal. Baseado nesse contexto, um conjunto de dados pode apresentar mais de uma moda. Nesse caso, dizemos ser multimodais, caso contrário, quando não existe um valor predominante, dizemos que é amodal.

Exemplificando

Calcular a moda para as idades dos candidatos à presidência de um clube desportivo:

65, 87, 49, 58, 65, 65, 67, 83, 87, 79.

Observe que, Mo = 65 (aparece 3 vezes).

A moda é especialmente útil para dados qualitativos. Não é possível analisar média ou mediana de dados não ordenados, como cidade ou preferência musical. Então a Moda entra em ação.

Mediana (Md)

A mediana é o valor tal que mais da metade dos dados é maior ou igual a ela, e mais da metade dos dados é menor ou igual a ela. A mediana é uma medida de posição. É, também, uma separatriz, pois divide o conjunto em duas partes iguais, com o mesmo número de elementos. O valor da mediana encontra-se no centro da série estatística organizada em ordem crescente, de tal forma que o número de elementos situados antes desse valor (mediana) é igual ao número de elementos que se encontram após esse mesmo valor (mediana).

Para o cálculo da mediana, temos duas considerações a fazer:

1. O número de observações (n) é ímpar. A mediana será o valor da variável que ocupa a posição de ordem $\frac{n+1}{2}$.

Exemplificando

Calcular a mediana dos valores:

9, 12, 8, 6, 14, 11, 5.

Em primeiro lugar, vamos organizar os dados em ordem crescente:

5, 6, 8, 9, 11, 12, 14

Observe que n = 7 (ímpar)

Logo, a mediana será dada pelo elemento que divide o Rol em duas partes iguais.

Md = 9

2. O número de observações (n) é par. Não existe, portanto, um valor que ocupe o centro, convencionando-se que a mediana será a média aritmética dos valores que ocupam as posições de ordem:

$$\frac{n}{2} \quad \text{e} \quad \frac{n}{2} + 1$$

Exemplificando

Calcular a mediana dos valores já ordenados: 6, 8, 9, 11, 12, 14.

n = 6 (par)

$$\frac{n}{2} = 3 \quad \text{e} \quad \frac{n}{2} + 1 = 4$$

A mediana será dada pela média aritmética entre o 3º e 4º elementos da sequência:

$$md = \frac{9 + 11}{2} = 10$$

Para calcularmos a mediana quando os dados estão agrupados em classes, não levamos em consideração se n é par ou ímpar e procedemos do seguinte modo:

1) Calcula-se n/2.

2) Pela frequência acumulada, identifica-se a classe que contém a mediana.

3) Aplica-se a fórmula:

$$md = li_{md} + \frac{\left(\frac{n}{2} - fac\right) \cdot h}{ni_{md}}$$

li_{md}=*limite inferior da classe md*

n=número total de elementos da amostra

f_{ac}=*frequência acumulada da classe anterior à classe md*

h=amplitude da classe md

ni_{md}=*frequência da classe md*

Média Geométrica (M_G): dados agrupados e não agrupados em classes.

Dados não tabelados

A média geométrica de um conjunto de N números x_1, x_2, x_3,... x_n é a raiz de ordem N do produto desses números:

$$M_G = \sqrt[N]{x_1 \cdot x_2 \cdot x_3 \cdots x_n}$$

$$M_G = \sqrt[N]{{x_1}^{n_1} \cdot {x_2}^{n_2} \cdot {x_3}^{n_3} \cdots {x_n}^{n_n}}$$

$$x_1: valor\ observado$$

$$n_1: número\ de\ observações\ da\ classe$$

$$M_G = \sqrt[N]{{\overline{x_1}}^{n_1} \cdot {\overline{x_2}}^{n_2} \cdot {\overline{x_3}}^{n_3} \cdots {\overline{x_n}}^{n_n}}$$

$$\overline{x_1}: ponto\ médio\ do\ intervalo\ de\ classe$$

$$n_1: número\ de\ observações\ de\ classe$$

Exemplificando

A média geométrica é muito utilizada nas situações envolvendo aumentos sucessivos. Por exemplo, vamos considerar um aumento de salário sucessivo de 15% no primeiro mês, 12% no segundo mês e 21% no terceiro mês. Vamos determinar a média geométrica dos aumentos, mas, para isso, as taxas percentuais devem ser transformadas em taxas unitárias; observe:

100%+15% = 1,15 100%+12% = 1,12 100%+21% = 1,21

$$\sqrt[3]{1{,}15 \times 1{,}12 \times 1{,}21} = 1{,}1594$$

O valor 1,1594 corresponde à taxa média de 15,94% de todos os aumentos sucessivos. Isso indica que a aplicação três vezes consecutivas da taxa de 15,94% corresponderá ao aumento sucessivo dos percentuais de 15%, 12% e 21%. Suponhamos que o salário reajustado seja de R$ 600,00. Acompanhe os aumentos utilizando as duas opções de reajustes:

1ª opção	2ª opção
600,00 + 15% = 690,00 690,00 + 12% = 772,80 772,80 + 21% = 935,09	600,00 + 15,94% = 695,64 695,64 + 15,94% = 806,53 806,53 + 15,94% = 935,09

Média Harmônica (M_H): dados agrupados e não agrupados em classes.

Sejam x_1, x_2, x_3,.....x_n, valores de x, associados às frequências absolutas n_1, n_2, n_3,.....n_n, respectivamente.

A média harmônica de X é definida por:

$$M_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

• Para dados não agrupados, n = 1.

• Para dados agrupados sem intervalo de classe, xi é o valor da variável.

• Para dados agrupados com intervalo de classe, xi é o ponto médio da classe.

Exemplificando

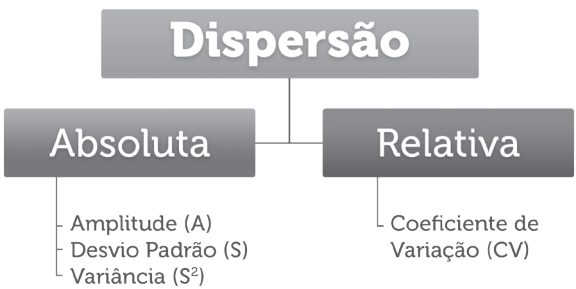
A média harmônica está relacionada ao cálculo matemático das situações envolvendo as grandezas inversamente proporcionais. Como exemplo, temos a relação entre velocidade e tempo. Suponha que, em uma determinada viagem, um carro desenvolva duas velocidades distintas: durante a metade do percurso ele manteve a velocidade de 50 km/h e durante a metade restante sua velocidade foi de 60 km/h. Vamos determinar a velocidade média do veículo durante o percurso. De acordo com a média harmônica, temos a seguinte relação:

$$MH = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{1}{\frac{1}{50} + \frac{1}{60}} \cong 54$$

A velocidade média do veículo durante todo o percurso será de aproximadamente 54 km/h. Caso calculássemos a velocidade média utilizando a média aritmética, chegaríamos ao resultado de 55 km/h. Esse valor demonstra que a velocidade e o tempo de percurso nos dois trechos seriam iguais. Mas precisamos considerar que no primeiro trecho o automóvel levou um tempo maior para o percurso, pois a velocidade era de 50 km/h e no segundo trecho o tempo decorrido foi menor, devido à velocidade de 60 km/h. Nesse momento, observamos a relação inversa entre velocidade e tempo e, para que não ocorra erro, é aconselhável, nessas condições, a utilização da média harmônica.

Medidas de Dispersão

Figura 1.17 | Medidas de dispersão



Fonte: O autor (2015)

Um aspecto importante no estudo descritivo de um conjunto de dados é o da determinação da variabilidade ou dispersão desses

dados, relativamente à medida de localização do centro da amostra. Supondo ser a média a medida de localização mais importante, será relativamente a ela que se definirá a principal medida de dispersão - a variância, apresentada a seguir.

São medidas estatísticas utilizadas para avaliar o grau de variabilidade, ou dispersão, dos valores em torno da média. Servem para medir a representatividade da média.

Amplitude Total

A amplitude total é a diferença entre o maior e o menor valor analisado em uma variável em ordem crescente. Vejamos, agora, como calcular amplitude total com dados agrupados e não agrupados (CRESPO, 2002).

$$AT=x_{(máx)}-x_{(mín)}$$

Foram medidos e ordenados os valores das variáveis A, B e C, então, vamos calcular a amplitude total.

A	80	80	80	80	80	80	80	80	80	80
B	76	77	78	79	80	80	81	82	83	84
C	55	65	70	75	80	85	85	90	95	100

Assim, aplicando a fórmula anterior para esses dados, obteremos os seguintes resultados:

$$AT_A=x_{(máx)}-x_{(mín)}=80-80=0$$

$$AT_B=x_{(máx)}-x_{(mín)}=84-76=8$$

$$AT_C=x_{(máx)}-x_{(mín)}=100-55=45$$

Nesse caso, podemos observar:

A variável A obteve uma amplitude total igual a 0, ou seja, uma dispersão nula. Então, significa que os valores não variam entre si. A variável B obteve uma amplitude igual a 8 e a variável C obteve uma amplitude total igual a 45. A utilização da amplitude total como medida de dispersão é muito limitada, pois só leva em consideração dois valores de todo o conjunto de dados. Assim,

quanto maior for o valor encontrado para a amplitude total, maior será a inconsistência ou a variação entre os valores da variável. Vamos ver agora a medida de dispersão variância e desvio-padrão.

Variância (s^2)

A variância é uma medida de dispersão que verifica a distância entre os valores da média aritmética.

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

Variância amostral

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

Variância populacional

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

Exemplificando

Em um clube de corrida, o treinador anotou o tempo gasto durante 5 dias de treinamento para analisar o desempenho dos corredores. A equipe é formada de três corredores: Jonas, Salomão e Davi.

Os tempos em minutos foram anotados na tabela a seguir:

Atletas	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5
Jonas	63	60	59	55	62
Salomão	54	59	60	57	61
Davi	60	63	58	62	55

O treinador está preocupado se esses tempos podem levá-los à vitória na competição que se aproxima. Ele precisa fazer cálculos de variância de cada atleta para analisar o desempenho e a possibilidade de vitória na competição.

Os dados que ele precisa levar em consideração para obter as análises são a média de tempo de cada atleta e a variância desses tempos.

Os cálculos da média aritmética de cada atleta estão abaixo:

$$Jonas \rightarrow \overline{x_J} = \frac{63 + 60 + 59 + 55 + 62}{5} = 59{,}8\ minutos$$

$$Salomão \rightarrow \overline{x_S} = \frac{54 + 59 + 60 + 57 + 61}{5} = 58{,}2\ minutos$$

$$Davi \rightarrow \overline{x_D} = \frac{60 + 63 + 58 + 62 + 55}{5} = 59{,}6\ minutos$$

Exemplificando

A média de cada atleta é conhecida pelo treinador. Agora, para calcular a variância de cada atleta, utilizaremos a seguinte fórmula:

$$Var = \frac{(dia\ 1 - \bar{x})^2 + (dia\ 2 - \bar{x})^2 + (dia\ 3 - \bar{x})^2 + (dia\ 4 - \bar{x})^2 + (dia\ 5 - \bar{x})^2}{total\ de\ dias}$$

$$Var_{jonas} = 7,76\ min \quad Var_{Salomão} = 6,16\ min \quad Var_{Davi} = 8,24\ min$$

A conclusão que o treinador tira com os cálculos de variância é que o atleta Davi tem tempos mais dispersos da média. Jonas tem tempos mais próximos da média dos outros atletas.

Desvio-padrão (S)

É a medida mais usada na comparação de diferenças entre conjuntos de dados, por ter grande precisão. É responsável por determinar a dispersão dos valores em relação à média e é calculado por meio da raiz quadrada da variância, conforme mostra a fórmula:

$$s = \sqrt{s^2} \quad \text{ou}$$

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

Exemplificando

Calcule o desvio-padrão para os tempos dos atletas.

$$s = \sqrt{\frac{(x_i - \bar{x})^2}{n}} = \sqrt{\frac{(x_i - \bar{x})^2 + (x_i - \bar{x})^2 + (x_i - \bar{x})^2 + (x_i - \bar{x})^2 + (x_i - \bar{x})^2}{n}}$$

$$s_{Jonas} = \sqrt{7,76} = 2,79 \quad s_{Salomão} = \sqrt{6,16} = 2,48 \quad s_{Davi} = \sqrt{8,24} = 2,87$$

O desvio-padrão indica qual é o "erro" se quiséssemos substituir um dos valores coletados pelo valor da média.

Coeficiente de Variação (CV)

É a medida relativa de dispersão útil para fazer comparação em termos relativos do grau de concentração. É calculado pela relação entre o desvio-padrão (s) e a média x da média de séries distintas.

$$CV = \frac{s}{\bar{x}} \times 100$$

Diz-se que uma distribuição tem: baixa dispersão quando CV ≤ 25%, média dispersão: 25%< CV<70% e alta dispersão: CV ≥ 70%. Quanto menor for o valor do coeficiente de variação, mais homogêneos serão os dados.

Exemplificando

Calcule o coeficiente de variação para cada atleta.

$$CV_{Jonas} = \frac{s}{\bar{x}} \times 100 = \frac{2{,}79}{59{,}8} \times 100 = 4{,}66\%$$

$$CV_{Salomão} = \frac{s}{\bar{x}} \times 100 = \frac{2{,}48}{58{,}2} \times 100 = 4{,}26\%$$

$$CV_{Davi} = \frac{s}{\bar{x}} \times 100 = \frac{2{,}87}{59{,}6} \times 100 = 4{,}81\%$$

A distribuição de dados analisada para os atletas possui baixa dispersão, pois os CV calculados estão abaixo de 25%.

Sem medo de errar

Estratifique em duas amostras:

Idades dos Homens entrevistados				
32	39	60	45	43
33	39	59	49	32
48	30	34	48	62
51	39	43	62	41
39	33	64	33	45
50	36	56	31	40
60	65	61	51	62
52	57	50	63	31
45	32	64	39	31
43	51	46	58	54
37	38	39	30	54
32	46	33	60	44
61	52	55	36	59
64	44	61	61	59
65	30	32	42	56
60	57	54	34	30
52	57	30	39	51
41	34	43	36	42
44	63	56	53	60

Amostra de homens ordenada				
30	30	30	30	30
31	31	31	32	32
32	32	32	33	33
33	33	34	34	34
36	36	36	37	38
39	39	39	39	39
39	39	40	41	41
42	42	43	43	43
43	44	44	44	45
45	45	46	46	48
48	49	50	50	51
51	51	51	52	52
52	53	54	54	54
55	56	56	56	57
57	57	58	59	59
59	60	60	60	60
60	61	61	61	61
62	62	62	63	63
64	64	64	65	65

$$média_{homens} = \frac{\sum idades_{homens}}{n_{homens}}$$

$$média_{homens} = \frac{4457}{95} = \mathbf{46{,}91\ anos}$$

$$mediana_{homens} = 46$$

$$moda_{homens} = 39$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n} = 123{,}51$$

Idade das Mulheres Entrevistadas			
56	44	47	49
51	61	60	57
62	32	37	58
35	49	59	55
58	47	50	53
49	37	54	55
40	37	38	35
41	45	32	45
46	41	53	48
30	44	31	45
47			

Mulheres			
30	31	32	32
35	35	37	37
37	38	40	41
41	44	44	45
45	45	46	47
47	47	48	49
49	49	50	51
53	53	54	55
55	56	57	58
58	59	60	61
62			

$$média_{mulheres} = \frac{\sum idades_{mulheres}}{n_{mulheres}}$$

$$média_{mulheres} = \frac{1913}{41} = \mathbf{46,65\ anos}$$

$$mediana_{mulheres} = 47$$

$$moda_{mulheres} = 37\ 45\ 47\ 49 \text{ (multimodal)}$$

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n} = 79{,}39$$

Coincidentemente, as médias de idades entre os gerentes homens e mulheres deram valores aproximados. A mediana expressa os valores centrais da amostra. Os valores mais ocorrentes de idades determinam a moda das amostras. Para os homens foi 39 e para as mulheres temos uma distribuição multimodal. A variância (s2) nos dois casos expressa quanto os valores estão longe da média. O desvio-padrão calculado indica qual é o "erro" se quiséssemos substituir um dos valores coletados pelo valor da média. A distribuição de dados analisada para os atletas possui baixa dispersão, pois os CV calculados estão abaixo de 25%.

De acordo com a teoria de Medidas de Tendência Central e Medidas de Dispersão, como podemos avaliar os resultados encontrados?

Pratique mais!

Desafiamos você a praticar o que aprendeu, transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

Seguidores do *Twitter*	
1. Competência de fundamento de área	Conhecer os fundamentos estatísticos básicos necessários à formação do profissional.
2. Objetivos de aprendizagem	Identificar as medidas de tendência central moda, média e mediana. Definir medidas de dispersão, como desvio-padrão e variância, sabendo interpretá-las de forma correta.

3. Conteúdos relacionados	Medidas de Tendência Central e Medidas de Dispersão.
4. Descrição da SP	Imagine que você tem 9 seguidores no Twitter e quer saber a média deles. Identificar as medidas de tendência central moda, média e mediana. Definir medidas de dispersão, como desvio-padrão e variância, sabendo interpretá-las de forma correta.  Cada um com o seguinte número de seguidores: 700 | 800 | 800 | 1000 | 1200 | 1300 | 1400 | 2000 | 2600
5. Resolução da SP	O cálculo da média será o seguinte: $média = \frac{700+800+800+1000+1200+1300+1400+2000+2600}{09}$ A média de seguidores, então, é aproximadamente 1311,1. Se você comparar com os outros dados, é uma média que resume bem o universo analisado. Afinal de contas, o que tem a menor quantidade de seguidores possui 700 e o que tem a maior quantidade possui 2600. Com os mesmos números de seguidores do Twitter, 1200 é a mediana, pois, dentre os 9 números, ocupa a posição central: 700 | 800 | 800 | 1000 | **1200** | 1300 | 1400 | 2000 | 2600 A mediana é bastante útil quando você possui uma grande quantidade de valores e muitos *outliers* (valores que fogem muito da tendência central, não sendo representativos do todo). A moda representa o valor mais comum. A moda é 800: é o valor que é repetido 2 vezes, mais do que os outros.

	700 | **800** | **800** | 1000 | 1200 | 1300 | 1400 | 2000 | 2600 Para analisar algumas variáveis quantitativas, a Moda pode ser bastante útil para identificar qual o tipo de ocorrência mais frequente. É especialmente útil quando a amplitude de valores possíveis é menor. No caso de seguidores, é bem grande, mas, no caso de "vídeos enviados" em um canal do YouTube, no qual a amplitude é menor, faz um pouco mais de sentido. A variância populacional pode ser calculada por: $$s^2 = \frac{\sum(x_i - \bar{x})^2}{n} \cong 349876{,}50$$ O desvio-padrão padrão dos seguidores é: $$s = \sqrt{s^2} = \sqrt{349876{,}50} \cong 591{,}50$$

Lembre-se

Média é o valor que aponta para onde mais se concentram os dados de uma distribuição.

Mediana de um conjunto de dados é o dado que fica no meio quando as entradas são colocadas em ordem crescente.

Moda é o valor que detém o maior número de observações, ou seja, o valor ou valores mais frequentes.

Desvio-padrão é a medida mais comum da dispersão estatística, mostra o quanto de variação ou "dispersão" existe em relação à média (ou valor esperado).

Faça você mesmo

Vamos agora saber mais alguns dados para nossa pesquisa com os seus colegas de trabalho (deve haver, no mínimo, 30 entrevistados).

1. Pergunte a idade de cada um deles.

2. Tabule.

3. Calcule a média, mediana, moda. Calcule também a variância e o desvio-padrão.

Faça valer a pena!

1. De um levantamento feito entre 100 famílias resultou a tabela a seguir.

Número de filhos	Número de famílias
0	18
1	23
2	28
3	21
4	7
5	3
Total	100

Determine o número médio de filhos, o número mediano de filhos e o número modal de filhos.

a. 3,85 filhos | 5 filhos | 1 filhos.
b. 1,25 filhos | 4 filhos | 2 filhos.
c. 1,95 filhos | 3 filhos | 5 filhos.
d. 1,85 filhos | 2 filhos | 2 filhos.
e. 1,53 filhos | 1 filhos | 3 filhos.

2. Para o conjunto seguinte, determinar, com aproximação centesimal, as seguintes medidas: a amplitude, a variância populacional, o desvio-padrão e o coeficiente de variação.

Conjunto de dados: 0,04 0,18 0,45 1,29 2,35

a. Amplitude = 0,31, variância = 0,24, desvio-padrão= 0,56 e CV = 59,90%.
b. Amplitude = 2,11, variância = 0,34, desvio-padrão= 0,36 e CV = 69,90%.
c. Amplitude = 1,31, variância = 0,44, desvio-padrão= 0,16 e CV = 79,90%.
d. Amplitude = 5,31, variância = 0,54, desvio-padrão= 0,76 e CV = 99,90%.
e. Amplitude = 2,31, variância = 0,74, desvio-padrão= 0,86 e CV = 99,90%

3. Determinar a moda do seguinte conjunto: 1, 6, 9, 3, 2, 7, 4 e 11:

a. 5
b. Amodal
c. 4 e 10
d. 18
e. 11

4. Determinar a mediana dos seguintes conjuntos:
Grupo 1: 9 | 14 | 2 | 8 | 7 | 14 | 3 | 21 | 1
Grupo 2: 0,02 | 0,25 | 0,47 | 0,01 | -0,30 | -0,5 |

a. Para o grupo 1, a mediana é 7. Para o grupo 2, a mediana é -0,30.
b. Para o grupo 1, a mediana é 3. Para o grupo 2, a mediana é 0,01.
c. Para o grupo 1, a mediana é 8. Para o grupo 2, a mediana é 0,015
d. Para o grupo 1, a mediana é 21. Para o grupo 2, a mediana é -0,5.
e. Para o grupo 1, a mediana é 9. Para o grupo 2, a mediana é 0,02.

5. A média aritmética entre dois valores é igual a 5 e a média geométrica é igual a 4. Qual a média harmônica entre esses dois valores?

a. 3,2.
b. 0,2.
c. 4,2.
d. 5,5.
e. 6,7.

6. Um livro com 50 páginas apresentou um número de erros de impressão por página conforme tabela a seguir:

Erros	Número de Páginas
0	25
1	20
2	3
3	1
5	1
Total	50

1. Qual o número médio de erros por página?
2. Qual o número modal de erros por página?

7. Dados dois grupos de pessoas, o grupo A com 10 elementos e o grupo B com 40 elementos, se o peso médio do grupo A for de 80 kg e o do grupo B for de 70 kg, então é verdade que o peso médio dos dois grupos considerados em conjunto é de 75 kg? Justifique.

Seção 1.4

Assimetria e curtose

Diálogo aberto

Após conhecermos as medidas de posição moda, mediana e média aritmética, veremos agora como elas se comportam uma em relação às outras. Isso envolve conhecermos a assimetria e a curtose que algumas bases de dados apresentam. Os objetivos de aprendizagem desta seção são identificar os dados de um conjunto quanto à assimetria, calculando o coeficiente de assimetria e classificar os dados de um conjunto quanto ao grau de achatamento da curva (curtose).

Duas distribuições podem se diferenciar uma da outra em termos de assimetria ou curtose, ou de ambas. A assimetria e o achatamento (nome técnico dado para curtose) têm importância devido a exposições teóricas relativas à dedução estatística que são comumente baseadas na hipótese de que populações são distribuídas normalmente. A precaução de erros para essa hipótese é feita utilizando as medidas de assimetria e de curtose. Veremos a medida de assimetria de Pearson, é baseada nas relações entre a média, mediana e moda. Essas três medidas são idênticas em valor para uma distribuição unimodal simétrica, mas, para uma distribuição assimétrica, a média distancia-se da moda, situando-se a mediana em uma posição intermediária, à medida que aumenta a assimetria da distribuição. Consequentemente, a distância entre a média e a moda poderia ser usada para medir a assimetria.

Para finalizar o relatório que deve ser apresentado pela agência de pesquisa contratada pela empresa multinacional do setor de alimentos sobre a entrevista com gerentes de contratação, será necessário analisar o estrato de gerentes que responderam "**Compromisso para toda a temporada**". A análise deve levar em consideração apenas os dados dos gerentes homens (95), os cálculos de assimetria ou curtose devem ser analisados e uma conclusão sobre o estrato deve ser estabelecida.

Assimile

Assimetria

Denomina-se **assimetria** o grau de afastamento de uma distribuição em relação ao eixo de simetria. Uma distribuição simétrica apresenta igualdade entre as medidas média, moda e mediana. Caso contrário, a distribuição é denominada **assimétrica**.

A assimetria de determinada base de dados possibilita analisar uma distribuição de acordo com as relações entre suas medidas de moda, média e mediana, quando observadas graficamente.

As medidas de assimetria indicam o grau de assimetria de uma distribuição de frequências unimodais em relação a uma linha vertical que passa por seu ponto mais elevado.

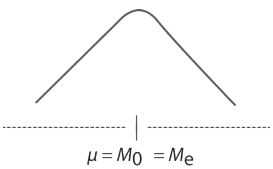
Uma distribuição com classes é **simétrica**, (Figura 1.18) quando: Média = Mediana = Moda

A Figura 1.19 mostra a assimetria positiva quando:

Moda ≤ Mediana ≤ Média.

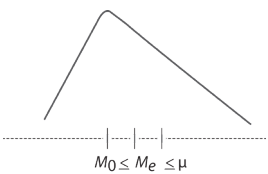
A Figura 1.20 mostra a assimetria negativa quando: Média ≤ Mediana ≤ Moda.

Figura 1.18 | Distribuição Simétrica



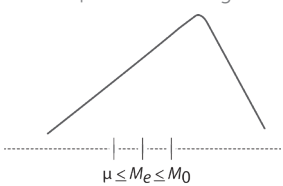
Fonte: O autor (2015).

Figura 1.19 | Assimetria positiva



Fonte: O autor (2015).

Figura 1.20 | Assimetria negativa

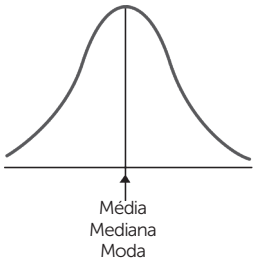


Fonte: O autor (2015).

Vamos detalhar melhor cada uma das distribuições:

Distribuição Simétrica

Figura 1.21 | Distribuição simétrica



Fonte: O autor (2015).

Uma distribuição é simétrica se tem o mesmo valor para a moda, a média e a mediana. As medidas estão no ponto central da distribuição.

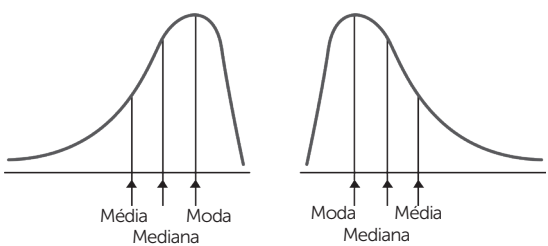
Em um gráfico, a distribuição simétrica é uma curva de frequências unimodal apresentando duas "caudas" simétricas em relação a uma linha vertical que passa por seu ponto mais alto (eixo de simetria).

Para facilitar a moda, a média aritmética e a mediana se localizam no ponto central da distribuição.

Distribuições Assimétricas

Apresentam uma curva com um único valor de moda e analisamos do ponto mais alto até a cauda, sendo essa para a direita ou para a esquerda. Para a direita caracteriza uma assimetria positiva e para a esquerda uma assimetria negativa.

Figura 1.22 | Distribuições assimétricas



Fonte: O autor (2015).

Coeficientes de Assimetria (AS)

Para quantificar o desvio de uma distribuição em relação a uma distribuição simétrica, usamos o coeficiente de assimetria. Ele nos permite comparar duas ou mais distribuições diferentes e saber qual delas é mais assimétrica. A curva será mais assimétrica, quanto maior for o coeficiente.

Coeficientes de Pearson

1º Coeficiente de Assimetria de Pearson (A):

$$A_{med} = \frac{3(\bar{x} - Med)}{s}$$

A_{Med} = Coeficiente de Assimetria de Pearson "A"

Med = Mediana

x = Média aritmética da amostra

s = Desvio-padrão da amostra

– Se $A_{med} < 0 \rightarrow$ a distribuição será Assimétrica Negativa;

– Se $A_{med} > 0 \rightarrow$ a distribuição será Assimétrica Positiva;

– Se $A_{med} = 0 \rightarrow$ a distribuição será Simétrica.

Exemplificando

Sendo a renda média de um trabalhador de uma empresa igual a R$ 1.194,50, a mediana de R$ 1.134,40 e o desvio-padrão de R$ 124,35, o coeficiente de assimetria será:

$$A_{med} = \frac{3 \cdot (1.194{,}50 - 1.134{,}40)}{124{,}35} = +1{,}45$$

Resposta: a curva é positivamente assimétrica ou assimétrica à direita.

2º Coeficiente de Assimetria de Pearson (B):

O Coeficiente de Pearson "B" mede o afastamento da simetria, expressando a diferença entre a média e a moda em relação ao desvio-padrão do grupo de medidas. O resultado desse coeficiente é obtido com o uso da seguinte fórmula:

$$A_{mod} = \frac{\bar{x} - Mod}{s}$$

Onde: A_{Mod} = Coeficiente de Assimetria de Pearson "B", Mod = Moda, x = Média aritmética da amostra e s = Desvio-padrão da amostra.

Da mesma forma que no caso anterior,, quando: A_{mod} = 0 a distribuição é simétrica. A_{mod} > 0; a distribuição é assimétrica positiva. A_{mod} < 0 a distribuição é assimétrica negativa.

Coeficiente Momento de Assimetria

Os momentos são muito importantes em Estatística para caracterizar distribuições de probabilidade. Eles dão uma ideia da tendência central, dispersão e assimetria de uma distribuição de probabilidades. Sejam m2 e m3 os momentos de segunda e de terceira ordem centrados na média, define-se o coeficiente momento de assimetria como sendo:

$$AS_m = \frac{m_3}{\sqrt{(m_2)^3}} = \frac{m_3}{s^3}$$

Coeficiente momento de assimetria (α3): é o terceiro momento abstrato.

$$\alpha_3 = \frac{M_3}{s^3}$$

O campo de variação do coeficiente de assimetria é: $-1 \leq \alpha_3 \leq +1$

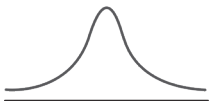
– Intensidade da assimetria:

$|\alpha_3| < 0{,}2$ - Simetria $0{,}2 < |\alpha_3| < 1{,}0$ - Assimetria fraca $|\alpha_3| > 1{,}0$ - Assimetria forte.

Curtose

Chamamos de curtose o grau de achatamento em uma distribuição de frequência que tem apenas uma moda, ou seja, uma unimodal, em relação à normal. Mede o agrupamento de valores da distribuição em torno do centro. Quanto maior o agrupamento de valores em torno do centro, maior será o valor da curtose. Os tipos de curtose são:

Quadro 1.5 | Tipos de curtose

 Leptocúrtica	**Leptocúrtica** – a medida de curtose é maior do que a da distribuição normal. A curva é mais alta que a curva da distribuição normal.
 Mesocúrtica	**Mesocúrtica** – a medida da curtose, que é igual à da distribuição normal, é chamada de curva padrão.
 Platicúrtica	**Platicúrtica** – a medida de curtose é menor do que a distribuição padrão. É uma curva mais achatada.

Fonte: Adaptado de Moretin (2010).

Coeficiente percentílico de curtose

É calculado pelas medidas de interquartil e pela amplitude entre o percentil 10º e o 90º. Seu valor para curva normal é 0,26367.

$$C_P = \frac{(Q_3 - Q_1)}{2(P_{90} - P_{10})}$$

A distribuição mesocúrtica se caracteriza quando temos o coeficiente de curtose igual a 0,263. A distribuição leptocúrtica tem o coeficiente de curtose menor do que 0,263. A distribuição platicúrtica tem o coeficiente de curtose maior do que 0,263.

Exemplificando

Sabendo-se que uma distribuição apresenta as seguintes medidas: Q1 = 24,4 cm, Q3 = 41,2 cm, P10 =20,2 cm e P90 =49,4 cm,temos:

$$C = \frac{41{,}2 - 24{,}4}{2(49{,}5 - 20{,}2)} = 0{,}2866$$

Como: 0,287>0,263, a distribuição é platicúrtica.

Sem medo de errar

Para finalizar o relatório que deve ser apresentado pela agência de pesquisa contratada pela empresa multinacional do setor de alimentos sobre a entrevista com gerentes de contratação, será necessário analisar o estrato de gerentes que responderam "**Compromisso para toda a temporada**". A análise deve levar em consideração apenas os dados dos gerentes homens (95); os cálculos de assimetria devem ser analisados e uma conclusão sobre o estrato deve ser estabelecida.

Amostra ordenada das idades dos homens entrevistados				
30	30	30	30	30
31	31	31	32	32
32	32	32	33	33
33	33	34	34	34
36	36	36	37	38
39	39	39	39	39
39	39	40	41	41
42	42	43	43	43
43	44	44	44	45

45	45	46	46	48
48	49	50	50	51
51	51	51	52	52
52	53	54	54	54
55	56	56	56	57
57	57	58	59	59
59	60	60	60	60
60	61	61	61	61
62	62	62	63	63
64	64	64	65	65

$$média_{homens} = \frac{\sum idades_{homens}}{n_{homens}}$$

$$média_{homens} = \frac{4457}{95} = \mathbf{46{,}91\ anos}$$

$$mediana_{homens} = 46$$

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n} = 123{,}51$$

$$s = \sqrt{s^2} = \sqrt{123{,}51} = 11{,}11$$

$$A_{med} = \frac{3(\bar{x} - Med)}{s} = \frac{3\,(46{,}91 - 46)}{11{,}11} = +0{,}24$$

A representação gráfica é:

Pelo coeficiente de Assimetria calculado acima temos uma distribuição assimétrica positiva.

Os valores calculados para média, mediana e moda possibilitam afirmar que é uma distribuição com assimetria positiva, pois:

$$\text{Moda} \leq \text{Mediana} \leq \text{Média}$$

A curva de assimetria mostra a realidade da amostra estudada? Como as medidas média, moda e mediana podem ser úteis na análise de uma amostra?

Lembre-se

A assimetria analisa a curva de frequência de forma horizontal, relacionando a sua característica com a configuração de uma distribuição simétrica. A curtose analisa a curva de frequência de forma vertical, relacionando a sua característica com a característica de uma distribuição normal.

As **medidas de assimetria e curtose** são medidas independentes e que não se influenciam mutuamente.

Avançando na prática

Pratique mais!

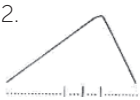
Instrução
Desafiamos você a praticar o que aprendeu, transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

<table>
<tr><th colspan="2">IMC – Índice de Massa Corpórea</th></tr>
<tr><td>1. Competência de fundamento de área</td><td>Conhecer os fundamentos estatísticos básicos necessários à formação do profissional.</td></tr>
<tr><td>2. Objetivos de aprendizagem</td><td>Identificar os dados de um conjunto quanto à assimetria, calculando o coeficiente de assimetria.</td></tr>
<tr><td>3. Conteúdos relacionados</td><td>Assimetria</td></tr>
<tr><td>4. Descrição da SP</td><td>No IMC – índice de massa corpórea medido em um grupo de 40 atletas homens e mulheres de uma academia foi calculado um desvio-padrão de 6,17. Foram retiradas de forma aleatória as seguintes amostras de IMC. Como está o desvio-padrão dessa amostra comparado ao desvio-padrão da população? Calcule o coeficiente de assimetria da amostra e interprete os resultados.
<table>
<tr><td>19,6</td><td>23,8</td><td>19,6</td><td>29,1</td><td>25,2</td></tr>
<tr><td>21,4</td><td>22</td><td>27,5</td><td>33,5</td><td>20,6</td></tr>
<tr><td>29,9</td><td>17,7</td><td>24</td><td>28,9</td><td>37,7</td></tr>
</table></td></tr>
<tr><td>5. Resolução da SP</td><td>Pelas fórmulas apresentadas na seção anterior, temos: Comparando com o valor para a população, que foi de 6,17, os valores encontrados para o desvio-padrão da amostra estão bem próximos.
<table>
<tr><td>Média</td><td>25,37</td><td>Mediana</td><td>24</td><td>Desvio-padrão</td><td>5,66</td></tr>
</table>
Para o coeficiente de simetria, temos:
$$A_{med} = \frac{3(\bar{x} - Med)}{s} = \frac{3\,(25{,}37 - 24)}{5{,}66} \cong +0{,}73$$
Distribuição Assimétrica Positiva.</td></tr>
</table>

Faça você mesmo

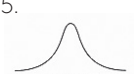
Classifique as curvas de distribuição segundo Assimetria e Curtose:

1. 

2. 

3. 

4. 

5. 

6. 

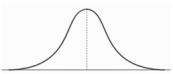
Faça você mesmo

1. Simétrica
2. Assimétrica Positiva
3. Assimétrica Negativa
4. Mesocúrtica
5. Platicúrtica
6. Leptocúrtica

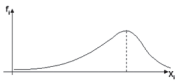
Faça valer a pena

Avaliação direcionada à compreensão dos aspectos conceituais dos conteúdos.

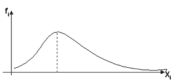
1. Analisando as curvas abaixo, marque a resposta correta.



(I)



(II)



(III)

a. A curva I é assimétrica.
b. A curva II é assimétrica positiva.
c. A curva I é simétrica.
d. A curva III é simétrica positiva.
e. Nenhuma das alternativas está correta.

2. Uma maternidade está analisando a idade das mulheres que tiveram o seu primeiro filho. Os dados obtidos são:
25 | 23 | 21 | 28 | 41 | 18 | 19 | 23 | 20 | 22 | 23
Como podemos classificar os dados em relação à assimetria?
Resposta: Média = 23,9, Mediana = 23, Moda = 23, Desvio-padrão = 6.

a. Distribuição assimétrica positiva.
b. Distribuição assimétrica negativa.
c. Distribuição simétrica.
d. Distribuição de curtose.
e. Distribuição simétrica de todas as distribuições.

3. Uma distribuição de frequência apresenta as seguintes medidas: Média = 48,1, Md = 47,5 e s = 2. Qual o seu coeficiente de assimetria?

a. 0,90.
b. 0,65.
c. 0,75.
d. 0,95.
e. 0,08.

4. Sabendo-se que uma distribuição apresenta as seguintes medidas, como podemos classificá-la?
Q1=24,4 cm; Q3= 41,2 cm; P10= 20,2 cm e P90= 49,5 cm.

a. Curtose – mesocúrtica..
b. Curtose – leptocúrtica.
c. Curtose – platicúrtica.
d. Assimétrica positiva.
e. Assimétrica negativa.

5. Considere os seguintes resultados relativos à distribuição A:

Distribuições	Q_1	Q_3	P_{10}	P_{90}
A	4	16	3	26

Classifique a distribuição.

a. Curtose – mesocúrtica.
b. Curtose – leptocúrtica.
c. Curtose – platicúrtica.
d. Assimétrica positiva.
e. Assimétrica negativa.

6. Considere as seguintes medidas, relativas a três distribuições de frequência:

Distribuições	Q_1	Q_3	P_{10}	P_{90}
A	814	935	772	1012
B	63,7	80,3	55,0	86,6
C	28,8	45,6	20,5	49,8

Calcule os respectivos graus de curtose e classifique cada uma das distribuições em relação à curva normal.

7. Considere os seguintes resultados, relativos à distribuição P

Distribuição	Q_1	Q_3	P_{10}	P_{90}
P	30	45	26	49

Determine o coeficiente de curtose e classifique a distribuição.

Referências

BARBETTA, P. A.; BORNIA, A. C. R. **Estatística para cursos de engenharia e informática**. 3. ed. São Paulo: Atlas, 2010.

CARVALHO, T. M. de. **Variabilidade espacial de propriedades físico-hídricas de um latossolo vermelho-amarelo através da geoestatística.** 1991. 84 p. Dissertação (Mestrado) - Escola Superior de Agricultura de Lavras, Universidade Federal de Lavras, Lavras, 1991.

GROSSI SAD, J. H. **Fundamentos sobre variabilidade dos depósitos minerais**. Rio de Janeiro: DNPM/CPRM-GEOSOL, 1986.

HINES, W.; MONTGOMERY, D. C.; GOLDSMAN, Dave; BORROR, Connie M. **Probabilidade e estatística na engenharia**. 4. ed. Rio de Janeiro: LTC, 2006.

JOHNSON, R.; KOBY, P. **Estatística**. São Paulo: Cengage Learning, 2013.

LARSON, R.; FARBER, B. **Estatística aplicada.** 4. ed. São Paulo: Pearson, 2010.

MARCONI, M. D. A.; LAKATOS, E. M. **Técnicas de pesquisa**: planejamento e execução de pesquisas, amostragens e técnicas de pesquisas, elaboração, análise e interpretação de dados. 3. ed. São Paulo: Atlas, 1996.

MOORE, D. S. **A estatística básica e sua prática**. 6. ed. Rio de Janeiro: LTC, 2014.

MORETTIN, L. G. **Estatística básica**: probabilidade e inferência. São Paulo: Pearson, 2010.

PINHEIRO, J. I. D. **Probabilidade e estatística**. Rio de Janeiro: Elsevier, 2012.

SPIEGEL, M. R. **Estatística**. 3. ed. São Paulo: Makron Books, 1993.

WALPOLE, R. E. **Probabilidade e estatística para engenheiros e ciências**. 8. ed. São Paulo: Pearson-Prentice Hall, 2009. v. 1.

WELLS, H. G. **Algumas citações interessantes**. Disponível em: <http://www.inf.ufsc.br/~marcelo/citar.html>. Acesso em: 25 jun. 2015.

Métodos Tabulares e Métodos Gráficos

Convite ao estudo

Nesta unidade, veremos conteúdos que são necessários para a realização dos métodos tabulares e métodos gráficos. Os objetivos desta unidade são: compreender as medidas separatrizes e sua utilização em estatística; construir e interpretar o boxplot; utilizar as tabelas de frequência e os diagramas de dispersão para melhor interpretação dos dados estatísticos; utilizar o coeficiente de correlação linear e a regressão linear para organizar os dados coletados e para a interpretação e análise desenvolvendo o raciocínio crítico sobre o fenômeno em questão.

Com esses objetivos, a competência geral da disciplina, que é conhecer os fundamentos estatísticos básicos necessários à formação do profissional, será desenvolvida nesta unidade.

A estatística nos auxilia em todos as áreas da nossa vida. Continuamente vemos a utilização de gráficos, porcentagens e pesquisas que nos dão um panorama sobre nossas situações cotidianas.

Você já se deparou com revistas especializadas em saúde que nos mostram uma porcentagem da população com um certo tipo de doença? Você já ficou tendencioso a não consumir algum tipo de alimento ou a consumir por causa de uma dessas pesquisas? Essas pesquisas têm muito a dizer sobre nossa rotina, sobre nosso estilo de vida e nossa expectativa de vida.

Falando de saúde, podemos considerar o sistema musculoesquelético que é muito importante para o ser humano. Além de nos ajudar em atividades atléticas, é responsável por

movimentos simples, como levantar de uma cadeira ou pegar um objeto em uma prateleira. Pode parecer bobagem, mas com o envelhecimento, atividades rotineiras tornam-se um desafio. Para se ter uma ideia, um jovem de 20 anos pode ter perdido 50% da sua massa muscular quando chegar aos 90 anos. E não se assuste com os noventa, se você tem 50 anos hoje provavelmente chegará a essa idade. Projeções preveem que em 2050 a expectativa de vida será próxima de 90 anos. Essa perda de massa muscular é responsável pela redução da força, aumento do risco de quedas e lentidão nos movimentos.

Essa preocupação com a massa muscular no envelhecimento levou um educador físico a fazer uma pesquisa com seus clientes. As informações levantadas pelo educador físico foram a idade e a quantidade de massa muscular. É esperado que a massa muscular de uma pessoa diminua com a idade.

Para estudar essa relação, o educador físico selecionou 18 mulheres, com idade entre 40 e 79 anos, e coletou informações sobre a idade e a massa muscular.

Você, será capaz, ao final desta unidade, de elaborar o relatório que conterá a tabela de idade dos clientes e a massa muscular medida, o diagrama de frequência de idades, o gráfico boxplot, o diagrama de dispersão com suas respectivas interpretações, o coeficiente de regressão e a reta de regressão linear. Todos os resultados apresentados auxiliarão o educador físico a tratar esse grupo de clientes a fim de terem menos perda de massa muscular ao longo do envelhecimento.

Com o estudo da unidade, poderemos esclarecer algumas perguntas: para uma academia que só atende mulheres, essa pesquisa é relevante? E se a academia atende homens e mulheres, essa pesquisa é representativa? A que conclusões podemos chegar ao analisarmos os resultados obtidos?

Pronto para começar?

Seção 2.1

Medidas separatrizes e *boxplot*

Diálogo aberto

As medidas separatrizes são valores que separam o rol (os dados ordenados) em quatro (quartis), dez (decis) ou em cem (percentis) partes iguais, para essas separações os dados devem estar ordenados. Medidas separatrizes são medidas intuitivas, de fácil compreensão e que também podem ser utilizadas para construir medidas de dispersão. Indicam limites para proporções de observações em um conjunto.

O boxplot, ou diagrama de caixa, é um gráfico que capta importantes aspectos de um conjunto de dados através do seu resumo dos cinco números, formado pelos seguintes valores: valor mínimo, primeiro quartil, segundo quartil, terceiro quartil e valor máximo.

Os objetivos de aprendizagem desta seção são compreender as medidas separatrizes e sua utilização em estatística e construir e interpretar o boxplot.

Com a preocupação com a perda de massa muscular, que é responsável pela redução da força, aumento do risco de quedas e lentidão nos movimentos, um educador físico fez uma pesquisa com seus clientes. As informações levantadas pelo educador físico foram a idade e a quantidade de massa muscular. É esperado que a massa muscular de uma pessoa diminua com a idade.

Para estudar essa relação, o educador físico selecionou 18 mulheres, com idade entre 40 e 79 anos, e coletou informações sobre a idade e a massa muscular (Y), conforme a tabela 2.1.

Tabela 2.1 | Dados da pesquisa: idade x massa muscular

Idade (X)	Massa muscular (Y)
71.0	82.0
64.0	91.0
43.0	100.0
67.0	68.0
56.0	87.0

73.0	73.0
68.0	78.0
56.0	80.0
76.0	65.0
65.0	84.0
45.0	116.0
58.0	76.0
45.0	97.0
53.0	100.0
49.0	105.0
78.0	77.0
73.0	73.0
68.0	78.0

Fonte: O autor (2015)

Você deve mostrar as medidas de quartis da amostra e montar o boxplot das idades das mulheres que estão sendo estudadas. Esses cálculos e representações serão importantes para a análise que auxiliará o educador físico.

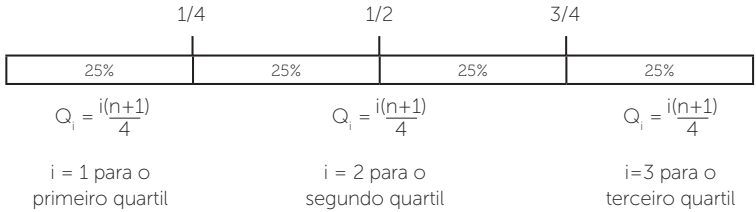
Assimile

Medidas Separatrizes
As medidas separatrizes são números que dividem a sequência ordenada de dados em partes que contêm a mesma quantidade de elementos da série.

As medidas **separatrizes** começam pela mediana, que divide a sequência ordenada em dois grupos, cada um deles contendo 50% dos valores da sequência. Além da mediana, as outras medidas separatrizes são: quartis, quintis, decis e percentis.

Quartis

Se uma série for dividida em quatro partes, o primeiro quartil será correspondente a 25% dos elementos e o segundo quartil a 50% de seus valores à direita. O Q_2 é a Mediana da série. O terceiro quartil Q_3 obedece à mesma regra dos anteriores.



Quintis

Ao dividir a série ordenada em cinco partes, cada uma ficará com 20% de seus elementos. Os elementos que separam esses grupos são chamados de quintis. Assim, o primeiro quintil, indicado por K1, separa a sequência ordenada, deixando 20% de seus valores à esquerda e 80% de seus valores à direita. De modo análogo são definidos os outros quintis.

Decis

Ao dividir a série ordenada em dez partes, cada uma ficará com 10% de seus elementos. Os elementos que separam esses grupos são chamados de decis. Assim, o primeiro decil, indicado por D1, separa a sequência ordenada, deixando 10% de seus valores à esquerda e 90% de seus valores à direita. Os outros decis são calculados da mesma forma.

Percentis

Ao dividir a série ordenada em cem partes, cada uma ficará com 1% de seus elementos. Os elementos que separam esses grupos são chamados de centis ou percentis. Assim, o primeiro percentil, indicado por P1, separa a sequência ordenada, deixando 1% de seus valores à esquerda e 99% de seus valores à direita. Do mesmo modo, definimos os outros percentis. Verifica-se que os quartis, quintis e decis são múltiplos dos percentis, então basta estabelecer a fórmula de cálculo de percentis. Todas as outras medidas podem ser identificadas como percentis, ou seja:

Tabela 2.2 | Percentis

Percentis	Quartis	Quintis	Decis
P_{10}			D1
P_{20}		K1	D2
P_{25}	Q1		
P_{30}			D3
P_{40}		K2	D4
P_{50}	Q2		D5
P_{60}		K3	D6
P_{70}			D7
P_{75}	Q3		
P_{80}		K4	D8
P_{90}			D9

Fonte: O autor (2015)

Cálculo da separatriz:

Identifica-se a medida que se pretende obter com o percentil correspondente, Pi. Calcula-se i% de n para localizar a posição do percentil i no Rol, ou seja:

$$P_i = \frac{i \times n}{100}$$

Boxplot

A partir das medidas separatrizes, constrói-se também um gráfico chamado gráfico de caixas (em inglês, *boxplot*), que ilustra os principais aspectos da distribuição, tomando por base essas medidas robustas.

O *boxplot* é um gráfico muito útil também na comparação de distribuições, é formado basicamente por um retângulo vertical (ou horizontal). O comprimento do lado vertical (ou horizontal) é dado pelo intervalo interquartil (em que estamos trabalhando com um retângulo vertical).

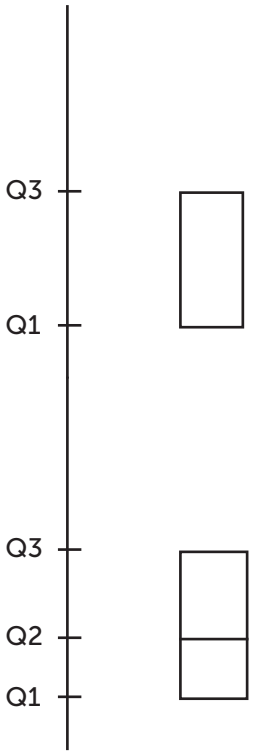
O tamanho do outro lado é indiferente, sugerindo-se apenas uma escala razoável. Na altura da mediana, traça-se uma linha, dividindo o retângulo em duas partes.

Note que aí já temos representados 50% da distribuição e também já temos ideia da sua **assimetria**. Para representar os 25% restantes em cada cauda da distribuição, temos que cuidar primeiro da presença de possíveis *outliers* ou valores discrepantes.

Um dado será considerado *outlier* se ele for menor que Q1 - 1,5 *IQ* ou maior que Q3 + 1,5 *IQ*, como mostra a figura a seguir.

Para representar o domínio de variação dos dados que não são *outliers*, traça-se a partir do retângulo, uma linha para cima e outra para baixo até o ponto mais remoto que não seja *outlier*. Esses pontos são chamados juntas.

O intervalo interquartil IQ é a distância entre o terceiro e o primeiro quartis, isto é:



$$IQ = Q3 - Q1$$

Pela definição dos quartis, sabe-se que entre os valores Q1 e Q3 sempre temos 50% das observações. Assim, quanto maior for o intervalo interquartil, mais dispersos serão os dados. Quanto aos *outliers*, eles são representados individualmente por um X (ou algum outro tipo de caractere), explicitando, de preferência, os seus valores, mas com quebra de escala no eixo.

O *boxplot* representa graficamente dados de forma resumida em um retângulo em que as linhas da base e do topo são o primeiro e o terceiro quartis, respectivamente. A linha entre essas é a mediana. O *boxplot*, além de apresentar a dispersão dos dados, torna-se útil também para identificar a ocorrência destes valores como sendo os que caem fora dos limites estabelecidos pelos valores adjacentes superior e inferior.

Complemente seus estudos

O Portal Action traz uma explicação sobre a construção do boxplot que vale a pena conhecer. Vamos lá? Disponível em: <http://www.portalaction.com.br/estatistica-basica/31-boxplot>. Acesso em: 8 jul. 2015.

Vocabulário

Assimetria – Que não tem simetria; não divisível em metade por um eixo longitudinal.

Rol – Lista, relação. Números ordenados.

Separatrizes – Qualquer valor de uma variável aleatória para o qual a função de distribuição assume valores múltiplos inteiros de uma fração dada.

Exemplificando

Visando ao aumento de peso de crianças carentes, uma dieta foi aplicada a 12 crianças. Os ganhos (valores positivos) e perdas (valores negativos) de peso, após a dieta, são descritos a seguir (em quilogramas):

11,2 / 6,3 / 7,8 / 5,9 / 5,6 / 4,6 / 2,5 /-0,7 / 3,0 / 6,2 / 6,0 / 3,6

Calcule as medidas separatrizes e construa o gráfico boxplot da distribuição de valores apresentados.

Dados ordenados:

-0,7	2,5	3,0	3,6	4,6	5,6	5,9	6,0	6,2	6,3	7,8	11,2

Medidas Separatrizes

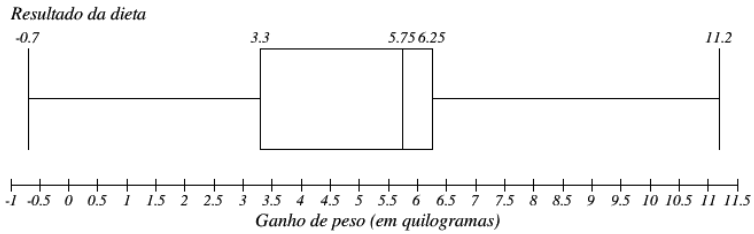
Q1 = 3,3 kg Q2 = 5,75 kg (mediana) Q3 = 6,25

Mínimo = -0,7

Máximo = 11,2

IQ = Q3 – Q1 = 6,25 – 3,3 = 2,95

Exemplificando



Faça você mesmo

Uma modista do São Paulo Fashion Week forneceu uma tabela de medidas de cintura das modelos e essas medidas foram tabuladas em centímetros da seguinte forma:

83	81	77	75	72	70	70	69	68	68	67	67
66	66	66	65	64	63	62	61	61	60	58	58

Calcule as medidas separatrizes e construa o gráfico boxplot da distribuição de valores apresentados.

1. Ordene os dados.

2. Calcule o primeiro quartil, o terceiro quartil, a mediana, o interquartil, o valor máximo e o valor mínimo.

3. Utilize o site indicado a seguir para plotar o boxplot <http://www.imathas.com/stattools/boxplot.html>. Acesso em: 8 jul. 2015.

Atenção

Quando estiver trabalhando com medidas separatrizes, utilize o rol de dados, ou seja, os dados ordenados.

Agora iremos mostrar as medidas de quartis da amostra e montar o boxplot das idades das mulheres que estão sendo estudadas.

Ordena-se os dados:

Tabela 2.3 | Dados da pesquisa ordenados

Idade (X)	Massa muscular (Y)
43.0	100.0
45.0	116.0
45.0	97.0
49.0	105.0
53.0	100.0
56.0	87.0
56.0	80.0
58.0	76.0
64.0	91.0
65.0	84.0
67.0	68.0
68.0	78.0
68.0	78.0
71.0	82.0
73.0	73.0
73.0	73.0
76.0	65.0
78.0	77.0

Fonte: O autor (2015)

Sendo n=18

$Q_1 = 1\ \frac{1}{4}\ _{(18+1)} = 53$

$Q_2 = 2\ \frac{1}{4}\ _{(18+1)} = 65$

$Q_3 = 3\ \frac{3}{4}\ _{(18+1)} = 71$

Calcula-se o IQ

$IQ = Q3 - Q_1 = 71-53 = 18$

Passo 1 - Calcula-se o 1º Quartil.

Passo 2 – Calcula-se o 3º Quartil.

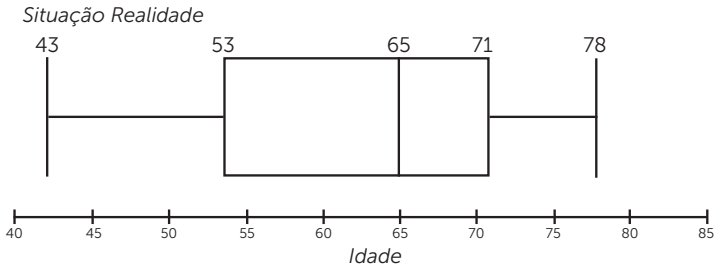
Passo 3 – Define-se a Mediana.

Passo 4 – Define-se o valor mínimo.

Passo 5 – Define-se o valor máximo.

Passo 6 – Calcula-se o interquartil.

Utilize o site indicado a seguir para gerar o boxplot: <http://www.imathas.com/stattools/boxplot.html>. Acesso em: 8 jul. 2015.



Lembre-se

O *boxplot* é gerado a partir das medidas separatrizes. Ele é chamado gráfico de caixas (em inglês, *boxplot*) e ilustra os principais aspectos da distribuição.

Avançando na prática

Pratique mais!	
Instrução Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.	
IPCA - Índice de Preços ao Consumidor Amplo	
1. Competência de fundamento de área	Conhecer os fundamentos estatísticos básicos necessários à formação do profissional.
2. Objetivos de aprendizagem	Compreender as medidas separatrizes e sua utilização em estatística e construir e interpretar o *boxplot*.
3. Conteúdos relacionados	Medidas Separatrizes e *boxplot*.

4. Descrição da SP

Na tabela a seguir, apresenta-se algumas medidas do IPCA (Índice de Preços ao Consumidor Amplo), que são variações mensais calculadas pelo IBGE para o ano de 2052, trata-se da inflação para os meses do ano.

fev	mar	abr	mai	jun	jul	ago	set	out	nov
1,05	1,10	0,56	0,30	0,19	1,09	0,56	0,31	1,09	0,95

Para esses dados, é necessário calcular as medidas separatrizes e criar o *boxplot* para a distribuição de valores apresentados. Interprete os dados apresentados no *boxplot*.

5. Resolução da SP

Vamos ordenar os valores:

0,19	0,30	0,31	0,56	0,56	0,95	1,05	1,09	1,09	1,10

Vamos calcular o 1º, 2º e 3º quartil.

$Q_1 = \frac{1(10+1)}{4} = 2{,}75$ Procuramos a posição 3 $\rightarrow$ $Q_1 = 0{,}31$

$Q_1 = \frac{2(10+1)}{4} = 5{,}5$ Procuramos a posição 5 $\rightarrow$ $Q_2 = 0{,}56$

$Q_1 = \frac{3(10+1)}{4} = 8{,}25$ Procuramos a posição 8 $\rightarrow$ $Q_1 = 1{,}09$

Vamos calcular o valor interquartil:

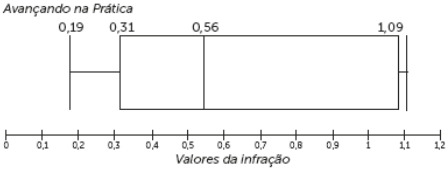
$IQ = Q_3 - Q_1 = 1{,}09 - 0{,}31 = 0{,}78$

Observando a distribuição, temos:

Valor min = 0,19 e Valor máx = 1,10

Criamos o *boxplot* demarcando um eixo com todos os valores encontrados acima.

Avançando na Prática



A caixa contém 50% dos dados. O limite superior é 1,09 e indica 75% dos dados e o limite inferior é 0,31 e indica 25% dos dados. A distância entre os pontos é conhecida como interquartil, que no nosso caso é 0,78.

A linha na caixa é a mediana, que calculamos em 0,56. A distribuição de dados é assimétrica, pois a linha não está no centro da caixa.

Faça você mesmo

Uma editora verificou os livros que estão na produção e, na última semana, os erros de editoração foram contabilizados por dia, e apresentados na tabela a seguir.

39	90	25	34	12	24	19

Calcule as medidas separatrizes e construa o gráfico boxplot da distribuição de valores apresentados.

Lembre-se

As medidas separatrizes são: **Quartis** - Ao dividir a série ordenada em quatro partes, cada uma ficará com 25% de seus elementos. **Quintis** - Ao dividir a série ordenada em cinco partes, cada uma ficará com 20% de seus elementos. **Decis** - Ao dividir a série ordenada em dez partes, cada uma ficará com 10% de seus elementos. **Percentis** - Ao dividir a série ordenada em cem partes, cada uma ficará com 1% de seus elementos.

O ***boxplot*** é um gráfico muito útil também na comparação de distribuições. É formado basicamente por um retângulo vertical (ou horizontal). O comprimento do lado vertical (ou horizontal) é dado pelo intervalo interquartil (em que estamos trabalhando com um retângulo vertical), medida que é calculada subtraindo Q1 de Q3.

Faça valer a pena!

1. Durante um dia inteiro de trabalho, foi contabilizado o número de vendas realizadas pelos vendedores. Os dados foram tabulados da seguinte forma:
Vendas: {4, 1, 8, 0, 11, 10, 7, 8, 6, 2, 9, 12}
Qual será o valor do primeiro quartil para a distribuição apresentada?
a) 3.
b) 4.
c) 5.
d) 6.
e) 7.

2. Para a mesma distribuição, qual será o valor do segundo quartil (mediana)?
a) 3,0.
b) 4,5.
c) 5,5.
d) 6,0.
e) 7,5.

3. Para a mesma distribuição, qual será o valor do terceiro quartil?
a) 7,0.
b) 8,5.
c) 9,5.
d) 10,5.
e) 15,0.

4. O valor de interquartil pode ser calculado por IQ = Q3-Q1. Para essa distribuição, qual é o valor de IQ?
a) 6,5.
b) 8,5.
c) 9,0.
d) 10,0.
e) 11,5.

5. Para a construção do boxplot, precisamos utilizar todos os cálculos do primeiro, segundo e terceiro quartis, como fizemos nos exercícios anteriores. Quais são os valores máximos e mínimos, respectivamente, para essa distribuição?
a) 3 e 12.
b) 1 e 10.
c) 5 e 6.
d) 10 e 4.
e) 12 e 0.

6. Construa o boxplot da distribuição.

7. Que conclusões tiramos ao analisar um boxplot?

Seção 2.2

Tabelas de frequências e diagrama de dispersão

Diálogo aberto

Uma vez que se conhece o conjunto de dados, sabe-se quais os valores que serão trabalhados e como essa distribuição pode ser classificada, podemos utilizar ferramentas para análises desses dados que facilitem a tomada de decisões.

As tabelas de frequências e os diagramas de dispersão são ferramentas que auxiliam essas análises, pois, pela definição, a distribuição de frequências é um arranjo tabular de um conjunto de dados em grupos, classes ou níveis, as frequências são as vezes que cada valor aparece na distribuição. O diagrama de dispersão é um gráfico em que pontos no espaço cartesiano XY são usados para representar simultaneamente os valores de duas variáveis quantitativas medidas em cada elemento do conjunto de dados.

Os diagramas de dispersão são indicados para análises estatísticas quando se tem interesse em mostrar a relação entre dois grupos de dados. O objetivo de aprendizagem desta seção é utilizar as tabelas de frequência e os diagramas de dispersão para melhor interpretação dos dados estatísticos.

Os dados levantados pela pesquisa do educador físico foram dispostos na tabela a seguir e mostram as idades das clientes e também a massa muscular.

Você deve organizar uma tabela de frequência para as idades com intervalos de classe de 5 anos.

Com essas informações, construa o diagrama de dispersão e interprete-o.

Como o diagrama de dispersão pode auxiliar na interpretação da pesquisa? A tabela de frequência tem qual importância para a análise de dados?

Ao final da seção, você será capaz de elaborar a tabela de frequência e o diagrama de dispersão para o relatório do educador físico.

Tabela 2.4 | Dados da Pesquisa – Idade x Massa Muscular

Idade (X)	Massa muscular (Y)
43	100
45	116
45	97
49	105
53	100
56	87
56	80
58	76
64	91
65	84
67	68
68	75
68	78
71	82
73	73
73	65
76	65
78	77

Fonte: O autor (2015).

Não pode faltar

Tabela de Frequência

Para encontrar as respostas de uma pesquisa, não basta apenas que sejam feitas as entrevistas ou os levantamentos de dados, é necessário também que eles estejam organizados de forma a facilitar o entendimento do leitor.

A primeira etapa após o levantamento dos dados é organizar uma tabela contendo todas as variáveis e suas respostas, mas isso ainda não é o suficiente, é preciso, com esses dados todos reunidos, montar uma Tabela de Frequências, ou seja, montar uma tabela para cada variável.

A Tabela de Frequências indica a frequência observada (relativa ou absoluta). Mostra a frequência com que cada observação aparece nos dados (também pode se referir a classes de observações).

Frequência absoluta: é definida pelo número de eventos analisados de um tipo.

Frequência relativa: é a porcentagem dos eventos que se tem interesse pelo total de eventos observados.

$$\frac{X_i}{n} \times 100$$

Frequência Cumulativa: é a medida de valores até um ponto e não mais de um único valor. Mede frequência absoluta ou relativa até um certo ponto e não apenas em um valor (LARSON, 2010).

Exemplificando

Em um estudo com mulheres que fazem exercícios todos os dias, queremos saber a quantidade de mulheres que está em cada categoria de exercício. Os dados foram tabulados da seguinte forma:

Tabela 2.5 | Estudo com mulheres

Exercício	Frequência Absoluta	Frequência Relativa	Frequência Cumulativa Relativa
Nenhum	185	$(\frac{185}{462})\times 100\% = 40{,}04\%$	40,04%
Mudando	213	$(\frac{213}{462})\times 100\% = 46{,}10\%$	86,14%
Baixo para moderado	49	$(\frac{49}{462})\times 100\% = 10{,}61\%$	97,75%
Alto	15	$(\frac{15}{462})\times 100\% = 3{,}25\%$	100,00%

Fonte: O autor (2015).

A distribuição de frequências visa representar um grande conjunto de informações, sem perder as suas principais características. Após a coleta de dados, é necessário sumarizar, sintetizar, representar e expor o fenômeno com a finalidade de se obter suas características quantitativas, visando à descrição numérica do fenômeno.

A ideia fundamental para sumarizar um conjunto de observações consiste na criação de grupos, classes ou níveis, com intervalos, geralmente regulares, contendo todas as observações. Os níveis, grupos e classes deverão ser mutuamente exclusivos e todos os valores deverão ser enquadrados nos respectivos intervalos.

A distribuição de frequências pode ser definida como um arranjo tabular de um conjunto em grupos, classes ou níveis com as suas respectivas frequências que representam o número de observações pertencentes a cada classe. A distribuição de frequências é uma série

cujos dados numéricos relativos a um fenômeno estão reunidos em intervalos de valores iguais ou não.

Na distribuição de frequências, os dados estatísticos estão dispostos ordenadamente em linhas e colunas, permitindo-se assim sua leitura no sentido horizontal e vertical. Além disso, o tempo, o local e a espécie do fenômeno não variam.

Uma tabela de frequências é uma tabela em que se procura fazer corresponder os valores observados da variável em estudo e as respectivas frequências. Essas tabelas de frequências podem representar tanto valor individual quanto valores agrupados.

Exemplificando

A distribuição de frequências apresentada na tabela a seguir é relativa aos salários de uma amostra de 100 empregados de uma construtora da capital de Minas Gerais.

Tabela 2.6 | Salários de uma amostra de empregados

Nº classes	Salários	Empregados
1ª	400 a 450	4
2ª	451 a 500	10
3ª	501 a 550	18
4ª	551 a 600	25
5ª	601 a 650	20
6ª	651 a 700	13
7ª	701 a 750	7
8ª	751 a 800	3
Total		100

Fonte: DRH

Os salários do pessoal da construtora incluem algumas categorias de trabalhadores, desde pedreiros, carpinteiros e pintores, numa amostra de 100 empregados. A tabela foi construída em 8 grupos salariais, com salários variando de R$ 400,00 a R$ 800,00. A primeira classe é composta de salários de R$ 400,00 a R$ 450,00, e assim por diante, variando de 50 em 50 reais.

Assimile

Diagramas de Dispersão

Diagrama ou gráfico de dispersão é uma ferramenta que indica a existência, ou não, de relações entre variáveis de um processo e sua intensidade, representando duas ou mais variáveis, uma em função da outra. Deve ser usada quando se necessita visualizar o que acontece com uma variável quando outra variável se altera, podendo identificar uma possível relação de causa e efeito entre elas.

O diagrama de dispersão é um gráfico em que pontos no espaço cartesiano XY são usados para representar simultaneamente os valores de duas variáveis quantitativas medidas em cada elemento do conjunto de dados.

A tabela e a figura a seguir mostram um esquema do desenho do diagrama de dispersão. Neste exemplo, foram medidos os valores de duas variáveis quantitativas, X e Y, em quatro indivíduos. O eixo horizontal do gráfico representa a variável X e o eixo vertical representa a variável Y.

O diagrama de dispersão é usado principalmente para visualizar a relação/associação entre duas variáveis, mas também é muito útil para:

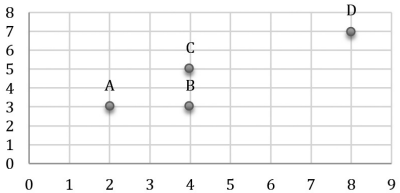
- Comparar o efeito de dois acontecimentos no mesmo indivíduo;
- Verificar o efeito antes/depois de um evento.

Tabela 2.7 | Exemplo de dados para dispersão

Indivíduos	Variável X	Variável Y
A	2	3
B	4	3
C	4	5
D	8	7

Fonte: O autor

Figura 2.1 | Diagrama de Dispersão



Fonte: O autor

Reflita

Diagramas de dispersão são gráficos que permitem a identificação entre causas e efeitos, para avaliar o relacionamento entre variáveis. O diagrama de dispersão é a etapa seguinte do diagrama de causa e efeito, pois verifica se há uma possível relação entre as causas, isto é, mostra-nos se existe uma relação, e em que intensidade.

Faça você mesmo

Dada a tabela a seguir, fazer diagramas de dispersão que auxiliem na construção de conclusões sobre a amostra estudada:

Tabela 2.8 | Dados pesquisados

Idade	Peso	Altura
17	50	1,50
18	55	1,58
20	72	1,62
25	62	1,65
17	70	1,71
38	83	1,72
54	80	1,78
64	72	1,80
37	52	1,55
41	95	1,90
28	62	1,65
19	79	1,82
46	85	1,82
74	79	1,90
58	85	1,90
60	89	2,00

Fonte: O autor (2015).

O primeiro diagrama deve ser Idade x Peso. O segundo deve ser Idade x Altura.

Vocabulário

Dispersão - Medida de variabilidade de uma distribuição em relação à média.

Quantitativas - Relativo ao indicativo da quantidade.

Sumarizar - Ato de reunir, de maneira resumida, os principais indicativos, assuntos e informações de forma a facilitar o que se pretender ler, estudar ou entender.

Complemente seus estudos

Caro aluno, utilize o link a seguir para aprender um pouco mais sobre os métodos tabulares e os métodos gráficos. O artigo traz exemplos que facilitam a sua compreensão sobre o assunto: Disponível em: <http://www.sboc.org.br/app/webroot/leitura-critica/LEITURA-CRITICA_C3.pdf>. Acesso em: 8 jul. 2015.

Atenção

Para criar os diagramas de dispersão mais facilmente, você pode utilizar o *software* Excel. No *link* há uma breve explicação de como podemos construir o diagrama utilizando o Excel. Disponível em: <https://youtu.be/k1N7skhL01M>. Acesso em: 8 jul. 2015.

Sem medo de errar

Para construirmos a tabela de frequência, precisamos organizar as idades de 5 em 5 anos e contar quantas idades estão nessa faixa etária.

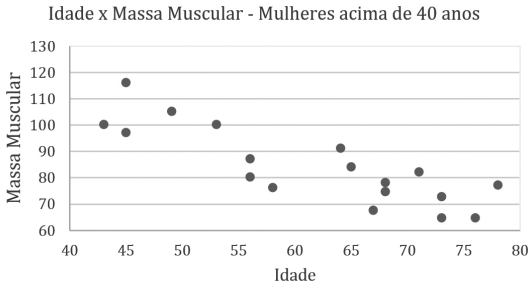
Tabela 2.9 | Pesquisa com Mulheres

Idades das Mulheres	Frequência f_i
40 – 45	3
46 – 50	1
51 – 55	1
56 – 60	3
61 – 65	2
66 – 70	3
71 – 75	3
76 – 80	2

Fonte: O autor (2015).

Para o Diagrama de dispersão, utilizamos as idades das mulheres no eixo X e as massas musculares no eixo Y.

Gráfico 2.1 | Diagrama de Dispersão Idade x Massa Muscular – mulheres acima de 40 anos



Fonte: O autor (2015).

Concluímos que ao observar o gráfico de dispersão entre as variáveis massa muscular e idade, vemos que há um forte indício de relação linear decrescente entre as variáveis em estudo. Nota-se que a massa muscular das pessoas diminui à medida que a idade aumenta. As mulheres na faixa dos 40 anos apresentaram maior massa que as mulheres de 80 anos.

Lembre-se

A **distribuição de frequências** visa representar um grande conjunto de informações, sem perder as suas principais características.

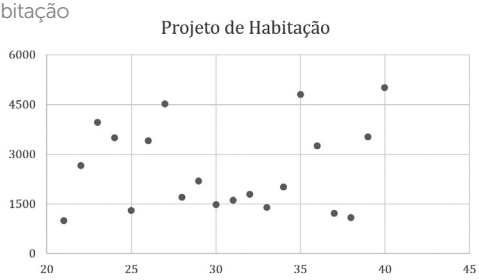
Os **diagramas de dispersão** são gráficos que permitem a identificação entre causas e efeitos, para avaliar o relacionamento entre variáveis.

Avançando na prática

Pratique mais!

Instrução
Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

<table>
<tr><th colspan="2">Programa de Habitação</th></tr>
<tr><td>1. Competência de fundamento de área</td><td>Conhecer os fundamentos estatísticos básicos necessários à formação do profissional.</td></tr>
<tr><td>2. Objetivos de aprendizagem</td><td>Utilizar as tabelas de frequência e os diagramas de dispersão para melhor interpretação dos dados estatísticos.</td></tr>
<tr><td>3. Conteúdos relacionados</td><td>Tabelas de Frequência e Diagramas de Dispersão.</td></tr>
<tr><td>4. Descrição da SP</td><td>A tabela mostra valores dos salários de vinte famílias que foram beneficiadas pelo Programa de Habitação Minha Casa Minha Vida. A partir dos dados apresentados, o governo precisa saber quantas famílias pertencem a cada faixa salarial, e para isso, você deve construir uma tabela de frequência com as faixas salariais: de 0 a 1500,00, de 1501,00 a 3000,00, de 3001,00 a de 4501,00 a 6000,00.
Tabela 2.10 | Idade do Comprador x Renda em R$
<table><tr><th>Idade do Comprador</th><th>Renda R$</th><th>Idade do Comprador</th><th>Renda R$</th></tr><tr><td>21</td><td>1000</td><td>29</td><td>2200</td></tr><tr><td>38</td><td>1100</td><td>22</td><td>2650</td></tr><tr><td>37</td><td>1200</td><td>26</td><td>3245</td></tr><tr><td>25</td><td>1300</td><td>36</td><td>3420</td></tr><tr><td>33</td><td>1400</td><td>24</td><td>3500</td></tr><tr><td>30</td><td>1500</td><td>39</td><td>3540</td></tr><tr><td>31</td><td>1600</td><td>23</td><td>3950</td></tr><tr><td>28</td><td>1700</td><td>27</td><td>4521</td></tr><tr><td>32</td><td>1800</td><td>35</td><td>4800</td></tr><tr><td>34</td><td>2000</td><td>40</td><td>5000</td></tr></table>Fonte: O autor (2015).
Construa um diagrama de dispersão e interprete-o.</td></tr>
<tr><td>5. Resolução da SP</td><td>Faça a tabela de frequência e separe os dados em faixas salariais de 0 a 1600,00, de 1601,00 a 3250,00 e de 3251,00 a 5000,00.
Tabela 2.11 | Faixa Salarial
<table><tr><th>Faixa Salarial</th><th>Frequência f_i</th></tr><tr><td>0 - 1500</td><td>6</td></tr><tr><td>1501 - 3000</td><td>6</td></tr><tr><td>3001 - 4500</td><td>5</td></tr><tr><td>4501 - 6000</td><td>3</td></tr></table>Fonte: O autor (2015).</td></tr>
</table>

5. Resolução da SP	O diagrama de dispersão pode ser dado por: Gráfico 2.2 | Diagrama de Dispersão – Projeto de Habitação  Fonte: O autor (2015). Concluímos que o diagrama de dispersão apresenta informações sobre análises bidimensionais. Temos dois dados que se relacionam entre si, a idade e a renda do comprador. A tabela de frequência mostra quantas vezes o dado se enquadra na classe estabelecida e a mesma informação pode ser observada no diagrama de dispersão.

Faça você mesmo

Com os mesmos dados apresentados anteriormente, você deve fazer a análise pela idade.

1. Faça a tabela de frequência da faixa etária, de 5 em 5 anos.

2. Apresente o diagrama de dispersão apenas para os dados dos compradores com menos de 30 anos.

3. Interprete a tabela e o diagrama.

Lembre-se

Frequência absoluta: Número de eventos observados de um tipo.

Frequência relativa: Dada em porcentagem (ou como fração). Se foram observados xi do tipo i, dentre n dados, a frequência relativa percentual será: $(\frac{X_i}{n})\times 100\%$

Frequência Cumulativa: Mede frequência absoluta ou relativa até um certo ponto e não apenas em um valor.

Faça valer a pena

Tabela 2.12 | Peso dos bebês nascidos no ano de 2008

Peso (gramas)	Contagem
Menos de 500	10.547
500 a 999	53.001
1000 a 1499	31.900
1500 a 1999	67.140
2000 a 2499	218.296
2500 a 2999	301.458
3000 a 3499	100.254
3500 a 3999	580.145
4000 a 4499	280.270
4500 a 4999	39.109

Fonte: O autor (2015).

No ano de 2008, foram levantados o peso e a contagem de bebês nascidos, nos Estados Unidos. Os dados foram apresentados na tabela anterior.

Utilize essas informações para responder às questões de 01 a 03.

1. Os dados da contagem correspondem a qual tipo de frequência?

a) Frequência Absoluta.
b) Frequência Relativa.
c) Frequência Cumulativa Relativa.
d) Frequência Cumulativa.
e) Frequência Assimétrica.

2. A frequência relativa para os bebês com peso de 3500 a 3999 gramas é aproximadamente:

a) 10%.
b) 25%.
c) 35%.
d) 50%.
e) 75%.

3. A frequência relativa referente aos bebês com peso de 2000 a 2499 gramas é aproximadamente:
a) 10%.
b) 20%.
c) 30%.
d) 40%.
e) 50%.

4. O diagrama de dispersão visa:
a) Identificar se existe uma tendência de variação conjunta entre duas ou mais variáveis.
b) Mostrar os dados para uma análise qualitativa.
c) Coletar dados sem tempo determinado, entre as variáveis que se deseja estudar as relações.
d) Verificar se as duas variáveis estão relacionadas, e se não há relação de causa e efeito.
e) Manter os padrões de dados para uma variável apenas.

5. Sobre o Diagrama de dispersão, pode-se afirmar que:
I. Diagrama de dispersão é uma ferramenta que indica a existência, ou não, de relações entre variáveis de um processo e sua intensidade, representando duas ou mais variáveis, uma em função da outra.
II. Diagrama de dispersão deve ser usado quando se necessita visualizar o que acontece com uma variável quando outra se altera, podendo identificar uma possível relação de causa e efeito entre elas.
III. Diagrama de dispersão é usado para representar simultaneamente os valores de duas variáveis quantitativas medidas em cada elemento do conjunto de dados.
Qual das alternativas indica as afirmativas corretas?

a) I e II.
b) I, II e III.
c) I e III.
d) Apenas a I.
e) II e III.

Utilize os dados a seguir para os exercícios 6 e 7. Os valores do metabolismo basal de 40 alunos foram tabulados. Os dados foram medidos em calorias por dia.

Tabela 2.13 | Pesquisa Idade x Metabolismo basal de 40 alunos

Idade	Metabolismo	Idade	Metabolismo	Idade	Metabolismo	Idade	Metabolismo	Idade	Metabolismo
12	910	16	950	16	1070	11	1000	18	1100
15	1090	14	1570	18	1670	10	1155	13	1290
17	1090	12	1250	15	1450	18	1478	17	1150
15	1547	15	1350	12	1680	16	1520	16	1230
15	990	14	1280	18	1130	13	1890	12	910
13	1380	15	1695	13	1220	12	1200	14	1960
13	1175	11	1348	18	1130	12	1370	15	2000
11	1210	11	1780	15	1950	18	1530	16	2100

Fonte: O autor (2015).

6. Faça a tabela de frequência utilizando os dados apresentados. As classes de frequência devem ser separadas de 300 em 300 calorias, começando com 900 calorias.

7. Faça um diagrama de dispersão metabolismo (x) e idade (y). Analise e estabeleça uma conclusão.

Seção 2.3

Coeficiente de correlação linear e o uso e aplicabilidade do coeficiente de correlação

Assimile

Correlação significa relação mútua entre dois termos, qualidade de correlativo, correspondência. Correlacionar significa estabelecer relação ou correlação entre; ter correlação.

Diálogo aberto

Necessitamos agora estudar o relacionamento entre duas ou mais variáveis, pois já sabemos calcular suas medidas individuais. Agora queremos verificar como uma variável influencia na relação com a outra.

Estudaremos dois tipos de associação entre duas variáveis. A primeira chamamos de experimental, em que as medidas são observadas pela imputação de valores ao acaso. A segunda chamamos de correlacional, em que não temos nenhum controle sobre as variáveis. Elas são analisadas naturalmente, sem ter interferência, e as duas variáveis são consideradas aleatórias. Quando os valores são ao acaso, não são tendenciosos e são definidos pela natureza.

Os objetivos de aprendizagem desta seção são entender o cálculo da correlação linear e estabelecer relações que possibilitem predizer uma ou mais variáveis em termos de outras.

Assim é que se fazem estudos para predizer as vendas futuras de um produto em função do seu preço, ou a perda de peso de uma pessoa em decorrência do número de semanas que se submete a uma dieta de 800 calorias por dia, ou a despesa de uma família com médico e remédios em função de sua renda, ou o consumo per capita de certos alimentos em função de seu valor nutritivo e do gasto com propaganda na TV, etc.

Naturalmente, o ideal seria que pudéssemos predizer uma quantidade exata em termos de outra, mas isso raramente é possível. Na maioria dos casos, devemos nos contentar com a predição de

médias, ou valores esperados. Por exemplo, não podemos predizer exatamente quanto ganhará um bacharel nos 10 anos subsequentes à sua formatura, mas, com base em dados adequados, é possível predizermos o ganho médio de todos os bacharéis nos 10 anos após a formatura. Analogamente, podemos predizer a safra média de certa variedade de trigo em termos do índice pluviométrico de julho, e a nota média de um calouro do curso de Direito em função do seu QI.

Assim, quando consideramos variáveis como peso e altura de um grupo de pessoas, ou uso de cigarro e incidência de câncer, procuramos verificar se existe alguma relação entre as variáveis de cada um dos pares e qual seria o grau dessa relação. Para isso, é necessário o conhecimento de novas medidas.

Os dados levantados pela pesquisa do educador físico foram dispostos em idades das clientes e também a massa muscular. Necessita-se estabelecer a correlação linear entre a idade e a massa muscular para colocar no relatório do educador físico. Isso nos permitirá estabelecer a relação de como a idade influencia na massa muscular das clientes da amostra estudada.

Para isso, utilize a tabela com os dados de idade (x) e massa muscular (y).

Tabela 2.14 | Dados Pesquisados – Idade x Massa Muscular

Idade (X)	Massa muscular (Y)
43	100
45	116
45	97
49	105
53	100
56	87
56	80
58	76
64	91
65	84
67	68
68	75
68	78

71	82
73	73
73	65
76	65
78	77

Fonte: O autor (2015).

Não pode faltar

Coeficiente de Correlação Linear

Apesar do diagrama de dispersão nos fornecer uma ideia do tipo e extensão do relacionamento entre duas variáveis X e Y, seria altamente desejável ter um número que medisse essa relação. Essa medida existe e é denominada de coeficiente de correlação. Quando se está trabalhando com amostras, o coeficiente de correlação é indicado pela letra r.

Tem-se uma variável estatística bidimensional quando, relativamente a cada elemento da população, se observa e estuda duas características distintas.

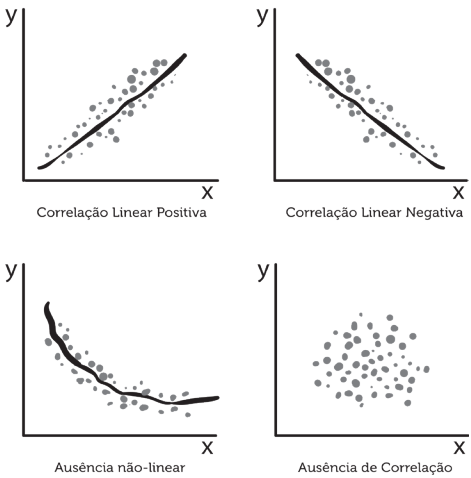
Para as variáveis estatísticas X e Y, a variável estatística bidimensional é representada por (X, Y).

Coeficiente de Correlação de Pearson:

A intensidade da associação linear existente entre as variáveis pode ser quantificada através do chamado coeficiente de correlação linear de Pearson:

$$r = \frac{\sum\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\left(\sum\left(x_i - \overline{x}\right)^2\left(\sum\left(y_i - \overline{y}\right)^2\right)\right)}} = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}}$$

Figura 2.2 | Gráficos de Correlação



Fonte: Adaptado de: Larson (2010)

Variáveis positivamente correlacionadas: No limite, isto é, se a correlação for "perfeita" - como é o caso se considerarmos a correlação da variável x consigo própria - o coeficiente de correlação será igual a 1. As variáveis estão negativamente correlacionadas: No limite, isto é, se a correlação for "perfeita", o coeficiente de correlação será igual a -1.

As variáveis não estão correlacionadas: No limite, isto é, em caso de "absoluta independência", o coeficiente de correlação será igual a 0.

Observação 1: não verificar correlação linear, não significa que não se verifique outro tipo de correlação, por exemplo, **exponencial**.

Observação 2: qualquer que seja a correlação verificada, correlação não significa causalidade.

As propriedades mais importantes do coeficiente de correlação são: o intervalo de variação da correlação se dá entre -1 a +1. É uma medida adimensional. O grau linear positivo da correlação entre X e Y se dá quando r é mais próximo de +1. O grau linear negativo da correlação entre X e Y se dá quando r é mais próximo de -1 (LARSON, 2010).

Complemente seus estudos

Para saber um pouco mais sobre o coeficiente de correlação de Pearson, você pode ler o artigo disponível em: <http://www.revista.ufpe.br/politicahoje/index.php/politica/article/viewFile/6/6>. Acesso em: 8 jul. 2015.

Assimile

Uso e aplicabilidade do coeficiente de Correlação

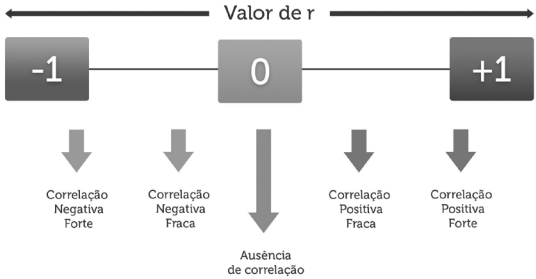
O principal objetivo da análise da correlação linear é medir a intensidade de uma relação linear entre duas variáveis.

A Correlação não é o mesmo que causa e efeito. Duas variáveis podem estar altamente correlacionadas e, no entanto, não haver relação de causa e efeito entre elas.

- Se duas variáveis estiverem amarradas por uma relação de causa e efeito elas estarão, obrigatoriamente, correlacionadas.
- O estudo de correlação pressupõe que as variáveis X e Y tenham uma distribuição normal.
- A palavra simples que compõe o nome correlação linear simples indica que estão envolvidas no cálculo somente duas variáveis.
- O coeficiente de correlação linear de Pearson mede a correlação em estatística paramétrica.

Análise do Diagrama de Dispersão para a correlação

Figura 2.3 | Diagrama de Dispersão para a correlação



Fonte: O autor (2015).

O diagrama de dispersão mostrará que a correlação será tanto mais forte quanto mais próximo estiver o coeficiente de –1 ou +1, e será tanto mais fraca quanto mais próximo o coeficiente estiver de zero.

Tabela 2.15 | Coeficientes de Correlação

Coeficiente de correlação	Correlação
r=1	Perfeita positiva
0,8 ≤ r < 1	Forte positiva
0,5 ≤ r < 0,8	Moderada positiva
0,1 ≤ r < 0,5	Fraca positiva
0 ≤ r < 0,1	Ínfima positiva
0	Nula
-0,1 < r < 0	Ínfima negativa
-0,5 < r ≤ -0,1	Fraca negativa
-0,8 < r ≤ -0,5	Moderada negativa
-1 < r ≤ -0,8	Forte negativa
r=-1	Perfeita negativa

Fonte: Larson (2010).

a) Correlação perfeita negativa (rxy = -1): quando os pontos estiverem perfeitamente alinhados, mas em sentido contrário, a correlação é denominada perfeita negativa.

b) Correlação negativa (-1 < rxy < 0): a correlação é considerada negativa quando valores crescentes da variável X estiverem associados a valores decrescentes da variável Y, ou valores decrescentes de X associados a valores crescentes de Y.

c) Correlação nula (rxy = 0): quando não houver relação entre as variáveis X e Y, ou seja, quando os valores de X e Y ocorrerem independentemente, não existe correlação entre elas.

d) Correlação positiva (0 < rxy < 1): será considerada positiva se os valores crescentes de X estiverem associados a valores crescentes de Y.

e) Correlação perfeita positiva (rxy = 1): a correlação linear perfeita positiva corresponde ao caso anterior, só que os pontos (X, Y) estão perfeitamente alinhados.

f) Correlação **espúria**: quando duas variáveis X e Y forem independentes, o coeficiente de correlação será nulo. Entretanto, algumas vezes, isso não ocorre, podendo, assim mesmo, o coeficiente apresentar um valor próximo de –1 ou +1. Nesse caso, a correlação é espúria. Todas as correlações são mostradas na tabela.

A correlação indica o comportamento conjunto de duas variáveis. Algumas aplicabilidades da correlação linear:

- O salário de um trabalhador está relacionado com a escolaridade, sendo em que grau a variável "salário médio do trabalhador" está ligada com a variável "escolaridade do trabalhador"?
- A quantidade de livros que uma pessoa já leu está relacionada com a sua escolaridade?
- Em que grau o peso de uma pessoa está relacionado com a sua altura?
- A estatura de uma pessoa está relacionada com a sua alimentação?

Vocabulário

Correlação – Relação de interdependência entre duas ou entre múltiplas variáveis.

Exponencial – Diz-se de uma quantidade ou variável que se apresenta em expoente, do cálculo relativo a essas quantidades, das equações em que elas existem e das curvas que as representam.

Espúria – Que não é certo, verdadeiro ou real; hipotético.

Exemplificando

Considere uma amostra aleatória, formada por 5 de 50 pacientes de um endocrinologista. Vamos verificar a correlação entre o consumo de açúcares por dia e o consumo de sal por dia. A tabela dispõe os valores para cada paciente.

Tabela 2.16 | Pacientes x Consumo de Açúcares e Sal

Números do Paciente	Consumo de Açucares (xi)	Consumo de Sal (yi)	xi . yi	xi2	yi2
1	5	6	30	25	36
8	8	9	72	64	81
24	7	8	56	49	64
38	10	10	100	100	100
44	6	5	30	36	25
Total	36	38	288	274	306

Fonte: O autor (2015).

Para calcular o coeficiente de correlação, temos:

$$r = \frac{n\sum x_i \cdot y_i - (\sum x_i)(\sum y_i)}{\sqrt{(n\sum x_i^2 - (\sum x_i^2)) \cdot (n\sum y_i^2 - (\sum y_i^2))}} = \frac{(5 \cdot 288) - (36 \cdot 38)}{\sqrt{(5 \cdot 274) - (36^2) \cdot (5 \cdot 306) - (38^2)}} = \frac{72}{79{,}77} = 0{,}902$$

O resultado indica uma correlação linear positiva altamente significativa entre as duas variáveis, consumo de açúcares e consumo de sal.

Faça você mesmo

Classifique os coeficientes de correlação segundo o diagrama a seguir:

Figura 2.4 | Diagrama de Dispersão da Correlação

Fonte: O autor (2015).

Faça você mesmo

a) -0,336

b) -0,985

c) 0,897

d) 0,495

e) 0

Atenção

1. O intervalo de variação vai de -1 a +1.

2. O coeficiente de correlação é uma medida adimensional, isto é, ele é independente das unidades de medida das variáveis X e Y.

3. Quanto mais próximo de +1 for "r", maior o grau de relacionamento linear positivo entre X e Y, ou seja, se X varia em uma direção, Y variará na mesma direção.

4. Quanto mais próximo de -1 for "r", maior o grau de relacionamento linear negativo entre X e Y, isto é, se X varia em um sentido, Y variará no sentido inverso.

5. Quanto mais próximo de zero estiver "r", menor será o relacionamento linear entre X e Y. Um valor igual a zero indicará ausência apenas de relacionamento linear.

Sem medo de errar

Calculando o coeficiente de correlação linear entre X e Y, denotamos as variáveis: Y = Massa Muscular e X = Idade (n=18).

Tabela 2.17 | Dados Pesquisados

Clientes	Idade (X)	Massa muscular (Y)	xi . yi	xi^2	yi^2
1	43	100	4300	1849	10000
2	45	116	5220	2025	13456
3	45	97	4365	2025	9409
4	49	105	5145	2401	11025
5	53	100	5300	2809	10000
6	56	87	4872	3136	7569
7	56	80	4480	3136	6400
8	58	76	4408	3364	5776
9	64	91	5824	4096	8281
10	65	84	5460	4225	7056
11	67	68	4556	4489	4624
12	68	75	5100	4624	5625
13	68	78	5304	4624	6084
14	71	82	5822	5041	6724
15	73	73	5329	5329	5329
16	73	65	4745	5329	4225
17	76	65	4940	5776	4225
18	78	77	6006	6084	5929
Total	1108	1519	91176	70362	131737

Fonte: O autor (2015).

$$r = \frac{n\sum x_i \cdot y_i - (\sum x_i)(\sum y_i)}{\sqrt{(n\sum x_i^2 - (\sum x_i^2)) \cdot (n\sum y_i^2 - (\sum y_i^2))}}$$

$$r = \frac{(18 \cdot 91176) - (1108 \cdot 1519)}{\sqrt{(18 \cdot 70362) - (1108^2) \cdot (18 \cdot 131737) - (1519^2)}}$$

$$r = \frac{-41884}{49828,1}$$

$$r = -0,84$$

Segundo o resultado da correlação obtida, pode-se notar que há uma forte correlação linear entre as variáveis massa muscular e idade. Nota-se que à medida que a idade da pessoa aumenta, a massa muscular diminui, o que é coerente com o gráfico de dispersão apresentado anteriormente.

Lembre-se

Correlação perfeita negativa → $r_{xy} = -1$.

Correlação negativa → $-1 < r_{xy} < 0$.

Correlação nula → $r_{xy} = 0$.

Correlação positiva → $0 < r_{xy} < 1$.

Correlação perfeita positiva → $r_{xy} = 1$.

Avançando na prática

Pratique mais!	
Instrução Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.	
Experimento no Laboratório de Biologia	
1. Competência de fundamento de área	Conhecer os fundamentos estatísticos básicos necessários à formação do profissional.
2. Objetivos de aprendizagem	Entender o cálculo da correlação linear e estabelecer relações que possibilitem predizer uma ou mais variáveis em termos de outras.
3. Conteúdos relacionados	Coeficiente de Correlação Linear
4. Descrição da SP	Durante 5 horas, foi medido o crescimento de uma bactéria em um laboratório de Biologia. A tabela abaixo mostra os valores das horas (x) e de crescimento (y). xi: 0 1 2 3 4 5 yi: 0 3 6 9 12 15 É preciso saber o coeficiente de correlação entre as horas observadas e o crescimento. Classifique a correlação e interprete o valor encontrado.
5. Resolução da SP	Para calcular o coeficiente de correlação, precisamos montar a tabela com os procedimentos:

Tabela 2.18 | Pacientes x Consumo de Açúcares e Sal

x_i	y_i	x_i^2	y_i^2	$x_i . y_i$
0	0	0	0	0
1	3	1	9	3
2	6	4	36	12
3	9	9	81	27
4	12	16	144	48
5	15	25	225	75
$\sum x_i = 15$	$\sum y_i = 45$	$\sum x_i^2 = 55$	$\sum y_i^2 = 495$	$\sum x_i \cdot y_i = 165$

Fonte: O autor (2015).

Vamos utilizar a fórmula do coeficiente de correlação de Pearson:

O valor de r é igual a 1, significando que as variáveis estão perfeitamente relacionadas e que a distribuição segue exatamente uma reta se fizemos o diagrama de dispersão.

$$r = \frac{n\sum x_i \cdot y_i - (\sum x_i)(\sum y_i)}{\sqrt{(n\sum x_i^2 - (\sum x_i^2)) \cdot (n\sum y_i^2 - (\sum y_i^2))}}$$

$$r = \frac{(6 \cdot 165) - (15 \cdot 45)}{\sqrt{\left((6 \cdot 55) - (15)^2\right) \cdot \left((6 \cdot 495) - (45^2)\right)}} \quad r = \frac{315}{\sqrt{105 \cdot 945}} = \frac{315}{315} = 1$$

Faça valer a pena!

1. Em uma clínica para mulheres, o endocrinologista fez uma pesquisa com 50 mulheres e analisou uma amostra de 5 mulheres com 50 anos de idade. As perguntas realizadas foram em relação ao nível de HDL – Colesterol bom e quantas horas semanais elas praticam exercícios físicos.

HDL (mg/dL)	40	50	55	60	65
Horas de exercícios físicos	0	2	3	4	6

É importante entender que em pessoas com índices de HDL acima de 50 mg/dL, as doenças cardiovasculares ocorrem com menor frequência. Qual é o coeficiente de correlação de Pearson?

a) 0,988
b) 0,855
c) 0,765
d) -0,534
e) -0,987

2. Como se classifica a correlação encontrada no exercício 1?

a) Correlação Negativa Forte.
b) Correlação Negativa Fraca.
c) Correlação Nula.
d) Correlação Positiva Forte.
e) Correlação Positiva Fraca.

3. Uma fábrica de automóveis apresentou a amostra:

Custo total de automóveis (milhões) Y	80	44	51	70	61
Produção X (mil unidades)	12	4	6	11	8

Que tipo de correlação se verifica entre o custo total e a produção apresentada pela fábrica de automóveis?

a) Correlação Negativa Forte.
b) Correlação Negativa Fraca.
c) Correlação Positiva Forte.
d) Correlação Positiva Fraca.
e) Correlação Nula.

4. Uma pesquisa sobre a escolaridade dos professores e a quantidade de livros que eles leram em um ano apresentou um coeficiente de correlação linear igual a -0,687. Qual é a conclusão que se pode tirar sobre essa pesquisa?

a) A pesquisa não apresenta relação entre as variáveis.
b) As unidades de medida das variáveis X e Y não são relacionáveis.
c) A pesquisa apresenta maior grau de relacionamento linear positivo entre X e Y, pois as quantidades de livros estão relacionadas à escolaridade dos professores
d) A pesquisa apresenta maior o grau de relacionamento linear negativo entre X e Y, pois as quantidades de livros lidos estão relacionadas à escolaridade dos professores.
e) Há uma correlação que indicará ausência de relacionamento linear.

5. Uma barra de ferro apresentou algumas medidas ao ser submetida ao aquecimento. A tabela a seguir mostra as temperaturas e as medidas:

Temperatura (°C)	10	15	20	25	30
Comprimento (mm)	1003	1005	1010	1011	1014

Qual é o coeficiente de correlação linear entre a temperatura (x) e o comprimento da barra (y)?

a) 0,841.
b) 0,801.
c) 0,777.
d) 0,983.
e) -0,987.

A polícia rodoviária costuma fazer bloqueios nas estradas para avaliar a condição dos motoristas, principalmente em feriados prolongados. A tabela a seguir mostra dados de uma avaliação feita pelos policiais rodoviários na Rodovia dos Bandeirantes nos feriados prolongados dos cinco primeiros meses de 2014. A quantidade de acidentes e a quantidade de motoristas alcoolizados são mostradas na tabela. Utilize os dados para os exercícios 6 e 7.

Tabela 2.19 | Acidentes de trânsito em 2014

Número de motoristas alcoolizados	Número de acidentes
100	35
254	90
140	33
115	45
98	29
707	232

Fonte: O autor

6. Determine o coeficiente de correlação.

7. Classifique a correlação e faça a interpretação dos resultados.

Seção 2.4

Coeficiente de determinação e regressão linear simples – método dos mínimos quadrados

Diálogo aberto

Na seção anterior, vimos que o principal objetivo da análise da correlação linear é medir a intensidade de uma relação linear entre duas variáveis. Nesta seção, veremos que a análise de regressão estuda o relacionamento entre uma variável chamada a variável dependente e outras variáveis chamadas variáveis independentes. Esse relacionamento é representado por um modelo matemático, isto é, por uma equação que associa a variável dependente com as variáveis independentes. Esse modelo é designado por modelo de regressão linear simples, em que define-se uma relação linear entre a variável dependente e uma variável independente.

Da mesma forma, como usamos a média para resumir uma variável aleatória, a reta de regressão é usada para resumir a estimativa linear entre duas variáveis aleatórias (LAPPONI, 1997).

Vamos estudar esse modelo nesta seção e nosso objetivo de aprendizagem é utilizar o coeficiente de correlação linear, o coeficiente de determinação e a regressão linear para organizarmos os dados coletados.

Para o relatório do estudo do educador físico sobre a diminuição da massa muscular com o envelhecimento, os dados coletados são referentes a 18 mulheres. Será necessário para o relatório mostrar a reta de regressão linear simples entre a variável dependente (y) - no nosso caso, a massa muscular - e a variável independente (x) - a idade das mulheres.

Você deve determinar o coeficiente de determinação, utilizando o coeficiente de correlação que foi calculado na seção anterior. E, com a reta de regressão estimada da variável massa muscular (y) em função da Idade (x), estime a massa muscular média de mulheres com 50 anos.

Não pode faltar

Coeficiente de Determinação

Como vimos na seção anterior, aprendemos a calcular o coeficiente de correlação r. O quadrado desse coeficiente é chamado coeficiente de determinação.

Assimile

O coeficiente de determinação indica a proporção de variação da variável independente que é explicada pela variável dependente, ou seja, é uma ferramenta que avalia a qualidade do ajuste. Também pode ser explicado pela relação da variação explicada pela variação total.

Quanto mais próximo da unidade o r^2 estiver, melhor é a qualidade do ajuste. O seu valor fornece a proporção da variável Y explicada pela variável X através da função ajustada.

$$r2 = \frac{\text{variação explicada}}{\text{variação total}}$$

É importante que você sabia interpretar o coeficiente de determinação corretamente, por exemplo, se o coeficiente de correlação é r = 0,9929, então o coeficiente de determinação será:

$r^2 = (0{,}9929)^2 = 0{,}9858 = 98{,}50\ \%$.

Isso significa que 98,50% da variação de y pode ser explicada pela relação entre x e y. O restante (1,5% da variação) não é explicada e é em razão de outros fatores ou a erro da amostra.

Associação não é causalidade

Suponha que encontremos uma associação ou correlação entre duas variáveis A e B. Podem existir diversas explicações do porquê elas variam conjuntamente, incluindo:

- Mudanças em A causam mudanças em B;
- Mudanças em B causam mudanças em A;
- Mudanças em outras variáveis causam mudanças tanto em A quanto em B.

A relação observada é somente uma coincidência.

A terceira explicação é frequentemente a mais apropriada. Isso indica que existe algum processo de conexão atuando, por exemplo, o número de pessoas usando óculos de sol e a quantidade de sorvete consumido num particular dia são altamente correlacionados. Isso não significa que usar óculos de sol causa a compra de sorvetes ou vice-versa.

É extremamente difícil estabelecer relações causais a partir de dados observacionais. Precisamos realizar experimentos para obter mais evidências de uma relação causal.

Complemente seus estudos

O *link* a seguir mostra mais alguns aspectos sobre a associação e causalidade. Acesse e estude um pouco mais sobre o tema. Disponível em: <http://www.galileu.esalq.usp.br/mostra_topico.php?cod=130>. Acesso em: 8 jul. 2015.

Regressão Linear

O objetivo da regressão linear é fazer a análise estatística, verificando a relação funcional de uma variável dependente com uma ou mais variáveis independentes. A regressão propõe uma equação que tenta explicar a variação da variável dependente pelas variáveis independentes.

A equação representa o fenômeno que está sendo estudado, podemos fazer um gráfico que já estudamos, que é o diagrama de dispersão, o qual verifica como os valores da variável dependente (Y) se comportam em relação à variável independente (X).

Os pontos do diagrama de dispersão ficam distanciados da curva do modelo matemático que podemos escolher. Para isso, podemos usar uma relação funcional para obtermos a equação estimada, de modo que as distâncias entre os pontos do diagrama e os pontos da curva do modelo escolhido sejam as menores possíveis.

Esse método descrito é chamado de Método dos Mínimos Quadrados (MMQ).

O Método dos Mínimos Quadrados faz a soma dos quadrados das distâncias entre os pontos do diagrama e os pontos da curva da equação estimada e os minimiza. Assim, uma relação funcional de X e Y ocorre para o modelo escolhido, mas com o mínimo de erro possível.

Reflita

O objetivo principal da análise de regressão é predizer o valor da variável dependente Y, dado que seja conhecido o valor da variável independente X.

O Método dos Mínimos Quadrados

O ajuste de curvas pelo método dos mínimos quadrados é relevante, pois, ao contrário do método gráfico, é um método que é independente da avaliação de quem está realizando o experimento. Esse método consiste em minimizar o erro quadrático médio, chamado de S. Para isso, utilizamos um conjunto de N medidas (xi e yi), dizendo que i são valores inteiros desde 1 a N. Assim, podemos calcular S da seguinte maneira:

$$S = \sum_{i=1}^{N} \Delta S_i = \sum_{i=1}^{N} (y - y_i)^2$$

Estabelecemos que y é o valor da curva ajustada calculada por (y=a·x+b).

Precisamos somar os valores de ΔSi para todas as N medidas e traçar uma reta, tornando a soma de ΔSi mínima.

A derivada de ΔS em relação a a é zero. E a derivada de ΔS em relação a b também é zero. Isso acontece razoavelmente para uma reta desejável que passa por todos os pontos experimentais.

O coeficiente linear da reta (b) e o coeficiente angular da reta (a) são dados por:

$$a = \frac{N \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \qquad b = \frac{\sum_{i=1}^{N} {x_i}^2 \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i y_i \sum_{i=1}^{N} x_i}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}$$

Sendo:

$$\sum_{i=1}^{N} x_i = x_i + x_2 + \cdots + x_4.$$

Assim, temos: y=ax+b

X – tempo – s	Y – posição – m
0,100	0,51
0,200	0,59
0,300	0,72
0,400	0,80
0,500	0,92

Em um autódromo de Kart, foram medidos os tempos e as posições dos carrinhos. Pelo método dos mínimos quadrados, determine a reta de regressão para as medidas. Defina o coeficiente de determinação (r2). Assuma como variável dependente (Y) os valores da posição e como variável independente (X) o tempo.

Resolução

Para o método dos mínimos quadrados, construa a tabela com os valores de xi, yi, xi.yi e x2 e as respectivas somatórias. N = 5.

x_i	y_i	$x_i.y_i$	x^2	y^2
0,100	0,51	0,051	0,010	0,26
0,200	0,59	0,118	0,040	0,35
0,300	0,72	0,216	0,090	0,52
0,400	0,80	0,32	0,160	0,64
0,500	0,92	0,46	0,250	0,85
$\sum x_i = 1{,}50$	$\sum y_i = 3{,}54$	$\sum x_i.y_i = 1{,}17$	$\sum x^2 = 0{,}55$	$\sum y^2 = 2{,}61$

Coeficiente de Determinação:

$$r = \frac{n\sum x_i \cdot y_i - (\sum x_i)(\sum y_i)}{\sqrt{(n\sum x_i^2 - (\sum x_i^2)) \cdot (n\sum y_i^2 - (\sum y_i^2))}} = 0{,}997 \qquad r^2 = 0{,}994$$

Calcularmos os índices a e b:

$$a = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2} = \frac{(5 \times 1{,}17) - (1{,}50 \times 3{,}54)}{(5 \times 0{,}55) - (1{,}50)^2} = \frac{5{,}85 - 5{,}31}{2{,}75 - 2{,}25} = \frac{0{,}54}{0{,}5} = 1{,}08$$

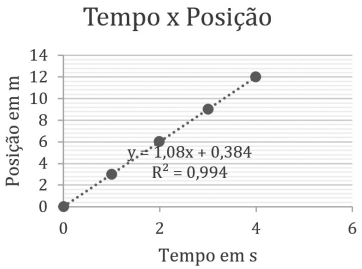
$$b = \frac{\sum_{i=1}^{N} x_i{}^2 \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i y_i \sum_{i=1}^{N} x_i}{N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2} = \frac{(0{,}55 \times 3{,}54) - (1{,}17 \times 1{,}5)}{(5 \times 0{,}55) - (1{,}50)^2} = \frac{0{,}192}{0{,}5} = 0{,}384$$

$$y = 1{,}08x + 0{,}384$$

Gerando a tabela, segundo a equação, para os valores da posição em função do tempo:

Exemplificando

x	y
0,100	0,49
0,200	0,60
0,300	0,71
0,400	0,82
0,500	0,92



Os dados experimentais são mostrados pelas esferas no gráfico. A reta de regressão linear mostra o método de mínimos quadrados para os dados apresentados.

Vocabulário

Causalidade - É o conjunto de todas as relações de causa e efeito.

Regressão linear - Uma equação que determina a relação entre as variáveis.

Atenção

O material disponível no *link* a seguir traz uma aula sobre regressão linear. Os exemplos apresentados estão bem detalhados e lhe ajudarão no estudo do tema. Disponível em: <http://www.ime.unicamp.br/~hlachos/RegresCorr.pdf>.

Ajuste uma reta de regressão para a relação entre as variáveis Y: massa muscular (dependente) e X: idade (independente). Determine o coeficiente de determinação e utilize o coeficiente de correlação que foi calculado na seção anterior. E, com a reta de regressão estimada da variável, Massa muscular (Y) em função da Idade (X), estime a massa muscular média de mulheres com 50 anos.

Tabela 2.20 | Dados Pesquisados

Clientes	Idade (X)	Massa muscular (Y)	xi . yi	xi^2	yi^2
1	43	100	4300	1849	10000
2	45	116	5220	2025	13456
3	45	97	4365	2025	9409
4	49	105	5145	2401	11025
5	53	100	5300	2809	10000
6	56	87	4872	3136	7569
7	56	80	4480	3136	6400
8	58	76	4408	3364	5776
9	64	91	5824	4096	8281
10	65	84	5460	4225	7056
11	67	68	4556	4489	4624
12	68	75	5100	4624	5625
13	68	78	5304	4624	6084
14	71	82	5822	5041	6724
15	73	73	5329	5329	5329
16	73	65	4745	5329	4225
17	76	65	4940	5776	4225
18	78	77	6006	6084	5929
Total	1108	1519	91176	70362	131737

Fonte: O autor (2015).

O coeficiente de correlação calculado na seção anterior foi:

$$r = \frac{n\sum x_i \cdot y_i - (\sum x_i)(\sum y_i)}{\sqrt{(n\sum x_i^2 - (\sum x_i^2)) \cdot (n\sum y_i^2 - (\sum y_i^2))}}$$

$$r = \frac{(18 \cdot 91176) - (1108 \cdot 1519)}{\sqrt{(18 \cdot 70362) - (1108^2) \cdot (18 \cdot 131737) - (1519^2)}}$$

$$r = \frac{-41884}{49828,1}$$

$$r = -0,84$$

O coeficiente de determinação é:

r2=(-0,84)2=0,71

Para calcular os índices a e b da reta de regressão, temos:

Calculam-se os índices a e b:

$$a = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{(18 \times 91176) - (1108 \times 1519)}{(18 \times 70362) - (1108)^2} = -1{,}08$$

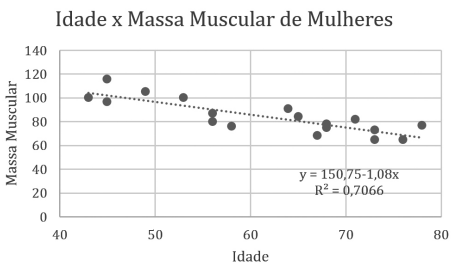
$$b = \frac{\sum_{i=1}^{N} x_i{}^2 \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i y_i \sum_{i=1}^{N} x_i}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{(70362 \times 1519) - (91176 \times 1108)}{(18 \times 70362) - (1108)^2} = 150{,}75$$

$$y = 150{,}75 - 1{,}08x$$

Com os valores determinados pela y pela equação anterior, teremos a seguinte tabela:

Tabela 2.21 | Dados ordenados

Idade (X)	Massa Muscular (Y)
43	104,31
45	102,15
45	102,15
49	97,83
53	93,51
56	90,27
56	90,27
58	88,11
64	81,63
65	80,55
67	78,39
68	77,31
68	77,31
71	74,07
73	71,91
73	71,91
76	68,67
78	66,51



Para as mulheres de 50 anos, teremos:

y= 150,75 - 1,08x

y= 150,75 - 1,08×50

y= 96,75

Fonte: O autor

A massa muscular estimada pela equação de regressão linear para mulheres de 50 anos é 96,75.

Assim, sendo o coeficiente de determinação r2 = 0,71, significa que se fizermos 1-0,71, encontramos que 0,29 ou 29% da variância da regressão não depende das variáveis estudadas.

Lembre-se

O **coeficiente de determinação** indica a proporção de variação da variável independente que é explicada pela variável dependente, ou seja, é uma ferramenta que avalia a qualidade do ajuste. Também pode ser explicado pela relação da variação total.

A regressão linear tem objetivo de fazer a análise estatística, verificando a relação funcional entre uma variável dependente com uma ou mais variáveis independentes. A regressão propõe uma equação que tenta explicar a variação da variável dependente pelas variáveis independentes.

Faça você mesmo

Em uma amostra aleatória, formada por 5 de 50 pacientes de um endocrinologista, vamos verificar a correlação entre consumo de açúcares por dia e o consumo de sal por dia. A tabela dispõe os valores para cada paciente.

Tabela 2.22 | Tabela de Frequência

Consumo de Açucares (xi)	Consumo de Sal (yi)
5	6
8	9
7	8
10	10
6	5
$\Sigma x_i = 36$	$\Sigma y_i = 38$

Fonte: O autor (2015).

Para os valores apresentados, determine o coeficiente de determinação e a equação de regressão linear pelo método dos mínimos quadrados e interprete os valores.

Avançando na prática

Pratique mais!

Instrução
Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

<table>
<tr><th colspan="2">Experimento de Biologia</th></tr>
<tr><td>1. Competência de fundamento de área</td><td>Conhecer os fundamentos estatísticos básicos necessários à formação do profissional.</td></tr>
<tr><td>2. Objetivos de aprendizagem</td><td>Entender o cálculo da correlação linear e estabelecer relações que possibilitem predizer uma ou mais variáveis em termos de outras.</td></tr>
<tr><td>3. Conteúdos relacionados</td><td>Coeficiente de Determinação e Regressão Linear.</td></tr>
<tr><td>4. Descrição da SP</td><td>Durante 5 horas, foi medido o crescimento de uma bactéria em um laboratório de Biologia. A tabela a seguir mostra os valores das horas (x) e de crescimento (y).
<table>
<tr><td>x_i</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td>y_i</td><td>0</td><td>3</td><td>6</td><td>9</td><td>12</td><td>15</td></tr>
</table>
Calcule o coeficiente de determinação, a equação de regressão linear e interprete os valores encontrados.</td></tr>
<tr><td>5. Resolução da SP</td><td>Para calcular o coeficiente de correlação, precisamos montar a tabela com os procedimentos:
Tabela 2.23 | Dados para o coeficiente de correlação
<table>
<tr><th>x_i</th><th>y_i</th><th>x_i^2</th><th>y_i^2</th><th>$x_i . y_i$</th></tr>
<tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>1</td><td>3</td><td>1</td><td>9</td><td>3</td></tr>
<tr><td>2</td><td>6</td><td>4</td><td>36</td><td>12</td></tr>
<tr><td>3</td><td>9</td><td>9</td><td>81</td><td>27</td></tr>
<tr><td>4</td><td>12</td><td>16</td><td>144</td><td>48</td></tr>
<tr><td>5</td><td>15</td><td>25</td><td>225</td><td>75</td></tr>
<tr><td>$\sum x_i = 15$</td><td>$\sum y_i = 45$</td><td>$\sum x_i^2 = 55$</td><td>$\sum y_i^2 = 495$</td><td>$\sum x_i \cdot y_i = 165$</td></tr>
</table>
Fonte: O autor
Vamos utilizar a fórmula do coeficiente de correlação de Pearson:
$$r = \frac{n\sum x_i \cdot y_i - (\sum x_i)(\sum y_i)}{\sqrt{(n\sum x_i^2 - (\sum x_i^2)) \cdot (n\sum y_i^2 - (\sum y_i^2))}}$$
$$r = \frac{(6 \cdot 165) - (15 \cdot 45)}{\sqrt{((6 \cdot 55) - (15)^2) \cdot ((6 \cdot 495) - (45^2))}}$$
$$r = \frac{315}{\sqrt{105 \cdot 945}} = \frac{315}{315} = 1$$
Sendo r = 1, o coeficiente de determinação (r2) também será 1.
Para a reta de regressão linear, calculamos os valores de índices a e b:</td></tr>
</table>

$$a = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{(5 \times 165) - (15 \times 45)}{(5 \times 55) - (15)^2} = \frac{150}{50} = 3$$

$$b = \frac{\sum_{i=1}^{N} x_i{}^2 \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i y_i \sum_{i=1}^{N} x_i}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{(55 \times 45) - (165 \times 15)}{(5 \times 55) - (15)^2} = 0$$

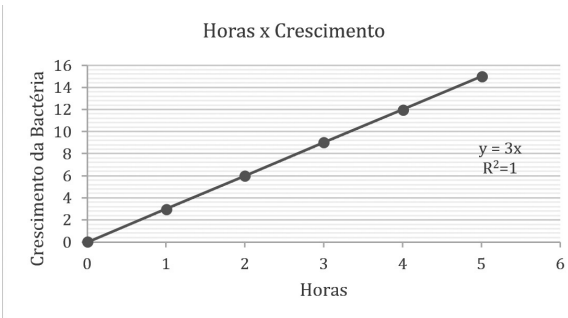
$$y = 3x$$

Para a equação, os valores de y são novamente calculados e a reta de regressão traçada:

Tabela 2.24 | Reta Regressão

x	y
0	0
1	3
2	6
3	9
4	12
5	15

Fonte: O autor.



Não há nenhuma variância da regressão entre as variáveis estudadas. As variáveis são perfeitamente relacionadas.

Faça valer a pena!

Em uma clínica para mulheres, o endocrinologista fez uma pesquisa com 50 pacientes e analisou uma amostra de 5 mulheres com 50 anos de idade. As perguntas realizadas foram em relação ao nível de HDL – Colesterol bom e quantas horas semanais elas praticam exercícios físicos. Utilize os seguintes dados para os exercícios 1 e 2.

HDL (mg/dL)	40	50	55	60	65
Horas de exercícios físicos	0	2	3	4	6

1. Qual é o valor do coeficiente de determinação do experimento?

a) 0,758.
b) 0,457.
c) 0,331.
d) 0,976.
e) 0,667.

2. Qual é a reta de regressão para o experimento?

a) y=3x-7,661.
b) y=2,41x-8,21.
c) y = 4,25x + 41,25.
d) y=4x-29,41.
e) y=9x-0,21.

Uma barra de ferro apresentou algumas medidas ao ser submetida ao aquecimento. A tabela a seguir mostra as temperaturas e as medidas. Utilize os seguintes dados para os exercícios 3 e 4.

Temperatura (°C)	10	15	20	25	30
Comprimento (mm)	1003	1005	1010	1011	1014

3. Qual é o valor do coeficiente de determinação do experimento?

a) 0,966.
b) 0,844.
c) 0,547.
d) 0,125.
e) 0,248.

4. Qual é a reta de regressão para o experimento?

a) y=0,33x+1000.
b) y=0,56x+997,4.
c) y=15x+590.
d) y=1000x+0,45.
e) y=22x+412.

5. Assinale a alternativa que mostra as afirmativas que estão corretas.
I. O coeficiente de determinação indica a proporção de variação da variável independente que é explicada pela variável dependente, ou seja, é uma ferramenta que avalia a qualidade do ajuste.
II. O coeficiente de determinação não é explicado pela relação da variação explicada pela variação total.
III. O objetivo da regressão linear é fazer a análise estatística, verificando a relação funcional entre uma variável dependente com uma ou mais variáveis independentes.
IV. A regressão propõe uma equação que tenta explicar a variação da variável dependente pelas variáveis independentes.

a) I, IV.
b) I, II.
c) I, III, IV.
d) I, II, III.
e) I, II, IV.

6. Os dados a seguir correspondem às variáveis renda familiar e gasto com alimentação (em unidades monetárias) para uma amostra de 25 famílias.

Tabela 2.25 | Dados para o coeficiente de correlação

Renda Familiar (X)	Gasto com Alimentação (Y)
3	1,5
5	2,0
10	6,0
10	7,0
20	10,0
20	12,0
20	15,0
30	8,0
40	10,0
50	20,0
60	20,0
70	25,0

70	30,0
80	25,0
100	40,0
100	35,0
100	40,0
120	30,0
120	40,0
140	40,0
150	50,0
180	40,0
180	50,0
200	60,0
200	50,0

Fonte: O autor.

Encontre o coeficiente de correlação e o coeficiente de determinação.

7. Com os dados apresentados no exercício 6:

a) Obtenha a equação de regressão do gasto com alimentação em função da renda familiar.
b) Qual é o significado prático do valor do coeficiente angular da reta de regressão?

Referências

BARBETTA, P. A.; BORNIA, A. C. R. **Estatística para cursos de engenharia e informática**. 3. ed. São Paulo: Atlas, 2010.

CARVALHO, T. M. de. **Variabilidade espacial de propriedades físico-hídricas de em um latossolo vermelho-amarelo através da geoestatística**. 1991. 84 p. Dissertação (Mestrado) - Escola Superior de Agricultura de Lavras, Universidade Federal de Lavras, Lavras, 1991.

GROSSI SAD, J. H. **Fundamentos sobre variabilidade dos depósitos minerais**. Rio de Janeiro: DNPM/CPRM - GEOSOL, 1986. 141p.

HINES, W. W. et al. **Probabilidade e estatística na engenharia**. 4. ed. Rio de Janeiro: LTC, 2006.

JOHNSON, R.; KOBY, P. **Estatística**. São Paulo: Cengage Learning, 2013.

LAPPONI, J. C. **Estatística usando Excel 5 e 7**. Rio de Janeiro: Elsevier. 2005.

LARSON, R.; FARBER, B. **Estatística aplicada**. 4. ed. São Paulo: Pearson, 2010.

MARCONI, M. D. A.; LAKATOS, E. M. **Técnicas de pesquisa**: planejamento e execução de pesquisas, amostragens e técnicas de pesquisas, elaboração, análise e interpretação de dados. 3. ed. São Paulo: Atlas, 1996.

MOORE, D. S. **A estatística básica e sua prática**. 6. ed. Rio de Janeiro: LTC, 2014.

MORETTIN, L. G. **Estatística básica: probabilidade e inferência**. São Paulo: Pearson, 2010.

PINHEIRO, J. I. D. **Probabilidade e estatística**. Rio de Janeiro: Elsevier, 2012.

SPIEGEL, M. R. **Estatística**. 3. ed. São Paulo: Makron Books, 1993. 643p.

WALPOLE, R. E. **Probabilidade e estatística para engenheiria e ciências**. 8. ed. São Paulo: Pearson-Prentice Hall, 2009.

Distribuições de Probabilidade Discretas e Contínuas

Convite ao estudo

A probabilidade é usada para estudar as chances que um fenômeno ou algum evento tem de acontecer ou se repetir. São exemplos clássicos de Probabilidade os jogos de azar, que são os jogos de cartas, as roletas e os dados. No Brasil, esses jogos são proibidos, mas em outros países, até mesmo em países vizinhos, os jogos de azar podem ser praticados livremente.

Nesta unidade, vamos estudar a probabilidade de modo a ajudá-lo a compreender como ela pode ser usada no nosso dia a dia. Quando fazemos um jogo de loteria, você sabia que podemos calcular a probabilidade de ganharmos o prêmio?

Uma distribuição de probabilidade mostra a chance que algo ou uma variável pode assumir, considerados alguns valores. Uma distribuição pode ser discreta com valores certos, como os jogos de dados e jogos de cartas, ou pode ser contínua.

Na seção 3.1, veremos o que são o Espaço amostral e os Eventos disjuntos. Na seção 3.2, estudaremos Definição da Distribuição Discreta de Probabilidade e a Distribuição de probabilidade binomial. Na seção 3.3, veremos os conceitos de Distribuição de Probabilidade de Poisson e Definição da Distribuição Contínua de Probabilidade. Na seção 3.4, veremos a Distribuição Normal e a Distribuição Normal Padrão.

Os objetivos de aprendizagem desta unidade são: definir o espaço amostral e compreender o que são eventos disjuntos; compreender a definição de Distribuição Discreta de Probabilidade e Probabilidade

Binomial; compreender a probabilidade de Poisson e a distribuição contínua de probabilidade; e, por último, saber como se comporta uma Distribuição Normal.

Todos esses conceitos nos ajudarão a alcançar a competência de fundamentos de área, que é conhecer os fundamentos estatísticos básicos necessários à formação do profissional.

Você está de férias em um país estrangeiro e foi conhecer os cassinos da região. Você escolheu um hotel-resort associado a um renomado cassino para passar seus dias de férias e se divertir com os jogos de azar. Com os conceitos de probabilidade, você será capaz de determinar a probabilidade de ganhar na mesa de dados e nos jogos de cartas que são seus preferidos, mas também não deixaremos de fora a máquina de caça-níquel e o bingo on-line.

Pronto para começar?

Seção 3.1

Espaço amostral e eventos disjuntos

Diálogo aberto

A probabilidade é usada para estudar as chances que um fenômeno ou algum evento tem de acontecer ou se repetir. São exemplos clássicos de probabilidade os jogos de azar, que são os jogos de cartas, as roletas e os dados. No Brasil, esses jogos são proibidos, mas em outros países, até mesmo em países vizinhos, os jogos de azar podem ser praticados livremente.

Todas as vezes que precisamos saber se um evento ocorrerá, utilizaremos os conceitos de probabilidade.

A competência que você terá ao final desta unidade é conhecer os fundamentos estatísticos básicos necessários à formação do profissional. Os objetivos desta seção são definir o espaço amostral e compreender o que são eventos disjuntos.

Os jogos de azar têm uma probabilidade entre a sorte e o azar. As probabilidades de sorte são bem menores que as de os jogadores terem azar. Os jogos são sustentáveis através da perda dos jogadores que financiam os que têm sorte, que, como já dizemos, são poucos.

Na essência do jogo de azar está a tomada de decisão sob condições de risco, para isso os jogadores conhecem os regulamentos. Os prêmios são estipulados pela combinação escolhida e pela probabilidade de acerto.

Você está de férias no exterior e foi conhecer os cassinos da região. Você escolheu um hotel-resort associado a um cassino para passar seus dias de férias e se divertir com os jogos de azar. Com os conceitos de probabilidade, você será capaz de determinar a probabilidade de ganhar na mesa de dados e nos jogos de cartas que são seus preferidos, mas também não deixaremos de fora a máquina caça-níquel e o bingo on-line. O primeiro jogo que você quer saber qual será a probabilidade de sair um número que você escolherá é a mesa de dados. Você começará com apenas um dado e seus números preferidos são 2 e 5. Encontre a probabilidade de sair esses dois números. Identifique qual é seu espaço amostral e quais são os eventos. Considere o dado sendo um dado honesto.

Não pode faltar

Espaço Amostral – É o conjunto de todos os resultados possíveis em um experimento aleatóriol.

Exemplificando

Por exemplo, para uma moeda que será lançada, o espaço amostral será o conjunto {cara, coroa}, pois são os dois resultados possíveis de se obter ao se jogar a moeda.

Pode ser representado pela letra S, da seguinte forma: S = {cara, coroa}.

Evento – Quando uma moeda é lançada, evento é a ocorrência desse fato. Serão os subconjuntos do espaço amostral.

Exemplificando

Para o exemplo do lançamento da moeda, os subconjuntos são:

A = {cara}

B = {coroa}

Tanto A quanto B estão contidos em S, por isso são chamados subconjuntos de S.

Classificação de Eventos

Podemos observar os seguintes tipos de eventos:

Evento Simples – Classificamos assim os eventos que são formados por um único elemento do espaço amostral.

A = { 5 } é a representação de um evento simples do lançamento de um dado cuja face para cima é divisível por 5. Nenhuma das outras possibilidades é divisível por 5.

Evento Certo – Ao lançarmos um dado, é certo que a face que ficará para cima, terá um número divisor de 720. Este é um evento

certo, pois 720 = 6! = 6 . 5 . 4 . 3 . 2 . 1. Obviamente, qualquer um dos números da face de um dado é um divisor de 720, pois 720 é o produto de todos eles.

O conjunto A = { 2, 3, 5, 6, 4, 1 } representa um evento certo, pois ele possui todos os elementos do espaço amostral S = { 1, 2, 3, 4, 5, 6 }.

Evento Impossível – No lançamento conjunto de dois dados, qual é a possibilidade de a soma dos números contidos nas duas faces para cima, ser igual a 15?

Este é um evento impossível, pois o valor máximo que podemos obter é igual a doze. Podemos representá-lo por A = {}.

Evento União – Seja A = { 1, 3 } o evento de ocorrência da face superior no lançamento de um dado, ímpar e menor ou igual a 3, e seja B = { 3, 5 }, o evento de ocorrência da face superior, ímpar e maior ou igual a 3. Então, C = { 1, 3, 5 } representa o evento de ocorrência da face superior ímpar, que é a união dos conjuntos A e B.

Note que o evento C contém todos os elementos de A e B.

Evento Intersecção – Seja A = { 2, 4 } o evento de ocorrência da face superior no lançamento de um dado, par e menor ou igual a 4, e seja B = { 4, 6 }, o evento de ocorrência da face superior, par e maior ou igual a 4. Então, C = { 4 } representa o evento de ocorrência da face 4 ao mesmo tempo no conjunto A e B.

Veja que o evento C contém apenas os elementos comuns a A e B.

Eventos Mutuamente exclusivos – Seja A = {1, 2, 3, 6} o evento de ocorrência da face superior no lançamento de um dado, um número divisor de 6, e seja B = { 5 }, o evento de ocorrência da face superior, um divisor de 5. Então, os eventos A e B são mutuamente exclusivos, pois, os eventos não possuem elementos em comum.

Evento Complementar – Seja A = {1, 3, 5} o evento de ocorrência da face superior no lançamento de um dado, um número ímpar, o seu evento complementar é AC = {2, 4, 6} , isto é, o evento de

ocorrência da face superior no lançamento de um dado, um número par.

Os elementos de A são todos os elementos do espaço amostral S que não estão contidos em A, então temos que A = S - AC e ainda que S = A + AC.

O conceito de Probabilidade

Para eventos aleatórios, existe a incerteza se um evento irá acontecer. Essa medida de chance ou probabilidade, que podemos esperar que o evento ocorra, designamos um número entre 1 e 0. Se são certos ou seguros que ocorrerá, dizemos que temos 100% de probabilidade ou 1, mas se sabemos que não ocorrerá o evento, podemos afirmar que sua probabilidade é zero.

Assimile

Axiomas da probabilidade, que são:

Para cada evento de A, temos $P(A) \geq 0$.

Para o evento certo ou garantido de S na classe $C \geq P(S)=1$.

Para qualquer evento considerado mutuamente exclusivo A1, A2, ..., na classe C, temos:

$P(A_1 \cup A_2 \cup ...) = P(A_1) + P(A_2) + ...$

Para dois eventos mutuamente exclusivos A_1, A_2, temos:

$P(A_1 \cup A_2) = P(A_1) + P(A_2)$

Teoremas de Probabilidade

Teorema 1 – Se $P(A_1) \leq P(A_2)$, $P(A_2 - A_1) = P(A_2) + P(A_1)$.

Teorema 2 – Para cada evento A $0 \leq P(A) \leq 1$, a probabilidade está entre 0 e 1.

Teorema 3 – O evento é impossível se a probabilidade é zero.

Teorema 4 – Se A' é complemento de A, então P(A')=1-P(A).

Teorema 5 – Se $A = A_1 \cup A_2 \cup \ldots A_n$ são eventos mutuamente exclusivos, então temos:

$$P(A)=P(A_1)+P(A_2)+\ldots+P(A_n)$$

Em particular, se A=S, o espaço amostral, teremos:

$$P(A)=P(A_1)+P(A_2)+\ldots+P(A_n) = 1$$

Teorema 6 – Se A e B são quaisquer dois eventos, então:

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

Para três eventos A1, A2, A3:

$$P(A_1\cup A_2\cup A_3)=$$

$$P(A_1)+P(A_2)+P(A_3)-P(A_1\cap A_2)-P(A_2\cap A_3)-P(A_1\cap A_3)+P(A1\cap A_2\cap A_3)$$

Serve para n eventos também.

Teorema 7 – Para quaisquer eventos A e B:

$$P(A)=P(A\cap B)+P(A\cap B')$$

Teorema 8 – Se um evento A deve resultar na ocorrência de um dos eventos mutuamente exclusivos A_1, A_2, ...An, teremos:

$$P(A)=P(A\cap A_1)+P(A\cap A_2)+\ldots+P(A\cap A_n)$$

Probabilidade Condicional

Antes da realização de um experimento, é necessário que já se tenha alguma informação sobre o evento que se deseja observar. Nesse caso, o espaço amostral se modifica e o evento tem a sua probabilidade de ocorrência alterada.

Fórmula de Probabilidade Condicional

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

Onde P(B|A) é a probabilidade de ocorrer B, condicionada pelo fato de já ter ocorrido A.

Complemente seus estudos

Para saber mais sobre o espaço amostral, o link a seguir poderá lhe auxiliar a estudar um pouco mais: <http://www.mspc.eng.br/matm/prob_est110.shtml>. Acesso em: 30 jul. 2015.

Faça você mesmo

Em uma caixa com 500 lâmpadas, há 20 defeituosas. Se A é o evento "lâmpada com defeito" e a referência é toda a caixa, qual é a probabilidade pela abordagem frequencial?

Vocabulário

Axiomas – verdades inquestionáveis universalmente válidas, muitas vezes utilizadas como princípios na construção de uma teoria ou como base para uma argumentação.

Empírica – Que faz alusão ao empirismo. Que se apoia exclusivamente na experiência e na observação.

Sem medo de errar

O primeiro jogo que você quer saber qual será a probabilidade de sair um número que você escolherá é a mesa de dados. Você começará com apenas um dado e seus números preferidos são 2 e 5. Encontre a probabilidade de sair esses dois números. Identifique qual é seu espaço amostral e quais são os eventos. Considere o dado sendo um dado honesto.

Seu espaço amostral é:

$S = \{1,2,3,4,5,6\}$

Se as probabilidades forem atribuídas igualmente aos pontos amostrais, teremos:

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$$

Você escolherá os números 2 e 5 para apostar. O evento de aparecer tanto 2 quanto 5 é indicado por 2 U 5.

Assim, a probabilidade de sair esses dois números na mesa de dados é:

$$P(2 \cup 5) = P(2) + P(5) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Interpretando essa probabilidade encontrada, há aproximadamente 33% de chances de sair 2 ou 5 na mesa de dados. De cada 3 jogadas, uma poderá ser um desses números.

Atenção

Para eventos aleatórios, existe a incerteza se um evento irá acontecer. A essa medida de chance ou probabilidade, em que podemos esperar que o evento ocorra, designamos um número entre 1 e 0.

Lembre-se

São os tipos de eventos possíveis de ocorrer:

Evento Simples.

Evento Certo.

Evento Impossível.

Evento União.

Evento Intersecção.

Eventos Mutuamente Exclusivos.

Evento Complementar.

Avançando na prática

Pratique mais	
Instrução Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.	
"Lançamento do Dado Honesto"	
1. Competência de fundamento de área	Conhecer os fundamentos estatísticos básicos necessários à formação do profissional.
2. Objetivos de aprendizagem	Definir o espaço amostral e compreender o que são eventos disjuntos.
3. Conteúdos relacionados	Espaço Amostral e Eventos Disjuntos.
4. Descrição da SP	Agora o dado foi lançado duas vezes. Qual será a probabilidade de na primeira jogada sair 4, 5 ou 6 e na segunda jogada 1, 2, 3 ou 4?
5. Resolução da SP	A será o evento "4, 5 ou 6 no primeiro lançamento" e B será o evento "1, 2, 3 ou 4 no segundo lançamento". Usaremos o conceito de probabilidade condicional. $P(A \cap B) \equiv P(A)P(B\|A) = P(A)P(B) = \left(\frac{3}{6}\right)\left(\frac{4}{6}\right) = \frac{1}{3}$ O segundo lançamento é independente do primeiro. Então usamos 3/6 para o primeiro evento, pois são 3 das 6 possibilidades igualmente possíveis. Para o segundo evento, usamos a probabilidade 4/6, pois são 4 das 6 possibilidades igualmente possíveis.

Lembre-se

Probabilidade Condicional

Fórmula de Probabilidade Condicional

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

Onde P(B|A) é a probabilidade de ocorrer B, condicionada pelo fato de já ter ocorrido A.

Você pode estudar mais sobre Probabilidade condicional no link: <http://www.ime.unicamp.br/~veronica/Coordenadas1s/aula5.pdf>. Acesso em: 30 jul. 2015.

Faça você mesmo

Em um sorteio de um carro em uma loja, foram colocadas em uma urna 100 bolas, enumeradas de 1 a 100. Para os 100 primeiros consumidores que estiveram na loja no dia da inauguração foi dado um bilhete com um número. Qual a probabilidade do seu número ser sorteado?

Faça valer a pena

1. Dados do DETRAN mostram que, em 2014, as vítimas fatais em decorrência de acidentes de carro foram 50 pessoas. O perfil das pessoas que vieram a óbito está no quadro a seguir.

Pedestre	22
Condutor de moto	12
Ciclista	8
Condutor de automóvel	3
Passageiro de ônibus	2
Passageiro de automóvel	1
Condutor de caminhão	1
Passageiro de moto	1

Fonte: Adaptado de: Correio Brasiliense, 20/7/2009.

Com os dados apresentados, qual é a probabilidade de uma vítima fatal ser um pedestre?
a) 12/50.
b) 16/25.
c) 8/50.
d) 11/25.
e) 1/25.

2. Ao se jogar dois dados, qual a probabilidade de se obter o número 7 como soma dos resultados?

a) 7/12.
b) 6/12.
c) 4/12.
d) 2/12.
e) 0.

3. O quadro a seguir apresenta o número estimado da população em cada região brasileira no ano de 2007, a porcentagem estimada de pessoas por região que possuem aparelho de telefone celular, e a multiplicação dessas duas quantidades por região (pop x cel), com duas casas decimais de precisão.

Região	Nº de habitantes da região (em 1.000.000) (pop)	Porcentagem de habitantes da região que possuem celular (cel)	Pop x cel
Sudeste	77,9	52%	40,51
Nordeste	51,5	44%	22,6
Sul	26,7	61%	16,29
Norte	14,6	43%	6,28
Centro-Oeste	13,3	60%	7,92
Total	184,0	-	93,66

Fonte: IBGE, TIC Domicílios do NIC.br.

De acordo com o quadro anterior, a probabilidade aproximada de um brasileiro que possui aparelho celular viver na região Norte ou na região Sul é:

a) 12,4%.
b) 20,2%.
c) 24,1%.
d) 35,8%.
e) 42,6%.

4. Analisando um lote de 360 peças para computador, o departamento de controle de qualidade de uma fábrica constatou que 40 peças estavam com defeito. Retirando-se uma das 360 peças, ao acaso, a probabilidade de esta peça NÃO ser defeituosa é:

a) 1/9.
b) 2/9.
c) 5/9.
d) 7/9.
e) 8/9.

5. Em uma pequena localidade, os amigos Arnor, Bruce, Carlão, Denílson e Eleonora são moradores de um bairro muito antigo que está comemorando 100 anos de existência. Dona Matilde, uma antiga

moradora, ficou encarregada de formar uma comissão que será a responsável pela decoração da festa. Para tanto, Dona Matilde selecionou, ao acaso, três pessoas entre os amigos Arnor, Bruce, Carlão, Denílson e Eleonora. Sabendo-se que Denílson não pertence à comissão formada, a probabilidade de Carlão pertencer à comissão é, em termos percentuais, igual a:

a) 30%.
b) 80%.
c) 62%.
d) 25%.
e) 75%.

6. Numa determinada zona eleitoral, sabe-se que 40% dos eleitores são do sexo masculino. Entre esses, 10% têm curso superior, ao passo que entre os eleitores do sexo feminino, 25% têm curso superior. Calcule a probabilidade de escolher um eleitor que seja do sexo feminino ou que não tenha curso superior.

7. Um grupo é formado por 10 pessoas, cujas idades são: 17, 19, 19, 20, 20, 20, 20, 21, 22 e 22. Escolhendo-se, aleatoriamente, uma pessoa do grupo, qual a probabilidade de que sua idade seja maior do que a moda?

Seção 3.2

Definição da distribuição discreta de probabilidade e distribuição de probabilidade binomial

Diálogo aberto

Na seção anterior, estudamos o que é o espaço amostral e quais são os eventos ou ocorrências de fatos que podemos estudar na probabilidade. Nesta seção, vamos estudar Definição da Distribuição Discreta de probabilidade e Distribuição de Probabilidade Binomial.

As distribuições de probabilidade farão a associação de uma probabilidade ao possível número que será resultado numérico de uma verificação, teste ou experimento.

As distribuições discretas de probabilidade expressam os valores finitos que as variáveis podem assumir. As distribuições chamadas binomiais são distribuições que expressam uma quantidade de sucessos em n ensaios independentes.

Para conhecer os fundamentos estatísticos básicos necessários à formação do profissional, que é a competência que estamos desenvolvendo, é necessário que você compreenda as distribuições de probabilidade que vamos estudar nesta seção. O objetivo desta seção é entender a teoria da probabilidade discreta e a distribuição binomial.

Você está de férias no exterior e foi conhecer os cassinos da região. Você escolheu um hotel-resort associado a um cassino para passar seus dias de férias e se divertir com os jogos de azar. Com os conceitos de probabilidade, você será capaz de determinar a probabilidade de ganhar na mesa de dados e nos jogos de cartas que são seus preferidos, mas também não deixaremos de fora a máquina caça-níquel e o bingo on-line. O segundo jogo que você experimentou no cassino foi o jogo de cartas. Você pode participar de cinco sorteios repetidos. A sua aposta é no naipe de espadas. Calcule a probabilidade de sair o naipe de espadas em cada um dos sorteios. Você assume a probabilidade de sucessos como sendo p=13/52 e a probabilidade de fracassos q=39/52, lembrando que p+q=1.

Não pode faltar

Definição da Distribuição Discreta de Probabilidade

Quando aplicamos a Estatística na resolução de problemas administrativos, verificamos que muitos problemas apresentam as mesmas características, o que nos permite estabelecer um modelo teórico para determinação da solução de problemas.

Os componentes principais de um modelo estatístico teórico são:

1. Os possíveis valores que a variável aleatória X pode assumir.
2. A função de probabilidade associada à variável aleatória X.
3. O valor esperado da variável aleatória X.
4. A variância e o desvio-padrão da variável aleatória X.

Há dois tipos de distribuições teóricas que correspondem a diferentes tipos de dados ou variáveis aleatórias: a distribuição discreta e a distribuição contínua.

Além de identificar os valores de uma variável aleatória, frequentemente podemos atribuir uma probabilidade a cada um desses valores. Quando conhecemos todos os valores de uma variável aleatória juntamente com suas respectivas probabilidades, temos uma distribuição de probabilidades.

A **distribuição de probabilidades** associa uma probabilidade a cada resultado numérico de um experimento, ou seja, dá a probabilidade de cada valor de uma variável aleatória. Por exemplo, no lançamento de um dado cada face tem a mesma probabilidade de ocorrência que é 1/6.

Como os valores das distribuições de probabilidades são probabilidades, e como as variáveis aleatórias devem tomar um de seus valores, temos as duas regras a seguir, que se aplicam a qualquer distribuição de probabilidades:

1. A soma de todos os valores de uma distribuição de probabilidades deve ser igual a 1.

$\sum P(x) = 1$, onde x toma todos os valores possíveis

2. A probabilidade de ocorrência de um evento deve ser maior

do que zero e menor do que 1. $0 \leq P(x) \leq 1$ para todo x.

No exemplo do lançamento de um dado, como todas as faces têm a mesma probabilidade de ocorrência, que é 1/6, ao somá-las obtemos o valor 1, que corresponde à primeira regra citada anteriormente. O valor 1/6 é maior do que zero e menor do que 1, assim, satisfaz a segunda regra citada.

Teoria de Probabilidade discreta

A **teoria de probabilidade discreta** lida com eventos que ocorrem em espaços amostrais enumeráveis.

Exemplificando

Lançamento de um dado, experimentos com baralhos de cartas, e uma caminhada aleatória são exemplos clássicos para a teoria da probabilidade discreta.

Definição clássica: inicialmente, a probabilidade de um evento a ocorrer foi definida como um número de casos favoráveis ao evento, sobre o número total de resultados possíveis.

Exemplificando

Se o evento é "ocorrência de um número par quando o dado é lançado", a probabilidade é dada por , uma vez que 3 faces das 6 têm números pares.

Definição moderna: a definição moderna começa com um conjunto chamado de espaço amostral, que se relaciona ao conjunto de todos resultados possíveis no sentido clássico, denotado por $S = \{x_1, x_2, \ldots\}$. Em seguida, é assumido que para cada elemento $x \in S$, um número intrínseco de "probabilidade" f(x) é associado, que satisfaz as seguintes propriedades:

1. $f(x) \in [0, 1]$ para todo $x \in S$
2. $\sum_{x \in \Omega} f(x) = 1$

Um evento é definido como qualquer subconjunto E do espaço amostral S. A probabilidade do evento é:

$$\mathbf{P(}\boldsymbol{E}\mathbf{)} = \sum_{x \in E} f(x)$$

Assim, a probabilidade de todo espaço amostral é 1, e a probabilidade do evento nulo é 0.

A função f(x) que transforma um ponto no espaço amostral no valor da "probabilidade" é chamada de **função de massa de probabilidade**, abreviada como **fmp (= pmf-probability mass function)**. A definição moderna não tenta responder como as funções de massa de probabilidade são obtidas, em vez disso, constrói uma teoria que assume sua existência.

Função de Densidade de Probabilidade

Distribuição discreta

Se X é uma variável que pode assumir um conjunto discreto de valores X_1, X_2, X_3, ..., Xk com respeito a probabilidades p_1, p_2, p_3,..., p_k, onde $p_1 + p_2 + p_3 + ... + p_k = 1$, dizemos que uma distribuição discreta de probabilidade para X foi definida. A função p(X), com os valores respectivos p_1, p_2, p_3, ..., p_k para $X = X_1, X_2, X_3, ..., X_k$ é chamada de **função de probabilidade**, ou **função de frequência**, de X, porque X pode assumir certos valores com"probabilidades dadas. Esta função é muitas vezes chamada de variável aleatória discreta. Uma variável aleatória é também conhecida como **variável de chance** ou **variável estocástica** (SPIEGEL, 2006, p. 130).

Distribuição de probabilidade binomial

A distribuição binomial é aplicada frequentemente para descrever controle estatístico de qualidade de uma população. Tem-se interesse principalmente em duas categorias: item defeituoso ou insatisfatório *versus* item bom ou satisfatório e sucesso e falhas que tenham ocorrido em uma amostra de tamanho fixo.

A distribuição binomial é aplicada a eventos provenientes de uma série de experimentos aleatórios, que constituem o chamado Processo de Bernoulli.

Processo de *Bernoulli*

Esse processo é análogo àquele de jogar uma moeda. As seguintes suposições se aplicam:

a) Cada experimento é dito ser uma tentativa. Existe uma série de tentativas, cada uma tendo dois resultados: sucesso ou falha.

b) A probabilidade de sucesso é igual a algum valor constante para todas as tentativas.

c) Os resultados sucessivos são estatisticamente independentes. A probabilidade de sucesso na próxima tentativa não pode variar, não importando quantos sucessos ou falhas tenham sido obtidos.

O processo de Bernoulli é comumente utilizado em aplicações de engenharia envolvendo controle de qualidade. Cada novo item criado no processo de produção pode ser considerado como uma tentativa, resultando em uma unidade com ou sem defeito. Esse processo não se limita a objetos, podendo ser usado em pesquisas eleitorais e de preferências dos consumidores por determinados produtos.

A Distribuição de Bernoulli é a distribuição de uma variavel aleatória X associada a um experimento de Bernoulli, em que se define X=1 se ocorre sucesso e X=0 se ocorre fracasso. Chamando de p a probabilidade de sucesso (0<p<1) e de q a probabilidade de fracasso, a distribuição de Bernoulli é:

x	0	1
Pr(X=x)	1-p	p

Certamente, as condições definidas de uma função de distribuição de probabilidades (fdp) são satisfeitas, uma vez que p>0, 1-p>0 e p+(1-p)=1. O valor de p é o único valor que precisamos conhecer para determinar completamente a distribuição. Ele é, então, chamado de parâmetro da distribuição de Bernoulli.

A função de distribuição acumulada é dada por:

$$F_x(x)\begin{cases} 0 & se\ x < 0 \\ 1-p & se\ 0 \leq x \leq 1 \\ 1 & se\ x \geq 1 \end{cases}$$

Esperança: pode ser calculada, sendo E(X) = p

Variância: pode ser calculada por Var(X) = p(1-p)

Reflita

Distribuições binomiais

O nome binomial é devido à fórmula, pois representa o termo geral do desenvolvimento do binômio de Newton.

A distribuição binomial é provavelmente a mais simples distribuição teórica possível e, portanto, frequentemente empregada em livros-texto para ilustrar o uso e as propriedades das distribuições teóricas mais gerais. Essa distribuição pertence a situações em que existem dois eventos possíveis de ocorrerem. Classicamente, estes dois eventos têm sido referidos como "sucesso" e "falha", mas essa atribuição é meramente arbitrária. De maneira mais geral, um dos eventos (digamos o "sucesso") é designado com o número 1, e o outro ("falha") com o número zero.

A variável aleatória de interesse, X, é o número de ocorrências do evento (dado pela soma de valores 1 ou 0) em um número de tentativas. O número de tentativas N pode ser qualquer inteiro positivo e a variável X pode tomar qualquer valor não negativo inteiro, variando de 0 (se o evento de interesse não ocorrer para todas as N tentativas) a N (se ocorrer em todas as ocasiões). A distribuição binomial pode ser usada para calcular probabilidades para cada um destes N+1 possíveis valores de X se as seguintes condições forem satisfeitas:

1) a probabilidade do evento ocorrer não mudar de tentativa para tentativa e

2) as saídas ou resultados das N tentativas forem mutuamente independentes.

Essas duas condições são, raramente, estritamente satisfeitas, mas situações reais podem ser próximas o suficiente a esta ideal, tal que a distribuição binomial fornece representações suficientemente acuradas.

A distribuição Binomial é o modelo probabilístico adequado para casos em que se consideram repetidas provas de Bernoulli, isto é, sucessões de experimentos aleatórios independentes, em cada um dos quais se observa a ocorrência ("sucesso") ou não ("fracasso") de um determinado acontecimento, de probabilidade p, constante de observação para observação. Seja a v.a.d. X: número de sucessos em n provas. A distribuição de probabilidade f(x) é dada por:

$f(x) = P(X=x) = P(x) = C_{n}^{x} p^{x}.q^{(n-x)}$

P(x) = é a probabilidade de que o evento se realize x vezes em n provas.

p = é a probabilidade de que o evento se realize em uma só prova = sucesso.

q = é a probabilidade de que o evento não se realize no decurso dessa prova = insucesso.

Parâmetros da Distribuição Binomial

Média = $\mu = n.p$

Desvio-padrão = $\sigma = \sqrt{(n \cdot p \cdot q)}$

Variância = $\sigma 2 = n \cdot p \cdot q$

Assimile

Um experimento de probabilidade binominal é composto por testes repetidos com as seguintes propriedades:

1. Existem n testes independentes idênticos repetidos.

2. Cada teste tem dois resultados possíveis (sucesso ou fracasso).

Assimile

3. P(sucesso) = p e P(fracasso) = q e p+ q = 1.

4. A variável aleatória binomial x é a contagem do número de testes bem-sucedidos que ocorreram; x pode assumir qualquer valor inteiro de zero a n.

Função de Probabilidade Binomial

Para um experimento binomial, considere que p representa a probabilidade de sucesso e q a probabilidade de fracasso de um único teste. Então, P(x) é a probabilidade de que haverá exatamente x sucessos em n testes:

$$P(x) = \binom{n}{k}(p^x)(q^{n-x}) \text{ para x } = 0,1,2,\ldots \text{n}$$

A fórmula tem três fatores básicos:

1. O número de formas que exatamente x sucessos podem ocorrer em n testes (n).
k

2. A probabilidade de exatamente x sucessos (px).

3. A probabilidade de que ocorra um fracasso nos (n-x) testes restantes de (qn-x).

Para calcular o coeficiente $\binom{n}{k}$, que é chamado de coeficiente binominal, utilizamos a seguinte fórmula:

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

Complemente seus estudos

No portal Action, você terá uma explicação prática sobre a Distribuição Binomial, para saber mais pesquise no *link*: <http://www.portalaction.com.br/probabilidades/51-distribuicao-binomial>. Acesso em: 30 jul. 2015.

Faça você mesmo

Dois alunos estão jogando um dado honesto. Um deles quer saber qual será a probabilidade em cinco lançamentos de sair 1 apenas uma vez.

Sem medo de errar

Calcule a probabilidade de sair o naipe de espadas em cada um dos sorteios. Você assume a probabilidade de sucessos como sendo p=13/52 e a probabilidade de fracassos q=39/52, lembrando que p+q=1.

Existem cinco sorteios repetidos: n = 5. Os sorteios são individuais e independentes, pois a carta sorteada é devolvida ao baralho e embaralhada novamente.

Para você, interessa se a carta "é de espadas" ou "não é de espadas".

p=P(de espadas) = 13/52

q=P(não é de espadas) = 39/52

x é o número de espadas registradas nos 5 sorteios. São valores possíveis (1, 2, 3, 4, 5).

A função de probabilidade binomial é:

$$P(x) = \binom{5}{x}\left(\frac{13}{52}\right)^{x}\left(\frac{39}{52}\right)^{5-x} = \binom{5}{k}\left(\frac{1}{4}\right)^{x}\left(\frac{3}{4}\right)^{5-x} = \binom{5}{x}(0{,}25)^{x}(0{,}75)^{5-x} \text{ para x } = 0, 1, \cdots, 5$$

Alteramos o x para cada sorteio e calculamos a probabilidade de sortearmos uma carta de espadas em cada um dos sorteios.

$$P(0) = \binom{5}{0}(0{,}25)^{0}(0{,}75)^{5-0} = 0{,}2373 \qquad P(1) = \binom{5}{1}(0{,}25)^{1}(0{,}75)^{5-1} = 0{,}3955$$

$$P(2) = \binom{5}{2}(0{,}25)^{2}(0{,}75)^{5-2} = 0{,}2637 \qquad P(3) = \binom{5}{3}(0{,}25)^{3}(0{,}75)^{5-3} = 0{,}0879$$

$$P(4) = \binom{5}{4}(0{,}25)^{4}(0{,}75)^{5-4} = 0{,}0146 \qquad P(5) = \binom{5}{5}(0{,}25)^{5}(0{,}75)^{5-5} = 0{,}0010$$

O número mais provável é o no sorteio 1, pois as cartas são repostas no baralho e embaralhadas novamente.

Atenção

Para um experimento binomial:

$$P(x) = \binom{n}{x}(p^x)(q^{n-x}) \text{ para x } = 0,1,2,\ldots \text{n}$$

A fórmula tem três fatores básicos:

O número de formas que exatamente x sucessos podem ocorrer em n testes $\binom{n}{x}$.

A probabilidade de exatamente x sucessos (px).

A probabilidade de que ocorra um fracasso nos (n-x) testes restantes de (qn-x).

Para calcular o coeficiente $\binom{n}{x}$, o qual deve ser sempre um inteiro positivo e é chamado de coeficiente binominal, utilizamos a seguinte fórmula:

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

Avançando na prática

Pratique mais	
Instrução Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.	
"Lançamento de Dados"	
1. Competência de fundamento de área	Conhecer os fundamentos estatísticos básicos necessários à formação do profissional.
2. Objetivos de aprendizagem	Entender a teoria da probabilidade discreta e a distribuição de probabilidade binomial.
3. Conteúdos relacionados	Distribuição de Probabilidade Binomial.
4. Descrição da SP	Dois amigos estão brincando fazendo apostas com um dado honesto. Um deles quer saber qual será a probabilidade em cinco lançamentos de sair um 3: (a) duas vezes, (b) no máximo uma vez e (c) pelo menos duas vezes.

5. Resolução da SP	A variável aleatória será X, sendo o número de vezes que um 3 aparece em cinco lançamentos do dado honesto. Probabilidade de um 3 em um único lançamento é p=1/6. Probabilidade de nenhum 3 em um único lançamento é q= 1-p = 5/6. $a)\ P(3\ ocorrer\ duas\ vezes) = P(x=2) = \binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^{5-2} = \frac{625}{3888}$ $b)\ P(3\ ocorrer\ no\ maximo\ uma\ vez) =$ $P(x \leq 1) = P(x=0) + P(x=1) =$ $\binom{5}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^{5-0} + \binom{5}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^{5-1} = \frac{3125}{7776} + \frac{3125}{7776} = \frac{3125}{3888}$ $c)\ P(3\ ocorrer\ pelo\ menos\ duas\ vezes) =$ $P(x \geq 2) = P(x=2) + P(x=3) + P(x=4) + P(x=5) =$ $\binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^{5-2} + \binom{5}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^{5-3} + \binom{5}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^{5-4} + \binom{5}{5}\left(\frac{1}{6}\right)^5\left(\frac{5}{6}\right)^{5-5}$ $= \frac{625}{3888} + \frac{125}{3888} + \frac{25}{7776} + \frac{1}{7776} = \frac{763}{3888}$ A probabilidade de sair um 3 é maior na condição em que ele ocorre no máximo uma vez em 5 jogadas.

Faça você mesmo

Calcule a probabilidade de sair o naipe de copas em cada um dos sorteios. A probabilidade de fracassos é q=39/52, lembrando que p+q=1. Calcule o p.

Faça valer a pena

1. Em um jogo com moedas honestas, qual será a probabilidade de ter 3 caras em 5 jogadas?

a) 0,3125.
b) 0,6911.
c) 0,4597.
d) 0,5214.
e) 0,2147.

2. Para o mesmo jogo com as moedas honestas, calcule agora a probabilidade de ocorrer menos de 3 caras em 5 jogadas.

a) 0,7.
b) 0,5.
c) 0,6.
d) 0,4.
e) 0,3.

3. Os pais sabem que a sua probabilidade de terem filhos com a pele morena é igual a ¼. Se o casal tiver 6 crianças, qual é a probabilidade de 3 delas terem a pele morena?

a) 0,52.
b) 0,26.
c) 0,48.
d) 0,13.
e) 0,66.

4. A probabilidade que você tem de acertar o alvo em um jogo de dardos é 0,3. Após 4 lançamentos, qual é a probabilidade de que você acerte o alvo pelo menos 3 vezes?

a) 0,0744.
b) 0,5123.
c) 0.
d) 0,0837.
e) 0,7892.

5. Um engenheiro da qualidade foi contratado para verificar a qualidade das peças que foram produzidas por uma indústria. Ele retirou uma amostra de 10 peças de forma aleatória e sabe que 20% das peças têm defeitos. Qual é a probabilidade de que não mais de 2 peças da amostra tenham defeitos?

a) 65,02%
b) 55,41%
c) 67,78%.
d) 71,03%.
e) 84,27%.

6. Quais são as condições para que se possa usar a distribuição binomial?

7. Agora o engenheiro extraiu uma amostra de 15 peças aleatórias na fabricação, sabendo que 85% das peças que são produzidas estão dentro da padronização. Calcule a probabilidade de 10 peças da amostra estarem dentro da padronização.

Seção 3.3

Distribuição de probabilidade de Poisson e definição da distribuição contínua de probabilidade

Diálogo aberto

Na seção anterior, apresentamos a Teoria da probabilidade discreta e a distribuição de probabilidade binomial, que considera que p representa a probabilidade de sucesso e q a probabilidade de fracasso de um único teste, o teste binomial.

Nesta seção, vamos estudar a Distribuição de Poisson, que é empregada em experimentos em que o interesse não é o número de sucessos em n tentativas, mas sim o número de sucessos durante um intervalo contínuo, podendo ser tempo, espaço, entre outros.

Para que você atinja, plenamente, a competência de fundamento de área, que é conhecer os fundamentos estatísticos básicos necessários à formação do profissional, a Distribuição de Probabilidade de Poisson e as Distribuições contínuas são muito importantes. Os objetivos de aprendizagem desta seção são: compreender o uso da Distribuição de Poisson e assimilar a definição da Distribuição contínua de probabilidade e sua utilização, para isso vamos entender a teoria aplicada a essas distribuições e desenvolver alguns exemplos.

Você está de férias no exterior e foi conhecer os cassinos da região. Você escolheu um hotel-resort associado a um cassino para passar seus dias de férias e se divertir com os jogos de azar. Com os conceitos de probabilidade, você será capaz de determinar a probabilidade de ganhar na mesa de dados e nos jogos de cartas que são seus preferidos, mas também não deixaremos de fora a máquina de caça-níquel e o bingo on-line. Agora você está na máquina de caça níquel, o operador da máquina disse que ela é honesta e há em média 2 ganhadores por dia. Qual a probabilidade de ter 5 ganhadores ou nenhum ganhador durante o dia?

Não pode faltar

Distribuição de Probabilidade de Poisson

Assimile

Foi o pesquisador S. D. Poisson quem descobriu, no início do século XIX, a distribuição de probabilidade em que a variável aleatória que tem essa distribuição é chamada de Poisson Distribuída.

A distribuição de Poisson é empregada em experimentos, nos quais não se está interessado no número de sucessos obtidos em n tentativas, como ocorre no caso da distribuição Binomial, mas sim no número de sucessos ocorridos durante um intervalo contínuo, que pode ser um intervalo de tempo, espaço etc.

Exemplificando

São alguns exemplos para a Distribuição de Probabilidade de Poisson:

Em um ano, a quantidade de suicídios em um município. Em um período de 15 minutos, o número de pessoas em um caixa no banco.

Quantidade de carros que passa no cruzamento da Avenida Paulista em um minuto, durante uma certa hora do dia.

É uma distribuição de probabilidade discreta que se aplica à ocorrência de eventos ao longo de intervalos especificados. A variável aleatória é o número de ocorrência do evento no intervalo. Os intervalos podem ser de tempo, distância, área, volume ou alguma unidade similar. É definida por:

$$f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Uma variável aleatória X para a distribuição de Poisson tem as seguintes propriedades:

$$X = \{0, 1, 2, \ldots\}$$
$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$
$$E(X) = \mu = \lambda$$
$$Var(X) = \sigma^2 = \lambda$$

Reflita

Uma distribuição de Poisson difere de uma distribuição binomial nestes aspectos fundamentais: 1. A distribuição binomial é afetada pelo tamanho da amostra n e pela probabilidade p, enquanto que a distribuição de Poisson é afetada apenas pela média λ; 2. Na distribuição binomial, os valores possíveis da variável aleatória X são 0; 1; 2; ...n, mas a distribuição de Poisson tem os valores de X de 0; 1; 2; ..., sem qualquer limite superior.

Obs.: o parâmetro λ é usualmente referido como taxa de ocorrência.

Definição de Distribuição Contínua de Probabilidade

É uma variável aleatória que assume valores contínuos, ou seja, pode ser valores reais dentro de um intervalo ou infinitos valores reais. Vejamos alguns exemplos de uma variável aleatória contínua: temperatura, precipitação, porcentagem, etc. Vamos denotar por *f(X)* a função de densidade de probabilidade e *F(X)* a função de distribuição de probabilidade. Quando calculamos a integral da função *f(X)* no intervalo (*a*,*b*) obtemos a probabilidade de *X* entre os valores *a* e *b*, isto é:

$$P(a \le X \le b) = \int_a^b f(X)dX$$

Para a função distribuição de probabilidade temos:

$$F(X) = \Pr ob(X \le b) = \int_{-\infty}^b f(X)dX$$

Uma função densidade de probabilidade precisa satisfazer as seguintes propriedades:

$$f(X) \ge 0$$

$$\int_{-\infty}^{+\infty} f(X)dX = 1$$

$$P\left(X = a\right) = 0 = \int_a^a f(X)dX$$

A probabilidade para que *X* tem valores entre *a* e *b* é dada por:

$$P\left(a \le X \le b\right) = \int_a^b f(X)dX = \int_{-\infty}^b f(X)dX - \int_{-\infty}^a f(X)dX = P\left(X \le b\right) - P\left(X \le a\right) = F(b) - F(\boldsymbol{a})$$

Complemente seus estudos

A distribuição binomial é afetada pelo tamanho da amostra n e pela probabilidade p, enquanto a distribuição de Poisson é afetada apenas pela média λ.

Faça você mesmo

Calcule a probabilidade de um valor da variável aleatória X se encontrar entre 110 e 150.

Sem medo de errar

Agora você está na máquina de caça níquel, o operador da máquina disse que ela é honesta e há em média 2 ganhadores por dia. Qual a probabilidade de ter 5 ganhadores ou nenhum ganhador durante o dia?

λ=2 ganhadores/dia em média.

$$P(X = 5) = \frac{e^{-2}2^5}{5!} = 0{,}036 = 3{,}6\% \textit{ de probabilidade de ter 5 ganhadores por dia.}$$

Para nenhuma chamada, temos:

$$P(X = 0) = \frac{e^{-2}2^0}{0!} = 0{,}14 \textit{ ou } 14\% \textit{ de probabilidade de não ter nenhum ganhador por dia.}$$

A conclusão que tiramos é que a probabilidade de não ter nenhum ganhador é maior que ter um número de ganhadores maior que a média, sendo vantajosa a máquina para o cassino.

Atenção

Sobre a Distribuição de Poisson, temos uma aula com explicação prática, disponível em <https://youtu.be/Y1u3pUdFqZQ>. Acesso em: 30 jul. 2015.

Lembre-se

Definição de Distribuição Contínua de Probabilidade:

$$P(a \leq X \leq b) = \int_a^b f(X)dX$$

Para a função distribuição de probabilidade, temos:

$$F(X) = Prob\,(X \leq b) = \int_{-\infty}^{b} f(X)dX$$

A probabilidade para que X tenha valores entre a e b é dada por:

$$P(a \leq X \leq b) = \int_a^b f(X)dX = F(b) - F(a)$$

Avançando na prática

Pratique mais	
Instrução Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.	
Corpo de Bombeiros	
1. Competência de fundamento de área	Conhecer os fundamentos estatísticos básicos necessários à formação do profissional.
2. Objetivos de aprendizagem	Compreender o uso da Distribuição de Poisson e assimilar a definição da Distribuição contínua de probabilidade e sua utilização.
3. Conteúdos relacionados	Distribuição de Probabilidade de Poisson e Definição da Distribuição Contínua de Probabilidade
4. Descrição da SP	Em uma cidade, o corpo de bombeiros recebe 3 chamadas em média por dia. Para que a escala de folga dos bombeiros fosse estabelecida, foi calculada a probabilidade de se ter 4 chamadas em um dia,, nenhuma chamada em um dia e 20 chamadas na semana.
5. Resolução da SP	Para 4 chamadas num dia, temos: λ=3 cham/dia em média. $P(X = 4) = \frac{e^{-\lambda}\lambda^k}{k!} = \frac{e^{-3}3^4}{4!} = 0{,}1680\ ou\ 16{,}80\%$

5. Resolução da SP	Para nenhuma chamada temos: $P(X = 0) = \frac{e^{-\lambda}\lambda^{k}}{k!} = \frac{e^{-3}3^{0}}{0!} = 0{,}0498\ ou\ 4{,}98\%$ Para 20 chamadas na semana: X = número de chamadas por dia. Y = número de chamadas por semana. $E(X) = \lambda' = 7\ x\ E(X) = 21\ chamadas\ por\ semana.$ $P(X = 0) = \frac{e^{-\lambda}\lambda^{k}}{k!} = \frac{e^{-21}21^{20}}{20!} = 0{,}0867\ ou\ 8{,}67\%$

Lembre-se

A distribuição de probabilidade discreta é aplicada para eventos que ocorrem ao longo de um intervalo de tempo.

A variável aleatória será o número de ocorrência do evento em questão em um intervalo, que pode ser tempo, volume, área, distância ou outra unidade similar.

É definida por:

$$f(x) = P(X=x) = \frac{\lambda^{x} e^{-\lambda}}{x!}$$

Faça você mesmo

Uma central telefônica tipo PABX recebe uma média de 5 chamadas por minuto. Qual a probabilidade deste PABX não receber nenhuma chamada durante um intervalo de 1 minuto?

Faça valer a pena!

1. Um levantamento na Polícia de uma determinada cidade mostrou a taxa de 0,02 homicídios por dia. Qual é a probabilidade de ocorrer 2 homicídios?

a) 0,0002.
b) 0,0134.

c) 0,0064.
d) 0,0247.
e) 0,0009.

2. Calcule a probabilidade da Polícia da mesma cidade do exercício anterior registrar 1 homicídio por dia:

a) 0,0031.
b) 0,3134.
c) 0,0826.
d) 0,0196.
e) 0,4325.

3. Uma empresa de pintura faz o serviço de forma mecânica e pode gerar algumas imperfeições durante o processo de pintura de peças automotivas. A chance de uma peça ter imperfeições na pintura é de 3%.
Determinado lote de 15 peças será entregue para o cliente. Qual a chance de haver, nesse lote, menos de 2 peças com imperfeições na pintura?

a) 0,17853.
b) 0,33961.
c) 0,24572.
d) 0,51231.
e) 0,92703.

4. Em uma escola, 10% dos estudantes preferem a cor vermelha para o uniforme, em detrimento da a cor azul. Qual é a probabilidade de que se escolhermos 10 estudantes, precisamente 2 preferirão a cor vermelha?

a) 0,3678.
b) 0,1937.
c) 0,2145.
d) 0,1678.
e) 0,1236.

5. A linha de produção de certa indústria farmacêutica tem 0,1% de chance de determinado medicamento ser produzido somente com o excipiente, ou seja, sem nenhum princípio ativo. Qual a probabilidade de que, em uma amostra de 600 medicamentos, mais de 2 deles estejam somente com o excipiente?

a) 2,3%.
b) 1,0%.
c) 3,5%.
d) 3,9%.
e) 4,5%.

6. A produção de móveis em uma indústria é feita dependendo da quantidade de pedidos que eles recebem dos vendedores das filiais através da página da indústria na internet. A taxa média de pedidos é 5 pedidos por hora. Calcule a probabilidade da indústria de móveis receber mais de 2 pedidos no prazo de uma hora. Para o caso de ocorrer um evento que impossibilite a fábrica de atender mais de dois pedidos por hora, avalie se a empresa deve aumentar o número de funcionários nesse período.

7. Qual a probabilidade da mesma indústria de móveis receber, em uma jornada de trabalho de 8 horas, 50 pedidos?

Seção 3.4

Distribuição normal

Diálogo aberto

Na seção anterior, estudamos a Distribuição de Probabilidade de Poisson e a Definição de Distribuição Contínua de Probabilidade. Aprendemos que uma distribuição de Poisson difere de uma distribuição binomial em aspectos fundamentais, como: o tamanho da amostra n e a probabilidade p afetam a Distribuição Binomial, já a Distribuição de Poisson é afetada apenas pela média λ.

A distribuição normal, conhecida também como distribuição gaussiana, é sem dúvida a mais importante distribuição contínua, por isso vamos estudá-la. Essa importância é devida a alguns fatores, entre os quais podemos citar o teorema central do limite, o qual é um resultado fundamental em aplicações práticas e teóricas, pois garante que mesmo que os dados não sejam distribuídos segundo uma normal, a média dos dados converge para uma distribuição normal, conforme o número de dados aumenta. Um exemplo para a distribuição normal é a altura de uma determinada população que em geral segue uma distribuição normal. Entre outras características físicas e sociais há um comportamento gaussiano, ou seja, segue uma distribuição normal. Para que você atinja a competência de fundamento de área, que é conhecer os fundamentos estatísticos básicos necessários à formação profissional, a Distribuição Normal é de extrema importância. O objetivo de aprendizagem desta seção é compreender a Distribuição Normal e sua utilização, para isso vamos entender a teoria aplicada a essas distribuições e alguns exemplos serão desenvolvidos passo a passo.

Você está de férias no exterior e foi conhecer os cassinos da região. Você escolheu um hotel-resort associado a um cassino para passar seus dias de férias e se divertir com os jogos de azar. Com os conceitos de probabilidade, você será capaz de determinar a probabilidade de ganhar na mesa de dados e nos jogos de cartas que são seus preferidos e o caça níquel. Enquanto você estava lá, uma fiscalização chegou para verificar se todas as bolas de uma máquina de jogo estavam conforme a padronização. A padronização diz

que o diâmetro médio de uma amostra de 200 bolas é de 0,502 polegadas com desvio-padrão de 0,005 polegadas. A tolerância máxima permitida do diâmetro é de 0,496 a 0,508 polegadas, caso contrário, serão consideradas com defeito. Supondo que os diâmetros são distribuídos normalmente, qual será o percentual de bolas com defeitos que a fiscalização encontrará?

Reflita

Regra Empírica

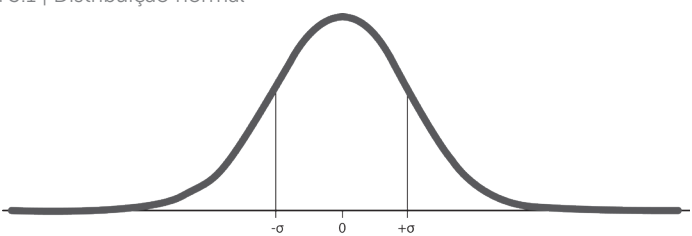
"Se uma variável é distribuída normalmente, então aproximadamente 68% dos dados estarão dentro do intervalo de um desvio-padrão da média; aproximadamente 95% dos dados estarão dentro do intervalo de dois desvios-padrão da média e aproximadamente 99,7% estarão dentro do intervalo de três desvios-padrão da média". (JOHNSON; KUBY, 2013, p. 46)

A distribuição normal é uma das distribuições teóricas mais empregadas. Muitas técnicas estatísticas assumem ou precisam da normalidade dos dados, como ocorre no cálculo da variância.

O teorema do limite central amplia a aplicação da Distribuição Normal e afirma que à medida que o tamanho da amostra aumenta, a distribuição amostral das médias amostrais tende para uma distribuição normal.

Para a distribuição normal, temos o seguinte aspecto gráfico:

Figura 3.1 | Distribuição normal



Fonte: O autor (2015).

A curva normal é um tipo de curva simétrica, suave, cuja forma lembra um sino. Essa distribuição só tem um valor para moda, ou seja, é unimodal, sendo seu ponto de frequência máxima, situado no meio da distribuição, em que a média, a mediana e a moda coincidem.

Assimile

São propriedades da Distribuição Normal:

- A variável aleatória X pode assumir todo e qualquer valor real.

- A representação gráfica é em forma de sino, em torno da média (μ).

- A área total ao redor das abcissas é igual a 1. Esse valor é a probabilidade da variável aleatória X assumir qualquer valor real.

- A probabilidade da variável aleatória X ter valores maiores que a média é igual à probabilidade de ter valores menores que a média (ambas as probabilidades são 0,5).

P (X > μ) = P (X < μ)=0,5.

Para calcular a probabilidade de uma variável aleatória assumir um valor em um determinado intervalo, temos a fórmula:

$$\Pr(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Para facilitar o cálculo dessa integral, foi criada uma metodologia que a reduz a um único caso de qualquer função de distribuição normal N(μ,σ) em uma única função de distribuição normal, em que μ=0 e o desvio-padrão σ=1.

A função de distribuição normal reduzida é designada como N(0,1).

Se o eixo vertical for deslocado para a direita até chegar ao centro, fizemos uma mudança de origem, pois o 0 passou a ocupar o lugar da média. Z é a nova variável que pode ser definida pela fórmula:

$$z = \frac{x - \mu}{\sigma}$$

O Quadro 3.1 a seguir mostra a probabilidade de Z tomar qualquer valor entre a média 0 e um dado valor z. Para construir o quadro foi usada a seguinte equação:

$$\Pr(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

Assim, temos:

$$\Pr(z_1 \leq z \leq z_2) = \int_{z_1}^{z_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}z^2} dz$$

Temos que se X é uma variável aleatória com distribuição normal de média μ e desvio-padrão σ, podemos escrever a probabilidade da seguinte forma:

P (μ < X < x) = P (0 < Z < z)

Aproximação da Binomial pela Normal:

Podemos definir a média pela distribuição binomial como sendo:

$\mu = n \cdot p$

Sendo μ a média procurada, n é o número de repetições do experimento e p é a probabilidade de sucesso no evento. Para a variância podemos ter:

$\sigma^2 = n \cdot p \cdot q$

σ^2 é a variância procurada, n e p têm a mesma definição anterior e q é a probabilidade de fracasso para o evento.

Quadro 3.1 | Distribuição Normal Padrão

P(Z<z)

z	0,0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633

1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014

-3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
-3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
-3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
-3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
-3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
-3,5	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
-3,6	0,0002	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
-3,7	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
-3,8	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
-3,9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Fonte: Johnson; Kuby (2013, p. 332 e 333).

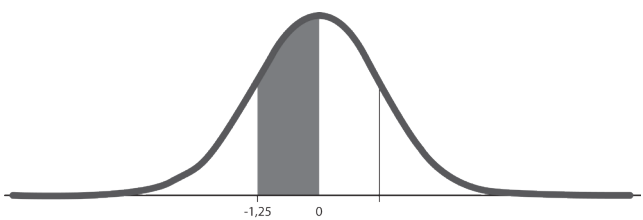
Complemente seus estudos

O Instituto de Matemática e Estatística da USP mostra uma aula sobre a Distribuição normal. Disponível em: <https://www.ime.usp.br/~chang/home/mae116/aulas/Aula%206_distribui%E7%E3o%20Normal.pdf>. Acesso em: 30 jun. 2015.

Exemplificando

Calcular a probabilidade da variável aleatória estar entre -1,25 e 0. A probabilidade corresponde à área destacada na figura 3.2.

Figura 3.2 | Distribuição Normal com variáveis de -1,25 a 0.



Fonte: O autor (2015).

Podemos utilizar o quadro para ver o valor de:

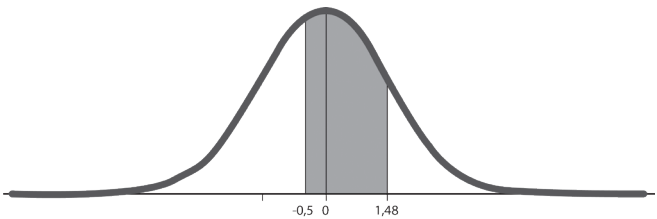
$P(0 < z < 1,25) = 0,3944$

Como a figura é simétrica, a mesma probabilidade de $P(-1,25 < z < 0) = P(0 < z < 1,25) = 0,3944$

Faça você mesmo

Procure a probabilidade correspondente à área destacada na figura 3.3.

Figura 3.3 | Distribuição Normal para variáveis de -0,5 a 1,48.



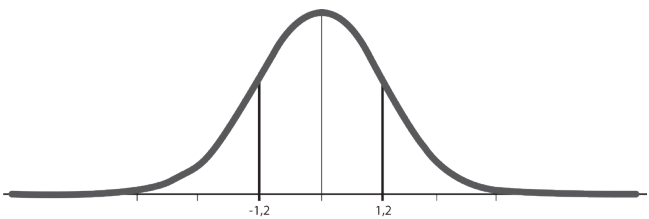
Fonte: O autor (2015).

Sem medo de errar

Enquanto você estava lá no cassino, uma fiscalização chegou para verificar se todas as bolas de uma máquina de jogo estavam conforme a padronização. A padronização diz que o diâmetro médio de uma amostra de 200 bolas é de 0,502 polegadas, com desvio-padrão de 0,005 polegadas. A tolerância máxima permitida do diâmetro é de 0,496 a 0,508 polegadas, caso contrário, serão consideradas com defeito. Supondo que os diâmetros são distribuídos normalmente, qual será o percentual de bolas com defeitos que a fiscalização encontrará?

$$0{,}496\ padronizado = \frac{(0{,}496 - 0{,}502)}{0{,}005} = -1{,}2$$

$$0{,}496\ padronizado = \frac{(0{,}508 - 0{,}502)}{0{,}005} = 1{,}2$$



Proporção de bolas sem defeitos:

= (área sob a curva normal entre z = -1,2 e z = 1,2)

= (duas vezes a área entre z=0 e z = 1,2)

= 2(0,3849) = 0,7698 ou 77%

O percentual de bolas com defeito é 100% - 77% = 23%

Lembre-se

Se X é uma variável aleatória com distribuição normal de média μ e desvio-padrão σ, podemos escrever a probabilidade da seguinte forma:

$P(\mu < X < x) = P(0 < Z < z)$

Avançando na prática

Pratique mais	
Instrução Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.	
"Dieta abaixa o colesterol?"	
1. Competência de fundamento de área	Conhecer os fundamentos estatísticos básicos necessários à formação do profissional.
2. Objetivos de aprendizagem	Compreender a Distribuição Normal e sua utilização
3. Conteúdos relacionados	Distribuição Normal
4. Descrição da SP	Em um hospital especializado em idosos, foi realizada uma pesquisa para determinar a eficiência de uma certa dieta na redução da quantidade de colesterol na corrente sanguínea. As pessoas foram submetidas a uma restrição de alguns alimentos por um intervalo de tempo bastante prolongado. Foram realizados exames nesse grupo de pessoas. A média de pessoas com colesterol considerado alto é 30 e o desvio-padrão é 10. Qual a probabilidade de 50 pessoas terem o colesterol alto?

5. Resolução da SP	$$z = \frac{x - \mu}{\sigma} = \frac{50 - 30}{10} = 2$$ Podemos utilizar a tabela para ver o valor de: P (0 < z < 2) = 0,4772

Faça você mesmo

Em uma fazenda que cria suínos, o peso médio dos animais é de 64kg, e o desvio-padrão é de 15 quilos. Se o peso segue uma distribuição, quantos animais terão seu peso entre 42 e 73 quilos?

Faça valer a pena!

1. Em uma loja de departamentos foi medido o tempo de atendimento aos clientes do setor de televendas. Os atendimentos seguem uma distribuição normal que tem média de 8 minutos e desvio-padrão de 2 minutos. Assinale a alternativa que mostra a probabilidade de um atendimento durar menos que 5 minutos.

a) 0,0324.
b) 0,0668.
c) 0,1458.
d) 0,7841.
e) 0,9451.

2. Em um frigorífico, as peças são processadas automaticamente. O processo segue uma distribuição normal com média de 7 minutos para cada peça e desvio-padrão de 2 minutos. Qual é a probabilidade do processamento durar entre 6 e 9 minutos?

a) 53,28%.
b) 12,78%.
c) 45,39%.
d) 32,87%.
e) 65,58%.

3. Uma enchedora automática de sabão líquido está regulada para que o volume médio de líquido em cada garrafa seja de 1000 cm^3 e desvio padrão

de 10 m^3. O volume segue uma distribuição normal. Qual é a porcentagem de garrafas em que o volume de líquido é menor que 990 cm^3?

a) 78,9%.
b) 98,1%.
c) 15,9%.
d) 41,3%.
e) 33,4%.

4. Uma fábrica de automóveis sabe que o motor de sua fabricação tem duração com distribuição normal com média de 150.000 km e desvio-padrão de 5.000km. Qual a probabilidade de que um carro, escolhido ao acaso dentre os fabricados por essa firma, tenha um motor que dure menos que 170.000 km?

a) 0.
b) 1.
c) 0,5.
d) 0,75.
e) 0,25.

5. O diâmetro de uma tubulação para gás comprimido segue a distribuição normal com média 25,08 pol. e desvio-padrão 0,05 pol. Se as especificações para essas tubulações são 25,00 $\pm$ 0,15 pol., determine o percentual das tubulações a serem fabricadas de acordo com as especificações.

a) 0,1475.
b) 0,2178.
c) 0,3217.
d) 0,9192.
e) 0,0331.

6. Uma indústria química mediu a concentração de um poluente em água liberada e as medições seguem uma distribuição normal, com média de 8 ppm e desvio-padrão de 1,5 ppm. Qual a chance de que, num dado dia, a concentração do poluente exceda o limite regulatório de 10 ppm?

7. Para as mesmas condições do exercício 4, qual será a probabilidade de durar entre 140000 km e 165000 km?

Referências

BARBETTA, P. A.; BORNIA, A. C. R. **Estatística para cursos de engenharia e informática**. 3. ed. São Paulo: Atlas, 2010.

CARVALHO, T. M. de. **Variabilidade espacial de propriedades físico-hídricas de em um latossolo vermelho-amarelo através da geoestatística**. 1991. 84 p. Dissertação (Mestrado) - Escola Superior de Agricultura de Lavras, Universidade Federal de Lavras, Lavras, 1991.

GROSSI SAD, J. H. **Fundamentos sobre variabilidade dos depósitos minerais**. Rio de Janeiro: DNPM/CPRM - GEOSOL, 1986. 141 p.

HINES, W. W. et al. **Probabilidade e estatística na engenharia**. 4. ed. Rio de Janeiro: LTC-Livros Técnicos e Científicos, 2006.

JOHNSON, R.; KUBY, P. **Estatística**. São Paulo: Cengage Learning, 2013.

LAPPONI, J. C. **Estatística usando Excel 5 e 7**. Rio de janeiro: Elsevier, 2005.

LARSON, R.; FARBER, B. **Estatística aplicada.** 4. ed. São Paulo: Pearson, 2010.

MARCONI, M. D. A.; LAKATOS, E. M. **Técnicas de pesquisa**: planejamento e execução de pesquisas, amostragens e técnicas de pesquisas, elaboração, análise e interpretação de dados. 3. ed. São Paulo: Atlas, 1996.

MOORE, D. S. **A estatística básica e sua prática**. 6. ed. Rio de Janeiro: LTC–Livros Técnicos e Científicos, 2014.

MORETTIN, L. G. **Estatística básica: probabilidade e inferência**. São Paulo: Pearson, 2010.

PINHEIRO, J. I. D. **Probabilidade e estatística**. Rio de Janeiro: Elsevier, 2012.

SPIEGEL, M. R. **Estatística**. 3. ed. São Paulo: Makron Books, 1993. 643 p.

SPIEGEL, M. R. **Estatística**. São Paulo: Makron Books, 1996.

WALPOLE, R. E. **Probabilidade e estatística para engenheiros e ciências**. 8. ed. São Paulo: Pearson-Prentice Hall, 2009. v.1.

Probabilidade e Estatística no Excel

Convite ao estudo

Chegamos à Unidade 4 de Probabilidade e essa unidade é uma parte muito importante para construirmos nosso conhecimento em estatística e probabilidade.

A competência de fundamento de área desta disciplina, vamos relembrar, é conhecer os fundamentos estatísticos básicos necessários à formação profissional. Os objetivos de aprendizagem são respectivamente conhecer a estatística descritiva calculada através do software Microsoft Excel; compreender a utilização de algumas funções do programa para os cálculos estatísticos; calcular os modelos de regressão e gráficos de dispersão no Excel e calcular a distribuição normal.

Você é o estagiário de um escritório de contabilidade com 20 funcionários, e seu chefe lhe deixou responsável por automatizar algumas operações que os funcionários do escritório utilizam com frequência.

Algumas operações solicitadas para automatização são: o cálculo das férias dos quatro gerentes da empresa; a média e o desvio-padrão salarial dos funcionários; o gráfico de dispersão salarial e a reta de regressão; e a distribuição normal do tempo que o escritório gasta na época do envio das declarações dos impostos de renda para a Receita Federal.

Seção 4.1

Estatística descritiva no Excel

Diálogo aberto

A competência de fundamento de área desta disciplina, vamos relembrar, é conhecer os fundamentos estatísticos básicos necessários à formação profissional. O objetivo de aprendizagem desta seção é conhecer a estatística descritiva calculada através do software Microsoft Excel.

Você é o estagiário de um escritório de contabilidade com 20 funcionários e seu chefe lhe deixou responsável por automatizar algumas operações que os funcionários do escritório utilizam com frequência.

A primeira operação que seu chefe lhe solicitou foi uma planilha com o cálculo das férias para os gerentes. O cargo de gerente tem o salário bruto de R$ 7.500,00 e você deve calcular o valor líquido, retirando o valor do imposto de renda, que é de 27,5%, e também o INSS, que é de 11%. Sabemos que o valor das férias é correspondente ao salário líquido acrescido de 1/3 do valor.

Não pode faltar

Estatística Descritiva no Excel

Assimile

Vamos apresentar fórmulas para que possamos resolver os cálculos de estatística utilizando o software Excel.

Para iniciar nossos estudos no Microsoft Excel, vamos procurar pelo ícone na barra de programas do seu computador:

Figura 4.1 | Ícone do Excel

Fonte: <https://upload.wikimedia.org/wikipedia/commons/thumb/8/86/Microsoft_Excel_2013_logo.svg/2000px-Microsoft_Excel_2013_logo.svg.png>. Acesso em: 18 ago. 2015.

Após abrir o programa, vamos fazer uma tabela:

Exemplificando

Fazendo uma tabela no Excel

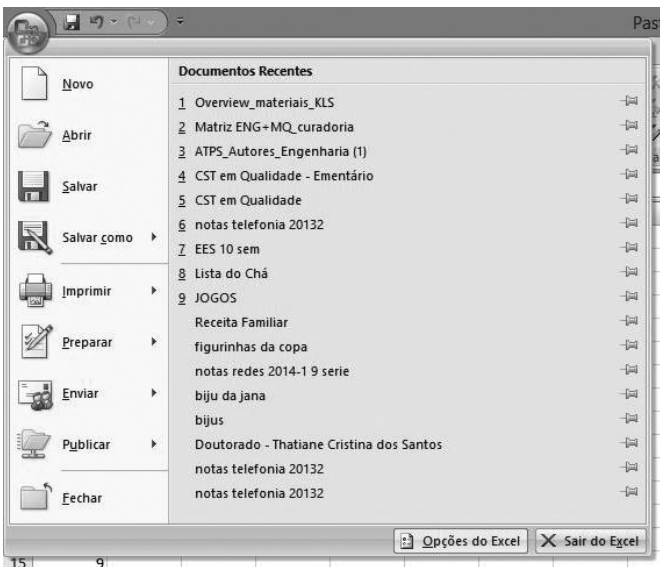
O Excel é uma matriz (com 1 milhão de linhas por 16 mil colunas na versão 2013). Para montar a tabela devemos reproduzir os seguintes passos:

1. Clicar na célula A1 e digitar o texto: **Notas**. Pressione ENTER.

2. Clicar nas células A2 a A16 e digitar os valores das notas. Pressione ENTER.

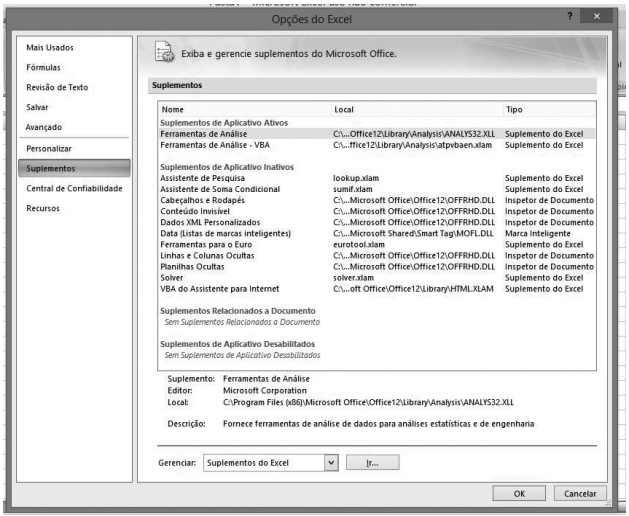
Para utilizar a Estatística descritiva após digitar os dados no Excel, ir no menu opções do Excel e acessar Suplementos, conforme a Figura 4.2.

Figura 4.2 | Acesso ao menu Opções do Excel



Fonte: O autor (2015).

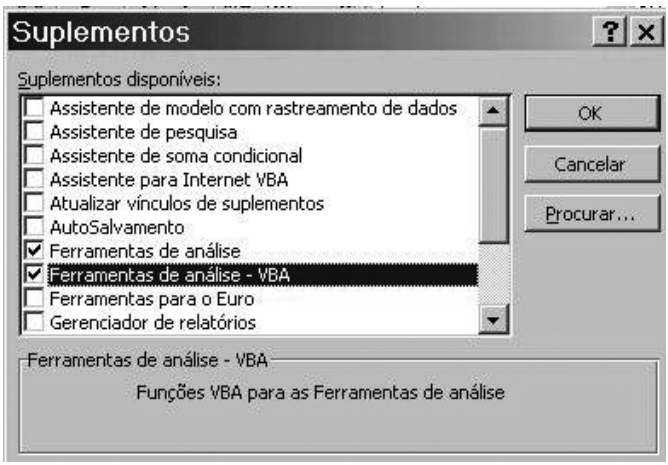
Figura 4.3 | Subitem Suplementos



Fonte: O autor (2015).

O próximo passo é selecionar Ferramentas de Análise e Ferramentas de Análise – VBA, conforme a Figura 4.4.

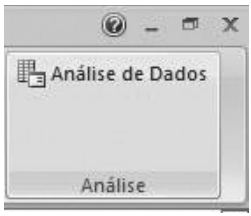
Figura 4.4 | Suplementos do Excel – Ferramentas de análise e Ferramentas de análise VBA



Fonte: O autor (2015).

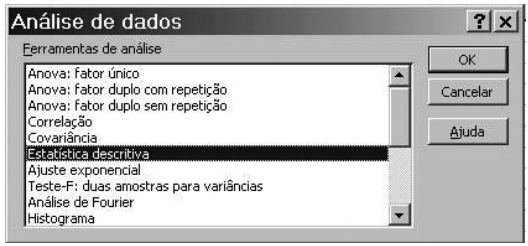
Agora, no Menu Ferramenta, selecione Ferramentas e Análise de Dados; em Análise de Dados, selecione Estatística Descritiva, conforme as Figuras 4.5a e 4.5b.

Figura 4.5a | Menu análise de dados



Fonte: O autor (2015).

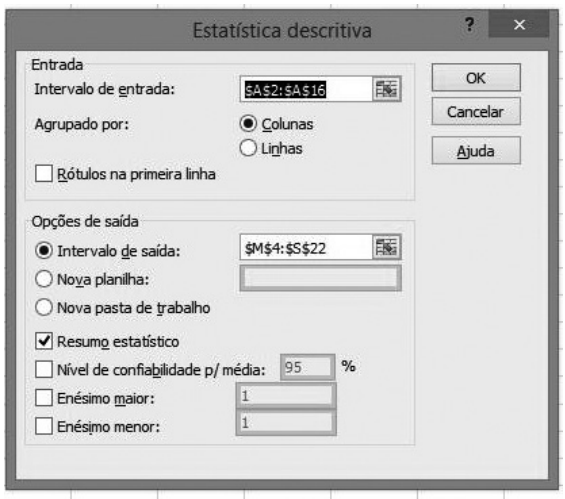
Figura 4.5b | Opção estatística descritiva



Fonte: O autor (2015).

Agora surgirá um quadro em que você deve selecionar os intervalos de entrada e os dados. Selecione o intervalo de saída e o lugar da planilha em que os dados calculados ficaram (os resultados). Marque o resumo estatístico e o intervalo de confiabilidade. No final, clique em OK.

Figura 4.6 | Estatística Descritiva do Excel



Fonte: O autor (2015).

A Tabela 4.1 mostra o resultado de saída através da Estatística Descritiva.

Tabela 4.1 | Estatística Descritiva calculada pelo Excel

Notas	
Média	7,533333
Erro-padrão	0,434979
Mediana	7
Modo	7
Desvio-padrão	1,684665
Variância da amostra	2,838095
Curtose	-0,21137
Assimetria	-0,2732
Intervalo	6
Mínimo	4
Máximo	10
Soma	113
Contagem	15

Fonte: O autor (2015).

Agora vamos construir um histograma para as notas:

Para fazer o histograma, construa o intervalo do bloco.

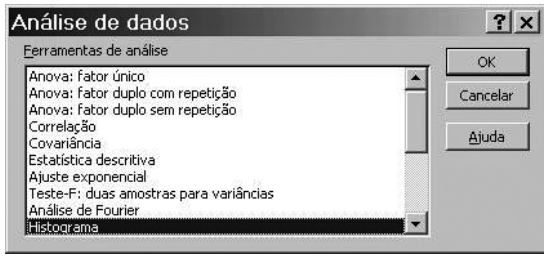
Tabela 4.2 | Dados para construção do histograma

Notas	Intervalo do bloco
9	7
7	8
8	9
6	10
7	
9	
4	
6	
7	
10	
6	
7	
8	

Fonte: O autor (2015).

Clique em Análise de dados e marque Histograma.

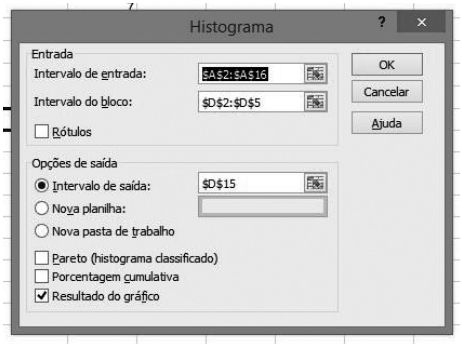
Figura 4.7 | Menu Análise de Dados



Fonte: O autor (2015).

Preencha os itens a seguir e clique em OK.

Figura 4.8 | Menu Histograma



Fonte: O autor (2015).

A saída será:

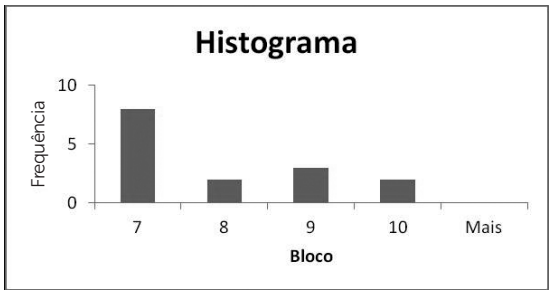
Tabela 4.3 | Saída para o histograma

Bloco	*Frequência*
7	8
8	2
9	3
10	2
Mais	0

Fonte: O autor (2015).

O histograma será:

Figura 4.9 | Histograma das notas



Fonte: O autor (2015).

Reflita

A opção "Análise de Dados" nos permite calcular um conjunto de valores das funções de "Estatística Descritiva" automaticamente.

Complemente seus estudos

Sobre a Estatística Descritiva no Excel e a utilização de outras fórmulas estatísticas sugerimos que você navegue pelo site: <http://biodigital.com.sapo.pt/Aula3_11032005.pdf>.

Faça você mesmo

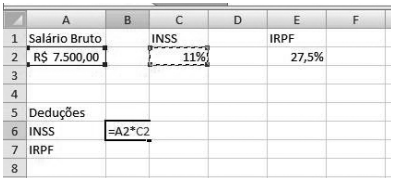
Repita o exemplo que demonstramos na aula de hoje em seu computador. Será importante para você assimilar como utilizamos a ferramenta Análise de Dados > Estatística Descritiva.

Sem medo de errar

A primeira operação que seu chefe lhe solicitou foi a criação de uma planilha com o cálculo das férias para os gerentes. O cargo de gerente tem o salário bruto de R$ 7.500,00, e você deve calcular o valor líquido, retirando o valor do imposto de renda, que é de 27,5%, e também o INSS, que é de 11%. Sabemos que o valor das férias é correspondente ao salário líquido acrescido de 1/3 do valor.

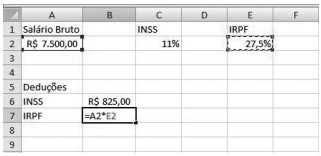
Salário bruto = R$ 7.500,00

Para calcular a dedução do INSS:



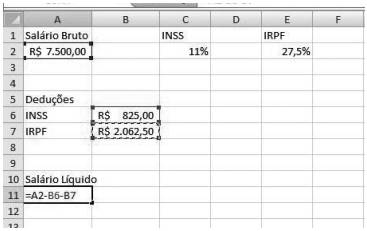
	A	B	C	D	E	F
1	Salário Bruto		INSS		IRPF	
2	R$ 7.500,00		11%		27,5%	
3						
4						
5	Deduções					
6	INSS	=A2*C2				
7	IRPF					
8						

Para calcular a dedução do IRPF:



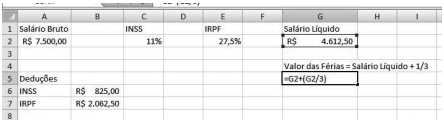
	A	B	C	D	E
1	Salário Bruto		INSS		IRPF
2	R$ 7.500,00		11%		27,5%
3					
4					
5	Deduções				
6	INSS	R$ 825,00			
7	IRPF	=A2*E2			
8					
9					

Para calcular o valor com as deduções:



	A	B	C	D	E	F
1	Salário Bruto		INSS		IRPF	
2	R$ 7.500,00		11%		27,5%	
3						
4						
5	Deduções					
6	INSS	R$ 825,00				
7	IRPF	R$ 2.062,50				
8						
9	Salário Líquido					
10	=A2-B6-B7					
11						
12						

Para cálculo das férias:



	A	B	C	D	E	F	G	H	I
1	Salário Bruto		INSS		IRPF		Salário Líquido		
2	R$ 7.500,00		11%		27,5%		R$ 4.612,50		
3									
4							Valor das Férias = Salário Líquido + 1/3		
5	Deduções						=G2+(G2/3)		
6	INSS	R$ 825,00							
7	IRPF	R$ 2.062,50							
8									

O valor das férias dos gerentes é R$ 6.150,00.

Atenção

As operações podem ser inseridas através do menu FÓRMULAS.

Lembre-se

Pode lhe auxiliar na programação das fórmulas uma breve visita ao site: <http://www.calculoexato.net/calculos-trabalhistas/como-calcular-ferias/>. As tabelas do Imposto de Renda e do INSS estão também nesse site.

Avançando na prática

Pratique mais!

Instrução

Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

<table>
<tr><th colspan="2">"pH do xampu"</th></tr>
<tr><td>1. Competência de fundamentos de área</td><td>Conhecer os fundamentos estatísticos básicos necessários à formação do profissional.</td></tr>
<tr><td>2. Objetivos de aprendizagem</td><td>Conhecer a estatística descritiva calculada através do software Microsoft Excel.</td></tr>
<tr><td>3. Conteúdos relacionados</td><td>Estatística Descritiva no Excel.</td></tr>
<tr><td>4. Descrição da SP</td><td>Foram feitas algumas medições do pH dos xampus de algumas marcas. Essas medições estão dispostas na tabela:
<table>
<tr><th>Medição</th><th>pH</th></tr>
<tr><td>1</td><td>5,12</td></tr>
<tr><td>2</td><td>5,20</td></tr>
<tr><td>3</td><td>5,15</td></tr>
<tr><td>4</td><td>5,17</td></tr>
<tr><td>5</td><td>5,16</td></tr>
<tr><td>6</td><td>5,19</td></tr>
<tr><td>7</td><td>5,15</td></tr>
</table>
São necessários os cálculos estatísticos descritivos para os valores de pH medidos para as amostras. Utilize o recurso da análise de dados.</td></tr>
<tr><td>5. Resolução da SP</td><td>
<table>
<tr><th>pH</th><th></th></tr>
<tr><td>Média</td><td>5,162857</td></tr>
<tr><td>Erro-padrão</td><td>0,010169</td></tr>
<tr><td>Mediana</td><td>5,16</td></tr>
<tr><td>Modo</td><td>5,15</td></tr>
<tr><td>Desvio-padrão</td><td>0,026904</td></tr>
<tr><td>Variância da amostra</td><td>0,000724</td></tr>
<tr><td>Curtose</td><td>-0,16506</td></tr>
<tr><td>Assimetria</td><td>-0,13645</td></tr>
<tr><td>Intervalo</td><td>0,08</td></tr>
<tr><td>Mínimo</td><td>5,12</td></tr>
<tr><td>Máximo</td><td>5,2</td></tr>
<tr><td>Soma</td><td>36,14</td></tr>
<tr><td>Contagem</td><td>7</td></tr>
</table>
</td></tr>
</table>

Faça você mesmo

Pergunte a idade de todas as pessoas que moram com você e, através da ferramenta de análise de dados > estatística descritiva, faça todos os cálculos estatísticos para as idades.

Faça valer a pena!

Com os dados da tabela abaixo, utilize no Excel a ferramenta de Análise de Dados e Estatística Descritiva e responda às questões de 1 a 7:

Uma microempresa tem 10 funcionários com os seguintes salários:

Funcionário	Salário (R$)
1	8.000,00
2	4.200,00
3	4.200,00
4	3.600,00
5	1.500,00
6	1.500,00
7	1.500,00
8	900,00
9	900,00
10	750,00

1. Qual é a média salarial da microempresa?

a) 8.000,00
b) 4.200,00
c) 2.705,00
d) 1.500,00
e) 900,00

2. Qual é o desvio-padrão amostral da empresa?

a) 2.309,45
b) 1.234,78
c) 4.578,63
d) 2.587,96
e) 9.854,65

3. Qual é o valor central de salário, ou seja, a mediana?

a) 8.000,00
b) 4.200,00
c) 2.705,00
d) 1.500,00
e) 900,00

4. Qual é a moda?

a) 8.000,00
b) 4.200,00
c) 2.705,00
d) 1.500,00
e) 900,00

5. Qual é o intervalo entre as amostras?
a) 8.000,00
b) 4.200,00
c) 2.705,00
d) 7.250,00
e) 900,00

6. Qual é o gasto total que a microempresa tem com a folha mensalmente?

7. Qual é o tipo de distribuição assimétrica que temos com a folha salarial apresentada?

Seção 4.2

Funções e pacotes estatísticos no software Excel

Diálogo aberto

A competência de fundamento de área desta disciplina novamente é conhecer os fundamentos estatísticos básicos necessários à formação profissional. O objetivo de aprendizagem desta seção é compreender a utilização de algumas funções do programa para os cálculos estatísticos.

Você é o estagiário de um escritório de contabilidade com 20 funcionários, e seu chefe lhe deixou responsável por automatizar algumas operações que os funcionários do escritório utilizam com frequência.

Agora seu chefe quer que você automatize o cálculo da média e o desvio-padrão salarial dos funcionários.

A tabela 4.4 mostra o valor do salário de cada funcionário:

Tabela 4.4 | Salário dos Funcionários

Funcionário	Salário (R$)		
1	7.500	11	1.150
2	7.500	12	1.150
3	7.500	13	1.000
4	7.500	14	1.000
5	1.500	15	900
6	1.500	16	900
7	1.500	17	830
8	1.200	18	750
9	1.200	19	750
10	1.200	20	750

Fonte: O autor (2015).

Não pode faltar

O Excel possui diversas funções que nos permitem calcular médias.

MÉDIA: Calcula a média aritmética dos números conhecidos.

MÉDIA.GEOMÉTRICA: Os n valores são multiplicados e é extraída a raiz de índice n do produto obtido.

MÉDIA.HARMÔNICA: É o recíproco da média aritmética dos mútuos.

MÉDIA.INTERNA: Calcula a média aritmética de um conjunto de dados, excluindo dos cálculos uma porcentagem dos valores extremos.

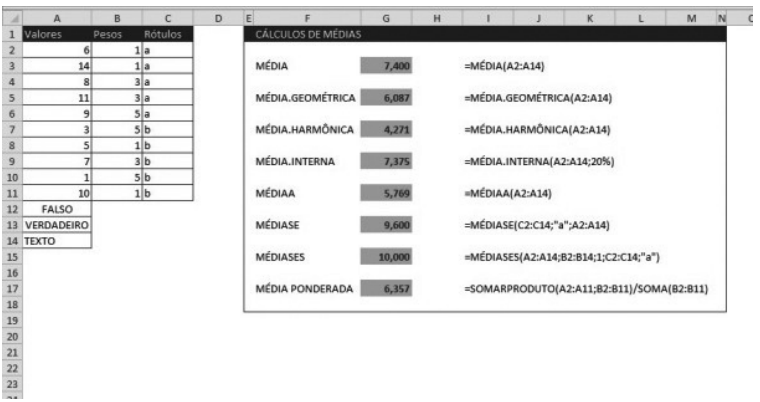
MÉDIAA: Calcula a média aritmética considerando valores de texto (0), valores FALSO (0) e VERDADEIRO (1), caso esses valores estejam presentes no conjunto informado.

MÉDIASE: Similar às funções CONT.SE e SOMASE, esta função calcula a média aritmética de um conjunto de dados com base em um critério informado.

MÉDIASES: Similar à função SOMASES, esta função calcula a média aritmética de um conjunto de dados com base em vários critérios.

MÉDIA PONDERADA: O Excel não possui uma função para cálculo direto da média ponderada, mas esta pode ser facilmente obtida pela combinação das funções SOMARPRODUTO e SOMA. A Figura 4.10 mostra as funções:

Figura 4.10 | Algumas funções do Excel



	A	B	C
1	Valores	Pesos	Rótulos
2	6	1	a
3	14	1	a
4	8	3	a
5	11	3	a
6	9	5	a
7	3	5	b
8	5	1	b
9	7	3	b
10	1	5	b
11	10	1	b
12	FALSO		
13	VERDADEIRO		
14	TEXTO		

CÁLCULOS DE MÉDIAS		
MÉDIA	7,400	=MÉDIA(A2:A14)
MÉDIA.GEOMÉTRICA	6,087	=MÉDIA.GEOMÉTRICA(A2:A14)
MÉDIA.HARMÔNICA	4,271	=MÉDIA.HARMÔNICA(A2:A14)
MÉDIA.INTERNA	7,375	=MÉDIA.INTERNA(A2:A14;20%)
MÉDIAA	5,769	=MÉDIAA(A2:A14)
MÉDIASE	9,600	=MÉDIASE(C2:C14;"a";A2:A14)
MÉDIASES	10,000	=MÉDIASES(A2:A14;B2:B14;1;C2:C14;"a")
MÉDIA PONDERADA	6,357	=SOMARPRODUTO(A2:A11;B2:B11)/SOMA(B2:B11)

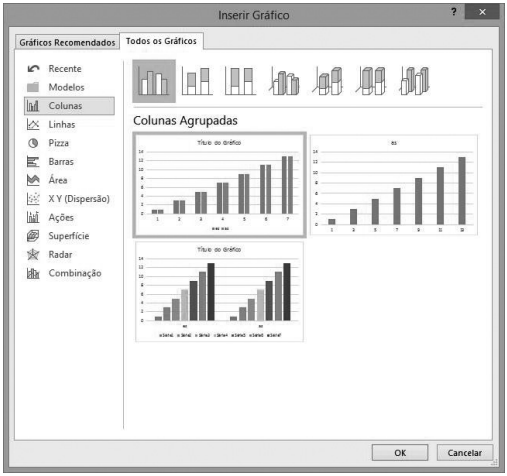
Fonte: usuariodoexcel.wordpress.com

Gráficos

O EXCEL permite que os gráficos sejam construídos de duas maneiras: através do Menu INSERIR, opção GRÁFICO; ou ainda pelo ícone "Assistente de Gráfico", na barra de ferramentas superior

(terceiro ícone à esquerda do controle de zoom), conforme mostra a Figura 4.11:

Figura 4.11 | Assistente de Gráfico

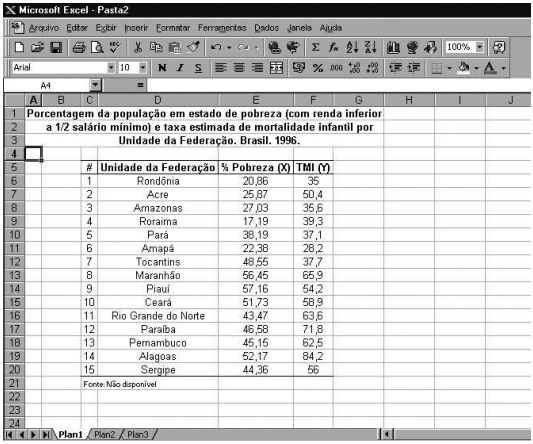


Fonte: O autor (2015).

Você pode selecionar o tipo e o subtipo de gráfico.

Exemplificando

Utilize as informações da Figura 4.12 para construir um gráfico de barras da Taxa de Mortalidade Infantil.



1	Porcentagem da população em estado de pobreza (com renda inferior			
2	a 1/2 salário mínimo) e taxa estimada de mortalidade infantil por			
3	Unidade da Federação. Brasil. 1996.			
4				
5	#	Unidade da Federação	% Pobreza (X)	TMI (Y)
6	1	Rondônia	20,86	35
7	2	Acre	25,87	50,4
8	3	Amazonas	27,03	35,6
9	4	Roraima	17,19	39,3
10	5	Pará	38,19	37,1
11	6	Amapá	22,38	28,2
12	7	Tocantins	48,55	37,7
13	8	Maranhão	56,45	65,9
14	9	Piauí	57,16	54,2
15	10	Ceará	51,73	58,9
16	11	Rio Grande do Norte	43,47	63,6
17	12	Paraíba	46,58	71,8
18	13	Pernambuco	45,15	62,5
19	14	Alagoas	52,17	84,2
20	15	Sergipe	44,36	56
21	Fonte: Não disponível			

Figura 4.11 | Assistente de Gráfico

Fonte: O autor (2015).

Para construir o gráfico de barras da Taxa de Mortalidade Infantil por Unidade da Federação escolha em "Tipo de gráfico" a opção "Barras" e em "Subtipo de gráfico" a opção mais adequada.

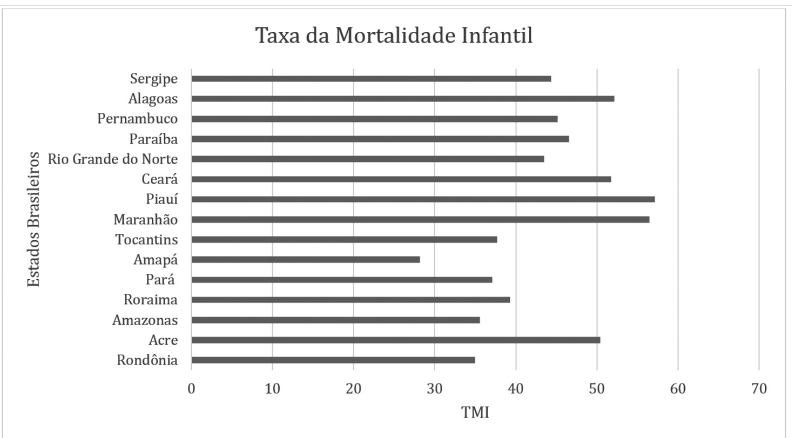
Siga pela opção "Avançar".

Na janela seguinte (etapa 2 do Assistente de gráfico – dados de origem do gráfico) duas informações devem ser fornecidas: Intervalo de dados e Sequência. Observe que o "Intervalo de dados" já está preenchido, porém, nem sempre é o intervalo desejado.

Ainda na etapa 2 do Assistente de gráfico, em Sequência, podemos notar as variáveis selecionadas para os eixos X e Y e nomear a legenda, que está apresentada como "Sequência 1" .

Ao avançar teremos a etapa 3 do Assistente de gráfico. É o passo para a formatação do gráfico: colocar um título e o nome das variáveis alocadas nos eixos horizontal e vertical, inserir legenda, etc., conforme mostra a Figura 4.13.

Figura 4.13 | Gráfico criado



Fonte: O autor (2015).

Assimile

Toda função no Excel, na linha de comando na planilha, deve começar por um sinal de igual.

Complemente seus estudos

O *link* mostra uma apostila com algumas funções do Excel.

Disponível em: <http://www.etepiracicaba.org.br/cursos/apostilas/aplicativos/formulas_excel.pdf>. Acesso em: 2 ago. 2015.

Faça você mesmo

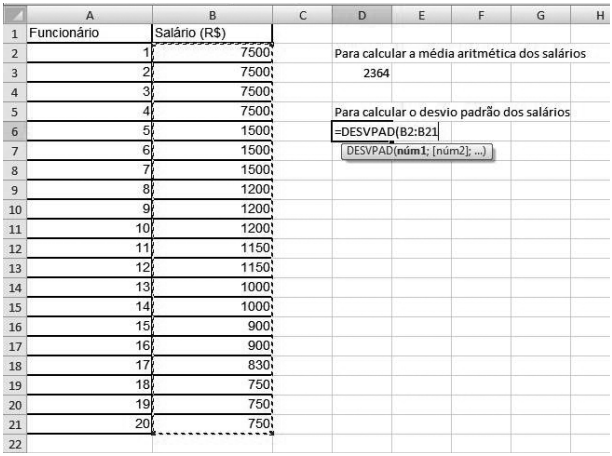
Repita o exercício que apresentamos no "Exemplificando" em seu computador no software Microsoft Excel.

Sem medo de errar

Você é o estagiário de um escritório de contabilidade com 20 funcionários, e seu chefe lhe deixou responsável por automatizar algumas operações que os funcionários do escritório utilizam com frequência. Agora seu chefe quer que você automatize o cálculo da média e o desvio-padrão salarial dos funcionários. A tabela a seguir mostra o valor do salário de cada funcionário:

	A	B	C	D	E	F	G	H
1	Funcionário	Salário (R$)						
2	1	7500		Para calcular a média aritmética dos salários				
3	2	7500		=MÉDIA(B2:B21)				
4	3	7500						
5	4	7500						
6	5	1500						
7	6	1500						
8	7	1500						
9	8	1200						
10	9	1200						
11	10	1200						
12	11	1150						
13	12	1150						
14	13	1000						
15	14	1000						
16	15	900						
17	16	900						
18	17	830						
19	18	750						
20	19	750						
21	20	750						
22								

Média dos Salários = R$ 2.364,00



	A	B	C	D	E	F	G	H
1	Funcionário	Salário (R$)						
2	1	7500		Para calcular a média aritmética dos salários				
3	2	7500		2364				
4	3	7500						
5	4	7500		Para calcular o desvio padrão dos salários				
6	5	1500		=DESVPAD(B2:B21				
7	6	1500		DESVPAD(**núm1**; [núm2]; ...)				
8	7	1500						
9	8	1200						
10	9	1200						
11	10	1200						
12	11	1150						
13	12	1150						
14	13	1000						
15	14	1000						
16	15	900						
17	16	900						
18	17	830						
19	18	750						
20	19	750						
21	20	750						
22								

Desvio-Padrão = R$ 2.645,169

Atenção

A média que calculamos é a Média Aritmética dos valores que colocamos dentro da função:

=MÉDIA(NUM1:NUMX)

Lembre-se

O *link* abaixo pode lhe ser útil nos seus estudos, pois apresenta algumas fórmulas estatísticas e como podemos utilizá-las. Disponível em : <https://www.ime.usp.br/~yambar/MAE116-Quimica/Estatistica%20Usando%20Excel.pdf>. Acesso em: 2 ago. 2015.

Avançando na prática

Pratique mais!

Instrução

Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

<table>
<tr><th colspan="2">"Cavalos de corrida"</th></tr>
<tr><td>1. Competência de fundamentos de área</td><td>Conhecer os fundamentos estatísticos básicos necessários à formação do profissional.</td></tr>
<tr><td>2. Objetivos de aprendizagem</td><td>Compreender a utilização de algumas funções do programa para os cálculos estatísticos.</td></tr>
<tr><td>3. Conteúdos relacionados</td><td>Funções e pacotes estatísticos no software Excel.</td></tr>
<tr><td>4. Descrição da SP</td><td>Um criador de cavalos de corrida pediu ao veterinário e treinador dos cavalos para medir os tempos de cada cavalo da sua criação fazendo o percurso da competição. O veterinário precisava montar uma tabela com os valores medidos e também a média, o desvio-padrão e a variância dos valores medidos.
A tabela com os tempos é:
<table>
<tr><th>Cavalo</th><th>Tempo (seg.)</th></tr>
<tr><td>1</td><td>135</td></tr>
<tr><td>2</td><td>190</td></tr>
<tr><td>3</td><td>145</td></tr>
<tr><td>4</td><td>120</td></tr>
<tr><td>5</td><td>110</td></tr>
<tr><td>6</td><td>184</td></tr>
<tr><td>7</td><td>165</td></tr>
<tr><td>8</td><td>173</td></tr>
</table>
</td></tr>
<tr><td>5. Resolução da SP</td><td>A média pode ser calculada no Excel digitando na barra de fórmulas:
=MÉDIA(B2:B9)
O desvio-padrão pode ser calculado por:
=DESVPAD(B2:B9)
A variância dos tempos:
=VAR(B2:B9)
Os valores encontrados são:
Média = 152,75
Desvio-padrão = 29,75
Variância = 885,64</td></tr>
</table>

Faça você mesmo

Faça um gráfico de barras para os tempos dos cavalos do exercício apresentado no "Avançando na Prática".

Faça valer a pena!

Utilize o Excel para responder às questões de 1 a 7. Os dados são apresentados na Tabela 4.5:

Tabela 4.5 | Pesquisa salarial

	A	B	C	D
1	Entrevistados	Idade	Escolaridade	Salário (R$)
2	1	35	Mestrado	3.200,00
3	2	45	Especialista	1.800,00
4	3	25	Mestrado	3.600,00
5	4	59	Especialista	1.200,00
6	5	29	Especialista	1.500,00
7	6	36	Doutorado	5.600,00
8	7	33	Especialista	2.200,00
9	8	38	Mestrado	3.400,00
10	9	42	Doutorado	6.200,00
11	10	47	Doutorado	6.500,00
12	11	56	Mestrado	4.000,00
13	12	52	Especialista	1.600,00

Fonte: O autor (2015).

1. Para calcular a média dos salários, qual função devemos utilizar?

a) =MÉDIA(D2:D13)
b) =MÉDIA(B2:B13)
c) =MÉDIA(C2:C13)
d) =MÉDIA(A2:D13)
e) =MÉDIA(B2:D13)

2. Para calcular a média das idades, qual função devemos utilizar?

a) =MÉDIA(D2:D13)
b) =MÉDIA(B2:B13)
c) =MÉDIA(C2:C13)
d) =MÉDIA(A2:D13)
e) =MÉDIA(B2:D13)

3. Para calcular a média dos salários dos doutorandos, qual função devemos utilizar?
a) =(A2+A4+A9+A12)/4
b) =(B2+B4+B9+B12)/4
c) =(C2+C4+C9+C12)/4
d) =(D2+D4+D9+D12)/4
e) =(A2+B4+C9+D12)/4

4. Se todos os entrevistados formassem a folha salarial de uma empresa, qual seria o montante gasto com salários por essa empresa?

a) 50.400,00
b) 80.200,00
c) 30.150,00
d) 75.400,00
e) 40.800,00

5. Para criar um gráfico de Idade x Salário, quais são as colunas que devemos selecionar?
a) A e B, respectivamente.
b) C e D, respectivamente.
c) B e D, respectivamente.
d) A e C, respectivamente.
e) A e D, respectivamente.

6. Calcule a média salarial para os Especialistas, Mestres e Doutores a partir da Tabela 4.5:

7. Monte um gráfico de colunas para as médias encontradas:

Seção 4.3

Modelos de regressão e gráficos de dispersão no Excel

Diálogo aberto

Conhecer os fundamentos estatísticos básicos necessários à formação do profissional é a competência de fundamentos de área desta disciplina. O objetivo de aprendizagem desta seção é calcular os modelos de regressão e gráficos de dispersão no Excel.

Você é o estagiário de um escritório de contabilidade com 20 funcionários, e seu chefe lhe deixou responsável por automatizar algumas operações que os funcionários do escritório utilizam com frequência.

Com os dados salariais que seu chefe apresentou, agora é necessário que você crie o gráfico de dispersão salarial e a reta de regressão. As idades também foram apresentadas. Esses dados são importantes para a elaboração do relatório final do estágio, que será exposto para o seu chefe e para o supervisor de estágios.

Tabela 4.6 | Salário dos Funcionários

Idade	Salário (R$)
55	7.500
50	7.500
45	7.500
43	7.500
29	1.500
23	1.500
26	1.500
25	1.200
23	1.200
21	1.200
20	1.150
18	1.150
19	1.000
18	1.000
19	900
18	900
17	830
16	750
16	750
17	750

Fonte: O autor (2015).

Não pode faltar

Utilizaremos o Excel e suas funcionalidades nativas, para determinar os parâmetros de uma regressão linear.

Sendo a regressão linear determinada por uma reta (Y = b + aX), calcularemos:

O coeficiente linear da reta (b)

O coeficiente angular da reta (a)

O coeficiente de determinação (r^2)

Exemplificando

Vamos usar os valores mostrados na Figura 4.14 para desenvolver o exemplo:

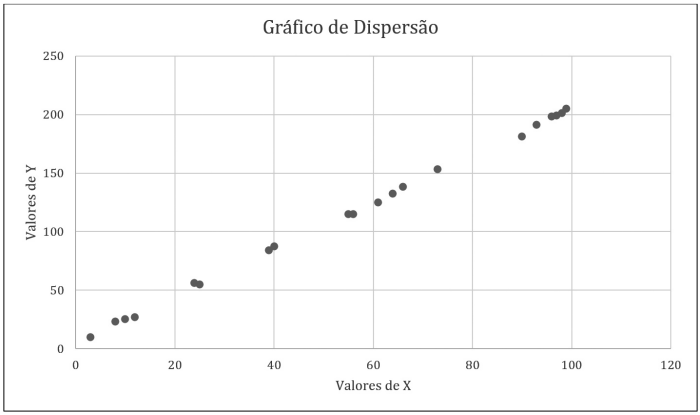
Figura 4.14 | Dados utilizados

	A	B	C
1	x	y	
2	3	10	
3	8	23	
4	10	25	
5	12	27	
6	24	56	
7	25	55	
8	39	84	
9	40	87	
10	55	115	
11	56	115	
12	61	125	
13	64	132	
14	66	138	
15	73	153	
16	90	181	
17	93	191	
18	96	198	
19	97	199	
20	98	201	
21	99	205	
22			

Fonte: O autor (2015).

Para criar o gráfico, vamos usar o tipo gráfico de dispersão, conforme mostra a Figura 4.15:

Figura 4.15 | Gráfico de dispersão



Fonte: O autor (2015).

Ao observamos o gráfico criado, os dados nos mostram que temos uma correlação linear. Os parâmetros da reta e o grau de adequação ao modelo podem ser calculados pelas funções:

INTERCEPÇÃO*(Val_Conhecidos_y;Val_Conhecidos_x)*: Cálculo de b

INCLINAÇÃO*(Val_Conhecidos_y;Val_Conhecidos_x)*: Cálculo de a

RQUAD*(Val_Conhecidos_y;Val_Conhecidos_x)*: Cálculo de r^2

Pelos dados que inserimos no Excel teremos as fórmulas:

=INTERCEPÇÃO($B$2:$B$21;$A$2:$A$21)

=INCLINAÇÃO($B$2:$B$21;$A$2:$A$21)

=RQUAD($B$2:$B$21;$A$2:$A$21)

A Figura 4.16 mostra os cálculos dos coeficientes:

Figura 4.16 | Cálculos dos coeficientes

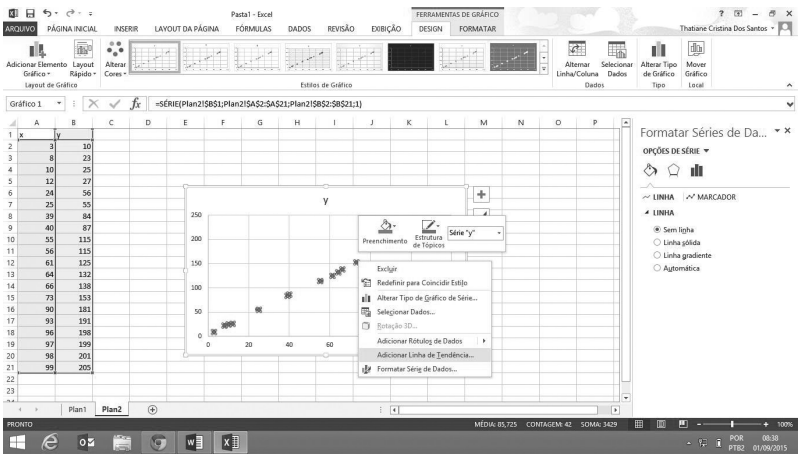
Coeficiente linear	5,3129
Coeficiente angular	1,9962
Coeficiente de determinação	0,9993
Equação da reta	1.9962x + 5,3129

Fonte: O autor (2015).

Também podemos determinar a equação e o valor de r^2 em um dos tipos de gráfico de dispersão.

Para Adicionar Linha de Tendência podemos clicar com o botão direito sobre um dos pontos do gráfico e ir em Adicionar Linha de Tendência, como mostra a Figura 4.17.

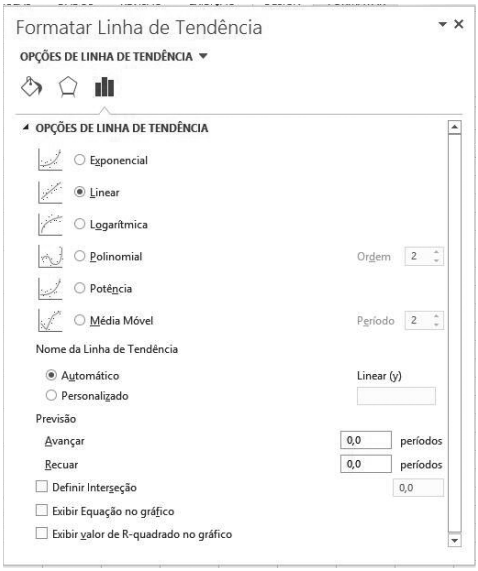
Figura 4.17 | Adicionar linha de tendência



Fonte: O autor (2015).

A Figura 4.18 mostra as opções possíveis para a linha de tendência. Para nosso exemplo, queremos a Linear no gráfico e o valor de r^2. Assim, clique nas duas opções a seguir.

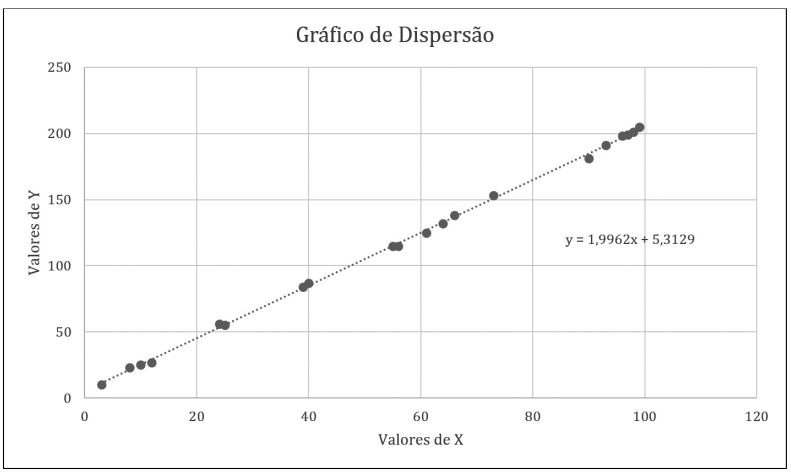
Figura 4.18 | Parâmetros da linha de tendência



Fonte: O autor (2015).

A Figura 4.19 mostra o gráfico com a reta e o valor de r^2.

Figura 4.19 | Reta e parâmetros da regressão



Fonte: O autor (2015).

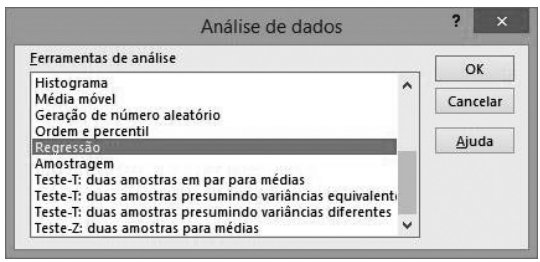
Outra forma de obter a regressão é usando o suplemento ***Análise de Dados***, que vimos anteriormente. A Figura 4.20 mostra o recurso Análise de Dados.

Figura 4.20 | Análise de Dados

Fonte: O autor (2015).

Podemos selecionar em análise de dados a função Regressão, como mostra a Figura 4.21.

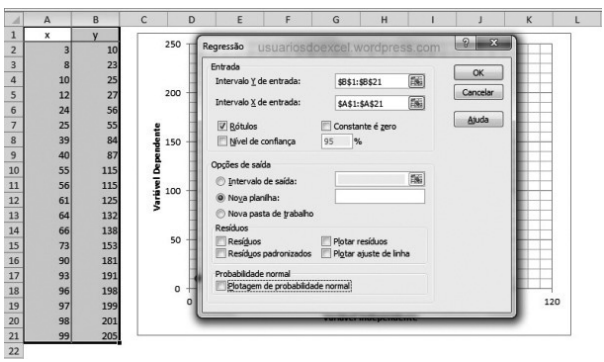
Figura 4.21 | Análise de Dados: Regressão



Fonte: O autor (2015).

Para definir os parâmetros dos cálculos básicos, deve-se usar a análise de resíduos e a adequação à distribuição normal, cálculos importantes para análises estatísticas mais aprofundadas.

Figura 4.22 | Parâmetros para cálculos



Fonte: O autor (2015).

Um relatório com os cálculos é gerado. Os resultados essenciais para a nossa análise foram destacados, como mostra a Figura 4.23.

Figura 4.23 | Relatório gerado

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	*Estatística de regressão*								
4	R múltiplo	0,999666036							
5	R-Quadrado	0,999332183							
6	R-quadrado ajustado	0,999295083							
7	Erro padrão	1,787015521							
8	Observações	20							
9									
10	ANOVA								
11		*gl*	*SQ*	*MQ*	*F*	*F de significação*			
12	Regressão	1	86016,51836	86016,51836	26935,51047	4,9014E-30			
13	Resíduo	18	57,48164053	3,193424474					
14	Total	19	86074						
15									
16		*Coeficientes*	*Erro padrão*	*Stat t*	*valor-P*	*95% inferiores*	*95% superiores*	*Inferior 95,0%*	*Superior 95,0%*
17	Interseção	5,3129437	0,783914218	6,77745546	2,38661E-06	3,666001041	6,959886359	3,666001041	6,959886359
18	x	1,996159717	0,012162775	164,1204146	4,9014E-30	1,970606674	2,021712759	1,970606674	2,021712759
19									

Fonte: O autor (2015).

Complemente seus estudos

No *link* a seguir é apresentado mais um exemplo sobre regressão linear no Excel. <http://www.escolaedti.com.br/regressao-linear-no-excel/>. Acesso em: 2 ago. 2015.

Faça você mesmo

Repita o "Exemplificando" no seu computador. Isso permitirá que você descubra as funções e formas de calcular a regressão linear utilizando o Excel.

Sem medo de errar

Com os dados salariais que seu chefe apresentou, agora é necessário que você crie o gráfico de dispersão salarial e a reta de regressão. Esses dados são importantes para a elaboração do relatório final do estágio, que será exposto para o seu chefe e para o supervisor de estágios.

Utilizando a ferramenta ANÁLISE DE DADOS > REGRESSÃO temos os seguintes cálculos:

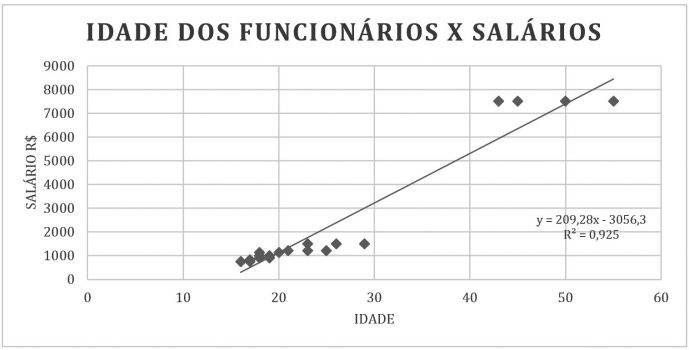
Tabela 4.7 | Cálculos: regressão pela análise de dados

Estatística de regressão	
R múltiplo	0,961776941
R-Quadrado	0,925014884
R-quadrado ajustado	0,920849044
Erro-padrão	744,1860706
Observações	20

Fonte: O autor (2015).

Podemos gerar o gráfico de dispersão, a reta de regressão linear e o coeficiente de determinação (r^2).

Figura 4.24 | Gráfico de Dispersão com a equação da reta e o coeficiente de determinação



Fonte: O autor (2015).

Podemos concluir que a maior parte dos funcionários tem entre 17 e 30 anos. Além disso, os salários mais altos são dos funcionários com mais idade.

Lembre-se

O vídeo a seguir lhe auxiliará a fazer a reta de regressão no Excel! <https://youtu.be/rx8uDzM5UYM>. Acesso em: 4 ago. 2015.

Avançando na prática

Pratique mais!

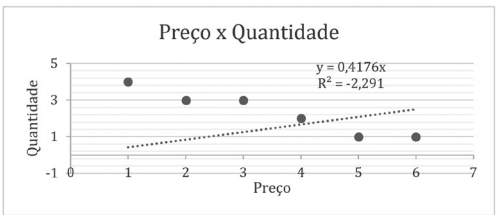
Instrução

Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

<table>
<tr><th colspan="2">“Pesquisa de Preços”</th></tr>
<tr><td>1. Competência de fundamentos de área</td><td>Conhecer os fundamentos estatísticos básicos necessários à formação do profissional.</td></tr>
<tr><td>2. Objetivos de aprendizagem</td><td>Calcular os modelos de regressão e gráficos de dispersão no Excel.</td></tr>
<tr><td>3. Conteúdos relacionados</td><td>Modelos de regressão e gráficos de dispersão no Excel.</td></tr>
<tr><td>4. Descrição da SP</td><td>Uma pesquisa foi realizada por 6 anos e, em relação aos seus resultados, é necessário construir o gráfico de dispersão com a linha de tendência e a devida equação da reta de regressão.

Os dados da pesquisa estão dispostos na tabela 4.8 a seguir:

Tabela 4.8 | Dados da Pesquisa
<table>
<tr><td></td><td>A</td><td>B</td><td>C</td></tr>
<tr><td>1</td><td>ANO</td><td>PREÇO (X)</td><td>QUANT. (Y)</td></tr>
<tr><td>2</td><td>1</td><td>4</td><td>2</td></tr>
<tr><td>3</td><td>2</td><td>6</td><td>1</td></tr>
<tr><td>4</td><td>3</td><td>3</td><td>3</td></tr>
<tr><td>5</td><td>4</td><td>5</td><td>1</td></tr>
<tr><td>6</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>7</td><td>6</td><td>2</td><td>3</td></tr>
</table>
Fonte: O autor (2015).</td></tr>
<tr><td>5. Resolução da SP</td><td>Selecione as células que contêm os dados de X e depois de Y;
Clique em menu inserir e gráfico; no tipo de gráfico, clique em dispersão;
Com o botão direito do mouse sobre o gráfico, selecione “Adicionar linha de tendência” e verifique se na aba tipo a opção linear está ativada.
O resultado está ilustrado na figura 4.25:

Figura 4.25 | Gráfico de dispersão para a pesquisa
 

Fonte: O autor (2015).</td></tr>
</table>

Faça valer a pena

Para as questões de 1 a 7, utilize a tabela 4.9 a seguir:

Uma bióloga observou um verme e seu crescimento. Os dados foram tabulados na tabela:

Tabela 4.9 | Dados de crescimento do verme

Horas de Vida	Peso (g)
1	3
2	6
3	8
4	9
5	12
6	15

Fonte: O autor (2015).

1. Qual é a fórmula para calcular o coeficiente linear da reta (b)?

a) =INTERCEPÇÃO($B$2:$B$7;$A$2:$A$7)
b) =INCLINAÇÃO($B$2:$B$7;$A$2:$A$7)
c) =RQUAD($B$2:$B$7;$A$2:$A$7)
d) =INTERCEPÇÃO($A$2:$A$7;$A$2:$A$7)
e) =INCLINAÇÃO($A$2:$A$7;$A$2:$A$7)

2. Qual é a fórmula para calcular o coeficiente angular da reta (a)?

a) =INTERCEPÇÃO($B$2:$B$7;$A$2:$A$7)
b) =INCLINAÇÃO($B$2:$B$7;$A$2:$A$7)
c) =RQUAD($B$2:$B$7;$A$2:$A$7)
d) =INTERCEPÇÃO($A$2:$A$7;$A$2:$A$7)
e) =INCLINAÇÃO($A$2:$A$7;$A$2:$A$7)

3. Qual é a fórmula para calcular o coeficiente de determinação?

a) =INCLINAÇÃO($A$2:$A$7;$A$2:$A$7)
b) =INTERCEPÇÃO($B$2:$B$7;$A$2:$A$7)
c) =INCLINAÇÃO($B$2:$B$7;$A$2:$A$7)
d) =INTERCEPÇÃO($A$2:a7;$A$2:$A$7)
e) =RQUAD($B$2:$B$7;$A$2:$A$7)

4. Qual é o valor encontrado para o coeficiente linear da reta (b)?

a) 0,8789
b) 0,6789
c) 0,9333
d) 0,9123
e) 0,8458

5. Qual é o valor encontrado para o coeficiente angular da reta (a)?
a) 2,8789
b) 4,7412
c) 3,9123
d) 2,2571
e) 2,8458

6. Qual é a equação da reta de regressão linear (Y = b + aX) para a pesquisa?

7. Faça o gráfico de dispersão com a reta de regressão linear por um dos métodos estudados:

Seção 4.4

Distribuição de probabilidade no Excel

Diálogo aberto

A competência de fundamento de área desta disciplina, vamos relembrar, é conhecer os fundamentos estatísticos básicos necessários à formação do profissional. O objetivo de aprendizagem desta seção é calcular a distribuição normal usando Excel.

Você é o estagiário de um escritório de contabilidade com 20 funcionários, e seu chefe lhe deixou responsável por automatizar algumas operações que os funcionários do escritório utilizam com frequência.

A operação que seu chefe quer automatizar agora é a distribuição normal do tempo que o escritório gasta na época do envio das declarações dos impostos de renda para a Receita Federal. Para as entregas dos impostos de renda o escritório trabalha em média 80 horas por semana, com desvio-padrão de 5 horas e primeiro registro atípico (X) = 92 horas. Calcule os parâmetros Z e a distribuição normal encontrando a área correspondente a valores de Z menores ou iguais a z, ou $P(Z \leq z)$.

Não pode faltar

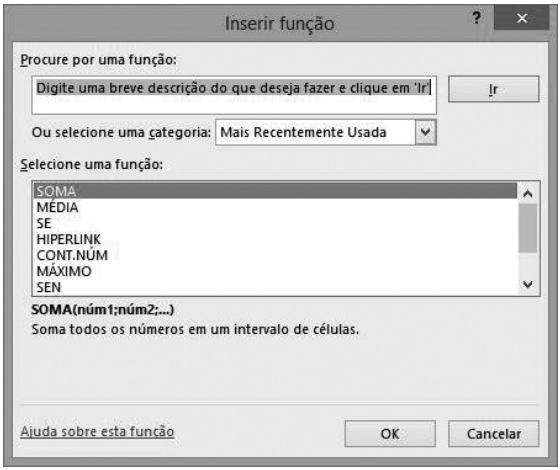
Distribuições de Probabilidade

Para se calcular as Probabilidades Binomiais no EXCEL, utiliza-se a função ***DISTRBINOM***.

Para calcular a Probabilidade Binomial no Excel devemos reproduzir os seguintes passos:

No menu **Fórmulas** selecione a opção **Função**, como na Figura 4.26:

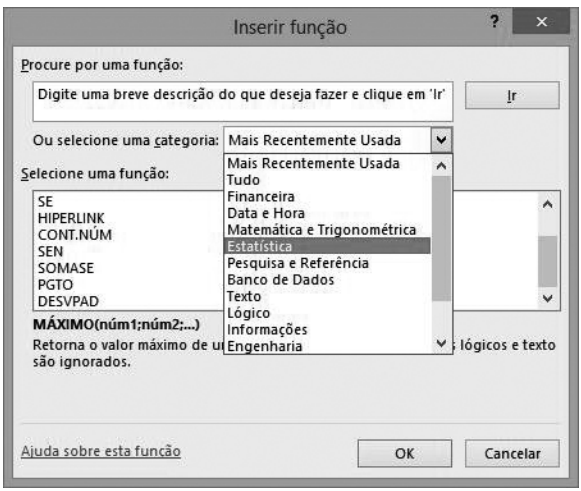
Figura 4.26 | Menu Fórmulas > Funções



Fonte: O autor (2015).

Para calcular a distribuição binomial, escolhemos em "Categoria da função" a opção "Estatística" e em "Nome da função" a opção DISTRBINOM, como mostra a Figura 4.27.

Figura 4.27 | Funções



Fonte: O autor (2015).

Ao clicar em OK, surgirá uma janela com os seguintes parâmetros:

Núm_s = Número de sucessos (neste caso, os valores são X = 0, 1, 2 ou 3);

Tentativas = Número de repetições independentes (n = 3, tamanho da amostra);

Probabilidade_s = Probabilidade de sucesso (p = 1/6 – probabilidade de sair a face 2 em 1 lance);

Cumulativo = *Falso* – para calcular a probabilidade de ter exatamente X sucessos.

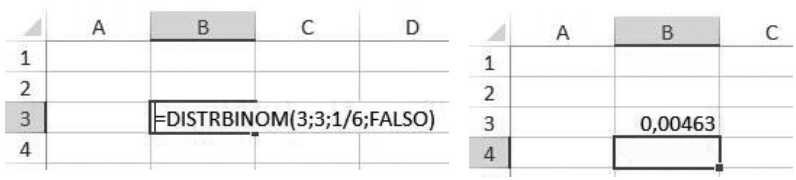
Verdadeiro – para calcular a probabilidade de ter X ou menos sucessos (probabilidade acumulada).

Exemplificando

Um dado é jogado 3 vezes. Qual a probabilidade de sair a face 2?

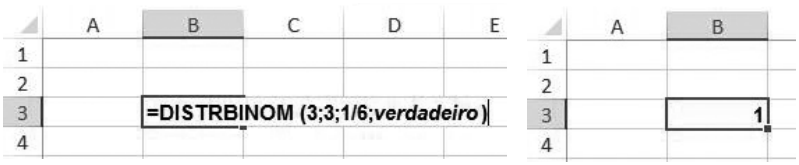
Uma maneira mais fácil de se calcular uma probabilidade binomial é digitando na célula a função e seus valores. Para a probabilidade da face 2 ocorrer 3 vezes, isto é, P(X = 3), podemos digitar na célula desejada: **=DISTRBINOM (3;3;1/6;falso)**, conforme mostra a Figura 4.28. Para a probabilidade da face 2 ocorrer 3 vezes ou menos, isto é, P(X < 3), podemos digitar na célula desejada: **=DISTRBINOM (3;3;1/6;verdadeiro),** conforme mostra a Figura 4.29.

Figura 4.28 | DISTRBINOM (3;3;1/6;falso)



Fonte: O autor (2015).

Figura 4.29 | DISTRBINOM (3;3;1/6;verdadeiro)



Fonte: O autor (2015).

Outras distribuições de probabilidade podem ser calculadas no Excel, tais como:

a) Probabilidades discretas, como a Poisson ou Hipergeométrica (DIST.HIPERGEO).

b) Probabilidades contínuas, como a normal (DIST.NORM) e t-student (DISTT).

Probabilidades Normais

Para se obter probabilidades sob a curva normal no Microsoft Excel, podem ser usadas 5 funções:

A função PADRONIZAR padroniza o valor X ~ normal (μ ,σ) para um valor Z ~ normal (0,1). Calcula o valor de Z, dados o valor de X, a média e o desvio-padrão, sendo que:

$$Z = \frac{X - \mu}{\sigma}$$

Exemplificando

Encontrar o valor de Z correspondente a X = 81, em que μ= 75 e σ= 6. Inicialmente digita-se os valores da média, do desvio-padrão e de X nas células B3, B4 e B5, como na Figura 4.30:

Figura 4.30 | Distribuição de valores

	A	B
1	**Calculando Probabilidades Normais**	
2		
3	**Média Aritmética**	**75**
4	**Desvio Padrão**	**6**
5	**Primeiro valor de X**	**81**
6	**Valor de Z**	

Fonte: O autor (2015).

A função PADRONIZAR encontra o valor de Z. A Figura 4.31 mostra como se padroniza um valor de X digitando a função na célula B6:

Figura 4.31 | Função Padronizar

	A	B
1	**Calculando Probabilidades Normais**	
2		
3	**Média Aritmética**	**75**
4	**Desvio Padrão**	**6**
5	**Primeiro valor de X**	**81**
6	**Valor de Z**	

Fonte: O autor (2015).

Exemplificando

Depois de digitar a função, tecla-se enter e verifica-se que o valor de Z é igual a 1.

A função **DIST.NORMP.N** faz o cálculo da área ou da probabilidade correspondente a um valor menor ou igual a um dado valor de Z (calcula a probabilidade acumulada até Z).

Exemplificando

Agora deseja-se encontrar a área correspondente a valores de Z menores ou iguais a 1, ou P(Z≤1). Assim, digita-se **=DISTR.NORMP.N(B5;B3;B4;VERDADEIRO)**. A Figura 4.32 mostra o valor encontrado para a fórmula:

Figura 4.32 | Valor encontrado para a fórmula DISTR.NORMP.N

B7 = =DIST.NORMP(B6)

	A	B
1	**Calculando Probabilidades Normais**	
2		
3	**Média Aritmética**	**75**
4	**Desvio Padrão**	**6**
5	**Primeiro valor de X**	**81**
6	**Valor de Z**	**1**
7	**P(Z<=1)**	**0,8413447**

Fonte: O autor (2015).

Observa-se que a probabilidade de Z ser menor ou igual a 1 (ou a probabilidade acumulada até 1) é 0,8413.

A função **INV.NORMP.N** faz uma função contrária da função anterior, ou seja, calcula o valor de Z para uma dada probabilidade.

Exemplificando

Precisamos encontrar o valor de z correspondente a uma área acumulada de 0,025, ou seja, P(Z ≤ z) = 0,025.

Para isso, digita-se **=INV.NORMP.N(0,025)** e encontra-se **z = -1,96**.

A função **DIST.NORM.N** é distinta da **DIST.NORMP.N**, pois ela calcula a área ou a probabilidade correspondente a um valor menor ou igual a um dado valor de X, ou seja, calcula a probabilidade acumulada até um valor não padronizado.

Exemplificando

Calcule usando o Excel a P(X ≤ 69), em uma distribuição normal com μ = 75 e σ = 6.

Digita-se =DIST.NORM.N (X ; μ ; σ ; VERDADEIRO), em que a opção "verdadeiro" retorna o valor acumulado, ou seja, a probabilidade de ser menor ou igual a 30. A Figura 4.33 ilustra o que vamos fazer para calcular:

Figura 4.33 | Função DIST.NORM.N

	A	B	C	D	E
1	**Calculando Probabilidades Normais**				
2					
3	**Média Aritmética**	**75**			
4	**Desvio Padrão**	**6**			
5	**Primeiro valor de X**	**81**			
6	**Valor de Z**	**1**			
7	**P(Z<=1)**	**0,8413447**			
8	**Segundo valor de X**	**69**			
9	**P(X<=69)**	**=DIST.NORM(B8;B3;B4;VERDADEIRO)**			
10					

Fonte: O autor (2015).

Temos como resultado P(X ≤ 69) = 0,1586553.

Complemente seus estudos

Para auxiliar no desenvolvimento de seus estudos no Excel, o link a seguir apresenta vídeos sobre os cálculos de distribuição de probabilidade no Excel:

Disponível em: <http://office.cursosguru.com.br/cursos/excel/curso-excel-2010-probabilidades/distribuicoes-de-probabilidade-excel-2010/>. Acesso em: 5 ago. 2015.

Faça você mesmo

Repita em seu computador, no software Microsoft Excel, os exercícios apresentados no "Exemplificando".

Sem medo de errar

A operação que seu chefe quer automatizar agora é a distribuição normal do tempo que o escritório gasta na época do envio das declarações dos impostos de renda para a Receita Federal. Para as entregas dos impostos de renda o escritório trabalha em média 80 horas por semana, com desvio-padrão de 5 horas e primeiro registro atípico (X) = 92 horas. Calcule os parâmetros Z e a distribuição normal encontrando a área correspondente a valores de Z menores ou iguais a z, ou P(Z≤z), acumulativa.

Para o cálculo de Z

=PADRONIZAR(X; MÉDIA; DESVIOPADRÃO)

Para o cálculo de P(Z<=2,4)

=DIST.NORM.N(X; MÉDIA; DESVIOPADRÃO; VERDADEIRO)

Usamos verdadeiro, pois queremos o valor cumulativo para a probabilidade normal.

Os resultados obtidos estão dispostos na tabela abaixo:

Cálculo das Probabilidade Normais	
Média Aritmética	80
Desvio-Padrão	5
Primeiro valor de X	92
Valor de Z	2,4
P(Z<=2,4)	0,991802

Lembre-se

A função PADRONIZAR padroniza o valor X ~ normal (μ, σ) para um valor Z ~ normal (0,1). Calcula o valor de Z, dados o valor de X, a média e o desvio-padrão.

A função DIST.NORMP.N faz o cálculo da área ou da probabilidade correspondente a um valor menor ou igual a um dado valor de Z (calcula a probabilidade acumulada até Z).

A função INV.NORMP.N faz uma função contrária da função anterior, ou seja, calcula o valor de Z para uma dada probabilidade.

A função DIST.NORM.N é distinta da DIST.NORMP.N, pois ela calcula a área ou a probabilidade correspondente a um valor menor ou igual a um dado valor de X.

Avançando na prática

<table>
<tr><th colspan="2">Pratique mais!</th></tr>
<tr><td colspan="2">Instrução
Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.</td></tr>
<tr><td colspan="2">"Pesquisa de Preços"</td></tr>
<tr><td>1. Competência de fundamentos de área</td><td>Conhecer os fundamentos estatísticos básicos necessários à formação do profissional.</td></tr>
<tr><td>2. Objetivos de aprendizagem</td><td>Calcular a distribuição normal usando Excel.</td></tr>
<tr><td>3. Conteúdos relacionados</td><td>Distribuição Normal no Excel.</td></tr>
<tr><td>4. Descrição da SP</td><td>Os dados de uma pesquisa com um atirador de flechas estão dispostos na tabela a seguir:
<table><tr><th colspan="2">Cálculo das Probabilidade Normais</th></tr><tr><td>Média Aritmética de acertos</td><td>30</td></tr><tr><td>Desvio-padrão</td><td>3</td></tr><tr><td>Primeiro valor de X</td><td>35</td></tr></table>
Precisamos padronizar o valor X ~ normal (μ ,σ) para um valor Z ~ normal (0,1). Calcule o valor de Z, dados o valor de X, a média e o desvio-padrão.
Calcule a distribuição normal acumulativa para o Z encontrado.</td></tr>
<tr><td>5. Resolução da SP</td><td>Calculamos Z usando a seguinte fórmula:
=PADRONIZAR(X;MÉDIA;DESVIOPADRÃO)
O valor encontrado foi:
Z=1,6667
Para calcularmos a distribuição normal temos:
=DIST.NORMAL.N(X;MÉDIA;DESVIOPADRÃO,CUMULATIVA)
Vamos colocar verdadeiro para cumulativo.
O valor encontrado foi:
P(Z<=1,6667) = 0,95221</td></tr>
</table>

Faça você mesmo

Repita os cálculos com média de acertos igual a 20, desvio-padrão de 5 e primeiro valor igual a 22.

Faça valer a pena!

Utilize as informações para resolver as questões de 1 a 7:

Uma fábrica de calçados produz por dia 45 pares de calçados, com desvio-padrão de 8 pares e o primeiro valor registrado igual a 48, conforme mostra a tabela abaixo:

		A	B
1	Cálculo das Probabilidade Normais		
2	Média Aritmética	45	
3	Desvio-padrão	8	
4	Primeiro valor de X	48	

1. Qual é o valor de Z?

a) 0,453
b) 0,375
c) 0,244
d) 0,687
e) 0,148

2. Qual é o valor da distribuição normal?

a) 0,64617
b) 0,78412
c) 0,54783
d) 0,14785
e) 0,98745

3. Para calcular o valor de Z utilizamos a função:

a) INV.NORMP.N
b) PADRONIZAR
c) DIST.NORMP.N
d) INV.NORMP.N
e) DIST.NORM.N

4. Quais são os parâmetros utilizados na função da questão 3?
a) (X; MEDIA, DESVPADRÃO)
b) (MEDIA, DESVPADRÃO)
c) (X; DESVPADRÃO)
d) (X; MÉDIA)
e) (MEDIA, DESVPADRÃO)

5. Para calcular a distribuição normal do exercício 2, qual função utilizamos?
a) INV.NORMP.N
b) PADRONIZAR
c) DIST.NORMP.N
d) INV.NORMP.N
e) DIST.NORM.N

6. Qual é a diferença entre as funções DIST.NORMP.N e DIST.NORM.N?

7. Para calcular a probabilidade acumulada até Z, qual é o parâmetro que deve ser alterado na função DIST.NORM.N?

Referências

BARBETTA, P. A.; BORNIA, A. C. R. **Estatística para cursos de engenharia e informática.** 3. ed. São Paulo: Atlas, 2010.

CARVALHO, T. M. **Variabilidade espacial de propriedades físico-hídricas de um latossolo vermelho-amarelo através da geoestatística**. 1991. 84 p. Dissertação (Mestrado) – Escola Superior de Agricultura de Lavras, Universidade Federal de Lavras, Lavras, 1991.

GROSSI SAD, J. H. **Fundamentos sobre variabilidade dos depósitos minerais**. Rio de Janeiro: DNPM/CPRM - GEOSOL, 1986. 141 p.

HINES, W. W. **Probabilidade e estatística na engenharia**. 4. ed. Rio de Janeiro: LTC-Livros Técnicos e Científicos, 2006.

JOHNSON, R.; KOBY, P. **Estatística**. São Paulo: Cengage Learning, 2013.

LAPPONI, J. C. **Estatística usando Excel 5 e 7**. Rio de Janeiro: Elsevier, 2005.

LARSON, R.; FARBER, B. **Estatística aplicada**. 4. ed. São Paulo: Pearson, 2010.

MARCONI, M. D. A.; LAKATOS, E. M. **Técnicas de pesquisa**: planejamento e execução de pesquisas, amostragens e técnicas de pesquisas, elaboração, análise e interpretação de dados. 3. ed. São Paulo: Atlas, 1996.

MOORE, D. S. **A estatística básica e sua prática**. 6. ed. Rio de Janeiro: LTC – Livros Técnicos e Científicos, 2014.

MORETTIN, L. G. **Estatística básica**: probabilidade e inferência. São Paulo: Pearson, 2010.

PINHEIRO, J. I. D. **Probabilidade e estatística**. Rio de Janeiro: Elsevier, 2012.

SPIEGEL, M. R. **Estatística**. 3. ed. São Paulo: Makron Books, 1993. 643 p.

WALPOLE, R. E. **Probabilidade e estatística para engenheiros e ciências**. 8. ed. São Paulo: Pearson-Prentice Hall, 2009. v. 1.

Anotações

Anotações

Anotações

Anotações

Anotações

ISBN 978-85-8482-225-6