Probabilidade e Estatística

Estácio

UNISEB
CENTRO UNIVERSITÁRIO

2015

Editorial

Autora do Original
Valéria Aparecida Ferreira

Sumário

Probabilidade e Estatística

Apresentação

Prezados(as) alunos(as)

Estatística é uma palavra de origem latina, que significou por muito tempo ciência dos negócios do Estado. Ela pode ser vista como uma Matemática Aplicada, uma disciplina da área das ciências exatas que tem aplicação em praticamente todas as áreas de estudo. Esse fato serve para desmistificar o temor vivido pelos alunos com relação ao ensino da matemática em si (aquela que nós aprendemos até o ensino médio). As dificuldades enfrentadas e a falta de conexão com a prática são talvez os fatores que mais contribuem para que este temor ocorra.

No entanto, o ensino da Estatística, mesmo provocando sentimentos semelhantes aos estudantes, proporciona a esses uma visão prática do conteúdo que está sendo abordado. Mais que isso, ele possibilita, a quem o está aplicando, a obtenção de importantes informações do fato que está sendo estudado. O conhecimento mínimo em Estatística se tornou pré-requisito para ler um jornal ou uma revista conceituada, pois muitas informações se encontram resumidas em tabelas ou gráficos que grande parte da população não tem condições de interpretar e por isso ignoram (ou não entendem) reportagens importantes para a formação de uma pessoa esclarecida social, econômica e politicamente.

Procuramos, aqui, apresentar a Estatística de forma clara e prática. Não com o intuito de formar especialistas nessa área, mas sim de proporcionar a você, futuro gestor, uma compreensão dos elementos básicos que compõem essa ciência, visando a aplicação na sua área de atuação. Não tivemos a intenção de esgotar o assunto, mas sim de apresentar os elementos necessários para que você realize uma leitura satisfatória da realidade que o cerca e das informações que têm a sua volta.

Estudaremos duas áreas da Estatística: Estatística Descritiva e a Probabilidade. No Capítulo 1 apresentaremos os conceitos básicos da Estatística bem como descrição de técnicas para organização e apresentação dos dados.

Nos Capítulos 2 e 3 aprenderemos a calcular e interpretar as medidas de posição e dispersão. E, introduziremos, nos Capítulos 4 e 5, conceitos de Probabilidade. Abordaremos o cálculo de probabilidades através do método clássico e frequencial e estudaremos a distribuição de probabilidade Binomial.
Muitos dos exemplos aqui apresentados são hipotéticos. São exemplos de situações que ocorrem de forma semelhante na realidade, mas os dados apresentados não são reais, foram criados apenas para ilustrar a aplicação do conteúdo apresentado.

Capítulo 1

Introdução à Estatística: Análise exploratória de dados

Neste primeiro capítulo, apresentaremos alguns conceitos básicos utilizados pela Estatística, além de fornecer recursos de organização, resumo e apresentação de dados através de tabelas e gráficos.

Quando realizamos uma coleta de dados, geralmente estamos lidando com uma quantidade muito grande de informações. Portanto, torna-se imprescindível a utilização de certas técnicas visando simplificar a leitura de tais informações. Para que se tenha uma visão do todo (sobre o fenômeno que está sendo estudado) precisamos, por exemplo, dispor as informações em tabelas ou apresentá-las em gráficos. É o que estaremos abordando num primeiro momento. Logicamente, há mais técnicas que podem ser aplicadas, mas elas serão vistas nos próximos capítulos.

Objetivos da sua aprendizagem

Após o estudo dos conceitos e técnicas apresentados neste capítulo, esperamos que você consiga identificar os diferentes tipos de variáveis que podem estar presentes em uma pesquisa, bem como, organizar, resumir e apresentar, através de tabelas e gráficos de frequências, as informações contidas em grandes conjuntos de dados.

Você se lembra?

Você se lembra de já ter visto tabelas em jornais, livros ou revistas, em que eram utilizados percentuais para indicar as frequências de ocorrências de respostas em uma pesquisa? Ou com os percentuais referentes à avaliação de um governo? Neste capítulo, veremos como (e para quê) construir tabelas dessa natureza, além de elaborar gráficos que representam os resultados dessas tabelas.

1.1 Breve histórico

O interesse por levantamento de dados não é algo que surgiu somente nos dias atuais. Há indícios de que 3000 anos A.C. já se faziam censos na Babilônia, China e Egito. Havia interesse dos governantes das grandes civilizações antigas por informações sobre suas populações e riquezas. Usualmente estas informações eram utilizadas para taxação de impostos e alistamento militar.

A palavra Estatística surgiu, pela primeira vez, no séc. XVIII. Alguns autores atribuem esta origem ao alemão Gottfried Achemmel (1719-1772), que teria utilizado pela primeira vez o vocábulo Estatística, em 1746.

Na sua origem, a Estatística estava ligada ao Estado. Na atualidade, a Estatística não se limita apenas ao estudo de dados demográficos e econômicos. Ela é empregada em praticamente todas as áreas do conhecimento, sempre que estiverem envolvidas coleta e análise de dados.

Conexão:
Para saber um pouco sobre a evolução histórica da Estatística, assista ao vídeo "História da Estatística" produzido pela Fundação Universidade de Tocantins, disponível em: http://www.youtube.com/watch?v=jCzMPL7Ub2k&feature=related

1.2 Definição de Estatística

A Estatística é uma ciência que trata de métodos científicos para a coleta, organização, descrição, análise e interpretação (conclusão) de um conjunto de dados, visando a tomada de decisões.

Podemos dividir a aplicação da Estatística basicamente em três etapas, que são descritas resumidamente a seguir:

1. Refere-se à coleta de dados, na qual devemos utilizar técnicas estatísticas que garantirão uma amostra representativa da população.

2. Depois dos dados coletados, devemos resumi-los em tabelas de frequências e/ou gráficos e, posteriormente, encontrar as medidas de posição e variabilidade (quantidades). Esta etapa também é conhecida como Estatística Descritiva ou Dedutiva.

3. Esta etapa envolve a escolha de um possível modelo que explique o comportamento dos dados para posteriormente se fazer a inferência dos dados para a população de interesse. Esta etapa também é chamada de Estatística Inferencial ou Indutiva. Nesta etapa, se faz necessário um conhecimento mais aprofundado, principalmente no que se refere aos tópicos de probabilidades.

A probabilidade fornece métodos para quantificar a incerteza existente em determinada situação, usando ora um número ora uma função matemática.

Podemos citar inúmeros exemplos da Estatística em várias áreas do conhecimento, mas só para convencê-lo da importância das técnicas estatísticas, vamos dar alguns exemplos:

1. Se estamos interessados em abrir um supermercado em um determinado local precisamos saber se fatores como sexo, grau de escolaridade, idade, estado civil, renda familiar, entre outros, interferem na abertura deste supermercado e os tipos de produtos que devem ser priorizados nesse estabelecimento, além de definir as estratégias de marketing mais eficientes.

2. Uma empresa, quando está interessada em lançar um novo produto no mercado, precisa saber as preferências dos consumidores. Para isso, é necessário realizar uma pesquisa de mercado.

3. O gestor precisa saber escolher uma amostra representativa de uma população de interesse para não perder muito tempo e, consequentemente, dinheiro da empresa em que trabalha.

4. Para se lançar um novo medicamento no mercado farmacêutico, é necessário a realização de várias experiências. O medicamento deve ser testado estatísticamente quanto à sua eficiência no tratamento a que se destina e quanto aos efeitos colaterais que pode causar, antes de ser lançado no mercado.

5. Para uma empresa, é muito importante fazer previsões de demanda de seus produtos. Para isto existem várias técnicas estatísticas como regressão linear, regressão logística, análise de séries temporais, etc.
6. Controles estatísticos de qualidade (ou controles estatísticos do processo) são indispensáveis em todos os tipos de empresas. Eles são realizados através de um conjunto de técnicas estatísticas, geralmente aplicadas por engenheiros de produção e administradores, para garantir o nível de qualidade exigido para a produção (ou serviço) dentro de uma indústria.

São inúmeras e diversificadas as aplicações de técnicas estatísticas que o gestor pode utilizar. Não conseguiremos falar sobre todas elas, mas apresentaremos os principais conceitos e técnicas que quando utilizados podem auxiliar na tomada de decisões.

Começaremos por apresentar alguns conceitos elementares bastante utilizados no processo estatístico.

População: é o conjunto total de elementos (objetos, itens, medidas, etc.) que têm determinada característica que se deseja estudar.

Amostra: é uma parte da população de interesse que se tem acesso para desenvolver o estudo estatístico. Se a amostra não for fornecida no estudo, devemos retirá-la da população através de técnicas de amostragem adequadas, para que os resultados fornecidos sejam confiáveis.

Estatística Descritiva: é a parte da estatística que trata da organização e do resumo do conjunto de dados por meio de gráficos, tabelas e medidas descritivas (quantidades).

Estatística Indutiva: é a parte que se destina a encontrar métodos para tirar conclusões (ou tomar decisões) sobre a população de interesse, geralmente, baseado em informações retiradas de uma amostra desta população.

Variável: é a característica de interesse no estudo. Vamos estudar dois tipos de variáveis: quantitativas e qualitativas.

Variáveis quantitativas: são aquelas cujas respostas da variável são expressas por números (quantidades). Podemos distinguir dois tipos de variáveis quantitativas: quantitativa contínua e discreta.

Variáveis quantitativas contínuas: são aquelas que podem assumir, teoricamente, infinitos valores entre dois limites (num intervalo), ou seja, podem assumir valores não inteiros. Por exemplo: altura (em metros) de alunos de uma determinada faixa etária, peso (em kg), salário, etc.

Variáveis quantitativas discretas: são aquelas que só podem assumir valores inteiros. Por exemplo: número de filhos por casal, número de livros em uma biblioteca, número de carros vendidos, etc.

Variáveis qualitativas: são as variáveis cujas respostas são expressas por um atributo. Podemos distinguir dois tipos de variáveis qualitativas: nominal e ordinal.

Variáveis qualitativas nominais: definem-se como aquelas em que as respostas são expressas por um atributo (nome) e esse atributo não pode ser ordenado. Por exemplo: tipo sanguíneo, religião, estado civil, etc.

Variáveis qualitativas ordinais: têm suas respostas expressas por um atributo (nome) e esse atributo pode ser ordenado. Por exemplo: grau de instrução, classe social, etc.

1.3 Distribuição de frequências

Para entendermos a ideia de distribuição de frequências, vamos analisar a seguinte situação: quando um pesquisador termina de coletar os dados para sua pesquisa, geralmente fica com muitos questionários em mãos (respondidos pelas pessoas que foram sorteadas para pertencer a sua amostra) ou com os dados digitados em alguma planilha eletrônica. O fato é que os dados "brutos" (sem tratamento) não trazem as informações de forma clara, por isso devemos tabular esses dados. Quando tabulamos os dados estamos resumindo as informações para melhor compreensão da variável em estudo. A esta tabulação damos o nome de distribuição de frequências (ou tabela de frequências).

Distribuição de frequências é uma tabela em que se resumem grandes quantidades de dados, determinando o número de vezes que cada dado ocorre (frequência) e a porcentagem com que aparece (frequência relativa).

Para facilitar a contagem do número de vezes que cada dado ocorre, podemos ordenar os dados. A uma sequência ordenada (crescente ou decrescente) de dados brutos damos o nome de Rol.

Os tipos de frequências com os quais iremos trabalhar são:

Frequência absoluta ou simplesmente frequência (f): é o nº de vezes que cada dado aparece na pesquisa.

Frequência relativa ou percentual (fr): é o quociente da frequência absoluta pelo número total de dados. Esta frequência pode ser expressa em porcentagem. O valor de (fr x100) é definido como fr (%).

Frequência acumulada (fa): é a soma de cada frequência com as que lhe são anteriores na distribuição.

Frequência relativa acumulada (f_{ra}): é o quociente da frequência acumulada pelo número total de dados. Esta frequência também pode ser expressa em porcentagem. O valor de (f_{ra} x100) é definido como f_{ra} (%).

Exemplo 1.1: Dada a tabela abaixo, vamos definir qual é a variável em estudo e qual o tipo de variável. Depois, completaremos a tabela de distribuição de frequências encontrando a frequência relativa (%).

Faixa de renda (em salários mínimos)	Número de operários (f)	Frequência relativa (%) (fr)
0 ├─ 2	43	39,09
2 ├─ 4	39	35,45
4 ├─ 6	16	14,55
6 ├─ 8	8	7,27
8 ├─ 10	4	3,64
Total	**110**	**100**

Tabela 1.1 – Distribuição de renda de operários de uma determinada empresa.

Em todos os nossos exemplos, na distribuição de frequências construída com intervalos de classes, vamos considerar que o intervalo de classe é fechado à esquerda e aberto à direita. Por exemplo, no caso dessa tabela, considerando a terceira classe de frequência, podemos dizer que os 16 operários que estão nesta classe recebem de 4 a menos que 6 salários mínimos por mês.

Resolução

A variável em estudo é a renda dos operários de uma determinada empresa. Esta variável é classificada como quantitativa contínua, pois pode assumir qualquer valor dentro de um intervalo numérico.

As frequências absolutas (f) são fornecidas no problema. As frequências relativas (f_r(%)) são encontradas dividindo cada frequência absoluta (de cada classe de frequência) pelo total de operários (110) e multiplicando por 100.

Uma distribuição de frequências apresenta, basicamente, as 3 colunas apresentadas na tabela 1.1. Desta maneira, conseguimos organizar de forma resumida um conjunto de dados.

Em alguns estudos podemos ter interesse em outras quantidades relacionadas à tabela, como, por exemplo, a frequência acumulada ou a frequência acumulada (%). Veremos mais adiante que a frequência acumulada é utilizada na construção de um gráfico denominado Ogiva. A tabela 1.2 apresenta a frequência acumulada e a frequência relativa acumulada (%).

Faixa de renda (em salários mínimos)	Número de operários (f)	f_r(%)	Frequência acumulada (f_a)	f_{ra} (%)
0 ⊢ 2	43	39,09	43	39,09
2 ⊢ 4	39	35,45	82	75,55
4 ⊢ 6	16	14,55	98	89,09
6 ⊢ 8	8	7,27	106	96,36
8 ⊢ 10	4	3,64	110	100,00
Total	**110**	**100**		

Tabela 1.2 – Distribuição das frequências acumuladas da variável faixa de renda.

A coluna frequência acumulada (f_a) de cada classe é obtida somando a frequência da respectiva classe com as que lhe são anteriores e a f_{ra} (%) é obtida dividindo a f_a pelo número total de dados e multiplicando por 100.

Exemplo 1.2: Uma determinada empresa resolveu traçar o perfil socioeconômico de seus empregados. Uma das variáveis estudadas foi o número de filhos, com idade inferior a 18 anos, de cada um dos empregados. A tabela 1.3 fornece a frequência e a frequência relativa (%) para cada valor obtido.

Número de filhos	Número de operários (f)	Fr (%)
0	6	13,33
1	11	24,44
2	13	28,89
3	7	15,56
4	5	11,11

Número de filhos	Número de operários (f)	Fr (%)
5	1	2,22
6	2	4,44
Total	45	100,00

Tabela 1.3 – Distribuição de frequências dos empregrados, segundo o número de filhos.

Para encontrarmos a f_a e a f_{ra} (%) seguimos o mesmo procedimento que foi utilizado na Tabela 1.2.

1.3.1 Agrupamento em classes

Como vimos no exemplo 1.1, para representar a variável contínua "renda", organizamos os dados em classes. Portanto, podemos dizer que a variável renda foi dividida em "5 classes de frequências".

Quando agrupamos em classes de frequências perdemos informações, pois não sabemos exatamente quais são os valores que estão contidos em cada uma das classes (a não ser que seja possível pesquisar esta informação no conjunto de dados brutos). Na análise das tabelas de frequências com intervalos de classes podemos identificar os seguintes valores:

Limite inferior (L_i): é o menor valor que a variável pode assumir em uma classe de frequência.

Limite superior (L_s): serve de limite para estabelecer qual o maior valor que a variável pode assumir em uma classe de frequência, mas, geralmente, os valores iguais ao limite superior não são computados naquela classe e sim na seguinte.

Ponto médio (P_m): é a média aritmética entre o L_i e o L_s da mesma classe, ou seja, $P_m = \dfrac{L_i + L_s}{2}$.

Amplitude (h): é a diferença entre o L_s e o L_i da classe, ou seja, $h = L_s - L_i$.

Amplitude total (h_t): é a diferença entre o L_S da última classe de frequência e o Li da primeira classe, ou seja: $h_t = L_s - L_i$.

Na construção de uma distribuição de frequências com intervalos de classes devemos determinar o número de classes que uma tabela deve ter e qual o tamanho (ou a amplitude) destas classes. Podemos usar o bom senso e escolher arbitrariamente quantas classes e qual a amplitude que estas classes devem ter. Em algumas situações, iremos tabular dados para comparar os resultados com informações de outras tabelas. Nesse caso, é melhor considerar as mesmas classes das tabelas que iremos comparar.

Quando não tivermos nenhuma referência sobre qual deve ser o número de classes a se trabalhar, podemos utilizar o critério que é sugerido por vários autores. Chama-se regra da raiz e será apresentado a seguir

Considere:

$$k \cong \sqrt{n} \quad \text{e} \quad h \cong \frac{R}{k}$$

onde k é o número de classes que vamos construir na tabela de frequências; n é o tamanho da amostra que estamos trabalhando; h é a amplitude de cada uma das classes e R é a amplitude total dos dados.

> Os valores de k e h devem ser arredondados sempre para o maior valor. Por exemplo, para uma amostra de tamanho n = 50 cujo menor valor é 4 e o maior valor é 445 temos que R = 441 (maior valor – menor valor). O número de classes seria dado por $k \cong \sqrt{n} = \sqrt{50} = 7{,}07106 \approx 8$ (maior inteiro depois de 7) e a amplitude (tamanho) de cada uma das 8 classes acima deverá ser $h \cong \frac{R}{k} = \frac{441}{8} = 55{,}125 \approx 56$ (maior inteiro depois de 55). Ou seja, deveríamos, para este exemplo, montar uma tabela com 8 classes e de amplitude 56. A tabela pode ser iniciada pelo menor valor do conjunto de dados.

Resumindo, para montar uma tabela de frequências com intervalos de classes devemos:

- Achar o mínimo e o máximo dos dados.
- Determinar as classes de frequências, o que na verdade nada mais é do que escolher intervalos de mesmo comprimento que cubra a amplitude entre o mínimo e o máximo. Para determinar o número de classes, usaremos $k \cong \sqrt{n}$ e para determinar o "tamanho" das classes usaremos $h \cong \frac{R}{K}$.

- Contar o número de observações que pertencem a cada intervalo de classe. Esses números são as frequências observadas da classe.
- Calcular as frequências relativas e acumuladas de cada classe.
- De modo geral, a quantidade de classes não deve ser inferior a 5 e nem superior a 25.

Um outro critério utilizado para construir distribuição de frequências com intervalos de classes é a regra de Sturges.

Neste critério, o número de classes a serem construídas é obtido utilizando a seguinte fórmula:

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

onde:
k:número de classes
n:total da amostra
log n: logaritmo na base 10 de n

A amplitude de cada intervalo de classe é obtida por:

$$h = \frac{\text{amplitude total}}{k} = \frac{R}{k}$$

Devemos arredondar o valor de k para o número inteiro mais próximo, pois o número de classes deve ser sempre inteiro. O arredondamento de h deve ser sempre efetuado para cima usando o mesmo número de casas decimais dos elementos da amostra para que nenhum elemento fique fora da tabela.

1.4 Métodos gráficos

O objetivo da utilização de gráficos em análise de dados é o de facilitar a compreensão do fenômeno estatístico por meio do efeito visual imediato que os gráficos proporcionam.

1.4.1 Tipos de gráficos

Existem vários tipos de gráficos. Os mais usados são: gráfico em linhas, diagramas de área (como por exemplo: gráfico em colunas, gráfico em barras e gráfico em setores) e gráficos para representar as distribuições de frequências construídas com intervalos de classes (como por exemplo: polígono de frequências, histograma e ogiva).

Vamos saber um pouco quando usar e como construir cada um destes gráficos.

Conexão:
Vamos refletir um pouco sobre a necessidade de abordagens pedagógicas para o ensino e a aprendizagem de gráficos acessando o endereço http://www.ufrrj.br/emanped/paginas/conteudo_producoes/docs_22/carlos.pdf.

1.4.1.1 Gráfico em linhas

Sempre que os dados estiverem distribuídos segundo uma variável no tempo (meses, anos, etc.), assim como sucede com os dados do exemplo 1.3 – figura 1.1, os dados podem, também, ser descritos através de um gráfico em linhas. Esse tipo de gráfico retrata as mudanças nas quantidades com respeito ao tempo através de uma série de segmentos de reta. É muito eficiente para mostrar possíveis tendências no conjunto de dados.

Exemplo 1.3: A tabela 1.4 fornece uma lista do número de assinantes de telefones celulares, em milhões, de 1997 a 2007, do país X. Construa um gráfico para resumir os dados da tabela a seguir.

Ano	Assinantes (em milhões)
1997	1,1
1998	1,3
1999	1,5
2000	1,9
2001	2,4
2002	2,6
2003	3,1
2004	7,4
2005	18,6
2006	21,5
2007	29

Tabela 1.4 – Assinantes de telefones celulares, em milhões, de 1997 a 2007.

O gráfico que melhor representa este conjunto de dados é o gráfico em linhas, já que os dados se reportam a uma série no tempo (série temporal). O gráfico está ilustrado na figura 1.1.

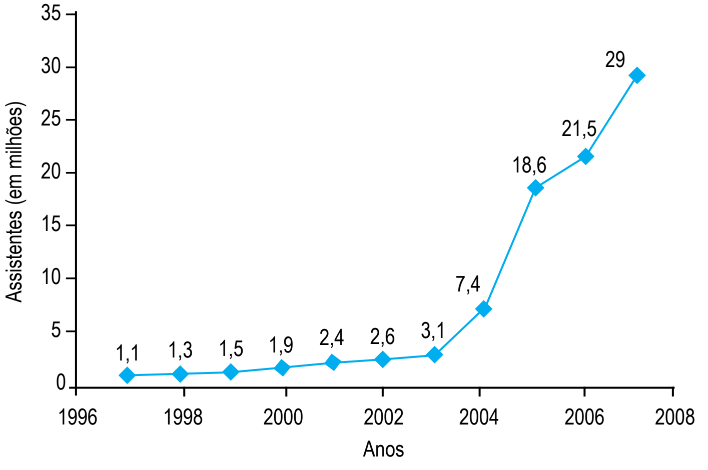


Figura 1.1 – Gráfico em linha para os dados de assinantes de telefones celulares.

1.4.1.2 Gráfico (ou Diagrama) em Barras (ou Colunas)

Os diagramas em barras (ou colunas) são bastante utilizados quando trabalhamos com variáveis qualitativas (dados categóricos). No eixo horizontal especificamos os nomes das categorias e no eixo vertical construímos uma escala com a frequência ou a frequência relativa. As barras terão bases de mesma largura e alturas iguais à frequência ou à frequência relativa. O gráfico em barras, quando as barras estão dispostas no sentido vertical, também é chamado de gráfico em colunas.

Quando construímos o gráfico de barras para variáveis qualitativas e as barras são arranjadas em ordem descendente de altura, a partir da esquerda para a direita, com o atributo que ocorre com maior frequência aparecendo em primeiro lugar, denominamos este gráfico de barras de Diagrama de Pareto.

Exemplo 1.4: Uma grande indústria de materiais de construção, com diversas lojas espalhadas pelo país, fez um levantamento das principais causas de perda de ativos durante o ano de 2007 e as informações estão dispostas na tabela seguinte.

Causas	Valor perdido (milhões de reais)
Má administração	5,2
Roubos de funcionários	3,9
Fraudes nas vendas	5,5
Assaltos às lojas	1,8
Perda do estoque	1,6
Atendimento ruim	0,8

Tabela 1.5 – Causas de perda de ativos durante o ano de 2007.

Graficamente, podemos representar este conjunto de dados de três formas diferentes: gráfico em colunas, gráfico em barras e o gráfico em setores (ou pizza ou circular), que será apresentado no próximo item.

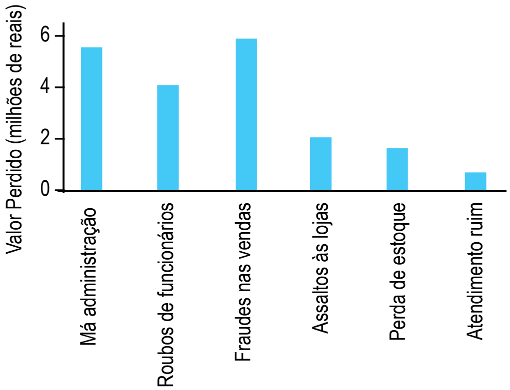


Figura 1.2a – Gráfico em colunas para a variável Causas de perdas de ativos.

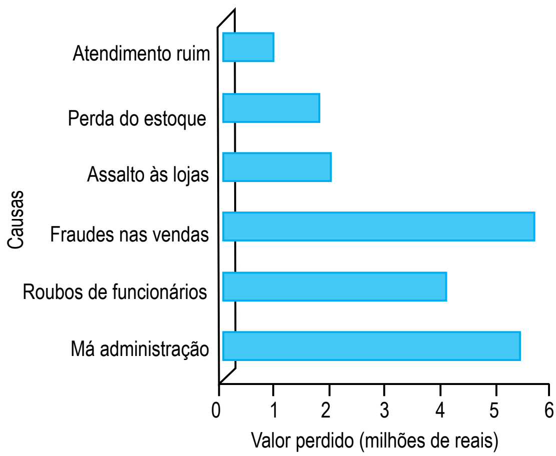


Figura 1.2b – Gráfico em barras para a variável Causas de perdas de ativos.

1.4.1.3 Gráfico (ou Diagrama) em Setores

O diagrama em setores, também conhecido como gráfico de pizza, é um dos gráficos mais utilizados para representar variáveis qualitativas (ou categóricas) e é bastante apropriado quando se deseja visualizar a proporção que cada categoria representa do total.

Vamos utilizar os dados do Exemplo 1.4 para mostrar um gráfico em setores.

Figura 1.3 – Gráfico em setores para a variável Causas de perdas de ativos.

Os gráficos que serão apresentados a seguir são gráficos construídos segundo uma distribuição de frequências com intervalos de classes. São eles: o histograma, o polígono de frequências e a ogiva.

1.4.1.4 Histograma

Um histograma é semelhante ao diagrama de barras, porém refere-se a uma distribuição de frequências para dados quantitativos contínuos. Por isso, apresenta uma diferença: não há espaços entre as barras. Os intervalos de classes são colocados no eixo horizontal enquanto as frequências são colocadas no eixo vertical. As frequências podem ser absolutas ou relativas.

Exemplo 1.5: A tabela a seguir apresenta o salário de funcionários de uma empresa no interior de Minas Gerais.

Salário (R$)	Freq. Absoluta (f)	Freq. Acumulada (fa)
400,00 ⊢ 800,00	38	38
800,00 ⊢1.200,00	18	56
1.200,00 ⊢ 1.600,00	12	68
1.600,00 ⊢ 2.000,00	8	76
2.000,00 ⊢ 2400,00	8	84
2.400,00 ⊢ 2.800,00	5	89
2.800,00 ⊢ 3.200,00	3	92
3.200,00 ⊢ 3.600,00	0	92
3.600,00 ⊢ 4.000,00	2	94
4.000,00 ⊢ 4.400,00	0	94
4.400,00 ⊢ 4.800,00	1	95
Total	95	

Tabela 1.6 – Distribuição de frequências dos salários dos funcionários de uma empresa no interior de Minas Gerais.

Como os dados da tabela 1.6 estão apresentados em intervalos de classes podemos representá-los graficamente através de um histograma ou do polígono de frequências, como mostram as figuras 1.4 e 1.5, respectivamente.

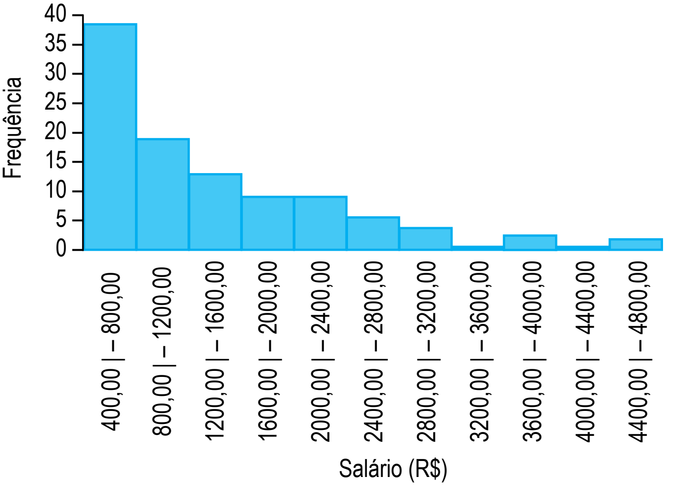


Figura 1.4 – Histograma dos salários dos funcionários de uma empresa no interior de Minas Gerais.

1.4.1.5 Polígono de Frequências

Podemos dizer que o polígono de frequências é um gráfico de linha de uma distribuição de frequências. No eixo horizontal são colocados os pontos médios de cada intervalo de classe e no eixo vertical são colocadas as frequências absolutas ou relativas (como no histograma). Para se obter as intersecções do polígono com o eixo das abscissas, devemos encontrar o ponto médio da classe anterior à primeira e o ponto médio da classe posterior à última.

Conexão:
Para se ter uma ideia da importância da organização dos dados em tabelas de frequências e da construção do histograma, leia "A Estatística na Prática" em: ANDERSON, David R.; SWEENEY, Denis J.; WILLIAMS, Thomas A. Estatística aplicada à administração e economia. São Paulo: Pioneira Thomson Learning, 2003, pp. 37 e 38.

O histograma e o polígono de frequências são gráficos alternativos e contêm a mesma informação. Fica a critério de quem está conduzindo o estudo a escolha de qual deles utilizar.

Considerando os dados do exemplo 1.5, temos o polígono de frequências representado pela figura 1.5.

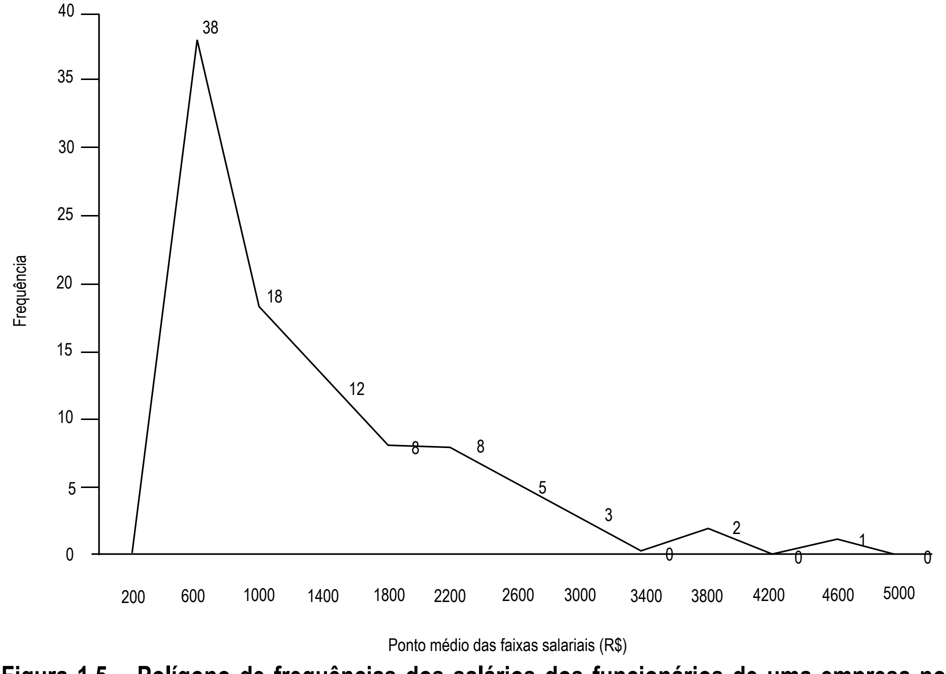


Figura 1.5 – Polígono de frequências dos salários dos funcionários de uma empresa no interior de Minas Gerais.

Para finalizarmos o estudo de gráficos, vamos apresentar um gráfico denominado ogiva.

1.4.1.6 Ogiva

Uma ogiva é um gráfico para uma distribuição de frequências acumuladas. Utilizando o exemplo 1.5, a terceira coluna traz a frequência acumulada dos dados e a ogiva fica representada pela figura 1.6.

> Para construir um gráfico de ogiva, devemos usar o limite superior de cada intervalo no eixo horizontal e a frequência acumulada no eixo vertical. A frequência acumulada relacionada com o limite inferior da primeira classe é sempre zero.

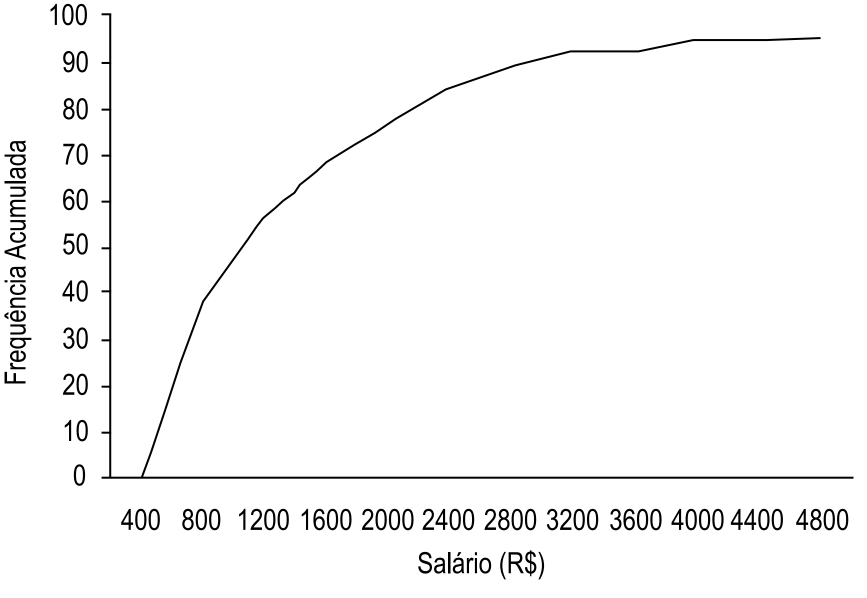


Figura 1.6 – Ogiva dos salários dos funcionários de uma empresa no interior de Minas Gerais.

Atividades

01. Classifique as variáveis a seguir em quantitativas (discretas ou contínuas) ou qualitativas (nominal ou ordinal).

a) Cor dos olhos.
b) Número de peças produzidas por hora.
c) Diâmetro externo.
d) Número de pontos em uma partida de futebol.
e) Produção de algodão.
f) Salários dos executivos de uma empresa.
g) Número de ações negociadas na bolsa de valores.
h) Sexo dos filhos.
i) Tamanho de pregos produzidos por uma máquina.
j) Quantidade de água consumida por uma família em um mês.
k) Grau de escolaridade.
l) Nível social.
m) Tipo sanguíneo.
n) Estado civil.

02. A seguir temos as idades dos funcionários de uma determinada empresa. Fazer uma distribuição de frequências, agrupando os dados em classes. OBS.: A tabela de distribuição de frequências deve ser completa com f, f_r e f_a.

Idades (dados brutos)

48	28	37	26	29	59	27	28	30	40	42	35	23	22	31
21	51	19	27	28	36	25	40	36	49	28	26	27	41	29

Baseado na tabela de frequências construída, responda:

a) Quantos são os funcionários com idade inferior a 33 anos?
b) Que porcentagem de funcionários tem idade igual ou superior a 47 anos?
c) Quantos são os funcionários com idade maior ou igual a 26 anos e não tenham mais que 40 anos?
d) Qual a porcentagem de funcionários com idade abaixo de 40 anos?
e) Qual a porcentagem de funcionários que têm no mínimo 40 anos?

03. Uma agência de turismo está interessada em saber o perfil dos seus clientes com relação à variável estado civil. Para isso, o gerente desta agência pediu ao funcionário do setor de vendas para fazer um gráfico que resuma estas informações. Construa o gráfico e interprete-o.

Estado civil	Número de clientes
Solteiro	2600
Casado	900
Viúvo	345
Separado	1200
Outros	1020
Total	6065

04. Um consultor estava interessado em saber quanto, geralmente, cada pessoa gastava em um determinado supermercado no primeiro sábado após receberem seus pagamentos (salários). Para isso ele entrevistou 50 clientes que passaram pelos caixas entre 13h e 18h, e anotou os valores gastos por cada um deles. Estes valores estão listados a seguir:

4,89	11,00	5,60	73,85	24,83	98,00	186,00	234,87	58,00	198,65
223,86	341,42	94,76	445,76	82,80	35,00	455,00	371,00	398,60	234,00
64,90	54,98	48,80	68,90	120,32	126,98	76,43	6,35	9,98	12,68
243,00	18,65	134,90	11,10	321,09	290,76	74,00	48,80	74,52	138,65
26,00	210,13	15,78	197,45	75,00	76,55	32,78	166,09	105,34	99,10

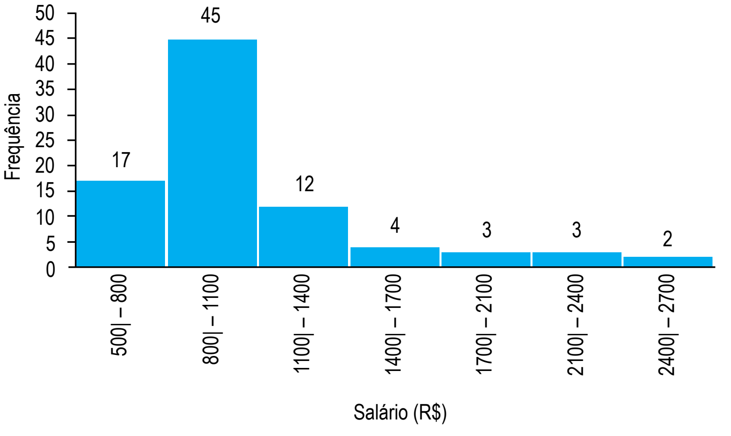
Analisando o conjunto de dados, responda os seguintes itens:

a) Qual é a variável em estudo? Classifique-a.

b) Construa uma tabela de frequências a partir do conjunto de dados brutos.

c) Construa um histograma e um polígono de frequências para a tabela construída no item b).

05. Analise o gráfico a seguir e responda:



a) Qual a variável em estudo? Classifique-a.

b) Quantos funcionários ganham entre R$ 800,00 (inclusive) e R$ 1.100,00 (exclusive)?

c) Qual o número de funcionários total desta empresa?

d) Qual a porcentagem de funcionários que ganham R$ 1.700,00 ou mais?

e) Qual a porcentagem de funcionários que ganham entre R$ 500,00 (inclusive) e não mais que R$ 1.100,00?

f) A partir do histograma, monte uma tabela de distribuição de frequências.

Reflexão

Estamos encerrando nosso primeiro capítulo. Vimos, aqui, alguns conceitos que serão fundamentais na compreensão do restante do conteúdo de Estatística. Já deve ter dado para perceber que, mesmo estando no início da disciplina, as aplicações práticas que você poderá fazer na sua área de atuação serão muitas. A compreensão e interpretação das mais variadas informações, com as quais nos deparamos em nosso cotidiano, dependem, em parte, do conhecimento de certos elementos estatísticos.

Estamos apenas no começo. Muitas técnicas (muito interessantes!) ainda serão abordadas. E lembre-se que o conhecimento e o domínio da Estatística certamente levarão você, futuro gestor, às decisões mais acertadas.

Leitura recomendada

Recomendamos a leitura do texto "Como analisar de forma simples um grande número de dados?", disponível no endereço http://www.klick.com.br/materia/20/display/0,5912,POR-20-91-931-,00.html que aborda de maneira clara alguns procedimentos que podem ser utilizados quando nos deparamos com situações em que precisamos resumir as informações de grandes conjuntos de dados.

Referências

ANDERSON, David R.; SWEENEY, Denis J.; WILLIAMS, Thomas A. **Estatística aplicada à administração e economia**. São Paulo: Pioneira Thomson Learning, 2003.

COSTA NETO, Pedro Luiz de Oliveira. **Estatística**, São Paulo: Edgard Blucher, 2002.

DOWNING, Douglas; CLARK, Jeffrey. **Estatística aplicada**. São Paulo: Saraiva, 2002.

FARIAS, Alfredo Alves de; SOARES, José Francisco; CÉSAR, Cibele Comini. **Introdução à estatística**. Rio de Janeiro: LTC, 2003.

MEMÓRIA, José M. P. **Breve História da Estatística**. Disponível em: <http://www.im.ufrj.br/~lpbraga/prob1/historia_estatistica.pdf>. Acesso em: 25 setembro 2014.

TRIOLA, Mario F.. **Introdução à estatística**. Rio de Janeiro: LTC, 1999.

VIEIRA, Sonia. **Elementos de estatística**. São Paulo: Atlas, 2003.

No próximo capítulo

Se até agora vimos como organizar, resumir e apresentar os dados (informações) em tabelas e gráficos, no próximo capítulo iremos incrementar esse processo através da inserção das medidas de posição e dispersão. São medidas que irão, de certa forma, representar o conjunto como um todo. Um exemplo bem conhecido de medida de posição (ou de tendência central) é a média e com relação à medida de dispersão podemos citar o desvio-padrão. Não são raras as situações em que a média é utilizada para representar a tendência central dos dados e o desvio-padrão para representar a variabilidade do conjunto de dados.

Além destas, veremos outras também importantes e com larga aplicação no estudo dos dados.

Capítulo 2

Medidas de posição

Nesse capítulo, aprenderemos como caracterizar um conjunto de dados através de medidas numéricas que sejam representativas de todo o conjunto.

As medidas de posição, também chamadas de medidas de tendência central, têm o objetivo de representar o ponto central de um conjunto de dados. As mais conhecidas são a média, a mediana e a moda. Além dessas medidas, podemos citar outras medidas de posição importantes, que não necessariamente são centrais. São ela os quartis, os decis e os percentis. Vamos estudar cada uma dessas medidas de posição (estatísticas).

Primeiramente, vamos fazer um estudo para os dados não tabulados, ou seja, quando os dados não estiverem na forma de distribuição de frequência. Em seguida, as mesmas medidas serão calculadas com base em dados tabulados.

Objetivos de sua aprendizagem

Por meio do estudo deste capítulo, esperamos que você seja capaz de calcular e de interpretar as medidas de posição aplicadas a conjuntos de dados.

Você se lembra?

Você se lembra das situações para as quais já calculou uma média? Que tipo de informação essa medida fornece? Para que serve? Para aplicar e interpretar medidas como ela, é necessário conhecê-las bem. Vamos, então, realizar um estudo detalhado da média e de outras medidas de mesma natureza.

2.1 Média

A **média aritmética** é a mais comum e mais simples de ser calculada dentre todas as medidas de posição mencionadas.

Para calculá-la, basta fazer a divisão da soma de todos os valores (x_1, x_2, ..., x_n) da variável pelo número deles (n):

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (2.1)$$

em que:

$\overline{x}$ = a média aritmética;

x_i = os valores da variável;

n = o número de valores.

Outro tipo de média que podemos encontrar é a média geométrica. Ela é muito utilizada no cálculo da taxa média de retorno de investimentos.

A média geométrica entre números reais $x_1 \cdot x_2, \ldots, x_n$ é definida como sendo a raiz n-ésima do produto dos n termos (ou, alternativamente) é o produto dos n termos elevado ao inverso do número de termos, ou seja:

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ \ldots x_n}$$

ou

$$\left(x_1 \cdot x_2 \cdot x_3 \ldots x_n\right)^{\frac{1}{n}} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

em que $\prod_{i=1}^{n} x_i$ indica o produtório de x_i, para i variando de 1 a n.

Em algumas circunstâncias não faz sentido calcular a média geométrica:

- Quando um dos valores do conjunto de dados for zero. Neste caso, o produto dos valores será zero e, consequentemente, G = 0.
- Quando o produto dos valores for negativo e o número total de observações for par. Neste caso, teríamos que calcular uma raiz de índice de par de um número negativo, o que é impossível no conjunto dos números reais.

A média geométrica será sempre menor ou igual a média aritmética.

2.2 Mediana (Md)

A mediana é outra medida de posição, dita mais robusta que a média, pois, da forma como ela é determinada, não permite que alguns valores muito altos ou muito baixos interfiram de maneira significativa em seu valor. Desta forma, se o conjunto de dados apresentar alguns poucos valores discrepantes em relação à maioria dos valores do conjunto de dados, em geral, é aconselhável usar a mediana em vez da média.

A mediana é a medida de posição mais frequentemente usada quando a variável em estudo for renda (R$), pois algumas rendas extremamente elevadas podem inflacionar a média. Neste caso, a mediana é uma melhor medida de posição central.

A mediana é encontrada ordenando os dados do menor para o maior valor e em seguida identificando o valor central destes dados ordenados. É uma medida que divide o conjunto de dados ao meio, deixando a mesma quantidade de valores abaixo dela e acima.

A determinação da mediana difere no caso do tamanho (n) do conjunto de dados ser par ou ímpar. Vejamos a seguir.

Se o número de elementos do conjunto de dados for ímpar, então a mediana será exatamente o valor "do meio", ou seja:

$$Md = x_{\left(\frac{n+1}{2}\right)} \qquad (2.2)$$

Se o número de elementos do conjunto de dados for par, então a mediana será exatamente a média "dos dois valores do meio", isto é:

$$Md = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}\right)+1}}{2} \qquad (2.3)$$

onde $x_{\left(\frac{n}{2}\right)}$, $x_{\left(\frac{n+1}{2}\right)}$ e $x_{\left(\frac{n}{2}\right)+1}$ indicam as posições onde os dados se encontram.

2.3 Moda (Mo)

A moda de um conjunto de dados é o valor (ou valores) que ocorre com maior frequência. A moda, diferentemente das outras medidas de posição, também pode ser encontrada quando a variável em estudo for qualitativa. Existem conjuntos de dados em que nenhum valor aparece mais vezes que os outros. Neste caso, dizemos que o conjunto de dados não apresenta moda.

Em outros casos, podem aparecer dois ou mais valores de maior frequência no conjunto de dados. Nestes casos, dizemos que o conjunto de dados é bimodal e multimodal, respectivamente.

Por conta das definições diferentes, a média, a mediana e a moda fornecem, muitas vezes, informações diferentes sobre o centro de um conjunto de dados, embora sejam todas medidas de tendência central.

No exemplo 2.1 apresentaremos os cálculos das medidas de posição para dados não tabelados (dados brutos).

Exemplo 2.1: Um gerente de banco quis estudar a movimentação de pessoas em sua agência na segunda semana de determinado mês. Ele constatou que no primeiro dia entraram 1.348 pessoas, no segundo dia, 1.260 pessoas, no terceiro, 1.095, no quarto, 832 e no último dia do levantamento, 850 pessoas. Encontre a média aritmética, a mediana e a moda para este conjunto de dados e interprete os resultados.

Resolução

A média aritmética é dada por:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1.348 + 1.260 + 1.095 + 832 + 850}{5} = \frac{5.385}{5} = 1.077$$

O número médio de pessoas que entraram na agência bancária na segunda semana do mês foi 1.077 pessoas. Isto quer dizer que, alguns dias entraram menos que 1.077 e outros dias entraram mais, ou seja, 1.077 é um valor em torno do qual o número de pessoas que entraram na agência, durante a segunda semana de cada mês, se concentra.

Para encontrar a mediana, devemos, primeiramente, ordenar os dados em ordem crescente (pode ser decrescente também):

832, 850, 1095, 1260, 1348

Como a quantidade de dados (n) é um número ímpar, a mediana é exatamente o valor que se encontra no meio do conjunto de dados. Nesse caso, a mediana é Md = 1095 pessoas. Isto significa que temos o mesmo número de observações menores ou iguais ao valor da mediana e o mesmo número de observações maiores ou iguais ao valor da mediana.

Este conjunto de dados não possui moda, pois não existe nenhum valor que "aparece" com mais frequência que os outros.

Agora, vamos fazer um estudo para os dados tabulados, ou seja, quando os dados estiverem na forma de uma distribuição de frequências.

Quando os dados estiverem tabulados, ou seja, na forma de distribuição de frequências, a maneira de se calcular a média aritmética muda um pouco. Como as frequências são números que indicam quantas vezes aparece determinado valor ou quantos valores têm em cada classe de frequência, elas funcionarão como "fatores de ponderação". Estas situações serão apresentadas nos exemplos 2.2 e 2.3, respectivamente.

Média Aritmética

No caso de dados tabulados, o cálculo da média aritmética é dada por:

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i \times f_i}{\sum_{i=1}^{k} f_i} \qquad (2.4)$$

Onde:

x_i é o valor da variável (ou o ponto médio de uma classe de frequência).

f_i é a frequência referente a cada valor (ou classe).

$\sum_{i=1}^{k}$ é a soma dos valores das frequências.

A expressão (2.4) apresentada anteriormente também é conhecida como fórmula da média ponderada.

No caso de distribuições de frequências que não apresentam intervalos de classes, a mediana e a moda são encontradas da maneira descrita nos itens 2.2 e 2.3, respectivamente.

Exemplo 2.2

Em um determinado mês, foi computado o número x de faltas ao trabalho, por motivos de saúde, que cada funcionário de uma determinada empresa teve. Os dados estão apresentados na tabela abaixo:

Número de Faltas	f
0	31
1	20
2	8
3	2
4	0
5	1
6	1
Total	**63**

Tabela 2.1: Número de faltas ao trabalho, por motivos de saúde.

Resolução

Média Aritmética

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i \times f_i}{\sum_{i=1}^{k} f_i} = \frac{(0\times31)+(1\times20)+(2\times8)+(3\times2)+(4\times0)+(5\times1)+(6\times1)}{63} = \frac{53}{63} \cong 0,84$$

ou seja, nesta empresa ocorreram, em média, 0,84 faltas por funcionário, por motivo de saúde.

Mediana

Como os dados estão tabelados, eles já se encontram ordenados. Para ficar mais fácil encontrar o valor da mediana, vamos incluir na distribuição de frequências uma coluna com as frequências acumuladas.

Número de Faltas	f	f_a
0	31	31
1	20	51
2	8	59
3	2	61
4	0	61
5	1	62
6	1	63
Total	**63**	

Agora, identificaremos a frequência acumulada imediatamente superior à metade do somatório das frequências simples:

$$\frac{\sum_{i=1}^{k} f_i}{2} = \frac{63}{2} = 3{,}15$$

A frequência acumulada imediatamente superior a 31,5 é $f_a = 51$. Portanto, o valor da mediana é o valor da variável associado à $f_a = 51$, ou seja,

$$Md = 1 \text{ falta}$$

Então, pelo menos 50% das observações são maiores ou iguais a 1 falta.

No caso do valor $\frac{\sum_{i=1}^{k} f_i}{2}$ ser exatamente igual a uma das frequências acumuladas f_a, o cálculo da mediana será a média aritmética entre dois valores da variável: x_i e $x_{(i+1)}$.

O valor da variável x_i será aquele cujo $\frac{\sum_{i=1}^{k} f_i}{2} = fa$ e o valor da variável x_{i+1} será aquele que está imediatamente após x_i na distribuição de frequência.

Moda

O valor que tem a maior frequência para este conjunto de dados é de $x = 0$, ou seja, é mais frequente encontrar funcionários que não faltam.

No caso do Exemplo 2.3 veremos que os dados estão agrupados em intervalos de classes. Quando o conjunto de dados for apresentado sob a forma agrupada perdemos a informação dos valores das

As medidas resumo calculadas quando os dados estiverem agrupados em intervalos de classes são apenas aproximações dos verdadeiros valores, pois substituímos os valores das observações pelo ponto do médio do intervalo de classe.

observações. Neste caso, vamos supor que todos os valores dentro de uma classe tenham seus valores iguais ao ponto médio desta classe.

Os cálculos da média, da moda e da mediana para tabelas de frequências agrupadas em classes estão apresentados a seguir.

Vale ressaltar que, sempre que possível, as medidas de posição e dispersão devem ser calculadas antes dos dados serem agrupados.

Exemplo 2.3

A tabela abaixo apresenta a distribuição de frequências do tempo de vida de 60 componentes eletrônicos (medido em dias) submetidos à experimentação num laboratório especializado. Calcular as medidas de posição.

Tempo de vida (dias)	f	Ponto Médio (x_i)
3 ├18	3	10,5
18 ├33	4	25,5
Tempo de vida (dias)	f	Ponto Médio (x_i)
33 ├48	4	40,5
48 ├63	8	55,5
63 ├78	10	70,5
78 ├93	28	85,5
93 ├108	2	100,5
108 ├123	1	115,5
Total	60	

Tabela 2.2: Tempo de vida de componentes eletrônicos.

Resolução

Neste tipo de tabela, como temos classes de frequências, devemos encontrar um valor que represente cada classe, para que possamos efetuar os cálculos. Por exemplo, considerando a primeira classe de frequência,

Tempo de vida (dias)	f	Ponto Médio (x_i)
3 ├18	3	10,5

sabemos que 3 componentes eletrônicos tiveram tempo de vida entre 3 e 18 dias, porém, não sabemos exatamente qual foi o tempo de vida de cada um. Se considerarmos o limite inferior da classe (3) para efetuarmos os cálculos, estaremos subestimando as estimativas. Por outro lado, se considerarmos o limite superior da classe (18) estaremos superestimando as estimativas. Portanto, vamos utilizar o ponto médio de cada classe para podermos fazer os cálculos sem grandes prejuízos. A terceira coluna da tabela apresentada contém os pontos médios calculados para cada intervalo de classe. O valor do ponto médio passa a ser o nosso valor x_i a ser utilizado nos cálculos. Vamos aprender como se faz:

Média Aritmética

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \times f_i}{\sum_{i=1}^{k} f_i} =$$

$$= \frac{(10,5\times 3)+(25,5\times 4)+(40,5\times 4)+(55,5\times 8)+(70,5\times 10)+(85,5\times 28)+(100,5\times 2)+(115,5\times 1)}{60}$$

$$= \frac{4155}{60} = 69,25$$

Podemos dizer, através da média aritmética, que os componentes eletrônicos têm uma duração média de 69 dias e 6 horas (69,25 dias).

Mediana

Como os dados estão tabelados em classes de frequências, o cálculo da mediana fica um pouquinho mais complicado. Agora, teremos que encontrar a mediana através da seguinte fórmula:

$$Md = l_{\inf md} + \frac{A_{md}}{F_{md}} \cdot \left(\frac{\sum f_i}{2} - F_{md-1} \right) \quad (2.5)$$

onde:

$l_{\inf md}$ é o limite inferior da classe que contém a mediana;

$\sum f_i$ é o número total de observações da distribuição de frequências;

F_{md-1} é a frequência acumulada da classe anterior à classe que contém a mediana;

f_{md} é o número de observações da classe que contém a mediana;

A_{md} é a amplitude do intervalo de classe que contém a mediana.

No cálculo da mediana para os dados da tabela 2.2 temos que primeiramente encontrar a classe que contém a mediana. Esta classe corresponde à classe associada à frequência acumulada imediatamente superior à $\dfrac{\sum f_i}{2}$.

Como $\dfrac{\sum f_i}{2} = \dfrac{60}{2} = 30$, temos que a classe que contém a mediana é de 78 ├93 (pois $f_a = 57$).

Tempo de vida	*f*	P · M x_1	*fa*
3|–18	3	10,5	3
18|–33	4	21,5	7
33|–48	4	40,5	11
48|–63	8	55,5	19
63|–78	10	70,5	29
78|–93	28	85,5	57
93|–108	2	10,5	59
108|–123	3	115,5	60
Total	**60**		

Classe que contém a mediana

Nº de observações da classe que contém a mediana

fa da classe anterior à classe que contém a Md

$\sum f_i$

Além disso, temos:

$\sum f_i$ = número total de observações da distribuição de frequências. Portanto, $\sum f_i = 60$.

F_{md-1} = frequência acumulada da classe anterior à classe que contém a mediana. Portanto, $F_{md-1} = 29$.

f_{md} = número de observações da classe que contém a mediana. Portanto, $f_{md} = 28$.

A_{md} = amplitude do intervalo da classe que contém a mediana. Portanto, $A_{md} = 93 - 78 = 15$.

Agora, basta substituirmos todos os valores encontrados na fórmula 2.5 para encontrarmos o valor da mediana:

$$\text{Md} = 78 + \frac{15}{28} \cdot \left(\frac{60}{2} - 29 \right) = 78 + \frac{15}{28} \cdot (30 - 29) = 78 + 0,54 \cong 78,5$$

Através da mediana podemos dizer que pelo menos 50% dos componentes eletrônicos avaliados têm duração igual ou inferior a 78 dias e 12 horas.

Moda

No cálculo da moda para dados agrupados devemos primeiramente identificar a classe modal, ou seja, a classe que apresenta a maior frequência.

Após a identificação da classe modal, utilizaremos a seguinte fórmula para calcular a moda bruta:

$$M_o = \frac{l + L}{2} \quad (2.6)$$

onde:

l:limite inferior da classe modal

L:limite superior da classe modal

Conexão:

Sugerimos os vídeos: “Novo Telecurso – E. Fundamental – Matemática – Aula 34 (parte 1)” e “Novo Telecurso – E. Fundamental – Matemática – Aula 34 (parte 2)” disponíveis, respectivamente em http://www.youtube.com/watch?v=SyWbYOtAIYc&NR=1 e http://www.youtube.com/watch?v=ejMyWfuSO5k que apresenta de modo bem prático a utilização das medidas de posição.

Agora, vamos comentar sobre outras medidas de posição, menos utilizadas, porém importantes em algumas situações. São elas: *quartis*, *decis* e *percentis*.

2.4 Medidas Separatrizes: Quartis, Decis e Percentis

Os quartis, decis e percentis são muito similares à mediana, uma vez que também subdividem a distribuição de dados de acordo com a proporção das frequências observadas.

Já vimos que a mediana divide a distribuição em duas partes iguais, então, os quartis (Q_1, Q_2 e Q_3), como o próprio nome sugere, divide a distribuição dos dados ordenados em quatro partes, sendo, Q_1 o quartil que separa os 25% valores inferiores dos 75% superiores, Q_2 o que divide o conjunto ao meio (é igual à mediana) e Q_3 o que separa os 75% valores inferiores dos 25% superiores. Não há um consenso universal sobre um procedimento único para o cálculo dos quartis, e diferentes programas de computador muitas vezes produzem resultados diferentes.

Perceba que o 2º quartil, o 5º decil e o 50º percentil representam a própria mediana, ou seja, todas estas medidas separatrizes (Q_2, D_5, e P_{50}), dividem a distribuição dos dados ao meio, deixando 50% dos dados abaixo delas e 50% acima.

Os decis, por sua vez, dividem a distribuição dos dados em 10 partes (D_i, i = 1, 2, ..., 9) e os percentis dividem a distribuição em 100 partes (P_i, i = 1, 2, ..., 99)

As medidas separatrizes, geralmente, só são calculadas para grandes quantidades de dados.

No Excel, por exemplo, temos a opção de pedir o cálculo de tais medidas.

Com os cálculos dos quartis, juntamente com os valores mínimo e máximo do conjunto de dados, podemos construir um gráfico chamado desenho esquemático ou *boxplot*. A análise deste gráfico é bastante útil no sentido de informar, entre outras coisas, a variabilidade e a simetria dos dados.

Conexão:
Para se entender quais são os procedimentos utilizados na construção de um boxplot, bem como sua interpretação, leia o texto: " Diagrama de Caixa (Boxplots)" em: TRIOLA, Mario F.. Introdução à estatística. 10.ed.Rio de Janeiro: LTC, 2008, pp.98 a 102

2.4.1 Cálculo dos quartis e dos percentis para dados não agrupados em classes

Como os quartis são medidas separatrizes precisamos, primeiramente, ordenar o conjunto de dados.

O primeiro quartil (Q_1) será o valor da variável que ocupar a posição $\frac{n}{4}$. O segundo quartil (Q_2) será o valor da variável que ocupar a posição $\frac{2n}{4}$ e o terceiro quartil (Q_3) será o valor da variável que ocupar a posição $\frac{3n}{4}$. Quando fazemos estas divisões para encontrar as posições dos quartis, pode acontecer do resultado ser um número inteiro ou um número fracionário. Então, adotaremos a seguinte convenção:

- Se a divisão resultar num número fracionário, arredonde-o para cima e o valor do quartil será a resposta da variável encontrada nesta posição.
- Se a divisão for um número inteiro, o quartil será a média aritmética da resposta da variável que ocupar a posição encontrada com a resposta da variável que ocupar a posição seguinte.

Exemplo 2.4

Um escritório que presta consultoria em administração levantou os tempos de espera de pacientes que chegam a uma clínica de ortopedia para atendimento de emergência. Foram coletados os seguintes tempos, em minutos, durante uma semana. Encontre os quartis.

2 5 10 11 3 14 8 8 7 12 3 4 7 3 4 2 6 7

Resolução:

Para encontrarmos os quartis, precisamos ordenar o conjunto de dados. Então:

2 2 3 3 3 4 4 5 6 7 7 7 8 8 10 11 12 14

• Posição do primeiro quartil (Q_1): $\frac{n}{4}=\frac{18}{4}=4,5$

Como a divisão resultou em um valor fracionário, vamos arredondar para 5. Portanto, o primeiro quartil é o valor que está na quinta posição.

$$Q_1 = 3$$

Então, pelo menos 25% das observações são menores ou iguais a 3 minutos.

• Posição do segundo quartil (Q_2): $\frac{2\,x\,n}{4}=\frac{2\,x\,18}{4}=9$

Como a divisão resultou em um valor inteiro, o segundo quartil será o resultado da média aritmética entre o valor que está na nona posição e o valor que está na décima posição.

$$Q_2 = \frac{6+7}{2} = 6,5$$

Temos que pelo menos 50% das observações são maiores ou iguais a 6,5 minutos.

• Posição do terceiro quartil (Q_3): $\frac{3\,x\,n}{4}=\frac{3\,x\,18}{4}=13,5$

Como a divisão resultou em um valor fracionário, vamos arredondar para 14. Portanto, o terceiro quartil é o valor que está na décima quarta posição.

$$Q_3 = 8$$

Neste conjunto de dados, pelo menos 25% das observações são maiores ou iguais a 8 minutos.

Agora que já aprendemos a calcular e interpretar os quartis para dados não agrupados, vamos passar para o conceito de percentis.

Da mesma forma que nos quartis, o conjunto de dados deve estar ordenado.

Quando dividimos o conjunto de dados em 100 partes, obtemos 99 percentis.

O percentil p_k será a resposta da variável que ocupar a posição

$$\frac{(k \text{ x } n)}{100}$$

Adotaremos a seguinte convenção:

- Se a divisão resultar num número fracionário, arredonde-o para cima e o valor do percentil será a resposta da variável encontrada nesta posição.
- Se a divisão for um número inteiro, o percentil será a média aritmética da resposta da variável que ocupar a posição encontrada com a resposta da variável que ocupar a posição seguinte.

Exemplo 2.5

Vamos encontrar o trigésimo quinto percentil do conjunto de dados do exemplo 2.4.

2 2 3 3 3 4 4 5 6 7 7 7 8 8 10 11 12 14

Resolução

O percentil p_{35} será a resposta da variável que ocupar a posição

$$\frac{(35 \text{ x } 18)}{100} = 6{,}3 \cdot$$

Como a divisão resultou em um valor fracionário, vamos arredondar para 7. Portanto, o trigésimo quinto percentil é o valor que está na sétima posição.

$$P_{35} = 4$$

Então, aproximadamente 35% das observações são menores ou iguais a 4 minutos.

2.4.2 Cálculo dos quartis e dos percentis para dados agrupados em classes

Para calcular os quartis quando os dados estão organizados em intervalos de classes, utilizaremos o mesmo procedimento descrito para o cálculo da mediana (Q_2) para dados agrupados em classes.

Para o cálculo do Q_1 utilizamos a seguinte fórmula:

$$Q_1 = l_{\inf q_1} + \frac{A_{q1}}{f_{q1}} \cdot \left(\frac{\sum fi}{4} - F_{q1-1} \right) \quad (2.7)$$

onde:

$l_{\inf_{q1}}$ é o limite inferior da classe que contém o primeiro quartil;

$\sum f_i$ é o número total de observações da distribuição de frequências;

F_{q1-1} é a frequência acumulada da classe anterior à classe que contém o primeiro quartil;

f_{q1} é o número de observações da classe que contém o primeiro quartil;

A_{q1} é a amplitude do intervalo de classe que contém o primeiro quartil.

De maneira semelhante, o cálculo do Q_3 será feito utilizando a seguinte fórmula:

$$Q_3 = l_{\inf q_3} + \frac{A_{q_3}}{f_{q_3}} \cdot \left(\frac{3 \text{ x} \sum fi}{4} - F_{q_3-1} \right) \quad (2.8)$$

onde:

$l_{\inf_{q3}}$ é o limite inferior da classe que contém o terceiro quartil;

$\sum f_i$ é o número total de observações da distribuição de frequências;

F_{q3-1} é a frequência acumulada da classe anterior à classe que contém o terceiro quartil;

f_{q3} é o número de observações da classe que contém o terceiro quartil;

A_{q3} é a amplitude do intervalo de classe que contém o terceiro quartil.

Exemplo 2.6

Vamos utilizar os dados do exemplo 2.3 para encontrar o primeiro e o terceiro quartil.

Tempo de vida	f	P · M x_i	f_a
3 ├18	3	10,5	3
18 ├33	4	25,5	7
33 ├48	4	40,5	11
48 ├63	8	55,5	19
63 ├78	10	70,5	29
78 ├93	28	85,5	57
93 ├108	2	100,5	59
108 ├123	1	115,5	60
Total	**60**		

Resolução

Primeiramente, temos que encontrar a classe que contém o primeiro quartil. Esta classe corresponde à classe associada à frequência acumulada imediatamente superior à $\frac{\sum fi}{4}$.

Como $\frac{\sum fi}{4} = \frac{60}{4}$, temos que a classe que contém o primeiro quartil é de 48 ├63 (pois $f_a = 19$).

Além disso, temos:

$\sum f_i$ = número total de observações da distribuição de frequências. Portanto, $\sum f_i = 60$.

F_{q1} = frequência acumulada da classe anterior à classe que contém o primeiro quartil. Portanto, $f_{q1} = 11$.

f_{q1} = número de observações da classe que contém o primeiro quartil. Portanto, $f_{q1} = 8$.

A_{q1} = amplitude do intervalo da classe que contém o primeiro quartil. Portanto, $A_{q1} = 63 - 48 = 15$.

Agora, basta substituirmos todos os valores encontrados na fórmula 2.7 e encontrar o valor do primeiro quartil:

$$Q_1 = 48 + \frac{15}{8} \cdot \left(\frac{60}{4} - 11 \right) = 48 + \frac{15}{8} \cdot (15 - 11) = 18 + 7,5 = 55,5$$

De acordo com o resultado obtido podemos esperar que aproximadamente 25% dos dados são menores ou iguais a 55,5, ou seja, aproximadamente 25% dos componentes eletrônicos têm duração inferior a 55 dias e 12 horas.

Agora, vamos encontrar a classe que contém o terceiro quartil. Esta classe corresponde à classe associada à frequência acumulada imediatamente superior à $\frac{3 \text{ x} \sum fi}{4}$.

Como $\frac{3 \text{ x} \sum fi}{4} = \frac{3 \text{ x } 60}{4}$, temos que a classe que contém o terceiro quartil é de 78 ├93 (pois f_a=57).

Além disso, temos:

$\sum f_i$ = número total de observações da distribuição de frequências. Portanto, $\sum f_i$ = 60.

F_{q3}= frequência acumulada da classe anterior à classe que contém o terceiro quartil. Portanto, F_{q3} = 29.
f_{q3}= número de observações da classe que contém o terceiro quartil. Portanto, f_{q3} = 28.
A_{q3} = amplitude do intervalo da classe que contém o terceiro quartil. Portanto, A_{q3} = 93 – 78 = 15.

Agora, basta substituirmos todos os valores encontrados na fórmula 2.8 e encontrar o valor do terceiro quartil:

$$Q_3 = 78 + \frac{15}{28} \cdot \left(\frac{3 \text{ x } 60}{4} - 29 \right) = 78 + \frac{15}{28} \cdot (45 - 29) = 78 + 8,57 \cong 86,57$$

De acordo com o resultado obtido podemos esperar que aproximadamente 75% dos dados são menores ou iguais a 86,57, ou seja, aproximadamente 75% dos componentes eletrônicos têm duração inferior a 86 dias e 14 horas.

Agora, vamos passar para o cálculo dos *percentis*.

No caso dos dados estarem organizados em intervalos de classes, os percentis são calculados utilizando a seguinte fórmula:

$$P_k = l_{inf_{pk}} + \frac{A_{pk}}{f_{pk}} \cdot \left(\frac{k \ x \sum f_i}{100} - F_{pk-1} \right) \quad (2.9)$$

em que k = 1,2,…,99.

O procedimento para encontrar as quantidades que devem ser substituídas na fórmula 2.9 são os mesmos que utilizamos para encontrar os quartis.

Exemplo 2.7

Vamos utilizar os dados do exemplo 2.3 para encontrar o décimo quinto percentil.

Tempo de vida	f	P · M x_i	f_a
3 ├ 18	3	10,5	3
18 ├ 33	4	25,5	7
33 ├ 48	4	40,5	11
48 ├ 63	8	55,5	19
63 ├ 78	10	70,5	29
78 ├ 93	28	85,5	57
93 ├ 108	2	100,5	59
108 ├ 123	1	115,5	60
Total	**60**		

Primeiramente, temos que encontrar a classe que contém o décimo quinto percentil. Esta classe corresponde à classe associada à frequência acumulada imediatamente superior à $\frac{15 \text{ x } \sum f_i}{100}$.

Como $\frac{15 \text{ x } \sum f_i}{100} = \frac{15 \text{ x } 60}{100} = 9$, temos que a classe que contém o décimo quinto percentil é de 33 ├48 (pois $f_a = 11$).

Além disso, temos:

$\sum f_i$ = número total de observações da distribuição de frequências. Portanto, $\sum f_i = 60$.

F_{p15} = frequência acumulada da classe anterior à classe que contém o décimo quinto percentil. Portanto, $F_{p15} = 7$.

f_{p15} = número de observações da classe que contém o décimo quinto percentil. Portanto, $f_{p15} = 4$.

A_{p15} = amplitude do intervalo da classe que contém o décimo quinto percentil. Portanto, $A_{p15} = 48 - 33 = 15$.

Agora, basta substituirmos todos os valores encontrados na fórmula 2.9 e encontrar o valor do décimo quinto percentil:

$$P_{15} = 33 + \frac{15}{4} \cdot \left(\frac{15 \text{ x } 60}{100} - 7 \right) = 33 + \frac{15}{4} \cdot (9 - 7) = 33 + 7,5 = 40,5$$

De acordo com o resultado obtido podemos esperar que aproximadamente 15% dos dados são menores ou iguais a 40,5, ou seja, aproximadamente 15% dos componentes eletrônicos têm duração inferior a 40 dias e 12 horas.

Conexão:

Agora que já abordamos como se calcula os quartis e percentis, faça uma leitura do texto: “Decis (D_k)” em TIBONI, Conceição G.R. Estatística básica - para os cursos de Administração, Ciências Contábeis, Tecnológicos e de Gestão. São Paulo: Atlas, 2010.

Atividades

01. Os dados abaixo referem-se ao número de horas extras de trabalho de uma amostra de 64 funcionários de uma determinada empresa localizada na capital paulista.

10	10	12	14	14	14	15	16
18	18	18	18	18	19	20	20
20	20	20	21	22	22	22	22
22	22	22	22	22	22	22	22
23	23	24	24	24	24	24	24
24	25	25	25	25	26	26	26
26	26	26	27	27	27	28	28
29	30	30	32	35	36	40	41

Pede-se:

a) Calcule e interprete as seguintes medidas descritivas calculadas para os dados brutos (dados não tabulados): média aritmética; mediana; moda.

b) Construir uma distribuição de frequências completa (com freq. absoluta, freq. relativa, freq. acumulada e ponto médio).

c) Com a tabela construída no item b), encontre as seguintes medidas: média aritmética; mediana; moda; 1º quartil; 7º decil; 99º percentil. Interprete os resultados.

d) Construa o histograma para este conjunto de dados.

02. Os dados abaixo representam as vendas mensais (em milhões de reais) de vendedores de gênero alimentícios de uma determinada empresa.

Vendas mensais (em milhões de reais)	Número de vendedores
0 ├ 1	6
1 ├ 2	12
2 ├ 3	20
3 ├ 4	48
4 ├ 5	14
5 ├ 6	10
Total	110

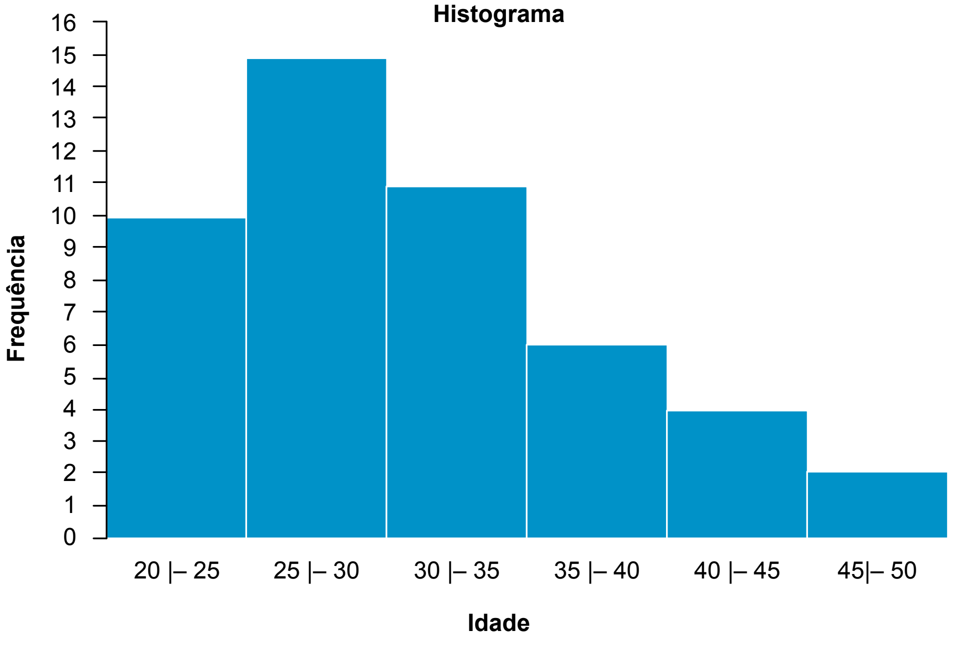
a) Qual a variável em estudo? Que tipo de variável é esta?
b) Encontre a média, a mediana e a moda e interprete os resultados.
c) Encontre as medidas separatrizes Q_3, D_1, e P_{80} e interprete os resultados.
d) Qual a porcentagem de vendedores com vendas mensais inferiores a 2 milhões de reais?
e) Qual a porcentagem de vendedores com vendas mensais superiores a 4 milhões de reais?
f) Qual a porcentagem de vendedores com vendas mensais entre 3 (inclusive) e 5 (exclusive) milhões de reais?
g) Qual a porcentagem de vendedores que vendem, pelo menos, 3 milhões de reais mensais?

03. Numa pesquisa realizada com 91 famílias, levantaram-se as seguintes informações com relação ao número de filhos por família:

número de filhos	0	1	2	3	4	5
frequência de famílias	19	22	28	16	2	4

Calcule e interprete os resultados da:
a) média aritmética
b) mediana
c) moda

04. O histograma abaixo representa a distribuição das idades dos funcionários de uma agência bancária. Com base no histograma abaixo, responda:



a) Qual a variável em estudo?

b) Quantos funcionários trabalham nesta agência bancária?

c) Quais são a média, a mediana e a moda para a idade dos funcionários desta agência? Interprete os resultados.

d) Qual o valor do primeiro quartil? Interprete o resultado.

e) Quantos funcionários têm menos que 30 anos?

f) Qual a porcentagem de funcionários com mais de 45 anos?

g) Qual a porcentagem de funcionários com no mínimo 30 anos?

05. Define-se a média aritmética de n números dados como o resultado da divisão por n da soma dos n números dados. Sabe-se que 4,2 é a média aritmética de 2.7; 3.6; 6.2; e x. Determine o valor de x.

Reflexão

Que a média é a medida de posição mais utilizada em nosso dia a dia talvez nem seria necessário dizer. Mas é preciso tomar certo cuidado quando utilizamos a média como parâmetro de um conjunto de dados. Você sabe que, se a média de sua turma em Estatística for igual a 7,0 (por exemplo), não quer dizer que toda ela, ou a maioria, teve bom desempenho nem que metade da turma teve desempenho igual ou superior a 7,0. Outras medidas, como vimos, podem complementar as informações dadas pela média.

Leitura recomendada

Como a média é uma medida descritiva muito utilizada no dia a dia, sugerimos que você ouça o áudio "Médias que interessam", da Série: Problemas e Soluções. Há dois módulos cujos conteúdos alertam par ao cuidado que se deve ter na interpretação da média, mostram uma aplicação da média ponderada e fazem uma análise crítica da utilização da média como uma informação única. O endereço para acesso é: http://m3.ime.unicamp.br/recursos/1315.

Referências

ANDERSON, David R.; SWEENEY, Denis J.; WILLIAMS, Thomas A. **Estatística aplicada à administração e economia**. São Paulo: Pioneira Thomson Learning, 2003.

BUSSAB, Wilton de O.; MORETTIN, Pedro A. **Estatística básica**. São Paulo: Saraiva, 2003.

COSTA NETO, Pedro Luiz de Oliveira. **Estatística**. São Paulo: Edgard Blucher, 2002.

DOWNING, Douglas; CLARK, Jeffrey. **Estatística aplicada**. São Paulo: Saraiva, 2002.

FARIAS, Alfredo Alves de; SOARES, José Francisco; CÉSAR, Cibele Comini. **Introdução à estatística**. Rio de Janeiro: LTC, 2003.

TIBONI, Conceição G. Rebelo. Estatística Básica - para os cursos de Administração, Ciências Contábeis, Tecnológicos e de Gestão. São Paulo: Atlas, 2010.

No próximo capítulo

Até agora estudamos estatísticas importantes de um conjunto de dados, tais como média, moda, mediana e medidas separatrizes. Estas medidas nos dão noção de posição central ou divisória do conjunto. No entanto, para que tenhamos informação mais completa do conjunto, é necessário estudar a sua variabilidade. As estatísticas que têm essa função são denominadas medidas de variabilidade ou de dispersão e serão abordadas no próximo capítulo.

Minhas anotações:

Capítulo 3

Medidas de dispersão

Estas medidas servem para indicar o quanto os dados se apresentam dispersos em torno da região central. Fornecem, portanto, o grau de variação existente no conjunto de dados. Dois ou mais conjuntos de dados podem, por exemplo, ter a mesma média, porém os valores poderão estar muito mais dispersos num conjunto do que no outro. Ou seja, podem ter maior ou menor grau de homogeneidade.

Objetivos de sua aprendizagem

Por meio do estudo deste capítulo, esperamos que você seja capaz de calcular e interpretar as medidas de dispersão aplicadas a conjuntos de dados, com o objetivo de avaliar o grau de homogeneidade.

Você se lembra?

Você se lembra de alguma vez em que saiu de casa tendo quase certeza de que ficaria preso em um engarrafamento no trânsito? Não é preciso ser muito observador para perceber que, em determinadas horas do dia, dependendo do dia da semana, o trânsito (nas grandes e nas médias cidades) estará congestionado. Talvez o melhor seria deixar para sair outra hora (se isto for possível). O fluxo de veículos, nesses momentos, apresenta certa homogeneidade, ou seja, quase sempre está intenso. Dificilmente, num dia como esses, você terá um fluxo acentuadamente menor (ou maior) do que o que você verifica todos os dias. Vamos estudar situações como essas, em que a informação sobre o grau de homogeneidade (ou heterogeneidade) nos ajudará a tomar a decisão mais adequada.

3.1 Exemplo Introdutório

Vamos analisar um exemplo bem simples que nos dá a ideia da importância de se conhecer as medidas de dispersão para a tomada de algumas decisões.

Exemplo 3.1: Imagine que estamos interessados em fazer uma viagem para Honolulu (Havaí) ou Houston (Texas) e para arrumar as malas necessitamos saber se a localidade a ser visitada faz calor, faz frio ou ambos. Se tivéssemos apenas a informação de que a temperatura média diária (medida durante um ano) das duas localizações fosse igual a 25 °C, poderíamos colocar na mala apenas roupas de verão? A resposta é não. Por exemplo, se estivéssemos interessados em viajar para o Havaí (em Honolulu), poderíamos levar apenas roupas de verão, pois a temperatura mínima observada durante um ano foi de 21 ºC e a máxima foi de 29º C. Porém, se resolvermos ir ao Texas (Houston), devemos tomar cuidado com a época, pois as temperaturas, durante um ano, variaram de 4 ºC (mínima) a 38 ºC (máxima). Com estas informações, concluímos que as temperaturas em Honolulu variam pouco em torno da média diária, ou seja, podemos levar uma mala apenas com roupas leves. Porém, em Houston, as temperaturas variam muito, com períodos de muito frio ou muito calor. Portanto, para ir à Houston sem perigo de sofrer com a temperatura, devemos analisar o período do ano para saber se a temperatura estará alta ou baixa.

As medidas de dispersão indicam o grau de variabilidade das observações. Estas medidas possibilitam que façamos distinção entre conjuntos de observações quanto à sua homogeneidade. Quanto menor as medidas de dispersão, mais homogêneo é o conjunto de dados.

Percebemos, através desse exemplo bem simples, que uma simples medida de dispersão (a amplitude, por exemplo) já ajudaria muito a tomar certos cuidados com a arrumação das bagagens.

Veremos, nos próximos itens, como calcular e interpretar as seguintes medidas de dispersão: amplitude, amplitude interquartil, desvio-padrão, variância e coeficiente de variação.

Primeiramente, vamos apresentar os cálculos das medidas de dispersão para dados não-tabulados, ou seja, quando os dados não estiverem na forma de distribuição de frequências.

3.2 Amplitude Total (R)

A amplitude total é a diferença entre o maior e o menor valor observado no conjunto de dados, ou seja:

$$R = x_{(\text{máximo})} - x_{(\text{mínimo})} \qquad (3.1)$$

A amplitude não é uma medida muito utilizada, pois só leva em conta dois valores de todo o conjunto de dados e é muito influenciada por valores extremos. No próximo item estudaremos uma medida de dispersão mais resistente a valores extremos.

3.3 Amplitude interquartil

A amplitude interquartil, ou distância interquartil, é uma medida de variabilidade que não é facilmente influenciada por valores discrepantes no conjunto de dados. Ela engloba 50% das observações centrais do conjunto de dados e seu cálculo é definido como:

$$\text{Amplitude interquartil} = Q_3 - Q_1 \qquad (3.2)$$

Agora, vamos estudar uma medida de dispersão muito utilizada e que leva em conta todos os valores do conjunto de dados: o desvio-padrão.

3.4 Desvio-Padrão (s)

Primeiramente, vamos entender qual é a definição da palavra desvio em estatística. Desvio nada mais é do que a distância entre qualquer valor do conjunto de dados em relação à média aritmética deste mesmo conjunto de dados.

O valor do desvio-padrão nunca é negativo. É zero apenas quando todos os valores do conjunto de dados são os mesmos. A unidade do desvio-padrão é a mesma unidade dos dados originais.

Existem várias medidas de dispersão que envolvem os desvios. São elas: o desvio-padrão (mais utilizada), a variância e o coeficiente de variação.

O desvio-padrão é a medida mais utilizada na comparação de diferenças entre grupos, por ser mais precisa e estar na mesma medida do conjunto de dados. Matematicamente, sua fórmula é dada pela raiz quadrada da média aritmética aproximada dos quadrados dos desvios, ou seja:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}} \qquad (3.3)$$

onde x_i é cada uma das observações do conjunto de dados, $\bar{x}$ é a média do conjunto de dados e n é o número total de observações do conjunto de dados.

Desenvolvendo a fórmula (3.3) chegamos a fórmula (3.4) que, para alguns casos, tornam os cálculos mais simples e rápidos.

$$s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} \qquad (3.4)$$

onde:

$\sum x_i^2$ é a soma de cada valor da variável ao quadrado;

$\left(\sum x_i\right)^2$ é o quadrado da soma de todos os valores da variável;
n é o número total de valores do conjunto de dados.

Como o desvio-padrão é uma medida de dispersão e mede a variabilidade entre os valores, temos que valores muito próximos resultarão em desvios-padrões pequenos, enquanto que valores mais espalhados resultarão em desvios-padrões maiores.

3.4.1 Uma regra prática para interpretar o desvio-padrão

Depois que calculamos o desvio-padrão, surge uma pergunta: como interpretá-lo?

Para conjuntos de dados que tenham distribuição em forma de sino, valem as seguintes considerações:

- Cerca de 68% das observações do conjunto de dados ficam a 1 desvio-padrão da média, ou seja, $(\bar{x} - s)$ e $(\bar{x} + s)$.

Conexão: Para se entender um pouco mais sobre o conceito de variabilidade, acesse o endereço http://www.gaussconsulting.com.br/si/site/05072.

- Cerca de 95% das observações do conjunto de dados ficam a 2 desvios-padrões da média, ou seja, $(\bar{x} - 2s)$ e $(\bar{x} + 2s)$.
- Cerca de 99,7% das observações do conjunto de dados ficam a 3 desvios-padrões da média, ou seja, $(\bar{x} - 3s)$ e $(\bar{x} + 3s)$.

3.5 Variância (s^2)

A variância de um conjunto de dados nada mais é do que o valor do desvio-padrão elevado ao quadrado, ou seja,

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad (3.5)$$

ou

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} \qquad (3.6)$$

A variância não é uma medida muito utilizada para mostrar a dispersão de um conjunto de dados, pois, expressa o seu resultado numa medida ao quadrado, não sendo possível interpretar o seu valor. Portanto, na análise descritiva dos dados, não vamos trabalhar com esta medida constantemente. Se um determinado problema fornecer a variância do conjunto de dados, basta calcularmos a raiz quadrada deste valor (variância) e obteremos o desvio-padrão, que é facilmente interpretado por estar na mesma medida do conjunto de dados.

3.6 Coeficiente de Variação (cv)

O coeficiente de variação (cv) é definido como o quociente entre o desvio-padrão e a média, e é frequentemente expresso em porcentagem. Ele mede o grau de variabilidade do conjunto de dados. Quando calculamos o desvio-padrão, obtemos um valor que pode ser grande ou pequeno, dependendo da variável em estudo. O fato de ele ser um valor considerado alto é relativo, pois dependendo da variável que está sendo estudada e da média, esta variação dos dados pode ser relativamente pequena. Então, o coeficiente de variação serve para calcular o grau de variação dos dados em relação à média aritmética. E é obtido através do seguinte cálculo:

$$cv = \frac{s}{\bar{x}} \cdot 100 \qquad (3.7)$$

onde s é o desvio-padrão e $\bar{x}$ é a média aritmética.

Alguns autores consideram a seguinte regra empírica para a interpretação do coeficiente de variação:

- Baixa dispersão: C.V ≤ 15%
- Média: C.V . 15% – 30%
- Alta: C.V ≥ 30%

Em geral, o coeficiente de variação é uma estatística útil para comparar a variação para valores originados de diferentes variáveis (por exemplo: peso, em Kg e altura, em cm), pois ele é adimensional.

3.7 Exemplo de aplicação das medidas de dispersão para dados não tabulados

Vamos exemplificar o cálculo da amplitude, da amplitude interquartil, do desvio-padrão, da variância e do coeficiente de variação utilizando o exemplo 2.1, que apresenta o conjunto de dados brutos.

Exemplo 3.2: Um gerente de banco quis estudar a movimentação de pessoas em sua agência na segunda semana de determinado mês. Ele constatou que no primeiro dia entraram 1.348 pessoas, no segundo dia, 1.260 pessoas, no terceiro, 1.095, no quarto, 832 e no último dia do levantamento, 850 pessoas. Encontre a amplitude, o desvio-padrão, a variância e o coeficiente de variação para este conjunto de dados e interprete os resultados.

Resolução

A amplitude é dada por:

$$R = x_{(\text{máximo})} - x_{(\text{mínimo})} = 1.348 - 832 = 516 \text{ pessoas.}$$

A diferença, no número de pessoas que entram na agência, entre o dia de maior movimento e o dia de menor movimento é de 516 pessoas.

Para encontrarmos a amplitude interquartil, precisamos calcular o primeiro e o terceiro quartil. Para isto, vamos seguir os procedimentos descritos no item 2.4.1.

- Posição do primeiro quartil (Q_1): $\frac{n}{4} = \frac{5}{4} = 1,25$

Como a divisão resultou em um valor fracionário, vamos arredondar para 2. Portanto, o primeiro quartil é o valor que está na segunda posição do conjunto de dados ordenados.

$$Q_1 = 850$$

- Posição do terceiro quartil (Q_3): $\frac{3 \text{ x } n}{4} = \frac{3 \text{ x } 5}{4} = 3,75$.

Como a divisão resultou em um valor fracionário, vamos arredondar para 4. Portanto, o terceiro quartil é o valor que está na quarta posição.

$$Q_3 = 1260$$

Então:

$$\begin{aligned} \text{Amplitude interquartil} &= Q_3 - Q_1 \\ &= 1260 - 850 \\ &= 410 \text{ pessoas} \end{aligned}$$

Então, a amplitude do intervalo que contém 50% das observações centrais é 410 pessoas.

O desvio-padrão é obtido através das fórmulas (3.3) ou (3.4). Como a média aritmética é um número inteiro e existem poucos dados, a fórmula (3.3) é mais rápida de ser calculada. Porém, fica a critério de cada um a utilização de uma ou de outra. Lembrando que a média aritmética encontrada anteriormente é igual a 1.077 e utilizando a fórmula (3.3), temos:

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

$$= \sqrt{\frac{(1.348-1.077)^2 + (1.260-1.077)^2 + (1.095-1.077)^2 + (832-1.077)^2 + (850-1.077)^2}{5-1}} =$$

$$= \sqrt{\frac{(271)^2 + (183)^2 + (18)^2 + (-245)^2 + (-227)^2}{4}} =$$

$$= \sqrt{\frac{(73.441)+(33.489)+(324)+(60.025)+(51.529)}{4}} =$$

$$= \sqrt{\frac{218.808}{4}} = \sqrt{54.702} \cong 233{,}88 \text{ pessoas}$$

Neste exemplo, entram na agência, em média, 1.077 pessoas por dia. O número de pessoas que entra na agência varia, mas, tipicamete, a diferença em relação à média foi de aproximadamente 234 pessoas.

A variância, como vimos, é obtida através das fórmulas (3.5) ou (3.6), ou simplestemente, como já temos o desvio-padrão, basta elevarmos o valor encontrado ao quadrado. Para o nosso exemplo, temos:

$$s^2 = (233{,}88 \text{ pessoas})^2 = 54699{,}85 \text{ pessoas}^2$$

Nesse caso, não há como interpretar a expressão pessoas^2. Por esse motivo, utilizamos o desvio-padrão no lugar da variância.

O coeficiente de variação, dado pela fórmula (3.7), é muito fácil de ser obtido desde que já conheçamos os valores da média aritmética e do desvio-padrão. Pela fórmula podemos observar que basta fazermos uma simples divisão. Para este exemplo, temos que:

$$cv = \frac{s}{\bar{x}} = \frac{233{,}88}{1.077} \cong 0{,}2172 \text{ ou } 21{,}72\%$$

Utilizando a regra empírica, podemos dizer que o conjunto de dados apresenta uma média dispersão.

Agora, vamos aprender a calcular as medidas de dispersão através de dados tabulados.

Quando os dados estiverem na forma tabulada, haverá uma pequena diferença no cálculo das médias de dispersão, pois agora será necessário considerar as frequências, que funcionarão como “fatores de ponderação”, referentes a cada valor da variável.

3.8 Desvio-padrão para dados tabulados

Se os dados estiverem tabulados, o desvio-padrão pode ser encontrado da seguinte forma:

$$s = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \times f_i}{n-1}} \qquad (3.8)$$

Desenvolvendo a fórmula (3.8) chegamos à fórmula (3.9) que também é utilizada para o cálculo do desvio-padrão:

$$s = \sqrt{\frac{\sum x_i^2 \times f_i - \frac{\left(\sum x_i \times f_i\right)^2}{n}}{n-1}} \qquad (3.9)$$

onde, para ambas as fórmulas (3.8) e (3.9), x_i representa cada uma das observações do conjunto de dados ou, se os dados estiverem agrupados em classes de frequências, x_i representa o ponto médio da classe, $\bar{x}$ é a média do conjunto de dados, f_i é a frequência associada a cada observação (ou classe de observações) do conjunto de dados e n é o número de total de observações no conjunto de dados.

3.9 Variância para dados tabulados

A variância de um conjunto de dados agrupados é dada por:

$$s^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \times f_i}{n-1} \qquad (3.10)$$

ou

$$s^2 = \frac{\sum x_i^2 \times f_i - \frac{\left(\sum x_i \times f_i\right)^2}{n}}{n-1} \qquad (3.11)$$

A **amplitude**, a **amplitude interquartil** e o **coeficiente de variação** não sofrem modificações significativas. A amplitude continua sendo a diferença entre o maior e o menor valor (se os dados estiverem em classes de frequências, R será a diferença entre o limite superior da última classe e o limite inferior da primeira classe). A amplitude interquartil continua sendo a diferença entre o terceiro e o primeiro quartil e o cálculo do coeficiente de variação é feito utilizando a fórmula (3.7), porém, se os dados estiverem em classes de frequências, o desvio-padrão e a média aritmética são obtidos utilizando x_i como o ponto médio da classe.

3.10 Exemplo de aplicação das medidas de dispersão para dados tabulados

Para demonstração dos cálculos para dados tabelados, vamos continuar utilizando os exemplos desenvolvidos no item 2.3 (Exemplos 2.2 e 2.3).

Exemplo 3.3: Em um determinado mês, foi computado o número x de faltas ao trabalho, por motivos de saúde, que cada funcionário de uma determinada empresa teve. Os dados estão apresentados na tabela a seguir:

Número de Faltas	f
0	31
1	20
2	8
3	2
4	0
5	1
6	1
Total	**63**

Tabela 3.1 – Número de faltas ao trabalho, por motivos de saúde.

Encontre a amplitude, o desvio-padrão, a variância e o coeficiente de variação para este conjunto de dados e interprete os resultados.

Resolução

A **amplitude** para este conjunto de dados é dada por:

$$R = x_{(máximo)} - X_{(mínimo)} = 6 - 0 = 6 \text{ faltas}$$

A maior diferença entre os números de faltas ao trabalho, por motivo de saúde, que funcionários de uma determinada empresa tiveram no período de um mês, é 6 faltas.

A amplitude interquartil é dada por:

$$\text{Amplitude interquartil} = Q_3 - Q_1 = 1 - 0 = 1 \text{ falta}$$

O desvio-padrão é obtido através das fórmulas (3.8) ou (3.9). Para exemplificar, vamos trabalhar com a fórmula (3.9). Para facilitar, vamos montar um quadro com os resultados que nos interessa para aplicar tal expressão.

Número de faltas x_1	**f_i**	**x_i x f_i**	$x_i^2 \cdot f_i$
0	31	0	0
1	20	20	20
2	8	16	32
3	2	6	18
4	0	0	0
5	1	5	25
6	1	6	36
Total (∑)	**63**	**53**	**131**

Substituindo os valores encontrados no quadro acima na fórmula 3.9, obtemos:

$$s = \sqrt{\frac{\sum x_i^2 \times f_i - \frac{\left(\sum x_i \times f_i\right)^2}{n}}{n\text{ -}1}} = \sqrt{\frac{131 - \frac{(53)^2}{63}}{63-1}} =$$

$$= \sqrt{\frac{131 - \frac{2809}{63}}{62}} \cong \sqrt{\frac{131 - 44,59}{62}} \cong$$

$$\cong \sqrt{1,3938} \cong 1,18 \text{ faltas}$$

Podemos dizer que, em média, ocorre aproximadamente 1 falta por funcionário, por mês. Na verdade, sabemos que esse número de faltas por funcionário varia em torno da média, mas, tipicamente, a diferença em relação à média é de, aproximadamente, 1 falta.

A **variância** é obtida através das fórmulas (3.10) ou (3.11), porém, como já temos o desvio-padrão, basta elevarmos o valor encontrado ao quadrado. Portanto, temos:

$s^2 = (1{,}18 \text{ faltas})^2 = 1{,}3924 \text{ faltas}^2$

Como 1,3924 faltas^2 não tem interpretação, utilizamos o desvio-padrão em vez da variância para interpretar o comportamento dos dados.

O coeficiente de variação para este exemplo é dado por:

$$cv = \frac{s}{\bar{x}} = \frac{1{,}18}{0{,}84} \cong 1{,}4048 \text{ ou } 140{,}48\%$$

O coeficiente de variação nos diz que este conjunto de dados apresenta uma alta dispersão.

Para finalizarmos, vamos fazer os cálculos para os dados agrupados em classes de frequências. Para isto, vamos utilizar o exemplo 2.3 que se encontra no item 2.3.

Exemplo 3.4: A tabela a seguir apresenta a distribuição de frequências do tempo de vida de 60 componentes eletrônicos (medido em dias) submetidos à experimentação num laboratório especializado.

Tempo de vida (dias)	f	Ponto Médio (x_i)
3 ├18	3	10,5
18 ├33	4	25,5
33 ├48	4	40,5
48 ├63	8	55,5
63 ├78	10	70,5
78 ├93	28	85,5
93 ├108	2	100,5
108 ├123	1	115,5
Total	60	

Tabela 3.2 – Tempo de vida de componentes eletrônicos.

Calcule a amplitude, o desvio-padrão, a variância e o coeficiente de variação para este conjunto de dados e interprete os resultados.

Resolução

A amplitude para este conjunto de dados é dada por:

$R = x_{(máximo)} – X_{(mínimo)} = 123 – 3 = 120$ dias.

A maior diferença entre os tempos de vida (em dias) dos componentes eletrônicos foi de 120 dias, ou seja, o componente com maior sobrevivência durou 120 dias a mais do que o componente que durou menos tempo.

Para o cálculo da amplitude interquartil vamos utilizar os resultados obtidos no exemplo 2.6. Portanto:

$$\text{Amplitude interquartil} = Q_3 – Q_1 = 86{,}57 – 55{,}5 = 31{,}07 \text{ dias}$$

De acordo como valor obtido, concluímos que 50% das observações centrais do conjunto de dados estão contidas em um intervalo cuja amplitude é 31,07 dias.

Para o cálculo do desvio-padrão, podemos utilizar as fórmulas (3.8) ou (3.9), onde o termo x_i é o ponto médio de cada classe de frequência. Como a média aritmética envolve valores decimais, é mais simples efetuar os cálculos através da fórmula (3.9). Como no exemplo anterior, vamos construir um quadro acrescentando as colunas que fornecerão os valores que precisamos para substituir na fórmula 3.9.

Classes de-frequências	f	Pm (x_i)	$x_i \cdot f_i$	$x_i^2 \cdot f_i$
3 ├18	3	10,5	31,5	330,75
18 ├33	4	25,5	102	2601
33 ├48	4	40,5	162	6561
48 ├63	8	55,5	444	24642
63 ├78	10	70,5	705	49702,5
78 ├93	28	85,5	2394	204687
93 ├108	2	100,5	201	20200,5
108 ├123	1	115,5	115,5	13340,25
Total	**60**		**4155**	**322065**

Com os valores obtidos, temos:

$$s=\sqrt{\frac{\sum x_i^2\times f_i-\frac{\left(\sum x_i\times f_i\right)^2}{n}}{n-1}}=\sqrt{\frac{322065-\frac{(4155)^2}{60}}{60-1}}=\sqrt{\frac{322065-\frac{17264025}{60}}{59}}\cong$$

$$\cong\sqrt{\frac{322065-287733,75}{59}}\cong\sqrt{581,89}\cong 24,12\text{ dias}$$

Em média, os componentes eletrônicos têm duração de 69 dias e 6 horas com uma variação de, aproximadamente, 24 dias e 3 horas para mais ou para menos com relação à média.

A **variância**, como já sabemos, é o desvio-padrão ao quadrado. Assim, temos:

$s^2 = (24,12 \text{ dias})^2 = 581,77 \text{ dias}^2$

Como 581,77 dias^2 não tem interpretação, utilizamos o desvio-padrão para interpretar o comportamento dos dados.

O **coeficiente de variação** para este exemplo é:

$$cv=\frac{s}{\bar{x}}=\frac{24,12}{69,25}\cong 0,3483 \text{ ou } 34,83\%$$

o que indica uma variabilidade alta no conjunto de dados.

Atividades

01. Vamos utilizar, entre outros exercícios, os mesmos da capítulo 2, porém encontrando as medidas de dispersão.

Os dados abaixo referem-se ao número de horas extras de trabalho de uma amostra de 64 funcionários de uma determinada empresa localizada na capital paulista.

10	10	12	14	14	14	15	16
18	18	18	18	18	19	20	20
20	20	20	21	22	22	22	22
22	22	22	22	22	22	22	22
23	23	24	24	24	24	24	24
24	25	25	25	25	26	26	26
26	26	26	27	27	27	28	28
29	30	30	32	35	36	40	41

a) Calcule e interprete as seguintes medidas de dispersão, calculadas para os dados brutos (dados não tabulados): amplitude, desvio-padrão, variância e coeficiente de variação e interprete os resultados.

b) Por meio da distribuição de frequências (dados tabulados) construída para este conjunto de dados (no capítulo anterior), encontre a amplitude, o desvio-padrão, a variância e o coeficiente de variação e interprete os resultados.

02. Os dados abaixo representam as vendas mensais (em milhões de reais) de vendedores de gênero alimentícios de uma determinada empresa.

Vendas mensais (em milhões de reais)	Número de vendedores
0 ├─ 1	6
1 ├─ 2	12
2 ├─ 3	20
3 ├─ 4	48
4 ├─ 5	14
5 ├─ 6	10
Total	110

Encontre a amplitude, o desvio-padrão, a variância e o coeficiente de variação e interprete os resultados

03. Os dados a seguir representam as notas de 5 disciplinas de um determinado candidato em um concurso público. São elas:

2, 5, 8, 8, 9

Calcule a amplitude, o desvio-padrão, a variância e o coeficiente de variação. Interprete os resultados.

04. Numa pesquisa realizada com 91 famílias, levantaram-se as seguintes informações com relação ao número de filhos por família:

número de filhos	0	1	2	3	4	5
frequência de famílias	19	22	28	16	2	4

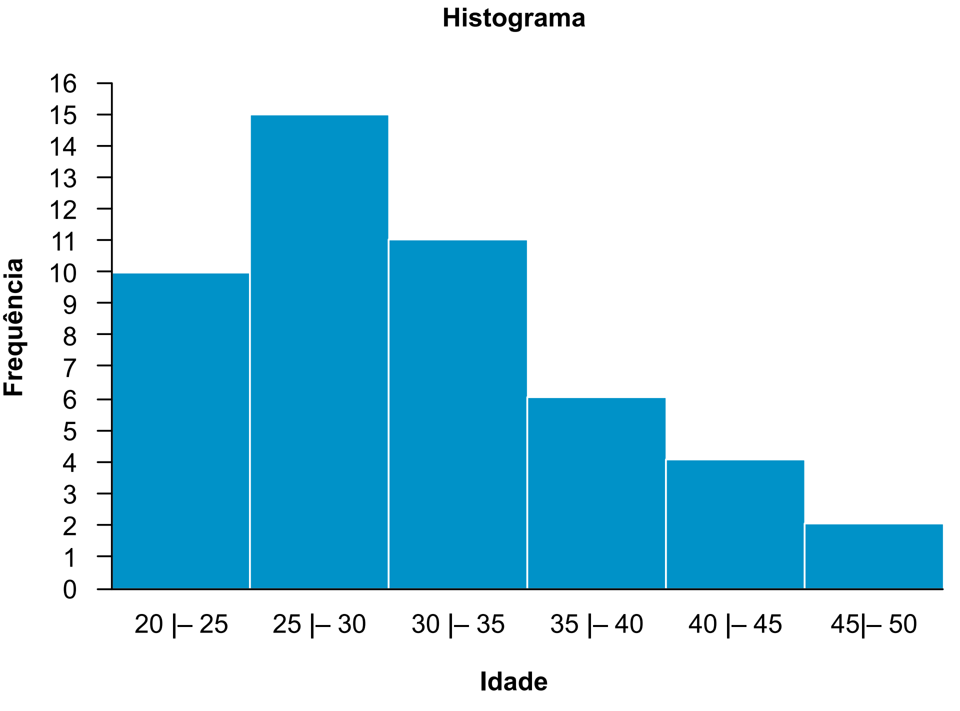
Calcule e interprete os resultados:

a) da amplitude;

b) do desvio-padrão;

c) do coeficiente de variação.

05. O histograma abaixo representa a distribuição das idades dos funcionários de uma agência bancária. Com base no histograma abaixo, responda:



Quais são a amplitude, o desvio-padrão, a variância e o coeficiente de variação para as idades dos funcionários? Interprete os resultados.

06. Um fabricante de caixas de cartolina fabrica três tipos de caixa. Testa-se a resistência de cada caixa, tomando-se uma amostra de 100 caixas e determinando-se a pressão necessária para romper cada caixa. Seguem os resultados dos testes:

Tipos de caixas	A	B	C
Pressão média de ruptura (bária)	15	20	30
Desvio-padrão das pressões (bária)	4	5	6

a) Que tipo de caixa apresenta a menor variação absoluta na pressão de ruptura?
b) Que tipo de caixa apresenta a maior variação relativa na pressão de ruptura?

Reflexão

Vimos, neste capítulo, que tão importante quanto conhecer a média de um conjunto de dados é determinar o seu grau de variabilidade (ou dispersão).

Por exemplo, num bairro nobre da capital paulista está uma das maiores favelas de São Paulo. Se analisarmos somente o valor da renda média do bairro certamente vamos concluir que o valor obtido é comparável às melhores economias do mundo. Porém, devemos levar em conta que a discrepância entre os diversos valores da renda deve ser muito grande. Então, para quantificar a variabilidade dos valores da variável em estudo é fundamental calcular as medidas de dispersão, particularmente o desvio-padrão.

Leitura recomendada

Aqui, sugerimos a leitura do artigo "E se todos fossem ao mesmo cinema ao mesmo tempo?" do professor Luiz Barco, disponível em http://super.abril.com.br/ciencia/lei-regularidade-estatistica-se-todos-fossem-ao-mesmo-cinema-ao-mesmo-tempo-439499.shtml. Ele retrata, de forma bem interessante, a questão da regularidade dos fenômenos relacionados ao comportamento social.

Referências

ANDERSON, David R.; SWEENEY, Denis J.; WILLIAMS, Thomas A. **Estatística aplicada à administração e economia**. São Paulo: Pioneira Thomson Learning, 2003.

BUSSAB, Wilton de O.; MORETTIN, Pedro A. **Estatística básica**. São Paulo: Saraiva, 2003.

COSTA NETO, Pedro Luiz de Oliveira. **Estatística**. São Paulo: Edgard Blucher, 2002.

DOWNING, Douglas; CLARK, Jeffrey. **Estatística aplicada**. São Paulo: Saraiva, 2002.

FARIAS, Alfredo Alves de; SOARES, José Francisco; CÉSAR, Cibele Comini. **Introdução à estatística**. Rio de Janeiro: LTC, 2003.

No próximo capítulo

Nos capítulos já vistos, estudamos formas de organizar e resumir dados por meio de distribuições de frequências e de medidas descritivas. São processos que, geralmente, nos passam informações sobre algo que já ocorreu. Tais informações também são de fundamental importância para que possamos prever o que irá acontecer no futuro. Para isso, no próximo capítulo, estudaremos a teoria de probabilidades, que, entre outras coisas, auxilia-nos na determinação de ocorrência de eventos futuros, tais como: vai chover amanhã, qual será minha receita no próximo mês, qual o nível de demanda de meu produto no próximo ano, entre muitos outros.

Minhas anotações:

Capítulo 4

Noções de Probabilidade

Nos capítulos anteriores vimos como organizar e descrever conjuntos de dados através de gráficos, tabelas e medidas resumo, tais como: medidas de posição e dispersão. Observamos que os resultados obtidos nos auxiliam na análise e interpretação dos dados.

Neste capítulo estudaremos conceitos básicos de probabilidade.

Objetivos da sua aprendizagem

Com o estudo dos conceitos abordados neste capítulo, você será capaz de identificar experimentos aleatórios e calcular as probabilidades de ocorrência de determinados eventos, através da definição de probabilidade e de suas propriedades.

Você se lembra?

Você se lembra do significado da palavra probabilidade? Sabe qual é o seu real sentido? E qual a sua importância em nosso dia a dia? Certamente você já deve ter feito perguntas cujas respostas dependiam do cálculo de probabilidades. Por exemplo:

- Qual a probabilidade de chover no próximo final de semana prolongado?
- Qual a probabilidade de se ganhar na Mega - Sena jogando um volante com seis números?
- Qual a probabilidade das vendas de determinado produto decrescer se aumentarmos o preço do produto?

O cálculo destas e outras probabilidades nos auxiliam na tomada de decisões.

Introdução

O cálculo efetivo de uma probabilidade depende frequentemente do uso dos resultados da análise combinatória. A análise combinatória é a parte da Matemática que desenvolve técnicas e métodos de contagem.

Apresentaremos, nos itens a seguir, um resumo dos principais resultados dessa área da Matemática elementar.

4.1 Princípio Fundamental da Contagem (PFC)

Para entendermos este conceito, vamos analisar o exemplo a seguir.

Exemplo 4.1

Um quiosque de praia no Rio de Janeiro lançou a seguinte promoção durante uma temporada de verão:

"Combinado de sanduíche natural e suco a R$ 10,00"

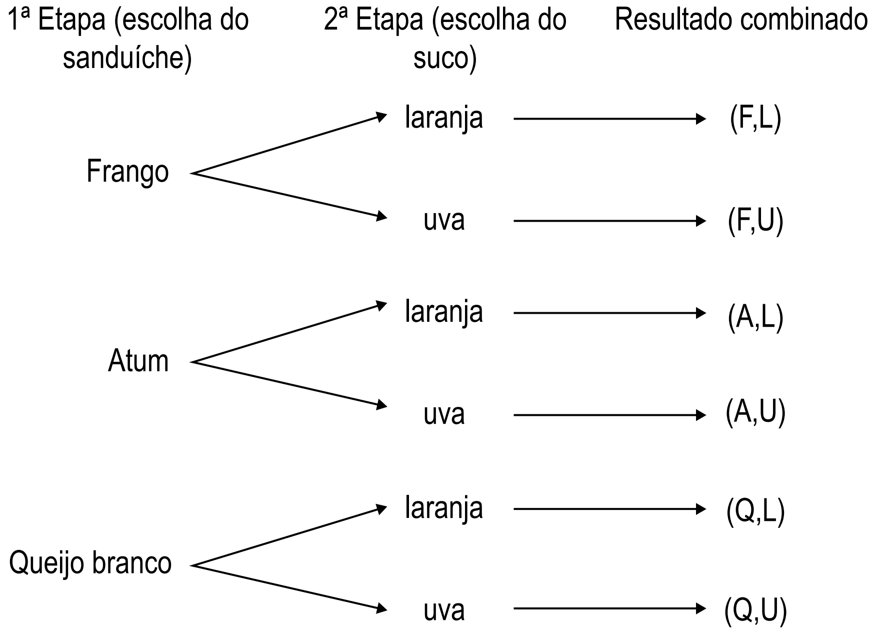
O cliente pode escolher três opções de sanduíche (frango, atum e queijo branco) e duas opções de suco (laranja e uva).

Considerando estas opções, de quantas formas distintas o cliente pode escolher seu combinado?

- O cliente poderá optar por três sabores do lanche: frango (F), atum (A) e queijo branco (Q).
- Para cada uma das possibilidades anteriores, a escolha do suco pode ser feita de duas maneiras: laranja (L) e uva (U).

A representação dessas possibilidades pode ser feita por meio de um diagrama conhecido como **diagrama de árvore**.

Temos:



Podemos observar que o número de combinados possíveis é $3 \cdot 2 = 6$.

Este exemplo nos ajuda a entender a definição a seguir.

Suponha que uma sequência ordenada seja formada por k elementos $(a_1, a_2, \ldots, a_k)$, em que:

- a_1 pode ser escolhido de n_1 maneiras distintas;
- a_2 pode ser escolhido de n_2 maneiras distintas, a partir de cada uma das possibilidades anteriores;
- $\vdots$
- a_k pode ser escolhido de n_k maneiras distintas, a partir das escolhas anteriores;

Então, o número de possibilidades para se construir a sequência $(a_1, a_2, \ldots, a_k)$ é:

$$n_1 \cdot n_2 \cdot \cdots n_k$$

Esse resultado é conhecido como **Princípio Fundamental da Contagem (PFC)** e serve de base para a resolução de problemas de contagem.

Exemplo 4.2

Quantos números de três algarismos podemos formar com os algarismos 0,1,2,3,4,5 e 6?

Resolução:

- O primeiro algarismo pode ser escolhido de seis maneiras distintas, pois o número que será formado não pode começar por zero. Observe que 021 = 21;
- O segundo algarismo pode ser escolhido de sete maneiras distintas, pois pode haver repetição de algarismo;
- O terceiro algarismo também pode ser escolhido de sete maneiras distintas.

Então, pelo PFC, a quantidade de números que podemos formar é:

$$6 \cdot 7 \cdot 7 = 294$$

Agora, passaremos ao conceito de fatorial de um número. Tal conceito é uma ferramenta de cálculo importante em Análise Combinatória.

4.2 Fatorial de um número natural

Definimos **fatorial de um número natural *n***, $n \geq 2$, e indicamos por n! (lemos " fatorial de n" ou " *n* fatorial"), o produto obtido pela multiplicação de n por todos os seus antecessores naturais positivos, ou seja:

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 3 \cdot 2 \cdot 1, \text{para } n \geq 2$$

Observação: Consideremos 0! = 1

Assim, temos, por exemplo:

$3! = 3 \cdot 2 \cdot 1 = 6$

$6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$

$7! = 7 \cdot \underbrace{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}_{6!} = 7 \cdot 6!$

Exemplo 4.3

Vamos encontrar o valor de $\dfrac{12!}{8!}$.

Resolução

Para encontrar o valor de 12!/8!, podemos desenvolver o fatorial do número maior (12) até chegarmos ao fatorial do número menor (8).

Então:

$$\frac{12!}{8!} = \frac{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8!}{8!} = 11880$$

Podemos dividir o 8! do numerador com o 8! do denominador, obtendo como resultado 1. Então, o resultado final será a multiplicação de $12 \cdot 11 \cdot 10 \cdot 9$.

Agora, vamos estudar métodos de contagem de determinados agrupamentos, baseados no PFC, que simplificarão a resolução de muitos problemas.

4.3 Arranjo

Dado um conjunto com *n* elementos distintos, chamamos arranjo dos n elementos, tomados *k* a *k*, a qualquer sequência ordenada de k elementos distintos escolhidos entre os n existentes.

Expressamos a definição acima da seguinte maneira:

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

Exemplo 4.4

Dado o conjunto A={1,2,3,4,5,6}, quantos arranjos desses seis elementos tomados dois a dois podemos formar?

Resolução

Vamos escrever todas as sequências ordenadas de dois elementos distintos escolhidos entre os seis elementos do conjunto A:

(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(2,1)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,4)	(3,5)	(3,6)	(4,1)	(4,2)	(4,3)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,6)	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)

Da sequência descrita acima, observamos que cada arranjo difere dos demais:

- pela natureza dos elementos escolhidos:

$$(1,3) \neq (3,5)$$

- pela ordem dos elementos escolhidos:

$$(1,3) \neq (3,1)$$

Já sabemos que a quantidade de arranjos que pode ser formada é 30. Utilizando a fórmula, obtemos:

$$A_{6,2} = \frac{6!}{(6-2)!} = \frac{6!}{4!} = \frac{6 \cdot 5 \cdot 4!}{4!} = 30$$

Na resolução do exemplo, listamos todos os possíveis arranjos somente para facilitar a compreensão do conceito.

A quantidade de arranjos que pode ser formada também pode ser obtida através do PFC:

$$6 \cdot 5 = 30$$

nº de opções para a escolha do 1º elemento do par

nº de opções para a escolha do 2º elemento do par

4.4 Permutação

Há situações em que devemos escolher n elementos distintos, entre os n disponíveis, para formar uma sequência ordenada (arranjo). Nestas situações, o nome dado a estes arranjos é **permutação**.

Assim, o número de permutações de n elementos distintos é dado por:

$$P_n = n!$$

Podemos observar que a permutação é um caso particular do arranjo, pois:

$$P_n = A_{n,n} = \frac{n!}{(n-n!)} = \frac{n!}{0!} = n!$$

Exemplo 4.5

Quantos são os anagramas (diferentes disposições das letras de uma palavra) da palavra PAZ?

Resolução

Vamos listar todos os anagramas:

PAZ PZA APZ AZP ZPA ZAP

O número de anagramas que pode ser formado é:

$$P_3 = 3! = 3 \cdot 2 \cdot 1 = 6$$

4.4.1 Permutação com elementos repetidos

Em uma permutação com elementos repetidos, a troca de posição desses elementos repetidos não altera o resultado do anagrama.

Então, se temos n elementos, dos quais $\mathbf{n_1}$ são iguais a $\mathbf{a_1}$ ($\mathbf{a_1}$ representa, por exemplo, uma letra), $\mathbf{n_2}$ são iguais a $\mathbf{a_2}$ ($\mathbf{a_2}$ representa outra letra), ..., $\mathbf{n_r}$ são iguais a $\mathbf{a_r}$, o número de permutações possíveis é dado por:

$$P^{n}_{n_1, n_2, ..., n_r} = \frac{n!}{n_1!n_2!...n_r!}$$

Para ficar mais claro este conceito, vamos analisar o exemplo a seguir.

Exemplo 4.6

Vamos determinar o número de anagramas formados com a palavra ARITMÉTICA.

Resolução

A palavra ARITMÉTICA possui 10 letras, sendo 2 letras iguais a A, 2 iguais a I e 2 letras T. Então, temos um caso de permutação com elementos repetidos.

O número de anagramas que podemos formar é:

$$P^{10}_{2,2,2} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2!}{2!2!2!} = \frac{1814400}{4} = 453600$$

4.5 Combinação

Na definição de arranjo vimos que, quando tomamos k elementos distintos de n existentes, formamos uma sequência ordenada.

Há casos em que só interessam os elementos que compõem a sequência, não importando a ordem em que ali figuram. Nestes casos, temos, então, o que se chama de combinação de n elementos tomado k a k:

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

Exemplo 4.7

Maria quer escolher dois sabores de torta doce para servir em sua festa de final de ano. A doceria oferece os seguintes sabores: limão (L), chocolate (C), morango (M) e floresta negra (F). De quantas formas distintas Maria poderá fazer essa escolha?

Resolução

Neste exemplo conseguimos perceber que **não** importa a **ordem** em que os sabores são escolhidos. Escolher, por exemplo, torta de chocolate e morango {C,M} é o mesmo que escolher torta de morango e chocolate {M,C}. Cada possível escolha de Maria representa, portanto, uma combinação de quatro sabores tomados dois a dois.

O número de formas distintas de Maria escolher os sabores é:

$$C_{4,2} = \frac{4!}{2!(4-2)!} = \frac{4 \cdot 3 \cdot 2!}{2!2} = 6$$

Vamos listar as possíveis formas para melhor entendimento do conceito de combinação:

{L,C} {L,M} {L,F} {C,M} {C,F} {M,F}

Agora que já relembramos conceitos da análise combinatória que são uteis no cálculo de probabilidades, vamos estudar como se calcula probabilidades em diversas situações. Antes disto, vamos conhecer brevemente um pouco da história da probabilidade.

4.6 Breve histórico

Quando estudamos a história da probabilidade, o nome de Gerolamo Cardano sempre é citado. Ele foi o primeiro homem na história a sistematizar dados e a entender a lógica de alguns processos que até então eram tidos como aleatórios para grande parte da humanidade.

A probabilidade que conhecemos e estudamos nos dias atuais surgiu em meados do século XVII, a partir dos estudos de De Mére, Blaise Pascal e Pierre de Fermat. Segundo BRUNI (2010), os estudos feitos por Pascal e Fermat sobre várias situações de jogos deram origem ao desenvolvimento da Teoria de Probabilidades – as Leis do Acaso.

Com o desenvolvimento das teorias de probabilidades, houve uma evolução da ciência atuarial e das aplicações no mercado de seguros. Bernoulli, em 1730, pesquisou, com base em estudos com recém-nascidos, um novo modo de se calcular o número esperado de sobreviventes após n anos.

A etapa moderna da Teoria das Probabilidades teve início em 1933, com Andrei Kolmogorov. Ele lançou as bases axiomáticas da probabilidade, baseada na Teoria dos Conjuntos, reduzindo a Teoria das Probabilidades à Teoria da Integração.

Conexão:

Para você conhecer um pouco mais sobre a história da probabilidade sugerimos ouvir o áudio "História da Probabilidade", disponível em http://m3.ime.unicamp.br/recursos/1253.

4.7 Experimento Aleatório, Espaço Amostral, Evento

Antes de passarmos à definição de probabilidade vamos apresentar alguns conceitos básicos necessários para efetuar seu cálculo.

- **Experimento Aleatório:** é uma situação ou acontecimento cujo resultado não pode ser previsto com certeza. Cada experimento poderá ser repetido inúmeras vezes sob condições essencialmente inalteradas. Embora não possamos afirmar qual será o resultado de um particular experimento, podemos descrever o conjunto dos possíveis resultados.
- **Espaço Amostral:** é o conjunto formado por todos os resultados do experimento aleatório. Indicamos este conjunto pela

letra grega ômega Ω. Cada elemento do espaço amostral é denominado ponto amostral.

- **Evento:** é um subconjunto do espaço amostral (indicado por letras maiúsculas do nosso alfabeto). O evento que possui somente um elemento é denominado evento simples.

Exemplo 4.8

Considere o experimento aleatório que consiste no lançamento de um dado. Neste experimento, o espaço amostral é definido como {1,2,3,4,5,6}. Alguns dos eventos que podem ser definidos neste experimento são:

A: saída de face par
A = {2,4,6}
B: saída de face ímpar
B = {1,3,5}
C: saída de face maior que 6
C = Ø. Neste caso Ø indica o conjunto vazio. Este evento é denominado evento impossível.
D: saída de face menor que 2
D = {1}, que é denominado evento simples.
E: saída de face menor ou igual a 6
E = {1,2,3,4,5,6}, que é o próprio espaço amostral Ω. Este evento é denominado evento certo.

Podemos observar, pelos eventos definidos anteriormente, que um evento pode não conter elementos (conjunto vazio), conter somente um elemento (evento simples), conter mais de um elemento e, finalmente, pode ser constituído por todos os elementos do espaço amostral, ou seja, o evento é o próprio Ω.

4.8 Operações com Eventos

4.8.1 União

Dados dois eventos A e B, temos que a união destes dois eventos é o evento que contém os pontos amostrais pertencentes a A, ou a B ou a ambos. Denotamos a união por A ∪ B. O diagrama de Venn, na figura 4.1, descreve a união dos eventos A e B.

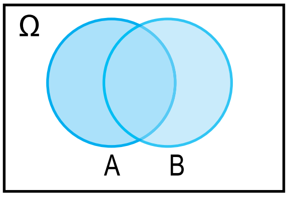


Figura 4.1 – União de dois eventos

4.8.2 Intersecção

A intersecção de dois eventos A e B, denotado por $A \cap B$, é o evento que contém os pontos amostrais que pertencem simultaneamente aos eventos A e B.

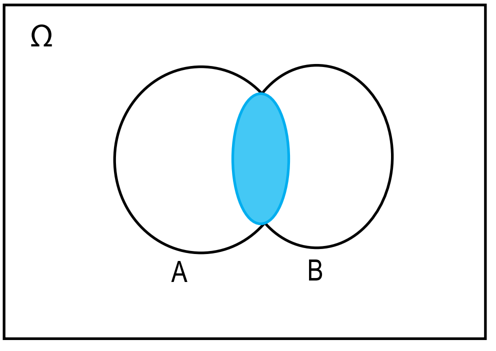


Figura 4.2 – Intersecção de dois eventos

Se $A \cap B = \emptyset$ temos que A e B são eventos mutuamente exclusivos, ou seja, quando um ocorre o outro não pode ocorrer.

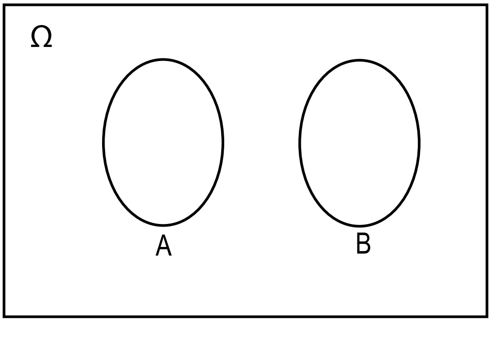


Figura 4.3 – Eventos mutuamente exclusivos

Observação

Quando estamos interessados na intersecção de dois eventos utilizamos a conjunção **e**, ou seja, queremos encontrar os elementos que pertencem ao evento A e ao evento B. No caso da união de dois eventos utilizamos a conjunção **ou**, ou seja, são elementos que pertencem ao evento A, ou ao B ou a ambos.

4.8.3 Complementação

O complemento do evento A, denotado por A^c , é definido como o evento que contém todos os pontos amostrais que não pertencem ao evento A, ou seja, $A^c = \Omega - A$

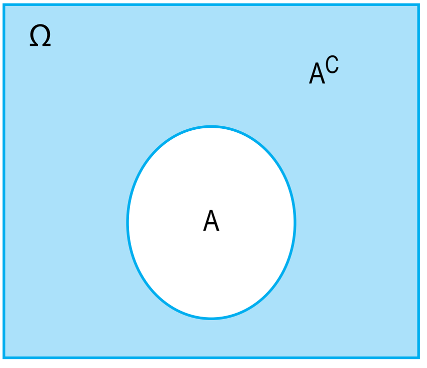


Figura 4.4 – Complementar do evento A

Exemplo 4.9

Considerando o experimento aleatório do exemplo 4.8 temos que $\Omega = \{1,2,3,4,5,6\}$.

Definindo os eventos A e B como:

A: saída de face par

$A = \{2,4,6\}$

B: saída de face menor ou igual a 4

$B = \{1,2,3,4\}$

Determinar $A \cup B$, $A \cap B$, A^c ,B^c , $A^c \cup B^c$, $A^c \cap B^c$, $A^c \cap B$, $B^c \cap A$

Resolução

$A \cup B = \{1,2,3,4,6\}$

$A \cap B = \{2,4\}$
$A^c = \{1,3,5\}$
$B^c = \{5,6\}$
$A^c \cup B^c = \{1,3,5,6\}$
$A^c \cap B^c = \{5\}$
$A^c \cap B = \{1,3\}$
$B^c \cap A = \{6\}$

4.9 Probabilidade

A probabilidade é um número que mede a possibilidade de ocorrência de um evento. O cálculo da probabilidade pode ser efetuado de três maneiras: através da **definição clássica** de probabilidade, através da **definição frequencial** de probabilidade e através do **método subjetivo**.

Vamos concentrar nossos estudos na definição clássica e frequencial. No método subjetivo, a probabilidade é estimada com base no conhecimento de circunstâncias relevantes. Por exemplo, dado o estado de saúde do paciente e a extensão dos ferimentos, um médico pode sentir que esse paciente tem uma chance de 95% de se recuperar completamente.

4.9.1 Definição Clássica

Aplicamos esta definição quando os pontos amostrais do espaço amostral são equiprováveis, ou seja, têm a mesma probabilidade de ocorrer. Por exemplo, quando jogamos um dado equilibrado todas as faces têm a mesma probabilidade de ocorrer, ou seja, 1/6.

Dado um evento A, a probabilidade de A, representada por P(A), é obtida através da definição clássica por:

$$P(A) = \frac{\text{número de resultados favoráveis ao evento A}}{\text{número total de resultados possíveis}} \qquad (4.1)$$

Exemplo 4.10

Retira-se uma carta de um baralho completo de 52 cartas. Qual a probabilidade de sair um rei?

Resolução

O espaço amostral deste experimento é:

$$\Omega = \{A_O, ..., K_O, A_E, ..., K_E, A_P, ..., K_P, A_C, ..., K_C\}$$

ou seja, temos 52 pontos amostrais igualmente prováveis de ocorrer. O evento A: sair um rei é o subconjunto A = $\{k_o, K_E, K_p, K_c\}$. Utilizando a definição clássica de probabilidade temos:

$$P(A) = \frac{4}{52}$$

Podemos expressar os resultados das probabilidades em forma de frações, decimais ou percentuais. Nesse caso, podemos indicar o resultado por

$P(A) = \frac{4}{52}, 0,0769$, ou, ainda, 7,69%.

4.9.2 Definição Frequencial

Vimos que a definição clássica de probabilidade só pode ser aplicada quando os pontos amostrais são igualmente prováveis de ocorrer. Em situações em que isto não ocorre podemos determinar a probabilidade através da definição frequencial. Esta definição baseia-se em observações repetidas do experimento aleatório. Seja A o evento de interesse. A probabilidade P(A) obtida através da definição frequencial é dada por:

$$P(A) = \frac{\text{número vezes que o evento A ocorreu}}{\text{número de repetições do experimento aleatório}} \qquad (4.2)$$

em que o número de repetições deve ser grande.

A ideia utilizada nesta definição é a mesma da frequência relativa definida no primeiro capítulo.

Exemplo 4.11

Uma loja de varejo tem registrado em seus arquivos que dos 2.000 televisores, de determinada marca, vendidas em certo período, 400 precisaram de reparos dentro da garantia de um ano. Qual é a probabilidade de que um consumidor que compre uma televisão dessa marca não precise utilizar a garantia?

Resolução

Pelas informações, temos que 1.600 televisores não precisaram de reparos durante a garantia. Sendo o evento A: a televisão não precisa de reparo durante a garantia e utilizando a teoria frequencial, temos:

$$P(A) = \frac{1.600}{2.000} = \frac{4}{5} = 0,8$$

Ou seja, o consumidor tem uma probabilidade 0,8 de não precisar usar a garantia.

Utilizamos aqui o conhecimento histórico para fazer uma previsão, ou seja, utilizamos a frequência relativa do evento, obtida de dados coletados, para estimar a probabilidade.

4.10 Regras Básicas de Probabilidade

Sejam A e B dois eventos do espaço amostral Ω. Então:

a) $0 \leq P(A) \leq 1$
b) $P(\Omega) = 1$
c) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
d) Se A e B forem mutuamente exclusivos, então $P(A \cup B) = P(A) + P(B)$
e) $P(A^c) = 1 - P(A)$

Exemplo 4.12

Retira-se uma carta de um baralho completo de 52 cartas. Qual a probabilidade de sair um nove ou uma carta de paus?

Resolução

Neste experimento aleatório, temos que o espaço amostral é formado por 52 pontos amostrais, ou seja, $\Omega = \{A_o,...,K_o, A_E,... K_E, A_P,... K_P, A_C,...,K_c\}$. Vale relembrar que todos os pontos amostrais são equiprováveis, com isso podemos utilizar a definição clássica de probabilidade. Devemos observar também que o enunciado nos pede para encontrar a probabilidade do evento sair nove ou do evento sair carta de paus, o que caracteriza a união de dois eventos. Portanto devemos utilizar $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Os eventos definidos pelo exercício são:

A: saída de uma carta nove

$$A: \{9_O, 9_E, 9_C, 9_P\}, \text{ portanto, } P(A) = \frac{4}{52}$$

B: saída de uma carta de paus

$$B: \{A_P, 2_P, ..., K_P\}, \text{ portanto, } P(B) = \frac{13}{52}$$

$$A \cap B = \{9_P\}, \text{ portanto, } P(A \cap B) = \frac{1}{52}$$

Assim,

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

$$P(A \cup B) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$$

$$P(A \cup B) = \frac{16}{52}$$

Exemplo 4.13

Uma urna contém 30 bolas vermelhas, 12 bolas azuis e 7 bolas pretas. Extraindo-se aleatoriamente uma bola, qual a probabilidade de ser:

a) vermelha

$$P(V) = \frac{30}{49}$$

b) azul

$$P(A) = \frac{12}{49}$$

c) azul ou preta

$$P(A \cup P) = P(A) + P(P) = \frac{12}{49} + \frac{7}{49} = \frac{19}{49}$$

Aqui não utilizamos $P(A \cup P) = P(A) + P(P) - P(A \cap P)$, pois não há a intersecção $A \cap P$ (uma bola não pode ser azul e preta), ou seja, A e B são eventos mutuamente exclusivos.

d) nem azul nem vermelha

$$P(P) = \frac{7}{49}$$

Exemplos 4.14

Em um congresso científico existem 25 administradores e 17 matemáticos. Qual a probabilidade de ser formar uma comissão com 8 membros, na qual figurem 5 administradores e 3 matemáticos?

Resolução

Vamos definir o evento A como: a comissão é formada por 5 administradores e 3 matemáticos.

O número total de comissões que conseguimos formar com 8 membros é:

$$\binom{42}{8} = 118.030.185$$

Agora, o número de comissões que conseguimos formar com 5 administradores e 3 matemáticos é:

$$\binom{25}{5} \cdot \binom{17}{3} = 36.128.400$$

Portanto:

$$P(A) = \frac{36.128.400}{118.030.185} = 0,3061 \text{ x } 100 = 30,61\%$$

O cálculo das combinações foi feito utilizando uma calculadora científica.

Exemplo 4.15

Em uma prova caíram dois problemas. Sabe-se que 132 alunos acertaram o primeiro, 86 erraram o segundo, 120 acertaram os dois e 54 acertaram apenas um problema. Qual a probabilidade de que um aluno, escolhido ao acaso:

a) não tenha acertado nenhum problema?
b) tenha acertado apenas o segundo problema?

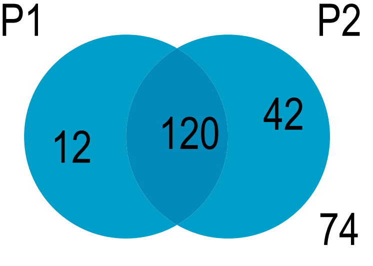
Resolução

Este exemplo é um exercício típico de cálculo de probabilidades envolvendo teoria dos conjuntos.

Do enunciado, temos:

- 120 acertaram os dois problemas;
- Se 120 acertaram os dois, 132 – 120 = 12 acertaram somente o primeiro;
- Se 54 acertaram apenas um e 12 acertaram somente o primeiro, 54 – 12 = 42 acertaram somente o segundo;
- A informação de que 86 erraram o segundo significa que esta quantidade acertou somente o primeiro ou não acertou nenhum deles. Se temos 12 que só acertaram o primeiro, 86 – 12 = 74 erraram os dois;
- Com esta análise conseguimos encontrar o número total de alunos que fizeram a prova: 120 + 12 + 42 + 74 = 248.

Podemos colocar estas informações no diagrama de Venn:



Então:

a) Definindo o evento A: o aluno não acertou nenhum problema, temos:

$$P(A) = \frac{74}{248} = 29{,}84\%$$

b) Definindo o evento B: o aluno acertou apenas o segundo problema, temos:

$$P(A) = \frac{42}{248} = 16,94\%$$

4.11 Probabilidade Condicional

Em muitas situações, podemos ter interesse em encontrar a probabilidade de ocorrência de um evento levando em conta que outro evento já ocorreu. Esta probabilidade recebe o nome de Probabilidade Condicional e é definida a seguir.

4.11.1 Definição de Probabilidade Condicional

Dados dois eventos A e B, a probabilidade condicional de A, dado que B ocorreu, é representada por P(A|B) e calculada por:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \qquad \text{com } P(B) > 0 \qquad (4.3)$$

Lemos a notação P(A|B) como a probabilidade de A ocorrer sabendo que B ocorreu.

Da definição apresentada obtemos a regra da multiplicação, de grande aplicação no cálculo de probabilidades, dada por:

$$P(A \cap B) = P(B).P(A \mid B) \qquad (4.4)$$

Exemplo 4.16

Há 100 candidatos para uma vaga numa empresa multinacional. Alguns têm curso superior, outros não. Alguns têm experiência no ramo, outros não. Os dados são:

	Possui curso superior	Não possui curso superior	Total
Com experiência anterior	35	45	80
Sem experiência anterior	15	5	20
Total	50	50	100

Considerando que o candidato escolhido para a vaga possui curso superior, qual a probabilidade de ele ter experiência anterior no ramo?

Resolução

Este exemplo se refere a um caso de probabilidade condicional, pois já sabemos que o candidato escolhido possui curso superior. Definindo os eventos e analisando o quadro temos:

A: ter experiência no ramo (definimos desta maneira, pois é a pergunta do exercício).

B: possui curso superior (definimos desta maneira, pois é o evento que sabemos que ocorreu).

$$P(B) = \frac{50}{100}$$

e

$$P(A \cap B) = \frac{35}{100}$$

Portanto:

$$P(A \mid B) = \frac{35/100}{50/100} = \frac{35}{50} = 0,7$$

Note que o valor que aparece no denominador é o total de casos do evento que sabemos que ocorreu. Neste exemplo sabíamos que o candidato escolhido tinha curso superior e o total de candidatos com este perfil é 50, justamente o valor que aparece no denominador do cálculo da probabilidade condicional.

Conforme TRIOLA (2008, pp. 138),

> [...] acreditar incorretamente que P(B|A) e P(A|B) sejam iguais ou usar um valor no lugar do outro é, às vezes, chamado confusão do inverso. Estudos mostram que médicos fornecem informações bastante enganosas quando eles confundem os inversos. Com base em estudos reais, eles tenderam a confundir P(câncer|teste positivo)

com P(teste positivo|câncer). Cerca de 95% dos médicos estimaram P(câncer|teste positivo) como cerca de 10 vezes mais alta, com a consequência de que os pacientes receberam diagnósticos enganosos e ficaram desnecessariamente angustiados pela informação incorreta.

4.12 Independência de eventos

Sejam A e B dois eventos do espaço amostral Ω, com $P(B) > 0$. O evento A é dito independente do evento B se

$$P(A \mid B) = P(A) \qquad (4.5)$$

ou seja, o evento A é independente do evento B se a probabilidade de A não é afetada pela ocorrência ou não de B.

Se A e B são independentes temos que:

$$P(A \cap B) = P(A).P(B) \qquad (4.6)$$

Exemplo 4.17

Numa caixa com 20 lâmpadas, 4 são defeituosas. Extraindo-se aleatoriamente duas lâmpadas, sem reposição, qual a probabilidade de:

a) nenhuma ser defeituosa;

b) ambas serem defeituosas.

c) considerando a extração das duas lâmpadas, com reposição, encontre a probabilidade de nenhuma delas ser defeituosa.

Resolução

Neste exemplo os eventos são dependentes, pois não há reposição das lâmpadas na caixa, ou seja, o resultado obtido na extração da segunda lâmpada é afetado pelo resultado obtido na primeira extração.

a) Nenhuma lâmpada ser defeituosa significa que as duas são perfeitas. Vamos indicar por P_1 primeira lâmpada ser perfeita e por P_2 segunda lâmpada ser perfeita. Como queremos encontrar a probabilidade da primeira ser perfeita e da segunda ser perfeita também devemos utilizar o conceito da regra da multiplicação,

$$P(P_1 \cap P_2) = P(P_1).P(P_2 \mid P_1)$$
$$P(P_1 \cap P_2) = \frac{16}{20}.\frac{15}{19} = \frac{240}{380} = 0,6316$$

b) Indicando por D_1 primeira lâmpada ser defeituosa e por D_2 segunda lâmpada ser defeituosa temos:

$$P(D_1 \cap D_2) = P(D_1).P(D_2 \mid D_1)$$
$$P(D_1 \cap D_2) = \frac{4}{20}.\frac{3}{19} = \frac{12}{380} = 0,0316$$

c) Vamos indicar por primeira lâmpada ser perfeita e por   segunda lâmpada ser perfeita. Neste caso, as duas extrações são independentes, pois o resultado da segunda extração não é afetado pelo primeiro resultado. Então:

$$P(P_1 \cap P_2) = P(P_1) \cdot (P_2)$$
$$P(P_1 \cap P_2) = \frac{16}{20} \cdot \frac{16}{20} = 64\%$$

Exemplo 4.18:

Para testar se um sistema especialista responde satisfatoriamente a um usuário, foram feitas cinco perguntas, cada uma com quatro alternativas de resposta. Se o sistema escolhe as alternativas aleatoriamente, qual é a probabilidade de ele responder corretamente a todas as cinco perguntas?

Resolução

Definindo os eventos:

P_1: acertar a pergunta 1
P_2 :acertar a pergunta 2
P_3: acertar a pergunta 3
P_4: acertar a pergunta 4
P_5: acertar a pergunta 5

Como o sistema escolhe as alternativas aleatoriamente, a probabilidade de ele acertar a pergunta é $\frac{1}{4}$ (pois de 4 alternativas, somente 1 é correta).

O fato do sistema ter escolhido uma alternativa (correta ou errada) em determinada pergunta não afeta as probabilidades de escolhas das outras perguntas. Portanto, os eventos são independentes. Então:

$$P(P_1 \cap P_2 \cap P_3 \cap P_4 \cap P_5) = P(P_1) \cdot P(P_2) \cdot P(P_3) \cdot P(P_4) \cdot P(P_5) = \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4}$$

$$= \frac{1}{1024} = 0{,}0977\%$$

Estudaremos, no próximo item, uma das relações mais importantes envolvendo probabilidades condicionais. A ideia central é atualizarmos valores prévios de probabilidades calculando as probabilidades adicionais, denominadas probabilidades posteriores.

4.13 Teorema de Bayes

Antes de apresentarmos a fórmula do teorema de Bayes, vamos analisar o exemplo a seguir.

Exemplo 4.19

Um certo programa pode ser usado com uma entre duas sub-rotinas A e B, dependendo do problema. A experiência tem mostrado que a sub-rotina A é usada 40% das vezes e B é usada 60% das vezes. Se A é usada, existe 75% de chance de que o programa chegue a um resultado dentro do limite de tempo. Se B é usada, a chance é de 50%. Se o programa foi realizado dentro do limite de tempo, qual a probabilidade de que a sub-rotina A tenha sido a escolhida?

Resolução

A probabilidade que temos que encontrar é uma probabilidade condicional, pois já sabemos que o programa foi realizado dentro do limite de tempo.

Então, de acordo com o que estudamos no item 4.11.1, podemos definir os seguintes eventos:

A: sub-rotina A foi a escolhida.

R: o programa foi realizado dentro do limite de tempo.

Pela fórmula da probabilidade condicional, queremos encontrar:

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

Agora, a interpretação da probabilidade que está no denominador é fundamental para o entendimento deste teorema.

A probabilidade do evento R ocorrer está associada à utilização de duas sub-rotinas.

Sabemos que o programa foi realizado dentro do limite de tempo, mas não sabemos qual sub-rotina foi utilizada. Com isto, precisamos considerar estas duas situações:

$R \cap A \rightarrow$ foi realizado dentro do limite de tempo e utilizada a rotina A;

$R \cap B \rightarrow$ foi realizado dentro do limite de tempo e utilizada a rotina B.

Então, o cálculo de $P(R)$ é dado por:

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

ou seja, a probabilidade do evento R ocorrer é a soma das duas situações possíveis.

As probabilidades informadas no enunciado são:

$P(A) = 0{,}40 \rightarrow$ probabilidade obtida através da definição frequencial;

$P(B) = 0{,}60 \rightarrow$ probabilidade obtida através da definição frequencial;

$P(R|A)=0{,}75 \rightarrow$probabilidade do programa ser realizado dentro do tempo,se (sabendo que) A é usada;

$P(R|B) = 0{,}50 \rightarrow$ probabilidade do programa ser realizado dentro do tempo,se (sabendo que) B é usada.

Como vamos encontrar $P(R \cap A)$ e $P(R \cap B)$? Utilizaremos a regra da multiplicação, abordada no item 4.11.1.

Esta regra nos diz que a probabilidade da intersecção dos dois eventos pode ser obtida da seguinte maneira:

$$P(R \cap A) = P(R|A) \cdot P(A)$$

e

$$P(R \cap B) = P(R|B) \cdot P(B)$$

As probabilidades que estão do lado direito da igualdade são fornecidas no enunciado. Então:

$$P(R \cap A) = P(R \mid A) \cdot P(A) = 0{,}75 \cdot 0{,}40 = 0{,}30$$

e

$$P(R \cap B) = P(R \mid B) \cdot P(B) = 0{,}50 \cdot 0{,}60 = 0{,}30$$

Finalmente, conseguiremos calcular a probabilidade procurada:

$$P(A \mid R) = \frac{P(A \cap R)}{P(R)} = \frac{P(A \cap R)}{P(R \cap A) + P(R \cap B)} = \frac{P(R \mid A) \cdot P(A)}{P(R \mid A) \cdot P(A) + P(R \mid B) \cdot P(B)}$$

$$P(A \mid R) = \frac{0{,}30}{0{,}30 + 0{,}30} = \frac{0{,}30}{0{,}60} = 0{,}5 = 50\%$$

No Teorema de Bayes, a quantidade do numerador sempre será um dos termos que está no denominador. Vale ressaltar também que $P(A \cap B) = P(B \cap A)$.

Agora, vamos formalizar o Teorema de Bayes.

Seja $C_1, C_2, \ldots, C_n$ uma partição do espaço amostral Ω, isto é,

- $C_i \cap C_j = \varnothing$, sempre que $i \neq j$,
- $C_1 \cup C_2 \cup \ldots \cup C_n = \Omega$.

Considere um evento qualquer A em Ω. Supomos conhecidas as probabilidades $P(C_i)$ e $P(A \mid C_i)$, $i = 1,2,\ldots,n$. A probabilidade de ocorrência do evento C_i, supondo-se a ocorrência do evento A, é dada por:

$$P(C_i \mid A) = \frac{P(C_i) \cdot P(A \mid C_i)}{\sum_{j=1}^{n} P(C_j) \cdot P(A \mid C_j)}$$

para todo $i = 1,2,\ldots,n$.

Atividades

01. Uma biblioteca acaba de receber, por doação, 40 novos livros, inclusive 15 romances históricos. Se 3 desses livros são escolhidos aleatoriamente, sem reposição, qual é a probabilidade de:

a) nenhum ser um romance histórico;
b) todos serem romances históricos;
c) pelo menos um ser romance histórico.

02. Como parte de uma campanha de promoção em São Paulo e no Rio de Janeiro, uma indústria de produtos de limpeza oferecerá um prêmio de R$ 50.000,00 a quem enviar seu nome em um formulário, com a opção de incluir um rótulo de um dos produtos da indústria. A distribuição dos 200.000 formulários recebidos está a seguir:

	Com rótulo	Sem rótulo
São Paulo	100.000	40.000
Rio de Janeiro	45.000	15.000

Escolhendo aleatoriamente um dos formulários e definindo os eventos A: o formulário escolhido é de São Paulo e B: o formulário escolhido tem um rótulo do produto, determine as seguintes probabilidades:

a) $P(A)$
b) $P(B)$
c) $P(A \mid B)$
d) $P(A^c \mid B^c)$
e) $P(A^c \mid B^c)$
f) $P(B \mid A^c)$

03. Uma loja de cosméticos tem os seguintes dados sobre a idade e o estado civil de 150 clientes.

	Estado Civil		
		Solteiro	Casado
Idade	≤30	70	20
	mais de 30	30	30

Selecionando aleatoriamente a ficha de um cliente, determinar:

a) probabilidade deste cliente ser solteiro;

b) probabilidade deste cliente ter mais de 30 anos;

c) se na ficha consta que o cliente é solteiro, qual é a probabilidade de ele ter mais de 30 anos;

d) probabilidade deste cliente ser casado sabendo que ele tem menos de 30 anos.

04. Uma agência de locação de carros fez um levantamento sobre o perfil dos seus clientes e obteve os seguintes resultados: 45% haviam alugado um carro no último ano por razões de negócios, 50% haviam alugado um carro no último ano por razões pessoais e 20% haviam alugado um carro no último ano tanto por razões de negócios como por razões pessoais.

a) Qual a probabilidade de que um cliente tenha alugado um carro durante o último ano por razões pessoais ou por razões de negócios?

b) Qual a probabilidade de que um cliente não tenha alugado um carro durante o último ano nem por razões pessoais nem por razões de negócios?

05. Um satélite em órbita tem três painéis solares, e todos eles devem permanecer ativos a fim de garantir o bom desempenho do aparelho. Os painéis funcionam independentemente uns dos outros. A chance de falha de cada um é 0,01. Qual a probabilidade de o satélite funcionar perfeitamente durante a missão? (Essa probabilidade é a chamada confiabilidade do sistema – Farias, Soares e César, pág. 65)

06. Uma rede local de computadores é composta por um servidor e cinco clientes (A, B, C, D e E). Registros anteriores indicam que dos pedidos de determinado tipo de processamento, realizados através de uma consulta, cerca de 10% vêm do cliente A, 15% do B, 15% do C, 40% do D e 20% do E. Se o pedido não for feito de forma adequada, o processamento apresentará erro. Usualmente, ocorrem os seguintes percentuais de pedidos inadequados: 1% do cliente A, 2% do cliente B, 0,5% do cliente C, 2% do cliente D e 8% do cliente E.

a) Qual é a probabilidade de o sistema apresentar erro?

b) Qual é a probabilidade de que o processo tenha sido pedido pelo cliente E, sabendo-se que apresentou erro?

Reflexão

O estudo de probabilidades, que iniciamos neste capítulo, tem muitas aplicações no dia a dia do gestor. Até o momento, nos preocupamos em apresentar a teoria referente ao assunto. No entanto, as aplicações que podemos fazer do cálculo de probabilidades são muito diversificadas: determinar a margem de erro e o grau de confiança de uma pesquisa, fazer previsões (com certo grau de confiança) de eventos futuros, auxiliar na tomada de decisões, calcular riscos de certos investimentos, etc.

Procure assimilar bem todos os procedimentos e conceitos apresentados neste capítulo para que possa acompanhar o desenvolvimento dos métodos que serão apresentados mais adiante.

Leitura recomendada

Recomendamos a leitura do texto “É possível quantificar o acaso?”, disponível no endereço http://www.klick.com.br/materia/20/display/0,5912,POR-20-89-957-,00.html, que apresenta uma interessante situação sobre o estudo de probabilidades.

Referências

ANDERSON, David R.; SWEENEY, Denis J.; WILLIAMS, Thomas A**. Estatística aplicada à administração e economia**. São Paulo: Pioneira Thomson Learning, 2003.

BARBETTA, Pedro A.; REIS, Marcelo M.; BORNIA, Antonio C. **Estatística**: para os cursos de engenharia e informática. São Paulo: Atlas, 2004.

BUSSAB, Wilton de O.; MORETTIN, Pedro A.. **Estatística básica**. São Paulo: Saraiva, 2003.

FARIAS, Alfredo Alves de; SOARES, José Francisco; CÉSAR, Cibele Comini. **Introdução à estatística**. Rio de Janeiro,: LTC, 2003.

FARBER, Larson. **Estatística aplicada**. São Paulo: Pearson Prentice Hall, 2004.

IEZZI, Gelson; DOLCE, Osvaldo; DEGENSZAJN David; PÉRIGO, Roberto. **Matemática**: volume único. São Paulo: Atual, 2007.

MORETTIN, Luiz G. **Estatística Básica** – Volume I – Probabilidade. 7.ed. São Paulo: Makron Books, 1999.

TOMAZ, Priscilla S.S. Anais do IX **Seminário Nacional de História da Matemática**. Disponível em: <http://www.each.usp.br/ixsnhm/Anaisixsnhm/Comunicacoes/1_Tomaz_P_S_S_Gerolamo_Cardano.pdf>. Acesso em: 30 set. 2014.

No próximo capítulo

No próximo capítulo apresentaremos as variáveis aleatórias, que são uma forma de associar valores aos resultados do experimento aleatório. Isso irá nos permitir ampliar a capacidade de aplicação da teoria de probabilidades.

Minhas anotações:

Capítulo 5

Variáveis aleatórias

O espaço amostral, em muitos experimentos, não consiste só em números como, por exemplo, o espaço amostral referente ao lançamento de uma moeda, que tem como pontos amostrais cara ou coroa. Em Estatística, muitas vezes, estamos interessados em resultados numéricos. Para transformar os resultados do espaço amostral em números utilizamos o conceito de variável aleatória.

Objetivos da sua aprendizagem

Que você seja capaz calcular o valor esperado e o desvio padrão de uma variável aleatória discreta e que consiga identificar as situações nas quais podemos aplicar o modelo de probabilidade binomial, bem como calcular as probabilidades associadas a tal modelo.

Você se lembra?

Você se lembra dos conceitos de média e desvio-padrão apresentados nos capítulos 2 e 3? E dos conceitos de probabilidade abordados no capítulo 4? Neste capítulo, combinaremos aqueles conceitos ao desenvolvermos as distribuições de probabilidade, que descrevem o que provavelmente acontecerá, ao invés do que realmente aconteceu.

5.1 Variável Aleatória

Uma variável aleatória (v.a.) é uma variável que associa um valor numérico a cada ponto do espaço amostral. Ela é denominada discreta quando pode assumir apenas um número finito ou infinito enumerável de valores e é dita contínua quando assume valores num intervalo da reta real.

É comum utilizarmos letras latinas para representarmos variáveis aleatórias.

5.2 Função discreta de probabilidade

Função Discreta de Probabilidade é a função que atribui a cada valor da v.a. sua probabilidade, ou seja,

$$P(X = x_i) = p(x_i), \qquad i = 1, 2, ..., n \tag{5.1}$$

Vamos considerar aqui que a v.a. discreta tem um número finito de valores possíveis.

A distribuição de probabilidade de uma v.a. discreta X é uma tabela que associa a cada valor de X sua probabilidade.

x	p(x)
x_1	$p(x_1)$
x_2	$p(x_2)$
x_3	$p(x_3)$
.	.
.	.
.	.
x_n	$p(x_n)$

Tabela 5.1 – Distribuição de probabilidade da v.a. X

Na tabela 5.1, os valores $x_1, x_2, x_3,..., x_n$ são aqueles que a v.a. pode assumir e $p(x_1)$, $p(x_2), p(x_3),..., p(x_n)$ suas respectivas probabilidades.

Uma distribuição de probabilidade deve satisfazer as seguintes condições:

$$0 \le p(x_i) \le 1, \qquad i = 1,2,...,n$$

$$\sum_{i=1}^{n} p(x_i) = 1$$

Exemplo 5.1: Vamos considerar o experimento aleatório que consiste no lançamento de três moedas. O espaço amostral deste experimento é:

Ω = {(c, c, c), (c, c, r), (c, r, c), (c, r, r), (r, c, c), (r, c, r), (r, r, c), (r, r,r)} onde c = cara e r = coroa.

Podemos definir a variável aleatória de interesse como sendo o número de coroas obtidas no lançamento das três moedas, ou seja, X: número de coroas. De acordo com a definição da variável aleatória podemos associar a cada ponto amostral um número, como mostra o quadro seguinte:

Resultados	X
c,c,c	0
c,c,r	1
c,r,c	1
c,r,r	2
r,c,c	1
r,c,r	2
r,r,c	2
r,r,r	3

Vemos que a cada resultado do experimento está associado um valor da v.a. X, a saber, 0, 1, 2 e 3. Temos que:

- X = 0, com probabilidade 1/8 se, e somente se, ocorre o resultado c,c,c;
- X = 1, com probabilidade 1/8 + 1/8 +1/8 = 3/8 se, e somente se, ocorrem os resultados c,c,r ou c,r,c ou r,c,c, que são mutuamente exclusivos;
- X = 2 com probabilidade 1/8 + 1/8 + 1/8 = 3/8 se, e somente se, ocorrem os resultados c,r,r ou r,c,r ou r,r,c, que são mutuamente exclusivos;
- X = 3 com probabilidade 1/8 se, e somente se, ocorre o resultado r,r,r.

Na tabela 5.2 apresentamos a distribuição de probabilidade da v.a. X.

x	p(x)
0	1/8
1	3/8
2	3/8
3	1/8

Tabela 5.2 – Distribuição de probabilidade da v. a. X = número de coroas

A distribuição de probabilidade satisfaz as condições $0 \leq p(x_i) \leq 1$ e $\sum_{i=1}^{4} p(x_i) = 1$, para i = 1, 2, 3, 4.

Exemplo 5.2: Uma companhia analisa diariamente o número de vendas de seus novos funcionários durante um período de experiência de 90 dias. Os resultados para um novo funcionário são apresentados a seguir.

Vendas por dia	Número de dias
0	15
1	17
2	12
3	18
4	8
5	10
6	9
7	1

Tabela 5.3: Distribuição do número de vendas do novo funcionário

a) Obtenha a probabilidade de cada resultado

b) Organize os dados em uma distribuição de probabilidade.

Resolução

a) Para encontrarmos a probabilidade de cada resultado, vamos usar a definição frequencial de probabilidade, ou seja, as probabilidades nada mais são que as frequências relativas de cada resultado:

Vendas por dia	Número de dias	Probabilidade
0	15	0,1667
1	17	0,1889
2	12	0,1333
3	18	0,2000
4	8	0,0889
5	10	0,1111
6	9	0,1000
7	1	0,0111

A primeira probabilidade foi encontrada fazendo $\frac{15}{90}$, a segunda probabilidade é $\frac{17}{90}$ e assim por diante.

b) Uma distribuição de frequência é uma tabela que contém os resultados da variável aleatória com suas respectivas probabilidades, ou seja:

Vendas por dia	Probabilidade
0	0,1667
1	0,1889
2	0,1333
3	0,2000
4	0,0889
5	0,1111
6	0,1000
7	0,0111
Total	**1**

Tabela 5.4: Distribuição de probabilidade do número de vendas por dia.

5.3 Valor esperado e variância de uma variável aleatória discreta

Para as distribuições de probabilidade podemos definir as mesmas medidas de tendência central e de dispersão estudadas nas distribuições de frequências.

A média de uma v.a. X também chamada de valor esperado ou esperança matemática é representada por E(X) e definida como:

$$E(X) = \sum_{i=1}^{n} x_i \cdot p(x_i) \qquad (5.2)$$

Observamos, pela definição de E(X), que para se calcular a média de uma v.a. precisamos multiplicar cada valor da v.a. por sua correspondente probabilidade e somar os produtos resultantes.

Podemos interpretar o valor esperado de uma v.a. como uma média ponderada dos x_i, onde os pesos são as probabilidades associadas.

Propriedades da média

Sejam a e b constantes e X uma variável aleatória. Então:

i) E(a) = a

ii) E(bX) = bE(X)

iii) E(X + a) = E(X) + a

iv) E(a + bX) = a + bE(X)

A variância de uma v.a. X é definida como:

$$\mathrm{Var}(X) = E\left(X^2\right) - \left[E(X)\right]^2 = \sum_{i=1}^{n} x_i^2 \cdot p(x_i) - \left[\sum_{i=1}^{n} x_i \cdot p(x_i)\right]^2 \qquad (5.3)$$

Já vimos anteriormente que o desvio-padrão (s) é a raiz quadrada da variância, portanto:

$$\sigma(x) = \sqrt{\mathrm{Var}(x)} \qquad (5.4)$$

Propriedades da variância

Sejam a e b constantes e X uma variável aleatória. Então:

i) Var(a) = 0

ii) Var(bX) = b^2Var(X)

iii) Var(X + a) = Var(X)

iv) Var(a + bX) = b^2Var(X)

Observação: Indicaremos a média e a variância de uma v.a. X por:

$$E(X) = \mu$$

$$\mathrm{Var}(X) = \sigma^2$$

Exemplo 5.3: Utilizando os dados do exemplo 5.1, vamos calcular a média e a variância da v.a. X: número de coroas.

x	p(x)	x · p(x)	x^2 · p(x)
0	1/8	0	0
1	3/8	3/8	3/8
2	3/8	6/8	12/8
3	1/8	3/8	9/8
Total	1	12/8	24/8

Resolução

Substituindo os valores do quadro apresentado nas respectivas fórmulas temos:

$$E(X)=\sum_{i=1}^{n} x_i \cdot p(x_i)=\frac{12}{8}=1,5$$

$$\begin{aligned}Var(X)&=E\left(X^2\right)-\left[E(X)\right]^2=\sum_{i=1}^{n} x_i^2 \cdot p(x_i)-\left[\sum_{i=1}^{n} x_i \cdot p(x_i)\right]^2\\&=3-(1,5)^2\\&=0,75\end{aligned}$$

Portanto, temos que o valor esperado do número de coroas, obtido no lançamento de 3 moedas, é 1,5 e o desvio-padrão é 0,866 $\left(\sqrt{0,75}\right)$.

Exemplo 5.4

Num teste de digitação, o tempo em minutos (T) que os candidatos levam para digitar um texto é modelado, de forma aproximada, pela seguinte função de probabilidade:

T	P(T)
3	0,1
4	0,1
5	0,2
6	0,2
7	0,2
8	0,1
9	0,1

O candidato recebe 4 pontos se faz em 9 minutos, 5 se faz em 8 minutos e assim por diante. Determine a média e a variância do número de pontos obtidos no teste.

Resolução

Como o exercício está pedindo a média e a variância do número de pontos obtidos, precisamos definir uma nova variável aleatória:

P: número de pontos obtidos no teste

O número de pontos obtidos depende do tempo que o candidato leva para digitar o texto. De acordo com o texto, se o candidato digitar o texto em 3 minutos, ele recebe 10 pontos. E, a probabilidade dele receber 10 pontos é a mesma probabilidade dele digitar em 3 minutos, ou seja, 0,1. Portanto:

T	P(T)
10	0,1
9	0,1
8	0,2
7	0,2
6	0,2
5	0,1
4	0,1

Definida a função distribuição de probabilidade da variável P, encontramos a média e a variância acrescentando duas colunas no quadro acima:

P	P(P)	$P \cdot P(P)$	$P^2 \cdot P(P)$
10	0,1	1	10
9	0,1	0,9	8,1
8	0,2	1,6	12,8
7	0,2	1,4	9,8
6	0,2	1,2	7,2
5	0,1	0,5	2,5
4	0,1	0,4	1,6
Total	**1**	**7**	**52**

Agora, basta substituirmos os valores encontrados nas respectivas fórmulas:

$$E(P) = \sum_{i=1}^{n} p_i \cdot p(p_i) = 7$$

$$\mathrm{Var}(P) = E(P^2) - [E(P)]^2 = \sum_{i=1}^{n} p_i \cdot p(p_i) - \left[\sum_{i=1}^{n} p_i \cdot p(p_i)\right]^2$$

$$= 52 - (7)^2 = 3$$

Então, o número médio de pontos obtidos no teste é 7 e a variância é 3 pontos^2. Para encontrarmos o desvio padrão, basta extrairmos a raiz quadrada da variância.

Exemplo 5.5

Na produção de uma peça são empregadas duas máquinas. A primeira é utilizada para efetivamente produzir as peças, e o custo de produção é de R$ 50,00 por unidade. Das peças produzidas nessa máquina, 90% são perfeitas. As peças defeituosas (produzidas na primeira máquina) são colocadas na segunda máquina para a tentativa de recuperação (torná-las perfeitas). Nessa segunda máquina o custo por peça é de R$ 25,00, mas apenas 60% das peças são de fato recuperadas. Sabendo que cada peça perfeita é vendida por R$ 90,00, e que cada peça defeituosa é vendida por R$ 20,00, calcule o lucro por peça esperado pelo fabricante.

Resolução

Neste exemplo, a variável aleatória de interesse é:

X: lucro

Para entendermos melhor as informações contidas no enunciado, vamos montar um quadro:

Custo	Venda	X (lucro)	P(X)	
50	90	40	0,9	peças perfeitas na 1ª máquina
50+25	90	15	0,06	peças perfeitas na 2ª máquina
50+25	20	-55	0,04	peças defeituosas

As informações contidas na primeira linha nos indicam que o lucro será de R$ 40,00, pois Lucro = Venda – Custo. Noventa por cento das peças produzidas na primeira máquina são perfeitas. Então, a probabilidade do lucro assumir o valor R$ 40,00 é de 0,9.

Na segunda linha estão as informações das peças que precisaram ir para a segunda máquina e que tornaram-se perfeitas. O custo, nesta situação, será de R$ 75,00, pois o custo por peça na segunda máquina é de R$ 25,00, adicionados aos R$ 50,00 de custo da primeira máquina. Agora, vamos entender o porquê do valor da probabilidade ser 0,06. Se 90% das peças produzidas pela primeira máquina são perfeitas, 10% são defeituosas.

Destas 10% de peças defeituosas, apenas 60% são recuperadas pela segunda máquina. Portanto, 0,10*0,6 = 0,06. Então, a probabilidade do lucro assumir o valor R$ 15,00 é de 0,06.

E, na terceira linha, temos a informação das peças defeituosas. Neste caso, a empresa terá prejuízo, pois o custo continuará sendo de R$ 75,00, e a peça será vendida por R$ 20,00. A probabilidade do lucro (que será um prejuízo) assumir o valor de R$ –55,00 é 0,04, pois das peças defeituosas, 40% não são recuperadas pela segunda máquina. Portanto, 0,10*0,4 = 0,04.

Agora, com os valores obtidos no quadro, conseguimos encontrar o lucro por peça esperado pelo fabricante.

Custo	Venda	X · P(x)
40	0,9	40 × 0,9 = 36,00
15	0,06	15 × 0,06 = 0,90
-55	0,04	– 55 × 0,04= –2,20
Total	1	34,70

O lucro esperado, por peça, é R$ 34,70.

5.4 Modelos probabilísticos para variáveis aleatórias Discretas

Algumas variáveis aleatórias aparecem com frequência em muitas situações práticas do nosso dia a dia. Um estudo detalhado dessas variáveis é muito importante para a construção de modelos probabilísticos com o objetivo de estimar seus parâmetros e calcular probabilidades. Uma das distribuições discretas de probabilidade mais importante é a distribuição binomial que será descrita a seguir.

5.4.1 Distribuição Binomial

Uma v.a. tem distribuição binomial se o experimento aleatório consiste em:

- n tentativas sob condições idênticas;
- cada tentativa é independente de todas as outras;
- há somente dois resultados possíveis em cada tentativa designados por sucesso(S) e fracasso(F);
- a probabilidade de sucesso é a mesma em cada tentativa.

Na distribuição binomial a v.a. X corresponde ao número de sucessos em n tentativas do experimento aleatório.

A função de probabilidade é definida como:

$$P(X = k) = \binom{n}{x} \cdot p^{k} \cdot q^{n-k} \qquad (5.5)$$

onde:

n é o número de tentativas do experimento aleatório;

p = P(S) é a probabilidade de sucesso em uma única tentativa;

q = P(F) é a probabilidade de fracasso em uma única tentativa;

$p + q = 1$.

$\binom{n}{k}$ é denominado número binomial e é dado pela fórmula $\binom{n}{k} = \frac{n!}{(n-k)!k!}$. Os termos n! e k! são denominados n fatorial e k fatorial e são dados pela multiplicação de todos os valores inteiros positivos entre entre 1 e n e entre 1 e k.

Por exemplo, $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$.

Quando a v.a. X tiver distribuição binomial, com parâmetros n e p, indicaremos por $X \sim b(n,p)$.

O valor esperado e a variância de uma v.a. binomial, com parâmetros n e p são, respectivamente, dados por:

$$\begin{aligned} E(X) &= n \cdot p \\ Var(X) &= n \cdot p \cdot q \end{aligned} \qquad (5.6)$$

Exemplo 5.6: Uma pesquisa mostrou que 60% das famílias residentes na grande São Paulo têm pelo menos dois carros. Determine a probabilidade de que dentre 15 famílias selecionadas aleatoriamente nesta região:

a) exatamente 5 tenham pelo menos dois carros;
b) de 8 a 10 tenham pelo menos dois carros.

Neste exemplo vamos definir a v.a. como tendo distribuição binomial, pois:

- o experimento está sendo realizado 15 vezes, ou seja, 15 famílias foram selecionadas para o estudo;
- há somente dois resultados possíveis: **sucesso**, se a família tem pelo menos dois carros e **fracasso**, se a família não tem pelo menos dois carros;
- as respostas são independentes umas das outras, ou seja, uma família ter pelo menos dois carros não afeta a probabilidade das outras famílias terem ou não pelo menos dois carros.

O primeiro passo para iniciar a resolução de problemas deste tipo é definir a v.a.. Neste caso, como estamos interessados no número de famílias com pelo menos dois carros, a v.a. é definida como:

X: número de famílias com pelo menos dois carros.

A v.a. X pode assumir os valores 0,1,2,...,15, pois, das 15 famílias selecionadas, pode acontecer de nenhuma ter pelo menos dois carros, 1 pode ter pelo menos dois carros, 2 podem ter pelo menos dois carros , assim por diante, até as 15 famílias com pelo menos dois carros.

a) Este item pede a probabilidade de que exatamente 5 tenham pelo menos dois carros, ou seja,

$$P(X = 5) = \binom{15}{5} \cdot (0,60)^5 \cdot (0,40)^{10}$$

e o número de tentativas é n = 15 famílias selecionadas. com P(sucesso) = 0,60, pois o enunciado nos informa que 60% das famílias têm pelo menos dois carros, P(fracasso) = 0,40

Fazendo os cálculos chegamos que

$$P(X=5)=\binom{15}{5}\cdot(0,60)^{5}\cdot(0,40)^{10}$$
$$=0,0245$$

b) A probabilidade pedida neste item pode ser escrita como:

$$\begin{aligned}P(8\le X\le 10)&=P(X=8)+P(X=9)+P(X=10)\\&=\binom{15}{8}\cdot(0,60)^{8}\cdot(0,40)^{7}+\binom{15}{9}\cdot(0,60)^{9}\cdot(0,40)^{6}+\binom{15}{10}\cdot(0,60)^{10}\cdot(0,40)^{5}\\&=0,177083662+0,206597605+0,185937845\\&=0,5696\end{aligned}$$

Observação: Estes cálculos são facilmente efetuados com o auxílio de uma calculadora científica.

Exemplo 5.7

Se 7% das peças produzidas por uma máquina são defeituosas, qual a probabilidade de que em dez peças escolhidas aleatoriamente:

a) não haja peças defeituosas;
b) pelo menos 3 peças sejam defeituosas;
c) exatamente 5 peças sejam defeituosas;
d) entre 2 e 4 peças sejam defeituosas.

Resolução

Temos aqui um experimento binomial com:

X: número de peças defeituosas

X = 0,1,2,3,...,10.

n = 10 peças

P(Sucesso) = p = 0,07

P(Fracasso) = q = 0,93

Note que a probabilidade de sucesso é que a peça seja defeituosa, pois a variável aleatória está definida como o número de peças defeituosas.

$$a)\ P(X = 0) = \binom{10}{0} \cdot (0,07)^0 \cdot (0,93)^{10} = 0,4840$$

$$b)\ P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5) + \ldots + P(X = 10)$$

Neste caso podemos simplificar os cálculos utilizando o complementar do evento, isto é,

$$P(X \geq 3) = 1 - P(X < 3) = 1 - \left[P(X = 0) + P(X = 1) + P(X = 2)\right]$$
$$= 1 - \left[\binom{10}{0} \cdot (0,07)^0 \cdot (0,93)^{10} + \binom{10}{1} \cdot (0,07)^1 \cdot (0,93)^9 + \binom{10}{2} \cdot (0,07)^2 \cdot (0,93)^8\right]$$
$$= 1 - [0,483982307 + 0,364287758 + 0,123387789]$$
$$= 1 - [0,971657854]$$
$$= 0,0283$$

$$c)\ P(X = 5) = \binom{10}{5} \cdot (0,07)^5 \cdot (0,93)^5 = 0,0003$$

$$d)\ P(2 \leq X \leq 4) = P(X = 2) + P(X = 3) + P(X = 4)$$
$$= \binom{10}{2} \cdot (0,07)^2 \cdot (0,93)^8 + \binom{10}{3} \cdot (0,07)^3 \cdot (0,93)^7 + \binom{10}{4} \cdot (0,07)^4 \cdot (0,93)^6$$
$$= 0,123387789 + 0,024766008 + 0,003262189$$
$$= 0,1514$$

Conexão:

O cálculo de probabilidades para variáveis aleatórias cuja distribuição é binomial também pode ser feita no Excel. Para compreender a função que deve ser utilizada, leia "Utilizando o Microsoft Excel para Obter Probabilidades Binomiais", que se encontra no livro Estatística: Teoria e Aplicações Usando Microsoft Excel em Português, pp. 196.

Atividades

01. Uma urna contém 3 bolas brancas e 7 bolas verdes. Três bolas são retiradas com reposição. Seja X: número de bolas verdes. Calcule E(X) e Var(X).

02. Uma companhia aérea tem as probabilidades

0,05	0,20	0,35	0,15	0,10	0,15

de receber 0, 1, 2, 3, 4 ou 5 reclamações sobre desvio de bagagem por dia. Quantas reclamações a companhia espera (valor esperado ou média) receber por dia?

03. Um comerciante tem a oportunidade de adquirir um embarque de seda pura por R$ 30.000,00. A probabilidade de ele vender essa seda por R$ 26.000,00 é de 0,40 e a probabilidade de ele vendê-la por R$ 35.000,00 é de 0,60. Qual é o lucro bruto esperado do comerciante?

04. O tempo T, em minutos, necessário para um operário processar certa peça é uma v.a. com a seguinte distribuição de probabilidade:

T	2	3	4	5	6	7
p(t)	0,1	0,1	0,3	0,2	0,2	0,1

Fonte: Bussab e Morettin, pag. 140.

a) Calcule o tempo médio de processamento.

Para cada peça processada, o operário ganha um fixo de R$ 2,00, mas, se ele processa a peça em menos de seis minutos, ganha R$ 0,50 em cada minuto poupado. Por exemplo, se ele processa a peça em quatro minutos, recebe a quantia adicional de R$ 1,00.

b) Encontre a distribuição, a média e a variância da v.a. G: quantia em R$ ganha por peça.

05. Um curso de treinamento aumenta a produtividade dos funcionários da área de atendimento ao consumidor em 80% dos casos. Se quinze funcionários participam desse curso, qual a probabilidade de:

a) exatamente quatro funcionários aumentarem a produtividade;
b) de 5 a 7 funcionários aumentarem a produtividade;
c) pelo menos dois funcionários não aumentarem a produtividade.

06. Um lote com máquinas digital é recebida por uma empresa. 30 aparelhos são inspecionados. O lote é rejeitado se pelo menos 3 máquinas apresentarem defeito. Sabendo-se que 1% das máquinas é defeituosa, calcule a probabilidade da empresa rejeitar todo o lote.

07. Um vendedor de seguros vende apólices a 5 homens, todos da mesma idade e de boa saúde. De acordo com as tabelas atuariais, a probabilidade de um homem, dessa idade particular, estar vivo daqui a 30 anos é de 2/3. Determinar a probabilidade de estarem ainda vivos daqui a 30 anos:

a) todos os 5 homens;
b) pelo menos 3;
c) apenas 2;
d) pelo menos 1 homem.

08. A probabilidade de um estudante, que ingressa em uma universidade, graduar-se, é de 0,4. Determinar a probabilidade de, entre 5 estudantes:

a) nenhum graduar-se;
b) um graduar-se;
c) pelo menos um graduar-se.

09. Um aluno marca ao acaso as respostas em um teste múltipla escolha com 10 questões e cinco alternativas por questão. Qual a probabilidade dele acertar exatamente 4 questões?

10. Uma pesquisa concluiu que 30% das mulheres brasileiras consideram a leitura sua atividade favorita de lazer. Você seleciona ao acaso seis mulheres e pergunta a elas se a leitura é sua atividade favorita de lazer. Obtenha a probabilidade de que:

a) exatamente duas delas respondam “sim”.
b) menos do que duas respondam “sim”.

Reflexão

Neste capítulo vimos que a variável aleatória fornece uma descrição numérica de um experimento aleatório. Com os resultados que uma variável aleatória pode assumir juntamente com suas respectivas probabilidade obtemos a distribuição de probabilidade e podemos calcular o valor esperado (média) e o desvio padrão para a variável aleatória. Podemos interpretar o valor esperado como uma média ponderada dos valores que a variável aleatória pode assumir. Os pesos são as probabilidades. Estudamos, também, uma distribuição discreta de probabilidade que tem muitas aplicações: distribuição binomial. Desde que o experimento em estudo satisfaça os requisitos necessários para que a variável aleatória tenha distribuição binomial, encontramos facilmente, com o auxílio de uma calculadora científica, as probabilidades de interesse.

Leitura Recomendada

Sugerimos que você assista ao vídeo "Revendo a Moratória", da série: Matemática na Escola. Ele apresenta uma aplicação do conceito de valor esperado. O endereço para acesso é http://m3.ime.unicamp.br/recursos/1170. Vale a pena conferir!

Referências

ANDERSON, David R.; SWEENEY, Denis J.; WILLIAMS, Thomas A. **Estatística aplicada à administração e economia**. São Paulo: Pioneira Thomson Learning, 2003.

BUSSAB, Wilton de O.; MORETTIN, Pedro A.. **Estatística básica.** São Paulo: Saraiva, 2003.

FARIAS, Alfredo Alves de; SOARES, José Francisco; CÉSAR, Cibele Comini. **Introdução à estatística**. Rio de Janeiro,: LTC, 2003.

FARBER, Larson. **Estatística aplicada**. São Paulo: Pearson Prentice Hall, 2004.

LEVINE, David M.; BERENSON, Mark L.; STEPHAN, David. **Estatística: Teoria e Aplicações Usando Microsoft Excel em Português**. Rio de Janeiro: LTC, 2000.

MAGALHÃES, Marcos Nascimento; LIMA, Antônio Carlos Pedroso de. **Noções de probabilidade e estatística**. São Paulo: Editora da Universidade de São Paulo, 2004.

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

Gabarito

Capítulo 1

1.

a) qualitativa nominal

b) quantitativa discreta

c) quantitativa contínua

d) quantitativa discreta

e) quantitativa contínua

f) quantitativa contínua

g) quantitativa discreta

h) qualitativa nominal

i) quantitativa contínua

j) quantitativa contínua

k) qualitativa ordinal

l) qualitativa ordinal

m) qualitativa nominal

n qualitativa nominal

2.

K ≈ 6 classes e amplitude da classe h ≈ 7

Tabela 1 – Distribuição de frequências das idades dos funcionários.

Idades	f	f_r	f_a
19| – 26	5	0,1667	5
26| – 33	13	0,4333	18
33| – 40	4	0,1333	22
40| – 47	4	0,1333	26
47| – 54	3	0,1000	29
54| – 61	1	0,0333	30
Total	**30**	**1**	

a) 18

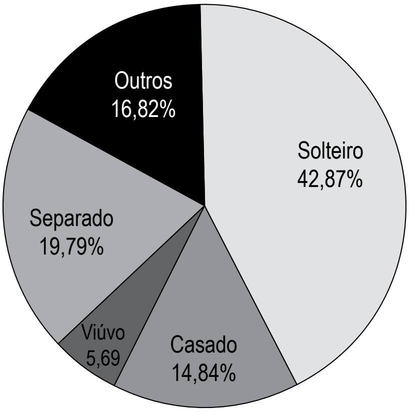
b) 13,33%

c) 17

d) 73,33%

e) 26,67%

3.



Através do gráfico, podemos dizer que aproximadamente 43% dos clientes desta agência de turismo são solteiros, 20% são separados, 17% têm outro tipo de estado civil, 15% são casados e apenas 5% são viúvos. Esta informação é importante na hora de lançar pacotes de viagens. A agência deve se lembrar que grande parte de seus clientes são solteiros. Também pode criar estratégias para trazer mais clientes casados ou viúvos, que provavelmente devem ter outro tipo de perfil.

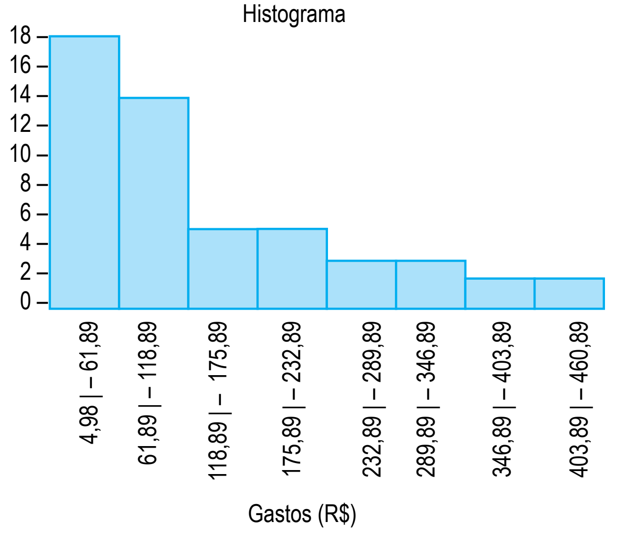
4.

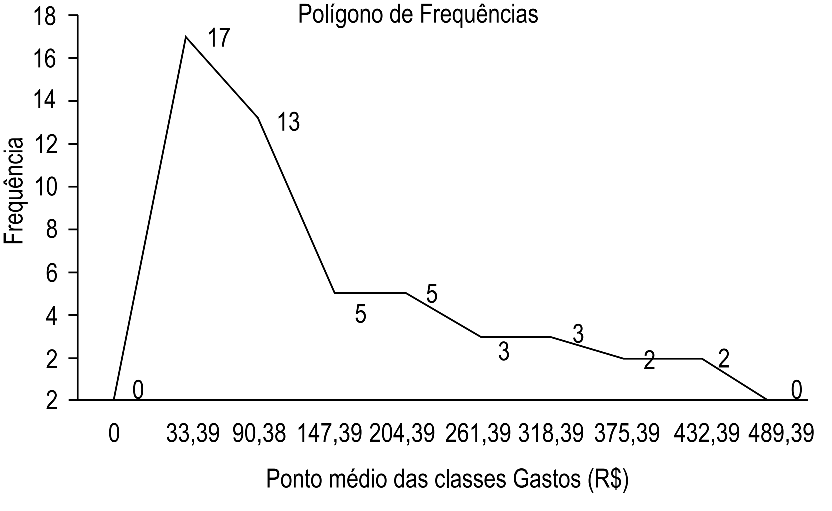
a) Valores gastos com supermercado. Variável quantitativa contínua.

b) Tabela 1: Distribuição de frequências para a variável Valores gastos com supermercado.

Classes (Gastos em R$)	f	f_r	f_a
4,89| – 61,89	17	0,34	17
61,89| – 118,89	13	0,26	30
118,89| – 175,89	5	0,10	35
175,89| – 232,69	5	0,10	40
232,89| – 289,89	3	0,06	43
289,89| – 346,89	3	0,06	46
346,89| – 403,89	2	0,04	48
403,89| 460,89	2	0,04	50
Total	50	1	

c)





5.

a) Salário de funcionários de uma empresa. Esta variável é classificada como quantitativa contínua.

b) 45 funcionários

c) 86 funcionários

d) 9,30%

e) 72,09%

f)

Salário (R$)	f	fr	fa
500,00| – 800,00	17	19,77	17
800,00| – 1.100,00	45	52,33	62
1.100,00| – 1.400,00	12	13,95	74
1.400,00| – 1.700,00	4	4,65	78
1.700,00| – 2.100,00	3	3,49	81
2.100,00| – 2.400,00	3	3,49	84
2.400,00| –2.700,00	2	2,33	86
Total	86	100,00	

Tabela 1 – Distribuição de frequências dos salários dos funcionários de uma empresa.

Capítulo 2

Antes das respostas, gostaríamos de deixar claro que as interpretações das questões ficam a cargo do estudante. Se ocorrerem dúvidas, entrar em contato com o tutor.

1.

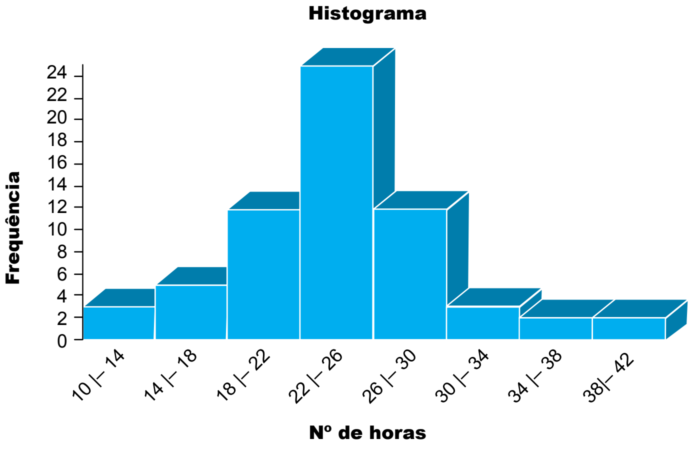
a) $\overline{X} \cong 23{,}1$, Md = 22,5 e Mo = 22

b)

Classes	f	*fr*	*fa*	Pm
10 ├ 14	3	0,0469	3	12
14 ├ 18	5	0,0781	8	16
18 ├ 22	12	0,1875	20	20
22 ├ 26	25	0,3906	45	24
26 ├ 30	12	0,1875	57	28
30 ├ 34	3	0,0469	60	32
34 ├ 38	2	0,0313	62	36
38 ├ 42	2	0,0313	64	40
Total	64	1,0000		

c) $\overline{X} \cong 24{,}6$ Md $\cong 23{,}9$ Md = 24 $Q_1 = 20{,}7$ $D_1 \cong 26$ $P_{99} = 40{,}7$

d)



2.

a) Vendas mensais. Variável quantitativa contínua.

b) $\overline{X} \cong 3{,}2$ Md = 3,4 Mo = 3,5

c) $Q_3 \cong 20{,}7$ $D_1 \cong 1{,}4$ $P_{80} = 4{,}1$

d) 16,36%

e) 21,82%

f) 56,36%

g) 65,45%

3.

a) $\overline{X} \cong 17{,}1$

b) Md = 2

c) Mo = 2

4.

a) Idade dos funcionários

b) 48

c) $\overline{X} \cong 30{,}9$ Md ≅ 29,7 Mo ≅ 27,8

d) $Q_1 \cong 25{,}7$

e) 25

f) 4,17%

g) 47,9%

5. 4,3

Capítulo 3

As interpretações das questões ficam a cargo do estudante. Se ocorrerem dúvidas, entrar em contato com o tutor.

1.

a) R = 31 $s \approx 6{,}1$ $s^2 \approx 37{,}2$ $cv \approx 0{,}2633$ ou 26,33%

b) R = 32 $s \approx 5{,}8$ $s^2 \approx 33{,}6$ $cv \approx 0{,}2358$ ou 23,58%

2. R = 6 $s \approx 1{,}24$ $s^2 \approx 1{,}54$ $cv \approx 0{,}3875$ ou 38,75%

3. R = 7 $s \approx 2{,}88$ $s^2 \approx 8{,}29$ $cv \approx 0{,}45$ ou 45%

4. R = 5 $s \approx 1{,}29$ $cv \approx 0{,}7588$ ou 75,88%

5. R = 30 $s \approx 6{,}9$ $s^2 \approx 47{,}6$ $cv \approx 0{,}2233$ ou 22,33%

6.

a) Caixa A (menor variação absoluta (s))

b) Caixa A (maior variação relativa (cv))

Capítulo 4

1.

a) 0,2328

b) 0,0461

c) 0,7672

2.

a) 0,7

b) 0,725

c) 0,6897

d) 0,7143

e) 0,2727

f) 0,75

3.

a) 0,6667

b) 0,4

c) 0,3

d) 0,2222

4.
a) 0,75 **b)** 0,25

5. 0,9703

6.
a) 0,02875 b) 0,5565

Capítulo 5

1. E(X) = 2,1 Var(X) = 0,63

2. 2,5

3. R$ 1.400,00

4.
a) 4,6 **b)** E(G) = 2,75 Var(G) = 0,4125

5.
a) 0,000011 **b)** 0,0042 **c)** 0,8329

6. 0,0033

7.
a) 0,1317 c) 0,1646
b) 0,7901 d) 0,9959

8.
a) 0,0778 b) 0,2592 c) 0,9222

9.
a) 0,0881

10.
a) 0,324135 **b)** 0,420175