

SÉRIE ACADÊMICA

ESTATÍSTICA APLICADA

LILIAN MALUF DE LIMA



SÉRIE ACADÊMICA

ESTATÍSTICA APLICADA

LILIAN MALUF DE LIMA

PIRACICABA • SÃO PAULO



©2021 PECEGE | Todos os direitos reservados. Permitida a reprodução desde que citada a fonte, mas para fins não comerciais. A responsabilidade pelos direitos autorais de texto e imagens desta obra são dos autores.

EXPEDIENTE EQUIPE

ORGANIZADORES

Carlos Shinoda

Daniela Flôres

Gabrielle de Souza Gomes

Haroldo José Torres da Silva

Maria Cecília Perantoni Fuchs Ferraz

Ricardo Harbs

PROJETO GRÁFICO E EDITORAÇÃO

Ana Paula Mendes Vidal de Negreiros

REVISÃO

Layane Rodrigues Vieira

Fernanda Latanze Mendes Rodrigues

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP) (CÂMARA BRASILEIRA DO LIVRO, SP, BRASIL)

M261e

Lima, Lilian Maluf.

Estatística Aplicada / Lilian Maluf de Lima - - Piracicaba, SP : Pecege Editora, 2021.

Série Acadêmica

ISBN: 978-65-86664-57-7

1. Probabilística. 2. Amostragem. 3. Levantamento. 4. Variáveis. I. Autor. II. Título.

III. Série.

CDD:311

FICHA CATALOGRÁFICA ELABORADA POR FELIPE MUSSARELLI CRB 9935/8

Os direitos autorais sobre as imagens utilizadas nesse material pertencem aos seus respectivos donos.

PREZADO(A) ALUNO(A),

Esse material foi desenvolvido no intuito de auxiliá-lo com os estudos nos cursos de **MBA** da **USP/ESALQ**, servindo como um referencial teórico básico e complementar às aulas oferecidas nos cursos.

Desejamos que esse material, de alguma forma, contribua para acrescentar novos conhecimentos, impulsionar o aprendizado e aprimorar as competências que já possui.

Bons estudos!!!

EQUIPE PECEGE



SOBRE A AUTORA

LILIAN MALUF DE LIMA

Graduada em Engenharia Agronômica pela Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ) da Universidade de São Paulo, com mestrado e doutorado em Ciências (Economia Aplicada) pela ESALQ/USP. Foi pesquisadora visitante em Wageningen University (WUR)- Agrotechnology and Food Sciences Organization e professora de métodos quantitativos em economia e administração na Pontifícia Universidade Católica de Campinas (PUC/Campinas) e na Universidade de Campinas, no Instituto de Economia (IE-UNICAMP). Atualmente é pós-doutoranda e professora do Departamento de Economia, Administração e Sociologia da ESALQ/USP. Sua área de atuação em termos de ensino e pesquisa é voltada ao uso de métodos quantitativos, com destaque para os trabalhos focados nos temas do agronegócio, comercialização, energia e comportamento do consumidor.

SUMÁRIO

1.	População e Amostra	9
1.1	Amostras probabilísticas	10
1.2	Amostras não-probabilísticas	12
1.3	Dados estatísticos primários e secundários	13
1.4	Variáveis quantitativas e qualitativas	14
	Recapitulando	16
	Referências	16
2.	Estatística Descritiva	16
2.1	Definição	16
2.2	Estatística Descritiva: métodos tabulares e gráficos	17
2.3	Estatística Descritiva: métodos numéricos	24
	Recapitulando	42
	Referências	43
3.	Probabilidades: experimentos, resultados experimentais e variáveis aleatórias	43
3.1	Definições	43
3.2	Variável aleatória (v.a.)	47
	Recapitulando	48
	Referências	49
4.	Distribuição de Probabilidade	49
4.1	Distribuições discretas de probabilidade	49
4.2	Distribuições contínuas de probabilidade	61
	Recapitulando	78
	Referências	79
	Apêndice	80

1. População e Amostra

A maioria dos trabalhos em estatística é realizada com o uso de **amostras** extraídas de uma **população**, na qual se deseja fazer um determinado estudo. O termo **população** refere-se a todos os elementos do grupo de interesse, ou seja, é o conjunto de todos os elementos que possuem uma característica em comum. O termo população não se aplica somente a moradores de determinada cidade, região ou país, mas também é utilizado para designar um conjunto de alunos de uma escola, trabalhadores de uma empresa, árvores de uma região, estabelecimentos ou animais em certa área, produtos em certo estabelecimento, entre outros.

Uma população tanto se refere a seres ou a objetos, como também aos atributos desses seres ou objetos. Dessa forma, tem-se, por exemplo, a população constituída pelo peso ou idade de indivíduos, pelo tamanho de animais, pela altura de árvores, pela potência de tratores e pelos tipos de veículos. Quando se levanta dados de toda a população, diz-se que está fazendo um **recenseamento**. O conjunto de dados obtidos em um recenseamento chama-se **censo**.

Uma amostra é um grupo de elementos extraídos da população. Dada uma população de **N** elementos, tem-se uma amostra dessa população se forem selecionados **n** elementos dela, com **n** sendo menor que **N** (Figura 1).

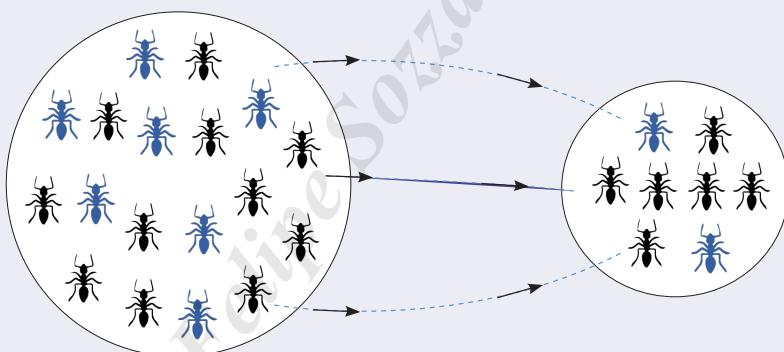


Figura 1. Diferença entre população e amostra

Alguns exemplos de população e amostra:

- **População:** todos os supermercados da cidade de São Paulo.
Amostra: 40 supermercados da capital paulista entrevistados para levantar preços de produtos de cesta básica.
- **População:** todos os agricultores filiados a uma cooperativa que possuem máquinas agrícolas de grande porte.
Amostra: trinta agricultores da cooperativa entrevistados sobre o uso dessas máquinas.

Sobre o tamanho da amostra, embora existam cálculos estatísticos que indiquem com bastante precisão o tamanho ideal, para fins práticos devemos considerar amostras que incluem no mínimo 30 elementos, conforme documentado em literatura correlata disponível em referências ao final deste capítulo.

Levantar os dados de todos os elementos que fazem parte da população é muitas vezes oneroso e de difícil execução. Dessa forma, quase sempre trabalhamos com uma amostra da população de interesse. É essencial que a amostra represente efetivamente a população da qual foi extraída. Se a amostragem for feita obedecendo certos critérios, podemos descobrir as características da população através dos dados levantados por meio de uma amostra. Chamamos esse processo de Inferência. Assim, a parte da estatística que procura deduzir informações relativas a uma população, mediante a utilização de amostras dela extraídas, é denominada **Inferência Estatística**.

Um critério importante a ser considerado numa amostragem diz respeito ao processo de seleção dos elementos da amostra. As amostras podem ser extraídas de diversas formas, sendo probabilísticas e não probabilísticas.

1.1 Amostras probabilísticas

Nas amostras probabilísticas sabemos a probabilidade de cada elemento selecionado. Dessa forma, as principais são:

Amostra Aleatória Simples (AAS): selecionar os elementos da população de forma que cada um deles tenha a mesma chance de ser selecionado.

Exemplo: considerando uma população finita de estabelecimentos (N) que vendem o mesmo produto em uma cidade, desejamos obter informações de preço e quantidade vendida deste produto nesta cidade. Devemos selecionar os estabelecimentos deste local para obtenção dessas informações. Ao obtermos a listagem desses locais de venda, escrevemos os nomes de cada estabelecimento da população em um cartão, misturando-os em uma urna e sorteando tantos cartões quanto desejarmos para obter uma amostra.

Quando a população é muito grande, isso torna-se inviável. Podemos ter uma AAS com reposição (uma unidade pode ser sorteada mais de uma vez) e sem reposição (unidade sorteada é removida da população).

Amostra aleatória estratificada (estratos): quando a amostra representa a estratificação da população.

Vamos considerar o levantamento de dados de venda de um produto agrícola para uma pesquisa. Para garantir que da amostra façam parte tipos

diferentes de estabelecimentos (supermercados, armazéns e quitandas, por exemplo), devemos fazer uma amostra estratificada. Para isso, devemos dividir os estabelecimentos em grupos de acordo com o tipo, e depois fazer uma amostragem por sorteio (AAS) para cada um desses grupos (estratos), vide Figura 2. É recomendável que a amostragem estratificada seja feita de forma proporcional. Por exemplo, se a população é composta de 100 estabelecimentos dos quais 20 são supermercados (20%), 30 são armazéns (30%) e 50 são quitandas (50%); em uma amostra de 30 estabelecimentos para serem entrevistados para coleta de informações do referido produto, devemos selecionar 6 supermercados (20%), 9 armazéns (30%) e 15 quitandas (50%). Assim, para a obtenção do grupo amostral de 6 supermercados, obtemos uma AAS de 6 dos 20 supermercados da população, por exemplo.

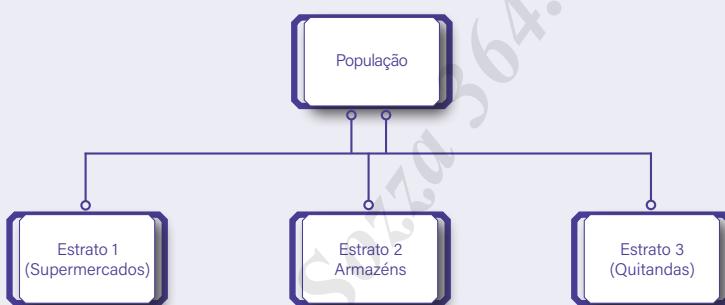


Figura 2. Diagrama de amostragem aleatória estratificada para o exemplo utilizado

Amostra por conglomerados: primeiramente a população é dividida em partes ou conglomerados e então se extrai uma amostra aleatória simples desses conglomerados. Cada conglomerado representa a população (teríamos subpopulações, pois representam a população inteira em pequena escala).

Utilizando o exemplo anterior de estabelecimentos que vendem determinado produto agrícola, em vez de termos um estrato para cada estabelecimento (supermercado, armazém ou quitanda), teríamos K conglomerados contendo os três tipos de estabelecimentos, que retratam a população (vide Figura 3).

Exemplos: amostragem por áreas em que os conglomerados são os quarteirões de uma cidade ou outras áreas bem definidas (podendo ser no campo). Esse tipo de amostragem requer um tamanho maior de amostra em comparação às amostragens anteriores.

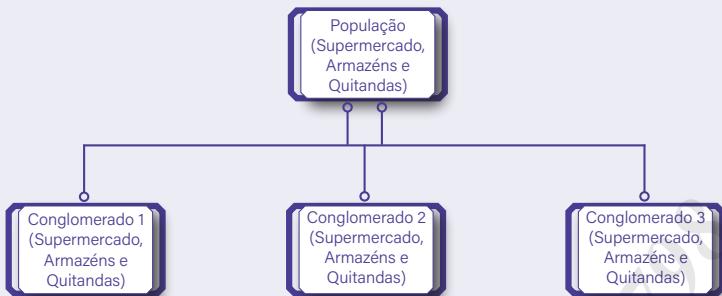


Figura 3. Diagrama de amostragem por conglomerados para o exemplo utilizado

Amostra sistemática: Quando selecionamos indivíduos sistematicamente.

Um exemplo dessa amostra acontece quando desejamos entrevistar 10 empregados de uma firma que contém 70 funcionários (população). Vamos supor que tenhamos a lista destes 70 funcionários numerados e, então, selecionaremos um funcionário a cada 7, totalizando os 10 empregados da amostra. Consideraremos para isso o total da população (70 empregados) dividido pelo número da amostra desejada (10 funcionários).

Esse tipo de amostragem apresenta vantagem sobre a amostra aleatória simples quando a população é muito grande. Como o primeiro elemento da amostra sistemática é uma escolha aleatória, podemos presumir que esse tipo de amostra tem as mesmas propriedades da MAS.

1.2 Amostras não-probabilísticas

Nas amostras não-probabilísticas as probabilidades de conhecimento e seleção da amostra são desconhecidas. Não há aleatoriedade para a escolha de um elemento da população. Quando aplicamos questionários de autopreenchimento, por exemplo, não temos controle estatístico dos dados coletados.

As principais são:

Amostra por conveniência: os elementos são selecionados para a amostra com base na conveniência/comodidade do pesquisador. Um professor que realiza pesquisas em uma Universidade pode utilizar estudantes voluntários para sua amostra, pois estão facilmente disponíveis por pouco ou nenhum

custo. Também, um gerente pode extrair uma amostra de um embarque de laranjas, selecionando-as casualmente de vários engradados. Rotular cada laranja e usar um método probabilístico de amostra seria impraticável.

Exemplos: estudos exploratórios são frequentemente utilizados em supermercados para testar produtos, para amostragem de animais selvagens capturados e amostragem de grupos de voluntários para pesquisas de consumidores, entre outros.

Amostra por julgamento: os elementos são selecionados para a amostra com base no julgamento da pessoa que realiza o estudo (pesquisador, por exemplo).

Apresenta maneira fácil de seleção da amostra, mas a qualidade dos resultados da amostra depende do julgamento da pessoa que seleciona.

Exemplo: um repórter pode tomar como amostra três senadores de um partido com o qual ele possui afinidade, julgando que eles refletem a opinião geral de todos os senadores.

É importante mencionar que devemos ter muita cautela ao tirar conclusões baseadas em amostras não probabilísticas para fazer inferências sobre populações.

1.3 Dados estatísticos primários e secundários

Quando utilizamos em nossas análises dados levantados e divulgados por instituições públicas ou privadas, dizemos que estamos trabalhando com dados secundários. Por outro lado, quando usamos dados levantados por nós para um estudo específico, estamos utilizando dados primários.

Dados estatísticos são divulgados por diversas instituições públicas e privadas.

Alguns exemplos de fontes secundárias de dados estatísticos:

➤ O Instituto Brasileiro de Geografia e Estatística (IBGE)¹ é uma instituição governamental que levanta e divulga dados de interesse do setor agrícola. Assim, tem-se várias publicações dessa instituição que podem ser de grande interesse para o produtor rural ou pesquisas da área, podendo-se citar: **Levantamento Sistemático da Produção Agrícola e Censo Agropecuário** - tratam de dados sobre previsão de safra, produção, valor, área, rendimento, pés colhidos, entre outras variáveis da agricultura. No caso da pecuária, trazem dados sobre rebanho, abate, etc. Também são divulgados por elas dados de horticultura e silvicultura; e Pesquisa de Orçamento Familiar (POF) - traz a quantidade adquirida de alimentos e bebidas e os dispêndios das famílias com itens de consumo alimentar, permitindo analisar os hábitos de consumo da população.

¹ <http://www.ibge.gov.br>

- O Ministério da Indústria, Comércio Exterior e Serviços fornece dados de exportação e Importação de produtos em geral, por meio do portal COMEX STAT².
- Índices de inflação e demais índices, podemos acessá-los no Instituto de Pesquisa Econômica Aplicada (IPEA) do Governo Federal³.
- Preços de produtos agrícolas com periodicidade diária, semanal ou mensal, em diferentes segmentos das cadeias agropecuárias, são coletados e divulgados por diversas instituições públicas e privadas, como por exemplo: Fundação Getúlio Vargas (FGV)⁴, Instituto de Economia Agrícola (IEA) do Governo do Estado de São Paulo⁵, Centro de Estudos Avançados em Economia Aplicada (CEPEA) da Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ) da Universidade de São Paulo (USP)⁶, Fundação Instituto de Pesquisas Econômicas (Fipe)⁷, entre outras.

- Preços de fretes de produtos agrícolas e pesquisas correlatas: Grupo de Pesquisa e Extensão em Logística Agroindustrial (ESALQ-LOG) da ESALQ/USP⁸.

Os dados levantados e divulgados por instituições públicas são, muitas vezes, também divulgados em publicações ou sites de instituições privadas.

1.4 Variáveis quantitativas e qualitativas

Quando os indivíduos ou elementos podem ser classificados em um número de categorias mutuamente exclusivas (se pertencer a uma, não pertence a outra), neste caso temos uma **variável qualitativa**. Assim, se fizermos o levantamento de um grupo de pessoas que consome determinado produto, podemos distribuir a resposta em duas classes: não consome e consome, por exemplo. As variáveis qualitativas podem ser nominais ou ordinais.

Variáveis quantitativas são expressas por números, exemplo: número de cabeças de gado em propriedades agrícolas, preço, renda, área de estabelecimentos agrícolas, número de pessoas ou estabelecimentos, etc. As variáveis quantitativas podem ser discretas ou contínuas. Um conjunto de dados numéricos pode referir-se a uma variável discreta ou a uma variável contínua. Vide na Figura 4, a classificação das variáveis.

² <http://comexstat.mdic.gov.br/pt/home>

³ <http://www.ipeadata.gov.br/Default.aspx>

⁴ <https://portalibre.fgv.br>

⁵ <http://www.iea.agricultura.sp.gov.br/out/index.php>

⁶ <https://www.cepea.esalq.usp.br/>

⁷ <https://www.fipe.org.br>

⁸ <https://esalqlog.esalq.usp.br>

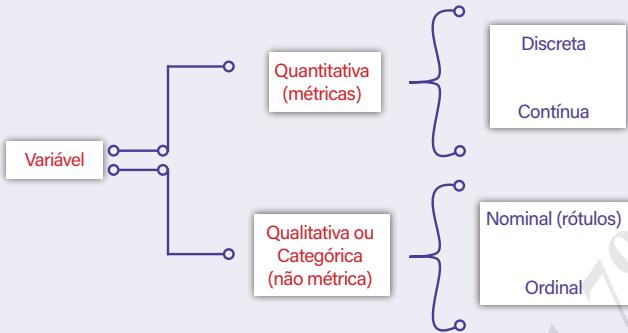


Figura 4. Tipos de variáveis (classificação)

1.4.1 Variáveis Quantitativas: discretas e contínuas

Variáveis discretas são aquelas que só podem assumir um número finito de diferentes valores dentro de um intervalo finito. Dados de contagem são sempre variáveis discretas. Assim tem-se, por exemplo, que o número de cabeças de gado em propriedades agrícolas é uma variável discreta, bem como número de indivíduos.

Variáveis contínuas podem assumir um número infinito de diferentes valores dentro de um intervalo finito. As variáveis econômicas medidas em unidades monetárias como, por exemplo, preço e renda, são variáveis contínuas. São também variáveis contínuas: a área dos estabelecimentos agrícolas, peso de objetos, etc.

1.4.2 Variáveis Qualitativas: nominais e ordinais

Variáveis nominais (rótulos): os dados são distribuídos em um número de categorias mutuamente exclusivas. Exemplos: tipo de religião, naturalidade de indivíduos, cores de veículos (sólida ou metálica), tipos de gêneros, rótulos usados para identificar o nome do produtor de maçãs de certo varejão, entre outras.

Variáveis ordinais: os dados são classificados em um número de categorias mutuamente exclusivas que podem ser ordenadas como “maior que”, “mais difícil que”, “superior à”, entre outros. Alguns exemplos: indivíduos conforme a carreira militar, status socioeconômico, grau de dureza dos minerais, classificação do calibre de frutas, por exemplo, de pêssegos vendidos na Companhia de Entrepostos e Armazéns Gerais de São Paulo (Ceagesp), com diâmetros: pequeno (2,5 a 4,5 cm), médio (4,5 a 5,6 cm) e grande (acima de 5,6 cm), etc. É comum associar números inteiros às variáveis qualitativas para poder analisá-las sob diferentes métodos estatísticos.

RECAPITULANDO

Se a amostra for representativa e a amostragem feita de forma correta, os resultados obtidos através da amostra podem ser considerados para toda a população. A população inclui todos os elementos que possuem uma característica em comum. Uma amostra é um grupo de elementos extraídos da população. Como é dispendioso, difícil e muitas vezes impraticável ter acesso à toda a população, costuma-se escolher uma amostra e estudá-la. Inferência: selecionamos uma amostra representativa de uma população e usamos os dados para fazer inferências sobre características da população da qual selecionamos a amostra. Tipos de amostragem: probabilística (amostra aleatória simples, amostra aleatória estratificada, amostra por conglomerados) e não-probabilística (amostra por conveniência e por julgamento). Recomenda-se o uso de amostras probabilísticas. Dados: são os fatos e números coletados, analisados e sintetizados para apresentação e interpretação. Todos os dados de um estudo denominam-se conjunto de dados. Dados usados como amostras podem ser secundários (coletados por instituições públicas e privadas) e primários (coletados pelo pesquisador). Variáveis: são características dos dados. Um conjunto de dados pode incluir uma ou mais variáveis. As variáveis classificam-se como quantitativas (discretas e contínuas) e qualitativas (nominais e ordinais). Por exemplo, preço e quantidade de soja classificam-se como variáveis quantitativas contínuas.

Referências

- Anderson, D.R.; Sweeney, D.J.; Williams, T. 2007. Estatística Aplicada à Administração e Economia. Pioneira Thompson Learning: São Paulo, SP, Brasil. p.10-13; 262-265.
- Bussab, W.O.; Morettin, P.A. 2006. Estatística Básica. Saraiva: São Paulo, SP, Brasil. p.9-16.; 255-283.
- Hoffmann, R. 2006. Estatística para Economistas. Pioneira Thomson Learning: São Paulo, SP, Brasil. p. 1-5.

2. Estatística Descritiva

2.1 Definição

No conceito comum, a estatística trata da coleta e apresentação de dados em tabelas ou gráficos ou, mais especificamente, da organização e apresentação de **contagens e medições**. Estatística **não** deve ser vista como uma coleção de números. As informações são obtidas com a finalidade de tomar decisões, exemplos de tomada de decisões: **obtenção de índice de audiência** para a avaliação de se manter ou não um programa no ar; **dados de quantidade vendida de um produto no tempo** para a realização de planejamento de estoques; **análise de preço de um produto** para avaliar a conveniência de produzir ou não, em maior ou menor escala;

análise do câmbio para avaliar impactos na balança comercial, entre outros.

Assim **estatística**, em seu conceito mais abrangente, refere-se ao conjunto de técnicas que auxiliam na tomada de decisão quando prevalecem condições de incerteza. Uma vez que os dados são coletados, a organização e a interpretação destes podem ocorrer pelo auxílio da estatística. Podemos dizer que formas gráficas, tabulares ou numéricas utilizadas para sumarizar e interpretar os dados coletados são conhecidas como **estatística descritiva**.

As fases do método de estatística descritiva podem ser visualizadas pela Figura 5:

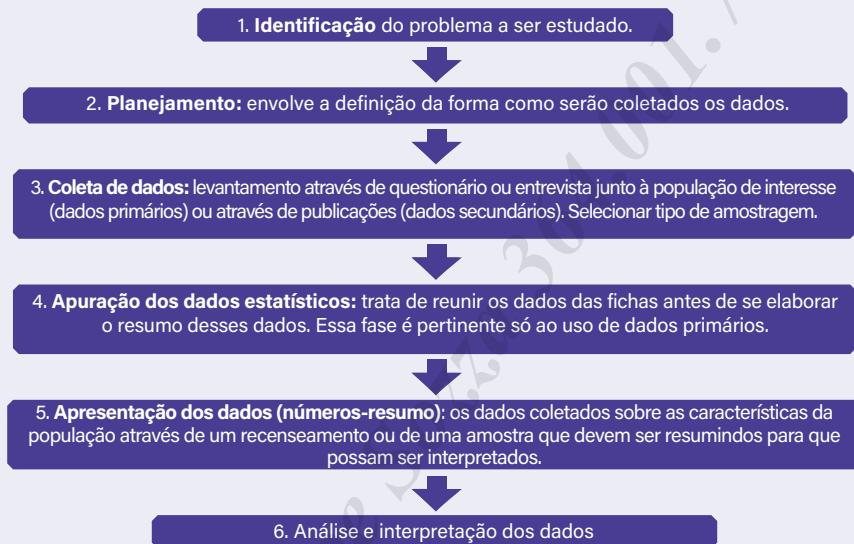


Figura 5. Fases do M todo de Estat stica Descritiva

As tr s  ltimas fases da estat stica descritiva (Figura 5) merecem maior detalhamento, sendo descritas a seguir em dois t picos: m todos tabulares e gr ficos e m todos num ricos.

2.2 Estat stica Descritiva: m todos tabulares e gr ficos

A etapa 4 da Figura 5 consiste na apura o dos dados estat sticos. Esse processo depende do tipo de vari vel: qualitativa ou quantitativa. Para a vari vel qualitativa, a apura o se d  por simples contagem. Por exemplo, “5 pessoas consomem um produto e 3 pessoas n o consomem o mesmo produto”. Para a vari vel quantitativa ´ necess rio que anotemos cada elemento, por exemplo: dados de pre o de um produto agr cola em estabelecimentos varejistas (vide Tabela 1).

Tabela 1. Tabulação de dados de preços de um produto agrícola por estabelecimentos de venda

Nome do Estabelecimento	Preço do Produto
	R\$
Supermercado A	1,45
Quitanda A	1,88
Armazém A	1,54
Supermercado D	1,55
Armazém F	1,33
Quitanda C	1,98

Nota: Dados hipotéticos

Após reunirmos os dados das fichas de coleta de dados (no caso de dados primários) ou compilar os dados secundários, devemos apresentá-los de forma resumida. Podemos fazer isso através de uma tabela de **Distribuição de Frequência**.

2.2.1 Distribuição de Frequência

Toda vez que dispomos de uma série de dados, podemos organizá-los em tabelas para facilitar a análise. Para exemplificar, vamos considerar que temos 320 dados referentes à mortalidade de aves em um determinado lote e em um determinado aviário (variável quantitativa discreta). Esses dados são apresentados a seguir (Tabela 2):

Tabela 2. Distribuição da quantidade de mortes de aves por causa (motivo)

Causa de mortalidade em aves	Mortalidade em aves (frequência absoluta)
Infecciosa	86
Metabólica*	146
Distúrbios locomotores	22
Inanção	45
Falha no manejo	10
Defeitos congênitos (má formação)	11
Total	320

Nota: *Frangos de corte criados em sistema intensivo tipo “dark house” que apresentaram síndrome de morte súbita e síndrome ascítica (causa déficit de oxigenação na ave).

Fonte: Adaptado de Bonfanti (2016).

A quantidade de vezes que um dado (mortalidade de aves em um lote conforme a causa) aparece é chamado de frequência, e a Tabela 2 é uma tabela de frequência. Note que a tabela de frequência nos dá uma ideia muito mais clara da distribuição da mortalidade de aves em um aviário conforme as suas causas,

em vez da apresentação dos dados individuais. Verificamos, por exemplo, que a maioria (86 + 146) das aves morreram devido às causas infecciosa e metabólica, enquanto a minoria foi devido à falha no manejo (10).

Veja que o total das frequências é 320, que corresponde às 320 aves para as quais temos dados sobre mortalidade ocasionadas por “causas”. É interessante, muitas vezes, calcular as frequências acumuladas e, nesse caso, a frequência acumulada de cada causa de morte em aves é igual a soma da frequência dessa causa com as das causas anteriores. Observa-se na Tabela 3, segunda coluna, a frequência acumulada para o caso da mortalidade de aves, verificando-se, por exemplo, que em 254 das 320 aves mortas as causas foram infecciosas, metabólicas e distúrbios locomotores.

Tabela 3. Distribuição da quantidade de mortes de aves por causa, incluindo frequência acumulada e relativa

Causa de mortalidade em aves	Frequência absoluta	Frequência acumulada	Frequência relativa	Frequência relativa acumulada
Infecciosa	86	86	0,269	0,269
Metabólica*	146	232	0,456	0,725
Distúrbios locomotores	22	254	0,069	0,794
Inanição	45	299	0,141	0,934
Falha no manejo	10	309	0,031	0,966
Defeitos congênitos (má formação)	11	320	0,034	1,000
Total	320	-	1,000	-

Nota: *Frangos de corte criados em sistema intensivo tipo “dark house” com síndrome de morte súbita e síndrome ascítica (causa déficit de oxigenação na ave).

Fonte: Adaptado de Bonfanti (2016).

Muitas vezes é útil calcular a frequência relativa e, para isso, basta dividir cada uma das frequências absolutas pela quantidade total de dados, 320 no nosso exemplo (correspondentes às 320 aves mortas). Chamemos de “classe” cada causa de morte das aves. A frequência relativa de uma classe multiplicada por 100 nos dá a porcentagem que o valor dessa classe aparece nos dados considerados. Assim tem-se, por exemplo, que em 45,6% das aves que morreram, a causa foi metabólica e, somente em 3,1% a causa foi falha no manejo (Tabela 3). A frequência relativa acumulada de cada classe é calculada somando a frequência relativa dessa classe com as das classes anteriores, apresentada na última coluna da Tabela 3. Notamos, por exemplo, que em 72,5% das aves que morreram, as causas foram infecciosas e metabólicas; ou ainda que em 93,4% das aves que morreram, as causas foram infecciosas, metabólicas, distúrbios locomotores e falhas no manejo.

Podemos visualizar ainda melhor essa distribuição traçando um gráfico da tabela de frequência. Uma distribuição de frequência para uma variável discreta pode ser representada por um gráfico de barras. Para construir esse gráfico, marcamos sobre o eixo horizontal os diferentes

valores assumidos pela variável (causas da mortalidade de aves em dado lote, no nosso exemplo). Em seguida, traçamos barras verticais, cujas bases se situam sobre o eixo horizontal, nos pontos marcados e cujas alturas são proporcionais às frequências. Este gráfico pode ser elaborado pelo Excel ou por meio de diversos softwares estatísticos, como o *R*, Eviews, Stata, por exemplo. Na Figura 6 encontra-se representada graficamente a distribuição de frequência correspondente à Tabela 2.

Pode-se observar, na Tabela 2 ou na Figura 6, que a quantidade de mortes de aves por “causas” que aparece com maior frequência é metabólica (146), que corresponde à barra de maior altura, seguida por infecciosa (86) e por inanição (45). A frequência relativa poderia ser utilizada alternativamente no gráfico apresentado na Figura 6:

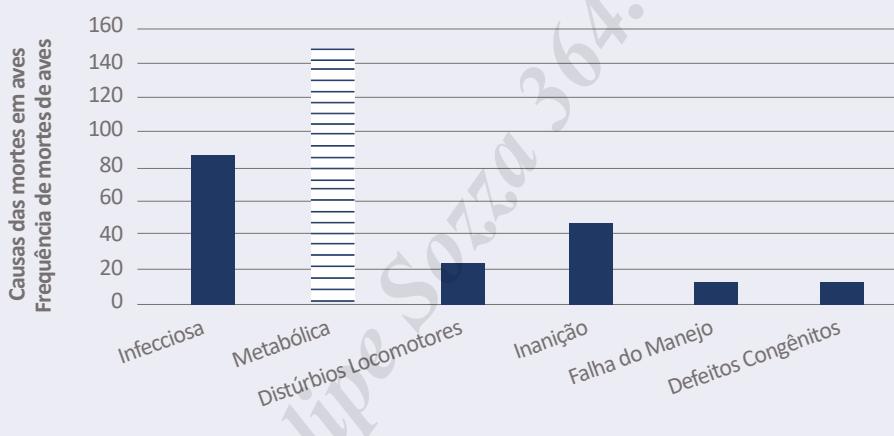


Figura 6. Gráfico de barras relativo à distribuição da quantidade de mortes de aves conforme a causa (Tabela 2)

Poderíamos considerar outro exemplo de variável quantitativa discreta: em vez de ser discriminada por um rótulo (causas de mortes em aves) fosse discriminada por número. Consideraremos uma amostra de 200 famílias de certa cidade e a quantidade de filhos por essas famílias. Sem a tabulação dos dados, teríamos algo como (dados originais) 200 informações de números de filhos de cada família, apresentadas assim: 3, 5, 0, 1, 1, 3, 2, etc. Esses números representam a quantidade de filhos de cada uma das 200 famílias que, tabulados, gerariam a Tabela 4, conforme os cálculos de frequências já explicitados:

Tabela 4. Distribuição das famílias conforme o número de filhos

Número de filhos por família	Frequência absoluta (número de famílias)	Frequência relativa	Frequência relativa acumulada
0	32	0,16	0,16
1	46	0,23	0,39
2	50	0,25	0,64
3	40	0,20	0,84
4	16	0,08	0,92
5	8	0,04	0,96
6	6	0,03	0,99
7	0	0,00	0,99
8	2	0,01	1,00
Total	200	1,00	-

Nota: Dados hipotéticos

Fonte: Adaptado de Hoffmann (2006).

Podemos observar pelas informações da Tabela 4:

- 32 das 200 famílias não possuem filhos (ou possuem zero número de filhos);
- 50 das 200 famílias possuem 2 filhos ou, em termos relativos, 25% das famílias apresentam 2 filhos, **exatamente**;
- nenhuma família possui 7 filhos;
- 92% das famílias possuem **até** 4 filhos;
- 64% das famílias possuem **até** 2 filhos.

Nos exemplos apresentados, consideramos variáveis quantitativas discretas, como quantidade de mortes de aves por “causas” e número de filhos por família, que só assumem valores inteiros (contagem). Vamos agora ver como se agrupam dados de uma variável **quantitativa contínua**, ou seja, de uma variável que pode assumir outros valores que não só os números inteiros. Para exemplificar, vamos supor que temos 40 dados de preço de venda de arroz agulhinha, em **R\$/saca de 60kg**, que foram informados por 40 produtores, numa pesquisa de cotação semanal de preços agrícolas. Esses dados hipotéticos são apresentados a seguir:

27,5; 25,3; 28,2; 27,4; 27,3; 24,3; 29,1; 26,3; 26,2; 26,9; 28,0; **31,5**; 27,2; 27,3; 30,9; 25,0; 25,9; 26,5; 30,0; 24,2; 27,4; 28,5; 26,2; 28,6; **24,0**; 25,3; 27,1; 26,7; 29,8; 28,9; 25,6; 29,7; 28,8; 27,7; 28,1; 27,5; 27,4; 27,3; 26,2; 26,1.

Para construir uma distribuição de frequência para uma variável contínua, devemos estabelecer intervalos de classe. Para isso, definimos primeiramente a quantidade de classes a ser considerada. Recomenda-se que essa quantidade esteja entre 5 e 20 e que as classes tenham a mesma amplitude, isto é, que a diferença entre o valor máximo e mínimo de cada classe seja igual. Veja que,

se definirmos uma quantidade muito pequena de classes, poderemos ter uma tabela muito resumida e perder algumas informações importantes. Por outro lado, se definirmos uma quantidade muito grande de classes, podemos ter um detalhamento desnecessário.

Vamos construir uma distribuição de frequência utilizando o nosso exemplo de preço do arroz. Verifica-se que o menor valor é 24 e o maior é 31,5 (em destaque na série supracitada). Subtraindo 24 de 31,5 ficamos com **7,5** (que é o valor da amplitude da amostra de preço). Vamos considerar que queiramos construir uma distribuição de frequência com 8 classes. A **amplitude de cada classe** será, então, a **amplitude da amostra** (7,5) dividida por 8 classes pré-definidas, que resulta em aproximadamente 1 ou **R\$1,00**. Na Tabela 5 apresentamos essa distribuição de frequência com intervalos de classe correspondendo a R\$1,00, considerando 8 classes. Nessa tabela incluímos uma coluna onde apresentamos o valor central de cada classe, além das colunas de frequência absoluta e de frequência relativa.

Tabela 5. Distribuição de preços hipotéticos de arroz agulhinha, em R\$/sc 60 kg, incluindo frequência relativa, para 8 classes

Intervalos de classe dos preços do arroz	Valor central de cada classe	Frequência absoluta	Frequência relativa
-- R\$/sc --	-- R\$/sc --		
24 - 25	24,5	3	0,075
25 - 26	25,5	5	0,125
26 - 27	26,5	8	0,200
27 - 28	27,5	11	0,275
28 - 29	28,5	7	0,175
29 - 30	29,5	3	0,075
30 - 31	30,5	2	0,050
31 - 32	31,5	1	0,025

Nota: Dados hipotéticos

Veja que temos que construir a tabela de forma que não haja indefinição em relação à classe onde vamos incluir um determinado valor limite, como o 25, por exemplo. Vamos usar a notação “a | - b” para o intervalo de números que inclui o extremo **a**, mas não inclua o extremo **b**. Dessa forma, o preço de R\$25,00, por exemplo, deve ser contabilizado na segunda classe e não na primeira.

Notamos pelos dados da Tabela 5 que preços de arroz dentro do intervalo de R\$27,00/sc (inclusive) e R\$28,00/sc são os mais frequentes, ou seja, a quarta classe é a mais frequente.

A distribuição de uma variável contínua pode ser representada por gráfico, esse gráfico pode ser um histograma. No histograma, cada retângulo

corresponde a uma dada classe, sendo a área de cada retângulo proporcional à frequência. A Figura 7 refere-se à distribuição de frequência apresentada na Tabela 5. Se a base do retângulo for uma unidade (que corresponde ao intervalo de classe de uma unidade), a frequência é proporcional à altura. Veja que no histograma marcamos os valores centrais no eixo horizontal.



Figura 7. Histograma relativo à distribuição de preços hipotéticos de arroz agulhinha, em R\$/sc 60 kg, incluindo frequência absoluta, para 8 classes com valor central de cada classe (Tabela 5)

Nota: Dados hipotéticos

Vamos considerar agora que desejamos construir, para os dados de preço do arroz, uma tabela de frequência com 5 classes (e não mais 8 classes como anteriormente). Para essa quantidade de classes temos $7,5/5 = 1,5$, que é a **amplitude dos intervalos** de classe a serem consideradas. A Tabela 6 mostra a distribuição dos preços do arroz agulhinha para 5 classes.

Tabela 6. Distribuição de preços hipotéticos de arroz agulhinha, em R\$/saca de 60 kg, incluindo frequência relativa, para 5 classes

Intervalos de classe dos preços do arroz	Valor central de cada classe	Frequência absoluta	Frequência relativa
--- R\$/sc ---	--- R\$/sc ---		
24,0 - 25,5	24,75	6	0,150
25,5 - 27,0	26,25	10	0,250
27,0 - 28,5	27,75	14	0,350
28,5 - 30,0	29,25	7	0,175
30,0 - 31,5	30,75	3	0,075

Nota: Dados hipotéticos

Muitas vezes, para simplificar a construção da distribuição de frequência, aproximamos uma variável contínua de uma variável discreta desconsiderando as casas decimais. Veja que isso é só uma aproximação da distribuição de frequência

de uma variável contínua.

Cabe destacar que existem fórmulas para calcular o número de classes, para que esse valor não seja arbitrário. As mais utilizadas, segundo literatura pesquisada, são a regra da Raiz Quadrada e a Regra de Scott. A primeira é calculada pela raiz quadrada do número referente ao tamanho da amostra; já a segunda é calculada pela raiz cúbica do tamanho da amostra multiplicado por 2. Considere n como o número de elementos da amostra. A Regra da Raiz Quadrada é dada por \sqrt{n} ; já a regra de Scott é dada por $\sqrt[3]{2n}$.

2.3 Estatística Descritiva: métodos numéricos

Já observamos que podemos resumir dados utilizando uma tabela de frequência. No entanto, muitas vezes, queremos resumir ainda mais esses dados apresentando um único valor que seja representativo de toda a série. Dessa forma, podemos fazer isso utilizando **métodos numéricos**.

As medidas numéricas de estatísticas descritivas podem ser usadas para resumir as informações de um conjunto de dados, podendo-se extrair mais análises. Essas medidas podem ser classificadas em:

- medidas de posição ou de tendência central;
- medidas de dispersão ou de variabilidade;
- mediadas de associação;
- medidas de formato ou da forma da distribuição.

Essas medidas podem ser usadas tanto para população como para amostra. Caso essas medidas sejam extraídas de uma população, são chamadas de parâmetros; se extraídas de uma amostra, são chamadas de estatística.

Assim, dentro do conceito de Inferência, podemos fazer o uso de uma estatística amostral (média amostral, variância amostral, desvio padrão amostral, por exemplo) para estimativas de parâmetros populacionais (média populacional, variância populacional, desvio padrão populacional, etc).

Define-se, portanto:

População: é o conjunto de todos os elementos ou resultados sob investigação.

Parâmetro: medida usada para descrever uma característica da população.

Amostra: é qualquer subconjunto da população.

Estatística: medida usada para descrever uma característica da amostra.

2.3.1 Medidas de posição ou de tendência central

Uma medida de tendência central de um conjunto de dados mostra o valor em torno do qual se agrupam as observações. As principais medidas de tendência central são a média aritmética (ou simplesmente média), a mediana

e a moda. Também são muito utilizadas as médias ponderada e geométrica. É mais fácil realizar análises comparativas a partir de um valor único que represente a série amostral.

Média aritmética

A média aritmética consiste na soma dos elementos da amostra dividido pelo número de elementos deste conjunto. A média de uma amostra costuma ser indicada por \bar{X} ; a média da população é indicada pela letra grega μ .

Exemplo: suponhamos o prazo médio de cobrança dos ativos de uma empresa, em dias (Tabela 7):

Tabela 7. Dados para o cálculo de média aritmética

	Ativo 1	Ativo 2	Ativo 3
Prazo (dias)	47	76	91

Nota: Dados hipotéticos

Sendo o tamanho da amostra “n” igual a três, temos que a média é dada por:

$$\bar{X} = \frac{47 + 76 + 91}{3} \cong 71,33$$

Como não resulta em um valor inteiro, poderemos arredondá-lo, considerando que se trata de uma cobrança. Logo, o prazo médio de cobrança dos ativos de uma empresa na condição exposta, corresponde a **71** dias.

Média ponderada

A média ponderada corresponde à média aritmética, considerando a atribuição de pesos em seu cálculo. Se um valor tem peso maior na amostra significa que ele entrará mais vezes na média. Assim, para o mesmo exemplo anterior, teremos (Tabela 8):

Tabela 8. Dados para o cálculo de média ponderada

	Ativo 1	Ativo 2	Ativo 3
Prazo (dias)	47	76	91
PESO	22.600	68.000	134.000

Nota: Dados hipotéticos

$$\bar{X} = \frac{(47 \times 22.600) + (76 \times 68.000) + (91 \times 134.000)}{(22.600 + 68.000 + 134.000)} = 82$$

Logo, o prazo médio de cobrança dos ativos de uma empresa, ponderado pelo preço dos ativos, na condição exposta, corresponde a 82 dias. Podemos notar que a média ponderada é maior que a aritmética neste exemplo, pois temos o maior peso, dado pelo preço, para o Ativo 3, o qual também dispõe do maior prazo, influenciando o valor da média ponderada para um nível mais elevado que a aritmética. Um outro exemplo de uso deste tipo de média é o cálculo da média da taxa de emissão de dióxido de carbono (CO_2) em diferentes regiões, ponderada pela quantidade de veículos automotores.

Média Geométrica

A média geométrica corresponde à raiz n -ésima da multiplicação dos valores da amostra que se pretende analisar. Por exemplo, a média geométrica para um aluno que tirou notas 4, 6 e 8, será (com $n = 3$):

$$\bar{X}_G = \sqrt[3]{4 \times 6 \times 8} \cong 5,8 \text{ aprox.}$$

A média geométrica é muito utilizada, por exemplo, para o cálculo da média de rendimento de ações.

Considerações para o cálculo de médias: a média para uma tabela de frequência de uma variável discreta é a soma dos números obtidos multiplicando os valores da variável pelas respectivas frequências relativas (média ponderada pela frequência). Para uma tabela de frequência de variável contínua, a média é a soma dos números obtidos multiplicando os valores centrais das classes pelas respectivas frequências relativas.

Mediana

A mediana corresponde ao valor que divide um conjunto de dados ao meio. Sendo os valores da amostra ordenados em ordem crescente (ou decrescente), a mediana é o valor tal que metade dos dados são iguais ou inferiores a esse valor e a outra metade são iguais ou superiores ao mesmo. Ao contrário da média, a mediana não é sensível aos valores extremos da amostra.

Exemplo 1: quando o tamanho da amostra “ n ” corresponde a um número ímpar.

Considere valores de vendas (em milhares de unidades) de um produto em 5 indústrias ($n=5$), expostos em “Amostra” na Figura 8.

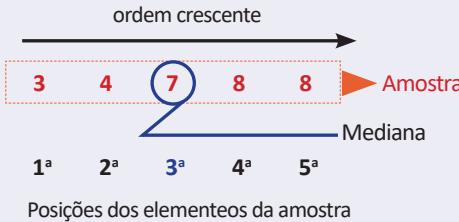


Figura 8. Valor da mediana quando o tamanho da amostra for um número ímpar

Neste caso, calculamos qual a posição em que se encontra o valor da mediana. Para isso, calcula-se $\frac{n+1}{2} = \frac{5+1}{2} = 3$. Logo, o valor da mediana ocupa a 3ª posição da amostra, estando seus elementos em ordem crescente. Este valor corresponde a 7 mil unidades de produtos vendidos, conforme exemplo utilizado.

Exemplo 2: quando o tamanho da amostra “ n ” corresponde a um número par.

Considere valores de preço de um produto (R\$) em 8 supermercados de São Paulo ($n=8$), expostos em “Amostra” na Figura 9:

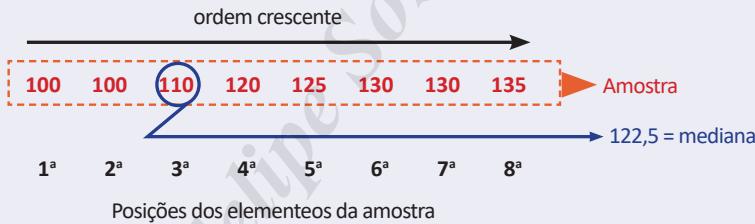


Figura 9. Valor da mediana quando o tamanho da amostra for um número par

Neste caso, também calculamos qual a posição que se encontra o valor da mediana. Para tanto, calculam-se as posições $\frac{n}{2}$ e $\left(\frac{n}{2}\right)+1$ dadas por $\frac{8}{2} = 4$ e $\left(\frac{8}{2}\right)+1 = 5$, respectivamente.

Então, localizamos os valores das 4ª e 5ª posições que são R\$120 e R\$125, respectivamente e calculamos a média dos mesmos. A média desses valores será $\frac{120 + 125}{2} = \text{R\$}122,5$ que será o valor da mediana.

É importante mencionar que o uso da mediana é preferível à média como medida de posição central quando a amostra apresentar muitos valores discrepantes (“outliers”). O cálculo da média se mostra sensível a esses valores.

Moda

A moda é a realização mais frequente do conjunto de valores observados (aquele que mais se repete na amostra de dados).

Por exemplo: em uma reunião do Comitê de Política Monetária do Banco Central [COPOM], uma entrevista apurou a opinião de 9 economistas a respeito de uma redução da taxa Selic (taxa básica de juro da economia). Os valores hipotéticos por eles anunciados foram:



Figura 10. Valor da moda para amostra de valores referentes à redução da Taxa Selic, conforme 9 economistas

Nota: Dados hipotéticos

Assim, de acordo com o exemplo supracitado (Figura 10), o valor da moda é de 0,25 (o que mais se repete). Um conjunto de dados pode ter mais de uma moda ou não ter moda, sendo essa uma limitação do uso dessa medida de tendência central. Nos casos multimodais, a moda quase nunca é considerada, porque relacionar três ou mais modas não seria especialmente útil para descrever a posição dos dados da amostra. Destaque-se que a moda é uma medida importante de posição de dados qualitativos.

Considerando dados quantitativos, numa tabela de frequência de uma variável discreta, a moda é o valor que aparece com maior frequência. Quando se tem uma tabela de frequência de uma variável contínua, a moda pode ser tomada como o valor central da classe de maior frequência. Assim, para os dados da Tabela 6, por exemplo, a moda é R\$27,75/sc.

Para entendermos a diferença entre a média e a moda, podemos considerar o exemplo sobre a mortalidade de aves por lote, em quatro lotes, por exemplo: Lote 1 com 25 mortes, Lote 2 com 17 mortes, Lote 3 com 22 mortes e Lote 4 com 17 mortes. Se alguém nos perguntar quantas aves morrem em média por lote e resolvemos calcular a média daqueles dados para dar a resposta, vamos dizer que morrem 20,25. Esse valor nunca será observado na série de dados, porque não se tem um número de mortes que não seja inteiro. Se, por outro lado, resolvemos responder a questão utilizando como medida de tendência central a moda, diremos 17 e não qualquer outro número.

Se utilizarmos a mediana para responder o valor médio de morte de aves por lote, na amostra do mesmo exemplo, encontráramos o valor de 19,5 ($(17 + 22)/2$). Por ser a amostra de tamanho par (4 lotes), após organizarmos os valores em ordem crescente, deveríamos realizar a média dos dois valores

centrais (17 e 22). No caso de um número par de dados, pode ocorrer (como foi em nosso exemplo) de a mediana, assim como a média, ser um número não observado na série de dados (não inteiro) já que ela é uma grandeza obtida pela média dos dois valores centrais. Assim, a não ser que os dois valores centrais sejam iguais, ou que a média desses dois valores centrais fosse um número inteiro para representar o número de morte de aves, a mediana aqui é um valor não observado na série e não indicado para o uso neste caso.

2.3.2 Medidas de dispersão ou de variabilidade

O resumo de um conjunto de dados por uma única medida representativa de posição central esconde toda a informação sobre a variabilidade do conjunto de informações. As medidas de dispersão são utilizadas para avaliar o grau de variabilidade de um conjunto de dados. Se um produtor vendesse todas as sacas de milho colhidas por R\$20,00 cada, não haveria dispersão alguma nos dados. Se, no entanto, o produtor vendesse metade das sacas colhidas por R\$22,00 e metade por R\$18,00, embora a média fosse também de R\$20,00, haveria dispersão.

Iremos discutir quatro medidas de dispersão. Na verdade, as duas últimas são construídas a partir da segunda:

- amplitude;
- variância;
- desvio padrão;
- coeficiente de variação.

Amplitude

Uma forma simples de medir a dispersão de uma série de dados é realizar a diferença entre o maior e o menor valor, essa grandeza é chamada de amplitude. Considerando os preços da saca de milho de R\$22,00 e R\$18,00, a amplitude é de R\$4,00. A amplitude nos dá ideia do afastamento entre o maior e o menor valor, mas não é, na verdade, uma boa medida de dispersão porque ela não leva em consideração os demais dados da série. A amplitude não nos dá qualquer informação sobre a distância entre quaisquer elementos da série de dados, exceto entre seus valores máximo e mínimo (extremos). Na Tabela 9, apresentamos, como exemplo, duas séries de dados com a mesma amplitude (500), podendo-se observar, no entanto, que a dispersão geral da série B é bastante superior à da série A.

Tabela 9. Exemplo de séries com a mesma amplitude

Série A		Série B
500		500
250		490
250		480
250		20
250		15
250		10

AMPLITUDE 500

Nota: Dados hipotéticos

Variância

Vimos que a amplitude não é indicada para medir a dispersão, pois necessitamos de uma medida que considere todos os dados da série e não somente os extremos. Para exemplificar, consideremos que o produtor de milho que vendia metade das sacas colhidas por R\$22,00 e metade por R\$18,00 tivesse colhido 8 sacas. O preço médio, como já dito, é R\$20,00 e é interessante determinar quão distante de R\$20,00 está cada preço de venda. Para tanto, subtraímos 20 de cada preço de venda obtendo os dados apresentados na segunda coluna da Tabela 10, que chamaremos de **desvios em relação à média**.

Tabela 10. Preços hipotéticos de venda de milho e desvios desses preços em relação à média

Preço de venda da saca de milho	Distância da média (desvios em relação à média)
R\$/sc	
22,00	2 (= R\$22,00 – R\$20,00)
22,00	2 (= R\$22,00 – R\$20,00)
22,00	2 (= R\$22,00 – R\$20,00)
22,00	2 (= R\$22,00 – R\$20,00)
18,00	-2 (= R\$18,00 – R\$20,00)
18,00	-2 (= R\$18,00 – R\$20,00)
18,00	-2 (= R\$18,00 – R\$20,00)
18,00	-2 (= R\$18,00 – R\$20,00)
Média dos preços de venda: R\$20,00	Soma dos desvios: zero

Nota: Dados hipotéticos

Somando todos os desvios em relação à média, teremos sempre um valor igual a zero e, portanto, a soma dos desvios em relação à média não pode ser uma medida de dispersão. Se, de outra forma, somarmos os quadrados dos desvios em relação à média, teremos um número positivo (Tabela 11). Observe que o quadrado de $2 = (2)^2 = 4$ e que o quadrado de $-2 = (-2)^2 = 4$. Notamos que esses valores são todos positivos sendo a soma dos mesmos um valor diferente de zero e, portanto, possível de ser quantificada.

Tabela 11. Quadrado dos desvios dos preços hipotéticos de venda da saca do milho

X_i = Preço de venda da saca de milho	Distância da média (desvios em relação à média)	Distância da média ao quadrado (desvios em relação à média ao quadrado)
----- R\$/sc -----		
22,00 ($=X_1$)	2	4
22,00 ($=X_2$)	2	4
22,00 ($=X_3$)	2	4
22,00 ($=X_4$)	2	4
18,00 ($=X_5$)	-2	4
18,00 ($=X_6$)	-2	4
18,00 ($=X_7$)	-2	4
18,00 ($=X_8$)	-2	4
Soma	0	32

Nota: Dados hipotéticos

A soma dos quadrados dos desvios poderia ser, então, uma boa medida de dispersão se não fosse pelo fato de o seu valor aumentar quando a quantidade de dados da série aumenta, mesmo quando os valores acrescentados não são distintos daqueles já existentes. Para contornar esse problema, podemos calcular uma média da soma do quadrado dos desvios a partir da divisão do valor obtido pelo número total de dados da série. Essa medida corresponde à **variância**.

Assim, a fórmula para calcular a variância de uma variável X quando estamos considerando uma população, sendo N o número de dados da população ($1, 2, 3, \dots, N$) é (Eq.(1)):

$$VarPop(X) = \frac{\sum_{i=1}^n (X_i - \mu)^2}{N} = \frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \dots + (X_N - \mu)^2}{N} \quad (1)$$

onde, μ é a média populacional da variável X , N é o número de dados da população e $i = 1, 2, 3, \dots, N$ representa o índice de cada observação X_i da população com N informações.

Na prática é difícil termos os dados de uma população, por exemplo, os preços de todos os produtores de arroz. Assim, quase sempre, nós trabalhamos com dados de uma amostra (por exemplo, amostra de 40 produtores de arroz). Quando esse for o caso, a fórmula para calcular a variância amostral é (Eq.(2)):

$$Var(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n - 1} \quad (2)$$

onde, n corresponde ao tamanho da amostra; \bar{X} é a média da amostra; X_i é cada valor da amostra que inicia em $i=1$ até $i=n$.

No exemplo apresentado na Tabela 11 temos então, para $n = 8$ (tamanho da amostra), o cálculo da variância amostral:

$$\begin{aligned} Var(X) &= \\ &= \frac{(22 - 20)^2 + (22 - 20)^2 + (22 - 20)^2 + (22 - 20)^2 + (18 - 20)^2 + (18 - 20)^2 + (18 - 20)^2 + (18 - 20)^2}{8 - 1} \\ &= \frac{32}{7} = 4,571 \end{aligned}$$

Assim, a variância amostral de 4,571 mede o grau de variabilidade do conjunto de dados de preço de venda de milho em relação ao seu preço médio (R\$20,00/sc).

Para um número grande de dados na amostra (acima de 400 observações), não há praticamente diferença em utilizar a segunda fórmula $Var(X)$ ou a primeira $VarPop(X)$, relativa à população. Para⁹ amostras abaixo desse valor, recomenda-se o uso da segunda fórmula dada por $Var(X)$.

Desvio padrão

Devemos lembrar que para calcular a variância temos no numerador a soma dos quadrados dos desvios e no denominador um valor que corresponde à quantidade de dados da amostra (sem unidade de medida). Assim, a unidade de medida da variância é o quadrado da unidade de medida da variável. Por exemplo, se os nossos dados se referirem a preços em reais (R\$), a unidade de medida da variância é R\$², o que pode causar problemas na interpretação. Para que isso não ocorra, podemos extrair a raiz quadrada da variância e temos uma grandeza cuja unidade de medida é igual à unidade de medida da variável que estamos trabalhando. Chamamos essa grandeza de desvio padrão. Tem-se, então, a seguinte fórmula para calcular o desvio padrão amostral (Eq.(3)):

⁹ Para maiores detalhes sobre a diferença das fórmulas de variância para população e amostra, vide propriedades dos estimadores em Bussab e Morettin (2002), p. 293.

$$DP(X) = \sqrt{Var(X)}$$

Data	Boi -- R\$/arroba*	Bezerro --- R\$/animal ----
26/09/2008	90,00	719,50
25/09/2008	89,50	719,00
24/09/2008	89,00	719,00
23/09/2008	88,00	718,00
22/09/2008	87,00	719,00
Média amostral	88,70	718,90
Variância amostral	1,45	0,30
Desvio Padrão amostral	1,20	0,55

precisaríamos de um valor para nos fornecer essa dimensão. Neste caso, deveríamos calcular o coeficiente de variação.

O coeficiente de variação (CV) é calculado da seguinte forma (Eq.(4)):

$$\text{Coeficiente de Variação (\%)} = \left(\frac{\text{Desvio padrão amostral}}{\text{média amostral}} \times 100 \right) \quad (4)$$

Ao dividirmos o desvio padrão pela média, temos uma grandeza adimensional, ou seja, sem unidade de medida. Multiplicando o valor obtido dessa divisão por 100, temos uma medida de dispersão em percentual.

O coeficiente de variação é muito útil porque nos mostra quão distantes (em média e em termos percentuais) estão os dados em relação à média da série. Assim, por exemplo, um coeficiente de variação de 20% indica que os dados da série variam, em média, 20% em relação à média da série. O CV é também uma medida muito utilizada para comparar variabilidades de diferentes amostras, com médias muito desiguais ou unidades de medida diferentes. Em literatura correlata, valores abaixo de 10% são considerados como baixos e desejáveis para CV. No caso de dados experimentais (agronômicos) esse valor pode chegar a 20% como satisfatório.

Uma tabela que apresenta valores de médias de variáveis deve também apresentar uma medida de dispersão dos dados que foram utilizados para calcular cada uma das médias. Quanto menor a dispersão dos dados utilizados no cálculo, mais representativa a média é dos dados. Embora não tão adequada como o coeficiente de variação, o simples fato de incluir a amplitude de variação, ou o valor máximo e mínimo, já é um indicativo da dispersão.

2.3.3 Medidas de associação

É comum estarmos interessados na relação entre duas variáveis, pois podemos desejar conhecer o grau da relação entre elas de modo que possamos prever melhor o resultado de uma delas quando conhecemos a realização da outra. Por exemplo, o peso de uma pessoa está relacionado com sua altura, a distância percorrida (km) está associada com o valor do frete de um produto (R\$), a quantidade de carne bovina consumida por uma família durante um mês pode depender de sua renda, entre outros. Podemos separar como medidas de associação: análises gráficas e numéricas.

Análise gráfica

Um instrumento bastante útil para verificar a associação entre duas variáveis quantitativas (expressas por número) é o gráfico de dispersão. Para exemplificar, consideraremos que um produtor de manga esteja interessado em saber se há relação entre o preço médio anual desse produto na sua região e o preço médio

desse produto em outra região tomada como referência. Para estabelecer essa relação, ele deve buscar séries históricas de preço de manga nas duas praças de interesse. Vamos supor que ele tenha as séries históricas anuais (10 anos), como as apresentadas na Tabela 13. Então, ele pode construir um gráfico com os pares de valores (X, Y) e inspecionar se os pontos apresentam uma tendência particular.

Tabela 13. Preços hipotéticos de manga em duas regiões de interesse (A e B)

Ano	Variável X	Variável Y
	Preço da manga na região A	Preço da manga na região B
1	1,20	2,10
2	2,10	4,10
3	4,10	3,50
4	4,00	6,40
5	6,00	7,40
6	6,00	4,50
7	7,30	7,30
8	8,30	9,00
9	9,50	7,50
10	9,50	6,20

Nota: Dados hipotéticos

Na Figura 11 estamos apresentando um exemplo desse gráfico construído com os dados hipotéticos da Tabela 13, representando por X, o preço da manga na região A, e por Y, o preço da manga na região B.

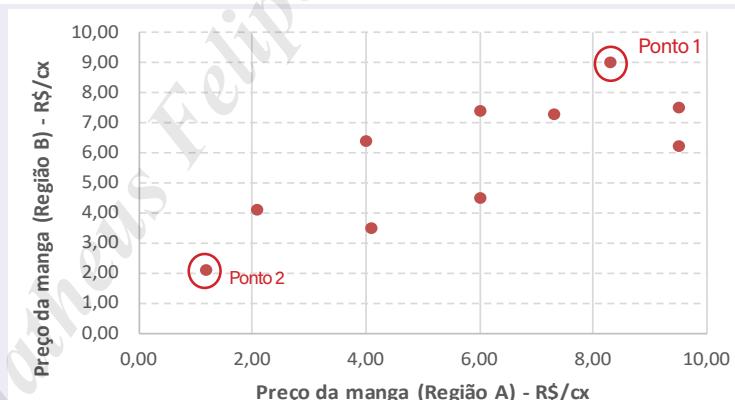


Figura 11. Dispersão dos pontos relacionando os preços de manga por caixa nas regiões A e B

Nota: Dados hipotéticos

Na Figura 11 observa-se que há uma relação positiva entre os preços da caixa de manga nas regiões A e B, porque existe uma tendência de um preço maior na região A (variável X) estar relacionado com um preço maior na região B (variável Y) (ponto 1, por exemplo) e um preço menor nessa região B estar relacionado com um preço menor na região A (ponto 2, por exemplo). Embora possam existir alguns pontos que não seguem esse padrão, na maior parte das vezes, um preço alto em A corresponde a um preço alto em B e um preço baixo em A corresponde a um preço baixo em B.

Análise numérica

Embora o gráfico de dispersão possa dar uma ideia da associação entre variáveis, muitas vezes é útil também quantificar essa relação. Existem muitos tipos de associações possíveis e aqui iremos apresentar o tipo de associação mais simples: a linear. Vamos definir uma medida que permite avaliar o quanto a nuvem de pontos no gráfico de dispersão se aproxima de uma reta (relação linear). Essa medida é definida de modo a variar entre -1 e 1 , e é denominada **coeficiente de correlação**. Feita a análise e obtido o coeficiente de correlação é possível saber, em primeiro lugar, se a relação é positiva (direta) ou negativa (inversa) e o grau de associação entre as variáveis: forte ou fraco.

Se o valor obtido para o coeficiente de correlação for próximo de 1 , então as variáveis são positivamente (diretamente) e fortemente relacionadas; se o valor for próximo de -1 , as variáveis são negativamente (inversamente) e fortemente relacionadas. Quando não houver relação **linear** entre as variáveis, o coeficiente de correlação obtido deve estar próximo de zero. Exemplos de padrões de associação (positiva, negativa ou ausente) entre variáveis são mostrados nos gráficos de dispersão da Figura 12.

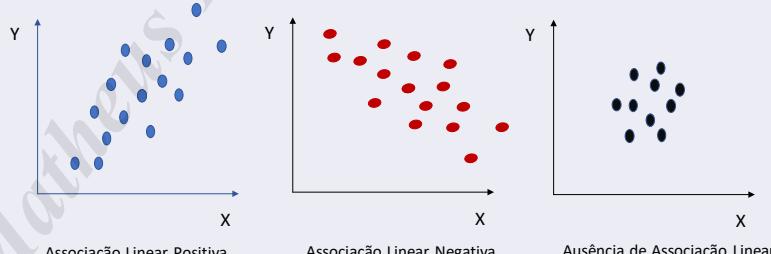


Figura 12. Tipos de associação entre duas variáveis

Nota: Dados hipotéticos

Sobre o grau do coeficiente (forte ou fraco) podemos classificá-lo conforme ilustração da Figura 13:

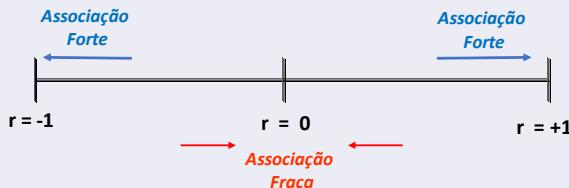


Figura 13. Níveis de associação dados pelo coeficiente de correlação

O cálculo do coeficiente de correlação é feito considerando a Equação 5:

$$Corr_{x,y} = r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} \quad (5)$$

Note que $x = (X_i - \bar{X})$ e $y = (Y_i - \bar{Y})$ correspondem às variáveis centradas (cada valor da variável subtraída da sua média).

onde, n é o tamanho das amostras tanto as referentes a variável X como da Y (elas possuem o mesmo tamanho em análise de associação); $i = 1, 2, \dots, n$; \bar{X} é a média da série X; \bar{Y} é a média da série Y.

Para exemplificar, vamos calcular o coeficiente de correlação (Tabela 14) para os dados hipotéticos de preços de manga nas regiões A e B, apresentados na Tabela 13.

Tabela 14. Cálculos intermediários necessários para obter o coeficiente de correlação entre os preços da manga nas regiões A (X_i) e B (Y_i)

Ano (i)	X_i	Y_i	$x = (X_i - \bar{X})$	$y = (Y_i - \bar{Y})$	xy	x^2	y^2
1	1,20	2,10	-4,60	-3,70	17,02	21,16	13,69
2	2,10	4,10	-3,70	-1,70	6,29	13,69	2,89
3	4,10	3,50	-1,70	-2,30	3,91	2,89	5,29
4	4,00	6,40	-1,80	0,60	-1,08	3,24	0,36
5	6,00	7,40	0,20	1,60	0,32	0,04	2,56
6	6,00	4,50	0,20	-1,30	-0,26	0,04	1,69
7	7,30	7,30	1,50	1,50	2,25	2,25	2,25
8	8,30	9,00	2,50	3,20	8,00	6,25	10,24
9	9,50	7,50	3,70	1,70	6,29	13,69	2,89
10	9,50	6,20	3,70	0,40	1,48	13,69	0,16
Soma	58,00	58,00	0,00	0,00	44,22	76,94	42,02
Média	$\bar{X} = 5,80$	$\bar{Y} = 5,80$	-	-	-	-	-

Nota: Dados hipotéticos de preços

Fonte: Elaborado pelo autor.

Note que os únicos valores do produto xy que aparecem com sinal negativo referem-se àqueles pontos nos quais o valor acima da média de X está relacionado ao valor abaixo da média de Y e o valor abaixo da média de X está relacionado ao valor acima da média de Y. Todos os demais têm sinal positivo.

Então, o coeficiente de correlação entre os preços da manga nas regiões A e B (dados por X e Y, respectivamente) vale:

$$\begin{aligned} \text{Corr}_{x,y} = r &= \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} = \frac{44,22}{\sqrt{\sum 76,94} \sqrt{\sum 42,02}} = \frac{44,22}{8,77 \times 6,48} = \frac{44,22}{56,83} \\ &= 0,78 \end{aligned}$$

Sendo o coeficiente de correlação + 0,78, temos que a associação dos preços de manga entre as regiões A e B pode ser considerada **forte** e **positiva** (ou direta).

Assim, para interpretação dos sinais do coeficiente de correlação, temos:

- Valores positivos: indicam a tendência de uma variável aumentar quando a outra aumenta (relação direta ou positiva).
- Valores negativos: indicam que valores altos de uma variável estão associados a valores baixos da outra (relação inversa ou negativa).

Algumas referências na literatura discriminam como “forte” valores iguais ou maiores que 0,8 até +1 no caso de valores positivos e, no caso de valores negativos como menores ou iguais a -0,8 até -1.

Note que esse conceito estatístico de correlação não indica “causa” de uma variável em relação à outra. O diagrama de dispersão e o coeficiente de correlação não apontarão a existência de causalidade das variáveis. A teoria econômica, no entanto, pode dar alguma indicação nesse sentido. Suponha que estejamos correlacionando a renda per capita da população ao consumo de um produto específico, como o leite, por exemplo. Se essas variáveis forem altamente correlacionadas, é plausível supor que o nível de consumo depende da renda, sendo “causa” de sua variação e não o contrário. Em alguns casos é difícil concluir sobre a causalidade das variáveis. Para exemplificar, vamos considerar que seja observada correlação entre aluguéis de fazendas e produção da lavoura. Podemos suspeitar que a produção das lavouras influencia os aluguéis. No entanto, a explicação pode ser outra: fazendeiros pagando aluguéis mais altos podem ser forçados a cultivar intensamente, e assim obtêm uma alta produção da lavoura.

2.3.4 Medidas de formato ou da forma da distribuição

Tanto a média como o desvio padrão podem não ser medidas isoladamente adequadas para representar um conjunto de dados, pois são afetados por valores extremos e, apenas com esses dois valores não temos a ideia da simetria ou assimetria da distribuição dos dados.

Observamos que histogramas podem ser calculados por meio de distribuições de frequência relativa, sendo notável a forma da distribuição: simétrica ou assimétrica (à esquerda ou à direita). Assim, temos que a medida chamada assimetria é uma medida numérica importante sobre a forma da distribuição que pode ser visualizada pelos histogramas.

Na Figura 14 temos ilustrações de histogramas que indicam a assimetria de três distribuições.

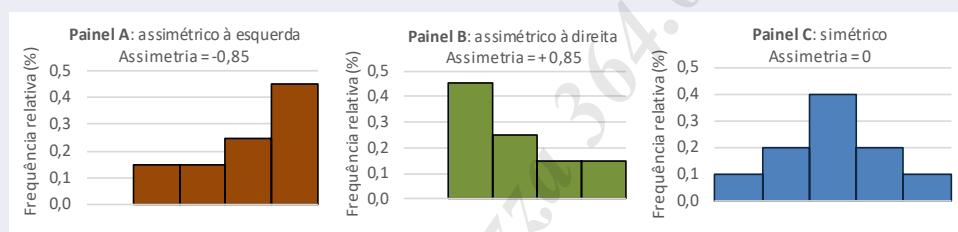


Figura 14. Ilustrações de histogramas com diferentes graus de assimetria

Nota: Dados hipotéticos

Os painéis A e B apresentam histogramas inclinados, sendo o primeiro inclinado à esquerda, e o segundo inclinado à direita. A assimetria de ambos equivale a -0,85 e +0,85, respectivamente. O cálculo da assimetria¹⁰ é complexo sendo prontamente calculado por meio de softwares estatísticos; os valores de assimetria da Figura 14 são meramente ilustrativos.

Para dados inclinados à esquerda, a assimetria é negativa; para dados inclinados à direita, a assimetria é positiva; se os dados são simétricos, a assimetria é nula.

Para uma distribuição simétrica (Painel C da Figura 14), a média, a mediana e a moda são iguais. Quando os dados possuem assimetria positiva (Painel B), a média geralmente será maior que a mediana que por sua vez é maior que a moda, ou seja, temos Média > Mediana > Moda; quando os dados possuem assimetria negativa (Painel A), geralmente a média será

¹⁰ A fórmula de cálculo da assimetria pode ser consultada em Anderson et al. (2007), apêndices 3.1 e 3.2.

menor que a mediana que por sua vez é menor que a moda, ou seja, temos Média < Mediana < Moda.

A distribuição de variáveis socioeconômicas é frequentemente assimétrica. A distribuição de salários em uma indústria, a distribuição de renda e a distribuição dos estabelecimentos agrícolas conforme sua área, no Brasil, mostram acentuada assimetria positiva. Nesse caso temos Média > Mediana > Moda, então qual medida devemos usar? Não há uma resposta geral, pois nesse caso, a escolha de uma das alternativas vai depender do foco da análise.

Como exemplo, usaremos o caso de um litígio salarial. Do ponto de vista do empregador (que deseja redução de custos), o salário médio seria a melhor opção; do ponto de vista dos empregados, o uso do salário modal ou mediano seria o mais conveniente, pois refletem melhor a situação da maioria dos empregados (menores valores).

É importante também, sobre o formato da distribuição, conhecermos a informação sobre como os dados se distribuem ao longo do intervalo entre o maior e o menor valor. Nesse contexto, temos a medida de **quantil**, que pode ser representada por percentis, quartis ou mediana.

Assim, para dados que não têm muitos valores repetidos, o p -ésimo percentil divide os dados da amostra em duas partes. Uma parte são valores menores que esse p -ésimo percentil e outra parte são valores maiores que esse p -ésimo percentil. Primeiramente vamos apresentar a fórmula de cálculo e, na sequência explicaremos a sua interpretação para melhor compreensão deste conceito.

A fórmula de cálculo do p -ésimo percentil é dada pela Equação 6:

$$i = \left(\frac{p}{100} \right) \times n \quad (6)$$

onde, n é o tamanho da amostra de dados; p é o percentil que se deseja encontrar; e, i é um índice que apontará em qual posição dentro da amostra estará o valor que corresponderá ao percentil procurado.

Vamos supor uma amostra de salários mensais (R\$) de 6 recém-formados em administração, **em ordem crescente**:

AMOSTRA DE SALÁRIOS: 2.710 2.850 2.920 **3.050** 3.130 3.325

Desejamos encontrar o **65º** percentil desses dados. Como calcular?

Calcularemos o índice dado pela Equação (6):

$$i = \left(\frac{p}{100} \right) \times n = \left(\frac{65}{100} \right) \times 6 = 3,9$$

Consideremos:

➤ Se o valor deste índice i não for um número inteiro, arredondaremos para cima. Esse número inteiro arredondado representa a posição do p -ésimo percentil.

➤ Se o valor deste índice i for um número inteiro, o p -ésimo percentil é dado pela média dos valores nas posições i e $i+1$.

Assim, para nosso exemplo, o índice i do 65º percentil resulta em **3,9**, que não é um número inteiro. Logo, arredondaremos para um número inteiro superior, e obteremos o valor **4**. Esse número indica a posição do salário na amostra em ordem crescente, que representa o 65º percentil. Assim, o 65º percentil da amostra de salário de recém-formados é o salário **R\$3.050,00**.

O que esse valor significa? Que aproximadamente **65%** dos recém-formados recebem salários menores que **R\$3.050,00**, ou dito de outra forma, que **35% (= 100% - 65%)** dos recém-formados recebem salários acima de **R\$3.050,00**.

E se o valor do índice i fosse um valor inteiro? Para este caso, vamos supor que desejamos encontrar o 50º percentil da amostra. Calculamos então o índice i :

$$i = \left(\frac{p}{100} \right) \times n = \left(\frac{50}{100} \right) \times 6 = 3,0$$

Neste caso, o índice i do 50º percentil resulta no valor **3**, que é um número inteiro. Então, na amostra em ordem crescente, tomamos os valores de salários que ocupam as posições 3ª e 4ª (i e $i+1$, respectivamente) e realizamos uma média com eles. Logo, temos da amostra: $(R\$2.920 + R\$3.050)/2 = R\$2.985,00$.

O que esse valor significa? Que aproximadamente **50%** dos recém-formados recebem salários menores que **R\$2.985,00**, ou dito de outra forma, que **50% (= 100% - 50%)** dos recém-formados recebem salários acima de **R\$2.985,00**. Notemos que o valor dado pelo 50º percentil corresponde à mediana.

Podemos também desejar dividir os dados da amostra em quatro partes, tendo cada parte aproximadamente 25% das observações (= um quarto das observações). Assim, temos:

*Q1 = primeiro quartil = 25º percentil.

*Q2 = segundo quartil = 50º percentil = mediana.

*Q3 = terceiro quartil = 75º percentil.

É importante notar que para o cálculo dos quartis ou percentis, usamos sempre a fórmula do índice i .

Vamos considerar outro exemplo para entendermos a utilidade dessa medida. Algumas universidades registram notas de exames de admissão em termos de percentis. Suponha que um candidato obtenha a nota 54 pontos de um exame de admissão. O desempenho desse estudante em relação aos outros que fizeram o mesmo exame pode não ser claro imediatamente. Considere que o 70º percentil das notas de todos os estudantes candidatos corresponde a 51 pontos. Isso significa que aproximadamente 70% dos estudantes tiveram pontuações menores que 51, ou ainda, que 30% dos estudantes tiveram notas superiores a 51. Então, o candidato que estamos analisando, cuja nota é de 54 pontos, está na minoria (30%) que apresentou notas mais elevadas (acima de 50 pontos). Parece ser um bom candidato a admissão, não?

RECAPITULANDO

Os métodos de estatística descritiva se resumem em dois grupos: métodos tabulares e gráficos e métodos numéricos. No contexto de métodos tabulares e gráficos, a distribuição de frequências de uma variável discreta é adequadamente representada por um gráfico de barras. Uma distribuição de frequências de uma variável contínua pode ser representada por um histograma. Os métodos numéricos são utilizados para resumir os dados de análise. Neste caso, usamos medidas de tendência central (média, moda e mediana, por exemplo) e medidas de dispersão (amplitude, variância, desvio padrão e coeficiente de variação). Uma medida de tendência central deve ser sempre acompanhada de uma medida de dispersão. Quanto menos dispersos são os dados em torno da medida de tendência central, mais representativa ela é daqueles dados. Também é importante considerarmos medidas de associação linear de duas variáveis pois podemos analisar se a relação entre elas é forte ou fraca, direta ou inversa: um exemplo é o coeficiente de correlação linear. Medidas de associação linear não necessariamente sugerem alguma relação de causa entre as variáveis em questão. Como complemento de análise de dados, é importante considerar medidas de formato, principalmente se os dados apresentam valores muito discrepantes. Essas medidas permitem obter informações sobre a assimetria da distribuição de frequência da amostra e sobre a posição que determinado valor da amostra está em relação aos demais, em termos de quartis ou percentis, por exemplo.

Referências

- Anderson, D.R.; Sweeney, D.J.; Williams, T. 2007. Estatística Aplicada à Administração e Economia. Pioneira Thompson Learning: São Paulo, SP, Brasil. p. 21-107
- Bonfanti, S.E. 2016. Principais causas de mortalidade em frangos de corte Griller criados em sistema intensivo Dark House. Disponível em: <<https://rd.uffs.edu.br/bitstream/prefix/418/1/BONFANTI.pdf>>. Acesso em: 29 jul. 2020.
- Bussab, W.O.; Morettin, P.A. 2006. Estatística Básica. Saraiva: São Paulo, SP, Brasil. p.9-51.
- Centro de Estudos Avançados em Economia Aplicada (CEPEA). 2008. Preços Agropecuários. Disponível em: <<https://www.cepea.esalq.usp.br/br>>. Acesso em: 19 ago. 2020.
- Hoffmann, R. 2006. Estatística para Economistas. Pioneira Thomson Learning: São Paulo, SP, Brasil. p. 23-52; 81-105.
- Sartoris, A. 2003. Estatística e Introdução à econometria. Saraiva: São Paulo, SP, Brasil. p. 31-48.

3. Probabilidades: experimentos, resultados experimentais e variáveis aleatórias

Observamos que a análise de um conjunto de dados por meio de técnicas numéricas e gráficas permite que tenhamos uma boa ideia da distribuição desse conjunto. Como já discutido, a distribuição de frequência nos permite avaliar a variabilidade das observações de um fenômeno. A partir dessas frequências podemos calcular medidas de estatística descritiva como medidas de posição e variabilidade. Todas essas medidas calculadas, juntamente com as frequências são estimativas de quantidades desconhecidas, associadas às populações das quais os dados foram extraídos na forma de amostras (que nos permitem realizar inferências). Quando consideramos as frequências relativas, estamos considerando estimativas de probabilidades de ocorrências de certos eventos de interesse. Diante de suposições adequadas, podemos criar um modelo teórico que reproduza de maneira razoável a distribuição de frequências que nos permite a obtenção de probabilidades (dentro de experimentos) que são úteis em processos de tomadas de decisões.

3.1 Definições

Para cada **experimento** teremos todos os seus resultados possíveis (**resultados experimentais ou eventos**). Assim, a Tabela 15 ilustra exemplos de experimentos e seus resultados.

Tabela 15. Exemplos de experimentos e resultados possíveis

Experimento	Resultados experimentais (=eventos)
Jogar uma moeda	Cara, coroa
Selecionar uma peça para inspeção	Defeituosa, não defeituosa
Fazer um contato de vendas	Comprar, não comprar
Lançar um dado	1,2,3,4,5,6
Jogar uma partida de futebol	Ganhar, perder, empatar

Fonte: Anderson et al. (2007).

Assim, ao especificar todos os resultados possíveis de um experimento, identificamos o seu **espaço amostral**.

Podemos atribuir probabilidades a esses resultados do espaço amostral que são úteis no auxílio às tomadas de decisões. Gerentes, por exemplo, fundamentam suas decisões em uma análise de incertezas, como por exemplo: quais as chances de queda de vendas se aumentarmos os preços? Qual é a probabilidade de o projeto ser concluído no prazo? Qual é a chance de um novo investimento ser lucrativo?

Então, **probabilidade** é uma medida numérica da possibilidade de um evento ocorrer. Podemos usar probabilidades como medidas do grau de incerteza associado aos eventos supracitados. Se houver probabilidades disponíveis, podemos determinar a possibilidade de cada um dos eventos acontecer.

Valores probabilísticos sempre são atribuídos em escala de zero a um, sendo mais próxima de zero, menor a probabilidade de um evento ocorrer e, mais próxima de um, maior a probabilidade de o evento ocorrer.

A probabilidade (P) é expressa pela razão entre as ocorrências que satisfazem o evento de interesse e todos os resultados possíveis (Eq.(7)):

$$P(\text{evento}) = \frac{\text{número de ocorrências do evento}}{\text{número total de resultados possíveis}} \quad (7)$$

Se utilizarmos como exemplo o primeiro experimento da Tabela 15, temos o experimento “jogar uma moeda”. O total de resultados (eventos) possíveis nesse experimento corresponde a dois: cara e coroa. Se quisermos conhecer a probabilidade de ocorrência do evento “sair cara”, por exemplo, teremos a $P(\text{cara}) = \frac{1}{2} = 0,5$ ou 50% (vide Equação 7). Para alguns experimentos, as probabilidades de seus eventos podem não ser conhecidas, como por exemplo, a probabilidade de comprar e de não comprar certo produto. Diante disso, podemos atribuir probabilidades aos resultados experimentais por meio de três abordagens: método clássico ou teórico, de frequência relativa ou empírico e o subjetivo.

Método Clássico ou Teórico: é apropriado quando todos os eventos (resultados experimentais) são igualmente prováveis. Neste caso, a probabilidade é estimada previamente a partir do conhecimento teórico sobre as chances de ocorrência de um evento (resultado experimental) dentro de um conjunto limitado (espaço amostral) de possibilidades igualmente prováveis. Se n resultados experimentais são possíveis, a probabilidade de $1/n$ é atribuída a cada resultado experimental. No lançamento de uma moeda, por exemplo, a probabilidade de obtermos cara é igual a probabilidade de obtermos coroa – são igualmente prováveis ($=1/2$).

Método da Frequência relativa ou empírico: quando a probabilidade é definida a partir de dados observados e não no conhecimento teórico prévio. É apropriado quando se tem dados disponíveis para estimar a proporção do tempo, por exemplo, em que o resultado experimental ocorrerá se o experimento for repetido inúmeras vezes. Consideremos o exemplo da Tabela 16, em que temos um estudo sobre o tempo de espera de pacientes para serem atendidos no pronto atendimento de um hospital. Um atendente registrou, por 20 dias consecutivos, o número de pacientes à espera deste atendimento, exposto na Tabela 16:

Tabela 16. Experimento com número de pessoas em espera em um hospital

Número de pessoas à espera	Número de dias em que o resultado ocorreu	Frequência relativa
0	2	$2/20 = 0,10$
1	5	$5/20 = 0,25$
2	6	$6/20 = 0,30$
3	4	$4/20 = 0,20$
4	3	$3/20 = 0,15$
-	20 = Total de dias	Soma = 1,00

Fonte: Anderson et al. (2007).

Os dados do experimento da Tabela 16 mostram que em dois dos 20 dias, nenhum (0) paciente esteve à espera de um pronto atendimento no hospital. Usando o método de frequência relativa teríamos a probabilidade de 10% ($=2/20$) de nenhum paciente estar à espera de atendimento nesses 20 dias de experimento; teríamos que a maior probabilidade (30%) seria atribuída à ocorrência de dois pacientes estarem à espera de atendimento, durante todo o experimento, e assim para os demais resultados.

Método subjetivo: quando a probabilidade é atribuída por um indivíduo, sem qualquer conhecimento objetivo, seja ele teórico ou empírico. É apropriado quando não se pode presumir realisticamente que os resultados

experimentais são igualmente prováveis e quando poucos dados relevantes estão disponíveis. Sob esse método, podemos esperar que diferentes pessoas atribuam diferentes probabilidades ao mesmo resultado experimental (evento). Por exemplo, a probabilidade de sucesso de um produto ser lançado, atribuída por seu criador com base em sua experiência profissional. Ele poderia associar essa probabilidade de sucesso como sendo 0,6 ou 60%. Outro profissional ou pesquisador, com base em informações de mercado, poderia atribuir a essa probabilidade um número mais conservador, por exemplo, 0,4 ou 40%.

Independentemente do método, para se **atribuir probabilidades** devemos sempre considerar dois requisitos básicos para tal finalidade:

1) A probabilidade atribuída a cada um dos resultados experimentais (eventos) deve situar-se entre zero e um (inclusive). Se considerarmos que E denota o **Evento** (ou **um resultado experimental**), esse evento terá $P(E)$ como sua probabilidade de ocorrência; então, teremos que $P(E)$ assumirá valores entre zero e um. Genericamente, se tivermos n resultados experimentais (ou eventos), representados por $i = 1, 2, \dots, n$, temos que E_i denota o i -ésimo resultado experimental e $P(E_i)$ denota as probabilidades dos i -ésimos eventos (Eq.(8)):

$$0 \leq P(E_i) \leq 1 \text{ para todo } i \quad (8)$$

2) A soma das probabilidades de todos os resultados experimentais (eventos), deve ser igual a um. Para n resultados experimentais (ou eventos), esse requisito pode ser escrito (Eq.(9)):

$$P(E_1) + P(E_2) + \dots + P(E_n) = 1 \text{ sendo } i = 1, 2, \dots, n \quad (9)$$

Por exemplo, em um experimento de **uma** jogada de moeda, temos dois eventos (ou dois resultados experimentais): cara e coroa. Então, cara seria o Evento 1 (E_1) e coroa, o Evento 2 (E_2). As respectivas probabilidades de ocorrência (pela fórmula de probabilidade) seriam, respectivamente $P(E_1)$ e $P(E_2)$, ou seja, $P(E_1)=1/2$ e $P(E_2)=1/2$. Assim os dois requisitos básicos seriam atendidos à medida que as probabilidades dos dois eventos (ou resultados experimentais) do experimento estão entre zero e um ($0 \leq 1/2 \leq 1$), e a soma de $P(E_1)$ e $P(E_2)$ resulta em valor 1,0 ($P(E_1) + P(E_2) = 1/2 + 1/2 = 1,0$).

3.2 Variável aleatória (v.a.)

A variável aleatória fornece um meio para se descrever os resultados experimentais (eventos) usando-se valores numéricos (vide Figura 15).

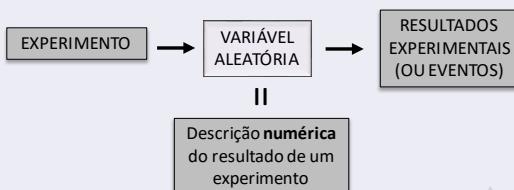


Figura 15. Definição de variável aleatória

Fonte: Elaboração pelo autor.

Devemos, portanto, considerar uma variável aleatória aquela que:

- associa um valor numérico a cada resultado experimental possível;
- cujo valor numérico depende do resultado do experimento;
- cada um de seus possíveis valores se associa a uma probabilidade;
- classifica-se em discreta ou contínua, dependendo dos valores numéricos que assume.

3.2.1 Variáveis aleatórias discretas

Uma variável aleatória que pode assumir um número finito de valores ou uma sequência infinita de valores que pertencem a um conjunto de números inteiros ($0, 1, 2, \dots$), é denominada **variável aleatória discreta**.

Alguns exemplos de variáveis aleatórias discretas:

Exemplo 1: certa variedade de planta exige o cumprimento de quatro condições edafoclimáticas (clima, temperatura, umidade e radiação) para sucesso de crescimento; X = variável aleatória discreta = número de condições edafoclimáticas que a variedade atinge em certas regiões ($0, 1, 2, 3, 4$) = **número finito de valores inteiros**. Em uma região B, a variedade atinge apenas duas condições edafoclimáticas ($x = 2$), em uma região C, a variedade não atinge nenhuma condição edafoclimática ($x = 0$) ou ainda em uma região D, a variedade atinge as quatro condições edafoclimáticas ($x = 4$).

Exemplo 2: X representa uma variável aleatória discreta que corresponde ao número de caminhões com carga agrícola que chegam no pedágio no período de 1 dia; X = **números inteiros e infinitos**.

As variáveis qualitativas binárias são um tipo específico de v.a. discreta,

cujos valores representam a ocorrência de uma qualidade e são representados por valores discretos (0 e 1), sendo 0 (zero) = ausência da qualidade e 1 (um) = presença de qualidade. **Exemplo:** pesquisa solicita ao indivíduo que relembrar a marca (ou o nome do produtor) de melão vendido na região nordeste brasileira. Então, **X = variável aleatória discreta e qualitativa = número finito de valores inteiros**, sendo: **X = 1** para os indivíduos que lembram e **X = 0** para os que não lembram.

3.2.2 Variáveis aleatórias contínuas

Quando os valores de uma variável aleatória fazem parte de um intervalo de números reais com infinitas possibilidades entre dois valores quaisquer, ela é denominada **variável aleatória contínua**.

A v.a. contínua assume qualquer valor numérico em um intervalo ou em uma coleção de intervalos, como por exemplo, peso, distância, temperatura, tempo, etc.

Alguns exemplos de variáveis aleatórias contínuas:

Exemplo 1: consideremos o serviço de mecânica de caminhões no atendimento de ocorrências em um trecho de 100 km. Seja X = v.a. contínua que representa o número de km até o local da próxima ocorrência ao longo do trecho de 100 km. Então, X = qualquer valor entre 0 e 100.

Exemplo 2: consideremos que X = v.a. contínua que representa o tempo para colheita de 1 ha de cana.

Exemplo 3: consideremos o enchimento de uma embalagem de defensivo agrícola que possui capacidade para 20 litros. Seja X = v.a. contínua que representa a quantidade real em litros de embalagens de defensivos. Então, X = qualquer valor entre 0 e 20 litros.

RECAPITULANDO

Probabilidade é uma medida numérica da possibilidade de um evento ocorrer. Podemos atribuir probabilidades aos resultados experimentais por meio de três abordagens: método clássico ou teórico, o de frequência relativa ou empírico, e o subjetivo. A variável aleatória fornece um meio para se descrever os resultados experimentais (eventos) usando-se valores numéricos. Ela pode ser discreta (assume um número finito de valores inteiros), por exemplo: gênero, idade, número de pessoas, etc.; ou contínua (faz parte de um intervalo de números reais com infinitas possibilidades entre dois valores quaisquer), por exemplo: peso, distância, temperatura, tempo, etc.

Referências

- Anderson, D.R.; Sweeney, D.J.; Williams, T. 2007. Estatística Aplicada à Administração e Economia. Pioneira Thompson Learning: São Paulo, SP, Brasil. p.129-143; 169-171.
- Bussab, W.O.; Morettin, P.A. 2006. Estatística Básica. Saraiva: São Paulo, SP. Brasil. p.103-106.; 128-137; 162-172.
- Hoffmann, R. 2006. Estatística para Economistas. Pioneira Thomson Learning: São Paulo, SP, Brasil. p. 57-58.
- Sartoris, A. 2003. Estatística e Introdução à econometria. Saraiva: São Paulo, SP, Brasil. p.1-31.

4. Distribuição de Probabilidade

A distribuição de probabilidade de uma variável aleatória (v.a.) descreve como as probabilidades estão distribuídas sobre os valores dessa v.a. Dada uma variável aleatória X , as respectivas probabilidades de ocorrência de cada um de seus valores descreverão a distribuição de probabilidade de v.a. discretas; já as probabilidades dos valores da v.a. estarem em intervalos de valores denotarão a distribuição de probabilidade de v.a. contínuas.

Destaque-se que a principal vantagem de definir uma variável aleatória e sua distribuição de probabilidade é que, uma vez que a distribuição de probabilidade seja conhecida, torna-se relativamente fácil determinar a probabilidade de uma série de eventos que podem ser de interesse do tomador de decisões. Os dois casos de distribuição de probabilidade (discreta e contínua) serão descritos a seguir, bem como respectivos exemplos.

4.1 Distribuições discretas de probabilidade

Para uma variável aleatória **discreta X** , a distribuição de probabilidade é definida por uma **função de probabilidade**, denotada por $f(X)$. Esta função de probabilidade fornece a probabilidade correspondente a cada um dos valores da variável aleatória.

Como exemplo de variável aleatória discreta e sua respectiva função de distribuição, consideremos as vendas de tratores em um dia, durante um experimento de 300 dias. Dados de vendas desse experimento apontam que 54 dias desse período de 300 apresentaram zero vendas cada um; que 117 dias desse experimento apresentaram uma venda cada um; que 72 dias tiveram duas vendas cada um; que 42 dias exibiram 3 vendas cada um; que 12 dias apresentaram 4 vendas cada e que apenas 3 dias apresentaram 5 vendas cada um. Esse detalhamento pode ser explicitado na Tabela 17, a seguir:

Tabela 17. Experimento de vendas diárias de máquinas agrícolas, durante 300 dias

X = v.a. discreta (vendas diárias de máquinas agrícolas)	Número de dias (parcela do experimento, em dias, que apresentou X vendas diárias)	f(X) Distribuição discreta de probabilidade
0	54	$f(0) = 54/300 = 0,18$
1	117	$f(1) = 117/300 = 0,39$
2	72	$f(2) = 72/300 = 0,24$
3	42	$f(3) = 42/300 = 0,14$
4	12	$f(4) = 12/300 = 0,04$
5	3	$f(5) = 3/300 = 0,01$
Total	-	1,00 ou 100%
	300 dias	

Fonte: Adaptado de Anderson et al. (2007).

Pelos dados da Tabela 17 temos que a variável aleatória discreta é representada por X, que corresponde ao número de vendas diárias de máquinas agrícolas. Também, que a distribuição de probabilidade pode ser calculada a partir dos valores de frequência relativa. Ressalte-se que a frequência relativa pode ser obtida neste caso, a partir da razão do número de dias, conforme o número de vendas diárias de X, pelo total de dias do experimento (300 dias).

Notamos pelos dados da Tabela 17 que as duas condições necessárias para que uma função de probabilidade discreta seja válida, são atendidas, ou seja: as probabilidades de ocorrência de cada valor da variável aleatória ($f(X)$) situam-se entre zero e um, sendo, portanto, positivas e, a soma dessas probabilidades resulta em 1 (ou 100%).

É interessante notar que podemos entender o processo de vendas de máquinas agrícolas analisando a distribuição de probabilidade. Por exemplo, qual seria a probabilidade de vender três ou mais máquinas agrícolas por dia? Resposta: essa probabilidade seria de 19% pois, deveríamos somar as probabilidades $f(3) + f(4) + f(5) = 0,14 + 0,04 + 0,01 = 0,19$ ou 19%. Ou ainda, qual seria a probabilidade de vender até duas máquinas agrícolas por dia? Resposta: essa probabilidade seria de 81% $f(0) + f(1) + f(2) = 0,18 + 0,39 + 0,24 = 0,81$ ou 81%. Devemos notar que a probabilidade de vender exatamente duas máquinas agrícolas seria 24%.

Poderíamos observar também por meio da distribuição de probabilidade discreta desse exemplo que a venda de uma máquina agrícola por dia é o evento (resultado experimental) mais provável (variável aleatória que apresenta o maior valor da distribuição de probabilidade, que é 39%).

Os resultados desse experimento apresentados na Tabela 17 poderiam também ser representados por meio de um gráfico de barras (Figura 16):

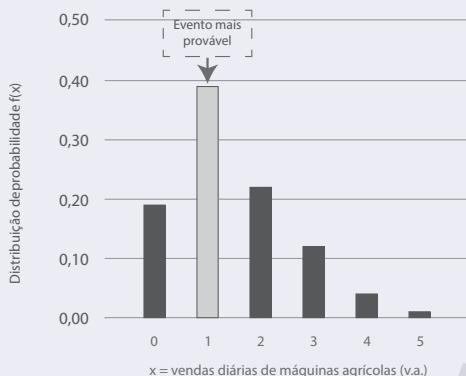


Figura 16. Representação gráfica da distribuição de probabilidade discreta para o número de máquinas agrícolas vendidas por um dia, no total de 300 dias

Fonte: Adaptado de Anderson et al. (2007).

4.1.1 Valor esperado ou Esperança

O valor esperado é a medida da posição central da variável aleatória (ou valor médio da variável aleatória). Para uma v.a. discreta, o valor esperado é dado pela média ponderada de todos os resultados experimentais, sendo os pesos iguais às probabilidades de ocorrência de cada um desses resultados. O procedimento é análogo à estimativa da média para distribuições de frequências.

É calculado pela Equação 10:

$$E(X) = \mu = \sum X f(X) \quad (10)$$

onde, X representa os valores da variável aleatória discreta e $f(X)$ corresponde aos valores das respectivas probabilidades de ocorrência (que constituem a distribuição de probabilidade discreta). Para o cálculo do valor esperado ($E(X)$) devemos multiplicar cada valor da v.a. pela respectiva probabilidade ($f(X)$) e então adicionar os produtos resultantes.

Assim, para o exemplo de vendas diárias de máquinas agrícolas (Tabela 17), para o cálculo do valor esperado das vendas diárias do experimento, teríamos:

$$\begin{aligned} E(X) &= \mu \\ &= (0 \times 0,18) + (1 \times 0,39) + (2 \times 0,24) + (3 \times 0,14) + (4 \times 0,04) + (5 \times 0,01) \\ &= 1,5 \end{aligned}$$

Sabemos, portanto, que, embora seja possível a realização de 0, 1, 2, 3, 4 e 5 vendas de máquinas agrícolas em qualquer um dos dias, ao longo do tempo, a concessionária pode prever a venda de uma média de 1,5 máquinas agrícolas por dia. Cabe comentar que o valor esperado não precisa ser um valor que a variável aleatória possa assumir. Aplicação: supondo 30 dias (1 mês) de operação da concessionária, podemos usar o valor esperado de 1,5 para prever vendas mensais médias de 30 (dias) x 1,5 (valor esperado) = 45 máquinas agrícolas.

4.1.2 Variância

A variância, como já descrito, é uma medida de variabilidade que, associada a uma medida de tendência central, pode fornecer a dispersão dos dados da amostra em relação à sua média ou valor esperado. Portanto, variância de uma v.a. é uma medida de quão dispersos os valores estão em torno do valor esperado.

No caso de uma v.a. discreta, a variância é a média ponderada dos desvios elevados ao quadrado que uma variável aleatória sofre a partir de sua média. Os pesos são as probabilidades ($f(X)$). É calculada pela Equação 11:

$$Var(X) = \sigma^2 = \sum (X - E(X))^2 \times f(X) \quad (11)$$

onde, X corresponde aos valores da variável aleatória discreta, $E(X)$ é o valor esperado ou a média e $f(X)$ representa os valores das respectivas probabilidades de ocorrência dos valores da v.a. (que constituem a distribuição de probabilidade de discreta). Para o cálculo da variância da v.a. discreta devemos multiplicar o valor dos desvios ao quadrado (obtidos pela diferença de cada variável aleatória em relação ao valor esperado) pela respectiva probabilidade ($f(X)$) e então adicionar os produtos resultantes.

Utilizando o exemplo da Tabela 17 de vendas diárias de máquinas agrícolas, teríamos o cálculo da variância detalhado na Tabela 18.

Tabela 18. Detalhamento do cálculo da variância da v.a. discreta X = vendas diárias de máquinas agrícolas

X = v.a. discreta	Desvios $((X - E(X))$)	Desvios ao quadrado $((X - E(X))^2$)	$f(X)$	$((X - E(X))^2 \times f(X))$
0	0-1,5 = -1,5	2,25	$f(0) = 0,18$	$2,25(0,18) = 0,4050$
1	1-1,5 = -0,5	0,25	$f(1) = 0,39$	$0,25(0,39) = 0,0975$
2	2-1,5 = +0,5	0,25	$f(2) = 0,24$	$0,25(0,24) = 0,0600$
3	3-1,5 = +1,5	2,25	$f(3) = 0,14$	$2,25(0,14) = 0,3150$
4	4-1,5 = +2,5	6,25	$f(4) = 0,04$	$6,25(0,04) = 0,2500$
5	5-1,5 = +3,5	12,25	$f(5) = 0,01$	$12,25(0,01) = 0,1225$
				Variância = Soma = 1,25

Fonte: Adaptado de Anderson et al. (2007).

Assim, de acordo com a Equação (11), a **variância** equivale a $\sum(X - E(X))^2 \times f(X) = 0,4050 + 0,0975 + \dots + 0,1225 = 1,25$. Como já relatado em outros capítulos, podemos extrair a raiz quadrada da variância e obter o

desvio padrão, que é uma medida preferida para descrever a variabilidade dos dados, uma vez que a variância é uma medida elevada ao quadrado, sendo, portanto, mais difícil de ser interpretada. Logo, o **desvio padrão** será $\sqrt{1,25} = 1,118$

Se tivéssemos outra amostra referente às vendas diárias de máquinas agrícolas para alguma comparação ou escolha, conforme a que apresentasse menor variabilidade dos dados, optaríamos, talvez, pela amostra que apresentasse o menor valor de desvio padrão. Não havendo outra amostra para comparação, poderíamos calcular o coeficiente de variação (= razão entre o desvio padrão e a média, multiplicada por 100); no caso da variável aleatória, seria a razão entre o desvio padrão e o valor esperado, multiplicada por 100. Realizando esse cálculo para o exemplo supracitado, obteríamos $(1,118/1,5) \times 100 = 74,53\%$, aproximadamente.

Os resultados indicam que há uma variabilidade elevada, ou seja, considerando o valor do coeficiente de variação, notamos que os dados da série de vendas diárias de máquinas variam 74,53% em relação a sua média. A dispersão média é de 1,25 máquinas agrícolas (variância), valor quase equivalente ao seu valor esperado (1,5). Ressalte-se que o coeficiente de variação (CV) é também útil para comparar variabilidades de diferentes amostras, com médias muito desiguais ou unidades de medida diferentes.

4.1.3 Distribuição uniforme de probabilidade discreta

O exemplo mais simples de distribuição de probabilidade discreta é o caso de uma **distribuição uniforme de probabilidade discreta**. Sua função pode ser definida pela Equação 12:

$$f(X) = \frac{1}{n} \quad (12)$$

onde, n representa o número de valores que a variável aleatória discreta pode assumir.

Por exemplo, considere o experimento de lançamento de um dado. Definimos X como a v.a. discreta que representa o número que aparece na face virada para cima do dado ao ser lançado. Assim, X pode assumir qualquer valor das faces do dado, ou seja, X pode assumir os valores: $X = 1,2,3,4,5,6$. Logo, temos $n = 6$ possíveis valores para a variável aleatória. Então a função de probabilidade discreta uniforme corresponde a $f(X) = 1/n = 1/6$; observamos que o valor da probabilidade de qualquer uma das faces acontecer (face virada para cima) é a mesma e equivale a $1/n$, ou seja, $1/6$ (por isso uniforme).

As distribuições de probabilidade discretas mais amplamente usadas são, de forma geral, especificadas por expressões matemáticas. A uniforme é a mais simples de todas e, as mais conhecidas e importantes são a distribuição binomial, a distribuição de Poisson e a distribuição hipergeométrica. Nesta abordagem discutiremos apenas a binomial.

4.1.4 Distribuição de probabilidade binomial

A distribuição de probabilidade binomial é uma distribuição de probabilidade discreta que está associada a um experimento de múltiplas etapas que chamaremos de experimento binomial.

Para que um experimento seja considerado como **experimento binomial**, as **quatro propriedades** a seguir **devem ser atendidas**:

1) O experimento apresenta uma sequência de n ensaios idênticos.

2) Dois resultados são possíveis em cada ensaio: sucesso e fracasso.

3) A probabilidade de um sucesso, denotado por p , não se modifica de ensaio para ensaio; consequentemente, a probabilidade de fracasso, denotada por $(1-p)$, não se modifica de ensaio para ensaio. Diante disso temos que a soma das probabilidades de sucesso e fracasso de um ensaio corresponde a 1 ou 100%, ou seja, $p + (1-p) = 1$ ou 100%.

4) Os ensaios são independentes.

Se o experimento incluir apenas as propriedades 2, 3 e 4, então teremos um evento de Bernoulli (número de ensaios = $n = 1$); se incluir as quatro propriedades, então teremos um experimento binomial.

Consideremos o seguinte exemplo para checarmos se é considerado experimento binomial: **um vendedor de seguros visita 10 produtores (em suas propriedades agrícolas)**. Consideremos como sucesso, o fato de o produtor comprar apólice de seguro para máquinas agrícolas (logo, fracasso seria não comprar). Suponhamos que a probabilidade de sucesso (de compra) é conhecida (**por meio de um método subjetivo**) e vale $p = 0,1$.

Vamos analisar as quatro propriedades que denotam um experimento binomial para esse exemplo:

1. O experimento apresenta uma sequência de n ensaios idênticos. É **VÁLIDA**, pois teremos 10 visitas do vendedor ou ensaios idênticos.
2. Dois resultados são possíveis em cada ensaio: sucesso e fracasso. É **VÁLIDA**, pois temos dois resultados possíveis (comprar e não comprar), sendo sucesso e fracasso, respectivamente.
3. A probabilidade de um sucesso, denotado por p , não se modifica de ensaio para ensaio. É **VÁLIDA**, pois a probabilidade de sucesso

- $(p = 0,1)$ e, por consequência, de fracasso $(1-p = 0,9)$ não se alteram de um evento para outro;
4. Ensaios são independentes. É **VÁLIDA**, pois o resultado de cada visita não influencia no resultado de outra.

Portanto, o experimento do exemplo é considerado um experimento binomial. É válido destacar que poderíamos ter o seguinte cenário: à medida que o dia fosse passando, a cada visita, o vendedor poderia se sentir cansado ou esgotado e então, a probabilidade de sucesso (considerada constante no tempo e igual a $p = 0,1$) poderia diminuir. Se isso fosse considerado em questão, então esse experimento não seria um experimento binomial, pois a propriedade 3 não seria validada.

Antes de utilizarmos outro exemplo e entendermos a utilidade da distribuição binomial, é válido considerar que, o que importa em um experimento binomial é o número de sucessos nos n ensaios, ou seja, a variável aleatória discreta (X) corresponderá ao número de sucessos nos n ensaios independentes. Se X é o número de resultados favoráveis em n ensaios, então X é uma variável aleatória discreta que pode assumir valores entre zero e $(n + 1)$. Para o exemplo das vendas de seguros supracitado, temos que X = produtor visitado a comprar apólice de seguro. Como são 10 visitas (ou 10 ensaios independentes), temos que X pode assumir valores de 0 a 10, ou seja, assumir 0 compra ou 10 compras (se todos os 10 produtores visitados comprarem o seguro). Logo, $X = 0,1,2,3,4,5,6,7,8,9,10$, ou seja, X assume 11 valores (pois contamos com o valor zero de compras); genericamente X pode assumir $n+1 = 10 + 1 = 11$ valores de compras. Dizemos então que X tem distribuição binomial com parâmetros n e p .

Consideremos um próximo exemplo, que atende as quatro propriedades do experimento binomial, para que possamos realizar algumas análises de possíveis resultados interessantes.

Exemplo para análise: consideremos 3 clientes que entram em uma loja, em momentos distintos e independentes, para comprar implementos agrícolas. Com base na experiência do gerente da loja (método subjetivo), sabe-se que a probabilidade de o cliente comprar vale 30%, ou seja, $p = 0,3$.

A pergunta para análise é: qual a probabilidade de 2 clientes realizarem a compra?

Antes de tudo, checamos as quatro propriedades e descobrimos que é um experimento binomial. Logo, a variável aleatória discreta X = número de clientes que efetuam a compra = número de sucessos. Temos que X assume 4 valores $(0,1,2,3)$ pois X assume $n+1$ valores $(3+1)$ e temos que $n = 3$ (ensaios independentes) e $p = 0,3$ (probabilidade de sucesso).

Primeiramente vamos descobrir quantas possibilidades teremos para $X = 2$: nesse caso teremos 3 possibilidades, ou seja, para que dois clientes efetuem a compra, temos a possibilidade de ser o 1º e o 2º ou o 1º e o 3º ou o 2º e o 3º. Para sermos didáticos, faremos um diagrama em árvore para entender essas possibilidades de 2 clientes efetuarem a compra, ou seja, de $X = 2$ (Figura 17).

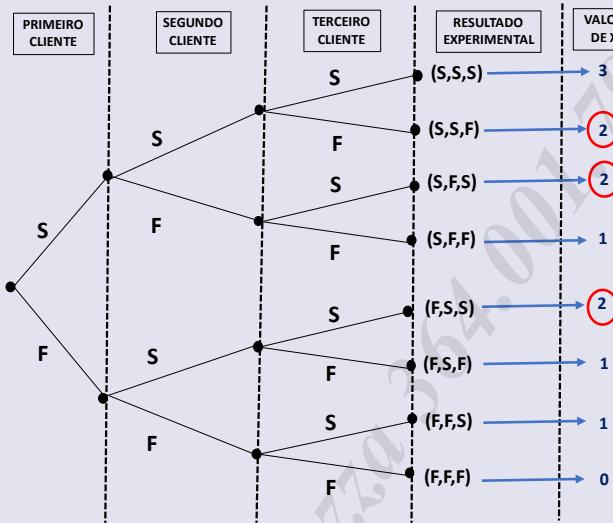


Figura 17. Diagrama em árvore para o problema de compras e não compras por 3 clientes

Nota: "S" representa sucesso (compra) e "F" representa fracasso (não compra)

Fonte: Adaptado de Anderson et al. (2007).

O diagrama da Figura 17 ilustra as possibilidades de X assumir os quatro valores (0,1,2,3). Também, ilustra as 3 possibilidades de $X = 2$ (circulado em cor vermelha), dadas pelos resultados experimentais (S,S,F) , (S,F,S) e (F,S,S) . Claro que para esse exemplo, conseguimos facilmente identificar 3 possibilidades de $X = 2$, mas, se tivéssemos um exemplo com o número de ensaios (n) muito grande e muitas possibilidades de ocorrência de um determinado valor de X , teríamos que utilizar uma fórmula combinatória, dada a seguir (Eq.(13)):

$$\binom{n}{X} = \frac{n!}{X!(n-X)!} \quad (13)$$

onde, n corresponde ao número de ensaios do experimento binomial; X corresponde ao número de sucessos.

Consideraremos também (Eq.(14)):

$$n! = n \times (n - 1) \times (n - 2) \times (n - 3) \times \cdots \times (2) \times (1) \text{ e } 0! = 1 \quad (14)$$

Temos que a Equação (13) é usada para calcular o número de resultados experimentais que fornecem exatamente X sucessos em n ensaios. Para o nosso exemplo, já visto pelo diagrama de árvore, poderíamos encontrar o número de resultados experimentais para $X = 2$, a partir de:

$$\binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{3 \times 2 \times 1}{2 \times 1 \times (1)} = \frac{6}{2} = 3$$

$$(S,S,F) = p \times p \times (1-p) = 0,3 \times 0,3 \times (1-0,3) = (0,3)^2 \times (0,7) = \mathbf{0,063}$$

Resultados experimentais*	Probabilidade do resultado experimental
(S,S,F)	$p \times p \times (1-p) = p^2 (1-p) = 0,3^2 (0,7) = \mathbf{0,063}$
(S,F,S)	$p \times (1-p) \times p = p^2 (1-p) = 0,3^2 (0,7) = \mathbf{0,063}$
(F,S,S)	$(1-p) \times p \times p = p^2 (1-p) = 0,3^2 (0,7) = \mathbf{0,063}$

Percebemos pelos resultados da Tabela 19, que todos os três resultados experimentais com dois sucessos e um fracasso têm exatamente a mesma probabilidade. Isso se mantém como regra, ou seja, em qualquer experimento binomial todas as sequências de resultados de ensaio que produzem X sucessos em n ensaios têm a mesma probabilidade de ocorrência.

Assim, podemos generalizar a probabilidade de cada sequência de ensaios, produzir X sucessos em n ensaios é dada por (Eq.(15)):

$$p^X(1-p)^{n-X}$$

$$f(X) = \binom{n}{X} p^X(1-p)^{n-X}$$

$$\binom{n}{X} = \frac{n!}{X!(n-X)!} \quad (17)$$

onde, p = probabilidade de sucesso em qualquer um dos ensaios; $(1-p)$ = probabilidade de fracasso em qualquer um dos ensaios.

Então, para o nosso exemplo, para responder à pergunta: qual a **probabilidade de 2 clientes realizarem compra?** Bastaria aplicar a Equação 16 da função de probabilidade binomial, e obteríamos:

$$f(X) = \binom{n}{X} p^X (1-p)^{n-X} = \binom{3}{2} 0,3^{0,2} (1-0,3)^{3-2} = 3 \times 0,068 = \mathbf{0,189 \text{ ou } 18,9\%}$$

Logo, se estamos convencidos de que uma situação exibe as propriedades de um experimento binomial (sendo um experimento binomial), e se conhecemos os valores de n e p , podemos usar a Equação 16 para calcular a probabilidade de X sucessos nos n ensaios.

Se tivéssemos 10 clientes entrando na loja para comprar implementos agrícolas, em vez de três, a função de probabilidade binomial é aplicável. Se desejássemos, nesse novo cenário, calcular a probabilidade de serem realizadas **exatamente** quatro vendas para dez clientes, considerando $p = 0,3$, teríamos:

$$\begin{aligned} f(X) &= \binom{n}{X} p^X (1-p)^{n-X} = f(4) = \binom{10}{4} 0,3^4 (1-0,3)^{10-4} = \\ &= \frac{10!}{4!(10-4)!} 0,3^4 (0,7)^6 = \mathbf{0,2001 \text{ ou } 20\% \text{ aprox.}} \end{aligned}$$

$$f(0) = \binom{10}{0} 0,3^0 (1-0,3)^{10-0} = 1 \times 1 \times (0,7)^{10} = \mathbf{0,028 \text{ aprox.}}$$

$$f(1) = \binom{10}{1} 0,3^1 (1-0,3)^{10-1} = 10 \times 0,3 \times (0,7)^9 = \mathbf{0,121 \text{ aprox.}}$$

Para valor esperado:

$$E(X) = \mu = n \times p$$

$$Var(X) = \sigma^2 = n \times p(1 - p)$$

$$E(X) = \mu = n \times p = 3 \times 0,3 = \mathbf{0,9}$$

$$Var(X) = \sigma^2 = n \times p(1 - p) = 3 \times 0,3 \times (1 - 0,3) = \mathbf{0,63}$$

$$DesPad(X) = \sigma = \sqrt{Var(X)} = \sqrt{0,63} = \mathbf{0,79}$$

Explorando um pouco esses resultados, poderíamos desejar conhecer o número esperado de clientes que fariam compra de implementos agrícolas, considerando uma quantidade de 1.000 clientes em vez de apenas três. Neste caso teríamos $E(X) = \mu = n \times p = 1.000 \times 0,3 = \mathbf{300}$. Assim, para aumentar o número esperado de vendas, os vendedores precisariam convencer mais clientes a entrar na concessionária (deveríamos aumentar o número de ensaios) e/ou aumentar a probabilidade (p) de um cliente individual qualquer realizar uma compra após entrar na concessionária.

Em relação a um cenário com 1.000 clientes e a mesma probabilidade de sucesso (compra), ou seja, $p = 0,3$, teríamos os valores para a variância e desvio padrão, respectivamente:

$$Var(X) = \sigma^2 = n \times p(1 - p) = 1.000 \times 0,3 \times (1 - 0,3) = \mathbf{210}$$

$$DesPad(X) = \sigma = \sqrt{Var(X)} = \sqrt{210} = \mathbf{14,49}$$

estamos considerando n maior (número de ensaios). Se calcularmos o coeficiente de variação de ambas as situações, teremos:

$$CV_{3\text{ clientes}} = \left(\frac{DesvPad(X)}{E(X)} \right) \times 100 = \frac{0,79}{0,90} \times 100 = 87\% \text{ } \textbf{aprox.}$$

$$CV_{1.000\text{ clientes}} = \left(\frac{DesvPad(X)}{E(X)} \right) \times 100 = \frac{14,49}{300} \times 100 = 4,83\% \text{ } \textbf{aprox.}$$

distribuída. Antes de apresentarmos a fórmula de cálculo da função de densidade de probabilidade, para melhor entendimento, utilizaremos um exemplo.

Exemplo: suponha que a v.a. seja X = tempo de voo de um avião que vai de a para b . Suponha que o tempo de voo possa ter qualquer valor no intervalo de 120 a 140 minutos. Uma vez que a variável aleatória X pode assumir qualquer valor desse intervalo, X é uma variável aleatória contínua, e não uma v.a. discreta. Diante de dados disponíveis, podemos concluir que a probabilidade de tempo de voo no intervalo de 1 minuto qualquer tenha a mesma probabilidade de tempo de voo em outro intervalo de 1 minuto, todos contidos no intervalo total de 120 a 140 minutos. Se considerarmos que cada um dos intervalos de 1 minuto é igualmente provável, dizemos que a variável aleatória contínua tem uma distribuição uniforme de probabilidade. Assim, a função de densidade de probabilidade, a qual define a distribuição uniforme de probabilidade, correspondente à variável aleatória “tempo de voo” é (Eq. (20)):

$$f(X) = \begin{cases} \frac{1}{b-a} & \text{para } a \leq X \leq b \\ 0 & \text{outro ponto qualquer} \end{cases} \quad (20)$$

Com os dados do exemplo teremos (Eq. (21)):

$$f(X) = \begin{cases} 1/20 & \text{para } 120 \leq X \leq 140, \text{ sendo } a = 120 \text{ e } b = 140 \\ 0 & \text{outro ponto qualquer} \end{cases} \quad (21)$$

A função de densidade de probabilidade de tempo de voo é a mesma para qualquer intervalo de tempo, dentro do intervalo estabelecido por a e b (120 e 140 minutos, respectivamente), ou seja, de $1/20$. A Figura 18 vai nos ajudar a compreender esse cálculo, apresentando um gráfico da função densidade de probabilidade.

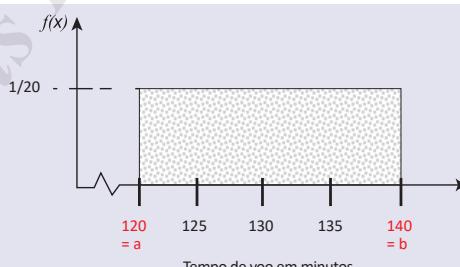


Figura 18. Função de densidade uniforme de probabilidade de tempos de voo
Fonte: Adaptado de Anderson et al. (2007).

Percebemos pela Figura 18 que a função de densidade de probabilidade vale $f(X) = 1/20$ e, como já descrito, ela não fornece a probabilidade de ocorrência de um intervalo da v.a. contínua diretamente. Por meio do cálculo da área sob a função de densidade de probabilidade, observaremos a probabilidade de ocorrência da v.a. X estar em algum intervalo específico. Por exemplo, qual seria a probabilidade de o tempo de voo situar-se entre 120 e 130 minutos? Ou seja, qual seria $P(120 \leq X \leq 130) = ?$.

Para encontrarmos essa probabilidade, bastaria calcular a área mais escura da Figura 19, demarcada por um retângulo, sendo sua base dada pela diferença entre os minutos (130 e 120) e sua altura dada pela função de densidade uniforme de probabilidade, já calculada por $1/20$.

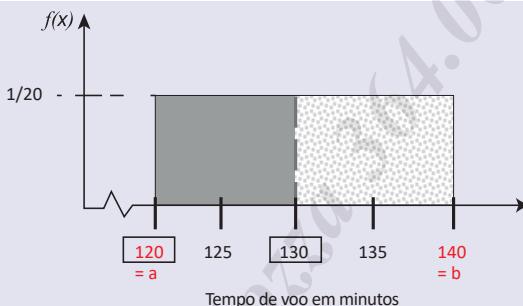


Figura 19. Área que fornece a probabilidade do tempo de voo entre 120 e 130 minutos
Fonte: Adaptado de Anderson et al. (2007).

Assim, a área do retângulo escuro, dada por base x altura, é calculada por $(130 - 120) \times (1/20) = 1/2$ ou $0,50$ ou **50%**. Visto que o tempo de voo precisa estar entre 120 e 140 minutos (a e b , respectivamente), sendo a distribuição de probabilidade descrita como uniforme neste intervalo, temos que $P(120 \leq X \leq 130) = 0,50$.

Dada a distribuição uniforme do tempo de voo, e usando a área como uma probabilidade, podemos responder a quaisquer questões probabilísticas sobre os tempos de voo. Por exemplo, qual seria a probabilidade de ocorrência de um tempo de voo entre 128 e 136 minutos? A largura desse intervalo é $(136 - 128 = 8)$ e a altura de $f(X) = 1/20$. Logo, multiplicando a base pela altura da área retangular denotada por esses valores, teríamos $8 \times 1/20 = 0,40$. A representação seria $P(128 \leq X \leq 136) = 0,40$ ou **40%**. Vale observar que $(120 \leq X \leq 140) = 1$ ou **100%** ou seja, a área **total** sob o gráfico de $f(X)$ é igual a **1** ou **100%**.

Portanto, a probabilidade de uma variável aleatória contínua assumir um valor dentro de um determinado intervalo entre a e b é definida como a área sob o gráfico da função de densidade de probabilidade ($f(X)$) que se encontra entre a e b . Uma vez que um ponto simples é um intervalo de largura zero, temos que

a probabilidade de uma v.a. contínua assumir de maneira exata qualquer valor exato em particular, vale zero (a área sob $f(X)$ seria zero).

Assim como no caso de v.a. discretas, em v.a. contínuas (X), a $f(X)$ correspondente à função de densidade de probabilidade deverá ser $f(X) \geq 0$ para todos os valores de X . Outro fato a ser observado é que no caso da função de probabilidade discreta, a soma das probabilidades deve ser igual a um; no caso de função de probabilidade contínua, isso também é observado à medida que a **área total** sob a função de densidade de probabilidade ($f(X)$) for **igual a um**.

4.2.1.1 Valor esperado e variância

Calcular o valor esperado e variância de uma variável aleatória contínua é análogo ao cálculo efetuado para v.a. discretas, no entanto envolve cálculo integral, disponível em literaturas mais avançadas. Apresentaremos aqui apenas as fórmulas do **valor esperado e da variância** para a **distribuição contínua uniforme de probabilidade**, a seguir (Eq. (21 e 22)):

Valor esperado:

$$E(X) = \frac{a + b}{2}$$

$$Var(X) = \frac{(b - a)^2}{12} \quad (22)$$

Sendo a o menor valor e b o maior valor que a v.a. pode assumir (do intervalo total estabelecido).

Para o exemplo do tempo de voo, temos que esse intervalo total corresponde a 120 e 140 minutos (a e b, respectivamente). Então o valor esperado, a variância, o desvio padrão e o coeficiente de variação (CV) valem, respectivamente:

$$E(X) = \frac{a + b}{2} = \frac{120 + 140}{2} = \frac{160}{2} = \mathbf{130 \text{ minutos}}$$

$$Var(X) = \frac{(b - a)^2}{12} = \frac{(140 - 120)^2}{12} = \mathbf{33,33}$$

$$DesvPad(X) = \sqrt{Var(X)} = \sqrt{\frac{(b - a)^2}{12}} = \sqrt{33,33} = \mathbf{5,77 \text{ minutos}}$$

$$CV(X) = \frac{DesvPad(X)}{E(X)} \times 100 = \frac{5,77}{130} \times 100 = \mathbf{4,44\% \text{ aprox.}}$$

Logo, o tempo médio de voo, diante de vários voos realizados, seria de **130 minutos**.

Pelo coeficiente de variação, teríamos que os tempos de voo nos intervalos de tempo possíveis entre 120 e 140 minutos variam cerca de 4,44% em relação a sua média (130 minutos).

4.2.2 Distribuição Normal de probabilidade

A Distribuição Normal de Probabilidade é a mais importante para descrever uma variável aleatória contínua, com ampla variedade de aplicações práticas como: altura e peso das pessoas, notas de exames, medições científicas, índices pluviométricos, preços, etc. Para melhor entendimento dessa distribuição, primeiramente relataremos as características da curva normal e então faremos menção ao cálculo das probabilidades de ocorrências da v.a. contínua estar em determinado intervalo.

4.2.2.1 Curva Normal

Para entendermos a forma da distribuição normal de probabilidade, voltemos à distribuição binomial. Vamos supor que, inicialmente temos um experimento binomial com dois ensaios ($n = 2$), probabilidade de sucesso $p = 0,5$. Como já visto, teremos $n + 1$ que a variável aleatória discreta pode assumir (X), ou seja, três valores. Isso é válido conforme aumentamos o número de ensaios (n) para a mesma probabilidade de sucesso (p). Se disponibilizarmos o histograma das distribuições binomiais considerando $n = 2, n = 5, n = 10$, teremos a seguinte representação na Figura 20:

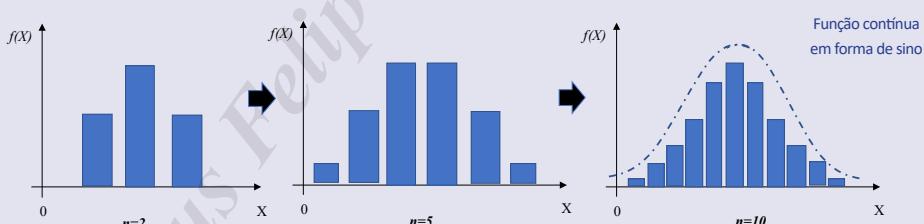


Figura 20. Histogramas de distribuição binomial, conforme o aumento do número de ensaios (n)

Fonte: Adaptado de Sartoris (2003).

Pela Figura 20 notamos que, se aumentássemos o número de ensaios (n) indefinidamente (2, 3, 10,), de forma que os retângulos do histograma se aproximassesem cada vez mais ou os pontos de um gráfico comum se colidissem, teríamos uma função contínua, cuja forma seria de sino.

O formato da distribuição normal de probabilidade, ilustrada na Figura 21, é de uma curva em forma de sino. Sua origem remonta a Gauss em seus trabalhos sobre erros de observações astronômicas (em 1810), designando essa distribuição como distribuição gaussiana ou distribuição normal:



Figura 21. Curva da função de densidade de probabilidade ($f(X)$) em forma de sino correspondente à distribuição normal de probabilidade (gaussiana)

Nota: μ representa a média e σ corresponde ao desvio padrão de X

Fonte: Adaptado de Bussab e Moretin (2002).

A função de densidade de probabilidade (f.d.p.) que define essa curva em forma de sino da distribuição normal de probabilidade é dada pela Equação 23:

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2}$$

$$X \sim N(\mu, \sigma)$$

- O ponto máximo da curva normal encontra-se na média μ , que também corresponde à mediana e à moda da distribuição.
- A média da distribuição pode ser qualquer valor numérico: negativo, nulo ou positivo (vide Figura 22).

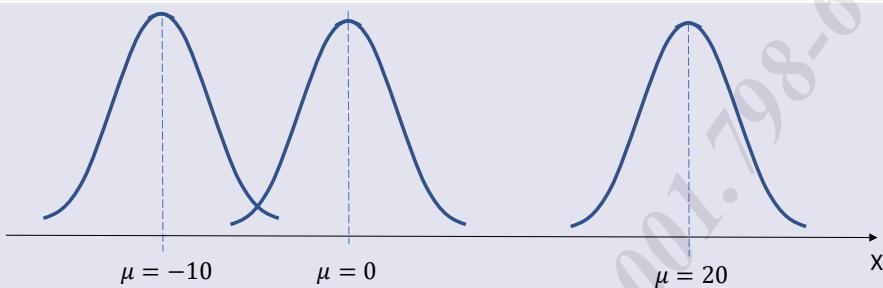


Figura 22. Curvas da função de densidade de probabilidade ($f(X)$) em forma de sino correspondente à distribuição normal de probabilidade para valores de médias negativo, nulo e positivo

Fonte: Adaptado de Anderson et al. (2007).

➤ A distribuição normal é simétrica (não é inclinada sendo assimetria = 0), sendo a forma da curva à esquerda da média uma imagem espelhada da forma da curva à direita da média. Os extremos da curva (caudas) tendem para o infinito e não tocam o eixo horizontal.

➤ O desvio padrão σ determina o quanto uma curva é achatada ou larga. Quanto maior o valor do desvio padrão, mais achatada será a curva da distribuição, exibindo maior variabilidade dos dados. Vide ilustração de duas curvas com diferentes achatamentos (valores de desvios-padrão) na Figura 23.

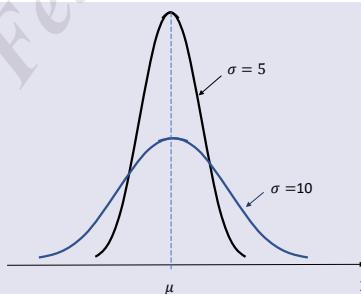


Figura 23. Curvas da função de densidade de probabilidade ($f(X)$) em forma de sino correspondente à distribuição normal de probabilidade para mesma média e valores distintos de desvio padrão

Fonte: Adaptado de Anderson et al. (2007).

➤ As probabilidades da variável aleatória normal são dadas por áreas sob a curva de densidade de probabilidade (f.d.p.). A área total sob a curva da distribuição normal de probabilidade é 1. Já que a distribuição é simétrica, a área da curva à esquerda da média vale 0,50 e, a área sob a curva à direita da média vale 0,50.

➤ As porcentagens dos valores de alguns intervalos comumente usados são representadas pela Figura 24, detalhados na sequência.

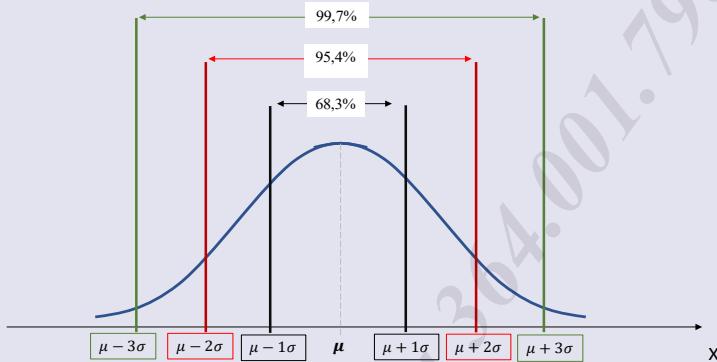


Figura 24. Porcentagens dos valores dos intervalos comuns da distribuição normal de probabilidade

Fonte: Adaptado de Anderson et al. (2007).

As porcentagens ilustradas pela Figura 24, podem ser detalhadas com:

- 68,3% aprox. dos valores de uma variável aleatória normal estão dentro de mais ou menos um desvio-padrão (σ) de sua média μ ;
- 95,4% aprox. dos valores de uma variável aleatória normal estão dentro de mais ou menos dois desvios-padrão (2σ) de sua média μ ;
- 99,7% aprox. dos valores de uma variável aleatória normal estão dentro de mais ou menos três desvios-padrão (3σ) de sua média μ .

O cálculo das probabilidades sob uma distribuição normal (sob a curva de função de densidade normal de probabilidade) pode se tornar trabalhoso, uma vez que envolve cálculo integral e, ao contrário da facilidade da função de distribuição uniforme, a área sob a f.d.p não será uma figura retangular, cujo cálculo é simples.

Assim, uma particular distribuição normal, conhecida como **distribuição normal padronizada**, que tem média zero e desvio-padrão igual a um, tem seus resultados dos cálculos das integrais já calculados e dispostos em uma tabela, chamada Tabela Z (vide Apêndice 1).

4.2.3 Distribuição Normal Padrão de Probabilidade

A distribuição normal padrão de probabilidade tem a mesma aparência geral das outras distribuições normais, porém com as propriedades especiais de $\mu = 0$ e $\sigma = 1$. A fórmula da função de densidade normal padrão de probabilidade é uma versão mais simples da Equação 25, a seguir:

$$f(X) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

a. Entre zero e um, ou seja $P(0 \leq Z \leq 1) = ?$ Vide ilustração na Figura 26.

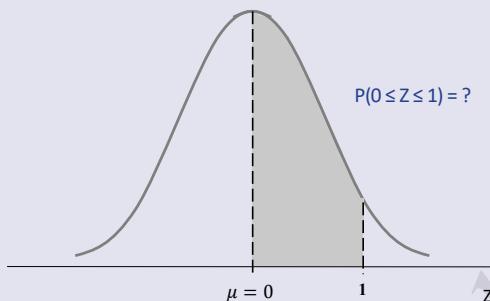


Figura 26. Ilustração da área da probabilidade a ser encontrada por meio de $P(0 \leq Z \leq 1)$

Estamos interessados na área entre $Z = 0$ e $Z = 1,00$. Então, precisamos encontrar na Tabela Z o valor correspondente a $Z = 1,00$. Primeiramente localizamos 1,0 na coluna à esquerda da Tabela Z e depois encontramos o valor 0,00 em sua linha superior (vide ilustração na Tabela 20). Cruzando esses valores no corpo da Tabela Z, descobrimos que a linha 1,0 e a coluna 0,00 se interceptam no valor 0,3413, o que nos dá a probabilidade desejada, ou seja, $P(0 \leq Z \leq 1) = 0,3413$.

Tabela 20. Parte da Tabela Z para o exemplo de cálculo de $P(0 \leq Z \leq 1)$

z	0,00	0,01	0,02
0,9			
1,0	0,3159	0,3186	0,3212
1,1	0,3413	0,3438	0,3461
1,2	0,3643	0,3665	0,3686
	0,3849	0,3869	0,3888

Fonte: Adaptado de Anderson et al. (2007).

A Tabela Z dispõe de probabilidades calculadas sempre a partir do valor da média para direita ou para a esquerda da distribuição, lembrando que a distribuição é simétrica, então a probabilidade (área sob a curva) da média (zero) para a

direita valo 0,5 e da média (zero) para a esquerda vale 0,5 também, ou seja $P(0 \leq Z \leq +\infty) = P(-\infty \leq Z \leq 0) = 0,50$.

Cabe ainda comentar que, dada a simetria da distribuição normal, $P(0 \leq Z \leq 1) = P(-1 \leq Z \leq 0) = 0,3413$.

b. Acima de (ou maior que) - 0,50, ou seja $P(Z \geq -0,5) = ?$

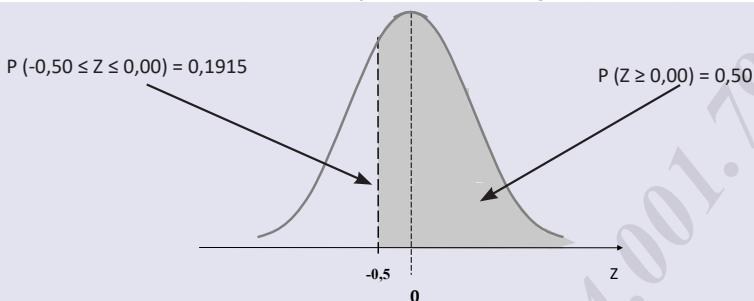


Figura 27. Ilustração da área da probabilidade a ser encontrada por meio de $P(Z \geq -0,5)$

Portanto, a área total do sombreamento (Figura 27) seria $P(Z \geq -0,5) = P(-0,50 \leq Z \leq 0,00) + P(Z \geq 0,00) = 0,1915 + 0,50 = 0,6915$.

c. Entre 1,00 e 1,58, ou seja $P(1,00 \leq Z \leq 1,58) = ?$

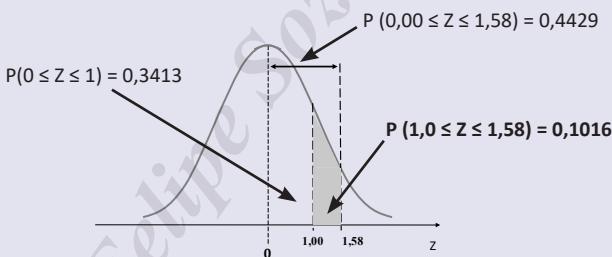


Figura 28. Ilustração da área da probabilidade a ser encontrada por meio de $P(1,0 \leq Z \leq 1,58)$

Portanto, a área total do sombreamento (Figura 28) seria $P(1,00 \leq Z \leq 1,58) = P(0,00 \leq Z \leq 1,58) - P(0,00 \leq Z \leq 1,00) = 0,4429 - 0,3413 = 0,1016$:

Agora que já entendemos o uso da Tabela Z a partir de valores da variável aleatória padronizada Z, vamos utilizar exemplos aplicados, mesmo com dados fictícios. Vamos transformar a variável aleatória contínua (X) de exemplos aplicados em v.a. padronizada (Z) e então, usar a Tabela Z para cálculo de probabilidades. Faremos a transformação da variável aleatória X em Z, pois existem valores tabelados para a variável Z, mas não para a variável aleatória (X).

Logo, para calcular probabilidades de qualquer distribuição normal, faremos a transformação da variável aleatória contínua X em padronizada (Z), antes de usar a Tabela Z. Essa transformação se dá por meio da Equação 26:

$$Z = \frac{X - \mu}{\sigma} \quad (26)$$

Vamos supor que uma v.a. tem distribuição normal, com média 100 e desvio padrão 10. Desejamos então saber qual é a probabilidade de a variável aleatória X estar entre 90 e 110, ou seja, desejamos encontrar $P(90 \leq X \leq 110)$. Devemos nos atentar que, agora, a nossa variável é X e não Z como em exemplos supracitados. Vamos transformar a variável x em padronizada (Z), usando a Equação 26. Como dados temos $\mu = 100$ e $\sigma = 10$.

Para $X = 90$, temos a seguinte padronização:

$$Z = \frac{X - \mu}{\sigma} = \frac{90 - 100}{10} = \frac{-10}{10} = -1$$

Para $X = 110$, temos a seguinte padronização:

$$Z = \frac{X - \mu}{\sigma} = \frac{110 - 100}{10} = \frac{10}{10} = +1$$

Então, para $P(90 \leq X \leq 110)$ padronizando X, temos, $P(-1 \leq Z \leq +1)$, conforme ilustrado na Figura 29:

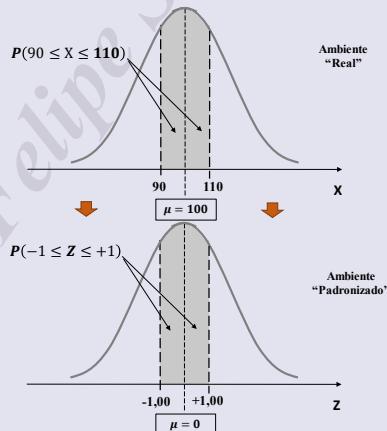


Figura 29. Ilustração da área da probabilidade a ser encontrada por meio de $P(-1,0 \leq Z \leq +1,00)$, partindo de $P(90 \leq X \leq 110)$

Utilizando a Tabela Z, encontraremos os valores das probabilidades de $Z = 0$ a $Z = -1$ e de $Z = 0$ a $Z = +1$. Na tabela teremos apenas os valores das

probabilidades de Z para valores de Z positivos; como a distribuição é uma normal simétrica, a mesma probabilidade que encontramos para os valores positivos de Z, encontraremos para os valores correspondentes negativos de Z.

Efetuando os cálculos, temos:

$$\begin{aligned} P(90 \leq X \leq 110) &= P\left(\frac{90 - 100}{10} \leq Z \leq \frac{110 - 100}{10}\right) = \\ &= P(-1,00 \leq Z \leq +1,00) = 0,3413 + 0,3413 = 0,6826 = \mathbf{68,27\%} \end{aligned}$$

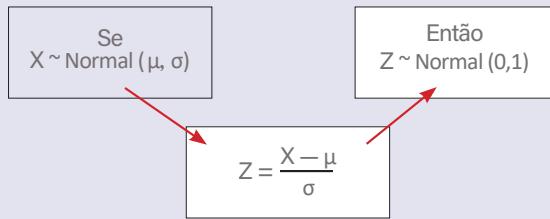


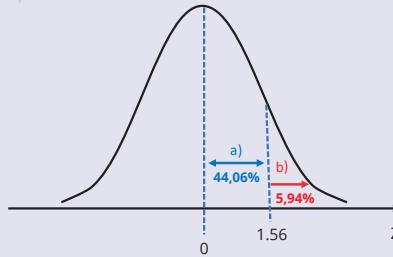
Figura 31. Ilustração da transformação da variável aleatória X em normal padronizada

Assim, teremos:

- a. Notemos aqui que o valor da média ($\mu = 0,18\%$) coincide com o valor inferior do intervalo no qual está contida a variável aleatória $P(0,18\% \leq X \leq 2,00\%)$. Isso é apenas coincidência do exemplo e em nada altera as formas de cálculo já vistas. O cálculo será dado por:

$$P(0,18\% \leq X \leq 2,00\%) = P\left(\frac{0,0018 - 0,0018}{0,0117} \leq Z \leq \frac{0,02 - 0,0018}{0,0117}\right) = \\ = P(0 \leq Z \leq 1,56) = 0,4406 = 44,06\%$$

$$P(X \leq 2,00\%) = P\left(Z \leq \frac{0,02 - 0,0018}{0,0117}\right) = \\ = P(Z \leq 1,56) = 0,5000 - 0,4406 = 5,94\%$$



Exemplo aplicado 2: suponhamos que os balancetes semestrais realizados em uma empresa mostraram que o lucro realizado se distribui normalmente com **média de \$48.000 e desvio padrão de \$8.000**. Qual é a probabilidade de que:

- No próximo semestre, o lucro seja maior que \$50.000, ou seja,
 $P(X \geq \$50.000)$

Para encontrar a área referente a essa probabilidade, devemos calcular:

$$P(X \geq \$50,000) = P(X \geq \$48,000) - P(\$48,000 \leq X \leq \$50,000)$$

Graficamente, teríamos a ilustração da Figura 33.

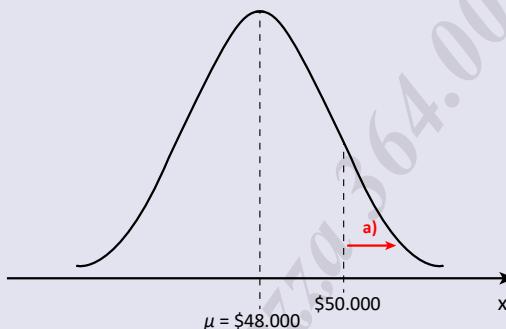


Figura 33. Ilustração da área da probabilidade a ser calculada para:

a) $P(X \geq \$50.000)$

Considerando a padronização da variável aleatória, o cálculo da probabilidade de “a” será dado por:

$$P(X \geq \$50.000) = P(X \geq \$48.000) - P(\$48.000 \leq X \leq \$50.000)$$

$$P\left(Z \geq \frac{\$50.000 - \$48.000}{\$8.000}\right) = P\left(Z \geq \frac{\$48.000 - \$48.000}{\$8.000}\right)$$

$$= P\left(\frac{\$48.000 - \$48.000}{\$8.000} \leq Z \leq \frac{\$50.000 - \$48.000}{\$8.000}\right)$$

$$= P(Z \geq 0,25) = P(Z \geq 0,00) - P(0,00 \leq Z \leq 0,25)$$

$$= P(Z \geq 0,25) = 0,5000 - 0,0987 = 0,4013 \text{ ou } 40,13\%$$

A Figura 34 facilita a visualização do cálculo das áreas para a solução do problema.

Figura 34. Ilustração da área da probabilidade encontrada por meio de $P(Z \geq 0,25)$, partindo de $P(X \geq \$50.00)$

Assim, no próximo semestre, a probabilidade de que o lucro seja maior que \$50.000 é de **40,13%**

- b. No próximo semestre, o lucro esteja entre \$40.000 e \$45.000, ou seja,
 $P(\$40.000 \leq X \leq \$45.000) = ?$

Para encontrar a área referente a essa probabilidade, devemos calcular:

$$P(\$40.000 \leq X \leq \$45.000) = P(\$40.000 \leq X \leq \$48.000) - P(\$45.000 \leq X \leq \$48.000)$$

$$\begin{aligned}
P(\$40.000 \leq X \leq \$45.000) &= \\
&= P(\$40.000 \leq X \leq \$48.000) - P(\$45.000 \leq X \leq \$48.000) = \\
&= P\left(\frac{\$40.000 - \$48.000}{\$8.000} \leq Z \leq \frac{\$45.000 - \$48.000}{\$8.000}\right) = \\
&= P\left(\frac{\$40.000 - \$48.000}{\$8.000} \leq X \leq \frac{\$48.000 - \$48.000}{\$8.000}\right) - \\
&\quad P\left(\frac{\$45.000 - \$48.000}{\$8.000} \leq X \leq \frac{\$48.000 - \$48.000}{\$8.000}\right) = \\
&= P(-1,00 \leq Z \leq -0,375) = P(-1,00 \leq Z \leq 0,00) - P(-0,375 \leq Z \leq 0,00) \\
&= P(-1,00 \leq Z \leq -0,375) = 0,3413 - 0,1443 = 0,197 \text{ ou } \mathbf{19,7\%}
\end{aligned}$$

Considerando a padronização da variável aleatória, o cálculo da probabilidade de “c” será dado por:

$$\begin{aligned}
 P(X \leq \$0) &= P(X \leq \$48.000) - P(\$0 \leq X \leq \$48.000) \\
 &= P\left(Z \leq \frac{\$0 - \$48.000}{\$8.000}\right) = P\left(Z \leq \frac{\$48.000 - \$48.000}{\$8.000}\right) - P\left(\frac{\$0 - \$48.000}{\$8.000} \leq Z \leq \frac{\$48.000 - \$48.000}{\$8.000}\right) \\
 &= P(Z \leq -6,00) = P(Z \leq 0,00) - P(-6,00 \leq Z \leq 0,00) \\
 &= P(Z \leq -6,00) = 0,50 - 0,50 = 0,00 \text{ ou } 0\% \text{ aproximadamente.}
 \end{aligned}$$

A Figura 38 facilita a visualização do cálculo das áreas para a solução do problema. Assim, a probabilidade de que no próximo semestre haja prejuízo é de praticamente **0%**.

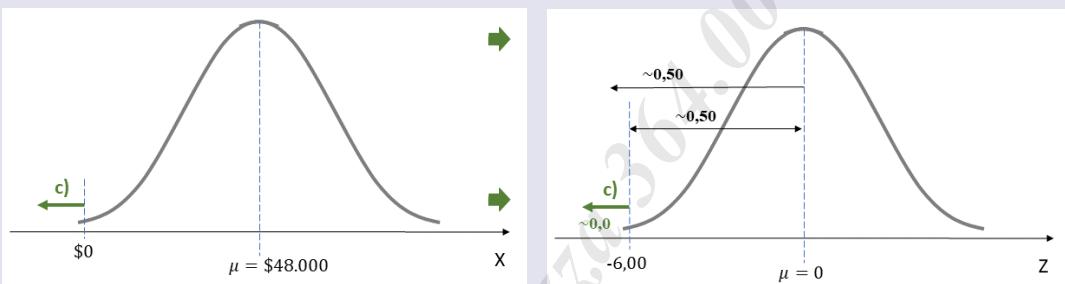


Figura 38. Ilustração da área da probabilidade encontrada por meio de partindo de $P(Z \leq -6,00)$, partindo de $P(X \leq \$0)$

Cabe comentar que quanto maior o tamanho da amostra de dados, podemos notar que a forma da distribuição amostral se torna aproximadamente uma normal (resumidamente é o que prega o Teorema do Limite Central – TLC). O consenso da literatura chega a um valor mínimo da amostra de 30 a 50 dados, para que esse teorema possa valer.

RECAPITULANDO

A distribuição de probabilidade de uma variável aleatória descreve como as probabilidades estão distribuídas sobre os valores dessa variável. A principal vantagem de definir uma variável aleatória e sua distribuição de probabilidade é que, uma vez que a distribuição de probabilidade seja conhecida, torna-se relativamente fácil determinar a probabilidade de uma série de eventos que podem ser de interesse do tomador de decisões.

Para uma variável aleatória discreta X , a distribuição de probabilidade é definida diretamente por uma função de probabilidade, denotada por $f(X)$. Para variável aleatória contínua, a contraparte da função de probabilidade é a função de densidade de probabilidade, também expressa por $f(X)$. Essa função não produz probabilidade

diretamente, como no caso da função de probabilidade das v.a. discretas. A probabilidade nesse caso será obtida pela área sob o gráfico de $f(X)$. Essa área será a probabilidade de a variável aleatória contínua X assumir um valor nesse intervalo. A distribuição de probabilidade discreta uniforme e contínua uniforme são as mais simples; para distribuição de probabilidade discreta, as mais conhecidas e importantes são a distribuição binomial, a distribuição de Poisson e a distribuição hipergeométrica. Para a distribuição de probabilidade contínua, as mais usuais são as distribuições normal e exponencial, sendo a primeira umas das mais importantes da estatística, com ampla variedade de aplicações práticas como: altura e peso das pessoas, notas de exames, medições científicas, índices pluviométricos, preços, etc. A distribuição normal padrão de probabilidade tem a mesma aparência geral das outras distribuições normais, porém com $\mu=0$ e $\sigma=1$.

Referências

- Anderson, D.R.; Sweeney, D.J.; Williams, T. 2007. Estatística Aplicada à Administração e Economia. Pioneira Thompson Learning: São Paulo, SP, Brasil. p. 169-191; 205-218.
- Bussab, W.O.; Morettin, P.A. 2006. Estatística Básica. Saraiva: São Paulo, SP. Brasil. p.140-143; 162-175.
- Gujarati, D. 2006. Econometria básica. Elsevier, Campus, Rio de Janeiro, RJ, Brasil. p. 698-719
- Hoffmann, R. 2006. Estatística para Economistas. p. 81-105. Pioneira Thomson Learning: São Paulo, SP, Brasil.
- Sartoris, A. 2003. Estatística e Introdução à econometria. Saraiva: São Paulo, SP. Brasil. p.57-65; 82-103.

Apêndice

Tabela Z. Distribuição Normal - Padrão

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000



 EDITORA
pecege

ISBN 978-65-86664-57-7

9 786586 664577