INSTITUTO FEDERAL SUL-RIO-GRANDENSE
UNIVERSIDADE ABERTA DO BRASIL
Programa de Fomento ao Uso das
TECNOLOGIAS DE COMUNICAÇÃO E INFORMAÇÃO NOS CURSOS DE GRADUAÇÃO - TICS

ESTATÍSTICA BÁSICA

Álvaro Nebel

Ministério da Educação
GOVERNO FEDERAL BRASIL PAÍS RICO É PAÍS SEM POBREZA

Estatística Básica

NEBEL, Álvaro

2012/1

Produzido pela Equipe de Produção de Material Didático da
Universidade Aberta do Brasil do Instituto Federal Sul-rio-grandense

EQUIPE DE PRODUÇÃO DE MATERIAL DIDÁTICO – UAB/IFSUL

Lisiane Corrêa Gomes Silveira
GESTORA DA EQUIPE DE DESIGN

Denise Zarnottz Knabach
Felipe Rommel
Helena Guimarães de Faria
Lucas Quaresma Lopes
Tabata Afonso da Costa
EQUIPE DE DESIGN

Catiúcia Klug Schneider
GESTORA DE PRODUÇÃO DE VÍDEO

Gladimir Pinto da Silva
PRODUTOR DE ÁUDIO E VÍDEO

Marcus Freitas Neves
EDITOR DE VÍDEO

João Eliézer Ribeiro Schaun
GESTOR DO AMBIENTE VIRTUAL DE APRENDIZAGEM

Giovani Portelinha Maia
GESTOR DE MANUTENÇÃO E SISTEMA DA INFORMAÇÃO

Anderson Hubner da Costa Fonseca
Carlo Camani Schneider
Efrain Becker Bartz
Jeferson de Oliveira Oliveira
Mishell Ferreira Weber
EQUIPE DE PROGRAMAÇÃO PARA WEB

SUMÁRIO

GUIA DIDÁTICO GD

APRESENTAÇÃO

Prezado(a) aluno(a),

Bem-vindo (a) ao espaço de estudo da Disciplina de Estatística Básica.

A estatística, antes de tudo, é um ramo da matemática aplicada. Tem como objetivo estabelecer métodos para coleta, organização, resumo, apresentação e análise dos dados, permitindo a obtenção de conclusões robustas e, finalmente, tomada de decisões. Serve de instrumento de apoio a vários outros campos do conhecimento, em especial a todos os ramos do conhecimento em que dados experimentais são manipulados.

Faz parte do grupo das ciências cujos primeiros passos remontam aos primórdios da história da humanidade e cujo desenvolvimento formal tende a estar em sintonia com a evolução do conhecimento humano. É uma ciência que está sempre absorvendo novas técnicas e contribuições de outras ciências, como novas descobertas e novas teorias.

A Estatística tem sido utilizada na pesquisa científica para a melhoria de recursos econômicos, para o aumento da qualidade e produtividade, nas questões judiciais, previsões e em muitas outras áreas do conhecimento humano. Assim, no dia-a-dia das pessoas, nas mais diferentes atividades, é comum recorrer-se à Estatística.

Nesta disciplina, o material está organizado de maneira a apresentar, em cada capítulo, uma introdução teórica, exemplos de aplicação e exercícios ou atividades para treinamento e fixação dos conteúdos.

Nas unidades, serão abordados os seguintes conteúdos: Introdução ao método estatístico; Estatística Descritiva: organização de dados, elaboração de tabelas e gráficos, distribuição de frequência, medidas de posição e dispersão; Probabilidade: conceitos e funções; Estatística Inferencial: distribuições, teoria da amostragem e da estimação, e conceitos de confiabilidade.

Esperamos que, através dos conteúdos e das atividades propostas, você possa adquirir conhecimentos e habilidades para aplicar a ferramenta estatística nas mais diversas atividades técnicas e científicas.

Lembre-se: há uma equipe que trabalha para que você supere suas dificuldades.

Bom estudo e boa sorte!

Objetivos

Objetivo Geral

Ao final desta disciplina o aluno será capaz de planejar, organizar e analisar dados coletados para estudos estatísticos. Além de ter o domínio dos conceitos básicos de estatística e probabilidade, de modo a poder aplicar estes conhecimentos na prática profissional e a ter embasamento para estudos mais avançados nesta área.

Habilidades

- Planejar e executar levantamento de dados para estudos estatísticos;
- Organizar e criticar os dados levantados, verificando quando os dados ainda são válidos ou não e a forma de tratá-los;

- Resumir os dados e apresentar de forma a facilitar às conclusões e tomadas de decisão;
- Ter conhecimento e domínio das principais estatísticas de posição e de dispersão;
- Identificar a necessidade de utilização de uma estatística robusta;
- Calcular probabilidade de eventos elementares;
- Ter conhecimento das principais funções de probabilidade (uniforme, binomial, normal, exponencial e de Poisson) e suas aplicações;
- Conhecer os diferentes tipos de amostragem e suas aplicações;
- Calcular os tamanhos de amostras em diferentes tipos de problemas e populações;
- Estimar médias e desvios numa amostragem;
- Ter noções de confiabilidade e suas aplicações.

Avaliação

Avaliação dos alunos

O rendimento dos alunos será avaliado através das atividades propostas no curso e do instrumento de avaliação que ocorrerá em encontro presencial.

Avaliação da disciplina

Formativa: ao longo de seu desenvolvimento, o programa e os materiais da disciplina serão analisados pelos alunos e equipe de professores.

Somativa: os alunos avaliarão a validade da disciplina para sua formação através de instrumento específico.

Programação

Primeira semana

As atividades a serem desenvolvidas na 1ª semana são:

1. Introdução ao método estatístico
2. Origem da estatística
3. Universo estatístico
4. Variáveis
5. Fases do método estatístico

Segunda semana

As atividades a serem desenvolvidas na 2ª semana são:

6. Estatística Descritiva
7. Apresentação de dados estatísticos
8. Tabelas
9. Gráficos

Terceira semana

A atividade a ser desenvolvida na 3ª semana é:

10. Distribuição de frequência

Quarta semana

A atividade a ser desenvolvida na 4ª semana é:

11. Medidas de posição

Quinta semana

As atividades a serem desenvolvidas na 5ª semana são:

12. Medidas de dispersão
13. Medidas de assimetria e de curtose

Sexta semana

As atividades a serem desenvolvidas na 6ª semana são:

14. Probabilidade
15. Conceito
16. Espaço amostral
17. Eventos complementares
18. Eventos independentes
19. Eventos mutuamente exclusivos

Sétima semana

As atividades a serem desenvolvidas na 7ª semana são:

20. Distribuição binomial
21. Distribuição normal

Oitava semana

As atividades a serem desenvolvidas na 8ª semana são:

22. Distribuição exponencial
23. Outros tipos de distribuição

Nona semana

As atividades a serem desenvolvidas na 9ª semana são:

24. Inferência Estatística
25. Teoria da amostragem
26. Teoria da estimação

Décima semana

A atividade a ser desenvolvida na 10ª semana é:

27. Estimativas da média e desvio-padrão

Décima primeira semana

As atividades a serem desenvolvidas na 11ª semana são:

28. Confiabilidade
29. Estimativas de confiabilidade

Décima segunda semana

A atividade a ser desenvolvida na 12ª semana é:

30. Estimativas de confiabilidade em sistemas

Currículo do Professor-Autor

Alvaro Luiz Carvalho Nebel

Possui graduação em Engenharia Agrícola pela Universidade Federal de Pelotas (1988), Pós-graduação em Administração de Empresas - FGV (1993), Licenciatura Plena em Formação Pedagógica para a Educação Profissional: Agropecuária, Mestrado em Agronomia pela Universidade Federal de Pelotas (2005) e Doutorado em Agronomia Ciências do Solo, pela Universidade Federal de Pelotas (2009). Atualmente é professor do Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense - IF-SUL, Campus Pelotas Visconde da Graça, no Curso Técnico em Agropecuária, Superior em Vitivinicultura e Enologia e no Curso de Educação à Distância - Biocombustíveis. Ministra as disciplinas de Climatologia Agrícola, Relação Solo-água-planta e Irrigação & Drenagem. Coordenador Pedagógico do Curso Técnico em Agropecuária (2010-2011). Desenvolve atividades de pesquisa e extensão relacionadas ao manejo do solo e da água em propriedades rurais de produção leiteira e irrigação de pastagens, em parceria com a EMATER, EMBRAPA-CPACT e UFPEL. Tem experiência na área de Engenharia Agrícola, com ênfase em Irrigação e Drenagem, atuando principalmente nos seguintes temas: manejo do solo e da água, recursos hídricos, irrigação e drenagem, variabilidade espacial e geoestatística.
<http://buscatextual.cnpq.br/buscatextual/visualizacv.do?id=K4785633Z0>

Referências

ANDERSON, David R. Estatística Aplicada à Administração e Economia, São Paulo: Cencage Learning, 2ª Edição, 2007.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. São Paulo, Ed. Saraiva. 540p. 2010.

CRESPO, A. A. **Estatística Fácil**. São Paulo, Ed. Saraiva. 218p. 2009.

FONSECA, J. S.; MARTINS, G. A. **Curso de Estatística**. São Paulo. 6ª Edição, Atlas, 267p. 1996.

FREITAS, E. A. Curso Técnico em Operações Comerciais. **Estatística Aplicada I**. EQUIPE SEDIS/ Universidade Federal do Rio Grande do Norte, 28p. 2008.

IBGE – Centro de Documentação e Disseminação de Informações. Normas de apresentação tabular / Fundação Instituto Brasileiro de Geografia e Estatística, Centro de Documentação e Disseminação de Informações. – 3.ed. – Rio de Janeiro : IBGE, 62p. 1993.

IBGE- Indicadores de Desenvolvimento Sustentável. Estudos e Pesquisas Informação Geográfica, n.7, Rio de Janeiro: IBGE, 443p. 2010.

IPEA - Instituto de Pesquisa Econômica Aplicada. Biocombustíveis no Brasil: Etanol e biodiesel. Secretaria de Assuntos Estratégicos da Presidência da República. Boletim n. 53, 57p. 2010.

ISO 31 – Grandezas e unidades, Parte 0 – Princípios gerais, Anexo B – **Guia para o arredondamento de números**, 3.ª Ed., 1992. MARTIN, Olivier. Da estatística política à sociologia estatística. Desenvolvimento e transformações da análise estatística da sociedade (séculos XVII-XIX). **Revista Brasileira de História** [online]. Vol.21, n.41, p. 13-34, 2001.

MILONE, G. **Estatística geral e aplicada**. São Paulo: Thomson, 483p. 2004.

MMA – Instituto do Meio Ambiente e dos Recursos Naturais Renováveis. Caderno setorial dos recursos hídricos: agropecuária. Ministério do Meio Ambiente, Secretaria dos Recursos Hídricos. Brasília: MMA, 96p. 2006.

SPIEGEL, Murray R. Estatística, São Paulo: Saraiva, 18ª edição, 2005.

STEVENSON, William J., Estatística Aplicada a Administração, São Paulo: Harbra, 2001

VIEIRA, S. **Estatística Básica**. Editora Cengage Learning. São Paulo. 176p. 2011.

A

Origem e histórico da Estatística, definições e introdução ao método estatístico

Unidade A
Estatística Básica

UNIDADE A

ORIGEM E HISTÓRICO DA ESTATÍSTICA, DEFINIÇÕES E INTRODUÇÃO AO MÉTODO ESTATÍSTICO

Objetivos

- Conhecer a origem e o histórico da Estatística.
- Compreender as características de cada uma das partes da Estatística.
- Descrever o método estatístico, identificando cada uma de suas etapas.

Palavras iniciais...

Olá! Este material é uma introdução a um dos importantes e vastos temas da matemática aplicada: a ESTATÍSTICA.

O material está organizado de maneira a apresentar, em cada capítulo, uma introdução teórica, exemplos de aplicação e exercícios ou atividades para treinamento e fixação dos conteúdos.

Estabeleça horários de estudo e faça um bom uso deste material.

Bom estudo e boa sorte!

Você verá por aqui...

O que é Estatística? Como evoluiu ao longo do tempo? As respostas a essas perguntas e muitas outras estão ao longo desta nossa primeira aula.

Aqui, você verá também alguns conceitos e definições iniciais necessários ao desenvolvimento do assunto, além de conhecer quais são as etapas do método estatístico.

Origem e histórico da estatística

Vamos começar nossos estudos de Estatística conhecendo um pouco de sua origem e histórico.

As primeiras tentativas de enumeração de indivíduos ou de bens começam com os grandes impérios da Antiguidade, cujas estruturas administrativas eram fortes: preocupados em gerir e administrar seu império do melhor modo, os poderes centrais procuraram conhecer melhor sua extensão territorial e o número de seus súditos. Foi assim que as civilizações egípcia, mesopotâmica e chinesa, como antes delas a civilização dos sumérios (5000 a 2000 a. C.), realizavam pesquisas censitárias das quais alguns traços chegaram até nós (MARTIN, 2001).

Vários povos já realizavam contagem populacional, utilizada para se obter informações sobre o número de habitantes, de nascimentos, de óbitos, faziam estimativas da riqueza individual e social, cobravam impostos e coletavam informações por processos que podem, atualmente, ser denominados de estatísticas.

No livro sagrado Chouking, de Confúcio, há citações de recenseamentos realizados na China nos anos de 2.275 a.C. e 2.238 a.C. que foram utilizados pelo seu imperador (Rei Yao) para investigar a quantidade de seus súditos e para descrever em números as condições econômicas (agricultura, indústria e comércio) e de poderio militar do seu império. Posteriormente, foram encontrados dados similares em livros da Dinastia Chinesa referentes a 1.100 a.C.

No ano 120, Menelaus apresenta tabelas estatísticas cruzadas, e em 620, em Constantinopla, surge um Primeiro Bureau de Estatística. Em 695 os árabes utilizam a média ponderada para a contagem de moedas e, em 826, utilizam cálculos estatísticos para a tomada de Creta.

Em 1.405, o persa Ghiyat Kâshî realiza os primeiros cálculos de probabilidade com a fórmula do binômio.

Com ambições mercantilistas, entre os séculos XVI e XVIII, vários governantes viram a necessidade de coletar informações de outras nações, como produção de bens, produção de alimentos e dados de comércio exterior, como forma de se obter o poder político através do poder econômico. E, também, passaram a realizar os primeiros levantamentos estatísticos com o objetivo de determinar leis sobre impostos e número de homens disponíveis para entrar em combate (FREITAS, 2008 – p.02).

A partir do século XV, começam a surgir as primeiras análises de fatos sociais, como batizados, casamentos, funerais, originando as primeiras tábuas e tabelas e os primeiros números relativos. Datam de 1.447 as primeiras tabelas de mortalidade elaboradas pelos sábios do Islã. A partir do século seguinte, esses estudos foram adquirindo proporções verdadeiramente científicas. Em 1.614, Napier cria os logaritmos; em 1.620 Descartes estabelece a Geometria Descritiva; em 1.629, Pierre de Fermat estabelece o Método de Máximo e Mínimo e a Teoria dos Números e, mais tarde, em 1.654, Fermat e Pascal estabelecem os Princípios do Cálculo das Probabilidades.

Em 1.800, é fundado na França o Bureau de Estatística. Em 1.805, Legendre estabelece o Método dos Mínimos Quadrados; em 1.812 Laplace publica sua Théorie Analytique des Probabilités, que foi a base da inferência estatística e a partir de 1.820 várias sociedades de estatística são criadas pelo mundo, destacando-se a American Statistical Association, em 1.839 e, em 1.853, é realizada a Primeira Conferência Internacional de Estatística em Bruxelas (Quetelet). Em 1.925 é lançado o clássico livro "Statistical Methods for Research Workers".

A Estatística é uma ciência que sempre está predisposta a incorporar técnicas, descobertas e teorias novas, próprias ou vindas de outras áreas do conhecimento humano. Prosseguindo com seu contínuo desenvolvimento e reestruturação, essa ciência hoje conta com uma poderosa parceira no trabalho com os dados pesquisados, que é a informática.

A evolução no processo de coleta, armazenamento e divulgação de informações estatísticas por meios informatizados tem sido seguida pelo surgimento de novas metodologias e técnicas de análise de dados estatísticos.

Assim, fizemos aqui um breve e conciso histórico de alguns fatos que demonstram a evolução da estatística ao longo do tempo. Obviamente, se fôssemos aplicar maior rigor nessa revisão, ocuparíamos algumas páginas para descrever todos os fatos que foram marcantes para o desenvolvimento da estatística. No

site da Associação Brasileira de Estatística é possível acessar uma extensa e abrangente cronologia elaborada por Gauss M. Cordeiro, no link *<http://www.redeabe.org.br/cronologia022006.doc>*. Também na página do Instituto de Matemática da UFRGS podemos acessar mais informações sobre a história da estatística no *<link http://paginas.ufrgs.br/mat/graduacao/estatistica/historia-da-estatistica>*.

Vale a pena dar uma olhada nestes materiais.

Introdução ao método estatístico

Definições:

a. O que é Estatística?

A estatística, antes de tudo, é um ramo da matemática aplicada. Tem como objetivo estabelecer métodos para coleta, organização, resumo, apresentação e análise dos dados, permitindo a obtenção de conclusões robustas e, finalmente, tomada de decisões. Serve de instrumento de apoio a vários outros campos do conhecimento, em especial a todos os ramos do conhecimento em que dados experimentais são manipulados.

Faz parte do grupo das ciências cujos primeiros passos remontam aos primórdios da história da humanidade e cujo desenvolvimento formal tende a estar em sintonia com a evolução do conhecimento humano, conforme afirma Milone (2004, p. 337). É, segundo o mesmo autor, uma ciência que está sempre absorvendo novas técnicas e contribuições de outras ciências, como novas descobertas e novas teorias.

A Estatística tem sido utilizada na pesquisa científica para a melhoria de recursos econômicos, para o aumento da qualidade e produtividade, nas questões judiciais, previsões e em muitas outras áreas do conhecimento humano (FREITAS, 2008 - p.03). Assim, no dia a dia das pessoas, nas mais diferentes atividades, é comum recorrer-se à Estatística.

Uma definição de Estatística bastante abrangente é:

> *A Estatística é uma ciência que reúne um conjunto de métodos adequados para a coleta, organização, descrição, análise e interpretação de dados, proporcionando extrair informações e estimativas a respeito dos mesmos e tomada de decisões razoáveis baseadas em tais análises.*

b. Objetivo da Estatística

A estatística fornece-nos as técnicas para extrair informação de dados, os quais são muitas vezes incompletos, na medida em que nos dão informação útil sobre o problema em estudo. Sendo assim, é objetivo da Estatística extrair informação dos dados para obter uma melhor compreensão das situações que representam.

Quando se aborda uma problemática envolvendo métodos estatísticos, estes devem ser utilizados mesmo antes de se recolher a amostra, isto é, deve-se planejar o procedimento que nos vai permitir recolher os dados, de modo que, posteriormente, se possa extrair o máximo de informação relevante para o problema em estudo, ou seja, para a população de onde os dados provêm. Quando de posse dos dados, procura-se agrupá-los e reduzi-los, sob forma de amostra, deixando de lado a aleatoriedade presente.

Seguidamente o objetivo do estudo estatístico pode ser o de estimar uma quantidade ou testar uma hipótese, utilizando-se técnicas estatísticas convenientes, as quais realçam toda a potencialidade da

Estatística, na medida em que vão permitir tirar conclusões acerca de uma população, baseando-se numa pequena amostra, dando-nos ainda uma medida do erro cometido.

Exemplo 1:

> *Para avaliar a conformidade de dimensões de uma determinada peça da qual se esteja produzindo uma grande quantidade não é necessário que se faça mensuração em cada peça produzida. Basta que sejam avaliadas algumas peças escolhidas ao acaso, para concluirmos se as dimensões das peças produzidas estão dentro dos padrões.*

c. Divisão da Estatística

Geral ou metodológica:

Descritiva

Indutiva ou Inferencial

Aplicada:

Mecânica Estatística, Demográfica, Econometria, entre outras.

Estatística Descritiva:

A estatística descritiva é a parte da estatística em que, a partir de um determinado conjunto de dados, organiza-os em tabelas, gráficos ou estabelece sumário através de medidas descritivas como a média, valor mínimo e máximo, desvio padrão, entre outras. É a parte da Estatística que tem por objetivo descrever os dados observados. Pode-se tomar como exemplo o conjunto de notas dos alunos de uma dada disciplina em um semestre letivo. A esse conjunto de notas denominamos de conjunto de dados, sendo a nota individual de cada aluno chamada de observação. A coleta, a organização e a descrição dos dados fazem parte da Estatística Descritiva.

Estatística Indutiva ou Inferencial:

A Estatística Indutiva ou Inferencial realiza a análise e a interpretação dos dados. Nesse caso, o conjunto de todos os dados de interesse é chamado de população. Uma parte retirada dessa população é chamada de amostra. Dessa forma, a partir de análise dos dados da amostra podemos estabelecer inferências e previsões sobre a população e tomar decisões. É a parte da Estatística que tem por objetivo obter e generalizar conclusões para a população a partir de uma amostra, através do cálculo de probabilidade.

Estatística Aplicada:

O desenvolvimento e o aperfeiçoamento de técnicas estatísticas, de obtenção e de análise de informações, permitem o controle e o estudo adequado de fenômenos, fatos, eventos e ocorrências, em diversas áreas do conhecimento. A Estatística Aplicada tem por objetivo fornecer métodos e técnicas aplicados a determinados campos do conhecimento em específico como, por exemplo, para as ciências médicas.

d. Universo Estatístico

A Estatística tem por objetivo o estudo dos fenômenos coletivos e das relações que existem entre eles. O fenômeno coletivo é aquele que se refere à população, ou universo, que compreende um grande número de elementos. Para a estatística interessam os fatos que englobam um conjunto de elementos, não

importando cada um dos elementos em particular.

O universo estatístico é o conjunto de dados, indivíduos, objetos, etc. e tais elementos são reunidos em subconjuntos denominados população. Assim, População é o conjunto constituído de elementos, de um mesmo universo, que apresentam pelo menos uma característica comum. A população, segundo o seu tamanho, pode ser finita ou infinita. É finita quando possui um número determinado de elementos e infinita quando possui um número infinito de elementos. Contudo, tal definição existe apenas no campo teórico, uma vez que na prática, nunca encontraremos populações com infinitos elementos e sim com grande número de componentes e tais populações são tratadas como infinitas.

e. Amostra

Na maioria das vezes, devido ao alto custo, ao intenso trabalho ou ao tempo necessário, limitamos as observações referentes a uma determinada investigação ou pesquisa a apenas uma parte da população, a qual denominamos de amostra. Amostra é um subconjunto finito da população.

f. Dados ou variáveis

Variável é uma característica ou condição dos elementos da população que estamos que estamos analisando. A variável pode assumir diferentes valores para os diferentes elementos da população, por isso é variável.

Dados são os valores coletados da variável que estamos avaliando.

g. Tipos de variáveis

Tipos de variáveis			
Qualitativa		Quantitativa	
Ordinal	Nominal	Contínua	Discreta

Variáveis qualitativas:

São variáveis que representam atributos ou qualidades. Dividem-se em nominais e ordinárias.

- **Qualitativas nominais:** cujos valores não têm uma relação de ordem entre si, ou seja, não podem ser hierarquizadas ou ordenadas. Ex: sexo, raça, grupo sanguíneo, etc..
- **Qualitativas ordinais**: cujos valores não são métricos, mas incluem relações de ordem. Ex: classe social (A, B, C, D), grau de instrução, níveis de peso (muito pesado, pesado, pouco pesado).

Variáveis quantitativas:

São variáveis que representam valores medidos ou contados. Podem ser classificadas ainda em contínuas ou discretas.

- **Quantitativa Continua** : são variáveis que podem assumir qualquer valor, inclusive fracionários, e resultam normalmente de uma mensuração. Ex: altura, peso, temperatura, idade, etc..
- **Quantitativa Discreta**: assumem valores inteiros, inclusive o zero, e resultam frequentemente de uma contagem. Ex: número de filhos, número de alunos, etc..

h. Arredondamento de números

Para o processo de arredondamento de números, as regras usuais (conforme ISO 31-0:1992, anexo B) devem ser utilizadas da seguinte forma:

1. Se o primeiro algarismo após o que queremos arredondar for de 0 a 4, conservamos o algarismo a ser arredondado e desprezamos os seguintes.
   Ex.: 5,74766 (para décimos) → 5,7

2. Se o primeiro algarismo após o que queremos arredondar for de 6 a 9, acrescenta-se uma unidade no algarismo a ser arredondado e desprezamos os seguintes.
   Ex.: 5,7766 (para décimos) → 5,8

3. Se o primeiro algarismo após o que queremos arredondar for 5, seguido apenas de zeros, conservamos o algarismo se ele for par ou aumentamos uma unidade se ele for ímpar, desprezando os seguintes.
   Ex.: 5,4500 (para décimos) → 5,4
   5,3500 (para décimos) → 5,4

   Se o 5 for seguido de outros algarismos dos quais, pelo menos um é diferente de zero, aumentamos uma unidade no algarismo e desprezamos os seguintes.
   Ex.: 5,2502 (para décimos) → 5,3
   5,3503 (para décimos) → 5,4

Fases do método estatístico

a. Definição do problema:

O que se pretende investigar, pesquisar, avaliar? É quando se define qual é o problema que quer se resolver, delimitando-se o objeto de estudo de forma viável em relação ao tempo e aos recursos disponíveis.

b. Planejamento:

Após a definição do problema a ser estudado, o planejamento consiste em determinar os procedimentos para a investigação e solução do problema. É a etapa onde são estabelecidos os detalhes mais importantes do estudo: o cronograma geral, a metodologia da coleta de dados, a definição do tamanho da amostra, entre outros.

c. Coleta ou levantamento dos dados:

É a obtenção, reunião e registro sistemático de dados.

d. Crítica dos dados:

Etapa onde são observadas as discrepâncias nos dados obtidos e, se necessário, decidir se haverá uma nova coleta de dados, descarte do dado discrepante ou complementação das informações sobre esse dado.

e. Apuração dos dados ou sumarização:

É o processo de sumarização dos dados obtidos mediante critério de classificação.

f. Apresentação dos dados:

Após a apuração, os dados são organizados conforme o objetivo e apresentados sob a forma de tabelas,

gráficos e medidas.

g. Análise e interpretação dos dados:

É a etapa onde são feitas medidas complementares a partir dos dados coletados que darão suporte a deduções e/ou induções sobre as informações obtidas, como também as conclusões sobre o estudo em questão. É o resultado do trabalho estatístico, com solução para o problema definido na primeira fase. Exige conhecimento técnico do processo, fenômeno ou evento que se está analisando.

Resumo

Nesta aula estudamos a origem e o histórico do desenvolvimento da Estatística, sua definição e divisões, conceituamos população estatística e amostra, definimos os tipos de variáveis existentes e as fases do método estatístico.

Exercícios

1. Conceitue população e amostra apresentando um exemplo.

2. Diferencie variáveis quantitativas de variáveis qualitativas.

3. Classifique as variáveis abaixo em qualitativas (nominal ou ordinal) ou quantitativas (contínuas ou discretas).

a. Grupo sanguíneo.
b. Grau de instrução.
c. Comprimento de peças produzidas em determinada máquina.
d. Número de alunos matriculados na disciplina de Estatística.
e. Temperaturas médias diárias registradas em Pelotas no ano de 2010.
f. Número de dias com temperatura média inferior a 10°C em Pelotas no ano de 2010.
g. Altura dos alunos matriculados na disciplina de Estatística.
h. Classe social dos alunos matriculados em Estatística.
i. Salário dos funcionários de uma fábrica de máquinas agrícolas.
j. Sexo dos funcionários da fábrica de máquinas agrícolas.

Referências

FONSECA, J. S.; MARTINS, G. A. **Curso de Estatística**. São Paulo. 6ª Edição, Atlas, 267p. 1996.

FREITAS, E. A. Curso Técnico em Operações Comerciais. **Estatística Aplicada I**. EQUIPE SEDIS/ Universidade Federal do Rio Grande do Norte, 28p. 2008.

ISO 31 – Grandezas e unidades, Parte 0 – Princípios gerais, Anexo B – **Guia para o arredondamento de números**, 3.ª Ed., 1992.

MARTIN, Olivier. Da estatística política à sociologia estatística. Desenvolvimento e transformações da análise estatística da sociedade (séculos XVII-XIX). **Revista Brasileira de História** [online]. Vol.21, n.41, p. 13-34, 2001.

MILONE, G. **Estatística geral e aplicada**. São Paulo: Thomson, 483p. 2004.

TICs

B

Estatística Descritiva - Apresentação de dados estatísticos, tabelas e gráficos

Unidade B
Estatística Básica

UNIDADE B

ESTATÍSTICA DESCRITIVA - APRESENTAÇÃO DE DADOS ESTATÍSTICOS, TABELAS E GRÁFICOS

Objetivos

- Identificar os diferentes tipos de séries estatísticas e sua representação.
- Conceituar tabelas e identificar seus elementos.
- Conceituar e identificar os diversos tipos de gráficos.

Você verá por aqui...

Nesta aula vamos trabalhar assuntos relacionados à organização de dados estatísticos e sua representação gráfica.

Ao final desta aula apresentamos uma relação de exercícios e atividades que têm o objetivo de fixar os conteúdos estudados.

Procure organizar seus horários, disponibilizando um bom tempo para as atividades e aproveite bem este material.

Começando...

Após a etapa de coleta dos dados, normalmente temos um conjunto extenso de valores e informações que precisam ser ordenados e organizados de tal forma que possamos ter uma visão global do fenômeno analisado. Representar o conjunto de valores por meio de tabelas ou gráficos adequados irá permitir uma boa caracterização das informações que temos, com as quais poderemos realizar diagnósticos e conclusões ou, ainda, fazer comparações com outros conjuntos semelhantes de dados.

Apresentação de dados estatísticos e suas representações

Uma **série estatística** consiste em um conjunto de dados ordenado segundo uma característica comum, ou seja, os dados referem-se a uma mesma variável. Uma série estatística, comumente, é representada através de uma tabela ou de um gráfico, conforme melhor ficar representado o conjunto de dados que queremos analisar. As características de cada uma dessas representações, tabelas ou gráficos, é o que veremos a seguir.

a. Tabelas

Uma Tabela deve apresentar um conjunto de dados de uma forma eficiente e organizada, facilitando sua compreensão e interpretação.

A tabela é composta por:

a) **Título** – localizado no topo da tabela, formado por um conjunto de informações, as mais completas possíveis, identificando o tipo de variável que está sendo analisado, o local e o período em que os dados foram coletados. O título deve responder a três questões: O quê? Onde? Quando?

b) **Cabeçalho** – parte superior da tabela que especifica o conteúdo das colunas

c) **Corpo** – é o conjunto de linhas e colunas que contém as informações sobre a variável em estudo. O espaço destinado a um único dado ou número é denominado casa ou célula.

d) **Coluna indicadora** – parte da tabela que especifica o conteúdo das linhas.

e) **Linhas** – são retas horizontais imaginárias que facilitam a leitura de dados.

f) **Elementos complementares** – são os elementos colocados no rodapé da tabela, tais como fonte, notas ou chamadas. A fonte identifica quem coletou originalmente os dados.

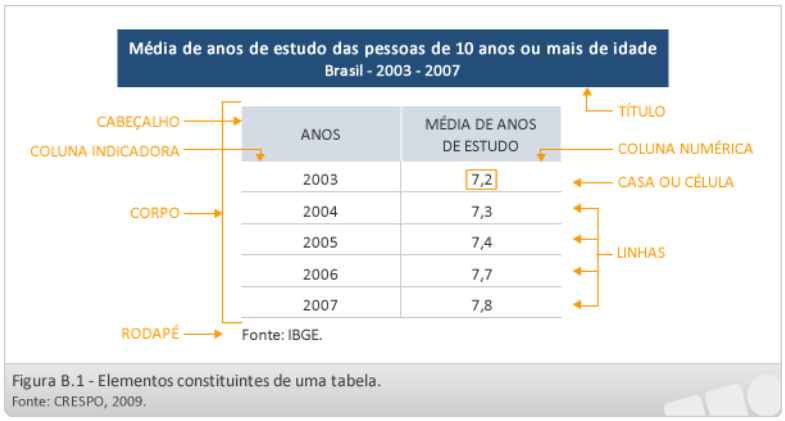


Figura B.1 - Elementos constituintes de uma tabela.
Fonte: CRESPO, 2009.

Conforme as normas para elaboração de tabelas (IBGE, 1993) nas células ou casas devem ser colocados:

- Um traço horizontal quando o valor é zero.
- Três pontos (...) quando não temos os dados.
- Um ponto de interrogação (?) quando temos dúvida quanto à exatidão de determinado valor.
- Zero, quando o valor é muito pequeno para ser expresso pela unidade utilizada. Se os valores são expressos em numerais decimais, precisamos acrescentar à parte decimal um número correspondente de zeros (0,0; 0,00; 0,000; ...).

Séries temporais ou cronológicas

As séries temporais ou cronológicas resultam tabelas que apresentam os valores da variável em estudo, **em determinado local**, discriminados segundo **intervalos de tempo** variáveis.

Exemplo 2:

Taxa de escolarização das pessoas de 18 e 19 anos de idade – Brasil – 2001/2008	
Ano	**Taxa de escolarização**
2001	51,4
2002	51,1
2003	51,7
2004	48,5
2005	47,6
2006	47,0
2007	45,0
2008	46,0

Fonte: IBGE, 2010.

Séries geográficas ou espaciais

As tabelas com séries geográficas ou espaciais apresentam os valores da variável em estudo, **em determinado tempo**, discriminados segundo **regiões**.

Exemplo 3:

Taxa de escolarização das pessoas de 18 e 19 anos de idade, por regiões – Brasil –2008	
Regiões	**Taxa de escolarização**
Norte	51,1
Nordeste	50,6
Centro-Oeste	47,1
Sudeste	42,9
Sul	41,4

Fonte: IBGE, 2010.

Séries específicas ou categóricas

As tabelas com séries específicas ou categóricas apresentam os valores da variável em estudo, **em determinado tempo e local**, discriminados segundo **categorias ou especificações.**

Exemplo 4:

Produtividade de algumas culturas na safra 2003/04 – Brasil	
Culturas	**Produtividade (kg/hectare)**
Algodão	3.098
Amendoim	2.213
Arroz	3.540
Aveia	1.374
Centeio	1.346
Feijão	700
Milho	3.291
Soja	2.339

Fonte: MMA, 2006.

Séries mistas ou de dupla entrada

São tabelas que apresentam a variação de valores de mais de uma variável, conjugando duas ou mais séries. Quando duas séries são conjugadas em uma mesma tabela temos uma tabela de dupla entrada, onde aparecem duas ordens de classificação: uma na linha (horizontal) e outra na coluna (vertical). Por apresentar mais de uma característica dos dados ao mesmo tempo, este tipo de tabela exige sempre mais de duas colunas. No exemplo a seguir, mostramos uma série categórica conjugada com uma série cronológica, que dá origem a uma série categórica-cronológica.

Exemplo 5:

Produção de Carne no Brasil, 1970-2000				
Origem	Produção de carne (milhões de toneladas)			
	1970	1980	1990	2000
Bovinos	1.845	2.850	4.115	6.579
Suínos	767	980	1.050	2.600
Frangos	366	1.370	2.356	5.980

Fonte: MMA, 2006.

Podem ser construídas tabelas com três ou mais entradas, embora não sejam tão comuns devido à dificuldade de representação.

b. Gráficos

Os dados estatísticos também podem ser representados por gráficos, cujo objetivo é o de produzir uma visualização mais rápida dos resultados ou do fenômeno que se investiga.

Algumas regras para a construção de gráficos, segundo Vieira (2011):

- O gráfico deve apresentar título e escala.
- O título deve ser colocado abaixo da ilustração.
- As escalas devem crescer da esquerda para a direita e de baixo para cima.
- As legendas explicativas devem ser colocadas preferencialmente a direita da figura.
- Os gráficos devem ser numerados, na ordem em que são citados no texto.

Requisitos fundamentais para a utilidade de um gráfico, segundo Crespo (2009):

- Simplicidade: o gráfico deve ser destituído de detalhes de importância secundária ou de traços desnecessários que possam levar o observador a uma análise morosa ou com erros.
- Clareza: o gráfico deve permitir a correta interpretação dos valores que estão representados.
- Veracidade: o gráfico deve expressar a verdade sobre o fenômeno em estudo.

Os principais tipos de gráficos são os **diagramas**, os **cartogramas** e os **pictogramas**.

Diagramas:

São gráficos geométricos de, no máximo, duas dimensões e que em sua construção geralmente utilizamos o sistema cartesiano. Entre os principais diagramas estão o gráfico em linha, o gráfico em colunas ou barras e o gráfico em setores. Vejamos cada um desses diagramas.

1. Gráfico em linha

Nesse tipo de gráfico utilizamos a linha poligonal para representar a série estatística. É o tipo de gráfico indicado para representar uma série temporal, principalmente quando o objetivo é mostrar a presença de flutuações nos dados em função da época em que foram medidos.

Exemplo 6: Para a construção do gráfico em linha utilizamos os dados da tabela apresentada no exemplo 2.

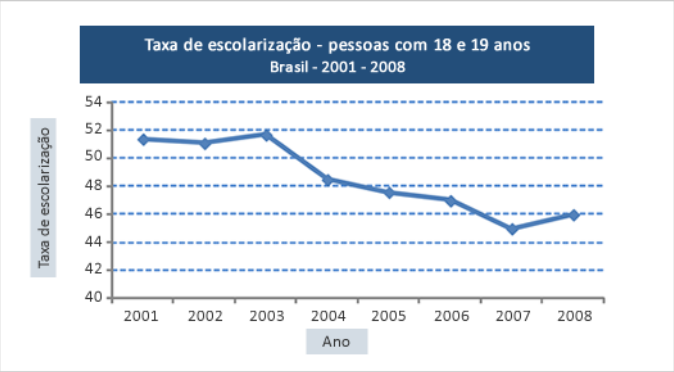


Figura B.2 - Taxa de escolarização/pessoas com 18 e 19 anos: Brasil - 2001-2008.
Fonte: IBGE, 2010.

Exemplo 7: Para este gráfico de múltiplas linhas utilizamos os dados da tabela apresentada no exemplo 5.

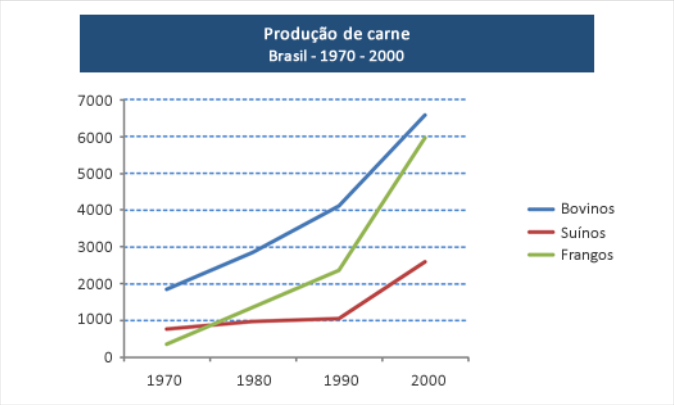


Figura B.3 - Produção de carne no Brasil - 1970 -2000.
Fonte: IBGE, 2010.

Podemos visualizar nos dois gráficos que as informações aparecem mais claras do que as apresentadas nas tabelas respectivas. Por exemplo, se nosso objetivo fosse o de demonstrar que o crescimento da produção de carne de frango foi o mais acentuado no período, em relação à produção de carne bovina ou de suínos, a Figura 2 seria mais conveniente do que a tabela apresentada.

2. Gráfico em colunas ou barras

Nesse tipo de gráfico representamos a série de dados por meio de retângulos dispostos verticalmente (em colunas) ou horizontalmente (em barras).

Se a representação for por meio de colunas, os retângulos têm a mesma base e suas alturas são proporcionais aos dados que representam, enquanto se optarmos por barras, os retângulos têm a mesma altura e os comprimentos são proporcionais aos respectivos dados.

Exemplo 8: Para este gráfico em colunas utilizamos os dados da tabela apresentada no exemplo 3.

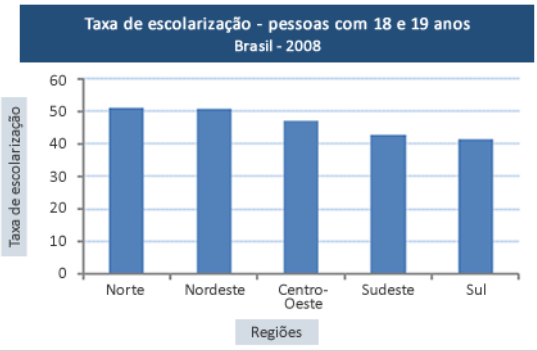


Figura B.4 - Taxa de escolarização por região/pessoas com 18 e 19 anos: Brasil - 2008.
Fonte: IBGE, 2010.

Neste tipo de gráfico ainda podemos apresentar o valor da variável ou atributo que estamos representando, conforme mostrado a seguir:

Exemplo 9: Gráfico em colunas apresentando o valor da variável ou atributo.

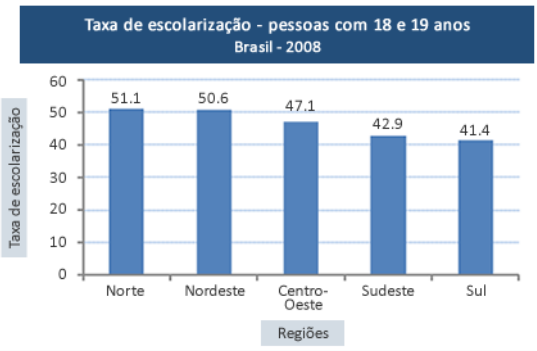


Figura B.5 - Taxa de escolarização por região/pessoas com 18 e 19 anos: Brasil - 2008.
Fonte: IBGE, 2010.

Da mesma forma, poderíamos ter apresentado os dados na forma de barras. Utilizamos este tipo de gráfico principalmente quando os rótulos dos dados (nomes das variáveis em estudo) têm nomes extensos, o que gera dificuldades para apresentar os dados na forma de gráficos em colunas.

Exemplo 10: Dados da tabela apresentada no exemplo 3, ilustrados em forma de gráfico de barras.

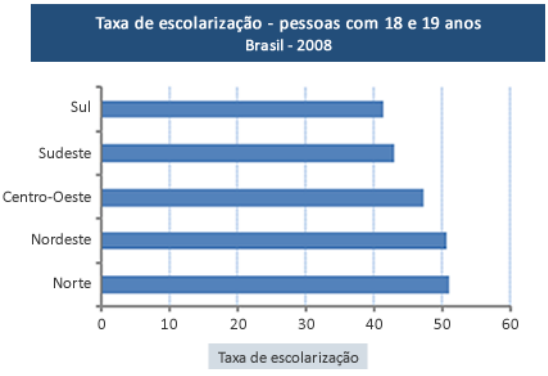


Figura B.6 - Taxa de escolarização por região/pessoas com 18 e 19 anos: Brasil - 2008.
Fonte: IBGE, 2010.

3. Gráfico em setores

Este tipo de gráfico é empregado quando desejamos ressaltar a participação de um determinado dado no total. É construído com base em um círculo que fica dividido em tantos setores quanto são os dados que iremos representar. Recomendamos a utilização desse tipo de gráfico somente quando todas as partes representadas efetivamente ocupam um setor mínimo, evitando com isso setores mal concebidos por apresentarem uma área quase imperceptível.

Exemplo 11: Para a elaboração do gráfico em setores da figura 6, utilizamos os dados da tabela apresentada no exemplo 5, somente para o ano de 2000.

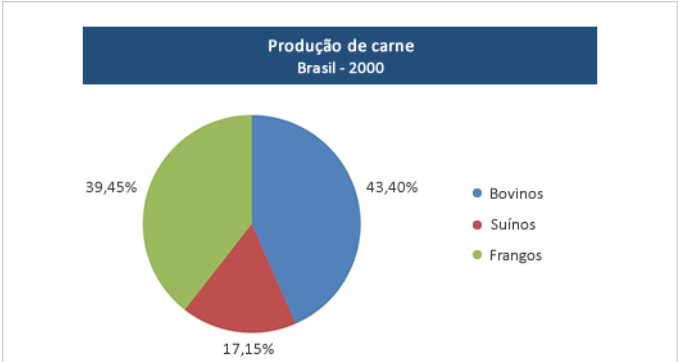


Figura B.7 - Produção de carne no Brasil no ano de 2000.
Fonte: MMA, 2006.

4. Diagramas em 3d

Os gráficos em colunas ou barras e os gráficos em setores podem ser apresentados em três dimensões, entretanto, temos que tomar algum cuidado para que o gráfico mantenha a clareza.

Exemplo 12: Gráfico em colunas apresentado em 3D:

Figura B.8 - Taxa de escolarização/pessoas com 18 e 19 anos: Brasil - 2008.
Fonte: IBGE, 2010.

Exemplo 13: Gráfico em setores apresentado em 3D:

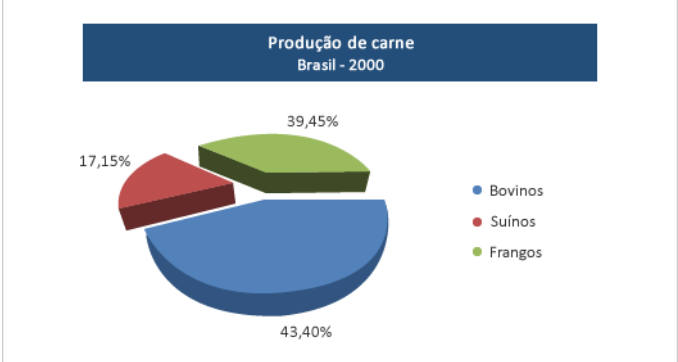


Figura B.9 - Produção de carne no Brasil no ano de 2000.
Fonte: MMA, 2006.

5. Gráfico polar:

O gráfico polar faz uso do sistema de coordenadas polares, sendo indicado para representar séries temporais que apresentam em seu desenvolvimento determinada periodicidade como, por exemplo, a variação da temperatura ao longo do dia, o consumo de energia elétrica durante o ano ou o mês, a variação da precipitação pluviométrica em um ano, entre outros.

Exemplo 14: Gráfico polar.

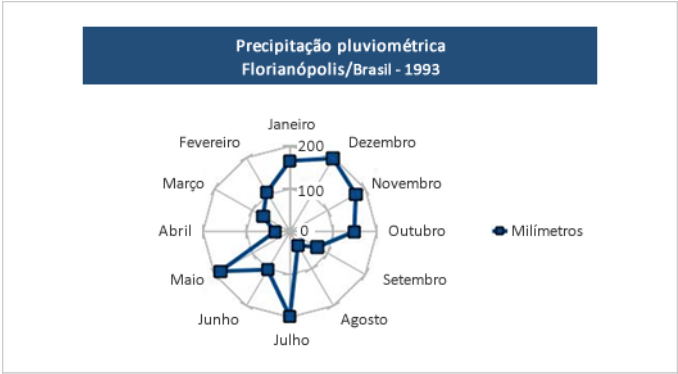


Figura B.10 - Precipitação pluviométrica em Florianópolis no ano de 1993.
Fonte: http://gilbertexbom.com/sig/exe2_estatistica.html

Cartogramas:

Neste tipo de gráfico o objetivo é o de representar os dados estatísticos diretamente relacionados com determinada área geográfica ou política.

Exemplo 15: Cartograma

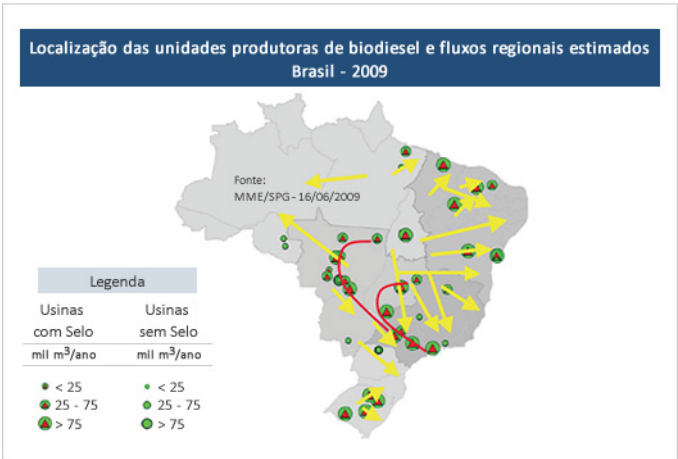


Figura B.11 - Localização das unidades produtoras de biodiesel e fluxos regionais estimados.
Fonte: IPEA, 2010.

Pictogramas:

São representações gráficas com apelo visual para atrair a atenção do leitor. Bastante utilizadas em jornais e revistas, apresentam dados estatísticos em forma de figuras.

Exemplo 16: Pictograma

Figura B.12 - Crescimento da inflação - Brasil Ð Agosto a Dezembro/2010.
Fonte: Revista Veja - 05/01/2011, p. 77.

Outras representações gráficas

Existem outras formas de representar séries de dados estatísticos, algumas são variações das que foram apresentadas e outras são apropriadas para representar medidas estatísticas, tais como diagramas de caixa (*Box plot*), diagramas de ramo e folhas, gráfico de pontos, entre outros. Nesta aula, consideramos os tipos mais comuns de representação gráfica de séries estatísticas sem, entretanto, abordar casos específicos utilizados em estatística aplicada.

Resumo

Nesta aula estudamos as formas de representar graficamente as séries estatísticas, através dos diversos tipos de tabelas e gráficos, bem como as normas e características construtivas de cada um foram abordadas e exemplificadas.

Exercícios

1. Com relação à construção de tabelas é correto afirmar que

a. o título deve possuir informações completas e estar localizado abaixo da tabela.
b. o título deve responder três questões: O quê? Como? Por quê?
c. o titulo deve conter o local e o período em que os dados foram coletados.
d. o título deve especificar de forma clara o conteúdo das colunas.
e. o título deve especificar de forma clara o conteúdo das colunas e das linhas.

2. Uma série estatística é denominada cronológica quando

a. o elemento variável é a espécie.
b. o elemento variável é o tempo.
c. o elemento variável é o local.
d. é o resultado da combinação de séries estatísticas de tipos diferentes.
e. os dados estão agrupados em subintervalos do intervalo observado.

3. Uma série estatística é denominada geográfica quando

a. o elemento variável é a espécie.
b. o elemento variável é o tempo.
c. o elemento variável é o local.
d. é o resultado da combinação de séries estatísticas de tipos diferentes.

e. os dados estão agrupados em subintervalos do intervalo observado.

4. Uma série estatística é denominada categórica quando

a. o elemento variável é a espécie.
b. o elemento variável é o tempo.
c. o elemento variável é o local.
d. é o resultado da combinação de séries estatísticas de tipos diferentes.
e. os dados estão agrupados em subintervalos do intervalo observado.

5. De acordo com as normas para representação tabular de dados, quando o valor de um dado é zero, deve-se colocar na célula correspondente:

a. zero (0)
b. três pontos (...)
c. um traço horizontal (-)
d. um ponto de interrogação (?)
e. um ponto de exclamação (!)

6. De acordo com as normas para representação tabular de dados, quando o valor de um dado é muito pequeno para ser expresso com o número de casas decimais utilizadas ou com a unidade de medida utilizada, deve-se colocar na célula correspondente:

a. zero (0)
b. três pontos (...)
c. um traço horizontal (-)
d. um ponto de interrogação (?)
e. um ponto de exclamação (!)

7. De acordo com as normas para representação tabular de dados, quando não possuímos o valor de um dado, devemos colocar na célula correspondente:

a. zero (0)
b. três pontos (...)
c. um traço horizontal (-)
d. um ponto de interrogação (?)
e. um ponto de exclamação (!)

8. Quando se deseja evidenciar a participação de um dado em relação ao total, o gráfico mais comumente utilizado é denominado

a. pictograma.
b. gráfico em colunas.
c. cartograma.
d. gráfico em setores.
e. gráfico em barras.

9. A representação gráfica encontrada em jornais e revistas que inclui figuras de modo a torná-las mais atraente é denominada

a. gráfico decorado.
b. gráfico em figuras.
c. cartograma.
d. gráfico em setores.
e. pictograma.

10. Considerando as regras para a construção de gráficos, assinale a afirmativa correta.

a. As escalas devem crescer da esquerda para a direita e de cima para baixo.
b. As legendas explicativas devem ser colocadas preferencialmente abaixo da figura.

c. Os gráficos devem ser numerados, na ordem em que são citados no texto.
d. O gráfico deve apresentar título e nota de rodapé.
e. O título deve ser colocado acima da ilustração.

Referências

CRESPO, A. A. **Estatística Fácil**. São Paulo, Ed. Saraiva. 218p. 2009.

IBGE – Centro de Documentação e Disseminação de Informações. Normas de apresentação tabular / Fundação Instituto Brasileiro de Geografia e Estatística, Centro de Documentação e Disseminação de Informações. – 3.ed. – Rio de Janeiro : IBGE, 62p. 1993.

IBGE- Indicadores de Desenvolvimento Sustentável. Estudos e Pesquisas Informação Geográfica, n.7, Rio de Janeiro: IBGE, 443p. 2010.

IPEA - Instituto de Pesquisa Econômica Aplicada. Biocombustíveis no Brasil: Etanol e biodiesel. Secretaria de Assuntos Estratégicos da Presidência da República. Boletim n. 53, 57p. 2010.

MMA – Instituto do Meio Ambiente e dos Recursos Naturais Renováveis. Caderno setorial dos recursos hídricos: agropecuária. Ministério do Meio Ambiente, Secretaria dos Recursos Hídricos. Brasília: MMA, 96p. 2006.

VIEIRA, S. **Estatística Básica**. Editora Cengage Learning. São Paulo. 176p. 2011.

TICs

C

Estatística Descritiva - Distribuição de Frequência

Unidade C
Estatística Básica

UNIDADE C

ESTATÍSTICA DESCRITIVA - DISTRIBUIÇÃO DE FREQUÊNCIA

Objetivos

- Conceituar representação de dados numéricos;
- Conhecer distribuições de frequência para dados estatísticos;
- Descrever os tipos e a importância da distribuição de frequência;
- Representar graficamente uma distribuição de frequência.

Você verá por aqui...

Nesta aula vamos trabalhar assuntos relacionados à apresentação de dados estatísticos em distribuição de frequência.

A construção da tabela de distribuição de frequência, sua representação gráfica e ainda, sua interpretação, será abordada de forma sequencial e com exemplos de aplicação para que você possa acompanhar o assunto e entender os conceitos.

Ao final da aula você terá um exercício de aplicação do conteúdo estudado.

Procure organizar seus horários, disponibilizando um bom tempo para as atividades e aproveite bem este material.

Começando...

Você deve estar lembrado que, na aula 1, vimos as fases do método estatístico e que, após a etapa de coleta dos dados, os mesmos precisam ser organizados conforme o objetivo que se quer e apresentados sob a forma de tabelas, gráficos e medidas. Na aula anterior, vimos como apresentar os dados em tabelas e gráficos. Agora, vamos ver a apresentação dos dados em distribuições de frequência. Para tal, o conjunto de dados obtidos precisa estar sumarizado, mediante um critério de classificação e organizado para apresentação.

Representação de uma amostra

Considere os seguintes dados, que expressam a produção diária de leite, em kg/dia, de um plantel de vacas da raça holandesa, anotados na ordem em que foram coletados.

23,5	24,7	18,9	24,6	18,9
17,0	18,9	20,9	23,5	25,5
20,9	20,9	18,9	18,5	18,2
18,0	18,5	17,5	18,2	17,5

Quando os dados estão listados sem nenhum outro tipo de ordenação do que a própria sequência em que foram coletados, denominamos de **dados brutos**. Assim, se observarmos os dados acima, vamos verificar que não é tão facilmente identificado o menor ou o maior valor encontrado nas medições, ou se os dados estão concentrados ao redor de um valor definido ou se são bem espalhados, dispersos.

Por outro lado, se organizarmos os dados em ordem crescente ou decrescente ficará mais fácil de ter uma ideia a respeito da distribuição dos valores. Esta organização dos dados em ordem crescente ou decrescente é o que denominamos ***rol***. A seguir, os dados são apresentados de forma ordenada crescente.

17,0	17,5	17,5	18,0	18,2
18,2	18,5	18,5	18,9	18,9
18,9	18,9	20,9	20,9	20,9
23,5	23,5	24,6	24,7	25,5

Com os dados ordenados (***rol***) podemos agora verificar facilmente que o menor rendimento foi de 17,0 kg/dia e que o maior foi de 25,5 kg/dia, já delimitando a **amplitude total** dos dados.

Amplitude total dos dados, ou ***range***, é a diferença entre o maior e o menor valor medido da variável em estudo.

No exemplo apresentado, a amplitude total é 8,5 (25,5 – 17,0 = 8,5).

Para descrever dados estatísticos resultantes de variáveis qualitativas ou quantitativas, utilizamos as **distribuições de frequência**. Para as variáveis qualitativas poderemos estabelecer apenas medidas de frequência de ocorrência, enquanto para as variáveis qualitativas poderemos utilizar diversas medidas estatísticas de posição e de dispersão, como veremos nas aulas seguintes.

Distribuição de frequência

Uma distribuição de frequência é uma série estatística na qual os dados estão organizados em grupos de classes ou categorias estabelecidas convenientemente.

As distribuições de frequência podem ser divididas em dois tipos:

- Distribuição de frequência sem intervalos de classe, ou distribuição pontual, onde todos os valores dos dados coletados são apresentados, e não há perdas de valores ou,
- Distribuição de frequência com intervalos de classe, onde os valores estão representados por faixas de magnitude.

Para o exemplo apresentado, valores diários de produção de leite de vinte vacas holandesas, a distribuição de frequência sem intervalos de classes, ou pontual, é:

Tabela 1 - Distribuição de frequência pontual para a produção de leite

Produção de leite ($kg.dia^{-1}$)	Frequência absoluta (f_i)
17.0	1
17.5	2
18.0	1
18.2	2
18.5	2
18.9	4
20.9	3
23.5	2

24.6	1
24.7	1
25.5	1
Total	**20**

Observe que na tabela todos os valores de dados obtidos estão apresentados e, portanto, não há perda de informação.

Frequência absoluta (F_i)

Definimos como frequência absoluta (f_i) o número de vezes que o dado aparece na amostra, ou, no caso de estarmos apresentando uma distribuição de frequência por classes, o número de elementos pertencentes a uma classe.

Distribuição de frequência com intervalos de classe para a produção de leite.

Tabela 2 - Distribuição de frequência com intervalos de classe para a produção de leite.

Produção de leite ($kg.dia^{-1}$)	Frequência absoluta (f_i)
17.0 ⊢ 18,7	8
18,7 ⊢ 20,4	4
20,4 ⊢ 22,1	3
22,1 ⊢ 23,8	2
23,8 ⊢ 25,5	3
Total	**20**

A soma das frequências absolutas é igual ao número total de dados:

$$\sum fi = n$$

Limites de classe

Na Tabela 2 aparece uma notação (⊢) que é utilizada para identificar os **limites da classe** e significa que estão incluídos os valores mínimos e excluídos os valores máximos, ou seja, na classe 18,7 ⊢ 20,4 estão computados os valores de 18,7 (inclusive) a 20,4 (exclusive). Esta é a notação que deve ser utilizada para identificar os limites de classe, de acordo com a Resolução 866/66 do IBGE, ou seja, **desta quantidade até menos aquela**.

Outras notações:

18,7 ⊣ 20,4 – excluído o valor correspondente ao limite inferior e incluído o valor correspondente ao limite superior.

18,7 ⊣⊢ 20,4 – incluídos os valores entre 18,7 e 20,4 (excluídos os valores 18,7 e 20,4).

O menor número é o **limite inferior** da classe ($\mathbf{l}_i$) e o maior número, o **limite superior da classe** (**L**i).

Número de Intervalos de Classe (k)

Para definirmos o número de classes em que os dados serão divididos, podemos utilizar as seguintes fórmulas, sendo n= tamanho da amostra:

$$k = 5 \rightarrow se \;\; n < 25;$$
$$k = \sqrt{n} \rightarrow se \;\; n \geq 25$$

Fórmula de Sturges:

$$k=1+3,32\ \ log(n)$$

Em nosso exemplo, da Tabela 2:

a) n= 20 → k = 5

b) k = 1 + 3,32 log (20); k = 5,29 arredondando k = 5,0.

Na montagem de uma tabela de distribuição de frequência não há uma fórmula exata para o número de intervalos de classes. As que apresentamos acima são para encontrarmos um primeiro valor. Via de regra, devemos usar no mínimo 5 e no máximo 15 classes. Menos de cinco classes perdemos muita informação e, mais do que 15 classes, a tabela fica muito extensa, dificultando a interpretação dos dados.

Amplitude do intervalo de classe (h)

Calculamos a amplitude de cada classe dividindo a amplitude total pelo número de classes (k). Assim, para o exemplo, temos amplitude total de 8,5 e k=5, logo, a amplitude de cada intervalo de classe será 1,7 (8,5 / 5 = 1,7).

Geralmente, mas não obrigatoriamente, iniciamos a primeira classe pelo menor valor do conjunto de dados, somando o valor da amplitude de classe para encontrar o limite superior, e assim sucessivamente, até a última classe que poderá, ou não, ter o maior valor da variável em estudo como o limite superior da classe.

Em uma tabela de distribuição de frequência com intervalos de classe ganhamos simplicidade, mas perdemos informação. No exemplo da Tabela 2 podemos observar que 8 vacas produziram entre 17,0 e 18,7 kg de leite por dia, mas não sabemos exatamente quanto cada uma produziu.

Frequência relativa (fr_i):

A frequência relativa é dada pela razão entre a frequência absoluta de cada classe e a frequência total ou soma das frequências absolutas:

$$f_r=\frac{f_i}{\Sigma f_i}$$

A utilização da frequência relativa facilita as comparações entre mais de um conjunto de dados com diferentes números de elementos.

A soma das frequências relativas é sempre igual a 1.

Na Tabela 3 pode ser observada a frequência relativa de cada classe para o exemplo dado anteriormente.

Frequência acumulada (F):

A frequência acumulada é a soma da frequência absoluta da classe em questão com as frequências absolutas das classes anteriores, sendo a frequência acumulada da última classe igual ao número total de observações.

$$F_3=f_3+f_2+f_1$$

A Tabela 3 apresenta uma distribuição de frequência, onde podem ser visualizados todos os tipos de

frequência.

Tabela 3 - Distribuição de frequência em classes para a variável produção de leite

Classe (i)	Produção de leite ($kg.dia^{-1}$)	f_i	x_i	fr_i	F_i	Fr_i	fp_i	Fp_i
1	17.0 ├ 18,7	8	17,85	0,40	8	0,40	40,0	40,0
2	18,7 ├ 20,4	4	19,55	0,20	12	0,60	20,0	60,0
3	20,4 ├ 22,1	3	21,25	0,15	15	0,75	15,0	75,0
4	22,1 ├ 23,8	2	22,95	0,10	17	0,85	10,0	85,0
5	23,8 ├ 25,5	3	24,65	0,15	20	1,0	15,0	100
		Σ=20		Σ=1,0			Σ=100	

Frequência acumulada relativa (Fr_i):

É dada pela frequência acumulada (F) da classe, dividida pela frequência total ($\sum f_i$) do conjunto de dados. A frequência acumulada da última classe é igual à unidade.

$$Fr_i = \frac{F_-}{\Sigma f_i}$$

Frequência percentual (fp_i):

A frequência percentual é obtida pela multiplicação da frequência relativa por cem (100):

$$fp_i = fr_i . 100$$

Frequência acumulada percentual (Fp_i):

A frequência acumulada percentual é obtida pela multiplicação da frequência acumulada relativa por cem (100):

$$Fp_i = Fr_i . 100$$

A frequência acumulada percentual da última classe é igual a 100.

Ponto médio de uma classe (x_i):

O ponto médio de uma classe, como diz o nome, é o ponto que divide o intervalo de classe em duas partes iguais. É dado pela soma dos limites inferior e superior da classe dividido por dois.

$$x_i = \frac{l_i + L_i}{2}$$

Assim, para o exemplo apresentado (Tabela 3), o ponto médio da classe 3 é dado por:

$$x_3 = (l_3 + L_3)/2 \rightarrow x_3 = 20,4 + 22,1 / 2 \rightarrow x_3 = 42,5 / 2 \rightarrow x_3 = 21,25.$$

O ponto médio de uma classe é o valor que a representa.

O conhecimento dos valores referentes aos vários tipos de frequência, como apresentado na Tabela 3, ajuda-nos a responder alguns questionamentos com relativa facilidade, tais como os seguintes:

a) Quantas vacas produzem menos do que 20,4 kg de leite por dia?

Esse valor refere-se ao limite superior da segunda classe e, portanto, a resposta é igual à frequência acumulada da segunda classe (F_2): 12 animais ou 60% (Fp_2).

b) Quantas vacas produzem mais do que 23,8 kg de leite por dia?

Esse valor é o limite inferior da quinta classe, logo: 3 animais produzem mais do que 23,8 kg de leite diários (f_5) ou 15,0% do total (fp_5).

Representação gráfica de uma distribuição de frequência:

Uma distribuição de frequência pode ser representada graficamente pelo histograma e pelo polígono de frequência.

Histograma

O histograma é formado por um conjunto de retângulos justapostos, cujas bases se localizam sobre o eixo horizontal, de tal modo que seus pontos médios coincidam com os pontos médios dos intervalos de classe (CRESPO, 2009).

As larguras dos retângulos são iguais às amplitudes dos intervalos de classe e as alturas devem ser proporcionais às frequências das classes e, dessa maneira, as alturas serão numericamente iguais às frequências. Os dados da Tabela 3 estão representados na Figura 1.

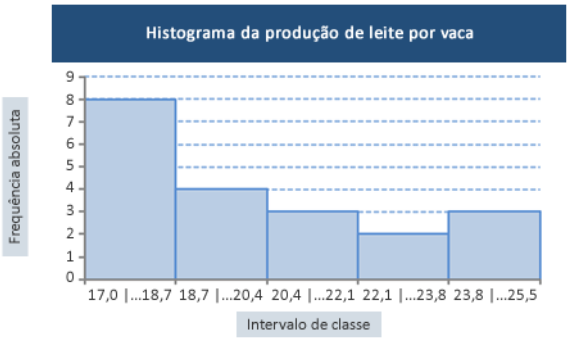


Figura C.1 - Histograma da distribuição de frequência para a produção de leite, conforme Tabela 3.

Polígono de frequência

O polígono de frequência é um gráfico em linha, sendo as frequências marcadas sobre os pontos médios dos intervalos de classe e unidas por segmentos de retas. Para fazer com que o gráfico inicie e termine sobre o eixo horizontal, criamos uma classe antes e depois da distribuição de frequência que queremos representar, com o mesmo intervalo entre classes, e definimos como ponto médio zero. A Figura 2 ilustra o procedimento e representa os dados da Tabela 3.

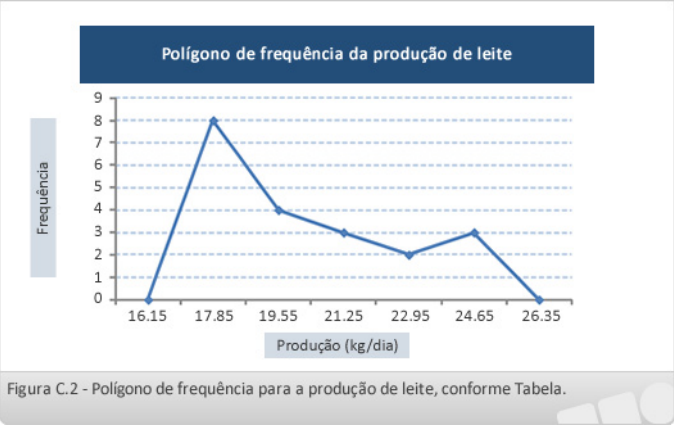


Figura C.2 - Polígono de frequência para a produção de leite, conforme Tabela.

Resumo

Nessa aula estudamos como organizar um conjunto de dados estatísticos; os conceitos de dados brutos e rol; como construir uma tabela de distribuição de frequência; os diversos tipos de frequência; como interpretar os dados em uma distribuição de frequência e, ainda, como representar graficamente as informações de uma distribuição de frequência.

Exercícios

Em uma indústria que fabrica peças mecânicas de reposição foi realizado um levantamento para detectar o índice de peças com defeitos em uma máquina. A contagem foi feita a partir de inspeção nas peças produzidas durante cada jornada de trabalho, durante 48 dias de avaliação. Os resultados foram tabulados por número de peças com defeito conforme a sequência de dias avaliados e estão apresentados no quadro abaixo.

12	14	19	8	18	12	11	13
2	17	13	6	21	13	16	4
1	14	18	16	34	15	14	1
0	16	0	2	22	9	21	13
28	15	17	11	11	19	20	16
13	19	8	12	8	18	10	9

A partir das informações responda às seguintes questões:

1. Organize os dados em rol.

2. Monte uma distribuição de frequência com intervalos de classe, adotando o limite inferior da primeira classe igual a zero.

3. Calcule o ponto médio de cada classe (x_i).
4. Calcule as frequências relativas (fr_i).
5. Calcule as frequências acumuladas (F_i).
6. Calcule as frequências acumuladas relativas (Fr_i).
7. Calcule as frequências percentuais (fp_i).
8. Calcule as frequências acumuladas percentuais (Fp_i).
9. Construa um histograma para a distribuição de frequência.

Referências

CRESPO, A. A. **Estatística Fácil**. São Paulo, Ed. Saraiva. 218p. 2009.

TICs

D

Estatística Descritiva - Medidas de Posição

Unidade D
Estatística Básica

UNIDADE D

ESTATÍSTICA DESCRITIVA - MEDIDAS DE POSIÇÃO

Objetivos

- Identificar as medidas de posição.
- Conceituar média, mediana e moda.

Você verá por aqui...

Nesta aula, estudaremos as medidas estatísticas que indicam a centralidade dos dados de uma distribuição e que indicam um valor que melhor representa o conjunto de dados.

Ao final desta aula, apresentamos uma relação de exercícios e atividades que têm o objetivo de fixar os conteúdos estudados.

Procure organizar seus horários, disponibilizando um bom tempo para as atividades e aproveite bem este material.

Começando...

Para ressaltar as tendências características de uma série estatística, isoladamente, ou em comparação com outras, necessitamos conhecer algumas medidas que nos permitam entender essas tendências, como as medidas de posição.

Medidas de posição

Medidas de posição, também chamadas de medidas de tendência central, referem-se à média, à moda e à mediana, que apresentam formas de obtenção e aplicação diferentes. As medidas de tendência central, ou de posição, fornecem um resumo dos dados estatísticos e dão ideia do centro em torno do qual os dados se distribuem, indicando, assim, um valor que melhor representa todo o conjunto de dados.

Para estudar essas medidas, precisamos antes conhecer alguns símbolos matemáticos que são utilizados em suas definições e cálculo.

Símbolos matemáticos

Para a representação dos valores de uma variável utilizamos x_1, x_2, x_3, ...x_n.

O subscrito indica a posição do valor da variável na sequência e, dessa forma, x_1 representa o primeiro valor observado, x_2 o segundo e assim por diante e x_i é o iésimo valor no conjunto de *n* valores.

A letra grega sigma (∑) é utilizada para indicar a soma dos *n* valores assumidos pela variável x_i , e lemos como "somatório de", conforme mostrado a seguir:

$$\Sigma x_i = x_1 + x_2 + x_3 + \ldots x_n$$

Média da amostra

A média aritmética, ou simplesmente média, é a medida de tendência central mais conhecida e utilizada para resumir a informação contida em um conjunto de dados (VIEIRA, 2011).

A média de um conjunto de dados é obtida somando todos os dados e dividindo o resultado pelo número total de dados.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Média de dados apresentados em tabela de distribuição de frequência

A média de dados discretos agrupados em uma tabela de distribuição de frequências é dada pelo somatório dos produtos dos valores da variável (x_i) pelas respectivas frequências (f_i), dividido pela soma das frequências.

$$\bar{x} = \frac{\Sigma\ x_i f_i}{\Sigma f_i}$$

Dados contínuos podem estar agrupados em classes e ser apresentados em tabelas de distribuição de frequências. Para calcularmos a média é necessário antes calcular o valor central de cada classe. Relembre que o valor central de cada classe, ou ponto médio da classe, é dado pela soma dos limites inferior e superior, dividida por dois.

$$\bar{x} = \frac{\Sigma\ x_i^* f_i}{\Sigma f_i}$$

Onde x_i^* é o valor central de cada classe ou ponto médio da classe.

Mediana da amostra

A mediana (Me) de um conjunto de dados é o valor **cuja posição** separa o conjunto de dados em duas partes iguais. Metade do número de elementos possui valor maior que a mediana e a outra metade possui valores menores do que a mediana.

Se o número de dados é ímpar, existe um único valor na posição central. Esse valor é a mediana dos dados.

Exemplo 1:

Sejam os valores 2, 3, 5, 6 e 7, a mediana tem valor 5.

Se fossemos calcular a média, essa seria igual a 4,6.

Se o número de dados é par, existem dois valores na posição central e a mediana é a média desses dois valores.

Exemplo 2:

Sejam os valores 2, 3, 5 e 6, a mediana é a média entre 3 e 5, logo, a mediana tem valor 4. A média calculada será 16 / 4 = 4.

Quando ocorrem dados **discrepantes** (valores muito maiores ou menores que os demais), esses valores podem **alterar a média**, distorcendo essa medida de posição. Para esses casos, o mais correto será usar a mediana para descrever a tendência central dos dados.

Exemplo 3:

Sejam os valores 2, 3, 5, 6, 7, 9, 9, 38.

A média calculada será: (2+3+5+6+7+9+9+38)/8 = 79/8 = 9,87.

É fácil notar que a média é maior do que 7 dos 8 dados que compõem a amostra.

A mediana será: a média dos valores 6 e 7, logo, a mediana é 6,5.

Se o valor discrepante (38) fosse substituído por um valor mais coerente com a série de dados, por exemplo, 11, o cálculo da média seria:

(2+3+5+6+7+9+9+11)/8 = 52/8 = 6,5. Note que o valor da mediana não seria alterado, permanecendo igual a 6,5.

O valor da mediana **pode coincidir ou não** com o valor de um elemento da série de dados. Quando o número de elementos da série é ímpar, haverá coincidência entre a mediana e um valor da série, entretanto, se o número de elementos é par, não haverá coincidência.

Moda

A moda (Mo) é o valor que ocorre com maior frequência em um conjunto de dados.

A moda é muito informativa quando o conjunto de dados é grande, mas se o conjunto de dados for relativamente pequeno (20 ou 30 observações), a moda não tem, em geral, sentido prático (VIEIRA, 2011).

A moda também pode ser utilizada para descrever dados qualitativos. Nesse caso, a moda é a categoria que ocorre com maior frequência, ou seja, a categoria que concentra a maior quantidade de dados.

Um conjunto de dados pode não ter moda, ou ter duas ou mais modas.

Exemplo 4:

Seja o conjunto de dados: 3, 5, 7, 6, 4, 9, 8. Este conjunto de dados não possui moda, pois todos os valores ocorrem uma única vez. Nesse caso, o conjunto apresenta uma **distribuição amodal**.

Exemplo 5:

Seja o conjunto de dados: 3, 5, 4, 6, 4, 9, 8. Nesse caso o conjunto apresenta moda igual a 4 e a d**istribuição é unimodal**, pois apresenta uma única moda.

Exemplo 6:

Seja o conjunto de dados: 2, 7, 7, 13, 15, 15, 22. Este conjunto apresenta duas modas, Mo_1= 7 e Mo_2= 15, sendo denominada **distribuição bimodal**.

Quando a distribuição apresenta mais de uma moda, como no exemplo 6, o histograma tem mais de um

pico. Conjunto de dados com três modas é denominado **trimodal** e com quatro ou mais modas é dito **multimodal**.

A moda é utilizada quando desejamos obter uma medida rápida e aproximada de posição ou quando a medida de posição deve ser o valor mais típico da distribuição (CRESPO, 2009).

Para calcular a moda de uma variável em uma série de dados, precisamos apenas da distribuição de frequências (contagem). Já para a mediana necessitamos minimamente ordenar as realizações da variável. Finalmente, a média só pode ser calculada para variáveis quantitativas (BUSSAB & MORETTIN, 2010).

As condições citadas limitam bastante o cálculo de medidas-resumos para as **variáveis qualitativas**. Para as variáveis **nominais** somente podemos trabalhar com a **moda** e para as variáveis ordinais, além da moda, podemos usar também a mediana.

Resumo

Nesta aula estudamos as medidas de posição, ou medidas de tendência central, definindo os conceitos e aplicações da média, da mediana e da moda, em uma série de dados estatísticos.

Exercícios

1. Dada a série de dados: 15, 40, 25, 50, 70, 55, a mediana será

a) 40.
b) 30.
c) 45.
d) 35.

2. Em uma série estatística, 50% dos dados situa-se

a) acima da média.
b) abaixo da moda.
c) acima da mediana.
d) abaixo da média.

3. Na distribuição de frequências apresentadas, o valor da média é:

x_i	1,0	3,0	4,0	5,0	6,0
fi	2	4	5	4	3

a) 4,5
b) 5,0
c) 3,0
d) 4,0

4. Na série de dados: 82, 86, 88, 88, 84, 90, 85, o valor da mediana e da moda é, respectivamente,

a) 88 e 90.
b) 88 e 88.
c) 84 e 88.
d) 86 e 88.

5. Dada a série de dados: 2, 5, 7, 3, 6, 4, 8, os valores da mediana, da moda e da média será, respectivamente,

a) 3, multimodal, 5.
b) 5, amodal, 4.
c) 4, multimodal, 5.
d) 3, amodal, 4.

Referências

BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. São Paulo, Ed. Saraiva. 540p. 2010.

CRESPO, A. A. **Estatística Fácil**. São Paulo, Ed. Saraiva. 218p. 2009.

VIEIRA, S. **Estatística Básica**. Editora Cengage Learning. São Paulo. 176p. 2011.

Estatística Descritiva - Medidas de Dispersão, Assimetria e Curtose

Unidade E
Estatística Básica

UNIDADE E

ESTATÍSTICA DESCRITIVA - MEDIDAS DE DISPERSÃO, ASSIMETRIA E CURTOSE

Objetivos

- Identificar as diferentes medidas de dispersão;
- Calcular as diversas medidas de dispersão: amplitude, variância, desvio padrão e coeficiente de variação;
- Conceituar assimetria e curtose;
- Calcular os coeficientes de assimetria e de curtose.

Você verá por aqui...

Nesta aula vamos conhecer as medidas de dispersão dos dados estatísticos, trabalhando os conceitos de mínimo, máximo, amplitude total, variância, desvio padrão, coeficiente de variação e os coeficientes de assimetria e curtose.

Ao final desta aula apresentamos uma relação de exercícios e atividades que têm o objetivo de fixar os conteúdos estudados.

Procure organizar seus horários disponibilizando um bom tempo para as atividades e aproveite bem este material.

Começando...

As medidas de tendência central – média, mediana e moda, como vimos na aula anterior, fornecem importantes informações a respeito do centro em torno do qual os dados estão dispersos, mas não quantificam o quanto os dados se dispersam, ou se distribuem, ao redor das medidas de centralidade. Para isso precisamos, além da medida de tendência central, uma medida de variabilidade ou dispersão.

Medidas de dispersão ou de variabilidade

Amplitude total

A amplitude total (AT), ou range, em um conjunto de dados é a diferença entre o maior e o menor valor observado.

Assim, podemos estabelecer que: AT = x(máx.) – x (mín.)

Tomando por exemplo os seguintes dados amostrais:

14, 21, 23, 21, 24, 24, 25, 26, 26, 27, 35

Calculamos a amplitude total como:

$$AT = 35 - 14 \rightarrow AT = 21$$

Podemos perceber que a amplitude total considera apenas o valor máximo e o valor mínimo, não importando todos os demais valores do conjunto de dados. Esta condição tem o inconveniente de expor a medida de amplitude total à presença de valores extremos ou atípicos, chamados *outliers*, o que quase sempre invalida a idoneidade do resultado (CRESPO, 2009).

Assim, devemos aperfeiçoar a descrição da variabilidade dos dados através de outras medidas de dispersão, como a variância ou o desvio padrão.

Variância

A **variância (s^2)** e o desvio padrão levam em consideração a totalidade dos valores assumidos pela variável em estudo, e assim, são índices de variabilidade bastante estáveis e geralmente os mais empregados. A variância expressa a média aritmética dos quadrados dos desvios. Definimos desvio como sendo a diferença entre um determinado valor da variável em estudo e a média dos valores totais.

Assim, a variância de uma **população** será calculada por:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}$$

Onde: x_i = valor de ordem *i* assumido pela variável

μ = média dos valores de *x*
s^2 = **variância populacional**
N = número de dados **da população**

Se os dados que estamos avaliando é uma amostra da população total de dados então a variância calculada é denominada **variância amostral**, e será calculada pela seguinte fórmula:

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

A diferença de cálculo entre as fórmulas está no denominador. O somatório dos quadrados dos desvios foi dividido por n-1 em lugar de N, porque proporciona uma melhor estimativa da variância populacional e porque, sendo nula a soma dos desvios, existem (n – 1) desvios independentes, e assim, conhecidos (n – 1) desvios o último está automaticamente determinado, pois a soma é zero.

Embora o processo de cálculo da média aritmética seja o mesmo para um conjunto de dados que representem uma amostra da população ou para todos os dados que compõem a população, utiliza-se o símbolo **μ** para representar a média da população, e **N** para o número de elementos da população.

Considerando os mesmo dados com os quais calculamos a amplitude total, vamos ver como obtemos os desvios e a variância dos dados, demonstrados na Tabela 1.

Tabela 1 – Cálculo dos desvios para obtenção da variância.

Ordem	Dados (x)	*Desvio* $\left(xi-\bar{x}\right)$	*Desvio²* $\left(xi-\bar{x}\right)^2$
1	14	-10,182	103,673
2	21	-3,182	10,125
3	21	-3,182	10,125
4	23	-1,182	1,397
5	24	-0,182	0,033
6	24	-0,182	0,033
7	25	0,818	0,669
8	26	1,818	3,305
9	26	1,818	3,305
10	27	2,818	7,941
11	35	10,818	117,029
n=11	$\sum$=266		$\sum$=254,33

A média será:

$\bar{x} = \frac{266}{11}$ $\bar{x} = 24{,}182$

E a variância **amostral**:

$$s^2 = 254{,}33/10 \longrightarrow \mathbf{s^2 = 25{,}43}$$

Quando a média não é exata e tem de ser arredondada, cada desvio fica ligeiramente afetado pelo erro devido ao arredondamento. O mesmo irá acontecer com os quadrados dos desvios. Uma fórmula alternativa, bastante usada para o cálculo da variância amostral, é:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]$$

Se estivermos calculando a variância populacional, substituímos 1/n-1 por 1/N.

Recalculando a variância dos dados apresentados na Tabela 1, através da fórmula alternativa, temos:

Tabela 2 – Cálculo da variância pela fórmula alternativa.

Ordem	(x_i)	$(x_i)^2$
1	14	196
2	21	441
3	21	441
4	23	529
5	24	576
6	24	576
7	25	625
8	26	676
9	26	676
10	27	729
11	35	1225
n=11	$\sum x_i = 266$	$\sum x_i^2 = 6.690$

$$s^2 = \frac{1}{10}\left[(6.690) - \frac{(266)^2}{11}\right]$$

E o novo valor da variância será:

$$\mathbf{s^2 = 25,76}$$

A diferença é devido ao arredondamento da média, como descrito anteriormente.

Se os desvios em relação à media são pequenos, podemos concluir que as observações estão aglomeradas em torno da média e a variabilidade dos dados é pequena. Por sua vez, se os desvios são grandes, os dados estão muito dispersos e a variabilidade é grande. A variância tem a capacidade de captar essas duas situações, portanto é um bom índice estimador da variabilidade dos dados.

Como a variância é calculada a partir dos quadrados dos desvios, seu resultado é um número em unidade quadrada em relação à variável sob estudo, o que, sob o ponto de vista prático, é um inconveniente. Como exemplo, vamos assumir que os dados apresentados na Tabela 2 representassem dias de ocorrência de geada nos últimos onze anos em Pelotas (dados fictícios). A variância então assume o valor de 25,76 dias^2.

Se extrairmos a raiz quadrada da variância, teremos uma medida da variabilidade dos dados e os próprios dados na mesma unidade. Esse novo índice de variabilidade é o que definimos como **desvio padrão**.

Desvio padrão

O desvio padrão (s ou σ) é definido como sendo a raiz quadrada da média aritmética dos quadrados dos desvios e, dessa forma, é dado pela raiz quadrada da variância.

A equação para o cálculo do **desvio padrão amostral** é:

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

E para o **desvio padrão populacional**:

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

Quando o desvio padrão é calculado usando todos os elementos da população é simbolizado pela letra grega **σ** (sigma), denominado *desvio padrão populacional*, sendo considerado um **parâmetro**.

Se for calculado a partir de uma amostra da população, é representado pela letra s, denominado desvio padrão amostral, e é considerado um **estimador**.

O desvio padrão dos dados apresentados na Tabela 2 será:

$$s = \sqrt{25{,}76} \text{ onde: s = 5,07 dias}$$

Tanto o desvio padrão como a variância são usados como medidas de dispersão ou variabilidade. O uso de uma ou de outra dependerá da finalidade que se tenha em vista. A variância é uma medida que tem pouca utilidade como estatística descritiva, porém é extremamente importante na inferência estatística e em combinações de amostras (CRESPO, 2009).

Coeficiente de variação

O coeficiente de variação (CV) é uma medida de variabilidade relativa que mede a dispersão dos dados em relação à média aritmética, sendo expressa pela razão entre o desvio padrão e a média, e multiplicada por cem. Assim, é uma medida adimensional expressa em percentual.

Tipo	Representação	unidade
CV amostral	$CV = \frac{s}{x} \times 100$	%
CV populacional	$CV = \frac{\sigma}{\mu} \times 100$	%

Exemplo:

Considere amostragens realizadas em duas escolas, com relação às notas obtidas por alunos da disciplina de estatística na primeira avaliação.

Medida	Escola A	Escola B
Média	6,5	6,5
Variância	1,167	0,416
Desvio Padrão	1,08	0,64
CV %	16,61	9,93

Embora o valor das médias tenha a mesma magnitude, é possível perceber que a Escola B apresenta menor dispersão dos dados, demonstrado nos menores valores de variância, desvio padrão e coeficiente de variação. O que está indicando que os alunos da escola B tiveram um desempenho mais homogêneo, com os dados variando menos em relação à média (menores desvios). Apesar de, no exemplo apresentado, a avaliação parecer um tanto óbvia, existem situações em que a magnitude dos valores das variáveis que se está comparando são diferentes e a conclusão a respeito dos dados é dificultada quando apenas a média e o desvio padrão são apresentados. O coeficiente de variação é a medida mais utilizada quando existe interesse em comparar variabilidade de diferentes conjuntos de dados.

Medidas separatrizes

Pelo que apresentamos na aula anterior, relativo às medidas de posição, e nessa aula com relação às medidas de dispersão, podemos concluir que tanto a média como o desvio padrão são afetados por valores extremos. Assim, quando a distribuição dos dados não é simétrica, ou seja, os dados não se distribuem de forma homogênea ao redor da média, temos uma distribuição assimétrica e precisamos conhecer outras medidas que permitam uma boa caracterização do conjunto de dados.

As medidas separatrizes, embora não sejam consideradas individualmente como medidas de tendência central, são baseadas em sua posição na série de dados. Assim, os ***quartis***, os ***percentis*** e os ***decis***, juntamente com a mediana, são medidas conhecidas como separatrizes.

Quartis

Os ***quartis*** são os valores de uma série de dados que a dividem em quatro partes iguais. Assim, os *quartís* Q_1, Q_2 e Q_3, dividem os dados em quatro partes, de tal forma que cada parte possui 25% dos dados, ou um quarto (1/4), daí sua denominação.

O primeiro quartil (Q_1) separa os dados de maneira que 25% dos valores são inferiores ou igual ao Q_1 e 75% dos dados são superiores ou igual ao Q_1.

Consideremos os seguintes dados:

5	5,5	6	7,0	7,5	8,5	9,5	11
Q_1		Q_2		Q_3			
25%		50%		75%		100%	

Na representação gráfica dos quartis podemos perceber que, o primeiro quartil situa-se entre os valores de 5,5 e 6, o segundo quartil entre os valores de 7,0 e 7,5, e é coincidente com a mediana, e o terceiro quartil situa-se entre os valores de 8,5 e 9,5.

Para calcular a posição dos quartis utilizamos a seguinte relação:

$$PosQ_i = \frac{i(n+1)}{4}$$

Assim, para os dados apresentados, o Q_2 será:

$$PosQ_2 = \frac{2(8+1)}{4}$$

$$Pos\ Q_2 = 4,5$$

Interpretando o resultado: Significa que o segundo quartil equivale ao valor da posição 4,5, ou seja, intermediária entre a 4ª e a 5ª posição. Precisamos então fazer a média entre os valores da quarta e da quinta posição, 7,0 e 7,5, respectivamente. O valor de Q_2 será 7,25 e significa que, 50% dos valores são inferiores ou iguais a 7,25 e 50% dos valores são iguais ou superiores a 7,25. Observe que o valor de Q_2 coincide com o valor da mediana.

Amplitude interquartílica

A amplitude interquartílica, denotada por q, é a diferença entre o terceiro quartil (Q_3) e o primeiro quartil (Q_1). Assim temos

$$q = Q_3 - Q_1$$

Apesar de ser uma medida pouca utilizada, a amplitude interquartílica apresenta uma característica interessante que é a resistência, ou seja, esta medida, ao contrário da amplitude total, não sofre nenhuma influência de valores discrepantes.

Percentis

Os **percentis** dividem a série de dados em cem partes iguais, cada uma com 1%, de tal maneira que o P50 corresponde ao Q_2 e a mediana.

Para determinar a posição dos percentis utilizamos a seguinte relação:

$$PosP_i = \frac{i(n+1)}{100}$$

Assim, o P_{50} será:

$$PosP_{50} = \frac{50(8+1)}{100}$$

Pos P_{50} = 4,5 (que é igual ao Q2 calculado).

Medidas de assimetria

Uma distribuição é dita simétrica quando a média, a mediana e a moda são coincidentes. Assim, em uma distribuição simétrica, o gráfico que representa a distribuição dos valores e sua frequência tem a forma de sino, conforme a figura 1.

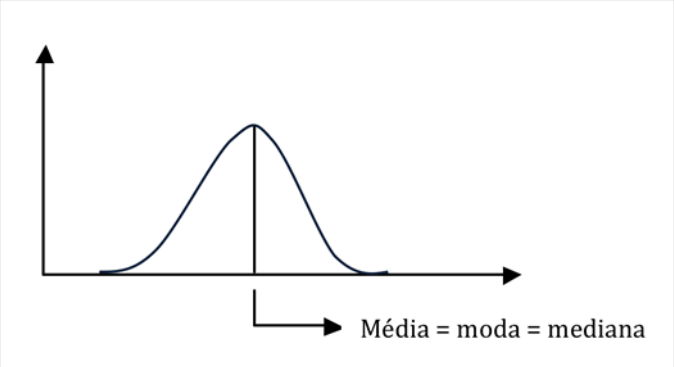


Figura E.1 - Representação de uma distribuição em forma de sino.

A partir da figura podemos observar que os dados apresentam o mesmo comportamento, ou distribuição, à direita e a esquerda da média e nesse caso, a média, a moda e a mediana são iguais ou muito próximas.

A assimetria é dita positiva (Figura 2-a) quando a cauda direita afasta-se mais do pico do que a cauda esquerda, e assim, a média é maior do que a mediana, a qual é maior do que a moda.

A distribuição de dados apresenta assimetria negativa (Figura 2-b) quando sua cauda esquerda afasta-se mais do pico do que a cauda direita, e nesse caso, a moda é maior do que a mediana, e esta maior do que a média.

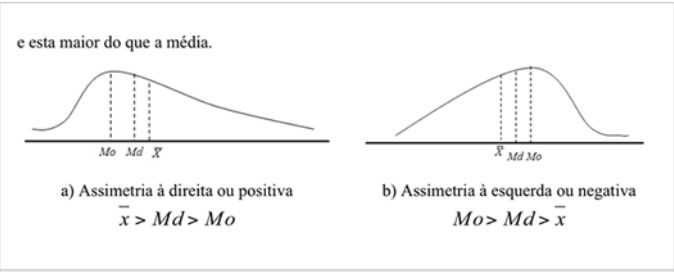


Figura E.2 - Representação de distribuições assimétrica positiva e negativa.

Há mais de uma forma para determinação da assimetria, sendo que a determinação pelo segundo critério de Pearson, ou coeficiente de assimetria de Pearson, é dada por:

$$As = \frac{3(\bar{x} - Md)}{s}$$

Onde: As = coeficiente de assimetria de Pearson

$\bar{x} = média$

Md = mediana

S = desvio padrão

O coeficiente de assimetria de Pearson tem as seguintes interpretações:

$As \leq -1$	Assimetria negativa
$-1 < As < -0,15$	Assimetria negativa moderada
$-0,15 \leq As \leq +0,15$	Simétrica
$+0,15 < As < +1$	Assimetria positiva moderada
$As \geq +1$	Assimetria positiva

Grau de Curtose

Denominamos de curtose o grau de achatamento ou alongamento da curva característica do conjunto de dados ou distribuição em relação a uma distribuição padrão, denominada curva normal.

A curva normal recebe o nome de mesocúrtica (Figura 3) e possui coeficiente de curtose (C) igual a 0,263.

Se uma curva é mais fechada que a normal (apresentando-se pontiaguda em sua parte superior), ela é chamada de leptocúrtica e possui coeficiente de curtose inferior a 0,263.

Se a curva apresentar um achatamento maior do que a curva normal, ela é chamada de platicúrtica e apresenta coeficiente de curtose maior do que 0,263.

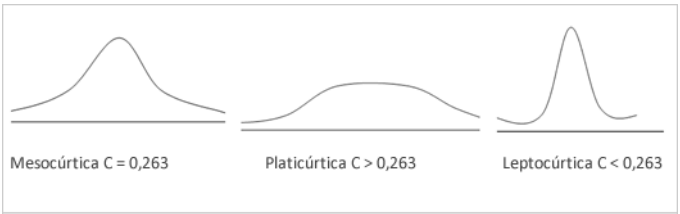


Figura E.3 - Nomenclaturas para o grau de curtose de uma curva de distribuição

Uma fórmula utilizada para o cálculo do coeficiente de curtose (C), conhecida como **coeficiente percentílico de curtose**, é apresentada a seguir:

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

Onde:

C = Coeficiente de Curtose;

Q_3 e Q_1 = quartil 3 e quartil 1, respectivamente;

P_{90} e P_{10} = percentil 90 e percentil 10, respectivamente.

Exercício Resolvido

1. Os resultados de 20 leituras de temperatura medidas no efluente da etapa de despolpa no processo de fabricação de pêssego em calda, são apresentados a seguir:

17,0	17,5	17,5	18,0	18,2
18,2	18,5	18,5	18,9	18,9
18,9	18,9	20,9	20,9	20,9
23,5	23,5	24,6	24,7	25,5

Calcule:

a) a média amostral

b) a amplitude total

c) a variância

d) o desvio padrão

e) o coeficiente de variação

f) a amplitude interquartílica

g) o coeficiente de assimetria e a classificação da assimetria

h) o coeficiente de curtose e a classificação da curtose

Respostas:

Organizando os dados em uma tabela, temos:

Ordem	(x_i)	$(x_i)^2$
1	17,0	289,00
2	17,5	306,25
3	17,5	306,25
4	18,0	324,00
5	18,2	331,24
6	18,2	331,24
7	18,5	342,25
8	18,5	342,25
9	18,9	357,21
10	18,9	357,21
11	18,9	357,21
12	18,9	357,21
13	20,9	436,81
14	20,9	436,81
15	20,9	436,81
16	23,5	552,25
17	23,5	552,25
18	24,6	605,16
19	24,7	610,09
20	25,5	650,25
n=20	Σ=403,50	Σ=8281,75

a) **a média** será:

$$\bar{x} = \frac{403,50}{20} \qquad \bar{x} = 20,175$$

b) **a amplitude total** será:

AT = 25,5 – 17,0 $\longrightarrow$ AT = 8,5

c) **a variância**:

usando a fórmula: $s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]$

$s^2 = 7,43$

$$s^2 = \frac{1}{19}\left[8281,75 - \frac{(403,50)^2}{20}\right]$$

d) **desvio padrão**:

$$s = \sqrt{7,43}$$

s = 2,725

e) **coeficiente de variação**:

CV = (2,725 / 20,175) x 100

CV = 13,51%

f) **amplitude interquartílica (q)**:

$q = Q_3 - Q_1$

$PosQ_3 = [3(20+1)] / 4$

$PosQ_3 = 15,75$

$Pos_{15} = 20,9$ e $Pos_{16} = 23,5$ logo, $Pos_{15,75} = 22,85$ e $Q_3 = 22,85$

$PosQ_1 = [1(20+1)] / 4$

$PosQ_1 = 5,25$

$Pos_5 = 18,2$ e $Pos_6 = 18,2$ logo, $Pos_{5,25} = 18,2$ e $Q_1 = 18,2$

q= 22,85 – 18,2

q = 4,65

g) **o coeficiente de assimetria e a classificação da assimetria:**

- para calcular o coeficiente de assimetria temos que conhecer o valor da mediana (Md):

$$As = \frac{3(\bar{x} - Md)}{s}$$

$$As = \frac{3(20{,}175 - 18{,}9)}{2{,}725}$$

As = 1,40 ⟶ assimetria positiva

h) **o coeficiente de curtose e a classificação da curtose:**

- para calcular o coeficiente de curtose (C), precisamos antes calcular P_{90} e P_{10} :

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

$$PosP_{10} = \frac{10(20+1)}{100}$$

$PosP_{10}$= 2,1 → $\underline{P_{10} = 17{,}5}$

$$PosP_{90} = \frac{90(20+1)}{100}$$

$PosP_{90}$= 18,9 → Pos_{18} = 24,6 e Pos_{19} = 24,7 logo, $\underline{P_{90} = 24{,}7}$

$$C = \frac{15{,}75 - 5{,}25}{2(24{,}7 - 17{,}5)}$$

C = 0,729 ⟶ platicúrtica (C > 0,263)

Na Figura 4 é apresentado o histograma dos dados de temperatura utilizados no exercício. Podemos observar a assimetria, com a cauda direita afastando-se mais do pico do que a cauda esquerda, e assim, a média (20,175) é maior do que a mediana (18,9) e do que a moda (18,9).

Podemos observar também, o maior achatamento da curva de distribuição dos dados em relação à curva normal, (platicúrtica) com coeficiente de curtose maior do que 0,263.

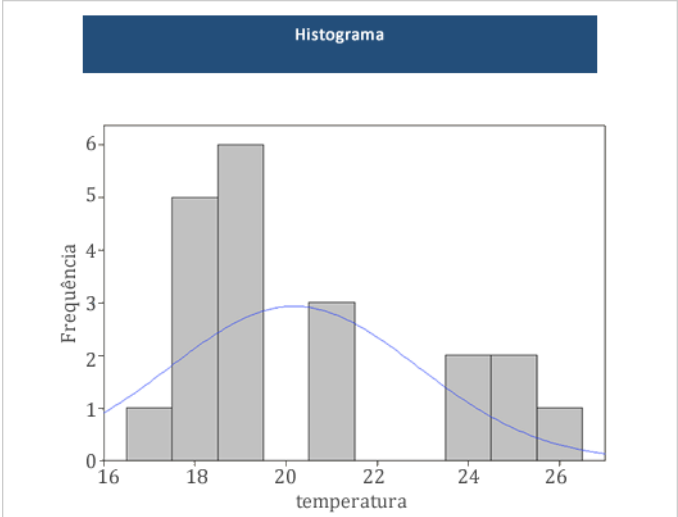


Figura E.4 - Histograma dos dados de temperatura utilizados no exercício 1.

Quadro resumo

Principais medidas de dispersão, assimetria e curtose	
Amplitude total	AT = x(máx.) – x (mín.)
Variância populacional	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$
Variância amostral	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
Variância amostral (média não exata)	$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]$
Desvio padrão amostral	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
Desvio padrão populacional	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}}$
Coeficiente de variação amostral	$CV = \frac{s}{\bar{x}} \times 100$
Quartís	$PosQ_i = \frac{i(n+1)}{4}$
Intervalo interquartílico	$q = Q_3 - Q_1$
Percentís	$PosP_i = \frac{i(n+1)}{100}$
Coeficiente de Assimetria de Pearson	$As = \frac{3(\bar{x} - Md)}{s}$
Coeficiente percentílico de Curtose	$C = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$

Exercícios

1. Os dados a seguir são relativos ao número de horas de funcionamento de um determinado tipo de lâmpada, até apresentar falhas.

441	275	470	341	535	497
285	327	389	400	512	295
420	415	360	430	600	915

Calcule:

a) a média amostral;
b) a amplitude total;
c) a variância;
d) o desvio padrão;
e) o coeficiente de variação;

f) a amplitude interquartílica;
g) o coeficiente de assimetria e a classificação da assimetria;
h) o coeficiente de curtose e a classificação da curtose.

Referências

CRESPO, A. A. **Estatística Fácil**. São Paulo, Ed. Saraiva. 218p. 2009.

Introdução à Probabilidade

Unidade F
Estatística Básica

UNIDADE F

INTRODUÇÃO À PROBABILIDADE

Objetivos

- Definir probabilidade;
- Identificar eventos complementares, independentes e mutuamente exclusivos;
- Calcular probabilidades para situações simples;
- Calcular probabilidades para combinação de eventos;

Introdução à probabilidade

As aplicações iniciais da matemática da probabilidade eram quase que exclusivamente relacionadas aos jogos de azar. Entretanto, a utilização de cálculos probabilísticos ultrapassou de muito o âmbito desses jogos. Embora pertencendo ao campo da Matemática, a inclusão da probabilidade nestes estudos se deve ao fato de que a maioria dos fenômenos de que trata a Estatística ser de natureza aleatória ou probabilística (CRESPO, 2009).

> A teoria da probabilidade permite que se calcule a chance de ocorrência de um determinado evento.

No estudo das probabilidades trabalhamos com experimentos, tais como lançar uma moeda ou um dado, ou ainda retirar uma determinada carta de um baralho. Um experimento é entendido como sendo qualquer processo que permite fazer observações, e são classificados em dois tipos:

Experimentos determinísticos

São experimentos em que o resultado é sempre o mesmo, ou se pode esperar que seja sempre o mesmo, apesar de ser repetido várias vezes em condições semelhantes. Como exemplo podemos citar um experimento para medir a temperatura de evaporação da água ao nível do mar. Por mais que se faça repetições a água irá ferver a 100°C.

Experimentos aleatórios

São os experimentos que, mesmo repetido várias vezes sob condições semelhantes, apresenta resultados imprevisíveis, entre os resultados possíveis, ou seja, são resultados explicados ao acaso. Como exemplo podemos citar, entre outros tantos, o lançamento de uma moeda ou o lançamento de um dado.

Um experimento aleatório é representado pela letra grega épsilon (ε), e o conjunto de resultados possíveis do experimento é denominado de **espaço amostral**, sendo representado pela letra **S**.

No experimento aleatório, lançamento de uma moeda, o resultado será cara ou coroa, que são os dois resultados possíveis, ou seja, o **espaço amostral** (**S**) de nosso **experimento aleatório** ($\boldsymbol{\varepsilon}$).

Se jogarmos uma moeda não viciada para cima, sempre sob as mesmas condições, não temos como prever se encontraremos como resposta “cara” ou “coroa”.

Figura F.1 - No experimento aleatório lançamento de uma moeda, o espaço amostral (S) é cara ou coroa.

Os escudos foram as primeiras moedas cunhadas no Brasil com a imagem do rei em uma das faces; na outra, trazia as Armas da Coroa portuguesa. Desse uso originou-se a expressão popular CARA/COROA, para indicar os dois lados das moedas. (Fonte: BCB, 2011).

O mesmo podemos dizer sobre um dado não viciado. Nunca saberemos se encontraremos como resposta “1”, “2”, “3”, “4”, “5” ou “6”.

Figura F.2 - O lançamento de um dado não viciado é um experimento aleatório.

Para esses dois experimentos, o espaço amostral, ou simplesmente o conjunto de resultados possíveis, representado por S, será:

No lançamento da moeda: S = {Cara, Coroa}

No lançamento do dado: S = {1, 2, 3, 4, 5, 6}

O **espaço amostral** pode ainda ser classificado, com relação ao seu número de elementos em:

- **Finito**: caso o número de elementos seja contável.
- **Infinito**: caso o número de elementos seja não enumerável.
- **Discreto**: conforme o número de elementos seja contável.
- **Contínuo**: se seus elementos são valores quaisquer entre dois números reais (expressos através de intervalos).

Evento

É um subconjunto qualquer do espaço amostral e é representado por uma das primeiras letras (maiúsculas) de nosso alfabeto (A, B, C, D, ...).

Exemplos:

No lançamento da moeda: evento A = cara; evento B= coroa.

No lançamento do dado (apresentamos apenas 3 dos diversos eventos que podem ser estabelecidos):

A= obter um número impar: A{1,3,5};

B= obter um número múltiplo de 2: B{2,4,6};

C= obter um número maior do que 3: C{4,5,6}.

Tipos de eventos

Quando estudamos um evento, em relação à probabilidade de ocorrência, ele pode ser classificado em evento certo, evento possível, evento impossível ou evento contingente. Observe a seguir as características de cada um desses tipos.

Evento certo:

se ele tem os mesmos elementos que o espaço amostral, ou seja, se A = S, temos que A é um evento certo. Exemplo: em um lançamento de dado ocorrer uma face menor do que 7, e logo, A = {1, 2, 3, 4, 5, 6}.

Evento possível:

Qualquer evento que não seja um conjunto vazio, mas não seja igual ao espaço amostral.

Evento impossível:

Quando o evento é um conjunto vazio, temos que ele é um evento impossível. Sendo B = Ø, significa que B é um evento impossível.

Evento contingente: são os eventos que nenhuma situação anterior poderia garantir sua ocorrência. Por exemplo, no lançamento de uma moeda, essa se posiciona de tal forma que o resultado não é cara nem é coroa.

Quando comparamos um evento com outro, de um mesmo espaço amostral, podemos classificá-los em:

Eventos complementares

Um evento é denominado complementar de outro, se é formado pelos elementos do espaço amostral que não pertencem ao segundo evento. Se, por exemplo, no lançamento de uma moeda, temos o evento A = {cara}, o evento B = {coroa} é seu complementar.

Sendo p a probabilidade de um evento e q a probabilidade de outro evento, eles são complementares se: p + q = 1.

Logo:

$$p = 1 - q$$

Exemplo:

Se a probabilidade de obter 2 no lançamento de um dado é 1/6, a probabilidade de não tirar 2 (ou seja, tirar qualquer outro número) é:

1 – 1/6 = 5/6

Eventos independentes

Dois eventos são independentes se a realização (ou não realização) de um dos eventos não afeta a probabilidade de realização do outro.

Se lançarmos, por exemplo, dois dados, o valor que obtivermos no 1º não afeta em nada o valor que obteremos no 2º, de modo que a probabilidade de que eles se realizem simultaneamente é igual ao produto das probabilidades de realização dos dois eventos.

Sendo p1 a probabilidade de realização do evento 1, e p2 a probabilidade de realização do evento 2, a probabilidade de ocorrência simultânea dos dois eventos é:

$$p = p1 . p2$$

Exemplo:

Ao lançarmos dois dados, a probabilidade de no primeiro ocorrer a face 4 voltada para cima é 1/6 e de ocorrer a face 2 voltada para cima no segundo dado é 1/6, logo, a probabilidade (p) de ocorrer 4 no primeiro dado e 2 no segundo dado é:

p1= 1/6

p2= 1/6

p= p1.p2 $\longrightarrow$ p= 1/6 x 1/6 $\longrightarrow$ p= 1/36.

Eventos mutuamente exclusivos

Dois ou mais eventos são ditos mutuamente exclusivos quando a realização de um exclui a possibilidade de realização do(s) outro (os).

No lançamento de uma moeda, o evento “cara” automaticamente exclui o evento “coroa”, uma vez que só há as duas possibilidades de ocorrência. As duas faces não podem ser obtidas no mesmo lançamento.

Assim, a probabilidade de que um OU outro evento se realize é a soma das probabilidades:

$$p = p1 + p2$$

A probabilidade de tirarmos cara **OU** coroa é 1/2 + 1/2 = 1

A probabilidade de tirarmos 5 **OU** 6 no lançamento de um dado é 1/6 + 1/6 = 2/6 = 1/3.

Podemos dizer que A e B são dois eventos mutuamente exclusivos se não apresentam elementos comuns.

Ou seja, considerando dois eventos A e B quaisquer, A e B são mutuamente exclusivos se $A \cap B = \emptyset$.

Dois eventos complementares são sempre mutuamente exclusivos, mas a recíproca nem sempre é verdadeira.

Resumo

Nessa aula definimos probabilidade e estudamos a aplicação de cálculos para estimar a chance de ocorrência de eventos complementares, independentes ou mutuamente exclusivos, em experimentos aleatórios.

Exercícios resolvido

1. Qual a probabilidade de obtermos uma dama de ouros ao retirarmos uma carta de um baralho de 52 cartas?

Resposta: p = 1/52

2. 2) Qual a probabilidade de obtermos uma dama de qualquer naipe ao retirarmos uma carta de um baralho de 52 cartas?

Resposta: p = 4/52 = 1/13

3. 3) De dois baralhos de 52 cartas, qual a probabilidade de retirarmos uma dama de cada baralho?

Resposta: como os dois eventos são independentes, temos p = p1 . p2
p= 1/13 . 1/13 = 1/169

4. 4) De um baralho de 52 cartas retiram-se duas cartas sem reposição. Qual a probabilidade da primeira ser a dama de ouros e a segunda a dama de copas?

Resposta: mais uma vez os eventos são independentes e temos p = p1 . p2. Importante observar que devido a retirada da primeira carta, restaram somente 51 cartas no baralho.
p= 1/52 . 1/51 = 1 / 2.652

Respostas:

1. Foi realizado um levantamento em 50 famílias de uma determinada localidade, sobre o número de filhos por família, sendo os resultados apresentados na tabela a seguir:

Número de Filhos	Frequência
0	8
1	20
2	12
3	6
4	3
5	1
	∑ = 50

Para o caso de se escolher aleatoriamente uma família, calcule:

a) a probabilidade de que tenha pelo menos um filho;
b) a probabilidade de que tenha três filhos ou mais;
c) a probabilidade de que tenha exatamente dois filhos.

2. Um número inteiro entre 1 e 20 é escolhido ao acaso.

a) Qual a probabilidade de que este número seja par?
b) Qual a probabilidade de que este número seja divisível por 3?
c) Qual a probabilidade de que este número seja par e divisível por 4?

3. Dois dados são lançados conjuntamente. Determine a probabilidade de:

a) a soma ser igual a 10 ou maior;
b) a soma ser 10;
c) a soma ser maior que 10.

Referências

BCB / Banco Central do Brasil. **História do Dinheiro no Brasil**. Disponível em http://www.bcb.gov.br/htms/album/p7.asp, acesso em 27/09/2011.

CRESPO, A. A. **Estatística Fácil**. São Paulo, Ed. Saraiva. 218p. 2009.

TICs

G

Modelos teóricos de distribuição de probabilidade

Unidade G
Estatística Básica

UNIDADE G

MODELOS TEÓRICOS DE DISTRIBUIÇÃO DE PROBABILIDADE

Objetivos

- Conceituar distribuição de probabilidade;
- Identificar as distribuições de probabilidade binomial e normal;
- Calcular probabilidades em distribuições binomial e normal.

Você verá por aqui...

Após os conceitos de probabilidade que vimos na aula anterior, vamos agora conceituar e identificar duas distribuições de probabilidade bastante comuns em eventos, quais sejam, a distribuição binomial e a distribuição normal.

Identificar a situação onde aplicamos uma ou outra distribuição e a metodologia para o cálculo de probabilidades serão trabalhados nessa aula.

Procure organizar seus horários disponibilizando um bom tempo para as atividades e aproveite bem este material.

Distribuição de probabilidade

Para definirmos uma distribuição de probabilidade precisamos antes definir variável aleatória.

Variável aleatória

Uma variável aleatória é uma função que associa um determinado valor numérico a cada ponto do espaço amostral de um experimento.

Assim, considerando o espaço amostral (S) relativo ao lançamento simultâneo de duas moedas, teremos (utilizamos ca= cara e co= coroa):

S = {(ca, co), (co, ca), (co, co), (ca, ca)}

Tomando X para representar o número obtido de “caras” nos dois lançamentos dos dados, podemos associar um número para X a cada ponto amostral, de acordo com a Tabela 1 a seguir:

Ponto amostral	X (número de caras)
(ca, co)	1
(co, ca)	1
(co, co)	0
(ca, ca)	2

Tabela 1- Número de “caras” no lançamento de dois dados.

A cada valor X(i) correspondem pontos do espaço amostral (S) e, para cada valor X(i) está associada uma probabilidade p(i) da ocorrência de tais pontos no espaço amostral.

Assim: $\sum p(i) = 1$

E os valores x_1, x_2, ..., xn e seus correspondentes p_1, p_2, ..., p_n definem uma distribuição de probabilidade.

Ao ser definida uma distribuição de probabilidade, estabelecemos uma correspondência unívoca entre os valores da variável aleatória **X** e os valores da variável **P**. Esta correspondência define uma **função** onde os valores X_i formam o **domínio da função** e os valores P_i formam o **conjunto imagem** da função.

Essa função é denominada **função de probabilidade** e representada por:

$$f(x) = P(X=x_i)$$

A função $P(X=x_i)$ determina a **distribuição de probabilidade** da **variável aleatória X**.

Distribuição binomial

Um experimento binomial tem as seguintes características:

- As n tentativas de um mesmo experimento são independentes, ou seja, o resultado de uma tentativa não afeta os resultados das sucessivas;
- Cada tentativa admite apenas dois resultados: sucesso ou fracasso, acertar ou errar, cara ou coroa, entre outros;
- No decorrer do experimento, a probabilidade p do sucesso e a probabilidade q do fracasso (q= 1- p) permanecem constantes.

Logo, se em **n** tentativas **independentes** de um experimento em que o resultado esperado só poderá ser **p ou q** (sucesso ou fracasso, respectivamente), e a probabilidade **p** for constante em todo o experimento, então a probabilidade de a variável aleatória **x** ter **k** (número de sucessos) nas **n** tentativas será obtida por:

$P(x = k) = C_{n,k} \cdot p_k \cdot q^{n-k}$

Onde:

P(x = k) é a probabilidade de que o evento se realize k vezes em n provas;

p é a probabilidade de que o evento se realize em uma só prova (sucesso);

q é a probabilidade de que o evento não se realize em uma só prova (fracasso);

$C_{n,k}$ é o coeficiente binomial de **n** sobre **k**, igual a $C_{n,k} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$

n! é o fatorial de n, por exemplo, se n=3, n! = 3x2x1

Essa função, denominada **lei binomial**, define a **distribuição binomial**.

Segundo CRESPO (2009), o nome **binomial** vem do fato de que $\binom{n}{k} p^k q^{n-k}$ é o termo geral de desenvolvimento do **binômio de Newton**.

Exemplo de aplicação:

Determine a probabilidade de ocorrer exatamente duas coroas no conjunto de dez lançamentos de uma moeda.

x= número de coroas

n = 10

p = 0,5

q= 0,5

k= 2

$$P(x=2) = C_{10,2} \times 0{,}5^2 \times 0{,}5^{10-2} = \frac{10!}{2!(10-2)!} \times 0{,}5^2 \times 0{,}5^8$$

P(x = 2) = 45 x 0,25 x 0,0039

P(x = 2) =0,0439

Multiplicando o resultado por 100, obtemos a probabilidade em porcentagem:

P(x = 2) =4,39%

Distribuição normal

Entre as distribuições teóricas de variável aleatória contínua, uma das mais empregadas é a **distribuição normal**. Uma das propriedades das distribuições normais é a simetria da curva, como pode ser visto na Figura 1.

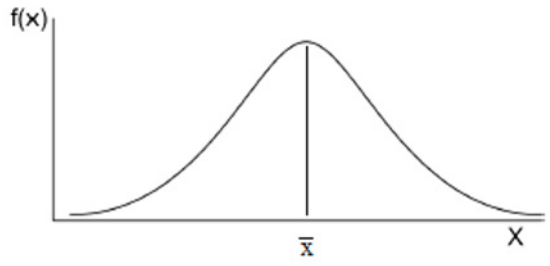


Figura G.1 - Representação de uma distribuição de frequência com curva normal.

Características da curva normal:

- A curva é simétrica em relação à média μ (x barra);
- A variável aleatória x pode assumir todo e qualquer valor real;
- A representação gráfica da distribuição normal é uma curva em forma de sino, simétrica em torno da média, que recebe o nome de curva normal ou de Gauss;
- A área total delimitada pela curva e pelo eixo das abscissas é 1 ou 100%;
- A curva normal é assintótica em relação ao eixo das abscissas, isto é, aproxima-se indefinidamente do eixo das abscissas sem, contudo, alcançá-lo;

- Como a curva é simétrica em torno da média, a probabilidade de ocorrer valor maior do que a média é igual a probabilidade de ocorrer valor menor do que a média, isto é :

$$P(X > \overline{x}) = P(X < \overline{x}) = 0{,}5$$

- Os pontos de inflexão da função são μ-σ e μ+σ:

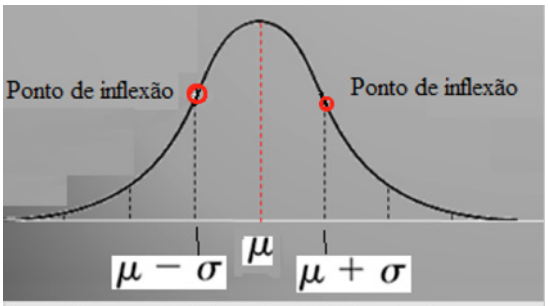


Figura G.2 - Representação dos pontos de inflexão da função curva normal.

Assim, quando trabalhamos com uma variável aleatória com distribuição normal, podemos obter a probabilidade de essa variável aleatória assumir um valor em um determinado intervalo. Para tanto, a fórmula para obter essa probabilidade é:

$$P(a \le X \le b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} \times e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Para utilizar essa função é necessária a aplicação de integração numérica, o que seria muito trabalhoso e exigiria conhecimentos que não serão tratados em nossas aulas. Entretanto, podemos contornar facilmente esse problema utilizando a variável normal padrão ou variável padronizada, **z**.

A variável padronizada z tem distribuição normal reduzida, ou seja, tem distribuição normal de média igual a zero e desvio padrão igual a um.

Sendo **x** uma variável aleatória com distribuição normal de média $\overline{x}$ e desvio padrão **s**, a variável padronizada z será dada por:

$$z = \frac{x - \overline{x}}{s}$$

As probabilidades associadas à distribuição normal padronizada não precisam ser calculadas, sendo encontradas em tabelas. A Tabela 3 apresenta a Distribuição Normal Padrão, com os valores calculados para a probabilidade **p** de valores entre a média zero e o valor $\mathbf{Z_c}$.

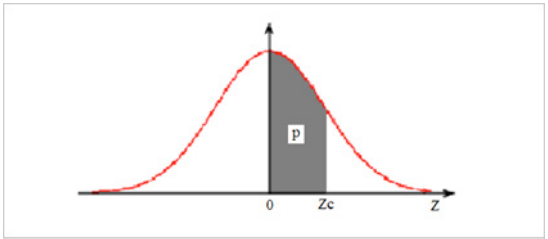


Figura G.2 - Representação da área sob a curva (probabilidade p) entre 0 e Z_C.

Z_c	Segundo decimal de Z_c									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.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

Tabela 3 - Distribuição Normal Padrão (probabilidade p, tal que $p=P(0<Z_c<Z)$

Exemplo de aplicação da tabela da curva normal padrão:

Para usar a Tabela 3, da curva normal padrão, devemos considerar o fato de que a curva é simétrica e

centrada na média. No corpo da Tabela estão os valores das probabilidades (área sob a curva entre os limites de zero e Z_c). Os valores de Z_c estão na margem esquerda e na margem superior da tabela, de tal forma que, na margem esquerda (primeira coluna) aparece o valor inteiro e a primeira casa decimal de Z_c, enquanto na primeira linha aparece o valor do segundo decimal de Z_c.

Consideremos o seguinte exemplo: Seja X a variável aleatória que representa o comprimento de determinada peça produzida por uma máquina, a qual tem distribuição normal de média igual a 2,0 cm e desvio padrão de 0,20cm. Precisamos conhecer a probabilidade de que uma peça produzida tenha comprimento entre 2,0 cm (a média) e 2,15 cm.

$$Zc = \frac{x - \bar{x}}{s} \qquad Zc = \frac{2{,}15 - 2{,}0}{0{,}2} \qquad Zc = 0{,}75$$

Para obter o valor de Z_c=0,75 basta que, na primeira coluna, localizemos o valor de 0,7 e na intersecção da linha que contém o valor de 0,7 com a coluna que contém o valor 0,05 encontraremos o valor de p= 0,2734 que corresponde a Z_c=0,75, conforme ilustrado abaixo.

Z_c	Segundo decimal de Zc									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.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

O valor 0,2734 corresponde a probabilidade de que uma peça produzida tenha dimensão entre a média (2,0cm) e 2,15cm, ou seja, 27,34% de probabilidade.

Observe que no exemplo apresentado, o valor do limite inferior de medida do parafuso coincide com o valor da média e, portanto, o valor de Zc para essa medida é zero, o que resulta em 0,2734 – 0 = 0,2734.

Se os valores procurados fossem entre 2,05cm e 2,15cm, teríamos que calcular o valor de Zc para 2,05cm, que resultaria em 0,25. Entrando com o valor de 0,25 na tabela iremos obter o valor de 0,097 para a probabilidade p. Assim, para essa faixa de valores, a probabilidade seria:

$P(2{,}05 \leq X \leq 2{,}15) = 0{,}2734 - 0{,}097 \rightarrow 0{,}1764 \; ou \; 17{,}64\%.$

Resumo

Nessa aula estudamos duas distribuições de probabilidade para variáveis aleatórias, a distribuição binomial e a distribuição normal, definindo o campo de aplicação de cada uma delas e as maneiras de realizar o cálculo de probabilidades através do conhecimento do modelo da distribuição da variável.

Exercícios

1. Uma moeda é lançada cinco vezes seguidas de forma independente. Calcule a probabilidade de serem obtidas três coroas nos cinco lançamentos.

2. Em uma industria, 40% dos funcionários possuem veículo próprio. Calcule a probabilidade de que, ao escolhermos aleatoriamente 12 funcionários, três ou menos possuam veículo próprio.

3. Os salários semanais dos operários de uma empresa são distribuídos normalmente em torno da média de R$ 500,00, com desvio padrão de R$ 40,00. Calcule a probabilidade de um operário ter um salário semanal situado entre R$ 490,00 e R$ 520,00.

4. A duração de um certo tipo de lâmpada é de, em média 850 acionamentos e desvio padrão de 40 acionamentos. Sabendo que a distribuição é normalmente distribuída, calcule a probabilidade de uma lâmpada durar:

a) entre 700 e 1.000 acionamentos;
b) mais de 800 acionamentos;
c) menos de 750 acionamentos.

Referências

CRESPO, A. A. **Estatística Fácil**. São Paulo, Ed. Saraiva. 218p. 2009.

Probabilidade - Distribuição de Poisson e Exponencial

Unidade H
Estatística Básica

UNIDADE H

PROBABILIDADE - DISTRIBUIÇÃO DE POISSON E EXPONENCIAL

Objetivos

- Identificar as distribuições de probabilidade de Poisson e exponencial;
- Calcular probabilidades em distribuição de Poisson e exponencial.

Você verá por aqui...

Nesta aula vamos conceituar e identificar outras duas importantes distribuições de probabilidade, quais sejam, a distribuição exponencial e a distribuição de Poisson.

Identificar a situação onde aplicamos uma ou outra distribuição e a metodologia para o cálculo de probabilidades será trabalhado nessa aula.

Procure organizar seus horários disponibilizando um bom tempo para as atividades e aproveite bem este material.

Distribuição de Poisson

A distribuição de Poisson é uma distribuição descontínua de probabilidade envolvendo dados que podem ser contados, como o número de ocorrências por unidade num intervalo de tempo, de área ou de distância.

É utilizada no cálculo da probabilidade do número de ocorrências de um determinado evento, em um intervalo contínuo de tempo ou espaço. São exemplos de distribuição de Poisson eventos como o número de chamadas telefônicas recebidas em uma delegacia num determinado tempo, quantidade de defeitos por metro em fios produzidos, entre outros.

Na distribuição de Poisson, a unidade de medida (tempo ou espaço) é contínua, mas a variável aleatória (número de ocorrências) é discreta. Portanto, não podemos efetuar a contagem da não ocorrência, ou seja, não podemos estimar o número de ligações telefônicas que deixaram de ser feitas para a delegacia em um determinado tempo.

Os possíveis valores que a variável aleatória x pode assumir na distribuição de Poisson são 1, 1, 2, 3...., sem limite superior.

Características da Distribuição de Poisson

- As ocorrências são independentes e aleatórias;
- A variável aleatória x é o número de ocorrências de um determinado evento ao longo de um intervalo de tempo ou espaço;

A função de probabilidade da variável x será dada pela relação:

$$P(x=k)=\frac{e^{-\mu}\times\mu^{k}}{k!}$$

Onde:

μ=λ.t

λ = coeficiente de proporcionalidade

t= tempo ou espaço

Exercício resolvido

O número de pedidos por telefone recebidos em uma pizzaria aos sábados a noite ocorrem a uma taxa de 8 pedidos por hora. Calcule:

a) quantos pedidos são esperados num período de quinze minutos;

x= número de pedidos

λ= 8 pedidos / 1 hora

t= 15 minutos

$\mu=\lambda.t \rightarrow \mu=8/60.15 \rightarrow \mu=2 pedidos$

Logo, são esperados 2 pedidos num período de quinze minutos.

b) qual a probabilidade de nenhum pedido ser solicitado em um intervalo de quinze minutos;

$$P(x=0)=\frac{e^{-2}\times 2^{0}}{0!}$$

P(x=0) = 0,1353

Logo, a probabilidade da pizzaria não receber nenhum pedido em um período de quinze minutos é 0,1353 ou 13, 53%.

c) qual a probabilidade de ocorrer pelo menos dois pedidos no período de quinze minutos;

Pelo menos dois pedidos é o mesmo que no mínimo dois pedidos, ou seja, dois, três ou n pedidos:

$P(x\geq 2)=P(x=2)+P(x=3)+P(x=4)+.....P(x=n)$

Como a distribuição de Poisson não possui limite superior, será impossível calcular por essa maneira. Assim, o cálculo pode ser pelo complementar, ou seja:

$P(x=0)+P(x=1)+P(x\geq 2)=1$

$Logo,\ P(x\geq 2)=1-[P(x=0)+P(x=1)]$

P(x=0)= 0,1353 (já calculado no item b.

$$P(x=1)=\frac{e^{-2}\times 2^{1}}{1!}$$

P(x=1) = 0,2706

E assim:

$P(x \geq 2)=1 - (0,1353+0,2706) \rightarrow P(x \geq 2)=0,5941$

A probabilidade da pizzaria receber ao menos dois pedidos num período de quinze minutos é de 0,5941 ou 59,41%.

d) qual a probabilidade de ocorrer exatamente dois pedidos em vinte minutos.

Será necessário recalcular a média, pois o período mudou de 15 minutos para 20 minutos.

$\mu=\lambda.t \rightarrow \mu=8/60.20 \rightarrow \mu=2,67 pedidos$

$$P(x = 2) = \frac{e^{-2,67} \times 2,67^2}{2!}$$

P(x=2) = 0,2468

A probabilidade da pizzaria receber exatamente dois pedidos em um período de vinte minutos é de 24,68%.

Distribuição exponencial

A distribuição exponencial envolve probabilidades ao longo do tempo ou da distância entre ocorrências num intervalo contínuo. Assim, a função de distribuição de probabilidade exponencial é usada como modelo do tempo entre falhas de equipamento elétrico, tempo entre a chegada de clientes em um supermercado, entre outras (STEVENSON, 2001).

A relação entre a distribuição exponencial e a de Poisson é que, enquanto na distribuição de Poisson podemos calcular o numero de ocorrências em um determinado tempo ou espaço, na distribuição exponencial estimamos o tempo ou espaço entre uma ocorrência e outra. Assim, na distribuição de Poisson estimamos a ocorrência da variável aleatória discreta e na distribuição exponencial a variável aleatória contínua.

Se um processo com distribuição de Poisson tem média de λ ocorrências durante um intervalo (de tempo ou espaço), então o espaço entre as ocorrências naquele intervalo será de 1/λ. Exemplificando, se as chamadas telefônicas ocorrem em média à razão de 6 por hora, então o tempo médio entre as chamadas será de 10 minutos.

As probabilidades exponenciais são expressas em termos de tempo ou distância entre ocorrências.

Para calcular a probabilidade de que o tempo ou espaço antes da primeira ocorrência seja maior que um determinado tempo ou espaço empregamos a seguinte fórmula:

$$P(T > t) = e^{-\lambda t}$$

Para calcular a probabilidade de que uma ocorrência ocorra em um intervalo igual a t ou antes de t,

usamos:

$$P(T \leq t) = 1 - e^{-\lambda t}$$

Exercício resolvido

Sabendo que o tempo médio de atendimento do pedido em uma pizzaria seja de 10 minutos e que esse tempo tenha distribuição exponencial, calcule:

a) a probabilidade de um cliente esperar mais do que 10 minutos;

$\lambda = 1 / 10 \rightarrow \lambda = 0{,}1$ por minuto

$$P(T > t) = e^{-\lambda t}$$

$$P(T > 10) = e^{-0{,}1(10)}$$

$$P(T > 10) = e^{-1}$$

P(T>10) = 0,368 ou 36,8%

b) a probabilidade de o pedido ser atendido em menos de 10 minutos;

$$P(T \leq t) = 1 - e^{-\lambda t}$$

$$P(T \leq 10) = 1 - e^{-1}$$

P(T≤10)= 1 – 0,368 → = 0,632 ou 63,2%

c) a probabilidade de o pedido ser atendido em no máximo 3 minutos.

$$P(T \leq 3) = 1 - e^{-0{,}1(3)}$$

$$P(T \leq 3) = 1 - e^{-0{,}3}$$

P(T≤3)= 1 – 0,741 → = 0,259 ou 25,9%

Resumo

Nessa aula estudamos duas distribuições de probabilidade, a distribuição de Poisson, descontínua, e a exponencial, contínua. Distribuições descontínuas de probabilidade envolvem variáveis aleatórias relativas a dados que podem ser contados, como o número de ocorrências, enquanto as distribuições contínuas de probabilidade apresentam um grande número de resultados possíveis incluindo valores inteiros e não-inteiros.

A distribuição de Poisson é útil para descrever as probabilidades do número de ocorrências num campo ou intervalo contínuo, geralmente tempo ou espaço.

A distribuição de probabilidade exponencial é útil para a determinação de probabilidades de tempo ou espaço entre ocorrências, quando a taxa de ocorrências tem distribuição de Poisson.

Exercícios

1. Os acidentes em uma empresa de construção civil têm aproximadamente a distribuição de Poisson, com média de 3 acidentes por mês. Determine a probabilidade de que em um determinado mês ocorram:

a) zero acidente;
b) 1 acidente;
c) 3 ou 4 acidentes.

2. Em um ambulatório de pronto atendimento são realizados 2,8 atendimentos por hora. Determine a probabilidade de que sejam atendidos 3 ou mais pacientes em:

a) um período de 30 minutos;
b) um período de 1 hora;
c) um período de 2 horas.

3. O tempo de atendimento em uma oficina é bem aproximado por uma distribuição exponencial com média de 4 minutos. Determine a probabilidade de:

a) o tempo de espera seja superior a 4 minutos;
b) o tempo de espera seja inferior a 4 minutos;
c) o tempo de espera seja de exatamente 4 minutos.

4. Determine a probabilidade de um determinado tipo de lâmpada de iluminação operar durante pelo menos 20.000 horas antes de apresentar uma falha, se o tempo médio entre falhas (1/λ) é:

a) 10.000
b) 20.000
c) 40.000

Referências

STEVENSON, W. J. Estatística aplicada à administração. Trad. Alfredo Alves de Farias. São Paulos, Ed. Harbra Ltda. 495p. 2001.

Inferência estatística: teoria da amostragem e teoria da estimação

Unidade I
Estatística Básica

UNIDADE

INFERÊNCIA ESTATÍSTICA: TEORIA DA AMOSTRAGEM E TEORIA DA ESTIMAÇÃO

Objetivos

- Conceituar os diferentes métodos de amostragem;
- Identificar o melhor método e tipo de amostragem em função da análise a ser realizada;
- Caracterizar os tipos de amostragem.

Introdução

A inferência estatística baseia-se na análise formal e técnica de uma amostra para formular julgamentos e decisões sobre o todo de onde foi retirada a amostra. A amostragem estatística envolve métodos formais e precisos, incluindo tipicamente uma afirmação probabilística. Assim, a probabilidade e a amostragem são estreitamente relacionadas e, juntas, formam o fundamento da teoria da inferência estatística (STEVENSON, 2001).

Na inferência estatística, a amostragem aleatória é de grande importância, uma vez que permite estimar o valor do erro possível, isto é, estabelecer a maior ou menor proximidade da amostra em relação à população, quanto a sua representatividade, o que não é possível pela amostragem não aleatória.

População e amostra

Na primeira aula do curso definimos o conceito de universo estatístico, o qual é constituído de todos os elementos comuns que compõem uma população. Assim, População é o conjunto constituído de elementos, de um mesmo universo, que apresentam pelo menos uma característica comum. A população, segundo o seu tamanho, pode ser finita ou infinita. É finita quando possui um número determinado de elementos e infinita quando possui um número infinito de elementos. Contudo tal definição existe apenas no campo teórico, uma vez que na prática, nunca encontraremos populações com infinitos elementos e sim com grande número de componentes e, tais populações são tratadas como infinitas.

Na maioria das vezes, devido ao alto custo, ao intenso trabalho ou ao tempo necessário, limitamos as observações referentes a uma determinada investigação ou pesquisa a apenas uma parte da população, a qual denominamos de amostra. Amostra é um subconjunto finito da população, escolhido sob diferentes tipos de amostragem, como veremos a seguir.

Censo e amostragem

Censo envolve a contagem ou análise de todos os elementos de uma determinada população, enquanto que a amostragem é o estudo de apenas uma parte da população. Para fazer a escolha entre censo e amostragem, devem ser levados em consideração diversos fatores: custo, facilidade de acesso aos

elementos da população, nível de precisão desejado e tempo para a realização do estudo estatístico.

A amostragem tem por finalidade fazer generalizações sobre todo o grupo sem precisar examinar cada um de seus elementos (STEVENSON, 2001).

Amostragem

Na maioria dos problemas de inferência estatística é impossível observar todos os elementos da população e assim, torna-se necessário analisar uma amostra da população. Para que as inferências assumidas a partir da amostra sejam válidas, é necessário que a amostra seja o mais representativa possível da população alvo.

Uma amostragem pode ser probabilística ou não probabilística. Para evitar o vício de seleção e garantir a representatividade, deve-se dispor da aleatoriedade na seleção de uma amostra, ou seja, utilizar um tipo de amostragem probabilística. Nem sempre isso é possível, nesse caso, usa-se um tipo de amostragem não probabilística, de acordo com as características do estudo envolvido.

A amostragem probabilística envolve todas as técnicas que usam métodos aleatórios na seleção dos elementos da amostra, atribuindo a cada um deles uma probabilidade de pertencer à amostra.

Sob o enfoque estatístico, a maneira correta de escolher uma amostra para tirar conclusões válidas sobre a população de estudo, a partir dos resultados obtidos na amostragem, é por amostragem probabilística.

Na Tabela 1 são apresentados, esquematicamente, os diversos tipos de amostragem probabilística e não probabilística.

Tipos de Amostragem	
Não Probabilística	**Probabilística**
Acidental ou conveni□ncia	Aleatória Simples
Intencional ou por julgamento	Aleatória Estratificada
Quotas ou proporcional	Conglomerado
	Sistemática

Tabela 1 – Tipos de amostragem.

Para a escolha do método de amostragem devemos considerar o tipo de pesquisa que estamos realizando, o acesso e a disponibilidade dos elementos da população, a variabilidade da população, disponibilidade de tempo, recursos financeiros e humanos disponíveis, entre outros fatores.

Amostragem não probabilística

Acidental ou conveniência

Nesse tipo de amostragem, os elementos que compõem a amostra são selecionados com base em sua semelhança presumida com a população e na sua disponibilidade imediata. Geralmente utilizada em pesquisas de opinião, em que os entrevistados são acidentalmente escolhidos. Tem a vantagem de ser rápida e de baixo custo, pela fácil seleção da amostra e coleta dos dados, entretanto é difícil avaliar sua representatividade em relação a população.

Ex: Pesquisas de opinião em praças públicas, ruas de grandes cidades, estabelecimentos comerciais.

Intencional ou por julgamento

De acordo com determinado critério, é escolhido intencionalmente um grupo de elementos que irão compor a amostra. O investigador se dirige intencionalmente a grupos de elementos dos quais deseja saber a opinião. É um modo relativamente fácil de selecionar uma amostra, no entanto, a qualidade dos resultados da amostra depende do julgamento de quem faz a seleção.

Ex: Numa pesquisa sobre preferência por determinado cosmético, o pesquisador se dirige a um grande salão de beleza e entrevista as pessoas que ali se encontram.

Quotas ou proporcional

Um dos métodos de amostragem mais comumente usados em levantamentos de mercado e em prévias eleitorais. Ele abrange três fases:

1ª - classificação da população em termos de propriedades que se sabe, ou presume, serem relevantes para a característica a ser estudada;

2ª - determinação da proporção da população para cada característica, com base na constituição conhecida, presumida ou estimada, da população;

3ª - fixação de quotas para cada entrevistador a quem tocará a responsabilidade de selecionar entrevistados, de modo que a amostra total observada ou entrevistada contenha a proporção e cada classe tal como determinada na 2ª fase.

Ex: Numa pesquisa sobre o "trabalho das mulheres na atualidade", provavelmente se terá interesse em considerar: a divisão cidade e campo, a habitação, o número de filhos, a idade dos filhos, a renda média, as faixas etárias etc.

Amostragem probabilística

Exige que cada elemento da população possua determinada probabilidade de ser selecionado. Normalmente possuem a mesma probabilidade. Assim, se N for o tamanho da população, a probabilidade de cada elemento ser selecionado será 1/N. Trata-se do método que garante cientificamente a aplicação das técnicas estatísticas de inferências. Somente com base em amostragens probabilísticas é que se podem realizar inferências ou induções sobre a população a partir do conhecimento da amostra.

Aleatória simples

É o processo mais elementar e freqüentemente utilizado, onde todos os elementos da população têm a mesma probabilidade de compor a amostra. É equivalente a um sorteio lotérico. Pode ser realizada numerando-se a população de 1 a n e sorteando-se, a seguir, por meio de um dispositivo aleatório qualquer, x números dessa seqüência, os quais corresponderão aos elementos pertencentes à amostra.

Ex: Vamos obter uma amostra, de 20%, representativa para a pesquisa da estatura de 90 alunos de uma escola:

1º - numeramos os alunos de 1 a 90.

2º - escrevemos os números dos alunos, de 1 a 90, em pedaços iguais de papel, colocamos na urna e após mistura retiramos, um a um, nove números que formarão a amostra.

OBS: quando o número de elementos da amostra é muito grande, esse tipo de sorteio torna-se muito trabalhoso. Neste caso utiliza-se uma Tabela de números aleatórios, construída de modo que os algarismos de 0 a 9 são distribuídos ao acaso nas linhas e colunas.

Aleatória Estratificada

Quando a população se divide em estratos (sub-populações), convém que o sorteio dos elementos da amostra leve em consideração tais estratos, daí obtemos os elementos da amostra proporcional ao número de elementos desses estratos.

Ex: Vamos obter uma amostra proporcional estratificada, de 10%, do exemplo anterior, supondo, que, dos 90 alunos, 54 sejam meninos e 36 sejam meninas. São, portanto dois estratos (sexo masculino e sexo feminino). Logo, temos:

SEXO	POPULAÇÃO	10%	AMOSTRA
MASCULINO	54	5,4	5
FEMININO	36	3,6	4
Total	90	9,0	9

Numeramos então os alunos de 01 a 90, sendo 01 a 54 meninos e 55 a 90, meninas e procedemos o sorteio casual com urna ou tabela de números aleatórios.

Conglomerado

Algumas populações não permitem, ou tornam extremamente difícil que se identifiquem seus elementos. Não obstante isso, pode ser relativamente fácil identificar alguns subgrupos da população. Em tais casos, uma amostra aleatória simples desses subgrupos (conglomerados) pode se colhida, e uma contagem completa deve ser feita para o conglomerado sorteado. Agrupamentos típicos são quarteirões, famílias, organizações, agências, edifícios etc.

Ex: Num levantamento da população de determinada cidade, podemos dispor do mapa indicando cada quarteirão e não dispor de uma relação atualizada dos seus moradores. Pode-se, então, colher uma amostra dos quarteirões e fazer a contagem completa de todos os que residem naqueles quarteirões sorteados.

Amostragem sistemática

Quando os elementos da população já se acham ordenados, não há necessidade de construir o sistema de referência. São exemplos os prontuários médicos de um hospital, os prédios de uma rua, etc. Nestes casos, a seleção dos elementos que constituirão a amostra pode ser feita por um sistema imposto pelo pesquisador.

Ex: Suponhamos uma rua com 900 casas, das quais desejamos obter uma amostra formada por 50 casas para uma pesquisa de opinião. Podemos, neste caso, usar o seguinte procedimento: como 900/50 = 18, escolhemos por sorteio casual um número de 01 a 18, o qual indicaria o primeiro elemento sorteado para a amostra; os demais elementos seriam periodicamente considerados de 18 em 18. Assim, suponhamos que o número sorteado fosse 4 a amostra seria: 4ª casa, 22ª casa, 40ª casa, 58ª casa, 76ª casa, etc.

Resumo

Nessa aula estudamos a diferença entre método de amostragem probabilístico e não-probabilístico e os diferentes tipos de amostragem em cada um dos métodos. Vimos que a amostragem probabilística permite maior representatividade da população e a inferência de conclusões, a partir da amostra, para o todo ou população. A amostragem não-probabilística tem menor custo e maior rapidez, entretanto não há garantia de representatividade da população.

Exercícios

1. Qual a principal diferença entre métodos de amostragem probabilísticos e não probabilísticos?
2. Quais as características de uma amostragem aleatória simples?
3. Quando utilizamos uma amostragem por quotas?
4. O que deve ser considerado para escolher o melhor método de amostragem?
5. Qual a diferença entre censo e amostragem?

Referências

STEVENSON, W. J. Estatística aplicada à administração. Trad. Alfredo Alves de Farias. São Paulo, Ed. Harbra Ltda. 495p. 2001.

J

Distribuições amostrais

Unidade J
Estatística Básica

UNIDADE 4

DISTRIBUIÇÕES AMOSTRAIS

Objetivos

- Conceituar distribuição amostral;
- Explicar como os parâmetros populacionais infuenciam as estatísticas amostrais;
- Explicar o Teorema do Limite Central e sua importância;
- Descrever as consequências do tamanho amostral.

Começando...

Nesta aula iremos estudar alguns exemplos de inferência estatística analisando os valores médios, os quais são uma característica específica da população. Em nosso estudo vamos considerar a média amostral, que é um descritor particular da amostra e que formam o que denominamos de distribuições amostrais da média. Para iniciar, apresentaremos primeiramente alguns conceitos essenciais.

Distribuições amostrais

Introdução

Para atingirmos o objetivo de entender o que seja uma distribuição amostral, e a sua utilidade na inferência estatística, vamos primeiramente estabelecer algumas definições e conceituações. Consideremos que tenham sido coletadas várias amostras de uma determinada população, através de um processo estatístico de amostragem. Assim, sabemos que:

• **Parâmetro**: é toda medida única, descritiva e numérica de uma população;

• **Estatística**: é o valor obtido através de cálculo, a partir de observações de uma amostra;

• Os valores obtidos de diversas médias amostrais, obtidas a partir de uma população, não são necessariamente iguais entre si, e não são necessariamente iguais ao valor da média da população;

• O conjunto das médias amostrais forma uma série de médias sobre a qual podemos calcular uma média e um desvio padrão;

• A variabilidade nos valores das diversas médias amostrais dá origem a uma distribuição de freqüências, que terá uma média das médias e um desvio padrão da variação das médias em torno da média (das médias);

• **Distribuição amostral** é a distribuição de frequências das médias amostrais;

• Cada média amostral, denominada estatística, é uma variável aleatória representada por ;

• **Erro padrão** é o desvio padrão da distribuição amostral, sendo representado por $\sigma_{\bar{x}}$.

• A média da população representada por μ é um parâmetro, enquanto o parâmetro desvio padrão da

população é representado por σ.

Na inferência estatística os parâmetros da população μ e σ serão considerados conhecidos. Na verdade estes parâmetros não são conhecidos, mas essa premissa é útil para o entendimento do conceito de distribuição amostral.

Exemplo 1:

Consideremos 10 grupos de entrevistadores os quais têm como tarefa calcular a média de renda familiar mensal de 50 famílias que vivem em um determinado bairro da cidade. Como todos os grupos levantam dados no mesmo bairro, ao concluírem a tarefa, estes entrevistadores terão formado uma série de 10 médias amostrais representadas como $\bar{x}_1; \bar{x}_2; \bar{x}_3; \ldots \bar{x}_{10}$.

Estas 10 médias amostrais calculadas pelos entrevistadores são denominadas de estatísticas e:

- É de se esperar que os 10 valores das médias amostrais sejam, em sua maioria, diferentes;
- Os valores das médias amostrais poderão ser diferentes do valor da média da população;
- Para as 10 médias amostrais podemos formar uma série e calcular sua média e seu desvio padrão.

Devido ao fato de o valor da média amostral ser uma variável, podemos obter uma distribuição amostral das médias, logo, os valores das médias amostrais têm sua própria distribuição de frequências.

Se outros 10 grupos de entrevistadores fizerem novas amostragens neste mesmo bairro em domicílios selecionados ao acaso, teremos novos valores de médias amostrais, em geral diferentes dos valores obtidos pelos 10 grupos anteriores.

Cada média amostral é uma estatística, sendo também uma variável aleatória que possui uma distribuição de frequências, com um valor próprio de média e de desvio padrão.

Teorema do Limite Central

À medida que o tamanho da amostra aumenta, a distribuição de frequências das médias amostrais tende a se aproximar cada vez mais da distribuição normal.

Se o tamanho n da amostra for suficientemente grande, a média de uma amostra aleatória retirada de uma população de dados, terá uma distribuição de frequências aproximadamente normal independentemente da população. Na prática, uma amostra é considerada suficientemente grande, se consistir de 30 ou mais observações (STEVENSON, 2001).

Já se a população tem distribuição normal, então a média amostral terá distribuição normal qualquer que seja o tamanho da amostra (Figura 1).

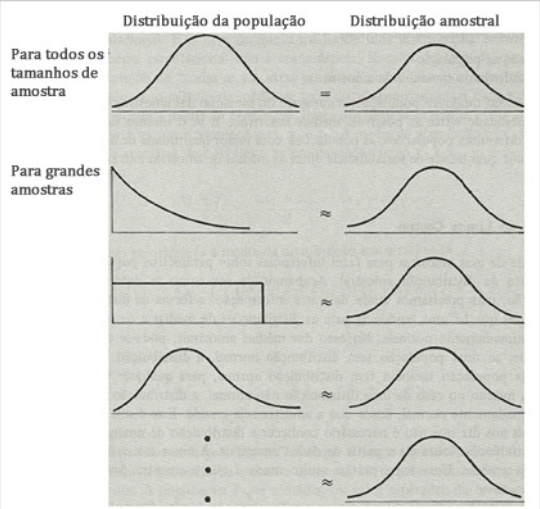


Figura J.1 - Ilustração do Teorema do Limite Central
(Fonte: Stevenson, 2001).

Pelo teorema do limite central pode-se afirmar então que a distribuição da média amostral é aproximadamente normal e que os valores da média e desvio padrão estão relacionados com os valores da média e desvio padrão da população, com a única restrição de que a amostra seja grande.

Em sentido estrito, o Teorema do Limite Central só se aplica a médias amostrais (STEVENSON, 2001).

Assim, se uma população de dados tem média μ e desvio padrão σ, da qual se retira uma amostra de tamanho n e média amostral $\overline{x}$, pode-se afirmar que:

O valor esperado das médias amostrais E [$\overline{x}$] é igual à média da população:

$$E[\ \overline{x}\ ] = \mu$$

O desvio padrão da distribuição amostral (denominado erro padrão) é igual:

$$\sigma \overline{x} = \frac{\sigma_x}{\sqrt{n}}$$

Onde:

$\sigma\ \overline{x}$ = desvio padrão da distribuição amostral

σ_x = desvio padrão da população

n = tamanho da amostra

Exemplo 2:

Consideremos uma população formada por 5 empresas aéreas que operam em um aeroporto de uma cidade, e que apresentam os seguintes números de vôos diários:

Empresa	A	*B*	*C*	*D*	*E*
Nº de vôos	2	4	6	8	10

Se pretendermos selecionar aleatoriamente uma amostra formada por duas empresas para avaliar o número médio de vôos da cidade, vamos ter um universo de 10 prováveis combinações.

Cada uma das empresas terá a mesma probabilidade de ser selecionada, e dependendo das empresas amostradas o número de vôos médio amostral pode ficar acima ou abaixo da média de vôos da população. Vamos então definir o espaço amostral e determinar o valor esperado das médias amostrais $\overline{x}$ de tamanho n = 2 retiradas da população:

Solução:

• A média de vôos da população formada pelas 5 empresas é igual a 6 vôos diários;

• Cada uma das 5 empresas aéreas tem probabilidade igual a 20% de ser sorteada.

• Espaço amostral:

Amostra	A -B	*A -C*	*A - D*	*A - E*	*B - C*	*B - D*	*B - E*	*C - D*	*C - E*	*D - E*
Média $\overline{x}$	3	4	5	6	5	6	7	7	8	9

• Distribuição de Frequências das médias amostrais

Média $\overline{x}$	3	*4*	*5*	*6*	*7*	*8*	*9*
Frequência	10	10	20	20	20	10	10

• Valor Esperado das Médias Amostrais

$E[\overline{x}] = 3 \times 0{,}1 + 4 \times 0{,}1 + 5 \times 0{,}2 + 6 \times 0{,}2 + 7 \times 0{,}2 + 8 \times 0{,}1 + 9 \times 0{,}1 = 6 = \mu$

• Distribuição amostral

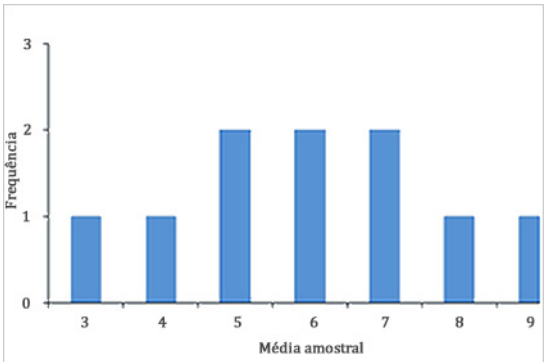


Figura J.2 - Distribuição amostral da média do número de vôos em amostras com n=2.

As distribuições amostrais tendem a produzir estatísticas amostrais representativas dos parâmetros populacionais. Apesar do fato de tenderem a apresentar certa variabilidade, podemos dizer que as estatísticas amostrais devem aproximar parâmetros populacionais de forma bastante satisfatória. Assim, esta característica de ser representativa resulta em estatísticas amostrais que tendem a se acumular na vizinhança dos verdadeiros parâmetros populacionais (STEVENSON, 2001). Podemos ver que a Figura 1 apresenta a distribuição amostral da média (amostral), a qual apresenta distribuição normal e com

os valores distribuídos ao redor da média amostral ($\overline{x}$ = 6,0) que coincide com o valor da média populacional (μ = 6,0).

Efeito do tamanho da amostra sobre a distribuição amostral

Na medida em que o tamanho da amostra aumenta, a distribuição amostral tende para a normalidade e a variabilidade amostral decresce, ou seja, em amostragens maiores as médias amostrais tendem a agrupar-se em torno da média populacional, o que é desejável quando se quer estimar a média da população a partir da média amostral, e a variabilidade é menor quando o tamanho da amostra é grande. A Figura 2 mostra esse efeito graficamente.

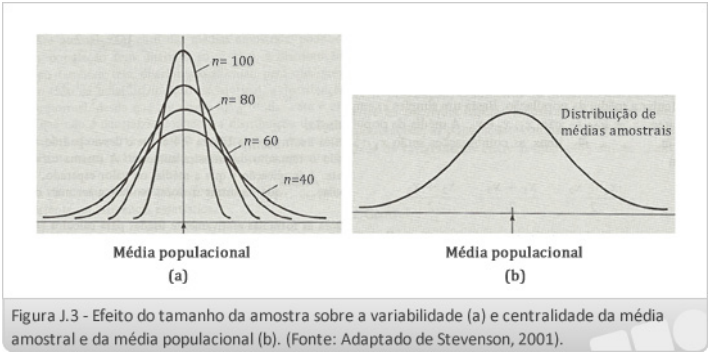


Figura J.3 - Efeito do tamanho da amostra sobre a variabilidade (a) e centralidade da média amostral e da média populacional (b). (Fonte: Adaptado de Stevenson, 2001).

Resumo

As distribuições amostrais são distribuições de probabilidade para estatísticas amostrais, e assim, indicam a probabilidade que possuem os diversos valores possíveis de uma estatística amostral. A finalidade da amostragem é permitir conhecer algo sobre a população sem precisar examinar a sua totalidade, sendo que as distribuições amostrais são a base para isto. A média de uma distribuição amostral e, por consequência, a média esperada da amostra, é igual a média da população, e os valores amostrais que têm maior probabilidade são os que estão mais próximos do verdadeiro valor populacional. Quanto maior a amostra, menor a dispersão entre os valores possíveis da amostra. O Teorema do Limite Central afirma que grandes amostras tendem a produzir distribuições amostrais aproximadamente normais, mesmo quando a população estudada não é normal, enquanto que para populações com distribuição normal as distribuições amostrais apresentarão normalidade, independente do tamanho da amostra.

Exercícios

1. Apresente uma definição para “distribuição amostral”.

2. Qual a relação entre o tamanho da amostra e a variabilidade de uma distribuição amostral de médias?

3. Explique porque repetidas amostras retiradas de uma mesma população tendem a variar entre si.

4. À vista do Teorema do Limite Central, quando é necessário saber se uma população investigada tem distribuição normal?

5. Uma população considerada grande, tem média 10,0 e desvio padrão 0,7. Extraímos uma amostra de 40 observações. Responda:

a) Qual a média da distribuição amostral?
b) Qual o desvio padrão da distribuição amostral?
[Respostas: a) 10,0 b) 0,11]

Referências

STEVENSON, W. J. **Estatística aplicada à administração**. Trad. Alfredo Alves de Farias. São Paulo, Ed. Harbra Ltda. 495p. 2001.

Estimação e Intervalos de Confiança

Unidade K
Estatística Básica

UNIDADE K

ESTIMAÇÃO E INTERVALOS DE CONFIANÇA

Objetivos

- Conceituar estimação e estimadores,
- Descrever e comparar estimativas pontuais e intervalares,
- Conceituar intervalo de confiança,
- Construir intervalos de confiança.

Você verá por aqui...

Nesta aula começaremos a trabalhar assuntos relacionados à inferência estatística e, para tanto, iremos definir e diferenciar o que é uma estimativa e o que são os estimadores. Vamosestudar o que é intervalo de confiança, como construímos um intervalo e qual a margem de erro que está associada em cada estimativa. Os assuntos estão organizados em uma sequência que permitirá que você vá se apropriando dos conhecimentos e, através dos exemplos, consiga entender e resolver os exercícios propostos ao final da aula.

Procure organizar seus horários disponibilizando um bom tempo para as atividades e aproveite bem este material.

Estimação, estimador e estimativa

Na inferência estatística, **estimação** consiste no processo de usar os dados de uma amostra (dados amostrais) para estimar valores de parâmetros populacionais desconhecidos tais como a média e o desvio padrão de uma população e a proporção populacional.

Estimador é um determinado variável estatística, calculado em função dos elementos da amostra e caracterizado por uma distribuição de probabilidade, que será utilizado no processo de estimação do parâmetro desejado.

Estimativa corresponde a cada um dos valores particulares assumido pelo estimador. Assim, as estatísticas amostrais são utilizadas como estimadores de parâmetros populacionais. Dessa maneira, uma média amostral é utilizada como estimativa de uma média populacional; um desvio padrão amostral serve como estimativa do desvio padrão da população, e a proporção de elementos com determinada característica comum, em uma amostra, serve para estimar a proporção da população que apresenta tal característica (STEVENSON, 2001).

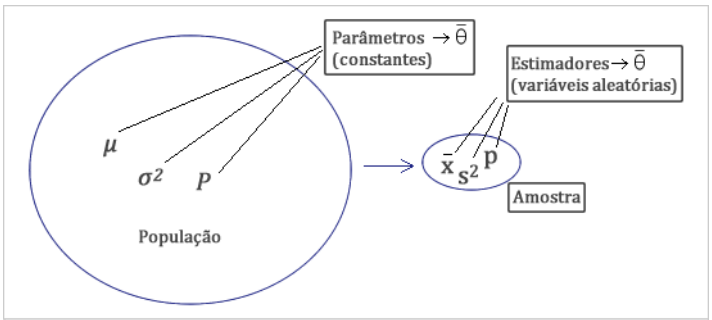


Figura K.1 -Estimativa de parâmetros populacionais a partir da amostra

A estimação pode ser realizada por processo **pontual** ou **intervalar**. Nos dois casos vamos observar n elementos extraídos (amostra) da população e para cada elemento vamos verificar o sucesso ou fracasso na presença da característica buscada.

Estimação Pontual

Ocorre quando é realizada uma única estimativa (um único valor) para um determinado parâmetro populacional.

Exemplo: usamos a média amostral para estimar a média populacional ou o desvio padrão

amostral para estimar o desvio padrão da população.

O estimador pontual p, também conhecido por proporção amostral, é definido como:

$$p = \frac{X}{n}$$

Onde:

X é o número de elementos da amostra que apresenta a característica desejada;

n é o tamanho da amostra coletada.

Suponhamos que em pesquisa realizada com 1.000 pessoas, 160 delas responderam positivamente sobre certo modelo de carro. Assim:

$$p = \frac{X}{n} = \frac{160}{1000} = 0{,}16$$

Entendemos então que 16% da população iriam escolher tal modelo de carro.

Estimação Intervalar

Ocorre quando fizemos uma estimativa de um intervalo de possíveis valores no qual se admite esteja o parâmetro populacional.

Por exemplo, a partir de uma média amostral igual a 25,0, podemos inferir que a média populacional esteja entre 24,0 e 26,0.

Nesse tipo de estimativa temos um intervalo de valores em torno do parâmetro amostral, no qual julgamos, com um risco conhecido de erro (e), estar o parâmetro (p) da população.

A este intervalo chamamos de **intervalo de confiança**, o qual está compreendido entre os valores do parâmetro menos o erro e do parâmetro mais o erro: [p-e; p+e].

Logo em seguida voltaremos a conceituar e aprofundar o conceito de intervalo de confiança.

Propriedades dos estimadores

Na maioria dos casos, estimar um determinado parâmetro populacional é tarefa relativamente simples, entretanto, a qualidade da estimativa irá depender muito da conveniente escolha que fizermos do estimador.

A escolha conveniente de um estimador deve ser feita buscando satisfazer às principais propriedades de um bom estimador, ou seja, **justeza, consistência e eficiência**.

Justeza

Um estimador T é dito justo quando a sua média (ou valor esperado E) for igual ao parâmetro α que pretendemos estimar, ou seja:

$$E[T] = \alpha$$

Consistência

Definimos como estimador consistente aquele que, na medida em que a amostra cresce, seu valor aproxima-se do valor do verdadeiro parâmetro que estamos estimando. Considerando que o estimador seja justo, a condição de consistência é equivalente a dizermos que a sua variância tende a zero quando o tamanho da amostra tende ao infinito, ou seja:

$$\lim n \to \infty \ \sigma 2(T) = 0$$

Eficiência

Definimos um estimador T como o estimador mais eficiente do parâmetro α se ele for justo e sua variância, para um mesmo tamanho da amostra, for menor que a de qualquer outro estimador justo. Assim, a eficiência é uma avaliação de quão próximo as estimativas individuais de T estão do parâmetro α.

Intervalos de confiança

O intervalo de confiança é uma estimativa intervalar que inclui uma afirmação probabilística. Essa informação indica a percentagem de intervalos que podemos esperar abranger o verdadeiro valor do parâmetro em seus limites. A amplitude de um intervalo de confiança depende de quatro itens:

- a dispersão dos valores populacionais,
- o nível de confiança indicado,
- o erro tolerável e,
- o tamanho da amostra.

Vamos recorrer aos conceitos estudados na Aula 7, Modelos teóricos de distribuição de probabilidade, onde conceituamos a variável padronizada Z que corresponde a um determinado valor afastado da média, na curva normal padronizada.

Uma das propriedades da curva normal padronizada é que 68% da estatística amostral está a menos de um desvio padrão de cada lado da média da distribuição amostral (a qual é igual a média da população, em se tratando de distribuição normal), 95,5% dos valores estão a dois desvios padrões em ambos os lados da média e que em torno de 0,3% da estatística amostral está a mais de três desvios padrões da média. A Figura 2 ilustra a distribuição normal padronizada com as respectivas probabilidades e desvios padrões da média.

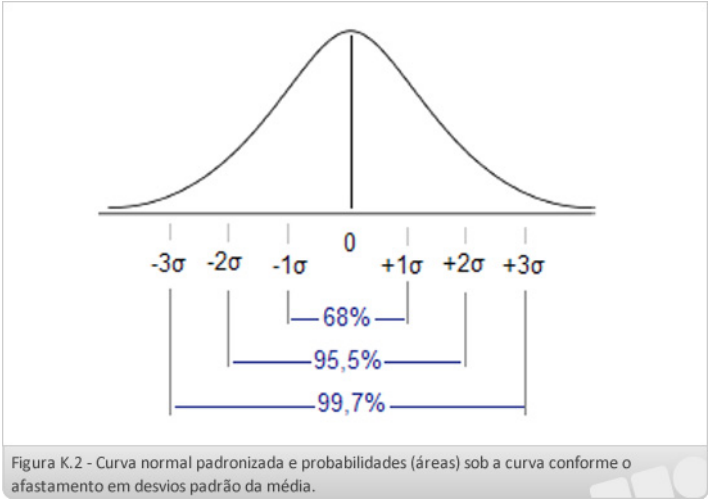


Figura K.2 - Curva normal padronizada e probabilidades (áreas) sob a curva conforme o afastamento em desvios padrão da média.

Vimos também na Aula 7, que a área sob a curva normal padronizada até um determinado valor Z não necessita de cálculo, apenas de consulta à Tabela da curva normal padrão, e que a área representa uma probabilidade.

Assim, se considerarmos que a média de uma amostra está a menos de 1,96 desvios padrões a contar da média verdadeira, podemos esperar estar certos 95% das vezes e errados 5% das vezes (consultando a Tabela da curva normal padrão, ao valor de 1,96 desvios corresponde um valor z= 0,475, logo: 0,475 x 2 = 0,95 e 1-α = 0,05).

Pode-se facilmente perceber que a média amostral poderá estar mais próxima da verdadeira média do que 1,96, ou mais afastada, sendo essa uma atribuição probabilística do intervalo em que o verdadeiro valor possa estar (STEVENSON, 2001).

A esse intervalo denominamos de **Intervalo de Confiança**, e a nossa confiabilidade é 1-P (erro). Assim, um intervalo de confiança de 95% carrega consigo um risco de erro igual a 5%.

O risco de erro diminui com o aumento do valor de Z, ou do número de desvios padrões afastados da média, entretanto aumenta o intervalo de valores que a média deverá estar.

Em geral, o nível de confiança é simbolizado por (1- α) X 100%, onde α é a proporção de caudas da distribuição que estão fora do intervalo de confiança.

Erro de estimação

Em um intervalo de estimação, o erro refere-se ao desvio (ou diferença) entre a média amostral e a verdadeira média populacional. Uma vez que o intervalo de confiança tem centro na média amostral, o erro máximo provável é igual à metade da amplitude do intervalo de confiança, sendo calculado como:

$$e = Z\frac{\sigma_x}{\sqrt{n}}$$

Logo, o intervalo de confiança será:

$$\bar{x} \pm Z \frac{\sigma_x}{\sqrt{n}}$$

Onde:

$\bar{x}$ = média amostral

Z = valor da variável padronizada, conforme o nível de confiança desejado

σ_x = desvio padrão populacional

n= número de observações (tamanho da amostra)

A utilização desse método, com a curva normal padronizada, é válido somente quando conhecemos o desvio padrão σ populacional, ou seja, para grandes amostras (n>30).

A fórmula apresentada para o erro nos mostra que há efetivamente três determinantes do tamanho ou da quantidade de erro na estimativa:

• A confiança desejada, representada pelo valor de Z;

• A dispersão da população (σ_x);

• O tamanho da amostra (n).

Os fatores Z e σ_x (numerador) têm efeito direto no erro, ou seja, quanto maior o nível de confiança desejado ou a dispersão da população, maior o erro potencial. O tamanho da amostra, n, por estar no denominador tem efeito inverso no erro, ou seja, para grandes amostras o erro potencial é menor.

Com a fórmula do erro, através de manipulações algébricas, podemos determinar a quantidade de erro associada à dispersão de uma população, ou o tamanho da amostra, ou ainda o intervalo de confiança, partindo do pressuposto de as demais variáveis estarem definidas.

Determinação do tamanho da amostra

Para determinar o tamanho da amostra que deve ser coletada para atender os demais pressupostos fazemos:

$$e = Z \frac{\sigma_x}{\sqrt{n}}$$

Logo: $\sqrt{n} = Z \frac{\sigma_x}{e}$

E por fim: $n = \left(z \frac{\sigma_x}{e} \right)^2$

Assim, o tamanho n da amostra necessária dependerá do grau de confiança desejado, da dispersão entre

os valores da população e do erro tolerável.

Exemplo 1:

Qual o tamanho necessário da amostra para produzir uma estimativa com intervalo de confiança de 95% para a verdadeira média populacional, com um erro tolerável de 1,0 unidades em qualquer dos sentidos, se o desvio padrão da população é 6,0 unidades?

Solução:

São determinados:

$\sigma_x = 6,0$

erro: $e = 1,0$

intervalo de confiança = 95%, logo, o valor de Z = 1,96.

Logo:

$$n = \left(z \frac{\sigma_x}{e} \right)^2$$

$$n = \left(1,96 \frac{6,0}{1,0} \right)^2$$

O que resulta em n = 138,29 (arredonda-se sempre para o próximo inteiro) n = 139

Na próxima aula veremos como estimar a média populacional nas situações em que o desvio padrão é conhecido e quando é desconhecido.

Resumo

Vimos nessa aula que a estimação envolve a avaliação do valor de um determinado parâmetro populacional com base em dados amostrais. As estimativas podem ser pontuais ou especificar um intervalo de valores em que julgamos estar o parâmetro populacional, o que definimos como estimativa intervalar. Os intervalos de confiança são estimativas intervalares que incluem uma afirmação probabilística que indica a percentagem de intervalos que podemos esperar abranger o verdadeiro valor do parâmetro em seus limites. A amplitude do intervalo de confiança é função da dispersão dos valores populacionais, do nível de confiança desejado, do erro tolerável e do tamanho da amostra.

Exercícios

1. Suponha que as alturas dos alunos do IFSUL tenham distribuição normal com σ = 12 cm. Foi retirada uma amostra aleatória de 100 alunos, obtendo-se . Construir, ao nível de confiança de 95%, o intervalo para a verdadeira altura média dos alunos.

2. Foram retiradas 25 peças da produção diária de uma máquina, encontrando- se para uma medida uma média de 5,2 mm. Sabendo que as medidas têm distribuição normal com desvio padrão populacional de 1,2 mm, calcular os intervalos de confiança para a média aos níveis de 90%, 95% e 99%.

3. Uma população tem desvio padrão igual a 10,0 unidades. Determine o tamanho da amostra para produzir, a um intervalo de confiança de 90%, um erro de 1,5 unidades.

4. Determine intervalos de confiança para cada um dos seguintes casos:

a) média amostral= 30,0; desvio padrão= 3,0; tamanho da amostra= 50; I.C.= 90%
b) média amostral= 30,0; desvio padrão= 3,0; tamanho da amostra= 50; I.C.= 95%
c) média amostral= 30,0; desvio padrão= 3,0; tamanho da amostra= 50; I.C.= 99%

5. No exercício anterior, em qual dos casos o intervalo é mais amplo? Explique o porquê.

Referências

STEVENSON, W. J. **Estatística aplicada à administração**. Trad. Alfredo Alves de Farias. São Paulo, Ed. Harbra Ltda. 495p. 2001.

Estimativa da média e do desvio padrão

Unidade L
Estatística Básica

UNIDADE L

ESTIMATIVA DA MÉDIA E DO DESVIO PADRÃO

Objetivos

- Descrever a estimativa da média populacional a partir de uma amostra;
- Explicar o uso de uma afirmação probabilística em uma estimativa;
- Explicar o uso das distribuições amostrais em uma estimativa;
- Construir intervalos de confiança para médias populacional;
- Descrever o erro associado a uma estimativa.

Você verá por aqui...

A estimação é o processo que consiste em utilizarmos dados amostrais para estimar os valores de parâmetros populacionais desconhecidos. A média e o desvio padrão de uma população são as características mais comuns e importantes de serem estimadas a partir de amostras aleatórias. Alguns critérios devem ser estabelecidos na estimação para que possamos conhecer o nível de confiança e o erro associado na estimativa, tais como a normalidade dos dados e o tamanho da amostra selecionada. A metodologia para realizar essas estimativas é o assunto que abordaremos durante esta aula. Procure organizar seus horários disponibilizando um bom tempo para as atividades e aproveite bem este material.

Estimativa da média de uma população

Para estimar a média de uma população a partir da média amostral precisamos antes saber se o desvio padrão populacional é conhecido ou não. O método usado é distinto para cada um dos casos e, portanto, iremos abordar primeiramente a estimativa da média quando o desvio padrão da população é conhecido.

Estimativa da média populacional com desvio padrão conhecido

Conforme vimos na aula anterior, quando o desvio padrão populacional é conhecido, as estimativas pontual e intervalar da média populacional são dados por:

Estimativa pontual	$\mu = \bar{x}$
Estimativa intervalar	$\mu = \bar{x} \pm Z\frac{\sigma}{\sqrt{n}}$

Onde σ é o desvio padrão populacional conhecido.

A estimativa intervalar da média populacional assume a hipótese que a distribuição amostral das médias amostrais é normal.

Conforme visto na aula 10, pela aplicação do Teorema do Limite Central, em grandes amostras a hipótese

da normalidade não precisa ser testada, entretanto, para amostras de 30 ou menos observações, é importante testar se a população amostrada tem distribuição normal, ou ao menos aproximadamente normal.

De outra forma essas técnicas não podem ser utilizadas (STEVENSON, 2001).

Exemplo 1:

Uma indústria de tubos de aço possui um processo de produção que opera de maneira contínua, através de um turno completo de produção. É projetado para que cada tubo tenha um comprimento de 11m, e o desvio padrão conhecido é de 0,02 m. A intervalos periódicos são selecionadas amostras para determinar se o comprimento médio do tubo ainda se mantém igual a 11m ou se algo de errado ocorreu no processo de produção para que tenha sido modificado o comprimento do tubo produzido. Se tal situação tiver ocorrido, deve-se adotar uma ação corretiva. Uma amostra aleatória de 100 tubos foi selecionada e verificou-se que o comprimento médio do tubo foi de 10,998 m. Estime o comprimento médio de todos os tubos deste processo de produção usando um intervalo de confiança de: a) 95 % e b) 99 %.

$$\frac{1-\alpha}{2} = \frac{0,95}{2} = 0,475$$

Na tabela da distribuição normal padronizada, para o valor de 0,4750 → z = 1,96

$$\pm Z\frac{\sigma}{\sqrt{n}} = \pm(1,96)\frac{0,02}{\sqrt{100}} = \pm 0,00392$$

10,998 + 0,00392 = 11,002

10,998 – 0,00392 = 10,994

Logo: 10,994 ≤ μ ≤ 11,002

b) 99%

$$\frac{1-\alpha}{2} = \frac{0,99}{2} = 0,495$$

Na tabela da distribuição normal padronizada, para o valor de 0,495 → z = 2,58

$$\pm Z\frac{\sigma}{\sqrt{n}} = \pm(2,58)\frac{0,02}{\sqrt{100}} = \pm 0,00516$$

10,998 + 0,00516 = 11,003

10,998 – 0,00516 = 10,993

Logo: 10,993 ≤ μ ≤ 11,003

A estimativa pontual da média μ = $\overline{x}$ = 10,998

Estimativa da média populacional quando o desvio padrão é desconhecido

Quando o desvio padrão populacional é desconhecido (o que geralmente ocorre), usamos o desvio padrão amostral como estimativa, substituindo σx por Sx nas equações para intervalos de confiança e de erros. Quando o tamanho da amostra é superior a 30 (n>30) o desvio padrão amostral fornece uma aproximação bastante razoável do verdadeiro valor, na maioria dos casos. Entretanto, para amostras de 30 ou menos observações, a aproximação normal não é adequada e devemos usar a distribuição t de *Student*, que é a distribuição correta quando utilizamos Sx (STEVENSON, 2001).

> A distribuição t de Student foi criada por W. S. Gossett, funcionário de uma cervejaria irlandesa no princípio do século XX. Como a empresa não permitia que seus funcionários publicassem trabalhos em seu próprio nome, Gossett adotou o pseudônimo de Student em seus trabalhos sobre a distribuição t. Por isso é que ela ficou conhecida como distribuição t de Student (Fonte: Stevenson, 2001).

A distribuição **t** tem sua forma bastante semelhante com a distribuição normal. A principal diferença entre as duas distribuições é que a distribuição **t** tem maior área nas caudas, o que acarreta um maior valor de t em relação ao correspondente valor z, para um dado nível de confiança.

A distribuição t, diferentemente da normal, não é padronizada e por isso há uma distribuição t (curva) ligeiramente diferente para cada amostra, conforme varia o número de observação (n).

Para amostras de pequeno tamanho (n ≤ 30), a distribuição **t** é mais sensível ao tamanho da amostra, e para amostras maiores essa sensitividade diminui.

Para grandes amostras é razoável usar valores de z para aproximar valores t, muito embora a distribuição t seja sempre teoricamente correta quando não se conhece o desvio padrão da população, independente do tamanho da amostra (STEVENSON, 2001).

Para consultar uma tabela **t** precisamos conhecer o **nível de confiança** desejado e o número de **graus de liberdade**.

O número de graus de liberdade está relacionado com a maneira como calculamos o desvio padrão amostral:

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

Onde: S_x = desvio padrão amostral

x = valor da variável (de cada amostra)

$\overline{x}$ =média amostral

n-1 = graus de liberdade

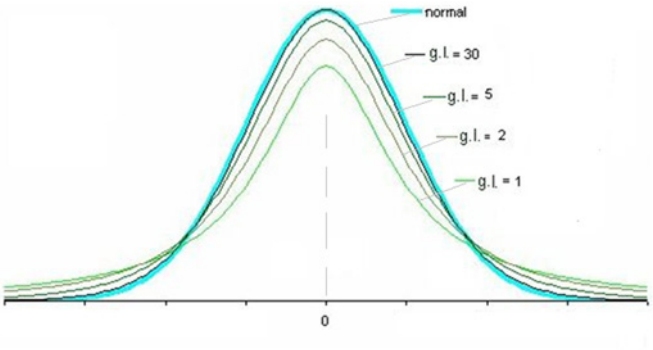
Logo, o número de graus de liberdade é igual a n-1, ou seja, o tamanho da amostra menos um.

Para facilitar o entendimento de graus de liberdade utilizaremos um exemplo apresentado em Stevenson (2001):

Suponhamos que queiramos três números cuja soma seja 10. O primeiro número pode ser tudo (mesmo negativo); o segundo número também. Mas o terceiro número está limitado à condição que a soma dos três deve ser 10. Escolhidos os dois primeiros valores, o terceiro está essencialmente determinado, não existe grau de liberdade para o terceiro valor. Há três números em jogo, mas liberdade só para dois.

O que é exigido é que a soma dos desvios em relação à média amostral seja zero, o que obriga um arredondamento do menor valor. Logo, o número de graus de liberdade é igual a **n-1**.

A Figura 1 apresenta a curva normal padronizada e a curva da distribuição t de *Student*, onde pode ser observada a maior área sobre as caudas para a distribuição t e a aproximação entre as duas curvas quando o número de graus de liberdade cresce (n=31, g.l.=30).



Comparação entre as distribuições normal e **t**, realçando a maior área nas caudas para a distribuição **t** e a similaridade das curvas quando aumenta o número de graus de liberdade (g.l.).

Figura L.1 - Comparações entre as distribuições

A Tabela 1 apresenta os valores t da distribuição de *Student*. Para utilizar a tabela precisamos conhecer, como escrito anteriormente, o nível de confiança desejado (que está representado na Tabela 1 pela área abaixo da curva para t_{α}), na primeira linha, e o número de graus de liberdade (n-1), que aparece na primeira coluna da tabela. O exemplo a seguir mostra como utilizar a tabela **t**.

Exemplo 2:

Queremos saber o valor de t para uma amostra de n=23, para um intervalo de confiança de 95%.

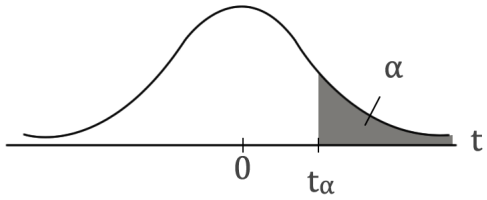
95% = 0,95 → 0,05 (nas duas caudas) = 0,05/2 = **0,025** (área na cauda superior)

n= 23 → número de graus de liberdade = n-1 → g.l.=**22**

Entrando na Tabela 1 com os valores de g.l.=22 e α = 0,025, encontramos o valor de **2,074** para **t**.

Convém salientar que a distribuição **t** só é teoricamente adequada quando a distribuição é normal. Na prática, quando n é maior do que 30 observações, a necessidade de admitir a normalidade diminui (STEVENSON, 2001).

Tabela 1 – Distribuição t de Student.



A tabela fornece os valores de t_α que correspondem a uma área α na cauda direita (superior) e a um número específico de graus de liberdade.

Graus de liberdade	Área α na cauda superior							
	0,10	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
1	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	1.371	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Fonte: Adaptado de Stevenson, 2001; original de Fisher, R. A. Statistical Methods for Research Workers, 1970.

Exemplo 3:

A seguinte amostra foi extraída de uma população com distribuição normal:

[9 – 8 – 12 – 7 – 9 – 6 – 11 – 6 – 10 – 9]

Estime o valor da média populacional com um intervalo de confiança de 95%.

Solução:

Média da amostra: $\overline{x}$ =8,7

Desvio padrão amostral: S_x=2

Como 1-α = 0,95 teremos: α = 1- 0,95 → 0,05 (nas duas caudas), logo, na Tabela 1 (unicaudal) teremos:

α = 0,05 / 2 → α = 0,025

Graus de Liberdade: n-1 → 10 – 1 , logo g.l. = 9

Consultando a Tabela **t** obtemos o valor de t_α = 2,262

Para o intervalo de confiança bicaudal temos: t_α = ±2,262

Substituindo na fórmula $\bar{x} \pm t_\alpha \frac{S_X}{\sqrt{n}}$

teremos:

$$\bar{x} \pm t_\alpha \frac{S_X}{\sqrt{n}} = 8{,}7 \pm (2{,}262)\frac{2}{\sqrt{10}} = 8{,}7 \pm 1{,}4306$$

Resposta: 7,269 ≤ μ ≤ 10,131

Amostragem de pequenas populações

Quando a população é **finita** e a amostra é constituída por **mais de 5%** da população devemos aplicar um **fator de correção finita** para modificar os desvios padrões das fórmulas.

O quadro abaixo resume as correções necessárias nas fórmulas:

Situação	Intervalo de confiança	Erro
σx conhecido	$\bar{x} \pm Z \frac{\sigma_X}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$	$Z \frac{\sigma_X}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
σx desconhecido	$\bar{x} \pm t_\alpha \frac{S_X}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$	$t_\alpha \frac{S_X}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

Para estimar o tamanho da amostra, conhecidos os demais parâmetros, também precisamos corrigir a maneira de calcular, conforme é mostrado no quadro a seguir:

Stevenson (2001) destaca que a não utilização dessas fórmulas quando for apropriado fazê-lo, pode resultar no dimensionamento de uma amostra que exceda o tamanho da população.

Situação	Tamanho da amostra (n)
σx conhecido	$n = \frac{z^2 \sigma_x^{\,2} N}{z^2 \sigma_x^{\,2} + e^2 (N-1)}$
σx desconhecido	$n = \frac{t^2 S_x^{\,2} N}{t^2 S_x^{\,2} + e^2 (N-1)}$

Stevenson (2001) destaca que a não utilização dessas fórmulas quando for apropriado fazê-lo, pode resultar no dimensionamento de uma amostra que exceda o tamanho da população.

Resumo

Para estimar a média populacional a partir de uma amostra utilizamos dois métodos em situações distintas. Se o desvio padrão populacional é conhecido usamos a distribuição normal padrão e, se não for conhecido, usamos a distribuição t de *Student*. O tamanho da amostragem também deve ser considerado na estimação e define a necessidade ou não de testar a normalidade dos dados. Quando o tamanho da amostra é superior a 5% da população, as fórmulas para as estimativas intervalares devem ser modificadas com fatores de correção finita.

Exercícios

1. Em uma pesquisa sobre as velocidades desenvolvidas por veículos em uma rodovia, um equipamento com radar monitorou a via por duas horas. No período de observação, 100 carros passaram pelo equipamento a uma velocidade média de 75 km/h, com desvio padrão de 15 km/h.

a) Estime a verdadeira média da população (estimativa pontual);
b) Estime um intervalo para a média populacional com um nível de confiança de 90%;
c) Repita o item b para um intervalo de confiança de 98%;
d) Qual o erro máximo associado ao intervalo encontrado no item c?

2. Foram retiradas 25 peças da produção diária de uma máquina, encontrando-se para uma determinada medida uma média de 5,2 mm. Sabendo que as medidas têm distribuição normal com desvio padrão populacional de 1,2 mm, estime os intervalos de confiança para a média aos níveis de 90%, 95% e 99%.

3. De uma distribuição normal obteve-se a seguinte amostra:

[25,2 - 26,0 - 26,4 - 27,1 - 28,2 - 28,4].

Determine o intervalo de confiança para a média da população a um nível de 90% e de 95%.

4. Em que condições é necessário saber que uma distribuição populacional é aproximadamente normal?

5. Em que condições é aplicável a distribuição **t**?

Referências

STEVENSON, W. J. **Estatística aplicada à administração**. Trad. Alfredo Alves de Farias. São Paulo, Ed. Harbra Ltda. 495p. 2001.