ESTATÍSTICA BÁSICA

Um resumo inteligente

José Antonio Farias Coelho

Centro Universitário
Estácio | FIC

José Antônio Farias Coelho

ESTATÍSTICA BÁSICA

Um resumo inteligente

1ª Edição

Centro Universitário Estácio | FIC

Fortaleza – Ceará
2017

Revisão Ortográfica
Marineide Meireles Nogueira

Normalização e Padronização
Luiza Helena de Jesus Barbosa

Capa
Janete Pereira do Amaral

Programação Visual e Diagramação
Janete Pereira do Amaral

Revisão de ABNT
Luiza Helena de Jesus Barbosa

Dados Internacionais de Catalogação na Fonte

C672c Coelho, José Antônio Farias
Estatística básica: um resumo inteligente / José Antônio Farias Coelho. Fortaleza: Centro Universitário Estácio do Ceará, 2020.
92f.; 30cm.

ISBN:

1. Estatística básica 2. Descritiva I. Coelho, José Antônio Farias II. Centro Universitário Estácio do Ceará
CDD 515.63

CENTRO UNIVERSITÁRIO ESTÁCIO DO CEARÁ
Pró-Reitoria de Pós-Graduação, Pesquisa e Extensão
Núcleo de Publicações Acadêmico-Científicas

CONSELHO EDITORIAL

Núcleo de Publicações Acadêmico-Científicas
Rua Vicente Linhares, 308 - Aldeota
CEP: 60.135-270 - Fortaleza – CE - Fone: (85) 3456-4100
www.publica-estaciofic.com.br

O AUTOR

José Antônio Farias Coelho

Graduado em Tecnologia da Construção Civil pela Universidade Estadual Vale do Acaraú (1990), cursou MBA em Comércio eletrônico na Universidade de Fortaleza (2004), Especialização em Gestão de APL na Universidade de Fortaleza (2010), Mestrado Acadêmico em Administração de Empresas na UECE (2013). Atualmente cursa Especialização em Metodologia do Ensino da Matemática na Estácio (2016).

Professor e coordenador no Centro Universitário Estácio do Ceará em cursos de graduação e pós graduação nas áreas de engenharia e arquitetura. Membro da Comissão Própria de Avaliação – CPA.

PREFÁCIO

Formatamos este livro em um conteúdo resumido, mas, didaticamente apresentado, na busca de facilitar a compreensão do leitor, contudo longe de ser o recurso final do aprendizado desta disciplina, que é ao mesmo tempo bela e complexa.

Há de se tornar público que, face à nossa formação acadêmica e relacionamento profissional, o presente livro recebeu preponderante influência do livro Estatística Fácil, do professor Antônio Arnot Crespo, o qual recomendamos a todos os alunos que aspiram a um aprofundamento e a um maior rigor no assunto.

Críticas e sugestões hão de surgir. E serão bem-vindas. Resta-nos o consolo de ter envidado esforços para empregar utilmente o nosso tempo. "A censura que nos for feita – se faz oportuno Souza Pinto – há de ser mitigada pelo censor se ele chegar a ter consciência de nossa boa vontade em acertar".

José Antônio Farias Coelho

SUMÁRIO BÁSICO

1 CONCEITOS BÁSICOS

"A Estatística é a ciência que estuda os processos de coleta, organização e análise de dados referentes a uma determinada população, assim como os métodos de tirar conclusões ou fazer predições com base nestes dados. Ela pode ser aplicada a diversas finalidades tais como controle de qualidade, pesquisas eleitorais, avaliação de riscos financeiros, entre muitas outras. etc."

A Estatística nos fornece ferramentas para aprender a partir dos dados coletados, e podermos tomar decisões com alto grau de confiabilidade.

O estudo da Estatística é muito importante para diversas áreas. Tomando como exemplo a área de saúde, verificamos que a mesma é bastante dependente de técnicas estatísticas para gerar novos conhecimentos e poder tomar decisões importantes.

Tais como:

- Quais grupos devem vacinar?
- Quais regiões precisam de maior atenção por parte das secretarias de saúde?

A estatística é a ciência que estuda os processos de coleta, organização e análise de dados referentes a uma determinada população, assim como os métodos de tirar conclusões ou fazer predições com base nestes dados. Entre suas finalidades estão a pesquisa eleitoral, o controle de qualidade e a avaliação de riscos financeiros.

1.1 Estatística e suas funções

O estudo dos fenômenos coletivos envolve uma série de etapas:

Estatística descritiva:

- Coleta,
- Organização,
- Descrição.

Estatística inferencial ou indutiva:

- Análise,
- Interpretação.

É importante salientar que a Estatística Inferencial utiliza-se da Estatística Descritiva, e também dos métodos de cálculo de probabilidades, para poder chegar a conclusões válidas e indicar possíveis decisões a serem tomadas com alto grau de confiabilidade.

1.2 População e amostra

O estudo dos fenômenos coletivos normalmente se dá a partir de amostras, que são subconjuntos das populações, que por sua vez, são o alvo do estudo a ser realizado.

Em estatística, "amostra", é o conjunto de todos os elementos que têm pelo menos uma característica em comum, não necessariamente sendo composta por pessoas.

Exemplo: Se vamos utilizar a estatística com a finalidade de fazer controle da qualidade de peças de roupas que foram compradas de determinado fornecedor, a população deste exemplo será composta de todas as peças de roupas que foram entregues por este fornecedor em determinado lote.

O controle de qualidade selecionará uma amostra desta população - algumas peças de roupa - para poder fazer os testes pertinentes e decidir se o lote de peças será aceito ou não.

A partir dos resultados obtidos de uma amostra toma-se uma decisão que irá afetar toda a população.

Outro exemplo: Se o nosso estudo tem como objetivo saber se determinado medicamento é eficaz em relação à determinada doença, a nossa população são todas as pessoas acometidas desta doença.

Porém o estudo vai ser realizado com um grupo bem menor de pessoas, também acometidas desta doença.

Uma amostra precisa ser representativa da população que queremos estudar, ou seja, ter as mesmas características da população.

Desta forma, se vamos fazer uma pesquisa com o objetivo de prever quem será o novo prefeito de determinada cidade, a população em questão será composta de todos os eleitores desta cidade e a amostra deve representar esta população.

Homens e mulheres devem estar representados, todas as classes sociais e assim por diante. A amostra precisa contemplar as mesmas características da população.

Enfim, o objetivo do estudo da amostra é poder generalizar estes resultados para toda a população.

1.3 Parâmetro e estatística

A estatística é utilizada para estimar um parâmetro, pois o interesse final é conhecer a população a partir da amostra. Neste caso, estamos utilizando a palavra estatística como a representação de um valor calculado a partir dos valores de uma amostra, diferentemente de estatística como ciência, que foi apresentada inicialmente.

Parâmetros: Características de uma população, ou seja, são valores calculados quando conseguimos obter os valores de todos os elementos da população (normalmente é muito difícil de conseguir).

Estatísticas: Características de uma amostra, ou seja, são valores calculados a partir dos dados de uma amostra.

1.4 Variáveis

Para se estudar determinada população é necessário que se defina quais são as características de interesse, ou seja, as variáveis que serão estudadas.

Por exemplo, se quiséssemos traçar o perfil dos doentes de determinada enfermidade, o que seria importante conhecer?

- Certamente sexo e idade parecem ter características interessantes.
- Será que classe social e nível de escolaridade também?
- E o número de pessoas que compõem a família do doente?
- Quais outras características seriam importantes?

Dependendo do objetivo da pesquisa as variáveis são diferentes. A variável (características), como percebemos pela própria palavra, pode assumir diferentes valores, ou seja, varia.

Os valores obtidos das variáveis são chamados = DADOS

- Por exemplo, sexo pode assumir o valor "masculino" ou "feminino".
- Nível de escolaridade pode assumir os valores "Fundamental incompleto", "Fundamental completo", "Médio incompleto" e assim por diante.
- Número de pessoas que compõe a família pode variar de 1 (a pessoa mora sozinha) até um número bem maior.

Se formos reparar nos valores destes dados, percebemos que têm características bem diferentes.

1.5 Tipos de variáveis

As variáveis são classificadas em:

Variáveis qualitativas

As variáveis qualitativas são as que se referem à qualidade e podem ser nominais (sem ordem) ou ordinais (com ordem).

As variáveis qualitativas nominais são aquelas que não seguem nenhuma ordem, como o exemplo das causas de internação. Exemplos:

- **População: Pacientes internados em determinado hospital.**
- Variável: Causas da internação.
- Valores: Infecção, pneumonia, doença cardiovascular, etc.
- **População: Pacientes internados em determinado hospital.**
- Variável: Sexo.
- Valores: Masculino ou feminino.
- **População: Doadores de sangue de determinado hemocentro.**
- Variável: Tipo de sangue.
- Valores: A, AB, B e O.
- **População: Visitantes de determinada exposição de arte ao ar livre.**
- Variável: Nível de escolaridade.
- Valores: Fundamental incompleto, Fundamental completo, Médio incompleto, Médio completo, etc.

As variáveis qualitativas são assim classificadas independente de os valores atribuídos a esta variável serem numéricos ou não. Por exemplo, se em determinado formulário pede-se para identificar o sexo masculino com o número 1 e o sexo feminino com o número 2, a variável sexo continua a ser classificada como qualitativa, pois os números, neste caso, não indicam quantidade mas uma qualidade (masculino ou feminino).

Variáveis quantitativas

Que podem ser:

<u>Quantitativas discretas</u>

As variáveis quantitativas discretas assumem valores definidos de um conjunto enumerável. Muitas vezes são valores inteiros, mas não necessariamente. Normalmente obtemos os valores das variáveis quantitativas discretas contando ou enumerando. Exemplos:

- **População: Pacientes internados em determinado hospital.**
- Variável: Número de vezes que foi internado.
- **População: Estudantes de determinada escola.**
- Variável: Número de irmãos.

Quantitativas contínuas

As variáveis quantitativas contínuas podem assumir qualquer valor dentro de um certo intervalo - menor valor possível para aquela característica (variável) e o maior valor possível.

Por exemplo, a altura de uma pessoa pode ser qualquer valor entre o comprimento do menor bebê nascido e a altura da maior pessoa já registrada, mas nunca será seis metros.

Normalmente, se obtém os valores das variáveis quantitativas contínuas através de medições, sendo estes valores pertencentes ao conjunto dos números reais. São exemplos:

- **População: Pacientes internados em determinado hospital.**
- Variável: Idade.
- **População: Estudantes de determinada escola.**
- Variável: Tempo de deslocamento até a escola.
- **População: Doadores de sangue de determinado hemocentro.**
- Variável: Peso.

1.6 Níveis de mensuração

Além da classificação das variáveis, temos os níveis de mensuração. A forma de medir os valores das variáveis vai determinar quais são as operações matemáticas possíveis de serem realizadas, e consequentemente quais medidas poderão ser calculadas e quais técnicas estatísticas poderão ser utilizadas.

Qualitativas

Variáveis qualitativas podem ser no máximo, ordenáveis. Explicando melhor, as variáveis qualitativas podem ser classificadas como:

Nominais: As variáveis nominais são aquelas que eu só consigo classificar os valores em grupos que não têm como ser colocados em ordem por nenhum critério.

Exemplo:

- População: Estudantes de determinada escola.
- Variável: Gênero de filme preferido.
- Valores: Comédia, Terror, Drama, Romance, etc.

Não temos como ordenar estes valores a partir de um critério natural da própria variável. Neste caso, o nível de mensuração é nominal.

Ordinais: As variáveis ordinais podem ser dispostas numa ordem que levam em conta um critério natural da variável. Esta ordem natural não quantifica o quanto um valor é mais ou menos que o outro, mas existe um critério de ordenação.

Exemplo:

- População: Estudantes de determinada escola.
- Variável: Nível de satisfação geral com a escola.
- Valores: Muito insatisfeito, Insatisfeito, Muito satisfeito.

Podemos ordenar as respostas do menor grau de satisfação para o maior grau de satisfação, mas não temos como quantificar exatamente estas diferenças. Neste caso o nível de mensuração é ordinal.

Quantitativas

As variáveis quantitativas podem permitir calcular proporções ou não. Podem ser classificadas como:

Intervalar: Este nível é similar ao ordinal (qualitativa), mas as magnitudes são conhecidas. Permite apenas números inteiros. Os dados não possuem um ponto inicial zero natural e intervalo entre uma categoria e outra dos valores desta variável é fixo. Permite apenas operações de adição e subtração.

Exemplos:

- Anos (1970, 1980, 1991 e 2000).(As magnitudes das diferenças entre os anos são conhecidas, mas o tempo não começou em zero).
- Número de alunos em uma sala. (Só é possível ter um número inteiro de alunos).

Razão: O nível de mensuração de razão é similar ao intervalar, mas possui algumas diferenças. O valores são números reais. Existe um ponto inicial zero natural, o que significa a possibilidade de ausência de valor medido. Dessa forma, é possível dizer que uma quantidade é maior que outra em X vezes. É o único nível de mensuração que permite qualquer tipo de operação aritmética.

Exemplos:

- Variáveis de peso, comprimento, Temperatura em graus Kelvin, tempo de reação, entre outros.

O quadro, abaixo, faz a comparação entre as variáveis qualitativas e quantitativas:

TIPO DE VARIÁVEL	NÍVEIS DE MENSURAÇÃO
Qualitativa	Nominal ou ordinal
Quantitativa - Discreta - Contínua	Intervalar ou razão

1.7 Atividade Proposta

Imagine que você vai se inscrever numa disciplina para cursar no próximo semestre. Esta disciplina vai ser disponibilizada em 2 turmas - turma A e turma B - e você poderá escolher entre uma delas.

Você faz uma pesquisa e descobre que neste semestre o professor da turma A reprovou 10 pessoas e o professor da turma B reprovou 5 pessoas. O professor A reprovou o dobro de alunos do que o professor B.

Com estas informações, qual turma você escolheria para se inscrever?

Tem certeza que você fez a melhor opção?

2 POPULAÇÃO E AMOSTRAS

As pesquisas quantitativas, normalmente, tratam uma grande quantidade de dados referentes a uma determinada POPULAÇÃO com objetivo de extrair informações destes dados. Normalmente não se pesquisa toda a população de interesse, mas um grupo menor chamado AMOSTRA.

As amostras precisam representar a população, ou seja, precisam ter as mesmas características da população, de modo que o estudo da amostra possa ser generalizado para toda a população. As diversas técnicas de amostragem têm como objetivo a obtenção de uma amostra representativa.

Porquê da utilização das amostras?

O primeiro e mais importante motivo é justamente o tamanho da população, que normalmente é muito grande, o que demandaria muito tempo e recursos financeiros.

O segundo é por conta de que o uso de amostras possibilita um estudo mais criterioso de cada unidade observacional, possibilitando um maior valor científico do que o estudo sumário de toda a população.

Como selecionar uma amostra?

Existem várias técnicas de amostragem que podem ser utilizadas para a composição de uma amostra. Estas técnicas estão classificadas a seguir.

2.1 Amostra aleatória, casual ou probabilística.

A característica das amostras aleatórias é possibilitar que todos os elementos da população a ser estudada tenham igual probabilidade de ser selecionados para compor a amostra. Para isto, todos os elementos da população precisam estar identificados de alguma maneira, por exemplo através de um número.

A amostra poderá ser composta a partir do sorteio de bolas numeradas presentes em uma urna.

Outra forma de selecionar uma amostra, e garantir a igual probabilidade de seleção de cada elemento da população, é selecionar os números que irão compor a amostra a partir de uma tabela de números aleatórios.

Uma amostra aleatória pode ser:

Simples: A amostra aleatória simples, também conhecida como casual simples ou aleatória, é obtida a partir de um sorteio, onde todos os elementos da população são considerados homogêneos para a variável que se quer observar.

Exemplo: Para obter a percentagem de pessoas que gostaram ou não de determinado filme, selecionaremos uma sessão, sortearemos alguns números e convidaremos as pessoas dos assentos sorteados para responder a pesquisa.

Estratificada: A amostra aleatória estratificada vai realizar o sorteio dentro de estratos previamente identificados. Os estratos correspondem a grupos que são internamente homogêneos, porém externamente heterogêneos.

Exemplo: Se vamos fazer uma prévia eleitoral e sabemos que a classe social A costuma votar diferentemente da classe social B, que também vota diferentemente da classe social C, vamos ter três estratos.

O sorteio será realizado em cada estrato separadamente e a amostra será composta pela reunião destes três sorteios. É importante ressaltar que a composição da amostra terá que ter a mesma proporção de elementos das classes A, B e C que a proporção na população.

Desta forma, se temos 10% da população sendo composta por pessoas da classe A, 30% da classe B e 60% da classe C, nossa amostra também terá de ser composta por 10% dos elementos da amostra da classe A, 30% da classe B e 60% da classe C.

Outro exemplo: Separamos uma população de sujeitos do sexo masculino em quatro amostras de acordo com a idade: amostra A – inferior a 20 anos; amostra B – entre 21 e 30 anos; amostra C – entre 31 e 40 anos; amostra D – entre 41 e 50 anos.

2.2 Amostra semi probabilística

Neste tipo de técnica, os elementos da amostra serão selecionados por procedimento parcialmente aleatório, como veremos a seguir. Os exemplos de técnicas semi probabilísticas são:

Amostragem sistemática

A amostragem sistemática define um sistema para a composição da amostra.

Por exemplo, se quero selecionar 10% de uma população de 1.000 pessoas onde cada pessoa é identificada por um número sequencial de três dígitos iniciando por 000. Podemos inicialmente sortear um número de 0 a 9, vamos supor o número 5, e a partir daí todos os números que terminam por 5 farão parte da minha amostra (vou somando 10 a cada número selecionado iniciando pelo primeiro número sorteado, ou seja, a cada dez pessoas seleciono uma – 10%).

Amostragem por conglomerados

A amostragem por conglomerados é obtida a partir de todos os elementos de determinados grupos selecionados, chamados de conglomerados.

Um conglomerado é um grupo qualquer, por exemplo, uma determinada escola, um determinado hospital, uma determinada empresa, etc.

Por exemplo, imaginemos uma pesquisa para saber o nível de ensino das escolas particulares. Podemos sortear três escolas particulares (três conglomerados). A minha amostra será composta de todos os alunos destas três escolas sorteadas.

Outro exemplo: queremos identificar as tendências de consumo de mulheres cariocas com idade entre 30 e 45 anos. Para esse estudo, dividiremos o total da população por zonas da cidade, selecionando uma amostra para cada zona.

Amostragem por quotas

A amostragem por quotas é semelhante à amostragem por estratos, porém os elementos dentro de cada quota não têm igual probabilidade de ser selecionados para a amostra, ou seja, a amostragem por quotas não é aleatória. A população precisa ser relativamente conhecida para a identificação das quotas a serem selecionadas.

2.3 Amostra não probabilística ou de conveniência

A amostragem por conveniência é realizada a partir da facilidade (conveniência) do pesquisador em relação aos elementos a serem selecionados para a sua amostra.

Por exemplo: Um médico vai selecionar para a sua pesquisa seus próprios pacientes, um professor vai selecionar seus próprios alunos ou os alunos da escola onde trabalha.

Sempre é bom ressaltar a necessidade de representatividade da amostra selecionada para a pesquisa. Não adianta nada um processamento dos dados impecável se os próprios dados não representarem a população que se supõe representar.

Outro exemplo, fim de saber a opinião dos estudantes brasileiros sobre o transporte público, escolheremos uma amostra entre os estudantes de duas universidades próximas.

Observação importante!!!

As técnicas de amostragem aleatórias são preferíveis, mas nem sempre são possíveis na prática. Devemos estar sempre atentos para evitar tendenciosidades.

A importância da forma de perguntar.

A forma de perguntar é importantíssima para que os dados sejam verdadeiros. Imagine que você está estudando Literatura. Especificamente o autor Jorge Amado. Você vai fazer uma pesquisa para saber se as pessoas, em geral, conhecem Jorge Amado, já leram seus livros, quais livros leram, etc. Inicialmente, você precisará definir quais são as informações que vai querer saber. Em seguida, como perguntar o que vai querer saber. Se você fosse este pesquisador. O que iria perguntar às pessoas? Como faria estas perguntas?

Uma pergunta poderia ser:

Quantos livros de Jorge Amado você já leu? O que você acha dessa pergunta? É adequada? Veja o que ele pode pensar: "quantos livros já li? Mas por esta pergunta, eu deveria ter lido pelo menos um..." Será que você não inibiria o seu entrevistado?

Pesquisas indicam que as pessoas não gostam de responder sobre hábitos de higiene. Também, não costumam dizer a verdade quando o assunto é sexo. O que fazer?

2.4 Atividade Proposta

Correlacione o conceito ao exemplo correto.

(1)	Amostragem aleatória estratificada	( )	Para obter a percentagem de pessoas que gostaram ou não de determinado filme, selecionaremos uma sessão, sortearemos alguns números e convidaremos as pessoas dos assentos sorteados para responder a pesquisa.
(2)	Amostragem aleatória simples	( )	A fim de saber a opinião dos estudantes brasileiros sobre o transporte público, escolheremos uma amostra entre os estudantes de duas universidades próximas.

(3)	Amostragem por conglomerado	( )	Queremos identificar as tendências de consumo de mulheres cariocas com idade entre 30 e 45 anos. Para esse estudo, dividiremos o total da população por zonas da cidade, selecionando uma amostra para cada zona.
(4)	Amostragem por conveniência	( )	Separamos uma população de sujeitos do sexo masculino em quatro amostras de acordo com a idade: amostra A – inferior a 20 anos; amostra B – entre 21 e 30 anos; amostra C – entre 31 e 40 anos; amostra D – entre 41 e 50 anos.

3 SÉRIES ESTATÍSTICAS, DADOS ABSOLUTOS E DADOS RELATIVOS

Um dos objetivos da estatística é organizar os dados de forma a permitir uma visualização de algum padrão de comportamento geral destes dados de forma mais fácil, simples e rápida do que a simples apresentação dos dados obtidos sem nenhum tratamento.

Uma das técnicas de organização de dados é apresentá-los em tabelas, chamadas de séries estatísticas, quando a distribuição dos dados estatísticos é organizada em função do tempo, do local ou da espécie.

Dados absolutos são os dados provenientes das contagens ou medidas realizadas no processo de coleta de dados e dados relativos são o resultado de relações obtidas a partir dos dados absolutos. As medidas relativas são muito utilizadas na área de saúde.

3.1 Série estatística

É toda tabela que apresenta a distribuição de um conjunto de dados estatísticos em função da época, do local ou da espécie.

Dependendo do fator que varia - época, local ou espécie - as séries estatísticas são classificadas em quatro tipos: Históricas, Geográficas, Específicas e Conjugadas.

Históricas

Séries históricas, cronológicas, temporais ou marchas são as séries onde a variável de interesse foi observada em momentos diferentes mantendo-se constante o local. Neste caso, não há discriminação em relação ao local da observação, o que varia é a época da observação.

Total de registro de casamento no Brasil entre 1980 e 2005

Ano	Número de registros
1980	948.164
1985	952.294
1990	777.460
1995	734.045
2000	732.721

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de População e Indicadores Sociais, Pesquisa de Assistência Médico-Sanitária - 2005.

Geográficas

Séries geográficas, espaciais, territoriais ou de localização são as séries onde a variável de interesse foi observada numa determinada época sendo organizada em função das regiões ou localizações da coleta de dados. Nesta caso, o momento da observação é determinado e observamos que a variação está nas regiões.

Internações no Brasil, no ano de 2004, em estabelecimentos de saúde da rede pública, por grandes regiões

Região	Número de internações	%
Norte	1.100.735	15,68%
Nordeste	2.260.057	32,18%
Centro-oeste	714.576	10,18%
Sudeste	2.315.352	32,97%
Sul	631.369	8,99%
Total:	**7.022.089**	**100,00%**

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de População e Indicadores Sociais, Pesquisa de Assistência Médico-Sanitária - 2005.

Específicas

Séries específicas ou categóricas são as séries onde a variável de interesse foi observada em determinada época e local, sendo discriminada segundo especificações ou categorias. Neste caso, época e local permanecem constantes e é a própria variável que varia dependendo da classificação das categorias.

Distribuição dos leitos da rede pública no Brasil, de acordo com a esfera administrativa. Brasil, 2005

Esfera administrativa	Número de leitos	%
Federal	17.189	11,54%
Estadual	61.699	41,42%
Municipal	70.078	47,04%
Total:	**148.966**	**100,00**

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de População e Indicadores Sociais, Pesquisa de Assistência Médico-Sanitária - 2005.

Conjugadas

Séries conjugadas, tabelas de contingência ou tabelas de dupla-entrada são tabelas onde fazemos uma relação entre duas variáveis, sendo o resultado da conjugação de duas séries.

Região	Número de internações		Total
	Rede pública	Rede Privada	
Norte	1.100.735	645.819	1.746.554
Nordeste	2.260.057	2.994.921	5.254.978
Centro-oeste	714.576	1.069.994	1.784.570
Sudeste	2.315.352	8.479.447	10.794.799
Sul	631.369	3.040.393	3.671.762
Total:	**7.022.089**	**16.230.574**	**23.252.663**

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de População e Indicadores Sociais, Pesquisa de Assistência Médico-Sanitária - 2005.

Este exemplo é a conjugação de uma série geográfica e uma série específica.

3.2 Dados absolutos e relativos

Dados brutos são os valores obtidos das variáveis a partir do processo de coleta de dados e que ainda não foram organizados.

Dados absolutos

Os dados absolutos referem-se à contagem dos dados brutos, sendo chamados de frequência absoluta.

Por exemplo, se uma pesquisa perguntou a 10 casais quantos filhos cada casal tinha, temos 10 respostas, que ao serem organizadas saberemos quantos casais não têm filhos, quantos casais tem apenas 1 filho, quantos casais tem 2 filhos e assim por diante. Estas respostas correspondem aos dados absolutos, ou frequências absolutas. Porém, os dados absolutos podem não dar uma visão muito boa do que está ocorrendo.

Dados relativos

Os dados relativos são o resultado de relações entre os dados absolutos que facilitam a comparação entre quantidades. Alguns exemplos são as percentagens, índices, coeficientes e taxas.

Por exemplo, vamos supor que você vá se inscrever na disciplina de Estatística que tem 2 turmas abertas com dois professores diferentes.

Sabendo que no período passado 10 alunos foram reprovados com o professor da turma A, e que 15 alunos foram reprovados com o professor da turma B, você se inscreveria na turma A sem pensar duas vezes?

Neste caso, você estaria comparando a frequência absoluta 10 com a frequência absoluta 15. E se a turma do professor A do período passado tivesse 20 alunos matriculados e a turma do professor B tivesse 60 alunos, a sua decisão seria a mesma?

Provavelmente não. Neste caso hipotético, 50% dos alunos do professor A do período passado foram reprovados contra 25% de reprovações do professor B. Muitas vezes é importante relacionar os dados absolutos obtendo dados relativos para podermos tomar decisões.

3.3 Percentagens

A percentagem é a relação entre um dado absoluto e o total de dados absolutos. Observe a tabela abaixo:

Distribuição dos leitos da rede pública no Brasil, de acordo com a esfera administrativa. Brasil, 2005

Esfera administrativa	Número de leitos	%
Federal	17.189	11,54%
Estadual	61.699	41,42%
Municipal	70.078	47,04%
Total:	**148.966**	**100,00%**

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de População e Indicadores Sociais, Pesquisa de Assistência Médico-Sanitária - 2005.

Os dados absolutos são:

- 17.189 Federal
- 61.699 Estadual
- 70.078 Municipal

Esses dados correspondem ao total de leitos da esfera administrativa federal, estadual e municipal respectivamente.

Calculamos as percentagens do seguinte modo: valor absoluto da esfera administrativa de interesse dividido pelo número total de leitos e multiplicamos por cem.

Exemplos dos cálculos:

- Esfera Federal = (17189 x 100)/148966 = 11,53888%
- Esfera Estadual = (61699 x 100)/148966 = 41,41818%
- Esfera Municipal = (70078 x 100)/148966 – 47,04295%

Porém, não é conveniente apresentar todas as casas decimais. A regra de arredondamento, segundo o IBGE apud (Arnot, 2009), orienta que: a última casa decimal que quisermos manter vai permanecer inalterada se o algarismo seguinte for 0, 1, 2, 3 ou 4.

Por exemplo: 47,04295 vai ser arredondado para 47,04 porque a terceira casa decimal é igual a 2 e quisemos manter até a segunda casa decimal. A última casa decimal que quisermos manter vai ser acrescentada de uma unidade se o algarismo seguinte for 5, 6, 7, 8 ou 9.

41,41818 vai ser arredondado para 41,42 porque a terceira casa decimal é igual a 8 e quisemos manter até a segunda casa decimal.

Desta forma os valores arredondados para as esferas federal, estadual e municipal foram 11,54%, 41,42% e 47,04% respectivamente. A interpretação destes valores significa que de cada 100 leitos 47 deles estão na esfera municipal.

3.4 Índices

Os índices são calculados dividindo-se duas grandezas de natureza diferentes, uma grandeza não inclui a outra. Veja os exemplos:

$$\text{Índice cefálico} = \frac{\text{Diâmetro transverso do crânio}}{\text{Diâmetro longitudinal do crânio}} x\ 100$$

$$\text{Densidade demográfica} = \frac{\text{População}}{\text{Superfície}}$$

$$\text{Renda per capita} = \frac{\text{Renda}}{\text{População}}$$

3.5 Coeficientes

Os coeficientes são calculados dividindo-se o número de ocorrências e o número total, sendo o número total composto da soma do número de ocorrências e não ocorrências. Veja os exemplos:

$$\text{Coeficiente de natalidade} = \frac{\text{Número de nascimentos}}{\text{População total}}$$

$$\text{Coeficiente de evasão escolar} = \frac{\text{Número de alunos evadidos}}{\text{Número inicial de matriculados}}$$

3.6 Taxas

As taxas são calculadas multiplicando-se os coeficientes por 10, 100, 1000, etc.

Taxa de natalidade = coeficiente de natalidade x 1.000

Taxa de evasão escolar = coeficiente de evasão escolar x 100

3.7 Atividade Proposta

Precisamos obter uma solução a 2% de determinado concentrado. Se quiséssemos meio litro desta solução, quando do concentrado, e quanto do diluente, teremos de utilizar?

a) 1 ml e 499 ml, respectivamente.
b) 2 ml e 498 ml, respectivamente.
c) 10 ml e 490 ml, respectivamente.
d) 20 ml e 480 ml, respectivamente.
e) 20 ml e 500 ml, respectivamente.

4 ORGANIZAÇÃO DE DADOS: DADOS ISOLADOS E AGRUPADOS SEM INTERVALOS DE CLASSES

Após a coleta de dados, temos os dados ainda sem organização que são chamados de dados brutos. A apuração de dados qualitativos é realizada contando a quantidade de respostas de cada categoria. A partir desta contagem, podemos organizar a informação em tabelas seguindo as normas do IBGE.

Tabela

É uma representação sintética das informações obtidas sobre o comportamento das variáveis coletadas através de processos de pesquisa. Vamos conhecer os componentes de uma tabela:

Título: Toda tabela precisa ter um título localizado no topo da tabela. O título tem como objetivo informar o significado dos dados apresentados na tabela, respondendo às perguntas: O quê?, Quando?, Onde?

Cabeçalho: O cabeçalho localiza-se logo abaixo do título e especifica o conteúdo das colunas, devendo estar destacado por traços horizontais.

Corpo: É o próprio conteúdo da tabela, sendo formado pelos dados dispostos em linhas e colunas.

Células: Cada célula é o encontro de uma linha com uma coluna e vai conter apenas uma informação.

Rodapé: O rodapé localiza-se abaixo do último traço horizontal da tabela e vai conter informações importantes que não foram informadas no título, tais como a fonte dos dados (qual foi a Instituição de Pesquisa que forneceu os dados e em qual data) e notas que podem acrescentar informações sobre a metodologia empregada, levantamento dos dados, apuração dos dados entre outras.

Observações importantes:

- A tabela não pode ser fechada nas laterais, mas pode ter traços verticais separando as colunas;
- A tabela deve ser delimitada por traços horizontais, mas não pode haver traços horizontais no corpo da tabela.

Exemplo:

Distribuição dos leitos da rede pública no Brasil, de acordo com a esfera administrativa. Brasil, 2005

Esfera administrativa	Número de leitos	%
Federal	17.189	11,54%
Estadual	.699	41,42%
Municipal	70.078	47,04%
Total:	148.966	100,00

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de População e Indicadores Sociais, Pesquisa de Assistência Médico-Sanitária - 2005.

4.1 Apresentação dos dados qualitativos

Para organizarmos dados qualitativos, precisamos contar quantas respostas de cada categoria foram obtidas.

Neste exemplo, vamos observar uma tabela com dados fictícios e organizaremos os dados coletados para a variável "Grau de satisfação" com o curso escolhido.

Inicialmente, temos que escrever um título que vá especificar o que será apresentado na tabela. Neste caso poderia ser: Grau de satisfação com o curso escolhido de 30 estudantes da Faculdade X, em 20/5/2010.

Em seguida, fazemos o cabeçalho e a coluna indicadora. Neste caso, a coluna indicadora vai conter as categorias MS, S e I que representam muito satisfeito, satisfeito e insatisfeito respectivamente.

O cabeçalho vai conter na primeira coluna Grau de Satisfação, na segunda coluna Número de estudantes e na terceira coluna a frequência relativa expressa em percentual.

A contagem deve ser feita sequencialmente a partir do ID = 1.

Vamos anotando cada resposta ao lado da categoria com um risco e no final contamos quantos riscos cada categoria obteve.

Podemos marcar cada grupo de cinco riscos para facilitar a contagem. Este procedimento facilita a contabilização total e permite que paradas no processo de contagem não impliquem em ter de recomeçar tudo.

A frequência relativa é muito útil para nos fornecer a proporção de unidades de cada categoria, podendo ser expressa em percentual. O cálculo da frequência relativa é obtido dividindo a frequência absoluta, que é a quantidade de dados de determinado valor, pelo total de valores. Se desejarmos expressá-la em percentagem, multiplica-se o resultado por 100.

$$\text{Frequência relativa (\%)} = \frac{\text{Frequência absoluta}}{\text{Tamanho da amostra}} \times 100\%$$

No nosso exemplo, o percentual que corresponde ao de alunos satisfeitos é de:

$$Fr\ (\%) = \frac{9}{30} \times 100\% = 30{,}0\%$$

Veja a tabela completa e observe que os riscos de auxílio para a contagem dos dados foram retirados.

Grau de satisfação com o curso escolhido de 30 estudantes da Faculdade X, em 20/5/2010

Grau de satisfação	Número de estudantes	Percentual
MS	9	30,0%
S	16	53,3%

I	5	16,7%
Total	30	100%

Fonte: Tabela com dados fictícios

4.2 Apresentação de dados quantitativos discretos

Quando queremos apresentar dados quantitativos discretos, devemos escrever os dados em ordem crescente na coluna indicadora e contamos quantas vezes cada dado se repetiu. Veja um exemplo de tabela para dados discretos:

Grau de satisfação com o curso escolhido de 30 estudantes da Faculdade X, em 20/5/2010.

Tamanho da Família	Número de estudantes	Percentual
1	2	6,7%
2	4	13,3%
3	8	26,7%
4	10	33,3%
5	5	16,7%
6	1	3,3%
Totais	**30**	**100%**

Fonte: Tabela com dados fictícios

Se os dados discretos tiverem uma amplitude muito grande não vamos poder apresentá-los em tabelas de dados isolados, como o exemplo anterior, porque teríamos uma quantidade de linhas muito grande.

Para estes casos os dados vão ser apresentados em tabelas de distribuição de frequências, assunto a ser abordado no próximo capítulo.

4.2 Atividade Proposta

1. Qual parte da tabela deve responder a três perguntas: O quê?, Quando?, Onde?

2. Quais são os erros da tabela abaixo?

Quantidade de Filhos	
Filhos	F
0	5
1	10
2	21
3	10
4	5
Total	30

5 ORGANIZAÇÃO DE DADOS: DADOS ISOLADOS E AGRUPADOS COM INTERVALOS DE CLASSES

Quando temos uma quantidade muito grande de dados quantitativos discretos com grande variação de valores ou quando os dados são quantitativos contínuos, não podemos organizá-los em tabelas de dados isolados, como vimos no capítulo anterior. Vamos precisar agrupá-los com intervalos de classe.

Regras para formatação

As regras para formatação de tabelas vistas na aula anterior permanecem válidas para as tabelas de distribuição de frequências, a diferença vai estar na forma de escrever a coluna indicadora e fazer a contagem dos dados. A seguir, são apresentados os passos para a formação da coluna indicadora.

Passo 1 - Encontre o maior e o menor valor da distribuição de dados

Passo 2 - Calcule a amplitude dos dados: Encontramos a amplitude dos dados na diferença entre o maior e o menor valor.

Passo 3 - Defina o número de classe: Não existem regras rígidas para definir o número de classes, mas de uma maneira geral, este número não deve ser menor do que 5 classes nem maior do que 20 classes. Um número muito pequeno de classes ocasiona muita perda de informação e um número muito grande vai dificultar o entendimento. Lembre-se de que a tabela deve facilitar a apresentação da informação.

Existem algumas fórmulas para o cálculo do número de classes. No nosso exemplo, usaremos a fórmula: $\mathbf{K \cong \sqrt{n}}$, onde K é o número de classes desejado e "n" é o número de dados da amostra.

Por exemplo, se uma amostra possui 30 elementos (dados), o número de classes que utilizaremos será calculado assim:

$$K \cong \sqrt{30}, \text{ ou seja, } k \cong 5{,}48$$

Como o número de classes, necessariamente, tem de ser um número inteiro, podemos escolher, 5 classes ou 6 classes.

Na equação anterior foi utilizado o símbolo ≅ (aproximadamente igual a) ao invés do símbolo da igualdade porque representam valores típicos a serem usados, porém não são rígidos, além do valor k ser obrigatoriamente um valor inteiro. Número de intervalos não pode ser um número fracionário.

Passo 4 - Calcule a amplitude: Sabendo o tamanho da amplitude total e o número de classes, calculamos a amplitude h (tamanho) de cada classe.

Por exemplo, se na mesma amostra anterior escolhemos o $k = 5$ e se nossa Amplitude Total for igual a 38, calcularemos assim o tamanho (intervalo) de cada classe.

$h \cong At/k$, ou seja, $h \cong 38/5$, onde teremos $h \cong 7{,}6$, porém como devemos sempre arredondar o tamanho (intervalo) da classe para cima, optaremos por $h \cong 8$.

Passo 5 - Organize as classes: Organizaremos as classes de maneira que a primeira classe contenha o menor valor observado.

Cada intervalo de classe tem seu limite inferior (li) e um limite superior (Li). No nosso caso, os intervalos de classe poderiam ser:

152 |--- 160

160 |--- 168

168 |--- 176

176 |--- 184

184 |--- 192

O símbolo |--- significa que o intervalo de classe é fechado à esquerda e aberto à direita, ou seja, o limite inferior faz parte da contagem dos dados e o limite superior não.

Neste caso, o limite inferior da primeira classe é li=152 e o limite superior da primeira classe é Li=160.

Na primeira classe do exemplo, como o intervalo é fechado à esquerda e aberto à direita, o valor 152 é contado dentro da classe, e o valor 160, na próxima, em que passa a ser o limite inferior da segunda classe.

Poderíamos, também, usar os intervalos de classe abaixo:

151 ---| 159

159 ---| 167

167 ---| 175

175 ---| 183

183 ---| 191

Neste caso, o símbolo ---| significa o oposto, ou seja, que o intervalo é aberto à esquerda e fechado à direita. O limite inferior não entra na contagem dos dados neste intervalo, mas o limite superior entra.

Pode-se usar qualquer uma das convenções, desde que cada dado só seja contado uma vez. Um mesmo dado não pode ser representado em dois intervalo de classe diferentes.

Passo 6 - Conte os dados: Definidos os intervalos, agora é só fazer a contagem dos dados de acordo com a convenção escolhida.

Por exemplo: Vejamos como ficaria uma tabela de distribuição de frequências com uma amostra das alturas, em centímetros, dos 30 estudantes da Faculdade X.

Adotando a primeira convenção.

Utilizaremos, k = 5 e h = 6, conforme calculado anteriormente.

Altura, em centímetros, dos 30 estudantes da Faculdade X pesquisados em 20/05/2010.

Altura	Número de estudantes	Percentual
152 |--- 160	4	13,3%

160 |--- 168	3	10,0%
168 |--- 176	8	26,7%
176 |--- 184	5	16,7%
184 |--- 192	10	33,3%
Total	**30**	**100,0%**

Fonte: Dados fictícios - Distribuição de frequências com intervalos fechado-aberto

Agora adotando a segunda convenção.

Altura, em centímetros, dos 30 estudantes da Faculdade X pesquisados em 20/05/2010.

Altura	Número de estudantes	Percentual
151 ---| 159	4	13,3%
159 ---| 167	3	10,0%
167 ---| 175	8	26,7%
175 ---| 183	5	16,7%
183 ---| 191	10	33,3%
Total	**30**	**100,0%**

Fonte: Dados fictícios - Distribuição de frequências com intervalos fechado-aberto

Os percentuais foram calculados como explicado no capítulo anterior, ou seja:

Fr(%)=(Frequência absoluta da classe/Total dos dados) × 100%

5.1 Atividade Proposta

Uma amostra das idades, dos 30 estudantes da Faculdade X. Resultou nestes dados:

19	19	19	20	20
20	20	20	22	22
22	25	25	27	27
28	32	32	32	32

33	35	39	40	43
47	51	55	55	58

1. Construa uma tabela de distribuição de frequências para as idades dos 30 alunos, considerando **k = 5**.

2. Agora, com os mesmos dados considere **k = 6** e construa uma tabela de distribuição de frequências.

3. Observe a tabela abaixo:

Nível de colesterol total no sangue dos funcionários da Empresa XYZ em outubro de 2010. (em mg/dl)

Nível de colesterol total	Número de funcionários	Percentual
180 |--- 190	3	5,5%
190 |--- 200	5	9,1%
200 |--- 210	8	14,5%
210 |--- 220	10	18,2%
220 |--- 230	12	21,8%
230 |--- 240	10	18,2%
240 |--- 250	7	12,7%
Total:	55	100,0%

Agora, com base na tabela, responda os itens abaixo:

a) Qual é o tamanho de cada classe (h)?

b) Quantas classes tem esta tabela?

c) Pela nomenclatura utilizada, o limite superior entra na contagem de cada classe?

6 GRÁFICOS

Uma forma de apresentar os dados estatísticos de maneira a proporcionar uma visão mais imediata do fenômeno é através de gráficos. Os gráficos têm a vantagem de apresentar o fenômeno de forma mais visual, proporcionando um entendimento mais fácil do fenômeno que está sendo representado.

6.1 Principais tipos de gráficos

Os principais tipos de gráfico são:

Cartogramas

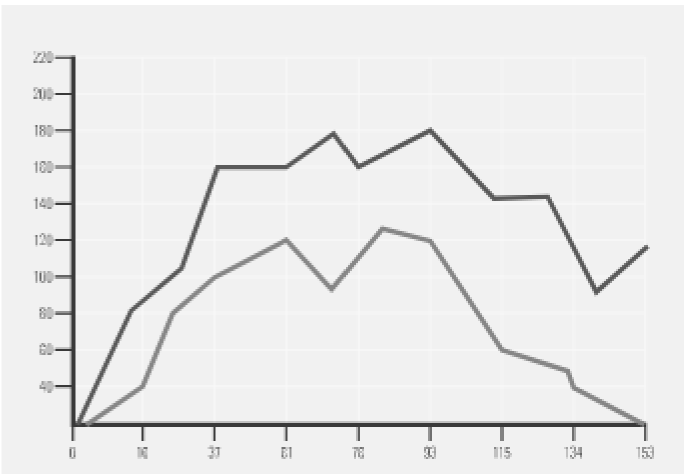
Os cartogramas representam a informação sobre um mapa.

Pictogramas

Os pictogramas utilizam imagens representativas do tema da informação a ser apresentada para construir o gráfico.

Diagramas

Os diagramas são gráficos geométricos, que em geral utilizam o sistema cartesiano para representação da informação.



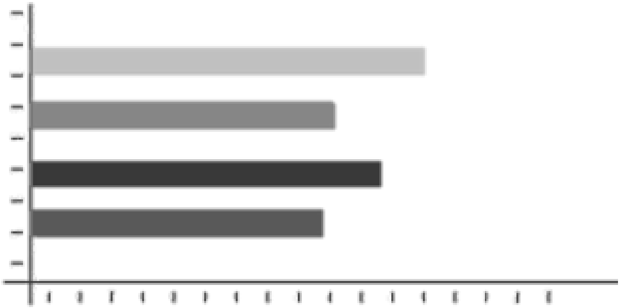
Gráficos em barras

Os gráficos em barras são utilizados para apresentar informações de variáveis qualitativas, sendo compostos por retângulos dispostos horizontalmente (em barras) ou verticalmente (colunas).

Barras horizontais

No caso das barras serem horizontais, a altura de todas as barras é a mesma e o comprimento vai ser dado pela frequência ou pela frequência relativa (normalmente em porcentagem) da categoria.

Exemplo de gráfico de barras horizontais

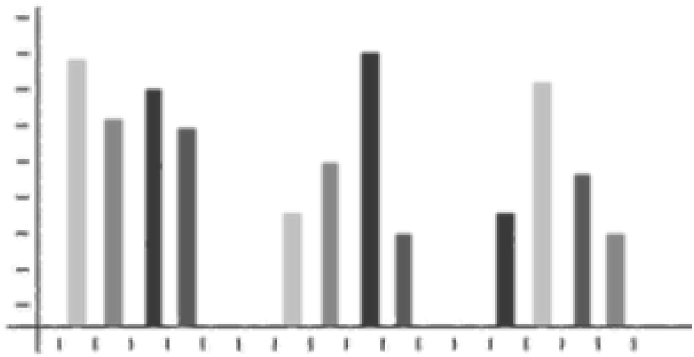


Fonte: próprio autor.

Barras verticais

No caso das barras serem verticais, a base será a mesma para todas as barras e a altura será dada pela frequência ou frequência relativa.

Exemplo de gráfico de barras verticais



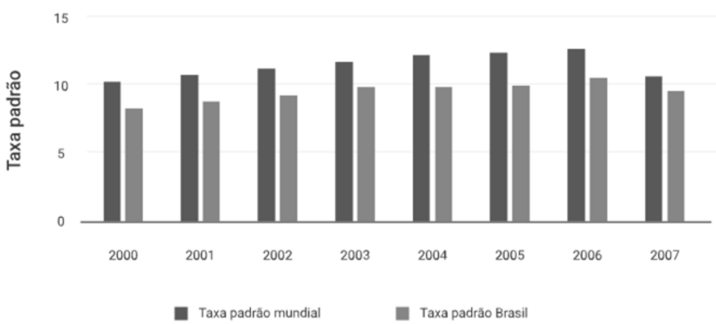
Fonte: próprio autor.

Este procedimento vai garantir a proporcionalidade entre as áreas dos retângulos e os dados. As distâncias entre as barras não deverá ser menor que a metade nem maior do que dois terços da altura das barras ou da base das colunas respectivamente.

Sempre que os textos forem extensos deve-se dar preferência ao gráfico de barras horizontais em relação ao gráfico de barras verticais.

Os gráficos em barras também podem apresentar mais de uma série ao mesmo tempo, como quando queremos comparar uma mesma variável em relação aos sexos masculino e feminino. Veja o exemplo:

Taxas padrão de mortalidade por câncer de mama, pelas populações mundial e brasileira, por 100.000 mulheres, entre 2000 e 2007



Fonte: IBGE (2010)

Este gráfico vai apresentar duas séries num mesmo gráfico, sendo que as duas séries apresentam informações sobre uma mesma variável (taxa padrão de mortalidade) para duas regiões diferentes, uma de abrangência nacional e a outra de abrangência mundial.

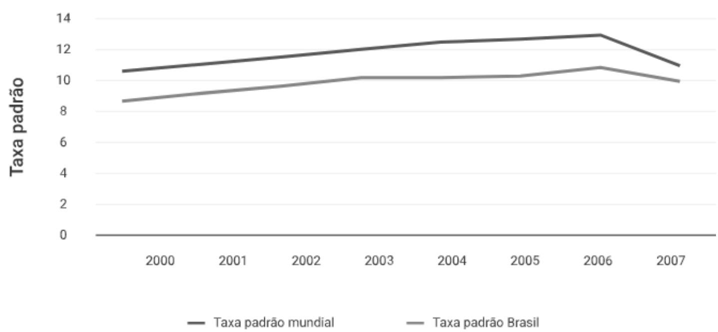
O gráfico em barras é bastante utilizado para apresentar séries temporais, porém quando o número barras começa a crescer muito é mais indicado utilizar os diagramas em linha.

Gráficos em linhas

Este tipo de gráfico apresenta os dados em pares ordenados utilizando os eixos cartesianos xy. No eixo x (eixo horizontal) colocam-se os dados da variável e no eixo y (vertical) colocam-se as frequências observadas correspondentes.

O gráfico é feito ligando-se estes pontos. Este tipo de gráfico é muito útil para comparação entre dois fenômenos e também para séries temporais mais longas, onde o gráfico em barras não vai mais ser indicado.

Taxas padrão de mortalidade por câncer de mama, pelas populações mundial e brasileira, por 100.000 mulheres, entre 2000 e 2007



Fonte: IBGE (2010)

Para dados discretos, pode-se usar o diagrama de linhas, que é diferente do diagrama em linhas recém-apresentado.

No diagrama de linhas, para cada par (dado, frequência) existe uma coluna bastante estreita como se fosse uma linha. Os pontos não são ligados para não passar a ideia de continuidade, já que os dados são discretos.

Gráficos em setores

Os gráficos em setores são indicados para variáveis qualitativas quando temos uma quantidade pequena de categorias e queremos comparar a participação de cada categoria com o total.

Veja a tabela abaixo:

Quantidade de internações por Região no período de abril/2010

Região	Internações
Centro-oeste	74.496
Norte	74.971
Sul	155.303
Nordeste	263.739
Sudeste	362.615
Total	**931.124**

Fonte: Ministério da Saúde - Sistema de Informações Hospitalares do SUS (SIH/SUS) - Site: http://tabnet.datasus.gov.br/ (agosto, 2010)Nota: Situação da base de dados nacional em 01/06/2010, sujeita a novas atualizações.

A construção de um gráfico em setores se dá com o desenho de um círculo e a determinação dos ângulos de cada setor através de uma regra de três simples e direta. Vamos construir um gráfico de setores com os dados da tabela acima.

Uma circunferência completa tem 360º, então para encontrarmos o ângulo de cada setor fazemos a correspondência de 360º com o total e o ângulo de cada setor com a frequência correspondente deste mesmo setor. No caso deste nosso exemplo:

360° — 931.124

Ângulo Sudeste — 362.615

Temos uma proporção onde aplicamos a regra fundamental das proporções que é o produto dos meios é igual ao produto dos extremos.

Ângulo Sudeste x 931124 = 360 x 362615

Ângulo Sudeste = 130541400/931124

Ângulo Sudeste ≅ 140º

360° — 931.124

Ângulo Nordeste — 263.739

Ângulo Nordeste x 931124 = 360 x 263739

Ângulo Nordeste = 94946040/931124

Ângulo Nordeste ≅ 102º

360° — 931.124

Ângulo Norte — 74.971

Ângulo Norte x 931124 = 360 x 74971

Ângulo Norte = 26989560/931124

Ângulo Norte ≅ 29º

360° — 931.124

Ângulo Sul — 155.303

Ângulo Sul x 931124 = 360 x 155303

Ângulo Sul = 55909080/931124

Ângulo Sul ≅ 60º

360° — 931.124

Ângulo Centro Oeste — 74.496

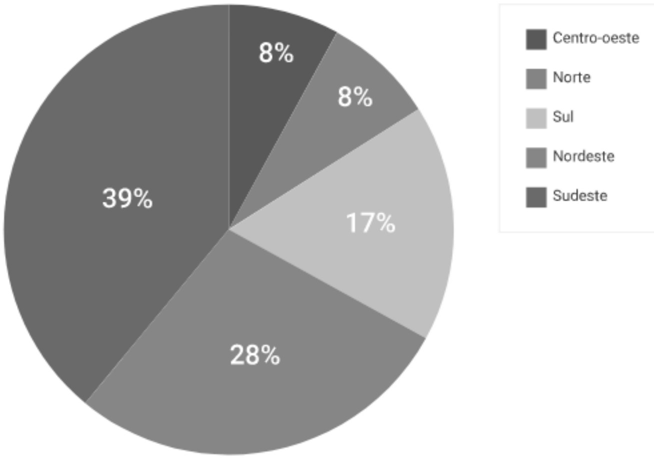
Ângulo Centro Oeste x 931124 = 360 x 74496

Ângulo Centro Oeste = 26818560/931124

Ângulo Centro Oeste ≅ 29º

Com todos os ângulos, basta desenhar um círculo e marcar os ângulos que teremos os setores proporcionais:

Internações, por Região do Brasil, no período de abril/2010



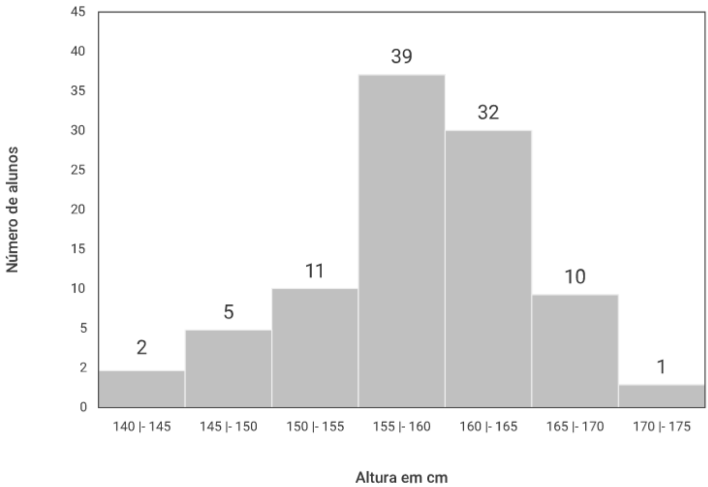
Fonte: site do DATASUS (2010)

Histogramas

O histograma é a representação gráfica de uma tabela de distribuição de frequências com intervalos de classe. Este gráfico é parecido com o gráfico em barras verticais, porém não existe espaço entre as colunas, para dar noção de continuidade.

Para construir o histograma, utiliza-se o sistema de eixos cartesianos, onde os intervalos de classe são representados no eixo horizontal e onde cada coluna tem a altura correspondente à frequência de seu intervalo de classe correspondente. Veja o exemplo:

Estrutura de 100 alunos da Escola X, em 2008



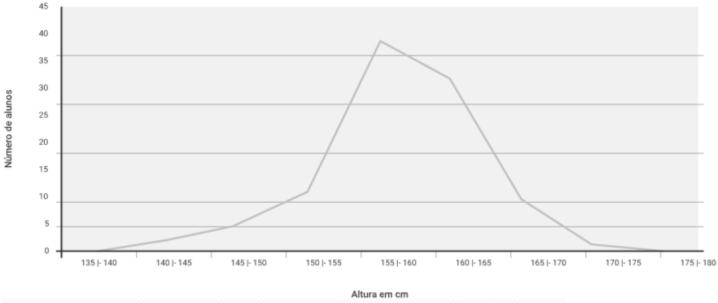
Fonte: próprio autor

Polígono de frequências

Os mesmos dados que estão no histograma podem ser apresentados no gráfico chamado de polígono de frequências. Neste gráfico, acrescenta-se uma classe antes e outra depois das classes originais da tabela de distribuição de frequências. Estas duas classes não contêm dados, mas servirão para "aterrissar o gráfico", ou seja, permitir que o gráfico toque o eixo x nas extremidades.

O polígono de frequências é desenhado marcando o ponto médio de cada coluna do histograma de frequências, mais o ponto médio das duas classes acrescentadas nas extremidades. Em seguida, ligam-se estes pontos e apaga-se o desenho do histograma que serviu de base para marcar os pontos médios. Veja um exemplo de polígono de frequências utilizando os mesmos dados apresentados para o exemplo de histograma.

Estrutura de 100 alunos da Escola X, em 2008



Fonte: próprio autor

6.2 Atividade Proposta

Observe a tabela:

Evolução dos registros de nascimentos.

Brasil 1996-2006

Ano	Número de registros
1996	2.627.272
1997	2.620.514
1998	2.699.407
1999	2.939.278
2000	2.861.748
2001	2.779.268
2002	2.803.054
2003	2.814.763
2004	2.813.704
2005	2.874.753
2006	2.799.128

Fonte: Site do IBGE (2010)

Qual dos gráficos abaixo é o mais indicado para representar a série da tabela.

a) Histograma

b) Polígono de frequências

c) Gráfico de setores

d) Gráfico em barras

e) Diagrama em linhas

7 MEDIDAS DE TENDÊNCIA CENTRAL COM DADOS ISOLADOS E AGRUPADOS SEM INTERVALO DE CLASSE

Além da organização de dados em tabelas e em gráficos, como vimos nos capítulos anteriores, é muito útil podermos representar uma grande quantidade de dados com um único valor. As medidas de tendência central calculam o que é médio ou típico de uma distribuição de valores. As médias de tendência central que veremos são: Moda, Média Aritmética e Mediana.

7.1 Moda

É o valor de maior frequência em uma distribuição de dados. A medida de tendência central mais simples é a moda. Podemos fazer analogia com a moda em relação ao vestuário. Quando dizemos que uma cor está na moda é porque quando saímos de casa vemos muitas pessoas vestidas com aquela cor, ou seja, aquela cor se sobressai.

Podemos ter moda para qualquer tipo de variável e nível de mensuração. Uma distribuição de frequências pode ser classificada em relação à moda como:

- **Curva modal**: A característica da curva modal é ter um único valor que se sobressai, ou seja, ter apenas uma moda.
- **Curva não modal**: Nas distribuições não modais todos os valores têm a mesma frequência e nas distribuições amodais não existe moda.
- **Curva bimodal/multimodal**: As distribuições bimodais e multimodais são aquelas onde temos 2 ou mais dados que se sobressaem respectivamente.

- **Curva antimodal**: Nas distribuições antimodais temos uma ante moda, que é o inverso da moda. Nestas distribuições o que é mais característico é o valor de menor frequência.

Veremos, agora, dois exemplos de "cálculo" de moda.

Exemplo 1

Quantidade de internações por Região no período de abril/2010

Esfera Administrativa	Número de leitos	%
Federal	17.189	11,54%
Estadual	61.699	41,42%
Municipal	70.078	47,04%
Total	**148.966**	**100%**

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de População e Indicadores Sociais, Pesquisa de Assistência Médico-Sanitária 2005.

Exemplo 2

Tamanho da família de 30 estudantes pesquisados na Faculdade X em 20/05/2010

Tamanho da família	f	%
1	2	6,7%
2	4	13,3%
3	8	26,7%
4	10	33,3%
5	5	16,7%
6	1	3,3%
Total	**30**	**100%**

Fonte: Dados fictícios

Repare que a moda é o valor que aparece com maior frequência e não o valor da frequência. Desta forma, a distribuição do exemplo 1 tem Mo = Municipal (não é 70.078) e a distribuição do exemplo 2 tem Mo = 4 (não é 10).

7.2 Média aritmética

A média aritmética é calculada de forma que a soma de todos os desvios em relação à média é igual a zero. Podemos dizer que a média aritmética é o centro da gravidade da distribuição. O símbolo utilizado para medição aritmética é $\overline{X}$.

A média aritmética é a medida de tendência central mais conhecida e utilizada, sendo calculada somando todos os valores e dividindo o resultado pela quantidade total de valores.

Agora, veremos o procedimento para cálculo da média aritmética.

$$\overline{X} = \frac{\sum_{i}^{n}=1\ Xi}{n}$$

$\overline{X}$ Símbolo da média aritmética

$\sum_{i}^{n}$ Xi = X1 + X2 + X3 + ... + Xn Somatório de todos os valores X desde o primeiro dado ($i = n$)

X Dados da variável em questão

i Índice de cada dado

n Número total de elementos

A média aritmética só pode ser calculada para dados quantitativos, pois não existe sentido em somar valores de dados qualitativos.

Exemplo1: Um professor perguntou a seus 10 alunos quantos irmãos cada um tinha. Ele obteve as seguintes respostas: 1, 3, 0, 2, 1, 1, 0, 4, 2, 3. Quantos irmãos, em média, os alunos deste professor tem?

Neste exemplo, X1=1, X2=3, X3=0 e assim sucessivamente.

Aplicando a fórmula, temos:

$\overline{X}$ = ∑in Xi = (1 + 3 + 0 + 2 + 1 + 1 + 0 + 4 + 2 + 3)/10

= 17/10 = 1,7

Se temos valores repetidos, podemos utilizar a fórmula para dados agrupados:

$$\overline{X} = \frac{\sum f_i X_i}{n}$$ fi é o valor da frequência absoluta de cada X_i.

No caso de utilização de tabelas prontas, o índice i corresponde a cada linha da tabela.

Nesse exemplo poderíamos ter feito o seguinte cálculo:

$$\overline{X} = \sum_{i}^{5}{}_{1} \frac{2x0-3x1-2x2+2x3-1x4}{10} = \frac{17}{10} = 1,7$$

Onde i varia de 1 a 5. Pois, temos 5 valores diferentes de dados: 0 a 4, que corresponderiam a 5 linhas em uma tabela, caso os dados estivessem dispostos em uma tabela.

Exemplo2: Vamos calcular a média aritmética da distribuição apresentada na tabela abaixo:

Tamanho da família de 30 estudantes pesquisados na Faculdade X em 20/05/2010

Tamanho da família	f	%
1	2	6,7%
2	4	13,3%
3	8	26,7%
4	10	33,3%
5	5	16,7%
6	1	3,3%
Total	**30**	**100%**

Fonte: Dados fictícios

Neste caso, podemos acrescentar uma coluna auxiliar na tabela para calcularmos os valores f_iX_i da tabela em cada linha.

Na parte inferior da tabela, nesta mesma coluna, colocamos o somatório dos resultados da linha.

O valor da média será calculado dividindo este somatório pelo total de valores (somatório da coluna f).

Desta forma, os cálculos ficaram:

$$\overline{X} = \frac{\sum_i^6 f_i X_i}{n} = \frac{105}{30} = 3,5$$

Propriedades da média aritmética:

- A soma das discrepâncias é igual a zero. Onde discrepância é a diferença de cada valor bruto menos a média aritmética, tendo como símbolo a letra "x" minúscula.

- Somando-se, ou subtraindo-se, todos os dados de uma distribuição por um mesmo valor, a nova média também será acrescida, ou diminuída deste mesmo valor.

- Multiplicando-se, ou dividindo-se, todos os dados de uma distribuição por um mesmo valor, a nova média também será multiplicada, ou dividida, por este mesmo valor.

Conferindo cada propriedade com os dados do exemplo 1 temos:

Calculando todas as discrepâncias (lembrando que $x = X_i - \overline{\mathbf{X}}$, onde X_i é cada valor bruto e $\overline{\mathbf{X}} = 1{,}7$):

(1 - 1,7), (3 - 1,7), (0 - 1,7), (2 - 1,7), (1 - 1,7), (1 - 0,7), (0 - 1,7), (4 - 1,7), (2 - 1,7), (3 - 1,7) = - 0,7; 1,3; - 1,7; 0,3; - 0,7; - 0,7; - 1,7; 2,3; 0,3; 1,3;

Somando as discrepâncias:

(-0,7)+(1,3)+(-1,7)+(0,3)+(-0,7)+(-0,7)+(-1,7)+(2,3)+(0,3)+(1,3)=0

Resultado: A soma das discrepâncias é igual a zero, comprovando a propriedade 1.

Vamos somar um valor constante (por exemplo, 2) a todos os valores e recalcular a média.

Valores: (1 + 2), (3 + 2), (0 + 2), (2 + 2), (1 + 2), (1 + 2), (0 + 2), (4 + 2), (2 + 2), (3 + 2).

Calculando a média:

$$\overline{X} = \sum_{i=1}^{10} \frac{3+5+2+4+3+3+2+6+4+5}{10} = \frac{37}{10} = 3,7$$

Comparando com a média anterior, vemos que:

$\overline{X} = 2 + \overline{X}$, pois 3,7 = 2 + 1,7

Comprovamos a propriedade 2.

Agora, vamos multiplicar por 2 todos os valores e recalcular a média.

Valores: (1 x 2), (3 x 2), (0 x 2), (2 x 2), (1 x 2), (1 x 2), (0 x 2), (4 x 2), (2 x 2), (3 x 2).

Calculando a média:

$$\overline{X} = \sum_{i=1}^{10} \frac{2+6+0+4+2+2+0+8+4+6}{10} = \frac{34}{10} = 3,4$$

Comparando com a média anterior, vemos que:

$\overline{X} = 2 + \overline{X}$, pois 3,4 = 2 x 1,7

Comprovamos a propriedade 3.

7.3 Mediana

Mediana é o valor central de uma distribuição de valores, sendo representada por **Md**.

A mediana divide os dados de uma distribuição ordenada em dois subgrupos com igual quantidade de elementos, sendo um subgrupo composto de 50% dos valores menores ou iguais à mediana e o outro subgrupo composto de 50% dos valores maiores ou iguais à mediana.

Para o cálculo da mediana de dados isolados vamos utilizar duas situações:

<u>Situação 1</u>

A quantidade de dados é par, não existindo um determinado termo no centro da distribuição. Nesta situação, o valor da mediana será calculado através da média aritmética dos dois termos centrais.

Vamos chamar os dois termos centrais de TC_1 e TC_2. Para sabermos os valores dos termos centrais, temos que ordenar os valores da distribuição e depois calcularmos a posição de cada um deles.

- Posição de $TC_1 = n/2$
- Posição de $TC_2 = (n+1)/2$

Sabendo a posição de cada um deles, encontramos seus valores e podemos calcular a mediana.

Mediana para n par = (valor de TC_1+ valor de TC_2)/2

Situação 2

A quantidade de dados é ímpar e temos um determinado termo no centro da distribuição. Neste caso, a mediana é o valor deste termo central.

Para encontrarmos o termo central, inicialmente ordenamos os valores e depois calculamos a posição do termos central (PTC).

Mediana para n ímpar = Valor do Próprio Termo Central.

Vejamos os seguintes exemplos:

Exemplo1: Vamos repetir o exemplo dado para o cálculo da média aritmética. Um professor perguntou a seus 10 alunos quantos irmãos cada um tinha, obtendo as seguintes respostas: 1, 3, 0, 2, 1, 1, 0, 4, 2, 3.

Qual é a mediana desta distribuição?

Inicialmente, temos que colocar os valores ordenados: 0, 0, 1, 1, 1, 2, 2, 3, 3, 4.

Observamos que temos um número par de dados (n = 10), logo dois termos centrais TC_1 e TC_2. Aplicando as fórmulas, encontramos a posição de cada um deles.

$TC_1 = 10/2 = 5^a$ Posição

$TC_2 = (10/2) + 1 = 6^a$ Posição

Olhando para a distribuição ordenada, vemos que o valor 1 ocupa a quinta posição e o valor 2 ocupa a sexta posição, logo:

Mediana = (1 + 2)/2 = 1,5

Exemplo2: Supondo que este professor tivesse apenas 9 alunos e que as respostas foram: 0, 0, 1, 1, 1, 2, 2, 3, 3.

PTC = (n + 1)/2 = (9 + 1)/2 = 10/2 = 5ª posição

Md = valor da 5º posição.

Md = 1

Estes dois exemplos tinham poucos dados, mas e se a minha distribuição estivesse sendo apresentada numa tabela de dados isolados? Como calcular a mediana?

Nestes casos, fazemos uma coluna auxiliar com as frequências acumuladas linha a linha. A partir desta coluna, encontramos a posição de cada valor da distribuição e calculamos a mediana.

Exemplo 3: Calcule a mediana da distribuição já apresentada anteriormente para o cálculo da moda e média.

a) Inicialmente vemos que n = 30 (par). Logo, temos que encontrar os dois termos centrais TC_1 e TC_2 para podermos calcular a mediana.

b) Calculamos a posição destes dois termos encontrando:

Posição de TC1 = 30/2 = 15ª posição

Posição de TC1 = (30/2) + 1 = 16ª posição

c) Sabemos que as tabelas já são construídas com os valores ordenados, mas temos que saber qual é o valor correspondente a 15ª e 16ª posições. Para isto, construímos a coluna das frequências acumuladas fac. Inicialmente, repetimos o valor da primeira frequência absoluta (neste caso, 2). Em seguida, somamos este valor (2) com o da próxima frequência absoluta (4), obtendo 6. Repetimos este procedimento, somando o resultado 6 com a próxima frequência absoluta 8, obtendo 14. E assim sucessivamente até a última linha de dados.

Repare que o valor da última célula da coluna de frequências acumuladas necessariamente tem que ser igual a n (no nosso exemplo igual a 30).

Tamanho da família de 30 estudantes pesquisados na Faculdade X em 20/05/2010

Tamanho da família	f	%	fac
1	2	-	2
2	4	-	2 + 4 = 6
3	8	-	6 + 8 = 14
4	10	-	14 + 10 = 24
5	5	-	24 + 5 = 29
6	1	-	29 + 1 = 30
Total	**30**	**-**	**

Fonte: Dados fictícios

d) A partir da coluna fac, temos como saber a posição que cada valor da tabela. A primeira célula da coluna fac mostra que o valor desta linha, no nosso exemplo $X_1 = 1$ ocupa desde a primeira posição até a posição correspondente ao valor desta célula (neste exemplo: desde a primeira posição até a segunda posição).

7.4 Posição relativa da média, mediana e moda

Podemos caracterizar uma distribuição como simétrica ou assimétrica:

Distribuição simétrica

As distribuições simétricas possuem valores de Moda, Média e Mediana iguais.

Distribuições simétricas: $\overline{\mathbf{X}}$ = Mo = Md

Distribuição assimétrica

Já nas distribuições assimétricas estes valores são diferentes. As distribuições assimétricas são classificadas em distribuição assimétrica positiva (ou à direita) e distribuição assimétrica negativa (ou à esquerda).

Distribuições assimétricas positivas: $\bar{X}$ Md > Mo

Distribuições assimétricas negativas: $\bar{X}$ < Md < Mo

Note que a mediana sempre se encontra entre a média e a moda.

7.4 Atividade Proposta

A partir da tabela a seguir, calcule:

Número de gols dos jogos da primeira fase do campeonato nacional do país XPTO no ano de 1999.

Número de gols	Jogos	%
0	5	20,8%
1	8	33,3%
2	3	12,5%
3	2	8,3%
4	1	4,2%
Total	**24**	**100%**

Fonte: Dados fictícios

a) A moda

b) A média

c) A mediana

8 MEDIDAS DE TENDÊNCIA CENTRAL COM DADOS AGRUPADOS EM INTERVALO DE CLASSE

Nesta aula, vamos ver como calcular as medidas de tendência central Moda, Média e Mediana para dados organizados em distribuições de frequências.

8.1 Moda para dados agrupados em intervalos de classe

Quando temos os valores organizados em intervalos de classe vamos ter que estimar qual seria o valor da moda dentro do intervalo de classe de maior frequência, chamado de classe modal.

A moda bruta vai ser calculada pelo valor central da classe modal, sem levar em consideração os valores das frequências das classes anterior e posterior.

$Mo = (li + Li)/2$

li = Limite inferior da classe modal.

Li = Limite superior da classe modal.

8.2 Média para dados agrupados em intervalos de classe

A média para dados agrupados é calculada de forma semelhante à calculada para dados isolados, sendo Xi substituído pelo valor médio de cada intervalo de classe.

Desta forma, construímos uma coluna auxiliar com os valores médios de cada intervalo de classe e calculamos a média como anteriormente.

Veja a fórmula:

$$\overline{X} = \frac{\sum f_i . X_i}{n}$$

8.3 Mediana para dados agrupados em intervalos de classe

A estimativa do valor da mediana a partir de sua classe é dada pela seguinte fórmula:

$$\mathrm{Md} = l_\mathrm{i} + \frac{\frac{\mathrm{n}}{2} - \mathrm{f}_{\mathrm{ai}-1}}{\mathrm{f}_\mathrm{i}}$$

li = Limite inferior da classe da mediana.

n = Número total de valores.

fai−1 = Frequência acumulada da classe anterior à classe da mediana.

fi = Frequência acumulada da classe anterior à classe da mediana.

A classe da mediana é aquela em que estão contidos 50% dos valores. Ela pode ser apresentada fazendo-se uma coluna de frequências acumuladas em percentual e identificando a primeira linha em que a frequência acumulada em percentual é igual ou superior a 50%.

Veja a tabela:

Altura, em centímetros, dos 30 estudantes da Faculdade X pesquisados em 20/05/2010

Idade	Número de estudantes	Percentual	fac%	fac
151 |- 159	4	13,3%	13,3%	4
159 |- 167	3	10,0%	23,3%	7
167 |- 175	8	26,7%	50,0%	15
175 |- 183	5	16,7%	66,7%	20
183 |- 191	10	33,3%	100%	30
Total	**30**	**100%**	**-**	**-**

Fonte: Dados fictícios

Vamos entender como é feito o cálculo da mediana de distribuição.

Primeiro, devemos identificar a classe da mediana. Para isso, criaremos uma coluna de frequências acumuladas.

Observe que, na tabela acima, temos exatamente o valor de 50% na coluna de frequências acumuladas em percentual no terceiro intervalo de classe.

Com a classe da mediana identificada podemos construir a coluna de frequências absolutas acumuladas.

Agora, vamos aplicar os resultados na fórmula:

$$\mathrm{Md} = l_i + \frac{\frac{n}{2} - f_{ai-1}}{f_i}$$

Onde:

li = 167cm.

h = 175 - 167 = 8

n/2 = 30/2 = 15

fa-1 = 7

fi = f3 = 8

$$\mathbf{Md = 167 + \frac{8(15-7)}{8} \therefore Md = 167 + \frac{8(8)}{8} \therefore Md = 167 + 8 \therefore Md = 175\,cm}$$

Repare que o valor da mediana, neste caso, foi exatamente igual ao limite superior da classe da mediana. Isto sempre ocorrerá quando o valor da frequência acumulada em percentual for exatamente igual a 50%.

8.4 Atividade Proposta

1. Calcule a moda bruta e a moda para a tabela a seguir.

Idade dos 30 estudantes da Faculdade X pesquisados em 20/05/2010

Idade	Número de estudantes	Percentual
19 |- 26	13	43,3%
26 |- 33	7	23,3%
33 |- 40	3	10%
40 |- 47	3	10%
47 |- 54	2	6,7%
54 |- 61	2	6,7%
Total	**30**	**100%**

Fonte: Dados fictícios

2. Calcule a média aritmética para a distribuição da tabela a seguir.

Idade dos 30 estudantes da Faculdade X pesquisados em 20/05/2010

Idade	Número de estudantes
19 |- 26	13
26 |- 33	7
33 |- 40	3
40 |- 47	3
47 |- 54	2
54 |- 61	2
Total	30

Fonte: Dados fictícios

3. Calcule a mediana da distribuição da tabela a seguir.

Idade dos 30 estudantes da Faculdade X pesquisados em 20/05/2010

Idade	Número de estudantes	fac%
19 |- 26	13	43,3%

26 |- 33	7	44,7%
33 |- 40	3	76,7%
40 |- 47	3	86,7%
47 |- 54	2	93,3%
54 |- 61	2	100%
Total	**30**	

Fonte: Dados fictícios

4. A partir da tabela a seguir, calcule:

a) Moda bruta.

b) Mediana.

c) Média.

Nível de colesterol total no sangue dos funcionários da Empresa XYZ em outubro de 2010. (em mg/dl)

Nível de colesterol total	Número de estudantes	Percentagem
180 |- 190	5	8,3%
190 |- 200	7	11,7%
200 |- 210	9	15%
210 |- 220	10	16,7%
220 |- 230	12	20%
230 |- 240	9	15%
240 |- 250	30	13,3%
Total	**60**	**100%**

Fonte: Dados fictícios

9 MEDIDAS DE DISPERSÃO OU VARIABILIDADE

As medidas de tendência central não são suficientes para representar bem uma distribuição de valores.

Vamos supor que você vá fazer uma viagem para uma cidade desconhecida onde só saiba a temperatura média da cidade na época em que você vai viajar, por exemplo 24℃.

Quais roupas você levará? Fica difícil decidir sem conhecer também alguma medida de dispersão, pois uma mesma cidade pode ter uma variação muito grande de temperatura ao longo de um dia.

Os dias podem ser bastante quentes e as noites muito frias e a temperatura média ser 24℃, mas também pode ter uma variação muito pequena em torno de 24℃ e a temperatura média também ser 24℃. Certamente a sua mala de roupas será bastante diferente neste dois casos. Neste capítulo conheceremos as medidas de dispersão e como utilizá-las.

Medida de dispersão são medidas que nos fornecem informação de como os valores variam, são chamadas de medidas de dispersão ou variabilidade e são bastante utilizadas na estatística.

Muitas medidas de dispersão vão comparar os dados em relação a um determinado valor (normalmente a média aritmética). As medidas que vamos estudar neste capítulo são: Amplitude Total, desvio médio, variância, desvio-padrão e coeficiente de variação.

9.1 Amplitude total

A amplitude total é a variação máxima encontrada nos dados, sendo calculada através da diferença entre o maior e menor valor.

Caso os dados estejam dispostos em tabelas de distribuição de frequências, a amplitude total vai ser calculada pela diferença entre o limite superior da maior classe e o limite inferior da menor classe.

9.2 Desvio médio

O desvio médio é a média de todas as discrepâncias.

Como a soma das discrepâncias sempre será igual a zero, pela propriedade da média aritmética, o cálculo do desvio médio utilizará o valor das discrepâncias sem sinal.

Desta forma, calcula-se o desvio médio somando todas as discrepâncias em módulo e dividindo este resultado pela quantidade de discrepâncias (que é igual à quantidade de dados, ou seja, n).

Dados isolados

$DM = (\sum |(Xi - \overline{X})|)/n$

Dados agrupados

$DM = (\sum fi\ |(X1 - \overline{X})|)/n$

Se os dados estiverem agrupados em intervalos de classe, Xi refere-se aos pontos médios de cada classe. Normalmente, o desvio médio não é muito utilizado na estatística, mas nos dá uma boa visão do que é uma medida de dispersão.

9.3 Variância

A variância é uma medida de dispersão bastante utilizada na estatística, sendo calculada a partir das discrepâncias elevadas ao quadrado (também com o objetivo de retirar o sinal das discrepâncias para poder somá-las).

Caso o nosso objetivo seja apenas a descrição dos dados, podemos assumir que os dados são a nossa população de interesse e a variância será representada pela letra grega σ^2, sendo calculada pela fórmula abaixo, onde a letra grega μ representa a média dos dados.

Dados isolados

$\sigma^2 = (\sum (Xi - \overline{X})^2)/n$

Dados agrupados

$\sigma^2 = (\sum fi\,(Xi - \overline{X})^2)/n$

Se o nosso objetivo for utilizar estes dados como uma amostra, para fazermos inferências em relação a um grupo maior, o cálculo da variância vai ser realizado pela fórmula abaixo, onde s^2 é a variância e $\overline{X}$ é a média da amostra.

Dados isolados

$s^2 = (\sum (Xi - \overline{X})^2)/(n - 1)$

Dados agrupados

$s^2 = (\sum fi\,(Xi - \overline{X})^2)/(n - 1)$

Observe que a modificação na fórmula é apenas o denominador, que passa a ser (n - 1) em substituição a n.

9.4 Desvio-padrão

O desvio padrão é calculado através da raiz quadrada da variância, sendo dado pela fórmula:

- Desvio-padrão da população (quando o interesse se restringe aos dados).
  $\sigma = \sqrt{\sigma^2}$

- Desvio-padrão da amostra (quando os dados serão utilizados para fazer inferências).

  $s = \sqrt{s^2}$

9.5 Propriedades do desvio-padrão

Analisaremos alguns exemplos a seguir.

Soma ou subtração

Somando-se, ou subtraindo-se, todos os dados de uma distribuição por um mesmo valor, o novo desvio-padrão não mudará.

Como todos os valores foram alterados da mesma diferença, a nova média também ficará alterada desta diferença (propriedade da média), implicando que as diferenças dos valores brutos em relação à média continuarão as mesmas gerando um mesmo desvio-padrão.

Multiplicação ou divisão

Multiplicando-se, ou dividindo-se, todos os dados de uma distribuição por um mesmo valor, o novo desvio-padrão também será multiplicado, ou dividido, por este mesmo valor.

Neste caso, ao multiplicar um valor por 2, por exemplo, se um escore possui valor 2 ao ser multiplicado por 2 passará a ser 4, porém se o escore possuir um valor 8, este passará a ser 16. A diferença entre 16 e 4 é muito maior do que inicialmente entre 2 e 4

9.6 Coeficiente de variação

O desvio-padrão não é uma boa medida para compararmos dois grupos muito diferentes.

Se temos o valor de 10kg como desvio-padrão de dois grupos diferentes, podemos dizer que os dois grupos possuem a mesma variabilidade?

Vamos supor que um dos grupos seja composto por pesos de crianças que calculamos ter média igual a 40kg. O segundo corresponde a um grupo de adultos com média de peso igual a 80kg.

Apesar dos desvios-padrões serem iguais, percebemos que 10kg correspondem a uma variação muito maior para um grupo com média de 40kg do que para um grupo com média de 80kg.

Se medirmos a altura e o peso de determinado grupo de pessoas, como saberemos qual variável tem maior variabilidade, se cada uma será expressa em uma medida diferente da outra (cm e kg)?

O coeficiente de variação é uma medida que vai permitir comparar variáveis de diferentes unidades, e também grupos onde as médias são muito diferentes. O coeficiente de variação é uma medida expressa em percentual e independente de unidade.

Veja a fórmula:

CV = (s/ $\overline{X}$) x 100

9.7 Atividade Proposta

1. Calcule a média, a amplitude total e o desvio médio dos dados da tabela.

Número de filhos dos pais presentes à primeira reunião da escola X em 2010.

N°. filhos	f	Percentual	fi Xi	(*X*i - Média)	|(*X*i - Média)|	fi | (*X*i - Média)|
1	7	30,43%	7	-1	1	1
2	11	47,83%	22	0	0	0
3	3	13,04%	9	1	1	1
4	2	8,70%	8	2	2	2

Total	23	100%	46			

Fonte: Dados fictícios

(Xi - Média) = discrepância

| (Xi - Média) | = módulo da discrepância

2. Calcule a variância e desvio-padrão dos dados da tabela, considerando que estes dados são uma amostra.

Número de filhos dos pais presentes à primeira reunião da escola X em 2010.

N°. filhos	f	Percentual	f i X i	(Xi - Média)	| (Xi - Média) |^2	f i | (Xi - Média) |^2
1	7	30,43%	7	-1	1	7
2	11	47,83%	22	0	0	0
3	3	13,04%	9	1	1	3
4	2	8,70%	8	2	4	8
Total	**23**	**100%**	**46**			**18**

Fonte: Dados fictícios

(X i - Média) = discrepância

| (X i - Média) | = módulo da discrepância

3. Calcule coeficiente de variação para os dados a seguir:

Média = 2

Desvio-padrão = 0,90

4. Calcule variância e desvio-padrão dos dados da tabela a seguir:

Número de filhos dos pais presentes à primeira reunião da escola X em 2010.

N°. filhos	f	Percentual

1	7	30,43%
2	11	47,83%
3	3	13,04%
4	2	8,70%
Total	**23**	**100%**

10 NOÇÕES DE CORRELAÇÃO

Será que existe relação entre altura e peso? Em nosso cotidiano, observamos o seguinte: à medida que a altura aumenta, as pessoas costumam ter pesos maiores. Este é um exemplo de relação entre duas variáveis.

O coeficiente de correlação é uma medida que quantifica o grau dessa relação. Mas afirmar que uma variável está relacionada à outra não implica a noção de causalidade. Por isso, o peso elevado de um indivíduo não depende, necessariamente, de sua altura. Neste capítulo , identificaremos as possibilidades e o grau de relação (forte ou fraco) entre duas variáveis.

Relação entre duas variáveis

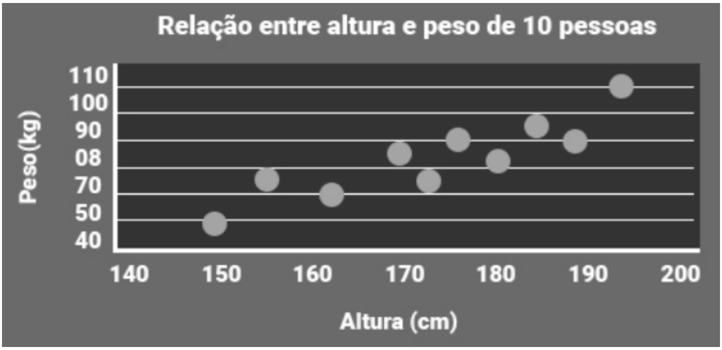
A relação entre duas variáveis pode ser observada em diagramas de dispersão. Quando queremos saber se uma variável tem relação com outra, medimos as duas variáveis de interesse para cada elemento da amostra.

Assim, se queremos saber se a altura tem relação com o peso, medimos a altura e o peso de cada pessoa. Então, cada uma irá fornecer um ponto a ser representado no gráfico de dispersão.

Tipos de correlação

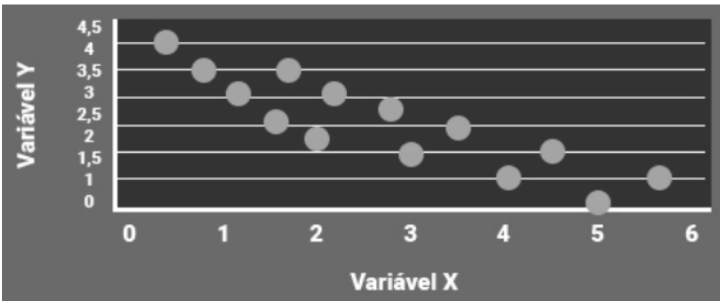
<u>Correlação linear positiva</u>

Quando um aumento de uma variável é acompanhado pelo aumento da outra variável, dizemos que há uma correlação positiva.



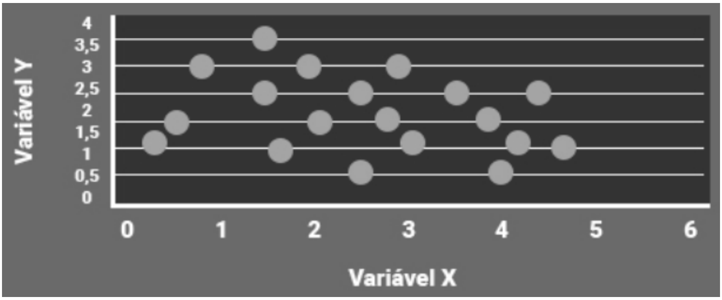
Correlação linear negativa

Gráfico com diversas marcações para duas variáveis: x e y. Iniciando na marcação y= 4 e x = 0,5. Quando y cai para 3,5 x aumenta para 0,9 e assim sucessivamente.



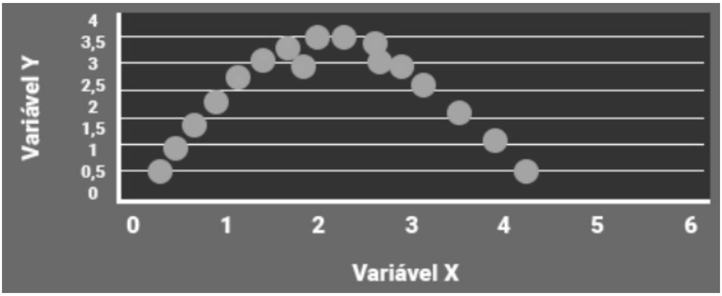
Correlação nula

Gráfico com diversas marcações para duas variáveis: x e y. Iniciando na marcação y= 4 e x = 0,5. Quando y cai para 3,5 x aumenta para 0,9 e assim sucessivamente.



Correlação não linear

Quando a taxa de aumento ou diminuição pode mudar como uma mudança de variáveis, causando um "padrão curvo" nos dados, dizemos que é uma correlação não linear.

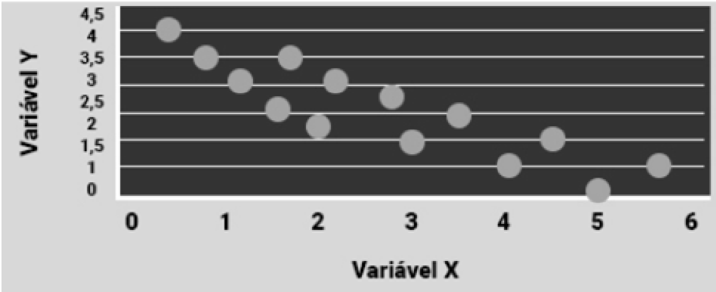


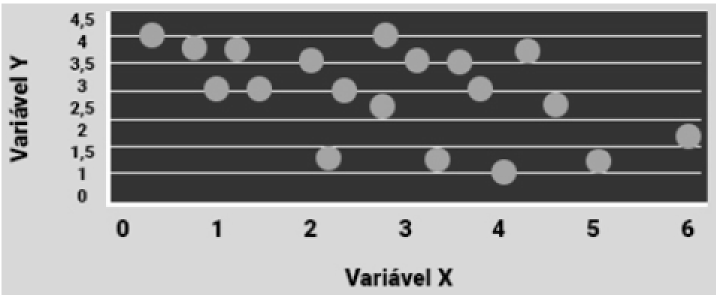
Grau de correlação - coeficiente linear de Pearson

O grau da correlação (r) pode ser quantificado através do coeficiente linear de Pearson. Dependendo do resultado encontrado, a relação pode ser classificada em:

$r = 0$	Não há correlação ou a correlação não é linear.
$0 < r < 0,3$	Correlação positiva muito fraca.
$0,3 \leq r < 0,6$	Correlação positiva fraca.
$0,6 \leq r \leq 1$	Correlação positiva forte.
$-0,3 < r < 0$	Correlação negativa muito fraca.
$-0,6 < r < -0,3$	Correlação negativa fraca.
$-1 \leq r \leq -0,6$	Correlação negativa forte.
$r = 1$	Correlação perfeita positiva.
$r = -1$	Correlação perfeita negativa.

Veja os exemplos de correlações negativas forte e fraca respectivamente:





O coeficiente de correlação linear de Pearson só tem significado se os dados possuem uma correlação linear, ou seja, estão dispostos no diagrama de dispersão em torno de uma reta.

Veja a fórmula:

$$r = \frac{n(\sum X_iY_i)-(\sum,X_i)(\sum,Y_i)}{\sqrt{\left[n\sum X^2\, i-(\sum,X_i)^2 \left[n\ \sum Y^2\, i-(\sum,Y_i)^2\right]\right]}}$$

n = Número de observações.

A tabela apresenta os dados para o cálculo do coeficiente de Pearson deste exemplo.

Exemplo para cálculo do coeficiente de correlação de Pearson

Altura	Peso	XY	X²	Y²
150	50	7500	22500	22500
155	65	10075	24025	4225
162	60	9720	26244	3600
169	75	12675	28561	5625
172	65	11180	29584	4225
175	80	14000	30625	6400
179	73	13067	32041	5329
183	85	15555	33489	7225
187	80	14960	34969	6400
192	100	19200	36864	10000
1724	**733**	**127932**	**298902**	**55529**

Sendo: n=10

$\sum x = 1.724$

$\sum y = 733$

$\sum xy = 127.932$

$\sum x \sum y = 1.724 \text{ x } 733 = 1.263.692$

$\sum x^2 = 298.902$

$(\sum x)^2 = (1724)^2 = 2.972.176$

$\sum y^2 = 55.529$

$(\sum y)^2 = (733)^2 = 537.289$

Aplicando estes valores na fórmula, temos:

$$\mathbf{r} = \frac{10(127.932)-(1724)(733)}{\sqrt{[10\ (298.902)-(2.972.176)][10(55.529)-(537.289)]}}$$

$$\mathbf{r} = \frac{(1.279.320)-(1.263.692)}{\sqrt{(16.844)(18.001)}}$$

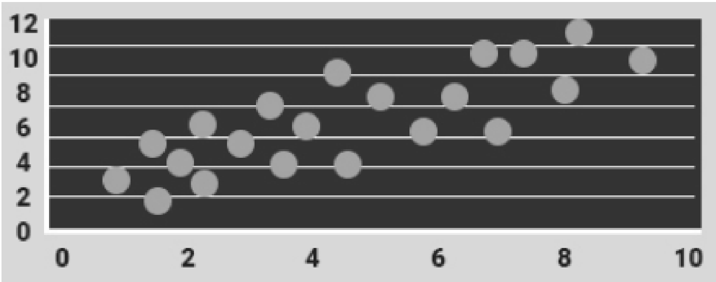
$$\mathbf{r} = \frac{(15.628)}{\sqrt{303.208.844}}$$

$$\mathbf{r} = \mathbf{0.8975}$$

Como r ≅ 0,90, podemos dizer que a correlação encontrada é positiva forte.

10.1 Atividade Proposta

1. Qual correlação é apresentada no gráfico a seguir?



a) Positiva fraca
b) Negativa fraca
c) Positiva forte
d) Negativa forte
e) Nula

Referências bibliográficas

Bussab WO, Morettin PA. Estatística básica. São Paulo: Saraiva; 2007.

CRESPO, Antonio Arnot. Estatística Fácil. 19ª edição. São Paulo: Editora Saraiva.2009.

FERREIRA, D. F. Fundamentos de Matemática Estatística / Daniel Furtado Ferreira. – 2. ed. Lavras : Ed. UFLA, 2013.

FERREIRA, D. F. Estatística básica. Lavras: Ed. Ufla, 2 ed. ampliada e revisada. 2009.

JACQUES-CALLEGARI, S. M. Bioestatística. Porto Alegre: Artmed, 2003.

LARSON. R. BETSY, F. Estatística Aplicada. São Paulo: Pearson Prentice Hall. 2004.

LEVIN, J.; FOX, J. A. Estatística para Ciências Humanas. 9. ed. São Paulo: Pearson, 2004.

MAGALHÃES, Marcos Nascimento; LIMA, Antônio Carlos Pedroso de. Noções de probabilidade e estatística. 6.ed., rev. São Paulo: EDUSP, 2005.

MITTELHAMMER, R. C. Mathematical statistics for economics and business. 2. ed. New York: Springer, 2013.

MORETTIN, Pedro Alberto; BUSSAB, Wilton de Oliveira. Estatística básica. 6. ed., rev. e atual. São Paulo: Saraiva, 2010.

SILVA, Marcos Noé Pedro da. "Estatística"; Brasil Escola. Disponível em: https://brasilescola.uol.com.br/matematica/estatistica-1.htm. Acesso em 28 de junho de 2020.

GABARITO DAS ATIVIDADES PROPOSTAS

Cap. 1

Faltam informações para que você possa escolher uma das turmas com segurança.
E se o professor A tinha 50 alunos e o professor B apenas 20 alunos?
Neste caso, o professor A reprovou 20% da turma e o professor B reprovou 25% da turma dele.
Você continuaria com a sua escolha?
Com este exemplo, vimos como podemos nos enganar se não tivermos uma visão crítica das informações que nos são passadas.
Este é apenas um exemplo de como as técnicas da estatística (no caso, cálculo de percentagens) podem nos ajudar a tomar decisões.

Cap. 2
Ordem das respostas: 2 – 4 – 1 – 2

Cap. 3
500 ml correspondem a meio litro e é a quantidade total da solução.
Logo, relacionamos 500 ml a 100% e calculamos quanto seriam 2% de 500 ml através de uma regra de três.
500 ml = 100%
Y (concentrado) = 2%
Multiplicando em X, obtemos:
100Y = 500 X 2
Y = 1000 ÷ 100
Y = 10 ml do concentrado.
Como o total da solução é 500 ml, a quantidade de diluente será obtida pela equação:
500 ml = diluente + concentrado
500 ml = diluente + 10 ml (obtido anteriormente)
Diluente = 500 - 10
Diluente = 490 ml

Cap. 4

1. Título: Toda tabela precisa ter um título localizado no topo da tabela. O título tem como objetivo informar o significado dos dados apresentados na tabela, respondendo às perguntas: O quê?, Quando?, Onde?

2. O título está bastante incompleto. Não diz onde, nem quando, foi realizada a pesquisa. Além disto, a especificação do que a tabela se propõe a apresentar não está clara. Quantidade de filhos de quem? Que quantidade de filhos é esta? Em termos de formatação, uma tabela estatística não pode ser fechada nas laterais.

Cap. 5

Item 1

Idade dos 30 alunos da Faculdade X pesquisados em 20/05/2010.

Altura	Número de estudantes	Percentual
19 |--- 27	13	43,3%
27 |--- 35	8	26,7%
35 |--- 43	3	10,0%
43 |--- 51	2	6,7%
51 |--- 59	4	13,3%
Total	**30**	**100,0%**

Fonte: Dados fictícios - Distribuição de frequências para as idades com k = 5

Item 2

Idade	Número de estudantes	Percentual
19 |--- 26	13	43,3%
26 |--- 33	7	23,3%
33 |--- 40	3	10,0%
40 |--- 47	2	10,0%
47 |--- 54	2	6,7%
54 |--- 61	3	6,7%
Total:	**30**	**100,0%**

Fonte: Dados fictícios - Distribuição de frequências para as idades com k = 6

Item 3

a) Qual é o tamanho de cada classe (h)?

10 (dez) mg/dl. O tamanho de cada classe é calculado subtraindo o limite superior da classe pelo limite inferior de cada classe. Neste exemplo todas as classes possuem o mesmo tamanho.

b) Quantas classes tem esta tabela?

7 (sete). O número de classes é igual ao número de linhas.

c) Pela nomenclatura utilizada, o limite superior entra na contagem de cada classe?

Não. A nomenclatura utilizada nesta tabela utiliza o símbolo |--- para a notação de intervalo. Este símbolo indica limite inferior fechado e limite superior aberto, de modo que a contagem dos dados com valores iguais ao limite inferior entra na contagem desta classe e a contagem dos dados com valores iguais ao limite superior vai entrar na contagem da classe posterior. Por exemplo, o valor 190 será contado na segunda classe (não na primeira).

Cap. 6
O gráfico em linha é mais adequado para representar séries geográficas longas do que o gráfico em barras. Histograma e polígono de frequências são gráficos para apresentar dados organizados em intervalos de classe, que não é o caso deste exemplo. O gráfico de setores é utilizado para comparar proporções de uma quantidade pequena de categorias.

Cap. 7
a) A moda.

Mo = 2; pois é o valor que tem a maior frequência absoluta (f = 8).

b) (∑ FiXi) ÷ N, ou seja, criamos uma coluna que é a multiplicação da coluna da frequência absoluta (jogos) pela coluna dos dados (neste caso, número de gols). A soma desta coluna fornece o número total de gols de todas as partidas. Para calcularmos a média, dividimos o número total de gols (neste caso, 43) pelo número total de partidas (neste caso, 24)Resultado: Média = 43/24 = 1,79.

Número de gols dos jogos da primeira fase do campeonato nacional do país XPTO no ano de 1999.

Número de gols	Jogos	%	FX
0	5	20,8%	0
1	5	20,8%	5
2	8	33,3%	16
3	3	12,5%	9
4	2	8,3%	8

5	1	4,2%	5
Total	**24**	**100%**	**43**

Fonte: Dados fictícios

c) A mediana.

Mediana é o valor que divide a distribuição ao meio, ou seja, metade dos valores estão abaixo (ou é igual à mediana) e a outra metade está acima (ou é igual à mediana). Como temos número par de valores (N = 24), a mediana vai estar no meio dos 2 valores centrais. A posição do primeiro termo central é dada dividindo o número total de valores (N = 24) por 2. Logo, o primeiro termo central ocupa a 12ª posição e o segundo termo central ocupa a 13ª posição. Para conhecermos o valor dos termos centrais calculamos a coluna das frequências acumuladas, como mostra a tabela a seguir.

Número de gols dos jogos da primeira fase do campeonato nacional do país XPTO no ano de 1999.

Número de gols	jogos	%	FX	Fac
0	5	20,8%	0	5
1	5	20,8%	5	10
2	8	33,3%	16	18
3	3	12,5%	9	21
4	2	8,3%	8	23
5	1	4,2%	5	24
Total	**24**	**100%**	**43**	

Fonte: Dados fictícios

Obs: Vemos que temos 10 jogos com número de gols até 1, ou seja até a posição 10 temos jogos com 1 ou menos gols. A partir da 11ª posição até a 18ª posição (linha marcada) temos jogos com 2 gols. A 12ª posição e a 13ª posição estão contidas nesta linha. Logo, a mediana é igual à média do valor da 12ª posição com o valor da 13ª posição. Como ambas as posições têm valor 2, a mediana também é igual a 2.

Cap. 8

1 Item

Classe modal: 19 |- 26

Moda bruta = (19 + 26)/2 = 45/2 = 22,5

li = 19

h = 26 - 19 = 7

fi = 13 (frequência absoluta da classe modal)

fi - 1 = 0 (frequência absoluta da classe anterior à classe modal, neste caso não existe, pois a classe modal é a primeira classe, não existindo classe anterior.)

fi + 1 = 7 (frequência absoluta da classe posterior à classe modal).

Aplicando a fórmula:

Mo = (19 + 7 (13 -))/(13 – 0) + (13 – 7)

Mo = 19 + (91/19)

Mo = 19 + 4,79

Mo = 23,79

2 Item

Inicialmente construímos a coluna auxiliar com os valores médios de cada classe.

X1 = (19 + 26)/2 = 22,5

X2 = (26 + 33)/2 = 29,5

X3 = (33 + 40)/2 = 36,5

O calcule deve ser feito para todas as classes.

Lembrando a fórmula da média:

$$\overline{X} = \frac{\sum f_i . X_i}{n}$$

Construímos, então, a coluna fiXi, com os resultados obtidos.

Idade dos 30 estudantes da Faculdade X pesquisados em 20/05/2010

Idade	Número de estudantes	Xi	FiXi
19 |- 26	13	22,5	292,5
26 |- 33	7	29,5	206,5
33 |- 40	3	36,5	109,5
40 |- 47	3	43,5	13,5
47 |- 54	2	50,5	101
54 |- 61	2	57,5	115
Total	**30**	**-**	**955**

Fonte: Dados fictícios

Assim, temos:

$$\overline{X} = \frac{\sum f_i . X_i}{n} = 955/30 = 31,83 \text{ anos}$$

3 Item

A classe da mediana (26 |- 33), segundo intervalo de classe, porque a frequência acumulada em percentual desta classe é 66,7% e a anterior era menor do que 50%. Lembre-se que a classe da mediana é a primeira classe onde a frequência acumulada em percentual é igual ou maior do que 50%.

Construímos a coluna de frequências absolutas acumuladas, encontrando fa -1 = 13.

Como h = 33 – 26 = 7 e a frequência absoluta do intervalo 26 |- 33 é 7, substituindo os valores na fórmula, temos:

Md = 26 + (7(30/2 - 13))/7 = 26 + 2 = 28

Idade dos 30 estudantes da Faculdade X pesquisados em 20/05/2010

Idade	Número de estudantes	fac%	fac
19 |- 26	13	43,3%	13
26 |- 33	7	66,7%	20
33 |- 40	3	76,7%	23
40 |- 47	3	86,7%	26
47 |- 54	2	93,3%	28
54 |- 61	2	100%	30
Total	**30**	**-**	**-**

Fonte: Dados fictícios

Item 4

a) Moda bruta

Moda bruta = (220+230) ÷ 2 = 225 mg/dl

Lembrando que classe modal é aquela de maior frequência: i = 5 : 220 ├ 230

Moda bruta é o ponto médio da classe modal.

b) Mediana.

Classe da mediana é aquela que contém o ponto onde estão 50% dos valores. Para encontrá-la, devemos fazer a coluna das frequências relativas em percentual acumuladas e ver onde se encontra o valor 50%.

Nível de colesterol total no sangue dos funcionários da Empresa XYZ em outubro de 2010. (em mg/dl)

Nível de colesterol total	Número de estudantes	Percentagem	fac%	fac
180 |- 190	5	8,3%	8,3%	5
190 |- 200	7	11,7%	20%	12
200 |- 210	9	15%	35%	21
210 |- 220	10	16,7%	51,7%	31
220 |- 230	12	20,0%	71,7%	43
230 |- 240	9	15,0%	86,7%	52
240 |- 250	30	13,3%	100,0%	60
Total	**60**	**100%**		

Fonte: Dados fictícios

Encontrando a classe da mediana, vamos aplicar a fórmula abaixo.

Md = 210 + (60 ÷ 2 – 21) ÷ 10
Md = 210 + (30-21) ÷ 10
Md = 210 + 9 ÷ 10
Md = 210,9 mg/dl

c) Média.

Como já sabemos, a fórmula da média é:

$$\overline{X} = \frac{\sum f_i . X_i}{n}$$

Para dados agrupados, Xi é o ponto médio de cada intervalo de classe. Constrói-se uma coluna para os pontos médios de cada classe. Esta coluna vai ser multiplicada pelas frequências absolutas respectivas de cada classe.

A soma desta última coluna (neste caso, 13.060) vai ser dividida pelo número de valores (neste caso, 60).

Encontrando Média = 13.060 ÷ 60 = 217,67mg/dl

Veja tabela seguinte para construção das colunas auxiliares:

Nível de colesterol total no sangue dos funcionários da Empresa XYZ em outubro de 2010. (em mg/dl)

Nível de colesterol total	Número de estudantes	Percentagem	Xi	fiXi
180 |- 190	5	8,3%	185	9125
190 |- 200	7	11,7%	195	1365
200 |- 210	9	15%	205	1845
210 |- 220	10	16,7%	215	2150
220 |- 230	12	20,0%	225	2700
230 |- 240	9	15,0%	235	2115
240 |- 250	30	13,3%	245	1960
Total	**60**	**100%**	**-**	**13060**

Fonte: Dados fictícios

Cap. 9

Item 1

Média = 46 ÷ 23 = 2
Amplitude total (AT) = 4 – 1 = 3
DM = 14 ÷ 23 = 0,6087

Item 2

Média = 46 ÷ 23 = 2
Variância da amostra = 18 ÷ (23 – 1) = 0,8182
Desvio-padrão da amostra = 0,9045

Item 3

CV = 0,90 x 100/2 = 45%

Item 4

Variância = $\sigma 2 = \Sigma fi\ (Xi - X)2/n$

O cálculo da variância é feito elevando cada discrepância ao quadrado, como na coluna 6 da tabela a seguir.

Em seguida, cada discrepância ao quadrado é multiplicada pelo valor correspondente à frequência absoluta de cada linha (coluna 7). O valor da variância é o resultado da soma da coluna 7 dividida pela soma da coluna F.

Número de filhos dos pais presentes à primeira reunião da escola X em 2010.

N°. filhos	**f**	**Percentual**	**f i X i**	**(Xi - Média)**	**| (Xi - Média) |^2**	**f i | (Xi - Média) |^2**
1	7	30,43%	7	-1	1	7
2	11	47,83%	22	0	0	0
3	3	13,04%	9	1	1	3
4	2	8,70%	8	2	4	8
Total	**23**	**100%**	**46**			**18**

$\sigma 2 = 18/23 = 0,78$

O desvio padrão é a raiz quadrada da variância.

Como $\sigma 2 = 0,78$, o desvio padrão $\sigma = \sqrt{\sigma^2} = \sqrt{0,78} = 0,88$.

Cap. 10

A medida em que x cresce, y também está crescendo. Dessa forma, a correlação é positiva. A correlação é forte, porque a dispersão dos dados é pequena.

CENTRO UNIVERSITÁRIO ESTÁCIO DO CEARÁ

Pró-Reitoria de Pós-Graduação, Pesquisa e Extensão

Núcleo de Publicações Acadêmico-Científicas