Estatística Descritiva

1 INTRODUÇÃO

A Estatística é uma ciência cujo campo de aplicação estende-se a muitas áreas do conhecimento humano. Entretanto, um equívoco comum que deparamos nos dias atuais é que, em função da facilidade que o advento dos computadores nos proporciona, permitindo desenvolver cálculos avançados e aplicações de processos sofisticados com razoável eficiência e rapidez, muitos pesquisadores consideram-se aptos a fazerem análises e inferências estatísticas sem um conhecimento mais aprofundado dos conceitos e teorias. Tal prática, em geral, culmina em interpretações equivocadas e muitas vezes errôneas...

Em sua essência, a Estatística é a ciência que apresenta processos próprios para coletar, apresentar e interpretar adequadamente conjuntos de dados, sejam eles numéricos ou não. Pode-se dizer que seu objetivo é o de apresentar informações sobre dados em análise para que se tenha maior compreensão dos fatos que os mesmos representam. A Estatística subdivide-se em três áreas: descritiva, probabilística e inferencial. A estatística descritiva, como o próprio nome já diz, se preocupa em descrever os dados. A estatística inferencial, fundamentada na teoria das probabilidades, se preocupa com a análise destes dados e sua interpretação.

A palavra estatística tem mais de um sentido. No singular se refere à teoria estatística e ao método pelo qual os dados são analisados enquanto que, no plural, se refere às estatísticas descritivas que são medidas obtidas de dados selecionados.

A estatística descritiva, cujo objetivo básico é o de sintetizar uma série de valores de mesma natureza, permitindo dessa forma que se tenha uma visão global da variação desses valores, organiza e descreve os dados de três maneiras: por meio de tabelas, de gráficos e de medidas descritivas.

A tabela é um quadro que resume um conjunto de observações, enquanto os gráficos são formas de apresentação dos dados, cujo objetivo é o de produzir uma impressão mais rápida e viva do fenômeno em estudo.

Para ressaltar as tendências características observadas nas tabelas, isoladamente, ou em comparação com outras, é necessário expressar tais tendências através de números ou estatísticas. Estes números ou estatísticas são divididos em duas categorias: medidas de posição e medidas de dispersão.

Para se obter bons resultados numa análise estatística, além dos métodos aplicados, também é necessário ter clareza nos conceitos utilizados. A seguir são apresentados alguns desses conceitos.

1.1 CONCEITOS FUNDAMENTAIS E DEFINIÇÕES

A estatística trabalha com dados, os quais podem ser obtidos por meio de uma população ou de uma amostra, definida como:

População: conjunto de elementos que tem pelo menos uma característica em comum. Esta característica deve delimitar corretamente quais são os elementos da população que podem ser animados ou inanimados.

Amostra: subconjunto de elementos de uma população. Este subconjunto deve ter dimensão menor que o da população e seus elementos devem ser representativos da população. A seleção dos elementos que irão compor a amostra pode ser feita de várias maneiras e irá depender do conhecimento que se tem da população e da quantidade de recursos disponíveis. A estatística inferencial é a área que trata e apresenta a metodologia de amostragem.

Em se tratando de conjuntos-subconjuntos, estes podem ser:

Finitos: possuem um número limitado de elementos.

Infinitos: possuem um número ilimitado de elementos.

Segundo Medronho (2003), elemento significa cada uma das unidades observadas no estudo.

Após a determinação dos elementos pergunta-se: o que fazer com estes? Pode-se medi-los, observá-los, contá-los surgindo um conjunto de respostas que receberá a denominação de variável.

Variável: é a característica que vai ser observada, medida ou contada nos elementos da população ou da amostra e que pode variar, ou seja, assumir um valor diferente de elemento para elemento.

Não basta identificar a variável a ser trabalhada, é necessário fazer-se distinção entre os tipos de variáveis:

Variável qualitativa: é uma variável que assume como possíveis valores, atributos ou qualidades. Também são denominadas variáveis categóricas.

Variável quantitativa: é uma variável que assume como possíveis valores, números.

Cada uma dessas variáveis pode ser sub-classificada em:

Variável qualitativa nominal: é uma variável que assume como possíveis valores,

atributos ou qualidades e estes não apresentam uma ordem natural de ocorrência.

Exemplo 01: meios de informação utilizados pelos alunos da disciplina Inferência Estatística do curso de Estatística da UEM: televisão, revista, internet, jornal.

Variável qualitativa ordinal: é uma variável que assume como possíveis valores atributos ou qualidades e estes apresentam uma ordem natural de ocorrência.

Exemplo 02: estado civil dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM: solteiro, casado, separado.

Variável quantitativa discreta: é uma variável que assume como possíveis valores números, em geral inteiros, formando um conjunto finito ou enumerável.

Exemplo 03: número de reprovas, por disciplina, dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM: 0, 1, 2, ...

Variável quantitativa contínua: é uma variável que assume como possíveis valores números, em intervalos da reta real e, em geral, resultantes de mensurações.

Exemplo 04: peso (quilogramas) dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM: 58, 59, 63.....

2 TABELA

É muito comum nos dias de hoje, devido ao uso de computadores, realizarem pesquisas em que a coleta de dados resulta em grandes coleções (quantidades) de dados para análise e torna-se quase impossível entendê-los, quanto ao(s) particular(es) objetivo(s) de estudo, se estes dados não estiverem resumidos. Em outras palavras, os dados na forma em que foram coletados não permitem, de maneira fácil e rápida, que se extraia informações. Torna-se difícil detectar a existência de algum padrão. É necessário "trabalhar os dados para transformá-los em informações, para compará-los com outros resultados, ou ainda para julgar sua adequação a alguma teoria" (Bussab, 2003, p.1). Montgomery (2003, p.14) afirma que "sumários e apresentações de dados bem constituídos são essenciais ao bom julgamento estatístico, porque permitem focar as características importantes dos dados ou ter discernimento acerca do tipo de modelo que deveria ser usado na solução do problema em questão".

Com o objetivo de levantar dados, para exemplificar a maioria das técnicas apresentadas, no dia 21/03/2005, um questionário (vide anexo I) foi aplicado aos alunos do 2º ano do curso de Estatística da Universidade Estadual de Maringá (UEM) matriculados na disciplina Inferência Estatística. As variáveis que compõem o questionário são:

Sexo: com categorias (1) se masculino e (2) se feminino

Id: idade em anos

Altura: altura em metros e centímetros

Peso: peso em quilos

Est.Civil: estado civil com categorias (1) se solteiro, (2) se casado e (3) se separado

Nºir.: número de irmãos

Transp.: meio de transporte mais utilizado com categorias (1) de coletivo e (2) se próprio

Procedência: município de procedência com categorias (1) se Maringá, (2) se outro município do Paraná e (3) se de outro Estado

Trabalho: relação do trabalho com o curso com categorias (1) não trabalho, (2) completamente relacionado, (3) parcialmente relacionado e (4) não relacionado

Inform: meio de informação mais utilizado com categorias (1) se TV, (2) jornal, (3) rádio, (4) revista e (5) internet

Disc.: número de disciplinas reprovadas no 1º ano da UEM.

Para se trabalhar estes dados são necessários, em primeiro lugar, tabulá-los e apresentá-

los na forma em que foram coletados (dados brutos) como na Tabela 01. Em geral, a 1ª coluna da tabela deve conter a identificação do respondente.

Tabela 01 - Informações sobre sexo, idade (anos), altura (metro e centímetro), peso (kg), estado civil, número de irmãos, transporte, procedência, relação do trabalho com o curso de Estatística, meio de informação e número de disciplinas reprovadas dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM - 21/03/2005.

Nº	Sexo	Id	Altura	Peso	Est.Civil	Nºir.	Transp.	Procedência	Trabalho	Inform	Disc.
1	F	20	1,60	58	Solteiro	1	Próprio	Maringá	Não Rel.	TV	2
2	F	26	1,65	59	Solteiro	2	Coletivo	Fora do Pr	Não trab.	Revista	0
3	F	18	1,64	55	Solteiro	2	Próprio	Maringá	Não trab.	TV	0
4	F	25	1,73	60	Solteiro	2	Coletivo	Outro no Pr	Não Rel.	TV	2
5	M	35	1,76	83	Casado	6	Coletivo	Outro no Pr	Não Rel.	TV	2
6	F	20	1,62	58	Solteiro	2	Coletivo	Outro no Pr	Não Rel.	Rádio	5
7	F	29	1,72	70	Solteiro	3	Coletivo	Maringá	Não trab.	TV	0
8	M	23	1,71	62	Separado	2	Próprio	Outro no Pr	Não Rel.	Internet	2
9	F	20	1,63	63	Solteiro	2	Próprio	Maringá	Não trab.	TV	1
10	M	20	1,79	75	Solteiro	2	Próprio	Fora do Pr	Não trab.	Internet	2
11	M	20	1,82	66	Solteiro	1	Próprio	Fora do Pr	Não trab.	TV	2
12	F	30	1,68	46	Solteiro	3	Próprio	Outro no Pr	Parc.Rel.	TV	4
13	F	18	1,69	64	Solteiro	1	Próprio	Maringá	Parc.Rel.	TV	0
14	M	37	1,82	80	Casado	2	Próprio	Maringá	Não Rel.	TV	3
15	M	25	1,83	62	Solteiro	1	Próprio	Outro no Pr	Não Rel.	TV	2
16	F	20	1,63	68	Solteiro	2	Coletivo	Maringá	Não trab.	TV	2
17	M	21	1,71	80	Solteiro	2	Coletivo	Maringá	Não Rel.	Internet	0
18	M	25	1,80	82	Casado	1	Próprio	Outro no Pr	Não Rel.	Internet	3
19	F	24	1,62	55	Solteiro	2	Próprio	Maringá	Não trab.	Jornal	2
20	M	19	1,74	58	Solteiro	2	Próprio	Maringá	Com.Rel.	TV	3
21	F	21	1,55	65	Solteiro	1	Próprio	Maringá	Não trab.	TV	1
22	M	22	1,73	62	Solteiro	0	Próprio	Maringá	Não trab.	Jornal	4

Fonte: Departamento de Estatística (DES)/UEM.

De acordo com Magalhães (2000), pode-se observar que a Tabela 01, tabela de dados brutos, contém muita informação, porém pode não ser muito rápido e prático obter estas informações. Por exemplo, não é imediato afirmar que existem mais homens que mulheres. Neste sentido, pode-se construir outra tabela para cada uma das variáveis que resumirá as informações ali contidas.

Segundo o mesmo autor, observa-se também que, ao usar programas computacionais e para facilitar/agilizar a digitação do banco de dados, às variáveis qualitativas associam-se valores numéricos e nem por isso a variável deixa de ser qualitativa. Cabe ao bom senso lembrar da natureza da variável.

Embora um certo volume de informação seja perdido quando os dados são resumidos, um grande volume pode também ser ganho. "Uma tabela talvez seja o meio mais simples de se resumir um conjunto de observações" (Pagano, 2004, p.10). "Deve ser usada quando é importante a apresentação dos valores" (Medronho, 2003, p.227), e sua leitura depende de quem a lê.

Todas as variáveis podem ser resumidas através de uma tabela, mas a construção é diferenciada dependendo do tipo de variável.

Denomina-se Tabela Simples à tabela que resume os dados de uma única variável qualitativa e Distribuição de Freqüências ao resumo de uma única variável quantitativa.

2.1 ELEMENTOS DA TABELA

Toda tabela deve ser simples, clara, objetiva e auto-explicativa. Segundo Milone (2004, p.25),

> os elementos fundamentais da tabela são: título, cabeçalho, coluna indicadora e corpo. O título aponta o fenômeno, época e local de ocorrência; o cabeçalho explica o conteúdo das colunas; a coluna indicadora detalha as linhas; o corpo mostra os dados. Complementarmente, tem-se: fonte, notas e chamadas. A fonte cita o informante (caracterizando a confiabilidade dos dados); as notas esclarecem o conteúdo e indicam a metodologia adotada na obtenção ou elaboração da informação; as chamadas clarificam pontos específicos da tabela.

A disposição de uma tabela pode ser generalizada como mostra a Figura 01 a seguir.

Tabela *rs* – Título respondendo as perguntas: o quê, onde e quando?

Coluna indicadora	Cabeçalho
Conteúdo da linha	Célula / Coluna

Fonte: Origem dos dados.
Nota: Informação esclarecedora.

Corpo da tabela

Figura 01 – Representação tabular dos dados.

Destaca-se que as tabelas devem ser numeradas em ordem crescente ou em que aparecem no texto, como é o caso de trabalhos científicos; as bordas superiores e inferior devem ser fechadas com traços horizontais enquanto às da esquerda e direita não, podendo ou não ser fechadas por traços verticais a separação das colunas no corpo da tabela. É conveniente também

que o número de casas decimais seja padronizado.

2.2 TABELA SIMPLES

Uma tabela simples contém as diferentes categorias observadas de uma variável qualitativa e suas respectivas contagens, denominadas freqüências absolutas. A contagem refere-se ao número de ocorrências de cada categoria e é realizada utilizando-se, por exemplo, a Tabela 01 ou o banco de dados.

Quanto à classificação, uma tabela simples pode ser temporal quando as observações são feitas levando-se em consideração o tempo; geográfica quando os dados referem-se ao local de ocorrência; específica (ou categórica) quando tempo e local são fixos; e comparativa quando a tabela resume informações de duas ou mais variáveis. A tabela comparativa é também denominada tabela cruzada ou de dupla ou mais entradas.

Os nomes da coluna indicadora e cabeçalho podem ser escritos iniciando-se com letras maiúsculas. Também é prática comum justificar à esquerda as diferentes categorias da variável qualitativa que se apresentam no conteúdo das linhas, iniciando-se com letras maiúsculas e podem ser dispostas na ordem em que aparecem nos questionários, ordem alfabética ou ordem decrescente de freqüência absoluta.

Exemplo 05: Tabela histórica.

Tabela 02 – Número de alunos matriculados na disciplina Probabilidade I do curso de Estatística da Universidade Estadual de Maringá.

Ano	Nº de Alunos
2000	40
2001	59
2002	63
2003	69
2004	71

Fonte: DES/UEM.
Nota: Os números de 2003 e 2004 correspondem a duas turmas.

Exemplo 06. Tabela geográfica, específica e comparativa construída a partir da Tabela 01.

Tabela 03 – Município de procedência dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

Município de Procedência	Nº de Alunos
Maringá	12
Outro no Paraná	7
Fora do Paraná	3
Total	22

Fonte: Tabela 01.

É comum e útil na interpretação de tabelas a inclusão de uma coluna contendo as freqüências relativas e/ou relativas em percentual. A freqüência relativa é obtida dividindo-se a freqüência absoluta de cada categoria da variável pelo número total de observações (número de elementos da amostra ou da população). Multiplicando-se este resultado por 100, obtém-se a freqüência relativa em percentual. Assim, a Tabela 5 torna-se:

Tabela 04 – Município de procedência dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

Município de Procedência	Nº de Alunos	Percentual
Maringá	12	55
Outro no Paraná	7	32
Fora do Paraná	3	13
Total	22	100

Fonte: Tabela 01.

Segundo Barbetta et al. (2004), as freqüências relativas em percentual são úteis ao se comparar tabelas ou pesquisas diferentes. Por exemplo, quando amostras (ou populações) têm números de elementos diferentes, a comparação através das freqüências absolutas pode resultar em afirmações errôneas enquanto que pelas freqüências relativas em percentual não, pois os percentuais totais são os mesmos.

Tabela 05 – Meio de transporte mais utilizado pelos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

Meio de transporte	Nº de Alunos
Coletivo	7
Próprio	15
Total	22

Fonte: Tabela 01.

Tabela 06 – Meio de transporte mais utilizado segundo o sexo dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

Meio de transporte	Sexo		Total
	Masculino	Feminino	
Coletivo	2	5	7
Próprio	8	7	15
Total	10	12	22

Fonte: Tabela 01.

Exercício 01. Construa tabelas simples, incluindo os percentuais, para as variáveis estado civil, relação do trabalho com o curso de graduação e meio de transporte mais utilizado referentes à Tabela 01. Construa também, uma tabela cruzada para as variáveis estado civil e meio de informação.

2.3 DISTRIBUIÇÃO DE FREQÜÊNCIA

Como já mencionado no início deste capítulo, dependendo do volume de dados, torna-se difícil ou impraticável tirar conclusões a respeito do comportamento das variáveis e, em particular, de variáveis quantitativas.

Pode-se, no entanto, colocar os dados brutos de cada uma das variáveis quantitativas em uma ordem crescente ou decrescente, denominado rol. A visualização de algum padrão ou comportamento continua sendo de difícil observação ou até mesmo cansativa, mas torna-se rápido identificar maiores e menores valores ou concentrações de valores no caso de variáveis quantitativas. Estes números (menor e maior valor observado) servem de ponto de partida para a construção de tabelas para estas variáveis. Vale destacar que para as variáveis qualitativas, pode-se também construir um rol em ordem temporal ou alfabética, por exemplo.

É a diferença entre o menor e maior valor observado da variável X, denominada amplitude total ($AT = x_{max} - x_{min}$), que definirá a construção de uma distribuição de freqüência pontual ou em classes.

O ideal é que uma distribuição de freqüência resuma os dados em um número de linhas que varie de 5 a 10.

2.3.1 Distribuição de freqüência pontual – sem perda de informação

A construção de uma distribuição de freqüência pontual é equivalente à construção de uma tabela simples, onde se listam os diferentes valores observados da variável, com suas freqüências absolutas, denotadas por F_i, onde o índice i corresponde ao número de linhas da tabela, como é mostrado na Tabela 7.

Tabela 07 – Número de irmãos dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

Número de irmãos	Contagem	Freqüência (F_i)
0	|	1
1	|||||	6
2	||||||||||	12
3	||	2
6	|	1
Total		22

Fonte: Tabela 01.

Observa-se que esta variável foi resumida em 5 linhas. Assim, i = 1,...,5, e, portanto, tem-se 5 valores para as freqüências absolutas. A freqüência absoluta da segunda linha, $F_2 = 6$, por exemplo, indica que seis alunos têm um irmão, enquanto apenas um afirmou ter seis irmãos. A soma de todas as freqüências absolutas deve ser igual ao número total de observações da variável, neste caso, 22. A segunda coluna desta tabela é uma coluna opcional em distribuições de freqüências.

Ainda, como colunas complementares em uma distribuição de freqüências e considerando i, a ordem da linha na tabela, tem-se:

- a **freqüência relativa**, denotada por f_i, e já definida anteriormente como:

$$f_i = \frac{F_i}{n}$$

onde n é o tamanho da amostra, devendo ser substituída por N se os dados forem populacionais. A soma das freqüências relativas de todas as categorias é igual a 1;

- a **freqüência relativa em percentual**, denotada por $f_i\%$, e definida como:

$$f_i\% = \frac{F_i}{n} \cdot 100,$$

representando o percentual de observações que pertencem àquela categoria. A soma das freqüências deve, agora, ser igual a 100%;

- a **freqüência absoluta acumulada**, denotada por F_{a_i}. Estas freqüências são obtidas somando-se a freqüência absoluta do valor considerado, às freqüências absolutas anteriores a este mesmo valor.

- a **freqüência acumulada relativa**, denotada por $f_{a_i}\%$ e definida como:

$$f_{a_i}\% = \frac{F_{a_i}}{n} 100$$

Uma tabela contendo todas estas freqüências é dita uma distribuição de freqüências completa. Desta forma, a Tabela 8 pode ser apresentada como:

Tabela 08 – Número de irmãos dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

Número de irmãos (x_i)	F_i	$f_i\%$	F_{a_i}	$f_{a_i}\%$
0	1	4,55	1	4,55
1	6	27,26	7	31,81
2	12	54,55	19	86,36
3	2	9,09	21	95,45
6	1	4,55	22	100,00
Total	22	100,00		

Fonte: Tabela 01.

Segundo Milone (2004), em se tratando das freqüências relativas em percentuais, arredondamentos se fazem necessários e devem ser feitos de maneira convencional. Neste tipo de aproximação opta-se sempre pelo menor erro. Por exemplo, se for necessário aproximar o número 0,483 para a ordem do centésimo, erra-se menos subtraindo 0,003 que adicionando 0,007 ao valor 0,483, portanto a aproximação correta é 0,48. Se a aproximação for do número 0,4853

para a ordem do centésimo, então o erro menor será para a adição de 0,0047 e não para a subtração de 0,0053, e a aproximação adequada é 0,49. Já no caso do número 0,485, o tamanho do erro de aproximação é o mesmo que se obtém quando feita para mais ou para menos (0,005), e neste caso, cabe ao pesquisador decidir qual aproximação é mais conveniente.

A soma de todas as freqüências relativas percentuais deve ser igual a 100. Entretanto, quando são feitas aproximações, tal fato pode não ocorrer. Para o caso em que for menor que 100, soma-se uma unidade ao dígito de interesse das maiores freqüências relativas até que a soma seja 100. Se for maior que 100, deve-se subtrair uma unidade das maiores freqüências relativas. Agora, se ocorrem empates ou se as maiores freqüências forem números inteiros, é conveniente trabalhar com as outras freqüências. O importante é que a distribuição dos dados não seja alterada.

2.3.2 Distribuição de freqüência em classes – com perda de informação

"A distribuição de freqüências em classes é apropriada para apresentar dados quantitativos contínuos ou discretos com um número elevado de possíveis valores" (Medronho, 2003, p231). É necessário dividir os dados em intervalos ou faixas de valores que são denominadas classes. Uma classe é uma linha da distribuição de freqüências. O menor valor da classe é denominado limite inferior (l_i) e o maior valor da classe é denominado limite superior (L_i). O intervalo ou classe pode ser representado das seguintes maneiras:

a) l_i |— L_i, onde o limite inferior da classe é incluído na contagem da freqüência absoluta mas o superior não;

b) l_i —| L_i, onde o limite superior da classe é incluído na contagem mas o inferior não;

c) l_i |—| L_i, onde tanto o limite inferior quanto o superior são incluídos na contagem;

d) l_i — L_i, onde os limites não fazem parte da contagem.

Pode-se escolher qualquer uma destas opções sendo o importante tornar claro no texto ou na tabela qual está sendo usada.

"Se houver muitos intervalos, o resumo não constituirá grande melhoria com relação aos dados brutos. Se houver muito poucos, um grande volume de informação se perderá. Embora não seja necessário, os intervalos são freqüentemente construídos de modo que todos tenham larguras iguais, o que facilita as comparações entre as classes". (Pagano, 2004, p.11).

Milone (2004, p.36) apresenta os seguintes critérios para a determinação do número de intervalos, denotado por k:

1. Raiz quadrada: $k = \sqrt{n}$

2. Log (Sturges): $k = 1 + 3{,}3 \log n$

3. ln (Milone): $k = -1 + 2 \cdot \ln n$

4. $k = 1 + 10^{d} AT$,

onde n é o número de elementos da amostra, AT é a amplitude total dos dados e d é o número de decimais de seus elementos.

Deve-se lembrar que sendo k o número de classes, o resultado obtido por cada um dos critérios deve ser o número inteiro mais próximo ao obtido.

Milone (2004) acrescenta ainda que, adotando o princípio de que os agrupamentos devem ter no mínimo cinco e no máximo 20 classes, o critério da raiz é valido para $25 \leq n \leq 400$, o do log para $16 \leq n \leq 572.237$ e o do ln para $20 \leq n \leq 36.315$.

Por outro lado, o pesquisador pode definir o número de classes baseando-se em sua experiência.

Determinado o número de classes da distribuição de freqüências, o próximo passo é determinar a amplitude de cada classe, h, que por uma questão de bom senso deveria ser um número com a mesma precisão dos dados.

A amplitude de classe, h, é definida por:

$$h = \frac{AT}{k}$$

e assim todas as classes terão a mesma amplitude, o que permitirá a construção de gráficos e cálculo de medidas descritivas.

No caso de uma distribuição de freqüência contínua, ou em classes, uma outra coluna pode ser acrescentada à tabela. É a coluna dos pontos médios, denotada por x_i e definida como a média dos limites da classe:

$$x_i = \frac{l_i + L_i}{2}, \quad i = 1, ..., k .$$

Estes valores são utilizados na construção de gráfico e na obtenção de medidas descritivas com o auxílio de calculadoras.

Exemplo 07. Considere a variável idade dos alunos da Tabela 01. A Tabela 09 apresenta a distribuição de freqüência adequada.

Tabela 09 – Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

Idade	F_i
18	2
19	1
20	6
21	2
22	1
23	1
24	1
25	3
26	1
29	1
30	1
35	1
37	1
Total	22

Fonte: Tabela 01.

Pode-se observar que a tabela possui 13 linhas e que muitas delas, seguidas, apresentam freqüência igual a 1, o que mostra que o resumo da idade não apresenta uma distribuição satisfatória dos dados.

Segundo Montgomery (2003), ao passar dos dados brutos, que é o mesmo que os dados apresentados numa distribuição de freqüências pontual, para uma distribuição de freqüência em classes, algumas informações são perdidas, pois não se tem mais as observações individuais. Por outro lado, essa perda é pequena quando comparada ao ganho de concisão e de facilidade de interpretação da distribuição de freqüência.

Assim, para a idade, tem-se:

$$\text{At} = 37 - 18 = 19 \text{ anos}$$

$$\text{k} = \sqrt{22} = 4,69 \cong 5 \text{ classes}$$

$$\text{h} = \frac{19}{5} = 3,8 \cong 4 \text{ anos}$$

e, a distribuição de freqüência é dada na Tabela 10.

Tabela 10 – Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

Idade	x_i	F_i	f_i %	F_{a_i}	f_{a_i} %
18 |---22	20	11	50,00	11	50,00
22 |---26	24	6	27,27	17	77,27
26 |---30	28	2	9,09	19	86,36
30 |---34	32	1	4,55	20	90,91
34 |---38	36	2	9,09	22	100,00
Total	-	22	100,00	-	-

Fonte: Tabela 01.

Nota-se que cada um dos valores observados deve pertencer a uma e somente uma classe. É usual que o limite inferior da primeira classe seja igual ao menor valor observado e que o maior valor pertença à última classe. Quando o limite superior da última classe coincidir com o maior valor observado é mais apropriado fechar este intervalo, contando o elemento nesta classe, do que abrir uma nova classe contendo apenas uma freqüência absoluta. Por outro lado, se o maior valor observado for inferior ao limite superior da classe, não há problemas, pois fixamos todas as classes com a mesma amplitude.

Nada impede que se construa uma tabela com amplitude de classes desiguais. Isto dependerá do objetivo do pesquisador. O que se recomenda é o cuidado na interpretação da tabela. O primeiro passo é calcular as amplitudes das classes (Δ_i) e apresentá-las numa coluna. Em seguida, calculam-se as densidades de freqüências de cada classe, dividindo-se F_i por Δ_i, para conhecer a concentração por unidade da variável. Pode-se, também, calcular as densidades das proporções para se conhecer o percentual de concentração em cada classe (f_i / Δ_i).

Exemplo 08. Considere os dados do exemplo 07. A distribuição de freqüências com intervalos de classes desiguais é apresentada na Tabela 11.

Tabela 11 – Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

Idade	Frequência F_i	Amplitude Δ_i	Densidade F_i / Δ_i	Proporção f_i	Densidade f_i / Δ_i
18 |---20	3	2	1,50	0,14	0,07
20 |---22	9	2	4,50	0,40	0,20
22 |---24	2	2	1,00	0,09	0,05
24 |---28	5	4	1,25	0,23	0,06
28 |---38	3	10	0,30	0,14	0,01
Total	22	-	-	1,00	-

Fonte: Tabela 01.

Uma outra forma de construir uma distribuição de freqüências com amplitudes de classes desiguais é a que se apresenta na Tabela 12, onde a última classe não apresenta limite superior especificado. Isto poderia, também, ocorrer na primeira classe, mas agora com o limite inferior não especificado.

Tabela 12 – Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá.

Idade	F_i	f_i %	F_{a_i}	f_{a_i} %
18 |---20	3	14	3	14
20 |---22	8	36	11	50
22 |---24	2	9	13	59
24 |---26	4	18	17	77
Acima de 26	5	23	22	100
Total	22	100	-	-

Fonte: Tabela 01.

Com este tipo de distribuição dificuldades podem ocorrer na construção de gráficos e no cálculo da média, por exemplo.

Exercício 02. Construa uma distribuição de freqüência completa para as variáveis da Tabela 01: a) número de disciplinas reprovadas no 1º ano do curso; b) peso.

3 GRÁFICOS

Gráfico é um recurso visual da Estatística utilizado para representar um fenômeno. Sua utilização em larga escala nos meios de comunicação social, técnica e científica, devem-se tanto à sua capacidade de refletir padrões gerais e particulares do conjunto de dados em observação, como à facilidade de interpretação e a eficiência com que resume informações dos mesmos.

Embora os gráficos forneçam menor grau de detalhes que as tabelas, estes apresentam um ganho na compreensão global dos dados, permitindo que se aperceba imediatamente da sua forma geral sem deixar de evidenciar alguns aspectos particulares que sejam de interesse do pesquisador. Uma representação gráfica coloca em evidência as tendências, as ocorrências ocasionais, os valores mínimos e máximos e também as ordens de grandezas dos fenômenos que estão sendo observados.

Todo gráfico, em sua versão final deve primar pela simplicidade, clareza e veracidade nas informações. Para atingir tal objetivo, a construção de um gráfico exige muito trabalho e cuidados. Segundo Silva (apud WALLGREN, 1996), a escolha da representação gráfica e, conseqüentemente, a escolha do tipo de gráfico mais adequado para representar um conjunto de dados deve ser feita com base nas respostas de questões como:

- Um gráfico realmente é a melhor opção?
- Qual é o público-alvo?
- Qual é o objetivo do gráfico?
- Que tipo de gráfico deve ser usado?
- Como o gráfico deve ser apresentado?
- Que tamanho o gráfico deve ter?
- Deverá ser usado apenas um gráfico?
- A qual meio técnico se deve recorrer?

Ao incluir um gráfico em um trabalho, sua identificação deve aparecer na parte inferior, precedido pela palavra Gráfico seguida de seu número de ordem de ocorrência no texto (algarismos arábicos), de seu respectivo título e/ou legenda explicativa de maneira breve e clara (dispensando a leitura do texto) e da fonte de onde se extraiu os dados. Uma regra básica para a elaboração adequada do título de qualquer gráfico, é verificar se o mesmo responde a três exigências: o quê, onde e quando.

Quando um gráfico for inserido em um texto, recomenda-se que este seja destacado tanto do texto que o precede, como do texto imediatamente subseqüente, por meio de três espaços simples.

O título é escrito em letras minúsculas, exceto a inicial da frase e dos nomes próprios. Deve ser separado da numeração do gráfico por um hífen seguido de um espaço. Caso seja composto por mais de uma linha, estas devem ser alinhadas sob a primeira letra da primeira linha do título.

Em casos onde a legenda se fizer necessário, como nos casos de gráficos comparativos, ela deve ser colocada à direita ou abaixo do gráfico.

A proporção da altura identificada no eixo vertical deve ser, aproximadamente, de 60% a 70% da largura da abscissa, ou seja, do eixo horizontal.

Dá-se preferência a pouca variação de cores. A variação de cores num mesmo gráfico é recomendada para o caso de gráficos comparativos.

No final das linhas que definem os eixos (abscissa e ordenada), devem ser expostas as unidades utilizadas nas escalas que mensuram as grandezas representadas.

Existem diversos tipos de gráficos. Porém, neste trabalho serão destacados aqueles de maior interesse pedagógico na representação das variáveis qualitativas e quantitativas.

3.1 GRÁFICOS PARA VARIÁVEIS QUALITATIVAS

3.1.1 Gráfico de barras

É um gráfico formado por retângulos horizontais de larguras iguais, onde cada um deles representa a intensidade de uma modalidade ou atributo.

É recomendável que cada coluna conserve uma distância entre si de aproximadamente 2/3 da largura da base de cada barra, evidenciando deste modo, a não continuidade na seqüência dos dados.

O objetivo deste gráfico é de comparar grandezas e é recomendável para variáveis cujas categorias tenham designações extensas.

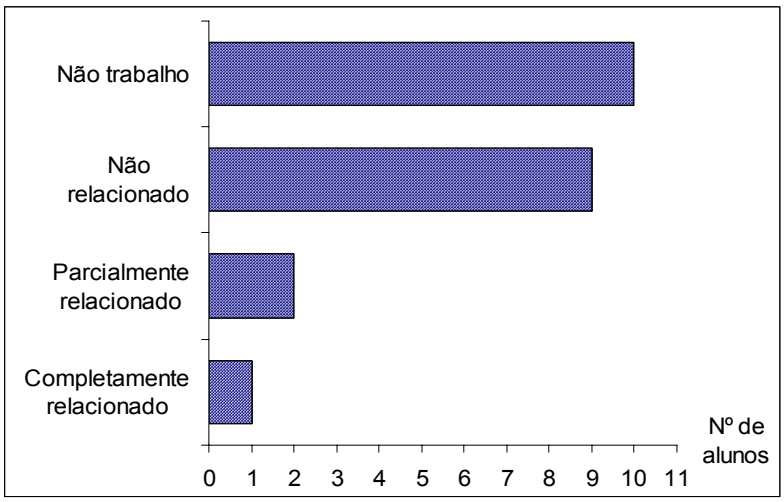


Figura 02 - Relação trabalho e curso dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, 21/03/2005.

Fonte: Tabela 01.

3.1.2 Gráfico de colunas

É o gráfico mais utilizado para representar variáveis qualitativas. Difere do gráfico de barras por serem seus retângulos dispostos verticalmente ao eixo das abscissas sendo mais indicado quando as designações das categorias são breves. Também para este tipo de gráfico deve ser preservada a distância entre cada retângulo de, aproximadamente, 2/3 da largura da base de cada coluna. O número de colunas ou barras do gráfico não deve ser superior a 12 (doze).

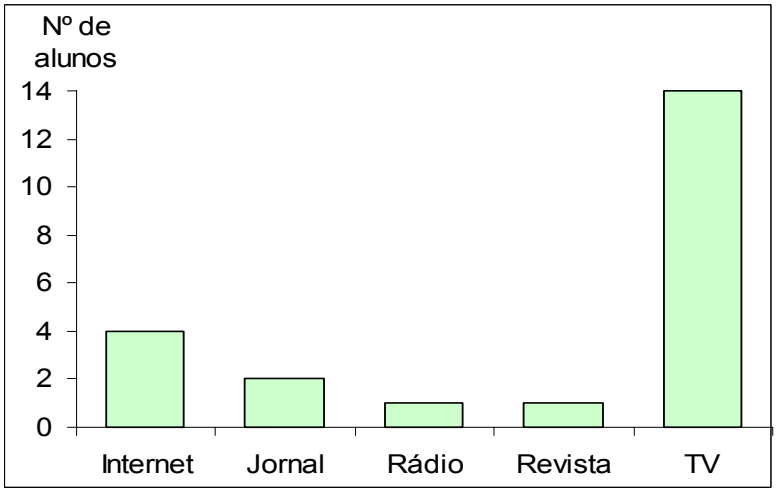


Figura 03 - Meios de informação utilizados pelos alunos da disciplina Inferência Estatística, curso de Estatística da UEM, 21/03/2005.

Fonte: Tabela 01.

Ao se descrever simultaneamente duas ou mais categorias para uma variável, é conveniente fazer uso dos gráficos de barras ou colunas justapostas (ou sobrepostas), chamados de gráficos comparativos. De acordo com as normas contidas em Gráficos (UFPR, 2001), este tipo de gráfico só deve ser utilizado quando apresentar até três elementos para uma série de no máximo quatro valores.

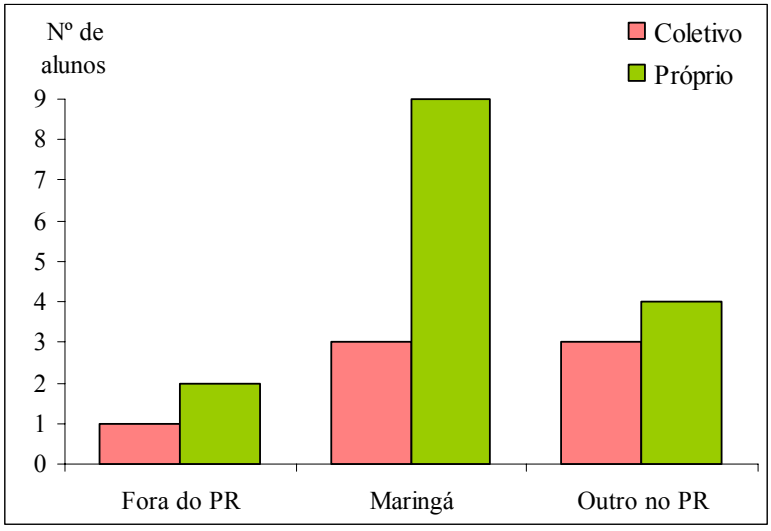


Figura 04 - Município de procedência segundo o tipo de transporte utilizado pelos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, 21/03/2005.
Fonte: Tabela 01.

3.1.3 Gráfico de setores

Tipo de gráfico onde a variável em estudo é projetada num círculo, de raio arbitrário, dividido em setores com áreas proporcionais às freqüências das suas categorias. São indicados quando se deseja comparar cada valor da série com o total. Recomenda-se seu uso para o caso em que o número de categorias não é grande e não obedecem a alguma ordem específica.

A Figura 05 mostra um gráfico de setores para a variável município de procedência que constam na Tabela 01. O procedimento para o cálculo do ângulo correspondente a cada categoria é feito por meio de simples proporções: 360º que corresponde a um círculo completo está para o total de alunos entrevistados, 22, assim como xº está para o total de alunos que pertencem à categoria desejada. Por exemplo, os 54% de alunos que residem no município de Maringá corresponderá a um ângulo x resultante da expressão $\frac{360^o}{22} = \frac{x^o}{12}$, cujo valor é aproximadamente 196º.

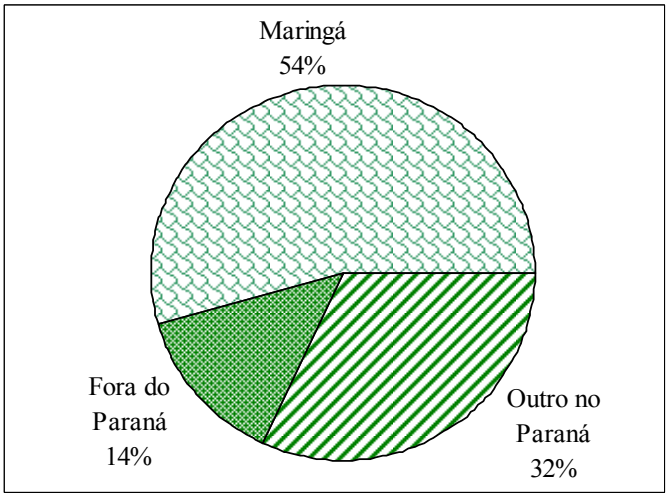


Figura 05 - Município de procedência dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, 21/03/2005.
Fonte: Tabela 01.

3.1.4 Gráfico de linhas

Sua aplicação é mais indicada para representações de séries temporais sendo por tal razão, conhecidos também como gráficos de séries cronológicas. Sua construção é feita colocando-se no eixo vertical (y) a mensuração da variável em estudo e na abscissa (x), as unidades da variável numa ordem crescente. Este tipo de gráfico permite representar séries longas, o que auxilia detectar suas flutuações tanto quanto analisar tendências. Também podem ser representadas várias séries em um mesmo gráfico.

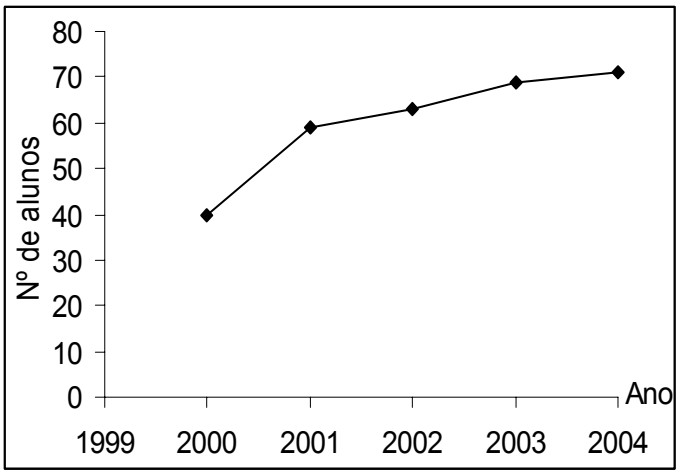


Figura 06 - Número de matrículas anuais na disciplina Probabilidade do curso de Estatística da UEM, 21/03/2005.
Fonte: Tabela 01.

Exercício 03: Considerando as informações sobre os alunos da disciplina de Inferência Estatística do curso de Estatística da Uem, contidas na Tabela 01, construa um gráfico adequado para representar as variáveis: a) sexo; b) estado civil; c) transporte; d) meios de informação segundo o sexo dos alunos.

3.2 GRÁFICOS PARA VARIÁVEIS QUANTITATIVAS DISCRETAS

3.2.1 Gráfico de bastões

Este gráfico é formado por segmentos de retas perpendiculares ao eixo horizontal (eixo da variável), cujo comprimento corresponde à freqüência absoluta ou relativa de cada elemento da distribuição. Suas coordenadas não podem ser unidas porque a leitura do gráfico deve tornar claro que não há continuidade entre os valores individuais assumidos pela variável em estudo.

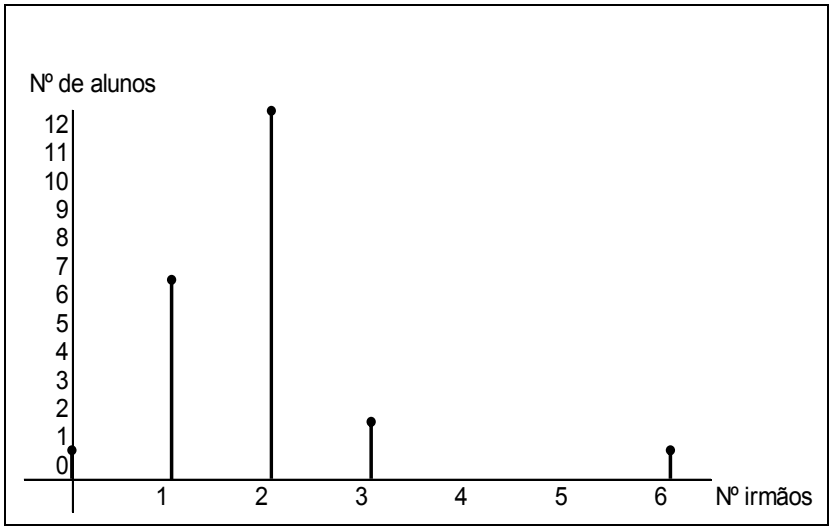


Figura 07 – Número de irmãos dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, 21/03/2005.

Fonte: Tabela 01.

3.2.2 Gráfico da freqüência acumulada

A Figura 08 mostra o gráfico para freqüência acumulada de uma variável quantitativa discreta. Na abscissa são alocados os valores assumidos pela variável número de irmãos e no eixo das ordenadas suas freqüências acumuladas. Observa-se que a leitura do gráfico exige alguns

cuidados básicos: caso o valor da variável esteja ou não incluído, sua freqüência acumulada difere. Se for de interesse saber quantos alunos tem dois ou menos irmãos (inclui-se dois irmão), a freqüência acumulada é de 19 alunos. Caso se queira apenas saber quantos alunos têm menos de dois irmãos (portanto o número dois não está incluso), sua freqüência acumulada é de 7 alunos.

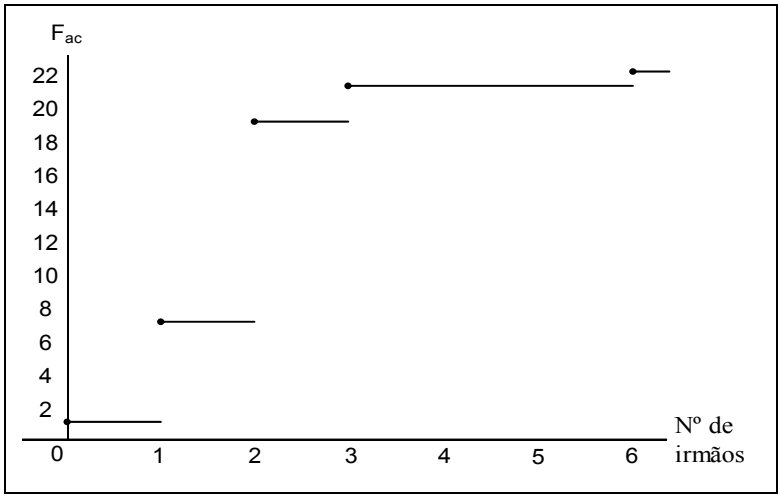


Figura 08 - Número acumulado de irmãos dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, 21/03/2005.

Fonte: Tabela 01.

3.3 GRÁFICOS PARA VARIÁVEIS QUANTITATIVAS CONTÍNUAS

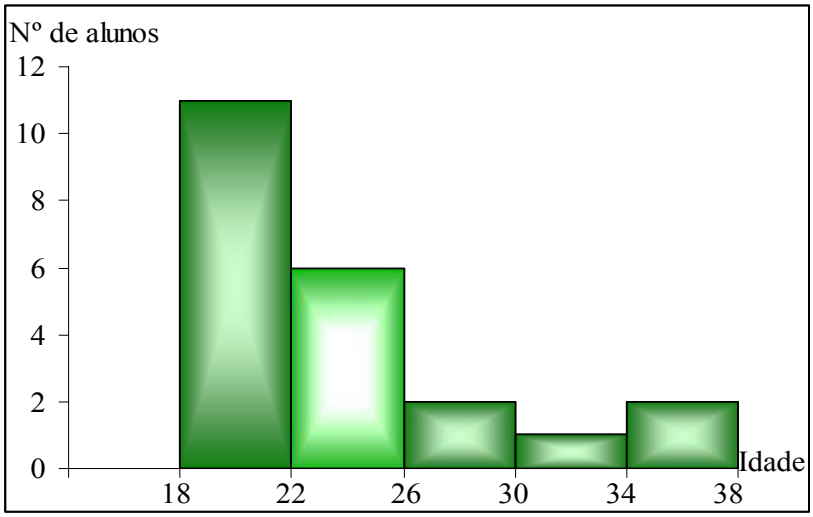
3.3.1 Histograma

É um gráfico de colunas justapostas que representa uma distribuição de freqüência para dados contínuos ou uma variável discreta quando esta apresentar muitos valores distintos.

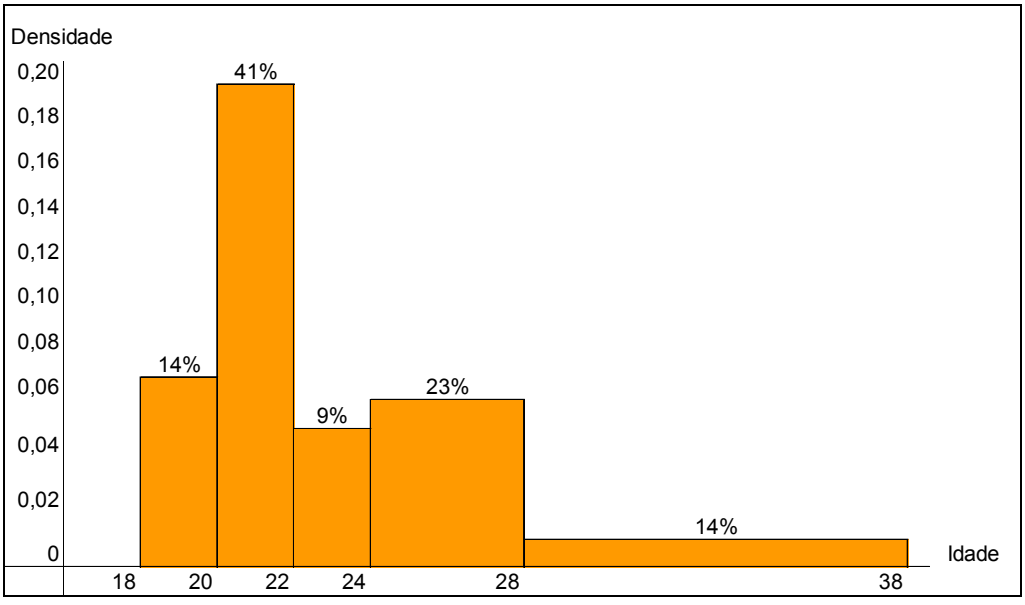
No eixo horizontal são dispostos os limites das classes segundo as quais os dados foram agrupados enquanto que o eixo vertical corresponde às freqüências absolutas ou relativas das mesmas.

Quando os dados são distribuídos em classes de mesma amplitude, Figura 09 (a), todas as colunas apresentam bases iguais com alturas variando em função das suas freqüências absolutas ou relativas. Neste caso, tem-se que a área de cada retângulo depende apenas da sua altura enquanto que no caso de dados agrupados em classes de dimensões diferentes, como mostra a Figura 9 (b), a área de cada coluna já não é mais proporcional à sua altura. Como a altura de cada classe precisa variar simultaneamente com sua largura, é necessário que a área de cada uma das colunas

permaneça em proporção conveniente, o que pode ser obtido dividindo-se as freqüências das classes pelas respectivas amplitudes e construindo-se o histograma a partir destas freqüências. Portanto, pode-se dizer que no primeiro caso, o eixo dos valores informa sobre a freqüência relativa de cada classe, no segundo caso, tal procedimento perde todo significado, e é necessário comparar as áreas para interpretar as informações que são expostas.



(a) Amplitudes iguais.



(b) Amplitudes desiguais.

Figura 09 - Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, 21/03/2005.

Fonte: Tabela 01.

3.3.2 Polígono de freqüência

É um gráfico de linha cuja construção é feita unindo-se os pontos de coordenadas de abscissas correspondentes aos pontos médios de cada classe e as ordenadas, às freqüências absolutas ou relativas dessas mesmas classes.

O polígono de freqüência é um gráfico que deve ser fechado no eixo das abscissas. Então, para finalizar sua elaboração, deve-se acrescentar à distribuição, uma classe à esquerda e outra à direita, ambas com freqüências zero. Tal procedimento permite que a área sob a linha de freqüências seja igual à área do histograma.

Uma das vantagens da aplicação de polígonos de freqüências é que, por serem gráficos de linhas, permitem a comparação entre dois ou mais conjuntos de dados por meio da superposição dos mesmos.

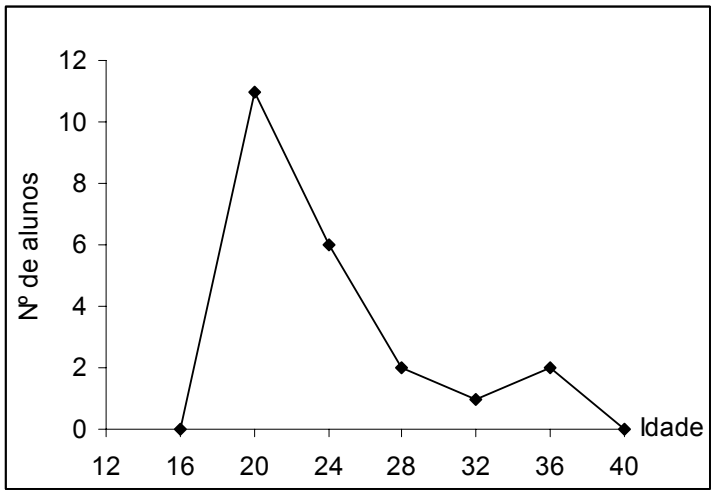


Figura 10 - Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, 21/03/2005.

Fonte: Tabela 01.

3.3.3 Gráfico da freqüência acumulada ou Ogiva

É um gráfico que permite descrever dados quantitativos por meio da freqüência acumulada. A ogiva é um gráfico de linha que une os pontos cujas abscissas são os limites superiores das classes, e, ordenadas suas respectivas freqüências acumuladas. Convém observa-se que o ponto

inicial desse gráfico é o limite inferior do primeiro intervalo, com freqüência acumulada zero, pois não existe qualquer valor inferior a ele.

Quando os dados contidos em cada classe são distribuídos uniformemente, pode-se estimar, a partir da ogiva, o número de elementos pertencentes a qualquer uma das classes que compõe a distribuição de freqüência dos dados e a quantidade ou porcentagem de elementos que estão abaixo de certo valor pertencente ao conjunto de dados.

Pela Figura 11, nota-se que não existem alunos com idade inferior a 18 anos enquanto que abaixo de 34 anos existem vinte alunos.

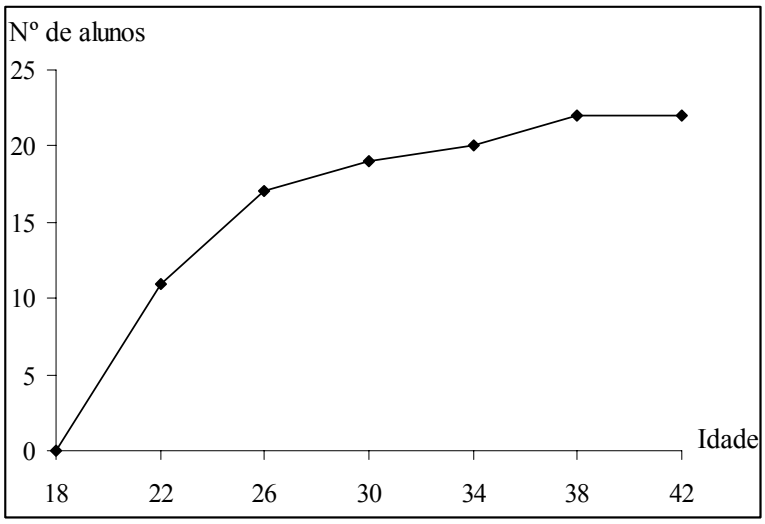


Figura 11 - Idade acumulada dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, 21/03/2005.

Fonte: Tabela 01.

Exercício 04: Considerando as informações sobre os alunos da disciplina de Inferência Estatística do curso de Estatística da UEM, contidas na Tabela 01, construa os gráficos adequados para as variáveis: a) peso; b) altura; c) número de reprovas no 1º ano de curso.

3.3.4 Ramo-e-Folhas

O diagrama Ramo-e-Folhas, criado por John Tukey, é um procedimento utilizado para armazenar os dados sem perda de informação. É utilizado para se ter uma idéia visual da distribuição dos dados. Cada valor observado, x_i, da variável X, deve consistir de no mínimo dois dígitos e a variável pode ser tanto quantitativa discreta como contínua.

Para construí-lo, divide-se cada número em duas partes. A primeira é denominada ramo e a segunda, folhas. O ramo consistirá de um ou mais dígitos iniciais se o valor da variável for um número inteiro e do número inteiro, se o valor da variável for um número com decimais. Nas folhas, colocam-se os dígitos restantes se o valor observado for número inteiro, ou os decimais, caso contrário. A Figura 12 (a) apresenta o ramo-e-folhas correspondente a variável idade do aluno da Tabela 01.

Observa-se que o ramo correspondente ao dígito 2 tem muitas folhas. Neste caso, a opção é dividir este ramo em dois: as folhas de 0 a 4 pertencerão a uma linha e as folhas de 5 a 9 pertencerão à outra linha. Os ramos são discriminados por um sinal no seu expoente, como na Figura 12 (b).

Ramo	Folha	Freqüência
1	8 8 9	3
2	0 0 0 0 0 0 0 1 1 2 3 4 5 5 6 9	16
3	5 7	2

(a) Sem divisão de ramos.

Ramo	Folha	Freqüência
1	8 8 9	3
2	0 0 0 0 0 0 0 1 1 2 3 4	12
2*	5 5 6 9	4
3	5 7	2

(b) Com divisão de ramos.

Figura 12 - Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

Fonte: Tabela 01.

Exercício 05: Construir o ramo-e-folhas para a variável altura, da Tabela 01.

4 MEDIDAS DESCRITIVAS

Uma outra maneira de se resumir os dados de uma variável quantitativa, além de tabelas e gráficos, é apresentá-los na forma de valores numéricos, denominados medidas descritivas. Estas medidas, se calculadas a partir de dados populacionais, são denominadas parâmetros e se calculadas a partir de dados amostrais são denominadas estimadores ou estatísticas.

As medidas descritivas auxiliam a análise do comportamento dos dados. Tais dados são provenientes de uma população ou de uma amostra, o que exige uma notação específica para cada caso, conforme mostra o Quadro 01.

Quadro 01: Notações de algumas estatísticas.

Medidas	Parâmetros	Estimadores
Número de elementos	N	n
Média	μ	$\overline{X}$
Variância	σ^2	S^2
Desvio padrão	σ	S

Classificam-se as medidas descritivas como: medidas posição (tendência central e separatrizes), medidas de dispersão, medidas de assimetria e de curtose.

4.1 MEDIDAS DE TENDÊNCIA CENTRAL

As medidas de tendência central são assim denominadas por indicarem um ponto em torno do qual se concentram os dados. Este ponto tende a ser o centro da distribuição dos dados. Reis (1998), afirma que:

> o valor a escolher depende das características dos dados. Por exemplo, num estudo agrícola sobre a produção de trigo por hectare de terra arável podemos estar interessados em conhecer o valor mais elevado da produtividade do solo agrícola das várias explorações analisadas. Num outro estudo sobre os resultados de uma turma de estudantes universitários talvez seja mais interessante conhecer o resultado médio obtido por 50% dos estudantes. Num outro estudo sobre os rendimentos per capta dos países da CEE, a comparação entre países será facilitada se calcularmos os rendimentos médios de cada país.

A seguir, são definidas as principais medidas de tendência central: média, mediana e moda.

4.1.1 Média aritmética

A média aritmética ($\overline{X}$) é a soma de todos os valores observados da variável dividida pelo número total de observações. Sob uma visão geométrica a média de uma distribuição é o centro de gravidade, representa o ponto de equilíbrio de um conjunto de dados. É a medida de tendência central mais utilizada para representar a massa de dados.

Seja (x_1, ..., x_n) um conjunto de dados. A média é dada por:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \quad \text{ou} \quad \overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

para dados populacionais ou amostrais, respectivamente. Caso os dados estejam apresentados segundo uma distribuição de freqüência, tem-se:

$$\mu = \frac{\sum_{i=1}^{k} x_i F_i}{N} \quad \text{ou} \quad \overline{X} = \frac{\sum_{i=1}^{k} x_i F_i}{n}.$$

Observe que no caso de dados agrupados a média é obtida a partir de uma ponderação, onde os pesos são as freqüências absolutas de cada classe e x_i é o ponto médio da classe i.

Citam-se a seguir, algumas propriedades da média aritmética:

1. a média é um valor calculado facilmente e depende de todas as observações;

2. é única em um conjunto de dados e nem sempre tem existência real, ou seja, nem sempre é igual a um determinado valor observado;

3. a média é afetada por valores extremos observados;

4. por depender de todos os valores observados, qualquer modificação nos dados fará com que a média fique alterada. Isto quer dizer que somando-se, subtraindo-se, multiplicando-se ou dividindo-se uma constante a cada valor observado, a média ficará acrescida, diminuída, multiplicada ou dividida desse valor.

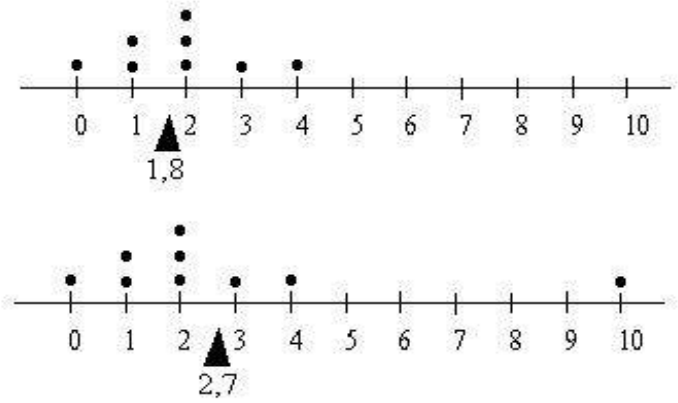
5. a soma da diferença de cada valor observado em relação à média é zero, ou seja, a soma dos desvios é zero.

$$\sum (x_i - \overline{x}) = 0$$

A propriedade 5, é de extrema importância para a definição de variância, uma medida de dispersão a ser definida posteriormente.

Destaca-se, ainda, que a propriedade 3, quando se observam no conjunto dados discrepantes, faz da média uma medida não apropriada para representar os dados. Neste caso, não existe uma regra prática para a escolha de uma outra medida. O ideal é, a partir da experiência do pesquisador,

decidir pela moda ou mediana. Para ilustrar, considere o número de filhos, por família, para um grupo de 8 famílias: 0, 1, 1, 2, 2, 2, 3, 4. Neste caso, a média é $\overline{x} = 1,875$ filhos por família. Entretanto, incluindo ao grupo uma nova família com 10 filhos, a média passa a ser $\overline{x} = 2,788$, o que eleva em 48,16% o número médio de filhos por família. Assim, ao observar a média, pode-se pensar que a maior parte das famílias deste grupo tem três filhos quando, na verdade, apenas uma tem três filhos.



Exemplo 09: Considerando a idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, a idade média é

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{20+26+18+\cdots+21+22}{22} = \frac{518}{22} = 23,5 \text{ anos}$$

Assim, a idade média dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá é 23,5 anos.

No entanto, ao considerar os dados agrupados como na Tabela 10, a média é:

$$\overline{X} = \frac{\sum_{i=1}^{5} x_i F_i}{n} = \frac{20 \cdot 11 + 24 \cdot 6 + \cdots + 36 \cdot 2}{22} = \frac{524}{22} = 23,8 \text{ anos.}$$

Nota-se que esta diferença ocorre devido ao fato de se utilizar os dados sem o conhecimento de seus valores individuais. Neste caso, tornou-se necessário representá-los pelos pontos médios de suas respectivas classes resultando numa certa perda de informação.

Exercício 06: Calcule a média aritmética para a variável altura dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM. a) utilizando os dados brutos; b) utilizando a distribuição de freqüência (dados agrupados).

4.1.2 Moda

A moda (M_o) é o valor que apresenta a maior freqüência da variável entre os valores observados. Para o caso de valores individuais, a moda pode ser determinada imediatamente observando-se o rol ou a freqüência absoluta dos dados. Por outro lado, em se tratando de uma distribuição de freqüência de valores agrupados em classes, primeiramente é necessário identificar a classe modal, aquela que apresenta a maior freqüência, e a seguir a moda é calculada aplicando-se a fórmula:

$$M_o = l_i + \frac{h(F_i - F_{i-1})}{(F_i - F_{i-1}) + (F_i - F_{i+1})}$$

onde

i é a ordem da classe modal;

l_i é o limite inferior da classe modal;

h é a amplitude da classe modal;

F_i é a freqüência absoluta da classe modal;

F_{i-1} é a freqüência absoluta da classe anterior à classe modal;

F_{i+1} é a freqüência absoluta da classe posterior à classe modal.

É relevante salientar que um conjunto de dados pode apresentar todos seus elementos com a mesma freqüência absoluta, e neste caso não existirá um valor modal, o que significa que a distribuição será classificada como amodal. Pode ocorrer, também, casos em que a seqüência de observações apresente vários elementos com freqüência iguais, implicando numa distribuição plurimodal.

O uso da moda é mais indicado quando se deseja obter, rapidamente, uma medida de tendência central. Um outro aspecto que favorece a utilização da moda é que seu valor não é afetado pelos valores extremos do conjunto de dados analisado.

Exemplo 10: A moda da idade dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, determinada pontualmente, é Mo= 20 anos. Isto significa que a idade mais freqüente entre estes alunos é de 20 anos.

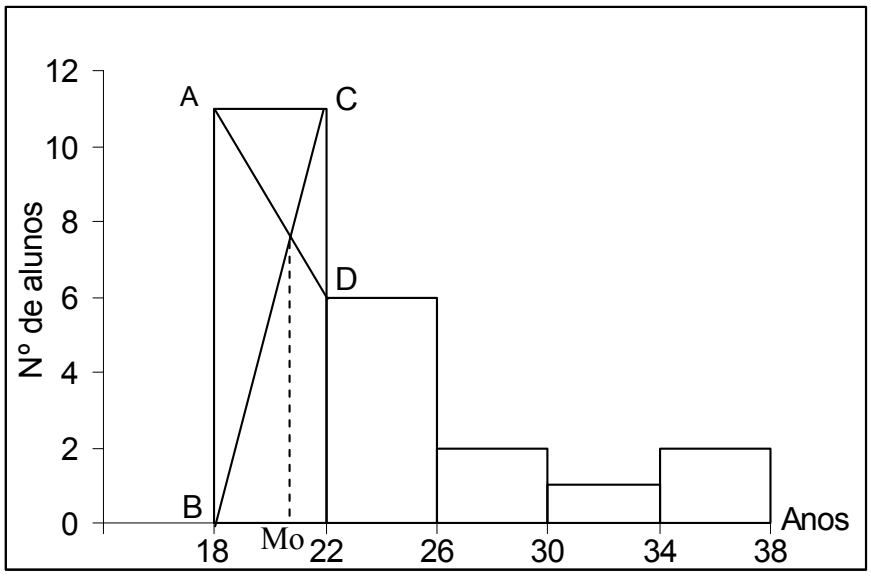
Ao considerar a distribuição apresentada na Tabela 10, a moda é

$$M_o = l_i + \frac{h \cdot (F_i - F_{i-1})}{(F_i - F_{i-1}) + (F_i - F_{i+1})} = 18 + \frac{4 \cdot (11-0)}{(11-0) + (11-6)} = 18 + \frac{44}{16} = 18 + 2,75 = 20,75 \text{ anos.}$$

A interpretação é análoga à determinada pontualmente.

Exercício 07: Calcule a moda para a variável altura dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM. a) utilizando os dados brutos; b) utilizando a distribuição de freqüência (dados agrupados).

Graficamente, utilizando-se um conjunto de dados hipotéticos, identifica-se a classe modal como aquela que apresenta o retângulo de maior altura (freqüência). A intersecção das retas que unem os pontos AD e os pontos BC, determina o ponto P que, projetado perpendicularmente no eixo da variável, corresponderá ao valor da moda M_o.



4.1.3 Mediana

A mediana (M_d) é o valor que ocupa a posição central da série de observações de uma variável, em rol, dividindo o conjunto em duas partes iguais, ou seja, a quantidade de valores inferiores à mediana é igual à quantidade de valores superiores a mesma.

Exemplo 11: Retomando o exemplo do número de filhos por famílias, verifica-se que:

Para o caso de oito famílias, n=8, a mediana é determinada como a seguir:

X	x_1	x_2	x_3	x_4		x_5	x_6	x_7	x_8
Valor observado	0	1	1	2	$\frac{x_4+x_5}{2}$	2	2	3	4

← 4 observações → Md=2 ← 4 observações →

Quando se acrescenta ao grupo uma outra família com 10 filhos o tamanho da amostra passa

a ser n=9. Neste caso, a mediana é:

X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Valor observado	0	1	1	1	2	2	3	4	10

← 4 observações → Md=2 ← 4 observações →

Observe que nos dois casos, por coincidência, a mediana manteve-se a mesma, M_d=2, significando que 50% das famílias possuem menos de 2 filhos ou 50% possuem mais de 2 filhos. Mostra-se assim, que a mediana não é influenciada por valores extremos.

Este procedimento pode tornar-se inadequado quando o conjunto de dados for composto por muitos elementos. Os passos a seguir indicam uma forma para o cálculo da mediana, independentemente do tamanho da amostra.

Ordenar as observações em ordem crescente ou decrescente (rol).

Calcular a posição, p, que a mediana ocupa no conjunto de dados:

$$p = 0,50(n+1)$$

Obter a mediana pela equação

$$M_d = x_{I_p} + F_p(x_{I_{p+1}} - x_{I_p})$$

onde I_p é a parte inteira de p e F_p a parte fracionária (ou decimal).

Exemplo 12: Considere o rol da idade dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM:

18, 18, 19, 20, 20, 20, 20, 20, 20, 21, 21, 22, 23, 24, 25, 25, 25, 26, 29, 30, 35, 37

A posição p da mediana é

$$p = 0,50(22+1) = 11,5\,.$$

Assim,

$$M_d = x_{11} + 0,5\cdot(x_{12} - x_{11}) = 21 + 0,5\cdot(22-21) = 21,5 \text{ anos},$$

logo, 50% dos alunos têm idade inferior a 21,5 anos.

Para os dados em distribuição de freqüências em classes, tem-se:

$$M_d = l_i + \frac{h\left(p - F_{a_{i-1}}\right)}{F_i}$$

onde:

$p = \frac{n}{2}$ indica a posição central da série;

i é a ordem da classe que contém o menor valor de F_{a_i}, tal que $F_{a_i} \geq p$;

$F_{a_{i-1}}$ é a freqüência acumulada da classe anterior à da mediana.

Exemplo 13: Ao considerar a distribuição apresentada na Tabela 10, a mediana é

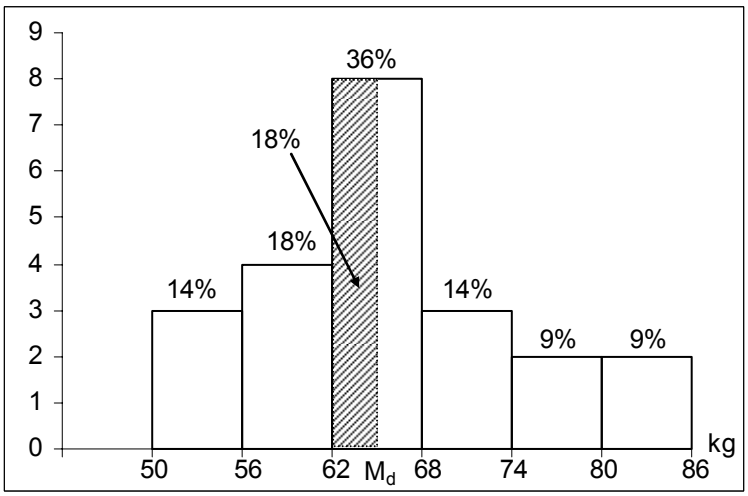
$$p = \frac{22}{2} = 11 \Rightarrow F_{a_i} \geq 11 \Rightarrow i = 1$$

$$M_d = l_i + \frac{h \cdot \left(p - F_{a_{i-1}}\right)}{F_i} = 18 + \frac{4 \cdot (11-0)}{11} = 18 + \frac{44}{11} = 18 + 4 = 22 \text{ anos.}$$

A idade mediana é 22 anos, ou seja, 50% dos alunos que cursam a disciplina Inferência Estatística do curso de Estatística da UEM têm idade inferior ou igual a 22 anos.

Exercício 08: Calcule a mediana para a variável altura dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM. a) utilizando os dados brutos; b) utilizando a distribuição de freqüência (dados agrupados).

Para ilustrar graficamente o cálculo da mediana, considere novamente um conjunto de pesos fictícios. Deve-se localizar no eixo da variável o ponto que divide o histograma ao meio. Isto é feito somando-se as áreas (freqüências relativas) até que se obtenha 50%. No histograma abaixo, a classe que contém a mediana é a classe de 62 a 68 kg, com freqüência relativa igual a 36%. Pode-se observar então que faltam 18%, 50%-(14%+18%) para completar 50% da distribuição. Tem-se então que o limite superior da base do retângulo hachurado é a mediana da distribuição.



Aplicando a proporcionalidade entre área e base do retângulo resultará na mediana:

$$\frac{68-62}{36\%} = \frac{M_d - 62\%}{18\%}$$

Portanto a mediana é igual a 65 kg.

4.2 MEDIDAS SEPARATRIZES

Estas medidas são valores que ocupam posições no conjunto de dados, em rol, dividindo-o em partes iguais e podem ser:

Quartil: Os quartis dividem o conjunto de dados em quatro partes iguais.

Quadro 02: Descrição dos quartis (dados amostrais).

Estatística	Notação	Interpretação	Posição
1º quartil	Q_1	25% dos dados são valores menores ou iguais ao valor do primeiro quartil.	p=0,25(n+1)
2º quartil	$Q_2 = M_d$	50% dos dados são valores menores ou iguais ao valor do segundo quartil.	p=0,50(n+1)
3º quartil	Q_3	75% dos dados são valores menores ou iguais ao valor do terceiro quartil.	p=0,75(n+1)

Decil: Os decis dividem o conjunto de dados em dez partes iguais.

Quadro 03: Descrição dos decis (dados amostrais).

Estatística	Notação	Interpretação	Posição
1º decil	D_1	10% dos dados são valores menores ou iguais ao valor do primeiro decil.	p=0,10(n+1)
2º decil	D_2	20% dos dados são valores menores ou iguais ao valor do segundo decil.	p=0,20(n+1)
3º decil	D_3	30% dos dados são valores menores ou iguais ao valor do terceiro decil.	p=0,30(n+1)
4º decil	D_4	40% dos dados são valores menores ou iguais ao valor do primeiro decil.	p=0,40(n+1)
5º decil	$D_5=Q_2=Md$	50% dos dados são valores menores ou iguais ao valor do segundo decil.	p=0,50(n+1)
6º decil	D_6	60% dos dados são valores menores ou iguais ao valor do terceiro decil.	p=0,60(n+1)
7º decil	D_7	70% dos dados são valores menores ou iguais ao valor do primeiro decil.	p=0,70(n+1)
8º decil	D_8	80% dos dados são valores menores ou iguais ao valor do segundo decil.	p=0,80(n+1)
9º decil	D_9	90% dos dados são valores menores ou iguais ao valor do terceiro decil.	p=0,90(n+1)

Percentil: Os percentis dividem o conjunto de dados em cem partes iguais. A seguir são apresentados alguns dos percentis mais usados:

Quadro 04: Descrição de alguns percentis (dados amostrais).

Estatística	Notação	Interpretação	Posição
5º Percentil	P_5	5% dos dados são valores menores ou iguais ao valor do primeiro percentil.	p=0,05(n+1)
10º Percentil	P_{10}	10% dos dados são valores menores ou iguais ao valor do décimo percentil.	p=0,10(n+1)
25º Percentil	$P_{25}=Q_1$	25% dos dados são valores menores ou iguais ao valor do percentil cinquenta.	p=0,25(n+1)
50º Percentil	$P_{50}= D_5 =Q_2 = Md$	50% dos dados são valores menores ou iguais ao valor do primeiro percentil.	p=0,50(n+1)
75º Percentil	$P_{75}=Q_3$	75% dos dados são valores menores ou iguais ao valor do primeiro percentil. (Q_3)	p=0,75(n+1)
90º Percentil	P_{90}	90% dos dados são valores menores ou iguais ao valor do percentil noventa.	p=0,90(n+1)
95º Percentil	P_{95}	95% dos dados são valores menores ou iguais ao valor do percentil noventa e cinco.	p=0,95(n+1)

Para os dados em rol, o cálculo das medidas separatrizes é a mesma que a da mediana, a saber:

$$S_k = x_{I_p} + F_p(x_{I_{p+1}} - x_{I_p})$$

onde I_p é a parte inteira de p e F_p a parte fracionária (ou decimal).

Para os dados em distribuição de freqüências em classes, o cálculo das medidas separatrizes é a mesma que a da mediana, a saber:

$$S_k = l_i + \frac{h\left(p - F_{a_{i-1}}\right)}{F_i}$$

onde:

$p = \frac{n}{4}k$, com $k = 1, 2, 3$, para determinação dos quartis;

$p = \frac{n}{10}k$, $k = 1, 2, \cdots, 9$ para o cálculo dos decis; e

$p = \frac{n}{100}k$, $k = 1, 2, \cdots, 99$ para os percentis;

i é a ordem da classe que contém o menor valor de F_{a_i}, tal que $F_{a_i} \geq p$;

$F_{a_{i-1}}$ é a freqüência acumulada da classe anterior à da separatriz.

Exemplo 14: Considerando o rol do exemplo 12, o terceiro quartil e o quadragésimo percentil são:

Terceiro quartil: $p = 0,75(22+1) = 17,25$ e

$Q_3 = S_3 = x_{17} + 0,25(x_{18} - x_{17}) = 25 + 0,25 \cdot (26 - 25) = 25,25$ anos.

Assim, pode-se afirmar que 75% dos alunos que cursam a disciplina Inferência Estatística do curso de Estatística da UEM têm idade inferior ou igual a 25,25 anos.

Quadragésimo percentil: $p = 0,40(22 + 1) = 9,2$ e

$P_{40} = S_{40} = x_9 + 0,20(x_{10} - x_9) = 20 + 0,20 \cdot (21 - 20) = 20,2$ anos.

Logo, 40% dos alunos que cursam a disciplina Inferência Estatística do curso de Estatística da UEM têm idade inferior ou igual a 20,2 anos.

Exemplo 15: Em continuação ao exemplo 14, da Tabela 10 tem-se:

Primeiro quartil: $p = \frac{n}{4}k = \frac{22}{4}1 = 5,5 \Rightarrow F_{a_i} \geq 5,5 \Rightarrow i = 1$

$$Q_1 = l_1 + \frac{h\left(p - F_{a_{1-1}}\right)}{F_1} = 18 + \frac{4(5,5 - 0)}{11} = 20 \text{ anos,}$$

Terceiro quartil: $p = \frac{n}{4}k = \frac{22}{4}3 = 16,5 \Rightarrow F_{a_i} \geq 16,5 \Rightarrow i = 2$

$$Q_3 = l_2 + \frac{h\left(p - F_{a_{2-1}}\right)}{F_2} = 22 + \frac{4 \cdot (16,5 - 11)}{6} = 25,67 \text{ anos}$$

Sétimo decil: $p = \frac{n}{10}k = \frac{22}{10}7 = 15,4 \Rightarrow F_{a_i} \geq 17 \Rightarrow i = 2$

$$D_7 = l_2 + \frac{h(p - F_{a_1})}{F_2} = 22 + \frac{2(15,4 - 11)}{6} = 23,47 \text{ anos}$$

Nonagésimo percentil: $p = \frac{n}{100}k = \frac{22}{100}90 = 19,8 \Rightarrow F_{a_i} \geq 19,8 \Rightarrow i = 4$

$$P_{90} = l_4 + \frac{h\left(p - F_{a_3}\right)}{F_4} = 30 + \frac{4(19,8 - 19)}{1} = 33,2 \text{ anos}$$

Conclui-se, que 25% dos alunos que cursam a disciplina Inferência Estatística do curso de Estatística da UEM têm idade inferior ou igual a 20 anos, 75% tem idade inferior a 25,67, 70% tem idade inferior a 22,8 e 90% tem idade inferior a 33,2 anos.

4.3 MEDIDAS DE DISPERSÃO

De acordo com Toledo (1985), fenômenos que envolvem análises estatísticas caracterizam-se por suas semelhanças e variabilidades. As medidas de dispersão auxiliam as medidas de

tendência central a descrever o conjunto de dados adequadamente. Indicam se os dados estão, ou não, próximos uns dos outros.

Desta forma, não há sentido calcular a média de um conjunto onde não há variação dos seus elementos. Existe ausência de dispersão e a medida de dispersão é igual a zero. Por outro lado, aumentando-se a dispersão, o valor da medida aumenta e se a variação for muito grande, a média não será uma medida de tendência central representativa.

Faz-se necessário, portanto, ao menos uma medida de tendência central e uma medida de dispersão para descrever um conjunto de dados.

As quatro medidas de dispersão que serão definidas a seguir são: amplitude total, amplitude interquartílica, desvio padrão e variância. Com exceção à primeira, todas têm como ponto de referência a média.

4.3.1 Amplitude Total

A amplitude total de um conjunto de dados é a diferença entre o maior e o menor valor observado. A medida de dispersão não levar em consideração os valores intermediários perdendo a informação de como os dados estão distribuídos e/ou concentrados.

$$At = x_{max} - x_{min}$$

Exemplo 16: A amplitude total da idade dos alunos que cursam a disciplina Inferência Estatística do curso de Estatística da UEM é

$$AT = 37\text{-}18 = 19 \text{ anos},$$

isto é, as idades dos alunos diferem em 19 anos.

4.3.2 Amplitude Interquartílica

A amplitude interquartílica é a diferença entre o terceiro e o primeiro quartil. Esta medida é mais estável que a amplitude total por não considerar os valores mais extremos. Esta medida abrange 50% dos dados e é útil para detectar valores discrepantes.

$$d_q = Q_3 - Q_1 .$$

Por outro lado, a amplitude semi-interquartílica é definida como a média aritmética da diferença entre a mediana e os quartis:

$$dq_m = \frac{Q_3 - Q_1}{2} .$$

Exemplo 17: A amplitude interquartílica da idade dos alunos que cursam a disciplina Inferência

Estatística do curso de Estatística da UEM considerando-se a Tabela 10 é:

$$dq = 25,67 - 20 = 5,67 \text{ anos}$$

A amplitude entre o terceiro e primeiro quartil, que envolve 50% (centrais) dos alunos, é de 5,67 anos.

Exemplo 18: Do exemplo 17, obtém-se a amplitude semi-interquartílica da idade dos alunos que cursam a disciplina Inferência Estatística do curso de Estatística da UEM:

$$dq_m = 2,84 \text{ anos.}$$

Observa-se que a distância entre a mediana e o quartil 1 (22-20) é 2. Como 2 < 2,84, isto indica que há uma concentração de dados à esquerda da mediana., e que os dados localizados a direita da mediana são mais dispersos.

4.3.3 Desvio-médio

A diferença entre cada valor observado e a média é denominado desvio e é dado por $(x_i - \mu)$ se o conjunto de dados é populacional, ou por $(x_i - \bar{x})$ se os dados são amostrais.

Ao somar todos os desvios, ou seja, ao somar todas as diferenças de cada valor observado em relação a média, o resultado é igual a zero (propriedade 5 da média). Isto significa que esta medida não mede a variabilidade dos dados. Para resolver este problema, pode-se desconsiderar o sinal da diferença, considerando-as em módulo e a média destas diferenças em módulo é denominada desvio médio:

$$d_m = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N} \qquad \text{ou} \qquad d_m = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n},$$

para dados populacionais ou amostrais, respectivamente. Caso os dados estejam apresentados segundo uma distribuição de freqüência, tem-se:

$$d_m = \frac{\sum_{i=1}^{N} |x_i - \mu| F_i}{N} \qquad \text{ou} \qquad d_m = \frac{\sum_{i=1}^{n} |x_i - \bar{x}| F_i}{n}.$$

4.3.4 Variância e desvio padrão

Enquanto não há nada conceitualmente errado em se considerar o desvio médio, segundo Pagano (2004), esta medida não tem certas propriedades importantes e não é muito utilizada. O mais comum é considerar o quadrado dos desvios em relação à média e então calcular a média. Obtém-se, assim a **variância** que é definida por:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \quad \text{ou} \quad S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1},$$

se os dados são populacionais ou amostrais, respectivamente. Caso os dados estejam apresentados segundo uma distribuição de freqüência, tem-se:

$$\sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \mu)^2 F_i}{N} \quad \text{ou} \quad s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 F_i}{n-1}.$$

Entretanto, ao calcular a variância observa-se que o resultado será dado em unidades quadráticas, o que dificulta a sua interpretação. O problema é resolvido extraindo-se a raiz quadrada da variância, definindo-se, assim, o **desvio padrão**:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \quad \text{ou} \quad S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}},$$

se os dados são populacionais ou amostrais e, se estiverem em distribuição de freqüências:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \mu)^2 F_i}{N}} \quad \text{ou} \quad S = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 F_i}{n-1}}.$$

É importante destacar que se duas populações apresentam a mesma média, mas os desvios padrão não são iguais, isto não significa que as populações têm o mesmo comportamento.

Exemplo 19: Considere três alunos cujas notas em uma disciplina estão apresentadas na Tabela 13. Observa-se que as médias das notas dos três alunos são iguais, porém, seus desvios em torno da média são diferentes. Isto quer dizer que seus desempenhos são diferentes. O aluno A é constante em seu desempenho, o segundo vai progredindo aos poucos e o terceiro diminui abruptamente seu desempenho. Em outras palavras, apesar dos três alunos terem o mesmo desempenho médio, a variabilidade difere.

Tabela 13. Notas, desvios e média dos alunos em uma disciplina.

Aluno	Notas	Soma	Média μ	$d=x_i-\mu$	$\lvert x_i-\mu\rvert$	$(x_i-\mu)^2$	$\sqrt{\sum(x_i-\mu)^2}$
A	8	40	8	0	0	0	$\sqrt{0}=0$
	8			0	0	0	
	8			0	0	0	
	8			0	0	0	
	8			0	0	0	
Total				0	0	0	
B	6	40	8	-2	2	4	$\sqrt{16}=4$
	6			-2	2	4	
	8			0	0	0	
	10			2	2	4	
	10			2	2	4	
Total				0	8	16	
C	10	40	8	2	2	4	$\sqrt{30}=5{,}48$
	10			2	2	4	
	10			2	2	4	
	5			-3	3	9	
	5			-3	3	9	
Total				0	12	30	

Como demonstrado no exemplo, geralmente, o desvio padrão é maior ou igual ao desvio médio, e isto devido ao fato de que para o cálculo do desvio-padrão cada desvio em torno da média é elevado ao quadrado, aumentando desproporcionalmente o peso dos valores extremos.

Exemplo 20: Retomando a idade dos alunos apresentada na Tabela 10, temos:

Desvio médio: $D_m = \dfrac{|20-23{,}8|11+\cdots+|36-23{,}8|2}{22} = 3{,}82$ anos

Variância: $s^2 = \dfrac{(20-23{,}8)^2 11+\cdots+(36-23{,}8)^2 2}{22-1} = 23{,}63$ anos

Desvio padrão: $s = \sqrt{23{,}63} = 4{,}86$ anos.

4.3.5 Coeficiente de Variação

O coeficiente de variação é uma medida de dispersão relativa definida como a razão entre o desvio padrão e a média:

$$CV = \frac{\sigma}{\mu}100 \quad \text{ou} \quad CV = \frac{S}{\overline{X}}100,$$

se os dados são populacionais ou amostrais.

A partir do coeficiente de variação pode-se avaliar a homogeneidade do conjunto de dados e, conseqüentemente, se a média é uma boa medida para representar estes dados. É utilizado, também, para comparar conjuntos com unidades de medidas distintas.

Uma desvantagem do coeficiente de variação é que ele deixa de ser útil quando a média está próxima de zero. Uma média muito próxima de zero pode inflacionar o CV.

Um coeficiente de variação superior a 50% sugere alta dispersão o que indica heterogeneidade dos dados. Quanto maior for este valor, menos representativa será a média. Neste caso, opta-se pela mediana ou moda, não existindo uma regra prática para a escolha de uma destas medidas. O pesquisador, com sua experiência, é que deverá decidir por uma ou outra. Por outro lado, quanto mais próximo de zero, mais homogêneo é o conjunto de dados e mais representativa será sua média.

Exemplo 21: Para idades apresentadas na Tabela 10, temos:

$$CV = \frac{4,86}{23,8}100 = 20,42\% .$$

Como CV<50%, pode-se afirmar que a média é uma medida descritiva representativa para a variável idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá do ano de 2002.

Exercício 09: Calcule as medidas de dispersão para a variável altura da Tabela 10.

4.3.6 Medidas de Assimetria

A medida de assimetria é um indicador da forma da distribuição dos dados. Ao construir uma distribuição de freqüências e/ou um histograma, está-se buscando, também, identificar visualmente, a forma da distribuição dos dados que é ou não confirmada pelo coeficiente de assimetria de Pearson (As) definido como:

$$A_S = \frac{\mu - M_O}{\sigma} \quad \text{ou} \quad A_S = \frac{\bar{X} - M_O}{S}$$

para dados populacionais e amostrais, respectivamente.

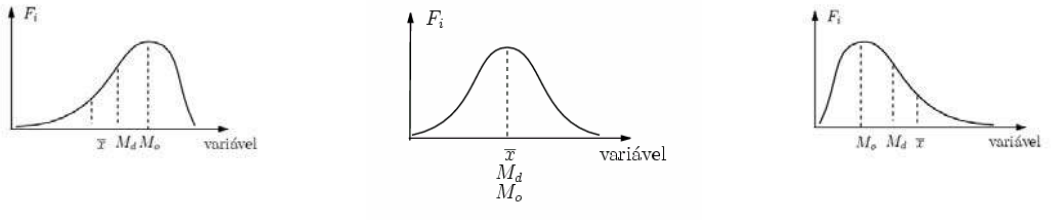
Uma distribuição é classificada como:

simétrica se média = mediana = moda ou As = 0;

assimétrica negativa se média ≤ mediana ≤ moda ou As < 0. O lado mais longo do

polígono de freqüência (cauda da distribuição) está à esquerda do centro.

assimétrica positiva se moda ≤ mediana ≤ média ou As > 0. O lado mais longo do polígono de freqüência está à direita do centro**.**



Assimétrica negativa Simétrica Assimétrica positiva

Figura 15 - Classificação quanto à forma da distribuição

Exemplo 22: A distribuição das idades apresentadas na Tabela 10 é classificada como assimétrica positiva, pois:

$$A_s = \frac{23{,}8 - 20{,}75}{3{,}44} = 0{,}89\,.$$

4.3.7 Medidas de Curtose

A medida de curtose é o grau de achatamento da distribuição, é um indicador da forma desta distribuição. É definido como:

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

A curtose ou achatamento é mais uma medida com a finalidade de complementar a caracterização da dispersão em uma distribuição. Esta medida quantifica a concentração ou dispersão dos valores de um conjunto de dados em relação às medidas de tendência central em uma distribuição de freqüências.

Uma distribuição é classificada quanto ao grau de achatamento como:

Leptocúrtica: quando a distribuição apresenta uma curva de freqüência bastante fechada, com os dados fortemente concentrados em torno de seu centro, K < 0,263.

Mesocúrtica: quando os dados estão razoavelmente concentrados em torno de seu centro, K= 0,263

Platicúrtica: quando a distribuição apresenta uma curva de freqüência mais aberta, com os dados fracamente concentrados em torno de seu centro, K > 0,263.

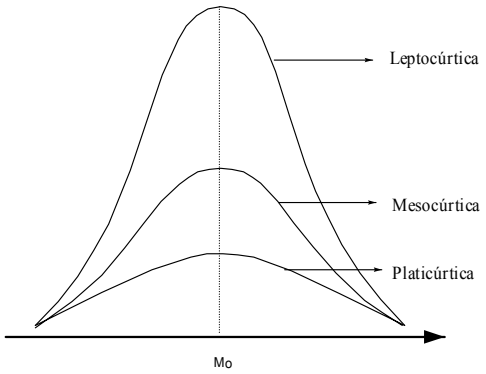


Figura 16 - Classificação da distribuição quanto à curtose.

Exemplo 23: Em relação ao grau de achatamento, a distribuição das idades apresentadas na Tabela 10 é classificada como leptocúrtica, pois:

$$K = \frac{(25,67 - 20)}{2(33,2 - 18,8)} = 0,1969 .$$

Quadro 5 - Resumo descritivo da variável idade (Tabela1)

Centrais	Separatrizes	Dispersão	Assimetria	Curtose
$\overline{x} = 23,8$	Q_1=20	AT=20	A_s=0,89	K=0,1969
$M_o = 20,8$	Q_3=25,67	dq=5,67		
$M_d = 22$	P_{10}=18,8	D_m=23,82		
	P_{90}=33,2	s^2=23,63		
		s=4,86		
		CV=20,42%		

Exercício 10: Determine e interprete as medidas de assimetria e curtose para a variável altura da Tabela 10.

4.4 BOX PLOT OU DESENHO ESQUEMÁTICO

O gráfico Box Plot (ou desenho esquemático) é uma análise gráfica que utiliza cinco medidas estatísticas: valor mínimo, valor máximo, mediana, primeiro e terceiro quartil da variável quantitativa. Este conjunto de medidas oferece a idéia da posição, dispersão, assimetria, caudas e dados discrepantes. A posição central é dada pela mediana e a dispersão pelo desvio interquartílico dq= Q3 - Q2. As posições relativas de Q1 , Q2 e Q3 dão uma noção da assimetria da distribuição. Os comprimentos das caudas são dados pelas linhas que vão do retângulo aos valores atípicos.

Segundo Triola (2004), um outlier ou ponto discrepante é um valor que se localiza distante de quase todos os outros pontos da distribuição. A distância a partir da qual considera-se um valor

como discrepante é aquela que supera 1,5dq. De maneira geral, são considerados outliers todos os valores inferiores Li= Q1-1,5dq ou os superiores a Ls=Q3+1,5dq.

Exemplo 24:A construção do gráfico Box Plot pode ser exemplificada tomando-se a variável idade da Tabela 01. Sua elaboração segue os seguintes passos:

Ordenar os dados em seqüência crescente.

18	18	19	20	20	20	20	20	20	21	21
22	23	24	25	25	25	26	29	30	35	37

Determinar as cinco medidas.

Mediana:

$i = 0,5(22+1) = 11,50$, logo, a mediana é

$M_d = x_{11,50}$

$M_d = x_{11} + 0,50(x_{12} - x_{11})$

$M_d = 21 + 0,50(22 - 21)$

$M_d = 21,50$

Primeiro quartil:

$i = 0,25(22+1) = 5,750$, temos então que o primeiro quartil é

$Q_1 = x_{5,75}$

$Q_1 = x_5 + 0,75(x_6 - x_5)$

$Q_1 = 20 + 0,75(20 - 20)$

$Q_1 = 20$

Terceiro quartil:

$i = 0,75(22+1) = 17,25$, temos então que o terceiro quartil é

$Q_3 = x_{17,25}$

$Q_3 = x_{17} + 0,25(x_{18} - x_{17})$

$Q_3 = 25 + 0,25(26 - 25)$

$Q_3 = 25,75$

Desvio interquartílico:

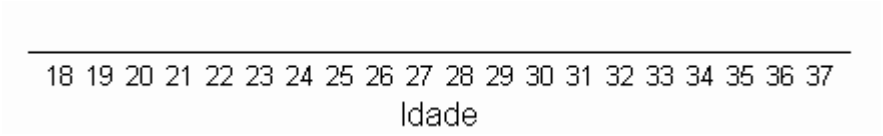
dq= Q3 – Q1 = 25,75 – 20,00 = 5,75

Limite inferior:

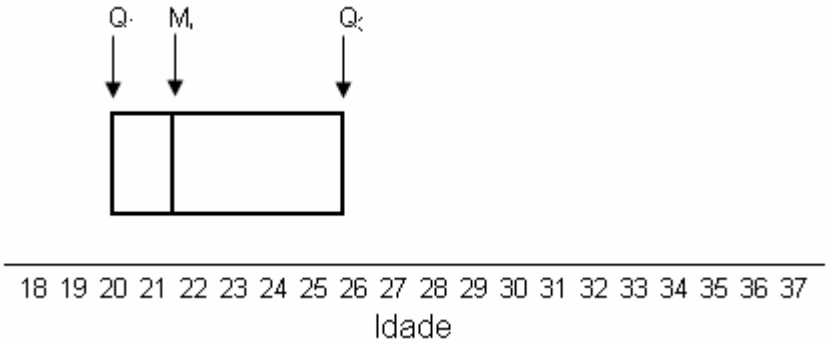
Li= Q1-1,5dq
Li = 20 – 1,5. 5,75 = 11,375

Limite superior:

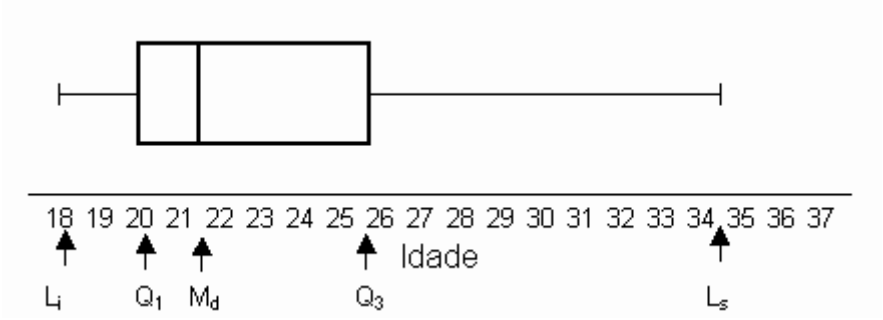
Ls= Q3+1,5dq
Ls= 25,75 + 1,5. 5,75 = 34,375
Construir uma escala com valores que incluam os valores máximo e mínimo dos dados.



Construir uma caixa (retangular) estendendo-se de Q1 a Q3, e trace uma linha na caixa no valor da mediana.



Traçar uma linha paralela à reta, com uma das extremidades alinhada ao limite inferior Li e a outra no centro do lado do retângulo correspondente ao primeiro quartil. Trace uma outra linha paralela à reta, com uma extremidade no centro do lado do retângulo correspondente ao terceiro quartil e a outra alinhada com o limite máximo Ls .



Identificar os pontos discrepantes

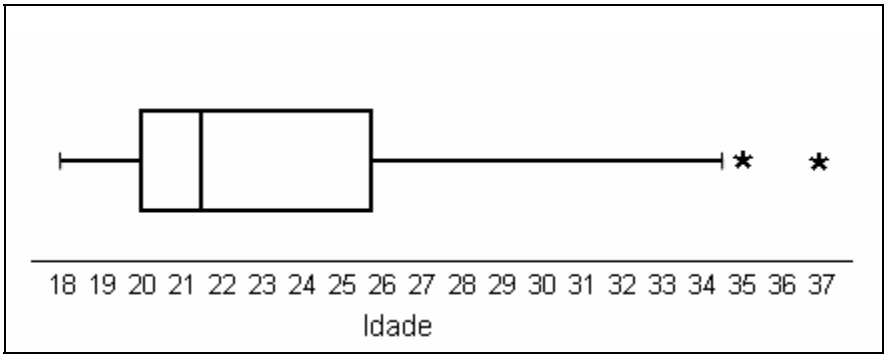


Figura 17 - Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá.

No conjunto de dados não existe aluno com idade inferior a 11,375, ou seja, não há aluno com idade considerada discrepante inferiormente. Entretanto, existem dois indivíduos cujas idades são superiores a 34,375, pontos estes considerados discrepantes neste conjunto de dados: as idades 35 e 37. Estes pontos são identificados no diagrama de caixas por meio de um asterisco na direção das linhas traçadas nos item v.

Note-se que no intervalo interquartílico (dentro do retângulo) existem 50% dos dados, dos quais, 25% estão entre a linha da mediana e a linha do primeiro quartil e os outros 25% estão entre a linha da mediana e a linha do terceiro quartil. Cada linha da cauda mais os valores discrepantes contêm os 25% restantes da distribuição. A Figura 17 mostra que a distribuição das idades dos alunos apresenta assimetria positiva, ou seja, dispersam-se para os valores maiores.

O gráfico Box Plot pode ser utilizado para fazer comparações entre várias distribuições. Essa comparação é feita através de vários desenhos esquemáticos numa mesma figura. Na Figura 18 é apresentado o gráfico para a variável idade classificada segundo o sexo do aluno. Nota-se que para o sexo feminino, não valores discrepantes e a distribuição apresenta assimetria positiva, com idade mediana inferior ao do sexo masculino.

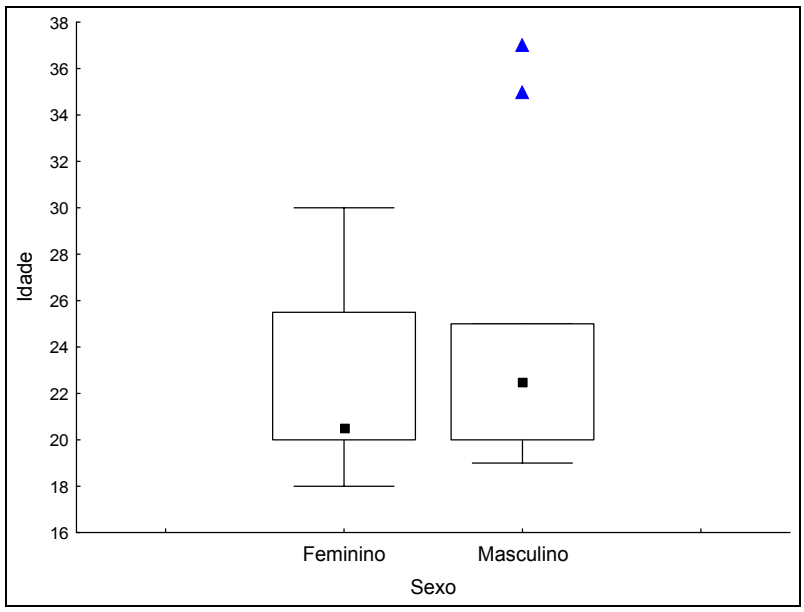


Figura 18 – Box plot da idade segundo o sexo dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá.

Exercício 11: Considere as variáveis peso, nº de reprovas na disciplina Inferência Estatística e nº de irmãos apresentados na Tabela 01. Determine e interprete os resultados, utilizando os dados em rol e em distribuição de frequências:

a) Média, mediana e moda.

b) Quartil 1, quartil 3; decil 4 e percentil 95.

c) Desvio médio, variância, desvio padrão e coeficiente de variação.

d) Medidas de assimetria e curtose.

e) Construir o box plot para cada uma das variáveis.

BIBLIOGRAFIA

BARBETTA, P. A. **Estatística Aplicada às Ciências Sociais**. Florianópolis: Editora da UFSC, 1998.

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

BUSSAB, W. O. e MORETTIN, P. A. **Estatística Básica**. São Paulo: Editora Saraiva, 2003.

CURTY, Marlene G.; CRUZ, Anamaria da C.; MENDES, Maria Tereza R. **Apresentação de trabalhos acadêmicos, dissertações e teses** (NBR 14724/2002). Maringá: Dental Press, 2002.

http://alea-estp.ine.pt/html/statofic/ SILVA, Ana Alexandrino da. Acesso em: 28 abr. 2005; às 21:03.

MAGALHÃES, M. N. e LIMA, A. C. P.de. **Noções de Probabilidade e Estatística**. São Paulo: IME-USP, 2000.

MEDRONHO, R. A., CARVALHO, D.M.de, BLOCH K.V., LUIZ, R.R. E WERNECK, G.L. **Epidemiologia**. São Paulo: Editora Atheneu, 2003.

MILONE, Giuseppe. **Estatística Geral e Aplicada**. São Paulo: Pioneira Thomson Learning, 2004.

MONTGOMERY, D.C. e RUNGER, G.C. **Estatística Aplicada e Probabilidade para Engenheiros.** Rio de Janeiro: Livros Técnicos e Científicos Editora S.A., 2003.

MÜLLER, Mary S.; CORNELSEN, Julce M. **Normas e padrões para teses, dissertações e monografias**. Londrina: Eduel, 2003.

Normas para apresentação de documentos científicos. Vol. 10, Gráficos. Curitiba: Editora da UFPR, 2001.

PAGANO, Marcello ; GAUVREAU, Kimberlee. **Princípios de Bioestatística**. Tradução da 2ª edição norte-americana. São Paulo: Pioneira Thomson Learning, 2004

REIS, Elizabeth. **Estatística descritiva**. Lisboa: Silabo, ed. 4, 1998.

SOARES, José F.; Alfredo A. FARIAS e CESAR, Cibele C. **Introdução à Estatística**. Rio de Janeiro: Livros Técnicos e Científicos Editora S.A., 1991.