CAPÍTULO 2 - Estatística Descritiva

Podemos dividir a Estatística em duas áreas: estatística indutiva (inferência estatística) e estatística descritiva.

Estatística Indutiva: (Inferência Estatística)

Se uma amostra é representativa de uma população, conclusões importantes sobre a população podem ser inferidas de sua análise.

A parte da estatística que trata das condições sob as quais essas inferências são válidas chama-se estatística indutiva ou inferência estatística.

Este assunto iremos tratar apenas no final desse curso. Neste capítulo, estudaremos a outra área da estatística, que é a Estatística Descritiva.

Estatística Descritiva

É a parte da Estatística que procura somente descrever e avaliar um certo grupo, sem tirar quaisquer conclusões ou inferências sobre um grupo maior.

A Estatística Descritiva pode ser resumida nas seguintes etapas:

- Definição do problema:
- Planejamento
- Coleta dos dados
  Crítica dos dados
- Apresentação dos dados
  tabelas
  gráficos
- Descrição dos dados

Nesse capítulo veremos como podem ser feitas tais apresentações (e descrições resumidas) dos dados.

Em estatística descritiva teremos portanto dois métodos que podem ser usados para a apresentação dos dados: métodos **gráficos** (envolvendo apresentação gráfica e/ou tabular) e métodos **numéricos** (envolvendo apresentações de medidas de posição e/ou dispersão).

Apresentação gráfica e tabular.

Os gráficos constituem uma das formas mais eficientes de apresentação de dados. Um gráfico é, essencialmente, uma figura constituida à partir de uma tabela, pois é quase sempre possível locar um dado tabulado num gráfico.

Enquanto as tabelas fornecem uma idéia mais precisa e possibilitam uma inspeção mais rigorosa aos dados, os gráficos são mais indicados em situações que objetivam dar uma visão mais rápida e fácil a respeito das variáveis às quais se referem os dados.

Embora a confecção de gráficos dependa muito da habilidade individual, algumas regras gerais são importantes. O leitor deve ficar atento e procurar saber sobre tais regras antes de se envolver na confecção de gráficos.

Existem vários tipos de gráficos que podem ser utilizados com o objetivo de descrever um conjunto de dados resumidamente. Alguns deles serão aqui exemplificados.

Vejamos, primeiro, uma forma tabular de apresentação de dados e, a seguir, veremos 3 tipos de apresentação gráfica.

- Distribuição de frequência

Organização tabular dos dados em classes de ocorrência, ou não, segundo suas respectivas frequências absolutas. Em alguns casos há também o interesse de se apresentar os dados em frequências relativas ou acumuladas.

A apresentação dos dados em tabelas obedecem a certas normas e recomendações. Essas normas são úteis para que as tabelas sejam feitas de modo que simplicidade, clareza e veracidade perdurem. Diferentes revistas costumam usar pequenas variações na confecção de suas tabelas. Uma observação importante é que as tabelas devem ter significado próprio, ou seja, devem ser entendidas mesmo quando não se lê o texto em que estão apresentadas. O mesmo é válido para as tabelas de distribuição de frequências.

exemplo:

Foram anotados os pontos finais dos alunos de INF 160, referentes ao segundo semestre de 1999. Foi feita a contagem e depois a organização dos dados na seguinte tabela:

Conceitos (Notas)	Número de alunos	Porcentagem
A (90 a 100)	14	7,07
B (75 a 89)	32	16,16
C (60 a 74)	50	25,25
R (< 60)	63	31,82
L [1/]	39	19,70
	198	100,00

FONTE: Departamento de Informática – UFV;
[1/] Reprovação por faltas.

- Diagrama de pontos (dot diagram)

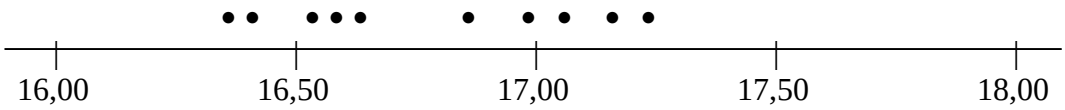
Este tipo de diagrama é muito útil para apresentar um pequeno conjunto de dados (até cerca de 20 observações). Assim podemos ver, de uma maneira rápida e fácil, a tendência central dos dados, além da sua distribuição ou variabilidade.

exemplo:

Considere o seguinte resultado de um experimento no qual o engenheiro testa adição de uma substância em cimento de construção para determinar seu efeito na força da tensão de aderência (em determinada unidade/cm^2):

16,85	16,40	17,21	16,35	16,52	17,04	16,96	17,15	16,59	16,57

Para esse conjunto de dados o diagrama de pontos seria:



Observe que os dados estão centrados num valor próximo de 16,8 e que os valores da tensão de aderência caem no intervalo de cerca de 16,3 até 17,2 ud/cm^2.

Este tipo de diagrama pode também ser usado para se comparar dois ou mais conjuntos de dados. Por exemplo suponha ter sido verificado a tensão de aderência em cimentos não modificados. Os resultados são apresentados abaixo.

17,50	17,63	18,25	18,00	17,86	17,75	18,22	17,90	17,96	18,15

Faça você mesmo o diagrama de pontos para os dois cojuntos de dados, ou seja, colocando ambos os conjuntos de dados no mesmo diagrama. Observe que o diagrama revela imediatamente que o cimento modificado parece ter uma menor força de tensão de aderência, mas que a variabilidade das medidas dentro de ambos os conjuntos de dados parece ser a mesma.

Testes estatísticos para verificar essas duas afirmativas podem ser realizados com esses dados apresentados, e serão discutidos no momento oportuno.

Quando o número de observações é pequeno, geralmente se torna difícil identificar algum padrão específico de variação. No entanto este tipo de diagrama pode ser útil em mostrar alguma característica incomum no conjunto de dados.

- Diagrama de ramos e folhas (stem-and-leaf diagram)

Quando o número de observações é relativamente grande, este diagrama pode ser de boa utilidade.

exemplo:

Barulho é medido em decibéis, representado por dB. Um decibel corresponde ao nível do som mais fraco que pode ser ouvido em um local silencioso por alguém com boa audição. Um sussurro corresponde a cerca de 30 dB; a voz humana em conversação normal corresponde a cerca de 70dB; um rádio em volume alto cerca de 100 dB; Desconforto para os ouvidos geralmente ocorre a cerca de 120 dB. Os dados abaixo correspondem aos níveis de barulho medidos em 36 horários diferentes em um determinado local.

82	89	94	110	74	122	112	95	100	78	65	60
90	83	87	75	114	85	69	94	124	115	107	88
97	74	72	68	83	91	90	102	77	125	108	65

o gráfico de ramos e folhas para o conjunto acima é:

6	0,5,5,8,9
7	2,4,4,5,7,8
8	2,3,3,5,7,8,9
9	0,0,1,4,4,5,7
10	0,2,7,8
11	0,2,4,5
12	2,4,5

- Histograma

Para alguns conjuntos de dados o número de valores distintos da variável em estudo é muito grande para serem considerados os tipos de apresentação gráfica apresentados acima. Em tais casos seria útil dividir os valores em grupos, ou intervalos de classe, e então plotar o número de valores dos dados correspondentes a cada intervalo de classe. Existem várias fórmulas para se estabelecer o número de classes, porém qualquer número de classes poderia ser utilizado, baseando-se nas seguintes observações:

(a) não escolher muito poucas classes, para evitar perda de informação sobre os dados;
(b) não escolher muitas classes, o que poderia fazer com que as frequências referentes a cada classe fossem tão pequenas a ponto de atrapalhar o discernimento de algum padrão de distribuição para a variável em estudo.

O que se faz na prática é tentar variados números de classes e verificar, com a ajuda de um computador, o número ideal para os dados em questão. Além disso, comumente usamos intervalos de classe de iguais amplitudes.

exemplo: (envolvendo distribuição de frequência e histograma, com algumas variações)

Suponhamos que uma empresa deseja avaliar a distribuição dos salários pagos por hora a seus funcionários. O estatístico da empresa possui os seguintes dados:

13,3	15,2	12,4	15,8	9,6	10,4	13,2	8,8	8,3	8,5	10,2
11,5	12,6	10,7	12,6	9,7	12,1	13,5	10,3	14,3	9,8	12,3
10,4	11,6	12,4	12,9	11,6	10,3	14,2	13,8			

Temos ai o que chamamos dados brutos.

Dados como estes poderiam ser agrupados em classes. Uma maneira de escolher o número de classes poderia ser usarmos um valor próximo à raiz quadrada do número de observações. Poderíamos usar, então, 5 classes. Tomando-se a diferença entre o maior e o menor valor do conjunto de dados, e dividindo pelo número de classes escolhido teríamos: (15,8 – 8,3)/5 = 1,5. Esse seria o valor para amplitude da classe, ou intervalo da classe. A seguinte tabela pode ser construida (com intervalo fechado à esquerda):

Classes	frequências
8,3 – 9,8	5
9,8 – 11,3	7
11,3 – 12,8	9
12,8 – 14,3	6
14,3 – 15,8	3
	30

Agora podemos ter uma idéia da distribuição dos salários. Apenas com essas informações poderíamos concluir que a classe de salários predominante na empresa é a terceira, ou seja, com salários de 11,3 a 12,8 salários mínimos.

Se quiséssemos obter maiores informações sobre os dados, poderíamos montar uma nova tabela, incluindo outros tipos de frequência, como: frequência acumulada (f_a), frequência relativa (f_r), e frequência acumulada relativa (f_{ar}).

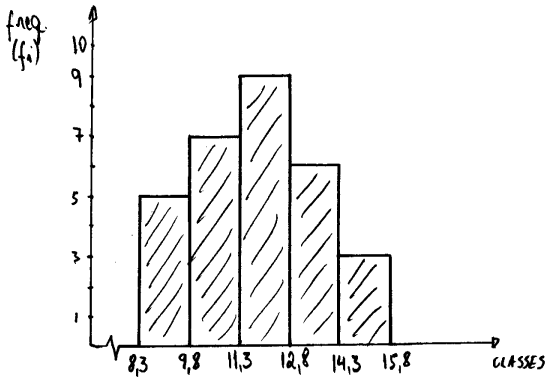
Classes	f_i	f_{ai}	f_{ri}	f_{ari}
8,3 – 9,8	5	5	0,17	0,17
9,8 – 11,3	7	12	0,23	0,40
11,3 – 12,8	9	21	0,30	0,70
12,8 – 14,3	6	27	0,20	0,90
14,3 – 15,8	3	30	0,10	1,00
	30		1,00	

Discussão: exemplos

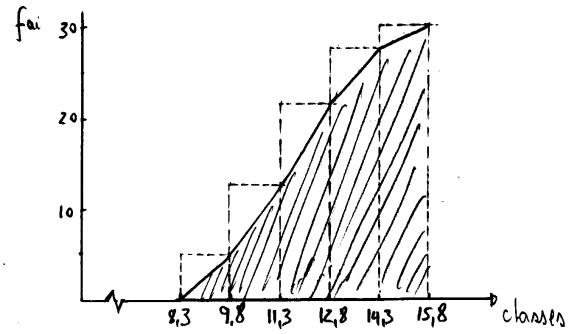
- na terceira coluna, a frequência acumulada 21 indica que , nessa empresa, 21 funcionários recebem salários/hora abaixo de 12,8 unidades;
- Podemos constatar, também, uma certa predominância de salários mais baixos. Realmente cerca de 70% da distribuição de salários concentra-se até o salário de 12,8 unidades;
- Os maiores salários serve a apenas 10% dos funcionários da empresa.;
- 40% dos funcionários (12 funcionários) recebem até 11,3 unidades, sendo 23% (ou seja, 7 funcionários) recebendo entre 9,8 e 11,3 unidades.

Essas informações preliminares, bem como outras, seriam impossíveis de serem obtidas se a população de funcionários fosse muito maior e os dados correspondentes não estivessem tabelados.

O histograma pode ser feito a partir das frequência simples de cada classe ou a partir das frequencias relativas. Bastaria informar corretamente o que seria usado no eixo vertical.



Algumas vezes há o interesse em plotar as frequências acumuladas, ou frequências acumuladas relativas. Nesse caso teríamos a chamada **Ogiva**, ou **ogiva percentual**, respectivamente (veja abaixo).



Medidas de posição e de dispersão.

Nesse tópico serão apresentadas algumas estatísticas úteis para resumir, de modo bastante conciso, as informações contidas em um conjunto do dados. Estatítica, nesse contexto, significa alguma quantidade numérica cujo valor é determinado pelos dados.

Medidas de Posição

Serão apresentadas algumas estatísticas usadas para descrever o centro de um conjunto de dados.

- **Média Aritmética**

Suponha termos um conjunto de n valores numéricos x_1, x_2, …, x_n. A *média aritmética* desses valores será dada por:

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

obs.: o cálculo da média pode ser frequentemente simplificado se observarmos que, para quaisquer contantes a e b

$$y_i = ax_i + b, \qquad i = 1,\ldots n.$$

de modo que a média amostral do novo conjunto de dados será:

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} (ax_i + b)}{n} = \frac{\sum_{i=1}^{n} ax_i + \sum_{i=1}^{n} b}{n} = a\overline{x} + b$$

exemplo:

Considere o seguinte conjunto de dados:

284, 280, 277, 282, 279, 285, 281, 283, 278, 277

encontre a média desses valores.

solução:

uma solução é a seguinte: ao invés de adicionar esses valores diretamente, fica mais fácil se subtraírmos 280 de cada um para obter os novos valores $y_i = x_i - 280$:

4, 0, -3, 2, -1, 5, 1, 3, -2, -3.

A média dos valores transformados será:

$$\overline{y} = 6/10 = 0{,}6.$$

Desse modo,

$$\bar{x} = \bar{y} + 280 = 280{,}6.$$

Algumas vezes queremos determinar a média de um conjunto de dados organizados em uma tabela de distribuição de frequências onde os *k* valores distintos de *X* (x_1, x_2, …, x_k) ocorrem nas respectivas frequências f_1, f_2, …, f_k. Nesse caso a média aritmética será dada por:

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}, \text{ onde } n = \sum_{i=1}^{k} f_i$$

Escrevendo a fórmula anterior como

$$\bar{x} = \frac{f_1}{n} x_1 + \frac{f_2}{n} x_2 + \cdots + \frac{f_k}{n} x_k$$

pode ser observado que a média amostral corresponde à *média ponderada* dos valores distintos de *X* na amostra, onde o peso dado a cada valor x_i nesse caso corresponde à proporção dos *n* valores iguais a x_i , com $i = 1$ *a* k.

exemplo:

a seguinte distribuição de frequência dá as idades de jovens em determinada lanchonete a determinada hora.

Idade	Frequência
15	2
16	5
17	11
18	9
19	14
20	13

encontre a média aritmética da idade dos indivíduos acima.

solução:

$$\bar{x} = (2.15 + 5.16 + 11.17 + 9.18 + 14.19 + 13.20)/54 \cong 18{,}24.$$

OBS.: se a tabela for organizada em classes de valores da variável, para o cálculo da média devemos substituir cada classe pelo seu ponto médio (média aritmética do limite superior e inferior da classe em questão) e calcular a média conforme discutido acima.

- **Mediana amostral**

Outra estatística usada para indicar o centro de um conjunto de dados é a *mediana amostral*, que pode ser definida, de maneira simplificada, como o valor intermediário do conjunto de dados, cujos *n* valores são dispostos em ordem crescente.

Se *n* for ímpar, a mediana será o valor que ocupa a posição $(n + 1)/2$; se *n* for par, a mediana será a média aritmética dos valores ocupando as posições $n/2$ e $n/2 + 1$.

exemplo:

encontre a mediana para os dados apresentados acima.

solução:
já que temos 54 observações, segue que a mediana amostral será a media dos valores ocupando as posições 27 e 28, quando essas 54 observações são organizadas em ordem crescente. Portanto a mediana será o valor 18,5.

OBS.: a escolha entre media e mediana depende do tipo de informação o pesquisador tenta obter dos dados. A media é afetada por valores extremos ocorrendo na distribuição, enquanto a mediana faz uso de apenas um ou dois valores centrais, não sendo, portanto, afetada por valores extremos.

- **Moda amostral**

Outra estatística que tem sido usada para indicar a tendência central de um conjunto de observações é a *moda amostral*. Ela é definida como o valor que ocorre com maior frequência. Podemos ter séries unimodais, bimodais ou multimodais, dependendo do número de valores modais ocorrendo na amostra.

exemplo:
encontre a moda para o mesmo exemplo acima.

solução:
a moda será o valor 19, pois esse valor ocorre com maior frequência na distribuição. Essa é uma distribuição unimodal.

Medidas de Dispersão

Essas medidas são úteis para complementar as informações fornecidas pelas medidas de posição. Descrevem a variabilidade ocorrendo no conjunto de dados sendo analisados.

- **Variância amostral**

A *variância amostral* de um conjunto de dados, x_1, x_2, …, x_n, é definida por

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

onde SQD_x corresponde à *soma de quadrados dos desvios de X*.

exemplo:
encontre a variância amostral para os dois conjuntos de dados abaixo:

A: 3, 4, 6, 7, 10 B: -20, 5, 15, 24

solução:
a média para o conjunto A é 6; portanto a variância será:

$$s^2 = [(-3)^2 + (-2)^2 + (0)^2 + 1^2 + 4^2]/4 = 7,5$$

a média para o conjunto B também é 6; portanto a variância de B será:

$$s^2 = [(-26)^2 + (-1)^2 + 9^2 + (18)^2]/3 \cong 360,67$$

Portanto, apesar dos dois conjuntos terem a mesma média, há maior variabilidade nos valores do conjunto B do que nos do conjunto A.

Para o cálculo da variância útil se faz a seguinte identidade algébrica:

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

Também, o cálculo da variância pode ser simplificado por notar que se:

$$y_i = ax_i + b, \qquad i = 1, \ldots, n$$

então, como visto atrás, $\bar{y} = a\bar{x} + b$ e, então

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=a^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$

ou seja, adicionando uma constante a cada valor do conjunto de dados não altera a variância amostral; enquanto multiplicando-se cada valor por uma constante, a nova variância amostral será igual a variância original multiplicada pelo quadrado da constante.

exemplo:

O conjunto de dados abaixo fornece o número mundial de acidentes aéreos fatais de aeronaves comerciais nos anos de 1985 a 1993.

Ano	1985	1986	1987	1988	1989	1990	1991	1992	1993
Acidentes	22	22	26	28	27	25	30	29	24

encontre a variância amostral do número de acidentes nesses anos.

solução:

considere o seguinte conjunto de dados resultante da subtração de 22 de cada valor original:

0, 0, 4, 6, 5, 3, 8, 7, 2

chamando esses valores de $y_1, y_2, \ldots, y_9$, teremos

$$\sum_{i=1}^{9}y_i=35, \qquad \sum_{i=1}^{9}y^2=203.$$

Portanto, já que a variância dos dados transformados corresponde exatamente à variância dos dados originais, usando-se a identidade algébrica acima teremos:

$$s^2=\frac{203-9(35/9)^2}{8}\cong 8{,}361$$

OBS.: se a cada valor de X tivermos associado sua frequência de ocorrência, então

$$s^2=\frac{\sum_i f_i(x_i-\bar{x})^2}{\sum_i f_i-1}=\frac{\sum f_i x_i^2-\frac{(\sum_i f_i x_i)^2}{\sum_i f_i}}{\sum_i f_i-1}$$

- **Desvio padrão amostral**

A raiz quadrada positiva da variância amostral é chamada de *desvio padrão amostral*, ou seja,

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=11}^{n} (x_i - \bar{x})^2}{n-1}}$$

Existem outras medidas também úteis para representar a dispersão dos dados. Poderíamos citar: Amplitude Total, Erro padrão da média, Coeficiente de variação.

- **Amplitude total**

A *amplitude total* é a diferença entre o maior e o menor valor da série. Tem a vantagem de ser rápido e fácil de ser calculada, porém fornece um número índice grosseiro da variabilidade de uma distribuição, por levar em conta apenas 2 valores de um conjunto.

- **Erro-padrão da média**

O *erro-padrão da média* mede a precisão da média. Sua fórmula é dada por:

$$s(\overline{X}) = \sqrt{V(\overline{X})} = \sqrt{\frac{s_X^2}{n}} = \frac{s_X}{\sqrt{n}}$$

- **Coeficiente de Variação**

O *coeficiente de variação* é uma medida de dispersão relativa. É uma medida útil para comparação, em termos relativos, do grau de concentração, em torno da média, de séries distintas. Por ser um número admensional permite a comparação de séries de variáveis com unidades diferentes. Sua fórmula é dada por:

$$C.V.\ (\%) = \frac{s_{(X)}}{\overline{X}} \cdot 100$$

OBS.: se existem duas amostras distintas A e B, e se desejamos saber qual delas é a mais homogênea, ou seja, de menor variabilidade, basta fazermos o seguinte: calculamos as médias e os desvios padrões de A e B, e:

- se $\overline{X}_A = \overline{X}_B$, então o próprio desvio padrão informará qual é a mais homogênea.
- se $\overline{X}_A \neq \overline{X}_B$, então a mais homogênea será a que tiver menor C.V.

OBS.: valores muito altos de C.V. indicam pequena representatividade da média.

exemplo:

Supor duas amostras:

A={1, 3, 5}

B={53, 55, 57}

Qual das duas é a mais homogênea?

solução:

$C.V._A = 2/3(100) = 66,7\%$

$C.V._B = 2/55(100) = 3,6\%$

Portanto a amostra B é a mais homogênea.

Exercícios Propostos

1) Considerando os dados amostrais abaixo, calcular: média aritmética, variância, desvio padrão, erro padrão da média e coeficiente de variação
   Dados: 2, 3, 5, 1, 2, 1, 4, 3, 3, 4, 3.
   R.:2,81; 1,56; 1,24; 0,37; 44,12%

2) Em certa região a temperatura média é 20 ^{0}C e a precipitação média é 700 mm. O desvio padrão para temperatura é 3 ^{0}C, enquanto que a variância para a precipitação é 1225 mm^2. Qual dos dois fenômenos apresenta maior variabilidade? Justifique.
   R.: a temperatura apresenta maior variabilidade relativa. Você justifica…

3) Um artigo retirado da revista *Technometrics* (Vol. 19, 1977, p. 425) apresenta os seguintes dados sobre a taxa de octanagem de várias misturas de gasolina:

88,5	87,7	83,4	86,7	87,5	91,5	88,6	100,3	96,5	93,3	94,7
91,1	91,0	94,2	87,8	89,9	88,3	87,6	84,3	86,7	84,3	86,7
88,2	90,8	88,3	98,8	94,2	92,7	93,2	91,0	90,1	93,4	88,5
90,1	89,2	88,3	85,3	87,9	88,6	90,9	89,0	96,1	93,3	91,8
92,3	90,4	90,1	93,0	88,7	89,9	89,8	89,6	87,4	88,4	88,9
91,2	89,3	94,4	92,7	91,8	91,6	90,4	91,1	92,6	89,8	90,6
91,1	90,4	89,3	89,7	90,3	91,6	90,5	93,7	92,7	92,2	92,2
91,2	91,0	92,2	90,0	90,7						

   (a) Construa o diagrama de folhas-e-ramos para esses dados
   (b) Construa a distribuição de frequência e o histograma. Use 8 intervalos de classe.
   (c) Construa a distribuição de frequência e o histograma, agora com 16 intervalos de classe.
   (d) Compare a forma dos dois histogramas em b e c. Ambos os histogramas mostram informações similares?

4) O seguinte conjunto de dados representa as “vidas” de 40 baterias de carro da mesma marca e mesmas características com aproximação até décimos do ano. As baterias tinham garantia para 3 anos.

2,2	4,1	3,5	4,5	3,2	3,7	3,0	2,6	3,4	1,6	3,1
3,3	3,8	3,1	4,7	3,7	2,5	4,3	3,4	3,6	2,9	3,3
3,9	3,1	3,3	3,1	3,7	4,4	3,2	4,1	1,9	3,4	4,7
3,8	3,2	2,6	3,9	3,0	4,2	3,5				

   (a) Construa a distribuição de frequência e o histograma;
   (b) Faça o gráfico da distribuição de frequências relativas acumuladas.
   (c) Calcule a média aritmética dos dados originais

(d) Usando a distribuição de frequência conforme obtido em a calcule a média novamente. Para tal, considere os pontos médios de cada classe (média entre os dois limites de cada classe) para serem os valores da variável no cálculo da média.
(e) Obtenha a variância para os dados originais conforme feito para a média em c.
(f) Obtenha a variância a partir da distribuição de frequência conforme feito para a média no ítem d.

obs.: use 7 intervalos de classe. Amplitude da classe igual a 0,5. E o início do intervalo mais baixo em 1,5.

5) Mostre que $\sum_i f_i (x_i - \bar{x})^2 = \sum f_i x_i^2 - \frac{(\sum_i f_i x_i)^2}{\sum_i f_i}$

6) Mostre que a soma de quadrados dos desvios (SQD) em relação à média é um mínimo. Dica: Considere *f(a)* a função que representa a SQD em relação a *a*. Ou seja, $f(a) = \sum_{i=1}^{n} (x_i - a)^2$. Usando seus conhecimentos de cálculo, mostre que *f(a)* será mínimo quando *a* for igual a média dos valores de X.

7) Calcule a média, mediana, e amplitude total dos valores dispostos no seguinte diagrama de ramos e folhas

6	0 5 5 8 9
7	2 4 4 5 7 8
8	2 3 3 5 7 8 9
9	0 0 1 4 4 5 7
10	0 2 7 8
11	0 2 4 5
12	2 4 5

UNIVERSIDADE FEDERAL DE VIÇOSA
--Departamento de Informática / CCE
INF 161 - Iniciação à Estatística / INF 162 – Estatística I
Lista de Exercícios: Estatística Descritiva

1) Os dados abaixo se referem a medidas tomadas em uma amostra de 10 cães:

Cão	1	2	3	4	5	6	7	8	9	10
Peso (kg)	23,0	22,7	21,2	21,5	17,0	28,4	19,0	14,5	19,0	19,5
Comprimento (cm)	104	105	103	105	100	104	100	91	102	99

Pede-se, para as características avaliadas, peso e comprimento, as estatísticas:

a) Média;

b) Variância;

c) Desvio-padrão;

d) Erro-padrão da média;

e) Coeficiente de variação;

f) Qual das duas características é a mais homogênea;

g) Mediana;

h) Moda.

2) Um pesquisador dispõe das seguintes informações, a respeito dos valores de uma amostra:

- a média de todos os valores é igual a 50,34;

- a soma dos quadrados dos valores é igual a 150.000;

- a amostra é constituída de 52 valores distintos.

Pergunta-se:

Com essas informações é possível obter alguma(s) medida(s) de dispersão dos valores amostrais? Em caso afirmativo, efetue os cálculos e obtenha a(s) respectiva(s) medida(s).

3) Considere os dados: 12, 17, 17, 17, 10, 10, 9, 9, 9, 12, 12, 6, 6, 6, 17, 17, 12, 12, 9, 9, 9, 12, 12, 12, 12. Supondo que sejam valores assumidos por uma variável aleatória discreta X, pede-se:

a) Média, mediana e moda;

b) Erro-padrão da média e C.V.(%).

4) Duas turmas A e B com $n_A = 50$ e $n_B = 80$ apresentaram médias $\overline{X}_A = 65$ e $\overline{X}_B = 70$ e variâncias $s_A^2 = 225$ e $s_B^2 = 235$. Qual é a turma mais homogênea?

5) A média de aprovação na disciplina de Estatística é 6 ou mais. Durante um período letivo foram realizadas quatro provas, sendo que a primeira prova teve peso dois, a segunda e a terceira o dobro do peso da primeira e a última igual ao peso da primeira. Os resultados, incluindo os de uma prova de substituição optativa, foram os seguintes:

Estudantes	1ª	2ª	3ª	4ª	Optativa
1	2,5	4,5	5,0	6,0	7,0
2	2,0	8,5	7,0	3,0	5,0
3	8,5	10,0	9,0	8,5	nc
4	3,5	5,5	8,5	7,5	6,5
5	3,0	5,0	6,0	4,5	5,0
6	6,0	3,0	4,0	5,0	2,0
7	8,0	1,5	2,0	9,0	5,0
8	1,5	2,0	1,0	2,5	nc
9	7,5	8,0	8,5	10,0	nc
10	5,5	4,5	5,0	4,5	2,5

Sabendo-se que a nota da prova optativa substitui a menor nota das provas precedentes, determine:

a) Média de cada estudante;

b) Para cada prova: média, moda, mediana, variância, desvio-padrão, erro-padrão da média e CV.

c) Para o período: média, variância, desvio-padrão, erro-padrão da média, CV.

d) Liste as provas em ordem crescente de homogeneidade.

RESPOSTAS

1.*a*) $\overline{X} = 20{,}58kg; \overline{Y} = 101{,}3cm$

b) $\hat{V}(X) = 14{,}2973kg^2; \hat{V}(Y) = 17{,}7889cm^2$

c) $s(X) = 3{,}7812kg; s(Y) = 4{,}2177cm$

d) $s(\overline{X}) = 1{,}1957kg; s(\overline{Y}) = 1{,}3338cm$

e) $CV_X = 18{,}37\%; CV_Y = 4{,}16\%$

f) Comprimento, pois é a que possui menor CV.

g) $Md_X = 20{,}35kg; Md_Y = 102{,}50cm$

h) $Mo_X = 19{,}0kg;\ Mo_Y = 100\,cm, 104cm\ e\ 105cm$

2. s^2 = 357,3723; s = 18,9043; CV = 37,55%; $s(\overline{X})$ = 2,6215

3. a) $\overline{X}$ = 11,4; Md = 12; Mo = 12 b) $s(\overline{X}) = 0{,}6904$; CV = 30,28%

4. Turma B

5. a)

Estudante	1	2	3	4	5	6	7	8	9	10
Média	5,33	6,50	9,17	7,00	5,25	3,83	5,17	1,67	8,42	4,50

b)

Arguição	1ª	2ª	3ª	4ª
$\overline{X}$	6,05	5,50	5,60	5,85
Mo	5	2; 4,5; 5	5; 8,5	2,5
Md	6,25	5,0	5,5	5,5
s^2	4,02	6,94	7,54	7,78
s	2,01	2,64	2,75	2,79
$s(\overline{X})$	0,63	0,83	0,87	0,88
CV(%)	33,16%	47,91%	49,05%	47,68%

c) $\overline{X} = 5{,}6833$; $s^2 = 6{,}2098$; $s = 4919$; $s(\overline{X}) = 0{,}2275$; CV = 43,85%

d) 3ª, 2ª, 4ª, 1ª