oi

Unidade II – Estatística descritiva

2.1. Apresentação de dados
  2.1.1 Séries estatísticas
  2.1.2 Tabelas
  2.1.3 Gráficos
2.2. Distribuições de freqüências e gráficos
  2.2.1 Tabelas de classificação simples
  2.2.2 Tabelas de classificação cruzada
2.3. Medidas descritivas
  2.3.1 Medidas de localização ou tendência central
  2.3.2 Medidas separatrizes
  2.3.3 Medidas de variação ou dispersão
  2.3.4 Medidas de formato
2.4. Análise exploratória de dados

Distribuição de freqüências e gráficos

- A distribuição de frequências é uma forma de resumir a informação sobre uma ou mais variáveis
- Organiza um conjunto de dados em classes, indicando a freqüência de observações em cada classe
- Além de resumir a informação, tem por finalidade:
  1. Representar a forma como se distribuem os valores das variáveis (localizaçao da maioria dos valores, simetria, número de picos e formato das caudas)
  2. Indicar qual modelo de distribuição de probabilidade poderia ser adequado para esses dados, pois fornece uma idéia empírica da distribuição da população
- Formato é muito sensível ao número de observações disponíveis

oi

Tabelas de distribuição de freqüências

- Tabelas de classificação simples
  → **uma variável**
- Tabelas de classificação dupla ou cruzada
  → **duas variáveis**

Tabelas de classificação simples

As características dessas tabelas variam de acordo com o tipo de variável em estudo.

⇨ Se a variável é do tipo categórica, devemos obter as freqüências para cada nível dessa variável.

⇨ Se a variável é do tipo numérica contínua, devemos primeiro construir intervalos de mesma amplitude e depois obter as freqüências para cada intervalo.

oi

Distribuição de freqüências para variáveis categóricas

Exemplo:

Variável em estudo: conceito na disciplina de Estatística

Dados brutos: ruim, médio, bom, médio, ruim, médio, ruim, médio, ruim, bom, médio, médio, bom, médio, médio, médio, ótimo, médio, bom, ótimo, bom, ótimo, médio, ótimo, médio, ruim, médio, ótimo, médio, médio, bom, ruim, bom, bom, médio, ruim, médio, médio, ótimo, médio, bom, ruim, ruim, bom, médio, médio, ruim, bom, médio, médio, bom, bom, bom, médio, ruim, bom, médio, médio, ruim, médio

⇨ Quando a variável for categórica ou numérica discreta (com poucos valores), a tabela de distribuição de freqüências apresentará a seguinte característica: cada valor da variável constituirá uma classe.

Construção da tabela

Para construir a tabela devemos seguir apenas dois passos:

1º passo. Ordenar as categorias ou valores da variável (colocar em rol). Cada categoria ou valor constituirá uma classe.

- O número da classe é representado por j, tal que j=1, 2, ..., k, onde k é o número total de classes.

2º passo. Contar o número de elementos em cada classe, ou seja, contar quantas vezes o dado está repetido.

oi

Exemplo 1. Variável categórica

Variável em estudo: conceito na disciplina de Estatística

Dados brutos: ruim, médio, bom, médio, ruim, médio, ruim, médio, ruim, bom, médio, médio, bom, médio, médio, médio, ótimo, médio, bom, ótimo, bom, ótimo, médio, ótimo, médio, ruim, médio, ótimo, médio, médio, bom, ruim, bom, bom, médio, ruim, médio, médio, ótimo, médio, bom, ruim, ruim, bom, médio, médio, ruim, bom, médio, médio, bom, bom, bom, médio, ruim, bom, médio, médio, ruim, médio

1º passo. Ordenar os níveis da variável

Número da classe (j)	Classe
1	Ruim
2	Médio
3	Bom
4	Ótimo

2º passo. Contar o número de elementos em cada classe

j	Classe	F_j
1	Ruim	12
2	Médio	27
3	Bom	15
4	Ótimo	6
	Σ	60

Os valores provenientes desta contagem, denotados por F_j, são denominados frequências absolutas das classes.

oi

Outras frequências importantes:

Frequência absoluta acumulada, denotada por F'_j, expressa o número de elementos acumulados em cada classe.

j	Classe	F_j	F'_j
1	Ruim	12	12
2	Médio	27	39
3	Bom	15	54
4	Ótimo	6	60
	Σ	60	-

Outras frequências importantes:

Frequência relativa, denotada por f_j, expressa a proporção de elementos em cada classe.

j	Classe	F_j	F'_j	f_j
1	Ruim	12	12	0,2
2	Médio	27	39	0,45
3	Bom	15	54	0,25
4	Ótimo	6	60	0,1
	Σ	60	-	1

oi

Outras frequências importantes:

Frequência relativa acumulada, denotada por f'_j, expressa a proporção de elementos acumulada em cada classe.

j	Classe	F_j	F'_j	f_j	f'_j
1	Ruim	12	12	0,2	0,2
2	Médio	27	39	0,45	0,65
3	Bom	15	54	0,25	0,90
4	Ótimo	6	60	0,1	1
	Σ	60	-	1	-

Frequências importantes:

F_j: frequência absoluta da classe j → número de elementos na classe j

F'_j: frequência absoluta acumulada da classe j → número de elementos acumulados na classe j

f_j: frequência relativa da classe j → proporção de elementos na classe j

f'_j: frequência relativa acumulada da classe j → proporção de elementos acumulados na classe j

oi

Interpretação considerando o contexto (contextualizando):

proporção de alunos que obtiveram até conceito Médio

número de alunos que obtiveram até conceito Bom

j	Classe	F_j	F'_j	f_j	f'_j
1	Ruim	12	12	0,2	0,2
2	Médio	27	39	0,45	0,65
3	Bom	15	54	0,25	0,90
4	Ótimo	6	60	0,1	1
	Σ	60	-	1	-

proporção de alunos que obtiveram conceito Ruim

número de alunos que obtiveram conceito Ótimo

Profa. Clause Piana 13

Exemplo 2. Variável numérica discreta

Pesquisa: Monitoramento de um canal de comunicação

Variável em estudo: número de erros em um conjunto de caracteres (*string*) de 1.000 *bits*. Foram avaliados 350 conjuntos.

Dados brutos: 2, 5, 6, 0, 4, 4, 3, 4, 2, 2, 3, 3, 5, 3, 5, 1, 2, 4, 2, 3, 5, 4, 3, 3, 2, 3, 0, 4, 4, 3, 4, 0, 2, 0, 2, 3, 3, 1, 2, 4, 2, ...

1º passo: Ordenar os valores da variável

Número da classe (j)	Classe
1	0
2	1
3	2
4	3
5	4
6	5
7	6

oi

Exemplo 2. Variável numérica discreta

Pesquisa: Monitoramento de um canal de comunicação

Variável em estudo: número de erros em um conjunto de caracteres (*string*) de 1.000 *bits*. Foram avaliados 350 conjuntos.

Dados brutos: 2, 5, 6, 0, 4, 4, 3, 4, 2, 2, 3, 3, 5, 3, 5, 1, 2, 4, 2, 3, 5, 4, 3, 3, 2, 3, 0, 4, 4, 3, 4, 0, 2, 0, 2, 3, 3, 1, 2, 4, 2, ...

2º passo. Contar o número de elementos em cada classe.

j	Classe	F_j
1	0	55
2	1	60
3	2	112
4	3	82
5	4	31
6	5	8
7	6	2
	Σ	350

j	Classe	F_j	F'_j	f_j	f'_j
1	0	55	55	0,1571	0,1571
2	1	60	115	0,1714	0,3286
3	2	112	227	0,32	0,6486
4	3	82	309	0,2343	0,8829
5	4	31	340	0,0886	0,9714
6	5	8	348	0,0229	0,9943
7	6	2	350	0,0057	1,0000
	Σ	350	-	1,0000	-

oi

Exercício proposto:

Os dados a seguir se referem ao número de ovos danificados em uma inspeção feita em 30 embalagens de uma dúzia cada, em um carregamento para o mercado de Lavras.

1	0	1	0	0	1	0	2	0	3
1	0	5	3	1	0	1	4	0	0
2	0	0	1	2	0	1	3	1	0

Construa a distribuição de freqüências para esses dados.

Resolução:

j	Classe	F_j	F'_j	f_j	f'_j
1	0	13	13	0,4333	0,4333
2	1	9	22	0,3	0,7333
3	2	3	25	0,1	0,8333
4	3	3	28	0,1	0,9333
5	4	1	29	0,03333	0,9667
6	5	1	30	0,03333	1,000
	Σ	30	-	1,000	-

oi

Distribuição de freqüências para variáveis contínuas

Exemplo:

Variável em estudo: valores gastos (em reais) pelas primeiras 50 pessoas que entraram num determinado Supermercado, no dia 01/01/2000.

Dados brutos:

32,03	19,54	45,40	25,13	46,69	18,36	13,78	15,23	36,37	15,62
17,00	27,65	85,76	38,64	86,37	24,58	20,16	93,34	48,65	22,22
23,04	42,97	28,06	52,75	3,11	8,88	9,26	10,81	12,69	28,38
18,43	61,22	41,02	44,67	19,50	17,39	39,16	44,08	38,98	19,27
26,24	28,08	59,07	82,70	26,26	24,47	54,80	70,32	50,39	20,59

As variáveis contínuas, em geral, assumem muitos valores diferentes uns dos outros.

⇨ Assim, as tabelas de distribuição de freqüências são construídas de modo que cada classe seja constituída por um intervalo de valores da variável.

⇨ Quando variáveis discretas assumem muitos valores diferentes é usual agrupar os dados discretos em intervalos de classe.

oi

Construção da tabela

1º passo. Ordenar o conjunto de dados: colocar os dados brutos em ordem crescente de grandeza (rol).

Dados ordenados:

3,11	8,88	9,26	10,81	12,69	13,78	15,23	15,62	17,00	17,39
18,36	18,43	19,27	19,50	19,54	20,16	20,59	22,22	23,04	24,47
24,58	25,13	26,24	26,26	27,65	28,06	28,08	28,38	32,03	36,37
38,64	38,98	39,16	41,02	42,97	44,08	44,67	45,40	46,69	48,65
50,39	52,75	54,80	59,07	61,22	70,32	82,70	85,76	86,37	93,34

Construção da tabela

1º passo. Ordenar o conjunto de dados: colocar os dados brutos em ordem crescente de grandeza (rol).

2º passo. Determinar o número de classes (k) da tabela.

De modo geral, esse valor não deverá ser inferior a 5 e nem superior a 15.
A perda de informação é inevitável.

- k muito grande ⇨ maior precisão e menor eficiência no resumo.
- k muito pequeno ⇨ resume demais e a precisão fica prejudicada.

oi

Construção da tabela

1º passo. Ordenar o conjunto de dados: colocar os dados brutos em ordem crescente de grandeza (rol).

2º passo. Determinar o número de classes da tabela.

- Regras para determinação do número de classes:

Arredondar para cima

$k = \sqrt{n}$ ← Regra empírica

$k = 1 + 3{,}32 \times \log n$ ← Fórmula de Sturges ($30 \leq n \leq 40$)

onde: k: número de classes
n: número de observações

Fórmula de Sturges

- É a mais antiga em uso.
- Pode funcionar de modo razoável apenas se $30 \leq n \leq 40$.
- Quando n é grande, tende a produzir um número pequeno de classes.
- Apresenta uma propriedade que, além de prever simetria na distribuição, indica que esse processo está mais preocupado com a estética da apresentação do que com a informação representada.
- Além disso, desconsidera totalmente a existência de valores atípicos na determinação.

Outros métodos para determinar o número de classes:

- Fórmula de Scott
- Fórmula combinada de Terrel e Scott
- Método de Shimazaki e Shinomoto
- Método de Freedman-Diaconis

Ver Sistema Galileu. Disponível em http://www.galileu.esalq.usp.br

oi

3º passo. Determinar a amplitude do intervalo de classe.

Para isso utilizamos a expressão $i = \frac{a_t}{k}$ ← Arredondar para cima

onde: i: amplitude do intervalo
a_t: amplitude total = ES – EI
- $X_{(1)}$ = Extremo Inferior
- $X_{(n)}$ = Extremo Superior

4º passo. Construir os intervalos de classe.

j	Classe
1	$x_{(1)} \vdash\!\!- x_{(1)} + i$
2	$x_{(1)} + i \vdash\!\!- x_{(1)} + 2i$
3	$x_{(1)} + 2i \vdash\!\!- x_{(1)} + 3i$
. . .	. . .
k	$x_{(1)} + (k-1)i \vdash\!\!- x_{(1)} + ki$

5º passo. Contar o número de observações em cada classe.

Na construção dos intervalos de classe, é importante observar que:

- ⇨ Recomenda-se o uso de intervalos de mesma amplitude, mas eventualmente uma amplitude variável poderá ser mais adequada ao contexto;
- ⇨ Deve ser garantido que todas as observações sejam classificadas;
- ⇨ As classes são mutuamente exclusivas, ou seja, uma observação pertence a uma única classe;
- ⇨ Com exceção da última classe, que é fechada à esquerda e à direita, os intervalos são fechados à esquerda e abertos à direita, de modo que um valor que coincida com o extremo superior será classificado na classe seguinte.

oi

Exemplo:

Os dados em rol abaixo (ordenação horizontal) se referem aos valores gastos (em reais) pelas primeiras 50 pessoas que entraram em um determinado Supermercado, no dia 01/01/2000.

3,11	8,88	9,26	10,81	12,69	13,78	15,23	15,62	17,00	17,39
18,36	18,43	19,27	19,50	19,54	20,16	20,59	22,22	23,04	24,47
24,58	25,13	26,24	26,26	27,65	28,06	28,08	28,38	32,03	36,37
38,64	38,98	39,16	41,02	42,97	44,08	44,67	45,40	46,69	48,65
50,39	52,75	54,80	59,07	61,22	70,32	82,70	85,76	86,37	93,34

Faça a distribuição de freqüências desses dados.

Resolução: Usando a fórmula empírica

$n = 50$

$k = \sqrt{n} = \sqrt{50} = 7{,}07 \cong 8$

$i = \frac{a_t}{k} = \frac{ES - EI}{k} = \frac{93{,}34 - 3{,}11}{8} = 11{,}28$

Ponto médio ou centro de classe

j	Classe	F_j	F'_j	f_j	f'_j	c_j
1	3,11 |— 14,39	6	6	0,12	0,12	8,75
2	14,39 |— 25,67	16	22	0,32	0,44	20,03
3	25,67 |— 36,95	8	30	0,16	0,60	31,31
4	36,95 |— 48,23	9	39	0,18	0,78	42,59
5	48,23 |— 59,51	5	44	0,10	0,88	53,87
6	59,51 |— 70,79	2	46	0,04	0,92	65,15
7	70,79 |— 82,07	0	46	0,00	0,92	76,43
8	82,07 |—| 93,35	4	50	0,08	1	87,71
	Σ	50	-	1	-	-

oi

Resolução: Usando a fórmula de Sturges

$n = 50$

$k = 1 + 3{,}32 \times \log n = 1 + 3{,}32 \times 1{,}7 = 6{,}64 \cong 7$

$$i = \frac{a_t}{k} = \frac{ES - EI}{k} = \frac{93{,}34 - 3{,}11}{7} = 12{,}89$$

j	Classe	F_j	F'_j	f_j	f'_j	c_j
1	3,11|— 16,00	8	8	0,16	0,16	9,56
2	16,00|— 28,89	20	28	0,4	0,56	22,45
3	28,89|— 41,78	6	34	0,12	0,68	35,34
4	41,78|— 54,67	8	42	0,16	0,84	48,23
5	54,67|— 67,56	3	45	0,06	0,9	61,12
6	67,56|— 80,45	1	46	0,02	0,92	74,01
7	80,45 |—| 93,34	4	50	0,08	1	86,90
	Σ	50	-	1	-	-

Exercício proposto:

Os dados se referem às notas dos alunos dos curso de Ciência e Engenharia da Computação da UFPel na primeira prova de Estatística Básica, no segundo semestre de 2013.

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

Faça a distribuição de freqüências desses dados.

oi

Representação gráfica

As distribuições de frequências podem ser representadas graficamente de duas formas distintas e exclusivas:

- **Histograma**
- **Polígono de frequências**

Histograma

Variável contínua (dados de mensuração)

O histograma consiste de um conjunto de retângulos contíguos cuja base é igual à amplitude do intervalo e a altura proporcional à freqüência das respectivas classes.

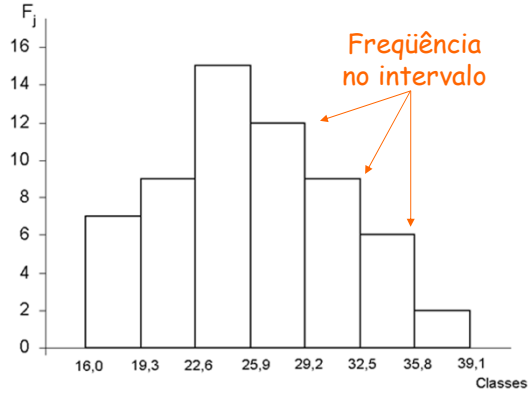


Figura 1. Frequência do peso ao nascer de 60 bovinos da raça Ibagé. UFPel, 2001.

oi

Outra maneira de construir o histograma

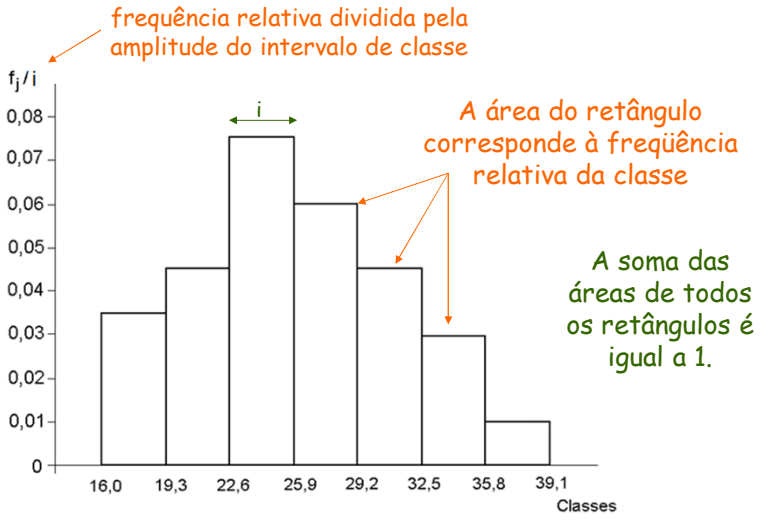


Figura 3. Histograma das freqüências relativas para o peso ao nascer de 60 bovinos da raça Ibagé. UFPel, 2001.

Histograma

Variável discreta (dados de enumeração)

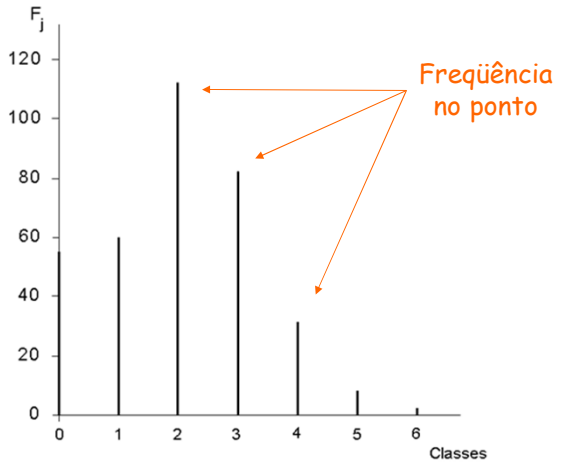


Figura 2. Frequência do número de erros em 350 conjuntos de caracteres (*strings*) de 1.000 *bits*.
Fonte: Dados fictícios.

oi

Variável categórica

Gráfico de colunas

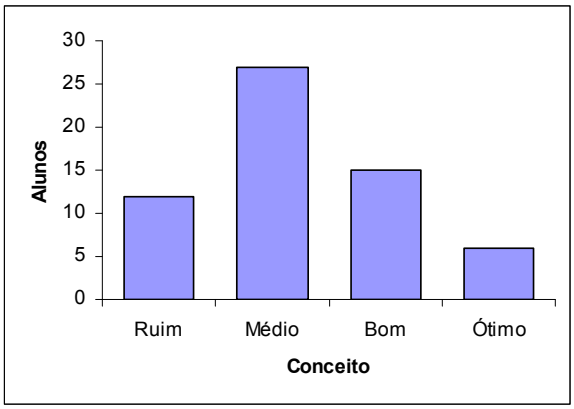


Figura 3. Conceito dos alunos na disciplina de Estatística. UFPel, 2001.

Variável categórica

Gráfico de setores

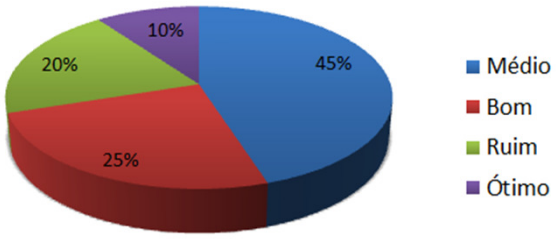


Figura 4. Conceito dos alunos na disciplina de Estatística. UFPel, 2001.

oi

Polígono de frequências

O polígono de frequências é constituído por segmentos de retas que unem os pontos cujas coordenadas são o ponto médio e a frequência de cada classe. Para fechá-lo toma-se uma classe anterior a primeira e uma posterior a última, uma vez que ambas possuem frequência zero.

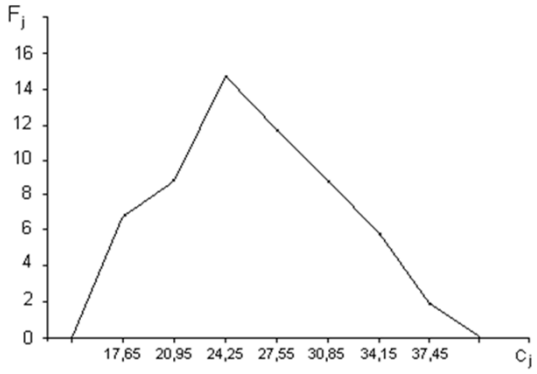
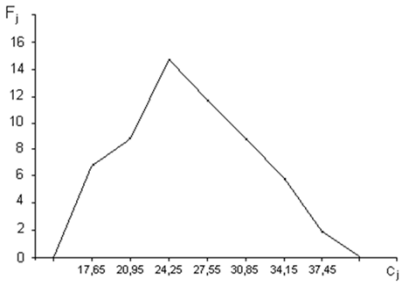
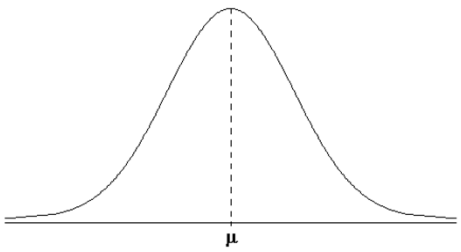


Figura 5. Polígono de freqüências para o peso ao nascer de 60 bovinos da raça Ibagé. UFPel, 2001.

Polígono de frequências



Curva normal



Formato da distribuição de frequências se assemelha ao formato da distribuição normal

oi

Ogivas → frequências acumuladas

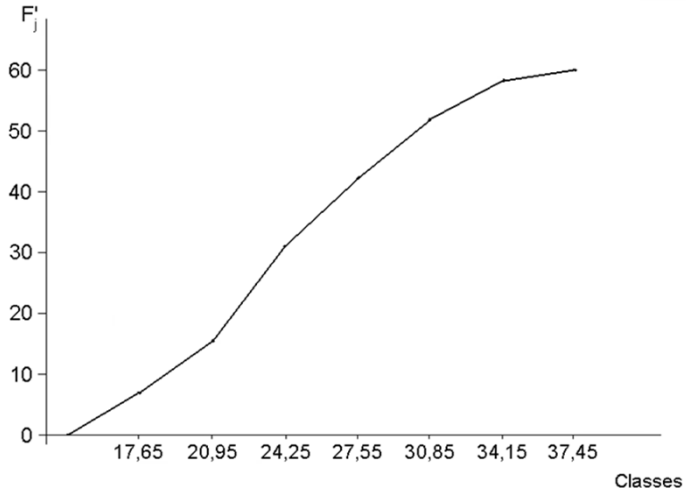


Figura 6. Frequências absolutas acumuladas do peso ao nascer de 60 bovinos da raça Ibagé. UFPel, 2001.

Exemplo:

j	Classe	F_j	F'_j	f_j	f'_j
1	3,11|— 14,39	6	6	0,12	0,12
2	14,39|— 25,67	16	22	0,32	0,44
3	25,67|— 36,95	8	30	0,16	0,60
4	36,95|— 48,23	9	39	0,18	0,78
5	48,23|— 59,51	5	44	0,10	0,88
6	59,51|— 70,79	2	46	0,04	0,92
7	70,79 |— 82,07	0	46	0,00	0,92
8	82,07 |—| 93,35	4	50	0,08	1
	Σ	50	-	1	-

oi

Usando a fórmula empírica

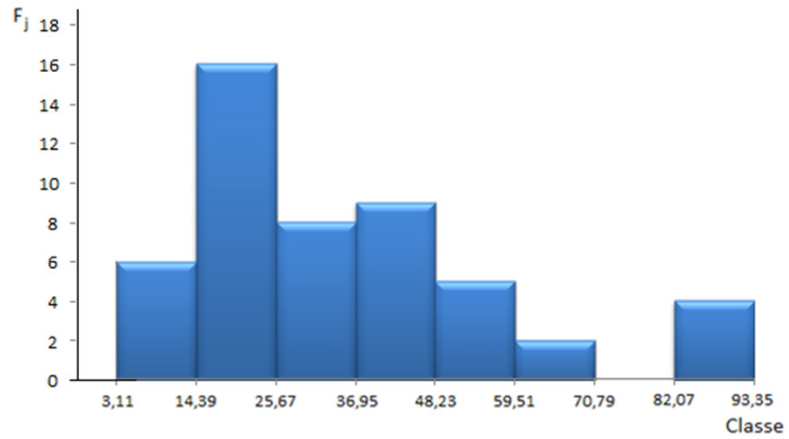


Figura 1. Frequência dos valores gastos (em reais) pelas primeiras 50 pessoas que entraram em um determinado Supermercado, no dia 01/01/2000.

Usando a fórmula de Sturges

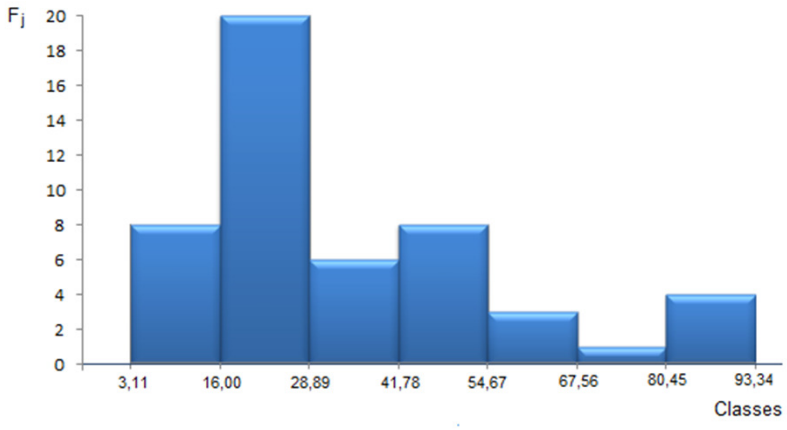


Figura 1. Frequência dos valores gastos (em reais) pelas primeiras 50 pessoas que entraram em um determinado Supermercado, no dia 01/01/2000.

oi

Exercício proposto:

j	Classe	F_j
1	4,5|— 5,22	2
2	5,22|— 5,94	5
3	5,94|— 6,66	3
4	6,66|— 7,38	8
5	7,38|— 8,10	6
6	8,10|— 8,82	9
7	8,82 |—| 9,54	7
	Σ	40

Construa um histograma para as frequências absolutas das notas dos alunos dos curso de Ciência e Engenharia da Computação da UFPel na primeira prova de Estatística Básica, no segundo semestre de 2013.

Bibliografia

FERREIRA, D.F. **Estatística básica**. Lavras: Editora UFLA, 2005.

MONTGOMERY, D.C.; RUNGER, G.C.; HUBELE. N.F. **Estatística Aplicada à Engenharia**. 2 ed. Rio de Janeiro: Editora LTC. 2004. 335p.

SILVEIRA JUNIOR, P. ; MACHADO, A.A. ; ZONTA, E.P.; SILVA, J.B. da. **Curso de Estatística** v.1. Pelotas: Universidade Federal de Pelotas, 1992, 135p.

Sistema Galileu de Educação Estatística. Disponível em: http://www.galileu.esalq.usp.br/topico.html