1 Estatística Descritiva

A estatística descritiva é parte da estatística que lida com a organização, resumo e apresentação de dados. Esta é feita por meio de:

- Tabelas;
- Gráficos;
- Medidas Descritivas (média, variância, entre outras).

1.1 Tipo de Variáveis

As variáveis podem ter valores numéricos ou não numéricos.

- Variáveis Qualitativas (ou categóricas) - são as características que não possuem valores quantitativos, mas, ao contrário, são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos
  - Variáveis nominais: não existe ordenação dentre as categorias.
    Exemplos: sexo, cor dos olhos, fumante/não fumante, doente/sadio.
  - Variáveis ordinais: existe uma ordenação entre as categorias.
    Exemplos: escolaridade ($1^o, 2^o, 3^o$ graus), estágio da doença (inicial, intermediário, terminal), mês de observação (janeiro, fevereiro,..., dezembro).
- Variáveis Quantitativas - são as características que podem ser medidas em uma escala quantitativa, ou seja, apresentam valores numéricos
  - Variáveis discretas: são aquelas variáveis que pode assumir somente valores inteiros num conjunto de valores. É gerada pelo processo de contagem
    Exemplos: número de filhos, número de empregados, número de processos.
  - Variáveis contínuas: são aquelas variáveis que podem assumir um valor dentro de um intervalo de valores. É gerada pelo processo de medição
    Exemplos: pressão arterial, idade, salário, atraso de transmissão de bytes por uma rede de internet.

1.2 Variáveis Qualitativas

Para resumir dados qualitativos, utiliza-se contagens, proporções, porcentagens, taxas por 1000, taxas por 1.000.000, etc, dependendo da escala apropriada. Por exemplo, se encontrarmos que 7 empresas com faturamento mensal acima de R$20.000,00 em uma amostra de 500 propriedades, poderíamos expressar isto como uma proporção (0,014) ou percentual (1,4%).

Freqüentemente o primeiro passo da descrição de dados é criar uma tabela de freqüências. Antes de montar a tabela de distribuição de freqüências temos algumas definições:

- Freqüência - medida que quantifica a ocorrência dos valores de uma variável a um dado conjunto de dados. As freqüências podem ser:

– Absoluta (fa) - contagem das observações de uma variável;

– Relativa (fr) - divisão da freqüência absoluta pelo total de observações

$$fr = \frac{fa}{n}$$

– Percentual (fp) - é a freqüência relativa multiplicada por 100

$$fp = 100 \times fr$$

Exemplo: Para adequar os produtos às preferências dos clientes, um provedor fez uma pesquisa sobre os provedores a qualidade dos serviços prestados utilizando uma amostra de 20 clientes, obtendo as seguintes variáveis:

Tabela 1: Variáveis observadas de 20 clientes de um provedor.

Amostra	Sexo	Qualidade	Amostra	Sexo	Qualidade
1	feminino	Boa	11	feminino	Ruim
2	feminino	Boa	12	feminino	Ruim
3	feminino	Boa	13	masculino	Boa
4	feminino	Boa	14	masculino	Boa
5	feminino	Boa	15	masculino	Ótimo
6	feminino	Ótimo	16	masculino	Regular
7	feminino	Ótimo	17	masculino	Regular
8	feminino	Regular	18	masculino	Ruim
9	feminino	Regular	19	masculino	Ruim
10	feminino	Ruim	20	masculino	Ruim

Neste é apresentado duas variáveis qualitativas sendo:

- Sexo - variável qualitativa nominal;
- Qualidade - variável qualitativa ordinal;

Para resumir separadamente cada variável podemos utilizar a tabelas simples, que são na maioria das vezes suficientes para descrever dados qualitativos especialmente quando existem poucas categorias.

Para a variável sexo, podemos utilizar as freqüências apresentadas na tabela 2:

Tabela 2: Distribuição de freqüência do sexo de 20 clientes de um provedor.

Sexo	Freqüência Absoluta (fa)	Freqüência Relativa (fr)	Freqüência Percentual (fp)
feminino	12	0,60	60%
masculino	8	0,40	40%
	20	1,00	100%

Para a variável qualidade no atendimento, além das freqüências utilizadas para a variável sexo, podemos utilizar mais duas freqüências:

- Freqüência Acumulada (FA)- obtida pelo soma das freqüências absolutas;
- Freqüência Percentual Acumulada (FP) - obtida pela soma das freqüências percentuais.

Tabela 3: Distribuição de freqüência qualidade no atendimento de um provedor de acordo com 20 clientes

Qualidade no Atendimento	Freqüência Absoluta (fa)	Freqüência Relativa (fr)	Freqüência Percentual (fp)	Freqüência Acumulada (FA)	Freqüência Percentual Acumulada (FP)
Ótima	3	0,15	15%	3	15%
Boa	7	0,35	35%	10	50%
Regular	4	0,20	20%	14	70%
Ruim	6	0,30	30%	20	100%
Total	20	1,00	100%	-	-

Dados qualitativos são usualmente bem ilustrados num simples gráfico de barras onde a altura da barra é igual à freqüência. O gráfico na Figura **??** apresenta as freqüências percentuais da Tabela 2.

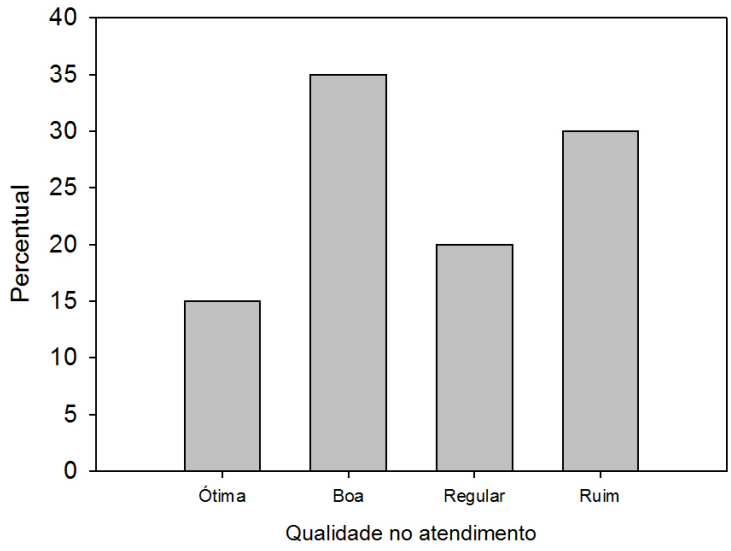

Figura 1: Qualidade no atendimento de um provedor de acordo com 20 clientes

Em alguns casos podemos estar interessados em resumir duas variáveis qualitativas ao mesmo tempo, neste caso vamos estudar a relação entre duas variáveis qualitativas que pode ser representada em uma tabulação cruzada. Nesta tabela conta-se quantos valores correspondem a cada par de possíveis resultados, para as duas variáveis. O resultado pode ser apresentado como freqüência absoluta ou relativa, em relação as colunas ou as linhas (nunca ambas).

O gráfico de barras, com barras justapostas de acordo com categorias diferentes, pode ser usado para apresentar a relação entre duas variáveis qualitativas.

Tabela 4: Distribuição de freqüência absoluta de 20 clientes de um provador de acordo com a qualidade de atendimento e o sexo

Qualidade	Sexo		Total
	Feminino	Masculino	
Boa	5	2	7
Ótimo	2	1	3
Regular	2	2	4
Ruim	3	3	6
Total	12	8	20

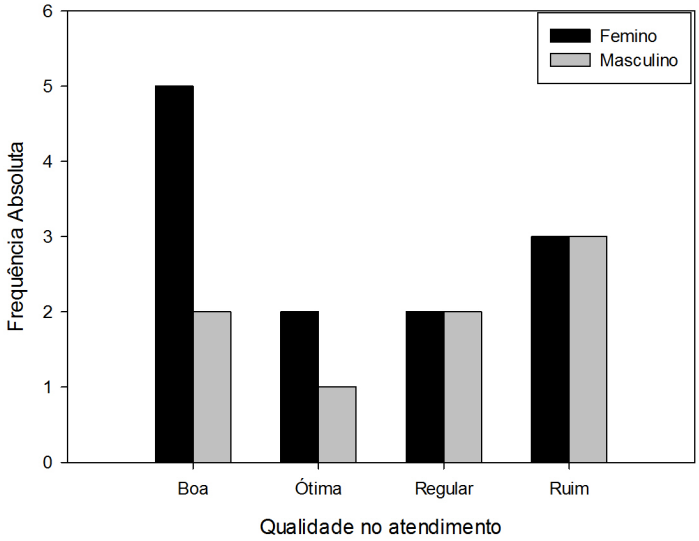

Figura 2: Distribuição de freqüência absoluta de 20 clientes de um provador de acordo com a qualidade de atendimento e o sexo

1.3 Variáveis Quantitativas

Da mesma forma que as variáveis qualitativas, podemos resumir dados quantitativos por meio de tabelas de freqüências, entretanto a distinção entre as variáveis quantitativas discretas e contínuas na forma de preparação destas tabelas.

A tabela de distribuição de freqüências de uma variável discreta é, em geral bastante semelhante à das variáveis qualitativas ordinais, pois os valores inteiros que a variável assume podem ser considerados como "categorias", ou "classes naturais".

Exemplo: Sejam dados referentes a um levantamento onde observou-se o numero de peças defeituosas em 25 maquinas de uma empresas.

Tabela 5: Número de peças defeituosas em 25 maquinas de uma empresa

3	5	7	1	3
6	5	5	5	3
8	5	2	6	2
4	4	4	3	5
6	2	2	4	5

Tabela 6: Distribuição de freqüências do número de peças defeituosas de 25 maquinas de uma empresa

Número de Minerais	Freqüência Absoluta (fa)	Freqüência Relativa (fr)	Freqüência Percentual (fp)	Freqüência Acumulada (FA)	Freqüência Percentual Acumulada (FP)
1	1	0,04	4%	1	4%
2	4	0,16	16%	5	20%
3	4	0,16	16%	9	36%
4	4	0,16	16%	13	52%
5	7	0,28	28%	20	80%
6	3	0,12	12%	23	92%
7	1	0,04	4%	24	96%
8	1	0,04	4%	25	100%
Total	25	1	100%		

Observa-se que a disposição da variável número de de peças defeituosas é semelhante a de uma variável qualitativa ordinal com 8 categorias e sua distribuição de freqüência pode ser vista na tabela 6. A representação gráfica pode ser feita por meio de um gráfico de barras conforme figura 4.

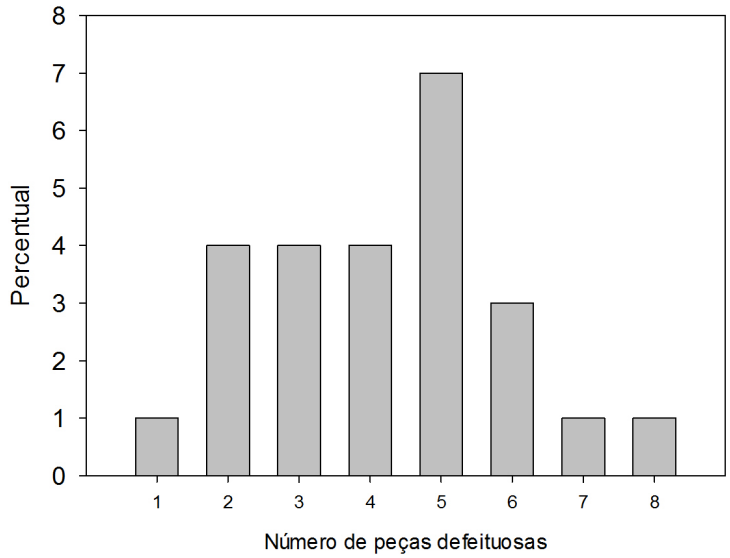

Figura 3: Número número de peças defeituosas de 25 maquinas de uma empresa

A construção de tabelas de distribuição de freqüências para variáveis quantitativas contínuas é feita agrupando os dados em classes e obtendo as freqüências observadas em cada classe. É importante notar que ao resumir dados referentes a uma variável contínua sempre se perde alguma informação já que não temos idéia de como se distribuem as observações dentro de cada classe. Para isso temos duas definições:

- Amplitude (A) - corresponde a diferença enter o maior valor e o menor valor de um conjunto de dados;
- Amplitude da classe (c) - consiste na diferença entre o limite superior e o limite inferior de uma classe em uma distribuição de freqüência.

O procedimento para construir tabelas de distribuição freqüências para variáveis quantitativas contínuas envolve os seguintes passos (algoritmo):

- Decidir sobre o numero de classes k, entre 5 e 20. Para que a decisão não seja totalmente arbitrária pode-se usar a raiz quadrada do total de valores como o número de classes, ou seja, $k \cong \sqrt{n}$
- Determinar a amplitude dos dados: A = Max - Min.
- Determinar a amplitude de classe c:
$$c = \frac{A}{k-1}$$
- Determinar o limite inferior da primeira classe LI_1:
$$LI_1 = Min - \frac{c}{2}$$
- Determinar o limite superior da primeira classe LS_1:
$$LS_1 = LI_1 + c$$
sendo que o limite inferior da segunda classe LI_2 é igual ao LS_1, e assim
$$LS_2 = LI_2 + c$$
e assim, sucessivamente todas as classes vão sendo construídas.
- Após a construção das classes, são contados quantos dados estão contidos em cada classe e se obtem as freqüências.

Tabela 7: Dados ordenados, relativos ao tempo em segundos para carga de um aplicativo num sistema compartilhado (30 observações).

6,94	7,27	7,46	7,97	8,03	8,37
8,56	8,66	8,88	8,95	9,30	9,33
9,55	9,76	9,80	9,82	9,98	9,99
10,14	10,19	10,42	10,44	10,66	10,88
10,88	11,16	11,80	11,88	12,25	12,34

$$\begin{aligned} k &= \sqrt{30} = 5,47 \approx 5 \\ A &= Max - Min = 12,34 - 6,94 = 5,40 \\ c &= \frac{A}{k-1} = \frac{5,40}{4} = 1,35 \\ LI_1 &= Min - \frac{c}{2} = 6,94 - \frac{1,35}{2} = 6,94 - 0,67 = 6,27 \end{aligned}$$

Tabela 8: Distribuição de freqüências, relativa ao ao tempo em segundos para carga de um aplicativo num sistema compartilhado.

Classes			Freqüência Absoluta (fa)	Freqüência Relativa (fr)	Freqüência Percentual (fp)	Freqüência Acumulada (FA)	Freqüência Percentual Acumulada (FP)
6,27	⊢	7,62	3	0,10	10%	3	10%
7,62	⊢	8,97	7	0,23	23%	10	33%
8,97	⊢	10,32	10	0,33	33%	20	67%
10,32	⊢	11,67	6	0,20	20%	26	87%
11,67	⊢	13,02	4	0,13	13%	30	100%
			30	1,00	100%		

Uma forma de representar graficamente à distribuição de freqüência das variáveis contínuas é por meio do histograma e do polígono de freqüência . Para elaboração deste gráfico é comum utilizar a chamada densidade de freqüência absoluta (dfa)

$$dfa = \frac{fr}{c}$$

O histograma é semelhante ao gráfico de barras verticais, no eixo vertical pode-se utilizar as freqüências ou densidades de freqüências e no eixo horizontal as classes. O polígono de freqüências é um gráfico de linhas em que no eixo vertical pode-se utilizar as freqüências ou densidades de freqüências e no eixo horizontal o ponto médio de cada classe.

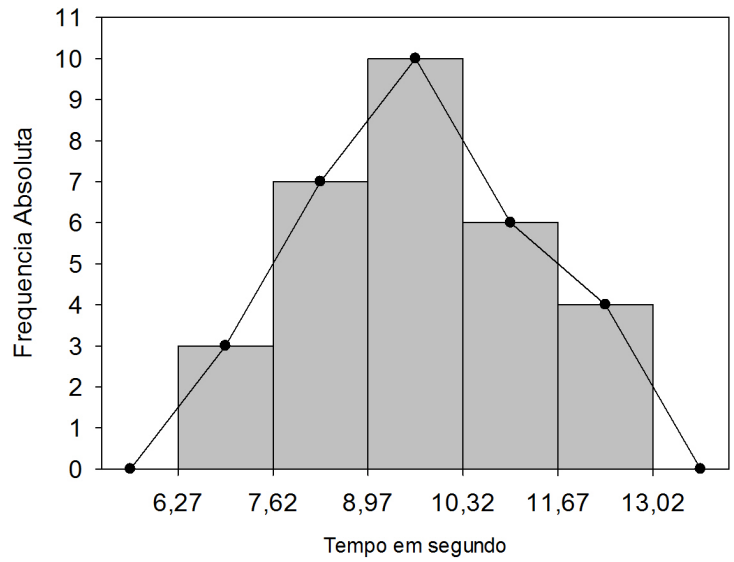

Figura 4: Histograma e Polígono de freqüências do relativa ao tempo em segundos para carga de um aplicativo num sistema compartilhado

Muitas vezes, a análise da distribuição de freqüências acumuladas é mais interessante do que a de freqüências simples, representada pelo histograma. O gráfico usado na representação gráfica da distribuição de freqüências acumuladas de uma variável contínua é a ogiva, apresentada na Figura 5. Para a construção da ogiva, são usadas as freqüências acumuladas (absolutas ou percentuais) no eixo vertical e os limites superiores

de classe no eixo horizontal.

O primeiro ponto da ogiva é formado pelo limite inferior da primeira classe e o valor zero, indicando que abaixo do limite inferior da primeira classe não existem observações. Daí por diante, são usados os limites superiores das classes e suas respectivas freqüências acumuladas, até a última classe, que acumula todas as observações. Assim, uma ogiva deve começar no valor zero e, se for construída com as freqüências relativas acumuladas, terminar com o valor 100.

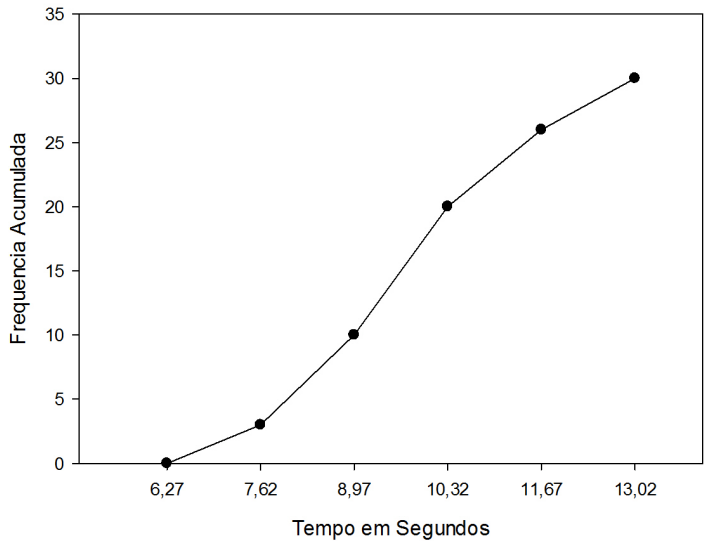

Figura 5: Ogiva para o tempo em segundos para carga de um aplicativo num sistema compartilhado