Tabelas ou distribuições de frequências

Simples e Cruzadas

- ✓ Em tempos de *Big Data* o mundo encheu-se de dados. As empresas têm dados de suas atividades, de seus funcionários, de seus clientes, etc. Mas para que esses dados sejam informativos, necessitamos organiza-los, resumi-los e apresenta-los de forma adequada.

- ✓ Estes dados podem estar estruturados de forma complexa. Entretanto o tipo de dado mais comum nas análises estatísticas são aqueles organizados de forma tabular ou de matrizes. As linhas dessas matrizes correspondem ao que se observou em cada elemento pesquisado, enquanto as colunas correspondem às características (variáveis) levantadas.

    - Por exemplo, na atualização das páginas de um *site*, podemos querer avaliar o perfil dos indivíduos que acessam esse *site*. Então, precisamos levantar, junto a cada indivíduo, algumas de suas características, tais como o sexo, a idade, o nível de instrução e o provedor utilizado. Ao realizar a pesquisa, podemos produzir uma matriz de dados, da seguinte forma:

Variáveis

Usuário	***Sexo***	***Idade (em anos)***	***Nível de instrução***	***Provedor***
1	*M*	*35*	*Superior*	*A*
2	*F*	*18*	*Fundamental*	*C*
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	*F*	*23*	*Médio*	*B*

Indivíduos ou casos (apontam para as linhas)

Dados (abrangem as linhas de dados)

- ✓ Tipologia de variáveis:
  As variáveis podem ser divididas em 4 tipos básicos:

  • Variáveis numéricas ou quantitativas: Consistem em medidas ou contagens numéricas. Para variáveis numéricas faz sentido somar seus valores (para obter um total geral, por exemplo), subtrair (para medir a diferença entre dois casos, por exemplo) ou tomar médias de seus valores. Elas podem ser divididas em:
  – discreta
  – contínua
  Variáveis quantitativas discretas assumem apenas alguns valores possíveis. Estes valores podem ser colocados numa lista enumerável. Exemplos: Número de filhos, número de acessos à plataforma.
  No caso de variáveis numéricas contínuas, seus valores podem assumir qualquer valor num intervalo da reta real. Exemplos: Altura, peso, salário.

• Variáveis categóricas ou qualitativas: Como o nome está dizendo, os valores possíveis de variáveis categóricas são categorias. Os valores são apenas rótulos indicando diferentes categorias em que os casos podem se classificados. Com estas variáveis categóricas, não faz sentido fazer operações aritméticas com seus valores. Assim, em princípio, nós não somamos, subtraimos ou tiramos médias de colunas na tabela que sejam variáveis categóricas. Elas podem ser divididas em:

– nominal

– ordinal

No caso de variáveis qualitativas ordinais, o valor é um rótulo para uma categoria dentre k possíveis e as categorias podem ser ordenadas. Existe uma ordem natural nos valores possíveis. Exemplos: Escolaridade, estágio da doença, classe social.

No caso de variáveis categóricas nominais, os seus valores possíveis são rótulos de categorias que não podem ser ordenadas. Exemplos: Profissão, sexo, religião.

1 Distribuição de freqüências

- Um dos primeiros passos para analisar um arquivo de dados, especialmente quando o número de observações for grande, é a distribuição de freqüências de cada variável.
  - Uma distribuição de freqüência é uma tabela que mostra categorias, valores ou intervalos de valores de acordo com as ocorrências.
- Observe os exemplos de distribuições de freqüências abaixo:

Tabela 1 – Distribuição de freqüências do provedor usado pelo visitante do site

Provedor	Freqüência	Porcentagem
A	10	25,0
B	17	42,5
C	7	17,5
D	6	15,0
Total	40	100,0

Tabela 2 – Distribuição de freqüência do número de defeitos encontrados em escadas no final da linha de produção

Número de defeitos encontrados	Freqüência	Porcentagem	Freqüência acumulada	Porcentagem acumulada
0	13	27,1	13	27,1
1	15	31,3	28	58,4
2	10	20,8	38	79,2
3	7	14,6	45	93,8
4	2	4,2	47	98
7	1	2,1	48	100
Total	48	100,0		

É possível observar que 14,6% das escadas apresentaram 3 defeitos e que 93,8% delas apresentaram no máximo 3 defeitos.

Observação importante: Veja que na Tabela 2 há uma coluna contendo as *freqüências acumuladas* e as *porcentagens acumuladas* que são obtidas somando-se os valores menores ou iguais ao valor considerado. Essa coluna é calculada para variáveis quantitativas, mas existem variáveis qualitativas (especialmente as ordinais) em que a freqüência acumulada e a porcentagem acumulada também têm utilização prática.

- Há situações em que uma variável discreta ou contínua apresenta uma quantidade muito grande de valores diferentes de modo que, se a tabela fosse construída nos mesmos moldes que as anteriores obteríamos praticamente os valores originais do banco de dados. A alternativa que vamos adotar consiste em construir intervalos de valores e contar o número de ocorrências em cada intervalo.
- Esses intervalos devem ser mutuamente exclusivos e, de preferência, ter o mesmo tamanho. Porém, em alguns casos, não é possível usar intervalos com a mesma amplitude. Por exemplo, se formos analisar os salários de uma empresa, provavelmente encontraremos os valores distribuídos em uma grande amplitude e a maioria deles concentrados na parte inferior da escala. Assim, pode ser conveniente usarmos intervalos menores para os valores iniciais e intervalos maiores (mais amplos) para valores finais.
- A Regra de Sturges é uma das regras mais utilizadas na Estatística para construção de uma tabela de frequências por intervalos. Isso porque a fórmula de Sturges nos fornece uma quantidade adequada de classes para os mais variados tamanhos de amostras. A regra de Sturges envolve os seguintes passos:

1. Cálculo da amplitude: $A = \max - \min$
2. Cálculo do número de *intervalos* da tabela: $k = 1 + 3{,}322.\log n$
3. Cálculo da amplitude dos *intervalos*: $A_k = \frac{A}{K}$

Exemplo 1: Distribuição de frequências por intervalos

Os dados, a seguir, representam o tempo (em segundos) que operadores gastam para montar um equipamento na linha de produção de uma empresa:

4,7	4,9	5,1	5,4	5,7	6	6,3	6,8	7,3	8,9
4,8	4,9	5,2	5,5	5,7	6,2	6,4	6,9	8,2	9,1
4,8	5	5,3	5,6	5,7	6,2	6,5	7	8,2	9,9
4,9	5	5,4	5,6	5,9	6,2	6,7	7,1	8,3	14,1
4,9	5	5,4	5,7	6	6,3	6,8	7,3	8,4	15,2

Utilizando os passos acima, temos:
A = 15,2 – 4,7 = 10,5
K = 1 + 3,322 . log 50 = 1 + 3,322 . 1,699 = 6,644 (pode-se escolher entre 6 ou 7 classes)
A_k= 10,5 / 6 = 1,75 (nesse caso, escolhendo 6 classes, cada classe deverá ter uma amplitude de 1,8 segundos, seguimos o mesmo número de casas decimais dos dados da amostra)

Tabela 3 - Distribuição de freqüência do tempo gasto (em segundos) para a montagem de um equipamento na linha de produção.

Tempo	Freqüência	Porcentagem	Freqüência acumulada	Porcentagem acumulada
4,7 ⊢ 6,5	32	64,0	32	64,0
6,5 ⊢ 8,3	11	22,0	43	86,0
8,3 ⊢ 10,1	5	10,0	48	96,0
10,1 ⊢ 11,9	0	0,0	48	96,0
11,9 ⊢ 13,7	0	0,0	48	96,0
13,7⊢ 15,5	2	4,0	50	100,0
Total	50	100,0	-	-

Observe que o símbolo ⊢ é utilizado para representar intervalo aberto/fechado, ou seja, o valor que se encontra do lado direito do símbolo não está incluído no intervalo (intervalo aberto) e o valor que se encontra do lado esquerdo do intervalo está incluído no intervalo (intervalo fechado). Portanto, a primeira classe de valores 4,7 ⊢ 6,5 indica que o intervalo vai de 4,7 a 6,49.

Os passos para construção de tabelas de freqüências não são uma imposição legal, portanto, o pesquisador é livre para criar o número de classes com as amplitudes que lhe forem mais convenientes.

Ao analisar variáveis quantitativas, normalmente, três informações principais são procuradas:

- ✓ Faixa em que os valores ocorrem com maior freqüência (faixa de valores típicos);
- ✓ Valores discrepantes, que podem ter sido originados de erros de mensuração ou digitação, mas também podem corresponder a elementos que apresentam comportamento muito diferente dos demais;
- ✓ Forma da distribuição, a fim de compará-la com modelos probabilísticos, o que nos permite usar técnicas mais avançadas de análise.

2 Tabelas de contingência ou Tabelas Cruzadas

- As tabelas de contingência são a forma usual de apresentar uma distribuição de freqüência conjunta de duas ou mais variáveis.
- Objetivos: Apresentar sob a forma de uma tabela de dupla entrada a informação de duas variáveis categorizadas.

Exemplo 2: Tabela cruzada

A Tabela 4 mostra o número de defeitos encontrados em uma amostra de pneus de veículos utilitários em relação ao lado e a posição.

Tabela 4 – Distribuição de freqüência do número de defeitos de acordo com a posição e o lado do pneu em veículos utilitários

Lado	Posição		Total
	Dianteira	Traseira	
Direito	32	28	60
Esquerdo	35	57	92
Total	67	85	152

Podemos calcular percentuais para cada célula da tabela, seja em relação ao total das linhas (lado do veículo), ao total das colunas (posição do pneu) ou ao total geral da tabela, o que possibilita uma avaliação da qualidade dos processos.

Por exemplo, para a frequência 32 que representa o número de defeitos identificados em pneus no lado direito e na posição dianteira do veículo, iremos calcular os três percentuais possíveis:

% da linha: $\frac{32}{60}.100 = 53{,}33\%$, ou seja, entre os defeitos localizados no lado direito do veículo, 53,33% deles estavam na posição dianteira.

% da coluna: $\frac{32}{67}.100 = 47{,}76\%$, ou seja, entre os defeitos localizados na posição dianteira, 47,76% deles estavam no lado direito do veículo.

% do total: $\frac{32}{152}.100 = 21{,}05\%$, ou seja, 21,05% dos defeitos identificados estavam na posição dianteira e no lado direito do veículo.

Veja a tabela com os percentuais da linha, da coluna e do total calculados a partir dos dados da Tabela 4.

Lado	Posição		Total
	Dianteira	Traseira	
Direito	32	28	60
% da linha	*53%*	*47%*	*100%*
% da coluna	*48%*	*33%*	*39%*
% total	*21%*	*18%*	*39%*
Esquerdo	35	57	92
% da linha	*38%*	*62%*	*100%*
% da coluna	*52%*	*67%*	*61%*
% total	*23%*	*38%*	*61%*
Total	67	85	152
% da linha	*44%*	*56%*	*100%*
% da coluna	*100%*	*100%*	*100%*
% total	*44%*	*56%*	*100%*

A Tabela 5 mostra o acréscimo de uma terceira variável (marca) aos dados da Tabela 4.

Tabela 5 – Distribuição de freqüência do número de defeitos de acordo com a posição, o lado e a marca do pneu em veículos utilitários.

Marca	Lado	Posição		Total
		Dianteira	Traseira	
A	Direito	32	28	60
	Esquerdo	35	57	92
B	Direito	27	23	50
	Esquerdo	43	55	98
Total		137	163	300

Quanto maior o número de variáveis inseridas simultaneamente numa tabela cruzada, maior deve ser o cuidado na montagem e interpretação dos dados. Na Tabela 5 temos que 110 defeitos ocorreram em pneus que estavam no lado direito do veículo e que 148 defeitos foram identificados em pneus da marca B.