Tabelas e Diagramas de Freqüência

O primeiro passo na análise e interpretação dos dados de uma amostra consiste na descrição (apresentação) dos dados em forma de tabelas, gráficos ou cartas.

Existem diversos tipos de gráficos e cartas que podem ser feitos com os programas de computador disponíveis atualmente. O tipo de gráfico a ser escolhido deve depender do que se quer comunicar. Um gráfico bem feito pode facilitar o entendimento da informação que se quer transmitir.

Atualmente, com a grande variedade de formatos de gráficos disponíveis nos programas de computador, pode-se ficar tentado a usar todo tipo de recurso para produzir gráficos com sombreamentos, rótulos em letras exóticas, em três dimensões, etc. É importante, porém, tomar cuidado com isso para não exagerar e tornar a figura poluída, o que dificulta o entendimento. A regra básica é manter a figura simples e clara.

Exemplo ilustrativo:
A versão digitalizada de um quadro de um famoso pintor foi projetada em uma tela para uma classe de 12 alunos do terceiro ano primário e pediu-se a cada um que escrevesse em uma folha de papel quantas cores diferentes eles podiam identificar no quadro. O tempo dado para o exercício foi 15 minutos e, após o seu término e a entrega das folhas ao professor, este contou, para cada aluno, a quantidade de cores identificada. Os resultados estão dados abaixo, onde cada aluno foi identificado por um número de identificação (ID) indo de 1 a 12:

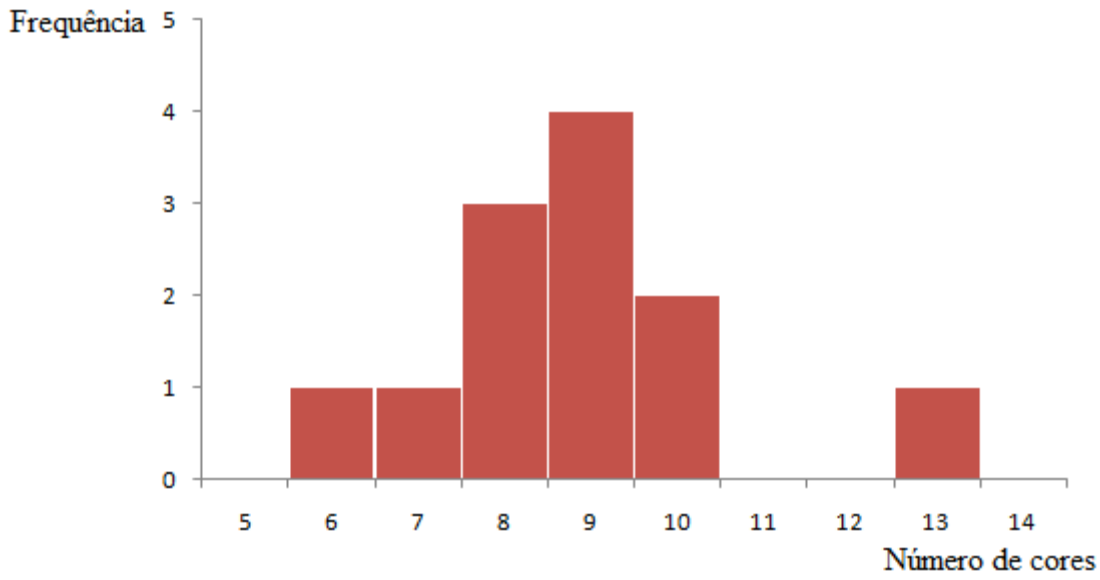
ID	1	2	3	4	5	6	7	8	9	10	11	12
Nº de cores	7	10	8	9	6	9	10	9	13	8	9	8

Para analisar os dados, pode-se tabular o número de vezes que cada quantidade de palavras ocorreu, que é a freqüência de cada quantidade:

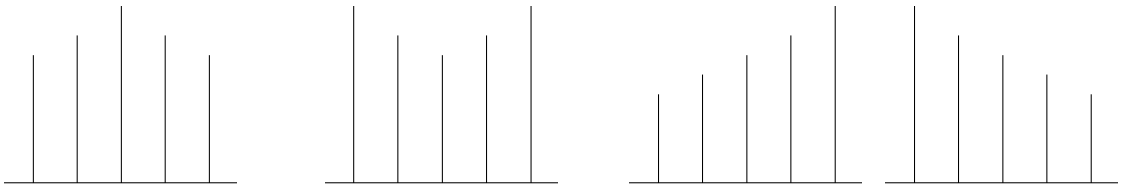
Nº de cores	Freqüência
5	0
6	1
7	1
8	3
9	4
10	2
11	0
12	0
13	1
14	0
Total	**12**

Vê-se que o número de cores percebido pelos alunos está mais concentrado na faixa entre 8 a 10 cores, mas existem alunos que identificaram 6 e 13 cores.

O exemplo acima corresponde ao que se chama de uma distribuição de freqüências. Pode-se apresentar os mesmos dados através de um diagrama de freqüências, onde os valores das freqüências são representados por barras cujas alturas são iguais às freqüências. Um gráfico desse tipo dá uma idéia da "forma" da distribuição:



Distribuições de freqüência podem ser simétricas ou assimétricas:



Simétricas **Assimétricas**

No caso do exemplo anterior, a distribuição de freqüências é aproximadamente simétrica.

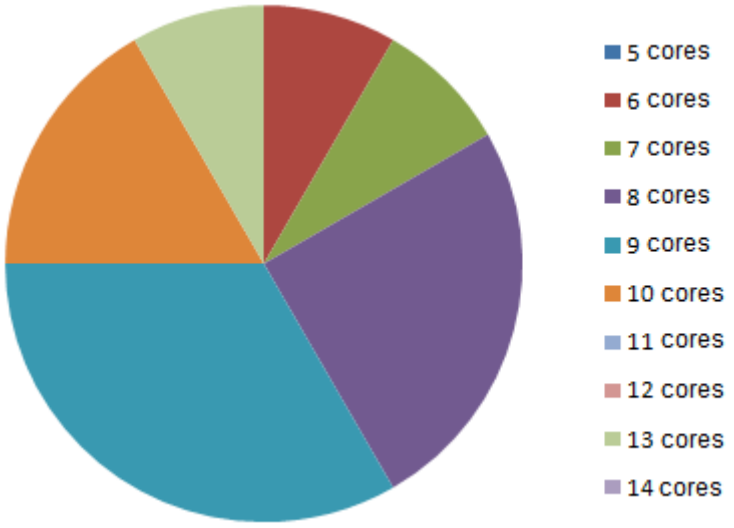
Outra maneira gráfica de se apresentar esses dados é usando os chamados diagramas de tipo pizza (ou torta), em que os tamanhos das "fatias" da pizza são proporcionais às freqüências de cada valor.

Por exemplo, para o caso em questão a circunferência da pizza equivaleria ao total de alunos, doze, o que nos permitiria calcular os segmentos de arco associados a cada setor (fatia) a partir de uma regra de três simples:

$$\begin{array}{rcr} 360 & — & 12 \\ x & — & 1 \end{array} \qquad \Rightarrow x = 360/12 = 30$$

Cada aluno corresponderia então a um setor da torta com 30^o, contado a partir de um eixo qualquer. Segundo os dados, há 4 alunos que identificaram 9 cores. Portanto, o setor correspondente a 9 cores teria um ângulo igual a $4x30^o = 120^o$. O setor correspondente a 10 cores teria um ângulo igual a $2x30^o = 60^o$ e assim por diante.

O diagrama de pizza dando as freqüências associadas a cada número de cores seria então (o eixo de onde se começou a contar os ângulos é o eixo correspondente a 12 hs e os ângulos foram contados no sentido horário):



Histogramas

Muitas vezes, o número de diferentes valores de uma dada variável medidos para uma amostra é muito grande.

Exemplo: Valores de acetilcolina nos glóbulos vermelhos do sangue (medidos em μmol/ml) de 35 trabalhadores rurais expostos a pesticidas (**Note que há poucos valores repetidos, e no máximo duas vezes**):

10,6 - 9,9 - 12,6 - 15,2 - 12,3 - 9,2 - 11,7 - 12,3 - 12,5 - 11,8 - 12,4 - 10,2 - 11,3 - 9,4 - 11,4 - 11,0 - 11,6 - 12,2 - 13,4 - 9,9 - 11,0 - 8,6 - 12,5 - 9,8 - 11,6 - 12,6 - 16,7 - 7,7 - 10,9 - 10,1 - 8,7 - 9,0 - 15,3 - 10,2 - 10,9

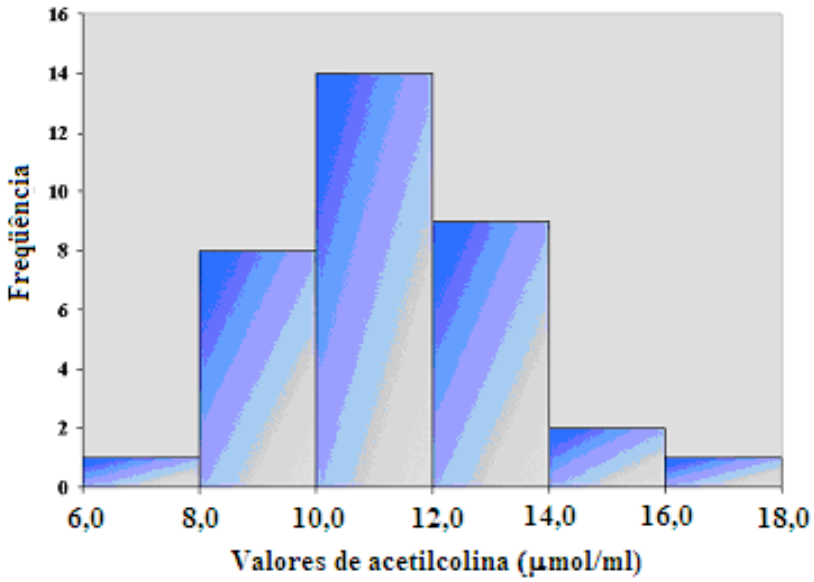
Nesses casos, é mais prático agrupar os possíveis valores da variável em classes e construir um diagrama de freqüências para as classes ao invés de para os valores individuais.

Com este processo de agrupamento perde-se o conhecimento sobre cada dado individual, mas ganha-se uma noção global sobre o comportamento da amostra (*a sua estrutura*), o que é muitas vezes mais importante.

Para o caso do exemplo, vamos agrupar os valores medidos em seis classes, indo de 6,0 (inclusive) até 18,0 (exclusive) de dois em dois. Obtemos então a seguinte distribuição de freqüências:

Valores de Acetilcolina	Freqüência
6,0 ⊢ 8,0	1
8,0 ⊢ 10,0	8
10,0 ⊢ 12,0	14
12,0 ⊢ 14,0	9
14,0 ⊢ 16,0	2
16,0 ⊢ 18,0	1
Total	**35**

A partir da distribuição de freqüências, pode-se montar o diagrama de freqüências para os dados agrupados. Um diagrama de freqüências para dados agrupados é chamado de histograma.



Note que o histograma permite uma rápida apreensão visual dos dados. Por exemplo, ele revela que a distribuição dos valores é aproximadamente simétrica em torno do intervalo "10,0 ⊢ 12,0 µmol/ml" e que esta é a faixa de valores mais freqüentes.

O uso da notação de intervalo fechado num extremo e aberto no outro, **(├)** ou **(┤)**, tem o objetivo de evitar ambigüidades no *posicionamento* de um dado dentro de um intervalo.

Segundo esta notação, o valor posicionado junto ao extremo fechado do intervalo (barra vertical) é considerado como incluído dentro do intervalo e o valor colocado junto ao extremo aberto do intervalo (sem barra vertical) é considerado como excluído do intervalo, devendo fazer parte de outro intervalo.

Por exemplo, vamos supor que uma das medidas de nível de acetilcolina tivesse o valor 10,0 μmol/ml. Em qual classe ele deveria ser colocado, na segunda ou na terceira?

Segundo a nossa definição, o limite superior da segunda classe exclui o valor 10,0 (teoricamente, ele vai até 9,99...). Já o limite inferior da terceira classe inclui o valor 10,0 (ele começa exatamente daí). Portanto, o valor 10,0 deve ser colocado na terceira classe.

Construção de Tabelas e Diagramas de Freqüências para Dados Agrupados

O primeiro passo na construção de uma tabela de freqüências é ordenar os dados por magnitude.

Para o exemplo dos valores de acetilcolina no sangue mostrados anteriormente, a ordenação por magnitude daria:

7,7 - 8,6 - 8,7 - 9,0 - 9,2 - 9,4 - 9,8 - 9,9 - 9,9 - 10,1 - 10,2 - 10,2 - 10,6 - 10,9 - 10,9 - 11,0 - 11,0 11,3 - 11,4 - 11,6 - 11,6 - 11,7 - 11,8 - 12,2 - 12,3 - 12,3 - 12,4 - 12,5 - 12,5 - 12,6 - 12,6 - 13,4 - 15,2 - 15,3 - 16,7

O próximo passo consiste em determinar o número de classes e as localizações dos seus intervalos.

Em geral, um número entre 5 e 8 classes será suficiente. Um número muito pequeno iria obscurecer detalhes sobre os dados e um número muito grande iria contra o espírito de se agrupar dados em classes.

Os dados do exemplo variam de 7,7 a 16,7, ou seja cobrem 9 unidades. Poderíamos agrupá-los em 5 classes de largura 2; ou 6 classes de largura 2 ou 1,5; ou ainda 7 classes de largura 2 ou 1,5 (não haveria muita diferença entre elas).

A localização dos limites das classes também é um pouco arbitrária: a primeira classe poderia ir de 7,5 a 9,5, ou de 7,0 a 9,0 (caso a escolha fosse de classes de largura 2 unidades); ou de 7,5 a 9,0 (caso se usasse classes de largura 1,5).

Às vezes é necessário tentar várias combinações até se encontrar a apresentação preferida.

Tente, como exercício para casa (não precisa entregar), agrupar os dados deste exemplo usando limites e larguras de classes diferentes dos usados e observe os resultados. Eles são, qualitativamente, muito diferentes entre si?

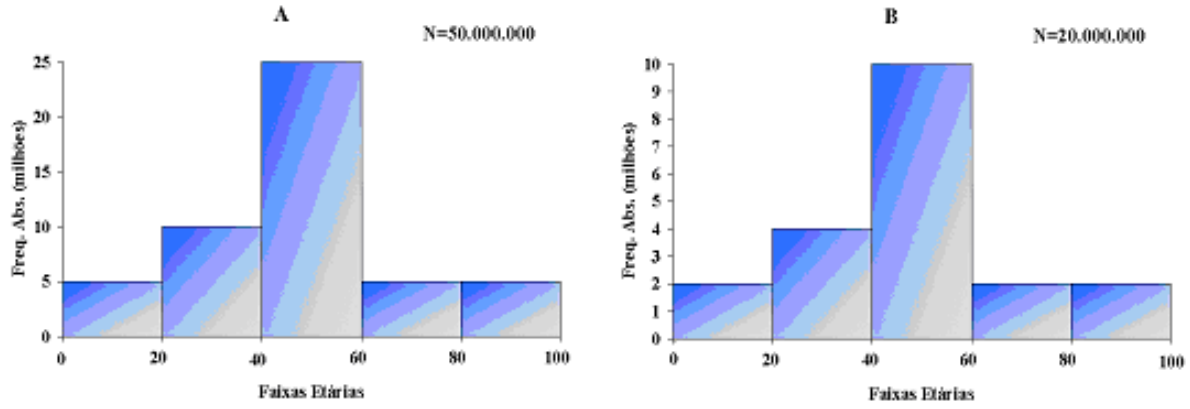
Frequências Relativas e Percentuais

Quando se quer comparar distribuições de freqüências obtidas para amostras distintas, deve-se utilizar frequências relativas, ou percentuais.

A frequência relativa é, por definição, a frequência absoluta dividida pelo número total de dados.

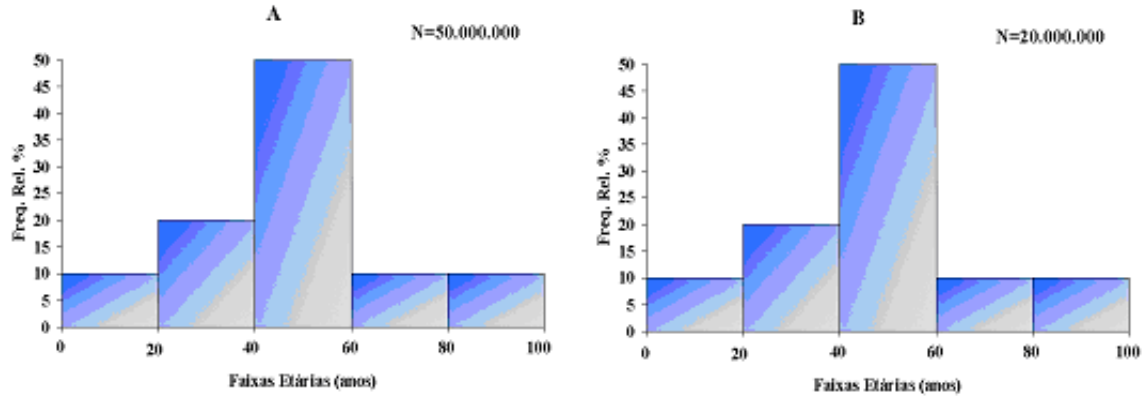
A frequência relativa percentual é obtida multiplicando-se a frequência relativa por 100.

Exemplo: Os histogramas abaixo mostram como as populações de dois países distintos se distribuem por faixa etária. O país A tem 50 milhões de habitantes e o país B tem 20 milhões (dados imaginários).



Segundo os histogramas (que consideram as freqüências absolutas), o país A tem mais velhos que o país B: 5 milhões acima dos 80 anos para o país A contra 2 milhões acima dos 80 anos para o país B. Será que podemos concluir que o país A é um país com população *relativamente* mais velha que o país B?

Para melhor comparar as distribuições populacionais dos dois países, vamos analisar os histogramas para as *frequências relativas percentuais*:



A porcentagem de pessoas acima de 80 anos no país A é exatamente igual à do país B. De fato, constata-se que as distribuições populacionais por faixa etária dos dois países são idênticas! Isto não é evidente a partir da comparação dos histogramas para as freqüências absolutas, o que indica que só se pode comparar duas distribuições quando se usam histogramas para dados relativos.

Note que a comparação entre os histogramas também só é possível porque os intervalos escolhidos para as duas amostras são iguais.

Freqüência Acumulada

Em uma tabela de freqüências de uma dada distribuição, costuma-se listar também a frequência acumulada e a frequência acumulada relativa.

A frequência acumulada para um dado valor é a soma das frequências dos valores menores ou iguais ao valor.

A frequência relativa acumulada para um dado valor é a soma das frequências relativas dos valores até o valor.

A frequência relativa acumulada percentual é a frequência relativa acumulada multiplicada por 100.

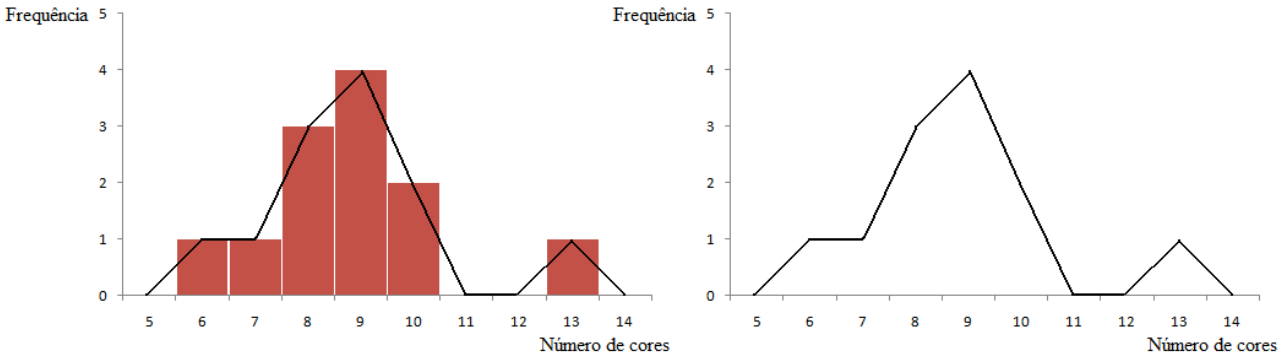
Usando todas as frequências já vistas, a tabela de frequências para o exemplo dos níveis de acetilcolina nos glóbulos vermelhos fica:

Acetilcolina nos glóbulos vermelhos do sangue (μmol/ml)	Frequência (f)	Frequência relativa (f_r)	Frequência relativa percentual (f_r%)	Frequência relativa acumulada ($f_r$ac)	Frequência relativa acumulada percentual ($f_r$ac%)
6,0 ├ 8,0	1	0,03	2,86	0,03	2,86
8,0 ├ 10,0	8	0,23	22,86	0,26	25,71
10,0 ├ 12,0	14	0,40	40,00	0,66	65,71
12,0 ├ 14,0	9	0,26	25,71	0,91	91,43
14,0 ├ 16,0	2	0,06	5,71	0,97	97,14
16,0 ├ 18,0	1	0,03	2,86	1,00	100,00
Total	35	1,00	100,00		

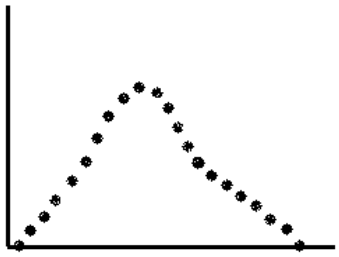
Poder-se-ia também listar as freqüências acumuladas acima de um determinado intervalo. Por exemplo, a frequência acumulada relativa acima do primeiro intervalo é igual a 0,97, acima do segundo é igual a 0,74, etc.

Outras maneiras de representar distribuições

- Polígono de freqüência: faça a união por segmentos de reta dos pontos médios das barras horizontais de um histograma. O exemplo abaixo ilustra a construção de um polígono de freqüências para o histograma do exemplo dos números de cores identificadas no quadro.



Note que se as larguras das classes forem muito pequenas teremos um polígono muito suave, quase contínuo, parecendo-se muito com uma curva (veja abaixo).



A utilização de polígonos de freqüência é muito útil quando se quer comparar dois ou mais histogramas, pois ela permite que se coloque mais de um histograma em um único gráfico sem tornar impossível a sua visualização.

- Diagrama de Ramo e Folhas:

Exemplo: Sejam os dados ordenados:

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

O diagrama de ramo-e-folhas para os dados é:

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

Um aspecto interessante de um diagrama de ramo-e-folhas é que ele combina as vantagens de um histograma (permite uma apreensão visual da forma da distribuição) sem que se percam os dados originais.

Observe que se tivéssemos acesso apenas ao diagrama de ramo-e-folhas do exemplo dado, sem conhecermos os dados originais, ainda assim seria possível reconstruir todos os dados (bastaria sabermos que os números que estão na coluna dos ramos correspondem às partes inteiras dos dados e que os que estão nas colunas das folhas correspondem às partes decimais).

Como outro exemplo seja o diagrama de ramo-e-folhas abaixo, dando a distribuição dos salários de 12 empregados em uma firma (na coluna dos ramos temos os algarismos dos milhares e das centenas e na coluna das folhas temos os algarismos das dezenas e das unidades; não há valores com centavos):

5	20	84	
6	50		
7	00	00	50
8	00	90	
9	00		
10			
11	00		
12	50	80	

O diagrama nos permite ver que os salários dos 12 funcionários não são distribuídos simetricamente e que a sua distribuição tem um pico em torno de 700 a 800 reais. Além disso, ele nos permite reconstruir os valores dos 12 salários: R$ 520; R$ 584; R$ 650; R$ 700; R$ 700; R$ 750; R$ 800; R$ 890; R$ 900; R$ 1100; R$ 1250 e R$ 1280.

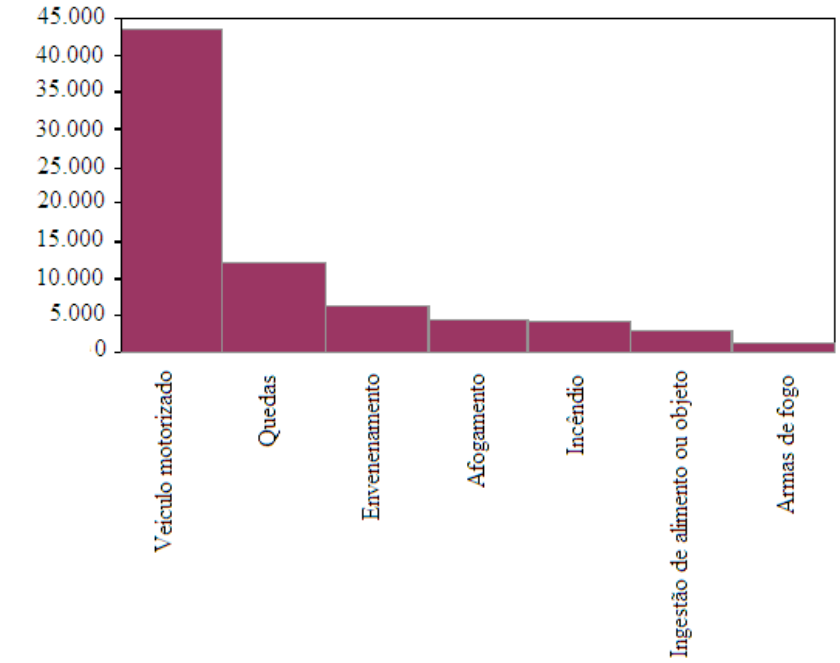
- Diagrama de Pareto:

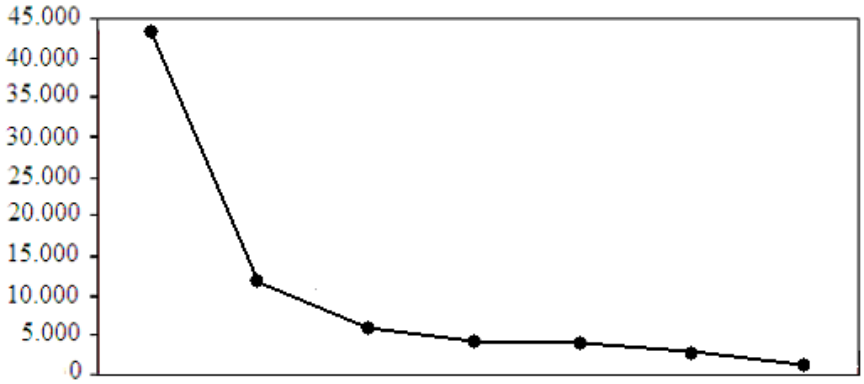
Um diagrama de Pareto é um tipo de histograma, com barras verticais dando as frequências de ocorrência dos eventos, só que para dados categóricos (não numéricos).

Como os dados são qualitativos, não temos como ordená-los numericamente em classes ao longo do eixo-*x*.

Em um diagrama de Pareto, as categorias são ordenadas ao longo do eixo-*x* de acordo com a sua frequência, da maior para a menor. Podemos dizer que um diagrama de Pareto organiza as categorias de acordo com o seu *ranking*, da categoria mais comum à menos comum.

Como exemplo, dá-se abaixo um diagrama de Pareto para as causas de morte acidental de uma população de 75.200 pessoas nos Estados Unidos em um ano recente (exemplo do livro do Triola). Note que ele permite ver claramente que muito mais pessoas morrem por acidentes com veículos motorizados do que por acidentes com armas de fogo. O mesmo diagrama está representado abaixo em termos de um polígono de freqüências.





- Tabela de Contingência:

Uma tabela de contingência é uma maneira de resumir dados categóricos para análise posterior. Ela representa dados relativos a duas categorias que podem estar relacionadas de forma tabular, com os sub-níveis de uma das categorias no topo da tabela (na horizontal) e os sub-níveis da outra categoria na lateral da tabela (na vertical). Cada célula na tabela corresponde a uma combinação mutuamente exclusiva de sub-níveis das duas categorias. Os valores colocados nas células são as *frequências de ocorrência* das combinações dos sub-níveis.

Por exemplo, considere um estudo feito com $N = 200$ pessoas que bebem cerveja regularmente para avaliar a preferência por três marcas de cerveja (vamos chamá-las de marcas A, B e C). A marca da cerveja é uma das categorias, sub-dividida em três tipos. Vamos supor que o estudo foi feito para se avaliar a preferência pelas marcas de cerveja em função do gênero (masculino ou feminino). O gênero, portanto, é a segunda categoria, com duas sub-divisões. A tabela de contingência para esse estudo seria uma como a dada abaixo.

	Marca de Cerveja			
Gênero	**A**	**B**	**C**	**Total**
Masculino	25 (22,7%)	41 (37,3%)	44 (40%)	110
Feminino	38 (42,2%)	33 (36,7%)	19 (21,1%)	90
Total	63 (31,5%)	74 (37%)	63 (31,5%)	200

Esta é uma tabela de contingência 2x3, porque tem duas linhas e três colunas. Em geral, uma tabela de contingência *n*x*m* tem *n* linhas e *m* colunas.

Observe que a tabela imediatamente nos informa muitas coisas. Ela nos informa que a maioria dos homens (44%) prefere a cerveja da marca C, enquanto que a maioria das mulheres (42,2%) prefere a cerveja da marca A. Ela também nos diz que, independentemente do sexo, a maioria das pessoas (37%) prefere a marca B.

Algumas vezes deseja-se estudar relações entre mais de duas categorias, por exemplo três. Em tais casos, como as tabelas de contingência são sempre entre duas categorias (pois não é possível representar mais de duas categorias em uma superfície bidimensional como a de uma folha de papel ou a tela de um computador), o que se faz é construir uma tabela de contingência entre duas categorias para cada uma das sub-divisões da terceira categoria (e da quarta categoria, etc).

Por exemplo, imagine que também se queira estratificar as 200 pessoas do estudo anterior por faixa etária. Por exemplo, vamos supor que as sub-categorias utilizadas para classificar a faixa etária sejam: < 25 anos; 25–40 anos; 40–65 anos; > 65 anos.
Os dados poderiam então ser representados como quatro tabelas de contingência 2x3, uma para cada faixa etária (veja o exemplo abaixo).

Pessoas com < 25 anos

Gênero	**Marca de Cerveja**			**Total**
	A	**B**	**C**	
Masculino	6 (20%)	12 (40%)	12 (40%)	30
Feminino	7 (29,2%)	8 (33,3%)	9 (37,5%)	24
Total	13 (24,1%)	20 (37%)	21 (38,9%)	54

Pessoas com 25–40 anos

Gênero	Marca de Cerveja A	B	C	Total
Masculino	8 (25%)	12 (37,5%)	12 (37,5%)	32
Feminino	12 (42,9%)	10 (35,7%)	6 (21,4%)	28
Total	20 (33,3%)	22 (36,7%)	18 (30%)	60

Pessoas com 40–65 anos

Gênero	Marca de Cerveja A	B	C	Total
Masculino	7 (25%)	11 (39,3%)	10 (35,7%)	28
Feminino	8 (40%)	7 (35%)	5 (25%)	20
Total	14 (29,2%)	18 (37,5%)	16 (33,3%)	48

Pessoas com > 65 anos

Gênero	Marca de Cerveja A	B	C	Total
Masculino	5 (25%)	8 (40%)	7 (35%)	20
Feminino	8 (44,4%)	6 (33,3%)	4 (22,2%)	18
Total	13 (34,2%)	14 (36,8%)	11 (29%)	38