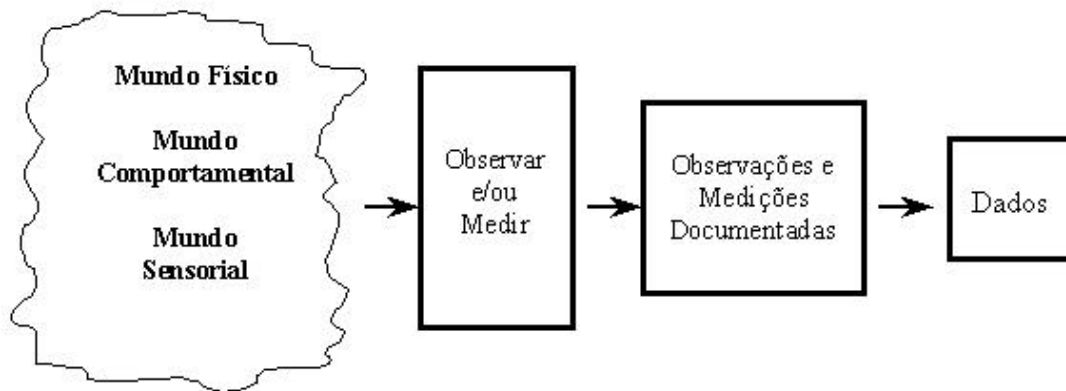


O Que São Dados?

Dados são observações documentadas ou resultados da medição. A disponibilidade dos dados oferece oportunidades para a obtenção de informações.

Os dados podem ser obtidos pela percepção através dos sentidos (por exemplo observação) ou pela execução de um processo de medição. A Figura abaixo ilustra as fontes dos dados.

Figura : Fontes dos Dados



Mundo Físico: Execução de uma tarefa
Medir: tempo gasto para realizar a tarefa
Instrumento de medição: cronômetro
Medida: minutos gasto na realização da tarefa

Mundo Comportamental: Realização de uma reunião
Medir: atitude das pessoas na reunião
Instrumento de medição: observação
Medida: atitude positiva, atitude negativa, atitude neutra

Mundo sensorial: Provar um alimento
Medir: aroma
Instrumento de medição: provadores
Medida: ruim, regular bom excelente

Tipos de Dados

Antes de iniciar a análise de uma base de dados, é preciso determinar corretamente que tipo de dados está disponível. Disso depende o tipo de análise a ser feito e a ferramenta a ser utilizada.

Existem muitas formas de classificar dados, dependendo do seu uso e da forma como os dados serão armazenados (em papel, em arquivo eletrônico, etc.). Vamos

apresentar algumas formas aqui, mas certamente o leitor poderá encontrar outras formas na literatura

Para exemplificar, considere um processo que utilize a base de dados cadastrais dos clientes de um Banco. A *unidade observacional* é o cliente. Para cada cliente, temos diversas características como Nome, Sexo, Estado Civil, Renda, etc. Cada uma dessas características é chamada tecnicamente de *variável*. Um cadastro, ou base de dados cadastrais é formado de casos ou registros (que no caso são os clientes) e variáveis. Para cada cliente temos o resultado das variáveis. Por exemplo, o primeiro cliente da base de dados tem o *nome* Fulano de Tal, o *Sexo* Masculino, o *Estado Civil* Solteiro, a *Renda* R\$2.500,00 por mês, etc.

Uma base de dados pode geralmente ser representada através de uma tabela, sendo que nas colunas estão as variáveis e nas linhas os casos.

Nome	Sexo	Estado Civil	Renda Mensal
João	M	C	1.200,00
Maria	F	S	1.350,00
Antonio	M	C	*

Uma base de dados é *univariada* quando contém apenas uma variável. *Bivariada* quando contém duas variáveis e *multivariada* quando contém três ou mais variáveis.

Dados no computador

Com respeito à inserção de dados em um software computacional, as variáveis em uma base de dados são classificadas tipicamente como *numéricas (N)*, *texto (T)* ou *data (D)*. Como exemplos de variáveis numéricas temos idade, renda, saldo da fatura, número de dependentes, etc. O sexo, estado civil, nome, rua, CEP, CIC, status do cliente são exemplos de variáveis alfanuméricas. A data de vencimento da conta e a data de nascimento são exemplos de variáveis do tipo *data*.

Variáveis numéricas são passíveis de serem manipuladas aritmeticamente. Embora a variável CEP seja representada numericamente, ela não é uma variável numérica. Podemos representar na base de dados a variável Sexo pelos símbolos 1 e 2, sendo 1=Masculino e 2=Feminino, mas mesmo assim a variável Sexo não é numérica. Não faz sentido, por exemplo, calcular a média dos sexos dos clientes, nem calcular a média dos CEP's!

Variáveis numéricas

Para variáveis numéricas, a ordenação e a magnitude são importantes. Variáveis numéricas podem ser classificadas como *discretas* ou *contínuas*. Variáveis discretas estão restritas a assumir apenas um número especificado de valores. Por exemplo, o número de dependentes de um cliente pode ser 0, 1, 2, 3, etc. Variáveis contínuas podem assumir, em princípio, qualquer valor em um determinado intervalo. O saldo da fatura pode ser qualquer valor entre R\$0,00 e R\$10.000,00, por exemplo.

O leitor deve estar atento ao fato de que uma variável contínua na maioria das vezes se apresenta como uma variável discreta em função da resolução do instrumento de medição. Por exemplo, a altura de uma pessoa é uma variável intrinsecamente contínua pois a altura de uma pessoa pode assumir qualquer valor digamos entre 30 e 220 cm.

Porém, se estamos usando uma fita métrica com escalas em centímetros, só podemos registrar alturas com saltos de um centímetro e o resultado da medição se apresenta sob a forma de uma variável discreta. Nesses casos, geralmente tratamos a variável como sendo contínua. Considere um outro exemplo: a idade de uma pessoa é uma variável intrinsecamente contínua. Porém podemos estar interessados em registrar apenas o número de anos completos de uma pessoa. Nesse caso, os valores medidos são: 0, 1, 2, 3 e assim por diante e podemos tratar essa variável como sendo discreta.

Variável texto

Variável texto é tipicamente utilizada para variáveis *classificatórias* ou *categóricas*. Por exemplo, o sexo de uma pessoa é uma variável classificatória, bem como o CEP, RG, etc. Um produto pode ser classificado como *Defeituoso* (ou *Não conforme*) ou *Não defeituoso* (*Conforme*). Um projeto pode estar *Atrasado*(A) ou *Não atrasado* (NA). Esses são alguns exemplos de variáveis classificatórias.

Uma variável classificatória pode ser *ordinal* quando as categorias possíveis para a variável apresentam uma ordenação. Quando a classificação não é ordenável a variável é *nominal*. Por exemplo, se representamos o sexo por 1 e 2 para Masculino e Feminino respectivamente, o valor 1 não é menor que 2, ou seja, a ordem não é importante. Por outro lado, se classificamos os clientes como 0, 1, 2, ..., 9 de acordo com o risco que ele proporciona ao banco (sendo 0 o de menor risco), então a ordem é importante e essa variável é ordinal.

Essas classificações são importantes do ponto de vista da análise estatística. A ferramenta estatística mais adequada a ser utilizada na análise dos dados depende do tipo de variável.

Variáveis contínuas podem ser *categorizadas*. Por exemplo, a renda de um cliente é uma variável contínua, mas podemos dividir a renda em faixas e criar uma variável categórica Faixa de renda em que, por exemplo, a faixa 1 corresponde a rendas de R\$1,00 a R\$500,00, faixa 2 de R\$501,00 a R\$1200,00, etc.

Variáveis numéricas também são chamadas de *dados quantitativos* e variáveis classificatórias são *dados qualitativos*.

Dados quantitativos são geralmente mais informativos. Por exemplo, saber que uma pessoa tem renda na faixa 2 não permite saber se sua renda está próxima ao limite inferior, ao superior, ou no meio. Se temos a renda da pessoa, podemos classificá-la na faixa de renda correspondente.

Dados de um Processo de Medição

Um processo de medição é um método que estabelece a relação entre a propriedade de interesse e um conjunto de classificações ou escala. Os sistemas de medição consistem em unidades-padrão para medição (isto é, pés metros para comprimento, horas para tempo, cedo ou tarde para tempo, etc) e procedimentos para a produção de valores nos termos destas unidades de medida. Os procedimentos podem incluir instrumentos físicos como um cromatógrafo gasoso, um velocímetro ou um calibrador. Os procedimentos também podem ser determinações subjetivas feitas pelas pessoas utilizando-se um ou mais dos sentidos humanos, que são registrados numa escala.

As entradas para o processo de medição são amostras dos materiais ou produtos, a realização de uma atividade, etc. O resultado de um processo de medição é denominado “medição” e será apresentado na forma de algum tipo de dado.

Um sistema simplificado de classificação de dados

Na indústria utiliza-se um sistema de classificação de dados simplificador. A medida de uma característica de um processo, produto ou serviço se apresenta sob a forma de uma variável:

contínua, também chamada simplesmente de *variável* (peso, altura, tempo de ciclo)

contagem (número de acidentes, número de defeitos, número de ligações perdidas em um Call Center)

classificatória (conforme/não conforme, defeituoso/não defeituoso, atrasado/não atrasado).

With count data, its important to know the **area of opportunity**: the boundaries that define when you’ll start and stop the count. This can be a given time period, a fixed area of product, or a certain number of units. The occurrences must be relatively rare compared to a relatively large area of opportunity

A Tabela abaixo lista alguns exemplos de tipos de dados de um processo de medição sob essa perspectiva.

Tabela : Exemplos de Tipos de Dados dos Processos de Medição

Dados Contínuos	Dimensão de uma peça mecanizada Viscosidade de um líquido Peso de uma pessoa Tempo para completar uma tarefa Custos de produção
Dados de Contagem	Número de partículas pretas numa amostra Número de erros numa página Número de acidentes num mês Número de vezes que uma tarefa foi concluída Contagem do estoque
Dados de Classificação	Página com erro ou sem erro Classificação de um produto (Tipo A ou Tipo B)

As medições podem ser transformadas de um tipo de dado para outro, entretanto geralmente isso só pode ser feito em uma direção. Os dados contínuos ou de contagem geralmente podem ser convertidos para dados de classificação ou hierarquização, mas não na direção inversa. Por exemplo, as medições dos horários de chegada das entregas poderiam ser registradas de acordo com o minuto mais próximo utilizando-se um relógio. Posteriormente estes dados contínuos poderiam ser convertidos em dados de classificação utilizando-se as definições “no horário”, “atrasado”. Porém se os dados originais foram registrados como uma classificação, não será possível transformar posteriormente estes dados em horários de chegada (dados contínuos).

Atividade – Classificação de variáveis

1. Classifique cada uma das seguintes variáveis de acordo com os dois sistemas de classificação apresentados nas colunas 2 e 3

Variáveis	Numérica (N) Texto (T) Data (D)	Contínua Contagem Classificatória (categórica)
Número de telefones por domicílio		
Número de chamadas de longo distância realizadas por mês		
Duração de cada chamada de longa distância		
Cor do telefone utilizado com mais frequência		
Se existe uma linha conectada ao modem na residência		
Quantia gasta com livros		
Tempo gasto na livraria por mês		
Se é filiado a algum Partido Político		
Caso seja, a que Partido Político é filiado		
Satisfação com um determinado produto		
Classificação de hotéis		
Quantidade de calorias de um produto alimentício		
Número de bolhas em uma garrafa de vidro		
Tempo médio de espera para se atendido em um Call Center		
Número de atendentes em um Call Center		
Número de ligações perdidas em um Call Center		
Motivos para ligações perdidas em um Call Center		
Fontes de consumo de água em uma residência		
Consumo de água em uma residência		