Roteiro de aulas - Classificação de variáveis e tratamento de dados

Disciplina: *Probabilidade e Estatística*

Professor: André Oliveira

01 de setembro de 2025

Classificação das variáveis e Tratamento de dados

A classificação adequada de uma variável é fundamental, pois orienta a seleção da metodologia estatística mais apropriada para o tratamento de dados e apresentação de dados de uma pesquisa. A classificação permite, de forma precisa, determinar o tipo de gráfico mais adequado e a metodologia inferencial mais indicada para a análise daquela variável em estudo. Conhecer a classificação da variável define o tipo de representação gráfica (histograma, *boxplot*, gráfico de barras ou de setores) e o método (modelo) de análise da mesma. Uma **variável** corresponde a uma característica mensurável sob estudo, que assume diferentes valores para diferentes elementos. Em contraposição a uma variável, o valor de uma constante é fixo. É uma característica a ser observada em cada elemento da amostra ou uma população. É importante ressaltar que uma variável não são necessariamente numéricos, uma vez que podem dizer respeito a atributos qualitativos observados na população. Uma **população** é composta por todos os indivíduos de interesse do pesquisador, enquanto uma amostra consiste em um **subconjunto** dessa população. Entretanto para que a amostra seja **representativa**, ela deve conter as características da população de interesse no estudo.

Observação ou Medição:

O valor de uma variável para um elemento específico é chamado de observação ou medição.

Conjunto de Dados

Um conjunto de dados é uma compilação de observações sobre uma ou mais variáveis e que pode ser chamado de banco de dados.

Tipos de variáveis:

Uma variável pode ser classificada como **quantitativa** ou **qualitativa.**

Variável Quantitativa:

Uma variável que pode ser mensurada numericamente é chamada de variável quantitativa. Os dados coletados de uma variável quantitativa são chamados de dados **quantitativos.**

Algumas variáveis (tais como número de *SSDs* vendidos por mês, tempo até o *notebook* apresentar defeito, quantidade de memória *RAM*) **podem** ser mensuradas numericamente, enquanto outras (tais como marcas de *SSDs*, tipos de sistema operacional, marca de *notebooks*) **não** podem ser expressas numericamente. O primeiro grupo de variévis são exemplos de variáveis **quantitativa**, já o segundo grupo de variáveis são variáveis **qualitativa**. Assim uma variável que pode ser mensurada numericamente é chamada de variável **quantitativa.** Os dados coletados sobre uma variável quantitativa são chamados de dados quantitativos.

- **Variáveis Discretas:** Os valores que determinada variável quantitativa pode assumir podem ser contáveis. Por exemplo, podemos contar o número de *Desktops* de uma empresa ou contar quantas pessoas trabalham no setor de recursos humanos de uma empresa de tecnologia. Uma variável que assume valores contáveis (inteiros) é chamada **variável discreta**. Observe que não existem **valores intermediários** possíveis entre valores consecutivos de uma **variável discreta.**
- **Variáveis Contínuas:** Algumas variáveis não podem ser contadas, entretanto podem assumir um valor numérico do conjunto dos números reais. Tais variáveis são chamadas de **variáveis contínuas.** Exemplos são as temperaturas de memória *RAM* ao final de um dia de trabalho dos *notebooks* de uma empresa, tempo usado de casa até o trabalho de uma pessoa ou o percentual de ouro nos componentes de uma placa-mãe.

Variáveis Qualitativas ou Categóricas:

Classificação de dados em categorias específicas, como tipos de sistemas operacionais (Windows, Linux, macOS), tipos de usuários, versões de software ou configurações de hardware. Incluem-se também os métodos de organização, armazenamento e linguagens de propagação. As variáveis qualitativas são divididas em dois grupos.

- **Ordinais:** cargo na empresa com níveis (gerente, operador de máquinas, motorista), estágio da doença com níveis (inicial, intermediário, terminal) e grau de escolaridade com níveis (1^o, 2^o, 3^o graus) classificação de um filme com níveis (excelente, ótimo, bom, regular, fraco, ruim, péssimo), pressão sanguínea com níveis (baixa, normal e alta) ou seja, **tem hierarquia** entre os níveis da variável.

- **Nominais:** sexo biológico com níveis (masculino e feminino), cor de pele com níveis (branca, negra, pardo), tipo sanguíneo com níveis (A, B, AB, O) ou seja, **não tem hierarquia** entre os níveis da variável.

Cabe destacar que uma variável originalmente quantitativa pode ser convertida em categórica (vice-versa), conforme a necessidade de tratamento e preparação dos dados para o modelo de análise estatística ou para aprendizado de máquina. Abaixo um diagrama que mostra a classificacação de variáveis com exemplos.

Para mostrar o conceito de **variável resposta** e **variável preditora (ou explicativa)**, considere o seguinte exemplo (situação):

Um técnico deseja verificar se a **temperatura** do ambiente influencia o desempenho da memória *RAM* dos Desktops.

A **variável resposta** é a característica que se mede, avalia ou observa no estudo, representando o fenômeno de interesse. Neste caso, a variável resposta é o **desempenho da memória RAM.** Já a **variável preditora (ou explicativa)** é a variável que pode **influenciar ou explicar** a variável resposta. No exemplo, a variável candidata a **variável preditora** é a **temperatura do ambiente**, selecionada pelo técnico como possível **fator** que afeta o desempenho da **memória *RAM***.

Em resumo, a **variável resposta** é o fator (variável) que se mede ou foi avaliado, enquanto a **variável preditora** é a é a candidada a influenciar ou explicar a variável resposta.

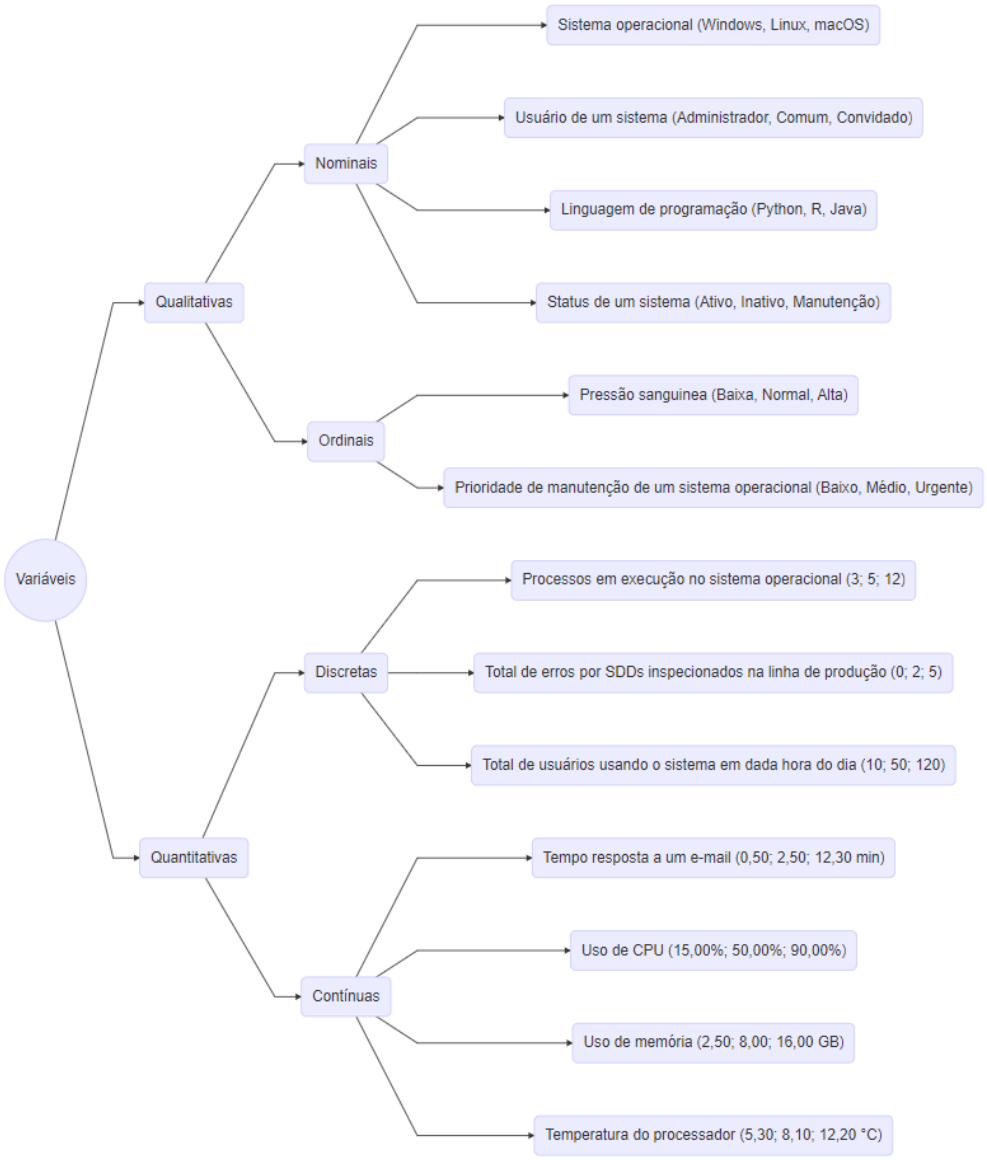

Figura 1: Diagrama de classificação de variáveis com exemplos

Tratamento de dados

Primeiramente há que definir o que são **dados** e o que são **informações.** Os dados são valores observados de uma variável (qualitativa ou quantitativa) alocados em uma planilha, em um banco de dados, servidores, SSDs ou até mesmo em um caderno por meio de anotações. Ao organizar estes **dados brutos** em um gráfico, em uma tabela passa a ser chamada de **informação** que será útil para tomada de decisões como foi descrito na primeira aula.

Para ilustrar o que são **dados** e o que são **informações** considere a situação. Paulo assinou um contrato com a empresa $NETVIX$ que oferece planos de internet de 60 $Megabyte$ (MB). Usando o aplicativo $FAST$ e com a finalidade de analisar a velocidade real da entrega prevista no contrato ele mensurou a velocidade após a instalação do modem por 79 dias consecutivos. Os **dados brutos**, em $Megabyte$ (MB), estão abaixo.

[51, 59, 56, 50, 57, 54, 51, 54, 60, 59, 52, 61, 53, 60, 57, 62, 53, 60, 52, 60, 62, 53, 51, 53, 56, 57, 54, 53, 58, 52, 55, 56, 63, 56, 52, 60, 61, 59, 51, 63, 57, 61, 61, 57, 54, 54, 51, 53, 57, 58, 63, 51, 51, 58, 53, 52, 52, 59, 52, 57, 61, 53, 56, 59, 51, 58, 61, 59, 63, 53, 62, 63, 51, 61, 63, 58, 50, 50, 63]

Colocando os dados da amostra em ordem crescente (ROL) tem a disposição abaixo.

[50, 50, 50, 51, 51, 51, 51, 51, 51, 51, 51, 51, 52, 52, 52, 52, 52, 52, 52, 53, 53, 53, 53, 53, 53, 53, 53, 53, 54, 54, 54, 54, 54, 55, 56, 56, 56, 56, 56, 57, 57, 57, 57, 57, 57, 57, 58, 58, 58, 58, 58, 59, 59, 59, 59, 59, 59, 60, 60, 60, 60, 60, 61, 61, 61, 61, 61, 61, 61, 62, 62, 62, 63, 63, 63, 63, 63, 63, 63]

Contando o número de vezes em que cada observação ocorreu nos **dados bruto** (f_i) é possível construir a **Tabela 1** de frequências e um **Figura 2** de distribuição, em que é possível identificar em nível *descritivo* as **informações** e analisar se a velocidade entregue é próxima da prevista em contrato.

Tabela 1: Distribuição da quantidade de Megabyte (MB) da amostra

Megabytes (MB)	fi
50	3
51	9
52	7
53	9
54	5
55	1
56	5
57	7
58	5
59	6
60	5
61	7
62	3
63	7

Fonte: Elaboração própria

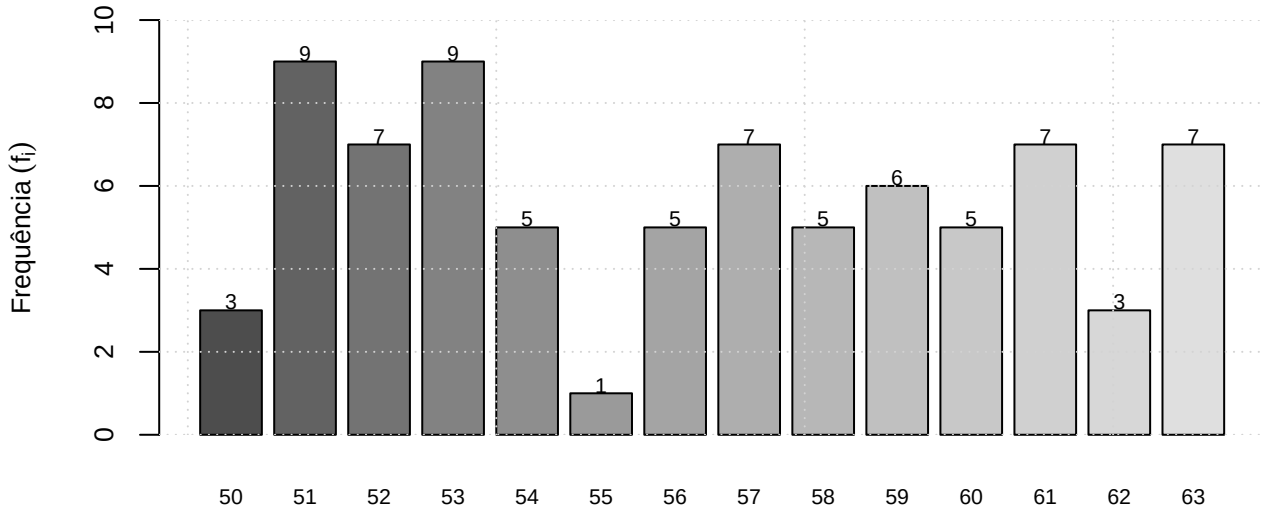

Figura 2: Quantidade de Megabyte (MB)

Fonte: Elaboração própria

O **Gráfico 2** e a **Tabela 1** nos mostram as mesmas ***informações*** sobre a distribuição a amostra. A partir da análise de ambos, observa-se que 57 das amostras coletadas estão abaixo de 60 MB, ou seja 72,15%, evidenciando assim que a velocidade de entrega está **abaixo** do contratado, e esta **informação** não estava clara (evidante) quando se olhava

(analisava) os dados na forma bruta. De forma geral dados brutos podem ser tratados e nos fornece informações (evidências) para tomada de decisão, neste caso será útil para o consumidor justificar que existem indícios de que o serviço contratado não está dentro do previsto em contrato. Claro que são evidências, pois o estudo foi com uma amostra e não com a população toda. A conclusão pode ser pode ser fundamentada por teste de hipóteses (TH) para uma média.

Agora considere que a empresa $TECINFO$ especializada em vendas de *notebooks* fez o balanco das vendas em 8 meses. Os dados de vendas estão estratificados pelas variáveis valor (em unidades de milhar), quantidade de memória RAM e *Marcas* como mostra a ***Tabela 2***.

Tabela 2: Dados da quantidade de vendas de notebooks

RAM	Marca	Valor
2 Gigabyte	Dell	5
2 Gigabyte	Avell	5
2 Gigabyte	ASUS	9
2 Gigabyte	Dell	7
2 Gigabyte	Avell	8
2 Gigabyte	ASUS	5
2 Gigabyte	Dell	5
2 Gigabyte	Avell	10
4 Gigabyte	ASUS	6
4 Gigabyte	Dell	7
4 Gigabyte	Avell	7
4 Gigabyte	ASUS	7
4 Gigabyte	Dell	10
4 Gigabyte	Avell	5
4 Gigabyte	ASUS	8
4 Gigabyte	Dell	5
8 Gigabyte	Avell	15
8 Gigabyte	ASUS	15
8 Gigabyte	Dell	20
8 Gigabyte	Avell	22
8 Gigabyte	ASUS	18
8 Gigabyte	Dell	17
8 Gigabyte	Avell	19
8 Gigabyte	ASUS	25

Fonte: Elaboração própria

A **Tabela 3** mostra uma distribuição das frequências (f_i) de vendas estratificada pelas variáveis marca e quantidade de memória *RAM*.

Tabela 3: Quantidade de vendas de notebooks estratificada pelas variáveis marca e quantidade de memória RAM

	ASUS	Avell	Dell
2 Gigabyte	2	3	3
4 Gigabyte	3	2	3
8 Gigabyte	3	3	2

Fonte: Elaboração própria

O **Gráfico 3** mostra uma distribuição das frequências (f_i) de vendas estratificada pelas **variáveis** marca e quantidade de memória *RAM*.

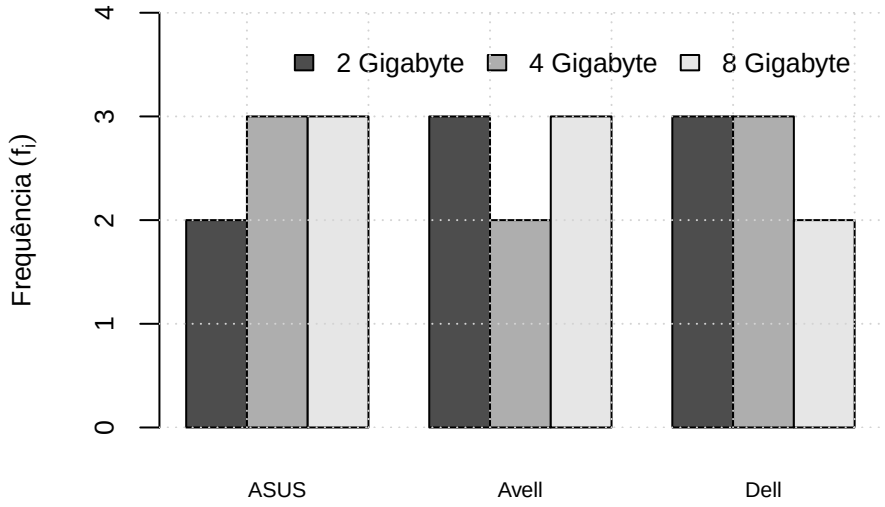

Figura 3: Distribuição de frequências de vendas (marca vs quantidade de memória RAM)

Fonte: Elaboração própria

Na **Tabela 3** e no **Gráfico 3** os dados foram apresentados sobre uma nova perpectiva em relação a **Tabela 2**, esta outra forma de apresentar possibilita novas informações disponíveis. O processamento de dados não se dá apenas por tabelas e gráficos, mas também por estatísticas descritivas (médias, variancias, desvios-padrão etc), taxas percentuais, teste de hipóteses (TH), métodos estatísticos, modelos probabilisticos que serão estudados nas próximas unidades deste curso.