

Relatório 5 - Prática: Estatística p/ Aprendizado de Máquina (I)

Beatriz Almeida Felício

Descrição da atividade

Tópicos Estatísticos e Ferramentas de Visualização de Dados

Tipos de Dados

1. **Numérico**

Dados numéricos são aqueles que podem ser quantificados e tratados matematicamente. Dividem-se em:

- **Discreto:** Valores que são contáveis, como número de filhos ou o resultado de um dado (1, 2, 3, etc.).
- **Contínuo:** Valores dentro de um intervalo infinito, como altura ou temperatura. Podem assumir qualquer valor em uma escala contínua.

2. **Categórico**

Dados categóricos representam categorias ou classificações e não possuem significado matemático. Exemplo: cores de carros (vermelho, azul, verde).

3. **Ordinal**

Dados ordinais misturam características de categóricos e numéricos, apresentando uma ordem específica. Por exemplo, níveis de escolaridade (fundamental, médio, superior). Embora possam ser ordenados, as diferenças entre os valores não são necessariamente consistentes.

Média, Mediana e Moda

- **Média:** É a soma de todos os valores dividida pelo número total de observações. Representa o valor típico, mas pode ser sensível a valores extremos (outliers). Fórmula:

$$\bar{x} = \frac{\sum x_i}{n}$$

- **Mediana:** É o valor central de um conjunto de dados organizados em ordem crescente. É menos influenciada por outliers, sendo uma medida de posição.
- **Moda:** O valor que aparece com maior frequência em um conjunto de dados. É útil para dados categóricos ou quando se deseja identificar tendências de repetição.

Variância e Desvio Padrão

- **Variância:** Mede o grau de dispersão dos dados em relação à média. É a média dos quadrados das diferenças entre cada valor e a média.
Fórmula:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

- **Desvio Padrão:** Representa a raiz quadrada da variância, indicando o quão espalhados os valores estão em torno da média. É mais intuitivo porque usa a mesma unidade dos dados.

$$s = \sqrt{s^2}$$

População vs Amostra

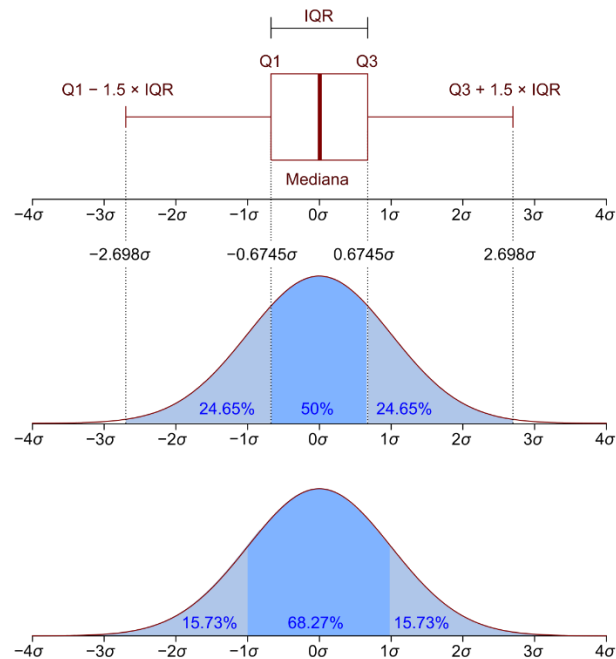
- **População:** Conjunto completo de elementos que se deseja estudar. Exemplo: altura de todas as pessoas de um país.
- **Amostra:** Subconjunto da população, usado para inferir características da população inteira, especialmente em casos onde medir todos os elementos seria inviável.



Função de Densidade e Massa de Probabilidade

1. **Função Densidade de Probabilidade (PDF)**
Descreve a probabilidade de uma variável contínua assumir valores em um intervalo. A área sob a curva da PDF entre dois pontos representa a probabilidade.

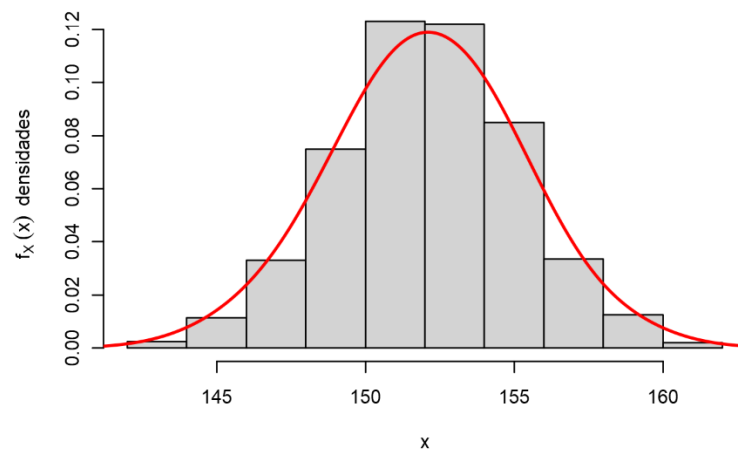
2. **Função de Massa de Probabilidade (PMF)**
 Utilizada para variáveis discretas, atribui probabilidades específicas a valores individuais. Por exemplo, a probabilidade de um dado mostrar um número específico.



Distribuições Estatísticas

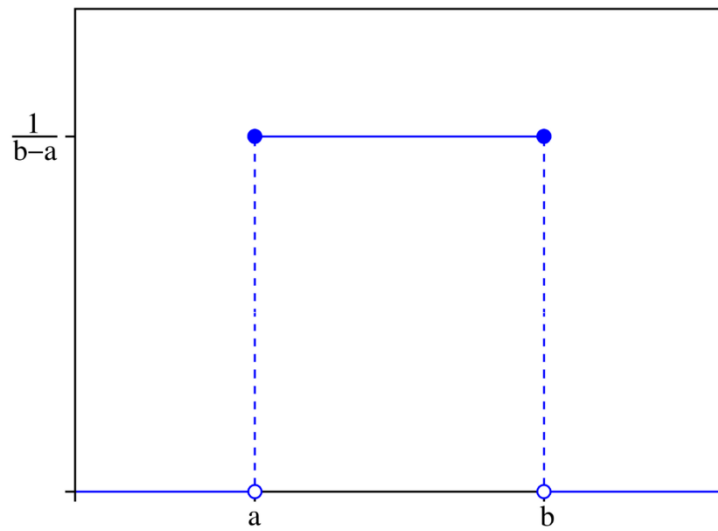
1. Normal

Também chamada de distribuição gaussiana, é simétrica e em formato de sino. Caracterizada por média (μ) e desvio padrão (σ). Aproximadamente 68% dos dados estão dentro de 1 desvio padrão da média.



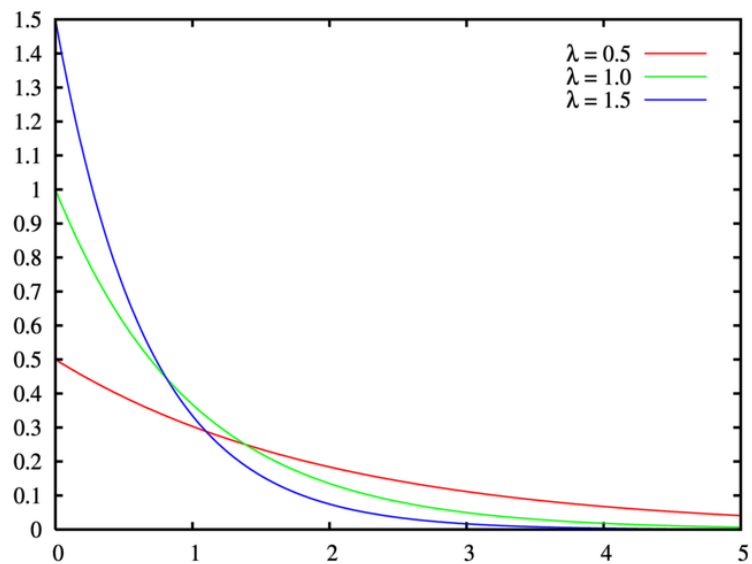
2. Uniforme

Todos os valores dentro de um intervalo têm a mesma probabilidade. Exemplo: rolar um dado justo.



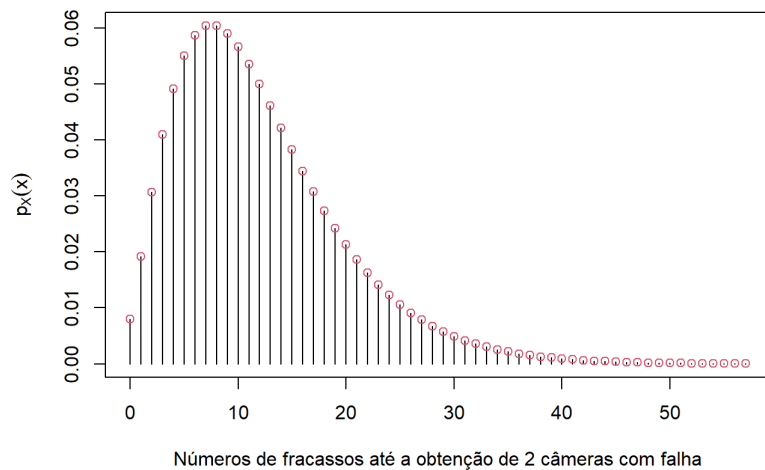
3. Exponencial

Modela o tempo entre eventos em um processo de Poisson, como o intervalo entre chegadas em um ponto de ônibus.



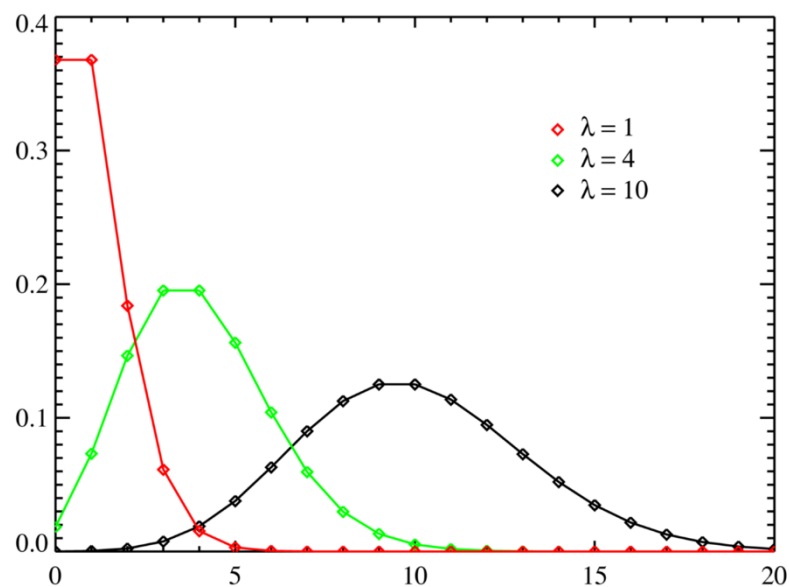
4. Binomial

Descreve o número de sucessos em N experimentos independentes, com uma probabilidade fixa pp de sucesso em cada um.



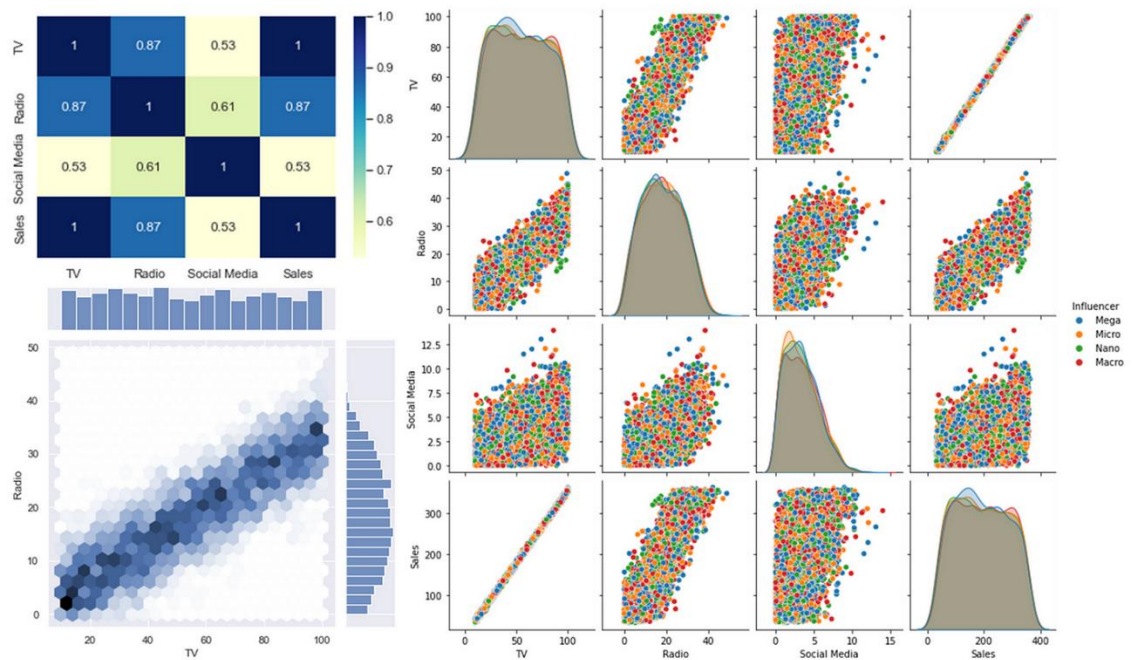
5. Poisson

Modela o número de eventos que ocorrem em um intervalo fixo de tempo ou espaço, dado uma taxa média de eventos.



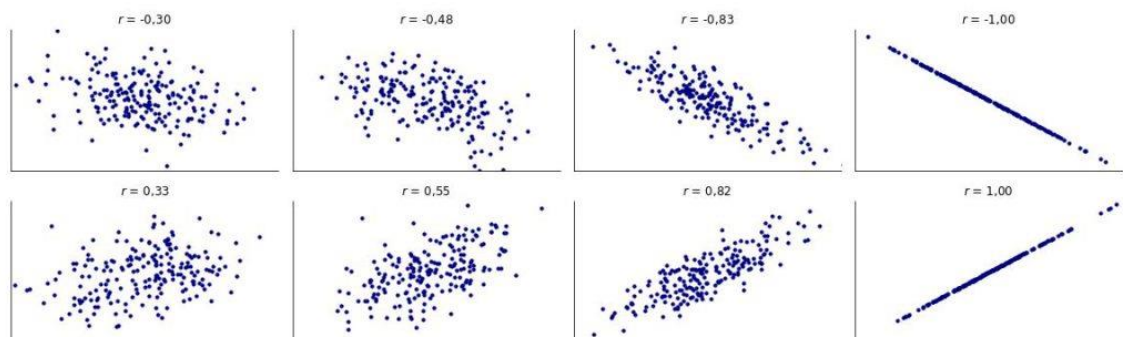
Percentis e Momentos

- **Percentis:** Dividem os dados em 100 partes iguais. Por exemplo, o 25º percentil representa o valor abaixo do qual 25% dos dados estão localizados.
- **Momentos:**
 1. **Primeiro Momento:** A média.
 2. **Segundo Momento:** Variância.
 3. **Terceiro Momento:** Assimetria da distribuição.
 4. **Quarto Momento:** Curtose, ou achatamento da distribuição.



Covariância e Correlação

- **Covariância:** Mede como duas variáveis variam juntas. Se positiva, indica que ambas aumentam juntas; se negativa, uma aumenta enquanto a outra diminui.
- **Correlação:** Normaliza a covariância, resultando em um valor entre -1 e 1. Correlação próxima a 1 ou -1 indica forte relação linear; 0 indica ausência de relação.



Probabilidade Condicional e Teorema de Bayes

1. Probabilidade

A probabilidade de um evento AA ocorrer dado que outro evento BB já ocorreu.
Fórmula:

Condicional

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(A)P(B|A)$$

2. **Teorema de Bayes**
Atualiza a probabilidade de uma hipótese com base em novas evidências.
Fórmula:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Conclusão

Os conceitos estatísticos apresentados constituem a base fundamental para a análise de dados, fornecendo as ferramentas e o entendimento necessários para explorar, interpretar e comunicar informações de maneira eficaz. A classificação dos tipos de dados é um passo inicial crucial, pois orienta a escolha das técnicas de análise mais adequadas. Medidas descritivas como média, mediana e moda oferecem resumos concisos dos dados, enquanto métricas de dispersão, como variância e desvio padrão, revelam a distribuição e a consistência dos valores. Além disso, a compreensão de distribuições de probabilidade e sua aplicação em modelagem estatística são indispensáveis para realizar inferências robustas e fundamentadas.

A visualização de dados, facilitada por ferramentas como Matplotlib e Seaborn, desempenha um papel essencial ao transformar resultados complexos em representações intuitivas e acessíveis, promovendo uma comunicação clara dos insights. Ao dominar esses conceitos e práticas, é possível não apenas realizar análises precisas, mas também aplicar esses conhecimentos em diversos cenários, desde a pesquisa científica até a tomada de decisões estratégicas em negócios, impulsionando inovação e eficiência.

Referências

11 **Funções densidade de probabilidades.** Disponível em: <https://www.lampada.uerj.br/arquivosdb/_book/variaveiscontinuas.html>. Acesso em: 4 dez. 2024.

ASTH, R. **C. Variância e desvio padrão.** Disponível em: <<https://www.todamateria.com.br/variancia-e-desvio-padrao/>>. Acesso em: 3 dez. 2024.

Links cursos. Disponível em: <https://docs.google.com/document/d/1rAvJ0FiWZG5fgecj4VhHqDJcgQWh_O5CrmZCw812GU/edit?tab=t.0>. Acesso em: 3 dez. 2024.

REIS, T. **Teorema de Bayes: Entenda o que é e como calcular.** Disponível em: <<https://www.suno.com.br/artigos/teorema-de-bayes/>>. Acesso em: 5 dez. 2024.