Resumo Teórico Aula 02
Data Science Experience

Matheus H. P. Pacheco

May 6, 2025

Contents

1 Introdução à Estatística

A *estatística* é a ciência que fornece ferramentas para lidar com a incerteza inerente aos dados. Sua origem remonta ao século XVII, com James Bernoulli e Abraham de Moivre, que estabeleceram fundamentos da teoria da probabilidade. Em um contexto moderno, a estatística se divide em dois grandes ramos:

1.1 População, Amostra e Parâmetros

- **População (N)**: conjunto completo de elementos de interesse (ex.: todos os clientes de uma empresa).
- **Amostra (n)**: subconjunto da população selecionado para análise. Deve ser *representativo* para inferir sobre a população.
- **Parâmetros**: características da população, como média μ e variância σ^2, geralmente *desconhecidos*.
- **Estatísticas**: estimativas calculadas na amostra, como média amostral $\bar{x}$ e variância amostral s^2.

1.2 Estatística Descritiva

Consiste em resumir e visualizar dados brutos para entender sua estrutura:

- *Medidas de tendência central*: média, mediana, moda.
- *Medidas de dispersão*: variância, desvio padrão, coeficiente de variação.
- *Representações gráficas*: histogramas, boxplots, diagramas de dispersão.

Essas ferramentas auxiliam na detecção de *outliers*, assimetrias e padrões iniciais.

1.3 Estatística Inferencial

Baseia-se em modelos probabilísticos e na teoria da amostragem para extrapolar conclusões da amostra para a população:

- **Estimativa pontual** e **intervalar** (intervalos de confiança).
- **Testes de hipótese** para verificar suposições sobre parâmetros (ex.: média, proporção).
- *Erro amostral* e *nível de confiança* $(1 - \alpha)$.

A inferência utiliza o *Teorema do Limite Central* e a *Lei dos Grandes Números* para garantir a normalidade assintótica de estimadores.

1.4 Fluxo de Trabalho em Ciência de Dados

No pipeline de Data Science, a estatística fundamenta as etapas iniciais:

1. **Coleta de Dados**: definição da amostragem e planos de experimento.
2. **Limpeza e Pré-processamento**: identificação de valores ausentes e anomalias.

3. **Análise Exploratória (EDA)**: aplicação de estatística descritiva para formular hipóteses.
4. **Modelagem**: escolha de modelos estatísticos (regressão, séries temporais, classificação).
5. **Validação e Interpretação**: uso de testes de hipótese e intervalos de confiança para avaliar desempenho.

1.5 Papel da Probabilidade

A probabilidade é a base teórica da estatística:

$$P(A) = \frac{\text{casos favoráveis}}{\text{casos possíveis}}, \quad P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

definindo modelos para variáveis aleatórias e distribuindo crenças sobre resultados incertos.

Em seguida, aprofundaremos nos conceitos de variáveis aleatórias, esperança e variância, essenciais para todas as análises estatísticas.

2 Tipos de Variáveis

As variáveis são atributos medidos em cada observação e determinam quais métodos estatísticos e técnicas de modelagem podem ser aplicados.

2.1 Escalas de Mensuração

Nominal: categorias sem ordem intrínseca. Ex.: gênero, cor dos olhos. Não permite operações aritméticas.

Ordinal: categorias com ordem, mas sem distância fixa entre níveis. Ex.: níveis de escolaridade (fundamental, médio, superior).

Intervalar: valores numéricos com diferença significativa, mas sem ponto zero absoluto. Ex.: temperatura em °C.

Razão (Ratio): como intervalar, porém com zero absoluto significativo. Ex.: peso, altura, renda.

Cada escala impõe restrições sobre estatísticas válidas (ex.: média aritmética só faz sentido em escalas intervalar e razão).

2.2 Variáveis Qualitativas

- **Nominais:** apenas identificação de categoria.
  - Representação em ciência de dados: codificação *one-hot*, *label encoding*.
  - Análise: frequências absolutas e relativas, tabelas de contingência, teste qui–quadrado de independência.
- **Ordinais:** mantêm ordenação, sem garantias de intervalos iguais.
  - Representação: *ordinal encoding*, escalonamento de scores.
  - Análise: mediana, percentis; testes não paramétricos como Mann–Whitney, Kruskal–Wallis.

2.3 Variáveis Quantitativas

Discretas: assumem valores inteiros contáveis (ex.: número de vendas).

- Modelos frequentes: distribuição Binomial, Poisson.
- Estatísticas: média, variância, teste z e t (se atendidos pressupostos).

Contínuas: podem assumir qualquer valor em um intervalo (ex.: altura, tempo de resposta).

- Modelos frequentes: Normal, Exponencial, Gamma.
- Estatísticas: média, variância, desvio padrão; testes de normalidade (Shapiro–Wilk, Kolmogorov–Smirnov).

2.4 Implicações para Ciência de Dados

- **Pré-processamento:** transformação de variáveis qualitativas em *features* numéricas, normalização/ padronização de variáveis contínuas.
- **Escolha de Modelos:** regressão linear requer variáveis numéricas e pressupostos de normalidade e homocedasticidade; árvores de decisão lidam naturalmente com misturas de qualitativas e quantitativas.
- **Avaliação de Performance:** métricas variam conforme o tipo de variável (RMSE para contínuas, acurácia/F1 para categóricas).

“‘latex

3 Medidas de Tendência Central

As medidas de tendência central buscam representar um valor típico de uma variável, seja em uma amostra ou em uma população.

3.1 3.1 Esperança Matemática e Média Aritmética

Para uma variável aleatória contínua ou discreta, o valor esperado (esperança) define o centro de massa da distribuição:

$$E[X] = \begin{cases} \sum_{i}^{\infty} x_i \, P(X = x_i), & X \text{ discreta}, \\ \int_{-\infty} x \, f_X(x) \, dx, & X \text{ contínua}, \end{cases}$$

Em amostras, estima-se $E[X]$ pela média amostral:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Propriedades:

- *Não viesado*: $E[\bar{X}] = E[X] = \mu$.
- *Variância da média*: $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$.
- *Linearidade do operador*: $E[aX + b] = aE[X] + b$.

3.2 3.2 Minimização do Erro Quadrático Médio

A média amostral surge como minimizadora do Erro Quadrático Médio (EQM):

$$\text{EQM}(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2.$$

Para encontrar o valor a que minimiza este critério, derivamos em relação a a e igualamos a zero:

$$\frac{d}{da}\text{EQM}(a) = -\frac{2}{n}\sum_{i=1}^{n}(x_i - a) \;=\; 0 \quad \Longrightarrow \quad a = \bar{x}.$$

3.3 3.3 Mediana

A mediana é definida como o valor m que minimiza a soma de desvios absolutos:

$$\min_{m}\sum_{i=1}^{n}|x_i - m|.$$

Para uma amostra ordenada $x_{(1)} \leq \cdots \leq x_{(n)}$,

$$\text{Mediana} = \begin{cases} x_{(\frac{n+1}{2})}, & n \text{ ímpar}, \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & n \text{ par}. \end{cases}$$

Robustez: A mediana é menos sensível a outliers que a média.

3.4 3.4 Moda

A moda é o valor que ocorre com maior frequência em um conjunto de dados. Para variáveis contínuas, define-se:

$$\hat{x}_{\text{mode}} = \arg\max_{x} f_X(x),$$

onde $f_X(x)$ é a densidade de probabilidade.

3.5 3.5 Outras Médias

- **Média Ponderada**:
$$\mu_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i},$$
usada quando diferentes observações têm importâncias distintas.

- **Média Geométrica**:
$$\left(\prod_{i=1}^{n} x_i\right)^{1/n},$$
adequada para taxas de crescimento e índices.

- **Média Harmônica**:
$$\frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}},$$
usada em médias de velocidades ou razões.

3.6 3.6 Relevância em Ciência de Dados

Em Data Science, as medidas de tendência central são fundamentais em:

- **Imputação de Dados**: média ou mediana para preencher valores ausentes.
- **Análise Exploratória (EDA)**: compreender o centro da distribuição de features.
- **Pré-processamento**: padronização (subtrair média e dividir por desvio) antes de modelagem.

‘‘‘

4 Medidas de Dispersão

As medidas de dispersão quantificam o espalhamento dos dados em relação à tendência central.

4.1 Variância

A variância mede o quadrado da distância média entre cada observação e o centro da distribuição.

Variância Populacional:

$$\sigma^2 = E\big[(X-\mu)^2\big] = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\,dx,$$

onde $\mu = E[X]$ e $f_X(x)$ é a densidade de X.

Variância Amostral:

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

com o denominador $n-1$ para garantir que $E[s^2] = \sigma^2$ (corrige viés).

4.2 Desvio Padrão

O desvio padrão é a raiz quadrada da variância, retornando à mesma unidade dos dados:

$$\sigma = \sqrt{\sigma^2}, \quad s = \sqrt{s^2}.$$

4.3 Coeficiente de Variação

O coeficiente de variação é uma medida relativa de dispersão:

$$CV = \frac{s}{\bar{x}} \times 100\%,$$

útil para comparar a variabilidade de conjuntos com médias diferentes.

4.4 Amplitude e Quartis

- **Amplitude (Range)**: diferença entre o maior e o menor valor: $R = x_{(n)} - x_{(1)}$.
- **Intervalo Interquartil (IQR)**: $IQR = Q_3 - Q_1$, onde Q_1 e Q_3 são os quartis primeiro e terceiro.
- **Boxplot**: representação gráfica do IQR, mediana e potenciais outliers.

4.5 Relevância em Ciência de Dados

- **Detecção de Outliers**: usar IQR e z-scores ($z_i = (x_i - \bar{x})/s$) para identificar observações atípicas.
- **Normalização e Padronização**: aplicação de z-score para algoritmos sensíveis à escala.
- **Análise de Dispersion em Features**: avaliar homocedasticidade e variância explicada em PCA.

5 Distribuições de Probabilidade

As distribuições de probabilidade descrevem como a probabilidade se distribui sobre os possíveis valores de uma variável aleatória, fornecendo a base para modelagem e inferência estatística.

5.1 Variáveis Aleatórias e Suas Funções

- **Variável Aleatória Discreta:** assume valores em um conjunto enumerável.

$$P(X = k), \quad \sum_{k} P(X = k) = 1.$$

- **Variável Aleatória Contínua:** assume valores em um intervalo contínuo.

$$f_X(x), \quad \int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1,$$

  onde $f_X(x)$ é a função densidade de probabilidade (FDP).

5.2 Principais Distribuições

Normal (Gaussiana)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big), \quad E[X] = \mu,\ \mathrm{Var}(X) = \sigma^2.$$

Usada em erro de medida, ruído em sinais, e como aproximação via Teorema do Limite Central.

Binomial

$$P(X = k) = \binom{n}{k} p^k\,(1-p)^{\,n-k}, \quad E[X] = np,\ \mathrm{Var}(X) = np(1-p).$$

Modela número de sucessos em n ensaios independentes com probabilidade p.

Poisson

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad E[X] = \lambda,\ \mathrm{Var}(X) = \lambda.$$

Aproxima binomial para n grande, p pequeno; aplicada em contagem de eventos raros.

Exponencial

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \quad E[X] = \frac{1}{\lambda},\ \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

Modela tempos de espera entre eventos de um processo de Poisson.

5.3 Separação Teórica

- **Função Massa vs. Densidade:** distingue-se soma (discreta) de integral (contínua).
- **Momentos:** $E[X^r] = \sum_k k^r P(X = k)$ ou $\int x^r f_X(x)\, dx$.
- **Função de Distribuição Acumulada (FDA):**

$$F_X(x) = P(X \leq x) = \begin{cases} \sum_{k \leq x} P(X = k), & \text{discreta,} \\ \int_{-\infty}^{x} f_X(t)\, dt, & \text{contínua.} \end{cases}$$

5.4 Noções de Aplicação em Ciência de Dados

- **Naive Bayes:** seleciona a distribuição de cada feature (Bernoulli, Multinomial, Gaussiana) para estimar $P(X \mid Y)$.
- **Inferência Bayesiana:** combina distribuição *a priori* e verossimilhança para obter a posteriori.
- **Modelos de Contagem:** Poisson e suas generalizações (Negativa Binomial) para dados de contagem.
- **Análise de Sobrevivência:** distribuições exponencial e Weibull para modelar tempos até um evento.
- **Teste de Aderência:** Kolmogorov–Smirnov e QQ-plots para verificar ajuste de dados a uma distribuição teórica.

6 Testes de Hipóteses

Os testes de hipótese são procedimentos formais para avaliar suposições sobre parâmetros populacionais com base em dados amostrais.

6.1 Elementos Fundamentais

- **Hipótese Nula (H_0):** afirmação inicial, geralmente de "sem efeito" ou "igualdade".
- **Hipótese Alternativa (H_1):** contrária a H_0, indica "efeito" ou "diferença".
- **Nível de Significância (α):** probabilidade máxima de rejeitar erroneamente H_0 (erro Tipo I).

- **Região Crítica:** valores de estatística de teste que levam à rejeição de H_0.
- **Estatística de Teste (T):** função dos dados que, sob H_0, possui distribuição conhecida.
- **p-valor:** $P(T \geq t_{\text{obs}} \mid H_0)$ (ou duas-faces, $P(|T| \geq |t_{\text{obs}}|)$).
- **Decisão:** rejeita-se H_0 se p-valor $< \alpha$; caso contrário, não se rejeita.

6.2 Erros em Testes

Tipo I: rejeitar H_0 sendo ela verdadeira. Probabilidade: α.

Tipo II: não rejeitar H_0 sendo H_1 verdadeira. Probabilidade: β.

Potência: $1 - \beta$, chance de detectar H_1 quando verdadeira.

6.3 Teste z para Uma Média (Conhecido)

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \ \sim \ N(0,1) \quad \text{sob } H_0 : \mu = \mu_0.$$

- Região crítica (unilateral à direita): $T > z_{1-\alpha}$.
- p-valor: $1 - \Phi(t_{\text{obs}})$, com Φ CDF da normal padrão.

6.4 Teste t para Uma Média (Desconhecido)

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \ \sim \ t_{\,n-1} \quad \text{sob } H_0 : \mu = \mu_0,$$

onde S^2 é a variância amostral. Usa-se $t_{1-\alpha,\,n-1}$ para definir a região crítica.

6.5 Teste Qui–Quadrado para Variância

$$T = \frac{(n-1)\,S^2}{\sigma_0^2} \ \sim \ \chi^2_{n-1} \quad \text{sob } H_0 : \sigma^2 = \sigma_0^2.$$

Rejeita-se H_0 se T estiver nos caudas da χ^2.

6.6 Teste de Proporções

Para proporção amostral $\hat{p}$ em n observações:

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \ \sim \ N(0,1) \quad (n \text{ grande, sob } H_0 : p = p_0).$$

6.7 p-Valor e Região Crítica

$$\text{p-valor} = P(T \geq t_{\text{obs}} \mid H_0) \quad \Longleftrightarrow \quad \text{p-valor} = \int_{t_{\text{obs}}}^{\infty} f_{T|H_0}(t)\,dt.$$

A região crítica é $T > t_{1-\alpha}$ (unilateral) ou $|T| > t_{1-\alpha/2}$ (bicaudal).

6.8 Relação com Intervalos de Confiança

Rejeitar $H_0 : \theta = \theta_0$ em um teste bilateral ao nível α equivale a θ_0 não pertencer ao intervalo de confiança de $100(1-\alpha)\%$ para θ.

6.9 Exemplo de Dedução

Para o teste z:

$$H_0 : \mu = \mu_0, \quad T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Sabemos $E[\bar{X}] = \mu_0$, $\mathrm{Var}(\bar{X}) = \sigma^2/n$, logo

$$E[T] = 0, \quad \mathrm{Var}(T) = 1.$$

Assim $T \sim N(0,1)$ e podemos determinar $z_{1-\alpha}$ tal que

$$P(T > z_{1-\alpha}) = \alpha.$$

6.10 Notas de Aplicação em Ciência de Dados

- Avaliar diferença de métricas (ex.: acurácia, tempo médio) entre modelos.
- Testar melhora significativa após ajuste de hiperparâmetros.
- Verificar independência de features com teste qui–quadrado antes de usar modelos paramétricos.

7 Correlação e Regressão Linear

A correlação e a regressão linear são ferramentas estatísticas para quantificar e modelar relações entre duas variáveis.

7.1 Covariância e Correlação de Pearson

A *covariância* entre X e Y mede como as variáveis variam conjuntamente:

$$\mathrm{Cov}(X,Y) \;=\; E\big[(X - E[X])(Y - E[Y])\big] \;=\; \iint (x - \mu_X)(y - \mu_Y)\, f_{X,Y}(x,y)\, dx\, dy.$$

Na prática, estima-se pela amostra:

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

A *correlação de Pearson* normaliza essa covariância para o intervalo $[-1, 1]$:

$$r \;=\; \frac{\mathrm{Cov}(X,Y)}{s_X\, s_Y},$$

onde s_X e s_Y são os desvios-padrão amostrais. - $r = 1$: correlação linear positiva perfeita. - $r = -1$: correlação linear negativa perfeita. - $r = 0$: ausência de relação linear.

7.2 Regressão Linear Simples

O modelo ajusta uma reta

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

minimizando o *Erro Quadrático* dos resíduos:

$$\mathrm{SSE}(\beta_0, \beta_1) = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2.$$

Derivando em relação a β_0 e β_1 e igualando a zero, obtemos o sistema

$$\frac{\partial \mathrm{SSE}}{\partial \beta_1} = 0, \quad \frac{\partial \mathrm{SSE}}{\partial \beta_0} = 0,$$

cuja solução é

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \, \bar{x}.$$

7.3 Noções em Ciência de Dados

- **Seleção de Features:** usar correlação para identificar multicolinearidade.
- **Interpretação de $\hat{\beta}_1$:** efeito médio de uma unidade de variação em X sobre Y.
- **Análise de Resíduos:** checar normalidade e homocedasticidade dos ε_i.
- **Qualidade de Ajuste:** índice $R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$ e teste F.

8 Intervalos de Confiança

Os intervalos de confiança (IC) fornecem um intervalo plausível para um parâmetro populacional com um grau de confiança $1 - \alpha$. Em vez de estimar apenas um valor pontual, o IC expressa a incerteza da estimativa.

8.1 Princípio Geral

Seja $\hat{\theta}$ um estimador pontual de θ (ex.: $\bar{X}$ para a média). Suponha que, para amostras grandes, a distribuição amostral de $\hat{\theta}$ seja aproximadamente normal:

$$\hat{\theta} \;\sim\; N\big(\theta,\ \mathrm{Var}(\hat{\theta})\big).$$

Então

$$\frac{\hat{\theta} - \theta}{\sqrt{\mathrm{Var}(\hat{\theta})}} \;\approx\; N(0, 1).$$

Para um nível de confiança $1 - \alpha$, temos

$$P\Big(-z_{\alpha/2} \le \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \le z_{\alpha/2}\Big) = 1 - \alpha,$$

onde $z_{\alpha/2}$ é o quantil padrão da normal.

8.2 Intervalo para a Média com σ Conhecido

Neste caso, $\mathrm{Var}(\bar{X}) = \sigma^2/n$, daí

$$IC_{1-\alpha}(\mu) : \bar{X} \pm z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} = \big[\,\bar{X} - z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}}\,,\ \bar{X} + z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}}\,\big].$$

Interpretação: em longas repetições de amostras, $100(1-\alpha)\%$ dos IC construídos conterão o valor verdadeiro μ.

8.3 Intervalo para a Média com σ Desconhecido

Quando σ é desconhecido, usamos a variância amostral S^2 e a distribuição t de Student:

$$IC_{1-\alpha}(\mu) : \bar{X} \pm t_{\alpha/2,\,n-1}\,\frac{S}{\sqrt{n}} = \big[\,\bar{X} - t_{\alpha/2,\,n-1}\tfrac{S}{\sqrt{n}}\,,\ \bar{X} + t_{\alpha/2,\,n-1}\tfrac{S}{\sqrt{n}}\,\big],$$

onde $t_{\alpha/2,\,n-1}$ é o quantil da t-Student com $n-1$ graus de liberdade.

8.4 Derivação Intuitiva

1. A distribuição de $\bar{X}$ tem média μ e desvio padrão $\sigma/\sqrt{n}$ (ou $S/\sqrt{n}$).
2. Padronizando, obtemos uma variável com distribuição conhecida ($N(0,1)$ ou t).
3. Selecionamos os pontos simétricos $\pm z_{\alpha/2}$ (ou $\pm t_{\alpha/2}$) que deixam área α nas duas caudas.
4. Desfazendo a padronização, encontramos os limites do intervalo.

8.5 Tamanho de Amostra e Precisão

Para garantir uma margem de erro E ao nível $1-\alpha$:

$$E = z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} \quad \Longrightarrow \quad n = \Big(\frac{z_{\alpha/2}\,\sigma}{E}\Big)^2.$$

Quando σ não é conhecido, usa-se uma estimativa inicial de S.

8.6 Noções para Ciência de Dados

- **Validação de Modelos:** construão de IC para parâmetros de regressão para avaliar a estabilidade das previsões.
- **Comparação de Grupos:** usar IC de diferença de médias ou proporções para verificar se a diferença observada é estatisticamente relevante.
- **Métricas de Performance:** IC para métricas (ex.: acurácia, AUC) via bootstrap para capturar incerteza.