



Análise de Dados Aplicada à Computação

EXPERIMENTOS ESTATÍSTICOS

Prof. M.Sc. Howard Roatti

Introdução aos Testes Estatísticos

- **Objetivo:**

- (a) Os testes estatísticos são ferramentas usadas para testar uma hipótese ou afirmativa sobre uma população a partir de uma amostra.
- (b) Eles nos ajudam a determinar se os resultados observados em uma amostra são estatisticamente significativos e, portanto, se podemos generalizá-los para a população maior.

- **Hipótese:** A Probabilidade de uma Métrica Obtida em um Experimento sobre uma Amostra Representar a População

- **Tipos de Experimentos:** teste A/B, teste de hipótese, teste t e teste t pareado, entre outros...

Introdução aos Testes Estatísticos

- São ferramentas que ajudam a tomar decisões baseadas em dados.
- Uma hipótese estatística é uma afirmação sobre uma população, que pode ser testada usando dados de uma amostra. Exemplo:
 - **H0 (Hipótese Nula)** → a média de idade dos alunos de uma escola é de 23 anos?
- A partir dessa hipótese, fazemos um teste estatístico para verificar se os dados de uma amostra são consistentes com essa hipótese.
- Envolve estabelecer um nível de significância, que é a probabilidade de obter um resultado tão extremo ou mais extremo do que o que foi observado, assumindo que a H0 é verdadeira.

Introdução aos Testes Estatísticos

- Se o resultado observado for altamente improvável de ocorrer sob a **H0**, então podemos **Rejeitar a Hipótese Nula** e concluir que a **H1**, **Hipótese Alternativa**, (em nosso exemplo, a média de idade é diferente de 23 anos) é mais provável de ser verdadeira.
- Os testes estatísticos também podem nos ajudar a determinar se as diferenças observadas entre amostras são estatisticamente significativas ou apenas o resultado do acaso.
- Eles podem ser usados para comparar duas ou mais amostras, para determinar a associação entre duas variáveis e para testar a relação entre uma variável dependente e uma ou mais variáveis independentes.

Teste A/B

- O teste A/B é um método estatístico utilizado para comparar duas versões diferentes de um elemento, como uma página de um site, um anúncio ou um e-mail.
- O objetivo é determinar qual versão é mais efetiva em alcançar um objetivo específico, como aumentar as conversões ou reduzir a taxa de rejeição.
- Para realizar um teste A/B, é preciso dividir uma amostra de usuários em dois grupos aleatórios: o grupo A e o grupo B.
 - O grupo A recebe a versão original do elemento; e
 - o grupo B recebe a versão modificada.

Teste A/B

- Em seguida, é preciso medir um indicador de desempenho, como a taxa de conversão, para cada grupo durante um período de tempo determinado.
- É importante que os dois grupos sejam expostos às mesmas condições externas durante o teste, para garantir que quaisquer diferenças de desempenho sejam causadas pela variação do elemento testado.
- Após coletar os dados, é possível aplicar um teste estatístico para determinar se a diferença de desempenho é estatisticamente significativa ou não.
- Se a diferença for significativa, podemos concluir que a versão que teve o melhor desempenho é a vencedora e deve ser implementada permanentemente.

Teste A/B – Exemplo Prático

- Suponha que temos um e-commerce e queremos testar duas variações de um botão "**Comprar Agora**" na página do produto para ver qual deles leva a uma taxa maior de conversão. A primeira variação tem um **texto mais agressivo** e a segunda variação tem um **texto mais suave**.
- Vamos supor que temos uma base de dados de usuários que visitam a página do produto e que já coletamos informações sobre as conversões. Podemos usar o pacote scipy do Python para realizar o teste A/B.

Teste A/B – Exemplo Prático

- Para começar, vamos importar as bibliotecas necessárias:

```
import numpy as np  
from scipy.stats import ttest_ind
```

- Vamos criar duas amostras aleatórias para representar os grupos A e B.
- Para isso, vamos simular as conversões com números aleatórios.
- Vamos supor que temos 1000 usuários em cada grupo:

```
np.random.seed(123)  
group_a_conversions = np.random.binomial(100, 0.03, size=1000)  
group_b_conversions = np.random.binomial(100, 0.04, size=1000)
```


Teste A/B – Exemplo Prático

- Neste exemplo, estamos simulando as conversões como uma distribuição binomial com uma taxa de conversão de 3% para o grupo A e 4% para o grupo B.
- Em seguida, podemos calcular a média e o desvio padrão de cada grupo:

```
group_a_mean = np.mean(group_a_conversions)
group_b_mean = np.mean(group_b_conversions)
group_a_std = np.std(group_a_conversions)
group_b_std = np.std(group_b_conversions)
```

Teste A/B – Exemplo Prático

- Agora, podemos aplicar o **teste t** para determinar se a diferença de desempenho entre os dois grupos é estatisticamente significativa.

```
t, p = ttest_ind(group_a_conversions, group_b_conversions)
```

- O valor **t** é o valor estatístico do **teste t** e o **valor p** é o valor p, que **representa a probabilidade de obter um resultado tão extremo quanto o observado**, assumindo que não há diferença real entre os grupos.

```
print(f'Group A mean: {group_a_mean:.2f}, std: {group_a_std:.2f}')  
print(f'Group B mean: {group_b_mean:.2f}, std: {group_b_std:.2f}')  
print(f't: {t:.2f}, p: {p:.2f}')
```

Teste A/B – Exemplo Prático

- Por fim, podemos interpretar os resultados.
- Se o **valor p** for **menor que** um nível de significância escolhido (**geralmente 0,05**), podemos rejeitar a Hipótese Nula de que não há diferença real entre os grupos e concluir que há uma diferença estatisticamente significativa.
- Nesse caso, podemos escolher a variação que teve o melhor desempenho como a vencedora.
- É importante que os resultados dos testes A/B sejam interpretados com cautela e que sejam levados em consideração outros fatores, como as limitações técnicas, a usabilidade e a experiência do usuário.

Teste de Hipótese

- O teste de hipótese é uma técnica estatística que permite testar uma afirmação sobre uma população ou um conjunto de dados. Ele envolve a formulação de duas hipóteses:
 - a hipótese nula (**H0**); e
 - a hipótese alternativa (**H1**).
- A **hipótese nula** é a afirmação que queremos testar e é geralmente a afirmação de que não há diferença ou efeito entre as amostras ou populações.
- A **hipótese alternativa** é a afirmação contrária à hipótese nula, ou seja, é a afirmação de que há uma diferença ou efeito significativo.

Teste de Hipótese

- Assim como no exemplo prático anterior, escolhemos um **nível de significância** para o teste, que **representa a probabilidade máxima de rejeitarmos a hipótese nula** quando ela é realmente verdadeira. Geralmente, esse nível é fixado em **5%** ou **1%**.
- Existem dois tipos de erros que podem ocorrer ao realizar um teste de hipótese: o erro tipo I e o erro tipo II.
 - O **erro tipo I** ocorre quando rejeitamos a hipótese nula quando ela é verdadeira.
 - Já o **erro tipo II** ocorre quando não rejeitamos a hipótese nula quando ela é falsa.

Teste de Hipótese

- Para realizar um teste de hipótese, geralmente usamos um teste estatístico adequado para o tipo de dados que estamos analisando. Por exemplo, se estamos analisando **dados contínuos**, podemos usar o **teste t de Student**. Se estamos analisando **dados categóricos**, podemos usar o **teste do qui-quadrado**.

Teste de Hipótese

- Um conceito importante no teste de hipótese é o **valor-p** (***p-value***), que é a probabilidade de obtermos um resultado tão extremo ou mais extremo do que o observado, supondo que a Hipótese Nula seja verdadeira.
- Se o **valor-p** for muito baixo (geralmente menor que **0,05** ou **0,01**), podemos **rejeitar a hipótese nula** e **concluir que há evidências estatísticas significativas para aceitar a hipótese alternativa**.

Teste de Hipótese

- A **Significância Estatística** é uma medida de quão confiáveis são os resultados obtidos por meio do teste de hipótese. Ela **está relacionada ao valor-p** e ao **nível de significância** escolhido para o teste. Quanto menor for o valor-p e menor for o nível de significância, mais significativo é o resultado e mais confiáveis são as evidências estatísticas para rejeitar a hipótese nula.
- A confiabilidade dos resultados também depende do **tamanho da amostra** e do **grau de incerteza** associado aos dados. Quanto maior for o tamanho da amostra e menor for a incerteza, mais confiáveis são os resultados obtidos por meio do teste de hipótese.

Teste de Hipótese

- A probabilidade de cometer o **erro do tipo I** é geralmente determinada pelo **nível de significância** escolhido para o teste.
- Por exemplo, se escolhermos um nível de significância de 0,05, estamos dizendo que estamos dispostos a aceitar um erro do tipo I de 5%. Ou seja, **há uma probabilidade de 5% de rejeitarmos a hipótese nula, mesmo que ela seja verdadeira.**

Teste de Hipótese

- Por outro lado, a probabilidade de cometer o **erro do tipo II** depende do **tamanho da amostra**, do tamanho do **efeito**, do **nível de significância** e do **poder do teste**.
- O **poder do teste** é a **probabilidade de rejeitar corretamente a hipótese nula, quando ela é falsa**. Portanto, quanto maior o poder do teste, menor é a probabilidade de cometer o erro do tipo II.
- Para aumentar a confiabilidade dos resultados, é importante escolher um tamanho de amostra adequado e um nível de significância apropriado.

Teste de Hipótese

- Existem vários métodos de teste de hipótese disponíveis, cada um com suas próprias suposições e limitações. Alguns dos testes mais comuns incluem:
- O teste **t de Student** é um teste paramétrico usado para comparar as médias de duas amostras independentes ou pareadas.
- O teste de **ANOVA** (Análise de Variância) é usado para comparar as médias de mais de duas amostras independentes.
- O teste de **qui-quadrado** é usado para testar a associação entre variáveis categóricas.

Teste de Hipótese

- O teste de **Wilcoxon** é um teste **não paramétrico** usado para comparar as médias de duas amostras pareadas. Ele é usado **quando as suposições do teste t de Student não são atendidas**, como quando as amostras não seguem uma distribuição normal.
- Entre outros testes...

Teste t de Student

- O **teste t** é baseado no cálculo da **estatística t**, que é **uma medida da diferença entre as médias das amostras em relação à variação dentro das amostras**. A estatística t é **calculada como a diferença entre as médias dividida pelo erro padrão da diferença**. O erro padrão da diferença leva em conta o tamanho da amostra, o desvio padrão e a correlação entre as amostras.
- Assim como no teste de hipótese, é necessário **definir H0 e H1 antes de realizar o teste t**. A hipótese nula é a de que as médias das amostras são iguais, enquanto a hipótese alternativa é a de que as médias são diferentes.

Teste t de Student

- O resultado do teste t é um **valor de p**, que **indica a probabilidade de observar uma diferença entre as médias das amostras tão grande ou maior que a observada**, assumindo que a hipótese nula seja verdadeira. Se o valor de p for menor do que um determinado nível de significância, geralmente 0,05, então a hipótese nula é rejeitada e a hipótese alternativa é aceita.
- Na prática, o teste t é **uma variação do teste z**, que é utilizado para testar hipóteses envolvendo proporções ou desvios padrão conhecidos.

Teste t de Student

- Ele é um teste estatístico utilizado para avaliar se as médias de duas **amostras independentes** são significativamente diferentes uma da outra.
- Amostras independentes são duas amostras que não têm nenhuma relação entre si.

Teste t – Exemplo Prático

- Suponha que você tenha duas amostras, uma de um grupo de controle e outra de um grupo de tratamento. Você deseja saber se a média do grupo de tratamento é significativamente diferente da média do grupo de controle.
- Primeiro, vamos importar a biblioteca necessária e criar as amostras:

```
import numpy as np  
amostra_controle = np.array([10, 11, 12, 14, 15, 15, 16, 17, 17, 18])  
amostra_tratamento = np.array([12, 13, 14, 14, 15, 16, 17, 18, 18, 20])
```


Teste t – Exemplo Prático

- Em seguida, podemos calcular as médias e as variâncias amostrais de cada grupo:

```
media_controle = np.mean(amostra_controle)
media_tratamento = np.mean(amostra_tratamento)

variancia_controle = np.var(amostra_controle, ddof=1)
variancia_tratamento = np.var(amostra_tratamento, ddof=1)
```

- Note que estamos usando o argumento `ddof=1` na função `np.var` para **ajustar a variância amostral para um grau de liberdade a menos**, que é necessário quando se está trabalhando com uma amostra.

Teste t – Exemplo Prático

- Agora, podemos calcular a estatística de teste t e o valor-p:

```
from scipy.stats import ttest_ind  
  
t, p = ttest_ind(amostra_tratamento, amostra_controle, equal_var=False)
```

- Estamos usando a função **ttest_ind** do módulo **scipy.stats** para calcular a estatística de **teste t** e o **valor-p**. O argumento **equal_var=False** indica que estamos assumindo que as variâncias das amostras não são iguais.
- Por fim, podemos imprimir os resultados:

```
print(f"Média do grupo de controle: {media_controle:.2f}")  
print(f"Média do grupo de tratamento: {media_tratamento:.2f}")  
print(f"Teste t: {t:.2f}")  
print(f"Valor-p: {p:.3f}")
```

Teste t – Exemplo Prático

```
In [17]: print(f"Média do grupo de controle: {media_controle:.2f}")
Média do grupo de controle: 14.50

In [18]: print(f"Média do grupo de tratamento: {media_tratamento:.2f}")
Média do grupo de tratamento: 15.70

In [19]: print(f"Teste t: {t:.2f}")
Teste t: 1.02

In [20]: print(f"Valor-p: {p:.3f}")
Valor-p: 0.321
```

- Podemos concluir que, com um nível de significância de 5%, não há evidências suficientes para rejeitar a hipótese nula de que as médias dos grupos são iguais.

Teste t – Exemplo Prático 2

- Suponha que você tenha duas amostras, uma de um grupo de controle e outra de um grupo de tratamento. Você deseja saber se a média do grupo de tratamento é significativamente diferente da média do grupo de controle.
- Primeiro, vamos importar a biblioteca necessária e criar as amostras:

```
import numpy as np

amostra_controle = np.array([10, 12, 12, 13, 14, 15, 15, 16, 16, 18])
amostra_tratamento = np.array([14, 15, 15, 16, 16, 17, 17, 18, 19, 20])
```

Teste t – Exemplo Prático 2

```
media_controle = np.mean(amostra_controle)
media_tratamento = np.mean(amostra_tratamento)

variancia_controle = np.var(amostra_controle, ddof=1)
variancia_tratamento = np.var(amostra_tratamento, ddof=1)

from scipy.stats import ttest_ind

t, p = ttest_ind(amostra_tratamento, amostra_controle, equal_var=False)

print(f"Média do grupo de controle: {media_controle:.2f}")
print(f"Média do grupo de tratamento: {media_tratamento:.2f}")
print(f"Teste t: {t:.2f}")
print(f"Valor-p: {p:.3f}")
```

Teste t – Exemplo Prático

```
In [10]: print(f"Média do grupo de controle: {media_controle:.2f}")  
Média do grupo de controle: 14.10  
  
In [11]: print(f"Média do grupo de tratamento: {media_tratamento:.2f}")  
Média do grupo de tratamento: 16.70  
  
In [12]: print(f"Teste t: {t:.2f}")  
Teste t: 2.71  
  
In [13]: print(f"Valor-p: {p:.3f}")  
Valor-p: 0.015
```

- Podemos concluir que, com um nível de significância de 5%, há evidências suficientes para rejeitar a hipótese nula de que as médias dos grupos são iguais e aceitar a hipótese alternativa de que as médias são diferentes. **Nesse caso, podemos concluir que o grupo de tratamento apresentou uma média significativamente maior do que o grupo de controle.**

Teste t Pareado

- O teste t pareado é **uma variação do teste t para amostras dependentes**. Ele é usado para testar a diferença entre duas médias relacionadas ou pareadas de uma única amostra. Essas amostras podem ser medidas antes e depois de um experimento ou emparelhadas de alguma outra forma.
- O teste t pareado é **baseado na diferença entre as observações emparelhadas e na distribuição t de Student**. A H_0 é que não há diferença significativa entre as médias das duas variáveis pareadas. A H_1 é que há uma diferença significativa entre as médias das duas variáveis pareadas.

Teste t Pareado

- Assim como o teste t independente, **o teste t pareado também usa o valor-p** para determinar a significância estatística dos resultados.
- Um exemplo prático do teste t pareado pode ser a **comparação da pressão arterial sistólica** de um paciente antes e depois de um tratamento. Nesse caso, **as observações são emparelhadas porque se referem ao mesmo paciente**. O objetivo do teste seria determinar se houve uma diferença significativa na pressão arterial sistólica antes e depois do tratamento.

Teste t Pareado – Ex. Prático

- Para realizar o teste t pareado no Python, podemos utilizar a função **ttest_rel** da biblioteca **scipy.stats**. Essa função calcula o **valor-p** do teste t pareado para um conjunto de amostras emparelhadas.

```
import scipy.stats as stats

# dados emparelhados
antes = [120, 128, 126, 124, 130]
depois = [118, 126, 124, 122, 128]

# realiza o teste t pareado
t_stat, p_value = stats.ttest_rel(antes, depois)

# exibe o resultado
print('t = %.2f, p = %.4f' % (t_stat, p_value))
```

Teste t Pareado

- A **interpretação** do resultado do teste t pareado **depende do valor-p**. Se o valor-p for menor que o nível de significância escolhido (por exemplo, **0.05**), podemos rejeitar a hipótese nula e concluir que há evidências suficientes para suportar a hipótese alternativa.
- Por outro lado, **se o valor-p for maior que o nível de significância, não podemos rejeitar a hipótese nula e concluimos que não há evidências suficientes para suportar a hipótese alternativa**, ou seja, que não houve uma diferença significativa na pressão arterial sistólica antes e depois do tratamento.

Testes de Wilcoxon

- É utilizado para comparar duas **amostras independentes não normalmente distribuídas**.
- Esse teste é uma alternativa ao teste t independente e apresenta a **vantagem de não requerer a suposição de normalidade na distribuição das amostras**.
- O teste de Wilcoxon pode ser utilizado, por exemplo, para **comparar a eficácia de dois tratamentos médicos em uma amostra de pacientes não normalmente distribuída**. Ele avalia se há diferença significativa entre as medianas das duas amostras e pode ser realizado tanto para amostras pequenas quanto para amostras grandes.

Testes de Wilcoxon

- O teste de Wilcoxon **é uma ótima escolha para comparar dois algoritmos**, especialmente quando as distribuições dos dados não são normais ou quando os dados são ordinais.
- O teste de Wilcoxon **pode ser utilizado para comparar a mediana de duas amostras pareadas** (comparando o desempenho de dois algoritmos). O **teste de Wilcoxon pareado** é uma versão do teste de Wilcoxon que é aplicado **quando as duas amostras são relacionadas**, como por exemplo, quando as mesmas instâncias são executadas pelos dois algoritmos.

Testes de Wilcoxon

- O teste de Wilcoxon pareado é executado da seguinte forma: primeiro, as diferenças entre as observações de cada par são calculadas. Em seguida, a diferença entre cada par é ranqueada em ordem crescente (ignorando os valores de zero) e soma-se as classificações para diferenças positivas e negativas. Por fim, **o teste verifica se há diferença significativa entre as classificações positivas e negativas.**
- Se o valor p resultante for menor que o nível de significância estabelecido, então a H_0 é rejeitada e podemos concluir que há diferença estatisticamente significativa entre as amostras. Caso contrário, não há evidência suficiente para rejeitar a H_1 .

Testes de Wilcoxon – Ex. Prático

- Suponha que você desenvolveu um algoritmo de compressão de arquivos e deseja comparar sua performance com um algoritmo já existente no mercado. Você coletou dados de tempo de execução de ambos os algoritmos em um conjunto de arquivos de teste.
- A **hipótese nula** é que **não há diferença significativa** entre as médias dos tempos de execução dos dois algoritmos, enquanto a **hipótese alternativa** é que **o algoritmo desenvolvido é mais rápido**.

Testes de Wilcoxon – Ex. Prático

```
import numpy as np
from scipy.stats import wilcoxon

meu_algoritmo = np.array([15.4, 16.2, 14.8, 15.5, 14.9, 15.7, 16.1, 15.3, 15.6, 16.0])
algoritmo_referencia = np.array([16.9, 17.5, 17.1, 18.0, 17.2, 16.7, 17.9, 17.8, 16.8, 17.3])

stat, p = wilcoxon(meu_algoritmo, algoritmo_referencia)
print('Estatística de teste: %.3f' % stat)
print('Valor de p: %.3f' % p)

# Interpretando o resultado
alpha = 0.05
if p > alpha:
    print('Não rejeitamos a hipótese nula de que os tempos de execução são iguais')
else:
    print('Rejeitamos a hipótese nula e concluímos que o algoritmo desenvolvido é mais rápido')
```

Resumo

- Nesta aula, aprendemos sobre alguns dos testes estatísticos mais comuns utilizados em experimentos e análises de dados. Começamos com uma introdução aos testes estatísticos e a importância deles na tomada de decisão. Em seguida, vimos o teste A/B, que é muito utilizado em experimentos de marketing digital para comparar duas versões de uma página ou campanha.
- Depois, aprendemos sobre o teste de hipótese e o p-value, que são fundamentais para validar se uma diferença observada entre dois grupos é estatisticamente significativa. Discutimos também a importância da confiabilidade nos resultados e do estabelecimento de um nível de significância adequado.

Resumo

- Em seguida, apresentamos o teste t, que é utilizado para comparar as médias de duas amostras independentes. Vimos como calcular o valor t e como interpretá-lo para determinar se as médias são estatisticamente diferentes.
- Também aprendemos sobre o teste t pareado, que é utilizado para comparar as médias de duas amostras pareadas, como antes e depois de um tratamento.
- Por fim, discutimos o teste de Wilcoxon, que é um teste não paramétrico utilizado quando as condições de aplicação do teste t não são atendidas.

Resumo

- É importante lembrar que os testes estatísticos apresentados nesta aula são apenas alguns dos muitos existentes, e que é fundamental escolher o teste correto para cada situação. Além disso, é essencial levar em consideração o tamanho da amostra, o nível de significância e outros fatores que podem afetar os resultados.

Seja curioso...

- Pesquise sobre os testes estatísticos:
 - Kolmogorov-Smirnov
 - Anderson-Darling
 - Shapiro-Wilk
 - qui-quadrado (χ^2)

Referências

- Bruce, P.; Bruce, A.; **Estatística Prática para Cientista de Dados: 50 Conceitos Essenciais**; Rio de Janeiro; Alta Books; 2019.
- Morettin, P. A.; Bussab, W. O.; **Estatística Básica**. 8 ed. São Paulo: Saraiva, 2013.
- <https://blog.minitab.com/pt/como-entender-os-testes-t-valores-t-e-distribuicoes-t>
- <https://blog.minitab.com/pt/entendendo-analise-de-variancia-anova-e-o-teste-f>
- https://www.scielo.br/pdf/jbpneu/v41n5/pt_1806-3713-jbpneu-41-05-00485.pdf
- <https://www.voitto.com.br/blog/artigo/teste-de-hipotese>

Análise de Dados Aplicada à Computação

PROF. M.SC HOWARD ROATTI