Inferência estatística com R e Python

Sumário

1 Introdução

Este projeto foi realizado na monitoria da disciplina de Inferência Estatística, com o objetivo de fornecer aos alunos exemplos práticos que possam orientar a aplicação dos conceitos estudados. Cada tópico apresenta uma breve explicação, códigos em R e Python, e uma interpretação dos resultados, visando facilitar a compreensão e a prática dos conteúdos abordados em aula.

2 Principais Bibliotecas

As principais bibliotecas para a aplicação de inferência estatística em python são:

1. **SciPy**: O módulo scipy.stats oferece uma ampla gama de ferramentas para realizar testes estatísticos, cálculos de distribuição de probabilidade, ajuste de modelos de dados, análise de regressão, e mais. Exemplos incluem testes t, ANOVA, testes qui-quadrado e funções de distribuição.

2. **Statsmodels**: Focada em modelagem estatística, a statsmodels oferece funcionalidades para realizar regressões lineares, logísticas, e de séries temporais, além de testes de hipóteses, análise de resíduos, e estatísticas descritivas. Ideal para análise econométrica e modelos GLM (Modelos Lineares Generalizados).

3. **Pandas**: Embora seja conhecida principalmente pela manipulação de dados, o pandas tem suporte básico para operações estatísticas, como medidas descritivas (média, desvio padrão, variância) e algumas funcionalidades de inferência simples, como correlação e regressão.

4. **Scikit-learn**: Usada principalmente para aprendizado de máquina, mas também oferece funcionalidades de inferência estatística, como validação cruzada, testes de hipótese e métodos de avaliação de desempenho de modelos. Tem módulos para regressão linear, análise de componentes principais (PCA), e outras análises estatísticas.

Em R, há várias bibliotecas robustas para a aplicação de inferência estatística. A seguir estão algumas das principais:

1. **Stats**: Faz parte do núcleo do R e inclui a maioria dos testes estatísticos clássicos, como testes t, ANOVA, regressão linear, testes de qui-quadrado, e funções de distribuição de probabilidade. É a base para a maior parte das funções estatísticas em R

2. **Cars**: O pacote car (Companion to Applied Regression) oferece ferramentas adicionais para a análise de regressão, incluindo diagnósticos de regressão, testes de hipótese multivariados, e ANOVA para modelos lineares e não lineares. Muito útil para análise de regressão mais avançada.

3. **MASS**: Um dos pacotes mais clássicos em R, oferece ferramentas para análise estatística avançada, como modelos lineares generalizados (GLMs), análise discriminante, regressão logística, e análise de componentes principais. Inclui também métodos de ajuste robusto.

4. **Infer**: Um pacote relativamente novo e muito intuitivo para realizar inferência estatística, que facilita a condução de testes de permutação, bootstrap, e construção de intervalos de confiança por meio de sintaxe simplificada.

Essas bibliotecas são apenas alguns exemplos, mas muitas outras podem ser utilizadas.

3 Intervalo de Confiança

Um intervalo de confiança fornece um intervalo de valores dentro do qual se espera que o valor de um parâmetro populacional caia, com um certo nível de confiança (geralmente 95%). Por exemplo, se tivermos uma amostra de alturas, podemos calcular um intervalo de confiança para a média da altura na população.

Exemplo em R

```r
# Gerar uma amostra de alturas
alturas <- rnorm(100, mean = 170, sd = 10)
media <- mean(alturas)
erro_padrao <- sd(alturas) / sqrt(length(alturas))

# Intervalo de Confianca de 95%
intervalo <- c(media - 1.96 * erro_padrao, media + 1.96 *
    erro_padrao)
print(intervalo)
```

Exemplo em Python

```python
import numpy as np
from scipy import stats

# Gerar uma amostra de alturas
alturas = np.random.normal(loc=170, scale=10, size=100)
media = np.mean(alturas)
erro_padrao = stats.sem(alturas)

# Intervalo de Confianca de 95%
intervalo = stats.t.interval(0.95, len(alturas)-1, loc=media,
    scale=erro_padrao)
print(intervalo)
```

Interpretação

O intervalo de confiança de 95% significa que, se repetíssemos o experimento muitas vezes, 95% dos intervalos calculados conteriam a verdadeira média populacional. Por exemplo, se o intervalo for (168, 172), podemos dizer que estamos 95% confiantes de que a média verdadeira da altura na população está entre 168 e 172.

4 Teste de Hipóteses

Os testes de hipóteses são usados para testar uma afirmação sobre um parâmetro populacional com base em dados de amostra. Um exemplo comum é o teste t para comparar a média da amostra com um valor específico.

Exemplo em R

```r
# Teste t para a media
t.test(alturas, mu = 170)
```

Exemplo em Python

```python
# Teste t para a media
stats.ttest_1samp(alturas, popmean=170)
```

Interpretação

O teste t retorna um valor de p (p-value). Se o p-value for menor que 0,05 (comumente usado como nível de significância), rejeitamos a hipótese nula. Isso significa que a média da amostra é significativamente diferente de 170. Caso contrário, não rejeitamos a hipótese nula e consideramos que não há evidências suficientes para afirmar que a média é diferente de 170.

5 ANOVA (Análise de Variância)

ANOVA é usada para comparar as médias de três ou mais grupos para verificar se pelo menos um deles é significativamente diferente dos outros. É uma extensão do teste t para mais de dois grupos.

Exemplo em R

```r
# Criando tres grupos de alturas
grupo1 <- rnorm(50, mean = 170, sd = 10)
grupo2 <- rnorm(50, mean = 165, sd = 10)
grupo3 <- rnorm(50, mean = 175, sd = 10)

# Realizando ANOVA
dados <- data.frame(alturas = c(grupo1, grupo2, grupo3),
                    grupo = factor(rep(1:3, each=50)))
resultado_anova <- aov(alturas ~ grupo, data = dados)
summary(resultado_anova)
```

Exemplo em Python

```python
from scipy.stats import f_oneway

# Criando tres grupos de alturas
grupo1 = np.random.normal(170, 10, 50)
grupo2 = np.random.normal(165, 10, 50)
grupo3 = np.random.normal(175, 10, 50)

# Realizando ANOVA
resultado_anova = f_oneway(grupo1, grupo2, grupo3)
print(resultado_anova)
```

Interpretação

ANOVA retorna um valor de p (p-value). Se o p-value for menor que 0,05, rejeitamos a hipótese nula de que todas as médias dos grupos são iguais. Isso significa que pelo menos um dos grupos é significativamente diferente dos outros. Caso o p-value seja maior que 0,05, não temos evidências suficientes para afirmar que há diferença significativa entre as médias dos grupos.

6 Regressão Linear

A regressão linear é usada para modelar a relação entre uma variável dependente e uma ou mais variáveis independentes. Pode ser usada para prever valores e analisar a força da relação entre variáveis.

Exemplo em R

```r
# Dados de exemplo
dados <- data.frame(
  x = rnorm(100, mean = 5, sd = 2),
  y = 3 * rnorm(100, mean = 5, sd = 2) + 4
)

# Ajustando o modelo
modelo <- lm(y ~ x, data = dados)
summary(modelo)
```

Exemplo em Python

```python
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression

# Dados de exemplo
np.random.seed(0)
x = np.random.normal(5, 2, 100).reshape(-1, 1)
y = 3 * np.random.normal(5, 2, 100) + 4

# Ajustando o modelo
modelo = LinearRegression().fit(x, y)
print(f'Coeficiente:{modelo.coef_[0]},Intercepto:{modelo.
    intercept_}')
```

Interpretação

O coeficiente da regressão indica quanto a variável dependente muda em média para cada unidade de aumento na variável independente. O intercepto é o valor da variável dependente quando a variável independente é zero. Um coeficiente positivo indica uma relação direta entre as variáveis, enquanto um negativo indica uma relação inversa.

7 Teste Qui-quadrado

O teste qui-quadrado é usado para determinar se há uma associação significativa entre duas variáveis categóricas.

Exemplo em R

```r
# Tabela de contingencia
tabela <- matrix(c(10, 20, 30, 40), nrow = 2)
colnames(tabela) <- c("Categoria1", "Categoria2")
rownames(tabela) <- c("GrupoA", "GrupoB")

# Teste qui-quadrado
chisq.test(tabela)
```

Exemplo em Python

```python
import numpy as np
from scipy.stats import chi2_contingency

# Tabela de contingencia
tabela = np.array([[10, 20], [30, 40]])

# Teste qui-quadrado
chi2, p, _, _ = chi2_contingency(tabela)
print(f'Qui-quadrado:{chi2},p-valor:{p}')
```

Interpretação

Se o p-value do teste qui-quadrado for menor que 0,05, rejeitamos a hipótese nula de que as variáveis são independentes. Isso indica que há uma associação significativa entre as duas variáveis. Caso contrário, não há evidências suficientes para afirmar que existe uma associação.

8 Correlação

A correlação mede a força e a direção da relação linear entre duas variáveis. O coeficiente de correlação varia de -1 a 1, onde -1 indica uma correlação negativa perfeita, 0 indica nenhuma correlação e 1 indica uma correlação positiva perfeita.

Exemplo em R

```r
# Gerar duas variaveis aleatorias
x <- rnorm(100)
y <- rnorm(100)

# Coeficiente de correlacao de Pearson
cor(x, y)
```

Exemplo em Python

```python
# Gerar duas vari veis aleat rias
x = np.random.normal(size=100)
y = np.random.normal(size=100)

# Coeficiente de correlacao de Pearson
correlacao = np.corrcoef(x, y)[0, 1]
print(correlacao)
```

Interpretação

Um coeficiente de correlação próximo de 1 ou -1 indica uma forte correlação positiva ou negativa, respectivamente. Um coeficiente próximo de 0 indica que as variáveis não têm uma relação linear. A interpretação deve considerar o contexto do estudo e a chance de que a correlação seja por acaso.