Métodos de Estatística Aplicada com Python

Aula 8

Carlos Góes[1]

[1]Pós-Graduação em Ciência de Dados
Instituto de Educação Superior de Brasília

2017

Sumário

Sumário

Introdução ao teste de hipótese

Intuição

- Em estatística, muitas vezes queremos testar uma hipótese
- Podemos, por exemplo, comparar a performance média de dois grupos e, sabando da variabilidade dos estimadores, entender o quão confiável é a diferença entre eles

Introdução ao teste de hipótese

Intuição

- Os passos essenciais, portanto, são:
    - definir uma hipótese a ser testada;
    - definir qual o nosso grau de incerteza tolerável para aceitar tal hipótese

Introdução ao teste de hipótese

Intuição

- A notação que utilizamos é:

$$\begin{aligned} H_0 &: \textit{hipótese nula} \\ H_a &: \textit{hipótese alternativa} \end{aligned}$$

- H_0 é a hipótese que vamos testar. Se a rejeitamos, aceitamos a hipótese alternativa.

Introdução ao teste de hipótese

Intuição

- Exemplo 1:

$$
\begin{aligned}
H_0 &: P(\textit{câncer}|\textit{fumante}) > P(\textit{câncer}|\textit{não} - \textit{fumante}) \\
H_a &: P(\textit{câncer}|\textit{fumante}) \leq P(\textit{câncer}|\textit{não} - \textit{fumante})
\end{aligned}
$$

- Se testamos e rejeitamos H_0, aceitamos H_a.
- Se testamos e não-rejeitamos H_0, aceitamos H_0.
- Note que não estamos provando que H_0 é correta, simplesmente não a rejeitamos, dados nossos critérios de grau de incerteza tolerável.

Sumário

Estatística-t

Definição

- Para saber o quão distante o valor de corte está, em erros padrão, da nossa estimativa, podemos calcular a "estatística t":

$$t_0 = \frac{\hat{\theta} - \theta_0}{s_{\hat{\theta}}} = \frac{estimativa \quad - \quad valor \quad de \quad corte}{erro \quad padrão} \tag{1}$$

Estatística-t

Definição

- Podemos calcular isso no python. Primeiro temos que criar uma fórmula para calcular o erro padrão:

```
import numpy as np
from scipy import stats

aleatorio = lambda media, desvio_padrao, n_obs:
np.random.normal(media, desvio_padrao, n_obs)

def erro_padrao(amostra):
    amostra = amostra[~np.isnan(amostra)]

    numerador = np.std(amostra, ddof=1)
    denominador = np.sqrt(len(amostra))

    return numerador / denominador
```

Estatística-t

Definição

- Depois, conferir que o erro padrão sempre diminui à medida que o tamanho da amostra aumenta

```
media = 0
desvio_padrao = 10

for i in [10, 100, 1000, 10000]:
    ep = erro_padrao(aleatorio(media, desvio_padrao, n_obs=i))
    print("n_obs: {}; erro padrão: {:.2f}".format(i, ep))
```

Estatística-t

Definição

- Com base nisso, criamos uma fórmula para calcular a estatística-t:

```
def t_stat_media(amostra, valor_corte):
    media = np.mean(amostra)
    numerador = (media - valor_corte)
    denominador = erro_padrao(amostra)
    t_stat = numerador / denominador

    return t_stat, media
```

Estatística-t

Definição

- E ver qual a estatística-t para distintos valores de corte:

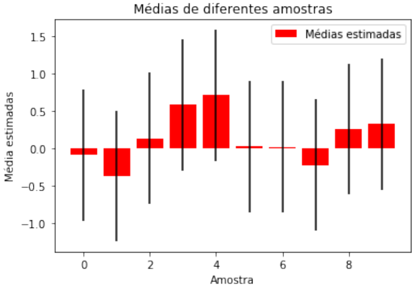
```
x = aleatorio(0, 100, 1000)

for i in np.linspace(0,25,26):
    est_t, media = t_stat_media(x, i)
    print("Média: {:.2f}; valor de corte: {:.2f}; estatística-t: {:.2f}"
          .format(media, i, est_t))
```

Sumário

P-valor

Definição

- Quando discutimos intervalos de confiança, obtemos uma série de valores críveis, dada a incerteza amostral, para dada estatística.
- Por exemplo, 95% das médias estimadas para cada uma das amostras abaixo estão no intervalo representado pela linha preta
- Todos os valores no intervalo são valores críveis para as médias



P-valor

Definição

- Já quando calculamos o p-valor, estamos preocupados em avaliar uma única possibilidade, qual seja: qual a probabilidade de determinada estatística ser (ou não ser), de fato, aquela que estimamos?
- Em testes de significância estatística, tentamos medir o grau de evidência que os dados fornecem frente a um valor especificado ou suposto do parâmetro

P-valor

Definição

- O p-valor denota a probabilidade de observar um resultado igual ou maior, em magnitude, ao observado, dado que a hipótese nula é correta:

$$Pr(X \geq x|H)$$

- O que isso quer dizer?

P-valor

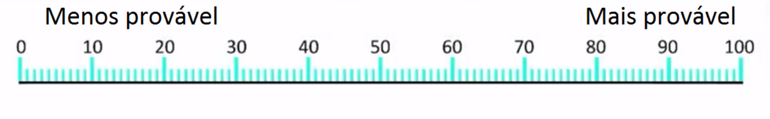
Intuição

- O que isso quer dizer?
- Se você observa, por exemplo, uma diferença na probabilidade de câncer entre fumantes e não fumantes, qual seria a probabilidade de você observar um valor absoluto igual ou maior do que esse, se a hipótese de que não há diferença for correta?
- Em outras palavras: qual a probabilidade de você encontrar uma diferença da magnitude medida (ou maior), se na verdade não existir diferença entre os grupos?
- Se esse valor for muito baixo, podemos rejeitar a hipótese nula - e provavelmente a diferença existe.

P-valor

Intuição

- O quão provável é que cheguemos a um resultado por fatores aleatórios se, de fato, a hipótese nula for verdadeira?



- Um resultado é considerado *estatisticamente significante* se for improvável que ele seja fruto puramente de fatores aleatórios.
- Se o p-valor for menor que 5%, podemos esperar que a associação de interesse estaria presente em 95% do casos, depois de várias amostragens seguidas.

P-valor

Intuição

- Qual teste pode ser feito?
- Depende do que estamos querendo testar...
  - Se o valor real é maior que o valor de corte (lado direito)?
  - Se o valor real é menor que o valor de corte (lado esquerdo)?
  - Se o valor real é maior ou menor que o valor de corte (bilateral)?

Hipótese alternativa	Evidência contra $H_0 : \theta = \theta_0$ fornecida por	P-valor
$H_a : \theta > \theta_0$	$\hat{\theta} - \theta_0$ mto positiva	$P = pr(X \geq x \vert H_0)$
$H_a : \theta < \theta_0$	$\hat{\theta} - \theta_0$ mto negativa	$P = pr(X \leq x \vert H_0)$
$H_a : \theta \neq \theta_0$	$\vert\hat{\theta} - \theta_0\vert$ mto grande	$P = 2 \cdot pr(X \leq -\vert x\vert \, \vert H_0)$

P-valor

Interpretação

- E como entender o resultado do p-valor?
- Não existe uma regra, mas a tabela abaixo traz uma aproximação de como, tradicionalmente, se faz essa interpretação:

P-valor	Interpretação
$p > 0.10$	Nenhuma evidência contra H_0
$0.05 < p \leq 0.10$	Evidência fraca contra H_0
$0.01 < p \leq 0.05$	Alguma contra H_0
$0.001 < p \leq 0.01$	Evidência forte contra H_0
≤ 0.01	Evidência muito forte contra H_0

- O importante é entender que quanto menor o p-valor, maior a força da evidência que temos para rejeitar H_0

P-valor

Aplicação

- Vamos extender nossa função de estatística-t para calcular um p-valor?

```
def p_valor(amostra, valor_corte):
    amostra = amostra[~np.isnan(amostra)]

    t_stat, media = t_stat_media(amostra, valor_corte)

    # Calcular p-value
    tt_stat = -1 * np.abs(t_stat)
    p_value = 2 * stats.t.cdf(tt_stat, df=(len(amostra)-1) )

    return t_stat, media, p_value

for i in np.linspace(0,25,26):
    est_t, media, p_val = p_valor(x, i)
    print("Média: {:.2f}; valor de corte: {:.2f}; estatística-t: {:.2f};
    p-valor: {:.2f}".format(media, i, est_t, p_val))
```

P-valor

Aplicação

- No lugar disso tudo, podemos simplesmente utilizar o scipy

```
from scipy import stats

media = np.mean(x)

for i in np.linspace(0,25,26):
    est_t, p_val = stats.ttest_1samp(x, i)
    print("Média: {:.2f}; valor de corte: {:.2f};
    estatística-t: {:.2f}; p-valor: {:.4f}"
          .format(media, i, est_t, p_val))
```

Sumário

Introdução ao teste de hipótese

Teste-t

- Uma hipótese pode ser: "O peso médio de homens ($\bar{p}_h$) é diferente do que o peso médio de mulheres ($\bar{p}_m$)"

$$
\begin{aligned}
H_0 &: \bar{p}_h - \bar{p}_m = 0 \\
H_A &: \bar{p}_h - \bar{p}_m \neq 0
\end{aligned}
$$

- Como calcular isso?

Introdução ao teste de hipótese

Teste-t

- Não é muito diferente do que já fizemos até agora.
- Calculamos uma estatística-t entre a diferença e o valor de corte (zero):

$$t_0 = \frac{(\bar{p}_h - \bar{p}_m) - 0}{\sqrt{\frac{s_h^2 + s_m^2}{n}}} \tag{2}$$

- Depois podemos simplesmente utilizar essa estatística-t e o número de graus de liberdade para encontrar um p-valor
- Importante: a fórmula acima presume que as amostras a serem comparadas têm variância e tamanho iguais. Em outros casos, a matemática é um pouco diferente.
- Na prática, o computador vai fazer o cálculo.

Introdução ao teste de hipótese

Teste-t

- Primeiro, vamos importar um arquivo do excel:

```
import pandas as pd

dfiq = pd.read_excel('https://github.com/omercadopopular/
cgoes/blob/master/StatsPython/
data/brain_size.xlsx?raw=true')

print(dfiq)
```

Quais são os problemas com essa importação?

- Valores faltando
- Nomes e unidades em inglês

Introdução ao teste de hipótese

Teste-t

- Primeiro, vamos importar um arquivo do excel, dizendo para o pandas que os valores faltantes estão sinalizados com um ponto:

```
import pandas as pd

dfiq = pd.read_excel('https://github.com/omercadopopular/
cgoes/blob/master/StatsPython/
data/brain_size.xlsx?raw=true', na_values=".")
```

- Depois, alteramos os nomes das colunas:

```
dfiq.columns = ['sexo', 'FSIQ', 'VIQ', 'PIQ', 'peso', 'altura', 'MRI_Count']
```

- Transformamos as escalas

```
lb_para_kg = lambda x: x / 2.2
in_para_cm = lambda x: x * 2.54

dfiq['peso'] = [lb_para_kg(pes) for pes in dfiq['peso']]
dfiq['altura'] = [in_para_cm(alt) for alt in dfiq['altura']]
```

Introdução ao teste de hipótese

Teste-t

- E alteramos os rótulos dos grupos:

```
dfiq['sexo'] = [string.replace("Female", "Feminino")
.replace("Male", "Masculino") for string in dfiq['sexo']]
```

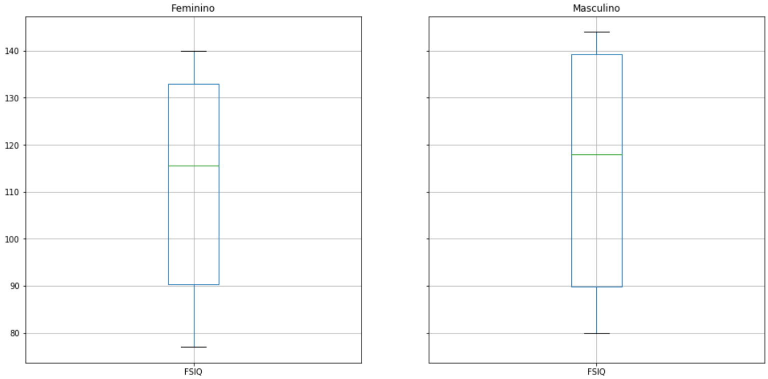
- A partir daí, usamos groupby para criar sumarizações com base nos cortes por "sexo":

```
grupos = dfiq.groupby('sexo')

print(grupos.mean(), grupos.median())

print(grupos.describe().T)

grupos.boxplot(column=['peso'])
grupos.boxplot(column=['FSIQ'])
```

Introdução ao teste de hipótese

Teste-t

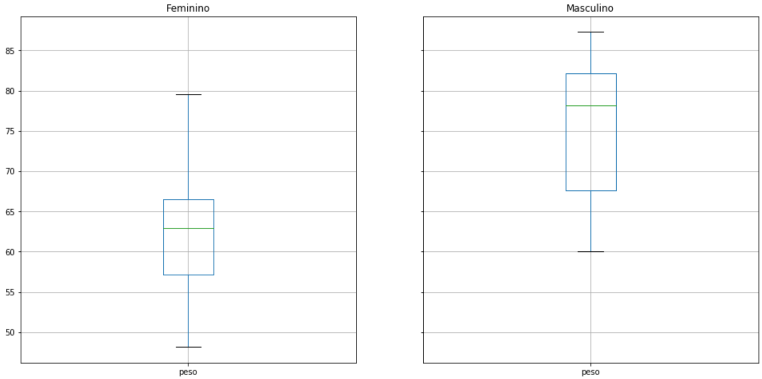
- Olhando para o box-plot, as diferenças em média de QIs parecem confiáveis?



Introdução ao teste de hipótese

Teste-t

- Olhando para o box-plot, as diferenças em média de peso parecem confiáveis?



Introdução ao teste de hipótese

Teste-t

- Como fazer achar um p-valor?
- Primeiramente, criamos dois DataFrames, com os dados de cada grupo.

```
from scipy import stats

masc = dfiq[ dfiq['sexo'] == 'Masculino'].dropna()
fem = dfiq[ dfiq['sexo'] == 'Feminino'].dropna()
```

- Depois, fazemos testes-t para independência de médias, utilizando o scipy.

```
t_stat, p_valor = stats.ttest_ind(masc['peso'], fem['peso'])
print("P-valor do teste-t de médias iguais de peso " +
      "para ambos os sexos: {:.2f}".format(p_valor))

t_stat, p_valor = stats.ttest_ind(masc['FSIQ'], fem['FSIQ'])
print("P-valor do teste-t de médias iguais de QI " +
      "para ambos os sexos: {:.2f}".format(p_valor))
```

Introdução ao teste de hipótese

Teste-f

- Outro teste que cumpre um papel similar é o Teste-F
- A maior diferença entre o Teste-F e o Teste-T é que ele serve para comparar as médias de *várias amostras diferentes*.
- Assim sendo, no Teste-F:

$$
\begin{aligned}
H_0 &: \mu_a = \mu_b = \ldots = \mu_N \\
H_a &: Nem \quad todas \quad as \quad médias \quad são \quad iguais
\end{aligned}
$$

Introdução ao teste de hipótese

Teste-f

- O teste-f é uma comparação de variâncias

$$F = \frac{variabilidade \quad entre \quad grupos}{variabilidade \quad dentro \quad dos \quad grupos}$$

- A intuição é a seguinte:
  - Quando as a variabilidade (desvio padrão) dentro dos grupos é pequena, a variabilidade entre grupos (diferença de médias domina), a estatística-f é maior, e as médias são concluídas como diferentes
  - Quando a variabilidade dentro dos grupos é muito grande, a estatística-f é menor, e as médias não podem ser concluídas como diferentes

Introdução ao teste de hipótese

Teste-f

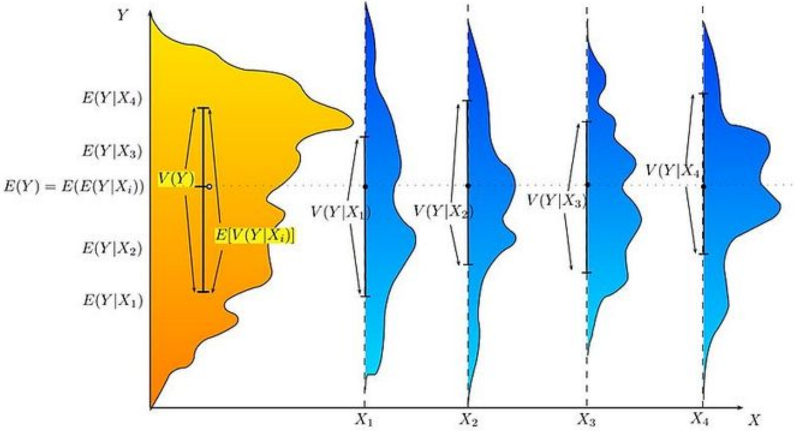


Figure 2: ANOVA : No fit

Introdução ao teste de hipótese

Teste-f

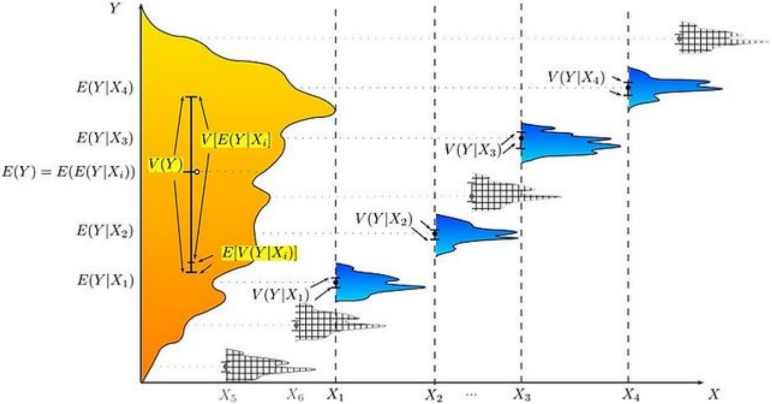


Figure 3: ANOVA : very good fit

Introdução ao teste de hipótese

Teste-f

- Formalmente, a estatística-f é calculada por

$$F = \frac{s_b^2}{s_w^2} \tag{3}$$

- em que:

$$s_b^2 = \frac{\sum_{i=1}^{K}(\bar{Y}_i - \bar{Y})^2}{K-1}, \quad para \quad K \quad grupos \tag{4}$$

- e:

$$s_w^2 = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(\bar{Y}_{ij} - \bar{Y}_i)^2}{N-K}, \quad para \quad N \quad obs \quad totais \tag{5}$$

Introdução ao teste de hipótese

Teste-f

- Podemos utilizar o scipy para calcular o teste-f

```
f_stat, f_pvalue = stats.f_oneway(masc['peso'], fem['peso'])
print("P-valor do F-teste de médias iguais de peso " +
      "para ambos os sexos: {:.2f}".format(f_pvalue))

f_stat, f_pvalue = stats.f_oneway(masc['FSIQ'], fem['FSIQ'])
print("P-valor do F-teste de médias iguais de peso " +
      "para ambos os sexos: {:.2f}".format(f_pvalue))
```

Introdução ao teste de hipótese

Teste-f

- Ou utilizar um outro pacote, que vai ser nossa principal ferramenda para regressão linear, chamado statsmodels

```
import statsmodels.formula.api as smf

model1 = smf.ols("peso ~ C(sexo)", data=dfiq.dropna()).fit()
print("P-valor do F-teste de médias iguais de peso " +
      "para ambos os sexos: {:.2f}".format(model1.f_pvalue))

model2 = smf.ols("FSIQ ~ C(sexo)", data=dfiq.dropna()).fit()
print("P-valor do F-teste de médias iguais de QI " +
      "para ambos os sexos: {:.2f}".format(model2.f_pvalue))
```