Lista 06 - Leonardo Santos Miranda

July 22, 2021

Questões 10, 19, 30 e 33 do capítulo 03 do livro de exercícios.

0.1 Questão 10

• Para verificar sua compreensão do problema, obtenha os números esperados que estão na Tabela 3.1.

[46]:
```
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt
from scipy.stats import chisquare
from scipy.stats import chi2_contingency

n, p = 8, 0.6#0.6 dado que é uma probabilidade >0.5
x = np.arange(0, n)
px = stats.binom.pmf(x,n,p)
print (px)
```

```
[0.00065536 0.00786432 0.04128768 0.12386304 0.2322432  0.27869184
 0.20901888 0.08957952]
```

• Calcule a estatística qui-quadrado neste problema (voce deve obter um valor de X2 = 91.87)

[47]:
```
esp = np.array([[165.22, 1401.69, 5202.65, 11034.65, 14627.60, 12409.87,6580.
↪24,1993.78,264.30]]).T
obs = np.array([[215, 1485, 5331, 10649, 14959, 11929,6678,2092,342]]).T
x2,array2 = chisquare(obs,esp)
test, p_value,dof, array = chi2_contingency(obs)
print ('x^2 =',x2,'|| p-value = ',1-p_value)
```

```
x^2 = [91.86947345] || p-value =  0.0
```

O primeiro valor retornado é o valor X^2 = 91.86947345, ou 91.87 arredondado. O segundo valor retornado é o valor P = 0, o que diz que os dados esperados e observados não veio de uma mesma distribuição • Qual a distribuição de referência desta estatística? Distribuição binomial • Qual o p-valor associado com esta estatística? O segundo valor retornado é o valor P = 0, o que diz que os dados esperados e observados não veio de uma mesma distribuição.

0.2 Questão 19

• Da Wikipedia: Zipf 's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Thus the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc. For example, in the Brown Corpus of American English text, the word "the"is the most frequently occurring word, and by itself accounts for nearly 7% of all word occurrences (69,971 out of slightly over 1 million). True to Zipf 's Law, the second-place word "of"accounts for slightly over 3.5% of words (36,411 occurrences), followed by "and"(28,852). Only 135 vocabulary items are needed to account for half the Brown Corpus. Explique como a equa¸c˜ao (3.2) implicaria que the frequency of any word is inversely proportional to its rank. A resposta ´e simples e direta, n˜ao tem nada sutil ou complicado aqui.

Resposta = Como a constante C (explicada no texto da pergunta) é dividida pelo K(rank) elevado a um número próximo de 1, isso implica que a palavra mais frequente vai aparecer aproxidamante duas vezes mais do que a 2° mais frequente, a 2° duas vezes mais frequente do que a 3° e etc...

• Use os dados da tabela acima para fazer o scatter-plot dos pontos (log(k), log(nk)). Em R, basta fazer summary(lm(y x)) onde y e x são os vetores com log(nk) e log(k), respectivamente. Qual o valor da inclinação?

Resposta = -0.999.

0.3 Questão 30

• Estime E(Y) usando a média aritmética X + 1 e obtenha assim uma estimativa de (theta)

Resposta = y = 1 + 0.9323 e (theta) = 0.696.

– Tabela –

Resposta dos valores esperados = Esp 336.96 117.16 54.36 28.37 23.74 Para (x >=5), P(X) = 1-0.585+0.20+0.09+0.04 = 0.745 Para (X>=5) = 0.745 * 576 = 23.745

• Embora seja óbvio que a distribuição logarítimica não se ajusta a estes dados, calcule a estatística qui-quadrado a partir das diferenças entre os valores observados e esperados nesta tabela.

[48]:
```
esp = np.array([229, 211, 93, 35, 7, 1]).T
obs = np.array([336.96, 117.16, 54.36, 28.37, 15.80,23.74]).T
x2,array2 = chisquare(obs,esp)
test, p_value,dof, array = chi2_contingency(obs)
print ('x^2 =',x2,'|| p-value = ',1-p_value)
```

```
x^2 = 638.1117802279664 || p-value =  0.0
```

• Obtenha o p-valor com o comando 1-pchisq(qq, df) onde qq ´e o valor da estat´ıstica quiquadrado e df ´e o n´umero de graus de liberdade.

Resposta = O p-valor é o segundo retorno da função anteriormente executada: 0, o que indica que os valores da tabela de observado e esperados nao foram obtidos usando a mesma distribuição.

0.4 Questão 33

[49]:
```
u = 120
o = 10
x = o*o
print (x-u/o)
```

```
88.0
```

[50]:
```
result=stats.norm.ppf(q=88,loc=0,scale=1)
print(result)
```

```
nan
```