Métodos de Estatística Aplicada com Python

Aula 9

Carlos Góes[1]

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

2017

Sumário

Sumário

Medidas de associação entre variáveis

Intuição

- Até agora o que vimos:
  - O que são parâmetros e estatísticas; populações e amostras;
  - Como resumir características de variáveis (média, mediana, moda, quantis, intervalos);
  - Como apresentar esses resumos graficamente;
  - Medidas de variabilidade (variância, desvio padrão);
  - Como entender probabilidades e funções probabilísticas;
  - Como medir a incerteza de nossas estimativas amostrais (erro padrão);
  - Como testar hipóteses quanto a médias e intervalos.

Medidas de associação entre variáveis

Intuição

- O que tudo isso tem em comum?
    - Todas elas trabalham com formas diferentes de descrever uma só variável.
    - Agora, vamos trabalhar com medidas de associação entre duas ou mais variáveis.

Sumário

Teste do Qui-Quadrado

Definição

- O teste do qui-quadrado serve para medir a distribuição de proporções de variáveis categóricas diferentes
- Ele tem duas funções, que vamos ver a seguir
    - A mensuração da *homogeneidade de distribuições* entre populações diferentes.
    - A mensuração da *independência* entre categorias diferentes e grupos específicos de uma mesma população.

Teste do Qui-Quadrado
Definição

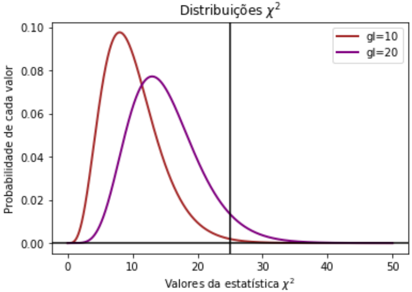
- O teste do qui-quadrado tem esse nome porque a estatística que construímos para o teste segue uma distribuição $\chi^2(gl)$, em que gl é o número de graus de liberdade.
- O teste se define da da seguinte maneira:

$$
\begin{aligned}
X^2 &= \sum_{n=1}^{N} \frac{(contagem \quad observada_i \quad - \quad contagem \quad esperada_i)^2}{contagem \quad esperada_i} \\
X^2 &= \sum_{n=1}^{N} \frac{(O_i - E_i)^2}{E_i} \qquad (1)
\end{aligned}
$$

Teste do Qui-Quadrado

Intuição

- O teste mede o quadrado da distância entre a contagem esperada e a contagem observada (numerador), normalizado pela contagem esperada (denominador).
- Essa estatística vai ser comparada com a distribuição (dados os graus de liberdade) e, assim como em outros teste de hipótese, vamos ver qual é a proporção acima da estatística de corte:



chi2.png

Teste do Qui-Quadrado

Teste de Uniformidade

- A primeira coisa para que podemos utilizar um teste de qui-quadrado é para tester se a distribuição de categorias diferentes (ou de populações diferentes) é uniforme.
- Assim sendo, a hipótese nula seria:

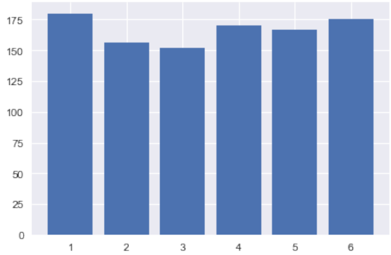
$$
\begin{aligned}
H_0 &: f_a = f_b = \ldots = f_N \\
H_a &: \neg H_0 \quad (nem \quad todas \quad as \quad frequencias \quad são \quad iguais)
\end{aligned}
$$

- Como sempre, se rejeitarmos a hipótese nula (H_0), aceitamos a hipótese alternativa

Teste do Qui-Quadrado

Teste de Uniformidade

- Vamos voltar ao nosso famoso exemplo do dado.
- Sabemos que para essa distribuição vai ser uniforme, por construção.



- Portanto, se fizermos um teste de qui-quadrado nessa distribuição de frequências, não poderemos rejeitar a hipótese nula de que todas as frequências são iguais (ou seja, a distribuição é uniforme)

Teste do Qui-Quadrado

Teste de Uniformidade

- Plotar o histograma e gravar a distribuição:

```
tamanho = 1000

dado = [random.randint(1,6) for i in range(0,tamanho)]

dist, bins, graf = plt.hist(dado,
                   bins=[0.5+i for i in range(0,7)],
                   rwidth=0.8)
plt.show()
```

- Usar a distribuição para fazer o teste:

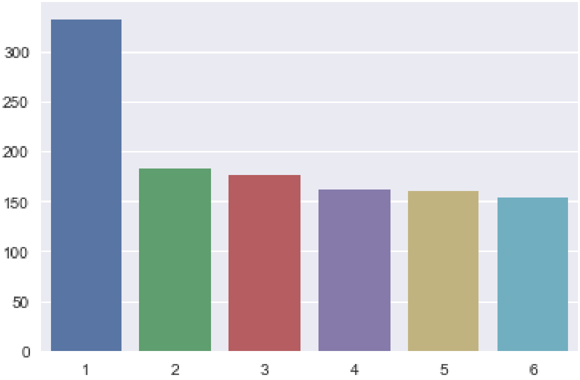
```
print('Distribuição: {} \n'.format(dist) + \
      str(stats.chisquare(dist)))
```

- Como o $p > 0.1$, não há evidência alguma de para rejeitar a hipótese nula.

Teste do Qui-Quadrado

Teste de Uniformidade

- Mas e se nosso dado fosse viciado, com maior frequência de seis?



Teste do Qui-Quadrado

Teste de Uniformidade

- Plotar o histograma e gravar a distribuição:

```
dist[0] = 2*dist[0]
eixox = list(range(1,7))
sns.barplot(x=eixox, y=dist)
```

- Usar a distribuição para fazer o teste:

```
print('Distribuição: {} \n'.format(dist) + \
      str(stats.chisquare(dist)))
```

- Como o $p < 0.00$, há fortíssima evidência alguma de para rejeitar a hipótese nula.

Teste do Qui-Quadrado

Teste de Independência

- Imagine que uma população ou amostra de divida, integralmente, em várias características mutuamente excludente.
- Imagine também que se somarmos todas essas características, chegamos ao conjunto total da população:
  - Exemplo: raça dos brasileiros.
  - $P(negro \quad ou \quad branco \quad ou \quad outro) = 1$
- Imagine ainda que podemos ter subgrupos dessa população e, em cada um dos subgrupos, os indivíduos também se dividem nessas categorias.
  - Exemplo: estados brasileiros.

Teste do Qui-Quadrado

Teste de Independência

- Se os subgrupos forem independentes das categorias, o que deveríamos esperar?
  - Que a proporção de cada categoria no total da população fosse mais ou menos igual em cada subgrupo
  - Ou seja, que a proporção de cada categoria *não dependa* do subgrupo que esteja sob análise.

Teste do Qui-Quadrado

Teste de Independência

- A hipótese nula é, portanto, que, para cada grupo $g = \{1, \ldots G\}$ e categoria $c = \{1, \ldots C\}$:

$$
\begin{aligned}
H_0 &: p_{1,c} = \ldots = p_{G,c} \quad \forall \quad c \\
H_a &: \neg H_0
\end{aligned}
$$

- Traduzindo:

$$
\begin{aligned}
H_0 &: p_{am,raça} = p_{df,raça} = p_{pb,raça} = \ldots = p_{rs,raça} \quad \forall \quad raça \\
H_a &: \neg H_0 \quad (Não \quad H_0)
\end{aligned}
$$

Teste do Qui-Quadrado

Teste de Independência

- Vamos usar dados do Censo da Educação Superior e ver se a distribuição racial é independente dos cursos.
- Primeiro carregamos os dados

```
arquivo = r"C:\Users\CarlosABG\Documents\IESB\aula 9\DM_ALUNO\cesdf.csv"
cesdf = pd.read_csv(arquivo, sep='|', encoding='latin_1')
print(cesdf)
```

- Podemos ver quantos estudantes estão listados:

```
cesdf.shape
```

- Quantos alunos há em cada curso:

```
cesdf.groupby('NO_CURSO').count()
```

- E listar cursos específicos:

```
cesdf.groupby('NO_CURSO').count().loc[['DIREITO', 'MEDICINA',
'CIÊNCIAS ECONÔMICAS', 'ESTATÍSTICA']]
```

Teste do Qui-Quadrado

Teste de Independência

- Consultando o dicionário de daodos, sabemos o que significam os números que representam a cor/raça dos estudantes:

34	29	CO_COR_RACA_ALUNO	Código da cor/raça do aluno	Num	8	1. Branca 2. Preta 3. Parda 4. Amarela 5. Indígena 6. Não dispõe da informação 0. Aluno não quis declarar cor/raça

Teste do Qui-Quadrado

Teste de Independência

- Com isso, construímos um dicionário com essas informações, e substituímos os números por strings:

```
racedict = {
    1: 'Branca',
    2: 'Preta',
    3: 'Parda',
    4: 'Amarela',
    5: 'Indígena',
    6: np.nan,
    0: np.nan
    }

cesdf['CO_COR_RACA_ALUNO'] = [racedict[int(aluno)]
for aluno in cesdf['CO_COR_RACA_ALUNO']]
```

Teste do Qui-Quadrado

Teste de Independência

- Finalmente, construímos uma tabela que faz uma contagem de frequência por raça e curso:

```
tabulacao = pd.crosstab(cesdf['NO_CURSO'], cesdf['CO_COR_RACA_ALUNO'])
print(tabulacao)
```

- Excluímos os cursos pequenos (menor que 5000 estudantes):

```
tabulacao = tabulacao[ tabulacao.sum(axis=1) > 5000 ]
print(tabulacao)
```

- E fazemos o teste de qui-quadrado:

```
chi2, p, ddof, expected = stats.chi2_contingency(tabulacao)
print('Teste de Chi-Quadrado \n' + \
      'Chi-quadrado: {:.2f} p-valor: {:.2f} \n'.format(chi2, p) +
      'Graus de liberdade: {}'.format(ddof))
```

- Como o $p < 0.00$, há fortíssima evidência alguma de para rejeitar a hipótese nula.

Teste do Qui-Quadrado

Uniformidade = Independência?

- Há um jeito simples de compreender intuitivamente como esses dois conceitos estão relacionados.
- Vamos primeiro calcular a porcentagem de negros em cada curso:

```
cesdf['NEGRO'] = (cesdf['CO_COR_RACA_ALUNO'] == 'Preta')
negro = cesdf.groupby('NO_CURSO')['NEGRO'].mean().sort_values()
print(negro)
```

- E plotar pontos que representam os cursos:

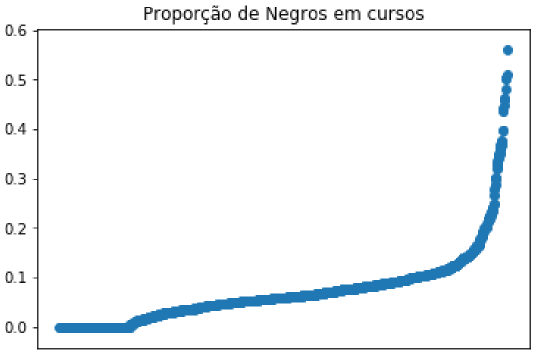
```
y_pos = range(0, len(negro))

plt.scatter(y_pos, negro)
plt.title('Proporção de Negros em cursos')
plt.tick_params(
    axis='x',
    bottom='off',
    labelbottom='off')
plt.show()
```

Teste do Qui-Quadrado

Uniformidade = Independência?

- O resultado é esse:



- O que isso significa?
- Se a distribuição fosse uniforme, todas as bolhas estariam mais ou menos na mesma altura.

Sumário

Covariância

Definição

- Vocês lembram de como se calcula a variância de uma amostra?

$$var(y) = \frac{\sum_{n=1}^{N}(y_n - \bar{y})^2}{N-1} \tag{2}$$

- Perceba que:

$$var(y) = \frac{\sum_{n=1}^{N}(y_n - \bar{y})(y_n - \bar{y})}{N-1} \tag{3}$$

Covariância

Definição

- Covariância é uma extensão desse conceito:

$$covar(x, y) = \frac{\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{N - 1} \tag{4}$$

- O que isso significa?
    - Se os valores mais altos de x ($+$ acima de $\bar{x}$) corresponderem aos mais altos de y ($+$ acima de $\bar{y}$), enquanto os mais baixos de x corresponderem aos mais baixos de y a covariância será positiva (x, y terão associação positiva).
    - Se, ao contrário, os valores mais altos de x corresponderem aos mais baixos de y, enquanto os mais baixos de x corresponderem aos mais altos de y, a covariância será negativa (x, y terão associação negativa).

Covariância

Aplicação

- Vamos utilizar uma base de dados de passagens aéreas para tentar analisar covariância entre distância e preços
- Carregue a base de dados (que está no formato do programa estatístico Stata) e veja o cabeçalho:

```
file = 'https://github.com/omercadopopular/cgoes/blob/master/StatsPython/
data/wooldridge/airfare.dta?raw=true'
df = pd.read_stata(file)
print(df.head())
```

- Vamos excluir as variáveis que não interessam:

```
df = df.drop(['ldist', 'y98', 'y99', 'y00', 'lfare',
             'ldistsq', 'concen', 'lpassen'], axis=1)
```

- E alterar o nome de uma variável para português:

```
df = df.rename(columns = {'fare':'preco'})
print(df.head())
```

Covariância

Aplicação

- Como calcular a covariância entre preços e distância?
- Primeiro escrevemos uma fórmula de covariância:

```
def cov(x,y):
    if len(x) != len(y):
        return 'Variáveis de tamanho diferente'

    else:
        media_x = np.mean(x)
        media_y =  np.mean(y)

        numerador = np.sum((x - media_x) * (y - media_y))
        denominador = len(x) - 1

        return numerador / denominador
```

- Depois a aplicamos:

```
print('Variância dos preços: {:.2f} \n'.format(cov(df['preco'], df['preco'])) +
      'Variância das distâncias: {:.2f} \n'.format(cov(df['dist'], df['dist'])) +
      'Co-variância de preço e distâncias: {:.2f} \n'.format(cov(df['preco'], df['dist']))
     )
```

Covariância

Aplicação

- Ou podemos simplesmente utilizar o numpy:

```
np.cov(df['preco'],df['dist'])
```

Correlação

Definição

- Um jeito mais simples de entender essa relação de associação positiva é transformar a covariância entre duas variáveis num índice de correlação.
- O índice de correlação, denominado pela letra grega ρ (pronúncia: "rô"), é a covariação *normalizada*, de tal modo que

$$\rho_{x,y} \in \{-1, 1\}$$

Correlação

Definição

- Como que essa padronização é feita?
- Divindo a covariância pelo produto dos desvios padrão de x (s_x) e y (s_y):

$$\rho_{x,y} = \frac{cov(x,y)}{s_x s_y} \tag{5}$$

Correlação

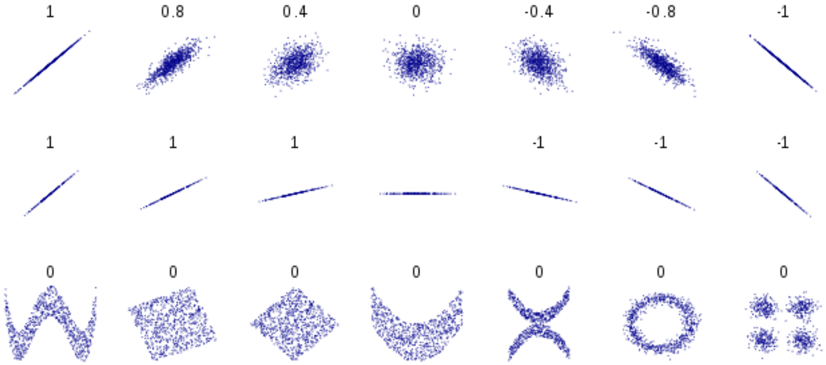
Intuição

- O que esse valor de -1 a $+1$ significa?
  - $+1$ é a correlação positiva máxima, de modo que uma série é uma transformação linear da outra (ex: $2x$ comparado com x)
  - -1 é a correlação negativa máxima, de modo que uma série é o inverso uma transformação linear da outra (ex: $-4x$ comparado com x).
  - Valores intermediários representam graus distintos de associação:

Correlação

Intuição

- Um jeito fácil de entender a correlação é por meio de diagramas de dispersão:



Covariância

Aplicação

- Vamos extender nosso programa para calcular a correlação:

```
def corr(x,y):
    numerador = cov(x,y)
    denominador = np.std(x, ddof=1) * np.std(y, ddof=1)

    return numerador / denominador
```

- E calcular o índice:

```
corr(df['preco'],df['dist'])
```

- Ou simplesmente usar o numpy:

```
    np.corrcoef(df['preco'],df['dist'])
```

Covariância

Aplicação

- Por último, podemos ver a relação entre as duas variáveis graficamente, por meio de um diagrama de dispersão:

```
plt.scatter('dist', 'preco',
            data=df, alpha=0.25)
plt.xlabel('Distância (milhas)')
plt.ylabel('Preço (dólares)')
plt.title('Passagens aéreas: relação entre distância e preço')
plt.show()
```

Correlação

Intuição

- Impressão da máquina: