Estatística Não Paramétrica

Teste de aderência Qui-quadrado e Kolmogorov-Smirnov (KS)

Danielle Ribeiro e Luiz Fernando Coelho Passos

Teste de Aderência

Os testes de aderência ou de qualidade de ajuste consistem em testar a adequabilidade de um modelo probabilístico a um conjunto de dados, onde

H_0: A população tem uma distribuição especificada (P = P_0)
H_1: A população não tem a distribuição especificada

Em que a distribuição especificada pode ser discreta ou contínua, com os valores dos parâmetros especificados, ou não, em H_0. O objetivo desse teste é saber se a distribuição de probabilidade considerada em H_0 é um modelo adequado à população amostrada.

Para realizar um teste de aderência é necessário se basear nos pressupostos de que a amostra aleatória obtida é independente e identicamente distribuída e de tamanho relativamente grande.

Teste Qui-Quadrado

É utilizado para testar a qualidade do ajuste, ao comparar a distribuição das frequências observadas (na amostra) com as frequências esperadas (obtidas sob a veracidade de H_0).

Quando trabalhamos com variáveis discretas e com variáveis qualitativas, a variável de interesse χ^2 está organizada em categorias ou classes $A_1, A_2, ..., A_n$, com probabilidades $p_i = P(\chi^2 \in Ai), i = 1, ..., n$.

Onde as hipóteses são,

$H_0 : p_1 = p_{10},\ p_2 = p_{20},\ \ldots\ ,\ p_n = p_{n0}$
$H_1 : p_j \neq p_{j0}$, para pelo menos um j, onde,

p_{i0} = probabilidade associada à categoria i, i = 1, . . . , k, calculada assumindo o modelo probabilístico em H_0.

O modelo probabilístico adequado para esta situação é o modelo multinomial.

Se χ^2 for uma v.a. contínua, podemos dividir o seu domínio de variação em intervalos e construir a distribuição de frequências correspondente.

Estatística de Teste

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad \sim \quad \chi^2_{(n-1)}$$

onde,

O_i = valor efetivamente observado para a classe A_i
E_i = valor esperado para a classe A_i, sob a veracidade de H_0,

$$E_i = np_{i0},\ i = 1, 2, ..., n$$

Região Crítica

$$RC = \{(c, \infty) \ \mid \ P(\ \chi^2_{(n-1)} > c) = \alpha\}$$

Quanto maior for a discrepância entre as frequências observadas e as frequências esperadas, maior é o valor observado da estatística de teste, e portanto, rejeita-se H_0, pois a região crítica é uma cauda à direita.

Exemplo 1 - v.a. discretas

Um estudo sobre a distribuição dos acidentes de trabalho numa indústria nos cincos dias da semana revelou que, em 150 acidentes:

```
dias = c("segunda","terça","quarta","quinta","sexta")
acidentes = c(30,40,20,25,33)
df = tibble(dias,acidentes)
df
```

```
## # A tibble: 5 x 2
##   dias    acidentes
##   <chr>       <dbl>
## 1 segunda        30
## 2 terça          40
## 3 quarta         20
## 4 quinta         25
## 5 sexta          33
```

O objetivo é testar a hipótese que os acidentes ocorrem com igual frequência nos cinco dias da semana (considere $\alpha = 5\%$).

Nossas hipóteses são:

$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = 1/5$

$H_1 : p_j \neq 1/5$, para pelo menos um j

```
Oi <- c(seg=32, ter=40, qua=20, qui=25, sex=33)
# esperados sob H_0
Ei <- sum(Oi)*1/length(Oi)
# estatística do teste
X2 <- sum((Oi-Ei)^2/Ei)
# graus de liberdade
nu <- length(Oi)-1
# p-valor do teste
pchisq(X2, df=nu, lower.tail=FALSE)
```

```
## [1] 0.09405103
```

Usando a função `chisq.test()`, temos:

```
chisq.test(Oi)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  Oi
## X-squared = 7.9333, df = 4, p-value = 0.09405
```

Conclusão

O p-valor obtido foi de 0.09405. Assim, não rejeitamos H_0 ao nível de significância de 5%, ou seja, podemos considerar que os acidentes ocorrem com igual frequência nos 5 dias da semana.

Exemplo 2 - v.a. contínuas

Considere os dados abaixo, que supostamente são uma amostra de tamanho n = 30 de uma distribuição normal, de média $\mu = 10$ e variância $\sigma^2 = 25$.

Nossas hipóteses são:

$H_0 : P = N(10, 25)$

$H_1 : P \neq N(10, 25)$

Os dados já estão ordenados:

```
df <- c(1.04,   1.73,   3.93,   4.44,   6.37,   6.51,
        7.61,   7.64,   8.18,   8.48,   8.57,   8.65,
        9.71,   9.87,   9.95,  10.01,  10.52,   10.69,
        11.72,  12.17,  12.61,  12.98,  13.03,  13.16,
        14.11,  14.60,  14.64,  14.75,  16.68,  22.14)
```

Categorizaremos os dados do exemplo da seguinte forma:

```
Oi <- c()
Oi[1]<-sum(df < 6.63)
Oi[2]<-sum(df > 6.63 & df < 10)
Oi[3]<-sum(df > 10 & df < 13.37)
Oi[4]<-sum(df > 13.37)

# esperados sob H_0
Ei <- sum(Oi)*1/length(Oi)
# estatística do teste
X2 <- sum((Oi-Ei)^2/Ei)

# Tabela
cbind(
  rbind(Oi, Ei),
  c(sum(Oi), Ei*4)
) %>% `colnames<-`(c("(-Inf;6.63]", "(6.63;10]", "(10;13.37]", "(13.37;+Inf)", "Total"))
```

```
##    (-Inf;6.63] (6.63;10] (10;13.37] (13.37;+Inf) Total
## Oi         6.0       9.0        9.0          6.0    30
## Ei         7.5       7.5        7.5          7.5    30
```

```
# graus de liberdade
nu <- length(Oi)-1
# p-valor do teste
pchisq(X2, df=nu, lower.tail=FALSE)
```

```
## [1] 0.7530043
```

Usando a função `chisq.test()`, temos:

```
chisq.test(Oi)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  Oi
## X-squared = 1.2, df = 3, p-value = 0.753
```

Conclusão

O p-valor obtido foi de 0.753. Assim, não rejeitamos H_0 ao nível de significância de 5%, ou seja, podemos considerar que temos uma amostra de uma normal com média 10 e variância 25.

Teste Kolmogorov-Smirnov

Também é usado para testar a adequabilidade de um modelo probabilístico a um conjunto de dados. Supondo que tenhamos uma amostra $X_1, ..., X_n$ de uma população P, sobre a qual estamos considerando uma v.a. X. Designemos por $f(x)$ a função densidade e por $F(x)$ a função de distribuição acumulada (f.d.a.) de X. Estimar $f(x)$ é equivalente a estimar $F(x)$ e o objetivo é testar se a amostra observada veio de uma distribuição de probabilidades especificada. Ou seja,

$H_0 : F(x) = F_0(x)$, para todo x.
H_1: caso contrário

Estatística de Teste

Considerando a função de distribuição empírica (f.d.e.), $F_e(x)$ como um estimador de $F(x)$, para todo valor x real. Espera-se que se $F_e(x)$ for um bom estimador de F(x) a curva das duas funções devem estar próximas. A partir disso, deriva-se a seguinte estatística de teste,

$$D = máx\ |F(x_i)\check{}F_e(x_i)|, \quad i = 1, ..., n$$

onde,

$F(x_1)$ representa a função de distribuição acumulada assumida para os dados (calculado sob a hipótese nula H_0)
$F_e(x_1)$ representa a função de distribuição acumulada empírica dos dados.

O valor encontrado na estatística de teste deve ser comparado com um valor crítico, fixado um nível de significância do teste. Se D for maior que o valor tabelado, rejeitamos H_0.

Exemplo

Retomemos o `Exemplo 2`, onde queríamos testar se 30 valores observados provinham de uma distribuição normal, com média 10 e variância 25.

Nossas hipóteses são:

$H_0 : F(x) = F_0(x), \forall x$

$H_1 : F(x) \neq F_0(x),$ para algum x,

onde $F_0(x)$ é a função de densidade acumulada da v.a. $X \sim N(10, 25)$.

O teste de Kolmogorov-Smirnov (KS) é calculado, no `R`, através da função `ks.test`. Assim, para testar se os valores são aderentes à distribuição normal:

```
ks.test(df, y="pnorm", mean=10, sd=5)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  df
## D = 0.11633, p-value = 0.769
## alternative hypothesis: two-sided
```

Conclusão

O p-valor obtido foi de 0.769. Assim, não rejeitamos H_0 ao nível de significância de 5%, ou seja, podemos considerar que temos uma amostra de uma normal com média 10 e variância 25.

Referência

- MORETTIN, Pedro A.; BUSSAB, Wilton O. **Estatística Básica**. São Paulo: Saraiva, 2010. cap. 14, 399-419.