MÉTODOS ESTATÍSTICOS

Testes de Hipóteses Não Paramétricos - Parte 1
Testes de Ajustamento

Licenciatura em Engenharia Informática

Departamento de Matemática
Escola Superior de Tecnologia de Setúbal
Instituto Politécnico de Setúbal
2021-2022

Testes de Hipóteses Não Paramétricos:

Testes de Ajustamento

- Os testes de ajustamento servem para testar a hipótese de que uma determinada amostra aleatória foi recolhida de uma população com uma determinada distribuição.
- Um teste deste tipo compara a hipótese nula (H_0) com a hipótese alternativa (H_1), tendo a seguinte forma:
  - H_0- Os dados provêm da população com a distribuição especificada
  - H_1- Os dados não provêm da população com a distribuição especificada

Princípios Básicos na Realização dos Testes de Ajustamento

1. São definidas duas **hipóteses**:
   - **Hipótese Nula** $= H_0$ - é a hipótese que indica a distribuição que se pretende testar.
   - **Hipótese Alternativa** $= H_1$ - é a hipótese que se contrapõe à hipótese nula, ou seja, que indica que a distribuição que foi colocada na hipótese nula não é válida.
2. É definida uma **Estatística Teste**, que é a base da realização do teste e consiste em comparar a amostra com o modelo teórico.
3. São construídas duas regiões:
   - **Região de Aceitação** $= RA$ - conjunto de valores para os quais H_0 é admissível.
   - **Região de Rejeição ou Região Crítica** $= RC$ - conjunto de valores para os quais H_0 não é admissível.

Princípios Básicos na Realização dos Testes de Hipóteses

4. A **regra de decisão** define as condições de rejeição ou não rejeição da hipótese nula:
   - Se o Valor Observado da Estatística de Teste sob a hipótese H_0 pertencer à Região de Aceitação, então Não se Rejeita H_0
   - Se o Valor Observado da Estatística de Teste sob a hipótese H_0 pertencer à Região Crítica, então Rejeita-se H_0
5. **Erros de decisão** - um teste de hipóteses nem sempre conduz a decisões corretas, a análise de uma amostra pode falsear as conclusões quanto à população. Um dos erros é o chamado **Erro de 1^a espécie** ou **Nível de significância do teste**:

$$\alpha = P\left[\text{rejeitar } H_0 \mid H_0 \text{ verdadeira}\right]$$

   para minimizar este erro fixa-se o seu valor.
6. As regiões de aceitação e de rejeição $(RA$ e $RC)$ são definidas à custa do valor fixado para o nível de significância (α).

Na prática, em vez de calcular a região crítica (RC) e a região de aceitação (RA), é usual calcular-se o **Valor-p** (ou **p-value**).

Valor-p (ou p-value)

É a probabilidade associada ao valor da estatística de teste, considerando H_0 verdadeira.

- Se o valor-p for pequeno significa que, no caso de H_0 ser verdadeira, estamos perante um evento muito raro, pouco provável de ocorrer, então deve optar-se por rejeitar H_0.

  Portanto, o valor-p também permite tomar decisões:

  - se valor-p $\leq \alpha$, então rejeita-se H_0
  - se valor-p $> \alpha$, então não se rejeita H_0

Testes de Ajustamento

- Existem diversos testes de ajustamento, só vamos ver os seguintes:
  - **Teste de ajustamento do Qui-Quadrado**
    - é válido para distribuições discretas e contínuas
    - amostras grandes
    - pode haver a necessidade de agrupar os dados
  - **Teste de ajustamento de Kolmogorov-Smirnov**
    - é válido para distribuições contínuas, muito usado para testar a distribuição Normal (também existe uma versão para distribuições discretas mas não será estudada)
    - pode ser aplicado a amostras pequenas
    - não é necessário agrupar os dados

Teste de ajustamento do Qui-Quadrado

Objetivo

Testar a adequabilidade de um modelo probabilístico a um conjunto de dados observados, ou seja, comparar a distribuição dos dados amostrais (frequências observadas) com a distribuição teórica que se associa à população de onde provém essa amostra.

Teste de ajustamento do Qui-Quadrado

Formulação das Hipóteses a Testar:

H_0- Os dados provêm da população com a distribuição teórica especificada

vs

H_1- Os dados não provêm da população com a distribuição teórica especificada

Estatística de Teste

A estatística de teste tem por base os desvios entre as frequências observadas (O_i) e esperadas (E_i). Supondo verdadeira a hipótese H_0, então

$$Q=\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}\sim\chi^2_{(k-1-r)}$$

onde r representa o número de parâmetros desconhecidos da distribuição proposta em H_0, estimados a partir da amostra.

Teste de ajustamento do Qui-Quadrado

Cálculo do Valor Observado da Estatística de Teste sob a Hipótese H_0

$$Q_{obs} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

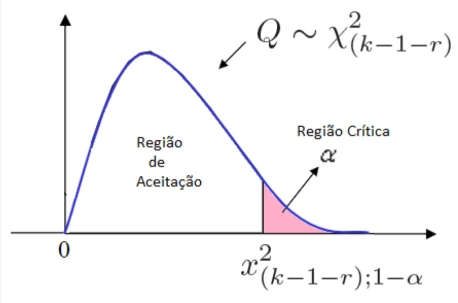
- De acordo com o tipo de dados da amostra, construir a respetiva tabela de frequências;
- k corresponde ao número de linhas da tabela de frequências
- **frequências observadas** $= O_i \rightarrow$ corresponde às frequências absolutas, n_i, das tabelas de frequências;
- **frequências esperadas** $= E_i = np_i \rightarrow$ frequência absoluta esperada referente à categoria ou classe i se H_0 for verdadeira, sendo
  - n é a dimensão da amostra
  - p_i a probabilidade da variável aleatória, com o modelo probabilístico definido na hipótese H_0, pertencer à categoria ou classe i.

Observação: Tem-se $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i = n$

Teste de ajustamento do Qui-Quadrado

Definição da Região de Aceitação e de Região Crítica

Um valor da estatística de teste Q elevado indica um desajuste entre a distribuição de frequências amostral e teórica:



- a Região de Aceitação é $RA = \left[0, x^2_{(k-1-r);1-\alpha}\right[$
- a Região Crítica é $RC = \left[x^2_{(k-1-r);1-\alpha}, +\infty\right[$

Teste de ajustamento do Qui-Quadrado

Regra de Decisão com base na Região Crítica

- Se o valor observado da estatística de teste não pertencer à Região Crítica,

$$Q_{obs} \notin RC$$

então, ao nível de significância α, **a hipótese** H_0 **não é rejeitada**, isto é, com base na amostra há evidências estatísticas que os dados provêm de uma população que possui a distribuição teórica definida na hipótese H_0.

- Se o valor observado da estatística de teste pertencer à Região Crítica,

$$Q_{obs} \in RC$$

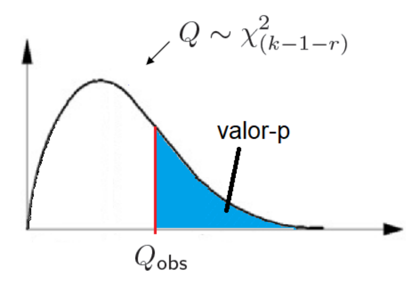
então, ao nível de significância α, **a hipótese** H_0 **é rejeitada**, isto é, com base na amostra há evidências estatísticas que os dados não provêm de uma população que possui a distribuição teórica definida na hipótese H_0.

Teste de ajustamento do Qui-Quadrado

Cálculo do valor-p

Considerando que H_0 é verdadeira, o valor-p indica a probabilidade do valor observado da estatística de teste ocorrer:

$$\textbf{valor-p} = P\left(Q \geq Q_{\text{obs}}\right)$$



O valor-p pode ser visto como o menor valor de α (nível de significância) para o qual os dados observados indicam que H_0 deve ser rejeitada.

Teste de ajustamento do Qui-Quadrado

Regra de Decisão com base no valor-p

- Se
$$\text{valor-p} > \alpha$$
então, ao nível de significância α, **a hipótese** H_0 **não é rejeitada**, isto é, com base na amostra há evidências estatísticas que os dados provêm de uma população que possui a distribuição teórica definida na hipótese H_0.

- Se
$$\text{valor-p} \leq \alpha$$
então, ao nível de significância α, **a hipótese** H_0 **é rejeitada**, isto é, com base na amostra há evidências estatísticas que os dados não provêm de uma população que possui a distribuição teórica definida na hipótese H_0.

Teste de ajustamento do Qui-Quadrado

Condições de aplicação do teste

- Não há mais de 20% das frequências esperadas inferiores a 5, isto é, $E_i < 5$ no máximo em 20% das células dos E_i.
- Todas as frequências esperadas devem ser maiores ou iguais a 1, isto é, $E_i \geq 1$ para todo $i = 1, \ldots, k$.
- Alguns autores acrescentam ainda que a dimensão da amostra deve ser maior que 20 e outros que deve ser maior que 30.

Observações

- As regras de aplicação do teste de ajustamento do Qui-Quadrado devem ser, tanto quanto possível verificadas, sob pena do teste não ser rigoroso. Ou seja, o teste de ajustamento do Qui-Quadrado não tem qualquer pressuposto, a infração das regras de aplicação apenas leva à perda de rigor.
- Quando as frequências esperadas não atingem os valores aconselhados agregam-se classes adjacentes de forma a obter novas classes que satisfaçam as condições.

Teste de ajustamento do Qui-Quadrado

Exemplo 1

Deseja-se verificar se um dado é equilibrado (não viciado), para tal lançou-se o dado 210 vezes e os resultados obtidos foram:

face do dado	Número de vezes que saiu a face
1	46
2	35
3	25
4	19
5	40
6	45

Teste a hipótese referida considerando um nível de significância de 5%.

Seja

$X-$ número da face virada para cima num lançamento de um dado com $D_X = \{1, 2, 3, 4, 5, 6\}$

Pretende-se verificar se o dado é equilibrado (não viciado),

ou seja,

Pretende-se verificar se a variável X segue uma distribuição Uniforme Discreta:

$$X \sim U_{(6)}$$

Hipótese a ser testada

Seja X a variável aleatória que representa o número da face virada para cima num lançamento de um dado

$$H_0 : X \sim U_{(6)} \qquad vs \qquad H_1 : X \nsim U_{(6)}$$

Dados

- Total de dados: $n = 46 + 35 + 25 + 19 + 40 + 45 = 210$
- Distribuição Uniforme discreta: $f(x) = P(X = x) = \begin{cases} \frac{1}{6} & , \; x \in \{1, 2, \ldots, 6\} \\ 0 & , \; \text{caso contrário} \end{cases}$
- Não é necessário estimar parâmetros: $r = 0$
- nível de significância $= \alpha = 0.05$

Construir a tabela de frequências observadas e esperadas

Domínio x_i	Frequências Observadas $O_i = n_i$	Probabilidade $p_i = f(x_i)$	Frequências Esperadas $E_i = n \times p_i$	Valor Observado da Estatística de Teste $\frac{(O_i - E_i)^2}{E_i}$
1	46	$f(1) = \frac{1}{6}$	$210 \times \frac{1}{6} = 35$	$\frac{(46-35)^2}{35} = 3.46$
2	35	$f(2) = \frac{1}{6}$	$210 \times \frac{1}{6} = 35$	$\frac{(35-35)^2}{35} = 0$
3	25	$f(3) = \frac{1}{6}$	$210 \times \frac{1}{6} = 35$	$\frac{(25-35)^2}{35} = 2.86$
4	19	$f(4) = \frac{1}{6}$	$210 \times \frac{1}{6} = 35$	$\frac{(19-35)^2}{35} = 7.31$
5	40	$f(5) = \frac{1}{6}$	$210 \times \frac{1}{6} = 35$	$\frac{(40-35)^2}{35} = 0.71$
6	45	$f(6) = \frac{1}{6}$	$210 \times \frac{1}{6} = 35$	$\frac{(45-35)^2}{35} = 2.86$
	$n = 210$	1	$n = 210$	$Q_{obs} = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = 17.2$

A estatística de teste, sob a hipótese H_0, tem distribuição Qui-Quadrado com

$$k-1-r = 6-1-0 = 5 \quad \text{graus de liberdade}$$

$$Q \sim \chi^2_{(5)}$$

Regra de Decisão através da Região Crítica

$Q_{obs} = 17.2$ e $RC = \left[x^2_{(k-1-r);1-\alpha}, +\infty\right[= \left[x^2_{(5);0.95}, +\infty\right[= [11.1, +\infty[$

Como $Q_{obs} = 17.2 \in RC$ então rejeita-se a hipótese H_0

Regra de Decisão através do valor-p

valor-$p = P(Q \geq Q_{obs}) = P(Q \geq 17.2) = 1 - P(Q < 17.2) = 1 - F(17.2)$

R: valor-$p = 1 - F(17.2) = 1 - 0.9959 = 0.0041$

Tabela em papel: valor-$p = 1 - F(17.2) \approx 1 - F(16.7) = 1 - 0.995 = 0.005$

Como valor-$p \leq 0.05 = \alpha$ então rejeita-se a hipótese H_0

Conclusão: Com base na amostra e ao nível de significância de 5%, o dado não parece ser equilibrado.

R

usar a função $chisq.test()$

e obtém-se

- $Q_{obs} = 17.2$
- graus de liberdade $= 5$
- valor-$p = 0.004136$

Como valor-$p \leq 0.05 = \alpha$ então rejeita-se a hipótese H_0

Conclusão: Com base na amostra e ao nível de significância de 5%, o dado não parece ser equilibrado.

Teste de ajustamento do Qui-Quadrado

Exemplo 2

A procura diária de um certo produto foi, em 40 dias escolhidos ao acaso, a seguinte:

Número de unidades	Número de dias
0	6
1	14
2	10
3	7
4	2
5	1

Será que tais observações foram extraídas de uma população com distribuição de Poisson, isto é, será de admitir que tal procura segue uma distribuição de Poisson? Teste a hipótese referida considerando um nível de significância de 1%.

Hipótese a ser testada

Seja X a variável aleatória que representa a procura diária do produto

$$H_0 : X \sim P(\lambda) \qquad vs \qquad H_1 : X \nsim P(\lambda)$$

Dados

- Total de dados: $n = 6 + 14 + 10 + 7 + 2 + 1 = 40$
- É necessário estimar λ, como $E[X] = \lambda$, então uma estimativa para λ é a média da amostra
- Estimativa de λ: $\overline{x} = \frac{0\times6+1\times14+2\times10+3\times7+4\times2+5\times1}{40} = 1.7$
- Distribuição Poisson: $X \sim P(1.7)$ (tabela)
- Número de parâmetros estimados: $r = 1$
- nível de significância $= \alpha = 0.01$

Construir a tabela de frequências observadas e esperadas

Domínio x_i	Frequências Observadas $O_i = n_i$	Probabilidade $p_i = f(x_i)$	Frequências Esperadas $E_i = n \times p_i$
0	6	$f(0) = 0.1827$	$40 \times 0.1827 = 7.309$
1	14	$f(1) = 0.3106$	$40 \times 0.3106 = 12.424$
2	10	$f(2) = 0.2640$	$40 \times 0.2640 = 10.56$
3	7	$f(3) = 0.1496$	$40 \times 0.1496 = 5.984$
4	2	$f(4) = 0.0636$	$\boxed{40 \times 0.0636 = 2.544^{(***)}}$
5 ou mais $^{(*)}$	1	$P(X \geq 5) = 0.0296^{(**)}$	$\boxed{40 \times 0.0296 = 1.184^{(***)}}$
	$n = 40$	1	$n = 40$

$^{(*)}$ o domínio da distribuição Poisson não tem fim

$^{(**)} P(X \geq 5) = 1 - P(X < 5) = 1 - P(X \leq 4) = 1 - F(4) = 1 - 0.9704 = 0.0296$

$^{(***)}$ Como falham as condições de aplicabilidade deste teste, mais de 20% $(6 \times 20\% = 1.2)$ das frequências esperadas são inferiores a 5, iremos agregar valores adjacentes.

Construir a tabela de frequências observadas e esperadas

Domínio x_i	Frequências Observadas $O_i = n_i$	Probabilidade $p_i = f(x_i)$	Frequências Esperadas $E_i = n \times p_i$	Valor Observado da Estatística de Teste $\frac{(O_i - E_i)^2}{E_i}$
0	6	0.1827	7.309	$\frac{(6-7.309)^2}{7.309} = 0.2344$
1	14	0.3106	12.424	$\frac{(14-12.424)^2}{12.424} = 0.1999$
2	10	0.2640	10.56	$\frac{(10-10.56)^2}{10.56} = 0.0297$
3	7	0.1496	5.984	$\frac{(7-5.984)^2}{5.984} = 0.1725$
4 ou mais	3	$P(X \geq 4) = 0.0932^{(*)}$	$40 \times 0.0932 = 3.728$	$\frac{(3-3.728)^2}{3.728} = 0.1422$
	$n = 40$	1	$n = 40$	$\sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} = 0.7787$

$^{(*)}\, P(X \geq 4) = 1 - P(X < 4) = 1 - P(X \leq 3) = 1 - F(3) = 1 - 0.9068 = 0.0932$

A estatística de teste, sob a hipótese H_0, tem distribuição Qui-Quadrado com

$$k-1-r = 5-1-1 = 3 \quad \text{graus de liberdade}$$

$$Q \sim \chi^2_{(3)}$$

Regra de Decisão através da Região Crítica

$Q_{obs} = 0.7787$ e $RC = \left[x^2_{(k-1-r);1-\alpha}, +\infty\right[= \left[x^2_{(3);0.99}, +\infty\right[= [11.3, +\infty[$

Como $Q_{obs} = 0.7787 \notin RC$ então não se rejeita a hipótese H_0

Regra de Decisão através do valor-p

valor-$p = P(Q \geq Q_{obs}) = P(Q \geq 0.7787) = 1 - P(Q < 0.7787) = 1 - F(0.7787)$

R: valor-$p = 1 - F(0.7787) = 1 - 0.1454 = 0.8546$

Tabela em papel: valor-$p = 1 - F(0.7787) \approx 1 - F(0.584) = 1 - 0.10 = 0.90$

Como valor-$p > 0.01 = \alpha$ então não se rejeita a hipótese H_0

Conclusão: Com base na amostra e ao nível de significância de 1%, conclui-se que a procura diária segue uma distribuição Poisson.

R

usar a função $chisq.test()$

e obtém-se

- $Q_{obs} = 0.77853$
- graus de liberdade (corrigido) $= 3$
- valor-p (corrigido) $= 0.8545952$

Como valor-$p > 0.01 = \alpha$ então não se rejeita a hipótese H_0

Conclusão: Com base na amostra e ao nível de significância de 1%, conclui-se que a procura diária segue uma distribuição Poisson.

Teste de ajustamento do Qui-Quadrado

Exemplo 3

Na tabela seguinte apresentam-se os tempos de falha (em horas) de uma determinada máquina:

1476	300	98	221	157
182	499	552	1563	36
246	442	20	796	31
47	438	400	279	247
210	284	553	767	1297
214	428	597	2025	185
467	401	210	289	1024

Será que tais observações foram extraídas de uma população com distribuição Exponencial, isto é, será de admitir que os tempos de falha seguem uma distribuição Exponencial? Teste a hipótese referida considerando um nível de significância de 10%.

Hipótese a ser testada

Seja X a variável aleatória que representa os tempos de falha em horas

$$H_0 : X \sim Exp(\theta) \qquad vs \qquad H_1 : X \nsim Exp(\theta)$$

Dados

- Total de dados: $n = 35$
- É necessário estimar θ, como $E[X] = \theta$, então uma estimativa para θ é a média da amostra
- Estimativa de θ: $\overline{x} = \frac{1476+300+98+\cdots+1024}{35} = 485.1714$ horas
- Distribuição Exponencial: $X \sim Exp(485.1714)$

$$F(x) = P(X \leq x) = \begin{cases} 0 & , x < 0 \\ 1 - e^{-\frac{x}{485.1714}} & , x \geq 0 \end{cases}$$

- Número de parâmetros estimados: $r = 1$
- Como a variável é contínua é necessário definir classes $\rightarrow$ Regra de Sturges: 6 classes e cada classe com amplitude 334.2
- nível de significância $= \alpha = 0.10$

Construir a tabela de frequências observadas e esperadas

Domínio: Classes $]x_i, x_{i+1}]$	Frequências Observadas $O_i = n_i$	Probabilidade $p_i = P(x_i < X \leq x_{i+1})$
$]-\infty, 354.2]$ $^{(*)}$	18	$P(X \leq 354.2) = F(354.2) = 0.5181$
$]354.2, 688.4]$	10	$P(354.2 < X \leq 688.4) = F(688.4) - F(354.2) = 0.2399$
$]688.4, 1022.6]$	2	$P(688.4 < X \leq 1022.6) = F(1022.6) - F(688.4) = 0.1205$
$]1022.6, 1356.8]$	2	$P(1022.6 < X \leq 1356.8) = F(1356.8) - F(1022.6) = 0.0605$
$]1356.8, 1691]$	2	$P(1356.8 < X \leq 1691) = F(1691) - F(1356.8) = 0.0304$
$]1691, +\infty[$ $^{(**)}$	1	$P(X > 1691) = 1 - P(X \leq 1691) = 1 - F(1691) = 0.0306$
	$n = 35$	1

$^{(*)}$ com base na amostra seria $[20, 354.2]$, mas a variável pode assumir qualquer valor em $\mathbb{R}$

$^{(**)}$ com base na amostra seria $]1691, 2025.2]$, mas a variável pode assumir qualquer valor em $\mathbb{R}$

Construir a tabela de frequências observadas e esperadas

Domínio: Classes $]x_i, x_{i+1}]$	Frequências Observadas $O_i = n_i$	Probabilidade $p_i = P(x_i < X \leq x_{i+1})$	Frequências Esperadas $E_i = n \times p_i$
$]-\infty, 354.2]$	18	0.5181	$35 \times 0.5181 = 18.1341$
$]354.2, 688.4]$	10	0.2399	$35 \times 0.2399 = 8.3965$
$]688.4, 1022.6]$	2	0.1205	$\boxed{35 \times 0.1205 = 4.2164^{(***)}}$
$]1022.6, 1356.8]$	2	0.0605	$\boxed{35 \times 0.0605 = 2.1173^{(***)}}$
$]1356.8, 1691]$	2	0.0304	$\boxed{35 \times 0.0304 = 1.0632^{(***)}}$
$]1691, +\infty[$	1	0.0306	$\boxed{35 \times 0.0306 = 1.0725^{(***)}}$
	$n = 35$	1	$n = 35$

$^{(***)}$ Falham as condições de aplicabilidade do teste, mais de 20% (6 classes $\times\ 20\% = 1.2$) das frequências esperadas são inferiores a 5, iremos agregar classes adjacentes.

Construir a tabela de frequências observadas e esperadas

Domínio: Classes $]x_i, x_{i+1}]$	Frequências Observadas $O_i = n_i$	Probabilidade $p_i = P(x_i < X \leq x_{i+1})$	Frequências Esperadas $E_i = n \times p_i$	Valor Observado da Estatística de Teste $\frac{(O_i - E_i)^2}{E_i}$
$]-\infty, 354.2]$	18	0.5181	18.1341	$\frac{(18-18.1341)^2}{18.1341} = 0.0010$
$]354.2, 688.4]$	10	0.2399	8.3965	$\frac{(10-8.3965)^2}{8.3965} = 0.3062$
$]688.4, +\infty[$	7	$P(X \geq 688.4) = 0.2420$	8.4695	$\frac{(7-8.4695)^2}{8.4695} = 0.2550$
	$n = 35$	1	$n = 35$	$\sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i} = 0.5622$

A estatística de teste, sob a hipótese H_0, tem distribuição Qui-Quadrado com

$$k-1-r = 3-1-1 = 1 \quad \text{graus de liberdade}$$

$$Q \sim \chi^2_{(1)}$$

Regra de Decisão através da Região Crítica

$Q_{obs} = 0.5622$ e $RC = \left[x^2_{(k-1-r);1-\alpha}, +\infty\right[= \left[x^2_{(1);0.90}, +\infty\right[= [2.71, +\infty[$

Como $Q_{obs} = 0.5622 \notin RC$ então não se rejeita a hipótese H_0

Regra de Decisão através do valor-p

valor-$p = P(Q \geq Q_{obs}) = P(Q \geq 0.5622) = 1 - P(Q < 0.5622) = 1 - F(0.5622)$

R: valor-$p = 1 - F(0.5622) = 1 - 0.5466 = 0.4534$

Tabela em papel: valor-$p = 1 - F(0.5622) \approx 1 - F(0.455) = 1 - 0.50 = 0.50$

Como valor-$p > 0.10 = \alpha$ então não se rejeita a hipótese H_0

Conclusão: Com base na amostra e ao nível de significância de 10%, conclui-se que os tempos de falha seguem uma distribuição Exponencial.

R

usar a função $chisq.test()$

e obtém-se

- $Q_{obs} = 0.56218$
- graus de liberdade (corrigido) $= 1$
- valor-p (corrigido) $= 0.453383$

Como valor-$p > 0.10 = \alpha$ então não se rejeita a hipótese H_0

Conclusão: Com base na amostra e ao nível de significância de 10%, conclui-se que os tempos de falha seguem uma distribuição Exponencial.

Teste de ajustamento de Kolmogorov-Smirnov

Objetivo

Testar a adequabilidade de um modelo probabilístico a um conjunto de dados observados, ou seja, comparar a função de distribuição teórica (referente à população) com a função de distribuição amostral (referente à amostra).

Formulação das Hipóteses a Testar:

$$H_0 -$$ A população possui certa distribuição teórica

$$vs$$

$$H_1 -$$ A população não possui certa distribuição teórica

Teste de ajustamento de Kolmogorov-Smirnov

Estatística de Teste

A estatística de teste tem por base a análise da proximidade entre a função de distribuição empírica ou da amostra, $F_S(x)$, e a função de distribuição teórica ou populacional, $F_T(x)$. Supondo verdadeira a hipótese H_0:

$$D = \sup_{x} |F_S(x) - F_T(x)|$$

A estatística D de Kolmogorov-Smirnov para uma amostra encontra-se tabelada. Na tabela encontra-se o quantil $D_{n;\alpha}$ para uma amostra de dimensão n e para um nível de significância α. Nesta tabela não é dada a função de distribuição mas o seu complementar, ou seja

$$D_{n;\alpha} \text{ é o quantil de probabilidade } \alpha \text{ tal que } P(D \geq D_{n;\alpha}) = \alpha$$

Em termos práticos:

$$D = \max_{x} \left\{ \left|F_S\left(x_{(i)}\right) - F_T\left(x_{(i)}\right)\right| ; \left|F_S\left(x_{(i-1)}\right) - F_T\left(x_{(i)}\right)\right| \right\}$$

Teste de ajustamento de Kolmogorov-Smirnov

Cálculo do Valor Observado da Estatística de Teste sob a Hipótese H_0

$$D_{obs} = \max_{x} \left\{ \left| F_S\left(x_{(i)}\right) - F_T\left(x_{(i)}\right) \right| ; \left| F_S\left(x_{(i-1)}\right) - F_T\left(x_{(i)}\right) \right| \right\}$$

- Ordenar a amostra.
- Construir a tabela de frequências sem definir classes.
- função de distribuição empírica ou da amostra $= F_S(x) \rightarrow$ corresponde à frequência relativa acumulada, F_i, das tabelas de frequências;
- função de distribuição teórica ou populacional $= F_T(x) \rightarrow$ corresponde ao valor da função de distribuição da variável aleatória, com o modelo probabilístico definido em H_0.

Teste de ajustamento de Kolmogorov-Smirnov

Definição da Região de Aceitação e de Região Crítica

Um valor da estatística de teste D elevado indica um desajuste entre a distribuição amostral e teórica:

- a Região de Aceitação é $RA = [0, D_{n;\alpha}[$
- a Região Crítica é $RC = [D_{n;\alpha}, +\infty[$

Regra de Decisão com base na Região Crítica

- Se $D_{obs} \notin RC$, então, ao nível de significância α, **a hipótese H_0 não é rejeitada**, isto é, com base na amostra há evidências estatísticas que os dados provêm de uma população que possui a distribuição teórica definida na hipótese H_0.
- Se $D_{obs} \in RC$, então, ao nível de significância α, **a hipótese H_0 é rejeitada**, isto é, com base na amostra há evidências estatísticas que os dados não provêm de uma população que possui a distribuição teórica definida na hipótese H_0.

Teste de ajustamento de Kolmogorov-Smirnov

Cálculo do valor-p

Considerando que H_0 é verdadeira, o valor-p indica a probabilidade do valor observado da estatística de teste ocorrer:

$$\textbf{valor-p} = P\left(D \geq D_{\text{obs}}\right)$$

Regra de Decisão com base no valor-p

- Se valor-p $> \alpha$, então, ao nível de significância α, **a hipótese** H_0 **não é rejeitada**, isto é, com base na amostra há evidências estatísticas que os dados provêm de uma população que possui a distribuição teórica definida na hipótese H_0.
- Se valor-p $\leq \alpha$, então, ao nível de significância α, **a hipótese** H_0 **é rejeitada**, isto é, com base na amostra há evidências estatísticas que os dados não provêm de uma população que possui a distribuição teórica definida na hipótese H_0.

O valor-p pode ser visto como o menor valor de α (nível de significância) para o qual os dados observados indicam que H_0 deve ser rejeitada.

Teste de ajustamento de Kolmogorov-Smirnov

Condições de aplicação do teste

- O teste de Kolmogorov-Smirnov será usado apenas para amostras aleatórias extraídas de populações contínuas.
- O teste de Kolmogorov-Smirnov só pode ser aplicado quando a distribuição indicada na hipótese nula está completamente especificada.
- Caso pretendessemos, por exemplo, efetuar um ajustamento de uma distribuição normal, sem especificar μ e σ, temos de recorrer a outro teste - Teste de Normalidade de Lilliefors (este teste não será abordado) ou, como vimos, ao teste de ajustamento do Qui-Quadrado com os dados agrupados.

Teste de ajustamento de Kolmogorov-Smirnov

Exemplo 4

Na tabela seguinte apresentam-se os tempos de falha (em horas) de uma determinada máquina:

1476	300	98	221	157
182	499	552	1563	36
246	442	20	796	31

Será que tais observações foram extraídas de uma população com distribuição Exponencial com média 730 horas? Teste a hipótese referida considerando um nível de significância de 10%.

Hipótese a ser testada

Seja X a variável aleatória que representa os tempos de falha em horas

$$H_0 : X \sim Exp(730) \qquad vs \qquad H_1 : X \nsim Exp(730)$$

Dados

- Total de dados: $n = 15$
- Distribuição Exponencial: $X \sim Exp(730)$

$$F(x) = P(X \leq x) = \begin{cases} 0 & , \ x < 0 \\ 1 - e^{-\frac{x}{730}} & , \ x \geq 0 \end{cases}$$

- nível de significância $= \alpha = 0.10$

- Construir a tabela de frequências com os valores da amostra por ordem crescente e sem definir classes:

x_i	n_i	f_i	$F_i = F_S(x_i)$	$F_T(x_i) = P(X \leq x_i)$	$\lvert F_S(x_i) - F_T(x_i)\rvert$	$\left\lvert F_S\left(x_{i-1}\right) - F_T(x_i)\right\rvert$
20	1	$\frac{1}{15}$	$\frac{1}{15} = 0.0667$	$F(20) = 0.0270$	$\lvert 0.0667 - 0.0270\rvert = 0.0396$	$\lvert 0 - 0.0270\rvert = 0.0270$
31	1	$\frac{1}{15}$	$\frac{2}{15} = 0.1333$	$F(31) = 0.0416$	$\lvert 0.1333 - 0.0416\rvert = 0.0918$	$\lvert 0.0667 - 0.0416\rvert = 0.0251$
36	1	$\frac{1}{15}$	$\frac{3}{15} = 0.2$	$F(36) = 0.0481$	$\lvert 0.2 - 0.0481\rvert = 0.1519$	$\lvert 0.1333 - 0.0481\rvert = 0.0852$
98	1	$\frac{1}{15}$	$\frac{4}{15} = 0.2667$	$F(98) = 0.1256$	$\lvert 0.2667 - 0.1256\rvert = 0.1410$	$\lvert 0.2 - 0.1256\rvert = 0.0744$
157	1	$\frac{1}{15}$	$\frac{5}{15} = 0.3333$	$F(157) = 0.1935$	$\lvert 0.3333 - 0.1935\rvert = 0.1398$	$\lvert 0.2667 - 0.1935\rvert = 0.0732$
182	1	$\frac{1}{15}$	$\frac{6}{15} = 0.4$	$F(182) = 0.2207$	$\lvert 0.4 - 0.2207\rvert = 0.1793$	$\lvert 0.3333 - 0.2207\rvert = 0.1127$
221	1	$\frac{1}{15}$	$\frac{7}{15} = 0.4667$	$F(221) = 0.2612$	$\lvert 0.4667 - 0.2612\rvert = 0.2055$	$\lvert 0.4 - 0.2612\rvert = 0.1388$
246	1	$\frac{1}{15}$	$\frac{8}{15} = 0.5333$	$F(246) = 0.2861$	$\lvert 0.5333 - 0.2861\rvert = 0.2473$	$\lvert 0.4667 - 0.2861\rvert = 0.1806$
300	1	$\frac{1}{15}$	$\frac{9}{15} = 0.6$	$F(300) = 0.3370$	$\lvert 0.6 - 0.3370\rvert = 0.2630$	$\lvert 0.5333 - 0.3370\rvert = 0.1963$
442	1	$\frac{1}{15}$	$\frac{10}{15} = 0.6667$	$F(442) = 0.4542$	$\lvert 0.6667 - 0.4542\rvert = 0.2125$	$\lvert 0.6 - 0.4542\rvert = 0.1458$
499	1	$\frac{1}{15}$	$\frac{11}{15} = 0.7333$	$F(499) = 0.4952$	$\lvert 0.7333 - 0.4952\rvert = 0.2381$	$\lvert 0.6667 - 0.4952\rvert = 0.1715$
552	1	$\frac{1}{15}$	$\frac{12}{15} = 0.8$	$F(552) = 0.5305$	$\lvert 0.8 - 0.5305\rvert =$ **0.2695**	$\lvert 0.6667 - 0.5305\rvert = 0.2028$
796	1	$\frac{1}{15}$	$\frac{13}{15} = 0.8667$	$F(796) = 0.6639$	$\lvert 0.8667 - 0.6639\rvert = 0.2027$	$\lvert 0.8 - 0.6639\rvert = 0.1361$
1476	1	$\frac{1}{15}$	$\frac{14}{15} = 0.9333$	$F(1476) = 0.8676$	$\lvert 0.9333 - 0.8676\rvert = 0.0657$	$\lvert 0.8667 - 0.8676\rvert = 0.0009$
1563	1	$\frac{1}{15}$	$\frac{15}{15} = 1$	$F(1563) = 0.8825$	$\lvert 1 - 0.8825\rvert = 0.1175$	$\lvert 0.9333 - 0.8825\rvert = 0.0509$

$$D_{obs} = max\left\{\left|F_S\left(x_i\right) - F_T\left(x_i\right)\right|; \left|F_S\left(x_{i-1}\right) - F_T\left(x_i\right)\right|\right\} = 0.2695$$

Regra de Decisão através da Região Crítica

$$RC = [D_{n;\alpha}, +\infty[= [D_{15;0.10}, +\infty[= [0.304, +\infty[$$

Como $D_{obs} = 0.2695 \notin RC$ então não se rejeita a hipótese H_0

Regra de Decisão através do valor-p

$$\text{valor-}p = P(D \geq D_{obs}) = P(D \geq 0.2695) \approx P(D \geq 0.266) = 0.20$$

Como valor-$p > 0.10 = \alpha$ então não se rejeita a hipótese H_0

Conclusão: Com base na amostra e ao nível de significância de 10%, conclui-se que os tempos de falha seguem uma distribuição Exponencial com média 730 horas.

R

usar a função $ks.test()$

e obtém-se

- $D_{obs} = 0.26946$
- valor-$p = 0.1881$

Como valor-$p = 0.1881 > 0.10 = \alpha$ então não se rejeita a hipótese H_0

Conclusão: Com base na amostra e ao nível de significância de 10%, conclui-se que os tempos de falha seguem uma distribuição Exponencial com média 730 horas.

Teste de ajustamento de Kolmogorov-Smirnov

Exemplo 5

Numa baía efetuaram-se 36 medições dos níveis de salinidade. Os valores obtidos aleatoriamente foram os seguintes:

75	92	80	80	84	72	84	77	81
77	75	81	80	92	72	77	78	76
77	86	77	92	80	78	68	78	92
68	80	81	87	76	80	87	77	86

Pretende-se testar, para um nível de significância de 5%, se os valores da salinidade nessa baía são normalmente distribuídos com média 80 e desvio padrão 6.95.

Hipótese a ser testada

Seja X a variável aleatória que representa os níveis de salinidade

$$H_0 : X \sim N(80, 6.95) \qquad vs \qquad H_1 : X \nsim N(80, 6.95)$$

Dados

- Total de dados: $n = 36$
- Distribuição Normal: $X \sim N(80, 6.95) \Leftrightarrow Z = \frac{X-80}{6.95} \sim N(0,1)$
- nível de significância $= \alpha = 0.05$

- Construir a tabela de frequências com os valores da amostra por ordem crescente:

x_i	n_i	f_i	$F_i = F_S(x_i)$	$F_T(x_i) = P(X \leq x_i)$	$\left\|F_S(x_i) - F_T(x_i)\right\|$	$\left\|F_S(x_{i-1}) - F_T(x_i)\right\|$
68	2	$\frac{2}{36}$	$\frac{2}{36} = 0.0556$	$\Phi\left(\frac{68-80}{6.95}\right) = 0.0421$	$\|0.0556 - 0.0421\| = 0.0134$	$\|0 - 0.0421\| = 0.0421$
72	2	$\frac{2}{36}$	$\frac{4}{36} = 0.1111$	$\Phi\left(\frac{72-80}{6.95}\right) = 0.1248$	$\|0.1111 - 0.1248\| = 0.0137$	$\|0.0556 - 0.1248\| = 0.0693$
75	2	$\frac{2}{36}$	$\frac{6}{36} = 0.1667$	$\Phi\left(\frac{75-80}{6.95}\right) = 0.2359$	$\|0.1667 - 0.2359\| = 0.0693$	$\|0.1111 - 0.2359\| = 0.1248$
76	2	$\frac{2}{36}$	$\frac{8}{36} = 0.2222$	$\Phi\left(\frac{76-80}{6.95}\right) = 0.2825$	$\|0.2222 - 0.2825\| = 0.0602$	$\|0.16667 - 0.2825\| = 0.1158$
77	6	$\frac{6}{36}$	$\frac{14}{36} = 0.3889$	$\Phi\left(\frac{77-80}{6.95}\right) = 0.3330$	$\|0.3889 - 0.3330\| = 0.0559$	$\|0.2222 - 0.3330\| = 0.1108$
78	3	$\frac{3}{36}$	$\frac{17}{36} = 0.4722$	$\Phi\left(\frac{77-80}{6.95}\right) = 0.3868$	$\|0.4722 - 0.3868\| = 0.0855$	$\|0.3889 - 0.3868\| = 0.0021$
80	6	$\frac{6}{36}$	$\frac{23}{36} = 0.6389$	$\Phi\left(\frac{80-80}{6.95}\right) = 0.5$	$\|0.6389 - 0.5\| = 0.1389$	$\|0.4722 - 0.5\| = 0.0278$
81	3	$\frac{3}{36}$	$\frac{26}{36} = 0.7222$	$\Phi\left(\frac{81-80}{6.95}\right) = 0.5572$	$\|0.7222 - 0.5572\| = \mathbf{0.1650}$	$\|0.6389 - 0.5572\| = 0.0817$
84	2	$\frac{2}{36}$	$\frac{28}{36} = 0.7778$	$\Phi\left(\frac{84-80}{6.95}\right) = 0.7175$	$\|0.7778 - 0.7175\| = 0.0602$	$\|0.7222 - 0.7175\| = 0.0047$
86	2	$\frac{2}{36}$	$\frac{30}{36} = 0.8333$	$\Phi\left(\frac{86-80}{6.95}\right) = 0.8060$	$\|0.8333 - 0.8060\| = 0.0273$	$\|0.7778 - 0.8060\| = 0.0282$
87	2	$\frac{2}{36}$	$\frac{32}{36} = 0.8889$	$\Phi\left(\frac{87-80}{6.95}\right) = 0.8431$	$\|0.8889 - 0.8431\| = 0.0458$	$\|0.8333 - 0.8431\| = 0.0097$
92	4	$\frac{4}{36}$	$\frac{36}{36} = 1$	$\Phi\left(\frac{92-80}{6.95}\right) = 0.9579$	$\|1 - 0.9579\| = 0.0421$	$\|0.8889 - 0.9579\| = 0.0690$

$$D_{obs} = max\left\{\left|F_S\left(x_i\right) - F_T\left(x_i\right)\right|; \left|F_S\left(x_{i-1}\right) - F_T\left(x_i\right)\right|\right\} = 0.1650$$

Regra de Decisão através da Região Crítica

$$RC = [D_{n;\alpha}, +\infty[= [D_{36;0.05}, +\infty[= [0.221, +\infty[$$

Como $D_{obs} = 0.1650 \notin RC$ então não se rejeita a hipótese H_0

Regra de Decisão através do valor-p

$$\text{valor-}p = P(D \geq D_{obs}) = P(D \geq 0.1650) \approx P(D \geq 0.174) = 0.20$$

Como valor-$p > 0.05 = \alpha$ então não se rejeita a hipótese H_0

Conclusão: Com base na amostra e ao nível de significância de 5%, conclui-se que os níveis da salinidade nessa baía são normalmente distribuídos com média 80 e desvio padrão 6.95.

R

usar a função $ks.test()$

e obtém-se

- $D_{obs} = 0.16502$
- valor-$p = 0.2808$

Como valor-$p = 0.2808 > 0.05 = \alpha$ então não se rejeita a hipótese H_0

Conclusão: Com base na amostra e ao nível de significância de 5%, conclui-se que os níveis da salinidade nessa baía são normalmente distribuídos com média 80 e desvio padrão 6.95.

Testes de ajustamento: Qui-Quadrado vs Kolmogorov-Smirnov

Vantagens e Desvantagens

- O teste de ajustamento do Qui-Quadrado é preferencialmente usado em amostras aleatórias extraídas de populações discretas enquanto o teste de Kolmogorov-Smirnov será apenas usado para amostras aleatórias extraídas de populações contínuas.
- O teste de Kolmogorov-Smirnov não requer o agrupamento dos dados, utiliza toda a informação contida no conjunto de dados (ao contrário do teste de ajustamento do Qui-Quadrado que é necessário a agregação dos dados em classes quando a variável é contínua).
- O teste de ajustamento do Qui-Quadrado está orientado essencialmente para grandes amostras, enquanto que o teste de Kolmogorov-Smirnov é aplicável a pequenas amostras.
- O teste de Kolmogorov-Smirnov só pode ser aplicado quando a distribuição indicada na hipótese nula está completamente especificada (o que não sucede com o teste de ajustamento do Qui-Quadrado, os parâmetros podem ser estimados).