MÉTODOS ESTATÍSTICOS

Testes de Hipóteses Não Paramétricos - Parte 1

Testes de Ajustamento

Engenharia Informática

Departamento de Matemática
Escola Superior de Tecnologia de Setúbal
Instituto Politécnico de Setúbal
2023-2024

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
  - **Teste de ajustamento de Kolmogorov-Smirnov**
    - é válido para distribuições contínuas (também existe uma versão para distribuições discretas mas não será estudada)
  - **Teste de Lilliefors**
    - adaptação do teste de Kolmogorov-Smirnov para testar a distribuição Normal (sem a especificar completamente)
  - **Teste de Shapiro-Wilk**
    - um dos testes mais usados para testar a distribuição Normal

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

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

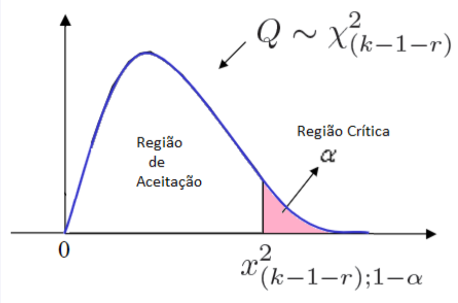
- De acordo com o tipo de dados da amostra, construir a respetiva tabela de frequências;
- k corresponde ao número de linhas da tabela de frequências
- **frequências observadas** $= O_i \rightarrow$ corresponde às frequências absolutas, n_i, das tabelas de frequências;
- **frequências esperadas** $= E_i = np_i \rightarrow$ 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 $RA = \left[0, x^2_{1-\alpha;(k-1-r)}\right[$
- a Região Crítica é $RC = \left[x^2_{1-\alpha;(k-1-r)}, +\infty\right[$

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,

$$Q_{obs} \in RC$$

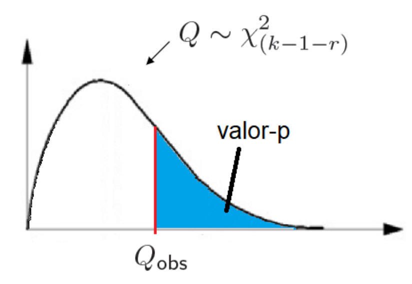
então, ao nível de significância α, **a hipótese** H_0 **é rejeitada**, isto é, com base na amostra há evidências estatísticas que os dados não provêm de uma população que possui a distribuição teórica definida na hipótese H_0.

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

Teste de ajustamento do Qui-Quadrado no R

- chisq.test()

Observação: Caso se pretenda testar um ajustamento de uma distribuição onde seja necessário estimar os parâmetros, então o valor-p que sai do chisq.test() não está correto, é necessário ajustar o valor-p ao facto de existirem parâmetros estimados.

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	$n = 210$	$Q_{obs} = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = 17.2$

Obs: Para o R só é necessário ter as colunas referentes às Frequências Observadas e à Probabilidade.

R

usar a função $chisq.test()$

com

$$chisq.test(x = O_i, p = p_i)$$

e obtém-se

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

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-$p = P(Q \geq Q_{obs}) = P(Q \geq 17.2) = 1 - P(Q < 17.2) = 1 - F(17.2) = 0.0041$

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 (ou ainda não se lançou o número de vezes suficiente para atingir o equilíbrio).

Teste de ajustamento do Qui-Quadrado

Exemplo 2

Pretende-se testar a eficácia de um novo tratamento em pacientes com uma determinada doença. Considera-se que cada paciente tem 60% de hipóteses de responder positivamente ao tratamento. O tratamento foi administrado a 50 grupos de 4 pacientes e os resultados obtidos foram os seguintes:

número de pacientes que respondem positivamente ao tratamento	número de grupos
1	5
2	12
3	19
4	14

Teste, para um nível de significância de 5%, se o número de pacientes que respondem positivamente ao tratamento pode ser modelado por uma variável aleatória que segue uma distribuição binomial com $n = 4$ e $p = 0.6$.

Hipótese a ser testada

Seja X a variável aleatória que representa o número de pacientes que respondem positivamente ao tratamento, num grupo de 4 pacientes

$$H_0 : X \sim B(4, 0.6) \qquad vs \qquad H_1 : X \nsim B(4, 0.6)$$

Dados

- Total de dados: $n = 5 + 12 + 19 + 14 = 50$
- Distribuição Binomial completamente especificada: $X \sim B(4, 0.6)$
- nível de significância $= \alpha = 0.05$
- Número de parâmetros estimados $= r = 0$
- número de linhas da tabela de frequências $= k = 5$, pois $D_X = \{0, 1, 2, 3, 4\}$
- Construir a tabela de frequências observada e esperada:

Construir a tabela de frequências observadas e esperadas

Domínio x_i	Frequências Observadas $O_i = n_i$	Probabilidade $p_i = f(x_i)$	Frequências $E_i = n \times p_i$
0	0	$f(0) = 0.0256$	$50 \times 0.0256 = 1.28$
1	5	$f(1) = 0.1536$	$50 \times 0.1536 = 7.68$
2	12	$f(2) = 0.3456$	$50 \times 0.3456 = 17.28$
3	19	$f(3) = 0.3456$	$50 \times 0.3456 = 17.28$
4	14	$f(4) = 0.1296$	$50 \times 0.1296 = 6.48$
	$n = 50$	1	$n = 50$

R

usar a função $chisq.test()$

com

$$chisq.test(x = O_i, p = p_i)$$

e obtém-se

- $Q_{obs} = 12.727$
- graus de liberdade $= 4$
- valor-$p = 0.01269$

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

$$k-1-r=5-1-0=4 \quad \text{graus de liberdade}$$

$$Q \sim \chi^2_{(4)}$$

Regra de Decisão através da Região Crítica

$$RC=\left[x^2_{1-\alpha;(k-1-r)},+\infty\right[=\left[x^2_{0.95;(5)},+\infty\right[=[9.4878,+\infty[$$

Como $Q_{obs}=12.7267 \in RC$ então rejeita-se a hipótese H_0

Regra de Decisão através do valor-p

valor-$p=P(Q \geq Q_{obs})=P(Q \geq 12.7267)=1-P(Q<12.7267)=1-F(12.7267)=0.0127$

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%, não há evidência estatística que os dados possam ser modelados por uma distribuição binomial com $n=4$ e $p=0.6$.

Teste de ajustamento do Qui-Quadrado

Exemplo 3

A procura diária de um certo produto foi, em 88 dias escolhidos ao acaso, a seguinte:

Número de unidades	Número de dias
0	14
1	22
2	18
3	15
4	10
5	9

1. Será que tais observações foram extraídas de uma população com distribuição de Poisson de média 2.1, isto é, será de admitir que tal procura segue uma distribuição de Poisson com média 2.1? 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

$$H_0 : X \sim P(2.1) \qquad \text{contra} \qquad H_1 : X \nsim P(2.1)$$

Dados

- Total de dados: $n = 14 + 22 + 18 + 15 + 10 + 9 = 88$
- Distribuição de Poisson completamente especificada: $X \sim P(2.1)$
- nível de significância $= \alpha = 0.01$

- número de parâmetros estimados $= r = 0$
- número de linhas da tabela de frequências $= k = 6$
- Construir a tabela de frequências observada e esperada:

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	14	$f(0) = 0.1225$	$88 \times 0.1225 = 10.7762$
1	22	$f(1) = 0.2572$	$88 \times 0.2572 = 22.6299$
2	18	$f(2) = 0.2700$	$88 \times 0.2700 = 23.7614$
3	15	$f(3) = 0.1890$	$88 \times 0.1890 = 16.6330$
4	10	$f(4) = 0.0992$	$88 \times 0.0992 = 8.7323$
5 ou mais $^{(*)}$	9	$P(X \geq 5) = 0.0621^{(**)}$	$88 \times 0.0621 = 5.4671$
	$n = 88$	1	$n = 88$

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

$^{(**)} P(X \geq 5) = 1 - P(X < 5) = 1 - P(X \leq 4) = 1 - F(4) = 1 - 0.9379 = 0.0621$

R

usar a função $chisq.test()$

com

$$chisq.test(x = O_i, p = p_i)$$

e obtém-se

- $Q_{obs} = 5.0063$
- graus de liberdade $= 5$
- valor-$p = 0.4151$

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

$$RC = \left[x^2_{1-\alpha;(k-1-r)}, +\infty\right[= \left[x^2_{0.99;(5)}, +\infty\right[= [15.0862, +\infty[$$

Como $Q_{obs} = 5.0063 \notin RC$ então não se rejeita a hipótese H_0

Regra de Decisão através do valor-p

valor-$p = P(Q \geq Q_{obs}) = P(Q \geq 5.0063) = 1 - P(Q < 5.0063) = 1 - F(5.0063) = 0.4151$

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 há evidência estatística que a procura diária segue uma distribuição de Poisson com média 2.1.

Teste de ajustamento do Qui-Quadrado

Exemplo 3

A procura diária de um certo produto foi, em 88 dias escolhidos ao acaso, a seguinte:

Número de unidades	Número de dias
0	14
1	22
2	18
3	15
4	10
5	9

2. 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

$$H_0 : X \sim P(\lambda) \qquad \text{contra} \qquad H_1 : X \nsim P(\lambda)$$

Dados

- Total de dados: $n = 14 + 22 + 18 + 15 + 10 + 9 = 88$
- Distribuição de Poisson não está completamente especificada: $X \sim P(\lambda)$
- É necessário estimar λ, como $E[X] = \lambda$, então uma estimativa para λ é a média da amostra
- Estimativa de λ: $\overline{x} = 2.1364$
- nível de significância $= \alpha = 0.01$

- número de parâmetros estimados $= r = 1$
- número de linhas da tabela de frequências $= k = 6$
- Construir a tabela de frequências observada e esperada:

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	14	$f(0) = 0.1181$	$88 \times 0.1225 = 10.3913$
1	22	$f(1) = 0.2523$	$88 \times 0.2572 = 22.1997$
2	18	$f(2) = 0.2695$	$88 \times 0.2700 = 23.7133$
3	15	$f(3) = 0.1919$	$88 \times 0.1890 = 16.8867$
4	10	$f(4) = 0.1025$	$88 \times 0.0992 = 9.0191$
5 ou mais $^{(*)}$	9	$P(X \geq 5) = 0.0658^{(**)}$	$88 \times 0.0621 = 5.7899$
	$n = 88$	1	$n = 88$

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

$^{(**)} P(X \geq 5) = 1 - P(X < 5) = 1 - P(X \leq 4) = 1 - F(4) = 1 - 0.9342 = 0.0658$

R

usar a função $chisq.test()$

com

$$chisq.test(x = O_i, p = p_i)$$

e obtém-se

- $Q_{obs} = 4.7289$
- graus de liberdade: estão ERRADOS
- valor-p : está ERRADO

Observação: como foi necessário estimar parâmetros, só o valor da estatística de teste é que está certo.

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

$$k-1-r = 6-1-1 = 4 \quad \text{graus de liberdade}$$

$$Q \sim \chi^2_{(4)}$$

Regra de Decisão através da Região Crítica

$$RC = \left[x^2_{1-\alpha;(k-1-r)}, +\infty\right[= \left[x^2_{0.99;(4)}, +\infty\right[= [13.2767, +\infty[$$

Como $Q_{obs} = 4.7289 \notin RC$ então não se rejeita a hipótese H_0

Regra de Decisão através do valor-p

valor-$p = P(Q \geq Q_{obs}) = P(Q \geq 4.7289) = 1 - P(Q < 4.7289) = 1 - F(4.7289) = 0.3163$

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 há evidência estatística que a procura diária segue uma distribuição Poisson.

Teste de ajustamento do Qui-Quadrado

Exemplo 4

O número de rifas vendidas por dia numa certa banca de uma feira foi, em 40 dias escolhidos ao acaso, a seguinte:

Número de rifas	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 o número de rifas vendidas por dia 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 número de rifas vendidas por dia

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

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: $X \sim P(1.7)$
- 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 $E_i = n \times p_i$
0	6	0.1827	7.309
1	14	0.3106	12.424
2	10	0.2640	10.56
3	7	0.1496	5.984
4 ou mais	3	$P(X \geq 4) = 0.0932^{(*)}$	$40 \times 0.0932 = 3.728$
	$n = 40$	1	$n = 40$

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

R

usar a função $chisq.test()$

com

$$chisq.test(x = O_i, p = p_i)$$

e obtém-se

- $Q_{obs} = 0.77853$
- graus de liberdade: estão ERRADOS
- valor-p : está ERRADO

Observação: como foi necessário estimar parâmetros, só o valor da estatística de teste é que está certo.

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

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

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

Regra de Decisão através do valor-p

valor-$p = P(Q \geq Q_{obs}) = P(Q \geq 0.7787) = 1 - P(Q < 0.7787) = 1 - F(0.7787) = 0.8546$

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 existe evidência estatística para afirmar que o número de rifas vendidas por dia segue uma distribuição de Poisson.

Teste de ajustamento do Qui-Quadrado

Exemplo 5

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
  (considera-se sempre a primeira classe a começar no mínimo observado e intervalos abertos à esquerda e fechados à direita)
- 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 $E_i = n \times p_i$
$]-\infty, 354.2]$	18	0.5181	18.1341
$]354.2, 688.4]$	10	0.2399	8.3965
$]688.4, +\infty[$	7	$P(X \geq 688.4) = 0.2420$	8.4695
	$n = 35$	1	$n = 35$

R

usar a função $chisq.test()$

com

$$chisq.test(x = O_i, p = p_i)$$

e obtém-se

- $Q_{obs} = 0.56253$
- graus de liberdade: estão ERRADOS
- valor-p : está ERRADO

Observação: como foi necessário estimar parâmetros, só o valor da estatística de teste é que está certo.

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

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

Como $Q_{obs} = 0.5625 \notin RC$ então não se rejeita a hipótese H_0

Regra de Decisão através do valor-p

valor-$p = P(Q \geq Q_{obs}) = P(Q \geq 0.5625) = 1 - P(Q < 0.5625) = 1 - F(0.5625) = 0.4532$

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 há evidência estatística 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 referente a dados contínuos

contra

$H_1 -$ A população não possui certa distribuição teórica referente a dados contínuos

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, $F_T(x)$:

$$D = \sup_x |F_S(x) - F_T(x)| \sim D_n$$

A estatística D de Kolmogorov-Smirnov 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.

Teste de ajustamento de Kolmogorov-Smirnov

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

Supondo verdadeira a hipótese H_0 e com base na amostra recolhida, em termos práticos, calcular o valor da estatística de teste consiste em calcular o seguinte valor:

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

Para calcular D_{obs} é necessário começar por construir a tabela de frequências sem definir classes onde:

- $\left(x_{(1)}, x_{(2)}, ..., x_{(n)}\right)$ corresponde à amostra recolhida ordenada;
- $F_S\left(x\right)$ corresponde à função de distribuição empírica ou da amostra (ou seja, corresponde às frequências relativas acumuladas das tabelas de frequências);
- $F_T\left(x\right)$ corresponde à função de distribuição teórica ou populacional (ou seja, 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 em 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 em H_0.

Como o R não tem uma função (simples) para calcular os quantis de probabilidade da distribuição de Kolmogorov-Smirnov, não vamos tomar decisões com recurso à região crítica.

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 (existem adaptações para distribuições discretas mas não serão lecionadas).
- O teste de Kolmogorov-Smirnov só pode ser aplicado quando a distribuição indicada na hipótese nula está completamente especificada.
- Caso se pretenda testar um ajustamento de uma distribuição normal sem especificar μ e σ, vamos recorrer a uma adaptação: **Teste de Normalidade de Lilliefors**.

  O valor da estatística de teste é igual, só o valor-p é ajustado pois é necessário estimar os parâmetros μ e σ.

Teste de ajustamento de Kolmogorov-Smirnov

Teste de ajustamento de Kolmogorov-Smirnov no R

- ks.test()

Teste de Normalidade de Lilliefors no R

- library(nortest)
- lillie.test()

Teste de ajustamento de Kolmogorov-Smirnov

Exemplo 6

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	$F_i = F_S(x_i)$	$F_T(x_i) = P(X \leq x_i)$	$\lvert F_S(x_i) - F_T(x_i)\rvert$	$\left\lvert F_S(x_{i-1}) - F_T(x_i)\right\rvert$
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	$\frac{1}{15}$	$\frac{15}{15} = 1$	$F(1563) = 0.8825$	$\lvert 1 - 0.8825\rvert = 0.1175$	$\lvert 0.9333 - 0.8825\rvert = 0.0509$

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

Agora para tomar uma decisão era necessário consultar a tabela da Estatística de Teste de Kolmogorov-Smirnov e construir a região crítica:

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

ou calcular o valor-p

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

Vamos recorrer ao R:

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 existe evidência estatística para afirmar que os tempos de falha podem seguir uma distribuição Exponencial com média 730 horas.

Teste de ajustamento de Kolmogorov-Smirnov

Exemplo 7

Numa baía efetuaram-se 54 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
74	93	79	81	83	71	83	78	80
76	76	80	82	91	72	76	79	75

1. 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: $n = 54$
- Distribuição Normal completamente especificada: $X \sim N(80, 6.95)$
- nível de significância $= \alpha = 0.05$
- Vamos recorrer ao R para efetuar o teste pretendido:

R

usar a função $ks.test()$

e obtém-se

- $D_{obs} = 0.16502$
- valor-$p = 0.1056$

Como valor-$p = 0.1056 > 0.05 = \alpha$ então não se rejeita a hipótese H_0

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

Teste de ajustamento de Kolmogorov-Smirnov

Exemplo 7

Numa baía efetuaram-se 54 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
74	93	79	81	83	71	83	78	80
76	76	80	82	91	72	76	79	75

2. Pretende-se testar, para um nível de significância de 1%, se os valores da salinidade nessa baía são normalmente distribuídos.

Hipótese a ser testada

Seja X a variável aleatória que representa os níveis de salinidade

$$H_0 : X \sim N(\mu, \sigma) \qquad \text{contra} \qquad H_1 : X \nsim N(\mu, \sigma)$$

Dados

- Total de dados: $n = 54$
- Distribuição Normal não está completamente especificada: $X \sim N(\mu, \sigma)$, μ e σ são desconhecidos
- nível de significância $= \alpha = 0.01$

Vamos recorrer ao R mas como a distribuição não está completamente especificada e queremos testar a distribuição Normal, vamos recorrer ao teste de Normalidade de Lilliefors:

(Observação: Se não pretendesse testar a distribuição Normal, então a única possibilidade era recorrer ao Teste de Ajustamento do Qui-Quadrado pois a distribuição não está completamente especificada.)

R

library(nortest)

usar a função $lillie.test()$

e obtém-se

- $D_{obs} = 0.14652$
- valor-$p = 0.005525$

Como valor-$p = 0.005525 \leq 0.01 = \alpha$ então Rejeita-se a hipótese H_0

Conclusão: Com base na amostra e ao nível de significância de 1%, conclui-se que não existe evidência estatística para afirmar que os níveis da salinidade nessa baía são normalmente distribuídos.

Observação: Se tivéssemos feito

ks.test(salinidade, "pnorm", mean=mean(salinidade), sd=sd(salinidade))

o valor observado da estatística de teste seria igual ao do teste de normalidade de Lilliefors, $D_{obs} = 0.14652$, mas o valor-p estava errado.

Teste de ajustamento de Shapiro-Wilk

Objetivo

Testar se um dado conjunto de observações pode ser considerado proveniente de uma população com distribuição Normal.

Formulação das Hipóteses a Testar:

H_0 : A população segue uma distribuição Normal

contra

H_1 : A população não segue uma distribuição Normal

Ou de forma equivalente:

Seja X a característica em estudo na população

$$H_0 : X \sim \text{ Normal} \qquad \text{contra} \qquad H_1 : X \nsim \text{ Normal}$$

Teste de ajustamento de Shapiro-Wilk

Estatística de Teste

A estatística de teste é

$$W = \frac{b^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \sim W_{(n)}$$

onde $(X_1, X_2, ...X_n)$ representa a amostra e b é a constante Shapiro-Wilk.

A estatística W de Shapiro-Wilk encontra-se tabelada. Na tabela encontra-se o quantil $W_{(n;\alpha)}$ para uma amostra de dimensão n e para um nível de significância α.

Teste de ajustamento de Shapiro-Wilk

Cálculo do Valor Observado da Estatística de Teste

Supondo que a Hipótese H_0 é verdadeira (ou seja, que a amostra vem de uma população com uma distribuição Normal) e com base na amostra recolhida obtém-se o valor observado da estatística de teste:

$$W_{obs} = \frac{b^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

a constante de Shapiro-Wilk, b, é calculada com base na amostra recolhida e com recurso a uma tabela (que não iremos utilizar).

A Estatística de Teste de Shapiro-Wilk é semelhante a uma correlação entre os valores da amostra e os valores da distribuição Normal, então tem-se

$$0 < W_{obs} \leq 1$$

só existindo para $n \geq 3$.

Teste de ajustamento de Shapiro-Wilk

Definição da Região de Aceitação e de Região Crítica

Um valor da estatística de teste W_{obs} próximo de 1 indica que é provável que a distribuição seja Normal, portanto tem-se:

- a Região de Aceitação é $RA = \left]W_{(n;\alpha)}, 1\right]$
- a Região Crítica é $RC = \left]0, W_{(n;\alpha)}\right]$

Regra de Decisão com base na Região Crítica

- Se $W_{obs} \notin RC$, 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 Normal.
- Se $W_{obs} \in RC$, 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 Normal.

Como o R não tem uma função (simples) para calcular os quantis de probabilidade da distribuição de Shapiro-Wilk, não vamos tomar decisões com recurso à região crítica.

Teste de ajustamento de Shapiro-Wilk

Cálculo do valor-p

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

$$\textbf{valor-p} = P\left(W \leq W_{\text{obs}}\right)$$

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 Normal.
- 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 Normal.

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 Shapiro-Wilk

Condições de aplicação do teste

- O teste de Shapiro Wilk só pode ser usado para testar a distribuição Normal.
- No caso da distribuição Normal e em comparação com o teste de ajustamento de Lilliefors (teste de ajustamento de Kolmogorov-Smirnov), o teste de ajustamento de Shapiro-Wilk só é considerado mais potente quando a amostra tem dimensão $n < 50$.

Teste de ajustamento de Shapiro-Wilk no R

- shapiro.test()

Teste de ajustamento de Shapiro-Wilk

Exemplo 8

Para avaliar os níveis de seca é usual medir os níveis de salinidade dos rios em determinadas localizações. Na tabela seguinte são apresentados os valores obtidos em 36 localizações:

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 1%, se os valores da salinidade são normalmente distribuídos.

Hipótese a ser testada

Seja X a variável aleatória que representa os níveis de salinidade

$$H_0 : X \sim \text{Normal} \qquad \text{contra} \qquad H_1 : X \nsim \text{Normal}$$

Dados

- Total de dados: $n = 36$
- nível de significância $= \alpha = 0.01$

Vamos recorrer ao R e como a mostra tem dimensão $36 < 50$ e pretende-se testar a normalidade, vamos recorrer ao test mais potente, ou seja, ao teste de ajustamento Shapiro-Wilks:

R

usar a função $shapiro.test()$

e obtém-se

- $W_{obs} = 0.9421$
- valor-$p = 0.05905$

Como valor-$p = 0.05905 \leq 0.01 = \alpha$ então não se rejeita a hipótese H_0.

Conclusão: Com base na amostra e ao nível de significância de 1%, conclui-se que existe evidência estatística para afirmar que os níveis da salinidade são normalmente distribuídos.

Testes de ajustamento - Vantagens e Desvantagens

- **Teste de ajustamento do Qui-Quadrado**

  Vantagens: permite testar distribuições discretas e contínuas. É preferencialmente usado em amostras aleatórias extraídas de populações discretas.

  Desvantagens: Para garantir que cumpre todas as condições convém ter grandes amostras. No caso das distribuições contínuas é necessário agrupar os dados em classes.

- **Teste de Ajustamento de Kolmogorov-Smirnov**

  Vantagens: permite testar qualquer distribuição contínua e não é necessário agrupar dados.

  Desvantagens: é necessário ter a distribuição completamente especificada.

  Observação: No caso da distribuição Normal a desvantagem é ultrapassada com o recurso ao **teste de Normalidade de Lilliefors**, que é considerado um dos testes mais potentes quando se tem amostras de dimensão $n \geq 50$.

Testes de ajustamento - Vantagens e Desvantagens

- **Teste de Ajustamento de Shapiro-Wilk**

  Vantagens: não necessita que seja especificada qual a distribuição Normal que se pretende testar, não é necessário agrupar dados e é considerado um dos testes mais potentes quando se tem amostras de dimensão $n < 50$.

  Desvantagens: só permite testar a distribuição Normal.