

Testes Ajustamento (Kolmogorov – Smirnov)

- TESTES DE AJUSTAMENTO

Condições	Teste de Ajustamento	R
Distribuição Discreta ou Distribuição Contínua (com recurso a classes)	Qui-Quadrado	chisq.test()
Distribuição Contínua (completamente especificada)	Kolmogorov-Smirnov	ks.test()
Normal e $n \geq 50$	Lilliefors	lillie.test() (library(nortest))
Normal e $n < 50$	Shapiro Wilks	shapiro.test()

Completamente especificada significa que conhecemos os parâmetros da distribuição.

Por exemplo, se houver suspeitas que uns dados provêm de uma população Exponencial com parâmetro 100, pode-se usar o teste Kolmogorov-Smirnov (temos uma distribuição contínua e com o parâmetro θ especificado)

Quando se pretende verificar a distribuição de uns dados, podem ser realizados vários testes. Existem outros, mas apenas falaremos nestes quatro.

Vejamos os seguintes exemplos:

1. Se quisesse verificar se uns dados são originários de uma população com distribuição Binomial, qual o teste a usar?
A Binomial é discreta. Logo deve-se usar o Qui-Quadrado.
Nos slides cap6- parte1 estão vários exemplos com o Qui-Quadrado.
2. Para ver se os dados são originários de uma distribuição Exponencial, sem conhecer o parâmetro, qual o teste?
Não conhecemos o parâmetro, logo não pode ser Kolmogorov-Smirnov (K-S). Exponencial não é distribuição normal, logo não se podem usar o Lilliefors nem o Shapiro-Wilks. Teria de ser o Qui-quadrado. Nos slides cap6- parte1 estão vários exemplos com o Qui-Quadrado.
3. Se pretendesse averiguar a Normalidade de uns dados sabendo os parâmetros μ e σ , qual o teste?
Como é contínua e está completamente especificada (conhece-se os parâmetros de uma normal, neste caso média e desvio padrão populacional) pode-se usar o K-S. (exemplos neste documento)
4. Se pretendesse averiguar a Normalidade de uns dados cuja amostra tem 40 elementos, qual o teste?
Temos uma normal (contínua), mas não se conhecem os parâmetros. Logo não pode ser K-S. Assim, como temos $n=40$, o teste mais indicado será o Shapiro-Wilks. (exemplos neste documento)

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

Cálculo do valor-p

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

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

D_{obs} é a estatística de teste

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 .

Observação:

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.

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, ou Shapiro-Wilks.**

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

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%.

Vamos fazer o exercício no RStudio. Mas primeiro, temos que o formular:

Hipótese a ser testada

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

$$H_0 : X \sim Exp(730) \quad vs \quad H_1 : X \not\sim Exp(730)$$

Ou seja, em H_0 colocamos a hipótese de ser uma exponencial com parâmetro 730 e em H_1 colocamos que não é uma exponencial com parâmetro 730.

Dados

- Total de dados: $n = 15$
- nível de significância = $\alpha = 0.10$

Como a distribuição é contínua e está completamente especificada, então vamos usar K-S.

No R:

```
# EXEMPLO 6
```

```
# H0:X segue uma distribuição Exponencial de média 730 horas
# contra
# H1:X Não segue uma distribuição Exponencial de média 730 horas

# amostra
amostra5 <- c(1476, 182, 246, 300, 499, 442, 98, 552, 20, 221, 1563, 796, 157, 36, 31)

# teste de Ajustamento de Kolmogorov-Smirnov
ks.test(amostra5, "pexp", rate=1/730)
```

“pexp” é a função distribuição de uma exponencial e Rate é $\frac{1}{\theta}$

O resultado é:

```
> ks.test(amostra5, "pexp", rate=1/730)
```

```
Exact one-sample Kolmogorov-Smirnov test
```

```
data: amostra5
D = 0.26946, p-value = 0.1881
alternative hypothesis: two-sided
```

Ou seja, o $p - value = 0.1881$

DECISÃO: Como $p - value$ é 0.1881, não é $\leq \alpha$, pois $\alpha = 0.1$, então não se rejeita 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.

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

$$H_0 : X \sim N(80, 6.95) \quad vs \quad H_1 : X \not\sim N(80, 6.95)$$

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:

```

# EXEMPLO 7

# amostra
amostra7 <- c(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)

#####
# EXEMPLO 7.1

# H0:X segue uma distribuição Normal de média 80 e desvio padrão 6.95
# contra
# H1:X Não segue uma distribuição Normal de média 80 e desvio padrão 6.95

# teste de Ajustamento de Kolmogorov-Smirnov
ks.test(amostra7, "pnorm", mean=80, sd=6.95)

```

Cujo resultado é

```
Asymptotic one-sample Kolmogorov-Smirnov test
```

```
data: amostra7
D = 0.16502, p-value = 0.1056
alternative hypothesis: two-sided
```

- valor- $p = 0.1056$

DECISÃO: Como $p - value$ é 0.1056, não é $\leq \alpha$, pois $\alpha = 0.05$, então não se rejeita 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.

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

$$H_0 : X \sim N(\mu, \sigma) \quad \text{contra} \quad H_1 : X \not\sim N(\mu, \sigma)$$

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 = $\alpha = 0.01$

Neste caso, não se conhecem os valores da média e do desvio padrão populacionais, logo pretende-se verificar se os dados provêm de uma população com distribuição Normal, ou seja, contínua, mas não está completamente especificada.

Como $n = 54 \geq 50$, então o teste adequado é o *Lilliefors*.

Observação: Se não se soubesse o tamanho da amostra, bastava fazer
`(length(salinidade))`

EXEMPLO 7.2

```
# H0:X segue uma distribuição Normal
# contra
# H1:X Não segue uma distribuição Normal

# teste de Ajustamento de Lilliefors
library(nortest)
lillie.test(amostra7)
```

DECISÃO: Como $p-value$ deu $0.005525 \leq \alpha$, pois $\alpha = 0.01$, então rejeita-se 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.

Note-se que:

(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.)

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

$$H_0 : X \sim \text{Normal} \quad \text{contra} \quad H_1 : X \not\sim \text{Normal}$$

Cálculo do valor-p

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

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

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 Normal.
- 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 Normal.

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 $n < 50$.

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

$$H_0 : X \sim \text{Normal} \quad \text{contra} \quad H_1 : X \not\sim \text{Normal}$$

Dados

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

Neste caso, não se conhecem os valores da média e do desvio padrão populacionais, logo pretende-se verificar se os dados provêm de uma população com distribuição Normal, ou seja, contínua, mas não está completamente especificada.

Como $n = 36 < 50$, então o teste a recorrer é o **Shapiro-Wilks**.

```
# EXEMPLO 8 Slides cap6-parte1

amostra8 <- c(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)

# H0:X segue uma distribuição Normal
# contra
# H1:X Não segue uma distribuição Normal

# teste de Ajustamento de Shapiro-Wilk
shapiro.test(amostra8)
```

DECISÃO: Como $p-value$ deu 0.05905 não é $\leq \alpha$, pois $\alpha = 0.01$, então não se rejeita 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.

Observação: Se $\alpha = 0.1$, então $p-value \leq \alpha$, então rejeita-se H_0 .