TESTE DE INDEPENDÊNCIA

6.3

MOTIVAÇÃO

- Quem tem bom desempenho em matemática também tem bom desempenho em física?
- Quem está exposto a um fator tem maior ou menor ocorrência de certa doença?
- A pane de um equipamento está associada a certa conduta?

Experimento

Notas de matemática e de física para 528 alunos.

		Matemática			
		Alta	Média	Baixa	Total
	Alta	56	71	12	139
Física	Média	47	163	38	248
	Baixa	14	42	85	141
	Total	117	276	135	528

Pergunta: Essas variáveis são independentes?

IDEIA

Se as variáveis forem independentes, então a distribuição conjunta delas é igual ao produto de suas marginais.

		Matemática			
		Alta	Média	Baixa	Total
	Alta	56	71	12	139
Física	Média	47	163	38	248
	Baixa	14	42	85	141
	Total	117	276	135	528

Estatisticamente

TESTE DE INDEPENDÊNCIA

Serve para testar se duas variáveis categóricas são independentes.

Vamos aprender o teste Qui-quadrado para independência de variáveis.

TESTE QUI-QUADRADO PARA INDEPENDÊNCIA

H_0: As variáveis são independentes
H_1: As variáveis não são independentes

$$e_{ij} = \frac{\text{total da linha i x total da coluna j}}{\text{total geral}}$$

Estatística de teste

$$Q^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

- o_{ij} é a frequência observada da linha i e coluna j
- e_{ij} é a frequência esperada da linha i e coluna j

em que r e s representam o número de linhas e de colunas, respectivamente.

Resultado: Se as frequências esperadas forem ao menos iguais a 5, então

$$Q^2 \sim \chi^2_{(r-1)(s-1)}.$$

Região crítica

$$RC = \{Q^2 : Q^2 \geq q_c\},$$

em que

$$P(\chi^2_{(r-1)(s-1)} \geq q_c) = \alpha$$

α é o nível de significância do teste.

Exemplo

Notas de matemática e de física para 528 alunos.

		Matemática			
		Alta	Média	Baixa	Total
	Alta	56	71	12	139
Física	Média	47	163	38	248
	Baixa	14	42	85	141
	Total	117	276	135	528

Pergunta: Essas variáveis são independentes?

Solução

		Matemática			
		Alta	Média	Baixa	Total
Física	Alta	56 (30,8)	71 (72,7)	12 (35,5)	139
	Média	47 (55,0)	163 (129,6)	38 (63,4)	248
	Baixa	14 (31,2)	42 (73,7)	85 (36,1)	141
	Total	117	276	135	528

$$Q^2_{\text{obs}} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 145,5$$

$$P(\chi^2_{(3-1)\times(3-1)} \geq q_c) = 0,05 \Rightarrow q_c = 9,49$$

$$RC = \{Q^2 : Q^2 \geq 9,49\} \Rightarrow Q^2_{\text{obs}} = 145,5 \in RC$$

Conclusão: Rejeitamos H_0, ou seja, há evidências para dizermos que as variáveis são dependentes.

Solução no R

```
> dados = as.table(matrix(c(56,71,12,47,163,38,14,42,85), byrow = T, nrow = 3))
> row.names(dados) = c("F Alta", "F Média", "F Baixa")
> colnames(dados) = c("M Alta", "M Média", "M Baixa")
> dados
        M Alta M Média M Baixa
F Alta      56      71      12
F Média     47     163      38
F Baixa     14      42      85
>
> chisq.test(dados)

	Pearson's Chi-squared test

data:  dados
X-squared = 145.78, df = 4, p-value < 2.2e-16
```

GRÁFICO DE PROBABILIDADE NORMAL

6.1

MOTIVAÇÃO

Muitos resultados estatísticos são baseados na suposição de normalidade dos dados.

Pergunta: Como verificar tal suposição?

Resposta

- Fazer um histograma.
  - Problema: requer um número razoável de amostras.
- Teste de Kolmogorov-Smirnov
  - Problema: requer um número razoável de amostras.
- Teste de Shapiro-Wilk ou Teste de Anderson-Darling
  - Uma alternativa para pequenas amostras.
- Análise gráfica: PP-plot ou QQ-plot
  - Uma alternativa para pequenas amostras.

GRÁFICO PAPEL DE PROBABILIDADE

- Ordenar o conjunto de dados.
- Para cada observação calcular

$$w_j = \frac{j - c}{n + 1 - 2c}$$

em que c = ⅜ se n ≤ 10, e c = ½, caso contrário.

- Construir o gráfico dos pontos $\mathbf{(x_{(j)}, \Phi^{-1}(w_j))}$.
- Se os pontos do gráfico se aproximarem de uma reta, então é aceitável o modelo normal.

Exemplo

(Montgomery & Runger, 2009) Foram coletadas dez observações sobre o tempo (em minutos) efetivo de serviço de uma bateria de um computador pessoal: 176, 191, 214, 220, 205, 192, 201, 190, 183, 185. Podemos dizer que é razoável supor que o tempo de serviço dessa bateria tem distribuição normal?

Exemplo

(Montgomery & Runger, 2009) Foram coletadas dez observações sobre o tempo (em minutos) efetivo de serviço de uma bateria de um computador pessoal: 176, 191, 214, 220, 205, 192, 201, 190, 183, 185. Podemos dizer que é razoável supor que o tempo de serviço dessa bateria tem distribuição normal?

j	1	2	3	4	5	6	7	8	9	10
$x_{(j)}$	176	183	185	190	191	192	201	205	214	220
w_j	0,06	0,16	0,26	0,35	0,45	0,55	0,65	0,74	0,84	0,94
$\Phi^{-1}(w_j)$	-1,55	-1,00	-0,66	-0,38	-0,12	0,12	0,38	0,66	1,00	1,55

Gráfico Papel de Probabilidade

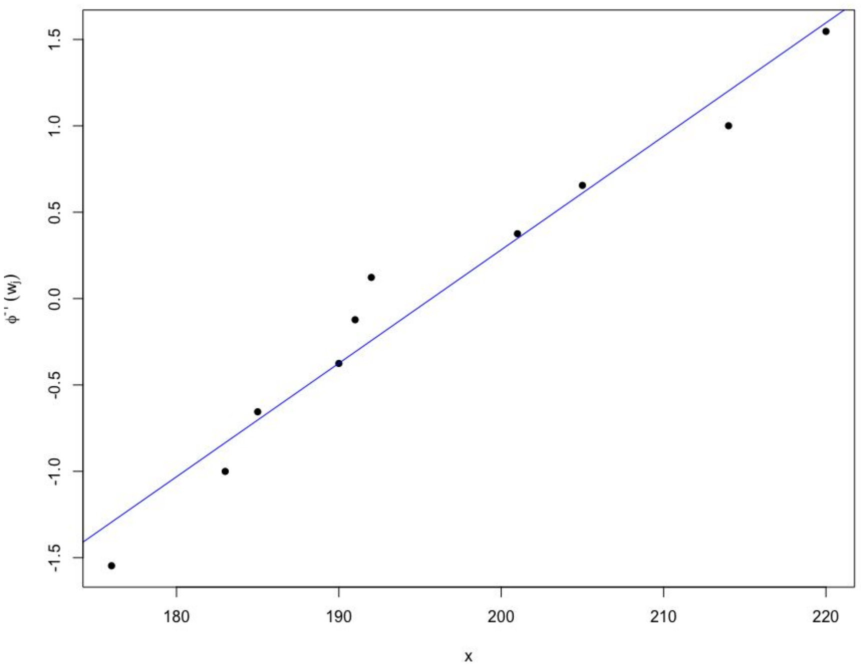


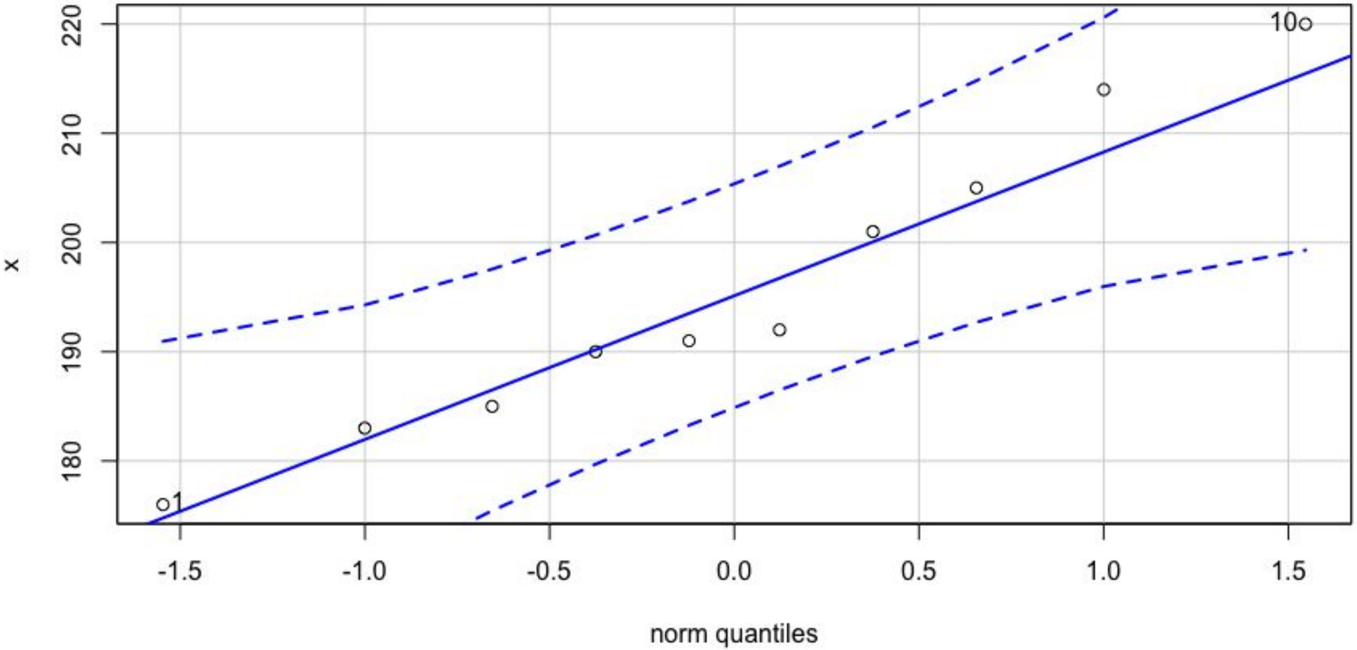
Gráfico QQ-plot

A diferença entre o PP-plot e o QQ-plot é a ordem dos pares ordenado. Para o QQ-plot usamos $(\Phi^{-1}(w_j), x_{(j)})$.

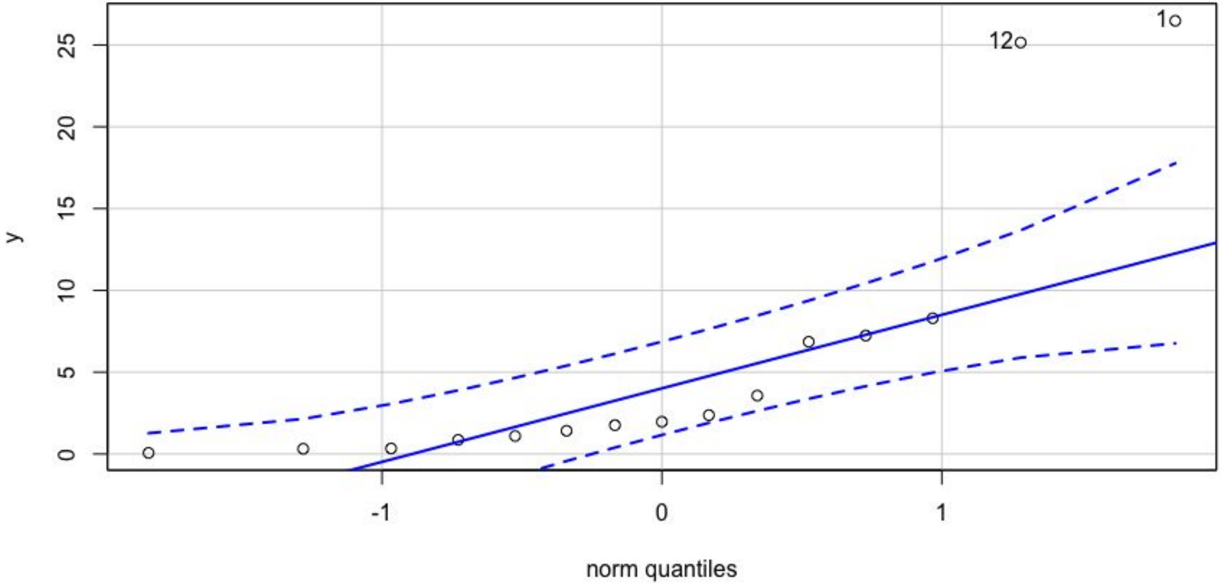
No R

```
>
> #install.packages("car")
> library("car")
>
> x = c( 176, 183, 185, 190, 191, 192, 201, 205, 214, 220)
> qqPlot(x, dist = 'norm')
[1] 10  1
```

Gráfico Papel de Probabilidade



Exemplo de não normalidade



Exemplo de não normalidade