Teste de Aderência

Exemplo

Um modelo genético especifica que animais de certa população devam estar classificados em quatro categorias, com probabilidades $p_1 = 0.656$, $p_2 = 0.093$, $p_3 = 0.093$ e $p_4 = 0.158$. Dentre 197 animais, obtivemos as seguintes frequências observadas: $O_1 = 125$, $O_2 = 18$, $O_3 = 20$ e $O_4 = 34$. Teste se esses dados estão de acordo com o modelo genético postulado.
Fonte: Morettin & Bussab, Estatística Básica 5ª edição, pág 395.

Teste de Aderência

Temos as probabilidades de cada categoria, e o número observado de indivíduos. Então o número esperado de indivíduos numa categoria é dado por $n \cdot p_c$, ou seja,
$E_1 = n \cdot p_1 = 197 \cdot 0.656 = 129.23$, $E_2 = E_3 = 197 \cdot 0.093 = 18.32$ e $E_4 = 197 \cdot 0.158 = 31.13$.

A estatística do teste de aderência é

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

Teste de Aderência

Nossa estatística observada Q_0 é dada por:

$$Q_0 = \frac{(125 - 129.23)^2}{129.23} + \frac{(18 - 18.32)^2}{18.32} + \frac{(20 - 18.32)^2}{18.32} + \frac{(34 - 31.13)^2}{31.13}$$

$$Q_0 = 0.5627$$

Note que a região crítica para os testes baseados na χ^2 é dada por $P(Q > q) = \alpha$. Para $\alpha = 0.05$ e 3 graus de liberdade, o valor de $q = 7.8147$. Portanto, não rejeitamos a hipótese de que os dados seguem a distribuição proposta.

Teste de Aderência

Exemplo

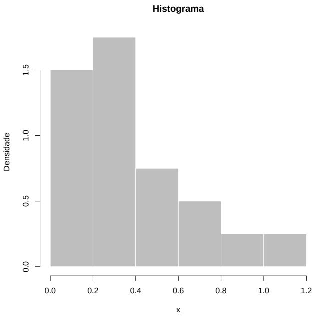
Teste, para o nível $\alpha = 0.01$, se os dados abaixo vem de uma distribuição exponencial com média 0.5.

0.378	0.391	0.458	0.063	0.009
1.007	0.470	0.368	0.831	0.387
0.228	0.389	0.627	0.480	0.093
0.123	0.089	0.646	0.093	0.400

Fonte: Morettin & Bussab, Estatística Básica 5ª edição, pág 409.

Teste de Aderência

Antes de aplicar o teste, observe o histograma dos dados:



É razoável dizer que eles têm distribuição exponencial?

Teste de Aderência

Para distribuições contínuas, devemos construir a tabela de valores esperados/observados a partir dos quartis esperados/observados. Note que queremos testar se os dados têm distribuição exponencial, com $\lambda = 2$. Temos que os quantis teóricos são dados por:

$$q_1(X) \text{ é tal que } \int_0^{q_1} \lambda e^{-\lambda x} = 0.25$$

$$q_2(X) = \text{Med}(X) \text{ é tal que } \int_0^{q_2} \lambda e^{-\lambda x} = 0.5$$

$$q_3(X) \text{ é tal que } \int_0^{q_3} \lambda e^{-\lambda x} = 0.75$$

Teste de Aderência

Para $\lambda = 2$, temos que $q_1 = 0.1438$, $q_2 = 0.3466$ e $q_3 = 0.6931$.

Defina as categorias A_1, A_2, A_3 e A_4, onde

- um elemento $x \in A_1$ se $x < q_1(X)$,
- $x \in A_2$ se $q_1(X) < x < q_2(X)$,
- $x \in A_3$ se $q_2(X) < x < q_3(X)$ e
- $x \in A_4$ se $x > q_3(X)$.

Se a hipótese nula é verdadeira (isto é, os dados têm distribuição exponencial(2)), então a proporção esperada de cada categoria é $1/4$.

Teste de Aderência

Construimos então a tabela com as frequências observadas e esperadas:

	A_1	A_2	A_3	A_4	Total
O_i	6	1	11	2	20
E_i	5	5	5	5	20

A estatística observada Q_0 é dada por:

$$Q_0 = \frac{(6-5)^2}{5} + \frac{(1-5)^2}{5} + \frac{(11-5)^2}{5} + \frac{(2-5)^2}{5} = 12.4$$

que tem 3 graus de liberdade. Como $12.4 > 7.814 = \chi^2_3(0.95)$, rejeito $H_0 \Rightarrow$ os dados não tem distribuição exponencial(2).

Teste de Independência

Exemplo

Um inspetor de qualidade toma uma amostra de 220 artigos num centro de distribuição. Se sabe que cada produto pode vir de uma de três fábricas e pode ou não estar defeituoso. O inspetor avalia todos os produtos e obtém os seguintes resultados:

	F_1	F_2	F_3	
D	8	15	11	34
ND	62	67	57	186
	70	82	68	220

Ser defeituoso independe da fábrica?

Teste de Independência

A hipótese de independência dos eventos é dada por $H_0 : P(A \cap B) = P(A)P(B)$. Obtemos a tabela esperada, sob a hipótese nula, calculando a probabilidade de cada casela:

$$E_{11} = \frac{70 \times 34}{220} = 10.810 \qquad E_{21} = \frac{70 \times 186}{220} = 59.180$$

$$E_{12} = \frac{82 \times 34}{220} = 12.673 \qquad E_{22} = \frac{82 \times 186}{220} = 69.327$$

$$E_{13} = \frac{68 \times 34}{220} = 10.509 \qquad E_{23} = \frac{68 \times 186}{220} = 57.490$$

Teste de Independência

A tabela esperada dos dados é simplesmente

	F_1	F_2	F_3	
D	10.81	12.67	10.51	34
ND	59.18	59.33	57.49	186
	70	82	68	220

Teste de Independência

A estatística observada do teste é:

$$Q_0 = \frac{(8-10.81)^2}{10.81} + \ldots + \frac{(57-57.49)^2}{57.49} = 1.398$$

Note que no teste de independência, temos $(r-1)(s-1)$ graus de liberdade, onde r e s são o número de linhas e de colunas. Então temos 2 graus de liberdade, e o *p-value* do teste é 0.497, ou seja, não rejeitamos a hipótese de independência entre o eventos "peça defeituosa" e "peça da fábrica i".

Teste de Homogeneidade

Exemplo

Suponha que o inspetor do exemplo anterior resolveu repetir o experimento, mas desta vez, ao invés de tomar 220 artigos ao acaso, resolveu colher uma amostra de exatamente 80 artigos selecionados ao acaso dentro de cada uma das três fábricas. Os dados colhidos foram:

	F_1	F_2	F_3	
D	8	15	11	34
ND	72	65	69	206
	80	80	80	240

Há diferença entre as fábricas na proporção de defeituosos?

Teste de Homogeneidade

A hipótese agora é $H_0 : P_1 = P_2 = P_3$, ou seja, cada fábrica tem a mesma população. Temos que as proporções de itens defeituosos são, respectivamente, $\hat{p}_1 = 8/80 = 0.1$, $\hat{p}_2 = 15/85 = 0.1875$ e $\hat{p}_3 = 0.1375$. As frequências esperadas são:

$$E_{11} = E_{12} = E_{13} = \frac{80 \times 34}{240} = 11.33$$

$$E_{21} = E_{22} = E_{23} = \frac{80 \times 206}{240} = 68.67$$

Teste de Homogeneidade

A estatística observada do teste é

$$Q_0 = \frac{(8-11.33)^2 + (15-11.33)^2 + (11-11.33)^2}{11.33}$$
$$+\frac{(72-68.67)^2 + (65-68.67)^2 + (69-68.67)^2}{68.67}$$
$$= 0.978 + 1.188 + 0.009 + 0.161 + 0.196 + 0.001 = 2.537$$

O quantil da χ^2 com 2 graus de liberdade, tal que $P(Q > q) = 0.05$ sob a hipótese nula, é 5.991. Como $Q_0 = 2.537 < 5.991$, não rejeitamos a hipótese das populações serem iguais (isto é, a probabilidade de uma peça ser defeituosa é a mesma nas três fábricas).

Análise de Regressão

Exemplo

Considere as seguintes variáveis, X e Z, que representam a idade e a acuidade visual, respectivamente.

x	z	x	z	x	z	x	z	x	z
20	90	25	100	30	70	35	90	40	90
20	100	25	90	30	90	35	80	40	90
20	80	25	80	30	90	35	70	40	60
20	90	25	90	30	80	35	90	40	80

Fonte: Morettin & Bussab, Estatística Básica 5ª edição, pág 441.

Exemplo

(a) Encontre a reta de quadrados mínimos $\hat{z}_i = \alpha + \beta x_i$, onde z mede a acuidade visual e x a idade do i-ésimo indivíduo.

(b) Interprete o significado de α e β nesse problema.

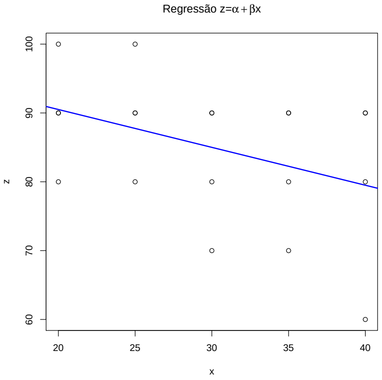
(a) Temos que os estimadores para α e β são dados por:

$$\hat{\alpha} = \bar{z} - \hat{\beta}\bar{x} \qquad \hat{\beta} = \frac{\sum_{i=1}^{n} x_i z_i - n\bar{x}\bar{z}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Note agora que $n = 20$, $\bar{x} = 30$, $\bar{z} = 85$, $\sum_{i=1}^{n} x_i z_i = 50450$ e $\sum_{i=1}^{n} x_i^2 = 19000$. Com isso, temos que:

$$\hat{\beta} = \frac{50450 - 20 \cdot 30 \cdot 85}{19000 - 20 \cdot 30^2} = -\frac{11}{20} \Rightarrow \hat{\alpha} = 85 + \frac{11}{20}30 = 101.5$$

(a) (cont.) O gráfico de dispersão dos dados, com a reta ajustada de regressão em azul, é dado por:



(b) Neste problema, a interpretação dos parâmetros é a seguinte:

- α é o intercepto. Ele representa a acuidade visual na idade $z = 0$, se fosse possível medí-la. Mas $z = 0$ não faz parte do intervalo $[20, 40]$ de idades observadas, e não faz sentido falar em acuidade visual de recém nascidos, então sua interpretação deve ser feita com cuidado.
- β é mais interessante. O fato dele ser negativo significa que, a medida que os indivíduos envelhecem, sua acuidade visual diminui. Ela diminui na razão de -0.55 por ano, pelo modelo ajustado.

Exemplo

Um estudo sobre duração de certas operações está investigando o tempo requerido (em segundos) para acondicionar objetos e o volume (em dm^3) que eles ocupam. Uma amostra foi observada e obtiveram-se os seguintes resultados:

Tempo	10,8	14,4	19,6	18,0	8,4
Volume	20,39	24,92	34,84	31,72	13,59
Tempo	15,2	11,0	13,3	23,1	
Volume	30,87	17,84	23,22	39,65	

Adaptado de: Morettin & Bussab, Estatística Básica 5ª edição, pág 446.

Análise de Regressão

Exemplo

(a) Estime a reta de regressão do tempo em função do volume.

(b) Faça o diagrama de dispersão dos dados.

(a) Temos que os estimadores para α e β são dados por:

$$\hat{\alpha} = \bar{v} - \hat{\beta}\bar{t} \qquad \hat{\beta} = \frac{\sum_{i=1}^{n} t_i v_i - n\bar{t}\bar{v}}{\sum_{i=1}^{n} t_i^2 - n\bar{t}^2}$$

Note agora que $n = 9$, $\bar{t} = 14.866$, $\bar{v} = 26.337$, $\sum_{i=1}^{n} t_i v_i = 3837.245$ e $\sum_{i=1}^{n} t_i^2 = 2165.26$. Com isso, temos que:

$$\hat{\beta} = \frac{3837.25 - 9 \cdot 14.87 \cdot 26.34}{2165.26 - 9 \cdot 14.87^2} = 1.78 \Rightarrow \hat{\alpha} = -0.107$$

Análise de Regressão

(a) (cont.) O gráfico de dispersão dos dados, com a reta ajustada de regressão em azul, é dado por: