Análise de Aderência e Associação

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}, \qquad (14.1)$$

2. Testes de Homogeneidade

Considere o seguinte exemplo.

Exemplo 14.2. Uma prova básica de Estatística foi aplicada a 100 alunos de Ciências Humanas e a 100 alunos de Ciências Biológicas. As notas são classificadas segundo os graus A, B, C, D e E (onde D significa que o aluno não recebe créditos e E indica que o aluno foi reprovado). Os resultados estão na Tabela 14.2.

Tabela 14.2: Resultados da aplicação de uma prova de Estatística a 100 alunos de Ciências Humanas e 100 alunos de Biologia.

Aluno de	Grau					Total
	A	B	C	D	E	
C. Humanas	15	20	30	20	15	100
C. Biológicas	8	23	18	34	17	100
Total	23	43	48	54	32	200

Exemplo 14.2 (continuação) Considerando P_1 como a população de alunos de Ciências Humanas e P_2 a dos alunos de Ciências Biológicas, nosso objetivo é testar a hipótese

$$H_0 : P_1 = P_2,$$

usando os resultados amostrais da Tabela 14.2. Para isso, precisamos encontrar os valores esperados n^*_{ij}, para aplicar a fórmula (14.1).

Tabela 14.7: Porcentagens estimadas das classes para cada população.

Aluno de	Grau					Total
	A	B	C	D	E	
C. Humanas	15	20	30	20	15	100
C. Biológicas	8	23	18	34	17	100
Total	11,5	21,5	24	27	16	100

Tabela 14.8: Freqüências absolutas sob H_0 (n^*_{ij}).

Aluno de	Grau					Total
	A	B	C	D	E	
C. Humanas	11,5	21,5	24	27	16	100
C. Biológicas	11,5	21,5	24	27	16	100
Total	23	43	48	54	32	200

Desse modo, encontramos os valores esperados n^*_{ij}, que podem ser substituídos em (14.1), obtendo-se

$$\chi^2_{obs} = \frac{(15-11{,}5)^2}{11{,}5} + \ldots + \frac{(15-16)^2}{16} + \frac{(8-11{,}5)^2}{11{,}5} + \ldots + \frac{(17-16)^2}{16} = 9{,}09.$$

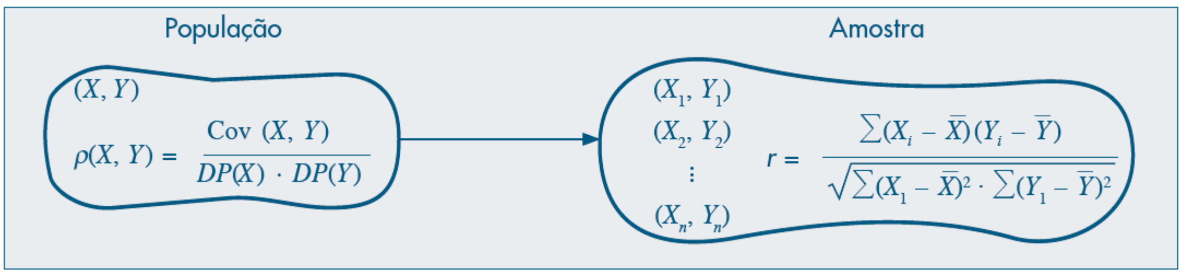
Da Tabela IV, com $\alpha = 0{,}05$ e 4 graus de liberdade encontramos $\chi^2_c = 9{,}488$, o que leva à não-rejeição de H_0, ou seja, a distribuição das notas é a mesma para as duas populações.

Observe que os valores esperados na Tabela 14.8 podem ser obtidos de $n^*_{ij} = (n_{i.}\, n_{.j})/n$.

14.5 Teste Para o Coeficiente de Correlação

O teste apresentado na seção anterior é adequado para averigüar a independência de duas variáveis qualitativas. Vimos, na seção 4.5, que para variáveis quantitativas o coeficiente de correlação é uma medida de associação mais adequada. Usualmente, podemos determinar o coeficiente de correlação para uma amostra, pois desconhecemos esse valor na população. Uma população que tenha duas variáveis não-correlacionadas pode produzir uma amostra com coeficiente de correlação diferente de zero. Para testar se a amostra foi colhida de uma população para a qual o coeficiente de correlação entre duas variáveis é nulo, precisamos obter a distribuição amostral da estatística r, definida em (4.7). Esquematicamente, temos a situação da Figura 14.2.

Figura 14.2: Coeficiente de correlação para população e amostra.



Seja $\rho = \rho(X, Y)$ o verdadeiro coeficiente de correlação populacional desconhecido. Vamos apresentar a distribuição amostral de r para duas condições da população: $\rho = 0$ e $\rho \neq 0$.

Exemplo 14.8. *Teste para* $\rho = \rho_0$. Durante muito tempo, o coeficiente de correlação entre a nota final num curso de treinamento de operários e sua produtividade, após seis meses do curso, resultou ser 0,50. Foram introduzidas modificações no curso, com o intuito de aumentar a correlação. Se o coeficiente de correlação de uma amostra de 28 operários submetidos ao novo curso foi 0,65, você diria que os objetivos da modificação foram atingidos?

A. *Hipóteses*

X: resultado no teste; Y: produtividade;

$H_0 : \rho(X, Y) = 0{,}50$;

$H_1 : \rho(X, Y) > 0{,}50$;

B. *Estatística do Teste*

R. Fisher sugeriu a seguinte transformação para a estatística r:

$$\xi = \frac{1}{2}\,\ell n\frac{1 + r}{1 - r}, \tag{14.7}$$

que tem uma distribuição muito próxima de uma normal $N(\mu_\xi, \sigma^2_\xi)$, com

$$\mu_\xi = \frac{1}{2}\,\ell n\frac{1 + \rho_0}{1 - \rho_0}, \quad \sigma^2_\xi = \frac{1}{n - 3}, \tag{14.8}$$

sendo n o tamanho da amostra $(X_1, Y_1), ..., (X_n, Y_n)$ e ρ_0 o valor do parâmetro populacional. A aproximação não vale para $\rho = -1$ ou $\rho = 1$. Além disso, para $\rho = 0$, temos um teste exato, que será visto no próximo exemplo. No nosso caso, sob a hipótese H_0, ξ terá distribuição aproximadamente normal, com

$$\mu_\xi = \frac{1}{2}\ \ell n\ \frac{1 + 0{,}5}{1 - 0{,}5} = 0{,}549, \quad \sigma_\xi^2 = \frac{1}{25} = 0{,}04.$$

C. *Região Crítica*

Como a hipótese alternativa sugere uma região crítica unilateral à direita, e como $\xi \sim N\,(0{,}549;\ 0{,}04)$, vem que a RC para ξ, no nível de significância $\alpha = 0{,}05$, será

$$\text{RC} = \{\xi : \xi > 0{,}549 + 1{,}654\sqrt{0{,}04}\} = \{\xi : \xi > 0{,}878\}.$$

D. *Resultado da Amostra*

Como $r = 0{,}65$, vem que

$$\xi_0 = \frac{1}{2}\ \ell n \frac{1 + 0{,}65}{1 - 0{,}65} = 0{,}774.$$

E. *Conclusão*

Como $\xi_0 \notin \text{RC}$, aceitamos H_0, ou seja, não existe evidência de que o coeficiente de correlação tenha aumentado.

Exemplo 14.9. *Teste para* $\rho = 0$. Queremos testar se existe ou não correlação entre o número de clientes e os anos de experiência de agentes de seguros. Sorteamos cinco agentes e observamos as duas variáveis. Os dados estão na Tabela 14.12. Qual seria a conclusão, baseando-se nesses dados?

Tabela 14.12: Anos de experiência para cinco agentes de seguros.

Agente	A	B	C	D	E
Anos de Experiência	2	4	5	6	8
Número de Clientes	48	56	64	60	72

A. *Hipóteses*

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

B. *Estatística do Teste*

Para amostras retiradas de uma população para a qual $\rho = 0$, pode-se provar que a estatística

$$T = r\sqrt{\frac{n-2}{1-r^2}} \tag{14.9}$$

tem distribuição t de Student com $n - 2$ graus de liberdade. No nosso exemplo, a estatística terá distribuição $t(3)$.

C. *Região Crítica*

Por ser um teste bilateral, consultando a Tabela V, teremos para $\alpha = 0{,}10$,

$$\text{RC} = (-\infty, -2{,}353] \cup [2{,}353, +\infty).$$

D. *Resultado da Amostra*

Calculando o coeficiente de correlação para os dados acima, obtemos $r = 0{,}95$; logo,

$$t_0 = (0{,}95)\sqrt{\frac{3}{1-(0{,}95)^2}} = 5{,}254.$$

E. *Conclusão*

Como $t_0 \in \text{RC}$, rejeitamos H_0, isto é, existe dependência entre anos de experiência e números de clientes.

Nesse caso seria conveniente construir um intervalo de confiança para ρ. Observe que, se $\rho \neq 0$, devemos usar a estatística ξ de (14.7). Portanto, se tomarmos por exemplo $\gamma = 0,95$, devemos procurar dois números ξ_1 e ξ_2 para ξ, tais que

$$P(\xi_1 < \xi < \xi_2) = 0,95.$$

Como $\xi \sim N(\mu_\xi, 1/2)$, podemos escrever

$$P\left(\frac{\xi_1 - \mu_\xi}{\sqrt{1/2}} < \frac{\xi - \mu_\xi}{\sqrt{1/2}} < \frac{\xi_2 - \mu_\xi}{\sqrt{1/2}}\right) = 0,95,$$

ou seja,

$$P(-1,96 < Z < 1,96) = 0,95,$$

com $Z \sim N(0,1)$. Logo, o intervalo para μ_ξ é

$$\mathrm{IC}(\mu_\xi;\ 0,95) = \xi_0 \pm 1,96\sqrt{1/2}.$$

Mas,

$$\xi_0 = \frac{1}{2}\,\ell n\,\frac{1 + 0,95}{1 - 0,95} = 1,832,$$

logo

$$\mathrm{IC}(\mu_\xi;\ 0,95) = 1,832 \pm 1,384 = (0,448;\ 3,216).$$

Como

$$\mu_\xi = \frac{1}{2}\,\ell n\,\frac{1 + \rho}{1 - \rho},$$

e uma expressão semelhante vale para os extremos do intervalo, podemos obter as operações inversas para encontrar os extremos do intervalo para ρ. Assim, de

$$0,448 = \frac{1}{2}\,\ell n\,\frac{1 + r}{1 - r}$$

obtemos

$$r = \frac{e^{0,896} - 1}{e^{0,896} + 1} = 0,420,$$

e de

$$3,216 = \frac{1}{2}\,\ell n\,\frac{1 + r}{1 - r}$$

obtemos

$$r = \frac{e^{6,432} - 1}{e^{6,432} + 1} = 0,997.$$

Finalmente, obtemos

$$\text{IC}(\rho;\ 0,95) = (0,420;\ 0,997).$$