Teste Qui-Quadrado para Independência
aula teórica das disciplinas MAE0110 e MAE0116 da USP
"Noções de Estatística"
Ministrante Prof. Dr. Vladimir Belitsky

23 de novembro de 2020

Exemplo 1. Um exemplo típico do tema Teste de Independência.

Os resultados da classificação de 100 pessoas segundo a cor dos olhos e a cor do cabelo foram os seguintes:

Cor do cabelo	Cor dos olhos		
	Castanhos	Azuis	Cinza
Claro	13	18	9
Escuro	37	12	11

Deseja-se verificar, com o nível de significância de 5%, se cor de olhos e cor de cabelos são atributos independentes.

Exemplo 1; formalismo.

O objetivo do exemplo formula-se assim:

deseja-se escolher entre as **hipóteses**

H: "os atributos "cor de olhos" e "cor de cabelos" são independentes";

A: "os atributos "cor de olhos" e "cor de cabelos" não são independentes (são dependentes, em outras palavras)".

Os valores 13, 18, 9, 37, 12, 11 chamam-se **frequencias observadas**. A notação genérica para estes é

$$o_{11}, o_{12}, o_{13}, o_{21}, o_{22}, o_{23}$$

(a indexação segue a regra usada em matrizes: o primeiro índice corresponde à linha da posição do valor, e o segundo número correpsonde à coluna).

Exemplo 1; o primeiro passo da solução.

No primeiro passo, calcula-se os "Totais" por linhas e por colunas e o "Total Global"que fica no canto direito de baixo:

Cor do cabelo	Cor dos olhos			**Total por linha**
	Castanhos	Azuis	Cinza	
Claro	13	18	9	40
Escuro	37	12	11	60
Total por coluna	50	30	20	100

Exemplo 1; o segundo passo da solução.

No segundo passo, as frequencias observadas são substituidas por **frequencias esperadas**; a notação genérica para estes é $e_{11}, e_{12}, e_{13}, e_{21}, e_{22}, e_{23}$. Estas são os valores hipotéticos que

- seriam observados caso H fosse válida,
- seriam observados se não hovesse aleatoriedade,
- dariam os mesmo Totais que as frequencias observadas.

Cor do cabelo	Cor dos olhos			**Total por linha**
	Castanhos	Azuis	Cinza	
Claro	e_{11}	e_{12}	e_{13}	40
Escuro	e_{21}	e_{22}	e_{23}	60
Total por coluna	50	30	20	100

Exemplo 1; o segundo passo da solução.

Eis a idéia central:

Se Cor dos Olhos não dependesse da Cor dos Cabelos, então as 50 pessoas com olhos castanhos seriam divididas em loiras e morenas nas mesmas proprções que as 30 pessoas com olhos azuis seriam divididas em loiras e morenas, e na mesmas proporções que as 20 pessoas com olhos cinza seriam divididas em loiras e morenas. Estas três proporções (iguais entre si) seriam obrigadas a coincidir também com as proporções totais de loiras e morenas. Já que no total, as loiras e as morenas estão em proporções como $\frac{40}{100}$ para $\frac{60}{100}$, então entre as 50 pessoas com olhos castanhos, deve ter $50 \times \frac{40}{100}$ loiras e $50 \times \frac{60}{100}$ morenas; entre as 30 pessoas com olhos azuis, deve ter $30 \times \frac{40}{100}$ loiras e $30 \times \frac{60}{100}$ morenas; e entre as 20 pessoas com olhos cinza, deve ter $20 \times \frac{40}{100}$ loiras e $20 \times \frac{60}{100}$ morenas.

Exemplo 1; o segundo passo da solução.

Cor do cabelo	Cor dos olhos			**Total por linha**
	Castanhos	Azuis	Cinza	
Claro	$\frac{40\times 50}{100}$	$\frac{40\times 30}{100}$	$\frac{40\times 20}{100}$	40
Escuro	$\frac{60\times 50}{100}$	$\frac{60\times 30}{100}$	$\frac{60\times 20}{100}$	60
Total por coluna	50	30	20	100

Observação: As frequencias esperadas não são obrigadas a serem números interios; na verdade, na maioria dos casos práticos, estes não serão inteiros.

Exemplo 1: o terceiro passo da solução.

No terceiro passo, calcula-se a distância entre as frequencias observadas e frequencias esperadas via a fórmula

$$(\chi^2)_{obs} = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

No caso do exemplo,

$$(\chi^2)_{obs} = \frac{\left(13 - \frac{40\times 50}{100}\right)^2}{\frac{40\times 50}{100}} + \frac{\left(18 - \frac{40\times 30}{100}\right)^2}{\frac{40\times 30}{100}} + \frac{\left(9 - \frac{40\times 20}{100}\right)^2}{\frac{40\times 20}{100}} +$$

$$+ \frac{\left(37 - \frac{60\times 50}{100}\right)^2}{\frac{60\times 50}{100}} + \frac{\left(12 - \frac{60\times 30}{100}\right)^2}{\frac{60\times 30}{100}} + \frac{\left(11 - \frac{60\times 20}{100}\right)^2}{\frac{60\times 20}{100}} = 9,288$$

Exemplo 1: o terceiro passo da solução.

O valor $(\chi^2)_{obs} = 9,288$ compara-se com o limiar ℓ que recorta a cauda da área $\alpha = 5\%$ da distribuição Qui-Quadrado com 2 graus de liberdade. O número de graus de liberdade calcula-se pela da seguinte maneira: toma-se a tabela sem a linha "totais por coluna" e sem a coluna "totais por linha", quer dizer, toma-se a tabela somente com os dados observados, e calcula-se o produto

$$(\text{número de linhas} - 1) \times (\text{número de colunas} - 1)$$

este valor é o número de gradus de liberdade para o problema. No caso agora tratado, a tabela com valores observados tem duas linhas e três colunas. Logo, o número correspondente de graus de liberdade é $(2-1)(3-1) = 2$.

Observe que o número de graus de liberdade calcula-se por métodos defirentes em Teste de Aderência e em Teste de Independência.

Exemplo 1: o terceiro passo da solução.

Pela tabela das distribuições Qui-Quadrado, $\ell = 5,992$ (a tabela está na transparência seguinte; o valor $5,992$ fica na intersecção da coluna "5%" com a linha marcada por "2 G.L.").

Como $\left(\chi^2\right)_{obs} = 9,288 > 5,992 = \ell$, então a hipótese nula (que alega a independência) é rejeitada.

Conclusão: O Teste de Independência, sendo aplicado aos dados do Exemplo 1 e com o nível de significância 5%, não confirmou a independência dos atributos "cor de olhos" e "cor de cabelos".

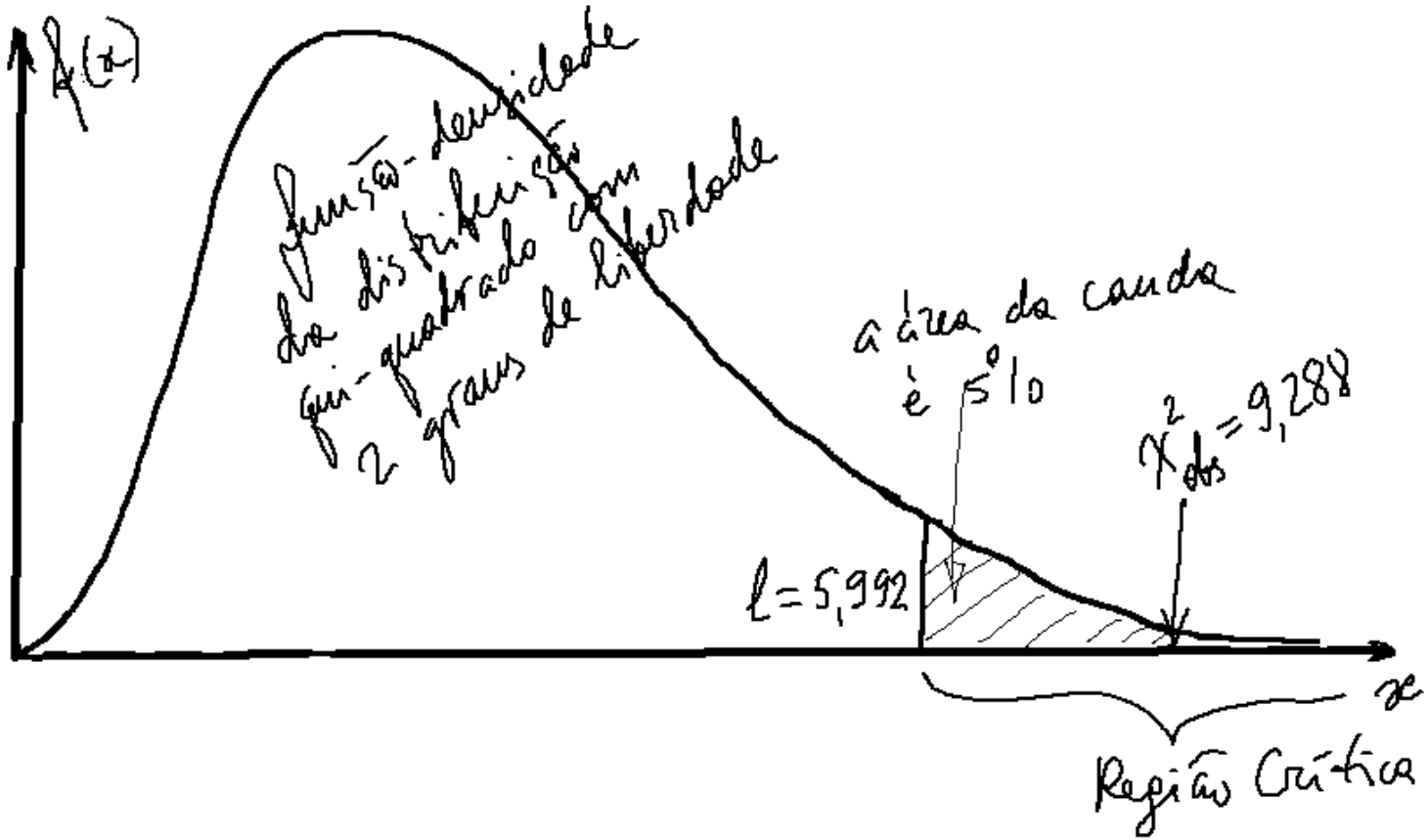
Tabela 1 - Distribuição Qui-Quadrado (χ^2)

Corpo da tabela fornece os valores de ℓ tais que $P(\chi^2 \geq \ell) = \alpha$

G.L.	α=99%	90%	80%	70%	10%	5%	1%	0.1%
1	0.000	0.016	0.064	0.148	2.705	3.841	6.635	10.828
2	0.020	0.211	0.446	0.713	4.605	5.992	9.210	13.816
3	0.115	0.584	1.005	1.424	6.251	7.815	11.345	16.266
4	0.297	1.064	1.649	2.195	7.779	9.488	13.277	18.467
5	0.554	1.610	2.342	3.000	9.236	11.070	15.086	20.515
6	0.872	2.204	3.070	3.828	10.645	12.592	16.812	22.458
7	1.239	2.833	3.822	4.671	12.017	14.067	18.475	24.322
8	1.647	3.490	4.594	5.527	13.362	15.507	20.090	26.125
9	2.088	4.168	5.380	6.393	14.684	16.919	21.666	27.877
10	2.558	4.865	6.179	7.267	15.987	18.307	23.209	29.588
11	3.054	5.578	6.989	8.148	17.275	19.675	24.725	31.264
12	3.571	6.304	7.807	9.034	18.549	21.026	26.217	32.910
13	4.107	7.042	8.634	9.926	19.812	22.362	27.688	34.528
14	4.660	7.790	9.467	10.822	21.064	23.685	29.141	36.123
15	5.229	8.547	10.307	11.721	22.307	24.996	30.578	37.697

Observação: Uma tabela mais detalhada da distribuição Qui-Quadrado está disponibilizada junto com listas de exercícios referentes ao tema Teste Qui-Quadrado.

Exemplo 1: a ilustração do terceiro passo da solução.



Notações e linguagem.

H chama-se **hipótese nula**, e A **hipótese alternativa**. O semi-eixo $[\ell, +\infty)$ chama-se **Região Crítica**. $(\chi^2)_{obs}$ chama-se **a observação** ou **o valor observado** da **Estatítsica Qui-Quadrado**, onde a palavra "estatística" possui seu sentido exato na Teoria Estatística, mas para nos, essa deve ser interpretada como "variável aleatória".
A regra de decisão pode ser reformulada assim:

se $(\chi^2)_{obs}$ não pertencer à Região Crítica $\Rightarrow$ aceitar H,
se $(\chi^2)_{obs}$ pertencer à Região Crítica $\Rightarrow$ rejeitar H.

A regra pode errar em dois sentidos: rejeitar H quando esta está válida (chama-se **Erro do Tipo I**), e rejeitar A quando esta está válida (chama-se **Erro do Tipo II**). A probabilidade do Erro do Tipo I é exatamente (e pela própria contrução da regra, alias) o **nível de significância**. A probabilidade do Erro do Tipo II não pode ser expressa por um número só, pois a hipótese A não especifica a forma da dependência: para cada forma de dependência, haveria seu valor da probabilidade do Erro do Tipo II.

Teoria geral.

Ao denotar por

$$X_{11}, X_{12}, \ldots, X_{21}, X_{22}, \ldots$$

as frequencias a serem vistas caso H for válida, tem-se que a variável aleatória

$$\chi^2 = \sum_{i,j} \frac{(X_{ij} - e_{ij})^2}{e_{ij}}$$

tem (aproximadamente) a distribuição Qui-Quadrado com o número de Graus de Liberdade igual a

(número das linhas da tabela-1)$\times$(número das colunas da tabela-1)

Isto leva (via o raciocínio parecido com aquele que desenvolvemos no Teste de Aderência) à seguinte

Regra de Decisão: para dado α, nível de significância do teste, achar a limiar ℓ que recorta a cauda direita com o peso α da distribuição Qui-Quadrado com o número apropreado de graus de liberdade; em posse de ℓ e $\left(\chi^2\right)_{obs}$, aceitar H caso $(\chi^2)_{obs} < \ell$, e aceitar A caso $(\chi^2)_{obs} \geq \ell$.