ME111 - Laboratório de Estatística

Aula 14 - Teste de Independência

Profa. Larissa Avila Matos

Teste de Independência

- Um dos principais objetivos de se construir uma tabela de contingência, com o objetivo de se analisar a distribuição conjunta de duas variáveis qualitativas, é descrever a associação entre elas.

- Ou seja, de certo modo esperamos que haja uma certa dependência entre as variáveis, por exemplo, sexo e curso de graduação. Desta forma, nosso foco é buscar evidências estatísticas de que duas variáveis possuem certo grau de associação.

- Como vimos no teste de aderência, devemos realizar um teste de hipóteses para investigar a existência de associação contra inexistência de associação.

Exemplo: Cor do cabelo versus Cor dos olhos

- Um estudo em 1899 examinou 6800 homens para verificar se a cor do cabelo e a cor dos olhos estavam relacionadas.

Cor dos olhos	Cor do cabelo				Total
	Castanhos	Pretos	Loiro	Ruivo	
Castanhos	438	288	115	16	857
Verdes	1387	746	946	53	3132
Azuis	807	189	1768	47	2811
Total	2632	1223	2829	116	6800

- Objetivo: Testar se cor dos olhos é independente da cor do cabelo. Ou seja, testar as seguintes hipóteses:

  H_0: cor de olhos e cor de cabelo são independentes

  H_1: cor de olhos e cor de cabelos não são independentes

- Em geral, os dados referem-se a mensurações de duas características (L e C) feitas em n unidades experimentais, que são apresentadas conforme a seguinte tabela:

	C				
L	C_1	C_2	$\cdots$	C_c	Total
L_1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\cdot}$
L_2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
L_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot c}$	$n = n_{\cdot\cdot}$

- Em resumo, o objetivo é testarmos a independência das variáveis L e C:

  H_0: As variáveis L e C são independentes

  H_1: As variáveis L e C não são independentes

- Quantas observações devemos ter em cada casela se L e C forem independentes?

- Quantas observações devemos ter em cada casela se L e C forem independentes?
- Se L e C forem independentes, temos que, para todos os possíveis (L_i e C_j):

$$P(L_i \cap C_j) = P(L_i) \times P(C_j), \qquad \begin{array}{l} i = 1, \ldots, r; \\ j = 1, \ldots, c. \end{array}$$

- Logo, o número esperado de observações com as características (L_i e C_j) entre as n observações sob a hipótese de independência, é dado por

$$E_{ij} = n_{..} \times p_{ij} = n_{..} \times p_{i.} \times p_{.j} = n_{..} \times \frac{n_{i.}}{n_{..}} \times \frac{n_{.j}}{n_{..}},$$

sendo p_{ij} a proporção de observações com as características (L_i e C_j).

- Assim,

$$E_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}}.$$

- O processo deve ser repetido para todas as caselas (ij).

- Para quantificar quão distante os frequências observadas estão das frequências esperadas, usamos a seguinte estatística:

  Estatística do Teste:

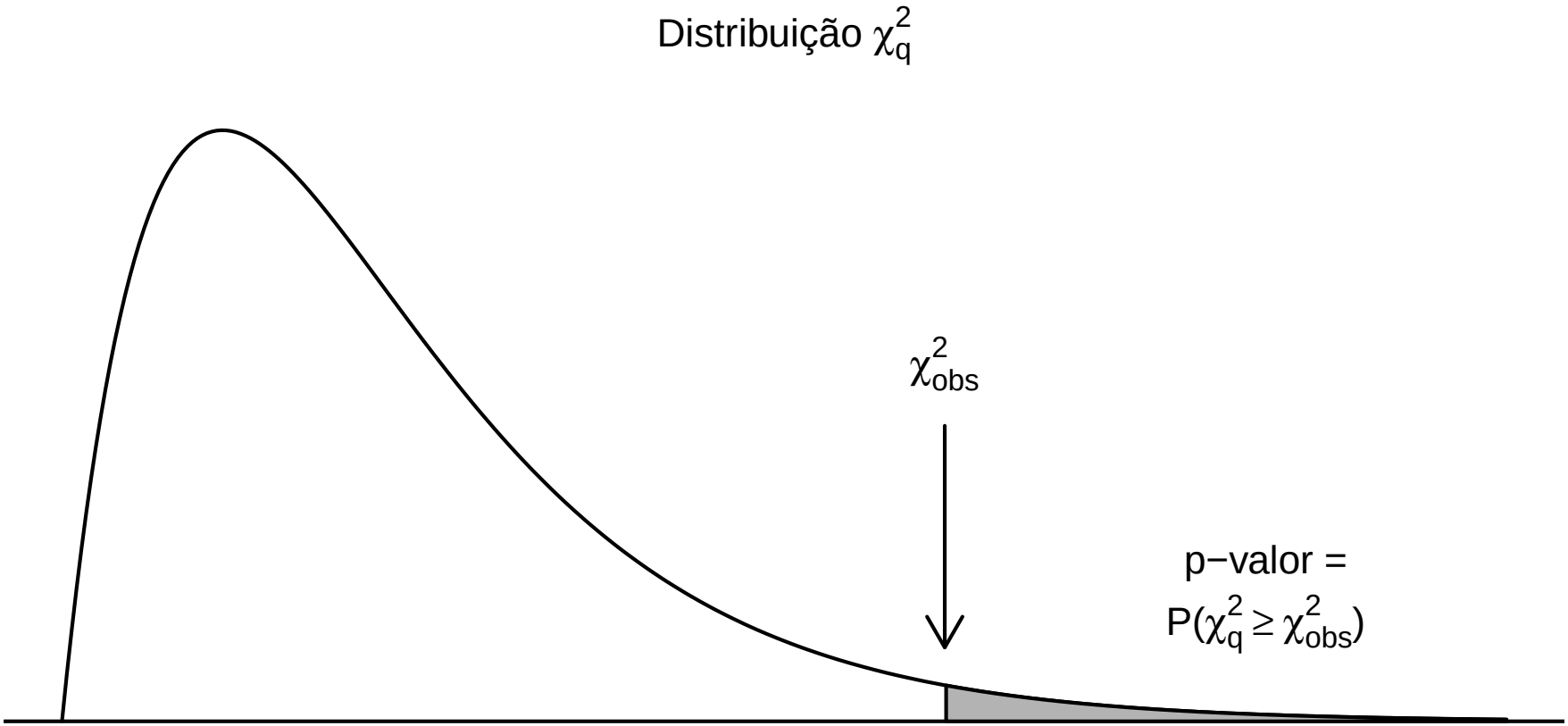
$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

  em que $O_{ij} = n_{ij}$ representa o total de observações na casela (ij).

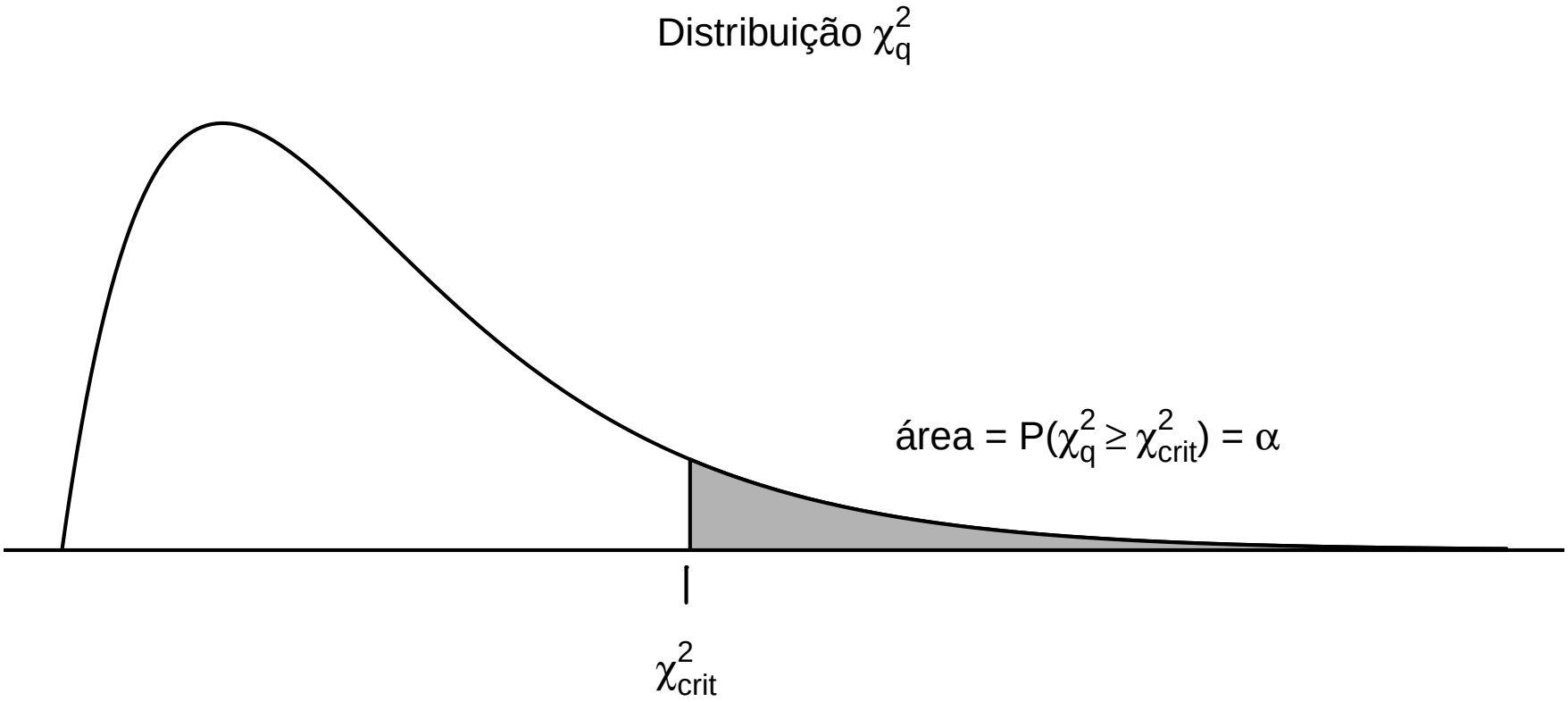
- Se H_0 é verdadeira: $\chi^2 \sim \chi^2_q$, com $q = (r-1) \times (c-1)$.

- Em outras palavras, se H_0 é verdadeira, a v.a. χ^2 segue uma distribuição aproximadamente Qui-quadrado com q graus de liberdade.

- Regra de decisão: Calcular o **p-valor** ou encontrar o **valor crítico**.
- **p-valor**: $P(\chi^2_q \geq \chi^2_{obs})$, em que χ^2_{obs} é o valor da estatística do teste calculada a partir dos dados.



- **Valor Crítico**: Para um nível de significância α, encontrar o valor crítico χ^2_{crit} na tabela Chi-quadrado tal que $P(\chi^2_q \geq \chi^2_{crit}) = \alpha$.



- **Conclusão**: Rejeitamos H_0 se

$$\text{p-valor} \leq \alpha \quad \text{ou} \quad \chi^2_{obs} \geq \chi^2_{crit}$$

Exemplo

```
cores=list("Olhos"=c("Castanhos","Verdes","Azuis"),
           "Cabelo"=c("Castanhos","Pretos","Loiro","Ruivo"))
x<-matrix(c(438,1387,807,288,746,189,115,946,1768,16,53,47),
          nrow = 3, ncol=4, dimnames=cores)
addmargins(x)
```

```
           Cabelo
Olhos       Castanhos Pretos Loiro Ruivo  Sum
  Castanhos       438    288   115    16  857
  Verdes         1387    746   946    53 3132
  Azuis           807    189  1768    47 2811
  Sum            2632   1223  2829   116 6800
```

- Valores esperados

```
n<-6800
n11<-(857*2632)/n
n12<-(857*1223)/n
n13<-(857*2829)/n
n14<-(857*116)/n
n21<-(3132*2632)/n
n22<-(3132*1223)/n
n23<-(3132*2829)/n
n24<-(3132*116)/n
n31<-(2811*2632)/n
n32<-(2811*1223)/n
n33<-(2811*2829)/n
n34<-(2811*116)/n
```

```
esperado<-matrix(c(n11,n21,n31,n12,n22,n32,n13,n23,n33,n14,n24,n34),
                 nrow = 3, ncol=4, dimnames=cores)
addmargins(esperado)
```

```
            Cabelo
Olhos        Castanhos    Pretos     Loiro     Ruivo  Sum
  Castanhos  331.7094  154.1340  356.5372  14.61941  857
  Verdes    1212.2682  563.2994 1303.0041  53.42824 3132
  Azuis     1088.0224  505.5666 1169.4587  47.95235 2811
  Sum       2632.0000 1223.0000 2829.0000 116.00000 6800
```

```r
observado<-x
addmargins(observado)
```

```
            Cabelo
Olhos        Castanhos Pretos Loiro Ruivo  Sum
  Castanhos        438    288   115    16  857
  Verdes          1387    746   946    53 3132
  Azuis            807    189  1768    47 2811
  Sum             2632   1223  2829   116 6800
```

```r
x.1<-((esperado-observado)^2/esperado)
x2<-sum(x.1)
x.1
```

```
           Cabelo
Olhos       Castanhos    Pretos     Loiro      Ruivo
  Castanhos  34.05899 116.26323 163.63011 0.13037624
  Verdes     25.18518  59.25713  97.81392 0.00343237
  Azuis      72.58450 198.22199 306.33978 0.01891411
```

```r
x2
```

```
[1] 1073.508
```

- Graus de liberdade:

```
gl<-(dim(x)[1]-1)*(dim(x)[2]-1)
gl
```

```
[1] 6
```

- Valor critico, considerando $\alpha = 0.05$

```
qchisq(0.95,df=gl)
```

```
[1] 12.59159
```

- Como, $\chi^2_{obs} = 1073.51 > 12.59 = \chi^2_{crit}$, rejeitamos H_0 a um nível de significância de 5%. Ou seja, a cor de olhos e cor de cabelos não são independentes.

- p-valor

```
p.valor=1-pchisq(x2,df=gl)
p.valor
```

```
[1] 0
```

- Como, p-valor$= 0 < 0.05 = \alpha$, rejeitamos H_0 a um nível de significância de 5%. Ou seja, a cor de olhos e cor de cabelos não são independentes.

```r
chisq.test(x)
```

```
    Pearson's Chi-squared test

data:  x
X-squared = 1073.5, df = 6, p-value < 2.2e-16
```

Exercício

- Um inspetor de qualidade toma uma amostra de 220 produtos 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:

	Fábrica			
	1	2	3	Total
Defeituoso	8	15	11	34
Não Defeituoso	62	67	57	186
Total	70	82	68	220

- Ser defeituoso independe da fábrica?