

# MATD49-Estatística não paramétrica

## Teste Exato de Fisher

# Idéia

- Proposto inicialmente por [Fisher, 1966], tem grande valor histórico;
- Usado em tabelas de contingência  $2 \times 2$ , para comparar 2 grupos de acordo com a presença de uma característica;
- É indicado quando o **tamanho das duas amostras independentes é pequeno** e consiste em determinar a **probabilidade exata de ocorrência de uma frequência observada**, ou de valores mais extremos;
- É necessário que as marginais das linhas e colunas sejam fixas (não aleatórias).

# Distribuição Hipergeométrica

- Considere um conjunto de  $n$  elementos dos quais  $r$  são do **Tipo A** e  $n - r$  são do **Tipo B**.
- Para um sorteio de  $c$  elementos ( $c < n$ ), feito ao acaso e **sem reposição**, defina a variável  $X$  como o número de elementos selecionados na amostra que são do **Tipo A**.
- $X$  tem distribuição Hipergeométrica se sua f.d.p. é dada por:

$$P(X = x) = \frac{\binom{r}{x} \binom{n-r}{c-x}}{\binom{n}{c}}, \quad 0 \leq x \leq \min\{c, r\}.$$

# Dados

- Considere  $n$  observações sumarizados em uma tabela de contingência  $2 \times 2$ .
- Os totais nas linhas correspondem a  $r$  e  $n - r$ . e os totais nas colunas correspondem a  $c$  e  $n - c$ . Os totais das linhas e das colunas é não aleatório.

Os dados são organizados na seguinte tabela de contingência  $2 \times 2$ :

	Característica de interesse		Total
	Presença	Ausência	
Grupo A	$x$	$r - x$	$r$
Grupo B	$c - x$	$n - r - c + x$	$n - r$
Total	$c$	$n - c$	$n$

- Cada observação é classificada dentro de uma única célula.
- Os totais das linhas e colunas são fixos, não aleatórios.
- Se os totais marginais são fixos, a probabilidade de observar determinada frequência na tabela  $2 \times 2$  tem distribuição hipergeométrica.

# Hipóteses do Teste

Sejam  $p_A$  e  $p_B$  as probabilidades dos grupos  $A$  e  $B$  apresentarem a característica de interesse, respectivamente (i.e. os percentuais das linhas).

## Bilateral:

$H_0 : p_A = p_B$ , Não existe diferença entre as proporções observadas nos dois grupos; os grupos são independentes; não existe associação entre os dois grupos.

$H_1 : p_A \neq p_B$ , as proporções são diferentes nos dois grupos.

## Unilateral à Direita:

$H_0 : p_A = p_B$

$H_1 : p_A > p_B$ , o grupo A tem maior incidência da característica de interesse que B.

## Unilateral à Esquerda:

$H_0 : p_A = p_B$

$H_1 : p_A < p_B$ , o grupo B tem maior incidência da característica de interesse que A.

# Estatística e regra de decisão I

**Estatística de teste:** A estatística de teste  $T$  é o número de observações na célula da linha 1 e coluna 1:

$$T = x,$$

que tem distribuição hipergeométrica,  $HG(n, r, c)$ , definida por:

$$P(T = x) = \frac{\binom{r}{x} \binom{n-r}{c-x}}{\binom{n}{c}}, 0 \leq x \leq \min\{c, r\}.$$

**Decisão:**

- Encontrar o  $p$  – valor usando a distribuição da estatística  $T$ .
- No caso do teste bilateral, o  $p$  – valor pode ser calculado a partir de diferentes critérios:
  - Uma abordagem usual é somar todas as probabilidades  $p(k) = P(X = k)$  para todos os  $k$  tais que  $p(k) \leq p(x_o)$ , com  $x_o$  o valor observado na tabela, isto é, o  $p$ -valor é a "probabilidade crítica"  $p = P(p(k) \leq p(x_o))$ ;

# Estatística e regra de decisão II

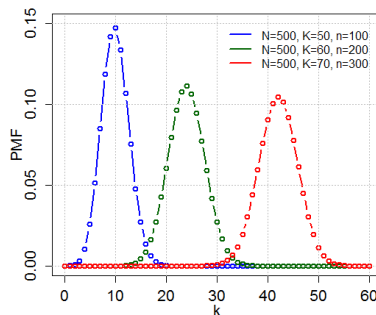


Figura: Distribuição Hipergeométrica. Fonte: wikipedia.

- Pode-se também considerar a soma das  $p(k)$  para tabelas que estão além da tabela observada com  $x_o$ :

$$P = P(|X - E(X)| \geq |x_o - E(X)|),$$

com  $E(X) = c * r/n$ , o que resulta no mesmo teste de qui-quadrado já visto anteriormente para tabelas  $2 \times 2$ ;

# Estatística e regra de decisão III

- Uma terceira abordagem é considerar  $p = 2\min(P(T \leq x), P(T \geq x))$ , mas isto pode, eventualmente, ser maior que 1;
- Outra possibilidade é pegar a menor probabilidade entre  $P(T \leq x)$  e  $P(T \geq x)$  e somar a ela uma probabilidade próxima, mas não maior que ela própria, na cauda do lado oposto.
- Veja [Agresti, 2007] para detalhes.
- Rejeitar  $H_0$  ao nível de significância  $\alpha$ , se o p-valor  $p \leq \alpha$ .



## Nota:

- O teste exato de Fisher só é válido para os dados amostrais, ou seja, não é válido para a população;
- Como este teste calcula uma probabilidade exata, é necessário saber o número de casos nas marginais da tabela  $2 \times 2$  antes dos dados serem analisados;
- É possível generalizar este teste para tabelas  $r \times c$  maiores do que  $2 \times 2$ . Veja [Freeman and Halton, 1951].

## Exemplo I

[Conover, 1996] 14 novos funcionários, 10 homens e 4 mulheres, todos com iguais competências, foram atribuídos entre dois setores: 10 como atendentes e 4 como representantes comerciais. A hipótese nula é de que homens e mulheres foram igualmente distribuídos para o cargo mais desejado de representante comercial contra a alternativa de que mulheres são privilegiadas para esta posição.

Os dados observados foram:

	Rep.Comercial	Atendente	Total
Homens	1	9	10
Mulheres	3	1	4
	4	10	14

**Tabela:** Distribuição segundo sexo e cargo ocupado.

## Exemplo II

As hipóteses são:

$$H_0 : p_H \geq p_M$$

$$H_a : p_H < p_M.$$

As tabelas iguais ou mais críticas do que a Tabela 1 são:

1	9	e	0	10
3	1		4	0

ou seja:

$$\begin{aligned}
 P(\text{rej } H_0 | H_0 \text{ verd}) &= P(X \leq 1) = P(X = 1) + P(X = 0) \\
 &= \frac{\binom{10}{1} \binom{4}{3}}{\binom{14}{4}} + \frac{\binom{10}{0} \binom{4}{4}}{\binom{14}{4}} \approx 0.041 < 0.05
 \end{aligned}$$

Logo, rejeitamos  $H_0$  ao nível 0.05.

# Aspectos computacionais

No R:

```
fisher.test()
```

No SAS, utilizaremos novamente o proc freq:

```
proc freq data=FatComp order=data;  
tables Exposure*Response / fisher;  
weight Cont;  
run;
```

## Exemplo no SAS I

Neste exemplo temos a incidência de doenças cardíacas em 23 pacientes de acordo com seu tipo de dieta.

Colesterol na dieta	Doença cardíaca		Total
	Sim	Não	
Alto	11	4	15
Baixo	2	6	8
Total	13	10	23

Desejamos avaliar se o percentual de doenças cardíacas é diferente entre as dietas.

## Exemplo no SAS II

No SAS, Utilizaremos o comando:

```
proc freq data=FatComp order=data;  
tables Exposure*Response / fisher;  
weight Cont;  
run;
```

# Teste de Mantel Haenszel

Este teste é indicado quando temos diversas tabelas  $2 \times 2$  sequenciais:

Grupos na tabela $i$	Característica		Total
	Presença	Não	
Grupo A	$x_i$	$r_i - x_i$	$r_i$
Grupo B	$c_i - x_i$	$n_i - r_i - c_i + x_i$	$n_i - r_i$
Total	$c_i$	$n_i - c_i$	$n_i$

respeitando as as mesmas suposições do teste exato de Fisher.

Além disso, as tabelas foram obtidas de a partir de amostras(experimentos) independentes.

# Hipóteses do Teste

Sejam  $p_{Ai}$  e  $p_{Bi}$  as probabilidades dos grupos  $A$  e  $B$  apresentarem a característica de interesse na  $i$ -ésima tabela de contingência, respectivamente (i.e. os percentuais das linhas).

## Bilateral:

$H_0$  :  $p_{Ai} = p_{Bi}$ , para todo  $i = 1, \dots, k$ . Não existe diferença entre as proporções observadas nos dois grupos; os grupos são independentes; não existe associação entre os dois grupos, para todo  $i$ .

$H_1$  : Ou  $p_{Ai} > p_{Bi}$  para algum(ns)  $i$  ou  $p_{Ai} < p_{Bi}$  para algum(ns)  $i$ .

## Unilateral à Direita:

$H_0$  :  $p_{Ai} = p_{Bi}$ , para todo  $i = 1, \dots, k$ .

$H_1$  :  $p_{Ai} \geq p_{Bi}$ , para todo  $i$  e  $p_{Ai} > p_{Bi}$ , para algum(ns)  $i$ .

## Unilateral à Esquerda:

$H_0$  :  $p_{Ai} = p_{Bi}$ , para todo  $i = 1, \dots, k$ .

$H_1$  :  $p_{Ai} \leq p_{Bi}$ , para todo  $i$  e  $p_{Ai} < p_{Bi}$ , para algum(ns)  $i$ .



# Estatística e decisão

A estatística do teste é:

$$T_{mh} = \frac{\sum_i x_i - \sum_i \frac{r_i c_i}{n_i}}{\sqrt{\sum_i \frac{r_i c_i (n_i - r_i)(n_i - c_i)}{n_i^2 (n_i - 1)}}, \quad (1)$$

que possui, sob  $H_0$ , distribuição aproximadamente normal padrão.

Rejeitamos  $H_0$  ao nível  $\alpha$  quando o valor da estatística  $T_{mh}$  for maior do que  $z_{1-\alpha}$ , isto é, para valores absolutos grandes da estatística.

# Exemplo

[Conover, 1996], adaptado.

Três grupos de pacientes de diferentes hospitais são tratados com um mesmo tratamento ou são alocados no grupo controle. Deseja-se saber se houve melhora nas as taxas de recuperação dos pacientes tratados.

	Grupo 1		Grupo 2		Grupo 3	
	Sucesso	Fracasso	Sucesso	Fracasso	Sucesso	Fracasso
Tratamento	10	1	9	0	8	0
Controle	12	1	11	1	7	3
Total	22	2	20	1	15	3

A estatística é


$$T_{mh} = \frac{(10 + 9 + 8) - \left( \frac{11 \cdot 22}{24} + \frac{9 \cdot 20}{21} + \frac{8 \cdot 15}{18} \right)}{\sqrt{\frac{11 \cdot 22 \cdot 13 \cdot 2}{24^2 \cdot 23} + \frac{9 \cdot 20 \cdot 12 \cdot 1}{21^2 \cdot 20} + \frac{8 \cdot 15 \cdot 10 \cdot 3}{18^2 \cdot 17}}} = \frac{1.6786}{1.1719} = 1.4323 < 1.65.$$


Logo, não rejeitamos  $H_0$  ao nível 0.05.


# Acknowledgements


Agradecemos ao prof. Anderson Ara pela disponibilização de seu material didático, no qual nos baseamos para a elaboração destes slides. Alguns trechos desta apresentação são replicados de seu material.

# Referências I

 Agresti, A. (2007).  
*An introduction to categorical data analysis.*  
Wiley-Interscience.

 Conover, W. J. (1996).  
*Practical nonparametric statistics.*  
John Wiley and sons, 3 ed. edition.

 Fisher, R. A. (1966).  
*The Design of Experiments.*  
Edinburgh: Oliver and Boyd., 8th edition.

 Freeman, G. H. and Halton, J. H. (1951).  
Note on an exact treatment of contingency, goodness of fit and other problems of  
significance.  
*Biometrika*, 38:141–149.