

Técnicas Computacionais em Probabilidade e Estatística I

Aula V

Chang Chiann

MAE 5704- IME/USP

1º Sem/2008

Análise de 2 conjunto de dados

Modelo Probabilístico:

2 conjuntos de dados \rightarrow amostras de duas v.a.s distintas.

Ex.: idade e peso

3 casos:

- 2 variáveis quantitativas;
- 2 variáveis qualitativas;
- 1 qualitativa e outra quantitativa.

Analisar dois conjuntos de dados por meio de:

- a) Métodos numéricos, ou seja, calcular medidas de posição e dispersão para cada conjunto de dados separadamente e, depois, medidas de associação entre os dois conjuntos;
- b) Métodos gráficos, a saber, aqueles já vistos para cada conjunto e, depois, gráficos para analisar as relações entre eles, como gráficos de dispersão e gráficos Q-Q(quantis-quantis).

Duas Variáveis Qualitativas

Os dados podem ser resumidos construindo-se uma tabela de distribuição de frequências, que quantifica a frequência das distintas categorias.

Variáveis qualitativas no arquivo ***PULSE***

Ran

Smokes

Sex

Activity

Variáveis qualitativas no arquivo *PULSE*

```
MTB > Tally 'Sex' 'Smokes' 'Activity';  
SUBC> Counts;  
SUBC> Percents.
```

Summary Statistics for Discrete Variables

Sex	Count	Percent	Smokes	Count	Percent
1	57	61,96	1	28	30,43
2	35	38,04	2	64	69,57
N=	92		N=	92	

Activity	Count	Percent
0	1	1,09
1	9	9,78
2	61	66,30
3	21	22,83
N=	92	

Podemos também construir tabelas de freqüências conjuntas (*tabelas de contingência*), relacionando duas variáveis qualitativas.

Exemplo 1: Há indícios de associação entre Sexo e Hábito de fumar?

	Hábito de Fumar		
Sexo	Fuma	Não Fuma	Total
Masculino	20	37	57
Feminino	8	27	35
Total	28	64	92

Qual o significado dos valores desta tabela?

Como concluir?

Verificar associação através da:

- porcentagem segundo as colunas, ou
- porcentagem segundo as linhas.

Sexo	Hábito de Fumar		Total
	Fuma	Não Fuma	
Masculino	71,43%	57,81%	61,96%
Feminino	28,57%	42,19%	38,04%
Total	100%	100%	100%

Qual o significado dos valores desta tabela?

Como concluir?

Variáveis Qualitativas

Gráficos

- Gráfico de setores
- Gráfico de barras

Gráfico de setores

Um círculo é dividido em tantos setores quantas forem as categorias da variável. A área de cada setor é proporcional à frequência da categoria

Arquivo *PULSE*— Gráfico de setores para a variável *Activity*

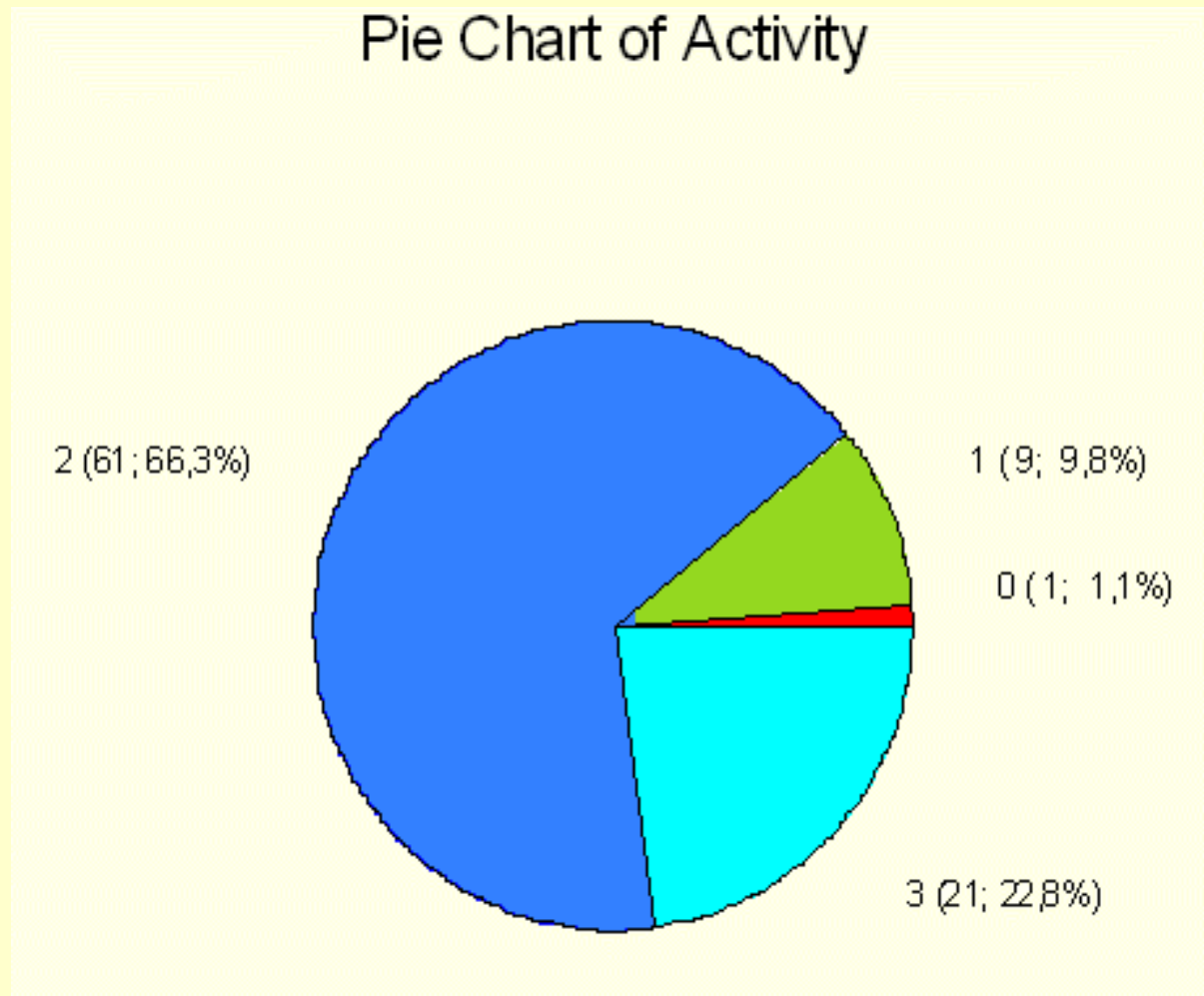
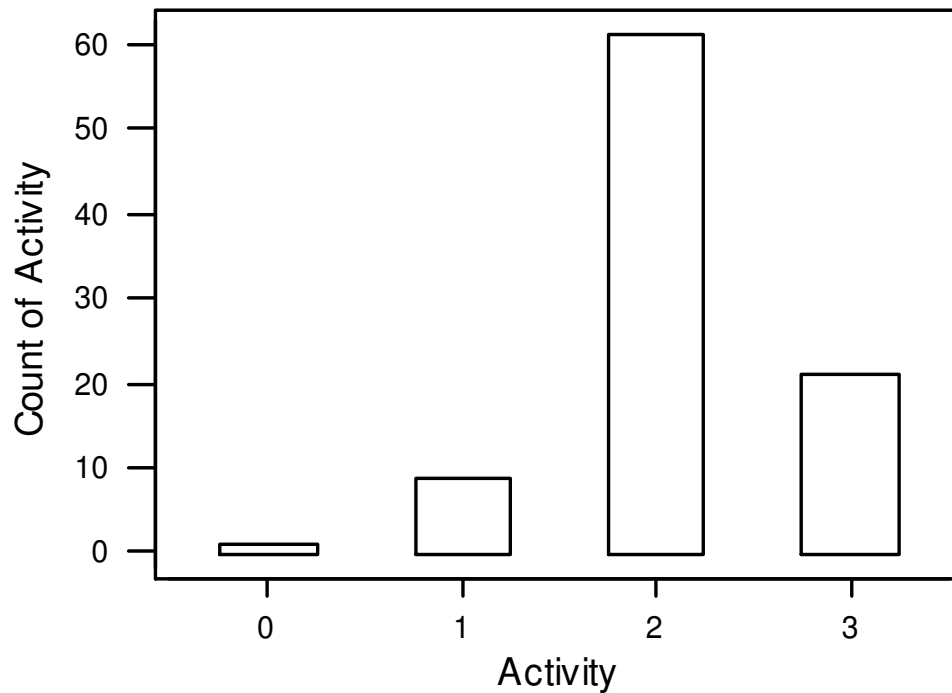


Gráfico de barras

Sobre um eixo, são representados retângulos, um para cada categoria da variável. A altura do retângulo é proporcional à frequência da categoria

Arquivo *PULSE*— Gráfico de barras para a variável *Activity*

MTB > Chart C8



Testes de Independência

Objetivo: Verificar se existe independência entre duas variáveis medidas nas mesmas unidades experimentais.

Exemplo: Deseja-se verificar se existe dependência entre a renda e o número de filhos em famílias de uma cidade.

- 250 famílias escolhidas ao acaso forneceram a tabela a seguir:

Renda (R\$)	Número de filhos				Total
	0	1	2	+ de 2	
menos de 2000	15	27	50	43	135
2000 a 5000	25	30	12	8	75
5000 ou mais	8	13	9	10	40
Total	48	70	71	61	250

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

$A \backslash B$	B_1	B_2	...	B_s	Total
A_1	n_{11}	n_{12}	...	n_{1s}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	...	n_{2s}	$n_{2\bullet}$
...
A_r	n_{r1}	n_{r2}	...	n_{rs}	$n_{r\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$...	$n_{\bullet s}$	n

Hipóteses a serem testadas – **Teste de independência:**

H : A e B são variáveis independentes

A : As variáveis A e B não são independentes

→ Quantas observações devemos ter em cada casela, se A e B forem independentes?

Se A e B forem independentes, temos que, para todos os possíveis pares $(A_i$ e $B_j)$:

$$P(A_i \cap B_j) = p_{ij} = P(A_i) \times P(B_j), \text{ para } i = 1, 2, \dots, r \text{ e } j = 1, 2, \dots, s.$$

Logo, o **número esperado de observações com as características $(A_i$ e $B_j)$** , 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 $(A_i$ e $B_j)$.

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

O processo deve ser repetido para todas as caselas (i, j) .

Distância entre os valores observados e os valores esperados sob a suposição de independência:

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

Estatística do teste de independência

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

Supondo H verdadeira,

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

sendo $q = (r - 1) \times (s - 1)$ graus de liberdade.

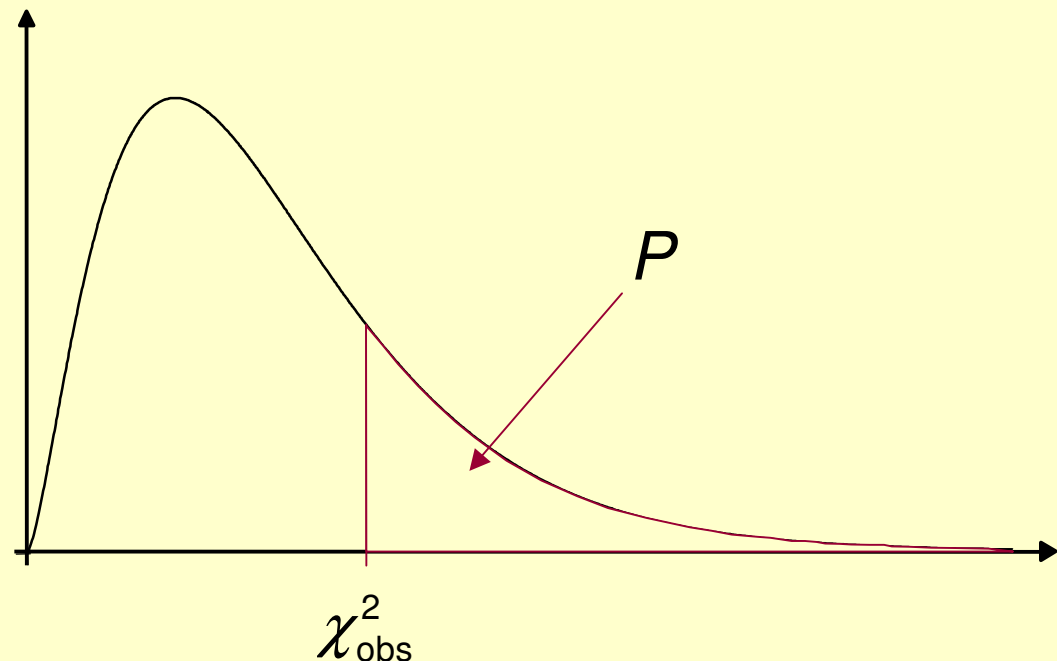
Regra de decisão:

Pode ser baseada no nível descritivo P , neste caso

$$P = P(\chi^2_q \geq \chi^2_{\text{obs}}),$$

em que χ^2_{obs} é o valor calculado, a partir dos dados, usando a expressão apresentada para χ^2 .

Graficamente:



Se, para α fixado, obtemos $P \leq \alpha$, rejeitamos a hipótese H de independência.

Exemplo (continuação):

Estudo da dependência entre renda e o número de filhos

- 250 famílias foram escolhidas ao acaso

Hipóteses H : O número de filhos e a renda são independentes

A : Existe dependência entre o número de filhos e a renda

Renda (R\$)	Número de filhos				Total
	0	1	2	+ de 2	
menos de 2000	15	27	50	43	135
2000 a 5000	25	30	12	8	75
5000 ou mais	8	13	9	10	40
Total	48	70	71	61	250

Exemplo do cálculo dos valores esperados sob H (independência):

- Número esperado de famílias sem filhos e renda menor que R\$ 2000:

$$E_{11} = \frac{48 \times 135}{250} = 25,92 .$$

Tabela de valores observados e esperados (entre parênteses)

Renda (R\$)	Número de filhos				Total
	0	1	2	+ de 2	
menos de 2000	15(25,92)	27(37,80)	50(38,34)	43(32,94)	135
2000 a 5000	25(14,40)	30(21,00)	12(21,30)	8(18,30)	75
5000 ou mais	8(7,68)	13(11,20)	9(11,36)	10(9,76)	40
Total	48	70	71	61	250

1 filho e renda de R\$ 2000
a R\$ 5000:

$$E_{22} = \frac{70 \times 75}{250} = 21,00$$

2 ou + filhos e renda de R\$ 5000 ou
mais:

$$E_{34} = \frac{61 \times 40}{250} = 9,76$$

Lembre-se:

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

Cálculo da estatística de qui-quadrado:

Renda (R\$)	Número de filhos				Total
	0	1	2	+ de 2	
menos de 2000	15(25,92)	27(37,80)	50(38,34)	43(32,94)	135
2000 a 5000	25(14,40)	30(21,00)	12(21,30)	8(18,30)	75
5000 ou mais	8(7,68)	13(11,20)	9(11,36)	10(9,76)	40
Total	48	70	71	61	250

$$\begin{aligned}
 \chi_{obs}^2 = & \frac{(15 - 25,92)^2}{25,92} + \frac{(25 - 14,40)^2}{14,40} + \frac{(8 - 7,68)^2}{7,68} + \frac{(27 - 37,80)^2}{37,80} + \\
 & + \frac{(30 - 21,00)^2}{21,00} + \frac{(13 - 11,20)^2}{11,20} + \frac{(50 - 38,34)^2}{38,34} + \frac{(12 - 21,30)^2}{21,30} + \\
 & + \frac{(12 - 21,30)^2}{21,30} + \frac{(9 - 11,36)^2}{11,36} + \frac{(43 - 32,94)^2}{32,94} + \frac{(8 - 18,30)^2}{18,30} + \\
 & + \frac{(10 - 9,76)^2}{9,76} = 36,62 .
 \end{aligned}$$

Determinação do número de graus de liberdade:

- Categorias de renda: $r = 3$
 - Categorias de nº de filhos: $s = 4$
- $\left. \begin{array}{l} \bullet \text{ Categorias de renda: } r = 3 \\ \bullet \text{ Categorias de nº de filhos: } s = 4 \end{array} \right\} \longrightarrow q = (r - 1) \times (s - 1) = 2 \times 3 = 6$

Logo, $\chi^2 \sim \chi_6^2$ e, supondo $\alpha = 0,05$, $P = P(\chi_6^2 \geq 36,62) = 0,000$

\therefore Como $P = 0,000 < \alpha = 0,05$, rejeitamos a independência entre número de filhos e renda familiar.

Os cálculos podem ser feitos diretamente no MINITAB:

Stat → Tables → Chi-Square test

Uma medida da relação entre duas variáveis qualitativas é o coeficiente de contingência de Pearson, dado por:

$$c = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Uma modificação de **C** é o coeficiente

$$T = \sqrt{\frac{\chi^2 / n}{(r-1)(s-1)}}$$

Que atinge o valor máximo(um) quando $r=s$.