Variáveis Multidimensionais

- Queremos estudar o comportamento conjunto de duas variáveis
  - Grau de Instrução: X
  - Região de Procedência: Y

	Ensino Fundamental	Ensino Médio	Ensino Superior	Total
Capital	4	5	2	11
Interior	3	7	2	12
Outra	5	6	2	13
Total	12	18	6	36

Variáveis Multidimensionais

- 4 indivíduos procedem da capital e possuem ensino fundamental
- Na última coluna, está representada a frequência absoluta da variável Y
- Na última linha está representada a frequência absoluta da variável X
- As frequências absolutas (parte interna da tabela) são chamadas de frequências absolutas conjuntas entre X e Y

Variáveis Multidimensionais

- Frequências Relativas (Proporções)
  - Em relação ao total de elementos (36)
  - em relação ao total de cada linha
  - em relação ao total de cada coluna
- A frequência relativa a ser utilizada depende do estudo que pretendemos fazer.

Variáveis Multidimensionais

- Distribuição das frequências relativas ao total (36)

	Ensino Fundamental	Ensino Médio	Ensino Superior	Total
Capital	0.11 (11%)	0.14 (14%)	0.06 (6%)	0.31 (31%)
Interior	0.08 (8%)	0.19 (19%)	0.06 (6%)	0.33 (33%)
Outra	0.14 (14%)	0.17 (17%)	0.05 (5%)	0.36 (36%)
Total	0.33 (33%)	0.50 (50%)	0.17 (17%)	1.00 (100%)

Variáveis Multidimensionais

- Distribuição das frequências relativas ao total por coluna

	Ensino Fundamental	Ensino Médio	Ensino Superior	Total
Capital	0.33	0.28	0.33	0.31
Interior	0.25	0.39	0.33	0.33
Outra	0.42	0.33	0.34	0.36
Total	1.00	1.00	1.00	1.00

Variáveis Multidimensionais

- Entre os empregados com ensino médio
  - 28% vêm da capital
  - 39% vêm do interior
  - 33% vêm de outros locais
- Permite comparar a distribuição de Y (procedência) conforme o grau de instrução: o grau de instrução está associado ao local de procedência ? ou equivalentemente no ensino médio (por exemplo) as proporções supracitadas são "iguais" ?

Independência de Variáveis

- A distribuição conjunta descreve a associação existente entre as variáveis
- Grau de dependência: como uma variável "explica" ou se se "associa" a outra

Independência de Variáveis

- Desejamos verificar se existe dependência entre o sexo (X) e a carreira escolhida (Y) por 200 alunos de Economia e Administração

	Masculino	Feminino	Total
Economia	85	35	120
Administração	55	25	80
Total	140	60	200

Independência de Variáveis

- Para fazer um estudo de dependência, serão utilizadas as frequências relativas ao total por coluna (observando que o número de estudantes do sexo masculino é diferente do número de estudantes do sexo feminino)

	Masculino	Feminino	Total
Economia	0.61	0.58	0.60
Administração	0.39	0.42	0.40
Total	1.00	1.00	1.00

Independência de Variáveis

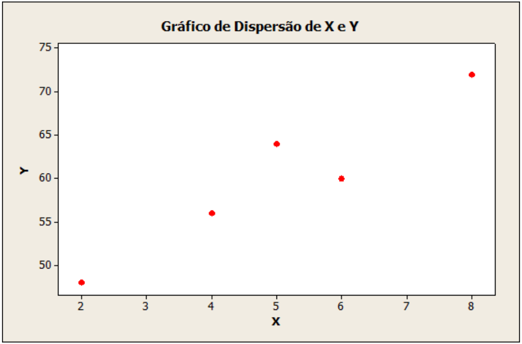
- Sem levar em conta os sexos (última coluna), 60% dos estudantes preferem economia e 40% administração;
- Se não houver dependência entre sexo e carreira escolhida, espera-se que quando observado para cada sexo, a escolha das carreiras tenha essas mesmas proporções;
- Sexo masculino: 61% dos estudantes na carreira de economia e 39% na de administração;
- Sexo Feminino: 58% dos estudantes na carreira de economia e 42% na de administração;
- Os dados indicam que não há dependência entre as variáveis, más ... isso precisa ser quantificado.

Diagramas de Dispersão

- Exemplo

Agente	Anos de Serviço (X)	N^o de Clientes (Y)
A	2	48
B	4	56
C	5	64
D	6	60
E	8	72
Total	25	300

Diagramas de Dispersão



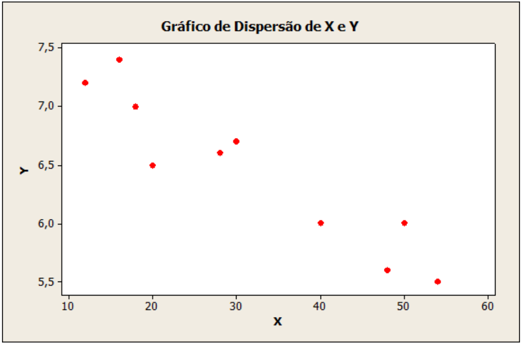
As variáveis anos de serviço e número de clientes tem uma grande dependência entre si.

Diagramas de Dispersão

- Exemplo
  - Renda Mensal Bruta (X)
  - % da Renda gasta com Assistência Médica (Y)

Família	X	Y
A	12	7.2
B	16	7.4
C	18	7.0
D	20	6.5
E	28	6.6
F	30	6.7
G	40	6.0
H	48	5.6
I	50	6.0
J	54	5.5

Diagramas de Dispersão



Nesse caso, a dependência entre X e Y é negativa.

Coeficiente de Correlação

- Objetivo: obter uma medida que permita quantificar a dependência que pode existir entre duas variáveis (positiva, negativa, muita ou pouca)
- Dado n pares de observações $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$:
  $Corr(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{DP(X)} \right) \left(\frac{y_i - \bar{y}}{DP(Y)} \right)$
- Essa medida leva em consideração todos os desvios $(x_i - \bar{x})$ e $(y_i - \bar{y})$ padronizados da forma $\left(\frac{x_i - \bar{x}}{DP(X)} \right)$ e $\left(\frac{y_i - \bar{y}}{DP(Y)} \right)$

Coeficiente de Correlação

- Propriedades:
  - $-1 \leq Corr(X, Y) \leq 1$
  - $Corr(X, Y)$ estiver próxima de 1: X e Y estão positivamente associados e existe associação linear entre as variáveis.
  - $Corr(X, Y)$ estiver próxima de -1: X e Y estão negativamente associados e existe associação linear entre as variáveis.

Coeficiente de Correlação

- Retomando o primeiro exemplo:
  - $\bar{x} = 5$
  - $DP(X) = 2$
  - $\bar{y} = 60$
  - $DP(Y) = 8$

Coeficiente de Correlação

Agente	X	Y	$x_i - \bar{x}$	$y_i - \bar{y}$	$\left(\frac{x_i-\bar{x}}{DP(X)}\right)$	$\left(\frac{y_i-\bar{y}}{DP(Y)}\right)$	$\left(\frac{x_i-\bar{x}}{DP(X)}\right)\left(\frac{y_i-\bar{y}}{DP(Y)}\right)$
A	2	48	-3	-12	-1.5	-1.5	2.25
B	4	56	-1	-4	-0.5	-0.5	0.25
C	5	64	0	4	0	0.5	0
D	6	60	1	0	0.5	0	0
E	8	72	3	12	1.5	1.5	2.25

Portanto: $Corr(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{DP(X)}\right)\left(\frac{y_i-\bar{y}}{DP(Y)}\right) = \frac{4.75}{5} = 0.95$