Estatística Descritiva

Prof. Lorí Viali, Dr.
viali@mat.ufrgs.br
http://www.mat.ufrgs.br/~viali/

4/5

Relacionamento entre Variáveis

Variáveis Qualitativas

Distribuição Conjunta

Suponha que se queira analisar o comportamento conjunto das variáveis **X = Grau de Instrução** e **Y = Região de procedência**. Neste caso, a distribuição de freqüências é apresentada como uma tabela de dupla entrada, que esta apresentada na tabela seguinte:

Exemplo (tabela um)

X / **Y**	Primeiro Grau	Segundo Grau	Superior	**Total**
Capital	4	5	6	**15**
Interior	11	4	3	**18**
Outra	2	3	2	**7**
Total	17	12	11	**40**

Cada elemento da tabela fornece a freqüência observada da realização simultânea das variáveis X e Y. Neste caso, foram observados 4 moradores da capital com primeiro grau, 6 com instrução superior, 7 moradores do interior com instrução do segundo grau e assim por diante.

A linha dos totais fornece a distribuição da variável X (grau de instrução) enquanto que o total das colunas fornece a distribuição da variável Y (região de procedência).

As distribuições separadas (das margens) são chamadas de **distribuições marginais** enquanto que a tabela um forma a **distribuição conjunta** das variáveis X e Y.

Ao invés de se trabalhar com as **freqüências absolutas**, pode-se obter as **freqüências relativas** (proporções), como foi feito no caso de uma única variável.

Mas agora existem três possibilidades de expressarmos a proporção de cada célula da tabela:

(1) em relação ao total geral;

(2) em relação ao total de cada linha;

(3) em relação ao total de cada coluna.

A tabela 2 apresenta a distribuição conjunta das freqüências relativas expressas como proporções do total geral.

Exemplo (tabela dois)

X / **Y**	Primeiro Grau	Segundo Grau	Superior	**Total**
Capital	10,0	12,5	15,0	**37,5**
Interior	27,5	10,0	7,5	**45,0**
Outra	5,0	7,5	5,0	**17,5**
Total	42,5	30,0	27,5	**100**

Neste caso pode-se afirmar que **10%** dos empregados vem da capital e tem instrução de primeiro grau. Os totais das margens fornecem **as distribuições** (em percentual) de cada uma das variáveis, consideradas individualmente. Assim **37,5%** dos pais vem da capital e, por exemplo, **30%** possuem segundo grau.

A tabela três apresenta a distribuição das proporções (em percentual) em relação ao total das colunas.

Exemplo (tabela três)

Y \ X	Primeiro Grau	Segundo Grau	Superior	**Total**
Capital	23,53	41,67	54,55	**37,5**
Interior	64,71	33,33	27,27	**45,0**
Outra	11,76	25,00	18,18	**17,5**
Total	**100,0**	**100,0**	**100,0**	**100**

Assim, pode-se ver que **25,53%** dos pais com instrução de primeiro grau vem da capital, **64,71%** vem do interior etc. Quantos aos pais com grau superior **54,55%** vem da capital, **27,27%** do interior etc. Esta distribuição serve para comparar a distribuição da procedência das pessoas conforme o grau de instrução.

De forma análoga, pode-se construir a distribuição das proporções em relação ao total de linhas.

Independência de variáveis

Um dos principais objetivos de se determinar a distribuição conjunta é descrever a associação existente entre as variáveis, isto é, quer-se conhecer o **grau de dependência existente entre elas**, de modo que se possa prever melhor o resultado de uma delas quando se conhece o resultado da outra.

Exemplo

Se fosse desejado estimar qual a renda média de uma família de Porto Alegre, a informação adicional sobre a classe social que essa família pertence permitirá que a estimativa seja mais precisa, pois se sabe que existe dependência entre os dois tipos de variáveis.

Quer-se identificar se existe ou não dependência entre sexo e curso escolhido, baseado em uma amostra de 200 alunos de Economia e Administração. Estes dados estão agrupados na tabela 4.

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística

Exemplo (tabela quatro)

Y \ X	Masculino	Feminino	**Total**
Economia	85	35	**120**
Administração	55	25	**80**
Total	**140**	**60**	**200**

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística

De início pode-se perceber que não é fácil tirar alguma conclusão, devido a diferença nos totais marginais. Desta forma, deve-se construir proporções segundo as linhas (ou colunas) para se poder fazer comparações.

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística

Exemplo (tabela cinco)

Y \ X	Masculino	Feminino	**Total**
Economia	61	58	**60**
Administração	39	42	**40**
Total	**100**	**100**	**100**

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística

Desta tabela pode-se observar que, independentemente de sexo, 60% dos alunos preferem Economia e 40% Administração (Pode-se ver pela coluna do total)

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística

Não havendo dependência entre as variáveis, seria esperado **as mesmas proporções para cada sexo**. Observando a tabela, pode-se ver que as proporções são bem próximos do que seria esperado, isto é, do sexo masculino **61%** preferem Economia e **39%** Administração, enquanto que do sexo feminino estas proporções são **58%** e **42%** respectivamente.

Estes resultados parecem indicar que não existe dependência entre as variáveis sexo e curso escolhido (pelo menos para estes dois cursos).

Exemplo

Suponha agora um mesmo tipo de exemplo, só que envolvendo alunos dos cursos de Física e Serviço Social, cuja distribuição conjunta está na tabela 6.

Exemplo (tabela seis)

Y \ X	Masculino	Feminino	**Total**
Física	100	20	**120**
Serviço Social	40	40	**80**
Total	**140**	**60**	**200**

Exemplo (tabela sete)

Y \ X	Masculino	Feminino	**Total**
Física	71	33	**60**
Serviço Social	29	67	**40**
Total	**100**	**100**	**100**

Comparando agora a distribuição das proporções pelos cursos, parece haver uma maior concentração de homens no curso de Física e de mulheres no de Serviço Social. Portanto, neste caso, as variáveis sexo e curso escolhido parecem ser dependentes.

Observe-se que se teria chegado as mesmas conclusões se tivesse sido utilizado o total de linhas ao invés do total de colunas. Quando existe dependência entre variáveis, sempre é conveniente **quantificar esta dependência.**

Dependência entre variáveis nominais

De um modo geral, a quantificação do grau de dependência entre duas variáveis é realizada pelos chamados **coeficientes de correlação ou associação**. Estas medidas descrevem através de um único número a dependência entre duas variáveis.

Para que a interpretação se torne mais fácil e intuitiva estes coeficientes normalmente variam de **zero** a **um** (ou de –1 a +1), e a proximidade de zero indica que as variáveis **são independentes**.

Existem várias formas de medir a dependência entre duas variáveis nominais. Uma delas é o denominado **coeficiente de contingência**, devido a Karl Pearson.

A análise da tabela sete mostrou que existe dependência entre as variáveis. Se houvesse independência o número esperado de estudantes masculinos de Física seria: (140.120)/200 = 84. Calculando os demais valores esperados poderíamos formar a tabela oito.

Exemplo (tabela oito)

Y \ **X**	Masculino	Feminino	**Total**
Física	84	36	**120**
Serviço Social	56	24	**80**
Total	**140**	**60**	**200**

Pode-se comparar as duas tabelas, isto é, os valores esperados com os observados, determinando-se os desvios existentes entre eles. Os resultados estão tabela nove.

Exemplo (tabela nove)

Y \ X	Masculino	Feminino
Física	16	-16
Serviço Social	-16	16

Uma vez obtidos os desvios de cada célula da tabela, pode-se obter os desvios relativos de cada célula. Para isto eleva-se cada resultado ao quadrado (para eliminar os valores negativos) e divide-se o resultado pelo valor esperado, isto é:

$(O_i - E_i)^2 / E_i$

Juntando os resultados de cada célula, tem-se uma medida do grau de afastamento, isto é, de dependência entre as duas variáveis. Esta medida é representada por χ^2 e lida **qui-quadrado**.

. Esta medida é representada por χ^2 e lida **qui-quadrado**. Para este exemplo, o valor desta medida seria:

χ^2 = 3,0476 + 7,1111 + 4,5714 + 10,6667 = 25,40.

No entanto, julgar a associação pelo expressão acima não é fácil, porque não se tem um padrão de comparação, para saber se este valor é alto ou não.

Por isto, utiliza-se uma outra medida, devida a Karl Pearson, e denominada de **Coeficiente de Contingência C**, definida por:

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

Onde n é o número de observações (tamanho da amostra).

Para o exemplo acima o coeficiente de Pearson será:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{25{,}3968}{25{,}3968 + 200}} = 0{,}34$$

Teoricamente este coeficiente é um número entre **zero** e **um**, sendo zero quando as variáveis forem independentes (não estiverem associadas). No entanto, mesmo quando existe uma associação perfeita entre as variáveis este coeficiente pode não ser igual a 1. Uma alteração possível é considerar o coeficiente:

$$C^* = \frac{C}{\sqrt{(t-1)/t}}$$

onde t é o mínimo entre o número de linhas e colunas.

Exercício

Determine o grau de associação entre o número de horas de estudo semanal e área de estudo, utilizando um levantamento feito entre 500 estudantes da Universidade Pindorama.

Exercício

Área / Tempo	Eng.	Hum.	**Sociais**	**Saúde**
Até 5	100	20	15	15
6 a 10	50	13	15	42
11 a 15	30	9	10	51
+ de 15	20	8	10	92