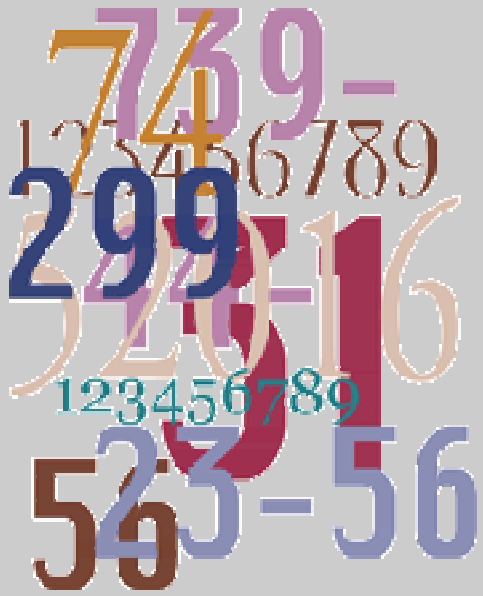
Material Didático

Série

Estatística Básica



Texto:

Análise Bidimensional

Prof. Lori Viali, Dr.

SUMÁRIO

ANÁLISE BIDIMENSIONAL

1. VARIÁVEIS BIDIMENSIONAIS QUALITATIVAS

1.1. INTRODUÇÃO

Até agora foi visto como se pode organizar, descrever e resumir informações representadas por uma única variável, mas este é apenas uma das situações possíveis. Pode-se ter 2, 3, ou mais variáveis. Neste caso a distribuição de freqüências conjunta das variáveis vai representar um papel importante na análise. Este estudo vai se deter basicamente nas variáveis bidimensionais, mas a extensão para mais de duas variáveis é imediata.

Exemplo 1.1

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 1.1 abaixo.

Tabela 1.1 - Distribuição conjunta das variáveis X e Y.

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 do corpo 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 1.1 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 aqui existem 3 possibilidades de

expressarmos a proporção de cada célula da tabela: (1) em relação ao total geral, (2) em relação ao total de cada linha e (3) em relação ao total de cada coluna.

A tabela 1.2 apresenta a distribuição conjunta das freqüências relativas expressas como proporções do total geral. 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, 45% são procedentes do interior e os restantes de outros estados. Da mesma forma pode-se constatar que 42,50% os pais tem primeiro grau, 30% o segundo grau e os restantes possuem formação superior.

Tabela 1.2 - Distribuição conjunta das variáveis X e Y.

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	7,50	5,0	**17,5**
Total	**42,50**	**30,0**	**27,5**	**100,0**

A tabela 1.3 apresenta a distribuição das proporções (em percentual) em relação ao total das colunas. Assim, pode-se afirmar que 25,53% dos pais com instrução de primeiro grau vem da capital, 64,71% vem do interior e 11,76% vem de fora do estado. Quantos aos pais com grau superior 54,55% vem da capital, 27,27% o interior e 18,18% de fora do estado. Este tipo de distribuição serve para comparar a distribuição da procedência das pessoas conforme o grau de instrução. De forma análogo, pode-se construir a distribuição das proporções em relação ao total de linhas.

Tabela 1.3 - Distribuição conjunta das variáveis X e Y.

X / Y	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**	**100,0**

1.2. 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.

Por exemplo, se for desejado estimar qual a renda média de uma família moradora de Porto Alegre, a informação adicional sobre qual a classe social que ela pertence permite que a estimativa seja mais precisa, pois se sabe que existe dependência entre os dois tipos de variáveis. Ou ainda, suponha que se queira advinhar o sexo de um estudante da cidade de PUC sorteado ao acaso. Como se sabe que aproximadamente metade dos estudantes da universidade são homens, não teríamos preferência em sugerir um ou outro sexo. No entanto, se for informado que este aluno estuda Pedagogia, então seremos inclinados a optar pelo sexo feminino, pois é que os alunos deste curso são quase que exclusivamente do sexo feminino. Agora se a informação fosse de que o aluno estuda Engenharia a sugestão seria outra, pois a grande maioria dos estudantes de Engenharia são do sexo masculino.

Vamos ver, então, como identificar se existe dependência entre duas variáveis.

Exemplo 1.2

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 1.4.

Tabela 1.4 - Distribuição conjunta dos alunos segundo o sexo (X) e o curso (Y)

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

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. Vamos supor que foram fixados os totais das colunas. Os resultados estão apresentados na tabela 1.5.

Tabela 1.5 - Distribuição conjunta dos alunos segundo o sexo (X) e o curso (Y)

X / Y	Masculino	Feminino	Total
Economia	61	58	**60**
Administração	39	43	**40**
Total	**100**	**100**	**100**

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

Não havendo dependência entre as variáveis, seria esperado as mesmas proporções para cada sexo. Observando a tabela, pode-se constatar que as proporções estão muito 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. Suponha agora um mesmo tipo de exemplo, só que envolvendo alunos dos cursos de Física e Serviço Social, cuja distribuição conjunta está na tabela 1.6.

Tabela 1.6 - Distribuição conjunta dos alunos segundo o sexo (X) e o curso (Y)

X / Y	Masculino	Feminino	Total
Física	100 (71)	20 (33)	**120 (60)**
Ciências Sociais	40 (29)	40 (67)	**80 (40)**
Total	**140 (100)**	**60 (100)**	**200 (100)**

Observe que as tabelas das porcentagens já foi calculada e colocada junto com a das freqüências absolutas. As percentagens foram calculadas, conforme exemplo anterior, em relação ao total das colunas.

Comparando agora a distribuição das proporções pelos cursos, independentes do sexo (coluna de total), com as distribuições diferenciadas por sexo (coluna de masculino e feminino), 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. Quando existe dependência entre variáveis, sempre é interessante quantificar esta dependência, que é que será visto adiante. Observe-se, também, que se teria chegado as mesmas conclusões se tivesse sido utilizado o total de linhas ao invés do total de colunas.

1.3. MEDIDA DO GRAU DE DEPENDÊNCIA ENTRE DUAS 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 medidas que medem a dependência entre duas variáveis nominais. Uma delas é o denominado **coeficiente de contingência**, devido a Karl Pearson.

Exemplo 1.3

Determinar o grau de dependência entre as variáveis da tabela 1.6, anterior.

A análise da tabela já mostrou que existe dependência entre as variáveis. Caso houvesse independência entre elas seria esperado que cada sexo apresentasse 60% de estudantes Física e 40% de estudantes de Ciências Sociais. Neste caso, o número esperado de estudantes masculinos de Física seria: 140x0,60 = 84 e o número esperado de estudantes masculinos de Ciências Sociais seria 140x0,40 = 56. Calculando os demais valores esperados poderíamos formar a tabela dos valores esperados. Tabela 1.7.

Tabela 1.7 – Valores esperados na tabela 1.6, caso as variáveis fossem independentes

X / Y	Masculino	Feminino	Total
Física	84 (60%)	36 (60%)	**120**
Ciências Sociais	56 (40%)	24 (40%)	**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 na tabela 1.8.

Tabela 1.8 – Desvios obtidos entre os valores e esperados, caso as variáveis fossem independentes.

X / Y	Masculino	Feminino
Física	100 - 84 = 16	20 - 36 = -16
Ciências Sociais	40 - 56 = -16	40 – 24 = 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:

$$(O_i - E_i)^2 / E_i$$

Assim, para a célula Física e Masculino, vai-se obter:

$(-16)^2 / 84 = 3,0476$ e para a célula Física e Feminino obtém-se: $(-16)^2 / 36 = 7,1111$.

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**. Para este exemplo, o valor desta medida seria:

$\chi^2 = 3,0476 + 7,1111 + 4,5714 + 10,6667 = 25,3968$.

Quanto maior for este valor, maior será o grau de associação entre as duas variáveis.

De um modo geral a expressão para avaliar o grau de dependência entre as duas variáveis é dado por:

$$\chi^2 = \sum(O_i - E_i)^2 / E_i$$

No entanto, julgar a associação pelo expressão acima não é muito 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).

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:

$C^* = C/[(t - 1)/t]^{1/2}$, onde t é o valor mínimo entre o número de linhas e colunas da tabela.

Para o exemplo acima o coeficiente de Pearson será:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{25,3968}{25,3968 + 200}} = 0,3357 = 0,34.$$

$C^* = 0,4747 = 0,47$.

2. VARIÁVEIS BIDIMENSIONAIS QUANTITATIVAS

2.1. O DIAGRAMA DE DISPERSÃO

Quando as variáveis envolvidas são do tipo quantitativo, pode-se usar o mesmo tipo de análise apresentada para as variáveis nominais e ordinais. A distribuição conjunta pode ser apresentada em tabelas de dupla entrada e através das distribuições marginais pode-se verificar se as variáveis estão ou não relacionadas. Também em certos casos será necessário agrupar os dados em classes ou valores da mesma forma que foi feita no estudo de uma única variável. No entanto, além desta forma de análise é possível a utilização de outros métodos quando as variáveis envolvidas são quantitativas.

Um procedimento bastante útil para estabelecer a associação entre duas variáveis quantitativas é o diagrama de dispersão, que nada mais é do que a representação dos pares de valores num sistema de eixos cartesianos.

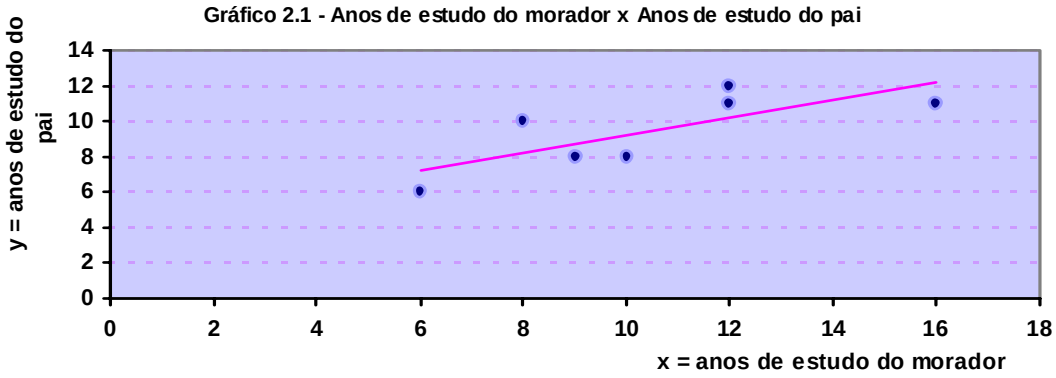
Exemplo 2.1

Na tabela 2.1 abaixo são apresentados os dados correspondentes ao número de anos de escola (X) dos pais e o número de anos de escola (Y) dos filhos de uma amostra de 6 habitantes da capital.

Tabela 2.1 – Anos de escola do pai e anos de escola do filho

Pai (X)	12	10	6	16	8	9	12
Filho (Y)	12	8	6	11	10	8	11

Fazendo o diagrama de dispersão destes valores obtém-se o gráfico abaixo.



Observando o diagrama de dispersão é possível ver que os dados estão seguindo uma dependência aparentemente linear com um relacionamento direto entre os valores (anos de estudo do morador) com o tempo de estudo o pai do morador. Assim à medida que a variável X aumenta a variável Y também aumenta.

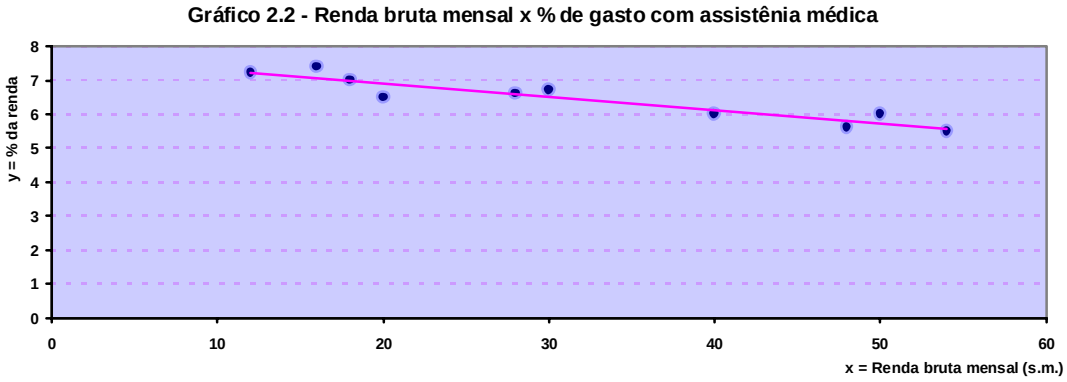
Exemplo 2.2

Considere-se, agora, a tabela 2.2 que retrata os valores da renda bruta mensal (em salários mínimos) de 10 famílias da classe média e o percentual desta renda gasto com assistência médica.

Tabela 2.2 – Renda bruta mensal e % de gastos com saúde

X = Renda bruta (s.m.)	12	16	18	20	28	30	40	48	50	54
(Y) = % gasto	7,2	7,4	7,0	6,5	6,6	6,7	6,0	5,6	6,0	5,5

Observando-se o gráfico de dispersão, pode-se perceber que existe uma tendência (linear) só que agora inversa, isto é, quanto maior a renda bruta mensal, menor é o percentual de gasto com assistência médica.



Exemplo 2.3

Considere-se, agora, o exemplo 2.3, que retrata os valores de 8 alunos (tabela 2.3) que foram submetidos a um teste de língua estrangeira e em seguida foi medido o tempo gasto por cada um para operar uma determinada máquina. Assim:

X = resultado obtido no teste (máximo 100 pontos)

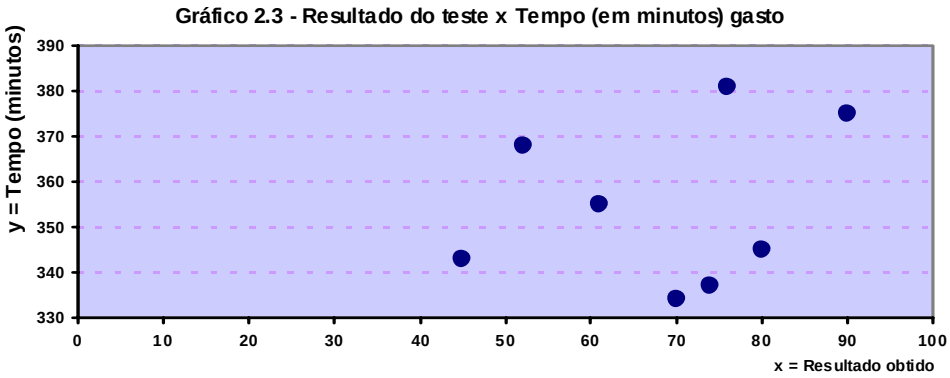
Y = tempo, em minutos, necessário para aprender a operar satisfatoriamente a máquina.

Tabela 2.3 – Resultado obtido no teste e tempo gasto para aprender

X = Resultado	45	52	61	70	74	76	80	90
(Y) = Tempo gasto	342	368	355	334	337	381	345	375

Observando-se o gráfico de dispersão, pode-se perceber que não existe uma tipo de tendência identificável entre as duas variáveis, isto é, quando uma cresce (decresce) a outra cresce (ou decresce).

Neste caso o conhecimento do resultado do teste não ajuda a prever o tempo gasto para operar a máquina.



Pelos exemplos, pode-se perceber que a representação gráfica de variáveis quantitativas ajuda no entendimento do relacionamento entre elas.

2.2. O COEFICIENTE DE CORRELAÇÃO

Observada uma associação entre duas variáveis quantitativas, pode-se então quantificar o valor desta associação. Existem vários tipos de associação possíveis e o que será visto aqui é a do tipo mais simples possível, isto é, o relacionamento linear. Quer dizer que vamos definir uma medida que mede o grau de associação dos pontos em torno de uma linha reta. Esta medida assumirá os valores no intervalo -1 a 1. Com zero indicando ausência de relacionamento **linear**, entre as variáveis. O fato de as variáveis não apresentarem relacionamento linear não implica que elas não apresentem outros tipos de relacionamento.

A determinação do coeficiente de correlação será feita com base nos valores da tabela 2.4 abaixo, que relaciona duas variáveis X = número de horas de estudo e Y = nota na prova de Estatística.

Tabela 2.4 – Ilustração do cálculo do coeficiente de correlação

Par	X	Y	X -	Y -	$(X -)/S_X = Z_X$	$(y-)/S_Y = Z_Y$	$Z_X.Z_Y$
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
Total	25	300	0	0	0	0	4,75

Os cálculos acima, mostram o seguinte procedimento para a determinação do coeficiente de correlação:

- Determinar as médias das variáveis X e Y.
- Determinar os desvios padrões das variáveis X e Y.
- Padronizar as variáveis, isto é, determinar Z_X e Z_Y .
- Obter os produtos dos valores padronizados.
- Obter a média dos produtos dos valores padronizados.

O coeficiente de correlação, isto é, o grau de relacionamento linear entre as variáveis X e Y será então:

r = 4,75 / 5 = 0,95, isto é, existe uma correlação muito forte entre as duas variáveis.

Definição

Dados n pares de valores de duas variáveis X e Y, o coeficiente de correlação entre elas, será anotado por **r** e calculado por:

$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X-\overline{X}}{S_X}\right)\left(\frac{Y-\overline{Y}}{S_Y}\right)$, ou seja, a média dos produtos dos valores padronizados (reduzidos) das variáveis X e Y.

Esta definição não é muito prática. Então na maioria das vezes é melhor utilizar a seguinte expressão alternativa para o cálculo do coeficiente de correlação.

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X-\overline{X}}{S_X}\right)\left(\frac{Y-\overline{Y}}{S_Y}\right) = \frac{\sum XY - n\overline{X}\overline{Y}}{\sqrt{\left(\sum X^2 - n\overline{X}^2\right)\left(\sum Y^2 - n\overline{Y}^2\right)}}$$

Exemplo 2.4

Na tabela 2.5 abaixo estão os dados referentes à percentagem da população economicamente ativa empregada no setor primário e o respectivo índice de analfabetismo para algumas regiões metropolitanas brasileiras. Verificar se existe correlação entre as duas variáveis.

Tabela 2.5 – População economicamente ativa empregada no setor primário e índice de analfabetismo

Regiões metropolitanas	Setor Primário	Índice de analfabetismo
São Paulo	2,0	17,5
Rio de Janeiro	2,5	18,5
Belém	2,9	19,5
Belo Horizonte	3,3	22,5
Salvador	4,1	26,5
Porto Alegre	4,3	16,6
Recife	7,0	36,6
Fortaleza	13,0	38,4

Fonte: Indicadores Sociais para Áreas Urbanas – IBGE – 1977

Os cálculos necessários para a determinação do coeficiente de correlação estão ilustrados na tabela 2.6 abaixo.

Tabela 2.6 – Cálculos para a determinação do coeficiente de correlação

Regiões	Setor (X)	Índice (Y)	XY	X^2	Y^2
A	2,0	17,5			
B	2,5	18,5			
C	2,9	19,5			
D	3,3	22,5			
E	4,1	26,5			
F	4,3	16,6			
G	7,0	36,6			
H	13,0	38,4			

$$r = \frac{\sum XY - n\overline{X}\,\overline{Y}}{\sqrt{\left(\sum X^2 - n\overline{X}^2\right)\left(\sum Y^2 - n\overline{Y}^2\right)}} =$$

Uma das possíveis interpretações do coeficiente de correlação é:

-1,00 ε correlação negativa perfeita.

-0,95 ε correlação negativa forte.

-0,50 ε correlação negativa moderada.

-0,10 ε correlação negativa fraca.

0,00 ε ausência de correlação.

0,10 ε correlação positiva fraca.

0,50 ε correlação positiva moderada.

0,95 ε correlação positiva forte.

1,00 ε correlação positiva perfeita.

2.3. REQUISITOS PARA A DETERMINAÇÃO E USO DO COEFICIENTE DE CORRELAÇÃO

Para a determinação do coeficiente de correlação de Pearson entre duas variáveis X e Y as seguintes condições devem ser levadas em consideração:

- O coeficiente de correlação de Pearson mede somente o relacionamento linear entre as variáveis;
- As variáveis devem ser mensuradas, no mínimo, a nível intervalar, de forma que se possa trabalhar com escores;
- Os valores utilizados devem ter sidos retirados aleatoriamente de uma população, a menos que não se tenha interesse em testar a significância deste coeficiente.
- Se for necessário testar a significância do coeficiente de correlação é necessário que as variáveis X e Y tenham sido extraídas de populações com distribuição normal.

2.4. A REGRESSÃO

2.4.1. INTRODUÇÃO

A regressão e a correlação são duas técnicas estreitamente relacionadas. A análise de correlação fornece um número que traduz o grau de relacionamento linear entre as duas variáveis, enquanto que a análise de regressão fornece uma equação (linear ou do primeiro grau) que descreve o relacionamento entre as duas variáveis. A equação pode ser usada para estimar ou predizer valores de uma das variáveis (variável explicada) conhecidos os valores da outra variável (variável explicativa)

Duas características da equação linear (parâmetros) precisam ser determinados para que se possa conhecer qual é a equação que relaciona duas variáveis X e Y. Uma equação linear tem a forma:

$$Y = a + bX$$

onde **b** é o coeficiente angular (parâmetro de regressão) da reta e **a** é o coeficiente linear (parâmetro linear). O primeiro fornece a inclinação da reta em relação ao eixo dos X e o segundo informa o ponto em que a reta corta o eixo dos Y. O coeficiente angular (**b**) indica a variação da variável Y por unidade de variação da variável X. Assim se o coeficiente angular de uma reta for 3, isto quer dizer que para unidade de variação de X, Y vai variar em 3 unidades.

A primeira providência a ser adotada antes de se tentar determinar uma equação de regressão é construir o diagrama de dispersão para verificar se os dados estão mais ou menos alinhados em torno de uma linha reta. Nem todo o relacionamento entre duas variáveis é do tipo linear. O relacionamento linear é apenas um dentre muitos outros possíveis.

2.4.2. DETERMINAÇÃO DA LINHA DE REGRESSÃO

Vamos supor que foram coletados "n" de valores das variáveis X e Y e que o relacionamento entre as duas variáveis seja do tipo linear, isto é:

$Y = a + bX + E$, onde E = termo erro.

O método para obter a equação de regressão é denominado de método dos mínimos quadrados e consiste em encontrar uma linha que passe pelos pontos de forma que as distâncias verticais de cada ponto dado até a linha sejam mínimas. Suponhamos que a equação desta linha seja:

$Y_c = a + bX$,

então a afirmação acima consiste em resolver a equação:

$\sum(Y - Y_c)^2 =$ mínimo,

onde Y é um valor observado de Y e Y_c é um valor calculado de Y, através da linha dos mínimos quadrados. Os valores de **a** e **b** que satisfazem a equação acima são obtidos através das seguintes expressões:

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} \quad e \quad a = \overline{Y} - b\overline{X}$$

Exemplo 2.5

Determine a equação que descreve a relação entre a freqüência de acidentes e o nível de esforço preventivo educacional com base nos dados abaixo:

Tabela 2.7 – Freqüência de acidentes e esforço educacional

Homens/Horas por mês com educação	2	5	4,5	8	9	1,5	3	6
Acidentes por milhão de homens/hora	7,0	6,4	5,2	4,0	3,1	8,0	6,5	4,4

A tabela 2.8 abaixo resume os cálculos necessários para determinar a equação de regressão dos acidentes em função das horas gastas em educação.

Tabela 2.8 – Cálculos para a determinação da equação de regressão

X	Y	XY	X2	Y2
2	7,0			
5	6,4			
4,5	5,2			
8	4,0			
9	3,1			
1,5	8,0			
3	6,5			
6	4,4			

Determinada a equação de regressão pode-se determinar o erro padrão da regressão, que é o desvio padrão dos erros da regressão. O erro é a diferença entre o valor dado de Y e o valor calculado Y_c, isto é, $E = Y - Y_c$. Este valor informa o quanto os pontos dados estão alinhados. Quanto menor o valor do erro padrão mais próximo da linha de regressão estão os pontos dados.

Tabela 2.9 – Cálculos para a determinação da equação de regressão

X	Y	Y_c	$E = Y - Y_c$	E^2
2	7,0			
5	6,4			
4,5	5,2			
8	4,0			
9	3,1			
1,5	8,0			
3	6,5			
6	4,4			

3. EXERCÍCIOS

(01) De um estudo numa determinada comunidade foram extraídas as seguintes informações:

- A proporção de pessoas solteiras é 0,4.
- A proporção de pessoas que recebem até 10 salários mínimos é 0,2.
- A proporção de pessoas que recebem até 20 salários mínimos é 0,7.
- A proporção de pessoas casadas entre os que recebem mais de 20 salários mínimos é 0,3.
- A proporção de pessoas que recebem até 10 salários mínimos entre os solteiros é de 0,3.

(1.1) Construa a distribuição conjunta das variáveis "estado civil" e "faixa salarial" e as respectivas distribuições marginais.

(1.2) Você diria que existe relação, entre as duas variáveis?

(02) Uma amostra de 200 habitantes de uma cidade foi escolhida ao acaso para analisar a atitude frente a um certo projeto do governo. O resultado está apresentado na tabela abaixo:

	Local de residência			
Opinião	**Urbano**	**Suburbano**	**Rural**	**Total**
A favor	30	35	35	**100**
Contra	60	25	15	**100**
Total	**90**	**60**	**50**	**200**

(2.1) Calcule as proporções em relação ao total das colunas.

(2.2) Você diria que a opinião independe do local de residência?

(2.3) Encontre uma medida de dependência entre as variações.

(03) A tabela, abaixo, mostra os resultados de um questionário para saber se adultos moradores nas proximidades de centros esportivos construídos pela prefeitura participam ou não das atividades programadas. Baseado nos resultados seriam possível dizer que a participação depende da cidade sendo considerada?

	Cidade		
Participa	**Porto Alegre**	**Caxias do Sul**	**Pelotas**
Sim	150	75	**115**
Não	250	225	**235**

(04) Uma pesquisa para verificar a tendência dos alunos a prosseguir os estudos, segundo sua classe social, mostrou os seguintes resultados:

Pretende continuar	Classe social		
	Alta	Média	Baixa
Sim	200	220	380
Não	200	280	720

(4.1) Você diria que a distribuição das respostas afirmativas é igual a de respostas negativas?

(4.2) Existe dependência entre os dois fatores? Dê uma medida que quantifique esta dependência.

(4.3) Se dos 400 alunos da classe alta 160 escolhessem continuar e 240 não, você mudaria sua conclusão? Justifique.

(05) Uma amostra de 5 casais foi colhida em um determinado bairro e seus salários anuais (em milhares de reais) estão na tabela abaixo.

	Casal	1	2	3	4	5
Salário	Homem (X)	10	13	15	17	20
	Mulher (Y)	9	10	13	13	15

(5.1) Encontre o salário anual médio dos homens e o desvio padrão do salário anual dos homens.

(5.2) Encontre o salário anual médio das mulheres e o desvio padrão do salário anual das mulheres.

(5.3) Construa o diagrama de dispersão.

(5.4) Encontre a correlação entre o salário anual dos homens e das mulheres.

(5.5) Qual o salário médio familiar? E a variância?

(5.6) Se o homem é descontado em 8% e a mulher em 6%, qual o salário líquido anual médio familiar? E a variância?

(06) Com relação aos valores da tabela em 5 (cinco) determine:

(6.1) A equação de regressão do salário das mulheres em função dos salários dos homens

(6.2) Faça uma previsão de quanto ganharia uma mulher cujo homem está ganhando 22000 anuais.

(6.3) Determine o erro padrão da regressão.

(07) Com relação aos valores do apêndice:

(7.1) Construa uma tabela de dupla entrada do estado civil em relação ao número de filhos.

(7.2) Determine qual o percentual de pais que são casados e possuem 3 filhos.

(7.3) Determine o percentual de pais solteiros.

(7.4) Dentre os casados qual o percentual dos que não possuem filhos.

(7.5) Determine o percentual de pais com 2 ou mais filhos.

(08) Com relação aos valores do apêndice, construa uma tabela de dupla entrada do estado civil em relação à educação e verifique se as variáveis são dependentes e quantifique esta dependência.

4. APÊNDICE

Tabela 3.1 - Informações sobre o estado civil, grau de instrução, número de filhos, renda (em salários mínimos) idade e procedência de uma amostra de 40 pais dos alunos do Educandário dona Virgulina Travessão.

Número	Estado Civil	Educação	Filhos	Renda	Idade	Procedência
01	Casado	Primeiro	4	4,00	40	Interior
02	Casado	Primeiro	3	5,55	29	Capital
03	Solteiro	Segundo	1	6,60	25	Capital
04	Viúvo	Primeiro	3	3,75	58	Outro
05	Desquitado	Segundo	2	2,90	36	Outro
06	Divorciado	Superior	1	8,85	37	Interior
07	Casado	Primeiro	3	2,25	34	Interior
08	Casado	Primeiro	2	3,20	39	Capital
09	Casado	Segundo	2	7,20	28	Capital
10	Casado	Superior	1	6,60	27	Capital
11	Casado	Superior	3	8,78	49	Outro
12	Casado	Primeiro	5	6,15	68	Interior
13	Desquitado	Segundo	6	6,00	58	Interior
14	Desquitado	Superior	2	9,10	47	Capital
15	Casado	Primeiro	1	8,60	32	Capital
16	Casado	Primeiro	2	3,45	36	Interior
17	Solteiro	Superior	2	4,88	41	Capital
18	Casado	Superior	6	5,45	46	Interior
19	Casado	Superior	3	4,30	37	Outro
20	Casado	Superior	3	6,00	49	Interior
21	Divorciado	Superior	2	5,00	31	Capital
22	Outro	Primeiro	4	3,65	44	Interior
23	Casado	Segundo	1	3,68	43	Capital
24	Casado	Superior	3	7,60	35	Capital
25	Divorciado	Primeiro	3	3,30	29	Interior
26	Outro	Segundo	1	2,50	29	Outro
27	Outro	Segundo	2	3,58	24	Interior
28	Casado	Primeiro	3	1,90	30	Interior
29	Casado	Segundo	3	6,78	58	Capital
30	Casado	Segundo	5	5,80	51	Interior
31	Casado	Superior	2	9,50	53	Capital
32	Casado	Primeiro	3	5,40	45	Interior
33	Casado	Primeiro	3	9,93	38	Interior
34	Solteiro	Segundo	2	8,80	28	Capital
35	Outro	Segundo	1	4,45	25	Outro
36	Divorciado	Primeiro	1	4,68	44	Interior
37	Desquitado	Segundo	2	3,56	33	Interior
38	Casado	Primeiro	3	4,87	29	Interior
39	Casado	Primeiro	2	2,44	44	Capital
40	Casado	Primeiro	3	3,25	38	Outro

Fonte: Dados hipotéticos

5. REFERÊNCIAS

[BUS86] BUSSAB, Wilton O, MORETTIN, Pedro A. *Estatística Básica*. 3° ed. São Paulo, Atual, 1986.

[LEV85] LEVIN, Jack. *Estatística Aplicada a Ciências Humanas.* São Paulo: Editora Harper & Row do Brasil Ltda., 1985.

[NET74] NETO, Pedro Luiz de Oliveira Costa. *Estatística*. São Paulo, Edgard Blücher, 1977.