Estatística e Probabilidade
Aula 4 - Análise bivariada

Professora: Josceli Tenório

Data: 31/03/2022

Introdução

Como calcular a frequência em análise com duas variáveis? Qual o total utilizar? Da coluna? Da linha? Ou o total geral?

Depende da análise desejada. A divisão pelo total geral expressa a composição do grupo por ambas características. A divisão pelo total da linha ou da coluna expressa um resultado condicionado à observação da linha ou coluna. A Figura 1 mostra um exemplo relacionando as variáveis grau de instrução (Y) e região de procedência (V). Verifique que os totais foram apresentados. É a denominada tabela de contingência.

V \ Y	Ensino Fundamental	Ensino Médio	Superior	Total
Capital	4	5	2	11
Interior	3	7	2	12
Outro	5	6	2	13
Total	12	18	6	36

Figura 1.Tabela de contingência: distribuição das frequências absolutas entre as variáveis Y e V.

A Figura 2 mostra a análise pela frequência das observações em relação ao total.

V \ Y	Ensino Fundamental	Ensino Médio	Superior	Total
Capital	11	14	6	31
Interior	8	19	6	33
Outro	14	17	6	36
Total	33	50	17	100

Figura 2. Tabela de contingência: frequência das observações em relação ao total da amostra em %.

A Figura 3 mostra a frequência considerando o total da linha e da coluna.

V \ Y	Ensino Fundamental	Ensino Médio	Superior	Total
Capital	33	28	33	31
Interior	25	39	33	33
Outro	42	33	33	36
Total	100	100	100	100

V \ Y	Ensino Fundamental	Ensino Médio	Superior	Total
Capital	36	45	18	100
Interior	25	58	17	100
Outro	38	46	15	100
Total	33	50	17	100

Figura 3. Análises pelo total da coluna ou da linha da amostra em %.

O que dizem os dados?

No exemplo, a distribuição pelo total das linhas mostra que, por exemplo, 36% dos funcionários da empresa que vieram da capital, terminaram o ensino fundamental.

Por outro lado, no exemplo da divisão pelos totais das colunas, temos que entre os funcionários com ensino médio, 39% vieram do interior.

A Figura 4 mostra as relações entre Y e V.

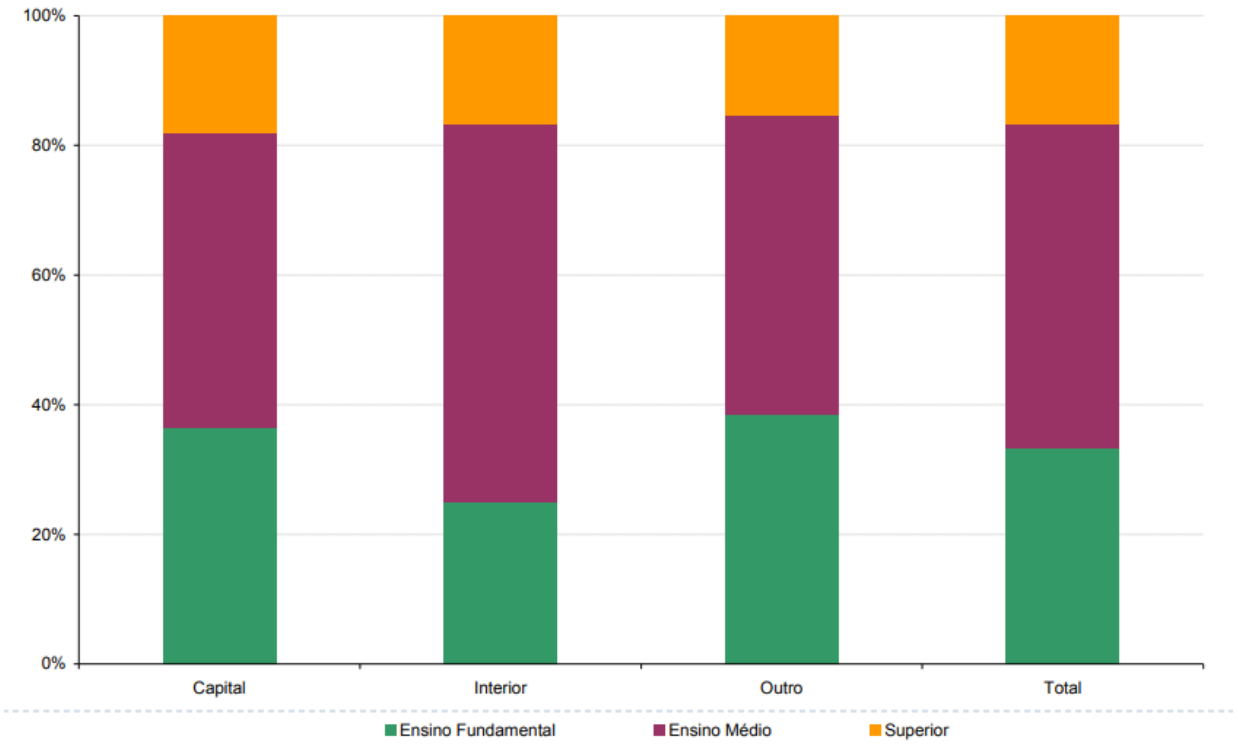


Figura 4. Distribuição do grau de instrução por região de procedência (em %).

Associação entre variáveis

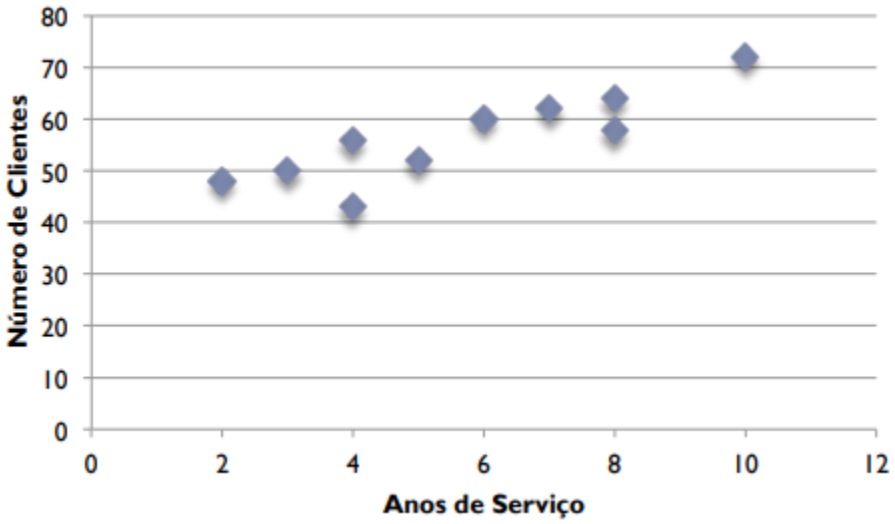
O objetivo de estabelecer a distribuição conjunta de duas variáveis é o de compreender a existência de alguma associação entre elas, ou o grau de dependência entre elas.

Associação entre variáveis quantitativas

Exemplo: Há alguma relação entre o número de clientes e os anos de serviço prestados pela empresa?

Tabela 1. Número de clientes e serviços prestados.

Agente	Anos de serviço (X)	Número de clientes
A	2	48
B	3	50
C	4	56
D	5	52
E	4	43
F	6	60
G	7	62
H	8	58
I	8	64
J	10	72



Por meio do gráfico de dispersão verificam-se possíveis relações entre duas variáveis numéricas.

Covariância

Dados n pares de valores (x1, y1)..., (xn, yn), chamaremos de covariância entre as variáveis X e Y, consideradas como população:

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n}$$

Vamos denominar a cov(X,Y) como $\mathbf{s_{xy}}$.

Ao considerar dados amostrais, a mesma correção aplicada ao cálculo da variância é utilizada e o denominador deverá ser **n-1**.

Se as variáveis estiverem correlacionadas de alguma forma, sua covariância será diferente de zero. De fato, se cov(X,Y)>0, então Y tende a aumentar à medida que X aumenta, e se cov(X,Y)<0, então Y tende a diminuir à medida que X aumenta. Observe que, embora as variáveis estatisticamente independentes sejam sempre não correlacionadas, a recíproca não é necessariamente verdadeira.

A covariância é a medida do afastamento simultâneo das respectivas médias.

Se as ambas variáveis aleatórias tendem a estar simultaneamente acima, ou abaixo, de suas respectivas médias, então a covariância tenderá a ser positiva e nos outros casos poderá ser negativa, como mostram os gráficos abaixo.

Características da covariância

A covariância de uma variável e ela mesma é a própria variância da variável, seja no caso de população ou amostra. Como Y = X, $\mathbf{s_{xy}}$ será:

$$\frac{\sum_{i=1}^{N}(X_i - \mu_X)\times(X_i - \mu_X)}{N} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)^2}{N};$$

Dessa forma teremos $\mathbf{s_{xy} = s_{xx}^2}$

A permutação das variáveis não altera o resultado da covariância, se os pares de valores não forem alterados: $\mathbf{s_{xy} = s_{yx}}$

Da mesma forma que a variância, a covariância é afetada pelos valores extremos da variável, ela não é uma medida resistente.

A unidade de medida é o resultado do produto das unidades dos valores das variáveis.

Coeficiente de correlação

Para facilitar o entendimento da relação entre duas variáveis e evitar a unidade de medida da covariância, foi definido o coeficiente de correlação r_{XY}.

Os valores de r_{XY} estão limitados entre os valores -1 e +1, e sem nenhuma unidade de medida.

O coeficiente de correlação busca auferir a direção da relação entre as variáveis, dentro de um intervalo determinado entre -1 e 1.

O objetivo do intervalo é discriminar a direção e a intensidade da relação:

- valores próximos de zero indicam ausência de relação entre as variáveis.
- valores próximos de 1 indicam forte relação positiva.
- valores próximos de -1 indicam forte relação negativa.

Um dos coeficientes de correlação mais conhecidos é o coeficiente de correlação de Pearson, porém sensível a uma relação linear entre duas variáveis. A relação que possibilita calcular o coeficiente de correlação amostral de Pearson entre as variáveis é:

r: Coeficiente de correlação amostral

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

onde :

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

são os desvios - padrão da variável x e y, respectivamente.

Se definirmos:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = (n-1)s_x^{\,2}$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1)s_y^{\,2}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

De forma simplificada, podemos escrever:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

Há outros coeficientes de correlação mais robustos que o coeficiente de correlação de Pearson. Isto é, mais sensíveis às relações não lineares.

Exemplo: Utilizando a tabela anterior e considerando como dados populacionais (n=10).

Agente	Anos de serviço (X)	Número de clientes	$x - \bar{x}$	$y - \bar{y}$	$\frac{x - \bar{x}}{dp(X)} = z_x$	$\frac{y - \bar{y}}{dp(Y)} = z_y$
A	2	48	-3,7	-8,5	-1,54	-1,05
B	3	50	-2,7	-6,5	-1,12	-0,80
C	4	56	-1,7	-0,5	-0,71	-0,06
D	5	52	-0,7	-4,5	-0,29	-0,55
E	4	43	-1,7	-13,5	-0,71	-1,66
F	6	60	0,3	3,5	0,12	0,43
G	7	62	1,3	5,5	0,54	0,68
H	8	58	2,3	1,5	0,95	0,18
I	8	64	2,3	7,5	0,95	0,92
J	10	72	4,3	15,5	1,78	1,91

Média: x = 5,7 e y = 56,5

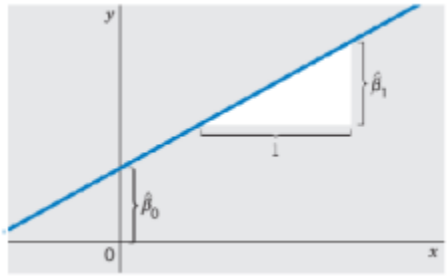
s_x^2 = 58,1/10 = 5,81 = > s_x = sqrt(5,81) = 2,41

s_Y^2 = 658,5/10 = 65,85 = > s_Y = sqrt(65,85) = 8,11

O coeficiente de correlação é:

r_{xy} = 8,77/10=0,877

Se existe correlação significativa e razões para usar uma variável para prever a outra, qual a equação da reta que melhor se ajusta ao gráfico de dispersão?



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Como obter os coeficientes a partir dos dados amostrais?

Melhor ajuste: **método dos mínimos quadrados**.

Minimiza-se a soma das distâncias verticais entre um ponto amostral e a reta procurada.

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2$$

Solução:

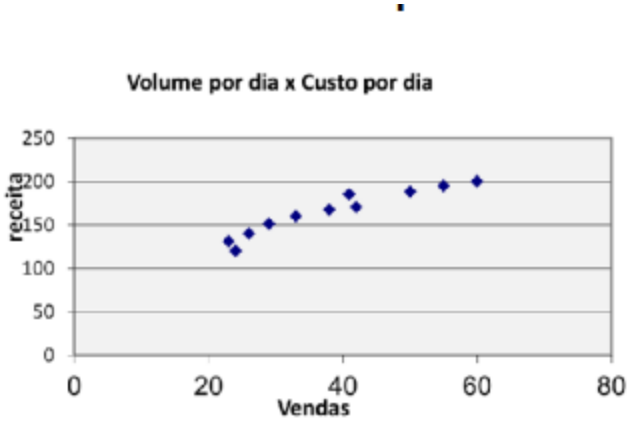
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Exemplo:

receita	vendas
131	23
120	24
140	26
151	29
160	33
167	38
185	41
170	42
188	50
195	55
200	60



Observa-se uma relação aproximadamente linear entre as variáveis.

Para obter a reta ajustada precisamos calcular os coeficientes linear e angular:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{421}{11} = 38.2727$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{1807}{11} = 164.2727$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 3229.182$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - y)^2 = 7104.182$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 1612.182$$

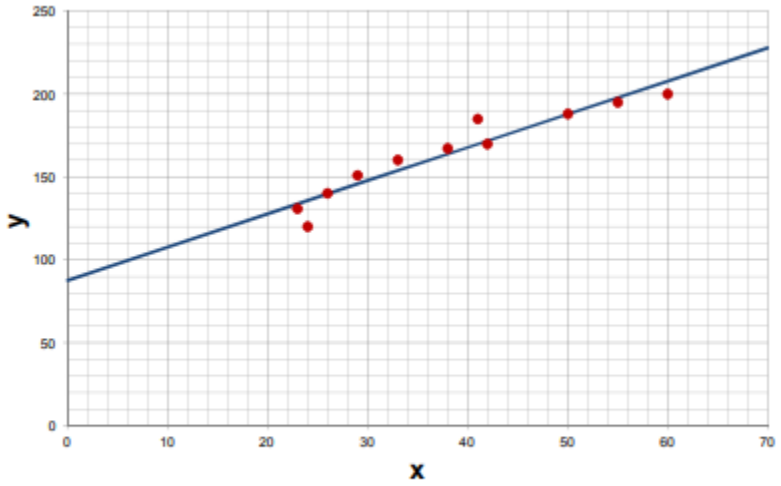
Temos :

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{1612.182}{\sqrt{3229.182}\sqrt{7104.182}} = 0.9542$$

$$\beta_1 = \frac{S_{xy}}{S_{xx}} = \frac{1612.182}{3229.182} = 2.003$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 87.613$$

A reta ajustada será:



$$y = 2.003x + 87.613$$

Cuidado!!!

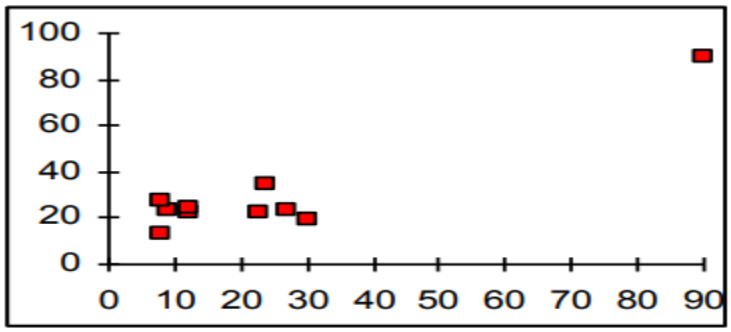
O coeficiente de correlação não mede a relação causa efeito entre as variáveis, apesar de que essa relação possa estar presente.

Uma correlação fortemente positiva entre as variáveis X e Y não autoriza afirmar que variações da variável X provocam variações na variável Y, ou vice-versa.

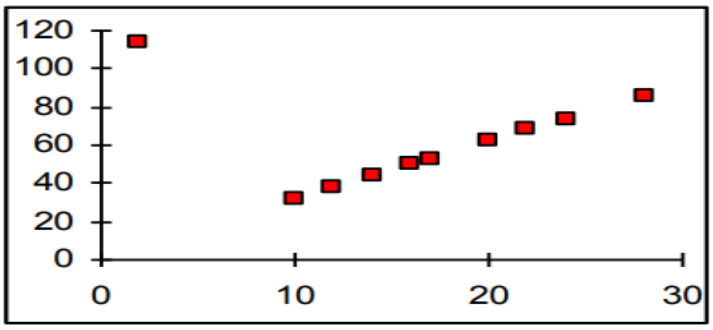
O coeficiente de correlação sozinho não identifica a relação causa-efeito entre as duas variáveis

Exemplos de anomalias

Para r próximo a + 1



Para r próximo a 0.



Aplicações

1. A partir da tabela de investimentos, crie tabelas de contingência, analise linhas e colunas e interprete os resultados.

Categoria	Investidor A	Investidor B	Investidor C
Ações	46,5	55	27,5
Bonds	32,0	44	19,0
RF	15,5	20	13,5
Poupança	16,0	28	7,0

2. A tabela mostra a evolução do PIB e do consumo da Alemanha entre 1999 e 2008, em milhões de euros correntes. Calcule a covariância, correlação e obtenha a reta de regressão.

	PIB	Consumo
1999	2012000	1175010
2000	2062500	1214160
2001	2113160	1258570
2002	2143180	1263460
2003	2163800	1284600
2004	2210900	1303090
2005	2243200	1324650
2006	2321500	1355140
2007	2422900	1373720
2008	2491400	1404570