Resumo 5 - Análise Bivariada (Bidimensional)

5.1. *Introdução*

O principal objetivo das análises nessa situação é explorar relações (similaridades) entre duas variáveis. A distribuição conjunta das freqüências será um instrumento poderoso para a compreensão do comportamento dos dados.

Quando consideramos duas variáveis (ou dois conjuntos de dados), podemos ter três situações:

(1) as duas variáveis são quantitativas;

(2) as duas variáveis são qualitativas; e

(3) uma variável é qualitativa e outra é quantitativa.

As técnicas de análise de dados nas três situações são diferentes. Quando as duas variáveis são quantitativas, as observações são provenientes de mensurações, e técnicas como gráficos de dispersão ou de quantis são apropriados. Quando as variáveis são qualitativas, os dados são resumidos em tabelas de contingências (dupla entrada), onde aparecerão as freqüências absolutas ou contagens de indivíduos que pertencem simultaneamente a categorias de uma e outra variável. Quando temos uma variável qualitativa e outra quantitativa, em geral analisamos o que acontece com a variável quantitativa quando os dados são categorizados de acordo com os diversos atributos da variável qualitativa. Mas podemos ter também o caso de duas variáveis quantitativas agrupadas em classes. Por exemplo, podemos querer analisar a associação entre renda e consumo de certo número de famílias e, para isso, agrupamos as famílias em classes de renda e classes de consumo. De modo geral, recaímos numa tabela de dupla entrada.

Em todas as situações, o objetivo é encontrar as possíveis relações ou associações entre as duas variáveis.

5.2. *Associação entre Variáveis Quantitativas*

Um dispositivo bastante útil para se verificar a associação entre duas variáveis quantitativas, ou entre dois conjunto de dados, é o gráfico de dispersão (ou diagrama de dispersão).

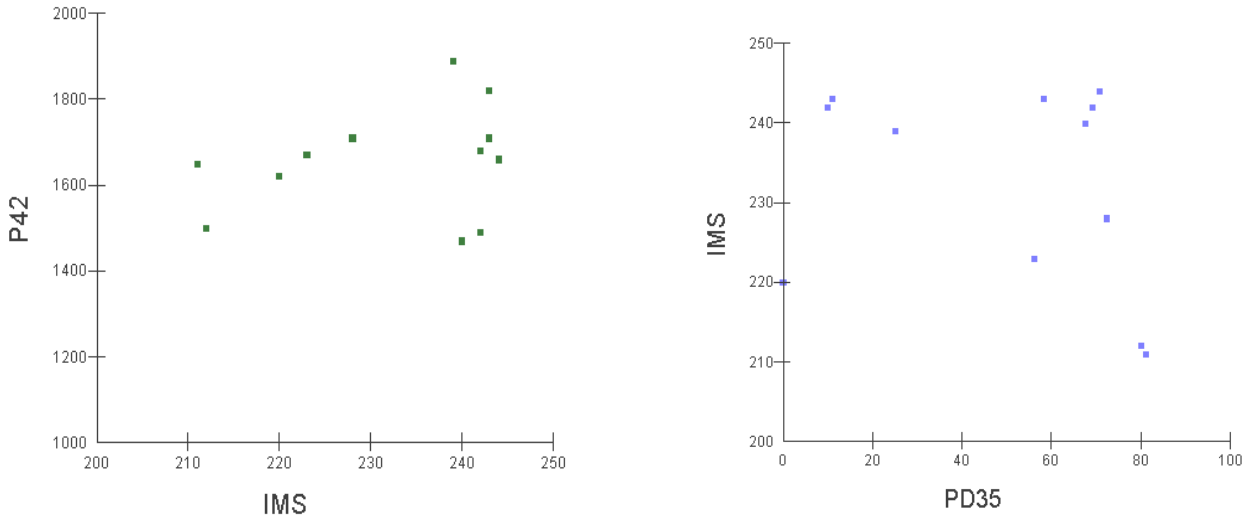
5.2.1. Diagrama de Dispersão

Os dados para um diagrama de dispersão consistem numa única amostra de indivíduos com duas medidas x e y feita em cada indivíduo. As medidas de cada indivíduo são representadas graficamente como um único ponto, tendo a medida x como abscissa e a y como ordenada.

Exemplo 1: Dados relativos a algumas características de importância econômica em linhagem materna de frango de corte, de uma amostra aleatória de 11 aves, do CNPSA/EMBRAPA. Usando seus conhecimentos de estatística, faça um resumo descritivo das variáveis.

Ave	p42	IMS	PD35	PD64
1	1820	243	11,11	55,66
2	1710	243	44,44	77,83
3	1470	240	58,33	75,35
4	1500	212	67,50	69,96
5	1490	242	80,00	77,00
6	1680	242	10,00	75,25
7	1890	239	69,23	80,18
8	1660	244	25,00	69,50
9	1710	228	70,83	65,61
10	1670	223	72,41	72,41
11	1650	211	56,10	61,48

P42: peso corporal aos 42 dias de idade (g); IMS: idade à maturidade sexual (dias); PD35: produção de ovos às 35 semanas de idade (%); PD64: produção de ovos às 64 semanas de idade (%).



5.2.2. Coeficiente de Correlação de Pearson

O *coeficiente de correlação de Pearson*, é uma medida da relação entre duas características numéricas, simbolizadas por X e Y. A fórmula para o coeficiente de correlação, simbolizada por r, é:

$$r = \frac{S_{XY}}{(S_X . S_Y)} ,$$

onde, S_{XY} representa a covariância entre X e Y. A covariância é uma medida que informará sobre a variabilidade conjunta de duas variáveis numéricas (quantitativas). Define-se como:

$$S_{XY} = \frac{\sum_{i=1}^{n} [(x_i - \bar{x})(y_i - \bar{y})]}{(n-1)} .$$

- Se $S_{XY} > 0$, as duas variáveis crescem ou decrescem conjuntamente.
- Se $S_{XY} < 0$, quando uma variável cresce, a outra tem tendência a decrescer.
- Se $S_{XY} = 0$, não há relação linear.

A covariância é afetada pelas unidades em que cada variável é medida, o coeficiente de correlação não. O coeficiente de correlação satisfaz $-1 \leqslant r \leqslant +1$.

Qual deve ser o tamanho do coeficiente de correlação??? Depende da aplicação. Por exemplo, quando as características físicas são medidas e se dispõe de bons dispositivos de medidas como em muitas ciências físicas, são possíveis correlações relativamente elevadas. Entretanto, as medições nas ciências biológicas freqüentemente envolvem características menos bem definidas e dispositivos de medidas imprecisos; em tais casos podem ocorrer correlações mais baixas. Colton (1974) fornece uma regra prática para a interpretação da dimensão de tais correlações:

Correlações entre 0 e 0,25 (ou -0,25) indicam relação pequena ou inexistente;

Correlações entre 0,25 e 0,50 (ou -0,25 e -0,50) indicam um grau razoável de relação;

Correlações entre 0,50 e 075 (ou -0,50 e -o,75) indicam uma relação moderada a boa;

Correlações maiores que 0,75 (ou -0,75) representam uma relação muito boa a excelente.

Exercício: Calcule o coeficiente de correlação entre as variáveis peso corporal aos 42 dias de idade e idade à maturidade sexual no exemplo1.

5.3. *Associação entre Variáveis Qualitativas*

Um dos principais objetivos de se construir uma distribuição conjunta de duas variáveis qualitativas é descrever a associação entre elas, isto é, conhecer o grau de dependência entre elas, de modo que se possamos prever melhor o resultado de uma delas quando conhecermos a realização da outra.

Exemplo 2: Queremos verificar se existe ou não associação entre sexo e uma determinada doença numa amostra de 200 animais. Esses dados estão na tabela a seguir:

Doença (Y)	Sexo (X)		Total
	Masculino	Feminino	
Doentes	85	35	120
Não doentes	55	25	80
Total	140	60	200

Inicialmente, verificamos que fica muito difícil tirar alguma conclusão, devido à diferença entre os totais marginais. Construindo as proporções segundo as linhas ou as colunas poderemos fazer comparações. Fixemos os totais das colunas, a distribuição conjunta (perfil coluna) está na tabela a seguir:

Doença (Y)	Sexo (X)		Total
	Masculino	Feminino	
Doentes	61%	58%	60%
Não doentes	39%	42%	40%
Total	100%	100%	100%

A partir dessa tabela podemos observar que, independentemente do sexo, 60% dos animais apresentaram a doença e 40% não. Não havendo dependência entre as variáveis, esperaríamos essas mesmas proporções para cada sexo. Observando a tabela, vemos que as proporções do sexo masculino (61% e 39%) e do feminino (58% e 42%) são próximas das

marginais (60% e 40%). Esses resultados podem indicar não haver dependência entre as duas variáveis, para o conjunto de animais considerado. Concluímos, então, que nesse caso, as variáveis sexo e a manifestação de uma determinada doença parecem ser NÃO ASSOCIADAS.

Exemplo 3: Queremos verificar se existe ou não associação entre vacinação e a manifestação de uma determinada doença em 200 ovinos da raça Hampshire Down. Esses dados estão na tabela a seguir:

Vacina (Y)	Doença (X)		Total
	Não contraíram	Contraíram	
Vacinados	100 (71%)	20 (33%)	120 (60%)
Não vacinados	40 (29%)	40 (67%)	80 (40%)
Total	140 (100%)	60 (100%)	200 (100%)

Comparando a distribuição conjunta (perfil coluna) das proporções pela vacinação, independentemente da doença, com as distribuições diferenciadas pela doença, observamos uma disparidade bem acentuada nas proporções. Parece haver maior concentração de animais vacinados que não contraíram a doença e de animais não vacinados que contraíram a doença. Nesse caso, as variáveis vacinação e manifestação de uma doença parecem ser ASSOCIADAS.

Quando existe associação entre as variáveis, sempre é interessante quantificar essa associação.

5.3.*1. Coeficiente de Contingência de Pearson*

Pearson definiu uma medida de associação, chamada coeficiente de contingência, com interpretação análoga ao coeficiente de correlação, dado por:

$$C = \sqrt{\frac{\chi^2}{(\chi^2 + n)}} \text{ , com } \chi^2 = \sum \sum \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right],$$

sendo

O_{ij} = freqüência observada na i-ésima categoria de X e j-ésima categoria de Y, e

E_{ij} = freqüência esperada na i-ésima categoria de X e j-ésima categoria de Y.

Contudo, o coeficiente descrito acima não varia entre 0 e 1. O valor máximo de C depende do número de linhas e colunas. Para evitar esse inconveniente, costuma-se definir um outro coeficiente, chamado de Coeficiente de Contingência Modificado, dado por:

$$C^{\circ} = \sqrt{\frac{(k\,\chi^2)}{[(k-1)(\chi^2 + n)]}},$$

onde k é o menor valor entre o número de linhas e o número de colunas da tabela. O coeficiente de contingência modificado satisfaz $0 \leqslant C^{\circ} \leqslant 1$.

Exercício: Calcule o coeficiente de associação para as variáveis vacinação e manifestação de uma doença no exemplo 3.

5.4. *Associação entre Variáveis Qualitativas e Quantitativas*

É comum nessas situações analisar o que acontece com a variável quantitativa dentro de cada categoria da variável qualitativa. Essa análise pode ser conduzida por meio de medidas-resumo, histogramas, box-plots, ou ramo-e-folhas.

Exemplo 4: Desejamos analisar o comportamento dos salários dentro de cada categoria de grau de instrução, ou seja, investigar o comportamento conjunto das variáveis X (salário) e Y (grau de instrução).

Grau de Instrução	n	média	dp	var	min	Q_1	med	Q_3	max
Fundamental	12	7,84	2,79	7,77	4,00	6,01	7,13	9,16	13,65
Médio	18	11,54	3,62	13,10	5,73	8,84	10,91	14,48	23,30
Superior	6	16,48	4,11	16,89	10,53	13,65	16,74	18,38	23,30
Todos	36	11,12	4,52	20,46	4,00	7,55	10,17	14,06	23,30

A leitura desses resultados sugere uma dependência dos salários em relação ao grau de instrução.

É conveniente ter uma medida que quantifique o grau de dependência entre as variáveis.

Com esse objetivo, convém observar que as variâncias podem ser usadas para construir essa medida. Sem usar a informação da variável categorizada, a variância calculada para a variável quantitativa para todos os dados mede a dispersão dos dados globalmente. Se a variância dentro de cada categoria for pequena e menor do que a global, significa que a variável qualitativa melhora a capacidade de previsão da quantitativa e portanto existe uma relação entre as duas variáveis.

Observe que para as variáveis X e Y, as variâncias de X dentro das três categorias são menores do que a global.

O grau de associação entre as duas variáveis pode ser definido como o ganho relativo na variância, obtido pela introdução da variável qualitativa, que satisfaz $0 \leqslant R^2 \leqslant 1$ é dado por:

$$R^2 = \frac{var(X) - [\overline{var}(X)]}{[var(X)]} \quad \text{ou} \quad R^2 = 1 - \frac{[\overline{var}(X)]}{[var(X)]},$$

onde $[\overline{var}(X)] = \dfrac{\sum_{i=1}^{k} (n_i var_i(X))}{\sum_{i=1}^{k} n_i}$ é a média das variâncias ponderada pelo número de observações, sendo k igual ao número de categorias e var_i (X) a variância de X dentro de cada categoria i, i = 1, 2, ..., k.

Exercício: Calcule o o grau de associação entre as variáveis salário e grau de instrução do exemplo 4.

5.5. *Coeficiente de Correlação de Postos de Spearman*

A correlação classificatória de Spearman (ou Correlação por postos), algumas vezes chamada de rho de Spearman, é freqüentemente usada para descrever a relação entre duas características ordinais. Usa apenas a ordem das observações e não o valor observado.

Este coeficiente não é sensível a assimetrias na distribuição, nem à presença de outliers, não exigindo portanto que os dados provenham de duas populações normais.

É também a estatística adequada para ser usada com variáveis numéricas como alternativa ao coeficiente de correlação de Pearson, quando esse último tem violada a condição de normalidade (simetria) e a de relação linear entre as variáveis.

Nos caso em que os dados não formam uma nuvem "bem comportada", com alguns pontos muito afastados dos restantes, ou em que parece existir uma relação crescente ou decrescente em formato de curva, o coeficiente ρ de Spearman é mais apropriado.

O cálculo da correlação classificatória de Spearman, simbolizada por ρ, envolve a colocação dos valores em ordem de classificação em cada uma das características, desde a mais baixa até a mais elevada; os postos são em seguida tratados como se fossem os verdadeiros valores.

O coeficiente de correlação de Spearman é definido por:

$$\rho = 1 - [6.\frac{(\sum d_i^2)}{(n\,(n^2-1))}] \ ,$$

onde d_i é a diferença entre cada posto de valor correspondentes de x e y e, *n* é o número de pares dos valores.

A correlação classificatória de Spearman pode variar de -1 a +1, como o coeficiente de correlação de Pearson; + 1 ou -1 indica o acordo perfeito entre as classes dos valores em vez daquela entre os próprios valores. Caso contrário, sua interpretação é semelhante ao r de Pearson.

Exemplo 5: Os dados a seguir são relativos a um estudo correlacional entre peso corporal, em Kg, de 12 borregos 2 dentes e 4 dentes da raça Hampshire Down.

Animais	Borregos 2 dentes (14 meses)	Borregos 4 dentes (20 meses)
1	60	80
2	58	72
3	63	80
4	51	83
5	54	72
6	55	92
7	48	69
8	70	88
9	65	79
10	53	82
11	62	85
12	52	79

Exercício: Calcule o coeficiente de correlação de Spearman para os dados do exemplo 5.