

Análise exploratória

Análises bivariadas

Prof. Me. Lineu Alberto Cavazani de Freitas

Departamento de Estatística
Laboratório de Estatística e Geoinformação



Análise exploratória

- ▶ Parte primordial de qualquer análise estatística é chamada **análise descritiva** ou **exploratória**.
- ▶ Consiste basicamente de **tabelas**, **resumos numéricos** e **análises gráficas** das variáveis disponíveis em um conjunto de dados.
- ▶ Trata-se de uma etapa de extrema importância e deve preceder qualquer análise mais sofisticada.
- ▶ As técnicas de análise exploratória visam **resumir** e **apresentar** as informações de um conjunto de dados brutos.

Análise exploratória

- ▶ A análise exploratória de dados é uma área relativamente nova.
- ▶ Nasceu do clássico livro **Exploratory Data Analysis** de **John Tukey** em 1977.
- ▶ Algo curioso é que Tukey tinha uma relação próxima com a Ciência da Computação e definiu os termos **bit** e **software**.

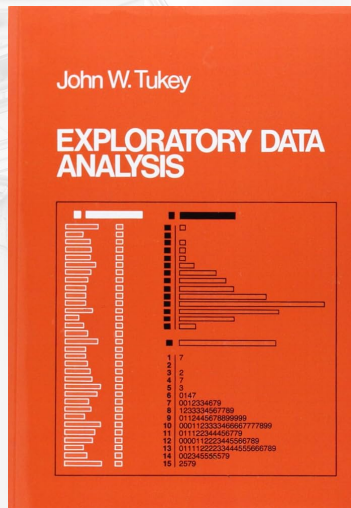


Figura 1. Capa do livro *Exploratory Data Analysis* de John Tukey.

Análise exploratória

- ▶ Como quase tudo em análise de dados, o **avanço computacional** permitiu com que a análise exploratória evoluísse substancialmente.
- ▶ Por exemplo: historicamente o processo de criação de um gráfico era reservado a pessoas qualificadas pois a produção de uma visualização era difícil.
- ▶ Hoje qualquer pessoa pode inserir dados em um aplicativo e gerar um gráfico.
- ▶ Este tipo de facilidade é importante para disseminação e democratização dos métodos, porém abre margem para certas práticas inadequadas.

Análise exploratória

- ▶ Tentar compreender um conjunto de dados sem algum método que permita resumir as informações é inviável.
- ▶ A análise exploratória é a primeira forma de tentarmos entender o que acontece nos nossos dados.
- ▶ Uma das tarefas é a etapa de consistência dos dados, isto é, verificar se os dados coletados são condizentes com a realidade.



Figura 2. Extraído de pixabay.com.

Análise exploratória

- ▶ O conjunto de técnicas aplicáveis está diretamente associado ao **tipo das variáveis de interesse** (quantitativas x qualitativas) e suas ramificações.
- ▶ Podemos conduzir análises focadas nas variáveis uma a uma (**análises univariadas**).
- ▶ Também podemos conduzir análises focadas em avaliar a relação entre as variáveis (**análises multivariadas**).



Figura 3. Extraído de pixabay.com.

Análise exploratória

Podemos fazer uso diversas técnicas, tais como

- ▶ Tabelas de frequência absolutas.
- ▶ Tabelas de frequência relativas.
- ▶ Tabelas de frequência acumuladas.
- ▶ Tabelas para múltiplas variáveis.
- ▶ Gráficos.
- ▶ Medidas de posição central.
- ▶ Medidas de posição relativa.
- ▶ Medidas de forma.
- ▶ Medidas de dispersão.
- ▶ Medidas de associação.

Análise exploratória

- ▶ Para ilustrar as técnicas de análise exploratória de dados, usaremos o conjunto de dados “milsa”.
- ▶ Este conjunto de dados aparece no livro “Estatística Básica” de W. O. Bussab e P. A. Morettin.
- ▶ Conjunto de dados hipotético de atributos de 36 funcionários da companhia “Milsa”.

O conjunto possui as seguintes variáveis:

- ▶ **Funcionário:** identificadora de funcionário.
- ▶ **Estado civil:** casado ou solteiro.
- ▶ **Instrução:** 1º grau, 2º grau, superior.
- ▶ **Filhos:** número de filhos.
- ▶ **Salário:** salário do funcionário.
- ▶ **Anos:** idade em anos completos.
- ▶ **Meses:** meses além dos anos completos.
- ▶ **Região:** capital, interior, outro.

Análise exploratória

Tabela 1. Primeiras linhas do conjunto de dados Milsa.

Funcionário	Estado civil	Instrução	Filhos	Salário	Anos	Meses	Região
1	solteiro	1o Grau	NA	4.00	26	3	interior
2	casado	1o Grau	1	4.56	32	10	capital
3	casado	1o Grau	2	5.25	36	5	capital
4	solteiro	2o Grau	NA	5.73	20	10	outro
5	solteiro	1o Grau	NA	6.26	40	7	outro
6	casado	1o Grau	0	6.66	28	0	interior
7	solteiro	1o Grau	NA	6.86	41	0	interior
8	solteiro	1o Grau	NA	7.39	43	4	capital
9	casado	2o Grau	1	7.59	34	10	capital
10	solteiro	2o Grau	NA	7.44	23	6	outro



Análise exploratória bivariada

Análise exploratória bivariada

- ▶ Em alguns casos podemos estar interessados na análise de **duas variáveis simultaneamente**.
- ▶ O objetivo é investigar a relação de **associação** entre as variáveis.
- ▶ **Tabelas, gráficos e coeficientes** específicos para relação entre variáveis podem ser usados.
- ▶ Tal como nas análises univariadas, as escolhas dependem dos tipos das variáveis.
- ▶ Considerando variáveis aos pares, as combinações podem ser:
 - ▶ Qualitativa x qualitativa.
 - ▶ Quantitativa x quantitativa.
 - ▶ Quantitativa x qualitativa.



Análise bivariada para variáveis qualitativas

Análise bivariada para variáveis qualitativas

- ▶ Neste tipo de situação avaliamos a **frequência** de observações para cada **combinação** de níveis das duas variáveis.
- ▶ Podem ser usadas **tabelas de frequências cruzadas**, também chamadas de **tabelas de dupla entrada**.
- ▶ Também é possível representar as frequências por meio de **recursos gráficos**.

Tabelas de frequências cruzadas

- ▶ As **linhas** dizem respeito aos **níveis** de uma variável.
- ▶ As **colunas** aos **níveis** da outra variável.
- ▶ As **células** mostram as **frequências** (absolutas ou relativas).
- ▶ As tabelas de dupla entrada também são chamadas de **distribuição conjunta**.
- ▶ As **margens** mostram as **frequências marginais** (de apenas uma das duas variáveis), também chamada de **distribuição marginal**.
- ▶ No caso de frequências relativas podem ser usados o **total geral** ou os totais **linha e coluna**.

Tabelas de frequências cruzadas

Tabela 2. Tabela de dupla entrada usando frequências absolutas.

	capital	interior	outro	Total
casado	7	8	5	20
solteiro	4	4	8	16
Total	11	12	13	36

Tabelas de frequências cruzadas

Tabela 3. Tabela de dupla entrada usando frequências relativas.

	capital	interior	outro	Total
casado	0.19	0.22	0.14	0.56
solteiro	0.11	0.11	0.22	0.44
Total	0.31	0.33	0.36	1.00

Tabelas de frequências cruzadas

Tabela 4. Tabela de dupla entrada usando frequências relativas aos totais linha.

	capital	interior	outro	Total
casado	0.35	0.40	0.25	1
solteiro	0.25	0.25	0.50	1
Total	0.31	0.33	0.36	1

Tabelas de frequências cruzadas

Tabela 5. Tabela de dupla entrada usando frequências relativas aos totais coluna.

	capital	interior	outro	Total
casado	0.64	0.67	0.38	0.56
solteiro	0.36	0.33	0.62	0.44
Total	1.00	1.00	1.00	1.00

Análise bivariada para variáveis qualitativas

- ▶ As **frequências cruzadas** podem ser representadas por meio de gráficos.
- ▶ Variações de **gráficos de barras** são as opções mais comuns.
- ▶ As possibilidades podem usar as **frequências absolutas, relativas** e permitem comparar a **composição** das variáveis.
- ▶ Gráficos para frequência para duas variáveis qualitativas:
 - ▶ Gráficos de barras lado a lado.
 - ▶ Gráfico de barras empilhadas.
 - ▶ Gráficos de barras empilhadas relativo.

Gráficos de barras lado a lado

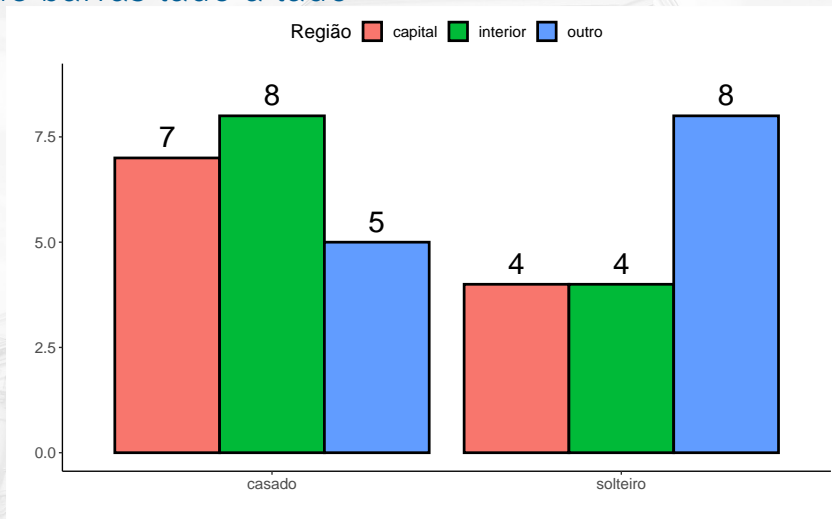


Figura 4. Gráfico de barras lado a lado.

Gráficos de barras lado a lado

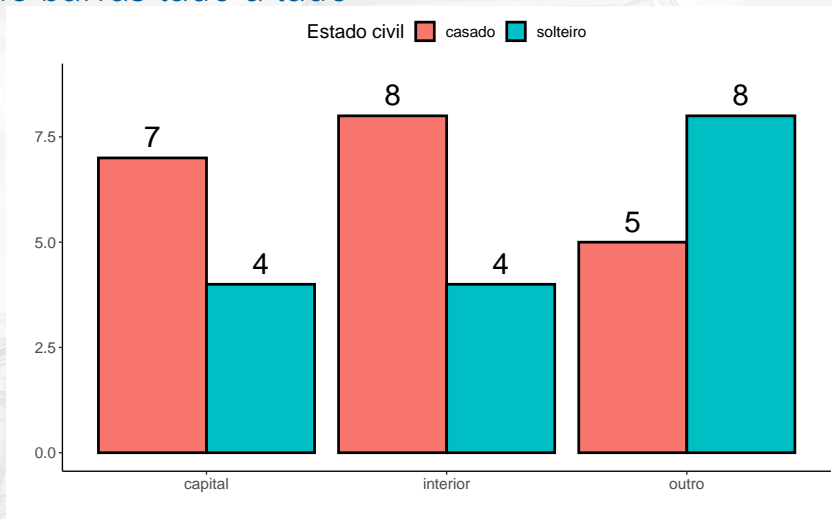


Figura 5. Gráfico de barras lado a lado.

Gráficos de barras empilhadas

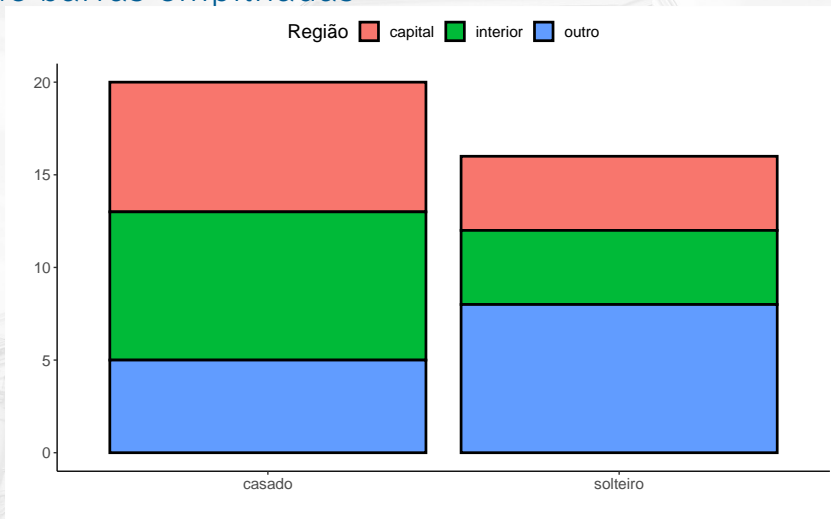


Figura 6. Gráfico de barras empilhadas.

Gráficos de barras empilhadas

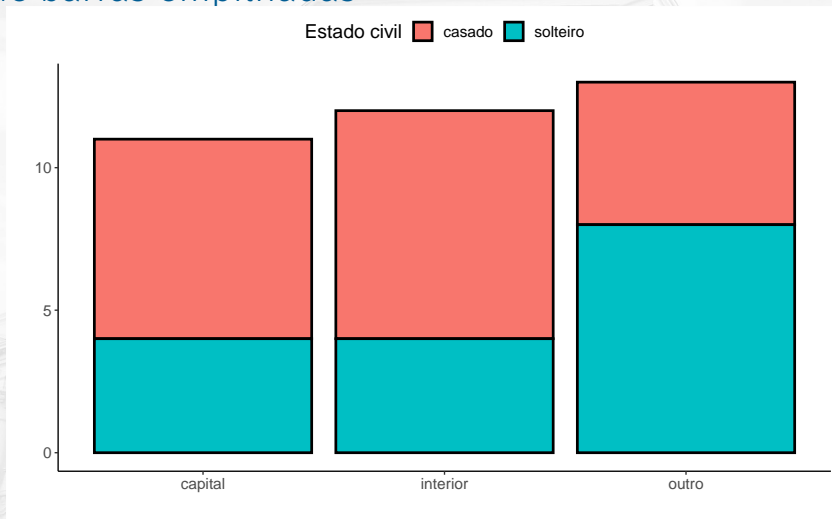


Figura 7. Gráfico de barras empilhadas.

Gráficos de barras empilhadas relativo

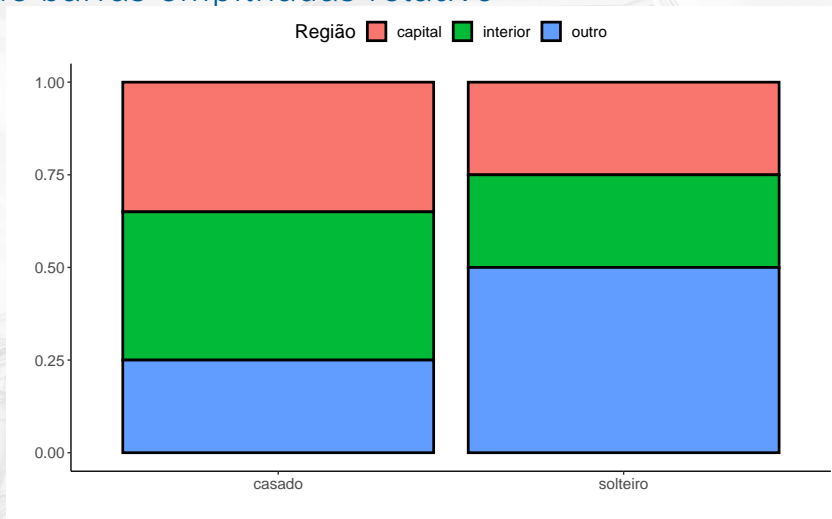


Figura 8. Gráfico de barras empilhadas relativo.

Gráficos de barras empilhadas relativo

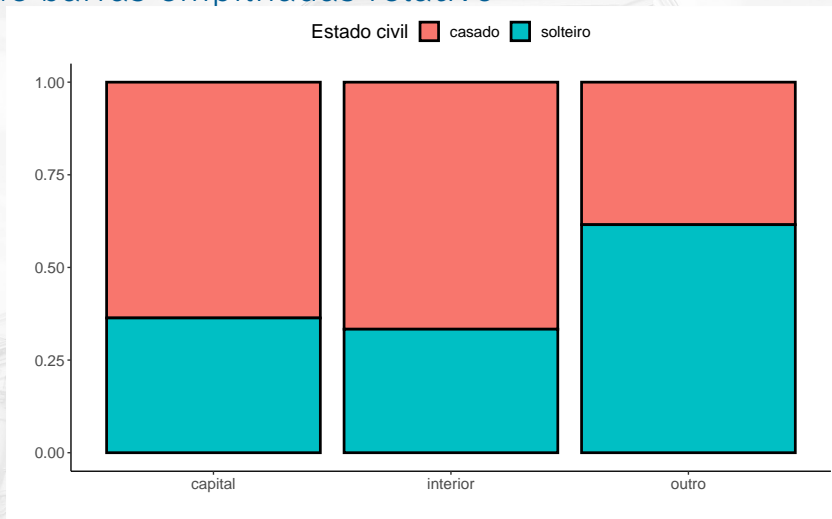


Figura 9. Gráfico de barras empilhadas relativo.

Medidas de associação para variáveis qualitativas

- ▶ Existem **medidas** que visam quantificar o **grau de associação** entre variáveis qualitativas.
- ▶ Uma dessas medidas é chamada de **Qui-quadrado**.
- ▶ Esta medida compara as **frequências observadas** em uma tabela de dupla entrada com as **frequências esperadas** caso não houvesse associação.
- ▶ Para obter a tabela de valores esperados basta, para cada casela, obter o produto entre o total da respectiva linha pelo total da respectiva coluna e dividir pelo total geral.

Medidas de associação para variáveis qualitativas

- ▶ O qui-quadrado é dado por:

$$Q = \sum_{i=1}^r \sum_{j=1}^s \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- ▶ Quanto mais próximo de 0, menor a evidência de associação.
- ▶ Como o valor é irrestrito, existem variações desta quantidade que visam ter os limites definidos.

Medidas de associação para variáveis qualitativas

Tabela 6. Valores observados.

	capital	interior	outro	Total
casado	7	8	5	20
solteiro	4	4	8	16
Total	11	12	13	36

Tabela 7. Valores esperados.

	capital	interior	outro	Total
casado	6.11	6.67	7.22	20
solteiro	4.89	5.33	5.78	16
Total	11.00	12.00	13.00	36

Medidas de associação para variáveis qualitativas

Tabela 8. $\frac{(o-e)^2}{e}$.

	capital	interior	outro
casado	0.13	0.27	0.68
solteiro	0.16	0.33	0.85

$$Q = \sum_{i=1}^r \sum_{j=1}^s \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 2,42$$



Análise bivariada para variáveis quantitativas

Análise bivariada para variáveis quantitativas

- ▶ Buscamos identificar **padrões** e **tendências** na análise das duas variáveis.
 - ▶ A medida que os valores de uma variável aumentam, a outra reduz?
 - ▶ A medida que os valores de uma variável aumentam, a outra aumenta?
 - ▶ A medida que os valores de uma variável aumentam, a outra se mantém estável?
- ▶ As principais técnicas são o **coeficiente de correlação** e o **diagrama de dispersão**.
 - ▶ O coeficiente é uma métrica que avalia a associação linear entre um par de variáveis numéricas.
 - ▶ O diagrama é um gráfico de pares ordenados.

Coeficiente de correlação linear de Pearson

- ▶ Usado para determinar se existe **relação linear** entre variáveis quantitativas.
 - ▶ Assume valores entre -1 e 1.
 - ▶ Se o valor é maior 0, então existe uma associação linear **positiva**.
 - ▶ Se o valor é menor que 0, então existe uma associação linear **negativa**.
 - ▶ Se o valor é igual a 0, então **não existe** uma associação linear.
- ▶ **CORRELAÇÃO NÃO IMPLICA EM CAUSALIDADE.**
 - ▶ O fato de existir uma correlação linear, seja positiva ou negativa, não implica que uma variável possui real influência nos desfechos da outra.
 - ▶ Causalidade causa correlação, mas correlação não implica em causalidade.

Covariância e correlação

- ▶ A covariância entre duas variáveis Y_1 e Y_2 é dada por:

$$\text{Cov}(y_1, y_2) = \frac{1}{n-1} \sum_{i=1}^n (y_{1i} - \bar{y}_1) \cdot (y_{2i} - \bar{y}_2).$$

- ▶ A partir da covariância podemos obter a correlação, que padroniza a medida pelas variâncias, fazendo com que, independente das variáveis, sempre seja um valor entre -1 e 1.

$$r = \frac{\sum_{i=1}^n (y_{1i} - \bar{y}_1) \cdot (y_{2i} - \bar{y}_2)}{\sqrt{\sum_{i=1}^n (y_{1i} - \bar{y}_1)^2} \cdot \sqrt{\sum_{i=1}^n (y_{2i} - \bar{y}_2)^2}} = \frac{\text{Cov}(y_1, y_2)}{\sqrt{V(y_1) \cdot V(y_2)}}.$$

Outros tipos de correlação

- ▶ A correlação de Pearson não serve para descrever associações que não sejam lineares.
- ▶ Existem outros tipos de correlação que servem inclusive para variáveis de outros tipos.
- ▶ Alguns exemplos são:
 - ▶ Correlação de Spearman.
 - ▶ Correlação de Kendall.
 - ▶ Ponto-bisserial.

Diagrama de dispersão

- ▶ O **diagrama de dispersão** é a principal ferramenta para visualizar duas variáveis quantitativas.
- ▶ Em um eixo são representados os valores de uma variável.
- ▶ No outro eixo os valores de uma segunda variável.
- ▶ Os pares ordenados são representados por pontos.

Diagrama de dispersão

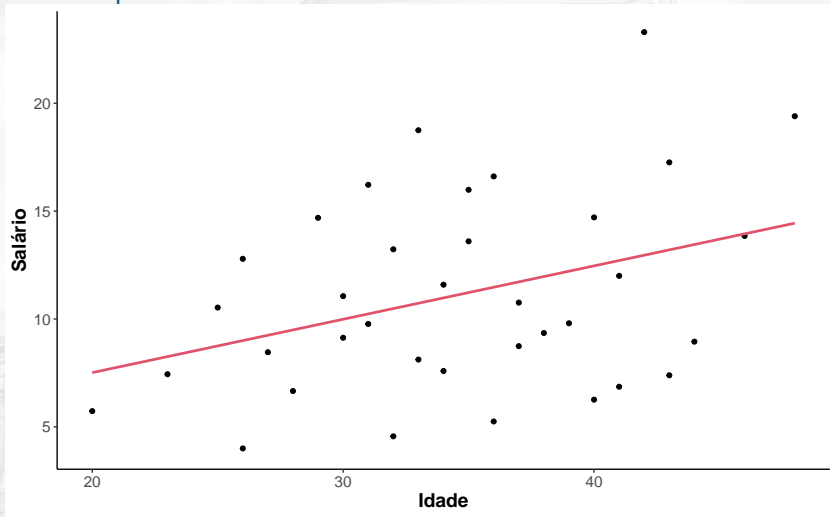


Figura 10. Diagrama de dispersão para o salário em função da idade.

Interpretação gráfica

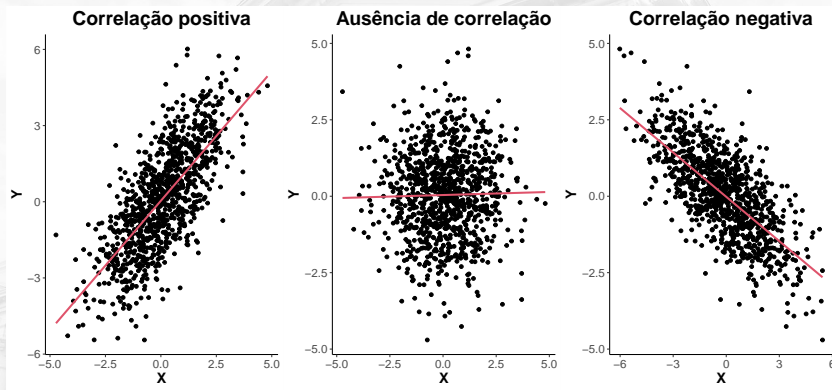


Figura 11. Avaliação de correlação usando diagramas de dispersão.

Covariância, correlação e diagrama de dispersão

Exemplo

- Considere as variáveis peso (Y_1) e altura (Y_2) de um conjunto de 10 indivíduos.

Y_1 : 60,09; 57,97; 54,12; 70,76; 59,74; 50,41; 58,19; 65,35; 71,18; 54,76

Y_2 : 1,54; 1,62; 1,52; 1,76; 1,63; 1,52; 1,65; 1,67; 1,66; 1,57

- $\overline{Y_1} = 60,26$; $\overline{Y_2} = 1,61$.
- $Var(Y_1) = 47,8$; $Var(Y_2) = 0,006$.
- Obtenha a covariância, coeficiente de correlação e o diagrama de dispersão.

Covariância, correlação e diagrama de dispersão

Exemplo

$$\text{Cov}(y_1, y_2) = \frac{1}{10 - 1} \{[(60,09 - 60,26) \cdot (1,54 - 1,61)] + \dots + [(57,76 - 60,26) \cdot (1,57 - 1,61)]\}$$

$$\text{Cov}(y_1, y_2) = 0,44$$

$$r = \frac{0,44}{\sqrt{47,8 \cdot 0,006}} = 0,82$$

Covariância, correlação e diagrama de dispersão

Exemplo - digrama de dispersão

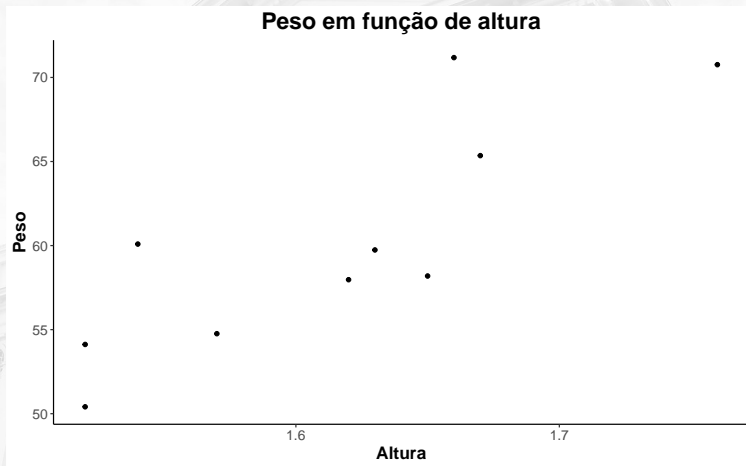


Figura 12. Diagrama de dispersão para peso e altura.

Covariância, correlação e diagrama de dispersão

Exemplo - digrama de dispersão

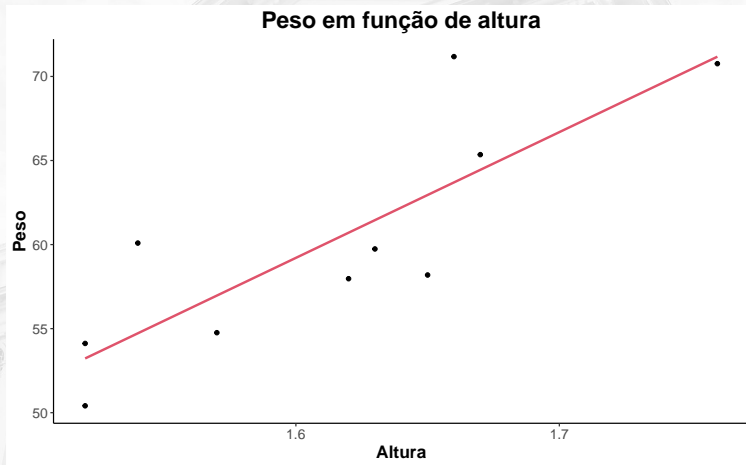
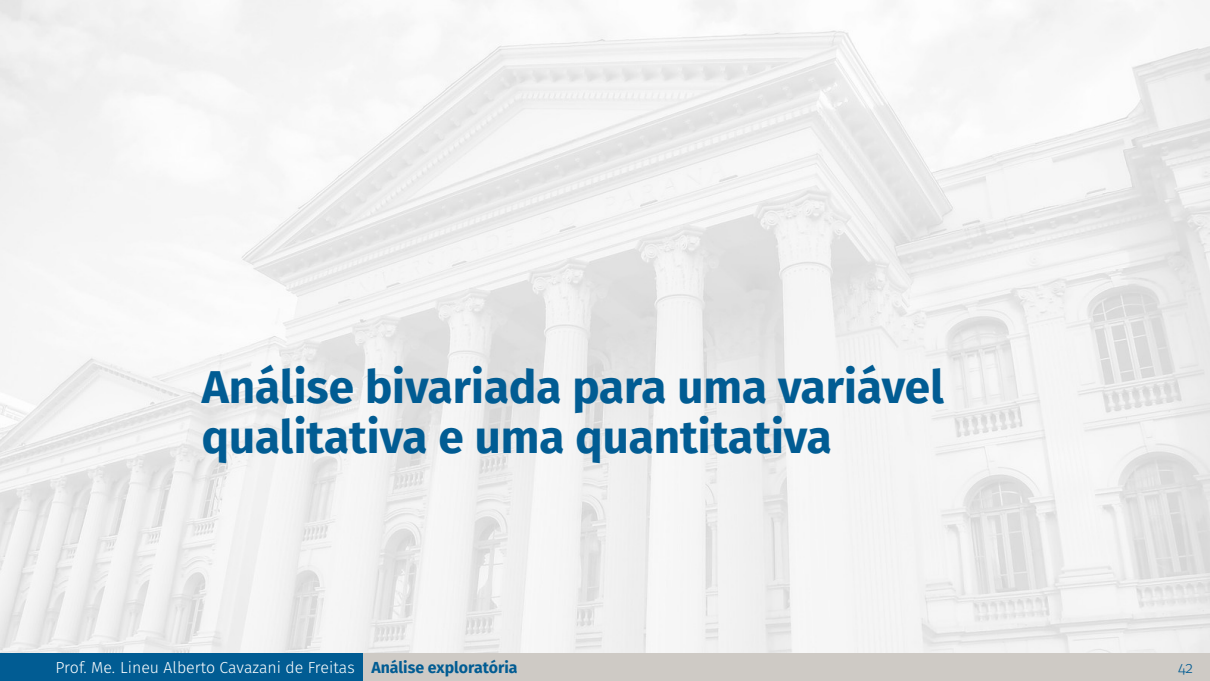


Figura 13. Diagrama de dispersão para peso e altura com linha de tendência linear.



Análise bivariada para uma variável qualitativa e uma quantitativa

Análise bivariada para uma variável qualitativa e uma quantitativa

- ▶ Neste caso estamos interessados em avaliar se os valores da variável numérica estão associados com os níveis da variável categórica.
- ▶ Podemos usar **medidas descritivas** para os valores dentro de cada um dos níveis da variável categórica.
- ▶ Para representar graficamente esta situação podemos criar um **box-plot** da variável numérica para cada nível do fator de interesse.

Tabela de medidas descritivas para níveis de um fator

Tabela 9. Medidas descritivas do salário em função da região.

Região	Média	Mediana	Desvio padrão
capital	11.46	9.77	5.48
interior	11.55	10.64	5.30
outro	10.45	9.80	3.15

Box-plot para níveis de um fator

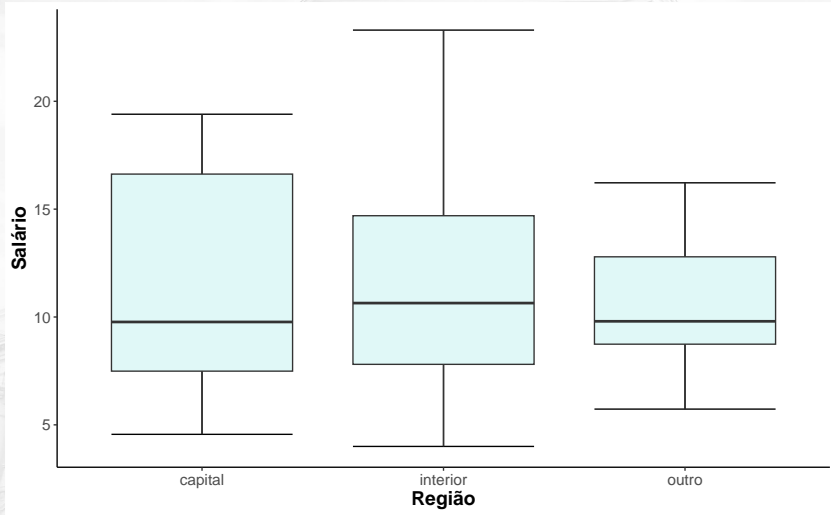


Figura 14. box-plots para o salário em função da região.



Outros tipos de gráficos e análises

Outros tipos de gráficos e análises

- ▶ Vimos as alternativas usuais para representação e análise de variáveis quantitativas e qualitativas.
- ▶ Contudo existem diversas situações particulares que exigem análises específicas.
- ▶ Algumas casos são: mapas, séries temporais, gráficos de perfil, nuvens de palavras.
- ▶ Também é possível trabalhar com gráficos que representam mais de duas variáveis ao mesmo tempo.
- ▶ Outra possibilidade é combinar gráficos.

O que foi visto:

- ▶ Análises bivariadas.
 - ▶ Qualitativa x qualitativa.
 - ▶ Quantitativa x quantitativa.
 - ▶ Quantitativa x qualitativa.

Próximos assuntos:

- ▶ Introdução à probabilidades.