INTRODUÇÃO À ANÁLISE MULTIVARIADA

1

Estrutura da Apresentação

I. Análise multivariada: conceitos e técnicas
II. Exame gráfico dos dados
III. Observações atípicas (*outliers*)
IV. Dados perdidos (*missing value*)
V. Suposições da análise multivariada
VI. Transformação de dados

2

Parte I

Análise multivariada: conceitos e técnicas

3

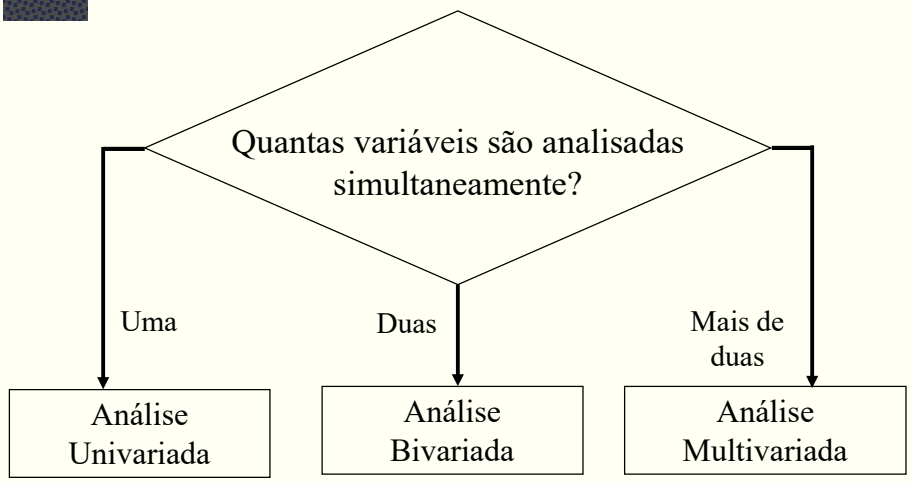
O que é análise multivariada?

"De um modo geral, refere-se a todos os métodos estatísticos que simultaneamente analisam múltiplas medidas sobre cada indivíduo ou objeto sob investigação. Qualquer análise simultânea de mais de duas variáveis de certo modo pode ser considerada análise multivariada."

(HAIR et al., 2005, p.26)

4

Nº de Variáveis X Análise



5

Tipos de Variáveis

- **Variáveis Quantitativas**
  - Variáveis Discretas
  - Variáveis Contínuas
- **Variáveis Qualitativas**
  - Variáveis Nominais
  - Variáveis Ordinais

6

Principais Técnicas de Análise Multivariada

- **Técnicas de Dependência:**
  - ✓ Regressão múltipla
  - ✓ Análise discriminante
  - ✓ Regressão logística
- **Técnicas de Interdependência:**
  - ✓ Análise fatorial
  - ✓ Análise de Cluster
  - ✓ MDS

7

Regressão múltipla

- Sua ideia-chave é a **dependência estatística** de uma variável em relação a duas ou mais variáveis independentes.
- Seus principais objetivos podem ser descritos como:
  - Encontrar a relação causal entre as variáveis.
  - Estimar os valores da variável dependente a partir dos valores conhecidos ou fixados das variáveis independentes.

8

Análise discriminante

- A variável dependente é qualitativa, podendo ser dicotômica (sim-não) ou multicotômica (alto-médio-baixo), e as variáveis independentes podem ser quantitativa ou qualitativa.
- Esta técnica estatística auxilia na identificação de quais variáveis conseguem diferenciar grupos ou categorias.

9

Regressão logística

- Técnica de análise multivariada que permite estabelecer a probabilidade de ocorrência de determinado evento para situações em que a variável dependente é qualitativa e de natureza dicotômica.
- Pode ser utilizada mesmo quando alguns dos pressupostos da análise discriminante não forem atendidos.

10

Análise fatorial

- É uma técnica multivariada de interdependência em que todas as variáveis são simultaneamente consideradas.
- Cada variável é relacionada com as demais, a fim de estudar as inter-relações existentes entre elas, buscando a redução ou sumarização dos dados.

11

Análise de Cluster

- É o nome dado ao grupo de técnicas multivariadas cuja finalidade primária é agregar objetos com base nas características que eles possuem.
- O objetivo é classificar uma amostra de indivíduos ou objetos em um pequeno número de grupos mutuamente excludentes, com base nas similaridades entre eles.

12

MDS

- O **Escalonamento Multidimensional (MDS)** é um procedimento que permite determinar a imagem relativa percebida de um conjunto de objetos, transformando os julgamentos de similaridade ou preferência em distâncias representadas no espaço multidimensional.

13

O Truque!

"O truque na estatística multivariada, se existe, não está nos cálculos, fácil e rapidamente feitos num computador com *software* adequado instalado. O truque consiste em escolher o método apropriado ao tipo de dados, usá-lo corretamente, saber interpretar os resultados e retirar deles as conclusões corretas."

(Reis, 2001, p.11)

14

Parte II

Exame gráfico dos dados

15

UMA PALAVRA DE ADVERTÊNCIA !

Se o pesquisador confia cegamente nessas técnicas para encontrar as respostas de suas questões sem ao menos atentar para as propriedades fundamentais dos **dados** que serão analisados, aumenta o risco de problemas sérios, tais como:

- ✓Uso indevido de técnicas
- ✓Violação de propriedades estatísticas
- ✓ Interpretação inadequada dos resultados

16

Examine seus dados...

- **Existe algum problema com meu banco de dados?**
- **Como solucionar esses problemas?**

17

Exemplo de dados

- Com intuito de exemplificar, no programa SPSS, temas abordados nesse capítulo, foi utilizado uma banco de dados que se encontra disponível em arquivo (**DemonstContEmpr.sav**).
- Esses dados foram retirados de demonstrações contábeis de empresas brasileiras.

18

Estatística Descritiva

- A Estatística descritiva está voltada para organizar, resumir e descrever os aspectos importantes de um banco de dados.
- Sintetizar os dados pode levar a perda de informações originais. Contudo, esta perda é pequena quando comparada ao ganho que se obtém com as interpretações que são proporcionadas.

19

Passos no SPSS
(Estatística descritiva das variáveis quantitativas)

1) *Analyze*
2) *Descriptive Statistics*
3) *Descriptives...*
4) *Variable(s)* (selecionar variáveis quantitativas)
5) *Options*... (selecionar opções desejadas)
6) *OK*

20

Relatório do SPSS
(Estatística descritiva das variáveis quantitativas)

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Patrimônio Líquido	100	33875	111110	71245,90	15312,14
Ativo Circulante	100	14575	60950	35311,25	10213,83
Passivo Circulante	100	12075	79350	50249,25	12942,80
Ativo Permanente	100	56425	152500	106094,25	24257,34
Ativo R. L. P.	100	1668	45036	19715,76	9971,79
Passivo E. L. P.	100	0	59658	34376,70	12916,70
LL em porcentagem	100	– 0,1173	0,0965	1,70E-02	3,13887E-02
Valid N (listwise)	100				

21

Onde:

N – Número de observações de cada variável.

Minimum – Corresponde ao menor valor encontrado para cada variável.

Maximum – Corresponde ao maior valor encontrado para cada variável.

Mean – Média aritmética não ponderada de cada variável.

Std. Deviation – Desvio-padrão de cada variável.

22

Média aritmética não ponderada

- A média é definida como a soma das observações dividida pelo número de observações.
- Se tivermos, por exemplo, **n** valores, temos:

$$Média = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

23

Desvio-Padrão

- É uma medida de dispersão.
- É a raiz quadrada da variância.
- Variância é definida como a média dos desvios ao quadrado em relação à média da distribuição

24

Como calcular a variância?

- Para uma amostra:

$$S^2 = \frac{\sum (x - \overline{X})^2}{n - 1}$$

- Para uma população finita:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

25

Passos no SPSS
(Estatística descritiva das variáveis qualitativas)

1) *Analyze*

2) *Descriptive Statistics*

3) *Frequencies...*

4) *Variable(s)* (selecionar variáveis qualitativas)

5) *Statistics...* (selecionar opções desejadas)

6) *OK*

26

Relatório do SPSS
(Estatística descritiva das variáveis qualitativas)

Tipo de SA

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Capital Aberto	60	60,0	60,0	60,0
	Capital Fechado	40	40,0	40,0	100,0
	Total	100	100,0	100,0	

Tamanho

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Pequena	34	34,0	34,0	34,0
	Média	32	32,0	32,0	66,0
	Grande	34	34,0	34,0	100,0
	Total	100	100,0	100,0	

27

Exame gráfico dos dados

- Examine a forma da distribuição da variável
- Examine a relação entre variáveis
- Examine as diferenças de grupos

28

Forma da distribuição

- Construindo um **histograma** é possível representar a frequência de ocorrências dentro de categorias de dados.
- Para avaliar normalidade, pode-se sobrepor à distribuição uma **curva normal**.
- O diagrama **ramo-e-folhas** é uma variante do histograma.

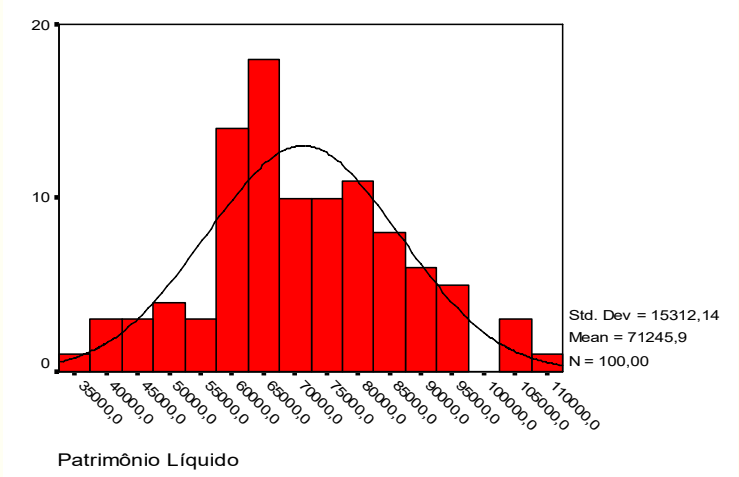
29

Passos no SPSS
(Histograma)

1) *Graphs*
2) *Histogram...*
3) *Variable* (selecionar a variável desejada)
4) *Display normal curve* (selecionar)
5) *Titles* (para definir título do gráfico)
6) *OK*

30

Relatório do SPSS
(Histograma)



31

Medidas

Forma

"É normal?"

32

Tipos principais de medidas

Assimetria

Curtose

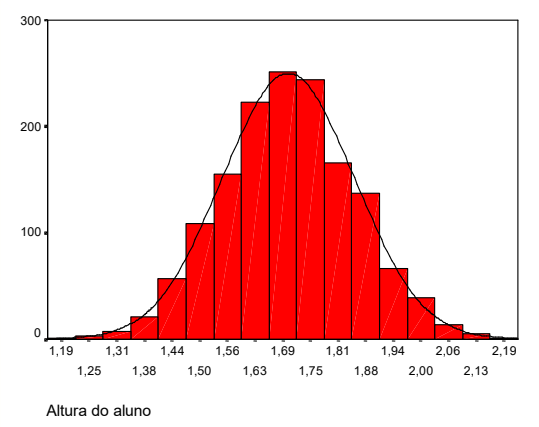
33

Assimetria

Analisa a concentração das distribuições de freqüência em torno do eixo

34

Afastamento ao eixo de simetria



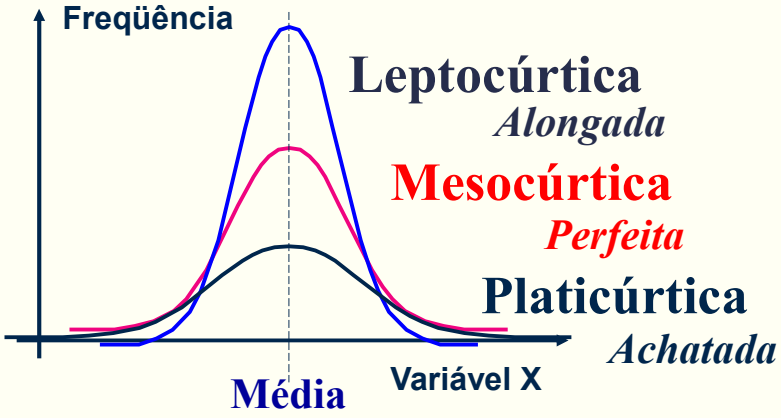
35

Curtose

Analisa o *achatamento* da curva

36

Analisando o achatamento



37

Calculando a curtose

$$K = \frac{Q_3 - Q_1}{2.(P_{90} - P_{10})}$$

k=0,263: distribuição mesocúrtica

distribuição nem chata nem delgada.

k > 0,263: distribuição leptocúrtica

distribuição delgada

k<0,263: distribuição platicúrtica

distribuição achatada

- $Q_3 = 3º$ quartil
- $Q_1 = 1º$ quartil
- $P_{90} = 90º$ percentil
- $P_{10} = 10º$ percentil

38

Calculando a Curtose

Descriptive Statistics

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Patrimônio Líquido	100	71245,90	15312,136	,194	,241	,092	,478
Valid N (listwise)	100						

Se o Erro Padronizado da Curtose for multiplicado por 3 e seu produto for maior que o valor absoluto da curtose, então a variável não terá problemas de curtose em sua análise.

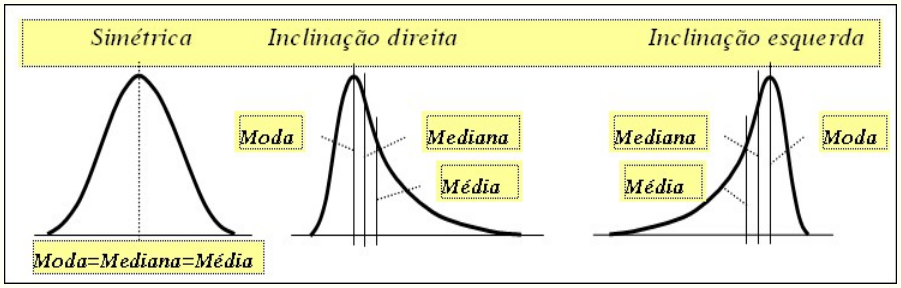
39

INCLINAÇÃO

Simétrica
Média=md=moda

Assimétrica Positiva
Média>md>moda

Assimétrica Negativa
Média<md<moda



40

- Na distribuição simétrica de frequências os valores de média, mediana e moda coincidem.
- As outras duas distribuições não são simétricas, e as medidas de tendência central têm posições relativas diferentes entre si, antecipando a forma da distribuição de frequências da amostra ou variável

41

Calculando a assimetria

Coeficiente de Pearson:

$$AS = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$$

42

Calculando a Assimetria

Descriptive Statistics

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Patrimônio Líquido	100	71245,90	15312,136	,194	,241	,092	,478
Valid N (listwise)	100						

Se o Erro Padronizado da Assimetria for multiplicado por 3 e seu produto for maior que o valor absoluto da assimetria, então a variável em questão é simétrica. Caso contrário é assimétrica.

43

Passos no SPSS
(Diagrama ramo-e-folhas)

1) *Analyze*
2) *Descriptive Statistics*
3) *Explore...*
4) *Dependent List* (Patrimônio Líquido – PL)
5) *Statistics...* (selecionar opções desejadas)
6) *Plots...* (selecionar *Stem-and-leaf*)
7) *OK*

44

Relatório do SPSS
(Diagrama ramo-e-folhas)

Patrimônio Líquido (Stem-and-Leaf Plot)

Frequency	Stem & Leaf
1,00	3 . 3
1,00	3 . 9
3,00	4 . 024
2,00	4 . 67
5,00	5 . 00114
3,00	5 . 668
19,00	6 . 0000000000222333333
19,00	6 . 5555555566667777799
9,00	7 . 011333444
11,00	7 . 55778889999
10,00	8 . 1111222244
5,00	8 . 56999
6,00	9 . 002334
2,00	9 . 66
3,00	10 . 555

45

Relação entre variáveis

- O método mais popular para examinar relações bivariadas é o **diagrama de dispersão**.
- Uma forte organização de pontos ao longo de uma linha reta caracteriza uma **relação linear**.
- Um formato particularmente adequado a técnicas multivariadas é a **matriz de dispersão**.

46

Passos no SPSS
(Matriz de dispersão)

1) *Graphs*

2) *Scatter...*

3) *Matrix* (selecionar)

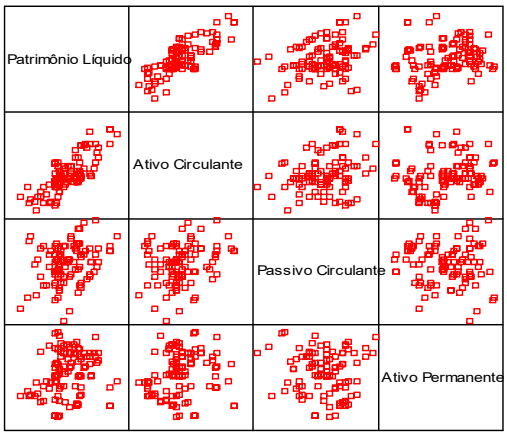
4) *Define*

5) *Matrix Variables* (Selecionar as variáveis PL, AC, PC e AP)

6) *OK*

47

Relatório do SPSS
(Matriz de dispersão)



48

Diferenças de grupos

- É preciso compreender como os valores estão distribuídos em cada grupo e se há diferenças suficientes para suportar significância estatística.
- Também é importante identificar observações ***outliers***.
- O método usado para essa tarefa é o **gráfico de caixas** (ou diagrama de extremos-e-quartis).

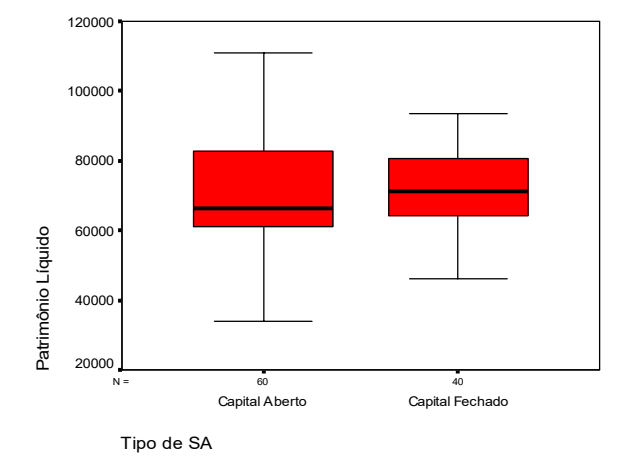
49

Passos no SPSS
(Gráfico de caixas)

1) *Graphs*
2) *Boxplot...*
3) *Simple* (selecionar)
4) *Summaries for groups of cases* (selecionar)
5) *Define*
6) *Variable* (Patrimônio Líquido – PL)
7) *Category Axix* (Tipo de S.A.)
8) *OK*

50

Relatório do SPSS
(Gráfico de caixas)



51

Parte III

Observações atípicas (*outliers*)

52

Observações atípicas (*outliers*)

São observações com uma combinação única de características identificáveis como sendo notavelmente diferentes das outras observações.

Não podem ser categoricamente caracterizadas como benéficas ou problemáticas.

É importante averiguar seu tipo de influência.

53

Classes de observações atípicas (*outliers*)

1º Erro de procedimento
(erro na entrada de dados ou uma falha na codificação)

2º Resultado de um evento extraordinário detectável

3º Observação extraordinária inexplicável

4º Observações com valores possíveis, mas com combinação extraordinária entre as variáveis.

54

Identificação de observações atípicas (*outliers*)

- ***Detecção Univariada*** – Casos que estão fora dos intervalos da distribuição, sendo que os principais passos deste procedimento são os seguintes:
  - ✓ Padronizar a variável para ter média 0 (zero) e desvio-padrão 1 (um).
  - ✓ Em pequenas amostras (n ≤ 80) ***outlier*** apresenta *score* ≥ 2,5.
  - ✓ Em grandes amostras ***outlier*** apresenta *score* ≥ 3,0.

55

Identificação de observações atípicas (*outliers*)

- ***Detecção Bivariada*** – Casos que estão fora do intervalo das outras observações, percebidos como pontos isolados no diagrama de dispersão (visualização gráfica).
- ***Detecção Multivariada*** – Casos com as maiores distâncias no espaço multidimensional de cada observação em relação ao centro médio das observações (visualização gráfica).

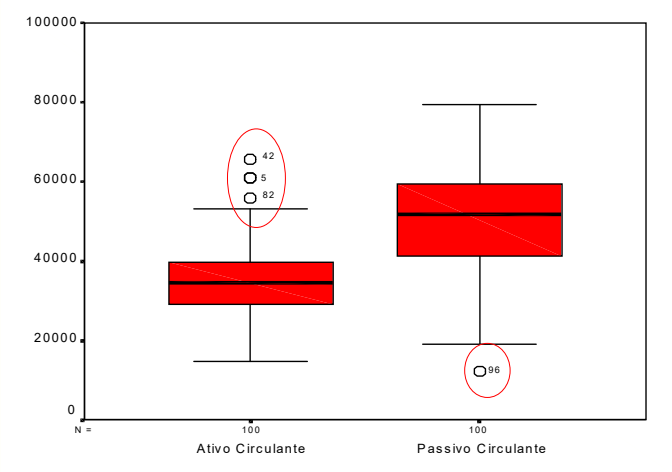
56

Passos no SPSS
(*Outliers*: detecção univariada)

1) *Graphs*
2) *Boxplot...*
3) *Simple* (selecionar)
4) *Summaries of separate variable* (selecionar)
5) *Define*
6) *Variable* (selecionar variáveis AC e PC)
7) *OK*

57

Relatório do SPSS
(*Outliers*: detecção univariada)



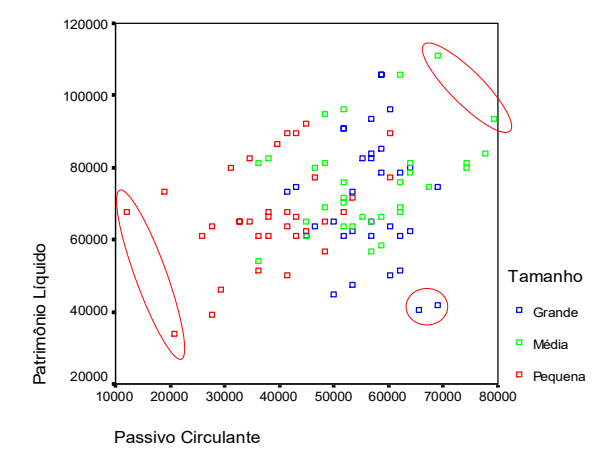
58

Passos no SPSS
(*Outliers*: detecção bivariada)

1) *Graphs*
2) *Scattert...*
3) *Simple*
4) *Y Axis* (variável PL)
5) *X Axis* (variável PC)
6) *Set markers by* (variável Tamanho)
7) *OK*

59

Relatório do SPSS
(*Outliers*: detecção bivariada)



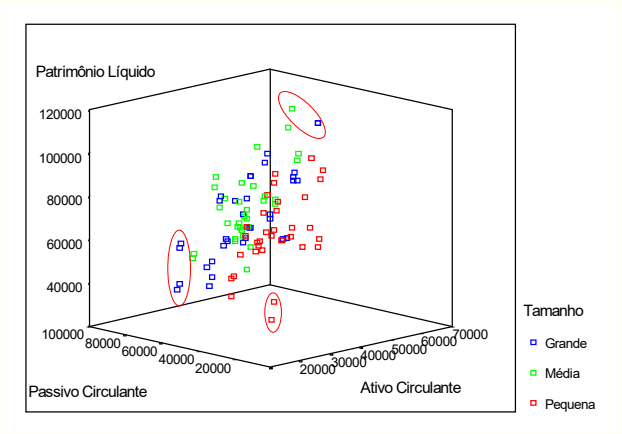
60

Passos no SPSS
(*Outliers*: detecção três dimensões)

1) *Graphs*
2) *Scatter...*
3) *3-D*
4) *Y Axis* (variável PL)
5) *X Axis* (variável PC)
6) *Z Axis* (variável AC)
7) *Set markers by* (variável Tamanho)
8) *OK*

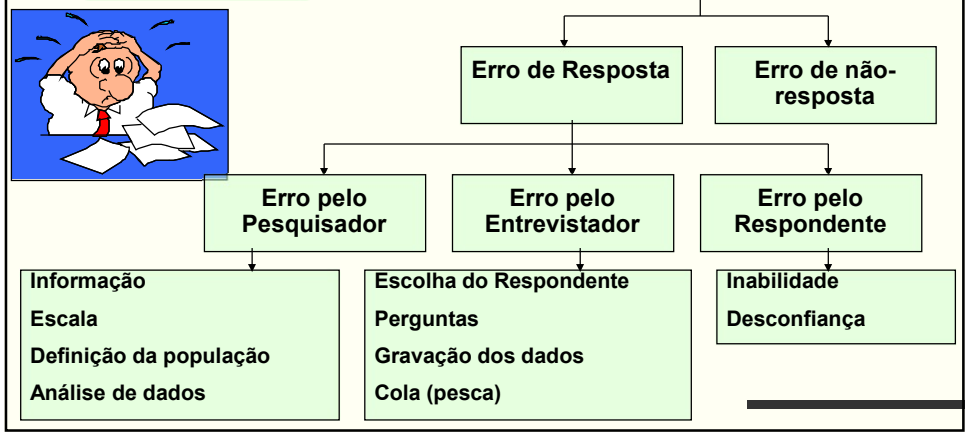
61

Relatório do SPSS
(*Outliers*: detecção três dimensões)



62

Fontes Potenciais de Erros em Pesquisa



63

Eliminação de observações atípicas (*outliers*)

Devem ser mantidas, a menos que exista prova demonstrável de que estão verdadeiramente fora do normal e que não são representativas de quaisquer observações na população.

Se as observações atípicas são eliminadas, o pesquisador corre o risco de melhorar a análise multivariada, mas limita sua generalidade.

64

Eliminação de observações atípicas (*outliers*)

Técnicas a serem implementadas:

Trimming ou Winsorizing (Hawkings, 1980)*

*Hawkings, D.M. Identification of Outliers. Chapman and Hall: London and New York, 1980.

65

Trimming: quando usar?

1. Erros de digitação
2. Erros de medida
3. Distribuição contaminada

66

Trimming: quando usar?

1. Remoção de outliers da base de dados

10, 15, 16, 17, 22

15, 16, 17

67

Winsorizing: quando usar?

Quando você acredita que está lidando com dados derivados de uma distribuição extremamente caudal.

68

Winsorizing: quando usar?

1. Substituição de outliers da base de dados

10, 15, 16, 17, 22

15, 15, 16, 17, 17

69

Eliminação de observações atípicas (*outliers*)

Trimming ou Winsorizing

Considerações Finais:

- Winsorizing ou trimming um ou dois dados que contam por menos de 5% de probabilidade e que não afetam a acuidade do p value.
- Quando for maior que 5%, pode ser necessário realizar outros testes de ajustes estatísticos.

70

Parte IV

Dados perdidos (*missing value*)

71

Dados Perdidos (*missing value*)

A preocupação primária do pesquisador é determinar as **razões** inerentes aos dados perdidos.

O pesquisador deve compreender os **processos** que conduzem os dados perdidos a fim de selecionar o curso de ação apropriado.

72

Padrão de Dados Perdidos

- Quando os dados perdidos ocorrem em um **padrão aleatório**, pode haver providências para minimizar seu efeito.
- As **ações corretivas** para dados perdidos somente poderão ser usadas se o processo de dados perdidos tiver um padrão aleatório, ou seja, quando o processo de dados perdidos for completamente ao acaso, pois, caso contrário, serão introduzidas tendências nos resultados.

73

Ações corretivas (remédios) para dados perdidos

- Incluir somente observações com dados completos
- Eliminar as observações e/ou variáveis problemáticas
- Utilizar métodos de atribuição

74

Incluir somente observações com dados completos

- ✓ Tratamento simples e direto.
- ✓ É conhecido como **abordagem de caso completo**.
- ✓ É mais apropriado quando a extensão de dados perdidos é pequena, a amostra é suficientemente grande e as relações nos dados são tão fortes que não podem ser afetadas por qualquer processo de dados perdidos.

75

Eliminar as observações e/ou variáveis problemáticas

- ✓ Pode-se descobrir que os dados perdidos estão concentrados em um pequeno subconjunto de casos e/ou variáveis, sendo que sua exclusão reduz substancialmente a extensão dos dados perdidos.
- ✓ O pesquisador sempre deve considerar os ganhos na eliminação de uma fonte de dados perdidos *versus* a eliminação de uma variável na análise multivariada.

76

Utilizar métodos de atribuição

- ✓ O método de atribuição é um processo de estimação de valores perdidos com base em valores válidos de outras variáveis e/ou observações na amostra.
- ✓ Principais métodos de atribuição:
  - Substituição por um caso
  - Substituição pela média
  - Atribuição por regressão
  - Substituição pela mediana
  - Substituição pela máxima esperança

77

Parte V

Suposições da análise multivariada

78

Suposições da análise multivariada

- A análise multivariada requer testes de suposições para as variáveis separadas e em conjunto.
- O foco agora será o exame de variáveis individuais.
- Nos capítulos posteriores serão abordados os métodos usados para avaliar as suposições inerentes às técnicas multivariadas específicas.

79

Suposições da análise multivariada

- As principais suposições são:
  - ✓ Normalidade
  - ✓ Homoscedasticidade
  - ✓ Linearidade
  - ✓ Multicolinearidade

80

Normalidade

- Os dados devem ter uma distribuição que seja correspondente a uma distribuição normal.
- Esta é a suposição mais comum na análise multivariada.
- Uma situação em que todas as variáveis exibem uma normalidade univariada ajuda a obter, apesar de não garantir, a normalidade multivariada.

81

Normalidade

- O teste diagnóstico de normalidade mais simples é uma verificação visual do histograma.
- *Kolmogorov-Smirnov*, *Jarque-Bera* e *Shapiro-Wilks* são exemplos de testes que tentam identificar se uma determinada variável possui distribuição normal.

82

Passos no SPSS
(Normalidade: teste Kolmogorov-Smirnov)

1) *Analyze*

2) *Nonparametric Tests*

3) *1-Sample K-S...*

4) *Test Variable List* (PL, PC, ARLP e LL)

5) *Test Distribution...* (selecionar opção Normal)

6) *Ok*

83

Relatório do SPSS
(Normalidade: teste Kolmogorov-Smirnov)

One-Sample Kolmogorov-Smirnov Test

		Patrimônio Líquido	Passivo Circulante	Ativo R. L. P.	LL em porcentagem
N		100	100	100	100
Normal Parameters [a,b]	Mean	71245,90	50249,25	19715,76	1,69501E-02
	Std. Deviation	15312,14	12942,80	9971,79	3,13887E-02
Most Extreme Differences	Absolute	0,101	0,086	0,095	0,164
	Positive	0,100	0,057	0,095	0,120
	Negative	– 0,101	– 0,086	– 0,068	– 0,164
Kolmogorov-Smirnov Z		1,012	0,862	0,945	1,636
Asymp. Sig. (2-tailed)		0,258	0,448	0,333	0,009

a. Test distribution is Normal.

b. Calculated from data.

84

Relatório do SPSS
(Normalidade: teste Kolmogorov-Smirnov)

Interpretação do relatório: Dado H_0 (a distribuição é normal) e H_1 (a distribuição não é normal), pode-se dizer que não existem evidências estatísticas para rejeitar H_0 (ao nível de significância de 5%) nas seguintes variáveis: Patrimônio Líquido, Passivo Circulante e Ativo R.L.P. (Sig. > 0,05), ou seja, nestes casos a distribuição é normal. Por outro lado, constatou-se que a variável LL em porcentagem não apresenta uma distribuição normal (Sig. < 0,05).

85

Homoscedasticidade

- A homoscedasticidade significa igualdade de variâncias entre as variáveis.
- Se as variáveis dependentes exibem iguais níveis de variância através da escala de previsão, a variância dos resíduos deve ser constante.
- Quando a variância dos termos de erro (ε) parece constante, diz-se que os dados são homoscedásticos.

86

Homoscedasticidade

- Para diagnosticar a homoscedasticidade podem ser utilizados testes estatísticos, tais como: Pesaran-Pesaran, Quandt-Goldfeld, Glejser e Park e Levene's Test para dados normais e não-normais.
- Os testes estatísticos relativos a esta suposição serão tratados no capítulo de regressão linear múltipla.

87

Transformações para obter normalidade

- Distribuição assimétrica positiva:

  Emprega-se o logaritmo das variáveis.

- Distribuição assimétrica negativa:

  Emprega-se a raiz quadrada das variáveis.

- Distribuição achatada:

  Emprega-se o inverso das variáveis (1/y e 1/x).

88

Transformações para obter homoscedasticidade

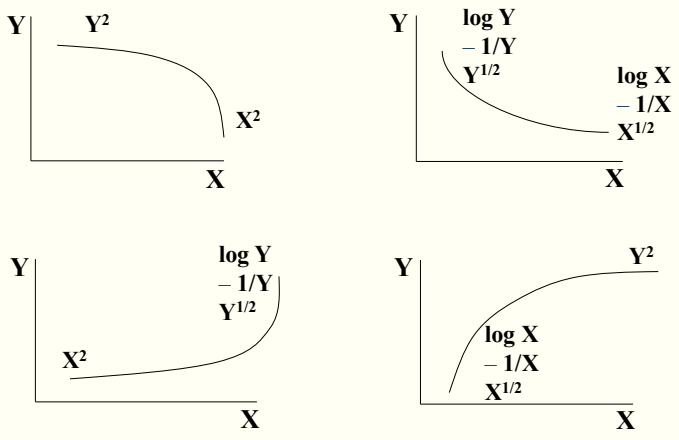
- Distribuição dos resíduos:

  Emprega-se logaritmo, raiz quadrada, inverso etc.

- A transformação deverá ser testada para verificar se o remédio utilizado é eficiente.

89

Transformações para obter linearidade



90

Passos no SPSS
(Transformação de dados)

1) *Transform*
2) *Compute...*
3) *Target Variable* (definir nome para a nova variável transformada)
4) *Numeric Expression* (inserir função matemática da transformação)
5) *Functions* (no caso de utilizar uma função de transformação do SPSS)
6) *Ok*

91

Passos no SPSS
(Transformação de dados)

Observação: É importante salientar que essa rotina do SPSS não tem como finalidade emitir relatórios, mas criar uma nova variável no arquivo de banco de dados que estiver sendo utilizado. Os números dessa nova variável corresponderão aos valores transformados, com base na função matemática que for empregada.

92