

Conteúdo da Aula

1. Estatística e Análise Exploratória de Dados.
2. Medidas de posição central.
3. Medidas de dispersão.
4. Medidas de ordenamento e forma.
5. Probabilidade.
6. Distribuições de probabilidades.
7. Amostragem e Estimação.
8. Correlação e Regressão.
9. Preparação de dados.
10. Análise Discriminante.



1

Capítulo

1

Estatística e análise exploratória de dados

2

1

Quatro objetivos do capítulo

- Entender as origens e os propósitos da Estatística
- Compreender a importância da análise de dados
- Classificar variáveis e casos
- Diferenciar variáveis qualitativas e quantitativas



3

Para entender a ...

Estatística

Status → *Estado*

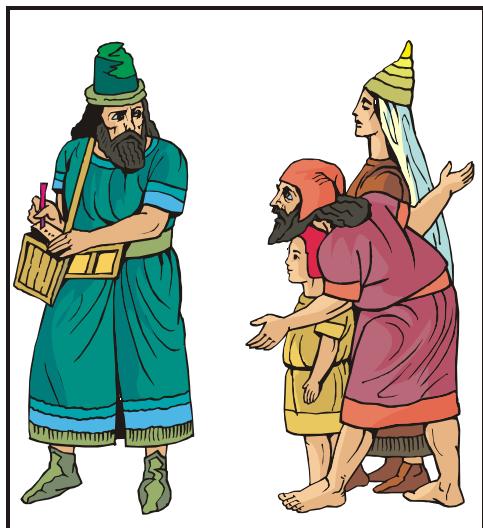
Poder público
Caracterização dos
dados



4

2

Uma origem controversa



**Estatística
para cobrar
IMPOSTOS**

5

E depois

**Facilitar a
análise
de
DADOS**



6

3

OBJETIVO DA ESTATÍSTICA

- O objetivo da *Estatística Descritiva* é organizar, resumir, analisar e interpretar observações disponíveis.
- O objetivo da *Inferência Estatística* é obter respostas corretas de questões específicas, atendendo a um determinado grau de acerto.

7

ORIGEM DOS DADOS

- A Estatística lida com dados, números dentro de um contexto.
- Entretanto, a utilização de estatística é mais do que trabalhar com números, pois embora a organização dos números e a construção de gráficos possa ser mecanizada com softwares e modelos, as idéias e os bons julgamentos, por enquanto, não podem ser automatizados.
- O analista deve ter o hábito de perguntar, por exemplo, o que mostram os resultados dentro de um determinado contexto? Quais as respostas que os dados podem dar a perguntas específicas?

8

Uma pergunta básica ...

**Como
entender os
DADOS?**



9

Uma representação didática ...

Dados

Estatística

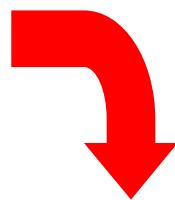
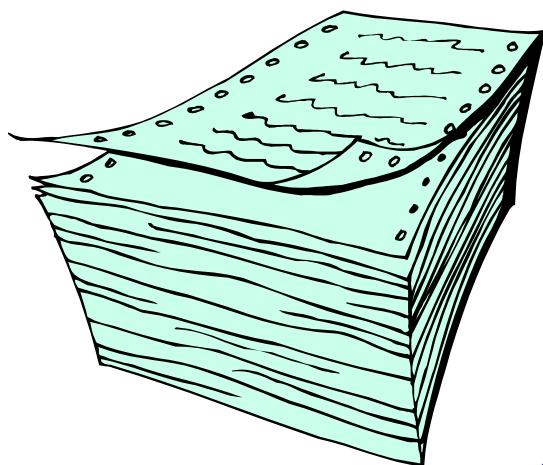
Informação

Decisão



10

Analisando ...



**Frequências
Média
Mediana
Desvio-Padrão
Coef. de Variação**

11

POPULAÇÃO TOTAL E PROPORÇÃO DA POPULAÇÃO POR SEXO, GRANDES GRUPOS DE IDADE E SITUAÇÃO DE DOMICÍLIO				
	1980	1990	1996	2000
População total	119.002.706	146.825.475	157.070.163	169.799.170
Por sexo (%)				
Homens	49,68	49,36	49,3	49,22
Mulheres	50,31	50,63	50,69	50,78
Por grandes grupos de idade (%)				
0-14 anos	38,2	34,72	31,54	29,6
15-64 anos	57,68	60,45	62,85	64,55
65 e mais	4,01	4,83	5,35	5,85
Por situação do domicílio (%)				
Urbana	67,59	75,59	78,36	81,25
Rural	32,41	24,41	21,64	18,75

12

- Da tabela podemos deduzir como essas proporções evoluíram com o passar do tempo, as tendências de crescimento, mas não é possível medir a força dessas tendências.
- Uma forma de analisá-las é medir a variação dos crescimentos durante os anos definidos nas colunas da tabela.

13

Taxa de crescimento - Média geométrica anual					
	1990/1980	1996/1990	2000/1996	2000/1980	2000/1990
População total	2,12 %	1,13 %	1,97 %	1,79 %	1,46 %
Por sexo					
Homens	-0,065 %	-0,020 %	-0,041 %	-0,047 %	-0,028 %
Mulheres	0,063 %	0,020 %	0,044 %	0,047 %	0,030 %
Por grandes grupos de idade					
0-14 anos	-0,95 %	-1,59 %	-1,57 %	-1,27 %	-1,58 %
15-64 anos	0,47 %	0,65 %	0,67 %	0,56 %	0,66 %
65 e mais	1,88 %	1,72 %	2,26 %	1,91 %	1,93 %
Por situação do domicílio					
Urbana	1,12 %	0,60 %	0,91 %	0,92 %	0,72 %
Rural	-2,79 %	-1,99 %	-3,52 %	-2,70 %	-2,60 %

14

ANÁLISE DOS RESULTADOS

- A população total continua crescendo; entretanto, a média geométrica da taxa de crescimento anual diminui, pois durante os anos 80 e 90 a média geométrica foi de 2,12% ao ano. Durante os anos 90 e 2000 foi de 1,46% ao ano.

15

- Quanto à classificação por sexo, a população de mulheres continua sendo maior que a dos homens, com tendência a aumentar essa diferença. De 1980 a 2000, a população de homens tem diminuído, com taxa média geométrica de -0,047% ao ano, e a população de mulheres tem aumentado, curiosamente, com taxa média geométrica de +0,047% ao ano.

16

- Quanto à classificação por grandes grupos de idade, entre 1980 e 2000, a população entre 0 e 14 anos diminuiu, com taxa média geométrica de -1,27% ao ano; a população entre 15 e 64 anos aumentou, com taxa média geométrica de 0,56% ao ano, e a população com mais de 65 anos aumentou, com taxa média geométrica 1,91% ao ano.

17

- Quanto à classificação por situação de domicílio, entre 1980 e 2000, a população com domicílio urbano aumentou com taxa média geométrica de crescimento positiva de 0,9% ao ano, e a população com domicílios rurais diminuiu, com taxa média geométrica de crescimento negativa de -2,7% ao ano.

18

PROJEÇÕES

- A análise dos resultados não se esgota nas poucas medidas que foram realizadas na planilha do Censo 2000. A partir dos resultados, surgem perguntas relacionadas com:
- As causas que geraram esses resultados. Por exemplo, enumerando as causas que vêm provocando a diminuição da população jovem e aumentando a população adulta, com destaque às pessoas com mais de 65 anos.

19

- As projeções futuras que podemos extrair desses resultados. Olhando para o futuro, também poderíamos enumerar as possíveis consequências dessas tendências. Um resultado rápido das consequências futuras pode ser resumido da seguinte forma: ***em longo prazo a população será mais velha e crescerá menos.***

20

DECISÕES

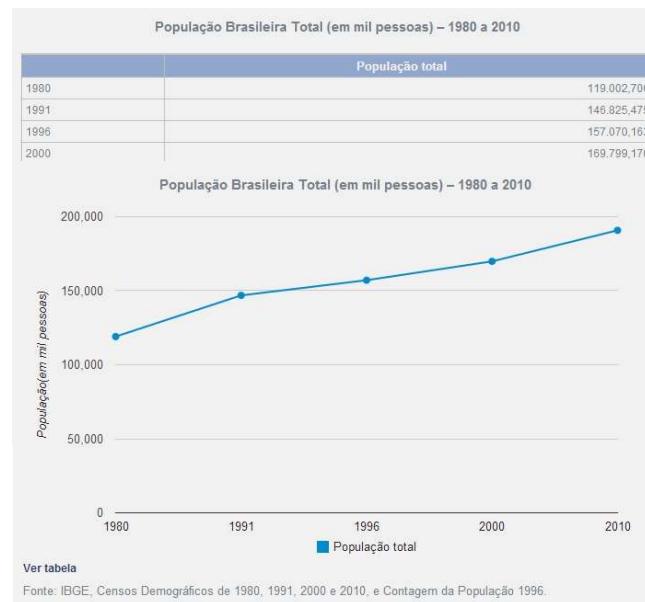
- Nas empresas é necessário prever as vendas, os estoques, os custos, o fluxo de caixa etc. para um determinado período, como é o orçamento anual do próximo ano.
- Na administração pública, se faz necessário prever o número de habitantes, a arrecadação, os custos dos serviços prestados etc.
 - *“O gestor público tem o dever de governar com olho no futuro, antecipando-se em dar respostas a problemas que explodirão depois de seu mandato.”*

21

- As tendências dos índices mostram riscos, oportunidades e desafios para as empresas.
- Enquanto o *cliente* dos serviços da administração pública é formado praticamente por todos os habitantes do país, o *cliente* das empresas privadas é uma parte desses habitantes.
- Por exemplo, o gerente de marketing necessita determinar o tamanho do mercado de seu novo produto, mas a população desse produto nem sempre coincide com a população do país.

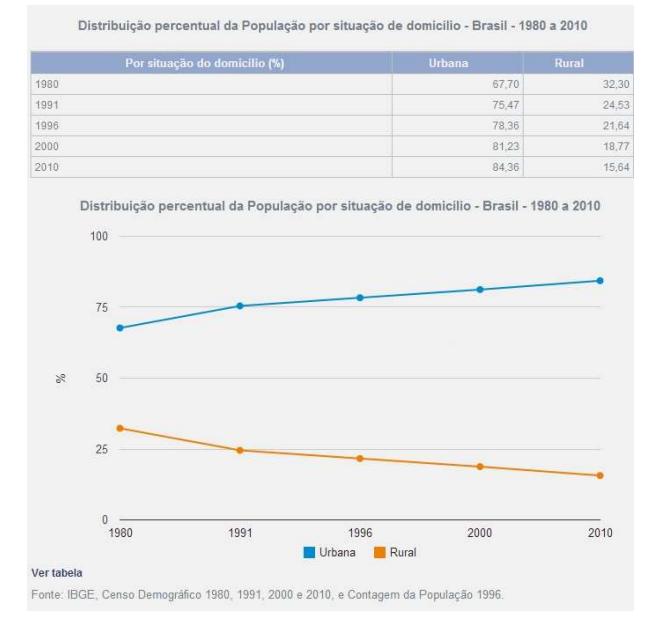
22

Dados IBGE 2013 – Os resultados confirmam?



23

Dados IBGE 2013 – Os resultados confirmam?



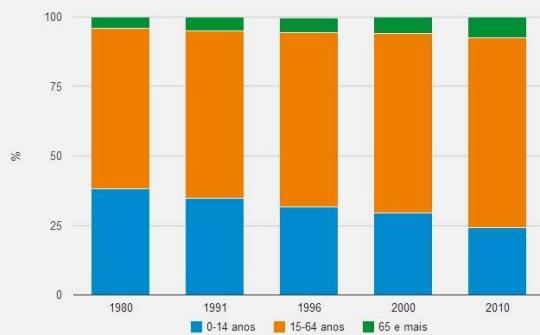
24

Dados IBGE 2013 – Os resultados confirmam?

Distribuição percentual da População por grandes grupos de idade Brasil - 1980 a 2010

Por grandes grupos de idade (%)	0-14 anos	15-64 anos	65 e mais
1980	38,20	57,68	4,01
1991	34,72	60,45	4,83
1996	31,54	62,85	5,35
2000	29,60	64,55	5,85
2010	24,08	68,54	7,38

Distribuição percentual da População por grandes grupos de idade Brasil - 1980 a 2010



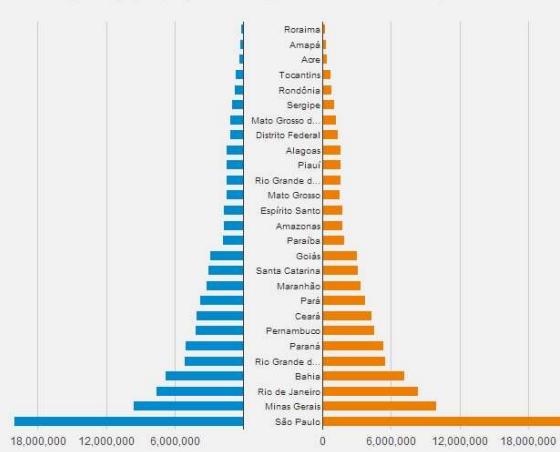
[Ver tabela](#)

Fonte: IBGE, Censo Demográfico de 1980, 1991, 2000 e 2010, e Contagem da População 1996.

25

Dados IBGE 2013 – Os resultados confirmam?

Distribuição da população por Sexo segundo Unidades da Federação - Brasil - 2010



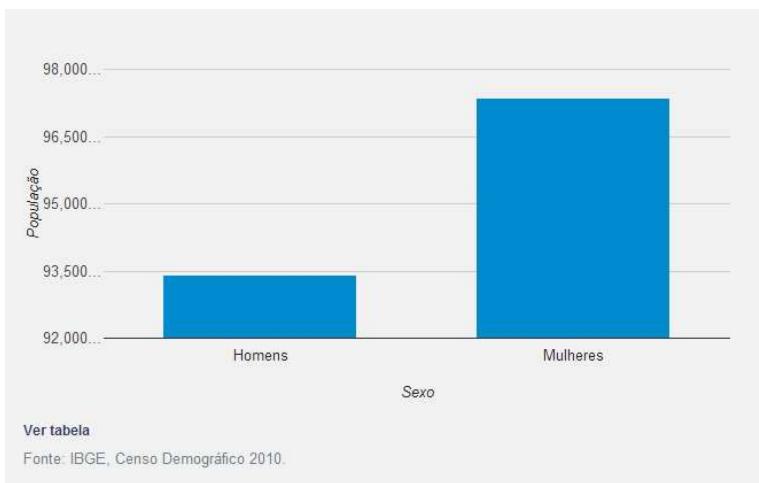
[Ver tabela](#)

Fonte: IBGE, Censo Demográfico 2010.

26

13

Dados IBGE 2013 – Os resultados confirmam?



27

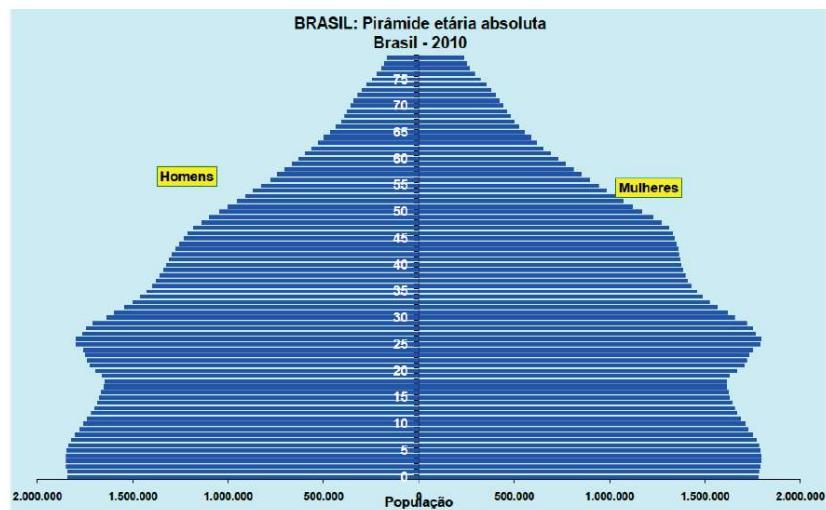
Dados IBGE 2013 – Os resultados confirmam?

Distribuição da população por Sexo segundo Unidades da Federação - Brasil - 2010

UF	Homens	Mulheres	Diferença	% Diferença
Acre	368.324	365.235	-3.089	-0,84%
Alagoas	1.511.767	1.608.727	96.960	6,41%
Amapá	335.135	334.391	-744	-0,22%
Amazonas	1.753.179	1.730.806	-22.373	-1,28%
Bahia	6.878.266	7.138.640	260.374	3,79%
Ceará	4.120.088	4.332.293	212.205	5,15%
Distrito Federal	1.228.880	1.341.280	112.400	9,15%
Espírito Santo	1.731.218	1.783.734	52.516	3,03%
Goiás	2.981.627	3.022.161	40.534	1,36%
Maranhão	3.261.515	3.313.274	51.759	1,59%
Mato Grosso	1.549.536	1.485.586	-63.950	-4,13%
Mato Grosso do Sul	1.219.928	1.219.928	0	0,00%
Minas Gerais	9.641.877	9.955.453	313.576	3,25%
Paraná	5.130.994	5.313.532	182.538	3,56%
Paraíba	1.824.379	1.824.379	0	0,00%
Pará	3.821.837	3.759.214	-62.623	-1,64%
Pernambuco	4.230.681	4.565.767	335.086	7,92%
Piauí	1.528.422	1.589.938	61.516	4,02%
Rio Grande do Norte	1.548.887	1.619.140	70.253	4,54%
Rio Grande do Sul	5.205.057	5.488.872	283.815	5,45%
Rio de Janeiro	7.625.679	8.364.250	738.571	9,69%
Rondônia	795.157	767.252	-27.905	-3,51%
Roraima	228.859	221.620	-7.239	-3,16%
Santa Catarina	3.100.360	3.148.076	47.716	1,54%
Sergipe	1.005.041	1.062.976	57.935	5,76%
São Paulo	20.077.873	21.184.326	1.106.453	5,51%
Tocantins	702.424	681.021	-21.403	-3,05%
TOTAL	93.406.990	97.221.871	3.814.881	4,08%

28

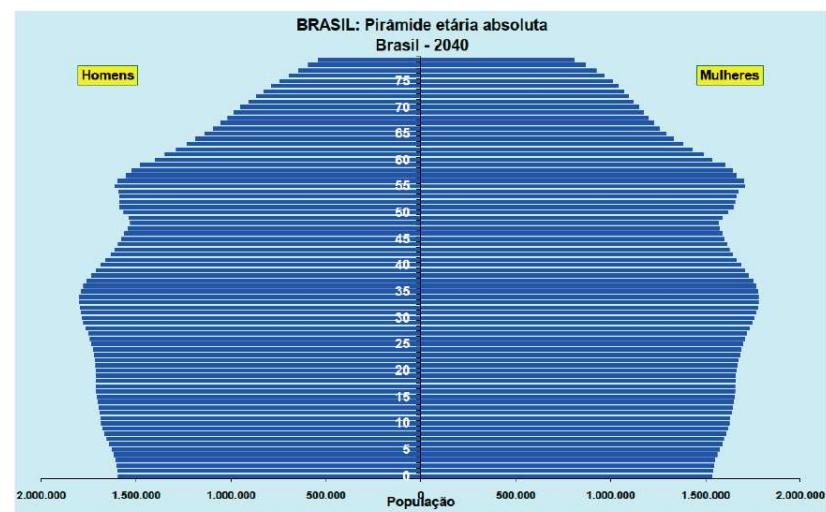
BRASIL: PIRÂMIDE POPULACIONAL - 2010



29

29

BRASIL: PIRÂMIDE POPULACIONAL - 2040



30

Assim ... é importante saber ...

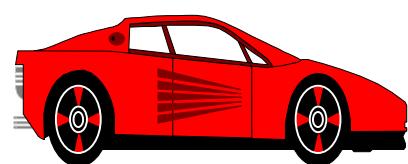
O que são dados?



31

Definição didática

- Conjunto de valores formados a partir do cruzamento de casos com variáveis



Casos	Variáveis				
	Cód.	Modelo	Ano	Cilindradas	Preço
1	Brasília	1978	4.000	\$ 4	
2	Fusquinha	1969	6.000	\$ 5	
3	Variant	1980	5.000	\$ 3	

32

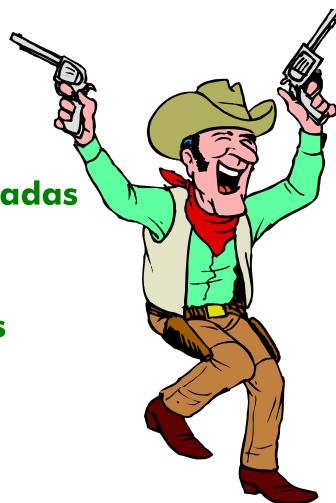
16

O que são variáveis?

- Especificações dos valores coletados

Variáveis

{ **Qualitativas**
não podem ser operadas algebricamente
Quantitativas
podem ser operadas algebricamente



33

Subdividindo as variáveis ...

- Qualitativas
 - Nominais
 - Ordinais
- Quantitativas
 - Contínuas
 - Discretas



34

DADOS E VARIÁVEIS

- Respostas de Pesquisas. **Quem aplica a pesquisa não tem nenhum controle intencional sobre os fatores que influenciam as respostas, por exemplo, a contagem de habitantes de um país, o cadastro dos clientes de um banco, a aceitação de um produto por um determinado tipo de consumidor etc.**

35

- Respostas de Experimentos. **Quem aplica o experimento tem controle intencional sobre os fatores que influenciam as respostas, por exemplo: o teste de estabilidade de produtos perecíveis frente a diferentes valores de temperatura e umidade; o desgaste de componentes de equipamentos mecânicos em condições especificadas e fora delas etc.**

36

- ***Unidade elementar*** é qualquer pessoa, objeto ou coisa que faça parte de uma população.
 - **Dado** é o resultado de investigação, cálculo ou pesquisa (Houaiss, 2005).
 - **Variável** é toda característica que pode assumir diversos valores conforme pessoa, objeto ou coisa.

37

CLASSIFICAÇÃO DOS DADOS

- Dados quantitativos. Referem-se a quantidades medidas numa escala numérica, em geral acompanhadas de alguma unidade de medida. Podem ser de dois tipos:
 - Dados discretos. Referem-se aos valores numéricos que assumem somente números inteiros positivos 0, 1, 2, 3 Os dados discretos resultam, em geral, de contagens. Por exemplo: a quantidade de vendas diárias de uma empresa, o número de filhos das famílias de uma região do país, o número de movimentos da conta corrente dos clientes de um banco comercial, a quantidade de peças defeituosas num lote de produção, o número de transações financeiras com erro de lançamentos, o número de acidentes nas estradas durante as férias anuais de verão etc.

38

19

- Dados contínuos. Referem-se aos valores numéricos que assumem qualquer valor do conjunto dos números reais. Os dados contínuos resultam, em geral, de medições que podem ter grande precisão. Por exemplo: o valor das vendas diárias de uma empresa, a estatura dos alunos da terceira série, o valor dos depósitos e retiradas da conta corrente dos clientes de um banco comercial, o consumo mensal de energia elétrica, o tempo necessário para realizar uma tarefa repetitiva, o tempo de espera para ser atendido num serviço de saúde pública etc.

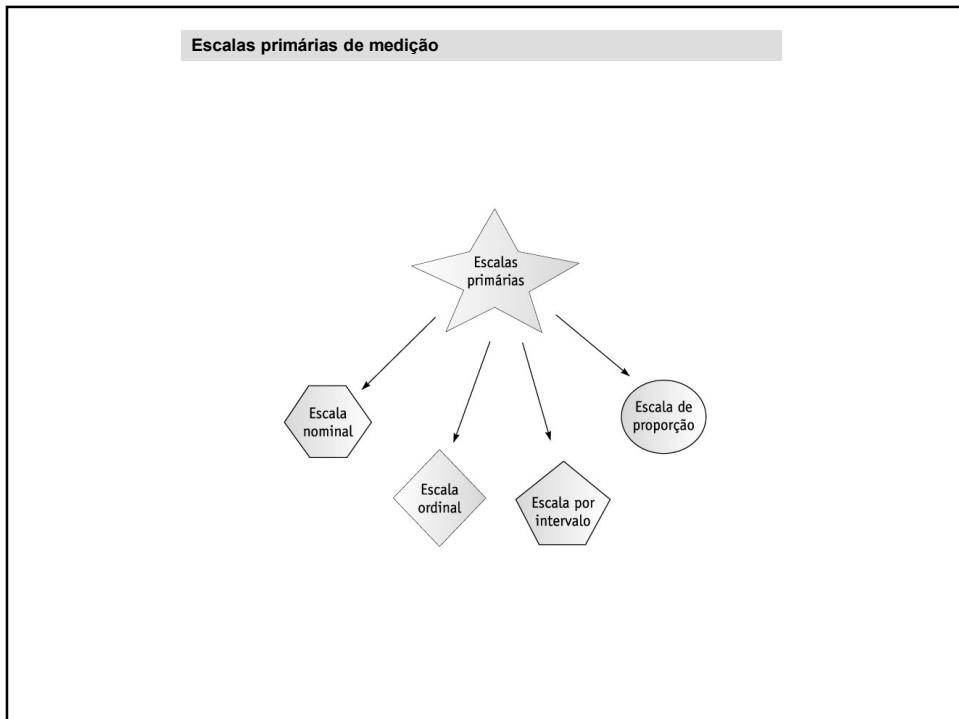
39

Dados qualitativos. Referem-se às observações não-numéricas, e são classificados em nominais e ordinais:

- Dados nominais. Esses dados não têm ordenamento nem hierarquia. Por exemplo, o sexo dos funcionários registrados no cadastro da empresa, o estado civil, o nome das empresas que têm ações negociadas na bolsa de valores, a cidade de residência do respondente etc.
- Dados ordinais. Esses dados são equivalentes aos nominais, porém incluindo uma ordem, uma hierarquia. Por exemplo, o cargo dos funcionários registrados no cadastro da empresa: presidente, diretor, gerente etc; a resposta a um questionário de pesquisa em que há uma escala para escolher: bom, regular e ruim; as posições das cinqüenta maiores empresas por vendas durante um ano: primeira, segunda etc.

40

20



41

ESCALA DE MEDIÇÃO

- Escala Nominal. Valores numéricos numa escala nominal apenas dão nome a uma categoria ou classe; os números são utilizados somente para diferenciar objetos, categorias ou nomes. Por exemplo, numa pesquisa de mercado realizada em estados brasileiros, a variável *estado de nascimento* do entrevistado foi codificada da seguinte forma: 1=Ceará, 2=Piauí, 3=Pernambuco, 4=Alagoas e 5=Bahia. Embora o código tenha transformado um nome num número, esse número não mantém todas as propriedades dos números, por exemplo, não se pode estabelecer relações do tipo $3>2$ ou $1+2=3$ ou $3-2=1$, como o leitor pode confirmar substituindo cada número pelo *estado* correspondente.

42

- Escala Ordinal. Valores numa escala ordinal dão nome e ordem a um objeto, categoria ou classe; os números são utilizados para diferenciar em ordem de superioridade, seguindo algum critério de hierarquia. Por exemplo, numa pesquisa, a variável *instrução do entrevistado* foi codificada assim: 1=Sem Instrução, 2=Ensino Fundamental, 3=Ensino Médio, 4=Ensino Superior, 5=Especialista, 6=Mestre e 7=Doutor. Nesse caso, na transformação de um nome num número, o número mantém algumas propriedades dos números, por exemplo, é possível estabelecer relações do tipo $3 > 2$ (o grau de instrução 3 é maior que o grau de instrução 2), porém não se pode estabelecer relações do tipo $2+3=5$, como o leitor pode confirmar substituindo cada número pelo grau de instrução correspondente.

43

- Escala de Intervalos. Valores numa escala de intervalos eliminam a limitação da escala ordinal estabelecendo intervalos iguais em que é possível ordenar as medições e, ao mesmo tempo, explicar quanto uma observação difere da outra. Por exemplo, o aumento da temperatura de ontem para hoje é de cinco graus, de 20 para 25 graus centígrados. Podemos dizer que hoje está mais quente do que ontem. Essa escala de medida tem uma unidade de medida, um zero arbitrário e a distância entre duas medições nessa escala tem um significado preciso. Outro exemplo de escala de intervalos são os tempos dos calendários, gregorianos e outros tipos.

44

- Escala Proporcional. **Valores numa escala proporcional eliminam a limitação da escala intervalar, estabelecendo um zero da própria categoria, denominado zero absoluto.** Por exemplo, peso zero claramente significa falta de peso, o peso de uma caixa de 86 kg é o dobro do de uma caixa de 43 kg, e trinta e três peças rejeitadas de um lote de produção representam o triplo do lote de produção com onze peças rejeitadas.
 - **O zero da escala de graus centígrado é o ponto de congelamento da água no nível do mar; entretanto, essa temperatura medida na escala de graus Fahrenheit é 32 graus F°.**

45

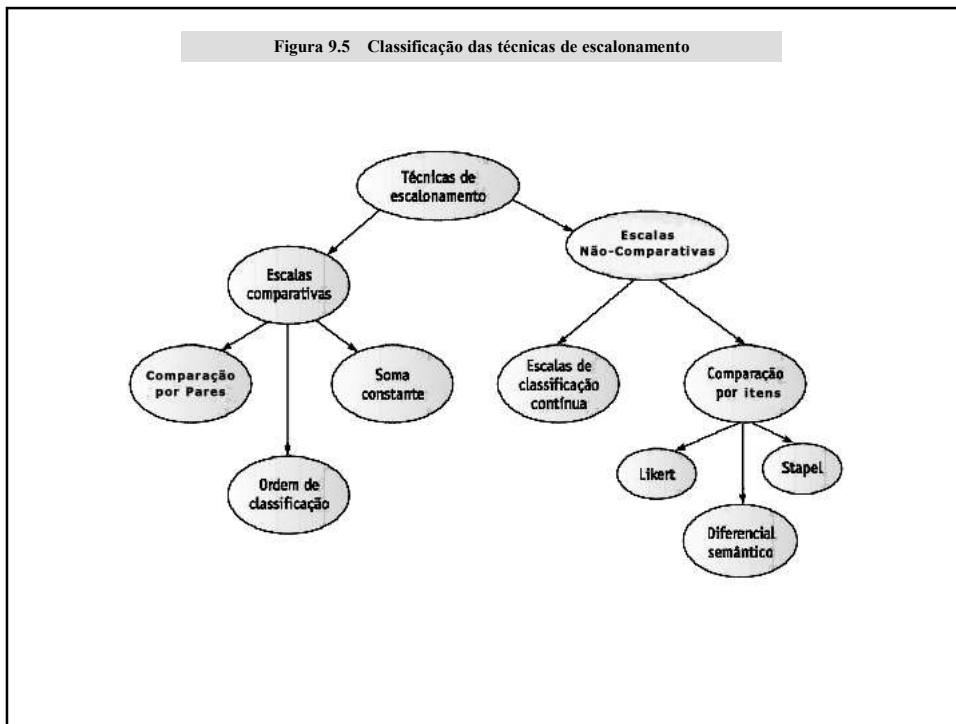
Escalas primárias de medição

Escala					
<i>Nominal</i>	Números atribuídos aos corredores	17	21	13	Chegada
<i>Ordinal</i>	Ordem de classificação dos vencedores	Terceiro lugar	Segundo lugar	Primeiro lugar	
<i>Por intervalo</i>	Classificação do desempenho numa escala de 0 a 100	74	90	97	
<i>De proporção</i>	Tempo de chegada, em segundos	16,1	14,0	13,2	

46

Escalas primárias de medição				
ESCALA PRIMÁRIA	CARACTERÍSTICAS BÁSICAS	EXEMPLOS COMUNS	EXEMPLOS DE MARKETING	ESTATÍSTICAS ADMISSÍVEIS
Nominal	Números identificam e classificam os objetos	Números do Seguro Social, numeração dos jogadores de futebol	Números das marcas, tipos de loja, classificação por sexo	Porcentagens, modo
Ordinal	Números indicam as posições relativas dos objetos, mas não a importância das diferenças entre eles	Classificações de qualidade, classificações das equipes em um torneio	Classificações de preferências, posição no mercado, classe social	Percentil, média
Por intervalo	As diferenças entre os objetos podem ser comparadas; ponto zero é arbitrário	Temperatura (Fahrenheit, Celsius)	Atitudes, opiniões, números de proporções	Amplitude, meio, desvio-padrão
De proporção	Ponto zero é fixo; os valores das proporções das escalas podem ser computados	Comprimento, peso	Idade, renda, custos, vendas, participação no mercado	Média geométrica (todas)

47



48

Figura 9.6 Escalonamento de comparação por pares

Instruções

Vamos apresentar dez pares de marcas de xampu. Para cada par, indique qual das marcas de xampu no par você preferiria para uso pessoal.

Forma de registro

	Jhirmack	Finesse	Vidal Sassoon	Head & Shoulders	Pert
Jhirmack		0	0	1	0
Finesse	1 ^a		0	1	0
Vidal Sassoon	1	1		1	0
Head & Shoulders	0	0	0		0
Pert	1	1	0	1	
Número de vezes preferida	3 ^b	2	0	4	1

^a1 em um box específico significa que a marca nessa coluna teve preferência sobre a marca na fileira correspondente do entrevistado. Um 0 significa que a marca da fileira teve preferência sobre a marca da coluna.

^bO número de vezes que uma marca teve preferência é obtido somando o 1 de cada coluna.

49

Figura 9.7 Escalonamento por ordem de classificação

Instruções

Classifique as várias marcas de creme dental em ordem de preferência. Comece escolhendo a marca que você mais gosta e atribua a ela o número 1. Em seguida, encontre sua segunda marca preferida e atribua a ela o número 2. Continue com esse procedimento até que tenha classificado todas as marcas de creme dental em ordem de preferência. A marca menos preferida deve receber uma classificação de 10.

Duas marcas não devem receber o mesmo número de classificação.

Os critérios de preferência são totalmente seus. Não existe resposta certa ou errada. Apenas tente ser consistente.

Marca	Ordem de classificação
1. Crest	_____
2. Colgate	_____
3. Aim	_____
4. Mentadent	_____
5. Macleans	_____
6. Ultra Brite	_____
7. Close Up	_____
8. Pepsodent	_____
9. Plus White	_____
10. Stripe	_____

50

Figura 9.8 Escalonamento de soma constante

Instruções

Abaixo estão oito atributos de um sabonete. Reparta 100 pontos entre os atributos, de modo que sua divisão reflete a importância relativa que você confere a cada atributo. Quanto mais pontos um atributo receber, mais importante será. Se um atributo não for importante, não lhe dê nenhum ponto. Se um atributo for duas vezes mais importante que outro, ele deverá receber duas vezes mais pontos.

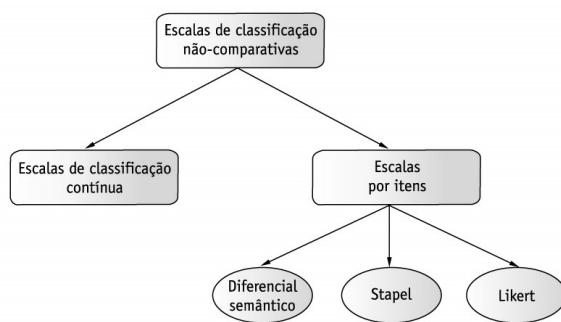
Forma

MÉDIA DE RESPOSTA DOS TRÊS SEGMENTOS

Atributo	Segmento I	Segmento II	Segmento III
1. Suavidade	8	2	4
2. Espuma	2	4	17
3. Redução	3	9	7
4. Preço	53	17	9
5. Fragrância	9	0	19
6. Embalagem	7	5	9
7. Hidratante	5	3	20
8. Poder de limpeza	13	60	15
Soma	100	100	100

51

Figura 10.3 Categorias de escalas de classificação não-comparativas



52

Escalas de Likert

- Concordo totalmente
- ...
- Discordo totalmente



53

Tabela 10.1 Escalas não-comparativas básicas

ESCALA	CARACTERÍSTICAS BÁSICAS	EXEMPLOS	VANTAGENS	DESVANTAGENS
Escala de classificação contínua	Colocar uma marca em uma linha contínua	Reação à propaganda na TV	Fácil de construir	A pontuação pode ser trabalhosa, a não ser que se use um computador
ESCALAS POR ITENS				
Escala de Likert	Grau de concordância em uma escala de 1 (discordo muito) a 5 (concordo muito)	Medição de atitudes	Fácil de construir, aplicar e compreender	Consumo mais tempo
Diferencial semântico	Escala de 7 pontos com rótulos opostos	Imagens de marca, produto e empresa	Versátil	Difícil de encontrar adjetivos opostos apropriados
Escala de Stapel	Escala unipolar de 10 pontos, -5 a +5, sem um ponto neutro (zero)	Medição de atitudes e imagens	Fácil de construir, aplicada pelo telefone	Confusa e difícil de aplicar

54

TIPOS DE VARIÁVEIS

As variáveis podem ser obtidas de duas formas.

- Séries temporais. As observações são dados de uma mesma variável em diferentes períodos de tempo, por exemplo, o valor do PIB anual de um país, a taxa mensal de desemprego numa região, as cotações diárias de uma ação, a rentabilidade mensal de uma empresa, a demanda de energia elétrica diária na região nordeste medida às dezoito horas etc.
- Corte transversal numa data ou período. Se na coleta dos dados não for considerada a sequência temporal, por exemplo, amostras da quantidade produzida e do preço médio dos produtos; ou das vendas e do investimento em propaganda; a média de apartamentos vendidos durante o último mês pelas primeiras dez imobiliárias da cidade; o número de operações fechadas por cinco ações numa determinada data etc.

55

Extraia informações para a ...

Mercearia
Melhor
Preço
Ltda.



56

28

Mercearia Melhor Preço	
▪ Variáveis	Classificação
Nome	Qualitativa
Código	Qualitativa
Estado	Qualitativa
Número de funcionários	Quantitativa
Qualidade do atendimento	Qualitativa
Faturamento	Quantitativa
Volume	Quantitativa

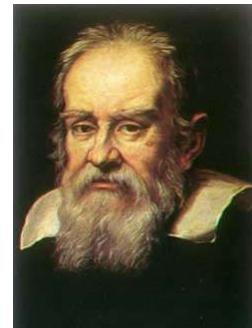
57



58

Para refletir ...

**“Meça o que for
possível ser medido;
o que não for, torne
mensurável.”**



Galileo Galilei
Séc. XVII

59

Três objetivos do capítulo

- Entender a importância das medidas
- Compreender as razões do uso de medidas de posição central
- Reconhecer as principais medidas de posição central, seus pontos fortes e fracos



60

ORDENAMENTO DE DADOS

- Em algumas situações, o objetivo é conhecer a posição de um determinado valor numérico em relação aos restantes valores da amostra.
- Por exemplo, qual a posição de um determinado candidato a *trainee* comparando seu QI com os QIs dos outros candidatos que concorrem? O QI desse candidato é baixo ou alto? Quantos candidatos têm QI maior que o candidato sob análise? Ou, quão maior é o QI do candidato?
- Para responder perguntas desse tipo, primeiro os valores da série de dados devem estar ordenados em ordem crescente ou decrescente. Depois deve-se estabelecer um critério que permita definir a posição de um determinado valor da série dentro da própria série de valores numéricos.

61

ORDENAMENTO DE DADOS

Exemplo 3.1

Ordenar de forma crescente os valores da amostra registrada na tabela.

31	38	19	27	24	42	32	18	43	15	39
----	----	----	----	----	----	----	----	----	----	----

Solução. Depois de ordenar de forma crescente os onze valores numéricos da amostra, a seguir são associados os números 1, 2, ..., 11 aos valores ordenados como mostra a tabela seguinte.

Amostra	15	18	19	24	27	31	32	38	39	42	43
Ordem	1	2	3	4	5	6	7	8	9	10	11

Agora, o valor 15 tem a posição 1, o 19 a posição 3 e o 43 a posição final 11.

62

- De forma geral, o Exemplo 3.1 mostra que os n valores numéricos de uma amostra ordenada de forma crescente foram associados à série dos números naturais 1, 2, 3, ... até n . Foi estabelecida uma relação de ordem entre os valores numéricos da amostra.

63

Exemplo 3.2

Determinar a *ordem* de cada valor da amostra seguinte:

27	32	64	65	58	62	59	54	29	30	26	48	47
46	43	38	29	32	35	37	31	43	45	42	37	36

Solução. Depois de ordenar os valores da amostra de forma crescente, foi associada a série de números 1, 2, ..., 26 aos valores como mostra a tabela.

Amostra	26	27	29	29	30	31	32	32	35	36	37	37	38
Ordem	1	2	3	4	5	6	7	8	9	10	11	12	13
Amostra	42	43	43	45	46	47	48	54	58	59	62	64	65
Ordem	14	15	16	17	18	19	20	21	22	23	24	25	26

64

- O procedimento de ordenamento crescente utilizado no Exemplo 3.2 foi o mesmo que o do Exemplo 3.1.
- Entretanto, no primeiro o trabalho manual foi facilitado pelo pequeno tamanho da amostra. No último exemplo, o ordenamento manual é menos eficiente, pois é mais trabalhoso e está sujeito a erro de seleção dos valores da amostra.
- O comando de classificação do Excel ajudará a ordenar séries de valores em ordem crescente ou decrescente, como mostra o Exemplo 3.3.

65

ORDENAMENTO com Excel

	A	B	C	D	E	F	G	H
1	Exemplo 3.3							
2								
3								
4								
5	27	27						
6	32	32						
7	64	64						
8	65	65						
9	58	58						
10	62	62						
11	59	59						
12	54	54						
13	29	29						
14	30	30						
15	26	26						
16	48	48						
17	47	47						
18	46	46						
19	43	43						
20	38	38						

Classificar

Classificar por: **Ordenada** Crescente Decrescente

Em seguida por: Crescente Decrescente

E depois por: Crescente Decrescente

Minha lista tem: Linha de cabeçalho Nenhuma linha

OK

	A	B	C
1	Exemplo 3.3		
2			
3			
4			
5	27	26	
6	32	27	
7	64	29	
8	65	29	
9	58	30	
10	62	31	
11	59	32	
12	54	32	
13	29	35	
14	30	36	
15	26	37	
16	48	37	
17	47	38	
18	46	42	
19	43	43	
20	38	43	

66

Algumas estatísticas

■ Medidas

Posição Central

Dispersão

*Ordenamento e
posição*

Forma



67

As EstatísticaS

■ Medidas úteis para a decisão

■ “Olhe para o centro” ...

■ Medidas de posição central

■ Mediana

■ Moda

■ Média ou Valor Esperado



68

34

Mediana

- Valor central de uma série ordenada de dados (Rol)

{3; 7; 9; 10; 4; 8; 2}

Ordenando no Rol

{2; 3; 4; 7; 8; 9; 10}

3 menores

n par?

mediana = 6

{2; 3; 4; 8; 9; 10}



69

MEDIANA

- A mediana Md é uma medida de tendência central cuja definição coincide com o percentil 50%. A mediana Md é um valor localizado na posição central tal que 50% dos valores são menores do que Md e os restantes 50% são maiores.
- Depois de ordenar os n valores da variável de forma crescente, a Md é determinada de acordo com o tipo do número n :
 - Se n for um número ímpar, a Md será o valor da variável situado na posição $(n+1)/2$.
 - Se n for um número par, a Md será igual ao resultado de dividir por dois a soma dos valores das posições $(n/2)$ e $(n/2)+1$. Nesse caso, a Md poderá não ser um valor da variável.

70

Exemplo 3.10

Calcular a *mediana* da amostra do Exemplo 3.1.

Solução. Para facilitar o trabalho, os dados da amostra são repetidos a seguir.

31	38	19	27	24	42	32	18	43	15	39
----	----	----	----	----	----	----	----	----	----	----

A tabela seguinte mostra os onze valores da amostra ordenados de forma crescente, identificando o valor da mediana dentro de um círculo.

15	18	19	24	27	31	32	38	39	42	43
----	----	----	----	----	----	----	----	----	----	----

Como a quantidade de dados da amostra $n=11$ é um número ímpar, o valor da mediana é $Md=31$ que corresponde ao dado da posição $6=(11+1)/2$.

71

Função do Excel

MED(núm1; núm2; ... ; núm30)

- A função estatística **MED(núm1; núm2; ... ; núm30)** retorna a mediana dos valores numéricos *núm1; núm2; ... ; núm30*. Cada um desses *núm* pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados.

	A	B	C	D	E	F
1		Exemplo 3.10				
2						
3		Amostra				
4		31				
5		38				
6		19				
7		27				

Mediana	31,00
----------------	-------

=MED(B4:B14)

72

Exemplo 3.11

Calcular a *mediana* da amostra do Exemplo 3.2.

Solução. Para facilitar o trabalho, os dados da amostra são repetidos a seguir.

27	32	64	65	58	62	59	54	29	30	26	48	47
46	43	38	29	32	35	37	31	43	45	42	37	36

A tabela seguinte mostra os vinte e seis valores da amostra ordenados de forma crescente, identificando os valores que fazem parte do cálculo da mediana dentro de um círculo.

26	27	29	29	30	31	32	32	35	36	37	37	38
42	43	43	45	46	47	48	54	58	59	62	64	65

Como a quantidade de dados $n=26$ é um número par, o valor da mediana será igual ao resultado de dividir por dois a soma dos valores das posições $(n/2)=13$ e $(n/2)+1=14$. O valor da mediana é $Md=40$ resultado obtido de $(38+42)/2$.

73

Propriedades da Mediana

- A quantidade de dados da amostra do Exemplo 3.10 é 11. Portanto, acima da $Md=31$ há cinco dados da amostra, assim como abaixo dela; a mediana é um valor da amostra.
- No Exemplo 3.11, acima da $Md=40$ há treze dados da amostra e abaixo dela também há treze dados, entretanto, a Md não é um valor da amostra.
- A mediana Md divide a área da distribuição de frequências em duas partes iguais a 50%.
- A mediana é uma medida resistente, ela é menos sensível à presença de valores suspeitos, dados bastante diferentes da maioria dos dados coletados na mesma amostra. Por exemplo, se o maior valor da amostra for duplicado, o valor Md não será alterado, pois está relacionada apenas com a ordem da série de valores.

74

MEDIANA

MEDIANA

Vantagens	Desvantagens
Fácil de calcular.	Difícil de incluir em funções matemáticas.
Não é afetada pelos dados extremos da amostra.	Não utiliza todos os dados da amostra.
É um valor único.	
Pode ser aplicada nas escalas: ordinal, intervalar e proporcional.	

75

O que é mais freqüente

Será que
está na
moda
???



76

Moda

- Valor que se repete com maior frequência

$\{2; 3; 4; 7; 7; 9; 10\}$

unimodal

$\{2; 2; 4; 7; 7; 9; 10\}$

bimodal ou multimodal

$\{2; 3; 4; 7; 8; 9; 10\}$

amodal



77

Função do Excel

MODO(núm1; núm2; ... ; núm30)

- A função estatística MODO(núm1; núm2; ... ; núm30) retorna a moda dos valores numéricos núm1; núm2; ... ; núm30. Cada um destes núm pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados.

	A	B	C	D	E	F	
1	Exemplo 3.12						
2							
Amostra							
4		14	14	14	13		
5		12	14	15	15		
6		13	15	13	16		
7		11	17	12	12		
8		12	14	14	12		
9		13	11	13			
10		16	13	14			
11							
12	Moda		14,00		$=MODO(B4:B10;C4:C10;D4:D10;E4:E8)$		
13							

78

39

MODA

MODA

Vantagens	Desvantagens
Fácil de calcular.	Pode estar afastada do centro dos dados.
Não é afetada pelos dados extremos da amostra.	Difícil de incluir em funções matemáticas.
Pode ser aplicada em qualquer escala: nominal, ordinal, intervalar e proporcional.	Não utiliza todos os dados da amostra.
	A amostra pode ter mais de uma moda.
	Algumas amostras podem não ter moda.

79

Média ... Aritmética Simples

- Mais usual das medidas estatísticas
- Relação entre soma e contagem
- Centro geométrico de um conjunto de dados

$$\text{média} = \frac{\text{soma}}{\text{contagem}}$$

$$\mu \text{ ou } \bar{x} = \frac{\sum_{i=1}^n x}{n}$$



80

40

Símbolos de diferentes médias

μ

População

\bar{x}

Amostra



81

MÉDIA

- A medida de posição mais utilizada é a *média aritmética* ou simplesmente *média* de uma amostra ou variável.
- *Média* \bar{X} é o resultado de dividir a soma dos valores das observações $X_1, X_2, \dots, X_i, \dots, X_n$ pela quantidade de dados n :

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

82

Exemplo 3.14

	A	B	C	D	E	F	G	H
1		Exemplo 3.14						
2								
3		Amostra						
4		31						
5		38						
6		19						
7		27						
8		24						
9		42						
10		32						
11		18						
12		43						
13		15						
14		39						
15								

$$\bar{X} = \frac{\sum_{i=1}^{11} X_i}{11} = \frac{31 + 38 + \dots + 39}{11} = 29,82$$

Média | 29,82 | =SOMA(B4:B14)/CONT.NÚM(B4:B14)

Média | 29,82 | =MÉDIA(B4:B14)

83

Função do Excel

MÉDIA(*núm1; núm2; ... ; núm30*)

- A função estatística MÉDIA(*núm1; núm2; ... ; núm30*) retorna a média aritmética dos valores numéricos *núm1; núm2; ... ; núm30*. Cada um desses *núm* pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados. No Exemplo 3.14, a amostra do intervalo B4:B14 foi registrada no primeiro argumento *núm1*. Se o nome da função MÉDIA for inserido com letras minúsculas ou maiúsculas ou sem o acento ortográfico, o Excel aceitará e registrará a função com letras maiúsculas e com o acento ortográfico.

84

1^a. Propriedade da MÉDIA

- A soma dos desvios de uma amostra ou variável é sempre igual a zero.

$$\sum_{i=1}^n (X_i - \bar{X}) = 0$$

- Essa propriedade é útil para verificar ou confirmar o resultado do cálculo da média de uma amostra ou variável, como também no desenvolvimento de provas matemáticas que apresentam a soma de desvios com relação à média.

85

2^a. Propriedade da MÉDIA

- Ao somar ou subtrair uma constante a todos ou de todos os valores de uma série de dados, a média também será somada ou subtraída dessa mesma constante.

3^a. Propriedade da MÉDIA

- Ao multiplicar ou dividir por uma constante todos os valores de uma série de dados, a média também será multiplicada ou dividida por essa mesma constante.

86

Outras Propriedades da Média

- Todos os valores da variável são incluídos no cálculo da média.
- A média é um valor único.
- A média está posicionada de forma equilibrada entre os valores ordenados da amostra. De outra maneira, os valores da amostra se distribuem ao redor da média.
- A média não é uma medida resistente como a mediana ou a moda, pois ela é sensível à presença de dados suspeitos ou extremos, dados com valores bastante diferentes da maioria dos dados coletados na mesma amostra.
- Nas amostras ou variáveis com histograma simétrico, os valores da mediana, a moda e a média coincidem, seus valores são iguais.

87

MÉDIA

MÉDIA

Vantagens	Desvantagens
Fácil de compreender e aplicar.	É afetada pelos dados extremos da amostra.
Utiliza todos os dados da amostra.	É necessário conhecer todos os dados da amostra.
É um valor único.	
Fácil de incluir em funções matemáticas.	
Pode ser aplicada nas escalas: intervalar e proporcional.	

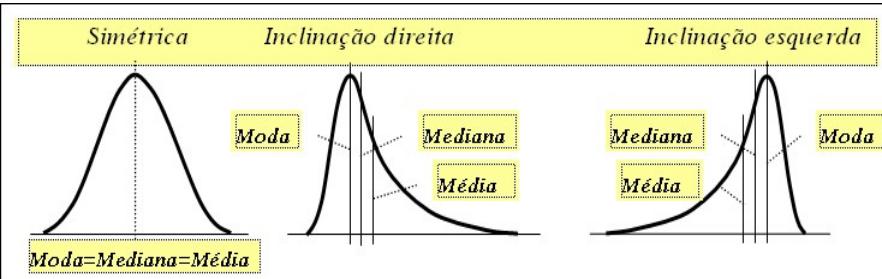
88

INCLINAÇÃO

Simétrica
Média=md=moda

Assimétrica Positiva
Média>md>moda

Assimétrica Negativa
Média<md<moda



89

- 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

90

- Na figura do meio a distribuição tem inclinação para a direita, simplesmente *inclinação direita* ou *positiva*. A moda está na posição do pico da distribuição e a mediana, que divide a distribuição em duas áreas iguais, se situa à direita da moda, pois a distribuição tem inclinação para a direita.
 - Como a média é uma medida afetada pelos dados extremos da amostra se situará à direita da mediana.
 - Utilizando os valores das medidas se verificará a seguinte relação: *Média > Mediana > Moda*.
 - Como nem sempre uma amostra ou variável terá moda, a análise da forma de distribuição poderá ser realizada com as outras duas medidas, *Média > Mediana*. Ou seja, se a média é maior que a mediana, a distribuição deve ter inclinação direita.

91

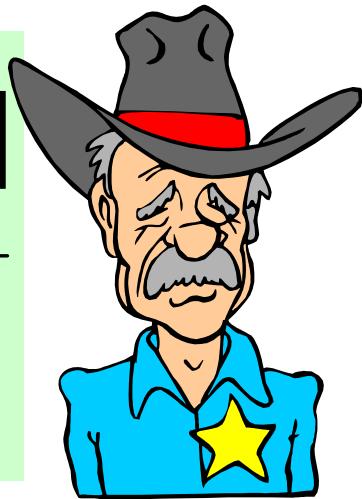
- De forma equivalente, na distribuição da direita na Figura 3.5, a distribuição tem inclinação para a esquerda, simplesmente *inclinação esquerda* ou *negativa*. A moda está na posição do pico da distribuição e a mediana, que divide a distribuição em duas áreas iguais, se situa à esquerda da moda, pois a distribuição tem inclinação para a esquerda.
 - Como a média é uma medida afetada pelos dados extremos da amostra se situará à esquerda da mediana.
 - Utilizando os valores das medidas se verificará a seguinte relação: *Média < Mediana < Moda*.
 - Como nem sempre uma amostra ou variável terá moda, a análise da forma de distribuição poderá ser realizada com as outras duas medidas, *Média < Mediana*. Ou seja, se a média é menor que a mediana a distribuição deve ter inclinação esquerda.

92

Média aritmética ponderada

É preciso considerar as frequências ...

$$\bar{x}_w = \frac{\sum_{i=1}^n [x_i \cdot f(x_i)]}{\sum_{i=1}^n f(x_i)}$$



93

MÉDIA PONDERADA

- O cálculo da média de uma amostra é realizado com todos os dados da amostra.
- Todos os dados recebem a mesma importância ou o mesmo peso, eles têm uma distribuição uniforme e discreta.
- Entretanto, os valores repetidos poderiam ser agrupados, como mostra o cálculo da média do exemplo abaixo.

$$\bar{X} = \frac{1}{26} \times (2 \times 11 + 5 \times 12 + 6 \times 13 + 7 \times 14 + 3 \times 15 + 2 \times 16 + 1 \times 17)$$

94

MÉDIA PONDERADA

Imagine o seguinte conjunto de dados:

i	Notas na Prova de Estatística	Peso	Resultado	
1	8	1	8×1	8
2	4	2	4×2	8
3	6	2	6×2	12
4	10	3	10×3	30
Soma	28	8	58	58
	7	Média Simples		7,25
				Média Ponderada

95

MÉDIA PONDERADA

O capital de uma empresa foi captado por três fontes, ações, financiamento de longo prazo e debêntures, cada um com seu capital próprio definido por uma taxa anual de juros. O objetivo é calcular o custo médio ponderado do capital captado pela empresa, considerando os seguintes valores a seguir:

Capital da empresa	Participação	Taxa de Juros (a.a.)
Acionistas	R\$1.000.000,00	12%
Financiamentos	R\$600.000,00	8%
Debêntures	R\$400.00,00	14%

96

48

MÉDIA PONDERADA - Solução

Capital da empresa	Participação	Taxa de Juros (a.a.)
Acionistas	R\$1.000.000,00	12%
Financiamentos	R\$600.000,00	8%
Debêntures	R\$400.00,00	14%

$$=((1000000*12\%)+(600000*8\%)+(400000*14\%))/(1000000+600000+400000)$$

$$=0,112 \text{ ou } 11,2\%$$

97

Algumas conclusões importantes:

- O cálculo da média ponderada é um caso particular do cálculo da média aritmética.
- Os pesos formam a distribuição de frequências relativas da variável.
- No cálculo da média aritmética, a quantidade de dados da variável é conhecida, entretanto no caso da média ponderada a quantidade de valores da variável não é explícita.
- Uma vantagem do procedimento da média ponderada é poder definir os pesos de cada dado numa previsão, sabendo que a soma dos pesos deve ser sempre igual a um ou 100%.

98

49

	A	B	C	D	E	F	G
1	Funções de tendência central						
2							
3	Amostra		Dados informados como				
4	14	14	Função Matemática	Intervalo	Matriz		
5	12	15	SOMA	352,00	352,00		
6	13	13		352,00	352,00		
7	11	12	Funções Estatísticas				
8	12	14	MÉDIA	13,54	13,54		
9	13	13		13,54	13,54		
10	16	14	MED	13,50	13,50		
11	14	13		13,50	13,50		
12	14	15	MODO	14,00	14,00		
13	15	16		14,00	14,00		
14	17	12					
15	14	12					
16	11						
17	13						
18							
19							

99

Média geométrica

Raiz enésima do produtório

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^n x_i}$$



100

Média geométrica

- Poder ser empregada na análise de dados agrupados ou não. Torna-se muito útil em situações que buscam analisar, por exemplo, um certo padrão ou razão de crescimento.

Mês	Vendas	Razão
Janeiro	10.000	
Fevereiro	14.000	1,4
Março	16.800	1,2
Abril	21.840	1,3
Maio	24.024	1,1

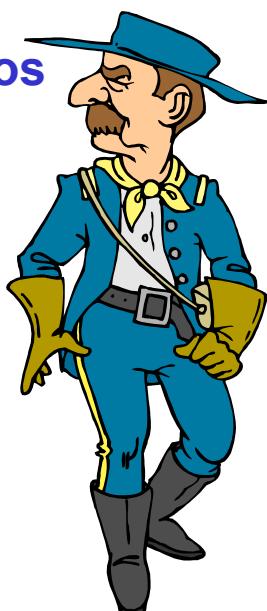
$$\sqrt[4]{1,4 \times 1,2 \times 1,3 \times 1,1} = 1,245 \Rightarrow \text{Média geométrica do crescimento}$$
$$= \text{media.geométrica(intervalo)} = 16.529,78 \Rightarrow \text{Média geométrica bruta}$$
$$10.000 \times 14.000 \times 16.800 \times 21.840 \times 24.024 = 1.234.057.144.320.000.000.000 = x \Rightarrow x^{(1/5)} = 16.529,78$$

101

Média harmônica

Inverso da média dos inversos

$$\bar{x}_h = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$



102

Média harmônica

- A utilização prática da média harmônica concentra-se em cálculos que envolvem, por exemplo, a obtenção de velocidades ou tempos médios.

Mês	Vendas
Janeiro	30
Fevereiro	60
Março	40
Abril	90

$$=\text{MÉDIA.HARMÔNICA}(30;60;40;90) = 46,45$$

103

Maior problema da média ...

**Maldição
dos
extremos**
ou outliers

**Extremos distorcem
algunas medidas**



104

Solução para o problema ...



105

Pesquisa sobre remuneração

- Empresa paga R\$400,00 aos estagiários de Administração
- Quer saber ...

É muito ou pouco?

- Coletou amostra de dados
- Dados:
 $\{300; 350; \underline{6000}; 340; 310; 380\}$

Pouquíssimo!!
!



$$\$1.280,00 : \frac{7680}{6}$$

106

Organizando os dados ...

- **Dados:**

{300; 350; 6000; 340; 310; 380}

- **Rol:**

{300; 310; 340; 350; 380; ~~6000~~}

\$400,00

Extremo distorce a média!

- **Rol sem extremo:**

{300; 310; 340; 350; 380}

Alto!

Média = 1680/5 = \$336,00

107

QUE MEDIDA UTILIZAR?

- Quando se procura conhecer valores totais, a média é utilizada. Por exemplo, em controle de qualidade a média é utilizada para determinar se o processo opera ao redor de um valor esperado ou alvo. Também, dá-se preferência à média pelas suas propriedades matemáticas.

108

- Se a amostra apresentar valores extremos, uma distribuição com acentuada inclinação, a mediana será mais adequada, pois não é afetada pelos dados extremos, como a média. Se quisermos conhecer o valor típico dos salários de uma determinada categoria de trabalhadores será utilizada a mediana. Por exemplo, se os salários pesquisados da categoria são: \$500, \$1.800, \$2.000, \$2.200 e \$2.500, a mediana é \$2.000 e a média é \$1.800. Portanto, o valor da média tende na direção dos valores extremos e a mediana não é afetada por esses valores extremos.
- A moda é um valor típico de uma amostra ou variável. Por exemplo, na distribuição do consumo de um mesmo produto com diferentes apresentações, a moda mostra a apresentação mais consumida, como é o caso do número de calçados, o tamanho de calças etc.

109

Encontrando o centro dos dados

- Fundo de investimento, com retornos : {7, 3 e 2}
- Média *ou soma por contagem*

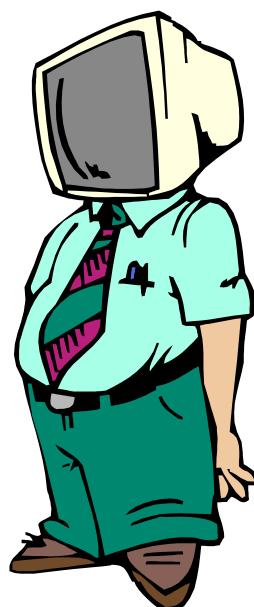
$$\text{Média} = (7 + 3 + 2) / 3 = \underline{\textcolor{red}{4}}$$

- Mediana *ou centro da série ordenada*

$$\text{Mediana} = \{2, \underline{\textcolor{red}{3}}, 7\}$$

- Moda *ou valor que mais se repete*

Amodal ou sem moda



110

Capítulo

3

Medidas de dispersão

111

Outras EstatísticaS

- Outras medidas úteis para a decisão
- “Cuidado com os lados” ...
- Medidas de dispersão
 - Amplitude
 - Desvio-médio
 - Variância
 - Desvio-padrão



112

- No Capítulo 2 foi mostrado que a média e a mediana determinam um valor central de uma amostra ou variável. Enquanto a mediana localiza a posição do dado ou observação situada no centro da amostra ordenada de forma crescente, e sem considerar os valores da variável, a média determina o valor central considerando todos os valores da variável.
- Por exemplo, as amostras $X=\{28, 29, 30, 31, 32\}$ e $Y=\{21, 25, 29, 34, 41\}$ têm o mesmo número de dados e, também, a mesma média 30. Entretanto, os desvios são diferentes, pois os desvios da variável X são -2, -1, 0, 1 e 2, e os desvios da variável Y são -9, -5, -1, 4 e 11.
- A comparação dessas duas amostras aponta a variabilidade ou dispersão de seus dados com relação à média como uma medida importante para descrever uma amostra ou variável.

113

- Você deve lembrar que, se não houver variabilidade, a maior parte das medidas estatísticas não tem utilidade.
- Há várias formas de medir a variabilidade dos dados de uma variável.
- Uma primeira tentativa é medir o intervalo ou *range* de variação, definido como o resultado da diferença entre os valores máximo e mínimo da amostra ou variável.

114

Encontrando os lados dos dados

- Notas em provas de Matemática: {7, 3 e 2}
- Amplitude
- O Maior menos o menor
- Range ou intervalo

$$R = \{ \text{Maior} - \text{Menor}$$
$$R = 7 - 2 = 5$$

*Problema:
apenas extremos
são considerados*



115

Exemplo 4.1

Determinar o intervalo de variação da amostra seguinte.

31	38	19	27	24	42	32	18	43	15	39
----	----	----	----	----	----	----	----	----	----	----

Solução. Os valores mínimo e máximo são, respectivamente, 15 e 43. O intervalo ou range de variação dos dados da amostra é $43 - 15 = 28$.

O resultado do Exemplo 4.1 mostra que os dados da amostra se distribuem dentro do intervalo de variação igual a 28. O conhecimento desse intervalo não auxilia muito na tentativa de medir a dispersão dos dados da variável, pois seu cálculo envolve apenas os valores extremos, deixando de considerar os outros valores da variável, que também são importantes.

116

Desvio-médio

- Desvio-médio ou afastamento médio em relação à média

$$DM = \frac{\sum_{i=1}^n (x_i - \bar{x})}{n}$$

Série Desvios

Média = 4

2	-2
3	-1
7	3
Soma	0
Média	0

É preciso
calcular os
desvios
ABSOLUTOS

117

DESVIO MÉDIO ABSOLUTO

- No Capítulo 2 vimos que os desvios dos dados de uma amostra ou variável medem sua dispersão ao redor de sua média. Portanto, a tentativa inicial de quantificar a variabilidade seria calcular a soma de todos os desvios. No entanto, pela primeira propriedade da média, a soma dos desvios é sempre igual a zero.
- Tentando manter o conceito desvio como medida de variabilidade, podemos utilizar a média dos valores absolutos dos desvios, procedimento denominado como *desvio absoluto médio* ou simplesmente *DAM*

118

O Desvio absoluto médio-DAM é obtido da expressão:

$$DAM = \frac{1}{n} \times (|X_1 - \bar{X}| + |X_2 - \bar{X}| + \dots + |X_n - \bar{X}|)$$

$$DAM = \frac{1}{n} \times \sum_{i=1}^n |X_i - \bar{X}|$$

onde X_i é um valor genérico e \bar{X} é a média da variável ou amostra.

119

Desvio-médio absoluto

- Desvio-médio absoluto ou afastamento médio absoluto em relação à média

$$DMA = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Série Desv. Abs.

Média = 4

2	2
3	1
7	3
Soma	6
Média	2

Calculamos os
MÓDULOS

120

Função do Excel

DESV.MÉDIO(núm1; núm2; ... ; núm30)

- A função estatística DESV.MÉDIO retorna o **desvio absoluto médio** dos valores numéricos *núm1; núm2; ... ; núm30*. Cada um desses *núm* pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados.

A	B	C	D	E	F	G	H
1	Exemplo 4.2						
2							
3	Amostra	Desvio	Desvio Absoluto		Resultados		
4	31	1,18	1,18				
5	38	8,18	8,18		Média	29,82	
6	19	-10,82	10,82		Soma dos Desvios Absolutos	92,18	
7	27	-2,82	2,82		DAM	8,38	
8	24	-5,82	5,82		Função DESV.MÉDIO	8,38	
9	42	12,18	12,18				=DESV.MÉDIO(B4:B14)
10	32	2,18	2,18				
11	18	-11,82	11,82				
12	43	13,18	13,18				
13	15	-14,82	14,82				
14	39	9,18	9,18				
15							

121

- Comparado com a tentativa de medir a variabilidade com o *intervalo*, o *DAM* é a média dos desvios absolutos e utiliza todos os valores da variável ou amostra.
- Entretanto, o valor absoluto dos desvios é um resultado difícil de compreender e não aceita tratamento matemático com as propriedades, por exemplo, do quadrado do desvio que será utilizado a seguir.

122

Variância

- Dispensa o uso do MÓDULO
- Usa o desvio ao quadrado

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

Série Desvio²

Média = 4	
2	4
3	1
7	9
Soma	14
Média	4,67

Um problema
DIMENSIONAL

123

VARIÂNCIA

- Mantendo os *desvios* para medir a variabilidade de uma variável, o procedimento recomendado é utilizar a soma dos quadrados dos desvios, pois seu resultado é um valor mínimo, como mostrou a segunda propriedade da média apresentada no Capítulo 2.

124

- Seja a variável $X = X_1, X_2, \dots, X_N$ uma população. Define-se a variância σ_X^2 da variável X da população contendo N dados:

$$\sigma_X^2 = \frac{1}{N} \times ((X_1 - \mu_X)^2 + (X_2 - \mu_X)^2 + \dots + (X_n - \mu_X)^2)$$

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

- Seja a variável $X = X_1, X_2, \dots, X_n$ uma amostra. Define-se a variância s_X^2 da variável X da amostra contendo n dados:

$$s_X^2 = \frac{1}{n-1} \times ((X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2)$$

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

125

	A	B	C	D	E	F	G	H	I
1	Exemplo 4.3								
2									
3	Amostra	Desvio	(Desvio)²		Resultados				
4	31	1,18	1,40		n	11	=CONT.NÚM(B4:B14)		
5	38	8,18	66,94		Média	29,82	=MÉDIA(B4:B14)		
6	19	-10,82	117,03		Soma (Desvios)²				
7	27	-2,82	7,94		Fórmula	997,64	=SOMA(D4:D14)		
8	24	-5,82	33,85		Função SOMAQUAD	997,64	=SOMAQUAD(C4:C14)		
9	42	12,18	148,40		Variância amostra				
10	32	2,18	4,76		Fórmula	99,76	=G6/(G4-1)		
11	18	-11,82	139,67		Função VAR	99,76	=VAR(B4:B14)		
12	43	13,18	173,76		Variância população				
13	15	-14,82	219,58		Fórmula	90,69	=G6/G4		
14	39	9,18	84,31		Função VAR	90,69	=VARP(B4:B14)		
15									
16									
17									
18									

126

Cálculo da variância da amostra. Com os resultados parciais obtidos se pode calcular o valor da variância da amostra $S_x^2 = 99,76$ utilizando:

- Manualmente a fórmula $S_x^2 = \frac{\sum_{i=1}^{11} (X_i - \bar{X})^2}{11-1} = \frac{997,64}{10} = 99,76$
- Registrando a fórmula =G8/(G4-1) na célula G12 da planilha.
- Utilizando a função estatística VAR e registrando a fórmula =VAR(B4:B14) na célula G13.

Cálculo da variância da população. Com os resultados parciais obtidos se pode calcular o valor da variância da amostra $\sigma_x^2 = 90,69$ utilizando:

- Manualmente a fórmula $\sigma_x^2 = \frac{\sum_{i=1}^{11} (X_i - \mu_X)^2}{11} = \frac{997,64}{11} = 90,69$
- Registrando a fórmula =G8/G4 na célula G16 da planilha.
- Utilizando a função estatística VARP e registrando na célula G17 a fórmula =VARP(B4:B14).

127

- O procedimento de cálculo utilizando a soma dos quadrados dos desvios é bastante trabalhoso.
- A Função SomaQuad no Excel é um procedimento de cálculo da variância que utiliza somente os dados da amostra e os quadrados desses dados, não sendo necessário utilizar a média e os desvios.
- Entretanto, esse procedimento de cálculo perde força quando comparado com a utilização das funções estatísticas do Excel.

128

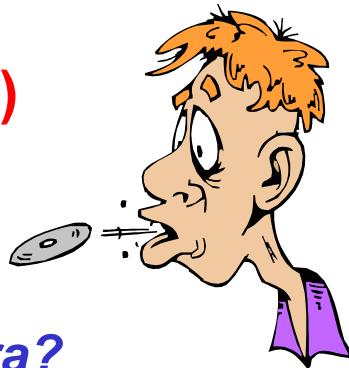
Desvio-padrão

- Resolve o problema dimensional da variância
- Raiz da variância

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

**Desvio = Raiz (4,67)
= 2,16**

*Ops ...
População ou amostra?*



129

DESVIO PADRÃO

- Para definir a variância, nos valemos da segunda propriedade da média: a soma dos quadrados dos desvios é sempre um valor mínimo, como foi apresentado no Capítulo 2.
- Uma desvantagem da variância é sua unidade de medida, o quadrado da unidade de medida dos dados da amostra ou variável, outra desvantagem é ampliar os desvios, pois são elevados ao quadrado.
- Como a unidade de medida da variância não explica nada sobre as características dos valores da amostra, é definido o desvio padrão que mantém a unidade de medida dos valores da variável.

130

O desvio padrão da variável X é a raiz quadrada positiva de sua variância. Dessa maneira:

- O desvio padrão considerado como população é: $\sigma_X = +\sqrt{\sigma_X^2}$.
- O desvio padrão considerado como amostra é: $S_X = +\sqrt{S_X^2}$.

131

	A	B	C	D	E	F	G	H	I
Exemplo 4.4									
Amostra Desvio $(Desvio)^2$									
31	1,18	1,40							
38	8,18	66,94							
19	-10,82	117,03							
27	-2,82	7,94							
24	-5,82	33,85							
42	12,18	148,40							
32	2,18	4,76							
18	-11,82	139,67							
43	13,18	173,76							
15	-14,82	219,58							
39	9,18	84,31							
Resultados									
		n	11						
		Média	29,82						
		Soma (Desvios) ²	997,64						
		Variância amostra	99,76						
		Variância população	90,69						
Desvio padrão amostra									
		Fórmula	9,99	=RAIZ(G7)					
		Função DESVPAD	9,99	=DESVPAD(B4:B14)					
Desvio padrão população									
		Fórmula	9,52	=RAIZ(G8)					
		Função DESVPADP	9,52	=DESVPADP(B4:B14)					

132

Cálculo do desvio padrão da amostra. O valor da desvio padrão da amostra $S_x = 9,99$ pode ser obtido:

- Manualmente a fórmula $S_x = +\sqrt{S_x^2} = +\sqrt{99,76} = 9,99$
- Registrando a fórmula $=RAIZ(G7)$ na célula G11 da planilha.
 - A função matemática **RAIZ(número)⁵** retorna a raiz quadrada positiva do argumento *número* que deve ser qualquer número positivo.
- Utilizando a função estatística DESVPAD e registrando na célula G12 a fórmula $=DESVPAD(B4:B14)$.

Cálculo do desvio padrão da população. O valor da desvio padrão da população $\sigma_x = 9,52$ pode ser obtido:

- Manualmente a fórmula $\sigma_x = +\sqrt{\sigma_x^2} = +\sqrt{90,69} = 9,52$
- Registrando a fórmula $=RAIZ(G8)$ na célula G15 da planilha.
- Utilizando a função estatística DESVPADP e registrando na célula G16 a fórmula $=DESVPADP(B4:B14)$.

133

Funções do Excel, para amostra

VAR(núm1; núm2; ... ; núm30)

- **A função estatística VAR retorna a variância da amostra dos valores numéricos núm1; núm2; ... ; núm30.**

DESVPAD(núm1; núm2; ... ; núm30)

- **A função estatística DESVPAD retorna o desvio padrão da amostra dos valores numéricos núm1; ... ; núm30.**
- **Nas duas funções, cada um dos argumentos núm pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados. No exemplo, a amostra do intervalo B4:B14 foi registrado no primeiro argumento núm1.**

134

Funções do Excel para população

VARP(*núm1; núm2; ... ; núm30*)

- A função estatística VARP retorna a **variância da população** dos valores numéricos *núm1; núm2; ... ; núm30*.

DESVPADP(*núm1; núm2; ... ; núm30*)

- A função estatística DESVPADP retorna o **desvio padrão da população** dos valores numéricos *núm1; ... ; núm30*.
- Nas duas funções, cada um dos argumentos *núm* pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados.

135

Para sempre lembrar

- Calcule amplitude, desvio-médio absoluto, variância e desvio-padrão da série:

{10; -2; 5; 7}



136

Calculando Amplitude e Desvio-Médio

<u>X_i</u>	<u>X_i</u>	<u>$X_i - \bar{X}_i$</u>
10	-2	$ -2 - 5 = 7$
-2	5	$ 5 - 5 = 0$
5	7	$ 7 - 5 = 2$
7	10	$ 10 - 5 = 5$

Média = 5 Soma = 14
Amplitude = 12 Desv. Médio Abs. = 3,5

137

Calculando Variância

<u>X_i</u>	<u>$(X_i - \bar{X}_i)^2$</u>
-2	$(-2 - 5)^2 = 49$
5	$(5 - 5)^2 = 0$
7	$(7 - 5)^2 = 4$
10	$(10 - 5)^2 = 25$

Média = 5 Soma = 78
Variância = 19,5

138

Calculando o Desvio-Padrão

- Desvio-padrão = raiz (variância)
- Desvio = raiz (19,50)
- Desvio = 4,4159



139

Algumas formulazinhas

	Populacional	Amostral
Variância	$\sigma^2 = \frac{\sum (X_i - \bar{X})^2}{n}$	$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$
Desvio-Padrão	$\sigma = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n}}$	$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$

140

A herança do João...
Vamos calcular?

Nosso amigo
recebeu \$400
mil de herança
e deseja
aplicar...



141

Os dados ...

Mês	Retornos % do fundo A	Retornos % do fundo B
Ago	1	5
Set	15	11
Out	8	8
Nov	13	9
Dez	3	7

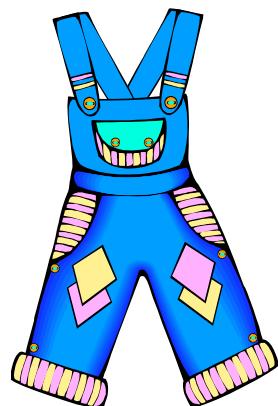
Calcule:

- a) Média
- b) Mediana
- c) Moda
- d) Amplitude
- e) Variância (Pop.)
- f) Desvio-padrão (Pop.)

142

Estatísticas de A

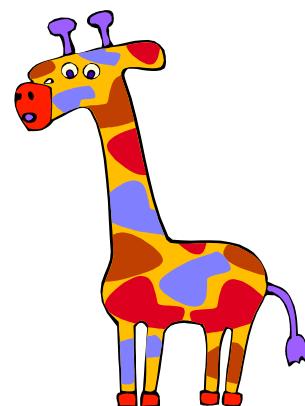
Mês	A	$(A - M)^2$
Ago	1	49
Set	15	49
Out	8	0
Nov	13	25
Dez	3	25
Soma	40	148
Contagem	5	5
Soma/Cont.	8	29,6
	<i>Média</i>	<i>Variância</i>
Mediana	8	
Moda	-	↓
Desvio		5,44



143

Estatísticas de B

Mês	B	$(A - M)^2$
1	5	9
2	11	9
3	8	0
4	9	1
5	7	1
Soma	40	20
Contagem	5	5
Soma/Cont.	8	4
	<i>Média</i>	<i>Variância</i>
Mediana	8	
Moda	-	↓
Desvio		2,00



144

Concluindo ...

	A	B
Média	8	8
Desvio	5,44	2,00

**Um fundo mais seguro é
aquele que ...**

**oferece um menor
risco em perdas!!!**

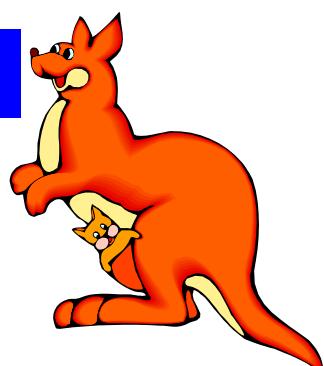
145

Significado do desvio-padrão

- De um modo geral, o desvio-padrão representa a mais clássica medida de dispersão da estatística. Sua associação ao valor da média, somado ou subtraído, permite encontrar e determinar as frequências relativas dos valores analisados.

Coeficiente de variação

$$CV = \frac{\sigma}{\mu} \text{ ou } \frac{s}{\bar{x}}$$



146

Coeficiente de variação

$$CV = \frac{\sigma}{\mu} ou \frac{s}{\bar{x}}$$

A princípio considera-se que quanto menor o CV, mais homogêneos são os dados.

Baixo: CV inferior a 10%,

Médio: CV entre 10 e 20%,

Alto: CV entre 20 e 30%,

Muito Alto: para valores acima de 30%

147

Capítulo

4

Medidas de ordenamento e forma

148

Medidas

Ordenamento

“Posição relativa”

149

Dividem a série ordenada

- **Mediana**
Divide ao meio
- **Quartis**
Dividem em quatro
- **Decis**
Dividem em dez
- **Centis ou Percentis**
Dividem em cem



150

Dividem a série ordenada

Quartis

**Dividem a distribuição
ordenada em quatro
partes iguais**

$$Q_{nq} = x_{\left[\frac{nq \cdot n}{4} + \frac{1}{2} \right]}$$



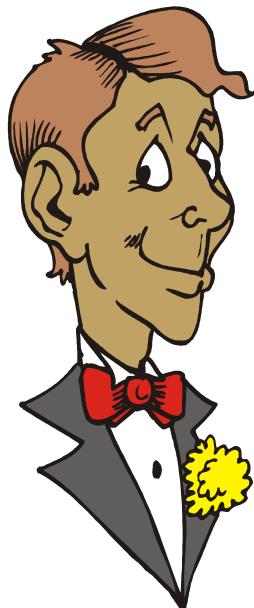
151

Dividem a série ordenada

▪ **Decis**

**Dividem a distribuição ordenada
em dez partes iguais**

$$D_{nd} = x_{\left[\frac{nd \cdot n}{10} + \frac{1}{2} \right]}$$



152

Dividem a série ordenada

▪ **Centis ou Percentis**

**Dividem a distribuição
ordenada em cem
partes iguais**

$$P_{nd} = x_{\left[\frac{np.n}{100} + \frac{1}{2} \right]}$$



153

Medidas

Forma

“É normal?”

154

Tipos principais de medidas

Assimetria

Curtose

155

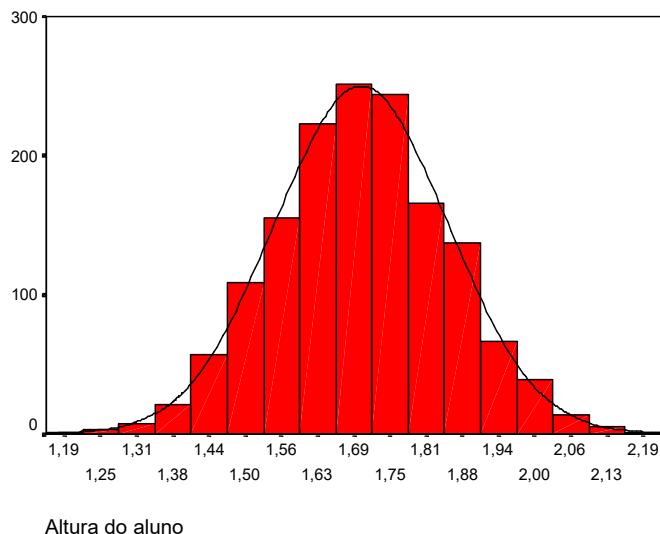
Assimetria

**Analisa a
concentração
das
distribuições de
frequência em
torno do eixo**



156

Afastamento ao eixo de simetria



157

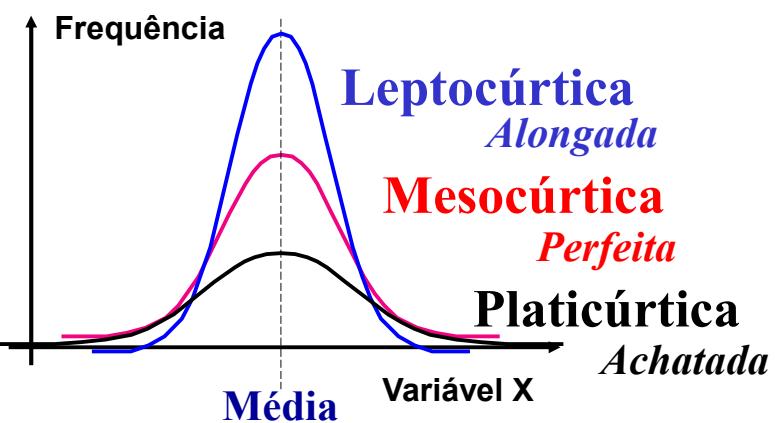
Curtose

**Analisa o
achatamento
da curva**



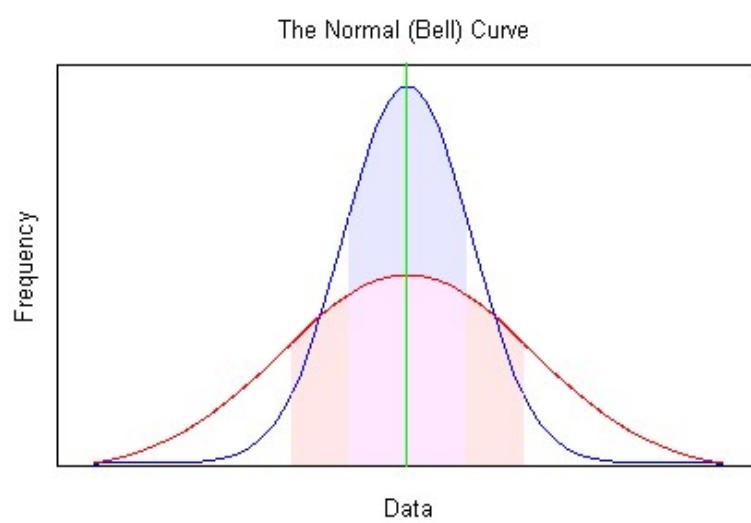
158

Analizando o achatamento



159

Diferentes curtoses



160

Calculando a curtose

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

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

$k=0,263$: distribuição mesocúrtica

distribuição nem chata nem delgada.

$k > 0,263$: distribuição leptocúrtica

distribuição delgada

$k < 0,263$: distribuição platicúrtica

distribuição achatada

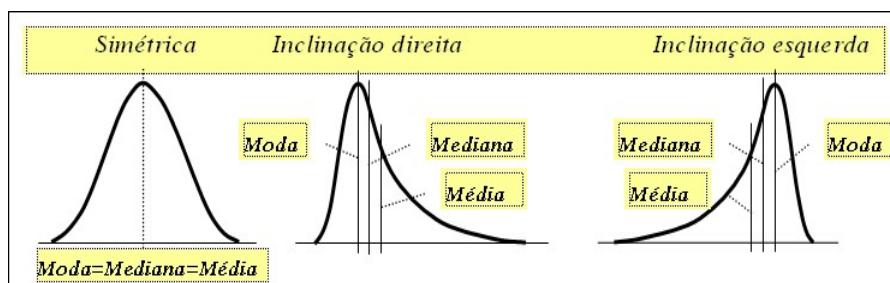
161

INCLINAÇÃO

Simétrica
Média=md=moda

Assimétrica Positiva
Média>md>moda

Assimétrica Negativa
Média<md<moda



162

- 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

163

Calculando a assimetria

Coeficiente de Pearson:

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

164

Subcapítulo

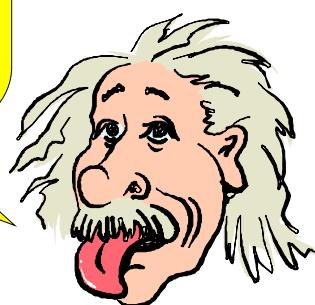
4.1

Gráficos

165

Para pensar ...

“Nem tudo o que pode ser contado conta,
e nem tudo o que conta pode ser contado”.



Einstein

166

Gráficos principais

- Uma variável quantitativa
 - Histograma
- Uma qual outra quanti
 - Boxplot
- Duas variáveis quantitativas
 - Dispersão
- Variáveis qualitativas
 - Colunas
 - Barras
 - Pizza (todo = 100%)

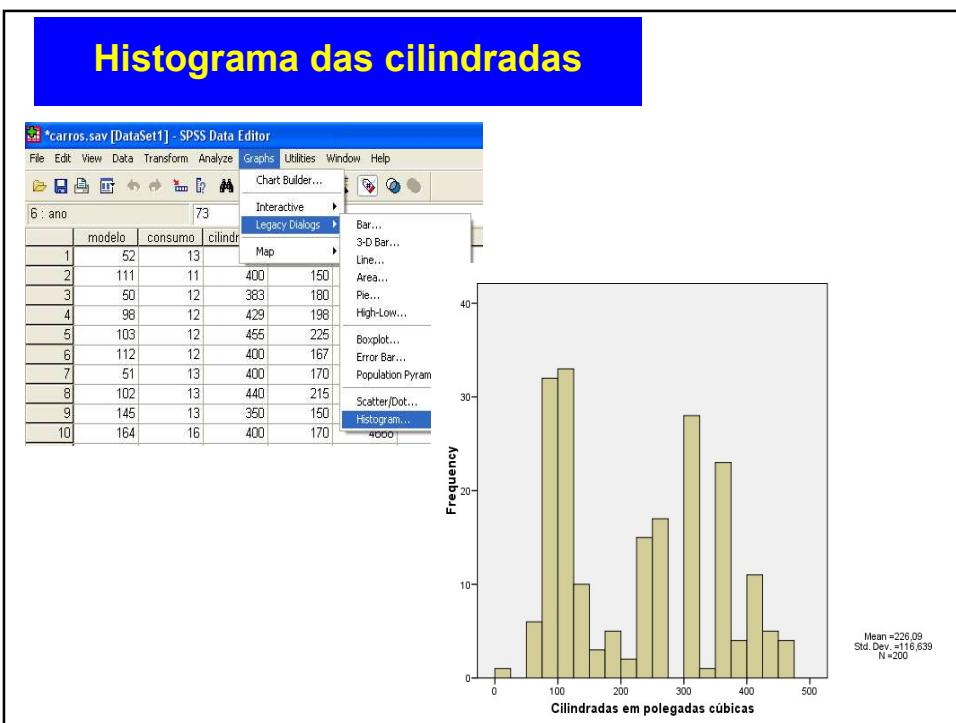


167

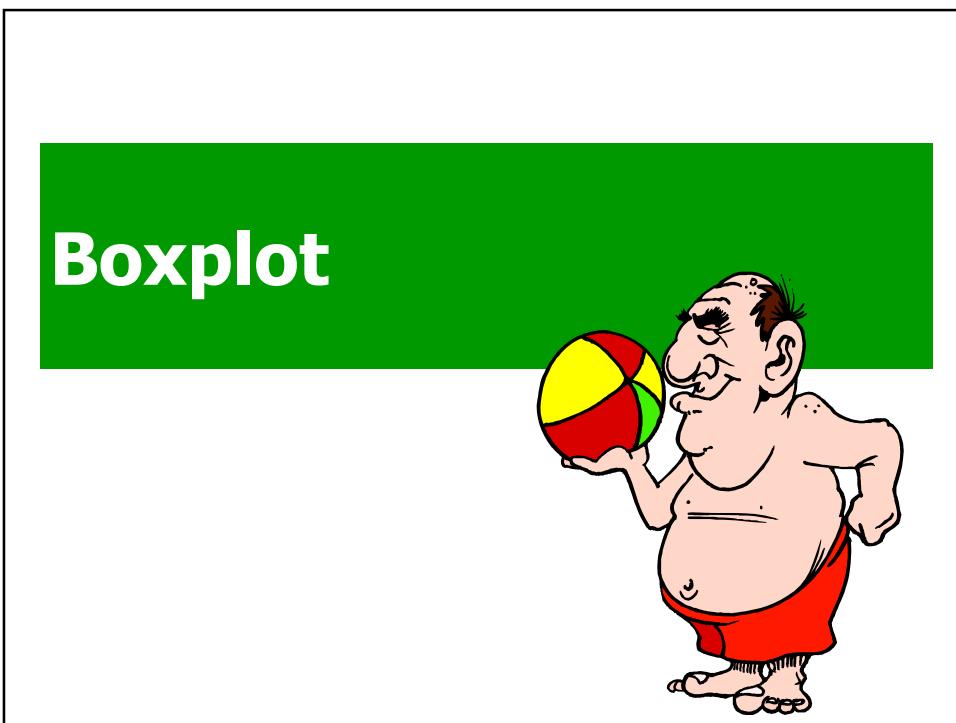
Histograma



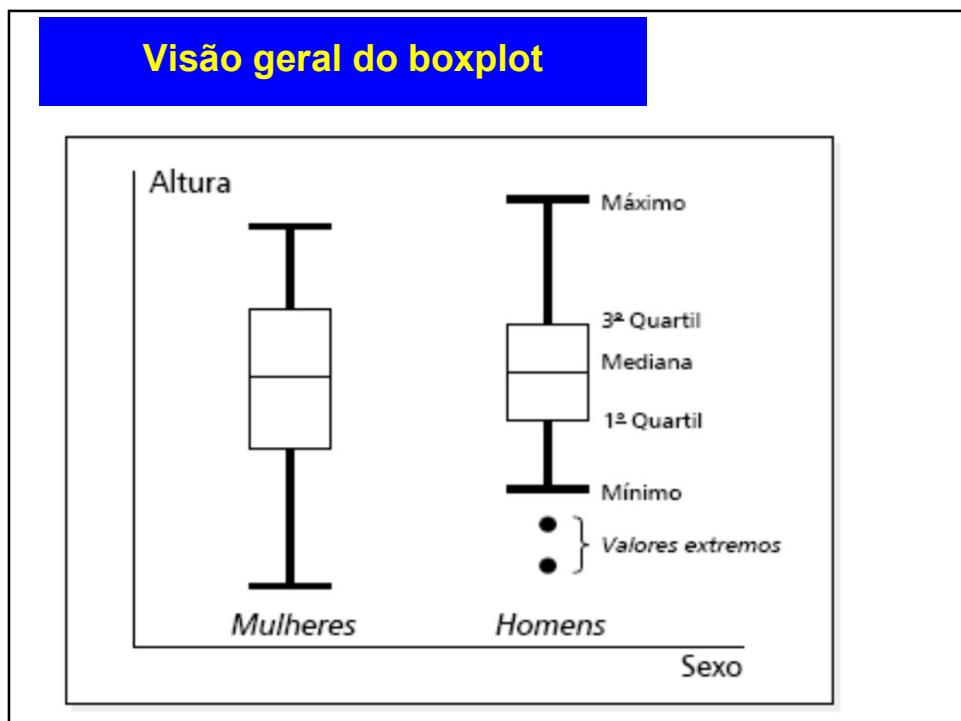
168



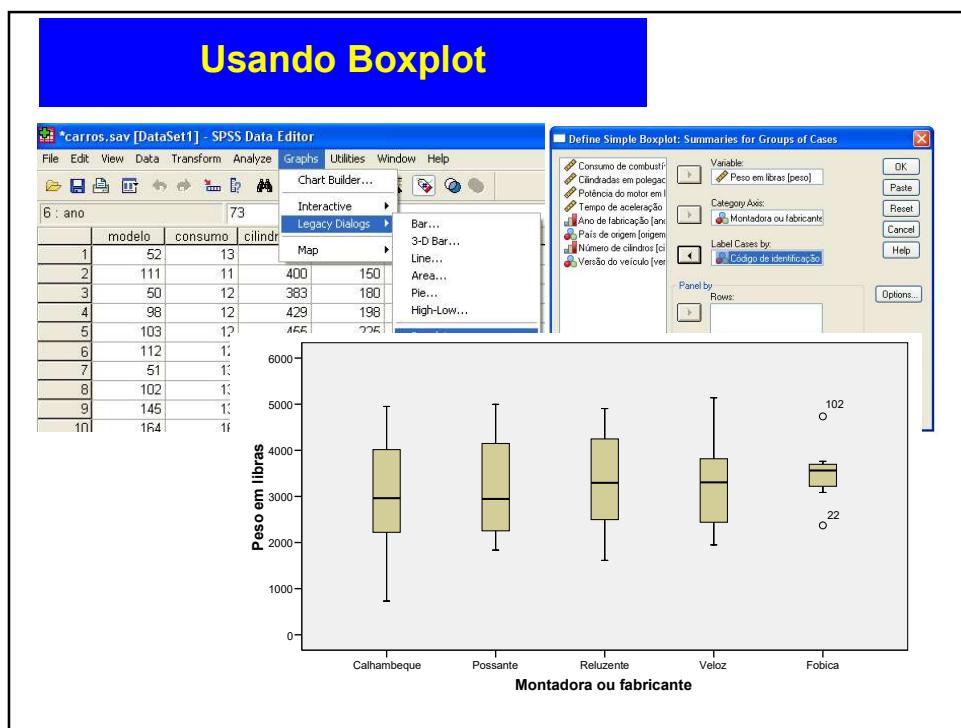
169



170



171



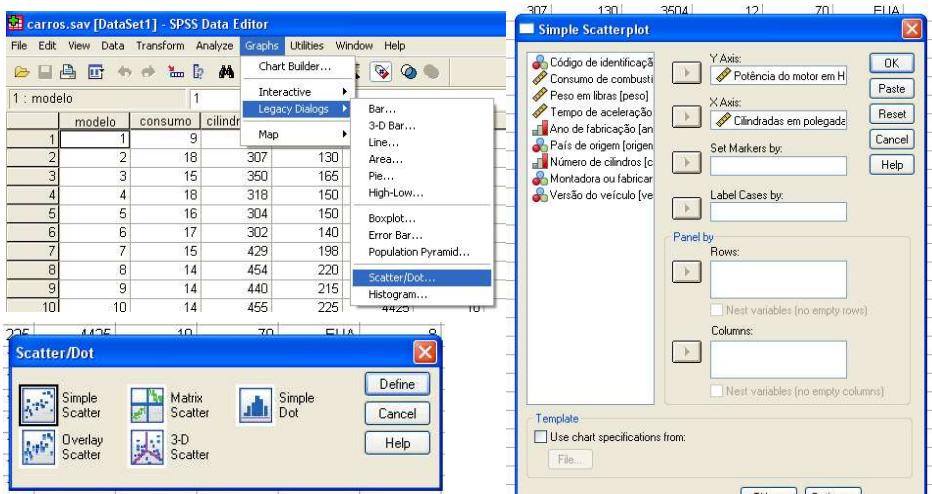
172

Gráfico ou diagrama de dispersão



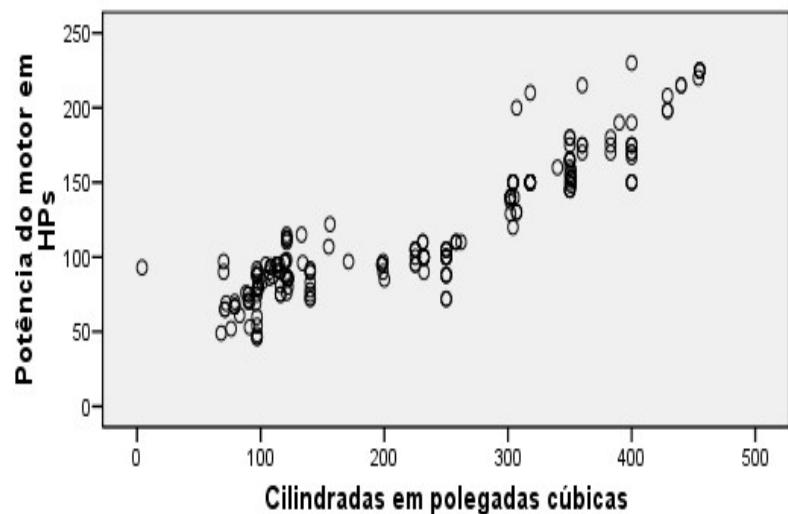
173

Gráficos de Dispersão



174

Resultado da dispersão



175

Capítulo

5

Probabilidade

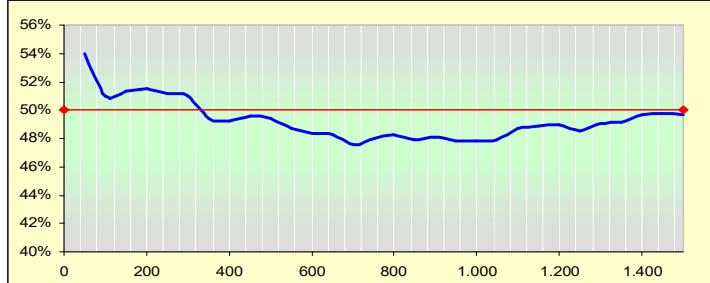
176

- Este capítulo ajudará a descrever a informação amostrada, facilitará a apresentação desses resultados e fornecerá uma ferramenta útil para realizar inferências sobre a população de onde foi extraída a amostra.
- Pela própria experiência, o resultado do lançamento de uma moeda pode ser cara ou coroa, descartando a moeda falsa com duas caras, ou duas coroas, ou aquela que possa ficar de pé apoiada na sua borda.

177

- Além disso, periodicamente recebemos informações como a seguinte: na pesquisa de intenção de voto para o segundo turno da eleição para governador, 43% dos eleitores da amostra preferem o candidato A, 37% dos eleitores preferem o candidato B e os restantes 20% dos eleitores não sabem.
- Qual a característica comum do lançamento de uma moeda e da pesquisa de intenção de voto? O resultado não pode ser previsto com antecedência! Por quê? Porque o resultado variará toda vez que lançarmos uma moeda ou extraímos outra amostra para a pesquisa de intenção de voto.
- Entretanto, se o lançamento da moeda for repetido um número muito grande de vezes, perceberemos uma tendência dos resultados.

178



- Esse gráfico, um dos muitos gráficos possíveis, representa a proporção de caras numa simulação de 1.500 lançamentos de uma moeda. O resultado dessa simulação em particular mostra que a proporção de caras tende a 50%, lembrando que esse gráfico foi especialmente escolhido para esta apresentação, pois, tecnicamente, a simulação de 1.500 lançamentos é um número pequeno de tentativas.

179

EXPERIMENTOS E EVENTOS

- Todo processo desenvolvido para realizar observações e obter dados com um determinado objetivo é denominado **experimento**.
- O conjunto formado por todos os resultados possíveis de um experimento é denominado **espaço amostral** do experimento.
- Um experimento é **aleatório** quando pode resultar em um dos resultados do espaço amostral sem que se seja possível predizer com certeza qual o resultado que será observado.

180

Se apesar de conhecer todos os resultados de um experimento não for possível antecipar seu resultado, esse experimento é denominado *experimento aleatório*.

Espaço amostral é o conjunto de todos os possíveis e diferentes resultados de um experimento aleatório.

A análise de um experimento aleatório começa pela identificação de todos seus resultados possíveis.

Por exemplo, no experimento do lançamento de duas moedas, seu espaço amostral é formado pelos quatro resultados possíveis CaCa, CaCo, CoCa e CoCo, ou o conjunto S dos resultados possíveis $S=\{CaCa, CaCo, CoCa, CoCo\}$. Cada resultado desse espaço amostral S é denominado *ponto amostral*.

181

Evento elementar é um resultado único do espaço amostral.

Evento é um subconjunto formado por um ou mais resultados do espaço amostral.

Um subconjunto do espaço amostral S é denominado evento.

Por exemplo, o *evento dos resultados que têm exatamente apenas uma cara* é descrito pelo subconjunto do espaço amostral $A=\{CaCo, CoCa\}$.

É importante observar que um evento pode ser partido (dividido) em seus eventos elementares.

182

- Os resultados possíveis do lançamento de uma moeda são apenas dois, os eventos elementares *Cara-Ca* e *Coroa-Co*.
- Pela própria característica do experimento, se o resultado de um lançamento for cara, esse resultado não poderá ser coroa ao mesmo tempo, pois são eventos *mutuamente excludentes*.
- A união de eventos elementares forma o espaço amostral, pois são eventos *coletivamente exaustivos*. Portanto, verifica-se que os eventos *A* e *B* pertencentes ao mesmo espaço amostral *S*:
 - São *mutuamente excludentes* se sua interseção $A \cap B$ for vazia, pois os dois eventos não têm nenhum elemento em comum.
 - São *coletivamente exaustivos* se a união $A \cup B$ dos eventos formarem o espaço amostral *S*, em que cada evento pode ter elementos repetidos no outro evento.

183

Exemplo 5.1

Analizar os resultados do lançamento de uma moeda.

Solução. Como o espaço amostral do lançamento de uma moeda tem apenas dois eventos, os eventos elementares *Ca* e *Co* são eventos mutuamente excludentes, eventos complementares e eventos coletivamente exaustivos.

Exemplo 5.2

A nota final do curso de *estatística* pode ser: conceito *A*, ou conceito *B* ou conceito *C*. Analisar os resultados dessas notas.

Solução. O espaço amostral da nota final de *estatística* está formado por três eventos elementares: conceito *A*, ou conceito *B* e conceito *C*. Os três conceitos são eventos mutuamente excludentes e eventos coletivamente exaustivos, pois quando agrupados formam o espaço amostral de todos os conceitos. Não são eventos complementares, pois o complemento do conceito *A* é a união do conceito *B* e do conceito *C*.

184

PROBABILIDADE

- Depois de apresentar os conceitos de experimento e eventos, o objetivo é dirigido para a avaliação do sucesso de ocorrer um determinado evento do espaço amostral de um experimento aleatório.
- Por exemplo, no lançamento de uma moeda um número muito grande de vezes, o sucesso de ocorrer o evento *Cara* é medido pela probabilidade $P(\text{Cara})$, um valor dentro do intervalo $(0, 1)$ incluindo ambos os limites.

185

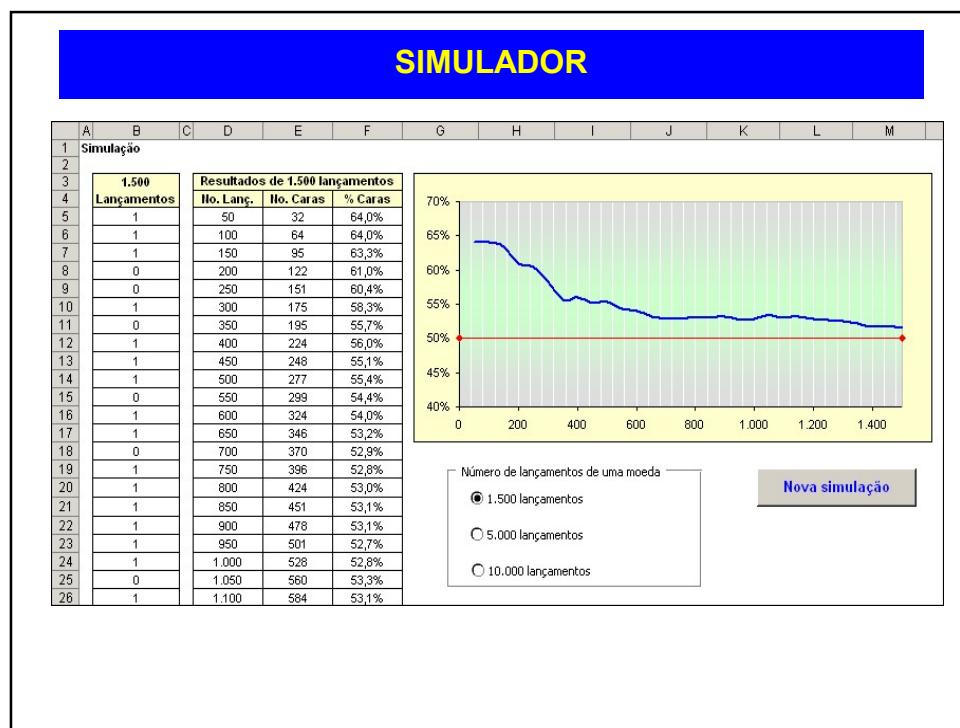
A probabilidade de sucesso $P(A)$ do evento A é um número entre zero e um. Tendo presente que a probabilidade $P(A)$ está associada à proporção de sucessos do evento A:

- Se $P(A)=0$ o evento A nunca ocorrerá, pois é um evento impossível.
- Se $P(A)=1$ o evento A sempre ocorrerá, pois é um evento certo.

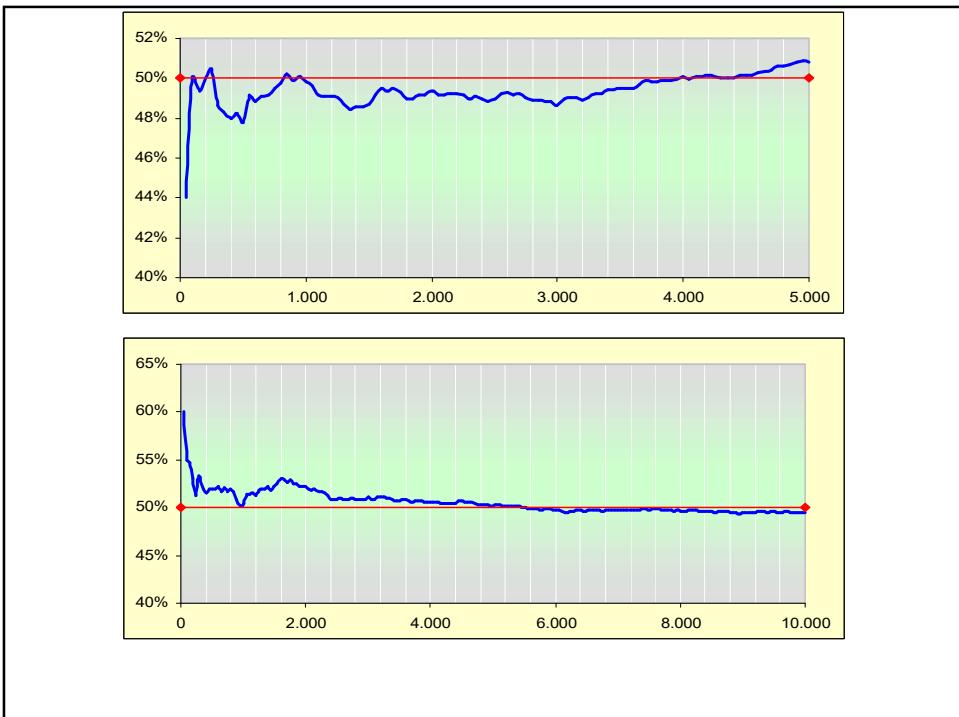
186

$P(A)$	Significado de $P(A)$
1	Sempre ocorre
0,90	Ocorre 90% das vezes e não ocorre em 10% das vezes
0,50	Ocorre 50% das vezes e não ocorre em 50% das vezes
0,15	Ocorre 15% das vezes e não ocorre em 85% das vezes
0	Nunca ocorre

187



188



189

- Os exemplos obtidos com o modelo de simulação mostram que 1.500 ou 10.000 lançamentos podem apresentar resultados parecidos, o que nos faz pensar que a quantidade de lançamentos não tem tamanho ou que há algum conceito que está fugindo ao nosso raciocínio.

190

Voltemos a Peter Bernstein

- **Suponha que você atire uma moeda repetidamente. A lei dos grandes números não diz que a média de suas jogadas se aproximará de 50% à medida que você aumentar o número de jogadas; a matemática elementar diz isso, poupando-lhe a tediosa tarefa de atirar a moeda repetidamente.**
- **Pelo contrário, a lei dos grandes números enuncia que aumentar o número de jogadas aumentará igualmente a probabilidade de que a razão entre as caras e o total de jogadas se desviará de 50% abaixo de uma quantidade especificada, por menor que seja.**

191

- **Não se está em busca da média real de 50%, mas da probabilidade de que o erro entre a média observada e a média real seja inferior a, digamos, 2%; em outras palavras, de que o aumento do número de jogadas aumente a probabilidade de que a média observada não se desvie em mais de 2% da média real.**
- **Isso não significa que não haverá erro após um número infinito de jogadas.**
- **Tudo o que a lei nos informa é que a média de um grande número de jogadas diferirá por menos de que certa quantidade especificada da média real mais provavelmente do que a média de um pequeno número de jogadas.**
- **Além disso, sempre haverá uma possibilidade de que o resultado observado difira da média real por uma quantidade maior do que o limite especificado.**

192

Lei de Benford

- O professor Dr. Theodore P. Hill pede sempre uma lição de casa especial para seus alunos de matemática, no Instituto de Tecnologia da Geórgia.
- Parte deles deve lançar uma moeda duzentas vezes e registrar fielmente seu resultado, enquanto a outra simplesmente deve fingir que jogou a moeda e inventar um resultado para os duzentos supostos arremessos.
- No dia seguinte, para espanto dos alunos, o Professor Hill consegue, com uma breve olhada nos trabalhos, apontar quase todos os que fraudaram os lançamentos.

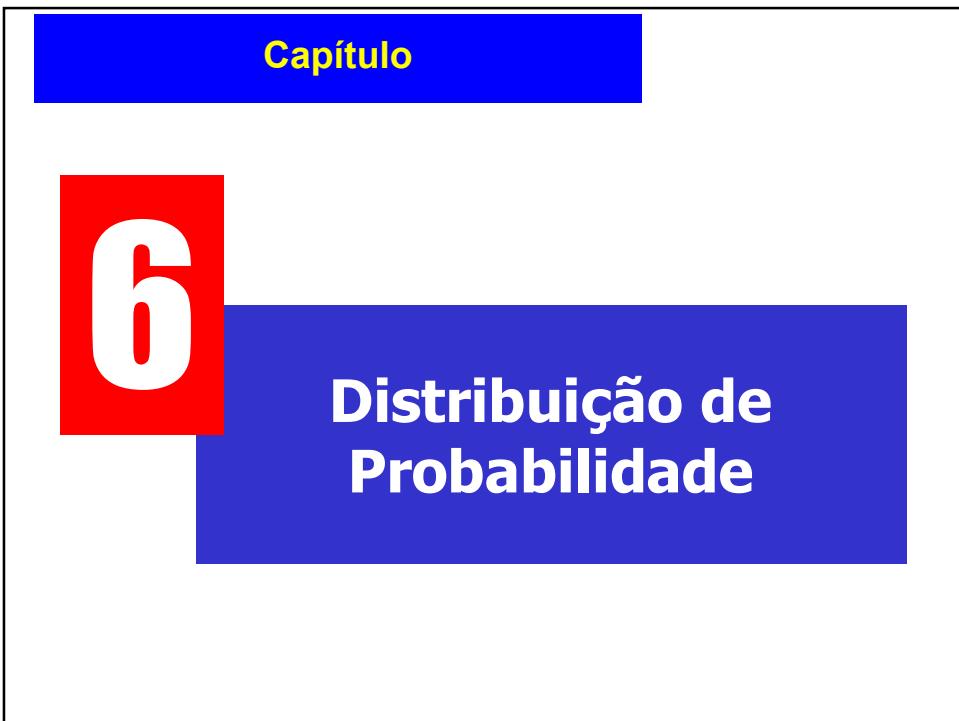
193

- A verdade, disse ele em uma entrevista, é que a maioria das pessoas não sabe quais são as reais probabilidades de um exercício como esse e, portanto, não consegue inventar dados convincentes.
- As previsões de probabilidades são muitas vezes surpreendentes.
- No caso da experiência com o lançamento da moeda, em algum ponto de uma série de duzentos arremessos de moeda, ou cara ou coroa aparecerá seis ou mais vezes seguidas. Aqueles que fraudaram um resultado não sabiam disso e evitaram simular longas seqüências de caras ou coroas, porque, erroneamente, pensaram ser improvável.”
- As seqüências de 0s e 1s do slide seguinte foram obtidas com o Simulador do Lançamento de uma Moeda, numa única simulação.

194

1	1	1
0	1	0
0	0	0
0	0	0
1	1	1
1	1	0
0	1	0
0	1	0
1	0	0
1	0	0
0	1	0
0	1	0
1	1	0
0	1	0
1	1	0
0	1	1
0	0	0
0	0	0
1	0	0
0	1	0
0	0	0
0	0	1
0	0	0
0	0	0

195



196

Recordar é viver ...

- Estatística ... Dados -> Informação
- Variáveis
 - Qualitativas: tabulação
 - Quantitativas: medidas estatísticas
- Medidas usuais
 - Posição central: média
 - Dispersão: desvio-padrão
 - Cuidado sempre presente: Extremos

197

Distribuições teóricas

Distribuição Normal

“Curva normal dos erros”



198

A curva normal dos erros

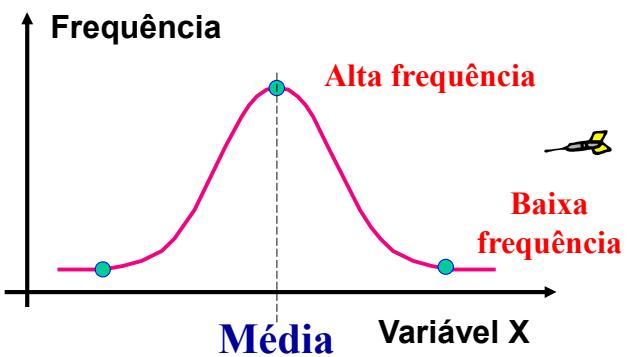
Usando a média e o desvio



199

Médias, desvios e sinos ...

■ Uso da curva normal

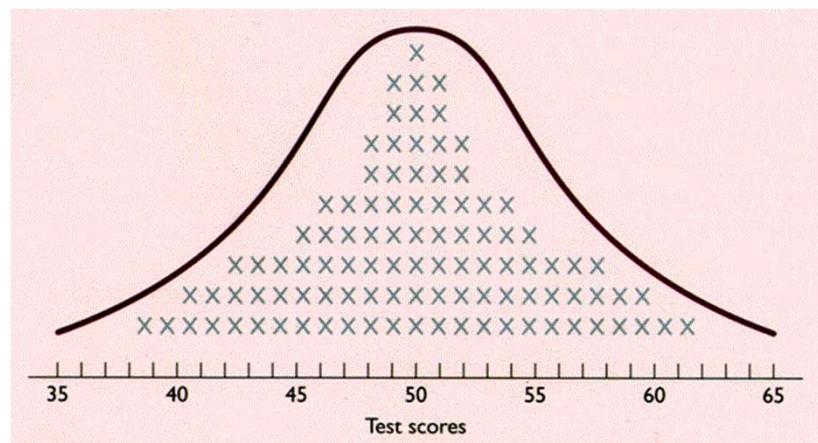


Área sob a curva permite obter as probabilidades

200

100

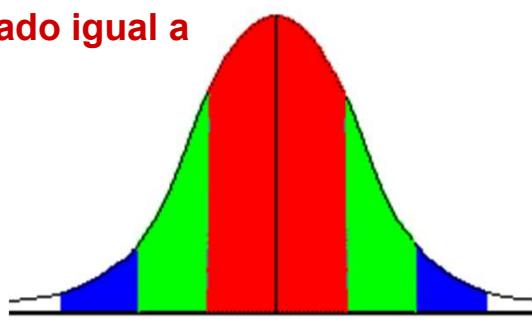
Exemplos de curvas



201

Características da curva

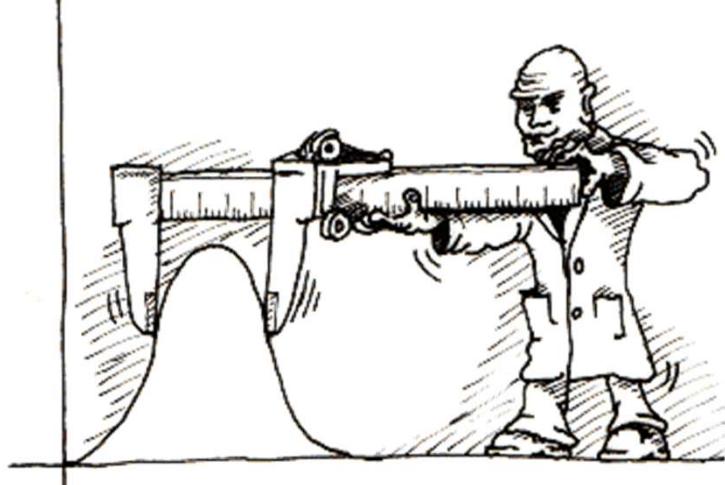
- Na teoria, prolonga-se de – infinito a + infinito
 - Área sob toda a curva igual a 100%
- Simétrica
 - Área de cada lado igual a 50%



202

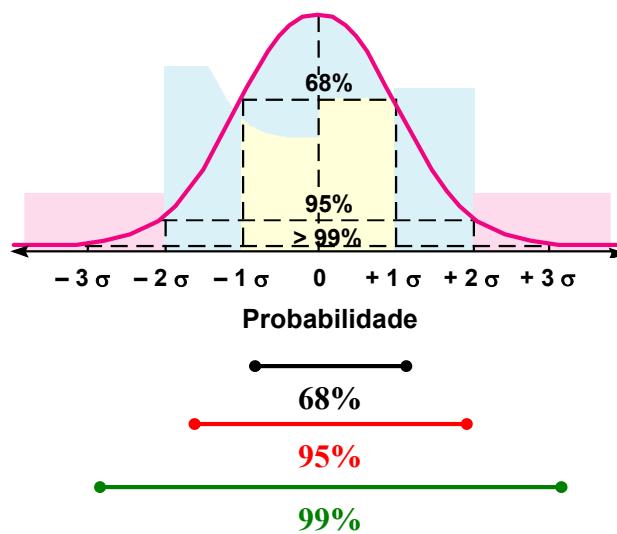
101

Um sino de múltiplos usos



203

Áreas sob a curva normal



204

102

Uma contribuição importante

**Eu encontrei a função
matemática da curva!**

*Áreas sob a curva poderiam
ser obtidas pelo cálculo
das integrais definidas*



Que trabalho!!!

205

Ainda bem!

- Mas ... ainda bem as áreas já estão calculadas em tabelas padronizadas
- Tabelas permitem obter de forma rápida e simples os valores das áreas sob a curva
- Para isso é preciso calcular valores padronizados da variável

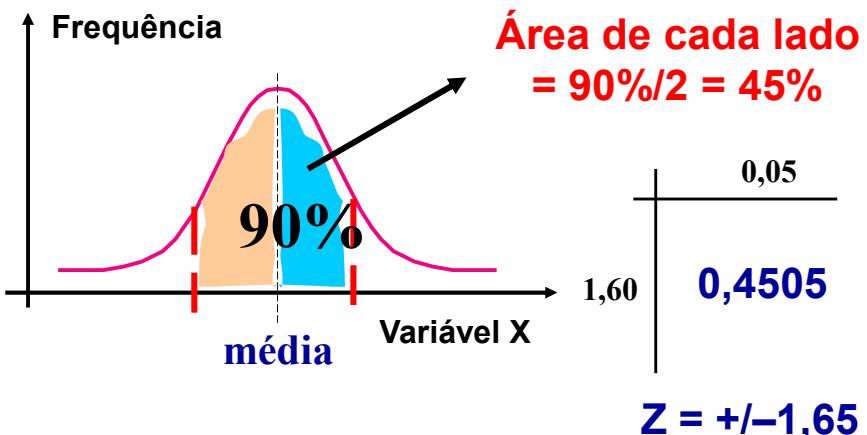


206

103

Um procedimento invertido

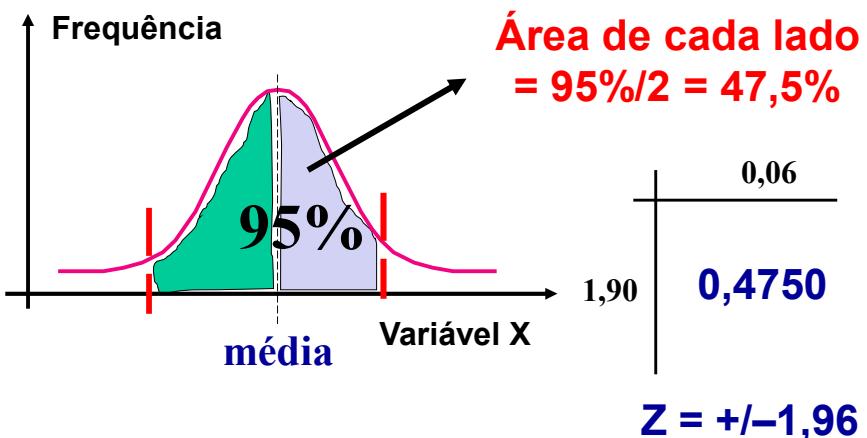
Calcule o valor de Z para área central igual a 90%



207

Um procedimento invertido

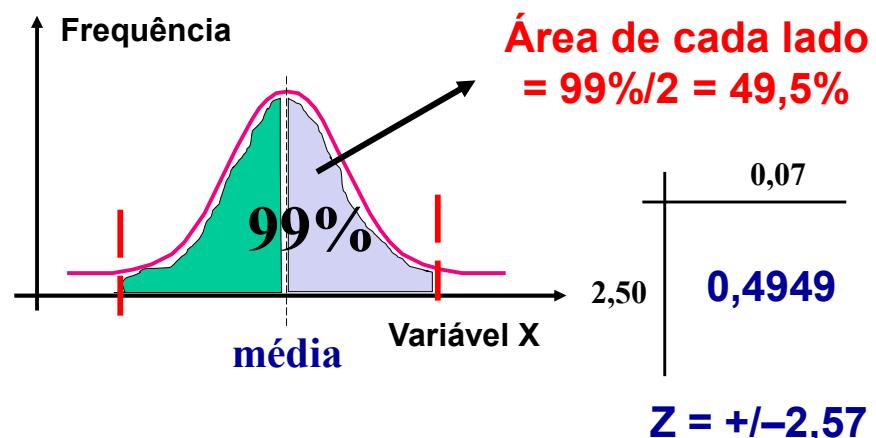
Calcule o valor de Z para área central igual a 95%



208

Um procedimento invertido

**Calcule o valor de Z para área
central igual a 99%**



209

Anexos ...

Distribuição Normal Padronizada



210

105

Tabelas de Z (1)

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,10	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,20	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,30	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,40	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,50	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,60	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,70	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,80	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,90	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389

211

Tabelas de Z (2)

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,00	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,10	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,20	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,30	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,40	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,50	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,60	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,70	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,80	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,90	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767

212

Tabelas de Z (3)

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
2,00	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,10	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,20	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,30	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,40	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,50	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,60	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,70	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,80	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,90	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986

213

Capítulo



Amostragem e Estimação

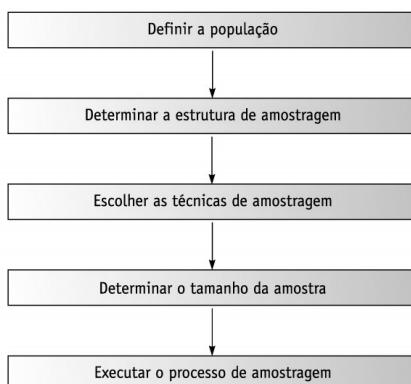
214

Tabela 12.1 Amostra versus censo

	CONDIÇÕES QUE FAVORECEM O USO DE	
	Amostra	Censo
1. Orçamento	Pequeno	Grande
2. Tempo disponível	Curto	Longo
3. Tamanho da população	Grande	Pequeno
4. Variação na característica	Pequena	Grande
5. Custo do erro de amostragem	Baixo	Alto
6. Custo do erro de não-amostragem	Alto	Baixo
7. Natureza da medição	Destrutiva	Não-destrutiva
8. Atenção a casos individuais	Sim	Não

215

Figura 12.3 Processo de elaboração da amostragem



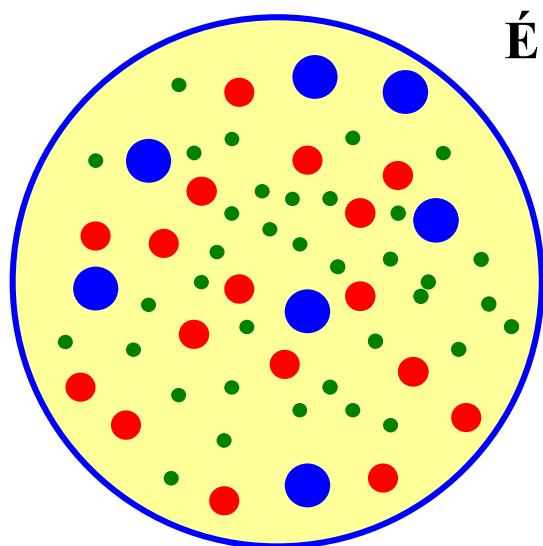
216

Conceito básico

- Inferir significa generalizar
- Com parte do todo (amostra)
- ... tento entender ...
- O próprio todo

217

Uma parte do maior ...



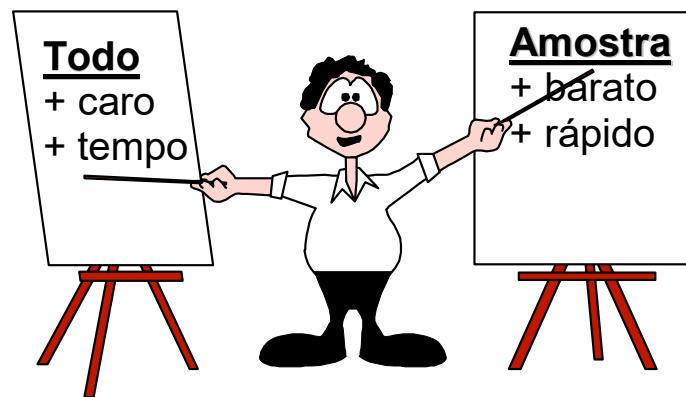
É representativa
do todo?

Amostra

Universo ou População

218

Para pensar ...



219

Amostragem População x Amostra

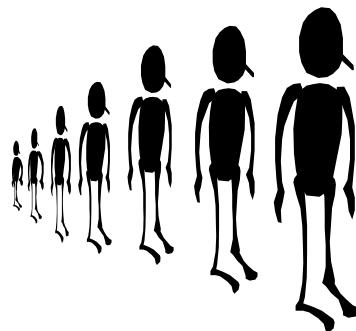
- **População** - conjunto dos elementos que se deseja estudar.
- **Amostra** - subconjunto da população.

220

110

POPULAÇÃO

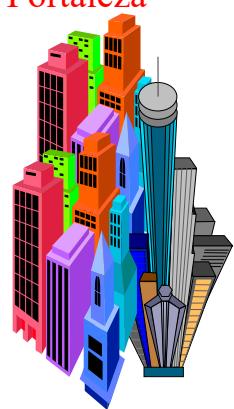
- Conjunto de elementos com pelo menos uma característica em comum observável.



Característica X observável

221

POPULAÇÃO:
moradores de
Fortaleza



AMOSTRA:
uma parte dos
Fortalezenses



222

111

POPULAÇÃO

Devemos definir cuidadosamente os elementos, itens, objetos, etc., que pertencerão a população.

CARACTERÍSTICAS

Incluir características populacionais de real interesse para o estudo. Que atendam os objetivos do estudo.

223

EXEMPLO

- **Objetivo** - estudar a situação econômica dos alunos da Faculdade XPTO.
- População: todos os alunos da Faculdade XPTO.
- **Amostra** - formada de n elementos da população.

224

112

ERRO AMOSTRAL

- Por se estudar só uma parte dos dados (amostra)?
 - Há erros de mensurações?

População

- **Finita** - Consiste de um número finito ou fixo de elementos, medidas, observações. Alunos da Faculdade XPTO, funcionários da Enel (antiga Coelce), eleitores da cidade de Quixeré etc.
- **Infinita** - Pelo menos hipoteticamente, um número infinito de elementos. Nascimentos em uma cidade, produção de uma máquina etc

225

População e Amostra

- **Censo**: Estudo através do exame de todos os elementos da população.
- **Amostragem**: Estudo por meio do exame de uma amostra.

Por que fazer amostragem ao invés de censo?

- Economia
- Menor tempo (atualização)
- Maior qualidade nos dados levantados (precisão)
 - População infinita.
- Mais fácil, com resultados satisfatórios.

226

Quando fazer censo?

- População pequena (tamanho da amostra grande em relação ao da população).
 - Quando se exige o resultado exato.
 - Quando já se dispõe dos dados da população.

227

Figura 12.6 Classificação das técnicas de amostragem



228

Modelos probabilísticos

- Amostragem aleatória simples
- Amostragem sistemática
- Amostragem estratificada



229

Modelos não probabilísticos

- Amostragem accidental ou por conveniência
- Amostragem por julgamento
- Amostragem intencional, proposital ou autogerada



230

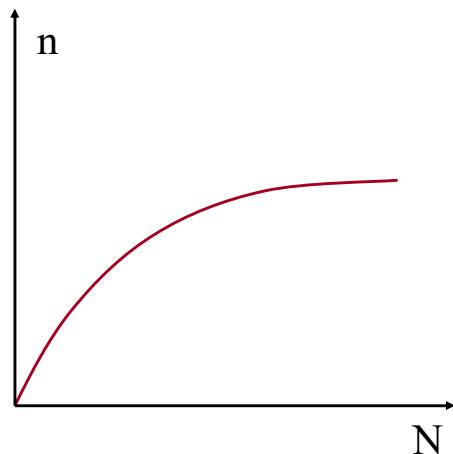
Tabela 12.3 Pontos fortes e fracos das técnicas básicas de amostragem		
TÉCNICA	PONTOS FORTES	PONTOS FRACOS
Amostragem não-probabilística		
Amostragem por conveniência	Menos cara, consome menos tempo, mais conveniente	Tendenciosidade de seleção, amostra não representativa, não recomendada para a pesquisa descritiva ou causal
Amostragem por julgamento	Não é cara, não consome muito tempo e é conveniente	Não permite a generalização, subjetiva
Amostragem por cota	Amostra pode ser controlada para certas características	Tendenciosidade de seleção, nenhuma garantia de representatividade
Amostragem autogerada	Consegue estimar características raras	Consome muito tempo
Amostragem probabilística		
Amostragem aleatória simples (AAS)	De fácil compreensão, resultados projetáveis	Difícil de construir a estrutura de amostragem, cara, baixa precisão, nenhuma garantia de representatividade
Amostragem sistemática	Pode aumentar a representatividade, mais fácil de implementar do que a AAS, estrutura de amostragem não é necessária	Pode diminuir a representatividade
Amostragem estratificada	Inclui todas as subpopulações importantes, precisão	Difícil de escolher variáveis relevantes de estratificação, não é viável estratificar com muitas variáveis, cara
Amostragem por grupo	Fácil de implementar, eficaz no custo	Imprecisa, difícil de computar e de interpretar os resultados

231

Tabela 12.4 Escolhendo a amostragem não-probabilística em comparação com a probabilística		
FATORES	CONDIÇÕES QUE FAVORECEM O USO DE	
	AMOSTRAGEM NÃO-PROBABILÍSTICA	AMOSTRAGEM PROBABILÍSTICA
Natureza da pesquisa	Exploratória	Conclusiva
Magnitude relativa dos erros de amostragem e de não-amostragem	Erros de não-amostragem são maiores	Erros de amostragem são maiores
Variabilidade da população	Homogênea (baixa)	Heterogênea (alta)
Considerações estatísticas	Desfavoráveis	Favoráveis
Considerações operacionais	Favoráveis	Desfavoráveis

232

Tamanho da amostra (n)
e Tamanho da população (N)

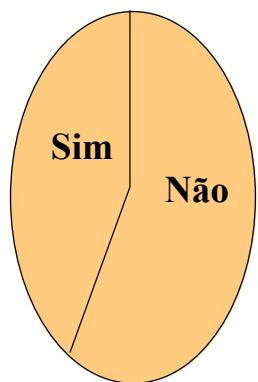


233

Amostragem

A amostra deve ser *representativa*!

População



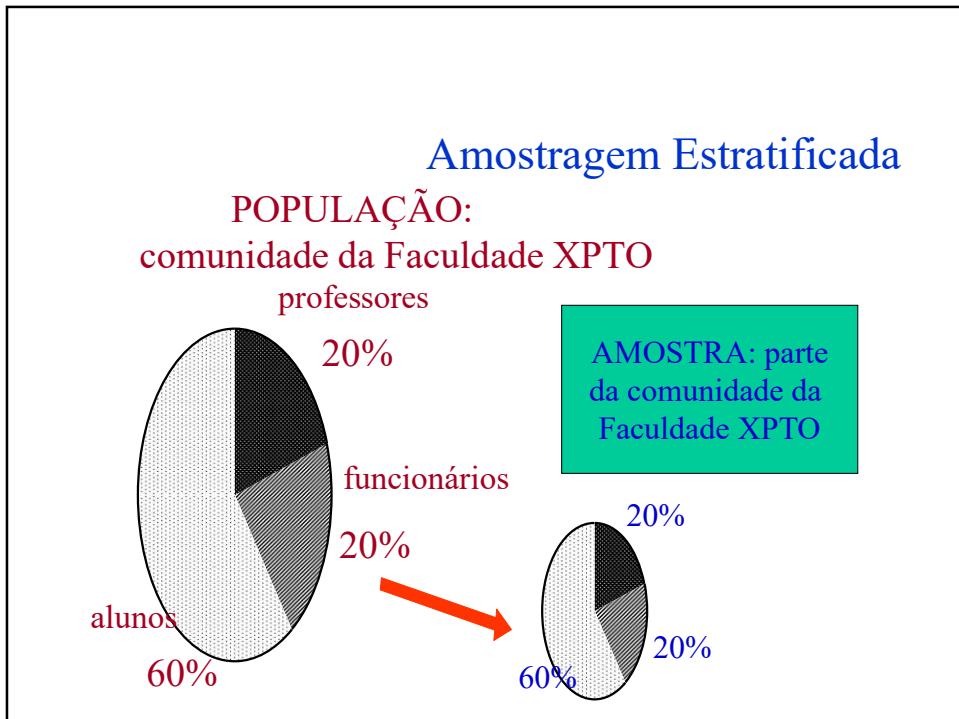
Plano de
amostragem



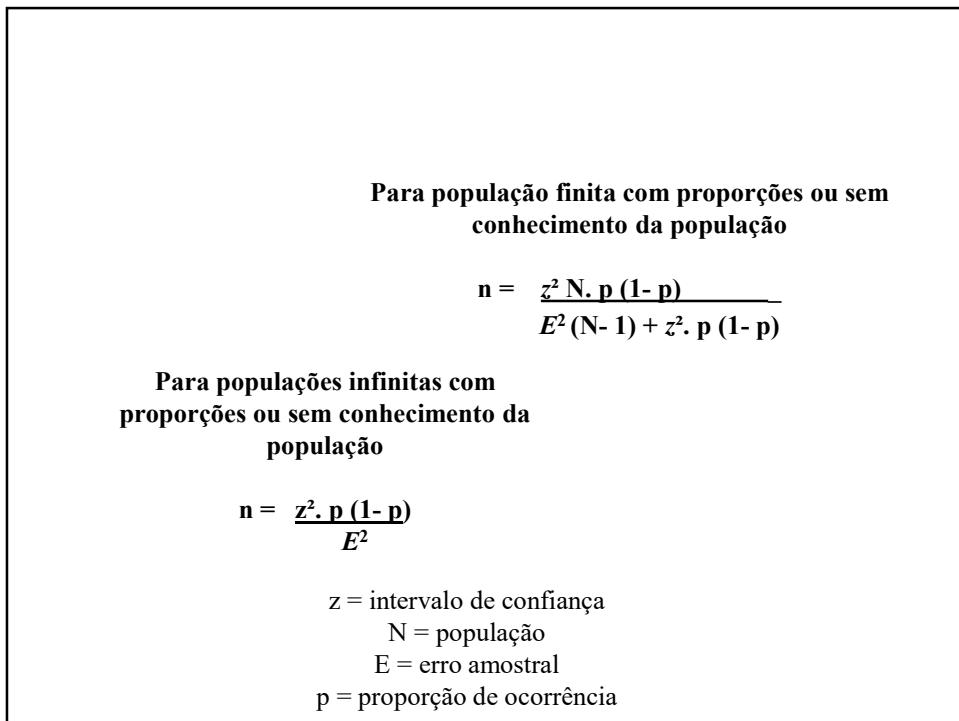
Amostra



234



235



236

Vamos calcular!

237

Símbolos de diferentes médias

\bar{x}

Amostra



μ

População

238

119

Sendo amostra representativa ...

É possível
calcular o erro
inferencial!!!

*Afastamento da
medida da amostra
em relação à
medida do todo*



239

Pesquisa Ibope sobre o Referendo (14.10.2005)

- A pesquisa quis saber como os eleitores responderiam à pergunta: "O comércio de armas de fogo e munição deve ser proibido no país?". Os que disseram Não à pergunta somaram 49%. E os que disseram Sim, 45%. Não souberam ou não opinaram 6%.
- Segundo o Ibope, o eleitorado está dividido entre proibir ou manter o comércio legal de armas de fogo e munição no Brasil. Considerando a **margem de erro de 2,2 pontos percentuais** para mais ou para menos, o resultado está no limite do empate técnico. Com a margem de erro, o Não, que aparece com 49%, ficaria entre 46,8% e 51,2%. O Sim, com 45%, ficaria entre 42,8% e 47,2%.
- O Ibope ouviu 2.002 mil eleitores entre os dias 11 e 13 de outubro. A pesquisa foi registrada no Tribunal Superior Eleitoral com o número 1.688.

240

Ainda as notícias ...

- Ibope: Com menos vantagem, Candidato A ainda lidera pesquisas

**Fortaleza/Ibope: Pesquisa dá vitória ao Candidato B
... A margem de erro é de três pontos percentuais para mais ou para menos.
O Ibope ouviu 1.068 eleitores.**

241

Calculando o erro inferencial

- **Erro será função:**
 - a) Tamanho da amostra**
 - b) Dispersão dos dados**
 - c) Nível de confiança desejado para o estudo**



242

Calculando o erro inferencial

Variáveis quantitativas



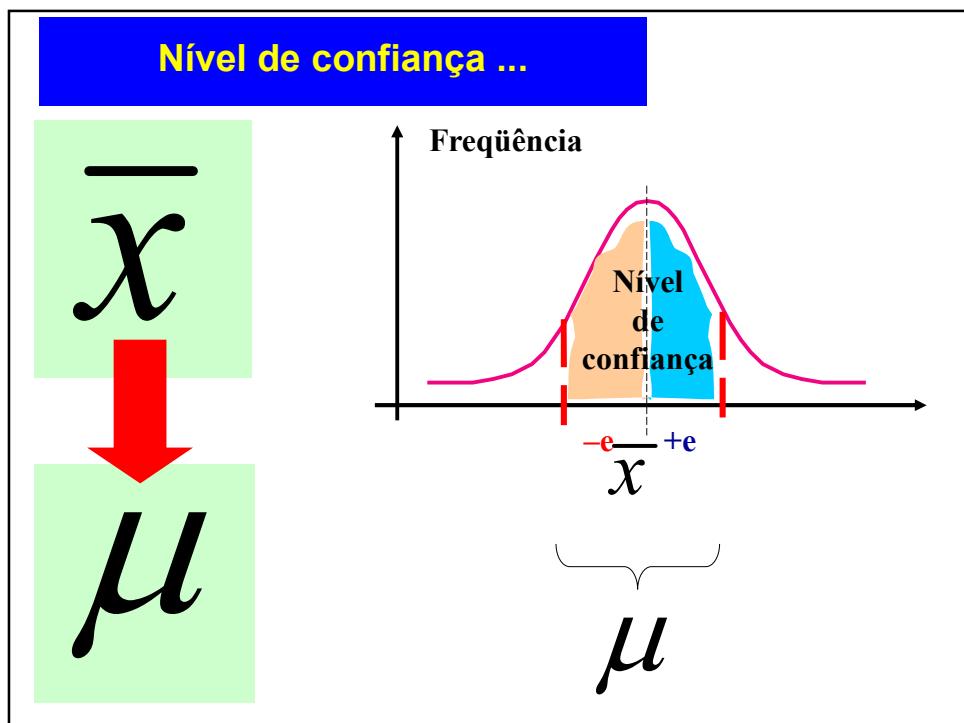
243

Entendendo o erro

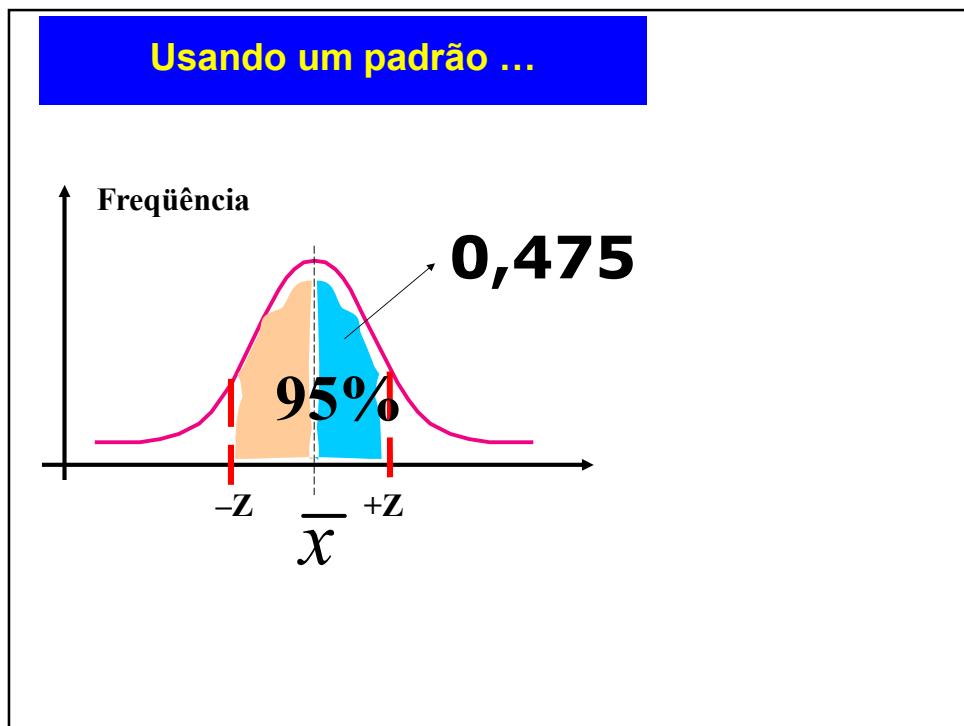
**Tamanho
Nível de confiança
Dispersão dos dados**

$$e = Z \frac{\sigma}{\sqrt{n}}$$

244



245

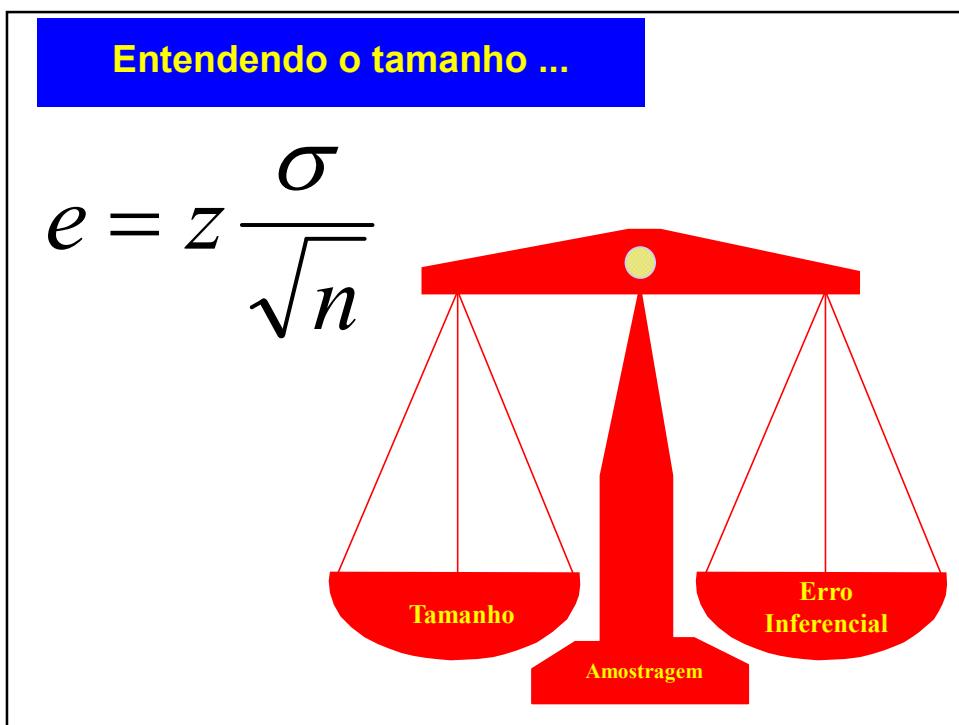


246

Tabelas de Z											
Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	
1,00	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621	
1,10	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830	
1,20	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015	
1,30	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177	
1,40	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319	
1,50	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441	
1,60	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545	
1,70	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633	
1,80	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706	
1,90	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767	

Para NC = 95%, Z = +/− 1,96

247



248

Tamanho e universo

N	Erro inferencial									
	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
Nível de confiança igual a 95%										
10	10	10	10	10	10	10	10	10	10	10
50	50	49	48	47	45	43	40	38	36	34
100	99	97	92	86	80	73	67	61	55	50
250	244	227	203	177	152	130	111	95	81	70
500	476	414	341	274	218	175	142	116	96	81
1.000	906	707	517	376	278	211	165	131	107	88
10.000	4900	1937	965	567	370	260	193	148	118	96
100.000	8763	2345	1056	597	383	267	196	150	119	96
1.000.000	9513	2396	1066	600	384	267	196	151	119	97
10.000.000	9595	2401	1067	601	385	267	196	151	119	97

249

Exemplo 1.5

- O objetivo da auditoria interna da empresa é verificar se o Setor de Contas a Pagar cumpre com as rotinas estabelecidas pela empresa para pagamento de fornecedores. Deve-se estabelecer o procedimento de seleção de quinze processos dos últimos seiscentos realizados.

250

Exemplo 1.7

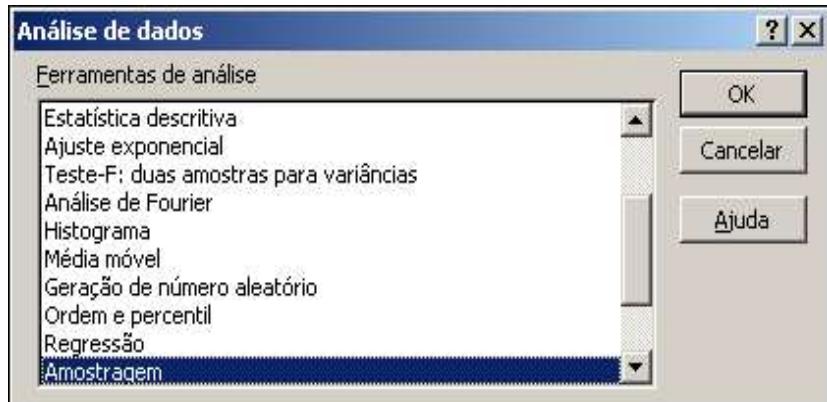
- A tabela seguinte registra a relação das 50 Maiores Empresas Privadas por Vendas do Brasil. O objetivo é retirar uma amostra aleatória sem reposição de tamanho 10 utilizando a tabela de números aleatórios. A tabela das maiores empresas está registrada na planilha 50 Maiores da Exame.

251

	A	B	C	D
1	As 50 primeiras empresas privadas por vendas em 2002 - Revista Exame			
2				
3	Ordem	Empresa - Ramo		Vendas
4	1	TELEMAR - Telecomunicações		\$ 6.303,7
5	2	TELEFÔNICA - Telecomunicações		\$ 5.480,5
6	3	CBB/AMBEV - Alimentos, bebidas e fumo		\$ 5.329,8
7	4	VOLKSWAGEN - Automotivo		\$ 5.295,2
8	5	PETRÓLEO IPIRANGA - Atacado e comércio exterior		\$ 4.214,1
9	6	SHELL - Atacado e comércio exterior		\$ 4.096,8
10	7	GEN	E F G	H
11	8	CAR	1	I
12	9	BRA	2	
13	10	GRU	3	
14	11	EMB	4	
15	12	VALE	5	
16	13	BUN	6	
17	14	FIAT	7	
18	15	ELET	8	
			9	
			10	
			11	
			12	
			13	
			14	
Amostragem com Reposição				
Amostra	Empresa - Ramo		Vendas	
1	34	CASAS BAHIA - Comércio varejista		\$ 1.690,70
2	40	COPESUL - Química e petroquímica		\$ 1.465,80
3	8	CARREFOUR - Comércio varejista		\$ 4.044,90
4	14	FIAT - Automotivo		\$ 3.121,40
5	5	PETRÓLEO IPIRANGA - Atacado e comércio exterior		\$ 4.214,10
6	18	NESTLÉ - Alimentos, bebidas e fumo		\$ 2.762,70
7	5	PETRÓLEO IPIRANGA - Atacado e comércio exterior		\$ 4.214,10
8	47	SONAE - Comércio varejista		\$ 1.156,50
9	23	FORD MOTOR - Automotivo		\$ 2.387,60
10	32	SADIA - Alimentos, bebidas e fumo		\$ 1.760,40

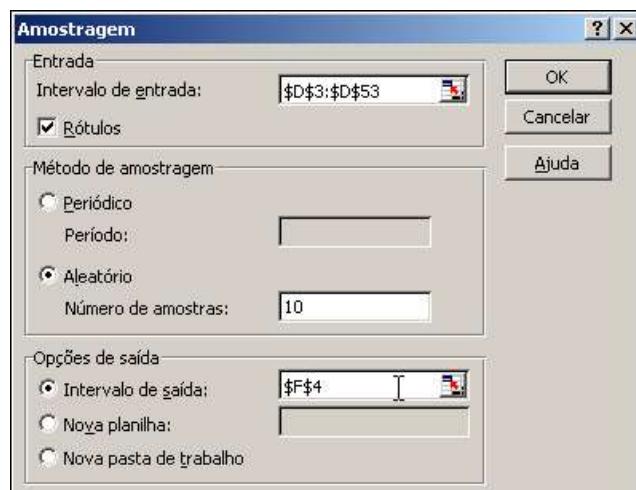
252

FERRAMENTAS DE ANÁLISE



253

F de A - Histograma



254

A	B	C	D	E	F	G	H
As 50 primeiras empresas privadas por vendas em 2002 - Revista Exame				Ferramenta de Análise Amostragem			
3	Ordem	Empresa - Ramo	Vendas	Vendas			
4	1	TELEMAR - Telecomunicações	\$ 6.303,7	5480,5			
5	2	TELEFÔNICA - Telecomunicações	\$ 5.480,5	4096,8			
6	3	CBB/AMBEV - Alimentos, bebidas e fumo	\$ 5.329,8	1127,2			
7	4	VOLKSWAGEN - Automotivo	\$ 5.295,2	6303,7			
8	5	PETRÓLEO IPIRANGA - Atacado e comércio exterior	\$ 4.214,1	3837,5			
9	6	SHELL - Atacado e comércio exterior	\$ 4.096,8	1465,8			
10	7	GENERAL MOTORS - Automotivo	\$ 4.092,7	5295,2			
11	8	CARREFOUR - Comércio varejista	\$ 4.044,9	2805,2			
12	9	BRASIL TELECOM - Telecomunicações	\$ 3.975,9	1886,1			
13	10	GRUPO PÃO DE AÇÚCAR - Comércio varejista	\$ 3.837,5	1465,8			
14	11	EMBRATEL - Telecomunicações	\$ 3.668,3				
15	12	VALE DO RIO DOCE - Mineração	\$ 3.418,0				
16	13	BUNGE ALIMENTOS - Alimentos, bebidas e fumo	\$ 3.158,1				
17	14	FIAT - Automotivo	\$ 3.121,4				
18	15	ELETROPAULO METROPOLITANA - Serviços públicos	\$ 3.078,0				
19	16	EMBRAER - Automotivo	\$ 2.945,3				
20	17	TEXACO - Atacado e comércio exterior	\$ 2.805,2				
21	18	NESTLÉ - Alimentos, bebidas e fumo	\$ 2.762,7				
22	19	CARGILL - Alimentos, bebidas e fumo	\$ 2.709,1				
23	20	ESSO - Atacado e comércio exterior	\$ 2.688,5				
24	21	ITAIPÚ BINACIONAL - Serviços públicos	\$ 2.529,6				
25	22	UNILEVER - Farmacêutico, higiene e cosméticos	\$ 2.456,9				
26	23	FORD MOTOR - Automotivo	\$ 2.387,6				
27	24	SOUZA CRUZ - Alimentos, bebidas e fumo	\$ 2.375,9				

255

AMOSTRAGEM SEM REPOSIÇÃO

Exemplo 1.9

- Construir um modelo para extrair uma amostra probabilística simples sem reposição de dez empresas da tabela das cinquenta primeiras empresas privadas por vendas.

256

A	B	C	D	E	F	G	H	I
1	As 50 primeiras empresas privadas por vendas em 2002 - Revista Exame							
3	4	5	6	7	8	9	10	11
Ordem	Empresa - Ramo	Vendas						
1	TELEMAR - Telecomunicações	\$ 6.303,7						
2	TELEFÔNICA - Telecomunicações	\$ 5.480,5						
3	CBB/AMBEV - Alimentos, bebidas e fumo	\$ 5.329,8						
4	VOLKSWAGEN - Automotivo	\$ 5.295,2						
5	PETRÓLEO IPIRANGA - Atacado e comércio exterior	\$ 4.214,1						
6	SHELL - Atacado e comércio exterior	\$ 4.096,8						
7	GENERAL MOTORS - Automotivo	\$ 4.092,7						
8	CARREFOUR - Comércio varejista	\$ 4.044,9						
9	BRASÍLIA TELECOM - Telecomunicações	\$ 3.975,9						
10	GRUPO PÃO DE AÇÚCAR - Comércio varejista	\$ 3.837,5						
11	EMBRATEL - Telecomunicações	\$ 3.668,3						
12	VALE DO RIO DOCE - Mineração	\$ 3.418,0						

Amostragem sem Reposição			Nova Amostragem
Amostra	Empresa - Ramo	Vendas	
1	ITAIPI BINACIONAL - Serviços públicos	\$ 2.529,6	
2	KLABIN PAPEL CELULOSE - Papel e celulose	\$ 1.155,1	
3	SHELL - Atacado e comércio exterior	\$ 4.096,8	
4	BUNGE FERTILIZANTES - Química e petroquímica	\$ 1.297,5	
5	BRASIKEM - Química e petroquímica	\$ 1.793,3	
6	NESTLÉ - Alimentos, bebidas e fumo	\$ 2.762,7	
7	CARREFOUR - Comércio varejista	\$ 4.044,9	
8	SOUZA CRUZ - Alimentos, bebidas e fumo	\$ 2.375,9	
9	EMBRATEL - Telecomunicações	\$ 3.668,3	
10	PONTO FRIÓ - Comércio varejista	\$ 1.153,3	

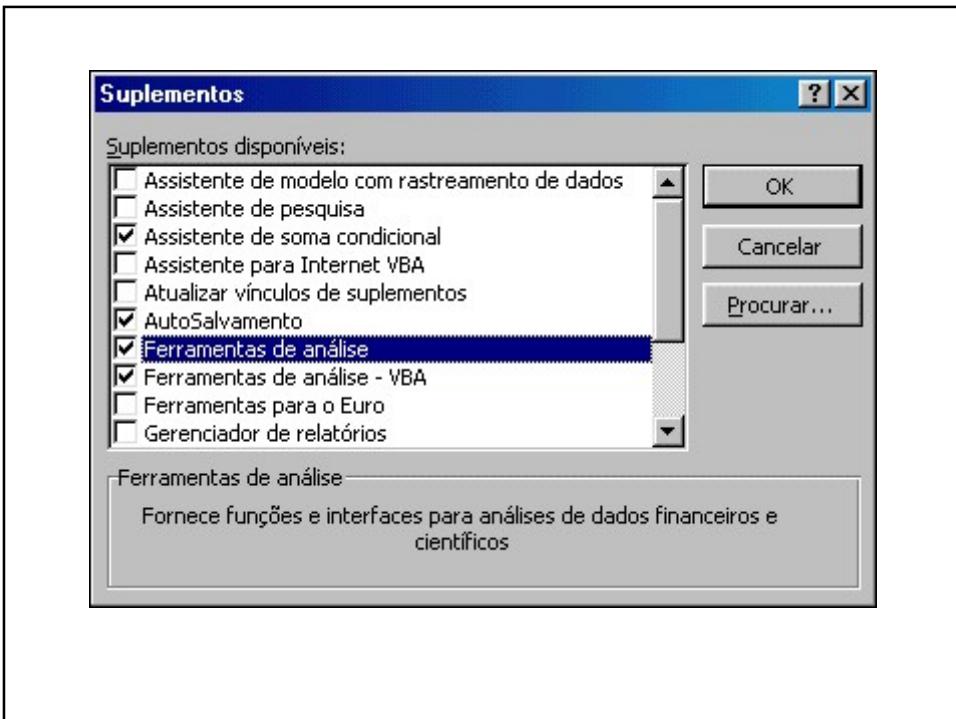
257

PREPARANDO O EXCEL

Excel

- No menu Ferramentas, escolha Suplementos. O Excel apresentará a caixa de diálogo Suplementos com os Suplementos disponíveis.
- Os suplementos Ferramentas de análise e Ferramentas de análise-VBA devem estar selecionados, como mostra a figura seguinte.
- Aproveite e também selecione o suplemento Solver.

258

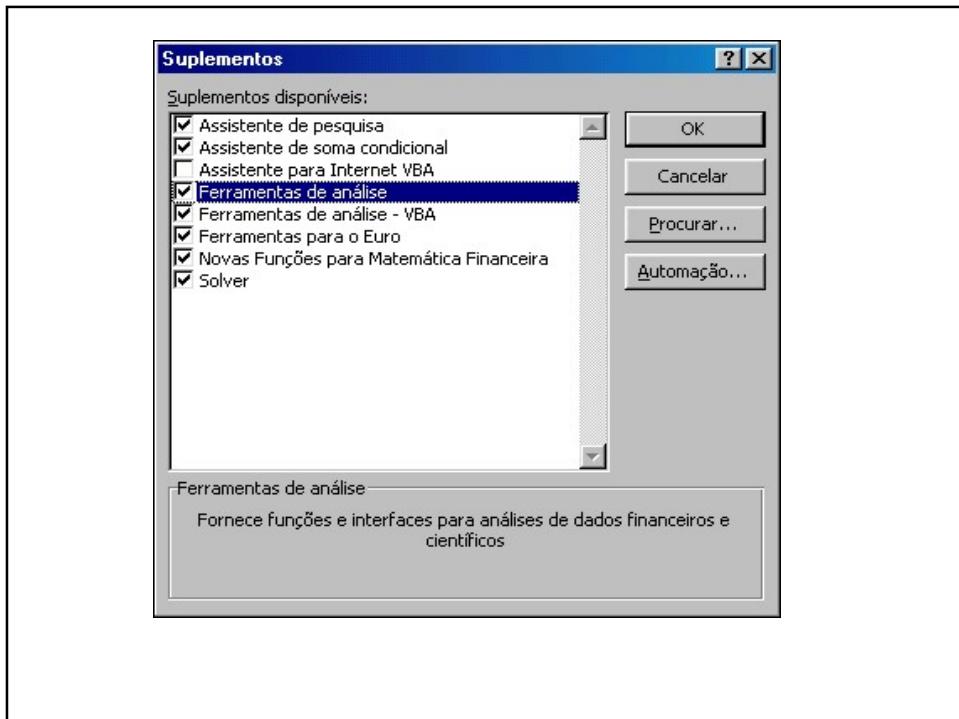


259

Excel versões 2007 e 2010.

- No menu Arquivos/Opções escolha Suplementos. O Excel apresentará a caixa de diálogo Suplementos com os Suplementos disponíveis.
- Os suplementos Ferramentas de análise e Ferramentas de análise-VBA devem estar selecionados, como mostra a figura seguinte. Depois de clicar em OK, as ferramentas de análise, bem como as funções especiais, estarão sempre disponíveis quando o aplicativo Excel for carregado.
- Aproveite e também selecione o suplemento Solver que será utilizado neste livro.

260



261

Para todas as versões do Excel.

- **Se os suplementos Ferramentas de análise, Ferramentas de análise-VBA e Solver não aparecerem na caixa de diálogo Suplementos, então os dois suplementos não foram instalados junto com o Excel. O leitor deverá instalar esses arquivos incluídos no programa de instalação do Excel ou Microsoft Office correspondente.**

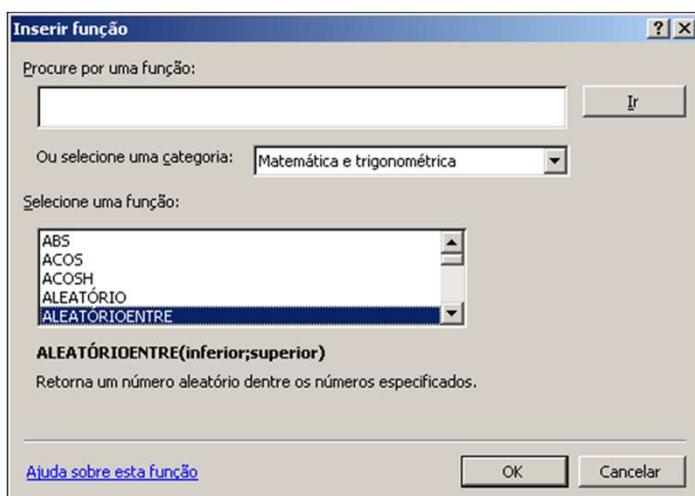
262

PREPARANDO O EXCEL



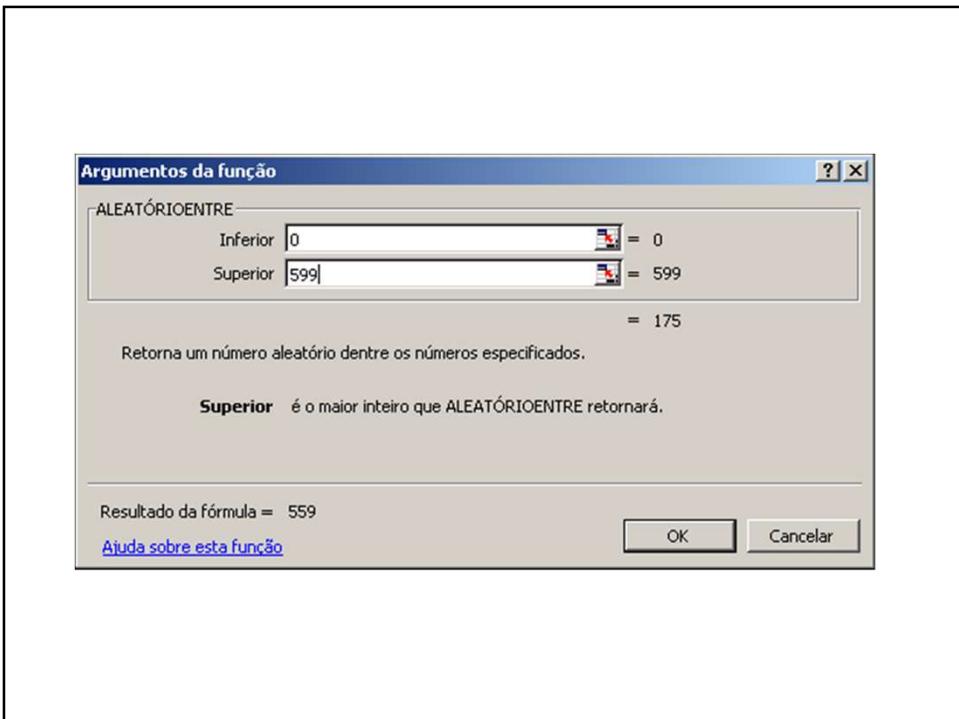
263

COMO REGISTRAR UMA FUNÇÃO



264

132



265

Amostragem sem Reposição									
	A	B	C	D	E	F	G	H	I
1 As 50 primeiras empresas privadas por vendas em 2002 - Revista Exame									
2									
3	Ordem	Empresa - R\$	Vendas		N.Aleat.		Amostra selecionada		
4	1	TELEMAQ	\$ 6.303,7		0,6860785		TEXACO - Atacado e comércio exterior	\$ 2.805,2	
5	2	TELEFÔNICA	\$ 5.480,5		0,1880333		COPERSUCAR - Atacado e comércio exterior	\$ 1.550,5	
6	3	CBBIA/AME	\$ 5.329,8		0,8861437		VOLKSWAGEN - Automotivo	\$ 5.295,2	
7	4	VOLKSW	\$ 5.295,2		0,0290058		PONTO Frio - Comércio varejista	\$ 1.153,3	
8	5	PETRÓLIO	\$ 4.214,1		0,0773903		SONAE - Comércio varejista	\$ 1.156,5	
9	6	SHELL -	\$ 4.096,8		0,261369		TELESP CELULAR - Telecomunicações	\$ 1.752,1	
10	7	GEREPA	\$ 4.092,7		0,1166721		COSIPA - Siderurgia e metalurgia	\$ 1.340,0	
11	8	CARREFOUR	\$ 4.044,9		0,0482277		KLabin PAPEL CELULOSE - Papel e celulose	\$ 1.155,1	
12	9	BRASIL	\$ 3.975,9		0,8553086		SHELL - Atacado e comércio exterior	\$ 4.096,8	
13	10	GRUPO	\$ 3.837,5		0,19673		CPFL - Serviços públicos	\$ 1.551,2	
14	11	EMBRAT	\$ 3.668,3		0,2965299				

266

Capítulo

8

Correlação e Regressão Linear

267

Duas variáveis quantitativas

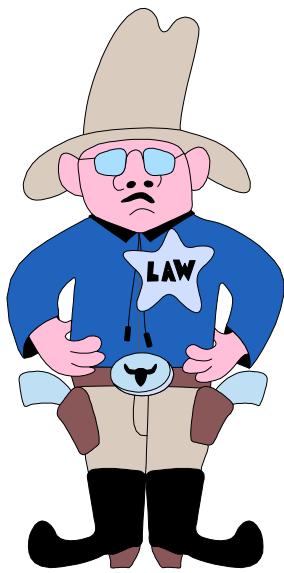
X

independente



Y

dependente



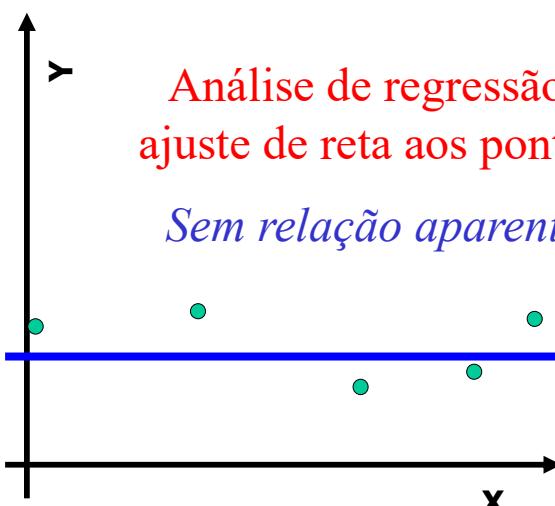
268

134

Sem relação ...

Análise de regressão:
ajuste de reta aos pontos

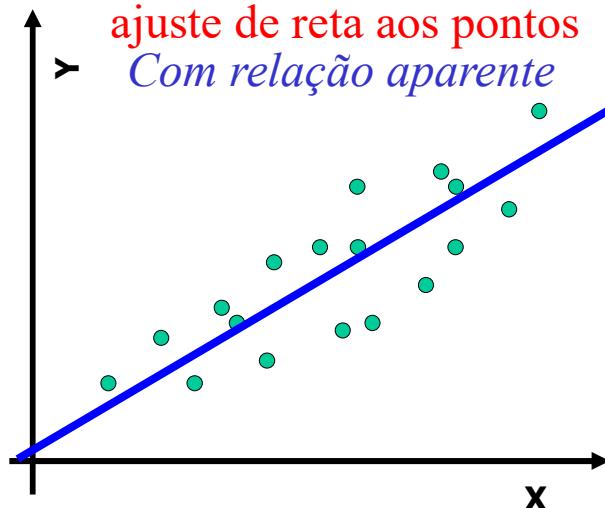
Sem relação aparente



269

Com relação

Análise de regressão:
ajuste de reta aos pontos
Com relação aparente



270

135

Cuidado com as variáveis!

Bebeu ...



refrigerante com vodka ficou bêbado
refrigerante com whisky ficou bêbado
refrigerante com vinho ficou bêbado
refrigerante com conhaque ficou bêbado
refrigerante com cerveja ficou bêbado
refrigerante com tequila ficou bêbado

...logo ...

271

- O coeficiente de correlação não mede a relação causa-efeito entre duas variáveis, apesar de que essa relação possa estar presente.
 - Por exemplo, uma correlação fortemente positiva entre as variáveis X e Y não significa 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; entretanto, numa regressão linear a relação causa-efeito deve ser definida no início da análise.
- Este inicia com a apresentação da relação linear simples entre duas amostras ou variáveis aleatórias.

272

136

- Na regressão linear simples será deduzida e analisada a reta que melhor explica essa relação, tendo previamente definido a variável independente e a variável dependente.
- Todos os dias, a *mídia* se encarrega de informar resultados de análises e pesquisas do tipo:
 - O valor da empresa depende do lucro futuro, a taxa de juro depende da inflação.
 - O salário depende da escolaridade do trabalhador etc.

273

- O objetivo da análise de regressão é encontrar uma função linear que permita:
 - Descrever e compreender a relação entre uma variável dependente e uma ou mais variáveis independentes.
 - Projetar ou estimar uma variável em função de uma ou mais variáveis independentes; por exemplo, as vendas para diferentes valores de investimento em propaganda, a demanda em função do preço unitário e do investimento em propaganda etc.

274

Exemplo 8.1

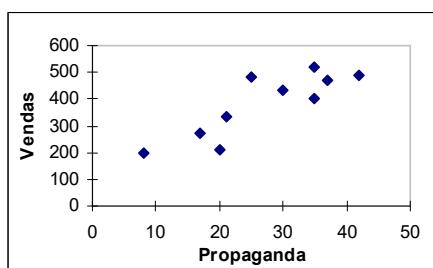
- O objetivo do diretor de vendas de uma rede de varejo é analisar a relação entre o investimento realizado em propaganda e as vendas das lojas da rede, para realizar projeções de vendas de futuros investimentos em propaganda.
- A tabela seguinte registra uma amostra representativa extraída dos registros históricos das lojas de tamanho equivalente, com os valores de Propaganda e Vendas em milhões.
- Analisar a possibilidade de definir um modelo que represente a relação entre as duas variáveis ou amostras.

Propaganda	30	21	35	42	37	20	8	17	35	25
Vendas	430	335	520	490	470	210	195	270	400	480

275

Solução

- Para analisar a relação entre as duas variáveis na planilha Exemplo 8.1, foi construído o gráfico de dispersão das vendas anuais em função do investimento anual em propaganda. Nesse gráfico pode-se ver que, nos últimos dez anos, o aumento de investimento em propaganda gerou aumento das vendas, e vice-versa.



276

- O gráfico de dispersão mostra que as vendas e o investimento em propaganda estão correlacionados de forma positiva, com um coeficiente de correlação próximo de +1.
- Uma reta como a linha tracejada no gráfico de dispersão acima poderá ser utilizada para realizar projeções das vendas futuras em função do investimento em propaganda.
 - A linha tracejada foi ajustada tentando equilibrar os pontos acima da reta com os pontos abaixo dela.
 - Essa reta é uma das muitas possíveis retas que poderiam ser ajustadas.

277

Modelo do Ajuste de uma Reta

- O ajuste de uma reta é um modelo linear que relaciona a variável dependente y e a variável independente x por meio da equação de uma reta do tipo: $y = a + bx$
- É importante observar que, da mesma forma como a média resume uma variável aleatória, a reta de regressão resume a relação linear entre duas variáveis aleatórias e, consequentemente, da forma como a média varia entre amostras do mesmo tamanho extraídas da mesma população, as retas também variarão entre amostras da mesma população.

278

- O objetivo do Exemplo 8.1 é ajustar uma reta a partir dos valores das amostras retiradas da população, considerando que o investimento em propaganda é a variável independente x , e as vendas anuais, a variável dependente y .
- Uma primeira forma de fazer isso é ajustar manualmente essa reta tentando equilibrar os pontos acima e abaixo dessa reta, como foi feito no gráfico do Exemplo 8.1.
- Como esse procedimento permite o ajuste de diversas retas, é necessário estabelecer um objetivo de eficiência de ajuste possível de medir, como é mostrado a seguir.

279

- Uma primeira forma é ajustar uma reta horizontal de valor igual à média dos valores da variável dependente y , que é uma reta de regressão com $b=0$.
 - Esse critério não necessita de regressão, entretanto, será uma referência útil para medir o grau de explicação da reta de regressão.
- Outra forma é ajustar uma reta que divida os pontos observados de forma que a soma dos desvios seja nula.
 - Entretanto, como há muitas retas que cumprem com essa condição, esse critério não poderá ser utilizado.
- Outra forma é ajustar uma reta de forma que minimize a soma dos quadrados dos desvios, lembrando a definição de variância.

280

Coeficientes de Regressão

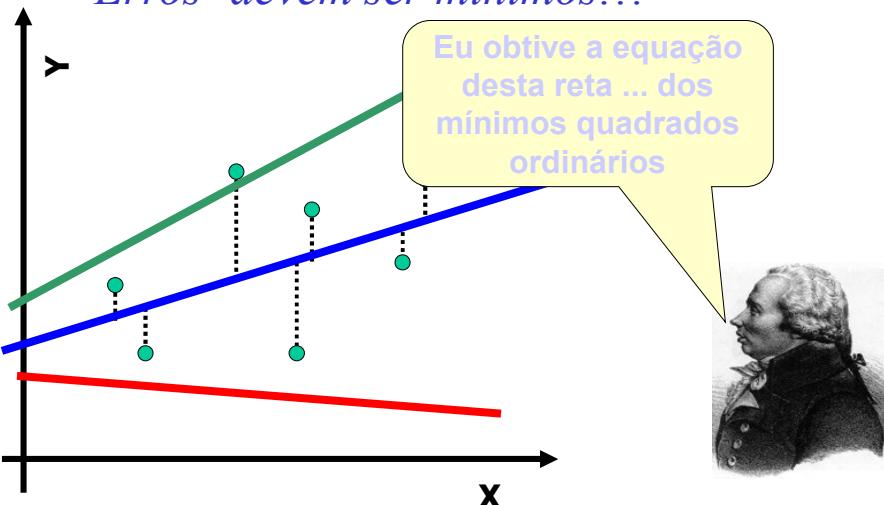
A reta de regressão é representada pela equação $\hat{y} = a + bx$, sendo \hat{y} a variável dependente e x a variável independente. Os coeficientes a e b são os coeficientes de regressão com o seguinte significado:

- O coeficiente b é a declividade da reta e define o aumento ou diminuição da variável y por unidade de variação da variável x .
- A constante a é o intercepto y , sendo igual ao valor de \hat{y} para $x=0$.

281

Erros quadráticos mínimos

Erros² devem ser mínimos!!!



282

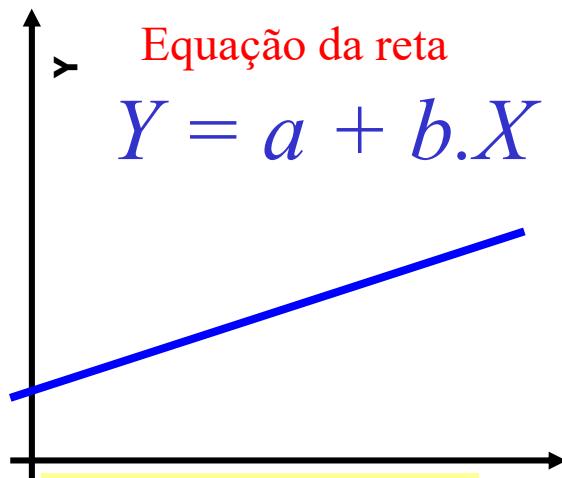
Legendre, Adrien-Marie (1752-1833)

Matemático francês, discípulo de Euler e Lagrange. É autor de um clássico trabalho de geometria, *Éléments de géométrie*. Também fez importantes contribuições em equações diferenciais, cálculo, teoria das funções e teoria dos números.



283

Definindo a equação da reta



Equação da reta

$$Y = a + b \cdot X$$

É preciso obter os somatórios!

$$a = \frac{\sum y - b \sum x}{n}$$

$$b = \frac{n(\sum xy) - (\sum x \sum y)}{n(\sum x^2) - (\sum x)^2}$$



284

142

Função do Excel

INTERCEPÇÃO(val_conhecidos_y; val_conhecidos_x)

- A função estatística INTERCEPÇÃO retorna o coeficiente de regressão a da reta de regressão linear considerando os valores das amostras informados nos argumentos *val_conhecidos_y* e *val_conhecidos_x*.
 - **Ao utilizar essa função, deve-se tomar o cuidado de fornecer os valores na ordem correta, o primeiro argumento *val_conhecidos_y* se refere aos valores da variável dependente *y*, e o argumento *val_conhecidos_x*, aos valores da variável independente *x*.**
 - **Os dois argumentos desta função devem ser números ou nomes, matrizes ou referências que contenham números.**

285

Função do Excel

INCLINAÇÃO(val_conhecidos_y; val_conhecidos_x)

- A função estatística INCLINAÇÃO retorna o coeficiente b da reta de regressão linear considerando os valores das amostras informados nos argumentos *val_conhecidos_y* e *val_conhecidos_x*.
 - **Ao utilizar esta função, deve-se tomar o cuidado de fornecer os valores na ordem correta, o primeiro argumento *val_conhecidos_y* se refere aos valores da variável dependente *y*, e o argumento *val_conhecidos_x*, aos valores da variável independente *x*.**
 - **Os dois argumentos desta função devem ser números ou nomes, matrizes ou referências que contenham números.**

286

Uma análise de vendas e gastos

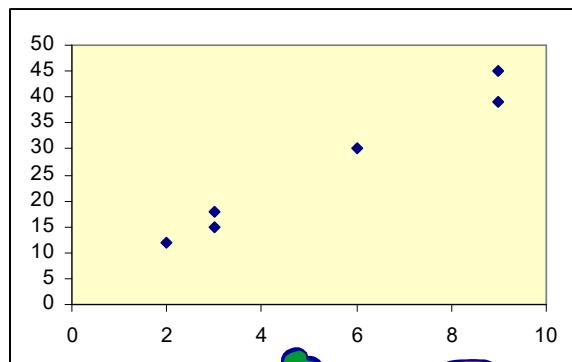
Mês	Vendas	Gastos
Jan.	3	18
Fev.	9	39
Mar.	6	30
Abr.	3	15
Mai.	2	12
Jun.	9	45



287

Uma análise gráfica ...

Vendas	Gastos
3	18
9	39
6	30
3	15
2	12
9	45



X → Y



288

Calculando os somatórios

Vendas X	Gastos Y	X ²	Y ²	XY
3	18	9	324	54
9	39	81	1.521	351
6	30	36	900	180
3	15	9	225	45
2	12	4	144	24
9	45	81	2.025	405

$$\sum \quad 32 \quad 159 \quad 220 \quad 5.139 \quad 1.059$$

289

Calculando b

n	ΣX	ΣY	ΣX^2	ΣY^2	ΣXY
6	32	159	220	5.139	1.059

$$b = \frac{6(1059) - (32 \cdot 159)}{6(220) - (32)^2}$$



$$b = 4,27703$$

290

Calculando a

n	ΣX	ΣY	ΣX^2	ΣY^2	ΣXY
6	32	159	220	5.139	1.059

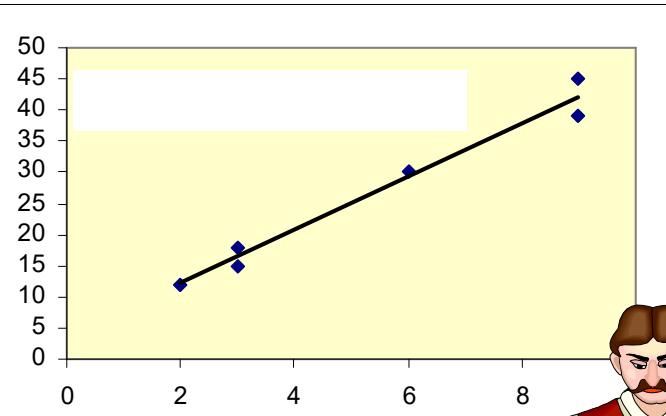
$$a = \frac{159 - 4,42770}{6} \cdot 32$$



$$a = 3,6892$$

291

No gráfico ...



292

Assim ...

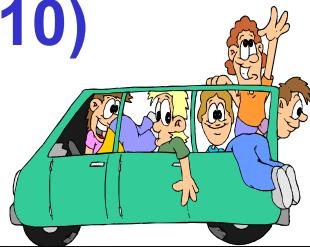
- Equação obtida ...

$$y = 3,6892 + 4,277x$$

- Para vendas previstas iguais a **10** unidades ...

$$y = 3,6892 + 4,277(10)$$

$$y = 46,4592$$



293

Projeção

- Uma das aplicações da regressão linear é projetar valores da variável dependente para valores definidos da variável independente.

294

Exemplo 8.2

- Utilizando a reta de regressão linear do exemplo a seguir, projetar as vendas para investimentos em propaganda de 20, 30 e 45 milhões.

295

296

Função do Excel

PREVISÃO(x; val_conhecidos_y; val_conhecidos_x)

- A função estatística PREVISÃO retorna o valor projetado \hat{y} para o valor registrado no argumento x considerando a reta de regressão linear simples correspondente aos valores das amostras informados nos argumentos $val_conhecidos_y$ e $val_conhecidos_x$.
 - Ao utilizar esta função, deve-se tomar o cuidado de fornecer os valores na ordem correta, o argumento $val_conhecidos_y$ se refere aos valores da variável dependente y , e o argumento $val_conhecidos_x$ aos valores da variável independente x .
 - Os dois argumentos desta função devem ser números ou nomes, matrizes ou referências que contenham números.

297

Campanhas de prevenção

- Um importante município do Ceará investiu em campanhas de acidentes de trânsito por meio de palestras.
- Analise a associação entre os dados.



298

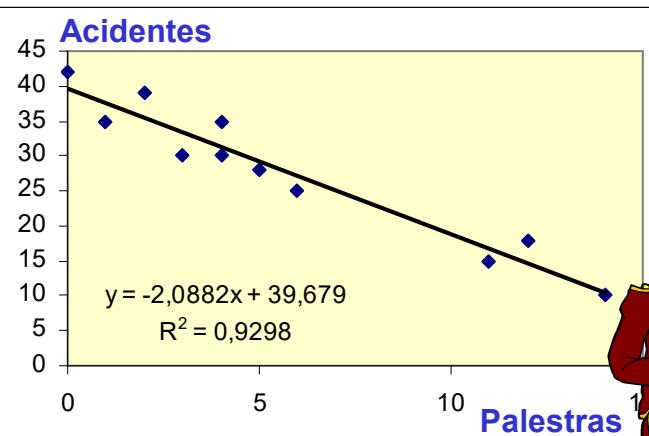
Dados do exercício

Mês	Palestras	Acidentes
Jan.	1	35
Fev.	2	39
Mar.	0	42
Abr.	5	28
Mai.	12	18
Jun.	4	35
Jul.	3	30
Ago.	11	15
Set.	14	10
Out.	6	25
Nov.	4	30



299

Resposta das campanhas



300

A previsão das receitas do hotel

- Um hotel gostaria de projetar suas receitas futuras com base em um modelo de regressão.
- Quais as receitas previstas para os dois próximos semestres?



301

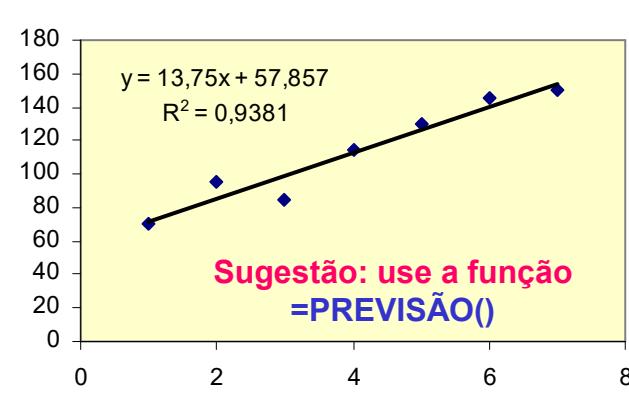
Dados passados ...

Semestre	Receitas
1	70
2	95
3	85
4	115
5	130
6	145
7	150



302

Resposta da previsão



**Valores previstos iguais
a 168 e 182**



303

Análise de correlação ...

**Estuda a
qualidade do
ajuste linear
feito para os
pontos**

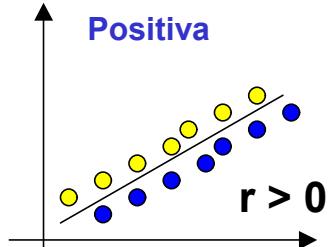


304

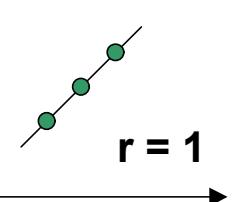
152

Diferentes níveis de aproximação

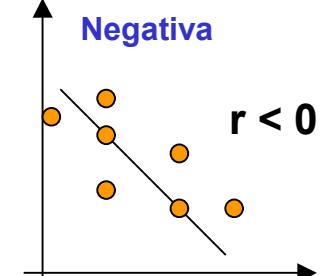
Positiva



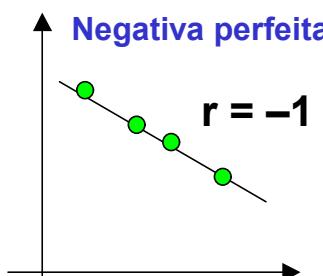
Positiva Perfeita



Negativa



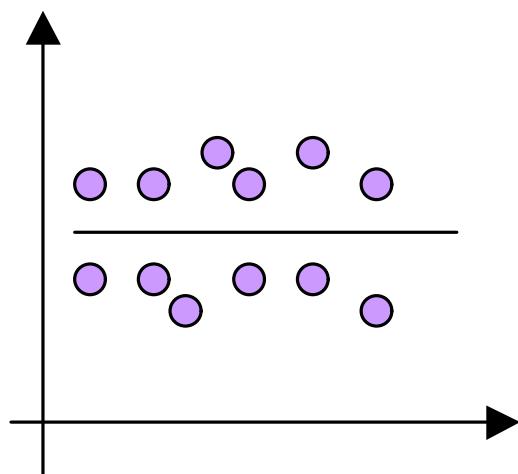
Negativa perfeita



305

Correlação inexistente

$r = 0$



306

153

Coefficiente de determinação

$$r^2 = \frac{\text{Variação explicada}}{\text{Variação total}}$$

$$r^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$



307

Coefficiente de Pearson

$$r = \pm \frac{n \sum xy - \sum x \cdot \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$



308

Coeficiente de Determinação

- O **coeficiente de determinação r^2** é definido como a relação que mede a proporção da variação total da variável dependente, que é explicada pela variação da variável independente:

$$r^2 = \frac{\text{Variação explicada}}{\text{Variação total}}$$

- Substituindo as expressões matemáticas na expressão anterior temos:

$$r^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

309

Coeficiente de Determinação

- A expressão mostra que o coeficiente de determinação r^2 é sempre um número positivo entre zero e um.
- Da própria fórmula pode-se deduzir que quanto maior for r^2 melhor será o poder de explicação da reta de regressão.

310

Função do Excel

RQUAD(*val_conhecidos_y; val_conhecidos_x*)

- A função estatística RQUAD retorna o coeficiente de determinação da reta de regressão considerando os valores das amostras informados nos argumentos *val_conhecidos_y* e *val_conhecidos_x*.
 - Ao utilizar a função RQUAD, deve-se tomar o cuidado de fornecer os valores na ordem correta, o primeiro argumento *val_conhecidos_y* se refere aos valores da variável dependente *y*, e o argumento *val_conhecidos_x*, aos valores da variável independente *x*.
 - Os dois argumentos desta função devem ser números ou nomes, matrizes ou referências que contenham números.

311

Coeficiente de Determinação

- O coeficiente de determinação r^2 , também denominado *r-quadrado*, é sempre um número positivo dentro do intervalo (0; 1) e deve ser interpretado como a proporção da variação total da variável dependente *y*, que é explicada pela variação da variável independente *x*.
- Observe que o coeficiente de correlação mede as variações dos dados da amostra *y* com relação aos valores projetados da reta, sempre na direção de *y*.
 - No exemplo em que o r^2 seja igual a 0,7385, pode-se dizer que 73,85% das variações das vendas podem ser explicadas pela variabilidade do investimento em propaganda, ficando 26,15% sem explicação.

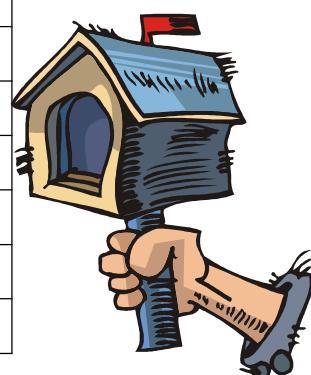
312

- Embora na determinação do coeficiente de correlação não seja necessário separar as variáveis entre independente e dependente, há uma relação importante entre correlação e regressão. Uma delas é a declividade da reta de regressão, que é função do coeficiente de correlação.
- Demonstra-se também que o coeficiente de determinação é igual ao quadrado do coeficiente de correlação, e vice-versa.

313

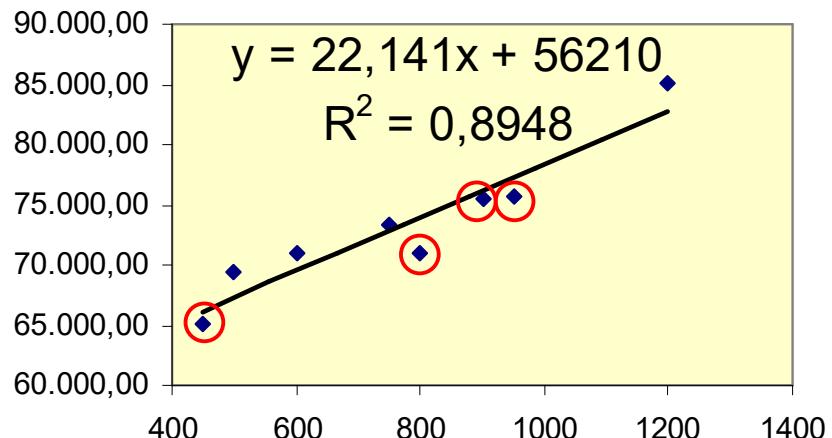
Dados das lojas

Loja	Tamanho em m ²	Vendas em \$
1	1200	85.000,00
2	800	71.000,00
3	600	71.000,00
4	450	65.000,00
5	900	75.500,00
6	950	75.600,00
7	750	73.250,00
8	500	69.500,00



314

Resultado da regressão



315

Empregando o modelo ...

Loja	Tamanho em m ²	Vendas em \$
1	1200	85.000,00
2	800	71.000,00
3	600	71.000,00
4	450	65.000,00
5	900	75.500,00
6	950	75.600,00
7	750	73.250,00
8	500	69.500,00

Previsto	Diferença

$$y = 56210 + 22,141x$$

316

Exercício

- A Oi, empresa de telefonia, resolveu analisar a relação entre idade do seu consumidor e sua conta média mensal. Analisou os dados de uma amostra formada pelos consumidores apresentados seguir.

Idade (em anos)	32	17	26	36	34	53	31	29
Conta média (em \$/mês)	85	20	50	82	77	200	90	60

Pede-se: (a) construa um modelo de ajuste linear entre os pontos; (b) calcule o coeficiente de correlação; (c) para uma idade de 20 anos e 40 anos, qual deve ser sua conta média pelo modelo linear; e (d) interprete os resultados encontrados.

317

Exercício

- O setor de alimentos resolveu analisar a relação entre a renda familiar e o consumo de produtos premium. Analisou os dados de uma amostra formada pelos consumidores apresentados seguir.

Renda familiar (em R\$)	1500	2500	10000	15000	3500	5000	3000	8000
Consumo prod. premium	3	5	8	10	4	7	3	8

Pede-se: (a) construa um modelo de ajuste linear entre os pontos; (b) calcule o coeficiente de correlação; (c) para uma renda de R\$12.000,00 e de R\$20.000,00, qual deve ser o consumo médio pelo modelo linear; e (d) interprete os resultados encontrados.

318

Introdução

- **Muitas vezes há nas empresas a necessidade de descrever o comportamento de certas variáveis importantes para a tomada de decisões, tais como: custos, receitas, despesas e resultados.**
- **As variáveis relevantes podem ser previstas intuitivamente, utilizando uma pesquisa de mercado, por exemplo. Mas isso só resolve o problema no curto prazo,**

319

Introdução

- Estas duas técnicas compreendem a análise de dados amostrais para obter informações sobre se duas ou mais variáveis são relacionadas e qual é a natureza desse relacionamento.
- A análise de regressão, bastante empregada nas áreas de negócios, é utilizada principalmente com o propósito de previsão.
- Consiste em determinar uma função matemática que busca descrever o comportamento de determinada variável dependente com base nos valores de uma ou mais variáveis independentes.
- A análise de correlação visa medir a força ou o grau de relacionamento entre variáveis.

320

Regressão linear simples

- Considere o seguinte exemplo:
- A empresa Previpeças S. A., que é fabricante de autopeças, deseja projetar as quantidades de peças a serem vendidas no próximo ano. Como a empresa entende que a quantidade de peças vendidas pode ser explicada por seu preço, pretende definir um modelo que relacione estas variáveis.
- As quantidades de peças vendidas nos últimos anos, bem como seus respectivos preços de venda, são mostradas na tabela a seguir.

321

Regressão linear simples

Tabela 1: Quantidade de peças vendidas e respectivos preços

Anos	Quantidade (Q) (1.000 un.)	Preço (P) (\$ 1.000,00)
1	2	4
2	1	6
3	3	3
4	1	5
5	4	1
6	3	2

Fonte: Corrar e Theóphilo (2004, p. 76).

322

Regressão linear simples

- Em um primeiro momento, é possível supor que a previsão desejada poderia ser feita a partir da média da quantidade de peças vendidas nos últimos anos.

$$\bar{Q} = \frac{2+1+3+1+4+3}{6} = \frac{14}{6} = 2,33 \text{ mil unidades.}$$

- Será que a média histórica de vendas proporciona previsão adequada das vendas futuras?
- A resposta a essa pergunta pode ser conseguida analisando-se, inicialmente, o Diagrama de Dispersão correspondente, como mostra a figura 1.
- O diagrama de dispersão é um gráfico bidimensional, por meio do qual se pode analisar o comportamento das variáveis em estudo.

323

Regressão linear simples

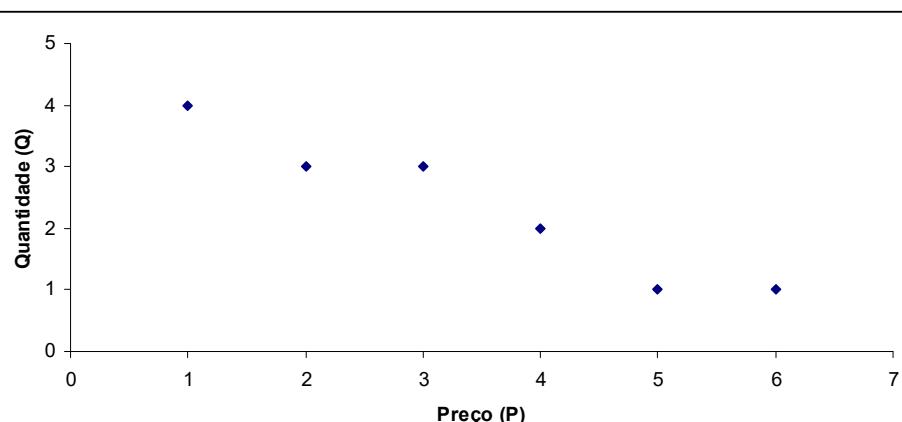


Figura 1: Diagrama de dispersão Preço x Quantidade.
Fonte: Corrar e Theóphilo (2004, p. 77).

324

Regressão linear simples

- Pela análise do gráfico, pode-se observar que a quantidade de peças vendidas apresenta a tendência de declínio à medida que os preços aumentam.
- Esse é um indicativo da existência de relação entre as variáveis.
- Por essa razão, a média histórica não é adequada ao propósito de se prever as quantidades vendidas, pois utilizar a média como modelo de previsão significa dizer que a quantidade de vendas permanece constante e igual a 2,33 mil unidades, independente do comportamento da variável preço, como mostra a figura 2.

325

Regressão linear simples

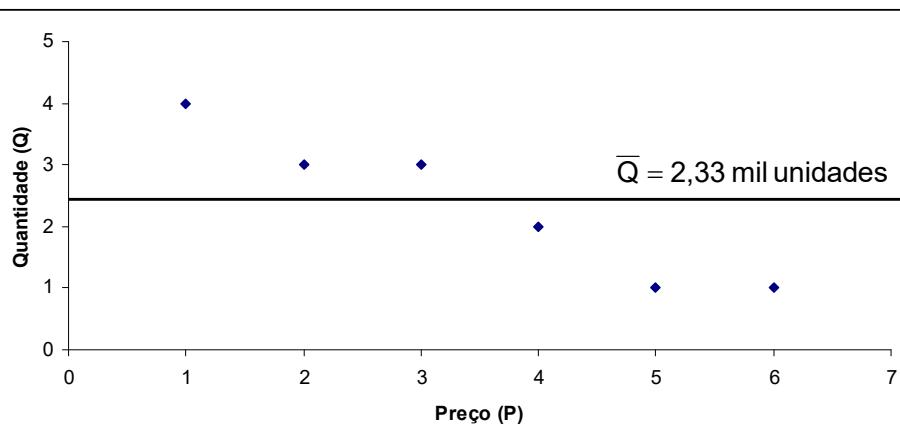


Figura 2: Ajustamento utilizando a quantidade média de vendas.
Fonte: Corrar e Theóphilo (2004, p. 76).

326

Regressão linear simples

- O comportamento entre duas variáveis pode assumir diversas formas, que vão desde uma relação linear até complicadas formas não lineares.
- No exemplo estudado, o comportamento entre as variáveis tende para uma relação linear.
- Por isso, o próximo passo consiste em buscar determinar a respectiva equação de regressão linear simples.
- Equação da reta: toda reta pode ser representada pela seguinte expressão matemática: $y = a + b.x$, sendo que x e y são variáveis, e “ a ” e “ b ” são seus respectivos coeficientes.
- Para exemplificar, é mostrada na figura 3 a seguir o gráfico correspondente à equação linear $y = 4 + 3x$.

327

Regressão linear simples

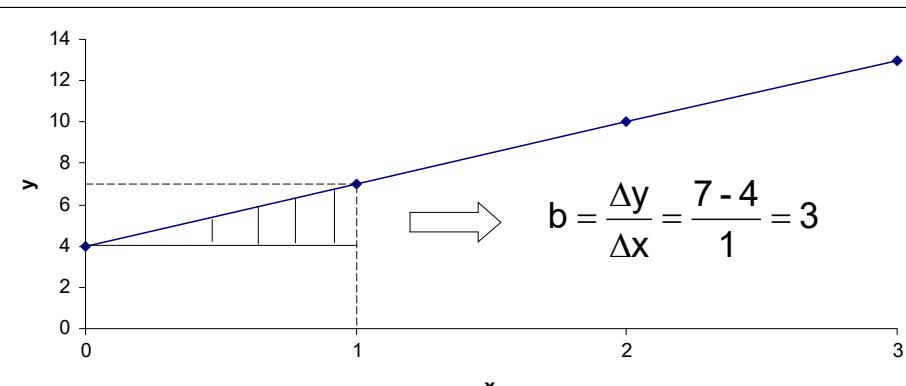


Figura 3: Gráfico da reta $y = 4 + 3x$.
Fonte: Corrar e Theóphilo (2004, p. 76).

328

Regressão linear simples

- O coeficiente “a” representa o ponto em que a reta de regressão intercepta o eixo vertical y. Isso ocorre quando x é igual a zero:
 - $y = 4 + 3.(0)$
 - $y = 4$
- Por sua vez, o coeficiente “b” representa o variação de y por unidade de variação de x. Na equação $y = 4 + 3x$, o coeficiente angular é 3; isso significa que, a cada variação de uma unidade de x, correspondem 3 unidades de variação em y.
- De uma maneira geral, tem-se um modelo de regressão linear simples quando uma relação linear entre duas variáveis, x e y, pode ser satisfatoriamente definida pela seguinte equação matemática:

329

Regressão linear simples

$$Y = A + B.X + U$$

- Cujos parâmetros são:
 - ✓ Y = variável dependente;
 - ✓ X = variável independente ou explicativa;
 - ✓ A = coeficiente linear ou intercepto da reta;
 - ✓ B = coeficiente angular ou declividade da reta;
 - ✓ U = erro aleatório na população.
- Esta equação representa o modelo de regressão da população. O que se pretende é estimar os verdadeiros parâmetros populacionais da amostra disponível. A equação a seguir é uma estimativa da equação populacional:

$$\hat{y} = a + b.x$$

330

Regressão linear simples

- Onde:
- \hat{y} = estimativa da variável dependente Y;
- a = estimativa do coeficiente linear A;
- b = estimativa do coeficiente angular B;
- x = valores amostrais da variável explicativa X.
- Para que se possa utilizar a análise de regressão com o propósito de se fazer previsões, deve-se calcular os coeficientes “a” e “b” da equação, que são utilizados como estimativas dos parâmetros populacionais A e B.

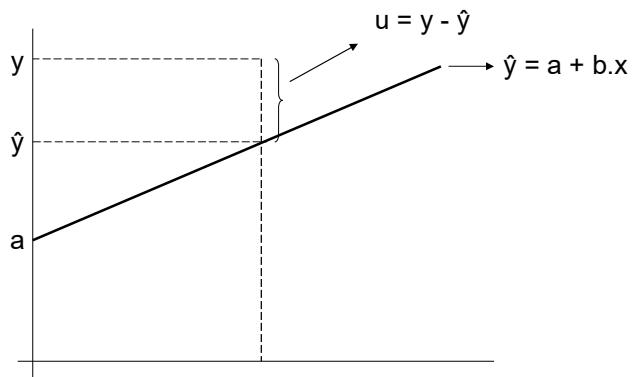
331

Método dos mínimos quadrados

- Para cada valor de x, podem existir um ou mais valores de y na amostra. Denominam-se esses valores de y observados.
- Por outro lado, para cada valor de x existirá um único valor de y pertencente à reta de regressão, denominado y projetado ou estimado (\hat{y}).
- Portanto, para cada valor de x pode-se ter um ou mais valores de y observados diferentes de \hat{y} estimado. Essas diferenças são chamadas de resíduos ou desvios.
- Adota-se o símbolo (u) para o resíduo: $u = |y - \hat{y}|$.
- A figura a seguir mostra a representação gráfica do resíduo.

332

Método dos mínimos quadrados



Representação gráfica do resíduo.

333

Método dos mínimos quadrados

- O objetivo da análise de regressão linear simples é o de obter a reta que melhor se ajuste aos dados observados.
- Para obter a reta, precisamos estimar os coeficientes “a” e “b”. Existem diversos métodos utilizados para esse fim. O mais usual deles é denominado método dos mínimos quadrados.
- Esse método parte de princípio de que a reta que melhor se ajusta aos dados é aquela para a qual as diferenças entre os valores observados e os valores projetados são as menores possível. Isto é, a reta deve ser tal que a soma dos resíduos:
$$\Sigma u = \Sigma |y - \hat{y}|$$
 seja mínima.
- O problema de se adotar esse método é que, como existem resíduos positivos e negativos, seu somatório será sempre igual a zero. Por isso, toma-se a soma dos desvios elevados ao quadrado.

334

Régressão linear múltipla

- Onde:
 - ✓ A = intercepto de “Y”;
 - ✓ B_i = coeficientes angulares;
 - ✓ U = erro aleatório em “Y” para a observação “i”.
- Como na regressão linear simples, deseja-se estimar os coeficientes B_i ($i = 1, 2, 3, \dots, k$) a partir dos valores de b_i , utilizando o método dos mínimos quadrados.
- Cada coeficiente b_i representa a variação provocada em y pelo aumento de uma unidade de x_i , consideradas constantes todas as outras variáveis independentes.
- Utiliza-se a ferramenta de análise de dados de regressão do Excel®.

335

Exemplo

- A Companhia Multifator deseja analisar o comportamento dos Custos Indiretos de Fabricação (cif), em função das variáveis: Horas de Mão-de-Obra Direta (hmod) e Horas-Máquina (hm).

Período	CIF	HMOD	HM
1	350	4	10
2	400	8	14
3	470	12	16
4	550	10	26
5	620	15	31
6	380	7	12
7	290	6	13
8	490	10	21
9	580	11	26
10	610	13	24
11	560	12	23
12	420	8	12
13	450	11	19
14	510	12	19
15	380	5	11

336

Exemplo

- Análise do modelo de regressão múltipla:
- Primeiramente, analisa-se os modelos de regressão simples entre cada uma das variáveis independentes (hm e hmod) e a variável dependente (cif).
 - Inicialmente estuda-se o comportamento dos (cif) em função da variável explicativa (hm). O relatório de Resumo dos Resultados da regressão é mostrado a seguir.

337

Exemplo

Estatística de regressão

R múltiplo	0,9198
R-Quadrado	0,8461
R-quadrado ajustado	0,8343
Erro padrão	40,9213
Observações	15

338

Exemplo

ANOVA

	gl	SQ	MQ	F	de significação
Regressão	1	119724,13	119724,1	71,49613	1,21328E-06
Resíduo	13	21769,20	1674,554		
Total	14	141493,33			

	Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores
Interseção	208,8764	32,7140	6,384919	2,4E-05	138,2020829
HM	14,1763	1,6765	8,455538	1,21E-06	10,55433973

339

Exemplo

- Analisando-se o relatório de resumos de resultados para o modelo $\hat{cif} = f(hm)$, tem-se que:
- $R^2 = 0,8461$: indica que existe forte correlação linear entre as variáveis, pois 84,61% da variação de (cif) são explicados pela variação de (hm);
- Valor P de aproximadamente zero: indica que o relacionamento linear entre as variáveis (cif) e (hm) é significativo, ao nível de 5%.
- O modelo proposto é $\hat{cif} = 208,88 + 14,18.hm$.
- Conclui-se que o modelo proposto é adequado para prever o comportamento da variável dependente (cif).
- Em seguida, analisa-se o comportamento da variável dependente (cif) em função da variável explicativa (hmod).

340

Exemplo

Estatística de regressão

R múltiplo	0,882913791
R-Quadrado	0,779536763
R-quadrado ajustado	0,762578052
Erro padrão	48,98514557
Observações	15

341

Exemplo

ANOVA

	gl	SQ	MQ	F	de significação
Regressão	1	110299,255	110299,3	45,96675	1,30171E-05
Resíduo	13	31194,0783	2399,544		
Total	14	141493,333			

	Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores
Interseção	200,82139	41,762202	4,8086	0,000341	110,5996422
HMOD	28,108882	4,1459274	6,7798	1,3E-05	19,15215091

342

Exemplo

- A interpretação do relatório para o modelo $\hat{cif} = f(hmod)$ é:
- O R^2 indica que 77,95% da variação do (cif) são explicadas pela variação de (hmod), o que revela forte associação linear entre as variáveis;
- O valor P indica que há um relacionamento linear significativo entre as variáveis (cif) e (hmod), ao nível de 5%.
- O modelo proposto é $\hat{cif} = 200,82 + 28,11.hmod$.
- Pode-se afirmar que esse é também um modelo adequado para prever o comportamento da variável (cif).
- Cada modelo individualmente é adequado para explicar o comportamento da variável (cif), mas será que o modelo de regressão múltipla é tão bom ou melhor que ambos?

343

Exemplo

Estatística de regressão

R múltiplo	0,940734411
R-Quadrado	0,884981232
R-quadrado ajustado	0,865811437
Erro padrão	36,82660828
Observações	15

344

Exemplo

ANOVA

	gl	SQ	MQ	F	F de significação
Regressão	2	125218,9444	62609,47	46,1654	2,31533E-06
Resíduo	12	16274,38892	1356,199		
Total	14	141493,3333			

	Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores
Interseção	184,8836013	31,76204673	5,820897	8,2E-05	115,6800465
HMOD	11,74602193	5,835472408	2,012866	0,067121	-0,968380201
HM	9,369382026	2,824833003	3,316791	0,006147	3,214599645

345

Exemplo

- A equação de regressão múltipla assume a forma:
- $cif = 184,884 + 11,746.hmod + 9,369.hm$
- R^2 representa o coeficiente de determinação múltipla, que é uma medida de quão bem a equação de regressão múltipla se ajusta aos dados amostrais. Um ajuste perfeito resultaria em $R^2 = 1$. Um ajuste muito ruim resulta em um valor de R^2 próximo de zero.
- É recomendado que na regressão múltipla se utilize o coeficiente de determinação ajustado $R^2_{ajustado}$ em vez do coeficiente R^2 . Isto porque sempre que uma variável explicativa é adicionada ao modelo, o valor de R^2 cresce, mesmo que essa variável não seja estatisticamente significativa.

346

Exemplo

- Valor P: o valor P é uma medida da significância global da equação de regressão múltipla. Um pequeno valor P indica que a equação de regressão múltipla tem boa significância geral e é adequada para previsões.
- Assim como o valor de $R^2_{ajustado}$, esse valor P é uma boa medida de quão bem a equação se ajusta aos dados amostrais.
- Para que um modelo de regressão múltipla seja eficiente e adequado, cada uma das variáveis explicativas do modelo de regressão deve se relacionar significativamente com a variável dependente y.
- Como já exposto, uma das formas de se demonstrar isso é utilizar o teste do valor P.

347

Exemplo

- Para o exemplo apresentado, pode-se concluir que:
- Para a variável explicativa (hm), tem-se que o valor P = 0,006. Este valor é menor que o valor para o teste de significância, que é 0,05, indicando que há relacionamento linear significativo entre as variáveis (cif) e (hm);
- Para a variável explicativa (hmod), tem-se que o valor P = 0,0671 > 0,05, indicando que não há relacionamento linear significativo entre as variáveis (cif) e (hmod).
- Portanto, pode-se concluir que a variável (hmod) não contribui para o aumento da eficiência de previsão do modelo resultante da regressão linear múltipla. Por isso, deve-se excluir essa variável do modelo, considerando para a previsão do (cif):

$$\hat{cif} = 208,88 + 14,18.hm$$

348

Capítulo

9

Preparação de Dados

349

Parte II

Exame gráfico dos dados

350

175

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

351

Examine seus dados...

➤ Existe algum problema com meu banco de dados?

➤ Como solucionar esses problemas?



352

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 (DemonstContEmpresas).
- Esses dados foram retirados de demonstrações contábeis de empresas brasileiras.

353

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.

354

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**

355

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
<i>Valid N (listwise)</i>	100				

356

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.

357

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:

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

358

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

359

Como calcular a variância?

- Para uma amostra:

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

- Para uma população finita:

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

360

180

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**

361

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	

362

Exame gráfico dos dados

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

363

Forma da distribuição

- Construindo um histograma é possível representar a freqüência 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.

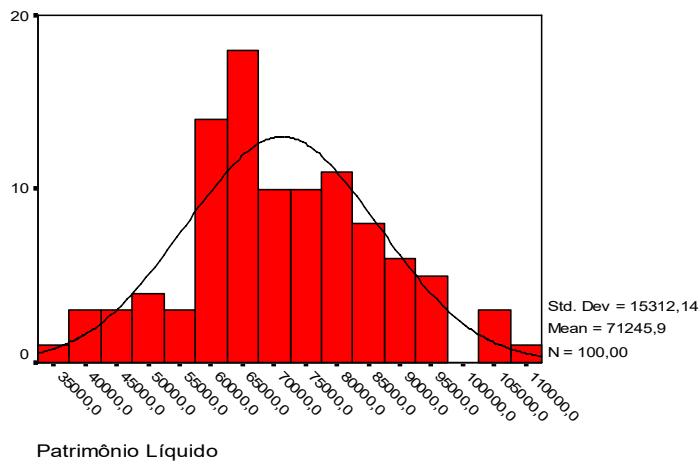
364

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*

365

Relatório do SPSS (Histograma)



366

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.

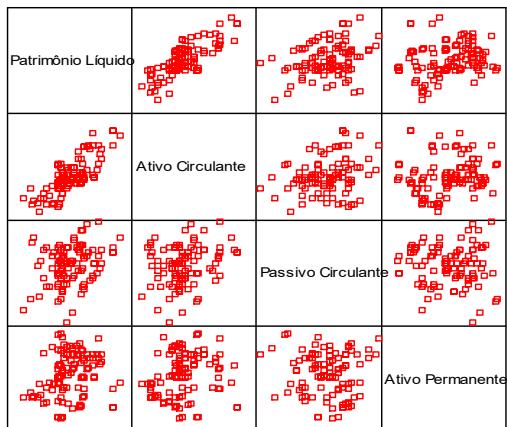
367

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*

368

Relatório do SPSS (Matriz de dispersão)



369

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).

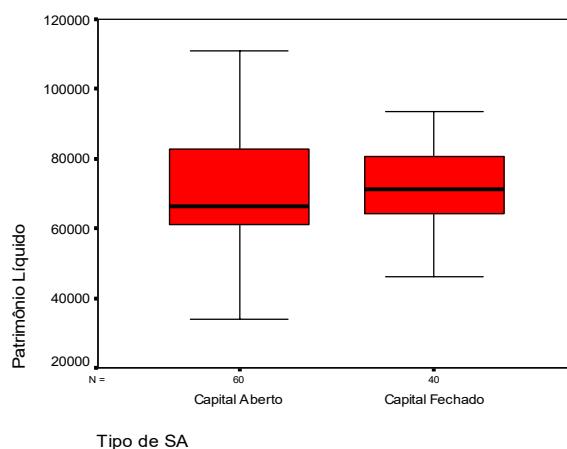
370

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**

371

Relatório do SPSS (Gráfico de caixas)



372

Parte III

Observações atípicas (*outliers*)

373

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.

374

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.

375

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 ($N \leq 80$) *outlier* apresenta $score \geq 2,5$.
- ✓ Em grandes amostras *outlier* apresenta $score \geq 3,0$.

376

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).

377

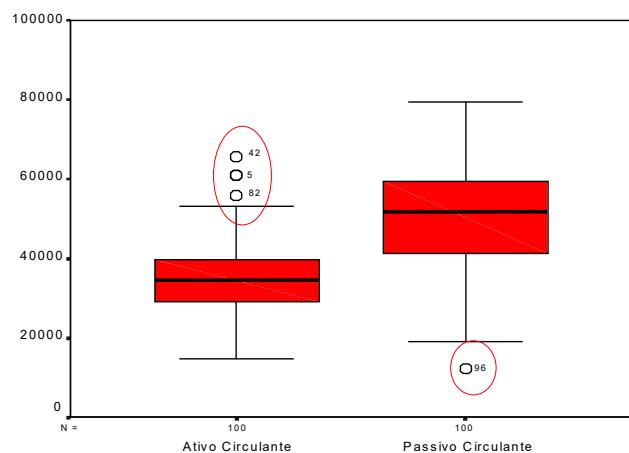
Passos no SPSS (*Outliers*: detecção univariada)

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

378

189

Relatório do SPSS **(Outliers: detecção univariada)**



379

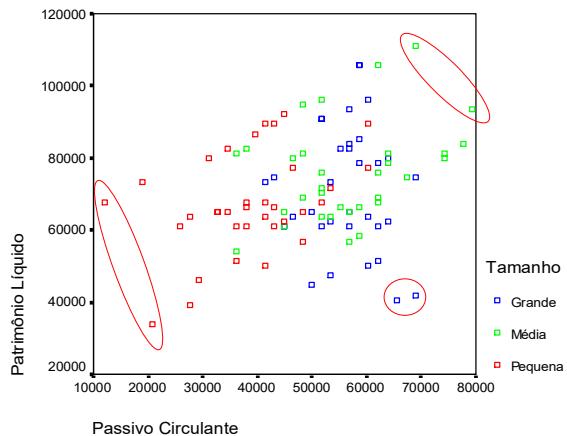
Passos no SPSS **(Outliers: detecção bivariada)**

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

380

190

Relatório do SPSS (Outliers: detecção bivariada)



381

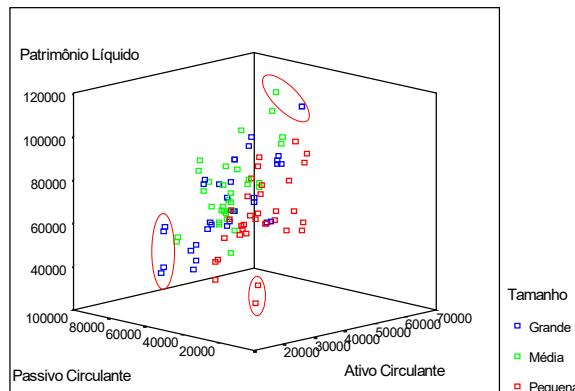
Passos no SPSS (Outliers: detecção três dimensões)

- 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*

382

191

Relatório do SPSS (Outliers: detecção três dimensões)



383

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.

384

Parte IV

Dados perdidos (*missing value*)

385

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.

386

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.

387

Ações corretivas (remédios) para dados perdidos

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

388

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.

389

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.

390

Utilizar métodos de atribuição

- ✓ O método de atribuição é um processo de estimativa 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

391

Conceitos

Técnica estatística que permite analisar a relação entre uma única variável dependente e uma ou mais variáveis independentes e fazer projeções.

- ☞ Vendas de produtos X investimentos em propaganda;
- ☞ Gastos familiares X renda familiar X nº de membros da família;
- ☞ Demanda X preço X renda *per capita* X crescimento da população;
- ☞ Salário X produtividade X tempo de casa.

392

392

Capítulo

10

Análise Discriminante

393

Introdução

- Existem muitas situações da realidade que podem ser vistas de forma estratificada, classificada ou agrupada.
- Ex.: Clientes rentáveis e não rentáveis; classes sociais A, B e C; consumidores dos produtos 1, 2, e 3; produtos que vendem bem nas regiões 1, 2 e 3 e não vendem nas regiões 4 e 5.
- Sabe-se que cada elemento integrante dos grupos ou estratos exemplificados possui uma gama de informações próprias, como idade, renda, forma, qualidade, altura, peso.
- Pode-se desejar, através da utilização de uma amostra, predizer os grupos aos quais pertencem novas observações, ou ainda que características são mais importantes na distinção entre os elementos de um e outro grupo.

394

197

Introdução

- Para abordar problemas com as características descritas, tem sido utilizada uma técnica estatística denominada Análise Discriminante.
- A análise discriminante é uma das técnicas de análise estatística multivariada, um ramo da análise estatística que se ocupa da investigação simultânea de duas ou mais variáveis.
- A análise estatística multivariada abrange também regressão múltipla, análise de variância e covariância multivariada, análise combinada ou conjunta (*conjoint*), análise canônica de correlação, análise de estratos, dentre outras.
- A análise discriminante trabalha com variáveis categóricas como variável dependente ou explicada e variáveis métricas para variáveis independentes ou explicativas

395

Análise discriminante

- Essa técnica usa informações disponíveis de variáveis métricas independentes para estimar o valor de uma variável dependente categórica.
- Em essência, busca desenvolver uma regra para prever a qual grupo, definido *a priori*, pertence uma nova observação, considerando os valores assumidos pelas variáveis independentes. A regra desenvolvida não tem a pretensão de ser infalível, mas apontar a alternativa de maior chance.
- A diferença básica entre a análise discriminante e a análise de regressão consiste no fato de que na primeira a variável dependente é categórica ou nominal, enquanto na segunda a variável dependente é numérica. Em ambas, porém, as variáveis independentes são numéricas.

396

Análise discriminante

- Em síntese, o objetivo da técnica é identificar a que classe pertence cada um dos elementos de um conjunto que trabalha com variáveis relacionadas a esses elementos que se pressupõem ser explicativas da definição da classe a que pertencem.
- Exemplos de aplicação da análise discriminante:
- *Credit scoring*: créditos efetuados aos clientes de uma instituição financeira podem pertencer a dois grupos distintos. Créditos que foram honrados pelos devedores e créditos que assumiram o estado de liquidação duvidosa.
- *Insurance rating*: pode ser usada para prever a classificação de risco (baixo, médio ou alto) de um novo cliente de uma seguradora. Informações como há quanto tempo dirige, número de vezes que se envolveu em acidentes, estado civil podem ser usadas como variáveis independentes para classificar a variável nominal risco.

397

Análise discriminante – dois grupos

- Considere o exemplo hipotético com duas variáveis independentes e dois grupos para discriminação das informações.
- O Banco Varejo tem como estratégia a ampliação de sua atuação no segmento de varejo. Para otimizar seus esforços de marketing precisa identificar, a priori, que característica tem o cliente de varejo que oferece melhores margens de contribuição. Assim, toma-se uma amostra aleatória de 20 observações pertencentes a dois grupos (grupo 1 margem de contribuição satisfatória e grupo 2 margem de contribuição não satisfatória) e os dados correspondentes das variáveis renda e número de dependentes, conforme a tabela a seguir.

398

199

Análise discriminante – dois grupos

Observ.	Grupo	Renda	Depend.	Observ.	Grupo	Renda	Depend.
1	1	3.400	3	11	2	3.800	5
2	1	2.400	2	12	2	3.400	5
3	1	2.700	2	13	2	2.000	3
4	1	2.300	2	14	2	1.100	3
5	1	3.100	1	15	2	1.800	3
6	1	2.200	2	16	2	1.100	2
7	1	4.900	5	17	2	1.000	2
8	1	2.700	3	18	2	2.600	4
9	1	3.400	4	19	2	600	2
10	1	4.200	5	20	2	1.000	5
Média 1		3.130	2,9	Média 2		1.840	3,40

399

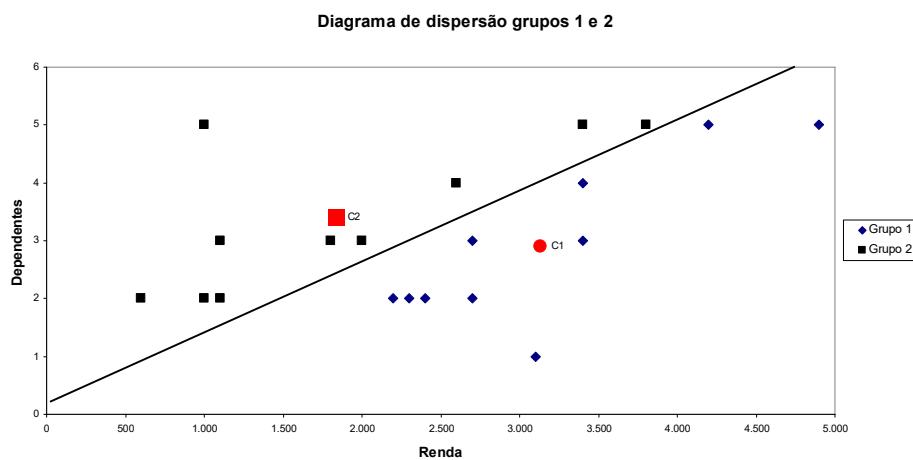
Análise discriminante – dois grupos

- O problema consiste em criar uma regra, baseada nos dados das variáveis renda e número de dependentes, capaz de identificar a qual grupo pertencerá um novo potencial cliente.
- Como o exemplo tem o número de variáveis limitado a duas, e com o único objetivo de facilitar o entendimento da técnica, pode-se visualizar os dados em um diagrama de dispersão.
- Do diagrama constam os pontos centróides dos dois grupos, indicados por C1 – centróide do grupo 1 – e C2 centróide do grupo 2. Esses pontos têm como coordenadas as médias das variáveis do grupo.
- Se fosse admitido que os pontos de observação gravitam em torno de seus respectivos centróides, seria possível dividi-los pelo traçado de uma reta que separasse da melhor forma possível os dois grupos.

400

200

Análise discriminante – dois grupos



401

Análise discriminante – dois grupos

- A regra de classificação poderia ser a seguinte: se as coordenadas da nova observação determinarem um ponto acima da reta, esta pertencerá ao grupo 2; se abaixo da reta, ao grupo 1.
- Isso sugere que problemas de análise discriminante podem ser equacionados por regressão linear. Softwares estatísticos reconhecidos como SPSS, SAS e STATGRAPH apresentam a regressão linear como forma de resolver problemas de análise discriminante.
- A solução do problema proposto requer uma função discriminante linear do tipo:

$$Z = a + b_1 X_1 + b_2 X_2$$

- Onde:

402

201

Análise discriminante – dois grupos

- ✓ Z representa o valor estimado da variável dependente ou escore discriminante;
- ✓ a é o intercepto da reta que representa a função discriminante;
- ✓ b_i são os coeficientes discriminantes das variáveis independentes ($i = 1, 2, \dots$); e
- ✓ X_i corresponde aos valores das variáveis independentes ($i = 1, 2, \dots$).
- Essa função resume em Z a capacidade máxima de explicação conjunta das duas variáveis independentes. Para o cálculo dos coeficientes discriminantes, utiliza-se a ferramenta regressão do Excel®, através, do menu ferramentas/análise de dados e a opção “regressão” da caixa de diálogo análise de dados.

403

Análise discriminante – dois grupos

RESUMO DOS RESULTADOS

Estatística de regressão

R múltiplo	0,79107278
R-Quadrado	0,625796144
R-quadrado ajustado	0,581772161
Erro padrão	0,331752856
Observações	20

404

Análise discriminante – dois grupos

ANOVA

	gl	SQ	MQ	F	F de significação
Regressão	2	3,128980719	1,564490359	14,21489152	0,000235191
Resíduo	17	1,871019281	0,110059958		
Total	19	5			
Coeficientes		Erro padrão	Stat t	valor-P	95% inferiores
Interseção	1,672058423	0,213251584	7,840778436	4,79387E-07	1,222136913
Renda	-0,000387728	7,50573E-05	-5,165768841	7,75778E-05	-0,000546085
Dependentes	0,251252931	0,067243373	3,736471259	0,001642472	0,109381816

405

Análise discriminante – dois grupos

- Com base nos coeficientes da regressão, pode-se escrever a equação que será utilizada para predizer os grupos a que pertencem as observações:
$$Z = 1,672058 - 0,000388X_1 + 0,251253X_2$$
- Essa equação é uma função linear e é denominada função discriminante. Os coeficientes de X_1 e X_2 são chamados coeficientes discriminantes e constituem a ponderação das variáveis independentes quanto à capacidade de predição.
- Os valores calculados pela equação, com a substituição dos valores X_1 – renda – e X_2 – dependentes –, são denominados escores discriminantes e não são, necessariamente, números inteiros. Esses pontos serão comparados a um ponto de corte para então se proceder à identificação das observações com os grupos a que pertencem.

406

Análise discriminante – dois grupos

Observação	Grupo	Renda	Dependentes	Escore discriminante	Grupo predito	Classificação incorreta
1	1	3.400	3	1,11	1	
2	1	2.400	2	1,24	1	
3	1	2.700	2	1,13	1	
4	1	2.300	2	1,28	1	
5	1	3.100	1	0,72	1	
6	1	2.200	2	1,32	1	
7	1	4.900	5	1,03	1	
8	1	2.700	3	1,38	1	
9	1	3.400	4	1,36	1	
10	1	4.200	5	1,30	1	
Média 1		3.130	2,9	1,19		

407

Análise discriminante – dois grupos

Observação	Grupo	Renda	Dependentes	Escore discriminante	Grupo predito	Classificação incorreta
11	2	3.800	5	1,45	1	SIM
12	2	3.400	5	1,61	2	
13	2	2.000	3	1,65	2	
14	2	1.100	3	2,00	2	
15	2	1.800	3	1,73	2	
16	2	1.100	2	1,75	2	
17	2	1.000	2	1,79	2	
18	2	2.600	4	1,67	2	
19	2	600	2	1,94	2	
20	2	1.000	5	2,54	2	
Média 2		1.840	3,4	1,81		
Ponto de corte					1,50	

408

Análise discriminante – dois grupos

- O ponto de corte é a média das médias dos escores discriminantes de cada grupo.
- No exemplo, é 1,50, resultado da média entre 1,19 e 1,81. A regra de atribuição está baseada no ponto de corte. Se o escore discriminante for igual ou menor do que o ponto de corte a observação pertence ao grupo 1; se maior, pertence ao grupo 2.
- Por exemplo, na primeira observação, o escore discriminante é: $1,672058 - (0,000388 \times 3400) + (0,251253 \times 3) = 1,106617 \approx 1,11$, pertecendo então ao grupo 1 por ser menor que o ponto de corte, ou escore crítico, 1,50.

409

Avaliação do poder discriminatório da função

- Um aspecto relevante é a avaliação da capacidade de discriminação da função pelo exame do desempenho desta com relação à alocação correta dos elementos entre os grupos.
- Isto pode ser visto resumidamente através da matriz de classificação a seguir:

		Grupos preditos		Totais
		1	2	
Grupos originais	1	10	0	10
	2	1	9	10
Totais		11	9	20

410

Avaliação do poder discriminatório da função

- Como pode ser notado na tabela, os escores discriminantes permitiram a identificação correta da quase totalidade da amostra (95% – 19 em 20). Os elementos preditos constam na linha diagonal da matriz. Somente a observação de número 11 foi classificada incorretamente.
- Sabe-se, contudo, que o alto índice de classificação correta decorre do fato de se estar aplicando a equação na mesma amostra que serviu para estimar os parâmetros.
- É aconselhável que a amostra inicial seja dividida em duas, de maneira que uma parte sirva para estimar a função e a outra seja utilizada para validar os dados.
- A eficiência das previsões, porém, só seria mensurada adequadamente com novos dados.

411

Avaliação do poder discriminatório da função

- Outro ponto a ser observado é a comparação do percentual de acerto conseguido com a função discriminante com as chances de acerto sem o uso da função.
- Se, por exemplo, um dos grupos contivesse 20% dos elementos e o outro 80%, bastaria classificar todos os elementos no segundo grupo para obter 80% de acerto.
- Nesse caso, a função discriminante só teria boa performance se conseguisse classificar corretamente um número superior a 80% dos elementos.
- Esse critério da avaliação da função é denominado de critério de máxima chance.
- Um segundo critério utilizado é o critério de chance proporcional obtido pela seguinte fórmula:

412

Avaliação do poder discriminatório da função

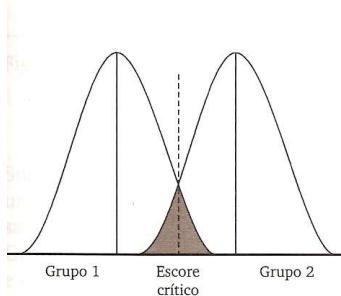
$$C_{\text{pro}} = p^2 + (1 - p)^2$$

- Onde:
 - ✓ C_{pro} = critério de chance proporcional;
 - ✓ p = proporção de elementos do grupo 1;
 - ✓ $(1 - p)$ = proporção de elementos do grupo 2.
- Para o caso onde os grupos têm 20% e 80%, respectivamente, o critério de chance proporcional seria $C_{\text{pro}} = (0,2^2 + 0,8^2) = 0,68$ ou 68%.
- Assim, se a função classificar mais de 68%, pode-se concluir que é válido utilizá-la.

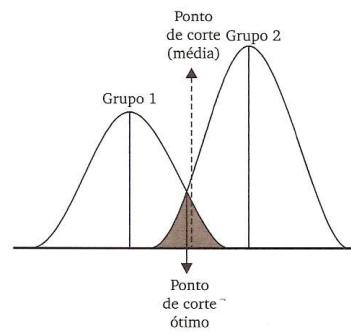
413

Escore crítico ou ponto de corte

- Três aspectos devem ser observados quando se extraem amostras das populações para a Análise Discriminante. O primeiro é o número de observações retiradas das populações. A figura a seguir mostra a distribuição normal das variáveis independentes.



Amostras de tamanhos iguais



Amostras de tamanhos diferentes

414

Escore crítico ou ponto de corte

- A figura mostra que, com amostras de tamanhos diferentes, o escore crítico, considerando apenas as médias, não é o ponto ótimo de discriminação entre os grupos.
- Um bom escore crítico deve dividir a zona de indefinição – região sombreada – em partes iguais. Isso não ocorre no caso de amostras de tamanhos diferentes. Faz-se necessário encontrar outra forma de calcular o escore crítico para melhorar a divisão da zona sombreada.
- Uma das formas de se conseguir isto é calcular o escore crítico usando uma média ponderada em relação ao número de elementos de cada amostra, como se segue:

$$Z_{EC} = \frac{n_1 Z_2 + n_2 Z_1}{n_1 + n_2}$$

415

Escore crítico ou ponto de corte

- Onde:
 - ✓ Z_{EC} = escore crítico para n diferentes;
 - ✓ n_1 = número de observações do grupo 1;
 - ✓ n_2 = número de observações do grupo 2;
 - ✓ Z_1 = centróide do grupo 1;
 - ✓ Z_2 = centróide do grupo 2.
- Para melhor entendimento, toma-se os dados do exemplo do Banco Varejo e supõe-se que os grupos sejam de tamanhos diferentes.
- Pode-se observar que o uso da fórmula ponderou os centróides pelos diferentes tamanhos dos grupos, afastando o ponto de corte da média do grupo com maior número de observações e melhorando a divisão da área sombreada.

416

208

Escore crítico ou ponto de corte

Observação	Grupo	Renda	Dependentes	Escore discriminante
1	1	3.400	3	1,27
2	1	2.400	2	1,28
3	1	2.700	2	1,19
4	1	2.300	2	1,31
5	1	3.100	1	0,76
6	1	2.200	2	1,34
7	1	4.900	5	1,40
8	1	2.700	3	1,49
Média 1		2.963	2,50	1,25

417

Escore crítico ou ponto de corte

Observação	Grupo	Renda	Dependentes	Escore discriminante
9	2	3.400	4	1,57
10	2	4.200	5	1,62
11	2	3.800	5	1,74
12	2	3.400	5	1,87
13	2	2.000	3	1,71
14	2	1.100	3	1,99
15	2	1.800	3	1,77
16	2	1.100	2	1,69
17	2	1.000	2	1,72
18	2	2.600	4	1,82
19	2	600	2	1,85
20	2	1.000	5	2,63
Média 2		2.167	3,58	1,83
Ponto de Corte (Média)				1,54
Ponto de Corte Ótimo				1,48

418

Escore crítico ou ponto de corte

- O segundo ponto a ser observado é o tamanho dos grupos na população. Supondo que o tamanho dos grupos seja igual, a probabilidade de se retirar um elemento do grupo 1 ou do grupo 2 é igual.
- No entanto, quando os tamanhos são diferentes – supondo, por exemplo, que existam na população 40% do grupo 1 e 60% do grupo 2 –, a probabilidade de retirar um elemento do grupo 1 e outro do grupo 2 é diferente.
- Em consequência, as probabilidades de alocar de forma incorreta um elemento são diferentes, como no caso de tamanhos de amostras diferentes. Deve-se considerar estas probabilidades no cálculo do escore crítico.
- O terceiro aspecto se refere ao custo de classificar erroneamente uma observação. Se o custo de classificar, por exemplo, um elemento do grupo 1 no grupo 2 for diferente do custo de alocar um elemento do grupo 2 no grupo 1, o escore crítico precisa levar isso em consideração também.

419

Escore crítico ou ponto de corte

- **Erro do Tipo I => a seleção de um cliente para a concessão de crédito, que poderia ter problemas financeiros não capturados pela equação discriminante.**
- **Erro do Tipo II => Perder um bom cliente em função, também, de uma incorreta avaliação. Este erro é menos “prejudicial” ao credor, à primeira vista: perde-se o cliente, mas não o capital! O valor da perda tende a ser zero. Entretanto, poderá o cliente reduzir suas operações, o que leva a instituição a perder, no geral, bem mais.**

420

Análise discriminante múltipla

- Os problemas que envolvem mais de dois grupos de classificação das observações são tratados pela análise discriminante dita múltipla.
- Utilizando-se o mesmo exemplo hipotético do Banco Varejo, toma-se uma amostra aleatória de 20 observações classificadas em 3 grupos:
 - ✓ Grupo 1: margem de contribuição satisfatória;
 - ✓ Grupo 2: margem de contribuição aceitável;
 - ✓ Grupo 3: margem de contribuição não satisfatória.
- E os dados correspondentes das variáveis renda e número de dependentes.

421

Análise discriminante múltipla

Observ.	Grupo	Renda	Depend.
1	1	3.400	3
2	1	2.400	2
3	1	2.700	2
4	1	2.300	2
5	1	3.100	1
6	1	2.200	2
Média 1		2.683	2

422

Análise discriminante múltipla

Observ.	Grupo	Renda	Depend.
7	2	4.900	5
8	2	2.800	3
9	2	3.400	4
10	2	4.200	5
11	2	3.800	5
12	2	3.400	5
13	2	2.000	3
Média 2		3.500	4,29

423

Análise discriminante múltipla

Observ.	Grupo	Renda	Depend.
14	3	1.100	3
15	3	1.800	3
16	3	1.100	2
17	3	1.000	2
18	3	2.600	4
19	3	600	2
20	3	1.000	5
Média 3		1.314	3,00

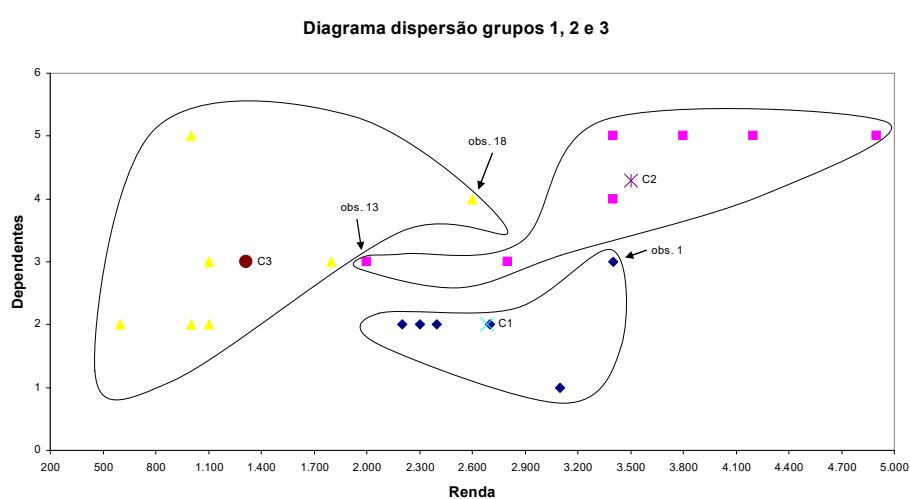
424

Análise discriminante múltipla

- O problema continua o mesmo, isto é, criar uma regra, baseada nos dados das variáveis renda e número de dependentes, capaz de identificar a qual grupo pertencerá uma nova observação. Agora, porém, o número de grupos é 3.
- Uma visualização gráfica, possível somente porque as variáveis independentes são limitadas a duas, é mostrada a seguir.
- A idéia básica é a mesma que envolve dois grupos. Assim, deverá ser encontrada uma regra que classifique os dados, levando em conta os pontos centróides rotulados no diagrama como C1, C2 e C3.
- A observação da figura evidencia que os pontos, em geral, gravitam em torno dos respectivos centróides.

425

Análise discriminante múltipla



426

Análise discriminante múltipla

- Uma solução possível seria a mensuração das distâncias entre os pontos observados e os centróides. Assim, as observações seriam classificadas segundo a maior proximidade de um ponto centróide.
- Medir as distâncias de um ponto observado aos centróides para alocar a observação ao grupo cujo centróide estiver mais próximo é, porém, deficiente sob o ponto de vista estatístico, por não considerar a variância das variáveis independentes.
- Uma forma de contornar esse aspecto é conhecida como Medida da Distância de Mahalanobis, apresentada a seguir, que considera as variâncias da variáveis independentes.
- De sua aplicação resulta a avaliação da distância de uma observação ao centróide de cada grupo, considerada a variância; a observação pertencerá ao grupo do qual estiver mais próximo.

427

Análise discriminante múltipla

$$D = \sqrt{\sum \frac{(x - \bar{x})^2}{s^2}}$$

- Onde:
- D é distância da observação ao centróide do grupo;
- x é o valor assumido pela variável independente na observação;
- \bar{x} é o valor médio da variável independente considerada dentro de um grupo;
- s^2 é a variância da variável independente considerada dentro de um grupo.
- Como exemplo, calcula-se as distâncias da observação 1:

$$D_{11} = \sqrt{\frac{(3.400 - 2.683)^2}{229.667} + \frac{(3 - 2)^2}{0,400}} = 2,18 \quad D_{13} = \sqrt{\frac{(3.400 - 1.314)^2}{448.095} + \frac{(3 - 3)^2}{1,3333}} = 3,12$$

$$D_{12} = \sqrt{\frac{(3.400 - 3.500)^2}{883.333} + \frac{(3 - 4,29)^2}{0,905}} = 1,36$$

Observando a menor distância, a observação será atribuída ao grupo 2.

428