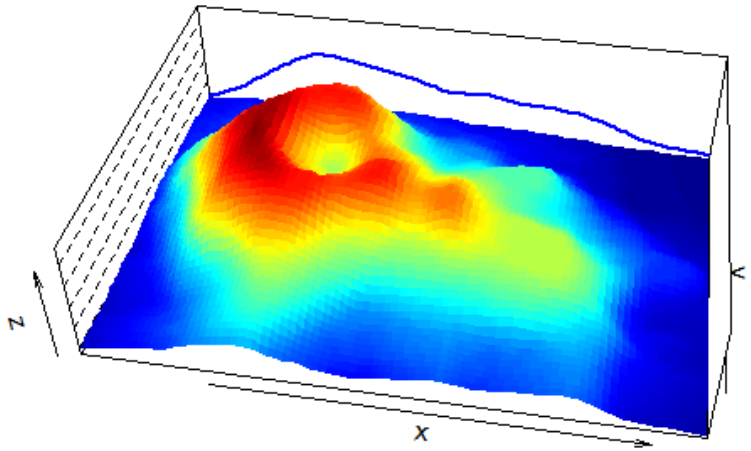


Estatística

Profa. Dra. Juliana Garcia Cespedes
PEDS - Programa de Excelência em Data Science
ITA+ITAÚ

Apresentação

Juliana Garcia Cespedes
Doutora em Estatística

Graduada em Matemática pela UNESP - Campus Rio Claro.
Mestre em Estatística e Experimentação Agronômica pela ESALQ - USP (2004)
Doutorado em Estatística e Experimentação Agronômica pela ESALQ - USP (2008).

Atualmente é professora adjunta da UNIFESP - São José dos Campos.
Docente do programa de Pós-Graduação em Pesquisa Operacional ITA/UNIFESP.

Áreas de Pesquisa:
- Inferência Bayesiana
- Modelos de Fronteira Estocástica
- Planejamento de Experimentos
- Probabilidade e Estatística Aplicadas

Ementa

- Estatística descritiva.
- Probabilidade.

Profª. Juliana

- Variáveis aleatórias.
- Modelos probabilísticos.
- Planejamento amostral.
- Intervalo de confiança
- Teste de hipóteses.
- Regressão linear e logística.
- Processos estocásticos.

Prof. Mauri
Profª. Denise
Prof. Marujo

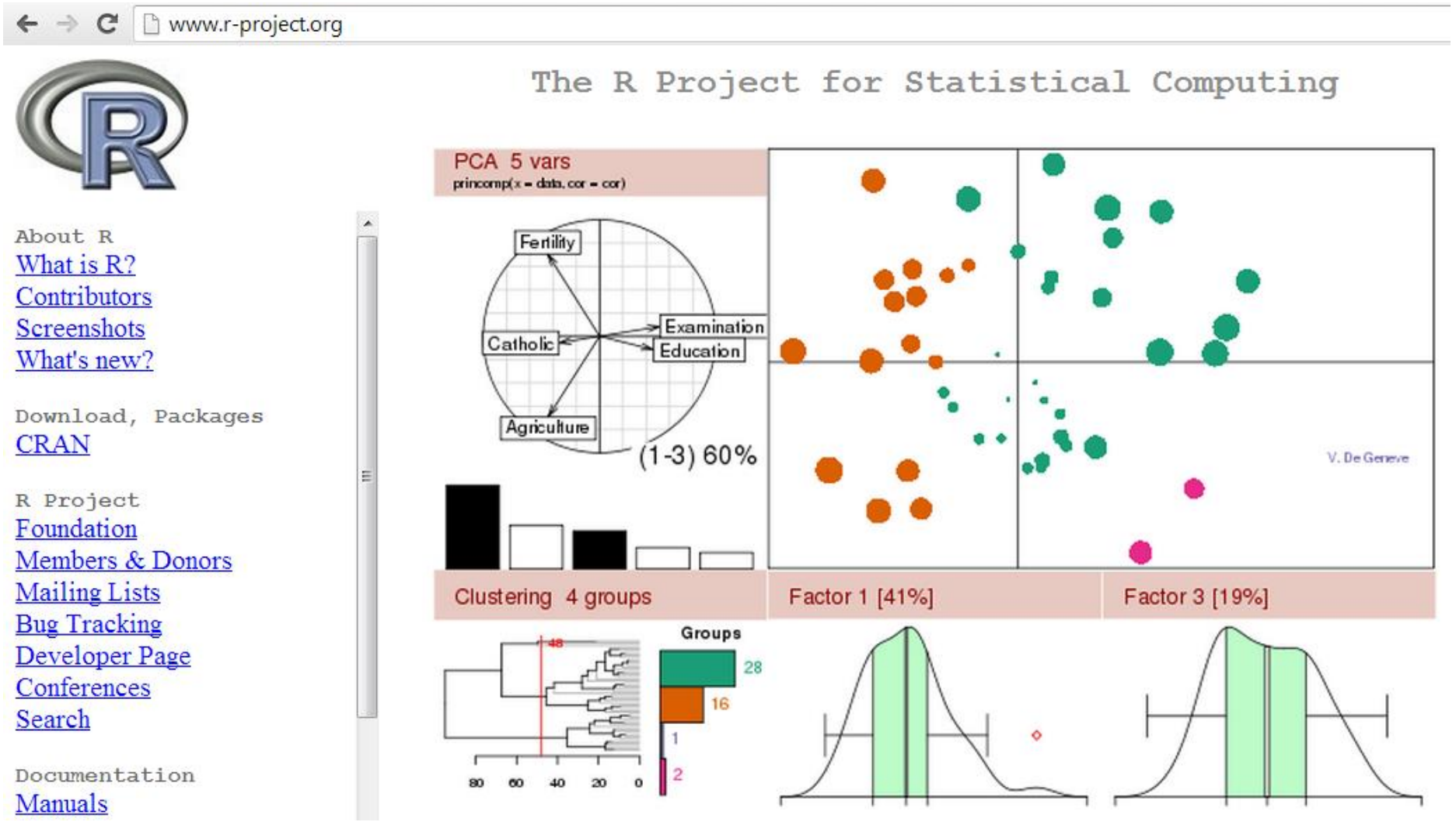
- Estatística Bayesiana.

Profª. Juliana

Bibliografia: Profa. Juliana

- BUSSAB, W. O; MORETTIN, P. A. ***Estatística Básica***, 6ª Ed. Editora Saraiva, 2010
- MAGALHÃES, M.N.; LIMA, A.C.P. ***Noções de probabilidade e estatística***. 7ª Ed. EDUSP, 2010
- CASELLA, G.; BERGER, R. ***Inferência Estatística***. Cengage Learning, 2010.
- GELMAN, A.; CARLIN, H.S.; RUBIN, D.B. ***Bayesian Data Analysis***. Second Edition, Chapman & Hall, 2003

Programa Computacional



O que é estatística?

- *“There are three kinds of lies — lies, damnable lies, and statistics”*

- *Há três espécies de mentiras: mentiras, mentiras deslavadas e estatísticas.*

- ***Benjamin Disraeli (1804 — 1881)*** *citado em "Hearings" - Página 427, United States. Congress. House. Committee on Education - 1928*

Estatística

- **Estatística:** é a ciência que tem por objetivo **planejar, coletar, tabular, analisar e interpretar informações** e delas extrair conclusões que permitam **tomar decisões** acertadas mediante incertezas.

Estatística Descritiva versus Inferência

- A **estatística descritiva** é um conjunto de técnicas destinadas a descrever e resumir os dados.

- **Inferência estatística** é o estudo de técnicas que possibilitam extrapolar informações e conclusões obtidas a partir de subconjuntos de valores (amostra) para a população inteira.

População e amostra

População é a totalidade dos elementos ou de um atributo dos elementos que estão sob investigação.

Amostra é qualquer subconjunto da população.

Exemplo: Estudar os efeitos colaterais de determinado anticoncepcional oral.

Quem é a população?

- **População seriam todas as mulheres**.

- Acontece que não é sempre fácil ou possível relacionar os efeitos da droga com toda a população. Pode ser, por exemplo, que determinado subgrupo feminino seja mais sensível a determinado efeito.

- Suponha que esse subgrupo seja o das **mulheres hipertensas**.

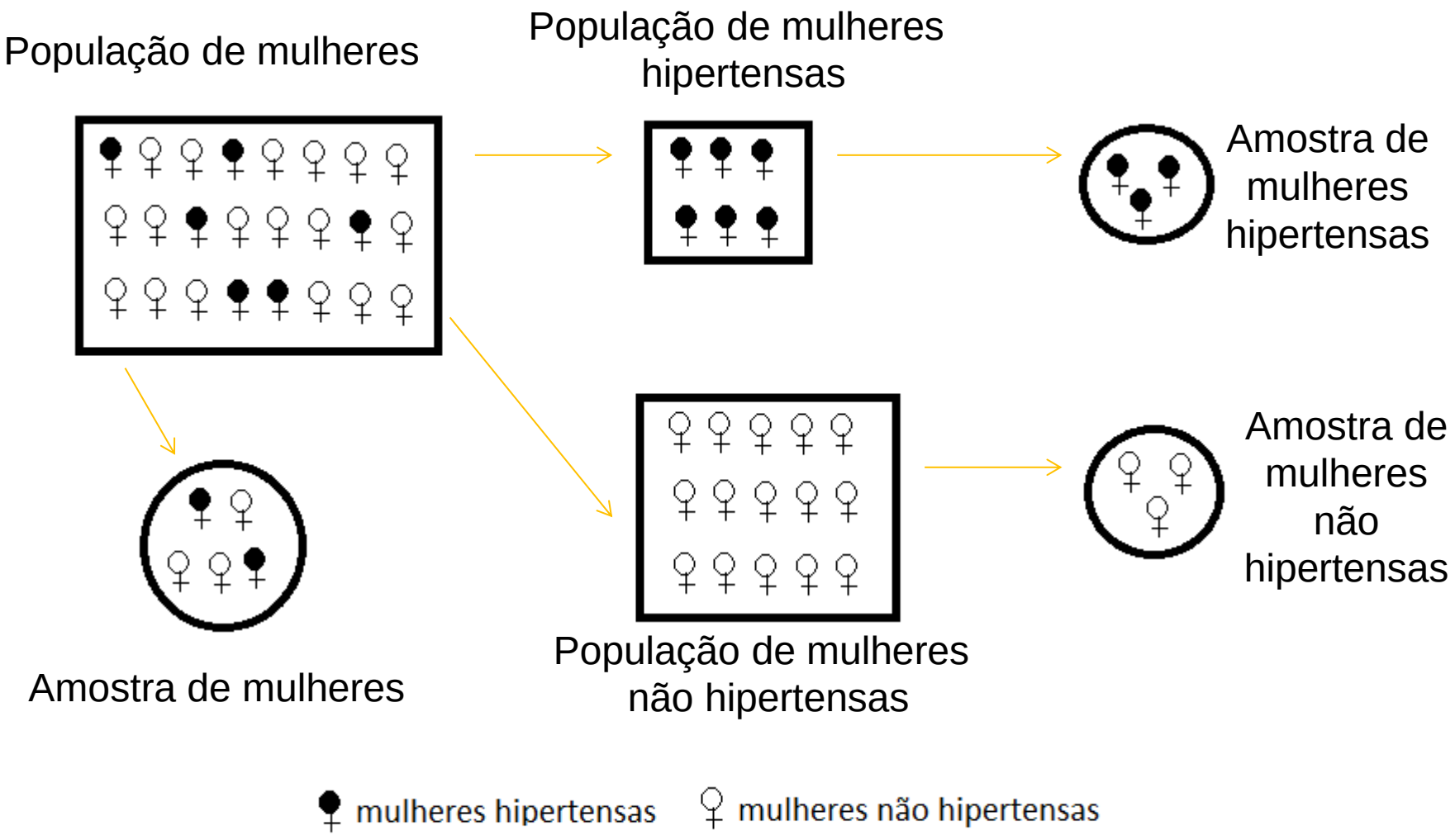
- Esse subgrupo é uma **amostra** ou uma **nova população**?

- Ao adicionar a característica “hipertensas” às mulheres, definiu-se uma **nova população**.

- Não se deve pensar que o novo grupo é uma amostra do primeiro, simplesmente porque é um subconjunto dele.

- A razão disto é que o grupo “**mulheres hipertensas**” possui todas as mulheres que tem pressão arterial elevada, portanto é uma população e costumamos chamá-la de **população objetivo**.

População e amostra



Medidas descritivas
Aula 1

Problema

Uma empresa está interessada em fazer um levantamento sobre alguns aspectos socioeconômicos dos seus empregados. Para isso, ele coletou sete informações sobre 36 colaboradores da seção de orçamentos:

1. Estado Civil;
2. Grau de instrução;
3. Sexo;
4. Número de filhos;
5. Idade;
6. Região de procedência;
7. Salário.

Colaborador	Estado Civil	Grau de instrução	Sexo	Número de filhos	Idade		Procedência	Salário (Salário mínimo)
					Ano	Mês		
1	solteiro	fundamental	masculino	NA	26	3	interior	4
2	casado	fundamental	feminino	1	32	10	capital	4.56
3	casado	fundamental	feminino	2	36	5	capital	5.25
4	solteiro	medio	masculino	NA	20	10	outra	5.73
5	solteiro	fundamental	feminino	NA	40	7	outra	6.26
6	casado	fundamental	masculino	0	28	0	interior	6.66
7	solteiro	fundamental	feminino	NA	41	0	interior	6.86
8	solteiro	fundamental	feminino	NA	43	4	capital	7.39
9	casado	medio	masculino	1	34	10	capital	7.59
10	solteiro	medio	feminino	NA	23	6	outra	7.44
11	casado	medio	masculino	2	33	6	interior	8.12
12	solteiro	fundamental	masculino	NA	27	11	capital	8.46
13	solteiro	medio	feminino	NA	37	5	outra	8.74
14	casado	fundamental	feminino	3	44	2	outra	8.95
15	casado	medio	feminino	0	30	5	interior	9.13
16	solteiro	medio	masculino	NA	38	8	outra	9.35
17	casado	medio	feminino	1	31	7	capital	9.77
18	casado	fundamental	feminino	2	39	7	outra	9.8
19	solteiro	superior	masculino	NA	25	8	interior	10.53
20	solteiro	medio	feminino	NA	37	4	interior	10.76
21	casado	medio	feminino	1	30	9	outra	11.06
22	solteiro	medio	feminino	NA	34	2	capital	11.59
23	solteiro	fundamental	masculino	NA	41	0	outra	12
24	casado	superior	masculino	0	26	1	outra	12.79
25	casado	medio	feminino	2	32	5	interior	13.23
26	casado	medio	feminino	2	35	0	outra	13.6
27	solteiro	fundamental	feminino	NA	46	7	outra	13.85
28	casado	medio	feminino	0	29	8	interior	14.69
29	casado	medio	masculino	5	40	6	interior	14.71
30	casado	medio	feminino	2	35	10	capital	15.99
31	solteiro	superior	feminino	NA	31	5	outra	16.22
32	casado	medio	feminino	1	36	4	interior	16.61
33	casado	superior	masculino	3	43	7	capital	17.26
34	solteiro	superior	masculino	NA	33	7	Capital	18.75
35	casado	medio	masculino	2	48	11	Capital	19.4
36	casado	superior	feminino	3	42	2	Interior	23.3

Fonte: Adpatação de Bussab e Morettin (2010).

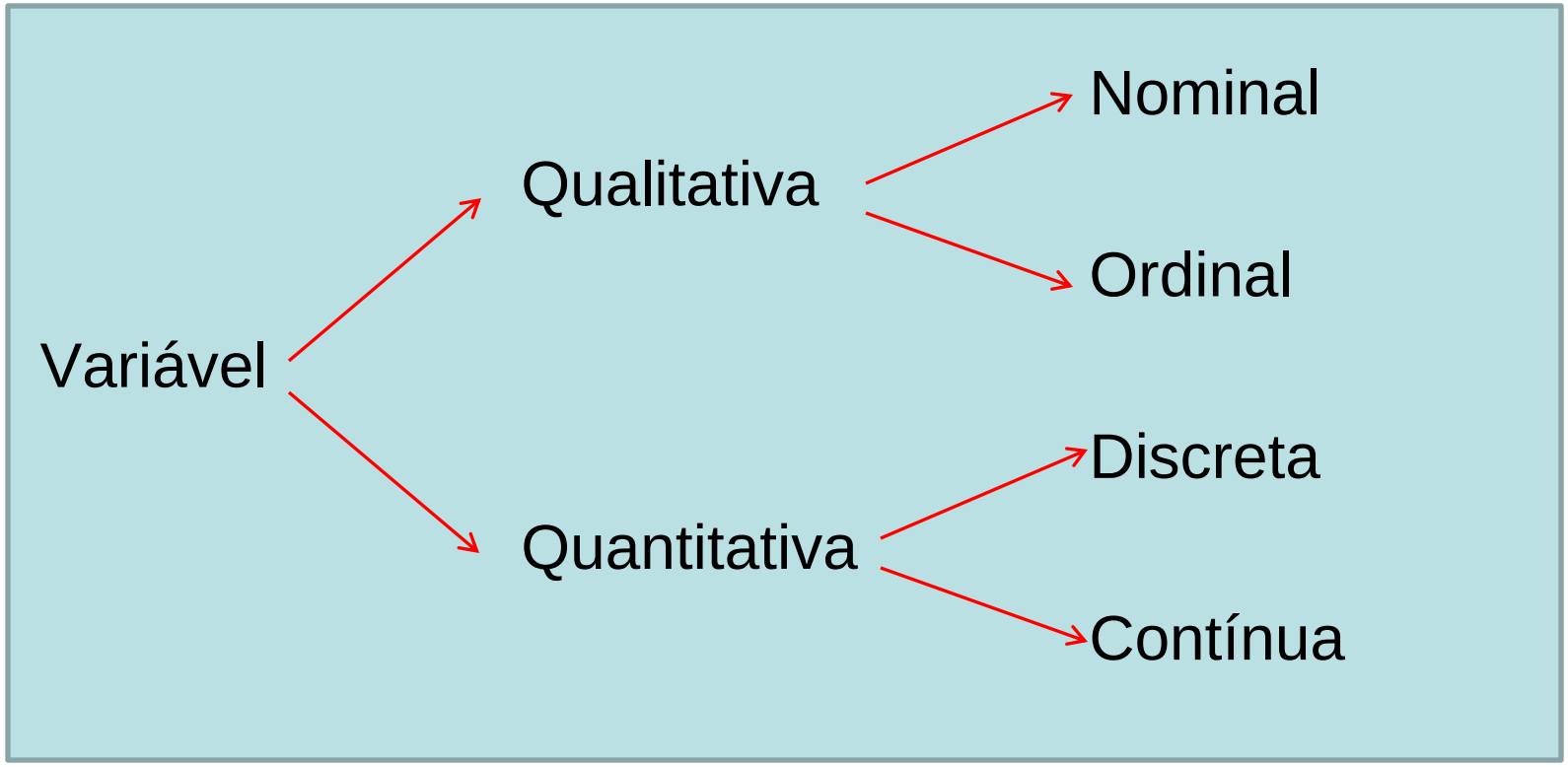
Tipos de variáveis

- Cada informação em estudo chama-se Variável.
- Para representar cada uma das variáveis em estudo atribui-se uma letra maiúscula, por exemplo, a letra X, para identificá-las.

Característica	Representação
Estado civil	X
Grau de instrução	Y
Sexo	Z
Número de filhos	F
Salário	S
Idade	U
Procedência	V

Tipos de variáveis

- As VARIÁVEIS são classificadas de acordo com o seu conteúdo e para cada tipo existe um tratamento estatístico diferente.



Tipos de variáveis

- VARIÁVEL QUALITATIVA:
  - Apresentam como possíveis realizações uma qualidade (ou atributo) do indivíduo pesquisado, por exemplo, sexo, educação, estado civil, etc.
    - **Nominal:** não existe nenhuma ordenação nas possíveis realizações.

      (sexo, estado civil)
    - **Ordinal:** existe uma ordem nos seus resultados.

      (educação, classe social)
- VARIÁVEL QUANTITATIVA:
  - Apresentam como possíveis realizações números resultantes de uma contagem ou mensuração, exemplo, número de filhos, tempo de internação, idade, etc.
    - **Discreta**: os valores pertencem a um conjunto finito ou enumerável de números. (número de filhos)
    - **Contínua**: os valores pertencem a um intervalo de números reais.

      (peso, altura)

Exercício

- Considere que uma empresa deseja conhecer o perfil de seus clientes. Para isso, observou-se as seguintes variáveis:

  - Região de Procedência
  - Tempo de relacionamento
  - Segmento
  - Faturamento médio
  - Satisfação do cliente.

Como classificar essas variáveis?

Exercício

- Considere a variável tempo de relacionamento

Como classificar essa variável?

Exercício

- Considere a variável tempo de relacionamento

Como classificar essa variável?

Em um estudo, essa variável pode ser **quantitativa** (numérica), empregando medidas expressas em anos, dias ou meses.

Em outra situação ela pode ser **qualitativa,** classificando como: *Cliente há mais de um ano, ou menos de um ano,* ou ainda: 1 (mais de um ano) ou 0 (caso contrário).

Distribuição de frequências

- Quando se estuda uma variável, o maior interesse do pesquisador é conhecer o comportamento dessa variável, analisando a ocorrência de suas possíveis realizações.

- É fácil analisar as variáveis, conhecer o seu comportamento, utilizando a Tabela 1?

Colaborador	Estado Civil	Grau de instrução	Sexo	Número de filhos	Idade		Procedência	Salário (Salario mínimo)
					Ano	Mês		
1	solteiro	fundamental	masculino	NA	26	3	interior	4
2	casado	fundamental	feminino	1	32	10	capital	4.56
3	casado	fundamental	feminino	2	36	5	capital	5.25
4	solteiro	medio	masculino	NA	20	10	outra	5.73
5	solteiro	fundamental	feminino	NA	40	7	outra	6.26
6	casado	fundamental	masculino	0	28	0	interior	6.66
7	solteiro	fundamental	feminino	NA	41	0	interior	6.86
8	solteiro	fundamental	feminino	NA	43	4	capital	7.39
9	casado	medio	masculino	1	34	10	capital	7.59
10	solteiro	medio	feminino	NA	23	6	outra	7.44
11	casado	medio	masculino	2	33	6	interior	8.12
12	solteiro	fundamental	masculino	NA	27	11	capital	8.46
13	solteiro	medio	feminino	NA	37	5	outra	8.74
14	casado	fundamental	feminino	3	44	2	outra	8.95
15	casado	medio	feminino	0	30	5	interior	9.13
16	solteiro	medio	masculino	NA	38	8	outra	9.35
17	casado	medio	feminino	1	31	7	capital	9.77
18	casado	fundamental	feminino	2	39	7	outra	9.8
19	solteiro	superior	masculino	NA	25	8	interior	10.53
20	solteiro	medio	feminino	NA	37	4	interior	10.76
21	casado	medio	feminino	1	30	9	outra	11.06
22	solteiro	medio	feminino	NA	34	2	capital	11.59
23	solteiro	fundamental	masculino	NA	41	0	outra	12
24	casado	superior	masculino	0	26	1	outra	12.79
25	casado	medio	feminino	2	32	5	interior	13.23
26	casado	medio	feminino	2	35	0	outra	13.6
27	solteiro	fundamental	feminino	NA	46	7	outra	13.85
28	casado	medio	feminino	0	29	8	interior	14.69
29	casado	medio	masculino	5	40	6	interior	14.71
30	casado	medio	feminino	2	35	10	capital	15.99
31	solteiro	superior	feminino	NA	31	5	outra	16.22
32	casado	medio	feminino	1	36	4	interior	16.61
33	casado	superior	masculino	3	43	7	capital	17.26
34	solteiro	superior	masculino	NA	33	7	Capital	18.75
35	casado	medio	masculino	2	48	11	Capital	19.4
36	casado	superior	feminino	3	42	2	Interior	23.3

Distribuição de frequências

- Podemos resumir as informações contidas na Tabela 1 construindo tabelas de frequências para cada uma das variáveis pesquisadas.

- Uma tabela de frequências possui informações sobre o número de pesquisados, a porcentagem, a proporção e a proporção acumulada de cada classe da variável analisada.

Distribuição de frequências

- Tabela de frequências da variável grau de instrução:

  (qualitativa nominal, ordinal ou quantitativa discreta em alguns casos)

Grau de instrução	Frequência f_i	Frequência relativa fr_i	Frequência acumulada fa_i	Porcentagem $100\ fr_i$
Fundamental	12	12/36 =0,333	0,333	33,3%
Médio	18	18/36 =0,500	0,833	50,0%
Superior	6	6/36 =0,167	1,000	16,7%
Total	36	36/36 =1,000		100%

Distribuição de frequências

- Observando os resultados da segunda coluna, conclui-se que dos 36 colaboradores, 12 têm ensino fundamental, 18 ensino médio e 6 possuem curso superior.

- A frequência relativa e a porcentagem são bastante úteis quando se decide comparar o resultado de pesquisas distintas.

- A frequência acumulada é utilizada para verificar onde encontra-se a maior parte da população pesquisada.

Distribuição de frequências

- Quando a variável é contínua, a construção da tabela de frequências exige certo cuidado.

- Se indicarmos cada classe que aparece na variável salário em uma tabela de frequências, não resumiremos as 36 observações num grupo menor, pois não existem observações repetidas.

- A solução é agrupar os dados por faixas de salário.

Distribuição de frequências

- Tabela de frequências do salário:

Salário	Frequência f_i	Frequência relativa fr_i	Frequência acumulada fa_i	Porcentagem $100\ f_i$
[4,00; 8,00)	10	10/36 =0,278	0,278	27,78%
[8,00; 12,00)	12	12/36 =0,333	0,611	33,33%
[12,00; 16,00)	8	8/36 =0,222	0,833	22,22%
[16,00; 20,00)	5	5/36 =0,139	0,972	13,89%
[20,00; 24,00)	1	1/36 =0,029	1,000	2,78%
Total	36	1		100%

Distribuição de frequências

- Procedendo deste modo, perde-se alguma informação. Por exemplo, não sabemos quais são os oito salários da classe de 12 a 16, a não ser que tenhamos a tabela original.
- Uma forma de interpretação é dizer que todos os oito salários são iguais ao ponto médio, 14.
- Sugere-se entre 5 a 15 classes da mesma amplitude.

Como definir o número de classes?

- Usando a equação de Sturges:

$$C = 1 + 3{,}3 * \log_{10}(N)$$

em que C é o número de classe e N é o número de dados.

Considerando N=36, tem-se C=6,14

Tabela de Sturges

Número de casos - N	Número de classes - Sturges
1	1
2	2
3 – 5	3
6 – 11	4
12 – 22	5
23 – 45	6
46 – 90	7
91 – 181	8
182 – 362	9
363 – 725	10

Outras formas

- Classes desiguais. Critério subjetivo.
- Classes de mesmo tamanho; número de classes pré-fixado.
- Critério da raiz quadrada:

$$C = \sqrt{N}$$

- Critério da desigualdade:

  O valor de *C* é o menor inteiro tal que $2C \geq N$

Amplitude de intervalo

- Para determinar a amplitude do intervalo basta subtrair o maior valor da variável do menor e dividir pelo número de classes.

$$\Delta = \frac{x_{\max} - x_{\min}}{C}$$

$$\Delta = \frac{23{,}30 - 4}{6} = 3{,}22 \approx 4$$

Exercício

- Considere que uma empresa deseja conhecer o perfil de seus clientes. Para isso, observou-se as seguintes variáveis:

  - Região de Procedência;
  - Tempo de relacionamento;
  - Segmento;
  - Faturamento médio;
  - Relacionamento com a empresa.

- Construa a tabela de frequências para as variáveis Faturamento médio e Segmento.

Programa R

- Salvar a pasta onde estão os dados do arquivo Excel na extensão CSV separado por vírgula.

- Obs: Verificar como está separada a casa decimal em seu arquivo (vírgula ou ponto).

Programa R

```
#Leitura dos dados no Programa R
dados<-read.csv2("C:/Aula Itau/dados_exer1.csv", sep=";", header=T)

# Distribuição de frequências - Variável Qualitativa
# Frequência absoluta
fa<-table(dados$seg)

#Frequência relativa
fr<-prop.table(fa)

#Frequência acumulada
fac<-cumsum(fr)
```

Programa R

#Montando a tabela

Tabela<-cbind(fa,fr,fac,por=100*fr)

Tabela

round(Tabela,2)

```
> Tabela<-cbind(fa,fr,fac,por=100*fr)
> Tabela
                          fa          fr          fac         por
                           1 0.006802721 0.006802721   0.6802721
Hoteis                    46 0.312925170 0.319727891  31.2925170
Locadora de automóveis    22 0.149659864 0.469387755  14.9659864
Restaurantes              78 0.530612245 1.000000000  53.0612245
> round(Tabela, 2)
                          fa   fr  fac   por
                           1 0.01 0.01  0.68
Hoteis                    46 0.31 0.32 31.29
Locadora de automóveis    22 0.15 0.47 14.97
Restaurantes              78 0.53 1.00 53.06
> 
```

Programa R

```
# Distribuição de frequências - Variável Quantitativa contínua
#Definir o número de classes:
 range(dados$fat_mil)
 nclass.Sturges(dados$fat_mil)

#Calcular a frequencia absoluta
 fa.var <- table(cut(dados$fat_mil, seq(0.1, 193.9, l = 9)))
 fa.var

#Retirar outlier e calcular a frequencia novamente
dados1<-dados[-146,]
fa.var <- table(cut(dados1$fat_mil, seq(0.1, 46.4, l = 9),include.lowest = T))
fa.var
```

Programa R

```
# Frequencia relativa
fr.var<- prop.table(fa.var)
Tabela<-cbind(fa.var,fr.var)
Tabela
round(Tabela,2)
```

```
> round(Tabela,2)
             fa.var fr.var
[0.1,5.89]       66    0.46
(5.89,11.7]      51    0.35
(11.7,17.5]      17    0.12
(17.5,23.2]       6    0.04
(23.2,29]         1    0.01
(29,34.8]         0    0.00
(34.8,40.6]       3    0.02
(40.6,46.4]       1    0.01
>
```

Gráficos para variáveis qualitativas

- Existem vários gráficos para representar variáveis qualitativas, os mais usados são: gráfico em barras e composição em setores, “pizza”.

GRÁFICO EM BARRAS

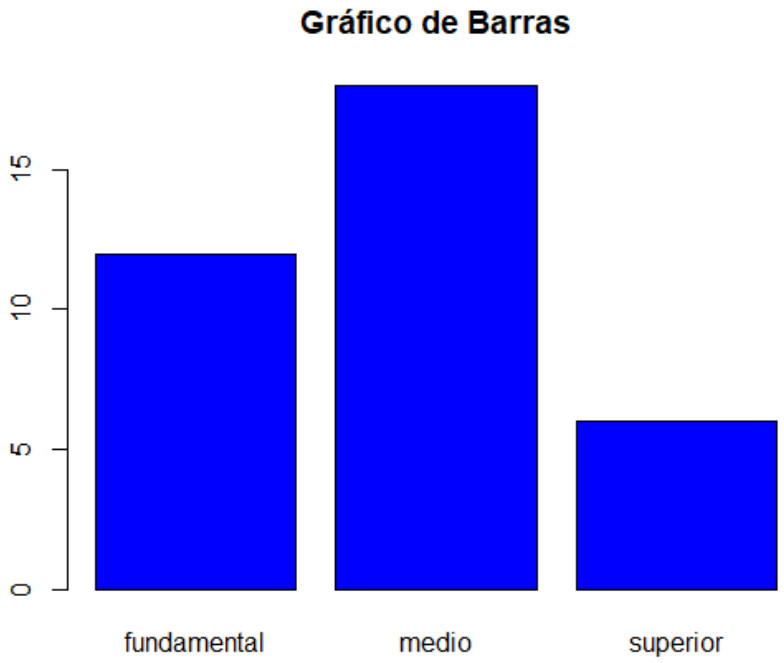
Consiste em construir retângulos ou barras, em que uma das dimensões é proporcional á magnitude a ser representada (frequência absoluta ou relativa) e a outra igual para todas as barras.

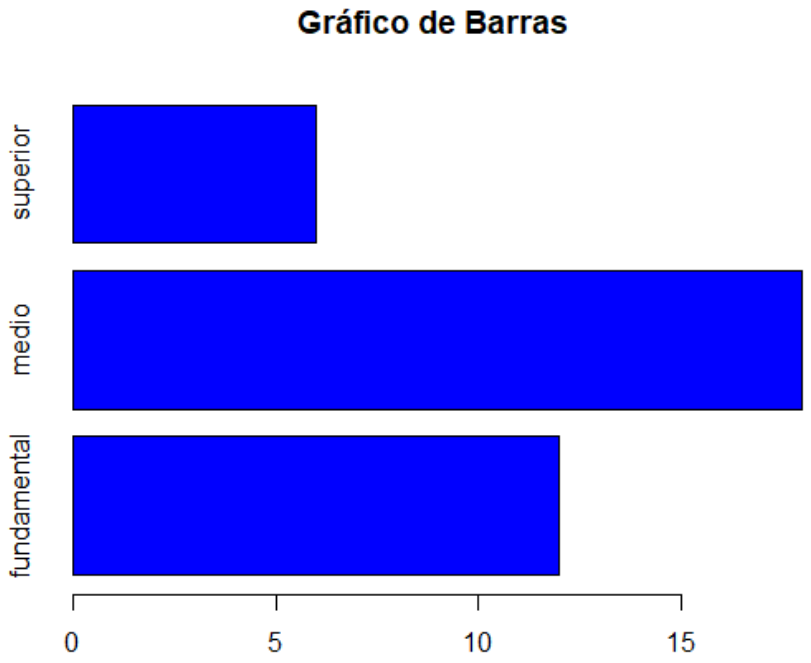
Gráficos para variáveis qualitativas

- Programa R:

barplot(table(nome), col="blue", main="Gráfico de Barras")

barplot(table(nome), col="blue", main="Gráfico de Barras", horiz=T)

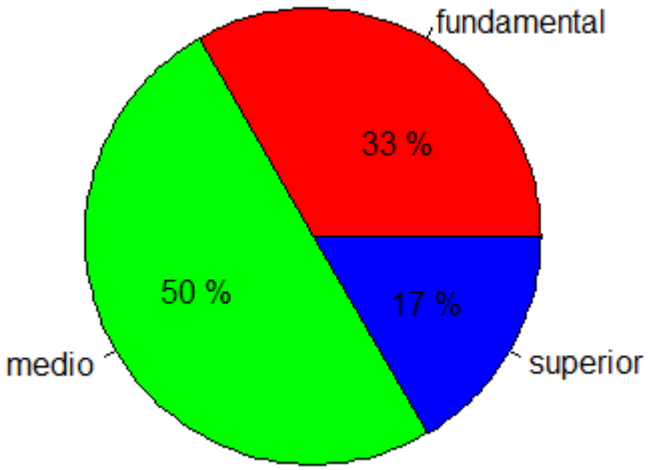




Gráficos para variáveis qualitativas

COMPOSIÇÃO EM SETORES

Representa a composição, geralmente em porcentagem, de partes de um todo. Consiste de um círculo de raio arbitrário, dividido em setores, que correspondem às partes de maneira proporcional



Programa R:
pie(table(nome))

Gráficos para variáveis quantitativas

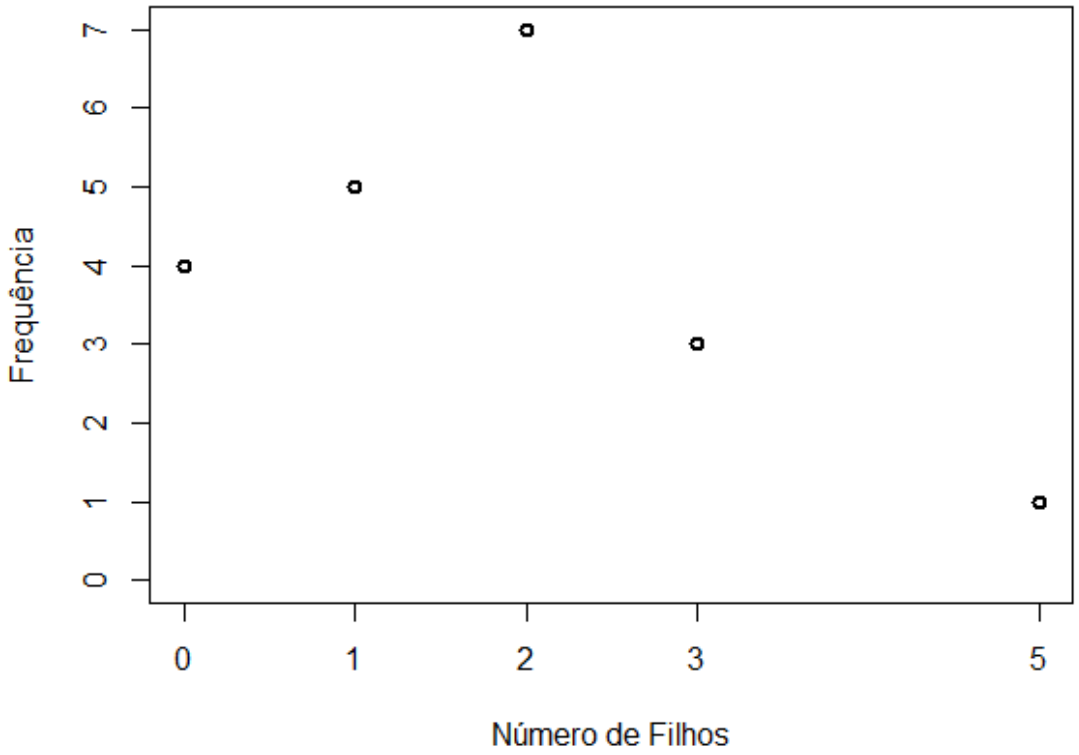
Consideram-se mais representações gráficas para variáveis quantitativas, tais como, gráfico de dispersão, ramo e folhas, histograma.

GRÁFICOS DE DISPERSÃO UNIDIMENSIONAL

1) Os pontos repetidos são empilhados um em cima do outro.

Gráficos para variáveis quantitativas

plot(table(name), type="p")



Gráficos para variáveis quantitativas

HISTOGRAMA

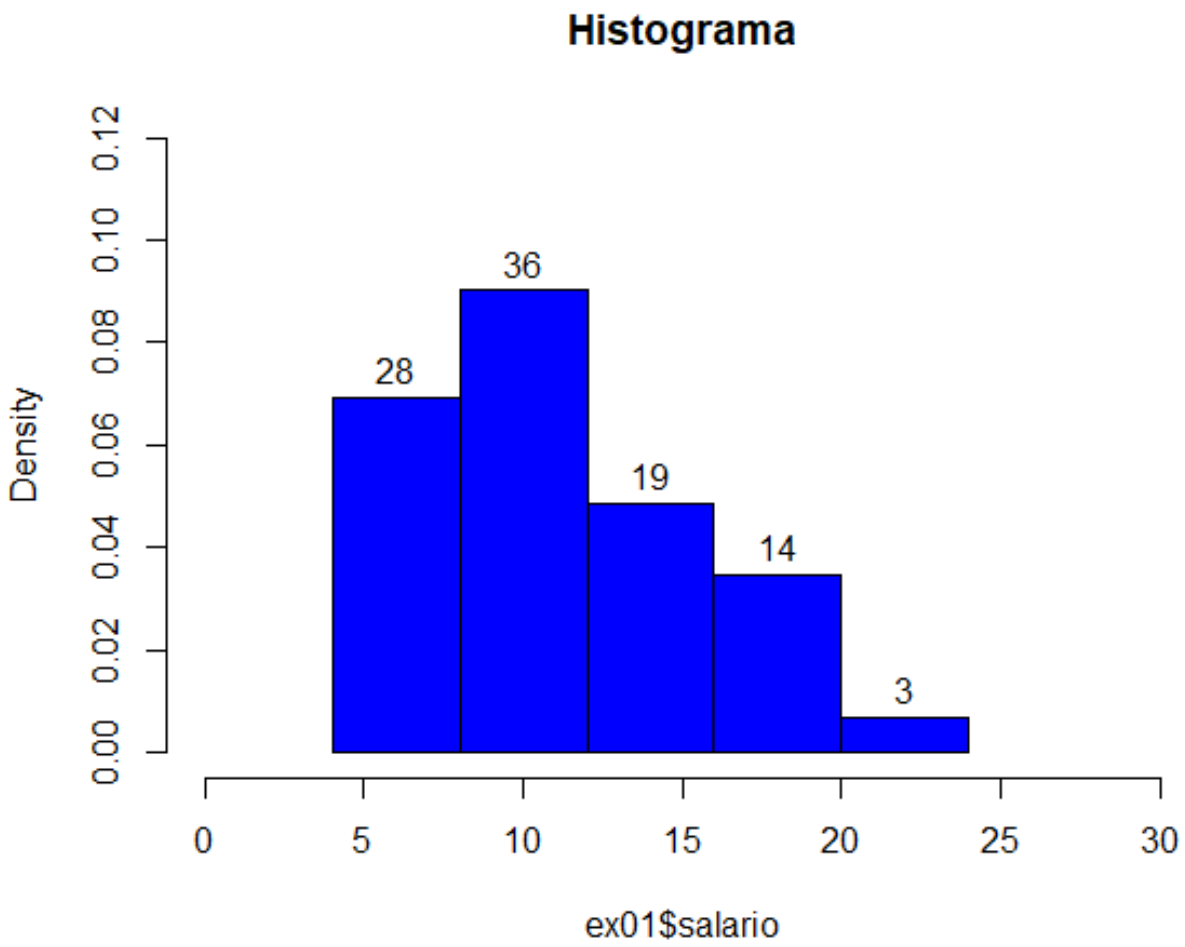
Gráfico de barras contínuas, com as bases proporcionais aos intervalos das classes e a área de cada retângulo proporcional à respectiva frequência.

Pode-se considerar tanto a freq. absoluta como a relativa.

Para que a área do retângulo seja proporcional a f_i, a sua altura deve ser proporcional a f_i/Δ_i, em que Δ_i representa a amplitude do i-ésimo intervalo.

Gráficos para variáveis quantitativas

hist(nome, breaks=c(4,8,12,16,20,24),include.lowest=T)



Ramo e folhas

- Tanto o histograma como os gráficos em barras dão uma idéia da forma da distribuição.
- Um procedimento alternativo para resumir um conjunto de valores, com o objetivo de obter uma idéia da forma de sua distribuição é o ramo e folhas. A vantagem é que não perde-se a informação sobre os dados.

- Não existe uma regra fixa para construir o gráfico, a idéia básica é dividir cada observação em duas partes: a primeira (o **ramo**) é colocada à esquerda de uma linha vertical, a segunda (a **folha**) é colocada à direita. Para a variável salários as observações 4,00 e 4,56, o ramo é o 4 e 00 e 56 são as folhas.

```
> stem(ex01$salario, scale=2)

  The decimal point is at the |

   4 | 06
   5 | 37
   6 | 379
   7 | 446
   8 | 157
   9 | 01488
  10 | 58
  11 | 16
  12 | 08
  13 | 269
  14 | 77
  15 |
  16 | 026
  17 | 3
  18 | 8
  19 | 4
  20 |
  21 |
  22 |
  23 | 3
```

stem(nome, scale=2)

Exercício 1

- Represente graficamente as variáveis:
  - Tempo de relacionamento em dias;
  - Relacionamento com a empresa;
  - Segmento.

  Use o programa R

Exercício 2

Suponha que duas empresas desejam contratá-lo e após considerar as vantagens de cada uma, você vai escolher aquela que lhe pagar melhor. Após certa pesquisa, você consegue a distribuição de salário das empresas, dadas pelos gráficos. Com base nas informações de cada gráfico, qual seria a sua decisão?