Estatística Básica

Introdução à Análise Exploratória de Dados

Renato Dourado Maia

Instituto de Ciências Agrárias

Universidade Federal de Minas Gerais

UFMG

ISTITUTO DE CIÊNCIAS AGRÁRIAS DA UFMG

Organização de Dados

- Hoje, serão discutidos alguns procedimentos que podem ser utilizados para organizar e descrever um conjunto de dados, seja em uma população ou em uma amostra.
- Poderá ser percebido como os conceitos relacionados à Teoria das Probabilidades aparecem naturalmente...

Organização de Dados

- A questão inicial é:
  - Dado um conjunto de dados, como "tratar" os valores, numéricos ou não, a fim de se extrair informações a respeito de uma ou mais características de interesse?

BASICAMENTE, SERÃO UTILIZADAS TABELAS DE FREQUÊNCIAS E GRÁFICOS, SENDO QUE A NATUREZA DOS DADOS DEVE SER SEMPRE CONSIDERADA.

Exemplo

- Suponha que um questionário foi aplicado aos alunos do primeiro ano de uma escola, fornecendo as seguintes informações:
  - Id: identificação do aluno.
  - Turma: turma em que o aluno foi alocado (A ou B).
  - Sexo: F se feminino, M se masculino.
  - Idade: idade em anos.
  - Alt: altura em metros.
  - Peso: peso em quilogramas.
  - Filhos: número de filhos na família.
  - Fuma: hábito de fumar (sim ou não).

Exemplo

- Toler: tolerância ao cigarro:
  - (I) indiferente, (P) incomoda pouco e (M) incomoda muito.
- Exerc: horas, por semana, de atividade física.
- Cine: número de vezes, por semana, em que vai ao cinema.
- OpCine: opinião a respeito das salas de cinema da cidade:
  - (B) regular e boa e (M) muito boa.
- TV: horas gastas assistindo TV, por semana.
- OpTV: opinião a respeito da qualidade da programação na TV:
  - (R) ruim, (M) média, (B) boa e (N) não sabe.

Organização de Dados

- O conjunto de informações disponíveis, após a tabulação do questionário ou pesquisa de campo, é denominado tabela de dados brutos, e contém os dados da maneira que foram coletados inicialmente. Vejamos a tabela em formato texto, e em formato planilha eletrônica...

Fonte do Exemplo: Noções de Probabilidade e Estatística / Marcos Nascimento Magalhães, Antônio Carlos Pedroso de Lima – 6 ed. rev. – São Paulo: Editora da Universidade de São Paulo, 2005.

Site com arquivos de dados e outras informações:

http://www.ime.usp.br/~noproest

Classificação de Variáveis

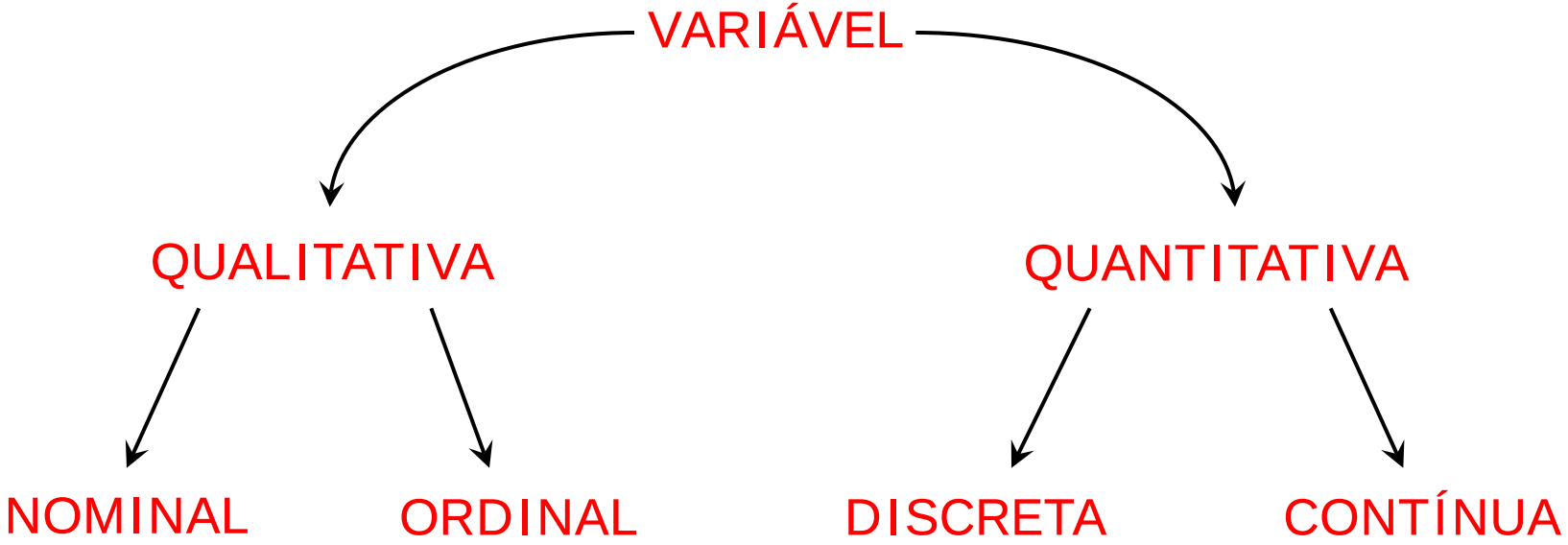


Tabela de Dados Brutos

- A tabela de dados brutos contém muita informação, mas normalmente não é prática para respondermos às questões de interesse:
  - Da tabela não conseguimos de imediato, por exemplo, dizer se os alunos se incomodam muito ou pouco com os fumantes...
- Uma alternativa interessante é construir, para cada variável, uma tabela com as informações resumidas. Essa tabela será denominada tabela de frequência, e contará os valores das variáveis e suas frequências absoluta e relativa.

Tabela de Frequência

- Para variáveis cujos valores possuem ordenação natural (qualitativas ordinais e quantitativas em geral), faz sentido a inclusão de uma coluna contendo as frequências acumuladas, para que se possam estabelecer pontos de corte com uma determinada frequência nos valores da variável...
- Tomemos como exemplo a variável Idade...

Tabela de Frequência

Tabela de Frequência para a Variável Idade

Idade	n_i	f_i	f_{ac}
17	9	0,18	0,18
18	22	0,44	0,62
19	7	0,14	0,76
20	4	0,08	0,84
21	3	0,06	0,90
22	0	0	0,90
23	2	0,04	0,94
24	1	0,02	0,96
25	2	0,04	1,00
Total	$n = 50$	1	

NOTAÇÃO

n_i → Frequência (Absoluta)

$f_i = \frac{n_i}{n}$ → Frequência Relativa

f_{ac} → Frequência Acumulada

90% dos alunos têm idade até 21 anos.

Tabela de Frequência

- A variável peso é quantitativa contínua e assim, teoricamente, seus valores podem ser qualquer número real num certo intervalo.
- Não é viável construir a tabela de frequência tal como fizemos nos casos anteriores!
- A solução é utilizar classes ou faixas de valores:
  - Não há regra formal quanto ao total de faixas, mas normalmente são utilizadas de 5 a 8, de mesma amplitude, sendo que amplitudes diferentes são interessantes para os extremos.

Tabela de Frequência

Tabela de Frequência para a Variável Peso

Peso	n_i	f_i	f_{ac}
40,0|--50,0	8	0,16	0,16
50,0|--60,0	22	0,44	0,60
60,0|--70,0	8	0,16	0,76
70,0|--80,0	6	0,12	0,88
80,0|--90,0	5	0,10	0,98
90,0|--100,0	1	0,02	1,00
Total	50	1	

Tabela de Frequência

- Quando uma variável é por natureza discreta, mas o conjunto de possíveis valores é muito grande, o caminho adequado é tratá-la como contínua e utilizar faixas.

- Um exemplo desse caso é a variável TV, que assume valores inteiros entre 0 e 30.

Tabela de Frequência

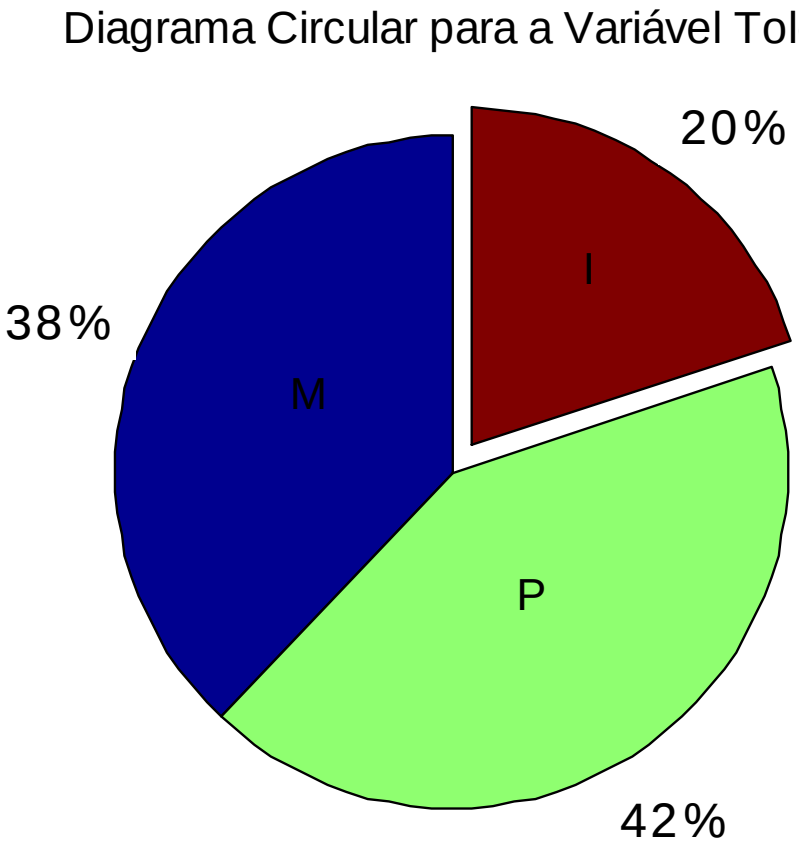
Tabela de Frequência para a Variável TV

TV	n_i	f_i	f_{ac}
0|--6	14	0,28	0,28
6|--12	17	0,34	0,62
12|--18	11	0,22	0,84
18|--24	4	0,08	0,92
24|--36	4	0,08	1,00
Total	50	1	

Gráficos

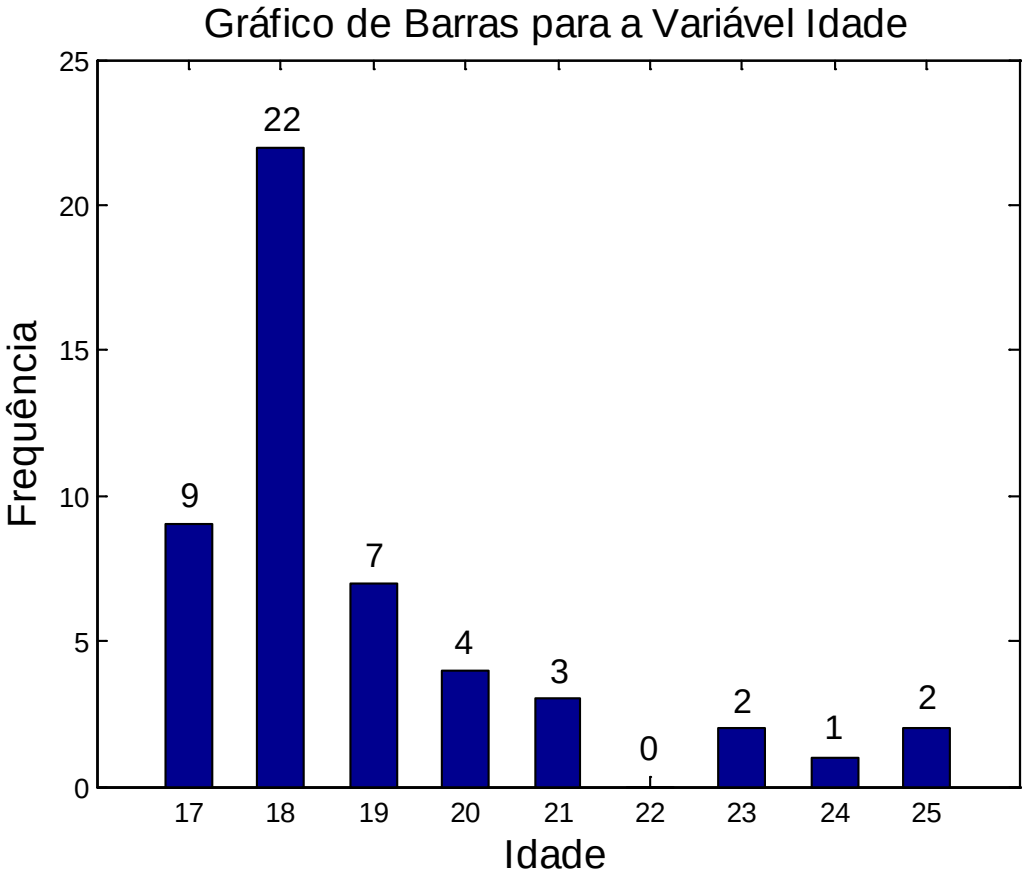
- A organização dos dados em tabelas de frequência proporciona um meio eficaz de estudo do comportamento de características de interesse.
- Em muitas vezes, a informação contida nas tabelas pode ser mais facilmente visualizada por meio de gráficos.
- Consideraremos três tipos básicos de gráficos: disco, pizza ou diagrama circular, barras e histograma.

Diagrama Circular



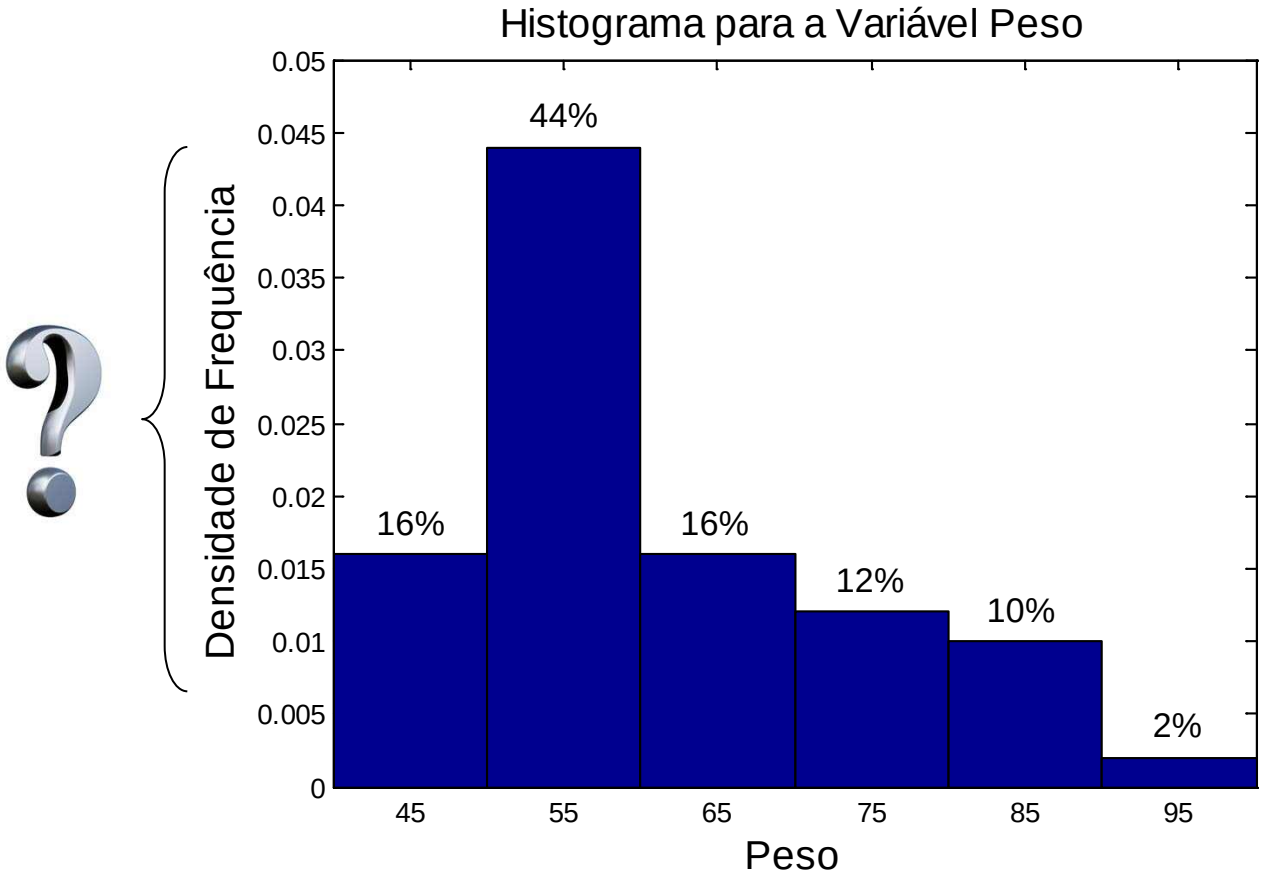
Adapta-se muito bem às variáveis qualitativas!

Gráfico de Barras



Interessante para variáveis discretas ou qualitativas ordinais!

Histograma



Adapta-se muito bem às variáveis contínuas!

Histograma

- Por que utilizar a densidade de frequência, e não a frequência absoluta?
  - Para evitar distorções, quando da utilização de amplitudes diferentes para as faixas, e em função da relação entre o histograma e a função densidade de probabilidade, que estudaremos posteriormente.

Para Refletir:

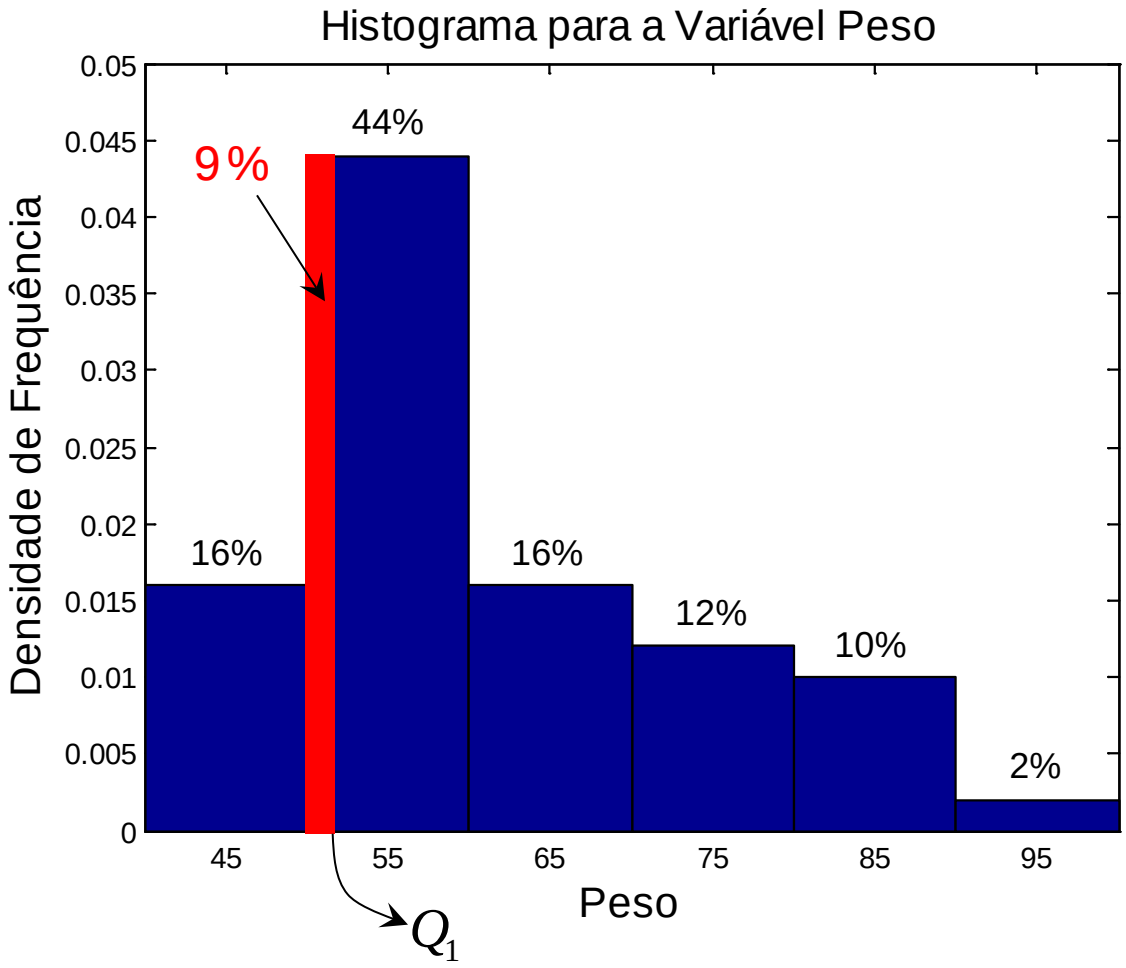
"*You can't connect the dots looking forward; you can only connect them looking backwards. So you have to trust that the dots will somehow connect in your future*".

Trecho do discurso de Steve Jobs para os formandos de Stanford em 12/06/2005

Quartis

- Primeiro Quartil:
  - Das observações ordenadas, 25% estão abaixo do valor do primeiro quartil.
- Segundo Quartil (mediana):
  - Das observações ordenadas, 50% estão abaixo do valor do primeiro quartil.
- Terceiro Quartil:
  - Das observações ordenadas, 75% estão abaixo do valor do primeiro quartil.

Histograma e o 1º Quartil

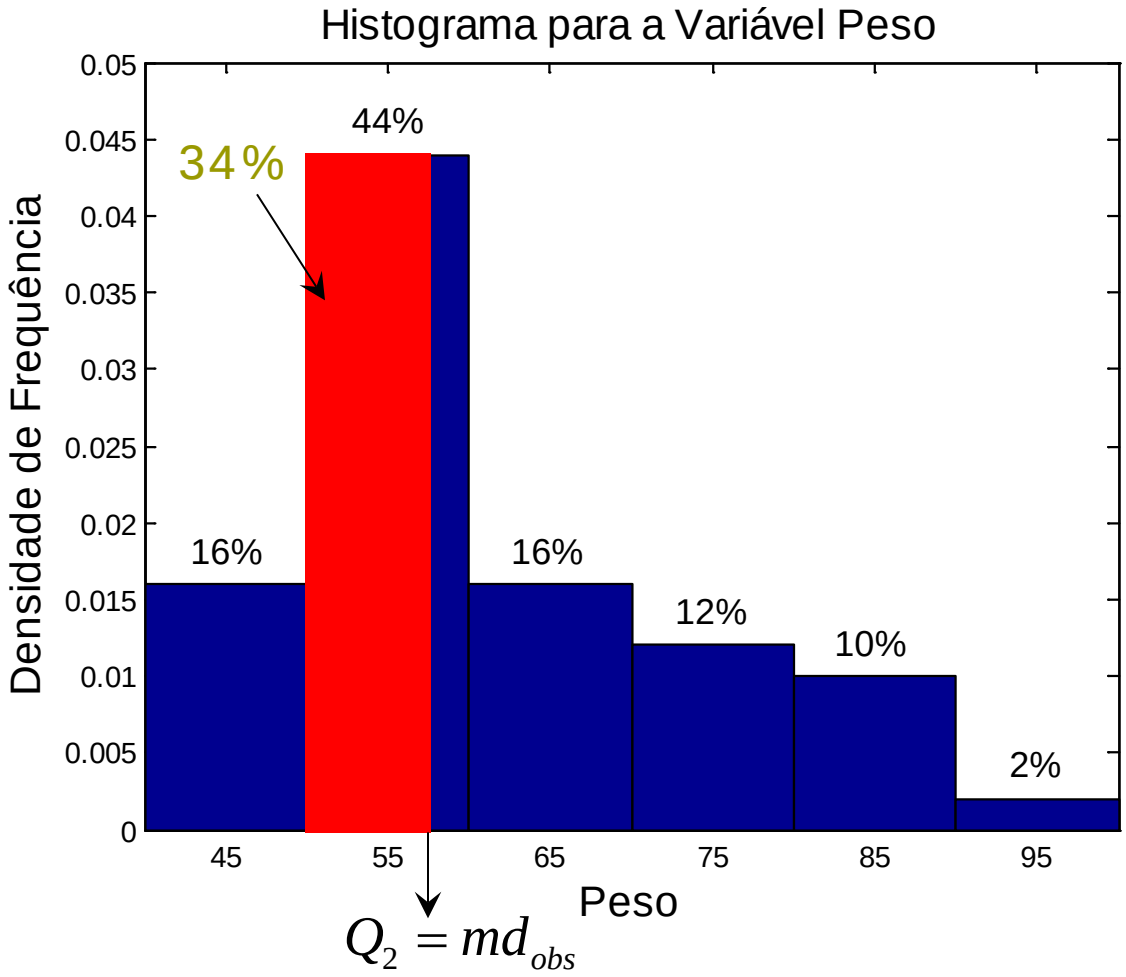


$$\frac{Q_1 - 50}{0,09} = \frac{60 - 50}{0,44}$$

$$\Downarrow$$

$$Q_1 = 52,05\ kg$$

Histograma e o 2º Quartil (Mediana)



$$\frac{md_{obs} - 50}{0,34} = \frac{60 - 50}{0,44}$$

$$\Downarrow$$

$$md_{obs} = 57,73\ kg$$

Histograma e o 3º Quartil

- Seguindo o mesmo procedimento realizado para o cálculo da mediana e do primeiro quartil, em-contra-se o valor de 69,38 kg para o terceiro quartil.

- Ao se utilizar o histograma para calcular os quar-tis, assume-se que as observações da variável em cada faixa são homogeneamente distribuídas. Apesar da suposição de homogeneidade não ser sempre verificada, ela é bastante razoável em muitas situações, e pode ser uma boa aproxima-ção da realidade.

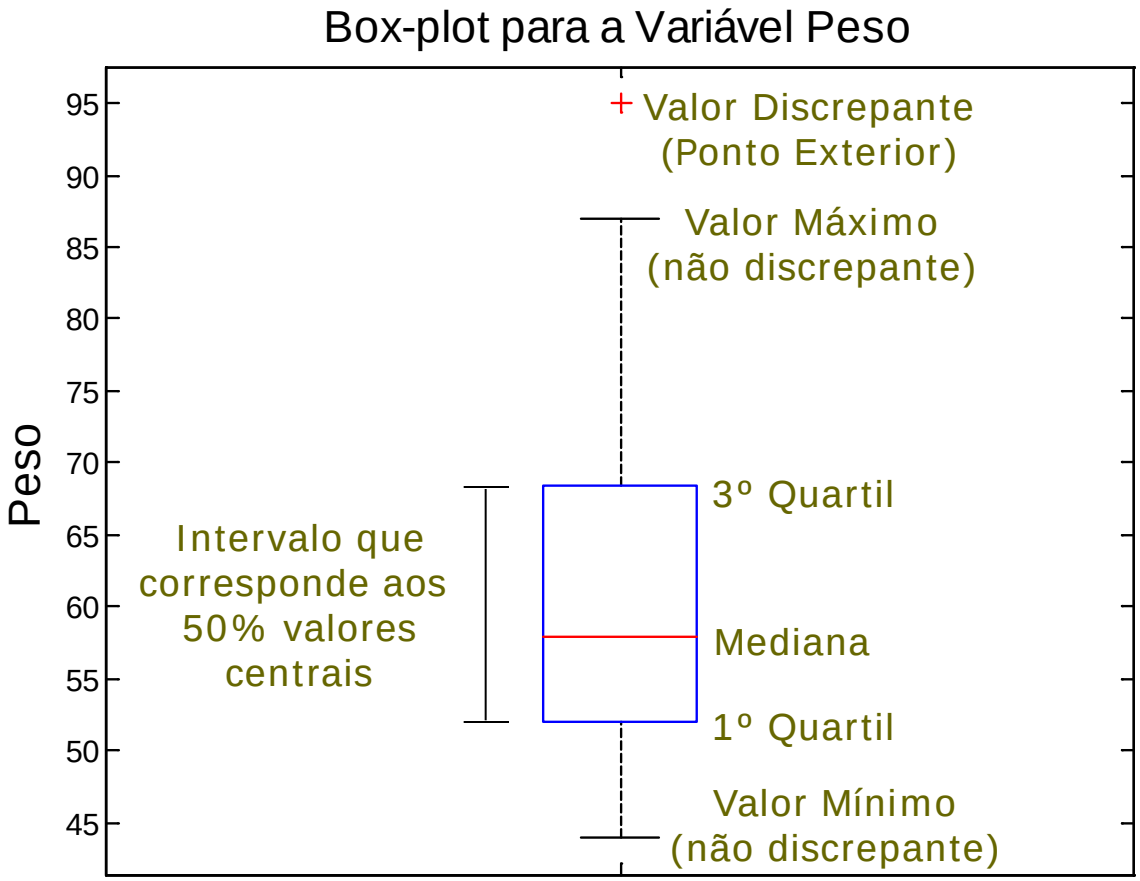
Quartis

- Para o cálculo de quartis e medianas utilizando a tabela de dados brutos, é necessário ordenar as observações e escolher os valores que dividem os dados nas proporções desejadas.
- Eventualmente, será necessário tomar médias de valores vizinhos.
- No caso de tabelas de frequências, os dados já estão ordenados, e o procedimento é similar.

Quartis e *Box-Plot*

- Uma representação gráfica interessante que envolve os quartis é o *box-plot* ou gráfico de caixa.
- Vamos analisar o *box-plot* da variável Peso, cujos quartis já calculamos...

Histograma e o 3º Quartil



Histograma e o 3º Quartil

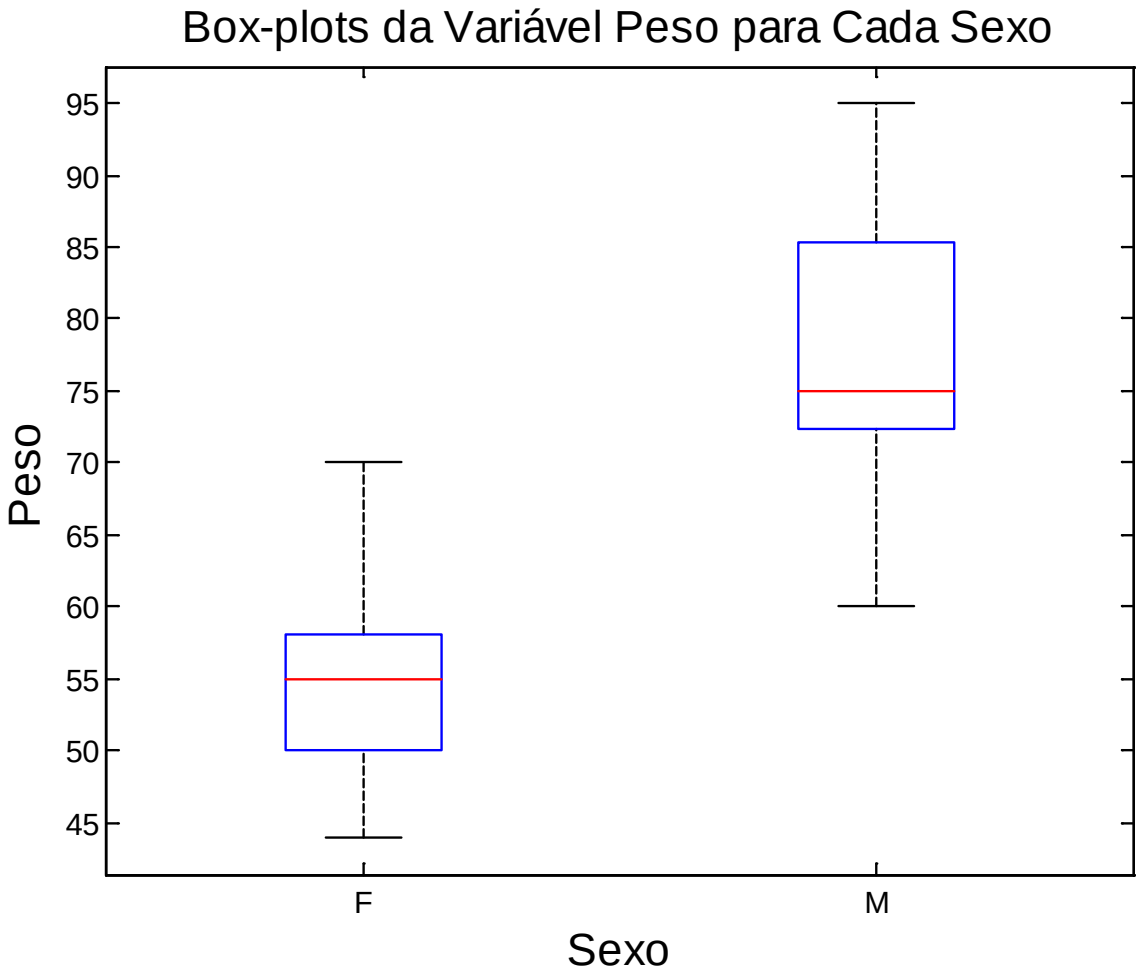
$DEQ\ (ou\ IQR) = Q_3 - Q_1 \rightarrow$ Distância entre Quartis (*Iterquartile Range*)

Observações acima de $Q_3 + (1,5) \cdot DEQ$ ou abaixo de $Q_1 - (1,5) \cdot DEQ$

Pontos exteriores ou valores atípicos *(outliers),* representados, comumente, por asteriscos.

Há diferentes convenções para a construção de *box-plots*, principalmente no que se refere à representação dos pontos exteriores: *mild outlier,* representado por uma circunferência, *extreme outlier,* representado por um círculo. Ao interpretar um *box-plot*, é importante conhecer a convenção utilizada!

Histograma e o 3º Quartil



Interpretações?

Cenas do Próximo Capítulo...

- Comentamos, no início do curso, que o desenvolvimento de ferramentas computacionais foi de extrema importância para a difusão e utilização de métodos estatísticos.

- Na próxima aula, discutiremos um pouco sobre ferramentas computacionais e brincaremos um pouco com um *software* chamado R. Para os mais curiosos, mais informações podem ser obtidas no *link* apresentado a seguir:

  - http://www.r-project.org/