Introdução à Análise Exploratória de Dados

Fonte: http://www.each.usp.br/lauretto/SIN5008_2011/aula01/aula1.pdf

Estatística Computacional SIN5008
Delhi Paiva e Marcelo Lauretto

Análise Exploratória de Dados (AED):

Finalidade:

Examinar os dados antes da aplicação de qualquer técnica estatística.

Objetivo:

Obter entendimento básico dos dados e das relações existentes entre as variáveis analisadas.

Fluxo de dados:

Após a coleta e a digitação de dados em um banco de dados o próximo passo a realizar é a *análise descritiva*.

Análise descritiva:

- Familiarizar-se com os dados, organizá-los e sintetizá-los
- Obter informações dos dados para responder as questões estudadas.

Etapas da AED

Para realizar uma AED recomenda-se seguir as seguintes etapas:

- preparar os dados para serem acessíveis a qualquer técnica estatística
- realizar um exame gráfico da natureza das variáveis individuais a analizar e uma análise descritiva que permita quantificar alguns aspectos gráficos dos dados
- realizar um exame gráfico das relações entre as variáveis analisadas e uma análise descritiva que quantifique o grau de inter-relação entre elas
- identificar os possíveis casos atípicos (***outliers***);
- avaliar, se necesário, a presença de dados ausentes (***missing values***);
- avaliar, necesário, algumas suposições básicas, como normalidade, linearidade e homocedasticidade.

Etapas da AED

A AED extrai informações de um conjunto de dados sem o peso das suposições de um modelo probabilístico. As técnicas gráficas desempenham um importante papel nesta forma de abordagem.

Estratégia de análise de dados mais conhecidas:

- ✓Estatística Clássica
- ✓Estatística Bayesiana
- ✓Análise Exploratória de Dados

De acordo com tabela, diferentemente do que é feito na Estatística Clássica e Estatística Bayesiana, na Análise Exploratória de Dados não há a imposição de um modelo aos dados, mas sim um trabalho de mineração nos dados que pode eventualmente indicar qual o melhor modelo.

Tabela 1. Estratégias de Análise

Abordagem	Estratégia
Estatística Clássica	Problema → Dados → Modelo → Análise
Estatística Bayesiana	Problema → Dados → Modelo Priori → Análise
EDA	Problema → Dados → Análise → Modelo

Técnicas Gráficas e Resumos Numéricos

Os gráficos constituem uma das formas mais eficientes de apresentação de dados. Um gráfico é, essencialmente, uma figura constituída a partir de uma tabela, pois é quase sempre possível localizar um dado tabulado num gráfico.

Enquanto as tabelas fornecem uma idéia mais precisa e possibilitam um tratamento mais rigoroso aos dados, os gráficos são mais indicados em situações cujo objetivo é dar uma visão mais rápida e fácil das variáveis às quais se referem os dados.

Portanto, a qualidade na representação gráfica deve ser pautada na **clareza, simplicidade e autoexplicação** . As técnicas gráficas desempenham um papel fundamental na AED.

Escalas de Mensuração

As técnicas a serem utilizadas dependem da natureza de mensuração das variáveis de interesse:

- **Nominal:** as variáveis são medidas em classes discretas, mas não é possível estabelecer ordem.
- **Ordinal:** as variáveis são medidas em classes discretas entre as quais é possível definir uma ordem, segundo uma relação descritível mas não quantificável.
- **Intervalar:** as variáveis assumem valores quantitativos, não possuem zero absoluto, i.e. não possuem uma medida de ausência de atributo.
- **Razão:** as variáveis assumem valores quantitativos, cuja relação exata entre estes é possível definir porque esta escala possui um zero absoluto.

Escalas de Mensuração

O tipo da análise que pode ser realizado depende da escala de medida da variável analisada. Na Tabela 2 se sugerem as representações gráficas e resumos descritivos numéricos mais recomendáveis para realizar essa análise.

Tabela 2. Representação e descrição

Escala de medida	Representações Gráficas	Medidas de tendência central	Medidas de dispersão
Nominal	Diagrama de barras Diagrama de linhas Diagrama de pizza	Moda	
Ordinal	Boxplot	Mediana	Intervalo Interquartílico
Intervalo	Histogramas Polígono de frequências	Média	Desvio padrão
Razão		Média Geométrica	Coeficiente de Variação

Tipos de variáveis

Variável:

Qualquer característica associada a uma população

Classificação:

- **Qualitativa:** são aquelas que apresentam como possíveis realizações uma qualidade ou atributo do indivíduo pesquisado
  - Nominal: sexo, cor dos olhos
  - Ordinal: classe social, grau de instrução
- **Quantitativa:** são aquelas que apresentam como possíveis realizações números resultantes de uma contagem ou mensuração
  - Contínua: peso, altura
  - Discreta: número de filhos, número de carros

Variáveis Quantitativas

Medidas de posição: valor ao redor do qual os dados estão distribuídos.

- Máximo (max): a maior observação
- Mínimo (min): a menor observação
- Moda (Mo): é o valor (ou atributo) que ocorre com maior frequência.
- Média ($\bar{X}$): soma de todos os valores da variável dividida pelo número de observações.
- Mediana (Me): valor que deixa 50 % das observações à sua esquerda
- Quartis: divide um conjunto de valores dispostos em forma crescente em quatro partes.
  - Primeiro Quartil (Q1): valor que deixa 25 % das observações à sua esquerda.
  - Terceiro Quartil (Q3): valor que deixa 75 % das observações à sua esquerda.

Variáveis Quantitativas

Medidas de Dispersão: A finalidade é encontrar um valor que resuma a variabilidade de um conjunto de dados

- Amplitude: diferença entre o valor máximo e o valor mínimo
- Intervalo-Interquartil: É a diferença entre o terceiro quartil e o primeiro quartil, ou seja, Q3 - Q1
- Variância: média dos quadrados dos desvios em relação à média aritmética
- Desvio Padrão: mede a variabilidade independente do núemro de observações e com a mesma unidade de medida da média
- Coeficiente de Variação: mede a variabilidade numa escala percentual independente da unidade de medida ou da ordem de grandeza da variável

$$CV = \frac{s}{\bar{X}} 100\ \%$$

Exame Gráfico dos Dados

Distribuição:

Histograma, ramo-e-folhas

Relação entre as variáveis:

Diagrama de dispersão

Diferenças entre grupos:

Box-plot (observações atípicas podem aparecer somente após agrupamento)

Descrição dos dados

É importante conhecer e saber construir os principais tipos de tabelas, gráficos e medidas resumo para realizar uma boa análise descritiva dos dados. Cada ferramenta fornece um tipo de informação e o seu uso depende, em geral, do tipo de variável que está sendo investigada.

Tabela 3. Tipo de variável e representação

variável qualitativa*	variável quantitativa
tabela de frequências	medidas de posição: média, mediana, moda
gráfico de barras	medidas de dispersão: variância, desvio-padrão,
diagrama circular (pizza)	amplitude, coeficiente de variação
	tabela de frequências
	histograma
	boxplot
	gráfico de linha ou sequência
	polígono de frequências

*Esta abordagem também pode ser interessante para as variáveis quantitativas discretas.

Tabela de frequências

Como o nome indica, conterá os valores da variável e suas respectivas contagens, as quais são denominadas frequências absolutas ou simplesmente, frequências.

No caso de variáveis qualitativas ou quantitativas discretas, a tabela de frequência consiste em listar os valores possíveis da variável, numéricos ou não, e fazer a contagem na tabela de dados brutos do número de suas ocorrências.

A frequência do valor i será representada por n_i, a frequência total por n e a frequência relativa por $h_i = h_i / n$.

Tabela de frequências

Para variáveis cujos valores possuem ordenação natural (qualitativas ordinais e quantitativas em geral), faz sentido incluirmos também uma coluna contendo as frequências acumuladas N_i e H_i, obtidas pela soma das frequências de todos os valores da variável, menores ou iguais ao valor considerado.

No caso das variáveis quantitativas contínuas, que podem assumir infinitos valores diferentes, a tabela de frequência precissa de classes ou faixas de valores e contamos o número de ocorrências em cada faixa.

Apesar de não adotarmos nenhuma regra formal para estabelecer as faixas, utilizaremos em geral, de 5 a 8 faixas com mesma amplitude. Eventualmente, faixas de tamanho desigual podem ser convenientes para representar valores nas extremidades da tabela.

Tabela de frequências

Classes	Intervalos	Frequência absoluta	Frequência relativa	Frequência absoluta acumulada	Frequência relativa acumulada
C	$(LI_i - LS_i)$	n_i	h_i	N_i	H_i
c_1	$(LI_1 - LS_1)$	n_1	$h_1 = \frac{n_1}{n}$	$N_1 = n_1$	$H_1 = \frac{N}{n_1} = h_1$
...	...	...	...	...	...
c_j	$(LI_j - LS_j)$	n_j	$h_j = \frac{n_j}{n}$	$N_j = n_1 + n_2 + ... + n_j$	$H_j = \frac{N}{n_j} = h_1 + h_2 + ... + h_j$
...	...	...	...	...	...
c_k	$(LI_k - LS_k)$	n_k	$h_k = \frac{n_k}{n}$	$N_k = n$	$H_k = 1$

Medidas de posição no caso de dados agrupados

- **Média:** Sejam $y_1, y_2, \ldots, y_k$ os pontos médios de cada intevalo de classe de uma distribuição de frequência de k classes

$$\bar{Y} = \frac{\sum_{j=1}^{k} n_i y_i}{n}$$

- **Mediana:**

$$Me = LI_j = c\left(\frac{(n/2) - N_{j-1}}{n_j}\right)$$

- **Moda:**

$$Mo = LI_j + c\left(\frac{n_j - n_{j-1}}{(n_j - n_{j-1}) + (n_j - n_{j+1})}\right)$$

Exemplo

Adaptado do dataset *Household Expenditures* (Aitchison, 1986):
Gastos domiciliares de 38 domicílios (HK$) em quatro grupos de despesas:

1. Moradia, gás, luz, etc
2. Alimentação, incluindo bebidas e tabaco
3. Outros bens, incluindo vestuário e bens duráveis
4. Servicos, incluindo transporte e veículos

Dataset:

- sex: sexo do chefe da família (male/female)
- children: número de filhos (adaptado por Marcelo Lauretto)
- housing, foodstuffs, othergoods, services: gastos mensais em cada grupo de despesas

Gráfico de barras

Para construir um gráfico de barras, representamos os valores da variável no eixo das abscissas e suas frequências ou porcentagens no eixo das ordenadas. Para cada valor da variável desenhamos uma barra com altura correspondendo à sua frequência ou porcentagem.

Este tipo de gráfico é interessante para as variáveis qualitativas ordinais ou quantitativas discretas, pois permite investigar a presença de tendência nos dados.

Gráfico de Barras

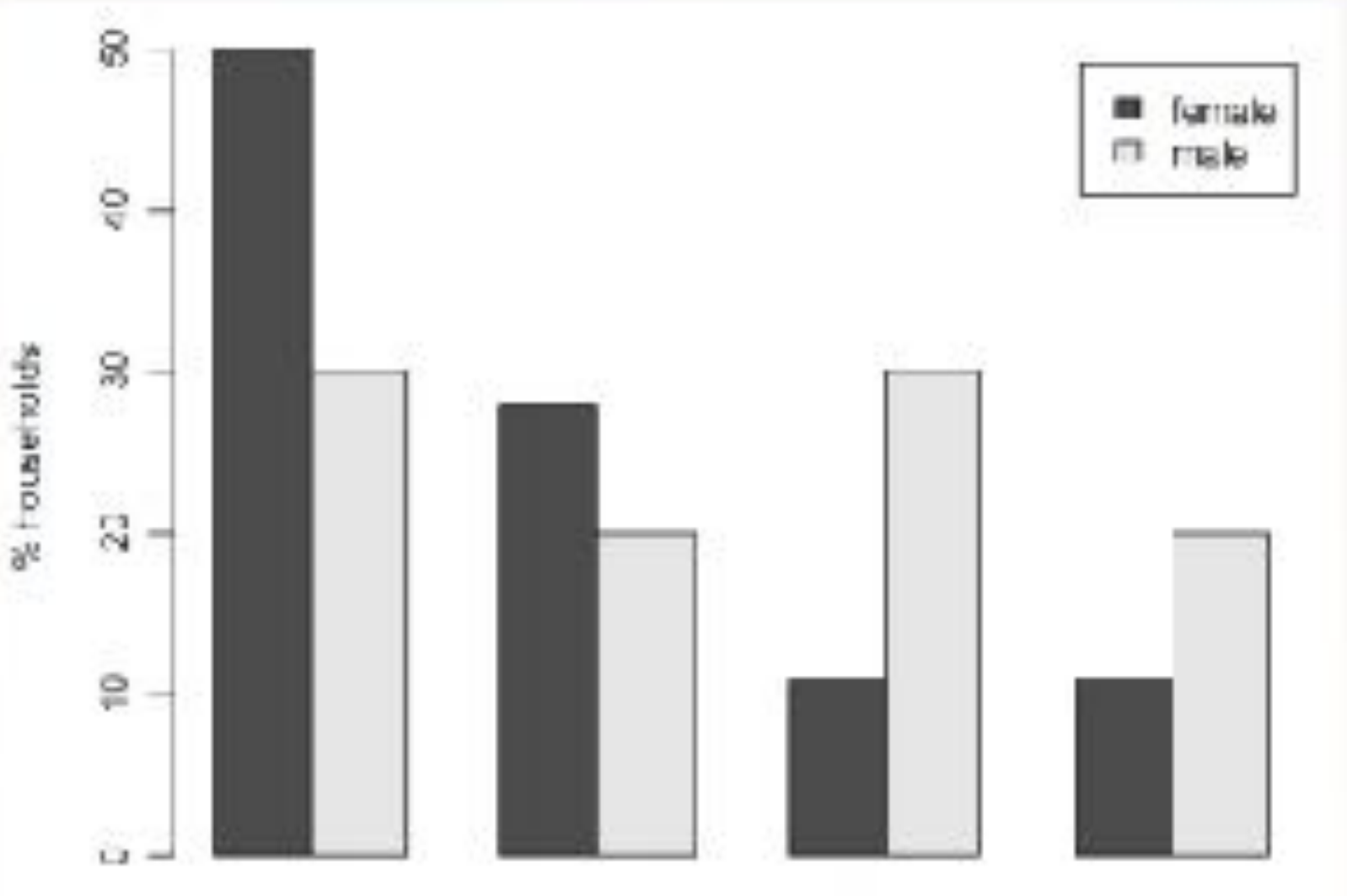


Figura: Número de filhos por sexo do chefe de familia

Diagrama Circular

Para construir um diagrama circular ou gráfico de pizza, repartimos um disco em setores circulares correspondentes às porcentagens de cada valor (calculadas multiplicando-se a frequência relativa por 100). Este tipo de gráfico adapta-se muito bem para as variáveis qualitativas nominais.

Diagrama Circular

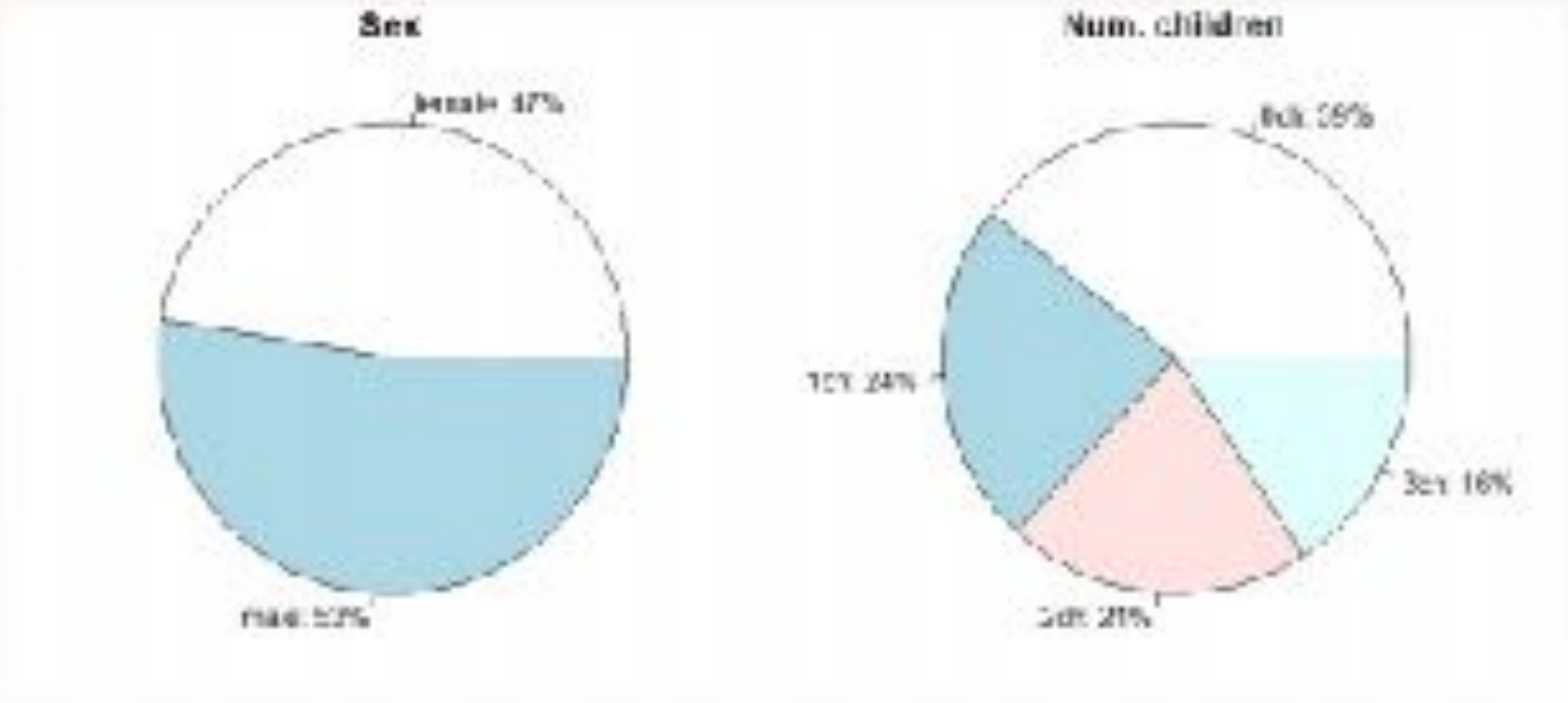


Figura: Esq: Percentual de domicílios por sexo do chefe de familia;
Dir: Percentual de domicílios por número de filhos

Diagrama Circular

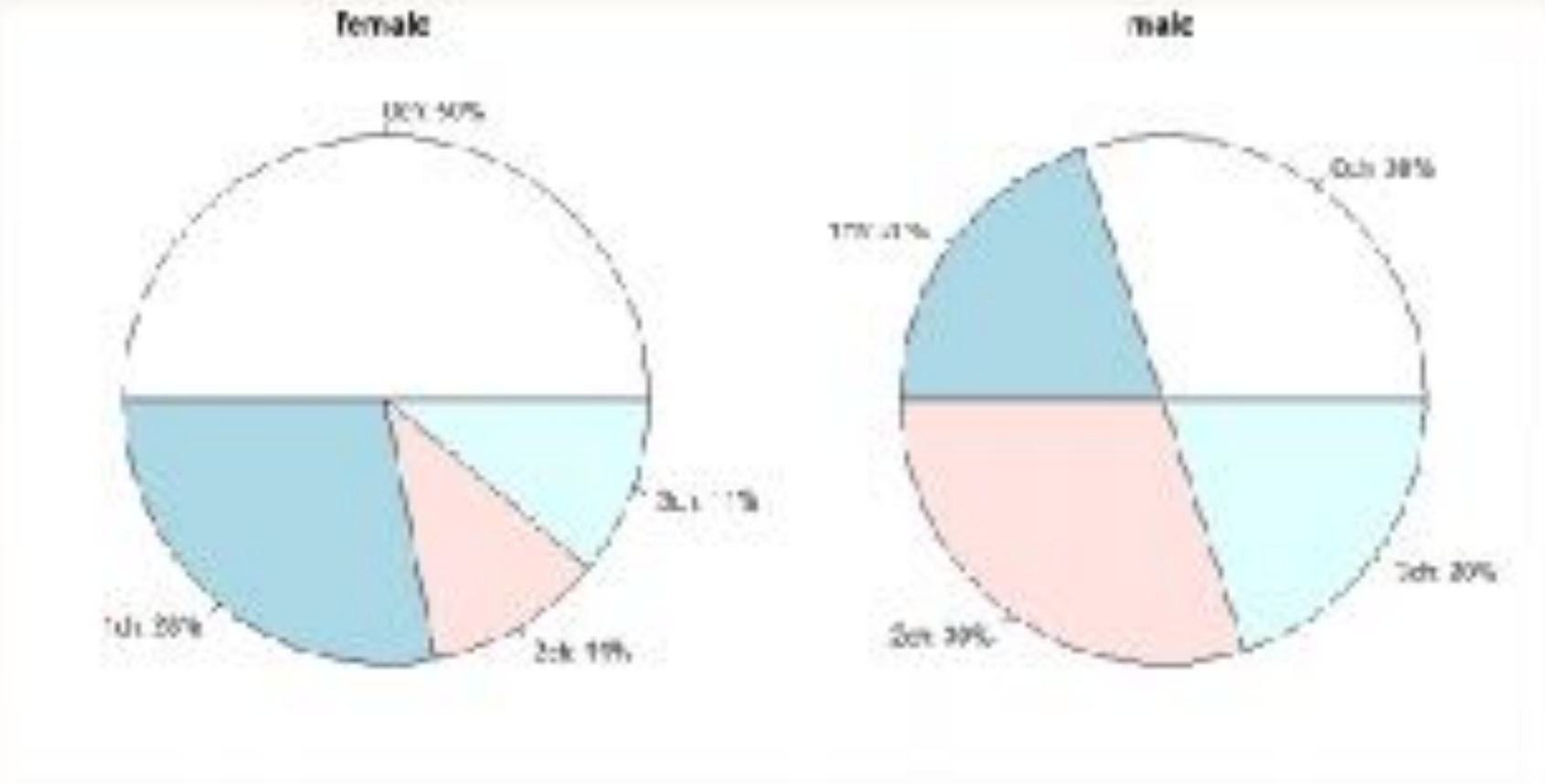


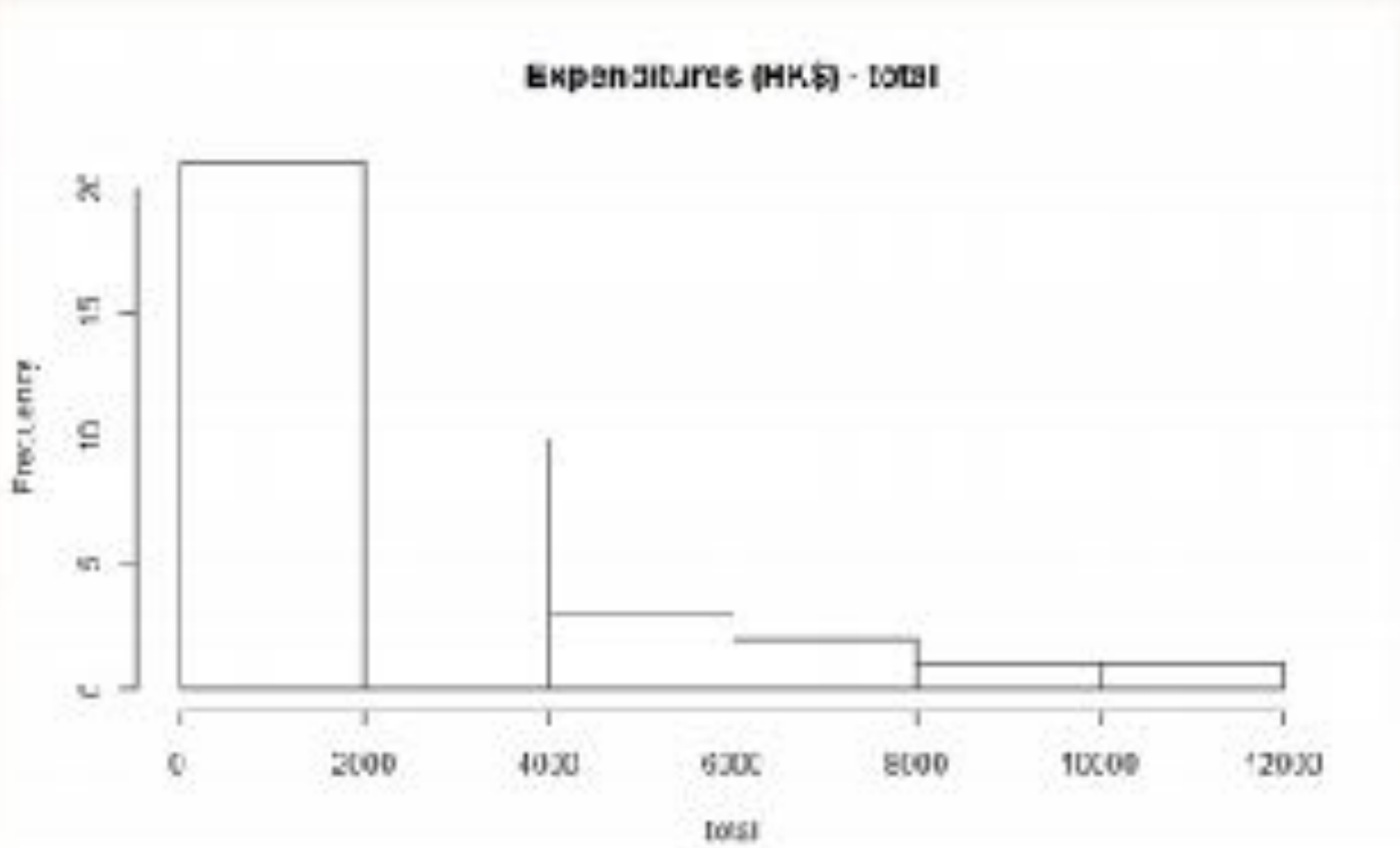
Figura: Percentuais de domicílios por número de filhos (segmentação por sexo)

Histograma

O histograma consiste em retângulos contíguos com base nas faixas de valores da variável e com área igual à frequência relativa da respectiva faixa. Desta forma, a altura de cada retângulo é denominada densidade de frequência ou simplesmente densidade definida pelo quociente da área pela amplitude da faixa.

Alguns autores utilizam a frequência absoluta ou a porcentagem na construção do histograma, o que pode ocasionar distorções (e, consequentemente, más interpretações) quando amplitudes diferentes são utilizadas nas faixas.

Histograma


Expenditures (HK$) - total
Frequency
0
5
10
15
20
0
2000
4000
6000
8000
10000
12000
total

Boxplot

Para construí-lo, desenhamos uma *caixa* com o nível superior dado pelo terceiro quartil ($Q\,3$) e o nível inferior pelo primeiro quartil ($Q\,1$) . A mediana ($Q\,2$) é representada por um traço no interior da caixa e segmentos de reta são colocados da caixa até os valores máximo e mínimo, que não sejam observações discrepantes.

O critério para decidir se uma observação é discrepante pode variar; chamaremos de discrepante os valores maiores do que $Q\,3 + 1\,,\,5 * (\,Q\,3 - Q\,1\,)$ ou menores do que $Q\,1 - 1\,,\,5 * (\,Q\,3 - Q\,1\,)$.

O Boxplot fornece informações sobre posição, dispersão, assimetria, caudas e valores discrepantes.

Boxplot

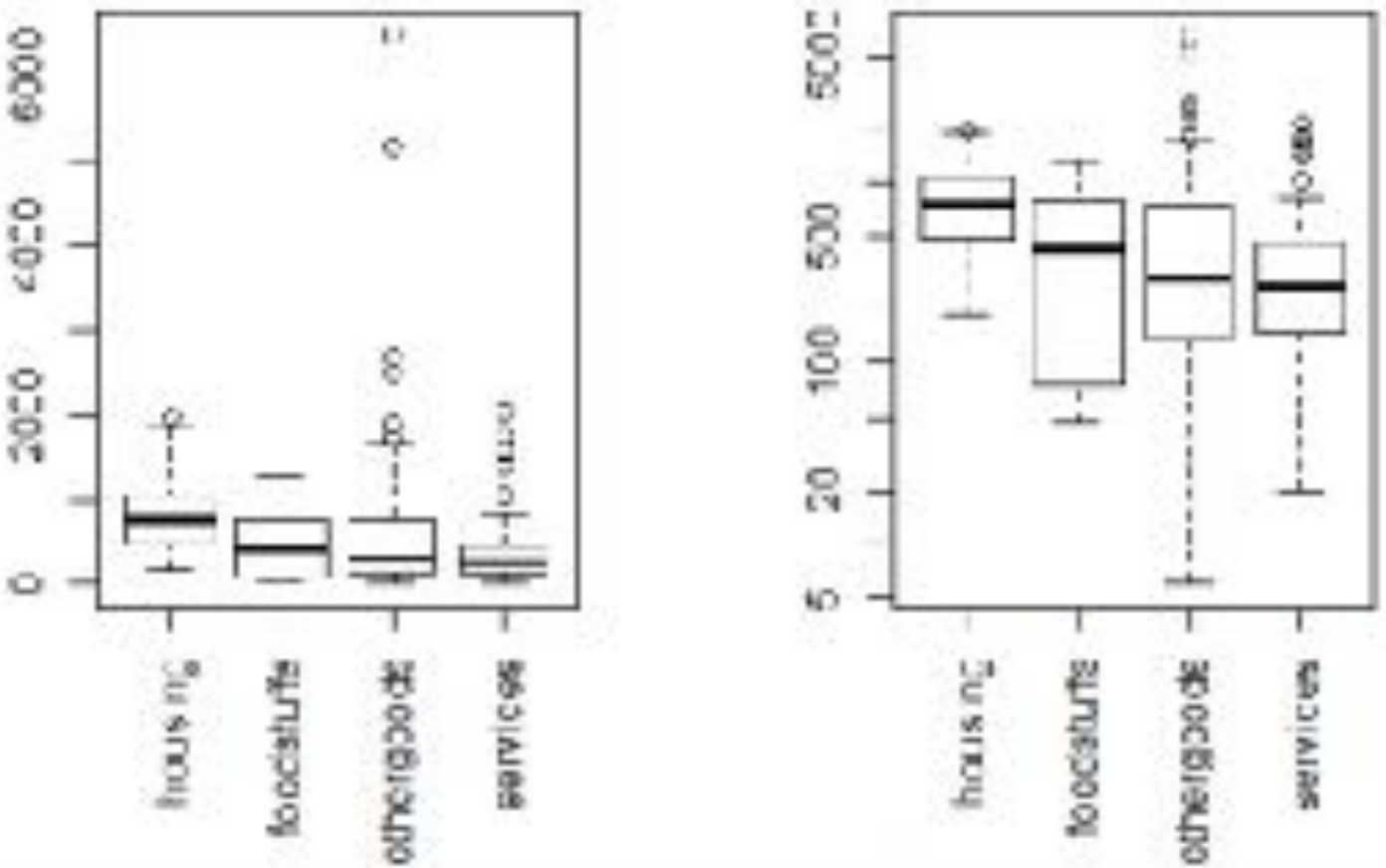


Figura: Consumos domiciliares por grupo de despesa, em escala natural (esq) e logarítmica (dir)

Gráfico de linha ou sequência

Adequados para apresentar observações medidas ao longo do tempo, enfatizando sua tendência ou periodicidade.

Polígono de frequências

Semelhante ao histograma, mas construído a partir dos pontos médios das classes.

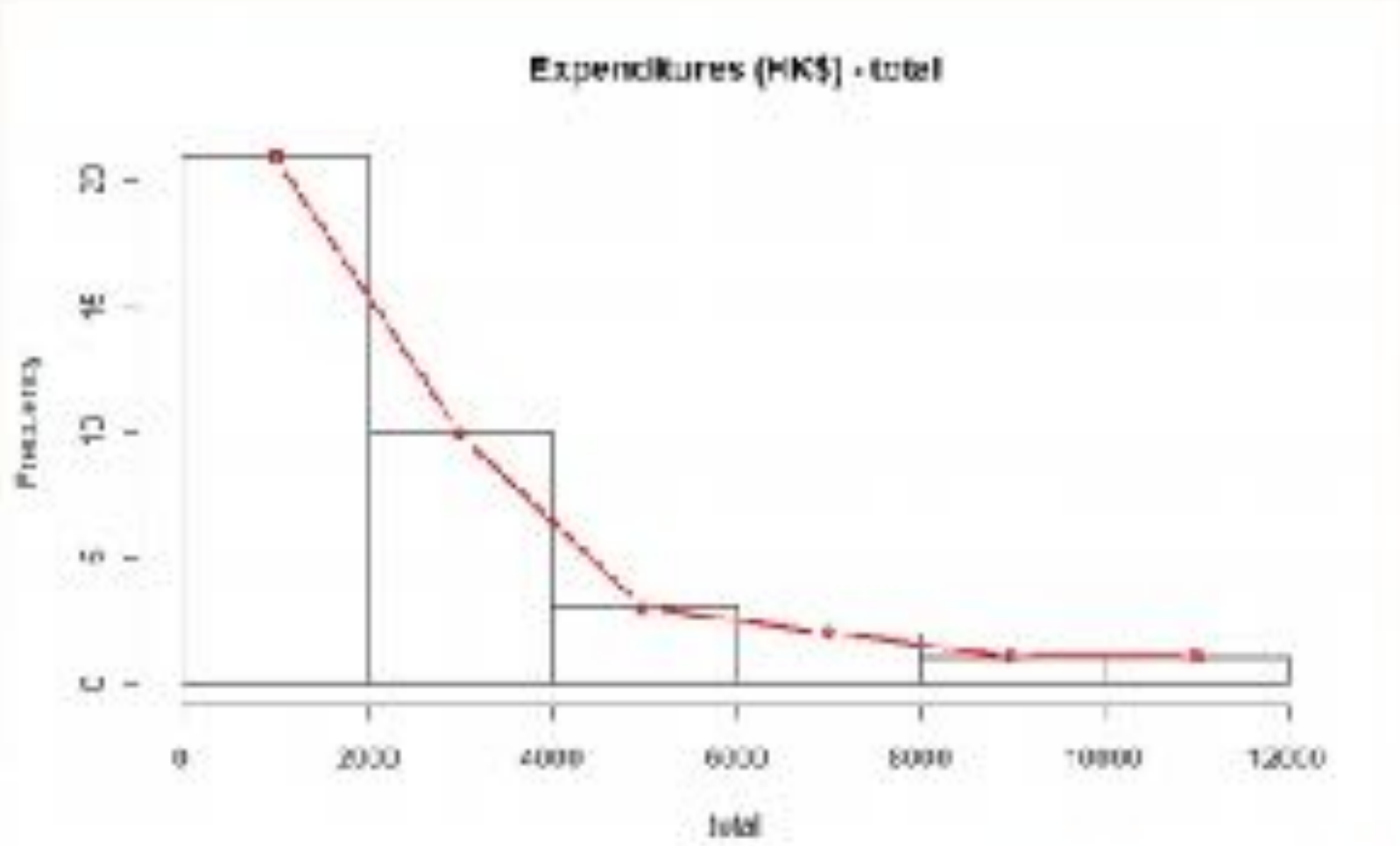


Diagrama de dispersão

Adequado para descrever o comportamento conjunto de duas variáveis quantitativas. Cada ponto do gráfico representa um par de valores observados.