Análise Exploratória de Dados

PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul

- Aula 10 -
Coleta, Preparação e Análise de Dados

Prof. Me. Lucas R. C. Pessutto

Slides adaptados do material do Prof. Lucas Silveira Kupssinskü e do Prof. Luan Fonseca Garcia

Análise de Dados

- Para conduzir uma mineração ou análise de dados de sucesso é ***essencial*** que conheçamos nossos dados.
  - Quais **tipos** de **atributos** ou **campos**?
  - Que **tipos** de **valores** cada atributo tem?
  - Como esses valores estão **distribuídos**?
  - Como podemos medir a **similaridade** entre determinados dados em relação a outros?
- Este tipo de insight facilita a análise subsequente dos dados!

Descrição Estatística de Dados

Descrição Estatística

- Precisamos estar familiarizados com os dados para poder realizar um pré-processamento efetivo.
- Descrições estatísticas permitem identificar propriedades dos dados e destacar quais valores devem ser tratados como ruído ou outliers.
  - Qual a tendência central dos meus dados?
  - Como está a dispersão dos meus dados? (Como eles estão espalhados?)
  - Qual a correlação entre eles?
  - Como visualizar isso?

Tendência Central - Média

- Medida mais simples (e efetiva) para calcularmos o centro de um conjunto de dados é a média aritmética.

Média simples:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

- Podemos associar a cada valor um peso que reflete sua importância, frequência, etc.

Média ponderada:

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}$$

Tendência Central - Média

- Problema: sensível a outliers. Um pequeno número de outliers extremos podem "corromper" a média.
  - Em uma turma onde 5 alunos tiraram 9 na prova e dois alunos tiraram 0 a média da turma seria 6.42

- Possível solução: eliminar valores extremos.
  - Ordenar os dados e descartar 2% dos valores mais altos e mais baixos antes de calcular a média.
- Chamada de **média truncada.**

Tendência Central - Mediana

- Em conjuntos de dados desbalanceados a média pode não ser uma boa métrica.
- **Conjunto desbalanceado** é aquele em que temos um conjunto de dados assimétrico, ou seja, não temos uma distribuição igualitária dos diferentes tipos de classes nos nossos dados.
- Nestes casos, uma métrica mais informativa é a **mediana**.
  - Valor que separa o conjunto de dados “ao meio”.

Tendência Central - Mediana

- Ordenando nossos dados, é o valor que divide o conjunto de dados em duas metades de mesmo tamanho.
- Para um conjunto de tamanho ímpar, é o exato valor central.
- Para um conjunto de tamanho par, pode ser qualquer um dos dois valores centrais ou a média deles.
- Ex: Notas dos alunos ordenadas:
  - 0, 0, 9, **9**, 9, 9, 9.

Mediana

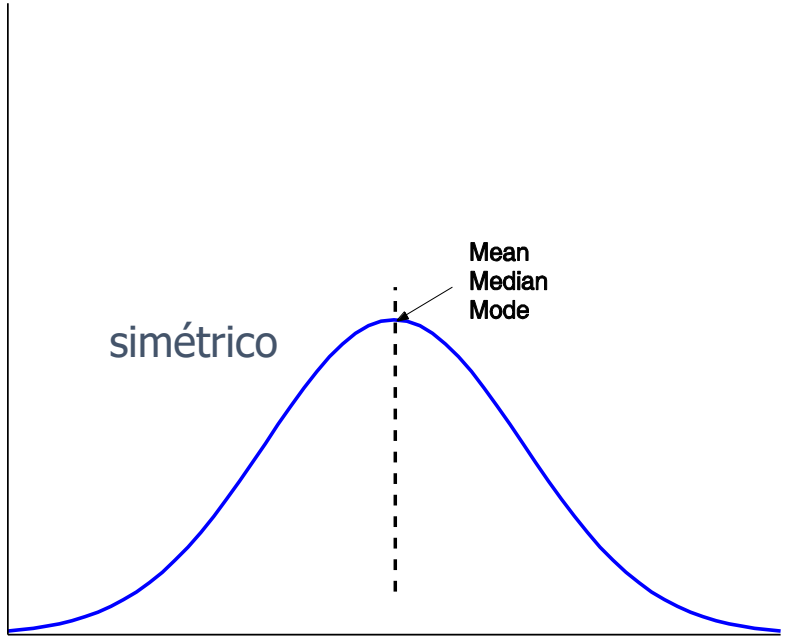
Tendência Central - Moda

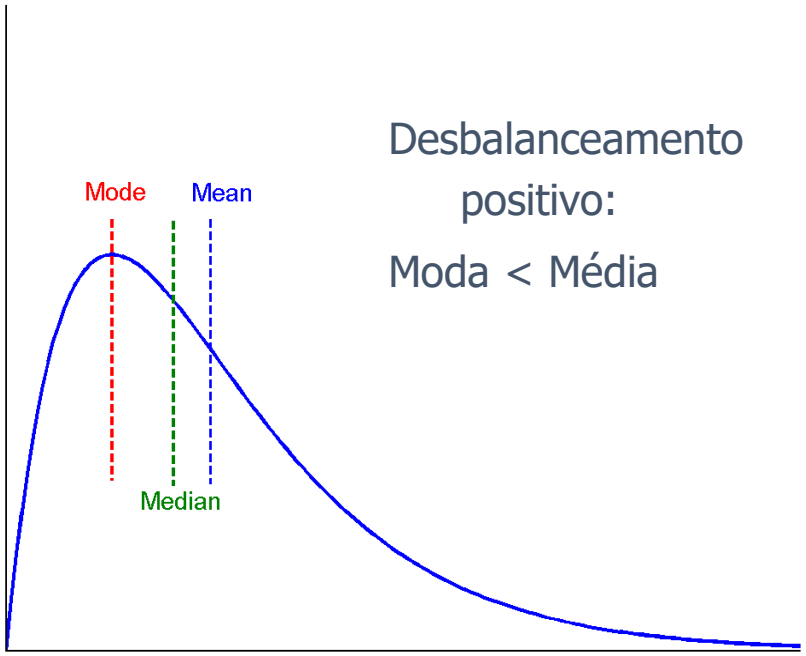
- Tanto média quanto mediana só podem ser calculadas para atributos numéricos.

- A **moda** de um conjunto é o valor que ocorre com a maior frequência em um conjunto.
  - Pode ser um atributo qualitativo ou quantitativo!
- Notas de alunos:
  - {A, A, B, B, B, C, A, B, B, C, C, C}
  - B ocorre 5 vezes, B é a moda deste conjunto.

Dados Balanceados vs Não-Balancados

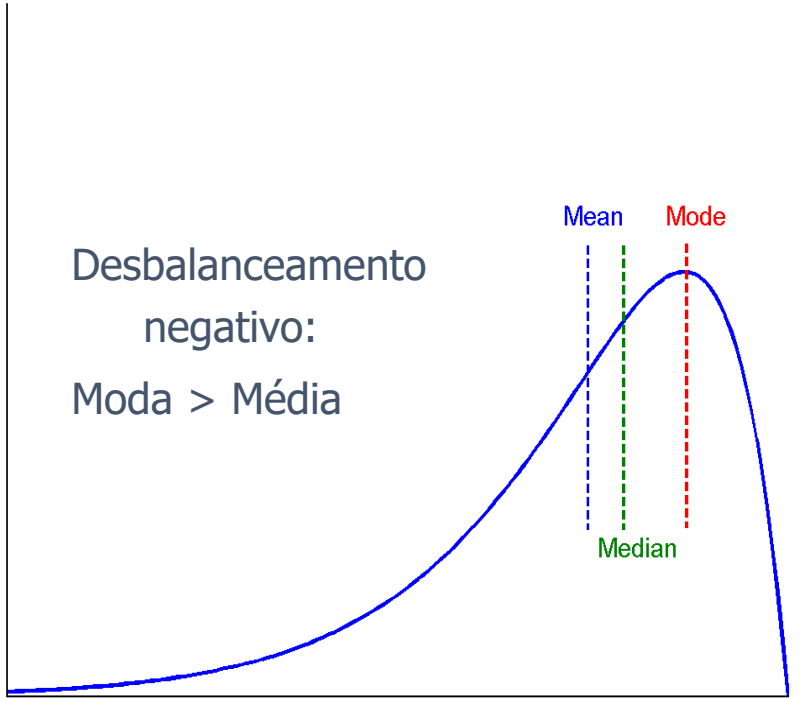
- Média, mediana e moda de dados balanceados e positivamente ou negativamente desbalanceados





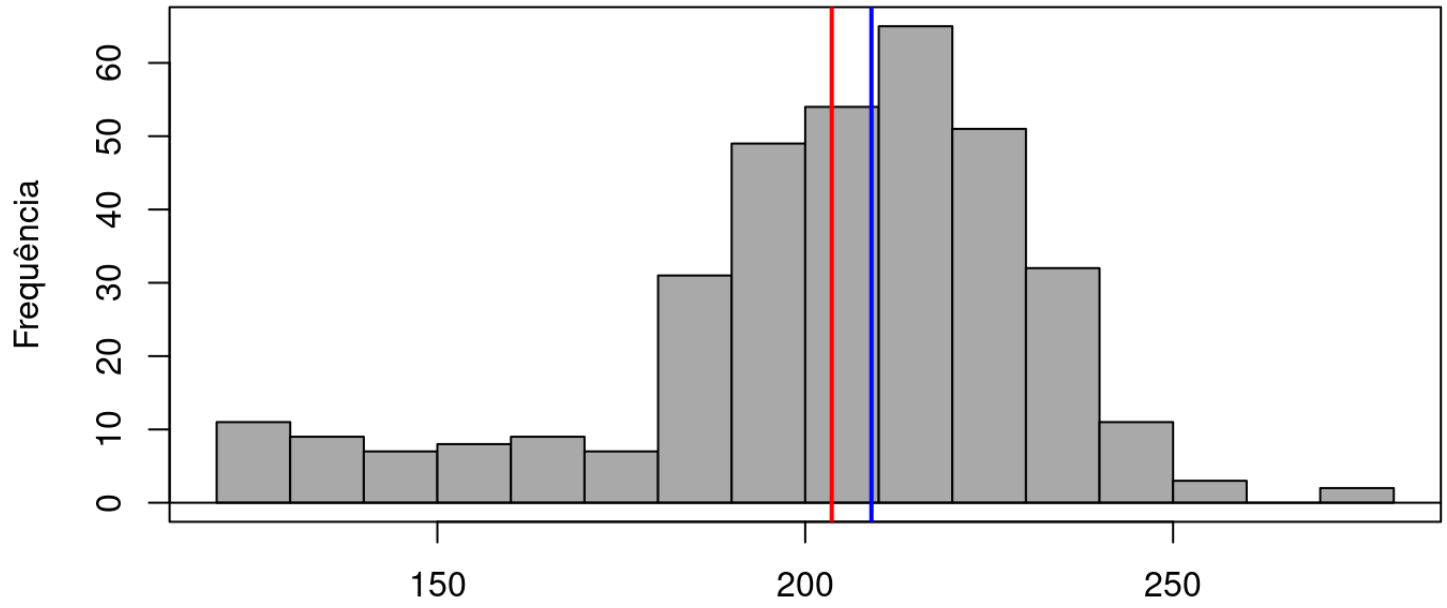
Desbalanceamento positivo:

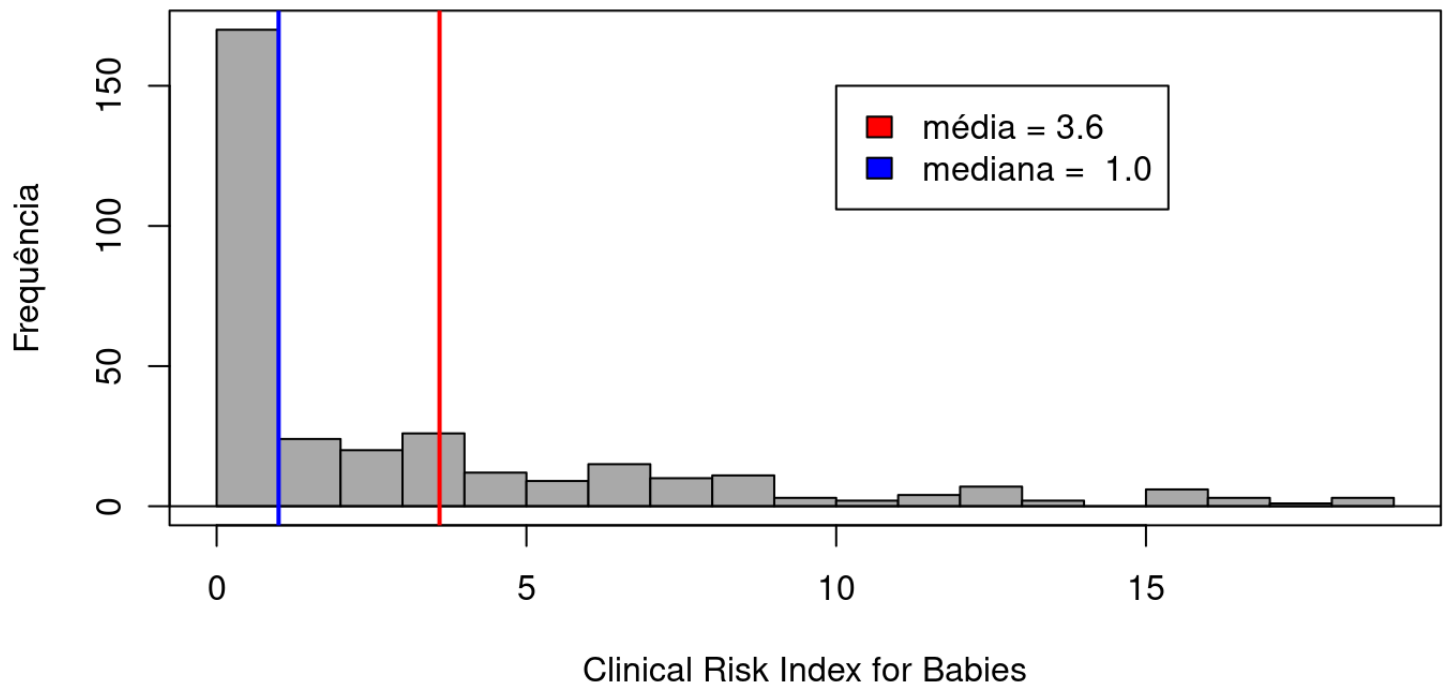
Moda < Média



Desbalanceamento negativo:

Moda > Média


Frequência
0
10
20
30
40
50
60
150
200
250
Idade Gestacional (dias)


Frequência
0
50
100
150
média = 3.6
mediana = 1.0
0
5
10
15
Clinical Risk Index for Babies

Dispersão dos dados

Dispersão

- Medidas de dispersão ajudam a compreender a variabilidade dos nossos dados.

- Um conjunto de medidas bastante útil para identificar *outliers* é o chamado **resumo dos cinco números** (five-number summary).
  - Intervalo (range)
  - Quartis
  - Intervalo interquartis

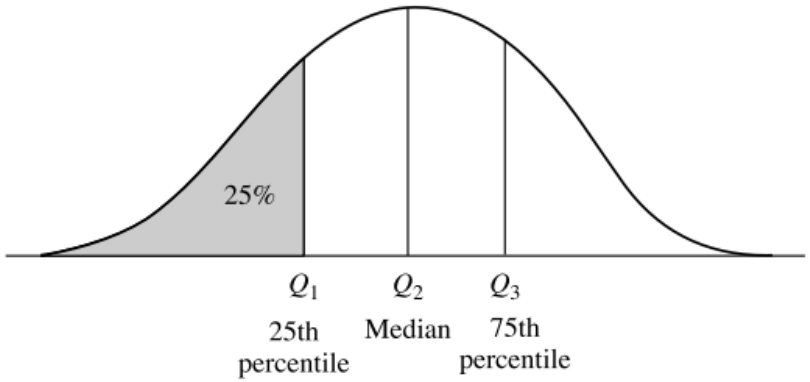
Intervalo (Range) e Quantis

- Intervalo:
  - É a diferença entre o **maior** valor e o **menor** valor de um conjunto de dados.

- Quantil:
  - Pontos entre intervalos regulares em uma distribuição de dados que dividem o conjunto em vários subconjuntos de mesmo tamanho.
  - 2-Quantil: divide o conjunto em 2 (mediana)
  - Quartil: Divide o conjunto em 4.
  - Percentil: Divide o conjunto em 100.

Quartil

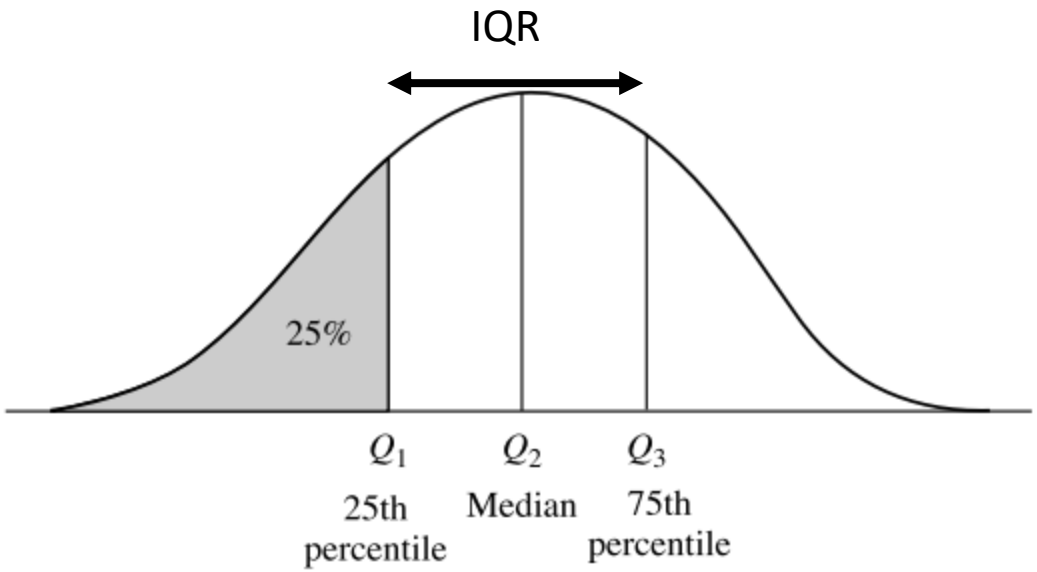
- Três valores que dividem o dataset em 4 subconjuntos de mesmo tamanho.
- Os quartis nos dão uma indicação do centro, da variabilidade e da forma de uma distribuição.
  - O primeiro quartil (Q1), corta a distribuição nos 25% menores valores, o segundo (Q2) em 50% e o terceiro (Q3) em 75%.
  - Q1 = 25-percentil. Q3 = 75-percentil...



Intervalo inter-quartis

- O intervalo inter-quartis (interquartile range – IQR) é a distância entre o terceiro quartil (Q3) e o primeiro quartil (Q1).

IQR = Q3 – Q1



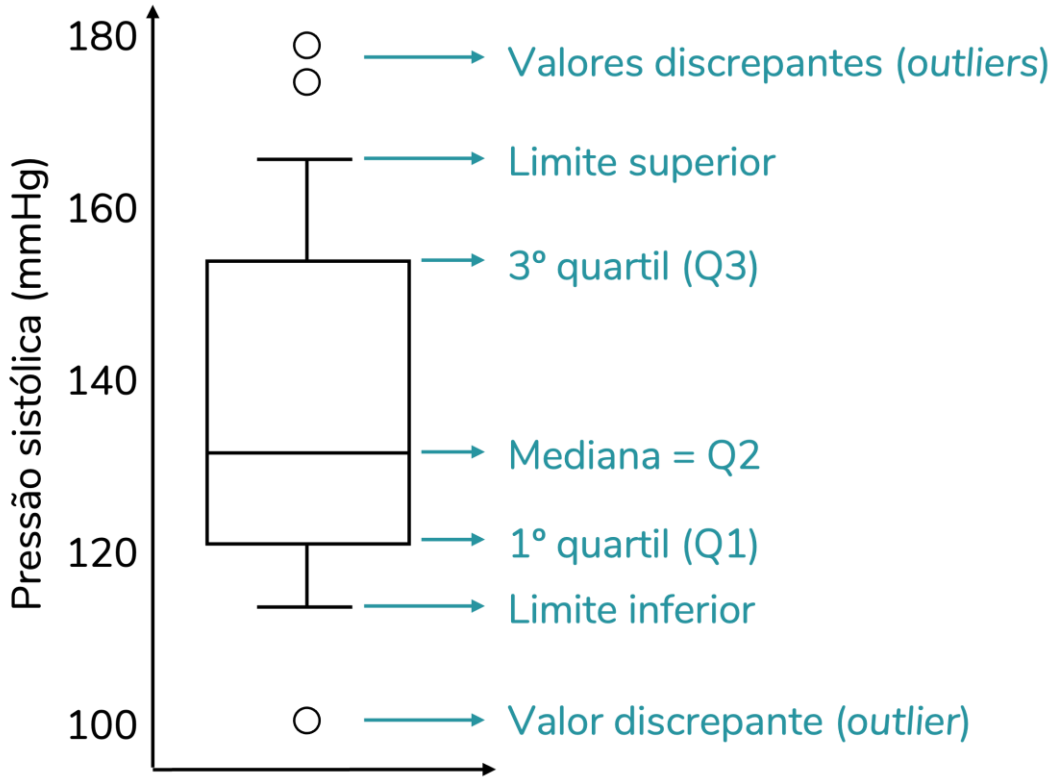
Resumo dos cinco números

- Analisar estas medidas por si só é pouco informativas.
  - Analisando só o Q2 (a mediana), não sabemos se temos dados muito variados.
  - Analisando apenas o intervalo, não sabemos se os dados são balanceados ou não.
  - Analisando apenas os quartis (Q1, Q2 e Q3), não conhecemos as “pontas” da nossa distribuição.

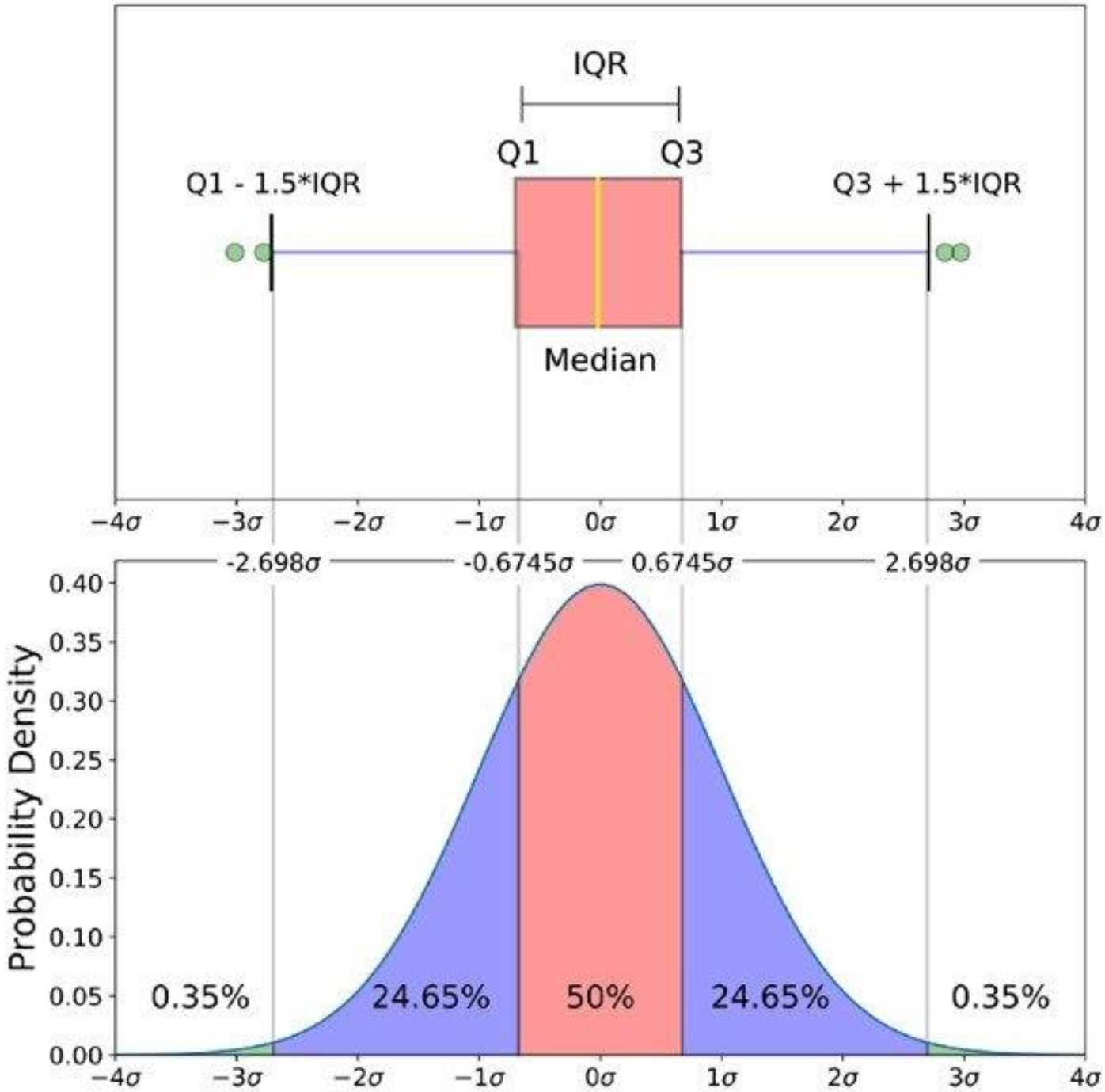
- O mais comum é analisarmos todos estes número em conjunto.
  - Utilizamos box plots para isso!

Boxplot

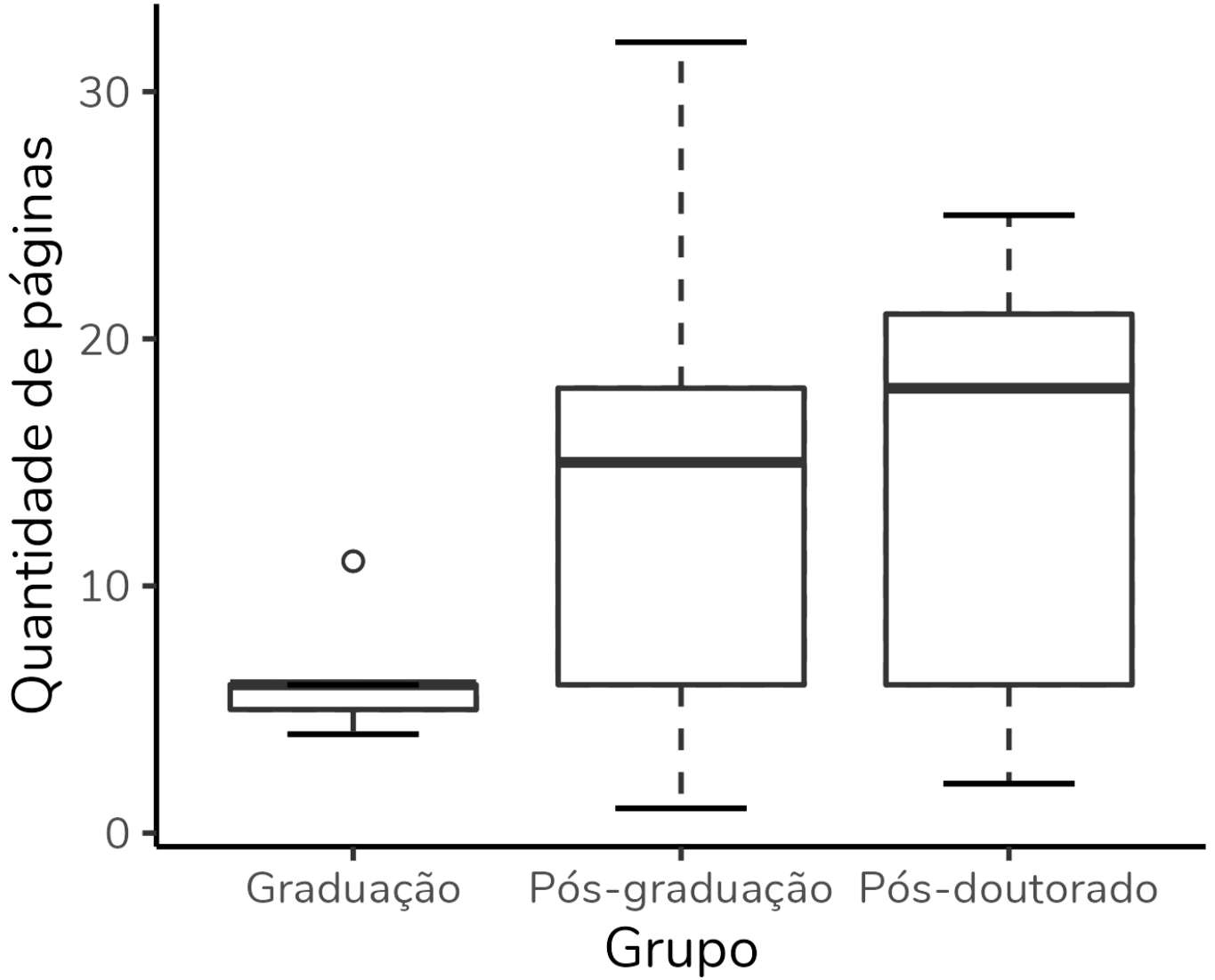
- Forma gráfica de representar uma distribuição.
- As extremidades do retângulo são os quartis Q1 e Q2.
- A linha dentro do retângulo é a mediana (Q2).
- As duas linhas fora do retângulo (whiskers) são os valores mínimo e máximo.



Boxplot e a Distribuição Normal



Boxplot


Quantidade de páginas
30
20
10
0
Graduação
Pós-graduação
Pós-doutorado
Grupo

Variância e Desvio Padrão

- Também são medidas de dispersão, porém em torno da média, e não da mediana.
- Seja μ a média aritmética de um conjunto

Variância σ^2: expectativa para uma variável aleatória do desvio quadrado da média

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

Desvio padrão σ :expectativa do desvio de uma variável aleatória

$$\sigma = \sqrt{\sigma^2}$$

O que podemos medir em atributos?

- Momento

$$\text{momento}_k(x_j) = \frac{\sum_{i=1}^{N}(x_j^{(i)} - \bar{x}_j)^k}{(N-1)}$$

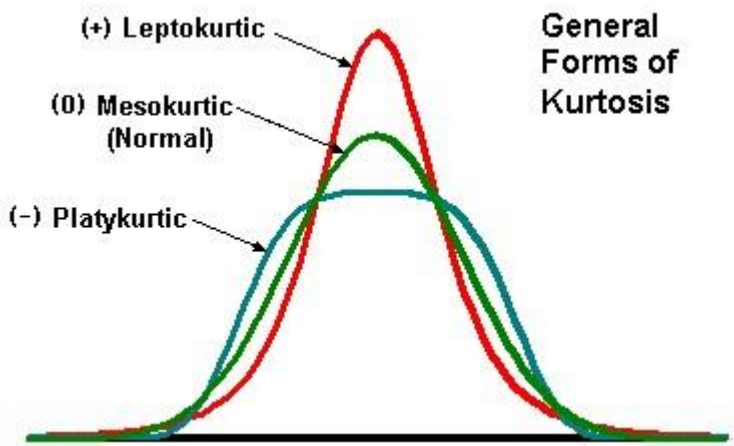
O que podemos medir em atributos?

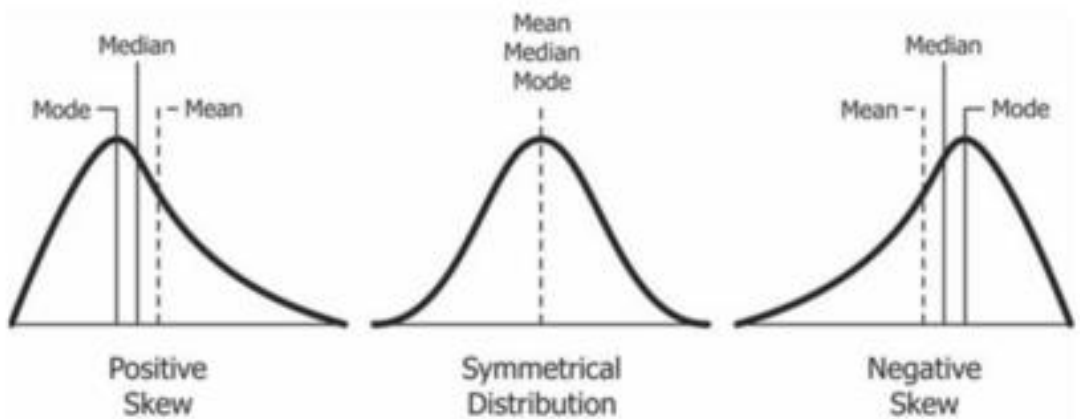
$$\text{momento}_k(x_j) = \frac{\sum_{i=1}^{N}(x_j^{(i)} - \bar{x}_j)^k}{(N-1)}$$

quando $k = 2$, tem-se a variância, que é o segundo momento central;

quando $k = 3$, tem-se a obliquidade, que é o terceiro momento central;

quando $k = 4$, tem-se a curtose, que é o quarto momento central.





Análise Multivariada de dados

Medidas de mais de um atributo

- Covariância

$$\operatorname{cov}(\mathbf{x}_j, \mathbf{x}_k) = \frac{1}{N-1} \sum_{i=1}^{N} (x_j^{(i)} - \bar{x}_j)(x_k^{(i)} - \bar{x}_k)$$

Covariância

- Se Cov(A,B) > 0, dizemos que temos uma **covariância positiva** e ambas variáveis se movem juntas.
- Se Cov(A,B) < 0, dizemos que temos uma **covariância negativa** e ambas variáveis se movem em posições opostas.
- Duas variáveis **independentes** entre si possuem Cov(A,B) = 0

- Variância é um caso especial da covariância onde temos Cov(A,A).

Exemplo covariância

- Suponha que duas ações (X_1,X_2) tenham os seguintes valores em uma semana:
  - (2, 5), (3, 8), (5, 10), (4, 11), (6, 14)
- Se as ações forem afetadas pelas mesmas tendências, os preços subirão ou cairão juntos?

$$\overline{X_1} = \frac{20}{5} = 4 \quad \overline{X_2} = \frac{48}{5} = 9{,}6 \qquad \operatorname{cov}(\mathbf{x}_j, \mathbf{x}_k) = \frac{1}{N-1}\sum_{i=1}^{N}(x_j^{(i)} - \bar{x}_j)(x_k^{(i)} - \bar{x}_k)$$

$$cov(X_1, X_2)$$
$$= \frac{1}{5-1}[(2-4)(5-9{,}6) + (3-4)(8-9{,}6) + (5-4)(10-9{,}6)$$
$$+ (4-4)(11-9{,}6) + (6-4)(14-9{,}6)]$$
$$= \frac{1}{4}[9{,}2 + 1{,}6 + 0{,}4 + 0 + 8{,}8] = 5$$

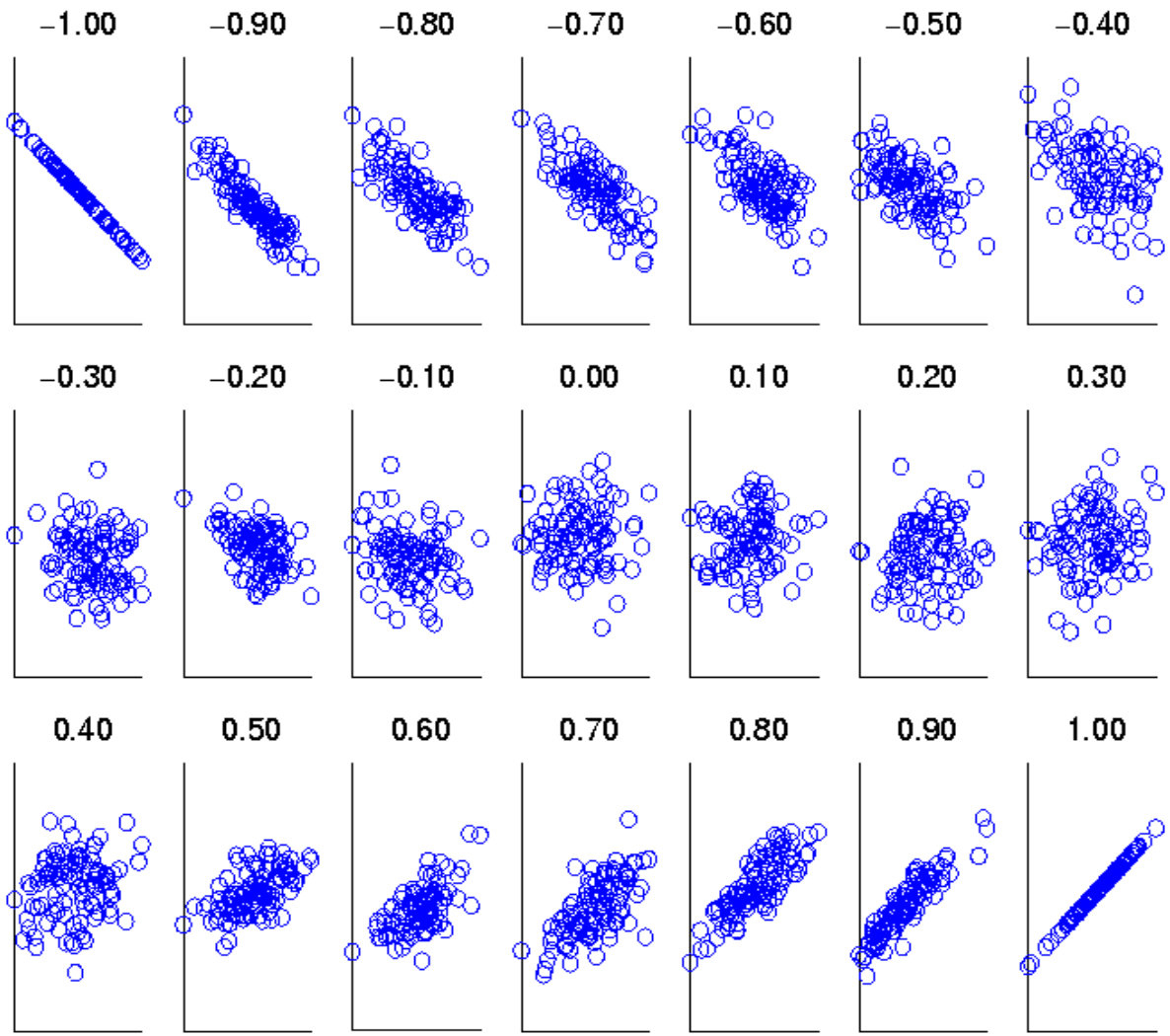
- Sendo assim, X_1 e X_2 subirão ou cairão juntas porque $Cov_{1,2} > 0$ (positivo)

Medidas de mais de um atributo

- Correlação

$$\text{corr}(\mathbf{x}_j, \mathbf{x}_k) = \frac{\text{cov}(\mathbf{x}_j, \mathbf{x}_k)}{\sigma_{\mathbf{x}_j} \sigma_{\mathbf{x}_k}}$$

Visualizando diferentes coeficientes de correlação



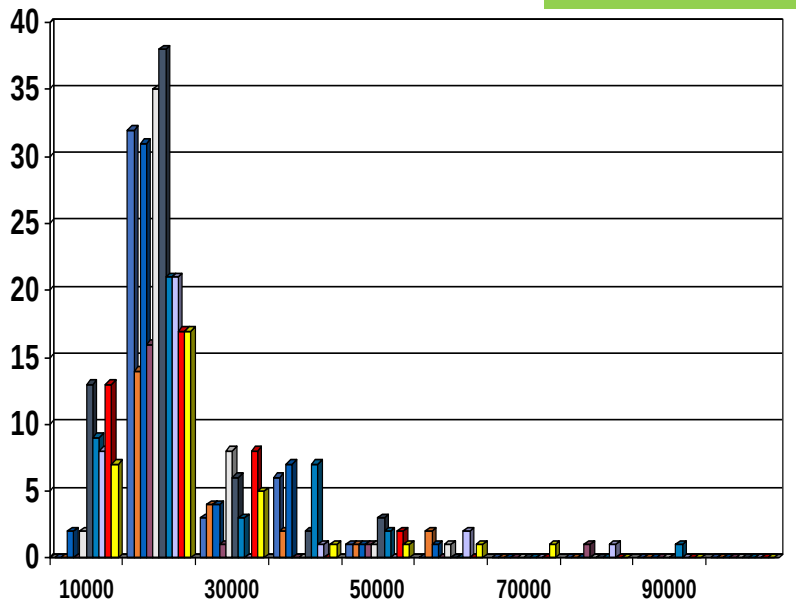
- Intervalo do coeficiente: [–1, 1]
- Vários scatter plots mostrando conjunto de pontos e o coeficiente de correlação entre eles mudando de –1 até 1

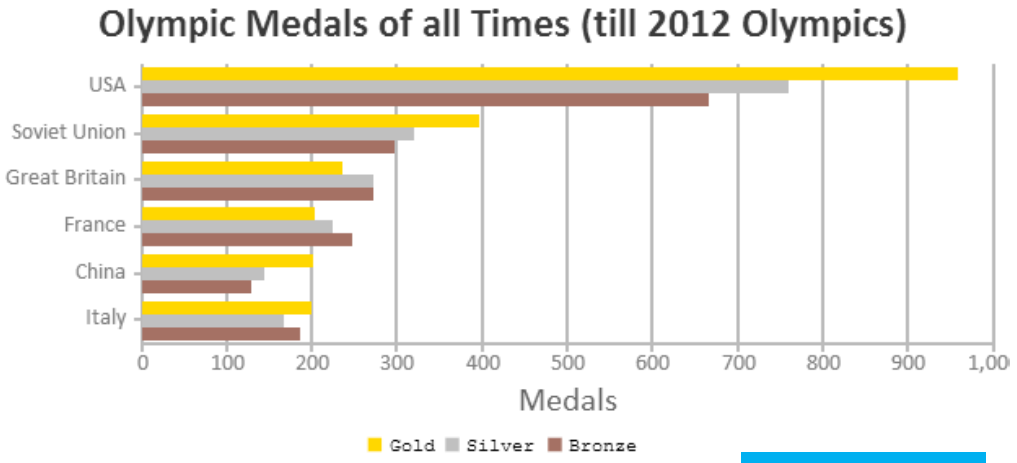
Outras visualizações

Histogramas

- Visualização gráfica em barras de um conjunto de dados tabulado e dividido em classes.
  - Dados são divididos em subconjuntos consecutivos (bins), cada barra representa a frequência da ocorrência em um subconjunto.
- Histogramas são **diferentes** de gráficos de barra (bar chart)!
  - Histogramas são quantitativos, gráficos de barra qualitativos.
  - A ordem das barras é relevante no histograma, no outro não é.

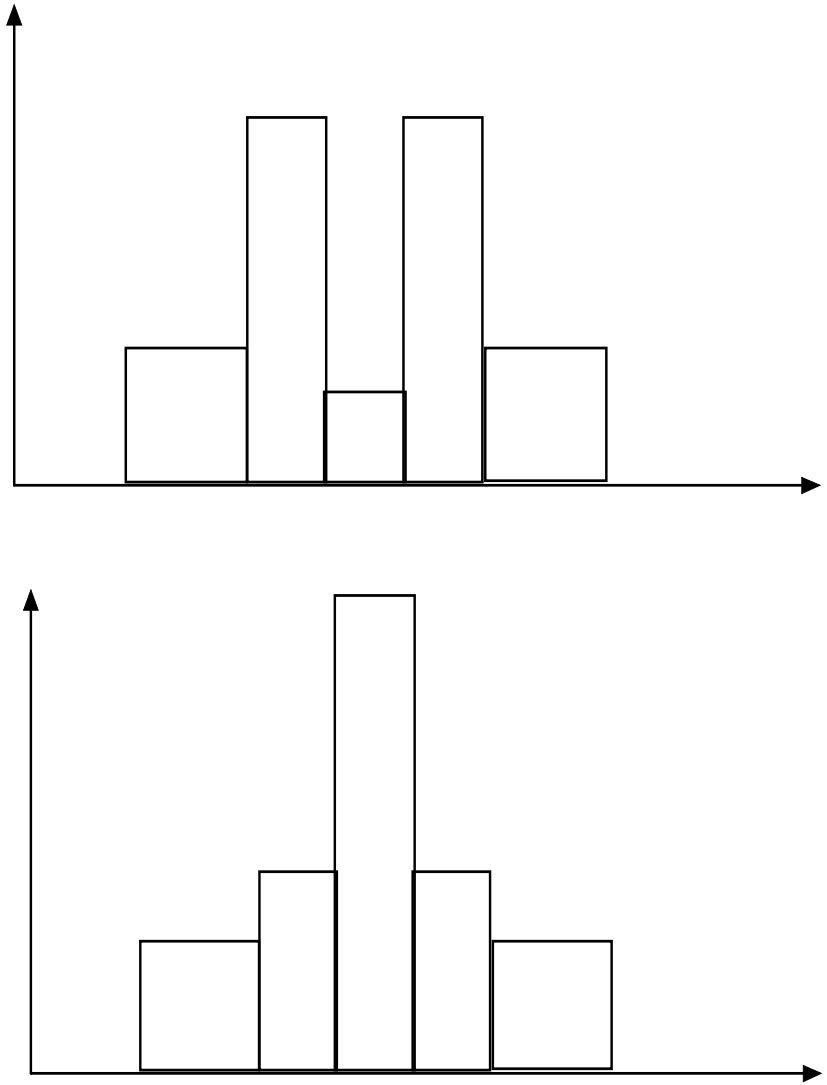
Histogram





Bar chart

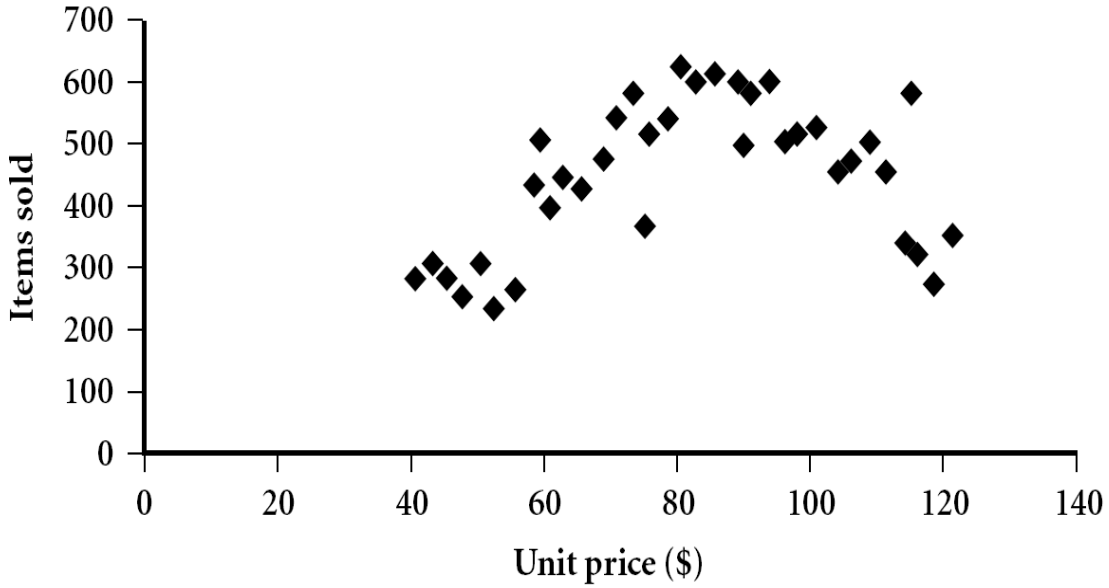
Histogramas e Boxplots

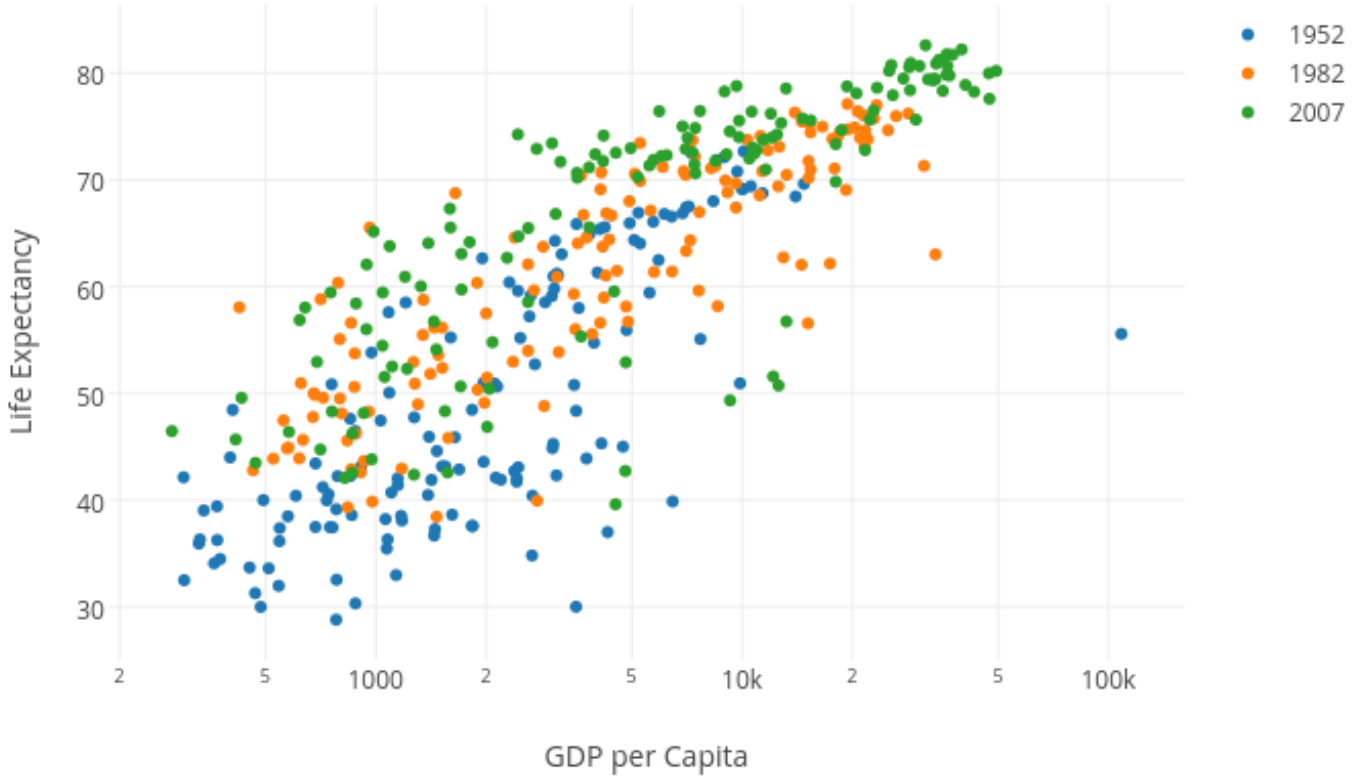


- Em geral histogramas são mais informativos que boxplots.
- Os dois histogramas na esquerda podem ter o mesmo boxplot.
  - Mesmo valores para min, Q1, mediana, Q3 e max.
- São distribuições bem diferentes!

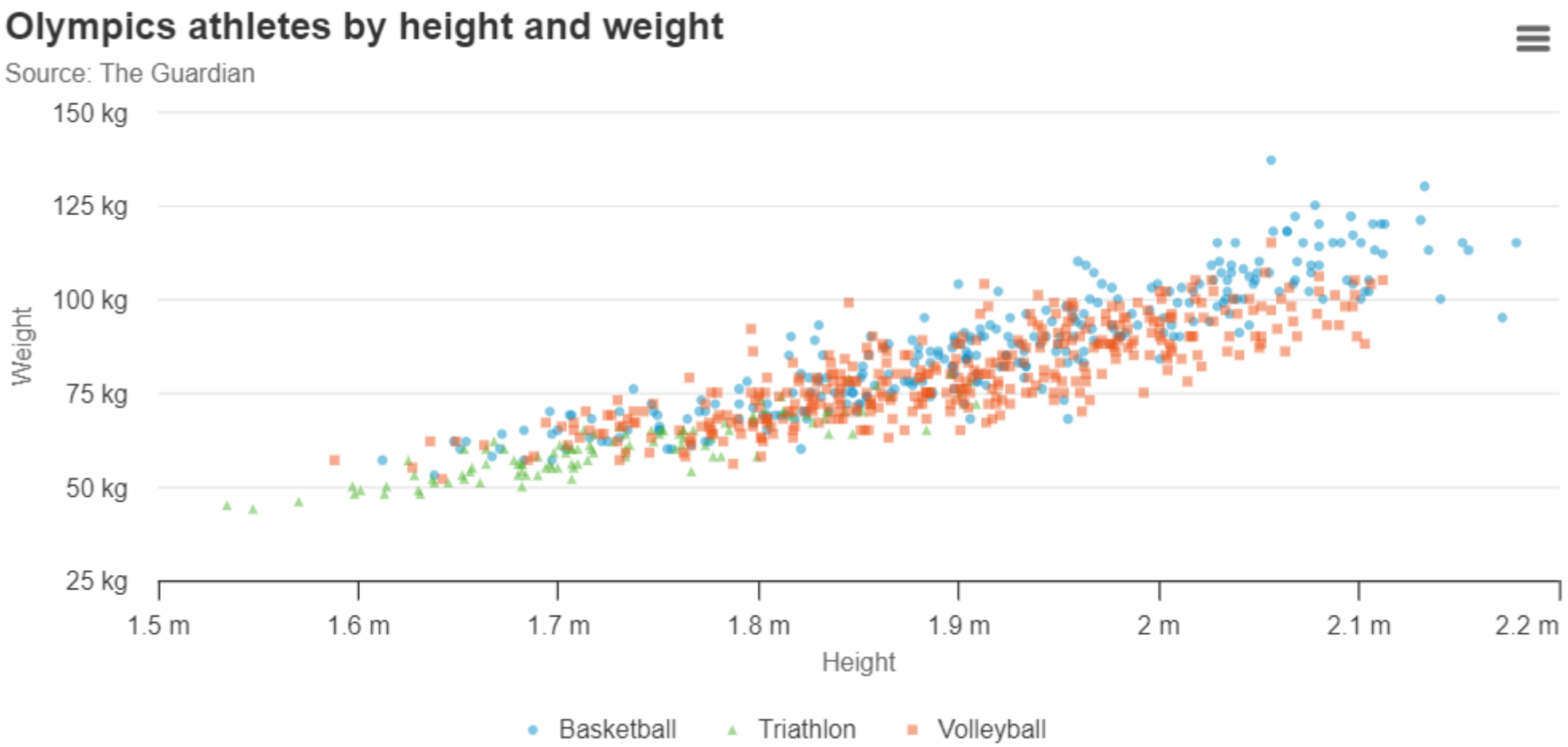
Scatter plot (Gráfico de dispersão)

- Primeira visualização em dados bivariados para olhar clusters, outliers, etc.
- Cada par de valores é tratado como um par de coordenadas e plotado como pontos no plano.

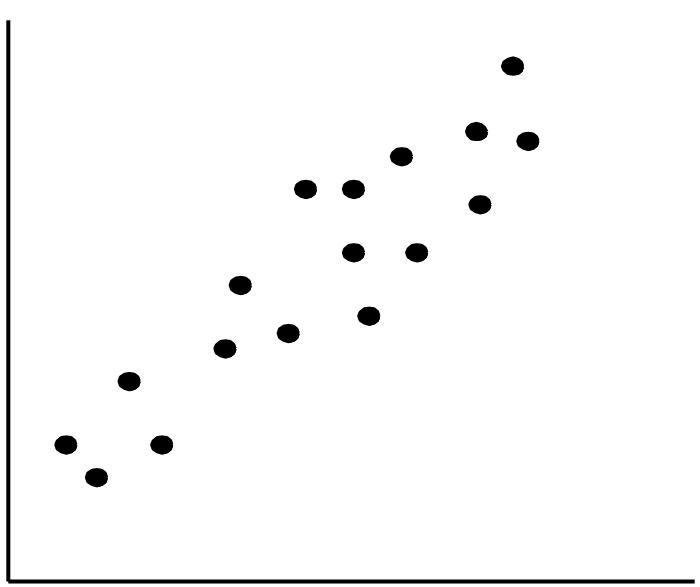


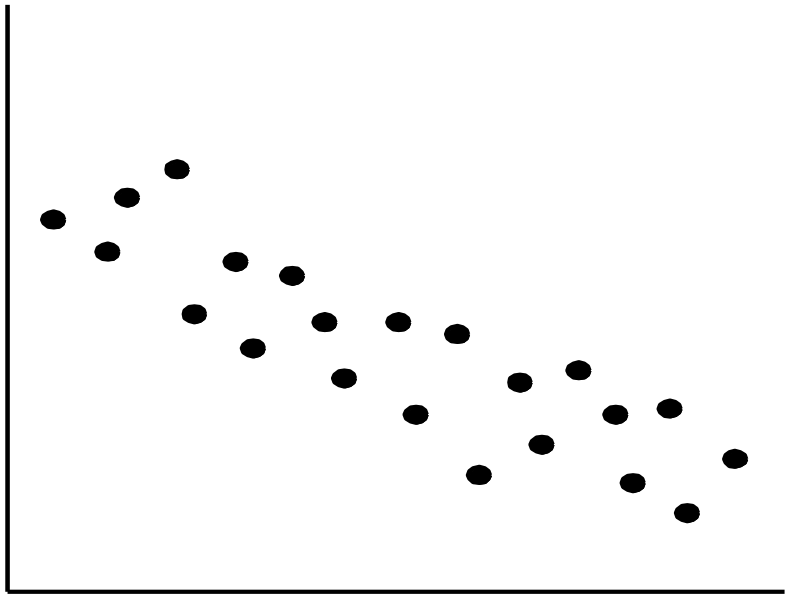


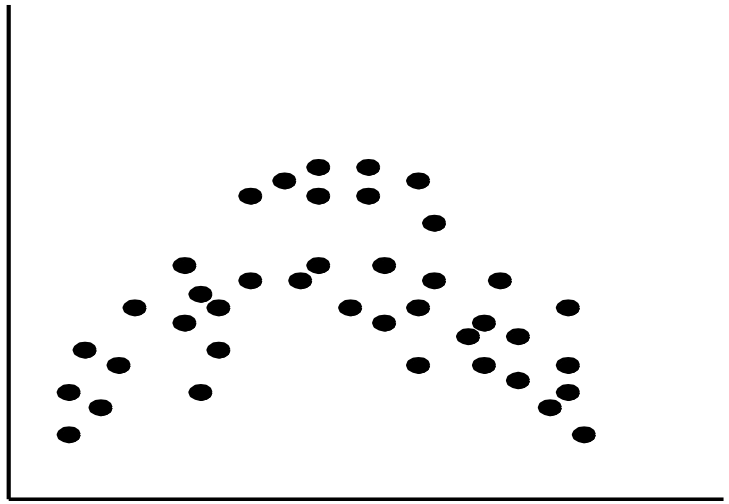
Visualizando correlações: Gráfico de Scatter Plot

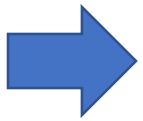


Correlações Positivas e Negativas



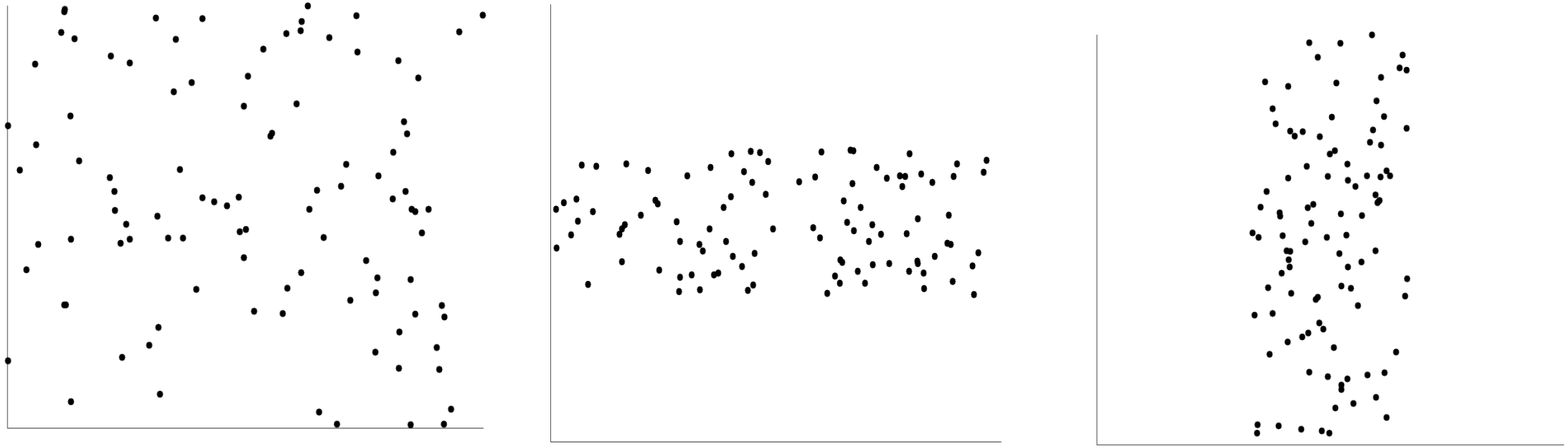






- Metade esquerda correlacionada positivamente
- Metade direita correlacionada negativamente

Dados não correlacionados



Medidas de Similaridade e Distância

Medidas de Proximidade

- Matriz de Dados vs Matriz de Dissimilaridade
- Medidas de proximidade para atributos:
  - Numéricos
  - Binários Simétricos e Assimétricos
  - Nominais
  - Ordinais
- Medidas de proximidade para objetos com atributos mistos.
- Outras medidas de similaridade
  - Cosseno
  - Informação Mútua

Similaridade, Dissimilaridade e Proximidade

- **Medida ou função de similaridade**
  - Função que quantifica a similaridade entre dois objetos
  - Mede o quanto dois objetos são parecidos: quanto maior o valor, mais similares
  - Geralmente no intervalo [0,1]: 0: não similar; 1: completamente similar
- **Medida ou função de dissimilaridade** (ou **distância**)
  - Medida do quão diferente são dois objetos de dados
  - "Inverso" da similaridade: Quanto maior o valor, mais distantes dois objetos
  - Intervalo [0, 1] ou [0, ∞) , dependendo da medida
- **Proximidade** se a qualquer um dos dois tipos de medidas
- Em geral é mais comum calcular a distância e considerar que

$$sim(a,b) = 1 - d(a,b)$$

Matriz de Dados e Matriz de Dissimilaridade

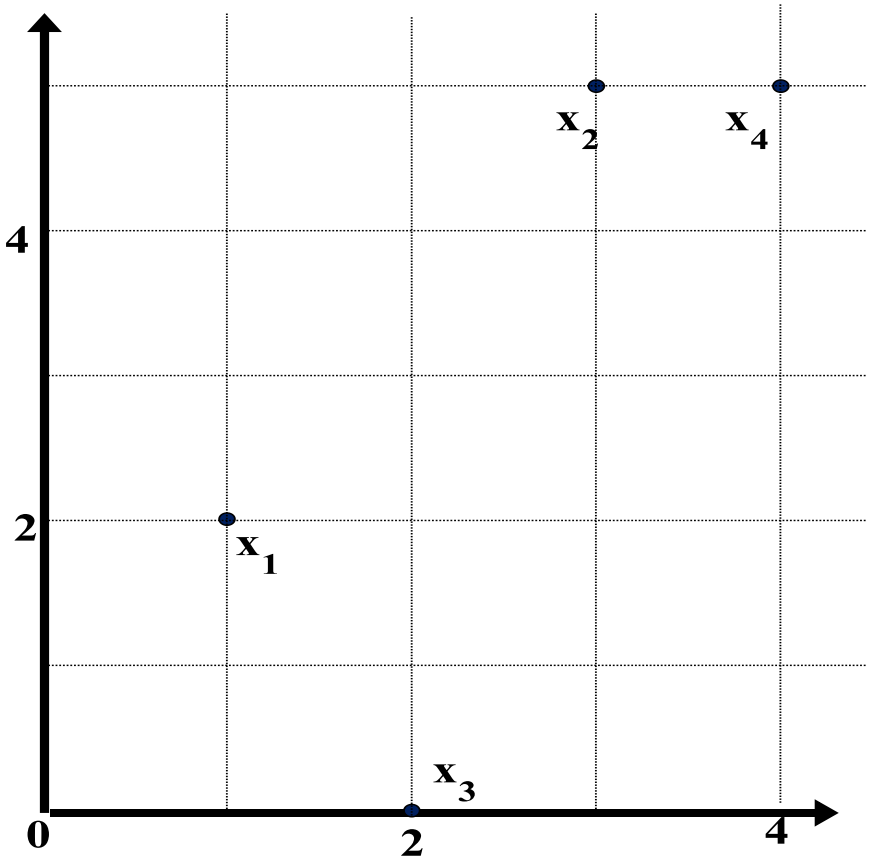
- Matriz de dados
  - Uma matriz de *n* objetos com *l* dimensões

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1l} \\ x_{21} & x_{22} & \cdots & x_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nl} \end{pmatrix}$$

- Matriz de dissimilaridade (ou distância)
  - n objetos, mas só registra distância d(i,j) (por que distância é tipicamente simétrica)
  - Funções de distância são diferentes para atributos de tipos diferentes (nominais, ordinais, binários, numéricos etc.)

$$\begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \cdots & 0 \end{pmatrix}$$

Exemplo



Matriz de Dados

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Matriz de distância (Euclidiana)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Atributos numéricos: Minkowski Distance

- Minkowski distance: Medida popular de distância

$$d(i, j) = \sqrt[p]{| x_{i1} - x_{j1} |^p + | x_{i2} - x_{j2} |^p + \cdots + | x_{il} - x_{jl} |^p}$$

onde $i = (x_{i1}, x_{i2}, \ldots, x_{il})$ e $j = (x_{j1}, x_{j2}, \ldots, x_{jl})$ são dois objetos com *l dimensões*, e *p* é o grau (também chamada de L-*p* norm)

- Propriedades interessantes:
  - d(i, j) > 0 se i ≠ j, e d(i, i) = 0 (Sempre positiva)
  - d(i, j) = d(j, i) (Simétrica)
  - d(i, j) ≤ d(i, k) + d(k, j) (Desigualdade triangular)
- Qualquer função de distância que respeite essas propriedades é uma **métrica**

Casos especiais da distância de Minkowski

- p = 1: (L_1 norm) Distância de Manhattan (ou city block)

$$d(i, j) = | x_{i1} - x_{j1} | + | x_{i2} - x_{j2} | + \cdots + | x_{il} - x_{jl} |$$

- p = 2: (L_2 norm) Distância Euclidiana

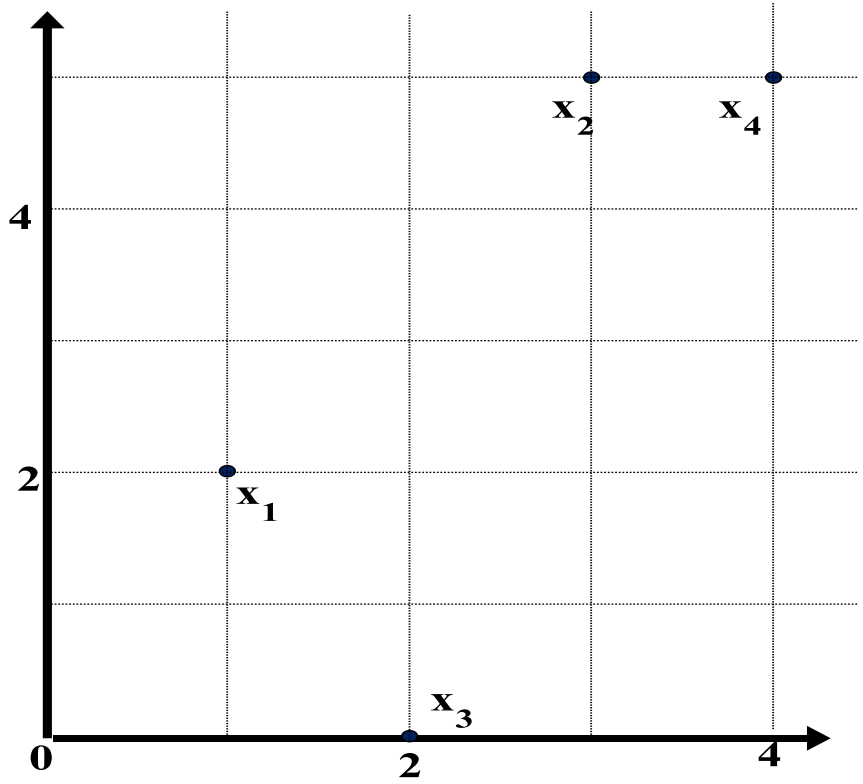
$$d(i, j) = \sqrt{| x_{i1} - x_{j1} |^2 + | x_{i2} - x_{j2} |^2 + \cdots + | x_{il} - x_{jl} |^2}$$

- $p \to \infty$: (L_{max} norm, L_∞ norm) Distância de Chebyshev
  - A diferença máxima entre todos os pares de atributos

$$d(i, j) = \lim_{p \to \infty} \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p} = \max_{f=1}^{l} |x_{if} - x_{jf}|$$

Exemplo:

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidiana (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Chebyshev (L_∞)

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Medidas para atributos binários

- Matriz de contigência para atributos binários

Objeto *i* \ Objeto *j*	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

- Distância para variáveis binárias simétricas:

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

- Distância para variáveis binárias assimétricas:

$$d(i, j) = \frac{r + s}{q + r + s}$$

- Coeficiente de Jaccard (medida de similaridade para variáveis assimétricas)

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

Exemplo: Dissimilaridade entre variáveis assimétricas

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gênero não é considerado (é simétrico)
- Todos outros atributos são assimétricos
- Considerando Y e P = 1, e N = 0
- Distância: $d(i, j) = \frac{r + s}{q + r + s}$

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Mary

Jack	1	0	$\sum_{row}$
1	2	0	2
0	1	3	4
$\sum_{col}$	3	3	6

Jim

Jack	1	0	$\sum_{row}$
1	1	1	2
0	1	3	4
$\sum_{col}$	2	4	6

Mary

Jim	1	0	$\sum_{row}$
1	1	1	2
0	2	2	4
$\sum_{col}$	3	3	6

Medidas para atributos nomimais

- Example: Cor (red, yellow, blue, green), profissão, etc.
- **Método 1**: Simple matching $$d(i, j) = \frac{p - m}{p}$$
  - *m*: # de matches, *p*: # total de variáveis
- **Método 2**: Converter atributo nominal para vários atributos binários e calcular como distância binária
  - Cor_vermelha: 0 ou 1
  - Cor_amarela: 0 ou 1
  - Cor_azul: 0 ou 1
  - Cor_verde: 0 ou 1

Exemplo para nominal

$$d(i, j) = \frac{p - m}{p}$$

m: # de matches, *p*: # total de variáveis

p = 1 (variável teste 1)

ID	Teste 1 (nominal)	Teste 2 (ordinal)	Teste 3 (numérico)
1	Código A	Excelente	45
2	Código B	Razoável	22
3	Código C	Bom	64
4	Código A	Excelente	28

Matriz de Dissimilaridade =

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Variáveis ordinais

- Ordenação é importante.
  - Podemos pensar na ordem como um ranqueamento (pequeno < médio < grande)
- Abordagem: substituir os valores de um atributo ordinal pelo valor do seu rank.
  - Ordenar os valores e atribuir um número para cada valor.
    - Ex: {pequeno, médio, grande} -> {1, 2, 3}
  - Calcular distância como se fosse atributo numérico.

$$r_{if} \in \{1, \ldots, M_f\}$$

- **Problema**:
  - Quando temos vários atributos nominais, aquele que tiver mais rankings vai influenciar mais na medida.
- **Solução**: mapear todos atributos para um intervalo de [0, 1] e substituir pelos valores normalizados

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Exemplo para ordinal

ID	Teste 1 (nominal)	Teste 2 (ordinal)	Teste 3 (numérico)
1	Código A	Excelente	45
2	Código B	Razoável	22
3	Código C	Bom	64
4	Código A	Excelente	28

- Ordenar valores de *f*:
  - {Razoável, Bom, Excelente}
- Atribuir um ranking $r_{if} \in \{1, \ldots, M_f\}$
  - {1, 2, 3}
- Normalizar os rankings em [0,1] $z_{if} = \frac{r_{if} - 1}{M_f - 1}$
  - {0,0.5,1}
- Substituir x_{if} pelo valor normalizado
- Calcular distância numérica
  - Distância Euclidiana:

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Teste 2 (ordinal)
1
0
0.5
1

Pares (1,2) e (4,2)
são os mais dissimilares

Proximidade com atributos mistos

- Um mesmo dataset pode ter **tipos diferentes** de atributos
  - Nominal, binário (ass)simétrico, numérico e ordinal
- Como calcular a distância entre objetos com atributos diferentes?
- **Abordagem simples**: **agrupar** atributos do mesmo tipo e calcular proximidade entre eles.
  - Problema é que precisamos que as diferentes análises tenham resultados compatíveis, o que é pouco provável em dados reais.
- **Outra abordagem**: calculo único, utilizando uma média **ponderada** das distâncias específicas de cada tipo de atributo.

Proximidade com atributos mistos

- Média ponderada onde o valor do peso *w* depende do tipo de atributo *f* dos *p* atributos existentes:

$$d(i,j) = \frac{\sum_{f=1}^{p} w_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} w_{ij}^{(f)}}$$

- $w^{(f)} = 0$ se
  - x_{if} ou x_{jf} estiverem faltando os valores (campo vazio)
  - $x_{if} = x_{jf} = 0$ e *f* for binário ou nominal.
- $w^{(f)} = 1$ para todos outros casos
  - Se *f* for ordinal, ordenar e normalizar os rankings e tratar como numérico
  - Se *f* for numérica, calcular a distância normalizada $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_f - min_f}$

Exemplo misto

- Já calculamos para Testes 1 e 2, falta a distância normalizada para Teste 3.

ID	Teste 1 (nominal)	Teste 2 (ordinal)	Teste 3 (numérico)
1	Código A	Excelente	45
2	Código B	Razoável	22
3	Código C	Bom	64
4	Código A	Excelente	28

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_f - min_f}$$

$$d_{ij}^{(teste\ 3)} = \begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

$$d_{ij}^{(teste\ 1)} = d_{ij}^{(teste\ 3)} =$$

$$d_{ij}^{(teste\ 2)} = \begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

$$d(i,j) = \frac{\sum_{f=1}^{p} w_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} w_{ij}^{(f)}}$$

Teste 1 Teste 2 Teste 3

$$d(3,1) = \frac{1(1) + 1(0.50) + 1(0.45)}{3} = 0.65$$

$$d_{ij} = \begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}$$

Outras medidas

- Existem diversas outras medidas, mas as que vimos já funcionam para grande parte dos casos.

- Uma medida que vale a pena conhecer é a **similaridade do cosseno**.
  - Tipicamente é utilizada em tarefas de PLN e bioinformática, onde dados são esparsos.
  - Calcula a similaridade entre dois vetores.

$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$

  - Onde *x* e *y* são vetores em um produto vetorial e ||*x*|| e ||y|| são a norma euclidiana de *x* e *y*.

Exemplo

- Documentos representados como *bag of terms:*

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Similaridade entre documentos 1 e 2.

d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$d_1 \bullet d_2$ = 5 X 3 + 0 X 0 + 3 X 2 + 0 X 0 + 2 X 1 + 0 X 1 + 0 X 1 + 2 X 1 + 0 X 0 + 0 X 1 = 25

$$\| d_1 \| = \sqrt{5\times5+0\times0+3\times3+0\times0+2\times2+0\times0+0\times0+2\times2+0\times0+0\times0} = 6.481$$

$$\| d_2 \| = \sqrt{3\times3+0\times0+2\times2+0\times0+1\times1+1\times1+0\times0+1\times1+0\times0+1\times1} = 4.12$$

cos(d_1, d_2) = 25/ (6.481 X 4.12) = 0.94

Informação Mútua

- Medida de dependência mútua entre duas variáveis.
  - Quantifica a “quantidade de informação” que podemos obter sobre **uma** variável aleatório observando **outra** variável aleatória.

- Relacionada com o conceito de entropia (quantidade de informação observada em **uma** variável aleatória)

- Entropia da variável *V* com *k* valores: $\mathrm{H}(V) = -\sum_k P(v_k) \log_2 P(v_k)$

- Informação Mutua = $I(A, B) = H(A) + H(B) - H(A, B)$

Informação Mútua

- Entropia da variável *V* com *k* valores:

$$\mathrm{H}(V) = -\sum_{k} P(v_k) \log_2 P(v_k)$$

$$\mathrm{H}(A, B) = -\sum_{k} \sum_{i} P(A_k, B_i) \log_2 P(A_k, B_i)$$

$$\mathrm{P}(k, i) = P(k) \times P(i)$$

- Informação Mútua = $I(A, B) = H(A) + H(B) - H(A, B)$

Resumo

- Tipos de atributos (ou features, ou dimensões)
  - Nominal, binário, ordinal, numérico, discreto vs. contínuo
- Estatísticas dos dados
  - Tendência central, dispersão, covariância e correlação, gráficos
- Similaridade/Dissimilaridade
  - Medidas para cada tipo de atributo e todos misturados
  - Cosseno para vetores esparsos
  - Informação mútua