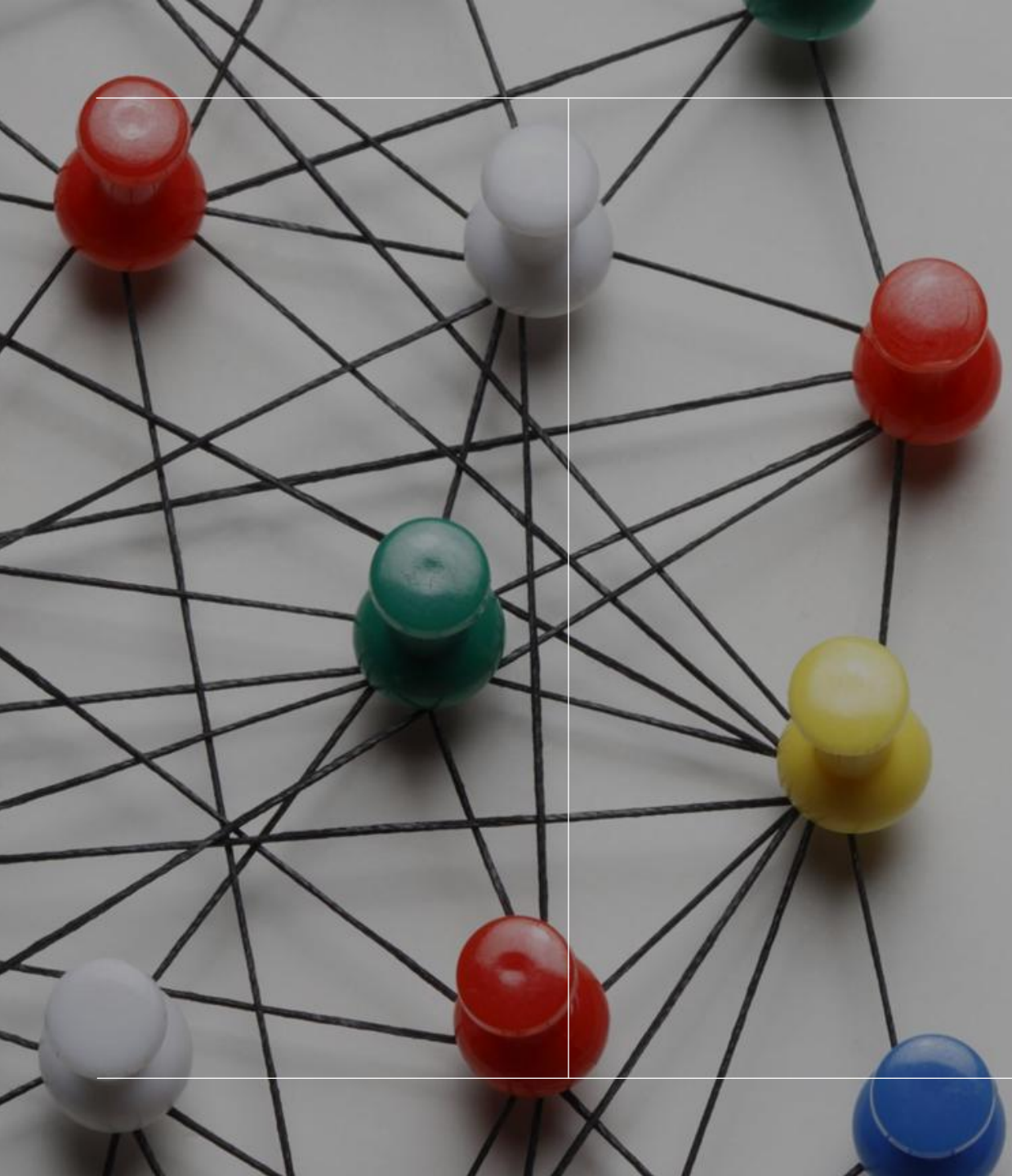
Aula 3
Análise Exploratória de Dados

PROF. ME. NATAN KLEIN
IA & ANÁLISE DE DADOS
CIÊNCIA DA COMPUTAÇÃO
ATITUS EDUCAÇÃO

Conteúdos

- Tipos de Dados
  - Modelagem
  - Definição
  - Categóricos
  - Numéricos
  - Escala de mensuração
- Localidade
  - Média
  - Mediana
  - Moda
- Análises de frequência
  - Gráficos
  - Tabelas de frequência
- Espalhamento
  - Intervalo
  - Desvios
  - Soma dos quadrados
  - Variância
  - Desvio-Padrão
- Distribuições
  - Métricas
  - Histogramas
- Partição
  - Quartis
  - Percentis
  - Box-plots



Tipos de dados

Tipos de dados: modelagem

- O mundo real é formado por objetos que diferem entre si nas suas características – *i.e.*, apresentam diferenças individuais.

- Essas características existem enquanto categorias ou quantidades que variam nos valores que assumem de objeto para objeto – por isso as chamamos de variáveis.

- Supomos que existem padrões por trás dessa variabilidade e que conhecê-los pode nos ajudar a explicar e prever o mundo. Para conhecer os padrões, precisamos de vários exemplos de objetos onde mensuramos as variáveis.

Tipos de dados: modelagem

- **Por exemplo:** fotos postadas nos stories no Instagram e número de likes recebidos são variáveis que assumem diferentes valores entre os usuários das redes sociais.

- Podemos identificar padrões nos dados construindo modelos: representações simplificadas da realidade. *e.g. "ponte".*

- Toda teoria é um modelo (*e.g.* ciclo vital conjugal), alguns modelos são matemáticos (*e.g.* juros compostos) e os modelos com que vamos trabalhar são empíricos, estatísticos e probabilísticos.

Tipos de dados: definição

- Os atributos, características ou variáveis podem ter natureza:
  - Qualitativa ou categórica
  - Quantitativa ou numérica

- Saber distinguir as variáveis é fundamental! Diferentes análises são usadas com diferentes tipos de variáveis.

Tipos de dados: categóricos

- Rótulos (*labels*) que atribuímos a conjuntos de coisas, categorias mutuamente exclusivas.
- Uma categoria é um conjunto de elementos que compartilham uma propriedade. Espécies, tipos. Não podem ser mensuradas.
- Uma categoria pode ter um ou mais níveis ou subcategorias. Sexo, por exemplo, costuma ter dois níveis (masc, fem), mas pode ter 3 (masc, fem, intersexual).
- Exemplos:
  - Cargos em uma empresa
  - Sexo
  - Doenças
  - Cor da pele

Tipos de dados: categóricos

- Variável binária: propriedade categórica com dois níveis, em geral, sim-não, presente-ausente. Costuma ser codificada em 0 e 1.

- Variável politômica: propriedade categórica com três ou mais níveis.

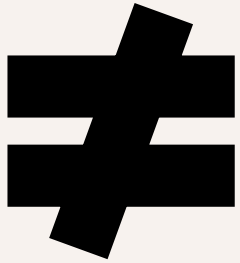
- As variáveis categóricas podem ter categorias ordenadas ou não:
  - **Variável nominal:** as subcategorias não apresentam ordem. *e.g.*sexo, cor dos olhos.
  - **Variável ordinal:** as subcategories apresentam uma ordem significativa. *e.g.* nível de escolaridade, classificação de satisfação.

Tipos de Dados: numéricos

- Coisas que existem como quantidades que podem ser mensuradas no mundo real.
- As variáveis quantitativas ou numéricas podem ser:
  - **Discretas:** contagens, números inteiros (0, 1, 2, ...).
  - **Contínuas:** medições sem restrição de precisão, números reais (1, 1.1, 1.2, 1.22, 1.223 …).
- Exemplos:
  - Idade
  - Pressão arterial
  - Altura
  - Dias desde a matrícula

Tipos de Dados: escala de mensuração

Natureza da variável



Escala de mensuração da variável

Tipos de Dados: escala de mensuração

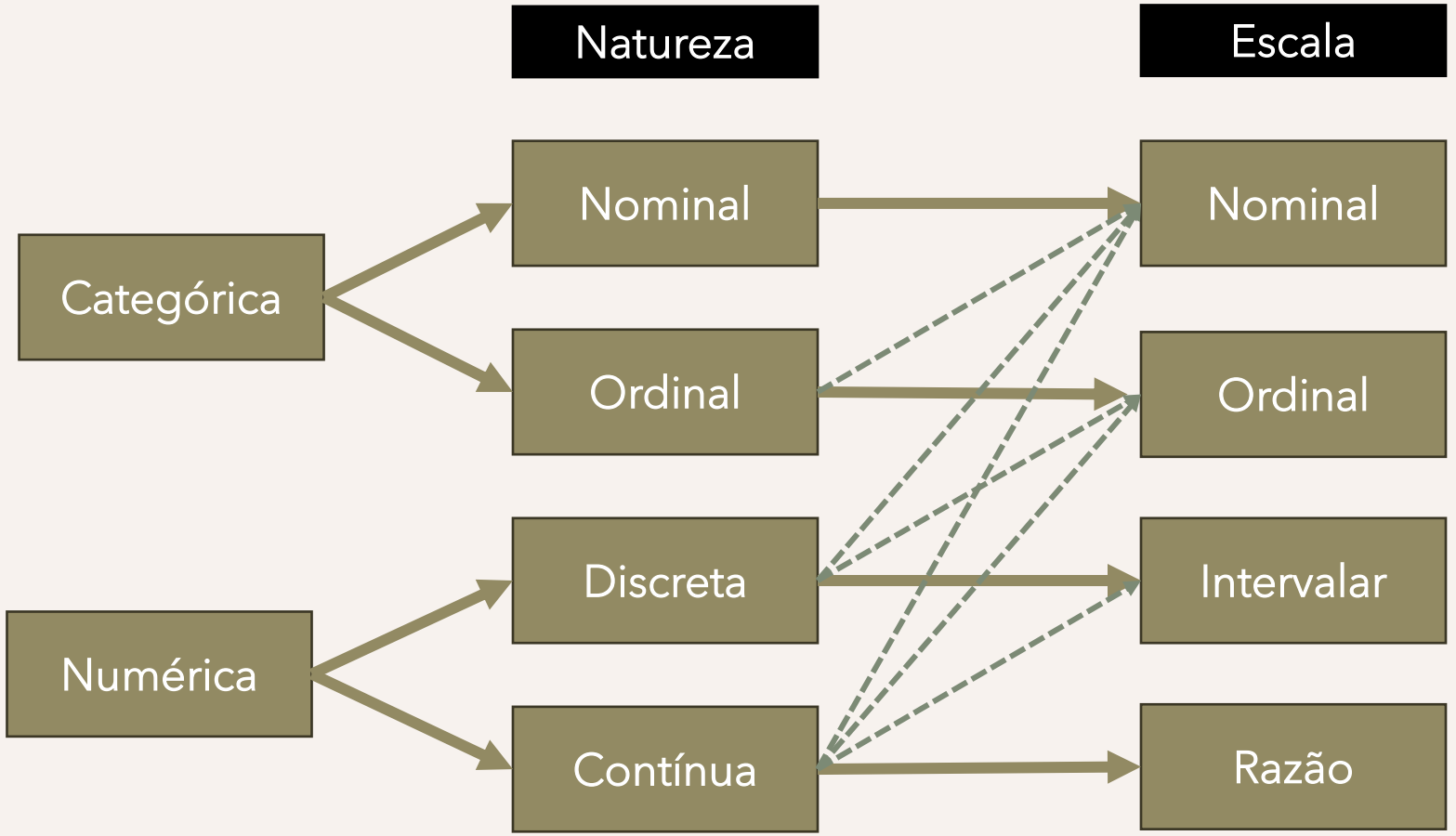
- Mensuração é o processo de atribuir números a propriedades de objetos do mundo real de acordo com uma regra especificada.

- A natureza das variáveis determina a informação contida nelas e especifica qual escala de mensuração pode ser utilizada.

Tipos de Dados: escala de mensuração

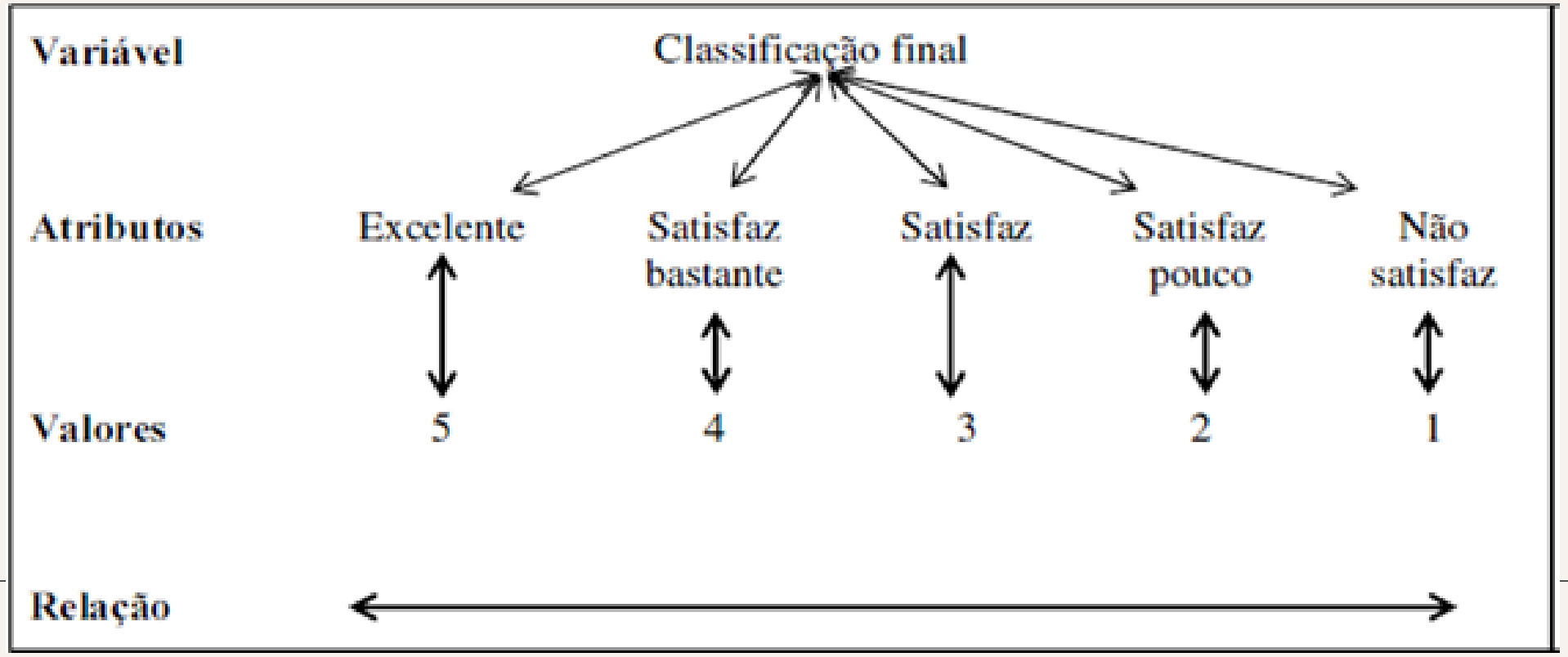
- Variáveis categóricas possuem escalas de mensuração:
  - **Nominal:** o valor dos números atribuídos é meramente nominal, *i.e.*, um "nome" para uma categoria.
  - **Ordinal:** além de valor nominal, a ordem de valor dos números expressa uma ordem natural nas categorias.
- Variáveis numéricas possuem escalas de mensuração:
  - **Intervalar/escalar:** assume que os intervalos entre os números são iguais, mas não permite avaliações de proporcionalidade ("dobro") porque o ponto "0" é arbitrário.
  - **De razão:** tem um ponto "0" significativo, *i.e.*, que indica ausência ou inexistência, permitindo avaliações de proporcionalidade ou razão.

Tipos de Dados: escala de mensuração



Tipos de Dados: escala de mensuração

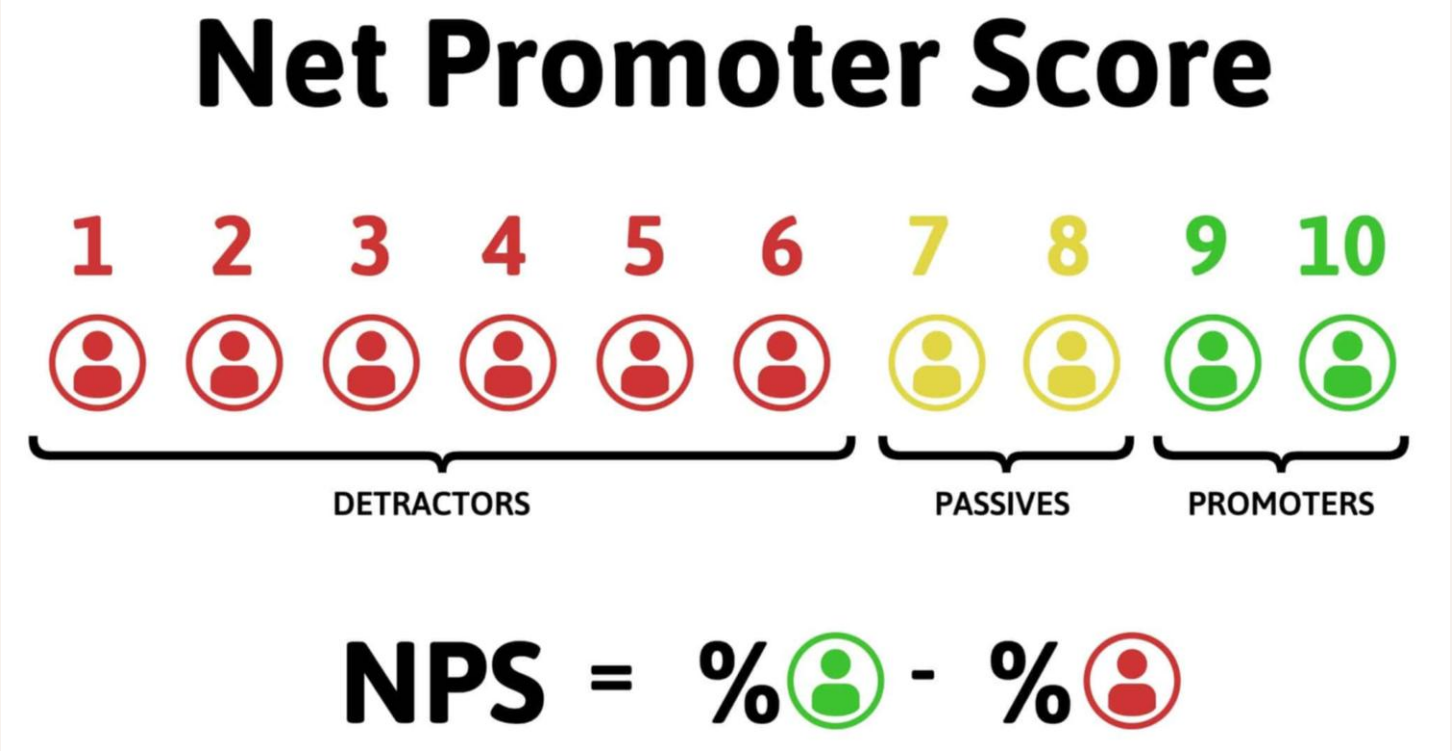
- Variáveis ordinais podem ser mensuradas de modo intervalar, como nas escalas tipo Likert. Quanto mais pontos na escala, mas ela se aproxima de uma variável discreta.



CONCORDÂNCIA	FREQUÊNCIA	SATISFAÇÃO
1. Discordo totalmente 2. Discordo parcialmente 3. Nem concordo, nem discordo 4. Concordo parcialmente 5. Concordo totalmente	1. Nunca 2. Raramente 3. Ocasionalmente 4. Frequentemente 5. Muito frequentemente	1. Muito insatisfeito 2. Parcialmente insatisfeito 3. Nem satisfeito, nem insatisfeito 4. Parcialmente satisfeito 5. Muito satisfeito
PERCEPÇÃO	**PROBABILIDADE**	**IMPORTÂNCIA**
1. Muito ruim 2. Ruim 3. Regular 4. Bom 5. Ótimo	1. Improvável 2. Pouco provável 3. Neutro 4. Provável 5. Muito provável	1. Nada importante 2. Pouco importante 3. Moderadamente importante 4. Importante 5. Muito importante

Tipos de Dados: escala de mensuração

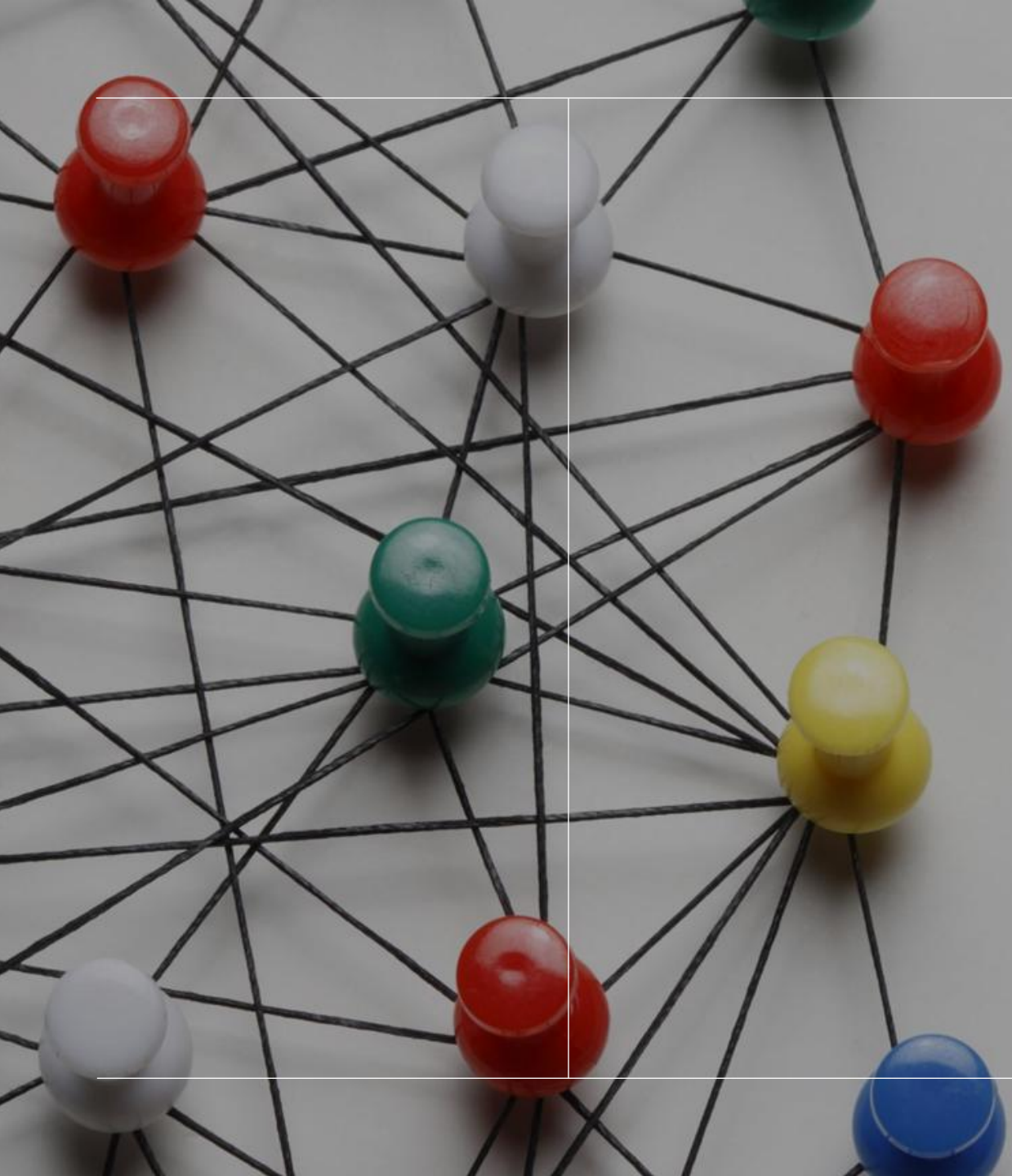
- Outra escala frequentemente usada é a Net Promoter Score (NPS):



Tipos de Dados: escala de mensuração

- Podemos criar escores itervalares a partir de múltiplas questões tipos Likert ou de questões de múltipla escolha:
  - Notas de provas
  - Escalas
  - Índice socioeconômico

- Podemos discretizar variáveis numéricas, criando categorias:
  - Faixa etária
  - Faixa de renda



Análises de frequência

Análises de frequência

- Variáveis categóricas são analisadas com:

  - **Frequências absolutas (*n*):** o número de ocorrências de cada subcategoria é contado e comparado entre as diferentes subcategorias.

  - **Frequências relativas (%):** o percentual que as ocorrências de cada subcategoria representa é computado e comparado entre as diferentes subcategorias

Análises de frequência

- Podemos analisar frequências absolutas por meio de gráficos de barras (vertical ou horizontal). Esses gráficos representam visualmente a distribuição dos dados para uma variável.

- Podemos analisar frequências relativas e absolutas por meio de tabelas de frequência.

- A escolha de gráficos ou tabelas depende de muitos fatores. Em geral usa-se tabelas, mas se houver muitas categorias, um gráfico pode ser mais adequado.

Análises de frequência: gráficos

Figura 1 – Proporção de participantes por sexo

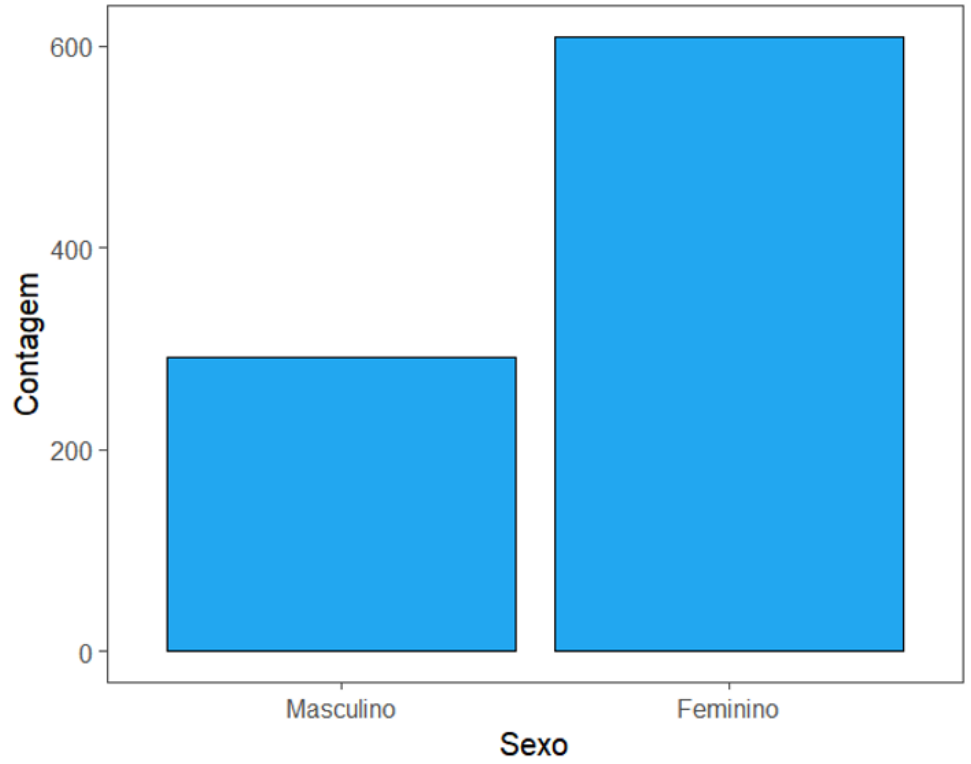
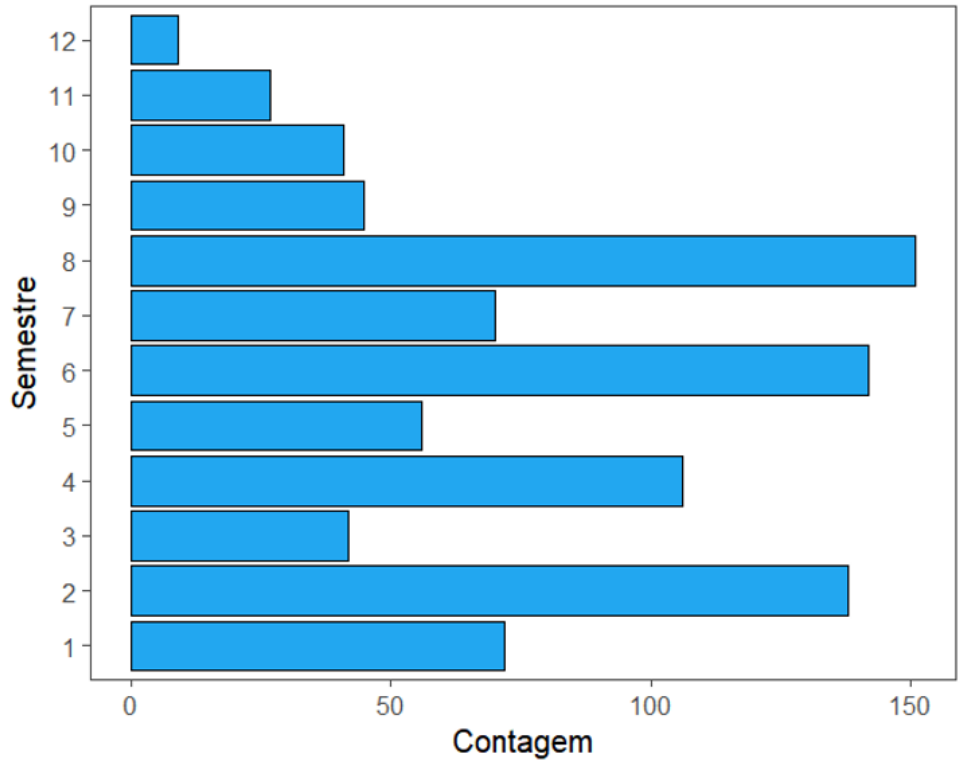


Figura 3 – Proporção de estudantes por semestre do curso



Análises de frequência: tabelas de frequência

- Cálculo de <u>percentuais</u>:

$$f_i = \frac{n_i}{N}$$

$$f_i = \frac{n_i}{N} * 100$$

Sexo	%	*n*
Masculino	63,1	590
Feminino	36,9	345
Total	100,00	935

- f_i frequência da i-ésima subcategoria (i = índice, posição)
- n_i número de ocorrências da i-ésima subcategoria
- N número total de casos

Análises de frequência: exercícios “na mão”

- Crie uma tabela com as frequências absolutas e relativas das subcategorias dos objetos listados abaixo e esboce um gráfico de barras vertical.

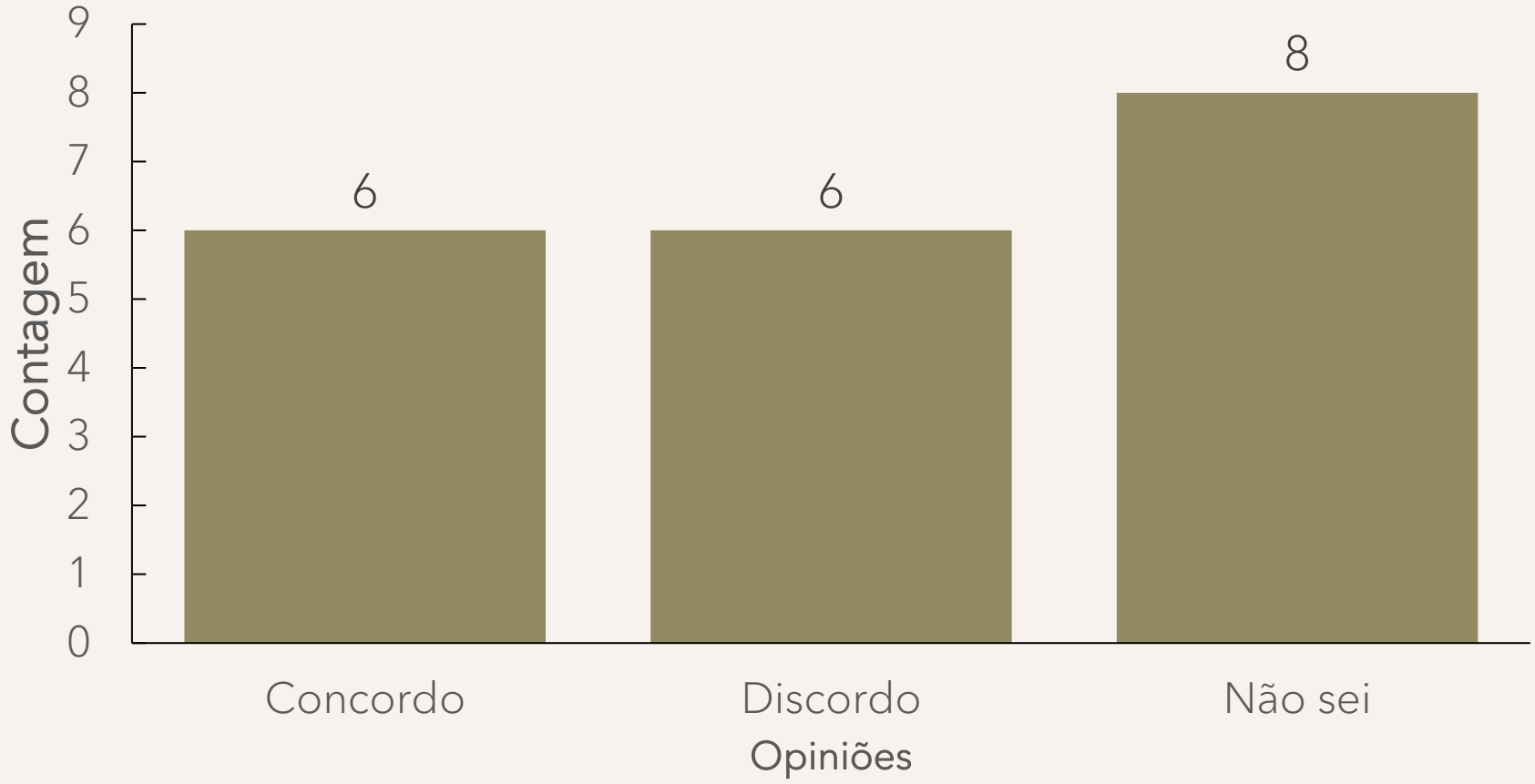
[concordo, concordo, concordo, discordo, discordo, concordo, não sei, não sei, discordo, concordo, concordo, não sei, não sei, não sei, não sei, não sei, não sei, discordo, discordo, discordo]

Análises de frequência: exercícios “na mão”

- Subcategorias: $\{f_{concordo}, f_{discordo}, f_{nãosei}\}$
- $f_{concordo} = \frac{6}{20} = 0{,}3 * 100 = 30\%$
- Concordo + discordo = opinião = 60%

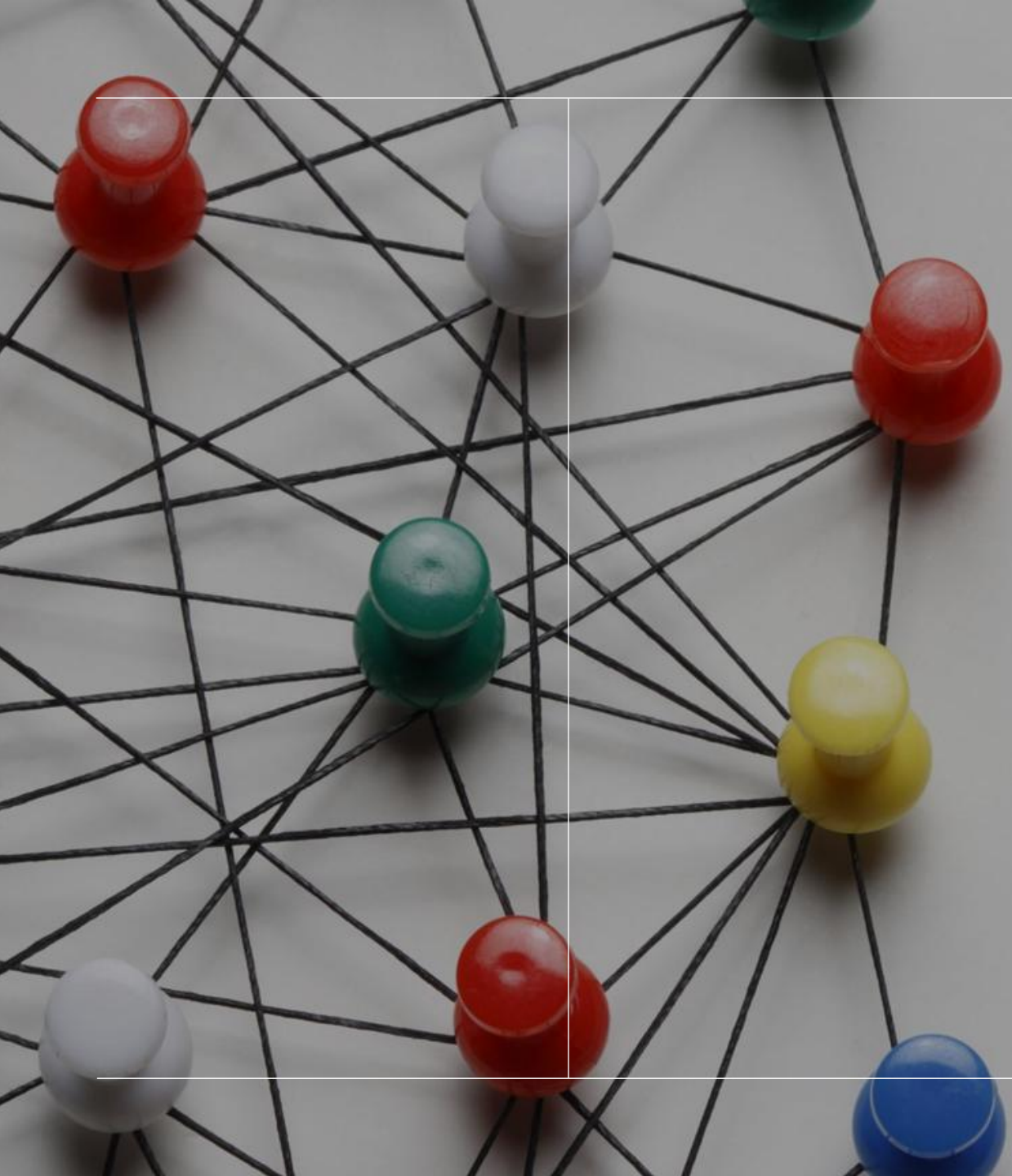
Subcategoria	%	*n*
Concordo	30,0%	6
Discordo	30,0%	6
Não sei	40,%	8
Total	100,0	20

Análises de frequência: exercícios “na mão”



Análises de frequência: exercícios no JASP

- Analise as frequências das variáveis por meio de gráficos de barras e tabelas de frequência:
  - sexo
  - tipo_trabalho
  - status_fumante
  - avc
  - avc X status_fumante



Distribuições

Distribuições: métricas

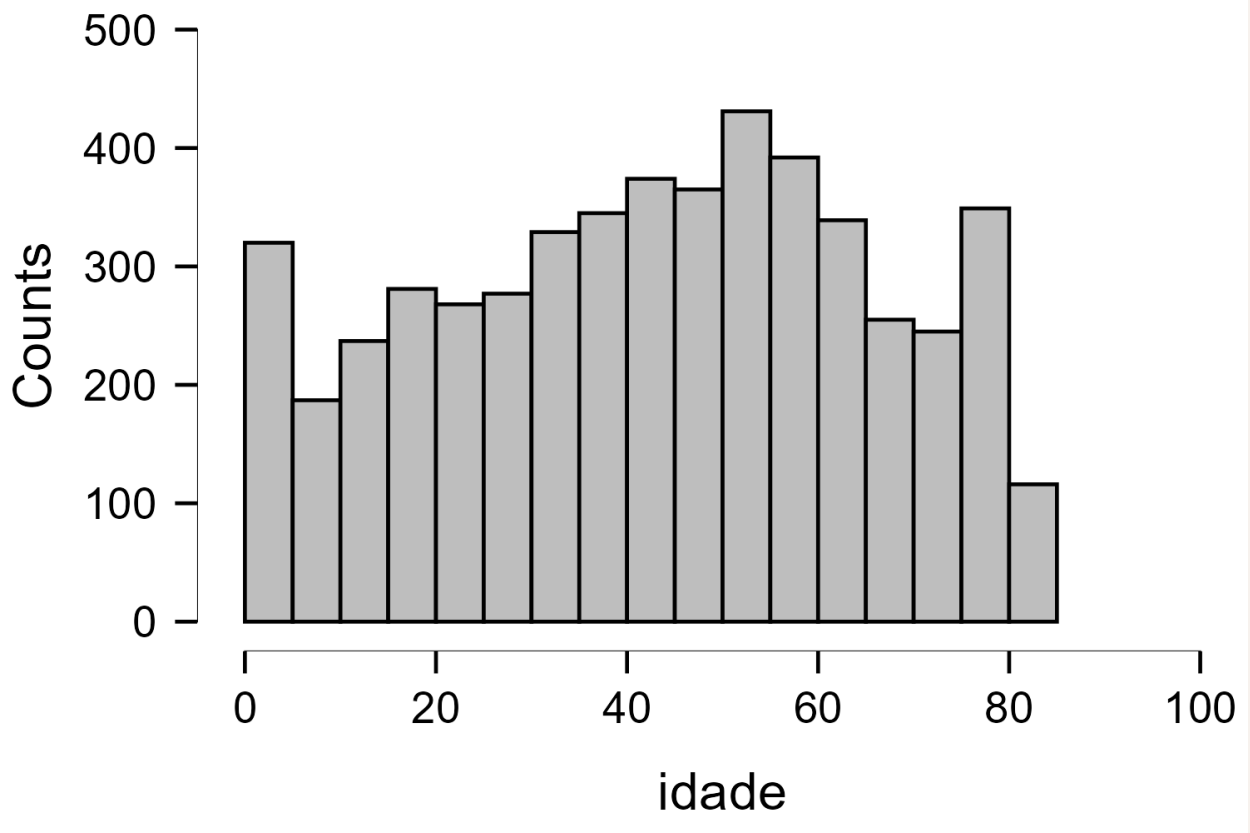
- As variáveis quantitativas ou numéricas são analisadas com medidas de localidade e espalhamento.

- Todas as variáveis quantitativas apresentam uma distribuição, representada visualmente por um gráfico chamado histograma de frequências.

- As medidas de localidade buscam resumir o centro ou tendência central dessas distribuições, onde os valores mais frequentes se agrupam. Já as medidas de espalhamento buscam resumir a variabilidade dos valores da distribuição em torno do centro.

Distribuições: histograma

- Os valores ordenados da variável são apresentados no eixo-x.
- A frequência de cada categoria de valores da variável é representada no eixo-y, pela altura da barra.
- As categorias de valores, *bins* (cestas, agrupamentos), discretizam a variável numérica. Podemos especificar o número de *bins* ou usar uma fórmula automática.



Distribuições: histograma

- Passo a passo para construir um histograma de frequências:
  - Escolha uma variável quantitativa com muitos valores diferentes
  - Ordene os dados do menor para o maior
  - Conte o número de valores diferentes
  - Encontre um fator que divide esse número e resulta em menos de 30 ou use uma regra teórica para dividir os bins (5 em 5 anos)
  - Use esse fator para definir o número de *bins*
  - Faça a contagem do número de valores em cada *bin*
  - Plote um gráfico de barras com as barras contíguas umas às outras

Distribuições: exercício “na mão”

[10, 11, 7, 20, 21, 15, 3, 2, 3, 4, 5, 6, 6, 5, 5, 5, 5, 11, 12, 13, 15, 17, 18, 19, 20, 13, 12, 1, 1, 1, 9, 9, 9, 8, 7]

N = 35

Use bins de 5 elementos cada

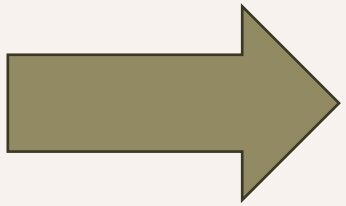
Distribuições: exercício “na mão”

[1, 1, 1, 2, 3, 3, 4, 5, 5, 5, 5, 5, 6, 6, 7, 7, 8, 9, 9, 9, 10, 11, 11, 12, 12, 13, 13, 15, 15, 17, 18, 19, 20, 20, 21]

Valores	*n*	Valores	*n*
1	3	11	2
2	1	12	2
3	2	13	2
4	1	15	2
5	5	17	1
6	2	18	1
7	2	19	1
8	1	20	2
9	3	21	1
10	1		

Distribuições: exercício “na mão”

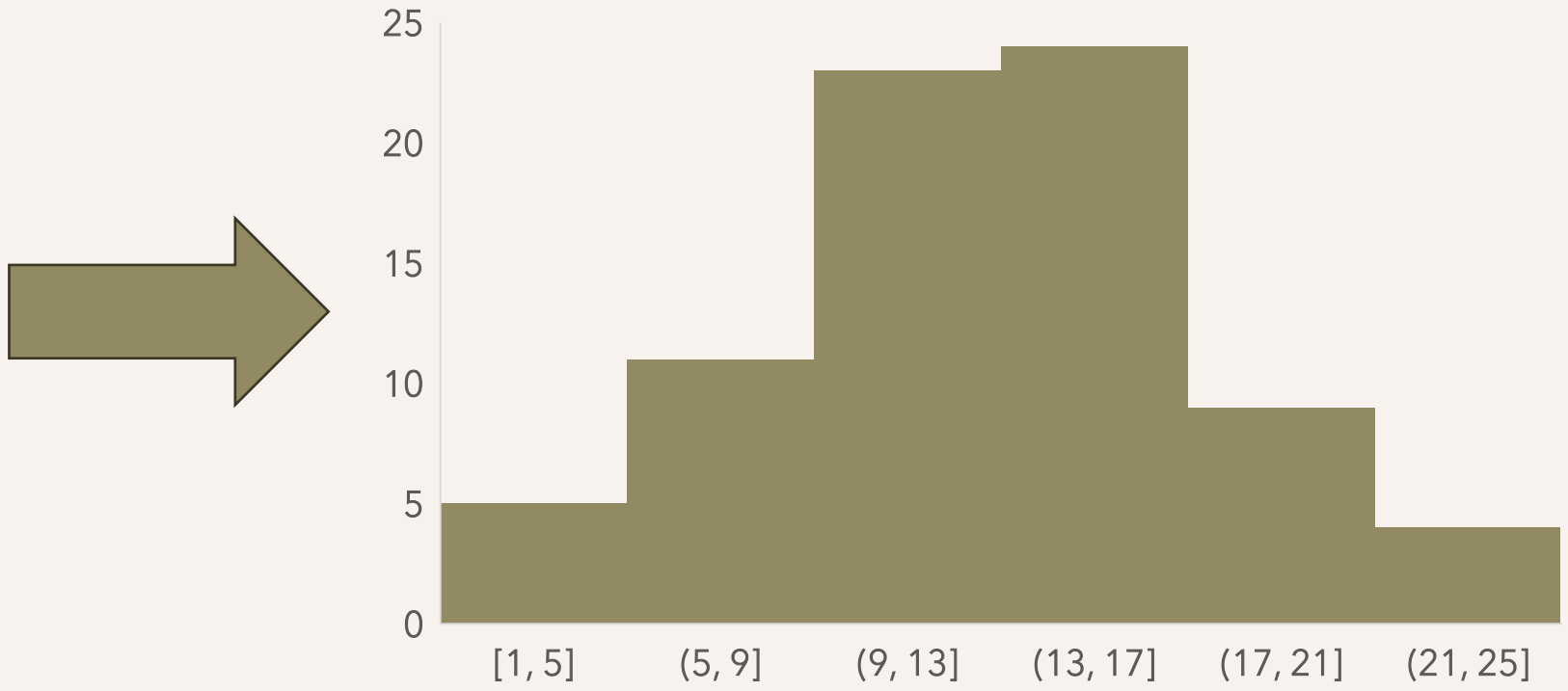
Valores	*n*	Valores	*n*
1	3	11	2
2	1	12	2
3	2	13	2
4	1	15	2
5	5	17	1
6	2	18	1
7	2	19	1
8	1	20	2
9	3	21	1
10	1		



bins	*n*
0-4	7
5-9	13
10-14	7
15-19	5
20-24	3

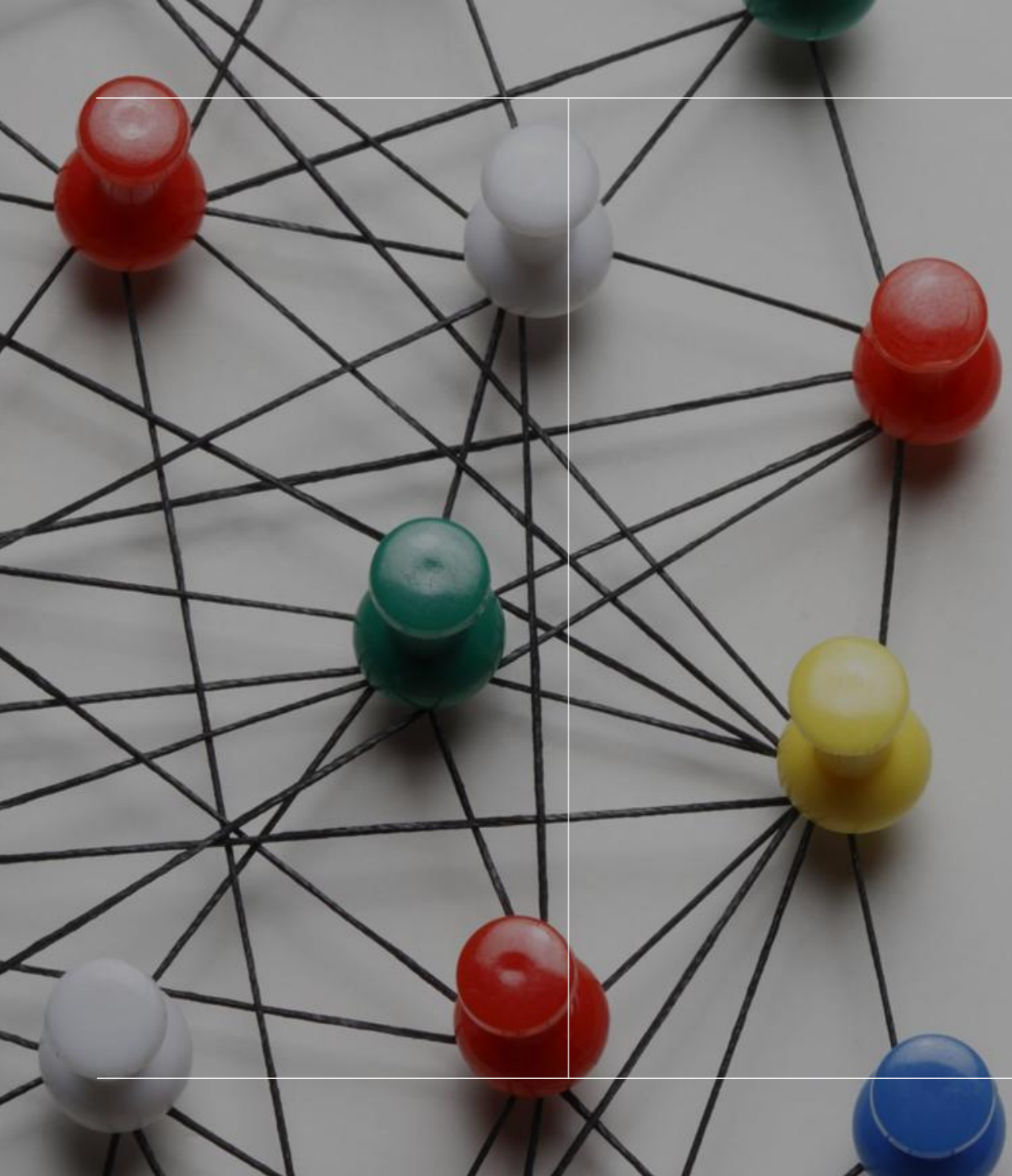
Distribuições: exercício “na mão”

bins	*n*
0-4	7
5-9	13
10-14	7
15-19	5
20-24	3



Distribuições: exercícios no JASP

- Analise os histogramas de frequência das seguintes variáveis:
  - idade
  - imc
  - nivel_glicose_med
  - imc X avc



Medidas de localidade

Medidas de localidade

- Uma distribuição representa a informação contida em uma variável numérica, mas, como resumir essa informação?

- Podemos usar métricas ou estatísticas descritivas para resumir a informação contida na distribuição.

- As principais medidas de localidade ou de tendência central são:
  - Média
  - Mediana
  - Moda

Medidas de localidade: média

- Medida de localidade mais usada.
- Soma todos os valores e divide pelo número de valores. Usa o máximo de informação.
- Funciona como uma estimativa do valor mais frequente em uma distribuição, o *"valor esperado"* (nome da média na teoria da probabilidade).

Medidas de localidade: média

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

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- $\bar{x}$ x-barra, símbolo da media
- $\sum_{i=1}^{n} x_i$ sigma indica somatório de todos os “x”, de *i* à *n*

Medidas de localidade: média

- Calcule a média do seguinte conjunto de dados:

$$[5, 5, 17, 18, 7]$$

Medidas de localidade: média

$$\bar{x} = \frac{5 + 5 + 17 + 18 + 7}{5} = \frac{52}{5} = 10.4$$

Medidas de localidade: média

- Uma limitação da média é sua sensibilidade a valores extremos *(outliers)*, que "puxam" a média na sua direção.

- Calcule a média [1, 2, 3, 4, 4, 4, 2, 5, 20]: 5
- Calcule a média [1, 2, 3, 4, 4, 4, 2, 5, 200]: 25

- Por exemplo, a média não é uma boa estimativa da renda da população!

Medidas de localidade: mediana

- A mediana estima a tendência central de uma distribuição a partir do valor que está "no meio", *i.e.*, ela não lida com os valores em si, mas com a posição deles. Isso torna a mediana "robusta" aos outliers, mas menos informativa.

- Para calcular a mediana:
  - Ordene o conjunto de dados do menor para o maior valor
  - Se o número de valores for ímpar, faça $(n + 1)/2$ e pegue o valor que está no meio.
  - Se o número de valores for par, faça $n/2$ e calcule a média do valor que está no meio e do seu vizinho à direita.

Medidas de localidade: mediana

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & \text{se } n \text{ é ímpar} \\ \dfrac{x_{n/2} + x_{(n/2)+1}}{2} & \text{se } n \text{ é par} \end{cases}$$

Medidas de localidade: mediana

- Calcule a mediana do seguinte conjunto de dados:

$$[5, 5, 17, 18, 7]$$

Medidas de localidade: mediana

- Dados ordenados: [5, 5, 7, 17, 18].
- $n = 5$, logo n é impar.
- (5 + 1)/2 = 3, portanto a mediana é o valor que está na terceira posição do conjunto de dados ordenado.
- Mediana = 7.

Medidas de localidade: moda

- A moda é o valor que se repete com mais frequência na distribuição.
- A moda do conjunto [5, 5, 7, 17, 18] é 5, porque se repete 2 vezes.
- Em um histograma, a moda é a barra mais alta, o valor modal.

Medidas de localidade: exercício no JASP

- Calcule a média, mediana e moda das variáveis:
  - idade
  - imc
  - nivel_glicose_med
  - imc X avc

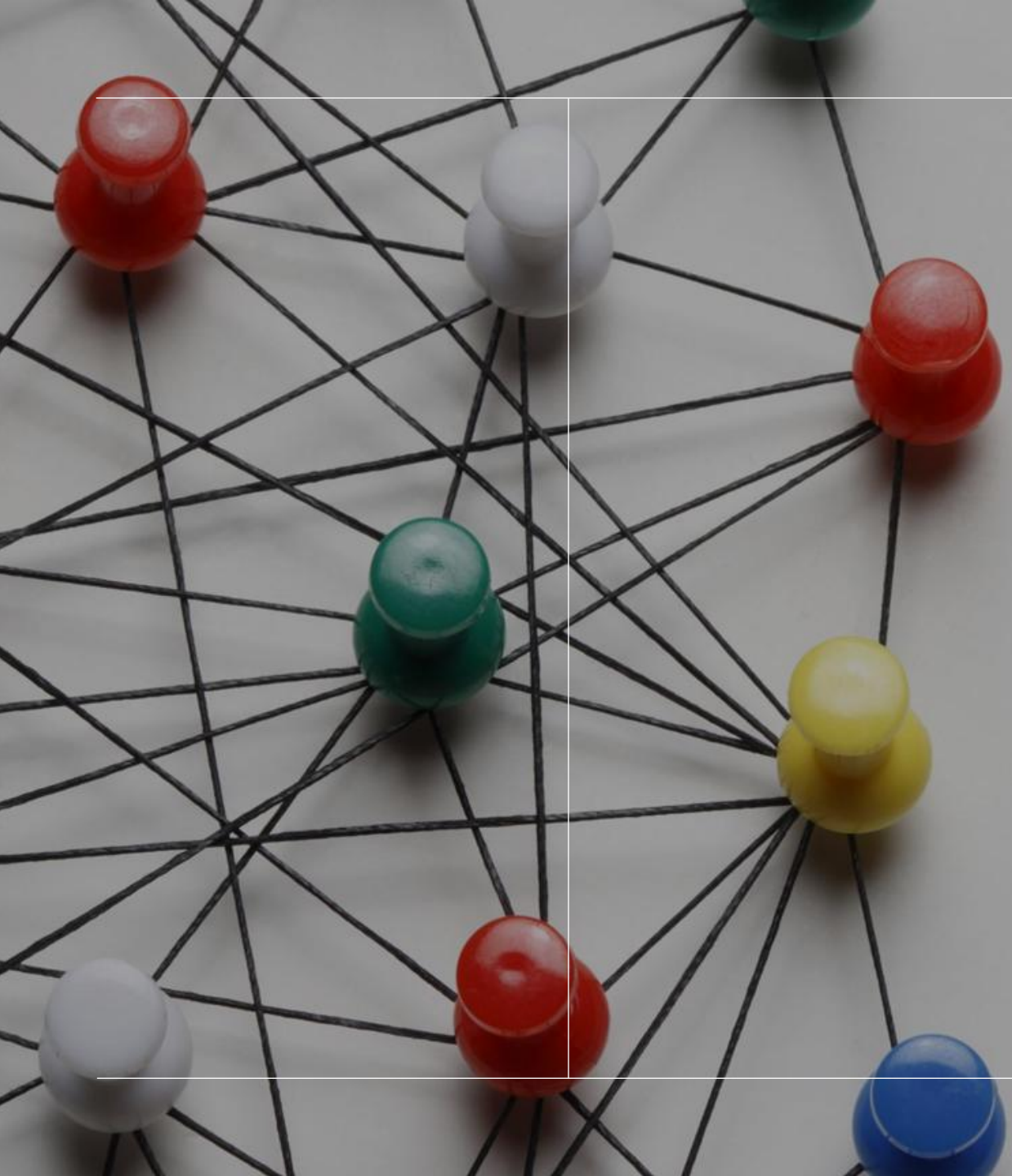
Medidas de localidade: outras médias

- Existem variações da média:
  - Média ponderada

$$\overline{x_w} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - Média aparada

$$\overline{x_{aparada}} = \frac{\sum_{i=k}^{n-k} x_i}{n - 2k}$$



Medidas de espalhamento

Medidas de espalhamento

- Medidas de localidade nos permitem resumir a tendência central dos dados, i.e., como eles se agrupam.

- Medidas de espalhamento nos ajudam a resumir a variabilidade dos dados.

- Queremos compreender como os valores de uma variável se assemelham uns aos outros e se diferenciam uns dos outros.

Medidas de espalhamento: intervalo

- A medida mais simples de espalhamento é a amplitude ou intervalo:

$$intervalo(x) = max(x) - min(x)$$

- Em geral é pouco usada, simples demais.

Medidas de espalhamento: desvios

- A média é um modelo do centro de uma distribuição.
- Podemos compreender a variabilidade de uma variável analisando o quanto os valores se distanciam da média, i.e., podemos calcular os *desvios:*

$$x_i - \bar{x}$$

- Valores positivos indicarão que um ponto nos dados está acima da média, enquanto valores negativos indicarão um ponto nos dados abaixo da média.

x	Desvio
5	0.7
4	-0.3
3	-1.3
6	1.7
7	2.7
1	-3.3
1	-3.3
9	4.7
3	-1.3
Média = 4,3	Soma = 0

Medidas de espalhamento: soma dos quadrados

- Como a soma dos desvios é 0, podemos usar dois truques matemáticos para somar os desvios:
  - Elevar ao quadrado: $(x_i - \bar{x})^2$
  - Aplicar o valor absoluto (transformar negativos em positivos): $|x_i - \bar{x}|$

- Elevando cada desvio ao quadrado e somando, temos a soma dos quadrados, uma medida geral da variabilidade nos dados:

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

x	Desvio	Desvio^2
5	0.7	0.49
4	-0.3	0.09
3	-1.3	1.69
6	1.7	2.89
7	2.7	7.29
1	-3.3	10.89
1	-3.3	10.89
9	4.7	22.09
3	-1.3	1.69
Média = 4,3	Soma = 0	Soma = 58.01

Medidas de espalhamento: variância

- Podemos melhorar nossa medida de espalhamento ao calcular a média com a soma dos quadrados, nos trazendo a variância:

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

Ou:

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

Var = 7,25

x	Desvio	Desvio^2
5	0.7	0.49
4	-0.3	0.09
3	-1.3	1.69
6	1.7	2.89
7	2.7	7.29
1	-3.3	10.89
1	-3.3	10.89
9	4.7	22.09
3	-1.3	1.69
Média = 4,3	Soma = 0	Soma = 58,01

Medidas de espalhamento: desvio-padrão

- A variância nos fornece uma métrica do desvio médio ao quadrado nos dados. Se tirarmos a raiz quadrada da variância, anulamos o efeito de elevar ao quadrado e obtemos o desvio-padrão, uma medida de variabilidade que está na mesma unidade dos nossos dados.

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

DP = 2,69

Medidas de espalhamento: exercício “na mão”

- Com a variável à esquerda, calcule:
  - Intervalo
  - Média
  - Mediana
  - Soma dos quadrados
  - Variância
  - Desvio-padrão

$$intervalo(x) = max(x) - min(x)$$

$$desvio = x_i - \bar{x}$$

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

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

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

Índice	Variável X
1	20
2	12
3	13
4	14
5	10
6	2
7	0
8	1
9	8
10	9
11	4
12	17

Medidas de espalhamento: exercício “na mão”

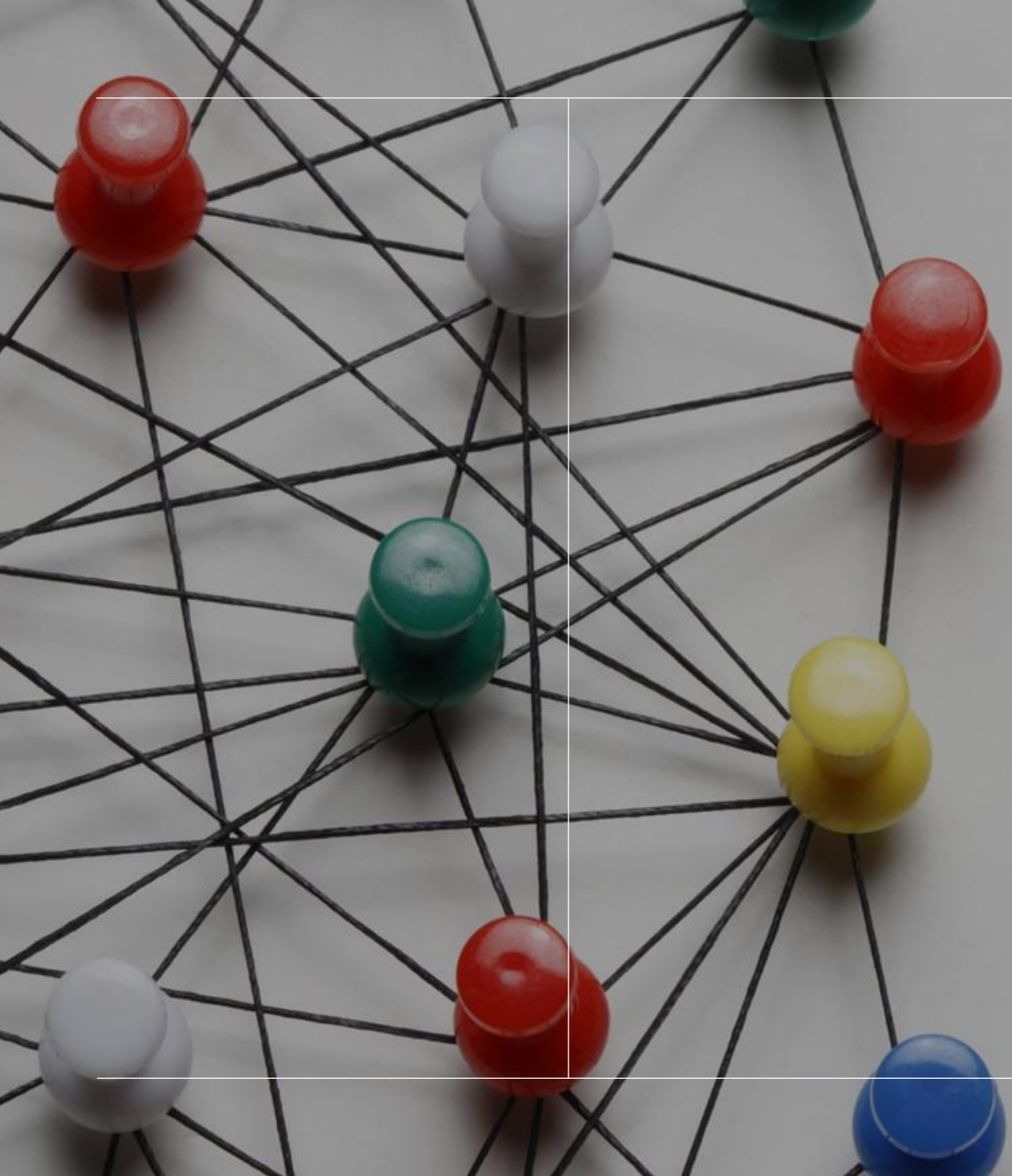
- Com a variável à esquerda, calcule:
  - Intervalo
  - Média
  - Mediana
  - Soma dos quadrados
  - Variância
  - Desvio-padrão

Métrica	Valor
Intervalo	20
Média	9,17
Mediana	9,5
SQ	455,67
Variância	41,42
Desvio-Padrão	6,43

Índice	Variável X
1	20
2	12
3	13
4	14
5	10
6	2
7	0
8	1
9	8
10	9
11	4
12	17

Medidas de espalhamento: exercício no JASP

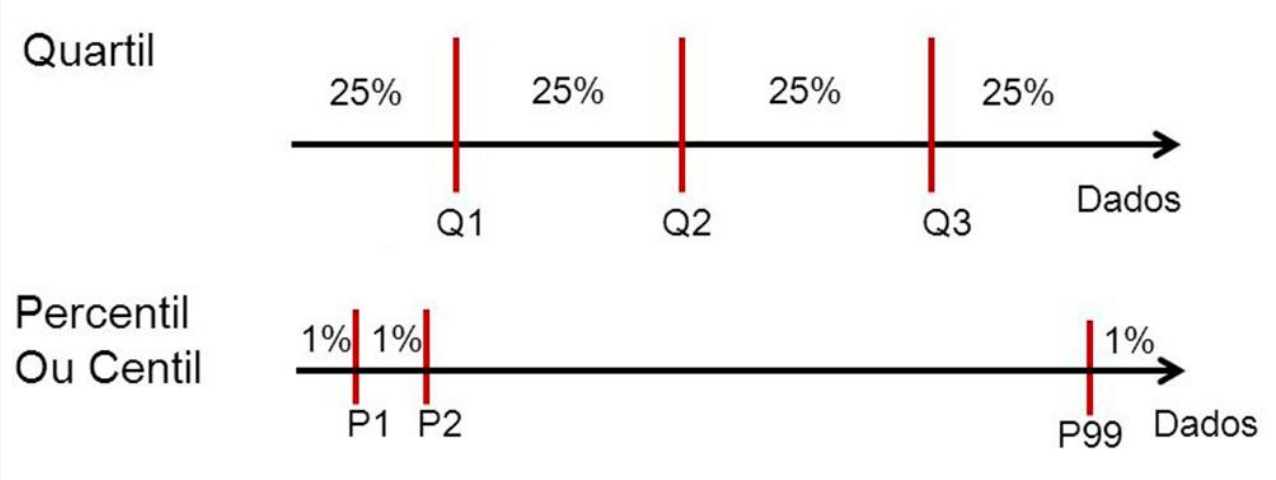
- Calcule o intervalo, variância e desvio-padrão:
  - idade
  - imc
  - nivel_glicose_med
  - imc X avc



Medidas de partição

Medidas de partição: quartis

- É possível cortar a distribuição dos dados em partes com proporções iguais:
  - **Quartis:** 4 partes, cada uma representando 25% dos dados
  - **Decis:** 10 partes, cada uma representando 10% dos dados
  - **Percentis:** 100 partes, cada uma representando 1% dos dados



Medidas de partição: quartis

- Passo a passo para calcular quartis:
  - Ordene o conjunto de dados
  - Calcule a mediana (valor de meio se *n* ímpar; média dos valores do meio se *n* par), este é o Q_2
  - Calcule a mediana da metade de baixo, este é o Q_1
  - Calcule a mediana da metade de cima, este é o Q_3

Medidas de partição: percentis

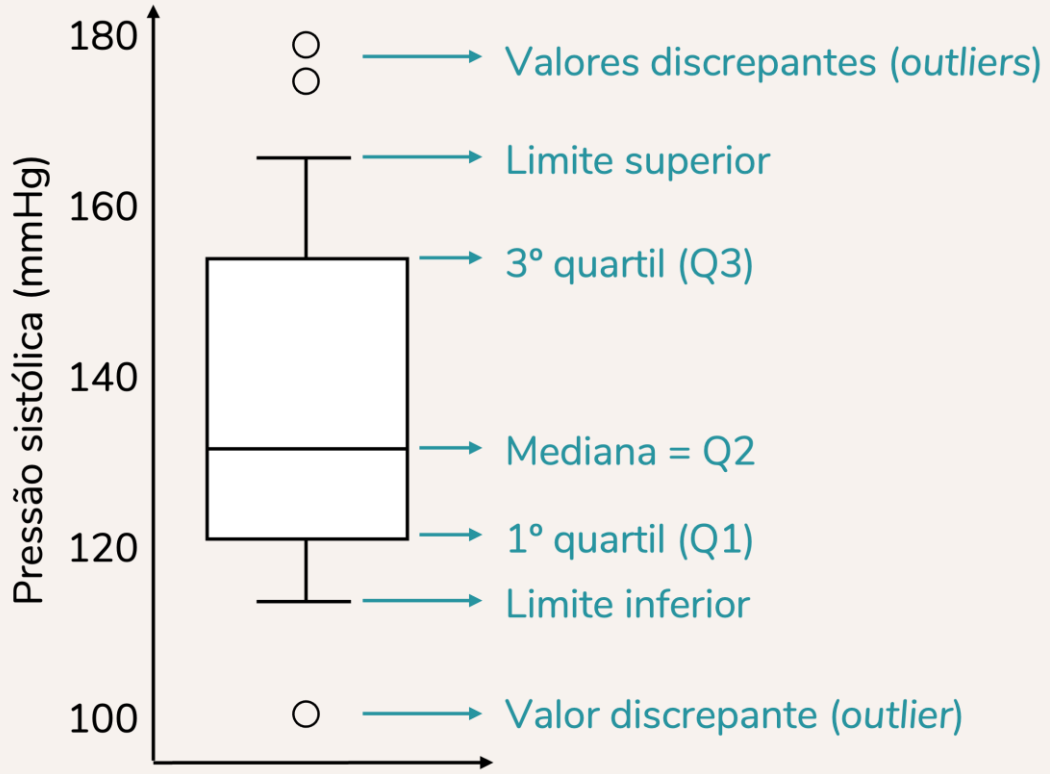
- Os algoritmos para calcular decis e percentis são mais complexos! O que nos importa é saber interpretá-los.

- Quartis, decis e percentis são valores dos nossos dados que marcam posições em relação a proporção dos dados. Um valor acima do valor que marca o Q_3 está acima de 75% dos dados. Um valor acima do valor que marca o P_{99} (percentil 99) está acima de 99% dos dados.

- Os testes de atenção no trânsito usam Percentis para interpretar os escores.

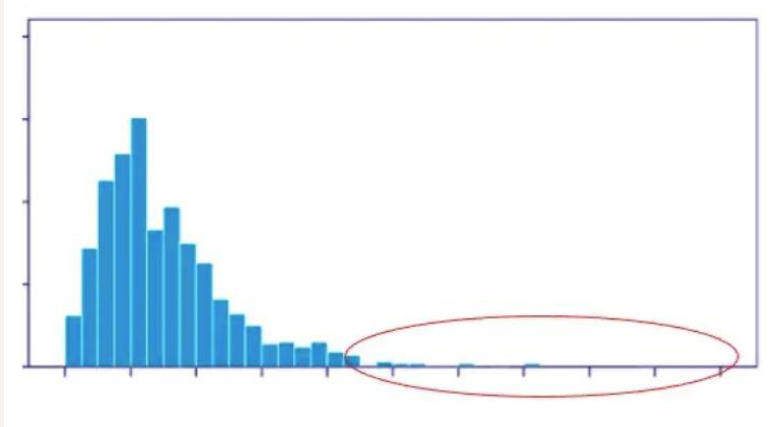
Medidas de partição: box-plots

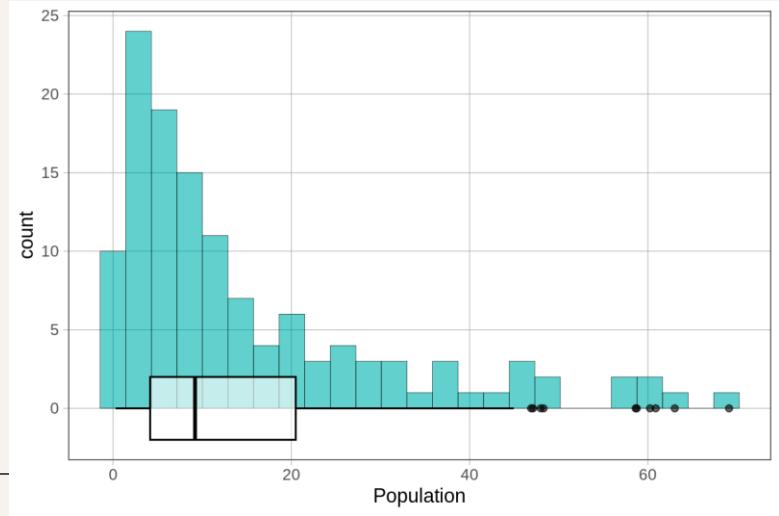
- Os *box-plots* são gráficos que nos ajudam a resumir a distribuição dos dados em uma variável numérica e detectar *outliers*.



Medidas de partição: outliers

- Os *outliers* são valores extremos, incomuns, muito diferentes da maioria.
- Por essas caractrísticas, podem ser um objeto de análise – detecção de anomalias.
- Em geral, os *outliers* impedem a detecção dos padrões que queremos identificar nos dados, sendo considerados viéses. Assim, eliminamos ou tratamos esses *outliers* antes de rodar nossas análises.





Medidas de partição: outliers

- A melhor forma de detector *outliers* é pelo box-plot, os pontos fora da caixa, detectados pela "regra do box-plot".

- Regra do box-plot:

$$LS = Q_3 + (1{,}5 * AIQ)$$
$$LI = Q_1 - (1{,}5 * AIQ)$$

Onde,

$$AIQ = Q_3 - Q_1$$

e

$$x_i > LS = outlier, x_i < LI = outlier$$

LS = Limite Superior

LI = Limite Inferior

AIQ = Amplitude Inter-Quartil

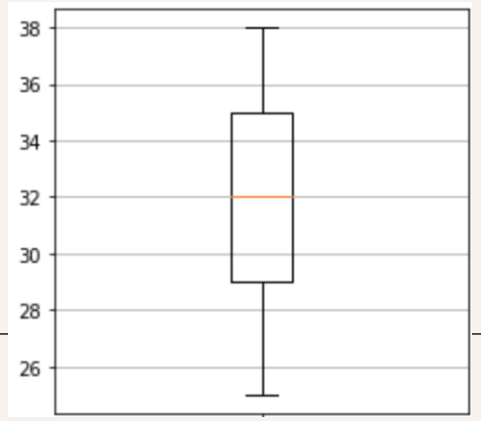
Medidas de partição: exercício “na mão”

- Calcule:
  - Valor mínimo e valor máximo
  - Quartis 1, 2 e 3
  - Amplitude Inter-Quartil
  - Limites do box-plot
  - Desenhe um box-plot

$$\text{Mdn(x)} = \begin{cases} x_{(n+1)/2} & \text{se } n \text{ é ímpar} \\ \dfrac{x_{n/2} + x_{(n/2)+1}}{2} & \text{se } n \text{ é par} \end{cases}$$

$$LS = Q_3 + (1{,}5 * AIQ)$$
$$LI = Q_1 - (1{,}5 * AIQ)$$

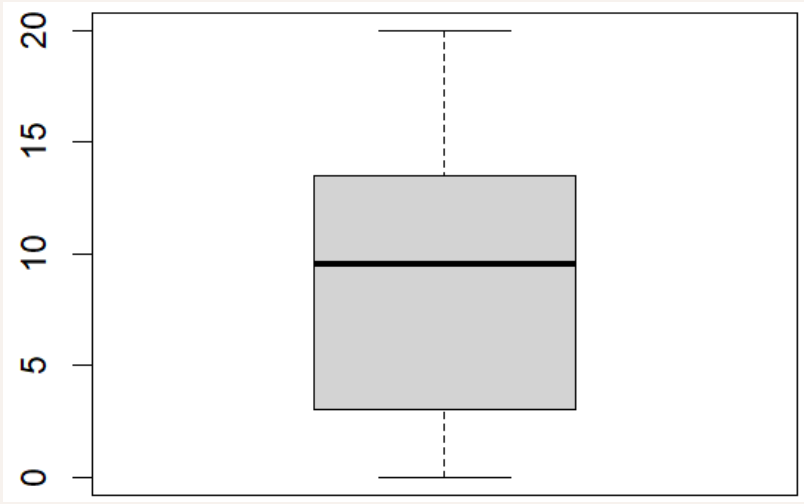
$$AIQ = Q_3 - Q_1$$



Índice	Variável X
1	20
2	12
3	13
4	14
5	10
6	2
7	0
8	1
9	8
10	9
11	4
12	17

Medidas de partição: exercício “na mão”

- Calcule:
  - Valor mínimo e valor máximo
  - Quartis 1, 2 e 3
  - Amplitude Inter-Quartil
  - Limites do box-plot
  - Desenhe um box-plot



Métrica	Valor
Min-Max	0 - 20
Q1	3,50
Q2	9,50
Q3	13,25
AIQ	9,75
LI	-11,1
LS	27,9

Medidas de partição: exercício no JASP

- Plote um box-plot para cada uma das variáveis abaixo:
  - idade
  - imc
  - nivel_glicose_med
  - imc X avc

Para se aprofundar...