A natureza da Estatística

A natureza da Estatística

O material dessa apresentação faz parte do primeiro capítulo de Estatística do Projeto Livro Aberto de Matemática para o Ensino Médio (www.umlivroaberto.com).

umlivroaberto

Introdução

- A todo momento somos expostos a informações sobre pesquisas científicas comprovando estatisticamente que tal substância causa uma patologia, sobre pesquisas de opinião, sobre projeções da estrutura etária da população daqui a 50 anos e outros temas de natureza incerta.
- Ter espírito crítico para informações sujeitas à incerteza a fim de poder interpretá-las e, quando necessário, poder escolher, entre diferentes opções, aquela que parece melhor diante da incerteza é uma habilidade necessária.
- A Estatística é uma disciplina fundamental para todos os estudantes e, certamente, tem grande responsabilidade na formação crítica do cidadão.

Introdução

- Existem várias definições para a Estatística.
- **Estudo das populações, das variações, e dos métodos de redução de dados.** (R. Fisher – 1890-1962)
- **Ciência de aprendizagem a partir dos dados.** (J. Kettenring)
- Uma boa definição : **Arte e ciência de coletar, organizar, analisar e interpretar dados, para que se tomem decisões sob incerteza**.

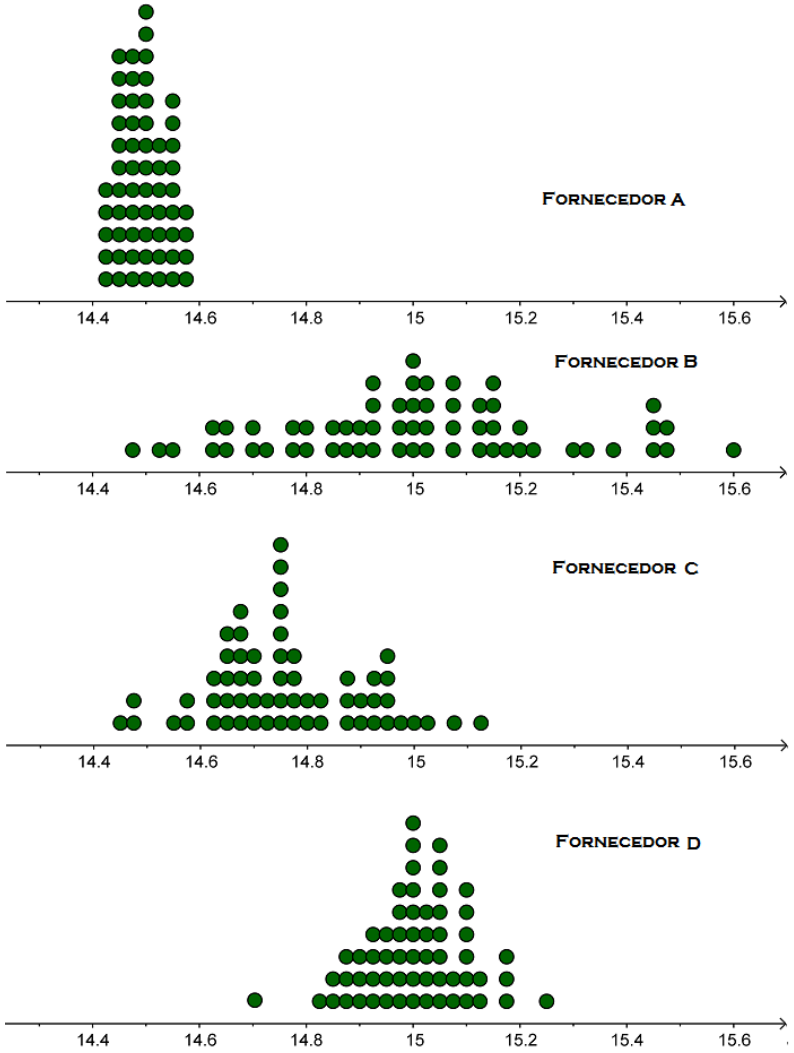
Introdução

- Existem formas diferentes de abordar a Estatística como uma disciplina. A abordagem voltada para **modelagem** será usada nesse material, adaptando-a para o nível do Ensino Médio.
- O propósito fundamental é, a partir de um conjunto de dados coletados, **visualizar estruturas**. Os primeiros passos nesse sentido envolvem a organização dos dados em tabelas de frequências e a construção de gráficos.

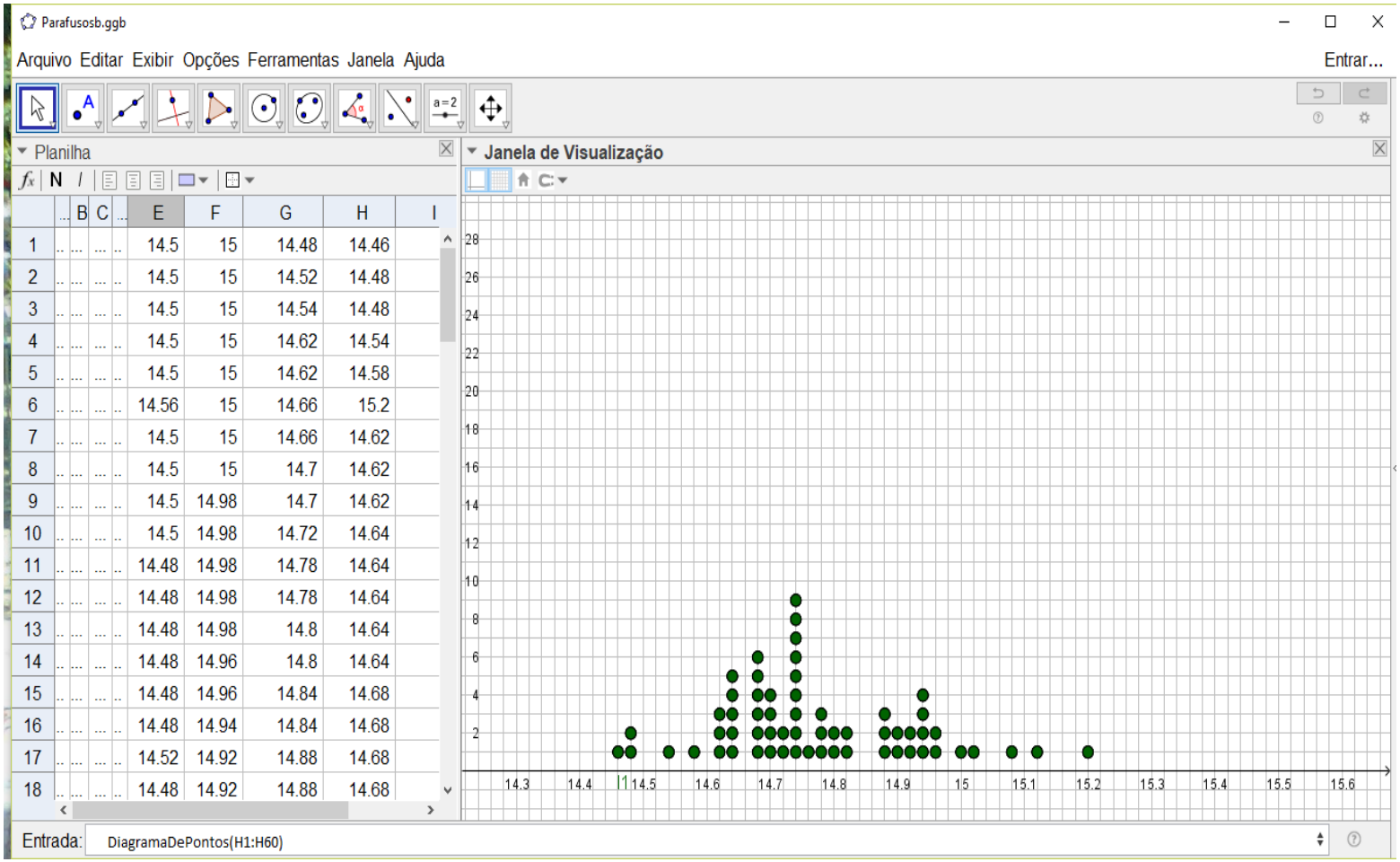
Distribuição

- É uma coleção de propriedades de um conjunto de dados como um todo, não de um particular valor do conjunto.
- Uma distribuição consiste de todos os valores diferentes nos dados, incluindo as frequências (ou probabilidades) associadas a cada valor.
- Variação e distribuição estão relacionadas a outras noções estatísticas fundamentais tais como "centro" (modeladas pela média, mediana, ou moda), dispersão (modeladas pelo desvio-padrão, ou variância, etc) e forma (por exemplo, bi-modal, uniforme, simétrica, assimétrica à direita, etc).

1) Escolha do fornecedor



GeoGebra: DiagramaDePontos

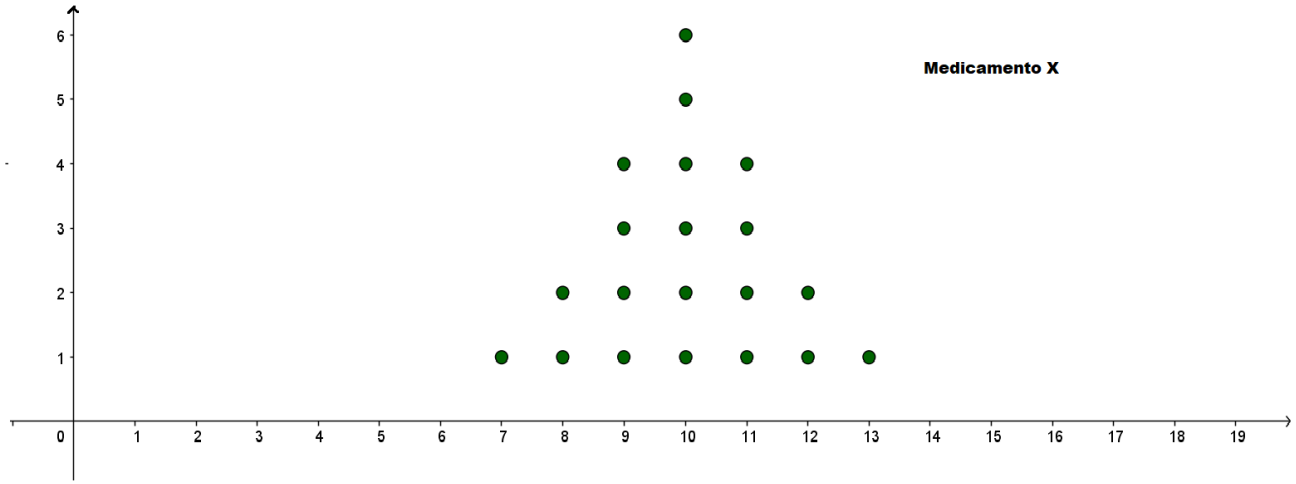


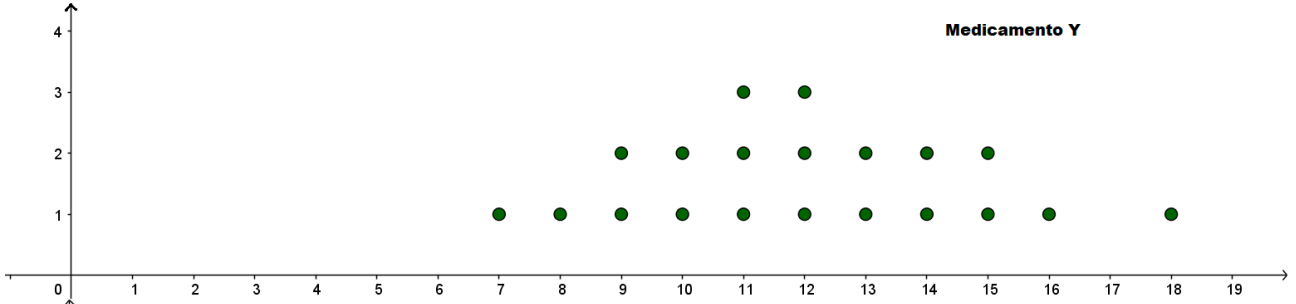
	E	F	G	H
1	14.5	15	14.48	14.46
2	14.5	15	14.52	14.48
3	14.5	15	14.54	14.48
4	14.5	15	14.62	14.54
5	14.5	15	14.62	14.58
6	14.56	15	14.66	15.2
7	14.5	15	14.66	14.62
8	14.5	15	14.7	14.62
9	14.5	14.98	14.7	14.62
10	14.5	14.98	14.72	14.64
11	14.48	14.98	14.78	14.64
12	14.48	14.98	14.78	14.64
13	14.48	14.98	14.8	14.64
14	14.48	14.96	14.8	14.64
15	14.48	14.96	14.84	14.68
16	14.48	14.94	14.84	14.68
17	14.52	14.92	14.88	14.68
18	14.48	14.92	14.88	14.68

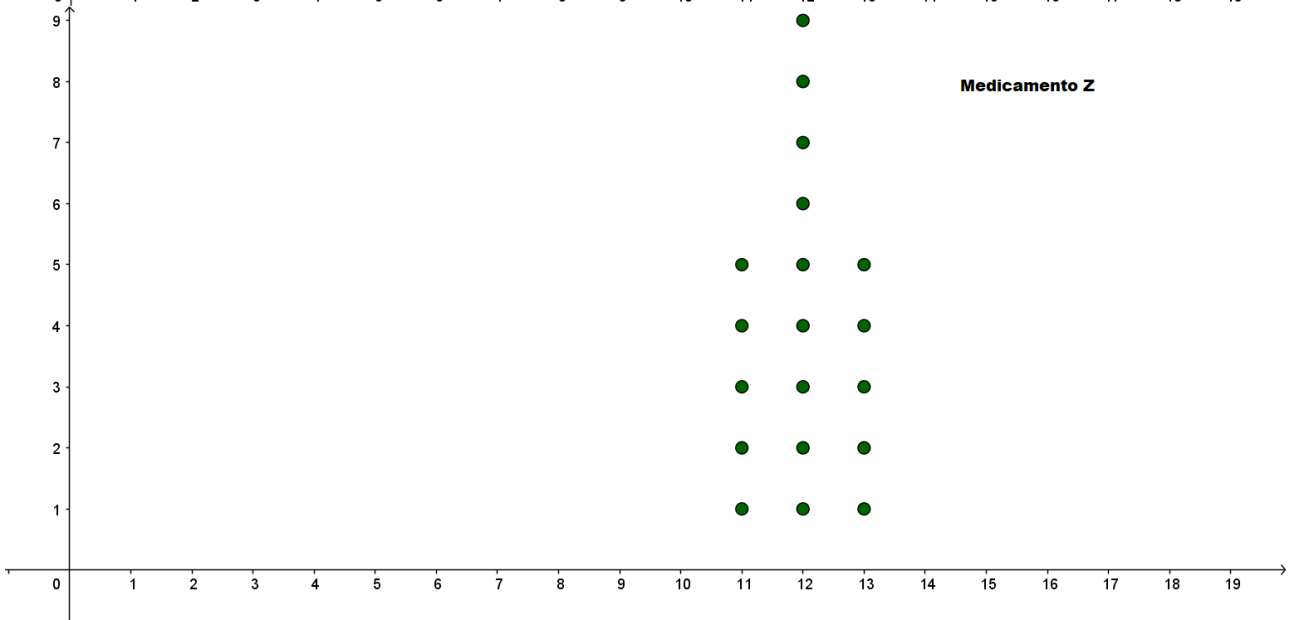
2) Comparação de tratamentos

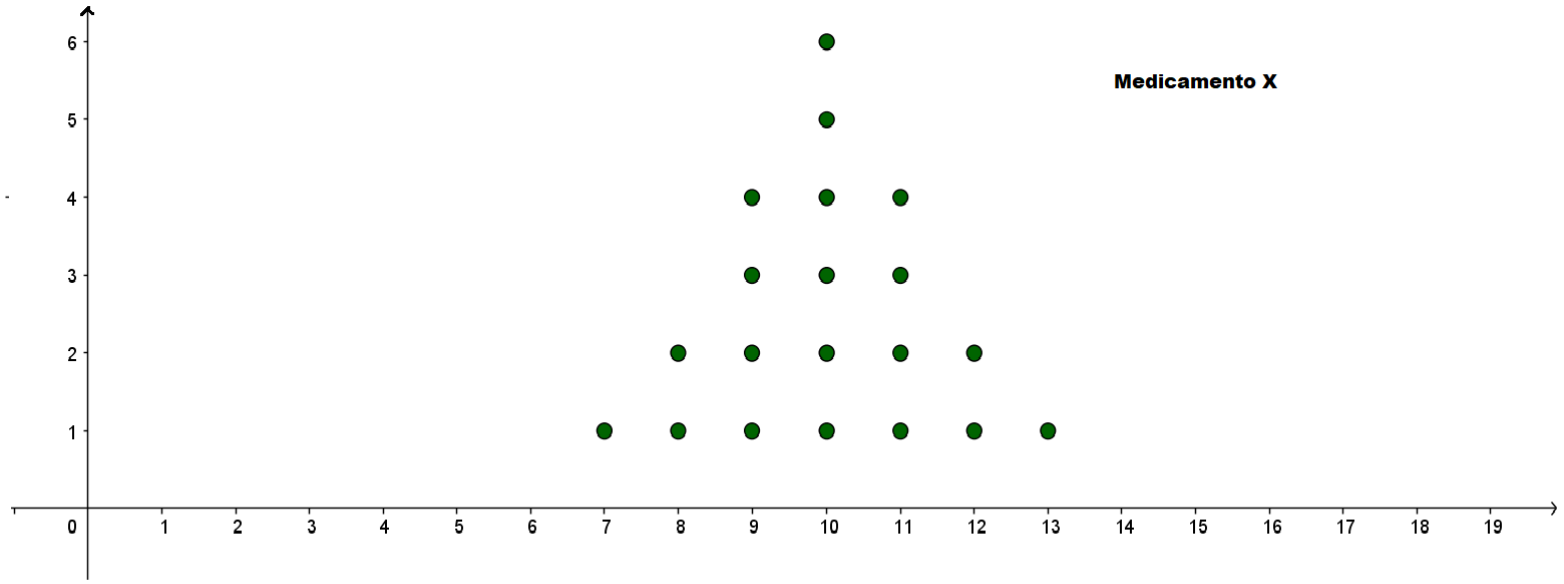
Deseja-se comparar três medicamentos, X, Y e Z, no tratamento da dor de cabeça. Para isso 60 pacientes com perfis similares foram separados aleatoriamente em três grupos de 20 cada. Para cada grupo, será ministrado um dos medicamentos e observado o tempo de cura da dor de cabeça (em minutos). No quadro a seguir estão dispostos os dados obtidos.

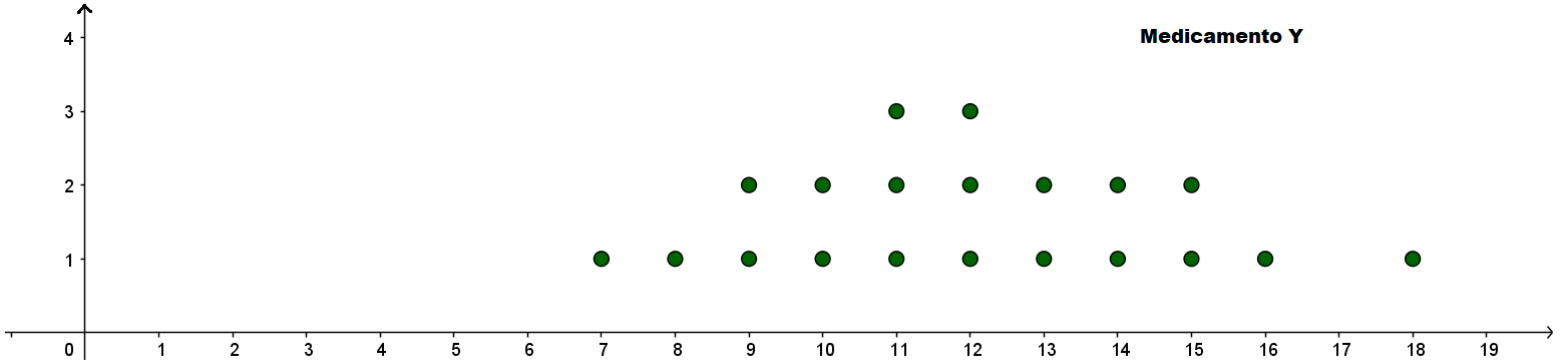
medicamento	tempo (em minutos)																				soma
X	7	8	8	9	9	9	9	10	10	10	10	10	10	11	11	11	11	12	12	13	200
Y	7	8	9	9	10	10	11	11	11	12	12	12	13	13	14	14	15	15	16	18	240
Z	11	11	11	11	11	12	12	12	12	12	12	12	12	12	12	13	13	13	13	13	240

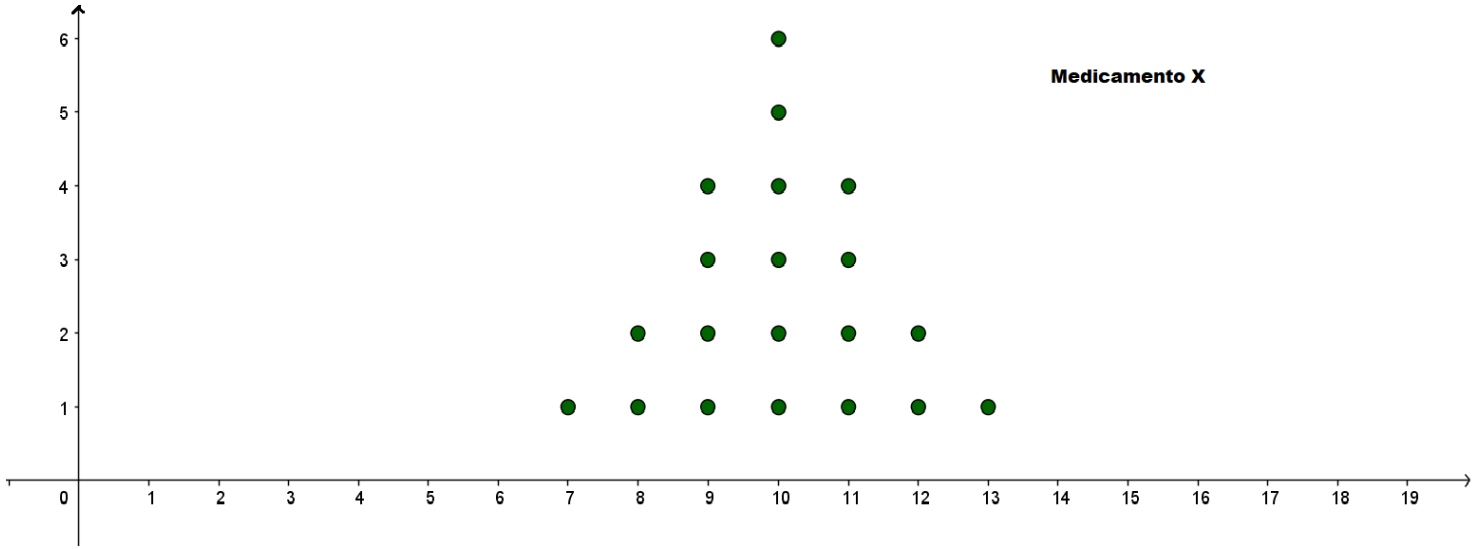

Medicamento X

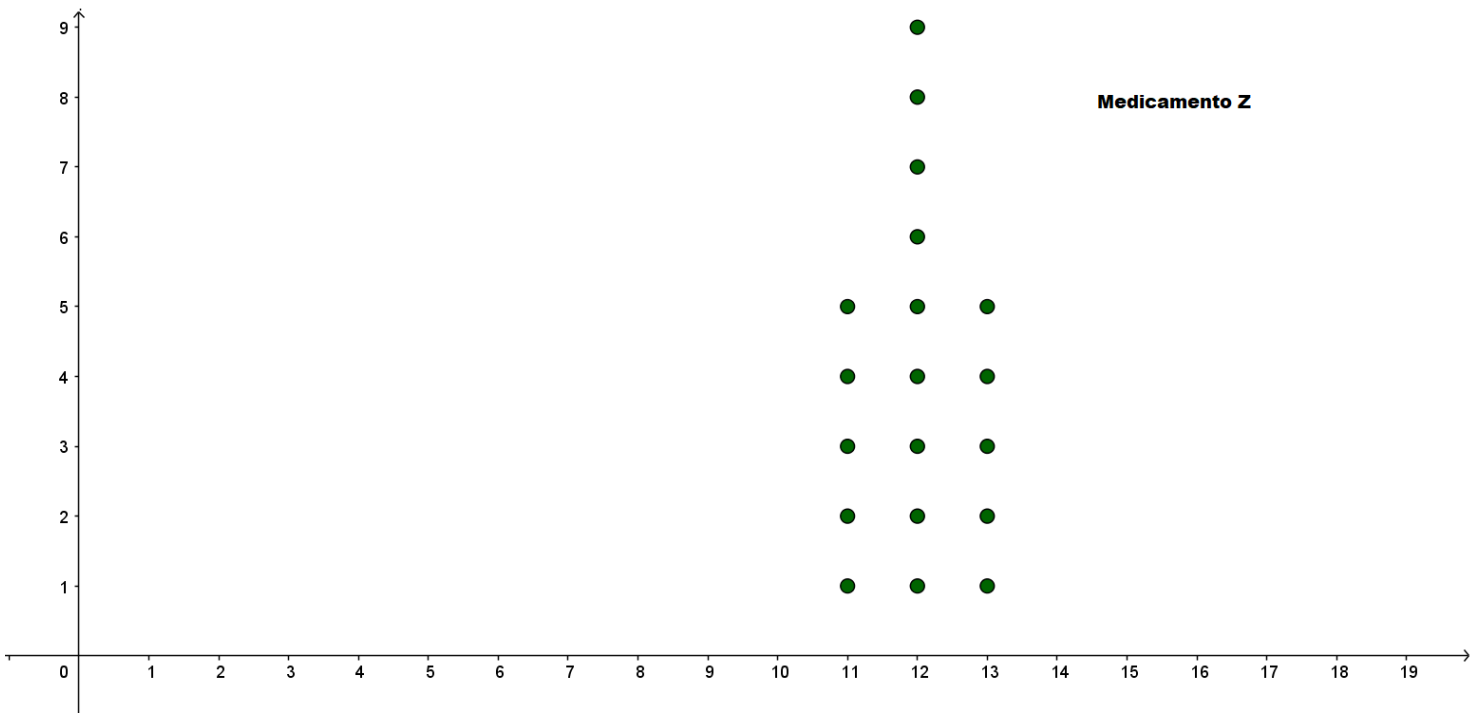

Medicamento Y


Medicamento Z


Medicamento X
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
1 2 3 4 5 6


Medicamento Y
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
1 2 3 4


Medicamento X
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
1 2 3 4 5 6


Medicamento Z
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
1 2 3 4 5 6 7 8 9

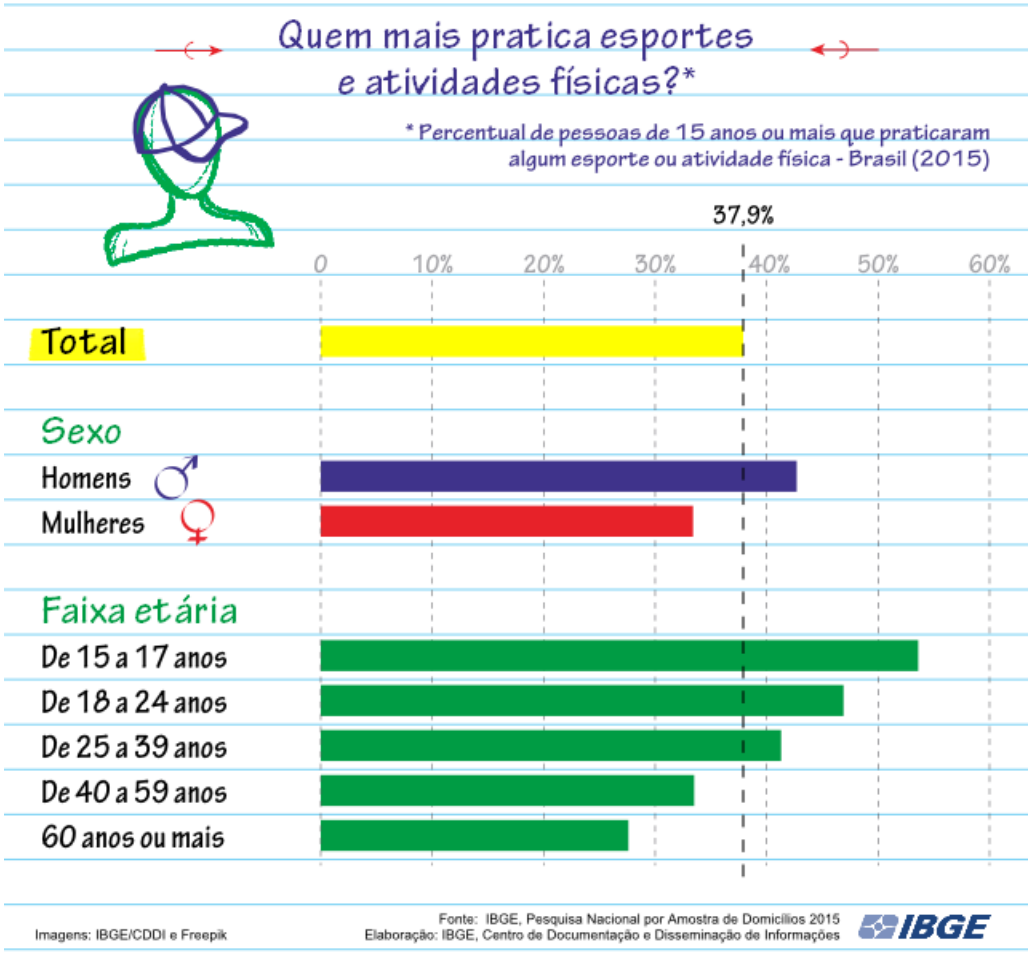
3) Pesquisa sobre a prática de esportes e atividade física (Fonte: IBGE, Suplemento da PNAD/2015, https://vamoscontar.ibge.gov.br/)

- A Pesquisa Nacional por Amostra de Domicílios (PNAD), realizada pelo IBGE obtém informações anuais sobre características demográficas e socioeconômicas da população. Com periodicidade variável, a PNAD obtém informações sobre migração, fecundidade, entre outras, tendo os domicílios como unidade de coleta da informação. Temas específicos abrangendo aspectos demográficos, sociais e econômicos também são investigados.
- A seleção da amostra é cuidadosamente planejada de modo que seja adequado estender os resultados obtidos para a população.
- Quando dispomos de dados provenientes de um subconjunto da população (amostra) sempre podemos descrever os dados nos restringindo apenas ao subconjunto. Se quisermos estender nossas conclusões para a população, será necessário o uso de outras tecnologias que permitam calcular as incertezas associadas a essas extensões.

3) Pesquisa sobre a prática de esportes e atividade física (Fonte: IBGE, Suplemento da PNAD/2015)

- Na PNAD 2015 foi realizada a investigação de um tema específico chamado "Suplemento de Práticas de Esporte e Atividade Física" no qual foram investigadas as pessoas moradoras de 15 anos ou mais de idade, em seu tempo livre, no período de referência de 365 dias. As informações levantadas nessa pesquisa foram obtidas por meio de um questionário no qual se perguntou:
- se a pessoa moradora havia praticado esporte, e em caso afirmativo, a respectiva modalidade.
- Independente da resposta anterior, também se perguntou se a pessoa praticava alguma atividade física que não considerava como esporte, informando, em caso positivo, também a modalidade.
- Outras informações levantadas nessa pesquisa foram: motivação para a prática da atividade física, local onde é praticada a atividade, frequência na qual a atividade é praticada, duração da atividade; e a participação em competições.

Análise de Infográfico



voltar

Conceitos básicos: população e amostra

- O conceito de **população** em Estatística é mais abrangente: é o conjunto de todos os elementos com pelo menos uma característica em comum. Observe que é exatamente essa característica em comum que vai definir o universo (a população) de uma pesquisa.
- **Amostra**: é um subconjunto não vazio da população.

Etapas da análise estatística

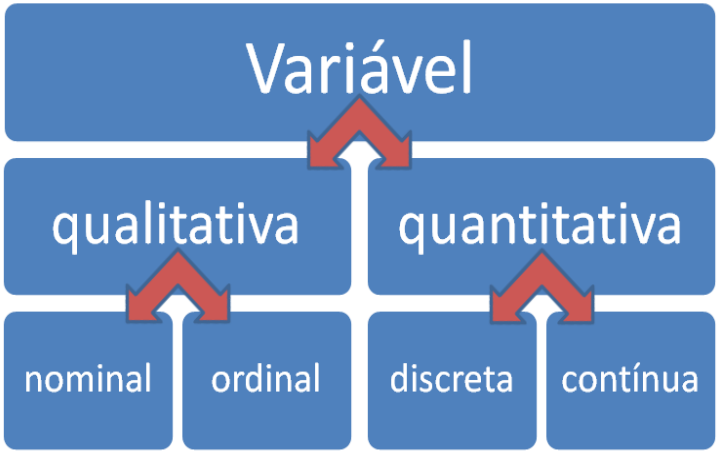
- **Estatística Descritiva**: corresponde a uma exploração das informações que podem ser retiradas dos dados amostrais de modo a reconhecer estruturas que possibilitem futuramente inferir sobre parâmetros de interesse.
- **Estatística Inferencial:** consiste em estabelecer modelos probabilísticos para que se possa fazer afirmações sobre a população com algum nível de confiança.
- Quando se realiza um censo, levantamento de dados de toda a população, não existe a necessidade de fazer uma inferência estatística. No entanto, muitas vezes a realização de um censo é inviável, por várias razões como custo muito alto, tempo muito longo, entre outras.

Conceitos básicos: parâmetro e estimador

- **Parâmetro**: é uma característica numérica da população.

- **Estimador**: é uma função que produz estimativas de parâmetros usando as variáveis observadas na amostra.

Classificação de variáveis



As técnicas e procedimentos estatísticos de análise de dados dependem do tipo de variável investigado. Nesse sentido é importante reconhecer a natureza de cada variável investigada para posterior tratamento da informação obtida.

Observações sobre a classificação de variáveis

- Uma variável quantitativa pode ser tratada como qualitativa, por exemplo, a idade trabalhada em faixas etárias torna-se uma variável qualitativa ordinal.
- Por outro lado, também podemos transformar uma variável qualitativa em quantitativa. Considere a variável "prática de atividades físicas" que tem como respostas "Sim" ou "Não". Esse tipo de variável com apenas duas respostas é chamado variável binária e tem uma representação numérica natural. Podemos atribuir o número 1 para a resposta "Sim" e o número 0 para a resposta "Não". Essa estratégia permite somar todas as respostas. Observe que a soma representará o número de pessoas na amostra que praticam atividade física e a "média" representará a proporção de pessoas na amostra que praticam atividade física.

Gráficos para variáveis qualitativas

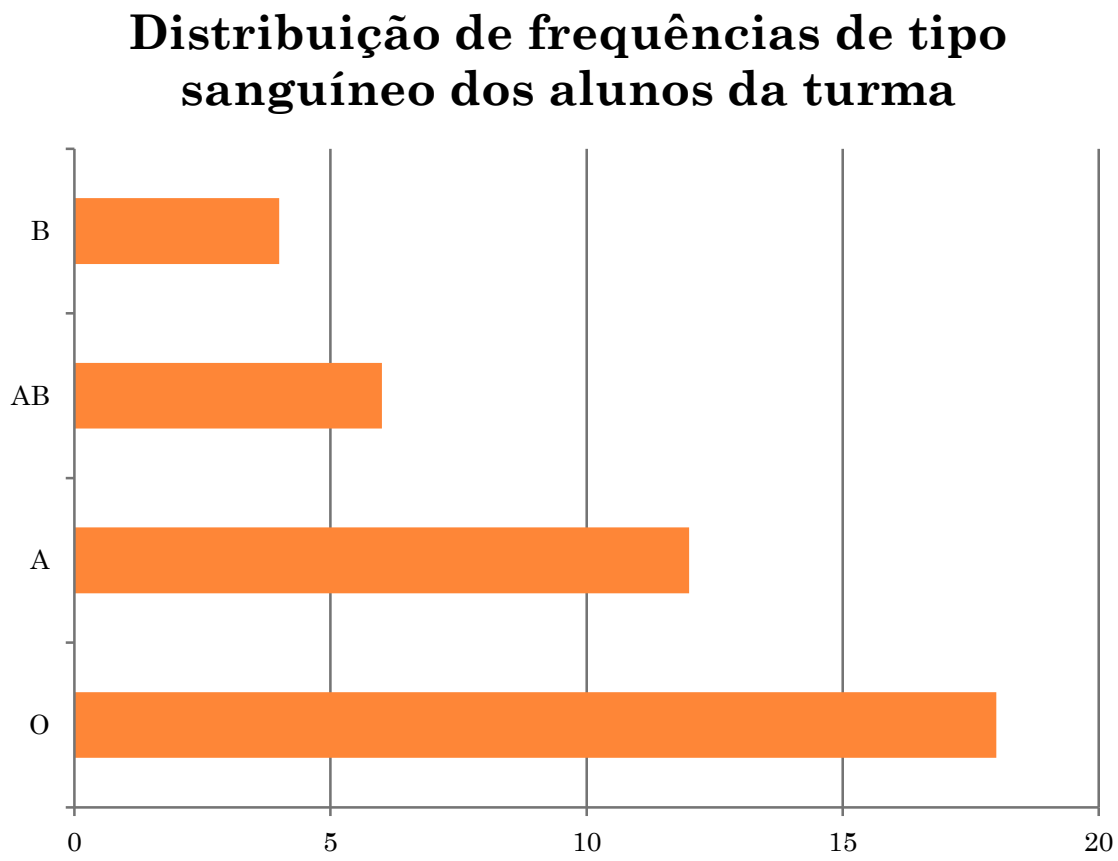
- A primeira etapa na descrição de um conjunto de informações de uma variável qualitativa é organizar as respostas obtidas em função das frequências nas quais elas ocorreram. Essas informações de frequências podem ser representadas pela frequência absoluta (contagem de casos), frequência relativa ou porcentagem (contagem de casos sobre o número total de observações).
- Para as variáveis qualitativas, o gráfico mais comum é o **gráfico de barras** em que as barras com larguras iguais são igualmente espaçadas e, seus comprimentos, são dados pelas respectivas frequências, cada barra representando uma resposta da variável.

Exemplo: Tipo sanguíneo

Numa turma de um colégio foram observados os tipos sanguíneos de seus 40 alunos. Verificou-se que 18 alunos têm sangue tipo “O”, 12, tipo “A”, 6, tipo “AB” e 4, tipo “B”. Nesse exemplo, temos que as frequências absolutas para os tipos sanguíneos “O”, “A”, “AB” e “B” foram, respectivamente, 18, 12, 6 e 4.

tipo sanguíneo	frequência absoluta	frequência relativa	porcentagem (%)
O	18	0,45	45
A	12	0,30	30
AB	6	0,15	15
B	4	0,10	10
total	40	1,00	100

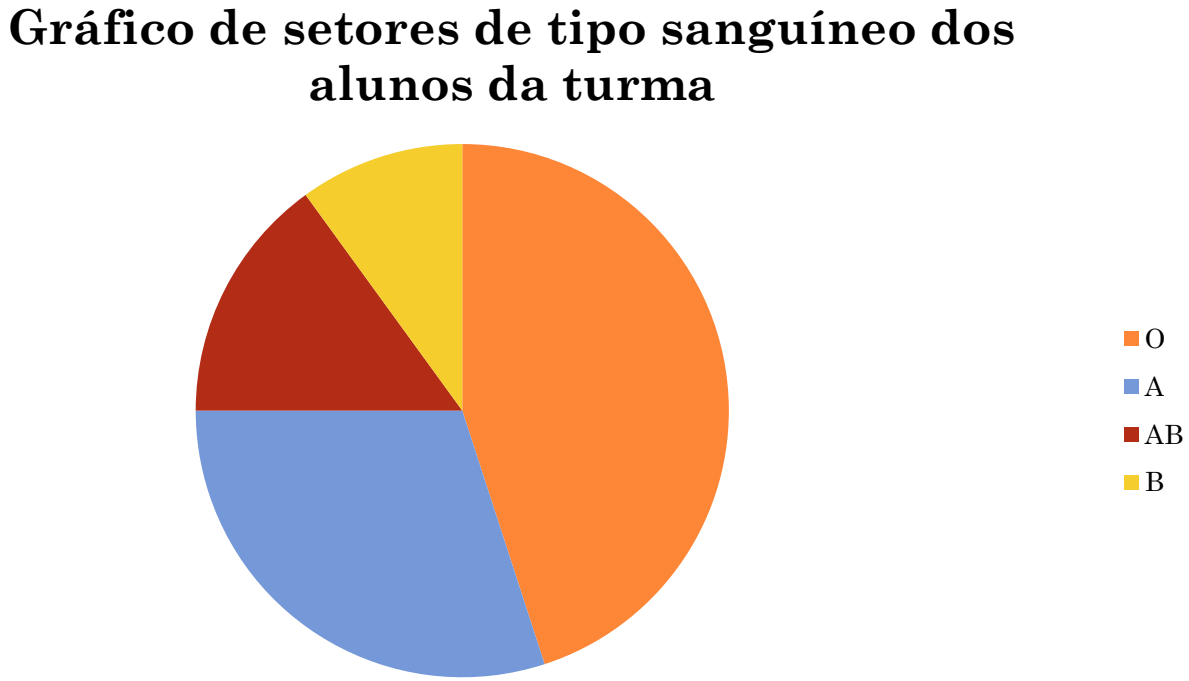
Exemplo: Gráfico de Barras



Gráficos para variáveis qualitativas

- Gráficos de área também são usados para variáveis qualitativas e, o mais comum, é o **gráfico de setores**. Cada setor representa uma resposta da variável qualitativa em que a medida da área do setor em relação à medida da área do círculo é igual à frequência relativa na qual a resposta ocorreu.
- Outra possibilidade de gráfico de área é o **gráfico de retângulos** no qual o retângulo maior é subdividido em retângulos cujas áreas relativas correspondem às porcentagens das respostas que eles representam.

Exemplo: Gráfico de Setores



Exemplo: Gráfico de Retângulos

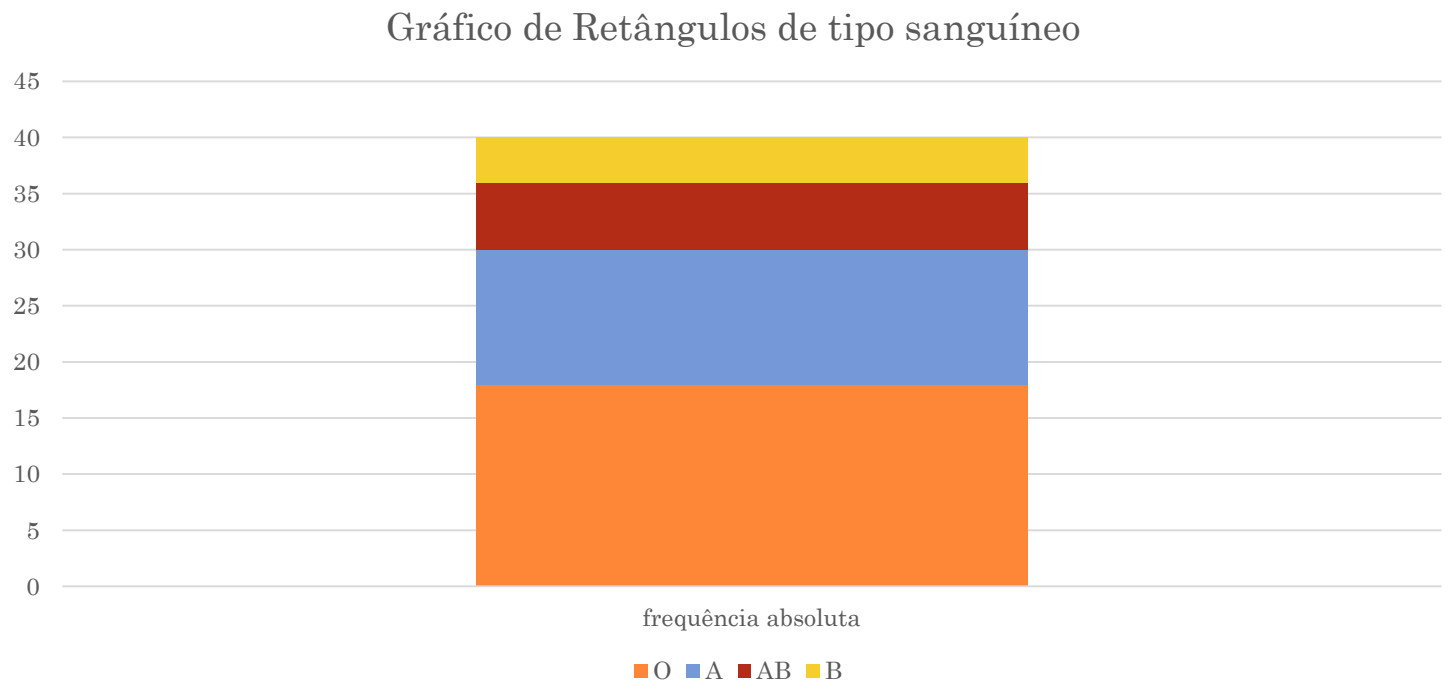


Gráfico de Barras versus Gráfico de Setores

- Se o número de respostas diferentes é grande, maior que 4, ou se as diferenças nas frequências das respostas são pequenas, por exemplo uma tem porcentagem 22% e a outra tem porcentagem 25%, o gráfico de setores não será adequado, pois pequenas diferenças de ângulos não são perceptíveis, enquanto que no gráfico de barras é fácil perceber pequenas diferenças de comprimento das barras.

EXEMPLOS

tipo sanguíneo	frequência absoluta
O	11
A	12
AB	13
B	5
total	40

Gráfico de setores: distribuição de frequências de de tipo sanguíneo

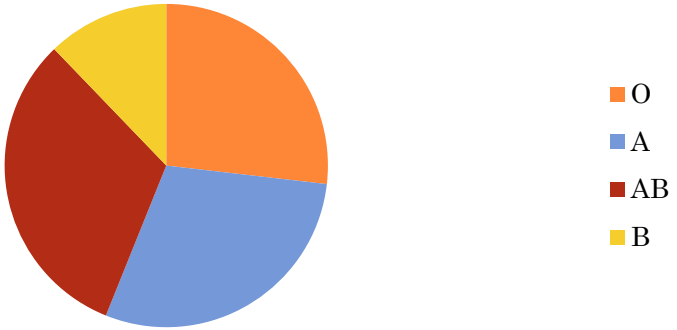
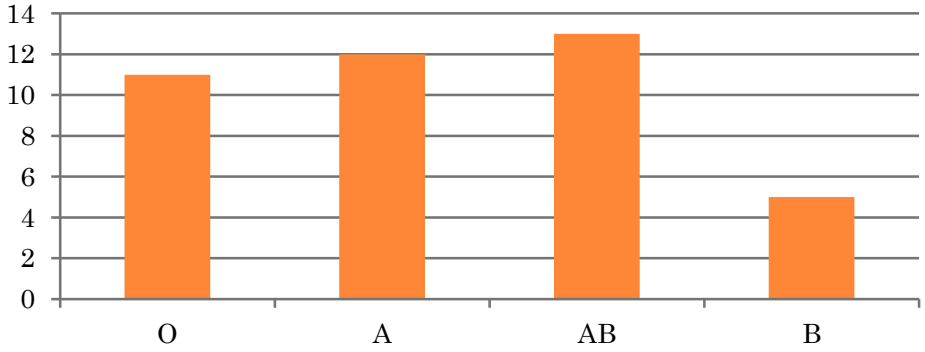
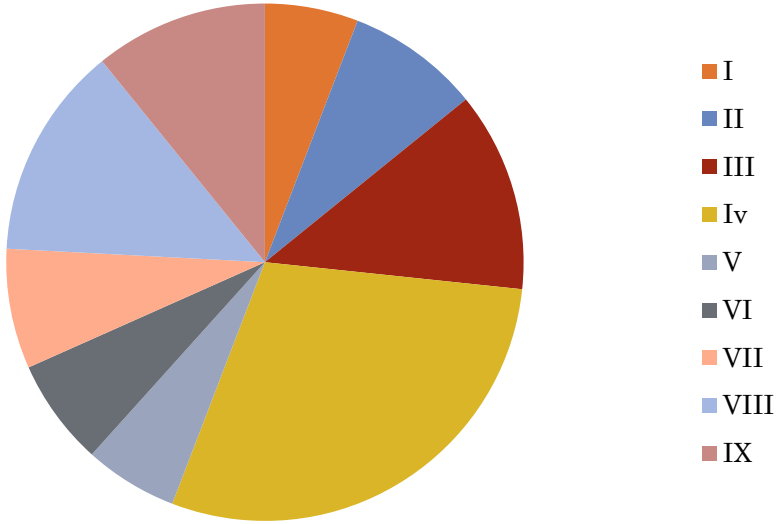


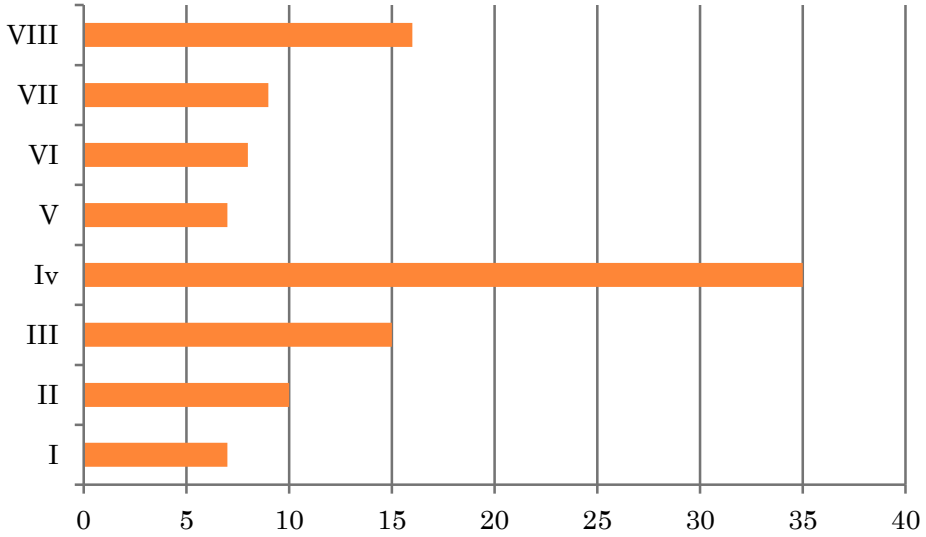
Gráfico de barras: distribuição de frequências de tipo sanguíneo



Exemplos

Classificação	frequência
I	7
II	10
III	15
IV	35
V	7
VI	8
VII	9
VIII	16
IX	13

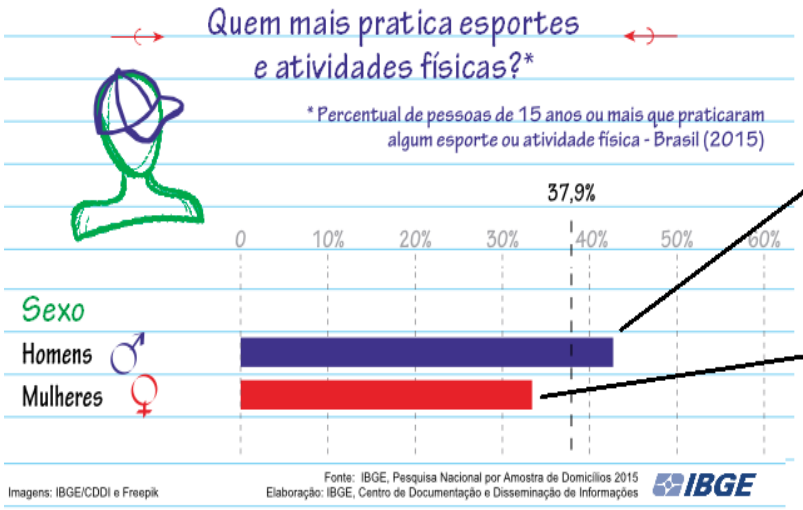




EXEMPLOS: COMPARAÇÕES

Para fazer comparações múltiplas, o gráfico de setores não é adequado. Observe que no infográfico I, os gráficos separados por sexo e por faixa etária, são gráficos de barras múltiplas.

Exemplo de gráfico de barras múltiplas



Pouco mais de 40% dos homens de 15 anos ou mais pratica atividade física, o complemento, pouco menos de 60% dos homens, não pratica.

Pouco mais de 30% das mulheres de 15 anos ou mais pratica atividade física, o complemento, pouco menos de 70% das mulheres não pratica.

Gráficos para variáveis quantitativas: Exemplo

- Um arranjo de oito radiotelescópios (A, B, C, D, E, F, G e H) como ilustrado na figura detectou sinais cujos oito registros de tempo para cada radiotelescópio se encontram no quadro a seguir.

A	B	C	D	E	F	G	H
3,03	4,37	5,04	5,73	4,03	5,37	6,04	6,74
3,38	4,46	5,11	5,84	4,38	5,46	6,11	6,84
3,69	4,55	5,19	5,95	4,60	5,55	6,19	6,96
3,78	4,63	5,29	6,08	4,78	5,64	6,29	7,08
3,92	4,71	5,36	6,23	4,92	5,72	6,36	7,23
4,04	4,79	5,45	6,41	5,04	5,79	6,45	7,40
4,16	4,87	5,54	6,62	5,16	5,87	6,54	7,63
4,27	4,95	5,64	6,97	5,26	5,95	6,64	7,97

Gráficos para variáveis quantitativas: **Exemplo**

- Como construir uma tabela de frequências desses dados uma vez que os registros de tempo são todos distintos? Como você faria para visualizar o comportamento de uma variável com estas características?
- A natureza quantitativa de uma variável contínua pode muitas vezes levar a resultados que praticamente não se repetem. Eles podem ser todos diferentes, como é observado no exemplo.
- Com o objetivo de identificar alguma estrutura no comportamento deste tipo de variável é necessário agrupar os valores em intervalos de classe.

Quantos intervalos de classe considerar no agrupamento dos dados?

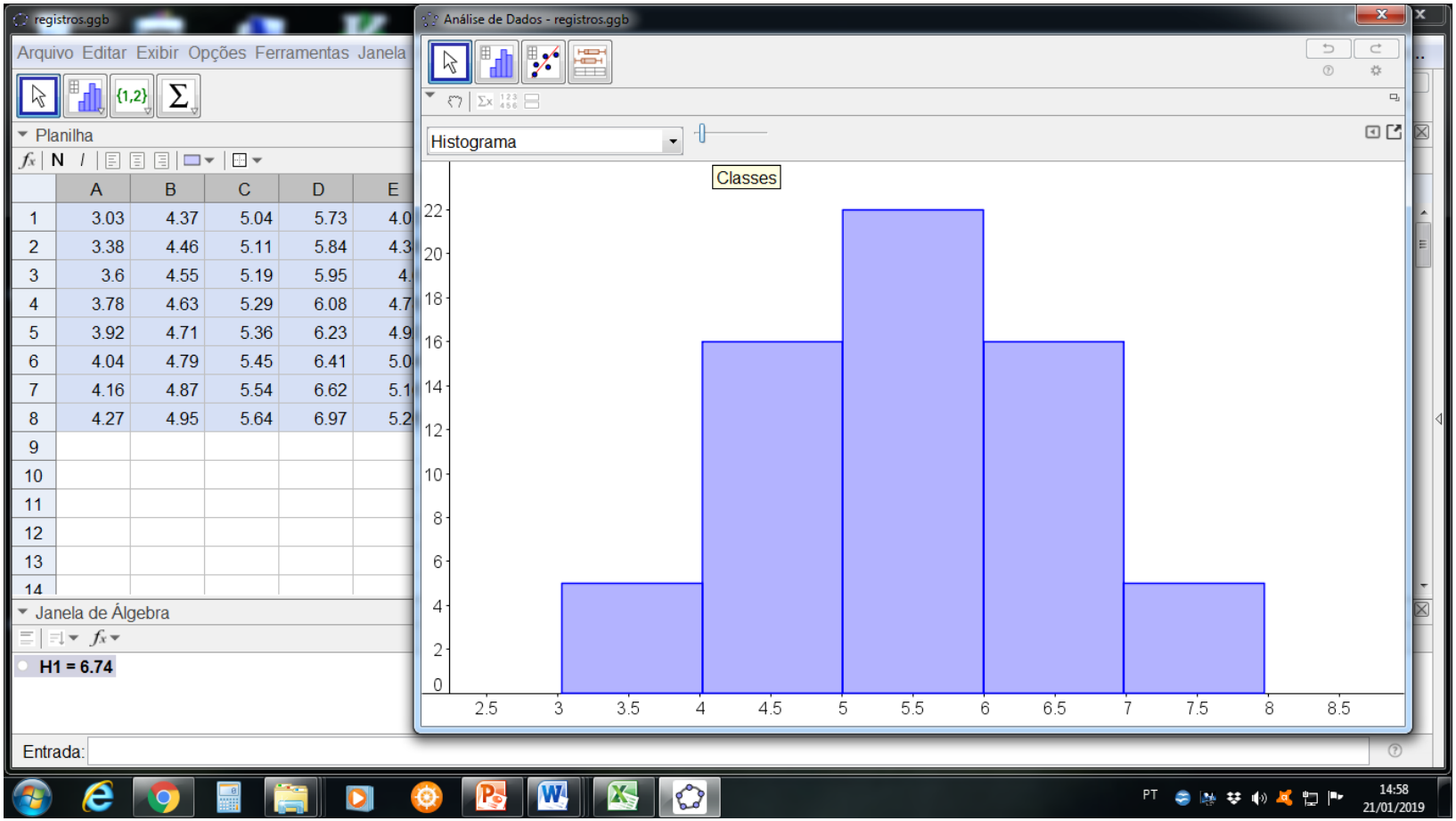
Não existe uma única resposta para essa questão. No entanto, devemos evitar tanto usar um número reduzido de intervalos, quanto usar um número grande de intervalos.

Por exemplo, se usarmos um único intervalo, o histograma seria representado por um único retângulo que nada informaria sobre o comportamento dos dados. Por outro lado, se o número de intervalos for igual ou superior ao número de observações, o histograma potencialmente teria apenas classes com uma única observação e o objetivo de visualizar estruturas dos dados em análise se perderia.

Quantos intervalos de classe considerar no agrupamento dos dados?

Embora não exista uma resposta única sobre quantos intervalos considerar, alguns autores sugerem usar o número inteiro mais próximo da raiz quadrada do número de observações, outros sugerem usar de 5 a 15 intervalos de amplitudes iguais. No GeoGebra, por exemplo, a função que constrói histogramas permite trabalhar com 3 a 20 intervalos.

GeoGebra: Construção do Histograma



	A	B	C	D	E
1	3.03	4.37	5.04	5.73	4.0
2	3.38	4.46	5.11	5.84	4.3
3	3.6	4.55	5.19	5.95	4.
4	3.78	4.63	5.29	6.08	4.7
5	3.92	4.71	5.36	6.23	4.9
6	4.04	4.79	5.45	6.41	5.0
7	4.16	4.87	5.54	6.62	5.1
8	4.27	4.95	5.64	6.97	5.2
9					
10					
11					
12					
13					
14					

Histograma

- Representação gráfica da distribuição de frequências de uma variável quantitativa contínua agrupada em intervalos, usando retângulos adjacentes.
- Cada retângulo no histograma corresponde a um intervalo considerado e a razão da área desse retângulo em relação à área total do histograma deve ser igual à frequência relativa de casos desse intervalo.

Histograma

- Quando os intervalos têm **amplitudes iguais**, pode-se trabalhar tanto na escala das frequências como na escala das densidades de frequência (frequência sobre amplitude).
- Quando os intervalos têm **amplitudes desiguais**, a escala a ser usada deve ser a das densidades de frequência (relativa ou absoluta), para que a estrutura resultante não seja distorcida, ou seja para que a razão da área de cada retângulo em relação à área total do histograma seja igual à frequência relativa de casos desse intervalo.

Histograma versus Gráfico de barras

- O gráfico de barras não é um histograma, apesar de suas representações serem parecidas.
- Os gráficos de barras são úteis para representar distribuição de frequências de variável qualitativa. Nesse gráfico só há um eixo com escala que corresponde aos valores das frequências das categorias (respostas) da variável. As barras podem ser tanto verticais como horizontais e são apresentadas de forma igualmente espaçada. Cada barra representa uma resposta da variável qualitativa e a altura da barra corresponde à frequência daquela resposta.

Histograma versus Gráfico de barras

- O gráfico de barras também pode ser usado para representar uma variável quantitativa discreta, sendo que nesse caso, as posições das barras correspondem aos valores assumidos pela variável. Pela natureza discreta da variável, as barras não são adjacentes e, pela natureza quantitativa da variável, o posicionamento das barras não é livre.
- Os histogramas são úteis para representar a distribuição de frequências de uma variável quantitativa contínua cujos valores foram agrupados em intervalos. No histograma, o eixo das abscissas (horizontal) representa a escala da variável contínua e, o eixo das ordenadas (vertical) representa a escala da frequência ou densidade de frequência que é definida como a razão entre a frequência e a amplitude do intervalo.

Gráfico de linha

- Representação útil quando os dados são quantitativos e coletados ao longo do tempo (série temporal).
- Esse gráfico é construído marcando-se no plano Cartesiano os pontos (x,y) em que abscissa x representa o tempo e, a ordenada y, a variável quantitativa. Os pontos consecutivos são unidos por segmentos.

GRÁFICO DE LINHA: EXEMPLO

- O quadro a seguir fornece a média das temperaturas máximas para cada mês nos anos de 1991 a 2000 da cidade de Porto Alegre em graus centígrados.
- Fonte: http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep

Médias das temperaturas máximas em Porto Alegre: 1991 a 2000

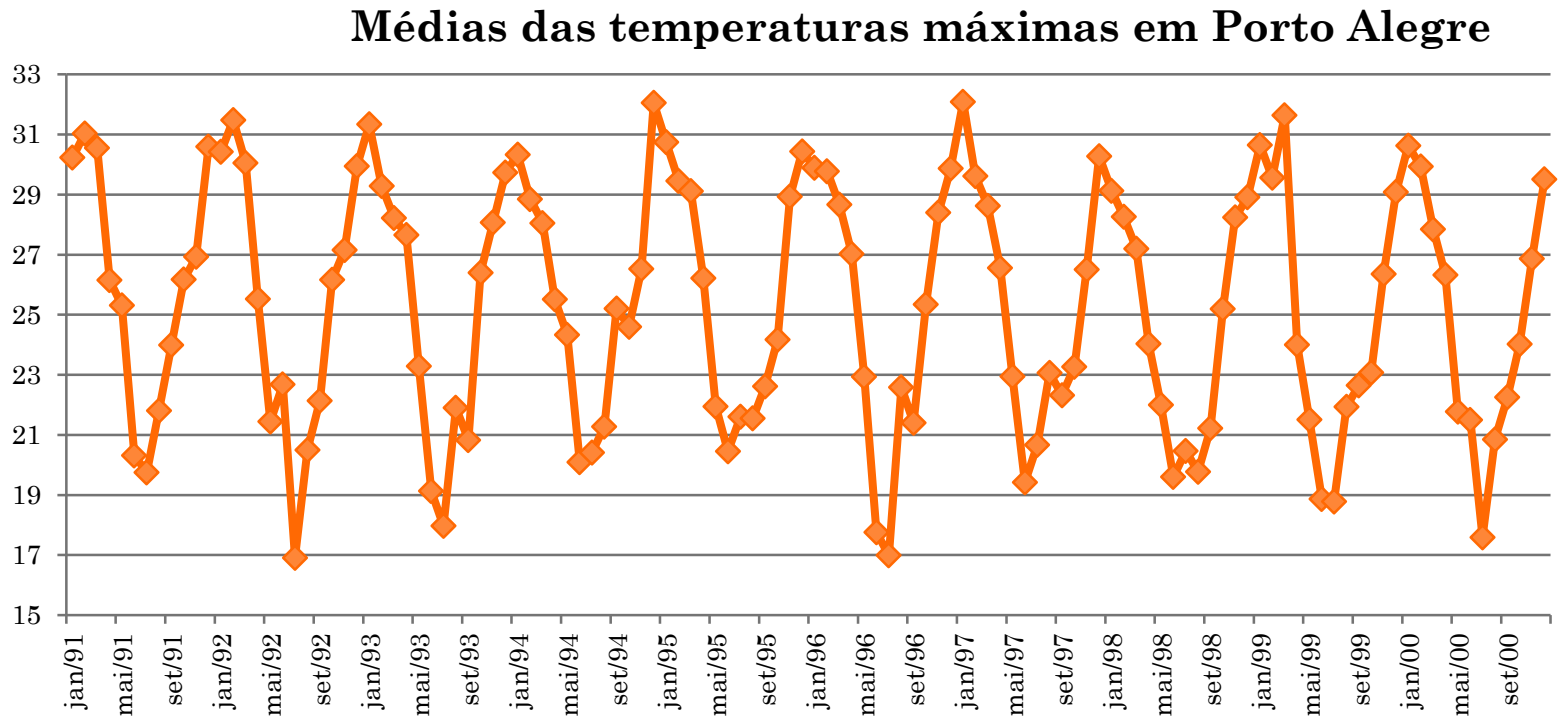
Mês	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
1	30,23	30,43	31,34	30,33	30,74	29,89	32,09	29,13	30,65	30,63
2	31,03	31,48	29,28	28,85	29,46	29,78	29,62	28,26	29,56	29,93
3	30,55	30,05	28,22	28,05	29,12	28,67	28,63	27,20	31,64	27,85
4	26,15	25,52	27,66	25,51	26,22	27,03	26,56	24,03	24,00	26,32
5	25,31	21,44	23,29	24,33	21,95	22,94	22,95	22,00	21,51	21,78
6	20,32	22,68	19,13	20,09	20,45	17,76	19,42	19,60	18,87	21,50
7	19,75	16,91	17,97	20,41	21,60	16,99	20,67	20,47	18,78	17,59
8	21,81	20,50	21,90	21,28	21,55	22,59	23,06	19,77	21,94	20,85
9	23,99	22,14	20,83	25,21	22,62	21,40	22,32	21,22	22,65	22,25
10	26,17	26,16	26,40	24,60	24,17	25,34	23,27	25,19	23,07	24,02
11	26,93	27,16	28,07	26,53	28,93	28,40	26,51	28,24	26,36	26,87
12	30,60	29,95	29,73	32,05	30,44	29,87	30,28	28,91	29,08	29,51

Gráfico de linha

- No link a seguir é possível construir o gráfico de temperaturas médias para cada ano:

https://goo.gl/ceZfvh

Também é possível obter o gráfico ao longo dos 10 anos.


Médias das temperaturas máximas em Porto Alegre
33
31
29
27
25
23
21
19
17
15
jan/91
mai/91
set/91
jan/92
mai/92
set/92
jan/93
mai/93
set/93
jan/94
mai/94
set/94
jan/95
mai/95
set/95
jan/96
mai/96
set/96
jan/97
mai/97
set/97
jan/98
mai/98
set/98
jan/99
mai/99
set/99
jan/00
mai/00
set/00

Noções básicas sobre seleção de amostras

- Quando queremos estender nossas observações provenientes de uma amostra para a população é importante ter cuidado na sua seleção, pois ela deve ser representativa da população. Os métodos de seleção de amostras podem ser probabilísticos ou não probabilísticos.
- O primeiro tipo é fundamental para que seja possível avaliar a incerteza das conclusões devido à amostragem tais como margem erro e nível de confiança. Nesse tipo de seleção de amostra, conhecemos a probabilidade de seleção dos elementos da população na amostra.

Noções básicas sobre seleção de amostras

Entre os métodos probabilísticos mais comuns destacam-se

- **amostragem aleatória simples**: todas as amostras de igual tamanho têm probabilidades iguais de serem selecionadas.
- **amostragem estratificada**: a população é dividida em grupos de elementos homogêneos (similares nas características a serem investigadas) e os grupos são heterogêneos entre si. A amostra é composta por amostras aleatórias simples de cada grupo, em geral, proporcionalmente aos tamanhos dos grupos.
- **amostragem por conglomerados**: a população é subdividida em conglomerados (subpopulações). Uma amostra aleatória simples de conglomerados é obtida e, em seguida, todos os elementos dos conglomerados escolhidos são observados.

NOÇÕES BÁSICAS SOBRE SELEÇÃO DE AMOSTRAS

Os casos mais comuns de **métodos não probabilísticos** são

- **amostragem por conveniência -** caracteriza-se por não ter um plano particular de amostragem. O objetivo nesse caso não seria generalizar conclusões e sim descrever as características principais do grupo de estudo.
- **amostragem por julgamento** - os elementos da amostra são escolhidos por um especialista no assunto sob investigação.
- A desvantagem dos métodos não probabilísticos está na impossibilidade de avaliar incertezas devido à amostragem.

Cenas do próximo capítulo

Após organizar os dados em tabelas e gráficos, os próximos passos envolvem resumir a informação obtida por meio de algumas medidas. Por exemplo, a partir de um conjunto de dados quantitativos pretende-se responder as seguintes questões:

- É possível encontrar valor(es) para resumir as observações? Qual(is) seria(m) este(s) valor(es)? Como encontrá-lo(s)?
- Como medir se os dados estão "próximos" ou "afastados" uns dos outros?
- Como você classifica a forma do gráfico construído para representar os dados?
- Existe algum valor muito diferente dos demais? Como identificá-lo?

As respostas a essas questões são tratadas no segundo capítulo de Estatística do Projeto Livro Aberto de Matemática, “Medidas de Posição e Dispersão”.