

Introdução à análise exploratória de dados

Luan D. Fiorentin

Universidade Federal do Paraná
Departamento de Estatística
Laboratório de Estatística e Geoinformação

13/08/2019



Sumário

1 Ideias gerais

2 Amostragem

- Tipos de amostragem.
- Métodos de amostragem.
- Erros na amostragem.

3 Análise exploratória de dados

- Organização dos dados.
- Tabelas de frequência.
- Representação gráfica.

4 Exercícios recomendados

O que é estatística?

- Estatística é um conjunto de técnicas para, sistematicamente:
 - Planejar a coleta de dados oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento;
 - Descrever, analisar e interpretar dados;
 - Extrair informações para subsidiar decisões;
 - Avaliar evidências empíricas sob hipóteses de interesse.
- Exemplos de aplicações:
 - Opinião da população brasileira sobre o novo governo.
 - Avaliar a efetividade de uma nova droga para a cura do câncer.
 - Entender os hábitos de compra dos clientes de uma loja virtual.
 - Recomendação personalizada de produtos.
 - Comparar a produtividade da soja sob diferentes formas de cultivo, adubação, etc.

Divisões básicas da estatística

- Divisões essenciais em Estatística e seus principais objetivos.
 - **Estatística descritiva ou exploratória:**
 - Consistência dos dados e interpretações iniciais.
 - Visualização dos dados e relações entre variáveis.
 - **Probabilidade:**
 - Fornece ferramentas para lidar/quantificar incerteza.
 - **Inferência estatística:**
 - Estimação de quantidades desconhecidas.
 - Formular e testar hipóteses.
 - Extrapolar para a população resultados obtidos na amostra.

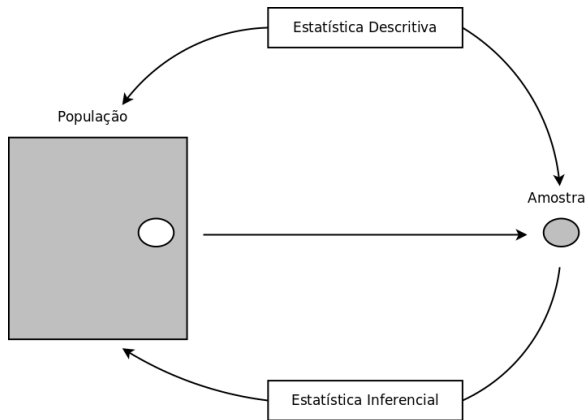
População e amostra

- Conceitos fundamentais:
 - **População:** Conjunto de todos os elementos sob investigação.
 - **Amostra:** Subconjunto da população.
 - **Variável** de interesse: característica a ser observada em cada indivíduo da amostra.

Exemplos em detalhes

- Opinião da população brasileira sobre o novo governo.
 - **População:** Todos os habitantes do Brasil? Outras opções?
 - **Amostra:** Algum subconjunto da população. Qualquer um será? Como selecionar?
 - **Variável de interesse:** Opinião sobre o novo governo. Como medir isso? Gosta? sim ou não.
- Avaliar a efetividade de uma nova droga para a cura do câncer.
 - **População:** Todos os seres humanos? Apenas os já doentes? Como levar em conta questões de raça, culturas, etc ...
 - **Amostra:** E agora?
 - **Variável de interesse:** Curou ou não curou? Será que isso é possível?
- Entender os hábitos de compra dos clientes de uma loja virtual.
 - **População:** Todos os clientes da loja virtual.
 - **Amostra:** Preciso de amostra?
 - **Variável de interesse:** E agora? Como caracterizar hábito de compra?

Ideia final



Etapas da análise estatística

- Definir a **população de interesse**.
 - População factível.
- Estabelecer os objetivos (questões) de pesquisa.
 - Definir critérios objetivos sobre quais dados coletar.
 - Postular a análise estatística a ser utilizada.
- Definir o método para coletar as amostras.
 - Fonte de dados secundários (IBGE, IPEA, etc);
 - Banco de dados da empresa;
 - Pesquisas amostrais;
 - Experimentos em laboratórios, etc.
- Análise dos dados.
 - Análise descritiva e exploratória (o que aconteceu na amostra?).
 - Análise inferencial (o que acontece na população?).

Planejamento da coleta de dados

- Definição do experimento.
 - Variáveis respostas/interesse.
 - Variáveis de controle (o que afeta a resposta?).
 - Desenho do experimento e randomização.

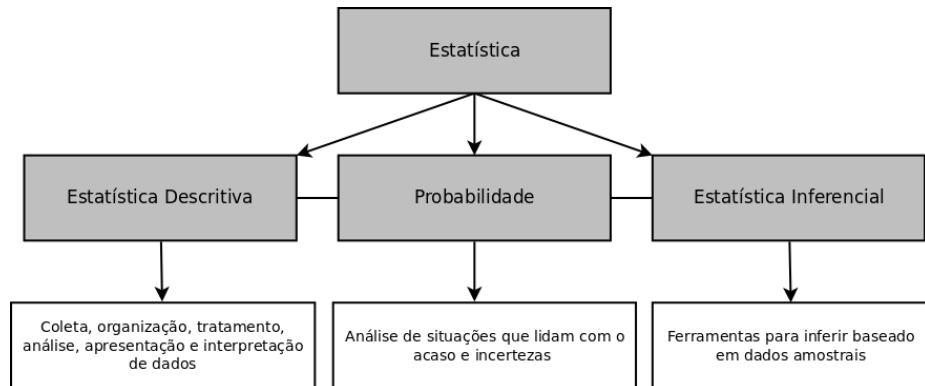
Planejamento da coleta de dados

- Definição do experimento.
 - Variáveis respostas/interesse.
 - Variáveis de controle (o que afeta a resposta?).
 - Desenho do experimento e randomização.
- Coleta de dados por amostragem.
 - Definição da população e característica de interesse.
 - Definição do plano amostral:
 - Aleatória simples (com ou sem reposição) ou sistemática;
 - Estratificada, por estratos da população (segundo uma característica);
 - Conglomerados, por grupos de indivíduos da população (subpopulações);
 - Amostragem complexa (combina anteriores).

Planejamento da coleta de dados

- Definição do experimento.
 - Variáveis respostas/interesse.
 - Variáveis de controle (o que afeta a resposta?).
 - Desenho do experimento e randomização.
- Coleta de dados por amostragem.
 - Definição da população e característica de interesse.
 - Definição do plano amostral:
 - Aleatória simples (com ou sem reposição) ou sistemática;
 - Estratificada, por estratos da população (segundo uma característica);
 - Conglomerados, por grupos de indivíduos da população (subpopulações);
 - Amostragem complexa (combina anteriores).
- Coleta de dados observacionais. Exemplos:
 - Presença de seres vivos num ambiente;
 - Fenômenos climáticos; Fluxo de usuários em um website.

Resumo: Objetivos e etapas da análise estatística



**** A forma de coleta dos dados é um tópico em si. No entanto neste curso básico ela será discutida de forma superficial dentro do tópico de análise exploratória.****

Sumário

- 1 Ideias gerais
- 2 Amostragem
 - Tipos de amostragem.
 - Métodos de amostragem.
 - Erros na amostragem.
- 3 Análise exploratória de dados
 - Organização dos dados.
 - Tabelas de frequência.
 - Representação gráfica.
- 4 Exercícios recomendados

Definições de amostragem

- Quando fazemos uma pesquisa, ou utilizamos algum mecanismo para obter informações, um dos objetivos principais é **coletar dados de uma pequena parte** de um grande grupo e aprender então alguma coisa sobre esse grupo maior.

Definições de amostragem

- Quando fazemos uma pesquisa, ou utilizamos algum mecanismo para obter informações, um dos objetivos principais é **coletar dados de uma pequena parte** de um grande grupo e aprender então alguma coisa sobre esse grupo maior.
- **População:** conjunto de indivíduos, objetos ou produtos que contém a característica que temos interesse. Exemplo:
 - Característica: altura dos estudantes da UFPR;
 - População: todos os estudantes da UFPR.
- **Amostra:** subconjunto da população, em geral com dimensão bem menor, que também possui a característica de interesse. Exemplo:
 - Característica: altura dos estudantes da UFPR;
 - Amostra: 100 estudantes selecionados ao acaso.

Definições de amostragem

- **População → Censo → Parâmetro:**

- Uma medida numérica que descreve alguma característica da população, usualmente representada por letras gregas: θ , μ , σ , ...

- **População → Amostra → Estatística:**

- Uma medida numérica que descreve alguma característica da amostra, usualmente denotada pela letra grega do respectivo parâmetro com um acento circunflexo: $\hat{\theta}$, $\hat{\mu}$, $\hat{\sigma}$ ou por letras do alfabeto comum: \bar{x} , s , ...

- **Exemplo:**

- Média Populacional: μ .
- Média Amostral: \bar{x} .

Exemplo

- **População:** Todos os alunos de uma única turma.
- **Característica** de interesse: idade dos alunos em anos.
- **Censo:** 22 21 24 23 20 22 21 25 24 24 23 19 25 24 23 23 20 21 23 20 23 22 23 23 25
 - Média Populacional: $\mu = 22,5 \rightarrow$ Parâmetro
- **Amostra:** 25 24 23 23 25
 - Média Amostral: $\mu = 24,0 \rightarrow$ Estimativa

Por que fazer amostragem?

- Parâmetros populacionais desconhecidos.
- Impossibilidade de realização de um censo.
- Mais barato, mais rápido.
- Importante: **Não existe nenhuma técnica estatística capaz de salvar uma amostra mal coletada!**

Por que fazer amostragem?

- Parâmetros populacionais desconhecidos.
- Impossibilidade de realização de um censo.
- Mais barato, mais rápido.
- Importante: **Não existe nenhuma técnica estatística capaz de salvar uma amostra mal coletada!**
- Em geral, uma amostra deve ser um subconjunto representativo da população aleatória (de alguma forma).
- Existem diversas maneiras para se retirar uma amostra de uma população → Teoria da Amostragem.
- Fonte de dados:
 - Planejamento de experimentos;
 - Levantamentos observacionais;
 - Levantamentos amostrais;

Planejamento de experimentos:

- Aplica um tratamento, e passa a observar seu efeito entre o objeto de estudo.
- Portanto, requer a **interferência do pesquisador sobre a população**, bem como o controle de fatores externos, com o intuito de medir o efeito desejado.
- Exemplos: Estudo do efeito de um novo medicamento, experimentos agrônômicos. . .

Levantamentos observacionais:

- Observa e mede características, mas não modifica o objeto de estudo.
- Os dados são coletados **sem que o pesquisador tenha controle** sobre as informações obtidas.
- Exemplo: Verificar o valor das vendas de uma empresa em um certo período (não há como “selecionar” as vendas), ...

Levantamentos amostrais:

- A amostra é obtida a partir de uma população bem definida, bem meio de processos bem definidos pelo pesquisador.
 - **Probabilísticos:** Cada elemento da população possui a mesma probabilidade de ser selecionado para compor a amostra → mecanismos aleatórios de seleção.
 - **Não Probabilísticos:** A seleção da amostra depende do julgamento do pesquisador. Há uma escolha deliberada dos elementos para compor a amostra → mecanismos não aleatórios de seleção.

Métodos não probabilísticos:

- **Amostragem por conveniência:** elementos selecionados por serem imediatamente disponíveis.
 - Exemplo: Uma repórter entrevistando pessoas na rua.
- **Amostragem por julgamento:** uma pessoa experiente no assunto escolhe intencionalmente os elementos a serem amostrados.
 - Exemplo: Novo produto “testado” entre funcionários.
- **IMPORTANTE:** Na amostragem não probabilística, os elementos da população não tem a mesma probabilidade de serem selecionados, portanto não há garantias da representatividade da população!

Métodos probabilísticos:

- 1 Amostragem Aleatória Simples (AAS).
- 2 Amostragem Sistemática.
- 3 Amostragem Estratificada.
- 4 Amostragem por Conglomerado.

Métodos probabilísticos:

1 Amostragem Aleatória Simples (AAS):

- **Todas as possíveis amostras de tamanho n tem a mesma chance de serem escolhidas** (de uma população com N elementos).
 - Exemplos:
 - Selecionar 10 estudantes de uma sala por sorteio e perguntar a idade.
 - Gerar uma amostra aleatória de 1000 números de matrícula de estudantes da UFPR (no computador!) e perguntar a idade.

Métodos probabilísticos:

1 Amostragem Aleatória Simples (AAS):

- **Todas as possíveis amostras de tamanho n tem a mesma chance de serem escolhidas** (de uma população com N elementos).
 - Exemplos:
 - Selecionar 10 estudantes de uma sala por sorteio e perguntar a idade.
 - Gerar uma amostra aleatória de 1000 números de matrícula de estudantes da UFPR (no computador!) e perguntar a idade.
- É o método mais simples para selecionarmos uma amostra probabilística de uma população.
- Serve de base para outros procedimentos amostrais, planejamento de experimentos e estudos observacionais.
- Utilizando-se um procedimento aleatório, sorteia-se um elemento da população. Repete-se o processo até que sejam sorteadas as n unidades na amostra.

Métodos probabilísticos:

Amostragem Aleatória Simples

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Unidades Amostrais	
	Amostradas
	Não Amostradas

Métodos probabilísticos:

① Amostragem Aleatória Simples (AAS):

- **Com reposição:** o mesmo elemento da população pode ser amostrado mais de uma vez. A probabilidade de seleção não se altera.
- **Sem reposição:** cada elemento da população é amostrado uma única vez. A probabilidade de seleção se altera.
- **Atenção!**
 - Na prática, em populações infinitas (muito grandes), a reposição ou não é irrelevante.

Métodos probabilísticos:

① Amostragem Aleatória Simples (AAS):

- Do ponto de vista da quantidade de informação contida na amostra, a amostragem sem reposição é mais adequada.
- No entanto, a **amostragem com reposição conduz a um tratamento teórico mais simples**, pois ele implica que tenhamos **independência entre as unidades** selecionadas.
- Portanto, na maioria dos casos quando nos referenciarmos a uma AAS, estamos nos referenciando a uma amostragem aleatória simples com reposição.

Métodos probabilísticos

2 Amostragem Sistemática:

- Utilizada quando os elementos estão dispostos de maneira organizada (ex.: fila, lista) e aleatória.
- Escolhe um ponto de partida e seleciona-se cada k -ésimo elemento da população (ex.: o 50º elemento).
- Exemplo: Em uma fábrica de lâmpadas, a cada 100 peças produzidas, uma é retirada para teste.

Métodos probabilísticos:

Amostragem Sistemática

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Unidades Amostrais	
	Amostradas
	Não Amostradas

Métodos probabilísticos:

3 Amostragem Estratificada:

- Indicada quando a população está dividida em grupos distintos, denominados estratos.
- Dentro de cada estrato é realizada uma amostragem aleatória simples.
- O tamanho da amostra pode ou não ser proporcional ao tamanho do estrato.
- Exemplo: Uma comunidade universitária com 8000 indivíduos está estratificada da seguinte forma:

Estrato	População	Amostra
Professores	800	80
Funcionários	1200	120
Estudantes	6000	600

Métodos probabilísticos:

Amostragem Estratificada

Estrato 1	1	2	3	4	5	6	7	8	9	10
	11	12	13	14	15	16	17	18	19	20
Estrato 2	21	22	23	24	25	26	27	28	29	30
	31	32	33	34	35	36	37	38	39	40
Estrato 3	41	42	43	44	45	46	47	48	49	50
Estrato 4	51	52	53	54	55	56	57	58	59	60
Estrato 5	61	62	63	64	65	66	67	68	69	70
	71	72	73	74	75	76	77	78	79	80
Estrato 6	81	82	83	84	85	86	87	88	89	90
Estrato 7	91	92	93	94	95	96	97	98	99	100

Unidades Amostrais

Amostradas

Não Amostradas

Métodos probabilísticos:

4 Amostragem por Conglomerado:

- A área da população é dividida em seções (ou conglomerados, ex.: bairros, quarteirões).
- Os conglomerados são selecionados aleatoriamente.
- Dentro de um conglomerado, todos os elementos são amostrados.

Métodos probabilísticos:

Amostragem por Conglomerado

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Unidades Amostrais	
	Amostradas
	Não Amostradas

Erros amostrais *versus* Erros não amostrais

- **Erros amostrais** é diferença entre o resultado da amostra e o verdadeiro valor da população.
 - Ocorre pois as amostras são **aleatórias**!
 - Não importa quão bem a amostra seja coletada, os erros amostrais sempre irão ocorrer.
 - Cada vez que uma amostra aleatória for retirada de uma população, um resultado diferente será observado.

Erros amostrais *versus* Erros não amostrais

- **Erros amostrais** é diferença entre o resultado da amostra e o verdadeiro valor da população.
 - Ocorre pois as amostras são **aleatórias**!
 - Não importa quão bem a amostra seja coletada, os erros amostrais sempre irão ocorrer.
 - Cada vez que uma amostra aleatória for retirada de uma população, um resultado diferente será observado.
- **Erros não amostrais** ocorre quando os dados amostrais são coletados incorretamente, devido a uma amostra tendenciosa, instrumento de medida defeituoso, anotações erradas, entre outras, ...
 - Atenção: Os erros não amostrais não devem existir, ou devem ser minimizados.

Exemplo de erros amostrais

- Selecione uma amostra de tamanho $n = 5$ das idades dos estudantes de uma sala:
22 21 24 23 20 22 21 25 24 24 23 19 25 24 23 23 20 21 23 20 23 22 23 23 25
- Repita 5 vezes (tente ser o mais aleatório possível!), calcule a média de cada amostra, e compare com a média populacional $\mu = 22,5$.

Amostra	\bar{x}	$\epsilon = \bar{x} - \mu$
23 23 23 24 23	23.2	+0.7
24 22 20 20 20	21.2	-1.3
21 20 19 22 25	21.4	-1.1
22 23 25 20 22	22.4	-0.1
21 20 22 24 20	21.4	-1.1

- O que isso nos diz a respeito das médias amostrais? E da variabilidade das médias amostrais?
- E se fizemos uma “média das médias” de todas as amostras? Voltaremos aqui mais tarde!

Sumário

- 1 Ideias gerais
- 2 Amostragem
 - Tipos de amostragem.
 - Métodos de amostragem.
 - Erros na amostragem.
- 3 Análise exploratória de dados
 - Organização dos dados.
 - Tabelas de frequência.
 - Representação gráfica.
- 4 Exercícios recomendados

Exemplo

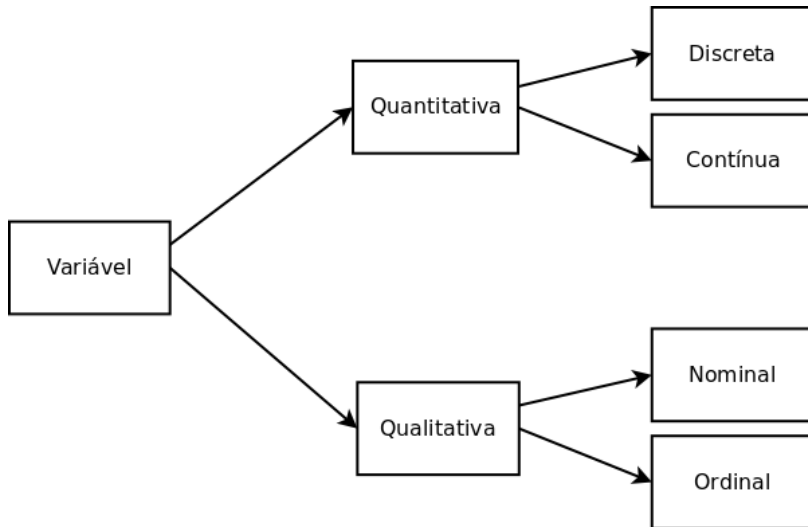
Pesquisa foi realizada com alunos. Variáveis:

- **Id**: identificação do aluno; **Turma**: A ou B;
- **Sexo**: feminino (F) ou masculino (M);
- **Idade**: em anos; **Alt**: altura em metros;
- **Peso**: em quilogramas; **Filhos**: n^o de filhos na família;
- **Fuma**: hábito de fumar: sim (S) ou não (N);
- **Toler**: tolerância ao cigarro: (I) indiferente; (P) incomoda pouco; (M) incomoda muito;
- **Exerc.**: horas de atividade física, por semana;
- **Cine**: n^o. de vezes que vai ao cinema por semana;
- **OpCine**: opinião a respeito das salas de cinema na cidade: (B) regular a boa; (M) muito boa;
- **TV**: horas gastas assistindo TV, por semana;
- **OpTV**: opinião a respeito da qualidade da programação na TV: (R) ruim; (M) média; (B) boa; (N) não sabe.

Organização dos dados

- A partir de um conjunto de dados coletado, a questão é:
 - Como extrair informações a respeito de uma ou mais características de interesse?
- Basicamente temos duas opções:
 - Tabelas de frequência;
 - Gráficos.
- O importante é levar em consideração a **natureza dos dados**.

Tipos de variáveis



Tipos de variáveis

- Variável qualitativa: são as características que não possuem valores quantitativos, mas são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos.
 - Nominal: não há uma ordenação entre as categorias.
 - Ordinal: existe uma ordem natural entre as categorias.
- Variável quantitativa: são as características que podem ser medidas em uma escala quantitativa, ou seja, apresentam valores numéricos que fazem sentido.
 - Discreta: os possíveis valores formam um conjunto finito ou enumerável de valores. É resultado de algum processo de contagem.
 - Contínua: os possíveis valores pertencem a um intervalo de números reais. É resultado de algum processo de mensuração.

Organização de Dados

- Uma típica **tabela de dados brutos** contém:
 - Variáveis (características, medições, etc) nas colunas.
 - Sujeito (indivíduo, objetos, etc) nas linhas.

Id	Turma	Sexo	Idade	Alt	Peso	Filhos	Fuma	Toler	Exerc	Cine	OpCine	TV	OpTV
1	A	F	17	1.60	60.5	2	NAO	P	0	1	B 16	R	
2	A	F	18	1.69	55.0	1	NAO	M	0	1	B 7	R	
3	A	M	18	1.85	72.8	2	NAO	P	5	2	M 15	R	
4	A	M	25	1.85	80.9	2	NAO	P	5	2	B 20	R	
5	A	F	19	1.58	55.0	1	NAO	M	2	2	B 5	R	
6	A	M	19	1.76	60.0	3	NAO	M	2	1	B 2	R	

- Tipos de variáveis:
 - Qualitativa nominal: Turma, Sexo, Fuma.
 - Qualitativa ordinal: Toler, OpCine, OpTV.
 - Quantitativa discreta: Idade, Filhos, Exerc, Cine, TV.
 - Quantitativa contínua: Alt, Peso.

Tabelas de frequência

- A tabela de dados brutos pode ser muito longa, portanto será difícil extrair alguma informação.
- As **tabelas de frequência** ajudam a resumir a informação da variável de interesse.
- Vamos usar 3 tipos de frequência:
 - Frequência **absoluta**: contagem de cada valor observado. Representado por n_i o número de indivíduos com a característica i .
 - Frequência **relativa**: número de indivíduos com a característica i dividido pelo total de indivíduos n , ou seja $f_i = \frac{n_i}{n}$.
 - Frequência **acumulada**: frequência (absoluta ou relativa) acumulada até um certo valor, obtida pela soma das frequências de todos os valores da variável, menores ou iguais ao valor considerado.

Tabela de frequência - qualitativa nominal

- Considerando a variável Sexo

	n_i	f_i
F	37	0.74
M	13	0.26
Sum	50	1.00

- Neste caso não faz sentido usar frequência acumulada.

Tabela de frequência - qualitativa ordinal

- Considerando a variável OpTV

	n_i	f_i	f_{ac}
R	39	0.78	0.78
M	1	0.02	0.80
B	3	0.06	0.86
N	7	0.14	1.00
Sum	50	1.00	

Tabela de frequência - quantitativa discreta

- Considerando a variável Idade

	n_i	f_i	f_{ac}
17	9	0.18	0.18
18	22	0.44	0.62
19	7	0.14	0.76
20	4	0.08	0.84
21	3	0.06	0.90
22	0	0.00	0.90
23	2	0.04	0.94
24	1	0.02	0.96
25	2	0.04	1.00
Sum	50	1.00	

Tabela de frequência - quantitativa contínua

- No caso de quantitativas contínuas não faz sentido contar cada valor pois podem existir muitos (potencialmente infinito).
- A solução é criar **classes** ou **faixas de valores**, e contar o número de ocorrências dentro destas classes.
- Para definir as classes:
 - Defina a amplitude da classe, de maneira que se obtenham de 5 a 8 classes (de mesma amplitude).
 - Identifique os valores máximo e mínimo da variável e construa as classes de maneira que inclua todos os valores.
- As classes de valores podem seguir um dos formatos:

Classe	Notação	Denominação	Resultado
$[a, b)$	$a \vdash b$	Fechado em a, aberto em b	Inclui a, não inclui b
$(a, b]$	$a \dashv b$	Aberto em a, fechado em b	Não inclui a, inclui b

Tabela de frequência - quantitativa contínua

- Considerando a variável Peso
 - Foram construídas 6 classes de amplitude 10.
 - As classes são do tipo $[a, b)$ ou $a \vdash b$.

	n_i	f_i	f_{ac}
[40, 50)	8	0.16	0.16
[50, 60)	22	0.44	0.60
[60, 70)	8	0.16	0.76
[70, 80)	6	0.12	0.88
[80, 90)	5	0.10	0.98
[90, 100)	1	0.02	1.00
Sum	50	1.00	

Tabela de frequência - quantitativa discreta (muitos valores)

- Considerando a variável TV
- Apesar de ser discreta, o número de valores únicos é muito grande e não seria útil contar as frequências de cada valor.
- Neste caso, utiliza-se o mesmo procedimento usado para quantitativas contínuas.
 - Foram construídas 6 classes de amplitude 6¹.

	n_i	f_i	f_{ac}
[0, 6)	14	0.28	0.28
[6, 12)	17	0.34	0.62
[12, 18)	11	0.22	0.84
[18, 24)	4	0.08	0.92
[24, 30)	3	0.06	0.98
[30, 36)	1	0.02	1.00
<i>Sum</i>	50	1.00	

¹Obs.: no livro a tabela tem 5 classes, pois a última tem comprimento 12

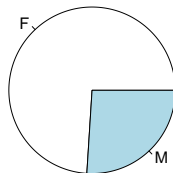
Representação gráfica

- Podemos visualizar as tabelas através de gráficos.
- Existe um tipo de gráfico adequado para cada tipo de variável.
- Cuidado deve ser tomado com representações visuais pois um gráfico desproporcional pode gerar interpretações distorcidas.
- As principais representações gráficas são:
 - Diagrama circular (setores ou “pizza”);
 - Gráfico de barras;
 - Histogramas;
 - Boxplots.

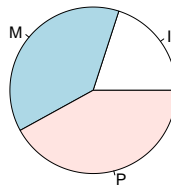
Diagrama circular

- Adequado para variáveis qualitativas nominal e ordinal.

Sexo



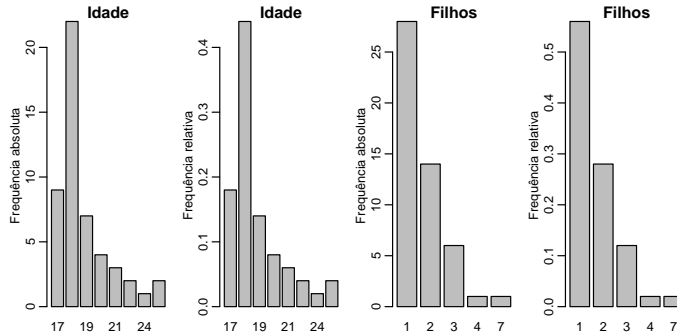
Toler



- O uso deste tipo de gráfico deve ser evitado, pois pode ser de difícil interpretação.

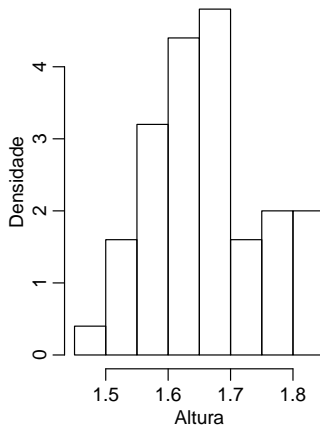
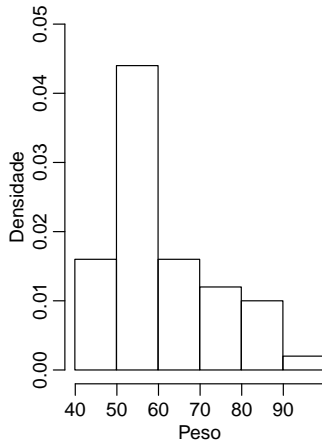
Gráfico de barras

- Adequado para variáveis qualitativas nominal/ordinal e quantitativa discreta (poucos valores distintos).
- Podem ser usadas as frequências absolutas ou relativas.



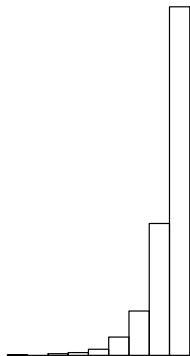
Histograma

- Adequado para quantitativa contínua.

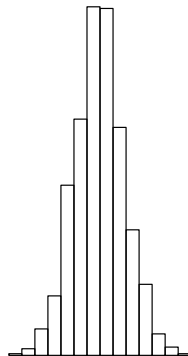


Tipos de assimetria

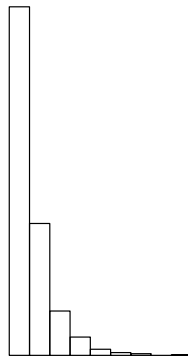
Assimétrico à esquerda



Simétrico



Assimétrico à direita

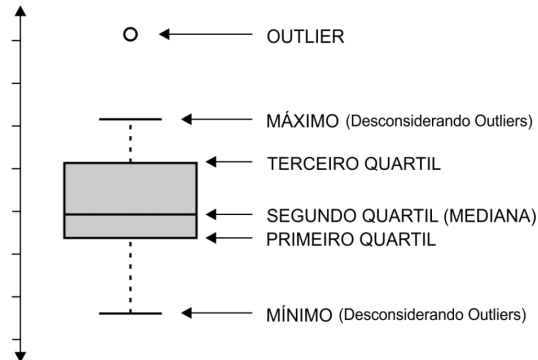


Mediana e quartis

- **Mediana:** valor da variável que divide o conjunto de dados ordenado em dois subgrupos de mesmo tamanho.
- **Quartis:** valores da variável que divide o conjunto de dados ordenados em quatro subgrupos de mesmo tamanho.
- **Posição** dos quartis:
 - $Q_1 = 0.25 \cdot (N + 1)$ e arredonde.
 - $Q_2 =$ média dos valores nas posições $(N/2)$ e $(N/2) + 1$ se N par e $Q_2 = (N + 1)/2$ se N ímpar.
 - $Q_3 = 0.75 \cdot (N + 1)$ e arredonde.
- Exemplo: Considere o conjunto de dados: 8.43(1), 8.65(2), 9.96(3), 10.91(6), 10.46(4) e 10.83(5).
 - $Q_1 = 0.25 \cdot 7 = 1.75 \approx 2$, ou seja 8.65.
 - $Q_2 =$ média dos valores nas posições 3 e 4, ou seja, $(9.96 + 10.46)/2 = 10.21$.
 - $Q_3 = 0.75 \cdot 7 = 5.25 \approx 5$, ou seja, 10.83.

Boxplot

- Adequado para quantitativa contínua.
- Pode ser usado também para quantitativa discreta com muitos valores.



Boxplots

- Excelente para explorar relações entre qualitativas e contínuas.

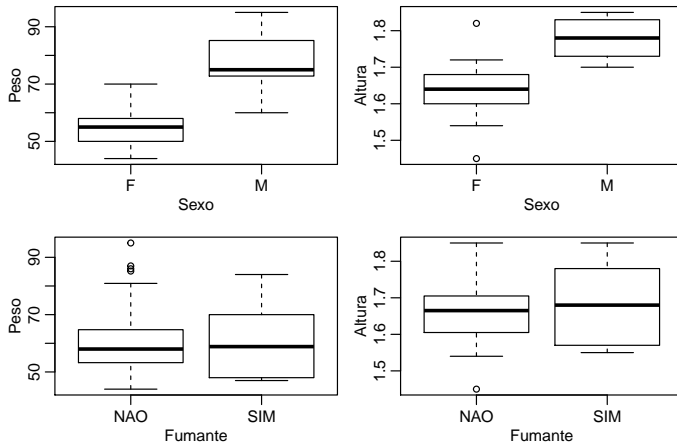


Diagrama de dispersão

- Adequado para explorar a relação entre variáveis quantitativas.

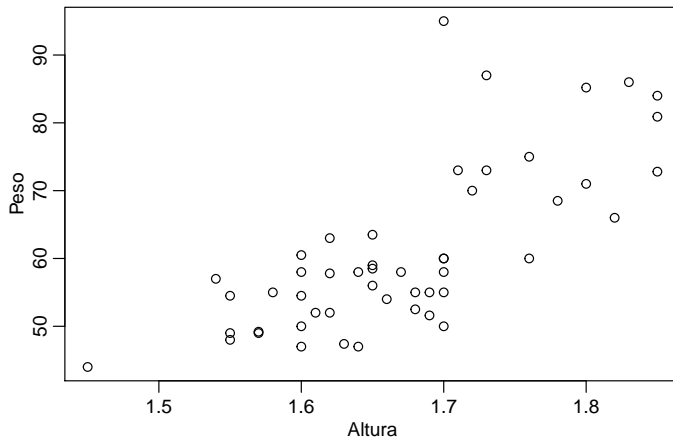


Diagrama de dispersão

- Exemplos de comportamentos do diagrama de dispersão.

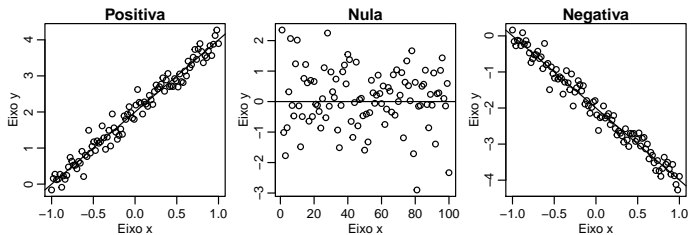


Diagrama de dispersão

- Formas não lineares.

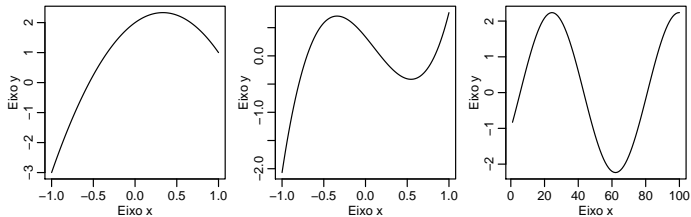
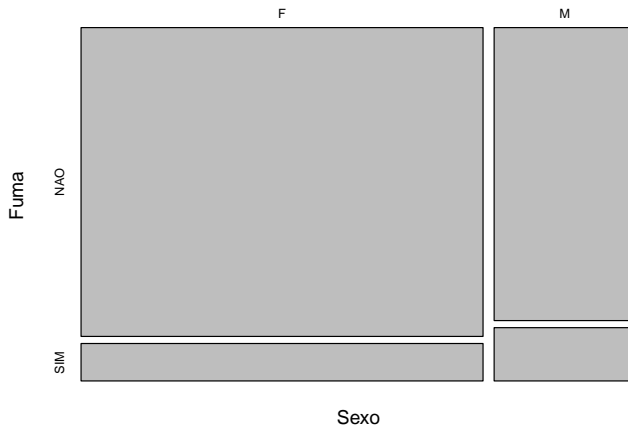


Gráfico de mosaico

- Adequado para explorar a relação entre variáveis qualitativas (nominais ou ordinais).



Resumo

- Qualitativa nominal ou ordinal:
 - Gráfico de setores.
 - Gráfico de barras.
- Quantitativa discreta:
 - Gráfico de barras (poucos valores).
 - Histograma ou boxplot (muitos valores).
- Quantitativas contínuas:
 - Histograma ou boxplot.
- Explorando relações:
 - Quantitativa vs Quantitativa: Diagrama de dispersão.
 - Qualitativa vs Quantitativa: Boxplots.
 - Qualitativa vs Qualitativa: Gráfico de mosaico.

Sumário

1 Ideias gerais

2 Amostragem

- Tipos de amostragem.
- Métodos de amostragem.
- Erros na amostragem.

3 Análise exploratória de dados

- Organização dos dados.
- Tabelas de frequência.
- Representação gráfica.

4 Exercícios recomendados

Exercícios recomendados

- Seção 1.1: Ex. 1, 2 e 3.
- Seção 1.2: Ex. 1, 2, 3 e 4.
- Seção 1.4: Ex. 1, 2, 3, 5 (troque diagrama circular por gráficos de barras), 8, 9, 12, 17 e 22.