Bioestatística

Prof. Ismael Bastos

O que é Estatística

Entendemos a Estatística como um conjunto de técnicas que permite, de forma sistemática, organizar, descrever, analisar e interpretar dados oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento.

O que é Bioestatística

A Bioestatística nada mais é que um ramo da estatística que visa aplicar conceitos estatísticos para o campo das Ciências Biológicas.

Áreas da Estatística

Podemos dividir a Estatística em três áreas:

- Estatística Descritiva
- Probabilidade
- Inferência Estatística

Áreas da Estatística

- **Estatística Descritiva** (*Conheça seus dados*) : Etapa inicial em qualquer análise de dados. É um conjunto de métodos estatísticos que tem por objetivo a organização, descrição e o resumo de dados.

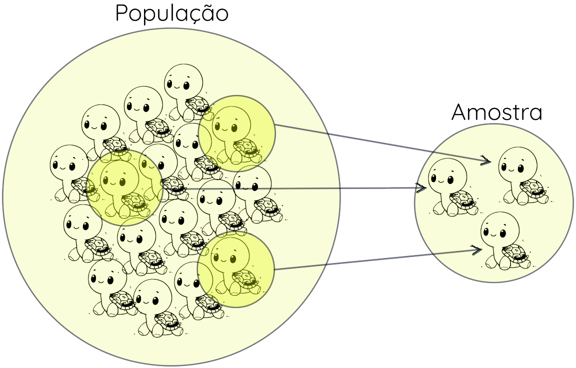
Áreas da Estatística

- **Estatística Descritiva** (*Conheça seus dados*) : Etapa inicial em qualquer análise de dados. É um conjunto de métodos estatísticos que tem por objetivo a organização, descrição e o resumo de dados.
- **Probabilidade** (*Qual a incerteza associada aos dados?*) : Auxilia na modelagem de fenômenos aleatórios, ou seja, aqueles em que há incerteza.

Áreas da Estatística

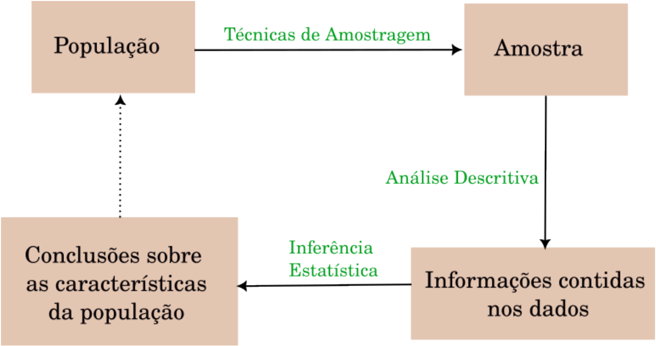
- **Estatística Descritiva** (*Conheça seus dados*) : Etapa inicial em qualquer análise de dados. É um conjunto de métodos estatísticos que tem por objetivo a organização, descrição e o resumo de dados.
- **Probabilidade** (*Qual a incerteza associada aos dados?*) : Auxilia na modelagem de fenômenos aleatórios, ou seja, aqueles em que há incerteza.
- **Inferência Estatística** (*Quais conclusões podemos tirar a partir destes dados?*): É um método estatístico que possibilita tirar conclusões sobre uma população a partir de informações contidas na amostra.

População x Amostra



Por que precisamos selecionar uma amostra?

Áreas da Estatística



- Trabalho 1 valendo 10 pontos
- Trabalho 2 valendo 10 pontos
- Prova 1 valendo 10 pontos
- Prova 2 valendo 10 pontos

$$\text{Nota Final} = 0,3 \cdot T_1 + 0,2 \cdot P_1 + 0,3 \cdot T_2 + 0,2 \cdot P_2$$

Funcionamento da disciplina

- Aulas expositivas com conteúdo em slide.
- Resolução de exemplos e exercícios no quadro *.
- Uso da linguagem de programação R.
- Disponibilização de informações da disciplina através do site da disciplina e do SIGA.
- Listas de exercício
- Alguns códigos estarão presentes nos slides, mas alguns outros serão disponibilizados separadamente.

*: A resolução de exemplos e exercícios não será disponibilizada no site.

- Por que usar a linguagem R para estatística?
- Instalação do R e R Studio.
- Posit Cloud.

Perguntas iniciais

- Como garantir que uma amostra representa bem a população?
  - Qual metodologia devo utilizar para selecionar minha amostra? **Exemplo**: Quero conduzir uma pesquisa eleitoral na cidade de São Paulo, como devo escolher as pessoas que vou entrevistar?

Perguntas iniciais

- Como garantir que uma amostra representa bem a população?
  - Qual metodologia devo utilizar para selecionar minha amostra?
    Exemplo: Quero conduzir uma pesquisa eleitoral na cidade de São Paulo, como devo escolher as pessoas que vou entrevistar?
  - Qual tamanho de amostra selecionar?
    Exemplo: Sei que o tamanho da população de São Paulo é de aproximadamente $11,45$ milhões de habitantes, quantas pessoas devo escolher?

Amostragem aleatória simples

Amostragem aleatória simples: Consideremos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória simples é feita da seguinte forma:

1. Numeramos os itens da população com número de 1 até N.

Amostragem aleatória simples

Amostragem aleatória simples: Consideremos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória simples é feita da seguinte forma:

1. Numeramos os itens da população com número de 1 até N.
2. Escrevemos cada um desses números em uma pedaço de papel.

Amostragem aleatória simples

Amostragem aleatória simples: Consideremos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória simples é feita da seguinte forma:

1. Numeramos os itens da população com número de 1 até N.
2. Escrevemos cada um desses números em uma pedaço de papel.
3. Colocamos esses papeis em uma urna bem misturados.

Amostragem aleatória simples

Amostragem aleatória simples: Consideremos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória simples é feita da seguinte forma:

1. Numeramos os itens da população com número de 1 até N.
2. Escrevemos cada um desses números em uma pedaço de papel.
3. Colocamos esses papeis em uma urna bem misturados.
4. Tiramos os n papeis correspondentes à amostra.

Amostragem aleatória simples

Amostragem aleatória simples: Consideremos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória simples é feita da seguinte forma:

1. Numeramos os itens da população com número de 1 até N.
2. Escrevemos cada um desses números em uma pedaço de papel.
3. Colocamos esses papeis em uma urna bem misturados.
4. Tiramos os n papeis correspondentes à amostra.

Pergunta: Essa retirada é com ou sem reposição?

Amostragem aleatória simples - Exemplo

O estudo O perfil socioeconômico e a percepção ambiental dos pescadores da Lagoa de Apodi, Rio Grande do Norte, Brasil propõe uma amostragem aleatória simples como metodologia de amostragem.

Amostragem aleatória simples - Exemplo no R

Nesse exemplo do R estamos assumindo que temos uma população de 10 indivíduos cujas idades estão armazenadas no vetor *populacao* e desejamos obter uma amostra de tamanho 3 utilizando a amostragem aleatória simples.

Amostragem aleatória simples no R

Nesse exemplo do R estamos assumindo que temos uma população de 10 indivíduos cujas idades estão armazenadas no vetor *populacao* e desejamos obter uma amostra de tamanho 3 utilizando a amostragem aleatória simples.

Amostragem aleatória simples no R

A função *sample* gera uma amostra aleatória simples de tamanho n, que nesse caso é 3.

Amostragem aleatória simples - Exemplo no R

Nesse exemplo do R estamos assumindo que temos uma população de 10 indivíduos cujas idades estão armazenadas no vetor *populacao* e desejamos obter uma amostra de tamanho 3 utilizando a amostragem aleatória simples.

Amostragem aleatória simples no R

A função *sample* gera uma amostra aleatória simples de tamanho n, que nesse caso é 3.

Obsevação: A função *sample* por padrão gera uma amostra selecionada sem reposição, para gerar uma amostra com reposição, basta adicionar o argumento $replace = TRUE$.

Amostragem aleatória estratificada: Consideremos que temos uma população (heterogênea) de tamanho N e desejamos uma amostra de tamanho n A amostragem aleatória estratificada é feita da seguinte forma:

1. Dividimos a população em estratos (com base em Idade, Renda, Sexo, Escolaridade ou outra variável).

Amostragem aleatória estratificada

Amostragem aleatória estratificada: Consideremos que temos uma população (heterogênea) de tamanho N e desejamos uma amostra de tamanho n A amostragem aleatória estratificada é feita da seguinte forma:

1. Dividimos a população em estratos (com base em Idade, Renda, Sexo, Escolaridade ou outra variável).
2. Determinamos o tamanho da amostra que iremos retirar de cada estrato, podendo ser feito de duas principais formas:

Amostragem aleatória estratificada: Consideremos que temos uma população (heterogênea) de tamanho N e desejamos uma amostra de tamanho n A amostragem aleatória estratificada é feita da seguinte forma:

1. Dividimos a população em estratos (com base em Idade, Renda, Sexo, Escolaridade ou outra variável).
2. Determinamos o tamanho da amostra que iremos retirar de cada estrato, podendo ser feito de duas principais formas:
   - **Proporcional**: A quantidade de indivíduos amostrados em cada estrato é proporcional ao tamanho do estrato.
   - **Igualitária**: Selecionamos o mesmo número de indivíduos em cada estrato, não levando em conta o tamanho.

Amostragem aleatória estratificada

Amostragem aleatória estratificada: Consideremos que temos uma população (heterogênea) de tamanho N e desejamos uma amostra de tamanho n A amostragem aleatória estratificada é feita da seguinte forma:

1. Dividimos a população em estratos (com base em Idade, Renda, Sexo, Escolaridade ou outra variável).
2. Determinamos o tamanho da amostra que iremos retirar de cada estrato, podendo ser feito de duas principais formas:
   - **Proporcional**: A quantidade de indivíduos amostrados em cada estrato é proporcional ao tamanho do estrato.
   - **Igualitária**: Selecionamos o mesmo número de indivíduos em cada estrato, não levando em conta o tamanho.
3. Realizamos a amostragem aleatória simples em cada estrato.

Amostragem aleatória estratificada - Exemplo

O estudo Hipertensão Arterial e Diabetes Mellitus entre trabalhadores da saúde: associação com hábitos de vida e estressores ocupacionais propõe uma amostragem aleatória estratificada como metodologia de amostragem.

Amostragem aleatória estratificada - Exemplo no R

Amostragem aleatória estratificada (Proporcional) no R

Amostragem aleatória estratificada: Consideremos que temos uma população (heterogênea) de tamanho N e desejamos uma amostra de tamanho n A amostragem aleatória por conglomerados é feita da seguinte forma:

1. Dividimos a população em conglomerados (como, por exemplo, escolas, unidades de saúde, bairros, etc.).

Amostragem aleatória por conglomerados

Amostragem aleatória estratificada: Consideremos que temos uma população (heterogênea) de tamanho N e desejamos uma amostra de tamanho n A amostragem aleatória por conglomerados é feita da seguinte forma:

1. Dividimos a população em conglomerados (como, por exemplo, escolas, unidades de saúde, bairros, etc.).
2. Determinamos o tamanho da amostra que iremos retirar de cada estrato, podendo ser feito de duas principais formas:

Amostragem aleatória por conglomerados

Amostragem aleatória estratificada: Consideremos que temos uma população (heterogênea) de tamanho N e desejamos uma amostra de tamanho n A amostragem aleatória por conglomerados é feita da seguinte forma:

1. Dividimos a população em conglomerados (como, por exemplo, escolas, unidades de saúde, bairros, etc.).
2. Determinamos o tamanho da amostra que iremos retirar de cada estrato, podendo ser feito de duas principais formas:
   - **Proporcional**: A quantidade de indivíduos amostrados em cada estrato é proporcional ao tamanho do estrato.
   - **Igualitária**: Selecionamos o mesmo número de indivíduos em cada estrato, não levando em conta o tamanho.

Amostragem aleatória estratificada: Consideremos que temos uma população (heterogênea) de tamanho N e desejamos uma amostra de tamanho n A amostragem aleatória por conglomerados é feita da seguinte forma:

1. Dividimos a população em conglomerados (como, por exemplo, escolas, unidades de saúde, bairros, etc.).
2. Determinamos o tamanho da amostra que iremos retirar de cada estrato, podendo ser feito de duas principais formas:
   - **Proporcional**: A quantidade de indivíduos amostrados em cada estrato é proporcional ao tamanho do estrato.
   - **Igualitária**: Selecionamos o mesmo número de indivíduos em cada estrato, não levando em conta o tamanho.
3. Realizamos a amostragem aleatória simples em cada estrato.

Outros tipos de amostragem

Considere, agora, que um aluno esteja interessado em avaliar a opinião dos alunos da UFRJ sobre o serviço de transporte entre os diversos campi, oferecido pela administração da universidade. Como ele não tem **condições** nem **tempo** de selecionar uma amostra de todos os alunos a UFRJ, decide entrevistar seus colegas de turma.

Outros tipos de amostragem

Considere, agora, que um aluno esteja interessado em avaliar a opinião dos alunos da UFRJ sobre o serviço de transporte entre os diversos campi, oferecido pela administração da universidade. Como ele não tem **condições** nem **tempo** de selecionar uma amostra de todos os alunos a UFRJ, decide entrevistar seus colegas de turma.

- A decisão desse aluno é razoável? Ou seja, essa amostra é representativa?
- Qual a diferença desse método para os que estudamos anteriormente?

Outros tipos de amostragem

Considere, agora, que um aluno esteja interessado em avaliar a opinião dos alunos da UFRJ sobre o serviço de transporte entre os diversos campi, oferecido pela administração da universidade. Como ele não tem **condições** nem **tempo** de selecionar uma amostra de todos os alunos a UFRJ, decide entrevistar seus colegas de turma.

- A decisão desse aluno é razoável? Ou seja, essa amostra é representativa?
- Qual a diferença desse método para os que estudamos anteriormente?

Esse tipo de amostragem é chamado de **amostragem por conveniência** e é um tipo de amostragem não-probabilística.

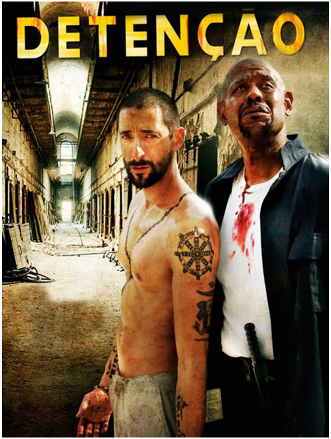
Problemas dos métodos de amostragem discutidos

- Como vou ter acesso a lista de todas as pessoas da população?

Problemas dos métodos de amostragem discutidos

- Como vou ter acesso a lista de todas as pessoas da população?
- Será que a informação sobre a proporção de elementos em cada estrato é confiável?

O problema do viés - Experimento de aprisionamento de Stanford (1971)



O problema do viés - Experimento de aprisionamento de Stanford (1971)

- Participantes de uma comunidade local foram convidados através de um anúncio de jornal que recrutava estudantes do sexo masculino para participar de um "**estudo psicológico da vida em uma prisão**".

O problema do viés - Experimento de aprisionamento de Stanford (1971)

- Participantes de uma comunidade local foram convidados através de um anúncio de jornal que recrutava estudantes do sexo masculino para participar de um "**estudo psicológico da vida em uma prisão**".
- Caso o estudante fosse aprovado, receberia $15 por dia (equivalente a aproximadamente $120 atualmente).

O problema do viés - Experimento de aprisionamento de Stanford (1971)

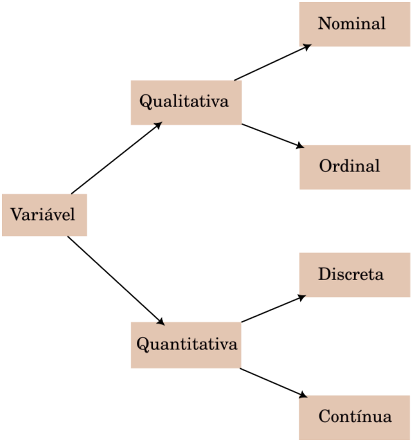
- Participantes de uma comunidade local foram convidados através de um anúncio de jornal que recrutava estudantes do sexo masculino para participar de um "**estudo psicológico da vida em uma prisão**".
- Caso o estudante fosse aprovado, receberia \$15 por dia (equivalente a aproximadamente \$120 atualmente).
- Os participantes foram submetidos a testes psicológicos e alocados aleatoriamente na posição de prisioneiro ou carcereiro.
- **Conclusão do Experimento**: Pessoas "boas" podem ter um "mal" comportamento quando colocadas em posições de poder ou sob condições de estresse.

Exemplos no nosso cotidiano

- Pesquisa eleitoral sobre intenções de voto para 2026 - UOL.
- Pesquisa eleitoral sobre intenções de voto para 2026 - SP - CNN.
- Pesquisa eleitoral sobre intenções de voto para 2026 - SP - Paraná Pesquisas.
- Pharmacokinetics and pharmacodynamics of cisatracurium in young and elderly adult patients.
- The Women's Health Initiative Randomized Trials and Clinical PracticeA Review

- **Variável:** Característica analisada (Nome da coluna)
  - **Quantitativa:** Assume valores numéricos
    - Discreta: Assume valores discretos, valores inteiros.
    - Contínua: Assume valores no intervalo dos números reais.
  - **Qualitativa:** Assume valores que representam atributos e/ou qualidades
    - **Ordinal**: Valores assumem ordem, indicando intensidade crescentes de realização
    - **Nominal**: Valores não assumem ordem

Tipos de variáveis



Exemplo inicial

Est. Civil	Instrução	Filhos	Salário	Idade
Solteiro	Fundamental	0	4,00	26
Casado	Médio	1	4,56	32
Casado	Superior	3	6,20	34
Solteiro	Médio	2	5,23	21
Casado	Superior	2	3,23	24
Solteiro	Médio	1	7,90	56
Casado	Fundamental	4	6,45	67
Casado	Fundamental	0	4,56	34
Solteiro	Médio	1	6,78	56
Solteiro	Superior	0	3,56	34

Tipos de variáveis

Atenção: Ao analisar as variáveis estamos olhando para a informação trazida e não para o valor em si.

Atenção: Ao analisar as variáveis estamos olhando para a informação trazida e não para o valor em si.

Est. Civil	Instrução	Filhos	Salário	Idade
1	1	0	4,00	26
2	2	1	4,56	32
2	3	3	6,20	34
1	2	2	5,23	21
2	3	2	3,23	24
1	2	1	7,90	56
2	1	4	6,45	67
2	1	0	4,56	34
1	2	1	6,78	56
1	3	0	3,56	34

Tipos de variáveis

Atenção: Ao analisar as variáveis estamos olhando para a informação trazida e não para o valor em si.

Est. Civil	Instrução	Filhos	Salário	Idade
1	1	0	4,00	26
2	2	1	4,56	32
2	3	3	6,20	34
1	2	2	5,23	21
2	3	2	3,23	24
1	2	1	7,90	56
2	1	4	6,45	67
2	1	0	4,56	34
1	2	1	6,78	56
1	3	0	3,56	34

Nesse caso a tabela acima é equivalente a vista no slide anterior e a interpretação das variáveis permanece a mesma.

Vamos para um exemplo real. Abrir base de dados do curso.

Tabela de frequência

Uma tabela de frequência relaciona categorias ou classes de valores juntamente com as frequências do número de valores que se enquadram em cada categoria ou classe.

Denotaremos por:

- n_i: Frequência absoluta
- f_i: Frequência relativa

onde, $f_i = \dfrac{n_i}{n}$

Exemplo 1

Construa a tabela de frequência para a variável formação utilizando os dados da tabela abaixo

idade	experiência(anos)	sexo	cidade	formação	salário (R$)
20	2	1	Rio de Janeiro	Farmácia	6000
25	5	0	São Paulo	Engenharia Química	5000
35	7	1	Rio de Janeiro	Farmácia	10000
40	15	1	Rio de Janeiro	Engenharia Química	6000
45	25	1	Rio de Janeiro	Farmácia	7500
22	1	0	São Paulo	Farmácia	1740
34	7	0	Rio de Janeiro	Farmácia	7000
56	25	1	São Paulo	Engenharia Química	6500
19	0	1	Rio de Janeiro	Engenharia Química	1990
29	9	0	São Paulo	Engenharia Química	8000

Tabela de frequência

Formação	n_i	f_i
Farmácia	5	0,5
Engenharia Química	5	0,5
total	n=10	1

No exemplo a seguir é criado o vetor *populacao* contendo 10 indivíduos, sendo armazenado o sexo (M - Masculino, F - Feminino).

Tabela de frequência no R (Frequência Absoluta)

Tabela de frequência - Exemplo no R

No exemplo a seguir é criado o vetor *populacao* contendo 10 indivíduos, sendo armazenado o sexo (M - Masculino, F - Feminino).

Tabela de frequência no R (Frequência Absoluta)

Ao executar o código percebemos que a função *table* retorna apenas uma tabela contendo a frequência absoluta.

Tabela de frequência - Exemplo no R (Frequência Relativa)

Uma forma de mostrar a frequência relativa é apresentada no código abaixo.

Tabela de frequência no R (Frequência Relativa)

Uma forma de mostrar a frequência relativa é apresentada no código abaixo.

Tabela de frequência no R (Frequência Relativa)

Atenção: Perceba que no código acima, a função que gera a tabela de frequência relativa (*prop_table*) recebe como argumento a própria tabela de frequência absoluta.

Tabela de frequência - Exemplo no R

- Como gero uma tabela contendo as duas informações?
- Como eu gero uma tabela de frequência a partir de dados reais?

Pausa - Aprendendo sobre DataFrames no R

Nome dos arquivos de código:

- Como criar e trabalhar com DataFrames em R: **criacao_dataframe.R**
- Como criar uma tabela de frequência completa no R: **tabela_freq.R**.
- Como ler bases de dados no R: **leitura_base_dados.R**

Pausa - Aprendendo sobre DataFrames no R

Nome dos arquivos de código:

- Como criar e trabalhar com DataFrames em R: **criacao_dataframe.R**
- Como criar uma tabela de frequência completa no R: **tabela_freq.R**.
- Como ler bases de dados no R: **leitura_base_dados.R**

Para aprender um pouco mais, recomendo a leitura da página 30 do livro **Uma introdução à programação com o R**.

Frequência acumulada e divisão em classes

Comumente se torna útil criar tabelas de frequência agrupadas por classes (intervalos) específicas. Tomemos como exemplo a variável idade, podemos agrupar os dados através da criação de faixas etárias.

idade	experiência(anos)	sexo	cidade	formação	salário (R$)
20	2	1	Rio de Janeiro	Administração	6000
25	5	0	São Paulo	Engenharia de Produção	5000
35	7	1	Rio de Janeiro	Administração	10000
40	15	1	Rio de Janeiro	Engenharia de Produção	6000
45	25	1	Rio de Janeiro	Administração	7500
22	1	0	São Paulo	Administração	1740
34	7	0	Rio de Janeiro	Administração	7000
56	25	1	São Paulo	Engenharia de Produção	6500
19	0	1	Rio de Janeiro	Engenharia de Produção	1990
29	9	0	São Paulo	Engenharia de Produção	8000

Faixa etária	n_i	f_i	f_{ac}
[15, 25)	3	0,3	0,3
[25, 35)	3	0,3	0,6
[35, 45)	2	0,2	0,8
[45, 55)	1	0,1	0,9
[55, 65)	1	0,1	1
total	10	1	

- Qual amplitude escolher?
- Todos os intervalos devem ter o mesma amplitude (tamanho)?

Frequência acumulada e divisão em classes

- Geralmente a divisão em classes é bastante útil para estudar variáveis quantitativas contínuas, pois geralmente os valores não se repetem.
- Tomemos como exemplo a variável salário da tabela do slide anterior, nela apenas o valor 6000 se repete, todos os demais são únicos. Se contruísimmos uma tabela de frequência comum, a maior parte dos valores teria frequência absoluta igual a 1.

Exemplo 7: Em uma turma do curso de Matemática foi feito o registro da idade de cada um dos estudantes dessa turma.

22	23	44	33	20	27	19	54	37	22
25	29	40	23	20	30	24	39	28	21

Construa uma tabela de frequência partindo da menor idade para a variável idade. Defina a amplitude como sendo igual a 5 e inclua apenas o limite inferior dos intervalos.

Exemplo 7: Em uma turma do curso de Matemática foi feito o registro da idade de cada um dos estudantes dessa turma.

22	23	44	33	20	27	19	54	37	22
25	29	40	23	20	30	24	39	28	21

Construa uma tabela de frequência partindo da menor idade para a variável idade. Defina a amplitude como sendo igual a 5 e inclua apenas o limite inferior dos intervalos.

Idade	n_i	f_i	f_{ac}
[19, 24)	8	0,4	0,4
[24, 29)	4	0,2	0,6
[29, 34)	3	0,15	0,75
[34, 39)	1	0,05	0,8
[39, 44)	2	0,1	0,9
[44, 49)	1	0,05	0,95
[49, 54)	0	0	0,95
[54, 59)	1	0,05	1
total	n=20	1	

Gráfico de setores (pizza)

Consiste em dividir um círculo (pizza) em diferentes setores (fatias), cada um representando a proporção do elemento analisado em relação ao conjunto de estudo.

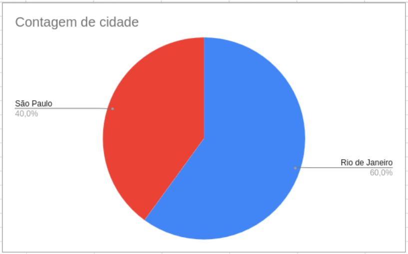


Figura: Gráfico de pizza para a variável cidade

Gráfico de barras

Utiliza o plano cartesiano com os valores da variável no eixo das abcissas e as frequências ou porcentagens no eixo das ordenadas.

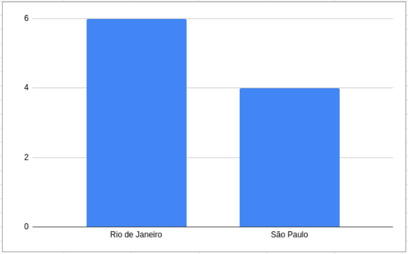


Figura: Gráfico de barras para a variável cidade

Histograma

O histograma é bastante semelhante ao gráfico de barras, de maneira informal, o histograma nada mais é que uma representação gráfica da uma table agrupada em classes.

Faixa etária	n_i	f_i	f_{ac}
[15, 25)	3	0,3	0,3
[25, 35)	3	0,3	0,6
[35, 45)	2	0,2	0,8
[45, 55)	1	0,1	0,9
[55, 65)	1	0,1	1
total	10	1	

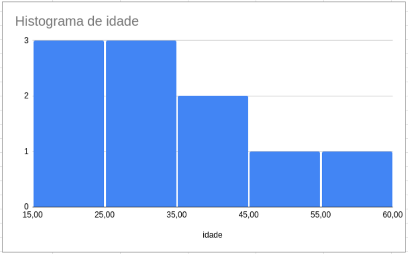


Figura: Histograma da variável idade

Medidas de Posição

Podemos ter o interesse em resumir nosso conjunto de dados, apresentando um ou mais valores que de certa forma representem todo o conjunto de dados.
Para determinar esses valores, utiliza-se as **medidas de posição central**: Média, Mediana e Moda.

Notações comumente adotadas

Em geral, é usual definir nosso conjunto de dados em estudo pela letra X e cada elemento desse conjunto de dados por X_i. Também é comum definir por n o tamanho total do conjunto de dados

Exemplo 2

Seja

$$90, 95, 45, 87, 56, 78, 89, 76, 55, 54, 56$$

O conjunto do peso de cada um dos alunos de uma determinada classe. Nesse caso, temos:

- $X = \{90, 95, 45, 87, 56, 78, 89, 76, 55, 54, 56\}$
- $n = 11$
- $X_1 = 90, X_2 = 95, \ldots, X_{11} = 56$

Média Amostral

A média aritmética, ou simplesmente média é uma medida que consiste em somar todos os valores de uma amostra ou população e dividir pelo total de elementos presentes na amostra. Denotaremos a média de um conjunto X qualquer por $\bar{X}$. A expressão para seu cálculo é dado a seguir:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{X_1 + \cdots + X_n}{n}$$

Média

Exemplo 3

Suponha que coletamos a altura de 10 *alunos de uma turma, as alturas obtidas foram as seguintes:*

$$1,90; 1,67; 1,87; 1,55; 1,76; 1,87; 1,95; 1,66; 1,75; 1,60$$

Logo,

$$\bar{X} = \frac{1,90 + 1,67 + 1,87 + 1,55 + 1,76 + 1,87 + 1,95 + 1,66 + 1,75 + 1,60}{10} = 1,76$$

Exemplo 4

Um aluno da disciplina matemática financeira resolveu investir na bolsa de valores, sua estratégia consistia em comprar uma ação na segunda-feira e em seguida vende-la na sexta-feira, a seguir são apresentados os ganhos, e perdas, em reais que ele teve ao longo de 8 semanas:

Semana 1	Semana 2	Semana 3	Semana 4	Semana 5	Semana 6	Semana 7	Semana 8
1	1,50	-2	2,30	-0,40	-1,60	-0,20	-0,60

Calcule o lucro médio que esse aluno teve. Essa estratégia foi vencedora?

Exemplo 4

Um aluno da disciplina matemática financeira resolveu investir na bolsa de valores, sua estratégia consistia em comprar uma ação na segunda-feira e em seguida vende-la na sexta-feira, a seguir são apresentados os ganhos, e perdas, em reais que ele teve ao longo de 8 semanas:

Semana 1	*Semana 2*	*Semana 3*	*Semana 4*	*Semana 5*	*Semana 6*	*Semana 7*	*Semana 8*
1	*1,50*	*-2*	*2,30*	*-0,40*	*-1,60*	*-0,20*	*-0,60*

Calcule o lucro médio que esse aluno teve. Essa estratégia foi vencedora?

Temos que:

$$\bar{X} = \frac{1 + 1,50 - 2 + 2,30 - 0,40 - 2,80 - 0,20 - 0,60}{8} = \frac{-1.20}{8} = -0,15$$

Mediana

A mediana é definida como o valor tal que 50% dos valores da variável estão acima da mediana e 50% estão abaixo.

Importante: Antes de calcular a mediana, é necessário ordenar os dados.

Mediana

A mediana é definida como o valor tal que 50% dos valores da variável estão acima da mediana e 50% estão abaixo.
Importante: Antes de calcular a mediana, é necessário ordenar os dados.

Exemplo 5

Suponhamos que o peso de 7 crianças de uma determinada turma escolar tenha sido coletada, os pesos coletados são exibidos a seguir:

$$28, 34, 45, 39, 42, 33, 50$$

Obtenha a mediana desse conjunto de dados

Primeiramente vamos ordenar os dados:

$$28, 33, 34, 39, 42, 45, 50$$

Logo, a mediana será o valor 39

Exemplo 6

Retomemos o exemplo das alturas dos alunos, temos os seguintes dados:

$$1,90; 1,67; 1,87; 1,55; 1,76; 1,87; 1,95; 1,66; 1,75; 1,60$$

Para calcular a mediana, primeiramente vamos ordenar os dados:

$$1,55; 1,60; 1,66; 1,67; 1,75; 1,76; 1,87; 1,87; 1,90; 1,95$$

Temos 10 elementos, logo, não conseguimos encontrar o valor que divide exatamente ao meio o conjunto de dados, nesse caso, a mediana é dada pela média dos elementos que ocupam a posição $\frac{n}{2}$ e $\frac{n}{2}+1$. Logo, denotando a mediana pela letra M, temos que:

$$M = \frac{1,75 + 1,76}{2} = 1,755$$

Mediana

Em resumo temos que, caso o tamanho do conjunto de dados seja impar, então a mediana é dada pelo elemento que ocupa a posição $\frac{n}{2}$ arredondado para cima, ou seja, o elemento que fica exatamente no meio. Caso o tamanho do conjunto de dados seja par, então fazemos a média dos elementos das posições $\frac{n}{2}$ e $\frac{n}{2}+1$. De maneira matemática, temos:

$$M = \begin{cases} x_{\frac{n+1}{2}} & \text{se n é impar} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{se n é par} \end{cases}$$

Onde M representa a mediana, x_i o i-ésimo elemento do conjunto de dados ordenado e n o total de elementos no conjunto de dados.

Observações

- Quando a média é próxima a mediana dizemos que isso é um indicativo de que a amostra não apresenta pontos aberrantes (possíveis outliers).
- A média é fortemente influenciada pela presença de pontos aberrantes, enquanto a moda não é.

Moda

A moda é simplesmente o valor que mais se repete, ou seja, o que apresenta maior frequência.

Importante: Caso existam dois valores que aparecem com maior frequência então teremos dua modas, se existirem três, então teremos três modas e assim sucessivamente.

Moda

A moda é simplesmente o valor que mais se repete, ou seja, o que apresenta maior frequência.
Importante: Caso existam dois valores que aparecem com maior frequência então teremos dua modas, se existirem três, então teremos três modas e assim sucessivamente.

Exemplo

Retomemos nossa tabela

idade	experiência(anos)	sexo	cidade	formação	salário (R$)
20	2	1	Rio de Janeiro	Administração	6000
25	5	0	São Paulo	Engenharia de Produção	5000
35	7	1	Rio de Janeiro	Administração	10000
40	15	1	Rio de Janeiro	Engenharia de Produção	6000
45	25	1	Rio de Janeiro	Administração	7500
22	1	0	São Paulo	Administração	1740
34	7	0	Rio de Janeiro	Administração	7000
56	25	1	São Paulo	Engenharia de Produção	6500
19	0	1	Rio de Janeiro	Engenharia de Produção	1990
29	9	0	São Paulo	Engenharia de Produção	8000

Se olharmos para a variável experiência, temos que a moda é 7 e 25

Moda

Exemplo 7

Retomemos o exemplo das alturas

$$1,90; 1,67; 1,87; 1,55; 1,76; 1,87; 1,95; 1,66; 1,75; 1,60$$

Temos que a moda é $1,87$

Quartis

Os quartis dividem o conjunto de dados em quatro partes iguais.

- O primeiro quartil, denotado por $Q1$, é o valor o qual 25% dos dados se encontram antes dele.
- O segundo quartil, denotado por $Q2$, é o valor o qual 50% dos dados se encontram antes dele.
- O terceiro quartil, denotado por $Q3$, é o valor o qual 75% dos dados se encontram antes dele.

Observação:Perceba que o segundo quartil é exatamente a mediana.

Quartis

Para obter a posição dos quartis, deve-se seguir as seguintes regras

- Posição do primeiro quartil($Q1$) $= \dfrac{n+1}{4}$
- Posição do segundo quartil($Q2$) = Calcular a mediana
- Posição do terceiro quartil($Q3$) $= \dfrac{3 \cdot (n+1)}{4}$

Observações

- Os dados devem estar ordenados
- Caso Q1 ou Q2 não sejam inteiros, fazer a média entre o elemento referente à parte inteira e seu sucessor.

Quartis

Exemplo 8

Retomemos o exemplo das alturas

$$1,90; 1,67; 1,87; 1,55; 1,76; 1,87; 1,95; 1,66; 1,75; 1,60$$

Calcule os quartis para esse conjunto de dados

Exemplo 8

Retomemos o exemplo das alturas

$$1,90; 1,67; 1,87; 1,55; 1,76; 1,87; 1,95; 1,66; 1,75; 1,60$$

Calcule os quartis para esse conjunto de dados

Assim como fizemos para mediana, é necessário primeiramente ordenar os dados.

$$1,55; 1,60; 1,66; 1,67; 1,75; 1,76; 1,87; 1,87; 1,90; 1,95$$

Temos que $x_{Q1} = \dfrac{10+1}{4} = 2,75$
Como $2,75$ não é um número inteiro, então devemos fazer a média entre os elementos que ocupam a segunda e terceira posição. Logo,

$$Q1 = \frac{1,60 + 1,66}{2} = 1,63$$

Quartis

$$x_{Q3} = \frac{3 \cdot (10 + 1)}{4} = 8,25$$

Como $8,25$ não é um número inteiro, então devemos fazer a média entre os elementos que ocupam a oitava e nona posição. Logo,

$$Q3 = \frac{1,87 + 1,90}{2} = 1,88$$

Boxplot

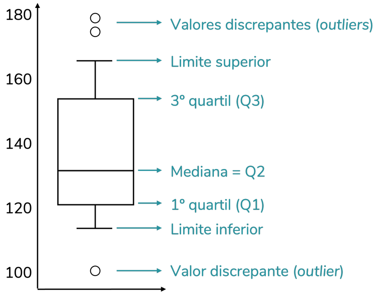
É um gráfico construído com base no resumo de cinco medidas, constituído por:

- Limite inferior (LI)
- 1º Quartil (Q_1)
- 2º Quartil (Q_2)
- 3º Quartil (Q_3)
- Limite superior (LS)

Para o cálculo do limite máximo e mínimo do boxplot, temos:

- $LI = Q_1 - 1,5(Q_3 - Q_1)$
- $LS = Q_3 + 1,5(Q_3 - Q_1)$

Boxplot



180
160
140
120
100
Valores discrepantes (*outliers*)
Limite superior
3º quartil (Q3)
Mediana = Q2
1º quartil (Q1)
Limite inferior
Valor discrepante (*outlier*)

Boxplot

Exemplo: Desenhe o boxplot para o exemplo das alturas.

Motivação

Tomemos o seguinte exemplo:

Exemplo 9

Em duas escolas foi verificado os seguintes resultados para as notas da prova de matemática dos cinco melhores alunos da turma :

Escola A: 9, 9, 9, 9, 9

Escola B: 10, 9,5, 9, 8,5, 8

Calcule a nota média desses alunos:

Motivação

Tomemos o seguinte exemplo:

Exemplo 9

Em duas escolas foi verificado os seguintes resultados para as notas da prova de matemática dos cinco melhores alunos da turma :
Escola A: 9, 9, 9, 9, 9
Escola B: 10, 9,5, 9, 8,5, 8
Calcule a nota média desses alunos:

Denotando por $\bar{X}_A$,M_A, $\bar{X}_B$,M_B a média e mediana de A e B respectivamente:

$$\bar{X}_A = \frac{9+9+9+9+9}{5} = 9$$

$$\bar{X}_B = \frac{10+9,5+9+8,5+8}{5} = 9$$

$$M_A = M_B = 9$$

As vezes podemos estar interessados em como os dados se comportam, ou seja, estamos interessados em como eles variam. Nesse sentido, introduziremos três novas medidas, o desvio médio, a variância e o desvio padrão.

Desvio médio

O desvio médio tem como objetivo verificar quanto os valores estão se distanciando da média, sua expressão é dada pela seguinte expressão:

$$dm(X) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Exemplo 10

Calcule o desvio médio para as notas do exemplo anterior.

O desvio médio tem como objetivo verificar quanto os valores estão se distanciando da média, sua expressão é dada pela seguinte expressão:

$$dm(X) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Exemplo 10

Calcule o desvio médio para as notas do exemplo anterior.

$$dm_A(X) = 0$$

$$dm_B(X) = \frac{|10-9| + |9,5-9| + |9-9| + |8,5-9| + |8-9|}{5} = \frac{3}{5} = 0,6$$

Variância

A variância é bastante semelhante ao desvio médio, no entanto, oferece uma função convexa.

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

Exemplo 11

Calcule a variância para as notas do exemplo anterior.

Variância

A variância é bastante semelhante ao desvio médio, no entanto, oferece uma função convexa.

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

Exemplo 11

Calcule a variância para as notas do exemplo anterior.

$$s_A^2(X) = 0$$

$$s_B^2(X) = \frac{(10-9)^2 + (9,5-9)^2 + (9-9)^2 + (8,5-9)^2 + (8-9)^2}{5}$$
$$= \frac{2,5}{5} = 0,5$$

Observe que a variância é sempre um número **positivo**

Desvio padrão

O desvio padrão consiste em tomar a raiz quadrada da variância. Indica a variação dos dados em relação à média, ou seja, o erro (desvio) cometido ao substituir nossa observação pela média.

$$s = \sqrt{Var(X)}$$

- A variância não tem interpretação prática direta.
- Já o desvio padrão tem, pois ele volta o valor para a unidade de medida dos dados.

Coeficiente de variação

Ainda assim, o desvio padrão não nos da uma base se a variação é alta ou baixa. Por exemplo, um desvio padrão de 0,6 é baixo? Um desvio padrão de 240.000 é alto? Nesse contexto surge o coeficiente de variação.

Coeficiente de variação

O cálculo do coeficiente de variação (CV) é dado pela seguinte expressão:

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

- Quanto menor for o valor do CV, mais homogêneo é um conjunto de dados

Podemos adotar o seguinte critério:

- $CV < 0,10 \implies$ variabilidade baixa
- $0,10 \leq CV < 0,20 \implies$ variabilidade intermediária
- $0,20 \leq CV < 0,30 \implies$ variabilidade alta
- $CV \geq 0,30 \implies$ variabilidade muito alta

Observação: Só podemos calcular o CV quando a média amostral for diferente de zero.