Estatística Basica I

Universidade Estadual Paulista, Júlio de Mesquita Filho - UNESP

clayton.pereira@unesp.br

Por que estudar Estatística

Referências

Referências e Fontes das Imagens

- Python Data Science Handbook

Por que estudar Estatística?

- A estatística está em tudo:
  - Nas recomendações da Netflix, nos resultados das eleições, nas métricas de um modelo de IA.

Por que estudar Estatística?

- A estatística está em tudo:
  - Nas recomendações da Netflix, nos resultados das eleições, nas métricas de um modelo de IA.
- Ela nos ajuda a responder:
  - O que os dados estão dizendo?
  - Essas diferenças são significativas?
  - Posso confiar nessa previsão?

Por que estudar Estatística?

- A estatística está em tudo:
  - Nas recomendações da Netflix, nos resultados das eleições, nas métricas de um modelo de IA.
- Ela nos ajuda a responder:
  - O que os dados estão dizendo?
  - Essas diferenças são significativas?
  - Posso confiar nessa previsão?
- Técnicas essenciais:
  - Medidas de tendência central, variabilidade, distribuição, probabilidade, inferência e visualização.

Por que estudar Estatística?

- A estatística está em tudo:
  - Nas recomendações da Netflix, nos resultados das eleições, nas métricas de um modelo de IA.
- Ela nos ajuda a responder:
  - O que os dados estão dizendo?
  - Essas diferenças são significativas?
  - Posso confiar nessa previsão?
- Técnicas essenciais:
  - Medidas de tendência central, variabilidade, distribuição, probabilidade, inferência e visualização.
- Na prática:
  - Usamos estatística para tomar decisões baseadas em dados reais.

"A Estatística é o que transforma dados em conhecimento."

Introdução

- Ciência de **aprender com dados**
  - Ajuda a usar os métodos adequados para coletar os dados, empregar a análise correta e apresentar os resultados de forma eficaz
- Os dados que estudamos são observações (amostras) de uma ou mais variáveis.
- Uma **variável** (aleatória) é aquilo que é observado para estudar um determinado fenômeno (idade, sexo, peso, etc.)
- A Estatística provê meios para classificar, resumir, organizar, analisar e interpretar dados.
- Envolve: descrever Conjuntos de Dados e tirar conclusões (fazer estimativas, decisões, previsões, etc. a cerca de conjuntos de dados

Introdução

- **Regra:**
  - A estatística deve simplificar e não complicar a interpretação dos dados, caso isso aconteça, volte, pois há algo de errado!!!

Estatística Analítica

- Estatística descritiva:
  - Se concentram na descrição das características visíveis de um conjunto de dados (uma população ou amostra)
- Estatística Inferencial:
  - Se concentram em fazer previsões (inferir) ou generalizações sobre um conjunto de dados maior (população), com base em amostras desses dados.
  - Intervalo de Confiança;
  - Teste de Hipótese;
  - Comparação entre grupos

Estatística Analítica

- População:
  - Pode ser definida como a totalidade de elementos que compõem um determinado conjunto, tendo obrigatoriamente alguma característica que conecte esses elementos;
- Amostra:
  - Pode ser definida como uma parte (ou subconjunto) dos elementos que compõem a população.

Exemplo:

- Pessoas residentes em Bauru;
- Cães que vivem em um determinado canil;
- Alunos matriculados em uma determinada disciplina

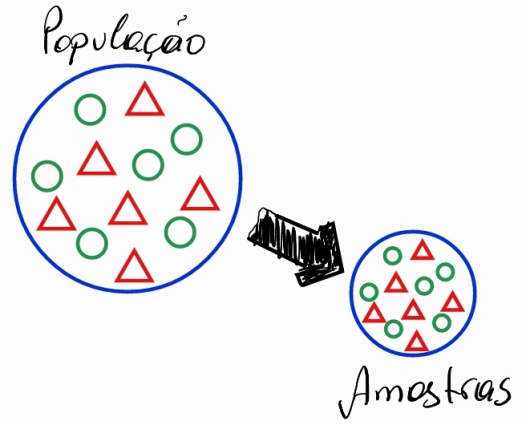


Figure: Exemplo de População X Amostras

Exemplo:

- Os elementos de uma população podem apresentar inúmeras condições ou características (variáveis) que podem ser observadas, contadas ou medidas
  - Cor do olho
  - Peso em indivíduos
  - Estado Civil
  - Escolaridade ...

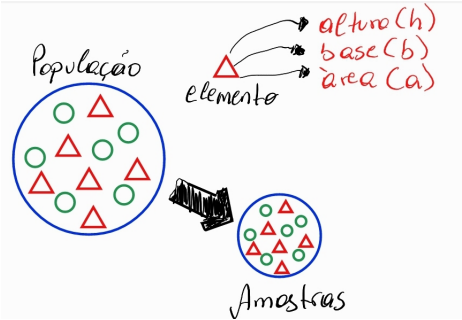


Figure: Exemplo de População X Amostras

Erro Amostral:

- **Devemos ter em mente que uma amostra nunca representará perfeitamente uma população.**

Erro Amostral:

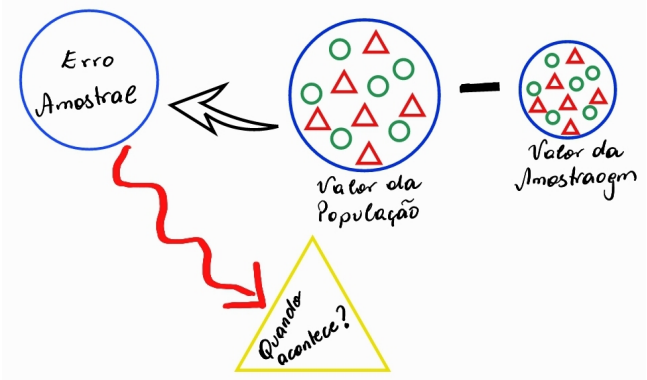
- A intensão de realizarmos a análise em amostras ao invés da população é tornar essa tarefa menos **"custosa"** em tempo e dinheiro, além de ser considerado extremamente desnecessário se utilizarmos as técnicas corretas;
- Uma amostra nunca irá representar perfeitamente uma população.
- O **erro** é representado pela diferença entre o resultado amostral e o verdadeiro valor populacional;

□ Problema:

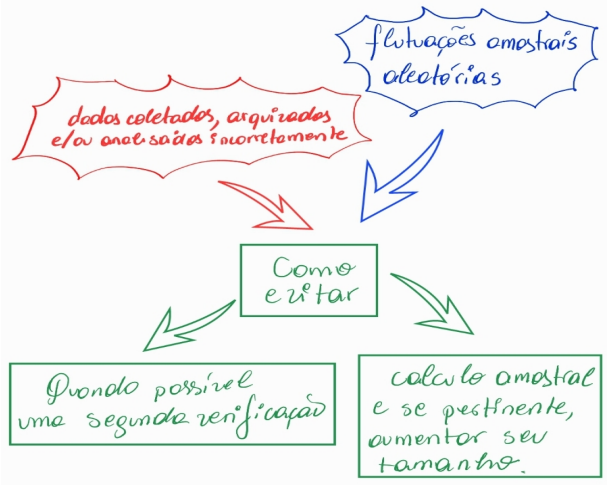
- Se inferirmos sobre características da população a partir dos dados amostras, pode-nos levar a tirar conclusões erradas.

Erro Amostral:

- Um fator óbvio é que: quanto **menor** o erro amostral, melhor.



Erro Amostral:



Entendendo as Variáveis

- Os dados são as informações que você coleta para aprender, tirar conclusões e testar hipóteses;
- O tipo de informação determina o que você pode aprender com ela;
- Diferentes tipos de variáveis permitem registram diversos tipos de informações
- As variáveis podem ser classificadas em dois principais tipos:
  - **Categóricas (qualitativas):** Quando os dados são distribuídos em categorias.
  - **Numéricas (quantitativas):** Quando os dados são expressos por números.

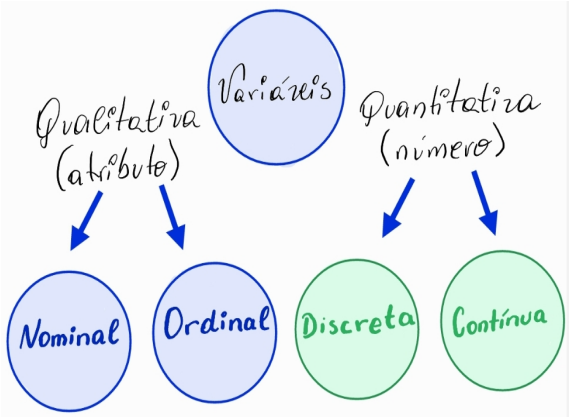
Variáveis Qualitativas vs Quantitativas

- **Variáveis qualitativas**: As informações representam características que você não mede com números. Em vez disso, as observações caem dentro de um número finito de grupos.
  - **Ordinais**: Quando os dados são distribuídos em categorias COM ordenação, ex: Grau de gravidade de uma doença, escolaridade.
  - **Nominais**: Quando os dados são distribuídos em categorias SEM ordenação, ex: Presença de um sintoma

Variáveis Qualitativas vs Quantitativas

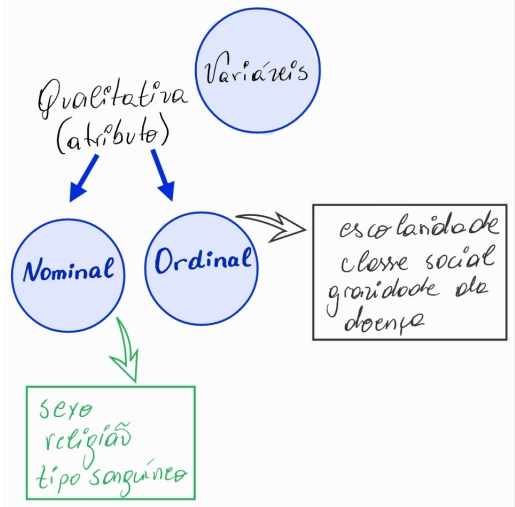
- **Variáveis quantitativas**: As informações são registradas como números e representam uma medição objetiva ou uma contagem
  - **Discretas:** Quando representa uma contagem, assumindo valores inteiros, ex: Número de cirurgias, número de filhos.
  - **Contínuas:** Quando representa uma medição, podendo assumir valores fracionários, ex: Idade, Pressão Arterial

Variáveis Qualitativas vs Quantitativas



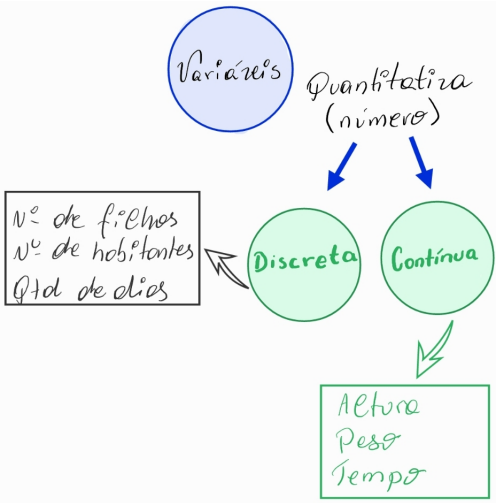
Exemplo:

- Apesar de não indicado na maioria das vezes, podemos transformar variáveis quantitativas em qualitativas.



Exemplo:

- Caso não queira usar os dados em seu estado bruto, quantitativo, podemos transformar a variável em qualitativa categorizando-a em grupos.



Distribuição de Frequência

- Organizamos nossos dados em ordem crescente para que possamos encontrar algumas medidas de frequência no comportamento dos dados:

```
In [5]:  1 freq

Out[5]: array([[12.58, 12.97, 13.51, 13.53, 14.47],
               [14.51, 14.53, 14.58, 15.17, 15.23],
               [15.29, 15.37, 15.83, 15.98, 16.01],
               [16.11, 16.47, 16.83, 16.97, 17.05]])
```

Distribuição de Frequência Buscamos então pelos seguintes valores:

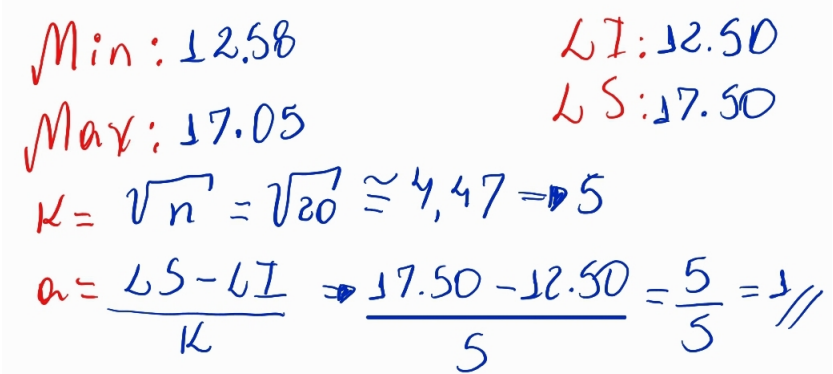
- Min:
- Max:
- LI - (Limite Inferior):
  - Utilizado para calcular a amplitude entre as classes, sendo esse uma aproximação pouco menor que o valor mínimo
- LS - (Limite Superior):
  - Mesma situação do LI porém, agora com uma aproximação pouco maior que o valor máximo
- k - Número de classes (regra de Sturges):
  - Qual a quantidade de agrupamentos possíveis dentro do nosso conjunto de dados
- a - Amplitude entre as classes k

Distribuição de Frequência (Regra de Sturges)

- A **regra Sturges** é um critério usado para determinar o número de classes ou intervalos necessários para representar graficamente um conjunto de dados estatísticos. Esta regra foi enunciada em 1926 pelo matemático alemão *Herbert Sturges*.
- A regra Sturges é amplamente usada, especialmente na área de estatística, especificamente para criar histogramas de frequência.

Distribuição de Frequência

- Teremos então os seguintes resultados:



Ou também poderíamos fazer assim...

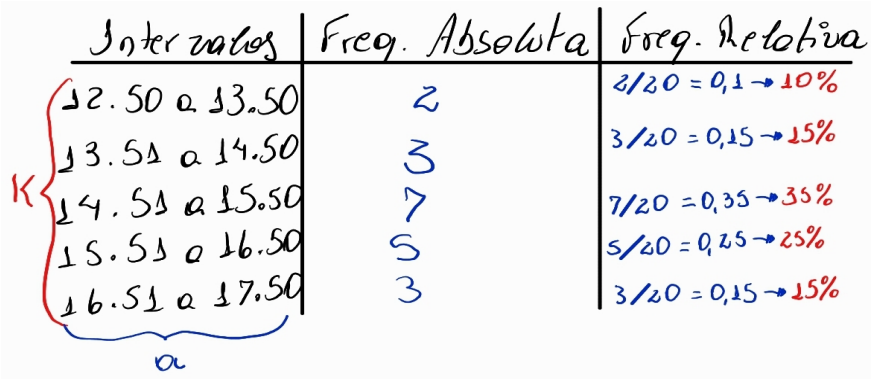
$$k = 1 + \frac{10}{3}\log_{10} n$$

In [53]:
```
1 n = len(freq2)
2 k = 1+(10/3)*np.log10(n)
3 k
```

Out[53]: 5.336766652213271

Distribuição de Frequência

- Teremos então os seguintes resultados:



Distribuição de Frequência

- Se fizermos isso com o comando **cut**:

In [38]:
```
freq2 = [12.58, 12.97, 13.51, 13.53, 14.47, 14.51, 14.53, 14.58, 15.17, 15.23,
                15.29, 15.37, 15.83, 15.98, 16.01, 16.11, 16.47, 16.83, 16.97, 17.05]
```

In [39]:
```
classes = [12.50, 13.50, 14.50, 15.50, 16.50, 17.50 ]
```

In [40]:
```
labels = ['A','B','C','D','E']
```

In [41]:
```
pd.cut(x = freq2,
       bins=classes,
       labels=labels,
       include_lowest=False)
```

Out[41]:
```
['A', 'A', 'B', 'B', 'B', ..., 'D', 'D', 'E', 'E', 'E']
Length: 20
Categories (5, object): ['A' < 'B' < 'C' < 'D' < 'E']
```

Distribuição de Frequência

- Vamos conferir?

In [42]:
```
pd.value_counts(pd.cut(x = freq2,
      bins=classes,
      labels=labels,
      include_lowest=False))
```

Out[42]:
```
C    7
D    5
B    3
E    3
A    2
dtype: int64
```

Distribuição de Frequência com Percentual

- Vamos conferir?

In [43]:
```
percent = pd.value_counts(pd.cut(x = freq2,
      bins=classes,
      labels=labels,
      include_lowest=False),
    normalize=True)
percent
```

Out[43]:
```
C    0.35
D    0.25
B    0.15
E    0.15
A    0.10
dtype: float64
```

Distribuição de Frequência com Percentual

- Agora em um Dataframe

In [49]:

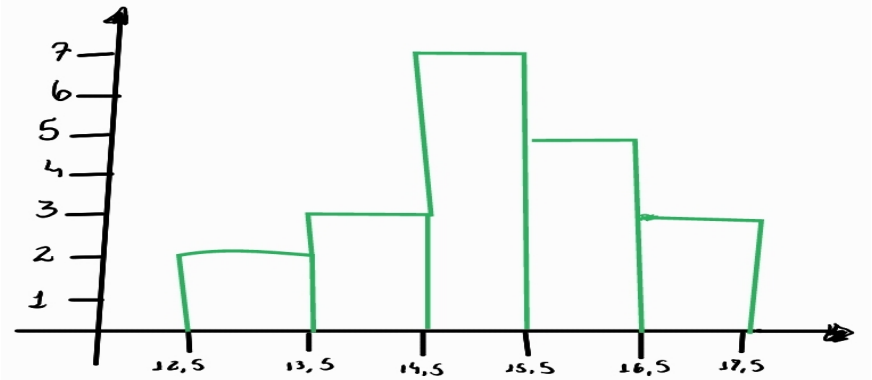
```
# Transformando em um Dataframe

frame_freq = pd.DataFrame({'Frequência':freq_all, 'Porcentagem %':percent})
frame_freq.sort_index(ascending=True)
```

Out[49]:

	Frequência	Porcentagem %
A	2	0.10
B	3	0.15
C	7	0.35
D	5	0.25
E	3	0.15

Distribuição de Frequência com Percentual

- Agora a representação em um Histograma

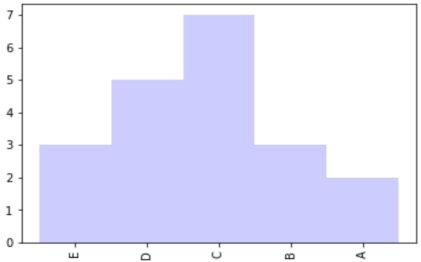


Distribuição de Frequência Python

- Em um Histograma gerado pelo Seaborn

```
In [73]:
1 frame_freq['Frequência'].sort_index(ascending=False).plot.bar(width= 1,
2         color= 'blue', alpha = 0.2, figsize= (6, 4))
```

```
Out[73]: <matplotlib.axes._subplots.AxesSubplot at 0x7f22aa6e53c8>
```

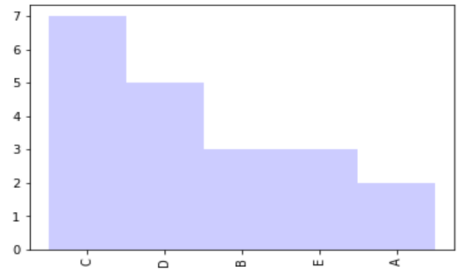


Distribuição de Frequência Python

- Em um Histograma gerado pelo Seaborn ordenado

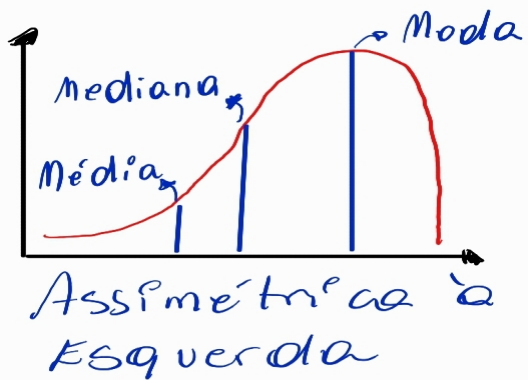
```
In [70]:
1 frame_freq['Frequência'].plot.bar(width= 1,
2         color= 'blue', alpha = 0.2, figsize= (6, 4))
```

```
Out[70]: <matplotlib.axes._subplots.AxesSubplot at 0x7f22aa87ac50>
```



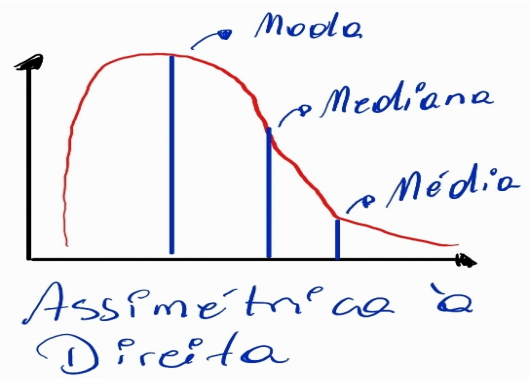
Distribuição de Frequência com Percentual

- Curva Assimétrica a Esquerda



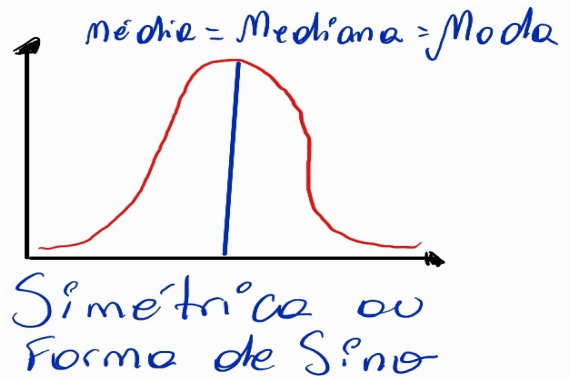
Distribuição de Frequência com Percentual

- Curva Assimétrica a Direita



Distribuição de Frequência com Percentual

- Curva Simétrica



Dúvidas?