# CEFET/RJ
# Bacharelado em Ciência da Computação
# GCC1625 - Inferência Estatística
# Trabalho 01

Prof. Eduardo Bezerra (ebezerra@cefet-rj.br)

agosto/2023

# Sumário

# 1 Teorema do Limite Central

Nesta parte, você irá realizar uma simulação computacional envolvendo a distribuição exponencial[1] e o Teorema do Limite Central.

A função de densidade de uma variável aleatória que segue a distribuição exponencial é a seguinte:

$$f(x;\lambda) = \begin{cases} \lambda e^{-(\lambda x)} & x \geq 0, \\ 0 & x < 0. \end{cases} \tag{1}$$

Na função acima, $\lambda$ é o parâmetro da distribuição exponencial, conhecido como taxa (*rate*). A média da distribuição exponencial é $1/\lambda$ e o desvio padrão é também $1/\lambda$.

A geração de números aleatórios que seguem a distribuição exponencial pode ser simulada tanto em `R` quanto em `Python`; veja a Listagem 1 e a Listagem 2. Nessas duas listagens considere o seguinte:

- lambda indica o parâmetro da distribuição exponencial;
- $n$ indica a quantidade de valores a gerar.

Listagem 1: Geração de números aleatórios seguindo a distribuição exponencial em R.

```
amostra = rexp(n, rate = lambda)
```

Listagem 2: Geração de $n$ números aleatórios seguindo a distribuição exponencial em Python.

```
from scipy.stats import expon
amostra = expon.rvs(size = n, scale=1/lambda)
```

Em sua investigação, defina o valor de lambda como igual a 0,2 para todas as simulações que você realizar. Sua investigação deve abranger a distribuição de médias de 40 exponenciais i.e., o tamanho de suas amostras deve ser $n = 40$). Além disso, sua investigação deve usar 1000 simulações (i.e., a quantidade de amostras deve ser igual a 1000).

[1] `https://en.wikipedia.org/wiki/Exponential_distribution`

(i) Crie uma amostra de tamanho 1000 a partir da distribuição exponencial, usando $\lambda = 0.2$. Em seguida, crie um histograma com os elementos dessa amostra.

(ii) Crie um histograma da distribuição amostral para a variável $\overline{x}$, a média amostral (*sampling distribution of the mean*). Construa esse histograma usando os dados resultantes das 1000 simulações. O gráfico que você deve produzir deve ser semelhante ao apresentado na Figura 1. Como sugestão, use ou a biblioteca `matplotlib`[2] (para a linguagem Python) ou a biblioteca `ggplot`[3] (para a linguagem R) para produzir esse gráfico. Repare que, assim como a figura abaixo, seu gráfico deve mostrar que a distribuição amostral é aproximadamente normal.

(iii) Agora, calcule a média e variância aproximadas para a variável $\overline{x}$ e use o TLC para obter aproximações para a média e a variância da população subjacente. Os valores que você obteve são próximos aos valores teóricos? Explique.

# 2 Distribuição amostral da média amostral

Considere uma população de sacos de batatas de 2Kg cada. Nessa população, considere que a *característica* (i.e., a *variável*) de interesse é a quantidade de batatas contida em cada saco. Sendo assim, a população subjacente corresponde a um conjunto de valores numéricos inteiros positivos (correspondentes às quantidades de batatas em cada saco).

Nesta parte do trabalho, você irá abordar a situação descrita acima como um problema de Probabilidade em vez de um problema de Estatística. Em um problema de Probabilidade, normalmente supõe-se que é conhecida a distribuição da população (o que não acontece em um problema de Estatística). Sendo assim, considere que a variável aleatória correspondente é a quantidade de batatas em um saco da população, e que essa variável segue uma distribuição uniforme discreta nos inteiros de 5 até 15. Isso significa que em cada saco podem ser encontradas no mínimo 5 e no máximo 15 batatas. Significa também que cada saco de batatas tem igual probabilidade de conter 5, 11, ..., 15 batatas.

[2] `https://matplotlib.org`
[3] `https://ggplot2.tidyverse.org`

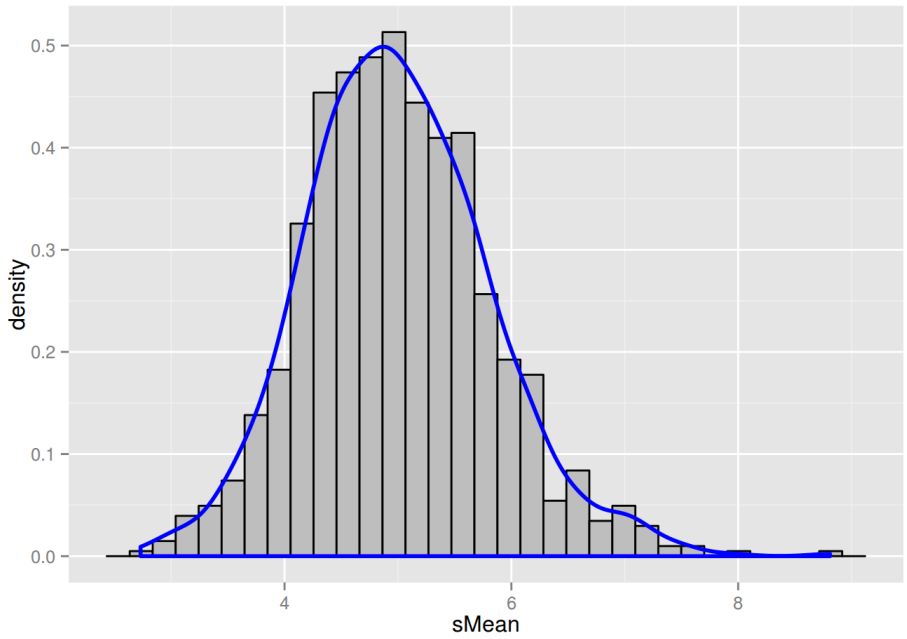


Figura 1: Modelo para a distribuição amostral de $\overline{x}$

(i) Esboce um gráfico (histograma) da distribuição da população com relação à característica *quantidade de batatas*. Dica: observe que esta população é modelada por uma distribuição discreta uniforme. Veja detalhes nesta entrada da Wikipedia: `https://en.wikipedia.org/wiki/Discrete_uniform_distribution`.

(ii) Encontre a média ($\mu$), a variância ($\sigma^2$) e o desvio padrão ($\sigma$) da distribuição da população. Dica: nome mesmo link da Wikipédia fornecido no item acima, você encontra as fórmulas que você deve usar para computar os valores solicitados neste item.

(iii) Considere que o plano amostral utilizado é uma amostragem aleatória simples feita com substituição (*simple random sample with replacement*). Considere tomar amostras de tamanho 2 dessa população e calcular a média de cada amostra. Feito isso, para cada amostra, você vai ter calculado uma *estatística pontual* (*point statistic*) da variável $\overline{x}$. Se você fizer isso para **todas** as possíveis amostras aleatórias de tamanho 2, a distribuição de todas as estatísticas pontuais resultantes

é denominada *distribuição amostral da média* (*sampling distribution of the sample mean*) para $n = 2$. Para a maioria dos casos práticos, dada uma população, não é possível gerar todas as possíveis amostras (de determinado tamanho) dessa população. Entretanto, para este caso particular, isso é possível, porque a população é finita e pequena. Sendo assim, encontre essa distribuição aplicando os passos a seguir.

- Primeiro, implemente uma função para produzir a lista de todos os possíveis valores da estatística. Para isso, gere todas as amostras possíveis de $n = 2$ elementos. Uma amostra possível é o par (10, 10); outra amostra possível é o par (15,12). Em seguida, para cada amostra gerada, compute a média de seus dois elementos. Por exemplo, para as duas amostras anteriores, as médias são 10 e 13,5, respectivamente. Repare que várias amostras diferentes geram o mesmo valor para a média. Por exemplo, as amostras (12, 15), (13, 14), (14, 13), (15, 12) todas geram a mesma média 13,5.
- Em seguida, usando o resultado da função acima, esboce um gráfico (histograma) da distribuição amostral de $\bar{x}$.

(iv) Encontre a média ($\mu_{\overline{x}}$), a variância ($\sigma^2_{\overline{x}}$) e o desvio padrão ($\sigma_{\overline{x}}$) da distribuição amostral da média amostral para $n = 2$. Dica: Compute esses valores aplicando as funções `numpy.mean` e `numpy.var` em Python (ou `mean` e `var` em R). Aplique essas funções sobre a distribuição que você produziu no item (iii).

(v) O Teorema do Limite Central apresenta uma teoria sobre os valores do desvio padrão e da média da distribuição amostral da média amostral $\overline{x}$. Use essa teoria e os resultados do item (ii) acima para encontrar o desvio padrão e a média da distribuição amostral da média para $n = 2$. Como forma de validação, você deve encontrar os mesmos valores encontrados no item (iv).

(vi) Suponha, por um momento, que você não conhece a distribuição de probabilidades da população e que deseja estimar a média da população a partir de uma amostra de tamanho 2 tomada aleatoriamente dessa população. Qual estatística você calcularia sobre essa amostra para estimar a média da população? Você acha que essa estatística seria um

bom estimador da população? Que outra estatística poderia ser um melhor estimador? Por quê?

(vii) Suponhamos que estivéssemos interessados na distribuição amostral da média para amostras de tamanho $n = 9$ e queiramos realizar os mesmos passos que em (iii) e (iv) acima.

(a) Para gerar a distribuição amostral teórica, teríamos que gerar todas as possíveis amostras de tamanho $n = 9$. Quantas amostras possíveis de tamanho $n = 9$ existem? Essa tarefa (gerar a distribuição amostral teórica neste caso) é factível de ser feita manualmente, ou mesmo usando um computador?

(b) Crie um histograma da distribuição amostral empírica (que é uma aproximação da distribuição amostral teórica). Dica: gere uma quantidade grande de amostras para obter uma aproximação adequada.

# 3 Distribuição amostral da diferença de médias

Considere que um pesquisador tenha desenvolvido um medicamento que supostamente melhora a memória. Considere duas populações hipotéticas: o desempenho das pessoas em um teste de memória se elas tiverem tomado o medicamento e o desempenho das pessoas se não tiverem. Suponha que a média ($\mu_1$) e a variância ($\sigma_1^2$) da distribuição das pessoas que tomam o medicamento sejam 50 e 25, respectivamente, e que a média ($\mu_2$) e a variância ($\sigma_2^2$) da distribuição das pessoas que não tomam o medicamento sejam 40 e 24, respectivamente. Segue-se que o medicamento, em média, melhora o desempenho no teste de memória em 10 pontos. Essa melhora de 10 pontos é para toda a população. Agora, considere a distribuição amostral da diferença entre as médias. Essa distribuição pode ser entendida pensando no seguinte plano amostral:

1. Produzir uma amostra de $n_1$ escores da população de pessoas que tomam o medicamento e computar a média. Essa média será designada como $M_1$.

2. Em seguida, produzir uma amostra de $n_2$ escores da população de pessoas que não tomam o medicamento e computar a média. Essa média será designada como $M_2$.

3. Calcular a diferença entre $M_1$ e $M_2$. Essa diferença será chamada de $M_d$, onde o $d$ significa "diferença". Esta é o estatística cuja distribuição amostral é de interesse.

A distribuição amostral pode ser aproximada repetindo o plano amostral acima várias vezes e plotando os valores de $M_d$. A distribuição de frequência (histograma) resultante seria uma aproximação da distribuição amostral. A média ($\mu_{M_d}$) e a variância ($\sigma^2_{M_d}$) da distribuição amostral de $M_d$ são:

$$\mu_{M_d} = \mu_1 - \mu_2$$

$$\sigma^2_{M_d} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Para o exemplo atual,

$$\mu_{M_d} = 50 - 40 = 10$$
$$\sigma^2_{M_d} = \frac{25}{n_1} + \frac{24}{n_2}$$

Se $n_1 = 10$ e $n_2 = 8$, então

$$\sigma^2_{M_d} = \frac{25}{10} + \frac{24}{8} = 5.5$$

Finalmente, o erro padrão de $M_d$ é simplesmente a raiz quadrada da variância da distribuição amostral de $M_d$:

$$\sigma_{M_d} = \sqrt{\frac{25}{10} + \frac{24}{8}} = 2.35$$

Agora, com base nas informações fornecidas acima, responda aos itens a seguir.

(i) Utilizando o plano amostral descrito acima, produza um histograma que aproxima a distribuição amostral de $M_d$. *Dica: repita os passos do plano amostra descrito 10000 vezes. Em seguida, crie o histograma solicitado com os 10000 valores produzidos.*

(ii) Usando a aproximação da distribuição amostral de $M_d$ obtida no item anterior, compute aproximações para a média e o desvio padrão dessa estatística. Os valores que você obteve são próximos aos fornecidos acima? Explique.

(iii) Uma vez conhecidos a média e o erro padrão da distribuição amostral de uma estatística, é possível responder a diversas perguntas. Para o caso da estatística aqui mencionada (*diferença entre as médias*), responda a seguinte pergunta: *Se um experimento com o medicamento para a memória descrito for realizado, qual é a probabilidade de a média do grupo de 10 sujeitos que receberam o medicamento ser 15 ou mais pontos maior do que a média dos 8 sujeitos que não receberam o medicamento?*

# 4 Intervalo de confiança para média populacional

Imagine que você seja um candidato a emprego tentando apresentar suas habilidades a um recrutador em um processo seletivo de uma empresa. Em qual das duas condições abaixo você teria mais chances de conseguir o emprego?

- você grava um áudio com um breve discurso descrevendo suas habilidades para o recrutador;
- você escreve um breve discurso para que o recrutador o leia.

A questão de pesquisa acima foi levantada por Schroeder and Epley (2015). Nesse artigo, os autores concluíram que a maneira como uma pessoa fala (ou seja, tom vocal, cadência, etc.) comunica informações sobre seu intelecto melhor do que suas palavras escritas (mesmo que sejam as mesmas palavras usadas no discurso falado).

Para examinar a questão de pesquisa descrita acima, os autores designaram aleatoriamente 39 recrutadores profissionais de empresas da Fortune 500[4] para uma de duas condições.

- Na *condição de áudio*, os participantes ouvem gravações de áudio do discurso falado de um candidato a emprego.

[4] `https://en.wikipedia.org/wiki/Fortune_500`

- Na *condição de transcrição*, os participantes lêem uma transcrição do discurso do candidato a emprego.

Depois de ouvir ou ler o discurso, os participantes classificaram os candidatos a emprego em três dimensões: inteligência, competência e foco. Essas classificações foram então usadas para criar uma única medida do intelecto do candidato, com pontuações mais altas indicando que os recrutadores classificaram os candidatos como superiores em intelecto. Os participantes também avaliaram sua impressão geral do candidato ao emprego (uma combinação de dois itens medindo impressões positivas e negativas). Por fim, os participantes indicaram qual o potencial de recomendar a contratação do candidato (0 - nada provável, 10 - extremamente provável).

O conjunto de dados fornecido para realizar essa parte do trabalho está no arquivo `SchroederEpley2015data.txt`. Nesse conjunto de dados, há várias colunas. Contudo, há duas colunas de interesse para esta parte do trabalho.

- a coluna `CONDITION` indica o condição à qual cada recrutador foi alocador. O valor 1 indica que o recrutador foi alocado na *condição de áudio*; O valor 0 indica que o recrutador foi alocado na *condição de transcrição*.

- a coluna `Intellect_Rating` indica a avaliação que cada recrutador atribuiu ao candidato. Essa avaliação é um valor inteiro entre 0 e 10.

Repare que a coluna `CONDITION` permite dividir a coleção de valores de avaliações em duas amostras distintas, que vamos chamar de $S_a$ e de $S_t$. As amostras $S_a$ e $S_t$ correspondem aos valores de avaliação atribuídos por recrutadores nas condições de áudio e de transcrição, respectivamente.

Agora, com base nas informações fornecidas acima, responda aos itens a seguir.

(i) Compute a média e o tamanho (quantidade de observações) tanto para $S_a$ quanto $S_t$.

(ii) Construa um *boxplot*[5] para apresentar graficamente as duas amostras. Seu gráfico deve ser semelhante ao apresentado na Figura 2. Em Python, você pode usar a biblioteca Seaborn[6]. Em R, você pode usar

[5] `https://en.wikipedia.org/wiki/Box_plot`

[6] `https://seaborn.pydata.org/generated/seaborn.boxplot.html`

a biblioteca GGPLOT2[7]. Forneça uma análise das informações fornecidas pelo gráfico.

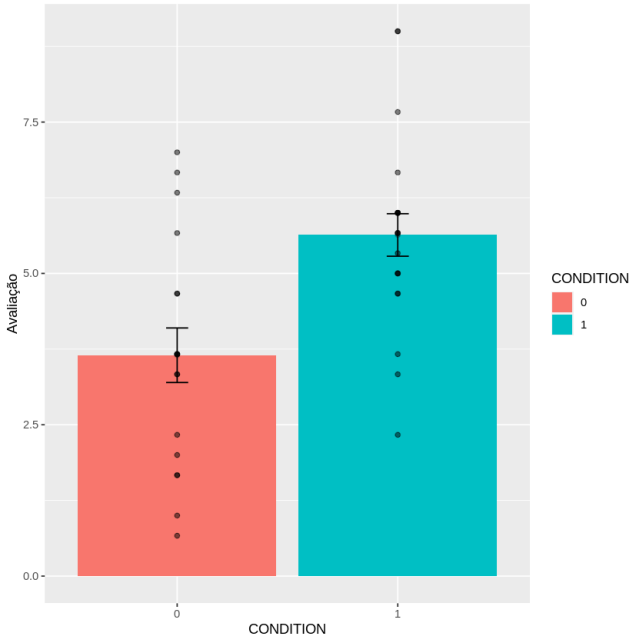


Figura 2: Boxplot para $S_a$ e $S_t$

(iii) Aplique um teste de normalidade em ambas as amostras, $S_a$ e $S_t$. Apresente uma análise do resultado obtido.

(iv) Construa um intervalo de confiança para a média de avaliações produzidas por duas populações distintas, $P_1$ e $P_2$, descritas abaixo.

- $P_1$: todos os recrutadores que ouvem os áudios
- $P_2$: todos os recrutadores que leem os transcritos

Note que a resposta a este item devem ser dois intervalos de confiança. Nos dois casos, use o nível de confiança de 99%.

(v) Com base nos intervalos de confiança obtidos no item anterior, você acha que a conclusão a que chegaram os autores em Schroeder and Epley (2015) é válida? Justifique sua resposta.

[7] `https://www.r-graph-gallery.com/boxplot.html`

# 5 Intervalo de confiança para média populacional

Essa parte do trabalho envolve computar um intervalo de confiança para a média das alturas dos estudantes de uma universidade. Você deve realizar essa parte do trabalho usando como amostra os dados contidos no conjunto de dados fornecido no arquivo `survey.csv`. Esse conjunto de dados contém o resultado de uma pesquisa feita com uma amostra de estudantes em uma universidade australiana. Os atributos desse conjunto de dados são descritos a seguir.

- `Sex`. O sexo do aluno. (Fator com os níveis *Male* e *Female*.)
- `Wr.Hnd`. vão (distância da ponta do polegar à ponta do dedo mínimo da mão aberta) da mão que escreve, em centímetros.
- `NW.Hnd`. extensão da mão que não escreve.
- W.Hnd. mão de escrita. (*Left* ou *Right*.)
- `Fold`. "Cruze os braços! Qual está no topo? (*R on L*, *L on R*, *Neither*.)
- `Pulse`. taxa de pulso do aluno (batimentos por minuto).
- `Clap`. 'Bata palmas! Qual mão está para cima? (*Right*, *Left*,*None*.)
- `Exer`. quantas vezes o aluno se exercita. (*Freq* (frequentemente), *Some*, *None*.)
- `Smoke`. Quanto o aluno fuma. (*Heavy*, *Regul* (regularmente), *Occas* (occasionalmente), *Never*.)
- `Height`. altura do aluno em centímetros.
- M.I. se o aluno expressou a altura em unidades imperiais (pés/polegadas) ou métricas (centímetros/metros). (*Metric*, *Imperial*.)
- `Age`. Idade do aluno em anos.

No R, o conjunto de dados *survey* pertence ao pacote `MASS`, que deve ser pré-carregado no espaço de trabalho do `R` antes da sua utilização. Para isso, utilize os comandos na Listagem 3.

Listagem 3: Carga do conjunto de dados *survey* em R.

```
library(MASS)    # faz a carga do pacote MASS
head(survey)     # detalhes acerca desse conjunto de
   dados
help(survey)     # documentacao associada
```

No Python, o arquivo `survey.csv` deve ser carregado inicialmente. Você pode fazer isso por meio da biblioteca Pandas, conforme exemplo na Listagem 4.

Listagem 4: Carga do conjunto de dados *survey* em Python.

```
import pandas as pd
df_survey = pd.read_csv('survey.csv')
df_survey.head()
```

A variável de interesse está na coluna `Height` do conjunto de dados. Inicialmente, você deve eliminar valores faltantes (*missing values*) nessa coluna. Para isso, pesquise sobre a função `na.omit`[8] da linguagem R, ou sobre a função `drop.na`[9] da biblioteca Pandas em Python.

Agora, realize o que se pede a seguir.

(i) Usando a distribuição t de Student, calcule um intervalo de confiança no nível de 95% para a altura média dos estudantes da universidade.

(ii) Construa outro intervalo de confiança, desta vez usando o $z$-score (em vez do $t$-score que você usou anteriormente).

(iii) Apresente uma análise comparativa dos dois intervalos de confiança obtidos.

# 6 Intervalo de confiança para proporção populacional

Nesta parte do trabalho, você deve considerar a mesma amostra contida no conjunto de dados *survey*. Dessa vez, você deveproduzir um intervalo

[8] `https://www.rdocumentation.org/packages/data.table/versions/1.13.2/topics/na.omit.data.table`

[9] `https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.dropna.html`

de confiança para a proporção de alunos da universidade que são canhotos. Sendo assim, dessa vez a coluna de interesse é `W.Hnd`.

(i) Primeiramente, se certifique de que existem pelo menos 10 estudantes destros e pelo menos 10 estudantes canhotos, para que você possa realizar a construção do intervalo de confiança de forma satisfatória.

(ii) Produza o intervalo de confiança solicitado, usando o nível de confiança 90%. Junto com o resultado, forneça também uma análise.

(iii) Produza o intervalo de confiança solicitado, usando o nível de confiança 95%. Junto com o resultado, forneça também uma análise e compare com o resultado obtido no item anterior.

(iv) Repita os itens (i), (ii) e (iii), desta vez considerando o atributo `Sex`.

(v) A amostra correspondente ao conjunto de dados *survey* tem tamanho suficiente para produzir um intervalo de confiança para a característica `Sexo` (coluna `Sexo`) com um erro amostral de 2 pontos percentuais? Se sim, construa esse intervalo de confiança usando nível de confiança 90%. Se não, explique.

# Especificação da entrega

Você pode desenvolver esse trabalho em duas linguagens alternativas, `R` ou `Python`. Independente da linguagem que escolher, você deve preparar um explicar sua implementação, análise e conclusões de cada parte desse trabalho. Além disso, certifique-se de fornecer respostas para cada uma das perguntas formuladas em cada parte deste trabalho.

Seu trabalho deve necessariamente ser produzido como um único *notebook* Jupyter[10]. Como sugestão, você pode usar a plataforma Google Colab[11] para produzir seu trabalho. Essa plataforma permite criar *notebooks* em ambas as linguagens.

Você deve necessariamente organizar seu *notebook* em seções que reflitam as seções apresentadas no enunciado deste trabalho. Sendo assim, use como ponto de partida o exemplo apresentado na Figura 3. Repare que, para cada item do trabalho, você deve transcrever o enunciado correspondente para o notebook. Deve também criar duas células, uma de código e outra de texto, para apresentar a solução para o item.

**IMPORTANTE**: Tão relevante quanto a implementação (seja em R ou Python) de cada parte deste trabalho é sua explicação sobre ela. Nesse sentido, você deve também apresentar suas análises e conclusões para cada item do trabalho.

- Um item que apresente apenas código (em R ou em Python), sem a explicação do mesmo, não receberá a totalidade da pontuação correspondente.
- Um item que apresente apenas um valor numérico como resposta (ou apenas um "sim" ou "não"), sem uma descrição sobre como a resposta foi obtida, não receberá a totalidade da pontuação correspondente.

O *notebook* Jupyter resultante do seu trabalho deve necessariamente ser definido com nome que siga o padrão `IE_T1_SEU_NOME_COMPLETO.ipynb`. Um exemplo: `IE_T1_EDUARDO_BEZERRA_DA_SILVA.ipynb`. Siga à risca esse padrão de nomenclatura.

[10] `http://jupyter.org/`

[11] `https://colab.research.google.com`

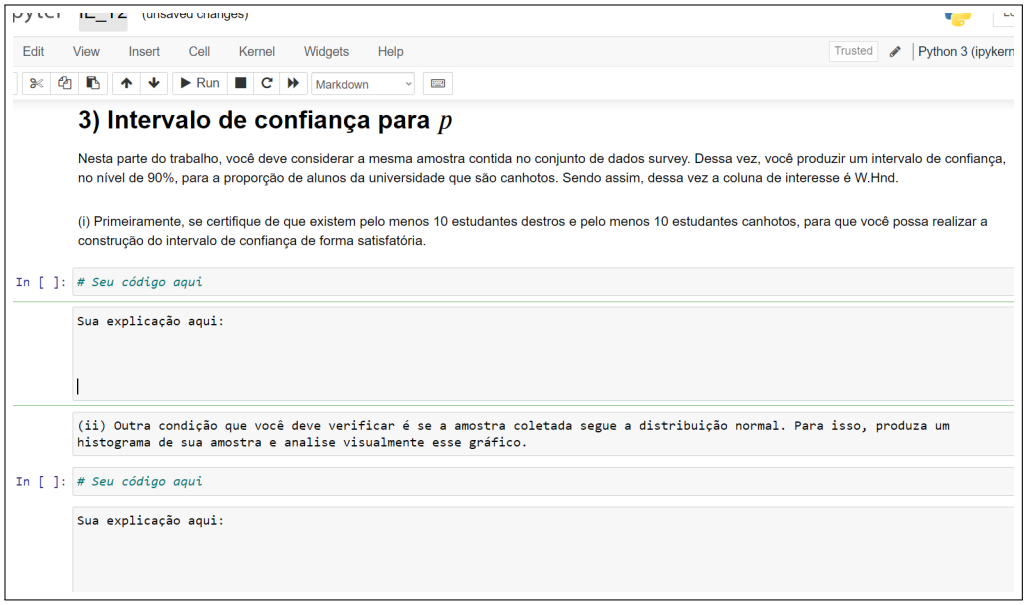


Figura 3: Modelo a ser seguido para apresentação da solução de cada parte do trabalho.

# Referências

Juliana Schroeder and Nicholas Epley. The sound of intellect: Speech reveals a thoughtful mind, increasing a job candidate's appeal. *Psychological Science*, 26(6):877–891, 2015. doi: 10.1177/0956797615572906. URL `https://doi.org/10.1177/0956797615572906`. PMID: 25926479.