GCC1625 - INFERÊNCIA ESTATÍSTICA

Trabalho 2

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

CEFET/RJ
Escola de Informática e Computação
Bacharelado em Ciência da Computação

Outubro/2025

Sumário

1 Máquina desregulada?

Considere que uma máquina de preenchimento de garrafas está configurada para preencher cada garrafa com 500 ml de vinho. O fabricante suspeita de que essa máquina está preenchendo as garrafas com valores a menor. Ele então coleta uma amostra de 20 garrafas preenchidas a partir da linha de produção e mede cuidadosamente o volume em cada uma delas. Os resultados obtidos nessas medições estão a seguir (valores em miligramas):

484.11, 459.49, 471.38, 512.01, 494.48, 528.63, 493.64, 485.03, 473.88, 501.59, 502.85, 538.08, 465.68, 495.03, 475.32, 529.41, 518.13, 464.32, 449.08, 489.27

Nessa parte do trabalho você deve verificar a alegação de que a máquina está desregulada, conforme a suspeita do fabricante.

(i) Os procedimentos paramétricos de testes de hipóteses sobre uma amostra que estudamos em nosso curso presumem que a amostra a ser usada é proveniente de uma distribuição aproximadamente normal. Para a amostra fornecida verifique essa hipótese usando dois dos testes de normalidade que estudamos no curso, o teste de Shapiro-Wilk e o teste usando QQ-plot. Nessa verificação defina o nível de significância como $\alpha = 0.05$.

(ii) Agora, aplique o teste de hipóteses. Repare que a variância da população é desconhecida; sendo assim, certifique-se de usar o procedimento de teste de hipóteses adequado para essa situação. Apresente sua análise para o nível de significância igual a 1%. Apresente e descreva claramente a aplicação dos quatro passos do procedimento.

2 Um estilo diferente ajuda?

O conjunto de dados fornecido no arquivo golfe.csv lista as pontuações de uma rodada para 75 membros selecionados aleatoriamente em um campo de golfe, primeiro usando seus próprios tacos originais e, dois meses depois, usando novos tacos com um estilo (*design*) experimental. Queremos verificar a alegação de que esse novo estilo de taco ajuda no desempenho dos jogadores.

(i) Os procedimentos paramétricos de testes de hipóteses sobre duas amostras que estudamos em nosso curso presumem que as amostras a serem usadas são provenientes de uma distribuição aproximadamente normal. Para cada uma das amostras fornecidas, verifique essa hipótese usando dois dos testes de normalidades que estudamos no curso. Para isso, use um nível de significância igual a 5%.

(ii) Agora, aplique o teste de hipóteses propriamente dito. Apresente e descreva claramente a aplicação dos quatro passos do procedimento. Repare que a variância da população não é conhecida. Sendo assim, certifique-se de usar o teste de hipóteses adequado para essa situação. Apresente sua análise para o nível de significância igual a 1%.

3 Transmissões: automática *versus* manual

O conjunto de dados mtcars apresenta informações sobre modelos de carros. Esse conjunto de dados contém várias variáveis. Entretanto para essa parte do trabalho, duas delas são relevantes:

- `am` - essa variável categórica indica o tipo de sistema de transmissão usado no modelo de automóvel (0 = automatic, 1 = manual). Com base nessa variável, podemos formar duas amostras independentes: veículos com transmissão automática e veículos com transmissão manual.
- `mpg` (*miles per gallon*) — variável quantitativa contínua que mede o consumo de combustível, indicando quantas milhas o veículo percorre por galão de combustível. Valores maiores indicam maior eficiência energética.

Utilizando essas informações, responda aos itens a seguir com base nos métodos estudados em aula.

(i) Realize uma análise gráfica exploratória comparando o consumo médio de combustível dos dois grupos de veículos. Utilize ao menos dois tipos de gráfico (como boxplots e histogramas sobrepostos) e comente os padrões observados.

(ii) Os procedimentos paramétricos de testes de hipóteses sobre duas amostras que estudamos em nosso curso presumem que as amostras a serem usadas são provenientes de uma distribuição aproximadamente normal. Para cada uma das amostras fornecidas, verifique essa hipótese usando dois dos testes de normalidades que estudamos no curso. Para isso, use um nível de significância igual a 5%.

(iii) Presumindo que os dados da coluna `mpg` seguem a distribuição normal, determine um intervalo de confiança no nível 95% da diferença entre as médias dos modelos de carros que possuem transmissão automática e os que possuem transmissão manual.

(iv) Suponha que, ao analisar os dados fornecidos, alguém levantou a alegação de que carros automáticos e manuais não apresentam a mesma eficiência relativa a consumo de combustível. Em particular, existe a suspeita de que carros com transmissão manual consumem (em média) menos combustível que suas contra-partidas com transmissão automática. Aplique um teste de hipóteses para verificar essa suspeita. Apresente e descreva claramente a aplicação dos quatro passos do procedimento. Repare que são fornecidas duas amostras independentes. Sendo assim, certifique-se de usar o teste de hipóteses adequado para essa situação. Apresente sua análise para o nível de significância igual a 5%.

4 Teste ANOVA

4.1 Situação-problema I

Nesta parte, você deve estudar as condições que permitem usar o método de análise de variância (ANOVA) para determinar se um grupo de populações tem uma média comum. Os dados apresentados nas figuras 1, 2 e 3 correspondem às estimativas de milhas percorridas por galão obtidas para amostras de modelos de carros de 1993, conforme relatado pelo *Consumer Reports: The 1993 Cars - Annual Auto Issue* (abril de 1993).

(i) Se uma ou mais das amostras não passam no teste de normalidade, então não podemos usar o método ANOVA. Sendo assim, antes de poder usar esse teste, você deve verificar as condições de aplicabilidade dele. Primeiro verifique, se as amostras são aproximadamente normalmente distribuídas. Em seguida, usando o teste de Levene, verifique se as amostras possuem variâncias iguais do ponto de vista estatístico.

Audi 90 -- 20
Chevy Cavalier -- 25
Chevy Corsica -- 25
Chrysler LeBaron -- 20
Dodge Spirit -- 22
Ford Tempo -- 22
Honda Accord -- 24
Mazda 626 -- 26
Mercedes-Benz 190E -- 20
Nissan Altima -- 24
Olds Achieva -- 24
Pontiac Sunbird -- 23
Saab 900 -- 20
Subaru Legacy -- 23
Volkswagen Passat -- 21
Volvo 240 -- 21

Figura 1: Carros compactos - modelo e MPG

(ii) Agora que você verificou as condições aplicabilidade do ANOVA, aplique esse teste para verificar a hipótese nula de que as três populações têm médias estatisticamente iguais. Declare as hipóteses nula e alternativa. A seguir, descreva sua conclusão. Use nível de significância de 5%.

Para sua comodidade, as observações das amostras (para cada uma das três categorias) são fornecidas na Listagem 1 como variáveis *numpy array*.

Listing 1: Amostras para aplicação do ANOVA.

```
compactos = np.array ([20, 25, 25, 20, 22, 22, 24, 26, 20, 24, 24, 23, 20,
    23, 21, 21])

medios = np.array ([18, 19, 22, 22, 19, 16, 21, 21, 21, 20,17, 18, 18, 17,
    19, 19, 18, 21, 23, 19, 22, 20])

grandes = np.array ([19, 16, 16, 17, 20, 20, 20, 18, 18, 19, 19, 15, 18,
   17, 15, 18, 17, 18, 18, 17])
```

4.2 Situação-problema II

Em http://www.flatworldknowledge.com/sites/all/files/data9.xls, você encontra dados que registram os custos dos materiais (livro didático, manual de solução, taxas de laboratório e assim por diante) em cada um dos dez cursos diferentes em cada um dos três assuntos diferentes, química, ciência da computação e matemática. Verifique, ao nível de significância

Acura Legend -- 18
Audi 100 -- 19
BMW 535i -- 22
Buick Century -- 22
Buick Riviera -- 19
Cadillac Seville -- 16
Chevy Lumina -- 21
Dodge Dynasty -- 21
Ford Taurus -- 21
Hyundai Sonata -- 20
Infiniti Q45 -- 17
Lexus ES300 -- 18
Lexus SC300 -- 18
Lincoln Continental -- 17
Mercedes-Benz 300E -- 19
Mercury Cougar -- 19
Mitsubishi Diamante -- 18
Nissan Maxima -- 21
Olds Cutlass Ciera -- 23
Pontiac Grand Prix -- 19
Toyota Camry -- 22
Volvo 850 -- 20

Figura 2: Carros médios - modelo e MPG

Buick LeSabre -- 19
Buick Roadmaster -- 16
Cadillac Deville -- 16
Chevy Caprice -- 17
Chrysler Concorde -- 20
Chrysler Imperial -- 20
Eagle Vision -- 20
Ford Crown Victoria -- 18
Lincoln TownCar -- 18
Olds Eighty-Eight -- 19
Pontiac Bonneville -- 19
Chevy Astro -- 15
Chevy Lumina APV -- 18
Dodge Caravan -- 17
Ford Aerostar -- 15
Mazda MPV -- 18
Nissan Quest -- 17
Olds Silhouette -- 18
Toyota Previa -- 18
Volkswagen Eurova -- 17

Figura 3: Carros grandes e Vans - modelo e MPG

de 1%, se os dados fornecem evidências suficientes para concluir que os custos médios nas três disciplinas não são todos iguais. Apresente os detalhes de aplicação todos os quatro passos do teste de hipóteses.

5 Testes Qui-quadrado

5.1 Distribuição χ^2

Suponha que uma variável aleatória Y siga a distribuição χ^2 com k graus de liberdade, isto é, $Y \sim \chi^2_{(k)}$. Por meio das funções apropriadas de `R`[1] ou de `Python`[2], compute o que se pede a seguir. Considere que $k = 13$.

(i) $\Pr(Y > 2,56)$

(ii) $\Pr(2,56 < Y < 4,87)$

(iii) O valor de y tal que $\Pr(Y < y) = 0,95$

5.2 Teste χ^2 (situação-problema I)

Considere novamente o conjunto de dados denominado `mtcars`. Para este conjunto de dados, descubra se as variáveis `cyl` e `carb` são ou não dependentes. Para isso, utilize o teste χ^2. Apresente o desenvolvimento, isto é, os comandos em `R` ou `Python` que você utilizou para chegar à conclusão.

5.3 Teste χ^2 (situação-problema II)

A fabricante das balinhas Zuzuba produz balinhas de diferentes cores. Esse fabricante alega que cada pacote produzido contém quantidades de balinhas de cada cor que não diferem significativamente das que são apresentadas na segunda coluna da tabela Tabela 1. Para testar essa alegação, um auditor comprou um pacote de Zuzubas em uma loja perto de sua casa e contou as quantidades de cada cor. Os dados levantados pelo auditor estão na terceira coluna da Tabela 1. Se o fabricante estiver correto, então não deve haver diferença significativa entre as quantidades de diversas cores de Zuzubas que ela alega depositar em cada pacote e as quantidades que o auditor encontrou.

cor	**esperado**	**observado**
vermelho	18	24
verde	19	16
roxo	16	13
azul	6	20
laranja	24	20
amarelo	17	14

Tabela 1: Valores esperados e observados para diferentes cores de balinhas Zuzuba.

[1] https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/Chisquare
[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2.html

(i) Utilizando o `R`[3] ou `Python`[4], apresente dois gráficos de setores (*pie charts*), um para as quantidades esperadas e outro para as quantidades observadas de cores.

(ii) Também utilizando `R` ou `Python`, teste a hipótese nula de que a alegação do fabricante é verdadeira, usando nível de significância $\alpha = 0,05$. Apresente os seguintes valores: graus de liberdade, valor da estatística e o valor-p. Apresente também a sua conclusão, contra ou a favor da hipótese nula, justificando sua resposta.

5.4 Teste χ^2 (situação-problema III)

Considere um caso hipotético em que se deseja testar a eficácia de um medicamento para um determinado problema médico. Suponha que temos 105 pacientes em estudo e 50 deles foram tratados com a droga. Os restantes 55 pacientes foram mantidos como amostras de controle. O estado de saúde de todos os pacientes foi verificado após uma semana. Os dados e resultados para todos esse indivíduos podem ser encontrados no arquivo `treatment.csv`[5].

A tabela de contingência para o conjunto de dados fornecido pode ser produzida por meio dos comandos na Listagem 2 e na Listagem 3.

Listing 2: Produção de uma tabela de contingência com R.

```
url <- "treatment.csv"
df <- read.csv(url)
tbl = table(df$treatment, df$improvement)
```

Listing 3: Produção de uma tabela de contingência com R.

```
import pandas as pd
data = pd.read_csv('treatment.csv')
data_crosstab = pd.crosstab(data['treatment'],
                            data['improvement'],
                               margins = False)
print(data_crosstab)
```

Nesta situação problema, temos duas variáveis discretas, uma que indica se o paciente foi tratado com o medicamento (*treated* ou *not-treated*), e outra que indica se o paciente melhorou ou não (*improved* ou *not-improved*) Utilizando `R` ou `Python`, verifique a alegação de que as duas variáveis são dependentes, usando nível de significância $\alpha = 0,05$.

5.5 Teste χ^2 (situação-problema IV)

Em http://www.flatworldknowledge.com/sites/all/files/data4.xls, está disponível um conjunto de dados que registra o resultado de 500 arremessos de um dado de seis lados. Verifique, ao nível de significância de 10%, se há evidência suficiente para concluir que o dado não é "justo" (ou "balanceado"), ou seja, que a distribuição de probabilidade difere da probabilidade 1/6 para cada das seis faces do dado. Forneça a descrição detalhada dos quatro passos de aplicação do teste.

[3] https://www.statmethods.net/graphs/pie.html

[4] https://matplotlib.org/3.1.1/gallery/pie_and_polar_charts/pie_features.html

[5] Dados obtidos no seguinte endereço: https://raw.githubusercontent.com/selva86/datasets/master/treatment.csv

6 Teste de Permutação

A Figura 4 mostra os resultados de um experimento no qual 7 de 16 camundongos foram selecionados aleatoriamente para receber um novo tratamento médico, enquanto os 9 restantes foram atribuídos ao grupo sem tratamento (controle). O tratamento tinha como objetivo prolongar a sobrevivência após uma cirurgia de teste. Em particular, a coluna "Data" mostra o tempo de sobrevivência após a cirurgia, em dias, para todos os 16 camundongos. Essa mesma figura também apresenta, para cada amostra: tamanho, média, desvio padrão.

Utilize o teste de permutação para responder à seguinte pergunta de pesquisa (use nível de significância igual a 5%): *O tratamento prolongou a sobrevivência?*. Você deve apresentar a declaração das hipóteses, descreva como calculou a estatística de deste e o p-valor, a finalmente apresente sua conclusão.

Group	Data			Sample Size	Mean	Estimated Standard Error
Treatment	94	197	16			
	38	99	141			
	23			(7)	86.86	25.24
Control	52	104	146			
	10	50	31			
	40	27	46	(9)	56.22	14.14
				Difference:	30.63	28.93

Figura 4: Mouse data - retirada de [1]

7 Bootstrap

Essa parte do trabalho é uma adaptação do Problema 9 na seção 5.4 de *An Introduction to Statistical Learning*[6]. O conjunto de dados usado aqui é o denominado Boston. Uma descrição desse conjunto de dados pode ser encontrada em `http://lib.stat.cmu.edu/datasets/boston`.

(i) Com base neste conjunto de dados, forneça uma estimativa pontual para a média populacional da variável medv. Chame essa estimativa $\hat{\mu}$.

(ii) Forneça uma estimativa do erro padrão de $\hat{\mu}$. Interprete o resultado.

(iii) Agora estime o erro padrão de $\hat{\mu}$ usando o método Bootstrap. Como essa estimativa se compara com sua resposta de (ii)?

[6] `https://www.statlearning.com`

(iv) Com base em sua estimativa de bootstrap de (iii), forneça um intervalo de confiança de 95% para a média de medv. Compare-o com os resultados obtidos usando t-test sobre o atributo `medv`.

(v) Com base neste conjunto de dados, forneça uma estimativa, $\hat{\mu}_{med}$, para a mediana populacional de `medv`.

(vi) Agora você deve estimar o erro padrão de $\hat{\mu}_{med}$. Infelizmente, não há uma fórmula simples para calcular o erro padrão da mediana. Em vez disso, estime o erro padrão da mediana usando o método bootstrap. Comente suas descobertas.

(vii) Forneça uma estimativa para o décimo percentil do atributo `medv`. Chame essa quantidade de $\hat{\mu}_{0.1}$.

(viii) Use o método bootstrap para estimar o erro padrão de $\hat{\mu}_{0.1}$. Comente suas descobertas.

8 Teste de Permutação *vs* Bootstrap

Uma empresa quer saber se é eficiente ensinar novas ferramentas aos seus funcionários usando cursos pela internet. A empresa seleciona aleatoriamente 7 trabalhadores e os atribui a dois grupos de tamanhos 4 e 3. O primeiro grupo frequentou aulas tradicionais, e o segundo frequentou cursos pela internet. Após a realização dos cursos, foi aplicado um teste aos trabalhadores, cujos resultados foram:

- Cursos na Internet: 37, 49, 55, 57
- Cursos tradicionais: 23, 31, 46

Verifique se os cursos da Internet são mais efetivos do que os cursos tradicionais. Para isso, aplique um teste de permutação e um teste de bootstrap. Use o nível de significância $\alpha = 0.1$. Os dois testes levam à mesma conclusão?

9 Especificação da entrega

1. 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.

2. Já é fornecido junto com este enunciado um notebook Jupyter para você usar como ponto de partida neste trabalho. Você encontra esse arquivo nesse link.

3. Certifique-se de fornecer respostas para cada uma das perguntas formuladas em cada parte deste trabalho.

4. Seu trabalho deve necessariamente ser produzido como um único *notebook* Jupyter[7]. Como sugestão, você pode usar a plataforma Google Colab[8] para produzir seu trabalho. Essa plataforma permite criar *notebooks* em ambas as linguagens.

5. Eventuais conjuntos de dados necessários neste trabalho podem ser encontrados no repositório do curso[9]. Referencie diretamente esses conjuntos de dados a partir do código que produzir em seu notebook Jupyter.

6. 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 6. 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.

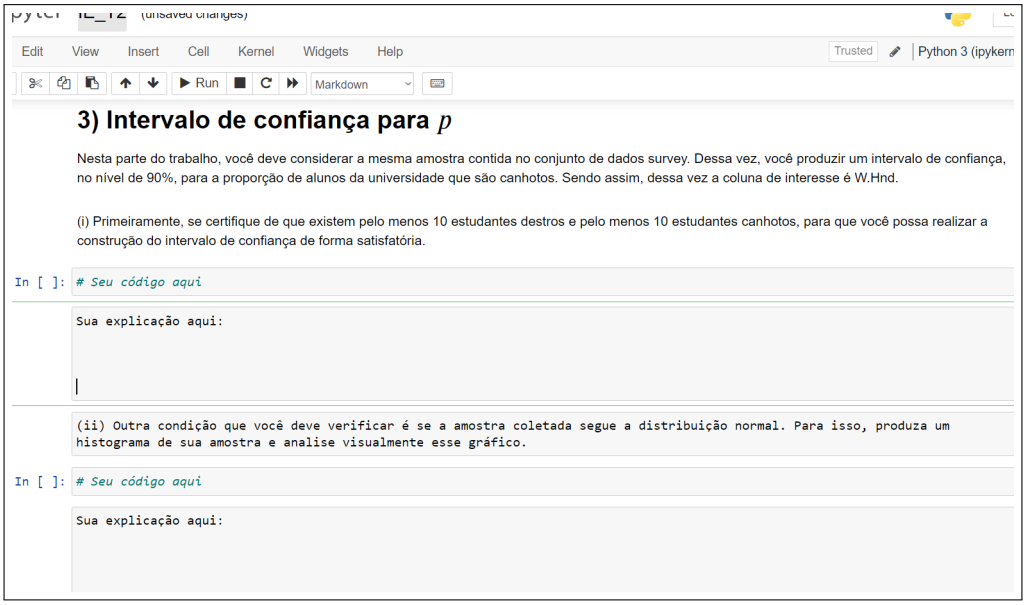

Figura 5: Modelo a ser seguido para apresentação da solução de cada parte do trabalho.

[7] http://jupyter.org/
[8] https://colab.research.google.com
[9] https://github.com/AILAB-CEFET-RJ/gcc1625/tree/main/data

7. Você deve também elaborar um vídeo (cuja duração aproximada foi especificada no primeiro dia de aula) no qual você deve explicar os aspectos mais importantes de cada parte do seu trabalho. Nesse vídeo, você também deve demonstrar a execução de cada parte e apresentar uma análise dos resultados obtidos. O link para acesso a esse vídeo deve estar contido na primeira célula (de texto) do notebook Jupyter.

8. 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.

9. O *notebook* Jupyter resultante do seu trabalho deve ser definido com nome que siga o padrão de nomenclatura `GCC1625_T2_SEU_NOME_COMPLETO.ipynb`. Um exemplo de uso dessa padrão: `GCC1625_T2_EDUARDO_BEZERRA_DA_SILVA.ipynb`. Siga à risca esse padrão de nomenclatura.

O que deve ser entregue

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 6. 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.

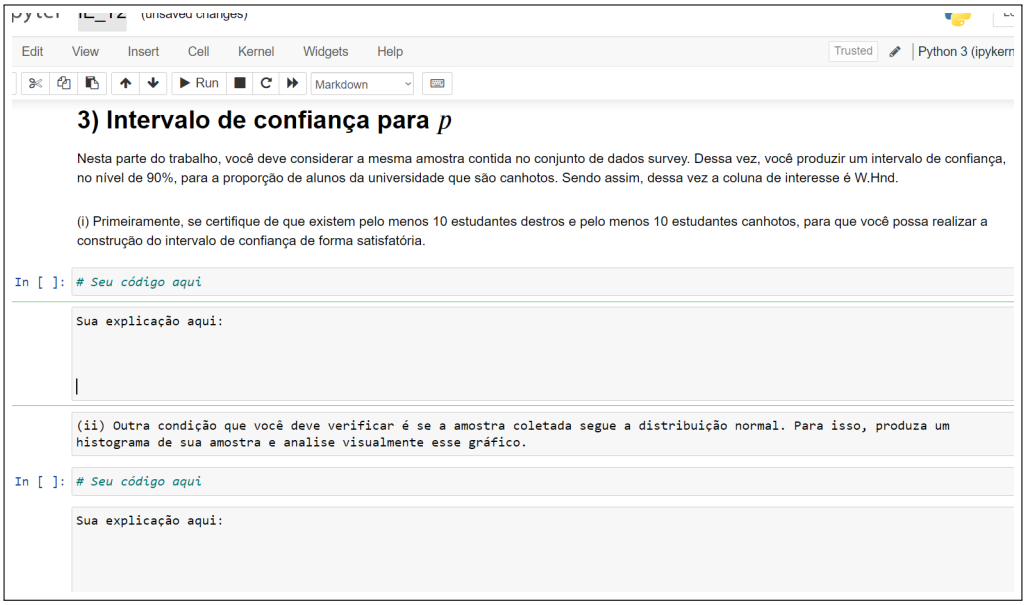

Figura 6: Modelo a ser seguido para apresentação da solução de cada parte do trabalho.

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.

[10] http://jupyter.org/
[11] https://colab.research.google.com

O *notebook* Jupyter resultante do seu trabalho deve necessariamente ser definido com nome que siga o padrão `GCC1625_T2_SEU_NOME_COMPLETO.ipynb`. Um exemplo: `GCC1625_T2_EDUARDO_BEZERRA_DA_` Siga à risca esse padrão de nomenclatura.

Referências

[1] Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. CRC press, 1994.