Materiais :

Q: Qual o link para os materiais do professor Filipe?

R: Os materiais utilizados pelo professor Filipe encontram-se em:

Códigos:

https://github.com/filipezabala/fdepcdd

Apresentação\Livros:

http://filipezabala.com/fdepcdd/

http://www.filipezabala.com/materiais/ecnrs.pdf.

Q: Onde na apostila eu encontro o capítulo 7.1 correlação. Ocorre que acessei no link http://filipezabala.com/fdepcdd/ e o menu é outro.

R: O professor Zabala acessa uma apostila de estatística clássica, que se encontra neste link: http://filipezabala.com/materiais/ecnrs.pdf na página 101.

Q: Estou seguindo o livro no http://filipezabala.com/fdepcdd/ e fazendo os exercícios por lá. Porém, as respostas que deveria estar no capítulo 10 não aparecem. Existe algum problema no site? Tentei usar o livro disponível em http://filipezabala.com/materiais/ecnrs.pdf mas não existe uma perfeita sincronia no conteúdo entre o site e o livro.

R: O material online ainda está sendo elaborado, portanto ainda não tenho todas as soluções. Nesse caso ao longo das resoluções o professor irá atualizando o material.

Q: O Link para o livro do John Neter está quebrado. Há outro lugar que disponibilize o livro?

R: Realmente o link externo foi removido, mas o link da UNICAMP disponibiliza o livro: https://www.ime.unicamp.br/~dias/John%20Neter%20Applied%20linear%20regression%20models.pdf?msclkid=66aaa682b8d011ec8f4cfa591ce1539b.

Ferramentas:

Q: Ao utilizar o R é necessário instalar pacotes adicionais para determinadas funções?

R: O R base contém um conjunto de pacotes que permite realizar uma série de cálculos e visualizações.

Os pacotes adicionais, tal como "coronavirus" disponibilizado pela universidade John Hopkins, devem ser instalados pela função "install.packages" ou mesmo no menu "Packages.Install" do RStudio.

No caso do 'coronavirus', por ser atualizado diariamente seu uso deve sempre ser precedido pelo comando de atualização 'coronavirus::update_dataset(silence = FALSE)'.

Q: Estou tendo problema para instalar o pacote *desempatetecnico* , como procedo?

R: Para evitar erros de instalação de pacotes deve-se tentar instalar ou abrir o R como administrador. Outra possibilidade pode ser também passar uns parâmetros no comando, segundo esse link do StackOverflow: https://stackoverflow.com/questions/63056269/error-lazy-loading-failed-for-package-rstan-on-centos-7

Conceitos e Exercícios:

Q: O professor Rômulo apresentou alguns cases de estudo sobre empresas e como conseguiram reduzir seus gastos, mas qual a média de valores poupados ou de renda adquirida?

R: Segundo o professor Rômulo, estes ganhos variam bastante, houve ganhos de 200 mil reais mês a 2 milhões de reais mês, em média.

Q: Para a tomada de decisão podemos utilizar o p-value comparando com a significância do teste.

Exemplo: com uma taxa de significância em 5% (0.05) e um p-value = 0.08 podemos aceitar H0.

Como o poder do teste pode influenciar essa aceitação em H0 com base na comparação feita acima?

R: Se p for maior que o nível de significância, idealmente a decisão seria "manter", "reter" ou "não rejeitar H0". Uma boa forma de visualizar a associação entre as componentes dos procedimentos de significância é neste site: https://rpsychologist.com/d3/nhst/

Q: Onde posso encontrar a resolução do exercício do coronavírus disponibilizado pelo professor Filipe?

R: As respostas dos exercícios propostos pelo professor Filipe podem ser encontradas a partir da página 140 do material disponível em: http://filipezabala.com/materiais/ecnrs.pdf

Q: Como eu poderia estar classificando a variável da letra "a. Participação de mercado (market share)" do exercício 2.1?

R: A participação de mercado ('market share') é um valor em percentual, pois vem da divisão da parte associada a uma empresa em questão (que pode ser medida em valores monetários, número de clientes, cidades/estados de atuação etc.) pelo total do mercado. Sendo assim, a classificação mais apropriada é quantitativa (é um número) contínua (pois admite casas decimais, em termos laicos).

Q: Referente aos Exemplos 2.45 e 2.46 das páginas 33 e 34 do material (http://filipezabala.com/materiais/ecnrs.pdf), por que há um código com fator de correção na fórmula que não possui fator de correção?

R: A função var do R calcula a variância amostral, aquela que divide por n-1 ou possui fator de correção. Por este motivo não preciso definir o fator de correção ao utilizá-la para calcular var.a.

Já para o cálculo da variância populacional var.p, como não há uma função específica para isso na linguagem precisei multiplicar a variância amostral obtida via var pelo inverso do fator de correção, ou (1/fator de correção).

Essa discussão fica resumida na Equação (30) da página 34.

Q: Ao quebrar uma tabela e criar uma quantidade de classificações o que seria o tamanho da classificação?

R: O conceito do final do vídeo da parte 3 da aula 3, é o agrupamento dos valores em intervalos de classe para variáveis contínuas. Para isso o professor verifica o melhor intervalo de cada grupo (tamanho da classificação) de forma que a quantidade de amostras em cada grupo fique balanceada.

Q: Em uma amostra de 5 empresas brasileiras de importação de rolamentos, constatou-se que elas gastaram R$65,000,000.00 (sessenta e cinco milhões de reais) em compra de rolamentos

da China.

a) Qual a estimativa por ponto do gasto médio das importações de rolamentos de empresas do ramo no Brasil?

b) Sabendo que o desvio padrão amostral de R$1,500,000.00, encontre um intervalo de 90% de confiança para o gasto médio das importações de rolamentos de empresas do ramo no Brasil.

R: O valor de 1.4 MM saiu da tabela t de *Student*

com 5-1=4 graus de liberdade para alfa = 5%, sabe-se que t=2.132. Assim a margem de erro é dada por: e = 2.132*1500000/sqrt(5)

= 1430189 ≈ 1.4MM

Q: Para uma população normal com variância conhecida, responda:

a) Qual o nível de confiança para o intervalo x ± 2.14σ√n?

b) Quais os valores de z que levam a um intervalo de 94% de confiança?

R: Temos IC para média com σ conhecido, logo a distribuição subjacente é normal. Portanto no item a temos que a confiança 1-α = Φ(2.14) - Φ(-2.14) ≈ 0.9838-0.0162 = 0.9676.

No item b, como temos 94% de confiança, temos 6% de "desconfiança", 3% para cada lado. Assim da tabela de normal, |z|≈1.88. Mais precisamente qnorm(.03) = -1.880794, ou qnorm(.97)=1.880794.

Q: No exercício 2.2, letra b, o que está sendo pedido sobre interpretar o x(4).

R: Um rol indica um vetor\lista ordenado conforme está no material: Quando se ordenam estes dados – em ordem crescente ou decrescente – obtém-se um rol, dando origem às estatísticas de ordem. A cada passo ou índice desse vetor\rol, há um valor associado, como no exemplo, há um índice indicando a quantidade de passos até a lixeira.

No exercício o professor propôs uma lista: 10, − 4 , 5 , 7 , 1 , 3 , 9 . Nesse caso, você deve obter um rol (letra a) e depois interpretar o passo\índice 4.

Q: A respeito do *p-hacking* , do p - *value* , o valor do " *ro* " deu muito baixo, com 10^-9.

Como rejeito o Ho e fico com o H1, se está mais próximo de 0?

R: Na verdade, o que o professor Filipe Zabala está explicando é uma forma de contornar a métrica *p-value (p-hacking)*. Como ele explica, você fixa um alfa ("número mágico" ou número de significância 5%), caso esteja abaixo desse número alfa, rejeita-se o H0, pois assim, há uma hipótese de diferença, ou seja, o tratamento teve efeito. O *p-hacking* então, aumenta o valor da amostra, pois o *p-value* é extremamente sensível ao aumento da amostra, quanto maior o valor de amostra, menor será o *p-value*.

Q: Tive dúvidas no exercício extra 9. letra a) Como calcular o nível de confiança 1-alfa

R: A confiança é a área entre -2.14 e +2.14 em uma normal padrão (pois o desvio padrão populacional é conhecido). Logo Φ(2.14)-Φ(-2.14)=0.9838-0.0162=0.9676.

Q: Resolução do exercício 8. da seção 5.4.

R: Primeiramente uma correção: "O peso líquido MÉDIO encontrado foi de 15.95g..."

As hipóteses são H0: \mu = 16.0g vs H1: \mu ≠ 16.0 (bilateral).

A estatística de teste é Tt = (15.95-16)/(0.15/√(32)) ≈ -1.89.

Como -2.03 < -1.89 < +2.03, não se rejeita H0.

Pela tabela abaixo, pode-se notar que 1.89 está entre os valores 1.697/1.684 (referente a 5% entre 30 e 40 gl) e 2.042/2.021 (referente a 2.5% entre 30 e 40 gl). Assim pode-se dizer que 0.025 < Pr(T > 1.89) < 0.05, portanto o valor-p (pelo fato de o teste ser bilateral) está entre 0.025*2 = 5% e 0.05*2 = 10%.

De forma exta, valor-p = 2*Pr(T > 1.89) ≈ 0.0684 = 6.84%.

Código R:

Seção 5.4, Ex. 8

n <- 32

xbar <- 15.95

s <- 0.15

mu0 <- 16.0

alfa <- 0.05

(Tt <- (xbar-mu0)/(s/sqrt(n)))

qt(.025, n-1)

2*pt(Tt, n-1)

Q: Como fazer o exercício 7.3 visto que, as variáveis não são binárias? Como aplico um modelo de regressão logística com estas variáveis? (*idade, tipoTrabalho, educação, anosestudo*)

Exercício 7.3: Considere o conjunto de dados apresentado por Ronny Kohavi e Barry Becker, disponível em https://archive.ics.uci.edu/ml/datasets/adult . Considere um modelo de regressão logística para avaliar as características que mais impactam no salário das pessoas (acima ou abaixo de 50 mil dólares).

R: Nesse caso, é possível utilizar variáveis do tipo numéricas como a de idade e *anosestudo* assim como é possível criar intervalos para elas e em seguida utilizar variáveis *hot-encode (true\false)*. Quanto a variáveis de tipo também, é possível utilizar hot encode e criar categorias a elas.

Q: No capítulo 2 - Estatística Descritiva, o professor deixa exercícios 2.1 ao 2.7 para praticarmos o conteúdo apresentado na aula, mas não tem nenhum material auxiliar onde eu possa comparar o exercício correto com as minhas respostas, assim eu observar onde errei ou não?

R: O material online ainda está sendo elaborado, portanto ainda não tenho todas as soluções. Nesse caso ao longo das resoluções o professor irá atualizando o material. As respostas que contém estão a partir da página 140 do material http://filipezabala.com/materiais/ecnrs.pdf .

Q: A Curva Laffer representação o % versus o tempo, correto? No exemplo da aula o X esta representando o tempo e o Y esta representando o %, pode ser o inverso?

R: Sim.

Q: Na parte de Tabela de frequência univariada contínua estou com uma dúvida. Há um trecho que mostra a tabela agrupada pela regra de Sturges. Entendi como achar em quantos grupos como a regra sugere dividir porém, não entendi como achar quais os grupos em si. Por exemplo, na tabela mostrada há os valores: 1.50 ⊢ 1.55, 1.55 ⊢ 1.60, etc... Não localizei onde fala sobre como definir quais devem ser as classes de amplitude.

R: Sobre a amplitude referida, é a Amplitude de classes, e é explicada em "Amplitude (h) e quantidade (k) de classes" que utiliza a amplitude das amostras também.

Q: Sobre o exercício 2.2.Considere o conjunto de dados 10,−4,5,7,1,3,9.A. Obtenha o rol Levando em consideração que o ROL é a ordenação dos dados coletados em ordem crescente ou decrescente, quando não é explicito na questão qual ordenação deve-se realizar?

R: Geralmente se usa em ordenação ascendente.

**FAQ gerado com base em comentários até o dia 21/04/2022.*