Trabalho No.2

CE313 - Estatística não-paramétrica

2 de julho de 2024

Redigir de maneira individual e entregar na área correspondente no sistema **Microsoft Teans** um relatório eletrônico até o dia **24 de julho de 2024**.

1. Um grupo de pesquisadores na Austrália conduziu uma intervenção dietética curta, de três semanas, em um experimento controlado randomizado. No estudo, 75 estudantes em idade universitária com sintomas de depressão elevados e hábitos alimentares relativamente inadequados foram aleatoriamente designados para um grupo de intervenção com dieta saudável ou para um grupo de controle.

   Os pesquisadores registraram a mudança ao longo do período de três semanas em duas escalas numéricas diferentes de depressão, a escala CESD e a escala DASS. A escala CESD (Centro de Estudos Epidemiológicos de Depressão) baseia-se mais em observações clínicas, enquanto a DASS (Escala de Depressão, Ansiedade e Estresse) depende mais de informações autorreferidas. Eles também registraram o índice de massa corporal (IMC) no início e no final do período de 21 dias.

   Os dados estão contidos no arquivo **DietDepression.csv** e disponibilizados da seguinte forma:

```
> dados = read.csv2("http://leg.ufpr.br/~lucambio/
                                    CE313/20241S/DietDepression.csv")
```

   Fonte: Francis HM, et al., "A brief diet intervention can reduce symptoms of depression in young adults - A randomised controlled trial,"

   PLoS ONE, 14(10), October 2019.

   O arquivo de dados **dados** dispõem das seguintes informações:

   Group: Controle (Control) ou Dieta (Diet)

   CESD1: Pontuação de depressão CESD no Dia 1

   CESD21: Pontuação de depressão CESD no dia 21

   CESDDiff: Mudança na pontuação de depressão CESD

   DASS1: Pontuação de depressão DASS no Dia 1

   DASS21: Pontuação de depressão DASS no dia 21

   DASSDiff: Mudança na pontuação de depressão DASS

   BMI1: Índice de Massa Corporal (IMC) no Dia 1

   BMI21: Índice de Massa Corporal (IMC) no Dia 21

   BMIDiff: Mudança no Índice de Massa Corporal (IMC)

Utilizando técnicas de reamostragem, queremos saber:

(a) Existem diferenças entre as formas de medir a depressão CESD e DASS? Em geral e entre os grupos?

(b) As mudanças no IMC são aleatórias ou não? Em geral e entre os grupos? Existem diferenças entre os valores dos IMC entre os grupos?

2. Considere o seguinte conjunto de dados:

```
> beans = read.csv("http://leg.ufpr.br/~lucambio/
                                  CE313/20241S/Beans_Dataset.csv")
```

Sete tipos diferentes de feijão seco foram utilizados nesta pesquisa, levando em consideração características como forma, formato, tipo e estrutura da situação do mercado. Um sistema de visão computacional foi desenvolvido para distinguir sete diferentes variedades registradas de feijão seco com características semelhantes, a fim de obter uma classificação uniforme de sementes.

Para o modelo de classificação, foram obtidas imagens de 13.611 grãos de 7 feijões cadastrados diferentes com câmera de alta resolução. As imagens de feijão obtidas por sistema de visão computacional foram submetidas às etapas de segmentação e extração de características, totalizando 16 características; 12 dimensões e 4 formatos foram obtidos a partir dos grãos.

Fonte: Dry Bean Dataset. (2020). UCI Machine Learning Repository.

https://doi.org/10.24432/C50S4B.

Informações:

Area: (A): A área de uma zona de bean e o número de pixels dentro de seus limites pixels

Perimeter: (P) A circunferência do feijão é definida como o comprimento de sua borda.

MajorAxisLength: (L) A distância entre as extremidades da linha mais longa que pode ser traçada a partir de um feijão

MinorAxisLength: (l) A linha mais longa que pode ser traçada a partir do feijão perpendicular ao eixo principal.

AspectRation: (K) Define a relação entre L e l.

Eccentricity: (Ec) Excentricidade da elipse tendo os mesmos momentos da região.

ConvexArea: (C) Número de pixels no menor polígono convexo que pode conter a área de uma semente de feijão.

EquivDiameter: (Ed) O diâmetro de um círculo com a mesma área que a área de uma semente de feijão.

Extent: (Ex) A proporção entre os pixels na caixa delimitadora e a área do bean.

Solidity: (S) Também conhecida como convexidade. A proporção entre os pixels na casca convexa e aqueles encontrados nos feijões.

Roundness: (R) Calculado com a seguinte fórmula: $(4piA)/(P^2)$

Compactness: (CO) Mede a redondeza de um objeto: Ed/L

ShapeFactor1: (SF1)

ShapeFactor2: (SF2)

ShapeFactor3: (SF3)

ShapeFactor4: (SF4)

Class: SEKER, BARBUNYA, BOMBAY, CALI, DERMASON, HOROZ e SIRA

(a) Queremos verificar se a distribuição da variável Perimeter é normal, de maneira global e considerando as sub-amostras de Perimeter, segundo as diferentes categorias em Class.

(b) Percebemos que a quantidade de observações é consideravelmente grande, do qual inferimos que as funções implementando os diversos testes de bondade de ajuste estudados não devem funcionar adequadamente. Por esse motivo, propomos as seguintes alternativas de trabalho:

(i) Selecionar uma quantidade grande B de sub-amostras com reposição, digamos $B = 10000$, de tamanho 100 cada uma e verificar a bondade de ajuste à normalidade de cada sub-amostra. Contar o número se amostras não conformes com a distribuição normal, ou seja, nas quais o teste rejeita a normalidade da sub-amostra e avaliar, com essas informações, a bondade de ajuste da amostra original. Isto justifica-se teoricamente porque, dado um conjunto de variáveis aleatórias independentes $X_1, \cdots, X_n$, com distribuição normal, então, qualquer sub-coleção $X_{i_1}, \cdots, X_{i_k}$, $k < n$ também é formada por variáveis aleatórias independentes com distribuição normal.

(ii) Utilizar testes de bondade de ajuste desenvolvidos para amostras grandes, por exemplo, veja o artigo em:
leg.ufpr.br/ lucambio/CE050/20211S/normality%20test%20large%20samples.pdf
The performance of univariate goodness-of-fit tests for normality based on the empirical characteristic function in large samples, escrito por J. M. VAN ZYL (2016). Department of Mathematical Statistics and Actuarial Science, University of the Free State, Bloemfontein, South Africa.
O teste proposto neste artigo, resumidamente, requer os seguintes passos:
Padronizar os dados. Isto pode ser feito utilizando a função base R `scale`. Os dados originais $x_1, \cdots, x_n$, padronizados são obtidos definidos como $z_i = (x_i - \overline{x}_n)/\sqrt{s_n^2}$, sendo $\overline{x}_n$ e s_n^2 a média e variância amostrais.
Avaliar a função característica empírica nos dados padronizados, ou seja, avaliar a função $\widehat{\phi}_S(t) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{e}^{itz_i}$. Isto pode ser feito, por exemplo, com o auxílio da função R `ecf`, no pacote `empichar`.
Avaliar a estatística de teste

$$\nu_n(1) = \log\left(|\widehat{\phi}_S(1)/\mathrm{e}^{-1/2}|\right),$$

onde $\sqrt{n}\,\nu_n(1) \underset{n\to\infty}{\longrightarrow} N(0, 0.0431)$, assintoticamente. O valor absoluto, na expressão acima, denota o módulo de um número complexo se o argumento for complexo.
Rejeita-se a normalidade se

$$\left|\nu_n(1)/\sqrt{0.0451/n}\right| = \left|4.8158\sqrt{n}\,\nu_n(1)\right| > z_{1-\alpha/2},$$

sendo $z_{1-\alpha/2}$ o percentil da distribuição normal padrão com α nível de significância.

3. A Altimetria é a parte da Topometria que estuda os processos de medida de distâncias verticais ou diferenças de níveis entre diversos pontos do terreno. Neste estudo, foram medidas as altitudes em determinados pontos do Primeiro Planalto Paranaense. Essas medidas geraram uma planilha onde constam as altitudes e as frequências dessas altitudes. Esses dados estão disponíveis em:

```
> altimetria = read.csv("http://leg.ufpr.br/~lucambio/
                    CE313/20241S/Altimetria.csv", sep=";", header = TRUE)
```

Os dados disponíveis correspondem a medidas nas altitudes até os 410 metros em vários pontos do Primeiro Planalto Paranaense. As variáveis do conjunto são Altitude, medida em metros e Freq, a frequência ou quantidade de vezes que uma determinada altitude aparece no ponto onde foi medida, por exemplo, a latitude de 10 metros apareceu 49.946 vezes, nestes dados as altitudes aparecem em intervalos de frequência de 10 metros.

Suspeita-se da existência de dois níveis de erosão. Utilize uma técnica estatística para identificar e, se assim for, confirmar os níveis de erosão.

4. Seja $X_1, \cdots, X_n \sim$ Uniforme$(0, \theta)$, onde $\theta > 0$ desconhecido. A quantidade pivotal usual para encontrar intervalos de confiança para θ é

$$\xi_n = n(\theta - X_{(n)})/\theta \sim \text{Exponencial}(1)\cdot$$

Seja $X_1^*, \cdots, X_n^*$ uma amostra Bootstrap. O estimador de momentos de ξ_n é

$$\xi_n^* = n(X_{(n)} - X_{(n)}^*)/X_{(n)}\cdot$$

De acordo com o principio do Bootstrap, a distribuição de ξ_n^* é a distribuição Bootstrap de ξ_n.

Selecione B amostras Bootstrap e verifique a bondade do ajuste da distribuição de ξ_n^* à distribuição de referência da estatística ξ_n, quando $n = 50$ e $\theta = 1$.

Observação: A convergência à distribuição de referência é devagar, bem devagar. Sugere-se escolher valores muito altos de B, como 10.000, 20.000 ou mais, valores muito altos de B para amostras relativamente pequenas. Isto deve-se à limitações do Bootstrap quando a distribuição dos dados não é unimodal.

5. O conjunto de dados **Data-pizza.csv** contém dados de um serviço de entrega de pizza em Londres, entregando pizzas em três áreas. Cada registro define um pedido/entrega e as propriedades correspondentes. Supõe-se que uma pizza tenha um gosto bom se a temperatura for alta o suficiente, digamos 45 Celsius. Então pode ser interessante para o serviço de entrega de pizza minimizar o tempo de entrega.

Um conjunto ou quadro de dados contêm 1.209 observações sobre as 17 variáveis a seguir.

index: um vetor numérico, indexando os registros (sem faltas aqui).

date: a data de entrega

week: inteiro, o número da semana

weekday: inteiro, o dia da semana

area: fator, os três distritos de Londres: Brent, Camden, Westminster

count: número inteiro, o número de pizzas entregues

rabate: lógico, TRUE se uma gorjeta foi dada

price: numérico, o preço total da(s) pizza(s) entregue(s)

operator: um fator com níveis Allanah, Maria, Rhonda

driver: um fator com níveis Carpenter, Carter, Taylor, Butcher, Hunter, Miller, Farmer

delivery_min: numérico, o tempo de entrega em minutos (decimal)

temperature: numérico, a temperatura da pizza em graus Celsius quando entregue ao cliente

wine_ordered: número inteiro, 1 se o vinho foi pedido, 0 se não

wine_delivered: número inteiro, 1 se o vinho foi entregue, 0 se não

wrongpizza: lógico, TRUE se uma pizza errada foi entregue

quality: fator pedido com níveis low, medium, high, definindo a qualidade da pizza quando entregue

Leitura dos dados:

```
> data.pizza = read.csv("http://leg.ufpr.br/~lucambio/
                    CE313/20241S/Data-pizza.csv", sep=",", header = TRUE)
```

Queremos saber se a ocorrência de NA nas variáveis temperature e quality são aleatórias ou não, globalmente, por região e operador. Saber isso permite inferir a qualidade do serviço. Acredita-se que, caso não seja aleatória a ocorrência de dados NA, a informação esteja sendo omitida propositalmente para mascarar problemas de qualidade nas entregas.