Lista 5 - Introdução a Análise de Dados
Dataframes
Gabarito

Guilherme Masuko

May 2023

Para essa lista, utilizaremos uma base de dados própria do R chamada "Motor Trend Car Road Tests". Utilize os comandos `data(mtcars)` e `View(mtcars)` para importar e visualizar os dados respectivamente. Para conhecer melhor essa base de dados (entender o significado de cada coluna, por exemplo), deve-se usar o seguinte comando `?mtcars`.

Questão 1

Encontre, utilizando código, o carro mais econômico dessa amostra (maior milhas por galão (*mpg*)).

Solução

```
# pegando o valor máximo de milhas por galão do dataframe mtcars
max_mpg <- max(mtcars$mpg)
max_mpg

# selecionando a linha de dado que contém milhas por galão idêntico
   ao máximo
mtcars[mtcars$mpg == max_mpg, ]
```

Questão 2

Encontre, utilizando código, o carro menos econômico dessa amostra (menor milhas por galão (*mpg*)).

Solução

```
# pegando o valor mínimo de milhas por galão do dataframe mtcars
min_mpg <- min(mtcars$mpg)
min_mpg
```

```
# selecionando a linha de dado que contém milhas por galão idêntico
   ao mínimo
mtcars[mtcars$mpg == min_mpg, ]
```

Questão 3

Dentro da categoria de carros mais potentes (digamos, $hp > 120$), qual é o carro mais econômico?

Solução

```
# criando um dataframe apenas com os carros mais potentes
mtcars_mais_potentes <- mtcars[mtcars$hp > 120,]

# pegando o valor máximo de milhas por galão do dataframe
   mtcars_mais_potentes
max_mpg_potentes <- max(mtcars_mais_potentes$mpg)
max_mpg_potentes

# selecionando a linha de dado que contém milhas por galão idêntico
   ao máximo
mtcars_mais_potentes[mtcars_mais_potentes$mpg == max_mpg_potentes, ]
```

Questão 4

E dentro da categoria de carros menos potentes (digamos, $hp \leq 120$), qual é o carro menos econômico?

Solução

```
# criando um dataframe apenas com os carros menos potentes
mtcars_menos_potentes <- mtcars[mtcars$hp <= 120,]

# pegando o valor mínimo de milhas por galão do dataframe
   mtcars_menos_potentes
min_mpg_potentes <- min(mtcars_menos_potentes$mpg)
min_mpg_potentes

# selecionando a linha de dado que contém milhas por galão idêntico
   ao mínimo
mtcars_menos_potentes[mtcars_menos_potentes$mpg ==
   min_mpg_potentes, ]
```

Questão 5

Existe um *trade-off* entre economia e potência de carros? Siga as intruções a seguir para responder essa pergunta.

a) Crie uma função que recebe dois vetores de mesmo tamanho e retorne a covariância[1] (amostral) entre esses dois vetores.

A covariância (populacional) entre duas variáveis é obtida através de

$$\mathbb{C}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

Mas como nossa base não tem todos os carros do mundo, vamos utilizar um estimador para a covariância verdadeiro (populacional), esse estimador é a covariância amostral, obtida através de

$$\widehat{\mathbb{C}(X, Y)} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)$$

onde μ_X e μ_Y são as médias amostrais da variáveis X e Y, respectivamente.

Obs: O aluno deve utilizar a formula do estimador para criar a função.

Solução

```
# função
covariancia <- function (vetor1, vetor2) {
  # tamanho do vetor
  n <- length(vetor1)
  # média do primeiro vetor
  mean_1 <- mean(vetor1)
  # média do segundo vetor
  mean_2 <- mean(vetor2)

  # vetor de desvios em relação a média dos vetores 1 e 2
  desvio_vetor_1 <- vetor1 - mean_1
  desvio_vetor_2 <- vetor2 - mean_2

  # produto dos vetores de desvios
  produto_desvios <- desvio_vetor_1 * desvio_vetor_2

  # somatório
  soma <- sum(produto_desvios)
```

[1] <https://en.wikipedia.org/wiki/Covariance>

```
  # divisão
  cov <- soma/(n-1)

  # retorno
  return(cov)
}

# teste
covariancia(mtcars$mpg, mtcars$hp)

# função built-in
cov(mtcars$mpg, mtcars$hp)
```

b) Crie uma função que receba dois vetores de mesmo tamanho como parâmetros e retorne a correlação[2] (amostral) entre eles.

A correlação (populacional) entre duas variáveis é obtida através de

$$\rho_{X,Y} = \frac{\mathbb{C}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

onde σ_X e σ_Y são os devios-padrão populacionais da variáveis X e Y, respectivamente.

E seu estimador é a correlação amostral, obtida através de

$$\widehat{\rho_{X,Y}} = \frac{\widehat{\mathbb{C}(X,Y)}}{\widehat{\sigma_X} \cdot \widehat{\sigma_Y}}$$

onde $\widehat{\sigma_X}$ e $\widehat{\sigma_Y}$ são os devios-padrão amostrais da variáveis X e Y, respectivamente.

Obs: O aluno deve utilizar a formula do estimador para criar a função. **Solução**

```
# função
correlacao <- function (vetor1, vetor2) {
  # covariância
  cov <- covariancia(vetor1, vetor2)

  # desvios-padrão dos vetores 1 e 2
  sd1 <- sd(vetor1)
```

[2]<https://en.wikipedia.org/wiki/Pearson_correlation_coefficient>

```
  sd2 <- sd(vetor2)

  # correlação
  corr <- cov/(sd1*sd2)

  # retorno
  return(corr)
}
```

c) Calcule a correção entre as duas variáveis economia e potência do carro (representadas por mpg e hp, respectivamente) para responder a pergunta principal.

Solução

```
# calculando correlação
correlacao(mtcars$mpg, mtcars$hp)

# função built-in
cor(mtcars$mpg, mtcars$hp)

# teste estatístico
cor.test(mtcars$mpg, mtcars$hp)
```

Questão 6

Crie uma coluna (wt_kg) que contenha o peso de libras (wt) convertido em quilogramas. Note que a medida da coluna wt equivale a 1000 libras. A fórmula de conversão é:

$$\text{peso em kg} = \frac{\text{peso em libras}}{2.2046}$$

Solução

```
rm(list=ls())

# aplicando a fórmula e multiplicando por 1000
mtcars$wt_kg <- (mtcars$wt/2.2046) * 1000
```

Questão 7

Qual é o peso médio, em quilogramas, dos carros?

Solução

```
# tomando o peso médio em quilogramas dos carros
mean_kg <- mean(mtcars$wt_kg)
mean_kg
```

Questão 8

Qual é o peso médio, em quilogramas, dos carros automáticos ($am = 0$)?

Solução

```
# criando um dataframe apenas com carros automáticos
mtcars_auto <- mtcars[mtcars$am == 0, ]
mtcars_auto

# tomando o peso médio em quilogramas dos carros automáticos
mean_kg_auto <- mean(mtcars_auto$wt_kg)
mean_kg_auto
```

Questão 9

Qual é o peso médio, em quilogramas, dos carros manuais ($am = 1$)?

Solução

```
# criando um dataframe apenas com carros manuais
mtcars_manual <- mtcars[mtcars$am == 1, ]
mtcars_manual

# tomando o peso médio em quilogramas dos carros manuais
mean_kg_manual <- mean(mtcars_manual$wt_kg)
mean_kg_manual
```

Questão 10

Qual é a correlação entre essas duas variáveis, transmissão (am) e peso em quilogramas (wt_kg)? O que isso significa?

Solução

```
# calculando a correlação
correlacao(mtcars$am, mtcars$wt_kg)

# função built-in
cor(mtcars$am, mtcars$wt_kg)

# teste estatístico
```

```r
cor.test(mtcars$am, mtcars$wt_kg)
```

Probabilidade

A probabilidade de um evento $A \subset \Omega$ ocorrer é

$$\begin{aligned}\mathbb{P}(A) &= \frac{n(A)}{n(\Omega)} \\ &= \frac{\text{número de elementos do evento}}{\text{número de elementos do espaço amostral}}\end{aligned}$$

Questão 11

Qual a probabilidade (amostral) de pegarmos (em nossa amostra) um carro manual mais pesado que o peso médio dos carros automáticos?

Solução

```r
# quantidade de carros manuais
n_carros_manual <- nrow(mtcars_manual)

# quantidade de carros manuais mais pesados que o peso médio dos
   carros automáticos
n_carros_manual_pesados <- nrow(mtcars_manual[mtcars_manual$wt_kg >
   mean_kg_auto,])

# probabilidade em porcentagem
round((n_carros_manual_pesados / n_carros_manual) * 100, 2)
```

Questão 12

Qual a probabilidade (amostral) de pegarmos (em nossa amostra) um carro automático mais pesado que o peso médio dos carros manuais?

Solução

```r
# quantidade de carros automáticos
n_carros_auto <- nrow(mtcars_auto)

# quantidade de carros automáticos mais pesados que o peso médio
   dos carros manuais
n_carros_auto_pesados <- nrow(mtcars_manual[mtcars_auto$wt_kg >
   mean_kg_manual,])

# probabilidade em porcentagem
```

```
round((n_carros_auto_pesados / n_carros_auto) * 100, 2)
```

Questão 13

Qual a probabilidade (amostral) de pegarmos (em nossa amostra) um carro manual mais pesado que o carro automático mais leve?

Solução

```
# peso mínimo dos carros automáticos
min_kg_auto <- min(mtcars_auto$wt_kg)
min_kg_auto

# quantidade de carros manuais mais pesados que o carro automático
   mais leve
n_carros_manual_pesados <- nrow(mtcars_manual[mtcars_manual$wt_kg >
   min_kg_auto,])
n_carros_manual_pesados

# probabilidade em porcentagem
round((n_carros_manual_pesados / n_carros_manual) * 100, 2)
```

Questão 14

Qual a probabilidade (amostral) de pegarmos (em nossa amostra) um carro automático mais leve que o carro manual mais pesado?

Solução

```
# peso máximo dos carros manuais
max_kg_manual <- max(mtcars_manual$wt_kg)
max_kg_manual

# quantidade de carros automáticos mais leve que o carro manual
   mais pesado
n_carros_auto_leves <- nrow(mtcars_manual[mtcars_auto$wt_kg <
   max_kg_manual,])
n_carros_auto_leves

# probabilidade em porcentagem
round((n_carros_auto_leves / n_carros_auto) * 100, 2)
```