Learning Week3

211294029 李梦麟

2023-03-07

## 问题一 Rivers数据集

1. 打印（print）这个数据集；

print(rivers)

## [1] 735 320 325 392 524 450 1459 135 465 600 330 336 280 315 870  
## [16] 906 202 329 290 1000 600 505 1450 840 1243 890 350 407 286 280  
## [31] 525 720 390 250 327 230 265 850 210 630 260 230 360 730 600  
## [46] 306 390 420 291 710 340 217 281 352 259 250 470 680 570 350  
## [61] 300 560 900 625 332 2348 1171 3710 2315 2533 780 280 410 460 260  
## [76] 255 431 350 760 618 338 981 1306 500 696 605 250 411 1054 735  
## [91] 233 435 490 310 460 383 375 1270 545 445 1885 380 300 380 377  
## [106] 425 276 210 800 420 350 360 538 1100 1205 314 237 610 360 540  
## [121] 1038 424 310 300 444 301 268 620 215 652 900 525 246 360 529  
## [136] 500 720 270 430 671 1770

1. 计算这一数据集的元素个数、平均数、中位数、标准差、方差、最大值、最小值等描述性统计特征。要求使用两种方法计算，一种为R语言自带的内置基本函数，另一种要求使用外部包提供的描述性统计函数。打印所有的计算结果；

length(rivers)

## [1] 141

mean(rivers)

## [1] 591.1844

median(rivers)

## [1] 425

sd(rivers)

## [1] 493.8708

var(rivers)

## [1] 243908.4

min(rivers)

## [1] 135

max(rivers)

## [1] 3710

library(psych)  
psych::describe(rivers)

## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 141 591.18 493.87 425 490.95 214.98 135 3710 3575 3.15 13.07  
## se  
## X1 41.59

1. 将2中的基本函数的计算结果构建成一个名为rivers.Des.1的向量中，打印该向量；

rivers.Des.1 <- c(mean(rivers),median(rivers),sd(rivers),var(rivers),max(rivers),min(rivers))  
print(rivers.Des.1)

## [1] 591.1844 425.0000 493.8708 243908.4086 3710.0000 135.0000

1. 将2中的由外部包提供的描述性统计函数的计算结果构建成一个名为rivers.Des.2的数据框（dataframe）中，要求该数据框有两个变量：feature\_name（统计指标名称）和value（统计值）。打印该数据框。

des <- describe((rivers))  
feature\_name <- names(des)  
value <- c(des$vars,des$n,des$mean,des$sd,des$median,des$trimmed,des$mad,des$min,des$max,des$range,des$skew,des$kurtosis,des$se)  
rivers.Des.2 <- data.frame(feature\_name,value)  
print(rivers.Des.2)

## feature\_name value  
## 1 vars 1.000000  
## 2 n 141.000000  
## 3 mean 591.184397  
## 4 sd 493.870842  
## 5 median 425.000000  
## 6 trimmed 490.946903  
## 7 mad 214.977000  
## 8 min 135.000000  
## 9 max 3710.000000  
## 10 range 3575.000000  
## 11 skew 3.150068  
## 12 kurtosis 13.067766  
## 13 se 41.591428

## 问题二 women

1. 计算该数据集的行数与列数；

print(nrow(women))

## [1] 15

print(ncol(women))

## [1] 2

1. 打印该数据集的前6个观测，和最后6个观测；

print(head(women))

## height weight  
## 1 58 115  
## 2 59 117  
## 3 60 120  
## 4 61 123  
## 5 62 126  
## 6 63 129

print(tail(women))

## height weight  
## 10 67 142  
## 11 68 146  
## 12 69 150  
## 13 70 154  
## 14 71 159  
## 15 72 164

1. 计算height列的均值和方差

mean(women$height)

## [1] 65

var(women$height)

## [1] 20

1. 请选择height列的值大于60的行，形成一个新的数据集，名称为women.Height60；

women.Height60 <- subset(women,height >60)  
print(women.Height60)

## height weight  
## 4 61 123  
## 5 62 126  
## 6 63 129  
## 7 64 132  
## 8 65 135  
## 9 66 139  
## 10 67 142  
## 11 68 146  
## 12 69 150  
## 13 70 154  
## 14 71 159  
## 15 72 164

1. 将women数据集转化为一个列表类型的名叫women.list数据，要求列表中包含两个元素，分别是height和weight，打印该列表；

women.list <- list(women$height,women$weight)  
print(women.list)

## [[1]]  
## [1] 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72  
##   
## [[2]]  
## [1] 115 117 120 123 126 129 132 135 139 142 146 150 154 159 164

1. 将women数据集转化为一个矩阵类型的名叫women.matrix数据，打印该矩阵和该矩阵的转置矩阵；

women.matrix <- as.matrix(women)  
print(women.matrix)

## height weight  
## [1,] 58 115  
## [2,] 59 117  
## [3,] 60 120  
## [4,] 61 123  
## [5,] 62 126  
## [6,] 63 129  
## [7,] 64 132  
## [8,] 65 135  
## [9,] 66 139  
## [10,] 67 142  
## [11,] 68 146  
## [12,] 69 150  
## [13,] 70 154  
## [14,] 71 159  
## [15,] 72 164

print(t(women.matrix))

## [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]  
## height 58 59 60 61 62 63 64 65 66 67 68 69 70  
## weight 115 117 120 123 126 129 132 135 139 142 146 150 154  
## [,14] [,15]  
## height 71 72  
## weight 159 164

1. 请使用R自带的cor这一函数，计算women这一数据集height和weight列的相关系数；

cor(women$height,women$weight)

## [1] 0.9954948

1. cor这一函数中的一个参数为method，它的取值包括，person，kendall，spearman。请简略写下他们之间的区别（不超过100个字）。

person是指皮尔逊相关系数用于度量两个变量之间的相关程度，其值介于-1与1之间

Kendall用于有序分类变量属于等级相关系数。排序一致，则为1，排序完全相反则为-1

Spearman 为秩相关系数，无参数的等级相关系数，亦即其值与两个相关变量的具体值无关，而仅仅与其值之间的大小关系有关。