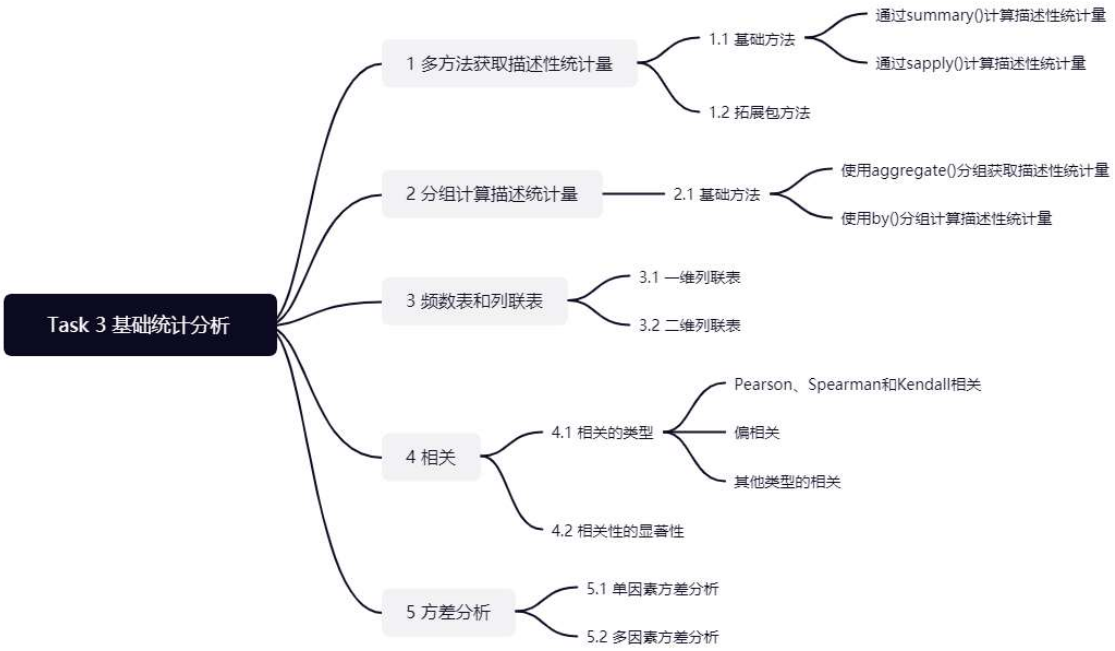
DataWhale 组队学习 R语言数据分析 Task03 基础统计分析

杨佳达

2021-07-15



基础统计分析

准备工作

如果没有相关的包，则使用 `install.packages('package_name')` 进行安装以下包。

```
library(pastecs)
library(psych)
library(ggm)
```

读取数据，使用H1N1流感数据集和波士顿房价数据集。

```
flu = read.table("./datasets/h1n1_flu.csv", header = TRUE, sep = '
housing = read.csv("./datasets/BostonHousing.csv", header = TRUE)
```

1 多种方法获取描述性统计量

1.1 基础方法

通过summary计算数值型变量的最大值、最小值、分位数以及均值，类别变量计算频数统计。

```
summary(flu[c("household_children", "sex")])
```

```
##  household_children     sex
##  Min.   :0.0000     Length:26707
##  1st Qu.:0.0000     Class :character
##  Median :0.0000     Mode  :character
##  Mean   :0.5346
##  3rd Qu.:1.0000
##  Max.   :3.0000
##  NA's   :249
```

```
summary(flu[c("h1n1_concern", "h1n1_knowledge")])
```

```
##   h1n1_concern   h1n1_knowledge
##  Min.   :0.000   Min.   :0.000
##  1st Qu.:1.000   1st Qu.:1.000
##  Median :2.000   Median :1.000
##  Mean   :1.618   Mean   :1.263
##  3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :3.000   Max.   :2.000
##  NA's   :92      NA's   :116
```

通过 sapply() 计算描述性统计量，先定义统计函数，在进行聚合计算。

```
mystats <- function(x, na.omit = FALSE) {
    if (na.omit)
        x <- x[!is.na(x)]
    m <- mean(x)
```

```
    n <- length(x)
    s <- sd(x)
    skew <- sum((x - m)^3/s^3)/n
    kurt <- sum((x - m)^4/s^4)/n - 3
    return(c(n = n, mean = m, stdev = s, skew = skew, kurtosis = k
}

sapply(flu[c("h1n1_concern", "h1n1_knowledge")], mystats)
```

```
##          h1n1_concern h1n1_knowledge
## n               26707          26707
## mean               NA             NA
## stdev              NA             NA
## skew               NA             NA
## kurtosis           NA             NA
```

1.2 拓展包方法

通过pastecs包中的 stat.desc()函数计算描述性统计量，可以得到中位数、平均数、平均数的标准误、平均数置信度为95%的置信区间、方差、标准差以及变异系数。

```
stat.desc(flu[c("household_children", "sex")])
```

```
##              household_children sex
## nbr.val            2.645800e+04  NA
## nbr.null           1.867200e+04  NA
## nbr.na             2.490000e+02  NA
## min                0.000000e+00  NA
## max                3.000000e+00  NA
## range              3.000000e+00  NA
## sum                1.414400e+04  NA
## median             0.000000e+00  NA
## mean               5.345831e-01  NA
## SE.mean            5.706247e-03  NA
## CI.mean.0.95       1.118455e-02  NA
## var                8.615057e-01  NA
## std.dev            9.281733e-01  NA
## coef.var           1.736256e+00  NA
```

通过psych包中的describe()计算描述性统计量。

```
describe(flu[c("household_children", "sex")])
```

```
##                    vars     n mean   sd median trimmed mad min
## household_children    1 26458 0.53 0.93      0    0.34   0   0
## sex*                  2 26707 1.41 0.49      1    1.38   0   1
##                    kurtosis   se
## household_children     1.04 0.01
## sex*                  -1.85 0.00
```

2 分组计算描述性统计

2.1 基础方法

使用aggregate()分组获取描述性统计

1. 分组计算不同性别收入贫困计数。
2. 是否属于查尔斯河的房价中位数平均值。

```
aggregate(flu[c("income_poverty")], by = list(sex = flu$sex), leng
```

```
##      sex income_poverty
## 1 Female          15858
## 2   Male          10849
```

```
aggregate(housing$medv, by = list(medv = housing$chas), FUN = mean
```

```
##   medv        x
## 1    0 22.09384
## 2    1 28.44000
```

使用 by() 分组计算描述性统计量

```
by(flu[c("income_poverty", "sex")], flu$sex, length)
```

```
## flu$sex: Female
```

```
## [1] 2
## ------------------------------------------------------------
## flu$sex: Male
## [1] 2
```

3 频数表和列联表

```
table(flu$sex)
```

```
##
## Female   Male
##  15858  10849
```

4 相关

4.1 相关的类型

Pearson、Spearman和Kendall相关

R可以计算多种相关系数，包括Pearson相关系数、Spearman相关系数、Kendall相关系数、偏相关系数、多分格（polychoric）相关系数和多系列（polyserial）相关系数。 1. 计算房价数据的相关系数，默认是Pearson相关系数。

```
cor(housing)
```

```
##                   X        crim          zn       indus
## X         1.000000000  0.40740717 -0.10339336  0.39943885 -0.00
## crim      0.407407172  1.00000000 -0.20046922  0.40658341 -0.05
## zn       -0.103393357 -0.20046922  1.00000000 -0.53382819 -0.04
## indus     0.399438850  0.40658341 -0.53382819  1.00000000  0.06
## chas     -0.003759115 -0.05589158 -0.04269672  0.06293803  1.00
## nox       0.398736174  0.42097171 -0.51660371  0.76365145  0.09
## rm       -0.079971150 -0.21924670  0.31199059 -0.39167585  0.09
## age       0.203783510  0.35273425 -0.56953734  0.64477851  0.08
## dis      -0.302210959 -0.37967009  0.66440822 -0.70802699 -0.09
## rad       0.686001976  0.62550515 -0.31194783  0.59512927 -0.00
## tax       0.666625924  0.58276431 -0.31456332  0.72076018 -0.03
## ptratio   0.291074227  0.28994558 -0.39167855  0.38324756 -0.12
## b        -0.295041232 -0.38506394  0.17552032 -0.35697654  0.04
```

```
## lstat    0.258464770  0.45562148 -0.41299457  0.60379972 -0.053
## medv    -0.226603643 -0.38830461  0.36044534 -0.48372516  0.175
##                 nox          rm         age         dis
## X        0.39873617 -0.07997115  0.20378351 -0.30221096  0.6866
## crim     0.42097171 -0.21924670  0.35273425 -0.37967009  0.6255
## zn      -0.51660371  0.31199059 -0.56953734  0.66440822 -0.3119
## indus    0.76365145 -0.39167585  0.64477851 -0.70802699  0.5951
## chas     0.09120281  0.09125123  0.08651777 -0.09917578 -0.0073
## nox      1.00000000 -0.30218819  0.73147010 -0.76923011  0.6114
## rm      -0.30218819  1.00000000 -0.24026493  0.20524621 -0.2098
## age      0.73147010 -0.24026493  1.00000000 -0.74788054  0.4566
## dis     -0.76923011  0.20524621 -0.74788054  1.00000000 -0.4945
## rad      0.61144056 -0.20984667  0.45602245 -0.49458793  1.0000
## tax      0.66802320 -0.29204783  0.50645559 -0.53443158  0.9102
## ptratio  0.18893268 -0.35550149  0.26151501 -0.23247054  0.4647
## b       -0.38005064  0.12806864 -0.27353398  0.29151167 -0.4444
## lstat    0.59087892 -0.61380827  0.60233853 -0.49699583  0.4886
## medv    -0.42732077  0.69535995 -0.37695457  0.24992873 -0.3816
##                 tax    ptratio           b      lstat       med
## X        0.66662592  0.2910742 -0.29504123  0.2584648 -0.22660
## crim     0.58276431  0.2899456 -0.38506394  0.4556215 -0.38830
## zn      -0.31456332 -0.3916785  0.17552032 -0.4129946  0.36044
## indus    0.72076018  0.3832476 -0.35697654  0.6037997 -0.48372
## chas    -0.03558652 -0.1215152  0.04878848 -0.0539293  0.17526
## nox      0.66802320  0.1889327 -0.38005064  0.5908789 -0.42732
## rm      -0.29204783 -0.3555015  0.12806864 -0.6138083  0.69535
## age      0.50645559  0.2615150 -0.27353398  0.6023385 -0.37695
## dis     -0.53443158 -0.2324705  0.29151167 -0.4969958  0.24992
## rad      0.91022819  0.4647412 -0.44441282  0.4886763 -0.38162
## tax      1.00000000  0.4608530 -0.44180801  0.5439934 -0.46853
## ptratio  0.46085304  1.0000000 -0.17738330  0.3740443 -0.50778
## b       -0.44180801 -0.1773833  1.00000000 -0.3660869  0.33346
## lstat    0.54399341  0.3740443 -0.36608690  1.0000000 -0.73766
## medv    -0.46853593 -0.5077867  0.33346082 -0.7376627  1.00000
```

2. 指定计算Spearman相关系数

```
cor(housing, method = "spearman")
```

```
##                   X        crim         zn       indus
## X        1.000000000  0.46103705 -0.1605047  0.32462127 -0.003
## crim     0.461037054  1.00000000 -0.5716602  0.73552374  0.041
## zn      -0.160504702 -0.57166021  1.0000000 -0.64281060 -0.041
```

```
## indus    0.324621271  0.73552374 -0.6428106  1.00000000  0.0898
## chas    -0.003759115  0.04153689 -0.0419370  0.08984138  1.0000
## nox      0.432491886  0.82146466 -0.6348284  0.79118913  0.0684
## rm      -0.035641354 -0.30911647  0.3610737 -0.41530129  0.0588
## age      0.208323439  0.70413998 -0.5444226  0.67948671  0.0677
## dis     -0.373498683 -0.74498614  0.6146265 -0.75707970 -0.0802
## rad      0.588480705  0.72780697 -0.2787672  0.45550745  0.0245
## tax      0.536928176  0.72904490 -0.3713945  0.66436139 -0.0444
## ptratio  0.297897432  0.46528319 -0.4484754  0.43371046 -0.1366
## b       -0.154474321 -0.36055532  0.1631351 -0.28583984 -0.0398
## lstat    0.257542491  0.63476026 -0.4900739  0.63874741 -0.0505
## medv    -0.273633481 -0.55889095  0.4381790 -0.57825539  0.1406
##                 nox          rm         age         dis
## X        0.43249189 -0.03564135  0.20832344 -0.37349868  0.5884
## crim     0.82146466 -0.30911647  0.70413998 -0.74498614  0.7278
## zn      -0.63482840  0.36107373 -0.54442256  0.61462654 -0.2787
## indus    0.79118913 -0.41530129  0.67948671 -0.75707970  0.4555
## chas     0.06842628  0.05881292  0.06779178 -0.08024808  0.0245
## nox      1.00000000 -0.31034391  0.79515291 -0.88001486  0.5864
## rm      -0.31034391  1.00000000 -0.27808202  0.26316822 -0.1074
## age      0.79515291 -0.27808202  1.00000000 -0.80160979  0.4179
## dis     -0.88001486  0.26316822 -0.80160979  1.00000000 -0.4958
## rad      0.58642870 -0.10749220  0.41798261 -0.49580647  1.0000
## tax      0.64952656 -0.27189846  0.52636644 -0.57433641  0.7048
## ptratio  0.39130908 -0.31292257  0.35538428 -0.32204056  0.3183
## b       -0.29666158  0.05366004 -0.22802200  0.24959532 -0.2825
## lstat    0.63682829 -0.64083156  0.65707079 -0.56426219  0.3943
## medv    -0.56260883  0.63357643 -0.54756169  0.44585685 -0.3467
##             ptratio           b       lstat       medv
## X        0.29789743 -0.15447432  0.25754249 -0.2736335
## crim     0.46528319 -0.36055532  0.63476026 -0.5588909
## zn      -0.44847543  0.16313510 -0.49007389  0.4381790
## indus    0.43371046 -0.28583984  0.63874741 -0.5782554
## chas    -0.13606462 -0.03981050 -0.05057483  0.1406122
## nox      0.39130908 -0.29666158  0.63682829 -0.5626088
## rm      -0.31292257  0.05366004 -0.64083156  0.6335764
## age      0.35538428 -0.22802200  0.65707079 -0.5475617
## dis     -0.32204056  0.24959532 -0.56426219  0.4458569
## rad      0.31832966 -0.28253261  0.39432245 -0.3467763
## tax      0.45334546 -0.32984308  0.53442319 -0.5624106
## ptratio  1.00000000 -0.07202734  0.46725885 -0.5559047
## b       -0.07202734  1.00000000 -0.21056185  0.1856641
## lstat    0.46725885 -0.21056185  1.00000000 -0.8529141
## medv    -0.55590468  0.18566412 -0.85291414  1.0000000
```

3. 城镇人均犯罪率与房价的相关系数

```
x <- housing
y <- housing[c("medv")]
cor(x, y)
```

```
##                medv
## X       -0.2266036
## crim    -0.3883046
## zn       0.3604453
## indus   -0.4837252
## chas     0.1752602
## nox     -0.4273208
## rm       0.6953599
## age     -0.3769546
## dis      0.2499287
## rad     -0.3816262
## tax     -0.4685359
## ptratio -0.5077867
## b        0.3334608
## lstat   -0.7376627
## medv     1.0000000
```

偏相关

偏相关是指在控制一个或多个定量变量时，另外两个定量变量之间的相互关系。使用ggm 包中的 pcor() 函数计算偏相关系数。

4.2 相关性的显著性检验

```
cor.test(housing[, c("crim")], housing[, c("medv")])
```

```
##
##  Pearson's product-moment correlation
##
## data:  housing[, c("crim")] and housing[, c("medv")]
## t = -9.4597, df = 504, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4599064 -0.3116859
## sample estimates:
```

```
##         cor
## -0.3883046
```

5 方差分析

方差分析（ANOVA）又称"变异数分析"或"F检验"，用于两个及两个以上样本均数差别的显著性检验。

5.1 单因素方差分析

从输出结果的F检验值来看，p<0.05比较显著，说明是否在查尔斯河对房价有影响。

```
fit <- aov(housing$medv ~ housing$chas)
summary(fit)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## housing$chas    1   1312  1312.1   15.97 7.39e-05 ***
## Residuals     504  41404    82.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.2 多因素方差分析

构建多因素方差分析，查看因子对房价的影响是否显著。

```
fit <- aov(housing$medv ~ housing$crim * housing$b)
summary(fit)
```

```
##                          Df Sum Sq Mean Sq F value   Pr(>F)
## housing$crim              1   6441    6441   96.05  < 2e-16 ***
## housing$b                 1   1697    1697   25.30 6.83e-07 ***
## housing$crim:housing$b    1    917     917   13.68 0.000241 ***
## Residuals               502  33662      67
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```