Biostatistics HomeWork 1

KIM SANG HYUN(202211545)

2025-03-14

Contents

0. Package and Data

```
library(ISLR)
data("NCI60")
unique(NCI60$labs)
```

```
##  [1] "CNS"          "RENAL"        "BREAST"       "NSCLC"        "UNKNOWN"
##  [6] "OVARIAN"      "MELANOMA"     "PROSTATE"     "LEUKEMIA"     "K562B-repro"
## [11] "K562A-repro"  "COLON"        "MCF7A-repro"  "MCF7D-repro"
```

```
dim(NCI60$data) # row = cancer cell lines(samples), col = gene expression
```

```
## [1]   64 6830
```

```
str(NCI60)
```

```
## List of 2
##  $ data: num [1:64, 1:6830] 0.3 0.68 0.94 0.28 0.485 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:64] "V1" "V2" "V3" "V4" ...
##   .. ..$ : chr [1:6830] "1" "2" "3" "4" ...
##  $ labs: chr [1:64] "CNS" "CNS" "CNS" "RENAL" ...
```

```
NCI60$data[1:5, 1:10]
```

```
##           1        2        3        4         5             6        7
## V1 0.300000 1.180000 0.550000 1.140000 -0.265000 -7.000000e-02 0.350000
## V2 0.679961 1.289961 0.169961 0.379961  0.464961  5.799610e-01 0.699961
## V3 0.940000 -0.040000 -0.170000 -0.040000 -0.605000  0.000000e+00 0.090000
## V4 0.280000 -0.310000  0.680000 -0.810000  0.625000 -1.387779e-17 0.170000
## V5 0.485000 -0.465000  0.395000  0.905000  0.200000 -5.000000e-03 0.085000
##           8           9          10
## V1 -0.315000 -0.45000000 -0.65498050
## V2  0.724961 -0.04003899 -0.28501950
## V3  0.645000  0.43000000  0.47501950
## V4  0.245000  0.02000000  0.09501949
## V5  0.110000  0.23500000  1.49001949
```

```
sum(is.na(NCI60$data))
```

```
## [1] 0
```

1. For each cancer cell line, compute average gene expression values. Identify two cell lines that have the largest and the smallest mean values. Also, include the maximun and minimum mean values.

```
sol_1 = function(x = NCI60$data, cell_name = NCI60$labs){

  avg_gene_expr_cancer_cell = apply(x, 1, mean)

  max_idx = which.max(avg_gene_expr_cancer_cell)
  min_idx = which.min(avg_gene_expr_cancer_cell)

  max_cell_name = cell_name[max_idx]
  min_cell_name = cell_name[min_idx]

  max_value = avg_gene_expr_cancer_cell[max_idx]
  min_value = avg_gene_expr_cancer_cell[min_idx]

  return(
     list(
      avg_gene_expr_cancer_cell = avg_gene_expr_cancer_cell,
      max_cell_name = max_cell_name,
      max_value = max_value,
      min_cell_name = min_cell_name,
      min_value = min_value
    )
  )

}

sol_1()
```

```
## $avg_gene_expr_cancer_cell
##          V1          V2          V3          V4          V5          V6
##  0.065301161  0.050764025  0.072186913  0.093828598  0.148587415  0.051984464
##          V7          V8          V9         V10         V11         V12
##  0.045946454  0.034898165  0.030765650  0.071969531  0.082987492  0.097510887
##         V13         V14         V15         V16         V17         V18
##  0.043871076  0.080305578  0.086590327  0.068228701  0.036918312  0.005414269
##          V19          V20          V21          V22          V23          V24
##  0.044108654 -0.010810123  0.022486014  0.021204167  0.022557379  0.028589256
##          V25          V26          V27          V28          V29          V30
##  0.055356775  0.076016757  0.037696115  0.030543813  0.052745532  0.029498474
##          V31          V32          V33          V34          V35          V36
##  0.065867052  0.045337045  0.035683095  0.008683965 -0.025011443 -0.067179793
##          V37          V38          V39          V40          V41          V42
## -0.073247688 -0.059364277 -0.142086454 -0.112343414 -0.079673926  0.016840221
##          V43          V44          V45          V46          V47          V48
## -0.022744231 -0.045290706  0.013481493  0.008972460  0.009429658  0.005350211
##          V49          V50          V51          V52          V53          V54
## -0.070572553 -0.045211715 -0.015887593  0.014101316  0.050248954 -0.033281552
##          V55          V56          V57          V58          V59          V60
## -0.043874535  0.016277763 -0.004496054 -0.019257425  0.040631693  0.065602038
##          V61          V62          V63          V64
##  0.040684766  0.072229801  0.021697977  0.039845104
##
## $max_cell_name
## [1] "BREAST"
##
## $max_value
##        V5
## 0.1485874
##
## $min_cell_name
## [1] "LEUKEMIA"
##
## $min_value
##        V39
## -0.1420865
```

2. For each gene, compute average gene expression values of 64 cancer cell lines, including “UNKNOWN” label. Identify top 5 gene that have the largest mean expression values and top 5 genes that have smallest mean expression values. Also, include their mean values with gene ID number(1~6830).

```
sol_2 = function(x = NCI60$data){

  avg_gene_expr_gene = apply(x ,2, mean)

  top_5_max_idx = order(avg_gene_expr_gene, decreasing = TRUE)[1 : 5]
  top_5_min_idx = order(avg_gene_expr_gene)[1 : 5]

  top_5_max_gene = avg_gene_expr_gene[top_5_max_idx]
  top_5_min_gene = avg_gene_expr_gene[top_5_min_idx]

  return(
    list(
      top_5_max_gene = top_5_max_gene,
      top_5_min_gene = top_5_min_gene
    )
  )
}

sol_2()
```

```
## $top_5_max_gene
##      6393       256       257      4700      6391 
## 1.1676457 1.1137491 1.0627335 0.9985928 0.9920303 
## 
## $top_5_min_gene
##       5869       5868       5984       3438        281 
## -0.8621881 -0.7442193 -0.7360845 -0.7223447 -0.7109384
```

3. Suppose that group "A" contains "BREAST" and "NSCLC", group "B" has "MELANOMA", "OVARIAN" and "PROSTATE", and group "C" has "LEUKEMIA", "RENAL" and "UNKNOWN". The other 6 cell lines belong to group "D". For each cancer group, compute the mean expression values and the standard deviation of gene experssion values.

```
cell_name = NCI60$labs

# grouping cells
A = c("BREAST","NSCLC")
B = c("MELANOMA", "OVARIAN", "PROSTATE")
C = c("LEUKEMIA", "RENAL", "UNKNOWN")

cell_group = ifelse(cell_name %in% A, "A",
               ifelse(cell_name %in% B, "B",
               ifelse(cell_name %in% C, "C", "D")))
```

```
groups = c("A", "B", "C", "D")

sol_3 = function(x = NCI60$data, group = groups){

  group_mean = sapply(group,
                      function(g) {
                        mean(x[cell_group == g, ])
                      })

  group_sd = sapply(group,
                    function(g) {
                      sd(x[cell_group == g, ])
                    })

  list(
    group_mean = as.data.frame(group_mean),
    group_sd   = as.data.frame(group_sd)
  )
}

sol_3()
```

```
## $group_mean
##    group_mean
## A 0.025053741
## B 0.040698588
## C 0.008992817
## D 0.005848171
##
## $group_sd
##    group_sd
## A 0.8006589
## B 0.7368874
## C 0.8443378
## D 0.7924379
```

4. For each cancer group defined in Q3, compute the sample SD of gene expression values of individual genes. Find genes whose SD is less than 0.2 or greater than 2 for each cancer group, i.e., $SD < 0.2$ or $SD > 2$. How many genes are overlapped by 4 different cancer groups? How many genes are overlapped by exactly 3 different cancer groups? or exactly 2 different cancer groups? Also, how many genes are uniquely identified by only one cancer group? Summarize your answer, using the following table.

```r
group_sd = lapply(groups,
                  function(grp) {
                    apply(NCI60$data[cell_group == grp, ], 2, sd)
                  })

names(group_sd) = c("A", "B", "C", "D")

filtered_genes = lapply(group_sd,
                        function(sd_value){
                          names(sd_value[sd_value < 0.2 | sd_value > 2])
                       })

unique_genes = unique(c(filtered_genes[["A"]],
                            filtered_genes[["B"]],
                            filtered_genes[["C"]],
                            filtered_genes[["D"]]))

num_overlapped = factor(apply(sapply(filtered_genes, function(g_list) unique_genes %in% g_list), 1, sum

sol_df = data.frame(
  x4 = length(num_overlapped[num_overlapped == "4"]),
  x3 = length(num_overlapped[num_overlapped == "3"]),
  x2 = length(num_overlapped[num_overlapped == "2"]),
  x1  = length(num_overlapped[num_overlapped == "1"])
)

rownames(sol_df) = "The number of genes"
colnames(sol_df) = c("4 groups", "3 groups", "2 groups", "1 groups")

sol_df
```

```
##                       4 groups 3 groups 2 groups 1 groups
## The number of genes       14       17       71      220
```

```
sol_4 = function(){
  group_sd = lapply(groups,
                  function(grp) {
                    apply(NCI60$data[cell_group == grp, ], 2, sd)
                  })

  names(group_sd) = c("A", "B", "C", "D")

  filtered_genes = lapply(group_sd,
                        function(sd_value){
                          names(sd_value[sd_value < 0.2 | sd_value > 2])
                        })

  unique_genes = unique(c(filtered_genes[["A"]],
                            filtered_genes[["B"]],
                            filtered_genes[["C"]],
                            filtered_genes[["D"]]))

  num_overlapped = factor(apply(sapply(filtered_genes, function(g_list) unique_genes %in% g_list), 1, su

  sol_df = data.frame(
    x4 = length(num_overlapped[num_overlapped == "4"]),
    x3 = length(num_overlapped[num_overlapped == "3"]),
    x2 = length(num_overlapped[num_overlapped == "2"]),
    x1  = length(num_overlapped[num_overlapped == "1"])
  )

  rownames(sol_df) = "The number of genes"
  colnames(sol_df) = c("4 groups", "3 groups", "2 groups", "1 groups")

  return(list(filtered_genes = filtered_genes, sol_df = sol_df))
}

sol_4()
```

```
## $filtered_genes
## $filtered_genes$A
##   [1] "16"   "111"  "112"  "113"  "134"  "196"  "243"  "245"  "248"  "251"
##  [11] "252"  "256"  "257"  "266"  "267"  "273"  "281"  "286"  "472"  "975"
##  [21] "1106" "1215" "1258" "1865" "2068" "2504" "2838" "2875" "2914" "2927"
##  [31] "3320" "3383" "3438" "3518" "3525" "3543" "3936" "3956" "3957" "4050"
##  [41] "4154" "4280" "4288" "4344" "4353" "4354" "4699" "4700" "4701" "5036"
##  [51] "5142" "5221" "5275" "5276" "5353" "5476" "5477" "5555" "5556" "5557"
##  [61] "5586" "5587" "5661" "5692" "5705" "5706" "5707" "5723" "5732" "5760"
##  [71] "5803" "5804" "5805" "5828" "5829" "5838" "5843" "5845" "5913" "5940"
##  [81] "5942" "5943" "5948" "5980" "6128" "6148" "6149" "6150" "6151" "6152"
##  [91] "6153" "6156" "6157" "6263" "6264" "6268" "6277" "6278" "6279" "6321"
## [101] "6328" "6356" "6391" "6392" "6393" "6415" "6416" "6419" "6429" "6430"
## [111] "6453" "6564" "6612" "6614" "6615" "6616" "6622" "6718"
##
## $filtered_genes$B
##  [1] "124"  "125"  "128"  "130"  "133"  "134"  "196"  "241"  "242"  "243" 
## [11] "252"  "256"  "257"  "286"  "287"  "408"  "416"  "561"  "580"  "581" 
## [21] "592"  "754"  "755"  "770"  "1067" "1110" "1508" "1664" "1888" "1896"
## [31] "1897" "2100" "2216" "2239" "2551" "2678" "2680" "2891" "3234" "3706"
## [41] "3713" "3957" "4093" "4094" "4280" "4288" "4289" "4304" "4306" "4308"
## [51] "4320" "4327" "4344" "4353" "4354" "4375" "4383" "4387" "4388" "4425"
## [61] "4426" "4699" "4700" "4701" "4716" "4971" "5094" "5275" "5276" "5353"
## [71] "5555" "5556" "5557" "5586" "5804" "5948" "6149" "6150" "6322" "6356"
## [81] "6391" "6392" "6393" "6434" "6554" "6635" "6710"
## 
## $filtered_genes$C
##   [1] "16"   "78"   "133"  "187"  "196"  "243"  "252"  "256"  "281"  "286" 
##  [11] "301"  "415"  "515"  "707"  "754"  "755"  "756"  "806"  "1199" "1229"
##  [21] "1387" "1388" "1389" "1390" "1391" "2068" "2070" "2074" "2080" "2081"
##  [31] "2082" "2083" "2102" "3234" "3248" "3282" "3372" "3373" "3490" "3491"
##  [41] "3518" "3525" "3894" "4085" "4131" "4154" "4245" "4344" "4354" "4699"
##  [51] "4700" "4701" "4716" "5127" "5221" "5270" "5301" "5336" "5392" "5481"
##  [61] "5489" "5496" "5506" "5510" "5586" "5587" "5588" "5705" "5712" "5721"
##  [71] "5729" "5732" "5758" "5760" "5774" "5796" "5803" "5804" "5805" "5867"
##  [81] "5868" "5869" "5870" "5872" "5878" "5884" "5899" "5902" "5910" "5917"
##  [91] "5921" "5927" "5928" "5937" "5940" "5941" "5942" "5943" "5946" "5948"
## [101] "5950" "5962" "5972" "5973" "5976" "5979" "5980" "5981" "5993" "6009"
## [111] "6010" "6017" "6018" "6035" "6039" "6046" "6084" "6085" "6086" "6087"
## [121] "6124" "6148" "6149" "6150" "6151" "6152" "6153" "6154" "6156" "6157"
## [131] "6169" "6243" "6272" "6274" "6277" "6278" "6279" "6288" "6289" "6382"
## [141] "6391" "6392" "6393" "6412" "6413" "6414" "6415" "6416" "6429" "6430"
## [151] "6592" "6596" "6635" "6644" "6646" "6688" "6689" "6710" "6717" "6817"
## 
## $filtered_genes$D
##   [1] "16"   "112"  "113"  "161"  "188"  "224"  "227"  "228"  "229"  "243" 
##  [11] "248"  "252"  "256"  "257"  "267"  "286"  "301"  "412"  "582"  "707" 
##  [21] "716"  "754"  "755"  "770"  "975"  "1187" "1380" "1382" "1388" "1389"
##  [31] "1390" "1391" "1393" "1396" "1613" "1716" "2080" "2081" "2096" "2104"
##  [41] "2302" "3212" "3424" "3936" "3957" "4010" "4057" "4060" "4093" "4094"
##  [51] "4119" "4154" "4231" "4344" "4472" "4612" "4644" "4699" "4700" "4701"
##  [61] "4703" "4704" "4706" "4845" "4994" "5031" "5142" "5472" "5481" "5646"
##  [71] "5680" "5691" "5692" "5696" "5705" "5706" "5707" "5732" "5804" "5805"
##  [81] "5838" "5867" "5868" "5869" "5870" "5916" "5917" "5937" "5948" "5980"
##  [91] "6068" "6149" "6274" "6391" "6392" "6393" "6415" "6416" "6612" "6635"
## [101] "6646" "6687" "6688" "6689"
## 
## 
## $sol_df
##                     4 groups 3 groups 2 groups 1 groups
## The number of genes       14       17       71      220
```