20250412

上海萤锐科技

目录

# 1. 简介

临床研究中如果课题中使用的样本量比较大多，经常需要在文章中使用一个表格展示样本的临床信息。手动统计这个表格是非常耗时且容易出错的事。今天小编就为大家介绍一款工具，其可以用来统计和绘制临床信息的表格，它就是今天的主角compareGroups(v4.9.1)[1]R包。

今天我们就使用TCGA-LIHC数据中的临床信息为大家展示如何使用compareGroups 统计并绘制临床信息表。

# 2. 加载需要的包

library(magrittr) # 使用其提供的管道操作符  
library(SummarizedExperiment)  
library(compareGroups)

# 3. 读取数据

读取已经下载好的TCGA-LIHC数据，该数据的保存格式是RangedSummarizedExperiment，我们可以使用colData函数从中提取样本的临床信息。

lihc <- readRDS("../../../disk14/data/dataset/TCGA/SummarizedExperiment/TCGA-LIHC.rds")  
sample\_info <- lihc %>% colData %>% as.data.frame()  
  
# 查看未经整理的临床数据  
sample\_info[1:5, 1:5]

barcode patient  
TCGA-FV-A3I0-01A-11R-A22L-07 TCGA-FV-A3I0-01A-11R-A22L-07 TCGA-FV-A3I0  
TCGA-DD-A3A6-11A-11R-A22L-07 TCGA-DD-A3A6-11A-11R-A22L-07 TCGA-DD-A3A6  
TCGA-DD-A3A6-01A-11R-A22L-07 TCGA-DD-A3A6-01A-11R-A22L-07 TCGA-DD-A3A6  
TCGA-BD-A3ER-01A-11R-A213-07 TCGA-BD-A3ER-01A-11R-A213-07 TCGA-BD-A3ER  
TCGA-CC-5261-01A-01R-A131-07 TCGA-CC-5261-01A-01R-A131-07 TCGA-CC-5261  
 sample shortLetterCode  
TCGA-FV-A3I0-01A-11R-A22L-07 TCGA-FV-A3I0-01A TP  
TCGA-DD-A3A6-11A-11R-A22L-07 TCGA-DD-A3A6-11A NT  
TCGA-DD-A3A6-01A-11R-A22L-07 TCGA-DD-A3A6-01A TP  
TCGA-BD-A3ER-01A-11R-A213-07 TCGA-BD-A3ER-01A TP  
TCGA-CC-5261-01A-01R-A131-07 TCGA-CC-5261-01A TP  
 definition  
TCGA-FV-A3I0-01A-11R-A22L-07 Primary solid Tumor  
TCGA-DD-A3A6-11A-11R-A22L-07 Solid Tissue Normal  
TCGA-DD-A3A6-01A-11R-A22L-07 Primary solid Tumor  
TCGA-BD-A3ER-01A-11R-A213-07 Primary solid Tumor  
TCGA-CC-5261-01A-01R-A131-07 Primary solid Tumor

## 3.1 整理临床数据

sample\_info <- sample\_info %>%   
 dplyr::filter(sample\_type\_id == "01") %>%   
 dplyr::select(sample\_type,   
 days\_to\_collection,   
 prior\_malignancy,   
 prior\_treatment,   
 ajcc\_pathologic\_stage,   
 ajcc\_pathologic\_t,  
 ajcc\_pathologic\_n,  
 ajcc\_pathologic\_m,  
 days\_to\_last\_follow\_up,   
 age = age\_at\_index,   
 race,   
 gender,  
 vital\_status)  
  
sample\_info <- sample\_info %>%   
 dplyr::mutate(  
 ajcc\_pathologic\_stage = sapply(ajcc\_pathologic\_stage, function(x){  
 switch(as.character(x),   
 "Stage I" = "I&II",   
 "Stage II" = "I&II",   
 "Stage III" = "III&IV",   
 "Stage IIIA" = "III&IV",   
 "Stage IIIB" = "III&IV",   
 "Stage IIIC" = "III&IV",   
 "Stage IV" = "III&IV",   
 "Stage IVA" = "III&IV",   
 "Stage IVB" = "III&IV",   
 "NA" = NA)  
 }),   
 ajcc\_pathologic\_t = sapply(ajcc\_pathologic\_t, function(x){  
 switch(as.character(x),   
 "T1" = "T1&2",   
 "T2" = "T1&2",   
 "T2a" = "T1&2",   
 "T2b" = "T1&2",   
 "T3" = "T3&4",   
 "T3a" = "T3&4",   
 "T3b" = "T3&4",   
 "T4" = "T3&4",   
 "TX" = NA,   
 "NA" = NA)  
 }),   
 ajcc\_pathologic\_n = sapply(ajcc\_pathologic\_n, function(x){  
 switch(as.character(x),   
 "N0" = "N-",   
 "N1" = "N+",   
 "NX" = NA,   
 "NA" = NA)  
 }),  
 ajcc\_pathologic\_m = sapply(ajcc\_pathologic\_m, function(x){  
 switch(as.character(x),   
 "M0" = "M-",   
 "M1" = "M+",   
 "MX" = NA)  
 }),   
 race = sapply(race, function(x){  
 switch(as.character(x),   
 "not reported" = NA,   
 x)  
 }),   
 vital\_status = sapply(vital\_status, function(x){  
 switch(as.character(x),   
 "Not Reported" = NA,   
 x)  
 }),   
 age = ifelse(age > 60, ">60", "<=60")  
 )

# 查看一下整理好的数据  
dplyr::glimpse(sample\_info)

Rows: 371  
Columns: 13  
$ sample\_type <chr> "Primary Tumor", "Primary Tumor", "Primary Tumo…  
$ days\_to\_collection <int> 70, 5304, 23, 110, 1777, 602, 3538, 28, 1224, 4…  
$ prior\_malignancy <chr> "no", "no", "yes", "no", "no", "no", "no", "no"…  
$ prior\_treatment <chr> "No", "No", "No", "No", "No", "No", "No", "No",…  
$ ajcc\_pathologic\_stage <chr> "I&II", "I&II", "I&II", "I&II", "I&II", "I&II",…  
$ ajcc\_pathologic\_t <chr> "T1&2", "T1&2", "T1&2", "T1&2", "T1&2", "T1&2",…  
$ ajcc\_pathologic\_n <chr> NA, "N-", NA, "N-", "N-", NA, "N-", "N-", "N-",…  
$ ajcc\_pathologic\_m <chr> "M-", "M-", NA, "M-", "M-", NA, "M-", "M-", "M-…  
$ days\_to\_last\_follow\_up <int> 848, NA, 1115, 12, 1900, 898, 2017, 366, 1553, …  
$ age <chr> ">60", ">60", ">60", "<=60", "<=60", "<=60", ">…  
$ race <chr> "white", "white", "white", "asian", "asian", "a…  
$ gender <chr> "female", "female", "male", "male", "male", "ma…  
$ vital\_status <chr> "Alive", "Dead", "Alive", "Dead", "Alive", "Ali…

# 4. 统计样本的基本信息

tbl <- compareGroups( ~ ., data = sample\_info)  
createTable(tbl)

--------Summary descriptives table ---------  
  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
 [ALL] N   
 N=371   
¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯   
sample\_type: Primary Tumor 371 (100%) 371   
days\_to\_collection 1043 (1186) 368   
prior\_malignancy: 371   
 no 336 (90.6%)   
 yes 35 (9.43%)   
prior\_treatment: 371   
 No 369 (99.5%)   
 Yes 2 (0.54%)   
ajcc\_pathologic\_stage: 347   
 I&II 257 (74.1%)   
 III&IV 90 (25.9%)   
ajcc\_pathologic\_t: 368   
 T1&2 275 (74.7%)   
 T3&4 93 (25.3%)   
ajcc\_pathologic\_n: 256   
 N- 252 (98.4%)   
 N+ 4 (1.56%)   
ajcc\_pathologic\_m: 270   
 M- 266 (98.5%)   
 M+ 4 (1.48%)   
days\_to\_last\_follow\_up 781 (741) 281   
age: 370   
 <=60 177 (47.8%)   
 >60 193 (52.2%)   
race: 361   
 american indian or alaska native 2 (0.55%)   
 asian 158 (43.8%)   
 black or african american 17 (4.71%)   
 white 184 (51.0%)   
gender: 371   
 female 121 (32.6%)   
 male 250 (67.4%)   
vital\_status: 370   
 Alive 240 (64.9%)   
 Dead 130 (35.1%)   
¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯

以上统计显示TCGA-LIHC数据中共包含371个肿瘤样本，其中男性250例，女性121例，且有9.43%的肿瘤患者之前得过肿瘤，但是仅0.54%的患者接受过治疗。

# 5. 统计性别之间的临床指标

tbl <- compareGroups(gender ~ ., data = sample\_info)  
createTable(tbl)

--------Summary descriptives table by 'gender'---------  
  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
 female male p.overall   
 N=121 N=250   
¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯   
sample\_type: Primary Tumor 121 (100%) 250 (100%) .   
days\_to\_collection 1222 (1390) 957 (1066) 0.068   
prior\_malignancy: 0.681   
 no 108 (89.3%) 228 (91.2%)   
 yes 13 (10.7%) 22 (8.80%)   
prior\_treatment: 0.547   
 No 120 (99.2%) 249 (99.6%)   
 Yes 1 (0.83%) 1 (0.40%)   
ajcc\_pathologic\_stage: 0.216   
 I&II 77 (69.4%) 180 (76.3%)   
 III&IV 34 (30.6%) 56 (23.7%)   
ajcc\_pathologic\_t: 0.714   
 T1&2 87 (73.1%) 188 (75.5%)   
 T3&4 32 (26.9%) 61 (24.5%)   
ajcc\_pathologic\_n: 0.098   
 N- 79 (96.3%) 173 (99.4%)   
 N+ 3 (3.66%) 1 (0.57%)   
ajcc\_pathologic\_m: 0.103   
 M- 85 (96.6%) 181 (99.5%)   
 M+ 3 (3.41%) 1 (0.55%)   
days\_to\_last\_follow\_up 755 (748) 792 (740) 0.706   
age: 0.232   
 <=60 52 (43.0%) 125 (50.2%)   
 >60 69 (57.0%) 124 (49.8%)   
race: <0.001   
 american indian or alaska native 0 (0.00%) 2 (0.82%)   
 asian 34 (28.8%) 124 (51.0%)   
 black or african american 3 (2.54%) 14 (5.76%)   
 white 81 (68.6%) 103 (42.4%)   
vital\_status: 0.064   
 Alive 70 (57.9%) 170 (68.3%)   
 Dead 51 (42.1%) 79 (31.7%)   
¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯

以上信息显示TCGA-LIHC中性别之间的样本在种族构成上有显著差别，其中女性肿瘤样本有大约68.6%来源于白人，而男性肿瘤样本约亚洲人和白人占比均在40%以上。

# 6. 筛选样本

使用compareGroups还可以在统计的同时对样本进行筛选。 compareGroups提供两种语法对样本进行筛选： 1. subset: 一次只能对单个变量筛选，影响所有变量  
2. selec: 一次只能对多个变量筛选，只影响对应变量

##　subset 我们首先看一下怎么使用subset筛选样本

# 只统计男性样本在大于60和小于60组间临床信息差异  
tbl <- compareGroups(age ~ ., data = sample\_info, subset = gender == "male")  
tbl

-------- Summary of results by groups of 'age'---------  
  
  
 var N p.value method selection   
1 sample\_type 249 . categorical gender == "male"  
2 days\_to\_collection 248 0.361 continuous normal gender == "male"  
3 prior\_malignancy 249 0.001\*\* categorical gender == "male"  
4 prior\_treatment 249 0.498 categorical gender == "male"  
5 ajcc\_pathologic\_stage 235 0.838 categorical gender == "male"  
6 ajcc\_pathologic\_t 248 1.000 categorical gender == "male"  
7 ajcc\_pathologic\_n 174 1.000 categorical gender == "male"  
8 ajcc\_pathologic\_m 182 1.000 categorical gender == "male"  
9 days\_to\_last\_follow\_up 198 0.455 continuous normal gender == "male"  
10 race 242 <0.001\*\* categorical gender == "male"  
11 gender 249 . categorical gender == "male"  
12 vital\_status 248 0.239 categorical gender == "male"  
-----  
Signif. codes: 0 '\*\*' 0.05 '\*' 0.1 ' ' 1

## 6.1 selec

使用selec同时对对个变量筛选

# gender只统计vital\_status是Alive状态的患者，vital\_status只统计男性样本  
tbl <- compareGroups(age ~ ., data = sample\_info,   
 selec = list(gender = vital\_status == "Alive",   
 vital\_status = gender == "male"))

Warning in chisq.test(xx, correct = FALSE): Chi-squared approximation may be  
incorrect  
Warning in chisq.test(xx, correct = FALSE): Chi-squared approximation may be  
incorrect  
Warning in chisq.test(xx, correct = FALSE): Chi-squared approximation may be  
incorrect

tbl

-------- Summary of results by groups of 'age'---------  
  
  
 var N p.value method   
1 sample\_type 370 . categorical   
2 days\_to\_collection 368 0.050\*\* continuous normal  
3 prior\_malignancy 370 0.010\*\* categorical   
4 prior\_treatment 370 1.000 categorical   
5 ajcc\_pathologic\_stage 346 0.219 categorical   
6 ajcc\_pathologic\_t 367 0.381 categorical   
7 ajcc\_pathologic\_n 256 1.000 categorical   
8 ajcc\_pathologic\_m 270 1.000 categorical   
9 days\_to\_last\_follow\_up 281 0.646 continuous normal  
10 race 360 <0.001\*\* categorical   
11 gender 239 0.358 categorical   
12 vital\_status 248 0.239 categorical   
 selection   
1 ALL   
2 ALL   
3 ALL   
4 ALL   
5 ALL   
6 ALL   
7 ALL   
8 ALL   
9 ALL   
10 ALL   
11 vital\_status == "Alive"  
12 gender == "male"   
-----  
Signif. codes: 0 '\*\*' 0.05 '\*' 0.1 ' ' 1

# 7. 总结

comapreGroups提供了三种函数接口，分别是创建统计统计结构，绘制统计表格，以及结果输出。 在上述实例中我们使用了其中的两个函数，分别是compareGroups，这个是这个包的核心函数，用于统计分析的，createTable是用来将统计结果以三线表的格式展示出来。compareGroups还提供了一个简化函数descrTable，该函数融合了上述两个函数的功能，比如实例一就可以直接使用descrTable得到。

descrTable( ~ ., data = sample\_info)

--------Summary descriptives table ---------  
  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
 [ALL] N   
 N=371   
¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯   
sample\_type: Primary Tumor 371 (100%) 371   
days\_to\_collection 1043 (1186) 368   
prior\_malignancy: 371   
 no 336 (90.6%)   
 yes 35 (9.43%)   
prior\_treatment: 371   
 No 369 (99.5%)   
 Yes 2 (0.54%)   
ajcc\_pathologic\_stage: 347   
 I&II 257 (74.1%)   
 III&IV 90 (25.9%)   
ajcc\_pathologic\_t: 368   
 T1&2 275 (74.7%)   
 T3&4 93 (25.3%)   
ajcc\_pathologic\_n: 256   
 N- 252 (98.4%)   
 N+ 4 (1.56%)   
ajcc\_pathologic\_m: 270   
 M- 266 (98.5%)   
 M+ 4 (1.48%)   
days\_to\_last\_follow\_up 781 (741) 281   
age: 370   
 <=60 177 (47.8%)   
 >60 193 (52.2%)   
race: 361   
 american indian or alaska native 2 (0.55%)   
 asian 158 (43.8%)   
 black or african american 17 (4.71%)   
 white 184 (51.0%)   
gender: 371   
 female 121 (32.6%)   
 male 250 (67.4%)   
vital\_status: 370   
 Alive 240 (64.9%)   
 Dead 130 (35.1%)   
¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯

其余的函数就是为了将生成的结果导出，共可以导出一下格式:

* csv
* html
* LaTeX
* PDF
* markdown
* word
* excel

也可以通过以下命令提取其中用于导出的函数。

ls("package:compareGroups") %>% grep("^export2", x = ., value = TRUE)

[1] "export2csv" "export2html"   
[3] "export2latex" "export2latex.cbind.createTable"  
[5] "export2latex.createTable" "export2md"   
[7] "export2pdf" "export2word"   
[9] "export2xls"

今天小编为大家介绍了一款临床信息统计的工具，希望能帮助大家在后续统计样本信息时减少烦恼！ 本公众号由上海萤锐科技有限公司运营，上海萤锐科技专注与单细胞、空间转录组、空间代谢组、空间蛋白质组以及宏基因组方向个性化分析和方案设计，欢迎大家交流咨询。后续我们还会更新更多空间转录组、空间蛋白质组、空间代谢组以及单细胞方面的内容，欢迎关注交流。



# 8. 参考文献

[1] [I. Subirana, H. Sanz, J. Vila, *Journal of Statistical Software* **2014**, *57*, 1](https://www.jstatsoft.org/v57/i12/).