目录

1 R 语言教程！compareGroups 神包制作描述性表一 1

1.1 R 包介绍 1

1.2 R 包安装 1

1.3 R 包参数 2

1 R 语言教程！compareGroups 神包制作描述性表一

第 95 期 R 语言教程！compareGroups 神包制作描述性表一

描述性表 1 在论文写作中占据着开篇起笔的作用。**对所用的数据进行描述和简单分析，为后续的模型构建提供数据可靠性信息**

本期介绍如何使用 compareGroups 神包来快速生成符合学术规范的表 1。并进行包括以下自定义设置：**1. 设置亚组 2. 设置非正态变量使用非参组间检验 3. 设置显示缺失值 4. 设置显示 OR 值 5. 设置使用自定义组间比较方法 6 数据导出**

1.1 R 包介绍

compareGroups 是一个在 CRAN 上可用的 R 包，**它可以生成描述性表格，展示几个变量的均值、标准差、分位数或频率。此外，还会使用适当的测试计算 p 值来检验组间差异。**

通过简单的代码，就能在 R 中生成美观、规范且可直接用于论文发表的描述性表格。这些表格还可以**导出到不同的格式，如 Word、Excel、PDF，或插入到 R-Sweave 或 R-markdown 文档中**。

在手册 https://cran.r-project.org/web/packages/compareGroups/vignettes/compareGroups_vignette.html里提供了非常友好的 R 包教程，描述了 compareGroups 的所有功能，并附有实际示例。

1.2 R 包安装

从 CRAN 中安装 R 包

```
install.packages("compareGroups")
```

或者从 github 安装最新版本

```
library(devtools)
devtools::install_github("isubirana/compareGroups")
```

1.3 R 包参数

1.3.1 查看数据

(不重要可以不看，知道是**包含多种数据类型的数据集即可**) 调用 R 包自带的 regicor 数据，包含 25 个变量。regicor（吉罗纳心脏登记）研究是一项横断面研究，参与者来自西班牙东北部地区。在此研究中，收集了参与者的各种数据集，包括人口统计信息（如年龄和性别）、人体测量数据（如身高、体重和腰围）以及脂质水平（包括总胆固醇和甘油三酯）。此外，参与者还完成了涵盖体育活动和生活质量等领域的问卷。

为了追踪健康结果，研究还收集了关于心血管事件和死亡的数据。这些信息是通过医院和官方登记册及报告，在超过 10 年的时间里获得的。

```
library(compareGroups)
library(bruceR) # 之前有介绍过，方便描述数据
```

```
# 方便起见，我们只分析前十个变量
data("regicor")

regicor <- regicor[,1:10]

str(regicor)
```

```
## 'data.frame':    2294 obs. of  10 variables:
##  $ id     : num  2.26e+03 1.88e+03 3.00e+09 3.00e+09 3.00e+09 ...
##   ..- attr(*, "label")= Named chr "Individual id"
##   .. ..- attr(*, "names")= chr "id"
##  $ year   : Factor w/ 3 levels "1995","2000",..: 3 3 2 2 2 2 2 1 3 1 ...
##   ..- attr(*, "label")= Named chr "Recruitment year"
##   .. ..- attr(*, "names")= chr "year"
##  $ age    : int  70 56 37 69 70 40 66 53 43 70 ...
##   ..- attr(*, "label")= Named chr "Age"
##   .. ..- attr(*, "names")= chr "age"
##  $ sex    : Factor w/ 2 levels "Male","Female": 2 2 1 2 2 2 1 2 2 1 ...
##   ..- attr(*, "label")= chr "Sex"
##  $ smoker : Factor w/ 3 levels "Never smoker",..: 1 1 2 1 NA 2 1 1 3 3 ...
##   ..- attr(*, "label")= Named chr "Smoking status"
##   .. ..- attr(*, "names")= chr "smoker"
##  $ sbp    : int  138 139 132 168 NA 108 120 132 95 142 ...
##   ..- attr(*, "label")= Named chr "Systolic blood pressure"
##   .. ..- attr(*, "names")= chr "sbp"
##  $ dbp    : int  75 89 82 97 NA 70 72 78 65 78 ...
```

```
##   ..- attr(*, "label")= Named chr "Diastolic blood pressure"
##   .. ..- attr(*, "names")= chr "dbp"
##  $ histhtn: Factor w/ 2 levels "Yes","No": 2 2 2 2 2 2 1 2 2 2 ...
##   ..- attr(*, "label")= Named chr "History of hypertension"
##   .. ..- attr(*, "names")= chr "histbp"
##  $ txhtn  : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 1 ...
##   ..- attr(*, "label")= chr "Hypertension treatment"
##  $ chol   : num  294 220 245 168 NA NA 298 254 194 188 ...
##   ..- attr(*, "label")= Named chr "Total cholesterol"
##   .. ..- attr(*, "names")= chr "chol"
```

1.3.2 生成描述性统计表

简单生成一个最简单的描述性统计表，发现定量资料用平均值标准差描述，分类资料用例数和占比描述

```
descrTable( ~ ., data = regicor)
```

```
##
## --------Summary descriptives table ---------
##
## ___________________________________________________________
##                                     [ALL]           N
##                                    N=2294
## ___________________________________________________________
## Individual id                1215817624 (1339538686) 2294
## Recruitment year:                                    2294
##     1995                          431 (18.8%)
##     2000                          786 (34.3%)
##     2005                         1077 (46.9%)
## Age                               54.7 (11.0)       2294
## Sex:                                                 2294
##     Male                         1101 (48.0%)
##     Female                       1193 (52.0%)
## Smoking status:                                      2233
##     Never smoker                 1201 (53.8%)
##     Current or former < 1y        593 (26.6%)
##     Former >= 1y                  439 (19.7%)
## Systolic blood pressure           131 (20.3)        2280
## Diastolic blood pressure          79.7 (10.5)       2280
## History of hypertension:                             2286
```

```
##     Yes                          723 (31.6%)
##     No                          1563 (68.4%)
## Hypertension treatment:                          2251
##     No                          1823 (81.0%)
##     Yes                          428 (19.0%)
## Total cholesterol                219 (45.2)        2193
## ------------------------------------------------------
```

1.3.3 设置分组变量

根据吸烟情况将人群分为三组，同时生成组间比较列（p.overall）。自动使用卡方检验 (分类变量) 和方差分析 (计量资料：两类时等价 t 检验)。

```
descrTable(`smoker`~ ., data = regicor)
```

```
##
## --------Summary descriptives table by 'Smoking status'---------
##
## ___________________________________________________________________________________________________
##                                Never smoker       Current or former < 1y       Former >= 1y      p.
##                                   N=1201                  N=593                   N=439
## ---------------------------------------------------------------------------------------------------
## Individual id            1229013133 (1337342152) 1534618659 (1372769742) 690225475 (1126583145)  <
## Recruitment year:                                                                                 <
##     1995                      234 (19.5%)             109 (18.4%)             72 (16.4%)
##     2000                      414 (34.5%)             267 (45.0%)             77 (17.5%)
##     2005                      553 (46.0%)             217 (36.6%)            290 (66.1%)
## Age                           56.5 (10.8)             50.6 (10.7)            55.3 (10.6)        <
## Sex:                                                                                              <
##     Male                      301 (25.1%)             410 (69.1%)            360 (82.0%)
##     Female                    900 (74.9%)             183 (30.9%)             79 (18.0%)
## Systolic blood pressure       132 (20.5)              128 (19.8)              133 (19.7)         <
## Diastolic blood pressure      79.5 (10.2)             78.8 (11.0)            81.2 (10.8)
## History of hypertension:                                                                          <
##     Yes                       421 (35.1%)             125 (21.2%)            162 (36.9%)
##     No                        777 (64.9%)             464 (78.8%)            277 (63.1%)
## Hypertension treatment:                                                                           <
##     No                        922 (77.9%)             525 (90.2%)            331 (77.2%)
##     Yes                       262 (22.1%)             57 (9.79%)              98 (22.8%)
## Total cholesterol             220 (46.7)              219 (44.7)              214 (42.6)
```

```
## ------------------------------------------------------------------------------------------------
```

1.3.4 删除某些变量不显示

如不希望描述性统计对 Id 和 year 进行描述，直接在 ~ 右侧的. 后使用减号进行删除（如需要的变量比较少，也可以手动写公式一个个加）

```
descrTable(`smoker`~ .-id-year, data = regicor)
```

```
##
## --------Summary descriptives table by 'Smoking status'---------
##
## ___________________________________________________________________________________
##                           Never smoker Current or former < 1y Former >= 1y p.overall
##                              N=1201            N=593             N=439
## -----------------------------------------------------------------------------------
## Age                       56.5 (10.8)       50.6 (10.7)       55.3 (10.6)   <0.001
## Sex:                                                                        <0.001
##     Male                  301 (25.1%)       410 (69.1%)       360 (82.0%)
##     Female                900 (74.9%)       183 (30.9%)        79 (18.0%)
## Systolic blood pressure    132 (20.5)        128 (19.8)        133 (19.7)   <0.001
## Diastolic blood pressure 79.5 (10.2)       78.8 (11.0)       81.2 (10.8)    0.001
## History of hypertension:                                                    <0.001
##     Yes                   421 (35.1%)       125 (21.2%)       162 (36.9%)
##     No                    777 (64.9%)       464 (78.8%)       277 (63.1%)
## Hypertension treatment:                                                     <0.001
##     No                    922 (77.9%)       525 (90.2%)       331 (77.2%)
##     Yes                   262 (22.1%)        57 (9.79%)        98 (22.8%)
## Total cholesterol          220 (46.7)        219 (44.7)        214 (42.6)    0.039
## -----------------------------------------------------------------------------------
```

1.3.5 亚组描述

subset=(逻辑判断) 来挑选出男性患者进行分析

```
descrTable(`smoker`~ .-id-year-sex, data = regicor,
           subset=(sex=="Male"))
```

```
##
## --------Summary descriptives table by 'smoker'---------
##
## ___________________________________________________________________________________
```

```
##                          Never smoker Current or former < 1y Former >= 1y p.overall
##                             N=301             N=410             N=360
## ------------------------------------------------------------------------------------
## Age                      55.0 (11.5)       52.7 (11.0)       56.8 (10.5)   <0.001
## Systolic blood pressure   133 (18.5)        133 (19.0)        136 (19.0)    0.048
## Diastolic blood pressure 81.3 (9.31)       81.2 (10.6)       82.3 (10.4)    0.253
## History of hypertension:                                                   <0.001
##     Yes                   85 (28.4%)       101 (24.8%)       145 (40.3%)
##     No                   214 (71.6%)       306 (75.2%)       215 (59.7%)
## Hypertension treatment:                                                    <0.001
##     No                   248 (83.5%)       357 (88.4%)       263 (75.1%)
##     Yes                   49 (16.5%)        47 (11.6%)        87 (24.9%)
## Total cholesterol         213 (44.0)        221 (41.9)        216 (43.3)    0.061
## ------------------------------------------------------------------------------------
```

1.3.6 自定义设置分组检验方法

设置 mehtod 参数值，如果不设置，默认所有变量符合正态分布。修改 age 为非正态后使用四分位数进行描述，同时使用非参检验进行分组比较

- **参数值为 1：正态分布分析** ：此值强制分析假设行变量遵循正态分布。
- **参数值为 2：连续非正态分析** ：选择此值意味着分析不假设行变量遵循正态分布，将其视为连续但非正态分布的变量。
- **参数值为 3：分类分析**：此值强制分析将行变量视为分类变量，无论其原始类型如何。
- **参数值为 4：Shapiro-Wilks 检验（正态检验）**：使用此值触发 Shapiro-Wilks 检验，以确定变量是否应在正态性假设下进行分析，还是非正态。这对于根据数据做出如何处理每个变量的决策非常有用。

```r
# descrTable(`smoker`~ .-id-year, data = regicor,method = 1) 假定所有变量符合正态分布

# 设置 age 变量为非正态，使用非参检验进行比较
descrTable(`sex`~ .-id-year, data = regicor, method=c(age = 2))
```

```
##
## --------Summary descriptives table by 'Sex'---------
##
## ___________________________________________________________________
##                                  Male            Female      p.overall
##                                 N=1101           N=1193
## -------------------------------------------------------------------
```

```
## Age                          54.0 [46.0;64.0] 55.0 [46.0;64.0]   0.851
## Smoking status:                                                  <0.001
##     Never smoker               301 (28.1%)      900 (77.5%)
##     Current or former < 1y     410 (38.3%)      183 (15.7%)
##     Former >= 1y               360 (33.6%)       79 (6.80%)
## Systolic blood pressure         134 (18.9)       129 (21.2)     <0.001
## Diastolic blood pressure       81.7 (10.2)      77.8 (10.5)     <0.001
## History of hypertension:                                         0.644
##     Yes                        341 (31.1%)      382 (32.1%)
##     No                         755 (68.9%)      808 (67.9%)
## Hypertension treatment:                                          0.096
##     No                         889 (82.5%)      934 (79.6%)
##     Yes                        189 (17.5%)      239 (20.4%)
## Total cholesterol               217 (42.7)       220 (47.4)      0.140
## ------------------------------------------------------------------------
```

1.3.7 不显示标签 label 值

有些数据集自带 label，可以通过 **include.label 设置是否显示标签**

```
descrTable(`sex`~ .-id-year, data = regicor, include.label= FALSE)
```

```
##
## --------Summary descriptives table by 'sex'---------
##
## ___________________________________________________________
##                               Male       Female    p.overall
##                              N=1101      N=1193
## ___________________________________________________________
## age                          54.8 (11.1) 54.7 (11.0)   0.840
## smoker:                                                <0.001
##     Never smoker             301 (28.1%) 900 (77.5%)
##     Current or former < 1y   410 (38.3%) 183 (15.7%)
##     Former >= 1y             360 (33.6%) 79 (6.80%)
## sbp                          134 (18.9)  129 (21.2)    <0.001
## dbp                          81.7 (10.2) 77.8 (10.5)   <0.001
## histhtn:                                               0.644
##     Yes                      341 (31.1%) 382 (32.1%)
##     No                       755 (68.9%) 808 (67.9%)
## txhtn:                                                 0.096
```

```
##     No                      889 (82.5%) 934 (79.6%)
##     Yes                     189 (17.5%) 239 (20.4%)
## chol                        217 (42.7)  220 (47.4)    0.140
## ------------------------------------------------------------
```

1.3.8 设置计量资料用四分位法描述

设置 **Q1 参数和 Q3 参数**设置如何描述非正态连续变量。如果设置成 0 和 1 就是最小值最大值描述

```
descrTable(`sex`~ .-id-year, data = regicor,method = c(age=2),
           Q1=0.025, Q3=0.975)
```

```
##
## --------Summary descriptives table by 'Sex'---------
##
## ___________________________________________________________________________
##                                  Male            Female      p.overall
##                                 N=1101           N=1193
## ___________________________________________________________________________
## Age                         54.0 [36.0;73.0] 55.0 [36.0;73.0]   0.851
## Smoking status:                                                 <0.001
##     Never smoker              301 (28.1%)      900 (77.5%)
##     Current or former < 1y    410 (38.3%)      183 (15.7%)
##     Former >= 1y              360 (33.6%)       79 (6.80%)
## Systolic blood pressure        134 (18.9)       129 (21.2)     <0.001
## Diastolic blood pressure      81.7 (10.2)      77.8 (10.5)     <0.001
## History of hypertension:                                        0.644
##     Yes                       341 (31.1%)      382 (32.1%)
##     No                        755 (68.9%)      808 (67.9%)
## Hypertension treatment:                                         0.096
##     No                        889 (82.5%)      934 (79.6%)
##     Yes                       189 (17.5%)      239 (20.4%)
## Total cholesterol              217 (42.7)       220 (47.4)      0.140
## ___________________________________________________________________________
```

1.3.9 生成 OR 和 HR 值

使用 show.ratio 变量来显示 OR 值，对于变量类型是 time-to-event 变量则输出 HR 值

```
descrTable(`sex`~ .-id-year, data = regicor,show.ratio = TRUE)
```

```
##
## --------Summary descriptives table by 'Sex'---------
##
## ____________________________________________________________________________________
##                                Male       Female           OR        p.ratio p.overall
##                               N=1101      N=1193
## ____________________________________________________________________________________
## Age                          54.8 (11.1) 54.7 (11.0) 1.00 [0.99;1.01]  0.840    0.840
## Smoking status:                                                                <0.001
##     Never smoker             301 (28.1%) 900 (77.5%)       Ref.        Ref.
##     Current or former < 1y   410 (38.3%) 183 (15.7%) 0.15 [0.12;0.19]  0.000
##     Former >= 1y             360 (33.6%) 79 (6.80%)  0.07 [0.06;0.10]  0.000
## Systolic blood pressure      134 (18.9)  129 (21.2)  0.99 [0.98;0.99] <0.001   <0.001
## Diastolic blood pressure     81.7 (10.2) 77.8 (10.5) 0.96 [0.96;0.97] <0.001   <0.001
## History of hypertension:                                                        0.644
##     Yes                      341 (31.1%) 382 (32.1%)       Ref.        Ref.
##     No                       755 (68.9%) 808 (67.9%) 0.96 [0.80;1.14]  0.612
## Hypertension treatment:                                                         0.096
##     No                       889 (82.5%) 934 (79.6%)       Ref.        Ref.
##     Yes                      189 (17.5%) 239 (20.4%) 1.20 [0.97;1.49]  0.086
## Total cholesterol            217 (42.7)  220 (47.4)  1.00 [1.00;1.00]  0.141    0.140
## ____________________________________________________________________________________
```

1.3.9.1 设置 OR 值的 ref 对照

使用 ref 参数设置变量的 ref 对照值。代码所示为把 smoker 的因子 level 为 3（值为"Former>=1y"）的设置为对照组计算其它组的 OR 值

同样作用的函数还有 ref.no 和 ref.y

```
descrTable(`sex`~ .-id-year, data = regicor,
           include.label = FALSE,show.ratio = TRUE,
           ref = c(smoker=3))
```

```
##
## --------Summary descriptives table by 'sex'---------
##
## ____________________________________________________________________________________
##                                Male       Female           OR        p.ratio p.overall
##                               N=1101      N=1193
## ____________________________________________________________________________________
## age                          54.8 (11.1) 54.7 (11.0) 1.00 [0.99;1.01]  0.840    0.840
```

```
## smoker:                                                                    <0.001
##     Never smoker           301 (28.1%) 900 (77.5%) 13.6 [10.4;18.0]  0.000
##     Current or former < 1y 410 (38.3%) 183 (15.7%) 2.03 [1.51;2.75] <0.001
##     Former >= 1y           360 (33.6%) 79 (6.80%)        Ref.        Ref.
## sbp                        134 (18.9)  129 (21.2)  0.99 [0.98;0.99] <0.001   <0.001
## dbp                        81.7 (10.2) 77.8 (10.5) 0.96 [0.96;0.97] <0.001   <0.001
## histhtn:                                                                    0.644
##     Yes                    341 (31.1%) 382 (32.1%)       Ref.        Ref.
##     No                     755 (68.9%) 808 (67.9%) 0.96 [0.80;1.14]  0.612
## txhtn:                                                                      0.096
##     No                     889 (82.5%) 934 (79.6%)       Ref.        Ref.
##     Yes                    189 (17.5%) 239 (20.4%) 1.20 [0.97;1.49]  0.086
## chol                       217 (42.7)  220 (47.4)  1.00 [1.00;1.00]  0.141    0.140
## ------------------------------------------------------------------------------------
```

1.3.10 不显示对照组的描述信息

使用 hide.no 来隐藏某些因子水平的描述。常用来隐藏掉 2 分类变量的否的信息

```
descTable(`sex`~ .-id-year, data = regicor,hide.no = "No")
```

```
##
## --------Summary descriptives table by 'Sex'---------
##
## ____________________________________________________________
##                               Male       Female    p.overall
##                              N=1101      N=1193
## ____________________________________________________________
## Age                        54.8 (11.1) 54.7 (11.0)   0.840
## Smoking status:                                      <0.001
##     Never smoker           301 (28.1%) 900 (77.5%)
##     Current or former < 1y 410 (38.3%) 183 (15.7%)
##     Former >= 1y           360 (33.6%) 79 (6.80%)
## Systolic blood pressure    134 (18.9)  129 (21.2)   <0.001
## Diastolic blood pressure   81.7 (10.2) 77.8 (10.5)  <0.001
## History of hypertension    341 (31.1%) 382 (32.1%)   0.644
## Hypertension treatment     189 (17.5%) 239 (20.4%)   0.096
## Total cholesterol          217 (42.7)  220 (47.4)    0.140
## ____________________________________________________________
```

1.3.11 同时显示总人群的描述

```
descrTable(`sex`~ .-id-year, data = regicor,show.all = TRUE)
```

```
##
## --------Summary descriptives table by 'Sex'---------
##
## ____________________________________________________________________________
##                                [ALL]         Male        Female     p.overall
##                                N=2294       N=1101       N=1193
## ----------------------------------------------------------------------------
## Age                          54.7 (11.0)  54.8 (11.1) 54.7 (11.0)    0.840
## Smoking status:                                                      <0.001
##     Never smoker             1201 (53.8%) 301 (28.1%) 900 (77.5%)
##     Current or former < 1y   593 (26.6%)  410 (38.3%) 183 (15.7%)
##     Former >= 1y             439 (19.7%)  360 (33.6%) 79 (6.80%)
## Systolic blood pressure       131 (20.3)  134 (18.9)  129 (21.2)    <0.001
## Diastolic blood pressure     79.7 (10.5)  81.7 (10.2) 77.8 (10.5)   <0.001
## History of hypertension:                                              0.644
##     Yes                      723 (31.6%)  341 (31.1%) 382 (32.1%)
##     No                       1563 (68.4%) 755 (68.9%) 808 (67.9%)
## Hypertension treatment:                                               0.096
##     No                       1823 (81.0%) 889 (82.5%) 934 (79.6%)
##     Yes                      428 (19.0%)  189 (17.5%) 239 (20.4%)
## Total cholesterol             219 (45.2)  217 (42.7)  220 (47.4)     0.140
## ----------------------------------------------------------------------------
```

1.3.12 结果导出

可导出各种格式,export2xls, export2latex, export2pdf, export2csv, export2md, export2word

{r} # 直接导出到 docx 中，其它函数语法差不多} #export2word(tab_0,file="1 data_summary.docx")

1.3.13 快速可视化

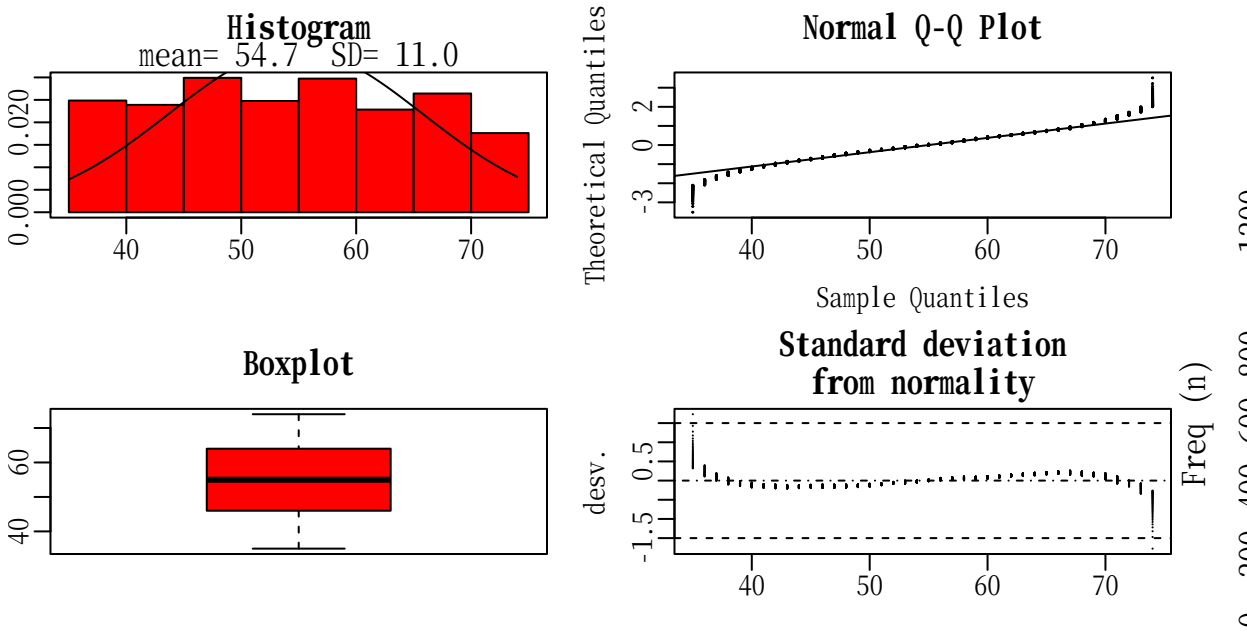
使用 plot 函数进行可视化，设置 file 和 type 参数进行保存例如

plot(res[c(1,2)], file="./figures/univar/", type="png")

```
tab0 <- descrTable(`sex`~ .-id-year, data = regicor)

plot(tab0)
```

Normality plots of 'Age'

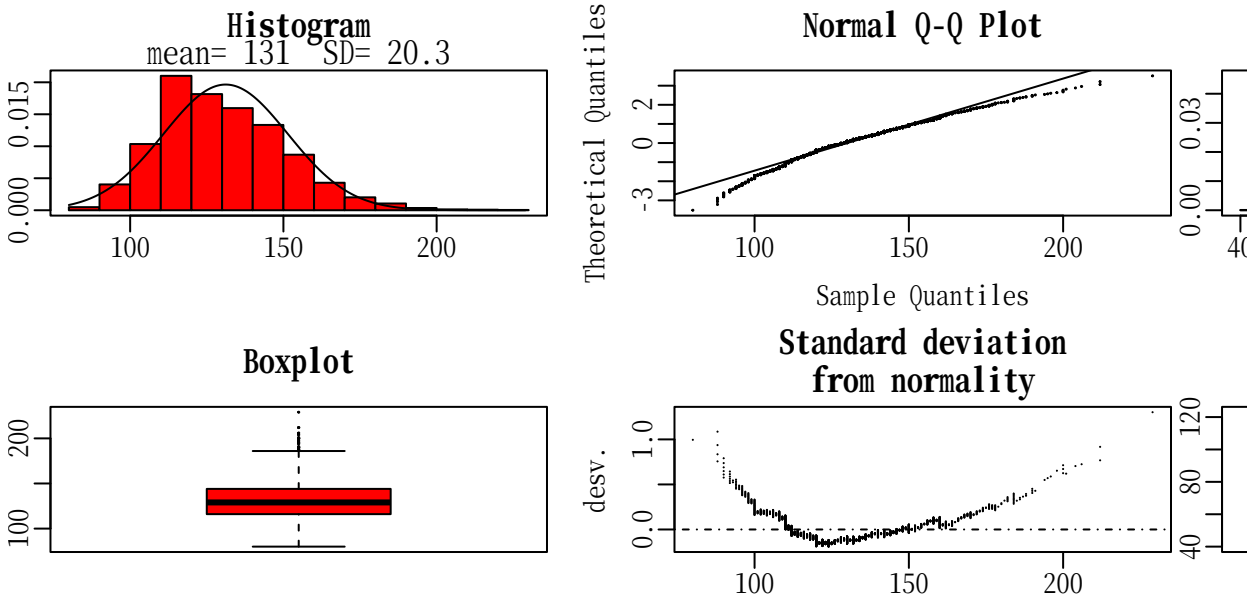


Shapiro-Wilks p-value: <0.001

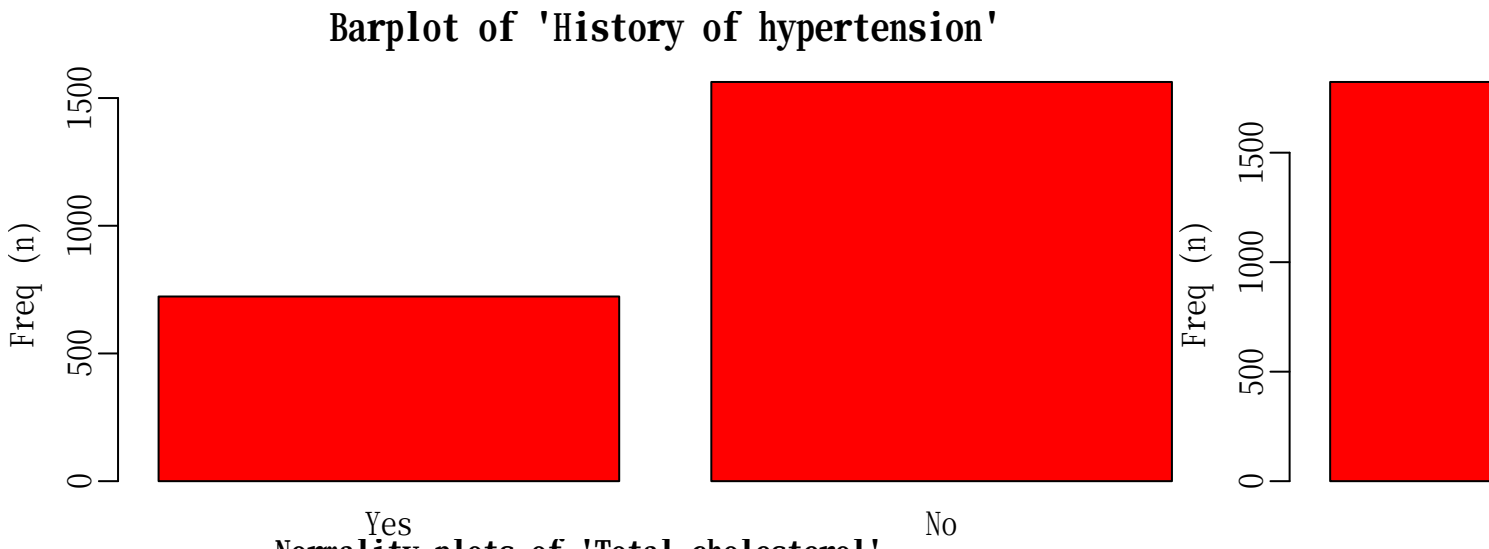
Never s

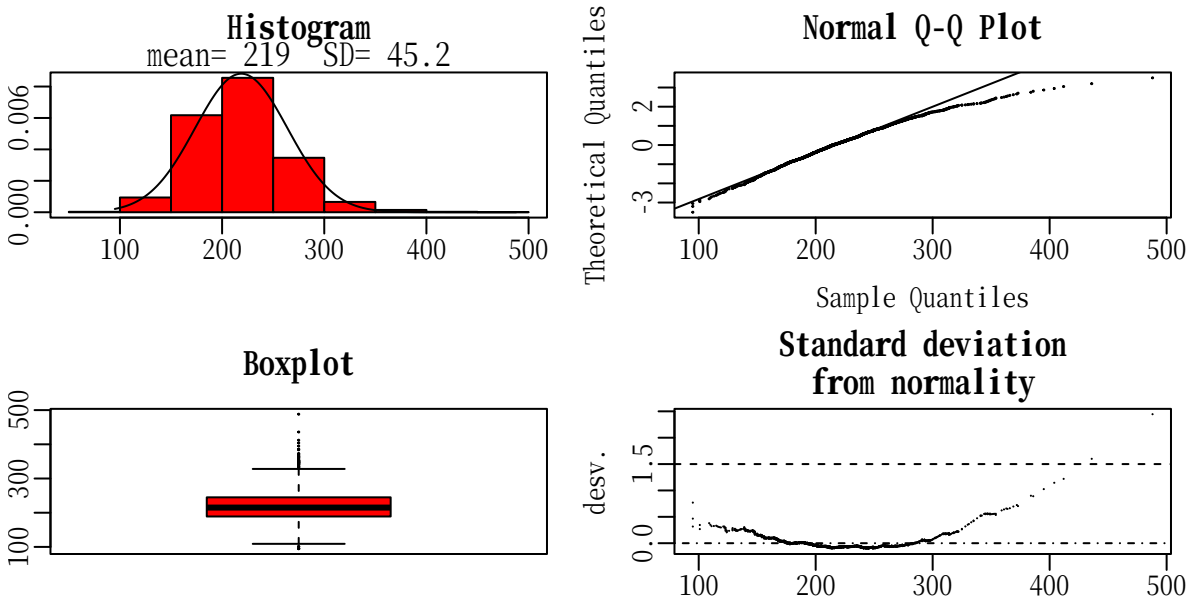
Normality plots of 'Systolic blood pressure'

Normal



Shapiro-Wilks p-value: <0.001





Shapiro-Wilks p-value: <0.001